Inferência em um modelo de regressão com resposta binária na presença de sobredispersão e erros de medição

Sandra Maria Tieppo

Inferência em um modelo de regressão com resposta binária na presença de sobredispersão e erros de medição

Sandra Maria Tieppo

Orientador: *Mário de Castro Andrade Filho*

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação - ICMC-USP, como parte dos requisitos para obtenção do título de Mestre em Ciências - Área de Ciências de Computação e Matemática Computacional.

"VERSÃO REVISADA APÓS A DEFESA"

Data da Defesa: 15/02/2007.

Visto do orientador:

USP - São Carlos.

Março / 2007.

Agradecimentos

Agradeço a Deus o dom da vida e nunca me abandonar.

Aos meus pais, Madalena e Osvaldo e ao meu irmão Marcelo, o apoio.

Ao Prof. Mário de Castro, a orientação deste trabalho e a disponibilidade.

Ao meu esposo, Sandro o incentivo e por sempre acreditar que eu conseguiria.

Agradeço aos meus professores do ICMC - USP e da UNIOESTE - Cascavel - PR.

A todos os meus amigos.

Aos funcionários do ICMC - USP.

Ao CNPq o apoio financeiro do presente trabalho.

Resumo

Modelos de regressão com resposta binária são utilizados na solução de problemas nas mais diversas áreas. Neste trabalho enfocamos dois problemas comuns em certos conjuntos de dados e que requerem técnicas apropriadas que forneçam inferências satisfatórias. Primeiro, em certas aplicações uma mesma unidade amostral é utilizada mais de uma vez, acarretando respostas positivamente correlacionadas, responsáveis por uma variância na variável resposta superior ao que comporta a distribuição binomial, fenômeno conhecido como sobredispersão. Por outro lado, também encontramos situações em que a variável explicativa contém erros de medição. É sabido que utilizar técnicas que desconsideram esses erros conduz a resultados inadequados (estimadores viesados e inconsistentes, por exemplo). Considerando um modelo com resposta binária, utilizaremos a distribuição beta-binomial para representar a sobredispersão. Os métodos de máxima verossimilhança, SIMEX, calibração da regressão e máxima pseudo-verossimilhança foram usados na estimação dos parâmetros do modelo, que são comparados através de um estudo de simulação. O estudo de simulação sugere que os métodos de máxima verossimilhança e calibração da regressão são melhores no sentido de correção do viés, especialmente para amostras de tamanho 50 e 100. Também estudaremos testes de hipóteses assintóticos (como razão de verossimilhanças, Wald e escore) a fim de testar hipóteses de interesse. Apresentaremos também um exemplo com dados reais.

Abstract

Regression models with binary response are used for solving problems in several areas. In this work we approach two common problems in some data sets and they need appropriate techniques to achieve satisfactory inference. First, in some applications, the same sample unity is utilized more than once, bringing positively correlated responses, which are responsible for the response variable variance be greater than an assumption binomial distribution, phenomenon known as overdispersion. On the other hand, also we find situations where the explanatory variable has measurement errors. It is known that the use of techniques which ignores this measurement errors brings inadequate results (e. g., biased and inconsistent estimators). Taking a model with binary response, we will use a beta-binomial distribution for modeling the overdispersion. The methods of maximum likelihood, SIMEX, regression calibration and maximum pseudo-likelihood were used in the estimation of the parameters, which are compared through a simulation study. The simulation study suggest that the maximum likelihood and regression calibration methods are better for bias correcting, especially for larger sample size. Likelihood ratio, Wald and score statistics are used in order to test hypothesis of interest. We will illustrate the techniques with an application to a real data set.

Sumário

Capítulo 1

Introdução

1.1 Motivação

Em certas aplicações lidamos com situações em que um determinado item é analisado e dele obtemos como resposta "sucesso" ou "insucesso" (sucesso é o evento de interesse). Estudaremos um modelo que associa a probabilidade de ocorrência de sucesso a uma característica do item (variável explicativa, X). Dispomos de uma amostra de n itens, cada item é examinado m_i vezes ($m_i \geq 2$) e, para cada repetição registramos $Z_{ij} = 0$ (insucesso) ou $Z_{ij} = 1$ (sucesso), $j = 1, \ldots, m_i$, $i = 1, \ldots, n$, portanto a variável resposta é binária. Consideramos $Y_i = \sum_{j=1}^{m_i} Z_{ij}$, totaliza o número de sucessos em m ensaios. Por exemplo, em um sistema de medição por atributo, bastante difundido na área industrial, uma peça é inspecionada (mais de uma vez) e classificada como defeituosa ou não-defeituosa. O interesse é associar a probabilidade de uma peça receber classificação defeituosa e uma característica da peça. Essa característica é medida com erro, cuja variância é determinada seguindo protocolos padronizados (ISO, 1997), podendo ser considerada conhecida. A Tabela 1.1 é uma forma típica de apresentação de dados dessa natureza. No cenário descrito acima, a estrutura proba-

Tabela 1.1: Apresentação dos dados.

Item	Número de repetições	Número de sucessos	Variável explicativa
1	m_1	$Y_1 = \sum_{j=1}^{m_1} Z_{1j}$	X_1
2	m_2	$Y_2 = \sum_{j=1}^{m_2} Z_{2j}$	X_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	m_i	$Y_i = \sum_{j=1}^{m_i} Z_{ij}$	X_i
$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	m_n	$Y_n = \sum_{j=1}^{m_n} Z_{nj}$	X_n

bilística da variável de interesse (Y) pode ser representada pela distribuição binomial, que é um caso particular de modelo linear generalizado (McCullagh & Nelder, 1989). Inferências são realizadas mais comumente adotando-se as funções de ligação logito e probito. Collett (2003) discute extensivamente o assunto, enfatizando aplicações. A utilização prática é facilitada pela disponibilidade de programas computacionais, como por exemplo o ambiente R (R Development Core Team, 2006).

As aplicações que originaram a proposição deste trabalho requerem duas extensões do modelo binomial. Primeiro, as observações disponíveis da variável explicativa estão contaminadas com erros de medição aditivos e, além disso, assumimos um modelo estrutural. Segundo, como a mesma unidade i é utilizada m_i vezes, as observações individuais W_{ij}, $j = 1, \ldots, m_i$, podem estar positivamente correlacionadas, de modo que a variância da variável resposta Y_i pode exceder o valor inerente à distribuição binomial – igual a $m_i\,\pi(x_i)\,[1-\pi(x_i)]$ –, fenômeno chamado de variação extra-binomial ou sobredispersão, tema central de Hinde & Demétrio (1998a) (vide também McCullagh & Nelder, 1989; Collett, 2003). Carroll *et al.* (2006) tratam de modelos não-lineares com erros de medição, propondo métodos aplicáveis aos modelos lineares generalizados.

Mais especificamente, Carroll *et al.* (1984) tratam de um modelo de regressão probito estrutural estimando os parâmetros pelo método de máxima verossimilhança. Schafer (1987, 1993) apresenta o modelo linear generalizado com uma ou mais covariáveis medidas com erro, com função de ligação canônica e covariáveis com distribuição normal utilizando o algoritmo EM. Freedman *et al.* (2004) utilizam um método denominado reconstrução de momentos para correção dos efeitos do erro de medida nas covariáveis de um modelo de regressão. Esse método é semelhante ao método de calibração da regressão (Carroll *et al.*, 2006) e é idêntico a este no modelo de regressão linear, mas apresenta resultados distintos para a regressão logística.

Em relação à sobredispersão em modelos lineares generalizados, Lin & Breslow (1996) propõem acomodar o fenômeno acrescentando efeitos aleatórios ao preditor linear, resultando em um modelo misto. Wang *et al.* (1998) levam essas idéias adiante e formulam uma nova classe de modelos bastante geral, denominada modelos lineares generalizados mistos com erros de medição - GLMMeMs. Utilizando uma abordagem bayesiana, Dey *et al.* (1997) propõem a criação de uma classe de modelos denominada modelos lineares generalizados com sobredispersão - OGLM's. Sugerem que esses modelos sejam ajustados sob uma perspectiva bayesiana utilizando distribuições a *priori* não informativas nas inferências. Favari (2006) também estudou um modelo de regressão binária com erro na variável explicativa, obtendo estimativas dos parâmetros através do algoritmo EM.

1.2 Conceitos básicos

Os modelos com erros de medição (também chamados de modelos com erros nas variáveis - MEV) foram propostos inicialmente na década de 70 do século XIX. Estes modelos podem ser interpretados como uma generalização dos modelos de regressão usuais nos quais a variável explicativa é medida sem erro. No MEV mais simples o

objetivo é fazer inferências, a partir de um conjunto de dados bivariados, sobre os parâmetros de uma reta ajustada entre as duas variáveis, ambas medidas com erro.

Nesta seção, apresentamos sucintamente, conceitos básicos sobre esses modelos, acompanhando a exposição de Cheng & Van Ness (1999). Fuller (1987) é outra referência importante. Segundo Carroll *et al.* (2006), o objetivo de modelar erros de medição é obter estimativas, aproximadamente, não viesadas dos parâmetros ajustando o modelo utilizando a variável medida com erro. Ao final deste capítulo dedicamos mais atenção aos modelos com resposta binária.

O modelo de regressão linear simples com uma variável explanatória é expresso por

$$Y = \beta_0 + \beta_1 X + e,$$

em que a variável explicativa (X) é fixa ou aleatória e o erro (e) tem média zero e é independente de X. Dado um conjunto de observações independentes, o intercepto (β_0) e o coeficiente angular (β_1) são estimados usando as técnicas de mínimos quadrados, de máxima verossimilhança ou algum procedimento robusto. O modelo de regressão linear simples com erros nas variáveis assume que as variáveis X e y estão relacionadas por

$$y = \beta_0 + \beta_1 X,$$

mas as variáveis X e y não são observáveis exatamente, só podem ser observadas com erros. Em vez de observarmos y e X diretamente, dispomos das variáveis $W = X + U$ e $Y = y + e$, em que X e os erros (U e e) não são correlacionados.

Aplicações nas quais a variável explicativa é medida na presença de erros são, possivelmente, mais comuns do que aquelas em que as medições são precisas. A maioria das variáveis médicas, tais como pressão sangüínea, batimentos cardíacos e temperatura corporal são medidas com erro. Variáveis agrícolas tais como teor de nitrogênio no solo e grau de infestação de pragas não podem ser medidas precisamente. Nas ciências econômicas, sociais e afins algumas variáveis só podem ser medidas com erros. Na

indústria, por exemplo, em uma fundição, também encontramos variáveis que contêm erros de medição, tais como o diâmetro de uma peça, a temperatura de fundição da peça e o teor de certa substância na peça.

Assumiremos que os erros $U_1, \ldots, U_n$ e $e_1, \ldots, e_n$ têm variâncias finitas e são descorrelacionados e, sem perda de generalidade, têm média zero. Resultados inferenciais exigem que essas variáveis sejam independentes e não somente não são correlacionadas. Freqüentemente supomos que os erros $(U_i,\, e_i)$ têm distribuição normal e a correlação nula implica automaticamente em independência. A suposição de normalidade dos erros torna-se importante por ter desdobramentos em relação à identificabilidade do modelo.

Existem três modelos principais dependendo das suposições que fizermos sobre a variável explicativa (X). O modelo funcional aditivo é da forma $W_i = X_i + U_i$, em que os X_i's são constantes desconhecidas; por outro lado, se os X_i's são variáveis aleatórias independentes, identicamente distribuídas e independentes dos erros de medição, o modelo é conhecido como um modelo estrutural, isto é,

$$W_i = X_i + U_i \quad \text{com} \quad X_i \overset{\text{iid}}{\sim} \mathsf{N}(\mu_X, \sigma_X^2), \tag{1.1}$$

com X_i e U_i independentes, $U_i \overset{\text{iid}}{\sim} \mathsf{N}(0, \sigma_U^2)$ para $i = 1, \ldots, n$.

As variáveis observáveis (W_i, Y_i) correspondem às variáveis não observáveis adicionadas de erros de medição (u_i, e_i), ou seja,

$$W_i = x_i + u_i \quad \text{e} \quad Y_i = y_i + e_i, \tag{1.2}$$

(u_i, e_i) são independentes e identicamente distribuídos com médias iguais a zero e matriz de covariâncias $\begin{bmatrix} \sigma_u^2 & 0 \\ 0 & \sigma_e^2 \end{bmatrix}$, $\quad i = 1, \ldots, n$ (Cheng & Van Ness, 1999).

Sabemos que o estimador de mínimos quadrados de β_1 da regressão de Y em relação a W não é consistente (Carroll *et al.*, 2006), mas converge para $\beta_1\, \sigma_X^2/(\sigma_X^2 + \sigma_U^2)$ no

modelo estrutural, de modo que ocorre atenuação no efeito da variável explicativa, pois

$$|\beta_1|\frac{\sigma_X^2}{\sigma_X^2+\sigma_U^2} \leq |\beta_1|.$$

Ignorar erros de medição pode trazer problemas tais como:

1. Atenuação do coeficiente β_1;
2. Aumento da var$(Y|W)$;
3. Os erros de medição distorcem as características dos dados, dificultando a análise gráfica do modelo;
4. A presença do erro de medida reduz o poder do teste da hipótese $\mathrm{H}_0 : \beta_1 = 0$.

O terceiro modelo, o modelo ultraestrutural (Dolby, 1976), assume que os X_i's são variáveis aleatórias independentes como em um modelo estrutural, mas não identicamente distribuídas, podendo ter diferentes médias (μ_i) e variâncias iguais (σ_X^2). O modelo ultraestrutural é uma generalização dos modelos funcional e estrutural. Se $\mu_1 = \mu_2 = \ldots = \mu_n = \mu$, o modelo ultraestrutural reduz-se ao modelo estrutural; ao passo que se $\sigma_X^2 = 0$, o modelo ultraestrutural reduz-se ao modelo funcional.

Por outro lado, segundo Carroll *et al.* (2006) a tradicional distinção entre modelo funcional e modelo estrutural, para algumas aplicações, não é tão relevante quanto a distinção entre *modelagem funcional* e *modelagem estrutural*, que facilita a escolha dos métodos de inferência baseada nas hipóteses assumidas a respeito da covariável.

De um modo geral a modelagem funcional pode ser vista como um conjunto de técnicas semiparamétricas. Mais especificamente, a modelagem funcional usa modelos paramétricos para a variável resposta, mas não assume hipótese alguma a respeito da distribuição da covariável não observada. Pode-se dizer que mesmo quando os X_i's representam uma amostra aleatória, a modelagem funcional é útil porque conduz a

procedimentos de estimação robustos, mesmo com a má especificação da distribuição dos X_i's.

Conforme Carroll *et al.* (2006), basicamente todos os pesquisadores de modelos com erros de medição chegaram à mesma conclusão: métodos baseados na função verossimilhança podem ter considerável valor, mas existe a possibilidade de inferências não robustas devido à má especificação da distribuição dos X_i's, que representa uma grande dificuldade desses métodos.

1.3 Objetivos e organização

Em um primeiro momento (Seção 2.2), serão avaliados os efeitos dos erros de medição na variável explicativa sobre alguns procedimentos inferenciais (bem estabelecidos na prática) quando há sobredispersão (Hinde & Demétrio, 1998a; Collett, 2003). Em um segundo passo, já incorporado o problema da sobredispersão à modelagem e diante de erros de medição, indicamos como obter inferências válidas. Essencialmente, o estudo de modelos nas condições das seções 2.2 e 2.4 compõem essa dissertação, sendo que nossa principal contribuição está na seção 2.4. Em um exemplo (Capítulo 4), trataremos de um experimento realizado em uma fábrica em que foi avaliado se cada uma das peças inspecionadas satisfaziam certa especificação. Um dos objetivos desse experimento consiste em estudar a associação entre o diâmetro de um furo de uma peça, denominado valor de referência, e a proporção de aceitação. Os métodos que serão utilizados para estimação dos parâmetros são os de máxima verossimilhança, quase-verossimilhança, máxima pseudo-verossimilhança, calibração da regressão e SIMEX. Também apresentaremos as estimativas do viés cometido na estimação dos parâmetros dos modelos.

Este trabalho está dividido em cinco capítulos, que compreendem a introdução, onde

fazemos uma abordagem geral a respeito dos modelos com erros de medição, técnicas inferencias utilizadas, estudo de simulações, aplicações e conclusão. No Capítulo 2 apresentamos os modelos binomial, com sobredispersão, beta-binomial e beta-binomial com erros de medição e os diversos métodos que serão usados nas estimativas dos parâmetros, bem como formas para determinar os erros padrão dessas estimativas. Ainda no Capítulo 2 apresentamos os testes de hipóteses usados no decorrer do trabalho. Os resultados obtidos em estudos de simulação compõem o Capítulo 3. Para complementar o trabalho, no Capítulo 4, apresentamos um exemplo ilustrativo com aplicações das técnicas expostas neste trabalho. Comentários gerais sobre essa dissertação e algumas propostas para trabalhos futuros compõem o Capítulo 5. Apresentamos ainda um Apêndice, que contém os elementos da função escore e da matriz de informação de Fisher do modelo beta-binomial.

Capítulo 2

Modelos e Inferência

Neste capítulo descreveremos o modelo binomial, que é a forma mais simples de modelar dados binários, modelo com sobredispersão, suas possíveis causas e formas de tratá-las. Apresentaremos o modelo beta-binomial, que pode ser usado para modelar a sobredispersão, e o modelo beta-binomial com erros de medição e também estimadores dos parâmetros desses modelos. Descrevemos o método SIMEX (Carroll *et al.*, 2006), suas características, aplicações e um método para cálculo dos erros padrão. Os testes de hipótese do parâmetro de sobredispersão e do coeficiente angular β_1 também são tratados neste capítulo.

Em muitos experimentos lidamos com situações em que um determinado item é analisado e dele obtemos como resposta "sucesso" ou "insucesso" (sucesso é o evento de interesse); portanto, a variável resposta é binária, isto é, admite apenas dois resultados. É comum encontrarmos situações práticas com esse tipo de variável resposta. Para ilustrar, citamos alguns exemplos recentes:

- Aplicações em economia podem ser encontradas em Verbeke & De Clercq (2006);
- A utilização de resposta binária em medicina (Draggalin & Fedorov, 2006);

- Kim *et al.* (2006) descrevem uma aplicação de resposta binária em biometria, e
- Dados binários também podem ser encontrados em computação, como em Li (2006).

2.1 Modelo binomial

A estrutura probabilística da variável de interesse (Y) pode ser representada pela distribuição binomial, ou seja,

$$Y_i|X_i = x_i \overset{\text{indep.}}{\sim} \mathsf{binomial}(m_i, \pi(x_i)), \tag{2.1}$$

e

$$g\big(\pi(x_i)\big) = \log\left(\frac{\pi(x_i)}{1-\pi(x_i)}\right) = \beta_0 + \beta_1\, x_i, \tag{2.2}$$

sendo que $g(\cdot)$ é a função de ligação e

$$\pi(x_i) = \frac{e^{\beta_0+\beta_1\, x_i}}{1+e^{\beta_0+\beta_1\, x_i}}, \quad i = 1, \ldots, n.$$

O modelo (2.1)-(2.2) constitui um caso particular de modelo linear generalizado (McCullagh & Nelder, 1989).

Sejam Y_i, $i = 1, \ldots, n$ o número de sucessos em m_i ensaios independentes, cada um com probabilidade de sucesso π_i. Dessa forma, $Y_i \sim \mathsf{binomial}(m_i, \pi_i)$. A função massa de probabilidade de Y_i é expressa da forma

$$f_{Y_i}(y_i) = \binom{m_i}{y_i} \pi_i^{y_i}(1-\pi_i)^{m_i-y_i}, \tag{2.3}$$

com $y_i = 0, \ldots, m_i$ e $i = 1, \ldots, n$. A esperança e a variância de Y_i são, respectivamente, $\mathrm{E}(Y_i) = \pi_i$ e $\mathrm{var}(Y_i) = m_i\pi_i(1-\pi_i)$. A esse modelo Williams (1982) denominou Tipo I.

Assumimos uma correlação constante e igual a ϕ entre Z_{ij} e Z_{ik}, $j \neq k$ de forma que

$$\mathrm{E}(Y_i) = m_i\, \pi_i$$

e

$$\begin{aligned}\mathrm{var}(Y_i) &= \sum_{j=1}^{m_i} \mathrm{var}(Z_{ij}) + \sum_{j=1}^{m_i} \sum_{k=1,k\neq j}^{m_i} \mathrm{cov}(Z_{ij}, Z_{ik}) \\ &= m_i\,\pi_i\,(1-\pi_i) + m_i\,(m_i-1)\,[\phi\,\pi_i\,(1-\pi_i)];\end{aligned}$$

logo,

$$\mathrm{var}(Y_i) = m_i\,\pi_i\,(1-\pi_i)[1+\phi\,(m_i-1)], \tag{2.4}$$

com $\pi_i = \pi(x_i)$, $i = 1,\ldots,n$, notando que $-1/(m_{(n)}-1) < \phi < 1$ e $m_{(n)} = \max\{m_1,\ldots,m_n\}$, ou seja, esta formulação cobre a possibilidade de uma correlação negativa (subdispersão). Se $\phi > 0$, a variância de Y_i ultrapassa a variância de uma variável aleatória com distribuição $\mathsf{binomial}(m_i,\pi_i)$. Se tivermos $m_1 = \cdots = m_n = m_0$, a expressão da variância se reduz a

$$\mathrm{var}(Y_i) = \; m_0\,\pi_i\,(1-\pi_i)[1+(m_0-1)\phi],$$

para $i = 1,\ldots,n$ obtendo-se o modelo com sobredispersão constante.

