Extensão do Método de Predição do Vizinho mais Próximo para o Modelo Poisson Misto

Helder Alves Arruda

Dissertação apresentada
ao
Instituto de Matemática e Estatística
da
Universidade de São Paulo
para
obtenção do título
de
Mestre em Ciências

Programa: Estatística

Orientador: Profª. Drª. Viviana Giampaoli

Durante o desenvolvimento deste trabalho o autor recebeu auxílio financeiro da CAPES/CNPq

São Paulo, março de 2017

Extensão do Método de Predição do Vizinho mais Próximo para o Modelo Poisson Misto

Esta versão da dissertação contém as correções e alterações sugeridas pela Comissão Julgadora durante a defesa da versão original do trabalho, realizada em 28/03/2017. Uma cópia da versão original está disponível no Instituto de Matemática e Estatística da Universidade de São Paulo.

Comissão Julgadora:

- Profª. Drª. Viviana Giampaoli (orientadora) - IME-USP
- Profª. Drª. Karin Ayumi Tamura - Marketdata Solutions Brasil
- Profª. Drª. Mariana Rodrigues Motta - UNICAMP

Resumo

ARRUDA, H. A. **Extensão do Método de Predição do Vizinho mais Próximo para o Modelo Poisson Misto**. 2017. Dissertação (Mestrado) - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2017.

Várias propostas têm surgido nos últimos anos para problemas que envolvem a predição de observações futuras em modelos mistos, contudo, para os casos em que o problema trata-se em atribuir valores para os efeitos aleatórios de novos grupos existem poucos trabalhos. Tamura, Giampaoli e Noma (2013) propuseram um método que consiste na computação das distâncias entre o novo grupo e os grupos com efeitos aleatórios conhecidos, baseadas nos valores das covariáveis, denominado Método de Predição do Vizinho Mais Próximo ou NNPM (Nearest Neighbors Prediction Method), na sigla em inglês, considerando o modelo logístico misto. O objetivo deste presente trabalho foi o de estender o método NNPM para o modelo Poisson misto, além da obtenção de intervalos de confiança para as predições, para tais fins, foram propostas novas medidas de desempenho da predição e o uso da metodologia Bootstrap para a criação dos intervalos. O método de predição foi aplicado em dois conjuntos de dados reais e também no âmbito de estudos de simulação, em ambos os casos, obtiveram-se bons desempenhos. Dessa forma, a metodologia NNPM apresentou-se como um método de predição muito satisfatório também no caso Poisson misto.

Palavras-chave: Modelo Poisson misto, Efeitos aleatórios, Predição, Vizinho mais próximo.

Abstract

ARRUDA, H. A. **An Extension of Nearest Neighbors Prediction Method for Mixed Poisson Model**. 2017. Dissertação (Mestrado) - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2017.

Many proposals have been created in the last years for problems in the prediction of future observations in mixed models, however, there are few studies for cases that is necessary to assign random effects values for new groups. Tamura, Giampaoli and Noma (2013) proposed a method that computes the distances between a new group and groups with known random effects based on the values of the covariates, named as Nearest Neighbors Prediction Method (NNPM), considering the mixed logistic model. The goal of this dissertation was to extend the NNPM for the mixed Poisson model, in addition to obtaining confidence intervals for predictions. To attain such purposes new prediction performance measures were proposed as well as the use of Bootstrap methodology for the creation of intervals. The prediction method was applied in two sets of real data and in the simulation studies framework. In both cases good performances were obtained. Thus, the NNPM proved to be a viable prediction method also in the mixed Poisson case.

Keywords: Mixed Poisson model, Random effects, Prediction, Nearest neighbors.

Sumário

Lista de Abreviaturas

BA	Base de ajuste
BP	Base de predição
MLG	Modelos Lineares Generalizados
MLGM	Modelos Lineares Generalizados Mistos
MP	Melhor Preditor
MPE	Melhor Preditor Empírico
MPL	Melhor Preditor Linear
MPLNV	Melhor Preditor Linear Não Viciado
NNPM	Nearest Neighbors Prediction Method
QMEP	Quadrados Médios do Erro de Predição
RE	Valor Relativo do Quadrado Médio dos Erros
RB	Viés Relativo

Capítulo 1

Introdução

1.1 Considerações iniciais

Desde sempre, os seres humanos buscam encontrar padrões em elementos da natureza com o intuito de compreender melhor e até medir os fenômenos que ocorrem a nossa volta. Tal busca por padrões pode ser entendida como a formulação de um modelo, e está presente nas mais variadas situações, algumas tão comuns que fazem com que as pessoas criem "modelos" quase que intuitivamente como, por exemplo:

1. Uma pessoa que olha para o céu, vê nuvens e diz que acha que vai chover.
2. Alguém olha um animal desconhecido, e supõe que ele seja perigoso, baseado nos dentes e garras afiados que ele possui.
3. Um aluno que acha que só vai bem em provas se estiver usando a sua meia da sorte.

Aparentemente, o modelo 3 insinua-se como um modelo pior do que os dois primeiros, visto que, suas motivações provavelmente tenham raízes puramente supersticiosas. No entanto, o que um modelo precisa ter para ser considerado certo?

Existe uma frase atribuída a George E. P. Box que talvez sirva de resposta a esse questionamento:

Essencialmente, todos os modelos estão errados, mas alguns são úteis.

E é devido a isso que almeja-se a criação de modelos estatísticos de modo que seja possível medir se os modelos supostos são bons, isto é, se são capazes de descrever aproximadamente o fenômeno ao qual eles se propõem.

Nessa ótica, os modelos lineares normais tornaram-se extremamente populares, graças a sua praticidade e simplicidade de resultados. Contudo, com o tempo foram surgindo problemas que os modelos lineares normais não conseguiam abarcar como, por exemplo, dados estruturados de forma hierárquica. Em tal estrutura, as unidades de um mesmo nível, agrupadas em uma unidade de nível mais alto, são raramente independentes, visto que estas unidades compartilham um mesmo ambiente ou apresentam características semelhantes.

Os Modelos Lineares Generalizados Mistos (MLGM), que são uma extensão dos Modelos Lineares Generalizados (MLG), apresentam-se como uma classe de modelos que incorpora efeitos aleatórios que são estimados individualmente para cada grupo, ou seja, os grupos são representados por efeitos aleatórios, e sua variabilidade entre-grupos é descrita por meio da variabilidade do intercepto aleatório e/ou da variabilidade da inclinação aleatória. Os MLGM também abrangeram o escopo de distribuições assumidas para a variável resposta, podendo adotar diversos tipos de distribuição pertencentes à família exponencial. O modelo particular da classe dos MLGM que considera a distribuição de Poisson será o objeto de estudo deste trabalho.

Assim sendo, o nosso interesse principal é predizer o componente aleatório, ou uma nova observação, em modelos Poisson mistos. Todavia, o trabalho restringir-se-á ao caso de predição para observações de grupos não presentes na amostra inicial, como será visto melhor nas seções seguintes.

1.2 Objetivos

Pode-se dividir um problema de predição para modelos mistos em dois casos: para grupos presentes na amostra ou para novos grupos. A literatura apresenta algumas propostas para questões do segundo tipo como, por exemplo, o Método de Predição do Melhor Preditor Empírico em Tamura (2012), no entanto, tal proposta envolve técnicas de integração muito custosas computacionalmente, podendo até se tornarem inviáveis dependendo da dimensão das integrais a serem solucionadas. Tendo em vista este quadro, Giampaoli *et al.* (2013) criaram um novo método, que não necessita da suposição de distribuição dos efeitos aleatórios considerando o modelo logístico misto. No entanto, duas importantes questões surgem a partir desse estudo:

- Apesar do método NNPM ter apresentado bons resultados para a predição pontual, não foi apresentada uma forma para a obtenção dos seus respectivos intervalos de confiança.
- Em Giampaoli *et al.* (2013), o método NNPM foi aplicado utilizando medidas de performance só definidas para o modelo logístico.

Portanto, os objetivos principais deste trabalho são estender o método NNPM para o modelo Poisson misto, analisando o seu comportamento em estudos de simulação e na aplicação em bancos reais, e propor uma forma para a construção de intervalos de confiança para as predições.

1.3 Contribuições

Abaixo, as principais contribuições desta dissertação:

- Proposição de três novas medidas de desempenho preditivo: *Porcentagem*, *Amplitude* e PA.
- A implementação do método de Bootstrap em conjunto com o método NNPM a fim de obter intervalos de confiança para as predições realizadas em grupos não presentes na amostra.
- Avaliação do comportamento do método NNPM no caso Poisson misto por meio de estudos de simulação.
- Ilustração do método NNPM para dois bancos com dados reais: uma na área de saúde e a outra na área de administração.

1.4 Organização do trabalho

Esta dissertação encontra-se dividida em oito capítulos.

O Capítulo 1 trata de uma esquematização do presente trabalho, contendo motivações iniciais e uma breve ambientação acerca dos temas que serão discutidos posteriormente, destacando os seus objetivos, além de uma seção que enumera as principais contribuições.

O Capítulo 2 serve como revisão dos temas base para a compreensão desta dissertação, ele se inicia com a distribuição de Poisson, até chegar na apresentação do modelo Poisson misto. Nas duas últimas seções, apresenta-se uma introdução ao método Bootstrap, adotado para a criação de intervalos de confiança e as medidas de desempenho preditivo Viés Relativo (RB) e o Valor Relativo do Quadrado Médio dos Erros (RE).

No Capítulo 3, o problema da predição em modelos mistos foi divido em dois casos: para grupos presentes na amostra e para novos grupos. Neste capítulo são discutidas algumas abordagens já existentes para ambas as situações.

O método de predição principal desta dissertação é apresentado no Capítulo 4, que surge como uma proposta alternativa para predição de observações oriundas de novos grupos. Este

capítulo é baseado no trabalho de Giampaoli *et al.* (2013), que desenvolveram o método NNPM para o caso logístico misto.

O Capítulo 5 possui o objetivo de estender o método NNPM para o caso Poisson misto, assim como, mostrar a utilização do método de Bootstrap para o problema de predição. A seção 5.2 propõe novas medidas de desempenho bastante úteis para modelos Poisson misto.

No Capítulo 6, é apresentado o uso de simulação para verificar se o método NNPM apresenta bons resultados para o caso Poisson misto, enquanto, no Capítulo 7, o método foi aplicado a dois bancos de dados reais.

As conclusões e sugestões para trabalhos futuros encontram-se no Capítulo 8.

Capítulo 2

Conceitos preliminares

Este Capítulo apresentará uma introdução aos tópicos estatísticos de maior importância para a compreensão desta presente dissertação: a distribuição de Poisson, Modelos Lineares Generalizados, Modelos Lineares Generalizados Mistos com resposta Poisson e uma introdução ao método Bootstrap.

2.1 Distribuição Poisson

Em muitos livros de estatística, como Johnson *et al.* (2005), a distribuição de Poisson é introduzida como um limite e, assim sendo, uma aproximação para a probabilidade do número de ocorrências de um dado evento raro quando uma grande quantidade de ensaios independentes é realizada.

Uma variável aleatória Y segue uma distribuição Poisson de parâmetro μ denotada por Y$\sim$Poisson(μ) se a probabilidade de ocorrência de y é

$$f(y) = \frac{e^{-\mu}\mu^{y}}{y!}, \tag{2.1}$$

em que:

- $Y = 0, 1, 2, 3...$
- $\mu > 0$.
- $E(Y) = Var(Y) = \mu$.

A distribuição de Poisson possui diversas aplicações, em geral muito utilizada em casos de contagem de eventos por unidade de tempo, e por também ser um caso limite para probabilidades no caso de amostragem sem reposição, foi descrita por Douglas (1980), p.5 como, em tradução livre:

Uma distribuição que desempenha uma função para as distribuições discretas similar a normal para as distribuições absolutamente contínuas.

Portanto, tais fatores servem de motivação da escolha desta distribuição para este trabalho.

2.2 Modelos Lineares Generalizados

Durante muito tempo os modelos lineares normais foram amplamente utilizados e, quando a suposição de normalidade não se verificava, buscava-se algum tipo de transformação na variável resposta. No entanto, tais transformações nem sempre são fáceis de serem obtidas, além do que, podem obscurecer as interconexões fundamentais entre as variáveis, assim sendo, muitas vezes se é de interesse tentar modelar ou até mesmo predizer com outro tipo de distribuição, por exemplo, a distribuição de Poisson.

Foi nesse contexto que Nelder e Wedderburn (1972) apresentaram os modelos lineares generalizados (MLGs), abrindo mais possibilidades para a distribuição da variável resposta e permitindo que ela pertença à família exponencial de distribuições, que será definida a seguir.

Suponha $Y_1, ..., Y_n$ variáveis aleatórias independentes, cada uma com a função de densidade de probabilidade da forma

$$f(y_i) = \exp[\phi \{y\theta_i - b(\theta_i)\} + c(y_i, \phi)], \tag{2.2}$$

em que $E(Y_i) = \mu_i = b'(\theta_i), Var(Y_i) = \phi^{-1}V_i, V = \frac{d\mu}{d\theta}$ é a função de variância e $\phi^{-1} > 0$ é o parâmetro de dispersão.

Os modelos lineares generalizados são definidos por 2.2 e pela componente sistemática

$$g(\mu_i) = \eta_i, \tag{2.3}$$

em que $\eta_i = \mathbf{x}_i^{\mathbf{T}}\boldsymbol{\beta}$ é o preditor linear, $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^{\mathbf{T}}, p < n$, é um vetor de parâmetros desconhecidos a serem estimados, $\mathbf{x}_i = (x_{i1}, ..., x_{ip})^{\mathbf{T}}$ representa os valores de p variáveis explicativas e $g(.)$ é uma função monótona e diferenciável, denominada função de ligação. Em outras palavras, o modelo supõe que uma função da média da variável resposta pode ser modelada a partir de uma combinação linear de variáveis explanatórias.

Um caso particular importante da função de ligação ocorre quando o parâmetro canônico (θ) coincide com o preditor linear, isto é, quando $\theta_i = \eta_i$. Mais detalhes sobre métodos de

estimação dos MLGs e algoritmos para a sua obtenção podem ser encontrados, por exemplo, em Paula (2013).

Como dito anteriormente, os MLGs abriram um leque maior de opções para a distribuição da variável resposta, dentre elas, é fácil notar que a densidade em 2.1 pode ser escrita da forma 2.2, como será demonstrado abaixo:

$$f(y) = \frac{e^{-\mu}\mu^y}{y!} = \exp\left\{\log(\frac{e^{-\mu}\mu^y}{y!})\right\} = \exp\left\{\log(\mu^y) + \log(e^{-\mu}) - \log(y!)\right\}$$
$$= \exp\left\{y\log(\mu) - \mu - \log(y!)\right\}.$$

Portanto, pode-se escrever

$$f(y) = \exp\left\{y\log(\mu) - \mu - \log(y!)\right\}, \tag{2.4}$$

em que $\log(\mu) = \theta, b(\theta) = e^\theta, \phi = 1$, $c(y, \phi) = -\log(y!)$ e $V(\mu) = \mu$. Assim o modelo Poisson pertence à família dos MLG.

