Modelos de Regressão Bivariados Bernoulli - Exponencial

Flávia Bolssone do Prado

São Carlos

2013

Universidade Federal de São Carlos
Centro de Ciências Exatas e de Tecnologia
Departamento de Estatística

Modelos de Regressão Bivariados Bernoulli - Exponencial

Flávia Bolssone do Prado

Orientador: Carlos A. R. Diniz

Dissertação apresentada ao Programa de Pós-Graduação em Estatística da Universidade Federal de São Carlos - PPGEs/UFSCar, como parte dos requisitos para a obtenção do título de Mestre em Estatística.

UFSCar - São Carlos-SP
Maio de 2013

Ficha catalográfica elaborada pelo DePT da Biblioteca Comunitária da UFSCar

P896mr

Prado, Flávia Bolssone do.
Modelos de regressão bivariados Bernoulli : exponencial / Flávia Bolssone do Prado. -- São Carlos : UFSCar, 2013.
82 f.

Dissertação (Mestrado) -- Universidade Federal de São Carlos, 2013.

1. Análise de regressão. 2. Modelos de regressão. 3. Estimadores de Bayes. 4. Distribuição exponencial. 5. Distribuição de Bernoulli. I. Título.

CDD: 519.536 (20ª)

UNIVERSIDADE FEDERAL DE SÃO CARLOS
Centro de Ciências Exatas e de Tecnologia
Programa de Pós-Graduação em Estatística
Via Washington Luís, Km 235 - C.P.676 - CGC 45358058/0001-40
FONE: (016) 3351-8292 – Email: ppgest@ufscar.br
13565-905 - SÃO CARLOS-SP - BRASIL

FOLHA DE APROVAÇÃO

Aluno(a) : Flávia Bolssone do Prado

DISSERTAÇÃO DE MESTRADO DEFENDIDA E APROVADA EM 05/04/2013
PELA COMISSÃO JULGADORA:

Presidente _____
Prof. Dr. Carlos Alberto Ribeiro Diniz (DEs-UFSCar/Orientador)

1º Examinador _____
Prof. Dr. Luis A. Milan (DEs-UFSCar)

2º Examinador _____
Prof. Dr. Ronaldo Dias (UNICAMP)

Agradecimentos

Agradeço, primeiramente, a Deus pelo dom da vida e por todas as oportunidades concedidas e alcançadas.

Aos meus pais, Valentim e Noraide, pelo incentivo e por sempre estarem dispostos quando precisei.

Aos colegas de mestrado e doutorado pela contribuição e paciência. Em especial a Daniela, Rosineide, Carolina e Cíntia, pela amizade, companheirismo e pelos momentos felizes que me proporcionaram.

Aos professores do Departamento de Estatística da Universidade Federal de São Carlos pela amizade e por todo conhecimento transmitido.

Em especial ao professor Dr. Carlos Alberto Ribeiro Diniz pela orientação e pelos conhecimentos compartilhados.

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pelo auxílio financeiro concedido desde o início deste trabalho.

Finalmente, agradeço a todos que de alguma maneira contribuíram para a conclusão deste trabalho.

Resumo

Neste trabalho desenvolvemos modelos de regressão para respostas bivariadas, discreta e contínua, com a variável discreta seguindo distribuição Bernoulli e a variável contínua, condicionada na discreta, seguindo distribuição exponencial. Um procedimento de ajuste, via abordagem Bayesiana, é utilizado para estimar os parâmetros do modelo e uma análise de resíduos Bayesianos é apresentada. Um estudo de simulação é descrito a fim de ilustrar a metodologia desenvolvida. Utilizamos três tamanhos amostrais diferentes para analisarmos os resultados. Aplicamos o modelo em um conjunto de dados reais relacionado a gastos com pacientes internados em hospitais, levando em consideração a utilização, ou não, de tratamento cirúrgico. A covariável disponível para a análise foi o número de dias de permanência do paciente hospitalizado.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Em várias áreas, tais como psicologia, medicina e finanças, é comum o uso de modelos bivariados que contenham ambas variáveis resposta discretas, ambas contínuas ou uma variável resposta discreta e outra contínua.

Modelos com as duas variáveis resposta discretas são apresentados em Jung & Winkelmann (1993), Van Ophem (1999) e Khafri *et al.* (2008), entre outros.

Jung & Winkelmann (1993) utilizaram um modelo Poisson bivariado motivado por mobilidade de emprego, levando em consideração dois aspectos: mudanças diretas de emprego para emprego e mudança de emprego depois de um período de desemprego, analisando dados de indivíduos do sexo masculino no mercado de trabalho alemão. Neste conceito de mobilidade de emprego a modelagem poderia ser feita separadamente para os dois aspectos, porém estes podem ser estreitamente relacionados, o que representa um risco nos históricos de trabalho individual. Esta nova modelagem, baseada na distribuição Poisson bivariada para os dois aspectos, permite estimativas simultâneas da estrutura de correlação entre as variáveis dependentes e os coeficientes de regressão.

Van Ophem (1999) apresentou um método de estimação para variáveis aleatórias discretas correlacionadas. Como exemplo de motivação utilizou dados de contagem

de uma análise que lida com o comportamento turístico e recreativo de indivíduos em viagens. Como variáveis resposta considerou o número de vezes em que um indivíduo foi passear e o número de atrações turísticas visitadas. Ambas as variáveis resposta foram consideradas seguindo distribuição Poisson e a possível correlação entre as mesmas é devido ambas serem obtidas de um mesmo indivíduo.

Khafri *et al.* (2008) apresentaram uma análise Bayesiana para um modelo de regressão Poisson bivariado, com dados de contagem, focados em modelos nos quais as observações não são correlacionadas entre os indivíduos mas são correlacionadas entre as respostas. Como exemplo de motivação consideraram tratamentos clínicos de casais inférteis levando em conta a fertilização *in vitro*, tomando como variáveis resposta o número de embriões obtidos e o número de óvulos maduros. O objetivo é encontrar o efeito de diferentes fatores clínicos e demográficos nas duas respostas, simultaneamente. A correlação ocorre pelo fato das duas respostas serem relacionadas a um mesmo casal.

Para modelos com ambas as variáveis resposta contínuas apresentamos, como exemplo, os trabalhos de Scollnik (2002) e Song *et al.* (2004).

Scollnik (2002) analisou dois modelos de regressão possíveis para dados bivariados de sinistros. A análise foi feita em relação a seguros contra acidentes, considerando valores de perda e valores de despesas alocadas com ajuste do sinistro. No primeiro modelo a variável associada ao valor da perda tem distribuição Pareto e a variável associada às despesas de ajuste, condicionada ao valor da perda, tem distribuição Gama. No segundo modelo ambas as variáveis, valor da perda e despesas de ajuste, condicionada ao valor da perda, têm distribuição Pareto. A análise conjunta para as duas variáveis do modelo é feita devido ao fato de que em seguros de acidentes as despesas alocadas com ajustes de sinistros estão diretamente relacionados ao pagamento da própria perda.

Song *et al.* (2004) apresentaram um estudo, com o objetivo de estimar a correlação entre as variáveis resposta de uma pesquisa sobre HIV, no qual a correlação é estimada através de equações de estimação generalizadas. Os dados disponíveis

na pesquisa, obtidos de mulheres grávidas, foram célula de contagem *CD4+*, útil pra detectar HIV positivo e saber como o organismo está reagindo ao vírus, e carga viral de HIV, usado como uma medida para verificar a gravidade de uma infecção viral. As duas variáveis resposta são assumidas normalmente distribuídas e a variável relacionada a carga viral de HIV é condicionada a variável relacionada a célula de contagem *CD4+*. O teste para carga viral é uma medida que fornece importante informação que é utilizada juntamente com célula de contagem *CD4+*, deste modo faz-se necessário o interesse na correlação de ambas as respostas.

Outra maneira de apresentar dados bivariados é com respostas mistas em que uma variável é discreta e a outra variável é contínua. O modelo mais simples que pode ser construído utilizando variáveis mistas é com uma variável discreta assumindo valores 0 ou 1 e a outra variável contínua. Nestes casos podemos utilizar o modelo de fatoração no qual a distribuição conjunta das respostas pode ser formulada de duas maneiras: primeiro com uma distribuição marginal para a variável discreta e uma distribuição condicional para a variável contínua, dado a resposta discreta ou considerando uma distribuição marginal para a variável contínua e uma distribuição condicional para a variável discreta, dado a resposta contínua.

Pesquisas relacionadas a análise conjunta de respostas mistas têm sido apresentadas apenas recentemente devido às dificuldades na especificação das distribuições conjuntas para o vetor de respostas e pela falta de modelos padrões. O nosso trabalho é desenvolvido considerando respostas bivariadas mistas, em que estão disponíveis a distribuição marginal da variável discreta e a distribuição condicional da variável contínua dado a variável discreta.

Um dos primeiros trabalhos na linha de respostas mistas é o de Olkin & Tate (1961) que, motivados por estudos na psicologia, desenvolveram um modelo, conhecido como modelo de localização, no qual a variável discreta é usada para indicar presença ou ausência de um atributo e segue distribuição binomial e a variável contínua, condicionada na variável discreta, segue distribuição normal. Consideraram, ainda, uma extensão multivariada em que o vetor resposta da variável discreta segue

distribuição multinomial e o vetor resposta da variável contínua segue distribuição normal multivariada. Como exemplo de motivação avaliaram características de membros de partidos políticos britânicos, tomando como variável discreta a origem dos indivíduos (nortistas ou sulistas) e como variável contínua escalas de pontuação atribuídas aos indivíduos em relação a dogmatismo e ao ato de tentar impor sua opinião. Como ambas as respostas, discreta e continua, são obtidas de um mesmo indivíduo existe a necessidade de analisá-las conjuntamente.

Em tais modelos o interesse pode estar em encontrar uma medida de associação entre a variável discreta, que assume valores 0 ou 1, e a variável contínua. Tate (1954) usa, para esta finalidade, o coeficiente de correlação produto-momento, que tem recebido o nome de coeficiente de correlação ponto-bisserial pela sua relação com o coeficiente de correlação bisserial de Karl Pearson. Neste trabalho também é descrito, com mais detalhes, o modelo de respostas mistas com distribuição binomial e distribuição normal, visto por Olkin & Tate (1961).

Como uma extensão do modelo de localização de Olkin & Tate (1961), Little & Schluchter (1985), motivados por pesquisas de risco, apresentaram procedimentos de máxima verossimilhança para análise de dados mistos contínuos e categóricos, com valores faltantes, baseados em um modelo normal multivariado para a variável contínua e um modelo poisson/multinomial para dados categóricos. Como exemplo de motivação foram analisados efeitos de distúrbios psicológicos dos pais em vários aspectos do desenvolvimento dos filhos. Analisaram a idade escolar das crianças de famílias classificadas pelo grupo de risco em que os pais pertencem. Tomaram como variáveis resposta o nível de pontuação em relação à leitura padronizada e à compreensão verbal padronizada. A análise dos dados feita de maneira conjunta ocorreu devido às repostas serem obtidas de uma mesma criança.

Em contraste com o modelo inicial de Olkin & Tate (1961), Cox (1972) descreveu um modelo no qual a distribuição da resposta contínua é normal e a variável discreta é condicionada à variável contínua e sua distribuição condicional é Bernoulli. Do mesmo modo, Catalano & Ryan (1992) usaram um conceito de variável latente para

obter a distribuição conjunta de um modelo misto com uma resposta contínua e uma resposta discreta. O modelo pode ser parametrizado de forma que permite escrever a distribuição conjunta como o produto da distribuição marginal da resposta contínua e da distribuição condicional da resposta binária, dada a resposta contínua. No modelo de Catalano & Ryan (1992) a variável discreta é associada a uma variável latente não observada e a variável contínua e a variável latente seguem distribuição normal. As estimativas dos parâmetros de ambos os modelos não são consistentes quando o modelo para a associação entre a resposta discreta e a resposta contínua for mal especificado. Como exemplo de motivação utilizaram um estudo de toxicidade em ratos tendo como resposta contínua o peso fetal e como resposta discreta a má formação do feto, condicionada ao peso fetal. A variável má formação tem um variável latente correspondente, não observada, em que o peso fetal e esta variável latente compartilham uma distribuição normal conjunta. A associação entre as variáveis resposta ocorre devido as características terem sido tomadas de um mesmo rato.

Ainda como uma extensão do modelo de localização de Olkin & Tate (1961), Fitzmaurice & Laird (1995) propuseram um modelo para respostas bivariadas discreta e contínua no qual a resposta discreta tem distribuição Bernoulli e a resposta contínua, condicionada na discreta, tem distribuição normal. Este modelo apresenta vantagens aos anteriores, uma é por seus parâmetros terem interpretações marginais e outra é que as estimativas de máxima verossimilhança dos parâmetros são consistentes mesmo se a associação entre as respostas discreta e contínua tenha sido mal especificada, o que era um problema nos modelos de Cox (1972) e Catalano & Ryan (1992). Esta dificuldade foi contornada simplesmente invertendo a escolha da variável condicionada.

O trabalho de Fitzmaurice & Laird (1995) é motivado por estudos de toxicidade no qual são realizados ensaios clínicos em ratas grávidas, avaliando em cada feto evidência de má formação e peso fetal. Devido às duas características avaliadas serem de um mesmo indivíduo (feto), as variáveis resposta discreta (existência de

má formação) e contínua (peso fetal) devem ser avaliadas conjuntamente.

Outros trabalhos na linha de variáveis mistas, discreta e contínua, que podemos citar são: Lauritzen & Wermuth (1989), Little & Rubin (1987), Laird (1995), Cox & Wermuth (1992), Zhao *et al.* (1992).

Nesta dissertação propomos modelos que são desenvolvidos seguindo o modelo de Fitzmaurice & Laird (1995). Trabalhamos com modelos bivariados nos quais a variável discreta segue distribuição Bernoulli e a variável contínua, condicionada na discreta, segue distribuição exponencial. O primeiro modelo é construído a partir de uma regressão linear para a média da distribuição condicional, com função de ligação logarítmica; o segundo modelo é construído obtendo a média da distribuição contínua condicional em função da variável discreta e das esperanças marginais e o terceiro modelo é construído seguindo o segundo modelo com a média condicional dependendo apenas da variável discreta e da esperança marginal da variável contínua. Os dois últimos modelos utilizam as funções de ligação logito e logarítmica para relacionar os parâmetros de interesse, de ambas regressões, às covariáveis disponíveis.

Nossos modelos se diferenciam dos já existentes pela utilização da distribuição normal para a variável contínua e pela aplicação de uma abordagem Bayesiana no processo de estimação.

Os parâmetros de interesse são os parâmetros relacionados à esperança marginal das respostas, enquanto que o parâmetro de associação entre as variáveis discreta e contínua é tratado como parâmetro perturbador. No exemplo de motivação as variáveis repostas são obtidas de um mesmo indivíduo e em iguais circunstâncias. Isto sugere a presença de correlação entre as duas respostas.

Uma abordagem Bayesiana, via Markov Chain Monte Carlo (MCMC), é utilizada para estimar os parâmetros dos modelos e uma análise de resíduos Bayesianos é descrita através de três tipos de resíduos, resíduo baseado na densidade preditiva condicional ordinária, resíduo baseado na distribuição a posteriori dos parâmetros do modelo e resíduo deviance Bayesiano.

Este trabalho está organizado em cinco capítulos.

No capítulo 2 encontramos as descrições dos modelos bivariados propostos anteriormente por outros autores.

No capítulo 3 apresentamos a construção dos modelos bivariados Bernoulli-Exponencial com a distribuição marginal sendo Bernoulli e a distribuição condicional sendo exponencial. Apresentamos, também, a metodologia desenvolvida para a estimação dos parâmetros e a descrição da análise de resíduos Bayesianos utilizada.

No capítulo 4 apresentamos um estudo de simulação, com objetivo de ilustrar a metodologia, uma análise de resíduos Bayesianos e aplicamos a metodologia em um conjunto de dados reais relacionado com gastos de planos de saúde com pacientes hospitalizados, levando em consideração a utilização, ou não, de tratamento cirúrgico.

Finalmente, no capítulo 5 apresentamos as conclusões finais do nosso trabalho.

Capítulo 2

Modelos Bivariados

Neste capítulo apresentamos os modelos de distribuições bivariadas descrito por vários outros autores nos trabalhos mencionados no capítulo anterior. Descrevemos um modelo para ambas as variáveis discretas, um modelo para ambas as variáveis contínuas e os principais modelos para respostas mistas com uma variável discreta e uma variável contínua.

2.1 Modelo bivariado com ambas variáveis discretas

Nesta seção apresentamos o trabalho de Khafri *et al.* (2008), que apresenta um modelo de Poisson bivariado.

2.1.1 Modelo de regressão de Poisson bivariado

Khafri *et al.* (2008) tomaram como observações y_{ij}, realizações da variável aleatória Y_{ij} com $i = 1, \ldots, n_j$, $j = 1, 2$, para o indivíduo i e a resposta j, sendo $\mathbf{Y}_i = (Y_{i1}, Y_{i2})$ o vetor de contagem para o indivíduo i em relação às duas respostas.

Neste trabalho o interesse está em modelos nos quais as observações são não correlacionadas entre os indivíduos mas são correlacionadas entre as respostas.

Pode-se assumir que Y_{ij} seja distribuído como uma variável aleatória de Poisson com parâmetro λ_{ij}, ou seja,

$$(Y_{ij}|Z_{ij}, b_{ij}) \sim \text{Poisson}(\lambda_{ij})$$

em que λ_{ij} é especificado como

$$\lambda_{ij} = exp(z'_{ij}\boldsymbol{\beta}_j + b_{ij}),$$

com $i = 1, 2, \ldots, n_j$, $j = 1, 2$, z_{ij} representando a variável explicativa, $\boldsymbol{\beta}_j$ um vetor $k \times 1$ de parâmetros e b_{i1} e b_{i2} representando os componentes que modelam a dependência entre Y_{i1} e Y_{i2}. Consideramos $\mathbf{b}_i = (b_{i1}, b_{i2})$, sendo que $\mathbf{b}_i \sim N_2(\mathbf{0}, \boldsymbol{\Sigma})$, em que $\boldsymbol{\Sigma}$ é uma estrutura de variância e covariância para acomodar a correlação entre b_{i1} e b_{i2}.

A função densidade de probabilidade para o vetor de contagem $\mathbf{Y}_i = (Y_{i1}, Y_{i2})$ é dada por

$$f(y_{i1}, y_{i2}|z_i, \boldsymbol{\beta}_j, \boldsymbol{\Sigma}) = \int \prod_{j=1}^{2} f(y_{ij}|\beta_j, b_{ij})\phi_j(\mathbf{b}_i|\mathbf{0}, \boldsymbol{\Sigma})d\mathbf{b}_i. \tag{2.1}$$

A integral em (2.1) não pode ser resolvida de uma forma analítica. Um método MCMC, em um contexto Bayesiano, pode ser utilizado.

Khafri *et al.* (2008) utilizaram uma análise Bayesiana hierárquica para o modelo, que foi escrita em dois níveis. As distribuições a priori para os parâmetros do modelo foram, para o primeiro nível de hierarquia, $\boldsymbol{\beta} \sim N_k(\mu_\beta, V_\beta^{-1})$ e $\boldsymbol{\Sigma}^{-1} \sim Wishart(\mu_\Sigma, V_\Sigma)$ e, para o segundo nível de hierarquia: $\boldsymbol{\mu}_\beta \sim N_k(\beta_0, \sigma_0)$, $V_\beta^{-1} \sim Wishart(\mu_{0_\beta}, V_{0_\beta})$ e $V_\Sigma^{-1} \sim Wishart(\mu_{0_\Sigma}, V_{0_\Sigma})$, com β_0, σ_0, μ_{0_β}, μ_{0_Σ}, V_{0_β} e V_{0_Σ} conhecidos, supondo que β_0 e σ_0 são independentes. $\boldsymbol{\lambda}_i = (\lambda_{i1}, \lambda_{i2})$ tem uma distribuição lognormal bivariada e $\mathbf{Y}_i = (Y_{i1}, Y_{i2})$ tem uma distribuição poisson-lognormal bivariada.

Dadas as observações, a distribuição a posteriori conjunta dos parâmetros do modelo é obtida combinando a função de verossimilhança e as distribuições a priori, via teorema de Bayes, ou seja

$$
\begin{aligned}
&\pi(\boldsymbol{\beta}_j, \mu_\beta, V_\beta, \mathbf{b}_i, \boldsymbol{\Sigma}|y_i, X) \\
&\propto L(\boldsymbol{\beta}_j, \mathbf{b}_i|y_i) P(\mathbf{b}_i|\boldsymbol{\Sigma}) P(\boldsymbol{\Sigma}^{-1}|\mu_\Sigma, V_\Sigma) g(V_\Sigma^{-1}) P(\boldsymbol{\beta}_j|\mu_\beta, V_\beta) g(\mu_\beta) g(V_\beta^{-1}) \\
&\propto \prod_{1=1}^{n} \prod_{j=1}^{2} exp(-exp(z_{ij}^{'}\boldsymbol{\beta}_j + b_{ij}))(exp(z_{ij}^{'}\boldsymbol{\beta}_j + b_{ij}))^{y_{ij}} \\
&\times \prod_{1=1}^{n} exp(-1/2 b_i^{\prime}\Sigma^{-1} b_i)|\Sigma^{-1}|^{1/2}|\sigma_0|^{-1/2} \\
&\times \frac{|\Sigma^{-1}|^{\frac{\mu_\Sigma - 2}{2}} exp(-1/2 tr(V_\Sigma^{-1}\Sigma^{-1}))}{|V_\Sigma|^{\frac{\mu_\Sigma}{2}}} \\
&\times \frac{|V_\Sigma^{-1}|^{\frac{\mu_{0_\Sigma} - 2}{2}} exp(-1/2 tr(V_\Sigma^{-1} V_{0_\Sigma}^{-1}))}{|V_{0_\Sigma}|^{\frac{\mu_{0_\Sigma}}{2}}} \\
&\times \prod_{j=1}^{2} |V_\beta|^{-1/2} exp(-1/2(\boldsymbol{\beta}_j - \boldsymbol{\mu}_\beta)^{'} V_\beta^{-1}(\boldsymbol{\beta}_j - \boldsymbol{\mu}_\beta)) \\
&\times exp(-1/2(\mu_\beta - \beta_0)^{'}\sigma_0^{-1}(\mu_\beta - \beta_0)) \\
&\times |V_\beta^{-1}|^{\frac{\mu_{0_\beta} - 2}{2}} . \frac{exp(-1/2 tr(V_{0_\beta}^{-1} V_\beta^{-1}))}{|V_{0_\beta}|^{\frac{\mu_{0_\beta}}{2}}}
\end{aligned}
$$

Como a distribuição a posteriori é analiticamente intratável foi utilizado, para a estimação dos parâmetros do modelo, um método de Gibbs Sampler para modelos hierárquicos.

Como exemplo de motivação, foram observados 268 casais inférteis que procuraram um centro cirúrgico de Teerã, no Irã, no período de julho de 2006 e março de 2007. Levaram em consideração a fertilização *in vitro*, tomando como variáveis resposta o número de embriões obtidos e o número de óvulos maduros.

Através dos resultados presentes no artigo foi observado que todas as distribuições a posteriores são simétricas em relação a suas médias, que as estimativas dos parâmetros sugerem que as informações femininas desempenham um papel importante na previsão do número de óvulos maduros e embriões obtidos e mostrou a falta de influência dos parâmetros do sexo masculino no número de embriões obtidos.

O modelo proposto não apresentou diferenças significativas nos resultados das técnicas de reprodução assistida em dois estágios de fertilização *in vitro*, realizados em diferentes épocas do ano. Com base na descrição dos efeitos de correlação, é

esperado que a especificação do modelo de regressão lognormal Poisson bivariado renda um modelo de predição superior, pelo fato do número de óvulos maduros e de embriões obtidos serem considerados correlacionados em um mesmo caso.

2.2 Modelo bivariado com ambas variáveis contínuas

Para ilustrar um modelo bivariado com ambas as variáveis respostas contínuas utilizamos o trabalho de Scollnik (2002).

2.2.1 Modelos de regressão para dados de perda bivariados

Scollnik (2002) apresentou, neste trabalho, dois modelos para ajuste de dados de perda, um modelo Pareto-gama e um modelo Pareto-Pareto, descritos abaixo.

A análise foi feita em relação a seguros contra acidentes, considerando valores de perda para a variável X e valores de despesas alocadas com ajuste do sinistro para a variável Y.

Considerado Ψ o vetor completo de parâmetros do modelo, uma abordagem clássica de estimação paramétrica foi utilizada para obtenção de $\widehat{\Psi}$.

Modelo de regressão Pareto-gama

Neste modelo é assumido uma distribuição Pareto para a variável X, com parâmetros α e θ, e uma distribuição condicional gama para a variável Y, com parâmetros δ e λ_x, com $\lambda_x = exp(\gamma + \beta log x)$.

A função densidade conjunta para este modelo é dada por

$$f(x, y|\Psi) = \frac{\alpha\theta^\alpha}{(x+\theta)^{\alpha+1}} \frac{\lambda_x^\delta y^{\delta-1}}{\Gamma(\delta) exp(y\lambda_x)},$$

com $\Psi = (\alpha, \theta, \delta, \gamma, \beta)$, em que α, θ e δ devem ser positivos e γ e β podem assumir valores positivos ou negativos.

É possível observar que a esperança condicional $E(Y|X = x, \Psi)$, neste caso, sempre existe e tem valor $\frac{\delta}{\lambda_x}$.