2.1.1 Máxima verossimilhança

Os processos iterativos utilizados nas simulações (Capítulo 3) tiveram o ponto inicial calculado com base em McCullagh & Nelder (1989) e Paula (2004), que mostram que as estimativas de máxima verossimilhança (MV) dos parâmetros $\boldsymbol{\theta}$ do modelo, com preditor linear η, podem sem calculadas pelo método de mínimos quadrados ponderados.

Para isso considera-se uma variável z, que desempenha o papel de uma variável dependente modificada e uma matriz de pesos $\mathbf{V} = \mathrm{diag}(v_1,\ldots,v_n)$, que muda a cada passo do processo iterativo. No caso do modelo logístico binomial $v_i = m_i\,\pi_i(1-\pi_i)$, e

$$z_i = \eta_i + \frac{(y_i - \eta_i\,\pi_i)}{m_i\,\pi_i(1-\pi_i)}, \quad i = 1,\ldots,n.$$

Esse processo pode ser resumido da seguinte forma:

1. Obtém-se uma estimativa inicial da proporção de sucessos $\left(\pi_i = \frac{y_i+0{,}5}{m_i+1}\right)$ das m_i repetições;

2. Neste passo tem-se $\eta_i = \log \dfrac{\pi_i}{1-\pi_i}$, como em (2.2), a matriz de pesos $\mathbf{V}$ e a variável modificada $\mathbf{z}$;

3. Estimam-se os parâmetros do modelo, utilizando a expressão

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^t\mathbf{V}\,\mathbf{X})^{-1}\mathbf{X}^t\mathbf{z}; \tag{2.5}$$

4. Recalcula-se o preditor linear η, utilizando a variável explicativa $\mathbf{X}$, os valores dos parâmetros obtidos em (2.5) e a expressão (2.2);

5. Obtém-se novamente a matriz de pesos $\mathbf{V}$ e a variável modificada $\boldsymbol{z}$;

6. Estima-se, novamente, $\widehat{\boldsymbol{\beta}}$, usando a expressão (2.5). Repetem-se os passos 4 - 6 até atingir a tolerância estipulada.

A matriz $\mathbf{I}(\widehat{\boldsymbol{\theta}}) = \mathbf{X}^t\,\mathbf{V}\,\mathbf{X}$ é um estimador da matriz de informação de Fisher, com

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}, \ \mathbf{V} = \text{diag}[m_1\hat{\pi}_1(1-\hat{\pi}_1), \ldots, m_n\hat{\pi}_n(1-\hat{\pi}_n)] \quad \text{e}$$

$\widehat{\boldsymbol{\Sigma}} = \mathbf{I}^{-1}(\widehat{\boldsymbol{\theta}})$ é uma aproximação da matriz de variâncias e covariâncias de $\widehat{\boldsymbol{\theta}}$ ($\widehat{\boldsymbol{\theta}} = (\widehat{\beta}_0, \widehat{\beta}_1)$ é o estimador de MV de $\boldsymbol{\theta}$).

2.2 Modelo com sobredispersão

Sobredispersão ou variação extra-binomial é um fenômeno comum que ocorre na modelagem de dados binários agrupados e cuja ocorrência é caracterizada quando a variação observada excede aquela assumida pelo modelo

(Hinde & Demétrio, 1998a). Especialmente em regressão logística, quando o desvio do modelo é maior do que os graus de liberdade, há indícios de sobredispersão.

Paula (2004) resume em apenas duas as possíveis causas da sobredispersão: correlação entre as réplicas binárias (que ocorre, por exemplo, quando uma peça é inspecionada várias vezes) ou variação entre as probabilidades de sucesso em um mesmo grupo. Menciona também que embora não seja tão simples a distinção entre os dois tipos de sobredispersão, os procedimentos estatísticos para tratá-las podem ser os mesmos.

Não levar em conta a existência de sobredispersão na análise dos dados pode levar à estimação incorreta dos erros padrão e, conseqüentemente, uma avaliação incorreta da significância individual dos parâmetros da regressão.

Encontramos na literatura diversas formas para resolver o problema da sobredispersão. Hinde & Demétrio (1998a) dividem essas abordagens em dois grupos:

1. assumir uma forma mais geral para a função de variância do modelo, possivelmente incluindo parâmetros adicionais;

2. assumir um modelo de dois estágios para a variável resposta, isto é, assumir que o parâmetro do modelo básico (por exemplo, a probabilidade de sucesso no modelo binomial) tem alguma distribuição de probabilidade.

Os modelos do tipo 1 podem não ter uma distribuição específica para a variável resposta, mas podem ser vistos como uma extensão do modelo básico. Para estimar os parâmetros da regressão pode-se usar quase-verossimilhança ou algum outro método que possibilite estimar o parâmetro adicional da função de variância.

Os modelos do tipo 2 são vistos como extensão dos modelos básicos e a estimativa dos parâmetros de regressão, em princípio, pode ser feita por máxima verossimilhança.

Um exemplo do modelo de modelo do tipo 2 é o modelo beta-binomial, que será

descrito a seguir.

2.3 Modelo beta-binomial

No modelo (2.1), supondo que $Y_i \,|\, P_i \overset{\text{indep.}}{\sim} \mathsf{binomial}(m_i, P_i)$ e $P_i \overset{\text{indep.}}{\sim} \mathsf{beta}(a_i, b_i)$, conclui-se que Y_i segue distribuição beta-binomial (Hinde & Demétrio, 1998a). Tomando $\mathrm{E}(P_i) = \pi_i = \frac{a_i}{a_i+b_i}$ e fazendo $c = a_i + b_i$ temos que $a_i = c\pi_i$ e $b_i = c(1-\pi_i)$, de forma que

$$f_{Y_i}(y_i) = \binom{m_i}{y_i} \frac{\Gamma(c\pi_i + y_i)\Gamma(c(1-\pi_i) + m_i - y_i)\Gamma(c)}{\Gamma(m_i + c)\Gamma(c\pi_i)\Gamma(c(1-\pi_i))}, \tag{2.6}$$

$y_i = 0, \ldots, m_i$ e $i = 1, \ldots, n$.

Lembrando que para $\alpha > 0$, $\Gamma(\alpha + 1) = \alpha\Gamma(\alpha)$, podemos reescrever o modelo beta-binomial, que está dividido em três casos:

1) $y_i \notin \{0,\, m_i\}$:

$$f_{Y_i}(y_i) = \binom{m_i}{y_i} \frac{\prod_{r=0}^{y_i-1}(c\pi_i + r) \prod_{s=0}^{m_i-y_i-1}(c[1-\pi_i] + s)}{\prod_{t=0}^{m_i-1}(c+t)};$$

2) $y_i = 0$:

$$f_{Y_i}(y_i) = \prod_{s=0}^{m_i-1}\left[\frac{c(1-\pi_i)+s}{c+s}\right]$$

e

3) $y_i = m_i$:

$$f_{Y_i}(y_i) = \prod_{t=0}^{y_i-1}\left[\frac{c\pi_i + t}{c+t}\right].$$

Fazendo $\phi = 1/(c+1)$, prova-se que a esperança e a variância de Y_i com distribuição beta-binomial são respectivamente, $\mathrm{E}(Y_i) = m_i\pi_i$ e $\mathrm{var}(Y_i) = m_i\pi_i(1-\pi_i)[1+\phi(m_i-1)]$, como em (2.4). A esse modelo Williams (1982) denominou Tipo II.

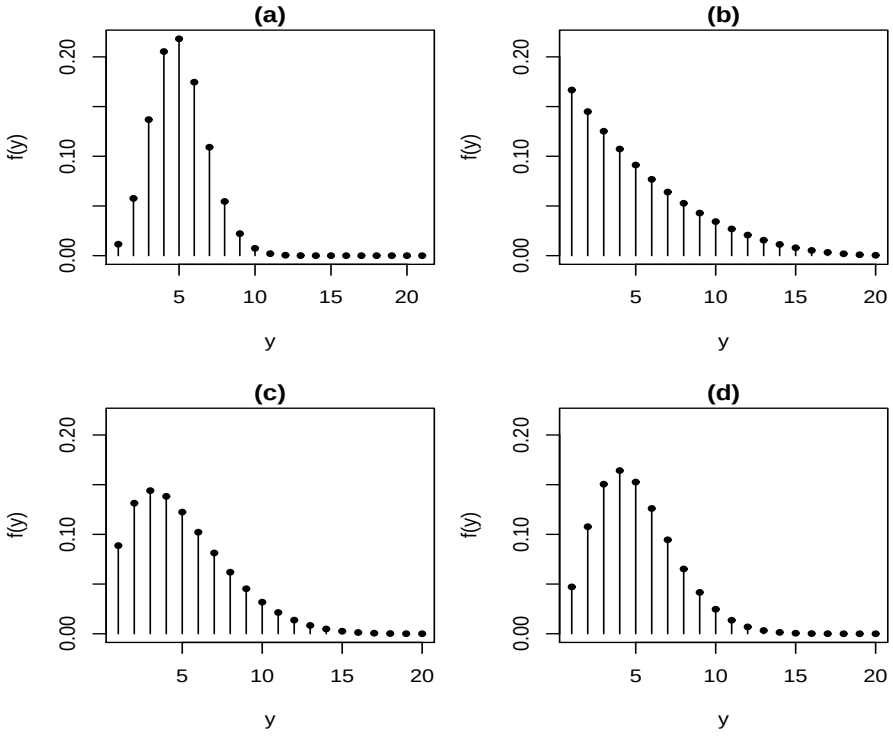

Figura 2.1: Função massa de probabilidade dos modelos binomial (a) e beta-binomial, (b), (c) e (d), com probabilidade de sucesso $\pi = 0,2$ e $m = 20$. No modelo beta-binomial os valores do parâmetro c são iguais a 5, 10 e 20, respectivamente.

Um caso especial do modelo beta-binomial é quando temos o parâmetro de sobredispersão $\phi = 0$, isto é, o modelo (2.6) se reduz ao modelo binomial (2.3).

Nas Figuras 2.1 e 2.2 podemos verificar o efeito da sobredispersão, pois à medida que o valor de c aumenta, ϕ se aproxima de zero e o modelo beta-binomial se aproxima do modelo binomial. Também podemos verificar uma distribuição menos concentrada dos valores da probabilidade. Essa variabilidade acarreta um aumento na variância do modelo, caracterizando a sobredispersão. Assim, pode-se concluir que o modelo beta-binomial é adequado para modelar sobredispersão.

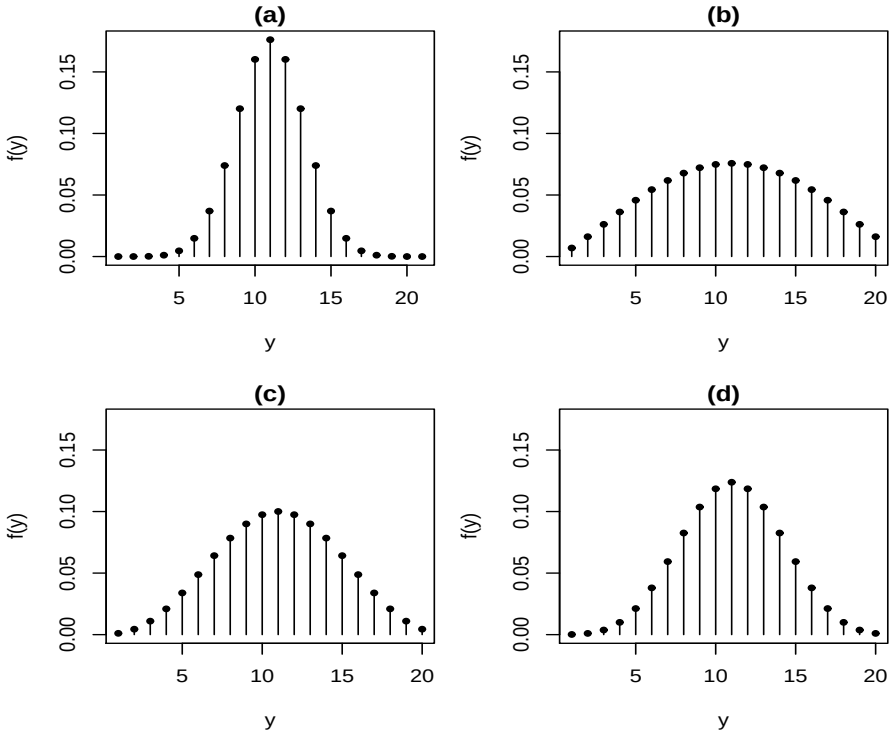

Figura 2.2: Função massa de probabilidade dos modelos binomial (a) e beta-binomial, (b), (c) e (d), com probabilidade de sucesso $\pi = 0,5$ e $m = 20$. No modelo beta-binomial os valores do parâmetro c são iguais a 5, 10 e 20, respectivamente.

2.3.1 Máxima verossimilhança

O método de máxima verossimilhança no modelo beta-binomial sem erros de medição será denominado neste trabalho de máxima verossimilhança ingênua (MVI).

O ajuste de um modelo é determinado pelo vetor $\widehat{\boldsymbol{\theta}}$ de estimativas dos parâmetros. Tendo uma distribuição para as observações (Y_i), a estimação dos parâmetros do modelo (β_0, β_1 e ϕ (ou c), denotados por $\boldsymbol{\theta}$) pode, em princípio, ser feita por máxima verossimilhança. Denotando por $l(\cdot)$ o logaritmo da função verossimilhança, que denominaremos log-verossimilhança, e ignorando a parte constante da distribuição beta-binomial, que não depende de π_i, podemos escrevê-la, no caso geral, como

1) $y_i \notin \{0, m_i\}$:

$$l(\boldsymbol{\theta}; y_i) = \sum_{r=0}^{y_i-1} \log(c\pi_i + r) + \sum_{s=0}^{m_i-y_i-1} \log(c(1-\pi_i) + s) - \sum_{t=0}^{m_i-1} \log(c + t).$$

2) $y_i = 0$:

$$l(\boldsymbol{\theta}; y_i) = \sum_{s=0}^{m_i-1} \{\log(c(1-\pi_i) + s) - \log(c + s)\}.$$

3) $y_i = m_i$:

$$l(\boldsymbol{\theta}; y_i) = \sum_{t=0}^{y_i-1} \{\log(c\pi_i + t) - \log(c + t)\}.$$

Portanto, a função log-verossimilhança do modelo beta-binomial pode ser escrita como

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

De uma forma resumida tem-se, a seguir, algumas propriedades do estimador $\widehat{\boldsymbol{\theta}}$, para um modelo com observações independentes, mas não identicamente distribuídas:
i) $\hat{\boldsymbol{\theta}} \xrightarrow{\mathrm{P}} \boldsymbol{\theta}$;

ii) $\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{\mathrm{D}} \mathsf{N}(\mathbf{0}, \mathbf{K}(\boldsymbol{\theta})^{-1})$, quando $n \to \infty$ e $\mathbf{K}(\boldsymbol{\theta})$ é a matriz de informação de Fisher esperada, com $\frac{1}{n}\sum_{i=1}^{n} \mathbf{K}_i(\boldsymbol{\theta}) \overset{n\to\infty}{\to} \mathbf{K}(\boldsymbol{\theta})$ e $\mathbf{K}_i(\boldsymbol{\theta})$ é a matriz de informação de Fisher esperada por observação, $i = 1, \ldots, n$.

A matriz de informação de Fisher esperada pode ser estimada por

$$\mathbf{K}(\widehat{\boldsymbol{\theta}}) = - \begin{bmatrix} \frac{\partial^2 l}{\partial c^2} & \frac{\partial^2 l}{\partial c \partial \beta_0} & \frac{\partial^2 l}{\partial c \partial \beta_1} \\ \frac{\partial^2 l}{\partial \beta_0 \partial c} & \frac{\partial^2 l}{\partial \beta_0^2} & \frac{\partial^2 l}{\partial \beta_0 \partial \beta_1} \\ \frac{\partial^2 l}{\partial \beta_1 \partial c} & \frac{\partial^2 l}{\partial \beta_1 \partial \beta_0} & \frac{\partial^2 l}{\partial \beta_1^2} \end{bmatrix}. \tag{2.7}$$

Os componentes da função escore e os elementos da matriz $\mathbf{K}(\widehat{\boldsymbol{\theta}})$ são apresentados no Apêndice A.

No modelo beta-binomial as estimativas iniciais de β_0 e β_1 foram obtidas por máxima verossimilhança no modelo binomial, como descrito na Subseção (2.1.1). Quanto ao parâmetro de sobredispersão (ϕ), sua estimativa inicial pode ser obtida através da expressão

$$\phi = \frac{X^2 - (n-p)}{\sum_{i=1}^{n}(m_i - 1)[1 - m_i\widehat{\pi}_i(1-\widehat{\pi}_i)h_i]}, \tag{2.8}$$

sendo que $h_i = \mathrm{var}(x_i^t\,\widehat{\beta}) = x_i^t(X^tVX)^{-1}x_i$ é a variância do preditor linear, X^2 é a estatística de Pearson do modelo binomial e p é o número de parâmetros do preditor linear. A expressão (2.8) deriva da igualdade entre a estatística X^2 de Pearson e a esperança dessa mesma estatística no modelo beta-binomial (Hinde & Demétrio, 1998a).

Neste trabalho as estimativas dos parâmetros são obtidas por um método denominado BFGS (Broyden - Fletcher - Goldfarb - Shanno), disponível em linguagem Ox (Doornik, 2002), que será utilizado em todos os métodos que requerem maximização. Este método faz parte de um conjunto de métodos utilizados para otimização de funções denominados quase-Newton. A teoria dos métodos quase-Newton é baseada no fato de que uma aproximação para a curvatura de uma função não-linear pode ser calculada sem forma explícita para a matriz hessiana.

A idéia fundamental é substituir a verdadeira inversa da matriz hessiana, utilizada nos métodos de Newton, por uma aproximação desta matriz sempre que se tornar muito difícil a obtenção da verdadeira inversa. Idealmente, essa aproximação converge para a inversa da matriz hessiana até a solução e o método completo se comporta como o método de Newton (Gill *et al.*, 1981). O método BFGS parece oferecer uma combinação de vantagens bastante atrativas para funções não-quadráticas, pois as direções geradas sempre são de crescimento da função dado que a matriz hessiana, no ponto de máximo, é definida negativa.

Torna-se importante destacar que, durante o processo de maximização no estudo de simulações, as amostras aleatórias nas quais o método BFGS não convergiu foram descartadas.

2.4 Modelo beta-binomial com erros de medição

Quando a variável explicativa (X) contém erros de medição, ou seja, observamos W ao invés de X e considerando o modelo estrutural, descrito em (1.1), a variável explicativa X tem distribuição normal com média μ_X e variância σ_X^2; $(W|X = x)$ tem distribuição normal com média x e variância σ_U^2, σ_U^2 conhecida e, além disso, consideramos que $(Y|X = x)$ tem distribuição beta-binomial (2.6). O descrito caracteriza um modelo beta-binomial com erros de medição. A seguir apresentaremos métodos de estimação dos parâmetros desse modelo.

2.4.1 Máxima verossimilhança

O primeiro método de estimação a ser tratado é o de máxima verossimilhança. A função log-verossimilhança correspondente à distribuição conjunta de ($\mathbf{Y}$, $\mathbf{W}$) tem a

forma

$$
\begin{aligned}
l(\boldsymbol{\theta}; \mathbf{Y}, \mathbf{W}) &= \sum_{i=1}^{n} \log[f_{Y,W}(y, w)] \qquad (2.9)\\
&= \sum_{i=1}^{n} \log \left[\int_{-\infty}^{\infty} f_{Y|X}(y|x; \phi, \beta_0, \beta_1)\, f_{W|X}(w|x; \sigma_U^2)\, f_X(x; \mu_X, \sigma_X^2) dx \right].
\end{aligned}
$$

A primeira componente desse integrando, $f_{Y|X}(y|x; \phi, \beta_0, \beta_1)$, é o modelo beta-binomial dado em (2.6); a segunda, $f_{W|X}(w|x; \sigma_U^2)$, é o modelo para $W|X$, que tem distribuição $\mathsf{N}(x, \sigma_U^2)$, σ_U^2 conhecida, e a última, $f_X(x; \mu_X, \sigma_X^2)$, é a função densidade da variável explicativa X, cuja distribuição é $\mathsf{N}(\mu_X, \sigma_X^2)$, conforme (1.1).

O cálculo de $l(\boldsymbol{\theta}; \mathbf{Y}, \mathbf{W})$ em (2.9) é bastante complexo e uma forma de contornar o problema é utilizar o método de quadratura gaussiana, disponível em linguagem Ox na biblioteca QuadPack (Piessens *et al.*, 1983).

Para o modelo binomial com erros de medição a função log-verossimilhança $l(\boldsymbol{\theta})$ é escrita como em (2.9) substituindo o modelo beta-binomial em $f_{Y|X}(y|x)$, pelo modelo binomial conforme (2.3). Essa função log-verossimilhança será utilizada nos testes de hipótese da sobredispersão, pois quando o parâmetro de sobredispersão (ϕ) é nulo o modelo beta-binomial se reduz ao modelo binomial.