2.3 Modelo de Regressão Poisson Misto

Suponha que um pesquisador selecionou algumas cidades brasileiras e colheu um conjunto de variáveis explicativas relacionadas a elas, a fim de averiguar quais características de uma dada cidade interferem na quantidade total de lojas de uma rede de comida rápida (fast food), mais detalhes sobre este exemplo são apresentados no Capítulo 7. A princípio, o uso do modelo definido na sessão anterior com resposta Poisson parece ser indicado, visto que, a variável resposta trata-se de uma contagem, no entanto, vale à pena ressaltar que 2.2 tem a suposição de independência entre todas as observações. Neste caso, seria razoável supor que haja independência entre lojas de fast food da mesma cidade?

A fim de abranger problemas desse tipo, Wedderburn (1974) estendeu a ideia dos MLGs para situações mais gerais incluindo dados correlacionados, ou seja, não independentes, os chamados Modelos Lineares Generalizados Mistos (MLGMs). Resultando no seguinte modelo para o problema apresentado no exemplo

$$g(\mu_{ij}) = \mathbf{x}_{ij}^{\mathbf{T}}\boldsymbol{\beta} + \mathbf{z}_{ij}^{\mathbf{T}}\boldsymbol{\alpha}_i, \tag{2.5}$$

com $i = 1, ..., q$ e $j = 1, ..., n_i$.

No qual:

- y_{ij} é o número de lojas observado na j-ésima rede de fast food da i-ésima cidade.
- $\boldsymbol{y}_i = (y_{i1}, ..., y_{in_i})^{\mathbf{T}}$ é o vetor com as n_i observações da i-ésima cidade.
- $\boldsymbol{x}_{ij} = (1, x_{1ij}, ..., x_{(p-1)ij})^{\mathbf{T}}$ é o vetor com os valores conhecidos das covariáveis associados a parte fixa do modelo.
- $\boldsymbol{z}_{ij} = (1, z_{1ij}, ..., z_{(k-1)ij})^{\mathbf{T}}$ é o vetor com os valores conhecidos das covariáveis associados a parte aleatória do modelo.
- $Y_{ij}|\boldsymbol{\alpha}_i$ são variáveis aleatórias condicionalmente independentes com distribuição Poisson(μ_{ij}).
- $\boldsymbol{\beta} = (\beta_0, ..., \beta_p)^{\mathbf{T}}$ é o vetor de parâmetros fixos da regressão.
- $\boldsymbol{\alpha}_i$ é um vetor k x 1 de variáveis aleatórias independentes com distribuição Normal de média $\mathbf{0}$ e matriz de covariância $\boldsymbol{\Sigma}$ $(N(\mathbf{0}, \boldsymbol{\Sigma}))$ em que $\boldsymbol{\alpha}_i$ é o vetor de efeitos aleatórios correspondente à i-ésima cidade.
- $g(.)$ é uma função de ligação estritamente monótona e duplamente diferenciável.

Vale ressaltar que no contexto de modelos mistos, a função densidade, condicional ao efeito aleatório, de um MLGM é dada por

$$f(y_{ij}|\boldsymbol{\alpha}_i) = \exp[a_{ij}\phi\{y_{ij}\theta_{ij} - b(\theta_{ij})\} + c(y_{ij}, a_{ij}\phi)], \tag{2.6}$$

em que y_{ij} é a variável resposta da j-ésima observação pertencente ao i-ésimo grupo, para $j = 1, ..., n_i$ e $i = 1, ..., q$. As quantidades $b(.)$, $c(.,.)$ e ϕ são associadas analogamente à família descrita em 2.2 e a_{ij} é um peso determinado a observação dentro do grupo. A particularização de 2.6 para a distribuição de Poisson pode ser facilmente obtida utilizando 2.4. Para maiores detalhes sobre este modelo, ver McCulloch *et al.* (2008).

Para a estimação de $\boldsymbol{\beta}$ é necessário lançar mão de métodos iterativos, posto que, os estimadores de máxima verossimilhança dos parâmetros não possuem expressão analítica. Uma alternativa é utilizar o método numérico de quadratura de Gaus-Hermite apresentado em Liu e Pierce (1994). Já para a predição dos efeitos aleatórios, alguns métodos usuais serão discutidos no Capítulo 3, enquanto nos Capítulos 4 e 5 uma nova proposta será apresentada e expandida, sendo este o objetivo principal desta presente dissertação.

2.4 Introdução ao método Bootstrap

Utilizar os métodos mais comuns de inferência estatística clássica pode culminar, por vezes, em algumas dificuldades como, por exemplo:

- Nem todos os estimadores têm sua distribuição amostral facilmente definida, mesmo quando se conhece a distribuição da variável aleatória em estudo.
- Algumas das suposições só valem para amostras grandes, o que dificulta a obtenção da distribuição amostral de um dado estimador quando isso não ocorrer.

Uma das formas de tentar contornar tais problemas é lançando mão de técnicas que utilizam o método de reamostragem, que se baseiam em calcular estimativas a partir de repetidas amostragens dentro da mesma amostra. Assim sendo, não se considera a distribuição amostral assumida de uma estatística e calcula-se uma distribuição empírica ao longo de uma quantidade grande de amostras.

Existem vários tipos de métodos baseados na reamostragem como, por exemplo, testes de aleatorização, validação cruzada ou jackknife. Na próxima seção será apresentado o método conhecido como Bootstrap, que será importante neste trabalho para a obtenção dos intervalos de confiança dos efeitos aleatórios.

2.4.1 Bootstrap

O Bootstrap é um método de reamostragem proposto por Bradley Efron em 1979. Assim como todo procedimento de reamostragem, o método descarta a distribuição amostral teórica assumida de uma estatística e calcula uma distribuição empírica. Mais informações e aplicações para cenários mais complexos que os descritos neste texto podem ser encontrados em Efron e Tibshirani (1994).

O princípio da reamostragem de Bootstrap baseia-se em, tendo como partida uma amostra original, gerar novas amostras de mesmo tamanho a partir do sorteio aleatório com reposição de seus elementos.

Assim sendo, suponha que sejam observados os seguintes pontos independentes $x_1, x_2, ..., x_n$, denotados pelo vetor $\mathbf{x} = (x_1, x_2, ..., x_n)$, dele se computa uma estatística de interesse $s(\mathbf{x})$. A amostra de Bootstrap $\mathbf{x}^* = (x_1^*, x_2^*, ..., x_n^*)$ é obtida selecionando-se aleatoriamente e com reposição n elementos dos dados originais $\mathbf{x}$. Tem-se abaixo um esquema de como obter um intervalo de confiança de $(1 - \alpha)$, onde $\alpha \in (0, 1)$, para uma dada estatística utilizando o método de Bootstrap

1. B amostras de Booststrap são obtidas do conjunto original de dados.
2. As replicações Booststrap $s(\mathbf{x}^{*\mathbf{1}}), s(\mathbf{x}^{*\mathbf{2}}), ..., s(\mathbf{x}^{*\mathbf{B}})$ são obtidas calculando o valor da estatística $s(\mathbf{x})$ em cada uma das amostras de Bootstrap.
3. Os percentis de $\frac{\alpha}{2}$ e $1 - \frac{\alpha}{2}$ são calculados das replicações de Bootstrap, culminando num intervalo de $(1 - \alpha)$ de confiança para a estatística de interesse $s(\mathbf{x})$.

O método adotado aqui é conhecido como intervalo de confiança percentil, contudo, existem outras técnicas de intervalos de confiança Bootstrap como, por exemplo, o intervalo de confiança de Bootstrap t e os métodos de Bootstrap de correção. Uma descrição maior desses procedimentos pode ser encontrado em Rizzo e Cymrot (2006).

2.5 Noções básicas sobre simulação

São variadas e diversas as situações em que se encontram dificuldades em se obter soluções analíticas acerca de algumas questões de interesse, e é neste contexto que tais problemas podem ser analisados sob a ótica da simulação.

Na estatística, o uso da simulação para a validação de modelos tem se tornado cada vez mais comum, como pode-se ser visto, por exemplo, em Law Averill e David (2000). De maneira bastante sucinta, o processo pode ser descrito da seguinte forma:

1. Visando descrever algum dado sistema, um modelo estatístico é proposto.
2. Os parâmetros que configuram o sistema são escolhidos e tomados como populacionais.
3. A realização do sistema é recriada por computador utilizando os parâmetros escolhidos na etapa anterior.
4. A partir do modelo proposto inicialmente e o sistema obtido no passo anterior, tenta-se "acertar" os parâmetros escolhidos em 2) utilizando algum método de estimação.

Como o passo 3) em geral depende de um processo aleatório, as etapas 3) e 4) podem ser repetidas muitas vezes, assim sendo, imagine que estamos interessados num modelo que possa estimar uma parâmetro d. Escolhendo um valor para d fixo na etapa 2), para cada repetição, uma estimativa diferente de d poderá ser obtida.

Se o método de estimação proposto para estimar o sistema for bom, espera-se que os valores de d obtidos em cada repetição sejam próximos do valor real escolhido na etapa 2). Contudo, como medir se estão próximos?

No tópico seguinte, será apresentado um exemplo relacionado à predição.

2.5.1 Exemplo de simulação

Em Boubeta *et al.* (2004), tinha-se o interesse de verificar a qualidade do método do melhor preditor empírico, que será detalhado no Capítulo 3. Foi considerado o modelo a seguir

$$log(\mu_{ij}) = \beta_0 + \beta_1 x_{ij} + \alpha_i, \tag{2.7}$$

$i = 1, ..., n$ e $j = 1, ..., n_i$.

Sendo:

- x_{ij} os valores das covariáveis associados à parte fixa do modelo.
- $Y_{ij}|\alpha_i$ variáveis aleatórias condicionalmente independentes com distribuição Poisson(μ_{ij}).
- β_0 e β_1 os parâmetros fixos da regressão.
- α_i variáveis aleatórias independentes com distribuição $\sim N(0, \sigma^2)$.

O objetivo era predizer as quantidades μ_{ij} e α_i, para tal, foi proposto um esquema de simulação, seguindo os quatro passos descritos na seção anterior:

- O passo 1) corresponde ao modelo proposto em 2.7.

- No passo 2), os valores das covariáveis x_{ij} e os valores populacionais de β_0 e β_1 são definidos e, para cada iteração k, os valores iniciais de α_i são gerados e denotados por $\alpha_i^{(k)}$.

- O passo 3) corresponde à geração dos valores $Y_{ij}|\alpha_i$, utilizando os valores definidos na etapa anterior.

- Munido dos $Y_{ij}|\alpha_i$ e do modelo proposto, para cada iteração k, serão feitas estimativas para β_0, β_1 e os valores preditos de α_i, denotados, respectivamente, por $\hat{\beta}_0^{(k)}$, $\hat{\beta}_1^{(k)}$ e $\hat{\alpha}_i^{(k)}$.

Foram feitas K repetições e, para cada uma delas, uma predição $\hat{\mu}_{ij}^{(k)}$ e $\hat{\alpha}_i^{(k)}$ foi obtida, com $k = 1, ..., K$.

Boubeta *et al.* (2004) utilizaram o Viés Relativo (RB) e o Valor Relativo do Quadrado Médio dos Erros (RE) para a avaliação das predições. Tais medidas estão definidas abaixo

- Viés Relativo (RB):
  $RB_i = \frac{\sum_{k=1}^{K}(\hat{\alpha}_i^{(k)} - \alpha_i^{(k)})}{K|\bar{\alpha}_i|}$
  $RB = \frac{1}{n}\sum_{i=1}^{n} RB_i$
- Quadrado Relativo do MSE (RE):
  $RE_i = \frac{\sqrt{\sum_{k=1}^{K}(\hat{\alpha}_i^{(k)} - \alpha_i^{(k)})^2}}{K|\bar{\alpha}_i|}$
  $RE = \frac{1}{n}\sum_{i=1}^{n} RE_i$

Em que $\alpha_i^{(k)}$ e $\hat{\alpha}_i^{(k)}$ são o valores iniciais e preditos, respectivamente, na k-ésima iteração, $\bar{\alpha}_i = \frac{1}{K}\sum_{k=1}^{K}\alpha_i^{(k)}$ e os RB_i e RE_i tratam-se de medidas de viés individual, enquanto o RB e o RE são medidas de viés geral.

O RB e RE também foram calculados analogamente em relação aos μ_{ij}.

Dessa forma, quanto mais próximos de zero forem os valores de RB e RE, melhor será o modelo considerado em termos de predição. Portanto, tais medidas permitem que comparações entre modelos possam ser feitas. As medidas RB e RE serão utilizadas posteriormente nesta dissertação para este fim.

Capítulo 3

Predição

Alguns métodos de predição já existentes na literatura serão apresentados neste Capítulo, com um enfoque maior no método do Melhor Preditor Empírico. Todavia, antes de adentrar nesse assunto, é importante notar que, em se tratando de predição, pode-se subdividir o problema em dois casos: predição de efeitos aleatórios para um grupo já presente na amostra e predição dos efeitos aleatórios de novos grupos. Neste último caso, há muito poucas publicações a respeito.

A fim de ilustrar a diferença entre essas duas situações, considere novamente o exemplo das redes de fast food da seção 2.3. Após ajustar um Modelo de Regressão Poisson Misto, um dos possíveis interesses do pesquisador é, dada uma nova rede, tentar prever de alguma forma o número total de lojas que esta franquia desenvolverá, contudo, podemos ter um dos seguintes casos:

1. A nova franquia será aberta em uma cidade considerada no ajuste do modelo.
2. A nova franquia pertence a uma nova cidade não presente no banco de dados original.

Na seção seguinte, serão discutidos métodos de predição para os efeitos aleatórios de um Modelo Linear Misto Generalizado para a situação em que o grupo está presente na amostra – a exemplo do caso 1 – e para a situação em que se trata de um novo grupo – a exemplo do caso 2.

