Regressão logística com erro de medida: comparação de métodos de estimação

Agatha Sacramento Rodrigues

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

Programa: Estatística

Orientadora: Prof[a]. Dr[a]. Silvia Lopes de Paula Ferrari

Durante o desenvolvimento deste trabalho a autora recebeu auxílio financeiro do CNPq

São Paulo, julho de 2013

Regressão logística com erro de medida: comparação de métodos de estimação

Esta versão da dissertação contém as correções e alterações sugeridas pela Comissão Julgadora durante a defesa da versão original do trabalho, realizada em 27/06/2013. Uma cópia da versão original está disponível no Instituto de Matemática e Estatística da Universidade de São Paulo.

Comissão Julgadora:

- Profª. Drª. Silvia Lopes de Paula Ferrari (orientadora) - IME-USP
- Prof. Dr. Carlos Alberto Ribeiro Diniz - UFSCar
- Prof. Dr. Mário de Castro Andrade Filho - ICMC-USP

Agradecimentos

Neste momento, me passa um filme na cabeça, relembrando todos os momentos, conversas, participações em congressos e trocas de emails que, de algum modo, constituem este trabalho.

Gostaria de agradecer e dedicar esse trabalho à minha família. Sempre tão presente em minha vida, me apoiando, dando força e compreendendo algumas ausências. Sem dúvidas, sem o grande esforço dos meus pais eu não estaria aqui, escrevendo os agradecimentos de uma dissertação de mestrado. Obrigada, minha base.

Agradeço meu namorado Nicholas Wagner Eugenio por estar sempre ao meu lado nesses três anos de muito companheirismo.

Agradeço a professora doutora Silvia Lopes de Paula Ferrari pela orientação acadêmica, paciência e compreensão na orientação deste trabalho. Em especial, por entender quando comecei a trabalhar concomitantemente ao mestrado. Serei sempre grata.

Agradeço ao professor doutor Carlos Alberto Ribeiro Diniz por algumas conversas sobre este trabalho, que é uma continuação do meu trabalho de conclusão de gradução, realizado sob a sua orientação.

Agradeço ao professor doutor Mário de Castro Andrade Filho e novamente ao professor doutor Carlos Alberto Ribeiro Diniz pelas contribuições e arguições durante a defesa.

Meus sinceros agradecimentos ao Ettore Enrico Delfino Ligorio e LCCA (Laboratório de Computação Científica Avançada da USP) por todo o apoio na utilização do *cluster* Puma. Ettore é um funcionário como poucos conheci na USP. Sempre disposto a ajudar e incansável no trabalho de aprender e instalar o *software* SAS nos *clusters*, uma vez que fui a primeira a utilizar o *software* no sistema de filas. Muito obrigada, Ettore.

Agradeço Guilherme Fernandes pelo auxílio com os dados da instituição financeira.

Por toda a explicação, conversas e trocas de emails, sempre muito atencioso.

Meus agradecimentos à professora doutora Annamaria Guolo pela trocas de emails tão importantes para a implementação do algoritmo EM-Monte Carlo.

Agradeço ao professor doutor Jalmar Manuela Farfán Carrasco que me auxiliou com sugestões no início deste trabalho e disponilizou o conjunto de dados da área médica.

Agradeço Paulo Henrique Ferreira da Silva por disponibilizar materiais de seu artigo.

Agradeço ao professor doutor Alexandre Patriota pela disponibilização de um conjunto de dados, que acabamos por não utilizar neste trabalho, e pelas conversas e trocas de emails.

Meus agradecimentos aos professores doutores Luis Gustavo Esteves, Anatoli Iambartsev, Julio Singer, Florencia Leonardi, Gilberto Alvarenga Paula e Lúcia Pereira Barroso que contribuíram com as disciplinas que cursei durante o mestrado.

Gostaria de agradecer a bolsa de estudos do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) que tive nos primeiros oito meses do mestrado e também o apoio financeiro do Instituto de Matemática e Estatística para a participação em congressos.

Agradeço a meus amigos e colegas do instituto Raony Cassab, Brian Alvarez, Vinicius Calsavara, Fábio Oki, Eliardo Costa, William Gonzalo, Andrés Rosso, Elisângela Rodrigues, Josemir Almeida, Bruno Monte, Victor Fossaluza, Soane Mota, Leandro Correia, Márcio Diniz, Melaine Oliveira, Andressa Cerqueira, Ana Paula Zerbeto, entre muitos outros que tive o imenso prazer de conviver nesses dois últimos anos.

Meus agradecimentos aos amigos e colegas que cultivei durante o trabalho no Instituto de Psicologia da USP. Sem dúvidas, o que aprendi nesses dezessete meses de trabalho complementam minha formação no mestrado. Em especial, agradeço ao Luiz Silva dos Santos, Vinicius David Frayze, Tania Kiehl Lucci, Cynthia Braga, Joana D Arc Lima, Valeria Campos, Maria Betânia da Costa, Wilma Maria Rodrigues e Alessandra Piccolo por todas as conversas e almoços muito divertidos.

Agradeço a Deus por tudo em minha vida. Agradeço por todo o trabalho dispendido no mestrado, pois, como consequência, conheci pessoas que quero levar para toda a vida.

A todos que de algum modo contribuíram para este momento, muito obrigada!

Resumo

Modelo de regressão logística com erro de medida: comparação de métodos de estimação.

Neste trabalho estudamos o modelo de regressão logística com erro de medida nas covariáveis. Abordamos as metodologias de estimação de máxima pseudoverossimilhança pelo algoritmo EM-Monte Carlo, calibração da regressão, SIMEX e *naïve* (ingênuo), método este que ignora o erro de medida. Comparamos os métodos em relação à estimação, através do viés e da raiz do erro quadrático médio, e em relação à predição de novas observações, através das medidas de desempenho sensibilidade, especificidade, verdadeiro preditivo positivo, verdadeiro preditivo negativo, acurácia e estatística de Kolmogorov-Smirnov. Os estudos de simulação evidenciam o melhor desempenho do método de máxima pseudoverossimilhança na estimação. Para as medidas de desempenho na predição não há diferença entre os métodos de estimação. Por fim, utilizamos nossos resultados em dois conjuntos de dados reais de diferentes áreas: área médica, cujo objetivo está na estimação da razão de chances, e área financeira, cujo intuito é a predição de novas observações.

Palavras-Chave: Calibração da regressão, Estimação por máxima pseudoverossimilhança, Medidas de desempenho na predição, Modelo de regressão logística, Modelos com erro de medida, SIMEX.

Abstract

Logistic regression model with measurement error: a comparison of estimation methods.

We study the logistic model when explanatory variables are measured with error. Three estimation methods are presented, namely maximum pseudo-likelihood obtained through a Monte Carlo expectation-maximization type algorithm, regression calibration, SIMEX and naïve, which ignores the measurement error. These methods are compared through simulation. From the estimation point of view, we compare the different methods by evaluating their biases and root mean square errors. The predictive quality of the methods is evaluated based on sensitivity, specificity, positive and negative predictive values, accuracy and the Kolmogorov-Smirnov statistic. The simulation studies show that the best performing method is the maximum pseudo-likelihood method when the objective is to estimate the parameters. There is no difference among the estimation methods for predictive purposes. The results are illustrated in two real data sets from different application areas: medical area, whose goal is the estimation of the odds ratio, and financial area, whose goal is the prediction of new observations.

Keywords: Logistic regression model, Maximum pseudo-likelihood estimation, Measurement error models, Predictive measures, Regression calibration estimation, SIMEX estimation.

Sumário

Capítulo 1

Introdução

Modelos de regressão são utilizados em diversas áreas com o intuito de relacionar uma variável resposta a variáveis explicativas (ou covariáveis). Em certas aplicações, a variável de interesse tem como resposta "sucesso" ou "fracasso" (o evento de interesse é o sucesso), ou seja, é uma variável binária e a sua variabilidade é explicada pelas covariáveis através de um modelo linear generalizado, como o modelo de regressão logística (McCullagh & Nelder, 1989). Na prática, é comum que pelo menos uma variável explicativa não seja observada de forma exata, mas sim com algum tipo de erro de medida. Como relatam Cunha & Colosimo (2003), são os seguintes os possíveis motivos para tal erro:

- Erro de respostas em métodos de coleta de dados, como entrevistas ou questionários, causado por confusão, ignorância, por falta de cuidado, gerados por falta de treinamento adequado ou mesmo pelo método usado para obter a resposta.
- Erro de coleta dos dados por falha nos equipamentos, devido a desgastes dos componentes, falta de calibração ou a condições ambientais, que geram variabilidade em instrumentos de leitura.
- Tempo e custo oriundos da observação da variável de interesse são inviáveis para o estudo.
- Processamento inadequado de dados ou perda de informações.
- Outros problemas que podem ocorrer após a coleta de dados.

Em diversos problemas práticos há covariáveis com erro de medida, principalmente nas áreas médica, epidemiológica e nutricional. Carroll *et al.* (2006, Cap. 1) apresentam diversos exemplos em que o erro de medida está presente em dados de resposta binária. Dentre esses, destacamos os dados do estudo epidemiológico NHANES-I (NHANES-I *Epidemiologic Study Cohort*). A variável resposta do i-ésimo indivíduo, Y_i, indica a presença de câncer de mama. Além das variáveis medidas sem erro, chamamos de Z_i, como idade, histórico de câncer na família e menopausa (sim ou não), variáveis nutricionais são de interesse no estudo, denotadas por X_i. Se estas variáveis fossem observadas sem erro, a análise do modelo de regressão logística usual poderia ser realizada (Hosmer & Lemeshow, 2000, Cap. 1 e 2), mas essas variáveis são de difícil medição e são observados os relatos de cada indivíduo sobre a alimentação nas últimas 24 horas, W_i. Desta maneira, W_i pode ser vista como covariáveis observadas com erro.

Na prática, muitas vezes W_i é utilizada no modelo, ignorando o fato de o interesse ser X_i e resultando em um estimador ingênuo (*naïve*). Ignorar erros de medida pode levar a estimadores viesados e até inconsistentes. O impacto do erro de medida tem sido objeto de estudo de muitos pesquisadores; vide, por exemplo, Stefanski & Carroll (1985), Carroll *et al.* (1984), Schneeweiss & Augustin (2006) e Gustafson (2004, Cap. 2). Como consequência, algumas técnicas de correção de estimadores na presença de erro de medida foram propostas nos últimos tempos. Fuller (1987) analisa o problema nos modelos de regressão linear e para o caso dos modelos não-lineares, uma revisão detalhada dos métodos de estimação pode ser vista em Carroll *et al.* (2006). Entre esses métodos estão o de máxima verossimilhança, máxima pseudoverossimilhança, SIMEX e calibração da regressão. O método de máxima verossimilhança produz estimativas ao maximizar a função de verossimilhança, o que leva a estimadores com boas propriedades (Schafer, 2002). Entretanto, em algumas situações, por razão da complexidade da função de verossimilhança, esse método é de difícil aplicação dado o imenso esforço computacional envolvido. Gong & Samaniego (1981) propõem a maximização de uma pseudoverossimilhança, que é uma modificação da função de verossimilhança e depende apenas dos parâmetros de interesse. Guolo (2010), por exemplo, utiliza o algoritmo numérico EM-Monte Carlo para estimar os parâmetros por este método e Carrasco *et al.* (2013) utilizam uma aproximação para a integral pelo método de quadratura de Gauss Hermite. Um método simples e de fácil implementação é abordado por Cook & Stefanski (1994) e Stefanski & Cook (1995), chamado de SIMEX (*Simulation-Extrapolation*), cuja ideia é

introduzir um erro de medida adicional aos dados por um esquema de reamostragem, de tal forma que se possa estabelecer uma relação entre o viés induzido pelo erro de medida e a variância do erro de medida adicionado e extrapolar essa relação, através de um modelo, para o caso correspondente à ausência de erro de medida. Tieppo (2007) utiliza este método no modelo de regressão beta-binomial com erro nas covariáveis. Carroll *et al.* (2006) propõem substituir a verdadeira covariável não observada por uma estimativa da esperança condicional de X_i dado W_i, usando réplicas ou dados de validação (*validation study*) ou de reprodutibilidade (*reproducibility study*). Após a substituição, as estimativas dos parâmetros são obtidas da maneira usual, como por máxima verossimilhança. Este método, conhecido como calibração da regressão, é utilizado em diversos trabalhos que estudam modelos de regressão com erro de medida. Podemos citar Rosner *et al.* (1989) que estudam o método usando dados de validação, Rosner *et al.* (1992) utilizam dados de reprodutibilidade e Thoresen & Laake (2007) utilizam réplicas de W_i para estimar a variância do erro de medida. Já Thoresen (2006) considera um estudo de simulação com múltiplas covariáveis medidas com erro e Rosner *et al.* (1990) aplicam a metodologia a dados epidemiológicos. Ainda em relação ao método da calibração da regressão, Spiegelman *et al.* (2011) sugerem uma adaptação do estimador na presença de erro de medida heteroscedástico.

As diferentes áreas de aplicação do modelo de regressão logística têm diferentes objetivos. Na área médica, a estimação dos parâmetros é importante, pois um resultado de interesse é, em geral, a razão de chances *(odds ratio)*. Na área financeira, por outro lado, o objetivo central é a predição de resultados futuros.

Diversos trabalhos estudam medidas de desempenho na predição do modelo logístico sem a presença de erro. Podemos citar Louzada *et al.* (2012), que estudam medidas de acerto da predição com correção dependente de *status* (*state-dependent*), e Diniz & Louzada (2012, Cap. 3), que estudam as medidas de desempenho na modelagem para eventos raros.

Sob a abordagem do erro de medida, há vasta literatura sobre o seu impacto na estimação dos parâmetros do modelo de regressão logística. Entretanto, o impacto do erro de medida nas medidas de desempenho da predição de novas observações não é abordado. Dentre os trabalhos levantados no estudo bibliográfico, apenas Rodrigues (2010) estudou as medidas de qualidade na predição em modelos de regresssão logística com erro de medida, cuja variância é considerada conhecida.

Este trabalho objetiva estudar o modelo de regressão logística tanto do ponto de vista da estimação quanto da predição, com o intuito de comparar os seguintes métodos de estimação: calibração da regressão, SIMEX, máxima pseudoverossimilhança e o método *naïve*.

No estudo de simulação, os métodos são comparados em relação ao viés e à raiz do erro quadrático médio na estimação. Para a predição, as seguinte medidas de desempenho são utilizadas: sensibilidade, especificidade, verdadeiro preditivo positivo, verdadeiro preditivo negativo e acurácia, como propõem Louzada-Neto *et al.* (2009) e a estatística de Kolmogorov-Smirnov, como propõem Diniz & Louzada (2012, Cap. 1).

1.1 Motivação

Neste trabalho são duas as motivações do estudo do modelo de regressão logística com erro de medida. Uma motivação oriunda de dados da área médica e outra tem origem em dados de uma instituição financeira. Esses casos práticos são descritos a seguir.

1.1.1 Dados da área médica

O Instituto Nacional do Coração dos Estados Unidos, conhecido como *National Heart, Lung, and Blood Institute*, em conjunto com a Universidade de Boston, liderou um ambicioso projeto de pesquisa em saúde nos Estados Unidos, o estudo *Framingham Heart Study*, que teve início nos anos 40 do século XX. Na época, pouco se sabia sobre as causas gerais de doenças cardíacas e acidente vascular cerebral, mas as taxas de mortalidade em razão dessas doenças aumentaram progressivamente desde o início do século e se tornaram uma epidemia americana. O objetivo do estudo era identificar fatores ou características comuns que contribuem para doenças cardiovasculares, seguindo o seu desenvolvimento ao longo de um período de tempo em um grande grupo de participantes que não haviam desenvolvido sintomas evidentes das doenças. Mais detalhes sobre o estudo podem ser obtidos em `http://www.framinghamheartstudy.org`.

Como motivação para aplicação da metodologia estudada em dados da área médica, consideramos um dos conjuntos de dados do estudo, cujo interesse é verificar o impacto de fatores de risco em doenças coronarianas em homens americanos. O estudo consiste na

observação de 1615 participantes de 31 a 65 anos de idade nos quais se observaram os possíveis fatores de risco para as doenças: idade (em anos), se fumante ou não, nível de colesterol (em mg/dL) e pressão sistólica (em mmHg) obtidos por uma série de exames. Oito anos depois foi observado se o participante desenvolveu ou não sintomas de doenças coronarianas. Como esta variável de interesse é binária, poderíamos utilizar o modelo de regressão logística usual (Hosmer & Lemeshow, 2000), mas a pressão sistólica verdadeira não pode ser medida exatamente, pois varia ao longo do dia para o mesmo indivíduo. Desta maneira, vemos por conveniente o uso do modelo de regressão logística com erro de medida na covariável.

1.1.2 Dados de uma instituição financeira

Outra motivação deste trabalho tem origem em dados financeiros. Uma instituição financeira brasileira tem o interesse em classificar clientes como adimplentes ou inadimplentes de acordo com um modelo estatístico. Muitas técnicas de modelagem têm sido objeto de estudos com o intuito de diminuir o erro de classificação e auxiliar os gestores na tomada de decisões. Dentre algumas covariáveis que possam predizer a classificação do cliente, a renda é um importante preditor, mas não é de conhecimento exato da instituição em questão. O que se tem disponível é a renda presumida desses clientes, ou seja, a observação da variável medida com erro. Desta maneira, temos o interesse em predizer uma variável binária (adimplente ou inadimplente) em que uma das covariáveis do modelo é observada com erro de medida. Mais uma vez, parece ser conveniente o uso do modelo de regressão logística com erro de medida na covariável.

1.2 Organização dos capítulos e objetivos específicos

No Capítulo 2 apresentamos uma breve revisão do modelo de regressão logística sem erro de medida, incluindo a estimação dos parâmetros, a interpretação da razão de chances e as medidas de desempenho da predição e a estatística de Kolmogorov-Smirnov. No Capítulo 3 apresentamos os principais conceitos do modelo com erro de medida, como a definição do modelo para o erro de medida, diferenças entre os modelos funcional e estrutural, o modelo geral sob o enfoque estrutural e as definições de erros diferencial e não diferencial. O Capítulo 4 aborda a teoria dos métodos de estimação considerados: calibração da regressão, SIMEX,

naïve, máxima verossimilhança e máxima pseudoverossimilhança.

O Capítulo 5 apresenta um estudo de simulação sob seis cenários para comparar os métodos de estimação do ponto de vista de estimação e predição. Destacamos que existem dois objetivos específicos no estudo. O primeiro é verificar o comportamento dos métodos de estimação com diferentes proporções de eventos da variável resposta. O que motivou este estudo é o fato de alguns trabalhos na literatura revelarem que o modelo de regressão logística usual subestima a probabilidade do evento de interesse quando este é construído utilizando conjuntos de dados extremamente desbalanceadas (King & Zeng, 2001). Diniz & Louzada (2012, p. 54 e 55), utilizando proporções de eventos de 1%, 15%, 30% e 50%, mostram que o modelo de regressão logística usual não é adequado para ajustar dados com desbalanceamento acentuado. E ainda, Louzada *et al.* (2012) mostram que as medidas de desempenho na predição diminuem conforme aumenta o desbalanceamento da amostra. O segundo objetivo específico é motivado pelo trabalho Spiegelman *et al.* (2011), que estudaram o método da calibração da regressão sob heteroscedasticidade do erro de medida, propondo uma correção para o estimador. Eles verificaram que o método da calibração da regressão usual (que não considera heteroscedasticidade) apresenta desempenho melhor ou igual ao estimador proposto. Desta maneira, queremos verificar como os outros métodos de estimação considerados se comportam na presença de erros de medida heteroscedásticos.

Nos Capítulos 6 e 7 aplicamos os resultados aos conjuntos de dados apresentados nas Seções 1.1.1 e 1.1.2, respectivamente. No Capítulo 8 apresentamos nossas conclusões e indicamos direções para novas pesquisas. No Apêndice A obtemos a variância dos estimadores considerados neste trabalho.

1.3 Suporte Computacional

O software utilizado é o `SAS`, versão 9.3 (`http://www.sas.com`). Utilizamos também o software `R` (R Development Core Team, 2013) para a realização de gráficos. O `R` é um software livre e está disponível em `http://www.r-project.org`.

A pesquisa foi desenvolvida com o auxílio do LCCA-Laboratório de Computação Científica Avançada da Universidade de São Paulo com a utilização do *cluster* Puma. Detalhes podem ser encontrados em `http://www.usp.br/lcca/tdi`.

Capítulo 2

Regressão logística

Em muitas pesquisas o objetivo é relacionar variáveis independentes (covariáveis) a uma variável dependente binária. Nesses casos, a regressão linear não é apropriada uma vez que a suposição de erros normalmente distribuídos é violada. Regressão logística é o método padrão na modelagem de uma resposta binária, isto é, a variável dependente possui apenas dois resultados: sucesso (ou evento ou evento de interesse), usualmente codificado como 1, e fracasso (que também chamaremos de não-evento), codificado como 0. A razão pela codificação em 0/1 é o fato de a variável aleatória com distribuição de Bernoulli ser igual a 1 se ocorre o evento e 0 se ocorre o não-evento.

Muitas situações práticas envolvem esse tipo de variável resposta. Aplicações em dados médicos podem ser encontrados em Hosmer & Lemeshow (2000, Cap. 1) e a utilização de regresão logística em biometria, em Kim *et al.* (2006). Breslow (1996) apresenta uma aplicação em epidemiologia. Já Copas & Loeber (1990) apresentam dados binários em pesquisas psicológicas e Verbeke & Clercq (2006), em economia.

2.1 Modelo de regressão logística

Sejam Y_i e X_i a i-ésima variável resposta e o vetor de covariáveis do i-ésimo indivíduo, respectivamente, com $i = 1, \ldots, n$. A distribuição de Y_i dada a observação x_i de X_i é de Bernoulli, ou seja,

$$Y_i | X_i = x_i \sim \mathrm{Ber}(\pi(\mathrm{x_i})).$$

Para simplificar a notação, escrevemos $\pi(x_i) = \pi_i$. A função de probabilidade de $Y_i|X_i = x_i$ é dada por

$$f_{Y|X}(y_i|x_i) = {\pi_i}^{y_i}(1-\pi_i)^{1-y_i}, \quad y_i = 0, 1;$$

x_i representa o vetor coluna de p covariáveis observadas do i-ésimo indivíduo. Tipicamente o primeiro elemento de x_i vale 1 para permitir que o modelo tenha intercepto não nulo. A probabilidade de sucesso, π_i, é relacionada com um vetor de $p+1$ parâmetros desconhecidos, $\beta = (\beta_0, \ldots, \beta_p)^\top$, através da ligação logito, como descrevem Hosmer & Lemeshow (2000, p. 31), isto é,

$$g(\pi_i) = \log \frac{\pi_i}{1-\pi_i} = x_i^\top \beta,$$

em que $g(.)$ é chamada de função de ligação. Consequentemente,

$$\pi_i = \frac{\exp(x_i^\top \beta)}{1 + \exp(x_i^\top \beta)}, \quad i = 1, \ldots, n. \tag{2.1}$$

2.2 Estimação

Para uma amostra de n observações $(Y_1, X_1), \ldots, (Y_n, X_n)$, a função de verossimilhança baseada na distribuição de $(Y_1, \ldots, Y_n)$ dada a observação de $(x_1, \ldots, x_n)$ de $(X_1, \ldots, X_n)$ é dada por

$$L(\beta) = \prod_{i=1}^{n} {\pi_i}^{y_i}(1-\pi_i)^{1-y_i},$$

e o logaritmo da função de verossimilhança é, portanto,

$$l(\beta) = \sum_{i=1}^{n} [y_i \log \pi_i + (1-y_i) \log (1-\pi_i)].$$

A ligação logito, função de ligação considerada aqui, é a ligação canônica do modelo de Bernoulli. As ligações canônicas garantem a concavidade de $l(\beta)$ e consequentemente garantem a unicidade da estimativa de máxima verossimilhança de β, se esta existir; vide Paula (2013, p. 8). Para encontrar o valor de β que maximiza $l(\beta)$, derivamos $l(\beta)$ com relação a $\beta = (\beta_0, \ldots, \beta_p)^\top$ e igualamos a zero, obtendo as equações de verossimilhança

$$\sum_{i=1}^{n} x_i(y_i - \pi_i) = 0,$$

em que π_i é dada por (2.1).

As equações de verossimilhança são não-lineares e requerem métodos iterativos para a solução do sistema de equações. Muitos, senão todos, programas estatísticos utilizam algum método numérico para estimação de β. Os *softwares* `R`, na função "glm.fit" e `SAS`, na *procedure* "proc logistic", utilizam o método escore de Fisher. Neste método, o processo iterativo é dado por

$$\beta^{(m+1)} = \beta^{(m)} + [I(\beta)^{-1}]^{(m)} U^{(m)}, \quad m = 0, 1, \ldots, \tag{2.2}$$

em que $U^{(m)}$ e $[I(\beta)^{-1}]^{(m)}$ são a função escore e a inversa da matriz de informação de Fisher no m-ésimo passo, respectivamente. Aqui, $U = X^\top(y - \pi)$ e $I(\beta) = X^\top V X$, em que

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p} \\ 1 & x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{np} \end{bmatrix}, \quad y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \pi = \begin{bmatrix} \pi_1 \\ \pi_2 \\ \vdots \\ \pi_n \end{bmatrix}$$

e

$$V = \mathrm{diag}[\pi_1(1-\pi_1),\ \pi_2(1-\pi_2), \ldots,\ \pi_{\mathrm{n}}(1-\pi_{\mathrm{n}})].$$

A partir de (2.2) chega-se ao processo iterativo de mínimos quadrados ponderados

$$\beta^{(m+1)} = (X^\top V^{(m)} X)^{-1} X^\top V^{(m)} z^{(m)}, \quad m = 0, 1, \ldots,$$

em que $z = \eta + V^{-1}(y - \pi)$ com $\eta = X\beta$. O processo iterativo é seguido até obter a convergência. Um critério de parada é, por exemplo,

$$\left| \frac{\beta_j^{(m+1)} - \beta_j^{(m)}}{\beta_j^{(m)}} \right| < \epsilon, \quad \forall j = 0, 1, \ldots, p.$$

Após a convergência, a probabilidade de sucesso estimada é

$$\widehat{\pi}_i = \frac{\exp(x_i^\top \widehat{\beta})}{1 + \exp(x_i^\top \widehat{\beta})}, \quad i = 1, \ldots, n.$$

em que $\widehat{\beta}$ é a estimativa de máxima verossimilhança de β. Mais detalhes do método escore de Fisher podem ser encontrados em Hilbe (2009, Cap. 3 e 4). Em grandes amostras, $\widehat{\beta} - \beta \sim \mathrm{N_{p+1}}(0, I(\beta)^{-1})$, aproximadamente.

2.3 Interpretação dos parâmetros – razão de chances

A função de ligação logito é a mais utilizada no ajuste do modelo de Bernoulli. Outras funções de ligação, como probito e complementar log-log, podem apresentar bom ajuste mas a ligação logito é capaz de fornecer uma interpretação conveniente dos parâmetros.

A chance de ocorrer o evento, dada a ligação logito em (2.1), é dada por

$$\frac{\pi_i}{1-\pi_i} = \exp(x_i^\top \beta) = \exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}).$$

Se uma variável independente contínua, por exemplo x_1, for acrescida de uma unidade, mantendo as outras variáveis independentes do modelo fixas, a chance do evento fica

$$\begin{aligned}
\frac{\pi_i^*}{1-\pi_i^*} &= \exp(\beta_0 + \beta_1(x_{i1}+1) + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}) \\
&= \exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \beta_1) \\
&= \exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}) \exp(\beta_1) \\
&= \frac{\pi_i}{1-\pi_i} \exp(\beta_1). \qquad (2.3)
\end{aligned}$$

Assim, a razão de chances (*odds ratio*) de (x_1+1) em relação a x_1, $\mathrm{OR}(\mathrm{x}_1+1, \mathrm{x}_1)$ é dada por

$$\mathrm{OR}(\mathrm{x}_1+1, \mathrm{x}_1) = \frac{\pi_\mathrm{i}^*/(1-\pi_\mathrm{i}^*)}{\pi_\mathrm{i}/(1-\pi_\mathrm{i})} = \exp(\beta_1),$$

ou seja, a chance do evento ocorrer entre os indivíduos que diferem na variável x_1 em 1 unidade é $\exp(\beta_1)$. Neste caso, a estimativa da razão de chances é $\widehat{OR}(x_1+1, x_1) = \exp(\widehat{\beta}_1)$. De uma forma geral, a estimativa da razão de chances com o acréscimo de c unidades, ou seja, substituindo $(x_{i1}+1)$ por $(x_{i1}+c)$ em (2.3), é dada por

$$\widehat{\mathrm{OR}}(\mathrm{x}_1+\mathrm{c}, \mathrm{x}_1) = \exp(\mathrm{c}\widehat{\beta}_1). \qquad (2.4)$$

Por (2.4), temos que $\log(\widehat{\mathrm{OR}}(\mathrm{x}_1+\mathrm{c}, \mathrm{x}_1)) = \mathrm{c}\widehat{\beta}_1$. Assim, um intervalo de $100(1-\alpha)\%$ de confiança aproximado para as estimativas da razão de chances é obtido ao calcular inicialmente um intervalo de confiança para β_1 e então transformar seus limites, ou seja,

$$\mathrm{IC}(100(1-\alpha)\%, \mathrm{OR}(\mathrm{x}_1+\mathrm{c}, \mathrm{x}_1)) = \exp\left[\mathrm{c}\widehat{\beta}_1 \mp \mathrm{z}_{1-\alpha/2}\ \mathrm{c}\widehat{\mathrm{EP}}(\widehat{\beta}_1)\right], \qquad (2.5)$$

em que $\widehat{\mathrm{EP}}(\widehat{\beta}_1)$ é a estimativa do erro padrão de $\widehat{\beta}_1$, obtida da raiz quadrada do segundo termo da diagonal principal de $I(\widehat{\beta})^{-1}$, em que $I(\widehat{\beta})^{-1}$ é a inversa da matriz de informação de Fisher estimada.

É muito comum que exista no modelo pelo menos uma variável independente que seja categórica. Nesses casos, variáveis auxiliares são utilizadas. Vale citar que se a variável tem k categorias, o modelo terá $k-1$ variáveis auxiliares referentes a essa variável.

Hosmer & Lemeshow (2000, p. 51) apresentam um exemplo em que a variável resposta indica a presença ou ausência de doença coronariana. Para simplificar, vamos considerar apenas uma variável independente, idade, com duas categorias: idade maior ou igual a 55 anos, codificada como 1 e idade menor que 55 anos, codificada como 0. Nesse caso, considerando as probabilidades de sucesso $\pi(0)$ e $\pi(1)$ respectivas a um indivíduo com idade menor que 55 anos e a um indivíduo com idade maior que 55 anos, a razão de chances fica

$$\mathrm{OR}(1,0) = \frac{\pi(1)/(1-\pi(1))}{\pi(0)/(1-\pi(0))} = \frac{\exp(\beta_0 + 1\beta_1)}{\exp(\beta_0 + 0\beta_1)} = \exp(\beta_1).$$

2.4 Predição

Em alguns casos, o objetivo do ajuste de um modelo de regressão logística é a predição. Uma instituição financeira, por exemplo, deseja prever se um indivíduo será inadimplente. Para isso, é necessário que o modelo tenha ótimo poder de discriminação, pois o erro de classificação, ou seja, aprovar um crédito a um cliente inadimplente ou recusá-lo a um adimplente, traz prejuízos para a empresa. A seguir, descrevemos medidas de qualidade da discriminação do modelo muito utilizadas na área financeira para avaliar o poder preditivo do modelo de regressão logística, a saber, a estatística de Kolmogorov-Smirnov e as medidas de desempenho sensibilidade, especificidade, verdadeiro preditivo positivo, verdadeiro preditivo negativo e acurácia.

2.4.1 A estatística de Kolmogorov-Smirnov

A estatística de Kolmogorov-Smirnov (estatística KS) tem origem no teste de hipóteses não paramétrico de Kolmogorov-Smirnov para testar se duas amostras provêm de populações com a mesma distribuição. Em outras palavras, o intuito é verificar se há indícios de que as duas populações são distintas em relação a uma característica de interesse.

Neste contexto, a estatística KS é utilizada para verificar se as populações do evento

(E) e do não-evento (N) estão bem discriminadas em relação às probabilidades de sucesso, pois é esperado que a população do evento tenha maior probabilidade de sucesso que a população do não-evento. Sejam $F_{Y=1}(\hat{\pi})$ e $F_{Y=0}(\hat{\pi})$ as funções de distribuição empírica dos dados de E e N, respectivamente. A estatística KS é dada por

$$\text{KS} = \max|F_{Y=1}(\hat{\pi}) - F_{Y=0}(\hat{\pi})|, \tag{2.6}$$

e, desta maneira, a estatística KS corresponde à distância máxima vertical entre os gráficos de $F_{Y=1}(\hat{\pi})$ e $F_{Y=0}(\hat{\pi})$ sobre o conjunto dos possíveis valores de $\hat{\pi}$.

Para uma amostra de população E, sejam $\hat{\pi}_{(1)}, \ldots, \hat{\pi}_{(n)}$ as probabilidades de sucesso estimadas ordenadas de forma crescente, em que n é o tamanho amostral. A função de distribuição empírica pode ser escrita como

$$F_{Y=1}(\hat{\pi}) = \begin{cases} 0, & \text{se } \hat{\pi} < \hat{\pi}_{(1)} \\ k/n, & \text{se } \hat{\pi}_{(k)} \leq \hat{\pi} < \hat{\pi}_{(k+1)}, \text{ k=1,...,n-1} \\ 1, & \text{se } \hat{\pi} \geq \hat{\pi}_{(n)} \end{cases}$$

E de forma análoga é obtida a função de distribuição empírica $F_{Y=0}(\hat{\pi})$.