Ainda para este modelo, foi obtida uma versão centralizada da função de densidade, tomando para a variável Y uma distribuição condicional gama com parâmetros δ e $\widetilde{\lambda}_x$, com $\widetilde{\lambda}_x = exp(\gamma + \beta[logx - k])$, em que k é igual ao valor médio observado de $log(x)$. Assim, a função densidade conjunta para este novo modelo é dada por

$$f(x, y|\Psi) = \frac{\alpha\theta^\alpha}{(x+\theta)^{\alpha+1}} \frac{\widetilde{\lambda}_x^\delta y^{\delta-1}}{\Gamma(\delta) exp(y\widetilde{\lambda}_x)}$$

Observa-se que as estimativas de máxima verossimilhança podem ser determinadas utilizando métodos convencionais. Poderia também utilizar o método de Newton-Raphson, que foi mostrado, em experiências, ser mais eficiente em razão de acelerar a convergência da simulação.

Modelos de regressão Pareto-Pareto

Neste outro modelo a variável X é assumida seguir distribuição Pareto com parâmetros α e θ e a variável Y é assumida seguir distribuição Pareto, condicionada na variável X, com parâmetros δ e λ_x, com $\lambda_x = exp(\gamma + \beta logx)$.

Assim, a função densidade conjunta para este modelo é dada por

$$f(x, y|\Psi) = \frac{\alpha\theta^\alpha}{(x+\theta)^{\alpha+1}} \frac{\delta\lambda_x^\delta}{(y+\lambda_x)^{\delta+1}},$$

com $\Psi = (\alpha, \theta, \delta, \gamma, \beta)$, em que α, θ e δ devem ser positivos e γ e β podem assumir valores positivos ou negativos.

Neste caso a esperança condicional $E(Y|X = x, \Psi)$ só existe quando $\delta > 1$ e tem valor $\frac{\lambda_x}{\delta - 1}$.

Com base neste modelo, foi obtida uma versão centralizada da função de densidade, tomando a distribuição condicional de Y como Pareto com parâmetros δ e $\widetilde{\lambda}_x$, com $\widetilde{\lambda}_x = exp(\gamma + \beta[logx - k])$, em que k é igual ao valor médio observado de $log(x)$, como anteriormente. A função densidade para este novo modelo é dada por

$$f(x, y|\Psi) = \frac{\alpha\theta^{\alpha}}{(x+\theta)^{\alpha+1}} \frac{\delta\widetilde{\lambda}_x^{\delta}}{(y+\widetilde{\lambda}_x)^{\delta+1}}.$$

As estimativas de máxima verossimilhança podem ser determinadas utilizando métodos padrões.

Para ambos os modelos o valor da estimativa de máxima verossimilhança, $\hat{\alpha}$, sugere que os dois primeiros momentos de X existem. Em relação a um teste de hipótese, há uma implicação de que as inferências preditivas condicionadas ao valor de $\hat{\alpha}$ podem estar sujeitas a uma variabilidade significativamente menor. Portanto, a perda total prevista pode ser radicalmente subestimada. Por esta razão, uma análise Bayesiana é adequada para explicar a incerteza inerente do parâmetro.

Uma abordagem Bayesiana, baseada em MCMC, implementada com a ajuda de WinBUGS, foi utilizada no processo de estimação e resultou em distribuições preditivas significativamente mais dispersas porque incorpora a incerteza do parâmetro que é efetivamente ignorada pela análise de máxima verossimilhança.

Os resultados apresentados por Scollnik (2002) mostraram que a distribuição preditiva Bayesiana para a variável perda, em cada caso de cobertura, é mais dispersa do que a distribuição preditiva correspondente das estimativas de máxima verossimilhança.

2.3 Modelo bivariado misto

Nesta secção apresentamos os principais modelos de trabalhos desenvolvidos para vetor bivariado de respostas mistas, com uma variável resposta seguindo distribuição discreta e outra seguindo distribuição contínua.

2.3.1 Modelo de Olkin & Tate (1961)

No modelo inicial, proposto por Tate (1954), a variável X tem distribuição binomial e a variável Y, condicionada à variável X, tem distribuição normal.

Os vetores independentes (X_α, Y_α), $\alpha = 1, \ldots, n$ são tais que

$$X \sim b(1,p),$$
$$(Y|x=1) \sim N(\mu_1,\sigma^2) \text{ e}$$
$$(Y|x=0) \sim N(\mu_0,\sigma^2).$$

Considerando $\Delta = \dfrac{(\mu_1-\mu_0)}{\sigma}$, a correlação entre as respostas é dada por

$$\rho_{X,Y} = \Delta\left[\frac{pq}{(1+pq\Delta^2)}\right]^{1/2},$$

com $q=(1-p)$.

Olkin & Tate (1961) apresentaram uma extensão multivariada análoga deste modelo. Seja $(Y_{1\alpha},\ldots,Y_{s\alpha},X_{0\alpha},\ldots,X_{k\alpha})$, $\alpha=1,\ldots,n$ uma sequência de vetores independentes, em que $\mathbf{X}=(X_0,X_1,\ldots,X_k)$ tem uma distribuição multinomial dada por

$$f(x_0,x_1,\ldots,x_k) = p_0^{x_0}p_1^{x_1}\ldots p_k^{x_k},\ x_m=0,1,$$

com $m=1,2,\ldots,k$.

A distribuição condicional de $\mathbf{Y}=(Y_1,\ldots,Y_s)$, dado $X_m=1$ é assumida $N_s(\boldsymbol{\mu}^{(m)},\boldsymbol{\Sigma})$, com vetor de médias $\boldsymbol{\mu}^{(m)}=(\mu_{1m},\ldots,\mu_{sm})$ e matriz de covariância $\boldsymbol{\Sigma}$, definida positiva.

Para facilidade no desenvolvimento do modelo foram considerados, separadamente, três casos: (i) $k=1,s>1$, (ii) $k>1,s=1$ e (iii) $k>1,s>1$.

É considerado o modelo $(Y_1,\ldots,Y_s|x_m=1,x_\nu=0,\nu\neq m=0,1,\ldots,k) \sim N_s(\boldsymbol{\mu}^{(m)},\boldsymbol{\Sigma})$ com médias condicionais dadas por

$$\begin{aligned} E(Y_1|x_m = 1) &= \mu_{1m} \\ \vdots & \quad \vdots \\ E(Y_s|x_m = 1) &= \mu_{sm} \end{aligned}$$

e covariâncias condicionais dadas por

$$\begin{aligned} Cov(Y_i, Y_j) &= \Psi_{ij} \\ Cov(X_m, X_s) &= \gamma_{ms} \\ Cov(Y_i, X_m) &= \delta_{im} \end{aligned}$$

com $i = 1, \ldots, s$, $j = 1, \ldots, s$, $m = 1, \ldots, k$ e $s = 1, \ldots, k$

Com isso, denotamos

$$Cov = \begin{pmatrix} Cov(Y_i, Y_j) & Cov(Y_i, X_m) \\ Cov(X_m, Y_i) & Cov(X_m, X_s) \end{pmatrix} = \begin{pmatrix} \mathbf{\Psi} & \mathbf{\Delta} \\ \mathbf{\Delta}' & \mathbf{\Gamma} \end{pmatrix}$$

em que

$$\mathbf{\Psi} = \mathbf{\Sigma} + \mathbf{U}\mathbf{D}_p\mathbf{U}'$$
$$\mathbf{\Delta} = \mathbf{U}\mathbf{D}_p$$
$$\mathbf{\Gamma} = \mathbf{D}_p - \mathbf{p}'\mathbf{p}$$

com $\mathbf{U} \equiv (\mu_{im} - \mu_i)$, $i = 1, 2, \ldots, s$, $m = 0, 1, \ldots, k$, $\mathbf{p} = (p_o, p_1, \ldots, p_k)$ e $\mathbf{D}_p = diag(p_o, p_1, \ldots, p_k)$. Tem-se que

A correlação múltipla entre $\mathbf{Y}$ e um subconjunto $(x_1, \ldots, x_l)$ de $(x_1, \ldots, x_k)$ pode ser:

$$\rho_y^2(x_1, \ldots, x_l) = \left(\sum_{m=0}^{l} p_m(\mu_m - \mu)^2 + \left[\sum_{m=0}^{l} p_m(\mu_m - \mu) \right]^2 \left[1 - \sum_{m=0}^{l} p_m \right]^{-1} \right) / \Psi_{11}$$

Distribuição para o caso (i) $k = 1, s > 1$

Sejam $(Y_{1\alpha}, \ldots, Y_{s\alpha}, X_{0\alpha}, X_{1\alpha})$, $\alpha = 1, \ldots, n$, n vetores independentes.

A distribuição condicional dos vetores $(Y_{1\alpha}, \ldots, Y_{s\alpha})$ é dada por

$$(Y_{1\alpha}, \ldots, Y_{s\alpha} | x_{m\alpha} = 1) \sim N_s(\boldsymbol{\mu}^{(m)}, \boldsymbol{\Sigma}),\ m = 0, 1.$$

Distribuição para o caso (ii) $k > 1, s = 1$

Sejam $(Y_\alpha, X_{0\alpha}, \ldots, X_{k\alpha})$, $\alpha = 1, \ldots, n$, n vetores independentes.

A distribuição condicional de Y_α é dada por

$$(Y_\alpha | x_{m\alpha} = 1) \sim N(\mu^m, \sigma^2),\ m = 0, 1, \ldots, k.$$

Distribuição para o caso (iii) $k > 1, s > 1$

Sejam $b_{im} = (\bar{y}_i^{(m)} - \bar{y}_i)$, $\mathbf{B} = (b_{im}) : s \times k + 1$, $\mathbf{D} = diag(n_0/n, \cdots, n_k/n)$, $h_{ij} = \sum_{m=0}^{k} n_m b_{im} b_{jm}/n$ e $\mathbf{H} = (h_{im}) : s \times s$. Então $\mathbf{BDB}'$ é uma estimativa de $\mathbf{UD}_p\mathbf{U}'$, lembrando que $U \equiv (\mu_{im} - \mu_i)$ e $D_p = diag(p_o, p_1, \ldots, p_k)$

Estimação

Sejam $h(y_\alpha, x_\alpha)$ a densidade do vetor $(Y_{1\alpha}, Y_{2\alpha}, \ldots, Y_{s\alpha}; X_{0\alpha}, X_{1\alpha}, \ldots, X_{k\alpha})$ e $f(x_\alpha)$ a densidade de $(X_{0\alpha}, X_{1\alpha}, \ldots, X_{k\alpha})$. Seja $\phi_m(y_\alpha)$ a densidade de um vetor $\mathbf{N}(\boldsymbol{\mu}^{(m)}, \boldsymbol{\Sigma})$. Assim,

$$f(x_\alpha) = p_0^{x_{0\alpha}} \ldots p_k^{x_{k\alpha}} \text{ e } h(y_\alpha, x_\alpha) = \sum_{m=0}^{k} x_{m\alpha} \phi_m(y_\alpha).$$

Portanto, a densidade de toda a amostra é

$$\prod_\alpha h(y_\alpha, x_\alpha) = \prod_\alpha (\sum_m x_{m\alpha} \phi_m(y_\alpha)) p_0^{x_{0\alpha}} \ldots p_k^{x_{k\alpha}}.$$

Dessa forma, a densidade conjunta é fatorada com um fator contendo os parâmetros p_0, p_1, ..., p_k e um outro fator contendo os parâmetros μ_{im} e σ_{ij}.

2.3.2 Modelo de Little & Schluchter (1985)

Little & Schluchter (1985) apresentaram procedimentos de máxima verossimilhança para análise de dados mistos contínuos e categóricos, com valores faltantes.

Sejam $\mathbf{X}_s$ o vetor de p variáveis contínuas e $\mathbf{Y}_s$ o vetor de q variáveis categóricas associadas ao indivíduo s, com $s = 1, \ldots, N$. Os dados analisados são incompletos de maneira que alguns $X_s^{'}$ e/ou alguns $Y_s^{'}$ podem ser faltantes. $\mathbf{W}_s$ é um vetor tal que $\mathbf{W}_s = \mathbf{E}_m$ se o indivíduo s pertencer a célula m da tabela de contingência e, $\mathbf{E}_m$ é um vetor com 1 no m-ésimo componente e zero nos demais componentes.

O modelo para os dados completos especifica a distribuição de (X_s, W_s), para $s = 1, \ldots, N$, em termos da distribuição condicional de X_s dado W_s e da distribuição marginal de W_s da seguinte maneira:

- Os N indivíduos observados apresentam uma distribuição multinomial com probabilidades $\pi_m = pr(W_s = E_m)$, para $s = 1, \ldots, N$ e $m = 1, \ldots, C$.
- Dado que $W_s = E_m$, X_s tem distribuição normal multivariada, $N_p(\boldsymbol{\mu}_m, \boldsymbol{\Omega})$, com vetor de médias $\boldsymbol{\mu}_m$ e matriz de covariâncias $\boldsymbol{\Omega}$.

Considerando $\boldsymbol{\Gamma} = [\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_C]$, a distribuição condicional de X_s dado W_s é $N_p(\boldsymbol{\Gamma} W_s, \boldsymbol{\Omega})$.

A função log verossimilhança, baseada nos dados completos, é dada por

$$L(\boldsymbol{\Gamma}, \boldsymbol{\Omega}, \boldsymbol{\pi}) = \sum_{s=1}^{N} log f(X_s | W_s, \boldsymbol{\Gamma}, \boldsymbol{\Omega}) + log[f(W_s | \boldsymbol{\pi})],$$

com $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_C)^{'}$.

No modelo para os dados incompletos define-se $\boldsymbol{\mu}_{s,obs,m} = E(X_{s,obs} | W_s = E_m)$, $m = 1, \ldots, C$ e $\boldsymbol{\Omega}_{s,obs} = cov(X_{s,obs} | W_s = E_m)$.

A densidade condicional de $X_{s,obs}$, dado que o valor do indivíduo s é conhecido, é uma mistura de densidades normais, ou seja,

$$f(X_{s,obs} | S_s, \boldsymbol{\Gamma}, \boldsymbol{\Omega}) = (\sum_{m \in S_s} \pi_m)^{-1} \sum_{m \in S_s} [\pi_m N_{p_s}(\boldsymbol{\mu}_{s,obs,m}, \boldsymbol{\Omega}_{s,obs})],$$

com $N_{p_s}(\boldsymbol{\mu}_{s,obs,m}, \boldsymbol{\Omega}_{s,obs})$ denotando a distribuição normal p_s-variada com média $\boldsymbol{\mu}_{s,obs,m}$ e matriz de covariância $\boldsymbol{\Omega}_{s,obs}$.

A função log verossimilhança para os dados incompletos, categóricos e contínuos, pode ser escrita como

$$\begin{aligned} L_I(\boldsymbol{\Gamma}, \boldsymbol{\Omega}, \pi) &= \sum_{s=1}^{n} \left\{ log f(X_{s,obs}|S_s, \boldsymbol{\Gamma}, \boldsymbol{\Omega}) + log \sum_{m \in S_s} \pi_m \right\} \\ &= -1/2 \sum_{s=1}^{n} p_s log(2\pi) - 1/2 \sum_{s=1}^{N} log|\boldsymbol{\Omega}_{s,obs}| \\ &+ \sum_{s=1}^{n} log \left[\sum_{m \in S_s} \pi_m exp\left\{ -1/2 (X_{s,obs} - \boldsymbol{\mu}_{s,obs,m})' \boldsymbol{\Omega}_{s,obs} (X_{s,obs} - \boldsymbol{\mu}_{s,obs,m}) \right\} \right]. \end{aligned}$$

Este trabalho é motivado por efeitos de distúrbios psicológicos dos pais em vários aspectos do desenvolvimento dos filhos. A idade escolar das crianças de famílias classificadas pelo grupo de risco em que os pais pertencem foi analisada. Como variáveis resposta foram considerados o nível de pontuação em relação à leitura padronizada e o nível de pontuação em relação à compreensão verbal padronizada.

A estimação foi feita por máxima verossimilhança, via algorítimo EM. Os estimadores de máxima verossimilhança obtidos para os dados completos foram $\hat{\pi} = N^{-1} \sum_s (W_s)$, $\hat{\boldsymbol{\Gamma}} = (\sum_s (X_s W_s'))(\sum_s (W_s W_s'))^{-1}$ e $\hat{\boldsymbol{\Omega}} = N^{-1} \sum_s [(X_s - \hat{\boldsymbol{\Gamma}} W_s)(X_s - \hat{\boldsymbol{\Gamma}} W_s)']$.

2.3.3 Modelo de Catalano & Ryan (1992)

Catalano & Ryan (1992) desenvolveram um modelo bivariado motivado por estudos de toxicidade em ratas grávidas, considerando má formação fetal como variável resposta discreta e peso fetal como variável resposta contínua. Este modelo permite respostas conjuntas discretas e contínuas, primeiramente assumindo independência entre as observações em diferentes fetos e depois considerando a dependência.

Modelos bivariados

Nos modelos apresentados a seguir, a variável Y_{1ij} é relacionada ao peso fetal e a variável Y_{2ij} é uma variável latente não observada correspondente a má formação,

para o feto j da ninhada i, com $j = 1, \ldots, n_i$ e $i = 1, \ldots, I$. Considera-se d_i a dose administrada na i-ésima rata grávida.

O modelo bivariado sugerido é dado por

$$\begin{aligned} Y_{1ij} &= \alpha_0 + \alpha_1 d_i + \epsilon_{1ij} \\ Y_{2ij} &= \beta_0 + \beta_1 d_i + \epsilon_{2ij} \end{aligned}$$

O parâmetro α corresponde ao peso fetal e parâmetro β corresponde à má formação.

Modelo bivariado independente

Suponha que os fetos na mesma ninhada são independentes e que

$$\epsilon_{ij} = \begin{pmatrix} \epsilon_{1ij} \\ \epsilon_{2ij} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \tau\sigma_1\sigma_2 \\ \tau\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}\right),$$

em que os resíduos ϵ_{ij} são independentes para todos os i e j.

Para um mesmos feto, $Y_{1_{ij}}$ e $Y_{2_{ij}}$ têm correlação constante τ para todo i e j. Denota-se por Y^*_{2ij} a variável indicadora observada para má formação, determinada pela variável latente Y_{2ij}, para o feto j da ninhada i, ou seja,

$$Y^*_{2ij} = \begin{cases} 1 & se \quad Y_{2ij} > 0 \\ 0 & se \quad Y_{2ij} \leq 0 \end{cases}$$

em que $Y^*_{2ij} = 1$ indica ocorrência de má formação.

A partir do modelo escrito inicialmente $\mathbf{Y}^*_2 = (Y_{2i1}, Y_{2i2}, \ldots, Y_{2in_i})$ segue um modelo probito, ou seja

$$P(Y^*_{2ij} = 1) = \Phi\left(\frac{\beta_0 + \beta_1 d_i}{\sigma_2}\right).$$

A distribuição conjunta de Y_{1ij} e Y^*_{2ij} é obtida pelo produto da distribuição marginal e da distribuição condicional, como segue

$$f_{Y_{1ij}, Y^*_{2ij}}(y_1, y^*_2) = f_{Y_{1ij}}(y_1) f_{Y^*_{2ij}|Y_{1ij}}(y^*_2|y_1).$$

$Y_{2ij}|Y_{1ij}$ segue distribuição normal com uma média que depende do resíduo do modelo para $\mathbf{Y}_1$

$$Y_{2ij}|Y_{1ij} \sim N(\mu_1, \sigma_2^2(1-\tau^2)),$$

com $\mu_1 = \beta_0 + \beta_1 d_i + \left(\frac{\sigma_2}{\sigma_1}\right)\tau e_{1ij}$, em que $e_{1ij} = Y_{1ij} - (\alpha_0 + \alpha_1 d_i)$ é o resíduo do modelo para o peso.

A partir da distribuição de $Y_{2ij}|Y_{1ij}$ obtêm-se a distribuição condicional de $Y^*_{2ij}|Y_{1ij}$ que também é um modelo probito, ou seja

$$P(Y^*_{2ij} = 1|Y_{1ij}) = \Phi\left(\frac{\mu_1}{\sqrt{\sigma_2^2(1-\tau^2)}}\right),$$

que pode ser reparametrizado, obtendo

$$P(Y^*_{2ij} = 1|Y_{1ij}) = \Phi(\beta^*_0 + \beta^*_1 d_i + \beta^*_2 e_{1ij}).$$

Como grandes valores da variável latente Y_{2ij} resultam em má formação, espera-se que Y_{1ij} e Y_{2ij} sejam negativamente correlacionados ($\tau < 0$).

Modelo bivariado correlacionado

Sejam τ a correlação constante entre observações de um mesmo feto e ρ_1 e ρ_2 as correlação separadas entre as observações em diferentes fetos, na mesma ninhada, relacionadas a peso e a má formação, respectivamente. Seja ρ_{12} a correlação entre a variável do peso e a variável latente de má formação, para diferentes fetos na mesma ninhada.

Considere, para a ninhada i, que $\mathbf{Y}_{1i} = (Y_{1i1}, \ldots, Y_{1in_i})'$ seja o vetor para o peso e $\mathbf{Y}_{2i} = (Y_{2i1}, \ldots, Y_{2in_i})'$ seja o vetor para a variável latente.

Tomando $\mathbf{Y}_i = \big((\mathbf{Y}_{1i_1}, \mathbf{Y}_{2i_1}), (\mathbf{Y}_{1i_2}, \mathbf{Y}_{2i_2}), \cdots, (\mathbf{Y}_{1i_{n_i}}, \mathbf{Y}_{2i_{n_i}})\big)$ como o vetor de observações bivariadas para a ninhada i, $\mathbf{Y}_i$ tem distribuição normal multivariada com média

$$E(\mathbf{Y}_i) = E\begin{pmatrix} \mathbf{Y}_{1i} \\ \mathbf{Y}_{2i} \end{pmatrix} = \begin{pmatrix} \mathbf{1} & d_i\mathbf{1} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{1} & d_i\mathbf{1} \end{pmatrix} \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \beta_0 \\ \beta_1 \end{pmatrix}$$

e matriz de covariância

$$Var(\mathbf{Y}_i) = \begin{pmatrix} \sigma_1^2[(1-\rho_1)\mathbf{I} + \rho_1\mathbf{J}] & \sigma_1\sigma_2[(\tau-\rho_{12})\mathbf{I} + \rho_{12}\mathbf{J}] \\ \sigma_1\sigma_2[(\tau-\rho_{12})\mathbf{I} + \rho_{12}\mathbf{J}] & \sigma_2^2[(1-\rho_2)\mathbf{I} + \rho_2\mathbf{J}] \end{pmatrix},$$

em que $\mathbf{I}$ é a matriz identidade e $\mathbf{J}$ é uma matriz de uns.

A distribuição condicional da variável latente, dado o vetor do peso fetal, é normal

$$\mathbf{Y}_{2i}|\mathbf{Y}_{1i} \sim N_{n_i}(\boldsymbol{\mu}_{2i}, \sigma_2^2\boldsymbol{\Sigma}_i)$$

na qual o j-ésimo elemento da média condicional é dado por

$$\mu_{2ij} = \beta_0 + \beta_1 d_i + \frac{\sigma_2}{\sigma_1}\left(\frac{\tau + (n_i-1)\rho_{12}}{1+(n_i-1)\rho_1}\right)\overline{e}_{1i} + \frac{\sigma_2}{\sigma_1}\left(\frac{\tau-\rho_{12}}{1-\rho_1}\right)(e_{1ij} - \overline{e}_{1i})$$

com $e_{1ij} = Y_{1ij} - (\alpha_0 + \alpha_1 d_i)$ sendo o resíduo do modelo para Y, $\overline{e}_{1i} = \overline{Y}_{1i} - (\alpha_0 + \alpha_1 d_i)$ sendo a média do resíduo e_{1ij} e

$$\boldsymbol{\Sigma}_i = \left[(1-\rho_2) - \frac{(\tau-\rho_{12})^2}{1-\rho_1}\right]\mathbf{I} + \left[\rho_2 - \frac{(1-\rho_1)(\tau^2 + (n_i-1)\rho_{12}^2) - (\tau-\rho_{12})^2}{(1-\rho_1)(1+(n_i-1)\rho_1)}\right]\mathbf{J}.$$

A distribuição condicional para o indicador de má formação observável, dado o peso fetal, segue um modelo probito correlacionado com

$$E(Y_{2ij}^*|Y_{1ij}) = \Phi\left(\frac{\mu_{2ij}}{\sigma_3}\right)$$

e

$$Var(Y_{2ij}^*|Y_{1ij}) = \Phi\left(\frac{\mu_{2ij}}{\sigma_{3i}}\right)\left[1 - \Phi\left(\frac{\mu_{2ij}}{\sigma_{3i}}\right)\right]$$

em que $\sigma_{3i}^2 = \sigma_2^2\left[1 - \dfrac{\tau^2(1-\rho_1) + (n_i-1)[\rho_1(\rho_{12}-\tau)^2 + (1-\rho_1)\rho_{12}^2]}{(1-\rho_1)(1+(n_i-1)\rho_1)}\right]$.

Estimação

Uma abordagem convencional para ajuste o da distribuição conjunta para os dados observados seria construir a verossimilhança bivariada e usar técnicas de máxima verossimilhança. Catalano & Ryan (1992) desenvolveram um procedimento de estimação baseado na fatoração da distribuição conjunta em dois componentes de regressão, a distribuição marginal de $\mathbf{Y}_1$ e a distribuição condicional de $\mathbf{Y}_2^*|\mathbf{Y}_1$, e aplicaram uma abordagem de equações de estimação generalizadas (GEE) em cada componente.