3.1 Alguns métodos de predição

Ao se trabalhar com predições, segundo Tamura (2012), o método que minimiza o erro quadrático médio da previsão é o mais utilizado para atribuir valores aos efeitos aleatórios. Assim sendo, seja $\zeta = \zeta(\boldsymbol{\beta}, \alpha_i)$ o termo em que se tem interesse de fazer a predição, com

$i = 1, ..., q$, deseja-se encontrar um preditor ζ' tal que a quantidade a seguir seja minimizada

$$E(\zeta' - \zeta)^2. \tag{3.1}$$

O método que visa minimizar 3.1 é conhecido por Melhor Preditor (MP), abaixo será mostrado que $\zeta" = E(\zeta|\boldsymbol{y})$, em que $\zeta"$ é o MP de ζ:

A quantidade 3.1 que se deseja minimizar, não é alterada se for somado e subtraído o termo $\zeta_0 = E(\zeta|\boldsymbol{y})$, isto é

$$E(\zeta' - \zeta)^2 = E(\zeta' - \zeta_0 + \zeta_0 - \zeta)^2,$$

dessa forma, é fácil ver que

$$E(\zeta' - \zeta)^2 = E[(\zeta' - \zeta_0)^2 + 2(\zeta' - \zeta_0)(\zeta_0 - \zeta) + (\zeta_0 - \zeta)^2].$$

Note que o termo $(\zeta_0 - \zeta)^2$ não envolve ζ' e que

$$E[(\zeta' - \zeta_0)(\zeta_0 - \zeta)] = E_{\boldsymbol{y}}[E_{\zeta|\boldsymbol{y}}[(\zeta' - \zeta_0)(\zeta_0 - \zeta)|\boldsymbol{y}]]$$

$$= E_{\boldsymbol{y}}[(\zeta'\zeta_0 - \zeta' E_{\zeta|\boldsymbol{y}}(\zeta|\boldsymbol{y}) - \zeta_0^2 + \zeta_0 E_{\zeta|\boldsymbol{y}}(\zeta|\boldsymbol{y}))]$$

$$= E_{\boldsymbol{y}}[\zeta'\zeta_0 - \zeta'\zeta_0 - \zeta_0^2 + \zeta_0^2] = 0.$$

Portanto, $E(\zeta' - \zeta)^2 = E(\zeta' - \zeta_0)^2 +$ termos que não dependem de ζ'. Como $E(\zeta' - \zeta_0)^2$ é sempre não negativo, encontra-se o mínimo escolhendo $\zeta' = \zeta_0 = E(\zeta|\boldsymbol{y})$, em outras palavras:

$$\zeta" = E(\zeta|\boldsymbol{y}), \tag{3.2}$$

é o MP de ζ.

Uma importante propriedade do MP é que ele é não viesado para ζ, no sentido que $E(E(\zeta|\boldsymbol{y})) = E(\zeta)$.

Para o caso dos MLGM, foi visto no Capítulo anterior em 2.6, que as respostas $y_{i1}, ..., y_{in_i}$ são condicionalmente independentes com densidade pertencente à família exponencial, e que $\alpha_1, ..., \alpha_q$ são variáveis aleatórias independentes com distribuição normal multivariada, no entanto, a fim de desenvolver 3.2 para um caso mais geral, será tomada uma distribuição

genérica $f_\alpha(.)$ para os efeitos aleatórios. Assim sendo, considere que S é um conjunto com $\{1, ..., q\}$, em que i pertence ao conjunto S. A esperança 3.2 pode ser descrita por

$$E(\zeta(\boldsymbol{\beta}, \alpha_S)|y_S) = \int \zeta(\boldsymbol{\beta}, \alpha_S) f(\alpha_S|y_S) d\alpha_S = \int \zeta(\boldsymbol{\beta}, \alpha_S) \frac{f(\alpha_S, y_S)}{f(y_S)} d\alpha_S =$$
$$\int \zeta(\boldsymbol{\beta}, \alpha_S) \frac{f(y_S|\alpha_S) f_\alpha(\alpha_S)}{f(y_S)} d\alpha_S = \frac{\int \zeta(\boldsymbol{\beta}, \alpha_S) f(y_S|\alpha_S) f_\alpha(\alpha_S) d\alpha_S}{f(y_S)}$$

como,

$$f(y_S) = \int f(y_S, \alpha_S) d\alpha_S = \int \frac{f(y_S, \alpha_S)}{f_\alpha(\alpha_S)} f_\alpha(\alpha_S) d\alpha_S = \int f(y_S|\alpha_S) f_\alpha(\alpha_S) d\alpha_S$$

tem-se que

$$\zeta" = \frac{\int \zeta(\boldsymbol{\beta}, \alpha_S) f(y_S|\alpha_S) f_\alpha(\alpha_S) d\alpha_S}{\int f(y_S|\alpha_S) f_\alpha(\alpha_S) d\alpha_S}. \tag{3.3}$$

No entanto, sabe-se que os $y_{ij}|\alpha_i$ são independentes, portanto, a função densidade condicional na forma da família exponencial é dada por

$$f(y_S|\alpha_S) = \prod_{i \in S} \prod_{j=1}^{n_i} f(y_{ij}|\alpha_i) = \prod_{i \in S} \prod_{j=1}^{n_i} \exp[a_{ij}\phi \{y_{ij}\theta_{ij} - b(\theta_{ij})\} + c(y_{ij}, a_{ij}\phi)]$$

e, da independência dos efeitos aleatórios, tem-se que a distribuição conjunta dos efeitos aleatórios pode ser escrita como:

$$f_\alpha(\alpha_S) = \prod_{i \in S} f_\alpha(\alpha_i).$$

Logo, pode-se escrever $\zeta"$ como

$$\zeta" = \frac{\int \zeta(\boldsymbol{\beta}, \alpha_S) \prod_{i \in S} \prod_{j=1}^{n_i} \exp[a_{ij}\phi \{y_{ij}\theta_{ij} - b(\theta_{ij})\} + c(y_{ij}, a_{ij}\phi)] \prod_{i \in S} f_\alpha(\alpha_i) d\alpha_i}{\int \prod_{i \in S} \prod_{j=1}^{n_i} \exp[a_{ij}\phi \{y_{ij}\theta_{ij} - b(\theta_{ij})\} + c(y_{ij}, a_{ij}\phi)] \prod_{i \in S} f_\alpha(\alpha_i) d\alpha_i}. \tag{3.4}$$

No entanto, o termo $\exp[c(y_{ij}, a_{ij}\phi)]$ encontra-se no numerador e no denominador de 3.4 e independe de α_i, assim sendo pode-se escrever

$$\zeta" = \frac{\int \zeta(\boldsymbol{\beta}, \alpha_S) \prod_{i \in S} \prod_{j=1}^{n_i} \exp[a_{ij}\phi \{y_{ij}\theta_{ij} - b(\theta_{ij})\}] \prod_{i \in S} f_\alpha(\alpha_i) d\alpha_i}{\int \prod_{i \in S} \prod_{j=1}^{n_i} \exp[a_{ij}\phi \{y_{ij}\theta_{ij} - b(\theta_{ij})\}] \prod_{i \in S} f_\alpha(\alpha_i) d\alpha_i}. \tag{3.5}$$

A dimensão das integrais envolvidas no denominador de 3.5 é de $r = dim(\alpha_i)$, enquanto no numerador é de ao menos sr, onde s denota a cardinalidade de S.

Se considerarmos o caso em que o subconjunto S é formado apenas pelo elemento i, 3.5 resume-se a

$$\zeta_i" = \frac{\int \zeta(\boldsymbol{\beta}, \alpha_S) \exp[\phi \sum_{j=1}^{n_i} a_{ij} \{y_{ij}\theta_{ij} - b(\theta_{ij})\}] f_\alpha(\alpha_i) d\alpha_i}{\int \exp[\phi \sum_{j=1}^{n_i} a_{ij} \{y_{ij}\theta_{ij} - b(\theta_{ij})\}] f_\alpha(\alpha_i) d\alpha_i}. \tag{3.6}$$

Para o caso Poisson misto e com $a_{ij} = 1$ chega-se, por fim, na seguinte expressão

$$\zeta_i" = \frac{\int \zeta(\boldsymbol{\beta}, \alpha_S) \exp[\sum_{j=1}^{n_i} \{y_{ij}\log(\mu_{ij}) - \mu_{ij}\}] f_\alpha(\alpha_i) d\alpha_i}{\int \exp[\sum_{j=1}^{n_i} \{y_{ij}\log(\mu_{ij}) - \mu_{ij}\}] f_\alpha(\alpha_i) d\alpha_i}. \tag{3.7}$$

Note que para se obter $\zeta"$ necessita-se dos verdadeiros valores de $\boldsymbol{\beta}$, o que não é viável na prática. Então, como alternativa, foi desenvolvido o Melhor Preditor Empírico (MPE), que substitui os valores de $\boldsymbol{\beta}$ por estimativas consistentes.

O método MP necessita também da suposição de que a distribuição de $\zeta|\boldsymbol{y}$ seja conhecida, existem outros métodos em que são feitas suposições mais fracas para a obtenção dos preditores, como pode ser visto em McCulloch *et al.* (2008), destacam-se:

- Método do Melhor Preditor Linear (MPL): que busca o melhor preditor dentro do conjunto dos preditores lineares. Tal método demanda apenas o conhecimento do primeiro e segundo momento de ζ e $\boldsymbol{y}$.

- Método do Melhor Preditor Linear Não Viciado (MPLNV): que busca o melhor preditor dentro do conjunto dos preditores lineares não viesados. Para este método é necessário se conhecer a $Var(\boldsymbol{y})$ e $Cov(\zeta, \boldsymbol{y})$.

Na literatura mais recente, métodos que também englobam o caso em que o interesse é a predição de grupos novos têm surgido.

Skrondal e Rabe-Hesketh (2009) propuseram utilizar a média do efeito aleatório para os novos efeitos aleatórios, isto é, atribuir o valor zero para as predições. Embora este método tenha como grande vantagem sua simplicidade, ele peca por ignorar a parte aleatória, o que

pode levar a conclusões errôneas.

Aplicar o método MPE, como em Tamura (2012), requer que métodos de integração numérica devam ser usados para resolver integrais multidimensionais, o que pode ser muito custoso computacionalmente.

Visando contornar estas dificuldades, Tamura e Giampaoli (2013) desenvolveram uma outra metodologia baseada em modelar o efeito aleatório do novo grupo a partir da dependência em relação às covariáveis agregadas no nível do grupo. Esta metodologia, denominada método de predição via modelos de regressão, envolve baixo custo computacional, contudo, nota-se que ela não considera possíveis correlações entre os efeitos aleatórios. No Capítulo 4, será apresentado um outro método que, além de possuir baixo custo computacional, também incorpora dependência entre os clusters.

Capítulo 4

Método de predição do vizinho mais próximo

Giampaoli *et al.* (2013) propuseram um novo método para a predição dos efeitos aleatórios para um novo grupo de um modelo logístico misto, usando a técnica dos vizinhos mais próximos ou, em inglês, Nearest Neighbors Prediction Method (NNPM). O método consiste na computação das distâncias entre o novo grupo e os grupos com efeitos aleatórios conhecidos baseados nos valores das covariáveis. Então, os efeitos aleatórios dos vizinhos mais próximos podem ser sumarizados por uma medida de centralidade afim de definir o valor dos efeitos aleatórios associados ao novo grupo.

Como em Giampaoli *et al.* (2013) foi considerado o caso logístico, as medidas de performance adotadas foram a área sob a curva ROC (AUC) e a estatística Kolmogorov-Smirnov (KS). Para maiores informações sobre estas medidas de desempenho ver Fawcett (2006) e Conover (1999).

O método NNPM foi aplicado a um banco de dados e também foram feitas simulações e os resultados obtidos foram satisfatórios tanto em Giampaoli *et al.* (2013), quanto em outro estudo realizado por Giampaoli *et al.* (2016b).

Tem-se então que o método NNPM apresentou bons resultados quando comparado a outros métodos, no entanto, as medidas de desempenho adotadas não podem ser usadas no modelo Poisson misto.

Definiremos o método de forma mais detalhada na seção seguinte.

4.1 O algoritmo NNPM

Considere um conjunto de grupos de um modelo misto em que o i-ésimo grupo é denotado por G_i com $i = 1, ..., q$. Para cada um deles existe um vetor de características associado g_i e um vetor de efeitos aleatórios estimados pelos métodos usuais de predição $\hat{\boldsymbol{\alpha}}_i = (\hat{\alpha}_{1i}, ..., \hat{\alpha}_{ki})$. Suponha agora um outro conjunto, no qual $G'_{i'}$ representa i'-ésimo novo grupo com $i' = 1, ..., q'$, sendo que cada um deles possui um vetor de características $g'_{i'}$ e um vetor de efeitos

aleatórios desconhecidos $\boldsymbol{\alpha_{i'}}$. O objetivo é predizer os valores dos efeitos aleatórios para o i'-ésimo novo grupo representado por $\boldsymbol{\alpha_{i'}}$. Abaixo segue a descrição do algoritmo:

1: Para i' de 1 até q' {

2: Para i de 1 até q {

3: Computar a distância $d_{(i',i)}$ entre $g'_{i'}$ e g_i;

4: }

5: Dispor os elementos $d_{(i',.)} = (d_{(i',1)}, d_{(i',2)}, ..., d_{(i',q)})$ em ordem crescente;

6: }

7: Para L de 1 até q {

8: Para i' de 1 até q' {

9: Computar uma medida de centralidade dos efeitos aleatórios conhecidos correspondentes aos L primeiros elementos ordenados de $d_{(i',.)}$ para produzir
$\tilde{\boldsymbol{\alpha}}_{i'} = (\tilde{\alpha}_{1.}, ..., \tilde{\alpha}_{k.})$;

10: Os valores preditos dos efeitos aleatórios $\tilde{\boldsymbol{\alpha}}_{i'}$ são inseridos no preditor linear do modelo logístico misto, fornecendo a probabilidade da resposta do i'-ésimo novo grupo no nível da observação;

11: }

12: }

13: Selecione o L que maximize a performance da predição no modelo logístico misto.

Em outras palavras tem-se então que o algoritmo funciona da seguinte forma:

1. Utilizando o valor das covariáveis, as linhas de 2-5 computam uma medida de distância entre um grupo novo $G'_{i'}$ e cada um dos grupos cujos efeitos aleatórios já foram preditos e as ordena em ordem crescente no vetor $d_{(i',.)}$. As linhas 1 e 6 fazem com que tal procedimento seja repetido para todos os grupos novos.

2. Para L indo de 1 a q, as linhas de 7-12 calculam uma medida de centralidade dos efeitos aleatórios conhecidos correspondentes aos L primeiros elementos ordenados de $d_{(i',.)}$, resultando em $\tilde{\boldsymbol{\alpha}}_{i'}$. Os valores $\tilde{\boldsymbol{\alpha}}_{i'}$ são inseridos no preditor linear no modelo logístico

misto e calcula-se o desempenho de predição.

3. Na linha 13, escolhe-se o valor de L que obteve a maior performance de predição. Portanto, os valores preditos para os efeitos aleatórios do novo grupo são obtidos de alguma medida central dos L grupos mais próximos, em que L é o número de vizinhos mais próximos que maximiza o desempenho de predição do modelo logístico misto.

Nota-se que o algoritmo proposto necessita de medidas de distância e de centralidade. Nas seções seguintes apresentamos algumas propostas de tais medidas, comuns na literatura, relacionadas a análise multivariada.

