POLINÔMIOS FRACIONÁRIOS EM MODELOS DE REGRESSÃO

Vinícius Alande

Dissertação apresentada à Universidade Estadual Paulista "Júlio de Mesquita Filho" para a obtenção do título de Mestre em Biometria.

BOTUCATU
São Paulo - Brasil
Março - 2012

POLINÔMIOS FRACIONÁRIOS EM MODELOS DE REGRESSÃO

Vinícius Alande

Orientadora: Prof. Dr. **Luzia A. Trinca**

Dissertação apresentada à Universidade Estadual Paulista "Júlio de Mesquita Filho" para a obtenção do título de Mestre em Biometria.

BOTUCATU
São Paulo - Brasil
Março - 2012

Dedicatória

À Deus pela saúde e força, à minha mãe por ser a base de tudo que sou hoje, aos mestres que passaram pela minha vida pregando o conhecimento e a melhora contínua.

Agradecimentos

À Deus por me dar a vida, mostrar o caminho correto a seguir, e suprir todas as minhas necessidades.

À minha família, Izabel, Aline e Gilberto pelo amor, educação, apoio, incentivo, compreensão nos momentos difíceis, por estarem do meu lado em todos os momentos da minha vida.

À minha orientadora, Profª Luzia Aparecida Trinca, por me mostrar o caminho da ciência, pela confiança depositada, pela forma e excelência que realiza seu trabalho e principalmente pela paciência.

À UNESP, ao programa de Pós-Graduação em Biometria, e ao Departamento de Bioestatística do Instituto de Biociências de Botucatu pela ótima estrutura física que me foi concebida, pelos apoios financeiros para eventos e congressos e pela oportunidade de realização do mestrado.

Ao corpo docente do departamento e funcionários, pela amizade, auxílio, incentivo e pelas horas de descontração no café.

Aos professores das bancas de qualificação e defesa por terem contribuido com sugestões, dicas e correções para o enriquecimento deste trabalho.

Aos amigos pós-graduandos e graduandos que fiz durante esse tempo que, sem dúvida, contribuiram e muito, para o desenvolvimento deste estudo, e para a minha estadia em Botucatu.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo apoio financeiro durante esses anos.

À todas as pessoas que direta ou indiretamente contribuiram para a realização deste trabalho.

"Bem-aventurado o homem que acha sabedoria, e o homem que adquire conhecimento. Porque é melhor a sua mercadoria do que artigos de prata, e maior o seu lucro que o ouro mais fino."

Provérbios 3:13-14

Sumário

Lista de Figuras

Página

Lista de Tabelas

Página

POLINÔMIOS FRACIONÁRIOS EM MODELOS DE REGRESSÃO

Autor: VINÍCIUS ALANDE
Orientadora: Prof. Dr. LUZIA A. TRINCA

RESUMO

Neste trabalho, a flexibilidade dos modelos de polinômios fracionários foi explorada como alternativa quando polinômios simples mostram falta de ajuste em modelos de regressão. Vários tipos de modelos de regressão foram considerados incluindo regressão logística, regressão logística com efeitos aleatórios, e regressão para resposta quantitativa contínua com aumento de variância. Três aplicações para dados biológicos foram realizadas: o exemplo dos rotíferos apresentado em Collett (1991), o estudo da tolerância para temperaturas de células de fungos apresentado em Theodoro et al. (2008) e o estudo da relação entre peso e comprimento de uma espécie de aranhas considerado em Stropa & Trinca (2005). Nos dois primeiros exemplos o modelo de regressão logística foi considerado e o ajuste mostrou sobredispersão dos dados, bem como falta de ajuste dos modelos simples. O uso de polinômios fracionários e a inclusão de efeitos aleatórios nas funções dos preditores lineares mostraram benefícios para ambos os problemas de modelagem. No terceiro exemplo,

a relação entre a variável resposta e a regressora foi não-linear com variância do erro não constante. O uso simultâneo de polinômios fracionários e transformações do tipo Box-Cox resultaram em funções preditivas razoáveis para o problema. A influência de pontos particulares foi explorada e todos os exemplos ilustraram que o processo de modelagem, na prática, requer cuidados nas inspeções das violações do modelo, considerações do problema em particular, e na tomada de decisões.

FRACTIONAL POLYNOMIALS IN REGRESSION MODELS

Author: VINÍCIUS ALANDE
Adviser: Prof. Dr. LUZIA A. TRINCA

SUMMARY

In this work the flexibility of fractional polynomial models were explored as alternative when simple polynomials show lack of fit in regression models. Several types of regression models were considered including logistic regression, mixed logistic regression, and regression for a continuous quantitative response with increasing variance. Three applications to biological data were shown: the rotifers example of Collett (1991); the study of tolerance to temperature of fungus cells of Theodoro et al. (2008); and the study of the relationship between weight and size of a specie of spiders of Stropa & Trinca (2005). In the first two examples the logistic model was considered and overdispersion as well as lack of fit of simple models were detected. The use of fractional polynomials and the inclusion of random effects in the linear predictor function showed benefits to both modeling problems. In the third example the relation was non-linear with nonconstant error variance. The simultaneous use of fractional polynomials and Box-Cox transformation resulted

in very reasonable prediction functions. Influence of particular points were explored. All examples illustrated that the modeling process in practice includes careful inspections of model violations, practical considerations and decisions.

1 INTRODUÇÃO

Várias são as áreas da ciência que adquirem conhecimento por meio da coleta de dados, seja por amostragem em estudos observacionais ou por experimentação. Em estudos observacionais é comum a coleta de dados de diversas variáveis para investigar a inter-relação entre elas e/ou identificar fatores que possivelmente afetam algum evento de interesse. Os estudos experimentais, por serem controlados, permitem a investigação de relação do tipo causa e efeito. Os Modelos de Regressão são ferramentas de grande potencial de aplicação em ambos os tipos de estudos. Por exemplo, na área médica, a partir de um Modelo de Regressão pode-se investigar a relação entre o tempo de sobrevivência de pacientes após determinado tratamento para o câncer e características como a idade, tamanho ou estágio do tumor, local do tumor, entre outros. Nesse contexto, tais variáveis são chamadas de fatores de prognóstico. Os Modelos de Regressão constituem uma classe ampla de modelos estatísticos, com origem no modelo mais simples possível que é aquele que simplifica a relação entre uma variável resposta quantitativa y por uma explicativa x através de uma reta. Essa ideia simples se estendeu muito e hoje abrange modelos com múltiplas variáveis explicativas, termos polinomiais, relações não lineares, respostas discretas, observações com censura, entre outras. Essa ampliação foi possível principalmente pelo desenvolvimento tecnológico que permitiu a construção de computadores e programas computacionais capazes de ajustar quase que qualquer tipo de modelo. Mas é importante lembrar que cada modelo, embora possivelmente útil para explicar ou aproximar um fenômeno desconhecido, tem como base algumas suposições e faz parte da análise de dados a investigação sobre a sua coerência. É de consenso que um modelo "bom" é aquele que explica o problema e é simples, parci-

monioso, produz resultados interpretáveis do ponto de vista do objetivo do estudo, é robusto a pequenas variações dos dados e permite previsões com margem de erro razoável para novos casos. Ou seja, na prática, encontrar um "bom" modelo é um desafio.

Um problema na construção de um modelo é a presença de não linearidade entre a variável resposta e a variável explicativa. Quando não há uma expressão não linear, originada do conhecimento físico do problema, para a relação, as estratégias usuais têm sido a inclusão de termos polinomiais no modelo ou a categorização das variáveis explicativas. A primeira estratégia pode resultar em modelos mais complicados do que o necessário e de difícil interpretação biológica, por exemplo, a inclusão de termos polinomiais de ordem 3 ou 4 pode não ser justificada na prática. A necessidade destes termos pode ser simplesmente devido a escolha "errada" da métrica para medir a variável explanatória (a escolha certa talvez um polinômio de primeira ordem resolvesse o problema). A categorização da(s) variável(ies) explicativa(s) envolve subjetivismo e perda de informação.

O problema da escolha "certa" da métrica tem sido considerado e no caso de apenas uma variável explicativa tem-se as usuais transformações, conforme a forma do gráfico de y versus x, recomendadas nos livros clássicos de regressão linear. No caso mais geral, o problema foi explorado por Box & Tidwell (1962) com a sugestão de uma família de transformações em x análoga à familia de Box-Cox (Box & Cox, 1964) que é aplicada em y. No entanto, a metodologia proposta não ganhou muito espaço nas aplicações devido ao custo computacional. Com o avanço computacional, a ideia de transformações Box & Tidwell (1962) foi estendida e formalizada na forma de uma classe de modelos envolvendo Polinômios Fracionários (PF) por Royston & Altman (1994). Essa classe de modelos inclui termos do tipo x^p em que p é escolhido a partir de um pequeno conjunto de valores inteiros e/ou não-inteiros pré-especificados. As funções polinomiais possíveis, englobam como casos particulares, os polinômios convencionais. A metodologia é aplicável tanto para modelos com erros supostos normais quanto para a classe mais geral de Modelos

Lineares Generalizados e modelos para tempo de sobrevivência.

Este trabalho foi desenvolvido com o objetivo de estudar a metodologia proposta por Royston & Altman (1994), conhecer ferramentas computacionais para o ajuste e explorar a flexibilidade de tais modelos aplicando-os a conjuntos de dados da área biológica. No Capítulo 2 são apresentados alguns conceitos básicos ligados aos métodos de estimação dos parâmetros dos Modelos de Regressão, assim como a apresentação da família de Modelos de Polinômios Fracionários, a metodologia de ajuste e de comparações entre modelos proposta por Royston & Sauerbrei (2008). Um exemplo da literatura foi reproduzido com o objetivo de esclarecimento didático. O capítulo 3 contempla três exemplos. O primeiro exemplo trata de um estudo sobre a densidade de uma espécie de microorganismos do zooplâncton, descrito por Collett (1991). O segundo se refere a um estudo da viabilidade de células do fungo *Paracoccidioides brasiliensis* sob diferentes temperaturas. Nestes dois exemplos foram utilizados um modelo logístico para ajustar os dados. O terceiro exemplo trata do estudo da relação entre o peso e o tamanho de aranhas femêas do tipo *Loxosceles gaucho*. Nesse caso, a relação observada foi não linear e heterogeneidade de variâncias. Explora-se então a aplicação de transformações, tanto na variável resposta como na explicativa a fim de obter um modelo final parcimonioso e coerente com as premissas usuais da modelagem.

2 MODELOS DE REGRESSÃO

Os Polinômios Fracionários (PF) podem ser utilizados em qualquer tipo de Modelo de Regressão com variáveis regressoras quantitativas, cuja parte preditiva envolve preditores lineares. Antes de introduzir a metodologia de PF, uma breve introdução, principalmente visando o estabelecimento da notação a respeito dos Modelos de Regressão, a começar pelos Modelos Lineares com Erros Normais, Modelos Lineares Generalizados (MLG), e em particular, o Modelo de Regressão Logística e o Modelo Binomial Misto.

2.1 Regressão Linear com Erros Normais

Dados n conjunto de valores de k variáveis regressoras $(X_1, \ldots, X_k)$ e uma variável aleatória resposta Y, o modelo de regressão linear múltiplo é escrito como

$$y_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \epsilon_i,$$

ou ainda, na forma matricial,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{1}$$

em que $\mathbf{y}$ é o vetor $(n \times 1)$ de respostas, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ \ldots, \ \beta_k)'$ é o vetor $(q = (k+1) \times 1)$ de parâmetros, $\mathbf{X}$ é a matriz $(n \times (k+1))$ cujas colunas são as variáveis regressoras e $\boldsymbol{\epsilon}$ é o vetor $(n \times 1)$ de erros aleatórios com $E(\boldsymbol{\epsilon}) = \mathbf{0}$. O modelo usual assume que $V(\boldsymbol{\epsilon}) = \mathbf{I}\sigma^2$.

Para ajustar um modelo de regressão aos dados observados deve-se estimar o vetor de parâmetros $\boldsymbol{\beta}$. Existem vários métodos de estimação como, por

exemplo, o de Máxima Verossimilhança (MV) e o de Mínimos Quadrados (MQ). O método MQ dos erros é bastante popular em modelos lineares devido a sua simplicidade mas também pelos estimadores serem equivalentes aos de MV no caso de normalidade e homocedasticidade dos erros. Se dispõe-se de apenas q observações, a determinação dos parâmetros se reduz a um problema matemático de resolução de um sistema de q equações com q incógnitas, não sendo possível fazer qualquer análise estatística. Portanto, deve-se ter $n > q$. Além disso, para obter as estimativas de minímos quadrados dos parâmetros, a matriz $\mathbf{X}'\mathbf{X}$ deve ser não-singular, isto é, seu posto deve ser igual a q. Nessa condição, sua matriz inversa $(\mathbf{X}'\mathbf{X})^{-1}$ existe, e a solução de minímos quadrados para $\hat{\boldsymbol{\beta}}$ é dada por

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

Esse estimador é não viesado, pois $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$, com $V(\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{X})^{-1}\sigma^2$. Pode-se mostrar que este estimador é também de variância mínima entre todos os outros estimadores não viesados, que são funções lineares de $\mathbf{y}$. Sob normalidade dos erros $\widehat{\boldsymbol{\beta}}$ coincide com o estimador de MV que, assintoticamente, segue a distribuição normal, ou seja, $\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, (\mathbf{X}'\mathbf{X})^{-1}\sigma^2)$.

Testes de hipóteses sobre os parâmetros e intervalos de confiança podem ser construídos utilizando-se as estatísticas clássicas do tipo F baseadas na análise de variância e do tipo t usuais (ver, por exemplo, Montgomery et al. (2006) Draper & Smith (2008)).

Na prática, para inferência utilizando o modelo ajustado, é necessário a investigação sobre a validade (ao menos aproximada) sobre as pressuposições do modelo. Violações comuns são a não linearidade e heterogeneidade de variâncias dos erros.

Caso a variável resposta não tenha distribuição normal uma das alternativas é buscar alguma transformação que amenize o problema. A longa experiência em análise de dados tem proporcionado recomendações sobre possíveis transformações de acordo com o tipo de dado, porém Box & Cox (1964) propuseram uma família de transformações que se popularizou como transformação Box-Cox. A

aplicação da metodologia de Modelos Lineares Generalizados, assunto da próxima seção, é uma outra alternativa, porém, na prática nem sempre se adequa aos dados. Sejam $y > 0$ e $\mathbf{X}$ as variáveis resposta e regressoras, respectivamente. A base fundamental da transformação Box-Cox é que existe uma potência y^λ que se relaciona linearmente com uma função de $\mathbf{X}$ de forma que os erros satisfaçam a condição de normalidade e variância constante. O método de transformação consiste na busca do valor para λ que maximiza a função de verossimilhança e é definido por:

$$y(\lambda) = \begin{cases} \frac{y^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \log(y) & \lambda = 0. \end{cases} \tag{2}$$

Encontrado $\hat{\lambda}$ utiliza-se $\mathbf{y}(\hat{\lambda}) = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ como modelo e procede-se da mesma forma para estimação de $\widehat{\boldsymbol{\beta}}$.

2.2 Modelos Lineares Generalizados

Nelder & Wedderburn (1972) propuseram uma teoria unificadora da modelagem estatística a que deram o nome de Modelos Lineares Generalizados (MLG), como uma extensão dos modelos lineares clássicos. Na realidade, eles mostraram que uma série de técnicas comumente estudadas separadamente podem ser reunidas sob o nome de Modelos Lineares Generalizados. Mostraram também que a maioria dos problemas estatísticos que surgem nas mais diversas áreas do conhecimento podem ser formulados, de uma maneira unificada, como nos Modelos de Regressão. Além de Nelder & Wedderburn (1972), na literatura, existem diversos livros clássicos que trata de MLGs, como exemplo, McCullagh & Nelder (1989), e em língua portuguesa pode-se encontrar Demétrio (2002) e Paula (2004). Esses modelos envolvem uma variável resposta e um conjunto de variáveis explicativas, sendo que a variável resposta, componente aleatório do modelo, tem uma distribuição pertencente à família exponencial na forma canônica (distribuição normal, gama e normal inversa, para dados contínuos; binomial para proporções; Poisson e binomial negativa para dados de contagens). As variáveis explicativas entram na forma de uma função

linear (componente sistemático) relacionada aos componentes aleatórios a partir de uma função de ligação, por exemplo, logarítmica para os modelos log-lineares, logito para regressão logística, e outras.

Seja Y uma variável aleatória e associado a ela um conjunto de q variáveis explicativas que como no modelo linear podem ser organizadas numa matriz $\mathbf{X}$. Para uma amostra de n observações os três componentes envolvidos no MLG são descritos a seguir.