Neste trabalho, vamos considerar uma adaptação da estatística KS muito utilizada por instituições financeiras na avaliação de modelos preditivos. Em modelos cujo objetivo é predizer se um cliente é inadimplente ou adimplente, o escore do cliente é obtido ao multiplicar por 1000 a sua probabilidade de sucesso estimada. Os escores são agrupados em nove categorias: menores que 200 (categoria 1), de 200 a 300 (categoria 2), ..., maior que 900 (categoria 9). Contamos o número de inadimplentes e adimplentes observados em cada categoria, calculamos $F_{Y=1}(c)$ e $F_{Y=0}(c)$ e obtemos a estatística KS dada em (2.6). Nesse caso, $F_{Y=1}(c)$ e $F_{Y=0}(c)$ correspondem às funções de distribuição empírica dos clientes inadimplentes e adimplentes, respectivamente, da c-ésima categoria, com $c = 1, \ldots, 9$.

O valor da estatística, multiplicada por 100%, pode variar de 0% a 100%. Quanto maior seu valor, maior é a separação entre o evento e não-evento; assim, valores altos da estatística KS são desejáveis.

2.4.2 Medidas de desempenho

Um procedimento importante na avaliação do poder de predição do modelo considerado refere-se à obtenção de medidas de desempenho. Isto pode ser feito, em geral, por

meio da sensibilidade, especificidade, acurácia e pela verificação dos valores preditivos positivos e negativos. Seja $\widehat{Y} = 1$ se um indivíduo selecionado ao acaso da população em estudo for classificado como evento e $\widehat{Y} = 0$ se classificado como não-evento. São definidas abaixo algumas medidas de desempenho.

- **Sensibilidade**: probabilidade de classificação correta do evento, ou seja,

$$\text{Se} = \text{P}(\widehat{Y} = 1 | Y = 1) = \frac{\text{P}(\widehat{Y} = 1; Y = 1)}{\text{P}(Y = 1)}.$$

- **Especificidade**: probabilidade de classificação correta do não-evento, ou seja,

$$\text{Es} = \text{P}(\widehat{Y} = 0 | Y = 0) = \frac{\text{P}(\widehat{Y} = 0; Y = 0)}{\text{P}(Y = 0)}.$$

- **Verdadeiro Preditivo Positivo**: probabilidade do indivíduo ser evento, dado que foi classificado como evento, ou seja,

$$\text{VPP} = \text{P}(Y = 1 | \widehat{Y} = 1) = \frac{\text{P}(Y = 1; \widehat{Y} = 1)}{\text{P}(\widehat{Y} = 1)}.$$

- **Verdadeiro Preditivo Negativo**: probabilidade do indivíduo ser não-evento, dado que foi classificado como não-evento, ou seja,

$$\text{VPN} = \text{P}(Y = 0 | \widehat{Y} = 0) = \frac{\text{P}(Y = 0; \widehat{Y} = 0)}{\text{P}(\widehat{Y} = 0)}.$$

- **Acurácia**: probabilidade de classificação correta,

$$\text{ACC} = \text{P}(Y = 1; \widehat{Y} = 1) + \text{P}(Y = 0; \widehat{Y} = 0).$$

Na Seção 2.4.4 mostramos como estimar essas medidas de desempenho.

2.4.3 Curva ROC

Discutimos uma maneira de obter a classificação do indivíduo como evento ou não-evento. Para isso, é necessário haver uma regra de decisão que transforme a probabilidade estimada pelo modelo, $\widehat{\pi}_i$, na predição: $\widehat{Y}_i = 0$ ou $\widehat{Y}_i = 1$. Denotando 1 como evento, é intuitivo pensar que se $\widehat{\pi}_i$ for grande, $\widehat{Y}_i = 1$ e se $\widehat{\pi}_i$ for pequeno, $\widehat{Y}_i = 0$. Mas quão grande deve ser $\widehat{\pi}_i$ para o i-ésimo indivíduo ser classificado como evento, ou seja, como podemos determinar um ponto de corte?

Uma forma bastante utilizada para determinar um ponto de corte é através da curva ROC (*Receiver Operating Characteristic Curve*) que consiste no gráfico dos pares de sensibilidade (eixo vertical) e o complementar da especificidade (eixo horizontal) para cada ponto de corte possível. As instituições financeiras, por exemplo, utilizam critérios financeiros na determinação do melhor ponto, como o quanto se perde em média ao aprovar um cliente que traz problemas de crédito e o quanto se deixa de ganhar ao não aprovar um cliente que não traria problemas para a instituição (Diniz & Louzada, 2012, p. 21). Partindo da suposição de que falsos negativos e falsos positivos trazem prejuízos equivalentes, escolhemos o ponto de corte referente à combinação ótima de sensibilidade e especificidade. Este ponto é o que mais se aproxima do canto superior esquerdo da curva ROC, já que é o ponto que conjuntamente maximiza a sensibilidade e minimiza o complementar da especificidade. Desta forma, considerando 99 pontos de corte possíveis, variando em 0,01, 0,02,..., 0,99, D_j é a distância entre o ponto (1,0) e o j-ésimo ponto de corte é dada por

$$D_{\mathrm{j}} = \sqrt{(1 - \mathrm{Se_j})^2 + (1 - \mathrm{Es_j})^2}, \qquad j = 0,01, 0,02, \ldots, 0,99.$$

Se $D_{\mathrm{pc}} = \min(D_{0,01}, D_{0,02}, \ldots, D_{0,99})$, então pc é o ponto de corte ótimo. Na Figura 2.1 apresentamos um exemplo de curva ROC destacando o ponto de corte ótimo.

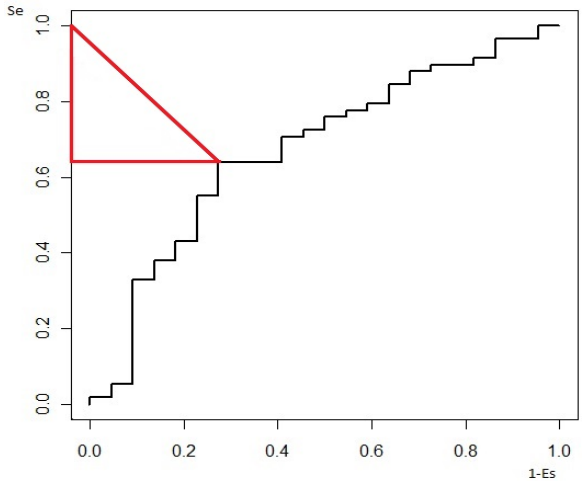

Figura 2.1: Curva ROC.

2.4.4 Estimativas das medidas de desempenho

Com o ponto de corte estabelecido pela curva ROC, cada indivíduo é classificado de acordo com a sua probabilidade de sucesso estimada quando comparada com o ponto de

corte. Se a probabilidade estimada for maior que o ponto de corte, o indivíduo é classificado como evento, caso contrário, como não-evento. Na Tabela 2.1 está a tabela de classificação do modelo. Abaixo, segue a definição das caselas desta tabela:

- verdadeiro positivo (VP): número de eventos classificados corretamente como eventos.
- verdadeiro negativo (VN): número de não-eventos classificados corretamente como não-eventos.
- falso positivo (FP): número de não-eventos classificados incorretamente como eventos.
- falso negativo (FN): número de eventos classificados incorretamente como não-eventos.

Tabela 2.1: Tabela de classificação do modelo.

	Observado	
Predito	1	0
1	VP	FP
0	FN	VN

Através dos resultados possíveis na classificação dos indivíduos, é possível estimar as probabilidades apresentadas na Seção 2.4.2. As estimativas da sensibilidade, especificidade, verdadeiros preditivos positivo e negativo e acurácia são, respectivamente,

$$\widehat{\text{Se}} = \frac{\text{VP}}{\text{VP} + \text{FN}}, \quad \widehat{\text{Es}} = \frac{\text{VN}}{\text{VN} + \text{FP}},$$

$$\widehat{\text{VPP}} = \frac{\text{VP}}{\text{VP} + \text{FP}}, \quad \widehat{\text{VPN}} = \frac{\text{VN}}{\text{VN} + \text{FN}} \quad \text{e} \quad \widehat{\text{ACC}} = \frac{\text{VP} + \text{VN}}{n}.$$

Neste trabalho as estimativas das medidas de desempenho serão apresentadas em porcentagem.

Capítulo 3

Modelos com erros de medida

Nesse capítulo discorremos sobre os principais conceitos nos modelos com erros de medida. São eles: os modelos para o erro de medida, características dos modelos funcional e estrutural, modelo geral para o erro de medida sob o enfoque estrutural e erros diferencial e não diferencial.

3.1 Modelos para o erro de medida

No problema de erro de medida é fundamental especificar o modelo para o erro de medida, ou seja, a relação da variável observada com a variável não observada. Carroll *et al.* (2006, Cap. 1) citam dois tipos de modelo, o modelo aditivo clássico e o modelo de Berkson. O modelo aditivo clássico considera que a variável observada é a soma da variável não observada e o erro de medida, ou seja,

$$W_i = X_i + U_i,$$

em que U_i é o erro de medida para o i-ésimo indivíduo. Já o modelo de Berkson considera a variável não observada como a soma da variável observada e o erro de medida,

$$X_i = W_i + U_i.$$

Se a escolha do modelo do erro for entre os dois apresentados, é necessário entender a diferença entre eles. Basicamente, o modelo clássico é preferível se a variável com erro for única para cada indivíduo, por exemplo, medição da pressão sanguínea. Mas se aos indivíduos

de um grupo ou estrato são dados os mesmos valores da variável, o modelo apropriado é o de Berkson. Por exemplo, trabalhadores de uma mina que recebem a mesma exposição à poeira, mas a verdadeira exposição é particular a cada indivíduo. Ainda, é fácil notar que variância de W_i é maior que a de X_i no modelo clássico e no modelo de Berkson acontece o contrário.

Em outras situações, a relação entre as variáveis observada e não observada pode não ser aditiva. Em Carroll *et al.* (2006, p. 13) está um exemplo de dados de sobrevida às explosões de Hiroshima e Nagasaki. Uma variável resposta é a aberração cromossômica e a verdadeira dose de radiação, X_i, não pode ser medida e no seu lugar, estimativas W_i são disponíveis. Ainda, é assumido que $W_i = 0$ se, somente se, $X_i = 0$. Além disso, assume-se que se X_i é positivo, este tem distribuição Weibull. Em símbolos, um modelo multiplicativo foi proposto, em que $W_i = X_i U_i$, $\log(U_i) \sim \mathrm{N}(\mu_\mathrm{u}, \sigma_\mathrm{u}^2)$.

Como relata Stefanski (2000), uma suposição incorreta sobre o modelo para o erro de medida pode causar problemas tanto quanto ignorá-lo. Desta maneira, a identificação correta da relação entre a variável não observada e a observada com erro é essencial para o sucesso do uso de modelos com erro de medida.

3.2 Modelos funcional e estrutural

Um dos conceitos relevantes acerca dos modelos com erro de medida é a distinção entre modelo funcional e estrutural. No modelo funcional, a quantidade não observada, X_i, $i = 1, \ldots, n$, é considerada uma sequência de constantes fixas desconhecidas ou parâmetros. Desta maneira, o número de parâmetros cresce com o tamanho da amostra e dizemos que há parâmetros incidentais no modelo. Já no modelo estrutural, a tal quantidade não observada é vista como uma variável aleatória, seguindo algum modelo paramétrico.

Carroll *et al.* (2006, p. 25) acreditam ser mais adequado considerar a distinção entre modelagem funcional e modelagem estrutural. De uma forma geral, na modelagem funcional a quantidade não observada pode ser fixa ou aleatória, mas neste caso, nenhuma suposição ou suposições mínimas são feitas sobre a sua distribuição. Já na modelagem estrutural, uma distribuição é assumida para a quantidade não observada e consequentemente, as estimativas e inferências dependem do modelo paramétrico escolhido. Neste contexto, Stefanski (2000)

diz que esta definição facilita a classificação dos métodos de estimação em métodos funcionais ou métodos estruturais baseada nas suposições feitas sobre a variável não observada.

3.3 Modelo geral para o erro de medida

Sob o enfoque estrutural, consideremos um modelo geral em que erro de medida está presente. Seja $f_{Y|X}(y_i|x_i;\beta)$ a função densidade de Y_i, variável resposta, relacionada com X_i, covariável(is), do i-ésimo indivíduo. No lugar de X_i, W_i é observado. Suponha que n realizações de Y_i e W_i estão disponíveis, em que $i = 1, \ldots, n$. Dada a função densidade conjunta de (Y_i, W_i, X_i), $f_{YWX}(y_i, w_i, x_i|\theta)$, que depende de um vetor de parâmetros, θ digamos, a função de verossimilhança para θ é dada por

$$L(\theta; y, w) = \prod_{i=1}^{n} \int f_{YWX}(y_i, w_i, x_i|\theta) dx_i.$$

A função de verossimilhança pode ser escrita em termos das distribuições condicionais, com $\theta = (\beta^\top, \delta^\top, \gamma^\top)^\top$, como

$$L(\theta; y, w) = \prod_{i=1}^{n} \int f_{Y|WX}(y_i|w_i, x_i; \beta) f_{W|X}(w_i|x_i; \delta) f_X(x_i|\gamma) dx_i. \tag{3.1}$$

Desta maneira, o logaritmo da função de verossimilhança, $l(\theta|y, w) = \log L(\theta|y, w)$ fica

$$l(\theta; y, w) = \sum_{i=1}^{n} \log \int [f_{Y|WX}(y_i|w_i, x_i; \beta) f_{W|X}(w_i|x_i; \delta) f_X(x_i|\gamma)] dx_i.$$

A função de verossimilhança permite a inclusão de covariáveis medidas sem erro, Z_i digamos, ao escrever as funções densidades como $f_{Y|WZX}(y_i|w_i, z_i, x_i; \beta)$, $f_{W|ZX}(w_i|z_i, x_i; \delta)$ e $f_{X|Z}(x_i|z_i; \gamma)$. Note que β é o parâmetro de interesse e $\lambda = (\delta^\top, \gamma^\top)^\top$ é o parâmetro de perturbação.

3.4 Erros de medida diferencial e não diferencial

Erro de medida não diferencial ocorre quando W_i não contém informação sobre Y_i além daquela disponível em X_i. Isto é, o erro é não diferencial se a distribuição de Y_i dado (X_i, W_i) depender apenas de X_i. Caso contrário, o erro de medida é diferencial.

Carroll *et al.* (2006, p. 36) apresentam um exemplo de dados com erro de medida não diferencial. A covariável de maior interesse é a pressão sanguínea de longo prazo, X_i, mas a pressão sanguínea só pode ser medida em um único dia, variável observada W_i. É plausível pensar que a pressão sanguínea medida em um dia não traz mais informação além daquela a longo prazo, ou seja, a informação de W_i não acrescenta informação quando X_i está disponível. Por outro lado, em estudos de caso-controle, o erro normalmente é diferencial. Por exemplo, em um estudo cuja variável resposta é o indicador da presença de câncer de mama, um preditor de interesse é a dieta da mulher antes do diagnóstico, X_i. Mas pelas características do estudo, o relato da dieta é feito depois do diagnóstico, W_i. Uma mulher que desenvolve o câncer deve mudar a sua dieta durante o tratamento e como consequência, a dieta reportada depois do diagnóstico é claramente correlacionada com a variável resposta, mesmo levando em conta a informação da dieta antes do diagnóstico.

Capítulo 4

Modelo de regressão logística com erros de medida

Neste capítulo introduzimos o modelo de regressão logística com erro de medida, combinando as informações contidas nos capítulos anteriores: regressão logística e modelos com erros de medida. Apresentamos e discutimos os seguintes métodos de estimação: máxima verossimilhança, máxima pseudoverossimilhança, SIMEX, calibração da regressão e *naïve*. Pela definição considerada na Seção 3.2, os métodos da calibração da regressão e SIMEX são métodos funcionais e os métodos de máxima verossimilhança e máxima pseudoverossimilhança são métodos estruturais. Erros padrão dos estimadores são apresentados no Apêndice A.

4.1 Modelo de regressão logística com erros de medida

No que segue, consideremos um modelo de regressão logística com erro de medida. Dado o modelo geral descrito na Seção 3.3, assumimos que $Y_i|(X_i = x_i, W_i = w_i) \sim \text{Ber}(\pi_i)$, em que π_i é a probabilidade de sucesso definida em (2.1). Neste trabalho, consideramos o erro de medida não diferencial, ou seja, admitimos que $f_{Y|W,X}(y_i|w_i, x_i) = f_{Y|X}(y_i|x_i)$. Sob o enfoque estrutural, assumimos que a distribuição da variável não observada é normal, ou seja, $X_i \sim N(\mu_x, \sigma_x^2)$ com μ_x e σ_x^2 representando a média e a variância, respectivamente. Ainda, a relação da variável observada com erro e da variável de interesse é representada pelo modelo aditivo clássico (Seção 3.1), ou seja, $W_i = X_i + U_i$ em que $U_i \sim N(0, \sigma_u^2)$ e σ_u^2 é

a variância do erro de medida.

Em alguns casos, consideramos σ_u^2 conhecida e em casos em que é desconhecida, assumimos que há réplicas de W_i para a estimação do parâmetro. Apesar de o modelo de regressão logística com erro de medida ser identificável sem informações adicionais (réplicas), o modelo de regressão probito não é (Carroll *et al.*, 2006, p. 184). A diferença entre esses dois modelos é tão pequena que não há informação útil sobre os parâmetros sem informações adicionais. Desta maneira, o modelo de regressão logística na presença de erro de medida é quase identificável e a estimação dos parâmetros sem informações adicionais é instável. Na presença de réplicas, o modelo com erro de medida tem a seguinte estrutura:

$$W_{ij} = X_i + U_{ij}, i = 1, 2, \ldots, n \ e \ j = 1, \ldots, r_i,$$

em que W_{ij} é a j-ésima réplica do i-ésimo indivíduo para a variável observada com erro e dado X_i, $W_{i1}, \ldots, W_{ir_i}$ são independentes. Seja $\overline{W}_i$ a média das réplicas do i-ésimo indivíduo e o desvio das observações em torno desta representa a variabilidade intrapessoal para cada indivíduo. Desta forma, a variância do erro de medida pode ser estimada por

$$\widehat{\sigma}_u^2 = \frac{1}{\sum_{i=1}^{r_i}(r_i - 1)} \sum_{i=1}^{n} \sum_{j=1}^{r_i} (W_{ij} - \overline{W}_i)^2, \tag{4.1}$$

em que

$$\overline{W}_i = \frac{1}{r_i} \sum_{j=1}^{r_i} W_{ij}.$$

Descrevemos agora os métodos de estimação que levam em conta o erro de medida e o método *naïve* (ingênuo).

4.1.1 Método naïve

O método *naïve* consiste em substituir a variável de interesse X_i pela variável observada W_i na estimação usual do modelo de regressão logística, como apresentado na Seção 2.2, ou seja, significa ajustar o modelo de regressão logística para os dados (W_i,Y_i), $i = 1, \ldots, n$.

4.1.2 Máxima verossimilhança e pseudoverossimilhança

O estimador de máxima verossimilhança do modelo com erro de medida é obtido por maximizar (3.1) em relação a θ. Embora este método de estimação apresente estimado-

res com boas propriedades no problema de erro de medida (Schafer, 2002), ele é de difícil aplicação dado o imenso esforço computacional em razão da complexidade da função de verossimilhança. Guolo (2010) sugere a maximização de uma função de pseudoverossimilhança, que simplifica a verossimilhança como função apenas dos parâmetros de interesse, utilizando o algoritmo EM-Monte Carlo.

Algoritmo EM-Monte Carlo

O algoritmo EM (*Expectation Maximization*) tem dois passos: o passo E avalia a esperança do logaritmo da função de verossimilhança completa e o passo M maximiza a esperança encontrada. Se essa esperança não tiver forma fechada, uma alternativa é aproximá-la por simulações Monte Carlo. A seguir apresentamos o algoritmo EM-Monte Carlo utilizado na maximização da função de verossimilhança e posteriormente a maximização da função de pseudoverossimilhança.

Máxima verossimilhança

Para o algoritmo EM, definimos $L(\theta; y, w, x)$ como a função de verossimilhança completa, dada por

$$L(\theta; y, w, x) = \prod_{i=1}^{n} f_{YWX}(y_i, w_i, x_i|\theta),$$

ou em função das distribuições condicionais,

$$L(\theta; y, w, x) = \prod_{i=1}^{n} f_{Y|WX}(y_i|w_i, x_i; \beta) f_{W|X}(w_i|x_i; \delta) f_X(x_i|\gamma),$$

em que θ é decomposto em β, γ e δ. O primeiro componente, $f_{Y|WX}(y_i|w_i, x_i; \beta)$, é a função de densidade condicional de $Y_i|(W_i, X_i)$, o segundo componente, $f_{W|X}(w_i|x_i; \delta)$, é a função de densidade condicional de $W_i|X_i$ e o último, $f_X(x_i|\gamma)$, é a função densidade da covariável X_i. Desta forma, o logaritmo da função de verossimilhança completa, $l(\theta; y, w, x) = \log L(\theta; y, w, x)$, é dado por

$$l(\theta; y, w, x) = \sum_{i=1}^{n} l(\theta; y_i, w_i, x_i) = \sum_{i=1}^{n} \log[f_{Y|WX}(y_i|w_i, x_i; \beta) f_{W|X}(w_i|x_i; \delta) f_X(x_i|\gamma)].$$

A cada iteração do algoritmo EM alternam-se um passo E e um passo M. Seja θ_r o valor de θ da r-ésima iteração. O $(r+1)$-ésimo passo E consiste em calcular

$$Q(\theta|\theta_r) = E[l(\theta; Y, W, X)|Y = y, W = w; \theta_r]. \tag{4.2}$$

Ainda, podemos decompor $Q(\theta|\theta_r)$ como a soma das esperanças dos logaritmos das densidades condicionais, ou seja,

$$\begin{aligned} Q(\theta|\theta_r) = E[l(\theta; Y|W, X)|Y = y, W = w; \theta_r] &+ E[l(\theta; W|X)|Y = y, W = w; \theta_r] \\ &+ E[l(\theta; X)|Y = y, W = w; \theta_r]. \end{aligned}$$

O passo M realiza a maximização de (4.2) com respeito a θ, resultando em uma nova estimativa θ_{r+1}. Dado um ponto inicial θ_0, a iteração entre o passo E e o passo M é repetida até a convergência.

A esperança em (4.2) não possui forma analítica e pode ser estimada por aproximações de Monte Carlo. Desta forma, seja $f_{X|YW}(x_i|y_i, w_i; \theta_r)$ a densidade de X_i dado o vetor (Y_i, W_i) e M amostras aleatórias $x^*_{r,1,i}, \ldots, x^*_{r,M,i}$ são simuladas dessa densidade. Então, a aproximação de Monte Carlo para Q é:

$$Q_m(\theta|\theta_r) = \frac{1}{M} \sum_{m=1}^{M} \sum_{i=1}^{n} [l(\theta; y_i|w_i, x^*_{r,m,i}) + l(\theta; w_i|x^*_{r,m,i}) + l(\theta; x^*_{r,m,i})]. \tag{4.3}$$

A distribuição de $X_i|Y_i, W_i$ normalmente não é conhecida em problemas de erro de medida. A solução é a utilização de amostragem de importância (*importance sampling*), utilizando a densidade $f_X(x_i|\theta_r)$ ou a densidade $f_{X|W}(x_i|w_i; \theta_r)$, que assumimos ter o mesmo suporte de $f_{X|YW}(x_i|y_i, w_i; \theta_r)$. A versão da amostragem de importância de (4.3) é

$$Q_m(\theta|\theta_r) = \frac{1}{M} \sum_{m=1}^{M} \sum_{i=1}^{n} k_{r,m,i}[l(\theta; y_i|w_i, x^*_{r,m,i}) + l(\theta; w_i|x^*_{r,m,i}) + l(\theta; x^*_{r,m,i})],$$

em que $k_{r,m,i}$ são os pesos da amostragem de importância. Se a densidade conhecida for $f_X(x_i|\theta_r)$, então $k_{r,m,i} = f_{X|YW}(x^*_{r,m,i}|y_i, w_i; \theta_r)/f_X(x^*_{r,m,i}|\theta_r)$ e $x^*_{r,1,i}, \ldots, x^*_{r,M,i}$ são obtidas por $f_X(x^*_{r,m,i}|\theta_r)$. No caso da densidade $f_{X|W}(x_i|w_i; \theta_r)$ ser a conhecida, temos $k_{r,m,i} = f_{X|YW}(x^*_{r,m,i}|y_i, w_i; \theta_r)/\ f_{X|W}(x^*_{r,m,i}|w_i; \theta_r)$. A expressão de $k_{r,m,i}$ pode ser ainda simplificada.

No primeiro caso,

$$\begin{aligned}
k_{r,m,i} &= \frac{f_{X|YW}(x^*_{r,m,i}|y_i,w_i;\theta_r)}{f_X(x^*_{r,m,i}|\theta_r)} = \frac{f_{YWX}(y_i,w_i,x^*_{r,m,i}|\theta_r)}{f_{YW}(y_i,w_i|\theta_r)f_X(x^*_{r,m,i}|\theta_r)} \\
&= \frac{f_{Y|WX}(y_i|w_i,x^*_{r,m,i};\theta_r)f_{W|X}(w_i|x^*_{r,m,i};\theta_r)f_X(x^*_{r,m,i}|\theta_r)}{f_{YW}(y_i,w_i|\theta_r)f_X(x^*_{r,m,i}|\theta_r)} \\
&= \frac{f_{Y|WX}(y_i|w_i,x^*_{r,m,i};\theta_r)f_{W|X}(w_i|x^*_{r,m,i};\theta_r)}{f_{YW}(y_i,w_i|\theta_r)} \\
&= \frac{f_{Y|WX}(y_i|w_i,x^*_{r,m,i};\theta_r)f_{W|X}(w_i|x^*_{r,m,i};\theta_r)}{\int f_{YWX}(y_i,w_i,x^*_{r,m,i}|\theta_r)dx^*_{r,m,i}} \\
&= \frac{f_{Y|WX}(y_i|w_i,x^*_{r,m,i};\theta_r)f_{W|X}(w_i|x^*_{r,m,i};\theta_r)}{\int f_{Y|WX}(y_i|w_i,x^*_{r,m,i};\theta_r)f_{W|X}(w_i|x^*_{r,m,i};\theta_r)f_X(x^*_{r,m,i}|\theta_r)dx^*_{r,m,i}}.
\end{aligned}$$

O peso $k_{r,m,i}$ pode ser aproximado por

$$k_{r,m,i} \approx \frac{f_{Y|WX}(y_i|w_i,x^*_{r,m,i};\theta_r)f_{W|X}(w_i|x^*_{r,m,i};\theta_r)}{\frac{1}{M}\sum_{m=1}^{M} f_{Y|WX}(y_i|w_i,x^*_{r,m,i};\theta_r)f_{W|X}(w_i|x^*_{r,m,i};\theta_r)}. \tag{4.4}$$

Analogamente, no caso da densidade $f_{X|W}(x_i|w_i;\theta_r)$ ser a conhecida,

$$k_{r,m,i} \approx \frac{f_{Y|WX}(y_i|w_i,x^*_{r,m,i};\theta_r)}{\frac{1}{M}\sum_{m=1}^{M} f_{Y|WX}(y_i|w_i,x^*_{r,m,i};\theta_r)}.$$

Máxima pseudoverossimilhança

Como descrevemos na Seção 3.3, decompomos θ em $(\beta^\top,\lambda^\top)^\top$, em que β é o parâmetro de interesse e $\lambda = (\delta^\top,\gamma^\top)^\top$ é o parâmetro de perturbação. Quando não é possível eliminar λ através de condicionamento e fatoração, uma possível alternativa é usar uma pseudoverossimilhança, como sugere Guolo (2010). Assim, a maximização de (3.1) é feita em dois passos: no primeiro, o parâmetro de perturbação é convenientemente estimado. Chamamos a estimativa de λ de $\widehat{\lambda} = (\widehat{\delta}^\top,\widehat{\gamma}^\top)^\top$. Em seguida, β é estimado ao maximizar a função de pseudoverossimilhança obtida fixando $\lambda = \widehat{\lambda}$. Assim, a estimativa de β é o valor de β que maximiza a função de pseudoverossimilhança dada por

$$\begin{aligned}
pL(\beta;y,w,\widehat{\lambda}) &= \prod_{i=1}^{n} pL(\beta;y_i,w_i,\widehat{\lambda}) \\
&= \prod_{i=1}^{n} \int f_{Y|WX}(y_i|w_i,x_i;\beta)f_{W|X}(w_i|x_i;\widehat{\delta})f_X(x_i|\widehat{\gamma})dx_i.
\end{aligned} \tag{4.5}$$

O mesmo processo do algoritmo EM-Monte Carlo apresentado para a maximização da função de verossimilhança é utilizado para a maximização da função de pseudoverossimi-

lhança. A função de pseudoverossimilhança completa é dada por

$$\begin{aligned} pL(\beta; y, w, x, \widehat{\lambda}) &= \prod_{i=1}^{n} pL(\beta; y_i, w_i, x_i, \widehat{\lambda}) \\ &= \prod_{i=1}^{n} f_{Y|WX}(y_i|w_i, x_i; \beta) f_{W|X}(w_i|x_i; \widehat{\delta}) f_X(x_i|\widehat{\gamma}). \end{aligned}$$

Suponha que uma estimativa de λ, $\widehat{\lambda}$, é obtida. Seja β_r o valor de β da r-ésima iteração e $pl(\beta; y, w, x, \widehat{\lambda}) = \log pL(\beta; y, w, x, \widehat{\lambda})$. O $(r+1)$-ésimo passo E é

$$Q(\beta|\beta_r; \widehat{\lambda}) = E[pl(\beta; Y, W, X)|Y = y, W = w; \beta_r, \widehat{\lambda}].$$

Ainda, podemos decompor $Q(\beta|\beta_r; \widehat{\lambda})$ como a soma das esperanças dos logaritmos das densidades condicionais, ou seja,

$$\begin{aligned} Q(\beta|\beta_r; \widehat{\lambda}) &= E[pl(\beta; Y|W, X)|Y = y, W = w; \beta_r, \widehat{\lambda}] + E[pl(\lambda; W|X)|Y = y, W = w; \beta_r, \widehat{\lambda}] \\ &+ E[pl(\lambda; X)|Y = y, W = w; \beta_r, \widehat{\lambda}]. \end{aligned} \tag{4.6}$$

A esperança em (4.6) pode ser simplificada ao remover os dois últimos termos, já que estes não dependem do parâmetro de interesse β. Assim, a aproximação de Monte Carlo para (4.6) é dada por

$$Q_m(\beta|\beta_r; \widehat{\lambda}) = \frac{1}{M} \sum_{m=1}^{M} \sum_{i=1}^{n} pl(\beta; y_i|w_i, x^*_{r,m,i}, \widehat{\lambda}),$$

em que $x^*_{r,1,i}, \ldots, x^*_{r,M,i}$ são M amostras aleatórias obtidas de $f_{X|YW}(x_i|y_i, w_i; \beta_r, \widehat{\lambda})$. Como descrito no algoritmo EM-Monte Carlo para a maximização da função de verossimilhança, a técnica de amostragem de importância pode ser útil quando não é possível a simulação de $f_{X|YW}(x_i|y_i, w_i; \beta_r, \widehat{\lambda})$.

A maximização da pseudoverossimilhança (4.5) assume que uma estimativa de λ seja disponível. A estimativa do parâmetro de perturbação λ pode ser obtida pela maximização da função de verossimilhança reduzida dada por

$$rL(\lambda; w) = \prod_{i=1}^{n} \int f_{W|X}(w_i|x_i, \delta) f_X(x_i|\gamma) dx_i. \tag{4.7}$$

Quando a integral em (4.7) não pode ser expressa na forma fechada, o algoritmo EM-Monte Carlo pode ser utilizado de forma análoga ao que apresentamos da maximização das funções de verossimilhança e de pseudoverossimilhança.

4.1.3 SIMEX

O método SIMEX consiste basicamente em duas etapas: simulação (SIM) e extrapolação (EX). A ideia do método é o fato de que o efeito do erro de medida em um estimador pode ser determinado por simulação, ao introduzir um erro de medida adicional aos dados utilizando um esquema de reamostragem, de tal forma que se possa estabelecer uma relação entre o viés induzido pelo erro de medida e a variância do erro de medida adicionado e extrapolar essa relação para o caso correspondente à ausência de erro de medida (Cook & Stefanski, 1994). Consideramos duas versões para o método SIMEX: a que considera a variância do erro de medida conhecida (ou convenientemente estimada) e a versão que leva em conta o desconhecimento da variância do erro de medida.

SIMEX com variância do erro de medida conhecida

Em situações em que o modelo para o erro de medida é aditivo clássico e a variância do erro de medida é conhecida, ou previamente e devidamente estimada, sejam as pseudoamostras definidas por

$$W_{d,i}(\zeta_b) = W_i + \zeta_b^{1/2} U_{d,i},\ d = 1, \ldots, D,\ i = 1, \ldots, n,\ b = 1, \ldots, B, \tag{4.8}$$

com $\zeta_b = \zeta_1, \ldots, \zeta_B$, em que $0 = \zeta_1 < \zeta_2 < \ldots < \zeta_B$ são valores definidos e conhecidos. Os pseudoerros, $U_{d,i}$, são mutuamente independentes, independentes dos dados observados e identicamente distribuídos com distribuição normal com média zero e variância σ_u^2.