Os procedimentos de estimação foram feitos em duas etapas. A primeira etapa é relacionada a ajustar uma regressão correlacionada do peso $\mathbf{Y}_1$ sobre a dose d e o tamanho da ninhada $(n_i - \bar{n})$. A segunda etapa é relacionada ao ajuste de uma regressão probito correlacionada de $\mathbf{Y}_2^*|\mathbf{Y}_1$, usando a dose, assim como resíduos do peso fetal médio e individual e covariáveis do tamanho da ninhada como variáveis explicativas e um parâmetro de correlação constante para explicar o efeito da ninhada.

A primeira etapa envolve a solução da equação de estimação dada por

$$\sum_{i=1}^{I} \mathbf{G}_{1i}^{'} \mathbf{V}_{1i}^{-1} (\mathbf{Y}_{1i} - \mathbf{G}_{1i}\boldsymbol{\alpha}) = \mathbf{0},$$

em que $\mathbf{V}_{1i}$ é a matriz de covariância de $\mathbf{Y}_{1i}$, G é a matriz de regressão, $n_i \times 3$, com uma coluna para o intercepto, uma coluna para a dose, d_i, e uma coluna para o tamanho da ninhada, $(n_i - \bar{n})$, e $\boldsymbol{\alpha}$ é o vetor dos parâmetros do modelo . Desta etapa, foram obtidas estimativas e variâncias dos parâmetros do modelo, α_0, α_1 e α_2, bem como uma estimativa de ρ_1.

A segunda etapa envolve solução de um conjunto de equações de estimação para o modelo probito correlacionado,

$$\sum_{i=1}^{I} \frac{\partial \mu_i}{\partial \boldsymbol{\beta}} \mathbf{V}_{2i}^{-1} (\mathbf{Y}_{2i}^* - \boldsymbol{\mu}_i) = \mathbf{0},$$

em que $\mathbf{V}_{2i}$ é a matriz de covariância para $\mathbf{Y}_{2i}^*|\mathbf{Y}_{1i}$, $\boldsymbol{\mu}_i$ é a média condicional de $\mathbf{Y}_{2i}^*|\mathbf{Y}_{1i}$ e $\boldsymbol{\beta}$ é o vetor de parâmetros. Estimativas das covariáveis residuais do peso,

$e_{1_{ij}} = Y_{1ij} - (\alpha_0 + \alpha_1 d_i)$ e $\bar{e}_{1i} = \overline{Y}_{1i} - (\alpha_0 + \alpha_1 d_i)$, foram obtidas dos resultados da primeira etapa.

2.3.4 Modelo de Fitzmaurice & Laird (1995)

Fitzmaurice & Laird (1995) motivados por estudos de toxicidade em ratas grávidas, avaliando em cada feto evidência de má formação e o peso fetal, desenvolveram um modelo bivariado com uma resposta binária para má formação e uma resposta contínua, condicionada na resposta discreta, para peso fetal.

Modelo bivariado

Sejam Y_i a variável resposta discreta, X_i a variável resposta contínua, $\mathbf{Z}_{1i}$ o vetor de covariáveis relacionado a Y_i e $\mathbf{Z}_{2i}$ o vetor de covariáveis relacionado a X_i, com $i = 1, \ldots, N$.

A distribuição marginal de Y_i é Bernoulli, ou seja

$$f(y_i|\mathbf{Z}_{1i}) = exp[y_i\theta_i - log\{1 + exp(\theta_i)\}],$$

com $\theta_i = log\left\{\dfrac{\mu_{1i}}{(1-\mu_{1i})}\right\} = \mathbf{Z}_{1i}\beta_1$ e $\mu_{1i} = \mu_{1i}(\beta_1) = E(Y_i) = Pr(Y_i = 1|\mathbf{Z}_{1i}, \beta_1)$.

A densidade conjunta de (X_i, Y_i) pode ser escrita como

$$f_{X_i,Y_i}(x_i, y_i) = f_{Y_i}(y_i) f_{X_i|Y_i}(x_i|y_i).$$

Assumindo que a distribuição de X_i, condicionada a Y_i, é normal, temos

$$f_{X_i|Y_i}(x_i|y_i) = (2\pi\sigma^2)^{-1/2} exp\left\{-\frac{1}{2\sigma^2}[x_i - \mu_{2i} - \gamma(y_i - \mu_{1i})]^2\right\},$$

com $\mu_{2i} = \mathbf{Z}_{2i}\beta_2$ e γ é um parâmetro para a regressão de X_i em Y_i, que induz associação ou correlação entre Y_i e X_i.

Devemos observar que a média de $X|Y$, escrita como $\mu_{2i} + \gamma(y_i - \mu_{1i})$ é obtida da distribuição normal bivariada, considerando ambas as variáveis seguindo distribuição normal, cuja média é obitida como $E(X) + \gamma(y - E(Y))$ e substituindo a distribuição

da variável Y por Bernoulli. Dessa mesma maneira buscamos proceder para o nosso modelo.

Sem perda de generalidade assume-se que o vetor de covariáveis $\mathbf{Z}_i$ prediz ambas respostas Y_i e X_i. Para simplificar a notação, considere $\boldsymbol{\alpha} = (\beta_2, \gamma)$ e $\mathbf{W}_i = [\mathbf{Z}_i, (Y_i - \mu_{1i})]$. Dessa forma, $E(X_i|Y_i) = \mathbf{W}_i\boldsymbol{\alpha}$.

Por construção, $E(X_i) = \mathbf{Z}_i\beta_2$ e ambos os parâmetros de regressão β_1 e β_2 têm interpretações marginais.

As equações de verossimilhança são dadas por

$$\begin{aligned}
\sum_{i=1}^{N} \frac{\partial l_i(x_i, y_i)}{\partial \beta_1} &= \sum_{i=1}^{N} (\mathbf{Z}_i'(y_i - \mu_{1i}) - \mathbf{Z}_i'\Delta_i(x_i - \mathbf{W}_i\boldsymbol{\alpha})\gamma\sigma^{-2}) \\
\sum_{i=1}^{N} \frac{\partial l_i(x_i, y_i)}{\partial \alpha} &= \sum_{i=1}^{N} (\mathbf{W}_i'(x_i - \mathbf{W}_i\boldsymbol{\alpha})\sigma^{-2}) \\
\sum_{i=1}^{N} \frac{\partial l_i(x_i, y_i)}{\partial \sigma^2} &= \Phi^{-1} \sum_{i=1}^{N} (c_i - \sigma^2)
\end{aligned}$$

com $\Delta_i = var(Y_i) = \mu_{1i}(1 - \mu_{1i})$, $C_i = (x_i - \mathbf{W}_i\boldsymbol{\alpha})^2$ e $\Phi = var(C_i)$

As equações de verossimilhança para $(\beta_1, \boldsymbol{\alpha}, \sigma^2)$ podem ser escritas na forma matricial, como segue

$$\sum_{i=1}^{N} \begin{pmatrix} \dfrac{\partial l_i}{\partial \beta_1} \\ \\ \dfrac{\partial l_i}{\partial \boldsymbol{\alpha}} \\ \\ \dfrac{\partial l_i}{\partial \sigma^2} \end{pmatrix} = \sum_{i=1}^{N} \begin{pmatrix} \mathbf{Z}_i'\Delta_i & -\gamma\mathbf{Z}_i'\Delta_i & 0 \\ 0 & \mathbf{W}_i' & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \Delta_i^{-1} & 0 & 0 \\ 0 & \sigma^{-2} & 0 \\ 0 & 0 & \Phi^{-1} \end{pmatrix} \begin{pmatrix} y_i - \mu_{1_i} \\ x_i - W_i\boldsymbol{\alpha} \\ c_i - \sigma^2 \end{pmatrix} =$$

$$\sum_{i=1}^{N} \begin{pmatrix} \dfrac{\partial E(Y_i)}{\partial \beta_1} & \dfrac{\partial E(Y_i)}{\partial \boldsymbol{\alpha}} & \dfrac{\partial E(Y_i)}{\partial \sigma^2} \\ \\ \dfrac{\partial E(X_i|Y_i)}{\partial \beta_1} & \dfrac{\partial E(X_i|Y_i)}{\partial \boldsymbol{\alpha}} & \dfrac{\partial E(X_i|Y_i)}{\partial \sigma^2} \\ \\ \dfrac{\partial E(C_i|Y_i)}{\partial \beta_1} & \dfrac{\partial E(C_i|Y_i)}{\partial \boldsymbol{\alpha}} & \dfrac{\partial E(C_i|Y_i)}{\partial \sigma^2} \end{pmatrix}' \times cov^{-1} \begin{pmatrix} Y_i \\ X_i|Y_i \\ C_i|Y_i \end{pmatrix} \begin{pmatrix} y_i - E(Y_i) \\ x_i - E(X_i|Y_i) \\ c_i - E(C_i|Y_i) \end{pmatrix}.$$

A covariância de $(\widehat{\beta_1}, \widehat{\beta_2}, \widehat{\gamma}, \widehat{\sigma^2})$ pode ser aproximada pelo inverso da matriz de informação de Fisher, ou seja,

$$Cov^{-1}(\widehat{\beta_1},\widehat{\beta_2},\widehat{\gamma},\widehat{\sigma^2}) \approx \sum_{i=1}^{N} \begin{bmatrix} (\Delta_i + \Delta_i^2\gamma^2\sigma^{-2})\mathbf{Z}_i^{'}\mathbf{Z}_i & (\sigma^{-2}\gamma\Delta_i)\mathbf{Z}_i^{'}\mathbf{Z}_i & 0 & 0 \\ (\sigma^{-2}\gamma\Delta_i)\mathbf{Z}_i^{'}\mathbf{Z}_i & (\sigma^{-2})\mathbf{Z}_i^{'}\mathbf{Z}_i & 0 & 0 \\ 0 & 0 & \Delta_i\sigma^{-2} & 0 \\ 0 & 0 & 0 & 2\sigma^4 \end{bmatrix}.$$

Modelo bivariado correlacionado

Fitzmaurice & Laird (1995) apresentaram também uma extensão deste modelo para permitir agrupamento. Neste modelo, cada um dos N clusters tem n_i vetores bivariados de respostas (X_{ij}, Y_{ij}), $j = 1, \ldots, n_i$, ou seja, os vetores de respostas para o i-ésimo cluster são dados por $(X_{i1}, Y_{i1}), (X_{i2}, Y_{i2}), \ldots, (X_{in_i}, Y_{in_i})$.

Além disso, cada unidade dentro de um cluster tem um vetor de covariáveis $\mathbf{z}_{ik}$, ou seja, a matriz de covariáveis para o cluster i pode ser escrita como $\mathbf{Z}_i = (\mathbf{z}_{i1}, \mathbf{z}_{i2}, \ldots, \mathbf{z}_{in_i})^{'}$.

Assume-se, assim, que o modelo para a média de $\mathbf{Y}_i$ e de $X_{ik}|\mathbf{Y}_i$ é dado por

$$logit(E[\mathbf{Y}_i]) = \mathbf{Z}_i\beta_1$$

e

$$E[X_{ik}|\mathbf{Y}_i] = \mathbf{z}_{ik}\beta_2 + \gamma_1(Y_{ik} - \mu_{1ik}) + \gamma_2 S_i,$$

em que $logit(E[\mathbf{Y}_i]) = (logit[\mu_{1i_1}], logit[\mu_{1i_2}], \ldots, logit[\mu_{1i_{n_i}}])^{'}$ e $S_i = \sum_{j=1}^{n_i}(Y_{ij} - \mu_{1_{ij}})$.

O vetor de parâmetros $\mathbf{\Gamma} = (\gamma_1, \gamma_2)$ induz a correlação entre Y_{ij} e X_{ij}, isto é, $(\gamma_1+\gamma_2)$ caracteriza a associação entre as respostas binária e contínua tomadas na mesma unidade dentro de um cluster, enquanto que γ_2 caracteriza esta associação para diferentes membros dentro do mesmo cluster.

Com o objetivo de simplificar a notação, considere $\boldsymbol{\alpha} = (\beta_2, \mathbf{\Gamma})$ um vetor de parâmetros e $\mathbf{W}_i = (\mathbf{W}_{i1}, \mathbf{W}_{i2}, \ldots, \mathbf{W}_{in_i})^{'}$ uma matriz de covariáveis com $\mathbf{W}_{ik} = [\mathbf{z}_{ik}, (Y_{ik} - \mu_{1_{ik}}), S_i]$.

São assumidas correlações intraclusters separadas, ρ_Y para a resposta binária e ρ_X para a resposta contínua, isto é, assume-se que

$$cov(\mathbf{Y}_i) = \mathbf{V}_{1i} \approx \Delta_i^{1/2}[(1-\rho_Y)\mathbf{I}_i + \rho_Y\mathbf{J}_i]\Delta^{1/2}$$

e

$$cov(\mathbf{X}_i|\mathbf{Y}_i) = \mathbf{V}_{2i} \approx \sigma^2[(1-\rho_X)\mathbf{I}_i + \rho_X\mathbf{J}_i],$$

em que Δ_i é uma matriz diagonal com elementos $var(Y_{ij})$, $\mathbf{I}_i$ é uma matriz identidade e $\mathbf{J}_i$ é uma matriz de $1's$.

A metodologia de equações de estimação generalizadas (GEE) pode ser utilizada para a estimação de $(\beta_1, \boldsymbol{\alpha})$, ou seja, pode-se construir um conjunto de equações de estimação generalizadas para $(\beta_1, \boldsymbol{\alpha})$ baseado nas equações escores, dado por

$$\sum_{i=1}^{n} \begin{pmatrix} \mathbf{Z}_i'\boldsymbol{\Delta}_i & -(\gamma_1+\gamma_2)\mathbf{Z}_i'\boldsymbol{\Delta}_i \\ \mathbf{0} & \mathbf{W}_i' \end{pmatrix} \begin{pmatrix} \mathbf{V}_{1_i}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{V}_{2_i}^{-1} \end{pmatrix} \begin{pmatrix} \boldsymbol{y}_i - \boldsymbol{\mu}_{1_i} \\ \boldsymbol{x}_i - \mathbf{W}_i\boldsymbol{\alpha} \end{pmatrix} = 0.$$

Seguindo esta abordagem pode-se estimar $(\beta_1, \boldsymbol{\alpha})$ por um algoritmo modificado de Fisher e $(\sigma^2, \rho_Y, \rho_X)$ por método de momentos.

Fitzmaurice & Laird (1995) apresentaram uma aplicação da metodologia através de dados de um estudo de desenvolvimento de toxicidade em fêmeas de ratos. O modelo assumido para esses dados, relacionando a média das respostas de má formação e peso fetal ao efeito linear da dose, é dado por

$$\begin{aligned} logit(E[Y_{ik}]) &= logit(\mu_{1_{ik}}) = \beta_{10} + \beta_{11}d_i, \\ E[X_{ik}|\mathbf{Y}_i] &= \beta_{20} + \beta_{21}d_i + \gamma_1(Y_{ik} - \mu_{1_{ik}}) + \gamma_2\bar{S}_i, \end{aligned}$$

em que d_i é a dose administrada na i-ésima rata e $\bar{S}_i = n_i^{-1} \sum_{j=1}^{n_i} (Y_{ij} - \mu_{1_{ij}})$.

Os parâmetros de regressão para a dose indicaram que a taxa de má formação aumentou com a dose, enquanto que a média de peso fetal diminuiu com o aumento da dose.

Capítulo 3

O Modelo de Regressão Bernoulli - Exponencial

3.1 Introdução

O modelo de regressão Bernoulli-Exponencial é um modelo bivariado no qual a variável resposta discreta segue distribuição Bernoulli e a variável resposta contínua, condicionada à observação da variável discreta, segue distribuição exponencial. Este modelo é motivado, entre outros, por gastos de planos de saúde com pacientes hospitalizados, levando em consideração a utilização ou não de tratamento cirúrgico.

Neste modelo a variável resposta discreta, Y_i, tem média μ_{1i} e a variável contínua, condicionada a Y_i, $X_i|Y_i$, tem média λ_i. Seja z_i a covariável disponível para o i-ésimo paciente e β_1, β_2 e γ os parâmetros desconhecidos, com γ sendo o parâmetro de associação entre a variável discreta e a variável contínua.

A função de probabilidade para a variável discreta é dada por

$$f_{Y_i}(y_i) = \mu_{1i}^{y_i}(1-\mu_{1i})^{1-y_i},\ y = 0 \text{ ou } y = 1$$

em que a média μ_{1i} é ligada à covariável por uma função de ligação logito, ou seja, $log\left(\frac{\mu_{1i}}{1-\mu_{1i}}\right) = z_i\beta_1$. Assim, $\mu_{1i} = \frac{e^{z_i\beta_1}}{1+e^{z_i\beta_1}}$.

A função de densidade para a variável contínua, condicionada a Y_i, é dada por

$$f_{X_i|Y_i}(x_i|y_i) = \frac{1}{\lambda_i} exp\left\{-\frac{x_i}{\lambda_i}\right\},$$

em que λ_i é construído como função da covariável disponível, z_i, da observação y_i, uma realização da variável aleatória Y_i, e da média marginal.

A função de probabilidade $f_{Y_i}(y_i) = \mu_{1i}^{y_i}(1-\mu_{1i})^{1-y_i}$ é reparametrizada como segue

$$\begin{aligned}
f_{Y_i}(y_i) &= \left(\frac{e^{z_i\beta_1}}{1+e^{z_i\beta_1}}\right)^{y_i}\left[1-\left(\frac{e^{z_i\beta_1}}{1+e^{z_i\beta_1}}\right)\right]^{1-y_i} \\
&= exp\left\{y_i\left[log\left(\frac{e^{z_i\beta_1}}{1+e^{z_i\beta_1}}\right)\right] + (1-y_i)\left[log\left(1-\frac{e^{z_i\beta_1}}{1+e^{z_i\beta_1}}\right)\right]\right\} \\
&= exp\left\{y_i\left[log\left(\frac{\frac{e^{z_i\beta_1}}{1+e^{z_i\beta_1}}}{\frac{1}{1+e^{z_i\beta_1}}}\right)\right] + log\left(\frac{1}{1+e^{z_i\beta_1}}\right)\right\} \\
&= exp\left\{y_i log(e^{z_i\beta_1}) + log\left(\frac{1}{1+e^{z_i\beta_1}}\right)\right\} \\
&= exp\left\{y_i z_i\beta_1 - log\left(1+exp(z_i\beta_1)\right)\right\}.
\end{aligned}$$

Desta forma, a função de probabilidade para Y_i pode ser escrita como

$$f_{Y_i}(y_i) = exp[y_i z_i\beta_1](1+exp(z_i\beta_1))^{-1}.$$

Descrevemos a seguir três diferentes modelos bivariados Bernoulli-Exponencial que se diferenciam pela construção do parâmetro λ_i, relacionado à variável X_i condicionada a Y_i. No Modelo 1 temos λ_i construído através de uma regressão linear, no Modelo 2 temos λ_i construído dependendo da variável aleatória Y_i e das médias marginais μ_{1i} e μ_{2i} e no Modelo 3 temos λ_i construído dependendo apenas da variável aleatória Y_i e da média marginal μ_{2i}.

3.2 Modelo 1

Nesta seção apresentamos um modelo no qual o parâmetro λ_i é obtido através de uma regressão linear, utilizando uma função de ligação logarítmica, dado por

$$log(\lambda_i) = z_i\beta_2 + \gamma y_i,$$

ou seja,

$$\lambda_i = exp(z_i\beta_2 + \gamma y_i).$$

Assim, a função de densidade para a variável X_i, condicionada a Y_i, para este modelo, é dada por

$$f_{X_i|Y_i}(x_i|y_i) = \frac{1}{exp(z_i\beta_2 + \gamma y_i)} exp\left\{-\frac{x_i}{exp(z_i\beta_2 + \gamma y_i)}\right\}.$$

A função de densidade conjunta de (X_i, Y_i) pode ser escrita como o produto da distribuição marginal de Y_i e da distribuição condicional de $X_i|Y_i$, ou seja,

$$f_{X_i,Y_i}(x_i, y_i) = f_{Y_i}(y_i) f_{X_i|Y_i}(x_i|y_i).$$

Dada uma amostra aleatória de tamanho n, $(y_1, x_1), (y_2, x_2), \ldots, (y_n, x_n)$, do vetor bivariado (Y, X) na qual a variável Y_i segue distribuição Bernoulli com média μ_{1i} e a variável X_i segue distribuição condicional exponencial com média λ_i, a função de verossimilhança conjunta é dada por

$$L(\beta_1, \beta_2, \gamma; \mathbf{y}, \mathbf{x}, \mathbf{z}) = \prod_{i=1}^{n} \left[f_{Y_i}(y_i) f_{X_i|Y_i}(x_i|y_i)\right]$$

$$= \prod_{i=1}^{n} \left[exp(y_i z_i \beta_1)(1 + exp(z_i\beta_1))^{-1} \frac{1}{exp(z_i\beta_2 + \gamma y_i)} exp\left(-\frac{x_i}{exp(z_i\beta_2 + \gamma y_i)}\right)\right],$$

com $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, valores observados da variável X, $\mathbf{y} = (y_1, y_2, \ldots, y_n)$, valores observados da variável Y e $\mathbf{z} = (z_1, z_2, \ldots, z_n)$, valores observados da covariável disponível.

3.3 Modelo 2

Nesta seção apresentamos a construção de λ_i feita de maneira que este parâmetro dependa da variável aleatória Y_i, de sua média marginal, μ_{1i} e da média marginal da variável X_i, μ_{2i} que é ligada à covariável por uma função de ligação logarítmica, ou seja, $\mu_{2i} = exp(z_i\beta_2)$. Assim temos,

$$\lambda_i = exp(z_i\beta_2) + \gamma y_i \mu_{1_i}.$$

A função de densidade para a variável X_i, condicionada a Y_i, para este modelo, é dada por

$$f_{X_i|Y_i}(x_i|y_i) = \frac{1}{exp(z_i\beta_2) + \gamma y_i \mu_{1_i}} exp\left[-\frac{x_i}{(exp(z_i\beta_2) + \gamma y_i \mu_{1_i})}\right].$$

A função de densidade conjunta de (X_i, Y_i) pode ser escrita como o produto da distribuição marginal de Y_i e da distribuição condicional de $X_i|Y_i$, ou seja,

$$f_{X_i,Y_i}(x_i, y_i) = f_{Y_i}(y_i) f_{X_i|Y_i}(x_i|y_i).$$

Dada uma amostra aleatória de tamanho n, $(y_1, x_1), (y_2, x_2), \ldots, (y_n, x_n)$, do vetor bivariado (Y, X) na qual a variável Y_i segue distribuição Bernoulli com média μ_{1i} e a variável X_i segue distribuição condicional exponencial com média λ_i, a função de verossimilhança conjunta é dada por

$$L(\beta_1, \beta_2, \gamma; \mathbf{y}, \mathbf{x}, \mathbf{z}) = \prod_{i=1}^{n} \left[f_{Y_i}(y_i) f_{X_i|Y_i}(x_i|y_i) \right]$$

$$= \prod_{i=1}^{n} \left[exp(y_i z_i \beta_1)(1 + exp(z_i \beta_1))^{-1} \frac{1}{exp(z_i \beta_2) + \gamma y_i \mu_{1_i}} exp\left(- \frac{x_i}{exp(z_i \beta_2) + \gamma y_i \mu_{1_i}} \right) \right],$$

com $i = 1, 2, \ldots, n$ e x_i, y_i e z_i valores observados da variável X, da variável Y e da covariável disponível, respectivamente.

3.4 Modelo 3

Nesta seção apresentamos um modelo no qual λ_i é construído da mesma forma do Modelo 2, considerando que este parâmetro dependa apenas da variável aleatória Y_i e da covariável. Para esse modelo temos,

$$\lambda_i = exp(z_i \beta_2) + \gamma y_i.$$

Com isso, a função de densidade para a variável X_i, condicionada a Y_i, para este modelo, é dada por

$$f_{X_i|Y_i}(x_i|y_i) = \frac{1}{exp(z_i \beta_2) + \gamma y_i} exp\left[- \frac{x_i}{(exp(z_i \beta_2) + \gamma y_i)} \right].$$

A função de densidade conjunta de (X_i, Y_i) pode ser escrita como o produto da distribuição marginal de Y_i e da distribuição condicional de $X_i|Y_i$, ou seja,

$$f_{X_i, Y_i}(x_i, y_i) = f_{Y_i}(y_i) f_{X_i|Y_i}(x_i|y_i).$$

Dada uma amostra aleatória de tamanho n, $(y_1, x_1), (y_2, x_2), \ldots, (y_n, x_n)$, do vetor bivariado (Y, X) na qual a variável Y_i segue distribuição Bernoulli com média μ_{1i} e a variável X_i segue distribuição condicional exponencial com média λ_i, a função de verossimilhança conjunta é dada por

$$L(\beta_1, \beta_2, \gamma; \mathbf{y}, \mathbf{x}, \mathbf{z}) = \prod_{i=1}^{n} \left[f_{Y_i}(y_i) f_{X_i|Y_i}(x_i|y_i) \right]$$

$$= \prod_{i=1}^{n} \left[exp(y_i z_i \beta_1)(1 + exp(z_i \beta_1))^{-1} \frac{1}{exp(z_i \beta_2) + \gamma y_i} exp\left(- \frac{x_i}{exp(z_i \beta_2) + \gamma y_i} \right) \right],$$

em que x_i é um valor observado da variável X, y_i é um valor observado da variável Y e z_i é um valor observado da covariável disponível, para $i = 1, 2, \ldots, n$.

3.5 Estimação

Para estimação dos parâmetros β_1, β_2 e γ consideramos uma abordagem Bayesiana via MCMC.