1. Componente Aleatório: representado por uma família de variáveis aleatórias independentes $Y_1, Y_2, \ldots, Y_n$, cada uma com função densidade ou função de probabilidades na forma dada por

$$f(y_i; \theta_i, \phi) = \exp\left(\phi[y_i\theta_i - b(\theta_i)] + c(y_i, \phi)\right), \tag{3}$$

chamada de família exponencial. Pode-se mostrar sob as condições usuais de regularidade que

$$E\left(\frac{\partial \ln f(y_i; \theta_i, \phi)}{\partial \theta_i}\right) = 0$$

e

$$E\left(\frac{\partial^2 \ln f(y_i; \theta_i, \phi)}{\partial \theta_i^2}\right) = -E\left[\left(\frac{\partial \ln f(y_i; \theta_i, \phi)}{\partial \theta_i}\right)^2\right],$$

para $\forall_i$, que $E(Y_i) = \mu_i = b'(\theta_i)$, $V(Y_i) = \phi^{-1}V(\mu_i)$, em que V_i é uma função de μ_i dada por $V(\mu_i) = \frac{d\mu_i}{d\theta_i}$ e ϕ^{-1} é o parâmetro de dispersão. V_i é chamada de função de variância, e como depende unicamente da média tem-se que o parâmetro natural θ_i pode ser expresso como

$$\theta_i = \int V_i^{-1} d\mu_i = q(\mu_i)$$

para $q(\mu_i)$ uma função conhecida de μ_i.

A função de variância desempenha um papel importante na família exponencial, uma vez que a mesma caracteriza a distribuição. Isto é, dada a função de

variância, tem-se uma classe de distribuições correspondentes, e vice-versa. Essa propriedade permite a comparação de distribuições por meio de testes simples para a função de variância. Para ilustrar, a função de variância definida por $V(\mu) = \mu(1-\mu)$, $0 < \mu < 1$, caracteriza a classe de distribuições binomiais com probabilidades de sucesso μ ou $1-\mu$.

Exemplificando, para melhorar o entendimento, sejam dois casos particulares; o primeiro assumindo uma variável aleatória com distribuição normal, e o segundo, com distribuição binomial.

Seja Y uma variável aleatória com distribuição normal com média μ e variância σ^2, ou seja, $Y \sim N(\mu, \sigma^2)$. A função densidade de Y é expressa na forma

$$\begin{aligned} f(y;\mu,\sigma^2) &= \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(y-\mu)^2\right) \\ &= \exp\left(\frac{1}{\sigma^2}\left(y\mu - \frac{\mu^2}{2}\right) - \frac{1}{2}(\ln 2\pi\sigma^2 + \frac{y^2}{\sigma^2})\right), \end{aligned}$$

em que $-\infty < \mu, y < \infty$, e $\sigma^2 > 0$. Logo, para $\theta = \mu$, $b(\theta) = \frac{\theta^2}{2}$, $\phi = \sigma^{-2}$ e $c(y,\phi) = \frac{1}{2}\ln\frac{\phi}{2\pi} - \frac{y^2\phi}{2}$, temos (3). Verifica-se facilmente que a função de variância é dada por $V(\mu) = 1$.

Seja Y a proporção de sucessos em n ensaios independentes, cada um com probabilidade de ocorrência μ. Supõe-se que $nY \sim B(n,\mu)$. A função de probabilidade nY é expressa por

$$\begin{pmatrix} n \\ ny \end{pmatrix} \mu^{ny}(1-\mu)^{n-ny} = \exp\left(\ln\begin{pmatrix} n \\ ny \end{pmatrix} + ny\ln\left(\frac{\mu}{1-\mu}\right) + n\ln 1 - \mu\right),$$

em que $0 < \mu, y < 1$. Obtem-se (3) fazendo $\phi = n$, $\theta = \ln\left(\frac{\mu}{1-\mu}\right)$, $b(\theta) = \ln(1+\exp^{\theta})$ e $c(y,\phi) = \ln\begin{pmatrix} n \\ ny \end{pmatrix}$. A função de variância aqui é dada por $V(\mu) = \mu(1-\mu)$.

2. Componente Sistemático ou Preditor Linear: como no modelo (1) as variáveis

explicativas entram na forma de uma combinação linear de seus efeitos

$$\eta_i = \sum_{j=0}^{k} x_{ij}\beta_j = \mathbf{x}_i'\boldsymbol{\beta} \Rightarrow \boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} \tag{4}$$

em que $\mathbf{x}_i = (x_{i0}, x_{i1}, \ldots, x_{ik})'$ é o vetor de covariáveis para a $i-$ésima observação ($x_{i0} = 1$ para $i = 1, 2, \ldots, n$) se o preditor inclui o intercepto, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)'$ o vetor de parâmetros e $\boldsymbol{\eta} = (\eta_1, \eta_2, \ldots, \eta_n)'$ o preditor linear.

3. Função de Ligação: uma função que liga o componente aleatório ao componente sistemático, ou seja, relaciona a média da resposta com o preditor linear, isto é,

$$\eta_i = g(\mu_i)$$

em que $g(.)$ é uma função monótona e diferenciável. Note que $g(.)$ transforma a esperança da variável resposta no preditor linear.

Assim, vê-se que para a especificação do modelo, os parâmetros θ_i da família exponencial não são de interesse direto (pois há um para cada observação) mas sim um conjunto menor de parâmetros $\beta_0, \beta_2, \ldots, \beta_q$ tais que uma combinação linear dos β's seja igual a alguma função do valor esperado de Y_i. Portanto, as decisões importantes na escolha do MLG são a escolha da distribuição da variável resposta, da matriz do modelo e da função de ligação. Se a função de ligação é escolhida de tal forma que $g(\mu_i) = \theta_i$, o preditor linear modela diretamente o parâmetro canônico e tal função de ligação é chamada ligação canônica. Isto resulta, frequentemente, em uma escala adequada para a modelagem com interpretação prática para os parâmetros de regressão, além de vantagens teóricas em termos da existência de um conjunto de estatísticas suficientes (Mood et al., 1974) para os parâmetros β's e alguma simplificação no algoritmo de estimação. A ligação identidade ($\eta = \mu$) para a distribuição normal, a ligação logarítmica ($\eta = \ln \mu$) para a distribuição Poisson, a ligação logística ($\eta = \ln \frac{\pi}{1-\pi}$) para a distribuição binomial, e a ligação recíproca

($\eta = \frac{1}{\mu}$) para a distribuição gama são exemplos de funções de ligação canônicas. A estimação dos parâmetros é feita pelo método da máxima verossimilhança e tem que ser resolvida de forma numérica por métodos iterativos do tipo Newton-Raphson, já que em geral as soluções não apresentam forma fechada.

Para estimar um dado parâmetro β constrói-se a função de verossimilhança que, tomando-se o log resulta em

$$l(\boldsymbol{\beta}, \mathbf{y}) = \log \prod_{i=1}^{n} (\exp\left(\phi[y_i\theta_i - b(\theta_i)] + c(y_i, \phi)\right)) = \sum_{i=1}^{n} \log(f(y_i; \theta_i, \phi)).$$

Derivando-se em relação a cada parâmetros tem-se a função escore que é calculada por

$$\begin{aligned}
\frac{\partial l(\boldsymbol{\beta}; \mathbf{y})}{\partial \beta_j} &= \sum_{i=1}^{n} \phi \left\{ y_i \frac{d\theta_i}{d\mu_i} \frac{d\mu_i}{d\eta_1} \frac{\eta_i}{\beta_j} - \frac{db(\theta_i)}{d\theta_1} \frac{d\theta_i}{d\mu_i} \frac{d\mu_i}{d\eta_i} \frac{d\eta_i}{d\beta_j} \right\} \\
&= \sum_{i=1}^{n} \phi \left\{ y_i V_i^{-1} \frac{d\mu_i}{d\eta_i} x_{ij} - \mu_i V_i^{-1} \frac{d\mu_i}{d\eta_i} x_{ij} \right\} \\
&= \sum_{i=1}^{n} \phi \left\{ \sqrt{\frac{\omega_i}{V_i}} (y_i - \mu_i) x_{ij} \right\},
\end{aligned}$$

em que $\omega_i = (d\mu_i/d\eta_i)^2/V_i$. Logo, pode-se escrever a função escore na forma vetorial

$$\mathbf{U}(\boldsymbol{\beta}) = \frac{\partial L(\boldsymbol{\beta}; y)}{\partial \boldsymbol{\beta}} = \phi \mathbf{X}' \mathbf{W}^{1/2} \mathbf{V}^{-1/2} (\mathbf{y} - \boldsymbol{\mu})$$

em que $\mathbf{X}$ é a matriz do modelo definida em (4), $\mathbf{W} = diag(\omega_1, \omega_2, \ldots, \omega_n)$ é a chamada matriz de pesos e $\mathbf{V} = diag(V_1, V_2, \ldots, V_n)$.

Assim, o processo iterativo de Newton-Raphson para a obtenção da estimativa de máxima verossimilhança de $\boldsymbol{\beta}$ é definido expandindo-se a função escore $\mathbf{U}(\boldsymbol{\beta})$ em torno de um valor inicial $\boldsymbol{\beta}^{(0)}$, tal que

$$\mathbf{U}(\boldsymbol{\beta}) = \mathbf{U}(\boldsymbol{\beta}^{(0)}) + \mathbf{U'}(\boldsymbol{\beta}^{(0)})|(\boldsymbol{\beta} - \boldsymbol{\beta}^{(0)}),$$

em que $\mathbf{U'}(\boldsymbol{\beta}^{(0)})$ denota a primeira derivada de $\mathbf{U}(\boldsymbol{\beta}^{(0)})$ com respeito a $\boldsymbol{\beta}$. Assim, repetindo o procedimento acima, chega-se ao processo iterativo

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} - \mathbf{U'}(\boldsymbol{\beta}^{(m)})^{-1} \mathbf{U}(\boldsymbol{\beta}^{(m)}),$$

$m = 0, 1, \ldots$. Como a matriz $\mathbf{U'(\boldsymbol{\beta})}$ pode não ser positiva definida, pode-se aplicar o método de *scoring* de Fisher, substituindo a matriz $\mathbf{U'(\boldsymbol{\beta})}$ pelo correspondente valor esperado, e assim chegar a um processo iterativo de mínimos quadrados reponderados,

$$\boldsymbol{\beta}^{(m+1)} = (\mathbf{X'W}^{(m)}\mathbf{X})^{-1}\mathbf{X'W}^{(m)}\mathbf{z}^{(m)}, \tag{5}$$

$m = 0, 1, \ldots$, em que $\mathbf{z} = \boldsymbol{\eta} + \mathbf{W}^{-1/2}\mathbf{V}^{-1/2}(\mathbf{y} - \boldsymbol{\mu})$. Note que $\mathbf{z}$ desempenha o papel de uma variável dependente modificada, enquanto $\mathbf{W}$ é a matriz de pesos que muda a cada passo do processo iterativo. A convergência ocorre em um número finito de passos, independente dos valores iniciais utilizados. É usual iniciar o processo iterativo com $\boldsymbol{\eta}^{(0)} = \mathbf{g(y)}$. Apenas para ilustrar, note que para o caso logístico binomial, tem-se $\boldsymbol{\omega} = n\boldsymbol{\mu}(1 - \boldsymbol{\mu})$ e variável dependende modificada dada por $\mathbf{z} = \boldsymbol{\eta} + (\mathbf{y} - n\boldsymbol{\mu})/n\boldsymbol{\mu}(1 - \boldsymbol{\mu})$. Como visto em (2.1), para o Modelo Linear com Erros Normais não é preciso recorrer ao processo iterativo para a obtenção da estimativa de máxima verossimilhança. Nesse caso, $\widehat{\boldsymbol{\beta}}$ assume forma fechada $\hat{\boldsymbol{\beta}} = (\mathbf{X'X})^{-1}\mathbf{X'y}$.

As informações sobre as propriedades assintóticas dos estimadores de máxima verossimilhança dos MLGs podem ser vistas em Fahrmeir & Kaufmann (1985).

Sem perda de generalidade, suponha que o logaritmo da função de verossimilhança seja agora definido por

$$l(\boldsymbol{\mu}; \mathbf{y}) = \sum_{i=1}^{n} l(\mu_i, y_i),$$

em que $\mu_i = g^{-1}(\eta_i)$ e $\eta_i = \mathbf{x}_i'\boldsymbol{\beta}$. Para o modelo saturado ($q = n$) a função $l(\boldsymbol{\mu}; \mathbf{y})$ é estimada por

$$l(\mathbf{y}; \mathbf{y}) = \sum_{i=1}^{n} l(y_i, y_i),$$

ou seja, a estimativa de máxima verossimilhança de μ_i fica nesse caso dada por $\hat{\mu}_i^0 = y_i$. Quando $q < n$, denota-se a estimativa de $l(\boldsymbol{\mu}; \mathbf{y})$ por $l(\hat{\boldsymbol{\mu}}; \mathbf{y})$. A qualidade do ajuste de um MLG é avaliada a partir da *deviance* ou função desvio dada por

$$D^*(\mathbf{y}; \hat{\boldsymbol{\mu}}) = \phi D(\mathbf{y}; \hat{\boldsymbol{\mu}}) = 2(l(\mathbf{y}; \mathbf{y}) - l(\hat{\boldsymbol{\mu}}; \mathbf{y})),$$

que é uma distância entre o logaritmo da função de verossimilhança do modelo saturado (com n parâmetros) e do modelo sob investigação (com q parâmetros) avaliado na estimativa de máxima verossimilhança $\widehat{\boldsymbol{\beta}}$. Um valor pequeno para a função desvio indica que, para um número menor de parâmetros, obtém-se um ajuste tão bom quanto o ajuste com o modelo saturado.

A Análise de Deviance (ANODEV) é uma generalização da Análise de Variância para os modelos lineares generalizados, visando obter, a partir de uma sequência de modelos, cada um incluindo mais termos que os anteriores, os efeitos de fatores, variáveis regressoras e suas interações. Dada uma sequência de modelos encaixados, utiliza-se a *deviance* como uma medida de discrepância do modelo e forma-se uma tabela de diferenças de *deviances*. Esses modelos precisam ter necessariamente a mesma distribuição para a variável resposta e as funções de ligações idênticas para serem comparados.

A ANODEV é realizada com base no teste da razão de verossimilhança que envolve a comparação dos valores do logaritmo da função de verossimilhança maximizada. Quando se tem um vetor de parâmetros, muitas vezes há o interesse no teste de hipótese de apenas um subconjunto deles. Seja, então, uma partição de parâmetros $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$ em que $\boldsymbol{\beta}_1$ de dimensão r é o vetor de interesse e $\boldsymbol{\beta}_2$ de dimensão $(q - r)$ o vetor coeficientes considerados de *nuisance*. Sejam as hipóteses $H_0 : \boldsymbol{\beta}_1 = \boldsymbol{\beta}_{1,0}$ e $H_a : \boldsymbol{\beta}_1 \neq \boldsymbol{\beta}_{1,0}$ sendo $\boldsymbol{\beta}_{1,0}$ um vetor de valores especificados para $\boldsymbol{\beta}_1$. Sabendo que $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\beta}}_2)$ é o estimador de máxima verossimilhança para $\boldsymbol{\beta}$ sem restrição e $\widehat{\boldsymbol{\beta}}_{2,0} = (\boldsymbol{\beta}_{1,0}, \widehat{\boldsymbol{\beta}}_{2,0})$ em que $\widehat{\boldsymbol{\beta}}_{2,0}$ é o estimador de $\boldsymbol{\beta}_2$ sob H_0, tem-se que o teste da razão de verossimilhanças compara o logaritmo da função de verossimilhança maximizada sem restrição, dada por $l(\widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\beta}}_2; \mathbf{y})$ com o valor sob H_0, dado por $l(\boldsymbol{\beta}_{1,0}, \widehat{\boldsymbol{\beta}}_{2,0}; \mathbf{y})$. Esse teste é geralmente preferível no caso de hipóteses relativas a vários coeficientes β's. Se a diferença for grande, então, H_0 é rejeitada. A estatística para esse teste é dada por:

$$\Lambda = -2 \ln \lambda = 2(l(\widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\beta}}_2; y) - l(\boldsymbol{\beta}_{1,0}, \widehat{\boldsymbol{\beta}}_{2,0}; \mathbf{y})),$$

que sob H_0 se distribui assintóticamente de acordo com uma distribuição χ^2_r. Para

amostras grandes, rejeita-se H_0, ao nível α de significância, se $\Lambda > \chi^2_{r,1-\alpha}$. No caso do modelo (1) este teste se reduz ao teste F com o numerador sendo o acréscimo na Soma de Quadrados devido a $\boldsymbol{\beta}_1$, quando $H_0 : \beta_1 = 0$.

Além do teste da razão de verossimilhança existem outros como o teste de Wald e o teste *score* (McCulloch et al., 2008). O teste de Wald é baseado na distribuição normal assintótica de $\widehat{\boldsymbol{\beta}}$ e é uma generalização da estatística t de Student (Wald, 1941). É, geralmente, o mais usado no caso de hipóteses relativas a um único coeficiente β_j. Tem como vantagem em relação ao teste da razão de verossimilhanças, o fato de não haver necessidade de se calcular $\widehat{\boldsymbol{\beta}}_{2,0}$. A estatística para esse teste é dada por

$$\mathbf{W} = (\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{1,0})'(\hat{V}(\widehat{\boldsymbol{\beta}}_1))^{-1}(\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{1,0}),$$

sendo que, para amostras grandes rejeita-se H_0, ao nível α de significância se $\mathbf{W} > \chi^2_{r,1-\alpha}$.

A partir da estatística de teste da razão de verossimilhanças, uma região de confiança para $\boldsymbol{\beta}_1$, com coeficiente de confiança $(1-\alpha)$, inclui todos os valores de $\boldsymbol{\beta}_1$ tais que

$$2(l(\widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\beta}}_2, \mathbf{y}) - l(\boldsymbol{\beta}_1, \widehat{\boldsymbol{\beta}}_{2,1}; \mathbf{y})) < \chi^2_{q,1-\alpha}$$

sendo $\widehat{\boldsymbol{\beta}}_{2,1}$ a estimativa de verossimilhança de $\boldsymbol{\beta}_2$ para cada valor de $\boldsymbol{\beta}_1$ que é testado ser pertencente ou não à região.