Para a d-ésima pseudoamostra ($d = 1, \ldots, D$) de um valor fixo de ζ_b, obtemos as estimativas $\widehat{\beta}^{(d)}(\zeta_b)$ pelo método de máxima verossimilhança através do ajuste do modelo de regressão logística considerando $(W_{d,i}(\zeta_b), Y_i),\ i = 1, \ldots, n$. As estimativas podem ser resumidas utilizando a média de $\widehat{\beta}^{(1)}(\zeta_b), \ldots, \widehat{\beta}^{(D)}(\zeta_b)$, resultando em $\widehat{\beta}(\zeta_b)$. Esta etapa do método é a simulação.

Vale notar que $\mathrm{Var}(W_i|X_i) = \sigma_u^2$ e pela relação dada em (4.8), $\mathrm{Var}(W_{d,i}(\zeta_b)|X_i) = (1+\zeta_b)\mathrm{Var}(W_i|X_i) = (1+\zeta_b)\sigma_u^2$. O ideal é um cenário sem erro de medida, o que corresponde a $(1 + \zeta_b)\sigma_u^2 = 0$, e assim $\zeta_b = -1$.

Desta maneira, a extrapolação implica na modelagem das estimativas $\widehat{\beta}(\zeta_1), \ldots, \widehat{\beta}(\zeta_B)$ em função de $\zeta_1, \ldots, \zeta_B$ e usando uma função apropriada para extrapolar para $\zeta_b = -1$, que neste caso chamamos de estimador SIMEX. Contudo, essa função é raramente conhecida e

por isso pode ser estimada por funções simples, como a função linear, quadrática e não-linear. A função linear é dada por

$$G_L(\zeta_b, \Gamma) = \gamma_1 + \gamma_2 \zeta_b.$$

Outra função muito utilizada por sua simplicidade é a função de extrapolação quadrática, que é dada por

$$G_Q(\zeta_b, \Gamma) = \gamma_1 + \gamma_2 \zeta_b + \gamma_3 \zeta_b^2.$$

Carroll *et al.* (2006, p. 109) discutem a utilização de outra função de extrapolação, dada por

$$G_{RL}(\zeta_b, \Gamma) = \frac{\gamma_1\gamma_3 + \gamma_2 + \gamma_1\zeta_b}{\gamma_3 + \zeta_b},$$

em que $\Gamma = (\gamma_1, \gamma_2, \gamma_3)^\top$ e $\gamma_3 = 0$ no modelo linear. Como $G_{RL}(\zeta_b, \Gamma)$ é a razão de duas funções lineares, esse modelo é chamado de extrapolação linear racional. A função $G_{RL}(\zeta_b, \Gamma)$ requer a utilização de um programa de mínimos quadrados não-lineares para o ajuste e Carroll *et al.* (2006, p. 110) apresentam uma forma de obter bons valores iniciais de Γ utilizando as estimativas do modelo quadrático.

SIMEX com variância do erro de medida desconhecida

Devanarayan & Stefanski (2002) apresentam uma versão do método SIMEX para as situações em que a variância do erro de medida é desconhecida. Nestes casos, a variância do erro de medida pode ser heteroscedástica. Para essa versão, r_i, número de réplicas de W_i do i-ésimo indivíduo, deve ser dois ou mais para poder estimar σ_{ui}^2 (se consideramos variância heteroscedástica, mas se assumirmos um modelo homoscedástico, a variância do erro de medida é σ_u^2). O modelo assumido é $W_{ij} = X_i + U_{ij}$, em que U_{ij}, $j = 1, \ldots, r_i$, são normalmente distribuídos com média 0 e variância σ_{ui}^2.

Como a variância do erro de medida é desconhecida, não podemos gerar as pseudoamostras como em (4.8). No entanto, podemos gerar pseudoamostras na forma de contrastes lineares aleatórios das réplicas, definidas por

$$W_{d,i}(\zeta_b) = \overline{W}_i + \left(\frac{\zeta_b}{r_i}\right)^{1/2} \sum_{j=1}^{r_i} c_{d,i,j} W_{ij},\ d = 1, \ldots, D,\ i = 1, \ldots, n \text{ e } b = 1, \ldots, B,$$

em que

$$c_{d,i,j} = \frac{A_{d,i,j} - \overline{A}_{d,i,.}}{\sqrt{\sum_{j=1}^{r_i} (A_{d,i,j} - \overline{A}_{d,i,.})^2}},$$

com $\sum_{j=1}^{r_i} c_{d,i,j} = 0$ e $\sum_{j=1}^{r_i} c_{d,i,j}^2 = 1$ e $A_{d,i,j} \sim N(0,1)$ independentes entre si e também independentes dos dados observados.

É fácil notar que $\text{Var}(W_{d,i}(\zeta_b)|X_i) = (1+\zeta_b)\sigma_u^2/r_i = (1+\zeta_b)\text{Var}(\overline{W}_i|X_i)$ e o cenário sem erro de medida é obtido com $\zeta_b = -1$.

O que difere esta versão para o método SIMEX da versão apresentada anteriormente é a forma como as pseudoamostras são geradas. Os passos seguintes são análogos à versão para o método SIMEX com variância conhecida.

Como esta versão do método SIMEX gera pseudoerros de dados observados, ela é chamada de SIMEX empírico.

4.1.4 Calibração da regressão

O método de calibração da regressão consiste em substituir a variável não observada X_i por alguma função de W_i (ou substituir por alguma função de (W_i, Z_i), se existir(em) covariável(is) medida(s) sem erro, Z_i), como a esperança condicional de X_i dado W_i. Após a substituição, estimam-se os parâmetros da maneira usual.

A seguir, apresentamos o método da calibração da regressão no caso de a variância do erro de medida ser desconhecida, mas estimada por réplicas de W_i, isto é, r_i réplicas, $W_{i1}, W_{i2}, \ldots, W_{ir_i}$, de W_i.

Como propõem Carroll *et al.* (2006, p. 71), a estimativa da esperança condicional de X_i dado $(\overline{W}_i, Z_i)$ é

$$\widehat{\text{E}}(X_i|\overline{W}_i = \overline{w}_i, \text{Z}_\text{i} = \text{z}_\text{i}) \approx \overline{\text{w}} + \begin{bmatrix} \widehat{\Sigma}_{xx}, & \widehat{\Sigma}_{xz} \end{bmatrix} \begin{bmatrix} \widehat{\Sigma}_{xx} + \widehat{\Sigma}_{uu}/r_i & \widehat{\Sigma}_{xz} \\ \widehat{\Sigma}_{xz}^\top & \widehat{\Sigma}_{zz} \end{bmatrix}^{-1} \begin{bmatrix} \overline{w}_i - \overline{w} \\ z_i - \overline{z} \end{bmatrix},$$

em que

$$\overline{w}_i = \frac{1}{r_i}\sum_{j=1}^{r_i} w_{ij}, \quad \overline{w} = \frac{1}{\sum_{i=1}^n r_i}\sum_{i=1}^n r_i\overline{w}_i, \quad \overline{z} = \frac{1}{n}\sum_{i=1}^n z_i,$$

$$\widehat{\Sigma}_{uu} = \frac{1}{\sum_{i=1}^{r_i}(r_i - 1)}\sum_{i=1}^n\sum_{j=1}^{r_i}(w_{ij} - \overline{w}_i)(w_{ij} - \overline{w}_i)^\top,$$

$$\widehat{\Sigma}_{zz} = \frac{1}{n-1}\sum_{i=1}^{n}(z_i - \overline{z})(z_i - \overline{z})^\top,$$

$$\widehat{\Sigma}_{xz} = \frac{1}{\sum_{i=1}^{n} r_i - (\sum_{i=1}^{n} r_i^2 / \sum_{i=1}^{n} r_i)} \sum_{i=1}^{n} r_i(\overline{w}_i - \overline{w})(z_i - \overline{z})^\top \quad \text{e}$$

$$\widehat{\Sigma}_{xx} = \frac{1}{\sum_{i=1}^{n} r_i - (\sum_{i=1}^{n} r_i^2 / \sum_{i=1}^{n} r_i)} \left\{ \left[\sum_{i=1}^{n} r_i(\overline{w}_i - \overline{w})(\overline{w}_i - \overline{w})^\top \right] - (n-1)\widehat{\Sigma}_{uu} \right\}.$$

A expressão $\widehat{\mathrm{E}}(\mathrm{X_i}|\overline{\mathrm{W}}_\mathrm{i} = \overline{\mathrm{w}}_\mathrm{i}, \mathrm{Z_i} = \mathrm{z_i})$ permite a existência de mais de uma variável observada com erro e sem erro de medida. No caso de apenas uma variável com erro de medida, X_i, e nenhuma variável sem erro de medida, a estimativa da esperança condicional de X_i dado $\overline{W_i} = \overline{w}_i$ é

$$\widehat{\mathrm{E}}(X_i|\,\overline{W}_i = \overline{w}_i) \approx \overline{\mathrm{w}} + \frac{\widehat{\sigma}^2_\mathrm{x}}{\widehat{\sigma}^2_\mathrm{x} + \widehat{\sigma}^2_\mathrm{u}/\mathrm{r_i}}(\overline{\mathrm{w}}_\mathrm{i} - \overline{\mathrm{w}}),$$

em que

$$\widehat{\sigma}^2_x = \frac{1}{\sum_{i=1}^{n} r_i - (\sum_{i=1}^{n} r_i^2 / \sum_{i=1}^{n} r_i)} \left\{ \left[\sum_{i=1}^{n} r_i(\overline{w}_i - \overline{w})^2 \right] - (n-1)\widehat{\sigma}^2_u \right\}, \tag{4.9}$$

e $\widehat{\sigma}^2_u$ é dado por (4.1). Se a variância do erro for conhecida, $\widehat{\sigma}^2_u/r_i$ é substituído por σ^2_u e a estimativa $\widehat{\sigma}^2_x$ fica $[\sum_{i=1}^{n}(\overline{w}_i - \overline{w})^2/(n-1)] - \sigma^2_u$.

Capítulo 5

Estudo de simulação

Realizamos um estudo de simulação com $R = 1000$ réplicas de amostras cujo objetivo é investigar o comportamento dos estimadores dos parâmetros do modelo de regressão logística quando uma covariável apresenta erro de medida, obtidos pelos métodos *naïve*, calibração da regressão (CR), máxima pseudoverossimilhança (MPV) e SIMEX, tanto na estimação dos parâmetros do modelo quanto no poder preditivo.

Na estimação, comparamos os métodos em relação às estimativas do viés e da raiz do erro quadrático médio (REQM) e correspondentes intervalos com 95% de confiança. Para $\beta = (\beta_0, \beta_1)$ temos que $\bar{\beta}_p = \sum_{j=1}^{R} \hat{\beta}_p^{(j)}/R$, $\widehat{vies}(\beta_p) = \bar{\beta}_p - \beta_p$ e $\widehat{EQM}(\beta_p) = \sum_{j=1}^{R} (\hat{\beta}_p^{(j)} - \beta_p)^2/R$, para $p = 0, 1$, são estimativas, respectivamente, da média, do viés e do erro quadrático médio do estimador de β_p, e $\hat{\beta}_p^{(j)}$ representa a estimativa de β_p na j-ésima amostra simulada. Um intervalo com 95% de confiança para o viés de β_p é dado por

$$IC(vies(\beta_p), 95\%) = \left[\widehat{vies}(\beta_p) - 1,96\frac{s_v}{\sqrt{R}};\quad \widehat{vies}(\beta_p) + 1,96\frac{s_v}{\sqrt{R}}\right],$$

em que $s_v = \sqrt{\sum_{j=1}^{R} (vi_j(\beta_p) - \widehat{vies}(\beta_p))^2/R}$ e $vi_j(\beta_p) = \hat{\beta}_p^{(j)} - \beta_p, \quad j = 1, \ldots, R$. Uma estimativa de $REQM(\beta_p)$ é $\widehat{REQM}(\beta_p) = \sqrt{\widehat{EQM}(\beta_p)}$ e um intervalo com 95% de confiança para $REQM(\beta_p)$ é dado por

$$IC(REQM(\beta_p), 95\%) = \left[\sqrt{\widehat{EQM}(\beta_p) - 1,96\frac{s}{\sqrt{R}}};\quad \sqrt{\widehat{EQM}(\beta_p) + 1,96\frac{s}{\sqrt{R}}}\right],$$

em que $s = \sqrt{\sum_{j=1}^{R} (EQ_j(\beta_p) - \widehat{EQM}(\beta_p))^2/R}$ e $EQ_j(\beta_p) = (\hat{\beta}_p^{(j)} - \beta_p)^2, \quad j = 1, \ldots, R$.

Para verificar o poder preditivo dos métodos em comparação, analisamos as medidas de desempenho sensibilidade, especificidade, VPP, VPN e acurácia, descritas na Seção 2.4.4,

e a estatística KS adaptada para categorias de escores, descrita na Seção 2.4.1. Assim, para cada simulação, apresentamos a mediana dos R valores de cada uma dessas medidas.

Consideramos amostras de tamanho $n = 45$, 150 e 430. Setenta por cento das observações foram utilizados para o ajuste do modelo (chamamos de amostra treinamento, da qual são obtidas as estimativas do viés e da REQM e a escolha do ponto de corte pela curva ROC, descrita na Seção 2.4.3) e as 30% restantes para a validação da predição (amostra teste, da qual obtemos as medidas de desempenho da predição e a estatística KS). Desta maneira, os tamanhos das amostras treinamento são $n_1 = 32$, 105 e 301 e os tamanhos das respectivas amostras testes são $n_2 = 13$, 45 e 129.

Seis cenários são considerados para as simulações. São eles:

- Cenário 1: Uma única covariável com erro, a variância do erro de medida é conhecida, o erro de medida é homoscedástico e a variável resposta tem a mesma proporção esperada de eventos e de não-eventos (amostra balanceada).
- Cenário 2: Uma única covariável com erro, a variância do erro de medida é desconhecida e estimada por réplicas da variável observada, o erro de medida é homoscedástico e a variável resposta tem a mesma proporção esperada de eventos e de não-eventos.
- Cenário 3: Uma única covariável com erro, a variância do erro de medida é desconhecida e estimada por três réplicas da variável observada, o erro de medida é homoscedástico e a proporção esperada de eventos na amostra é de 22%.
- Cenário 4: Uma única covariável com erro, a variância do erro de medida é desconhecida e estimada por três réplicas da variável observada, o erro de medida é homoscedástico e a proporção esperada de eventos na amostra é de apenas 10%.
- Cenário 5: Uma única covariável com erro, a variância do erro de medida é desconhecida e estimada por três réplicas da variável observada, o erro de medida apresenta baixa heteroscedasticidade e a variável resposta tem a mesma proporção esperada de eventos e de não-eventos.
- Cenário 6: Uma única covariável com erro, a variância do erro de medida é desconhecida e estimada por três réplicas da variável observada, o erro de medida apresenta alta heteroscedasticidade e a variável resposta tem a mesma proporção esperada de eventos e de não-eventos.

Para cada cenário e tamanho amostral, a variável não observada X_i é gerada com distribuição normal com média μ_x e variância σ_x^2. A variável resposta Y_i é gerada a partir de uma distribuição de Bernoulli com probabilidade de sucesso dada em (2.1). Além disso, geramos a variável observada com erro, W_i, a partir de sua distribuição condicional dado X_i, ou seja, com distribuição normal com média x_i e variância $\sigma_{ui}^2 = q(x_i)\sigma_u^2$, em que $q(x_i) = |x_i|^\rho$ e $i = 1, \ldots, n$. No cenário 5, consideramos $\rho = 0{,}5$ e no cenário 6, $\rho = 2$. Nos demais cenários, $\rho = 0$ e desta maneira, $\sigma_{ui}^2 = \sigma_u^2$.

Como definir uma alta ou baixa variância do erro de medida depende da variabilidade total dos dados, consideramos $h = \sigma_u^2/(\sigma_u^2 + \sigma_x^2)$ a variabilidade do erro relativamente à variância total dos dados. Assim, quanto maior o valor de h, mais a variabilidade do erro de medida representa da variabilidade total. Consideramos neste trabalho os valores de h 0,05, 0,25 e 0,5.

Para a maximização no passo M do algoritmo EM-Monte Carlo, necessário para obter as estimativas de MPV, utilizamos $M = 1000$ amostras de Monte Carlo e o método de otimização BFGS (Nocedal & Wright, 2006, Cap. 3). A regra de parada do algoritmo EM é dada por

$$\max_a \left(\frac{|\theta_{r+1,a} - \theta_{r,a}|}{|\theta_{r,a} + \epsilon_1|} \right) < \epsilon_2,$$

em que a é o indicativo do parâmetro do modelo na r-ésima iteração do algoritmo EM-Monte Carlo. Seguindo Booth & Hobert (1999) e Guolo (2010), tomamos $\epsilon_1 = 0{,}001$ e $\epsilon_2 = 0{,}005$. Aqui, o vetor dos parâmetros de interesse é $\beta = (\beta_0, \beta_1)^\top$ e o vetor de parâmetros de perturbação é $\lambda = (\mu_x, \sigma_x^2, \sigma_u^2)^\top$. Os estimadores de máxima verossimilhança para λ podem ser obtidos de forma fechada e as estimativas de $\lambda = (\mu_x, \sigma_x^2, \sigma_u^2)^\top$ são, respectivamente, $\widehat{\mu}_x = \overline{w}$, $\widehat{\sigma}_x^2$ dado por (4.9) e $\widehat{\sigma}_u^2$ dado por (4.1), com $r_i = 3$, nos cenários que há réplicas de W_i. Nas situações que σ_u^2 é conhecida, $\lambda = (\mu_x, \sigma_x^2)^\top$ e $r_i = 1$.

Para o método SIMEX, consideramos $D = 100$ pseudoamostras, $\zeta_b = \{0, 0{,}5, 1, 1{,}5, 2\}$ e o método de extrapolação linear. Nos cenários que consideramos a variância do erro de medida desconhecida, utilizamos duas versões do método SIMEX. Na primeira (SIMEX1), estimamos a variância do erro de medida por (4.1) e utilizamos essa estimativa no lugar de σ_u^2 na versão do SIMEX que considera a variância do erro de medida conhecida. A segunda versão (SIMEX2) é o SIMEX empírico, que descrevemos na Seção 4.1.3. No cenário 1 utilizamos o método SIMEX com variância do erro de medida conhecida.

O método de MPV é uma alternativa ao método de máxima verossimilhança (MV) por este ser de difícil aplicação dado o imenso esforço computacional em razão da complexidade da função de verossimilhança, como mencionado na Seção 4.1.2. Alguns trabalhos compararam os dois métodos. Carrasco (2012) e Carrasco *et al.* (2013), por exemplo, comparam os métodos de estimação em modelos de regressão beta. Já Guolo (2010) compara os métodos de MV e de MPV, maximizando as funções de verossimilhança e de pseudoverossimilhança pelo método EM-Monte Carlo, em que o modelo de regressão logística é um dos modelos em consideração. Os trabalhos observaram que os dois métodos apresentam resultados similares e por ter o método de MPV o menor custo computacional, os três trabalhos recomendam esse método. Para confirmar as conclusões obtidas nos trabalhos citados, nas Tabelas 5.1 e 5.2 estão o viés e a REQM dos estimadores de β_0 e β_1 obtidos pelos dois métodos em uma das situações dos cenários 1 e 2 (com os mesmos valores dos parâmetros dos respectivos cenários), respectivamente, com $n_1 = 105$ e $h = 0{,}25$. Observamos que os comportamentos dos métodos em relação ao viés e REQM são similares. Desta forma, nas simulações que seguem não consideramos o método de MV devido ao seu alto custo computacional e por haver indicações de que apresenta resultados parecidos com o método de MPV.

Tabela 5.1: Viés e REQM [intervalo de 95% de confiança] dos estimadores de β_0 e β_1; Cenário 1, $n_1 = 105$ e $h = 0{,}25$.

Métodos	Viés		REQM	
	β_0	β_1	β_0	β_1
MPV	1,451	−0,180	2,734	0,336
	[1,307; 1,594]	[−0,197; −0,162]	[2,607; 2,855]	[0,319; 0,351]
MV	1,573	−0,195	2,732	0,334
	[1,434; 1,712]	[−0,212; −0,178]	[2,606; 2,852]	[0,318; 0,349]

Tabela 5.2: Viés e REQM [intervalo de 95% de confiança] dos estimadores β_0 e β_1; Cenário 2, $n_1 = 105$ e $h = 0{,}25$.

Métodos	Viés		REQM	
	β_0	β_1	β_0	β_1
MPV	2,116	−0,263	2,969	0,366
	[1,985; 2,247]	[−0,279; −0,247]	[2,854; 3,080]	[0,352; 0,380]
MV	2,168	−0,269	2,954	0,364
	[2,042; 2,294]	[−0,285; −0,254]	[2,842; 3,063]	[0,350; 0,377]

Uma questão importante é o substituto de X_i na estimação da probabilidade de sucesso. Carroll *et al.* (2006, p. 38) afirmam ser a estimação *naïve* a apropriada para predição, argumentando que W_i é observada sem erro como uma medição de si mesmo e se tiver disponível um conjunto de dados (Y_i, W_i, Z_i), $i = 1, \ldots, n$, é possível ajustar um modelo conveniente para a variável resposta como uma função de (W_i, Z_i). Mas este argumento não é tão convincente se pensarmos que a probabilidade de sucesso é uma função dos parâmetros do modelo, que foram estimados levando em conta a presença do erro de medida.

Podemos pensar em um substituto de X_i para cada método de estimação em consideração. O método da CR tem um substituto natural, a esperança condicional estimada de X_i dado W_i, apresentada na Seção 4.1.4. Para o método SIMEX, não conseguimos chegar a nenhum substituto natural. Para o método de MPV uma possibilidade é substituir X_i por $\overline{x}_i^*$, em que $\overline{x}_i^* = \sum_{m=1}^{M} x_{m,i}^*/M$ e $x_{m,i}^*$ é gerado por uma distribuição normal com média $\widehat{\mu}_x$ e variância $\widehat{\sigma}_x^2$ do último passo do algoritmo EM-Monte Carlo (Seção 4.1.2). Mas essa substituição não levou a bons resultados, como podemos ver na Tabela 5.3. Esse resultado já havia sido observado por Rodrigues (2010) que usou uma substituição similar no método de máxima verossimilhança.

Como a esperança condicional estimada de X_i dado W_i é um preditor de X_i, utilizamos essa substituição em todos os métodos que corrigem o erro de medida.

Tabela 5.3: Medidas de desempenho e estatística KS obtidas por diferentes métodos de estimação; Cenário 2, $n_2 = 45$ e $h = 0{,}05$.

Medidas	MPV	CR	*naïve*
Se	50,0	89,4	89,4
Es	50,0	90,4	90,3
VPP	50,0	90,4	90,4
VPN	50,0	89,4	89,6
ACC	53,8	88,8	88,8
KS	14,2	83,3	83,1

A seguir, apresentamos o comportamento dos métodos de estimação nos seis cenários em consideração.

5.1 Cenário 1

Neste cenário consideramos que a variância do erro de medida, σ_u^2, é conhecida e comum para todas as observações, e seus valores são 0,15, 1 e 3. Os demais parâmetros, fixos para todas as simulações, são: $\mu_x = 8$, $\sigma_x^2 = 3$, $\beta_0 = -16$ e $\beta_1 = 2$.

5.1.1 Estimação

As Figuras 5.1-5.6 mostram os gráficos das estimativas pontuais e intervalares do viés e da REQM dos estimadores de β_0 e β_1 para os diferentes valores de h e n_1.

O estimador de MPV apresenta bom desempenho para todos os valores de h e é o melhor método na maioria das situações em relação ao viés e à REQM. Apesar de apresentar o menor viés quando, conjuntamente, o tamanho da amostra é pequeno ($n_1 = 32$) e $h = 0{,}5$, o estimador da CR apresenta o maior valor da REQM, mas para tamanhos amostrais maiores, o estimador tem bom desempenho para todos os valores de h considerados. Em geral, o método da CR tem bom desempenho e na maioria das vezes, só não é superior nas métricas de estimação que o estimador de MPV. O método *naïve* produz os piores estimadores tanto no viés quando em relação à REQM (com raras exceções não é o pior e, mesmo nelas, é

um dos piores). Observamos também que o estimador SIMEX tem os maiores valores de REQM e viés para estimar β_0 quando $h = 0{,}05$ e os tamanhos amostrais são $n_1 = 32$ e 105. Nos demais casos, apresenta comportamento intermediário e muitas vezes tem melhor desempenho apenas que o estimador *naïve*, com exceção quando $h = 0{,}25$, pois apresenta comportamento similar ao estimador de MPV, principalmente para β_0.

5.1.2 Predição

As Tabelas 5.4 e 5.5 mostram as medianas dos R valores das medidas sensibilidade (Se), especificidade (Es), VPP, VPN, ACC e estatística KS para os diferentes valores de h e n_2.

Em geral, não há diferença importante entre os métodos de estimação para as medidas sensibilidade, especificidade, VPP, VPN e acurácia. No entanto, em alguns casos, há pequenas diferenças nas medidas obtidas pelos métodos de estimação, como a situação $n_2 = 45$ e $h = 0{,}05$, onde o valor mediano da sensibilidade obtido pelo método SIMEX é 84,8% e dos demais métodos é 85,7%. No entanto, ao avaliar os gráficos *boxplot* das Figuras 5.7 e 5.8, observamos que as medidas de sensibilidade e especificidade apresentam comportamentos similares para todos os métodos de estimação. Isto acontece para as demais medidas de desempenho consideradas e para as medidas de desempenho nos próximos cenários.

Os valores da estatística KS obtidos pelo método SIMEX são os menores em todos os casos e a diferença do valor desta estatística em relação ao valor obtido pelos demais métodos diminui conforme h aumenta. É possível ver esse comportamento na Figura 5.9, onde estão os gráficos *boxplot* dos valores da estatística KS para $n_2 = 45$ e é similar para os demais tamanhos amostrais.

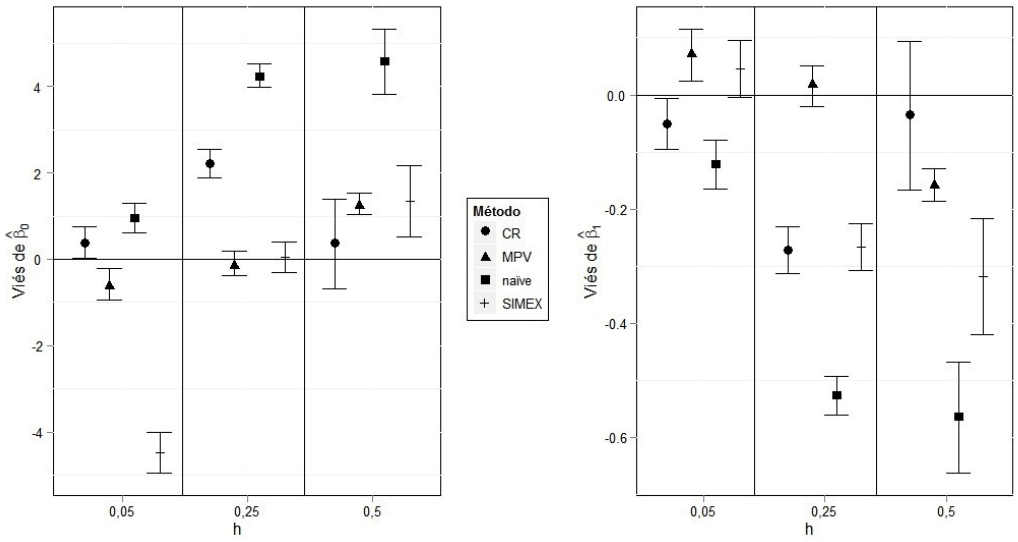

Figura 5.1: Intervalo de 95% de confiança para o viés dos estimadores de β_0 e β_1; cenário 1, $n_1 = 32$

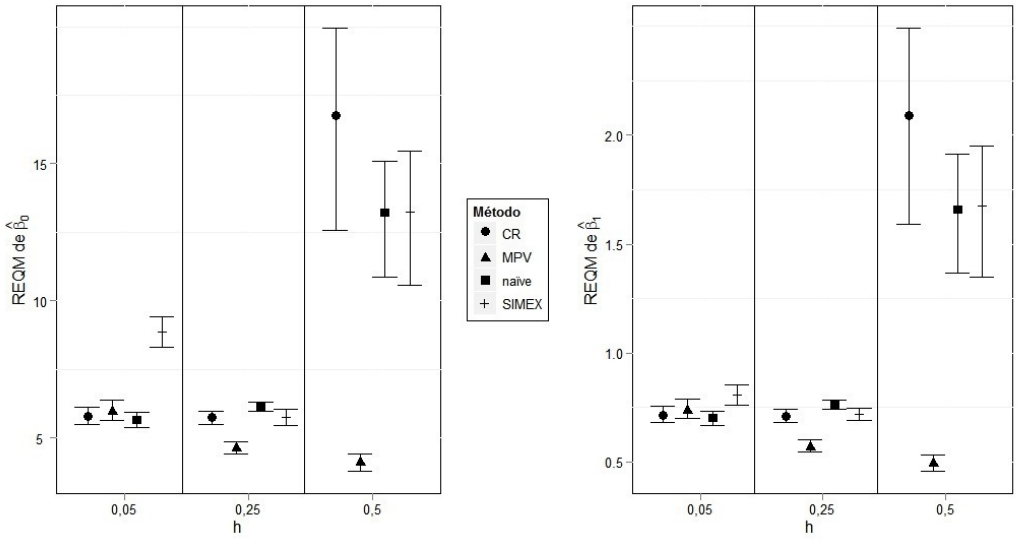

Figura 5.2: Intervalo de 95% de confiança para REQM dos estimadores de β_0 e β_1; cenário 1, $n_1 = 32$

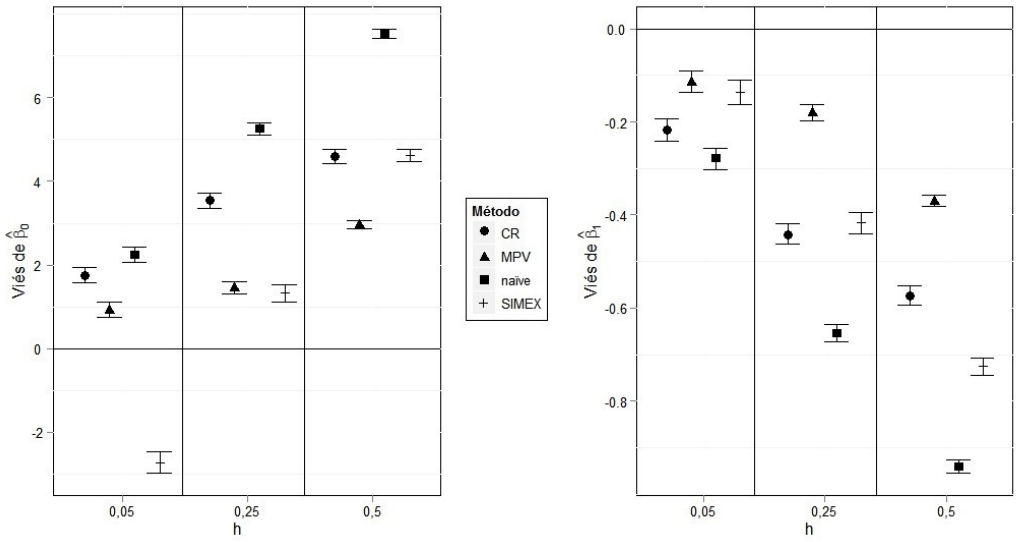

Figura 5.3: Intervalo de 95% de confiança para o viés dos estimadores de β_0 e β_1; cenário 1, $n_1 = 105$

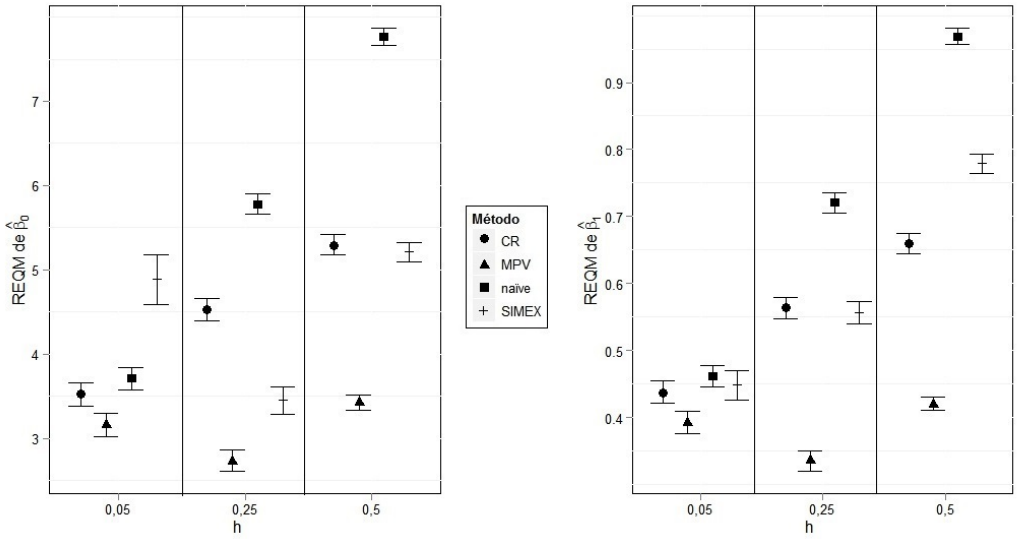

Figura 5.4: Intervalo de 95% de confiança para REQM dos estimadores de β_0 e β_1; cenário 1, $n_1 = 105$

