UNIVERSIDADE FEDERAL DE PERNAMBUCO

CENTRO DE CIÊNCIAS EXATAS E DA NATUREZA

PROGRAMA DE PÓS-GRADUAÇÃO EM ESTATÍSTICA

LARISSA DOS SANTOS LIMA

MODELO PROBIT COM ERRO DE CLASSIFICAÇÃO E ERRO DE MEDIDA DO TIPO BERKSON NORMAL ASSIMÉTRICO

Recife

2020

LARISSA DOS SANTOS LIMA

MODELO PROBIT COM ERRO DE CLASSIFICAÇÃO E ERRO DE MEDIDA DO TIPO BERKSON NORMAL ASSIMÉTRICO

Dissertação apresentada ao Programa de Pós-Graduação em Estatística do Centro de Ciências Exatas e da Natureza da Universidade Federal de Pernambuco, como requisito parcial à obtenção do título de mestre em Estatística.

Área de Concentração: Estatística Matemática

Orientadora: Betsabé G. Blas Achic

Recife

2020

Catalogação na fonte
Bibliotecária Monick Raquel Silvestre da S. Portes, CRB4-1217

L732m Lima, Larissa dos Santos
Modelo probit com erro de classificação e erro de medida do tipo Berkson normal assimétrico / Larissa dos Santos Lima. – 2020.
59 f.: il., fig., tab.

Orientador: Betsabé G. Blas Achic.
Dissertação (Mestrado) – Universidade Federal de Pernambuco. CCEN, Estatística, Recife, 2020.
Inclui referências.

1. Estatística matemática. 2. Erro de medida. 3. Assimetria. I. Achic, Betsabé G. Blas (orientador). II. Título.

519.5 CDD (23. ed.) UFPE- CCEN 2020 - 117

LARISSA DOS SANTOS LIMA

MODELO PROBIT COM ERRO DE CLASSIFICAÇÃO E ERRO DE MEDIDA DO TIPO BERKSON NORMAL ASSIMÉTRICO

Dissertação apresentada ao Programa de Pós-Graduação em Estatística da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Mestre em Estatística.
Área de concentração: Estatística Matemática

Aprovada em: 05 DE JUNHO DE 2020

BANCA EXAMINADORA

Prof.(º) Aldo William Medina Garay
UFPE

Prof.(º) Cláudio Tadeu Cristino
UFRPE

Prof.(º) Roberto Ferreira Manghi
UFPE

À minha família.

AGRADECIMENTOS

A Deus, por ser essencial em minha vida, meu guia, socorro presente na hora da angústia, alívio em momentos de tensão.

À Nossa Senhora das Graças, por todas as bênçãos e proteção.

À minha família, que com muito carinho e apoio, não mediram esforços para que eu chegasse até esta etapa da minha vida. Mãe, seu cuidado, esforço e dedicação ao longo da vida foi que deram, em alguns momentos, a esperança para seguir. Pai, sua presença significou segurança e certeza de que não estou sozinha nessa caminhada. Irmão, você é meu orgulho e o melhor presente que Deus me deu. Vó, obrigada por todo amor, carinho e por suas orações. A vocês que, muitas vezes, renunciaram aos seus sonhos para que eu pudesse realizar o meu, partilho a alegria deste momento.

Ao meu namorado Bruno Lucian, por compreender a importância desse momento para minha vida profissional e mesmo distante se fazer presente, sempre que possível. Obrigada por todo amor, apoio, carinho, incentivo, preocupação, por me ajudar durante a elaboração dessa pesquisa e por me incentivar a buscar sempre o melhor.

À minha amiga e irmã Karina Santos, por não fazer da distância um obstáculo. Obrigada por estar comigo em vários momentos.

Aos meus amigos do Rio de Janeiro, por valorizarem cada reencontro.