Para um parâmetro β_j o intervalo de confiança a $100(1-\alpha)\%$ obtido pela estatística de Wald é dado por $\hat{\beta}_j \pm z_{\frac{\alpha}{2}}\hat{\sigma}_{\hat{\beta}_j}$.

Para os Modelos Lineares e Lineares Generalizados existe uma extensa lista de referência que tratam da checagem das suposições usuais desses modelos, como análise de resíduos e diagnósticos, uma referência clássica é Mosteller et al. (1977) para modelos lineares e Collett (1991) para modelos logísticos.

2.2.1 Regressão Logística

A Regressão Logística é um modelo particular da classe dos Modelos Lineares Generalizados em que a variável resposta Y assume distribuição Binomial ou Bernoulli e a função de ligação é do tipo logito. Esse modelo tem grande aplicação nas mais diversas áreas pois se encaixa bem nos estudos em que a unidade de observação é classificada em duas categorias, genericamente tratadas como sucesso ou fracasso. Seja $Y \sim Binomial(n,\ \pi)$ em que n é o número de unidades, consideradas independentes, de observação a serem classificadas como sucesso e π a probabilidade de se observar sucesso na j-ésima unidade ($j = 1, 2, \ldots, n$). Y representa o número de sucessos em n. Se $n = 1$ tem-se que Y é uma variável binária. Suponha que N realizações ($y_i,\ i = 1, 2, \ldots, N$) de Y são tomadas, sendo que cada realização pode ser tomada sob diferentes condições de forma que π possa variar em i. Na prática é comum n variar em i também. Segue que $E(Y_i) = n_i\pi_i$ e $Var(Y_i) = n_i\pi_i(1-\pi_i)$. Sejam x_{ij} ($i = 1, 2, \ldots, N$ e $j = 1, 2, \ldots, k$) os valores das covariáveis que representam as condições na observação i, que possivelmente afetam π_i. O modelo linear logístico é dado por

$$logit(\pi_i) = log\frac{\pi_i}{1-\pi_i} = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki}.$$

Aplicando a transformação inversa obtém-se

$$\pi_i = \frac{e^{\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\ldots+\beta_k x_{ik}}}{1 + e^{\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\ldots+\beta_k x_{ik}}}.$$

Escrevendo $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_N)'$ tem-se

$$\boldsymbol{\pi} = \frac{e^{\boldsymbol{\eta}}}{1 + e^{\boldsymbol{\eta}}}$$

em que $\eta_i = \sum_{j=0}^{k} \beta_j x_{ij} = \mathbf{x}_i'\boldsymbol{\beta}$, e então $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$ ($x_{i0} = 1$ para qualquer i).

Particularizando a interpretação da relação entre $\boldsymbol{\pi}$ e uma variável explanatória X_j tem-se que ela é sempre sigmóide, crescente se $\beta_j > 0$ e decrescente se $\beta_j < 0$, enquanto que a função $logit(\boldsymbol{\pi})$ é linearmente relacionada a X_j.

Várias quantidades derivadas da expressão do modelo logístico são de grande interesse prático. Por exemplo

$$O_i = \pi_i/(1-\pi_i) = e_i^{\eta}$$

mede a chance de sucesso, cujo termo em Inglês, também bastante usado em Português, é *odds*. A razão de chances ou *odds ratio* é dada por

$$OR_{ii'} = \frac{O_i}{O_{i'}} \frac{\pi_i/(1-\pi_i)}{\pi_{i'}/(1-\pi_{i'})} = e^{\eta_i - \eta_{i'}}$$

que compara a chance do evento de interesse ocorrer sob as duas condições i e i' e é, portanto, interpretada como uma medida relativa do risco do evento ocorrer. Como π_i é função de $\mathbf{x}_i$ convém escrever $\pi_i = \pi(\mathbf{x}_i)$.

Em regressão linear, cada coeficiente da regressão está relacionado à variação em $E(y)$ quando do incremento de uma unidade no valor da variável regressora respectiva, dado todas as demais regressores mantidas constantes.

No modelo logístico a interpretação é similar, ilustrando o caso de apenas uma variável regressora para simplificar, leva ao resultado

$$\begin{aligned} \beta_1 &= logit(\pi(x+1)) - logit(\pi(x)) \\ &= \log(O(x+1)) - \log(O(x)) \\ &= \log\left(\frac{O(x+1)}{O(x)}\right) = \log(e^{\beta_1}). \end{aligned}$$

O mesmo resultado é obtido quando há várias variáveis regressoras e o valor de apenas uma delas é incrementado de uma unidade. Desta forma, tem-se que e^{β_j} representa a razão de chances quando X_j aumenta em uma unidade e os níveis das demais regressoras são mantidos constantes. Assim, β_j representa o $\log[OR(x_j+1;\ x_j)]$.

Os parâmetros $\boldsymbol{\beta}$ do Modelo Logístico podem ser facilmente estimados usando-se o método de máxima verossimilhança. Dados as observações y_i's, a função de verossimilhança é dada por

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{N} \begin{pmatrix} n_i \\ y_i \end{pmatrix} \pi_i^{y_i} (1-\pi_i)^{n_i - y_i}$$

que depende das probabilidades desconhecidas, que por sua vez dependem dos parâmetros $\boldsymbol{\beta}$. Assim, a função de verossimilhança pode ser escrita como uma função de $\boldsymbol{\beta}$. O problema é obter os valores de $\hat{\boldsymbol{\beta}}$ que maximize $L(\boldsymbol{\beta})$ ou equivalentemente $logL(\boldsymbol{\beta})$ $(l(\boldsymbol{\beta}))$. O logaritmo da função de verossimilhança é

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{N} \left\{ \log \begin{pmatrix} n_i \\ y_i \end{pmatrix} + y_i \eta_i - n_i \log(1 + e^{\eta_i}) \right\}.$$

Derivando-se $l(\boldsymbol{\beta})$ com respeito a cada β_j $(j = 0, 1, \ldots, k)$ tem-se $k+1$ equações não lineares

$$\frac{\partial \ln L(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{N} y_i x_{ij} - \sum_{i=1}^{N} n_i x_{ij} e^{\eta_i} (1 + e^{\eta_i})^{-1},$$

$j = 0, 1, \ldots, k$ que igualadas a zero e resolvidas numericamente resultam no estimatidor de $\boldsymbol{\beta}$ $(\widehat{\boldsymbol{\beta}})$. Uma vez obtido $\widehat{\boldsymbol{\beta}}$, as outras quantidades de interesse como $\boldsymbol{\eta}$, $\boldsymbol{\pi}$, e as razões de chances são estimadas substituindo-se os $\hat{\beta}_j$'s nas respectivas expressões. Por exemplo, o preditor linear é obtido por

$$\hat{\eta}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \ldots + \hat{\beta}_k x_{ik},$$

e a probabilidade estimada $\hat{\pi}_i$ é dada por

$$\hat{\pi}_i = \frac{\exp \hat{\eta}_i}{1 + \exp \hat{\eta}_i}.$$

A precisão de $\widehat{\boldsymbol{\beta}}$ é aproximada pela expressão da variância assintótica de estimadores de máxima verossimilhança, dada por

$$\widehat{Var(\widehat{\boldsymbol{\beta}})} = (\mathbf{X}' \widehat{\mathbf{V}} \mathbf{X})^{-1},$$

em que

$$\widehat{\mathbf{V}} = \begin{pmatrix} n_1^{-1} \hat{\pi}_1 (1 - \hat{\pi}_1) & 0 & \ldots & 0 \\ 0 & n_2^{-1} \hat{\pi}_2 (1 - \hat{\pi}_2) & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \ldots & n_N^{-1} \hat{\pi}_N (1 - \hat{\pi}_N) \end{pmatrix}.$$

Testes de hipóteses e construção de intervalos de confiança para $\boldsymbol{\beta}$ são usualmente baseados nas propriedades assintóticas de $\widehat{\boldsymbol{\beta}}$. Por exemplo, o intervalo com $100(1-\alpha)\%$ de confiança para β_j $(j = 0, 1, \ldots, k)$ baseado na estatística de Wald é da forma

$$\hat{\beta}_j \pm z_{\frac{\alpha}{2}} \sqrt{\widehat{Var(\hat{\beta}_j)}}.$$

Similarmente, um intervalo de confiança aproximado para a razão de chances $OR(x_j + 1, x_j)$ é dado por

$$\left[e^{\hat{\beta}_j - z_{\frac{\alpha}{2}} \sqrt{\widehat{Var(\hat{\beta}_j)}}};\ e^{\hat{\beta}_j + z_{\frac{\alpha}{2}} \sqrt{\widehat{Var(\hat{\beta}_j)}}} \right]. \tag{6}$$

2.3 Modelo Binomial Misto

Os modelos estatísticos apresentados nas seções anteriores são funções de parâmetros, ou seja, de coeficientes considerados fixos, e de um termo ou componente aleatório. São genericamente chamados de modelos de efeitos fixos. No entanto existem situações em que a natureza das variáveis explanatórias e/ou a estrutura de aleatorização (sorteio) utilizada no experimento, ou amostragem, impõem ao modelo mais de um termo aleatório. Um modelo linear que é função dos parâmetros (efeitos das regressoras) e de mais de um componente aleatório é chamado de Modelo Misto. Um exemplo simples e clássico é o caso do Modelo Linear Misto para um experimento em blocos em que os blocos são considerados de efeitos aleatórios já que é razoável se pensar que aqueles utilizados no experimento constituem uma amostra aleatória de uma população de blocos possíveis e não se tem interesse no efeito de nenhum deles em particular, apenas são utilizados como forma de controle da heterogeneidade nas respostas. O fato é que a incorporação de efeito aleatório no modelo induz uma estrutura de correlação entre unidades de observação sob uma mesma classificação do fator de efeito aleatório.

No caso do Modelo Binomial, dado π_i, a variância está automaticamente determinada $(Var(y_i) = \pi_i(1 - \pi_i))$. Porém, é comum a análise de dados indicar uma variabilidade maior do que a esperada, fenômeno este chamado de sobredispersão McCulloch et al. (2008), Collett (1991), Demétrio (2002) e Hinde &

Demétrio (1998). Muitas vezes isso ocorre pelo negligenciamento do plano de amostragem ou de delineamento na fase de modelagem. Por exemplo, num experimento em que y_i, a resposta na unidade experimental i, é o número de sucesso em n_i subunidades (ou observações), a rigor o Modelo Binomial não é apropriado já que a hipótese de independência entre as n_i *repetições* é irrealística.

A bibliografia especializada apresenta algumas alternativas para contornar o problema como o Modelo Beta-binomial que assume que π_i's são variáveis aleatórias com distribuição Beta. A flexibilidade desse modelo na incorporação de correlação entre observações que ocorrem agrupadas pode ser vista em Hinde & Demétrio (1998).

O Modelo Binomial Misto é outra alternativa de extensão do modelo de efeitos fixos bastante flexível para modelar dados de contagem na presença de vários fatores que acarretam no possível aumento de variabilidade, incorporando também uma estrutura de correlação entre as observações que naturalmente ocorrem agrupadas.

O Modelo Binomial Misto será introduzido supondo um experimento inteiramente aleatorizado. Deseja-se modelar a proporção de sucessos y_i em n_i na unidade experimental i $(i = 1, 2, \ldots, N)$ em função das condições experimentais $\mathbf{x}_i$. Um modelo razoável é dado por

$$\begin{aligned} E(y_i|b_i) &= \pi_i = \frac{e^{\beta_0+b_i+\beta_1 x_i}}{1+e^{\beta_0+b_i+\beta_1 x_i}} \\ n_i y_i|b_i &\sim Binomial(n_i, \pi_i) \\ b_i &\sim Normal(0, \sigma_b^2). \end{aligned} \tag{7}$$

Esse modelo resulta numa correlação não negativa entre observações do mesmo grupo, ou seja, observações pertencentes à mesma unidade experimental tendem a ser mais parecidas do que observações de unidades experimentais diferentes. Quanto maior o valor de σ_b^2 maior a correlação. O parâmetro σ_b^2 é chamado de componente de variância.

Assim, dado b_i's, os efeitos aletórios das unidades i's, as respostas se-

guem o modelo logístico no qual o intercepto varia com a unidade experimental, ou seja, permite-se heterogeneidade da probabilidade de sucesso nas unidades experimentais dentro de mesma condição experimental $\mathbf{x}$.

No caso de mais de uma variável regressora e mais de um fator com efeito aleatório, para o Modelo Binomial Misto o preditor linear é definido como

$$\boldsymbol{\eta} = \log\left(\frac{\boldsymbol{\pi}}{1-\boldsymbol{\pi}}\right) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b},$$

em que $\mathbf{b}$ é um vetor de efeitos aleatórios as quais as respostas estão sujeitas e $\mathbf{Z}$ é a matriz de colunas indicadoras dos fatores de efeitos aleatórios. Condicionando-se em $\mathbf{b}$, β_j tem a mesma interpretação do modelo de efeitos fixos.

O método de estimação dos parâmetros é o de Máxima Verossimilhança. Para modelos com componente de variância, os parâmetros a serem estimados são o vetor $\boldsymbol{\beta}$ e σ_b^2. Os efeitos aleatórios b_i's são variáveis aleatórias não observáveis. A verossimilhança é construída assumindo-se $\mathbf{b}$ dado e então eliminando-os da expressão tomando-se a integral, ou seja, obtendo-se

$$L(\boldsymbol{\beta}, \sigma^2) = \prod_{i=1}^{N} \int_{-\infty}^{\infty} \prod_{j=1}^{n_i} \frac{e^{(b_i + \mathbf{x}'\boldsymbol{\beta})y_{ij}}}{1 + e^{(b_i + \mathbf{x}'\boldsymbol{\beta})y_{ij}}} \frac{e^{-\frac{b_i^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}} db_i.$$

Na expressão acima y_{ij} é uma variável binária. A expressão deixa claro que a maximização é complicada, mesmo para modelos bem simples não há forma fechada para a integral e aproximações numéricas são utilizadas. Para modelos com apenas um fator aleatório o procedimento de Gauss-Hermite pode ser utilizado. Para modelos mais complicados outros métodos como por exemplo Monte Carlo é recomendado Agresti (2007).

Testes e intervalos de confiança para $\boldsymbol{\beta}$ são obtidos da maneira usual. O teste da razão de verossimilhanças pode ser usado para comparar modelos encaixados. Testes para os componentes de variância são mais difícies. Como σ_b^2 não pode ser negativo, a hipótese $H_0 : \sigma_b^2 = 0$ especifica o valor do parâmetro no limite do espaço paramétrico e portanto o teste da razão de verossimilhanças não é válido.

Para avaliação do ajuste no Modelo Misto destaca-se o cálculo dos resíduos condicionais e marginais que são úteis para detectar violações do modelo.

O resíduo condicional nada mais é do que a diferença entre a resposta observada e sua estimativa condicional, ou seja

$$\mathbf{r}_c = \mathbf{y} - (\mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{Z}\hat{\mathbf{b}}).$$

$\mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{Z}\hat{\mathbf{b}}$ é a estimativa condicional de $\mathbf{y}$, ou seja, dado $\mathbf{b} = \hat{\mathbf{b}}$. O vetor $\hat{\mathbf{b}}$ é a predição dos efeitos aleatórios. Como nos outros modelos, é comum padronizar o resíduo como forma de tornar a investigação mais fácil, Pinheiro & Bates (2000) sugerem usar $r_{ci}/\sqrt{\widehat{V(y_i)}}$.

Para a parte aleatória do modelo, o vetor $\mathbf{b}$ é suposto ter distribuição normal, portanto, o gráfico quantil quantil normal para os valores estimados de $\mathbf{b}$ é útil para checar esta suposição.

2.4 Polinômios Fracionários

Em Modelos de Regressão, a relação entre a variável resposta e uma ou mais regressoras pode ser não linear, que muitas vezes, pode ser aproximada por curvas quadráticas ou cúbicas. A representação dessas curvas pode ser feita incluindo termos de potências nas variáveis regressoras. No entanto, potências de baixa ordem fornecem formas de curvas limitadas e de alta ordem podem não ajustar bem aos dados devido aos valores extremos, assim como, complicar o modelo atrapalhando a interpretação. Royston & Altman (1994) propuseram uma extensão dessa família de curvas chamada de Polinômios Fracionários (PF), cujos termos são restritos a um pequeno conjunto de inteiros e não inteiros pré-definidos. Os termos são selecionados de forma que os polinômios convencionais são um subconjunto dessa família.

Os Modelos de Regressão utilizando PF têm aparecido na literatura de forma variada durante um longo período a partir de uso de regras ditadas pela prática. Royston & Altman (1994) forneceram uma descrição unificada e um grau de formalização para eles. Eles são vantajosos por terem flexibilidade considerável no ajuste de modelos em que não há linearidade entre a resposta e as regressoras ou na parte sistemática de um MLG. Na realidade a ideia surgiu com Box & Tidwell

(1962), porém as dificuldades computacionais da época impediram sua disseminação nas aplicações. A proposta de restringir valores das potências a um conjunto restrito acarreta na facilidade de ajuste utilizando um método padrão.