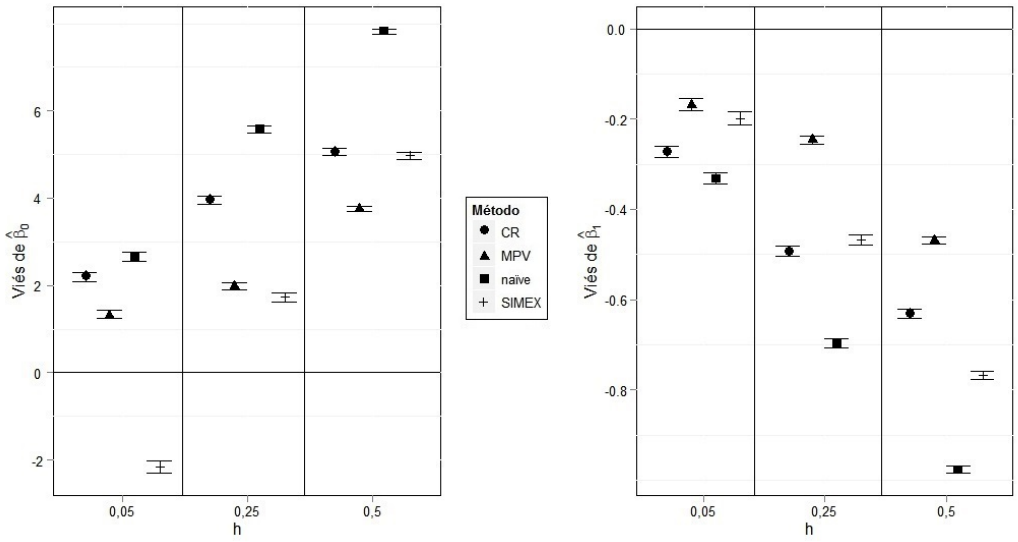

Figura 5.5: Intervalo de 95% de confiança para o viés dos estimadores de β_0 e β_1; cenário 1, $n_1 = 301$

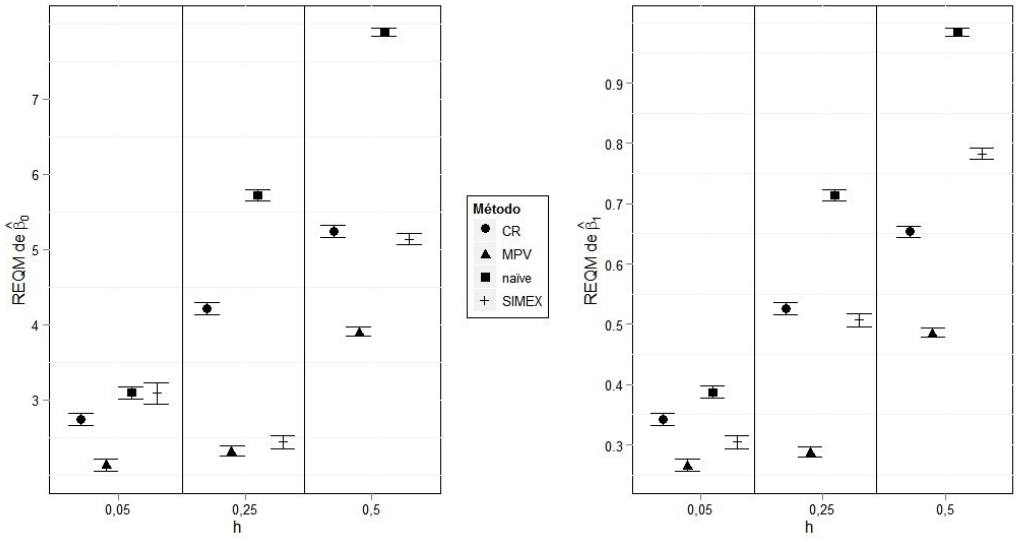

Figura 5.6: Intervalo de 95% de confiança para REQM dos estimadores de β_0 e β_1; cenário 1, $n_1 = 301$

Tabela 5.4: Mediana das medidas de desempenho na predição e estatística KS para o cenário 1

	Se			Es			VPP			VPN		
	$n_2 = 13$											
h	0,05	0,25	0,5	0,05	0,25	0,5	0,05	0,25	0,5	0,05	0,25	0,5
MPV	83,3	85,7	85,7	87,5	85,7	87,5	87,5	85,7	87,5	85,7	85,7	85,7
CR	83,3	85,7	85,7	87,5	85,7	87,5	87,5	85,7	85,7	85,7	85,7	85,7
SIMEX	83,3	85,7	85,7	88,8	85,7	87,5	88,8	85,7	87,5	83,3	85,7	85,7
naïve	83,3	85,7	85,7	87,5	85,7	87,5	87,5	85,7	85,7	85,7	85,7	85,7
	$n_2 = 45$											
h	0,05	0,25	0,5	0,05	0,25	0,5	0,05	0,25	0,5	0,05	0,25	0,5
MPV	85,7	85,7	85,7	86,2	85,9	86,3	86,3	85,7	85,7	85,7	85,7	85,7
CR	85,7	85,4	85,0	86,3	85,7	86,3	86,3	85,7	85,7	85,7	85,7	85,7
SIMEX	84,8	85,7	85,7	86,9	85,7	86,3	86,9	85,7	85,7	85,0	85,7	85,7
naïve	85,7	85,4	85,0	86,3	85,7	86,3	86,3	85,7	85,7	85,7	85,7	85,7
	$n_2 = 129$											
h	0,05	0,25	0,5	0,05	0,25	0,5	0,05	0,25	0,5	0,05	0,25	0,5
MPV	85,4	85,2	85,4	85,9	85,7	85,7	85,9	85,7	85,4	85,5	85,2	85,4
CR	85,4	85,1	85,4	85,9	85,7	85,5	85,9	85,7	85,4	85,5	85,2	85,4
SIMEX	85,0	85,2	85,4	86,2	85,7	85,7	86,1	85,7	85,3	85,2	85,2	85,4
naïve	85,4	85,1	85,4	85,9	85,7	85,5	85,9	85,7	85,4	85,5	85,2	85,4

Tabela 5.5: Continuação: Mediana das medidas de desempenho na predição para o cenário 1

	ACC			KS		
	$n_2 = 13$					
h	0,05	0,25	0,5	0,05	0,25	0,5
MPV	84,6	84,6	84,6	83,3	83,3	83,3
CR	84,6	84,6	84,6	83,3	83,3	83,3
SIMEX	84,6	84,6	84,6	50,0	71,4	80,0
naïve	84,6	84,6	84,6	83,3	83,3	83,3
	$n_2 = 45$					
h	0,05	0,25	0,5	0,05	0,25	0,5
MPV	84,4	84,4	84,4	76,7	76,6	76,7
CR	84,4	84,4	84,4	76,7	76,5	76,7
SIMEX	84,4	84,4	84,4	48,2	70,2	73,9
naïve	84,4	84,4	84,4	76,7	76,5	76,7
	$n_2 = 129$					
h	0,05	0,25	0,5	0,05	0,25	0,5
MPV	85,2	85,2	85,2	73,6	73,6	73,5
CR	85,2	85,2	85,2	73,6	73,4	73,5
SIMEX	85,2	85,2	85,2	49,1	68,5	72,6
naïve	85,2	85,2	85,2	73,6	73,4	73,5

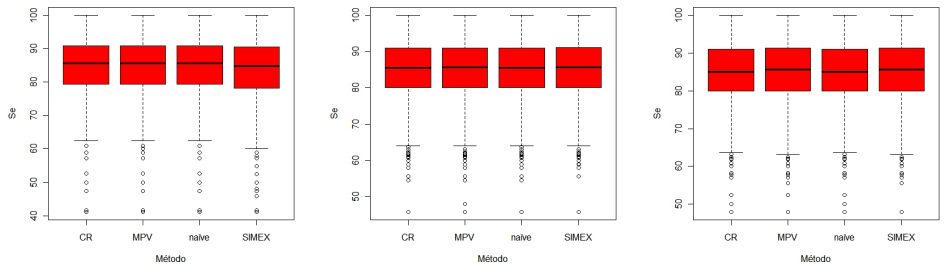

Figura 5.7: Gráficos *boxplot* dos valores de sensibilidade para $h = 0{,}05$, $0{,}25$ e $0{,}5$, respectivamente; cenário 1, $n_2 = 45$

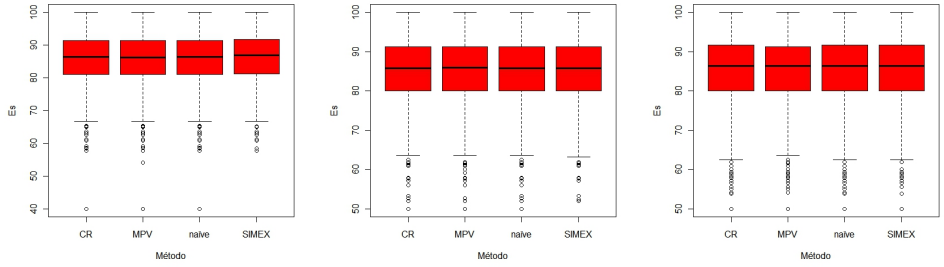

Figura 5.8: Gráficos *boxplot* dos valores de especificidade para $h = 0{,}05$, $0{,}25$ e $0{,}5$, respectivamente; cenário 1, $n_2 = 45$

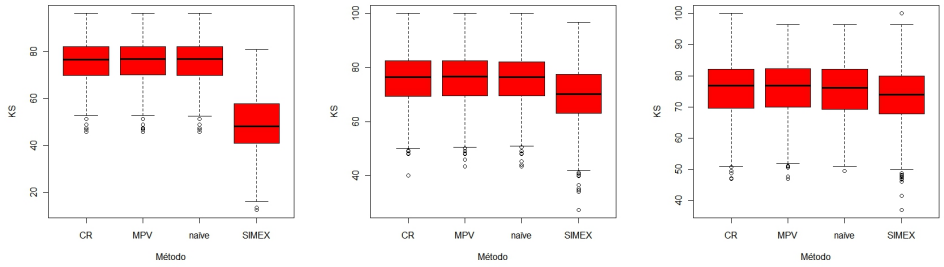

Figura 5.9: Gráficos *boxplot* dos valores da estatística KS para $h = 0{,}05$, $0{,}25$ e $0{,}5$, respectivamente; cenário 1, $n_2 = 45$

5.2 Cenário 2

Neste cenário estamos considerando uma situação mais real, ou seja, a variância do erro de medida é desconhecida, comum para todas as observações e estimada por (4.1) com três réplicas de W_i. Seus valores são 0,37, 2,33 e 7. Os demais parâmetros, fixos para todas as simulações deste cenário, são $\mu_x = 8$, $\sigma_x^2 = 7$, $\beta_0 = -16$ e $\beta_1 = 2$. A proporção esperada de eventos na amostra é de 50%.

5.2.1 Estimação

As Figuras 5.10-5.15 mostram os gráficos das estimativas pontuais e intervalares do viés e da REQM dos estimadores de β_0 e β_1 para os diferentes valores de h e n_1.

Em geral, o método de MPV apresenta o melhor desempenho em relação ao viés e REQM e o segundo melhor desempenho é obtido pelo método da CR, perdendo para as duas versões do método SIMEX apenas quando $h = 0{,}25$, em relação à REQM (em algumas situações) e ao viés, uma vez que os estimadores SIMEX só não são menos viesados que o estimador de MPV. As duas versões do método SIMEX apresentam comportamentos muito parecidos. O método *naïve* apresenta o pior desempenho tanto em relação ao viés quanto à REQM.

5.2.2 Predição

As Tabelas 5.6 e 5.7 mostram as medianas dos R valores das medidas sensibilidade (Se), especificidade (Es), VPP, VPN, ACC e estatística KS para os diferentes valores de h e n_2.

Observamos que, de uma maneira geral, não há diferença entre os métodos nas medidas de sensibilidade, especificidade, VPP, VPN, acurácia e estatística KS. Em alguns casos, há pequenas diferenças nos valores medianos obtidos pelos métodos de estimação. No entanto, assim como apresentado nas Figuras 5.7 e 5.8 do cenário anterior, as medidas apresentam comportamentos muito similares para todos os métodos de estimação. Ainda, os valores medianos das medidas diminuem conforme h aumenta, fixados o método de estimação e n_2.

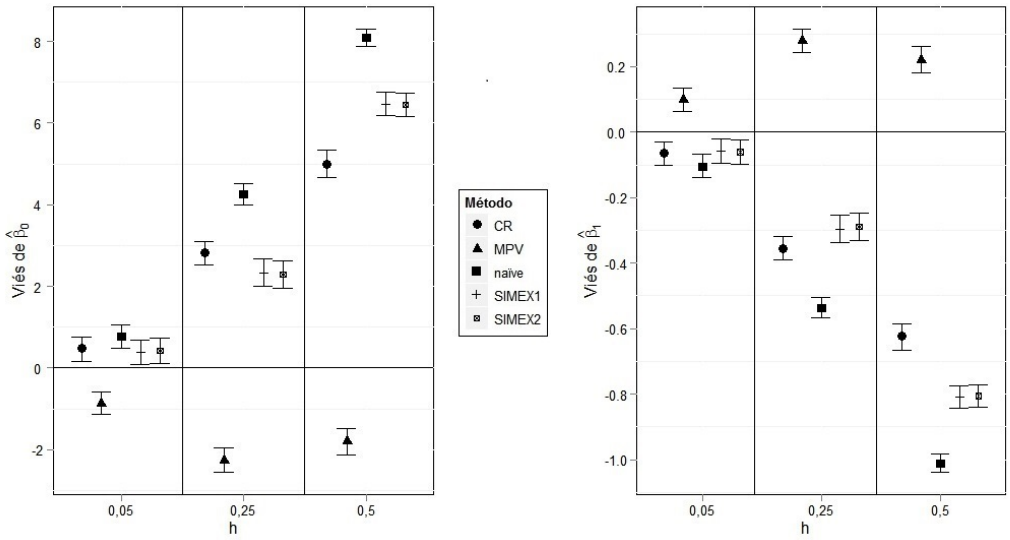

Figura 5.10: Intervalo de 95% de confiança para o viés dos estimadores de β_0 e β_1; cenário 2, $n_1 = 32$

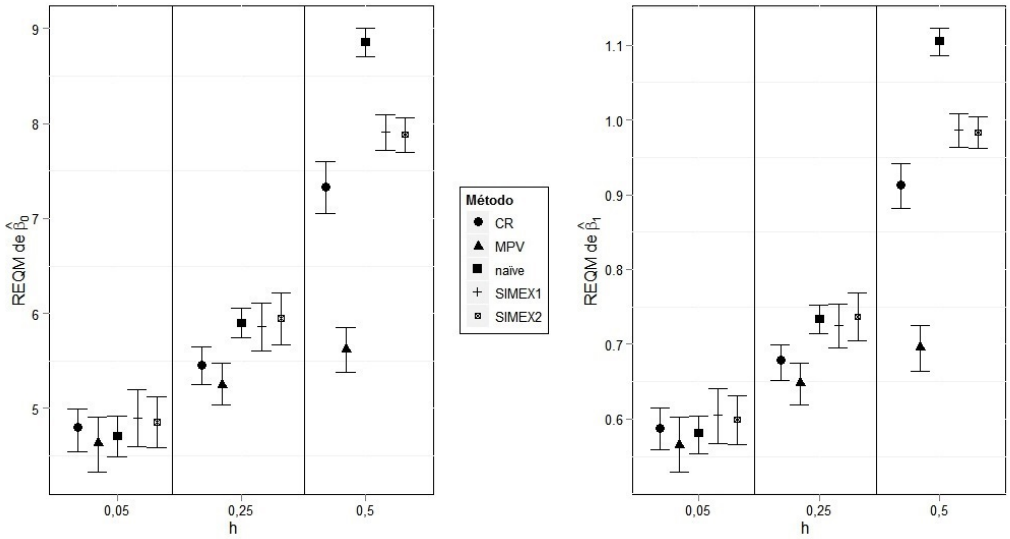

Figura 5.11: Intervalo de 95% de confiança para REQM dos estimadores de β_0 e β_1; cenário 2, $n_1 = 32$

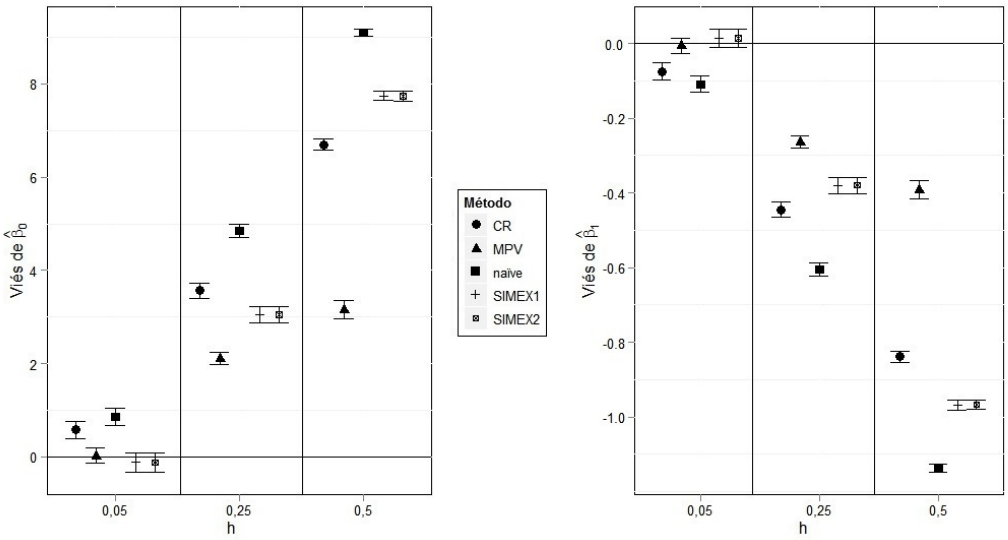

Figura 5.12: Intervalo de 95% de confiança para o viés dos estimadores de β_0 e β_1; cenário 2, $n_1 = 105$

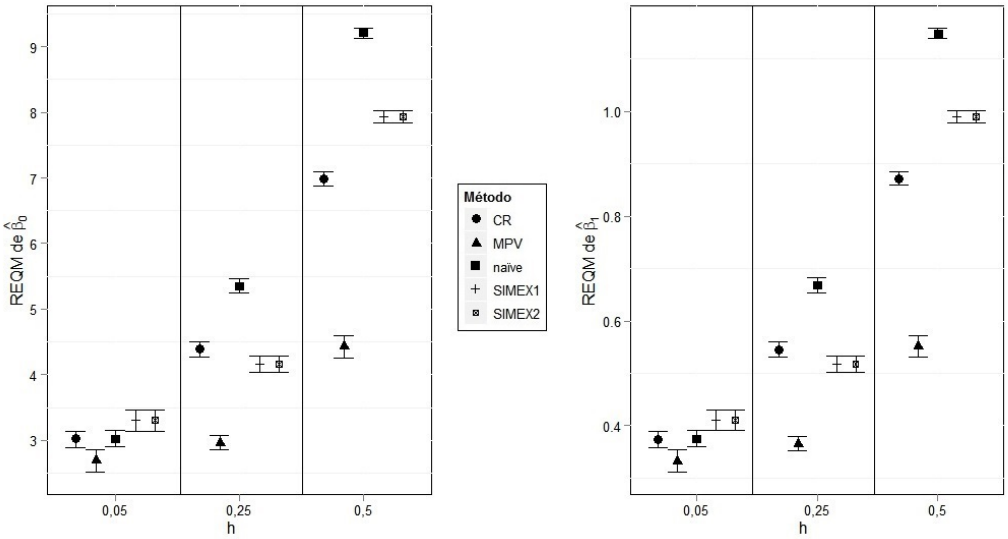

Figura 5.13: Intervalo de 95% de confiança para REQM dos estimadores de β_0 e β_1; cenário 2, $n_1 = 105$

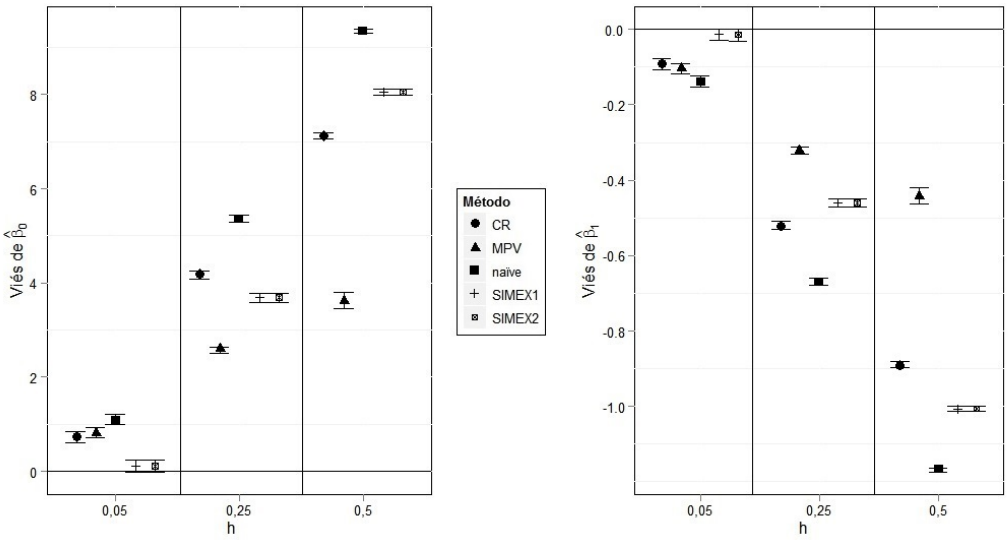

Figura 5.14: Intervalo de 95% de confiança para o viés dos estimadores de β_0 e β_1; cenário 2, $n_1 = 301$

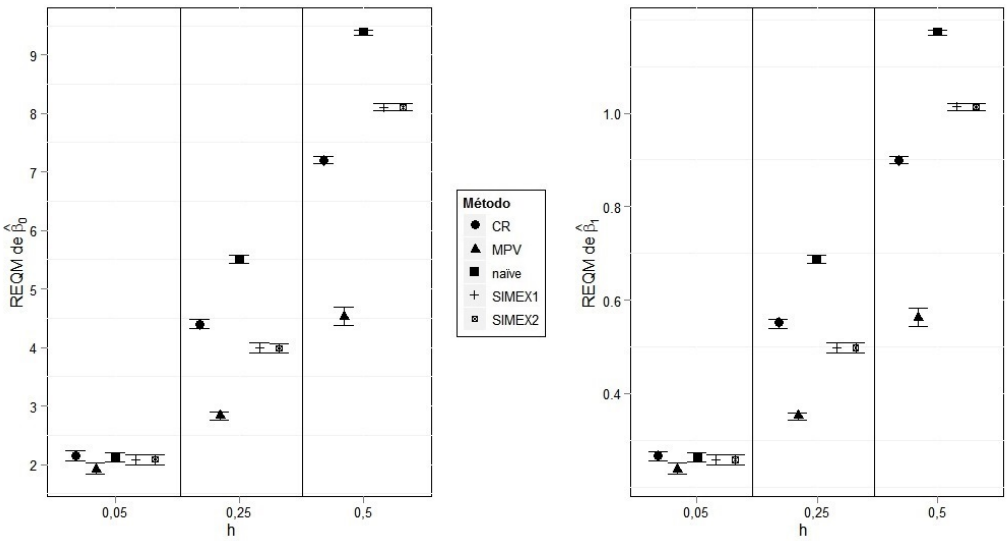

Figura 5.15: Intervalo de 95% de confiança para REQM dos estimadores de β_0 e β_1; cenário 2, $n_1 = 301$

Tabela 5.6: Mediana das medidas de desempenho na predição para o cenário 2

	Se			Es			VPP			VPN		
	$n_2 = 13$											
h	0,05	0,25	0,5	0,05	0,25	0,5	0,05	0,25	0,5	0,05	0,25	0,5
MPV	87,5	84,5	83,3	100,0	87,5	85,7	100,0	87,5	83,9	87,5	85,7	85,7
CR	87,5	83,3	83,3	100,0	87,5	85,7	100,0	87,5	83,3	87,5	85,7	85,7
SIMEX1	87,5	84,5	83,3	100,0	87,5	85,7	100,0	87,5	83,3	87,5	85,7	83,3
SIMEX2	87,5	85,7	83,3	100,0	87,5	85,7	100,0	87,5	83,3	87,5	85,7	85,7
naïve	87,5	84,5	83,3	100,0	87,5	85,7	100,0	87,5	83,3	87,5	85,7	85,7
	$n_2 = 45$											
h	0,05	0,25	0,5	0,05	0,25	0,5	0,05	0,25	0,5	0,05	0,25	0,5
MPV	89,4	85,7	82,6	90,4	86,9	83,3	90,4	86,3	82,3	89,4	86,3	82,6
CR	89,4	86,2	82,6	90,4	86,9	83,3	90,4	86,3	82,6	89,4	86,3	82,6
SIMEX1	89,4	85,7	82,6	90,0	86,9	83,3	90,4	86,3	82,4	89,6	86,3	82,6
SIMEX2	89,4	86,2	82,6	90,0	86,9	83,3	90,4	86,3	82,6	89,4	86,3	82,6
naïve	89,4	86,2	82,6	90,3	86,9	82,6	90,4	86,3	82,3	89,6	86,3	82,6
	$n_2 = 129$											
h	0,05	0,25	0,5	0,05	0,25	0,5	0,05	0,25	0,5	0,05	0,25	0,5
MPV	90,0	86,4	81,8	89,8	86,5	81,5	90,1	86,5	81,5	89,8	86,7	81,3
CR	90,0	86,4	81,6	89,8	86,5	81,6	90,1	86,5	81,5	89,8	86,7	81,3
SIMEX1	90,0	86,4	81,6	89,8	86,5	81,6	90,1	86,5	81,6	89,8	86,7	81,3
SIMEX2	90,0	86,4	81,6	90,0	86,5	81,7	90,1	86,5	81,5	89,8	86,7	81,2
naïve	90,0	86,4	81,5	89,8	86,4	81,5	90,0	86,5	81,5	89,8	86,7	81,3

Tabela 5.7: Continuação: Mediana das medidas de desempenho na predição para o cenário 2

	ACC			KS		
	$n_2 = 13$					
h	0,05	0,25	0,5	0,05	0,25	0,5
MPV	84,6	84,6	84,6	87,5	83,3	75,0
CR	84,6	84,6	84,6	87,5	83,3	75,0
SIMEX1	84,6	84,6	84,6	87,5	83,3	75,0
SIMEX2	84,6	84,6	84,6	87,5	83,3	75,0
naïve	84,6	84,6	84,6	87,5	83,3	75,0
	$n_2 = 45$					
h	0,05	0,25	0,5	0,05	0,25	0,5
MPV	88,8	86,6	82,2	83,1	77,6	68,2
CR	88,8	86,6	82,2	83,3	77,6	68,0
SIMEX1	88,8	86,6	82,2	83,1	77,5	68,0
SIMEX2	88,8	86,6	82,2	83,3	77,6	67,8
naïve	88,8	86,6	82,2	83,1	77,6	67,8
	$n_2 = 129$					
h	0,05	0,25	0,5	0,05	0,25	0,5
MPV	89,9	86,8	81,3	80,8	74,3	64,6
CR	89,9	86,8	81,3	81,0	73,8	64,3
SIMEX1	89,9	86,8	81,3	80,9	74,1	64,3
SIMEX2	89,9	86,8	81,3	80,9	74,0	64,3
naïve	89,9	86,8	81,3	80,8	73,8	64,1

5.3 Cenário 3

Como no cenário 2, a variância do erro de medida é desconhecida, comum para todas as observações e estimada por (4.1) com três réplicas de W_i. Seus valores são 0,37, 2,33 e 7. Os demais parâmetros, fixos para todas as simulações deste cenário, são $\mu_x = 8$, $\sigma_x^2 = 7$, $\beta_0 = -16$ e $\beta_1 = 1{,}55$. A proporção esperada de eventos na amostra é de 22% e não consideramos $n = 45$ por saber que teríamos poucos eventos na amostra.

5.3.1 Estimação

As Figuras 5.16-5.19 mostram os gráficos das estimativas pontuais e intervalares do viés e da REQM dos estimadores de β_0 e β_1 para os diferentes valores de h e n_1.

O estimador de MPV é o menos viesado para $h > 0{,}05$ e com menor REQM para, conjuntamente, $n_1 = 301$ e $h > 0{,}05$. O estimador da CR apresenta os menores valores de REQM quando $n_1 = 105$ e nestes casos, o estimador de MPV tem o segundo menor valor quando $h = 0{,}5$. Assim como no cenário 2, o estimador da CR é o segundo menos viesado, mas destacamos duas exceções. A primeira é quando $h = 0{,}25$, pois os estimadores SIMEX só não são menos viesados que o estimador de MPV para $n_1 = 301$ e apresentam valores similares a este estimador para $n_1 = 105$. A segunda exceção se deve ao fato de o estimador da CR ser o menos viesado para $h = 0{,}05$ quando $n_1 = 301$ e apresentar viés similar ao estimador *naïve*, o menos viesado, quando $n_1 = 105$. As duas versões do método SIMEX apresentam comportamentos muito parecidos e na maioria das vezes só são melhores em relação ao viés e REQM que o método *naïve*. Em geral, o método *naïve* apresenta o pior desempenho tanto em relação ao viés quanto em relação à REQM para $h > 0{,}05$.

5.3.2 Predição

As Tabelas 5.8 e 5.9 mostram as medianas dos R valores das medidas sensibilidade (Se), especificidade (Es), VPP, VPN, ACC e estatística KS para os diferentes valores de h e n_2.

Não há diferença importante entre os métodos de estimação nas medidas de sensibilidade, especificidade, VPP, VPN e acurácia para todo h e n_2. Ainda, observamos que os

valores dessas medidas diminuem conforme a variância do erro de medida aumenta. Os valores da estatística KS são similares para todos os métodos de estimação, com uma ressalva quando $h = 0{,}5$ e $n_2 = 45$, pois a mediana da estatística KS obtida pelo método da CR é três unidades maior do que o valor obtido pelo método *naïve*. Ainda, temos que os valores da estatística diminuem conforme h aumenta.