Aos meus amigos da pós, que se tornaram minha família aqui em Recife, obrigada pela acolhida e receptividade. Agradeço especialmente à minha turma de mestrado (Ranah, Cristine, Professor Luis Félix, Joás, José (Zé) Carlos, André, Jordan e Fernando), obrigada por todos os momentos que dividimos aqui. Por todos os dias de estudo intenso, por toda cumplicidade, pelos momentos de descontração e alegria. Gostaria de agradecer, particularmente, àqueles que pude conviver e conhecer melhor, Jairo e Cristine. Vocês foram fundamentais para minha formação. Obrigada por trazerem leveza à essa caminhada. Serei eternamente grata pela amizade de vocês.

Às minhas companheiras de moradia, Fernanda Franklin e Ranah Duarte, por partilharem momentos em que confidenciamos alegrias e tristezas. À Ranah, por ter dividido comigo muitos momentos desde a graduação e ter "topado"mais esta etapa. E à Fernanda, por permitir que pudéssemos construir uma amizade ao compartilhar histórias e vivências. Certamente me recordarei com muito carinho de todas as histórias e momentos que dividimos.

Aos amigos que fiz em Recife, adorei conhecê-los e os levarei em meu coração.

À comunidade R-Ladies Recife, que tive o prazer e a alegria de fundar com a ajuda de outras meninas. Minha eterna gratidão por me dar a oportunidade de me enxergar profissionalmente.

À professora Betsabé Blas, pela orientação e apoio nesse trabalho.

Aos professores do Programa de Pós-Graduação em Estatística da UFPE, que foram tão importantes para minha formação acadêmica. Em especial, ao professor Raydonal Ospina, por ser tão humano ao olhar para os alunos. Obrigada pela paciência diante das minhas falhas e dificuldades, pela constante disponibilidade e pela amizade.

Aos professores, Roberto Manghi, Cláudio Tadeu e Aldo Garay, por terem aceitado o convite para compor a banca e pelas contribuições para esse trabalho.

À Universidade Federal de Pernambuco, pelos recursos físicos.

À FACEPE, pelo apoio financeiro.

À Recife, sou grata pela acolhida e receptividade.

RESUMO

Nesta dissertação, foi estudado o modelo de regressão binária com erro de classificação, que está associado à variável resposta; e erro de medida. O problema de erro de medida está associado à variável independente, que é muitas vezes custoso ou impossível de mensurar. Por isso, faz-se necessário considerar uma variável substituta. Em modelos lineares, é frequentemente assumido que as observações seguem uma distribuição normal, porém nem sempre essa suposição é válida. Portanto, neste trabalho propomos um modelo de regressão binária sujeito a erro de classificação e erro de medida do tipo Berkson na variável preditora, e o erro de medida segue distribuição normal assimétrica. Tal distribuição foi introduzida por AZZALINI (1985) e é importante para modelar a assimetria da distribuição dos dados. Assim, os efeitos dos erros de medida e dos erros de classificação são investigados através de um estudo de simulação de Monte Carlo. Finalmente, foi apresentado e explorado uma aplicação em dados reais.

Palavras-chave: Erro de medida. Erro de classificação. Normal assimétrica. Assimetria. Erro do tipo Berkson. Simulação de Monte Carlo.

ABSTRACT

In this master's thesis, was studied the binary regression model with misclassification, that is associated to the response variable; and measurement error. The measurement error problem is associated with the independent variable, which is mostly costly or impossible to be measured. Therefore, it is necessary to consider a substitute variable. In linear models, was often assumed that the observations follow a normal distribution, but this assumption is not always valid. Therefore, was proposed the binary regression model is subject to misclassification and measurement error of Berkson type on the predictor variable, and the measurement errors follow a skew normal distribution. This distribution was introduced by AZZALINI (1985) and it is useful on modeling asymmetric data distribution. Thus, the misclassification and measurement error effects are investigated by a simulation study, using the Monte Carlo's simulation aproach. An application was also presented and explored, and finally the conclusion about this study was presented.