As estimativas iniciais de β_0, β_1 e ϕ podem ser aquelas obtidas no modelo beta-binomial sem erros de medição, ou ainda as estimativas iniciais sugeridas para esse modelo, Subseção (2.3.1). Para os parâmetros μ_X e σ_X^2 as estimativas iniciais podem ser $\overline{W}$ e $S_W^2 - \sigma_U^2$, respectivamente, sendo

$$
\overline{W} = n^{-1} \sum_{i=1}^{n} W_i \quad \text{e} \quad S_W^2 = (n-1)^{-1} \sum_{i=1}^{n} (W_i - \overline{W})^2.
$$

2.4.2 Máxima pseudo-verossimilhança

Estimativas de máxima verossimilhança muitas vezes não são possíveis de obter devido a dificuldades computacionais ou algébricas. Uma alternativa é a utilização do

método de máxima pseudo-verossimilhança (MPV), proposto inicialmente por Gong & Samaniego (1981) e complementado por Parke (1986). Seja $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ e seja $l(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ a função log-verossimilhança em uma amostra de tamanho n. Seja ainda $\tilde{\boldsymbol{\theta}}_2$ uma estimativa de $\boldsymbol{\theta}_2$ obtida por um outro método diferente da MPV, essas estimativas $\tilde{\boldsymbol{\theta}}_2$ são substituídas no vetor $\boldsymbol{\theta}$, o que acarreta uma diminuição no número de parâmetros a ser estimado, pois $\boldsymbol{\theta}_2$ agora é considerado constante.

No modelo em estudo, beta-binomial com erros de medição, estrutural, temos $\boldsymbol{\theta} = (\phi, \beta_0, \beta_1, \mu_X, \sigma_X^2)$, $\boldsymbol{\theta}_1 = (\phi, \beta_0, \beta_1)$, $\boldsymbol{\theta}_2 = (\mu_X, \sigma_X^2)$, $\tilde{\boldsymbol{\theta}}_2 = (\overline{W}, S_W^2 - \sigma_U^2)$ e a função log-verossimilhança é a mesma da equação (2.9).

O método de máxima pseudo-verossimilhança se aplica, nesse caso, pois $\tilde{\boldsymbol{\theta}}_2$ é um estimador consistente de $\boldsymbol{\theta}_2$. Desta forma, as estimativas de pseudo-verossimilhança são calculadas maximizando a função $l(\boldsymbol{\theta}_1, \tilde{\boldsymbol{\theta}}_2)$ em relação a $\boldsymbol{\theta}_1$, com valores iniciais dos parâmetros em $\boldsymbol{\theta}_1$ obtidos como no modelo beta-binomial com erros de medição.

O estimador $\tilde{\boldsymbol{\theta}}_1$ tem propriedades que dependem das propriedades de $\tilde{\boldsymbol{\theta}}_2$ (Gong & Samaniego, 1981). Assim o estimador de MPV será consistente, sob certas condições de regularidade, e quando $\tilde{\boldsymbol{\theta}}_2$ o for. A eficiência de $\tilde{\boldsymbol{\theta}}_1$ depende da eficiência relativa de $\tilde{\boldsymbol{\theta}}_2$. A distribuição assintótica de $\tilde{\boldsymbol{\theta}}_1$ é verificada, sob certas condições de regularidade, quando

$$\sqrt{n}(\tilde{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}_2) \xrightarrow{\text{D}} \mathsf{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\tilde{\boldsymbol{\theta}}_2}), \quad \text{quando}\, n \to \infty.$$

2.4.3 Método de calibração da regressão

A idéia central do método de calibração da regressão consiste em substituir a variável não observável $\mathbf{X}$ por alguma função de $\mathbf{W}$. Após a substituição as estimativas dos parâmetros são obtidas por algum método confiável, por exemplo, máxima verossimilhança (Carroll *et al.*, 2006).

De forma resumida, podemos dizer que o método de calibração da regressão está organizado da seguinte forma:

i) Determinar $\mathbf{X}$ como função de $\mathbf{W}$, usando por exemplo, $\mathrm{E}(X|W)$. Além das suposições do modelo em (1.1), assumimos que a distribuição conjunta de $\mathbf{X}$ e $\mathbf{W}$ é

$$\begin{pmatrix} X \\ W \end{pmatrix} \sim \mathsf{N}_2 \left(\begin{pmatrix} \mu_X \\ \mu_X \end{pmatrix}, \begin{pmatrix} \sigma_X & \sigma_X^2 \\ \sigma_X^2 & \sigma_X^2 + \sigma_U^2 \end{pmatrix} \right);$$

então,

$$\mathrm{E}(X|W) = \mu_X + \frac{\sigma_X^2 (W - \mu_X)}{\sigma_X^2 + \sigma_U^2}.$$

ii) Substitui-se a variável não observada ($\mathbf{X}$) por $\mathrm{E}(X|W)$ e estimam-se os parâmetros do modelo beta-binomial (2.6);

iii) Estimam-se os erros padrão das estimativas pelo método de reamostragem *bootstrap*, por exemplo.

Os erros padrão das estimativas $\widehat{\boldsymbol{\theta}}$ resultantes do método de calibração da regressão podem ser estimados por reamostragem. Nesse trabalho utilizamos a técnica do *bootstrap* não-paramétrico, que consiste em obter uma nova amostra a partir de $\mathbf{W}$ e $\mathbf{Y}$, com reposição. Sendo $\widehat{\boldsymbol{\theta}}$ o vetor de estimativas obtidas pelo método de calibração da regressão e $\widehat{\boldsymbol{\theta}}^{(q)}$ a estimativa obtida por esse mesmo estimador a partir da q-ésima amostra *bootstrap*, $q = 1, \ldots, Q$, (Q é o número de amostras *bootstrap*), $\bar{\boldsymbol{\theta}}$ é a média de $\widehat{\boldsymbol{\theta}}^{(1)}, \ldots, \widehat{\boldsymbol{\theta}}^{(q)}$. Então, a matriz de covariâncias de $\widehat{\boldsymbol{\theta}}$ pode ser estimada por

$$\widehat{\mathrm{var}}(\widehat{\boldsymbol{\theta}}) = (Q-1)^{-1} \sum_{q=1}^{Q} (\widehat{\boldsymbol{\theta}}^{(q)} - \bar{\boldsymbol{\theta}})(\widehat{\boldsymbol{\theta}}^{(q)} - \bar{\boldsymbol{\theta}})^t.$$

No estudo de simulação que realizamos foram utilizadas $Q = 100$ réplicas *bootstrap*, conforme sugerido por Thoresen & Laake (2000).

2.4.4 Método de simulação e extrapolação (SIMEX)

Uma das conseqüências de ignorarmos erros de medição na variável explicativa (como se $W = X$) é a obtenção de estimadores inconsistentes, problema que pode ser contornado recorrendo a técnicas adequadas, como, por exemplo, o método SIMEX de Cook & Stefanski (1994) (vide também Carroll *et al.*, 2006). Este método tem como idéia central o fato de que os efeitos do erro de medição podem ser determinados experimentalmente via simulações, não sendo necessárias suposições a respeito da distribuição da covariável não observada X_i, $i = 1, \ldots, n$ (Carroll *et al.*, 2006).

Podemos, inicialmente, pensar em um modelo de regressão linear simples em que a variável explicativa é medida com erros. Supomos que o modelo de regressão é $\mathrm{E}(Y|X) = \beta_0 + \beta_1 X$ e que $W = X + U$, como em (1.1). A estimativa de β_1 da regressão de Y em W converge para $\beta_1 \, \sigma_X^2 \, /(\sigma_X^2 + \sigma_U^2)$.

Em um modelo de regressão linear e com a finalidade de corrigir estimadores ingênuos (assim denominados aqueles que desconsideram o erro de medição na variável explicativa), Cook & Stefanski (1994) consideraram um modelo em que a variância do erro de medição é $(1 + \lambda)\sigma_U^2$. A esperança do estimador de mínimos quadrados,

$$E(\widehat{\beta}_1) = \frac{\beta_1 \, \sigma_X^2}{\sigma_X^2 + (1 + \lambda)\sigma_U^2}, \tag{2.10}$$

é uma função de uma nova componente ($\lambda \geq 0$) e pretende-se a partir dela, mediante uma etapa denominada extrapolação obter um estimador não viesado. O estimador será não viesado quando tivermos $\lambda = -1$, o que é intuitivo e tem a finalidade de eliminar da expressão (2.10) a variância σ_U^2.

Concretamente, no modelo (2.1)-(2.2)-(1.2)-(1.1), o método requer a geração de conjuntos de dados com erros de medida

$$W_{s,i}(\lambda) = W_i + \lambda^{1/2} U_{s,i}, \quad s = 1, \ldots, S, \quad i = 1, \ldots, n, \tag{2.11}$$

em que $U_{s,i}$ são mutuamente independentes, independentes dos dados observados e têm distribuição independente e identicamente distribuída, normal com média zero e variância σ_U^2, que é conhecida em nosso trabalho.

Para cada pseudo-amostra $\{W_{s,i}(\lambda), i = 1, \ldots, n\}$, obtemos as estimativas dos parâmetros do modelo pelo método de máxima verossimilhança. Estas estimativas podem ser resumidas utilizando-se a média (ou outra medida) desses valores.

Após a etapa de simulação, as estimativas obtidas (viesadas), são representadas em um gráfico como função da variância adicionada, permitindo estabelecer uma relação entre o viés e a variância do erro de medição.

A escolha do número de simulações (S) e do número de diferentes λ's (L), bem como a função de ajuste guia-se pela prática de uso do método SIMEX. Valores tais como $S = 1000$, $\lambda_L = 2$ e $L = 5$ já foram relatados. O gráfico da Figura 2.3 ilustra uma aplicação do método SIMEX. Como no conjunto de dados que motivou esse trabalho a variância dos erros de medição (σ_U^2) pode ser considerada conhecida, o método SIMEX é adequado. Vale ressaltar que a etapa de simulação é bastante informativa, pois revela diretamente os efeitos dos erros de medição sobre estimativas que ignoram estes erros.

Ao simularmos $W_{s,i}(\lambda)$ a variância total do erro adicionado é $\sigma_U^2\,(1 + \lambda)$, pois

$$W_{s,i}(\lambda) = W_i + \lambda^{1/2}\, U_{s,i} \; = X_i + U_i + \lambda^{1/2}\, U_{s,i},$$

$$\text{var}(U_i + \lambda^{1/2}\, U_{s,i}) = \sigma_U^2 \; + \lambda\, \sigma_U^2 \; = \sigma_U^2\,(1 + \lambda).$$

Podemos descrever o método SIMEX de um modo mais formal. Desta forma supomos a existência de um procedimento de estimação que leva de um conjunto de dados a um espaço paramétrico. Seja $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, espaço paramétrico, T o funcional que leva o conjunto de dados em $\boldsymbol{\Theta}$. Então, podemos definir os seguintes estimadores:

$$\widehat{\boldsymbol{\theta}}_{\text{verdadeiro}} = T(\boldsymbol{Y}, \boldsymbol{X}) \quad \text{e} \quad \widehat{\boldsymbol{\theta}}_{\text{ingênuo}} = T(\boldsymbol{Y}, \boldsymbol{W}).$$

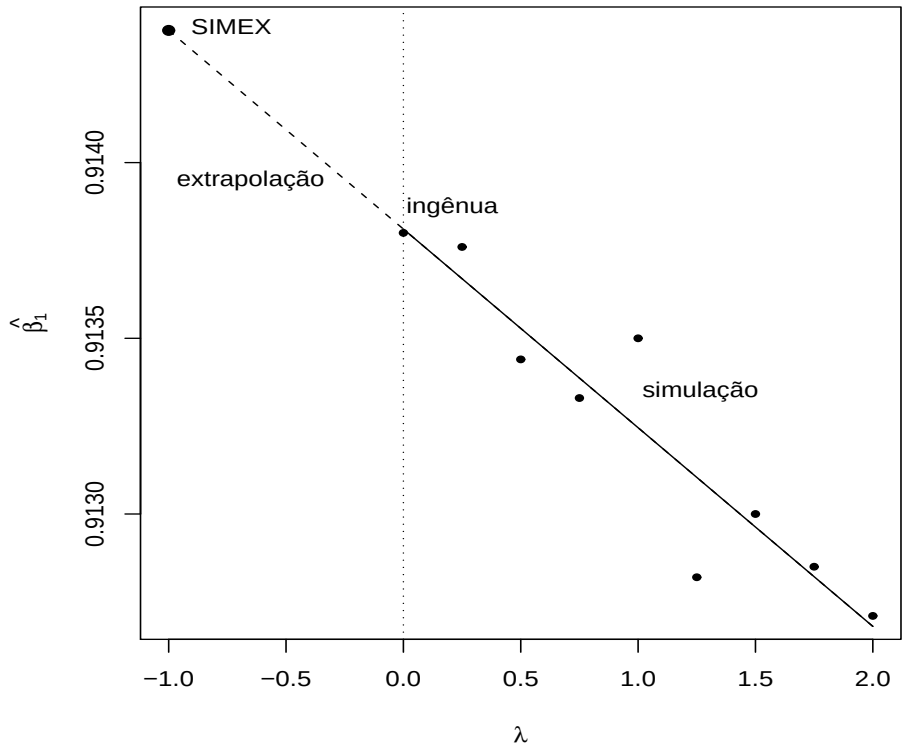

Figura 2.3: Estimação de um parâmetro pelo método SIMEX.

Como o estimador $\widehat{\boldsymbol{\theta}}_{\text{verdadeiro}}$ depende de $\mathbf{X}_i$ que é desconhecido, este não é um estimador no sentido estrito. Considerando $\lambda > 0$, definimos uma nova variável $W_{s,i}$, como em (2.11).

Definimos

$$\widehat{\boldsymbol{\theta}}_s(\lambda) = T(\mathbf{Y}, \mathbf{W}_s(\lambda)) \quad \text{e} \quad \widehat{\boldsymbol{\theta}}(\lambda) = E(\widehat{\boldsymbol{\theta}}_s(\lambda)|\{\mathbf{Y}, \mathbf{W}\}),$$

em relação à distribuição de $U_{s,i}$, sendo que $\widehat{\boldsymbol{\theta}}_s$ designa o estimador SIMEX de $\boldsymbol{\theta}$ correspondente à s-ésima amostra.

Note que $\widehat{\boldsymbol{\theta}}(0) = \widehat{\boldsymbol{\theta}}_s(0) = \widehat{\boldsymbol{\theta}}_{\text{ingênuo}}$. A estimação exata de $\widehat{\boldsymbol{\theta}}(\lambda)$, para $\lambda > 0$, geralmente não é factível, mas é possível fazer uma aproximação. Para isso deve-se simular $W_{s,i}$ como em (2.11), para um determinado S, calcular $\widehat{\boldsymbol{\theta}}(\lambda)$, como a média das estimativas $\widehat{\boldsymbol{\theta}}_s(\lambda)$, $s = 1, \ldots, S$. Essa etapa é denominada simulação.

Na estapa de extrapolação o objetivo é modelar $\widehat{\boldsymbol{\theta}}(\lambda)$, como uma função de λ, $\lambda \geq 0$,

e usar um modelo de extrapolação para $\lambda = -1$, que corresponde à ausência de erro de medição. O resultado da extrapolação é representado por $\widehat{\theta}_{\text{SIMEX}}$.

Apresentamos resumidamente os principais passos do método SIMEX:

1. Selecionamos um vetor $\boldsymbol{\Lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_L)$, com $0 = \lambda_1 < \lambda_2 < \ldots < \lambda_L$. Os elementos de $\boldsymbol{\Lambda}$ controlam o erro adicionado às observações e $\lambda = -1$ conduz às estimativas corrigidas no último passo;

2. Para cada λ, geram-se S conjuntos de números aleatórios $U_{s,i} \overset{\text{iid}}{\sim} \mathsf{N}(0, \sigma_U^2)$, independentes de W_i, $s = 1, \ldots, S$, $i = 1, \ldots, n$;

3. Calculamos $W_{s,i}(\lambda) = W_i + \lambda^{1/2} U_{s,i}$, para cada $\lambda > 0$ fixado;

4. Para cada pseudo-amostra $\{W_{s,i}(\lambda), i = 1, \ldots, n\}$ estimamos os parâmetros do modelo beta-binomial (2.6), resultando em $\widehat{\boldsymbol{\theta}}_s(\lambda)$, $s = 1, \ldots, S$;

5. Tomamos a média (ou outra medida resumo) das componentes de $\widehat{\boldsymbol{\theta}}_s(\lambda)$, $s = 1, \ldots, S$;

6. Na etapa de extrapolação, ajustamos um dos modelos – $\mathcal{G}_L(\lambda, \boldsymbol{\Gamma}) = \gamma_1 + \gamma_2\lambda$ (linear), $\mathcal{G}_Q(\lambda, \boldsymbol{\Gamma}) = \gamma_1 + \gamma_2\lambda + \gamma_3\lambda^2$ (quadrático) ou $\mathcal{G}_{RL}(\lambda, \boldsymbol{\Gamma}) = \gamma_1 + \frac{\gamma_2}{\gamma_3 + \lambda}$ (racional linear) – aos pontos obtidos na etapa de simulação e para cada componente de $\widehat{\boldsymbol{\theta}}$ (Carroll *et al.*, 2006);

7. Calculamos $\mathcal{G}(-1, \boldsymbol{\Gamma})$ para cada componente de $\widehat{\boldsymbol{\theta}}$ obtendo $\widehat{\boldsymbol{\theta}}_{\text{SIMEX}}$.

Os erros padrão das estimativas SIMEX podem ser calculados por uma abordagem de equações de estimação.

De acordo com Carroll *et al.* (1996), o método SIMEX produz estimadores aproximadamente consistentes. Podemos obter resultados não-viesados para o estimador

SIMEX usando a abordagem de equações de estimação da forma

$$\sum_{i=1}^{n} \boldsymbol{\Psi}(Y_i, X_i, \boldsymbol{\theta}) = \mathbf{0}.$$

No modelo em estudo (beta-binomial) $\Psi(\centerdot)$ é a função escore que está descrita no Apêndice A. Na etapa de simulação, para s fixado, a teoria assintótica mostra que

$$\sqrt{n}\{\widehat{\boldsymbol{\theta}}_{\mathrm{s}}(\lambda) - \boldsymbol{\theta}(\lambda)\} \approx -\boldsymbol{\mathcal{A}}^{-1}\{\sigma_U^2, \lambda, \boldsymbol{\theta}(\lambda)\} n^{-1/2} \sum_{i=1}^{n} \boldsymbol{\Psi}(\mathbf{Y}_i, \mathbf{W}_{\mathrm{s,i}}(\lambda), \boldsymbol{\theta}(\lambda)), \qquad (2.12)$$

com

$$\boldsymbol{\mathcal{A}}(\sigma_U^2, \lambda, \boldsymbol{\theta}(\lambda)) = \mathrm{E}\left(\frac{\partial}{\partial \boldsymbol{\theta}} \boldsymbol{\Psi}(\mathbf{Y}, \mathbf{W}_{\mathrm{s,i}}, \boldsymbol{\theta})\right).$$

No modelo beta-binomial $\dfrac{\partial}{\partial \boldsymbol{\theta}} \boldsymbol{\Psi}(\centerdot)$ é a matriz de informação de Fisher, descrita no Apêndice A.

Fazendo

$$\boldsymbol{\chi}_{\mathrm{S,i}}(\sigma_U^2, \lambda, \boldsymbol{\theta}(\lambda)) = S^{-1} \sum_{s=1}^{S} \boldsymbol{\Psi}(\mathbf{Y}_i, \mathbf{W}_{\mathrm{s,i}}(\lambda), \boldsymbol{\theta}(\lambda))$$

e dividindo por S a equação (2.12) temos a aproximação assintótica:

$$\sqrt{n}\{\widehat{\boldsymbol{\theta}}(\lambda) - \boldsymbol{\theta}(\lambda)\} \approx -\boldsymbol{\mathcal{A}}^{-1}(\sigma_U^2, \lambda, \boldsymbol{\theta}(\lambda))\, n^{-1/2} \sum_{i=1}^{n} \boldsymbol{\chi}_{\mathrm{S,i}}(\sigma_U^2, \lambda, \boldsymbol{\theta}(\lambda)). \qquad (2.13)$$

As parcelas de $\boldsymbol{\chi}_{\mathrm{S,i}}(\centerdot)$ são independentes e identicamente distribuídos com média zero.