4.2 Medidas de distância

Medidas de distância podem ser usadas para avaliar o grau de proximidade entre observações. Existem dois tipos de medidas de distância: similaridade, quanto maior o valor, maior a semelhança entre os objetos, e dissimilaridade, quanto maior o valor, mais diferentes são os objetos.

Giampaoli *et al.* (2013) optaram por utilizar as seguintes medidas de dissimilaridade: Euclidiana, City Block, Minkowski, e Mahalanobis, que serão definidas a seguir:

Seja uma matriz $\mathbf{X}$ (q x p) com q grupos e p covariáveis, ou seja, cada grupo j tem um vetor de covariáveis associado, dado por $\mathbf{x}_j^T = (x_{j1}, x_{j2}, ..., x_{jp})$, o objetivo é calcular uma medida de distância entre um grupo i' e i.

- Euclidiana: é a distância mais popular entre dois objetos e é definida por

$$d^e_{(i'i)} = \sqrt{(\mathbf{x}_{i'} - \mathbf{x}_i)'(\mathbf{x}_{i'} - \mathbf{x}_i)} \tag{4.1}$$

- City Block: é mais simples que a distância Euclidiana, e é definida da seguinte forma

$$d^{cb}_{(i'i)} = |\mathbf{x}_{i'} - \mathbf{x}_i| \tag{4.2}$$

- Minkowski: é similar à distância Euclidiana, mas ao invés de utilizar a raiz quadrada, usa-se a λ-ésima raiz e é definida por

$$d^{mi}_{(i'i)} = (\sum_{m=1}^{p} |x_{i'm} - x_{im}|^{\lambda})^{1/\lambda} \tag{4.3}$$

- Mahalanobis: é ponderada pela matriz de covariância $\mathbf{\Sigma}$ e definida por

$$d^{ma}_{(i'i)} = (\mathbf{x}_{i'} - \mathbf{x}_i)'\mathbf{\Sigma}^{-1}(\mathbf{x}_{i'} - \mathbf{x}_i) \tag{4.4}$$

Note que as medidas Euclidiana e City Block são casos particulares da medida de Minkowski, sendo obtidas fazendo, respectivamente, $\lambda = 2$ e $\lambda = 1$. A escolha do valor de λ depende apenas da ênfase que se deseja dar a distâncias maiores, ou seja, quanto maior o valor de λ, maior a sensitividade da métrica a distâncias maiores, como pode ser visto em Johnson e Wichern (2007). Já a distância de Mahalanobis se diferencia das demais por levar em consideração a correlação entre o conjunto de dados.

4.3 Medidas de centralidade

Giampaoli *et al.* (2013) consideraram as seguintes medidas de centralidade, usuais na literatura, mas apresentadas aqui em função dos efeitos aleatórios para uma melhor compreensão do método NNPM:

Considere os efeitos aelatórios $\hat{\boldsymbol{\alpha}}_1, \hat{\boldsymbol{\alpha}}_2, ..., \hat{\boldsymbol{\alpha}}_L$, deseja-se obter uma medida de centralidade entre os L efeitos.

- Média: a média aritmética dos L efeitos aleatórios é definida por

$$\tilde{\boldsymbol{\alpha}}_{i'}^{a} = \sum_{l=1}^{L} \hat{\boldsymbol{\alpha}}_l / L. \tag{4.5}$$

- Medoide: o efeito aleatório com menor distância em relação a $\tilde{\boldsymbol{\alpha}}_{i'}^{a}$, será denotado por $\tilde{\boldsymbol{\alpha}}_{i'}^{m}$.

- Distância inversamente proporcional (PID na sigla em inglês): uma média ponderada pelo inverso da distância dos efeitos aleatórios definida por

$$\tilde{\boldsymbol{\alpha}}_{i'}^{pid} = \sum_{l=1}^{L} \hat{\boldsymbol{\alpha}}_l \frac{1}{d_{i'l}} \frac{1}{\sum_{l=1}^{L} \frac{1}{d_{i'l}}}. \tag{4.6}$$

- PID Medoide: o efeito aleatório com menor distância em relação ao PID, denotado por $\tilde{\boldsymbol{\alpha}}_{i'}^{pidm}$.

- Distância exponencialmente proporcional (PED na sigla em inglês): uma média ponderada pela função exponencial dos efeitos aleatórios definida por

$$\tilde{\boldsymbol{\alpha}}_{i'}^{ped} = \sum_{l=1}^{L} \hat{\boldsymbol{\alpha}}_l e^{-d_{i'l}} \frac{1}{\sum_{l=1}^{L} e^{-d_{i'l}}}. \tag{4.7}$$

- PED Medoide: o efeito aleatório com menor distância em relação ao PED, denotado por $\tilde{\boldsymbol{\alpha}}_{i'}^{pedm}$.

- Mediana: a mediana dos L efeitos aleatórios denotada por $\tilde{\boldsymbol{\alpha}}_{i'}^{md}$.

A etapa do método NNPM correspondente à escolha do valor L e medidas de distância e centralidade que maximizam o desempenho de predição será denominada por: etapa de avaliação do modelo.

Após tal estágio e a averiguação de que o método possui bom desempenho preditivo, para a predição de novos dados será aplicado diretamente o método NNPM utilizando o valor de L e as medidas de distância e centralidade fixadas nesta fase.

Assim sendo, é importante que depois de certo período seja repetida a etapa de avaliação do modelo, a partir de uma nova coleta de dados, considerando uma Base de Ajuste e uma Base de Predição, esta forma de dividir uma base de dados será definida melhor na seção 5.2.2.

Munidos dos conceitos e definições das seções anteriores, nos próximos Capítulos, estenderemos o método NNPM para o modelo Poisson misto e será sugerida obtenção de intervalos de confiança a partir do método Bootstrap.

Capítulo 5

NNPM para o caso Poisson

5.1 Aplicação do método Bootstrap junto ao NNPM

A título de exemplificação, o método NNPM será aplicado a um banco de dados referente ao tratamento de diabetes. Este mesmo exemplo será retomado de forma mais detalhada no Capítulo 7, de modo que neste Capítulo ele será visto de maneira mais simplificada, apenas com o objetivo de mostrar como serão obtidos intervalos de confiança para os efeitos aleatórios no caso Poisson utilizando o método Bootstrap. A Tabela 5.1 contém as variáveis que serão consideradas:

Tabela 5.1: *Variáveis consideradas na aplicação do Bootstrap junto ao método NNPM*

Variável	Descrição
Resposta (y)	Número de casos de hipoglicemia que o paciente teve
Insulina (x_1)	Tipo de insulina ao qual cada paciente foi submetido
Dose (x_2)	Quantidade de insulina utilizada no tratamento
Dieta (x_3)	Se o paciente segue as recomendações de um nutricionista ou não
Medição (x_4)	A quantidade média de medições de glicemia durante o dia
Gênero (x_5)	Masculino ou feminino
Cidade (x_6)	Cidade ao qual o hospital em estudo pertence

Ao total são vinte cidades, quinze foram incluídas à Base de Ajuste, e o método NNPM será aplicado com o intuito de predizer as outras cinco. A Base de Ajuste será denotada da seguinte forma:

$(\boldsymbol{z}_1, \boldsymbol{z}_2, ..., \boldsymbol{z}_n)$, no qual n é o número de pacientes no estudo e z_i são todos os valores associados ao i-ésimo paciente, ou seja: $\boldsymbol{z}_i = (y_i, x_{1i}, x_{2i}, x_{3i}, x_{4i}, x_{5i}, x_{6i})$.

A seguir, será esquematizado o método de Bootstrap de forma análoga ao feito na seção 2.4.

1. Selecionar a amostra Bootstrap:

   É selecionado uma amostra com reposição de $(\boldsymbol{z}_1, \boldsymbol{z}_2, ..., \boldsymbol{z}_n)$ gerando: $(\boldsymbol{z}_1^*, \boldsymbol{z}_2^*, ..., \boldsymbol{z}_n^*)$

2. Predizer os efeitos:

   Prediz-se os efeitos aleatórios das cinco cidades da Base de Predição utilizando o método NNPM.

3. Repetição:

   Os passos anteriores foram repetidos cem vezes, gerando cem predições para cada um dos cinco efeitos aleatórios.

4. Intervalo:

   Para cada uma das cidades, organizamos os valores preditos para seu respectivo efeito em ordem crescente.

   Depois os valores de percentis de $\frac{\alpha}{2}$ e $(1 - \frac{\alpha}{2})$ são determinados, culminando num intervalo de $(1 - \alpha)$ de confiança para cada efeito.

5.2 Medidas de desempenho de predição

No Capítulo anterior, foi visto que para avaliar o desempenho do modelo precisa-se adotar alguma medida de performance de predição que nos permita escolher entre diferentes modelos. No entanto, quando se trata da distribuição de Poisson, não se pode utilizar a curva ROC ou a medida KS, visto que são métodos específicos para dados binários.

Nesta seção, serão apresentadas medidas que podem ser utilizadas para casos mais gerais.

5.2.1 Proposta I

Em Sheiner e Beal (1981), discute-se formas adequadas para a comparação da performance de modelos, assim como procedimentos não aconselhados. Uma importante medida com tal finalidade chama-se QMEP, definido da seguinte forma

$$\sum_{i}^{N} (Y_i - Y_i^*)^2/N, \tag{5.1}$$

em que N é o total de elementos preditos, Y_i é o valor real da variável resposta contido na Base de Predição e Y_i^* é a predição para esse termo.

Logicamente, quanto menor for o valor da medida QMEP, melhor será o desempenho preditivo.

5.2.2 Proposta II

Embora a medida de desempenho vista em 5.1 seja útil para se comparar métodos de predição, é importante também que existam medidas de desempenho alternativas, de modo que possamos compará-las e termos maior informação sobre a qualidade da predição.

As medidas RB e RE, vistas no seção 2.5, poderiam ser úteis, contudo, elas necessitam dos verdadeiros valores dos parâmetros, algo só possível no âmbito da simulação.

Ainda no caso em que os valores populacionais são conhecidos, a seguir, uma nova medida de desempenho será proposta

$$\sum_{i}^{N} \boldsymbol{I}_{A_i}(\mu_i)/N. \tag{5.2}$$

No qual, N representa novamente o número de elementos preditos, μ_i o valor da média populacional, isto é, $\mu_i = E(Y_i/\boldsymbol{\alpha_i})$, A_i trata-se do intervalo de Bootstrap feito para μ_i e a função $\boldsymbol{I}_{A_i}(\mu_i)$ é tal que

$$\boldsymbol{I}_{A_i}(\mu_i) = \begin{cases} 1, & \mu_i \in A_i \\ 0, & \mu_i \notin A_i. \end{cases}$$

Em outras palavras, a medida 5.2 refere-se à porcentagem de vezes que o intervalo de Bootstrap feito para predizer μ_i contém o verdadeiro valor.

Quando se deseja avaliar métodos de predição, é usual separar a banco de dados em duas partes: Base de Ajuste (BA) e Base de Predição (BP). A primeira é usada para o ajuste do modelo e aplicação do método e a última para a validação do modelo, isto é, usamos BA para predizer os valores de BP, podendo, assim, averiguar se os valores preditos se aproximam dos valores reais.

Dessa forma, embora não sejam conhecidos os valores populacionais, os valores da variável resposta são acessíveis.

Levando em consideração tais fatores, será proposto nesta dissertação a seguinte medida de desempenho, baseada em 5.2

$$Porcentagem = \sum_{i}^{N} \boldsymbol{I}_{A_i}(\mu_i^{MP})/N. \tag{5.3}$$

Sendo que, μ_i^{MP} é a predição de μ_i utilizando o método MPE calculado utilizando o resultado obtido em 3.5.

Como a medida 5.3 independe de valores populacionais, ela será adotada também como medida de desempenho para a utilização do método NNPM nas aplicações, no entanto, é importante se fazer as seguintes considerações:

- Apesar de o cálculo da predição pelo método MPE ser mais pesado computacionalmente, usá-lo apenas como medida de desempenho faz com que só seja necessário realizar tais cálculos na etapa de avaliação do método NNPM, isto é, na etapa em que serão selecionados a quantidade L de vizinhos e as medidas de distância e centralidade ótimas. Após a escolha de tais quantidades, a predição para os dados futuros passará a ser feita a partir destes.
- Utilizar o MPE apenas na etapa de avaliação do modelo faz com que não se tenha o problema dos valores da resposta serem desconhecidos, posto que, o banco de dados foi dividido em BA e BP.
- Ao contrário do que ocorre no método NNPM, o MPE necessita da especificação da distribuição dos efeitos aleatórios. Todavia, mesmo se for assumida a normalidade, McCulloch e Neuhaus (2011) e Neuhaus *et al.* (2013) mostraram que o MLGM é robusto quanto a fuga desta suposição.
- Vale lembrar que os intervalos de confiança de Bootstrap são feitos para os valores populacionais de μ_i e não valores preditos dele, o que faz com que 5.3 deva ser visto com cautela. No entanto, sabe-se que o MP é um preditor não viesado, portanto, para tamanhos grandes de amostra espera-se que o preditor se aproxime do valor populacional.

Evidentemente, valores de 5.3 próximos de 1 indicam bom desempenho preditivo, porém, caso a amplitude dos intervalos seja muito grande, os resultados obtidos serão pouco informativos.

Tendo em vista este problema, as seguintes medidas também serão considerados nesta dissertação

$$Amplitude = \sum_{i}^{N} (LS_i - LI_i)/N \tag{5.4}$$

e

$$PA = Porcentagem/Amplitude. \tag{5.5}$$

LS_i e LI_i representam, respectivamente, os limites superior e inferior dos intervalos de Bootstrap feitos para μ_i.

Espera-se, portanto, um bom desempenho preditivo quando a *Porcentagem* for próxima de 1 e a *Amplitude* relativamente pequena, analogamente, valores maiores de PA indicam melhor poder preditivo.

Para a aplicação do algoritmo NNPM explicitado na seção 4.1 com resposta Poisson, adotar-se-á apenas a medida de desempenho 5.1 devido à sua simplicidade e ganhos em aspectos de velocidade computacional, no entanto, após a predição os resultados serão avaliados utilizando tanto a medida QMEP como também as medidas de *Porcentagem*, *Amplitude* e PA.

Capítulo 6

Estudos de simulação

6.1 Aspectos iniciais

Como já foi dito nos Capítulos anteriores, o método NNPM foi aplicado em outros trabalhos apenas no modelo logístico, com o intuito de avaliar o desempenho do método NNPM para a predição dos efeitos aleatórios em um modelo Poisson misto, será realizado um estudo de simulação considerando o seguinte modelo

$$log(\mu_{ij}) = \beta_0 + \beta_1 x_{ij} + \alpha_i, \tag{6.1}$$

com $i = 1, ..., n$, $j = 1, ..., n_i$ e $\alpha_i \sim N(0, \sigma_a)$.