Keywords: Measurement error. Misclassification. Skew normal distribution. Berkson type error. Monte Carlo simulation. Asymmetry.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

O problema de erro de medida em modelos de regressão com resposta binária é um tema que vem sendo bastante discutido na literatura entre os pesquisadores. O erro de medida ocorre quando não se pode observar exatamente uma ou mais variáveis que entram em um modelo de interesse. Existem muitos motivos pelos quais esses erros ocorrem, sendo os mais comuns o erro do instrumento de medição e o erro de amostragem. Aplicações nas quais a variável explicativa é medida na presença de erros são, possivelmente, mais comuns do que aquelas em que as medições são precisas. A maioria das variáveis médicas, tais como pressão sanguínea, batimentos cardíacos e temperatura corporal são medidas com erro (TIEPPO, 2007).

CARROL *et al.* (1984) analizaram dados de KANNEL e GORDON (1968) e mostraram o efeito dos erros de medida na regressão binária, iniciando com esse estudo uma área de pesquisa. Por outro lado, ROY, BANERJEE e MAITI (2005) consideraram um modelo de regressão para variável resposta binária sujeita a erro de classificação e variáveis independentes sujeitas a erro de medida. Os autores consideraram que os erros de medida são do tipo Berkson e seguem uma distribuição normal. Nesse estudo, algumas covariáveis são não observáveis e então foram consideradas medidas substitutas para essas covariáveis. Para avaliar o efeito de ignorar o erro de classificação e/ou erro de medida na estimativa dos parâmetros de regressão, foi feito um estudo de simulação e para o ajuste do modelo foi utilizada a função de verossimilhança.

Neste trabalho X será denotado como a variável independente não observada ou verdadeira, e também será considerado uma variável substituta denotada por W, a qual é observada. Em experimentos de bioensaios, na área epidemiológica, os valores da variável X podem não ser exatamente observados. No contexto de erro de medida, BERKSON (1950) notou que nos casos em que a variável substituta W é corrigida pelo pesquisador, ou seja, quando é adicionado o erro à W, há uma diferença substancial da situação em que W não é corrigida. Foi mostrado que na regressão linear com erro na variável X, o estimador por mínimos quadrados é consistente quando W é fixo. Ao considerar o modelo de regressão binária com erro na variável X, o estimador por máxima verossimilhança não é consistente se W é ou não corrigido. Um exemplo apresentado foi um bioensaio, no qual os organismos são expostos ao aumento das concentrações especificadas do material a ser analisado e portanto, a dosagem é uma observação controlada correspondente à variável X.

BURR (1988) estudou os erros de medida do tipo Berkson no campo de bioensaios. Nesse estudo, foi considerado o modelo probit com erros do tipo Berkson normalmente distribuí-

dos. Para a estimação dos parâmetros, foi utilizado o estimador de máxima verossimilhança.

REEVES, COX e DARBY (1998) estudaram o modelo de regressão com erro de medida para variáveis explicativas, incorporando como caso particular os erros de medida do tipo Berkson, considerando erros aditivos ou multiplicativos. O estudo foi motivado por problemas epidemiológicos, levando-se em consideração uma variável resposta contínua, como também considerou-se um outro cenário para o modelo de regressão logística. Os procedimentos de estimação considerados relativamente simples propostos para uso com dados de coorte e dados de caso controle são verificados através de simulação, sob a suposição de erros de vários tipos (funcional, estrutural, Berkson, entre outros). Os resultados mostram que mesmo em situações onde a análise convencional fornece estimativas de inclinação (intercepto) que são em média atenuadas por um fator de aproximadamente 50%, as estimativas obtidas através de funções de probabilidade alterada propostas estão dentro de 5% de seus valores reais.

MCGLOTHLIN, STAMEY e SEAMAN (2008) consideraram uma análise bayesiana para modelar uma resposta binária sujeita ao erro de classificação. Além disso, presume-se que uma variável explicativa é não observável, mas tem-se uma variável substituta. Um modelo de regressão binário é desenvolvido para incorporar o erro de medição na covariável, bem como a classificação incorreta na variável resposta. Diferentemente dos métodos existentes, nenhum parâmetro do modelo precisa ser considerado conhecido. Um experimento de simulação, a partir do método de Monte Carlo explora as vantagens da abordagem e a metodologia desenvolvida é ilustrada usando dados de sobrevivência de bombas atômicas.