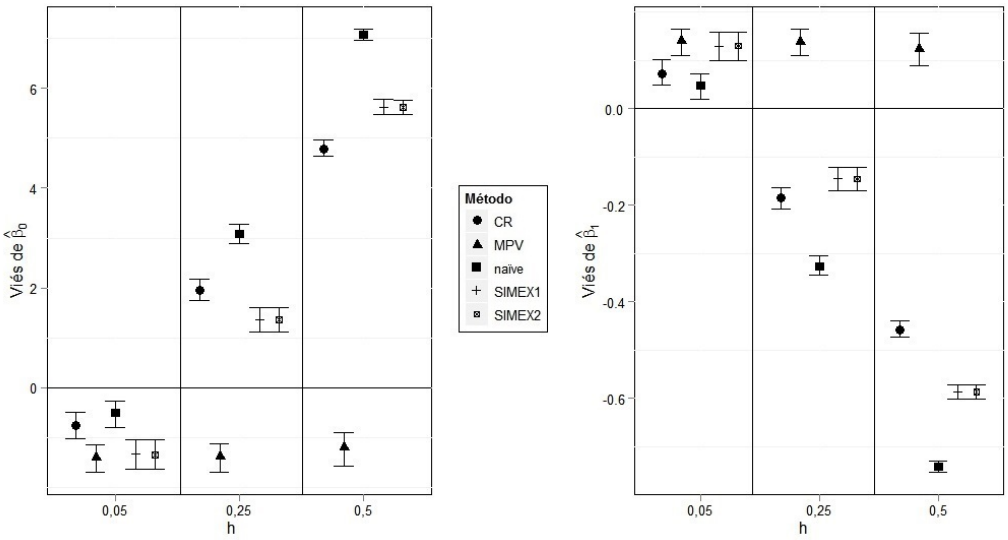

Figura 5.16: Intervalo de 95% de confiança para o viés dos estimadores de β_0 e β_1; cenário 3, $n_1 = 105$

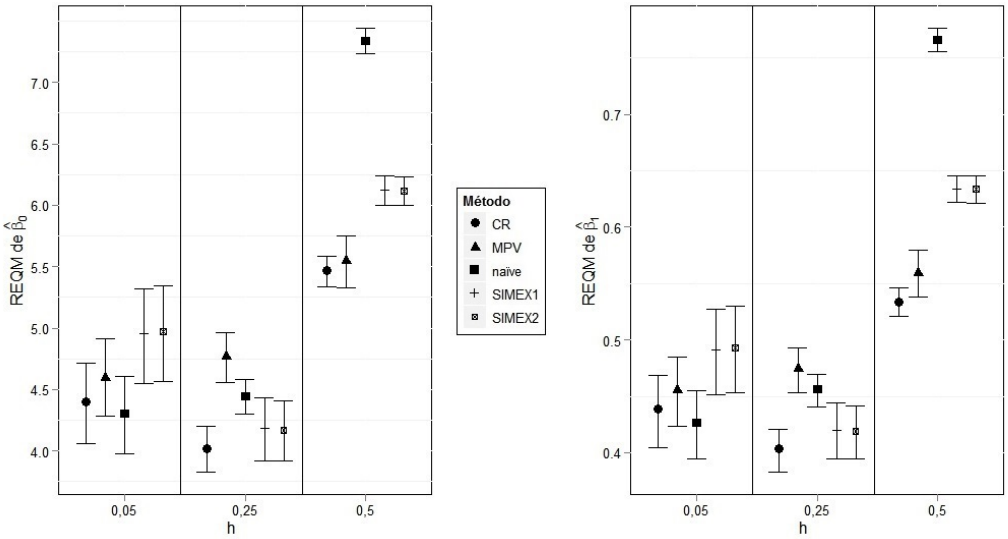

Figura 5.17: Intervalo de 95% de confiança para REQM dos estimadores de β_0 e β_1; cenário 3, $n_1 = 105$

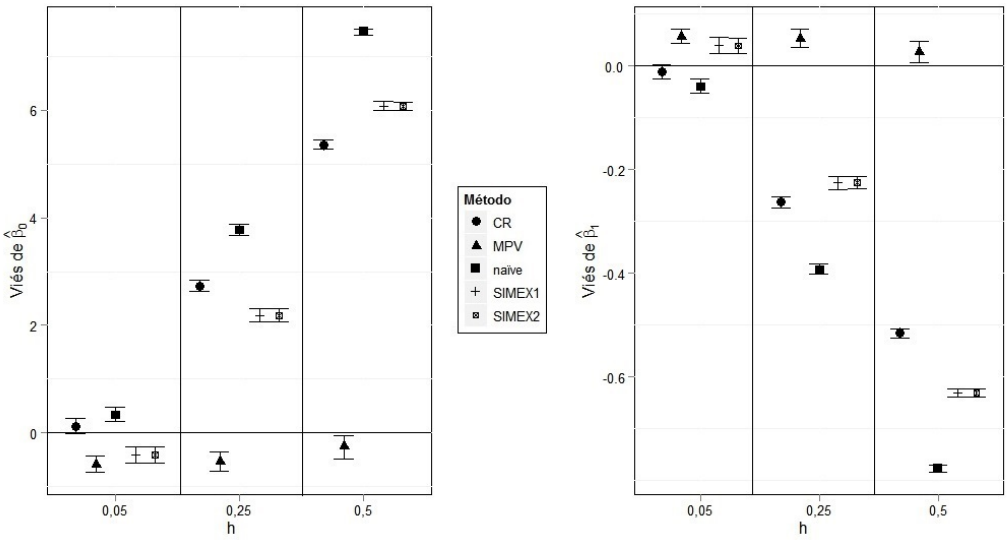

Figura 5.18: Intervalo de 95% de confiança para o viés dos estimadores de β_0 e β_1; cenário 3, $n_1 = 301$

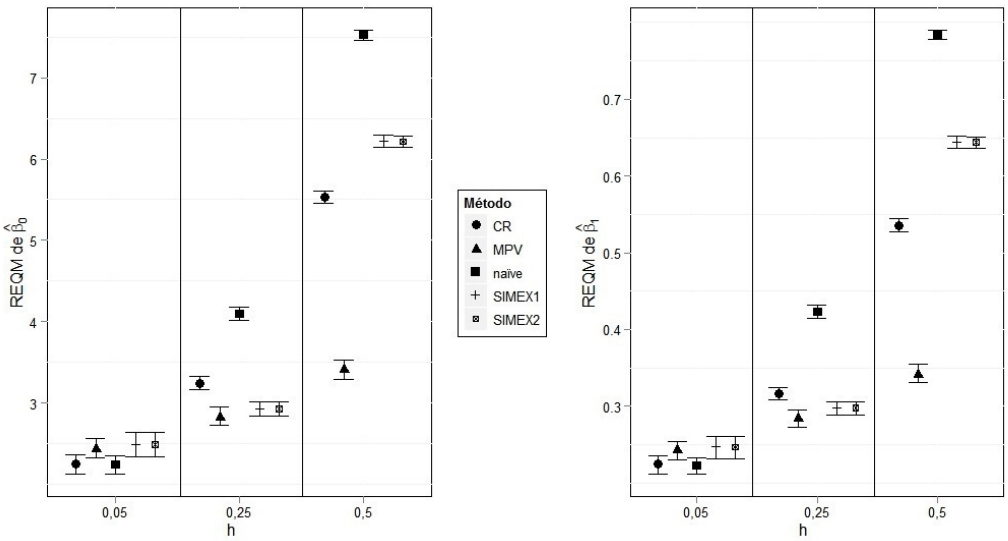

Figura 5.19: Intervalo de 95% de confiança para REQM dos estimadores de β_0 e β_1; cenário 3, $n_1 = 301$

Tabela 5.8: Mediana das medidas de desempenho na predição para o cenário 3

	Se			Es			VPP			VPN		
	$n_2 = 45$											
h	0,05	0,25	0,5	0,05	0,25	0,5	0,05	0,25	0,5	0,05	0,25	0,5
MPV	90,0	88,8	84,6	89,4	87,0	83,3	70,0	64,7	57,1	96,9	96,8	96,0
CR	90,0	88,8	84,6	89,4	87,0	83,3	70,0	64,4	57,1	96,9	96,8	96,0
SIMEX1	90,0	88,8	84,6	89,4	87,0	83,1	70,0	64,7	57,1	96,9	96,8	96,0
SIMEX2	90,0	88,8	84,6	89,5	87,1	83,3	70,0	64,7	57,1	96,9	96,8	96,0
naïve	90,0	88,8	84,6	89,6	86,8	83,3	70,0	64,2	57,1	96,9	96,8	96,0
	$n_2 = 129$											
h	0,05	0,25	0,5	0,05	0,25	0,5	0,05	0,25	0,5	0,05	0,25	0,5
MPV	89,4	87,0	82,6	88,4	85,4	81,3	67,5	61,0	53,8	96,8	96,1	94,8
CR	89,6	87,0	82,7	88,4	85,4	81,1	67,5	60,9	54,0	96,8	96,2	94,8
SIMEX1	89,6	87,0	82,7	88,3	85,5	81,1	67,5	61,2	53,8	96,8	96,1	94,9
SIMEX2	89,6	87,0	82,7	88,3	85,5	81,1	67,5	61,3	53,8	96,8	96,1	94,8
naïve	89,6	87,0	82,7	88,3	85,5	81,2	67,5	61,3	53,6	96,8	96,1	94,9

Tabela 5.9: Continuação: Mediana das medidas de desempenho na predição para o cenário 3

	ACC			KS		
	$n_2 = 45$					
h	0,05	0,25	0,5	0,05	0,25	0,5
MPV	88,8	86,6	82,2	80,4	75,9	67,2
CR	88,8	86,6	82,2	80,5	76,7	70,0
SIMEX1	88,8	86,6	82,2	80,4	76,2	68,0
SIMEX2	88,8	86,6	82,2	80,4	76,0	68,0
naïve	88,8	86,6	82,2	80,4	75,9	67,1
	$n_2 = 129$					
h	0,05	0,25	0,5	0,05	0,25	0,5
MPV	88,3	85,2	81,3	78,1	72,6	64,1
CR	88,3	85,2	81,3	78,2	73,0	65,2
SIMEX1	88,3	85,2	81,3	78,1	72,7	64,3
SIMEX2	88,3	85,2	81,3	78,1	72,6	64,2
naïve	88,3	85,2	81,3	78,1	72,6	64,2

5.4 Cenário 4

Ainda no estudo de amostras desbalanceadas, consideramos neste cenário os mesmos valores dos parâmetros do cenário 3, com exceção de β_1 que agora é igual a 1,35. Desta maneira, este cenário é o que apresenta amostras mais desbalanceadas, com proporção esperada dos eventos de 10%. Não consideramos $n = 45$ por saber que teríamos poucos eventos na amostra.

5.4.1 Estimação

As Figuras 5.20-5.23 mostram os gráficos das estimativas pontuais e intervalares do viés e da REQM dos estimadores de β_0 e β_1 para os diferentes valores de h e n_1.

O estimador de MPV apresenta o melhor desempenho em relação ao viés quando, conjuntamente, $n_1 = 301$ e $h > 0{,}05$ e quando $n_1 = 105$ e $h = 0{,}5$. O estimador da CR é o segundo menos viesado para todos os valores de h, com exceção quando, conjuntamente, $h = 0{,}25$ e $n_1 = 301$, pois os estimadores SIMEX são os segundos menos viesados. O estimador *naïve* apresenta o menor viés quando $h = 0{,}05$, com vieses parecidos aos obtidos pelo estimador da CR, mas é o mais viesado para os demais valores de h. Em relação à REQM, o método da CR é, em geral, o método com os menores valores e o método de MPV apresenta os maiores valores, principalmente para $n_1 = 105$. O método *naïve* apresenta os segundos maiores valores da REQM e é o método com maior REQM quando $n_1 = 301$ e $h = 0{,}5$.

5.4.2 Predição

As Tabelas 5.10 e 5.11 mostram as medianas dos R valores das medidas sensibilidade (Se), especificidade (Es), VPP, VPN, ACC e estatística KS para os diferentes valores de h e n_2.

Para as medidas de desempenho sensibilidade, especificidade, VPP, VPN e acurácia, não há diferença importante entre os métodos de estimação. Ainda, observamos que os valores dessas medidas diminuem conforme a variância do erro de medida aumenta. O valor mediano da estatística KS obtido pelo método da CR é o maior para todos os valores de

h e n_2, seguido, na maioria das vezes, pelo método de MPV. O método *naïve* apresenta os menores valores medianos da estatística KS e a diferença dos seus valores com relação aos valores obtidos pelo método da CR aumenta conforme h aumenta. No entanto, através da Figura 5.24, podemos observar que os gráficos *boxplot* dos valores da estatística KS de cada método de estimação apresentam comportamentos muito similares. Os valores da estatística KS diminuem conforme h aumenta, fixados n_2 e o método de estimação.

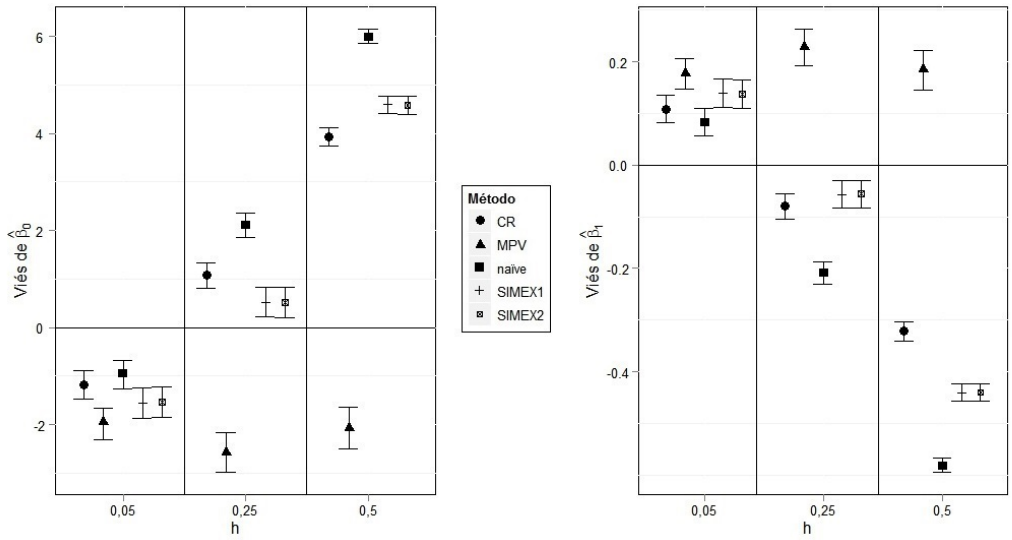

Figura 5.20: Intervalo de 95% de confiança para o viés dos estimadores de β_0 e β_1; cenário 4, $n_1 = 105$

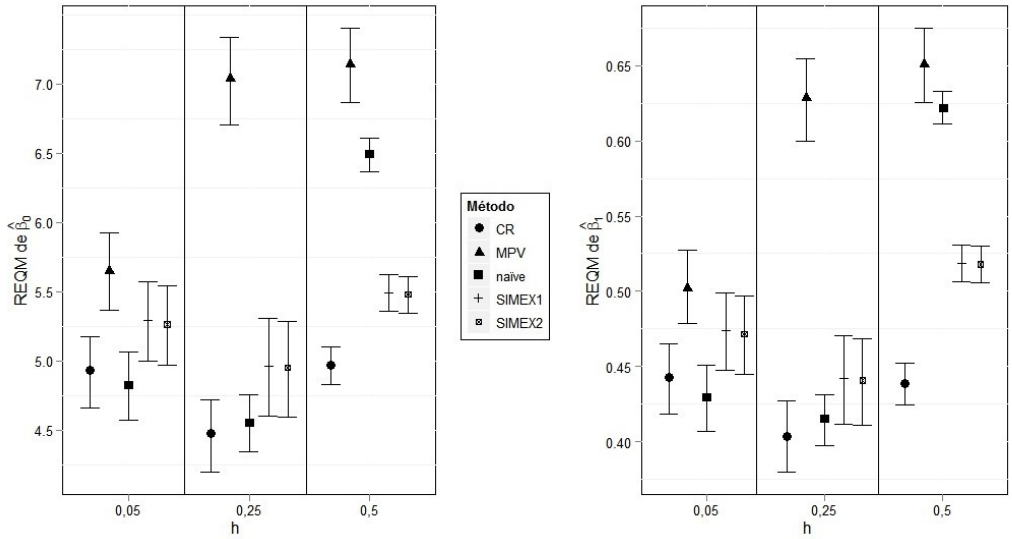

Figura 5.21: Intervalo de 95% de confiança para REQM dos estimadores de β_0 e β_1; cenário 4, $n_1 = 105$

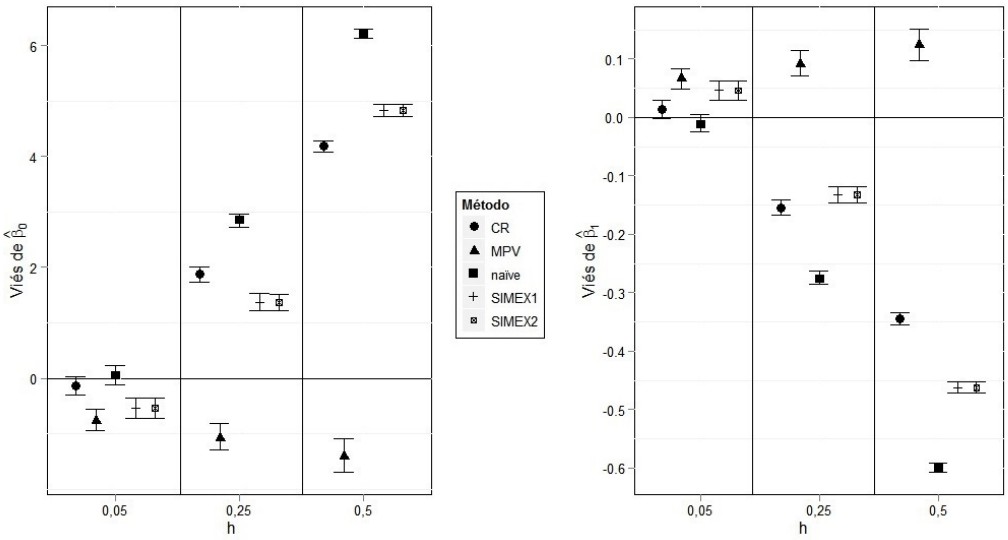

Figura 5.22: Intervalo de 95% de confiança para o viés dos estimadores de β_0 e β_1; cenário 4, $n_1 = 301$

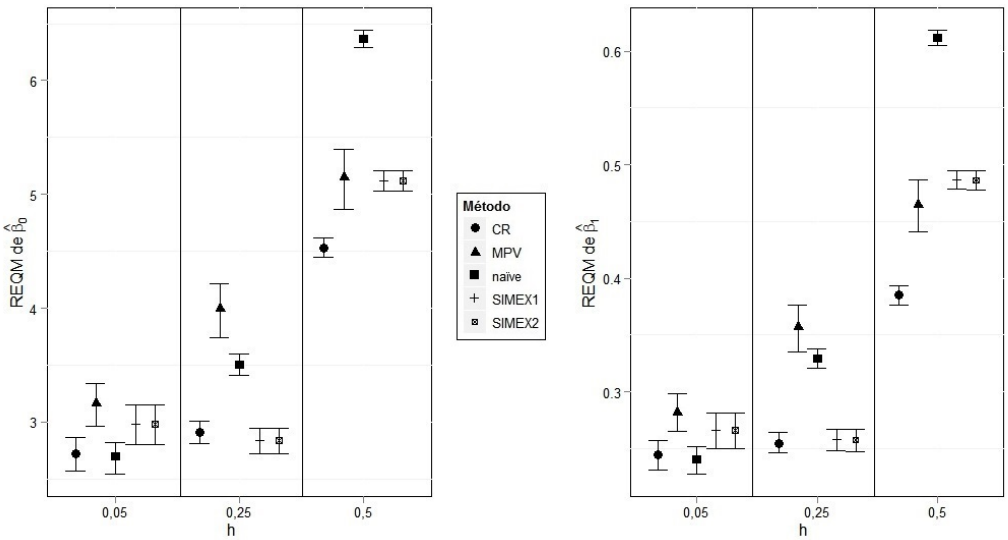

Figura 5.23: Intervalo de 95% de confiança para REQM dos estimadores de β_0 e β_1; cenário 4, $n_1 = 301$

Tabela 5.10: Mediana das medidas de desempenho na predição para o cenário 4

	Se			Es			VPP			VPN		
	$n_2 = 45$											
h	0,05	0,25	0,5	0,05	0,25	0,5	0,05	0,25	0,5	0,05	0,25	0,5
MPV	100,0	100,0	87,5	90,2	88,0	84,6	50,0	44,4	36,3	100,0	100,0	97,5
CR	100,0	100,0	87,5	90,2	88,0	84,6	50,0	44,4	36,3	100,0	100,0	97,5
SIMEX1	100,0	100,0	87,5	90,2	88,0	84,6	50,0	44,4	36,3	100,0	100,0	97,6
SIMEX2	100,0	100,0	87,5	90,2	88,0	84,6	50,0	44,4	36,3	100,0	100,0	97,5
naïve	100,0	100,0	87,5	90,2	88,0	84,6	50,0	44,4	36,3	100,0	100,0	97,5
	$n_2 = 129$											
h	0,05	0,25	0,5	0,05	0,25	0,5	0,05	0,25	0,5	0,05	0,25	0,5
MPV	90,0	88,8	84,6	88,6	86,5	83,0	46,1	40,6	34,6	99,0	98,9	98,0
CR	90,0	88,8	84,6	88,7	86,5	83,0	46,4	40,5	34,6	99,0	98,9	98,0
SIMEX1	90,0	88,2	84,6	88,6	86,5	83,0	46,1	40,6	34,4	99,0	98,9	98,0
SIMEX2	90,2	88,8	84,6	88,6	86,5	83,0	46,4	40,6	34,6	99,0	98,9	98,0
naïve	90,0	88,8	84,2	88,6	86,4	82,9	45,8	40,6	34,4	99,0	98,9	97,9

Tabela 5.11: Continuação: Mediana das medidas de desempenho na predição para o cenário 4

	ACC			KS		
	$n_2 = 45$					
h	0,05	0,25	0,5	0,05	0,25	0,5
MPV	91,1	88,8	84,4	75,0	70,5	66,1
CR	91,1	88,8	84,4	77,5	75,0	70,5
SIMEX1	91,1	88,8	84,4	75,0	70,5	64,1
SIMEX2	91,1	88,8	84,4	75,4	71,2	64,1
naïve	91,1	88,8	84,4	75,0	70,0	61,5
	$n_2 = 129$					
h	0,05	0,25	0,5	0,05	0,25	0,5
MPV	89,1	86,8	82,9	74,2	70,2	62,3
CR	89,1	86,8	82,9	75,0	72,5	66,0
SIMEX1	89,1	86,8	82,9	74,2	70,0	59,9
SIMEX2	89,1	86,8	82,9	74,2	70,0	59,9
naïve	88,3	86,8	82,9	74,2	69,0	58,3

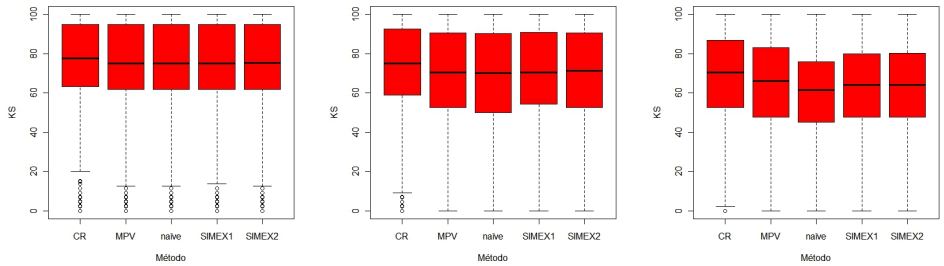

Figura 5.24: Gráficos *boxplot* dos valores da estatística KS para $h = 0{,}05$, $0{,}25$ e $0{,}5$, respectivamente; cenário 4, $n_2 = 45$

5.5 Cenário 5

Estudamos neste e no próximo cenário (cenário 6) o comportamento dos métodos de estimação sob a presença de erro de medida heteroscedástico. Ainda consideramos os mesmos métodos de estimação estudados nos cenários anteriores, que, com exceção do método SIMEX2, não consideram a presença de heteroscedasticade. O trabalho desenvolvido por Spiegelman *et al.* (2011) motivou o estudo deste e do próximo cenário, pois propuseram uma modificação para o estimador da calibração da regressão para corrigir os coeficientes de regressão para covariáveis medidas com erro heteroscedástico. Eles observaram que o estimador da calibração da regressão usual (o qual estamos considerando neste trabalho, que não leva em conta a heteroscedasticidade) se comporta melhor ou igual ao estimador proposto. Desta maneira, queremos verificar como os outros estimadores considerados neste trabalho se comportam na presença de erro de medida heteroscedástico.

Vale ressaltar que no trabalho de Spiegelman *et al.* (2011) foram utilizados dados de validação, ou seja, para uma parte da amostra, a variável X_i é observada. Desta maneira, no método de calibração da regressão usual, a esperança condicional de X_i dado (W_i, Z_i) é estimada usando os dados de validação (Rosner *et al.*, 1989).

Neste cenário, a variância do erro de medida é desconhecida e estimada por (4.1) com três réplicas de W_i. Na simulação dos dados, $\sigma_{ui}^2 = |x_i|^{0,5}\sigma_u^2$, o que chamamos de baixa heteroscedasticidade, e os valores de σ_u^2 são 0,37, 2,33 e 7. Os demais parâmetros, fixos para todas as simulações deste cenário, são $\mu_x = 8$, $\sigma_x^2 = 7$, $\beta_0 = -16$ e $\beta_1 = 2$.

5.5.1 Estimação

As Figuras 5.25-5.30 mostram os gráficos das estimativas pontuais e intervalares do viés e da REQM dos estimadores de β_0 e β_1 para os diferentes valores de h e n_1.

O estimador de MPV é o melhor método na presença de baixa heteroscedasticidade, tanto em relação ao viés quanto em relação à REQM. O estimador *naïve* é o mais viesado e só não é o método com maior REQM quando o tamanho da amostra é pequena (para todos os valores de h) e quando, conjuntamente, a variância do erro de medida é alta e o tamanho da amostra é intermediária, ou seja, quando $h = 0{,}5$ e $n_1 = 105$. Nessas situações, o estimador da CR é o método com maior REQM, mesmo apresentando bom desempenho no viés. As duas versões do estimador SIMEX apresentam comportamentos similares, sendo melhores, na maioria dos casos, apenas que o estimador *naïve*.

5.5.2 Predição

As Tabelas 5.12 e 5.13 mostram as medianas dos R valores das medidas sensibilidade (Se), especificidade (Es), VPP, VPN, ACC e estatística KS para os diferentes valores de h e n_2.

Não há diferenças que mereçam destaques nas medidas de sensibilidade, especificidade, VPP, VPN, acurácia e estatística KS entre os métodos de estimação considerados. Mas, para todo n_2, os valores medianos dessas medidas diminuem conforme a variância do erro de medida aumenta, ou seja, conforme h aumenta.

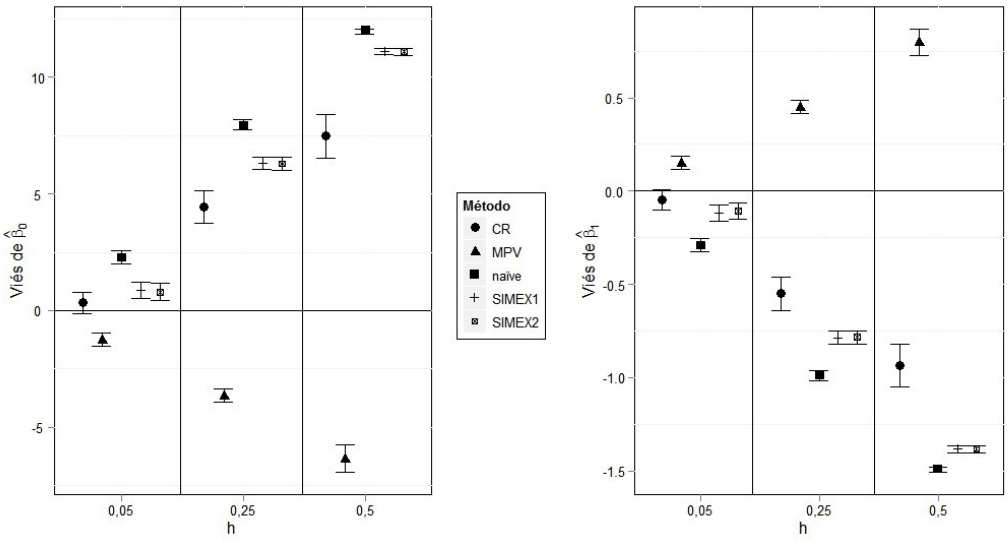

Figura 5.25: Intervalo de 95% de confiança para o viés dos estimadores de β_0 e β_1; cenário 5, $n_1 = 32$

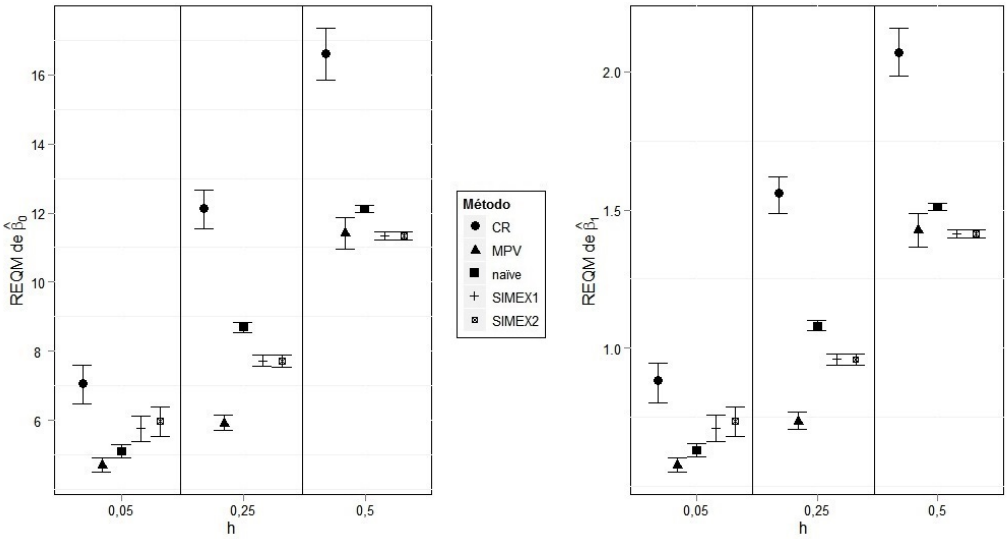

Figura 5.26: Intervalo de 95% de confiança para REQM dos estimadores de β_0 e β_1; cenário 5, $n_1 = 32$

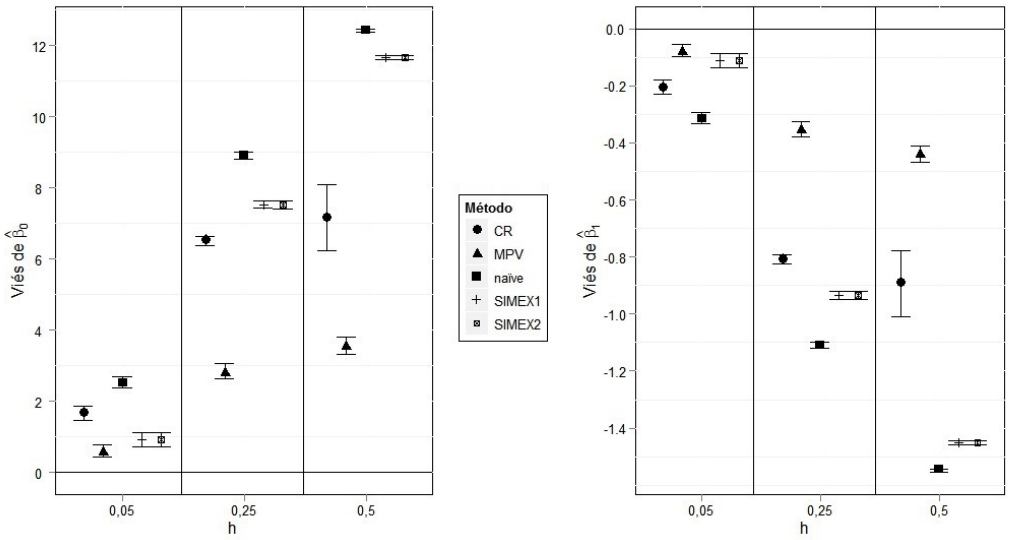

Figura 5.27: Intervalo de 95% de confiança para o viés dos estimadores de β_0 e β_1; cenário 5, $n_1 = 105$

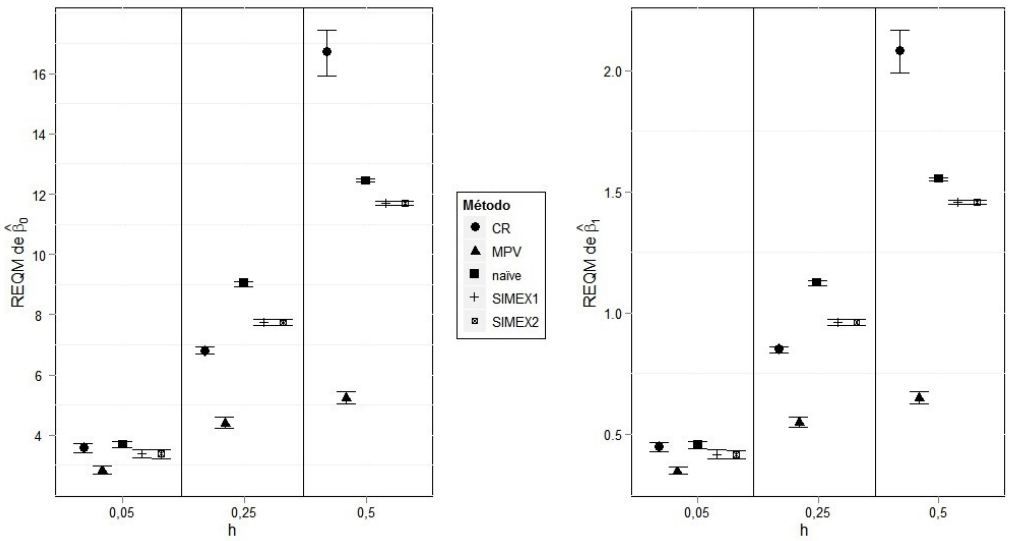

Figura 5.28: Intervalo de 95% de confiança para REQM dos estimadores de β_0 e β_1; cenário 5, $n_1 = 105$

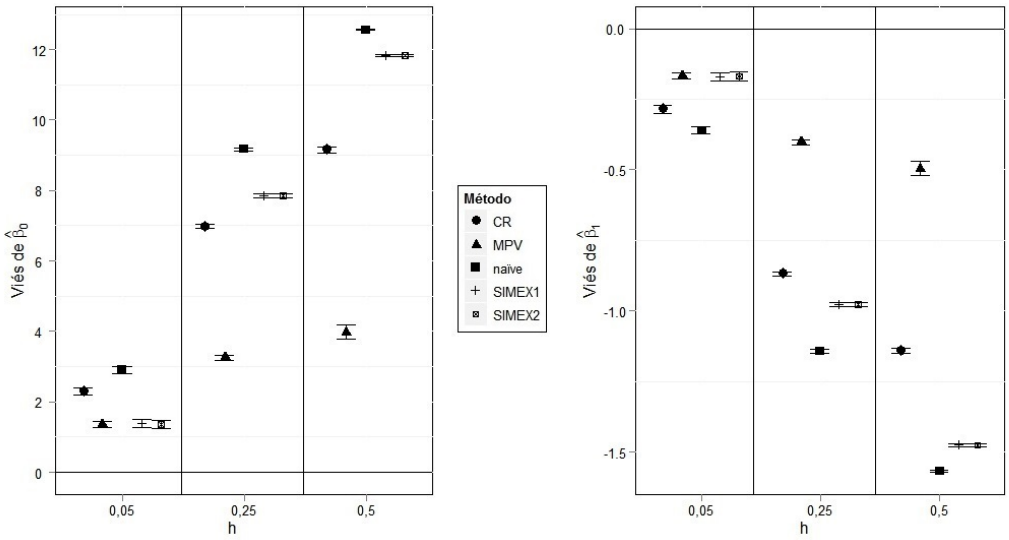

Figura 5.29: Intervalo de 95% de confiança para o viés dos estimadores de β_0 e β_1; cenário 5, $n_1 = 301$