Sendo que o modelo em 6.1 é um caso particular do modelo explicitado em 2.5, de modo que as mesmas notações e suposições utilizadas prevalecem. Relembrando que n representa o número de clusters e n_i trata-se do número de observações dentro de cada cluster.

Sem perda de generalidade, os estudos foram realizados adotando a distância Euclidiana como medida de distância e o PID Medoide como medida de centralidade. Foram considerados dois grupos de cenários:

- Cenário 1

  Aqui foi fixado o valor $n = 50$ e $n_i = 10$, as simulações foram feitas para os seguintes valores populacionais dos parâmetros:

  - $\beta_1 = 1; \beta_0 = 1$

  - $\beta_1 = 1; \beta_0 = -1$

  - $\beta_1 = 2; \beta_0 = 2$

  - $\beta_1 = 2; \beta_0 = -2$

  - $\beta_1 = 4; \beta_0 = 3$

- $\beta_1 = 4; \beta_0 = -3$

Sendo que para cada uma dessas combinações o desvio padrão σ_a assumirá os seguintes valores: $0,25; 0,50; 1; 2$.

- Cenário 2

Neste grupo de simulações, será comparada o desempenho de predição na simulação pelo método NNPM com os resultados obtidos em Boubeta *et al.* (2004), no qual os seguintes casos foram analisados:

- $n = 50$

- $n = 100$

- $n = 150$

Com os valores $\beta_1 = 0,6; \beta_0 = -2; \sigma_a = 0,5$ fixados.

Na seção 6.2, será explicitado o esquema de simulação utilizado para ambos os cenários, e nas seções 6.3 e 6.4, os resultados obtidos para cada um deles. Por se tratar de um estudo de simulação, não serão feitos intervalos de confiança de Bootstrap e nem calculadas as medidas de desempenho definidas no Capítulo 5, visto que é possível comparar os valores preditos diretamente com os seus respectivos valores populacionais através das medidas de desempenho RB e RE, vistas no Capítulo 2.

6.2 Estudos de simulação

Seguem os passos seguidos nas simulações:

1. Geração das covariáveis: Os valores de x_{ij} foram obtidos de uma normal de variância $0,25$ e média $(i - 25)/100$. As covariáveis foram geradas dessa forma para que elas possam conter informações diferentes referentes a um dado cluster.

2. Geração dos efeitos aleatórios: Os efeitos aleatórios α_i dos n cluster são obtidos de uma normal de média 0 e desvio padrão σ_a.

3. Cálculo das médias exatas:

   $\mu_{ij} = \exp(\beta_0 + \beta_1 x_{ij} + \alpha_i)$

4. Obtenção das observações da Base de Ajuste: Consideraremos que as primeiras $n_i/2$ observações de cada um dos n clusters fazem parte da Base de Ajuste, enquanto o restante da Base de Predição. Geramos tais valores a partir de uma distribuição Poisson com sua respectiva média calculada no passo 3.

5. Ajuste do modelo: Utilizamos a Base de Ajuste para as estimativas $\hat{\beta}_0$ e $\hat{\beta}_1$, e as predições de $\hat{\alpha}_i$.

   É importante ressaltar que dada a natureza da simulação, sabemos o cluster ao qual pertencem as observações que foram deixadas na Base de Predição, no entanto, veremos a capacidade do método NNPM de obter predições condizentes para os efeitos aleatórios correspondentes.

6. Utilização do método NNPM: Nesta etapa o valor dos efeitos aleatórios dos grupos da Base de Predição são preditos através do método do vizinho mais próximo.

O passo 2 ao 6 foi repetido $K = 1000$ vezes.

6.3 Resultados do Cenário 1

A Tabela 6.1 contém os valores RB e RE, definidos no Capítulo 2, considerando os valores populacionais do primeiro Cenário. Tais valores foram calculados por meio do método NNPM, sendo obtidos tanto para a predição dos efeitos, quanto para as médias.

Dos resultados, podem ser feitas as seguintes observações:

- Efeitos

  Os valores obtidos não foram muito satisfatórios no geral, tanto para os RB, que parecem oscilar em torno de 1, quanto para os RE, que apresentam alguns valores ainda mais altos. Destacam-se os resultados:

  - RB: no caso em que $\beta_1 = 2$ e $\beta_0 = 2$, o menor valor, $0,465$, foi atingido para o maior valor de σ_a, enquanto seria natural supor que menores valores de σ_a culminariam em melhor desempenho; para $\beta_1 = 4$ e $\beta_0 = 3$, o valor do RB em $\sigma_a = 0,25$ foi de $1,194$ o que é um pouco mais alto do que o esperado e quando $\sigma_a = 2$ foi obtido o maior resultado em módulo, $-3,193$, que ainda se trata de um valor razoável.

  - RE: para $\beta_1 = 1$ e $\beta_0 = -1$, e $\beta_1 = 2$ e $\beta_0 = -2$ os valores abaixam com o aumento de σ_a; para $\beta_1 = 4$ e $\beta_0 = 3$ obtiveram-se os valores mais altos desta medida com números superiores a 7, sendo $7,901$ quando $\sigma_a = 1$ e $7,352$ em $\sigma_a = 2$; para $\beta_1 = 4$ e $\beta_0 = -3$, os valores para $\sigma_a = 2$ e $\sigma_a = 0,5$ foram de $0,213$ e $4,825$, portanto, bem fora do esperado.

- Médias

  Os resultados obtidos foram predominantemente próximos de zero, o que indica um bom desempenho preditivo. Os valores RB e RE tornam-se mais precisos para menores valores de σ_a, no caso dos RB, quando $\sigma_a = 2$ os valores foram em módulo consideravelmente maiores.

Tabela 6.1: *Valores de RB e RE calculados para as predições feitas pelo método NNPM para o Cenário 1*

β_1	β_0	σ_a	Efeitos		Média	
			RB	RE	RB	RE
1	1	0,25	1,080	1,098	-0,018	0,008
		0,5	0,883	0,705	-0,090	0,019
		1	1,232	0,799	-0,486	0,062
		2	0,934	0,904	2,442	0,577
	-1	0,25	1,297	1,226	-0,023	0,010
		0,5	1,279	0,632	-0,087	0,019
		1	1,029	0,318	-0,402	0,057
		2	0,974	0,271	-3,084	0,727
2	2	0,25	0,822	1,373	-0,024	0,008
		0,5	1,428	1,895	-0,104	0,019
		1	0,759	2,006	-0,490	0,060
		2	0,465	1,891	-3,111	0,946
	-2	0,25	1,030	0,622	-0,030	0,012
		0,5	0,865	0,557	-0,083	0,021
		1	1,023	0,290	-0,382	0,056
		2	1,002	0,201	-2,777	0,693
4	3	0,25	1,194	2,386	-0,021	0,008
		0,5	0,400	4,432	-0,100	0,019
		1	-0,240	7,901	-0,488	0,064
		2	-3,193	7,352	-3,203	0,920
	-3	0,25	0,903	2,259	-0,020	0,018
		0,5	1,919	4,825	-0,106	0,025
		1	-0,717	2,533	-0,412	0,059
		2	0,948	0,213	-2,265	0,578

Percebe-se então que embora as predições para os Efeitos não tenham sido muito boas, na parte mais importante, que é a predição das Médias, o desempenho preditivo foi muito bom, exceto para o maior valor de σ_a.

6.4 Resultados do Cenário 2

Em Boubeta *et al.* (2004), foi considerado o problema de predição usual, isto é, para clusters já presentes na amostra. Utilizou-se então o método do MPE, a Tabela 6.2 representa a comparação entre os valores RB e RE encontrados no estudo de Boubeta, Lombardía e Morales (2004), com os obtidos na presente dissertação utilizando o método NNPM.

Pode-se notar que os valores de RB para a predição dos efeitos foram mais próximos de zero em Boubeta *et al.* (2004) do que quando utilizado o método NNPM, indicando que Boubeta *et al.* (2004) obteve melhores resultados neste caso. O contrário aconteceu na

Tabela 6.2: *Comparação dos valores RB e RE encontrados pelo método MPE em Boubeta et al.(2004), com os obtidos via NNPM, considerando o Cenário 2*

	RB				RE			
	Boubeta et al		NNPM		Boubeta et al		NNPM	
Tamanho	Efeitos	Médias	Efeitos	Médias	Efeitos	Médias	Efeitos	Médias
n=50	-0,02	-1,91	0,98	-0,07	0,64	0,65	0,64	0,02
n=100	0,05	-1,88	1,04	-0,09	0,64	0,64	0,67	0,02
n=150	0,05	-1,88	1,11	-0,10	0,64	0,64	1,00	0,03

predição das médias, que pode ser considerado o mais importante, o método NNPM mostrou-se mais satisfatório, isto é, com valores de RB mais próximos de zero.

Em relação à medida RE, os valores para a predição dos efeitos deram consideravelmente próximos. Para os resultados da média, novamente o método NNPM apresentou resultados melhores.

Vale lembrar que a comparação feita aqui é relativa, pois os valores das covariáveis utilizados nas simulações são diferentes.

Capítulo 7

Aplicações

Com o intuito de ilustrar o funcionamento do método NNPM, dois bancos de dados reais serão analisados neste Capítulo: um na área da saúde e o outro na área de administração de negócios.

O primeiro banco de dados trata de um estudo envolvendo pessoas com diabetes, no qual, o interesse é compreender melhor os fatores que influenciam no número de casos de hipoglicemia, um distúrbio provocado pela baixa concentração de glicose no sangue, que pode ter consequências sérias para o paciente como o desencadeamento de convulsões. Um levantamento multicêntrico foi realizado considerando pessoas de diferentes regiões do Brasil, é razoável supor que pacientes tratados numa mesma cidade sejam correlacionados por compartilharem condições ambientais, portanto, a cidade em que o paciente foi tratado entrará no modelo como efeito aleatório.

A segunda pesquisa busca compreender as características institucionais de uma cidade e suas relações com a distribuição de lojas fast food, tais conhecimentos podem auxiliar gestores como, por exemplo, na viabilidade de abrir novas lojas em determinado local. Novamente, a cidade em que cada franquia está instalada será considerada como fator de agrupamento.

Nas próximas seções, ambos os estudos serão apresentados com mais detalhes, juntamente com as análises e a utilização do método NNPM, que é o principal objetivo deste Capítulo.

7.1 Aplicação na área de diabetes: casos de hipoglicemia

A diabetes do tipo 1 ocorre quando o pâncreas perde a capacidade de produzir insulina em decorrência de um problema no sistema imunológico, fazendo com que nossos anticorpos ataquem as células que produzem esse hormônio. Como a insulina é necessária para levar o açúcar do sangue às células, onde a glicose poderá ser estocada ou usada como fonte de energia, a falta de insulina no sangue faz com que a glicose não entre nas células, se acumulando no sangue.

Tal quadro faz com que a pessoa com diabetes do tipo 1 deva manter uma vida saudável e o controle constante da glicemia, a fim de evitar possíveis complicações da doença como, por exemplo, a hipoglicemia.

A hipoglicemia é um distúrbio provocado pela baixa concentração de glicose no sangue, que pode afetar pessoas com diabetes ou não, no entanto, indivíduos com diabetes são mais propensos a sofrerem com esse tipo de problema. O que ocorre é que, em se tratando de pessoas com diabetes, as taxas de glicose presentes na corrente sanguínea estão mais altas que o normal. Para controlar esse problema, alguns medicamentos ou insulina são prescritos pelos médicos para ajudar na regulação dos níveis de açúcar no sangue.

A dose de insulina que a pessoa irá aplicar depende sempre da quantidade de carboidrato ingerido no dia, além de outros fatores como, por exemplo, atividade física. Pode acontecer, todavia, de um paciente acabar injetando mais insulina do que o necessário, levando-a a um quadro de hipoglicemia. Mesmo que a dose ministrada tenha sido a ideal para manter as taxas de glicose no sangue estáveis, se a pessoa não comer uma quantidade razoável durante as refeições ela também pode acabar apresentando um episódio de hipoglicemia, pois não terá produzido glicose suficiente.

Ignorar os sintomas da hipoglicemia pode acarretar em consequências graves, como a perda de consciência e convulsões, visto que, o cérebro necessita de glicose para funcionar apropriadamente. Assim sendo, é de vital importância estudos que possam ajudar a entender melhor os casos de hipoglicemia num indivíduo com diabetes.

Com este intuito, foi considerada uma base de dados multicêntrica obtida do National Brazilian Health Care System (NBHCS) entre o período de dezembro de 2008 e dezembro de 2010. O número de pacientes considerados no estudo de cada região foi determinado baseado na prevalência estimada de diabetes do tipo 1 no Brasil e na densidade populacional em cada região geográfica. Os dados foram recolhidos através de entrevistas durante as visitas clinicas, sendo que, só foram considerados pacientes com ao menos 12 meses de acompanhamento.

Como o foco maior desta dissertação é na área de predição de novos grupos e não seleção de modelos específicos, serão consideradas apenas as seguinte variáveis:

- Total de casos de hipoglicemia: é a variável resposta, que é a quantidade de casos de hipoglicemia registrada no último mês.
- Tratamento: variável categórica, que denota o tipo de insulina usado pelo paciente, sendo 11 tipos de tratamentos.
- Dieta: se o paciente seguiu a dieta recomendada,

  - Não

- Sim.

- Medição de glicemia: uma variável categorizada que mede quantas vezes a glicemia do paciente é medida em média por dia. As categorias são,

  - zero

  - entre zero e um

  - um

  - dois

  - três

  - quatro

  - cinco

  - seis

  - entre sete e oito

  - oito ou mais.

- Gênero: Feminino; Masculino.

- Dose: variável quantitativa, que mede a dose total de insulina aplicada.

- Idade: idade do paciente medida em anos.

- Cidade: variável em que será considerado o agrupamento, ou seja, o efeito aleatório. Ela representa a cidade em que cada paciente foi tratado,

  - Aracaju

  - Bauru

  - Belém

  - Belo Horizonte

  - Brasília

  - Campina Grande

  - Curitiba

  - Florianópolis

  - Fortaleza

  - Goiânia

  - Joinville

  - Londrina

  - Manaus

- Porto Alegre
- Ribeirão Preto
- Rio de Janeiro
- Salvador
- São José do Rio Preto
- São Luís
- São Paulo.

As cidades que corresponderão às Bases de Ajuste e de Predição foram selecionadas de maneira aleatória. A Base de Ajuste será composta de 1740 pacientes, divididos entre as cidades de Aracaju (21 pacientes), Bauru (50 pacientes), Belém (22 pacientes), Belo Horizonte (37 pacientes), Campina Grande (30 pacientes), Curitiba (65 pacientes), Florianópolis (25 pacientes) , Fortaleza (187 pacientes), Londrina (76 pacientes), Porto Alegre (329 pacientes), Ribeirão Preto (49 pacientes), Rio de Janeiro (293 pacientes), Salvador (130 pacientes), São José do Rio Preto (25 pacientes) e São Paulo (401 pacientes). As cidades de Brasília (39 pacientes), Goiânia (18 pacientes), Joinville (23 pacientes), Manaus (17 pacientes) e São Luís (9 pacientes) serão tomadas como fazendo parte da Base de Predição, totalizando 106 pacientes. Ou seja, usaremos as observações de BA a fim de predizer os efeitos aleatórios associados às cidades de BP.