STEFANSKI e CARROL (1987) estudaram a estimação em modelos lineares generalizados quando o vetor de variáveis explicativas apresenta erro de medida, e generaliza os resultados obtidos por STEFANSKI e CARROL (1985) para regressão logística. No caso em que tem-se um vetor de variáveis explicativas independentes e identicamente distribuídas com distribuição desconhecida, obtém-se funções escores eficientes. Já BICKEL e RITOV (1987) apresentam os resultados relacionados ao erro de medida do tipo estrutural.

A suposição que as observações seguem uma distribuição normal é uma prática comum em modelos lineares e não lineares. No entanto, em muitas aplicações essa suposição não é válida. Na prática, existem dados que seguem uma distribuição assimétrica. AZZALINI (1985) propôs a distribuição normal assimétrica, em que essa nova classe de funções de densidade depende de um parâmetro de assimetria, λ, de modo que quando $\lambda = 0$ tem-se como caso particular a distribuição normal.

AZZALINI e VALLE (1996) e AZZALINI e CAPITANIO (1999) estudaram uma

versão multivariada da distribuição normal assimétrica. Essa distribuição representa uma extensão matematicamente tratável da densidade normal multivariada com a adição de parâmetros para regular a assimetria. Os autores demonstram que a distribuição normal assimétrica multivariada possui uma flexibilidade razoável no ajuste real dos dados, enquanto mantém algumas propriedades da densidade normal. Essa abordagem é importante para aplicações práticas pois, no caso multivariado, há menos distribuições disponíveis para lidar com dados com distribuições não normais comparado ao caso univariado.

Na área epidemiológica, CARROL *et al.* (1984) desenvolveram um importante estudo de base e nesse artigo estudaram a probabilidade de desenvolver uma doença cardíaca. Os autores consideraram que a variável explicativa é sujeita ao erro de medida do tipo estrutural e mostram que se os erros de medida são grandes, ou seja, neste caso quando a variância do erro de medida for três vezes maior do que a variância da variável não observada, as estimativas de probabilidades do evento em questão podem conter erros substanciais, especialmente para grupos de alto risco. Nos casos de erro de medida grande, os autores investigaram estimadores por máxima verossimilhança condicional e suas propriedades, concluindo que à medida que a variância da variável não observada aumenta, a função de verossimilhança condicional melhora para a correção dos erros de medida.

O conceito de erro de classificação é conhecido por ser um grande problema em estudos epidemiológicos, devido à ocorrência da classificação incorreta de respostas binárias. Para investigar os efeitos desses erros, SPOSTO *et al.* (1992) descreveram uma análise estatística com o objetivo de analisar os riscos estimados em apresentar câncer ou não, devido à exposição a radiação dos sobreviventes da bomba atômica de Hiroshima e Nagasaki; também avaliaram se a variável observada, referente à não mortalidade por câncer, pode ser atribuída ao erro de classificação. As probabilidades de misclassificação (erro de classificação) são descritas por funções estimadas, a partir dos dados de autópsias dos membros do LSS (Life Span Study). Esse artigo apresentou uma aplicação dos tipos de métodos para a classificação incorreta do diagnóstico de mortalidade ou não-mortalidade, através do modelo de Poisson para dados de sobrevivência quanto a análise da variável resposta aos sobreviventes da bomba.