Na abordagem Bayesiana, os parâmetros seguem distribuições a priori, neste caso, não informativas. Através do Teorema de Bayes, pelo produto das distribuições a priori dos parâmetros com a função de verossimilhança, obtemos a distribuição a posteriori dos parâmetros. Tais distribuições podem ser intratáveis analiticamente e, por este motivo, utilizamos o método de aproximação Metrópolis.

As distribuições a priori para os parâmetros β_1, β_2 e γ, para os três modelos em estudo, são consideradas como $\beta_1 \sim N(\theta_1, \sigma_1^2)$, $\beta_2 \sim N(\theta_2, \sigma_2^2)$ e $\gamma \sim N(\theta_3, \sigma_3^2)$, com hiperparâmetros θ_1, σ_1^2, θ_2, σ_2^2, θ_3 e σ_3^2 conhecidos.

Considerando as funções de verossimilhança descritas anteriormente para cada modelo obtemos a sua distribuição a posteriori.

A distribuição a posteriori para o Modelo 1 é dada por

$$\begin{aligned}\pi(\beta_1,\beta_2,\gamma|\mathbf{y},\mathbf{x},\mathbf{z}) &\propto exp\left\{-\frac{1}{2}\left[\left(\frac{\beta_1-\theta_1}{\sigma_1}\right)^2+\left(\frac{\beta_2-\theta_2}{\sigma_2}\right)^2+\left(\frac{\gamma-\theta_2}{\sigma_2}\right)^2\right]\right\}\\ &\times \prod_{i=1}^{n}\left[exp(y_i z_i\beta_1)[1+exp(z_i\beta_1)]^{-1}\right]\\ &\times \prod_{i=1}^{n}\left[\frac{1}{exp(z_i\beta_2+\gamma y_i)}exp\left\{-\frac{x_i}{exp(z_i\beta_2+\gamma y_i)}\right\}\right].\end{aligned}$$

A distribuição a posteriori para o Modelo 2 é dada por

$$\begin{aligned}\pi(\beta_1,\beta_2,\gamma|\mathbf{y},\mathbf{x},\mathbf{z}) &\propto exp\left\{-\frac{1}{2}\left[\left(\frac{\beta_1-\theta_1}{\sigma_1}\right)^2+\left(\frac{\beta_2-\theta_2}{\sigma_2}\right)^2+\left(\frac{\gamma-\theta_3}{\sigma_3}\right)^2\right]\right\}\\ &\times \prod_{i=1}^{n}\left[exp(y_i z_i\beta_1)[1+exp(z_i\beta_1)]^{-1}\right]\\ &\times \prod_{i=1}^{n}\left[\frac{1}{exp(z_i\beta_2)+\gamma y_i\frac{e^{z_i\beta_1}}{1+e^{z_i\beta_1}}}exp\left\{-\frac{x_i}{exp(z_i\beta_2)+\gamma y_i\frac{e^{z_i\beta_1}}{1+e^{z_i\beta_1}}}\right\}\right].\end{aligned}$$

A distribuição a posteriori para o Modelo 3 é dada por

$$\begin{aligned}\pi(\beta_1,\beta_2,\gamma|\mathbf{y},\mathbf{x},\mathbf{z}) &\propto exp\left\{-\frac{1}{2}\left[\left(\frac{\beta_1-\theta_1}{\sigma_1}\right)^2+\left(\frac{\beta_2-\theta_2}{\sigma_2}\right)^2+\left(\frac{\gamma-\theta_2}{\sigma_2}\right)^2\right]\right\}\\ &\times \prod_{i=1}^{n}\left[exp(y_i z_i\beta_1)[1+exp(z_i\beta_1)]^{-1}\right]\\ &\times \prod_{i=1}^{n}\left[\frac{1}{exp(z_i\beta_2)+\gamma y_i}exp\left\{-\frac{x_i}{exp(z_i\beta_2)+\gamma y_i}\right\}\right].\end{aligned}$$

Em todos os modelos, como a distribuição a posteriori que obtivemos não possui uma forma fechada, aplicamos um procedimento MCMC, cujo algoritmo foi desenvolvido da seguinte maneira:

Passo 1: Inicialize: $\beta_1 = \beta_1^*$; $\beta_2^* = \beta_2^*$ e $\gamma = \gamma^*$

Nos próximos passos executamos o algoritmo para os três parâmetros separadamente, como segue

Passo 2: Geração do passeio aleatório

Gere r seguindo uma distribuição normal com média 0 e variância pequena.

Esta variância será definida de maneira diferente para cada tamanho amostral e para cada modelo, de maneira que consigamos manter a taxa de rejeição entre 0.6 e 0.7. Para valores da taxa de rejeição acima de 0.7 devemos diminuir a variância desta distribuição e para valores da taxa de rejeição abaixo de 0.6 devemos aumentar a variância. Os valores que utilizamos para a variância de r, em cada caso, são apresentados na tabela 3.1

Tabela 3.1: Variâncias das distribuições do passeio aleatório para β_1

Variâncias de r	$m = 50$	$m = 100$	$m = 500$
Modelo 1	1.36	0.87	0.38
Modelo 2	1.35	0.9	0.4
Modelo 3	1.34	0.92	0.4

Com isso, obtenha $\beta_1^{'}$, em que $\beta_1^{'} = \beta_1 + r$.

Como μ_1 é uma função de β_1, ou seja, $\mu_1 = \dfrac{exp(Z_i\beta_1)}{exp(1 + Z_i\beta_1)}$ devemos atualizá-lo como segue

Calcule $\mu_1^{'} = \dfrac{exp(Z_i\beta_1^{'})}{exp(1 + Z_i\beta_1^{'})}$.

Passo 3: Calcule $\alpha = min(1, A_{\beta_1})$, tal que $A_{\beta_1} = \dfrac{\pi(\beta_1^{'}, \beta_2, \gamma|y, x)}{\pi(\beta_1, \beta_2, \gamma|y, x)}$

Passo 4: Gere $u \sim U(0, 1)$

Se $u < \alpha$ então $\beta_1 = \beta_1^{'}$ e $\mu_1 = \mu_1^{'}$.

Se $u \geq \alpha$ então $\beta_1 = \beta_1$ e $\mu_1 = \mu_1$.

Passo 5: Gere s seguindo uma distribuição normal com média 0 e variância pequena, que será definida como no passo 2. Os valores que utilizamos para a variância de s, em cada caso, são apresentados na tabela 3.2

Tabela 3.2: Variâncias das distribuições do passeio aleatório para β_2

Variâncias de s	$m = 50$	$m = 100$	$m = 500$
Modelo 1	0.43	0.31	0.14
Modelo 2	0.6	0.45	0.2
Modelo 3	0.57	0.4	0.19

Dessa forma, obtenha $\beta_2^{'}$, em que $\beta_2^{'} = \beta_2 + s$.

Passo 6: Calcule $\alpha = min(1, A_{\beta_2})$, tal que $A_{\beta_2} = \dfrac{\pi(\beta_1, \beta_2^{'}, \gamma|y, x)}{\pi(\beta_1, \beta_2, \gamma|y, x)}$

Passo 7: Gere $u \sim U(0, 1)$

Se $u < \alpha$ então $\beta_2 = \beta_2^{'}$.

Se $u \geq \alpha$ então $\beta_2 = \beta_2$.

Passo 8: Gere t seguindo uma distribuição normal com média 0 e variância pequena, que será definida como no passo 2. Os valores que utilizamos para a variância de t, em cada caso, são apresentados na tabela 3.3

Tabela 3.3: Variâncias das distribuições do passeio aleatório para γ

Variâncias de t	$m = 50$	$m = 100$	$m = 500$
Modelo 1	0.61	0.45	0.19
Modelo 2	2.5	1.7	0.7
Modelo 3	0.55	0.45	0.18

Assim, obtenha $\gamma^{'}$, em que $\gamma^{'} = \gamma + t$.

Passo 9: Calcule $\alpha = min(1, A_{\gamma})$, tal que $A_{\gamma} = \dfrac{\pi(\beta_1, \beta_2, \gamma^{'}|y, x)}{\pi(\beta_1, \beta_2, \gamma|y, x)}$

Passo 10: Gere $u \sim U(0, 1)$

Se $u < \alpha$ então $\gamma = \gamma^{'}$.

Se $u \geq \alpha$ então $\gamma = \gamma$.

Repita os passos de 2 a 10.

3.6 Análise de Resíduos

Com o objetivo de constatar se o modelo ajusta-se bem aos dados e de detectar possíveis observações outliers, verificando se existe algum valor discrepante que possa influenciar no ajuste da função, realizamos um estudo de diagnósticos para a variável contínua X_i, condicionada à Y_i através de análise de resíduos e de um diagnóstico de influência.

As suposições iniciais consideradas na construção dos três modelos são: a variável aleatória Y segue distribuição Bernoulli, a variável aleatória X, condicionada a Y, segue distribuição exponencial e as observação são independentes. Para verificar a adequação do modelo aos dados realizamos uma análise de resíduos utilizando três tipos de resíduos Bayesianos, resíduo baseado na densidade preditiva condicional ordinária (CPO), resíduo baseado na distribuição a posteriori dos parâmetros

do modelo (Pires & Diniz (2012)) e resíduo deviance Bayesiano (Spiegelhalter *et al.* (2002)). Estes resíduos são descritos a seguir.

3.6.1 Resíduos baseados na densidade preditiva condicional ordinária (CPO)

Para o cálculo dos resíduos baseados na CPO é utilizada a densidade preditiva condicional ordinária que é construída como segue.

Densidade preditiva condicional ordinária

A CPO_i, para a i-ésima observação, é dada por

$$CPO_i = \left[\int_\Theta \frac{1}{\pi(x_i|D,\boldsymbol{\theta})}\pi(\boldsymbol{\theta}|D)d\boldsymbol{\theta}\right]^{-1}, \tag{3.1}$$

em que $D = (y, x, z)$ e $\boldsymbol{\theta} = (\beta_1, \beta_2, \gamma)$.

Encontramos o valor predito para x_i através da densidade preditiva condicional ordinária utilizando o seguinte método numérico:

i Geramos um conjunto de amostras de tamanho Q, $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_Q$, da distribuição a posteriori $\pi(\boldsymbol{\theta}|D)$ e, a partir da qual, determinamos $\hat{\beta}_{1_1}, \ldots, \hat{\beta}_{1_Q}$, $\hat{\beta}_{2_1}, \ldots, \hat{\beta}_{2_Q}$, $\hat{\gamma}_1, \ldots, \hat{\gamma}_Q$.

ii A integral em (3.1) não possui solução analítica. Para resolvê-la geramos um grid de valores na vizinhança de x_i e, para cada observação, retiramos uma amostra desse grid. Para cada $\widetilde{x}_i$ na amostra do grid, obtemos a estimativa de Monte Carlo para a CPO, dada por

$$\widehat{CPO_i} = \left\{\frac{1}{Q}\sum_{q=1}^{Q}\left[\frac{1}{exp(z_i\hat{\beta}_{2_q}) + \hat{\gamma_q}y_i\hat{\mu}_{1_{iq}}}\exp\left\{-\frac{\widetilde{x}_i}{\exp(z_i\hat{\beta}_{2_q}) + \hat{\gamma_q}y_i\hat{\mu}_{1_{iq}}}\right\}\right]^{-1}\right\}^{-1}.$$

iii O valor de $\widetilde{x}_i$ na amostra do grid que maximiza a $\widehat{CPO_i}$ é o valor predito para a variável resposta x_i.

Calculamos o resíduo baseado na CPO como $r_{pp_i} = x_i - \tilde{x}_i$, em que x_i é a i-ésima resposta observada da variável contínua X e $\tilde{x}_i$ é o i-ésimo valor predito. Padronizamos este resíduo da seguinte maneira

$$r_{spp_i} = \frac{r_{pp_i}}{\sqrt{\widehat{Var}(X_i|Y_i)}}\ , i = 1, \ldots, n$$

em que $\widehat{Var}(X_i|Y_i) = \hat{\lambda}_i^2$.

No resíduo baseado na CPO buscamos valores preditos de x_i que maximizam a CPO. Pelo fato de calcularmos o resíduo relacionado a variável $X|Y$, que assume distribuição exponencial, o valor de $\tilde{x}_i$ que maximiza a CPO será sempre o mínimo valor na vizinhança de x_i, como podemos observar no gráfico da Figura 3.1. Por este motivo os valores de x_i não servem para predizer nossos modelos e, portanto, não podemos usar este tipo de resíduo.

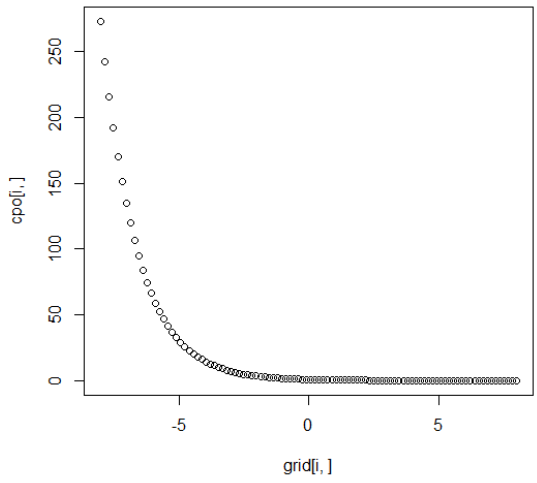

Figura 3.1: Gráfico da CPO versus valores na vizinhança de x_i para o tamanho amostral $n = 100$, do Modelo 2

3.6.2 Resíduos baseados na distribuição a posteriori dos parâmetros do modelo

Calculamos o resíduo padronizado baseado na distribuição a posteriori dos parâmetros do modelo como

$$r_{spd_i} = \frac{x_i - \widehat{E}(X_i|Y_i)}{\sqrt{\widehat{Var}(X_i|Y_i)}} = \frac{x_i - \hat{\lambda}_i}{\hat{\lambda}_i},$$

em que $\widehat{Var}(X_i|Y_i) = \left(\sum_{q=1}^{Q} \hat{\lambda}_{i_q}\right)^2$ e $\widehat{E}(X_i|Y_i) = \sum_{q=1}^{Q} \hat{\lambda}_{i_q}$.

3.6.3 Resíduo Deviance Bayesiano

Calculamos o resíduo deviance Bayesiano como

$$r_{di_i} = \pm\sqrt{D_i + pd_i},$$

em que, $\pm$ significa o sinal de $(x_i - \hat{\lambda_i})$, $D_i = -2log[P(X_i, Y_i; \hat{\beta_1}, \hat{\beta_2}, \hat{\gamma})]$,

$$pd_i = -2\left\{\frac{1}{Q}\sum_{q=1}^{Q}\left[log\frac{\pi(\hat{\beta}_{1_q}, \hat{\beta}_{2_q}, \hat{\gamma}_q|x_i, y_i)}{\pi(\hat{\beta}_{1_q}, \hat{\beta}_{2_q}, \hat{\gamma}_q)}\right] - log\frac{\pi(\hat{\beta_1}, \hat{\beta_2}, \hat{\gamma}|x_i, y_i)}{\pi(\hat{\beta_1}, \hat{\beta_2}, \hat{\gamma})}\right\}$$

e $\widehat{Var}(X_i|Y_i) = \hat{\lambda}^2$

A adequação do modelo ajustado é analisada através de gráficos dos resíduos versus valores esperados, avaliando se os pontos são dispersos de maneira aleatória e em uma faixa horizontal centrada no zero. Os pontos distantes da faixa horizontal, centrada no zero, em alguns desvios, que devem ser considerados de diferentes tamanhos para cada caso, são tidos como outliers.

No resíduo baseado na distribuição a posteriori dos parâmetros do modelo, por ser considerado de forma padronizada, os desvio são obtidos de maneira usual no intervalo $(-2, 2)$, considerando o intervalo de credibilidade de 95%, e são tidos como outliers os pontos que se encontram fora desse intervalo e se destacam em relação aos demais.

Em relação ao resíduo deviance Bayesiano, para que pontos sejam considerados como outliers devemos primeiramente identificar os pontos mais afastados da maioria e determinar os valores destes pontos Este resíduo é calculado como $\sqrt{D_i + pd_i}$, em que $D_i = -2log(densidade)$ e valores que se encontram na cauda da densidade podem fornecer alto valor de D_i e com isso, um alto valor do resíduo. Dessa forma observamos que o alto valor do resíduo não necessariamente indica um outlier e sim um ponto na cauda da densidade, comum em geração de dados exponenciais, uma vez que a variância é igual a média ao quadrado. Fazendo uma analogia com a distribuição normal ou com a teoria de seis sigma, em que o intervalo $(\mu - 3\sigma, \mu + 3\sigma)$ deve conter em torno de 99% dos pontos temos, no caso exponencial, o intervalo $(0, \lambda + 3\lambda)$, ou seja, $(0, 4\lambda)$.

Para o resíduos deviance Bayesiano, então, os desvios a serem tomados como limites para detectar outliers serão considerados em relação aos pontos que apresentarem valores de x_i maiores do que 4λ.

Os pontos discrepantes no modelo devem ser considerados como influentes, ou não, através de um diagnóstico de influência. Nestes casos, nem todo ponto outlier será influente e nem todo ponto influente será outlier.

Os gráficos dos resíduos podem apresentar, também, algumas tendências como uma aparência afunilada, indicando que a variância não é constante, apresentar uma faixa crescente de pontos, que pode significar que um termo linear deveria ser incluído no modelo, ou apresentar uma faixa em forma de parábola, representando que termos lineares e quadráticos deveriam ser incluídos no modelo. Outra informação que obtemos através destes gráficos é que observações distantes da tendência geral pode afetar o próprio ajuste do modelo.

Outro gráfico que podemos utilizar para análise de pontos discrepantes é o gráfico boxplot, que deve ser analisado conforme a variação de seus intervalos, que representa variação nos dados e pela presença de pontos extremos, que indicam pontos outliers.

3.6.4 Diagnóstico de influência

Apresentamos um caso Bayesiano de diagnóstico de influência baseado na divergência de Kullback-Leibler (Pires & Diniz (2012)). Utilizamos a estimativa da divergência $\hat{k}$ para calcularmos uma calibração p_i^* para verificar a influência ou não dos pontos do modelo.

$$\hat{k} = log\left[\frac{1}{Q}\sum_{q=1}^{Q}\frac{1}{f(x_i, y_i; \hat{\beta}_{1_q}, \hat{\beta}_{2_q}, \hat{\gamma}_q)}\right] + \frac{1}{Q}\sum_{q=1}^{Q} log[f(x_i, y_i; \hat{\beta}_{1_q}, \hat{\beta}_{2_q}, \hat{\gamma}_q)]$$

$$p_i^* = 0.5\left[1 + \sqrt{1 - exp\left(-2\hat{k}\right)}\right]$$

Com esse diagnóstico, buscamos analisar a influência de pontos discrepantes com o objetivo de identificar se estes pontos influenciam o nosso modelo. São considerados como pontos influentes aqueles que apresentam valores da calibração muito maiores do que 0.5, nos nossos modelos serão considerados influentes os que apresentarem valor de calibração maior do que 0.7. Uma análise futura para os pontos identificados como outliers e como influentes deve ser feita de modo a verificar a necessidade de retirar ou de manter esses pontos no modelo.

3.6.5 Pontos outliers

As observações que apresentam um grande afastamento das restantes, ou seja, os seus resíduos são muito maiores, em valores absolutos, do que os resíduos das demais, são chamadas de outliers.

Outliers podem ocorrer devido a erros na gravação dos dados. Se estes erros, de alguma forma, puderem ser corrigidos, isto deve ser feito, caso contrário, estas observações devem ser descartadas. Se os dados realmente são outliers, sem causa conhecida, o seu descarte pode ser prejudicial ao ajuste do modelo. Dessa maneira, eles não devem apenas ser ignorados, mas uma análise mais detalhada deve ser feita.

Sendo assim, analisamos estes pontos, primeiramente, verificando se temos muitos outliers, se isso ocorre temos indicativo de falta de ajuste do modelo e outros modelos devem ser considerados. Se não temos um número muito grande de outliers, devemos verificar se as observações discrepantes influenciam o ajuste do modelo. Caso encontramos pontos influentes, estes devem ser investigados. Um ponto é influente se sua exclusão do ajuste da regressão causa uma mudança significativa nos valores ajustados.

O fato de que uma observação fornece um outlier grande não é, evidentemente bom, mas não significa necessariamente que a observação é influente na adequação do modelo escolhido.

Em relação ao ponto influente, devemos retirá-lo dos dados e refazer a análise sem ele. Com isso, comparamos as duas análises, a com o ponto e a sem o ponto, e verificamos a mudança que este novo conjunto ocasiona nas estimativas dos parâmetros em relação as estimativas do conjunto completo.

Capítulo 4

Estudo de Simulação e Análise de Dados Reais

Neste capítulo apresentamos um estudo de simulação com o objetivo de ilustrar a metodologia desenvolvida. O estudo envolve uma simulação de valores gerados para diferentes tamanhos amostrais $n = 50$, $n = 100$ e $n = 500$ a fim de constatar o bom funcionamento dos três modelos desenvolvidos. Utilizamos distribuições a priori não informativas, caracterizadas pela ausência total ou quantidade mínima de informação, neste caso distribuições normais com média zero e variância grande, ou seja, $\beta_1 \sim N(0, 100)$, $\beta_2 \sim N(0, 100)$ e $\gamma \sim N(0, 100)$, para os três modelos considerados. Realizamos uma análise de resíduos Bayesiana considerando os dois tipos de resíduos apresentados no Capítulo 3, resíduo baseados na distribuição a posteriori dos parâmetros do modelo e resíduo deviance Bayesiano, e realizamos um diagnóstico de influência a fim de detectar pontos influentes no modelo. Uma análise de dados reais, considerando como variáveis resposta utilização, ou não, de processo cirúrgico e gastos com pacientes hospitalizados, é elaborada levando em conta os três modelos propostos.

4.1 Estudo de Simulação para o Modelo 1

Para o Modelo 1 geramos 200 amostras de tamanho 130000 para $n = 50$, descartando as primeiras 10000 iterações como "burn-in" e o restante das iterações espaçadas de 30 em 30 e geramos 200 amostras de tamanho 85000 para $n = 100$ e $n = 500$, descartando as primeiras 5000 iterações como "burn-in" e o restante das iterações espaçadas de 20 em 20. Assim, uma amostra final de

tamanho 4000 foi obtida para cada tamanho amostral. A convergência das cadeias foi verificada utilizando os valores dos parâmetros estabelecidos em $\beta_1 = 1.35$, $\beta_2 = -2.60$ e $\gamma = 0.67$. A covariável foi construída gerando valores de uma distribuição uniforme. O valor da covariável foi considerado de forma padronizada.

Simulamos também amostras com os valores dos parâmetros estabelecidos em $\beta_1 = 1.35$, $\beta_2 = 2.60$ e $\gamma = 0.67$ e em $\beta_1 = -1.35$, $\beta_2 = 2.60$ e $\gamma = 0.67$.

4.1.1 Resultados do estudo de simulação para o Modelo 1

Descrevemos a seguir o método utilizado e ilustramos alguns resultados do estudo de simulação para o Modelo 1. Para os três tamanhos amostrais simulados apresentamos tabelas de resumo a posteriori contendo valores de média, mediana, variância, intervalos de credibilidade, intervalos HPD, vício e erro quadrático médio, considerando os valores dos parâmetros estabelecidos em $\beta_1 = 1.35$, $\beta_2 = -2.60$ e $\gamma = 0.67$.

Na Tabela 4.1 apresentamos o resumo a posteriori para cada tamanho amostral $n = 50$, $n = 100$ e $n = 500$.

Tabela 4.1: Medidas descritivas para os parâmetros β_1, β_1 e γ

n = 50	Real	média	mediana	variância	IC		HPD		vício	eqm
β_1	1.35	1.6298	1.6040	0.0787	0.8355	2.5737	0.7937	2.5124	0.2798	0.1570
β_2	-2.60	-2.5965	-2.5934	0.0265	-2.9117	-2.2985	-2.9055	-2.2954	0.0035	0.0265
γ	0.67	0.6357	0.6291	0.0467	0.2463	1.0634	0.2360	1.0482	-0.0343	0.0479
n = 100	Real	média	mediana	variância	IC		HPD		vício	eqm
β_1	1.35	1.2191	1.2096	0.0172	0.7212	1.7717	0.7072	1.7501	-0.1309	0.0344
β_2	-2.60	-2.5966	-2.5954	0.0110	-2.8096	-2.3902	-2.8065	-2.3890	0.0034	0.0110
γ	0.67	0.6750	0.6714	0.0198	0.3935	0.9762	0.3880	0.9674	0.0050	0.0198
n = 500	Real	média	mediana	variância	IC		HPD		vício	eqm
β_1	1.35	1.2206	1.2186	0.0168	0.9942	1.4579	0.9917	1.4532	-0.1294	0.0336
β_2	-2.60	-2.6036	-2.6034	0.0022	-2.6987	-2.5104	-2.6977	-2.5103	-0.0036	0.0022
γ	0.67	0.6750	0.6744	0.0049	0.5439	0.8099	0.5431	0.8078	0.0050	0.0049

Pela tabela de resumo a posteriori verificamos que os valores estimados estão próximos dos valores reais. Notamos pouca variação das estimativas e que estas estão contidas nos intervalos de credibilidade. Neste modelo, observamos que quanto maior o valor de n mais o valor da média se aproxima do valor real e menores são os intervalos de credibilidade e HPD.