Para se ter uma noção preliminar acerca do número de casos de hipoglicemia, apresenta-se, na Figura 7.1, um gráfico de boxplots separando a distribuição da variável resposta para cada cidade de BP.

Foi ajustado o seguinte modelo para BA, com suposições análogas ao que foi visto no modelo 2.5

$$log(\mu_{ij}) = \beta_0 + \mathbf{x}_{ij}^{\mathbf{T}}\boldsymbol{\beta} + \alpha_i, \tag{7.1}$$

em que μ_{ij} é o valor esperado do número de casos de hipoglicemia para o j-ésimo paciente da i-ésima cidade condicionado ao intercepto aleatório da i-ésima cidade, α_i ; $\mathbf{x}_{ij}$ o vetor de covariáveis conhecidas do j-ésimo paciente da i-ésima cidade, sendo que, para as variáveis categóricas foi considerada a parametrização da casela de referência; β_0 é o intercepto fixo e $\boldsymbol{\beta}$ são os parâmetros desconhecidos associados ao vetor $\mathbf{x}_{ij}$.

Na próxima seção, apresentam-se o diagnóstico e as estimativas dos parâmetros obtidas do modelo 7.1.

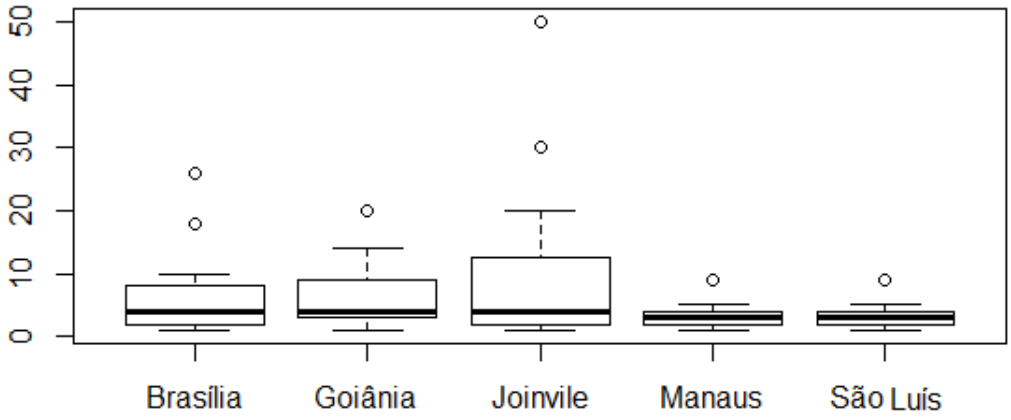

Figura 7.1: *Boxplots da distribuição do número de casos de hipoglicemia para cada cidade da BP.*

7.1.1 Estimação

Para cada variável categorizada, tomou-se a primeira categoria como casela de referência, as demais entraram no modelo como variáveis indicadoras. Os parâmetros associados às categorias de cada variável indicadora seguem abaixo:

- β_1: Tratamento 2
- β_2: Tratamento 3
- β_3: Tratamento 4
- β_4: Tratamento 5
- β_5: Tratamento 6
- β_6: Tratamento 7
- β_7: Tratamento 8
- β_8: Tratamento 9
- β_9: Tratamento 10
- β_{10}: Tratamento 11
- β_{11}: Dieta Sim
- β_{12}: Medição de glicemia entre zero e um

- β_{13}: Medição de glicemia um
- β_{14}: Medição de glicemia dois
- β_{15}: Medição de glicemia três
- β_{16}: Medição de glicemia quatro
- β_{17}: Medição de glicemia cinco
- β_{18}: Medição de glicemia seis
- β_{19}: Medição de glicemia entre sete e oito
- β_{20}: Medição de glicemia oito ou mais
- β_{21}: Gênero Masculino

Enquanto os parâmetros associados às variáveis quantitativas da parte fixa são:

- β_{22}: Dose
- β_{23}: Idade

Sendo que o β_0 trata-se do intercepto do modelo.

Foi utilizado o método de máxima verossimilhança para o ajuste do modelo, usando a aproximação de Laplace. As estimativas dos parâmetros, erros padrão e valores p do modelo ajustado podem ser observados na Tabela 7.1.

A partir da análise dos resultados obtidos, podem-se destacar alguns resultados:

- Para a variável Tratamento (parâmetros de β_1 ao β_{10}), considerando um nível de significância de 5%, tem-se que apenas as categorias Tratamento 3 (β_2), Tratamento 4 (β_3), Tratamento 10 (β_9) e Tratamento 11 (β_{10}) diferem da categoria Tratamento 1 (casela de referência), sendo que a categoria Tratamento 3 é a que ocasiona um maior aumento médio do número de casos de hipoglicemia.
- O sinal positivo de β_{11} (parâmetro associado à variável Dieta) indica que quem seguiu a dieta recomendada apresentou, em média, mais casos de hipoglicemia, o que vai contra ao que era esperado. Isso pode ter ocorrido por diversos motivos, dentre eles, o paciente pode achar que está seguindo a dieta recomendada quando na verdade está ingerindo uma quantidade menor, o que pode acarretar falta de açúcar no sangue.

Tabela 7.1: *Estimativas dos parâmetros, erro padrão e valor p para a Base de Ajuste*

Parâmetros	Estimativa	Erro Padrão	Valor P
β_0	0,963	0,181	<0,001
β_1	0,231	0,129	0,073
β_2	0,465	0,133	<0,001
β_3	0,412	0,141	0,003
β_4	0,082	0,128	0,525
β_5	0,332	0,240	0,165
β_6	0,123	0,124	0,320
β_7	0,185	0,129	0,150
β_8	0,255	0,174	0,144
β_9	0,349	0,123	0,004
β_{10}	0,325	0,123	0,008
β_{11}	0,148	0,027	<0,001
β_{12}	0,572	0,143	<0,001
β_{13}	0,446	0,116	<0,001
β_{14}	0,474	0,111	<0,001
β_{15}	0,596	0,110	<0,001
β_{16}	0,633	0,111	<0,001
β_{17}	0,866	0,114	<0,001
β_{18}	0,916	0,116	<0,001
β_{19}	0,802	0,122	<0,001
β_{20}	1,091	0,131	<0,001
β_{21}	-0,047	0,020	0,021
β_{22}	-0,006	0,000	<0,001
β_{23}	0,001	0,001	0,329

- Dos parâmetros de β_{12} até o β_{20} (parâmetros associados à variável Medição de glicemia), percebe-se que as estimativas tendem a aumentar para maiores quantidades de medições, ou seja, quanto maior a quantidade de medições de glicemia, maior é o valor esperado de casos de hipoglicemia. O que pode indicar que o paciente, por já ter histórico de hipoglicemia, realiza maior quantidade de medições para tentar se manter com níveis ideais de açúcar no sangue.

- Para a variável Gênero (β_{21}), como o valor p foi baixo $(2,1\%)$, a estimativa relativa à categoria Masculino é significativa, o que, dado ao seu sinal negativo, indica que pacientes homens têm, em média, menos ocorrências de hipoglicemia do que mulheres.

- O sinal negativo da estimativa da variável Dose (β_{22}) indica que com maiores doses de insulina aplicadas no paciente, uma menor quantidade de casos de hipoglicemia será esperada. Este resultado também é relativo, pois a quantidade de insulina depende do peso do paciente.

- Como o valor p da variável Idade (β_{23}) é alto, conclui-se que a variável Idade não é significante para hipoglicemia, em outras palavras, o aumento ou decréscimo de idade não influencia no número médio de casos de hipoglicemia.

Vale lembrar que os resultados acima não possuem valor conclusivo, uma vez que, não foi feita uma seleção de variáveis e nem considerou-se possíveis efeitos de interação. O objetivo real nesta dissertação é apenas apresentar as análises de predição, de modo que, o modelo aqui considerado trata-se de uma ilustração.

As predições dos efeitos aleatórios das cidades na BA, assim como a estimativa do desvio padrão, podem ser vistos na Tabela 7.2.

Tabela 7.2: *Predições dos efeitos aleatórios referentes à Base de Ajuste*

Efeitos	**Predições**
Aracaju	-0,275
Bauru	0,117
Belém	0,102
Belo Horizonte	-0,376
Campina Grande	-0,087
Curitiba	0,081
Florianópolis	0,348
Fortaleza	-0,272
Londrina	0,083
Porto Alegre	0,132
Ribeirão Preto	-0,283
Rio de Janeiro	0,208
Salvador	-0,371
São José do Rio Preto	0,394
São Paulo	0,224
Desvio padrão estimado do efeito 0,2585	

Predições maiores das cidades indicam que pacientes residentes em tais localidades apresentam valores esperados mais elevados de casos de hipoglicemia, assim sendo, a cidade de São José do Rio Preto (valor predito de 0,394) é a cidade mais agravante em termos de casos de hipoglicemia, enquanto a cidade de Belo Horizonte (valor predito de -0,376) é a que se espera menor número de ocorrências de hipoglicemia.

Para o diagnóstico do modelo 7.1 foram usados os quantis residuais normalizados aleatorizados definidos abaixo

$$r_{q_{ij}} = \Phi^{-1}(F(y_{ij}; \hat{\mu}_{ij})), \tag{7.2}$$

em que $i = 1, 2, ..., N$, Φ^{-1} é a função inversa da função de distribuição acumulada de uma variável com distribuição normal padrão e $F(y_{ij})$ é a função de distribuição acumulada de uma variável com distribuição Poisson(μ).

Rigby e Stasinopoulos (2005) usaram este tipo de resíduos a fim de comparar diferentes modelos, incluindo modelos mistos. Na Figura 7.2, são fornecidos os gráficos de diagnóstico. Os resíduos quantílicos se comportam satisfatoriamente, visto que, pelo gráfico (a) vemos uma disposição aleatória ao redor do zero, e pelo gráfico (b), nota-se que os resíduos se

dispõe ao longo da reta do QQ-plot normal. Pelo QQ-plot do gráfico (c), percebe-se que a suposição de normalidade dos efeitos aleatórios é plausível.

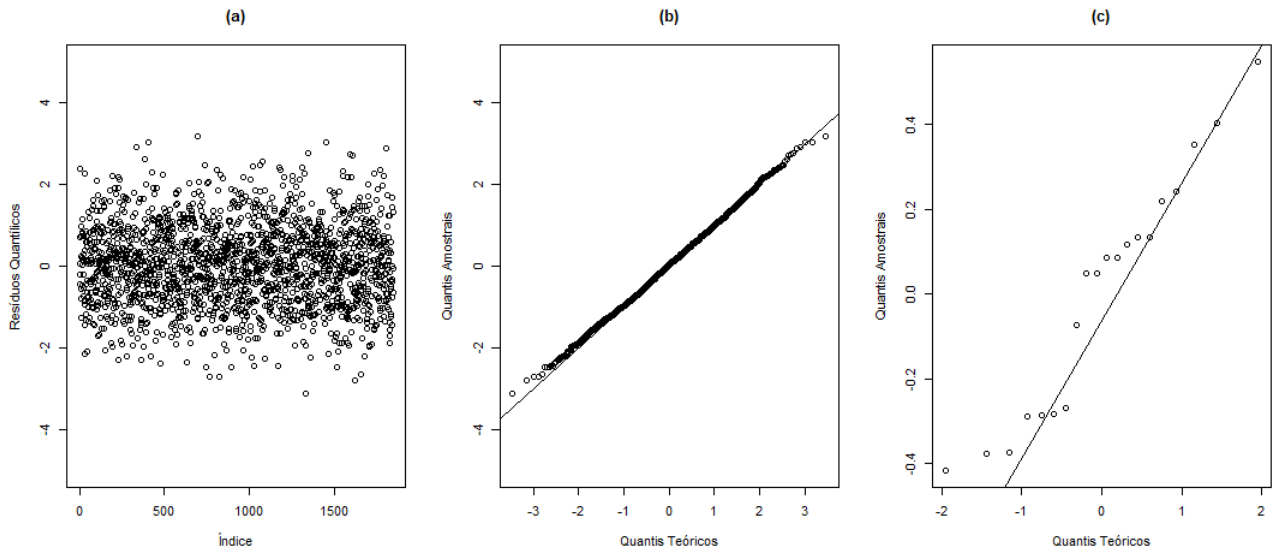

Figura 7.2: *Diagnóstico do modelo: (a) resíduos quantílicos vs índice, (b) QQ-plot dos resíduos quantílicos, (c) QQ-plot do efeito aleatório.*

7.1.2 Predição dos novos grupos

Para a predição de novos grupos foi utilizado o método NNPM considerando cada um dos cruzamentos das medidas de distância e centralidade apresentadas no Capítulo 4. Nos gráficos da Figura 7.3 estão apresentados os valores obtidos das medidas de desempenho QMEP, *Porcentagem*, *Amplitude* e PA definidas, respectivamente, em 5.1, 5.3, 5.4 e 5.5.

Para a medida QMEP, os valores mais baixos para a medida de centralidade Mediana indicam um melhor desempenho de predição.

Nota-se que, quanto a variável *Porcentagem*, com exceção dos resultados obtidos com a medida de centralidade Mediana, todos os cenários atingiram índices por volta de 80%, o que mostra um excelente desempenho preditivo.

Considerando a medida de centralidade Mediana, os resultados foram ainda melhores, aproximando-se de 100%, o que corrobora com o encontrado para a medida de desempenho QMEP.

Todavia, analisando o gráfico relativo à variável *Amplitude*, tem-se que as medidas de centralidade PED, PED Medoide e Mediana culminaram em intervalos de Bootstrap maiores do que as demais medidas de centralidade.