HEID *et al.* (2004) enfatizaram a importância da diferença entre o erro clássico e o erro do tipo Berkson, como também mostraram a diferença entre um modelo aditivo e um modelo multiplicativo. O erro clássico surge quando uma quantidade é medida por algum dispositivo/instrumento e as medições repetidas variam em torno do valor verdadeiro. Já o erro do tipo Berkson é utilizado quando a média do grupo é atribuída a cada indivíduo, adequando as

características do grupo. Os autores definiram claramente os preditores clássicos, e os preditores definidos usando dois exemplos de estudos alemães de caso-controle sobre câncer de pulmão e o caso de exposição aleatória ao gás radioativo, radônio. Nesse trabalho, os autores lidaram com o erro sendo variável aleatória, não-diferencial e homoscedástico. Ao enfatizar tais diferenças, os autores concluíram que a avaliação da exposição não deve ser apenas a mais precisa possível, mas também deve fornecer um modelo para os erros de medida com uma clara diferenciação entre os componentes clássicos, e do tipo Berkson.

A invenção do modelo probit se deu através de BLISS (1934), em que estudou experimentos biológicos do tipo dose-resposta para doses fixas e respostas aleatórias que refletem a distribuição individual de níveis de tolerância. A introdução do termo probit, uma abreviação de *probability unit*, é apresentada em BLISS (1935) como uma forma mais conveniente de expressar desvios da média de uma distribuição Normal (DINIZ, 2015).

BERKSON (1944) propôs o uso do modelo logístico em experimentos biológicos e assinalou o termo logit, em analogia à probit. Berkson alegava que estimações pelo método dos Mínimos Quadrados eram bem mais eficientes do que por Máxima Verossimilhança (BERKSON *et al.*, 1980), porém o fato do modelo probit possibilitar a interpretação baseada em desvios da média de uma distribuição normal, fez com que adeptos do modelo de BLISS (1934) não aceitassem a proposta de Berkson muito bem, mesmo já tendo exposto suas razões para o uso da função de ligação logit (BERKSON, 1951).

EUGENIO e OTHERS (2017) apresentou um modelo de regressão para dados binários com mistura de quatro funções de ligação, incluindo a probit. Os procedimentos de estimação frequentista são expostos e através de estudos de simulações mostrou-se que, em relação a outros modelos, a função de ligação proposta apresenta melhor desempenho nas estimações de proporções, ao passo que para previsões é igual às demais. Sua flexibilidade em poder ser tanto uma função de ligação simétrica quanto assimétrica é corroborada pelo resultados das análises de três bancos de dados reais, bem como pelas simulações. Mostrou-se ainda um caso em que, por não conseguir obter melhores resultados com as combinações de ligações, a mistura associa peso total a um de seus componentes.

Os avanços constantes da tecnologia tornaram os processos de mensuração cada vez mais precisos. Porém, não é realista supor que as variáveis são medidas sem erro e, muitas vezes, é custoso ter acesso aos seus verdadeiros valores. Neste trabalho faremos uma extensão do estudo realizado por ROY, BANERJEE e MAITI (2005) que consideraram que o erro de medida segue distribuição normal. Consideramos que o erro de medida segue uma distribuição normal

assimétrica. O modelo abordado será o modelo de regressão binária, na qual a variável resposta está sujeita ao erro de classificação e as covariáveis ao erro de medida do tipo Berkson. Assim, é necessário um valor substituto observado, W, para o valor verdadeiro não observado X.

1.1 OBJETIVOS

1.1.1 Objetivo Geral

O objetivo geral deste trabalho é propor modelos binários, considerando que o preditor de risco contínuo está sujeito ao erro de medida e que a variável resposta é sujeita ao erro de classificação. Como pressuposto do problema, tem-se que o erro de medida é do tipo Berkson e segue uma distribuição normal assimétrica.

1.1.2 Objetivos Específicos

A fim de alcançar o objetivo geral, os objetivos específicos deste trabalho são:

a) Investigar, a eficiência do uso do modelo proposto em relação aos modelos existentes. Os modelos considerados são:

   Modelo 1: Modelo sem erro de classificação e sem erro de medida (Naive);

   Modelo 2: Modelo incorporando erro de classificação (ROY, BANERJEE e MAITI (2005));

   Modelo 3: Modelo incorporando erro de medida (modelo proposto);

   Modelo 4: Modelo com erro de classificação e erro de medida (modelo proposto), que tem como submodelos os modelos 1, 2 e 3.

b) Mostrar a aplicabilidade do modelo proposto (modelo 4) em conjunto de dados reais.