2.4.1 Polinômios Fracionários para uma Variável Regressora

Nesta secção, considera-se um modelo de regressão com apenas uma variável regressora e positiva. Definindo $x^0 = \log x$, então transformações do tipo x^p para diferentes valores de p provenientes do conjunto $S = \{-2; -1; -0,5; 0; 0,5; 1; 2; 3\}$, podem ser razoáveis para tentar linearizar o modelo. Esse conjunto de valores incluem transformações usuais como $\frac{1}{x}$ e $\log x$ e também o modelo linear ($p = 1$).

Royston & Altman (1994) desenvolveram uma estrutura para essas transformações, definindo funções da forma $\beta_0 + \beta_1 x^p$ com p pertecente a S como polinômios fracionários de grau 1, denotado por PF1. Por exemplo, com p =0,5 a função do modelo é $\beta_0 + \beta_1\sqrt{x}$. Se um polinômio de grau 1 não for suficiente para encontrar a melhor transformação em x, eles definem os polinômios fracionários de grau 2, denotado por PF2, como $\beta_0 + \beta_1 x^{p_1} + \beta_2 x^{p_2}$, com p_1 e p_2 pertencentes ao conjunto S. Naturalmente, o polinômio de grau m, com m inteiro e maior que 1, é denotado por PFm. Os modelos PF1 e PF2 podem gerar uma variedade de curvas bastante atrativas do ponto de vista de aplicações e por esse motivo, será dado mais atenção a essas duas classes de polinômios fracionários.

Um modelo polinômial fracionário de grau 1 é definido como

$$\varphi_1^*(x; p) = \beta_0 + \beta_1 x^p = \beta_0 + \varphi_1(x; p),$$

em que $\varphi_1(x; p) = \beta_1 x^p$.

Para uma transformação de x de grau 2 (PF2) com potências $\mathbf{p} = (p_1, p_2)$ defini-se $x^{\mathbf{p}}$ o vetor linha com

$$x^{\mathbf{p}} = x^{(p_1,p_2)} = \begin{cases} (x^{p_1}, x^{p_2}), & p_1 \neq p_2 \\ (x^{p_1}, x^{p_1}\log x) & p_1 = p_2 \end{cases}.$$

Assim um modelo com vetor de parâmetros de regressão $\boldsymbol{\beta} = (\beta_1, \beta_2)'$ e vetor de potências $\mathbf{p}$ é

$$\varphi_2^*(x; \mathbf{p}) = \beta_0 + \beta_1 x^{p_1} + \beta_2 x^{p_2} = \beta_0 + x^{\mathbf{p}}\boldsymbol{\beta} = \beta_0 + \varphi_2(x; \mathbf{p}).$$

A constante β_0 é opcional e depende do contexto. Por exemplo, β_0 é usado nos modelos com erros normais e em geral nos modelos logísticos, mas não nos modelos de regressão de Cox.

Uma definição geral de uma função PFm com potências $\mathbf{p} = (p_1 \leq p_2 \leq \ldots \leq p_m)$ é facilmente escrita como uma relação de recorrência. Supondo $h_0(x) = 1$ e $p_0 = 0$, tem-se

$$\varphi_m^*(x; \mathbf{p}) = \beta_0 + \varphi_m(x; \mathbf{p}) = \sum_{j=1}^{m} \beta_j h_j(x)$$

em que

$$h_j(x) = \begin{cases} x^{p_j}, & p_j \neq p_{j-1} \\ h_{j-1}(x) \log x & p_j = p_{j-1} \end{cases}$$

para $j = 1, \ldots, m$. Note que φ^* inclui o termo β_0, enquanto a função φ o exclui. Por exemplo, para $m = 2$ e $\mathbf{p} = (-1, 2)$ tem-se $h_1(x) = x^{-1}$, $h_2(x) = x^2$, então o modelo pode ser escrito como $\beta_0 + \beta_1 \frac{1}{x} + \beta_2 x^2$. Para $\mathbf{p} = (2, 2)$ tem-se $h_1(x) = x^2$, $h_2(x) = x^2 \log x$ e o modelo $\beta_0 + \beta_1 x^2 + \beta_2 x^2 \log x$.

A prática mostra que raramente é necessário um polinômio com grau maior que 2 para ajustar uma regressora no Modelo de Regressão.

Os Modelos de Polinômios Fracionários podem ser ajustados via máxima verossimilhança. Fixado um valor da potência, a estimativa dos coeficientes de regressão consiste em encontrar um valor para $\boldsymbol{\beta}$ que maximiza a verossimilhança dos modelos com preditor linear $\beta_0 + x^{\mathbf{p}}\boldsymbol{\beta}$. Para estimar $\mathbf{p}$ o método de MV é aplicado ao conjunto S discreto. O melhor modelo ajustado é aquele cujo valor de $\mathbf{p}$ apresenta maior verossimilhança. Para PF1, oito modelos devem ser ajustados, enquanto 36 modelos são examinados para PF2 (28 se $p_1 \neq p_2$ mais 8 com $p_1 = p_2$).

Suponha que a cada um dos m elementos de $\mathbf{p}$ nos modelos PFm seja permitido variar continuamente no intervalo $(-\infty, \infty)$, em vez dos valores restritos

de S. Então $\beta_0 + \varphi_m(x; \mathbf{p})$ se torna um modelo não linear com $2m+1$ parâmetros (m potências, m coeficientes de regressão mais o intercepto). Seja $D(m, \mathbf{p})$ a deviance de um modelo PFm com potências $\mathbf{p}$, e $D(0)$ a deviance do modelo nulo (só com β_0 no preditor), e seja $\hat{\mathbf{p}}$ a Estimativa de Máxima Verossimilhança (EMV) para $\mathbf{p}$. Sob a hipótese nula $\boldsymbol{\beta} = 0$, a distribuição de $D(0) - D(m, \hat{\mathbf{p}})$ é aproximadamente χ^2 com $2m$ graus de liberdade (Royston e Altman, 1994). Este resultado se aplica assintoticamente a todos os modelos considerados neste estudo.

Seja $\mathbf{p}_{PF}$ a EMV das potências $\mathbf{p}$ restritas ao conjunto S. Então $D(m, \hat{\mathbf{p}}) \leq D(m, \mathbf{p}_{PF})$. A diferença da *deviance* $\Delta D_{PF} = D(0) - D(m, \mathbf{p}_{FP})$ também tem distribuição aproximadamente χ^2 com $2m$ graus de liberdade. Ambler e Royston (2001), a partir de estudos de simulação, comentam que a probabilidade de significância encontrada para o conjunto restrito é superestimada em relação ao valor p para o conjunto de valores reais, mas concluiram que, a versão restrita pode ser uma boa aproximação e traz vantagens no processo de estimação, evitando os problemas de convergência frequentes nos modelos não lineares. Ainda, a versão discretizada está estreitamente ligada à facilidade de interpretação do modelo ajustado.

Por similaridade, a diferença da *deviance* entre os modelos PFm e PF($m-1$) é aproximadamente distribuída como χ^2 com 2 graus de liberdade, sob a hipótese nula que os $\boldsymbol{\beta}$'s adicionais são iguais a zero. Para uma variável X, a comparação entre os modelos PF1 e PF2 requer 2 graus de liberdade, e entre os modelos PF1 e linear requer 1 grau de liberdade.

Um modelo do tipo PF1 está aninhado em um do tipo PF2 no sentido que, para cada modelo PF1 com potências p_1^*, há oito modelos PF2 com potências (p_1^*, p_2). Assim, os procedimentos para comparação de modelos aninhados podem ser aplicados de maneira usual.

Em Modelos de Regressão com Erros Normais, a distribuição F é usada ao invés da distribuição χ^2 para testar hipóteses (comparar modelos), melhorando as aproximações em amostras pequenas.

Intervalos de Confiança (IC) para a curva ajustada podem ser calcu-

lados por aproximações via métodos usuais ignorando o erro de estimação de $\mathbf{p}_{PF}$, ou via métodos de reamostragem como o *bootstrap*.

A escolha do melhor modelo PF1 ou PF2 pelo critério de deviance seria simples. Contudo, ter uma função padrão é importante para dar parcimônia, estabilidade e utilidade geral para as funções selecionadas. Na maioria dos algoritmos que implementam a modelagem via PF, a função padrão é a linear.

Para uma variável explicativa X, Royston e Altman (1994) sugerem o seguinte procedimento para selecionar a curva que melhor ajusta os dados. A variável X é incluida no modelo devido a significância do teste de comparação a um nível pré-estabelecido (ou devido algum outro critério definido a priori), e suas potências são determinadas pelos seguintes passos:

1. Teste o melhor modelo PF2 para X contra o modelo nulo (somente β_0) usando 4 graus de liberdade. Se o resultado do teste for não significativo, pare, concluindo que o efeito de X é não significativo ao nível α. Caso contrário, continue.

2. Teste o melhor modelo PF2 para X contra o modelo linear ($\beta_0+\beta_1 X$) usando 3 graus de liberdade. Se o resultado do teste é não significativo, pare, e o modelo final é uma reta. Caso contrário, continue.

3. Teste o melhor modelo PF2 para X contra o melhor modelo PF1 usando 2 graus de liberdade. Se o resultado do teste é não significativo, pare, e o modelo final é do tipo PF1. Caso contrário, o modelo final é FP2. Fim do procedimento.

O teste no passo 1 é um teste de associação global das respostas com X. O teste no passo 2 examina evidências de não-linearidade. E no passo 3, escolhe-se entre um modelo não-linear simples ou mais complexo. Um procedimento similar pode ser aplicado para modelos com grau m maior que 2. Quanto maior o valor de m mais complexo o modelo. Normalmente, é escolhido $m = 2$ e α =0,05.

Como em qualquer tipo de procedimento de seleção de variáveis ou de modelos, o método proposto está sujeito a probabilidade do erro tipo I ser mais

alta e/ou perda de poder. Ambler & Royston (2001) estudaram esse problema e concluiram que, no geral, o procedimento é conservador rejeitando a hipótese nula quando ela é verdadeira com menor probabilidade do que o nível de significância nominal.

No ajuste dos modelos, devido a problemas computacionais, recomenda-se corrigir a escala e centrar as variáveis quando a ordem de magnitude pode causar problemas numéricos. Além disso, como o modelo PF é definido para $X > 0$ e a variável regressora apresentar valores iguais a zero, sugere-se a adição de uma constante c, de forma que $x + c > 0$.

Nota-se que o procedimento de modelagem proposto pode ser aplicado utilizando qualquer programa ou pacote capaz de ajustar modelos de regressão lineares. Porém, existe um pacote automático desenvolvido por Gareth Ambler e Axel Benner chamado *mfp* (Ambler & Royston (2001) e Royston & Altman (1994)) disponível no software R (R Development Core Team, 2010). Exitem também ferramentas automáticas no SAS (Software, 2008) e STATA (Stata Statistical Software: Release 12., 2011).

2.4.2 Exemplo

Será apresentado, na sequência, um exemplo de Royston & Altman (1994) para esclarecer melhor a ideia da modelagem por Polinômios Fracionários apresentada até aqui. O exemplo tem com base o conjunto de dados identificado por *Whitehall I* que foi tratado por Royston & Altman (1994) e também Royston & Sauerbrei (2008). A análise desses dados foi reproduzida como forma de aprendizado e exploração das ferramentas computacionais disponíveis. Trata-se de um estudo de coorte prospectivo com 17260 servidores civis do governo britânico, cujo objetivo principal foi estudar a associação entre óbito e a pressão arterial sistólica (*sysbp*), num período de 10 anos. Então, para o exemplo, a resposta é binária (óbito no período (1670 casos) ou não) e a explicativa *sysbp* é quantitativa com média observada igual $136,1$ mmHg. A classe de modelos sugerida é a de Regressão Logística que tentará

explicar o logaritmo da chance de óbito por uma função de *sysbp*.

Dado o conjunto de potências possíveis S e considerando PF no máximo grau 2 como sugerem os autores, tem-se 8 modelos ajustados de grau 1 e 36 modelos ajustados de grau 2. Para escolher qual o melhor ajuste, utilizam-se as diferenças de deviances comparando com a função base definida pela a função linear. O melhor modelo é aquele que produz a maior diferença de deviance em relação à função linear.

A Tabela 1 apresenta as diferenças entre as deviances comparadas com a deviance do modelo linear para todos os 8 modelos de PF1 e todos os 36 modelos de PF2 considerando a variável *sysbp* como explicativa. A maior diferença de deviance mostra o melhor ajuste das potências dos modelos de PF1 e PF2. Como pode ser visto, os melhores modelos de PF1 e PF2 tem potências 2 e $(-2, -2)$, respectivamente.

Os passos do procedimento de seleção entre esses dois modelos e os resultados de cada teste são mostrados na Tabela 2. Para escolher o melhor entre os graus 1 e 2 do polinômio fracionário a sequência proposta de comparações é: (1) PF2 versus modelo nulo, resultado é significativo, (2) PF2 versus modelo linear, resultado é significativo, e (3) PF2 versus PF1, resultado também é significativo ao nível de 5% de significância. Portanto, baseado no critério usado, o melhor modelo é do tipo PF2. A Figura 1 mostra a diferença entre os ajustes do modelo PF1(2) e modelo PF2(-2,-2) indicando a flexibilidade do segundo modelo. Salienta-se que, embora o gráfico mostra alguns pontos bastante destacados dos demais, o logaritmo da *odds* de óbito observado foi calculado com base no número de óbitos dividido pelo número total de servidores para cada valor de *sysbp*. Em alguns casos, o número de réplicas foi bem baixo fazendo com que tais estimativas sejam bastante incertas.

Como padrão, os softwares disponíveis hoje para a técnica de PF automatizada (SAS, Stata e R) tranformam, se apropriado, os valores observados das covariáveis dividindo todos por uma constante para obter estabilidade numérica.

No exemplo, como o melhor modelo indicou PF2(-2,-2), para ajustar o

Tabela 1: Estimativas dos parâmetros do PF para os modelos tentativos e suas respectivas diferenças de deviances (D)

PF1		PF2								
p	D	p_1	p_2	D	p_1	p_2	D	p_1	p_2	D
-2	-79,19	-2	-2	26,22	-1	1	12,97	0	2	7,05
-1	-43,15	-2	-1	24,43	-1	2	7,80	0	3	3,74
-0,5	-29,40	-2	-0,5	22,80	-1	3	2,53	0,5	0,5	10,95
0	-17,37	-2	0	20,72	-0,5	-0,5	17,93	0,5	1	9,51
0,5	-7,45	-2	0,5	18,23	-0,5	0	16,00	0,5	2	6,80
1	0,00	-2	1	15,38	-0,5	0,5	13,93	0,5	3	4,41
2	6,43	-2	2	8,85	-0,5	1	11,77	1	1	8,46
3	0,98	-2	3	1,63	-0,5	2	7,39	1	2	6,61
		-1	-1	21,62	-0,5	3	3,10	1	3	5,11
		-1	-0,5	19,78	0	0	14,24	2	2	6,44
		-1	0	17,69	0	0,5	12,43	2	3	6,45
		-1	0,5	15,41	0	1	10,61	3	3	7,59

Modelo de Regressão Logística são requeridas duas covariáveis: X^{*-2} e $X^{*-2}\log X$, em que $X^* = sysbp/100$. Ainda, para facilidade de interpretação recomenda-se centrar as variáveis na média, que no caso, se equivale a obtenção de $(136, 1/100)^{-2} = 0,5399$ e $(136, 1/100)^{-2}\log(136, 1/100) = 0,1664$. Fazendo, então, $X_1 = X^{*-2} - 0,5399$ e $X_2 = X^{*-2}\log X^* - 0,1664$, tem-se que o preditor linear do modelo logístico é dado por $\beta_0 + \beta_1 X_1 + \beta_2 X_2$. Os parâmetros estimados, erros padrão e a matriz de variância-covariância das estimativas são apresentados na Tabela 3. A constante $\hat{\beta_0} = -2,388$ fornece o valor preditivo do logaritmo da chance de morte para um indivíduo da amostra que tenha pressão média (136, 1mmHg), e fornece a probabilidade de óbito estimada de $\exp(-2,388)/(1+\exp(-2,388)) = 0,084$.