Seja $\boldsymbol{\Lambda} = \{\lambda_1, \ldots, \lambda_L\}$ o conjunto de valores de λ usados nas simulações, como já mencionado, e $\widehat{\boldsymbol{\theta}}_*(\boldsymbol{\Lambda}) = \{\widehat{\boldsymbol{\theta}}^t(\lambda_1), \ldots, \widehat{\boldsymbol{\theta}}^t(\lambda_L)\}^t$, $\lambda \in \boldsymbol{\Lambda}$, vetor das estimativas dos parâmetros. Definimos

$$\boldsymbol{\Psi}_{\mathrm{S,i(1)}}(\sigma_U^2, \boldsymbol{\Lambda}, \boldsymbol{\theta}_*(\boldsymbol{\Lambda})) = \mathrm{vec}[\boldsymbol{\chi}_{\mathrm{S,i}}(\sigma_U^2, \lambda, \boldsymbol{\theta}(\lambda)),\ \lambda \in \boldsymbol{\Lambda}],$$

$$\boldsymbol{\mathcal{A}}_{11}(\sigma_U^2, \boldsymbol{\Lambda}, \boldsymbol{\theta}_*(\boldsymbol{\Lambda})) = \mathrm{diag}[\boldsymbol{\mathcal{A}}(\sigma_U^2, \lambda, \boldsymbol{\theta}(\lambda)),\ \lambda \in \boldsymbol{\Lambda}]$$

e

$$\boldsymbol{\mathcal{C}_{11}}\{\sigma_U^2, \boldsymbol{\Lambda}, \boldsymbol{\theta}_*(\boldsymbol{\Lambda})\} = \mathrm{Cov}[\boldsymbol{\Psi}_{\mathrm{S,1(1)}}\{\sigma_U^2, \boldsymbol{\Lambda}, \boldsymbol{\theta}_*(\boldsymbol{\Lambda})\}].$$

Usando a equação (2.13) temos que $\sqrt{n}\{\widehat{\boldsymbol{\theta}}_*(\boldsymbol{\Lambda}) - \boldsymbol{\theta}_*(\boldsymbol{\Lambda})\} \approx N(\mathbf{0}, \boldsymbol{\Sigma})$, com

$$\boldsymbol{\Sigma} = \boldsymbol{\mathcal{A}}_{11}^{-1}(.)\boldsymbol{\mathcal{C}}_{11}(\sigma_U^2, \boldsymbol{\Lambda}, \boldsymbol{\theta}_*(\boldsymbol{\Lambda}))\{\boldsymbol{\mathcal{A}}_{11}^{-1}(.)\}^t. \tag{2.14}$$

Define-se ainda $\boldsymbol{\mathcal{G}}^*(\boldsymbol{\Lambda}, \boldsymbol{\Gamma}^*)$, modelo ajustado para cada um dos parâmetros $\widehat{\boldsymbol{\theta}}(\lambda)$ e $\mathbf{R}(\boldsymbol{\Gamma}^*) = \widehat{\boldsymbol{\theta}}_*(\boldsymbol{\Lambda}) - \boldsymbol{\mathcal{G}}^*(\boldsymbol{\Lambda}, \boldsymbol{\Gamma}^*)$ resíduo do modelo ajustado.

Da soma dos quadrados dos resíduos do modelo, $\mathbf{R}(\boldsymbol{\Gamma}^*)^t\mathbf{R}(\boldsymbol{\Gamma}^*)$, obtemos $\widehat{\boldsymbol{\Gamma}}^*$. Essas estimativas são obtidas utilizando a equação de estimação $\mathbf{s}(\boldsymbol{\Gamma}^*)^t\mathbf{R}(\boldsymbol{\Gamma}^*) = \mathbf{0}$, onde $\mathbf{s}(\boldsymbol{\Gamma}^*) = \dfrac{\partial}{\partial(\boldsymbol{\Gamma}^*)}\,\mathbf{R}(\boldsymbol{\Gamma}^*)$.

A teoria assintótica mostra que $n^{-\frac{1}{2}}(\widehat{\boldsymbol{\Gamma}}^* - \boldsymbol{\Gamma}) \approx \mathsf{N}\{\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\Gamma}^*)\}$, com

$$\boldsymbol{\Sigma}(\boldsymbol{\Gamma}^*) = \mathbf{s}(\boldsymbol{\Gamma}^*)^{-t}\,\boldsymbol{\Sigma}\,\mathbf{s}^{-1}(\boldsymbol{\Gamma}^*)$$

e $\boldsymbol{\Sigma}$ é dada em (2.14). A notação $\mathbf{s}(\cdot)^{-t}$ designa a transposta da inversa da matriz $\mathbf{s}$.

Considerando que $\widehat{\boldsymbol{\theta}}_{\text{SIMEX}} = \boldsymbol{\mathcal{G}}^*(-1, \widehat{\boldsymbol{\Gamma}}^*)$, que são as estimativas dos parâmetros do modelo obtidas na ausência de erro de medição, a teoria assintótica nos mostra que $\sqrt{n}\{\widehat{\boldsymbol{\theta}}_{\text{SIMEX}} - \boldsymbol{\theta}\} \approx \mathsf{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\widehat{\boldsymbol{\theta}}_{\text{SIMEX}}})$, $\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\theta}}_{\text{SIMEX}}} = \boldsymbol{\mathcal{G}}^*_{\boldsymbol{\Gamma}^*}(-1, \boldsymbol{\Gamma}^*)\boldsymbol{\Sigma}(\boldsymbol{\Gamma}^*)\{\boldsymbol{\mathcal{G}}^*_{\boldsymbol{\Gamma}^*}(-1, \boldsymbol{\Gamma}^*)\}^t$, com $\boldsymbol{\mathcal{G}}^*_{\boldsymbol{\Gamma}^*}(\lambda, \boldsymbol{\Gamma}^*) = \dfrac{\partial}{\partial(\boldsymbol{\Gamma}^*)^t}\,\boldsymbol{\mathcal{G}}^*(\lambda, \boldsymbol{\Gamma}^*)$. A matriz de covariâncias assintótica, $\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\theta}}_{\text{SIMEX}}}$, é obtida pelo método delta (Sen & Singer, 1993).

2.5 Testes de hipóteses

A fim de testarmos a existência da sobredispersão realizamos testes de hipóteses confrontando o modelo com sobredispersão (beta-binomial) e o modelo sem sobredispersão (binomial). A estatística de teste utilizada é a da razão de verossimilhanças (RV) (Paul *et al.*, 1989). Testar a sobredispersão é equivalente a testarmos $\mathrm{H}_0 : \phi = 0$ contra $\mathrm{H}_1 : \phi > 0$. Sejam $l_0 = l(\beta_0, \beta_1, \mu_X, \sigma_X^2)$ e $l_1 = l(\phi, \beta_0, \beta_1, \mu_X, \sigma_X^2)$ as funções log-verossimilhança do modelo beta-binomial com erros de medição e do modelo binomial

com erros de medição, respectivamente. A estatística do teste (do logaritmo) da razão de verossimilhanças é dada por

$$\xi_{RV} = 2\,(l_1 - l_0).$$

Sob certas condições, sob a hipótese nula a estatística ξ_{RV} geralmente tem distribuição assintótica qui-quadrado com 1 grau de liberdade. No entanto, $\phi = 0$ está na fronteira do espaço paramétrico e quando isso acontece a teoria assintótica usual falha. Assim, a distribuição assintótica apropriada para essa estatística sob a hipótese nula é uma distribuição mista que tem função massa de probabilidade igual a $\frac{1}{2}$ em 0 e função densidade de uma variável $\frac{1}{2}\chi^2_{(1)}$ para os demais valores. O teste proposto por Paul *et al.* (1989) não considera a existência de erros de medição na covariável, porém estamos considerando que a variável explicativa contém erros de medição, então o teste que utilizamos é uma adaptação do teste proposto por Paul *et al.* (1989).

No sentido de testar a existência da sobredispersão foram utilizadas outras estatísticas de teste, como o teste C(α) que também foi proposto por Paul *et al.* (1989) e ainda os testes propostos por Kim & Margolin (1992) e Paula & Artes (2000), mas deixamos de apresentar os resultados obtidos pois não foram satisfatórios. Também tentamos uma estratégia utilizando o método SIMEX para a detecção da sobredispersão, que também se mostrou infrutífera.

Realizamos testes de hipóteses sobre o parâmetro β_1 em (2.2), das hipóteses $\mathrm{H}_0 : \beta_1 = \beta_{1,0}$ contra $\mathrm{H}_1 : \beta_1 \neq \beta_{1,0}$. As estatísticas de teste utilizadas foram a da razão de verossimilhanças (RV), Wald e escore (Paula, 2004) que têm as seguintes equações:

1. Razão de verossimilhanças:

$$\xi_{RV} = 2\,(l_1 - l_0);$$

2. Wald:

$$\xi_W = \frac{(\widehat{\beta}_1 - \beta_{1,0})^2}{\widehat{\mathrm{var}}(\widehat{\beta}_1)};$$

3. Escore:

$$\xi_{SR} = \left[\frac{\partial l_0}{\partial \beta_1}\right]^2 \widehat{\text{var}}_0(\widehat{\beta}_1),$$

sendo que l_0 é a função log-verossimilhança e $\widehat{\text{var}}_0(\widehat{\beta}_1)$ denota a variância assintótica de $\widehat{\beta}_1$, ambas sob H_0.

Sob a hipótese nula as estatísticas de teste (RV, Wald e escore) têm distribuição chi-quadrado com um grau de liberdade.

Importante ressaltar que no modelo beta-binomial com erros de medição a função escore bem como a matriz de informação de Fisher foram calculadas numericamente.

Os resultados obtidos a partir dos testes de hipóteses mencionados são apresentados no Capítulo 3.

Capítulo 3

Simulações

No Capítulo 2 apresentamos os seguintes métodos de estimação: máxima verossimilhança ingênua (MVI) e com erros (MVE), máxima pseudo-verossimilhança (MPV), calibração da regressão (CR) e SIMEX. Tendo em vista comparar esses métodos de estimação dos parâmetros do modelo, quanto à correção do viés e ao erro quadrático médio (EQM) das estimativas de β_0 e β_1, realizamos um estudo de simulação.

Os estudos de simulações deste trabalho foram desenvolvidos em linguagem Ox (Doornik, 2002) e os gráficos apresentados foram elaborados em R (R Development Core Team, 2006).

3.1 Cenário 1

Neste cenário consideramos que os verdadeiros valores dos parâmetros são $\beta_0 = 0$, $\beta_1 = 2$, $c = 5,5 (\phi = 0,15)$ e $X \sim \mathsf{N}(0,1)$. Esses valores foram escolhidos a partir de exemplos relatados na literatura consultada. Utilizamos 1000 réplicas de amostras de tamanhos 10, 20, 50 e 100. Também nos interessa observar o comportamento dos diversos métodos citados à medida que aumentamos a variância do erro de medição

(σ_U^2). Portanto, nesse estudo de simulação utilizamos quatro diferentes variâncias (0; 0,1; 0,2; 0,3). No método SIMEX as simulações foram realizadas com $S = 100$ réplicas SIMEX e $\mathbf{\Lambda} = (0;\ 0{,}5;\ 1;\ 1{,}5;\ 2)$. A medida resumo das estimativas SIMEX dos parâmetros foi a mediana e o modelo linear foi usado na etapa de extrapolação, por que apresentou um bom ajuste aos conjuntos de dados simulados e também por sua simplicidade.

Apresentamos as Figuras 3.1 - 3.4, representando as estimativas de β_1, para complementar os resultados apresentados nas Tabelas 3.1 - 3.4. Podemos observar que as estimativas desse parâmetro se aproximam do valor verdadeiro à medida que aumentamos o tamanho da amostra, especialmente para os métodos de MVE e CR com $n \neq 10$. Percebemos que para amostras pequenas ($n = 10$), o método SIMEX tem melhor desempenho, apresentando menor viés. Os métodos de MVI e MPV foram os que apresentaram os piores desempenhos para todos os tamanhos de amostra simulados. Também observamos nas estimativas forte influência do erro de medição, dado que quanto maior a variância (σ_U^2) mais distantes são as estimativas do valor verdadeiro, o que pode ser observado para $\widehat{\beta}_1$ em todos os métodos, mas especialmente para MVI, SIMEX e MPV. Constatamos que o método SIMEX corrige o viés das estimativas do parâmetro se comparado ao MVI, porém não o corrige totalmente, mesmo para tamanhos de amostras maiores ($n = 100$, por exemplo).

Conforme esperado, à medida que aumentamos o tamanho da amostra, os estimadores de máxima verossimilhança mostraram melhor comportamento, principalmente em termos de viés simulado. O efeito do aumento da variância do erro de medição também pode ser observado nas tabelas.

As Tabelas 3.1 - 3.4 contêm os resultados obtidos nas simulações. O viés simulado do estimador é obtido pela diferença entre a média das estimativas do parâmetro e o valor verdadeiro. Não apresentamos o viés simulado do estimador $\widehat{\beta}_0$ pois este é igual à

média das estimativas. O erro quadrático médio simulado (EQM), que também consta destas tabelas é calculado pela expressão

$$EQM = \frac{1}{R}\sum_{i=1}^{R}(\widehat{\theta}_i - \theta)^2,$$

em que R designa o número de réplicas amostrais e θ é um parâmetro do modelo. Não apresentamos as estimativas do método de CR, com $n = 10$ e $\sigma_U^2 = 0,3$ pois este se mostrou bastante instável em situação adversa.

Conforme esperado, pode-se notar que à medida que aumentamos o tamanho da amostra, as estimativas dos parâmetros dos modelos pelos métodos MVE e CR se aproximam dos valores verdadeiros. No método SIMEX as estimativas dos parâmetros se afastaram dos valores verdadeiros à medida que aumentamos o tamanho da amostra, o que também ocorreu com o método MVI. O método MPV teve uma pequena correção do viés quando aumentamos o tamanho da amostra. Comparando os métodos MVE e CR observamos que MVE apresentou menor viés absoluto e CR menor EQM.

As estimativas do parâmetro β_1 com $\sigma_U^2 = 0,3$ utilizando os métodos MVI, MVE, SIMEX e MPV são apresentadas nas Figuras 3.5 - 3.8 e para o método de CR estão expostas na Figura 3.9. Optamos por apresentar os histogramas das estimativas do método de calibração da regressão separado dos demais métodos usados porque para amostras pequenas ($n = 10$ e 20) foi o método que forneceu estimativas pontuais mais distantes do valor verdadeiro, principalmente para $\sigma_U^2 = 0,3$, que é a situação mais desfavorável. Por esse mesmo motivo, deixamos de apresentar o histograma para $n = 10$. Nesses histogramas verificamos que as estimativas pontuais são mais dispersas no método de MVE, para $n = 10$, e no método de MPV para os demais tamanhos da amostra. Apesar da dispersão das estimativas, podemos observar que a maior concentração delas está em torno do valor verdadeiro ($\beta_1 = 2$), especialmente para os métodos MVE e CR. A simetria das estimativas pode ser melhor observada para amostras maiores ($n = 50$ e 100) e para o método de MVE. O método de CR apresenta

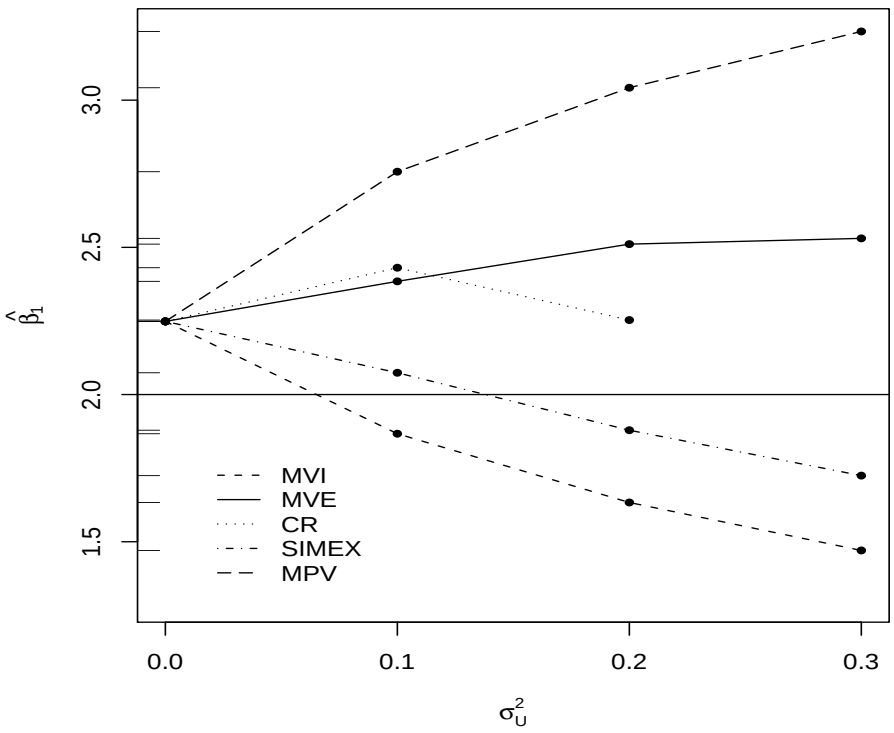

Figura 3.1: Média das estimativas de β_1 segundo diversos métodos, $n = 10$ (Cenário 1).

grande dispersão das estimativas, especialmente para amostras pequenas (por exemplo, $n = 20$), conforme gráfico (a) da Figura 3.9 e apresenta também estimativas muito distantes do valor verdadeiro, por exemplo, $\widehat{\beta}_1 = 1055$ para $n = 10$, em certas amostras. A simetria pode ser melhor observada para $n = 100$. Importante pontuar que esses histogramas foram obtidos a partir das estimativas do modelo para o maior valor de variância ($\sigma_U^2 = 0,3$) que é a situação mais desfavorável.

Pela análise anterior podemos concluir que os melhores resultados foram obtidos com os métodos MVE e CR e que MVE foi o melhor método para correção do viés (lembrando que tais estimadores são consistentes). Thoresen & Laake (2000) destacam que o método CR é uma boa alternativa para correção do viés se comparado ao método ingênuo (que desconsidera os erros de medição na variável explicativa) e ao estimador de máxima verossimilhança.

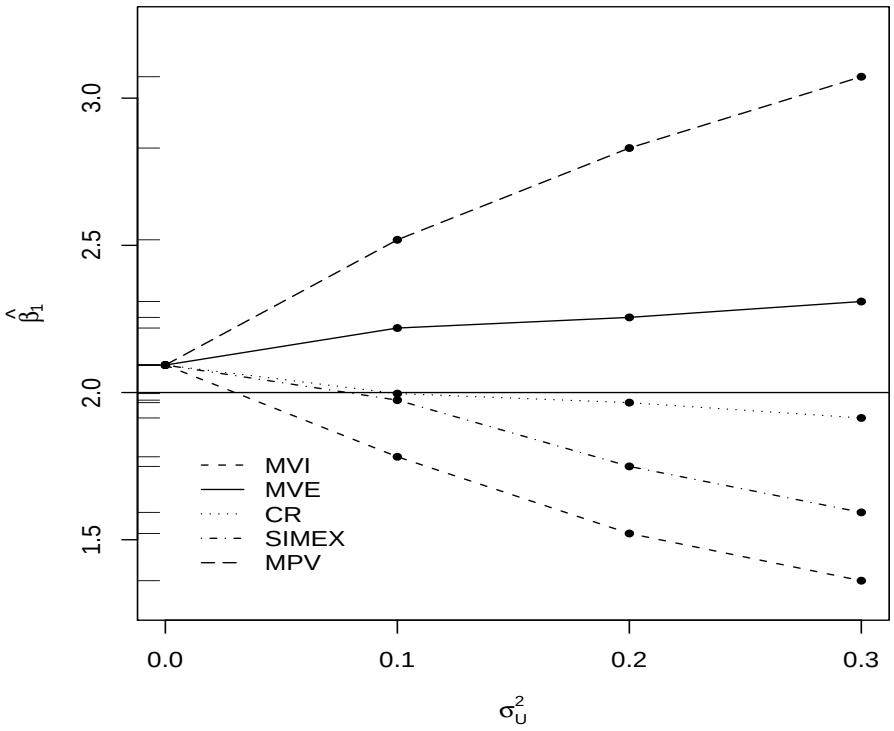

Figura 3.2: Média das estimativas de β_1 segundo diversos métodos, $n = 20$ (Cenário 1).

Os resultados dos testes de hipóteses da sobredispersão, que estão descritos no Capítulo 2, Seção 2.5, são apresentados na Tabela 3.5. Os resultados mostram que o teste RV foi bastante conservador, com taxas de rejeição inferiores ao nível de significância ($\alpha = 0,05$) especialmente para as amotras $n = 20$ e $n = 50$.