1.1.3 Suporte e aspectos computacionais

Para o desenvolvimento computacional dessa dissertação, foi utilizada a linguagem de programação R, através do software *R Studio* (versão 3.6.1) no desenvolvimento do código para o estudo de simulação, como também para análise de dados e produção de gráficos. É uma liguagem *open source*, ou seja, sem custos e está disponível para diferentes sistemas operacionais como Unix/Linux, Mac e Windows.

O desenvolvimento textual foi realizado com auxílio do sistema tipográfico LaTeX[1],

[1] Para mais informações sobre o LaTeX, acesse *https://www.latex-project.org/*.

que é um sistema de preparação de documentos e inclui recursos projetados para a produção de documentação técnica e científica.

Vale ressaltar que para esse estudo de simulação foi necessário a utilização de um servidor. Como a Universidade não possui uma estrutura para essa finalidade e por conta de tempo, foi preciso desenvolver uma máquina virtual na *Google Cloud Platform (GCP)* por conta própria e dentro dessa máquina foi instalado o *R Studio Server*[2] (versão 1.2.5019) para rodar as simulações. Nos testes realizados inicialmente, o tempo computacional estava muito alto ao rodar $R = 500$ e $n = 10.000$, chegando a registrar mais de uma semana para concluir a simulação para apenas um modelo. Então para tentar solucionar esse problema, foi utilizado a paralelização que reduziu esse tempo pela metade. Para abrir uma conta na GCP foi gratuito e para testar as funcionalidades, a plataforma disponibilizou US$ 300 para teste.

1.1.4 Organização da dissertação

Além deste capítulo introdutório que se inicia com um breve referencial teórico sobre erros de medida, erros de classificação e a distribuição normal assimétrica, esta dissertação apresenta mais quatro capítulos. No Capítulo 2 é apresentado os tipos de modelos com erro de medida e com erro de classificação, além das definições do modelo com erro de medida, distribuição normal assimétrica e suas propriedades. É apresentado o modelo de regressão binária com erro de classificação e erro de medida normal assimétrico que é a proposta neste trabalho. O Capítulo 3 apresenta o estudo de simulação para os quatro modelos que foram mencionados nos objetivos específicos. Já no Capítulo 4, é apresentada uma aplicação com dados reais que foram ajustados nos quatro modelos mencionados no decorrer do trabalho (Modelo Naive, modelo probit com erro de medida, modelo probit com erro de classificação e modelo probit com erro de classificação e erro de medida). Por fim, no Capítulo 5 encontram-se as considerações finais do estudo e trabalho futuro.

[2] Para mais infomações sobre a GCP e sobre o RStudio server acesse: https://cloud.google.com/ e https://rstudio.com/products/rstudio/download-server/.

2 MODELOS COM ERRO DE MEDIDA E ERRO DE CLASSIFICAÇÃO

Neste capítulo será definido os tipos de erros de medida: estrutural, ultraestrutural, funcional e Berkson. Como também serão apresentados os modelos com erro de classificação.

2.1 MODELOS COM ERRO DE MEDIDA

Os modelos de regressão com erros de medida geralmente, são definidos de maneira que a variável resposta Y_t seja uma função de preditores ou covariáveis em que as medições estão sujeitas a erros de medida.

Ao falar de regressão binária, estamos considerando que a variável resposta Y_t tem como valores 0 ou 1. O problema de erro de medida na regressão binária está associado à variável independente X que, muitas vezes, é muito custoso ou até mesmo impossível mensurá-la. No lugar do valor verdadeiro da variável preditora, X, considera-se o valor de uma variável substituta, W, que está associada a um erro de medida.