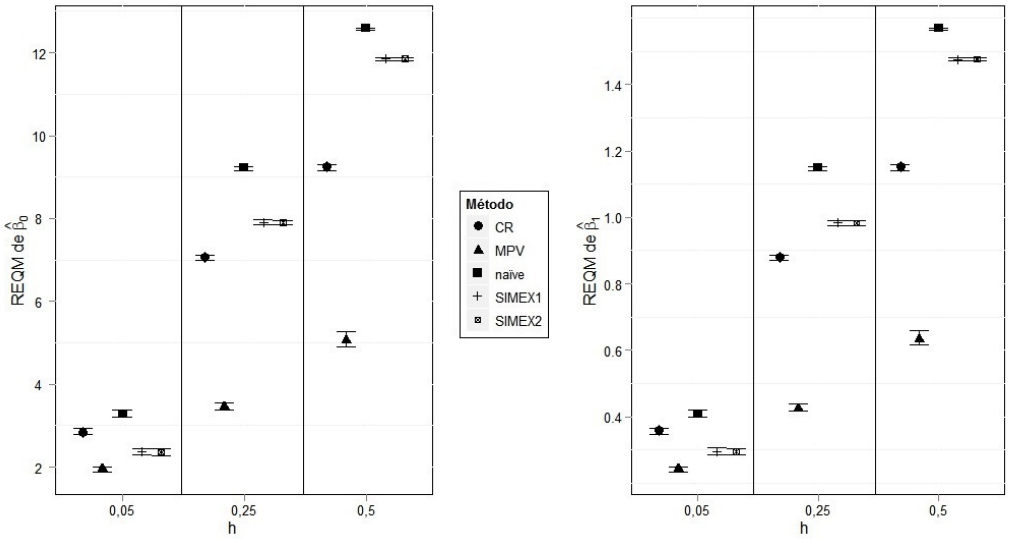

Figura 5.30: Intervalo de 95% de confiança para REQM dos estimadores de β_0 e β_1; cenário 5, $n_1 = 301$

Tabela 5.12: Mediana das medidas de desempenho na predição para o cenário 5

	Se			Es			VPP			VPN		
	$n_2 = 13$											
h	0,05	0,25	0,5	0,05	0,25	0,5	0,05	0,25	0,5	0,05	0,25	0,5
MPV	100,0	85,7	80,0	88,8	85,7	80,0	88,8	85,7	80,0	100,0	83,3	80,0
CR	95,4	85,7	77,7	88,8	85,7	80,0	88,8	85,7	80,0	91,2	83,3	77,7
SIMEX1	100,0	85,7	80,0	88,8	85,7	80,0	88,8	85,7	80,0	100,0	83,3	80,0
SIMEX2	100,0	85,7	77,7	88,8	85,7	80,0	88,8	85,7	80,0	95,8	83,3	80,0
naïve	100,0	85,7	78,8	88,8	85,7	80,0	88,8	85,7	80,0	91,6	83,3	77,7
	$n_2 = 45$											
h	0,05	0,25	0,5	0,05	0,25	0,5	0,05	0,25	0,5	0,05	0,25	0,5
MPV	88,8	82,6	76,1	89,4	83,3	76,0	89,4	82,6	75,0	88,8	83,3	76,0
CR	88,8	82,6	76,1	89,4	83,3	75,4	89,4	82,6	75,0	88,8	83,3	76,0
SIMEX1	88,8	82,6	76,1	89,4	83,3	75,0	89,4	82,6	75,0	88,8	83,3	76,0
SIMEX2	88,8	82,6	76,1	89,4	83,3	75,8	89,4	82,6	75,0	88,8	83,3	76,1
naïve	88,8	82,6	76,1	89,4	83,3	75,4	89,4	82,6	75,0	88,8	83,3	76,0
	$n_2 = 129$											
h	0,05	0,25	0,5	0,05	0,25	0,5	0,05	0,25	0,5	0,05	0,25	0,5
MPV	88,4	81,9	74,1	88,7	82,0	75,4	88,7	82,1	75,5	88,5	81,8	74,2
CR	88,3	81,9	74,2	88,7	82,0	75,3	88,8	82,0	75,4	88,5	81,8	74,3
SIMEX1	88,4	81,9	74,1	88,7	82,1	75,0	88,8	82,1	75,3	88,5	81,8	74,2
SIMEX2	88,3	81,9	74,2	88,7	82,1	75,0	88,7	82,1	75,4	88,5	81,8	74,3
naïve	88,3	81,9	74,2	88,7	82,0	75,0	88,8	82,0	75,4	88,5	81,7	74,3

Tabela 5.13: Continuação: Mediana das medidas de desempenho na predição para o cenário 5

	ACC			KS		
	$n_2 = 13$					
h	0,05	0,25	0,5	0,05	0,25	0,5
MPV	92,3	84,6	76,9	85,7	75,0	60,0
CR	92,3	84,6	76,9	85,7	75,0	63,8
SIMEX1	92,3	84,6	76,9	85,7	75,0	63,8
SIMEX2	92,3	84,6	76,9	85,7	75,0	63,8
naïve	92,3	84,6	76,9	85,7	75,0	63,8
	$n_2 = 45$					
h	0,05	0,25	0,5	0,05	0,25	0,5
MPV	88,8	82,2	75,5	80,9	68,4	54,0
CR	88,8	82,2	75,5	80,1	68,5	54,9
SIMEX1	88,8	82,2	75,5	80,9	68,4	53,7
SIMEX2	88,8	82,2	75,5	80,6	68,4	54,0
naïve	88,8	82,2	75,5	80,1	68,0	53,3
	$n_2 = 129$					
h	0,05	0,25	0,5	0,05	0,25	0,5
MPV	88,3	82,1	75,1	78,1	65,4	51,7
CR	88,3	82,1	75,1	78,1	65,0	51,6
SIMEX1	88,3	82,1	75,1	77,9	65,2	50,8
SIMEX2	88,3	82,1	75,1	77,9	65,2	50,8
naïve	88,3	81,3	75,1	77,9	65,0	50,5

5.6 Cenário 6

Consideramos as mesmas características e valores dos parâmetros do cenário 5, a exceção é que neste cenário $\rho = 2$, que chamamos de alta heteroscedasticidade.

5.6.1 Estimação

As Figuras 5.31-5.36 mostram os gráficos das estimativas pontuais e intervalares do viés e da REQM dos estimadores de β_0 e β_1 para os diferentes valores de h e n_1.

O estimador de MPV é o menos viesado e com os menores valores de REQM para todos os valores de h e n_1. O estimador da CR é o mais viesado, com duas exceções. A primeira é quando $h = 0{,}05$, para todos os tamanhos amostrais, e a segunda exceção é quando $h = 0{,}5$ e o tamanho da amostra é $n_1 = 301$, cujos maiores vieses são obtidos pelo estimador *naïve*. O método da CR só não é o estimador com maior valor da REQM quando $h = 0{,}05$ e os tamanhos amostrais são $n_1 = 105$ e 301, nesses casos o estimador *naïve* tem o maior valor da REQM. Nas situações em que o estimador da CR é o mais viesado e/ou com maior REQM, o estimador *naïve* é o segundo pior método. As duas versões do método SIMEX apresentam comportamentos similares e ocupam posição intermediária na comparação com os demais métodos, pois o método SIMEX é o segundo melhor nas situações que o método da CR é o pior e é o segundo pior método quando o método da CR tem bom desempenho.

5.6.2 Predição

As Tabelas 5.14 e 5.15 mostram as medianas dos R valores das medidas sensibilidade (Se), especificidade (Es), VPP, VPN, ACC e estatística KS para os diferentes valores de h e n_2.

Observamos que os valores medianos das medidas de desempenho sensibilidade, especificidade, VPP, VPN e acurácia obtidos pelo método da CR são os menores para $h > 0{,}05$. No entanto, ao avaliar os gráficos *boxplot* das Figuras 5.37 e 5.38 observamos que as medidas de sensibilidade e especificidade apresentam comportamentos similares para todos os métodos de estimação, apesar de o método da CR apresentar maior variabilidade para $h > 0{,}05$. Isto acontece para as demais medidas de desempenho consideradas.

Os valores medianos da estatística KS obtidos pelos métodos da CR e MPV são os maiores para todos os valores de h e a diferença dos valores medianos desses métodos com o obtido pelo método *naïve* aumenta conforme a variância do erro de medida aumenta. Para $n_2 = 45$ e $h = 0{,}5$, por exemplo, o valor mediano da estatística KS dos métodos de MPV e CR são, respectivamente, 25,7% e 25% e do método *naïve* é 20,6%. No entanto, através dos gráficos *boxplot* da Figura 5.39 observamos que os valores da estatística KS apresentam comportamentos similares para todos os métodos de estimação. Assim como as medidas de desempenho citadas anteriormente, os valores da estatística KS diminuem conforme h aumenta.

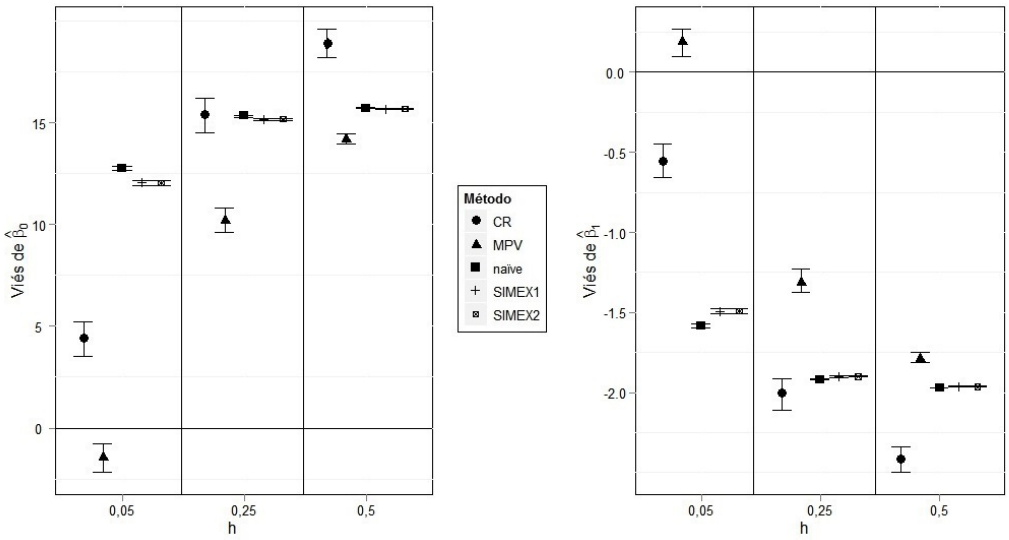

Figura 5.31: Intervalo de 95% de confiança para o viés dos estimadores de β_0 e β_1; cenário 6, $n_1 = 32$

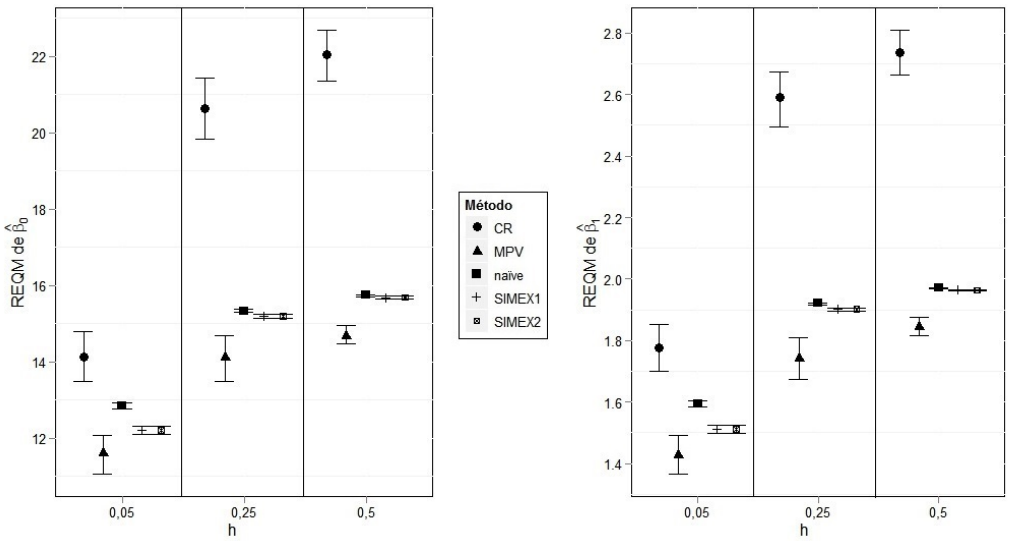

Figura 5.32: Intervalo de 95% de confiança para REQM dos estimadores de β_0 e β_1; cenário 6, $n_1 = 32$

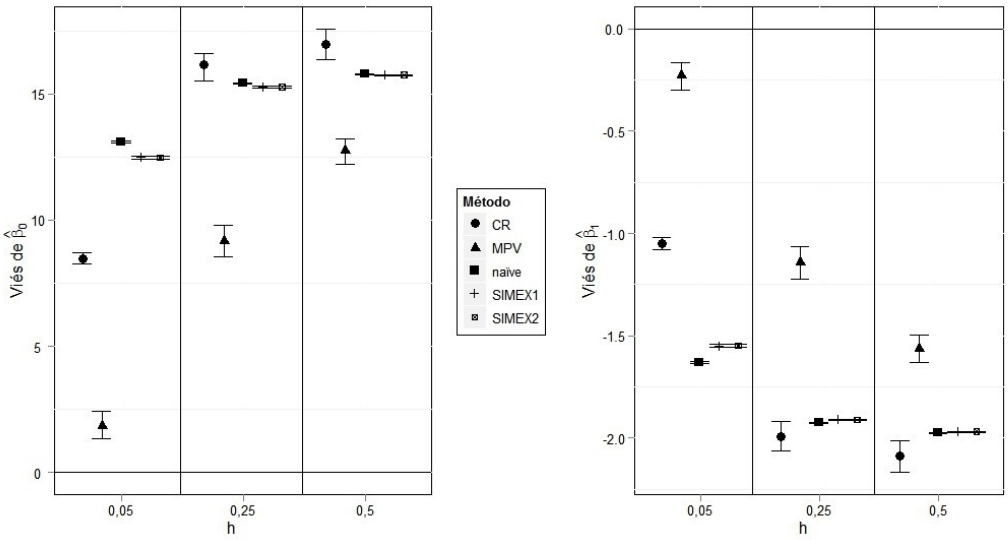

Figura 5.33: Intervalo de 95% de confiança para o viés dos estimadores de β_0 e β_1; cenário 6, $n_1 = 105$

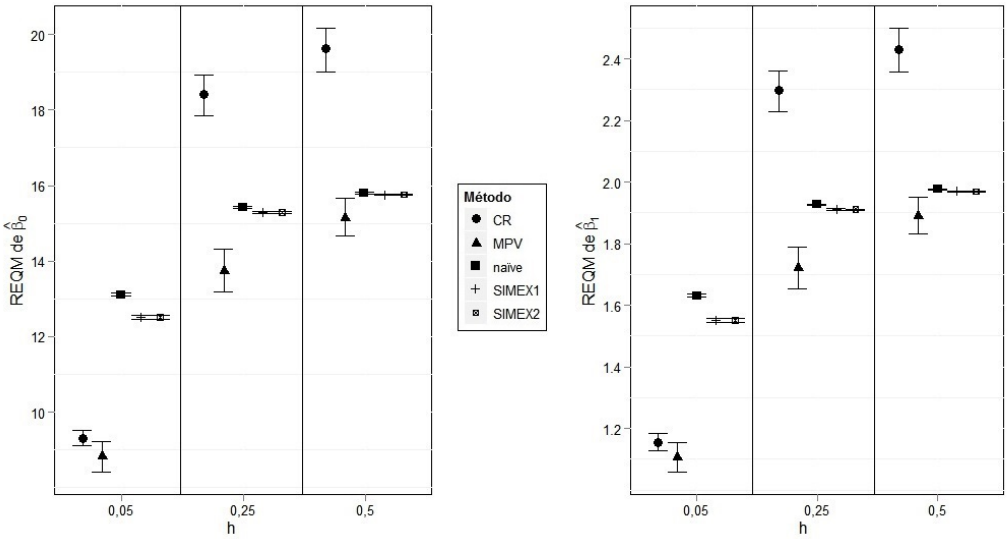

Figura 5.34: Intervalo de 95% de confiança para REQM dos estimadores de β_0 e β_1; cenário 6, $n_1 = 105$

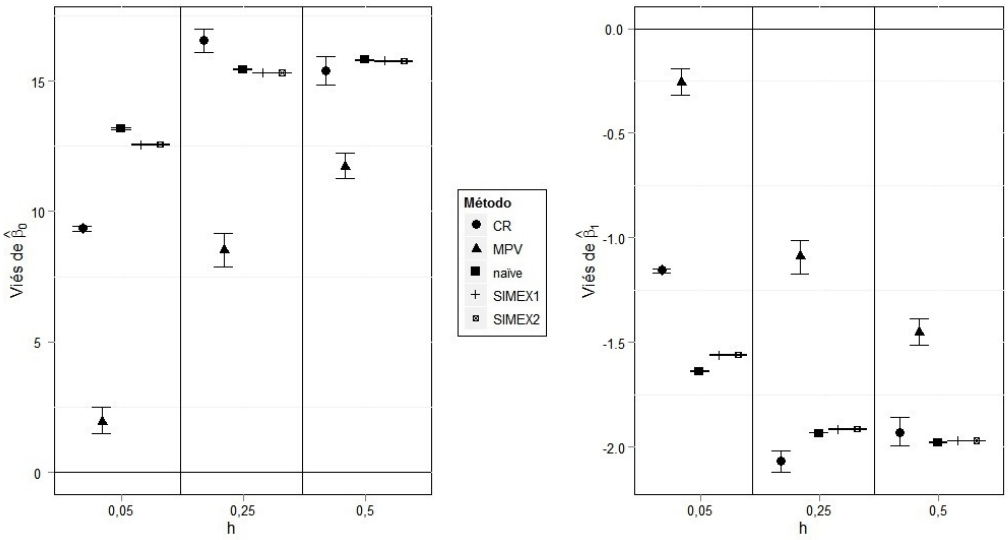

Figura 5.35: Intervalo de 95% de confiança para o viés dos estimadores de β_0 e β_1; cenário 6, $n_1 = 301$

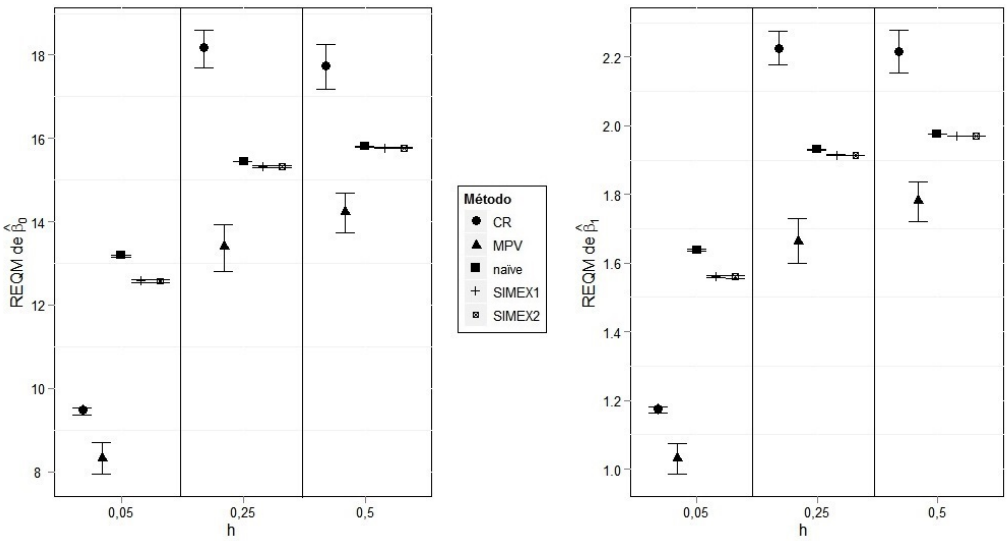

Figura 5.36: Intervalo de 95% de confiança para REQM dos estimadores de β_0 e β_1; cenário 6, $n_1 = 301$

Tabela 5.14: Mediana das medidas de desempenho na predição para o cenário 6

	Se			Es			VPP			VPN		
	$n_2 = 13$											
h	0,05	0,25	0,5	0,05	0,25	0,5	0,05	0,25	0,5	0,05	0,25	0,5
MPV	75,0	60,0	57,1	80,0	71,4	71,4	80,0	66,6	66,6	75,0	63,6	60,0
CR	75,0	50,0	50,0	77,7	66,6	66,6	77,7	60,0	58,3	75,0	57,1	55,5
SIMEX1	75,0	60,0	55,5	80,0	71,4	71,4	80,0	66,6	66,6	75,0	63,6	60,0
SIMEX2	75,0	60,0	55,5	80,0	71,4	71,4	80,0	66,6	66,6	75,0	63,6	60,0
naïve	75,0	60,0	55,5	80,0	71,4	71,4	80,0	66,6	66,6	75,0	63,6	60,0
	$n_2 = 45$											
h	0,05	0,25	0,5	0,05	0,25	0,5	0,05	0,25	0,5	0,05	0,25	0,5
MPV	72,7	59,0	53,3	76,1	70,0	68,1	75,0	66,6	62,5	73,9	62,9	59,0
CR	72,7	54,1	47,3	76,1	66,6	65,2	75,0	60,8	57,1	73,9	58,6	54,5
SIMEX1	72,7	59,0	52,3	76,1	69,5	68,1	75,0	66,6	62,5	73,9	62,9	58,6
SIMEX2	72,7	59,0	52,3	76,1	70,0	68,0	75,0	66,6	62,0	73,7	62,9	58,6
naïve	72,7	59,0	52,6	76,1	69,5	68,1	75,0	66,6	61,9	73,9	62,9	58,3
	$n_2 = 129$											
h	0,05	0,25	0,5	0,05	0,25	0,5	0,05	0,25	0,5	0,05	0,25	0,5
MPV	72,1	58,0	52,8	75,3	67,6	66,6	74,5	64,6	61,5	72,7	61,6	58,1
CR	72,3	56,2	47,1	75,3	66,6	65,2	74,4	63,0	57,4	72,9	60,0	54,5
SIMEX1	72,3	57,6	52,3	75,0	67,6	66,6	74,3	64,3	61,4	72,8	61,4	57,9
SIMEX2	72,3	57,8	52,3	75,0	67,7	66,6	74,3	64,4	61,5	72,8	61,4	57,9
naïve	72,2	57,6	52,3	75,3	67,6	66,1	74,2	64,4	61,4	72,7	61,4	57,8

Tabela 5.15: Continuação: Mediana das medidas de desempenho na predição para o cenário 6

	ACC			KS		
	$n_2 = 13$					
h	0,05	0,25	0,5	0,05	0,25	0,5
MPV	76,9	61,5	61,5	55,0	40,0	35,7
CR	76,9	61,5	53,8	57,1	40,0	35,0
SIMEX1	76,9	61,5	61,5	57,1	37,5	30,0
SIMEX2	76,9	61,5	61,5	60,0	37,7	30,0
naïve	76,9	61,5	61,5	57,1	37,5	28,5
	$n_2 = 45$					
h	0,05	0,25	0,5	0,05	0,25	0,5
MPV	73,3	64,4	60,0	51,6	31,5	25,7
CR	73,3	60,0	55,5	51,9	31,5	25,0
SIMEX1	73,3	64,4	60,0	51,1	29,1	21,5
SIMEX2	73,3	64,4	60,0	51,2	29,2	21,4
naïve	73,3	64,4	60,0	51,0	28,5	20,6
	$n_2 = 129$					
h	0,05	0,25	0,5	0,05	0,25	0,5
MPV	73,6	62,7	59,6	48,9	27,8	21,7
CR	73,6	62,0	56,5	48,8	28,0	21,7
SIMEX1	73,6	62,7	58,9	47,9	26,7	18,9
SIMEX2	73,6	62,7	59,6	47,9	26,7	18,9
naïve	73,6	62,7	58,9	47,6	25,8	17,9

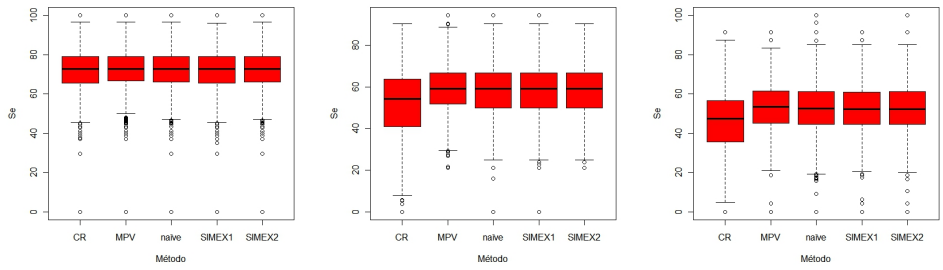

Figura 5.37: Gráficos *boxplot* dos valores de sensibilidade para $h = 0{,}05$, $0{,}25$ e $0{,}5$, respectivamente; cenário 6, $n_2 = 45$

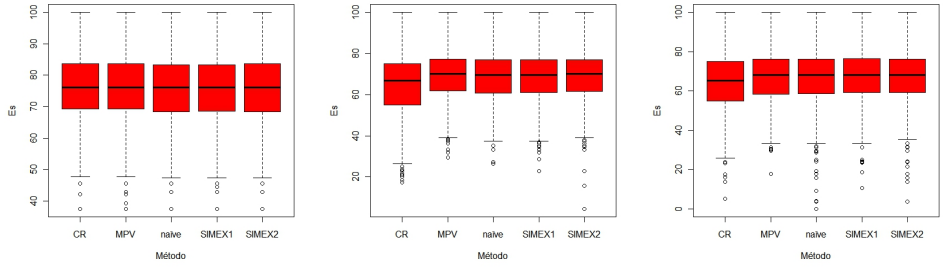

Figura 5.38: Gráficos *boxplot* dos valores de especificidade para $h = 0{,}05$, $0{,}25$ e $0{,}5$, respectivamente; cenário 6, $n_2 = 45$

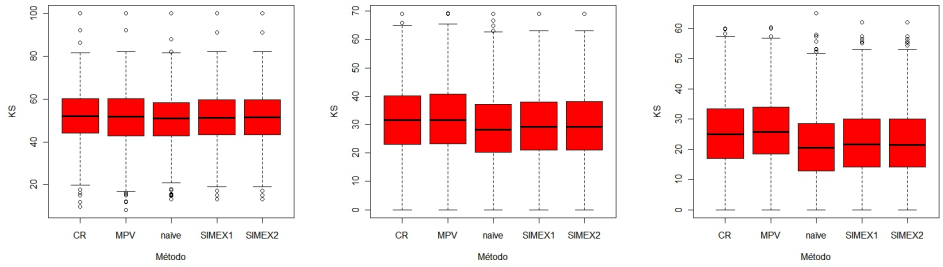

Figura 5.39: Gráficos *boxplot* dos valores da estatística KS para $h = 0{,}05$, $0{,}25$ e $0{,}5$, respectivamente; cenário 6, $n_2 = 45$

5.7 Considerações do estudo de simulação

Em geral, o estimador *naïve* é o mais viesado e com maiores valores de REQM. As duas versões do método SIMEX, com função de extrapolação linear, apresentam comportamentos na estimação muito similares e apresentam desempenho melhor, na maioria das vezes, apenas do que o método *naïve*. O método MPV obtém, na grande maioria das vezes, o melhor desempenho na estimação tanto em relação ao viés quanto em relação à REQM e o método da CR o segue, sendo o segundo método com melhor desempenho em geral. Mas valem algumas ressalvas, expostas a seguir.

O método de MPV piora o seu desempenho (principalmente na REQM) em relação aos demais métodos de estimação conforme diminui a proporção de eventos, ou seja, conforme aumenta o desbalanceamento da amostra. Com isso, o método da CR que apresenta, em geral, o segundo menor valor da REQM em amostras balanceadas, apresenta o menor valor de REQM em situações de desbalanceamento da amostra.

Na presença de heteroscedasticidade, de uma maneira geral, o método de MPV é o melhor método entre os comparados, mesmo não levando em conta a heteroscedasticidade na função de pseudoverossimilhança. O método da CR apresenta bom desempenho na presença de baixa heteroscedasticidade, em alguns casos sendo superior ao método de MPV em relação ao viés. Com relação à REQM, este método apresenta alto valor quando o tamanho amostral é pequeno ou quando o tamanho da amostra não é grande o suficiente em relação à variância do erro de medida, uma vez que para $h = 0{,}5$ o método da CR tem a maior REQM quando $n_1 = 105$ e apresenta o segundo menor valor de REQM quando $n_1 = 301$. Já quando a heteroscedasticidade é alta, o método da CR só apresenta bom desempenho em relação ao viés quando $h = 0{,}05$ e para a REQM quando o tamanho da amostra é grande ($n_1 = 105$ e $n_1 = 301$) e $h = 0{,}05$.

No estudo desenvolvido por Spiegelman *et al.* (2011), o método da CR usual é adequado mesmo na presença de heteroscedasticidade. Nas Tabelas 5.16 e 5.17 estão o viés e REQM dos estimadores de β_1 obtidos pelos métodos da CR e *naïve* sob a abordagem de dados de validação, com $n_1 = 301$ (tamanho da amostra de validação: 45), para $\rho = 0,5$ e $\rho = 2$, respectivamente. Como podemos observar, sob a abordagem de dados de validação, o método da CR é sempre melhor que o método *naïve* mesmo na presença de alta heteroscedasticidade, resultado diferente ao que encontramos no cenário 6 ($\rho = 2$), sob a abordagem de replicações.

Os valores de viés e REQM de todos os métodos aumentam conforme aumenta a heteroscedasticidade do erro de medida, principalmente na mudança de $\rho = 0{,}5$ para 2, fixados o método de estimação e valores de h e n_1.

No modelo aditivo clássico para o erro em modelo de regressão linear simples, $E(Y_i|X_i) = \beta_0 + \beta_1 X_i$, o viés causado pelo erro de medida é sempre na forma de atenuação. De fato, se $\widehat{\beta}_1$ representa o estimador de mínimos quadrados *naïve*, temos $\mathrm{E}(\widehat{\beta}_1) = (1 + \mathrm{h}^2)^{-1}\beta_1 < \beta_1$, em que $(1 + h^2)^{-1}$ é chamado de fator de atenuação. Em modelos de regressão logística, não há uma forma fechada para o viés induzido pelo erro de medida (Gustafson, 2004, p. 24) e diferentemente do modelo de regressão linear, o fator de atenuação não é o mesmo para todos os valores de β_1. No entanto, Gustafson (2004, p. 25) comparou as curvas do fator de atenuação em função de h e verificou que essas têm forte semelhança com as de regressão linear, apoiando a ideia de que o impacto do erro de medida em modelos de regressão logística é muito semelhante ao seu impacto em modelos de regressão linear. Em geral, obtemos neste trabalho resultados compatíveis com os relatados na literatura, inclusive o fato de o viés aumentar com o aumento da variância do erro de medida, como esperado.

Na predição de novas observações, não observamos diferenças importantes nas medidas de desempenho sensibilidade, especificidade, VPP, VPN e acurácia entre os métodos de estimação sob a presença de heteroscedasticidade e com diferentes proporções de eventos. A estatística KS apresenta diferença em seus valores medianos para alguns casos. São eles: proporção de eventos de interesse baixa, onde o método da CR apresenta maiores valores da estatística, e na presença de alta heteroscedasticidade, onde os valores medianos da estatística são maiores para os métodos de MPV e CR. No entanto, os gráficos *boxplot* dos valores da estatística são similares para todos os métodos de estimação.

Ao comparamos os resultados obtidos na predição dos cenários 2, 3 e 4, podemos observar a relação entre o comportamento das medidas de desempenho e o desbalanceamento da amostra, fixos o método de estimação e valores de h e n_2. Observamos que a medida VPP diminui conforme aumenta o desbalanceamento da amostra e a medida VPN aumenta. Mas o comportamento destas medidas deve ser avaliado com cautela, uma vez que elas sofrem o efeito da proporção de eventos, como citado por Louzada-Neto *et al.* (2009).

Ao comparamos os resultados obtidos na predição dos cenários 2, 5 e 6, podemos

observar o comportamento das medidas de desempenho na presença de erro de medida heteroscedástico. Assim, observamos que os valores obtidos da sensibilidade, especificidade, VPP, VPN e acurácia diminuem conforme aumenta a heteroscedasticidade do erro de medida, principalmente na mudança de $\rho = 0{,}5$ (cenário 5) para 2 (cenário 6).

Tabela 5.16: Viés e REQM [intervalo de 95% de confiança] dos estimadores de β_1 com $\rho = 0,5$ sob a abordagem de dados de validação.

Método	$h = 0{,}05$	$h = 0{,}25$	$h = 0{,}5$
	Viés		
CR	0,050	0,039	0,052
	[0,032; 0,067]	[0,022; 0,057]	[0,034; 0,070]
naïve	−0,509	−0,950	−1,225
	[−0,522; −0,497]	[−0,959; −0,942]	[−1,231; −1,218]
	REQM		
CR	0,282	0,282	0,295
	[0,266; 0,297]	[0,266; 0,298]	[0,279; 0,311]
naïve	0,547	0,961	1,229
	[0,536; 0,558]	[0,953; 0,969]	[1,223; 1,236]

Tabela 5.17: Viés e REQM [intervalo de 95% de confiança] dos estimadores de β_1 com $\rho = 2$ sob a abordagem de dados de validação.