Ilustra-se, a seguir, o cálculo da variância do erro padrão do valor de predição $\hat{\eta}$ deste modelo, ignorando no cálculo o fato de que as potências do PF são

Tabela 2: Aplicação dos procedimentos de testes para selecionar o melhor modelo

Modelo	deviance	$\hat{\mathbf{p}}$	Comparação	D	Valor p
PF2	10641,17	-2;-2	PF2 vs Nulo	33,57	$< 0{,}001$
			PF2 vs Linear	26,22	$< 0{,}001$
PF1	10660,96	2	PF2 vs PF1	19,79	$< 0{,}001$
Linear	10667,39	1			
Nulo	10973,74				

Tabela 3: Estimativas dos coeficientes e matriz de variância e covariância para o modelo logístico com preditor linear do tipo PF2 para pressão arterial sistólica

Parâmetro	Estimativa	Erro padrão (EP)	Matriz de Var-Cov		
$\hat{\beta}_1$	-5,433	0,321	0,103		
$\hat{\beta}_2$	-14,300	1,317	0,374	1,734	
$\hat{\beta}_0$	-2,388	0,032	0,005	0,024	0,001

estimadas, onde $\hat{\eta}$ estima o logaritmo da chance de um indivíduo morrer dado o valor da covariável. A variância de $\hat{\eta}$ com (X_1, X_2) é calculada de acordo com

$$\begin{aligned} V(\hat{\eta}(x_1, x_2)) &= V(\hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2) \\ &= V(\hat{\beta}_0) + V(\hat{\beta}_1)X_1^2 + V(\hat{\beta}_2)X_2^2 + 2Cov(\hat{\beta}_0, \hat{\beta}_1)X_1 \\ &\quad +2Cov(\hat{\beta}_0, \hat{\beta}_2)X_2 + 2Cov(\hat{\beta}_1, \hat{\beta}_2)X_1X_2. \end{aligned}$$

Por exemplo, com $X = 150$mmHg então $X^* = 1,5$, $X_1 = -0,095$, $X_2 = 0,013$, $\hat{\eta}(150) = -2,066$. A partir da equação acima, temos $V(\hat{\eta}(150)) = 0,000936$, $EP(\hat{\eta}(150)) = 0,03$. A probabilidade estimada de óbito é $\exp\hat{\eta}/(1 + \exp\hat{\eta}) = 0,1124$ e o intervalo de confiança aproximado é dado por

$$\frac{\exp\left(\hat{\eta} \pm 1,96EP(\hat{\eta})\right)}{1 + \exp\left(\hat{\eta} \pm 1,96EP(\hat{\eta})\right)} = (0,1066; 0,1185).$$

Finalmente pode-se calcular a estimativa *odds ratio* (razão de chances)

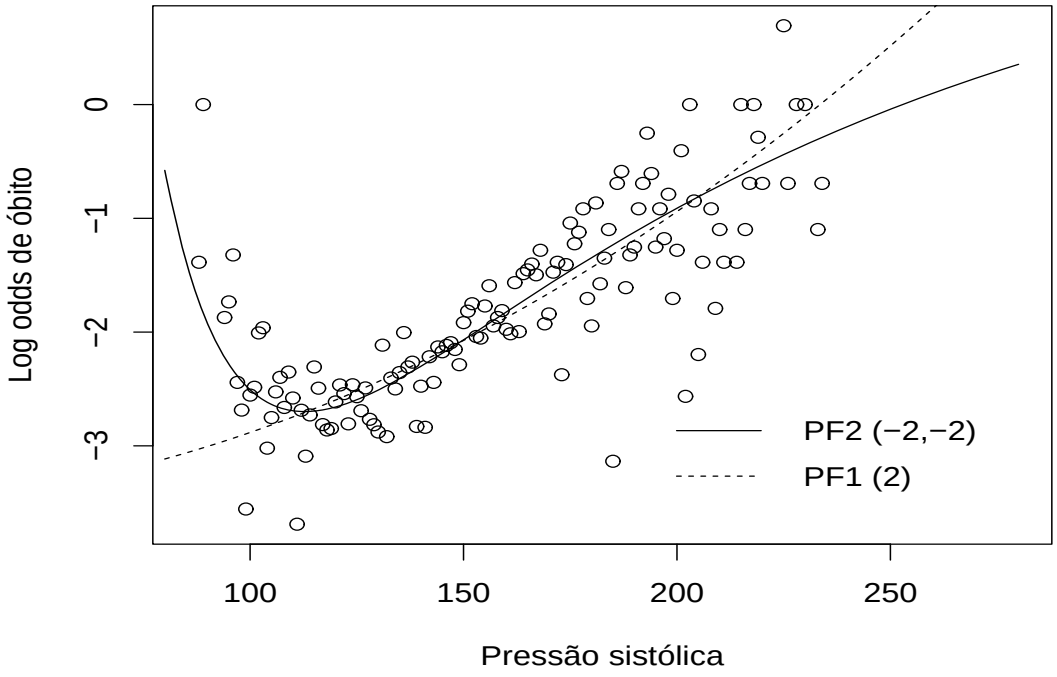

Figura 1 - Modelos ajustados por PF1(2) e PF2(-2,-2).

e seu intervalo de confiança para comparar um dado valor de X com um valor referência X^{ref}. Portanto,

$$\begin{aligned} LogOR &= (\hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2) - (\hat{\beta}_0 + \hat{\beta}_1 X_1^{ref} + \hat{\beta}_2 X_2^{ref}) \\ &= \hat{\beta}_1 (X_1 - X_1^{ref}) + \hat{\beta}_2 (X_2 - X_2^{ref}), \end{aligned}$$

onde $X_1^{ref} = X_{ref}^{-2} - 0,5399$, $X_2^{ref} = X_{ref}^{-2} \log X_{ref}^{*} - 0,1664$, $X_{ref}^{*} = X_{ref}/100$. Sua variância estimada é

$$\begin{aligned} V(LogOR) &= V(\hat{\beta}_1)(X_1 - X_1^{ref})^2 + V(\hat{\beta}_2)(X_2 - X_2^{ref})^2 \\ &\quad +2Cov(\hat{\beta}_1, \hat{\beta}_2)(X_1 - X_1^{ref})(X_2 - X_2^{ref}). \end{aligned}$$

Usando, por exemplo, $X = 150$mmHg e $X_{ref} = 105$mmHg, encontramos $X_1^{ref} = 0,3671$, $X_2^{ref} = -0,1221$, $logOR = 0,5691$, $V(logOR) = 0,007$,

$EP(logOR) = 0,08$. A razão de chances é dada por $\exp(0,5691) = 1,77$ com intervalo de confiança dado por $(1,50; 2,08)$.

Esta seção teve o objetivo de ilustrar a forma de seleção dos modelos proposta por Royston & Altman (1994) e a estimação dos parâmetros. Porém, lembra-se que realização do estudo de diagnóstico do modelo e análise de resíduo são fundamentais para colaborar com a seleção do modelo. Nas aplicações a seguir será feito um estudo mais detalhado desse assunto.

3 APLICAÇÕES

Para ilustrar a flexibilidade da técnica de Polinômios Fracionários, será apresentado neste capítulo três exemplos.

O primeiro é um exemplo descrito em Collett (1991), em que micro-organismos foram estudados para detectar em qual densidade relativa eles ficariam em suspensão em tubos com uma determinada solução. Para isso foi utilizado a metodologia da Regressão Logística usual, Regressão Logística transformada pelos PFs e o Modelo Binomial Misto.

O segundo exemplo trata de um estudo realizado com fungos, cujo objetivo foi estudar a tolerância de células cultivadas a diferentes temperaturas. Como no exemplo anterior, o modelo mais adequado encontrado foi o Modelo Binomial Misto transformado pelo PF.

O terceiro exemplo traz uma aplicação de PF a dados de comprimentos e pesos de aranhas fêmeas coletados no Jardim Botânico de Botucatu, o objetivo principal foi estudar a relação entre essas duas características e modelar o peso em função do tamanho. A modelagem envolveu transformações do tipo Box-Cox na variável peso e de PF na variável regressora.

3.1 Exemplo 1: Estimação da densidade de Rotíferos

Os dados deste experimento foram apresentados em Collett (1991) que utilizou um Modelo Binomial Misto devido à sobredispersão. O experimento envolveu duas espécies de microorganismos aquáticos, pertencentes ao filo *Rotifera*, comuns no zooplâncton de água doce. O objetivo foi determinar as densidades relativas de

cada espécie. Aqui será considerada apenas a espécie *Keratella cochlearis*.

As densidades relativas dos rotíferos foram obtidas usando um método indireto. Certas quantidades de microorganismos foram centrifugadas em tubos com soluções de diferentes densidades conhecidas. O número de indivíduos centrifugados por tubo variou. Ao final do experimento contou-se o número de indíviduos em suspensão em cada tubo. A densidade relativa da espécie é estimada pela densidade da solução que proporciona 50% dos animais em suspensão, ou seja, a quantidade conhecida por LD_{50}. Portanto considerando a densidade como variável regressora e a resposta binária (rotífero em suspensão ou não) foram ajustados modelos logísticos com preditor linear do tipo

$$\eta = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \varphi_m(x_i, \mathbf{p}),$$

em que π_i é a probabilidade de um rotífero ficar em suspensão na densidade x_i.

Embora a natureza do experimento leva diretamente à proposta de um Modelo Misto, aqui será apresentado uma sequência de estratégias que são comumente utilizadas na prática. Vários modelos foram ajustados, desde o mais simples até o Modelo Misto com variável regressora transformada por PF. O primeiro deles, um Modelo Logístico com efeito linear de X, ou seja, $m = 1$ e $p = 1$ (Modelo Original), resultou na equação

$$\hat{\eta} = -114{,}35(4{,}03) + 108{,}75(3{,}86)X,$$

em que os valores entre parênteses são os erros-padrão. Esse ajuste mostrou sobredispersão (*deviance*=300,19 e 18 graus de liberdade), possíveis pontos influentes e valores altos de resíduos como pode ser visto na Figura 2. Portanto considera-se que o ajuste não está adequado. O gráfico da esquerda da Figura 2 indica falta de ajuste no preditor linear.

O segundo modelo ajustado foi o logístico considerando a técnica de PF para tentar melhorar a falta de ajuste. O modelo encontrado é um modelo de segundo grau ($m = 2$) com a melhor transformação $p_1 = p_2 = 3$, referido como

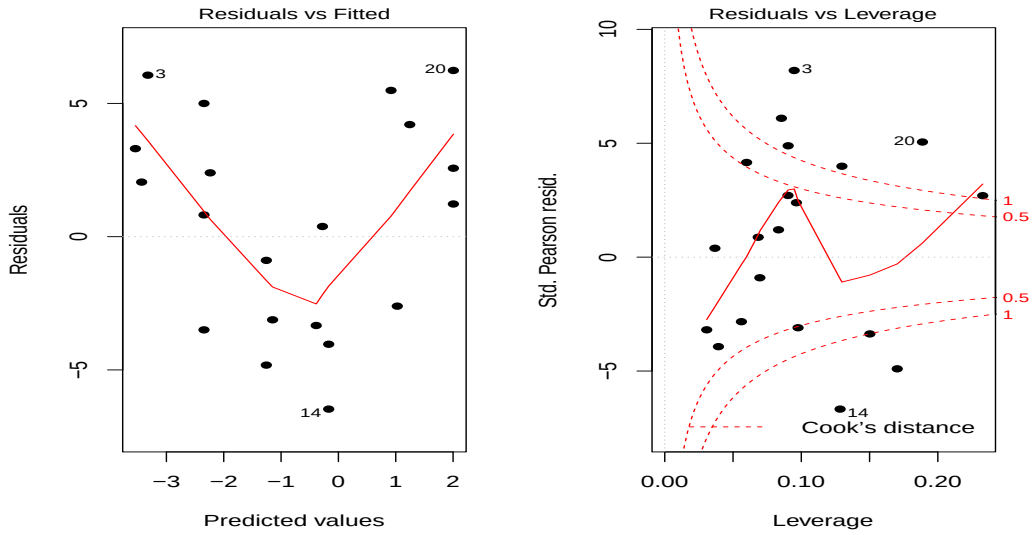

Figura 2 - Análise de resíduos para o Modelo Original, Exemplo 1.

Modelo PF, cuja expressão do preditor linear ajustado é

$$\hat{\eta} = 701{,}17(63{,}58) - 701{,}37(63{,}18)X^3 + 1947{,}84(168{,}61)X^3 \log X. \tag{8}$$

Esse modelo apresentou *deviance* de 143,19 e 15 graus de liberdade. A *deviance* praticamente diminuiu pela metade, porém, não o suficiente para resolver o problema de sobredispersão. O ajuste ainda apresenta alguns valores altos de resíduos para algumas densidades, e indica possíveis pontos influentes nas estimativas como pode ser visto na Figura 3.

A Figura 4 mostra a flexibilidade do Modelo PF para explicar a relação entre a probabilidade de suspensão dos rotíferos e a densidade relativa, já que a curva ajustada acompanha melhor os pontos do que a curva ajustada pelo modelo linear.

Como o planejamento do experimento já indicava a necessidade de inclusão de efeito aleatório para cada tubo (unidade experimental), fato também apontado pelas análises preliminares que indicaram variação extra-binomial nos dados, dois Modelos Mistos foram ajustados, um com X sem transformar e o outro incorporando a técnica de PF no Modelo Binomial Misto. A equação ajustada para

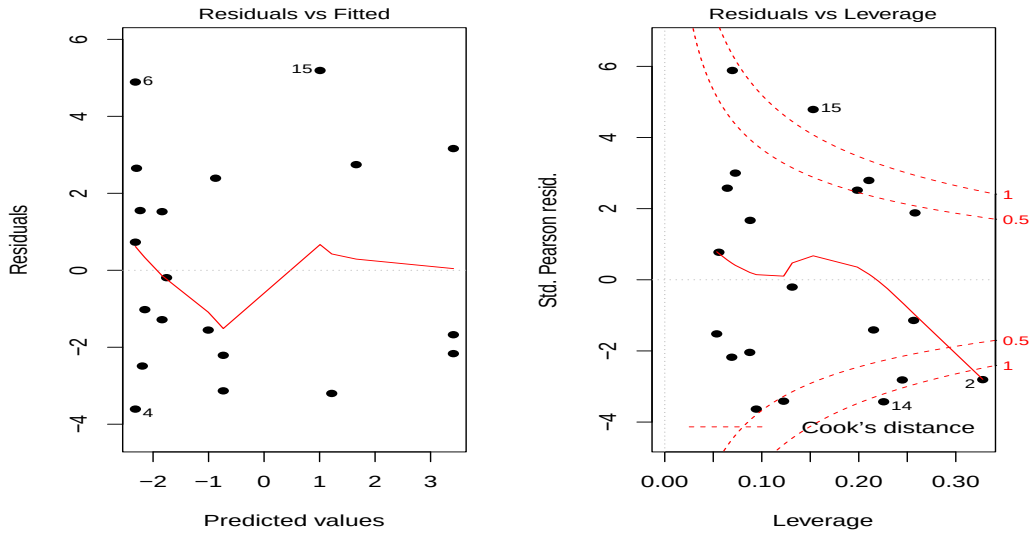

Figura 3 - Análise de resíduos para o Modelo PF, Exemplo 1.

a parte fixa do Modelo Misto mais simples foi

$$\hat{\eta} = -123{,}68(16{,}33) + 117{,}93(15{,}64)X, \tag{9}$$

com componente de variância estimada igual a $1,21$. A *deviance* resultante deste modelo é 76,43 com 17 graus de liberdade. Na análise de resíduos para este modelo (Figura 5) percebe-se uma observação com resíduo bem maior do que as demais assim como uma tendência de curvatura em função da densidade (X). A suposição de normalidade dos efeitos aleatórios parece razoável, porém com um ponto bem longe do esperado.

Transformações em X também podem ser usadas no Modelo Misto. Intuitivamente espera-se que a melhor transformação encontrada para o modelo só com efeitos de X seja também a melhor para o Modelo Misto. No entanto, a metodologia de busca do polinômio proposta por Royston & Sauerbrei (2008) pode ser estendida para o Modelo Misto. Para este exemplo, essa busca levou às mesmas potências encontradas no Modelo PF. Essa tranformação é de segundo grau, com

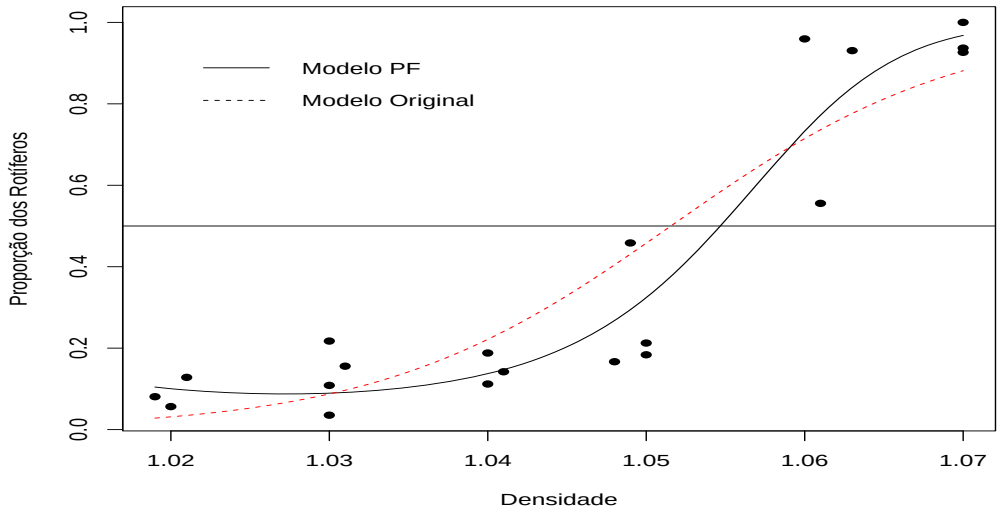

Figura 4 - Curvas das probabilidades ajustadas pelo Modelo Original e pelo Modelo PF, Exemplo 1.