As estatísticas de teste (RV, Wald e escore) para o teste do coeficiente angular ($\beta_1 = 0$), obtidas conforme descrito na Seção 2.5, são apresentadas na Tabela 3.6. Importante observar que os valores de variâncias utilizados nas simulações das estatísticas de teste (0,1; 0,2; 0,3; 0,4), diferem um pouco daqueles usados nas Tabelas 3.1 - 3.4. Naquele caso não utilizamos $\sigma_U^2 = 0,4$, pois o tempo computacional aumentaria consideravelmente. Também deixamos de usar $\sigma_U^2 = 0$ pois corresponde à ausência de erro de medição e incorporar os efeitos dos erros de medição é um dos objetivos destas estatísticas de teste. Observamos que os valores da variância do erro de medição (σ_U^2) influenciaram todas as estatísticas de teste, aumentando ou diminuindo as taxas de re-

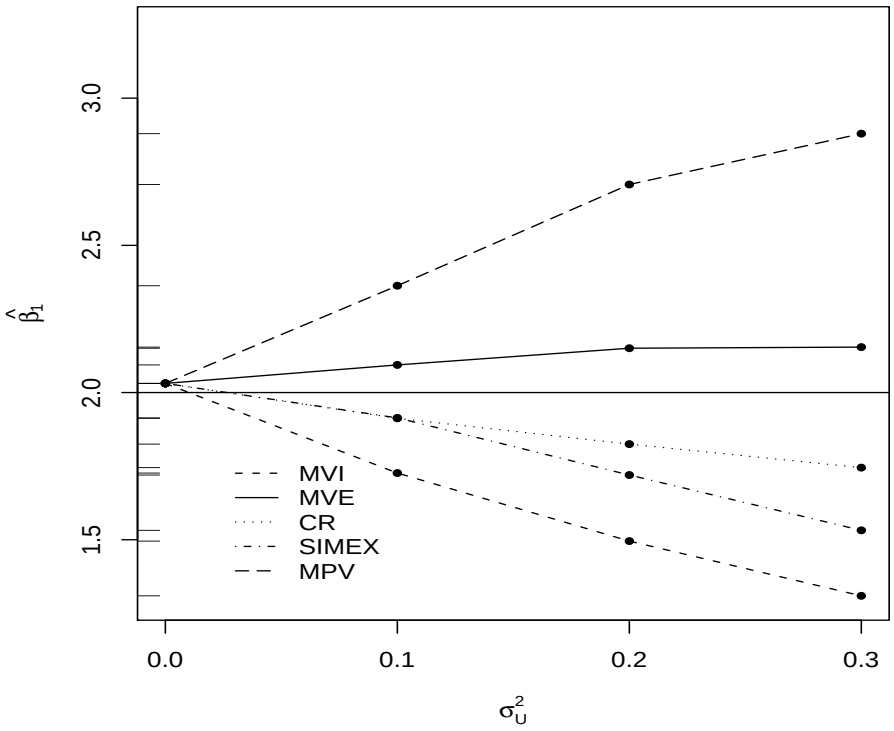

Figura 3.3: Média das estimativas de β_1 segundo diversos métodos, $n = 50$ (Cenário 1).

jeição, em todos os tamanhos de amostra, principalmente nas menores. As estatísticas RV, Wald e escore mostraram-se liberais nesta situação, com a maioria das taxas de rejeição acima do nível de significância (α), especialmente para $n = 10$ e $n = 20$. A maior diferença entre as taxas de rejeição e o nível de significância pode ser observada na estatística do escore, para $n = 10$ e $n = 20$. O tamanho da amostra tem forte influência nos resultados das estatísticas de teste, particularmente no teste do escore, aproximando as taxas de rejeição do valor nominal, à medida que aumentamos o tamanho da amostra. As estatísticas RV e Wald apresentaram os mesmos resultados, para $n = 100$, em todas as variâncias. Para os demais tamanhos de amostra também tiveram comportamento semelhante.

Apresentamos na Figura 3.10 gráficos de quantis para $n = 100$ e $\sigma_U^2 = 0,4$. Quanto mais os valores das estatísticas se aproximam dos quantis teóricos, mais próximos estarão os pares de pontos da reta identidade, representada pela linha tracejada. Cons-

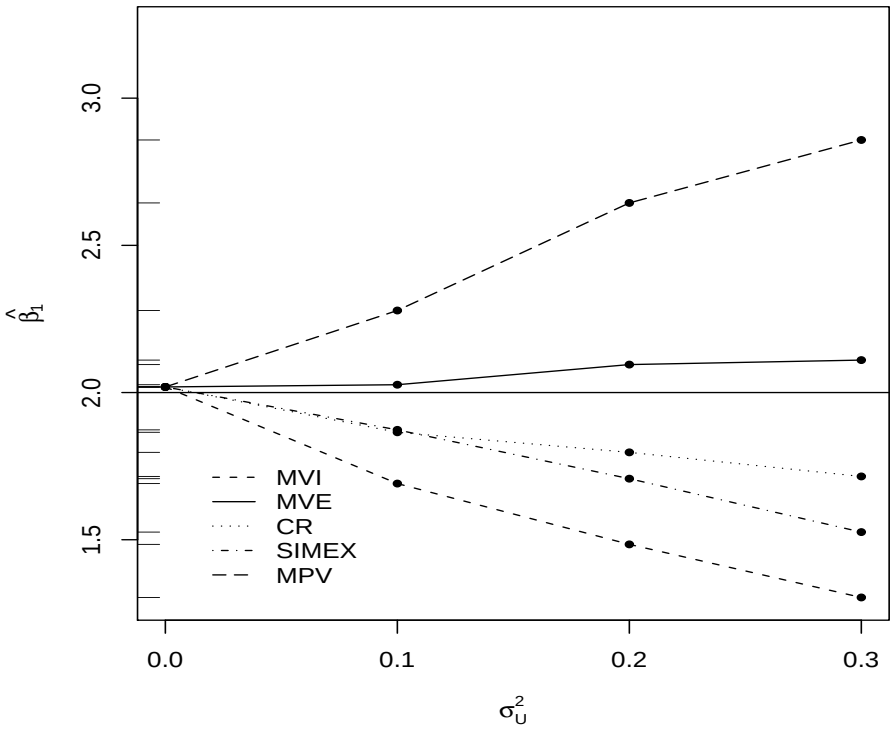

Figura 3.4: Média das estimativas de β_1 segundo diversos métodos, $n = 100$ (Cenário 1).

tatamos que, de um modo geral, os valores obtidos para as três estatísticas se aproximaram dos quantis da distribuição $\chi^2_{(1)}$, exceto para uma das amostras. A estatística de teste RV foi a que mais se aproximou dos quantis teóricos.

As simulações do poder empírico (%) dos testes (RV, Wald e escore), estão representadas nas Tabelas 3.7 e 3.8, para $H_0 : \beta_1 = 0,5$ e $H_0 : \beta_1 = 1,5$, respectivamente. Observamos na Tabela 3.7 que o poder aumenta conforme aumenta o tamanho da amostra, se aproximando de 100% para o maior tamanho da amostra. O teste do escore apresenta o maior poder para $n = 10$, 20 e 50 e para $n = 100$ todos os testes tiveram os mesmos resultados, exceto o do escore para variância 0,4. A variância do erro de medição tem influência sobre o poder do teste, fazendo com que esse valor diminua com o aumento do erro de medição, o que pode ser verificado para todos os métodos e tamanhos de amostra.

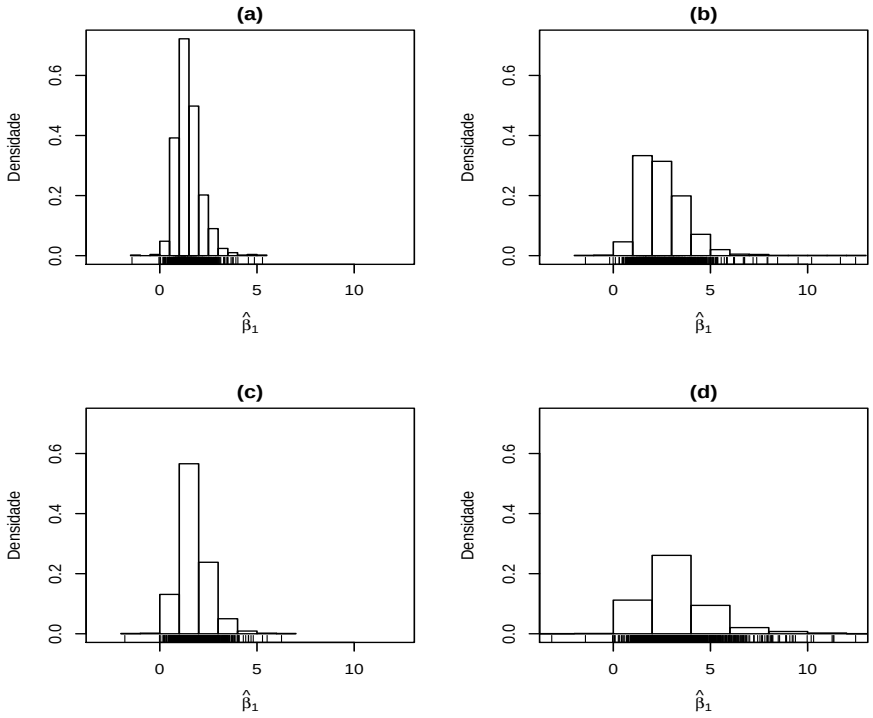

Figura 3.5: Cenário 1 - Histogramas das estimativas de β_1 para $n = 10$ e $\sigma_U^2 = 0,3$, utilizando os métodos de MVI (a), MVE (b), SIMEX (c) e MPV (d).

Na Tabela 3.8 além de observarmos a influência da variância do erro de medição e do tamanho da amostra, como na tabela anterior, vemos que a estatística de RV apresenta os maiores valores do poder, especialmente para $n = 10$ e $n = 20$. No entanto, os percentuais estão próximos em todos os testes e para todas as variâncias e tamanhos amostrais. Para $n = 50$ e $n = 100$ o poder no teste RV é de 100% para todas as variâncias. Podemos verificar que quando $\beta_1 = 1,5$ as taxas de rejeição, mesmo para $n = 20$, já estão próximas de 95%, o que para $\beta_1 = 0,5$ só acontece quando $n = 50$. De acordo com estes resultados e a análise anterior parece-nos que as estatísticas de teste mais conveniente, neste caso, são RV e Wald, que apresentaram resultados razoáveis mesmo para as amostras menores ao passo que a estatística do escore só apresentou bons resultados para as amostras 50 e 100.

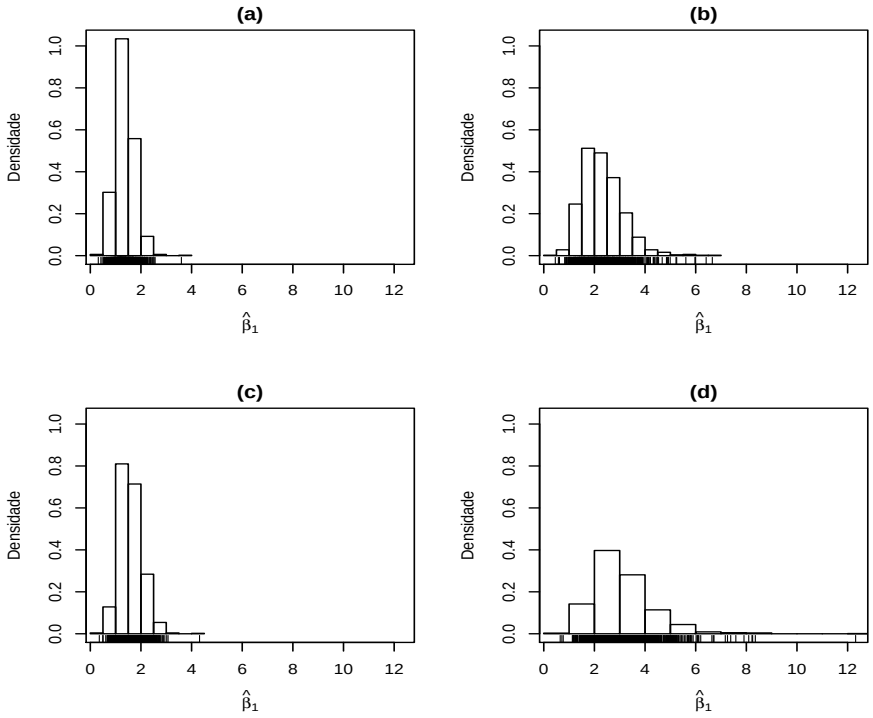

Figura 3.6: Cenário 1 - Histogramas das estimativas de β_1 para $n = 20$ e $\sigma_U^2 = 0,3$, utilizando os métodos de MVI (a), MVE (b), SIMEX (c) e MPV (d).

3.2 Cenário 2

Neste cenário realizamos estudo de simulação com os métodos CR e MVE, pois foram os que tiveram melhor performance no Cenário 1 e também para economizarmos tempo computacional, pois essas simulações foram bastente demoradas, sendo que algumas demoraram mais de 10 dias para serem concluídas. Estas simulações foram realizadas em um computador pessoal usando o sistema operacional Windows XP, processador Pentium 4 (Intel) de 3GHz e 256 MB de mémoria RAM. Neste estudo os valores verdadeiros dos parâmetros foram $\beta_0 = 1015,8$, $\beta_1 = -104,86$, $c = 5,4$ ($\phi = 0,16$), $\mu_X = 9,6$ e $\sigma_X^2 = 0,0015$. Os valores verdadeiros de β_0 e β_1 são as estimativas obtidas para o conjunto de dados da Tabela 4.1 quando ajustado o modelo binomial, o parâmetro c foi obtido conforme o Capítulo 2 (a partir do valor de ϕ),

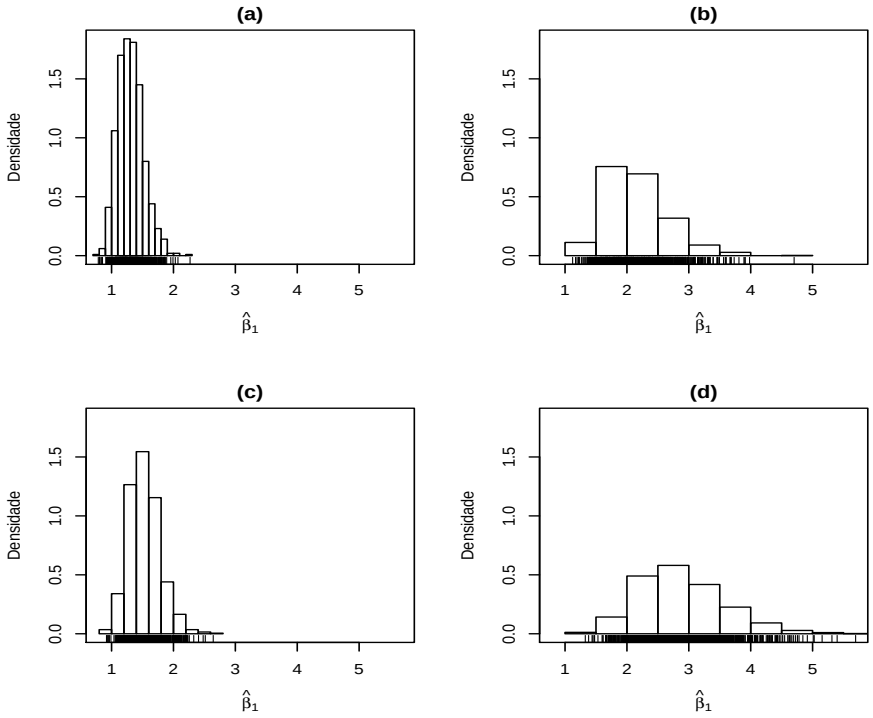

Figura 3.7: Cenário 1 - Histogramas das estimativas de β_1 para $n = 50$ e $\sigma_U^2 = 0,3$, utilizando os métodos de MVI (a), MVE (b), SIMEX (c) e MPV (d).

finalmente, μ_X e σ_X^2 representam a média e a variância amostral do conjunto de dados da aplicação da Tabela 4.1. O vetor de variâncias utilizado nestas simulações é $\sigma_U^2 = (0,0001; 0,0002; 0,0003; 0,0004)$.

As Figuras 3.11 - 3.14 representam os valores médios de $\widehat{\beta}_1$, para os métodos MVE e CR, para os diversos tamanhos de amostra e de variâncias, neste cenário. Podemos observar que ambos os métodos tiveram comportamento semelhante para $n = 10$ e $n = 20$, porém com estimativas bastante distantes do valor verdadeiro. À medida que aumentamos o tamanho da amostra os métodos se diferenciaram bastante quanto às estimativas apresentadas. No método MVE existe uma certa flutuação nas estimativas de β_1, para $n = 50$ e $n = 100$.

Constatamos que a variância do erro de medição (σ_U^2) tem maior influência no

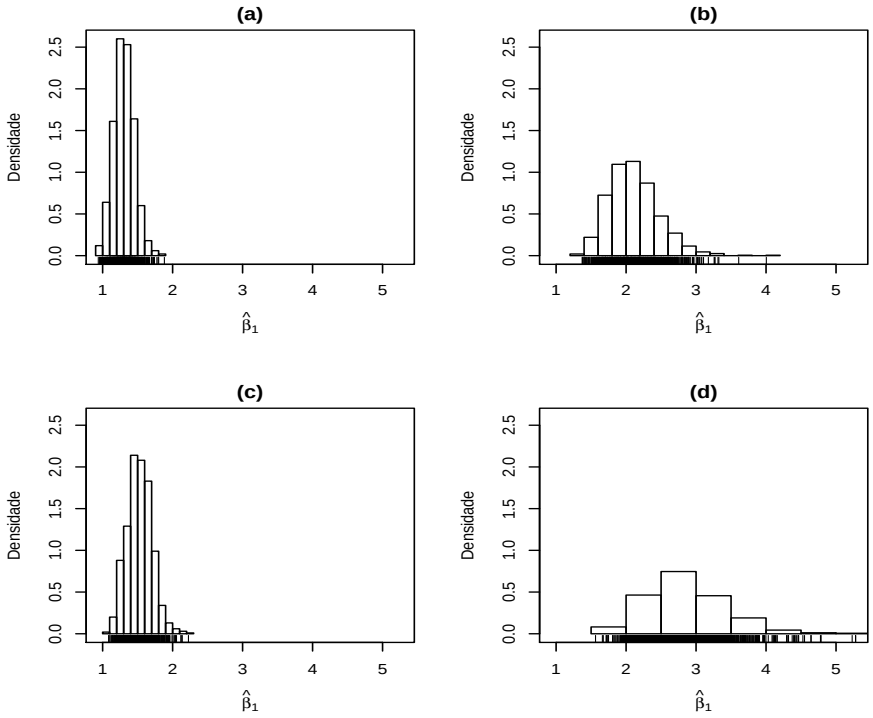

Figura 3.8: Cenário 1 - Histogramas das estimativas de β_1 para $n = 100$ e $\sigma_U^2 = 0,3$, utilizando os métodos de MVI (a), MVE (b), SIMEX (c) e MPV (d).

método CR do que no MVE, fazendo com que essas estimativas se afastem mais do valor verdadeiro, para todos os tamanhos de amostra.

Resultados obtidos como estimativas dos parâmetros, viés simulado e erro quadrático médio simulados (EQM), de β_0 e β_1, são apresentados nas Tabelas 3.9 e 3.10, onde constatamos que os menores vieses das estimativas dos parâmetros, em termos absolutos, foram obtidos no método MVE para $n = 50$, embora o EQM tenha sido alto, para $\widehat{\beta}_1$ e $\sigma_U^2 = 0,0002$, por exemplo, foi de 2972,40. Quando passamos de $n = 50$ para $n = 100$ houve um aumento nos vieses das estimativas.

As médias das estimativas que mais se aproximaram do valor verdadeiro foram $\widehat{\beta}_0 = 1047,00$ e $\widehat{\beta}_1 = -108,00$ com vieses 31,20 e -3,22, respectivamente. Os menores EQM resultaram do método CR para $n = 100$.

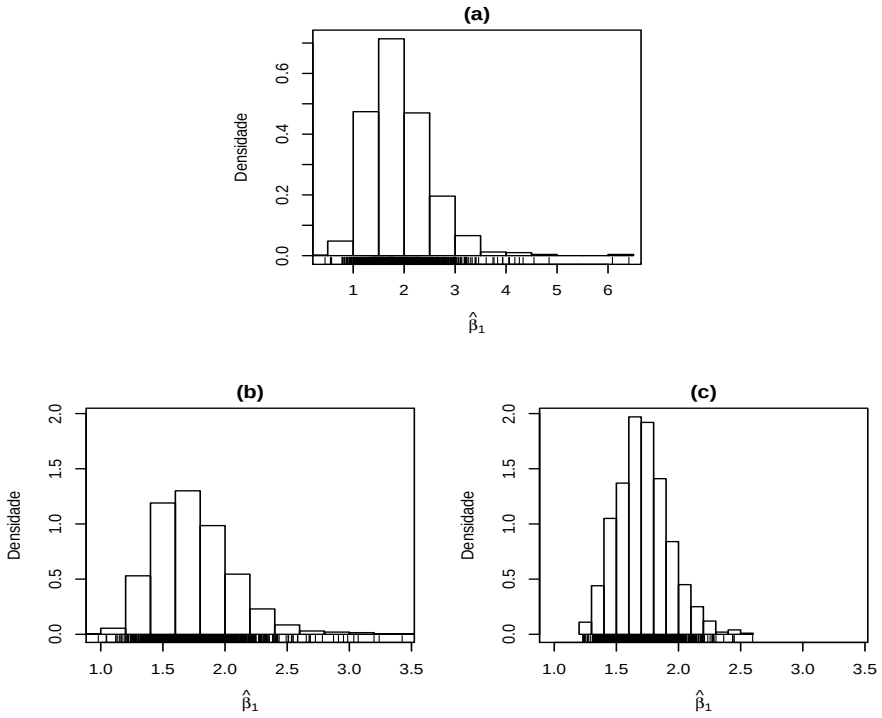

Figura 3.9: Cenário 1 - Histogramas das estimativas de β_1 para $n = 20$ (a), $n = 50$ (b) e $n = 100$, $\sigma_U^2 = 0,3$, para o método CR.