O gráfico da variável PA é interessante por levar em consideração tanto a porcentagem de vezes que o intervalo conteve a média, quanto a amplitude do intervalo. Neste aspecto,

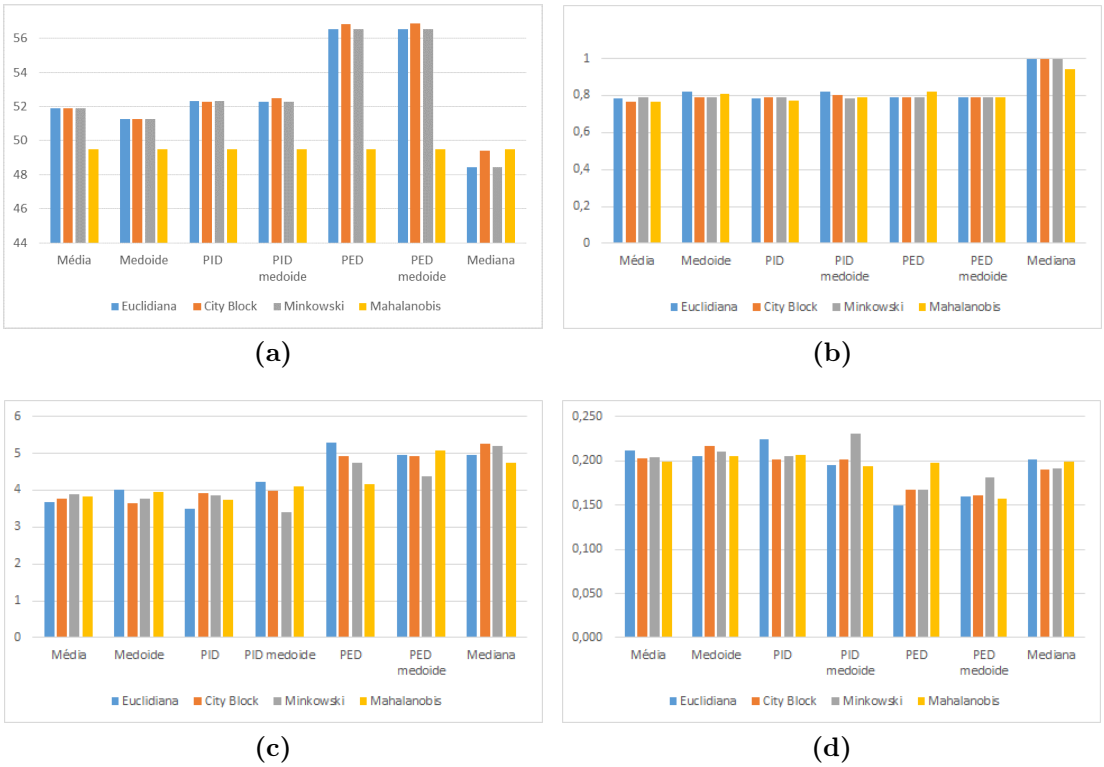

Figura 7.3: *Desempenho preditivo considerando diferentes tipos de medidas de distância e centralidade: (a) QMEP , (b) Porcentagem , (c) Amplitude, (d) PA*

percebe-se que o cenário que obteve o melhor resultado foi com o uso da medida de centralidade PID Medoide com a medida de distância de Minkowski, enquanto as medidas de centralidade PED e PED Medoide apresentaram um desempenho inferior.

Como a medida de centralidade PID Medoide com a medida de distância Minkowski apresentaram valores satisfatórios de QMEP e *Porcentagem*, além de intervalos de confiança mais precisos, estas serão as medidas adotadas para a aplicação do método NNPM.

Assim sendo, seguem na Tabela 7.3 os intervalos de confiança de 95% para os efeitos aleatórios correspondentes às cidades da Base de Predição, construídos tal qual foi visto na seção 5.1.

Tabela 7.3: *Intervalos de confiança para os efeitos aleatórios referentes à Base de Predição*

Efeitos	**Intervalos de confiança**	
	Limite inferior	**Limite superior**
Brasília	-0,040	0,246
Goiânia	-0,276	0,066
Joinville	-0,159	0,237
Manaus	-0,359	0,038
São Luís	-0,231	0,185

Da observação da Tabela 7.3, pode-se concluir que todos os intervalos de confiança para os efeitos aleatórios referentes à Base de Predição contém o valor zero, isto poderia ser de

alguma maneira esperado dado que o desvio padrão do efeito aleatório, visto na Tabela 7.2, foi relativamente pequeno (0,2585).

7.2 Aplicação na área de administração

Foi visto no exemplo da seção 2.3 que se pode ter o interesse de averiguar os fatores de uma cidade que podem influenciar na concentração de lojas e de marcas de redes de comida rápida. De fato, Robles (2015) realizou um estudo similar e que foi ampliado em Giampaoli *et al.* (2016a).

Giampaoli *et al.* (2016a) concentraram-se apenas em empresas localizadas no Brasil, dividindo-as em: de origem nacional ou estrangeira. Foram consideradas várias variáveis sociais, demográficas, de governança e de desenvolvimento com o intuito explicar a aglomeração de lojas nas cidades.

A pesquisa envolveu 103 marcas de redes de fast food estrangeiras e nacionais que atuavam no Brasil, no ano de 2015, distribuídas em 542 cidades, sendo que, estes dados foram fornecidos pela Associação Brasileira de Franchising (ABF) e coletados por Giampaoli *et al.* (2016a).

Nesta dissertação, serão consideradas apenas as lojas nacionais e cidades com mais de três franquias de comida rápida, além das seguintes variáveis:

- Total de lojas: é a variável resposta, correspondente à quantidade de lojas, por franquia, presente na cidade.
- IDHM - Renda: indicador da renda municipal per capita, isto é, a soma da renda de todos os residentes, dividida pelo número de pessoas que moram no município - inclusive crianças e pessoas sem registro de renda.
- Taxa de trabalhadores beneficiados: porcentagem de trabalhadores beneficiados pelo Programa de Alimentação do Trabalhador (PAT) - serviços próprios de alimentação, distribuição de alimentos, inclusive não preparados (cesta básica), cartões eletrônicos que permitam a aquisição de refeições ou produtos de gênero alimentício em estabelecimentos comerciais - em relação à população (estimativa) de 2015.
- IBEU: trata-se de um índice de bem-estar urbano.
- População: estimativa da população em 2015.
- Cidade: variável em que será considerado o agrupamento, ou seja, o efeito aleatório. Ela representa a cidade em que cada franquia está situada,

  - Aparecida de Goiânia

- Belém
- Belo Horizonte
- Betim
- Brasília
- Campinas
- Canoas
- Carapicuíba
- Cariacica
- Cascavel
- Caucaia
- Contagem
- Curitiba
- Diadema
- Duque de Caxias
- Florianópolis
- Fortaleza
- Goiânia
- Guarulhos
- Manaus
- Mauá
- Mogi das Cruzes
- Niterói
- Nova Iguaçu
- Osasco
- Porto Alegre
- Recife
- Rio de Janeiro
- Salvador
- Santo André
- Serra
- Suzano

- Vila Velha

- Vitória.

As cidades que serão utilizadas como Base de Ajuste são: Aparecida de Goiânia (4 franquias), Betim (8 franquias), Brasília (33 franquias), Campinas (31 franquias), Canoas (8 franquias), Carapicuíba (4 franquias), Cariacica (7 franquias), Cascavel (4 franquias), Caucaia (4 franquias), Contagem (12 franquias), Diadema (9 franquias), Duque de Caxias (5 franquias), Florianópolis (15 franquias), Fortaleza (30 franquias), Manaus (15 franquias), Mauá (9 franquias), Mogi das Cruzes (13 franquias), Niterói (13 franquias), Osasco (15 franquias), Porto Alegre (12 franquias), Rio de Janeiro (29 franquias), Santo André (21 franquias), Suzano (5 franquias) e Vila Velha (15 franquias). Para a Base de Predição as seguintes cidades foram selecionadas: Belém (11 franquias), Belo Horizonte (21 franquias), Curitiba (24 franquias), Goiânia (18 franquias), Guarulhos (17 franquias), Nova Iguaçu (5 franquias), Recife (29 franquias), Salvador (22 franquias), Serra (8 franquias) e Vitória (9 franquias).

Na Figura 7.4, tem-se um gráfico de boxplots com a distribuição do número total de lojas por franquia, dividido nas cidades da Base de Predição.

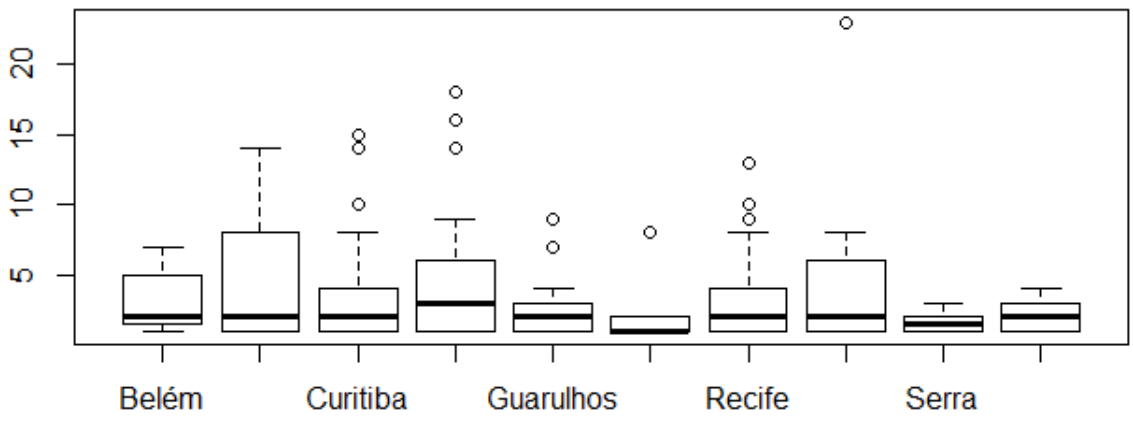

Figura 7.4: *Boxplots da distribuição do número de lojas para cada cidade da BP.*

O modelo ajustado na BA é similar ao utilizado para o número de casos de hipoglicemia em 7.1, sendo que: μ_{ij} é o valor esperado do número de lojas para a j-ésima rede de comida rápida da i-ésima cidade condicionado ao intercepto aleatório da i-ésima cidade, α_i ; $\mathbf{x}_{ij}$ o vetor de covariáveis conhecidas da j-ésima rede de comida rápida da i-ésima cidade; β_0 é o intercepto fixo e $\boldsymbol{\beta}$ são os parâmetros desconhecidos associados ao vetor $\mathbf{x}_{ij}$.

A fim de evitar problemas de escalas nas variáveis, foi tomado o logaritmo das variáveis Taxa de trabalhadores beneficiados e População.

O diagnóstico e as estimativas dos parâmetros do ajuste serão obtidas na próxima seção.

7.2.1 Estimação

Abaixo, seguem os parâmetros associados às variáveis da parte fixa:

- β_1: IDHM - Renda
- β_2: log(Taxa de trabalhadores beneficiados)
- β_3: IBEU
- β_4: log(População)

  As estimativas dos parâmetros, erros padrão e valores p encontram-se na Tabela 7.4.

Tabela 7.4: *Estimativas dos parâmetros, erro padrão e valor p para a Base de Ajuste*

Parâmetros	Estimativa	Erro Padrão	Valor P
β_0	-4,475	1,151	<0,001
β_1	5,355	1,660	0,001
β_2	0,325	0,073	<0,001
β_3	-3,971	1,647	0,016
β_4	0,130	0,051	0,010

Acerca das estimativas, serão ressaltados alguns resultados:

- Adotando um nível de significância de 5%, nota-se que as variáveis IDHM - Renda (β_1), Taxa de trabalhadores beneficiados (β_2) e População (β_4) são significativas para o modelo. Além disso, a partir do sinal positivo das estimativas, espera-se que com o aumento dos valores observados destas variáveis, também haja um crescimento na quantidade de lojas por franquia.
- Para a variável IBEU (β_3), obteve-se um valor p baixo ($1,6\%$), indicando que o índice interfere na variável resposta. No entanto, ao contrário das demais variáveis, o sinal negativo da estimativa mostra uma relação inversa com o número total de lojas por franquia, isto é, quanto maior o IBEU de uma cidade, menor é o número esperado de lojas por franquia. Isto pode estar associado a um número maior de marcas em cidades com maior “qualidade".

Faz-se importante lembrar que, tal qual os resultados obtidos na seção 7.1, a interpretação acima não é conclusiva, por questões já citadas anteriormente.

Na Tabela 7.5, encontram-se as predições dos efeitos aleatórios para as cidades de BA, assim como a estimativa do desvio padrão.

Tabela 7.5: *Predições dos efeitos aleatórios referentes à Base de Ajuste*

Efeitos	**Predições**
Aparecida de Goiânia	-0,020
Betim	0,011
Brasília	0,063
Campinas	0,303
Canoas	-0,170
Carapicuíba	-0,016
Cariacica	-0,059
Cascavel	0,134
Caucaia	0,039
Contagem	-0,013
Diadema	0,006
Duque de Caxias	0,221
Florianópolis	-0,044
Fortaleza	0,030
Manaus	-0,255
Mauá	-0,009
Mogi das Cruzes	-0,196
Niterói	-0,103
Osasco	0,021
Porto Alegre	-0,249
Rio de Janeiro	0,407
Santo André	0,148
Suzano	0,074
Vila Velha	-0,215
Desvio padrão estimado do efeito	0,2253

A cidade com efeito predito mais alto foi o Rio de Janeiro $(0,407)$, assim, das cidades da Base de Ajuste, espera-se que haja maior quantidade de lojas para franquias situadas na cidade do Rio de Janeiro. Já a cidade de Belo Horizonte foi a cidade com predição mais baixa $(-0,376)$, portanto, a cidade em que se espera um menor número de lojas por franquia.

Os gráficos de diagnóstico fornecidos pela Figura 7.5 novamente se mostram satisfatórios. Em (a) nota-se uma disposição aleatória ao redor do zero, em (b) os resíduos situam-se ao longo da reta do QQ-plot normal e em (c) vemos que a suposição de normalidade dos efeitos aleatórios é aceitável.

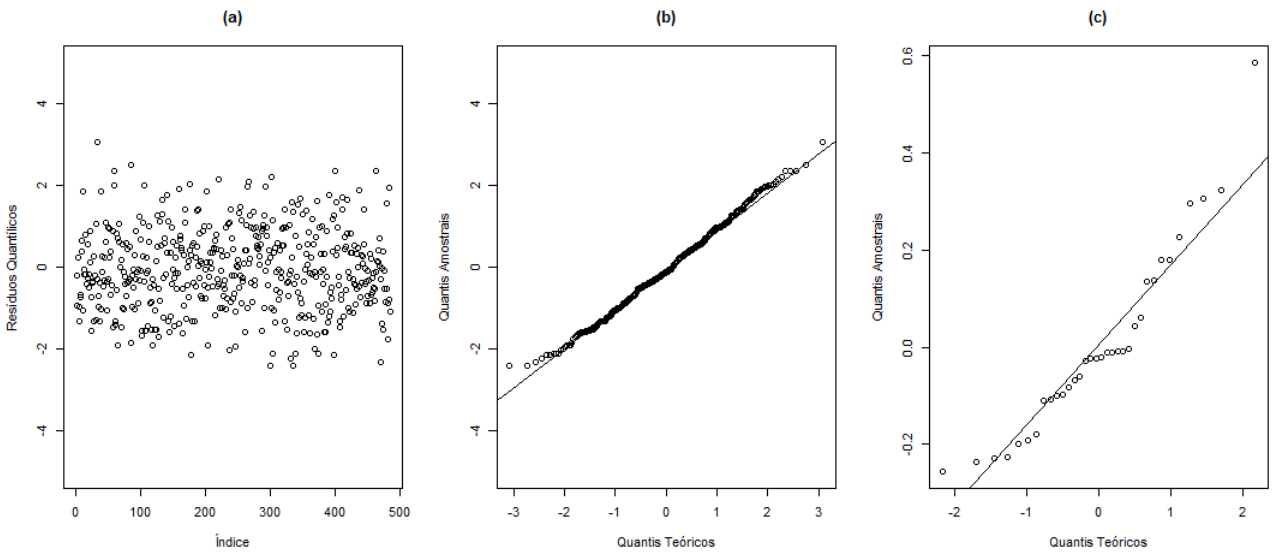

Figura 7.5: *Diagnóstico do modelo: (a) resíduos quantílicos vs índice, (b) QQ-plot dos resíduos quantílicos, (c) QQ-plot do efeito aleatório.*

7.2.2 Predição dos novos grupos

Na Figura 7.6, encontram-se as medidas de desempenho obtidas.