Na prática podemos encontrar esta e outras condições em que os erros de medida podem ser considerados. CUNHA e COLOSIMO (2003) relatam algumas dessas condições, como por exemplo:

- Erros provenientes de métodos e técnicas de coleta de dados como entrevista, observação ou questionário. Tais erros podem ser causados por confusão, ignorância, falta de cuidado, por falta de treinamento adequado ou mesmo pelo método usado para obter a resposta;
- Erros por processamento inadequado, como utilização de técnica de análise e processamento de dados de pouca confiança ou não apropriados para o problema estudado;
- Erros por armazenamento com pouca confiabilidade, falhas na entrada dos dados para o processamento, ou perda de informações;
- Erros causados pela incerteza pertencente aos equipamentos de medição.

CARROL *et al.* (2006), consideraram dois tipos de preditores da seguinte maneira: $\mathbf{X_1}$ representa o vetor de preditores que, para todos os efeitos práticos, são medidos sem erros; e $\mathbf{X_2}$ aqueles que não podem ser observados, exatamente, para todos os objetos de estudo. Logo, pode-se observar uma variável $\mathbf{W}$ que está relacionada à variável $\mathbf{X_2}$ não observável. Os parâmetros do modelo que relacionam $\mathbf{Y}$ e $(\mathbf{X_1}, \mathbf{X_2})$ não podem ser estimados diretamente, ajustando $\mathbf{Y}$ em relação à $(\mathbf{X_1}, \mathbf{X_2})$, pois $\mathbf{X_2}$ não é observado. Logo, o objetivo da modelagem com erros de medida é obter boas estimativas ao ajustar um modelo que relaciona $\mathbf{Y}$ em termos de $(\mathbf{X_1}, \mathbf{W})$. A realização desse objetivo requer uma análise cuidadosa pois, substituir $\mathbf{W}$ por $\mathbf{X_2}$

e não fazer modificações nos métodos de ajuste para essa substituição, pode levar à estimativas tendenciosas.

Quando temos um problema de erro de medida, é preciso especificar o modelo a ser utilizado. Há um modelo em que a variável com erro é única para cada indivíduo como, por exemplo, na medição dos batimentos cardíacos. E se para cada indivíduo do grupo é atribuido os mesmos valores da variável, o modelo utilizado para esse caso será o modelo Berkson. Por exemplo: moradores de determinadas regiões das cidades de Hiroshima e Nagasaki que receberam a mesma exposição à radiação, devido ao bombardeio ocorrido, mas a verdadeira exposição é particular a cada indivíduo. As consequências dos erros de medida têm recebido muita atenção na literatura. Devido a isso, os erros de medida do tipo Berkson vem sendo estudado por muitos autores como WHITTEMORE e KELLER (1988), REEVES, COX e DARBY (1998), KIM, YASUI e BURSTYN (2006), MCGLOTHLIN, STAMEY e SEAMAN (2008), entre outros. Vale ressaltar que a presença de erro de medida pode fazer com que as estimativas dos parâmetros de um modelo que não assume erro de medida sejam viesadas, o que pode levar a conclusões errôneas.

Existem situações em que a relação entre a variável observada e a variável não observada pode ser aditiva ($W = X + \delta$) ou multiplicativa ($W = X\delta$). CARROL *et al.* (2006) apresentam um exemplo com dados de sobrevida às explosões da bomba de Hiroshima e Nagasaki. Nesse caso, a variável resposta é a aberração cromossômica e a verdadeira dose de radiação, X, não pode ser medida diretamente. Portanto, em seu lugar considera-se a variável substituta W. Ainda, é considerado que $W = 0$ se e somente se $X = 0$. Além disso, considera-se que se a variável X_t é positiva, essa tem distribuição Weibull. Logo, o modelo multiplicativo proposto é dado por $W = X\delta$, em que $\log(\delta) \sim N(\mu_\delta, \sigma_\delta^2)$.

Assim, uma suposição incorreta sobre o modelo com o erro de medida pode causar tantos problemas quanto ignorá-lo. Desse modo, a identificação correta da relação entre a variável não observada e a variável observada com erro é essencial para o sucesso do uso de modelos com erro de medida (STEFANSKI, 2000).