Método	$h = 0{,}05$	$h = 0{,}25$	$h = 0{,}5$
	Viés		
CR	0,273	0,414	0,433
	[0,251; 0,295]	[0,388; 0,440]	[0,407; 0,459]
naïve	$-1{,}261$	$-1{,}618$	$-1{,}762$
	$[-1{,}267; -1{,}255]$	$[-1{,}621; -1{,}615]$	$[-1{,}763; -1{,}760]$
	REQM		
CR	0,446	0,585	0,600
	[0,422; 0,468]	[0,557; 0,612]	[0,570; 0,629]
naïve	1,264	1,619	1,762
	[1,259; 1,270]	[1,616; 1,622]	[1,760; 1,764]

Capítulo 6

Dados reais I - Aplicação na área médica

A fim de aplicar os métodos estudados em dados reais da área médica, consideramos os dados do estudo *Framinghanm Heart Study* (Carroll *et al.*, 1984). O interesse do estudo é verificar o impacto de fatores de risco em doenças coronarianas em homens americanos. As doenças coronarianas se desenvolvem em virtude do depósito de substâncias gordurosas nas paredes das artérias, restringindo o fluxo sanguíneo e são as causas de infarto (ou ataque cardíaco), uma das doenças de maior mortalidade no mundo, segundo a Organização Mundial da Saúde (`http://www.who.int`).

Desta forma, utilizamos o modelo de regressão logística para verificar o quanto os fatores de risco influenciam na probabilidade de surgimento de doenças coronarianas em homens. A variável resposta, Y_i, indica a presença ou ausência de doenças coronarianas (chamamos de chd) para o i-ésimo homem e as variáveis preditoras (fatores de risco) são idade do paciente (em anos), se fumante ou não, nível de colesterol (em mg/dL) e pressão sistólica (em mmHg). O estudo consiste na observação de 1615 homens de 31 a 65 anos de idade. Se as variáveis preditoras fossem todas observadas de maneira exata, poderíamos utilizar o modelo de regressão logística usual (Hosmer & Lemeshow, 2000). Mas a pressão sistólica verdadeira não pode ser medida exatamente, pois varia ao longo do dia para o mesmo indivíduo. Assim como fazem Carroll *et al.* (2006, p. 112), analisamos os dados considerando a pressão sistólica como variável com erro e as demais variáveis como covariáveis observadas sem erro, Z_i.

A variável pressão sistólica observada (chamamos de sbp) foi medida em dois exames em momentos distintos. Os valores observados nesses dois exames são considerados réplicas.

Carroll *et al.* (1984) sugerem uma transformação na variável com erro e utilizamos a transformação considerada por Carroll *et al.* (2006, p. 113): $W_{ji} = \log(sbp_{ji} - 50)$, $j = 1, 2$, que chamamos de $lsbp_{ji}$. Ainda, seja $mlsbp_i$ a média das duas réplicas do i-ésimo indivíduo da amostra, com $i = 1, \ldots, 1615$. No decorrer do texto, chamamos de *lsbp* a variável observada com erro transformada e *mlsbp* a variável oriunda da média das réplicas de *lsbp*.

Na Tabela 6.1 apresentamos as medidas descritivas da amostra em questão. Observamos que 8% dos homens amostrados apresentam doenças coronarianas, 77% são fumantes, eles têm quase 46 anos em média e nível de colesterol médio de 228 mg/dL. Também podemos observar que, em média, os valores de *lsbp* nas duas réplicas são próximos e próximos da média de *mlsbp*. Vale notar que os valores médios das covariáveis e a proporção de fumantes são maiores no grupo com doenças coronarianas do que no grupo sem doenças.

Tabela 6.1: Média (desvio padrão) ou *proporção de doentes ou **proporção de fumantes geral e por status de chd

Variável	geral	sem chd	com chd
chd	0,079*	-	-
fumante	0,773**	0,768**	0,828**
idade	45,861	45,493	50,133
	(8,588)	(8,566)	(7,669)
colesterol	228,404	227,202	242,367
	(41,452)	(41,244)	(41,470)
1ª réplica de *lsbp*	4,374	4,365	4,476
	(0,226)	(0,222)	(0,237)
2ª réplica de *lsbp*	4,355	4,345	4,464
	(0,229)	(0,226)	(0,237)
mlsbp	4,364	4,355	4,469
	(0,213)	(0,210)	(0,223)

O modelo para o erro de medida considerado é $lsbp_{ji} = X_i + U_{ji}$, em que X_i é a pressão sistólica verdadeira transformada do i-ésimo indivíduo. A estimativa de σ_u^2 é obtida por (4.1), resultando em $\widehat{\sigma}_u^2 = 0{,}012$. Ainda, temos que $mlsbp_i = X_i + \overline{U}_i$, em que $\overline{U}_i = \sum_{j=1}^{2} U_{ji}/2$ e a estimativa de σ_x^2 é $\widehat{\sigma}_x^2 = \widehat{\sigma}_{mlsbp}^2 - \widehat{\sigma}_u^2/2$, em que $\widehat{\sigma}_{mlsbp}^2 = 0{,}045$. Desta

maneira, temos que a estimativa de h é 0,138.

O valor de M e o método de otimização do algoritmo EM-Monte Carlo utilizados para a maximização da função de pseudoverossimilhança são os mesmos dos estudos de simulação apresentados no Capítulo 5. De forma análoga ao apresentado na Seção 3.3 para a função de verossimilhança, covariáveis observadas sem erro de medida, Z_i, foram incluídas na função de pseudoverossimilhança. Assumimos que a distribuição de X_i dado a observação z_i de Z_i é normal com média $\mu_{xi} = \alpha_0 + \alpha_1 z_{1i} + \alpha_2 z_{2i} + \alpha_3 z_{3i}$, com $\alpha = (\alpha_0, \alpha_1, \alpha_2, \alpha_3)^\top$, e variância σ_x^2. A estimativa de α é obtida pelo método de mínimos quadrados da regressão linear de Z_i em $mlsbp_i$, com $i = 1, \ldots, n$.

Utilizamos duas versões para o método SIMEX. Na primeira (SIMEX1) utilizamos a estimativa da variância do erro de medida como sendo o seu valor verdadeiro, embasando-nos no fato de existirem 1614 graus de liberdade para estimar a variância do erro de medida e assim, para propósitos práticos, a variância do erro de medida é considerada conhecida (Carroll *et al.*, 2006, p. 115). Neste caso, substituimos W_i da equação (4.8) por $mlsbp_i$. A segunda versão (SIMEX2) é o SIMEX empírico. Para cada versão do método SIMEX, consideramos três funções de extrapolação: linear (lin), quadrática (quad) e extrapolação linear racional (lr). Utilizamos $D = 2000$ pseudoamostras e $\zeta_b = \{0, 0{,}28, 0{,}57, 0{,}85, 1{,}14, 1{,}43, 1{,}71, 2\}$.

Na Tabela 6.2 estão as estimativas dos parâmetros associados a cada covariável do modelo calculadas pelos diferentes métodos de estimação. Entre parênteses estão os erros padrão estimados obtidos a partir da teoria apresentada no Apêndice A. Os valores dos erros padrão das estimativas dos parâmetros calculados pelos métodos de estimação diferem, em geral, apenas na segunda casa decimal. As estimativas dos parâmetros associados às covariáveis sem erro de medida variam, na maioria dos casos, apenas na terceira casa decimal. Já as estimativas de β_1, parâmetro relacionado à covariável com erro de medida, variam de 1,706 (estimativa obtida pelo método *naïve*) a 2,017 (estimativa obtida pelo método MPV).

As estimativas calculadas pelos métodos da CR e MPV para β_1 são bem próximas e se nos basearmos nos estudos de simulação apresentados no Capítulo 5, onde o MPV teve, em geral, o melhor desempenho na estimação e o método da CR o acompanhou na maioria das situações, as estimativas desses dois métodos parecem ser as mais confiáveis. A função de extrapolação linear racional é a função que apresentou estimativas para β_1 mais próximas

dos valores obtidos pelos métodos de MPV e CR, para as duas versões do método SIMEX.

Na Tabela 6.3 apresentamos as estimativas pontuais e intervalos de 95% de confiança para a razão de chances entre indivíduos que diferem na covariável *lsbp* em uma unidade, fixadas as demais covariáveis, obtidas por (2.4) e (2.5) ao considerar $c = 1$. Para todos os métodos de estimação, o valor 1 não está contido no intervalo de confiança para a razão de chances, mas observamos diferenças nas estimativas. A estimativa da razão de chances obtida pelo método *naïve* é 5,507 e as estimativas obtidas pelos métodos que fazem correção do erro de medida variam entre 6,459 e 7,522, este último é a estimativa da razão de chances obtida pelo método de MPV. Neste caso, ignorar o erro de medida não altera o sentido da relação entre as variáveis: quanto maior *lsbp*, maior a chance de ter doenças coronarianas, dadas as demais covariáveis fixas. No entanto, o método *naïve* subestima a razão de chances entre indivíduos que diferem na covariável *lsbp*.

Um resultado similar foi observado por Rosner *et al.* (1990) nos dados do estudo *Nurses' Health Study*. O objetivo do estudo era verificar se as covariáveis idade, gordura consumida, total de calorias consumidas e consumo de álcool influenciam na presença de câncer de mama. As três últimas covariáveis citadas foram medidas com erro e por isso foi considerado um modelo de regressão logística com múltiplas variáveis com erro de medida. Eles observam que para a variável consumo de álcool, a razão de chances estimada é 1,33 para o método sem correção (*naïve*) e para o método que faz correção do erro de medida a razão de chances é 1,62. Assim, como observamos nos dados estudados neste Capítulo, a razão de chances é subestimada ao ignorarmos o erro de medida.

Tabela 6.2: Estimativas dos parâmetros (erro padrão) do modelo de regressão logística

Métodos	Variáveis				
	intercepto	*lsbp*	idade	fumante	colesterol
naïve	−14,949	1,706	0,055	0,593	0,007
	(1,899)	(0,417)	(0,011)	(0,25)	(0,002)
MPV	−16,219	2,017	0,053	0,602	0,007
	(2,024)	(0,467)	(0,010)	(0,244)	(0,002)
CR	−16,172	2,010	0,053	0,601	0,007
	(2,006)	(0,459)	(0,011)	(0,249)	(0,002)
SIMEX1 lin	−15,586	1,865	0,054	0,597	0,007
	(2,013)	(0,448)	(0,011)	(0,250)	(0,002)
SIMEX1 quad	−15,796	1,918	0,053	0,599	0,007
	(2,091)	(0,469)	(0,011)	(0,250)	(0,002)
SIMEX1 lr	−15,836	1,928	0,053	0,601	0,007
	(2,125)	(0,478)	(0,011)	(0,250)	(0,002)
SIMEX2 lin	−15,608	1,869	0,054	0,598	0,007
	(2,015)	(0,449)	(0,011)	(0,250)	(0,002)
SIMEX2 quad	−15,859	1,933	0,053	0,600	0,007
	(2,082)	(0,467)	(0,011)	(0,250)	(0,002)
SIMEX2 lr	−15,913	1,946	0,053	0,601	0,007
	(2,107)	(0,474)	(0,012)	(0,250)	(0,002)

Tabela 6.3: Estimativas das razões de chances [IC de 95% de confiança] entre indivíduos que diferem na variável *lsbp* em uma unidade

Métodos	Razões de chances
naïve	5,507 [2,430; 12,481]
MPV	7,522 [3,008; 18,808]
CR	7,470 [3,034; 18,392]
SIMEX1 lin	6,459 [2,679; 15,570]
SIMEX1 quad	6,810 [2,712; 17,099]
SIMEX1 lr	6,879 [2,693; 17,578]
SIMEX2 lin	6,482 [2,688; 15,631]
SIMEX2 quad	6,911 [2,765; 17,274]
SIMEX2 lr	7,007 [2,767; 17,745]

Capítulo 7

Dados reais II - Aplicação na área financeira

Em instituições financeiras, a classificação adequada dos clientes é de vital importância para determinar a concessão de crédito. A concessão de crédito ganhou espaço nas empresas do setor financeiro, tornando-se uma das principais fontes de receita e, rapidamente, este setor percebeu a necessidade de aumentar o volume de recursos concedidos sem perder agilidade e qualidade. Nesse ponto, a contribuição da modelagem estatística foi essencial e os modelos estatísticos passaram a ser um importante instrumento para auxiliar os gestores na tomada de decisões. Para mais detalhes e histórico bibliográfico, sugerimos a leitura de Diniz & Louzada (2012).

Uma covariável importante para estes modelos é a renda do cliente. Se o cliente não pertence ao portfólio da instituição, é possível que a informação de sua renda não seja disponível. Além disso, a renda não é uma variável de fácil acesso e pode ser desconhecida mesmo o cliente fazendo parte do portfólio. O que se faz na prática é utilizar a renda presumida como covariável no modelo de interesse. Ainda na prática, o modelo é ajustado de forma usual, sem levar em conta que a renda presumida é uma covariável medida com erro.

Neste contexto, uma instituição financeira brasileira tem o interesse em um modelo de classificação dos clientes como inadimplentes ou adimplentes em que a renda é uma covariável medida com erro. A informação da renda é obtida por um modelo. Nesse modelo, que chamamos de modelos de renda presumida, era disponível a renda verdadeira do j-ésimo

indivíduo, y_j^*, com $j = 1, \ldots, n^*$, em que n^* é o número de indivíduos para os quais se tinha a informação da renda. Um modelo estatístico foi utilizado para estimar a renda em função de p^* covariáveis, sendo que x_j^* denota o vetor coluna das covariáveis do j-ésimo indivíduo cujo primeiro elemento vale 1 para permitir a inclusão do intercepto não nulo no modelo. Ainda, $g(\mu^*) = x_j^{*\top}\beta^*$, em que $g(.)$ é a função de ligação utilizada, que não conhecemos por questão de sigilo da instituição financeira, e β^* é o vetor de parâmetros do modelo de interesse. A estimativa da variância do erro, chamamos de $\widehat{\sigma}_e^2$, é dado por

$$\widehat{\sigma}_e^2 = \frac{\sum_{j=1}^{n*} (y_j^* - \hat{y}_j^*)^2}{n^*},$$

em que $\hat{y}_j^* = g^{-1}(x_j^{*\top}\widehat{\beta}^*)$ é o valor predito de y_j^* e desta maneira, $e_j = y_j^* - \hat{y}_j^*$ é o resíduo. O valor da estimativa $\widehat{\sigma}_e^2$ é a única informação disponível pela instituição financeira sobre o modelo de renda presumida.

No modelo de classificação, a informação de renda presumida dos n clientes do banco de dados é $W_i = g^{-1}(x_i^{*\top}\widehat{\beta}^*)$, $i = 1, \ldots, n$, em que $\widehat{\beta}^*$ é a estimativa dos parâmetros do modelo de renda presumida. Como discutido anteriormente, nos modelos de erro de medida, é necessário que alguma informação seja disponível sobre a variância do erro de medida, seja ela conhecida, estimada por réplicas ou dados de validação. No problema em questão, estamos considerando um modelo aditivo clássico para o erro de medida, ou seja, $W_i = X_i + U_i$. Assim, $U_i = W_i - X_i$, em que X_i é a renda verdadeira e W_i é a renda presumida. Como pode ser visto, tanto e_i quanto U_i representam o desvio da renda estimada em relação à renda verdadeira. Desta forma, consideramos que a estimativa da variância do erro de medida é a estimativa da variância do resíduo do modelo para renda, isto é $\sigma_u^2 = \widehat{\sigma}_e^2$.

Além da covariável medida com erro, renda presumida (em reais), consideramos duas covariável observadas sem erro: tempo desde a última consulta e escore de risco genérico. A covariável tempo desde a última consulta, que chamamos de tempo, é o tempo, em dias, desde a última consulta realizada por empresas de concessão de crédito, ou seja, o tempo desde que o cliente tentou realizar alguma compra e a empresa verificou se ele estava com pagamentos pendentes. O escore de risco genérico é um escore de risco de inadimplência do cliente cadastrado na instituição financeira que leva em conta seu histórico financeiro, assumindo valores de 262 a 984.

A amostra completa é composta por $n = 146551$ indivíduos, dos quais separamos 70% das observações em amostra treinamento completa ($n_1 = 102586$) e os 30% restantes em

amostra teste completa ($n_2 = 43965$). A prevalência, proporção de clientes inadimplentes na amostra, é de 8%.

Consideramos a técnica de seleção dependente de *status* (Louzada *et al.*, 2012). A seleção dependente de *status* (*state-dependent*) é uma técnica de modelagem estatística usada nos casos em que a amostra considerada para o ajuste do modelo, amostra completa, contém baixa proporção de eventos, que neste caso é o cliente ser inadimplente. A ideia da técnica é manter os clientes inadimplentes na amostra e selecionar aleatoriamente uma parcela de clientes adimplentes.

No problema em questão, Y_i é a variável resposta indicadora do *status* do cliente, em que $Y_i = 1$ se inadimplente e $Y_i = 0$ indica que o cliente é adimplente. A amostra completa é uma amostra aleatória com fração α da população e apenas uma fração γ de observações com $Y_i = 0$ é mantida. A probabilidade do i-ésimo cliente ter $Y_i = 1$ e ser incluído na amostra é $\alpha\pi_i$ e a probabilidade de ter $Y_i = 0$ e ser amostrado é $\gamma\alpha(1 - \pi_i)$. Então, se o evento $I_i = 1$ significa que o i-ésimo indivíduo é incluído na amostra, pelo Teorema de Bayes, é possível concluir que

$$\begin{aligned}\tilde{\pi}_i = P(Y_i = 1|I_i = 1) &= \frac{P(Y_i = 1, I_i = 1)}{P(Y_i = 1, I_i = 1) + P(Y_i = 0, I_i = 1)} \\ &= \frac{\alpha\pi_i}{\alpha\pi_i + \gamma\alpha(1 - \pi_i)} = \frac{\pi_i}{\pi_i + \gamma(1 - \pi_i)}.\end{aligned}$$

Com π_i dado por (2.1), $\tilde{\pi}_i$ fica

$$\tilde{\pi}_i = \frac{\exp(x_i^\top\beta)}{\exp(x_i^\top\beta) + \gamma} = \frac{1/\gamma\exp(x_i^\top\beta)}{1 + 1/\gamma\exp(x_i^\top\beta)} = \frac{\exp(x_i^\top\beta - \log(\gamma))}{1 + \exp(x_i^\top\beta - \log(\gamma))}. \tag{7.1}$$

Desta forma, $\tilde{\pi}_i$ é a probabilidade de sucesso corrigida pela seleção dependente de *status*. Como desejamos uma amostra balanceada, selecionamos o mesmo número de clientes adimplentes, resultando $\gamma = 0,08$. A estimativa da variância do erro de medida é $\widehat{\sigma}_u^2 = 22500$ e a estimativa de h é 0,073, o que indica baixa variância devido ao erro de medida. Consideramos os métodos de estimação *naïve*, MPV, CR e SIMEX (versão com variância do erro de medida conhecida) sob duas abordagens: considerando a correção da probabilidade de sucesso dada em (7.1) e não considerando-a, que chamamos de abordagem usual. O valor de M, o método de otimização do algoritmo EM-Monte Carlo, utilizados para a maximização da função de pseudoverossimilhança, os valores de D e ζ_b e a função de extrapolação para o método SIMEX são os mesmos dos estudos de simulação apresentados no Capítulo 5.

Nas Tabelas 7.1 e 7.2 estão as medidas de desempenho sensibilidade, especificidade, VPP, VPN, acurácia e estatística KS obtidas pelos métodos de estimação na abordagem usual e considerando a correção dependente de *status*, respectivamente. Observamos que não há diferença entre os métodos nas medidas de desempenho na predição nas duas abordagens. Ainda, fixado qualquer método de estimação e com exceção da estatística KS, as medidas de desempenho obtidas pelas duas abordagens são similares, como foi observado por Louzada *et al.* (2012) na situação sem erro de medida. Já a estatística KS diminui cerca de 17% para todos os métodos de estimação ao considerar a correção dependente de *status*.

Vale ressaltar que utilizamos a adaptação da estatística KS que considera as categorias de escores (Seção 2.4.1). O menor valor da estatística KS obtido pela correção dependente de *status* se deve ao fato de que o menor escore estimado é 600, independente do método de estimação, e por isso, as frequências de adimplentes e de inadimplentes estão concentradas em apenas quatro categorias (600 a 700, 700 a 800, 800 a 900 e maior que 900). Já ao utilizarmos a estatística sem a adaptação por categorias de escores, o valor obtido para as duas abordagens é 31%.

Na Figura 7.1 estão os gráficos das probabilidades de sucesso estimadas ordenadas obtidas pelos métodos de estimação usando a correção dependente de *status* (gráfico à esquerda) e na abordagem usual (gráfico à direita). É possível observar que as curvas das probabilidades estimadas obtidas pelos métodos de estimação se sobrepõem nas duas abordagens consideradas.

Apesar de não existir diferença entre os métodos de estimação nas curvas das probabilidades de sucesso estimadas ordenadas, observamos que, fixado o método de estimação, as curvas das probabilidades de sucesso estimadas das duas abordagens são diferentes, uma vez que a abordagem usual subestima a probabilidade de sucesso, como pode ser visto na Figura 7.2. No entanto, em ambas as abordagens não há mudança na posição das probabilidades de sucesso estimadas em relação ao respectivo ponto de corte (ponto de corte usando a correção dependente de *status* é 0,86 e sem a correção é 0,45), de modo que não influencia as medidas de desempenho, uma vez que são iguais nas duas abordagens (Tabelas 7.1 e 7.2).

Tabela 7.1: Medidas de desempenho na predição sem corrigir pela seleção dependente de *status*

Medidas	MPV	CR	SIMEX	*naïve*
Se	68,7	68,7	68,7	68,7
Es	62,1	62,0	62,0	62,0
VPP	64,4	64,4	64,4	64,4
VPN	66,5	66,5	66,5	66,5
ACC	65,4	65,4	65,4	65,4
KS	30,6	30,6	30,6	30,6

Tabela 7.2: Medidas de desempenho na predição com correção dependente de *status*

Medidas	MPV	CR	SIMEX	*naïve*
Se	68,7	68,7	68,7	68,7
Es	62,1	62,0	62,0	62,0
VPP	64,4	64,4	64,4	64,4
VPN	66,5	66,5	66,5	66,5
ACC	65,4	65,4	65,4	65,4
KS	25,4	25,5	25,5	25,5

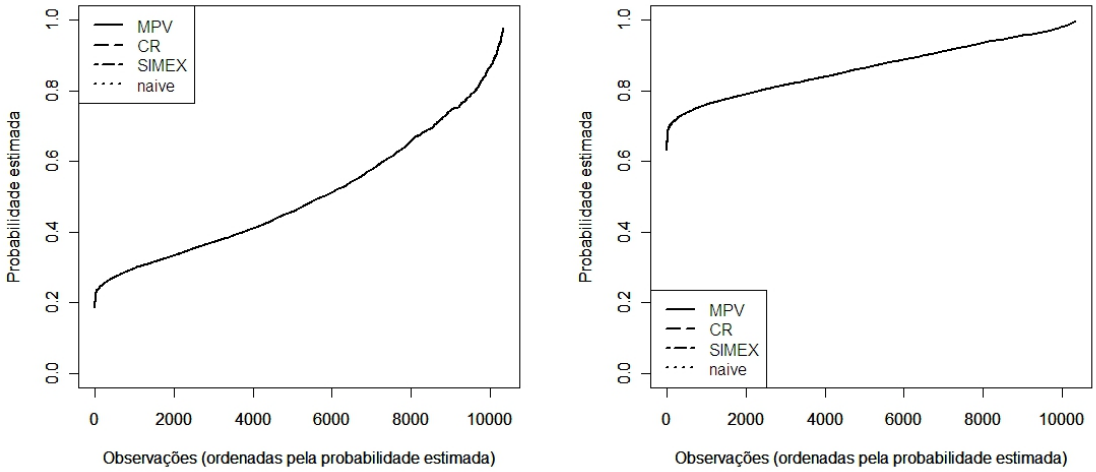

Figura 7.1: Gráfico das probabilidades de sucesso estimadas ordenadas dos métodos de estimação sem considerar correção dependente de *status* (à esquerda) e considerando a correção (à direita)

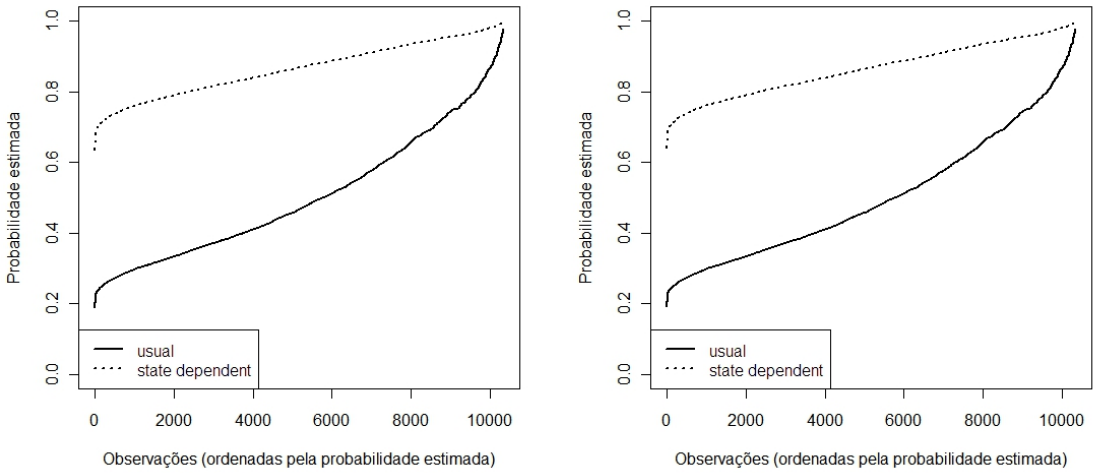

Figura 7.2: Gráfico das probabilidades de sucesso estimadas ordenadas do método de MPV (à esquerda) e do método *naïve* (à direita) considerando e não considerando a correção dependente de *status*

7.1 Discussão

Como observado, não encontramos diferença entre os métodos de estimação nas medidas de desempenho na predição, fato que, em geral, já havíamos observado no estudo de simulação do Capítulo 5. Entretanto, existe diferença entre os métodos na estimação dos parâmetros de interesse, também observado no estudo de simulação. Como a probabilidade de sucesso é uma função desses parâmetros, podemos entender que, apesar de os métodos que levam em conta o erro de medida não apresentarem diferença em relação às medidas de desempenho na predição quando comparados ao método *naïve*, existe diferença nas probabilidades de sucesso estimadas, fato que não foi observado nesta aplicação, vide gráficos da Figura 7.1. Por este motivo, levantamos aqui algumas hipóteses e alternativas para o ajuste dos dados sob a metodologia em questão.

Há disponível sobre o modelo de renda presumida apenas a estimativa da variância do erro, o que nos levou ao modelo proposto. Após a apresentação e discussão dos resultados obtidos com a instituição financeira, temos a proposta de estudar mais profundamente esses dados e verificar se outro modelo para o erro de medida é mais correto. Para isso, seria necessário ter disponíveis mais informações sobre o modelo para renda presumida e, se possível, ter informação sobre a renda verdadeira para alguns clientes da instituição.

Como o intuito do trabalho é comparar métodos de estimação na estimação e predição de novas observações, não entrando na questão de testes de hipóteses, as covariáveis consideradas no modelo foram pré-definidas e fixas para todos os métodos de estimação. Uma alternativa é realizar testes de hipóteses para a seleção das variáveis para cada método de estimação. Uma covariável pode ser significativa ao considerar um método de estimação, mas pode não ser ao considerar outro. Carrasco *et al.* (2013) observaram em uma aplicação do modelo de regressão beta com erro na covariável que um parâmetro é significativo a um nível de significância de 5% para alguns métodos de estimação e não para outros. Desta maneira, a proposta é selecionar as covariáveis do modelo para cada método de estimação e após isso, realizar o estudo de desempenho na predição.

Capítulo 8

Considerações Finais

Neste trabalho estudamos o modelo de regressão logística com erro de medida na covariável através do estudo teórico de métodos de estimação, estudo de simulação e duas aplicações em dados reais, uma da área médica e outra da área financeira. Apresentamos a seguir as principais considerações.

- No Capítulo 2 revisamos o modelo de regressão logística sem erro de medida e no Capítulo 3 apresentamos os principais conceitos do modelo com erro de medida nas covariáveis.

- No Capítulo 4 apresentamos os métodos de estimação *naïve*, máxima verossimilhança e máxima pseudoverossimilhança, calibração da regressão e SIMEX. Para o último método apresentamos as duas versões consideradas neste trabalho: com variância do erro de medida considerada conhecida e a versão cuja variância do erro de medida é estimada por réplicas da variável observada com erro. Ainda neste capítulo, apresentamos o algoritmo EM-Monte Carlo, utilizado para a maximização das funções de verossimilhança e de pseudoverossimilhança.

- No Capítulo 5 realizamos um estudo de simulação com o intuito de avaliar os métodos de estimação em relação à estimação e à predição de novas observações. Os resultados evidenciaram superioridade do método de máxima pseudoverossimilhança na estimação, seguido pelo método da calibração da regressão. Mas vale notar a piora no desempenho do método de máxima pseudoverossimilhança nos valores da REQM em relação aos demais métodos conforme a proporção dos eventos de interesse dimi-

nui. Em geral, o método SIMEX, com função de extrapolação linear, só apresentou melhores resultados que o método *naïve*, método com pior desempenho na estimação. Na predição, não houve diferença entre os métodos nas medidas de desempenho e estatística KS. Uma vez que há diferença entre os métodos na estimação dos parâmetros do modelo e a probabilidade de sucesso é uma função desses parâmetros, parece que ignorar o erro de medida não influencia na classificação de novas observações, mas influencia na estimação da probabilidade de sucesso.

- No Capítulo 6 consideramos a aplicação da metodologia estudada a um conjunto de dados reais de origem médica. Observamos que a razão de chances entre indivíduos com diferentes valores da variável observada com erro, fixadas as demais covariáveis, é subestimada quando ignoramos a presença de erros de medida.

- No Capítulo 7 apresentamos a aplicação a um conjunto de dados da área financeira. Observamos que não houve diferença entre os métodos de estimação nas medidas de desempenho na predição com e sem a correção dependente de *status*. Ainda neste capítulo apresentamos alternativas para o ajuste do modelo para o conjunto de dados em questão.

  Os dados da área financeira foram fornecidos pela empresa Serasa Experian.

8.1 Trabalhos futuros

Como trabalhos futuros, propomos:

- realizar as alternativas de análise do conjunto de dados da área financeira como discutimos na Seção 7.1;

- aplicar a metodologia estudada em dados da área médica no Brasil. Em específico, temos o contato e interesse de ambas as partes em aplicar a metodologia estudada em um conjunto de dados da nutrição do Instituto do Câncer do Estado de São Paulo;

- considerar outros valores de h no estudo de simulação, como $h = 0{,}10$, $0{,}15$, $0{,}2$ e $0{,}35$;

- estudar os métodos de estimação considerados neste trabalho em situações com duas covariáveis medidas com erro e correlacionadas. Uma motivação para este estudo é

ainda os dados do estudo *Framinghanm Heart*: além da pressão sistólica, podemos considerar que o nível de colesterol foi medido com erro;

- realizar um estudo mais aprofundado do método de MPV sob diferentes proporções de eventos de interesse;

- realizar um estudo de simulação quando a variável com erro de medida não tem distribuição normal, mas é erroneamente suposta na utilização dos métodos estruturais máxima verossimilhança e máxima pseudoverossimilhança. Thoresen (2006) realizou um estudo de simulação onde a variável com erro de medida é gerada por diversas distribuições não normais e verificou o desempenho do método da calibração da regressão, uma vez que a estimativa da esperança condicional de X_i dado (W_i, Z_i) apresentada na Seção 4.1.4 é exata quando a distribuição condicional de $X_i|(W_i, Z_i)$ é normalmente distribuída e é aproximada para os demais casos;

- estudar outras funções de extrapolação para o método SIMEX;

- considerar erro de medida na variável resposta.