Para o QMEP, tem-se que os melhores resultados (valores mais baixos) foram obtidos para a medida de centralidade Mediana, enquanto a medida de distância de Mahalanobis teve os piores resultados para todas as medidas de centralidade.

Para a variável *Porcentagem*, nota-se que a medida de distância de Mahalanobis comporta-se de maneira similar ao obtido com a variável QMEP, isto é, foi a que teve resultados mais baixos (próximos de 70%) para todas as medidas de centralidade, com exceção da Mediana (valor acima de 80%). Os demais cenários apresentaram resultados predominantemente em torno de 80%, sendo que para o cenário com medida de distância Euclidiana e medida de centralidade Mediana o resultado chegou próximo de 100% mostrando um ótimo desempenho.

Analisando a variável *Amplitude*, tem-se que a maior amplitude média dos intervalos ocorreu para o cenário medida de distância City Block e medida de centralidade PED Medoide, um pouco superior a 2,5, enquanto os intervalos mais precisos deram-se para distância Euclidiana e medida de centralidade Medoide, valor próximo de 2.

Para a variável PA, que leva em conta a precisão dos intervalos e a porcentagem de vezes que o intervalo conteve a média, o melhor valor foi obtido com a medida de distância Euclidiana e medida de centralidade Mediana.

Percebe-se que o desempenho correspondente ao cenário medida de distância Euclidiana e medida de centralidade Mediana apresentou os melhores resultados para as variáveis QMEP

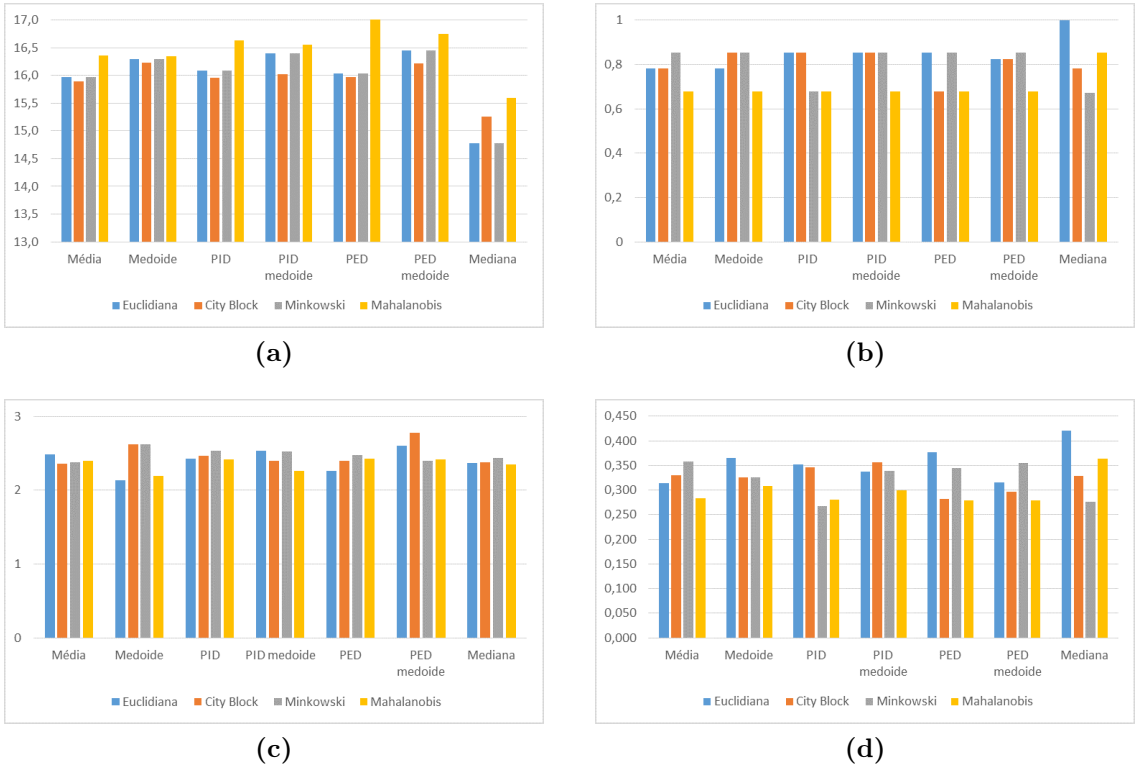

Figura 7.6: *Desempenho preditivo considerando diferentes tipos de medidas de distância e centralidade: (a) QMEP, (b) Porcentagem, (c) Amplitude, (d) PA.*

e *Porcentagem*, além de ter intervalos precisos o bastante para também apresentar o valor mais alto da variável PA. Portanto, este será o cenário adotado para a aplicação do método NNPM.

Seguem na Tabela 7.6 os intervalos de confiança de 95% para os efeitos aleatórios correspondentes às cidades da Base de Predição.

Tabela 7.6: *Intervalos de confiança para os efeitos aleatórios referentes à Base de Predição*

Efeitos	**Intervalos de confiança**	
	Limite inferior	**Limite superior**
Belém	-0,504	-0,027
Belo Horizonte	-0,418	0,013
Curitiba	-0,369	0,104
Goiânia	-0,042	0,741
Guarulhos	-0,504	-0,008
Nova Iguaçu	-0,329	0,160
Recife	-0,418	0,029
Salvador	-0,417	0,012
Serra	-0,167	0,325
Vitória	-0,265	0,082

Diferentemente da aplicação anterior, nota-se na Tabela 7.6 que, apesar do desvio do efeito aleatório ser de apenas 0,2253 (Tabela 7.5), os intervalos de confiança para os efeitos

aleatórios das cidades de Belém e Guarulhos não contém o zero, sinalizando a validez da inclusão do efeito aleatório.

Capítulo 8

Conclusões

8.1 Considerações finais

O objetivo deste trabalho foi estender o método NNPM para o modelo Poisson misto e analisar o comportamento preditivo para novos grupos por meio de aplicações e simulações. Além disso foi proposto o método de Bootstrap para a obtenção dos intervalos de confiança para as quantidades aleatórias preditas pelo método NNPM.

Como alternativa à curva ROC e a estatística KS, medidas de desempenho restritas a respostas binárias, foram propostas três novas medidas de desempenho: *Porcentagem*, *Amplitude* e PA. Apesar de aqui terem sido aplicadas para modelo Poisson misto, essas medidas podem ser utilizadas para qualquer que seja a distribuição da variável resposta considerada, e têm a vantagem de serem de fácil interpretação.

O método de Bootstrap para obtenção dos intervalos de confiança foi aplicado para dois bancos de dados reais, considerando todas as combinações possíveis das medidas de distância e de centralidade definidas no Capítulo 4.

No primeiro, referente ao banco de dados na área de diabetes, embora todas as configurações tenham apresentado excelentes medidas de *Porcentagem*, levando em consideração também a precisão dos intervalos por meio da variável PA, concluímos que a melhor desempenho preditivo ocorreu para a medida de centralidade PID Medoide e medida de distância de Minkowski.

Para o banco de dados na área de administração, encontramos resultados de *Porcentagem* bastante satisfatórios, e a medida de centralidade Mediana e medida de distância Euclidiana destacaram-se como o melhor valor de QMEP, *Porcentagem* e PA, sendo portanto o cenário escolhido para a aplicação do método NNPM.

Como foi observado nestes exemplos, medidas de centralidade e distância diferentes podem não levar aos mesmos resultados, por isso é importante utilizar as que sejam apropriadas

segundo o contexto do problema e realizar as devidas comparações.

Nos estudos de simulação, foram utilizadas as medidas RE e RB definidas em Boubeta *et al.* (2004) para a valoração preditiva. Para o cenário 1, a predição para as Médias foi bastante satisfatória, no entanto, o desempenho não tão bom para a predição dos Efeitos faz com que sejam necessários novos estudos de simulação considerando valores mais altos para σ_a. No cenário 2 a predição comportou-se de forma adequada, e obteve bons resultados mesmo se comparada às predições encontradas em Boubeta *et al.* (2004) através do MPE.

Assim sendo, nota-se que o método NNPM se apresenta como uma técnica muito promissora para modelos mistos com resposta Poisson, com um escopo de aplicações bem variadas, uma vez que é capaz de incorporar possíveis correlações entre os efeitos aleatórios e não necessita da suposição de distribuição de tais efeitos.

8.2 Sugestões para pesquisas futuras

Com o intuito de aprofundar os temas aqui abordados, tem-se abaixo algumas possibilidades de pesquisas futuras:

- Aplicar o método NNPM em modelos com mais de um efeito aleatório.
- Buscar medidas de desempenho alternativas para o modelo Poisson misto, que não necessitem o cálculo do MPE.
- Comparar o desempenho do NNPM para o modelo Poisson misto, com outras técnicas de predição para observações de novos grupos.
- Prosseguir os estudos utilizando outras distribuições para a variável resposta.

Referências Bibliográficas

Boubeta *et al.*(**2004**) Miguel Boubeta, María José Lombardía e Domingo Morales. Empirical best prediction in poisson mixed models application to poverty data. Em *Research Group on Modeling, Optimization and Statistical Inference (MODES)*, Department of Mathematics University of A Coruña, Poznan, Poland. Citado na pág. 10, 11, 32, 34, 56

Conover(1999) W.J Conover. *Pratical Nonparametric Statistics*. Second edição. Citado na pág. 19

Douglas(1980) James B Douglas. Analysis with standard contagious distributions. Citado na pág. 5

Efron e Tibshirani(1994) Bradley Efron e Robert J Tibshirani. *An introduction to the bootstrap*. CRC press. Citado na pág. 9

Fawcett(2006) Tom Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874. Citado na pág. 19

Giampaoli *et al.*(**2013**) Viviana Giampaoli, Karin Ayumi Tamura e Alexandre Noma. Nearest neighbors prediction method for mixed logistic regressions. Em *28th International workshop on Statistical Modelling*, Palermo, Italy. Citado na pág. 2, 4, 19, 21, 22

Giampaoli *et al.*(**2016a**) Viviana Giampaoli, Alberto Nunes Ferraz Junior e Isadora Avidos Cid Castro. Relatório de análise estatística sobre o projeto "aglomeração das redes de fast food no brasil.". Citado na pág. 47

Giampaoli *et al.*(**2016b**) Viviana Giampaoli, Karin A Tamura, Norma P Caro e Luiz J Simoes de Araujo. Prediction of a financial crisis in latin american companies using the mixed logistic regression model. *Chilean Journal of Statistics (ChJS)*, 7(1). Citado na pág. 19

Johnson *et al.*(**2005**) Norman L Johnson, Adrienne W Kemp e Samuel Kotz. *Univariate Discrete Distributions, Set*, volume 444. John Wiley & Sons. Citado na pág. 5

Johnson e Wichern(2007) R.A. Johnson e D.W. Wichern. *Applied Multivariate Statistical Analysis*. Sexta edição. Citado na pág. 22

Law Averill e David(2000) M Law Averill e Kelton W David. Simulation modeling and analysis. *Mc-Graw Hill*. Citado na pág. 10

Liu e Pierce(1994) Qing Liu e Donald A Pierce. A note on gauss—hermite quadrature. *Biometrika*, 81(3):624–629. Citado na pág. 8

McCulloch e Neuhaus(2011) Charles E McCulloch e John M Neuhaus. Prediction of random effects in linear and generalized linear models under model misspecification. *Biometrics*, 67(1):270–279. Citado na pág. 28

McCulloch *et al.***(2008)** Charles E McCulloch, Shayle R Searle e John M Neuhaus. *Generalized, Linear, and Mixed Models*. Wiley. Citado na pág. 8, 16

Nelder e Wedderburn(1972) John A Nelder e Robert WM Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society*, 135:370–384. Citado na pág. 6

Neuhaus *et al.***(2013)** John M Neuhaus, Charles E McCulloch e Ross Boylan. Estimation of covariate effects in generalized linear mixed models with a misspecified distribution of random intercepts and slopes. *Statistics in medicine*, 32(14):2419–2429. Citado na pág. 28

Paula(2013) Gilberto Alvarenga Paula. *Modelos de regressão: com apoio computacional*. IME-USP São Paulo. Citado na pág. 7

Rigby e Stasinopoulos(2005) Robert A Rigby e D Mikis Stasinopoulos. Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(3):507–554. Citado na pág. 44

Rizzo e Cymrot(2006) Ana Lucia Tucci Rizzo e Raquel Cymrot. Estudo e aplicações da técnica bootstrap, 2006. Citado na pág. 10

Robles(2015) F. Robles. Agglomeration and institutional determinants of franchised fast food networks in latin america. Em *9th Iberoamerican Academy of Management Conference. From Developing to Developed Economies: the future of entrepreneurship and innovation in Iberoamerica*, Universidad del Desarrollo, Santiago, Chile. Citado na pág. 47

Sheiner e Beal(1981) Lewis B Sheiner e Stuart L Beal. Some suggestions for measuring predictive performance. *Journal of pharmacokinetics and biopharmaceutics*, 9(4):503–512. Citado na pág. 26

Skrondal e Rabe-Hesketh(2009) Anders Skrondal e Sophia Rabe-Hesketh. Prediction in multilevel generalized linear models. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 172(3):659–687. Citado na pág. 16

Tamura(2012) Karin Ayumi Tamura. *Métodos de predição para modelo logıstico misto com k efeitos aleatórios*. Tese de Doutorado, Universidade de Sao Paulo. Citado na pág. 2, 13, 17

Tamura e Giampaoli(2013) Karin Ayumi Tamura e Viviana Giampaoli. New prediction method for the mixed logistic model applied in a marketing problem. *Computational Statistics & Data Analysis*, 66:202–216. Citado na pág. 17

Wedderburn(1974) Robert WM Wedderburn. Quasi-likelihood functions, generalized linear models, and the gauss—newton method. *Biometrika*, 61(3):439–447. Citado na pág. 7