Os histogramas dos três parâmetros do Modelo 1, $\beta 1$, $\beta 2$ e γ, foram analisados para os tamanhos amostrais $n = 50$, $n = 100$ e $n = 500$, respectivamente, apresentando simetria em torno da média

das estimativas que também pode ser percebida pela proximidade dos valores de média e mediana dos parâmetros.

Além destes resultados, realizamos teste de convergência de Geweke, obtivemos gráficos de convergência e gráficos de autocorrelação, que não foram apresentados aqui, porém apresentaram resultados satisfatórios.

4.1.2 Resultados da análise de resíduos para o Modelo 1

Para o Modelo 1 realizamos a análise de resíduos para os dois tipos de resíduos, resíduos baseados na distribuição a posteriori dos parâmetros do modelo e resíduos deviance Bayesianos. Apresentamos apenas os resultados obtidos para resíduos baseados na posteriori.

Para o cálculo dos resíduos deviance Bayesianos, neste modelo, é necessário admitir como restrições aos parâmetros $\beta_2 > 0$, $0 < \gamma < 1$ e a covariável deve ser assumir valores positivos.

A análise dos resíduos e o diagnóstico de influência foram ilustrados através do gráfico dos resíduos baseados na distribuição a posteriori dos parâmetros versus valores esperados e do boxplot dos resíduos baseados na distribuição a posteriori dos parâmetros, mostrados na Figura 4.1, e do gráfico de calibração, mostrado na Figura 4.2, respectivamente.

O estudo de diagnóstico foi realizado para os três tamanhos amostrais considerados no processo de estimação, porém apresentamos aqui apenas os resultados para o tamanho amostral $n = 100$, devido aos três tamanhos apresentarem as mesmas conclusões.

Na Figura 4.1 apresentamos o gráfico dos resíduos baseados na distribuição a posteriori dos parâmetros e o boxplot das amostras MCMC da distribuição a posteriori para os resíduos baseados na distribuição a posteriori dos parâmetros do modelo, respectivamente, que têm por objetivo verificar a presença de pontos discrepantes, investigando se são possíveis outliers. O gráfico dos resíduos mostra a dispersão dos pontos em torno de zero. Pequenos intervalos no boxplot indicam pequena variação nos dados e pontos dispersos dos demais podem indicar pontos outliers.

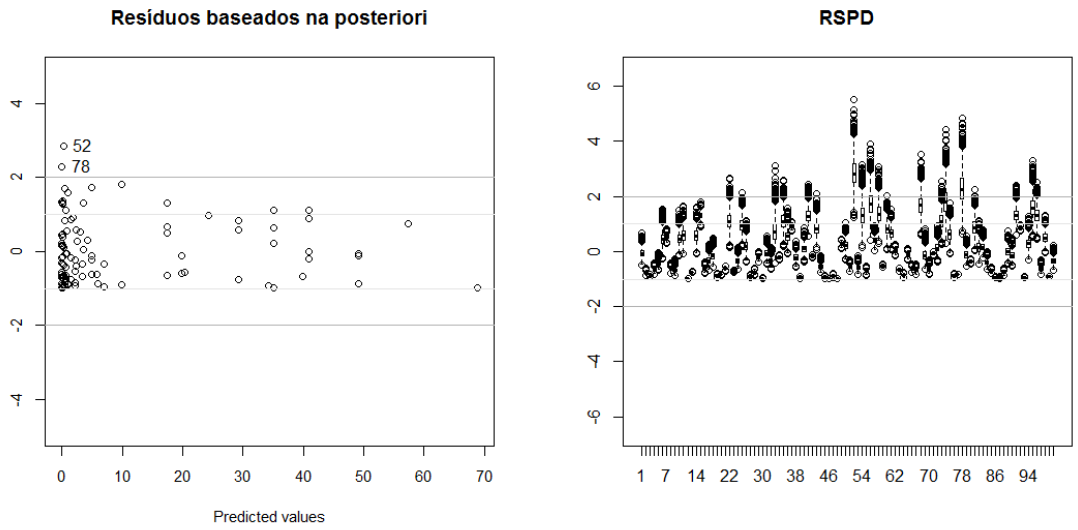

Figura 4.1: Gráfico dos resíduos baseados na distribuição a posteriori dos parâmetros versus valores esperados e Boxplot das amostras MCMC da distribuição a posteriori.

Pelo gráfico dos resíduos baseados na distribuição a posteriori dos parâmetros, considerando o intervalo com 95% de credibilidade, observamos a presença de dois pontos extremos, fora da faixa de intervalo $(-2, 2)$, o caso 52 e o caso 78, que se destacam em relação aos demais e por este motivo são considerados como outliers.

Pelo boxplot vemos que a maioria dos pontos apresenta pequenos intervalos, o que mostra pequena variação das estimativas, porém dois pontos apresentam intervalos maiores, o caso 52 e o caso 78, o que também nos leva a identificá-los como outlier.

Na Figura 4.2 encontramos os resultados para a calibração obtida como descrita no Capitulo 3, os quais utilizamos para detectar pontos influentes.

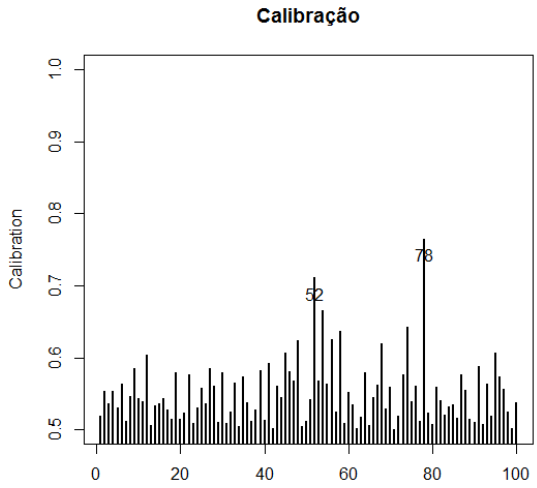

Figura 4.2: Gráfico da calibração p^*

Através dos resultados do valor da calibração são considerados pontos influentes aqueles que apresentarem valor da calibração muito maior do que 0.5. Pelo gráfico da Figura 4.2 observamos que os casos 52 e 78, detectados também como outliers neste caso, são pontos influentes no modelo. Com isso retiramos estes pontos do modelo e refazemos o ajuste para verificar o impacto que eles causa no modelo.

Na Tabela 4.2 apresentamos os valores dos parâmetros para os modelos com e sem os pontos. A tabela mostra o impacto causado pelos casos 52 e 78 na estimação dos parâmetros. É possível observar que a retirada individual dos pontos não retorna impactos, apenas quando retiramos os dois pontos conjuntamente observamos uma mudança um pouco maior no parâmetro γ.

Tabela 4.2: Mudança relativa da retirada do ponto do modelo

Mudança para retirada do ponto 52			
Parâmetros	com o ponto	sem o ponto 52	mudança relativa
β_1	1.2191	1.2108	-0.68%
β_2	-2.5966	-2.5946	-0.07%
γ	0.6750	0.6119	-9.35%
Mudança para retirada do ponto 78			
Parâmetros	com o ponto	sem o ponto 78	mudança relativa
β_1	1.2191	1.2060	-1.07%
β_2	-2.5966	-2.6161	0.75%
γ	0.6750	0.6386	-5.39%
Mudança para retirada dos pontos 52 e 78			
Parâmetros	com o ponto	sem os pontos 52 e 78	mudança relativa
β_1	1.2191	1.1956	-1.92%
β_2	-2.5966	-2.6264	1.15%
γ	0.6750	0.5779	-14.39%

Pela Tabela 4.2 não observamos mudanças significativas na retirada dos pontos, porém pelos gráficos dos resíduos sem os pontos notamos que a retirada de tais pontos leva ao não aparecimento de pontos outliers e de autos valores da calibração.

Nas Figuras 4.3, 4.4 e 4.5 ilustramos os gráficos para o resíduo baseado na distribuição a posteriori dos parâmetros e os gráficos da calibração, considerando os dados sem o caso 52, sem o caso 78 e sem os casos 52 e 78, respectivamente.

Pelos gráficos das Figuras 4.3, 4.4 e 4.5 e pela tabela de mudança relativa das estimativas observamos que a retirada dos pontos detectados como influentes não levou ao aparecimento de pontos outliers e nem a altos valores de calibração para nenhum dos demais pontos. Com isso notamos que, apesar destes pontos serem considerados influentes, eles não causam grandes impactos no modelo.

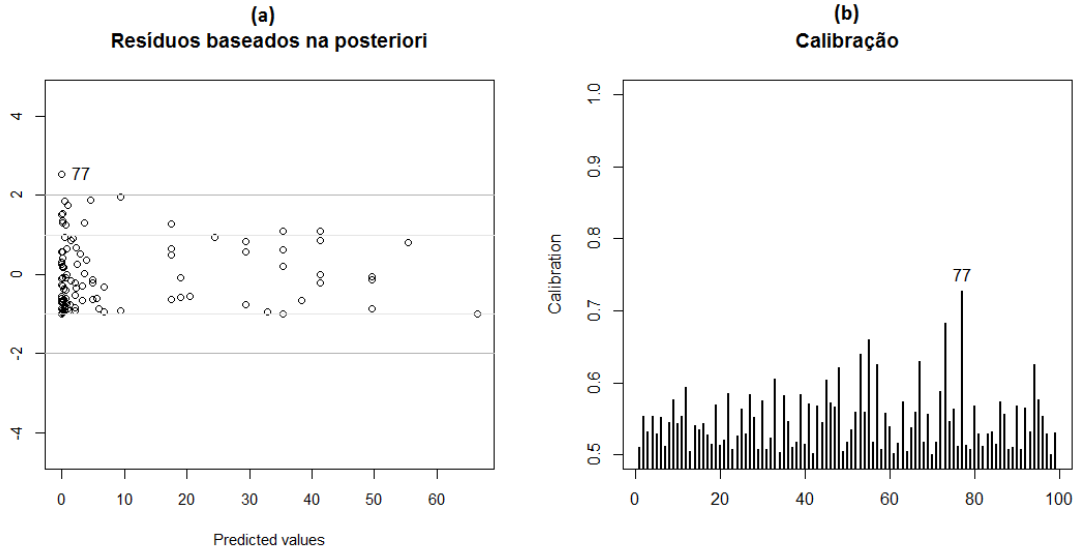

Figura 4.3: (a) Gráfico dos resíduos baseados na distribuição a posteriori dos parâmetros versus valores esperados; (b) Gráfico da calibração. Ambos para os dados sem o caso 52.

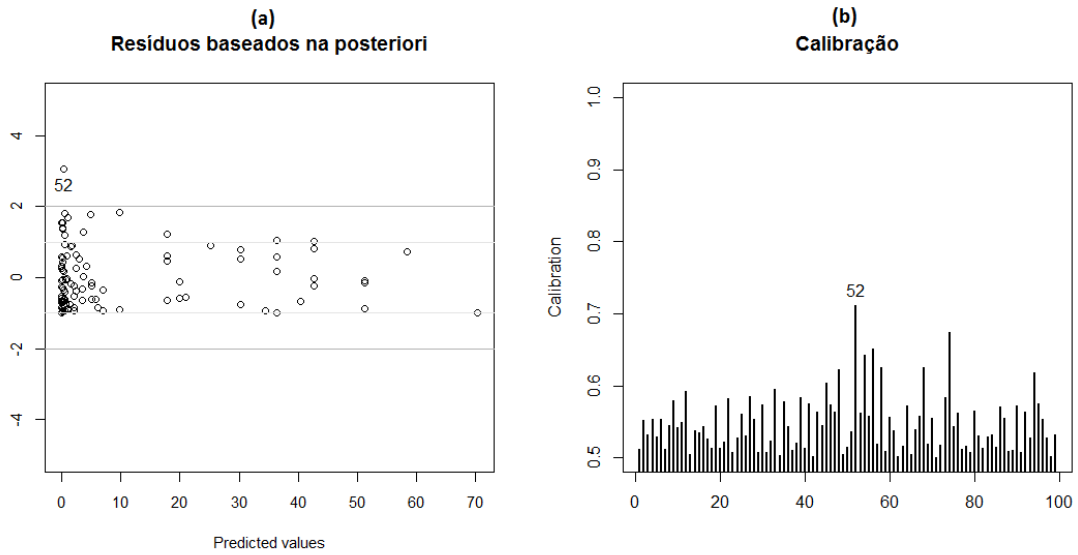

Figura 4.4: (a) Gráfico dos resíduos baseados na distribuição a posteriori dos parâmetros versus valores esperados; (b) Gráfico da calibração. Ambos para os dados sem o caso 78.

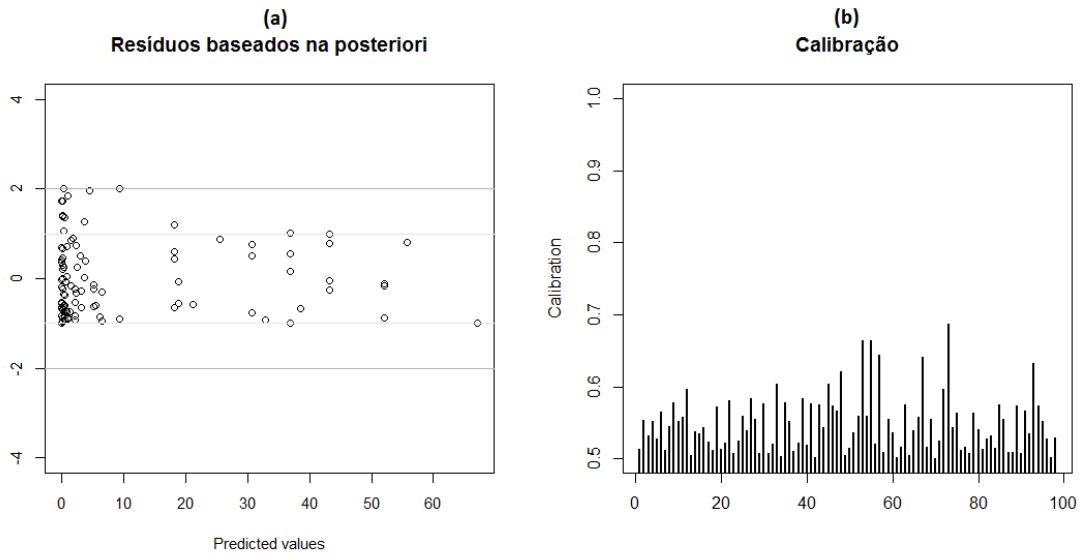

Figura 4.5: (a) Gráfico dos resíduos baseados na distribuição a posteriori dos parâmetros versus valores esperados; (b) Gráfico da calibração. Ambos para os dados sem o caso 52 e sem o caso 78.

Outros resultados

Realizamos o mesmo procedimento apresentado anteriormente considerando, também, os valores dos parâmetros fixados em $\beta_1 = 1.35$, $\beta_2 = 2.60$ e $\gamma = 0.67$ e em $\beta_1 = -1.35$, $\beta_2 = 2.60$ e $\gamma = 0.67$. Apresentamos como resultado as tabelas de resumo a posteriori para os três tamanhos amostrais $n = 50$, $n = 100$ e $n = 500$.

Considerando $\beta_1 = 1.35$, $\beta_2 = 2.60$ e $\gamma = 0.67$, temos os seguintes resultados.

Tabela 4.3: Medidas descritivas para os parâmetros β_1, β_2 e γ

n = 50	Real	média	mediana	variância	IC		HPD		vício	eqm
β_1	1.35	1.3177	1.3004	0.0011	0.6158	2.1197	0.5838	2.0735	-0.0323	0.0021
β_2	2.60	2.5892	2.5919	0.0265	2.2724	2.8905	2.2792	2.8942	-0.0108	0.0266
γ	0.67	0.6799	0.6723	0.0459	0.2605	1.1429	0.2484	1.1246	0.0099	0.0460
n = 100	Real	média	mediana	variância	IC		HPD		vício	eqm
β_1	1.35	1.6096	1.5962	0.0678	1.0362	2.2589	1.0144	2.2271	0.2596	0.1351
β_2	2.60	2.5882	2.5903	0.0124	2.3630	2.8017	2.3682	2.8047	-0.0118	0.0125
γ	0.67	0.6826	0.6792	0.0220	0.3851	1.0001	0.3797	0.9912	0.0126	0.0222
n = 500	Real	média	mediana	variância	IC		HPD		vício	eqm
β_1	1.35	1.5316	1.5292	0.0331	1.2762	1.8006	1.2728	1.7945	0.1816	0.0661
β_2	2.60	2.5950	2.5954	0.0024	2.4965	2.6914	2.4977	2.6917	-0.0050	0.0025
γ	0.67	0.6741	0.6733	0.0055	0.5335	0.8189	0.5327	0.8169	0.0041	0.0055

Considerando $\beta_1 = -1.35$, $\beta_2 = 2.60$ e $\gamma = 0.67$, temos os seguintes resultados.

Tabela 4.4: Medidas descritivas

n = 50	Real	média	mediana	variância	IC		HPD		vício	eqm
β_1	-1.35	-1.8717	-1.8360	0.2736	-2.9683	-0.9796	-2.8887	-0.9268	-0.5217	0.5458
β_2	2.60	2.5957	2.5925	0.0253	2.2917	2.9166	2.2877	2.9091	-0.0043	0.0253
γ	0.67	0.6482	0.6409	0.0493	0.2265	1.1120	0.2150	1.0944	-0.0217	0.0498
n = 100	Real	média	mediana	variância	IC		HPD		vício	eqm
β_1	-1.35	-1.9481	-1.9288	0.3596	-2.7250	-1.2807	-2.6785	-1.2489	-0.5981	0.7173
β_2	2.60	2.6110	2.6088	0.0154	2.3910	2.8428	2.3885	2.8380	0.0110	0.0156
γ	0.67	0.6795	0.6758	0.0305	0.3546	1.0258	0.3494	1.0172	0.0095	0.0306
n = 500	Real	média	mediana	variância	IC		HPD		vício	eqm
β_1	-1.35	-1.5383	-1.5356	0.0356	-1.8160	-1.2758	-1.8093	-1.2720	-0.1883	0.0711
β_2	2.60	2.6037	2.6034	0.0029	2.5083	2.7010	2.5078	2.6995	0.0037	0.0029
γ	0.67	0.6650	0.6642	0.0047	0.5294	0.8048	0.5287	0.8029	-0.0050	0.0048

Pelas Tabelas 4.3 e 4.4 observamos que para esses valores dos parâmetros, os valores das médias e das medianas estão próximos dos valores reais e que as estimativas estão contidas nos intervalos de credibilidade e possuem pouca variância. Por estes resultados notamos o bom funcionamento do modelo para outros valores dos parâmetros.

4.2 Estudo de Simulação para o Modelo 2

Geramos, para o Modelo 2, 200 amostras de tamanho 130000 para $n = 50$, descartando as primeiras 10000 iterações como "burn-in" e o restante das iterações espaçadas de 30 em 30 e geramos 200 amostras de tamanho 85000 para $n = 100$ e $n = 500$, descartando as primeiras 5000 iterações como "burn-in" e o restante das iterações espaçadas de 20 em 20. Com isso, uma amostra final de tamanho 4000 foi obtida para cada tamanho amostral. A convergência das cadeias foi verificada utilizando os valores dos parâmetros estabelecidos em $\beta_1 = 1.35$, $\beta_2 = -0.63$ e $\gamma = 2.61$. A covariável foi construída gerando valores de uma distribuição uniforme. O valor da covariável foi considerado de forma padronizada.

Simulamos também amostras com os valores dos parâmetros estabelecidos em $\beta_1 = 1.35$, $\beta_2 = 0.63$ e $\gamma = 2.61$ e em $\beta_1 = -1.35$, $\beta_2 = 0.63$ e $\gamma = 2.61$.

4.2.1 Resultados do estudo de simulação para o Modelo 2

O método utilizado para o Modelo 2 é descrito e alguns resultados do estudo de simulação são ilustrados. Para os três tamanhos amostrais simulados apresentamos tabelas de resumo a posteriori, considerando os valores dos parâmetros estabelecidos em $\beta_1 = 1.35$, $\beta_2 = -0.63$ e $\gamma = 2.61$.

Na Tabela 4.5 apresentamos o resumo a posteriori para os tamanhos amostrais considerados.

Tabela 4.5: Medidas descritivas para os parâmetros β_1, β_2 e γ

n = 50	Real	média	mediana	variância	IC		HPD		vício	eqm
β_1	1.35	1.6287	1.5985	0.0839	0.8116	2.6199	0.7648	2.5503	0.2788	0.1616
β_2	-0.63	-0.6569	-0.6512	0.0311	-1.0420	-0.3040	-1.0294	-0.2960	-0.0269	0.0318
γ	2.61	2.9835	2.8508	0.8398	1.5732	5.1579	1.4013	4.8213	0.3735	0.9793
n = 100	Real	média	mediana	variância	IC		HPD		vício	eqm
β_1	1.35	1.5819	1.5684	0.0562	1.0188	2.2205	0.9958	2.1877	0.2319	0.1100
β_2	-0.63	-0.6167	-0.6138	0.0205	-0.9043	-0.3456	-0.8965	-0.3408	0.0133	0.0207
γ	2.61	2.7001	2.6373	0.3690	1.7198	4.0419	1.6253	3.8865	0.0901	0.3771
n = 500	Real	média	mediana	variância	IC		HPD		vício	eqm
β_1	1.35	1.4788	1.4762	0.0172	1.2316	1.7404	1.2274	1.7336	0.1288	0.0338
β_2	-0.63	-0.6211	-0.6205	0.0034	-0.7420	-0.5038	-0.7406	-0.5035	0.0089	0.0035
γ	2.61	2.5515	2.5415	0.0549	2.1302	3.0290	2.1141	3.0058	-0.0585	0.0583

Pela tabela de resumo a posteriori observamos que os valores das médias se aproximam mais dos valores reais quando o valor de n aumenta. Podemos observar pouca variação das estimativas e que estas estão contidas nos intervalos de credibilidade. Verificamos que estes intervalos são menores

para valores de n maiores.

Os histogramas dos três parâmetros do Modelo 2 foram analisados e verificamos que apresentam simetria em torno da média para tamanho amostral maior, ocorrendo um pouco de assimetria quando o tamanho amostral é menor.

Foram obtidos, também, gráficos de convergência e de autocorrelação e foi realizado um teste de convergência de Geweke. Estes resultados não foram mostrados aqui, porém foram satisfatórios.

4.2.2 Resultados da análise de resíduos para o Modelo 2

A análise dos resíduos e o diagnóstico de influência foram ilustrados através dos gráficos dos resíduos versus valores esperados para os dois tipos de resíduos apresentados, resíduos baseados na distribuição a posteriori dos parâmetros e resíduos deviance Bayesianos, boxplot dos resíduos baseados na posteriori dos parâmetros e gráfico de calibração, mostrados nas Figuras 4.6, 4.7 e 4.8, respectivamente.

O estudo de diagnóstico foi realizado para os três tamanhos amostrais considerados no processo de estimação, porém apresentamos aqui apenas os resultados para o tamanho amostral $n = 100$, devido aos três tamanhos apresentarem as mesmas conclusões.

Na Figura 4.6 apresentamos o gráfico dos resíduos baseados na distribuição a posteriori dos parâmetros do modelo e o gráfico dos resíduos deviance Bayesianos, respectivamente, que têm por objetivo verificar a presença de pontos discrepantes, investigando se são possíveis outliers. Estes gráficos mostram a dispersão dos pontos em torno de zero.

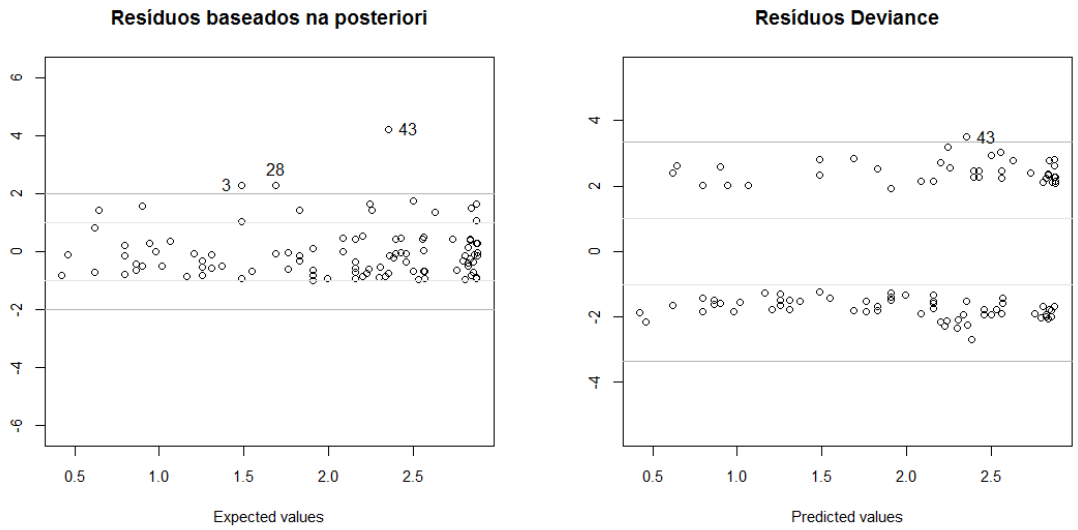

Figura 4.6: Gráfico dos resíduos baseados na distribuição a posteriori dos parâmetros versus valores esperados e gráfico dos resíduos deviance Bayesiano versus valores esperados.

A Figura 4.7 apresenta o boxplot das amostras MCMC da distribuição a posteriori para os resíduos baseados na distribuição a posteriori dos parâmetros do modelo. No Boxplot, pequenos intervalos indicam pequena variação nos dados e pontos com grande dispersão podem indicar pontos outliers.