Neste cenário encontramos situações em que não é possível comparar o EQM dos estimadores pois o viés é alto.

Para esse cenário, o método de máxima verossimilhança apresentou estimativas mais próximas dos valores verdadeiros quando $n = 50$ e também menor viés, contudo neste caso, os EQM dos parâmetros ficaram acima dos valores calculados para $n = 20$ e $n = 100$. As variâncias do erro de medição, apesar de parecerem pequenas, influenciaram as estimativas fazendo que aumentasse o viés, para todos os tamanhos de amostra em ambos os métodos.

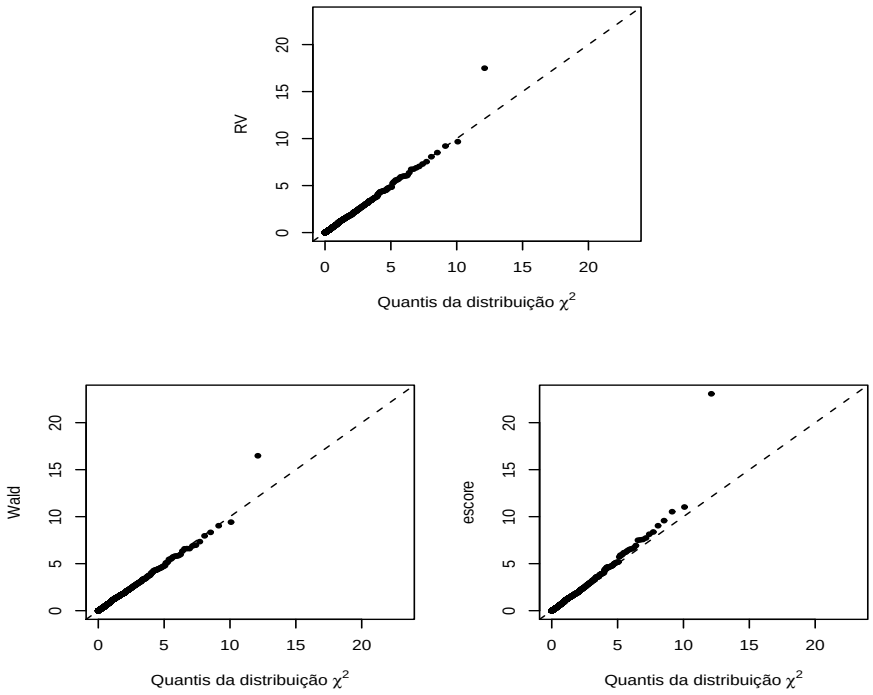

Figura 3.10: Cenário 1 - Gráficos de quantis para as estatísticas de teste (RV, Wald e escore) com $n = 100$ e $\sigma_U^2 = 0,4$.

Tabela 3.1: Cenário 1 - Média das estimativas dos parâmetros do modelo, viés simulado e erro quadrático médio simulado (EQM) utilizando os métodos de máxima verossimilhança ingênuo (MVI) e com erros (MVE), calibração da regressão (CR), SIMEX e máxima pseudo-verossimilhança (MPV), $n = 10$.

Estimadores	Variância (σ_U^2)	Estimativas		Viés	EQM	
		$\widehat{\beta}_0$	$\widehat{\beta}_1$	$\widehat{\beta}_1$	$\widehat{\beta}_0$	$\widehat{\beta}_1$
MVI	0	0,0130	2,2485	0,2485	0,1719	0,6041
	0,1	0,0157	1,8669	-0,1331	0,1958	0,4477
	0,2	0,0030	1,6333	-0,3667	0,2183	0,5571
	0,3	-0,0073	1,4704	-0,5296	0,2640	0,6883
MVE	0	0,0130	2,2485	0,2485	0,1719	0,6041
	0,1	0,0147	2,3845	0,3845	0,2772	1,0123
	0,2	0,0164	2,5110	0,5110	0,4115	1,8208
	0,3	-0,0122	2,5305	0,5305	0,5255	1,9278
CR	0	0,0130	2,2485	0,2485	0,1719	0,6041
	0,1	0,0047	2,4308	0,4308	0,2905	0,8163
	0,2	0,0091	2,2530	0,2530	0,5098	1,8501
SIMEX	0	0,0130	2,2485	0,2485	0,1719	0,6041
	0,1	0,0166	2,0740	0,0740	0,2330	0,5921
	0,2	0,0017	1,8788	-0,1212	0,2609	0,6231
	0,3	-0,0083	1,7246	-0,2754	0,3200	0,6718
MPV	0	0,0130	2,2485	0,2485	0,1719	0,6041
	0,1	0,0103	2,7570	0,7570	0,3929	1,9936
	0,2	0,0023	3,0422	1,0422	0,6452	3,0903
	0,3	-0,0013	3,2334	1,2334	1,0048	4,3748

Tabela 3.2: Cenário 1 - Média das estimativas dos parâmetros do modelo, viés simulado e erro quadrático médio simulado (EQM) utilizando os métodos de máxima verossimilhança ingênuo (MVI) e com erros (MVE), calibração da regressão (CR), SIMEX e máxima pseudo-verossimilhança (MPV), $n = 20$.

Estimadores	Variância (σ_U^2)	Estimativas		Viés	EQM	
		$\widehat{\beta}_0$	$\widehat{\beta}_1$	$\widehat{\beta}_1$	$\widehat{\beta}_0$	$\widehat{\beta}_1$
MVI	0	0,0018	2,0934	0,0934	0,0693	0,2069
	0,1	0,0025	1,7818	-0,2182	0,0858	0,2019
	0,2	0,0023	1,5212	-0,4788	0,0901	0,3673
	0,3	0,0017	1,3608	-0,6392	0,0945	0,5429
MVE	0	0,0018	2,0934	0,0934	0,0693	0,2069
	0,1	0,0037	2,2191	0,2191	0,1101	0,4164
	0,2	-0,0040	2,2553	0,2553	0,1354	0,6080
	0,3	0,0031	2,3092	0,3092	0,1651	0,7498
CR	0	0,0018	2,0934	0,0934	0,0693	0,2069
	0,1	0,0037	1,9965	-0,0035	0,0896	0,2080
	0,2	-0,0163	1,9657	-0,0343	0,2740	3,7756
	0,3	0,0024	1,9137	-0,0863	0,1223	0,3923
SIMEX	0	0,0018	2,0934	0,0934	0,0693	0,2069
	0,1	0,0036	1,9742	-0,0258	0,0965	0,2145
	0,2	0,0003	1,7491	-0,2509	0,1030	0,2637
	0,3	0,0023	1,5929	-0,4071	0,1082	0,3614
MPV	0	0,0018	2,0934	0,0934	0,06928	0,20693
	0,1	0,0083	2,5188	0,5188	0,13805	0,84331
	0,2	-0,0118	2,8306	0,8306	0,22293	1,6718
	0,3	0,0045	3,0728	1,0728	0,33115	2,5073

Tabela 3.3: Cenário 1 - Média das estimativas dos parâmetros do modelo, viés simulado e erro quadrático médio simulado (EQM) utilizando os métodos de máxima verossimilhança ingênuo (MVI) e com erros (MVE), calibração da regressão (CR), SIMEX e máxima pseudo-verossimilhança (MPV), $n = 50$.

Estimadores	Variância (σ_U^2)	Estimativas		Viés	EQM	
		$\widehat{\beta}_0$	$\widehat{\beta}_1$	$\widehat{\beta}_1$	$\widehat{\beta}_0$	$\widehat{\beta}_1$
MVI	0	0,0076	2,0310	0,0310	0,0256	0,0573
	0,1	-0,0041	1,7265	-0,2735	0,0310	0,1333
	0,2	0,0126	1,4954	-0,5046	0,0309	0,3035
	0,3	0,0017	1,3094	-0,6906	0,0328	0,5196
MVE	0	0,0076	2,0310	0,0310	0,0256	0,0573
	0,1	-0,0021	2,0938	0,0938	0,0377	0,1326
	0,2	0,0127	2,1506	0,1506	0,0502	0,2131
	0,3	0,0005	2,1543	0,1543	0,0592	0,2732
CR	0	0,0076	2,0310	0,0310	0,0256	0,0573
	0,1	-0,0023	1,9125	-0,0875	0,0318	0,0819
	0,2	0,0105	1,8248	-0,1752	0,0355	0,1107
	0,3	-0,0013	1,7449	-0,2551	0,0395	0,1664
SIMEX	0	0,0076	2,0310	0,0310	0,0256	0,0573
	0,1	-0,0038	1,9138	-0,0862	0,0344	0,0875
	0,2	0,0129	1,7198	-0,2802	0,0358	0,1495
	0,3	0,0009	1,5319	-0,4681	0,0375	0,2821
MPV	0	0,0076	2,0310	0,0310	0,0256	0,0573
	0,1	-0,0017	2,3628	0,3628	0,0449	0,3267
	0,2	0,0075	2,7067	0,7067	0,0875	0,8631
	0,3	-0,0006	2,8794	0,8794	0,1253	1,2432

Tabela 3.4: Cenário 1 - Média das estimativas dos parâmetros do modelo, viés simulado e erro quadrático médio simulado (EQM) utilizando os métodos de máxima verossimilhança ingênuo (MVI) e com erros (MVE), calibração da regressão (CR), SIMEX e máxima pseudo-verossimilhança (MPV), $n = 100$.

Estimadores	Variância (σ_U^2)	Estimativas		Viés	EQM	
		$\widehat{\beta}_0$	$\widehat{\beta}_1$	$\widehat{\beta}_1$	$\widehat{\beta}_0$	$\widehat{\beta}_1$
MVI	0	0,0027	2,0190	0,0190	0,0127	0,0298
	0,1	0,0059	1,6908	-0,3092	0,0147	0,1216
	0,2	0,0006	1,4842	-0,5158	0,0151	0,2911
	0,3	0,0052	1,3039	-0,6961	0,0157	0,5056
MVE	0	0,0027	2,0190	0,0190	0,0127	0,0298
	0,1	0,0069	2,0266	0,0266	0,0179	0,0522
	0,2	-0,0003	2,0950	0,0950	0,0217	0,0999
	0,3	0,0049	2,1102	0,1102	0,0262	0,1371
CR	0	0,0027	2,0190	0,0190	0,0127	0,0298
	0,1	0,0066	1,8650	-0,1350	0,0153	0,0507
	0,2	-0,0001	1,7968	-0,2032	0,0164	0,0806
	0,3	0,0053	1,7150	-0,2850	0,0177	0,1251
SIMEX	0	0,0027	2,0190	0,0190	0,0127	0,0298
	0,1	0,0069	1,8732	-0,1268	0,0165	0,0517
	0,2	0,0005	1,7074	-0,2926	0,0172	0,1218
	0,3	0,0057	1,5260	-0,4740	0,0177	0,2559
MPV	0	0,0027	2,0190	0,0190	0,0127	0,0298
	0,1	0,0108	2,2787	0,2787	0,0223	0,1605
	0,2	-0,0027	2,6441	0,6441	0,0382	0,6266
	0,3	0,0068	2,8580	0,8580	0,0655	1,0543

Tabela 3.5: Cenário 1 - Taxas de rejeição (%) de $H_0 : \phi = 0$ de acordo com o teste da razão de verossimilhanças para um nível de significância $\alpha = 0,05$.

Variância (σ_U^2)	Tamanho amostral (n)		
	20	50	100
0,1	2,4	2,8	2,9
0,2	2,0	1,8	3,9
0,3	1,4	3,3	3,1
0,4	2,4	2,1	3,0

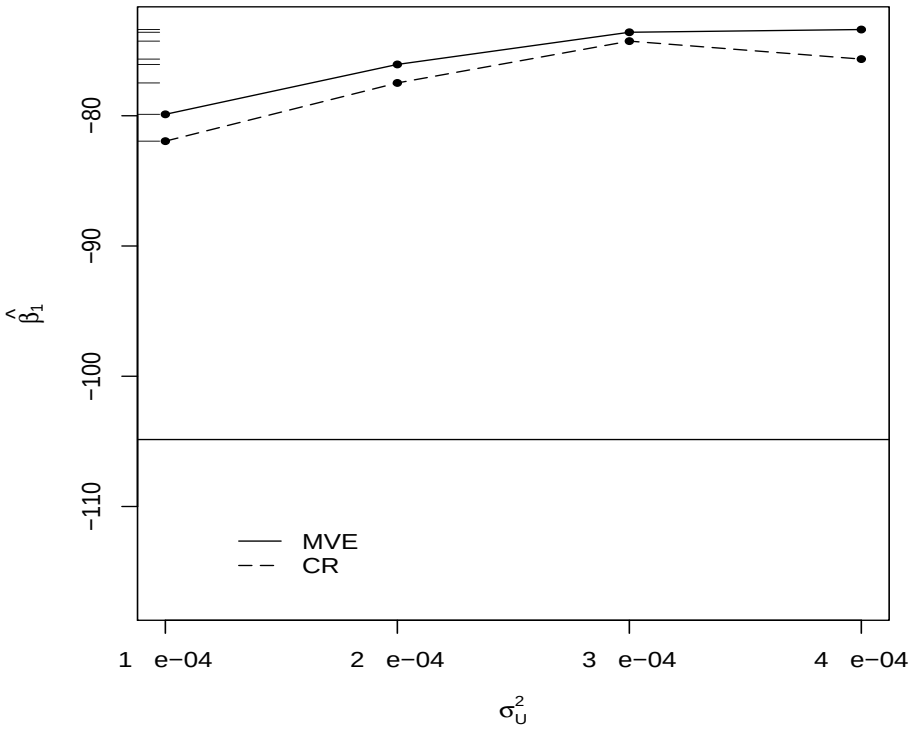

Figura 3.11: Média das estimativas de β_1 para os métodos MVE e CR, $n = 10$ (Cenário 2).

Tabela 3.6: Cenário 1 - Taxas de rejeição (%) da hipótese $H_0 : \beta_1 = 0$ de acordo com os testes da RV, Wald e escore para um nível de significância $\alpha = 0,05$.

Tamanho amostral (n)	Variância (σ_U^2)	Estatística de teste		
		RV	Wald	escore
10	0,1	7,5	8,0	10,4
	0,2	6,2	5,8	9,0
	0,3	7,4	7,0	11,0
	0,4	5,1	3,9	11,2
20	0,1	6,1	6,4	8,9
	0,2	5,8	5,6	8,6
	0,3	6,7	5,5	10,4
	0,4	6,1	4,4	10,0
50	0,1	4,7	5,0	6,5
	0,2	5,0	4,9	7,0
	0,3	4,1	3,7	4,8
	0,4	5,6	5,1	7,2
100	0,1	4,5	4,5	5,2
	0,2	4,1	4,1	4,4
	0,3	4,5	4,5	5,4
	0,4	4,6	4,6	5,5

Tabela 3.7: Cenário 1 - Poder empírico (%) dos testes RV, Wald e escore no teste de $H_0 : \beta_1 = 0,5$ com nível de significância $\alpha = 0,05$.

Tamanho amostral (n)	Variância (σ_U^2)	Estatística de teste		
		RV	Wald	escore
10	0,1	34,6	35,4	44,6
	0,2	30,0	28,4	37,9
	0,3	27,6	23,0	34,7
	0,4	22,4	16,3	28,2
20	0,1	57,5	60,0	63,9
	0,2	54,6	55,2	61,5
	0,3	52,0	50,0	56,9
	0,4	48,1	42,3	53,2
50	0,1	93,6	94,5	95,0
	0,2	88,5	89,3	90,1
	0,3	86,4	87,2	87,5
	0,4	84,1	83,9	86,6
100	0,1	99,9	99,9	99,9
	0,2	99,9	99,9	99,9
	0,3	99,7	99,7	99,7
	0,4	98,7	98,7	98,9

Tabela 3.8: Cenário 1 - Poder empírico (%) dos testes RV, Wald e escore no teste de $H_0 : \beta_1 = 1,5$ com nível de significância $\alpha = 0,05$.

Tamanho amostral (n)	Variância (σ_U^2)	Estatística de teste		
		RV	Wald	escore
10	0,1	90,2	85,6	85,3
	0,2	84,6	72,7	75,3
	0,3	79,5	59,4	64,8
	0,4	74,2	50,6	52,5
20	0,1	99,6	99,6	96,0
	0,2	98,9	97,9	95,0
	0,3	98,1	92,2	86,8
	0,4	96,1	85,4	78,4
50	0,1	100,0	100,0	99,9
	0,2	100,0	100,0	99,6
	0,3	100,0	99,9	98,4
	0,4	100,0	99,7	91,9
100	0,1	100,0	100,0	100,0
	0,2	100,0	100,0	100,0
	0,3	100,0	100,0	100,0
	0,4	100,0	99,9	98,2

Tabela 3.9: Cenário 2 - Média das estimativas dos parâmetros do modelo, viés simulado e erro quadrático médio simulado (EQM) utilizando os métodos de máxima verossimilhança com erros (MVE) e calibração da regressão (CR), $n = 10$ e 20.

Estimadores	Variância (σ_U^2)	Estimativas		Viés		EQM	
		$\widehat{\beta}_0$	$\widehat{\beta}_1$	$\widehat{\beta}_0$	$\widehat{\beta}_1$	$\widehat{\beta}_0(\times 10^5)$	$\widehat{\beta}_1$
$n = 10$							
MVE	0,0001	774,13	-79,89	-241,67	24,97	2,33	2496,4
	0,0002	737,25	-76,06	-278,55	28,80	2,52	2702,5
	0,0003	713,40	-73,59	-302,40	31,27	2,82	3025,9
	0,0004	711,29	-73,38	-304,51	31,48	2,97	3187,1
CR	0,0001	793,90	-81,95	-221,90	22,91	3,32	3555,7
	0,0002	750,70	-77,48	-265,10	27,38	3,63	3892,6
	0,0003	719,55	-74,27	-296,25	30,59	4,53	4864,4
	0,0004	732,70	-75,65	-283,10	29,21	4,73	5079,1
$n = 20$							
MVE	0,0001	953,05	-98,38	-62,75	6,48	1,94	2080,8
	0,0002	908,90	-93,81	-106,90	11,05	2,18	2339,7
	0,0003	908,10	-93,72	-107,70	11,14	2,57	2752,6
	0,0004	896,60	-92,54	-119,20	12,32	2,53	2706,9
CR	0,0001	906,48	-93,58	-109,32	11,28	2,52	2700,5
	0,0002	829,15	-85,59	-186,65	19,27	2,44	2615,9
	0,0003	836,07	-86,32	-179,73	18,54	3,36	3607,8
	0,0004	816,07	-84,28	-199,73	20,58	3,29	3532,6

Tabela 3.10: Cenário 2 - Média das estimativas dos parâmetros do modelo, viés simulado e erro quadrático médio simulado (EQM) utilizando os métodos de máxima verossimilhança com erros (MVE) e calibração da regressão (CR), $n = 50$ e 100.

Estimadores	Variância (σ_U^2)	Estimativas		Viés		EQM	
		$\widehat{\beta}_0$	$\widehat{\beta}_1$	$\widehat{\beta}_0$	$\widehat{\beta}_1$	$\widehat{\beta}_0(\times 10^5)$	$\widehat{\beta}_1$
$n = 50$							
MVE	0,0001	1073,00	-110,77	57,20	-5,91	1,97	2108,9
	0,0002	1063,10	-109,74	47,30	-4,88	2,77	2972,4
	0,0003	1094,50	-113,00	78,70	-8,14	2,60	2781,6
	0,0004	1047,00	-108,08	31,20	-3,22	2,58	2763,0
CR	0,0001	934,24	-96,43	-81,56	8,43	1,36	1462,5
	0,0002	873,87	-90,20	-141,93	14,66	1,65	1770,2
	0,0003	853,33	-88,11	-162,47	16,76	1,80	1926,1
	0,0004	814,61	-84,10	-201,19	20,76	2,02	2161,7
$n = 100$							
MVE	0,0001	1120,37	-115,67	104,57	-10,81	1,93	2062,3
	0,0002	1133,10	-116,99	117,30	-12,13	1,97	2102,5
	0,0003	1099,20	-113,48	83,40	-8,62	1,84	1964,4
	0,0004	1108,50	-114,45	92,70	-9,59	2,11	2259,6
CR	0,0001	921,18	-95,09	-94,62	9,77	0,83	893,8
	0,0002	873,42	-90,16	-142,38	14,70	0,98	1057,5
	0,0003	809,68	-83,58	-206,12	21,28	1,14	1225,5
	0,0004	763,61	-78,83	-252,19	26,03	1,28	1365,4

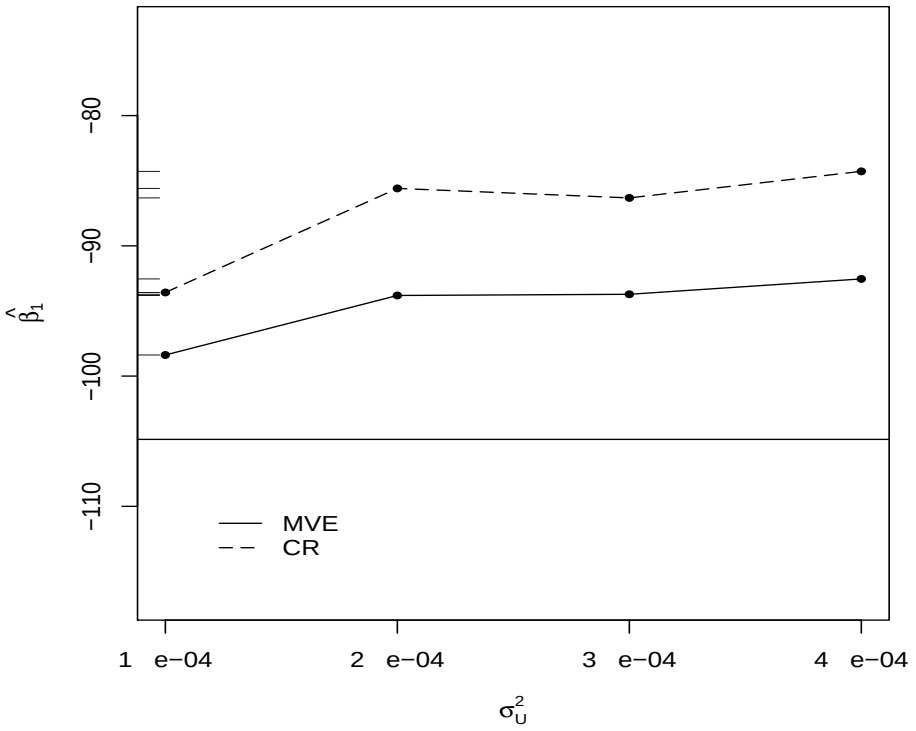

Figura 3.12: Média das estimativas de β_1 para os métodos MVE e CR, $n = 20$ (Cenário 2).