$p_1 = p_2 = 3$. A estimativa da parte fixa desse modelo é

$$\hat{\eta} = 690{,}2(183{,}2) - 690{,}6(181{,}9)X^3 + 1923{,}3(482{,}1)X^3 \log X, \tag{10}$$

e a componente de variância estimada é igual a 0,5654. A transformação ajudou a reduzir um pouco mais a *deviance* que passou de 76,43 com 17 graus de liberdade para 63,99 com 14 graus de liberdade. Os gráficos de resíduos (Figura 6) deste modelo dado a transformação na variável regressora não levanta muitas preocupações com relação à qualidade do ajuste. Vale notar que o efeito de X sobre a resposta é muito grande e portanto, todos os modelos o detecta sem problemas. No entanto, dado o objetivo do experimento ser o de estimar a LD_{50}, conclui-se que o melhor modelo é aquele que acompanha melhor os dados e nesse caso, como mostra a Figura 4, a transformação PF captura bem o comportamento. Para ilustrar as diferenças, a estimativa da LD_{50} pelo modelo mais simples foi igual a 1,0514 com o intervalo de 95% de confiança estimado em $(1,0506; 1,0526)$ e pelo modelo mais completo igual a 1,0541 com intervalo de confiança de $(1,0537; 1,0556)$.

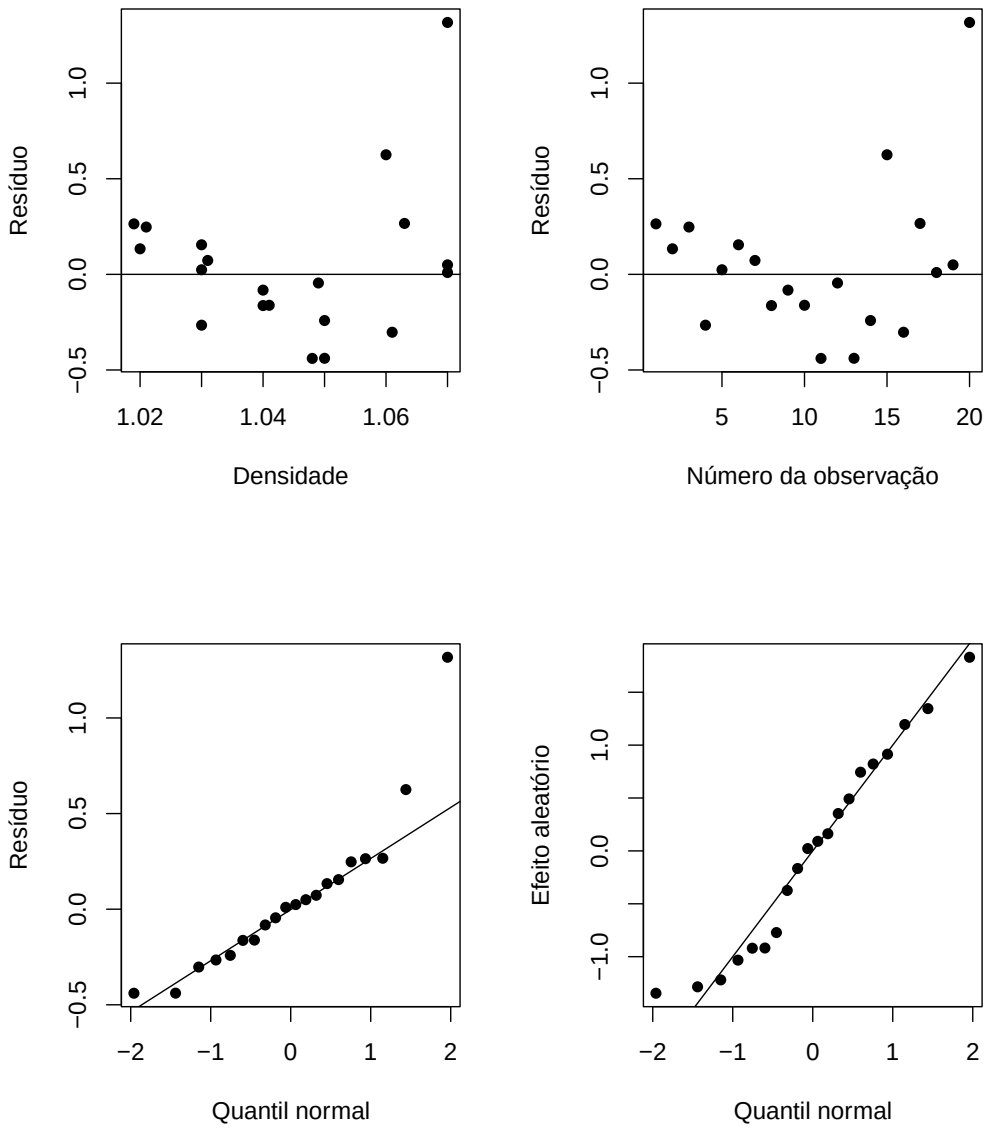

Figura 5 - Análise de Resíduos para o Modelo Binomial Misto com X linear, dado por (9), Exemplo 1.

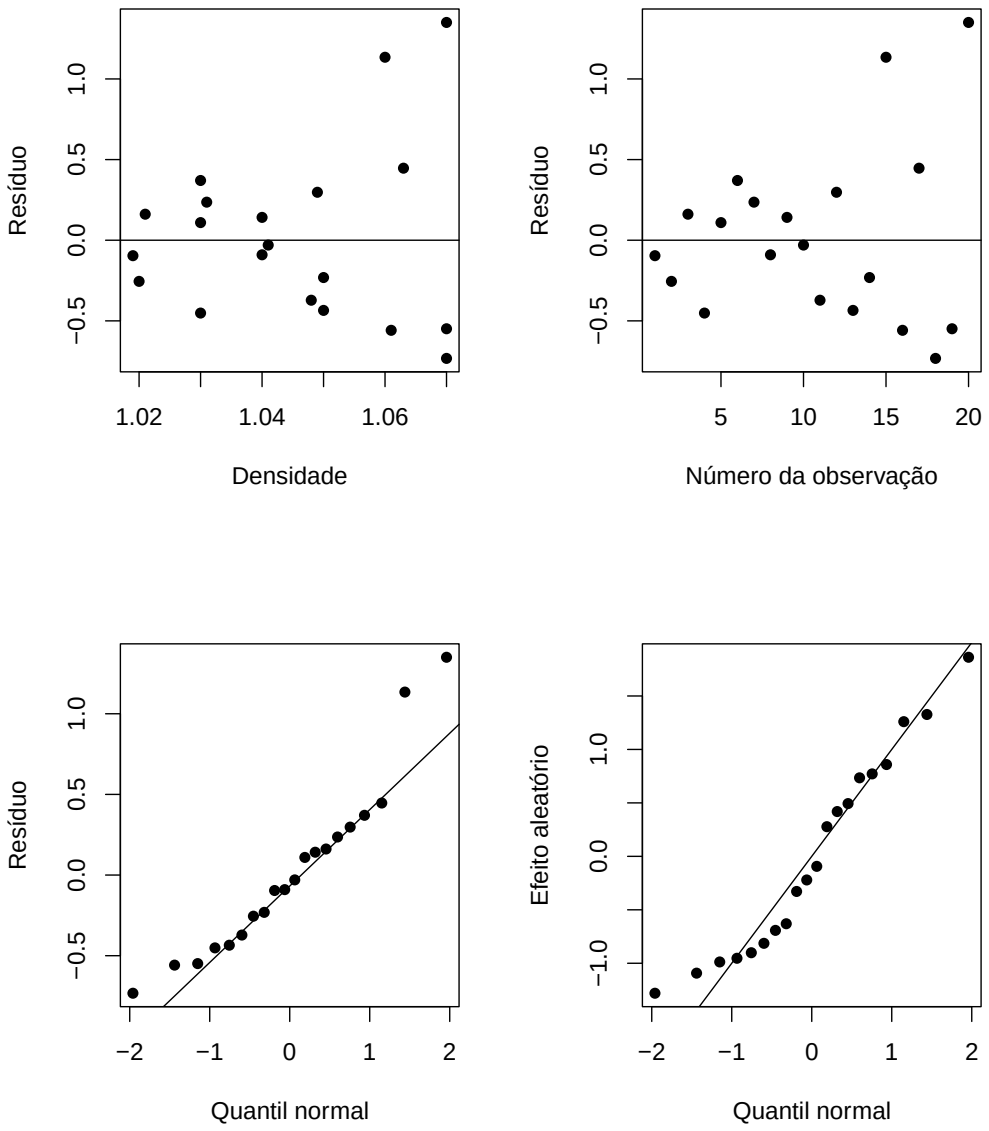

Figura 6 - Análise de resíduos para o Modelo Binomial Misto transformado, dado por (10), Exemplo 1.

3.2 Exemplo 2: Tolerância do *Paracoccidioides brasiliensis* a altas temperaturas

Esse conjunto de dados origina-se de um estudo experimental para investigar a tolerância de culturas de células de isolados do fungo *Paracoccidioides brasiliensis* a altas temperaturas. Esse fungo é o agente causador da Paracoccidioidomicose (PCM), uma micose sistêmica endêmica de regiões rurais da América Latina, considerado um importante problema de saúde pública no Brasil, por incapacitar e levar a óbito, principalmente homens na idade ativa. Amostras do fungo são encontradas em tecidos de humanos, cães e tatus. Theodoro et al. (2008) isolaram amostras do fungo em vários organismos e, dentre outros objetivos, estudaram a tolerância à temperatura de células cultivadas. Nesse exemplo foi utilizado apenas amostra de um dos isolados. Fragmentos de fungo sofreram pré-cultura padrão, após o qual amostras foram tomadas e cultivadas por mais 15 dias variando-se a temperatura em cinco níveis igualmente espaçados entre 37 e 41^oC. Para cada tratamento foram realizadas cinco réplicas. Em cada réplica contou-se o número de células viáveis e o número de células mortas da cultura.

Novamente o delineamento experimental sugere o uso de um Modelo Binomial Misto, porém seguindo a mesma sequência de estratégias de modelagem do exemplo anterior, inicialmente foi ajustado o Modelo de Regressão Logística usual sem transformação. Este modelo apresentou sobredispersão dos dados uma vez que sua *deviance* é de 230,34 com 23 graus de liberdade e ainda a análise de resíduos não foi satisfatória pois o ajuste mostrou valores altos para os resíduos, sendo alguns pontos influentes. Aplicando a técnica de PF, o melhor modelo se revelou de segundo grau ($m = 2$) com a transformação ($p_1 = p_2 = -2$), cuja expressão para o preditor linear ajustado é dado por:

$$\hat{\eta} = -551{,}8(53{,}25) + 228{,}9(23{,}13)X^{-2} + 153{,}7(15{,}96)X^{-2}\log X. \qquad (11)$$

Esse modelo apresentou *deviance* de 163,6 e 20 graus de liberdade. A

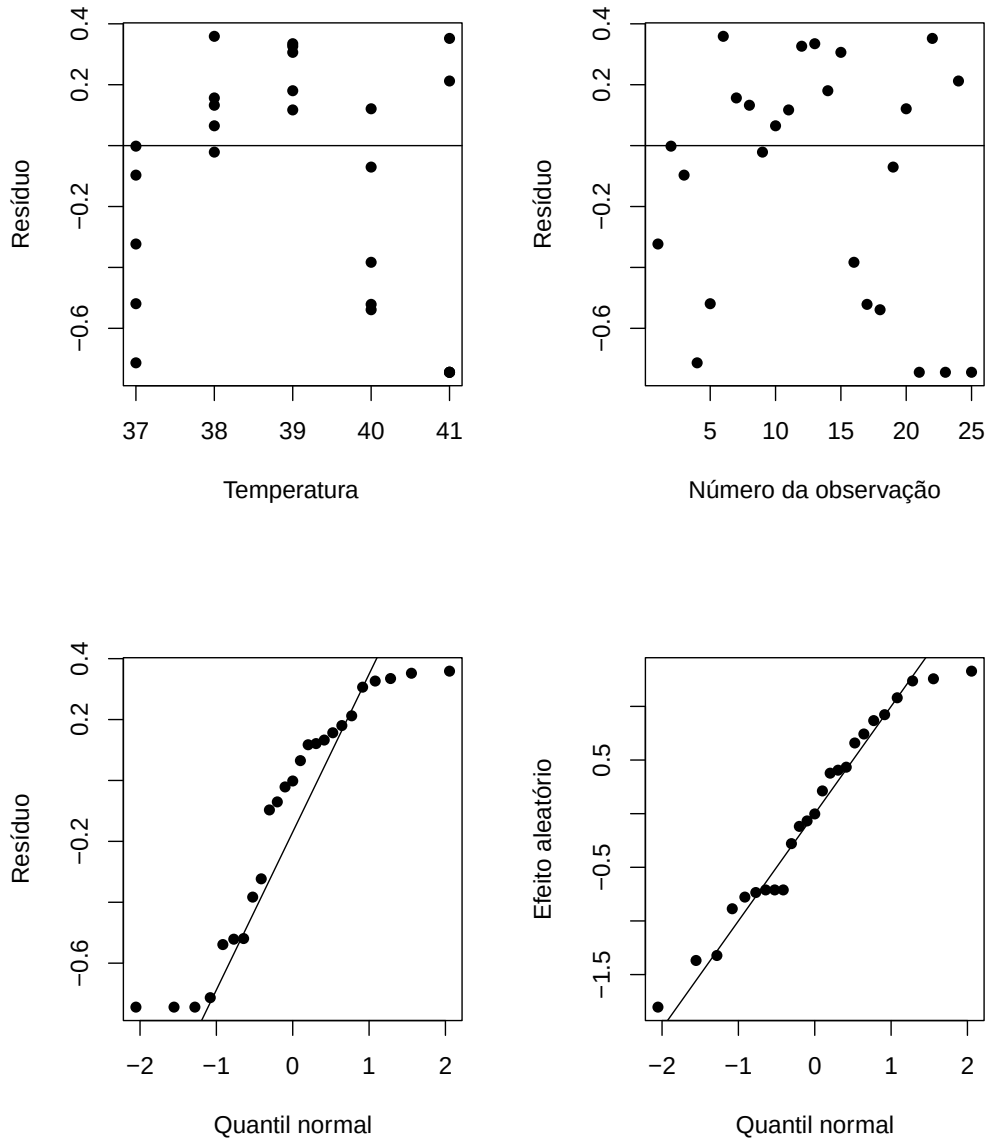

Figura 7 - Análise de Resíduos para o Modelo Binomial Misto com efeito linear de X, Exemplo 2.

deviance diminui, porém, ainda é bem mais alta do que o esperado. O ajuste ainda apresenta valores altos de resíduos para algumas temperaturas, indicando possíveis pontos influentes nas estimativas e falta de ajuste do modelo.

Aparentemente tem-se o mesmo problema de sobredispersão do exemplo anterior, e assim como Collett (1991) sugere, o próximo passo foi ajustar um modelo considerando as unidades experimentais com efeitos aleátorios. Para efeito de comparação ajustou-se o Modelo Binomial Misto com apenas o termo linear de temperatura. As seguintes estimativas para os parâmetros da parte fixa do modelo é dada por

$$\hat{\eta} = 90,43(8,38) - 2,31(0,21)X.$$

A estimativa para a componente de variância é 1,71, a *deviance* é 78,31 com 22 graus de liberdade. Apesar da *deviance* ter praticamente diminuido pela metade, a análise de resíduos mostrou indícios de falta de ajuste do modelo já que os pontos se apresentam de forma tendenciosa como pode ser visto na Figura 7. Aplicando a técnica de PF para tentar resolver o problema da falta de ajuste e melhorar os gráficos de checagem das suposições do modelo, a transformação indicada foi a mesma do modelo de efeitos fixos, ou seja um modelo de segundo grau com $p_1 = p_2 = -2$ como sendo as melhores potências para transformar a variável temperatura. O Modelo Binomial Misto com trasformação na temperatura apresentou as seguintes estimativas para os parâmetros da parte fixa

$$\hat{\eta} = -466,6(116,5) + 191,7(50,8)X^{-2} + 128,0(35,1)X^{-2}\log X. \tag{12}$$

A estimativa da componente de variância é 1,14, a *deviance* igual a 69,94 com 19 graus de liberdade. Para este modelo a análise de resíduos se apresentou de forma satisfatória as suposições iniciais (Figura 8), além da *deviance* diminuir um pouco mais em relação ao modelo anterior. O problema da falta de ajuste também parece ter sido sanado.