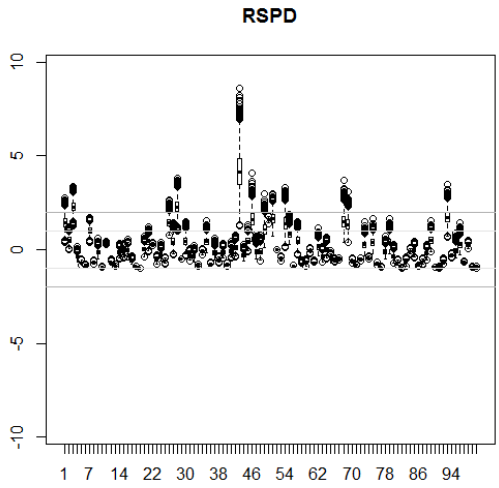

Figura 4.7: Boxplot das amostras MCMC da distribuição a posteriori

Pelo gráfico dos resíduos baseados na distribuição a posteriori dos parâmetros, considerando o intervalo com 95% de credibilidade, observamos a presença de alguns pontos extremos, fora da faixa de intervalo $(-2, 2)$, mas apenas um ponto, o caso 43, se destaca em relação aos demais e, por este motivo, o consideramos como outlier.

Pelo gráfico dos resíduos deviance Bayesianos, considerando os pontos cujos valores são menores do que quatro vezes o desvio padrão (λ), como descrito anteriormente, observamos que o caso 43 apresenta valor maior do que 4λ e seu resíduo encontra-se fora da faixa estabelecida em relação aos demais valores. Sendo assim, consideramos o caso 43 como outlier, também para este tipo de resíduo.

Pelo boxplot da Figura 4.7 vemos que a maioria dos pontos apresenta pequenos intervalos, o que mostra pequena variação das estimativas, porém um ponto apresenta intervalo maior, o caso 43, o que também nos leva a identificá-lo como outlier.

Na Figura 4.8 encontramos os resultados para a calibração obtida como descrita no Capitulo 3, pelos quais observamos presença, ou não, de ponto influente.

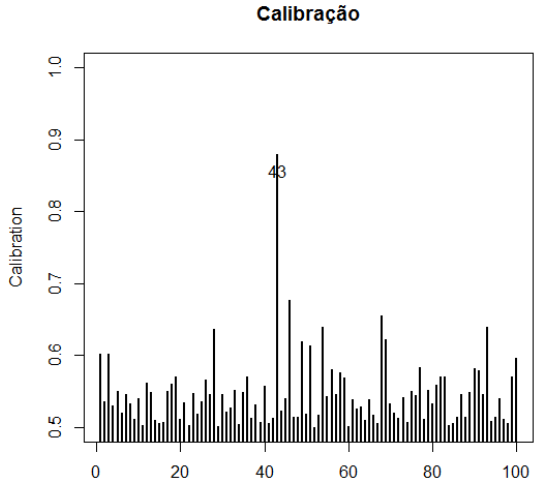

Figura 4.8: Gráfico da calibração p^*

Através dos resultados do valor da calibração são considerados pontos influentes aqueles que apresentarem valor da calibração muito maior do que 0.5. Pelo gráfico da Figura 4.8 observamos que apenas o caso 43, detectado também como outlier neste caso, é um ponto influente no modelo, por apresentar valor de calibração muito maior do que para os demais pontos. Com isso retiramos este ponto do modelo e refazemos o ajuste para verificar o impacto que ele causa no modelo.

Na Tabela 4.6 apresentamos os valores dos parâmetros para os modelos com e sem o ponto. A tabela mostra o impacto causado pelo caso 43 na estimação dos parâmetros.

Tabela 4.6: Mudança relativa da retirada do ponto do modelo

Parâmetros	com o ponto	sem o ponto	mudança relativa
β_1	1.5819	1.5747	-0.46%
β_2	-0.6167	-0.4359	-29.31%
γ	2.7001	3.3679	24.74%

Pela Tabela 4.6 observamos que a estimativa do parâmetro β_1 não sofre impacto com a retirada do ponto, mas para os parâmetros β_2 e γ há uma mudança um pouco mais significativa com a retirada do ponto. Pelo gráfico dos resíduos baseados na distribuição a posteriori dos parâmetros e pelo gráfico dos resíduos deviance Bayesianos observamos que a retirada do caso 43 não levou

ao aparecimento de outros pontos outliers e pelo gráfico da calibração observamos que a retirada do caso 43 não levou a altos valores de calibração para nenhum dos demais pontos. Com isso percebemos que torna-se interessante a retirada deste ponto do Modelo 2.

Na Figura 4.25 ilustramos os gráficos para os dois tipos de resíduos e na Figura 4.26 ilustramos o gráfico da calibração, ambas considerando os dados sem o caso 43.

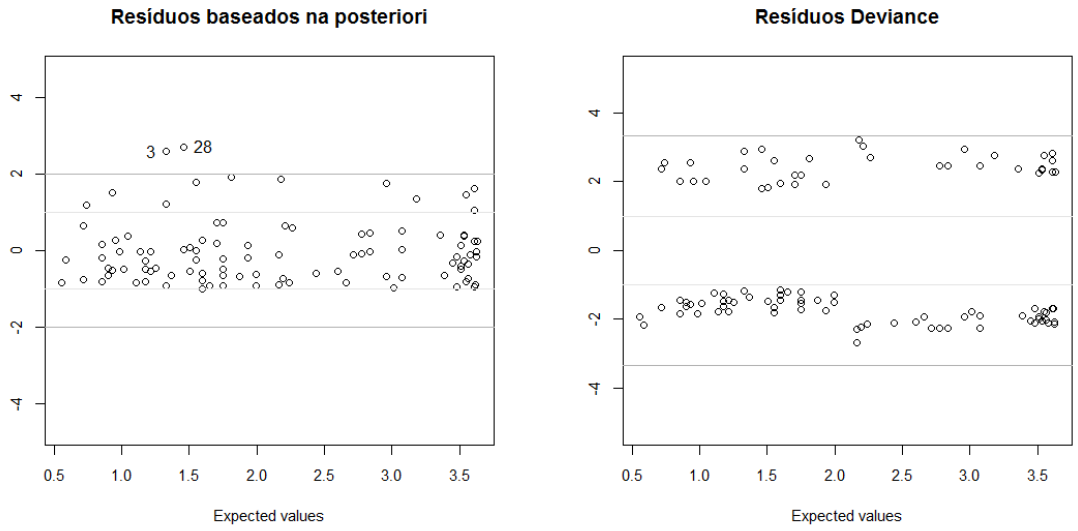

Figura 4.9: Gráfico dos resíduos baseados na distribuição a posteriori dos parâmetros versus valores esperados e gráfico dos resíduos deviance Bayesianos versus valores esperados, para os dados sem o caso 43

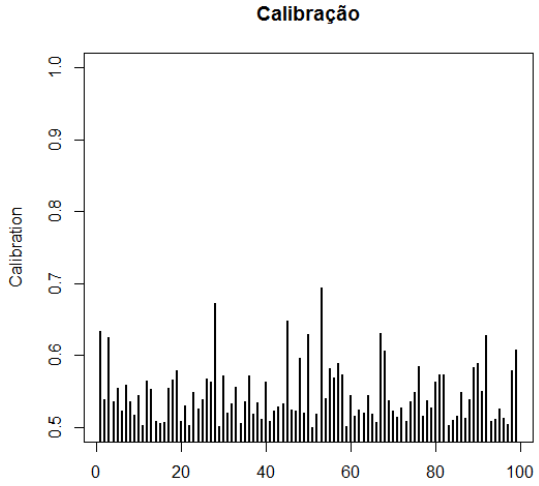

Figura 4.10: Gráfico da calibração p^*, para os dados sem o caso 43

Também, com a retirada do caso 43, observamos nos resultados que nenhum ponto apresentou valor de x_i maior do que 4λ.

Outros resultados

Os estudo de simulação para o Modelo 2 foi realizado considerando, também, os valores dos parâmetro fixados em $\beta_1 = 1.35$, $\beta_2 = 0.63$ e $\gamma = 2.61$ e $\beta_1 = -1.35$, $\beta_2 = 0.63$ e $\gamma = 2.61$, para os tamanhos amostrais $n = 50$, $n = 100$ e $n = 500$, a fim de verificarmos o bom funcionamento do modelo para outros valores dos parâmetros.

Considerando $\beta_1 = 1.35$, $\beta_2 = 0.63$ e $\gamma = 2.61$, obtivemos os seguintes resultados.

Tabela 4.7: Medidas descritivas para os parâmetros β_1, β_2 e γ

n = 50	Real	média	mediana	variância	IC		HPD		vício	eqm
β_1	1.35	1.5914	1.5651	0.0762	0.8209	2.5128	0.7784	2.4469	0.2414	0.1345
β_2	0.63	0.5811	0.5900	0.0406	0.1505	0.9643	0.1718	0.9770	-0.0489	0.0430
γ	2.61	3.1513	3.0373	0.9534	1.5128	5.4352	1.3624	5.1696	0.5413	1.2464
n = 100	Real	média	mediana	variância	IC		HPD		vício	eqm
β_1	1.35	1.3899	1.3790	0.0104	0.8845	1.9561	0.8663	1.9291	0.0399	0.0120
β_2	0.63	0.6213	0.6249	0.0179	0.3474	0.8754	0.3566	0.8813	-0.0087	0.0179
γ	2.61	2.8652	2.8074	0.5015	1.6754	4.3855	1.5914	4.2514	0.2552	0.5666
n = 500	Real	média	mediana	variância	IC		HPD		vício	eqm
β_1	1.35	1.5314	1.5290	0.0345	1.2831	1.7932	1.2792	1.7868	0.1814	0.0674
β_2	0.63	0.6285	0.6291	0.0031	0.5108	0.7428	0.5127	0.7434	-0.0015	0.0031
γ	2.61	2.6614	2.6489	0.1166	2.0623	3.3330	2.0424	3.3034	0.0514	0.1193

Considerando $\beta_1 = -1.35$, $\beta_2 = 0.63$ e $\gamma = 2.61$, temos os seguintes resultados.

Tabela 4.8: Medidas descritivas para os parâmetros β_1, β_2 e γ

n = 50	Real	média	mediana	variância	IC		HPD		vício	eqm
β_1	-1.35	-1.3874	-1.3652	0.0098	-2.2371	-0.6662	-2.1820	-0.6357	-0.0374	0.0112
β_2	0.63	0.6245	0.6162	0.0450	0.2047	1.0920	0.1919	1.0710	-0.0055	0.0450
γ	2.61	2.9939	2.8933	0.6284	1.7093	4.8467	1.5848	4.6076	0.3839	0.7758
$n = 100$	Real	média	mediana	variância	IC		HPD		vício	eqm
β_1	-1.35	-1.3527	-1.3412	0.0025	-1.9430	-0.8296	-1.9143	-0.8098	-0.0027	0.0025
β_2	0.63	0.6252	0.6220	0.0230	0.3603	0.9074	0.3559	0.9001	-0.0048	0.0230
γ	2.61	2.8528	2.7902	0.3321	1.8390	4.2207	1.7481	4.0731	0.2428	0.3911
$n = 500$	Real	média	mediana	variância	IC		HPD		vício	eqm
β_1	-1.35	-1.3950	-1.3924	0.0024	-1.6492	-1.1543	-1.6451	-1.1527	-0.0450	0.0044
β_2	0.63	0.6324	0.6317	0.0030	0.5180	0.7400	0.5174	0.7482	0.0023	0.0030
γ	2.61	2.6423	2.6305	0.0516	2.1862	3.1648	2.1668	3.1369	0.0323	0.0526

Por estes resultados observamos que os valores das médias e das medianas estão próximos dos valores reais, que há pouca variação das estimativas e que estas estão contidas nos intervalos de credibilidade. Pelas tabelas acima verificamos um bom funcionamento do modelo, também para outros valores dos parâmetros.

Constatamos, assim, a ocorrência destes mesmos resultados para três valores fixados de parâmetro. Com isso, verificamos o bom funcionamento do modelo para diferentes valores de parâmetros.

4.3 Estudo de Simulação para o Modelo 3

No Modelo 3 foram geradas 200 amostras de tamanho 130000 para $n = 50$, descartando as primeiras 10000 iterações como "burn-in" e o restante das iterações espaçadas de 30 em 30 e geramos 200 amostras de tamanho 85000 para $n = 100$ e $n = 500$, descartando as primeiras 5000 iterações como "burn-in" e o restante das iterações espaçadas de 20 em 20. Uma amostra final de tamanho 4000 foi obtida para cada tamanho amostral. A convergência das cadeias foi verificada utilizando os valores dos parâmetros estabelecidos em $\beta_1 = 1.35$, $\beta_2 = -2.60$ e $\gamma = 0.67$. A covariável foi construída gerando valores de uma distribuição uniforme e o valor da covariável foi considerado de forma padronizada.

Aamostras com os valores dos parâmetros estabelecidos em $\beta_1 = 1.35$, $\beta_2 = 2.60$ e $\gamma = 0.67$ e em $\beta_1 = -1.35$, $\beta_2 = 2.60$ e $\gamma = 0.67$ também foram simuladas.

4.3.1 Resultados do estudo de simulação para o Modelo 3

A seguir descrevemos o método utilizado e ilustramos alguns resultados do estudo de simulação para o Modelo 3. Para os três tamanhos amostrais simulados apresentamos tabelas de resumo a posteriori, considerando os valores dos parâmetros estabelecidos em $\beta_1 = 1.35$, $\beta_2 = -2.60$ e $\gamma = 0.67$.

Na Tabela 4.9 apresentamos o resumo a posteriori para os tamanhos amostrais $n = 50$, $n = 100$ e $n = 500$, respectivamente.

Tabela 4.9: Medidas descritivas para os parâmetros β_1, β_2 e γ

n = 50	Real	média	mediana	variância	IC		HPD		vício	eqm
β_1	1.35	1.1602	1.1404	0.0363	0.4653	1.9648	0.4348	1.9198	-0.1898	0.0723
β_2	-2.60	-2.5907	-2.5867	0.0328	-2.9647	-2.2386	-2.9554	-2.2334	0.0093	0.0329
γ	0.67	0.7731	0.7341	0.0608	0.4156	1.3554	0.3712	1.2564	0.1031	0.0715
n = 100	Real	média	mediana	variância	IC		HPD		vício	eqm
β_1	1.35	1.3916	1.3806	0.0018	0.8641	1.9789	0.8488	1.9563	0.0416	0.0035
β_2	-2.60	-2.5931	-2.5914	0.0200	-2.8579	-2.3386	-2.8523	-2.3358	0.0069	0.0201
γ	0.67	0.7031	0.6861	0.0205	0.4650	1.0389	0.4411	0.9972	0.0331	0.0216
n = 500	Real	média	mediana	variância	IC		HPD		vício	eqm
β_1	1.35	1.4104	1.4080	0.0037	1.1629	1.6716	1.1589	1.6650	0.0604	0.0073
β_2	-2.60	-2.5987	-2.5980	0.0038	-2.7196	-2.4817	-2.7177	-2.4808	0.0013	0.0038
γ	0.67	0.6774	0.6746	0.0034	0.5722	0.7984	0.5677	0.7918	0.0074	0.0035

Pelas tabelas de resumo a posteriori verificamos que os valores estimados estão próximos dos

valores reais, principalmente para amostras maiores, notamos pouca variação das estimativas e que estas estão contidas nos intervalos de credibilidade. Neste modelo, para valores maiores de n notamos que os valores das médias e das medianas estão mais próximos dos valores reais e que os intervalos de credibilidade são menores.

Construímos os histogramas dos três parâmetros do Modelo 3, $\beta 1$, $\beta 2$ e γ, para os tamanhos amostrais $n = 50$, $n = 100$ e $n = 500$. Estes histogramas não são apresentados aqui. Por eles verificamos uma pequena assimetria no parâmetro γ em torno da média, em especial para tamanhos amostrais menores. Para tamanho amostral maior e para os demais parâmetros notamos simetria em torno da média.

Além dos resultados apresentados nas tabelas de resumo a posteriori, realizamos um teste Geweke, obtivemos gráficos de convergência e gráficos de autocorrelação, que não foram apresentados aqui, porém foram resultados satisfatórios.

4.3.2 Resultados da análise de resíduos para o Modelo 3

Para o Modelo 3, assim como no Modelo 1, realizamos a análise de resíduos para os dois tipos de resíduos descritos no capítulo anterior e apresentamos apenas os resultados obtidos para resíduos baseados na posteriori. Para o cálculo dos resíduos deviance Bayesianos, também neste modelo, é necessário admitir como restrições aos parâmetros $\beta_2 > 0$, $0 < \gamma < 1$ e a covariável deve ser assumir valores positivos.

A análise dos resíduos e o diagnóstico de influência foram ilustrados através do gráfico dos resíduos baseados na distribuição a posteriori dos parâmetros versus valores esperados e boxplot dos resíduos baseados na distribuição a posteriori dos parâmetros, ilustrados na Figura 4.11 e gráfico de calibração ilustrado na Figura 4.12.

O estudo de diagnóstico foi realizado para os três tamanhos amostrais considerados no processo de estimação, porém apresentamos aqui apenas os resultados para o tamanho amostral $n = 100$, devido aos três tamanhos apresentarem mesmas conclusões.

Na Figura 4.11 apresentamos o gráfico dos resíduos baseados na distribuição a posteriori dos parâmetros do modelo e o boxplot das amostras MCMC da distribuição a posteriori para os resíduos baseados na distribuição a posteriori dos parâmetros do modelo, respectivamente, que têm por objetivo verificar a presença de pontos discrepantes, investigando se são possíveis outliers. O gráfico dos resíduos mostra a dispersão dos pontos em torno de zero e no boxplot pequenos intervalos indicam pequena variação nos dados e pontos com grandes intervalos podem indicar pontos outliers.

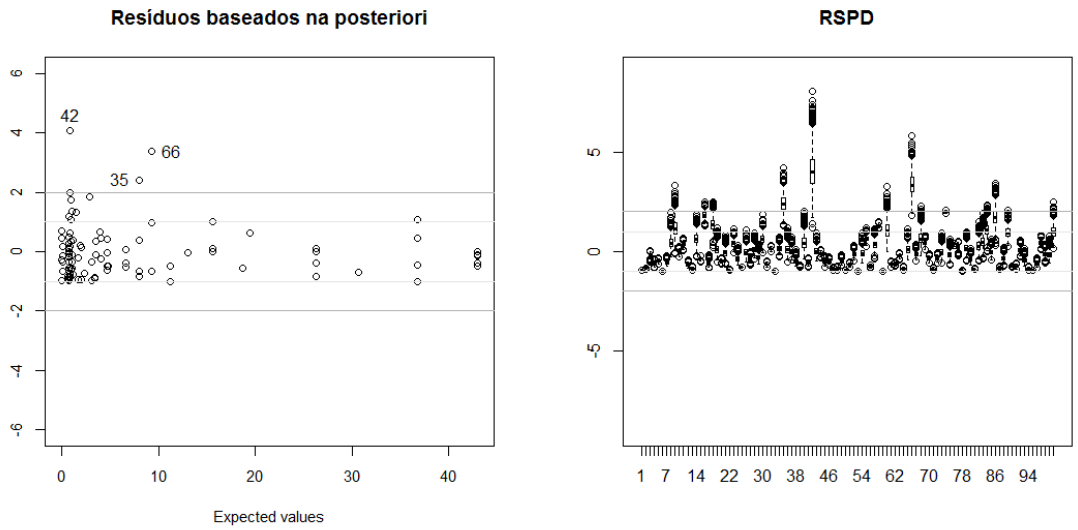

Figura 4.11: Gráfico dos resíduos baseados na distribuição a posteriori dos parâmetros versus valores esperados e boxplot das amostras MCMC da distribuição a posteriori.

Pelo gráfico dos resíduos baseados na distribuição a posteriori dos parâmetros, considerando o intervalo com 95% de credibilidade, observamos a presença de alguns pontos extremos, fora da faixa de intervalo $(-2, 2)$, mas apenas dois pontos se destacam dos demais pontos, o caso 42 e o caso 66 e, por este motivo, os consideramos como outliers.

Pelo boxplot da Figura 4.11 observamos que a maioria dos pontos apresenta pequenos intervalos, o que mostra pequena variação das estimativas, mas que os pontos 42 e 66 apresentam intervalos maiores, o que também nos leva a identificá-los como outliers.

Na Figura 4.12 apresentamos os resultados para a calibração que podem indicar presença de pontos influentes. São considerados pontos influentes aqueles que apresentarem valor da calibração muito maior do que 0.5.

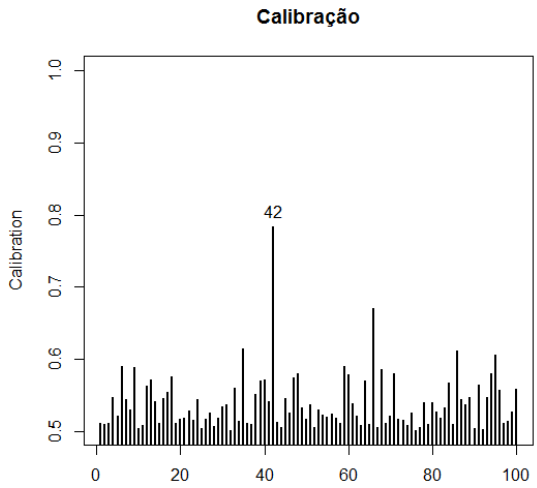

Figura 4.12: Gráfico da calibração p^*

Pelo gráfico da Figura 4.12 observamos que apenas o caso 42, detectado também como outlier neste caso, é um ponto influente no modelo, por apresentar valor de calibração muito maior do que para os demais pontos. Com isso retiramos este ponto do modelo e refazemos o ajuste para verificar o impacto que ele causa no modelo.

Na Tabela 4.10 apresentamos os valores dos parâmetros para os modelos com e sem o ponto influente. A tabela mostra o impacto causado pelo caso 42 na estimação dos parâmetros.

Tabela 4.10: Mudança relativa da retirada do ponto do modelo

Parâmetros	com o ponto	sem o ponto	mudança relativa
β_1	1.3916	1.3769	-1.06%
β_2	-2.5931	-2.5313	-2.38%
γ	0.7031	0.6276	-10.74%

Pela Tabela 4.10 verificamos que a retirada do caso 42 do modelo ocasionou um pequeno impacto na estimativa do parâmetro γ, mas não ocasionou grande impacto nas estimativas obtidas nos demais parâmetros.

Na Figura 4.31 ilustramos o gráfico dos resíduos baseados da distribuição a posteriori dos parâmetros e na Figura 4.14 ilustramos o gráfico da calibração, ambas considerando os dados sem o caso 42.

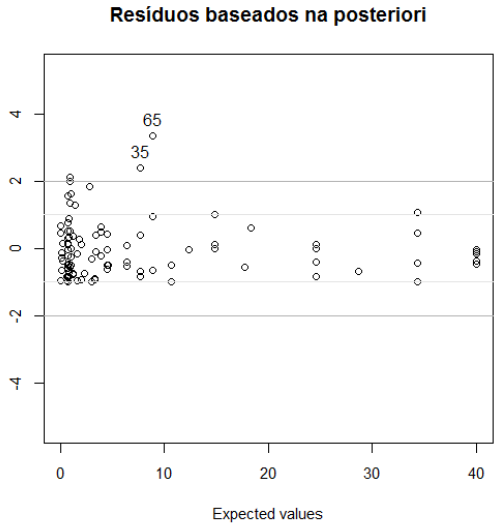

Figura 4.13: Gráfico dos resíduos baseados na distribuição a posteriori dos parâmetros versus valores esperados, para os dados sem o caso 42.

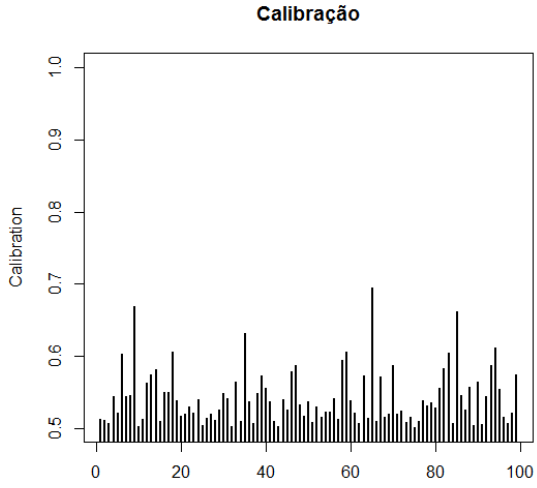

Figura 4.14: Gráfico da calibração p^*, para os dados sem o caso 42

Pelo gráfico dos resíduos baseados na distribuição a posteriori dos parâmetros do modelo observamos que a retirada do caso 42 não levou ao aparecimento de novos pontos outliers e pelo gráfico

da calibração observamos que a retirada do caso 42 não levou a altos valores de calibração para nenhum dos demais pontos. Com isso, a retirada do ponto não se torna necessária.

Outros resultados

No Modelo 3 realizamos, também, o estudo de simulação para os valores dos parâmetros fixados em $\beta_1 = 1.35$, $\beta_2 = 2.60$ e $\gamma = 0.67$ e em $\beta_1 = -1.35$, $\beta_2 = 2.60$ e $\gamma = 0.67$, para os tamanhos amostrais $n = 50$, $n = 100$ e $n = 500$, com o objetivo de constatar o bom funcionamento do modelo para outros valores dos parâmetros.

Considerando $\beta_1 = 1.35$, $\beta_2 = 2.60$ e $\gamma = 0.67$, obtivemos os seguintes resultados.