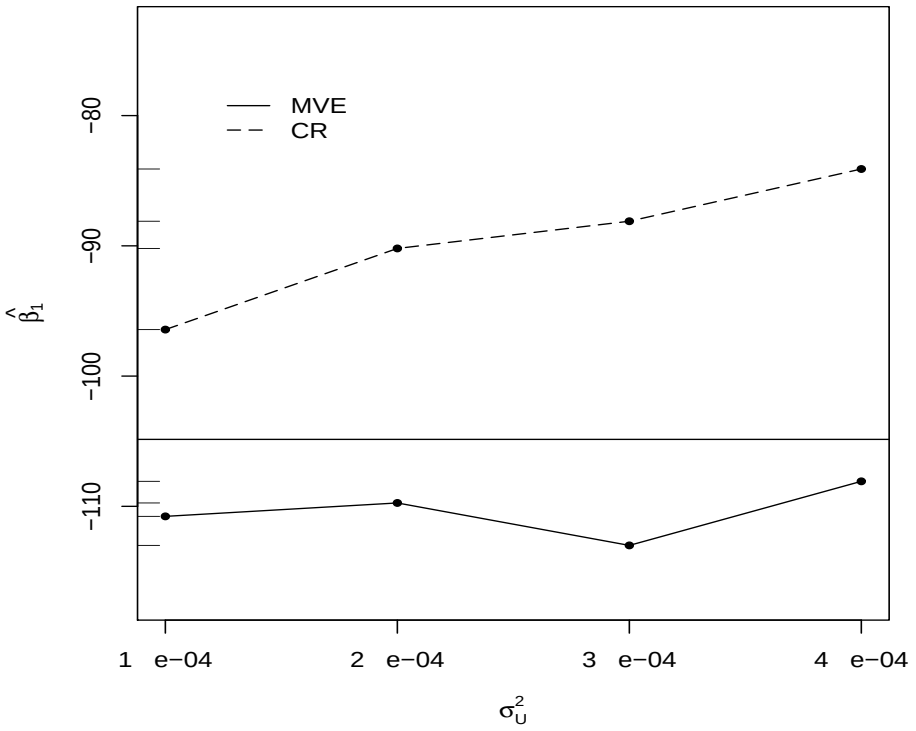

Figura 3.13: Média das estimativas de β_1 para os métodos MVE e CR, $n = 50$ (Cenário 2).

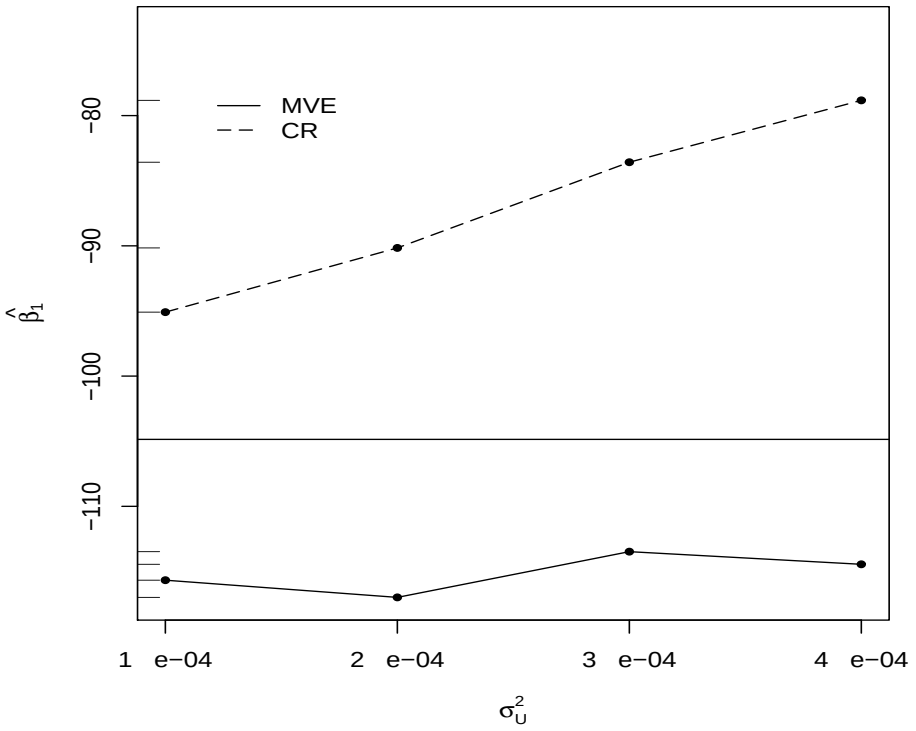

Figura 3.14: Média das estimativas de β_1 para os métodos MVE e CR, $n = 100$ (Cenário 2).

Capítulo 4

Aplicação

Neste capítulo desenvolvemos uma aplicação das técnicas apresentadas no Capítulo 2 e da experiência obtida no estudo de simulação (Capítulo 3). Peças em uma empresa são manufaturadas com base em especificações técnicas para diferentes características da qualidade, como diâmetro, composição química e massa. As especificações técnicas determinam intervalos para cada característica da qualidade. Por exemplo, o diâmetro de um furo em uma peça deve estar entre o LIE e o LSE, que são os limites inferior e superior de engenharia, respectivamente. Para avaliar a capacidade em atender a estas especificações é preciso medir (estimar) o valor destas características da qualidade para as peças produzidas, o que é feito com base em sistemas de medição apropriados. Em geral, classificamos os sistemas de medição em dois grupos: sistema de medição por variável e por atributo. Os sistemas de medição por variável são aqueles que determinam um valor numérico para a característica da qualidade da peça, enquanto os sistemas de medição por atributo classificam as peças em aceitas ou não, conforme suas especificações. Devido ao baixo custo de fabricação e manutenção e a facilidade de manuseio, os sistemas de medição por atributo são um dos sistemas de medição mais utilizados na indústria automobilística mundial.

4.1 Sistema de medição por atributo

Uma forma barata e rápida de as empresas controlarem seus processos e produtos consiste na utilização de sistemas de medição do tipo "passa" "não passa". Estes sistemas classificam as peças em defeituosas ou não, isto é, determinam se a característica da peça pertence ao intervalo (LIE, LSE). Apesar de baratos e ágeis, estes sistemas de medição são passíveis de falhas.

Para classificar a peça esse sistema de medição utiliza um dispositivo composto de duas faces, sendo que a face "passa" é confeccionada com algumas frações de unidade acima do LIE definido para a peça analisada, enquanto que a face "não passa" é confeccionada com algumas frações abaixo do LSE, conforme Figura 4.1. A peça será considerada defeituosa caso o diâmetro do furo seja menor que a face passa ou maior que a face não passa. Mais detalhes sobre o sistema de medição por atributo podem ser obtidos em Favari (2006).

4.2 Descrição do experimento

Os dados deste exemplo foram coletados de um experimento realizado em uma fábrica de peças automotivas onde foi avaliado se cada uma das oito peças inspecionadas atendiam a uma especificação de diâmetro do furo de uma peça (valor de referência, em mm). O método de inspeção é do tipo "passa" "não passa". As peças que satisfizessem essa condição seriam aceitas e as demais rejeitadas (defeituosas). Foram realizadas 20 inspeções em cada peça, utilizando o mesmo equipamento, por um mesmo técnico. Os resultados são apresentados na Tabela 4.1 e na Figura 4.3.

Um dos objetivos do experimento consiste em estudar a associação entre o diâmetro do furo de uma peça (chamada de valor de referência, em mm), medido em um equi-

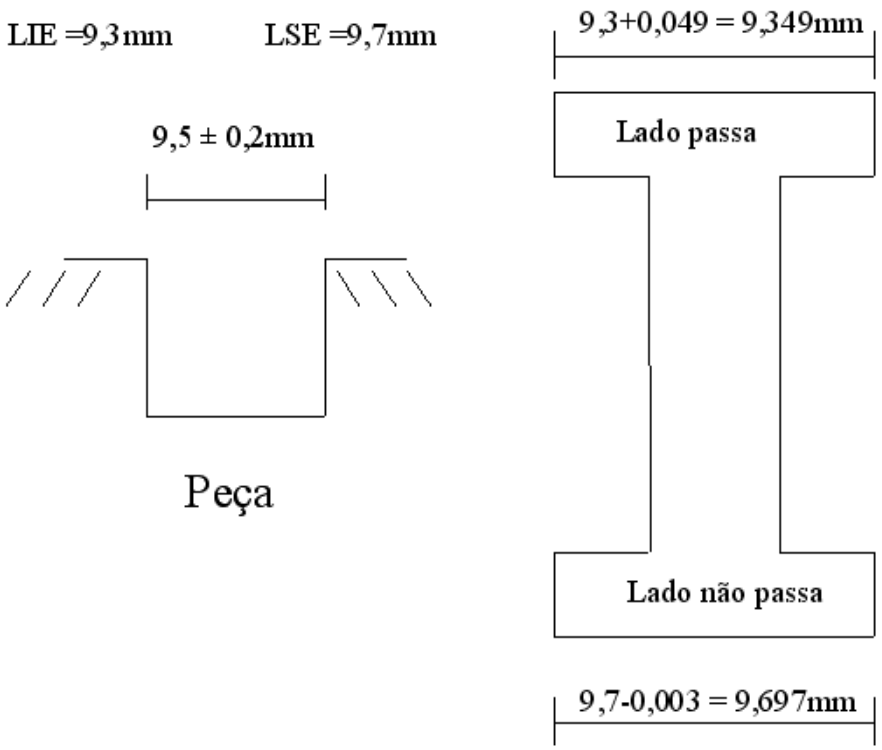

Figura 4.1: Sistema de medição "passa" "não passa".

pamento de precisão, e a proporção de aceitação. O erro de medição no valor de referência tem desvio padrão igual a $0,006$mm. Embora aparentemente pequeno, este desvio padrão traz consequências para a análise dos dados fazendo com que que diminua sensivelmente a proporção de intervalos que contêm o verdadeiro valor do parâmetro $\widehat{\beta}_1$, quando desconsiderado os erros de medição (Favari, 2006).

Para os dados do exemplo obtivemos as estimativas dos parâmetros do modelo pelos métodos de máxima verossimilhança (binomial, beta-binomial ingênuo, beta-binomial com erros de medição), SIMEX e calibração da regressão, que estão relacionadas na Tabela 4.2 juntamente com seus respectivos erros padrão. Nesta tabela MV representa o método de máxima verossimilhança no modelo binomial sem erros de medição. A maximização das funções foi obtida pelo método BFGS, com os valores iniciais $c = 2$, $\beta_0 = 2$ e $\beta_1 = -1$, que foram escolhidos sem nenhuma referência, embora tenhamos efetuado vários testes, com diferentes valores iniciais e sempre obtendo as mesmas

Tabela 4.1: Valor de referência e número de aceitações em 20 inspeções.

Item	Valor de referência (mm)	Número de aceitações
1	9,64	20
2	9,65	20
3	9,67	20
4	9,68	8
5	9,70	5
6	9,71	3
7	9,73	0
8	9,75	0

estimativas, ou seja, neste caso os valores iniciais não influenciaram nas estimativas. O desvio do modelo sem erros de medição usando a distribuição binomial foi de $15,44$ para 6 graus de liberdade, indicando sobredispersão, embora na Tabela 4.2 o erro padrão de $\widehat{\phi}$ seja próximo a $\widehat{\phi}$. Os erros padrão de $\widehat{\phi}$, para os métodos MVI e MVE, foram obtidos pelo método delta e para os demais métodos utilizamos o descrito no Capítulo 2. De acordo com a estimativa de máxima verossimilhança, a variância da distribuição binomial é multiplicada por $1 + \widehat{\phi}(20 - 1) = 4,04$.

No cálculo das estimativas SIMEX, a partir dos dados do exemplo obtivemos 1000 réplicas SIMEX, com $\lambda_L = 2$ e $L = 16$, usamos a mediana como medida resumo e ajuste linear, dadas as justificativas anteriores (Capítulo 3). Inclusive, apresentamos na Figura 4.2 um comparativo entre os ajustes linear e quadrático para a extrapolação, onde podemos observar que o ajuste linear não perde em qualidade para o ajuste quadrático, para esse conjunto de dados e ainda é mais simples. Constata-se que os menores erros padrão para $\widehat{\boldsymbol{\theta}}$ foram obtidos pelo método de CR e que as estimativas que mais diferem das demais são as referentes ao método MVE, para todos os parâmetros. O

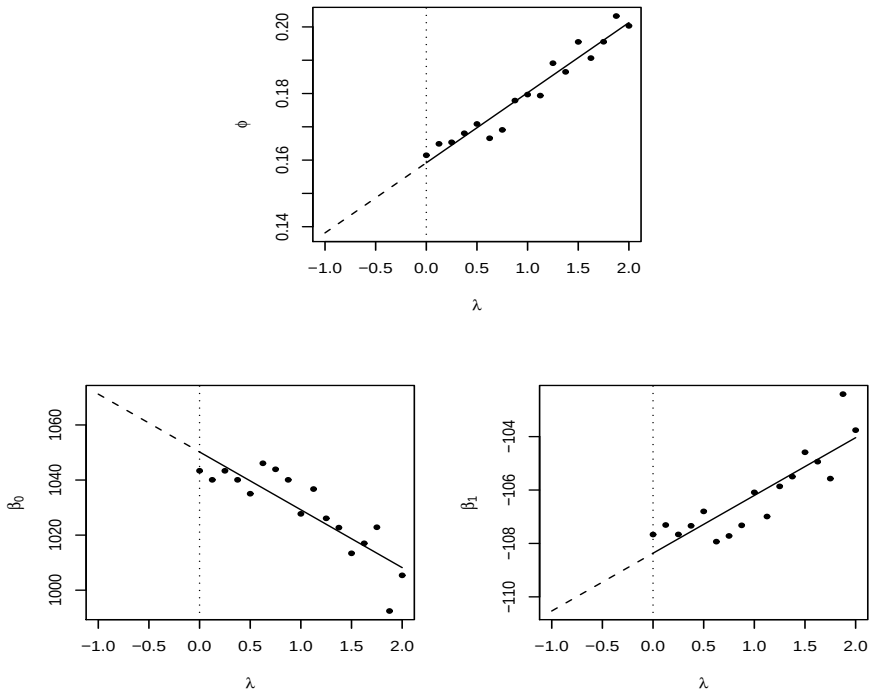

Figura 4.2: Resultados da etapa de simulação, ajuste do modelo linear e extrapolação (método SIMEX).

Tabela 4.2: Estimativas dos parâmetros e erros padrão.

Métodos	Parâmetros	Estimativas	Erros padrão
MV	β_0	1015,80	160,68
	β_1	-104,86	16,59
MVI	β_0	1043,40	290,54
	β_1	-107,67	29,98
	ϕ	0,16	0,14
MVE	β_0	798,35	213,73
	β_1	-82,33	22,05
	ϕ	0,088	0,099
	μ_X	9,70	0,015
	σ_X^2	0,0018	0,0009
CR	β_0	1073,40	66,80
	β_1	-110,77	6,87
	ϕ	0,16	0,091
SIMEX	β_0	1067,42	406,89
	β_1	-110,16	26,63
	ϕ	0,14	0,50

maior erro padrão de $\widehat{\beta}_1$ foi obtido com o método SIMEX. Neste exemplo, comparando MVE e MVI ocorre atenuação, ao contrário dos demais métodos que procuram corrigir os efeitos do erro de medição. O gráfico de dispersão entre as proporções observadas e os valores de referência, bem como os modelos ajustados são apresentados na Figura 4.3. Observamos que os pontos abaixo de $\widehat{\pi} = 0,3$ parecem melhor ajustados em todos os modelos, exceto o beta-binomial com erros de medição. Destacamos que existe uma diferença mais acentuada entre os modelos beta-binomial com erros de medição e os demais na parte central do gráfico. As estimativas obtidas com o método MVE são

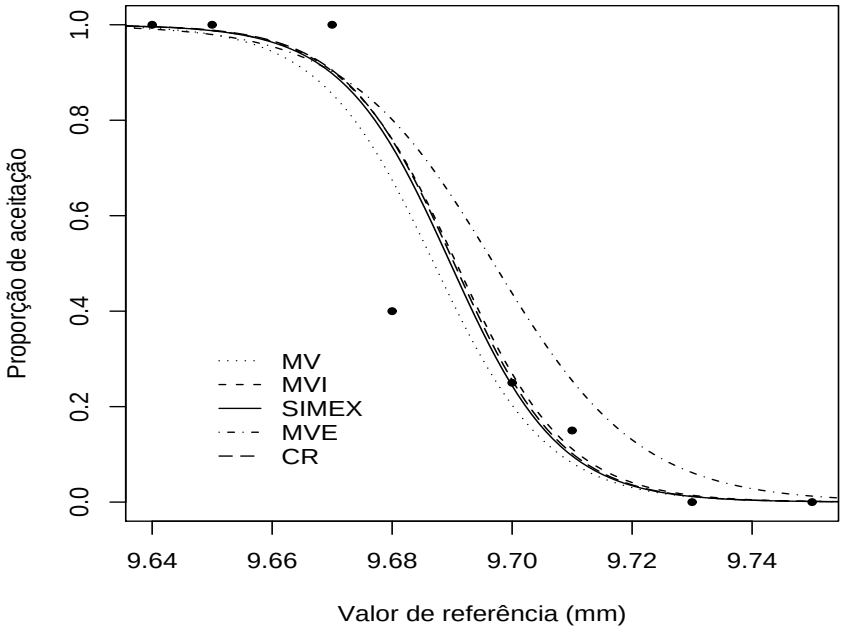

Figura 4.3: Proporção de aceitação *versus* valor de referência e modelos ajustados.

sempre menores (em valor absoluto) que as demais, o que justifica o afastamento dessa curva das outras, na Figura 4.3. Levando em conta o tamanho da amostra, a Figura 4.3, e considerando os resultados obtidos nas simulações, CR é o método que mais parece se adequar.

Capítulo 5

Conclusão

Neste trabalho estudamos uma extensão ao modelo de regressão para resposta binária usual. Foram tratados o problema da sobredispersão na variável resposta e a presença de erros de medição na variável explicativa. Em seguida destacamos alguns pontos.

No Cenário 1 das simulações, onde comparamos diversos métodos de estimação dos parâmetros, podemos verificar o melhor desempenho dos métodos MVE e CR, com estimativas menos viesadas, à medida em que aumentamos o tamanho da amostra. Podemos também analisar a influência dos erros de medição que provocaram um aumento do viés das estimativas dos parâmetros, com o aumento do erro de medição.

No Cenário 2, talvez a maior dificuldade tenha sido o tempo computacional, que conforme relatamos no Capítulo 3 em algumas situações as simulações foram muito demoradas. Observamos que os métodos (MVE e CR) diferiram bastante, quanto às estimativas, quando aumentamos o tamanho da amostra; também houve uma flutuação dos resultados simulados.

Tivemos dificuldade em encontrar na literatura testes de hipóteses para a sobredispersão quando a covariável contém erros de medição. Por isso o teste realizado foi uma

adaptação de um teste da RV de Paul *et al.* (1989). Este teste mostrou-se conservador para hipótese $H_0 : \phi = 0$.

As estatísticas de teste para $H_0 : \beta_1 = 0$ contra $H_1 : \beta_1 \neq 0$ que tiveram melhor desempenho foram as estatísticas RV e Wald, com taxa de rejeição próximas ao valor nominal, mesmo para amostras pequenas e para todas as variâncias, no entanto os melhores resultados foram obtidos para $n = 50$ e $n = 100$. O teste do escore não apresentou bons resultados para amostras pequenas, apresentando taxas de rejeição sob H_0 superiores ao valor nominal. Conforme esperado, na simulações do poder empírico das estatísticas de teste os resultados foram influenciados pelo tamanho da amostra e pelo valor de β_1. O aumento da variância do erro de medição contribuiu para diminuir o poder dos testes.