Portanto, concluímos que, dentre os modelos investigados, o que se mostrou mais apropriado para ajustar a probabilidade de células viáveis do fungo

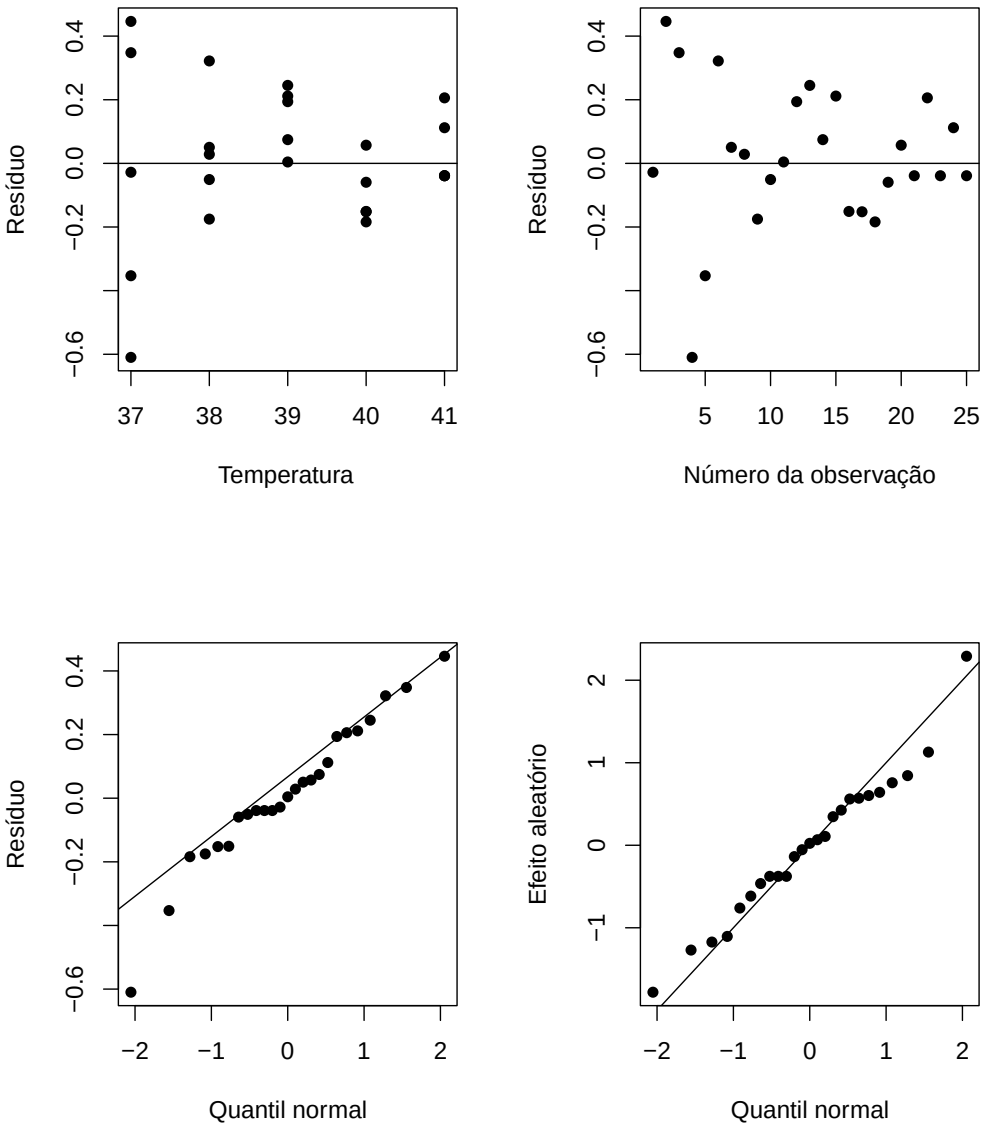

Figura 8 - Análise de resíduos para o Modelo Binomial Misto transformado, Exemplo 2.

em relação à temperatura foi o Modelo Binomial Misto com a variável temperatura transformada por um PF de segundo grau.

Para ilustrar as consequências nas interpretações dos resultados dos diversos modelos, intervalos de confiança para a razão de chances foram construídos. Vale notar que o Modelo Logístico usual assume razão de chances constante quando se aumenta a temperatura em uma unidade, independentemente da referência. Em alguns casos essa pode ser a razão da frequente falta de ajuste de tal modelo. Já o Modelo PF apresenta a flexibilidade de razões de chances dependentes da referência. Embora a interpretação se torna mais complexa, o modelo representa melhor a realidade.

Foram calculadas estimativas do risco relativo quando muda-se a temperatura de 39º para 40ºC, para o Modelo Linear sem transformação e para o Modelo Binomial Misto transformado. Para o Modelo Linear o risco é constante independente da temperatura, ou seja, o risco de morte da célula conforme aumenta-se a temperatura de 39º para 40ºC é 11,13. O intervalo de 95% de confiança para esta estimativa é $(9,39; 13,33)$. Para o Modelo Binomial Misto, o risco é bem maior, 17,15, com intervalo de confiança estimado em $(10,07; 29,20)$. Definindo, agora, a mudança da temperatura 38º para 39ºC, para o Modelo Misto, o risco diminui para 6,81 com intervalo de confiança de $(4,48; 10,34)$. Como esperado, o intervalo de confiança para o Modelo Misto pode ser mais amplo do que o linear, dependendo da região de interesse. Isto se deve ao fato de se considerar as unidades experimentais como um fator aleatório, ou seja, a variabilidade extra.

3.3 Exemplo 3: Relação Peso e Comprimento de Aranhas

A distinção direta entre aranhas fêmeas jovens e adultas (a partir das linhagens basais) dos gêneros *(Mygalomorphae e Haplogynae)* é impraticável em muitos estudos ecológicos, pois para a avaliação do seu estado reprodutivo a aranha necessita estar morta. Acredita-se que o estado reprodutivo esteja relacionado à relação peso-tamanho dos indivíduos. Nesse sentido, Stropa & Trinca (2005) estudaram a

relação entre o comprimento (mm) e peso (mg) do corpo de aranhas do tipo marrom (nome científico) coletadas no Jardim Botânico do Instituto de Biociências de Botucatu ($n = 289$ indivíduos). A natureza não linear da relação (Figura 9) e a forte evidência de heterogeneidade de variâncias levou os autores a aplicarem uma transformação *ad hoc* (logaritmica) no peso e estudar a relação através de Modelos de Regressão Segmentada. Nesta aplicação será feito um estudo diferenciado tentando explicar a relação entre as duas variáveis a partir de Polinômios Fracionários.

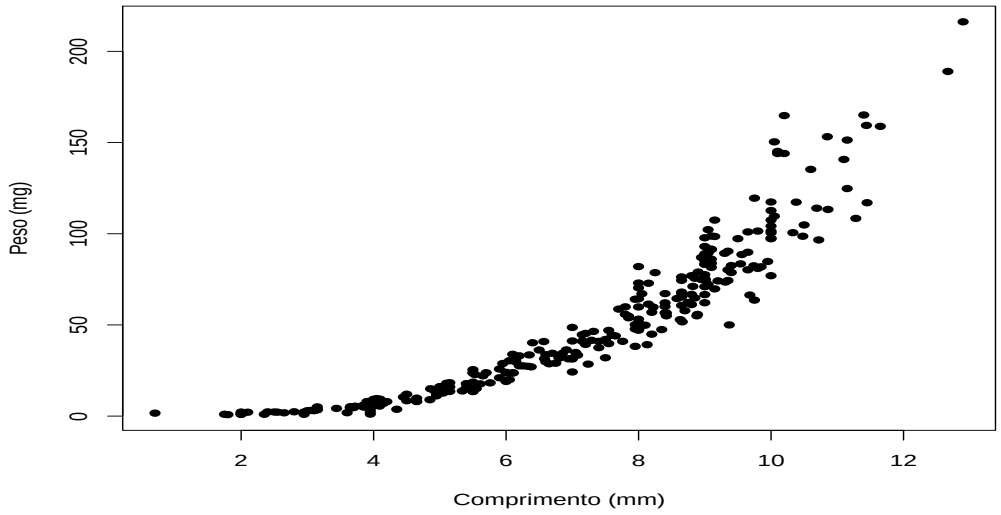

Figura 9 - Diagrama de dispersão do Comprimento (mm) e Peso (mg) de aranhas, Exemplo 3.

A metodologia de PF tem como objetivo corrigir problemas de falta de ajuste do modelo mas não consegue corrigir os desvios aos pressupostos do Modelo de Regressão clássico sobre a variável resposta. Embora a metodologia de Modelos Lineares Generalizados estende em muito as alternativas de modelagem, na prática, ainda pode-se encontrar dificuldades para ajustar um dado modelo. Isso ocorreu com esse exemplo e uma saída como a proposta de modelagem, se baseou no uso de Polonômios Fracionários conjuntamente com transformações na variável resposta.

Ignorando a heteregeneidade de variâncias, o primeiro passo foi utilizar

o procedimento de seleção do melhor modelo de PF que resultou em:

$$\hat{y} = 1{,}93(1{,}0400) + 0{,}10(0{,}0016)X^3. \tag{13}$$

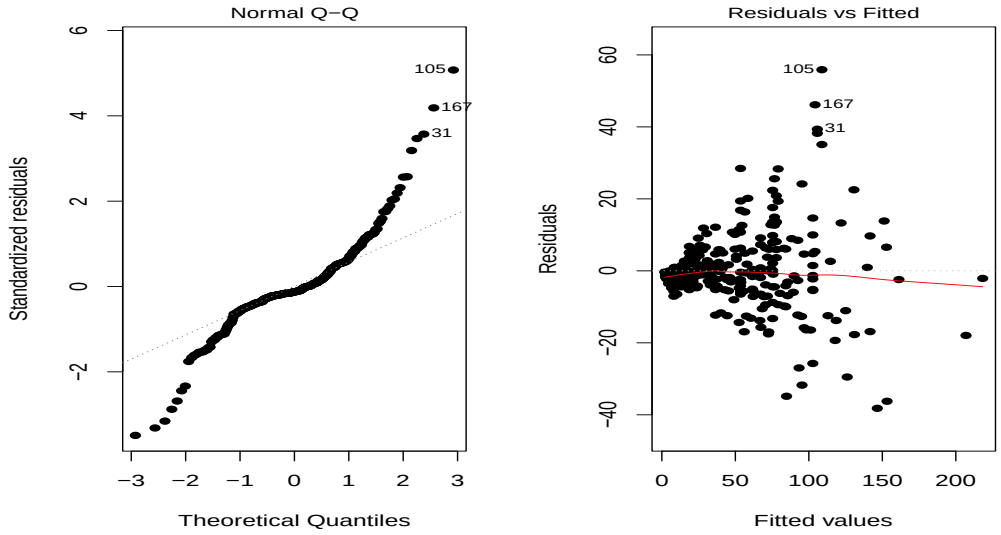

Figura 10 - Análise de resíduos para o modelo da equação (13), Exemplo 3.

Como esperado, a análise de resíduos mostrou, com mais destaque, indícios de que a variância está aumentando com os valores ajustados e que existe problemas com a normalidade dos erros (Figura 10). Na prática, aplicam-se transformações na resposta para contornar este tipo de problema. Surge então uma dúvida do que transformar primeiro: a variável resposta ou a regressora? No caso, como Y e X podem precisar de transformação, Royston & Sauerbrei (2008) sugerem que Y seja transformada primeiro e fixada essa transformação busca-se a melhor transformação em X. Porém, como λ de (2) é otimizado supondo a média da resposta dada por $X\hat{\boldsymbol{\beta}}$ e $X\hat{\boldsymbol{\beta}}$ não necessariamente é o melhor preditor, propõe-se um método iterativo de transformações. Então, a partir do modelo dado na equação (13) estimou-se λ por máxima verossimilhança, o que resultou em $\hat{\lambda} = 0{,}325$. Com a resposta transformada $y(0{,}325)$ foi determinado o melhor modelo PF ajustado, dado por:

$$\hat{y}(0{,}325) = -22{,}04(0{,}98) + 12{,}97(1{,}06)X^{-0{,}5} + 9{,}07(0{,}22)X^{0{,}5}. \tag{14}$$

Para esse modelo, parece que o problema de heteroscedasticidade de variância foi atenuado, mas não o de normalidade dos erros (Figura 11), mostrando observações com resíduos padronizados altíssimos.

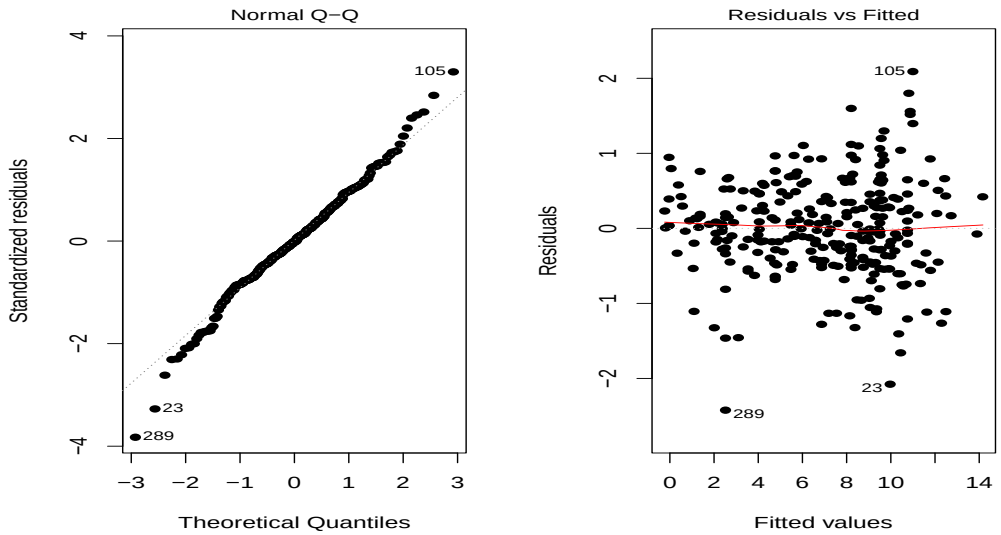

Figura 11 - Análise de resíduos para o modelo da equação (14), Exemplo 3.

Tabela 4: Estimativas dos parâmetros das transformações de Box-Cox e de PF, Exemplo 3

Passo	λ	m	p_1	p_2	*Deviance*
0	1	1	3		35210,04
1	0,325	2	-0,5	0,5	115,89
2	0,280	2	0	0	83,69
3	0,260	2	0	0	72,71

Assim, o processo iterativo deve buscar um novo valor para λ e sucessivamente, valores para as potências do PF, até que não haja mais mudanças nos parâmetros estimados ($\hat{p_1}$ e $\hat{p_2}$ ou $\hat{\lambda}$). A Tabela 4 resume os resultados deste processo

iterativo.

Foram necessários três passos para o processo convergir indicando a transformação y^{026} na resposta e o modelo quadrático para X na escala log, dado por:

$$\hat{y}(0{,}26) = -0{,}09(0{,}26) - 1{,}39(0{,}32)\log X + 2{,}31(0{,}10)(\log X)^2. \tag{15}$$

O intervalo de confiança para λ não incluiu os valores 0 e nem 0,5, os quais levariam às transformações logaritmica e raíz quadrada, costumeiramente utilizadas na prática.

O diagnóstico do ajuste do modelo ainda não é o ideal e mostra que existe uma observação (indivíduo 4) com alto *leverage* que pode ser influente nas estimativas (Figura 12). Para estudar a influência deste ponto em específico, fez-se sua exclusão e repetiu-se o processo iterativo de transformações de variáveis. A Tabela 5 resume os resultados dos parâmetros estimados para o conjunto de dados sem a observação 4.

Tabela 5: Estimativas dos parâmetros das transformações de Box-Cox e de PF sem a observação 4, Exemplo 3

Passo	λ	m	p_1	p_2	Deviance
	1	1	3		35210,00
1	0,330	2	-1	0,5	119,06
2	0,280	2	-2	0,5	82,90
3	0,270	2	-0,5	0	77,26
4	0,254	2	-0,5	0	69,13

Os resultados mostram certa diferença entre os dois modelos finais indicados nas Tabelas 4 e 5. A observação 4, foi influente na estimação do PF mas não muito importante na estimação de λ. É interessante notar que essa observação se refere a um indivíduo cujas medidas foram muito pequenas (ponto de menor comprimento que aparece bem separado dos demais na Figura 9). Como na prática

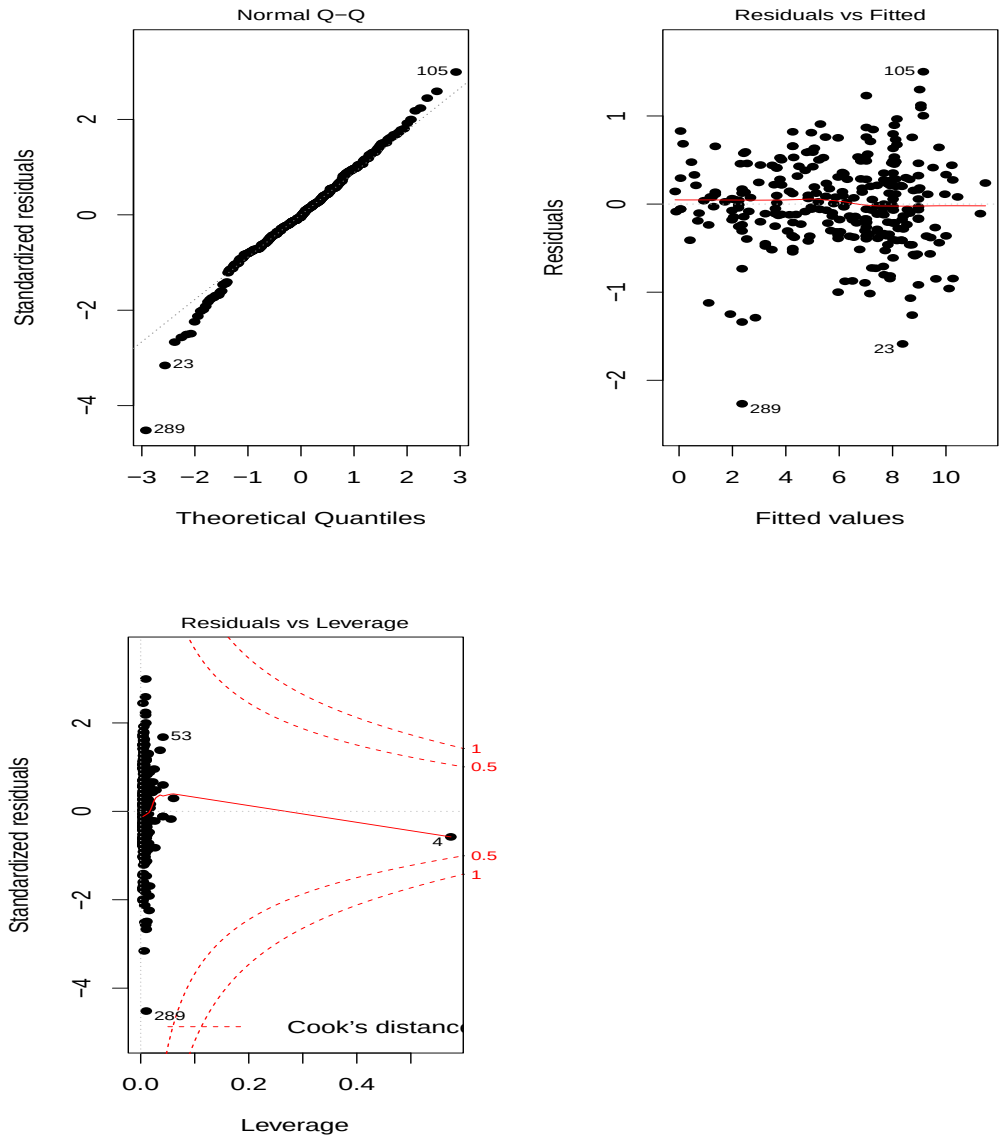

Figura 12 - Análise de resíduos para o modelo da equação (15), Exemplo 3.

a obtenção de medidas de organismos muito pequenos não é muito acurada, talvez, seja justificado a remoção desse indivíduo para o estudo. Portanto o modelo final do processo iterativo eliminando o ponto 4 é dado por:

$$\hat{y}(0{,}254) = -38{,}54(1{,}99) + 40{,}02(2{,}44)X^{-0\overline{5}} + 15{,}06(0{,}54)\log X. \tag{16}$$

O diagnóstico do modelo obtido é satisfatório (Figura 13), apesar de ainda apresentar uma observação com resíduo fora da faixa de valores esperada. A Figura 14 ilustra o ajuste desse modelo com os dados transformados em y. Embora as equações dos dois modelos PF pareçam muito diferentes fez-se um exercício para avaliar o quão diferentes seriam as predições do peso a partir de medidas do comprimento. Como a transformação Box-Cox foi praticamente a mesma nos dois casos, isso fica mais fácil e é ilustrado na Figura 15.