Tabela 4.11: Medidas descritivas para os parâmetros β_1, β_2 e γ

n = 50	Real	média	mediana	variância	IC		HPD		vício	eqm
β_1	1.35	1.0255	1.0121	0.1058	0.3754	1.7502	0.3553	1.7203	-0.3245	0.2111
β_2	2.60	2.6121	2.6113	0.0192	2.3047	2.9243	2.3041	2.9209	0.0121	0.0194
γ	0.67	0.8998	0.8185	0.1932	0.3599	1.9128	0.2827	1.7030	0.2298	0.2460
n = 100	Real	média	mediana	variância	IC		HPD		vício	eqm
β_1	1.35	1.3851	1.3741	0.0013	0.8566	1.9757	0.8389	1.9497	0.0351	0.0025
β_2	2.60	2.6029	2.6034	0.0106	2.3934	2.8104	2.3956	2.8107	0.0029	0.0106
γ	0.67	0.8541	0.7921	0.1526	0.3570	1.7071	0.2913	1.5496	0.1841	0.1865
n = 500	Real	média	mediana	variância	IC		HPD		vício	eqm
β_1	1.35	1.3664	1.3643	0.0003	1.1235	1.6216	1.1209	1.6161	0.0164	0.0005
β_2	2.60	2.6006	2.6006	0.0026	2.5071	2.6941	2.5080	2.6941	0.0006	0.0026
γ	0.67	0.6924	0.6837	0.0117	0.5042	0.9301	0.4908	0.9094	0.0224	0.0122

Considerando $\beta_1 = -1.35$, $\beta_2 = 2.60$ e $\gamma = 0.67$, temos os seguintes resultados.

Tabela 4.12: Medidas descritivas para os parâmetros β_1, β_2 e γ

n = 50	Real	média	mediana	variância	IC		HPD		vício	eqm
β_1	-1.35	-1.4614	-1.4415	0.0125	-2.3011	-0.7319	-2.2557	-0.6999	-0.1114	0.0250
β_2	2.60	2.5819	2.5751	0.0400	2.2182	2.9833	2.2079	2.9677	-0.0181	0.0404
γ	0.67	0.7437	0.7158	0.0406	0.4397	1.2082	0.4032	1.1378	0.0737	0.0461
n = 100	Real	média	mediana	variância	IC		HPD		vício	eqm
β_1	-1.35	-1.8567	-1.8388	0.2581	-2.5908	-1.2222	-2.5517	-1.1958	-0.5067	0.5149
β_2	2.60	2.5935	2.5912	0.0167	2.3341	2.8663	2.3309	2.8604	-0.0065	0.0167
γ	0.67	0.7112	0.6955	0.0175	0.4785	1.0343	0.4561	0.9957	0.0412	0.0192
n = 500	Real	média	mediana	variância	IC		HPD		vício	eqm
β_1	-1.35	-1.4333	-1.4312	0.0070	-1.6895	-1.1892	-1.6835	-1.1857	-0.0833	0.0139
β_2	2.60	2.5938	2.5932	0.0034	2.4776	2.7132	2.4766	2.7111	-0.0062	0.0035
γ	0.67	0.6773	0.6745	0.0033	0.5720	0.7985	0.5673	0.7918	0.0073	0.0034

Os resultados mostraram que, também para esses valores dos parâmetros, os valores obtidos das estimativas são próximos do valor real, que as estimativas possuem pouca variação e que estão contidas nos intervalos de credibilidade. Isto permite observarmos o bom funcionamento do modelo para diferentes valores dos parâmetros.

4.4 Análise de Dados Reais

Nesta seção analisamos um conjunto de dados reais para ilustrar a metodologia desenvolvida nos três modelos apresentados no capítulo anterior. Este conjunto de dados reais, fornecido por uma operadora de planos de saúde, está relacionado a gastos com pacientes hospitalizados, levando em conta a utilização ou não de centro cirúrgico. O conjunto de dados é composto de 150 internações.

Utilizamos como variável resposta discreta a necessidade ou não de procedimentos cirúrgicos e como variável resposta contínua gastos totais com o paciente hospitalizado. A covariável disponível é número de dias de permanência do paciente hospitalizado. Os resultados são apresentados na tabela de resumo a posteriori, para os três modelos.

Realizamos uma análise de resíduos e um diagnóstico de influência para os três modelos apresentados, considerando os resíduos descritos no Capítulo 3. Em relação ao Modelo 1 e ao Modelo 3, assim como nos dados simulados, para o cálculo dos resíduos deviance Bayesianos foi necessário

admitir restrições aos parâmetros $\beta_2 > 0$ e $0 < \gamma < 1$.

Nos três modelos aplicamos o processo MCMC com 130000 iterações, descartando as primeiras 10000 iterações como "burn-in" e o restante das iterações espaçadas de 60 em 60 observações com o objetivo de eliminar a autocorrelação existente entre os parâmetros. No final do processo obtivemos uma amostra de tamanho 2000.

4.4.1 Modelo 1

Os resultados para o Modelo 1, média, mediana, variância, intervalos de credibilidade e intervalos HPD dos parâmetros, são apresentados na Tabela 4.13. Gráfico de convergência, gráfico ACF e teste de convergência de Geweke foram realizados e apresentaram resultados satisfatórios. Os gráficos e o teste não são apresentados neste trabalho.

Tabela 4.13: Média, mediana, variância e intervalos de credibilidade dos parâmetros do Modelo 1

$n = 150$	média	mediana	variância	IC		HPD	
β_1	-0.0407	-0.0405	0.00012	-0.0635	-0.0212	-0.0620	-0.0202
β_2	0.0773	0.0773	0.000023	0.0681	0.0868	0.0676	0.0863
γ	0.4807	0.4747	0.0335	0.1362	0.8626	0.0960	0.8156

Com o ajuste do Modelo 1 é possível verificar, pela tabela de resumo a posteriori, que os valores da média e da mediana estão bem próximos, que há pouca variação das estimativas e que estas estão contidas nos intervalos de credibilidade.

O modelo final a ser considerado para a média da variável Y e da variável X condicionada a Y, neste caso, será dado por

$$E[Y_i] = \frac{e^{-0.0407 * z_i}}{1 + e^{-0.0407 * z_i}}$$

e

$$E[X_i|Y_i] = \exp\left(0.0773 z_i + 0.4807 y_i\right) .$$

A Figura 4.15 apresenta os gráficos para $E[Y_i]$ e para $E[X_i|Y_i]$ do Modelo 1 final.

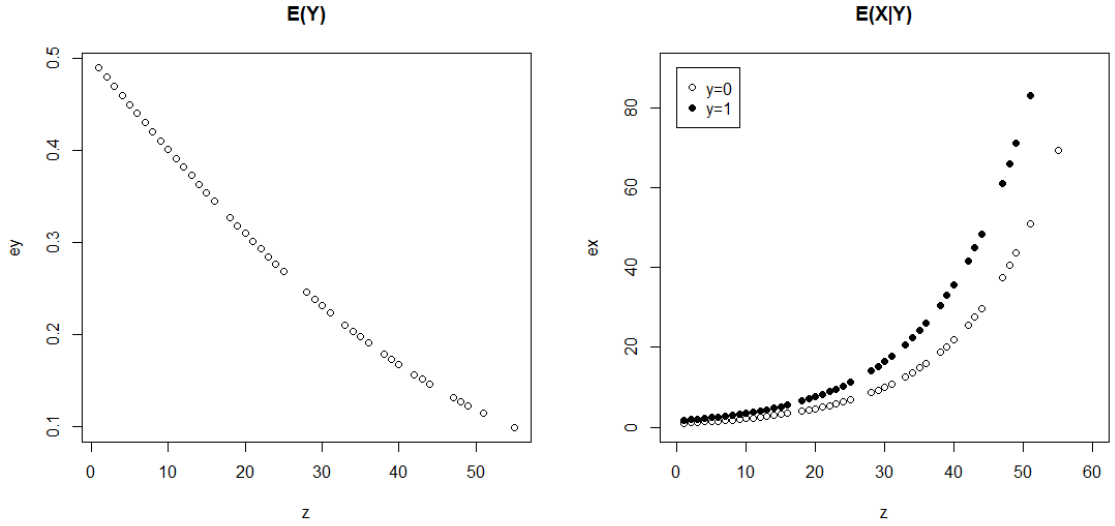

Figura 4.15: Gráficos do modelo final 1 para a média de Y e para a média de $X|Y$, respectivamente.

É possível observar pela Figura 4.15 que $E(X|Y)$, representando os gastos com o paciente, em ambas as curvas $y = 0$ e $y = 1$, têm o mesmo comportamento, porém a curva para $y = 1$, representando utilização de procedimento cirúrgico, apresenta valores maiores do que para $y = 0$.

Na figura 4.16 apresentamos os histogramas dos três parâmetros do Modelo 1, $\beta 1$, $\beta 2$ e γ, respectivamente, considerando o conjunto de dados reais. Sobre estes histogramas traçamos a curva da densidade a posteriori de cada parâmetro.

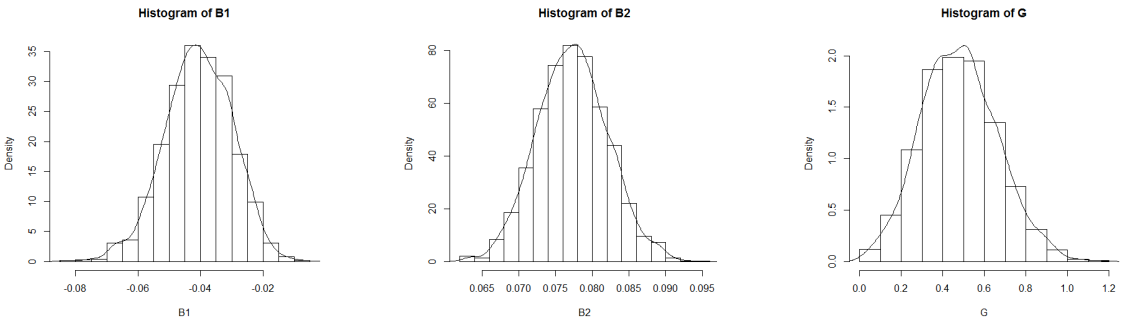

Figura 4.16: Histogramas dos parâmetros $\beta 1$, $\beta 2$ e γ.

A adequabilidade do modelo é verificada através de uma análise de resíduos, utilizando, para este modelo, o resíduo baseado na distribuição a posteriori dos parâmetros do modelo. Realizamos,

também um diagnóstico de influência para verificar a presença de pontos influentes.

Análise de resíduos

Para o Modelo 1 realizamos uma análise de resíduos considerando o resíduo baseado na distribuição a posteriori dos parâmetros do modelo e resíduos deviance Bayesianos, assim como nos dados simulados. A Figura 4.17 ilustra o gráfico do resíduo e o boxplot das amostras MCMC da distribuição a posteriori, para os resíduos baseados na distribuição a posteriori dos parâmetros do modelo. Analisamos estes gráficos a fim de verificar presença ou ausência de pontos extremos que podem ser considerados como outliers.

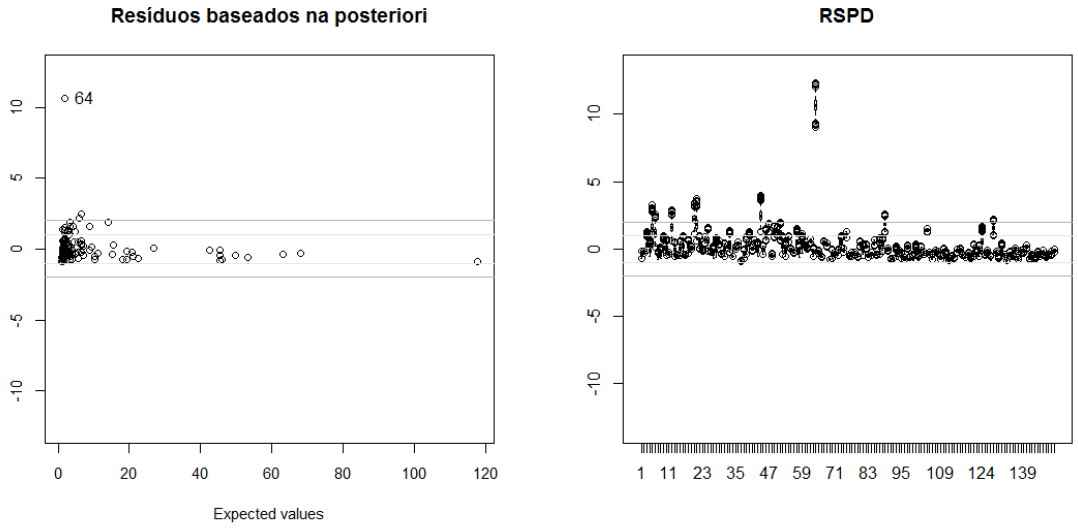

Figura 4.17: Gráfico dos resíduos baseados na distribuição a posteriori dos parâmetros do modelo versus valores esperados e boxplot das amostras MCMC da distribuição a posteriori.

A partir dos gráficos da Figura 4.17 percebemos a presença de um ponto extremo, fora da faixa horizontal de pontos centrada no zero, o caso 64, tanto no gráfico de resíduos como no boxplot, o que nos leva a considerá-lo como outlier. Os demais pontos do boxplot da Figura 4.17 apresentam pouca variação, o que indica pequena variação das estimativas.

Realizamos um teste de calibração, ilustrado na Figura 4.18, para verificar a existência de influência de algum ponto no modelo.

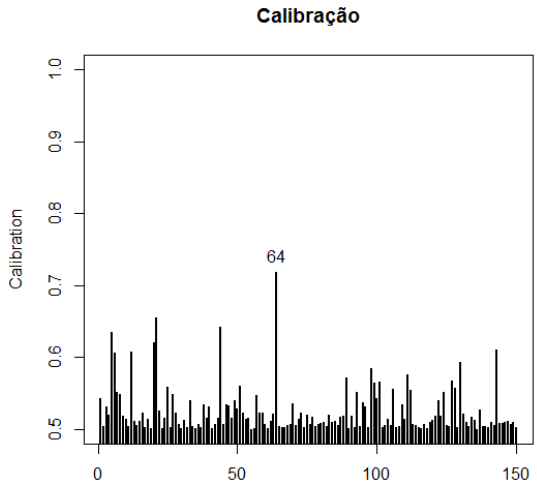

Figura 4.18: Calibração dos resíduos para os dados reais no Modelo 1.

Observamos, pelos resultados obtidos, que neste caso, apenas o ponto 64, detectado também como outlier, apresentou alto valor de calibração, o que nos leva a considerá-lo como influente no modelo. Um reajuste do modelo é então apresentado retirando este caso detectado como influente para verificar se tal ponto causa impacto no modelo.

A Tabela 4.14 apresenta os valores dos parâmetros para os modelos com e sem o ponto. A tabela mostra o impacto causado pelo caso 64 na estimação dos parâmetros.

Tabela 4.14: Mudança relativa da retirada do ponto do modelo

Parâmetros	com o ponto	sem o ponto	mudança relativa
β_1	-0.0407	-0.0398	-2,14%
β_2	0.0773	0.0748	-3.32%
γ	0.4807	0.5192	8.02%

Pela Tabela 4.14 e pelo gráfico de resíduos para os dados sem o ponto 64 verificamos que, apesar deste ponto ser um ponto outlier e se mostrar influente, ele não causa impacto na estimação dos parâmetros do modelo.

A Figura 4.19 apresenta o gráfico do resíduo baseado na distribuição a posteriori dos parâmetros e o gráfico da calibração, ambas considerando os dados sem o caso 64.

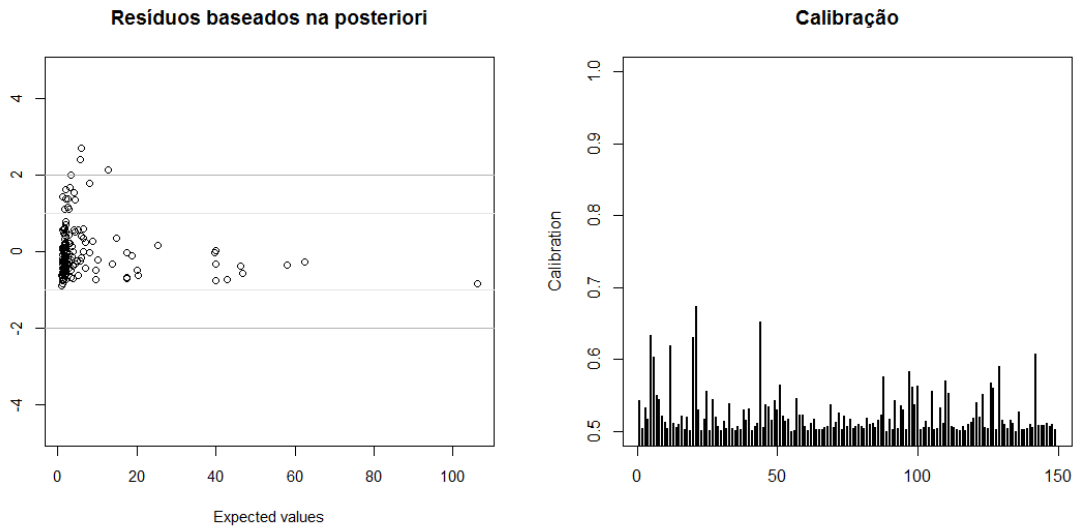

Figura 4.19: Gráfico dos resíduos baseados na distribuição a posteriori dos parâmetros versus valores esperados e gráfico da calibração p^*, ambos para os dados sem o caso 64

Pelo gráfico de resíduos baseados na distribuição a posteriori dos parâmetros observamos que a retirada do caso 64 não levou ao aparecimento de novos pontos outliers e pelo gráfico da calibração observamos que a retirada do caso 64 não ocasionou altos valores de calibração para nenhum dos demais pontos.

4.4.2 Modelo 2

A Tabela 4.15 apresenta o resumo a posteriori para os três parâmetros do modelo, contendo a média, a mediana, a variância e os intervalos de credibilidade e HPD.

Tabela 4.15: Média, mediana, variância e intervalos de credibilidade dos parâmetros do Modelo 2

$n = 150$	média	mediana	variância	IC		HPD	
β_1	-0.0031	-0.0031	0.000008	-0.0089	0.0024	-0.0089	0.0024
β_2	0.0200	0.0200	0.0000024	0.0169	0.0230	0.0170	0.0232
γ	0.7670	0.6388	0.3728	0.0296	2.2503	0.000032	1.9164

Com o ajuste do Modelo 2 é possível observar que os valores da média e da mediana estão próximos, que há pouca variação das estimativas e que estas estão contidas nos intervalos de credibilidade. Observamos, também, que para o parâmetro β_1, o intervalo de credibilidade contém o zero, o que pode indicar que a covariável não é significativa para este parâmetro nesse modelo.

Sendo assim, o modelo final a ser considerado para a média da variável Y e da variável X condicionada a Y, neste caso, é dado por

$$E[Y_i] = \frac{e^{-0.0031*z_i}}{1+e^{-0.0031*z_i}}$$

e

$$E[X_i|Y_i] = e^{0.02z_i} + 0.7670y_i\left(\frac{e^{-0.0031z_i}}{1+e^{-0.0031z_i}}\right)$$

A Figura 4.20 apresenta os gráficos que ilustram as curvas do modelo 2 final para $E[Y_i]$ e para $E[X_i|Y_i]$.

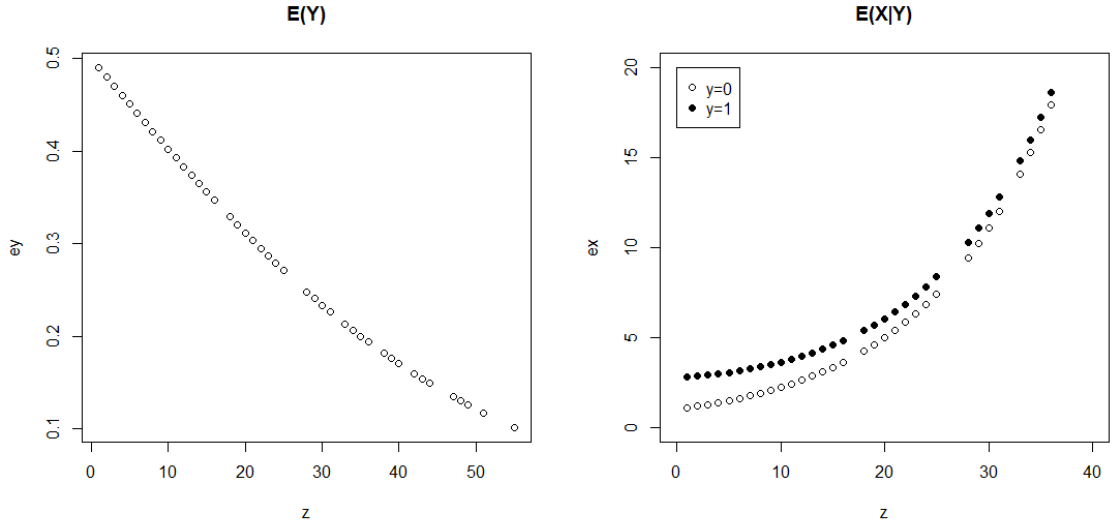

Figura 4.20: Gráficos do modelo final 2 para a média de Y e para a média de $X|Y$, respectivamente.

Observamos pela Figura 4.20 que para $E[X_i|Y_i]$, representando os gastos com o paciente, a curva $y = 1$, representado necessidade de procedimentos cirúrgicos, apresenta valores maiores do que a curva $y = 0$, embora ambas as curvas apresentem o mesmo comportamento.

Na figura 4.21 apresentamos os histogramas dos três parâmetros do Modelo 2, $\beta 1$, $\beta 2$ e γ, respectivamente, considerando o conjunto de dados reais. Sobre estes histogramas traçamos a curva da densidade a posteriori de cada parâmetro.

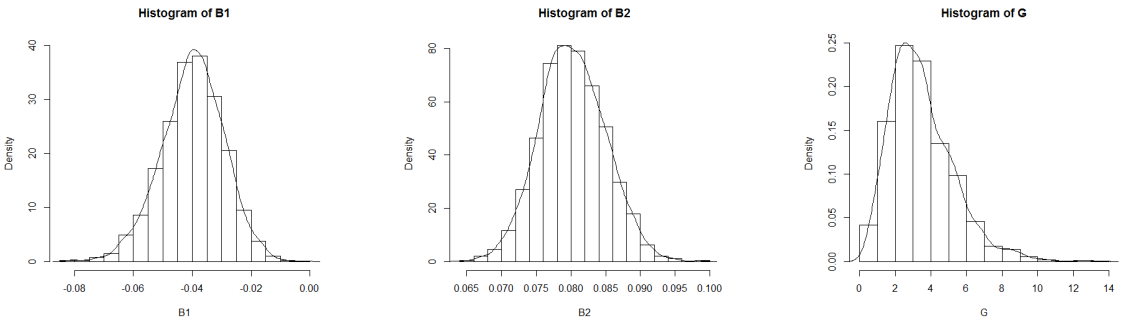

Figura 4.21: Histogramas dos parâmetros $\beta 1$, $\beta 2$ e γ.

Observamos nos histogramas da Figura 4.21 que para o parâmetro γ é possível notar uma assimetria em torno da média, já os parâmetros β_1 e β_1 apresentam simetria em torno de suas

médias.

Análise de resíduos

Em relação ao Modelo 2, realizamos uma análise de resíduos considerando os dois tipos de resíduos apresentados no Capítulo 3, resíduos baseados na distribuição a posteriori dos parâmetros do modelo e resíduos deviance Bayesianos. Na Figura 4.22 mostramos os gráficos para os dois tipos de resíduos e na Figura 4.23 ilustramos o boxplot das amostras MCMC da distribuição a posteriori para os resíduos baseados na distribuição a posteriori dos parâmetros do modelo. Por estes gráficos podemos analisar a presença ou ausência de pontos extremos que podem ser considerados como outliers.

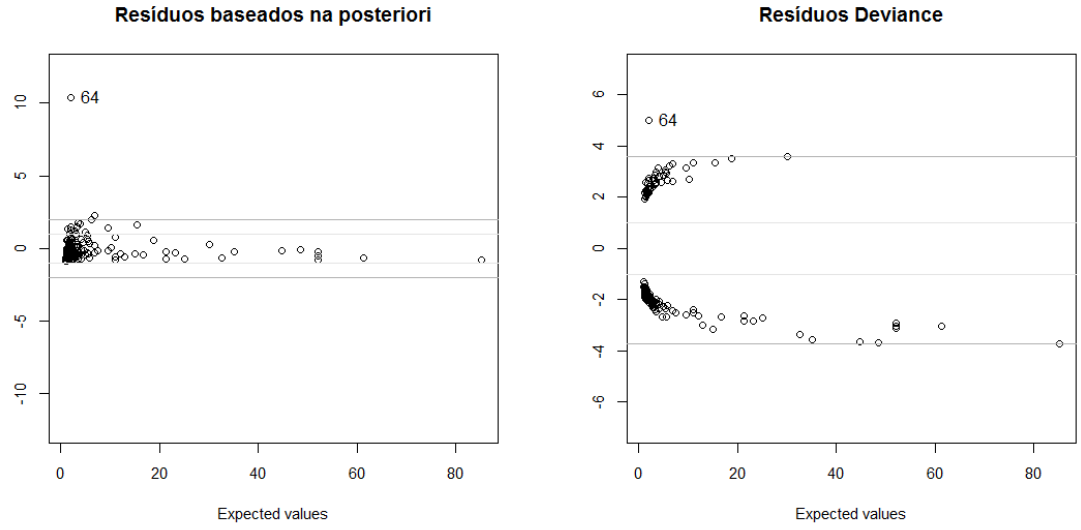

Figura 4.22: Gráfico dos resíduos baseados na distribuição a posteriori dos parâmetros do modelo versus valores esperados e gráfico dos resíduos deviance versus valores esperados.