Na aplicação vimos que um desvio padrão igual a $0,006$mm, embora aparentemente pequeno, faz com que diminua sensivelmente a proporção de intervalos que contêm o verdadeiro valor do parâmetro $\widehat{\beta}_1$, quando desconsiderado os erros de medição (Favari, 2006). Na aplicação o método CR apresentou resultados com menores erros padrão.

Como proposta de trabalhos futuros listamos os seguintes itens:

1. Estudo de funções de ligação assimétricas, por exemplo, ligação extremito (complemento log-log) e também função de ligação com um parâmetro (a estimar) (Bazán *et al.*, 2006);

2. Estudo das propriedades de tendência e repetitividade do sistema de medição "passa" "não passa". A tendência representa as diferenças entre as medições deste sistema e do sistema medição de referência e a repetitividade representa a variabilidade associada ao sistema de medição "passa" "não passa";

3. Utilizar modelos lineares generalizados mistos com erros de medição - GLMMeMs, propostos por Wang *et al.* (1998), para análise do viés e inferência funcional

usando o método SIMEX;

4. Adaptação de outras estatísticas de teste da hipótese $H_0 : \phi = 0$ para modelos onde a variável explicativa é medida com erros; e

5. Com o objetivo de analisar a influência da sobredispersão sobre os parâmetros do modelo realizar estudos de simulação de estatísticas de teste para diferentes valores do parâmetro de sobredispersão (ϕ).

Apêndice A

Função escore e matriz de informação de Fisher do modelo beta-binomial

No Capítulo 2 apresentamos a função log-verossimilhança para o modelo beta-binomial, a função escore e também a matriz de informação de Fisher.

Lembrando que $\frac{\partial \pi_i(x)}{\partial \beta_0} = \pi_i(1-\pi_i)$ e $\frac{\partial \pi_i(x)}{\partial \beta_1} = x_i\pi_i(1-\pi_i)$, apresentamos, inicialmente, os componentes da função escore:

i) $y_i \notin \{0, m_i\}$:

$$\frac{\partial l}{\partial c} = \sum_{i=1}^{n}\left\{\sum_{r=0}^{y_i-1}\frac{\pi_i}{c\pi_i+r} + \sum_{s=0}^{m_i-y_i-1}\frac{1-\pi_i}{c(1-\pi_i)+s} - \sum_{t=0}^{m_i-1}\frac{1}{c+t}\right\},$$

$$\frac{\partial l}{\partial \beta_0} = c\sum_{i=1}^{n}\left\{\sum_{r=0}^{y_i-1}\frac{\pi_i(1-\pi_i)}{c\pi_i+r} - \sum_{s=0}^{m_i-y_i-1}\frac{\pi_i(1-\pi_i)}{c(1-\pi_i)+s}\right\}$$

e

$$\frac{\partial l}{\partial \beta_1} = c\sum_{i=1}^{n}\left\{\sum_{r=0}^{y_i-1}\frac{x_i\pi_i(1-\pi_i)}{c\pi_i+r} - \sum_{s=0}^{m_i-y_i-1}\frac{x_i\pi_i(1-\pi_i)}{c(1-\pi_i)+s}\right\};$$

ii) $y_i = 0$:

$$\frac{\partial l}{\partial c} = \sum_{i=1}^{n}\sum_{t=0}^{m_i-1}\left\{\frac{1-\pi_i}{c(1-\pi_i)+t} - \frac{1}{c+t}\right\},$$

$$\frac{\partial l}{\partial \beta_0} = -c\sum_{i=1}^{n}\sum_{t=0}^{m_i-1}\frac{\pi_i(1-\pi_i)}{c(1-\pi_i)+t}$$

e

$$\frac{\partial l}{\partial \beta_1} = -c\sum_{i=1}^{n}\sum_{t=0}^{m_i-1}\frac{x_i\pi_i(1-\pi_i)}{c(1-\pi_i)+t};$$

iii) $y_i = m_i$:

$$\frac{\partial l}{\partial c} = \sum_{i=1}^{n}\sum_{r=0}^{y_i-1}\left\{\frac{\pi_i}{c\pi_i+r} - \frac{1}{c+r}\right\},$$

$$\frac{\partial l}{\partial \beta_0} = c\sum_{i=1}^{n}\sum_{r=0}^{y_i-1}\frac{\pi_i(1-\pi_i)}{c\pi_i+r}$$

e

$$\frac{\partial l}{\partial \beta_1} = c\sum_{i=1}^{n}\sum_{r=0}^{y_i-1}\frac{x_i\pi_i(1-\pi_i)}{c\pi_i+r}.$$

A matriz de informação de Fisher esperada pode ser estimada por $\mathbf{K}(\widehat{\boldsymbol{\theta}})$, como em (2.7) cujos elementos seguem abaixo:

i) $y_i \notin \{0, m_i\}$:

$$\frac{\partial^2 l}{\partial c^2} = \sum_{i=1}^{n}\left\{\sum_{t=0}^{m_i-1}\frac{1}{(c+t)^2} - \sum_{r=0}^{y_i-1}\frac{\pi_i^2}{(c\pi_i+r)^2} - \sum_{s=0}^{m_i-y_i-1}\frac{(1-\pi_i)^2}{(c(1-\pi_i)+s)^2}\right\},$$

$$\begin{aligned}\frac{\partial^2 l}{\partial c\partial \beta_0} &= \sum_{i=1}^{n}\sum_{r=0}^{y_i-1}\left\{\frac{\pi_i(1-\pi_i)}{c\pi_i+r} - \frac{c\pi_i^2(1-\pi_i)}{(c\pi_i+r)^2}\right\} \\ &\quad - \sum_{i=1}^{n}\sum_{s=0}^{m_i-y_i-1}\left\{\frac{\pi_i(1-\pi_i)}{c(1-\pi_i)+s} - \frac{c\pi_i(1-\pi_i)^2}{(c(1-\pi_i)+s)^2}\right\},\end{aligned}$$

$$\frac{\partial^2 l}{\partial c \partial \beta_1} = \sum_{i=1}^{n}\sum_{r=0}^{y_i-1}\left\{\frac{x_i\pi_i(1-\pi_i)}{c\pi_i+r} - \frac{cx_i\pi_i^2(1-\pi_i)}{(c\pi_i+r)^2}\right\} - \sum_{i=1}^{n}\sum_{s=0}^{m_i-y_i-1}\left\{\frac{x_i\pi_i(1-\pi_i)}{c(1-\pi_i)+s} - \frac{cx_i\pi_i(1-\pi_i)^2}{(c(1-\pi_i)+s)^2}\right\},$$

$$\frac{\partial^2 l}{\partial \beta_0^2} = \sum_{i=1}^{n}\sum_{r=0}^{y_i-1}\left\{\frac{c\pi_i(1-\pi_i)(1-2\pi_i)}{c\pi_i+r} - \left(\frac{c\pi_i(1-\pi_i)}{c\pi_i+r}\right)^2\right\} - \sum_{i=1}^{n}\sum_{s=0}^{m_i-y_i-1}\left\{\frac{c\pi_i(1-\pi_i)(1-2\pi_i)}{c(1-\pi_i)+s} + \left(\frac{c\pi_i(1-\pi_i)}{c(1-\pi_i)+s}\right)^2\right\},$$

$$\frac{\partial^2 l}{\partial \beta_0 \partial \beta_1} = \sum_{i=1}^{n}\sum_{r=0}^{y_i-1}\left\{\frac{cx_i\pi_i(1-\pi_i)(1-2\pi_i)}{c\pi_i+r} - x_i\left(\frac{c\pi_i(1-\pi_i)}{c\pi_i+r}\right)^2\right\} - \sum_{i=1}^{n}\sum_{s=0}^{m_i-y_i-1}\left\{\frac{cx_i\pi_i(1-\pi_i)(1-2\pi_i)}{c(1-\pi_i)+s} + x_i\left(\frac{c\pi_i(1-\pi_i)}{c(1-\pi_i)+s}\right)^2\right\}$$

e

$$\frac{\partial^2 l}{\partial \beta_1^2} = \sum_{i=1}^{n}\sum_{r=0}^{y_i-1}\left\{\frac{cx_i^2\pi_i(1-\pi_i)(1-2\pi_i)}{c\pi_i+r} - \left(\frac{cx_i\pi_i(1-\pi_i)}{c\pi_i+r}\right)^2\right\} - \sum_{i=1}^{n}\sum_{s=0}^{m_i-y_i-1}\left\{\frac{cx_i^2\pi_i(1-\pi_i)(1-2\pi_i)}{c(1-\pi_i)+s} + \left(\frac{cx_i\pi_i(1-\pi_i)}{c(1-\pi_i)+s}\right)^2\right\};$$

ii) $y_i = 0$:

$$\frac{\partial^2 l}{\partial c^2} = \sum_{i=1}^{n}\sum_{t=0}^{m_i-1}\left\{\frac{1}{(c+t)^2} - \left(\frac{1-\pi_i}{c(1-\pi_i)+t}\right)^2\right\},$$

$$\frac{\partial^2 l}{\partial c \partial \beta_0} = \sum_{i=1}^{n}\sum_{t=0}^{m_i-1}\left\{\frac{c\pi_i(1-\pi_i)^2}{(c(1-\pi_i)+t)^2} - \frac{\pi_i(1-\pi_i)}{c(1-\pi_i)+t}\right\},$$

$$\frac{\partial^2 l}{\partial c \partial \beta_1} = \sum_{i=1}^{n}\sum_{t=0}^{m_i-1}\left\{\frac{cx_i\pi_i(1-\pi_i)^2}{(c(1-\pi_i)+t)^2} - \frac{x_i\pi_i(1-\pi_i)}{c(1-\pi_i)+t}\right\},$$

$$\frac{\partial^2 l}{\partial \beta_0^2} = \sum_{i=1}^{n}\sum_{t=0}^{m_i-1}\left\{\frac{c\pi_i(1-\pi_i)(2\pi_i-1)}{c(1-\pi_i)+t} - \left(\frac{c\pi_i(1-\pi_i)}{c(1-\pi_i)+t}\right)^2\right\},$$

$$\frac{\partial^2 l}{\partial\beta_0\partial\beta_1} = \sum_{i=1}^{n}\sum_{t=0}^{m_i-1}\left\{\frac{cx_i\pi_i(1-\pi_i)(2\pi_i-1)}{c(1-\pi_i)+t} - x_i\left(\frac{c\pi_i(1-\pi_i)}{c(1-\pi_i)+t}\right)^2\right\}$$

e

$$\frac{\partial^2 l}{\partial\beta_1^2} = \sum_{i=1}^{n}\sum_{t=0}^{m_i-1}\left\{\frac{cx_i^2\pi_i(1-\pi_i)(2\pi_i-1)}{c(1-\pi_i)+t} - \left(\frac{cx_i\pi_i(1-\pi_i)}{c(1-\pi_i)+t}\right)^2\right\};$$

iii) $y_i = m_i$:

$$\frac{\partial^2 l}{\partial c^2} = \sum_{i=1}^{n}\sum_{r=0}^{y_i-1}\left\{\frac{1}{(c+t)^2} - \left(\frac{\pi_i}{c\pi_i+t}\right)^2\right\},$$

$$\frac{\partial^2 l}{\partial c\partial\beta_0} = \sum_{i=1}^{n}\sum_{r=0}^{y_i-1}\left\{\frac{\pi_i(1-\pi_i)}{c\pi_i+r} - \frac{c\pi_i^2(1-\pi_i)}{(c\pi_i+r)^2}\right\},$$

$$\frac{\partial^2 l}{\partial c\partial\beta_1} = \sum_{i=1}^{n}\sum_{r=0}^{y_i-1}\left\{\frac{x_i\pi_i(1-\pi_i)}{c\pi_i+r} - \frac{cx_i\pi_i^2(1-\pi_i)}{(c\pi_i+r)^2}\right\},$$

$$\frac{\partial^2 l}{\partial\beta_0^2} = \sum_{i=1}^{n}\sum_{r=0}^{y_i-1}\left\{\frac{c\pi_i(1-\pi_i)(1-2\pi_i)}{c\pi_i+r} - \left(\frac{c\pi_i(1-\pi_i)}{c\pi_i+r}\right)^2\right\},$$

$$\frac{\partial^2 l}{\partial\beta_0\partial\beta_1} = \sum_{i=1}^{n}\sum_{r=0}^{y_i-1}\left\{\frac{cx_i\pi_i(1-\pi_i)(1-2\pi_i)}{c\pi_i+r} - x_i\left(\frac{c\pi_i(1-\pi_i)}{c\pi_i+r}\right)^2\right\}$$

e

$$\frac{\partial^2 l}{\partial\beta_1^2} = \sum_{i=1}^{n}\sum_{r=0}^{y_i-1}\left\{\frac{cx_i^2\pi_i(1-\pi_i)(1-2\pi_i)}{c\pi_i+r} - \left(\frac{cx_i\pi_i(1-\pi_i)}{c\pi_i+r}\right)^2\right\}.$$

Referências Bibliográficas

Automotive Industry Action Group (AIAG) (2002). *MSA-3: Measurement Systems Analysis*. Automotive Industry Action Group (AIAG), Michigan.

Bazán, J. L., Bolfarine, H. & Branco, M. D. (2006). A Generalized Skew Probit Class Link for Binary Regression. Relatório Técnico RT-MAE 2006-05, IME - Universidade de São Paulo, São Paulo.

Carroll, R. J., Spiegelman, C. H., Lan, K. K. G., Bailey, K. T. & Abbott, R. D. (1984). On Errors-in-variables for Binary Regression Models. *Biometrika*, **71**(1), 19–25.

Carroll, R. J., Kuchenhoff, H., Lombard, F. & Stefanski, L. A. (1996). Asymptotics for the SIMEX in Nonlinear Measurement Error Models. *Journal of the American Statistical Association*, **91**(433), 242–250.

Carroll, R. J., Ruppert, D., Stefanski, L. A. & Crainiceanu, C. M. (2006). *Measurement Error in Nonlinear Models - A Modern Perspective*. Chapman & Hall/CRC, Boca Raton, FL, second edition.

Cheng, C.-L. & Van Ness, J. W. (1999). *Statistical Regression with Measurement Errors*. Arnold, London.

Collett, D. (2003). *Modelling Binary Data*. Chapman & Hall/CRC, Boca Raton, FL, second edition.

Cook, J. R. & Stefanski, L. A. (1994). Simulation-Extrapolation Estimation in Parametric Measurement Error Models. *Journal of the American Statistical Association*, **89**(428), 1314–1328.

Cox, D. R. & Hinkley, D. V. (1974). *Theoretical Statistics*. Chapman & Hall, London.

Dean, C. B. (1992). Testing for Overdispersion in Poisson and Binomial Regression-Models. *Journal of the American Statistical Association*, **87**(418), 451–457.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society Séries B-Methodological*, **39**(1), 1–38. With discussion.

Demétrio, C. G. B. (2002). *Modelos Lineares Generalizados em Experimentação Agronômica*. ESALQ - Universidade de São Paulo, Piracicaba.

Dey, D. K., Gelfand, A. E. & Peng, F. (1997). Overdispersed Generalized Linear Models. *Journal of Statistical Planning and Inference*, **64**(1), 93–107.

Dolby, G. R. (1976). The Ultraestructural Relation: A Synthesis of the Functional and Structural Relations. *Biometrika*, **63**(1), 39–50.

Doornik, J. A. (2002). *Object-Oriented Matrix Programming Using Ox*. Timberlake Consultants Press and Oxford, London, third edition.

Draggalin, V. & Fedorov, V. (2006). Design of Multi-centre Trials With Binary Response. *Statistics in Medicine*, **25**(16), 2701–2719.

Dávila, V. H. L. (2004). *Modelos Lineares Mistos Assimétricos*. Tese de doutorado, IME – Universidade de São Paulo, São Paulo.

Fahrmeir, L. (1988). A Note on Asymptotic Testing Theory for Nonhomogeneous Observations. *Stochastic Processes and their Applications*, **28**(2), 267–273.

Favari, D. F. (2006). *Uma Aplicação Industrial de Regressão Binária com Erros na Variável Explicativa*. Dissertação de mestrado, ICMC – Universidade de São Paulo, São Carlos.

Freedman, L. S., Fainberg, V., Kipnis, V., Midthune, D. & Carroll, R. J. (2004). A New Method for Dealing with Measurement Error in Explanatory Variables of Regression Models. *Biometrics*, **60**(1), 172–181.

Fuller, W. A. (1987). *Measurement Error Models*. Wiley, New York.

Gill, P. E., Murray, W. & Wright, M. H. (1981). *Practical Optimization*. Academic Press, London.

Gomes, P. L. S. (2004). *Uma Estratégia para Correção de uma Característica de Qualidade de uma Peça devido à Variação de Temperatura*. Dissertação de mestrado, ICMC – Universidade de São Paulo, São Carlos.

Gong, G. & Samaniego, F. J. (1981). Pseudo Maximum Likelihood Estimation: Theory and Applications. *The Annals of Statistics*, **9**(4), 861–869.

Hinde, J. & Demétrio, C. G. B. (1998a). *Overdispersion: Models and Estimation*. 13º SINAPE, Caxambu. Associação Brasileira de Estatística, São Paulo.

Hinde, J. & Demétrio, C. G. B. (1998b). Overdispersion: Models and Estimation. *Computational Statistics & Data Analysis*, **27**(2–3), 151–170.

ISO (1997). *ISO/IEC Guide 43-1: Guide to the Expression of Uncertainty in Measurement – Proficiency Testing by Interlaboratory Comparisons – Part I: Development and Operation of Proficiency Testing Schemes*. International Organization for Standardization, Geneva.

Kim, B. R., Carter, R., Rao, P., Ariet, M. & Resnick, M. (2006). Standardized Risk

and Description of Results From Multivariable Modeling of a Binary Response. *Biometrical Journal*, **48**(1), 54–66.

Kim, B. S. & Margolin, B. H. (1992). Testing Goodness of Fit of a Multinomial Model Against Overdispersed Alternatives. *Biometrics*, **48**(3), 711–719.

Lambert, D. & Roeder, K. (1995). Overdispersion Diagnostics for Generalized Linear Models. *Journal of the American Statistical Association*, **90**(432), 1225–1236.

Li, T. (2006). A Unified View on Clustering Binary Data. *Machine Learning*, **62**(3), 199–215.

Lin, X. & Breslow, N. E. (1996). Bias Correction in Generalized Linear Mixed Models with Multiple Components of Dispersion. *Journal of the American Statistical Association*, **91**(435), 1007–1016.

Luenberger, D. G. (1984). *Linear and Nonlinear Programming*. Addison - Wesley, Boston, second edition.

McCullagh, P. & Nelder, J. A. (1989). *Generalized Linear Models*. Chapman & Hall, London, second edition.

Parke, W. R. (1986). Pseudo-maximum-likelihood Estimation: The Asymptotic Distribution. *The Annals of Statistics*, **1**(14), 355–357.

Paul, S. R., Liang, K. Y. & Self, S. G. (1989). On Testing Departure from the Binomial and Multinomial Assumptions. *Biometrics*, **45**(1), 231–236.

Paula, G. A. (2004). *Modelos de Regressão com Apoio Computacional*. IME - Universidade de São Paulo, São Paulo.

Paula, G. A. & Artes, R. (2000). One-Sided Test to Assess Correlation in Linear Logistic Models using Estimating Equations. *Biometrical Journal*, **42**(6), 701–714.

Piessens, R., Doncker-Kapenga, E., Überhuber, C. & Kahaner, D. (1983). *QUADPACK, A Subroutine Package for Automatic Integration*. Springer-Verlag, New York.

Press, W. H., Teukolsky, S. A., Vetterling, W. T. & Flannery, B. P. (1994). *Numerical Recipes in Fortran: The Art of Scientific Computing*. Cambridge University Press, New York, second edition.

R Development Core Team (2006). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Schafer, D. W. (1987). Covariate Measurement Error in Generalized Linear Models. *Biometrika*, **74**(2), 385–391.

Schafer, D. W. (1993). Likelihood Analysis for Probit Regression with Measurement Errors. *Biometrika*, **80**(4), 899–904.

Sen, P. K. & Singer, J. M. (1993). *Large Sample Methods in Statistics: An Introduction With Applications*. Chapman & Hall, New York.

Thoresen, M. & Laake, P. (2000). A Simulation Study of Measurement Error Correction Methods in Logistic Regression. *Biometrics*, **56**(3), 868–872.

Verbeke, T. & De Clercq, M. (2006). The Income-environment Relationship: Evidence From a Binary Response Model. *Ecological Economics*, **59**(4), 419–428.

Wang, N., Lin, X., Gutierrez, R. G. & Carroll, R. J. (1998). Bias Analysis and SIMEX Approach in Generalized Linear Mixed Measurement Error Models. *Journal of the American Statistical Association*, **93**(441), 249–261.

Williams, D. A. (1982). Extra-Binomial Variation in Logistic Linear Models. *Applied Statistics*, **31**(2), 144–148.

Zavala, A. A. Z. (2001). *Análise Comparativa dos Algoritmos EM e SIMEX nos Modelos Lineares Mistos Aplicados a Análise de Regressão com Erros nas Variáveis*. Dissertação de mestrado, IME – Universidade de São Paulo, São Paulo.