As diferenças aparecem, praticamente, apenas nos extremos e são marcantes no extremo inferior. Essa análise salienta os perigos de se fazer extrapolações por Modelos de Regressão. Avaliando esse aspecto, embora a observação 4 tenha apresentado alto *leverage*, o modelo com os dados completos parece mais robusto para predições e talvez mais razoável na prática. No entanto, se essa região for considerada importante, mais observações devem ser tomadas para melhor esclarecimento da relação.

Esse exemplo mostrou que embora os PFs representem alternativas flexíveis para lidar com falta de ajuste, as transformações encontradas podem ser muito dependentes de observações influentes e, portanto, investigações cuidadosas devem ser feitas durante a modelagem. A aplicação mostrou benefício no processo iterativo para encontrar transformações tanto na variável resposta quanto na variável explicativa. No caso dos dados completos o modelo final pelo processo iterativo foi mais simples (quadrático na escala log) do que aquele resultante da aplicação da transformação Box-Cox antes de se realizar a busca do melhor PF. Outra alternativa ao processo iterativo, quando o número de observações é razoavelmente grande, é discretizar X num número de intervalos de classes e estimar λ utilizando como preditor linear o modelo as classes (fator qualitativo). Desta forma a transformação

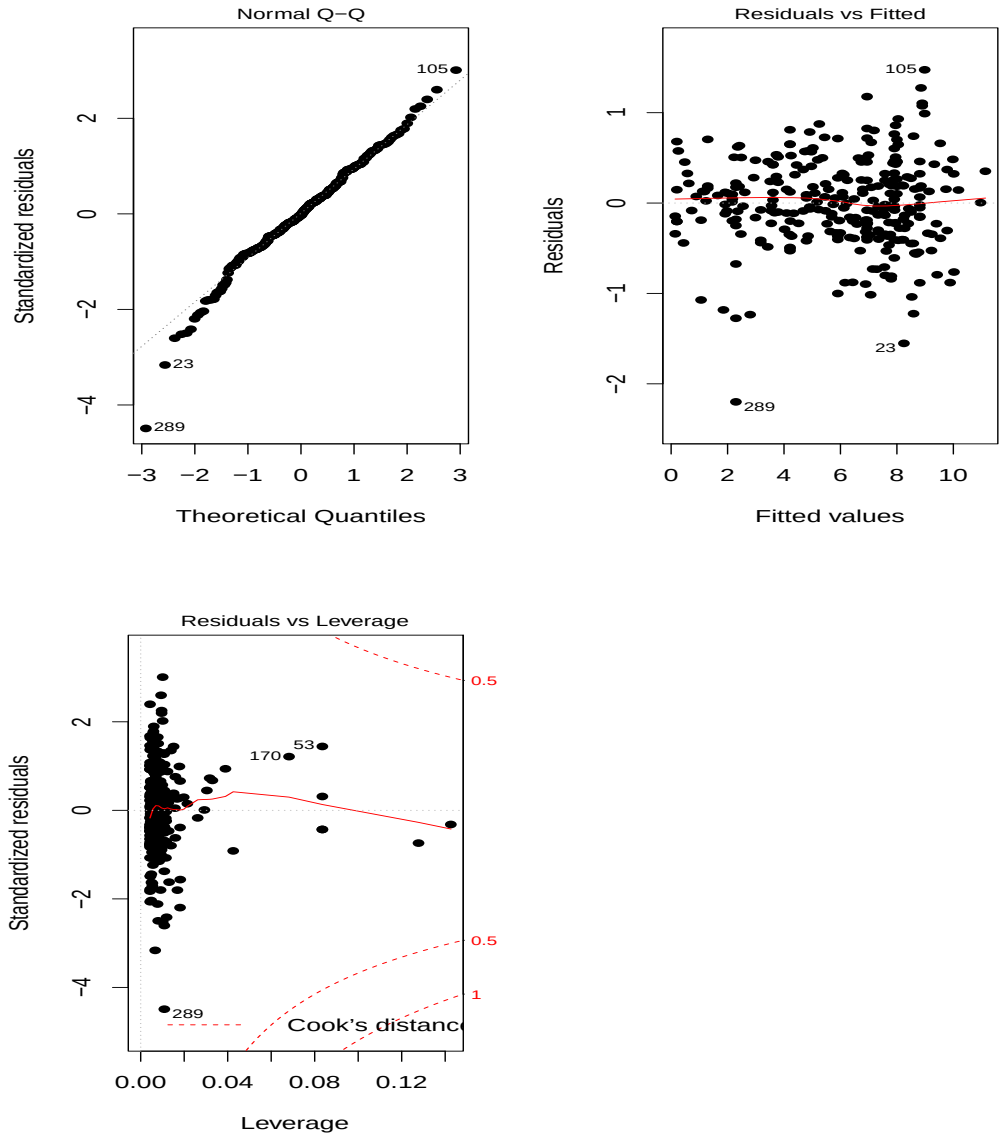

Figura 13 - Análise de resíduos para o modelo da equação (16) (sem a observação 4), Exemplo 3.

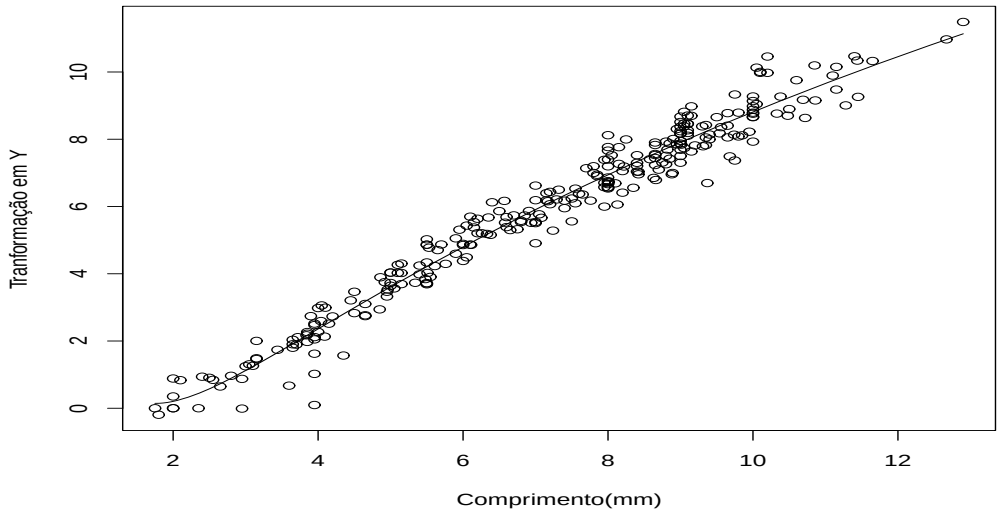

Figura 14 - Curva ajustada para o modelo da equação (16) considerando os dados transformados e removida a observação 4, Exemplo 3.

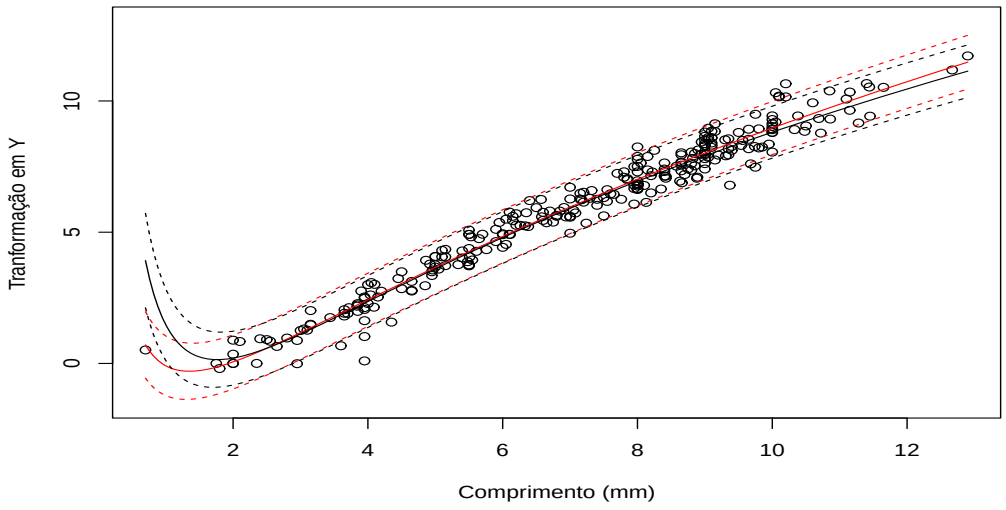

Figura 15 - Curvas ajustadas e intervalos de predição do peso (transformado) para os Modelos PF, com e sem a abservação 4, Exemplo 3.

Box-Cox é baseada na variabilidade de *erro puro* e, portanto, robusta a mudanças da parte preditiva. Essa alternativa foi testada resultando em $\hat{\lambda}$ =0,265 e portanto bem próxima das estimativas por iteração.

4 CONSIDERAÇÕES FINAIS

Esse trabalho considerou a aplicação de Polinômios Fracionários na análise de dados da área Biológica a partir de Modelos de Regressão. O uso de PF aumenta a flexibilidade dos modelos quando o preditor linear que inclui apenas efeitos simples das regressoras mostra falta de ajuste. Os três exemplos ilustrados se beneficiaram do uso de PF. Nos exemplos com o Modelo Logístico foi necessário a extensão para Modelos Mistos devido a presença de sobredispersão. Embora Royston & Altman (1994), Ambler & Royston (2001) e Royston & Sauerbrei (2008) tenham apresentado muitos exemplos práticos envolvendo uma ampla classe de modelos, o Modelo Misto não havia sido explorado. A estratégia de estimação das potências restrita a um conjunto enumerável de valores permite que a estimação possa ser realizada por qualquer *software* capaz de ajustar Modelos Mistos, basta repetir a tarefa para cada valor do conjunto e comparar as *deviances* dos modelos.

Para os três exemplos foi necessário um modelo de grau 2. Embora essa não seja a situação ideal, já que não se pode dizer que tal modelo é de simples interpretação, a análise indicou a necessidade da complexidade para explicar a variabilidade dos dados. No caso do estudo da tolerância à temperatura dos fungos a análise indicou que o risco relativo varia com a temperatura de referência. Esse resultado parece bastante coerente do ponto de vista biológico. Vale ressaltar que parâmetros do tipo potência são de difícil estimação, em geral, e que no caso, a variável regressora com apenas cinco níveis distintos, pode causar viés na estimação da potência. O erro de estimação da potência não foi considerado nesse trabalho mas seu estudo pode ser feito pela teoria de modelos não-lineares.

O estudo da relação entre peso e comprimento das aranhas salientou o

cuidado que se deve ter na avaliação do modelo ajustado. Foi ilustrado o problema da influência de certas observações na determinação do polinômio e o perigo de extrapolações a partir do modelo ajustado. Embora nos extremos os dois modelos estudados mostraram-se bem distintos, para a faixa de valores da regressora na qual os modelos ajustados são válidos, as predições foram muito similares, indicando que no caso de predição, os Modelos de PF são bastante úteis apesar de estarem sujeitos à alta influência de algumas poucas observações.

REFERÊNCIAS BIBLIOGRÁFICAS

AGRESTI, A. **An Introduction to Categorical Data Analysis**. 2. ed. NYC: Willey, 2007.

AMBLER, G.; ROYSTON, P. Fractional Polynomial Model selection procedures: investigation of Type I error rate. **Journal of Statistical Computation and Simulation**, v.69, p.89–108, 2001.

BOX, G. E. P.; COX, D. R. An analysis of transformations. **Journal of the Royal Statistical Society B**, v.26, p.211–243, 1964.

BOX, G. E. P.; TIDWELL, P. W. Transformation of the independent variables. **Technometrics**, v.4, p.531–550, 1962.

COLLETT, D. **Modelling Binary data**. London: Chapman and Hall, 1991. 369p.

DEMÉTRIO, C. G. B. **Modelos Lineares Generalizados em Experimentação Agronômica**. Piracicaba-SP: Esalq-USP, 2002. 113p.

DRAPER, N. R.; SMITH, H. **Applied Regression Analysis**. 3. ed. New York: Jhon Wiley e Sons, 2008. 706p.

FAHRMEIR, L.; KAUFMANN, H. Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. **Annals of Statistics**, v.13, p.1342–1368, 1985.

HINDE, J. P.; DEMÉTRIO, C. G. B. Overdispersion: models and estimation. **Comp. Stat. Data Anal.**, v.27, p.151–170, 1998.

MCCULLAGH, P.; NELDER, J. **Generalized Linear Models**. 2. ed. London: Chapman and Hall, 1989.

MCCULLOCH, C. E.; SEARLE, S. R.; NEUHAUS, J. M. **Generalized, Linear, and Mixed Models**. 2. ed. New York: John Wiley e Sons, 2008. 384p.

MONTGOMERY, D. C.; PECK, E. A.; VINING, G. G. **Introduction to Linear Regression Analysis**. 4. ed. Hardcover: Jhon Wiley e Sons, 2006. 640p.

MOOD, A. M.; GRAYBILL, F. A.; BOES, D. C. **Introduction to the Theory of Statistics**. 3. ed. New York: McGraw-Hill, 1974. 564p.

MOSTELLER, F.; TUKEY, J. W.; NEUHAUS, J. M. **Data analysis and regression**. New York: Addison-Wesley, 1977.

NELDER, J. A.; WEDDERBURN, R. W. M. Generalized Linear Models. **Journal of the Royal Statistical Society A**, v.135, n.3, p.128–141, 1972.

PAULA, G. A. **Modelos de regressão com apoio computacional**. São Paulo: IME-USP, 2004. 245p.

PINHEIRO, J. C.; BATES, D. M. Mixed-Effects Models in S and S-PLUS. **Springer**, 2000.

R DEVELOPMENT CORE TEAM. **R: A Language and Environment for Statistical Computing**. R Foundation for Statistical Computing, Vienna, Austria, 2010. ISBN 3-900051-07-0.

ROYSTON, P.; ALTMAN, D. G. Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling. **Journal of the Royal Statistical Society**, v.43, n.3, p.429–467, 1994.

ROYSTON, P.; SAUERBREI, W. **Multivariable Model-Building**. Chichester: Jhon Wiley e Sons, 2008. 303p.

SOFTWARE, S. S. A. SAS Institute Inc., Cary, NC, USA, versão 9.2 ed., 2008.

STATA STATISTICAL SOFTWARE: RELEASE 12. **StataCorp.** College Station, TX: StataCorp LP, 2011.

STROPA, A. A.; TRINCA, L. A. A model to assess the minimal size of adult female spiders. **Journal of health and envionmental**, v.1, p.3–11, 2005.

THEODORO, R. C.; BOSCO, S. M. G.; ARAÚJO, J. J.; CANDEIAS, J. M. G.; MACORIS, S. A. G.; TRINCA, L. A.; BAGAGLI, E. Dimorphism, thermal tolerance, virulence and Heat Shock Protein 70 transcription in different isolates of Paracoccidioides brasiliensis. **Mycopathologia**, v.165, p.355–365, 2008.

WALD, A. Asymptotically most powerfull tests of statistical hypotheses. **Annals of Mathematical Statistics**, v.12, p.1–19, 1941.