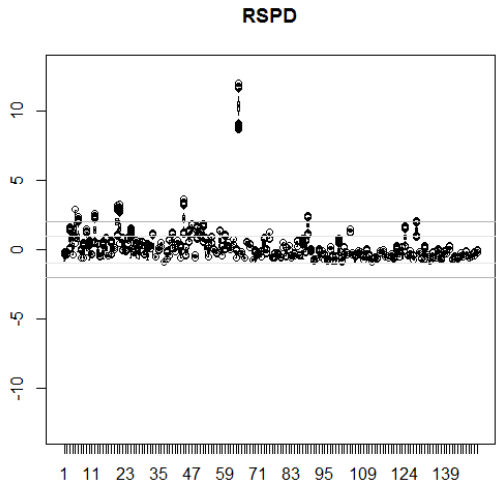

Figura 4.23: Boxplot das amostras MCMC da distribuição a posteriori

Pelos gráficos da Figura 4.22 e do boxplot da Figura 4.23 percebemos a presença de um ponto extremo, fora da faixa horizontal de pontos centrada no zero, o caso 64, o que nos leva a considerá-lo como outlier. Pelo boxplot notamos os pontos apresentam pequenos intervalos, indicando pequena variação das estimativas.

Realizamos um teste de calibração, também apresentado no Capítulo 3, ilustrado na Figura 4.24, para verificar se existe influência de algum ponto no modelo.

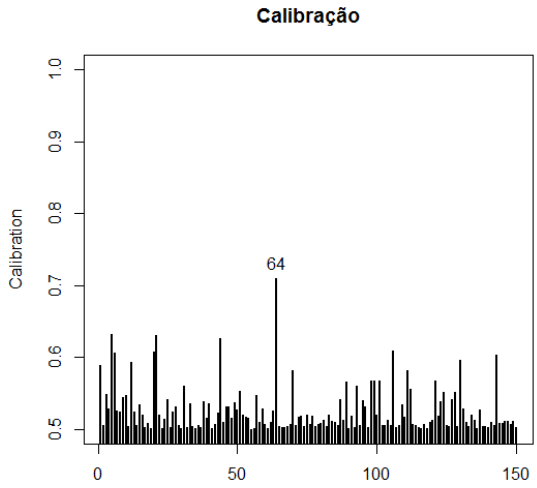

Figura 4.24: Calibração dos resíduos.

Através dos resultados apresentados notamos que, neste caso, apenas o ponto 64, visto nos gráficos dos resíduos como outlier, apresentou alto valor de calibração, o que pode nos levar a considerá-lo como influente no modelo. Um reajuste do modelo é então apresentado retirando este caso detectado como influente para verificar se tal ponto causa impactos no modelo.

A Tabela 4.16 apresenta os valores dos parâmetros para os modelos com e sem o ponto e mostra o impacto causado pelo caso 64 na estimação dos parâmetros.

Tabela 4.16: Mudança relativa da retirada do ponto do modelo

Parâmetros	com o ponto	sem o ponto	mudança relativa
β_1	-0.0398	-0.0391	-1.76%
β_2	0.0801	0.0780	-2.62%
γ	3.4940	3.6822	5.39%

Pela Tabela 4.16 e pelos gráficos de resíduos para os dados sem o ponto 64 verificamos que apesar deste ponto ser um ponto outlier e se mostrar influente ele não causa impacto na estimação dos parâmetros do modelo.

Na Figura 4.25 ilustramos os gráficos para os dois tipos de resíduos e na Figura 4.26 ilustramos o gráfico da calibração, ambas considerando os dados sem o caso 64.

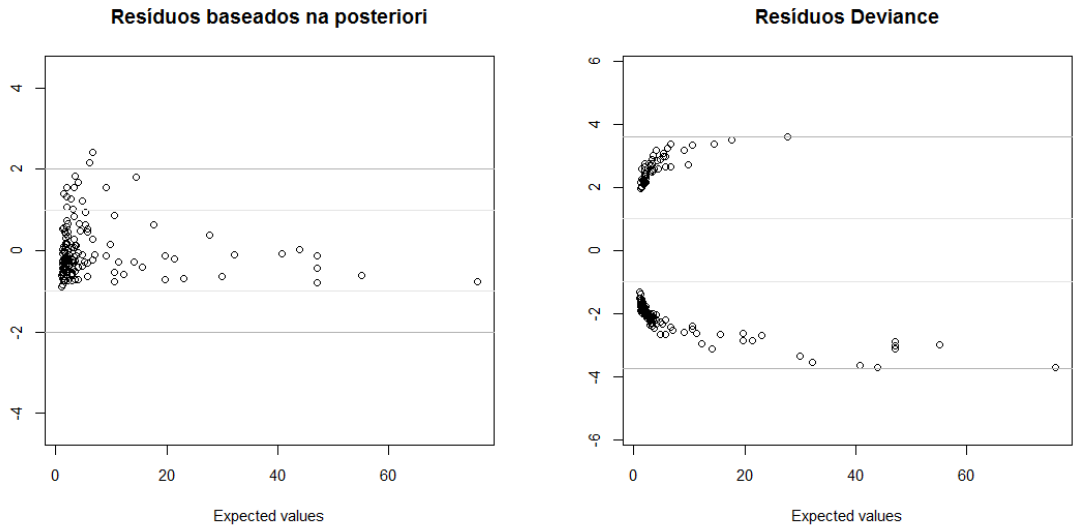

Figura 4.25: Gráfico dos resíduos baseados na distribuição a posteriori dos parâmetros versus valores esperados e gráfico dos resíduos deviance Bayesianos versus valores esperados, ambos para os dados sem o caso 64

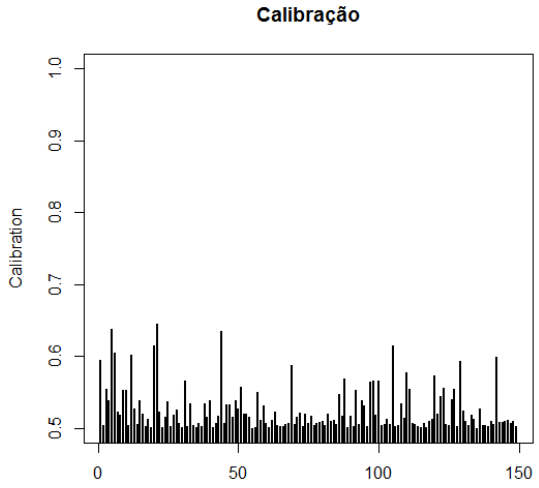

Figura 4.26: Gráfico da calibração p^*, para os dados sem o caso 64

Pelo gráfico de resíduos baseados na distribuição a posteriori dos parâmetros e pelo gráfico de

resíduos deviance Bayesianos observamos que a retirada do caso 64 não levou ao aparecimento de novos pontos outliers e pelo gráfico da calibração observamos que a retirada do caso 64 não levou a altos valores de calibração para nenhum dos demais pontos.

4.4.3 Modelo 3

Apresentamos o resumo a posteriori dos três parâmetros do Modelo 3, através da Tabela 4.17, contendo a média, a mediana, a variância e os intervalos de credibilidade e HPD.

Tabela 4.17: Média, mediana, variância e intervalos de credibilidade dos parâmetros do Modelo 3

$n = 150$	média	mediana	variância	IC		HPD	
β_1	-0.0409	-0.0405	0.00012	-0.0636	-0.0209	-0.0636	-0.0211
β_2	0.0796	0.0796	0.0000025	0.0694	0.0895	0.0693	0.0892
γ	1.6009	1.5152	0.5729	0.4338	3.3270	0.2557	3.0167

Com o ajuste do Modelo 3 é possível observar, pela Tabela 4.17, que os valores da média e da mediana estão bem próximos, que há pouca variação das estimativas e que estas estão contidas nos intervalos de credibilidade.

Sendo assim, o modelo final a ser considerado para a média da variável Y e da variável X condicionada a Y, neste caso, será dado por

$$E[Y_i] = \frac{e^{-0.0409*z_i}}{1+e^{-0.0409*z_i}}$$

e

$$E[X_i|Y_i] = e^{(0.0796z_i)} + 1.6009y_i$$

A Figura 4.27 apresenta os gráfico do modelo 3 final para $E[Y_i]$ e para $E[X_i|Y_i]$.

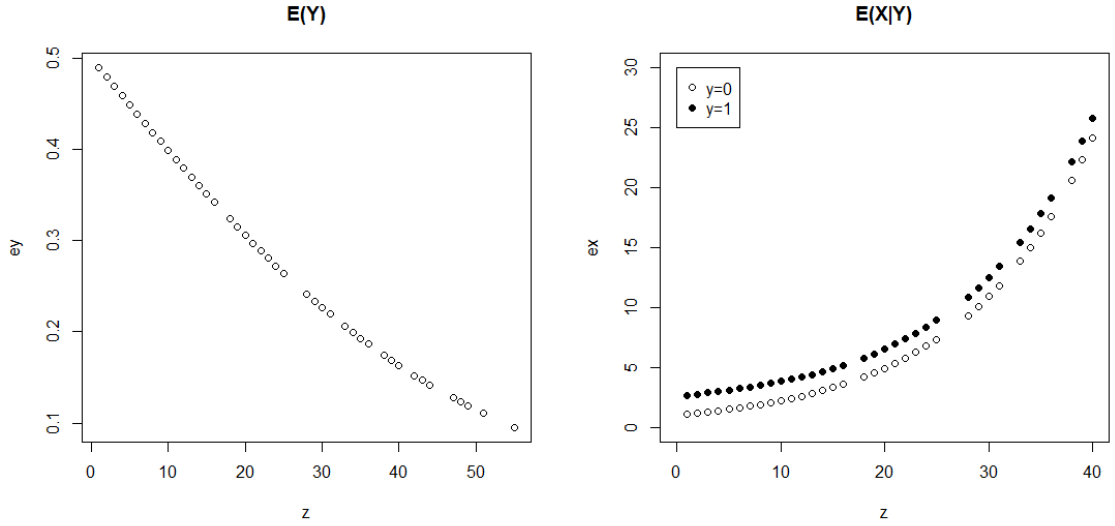

Figura 4.27: Gráficos do modelo final 3 para a média de Y e para a média de $X|Y$, respectivamente.

Pela Figura 4.27 observamos que os valores no modelo final para a média de $X|Y$, representando os gastos com os pacientes, são maiores para a curva de $y = 1$, representando a necessidade de procedimentos cirúrgicos, do que para a curva de $y = 0$, embora ambas as curvas apresentem o mesmo comportamento.

Na figura 4.28 apresentamos os histogramas dos três parâmetros do Modelo 3, $\beta 1$, $\beta 2$ e γ, respectivamente, considerando o conjunto de dados reais. Sobre estes histogramas traçamos a curva da densidade a posteriori de cada parâmetro.

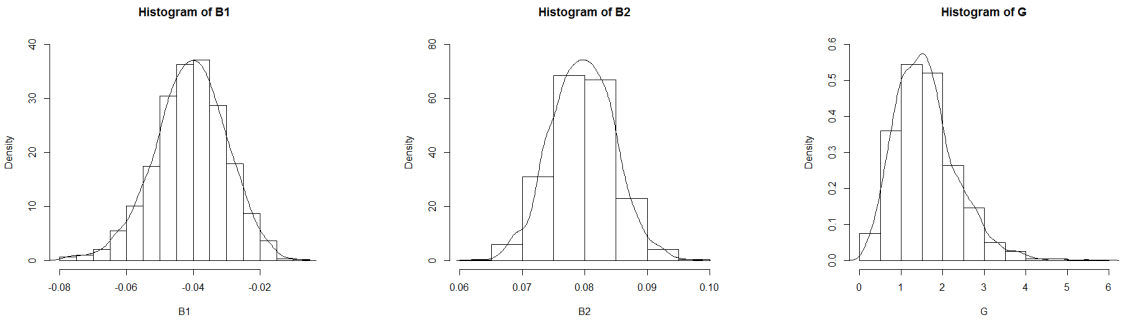

Figura 4.28: Histogramas dos parâmetros $\beta 1$, $\beta 2$ e γ.

Pelos histogramas da Figura 4.28 notamos uma leve assimetria dos parâmetros β_1 e γ em torno

de suas médias, já o parâmetro β_2 mostra-se simétrico em torno de sua média.

A adequação deste modelo é verificada através uma análise de resíduos, utilizando o resíduo baseado na distribuição a posteriori dos parâmetros, para este modelo.

Análise de resíduos

Realizamos uma análise de resíduos, para o Modelo 3, considerando os resíduos baseados na distribuição a posteriori dos parâmetros do modelo, assim como nos dados simulados. A Figura 4.29 ilustra o gráfico de resíduos e o boxplot das amostras MCMC da distribuição a posteriori dos parâmetros do modelo. Analisamos estes gráficos a fim de verificar presença ou ausência de pontos extremos que podem ser considerados como outliers.

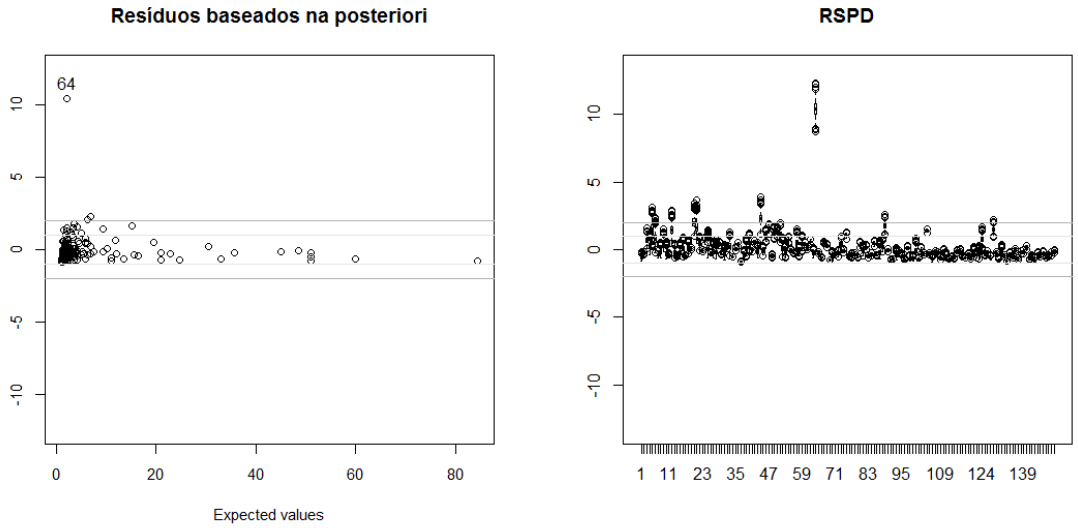

Figura 4.29: Gráfico dos resíduos baseados na distribuição a posteriori dos parâmetros do modelo versus valores esperados e boxplot das amostras MCMC da distribuição a posteriori.

A partir dos gráficos da Figura 4.29 observamos que o caso 64 é um ponto extremo, fora da faixa horizontal de pontos centrada no zero, tanto no gráfico de resíduos como no boxplot, o que nos leva a considerá-lo como outlier. Os pontos do boxplot da Figura 4.29 apresentam pequenos intervalos, o que indicam pequena variação nas estimativas.

Realizamos um teste de calibração, ilustrado na Figura 4.30, para verificar a existência de influência de algum ponto no modelo.

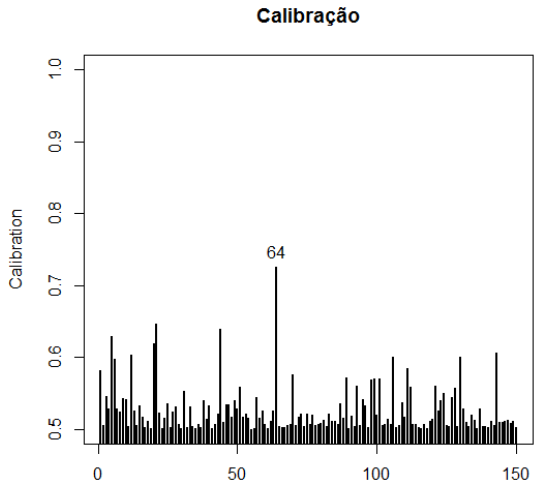

Figura 4.30: Calibração dos resíduos para os dados reais no Modelo 3.

Pelos resultados obtidos observamos que apenas o ponto 64, detectado também como outlier neste caso, apresentou alto valor de calibração, o que nos leva a considerá-lo, também, como ponto influente no modelo. Um reajuste do modelo é então apresentado retirando este caso detectado como influente para verificar se tal ponto causa impacto no modelo.

A Tabela 4.18 apresenta os valores dos parâmetros para os modelos com e sem o ponto, mostrando o impacto causado pelo caso 64 na estimação dos parâmetros.

Tabela 4.18: Mudança relativa da retirada do ponto do modelo

Parâmetros	com o ponto	sem o ponto	mudança relativa
β_1	-0.0409	-0.0405	-0.99%
β_2	0.0796	0.0797	0.10%
γ	1.6009	1.6174	1.03%

Pela Tabela 4.18 e pelo gráfico de resíduos para os dados sem o ponto 64 verificamos que apesar deste ponto ser um ponto outlier e se mostrar influente ele não causa impacto na estimação dos parâmetros do modelo.

A Figura 4.31 apresenta o gráfico de resíduos baseados na distribuição a posteriori dos parâmetros e o gráfico da calibração, ambos considerando os dados sem o caso 64.

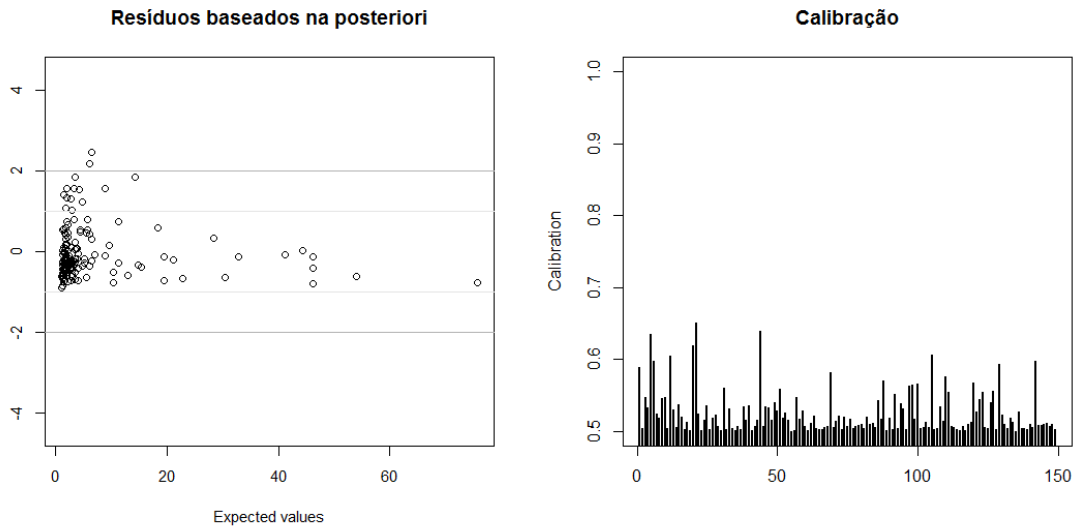

Figura 4.31: Gráfico dos resíduos baseados na distribuição a posteriori dos parâmetros versus valores esperados e gráfico da calibração p^*, ambos para os dados sem o caso 64

Através do gráfico de resíduos baseados na distribuição a posteriori dos parâmetros observamos que a retirada do caso 64 não levou ao aparecimento de novos pontos outliers e pelo gráfico da calibração observamos que a retirada do caso 64 não ocasionou altos valores de calibração para nenhum dos demais pontos.

4.4.4 Conclusões

Comparando os resultados obtidos para os três modelos analisados, embora todos apresentaram resultados parecidos, foi possível perceber que o Modelo 1 apresentou melhores resultados em relação aos outros dois modelos; as estimativas do Modelo 1, nos dados simulados, apresentaram valores mais próximos do valor real e os histogramas dos parâmetros β_1, β_2 e γ deste modelo apresentaram-se mais simétricos em torno de suas médias.

A vantagem do Modelo 1 e do Modelo 3 em relação ao Modelo 2 é a não utilização da média da variável discreta, μ_{1_i}, na construção da média da variável contínua, λ_i, fazendo com que λ_i não dependa desta média.

Capítulo 5

Considerações Finais

Esta dissertação apresenta modelos de regressão bivariados baseados em um modelo Bernoulli para a resposta discreta e um modelo exponencial para a resposta contínua, condicionada na resposta discreta, seguindo o trabalho de modelos bivariados de Fitzmaurice & Laird (1995). Os três modelos apresentados foram construídos com o objetivo de obtermos modelos bivariados com uma densidade marginal com distribuição Bernoulli e uma densidade condicional com distribuição exponencial.

Utilizamos uma abordagem Bayesiana para estimar os parâmetros do modelo. Apresentamos uma análise de resíduos considerando resíduos baseados na distribuição preditiva a posteriori, resíduos baseados na distribuição a posteriori dos parâmetros do modelo para o Modelo 1, Modelo 2 e Modelo 3 e resíduos deviance Bayesiano para o Modelo 2 e, um diagnóstico de influência para os três modelos.

Um estudo de simulação foi apresentado para ilustrar a metodologia considerando três tamanhos amostrais diferentes e, através dos resultados obtidos, foi possível verificar o bom funcionamento dos modelos desenvolvidos. Pelas análises de resíduos apresentadas observamos a adequabilidade dos modelos aos dados e detectamos a presença de alguns pontos extremos que foram investigados como possíveis outliers. Tais pontos foram constatados como outliers, porém não apresentaram influência nos modelos.

Realizamos uma analise de dados reais, baseada em um conjunto de dados de uma operadora de planos de saúde em relação a pacientes hospitalizados, considerando como resposta discreta utilização, ou não, de procedimentos cirúrgicos e como resposta contínua os gastos totais com o paciente hospitalizado. Assim como para os dados simulados observamos, também para os dados

reais, um bom funcionamento do modelo no processo de estimação dos parâmetros e adequabilidade aos dados.

A realização do processo de estimação em uma abordagem clássica não foi possível devido a inversibilidade da matriz de covariâncias nos três modelos, que seria utilizada em um processo iterativo de Fisher. Como proposta para futuros trabalhos pode-se buscar outras formas de obter a estimação dos parâmetros por uma abordagem clássica e a utilização de outras distribuições para as variáveis respostas.

Referências

Catalano, P. J. & Ryan, L. M. (1992). Bivariate latent variable models for clustered discrete and continuous outcomes. **87**(419), 651–658.

Cox, D. R. (1972). The Analysis of Multivariate Binary Data. *Journal of the Royal Statistical Society*, **21**(2), 113–120.

Cox, D. R. & Wermuth, N. (1992). Response Models for Mixed Binary and Quantitative Variables. *Biometrika*, **79**(3), 441–461.

Fitzmaurice, G. M. & Laird, N. M. (1995). Regression models for a bivariate discrete and continuous outcome with clustering. *Journal of the American Statistical Association*, **90**, 845–852.

Jung, R. C. & Winkelmann, R. (1993). Two aspects of labor mobility: a bivariate poisson regression approach. *Empirical Economics*, **18**(2), 543–556. The original publication is available at www.springerlink.com.

Khafri, S., Kazemnejad, A. & Eskandari, F. (2008). Hierarchical bayesian analysis of bivariate poisson regression model. *World Applied Sciences Journal*, **5**(4), 667–675.

Lauritzen, S. L. & Wermuth, N. (1989). Graphical Models for Associations between Variables, some of which are Qualitative and some Quantitative. *Annals of Statistics*, **17**(1), 31–57.

Little, R. J. A. & Rubin, D. B. (1987). *Statistical Analysis with Missing Data*. Wiley Series in Probability and Statistics. Wiley, New York, first edition.

Little, R. J. A. & Schluchter, M. D. (1985). Maximum Likelihood Estimation for Mixed Continuous and Categorical Data With Missing Values. *Biometrika*, **72**(3), 497–512.

Olkin, I. & Tate, R. F. (1961). Multivariate correlation models with mixed discrete and continuous variables. *Annals of Mathematical Statistics*, **32**, 448–465.

Pires, R. M. & Diniz, C. A. R. (2012). Correlated binomial regression models. *Comput. Stat. Data Anal.*, **56**(8), 2513–2525.

Scollnik, D. P. (2002). Regression models for bivariate loss data. *North American Actuarial Journal*, **6**(4), 67–80.

Song, J., Barnhart, H. X. & Lyles, R. H. (2004). A GEE Approach for Estimating Correlation Coefficients Involving Left-censored Variables. *Journal of Data Science*, **2**, 245–257.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of The Royal Statistical Society Series B-statistical Methodology*, **64**, 583–639.

Tate, R. F. (1954). Correlation Between a Discrete and a Continuous Variable. Point-Biserial Correlation. *The Annals of Mathematical Statistics*, **25**(3), 603–607.

Van Ophem, H. (1999). A general method to estimate correlated discrete random variables. *Econometric Theory*, **15**(02), 228–237.

Zhao, L. P., Prentice, R. L. & Self, S. G. (1992). Multivariate mean parameter estimation by using a partly exponential model. *Journal of the Royal Statistical Society, Series B*, **54**(3), 805–811.