Apêndice A

Variância dos estimadores

A.1 Variância do estimador de MPV

Para encontrar a distribuição assintótica do estimador de MPV para β, Carroll *et al.* (2006, Seção A.6.6) utilizam a teoria de equações de estimação. Para facilitar a notação e considerando variáveis observadas sem erro (Z_i), definimos $pl_i(\beta;\lambda) = \log[pL(\beta; y_i, w_i, z_i, \lambda)] = \log \int f_{YWZX}(y_i, w_i, z_i, x_i|\beta, \lambda)dx_i$ e $rl_i(\lambda) = \log[rL(\lambda; w_i, z_i)] = \log \int f_{WZX}(w_i, z_i, x_i|\lambda)dx_i$, $i = 1, \ldots, n$. Sob condições de regularidade apresentadas por Gong & Samaniego (1981), a distribuição assintótica de $\sqrt{n}(\widehat{\beta} - \beta)$ é normal com média 0 e matriz de covariâncias Σ (Guolo, 2010), em que

$$\Sigma = I_{\beta\beta}^{-1}(\Sigma_{\beta\beta} - I_{\beta\lambda}I_{\lambda\lambda}^{-1}\Sigma_{\lambda\beta} - \Sigma_{\lambda\beta}^{\top}I_{\lambda\lambda}^{-1}I_{\beta\lambda}^{\top} + I_{\beta\lambda}I_{\lambda\lambda}^{-1}\Sigma_{\lambda\lambda}I_{\lambda\lambda}^{-1}I_{\beta\lambda}^{\top})I_{\beta\beta}^{-1}. \tag{A.1}$$

A forma reduzida de Σ desenvolvida por Parke (1986) e considerada por Skrondal & Kuha (2012) foi utilizada e apresenta resultados semelhantes à expressão (A.1). Os blocos da matriz Σ são

$$I_{\beta\beta} = -\sum_{i=1}^{n} E\left\{\frac{\partial^2 pl_i(\beta;\lambda)}{\partial\beta\partial\beta^{\top}}\right\}\Bigg|_{\lambda=\widehat{\lambda}}, \quad I_{\lambda\lambda} = -\sum_{i=1}^{n} E\left\{\frac{\partial^2 rl_i(\lambda)}{\partial\lambda\partial\lambda^{\top}}\right\},$$

$$I_{\beta\lambda} = -\sum_{i=1}^{n} E\left\{\frac{\partial^2 pl_i(\lambda)}{\partial\beta\partial\lambda^{\top}}\right\}\Bigg|_{\lambda=\widehat{\lambda}}, \quad \Sigma_{\beta\beta} = \sum_{i=1}^{n} \frac{\partial pl_i(\beta;\lambda)}{\partial\beta}\Bigg|_{\lambda=\widehat{\lambda}} \left\{\frac{\partial pl_i(\beta;\lambda)}{\partial\beta}\Bigg|_{\lambda=\widehat{\lambda}}\right\}^{\top},$$

$$\Sigma_{\lambda\lambda} = \sum_{i=1}^{n} \frac{\partial rl_i(\lambda)}{\partial\lambda}\left\{\frac{\partial rl_i(\lambda)}{\partial\lambda}\right\}^{\top} \quad \text{e} \quad \Sigma_{\lambda\beta} = \sum_{i=1}^{n} \frac{\partial rl_i(\lambda)}{\partial\lambda}\left\{\frac{\partial pl_i(\beta;\lambda)}{\partial\beta}\Bigg|_{\lambda=\widehat{\lambda}}\right\}^{\top}.$$

Como apresenta Louis (1982), as derivadas de $pl_i(\beta;\lambda)$ com respeito a β é

$$\frac{\partial}{\partial\beta}pl_i(\beta;\lambda) = \frac{\int \frac{\partial}{\partial\beta} f_{YWZX}(y_i,w_i,z_i,x_i|\beta,\lambda)dx_i}{\int f_{YWZX}(y_i,w_i,z_i,x_i|\beta,\lambda)dx_i}.$$

Agora, multiplicamos e dividimos a integral do numerador por $f_{YWZX}(y_i,w_i,z_i,x_i|\beta,\lambda)$, obtendo

$$\begin{aligned}
\frac{\partial}{\partial\beta}pl_i(\beta;\lambda) &= \frac{\int \frac{\partial}{\partial\beta} f_{YWZX}(y_i,w_i,z_i,x_i|\beta,\lambda)\frac{f_{YWZX}(y_i,w_i,z_i,x_i|\beta,\lambda)}{f_{YWZX}(y_i,w_i,z_i,x_i|\beta,\lambda)}dx_i}{\int f_{YWZX}(y_i,w_i,z_i,x_i|\beta,\lambda)dx_i} \\
&= \frac{\int \frac{\frac{\partial}{\partial\beta} f_{YWZX}(y_i,w_i,z_i,x_i|\beta,\lambda)}{f_{YWZX}(y_i,w_i,z_i,x_i|\beta,\lambda)} f_{YWZX}(y_i,w_i,z_i,x_i|\beta,\lambda)dx_i}{\int f_{YWZX}(y_i,w_i,z_i,x_i|\beta,\lambda)dx_i} \\
&= \int \frac{\partial}{\partial\beta}\log f_{YWZX}(y_i,w_i,z_i,x_i|\beta,\lambda)\left[\frac{f_{YWZX}(y_i,w_i,z_i,x_i|\beta,\lambda)}{\int f_{YWZX}(y_i,w_i,z_i,x_i|\beta,\lambda)dx_i}\right]dx_i \\
&= \int \left[\frac{\partial}{\partial\beta}\log f_{YWZX}(y_i,w_i,z_i,x_i|\beta,\lambda)\right] f_{X|YWZ}(x_i|y_i,w_i,z_i;\beta,\lambda)dx_i \\
&= E\left\{\frac{\partial}{\partial\beta}\log f_{YWZX}(y_i,w_i,z_i,x_i|\beta,\lambda)\middle| y_i,w_i,z_i;\beta,\lambda\right\} \\
&= S_\beta(\beta;\lambda). \qquad \text{(A.2)}
\end{aligned}$$

Mas como a esperança de (A.2) não pode ser obtida de forma fechada, utilizamos uma aproximação de Monte Carlo. Assim,

$$\begin{aligned}
& E\left\{\frac{\partial}{\partial\beta}\log f_{YWZX}(y_i,w_i,z_i,x_i|\beta,\lambda)\middle| y_i,w_i,z_i;\beta,\lambda\right\} \\
&= \int \frac{\partial}{\partial\beta}\log f_{YWZX}(y_i,w_i,z_i,x_i|\beta,\lambda) f_{X|YWZ}(x_i|y_i,w_i,z_i;\beta,\lambda)dx_i \\
&= \int g(x_i) f_{X|YWZ}(x_i|y_i,w_i,z_i;\beta,\lambda)dx_i \approx \frac{1}{M}\sum_{m=1}^{M} k_{m,i}\, g(x^*_{m,i}), \qquad \text{(A.3)}
\end{aligned}$$

em que $k_{m,i}$ é o peso da amostragem de importância. Desta forma, a matriz $\Sigma_{\beta\beta}$ pode ser aproximada por

$$\sum_{i=1}^{n}\left\{\frac{1}{M}\sum_{m=1}^{M} k_{m,i}\frac{\partial}{\partial\beta}pl(\beta;y_i,w_i,z_i,x^*_{m,i},\lambda)\middle|_{\lambda=\hat\lambda}\right\}\left\{\frac{1}{M}\sum_{m=1}^{M} k_{m,i}\frac{\partial}{\partial\beta}pl(\beta;y_i,w_i,z_i,x^*_{m,i},\lambda)\middle|_{\lambda=\hat\lambda}\right\}^\top,$$

em que $x^*_{m,i}$ é gerado pela distribuição de $X_i|Z_i = z_i$ da última iteração do algoritmo MCEM. Na presença de covariáveis sem erro Z_i, o peso da amostragem de importância fica

$$k_{m,i} \approx \frac{f_{Y|WZX}(y_i|w_i,z_i,x^*_{m,i};\beta) f_{W|ZX}(w_i|z_i,x^*_{m,i};\beta)}{\frac{1}{M}\sum_{m=1}^{M} f_{Y|WZX}(y_i|w_i,z_i,x^*_{m,i};\beta) f_{W|ZX}(w_i|z_i,x^*_{m,i};\beta)}.$$

As matrizes $\Sigma_{\lambda\beta}$ e $\Sigma_{\lambda\lambda}$ são aproximadas de maneira análoga. A aproximação para $\Sigma_{\lambda\lambda}$ é dada por

$$\sum_{i=1}^{n}\left\{\frac{1}{M}\sum_{m=1}^{M}kr_{m,i}\frac{\partial}{\partial\lambda}rl(\lambda;w_i,z_i,x^*_{m,i})\right\}\left\{\frac{1}{M}\sum_{m=1}^{M}kr_{m,i}\frac{\partial}{\partial\lambda}rl(\lambda;w_i,z_i,x^*_{m,i})\right\}^{\top},$$

e a aproximação para $\Sigma_{\lambda\beta}$ é

$$\sum_{i=1}^{n}\left\{\frac{1}{M}\sum_{m=1}^{M}kr_{m,i}\frac{\partial}{\partial\lambda}rl(\lambda;w_i,z_i,x^*_{m,i})\right\}\left\{\frac{1}{M}\sum_{m=1}^{M}k_{m,i}\frac{\partial}{\partial\beta}pl(\beta;y_i,w_i,z_i,x^*_{m,i},\lambda)\right\}^{\top}.$$

O peso da amostragem de importância $kr_{m,i}$ é dado por

$$kr_{m,i}\approx\frac{f_{W|ZX}(w_i|z_i,x^*_{m,i};\lambda)}{\frac{1}{M}\sum_{m=1}^{M}f_{W|ZX}(w_i|z_i,x^*_{m,i};\lambda)}.$$

As matrizes $I_{\beta\beta}$, $I_{\lambda\lambda}$ e $I_{\beta\lambda}$ podem ser calculadas ao explorar a fórmula desenvolvida por Louis (1982). A matriz hessiana de $pl_i(\beta;\lambda)$ com respeito a β é

$$\begin{aligned}\frac{\partial^2}{\partial\beta\partial\beta^{\top}}pl_i(\beta;\lambda) &= \frac{\partial}{\partial\beta}\left\{\frac{\partial}{\partial\beta}pl_i(\beta;\lambda)\right\}\\ &= \frac{\partial}{\partial\beta}\left\{\frac{\int\frac{\partial}{\partial\beta}f_{YWZX}(y_i,w_i,z_i,x_i|\beta,\lambda)dx_i}{\int f_{YWZX}(y_i,w_i,z_i,x_i|\beta,\lambda)dx_i}\right\}.\end{aligned}$$

Calculando a derivada acima, temos

$$\begin{aligned}\frac{\partial^2}{\partial\beta\partial\beta^{\top}}pl_i(\beta;\lambda) &= \frac{\int\frac{\partial^2}{\partial\beta\partial\beta^{\top}}f_{YWZX}(y_i,w_i,z_i,x_i|\beta,\lambda)dx_i\int f_{YWZX}(y_i,w_i,z_i,x_i|\beta,\lambda)dx_i}{[\int f_{YWZX}(y_i,w_i,z_i,x_i|\beta,\lambda)dx_i]^2}\\ &\quad-\frac{\int\frac{\partial}{\partial\beta}f_{YWZX}(y_i,w_i,z_i,x_i|\beta,\lambda)dx_i\int\frac{\partial}{\partial\beta}f_{YWZX}(y_i,w_i,z_i,x_i|\beta,\lambda)dx_i}{[\int f_{YWZX}(y_i,w_i,z_i,x_i|\beta,\lambda)dx_i]^2}\\ &= \frac{\int\frac{\partial^2}{\partial\beta\partial\beta^{\top}}f_{YWZX}(y_i,w_i,z_i,x_i|\beta,\lambda)dx_i}{\int f_{YWZX}(y_i,w_i,z_i,x_i|\beta,\lambda)dx_i}\\ &\quad-\frac{\int\frac{\partial}{\partial\beta}f_{YWZX}(y_i,w_i,z_i,x_i|\beta,\lambda)dx_i}{\int f_{YWZX}(y_i,w_i,z_i,x_i|\beta,\lambda)dx_i}\frac{\int\frac{\partial}{\partial\beta}f_{YWZX}(y_i,w_i,z_i,x_i|\beta,\lambda)dx_i}{\int f_{YWZX}(y_i,w_i,z_i,x_i|\beta,\lambda)dx_i}\\ &= \frac{\int\frac{\partial^2}{\partial\beta\partial\beta^{\top}}f_{YWZX}(y_i,w_i,z_i,x_i|\beta,\lambda)dx_i}{\int f_{YWZX}(y_i,w_i,z_i,x_i|\beta,\lambda)dx_i}-S_{\beta}(\beta;\lambda)S^{\top}_{\beta}(\beta;\lambda).\end{aligned}\tag{A.4}$$

Multiplicando e dividindo o primeiro termo do lado direito da igualdade de (A.4), que chamamos de (I), por $f_{YWZX}(y_i,w_i,z_i,x_i|\beta,\lambda)$, temos

$$\begin{aligned}(I) &= \frac{\int\frac{\partial^2}{\partial\beta\partial\beta^{\top}}f_{YWZX}(y_i,w_i,z_i,x_i|\beta,\lambda)\frac{f_{YWZX}(y_i,w_i,z_i,x_i|\beta,\lambda)}{f_{YWZX}(y_i,w_i,z_i,x_i|\beta,\lambda)}dx_i}{\int f_{YWZX}(y_i,w_i,z_i,x_i|\beta,\lambda)dx_i}\\ &= \int\frac{\frac{\partial^2}{\partial\beta\partial\beta^{\top}}f_{YWZX}(y_i,w_i,z_i,x_i|\beta,\lambda)}{f_{YWZX}(y_i,w_i,z_i,x_i|\beta,\lambda)}f_{X|YWZ}(x_i|y_i,w_i,z_i;\beta,\lambda)dx_i\\ &= E\left\{\frac{\frac{\partial^2}{\partial\beta\partial\beta^{\top}}f_{YWZX}(y_i,w_i,z_i,x_i|\beta,\lambda)}{f_{YWZX}(y_i,w_i,z_i,x_i|\beta,\lambda)}\middle|y_i,w_i,z_i;\beta,\lambda\right\}.\end{aligned}$$

Ainda, temos que

$$
\begin{aligned}
& E\left\{\frac{\partial^2}{\partial\beta\partial\beta^\top}\log f_{YWZX}(y_i,w_i,z_i,x_i|\beta,\lambda)\middle| y_i,w_i,z_i;\beta,\lambda\right\} \\
&= E\left\{\frac{\frac{\partial^2}{\partial\beta\partial\beta^\top}f_{YWZX}(y_i,w_i,z_i,x_i|\beta,\lambda)}{f_{YWZX}(y_i,w_i,z_i,x_i|\beta,\lambda)}\middle| y_i,w_i,z_i;\beta,\lambda\right\} \\
-\quad & E\left\{\frac{\partial}{\partial\beta}\log f_{YWZX}(y_i,w_i,z_i,x_i|\beta,\lambda)\left[\frac{\partial}{\partial\beta}\log f_{YWZX}(y_i,w_i,z_i,x_i|\beta,\lambda)\right]^\top\middle| y_i,w_i,z_i;\beta,\lambda\right\}.
\end{aligned}
$$

Isolando o segundo termo do lado direito da igualdade acima,

$$
\begin{aligned}
& (I) = E\left\{\frac{\partial^2}{\partial\beta\partial\beta^\top}\log f_{YWZX}(y_i,w_i,z_i,x_i|\beta,\lambda)\middle| y_i,w_i,z_i;\beta,\lambda\right\} \\
+\quad & E\left\{\frac{\partial}{\partial\beta}\log f_{YWZX}(y_i,w_i,z_i,x_i|\beta,\lambda)\left[\frac{\partial}{\partial\beta}\log f_{YWZX}(y_i,w_i,z_i,x_i|\beta,\lambda)\right]^\top\middle| y_i,w_i,z_i;\beta,\lambda\right\}.
\end{aligned}
$$

Assim, temos que

$$
\begin{aligned}
& \frac{\partial^2}{\partial\beta\partial\beta^\top}pl_i(\beta;\lambda) = E\left\{\frac{\partial^2}{\partial\beta\partial\beta^\top}\log f_{YWZX}(y_i,w_i,z_i,x_i|\beta,\lambda)\middle| y_i,w_i,z_i;\beta,\lambda\right\} \\
+\quad & E\left\{\frac{\partial}{\partial\beta}\log f_{YWZX}(y_i,w_i,z_i,x_i|\beta,\lambda)\left[\frac{\partial}{\partial\beta}\log f_{YWZX}(y_i,w_i,z_i,x_i|\beta,\lambda)\right]^\top\middle| y_i,w_i,z_i;\beta,\lambda\right\} \\
-\quad & S_\beta(\beta;\lambda)S_\beta^\top(\beta;\lambda).
\end{aligned}
$$

A matriz $I_{\beta\beta}$ é dada por

$$
\begin{aligned}
I_{\beta\beta} &= -\sum_{i=1}^{n} E\left\{\frac{\partial^2}{\partial\beta\partial\beta^\top}pl_i(\beta;\lambda)\right\}\Bigg|_{\lambda=\hat\lambda} \\
&= -\sum_{i=1}^{n}\Bigg[E\left\{\frac{\partial^2}{\partial\beta\partial\beta^\top}\log f_{YWZX}(y_i,w_i,z_i,x_i|\beta,\hat\lambda)\middle| y_i,w_i,z_i;\beta,\hat\lambda\right\} \\
&+ E\left\{\frac{\partial}{\partial\beta}\log f_{YWZX}(y_i,w_i,z_i,x_i|\beta,\hat\lambda)\left[\frac{\partial}{\partial\beta}\log f_{YWZX}(y_i,w_i,z_i,x_i|\beta,\lambda)\right]^\top\middle| y_i,w_i,z_i;\beta,\hat\lambda\right\} \\
&- S_\beta(\beta;\hat\lambda)S_\beta^\top(\beta;\hat\lambda)\Bigg]. \qquad \text{(A.5)}
\end{aligned}
$$

Mas como as esperanças de (A.5) não podem ser obtidas de forma fechada, utilizamos

a mesma ideia apresentada em (A.3). Assim, a aproximação de $I_{\beta\beta}$ fica

$$\begin{aligned}
I_{\beta\beta} \;&\approx \\
-\;& \sum_{i=1}^{n}\left[\frac{1}{M}\sum_{m=1}^{M} k_{m,i}\frac{\partial^2}{\partial\beta\partial\beta^\top}\log f_{YWZX}(y_i,w_i,z_i,x^*_{m,i}|\beta,\hat{\lambda})\right. \\
+\;& \frac{1}{M}\sum_{m=1}^{M} k_{m,i}\frac{\partial}{\partial\beta}\log f_{YWZX}(y_i,w_i,z_i,x^*_{m,i}|\beta,\hat{\lambda})\left\{\frac{\partial}{\partial\beta}\log f_{YWZX}(y_i,w_i,z_i,x^*_{m,i}|\beta,\lambda)\right\}^\top \\
-\;& \frac{1}{M}\sum_{m=1}^{M} k_{m,i}\frac{\partial}{\partial\beta}\log f_{YWZX}(y_i,w_i,z_i,x^*_{m,i}|\beta,\lambda) \\
& \left.\times\left\{\frac{1}{M}\sum_{m=1}^{M} k_{m,i}\frac{\partial}{\partial\beta}\log f_{YWZX}(y_i,w_i,z_i,x^*_{m,i}|\beta,\lambda)\right\}^\top\right]. \qquad \text{(A.6)}
\end{aligned}$$

No mesmo caminho, a derivada de $rl_i(\lambda)$ com respeito a λ é

$$\frac{\partial rl_i(\lambda)}{\partial\lambda} = \frac{\int \frac{\partial}{\partial\lambda} f_{WZX}(w_i,z_i,x_i|\lambda)dx_i}{\int f_{WZX}(w_i,z_i,x_i|\lambda)dx_i}.$$

Ao multiplicarmos e dividirmos o integrando do numerador por $f_{WZX}(w_i,z_i,x_i;\lambda)$, obtemos

$$\frac{\partial}{\partial\lambda}rl_i(\lambda) = E\left\{\frac{\partial}{\partial\lambda}\log f_{WZX}(w_i,z_i,x_i|\lambda)\middle| w_i,z_i;\lambda\right\} = S_\lambda(\lambda).$$

Similarmente ao caso anterior,

$$\begin{aligned}
&\frac{\partial^2}{\partial\lambda\partial\lambda^\top}rl_i(\lambda) = E\left\{\frac{\partial^2}{\partial\lambda\partial\lambda^\top}\log f_{WZX}(w_i,z_i,x_i|\lambda)\middle| w_i,z_i;\lambda\right\} \\
+\;& E\left\{\frac{\partial}{\partial\lambda}\log f_{WZX}(w_i,z_i,x_i|\lambda)\left[\frac{\partial}{\partial\lambda}\log f_{WZX}(w_i,z_i,x_i|\lambda)\right]^\top\middle| w_i,z_i;\lambda\right\} \\
-\;& S_\lambda(\lambda)S_\lambda^\top(\lambda).
\end{aligned}$$

Analogamente ao apresentado em (A.6), a aproximação de $I_{\lambda\lambda}$ é

$$\begin{aligned}
I_{\lambda\lambda} \;&\approx \\
-\;& \sum_{i=1}^{n}\left[\frac{1}{M}\sum_{m=1}^{M} kr_{m,i}\frac{\partial^2}{\partial\lambda\partial\lambda^\top}\log f_{WZX}(w_i,z_i,x^*_{m,i}|\lambda)\right. \\
+\;& \frac{1}{M}\sum_{m=1}^{M} kr_{m,i}\frac{\partial}{\partial\lambda}\log f_{WZX}(w_i,z_i,x^*_{m,i}|\lambda)\left\{\frac{\partial}{\partial\lambda}\right\}\log f_{WZX}(w_i,z_i,x^*_{m,i}|\lambda)^\top \\
-\;& \frac{1}{M}\sum_{m=1}^{M} kr_{m,i}\frac{\partial}{\partial\lambda}\log f_{WZX}(w_i,z_i,x^*_{m,i}|\lambda) \\
& \left.\times\left\{\frac{1}{M}\sum_{m=1}^{M} kr_{m,i}\frac{\partial}{\partial\lambda}\log f_{WZX}(w_i,z_i,x^*_{m,i}|\lambda)\right\}^\top\right].
\end{aligned}$$

Já a aproximação para $I_{\beta\lambda}$ é

$$
\begin{aligned}
I_{\beta\lambda} \;\approx\; & \\
- \;& \sum_{i=1}^{n}\left[\frac{1}{M}\sum_{m=1}^{M} k_{m,i}\frac{\partial^2}{\partial\beta\partial\lambda^\top}\log f_{YWZX}(y_i,w_i,z_i,x^*_{m,i}|\beta,\lambda)\right. \\
+ \;& \frac{1}{M}\sum_{m=1}^{M} k_{m,i}\frac{\partial}{\partial\beta}\log f_{YWZX}(y_i,w_i,z_i,x^*_{m,i}|\beta,\hat\lambda)\left\{\frac{\partial}{\partial\lambda}\log f_{WZX}(w_i,z_i,x^*_{m,i}|\lambda)\right\}^\top \\
- \;& \frac{1}{M}\sum_{m=1}^{M} k_{m,i}\frac{\partial}{\partial\beta}\log f_{YWZX}(y_i,w_i,z_i,x^*_{m,i}|\beta,\hat\lambda) \\
& \left.\times\left\{\frac{1}{M}\sum_{m=1}^{M} kr_{m,i}\frac{\partial}{\partial\lambda}\log f_{WZX}(w_i,z_i,x^*_{m,i}|\lambda)\right\}^\top\right].
\end{aligned}
$$

A.2 Variância do estimador SIMEX

Vamos considerar a abordagem feita por Stefanski & Cook (1995) para estimar a variância do estimador SIMEX, que mostram ser dada por

$$\mathrm{Var}(\widehat\beta_{\text{simex}}) \approx \mathrm{Var}(\widehat\beta_{\text{verdadeiro}}) + \mathrm{Var}(\widehat\beta_{\text{simex}} - \widehat\beta_{\text{verdadeiro}}), \tag{A.7}$$

em que $\widehat\beta_{verdadeiro}$ é um estimador para β obtido com X_i disponível. A equação em (A.7) decompõe a variância do estimador SIMEX em um componente referente à variabilidade da amostra, $\mathrm{Var}(\widehat\beta_{\text{verdadeiro}}) = \tau^2$, e um componente referente à variabilidade do erro de medida, $\mathrm{Var}(\widehat\beta_{\text{simex}} - \widehat\beta_{\text{verdadeiro}})$.

Seja $\widehat\tau^2_{verdadeiro} = T_{var}[\{Y_i, Z_i, X_i\}_1^n] = \widehat{\mathrm{Var}}(\widehat\beta_{\text{verdadeiro}})$ um estimador para a variância de $\widehat\beta_{verdadeiro}$. Como X_i não é disponível, um estimador para τ^2 é dado por

$$\widehat\tau^2_d(\zeta_b) = T_{var}[\{Y_i, Z_i, W_{d,i}(\zeta_b)\}_1^n],$$

em que $T_{var}[Y_i, Z_i, W_{d,i}(\zeta_b)_1^n]$ é a variância estimada do estimador de máxima verossimilhança dos parâmetros do modelo de regressão logística considerando $W_{d,i}(\zeta_b)$ no lugar de X_i. Ao obter a média desses valores, para $d = 1,\ldots,D$ e $i = 1,\ldots,n$, chegamos à estimativa $\widehat\tau^2(\zeta_b)$.

Para obter a estimativa do segundo componente da variância, $\mathrm{Var}(\widehat\beta_{\text{simex}} - \widehat\beta_{\text{verdadeiro}})$, calculamos as diferenças

$$\Delta_d(\zeta_b) = \widehat\beta^{(d)}(\zeta_b) - \widehat\beta(\zeta_b),\ d = 1,\ldots,D,\ b = 1,\ldots,B,$$

onde $\widehat{\beta}(\zeta_b)$ é a média de $\widehat{\beta}^{(1)}(\zeta_b), \ldots, \widehat{\beta}^{(D)}(\zeta_b)$. Ainda, seja $s^2_\Delta(\zeta_b)$ definido por

$$s^2_\Delta(\zeta_b) = (D-1)^{-1} \sum_{d=1}^{D} \Delta_d(\zeta_b)\Delta_d^\top(\zeta_b),$$

ou seja, a matriz de covariâncias amostral de $\{\widehat{\beta}^{(d)}(\zeta_b)\}_{d=1}^D$. Isso porque o componente de variância que queremos estimar é dado por

$$\mathrm{Var}(\widehat{\beta}_{\mathrm{simex}} - \widehat{\beta}_{\mathrm{verdadeiro}}) = -\lim_{\zeta_\mathrm{b} \to -1} \mathrm{E}\{\mathrm{s}^2_\Delta(\zeta_\mathrm{b})\}.$$

Para mais detalhes, vide Carroll *et al.* (2006, Seção B.4) e Stefanski & Cook (1995).

Desta maneira, a diferença $\{\widehat{\tau}^2(-1) - s^2_\Delta(-1)\}$ é um estimador para $\mathrm{Var}(\widehat{\beta}_{\mathrm{simex}})$. Na estimação SIMEX, o passo de simulação consiste em obter $\widehat{\beta}(\zeta_b)$, $\widehat{\tau}^2(\zeta_b)$ e $s^2_\Delta(\zeta_b)$ para $\zeta_b = \zeta_1, \ldots, \zeta_B$. O modelo de extrapolação de $\widehat{\beta}(\zeta_b)$ para $\zeta_b = -1$ resulta em $\widehat{\beta}_{simex}$, um estimador para β, e a extrapolação da diferença $\widehat{\tau}^2 - s^2_\Delta$ para $\zeta_b = -1$ resulta em um estimador para $\mathrm{Var}(\widehat{\beta}_{\mathrm{simex}})$.

A.3 Variância do estimador da CR

Os erros padrão das estimativas $\widehat{\beta}$ resultantes do método de calibração da regressão podem ser obtidos por reamostragem, como a técnica *bootstrap*, a qual utilizamos neste trabalho. Seja $\widehat{\beta}$ a estimativa dos parâmetros β e $\widehat{\beta}^{(q)}$ a estimativa obtida pelo mesmo estimador a partir da q-ésima amostra *bootstrap*, $q = 1, \ldots, Q$ e seja também $\bar{\beta}$ a média de $\widehat{\beta}^{(1)}, \ldots, \widehat{\beta}^{(Q)}$. Desta maneira,

$$\widehat{\mathrm{Var}}(\widehat{\beta}) = (\mathrm{Q}-1)^{-1} \sum_{\mathrm{q}=1}^{\mathrm{Q}} (\widehat{\beta}^{(\mathrm{q})} - \bar{\beta})(\widehat{\beta}^{(\mathrm{q})} - \bar{\beta})^\top.$$

Foram utilizadas $Q = 1000$ réplicas *bootstrap*.

Referências Bibliográficas

Booth, J. G. & Hobert, J. P. Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *Journal of the Royal Statistical Society*, 61: 265–285, 1999.

Breslow, N. E. Statistics in epidemiology: the case-control study. *Journal of American Statistical Association*, 91:14–28, 1996.

Carrasco, J. M. F. *Modelos de Regressão Beta com Erro nas Variáveis*. Tese de Doutorado, Instiituto de Matemática e Estatística - IME-USP, 2012.

Carrasco, J. M. F., Ferrari, S. L. P. & Arellano-Valle, R. B. Errors-in-variables beta regression models. *arXiv: 1212.0870v2 [stat.ME]*, 2013.

Carroll, R. J., Spiegelman, C., Lan, K. K., Bailey, K. T. & Abbott, R. D. On errors-in-variables for binary regression models. *Biometrika*, 71:19–26, 1984.

Carroll, R. J., Ruppert, D., Stefanski, L. A. & Crainiceanu, C. M. *Measurement Error in Nonlinear Models - A Modern Perspective*. Chapman & Hall, 2nd ed., 2006.

Cook, J. R. & Stefanski, L. A. Simulation-extrapolation estimation in parametric measurement error models. *Journal of the American Statistical Association*, 89:1314–1328, 1994.

Copas, J. B. & Loeber, R. Relative improvement over chance (RIOC) for 2x2 tables. *British Journal of Mathematical and Statistical Psychology*, 43:293–307, 1990.

Cunha, W. J. & Colosimo, E. A. Intervalo de confiança bootstrap para modelos de regressão com erro de medida. *Revista de Matemática e Estatística*, 21:25–41, 2003.

Devanarayan, V. & Stefanski, L. A. Empirical simulation extrapolation for measurement error models with replicate measurements. *Statistics and Probability Letters*, 59:219–225, 2002.

Diniz, C. & Louzada, F. *Modelagem Estatística para Risco de Crédito.* Vigésimo SINAPE, Associação Brasileira de Estatística (ABE), 2012.

Fuller, W. A. *Measurement Error Models.* John Wiley, 1987.

Gong, G. & Samaniego, F. J. Pseudo maximum likelihood estimation: theory and applications. *The Annals of Statistics*, 9:861–869, 1981.

Guolo, A. Pseudo-likelihood inference for regression models with misclassified and mismeasured variables. *Statistica Sinica*, 21:1639–1663, 2010.

Gustafson, P. *Measurement Error and Misclassification in Statistics and Epidemiology : Impacts and Bayesian Adjustments.* Chapman & Hall, 2004.

Hilbe, J. M. *Logistic Regression Models.* Chapman & Hall, 2009.

Hosmer, D. W. & Lemeshow, S. *Applied Logistic Regression.* John Wiley & Sons, 2nd ed., 2000.

Kim, B. R., Carter, R. L., Rao, P. V., Ariet, M. & Resnick, M. B. Standardized risk and discription of results from multivariable modeling of a binary response. *Biometrical Journal*, 48:54–66, 2006.

King, G. & Zeng, L. *Logistic Regression in Rare Events Data.* Harvard University, Cambridge, 2001.

Louis, T. A. Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society, Ser. B*, 44:226–233, 1982.

Louzada, F., Ferreira-Silva, P. H. & Diniz, C. A. R. On the impact of disproportional samples in credit scoring models: an application to a Brazilian bank data. *Expert Systems with Applications*, 39:8071–8078, 2012.

Louzada-Neto, F., Amaral, G. J. A., Abreu, H. J., Guirado, L., Ferreira, M. R. P. & Silva, P. H. F. Medidas estatísticas da capacidade preditiva de modelos de classificação em *credit scoring*. *Revista Serasa Experian*, 2009.

McCullagh, P. & Nelder, J. A. *Generalized Linear Models*. Chapman & Hall, 2nd ed., 1989.

Nocedal, J. & Wright, S. J. *Numerical Optimization*. Springer, 2nd ed., 2006.

Parke, W. R. Pseudo maximum likelihood estimation: the asymptotic distribution. *The Annals of Statistics*, 14:355–357, 1986.

Paula, G. A. *Modelos de Regressão com Apoio Computacional*. Disponível em http://www.ime.usp.br/ giapaula/cursospos.htm, 2013.

Rodrigues, A. S. *Modelo de Regressão Logística com Erro de Medida*. Trabalho de Graduação, Departamento de Estatística - Universidade Federal de São Carlos - UFSCar, 2010.

Rosner, B., Willett, W. C. & Spiegelman, D. Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error. *Statistics in Medicine*, 8:1051–1069, 1989.

Rosner, B., Spiegelman, D. & Willett, W. C. Correction of logistic regression relative risk estimates and confidence intervals for measurement error: the case of multiple covariates measured with error. *American Journal of Epidemiology*, 132:734–745, 1990.

Rosner, B., Spiegelman, D. & Willett, W. C. Correction of logistic regression relative risk estimates and confidence intervals for random within-person measurement error. *American Journal of Epidemiology*, 136:1400–1413, 1992.

Schafer, D. W. Likelihood analysis and flexible structural modeling for measurement error model regression. *Journal of Statistical Computation and Simulation*, 72:53–61, 2002.

Schneeweiss, H. & Augustin, T. Some recent advances in measurement error models and methods. *Allgemeines Statistisches Archiv*, 90:183–197, 2006.

Skrondal, A. & Kuha, J. Improved regression calibration. *Psychometrika*, 77:649–669, 2012.

Spiegelman, D., Logan, R. & Grove, D. Regression calibration with heteroscedastic error variance. *The International Journal of Biostatistics*, 7:4, 2011.

Stefanski, L. A. Measurement error models. *Journal of the American Statistical Association*, 95:1353–1358, 2000.

Stefanski, L. A. & Carroll, R. J. Covariate measurement error in logistic regression. *The Annals of Statistics*, 13:1335–1351, 1985.

Stefanski, L. A. & Cook, J. R. Simulation-Extrapolation: the measurement error jacknife. *Journal of the American Statistical Association*, 90:1247–1256, 1995.

Thoresen, M. Correction for measurement error in multiple logistic regression: a simulation study. *Journal of Statistical Computation and Simulation*, 76:475–487, 2006.

Thoresen, M. & Laake, P. A simulation study of measurement error correction methods in logistic regression. *Journal of Statistical Computation and Simulation*, 77:683–694, 2007.

Tieppo, S. M. *Inferência em um Modelo de Regressão com Resposta Binária na Presença de Sobredispersão e Erros de Medição.* Dissertação de Mestrado, Instituto de Ciências Matemáticas e de Computação - ICMC-USP, 2007.

Verbeke, T. & Clercq, M. D. The income-environment relationship: evidence from a binary response model. *Ecological Economics*, 59:419–428, 2006.