Estudamos modelos de regressão em que não é possível observar X_t diretamente (FULLER, 1987). Então, ao invés de observar X_t, observamos a variável W_t que é composta pela variável não observada X_t mais o erro de medida δ_t. Assim, temos

$$W_t = X_t + \delta_t, \quad t = 1, 2, \ldots, n. \tag{2.1}$$

Podemos definir os tipos de erros de medida sob algumas suposições.

1. **Estrutural**: no erro de medida do tipo estrutural, consideramos uma distribuição de probabilidade para a covariável não observada. Ou seja, X_t são variáveis aleatórias independentes e identicamente distribuídas, sendo X_t independente de δ_t.
2. **Ultraestrutural**: no erro de medida do tipo ultraestrutural, proposto por DOLBY (1976), temos que X_t são variáveis aleatórias independentes, assim como no estrutural, porém não são identicamente distribuídas.
3. **Funcional**: já no erro de medida do tipo funcional, valores desconhecidos de X_t são vistos como parâmetros. Nesse caso, o número de parâmetros cresce com o tamanho da amostra e dizemos que há parâmetros incidentais no modelo (CARRASCO, 2012).

Um importante modelo de regressão com erros de medida foi introduzido por BERKSON (1950) e utilizado para muitas aplicações. Neste tipo de erro, consideramos a variável não observada (X_t) como a soma da variável observada (W_t) e o erro de medida (δ_t).

Assim, o **erro de medida do tipo Berkson** é definido por

$$X_t = W_t + \delta_t^*, \tag{2.2}$$

onde W_t é pré-fixado (ACHIC; SANDOVAL; YOSHIDA, 2007), δ_t^* e W_t são independentes.

Logo, o modelo definido por

$$Y_t = \beta_0 + \beta_1 X_t + e_t \quad (t = 1,2,...,n),$$

juntamente com a Equação (2.2) e suas suposições descritas acima, é conhecido como modelo de regressão com erros de medida do tipo Berkson (CHENG; NESS, 1999). Assim temos que

$$\begin{aligned} Y_t &= \beta_0 + \beta_1(W_t + \delta_t^*) + e_t \\ &= \beta_0 + \beta_1 W_t + (e_t + \beta_1 \delta_t^*). \end{aligned}$$

em que o erro, e_t, são variáveis aleatórias independentes com média zero e variância σ_e^2; δ_t^* são variáveis aleatórias independentes com média zero e variância σ_δ^2; e_t e δ_s^* tem correlação igual a zero, para todo t e s.

Vejamos, a seguir, um exemplo aplicado envolvendo o erro de medida do tipo Berkson.

Exemplo:

Como citado acima, em alguns experimentos é muito custoso mensurar a variável preditora X. Devido a isso, uma variável substituta W é considerada. Como no exemplo usado por WANG e OTHERS (2004), imagine um estudo epidemiológico que investiga a severidade de uma doença pulmonar, Y, nos residentes de uma cidade em relação a quantidade de poluentes do ar, X. Suponha que os poluentes são medidos por certas estações de observação espalhadas pela cidade. A exposição atual dos residentes aos poluentes X, pode variar aleatoriamente nos valores medidos W, nas estações de observação. Nesse caso, X pode ser representado por W acrescido de um erro aleatório. Então, um modelo razoável para representar os erros de medição é o modelo apresentado em (2.2).

2.1.1 Erros de medida não-diferenciais

Para uma melhor contextualização do tema, é importante apontar o conceito de erro de medida não-diferencial. O erro de medida não-diferencial ocorre se a distribuição de $Y|(\mathbf{X_1}, \mathbf{X_2}, W)$ depende somente de $(\mathbf{X_1}, \mathbf{X_2})$, ou seja, quando W não traz nenhuma informação sobre Y além das que estão disponíveis em $\mathbf{X_1}$ e $\mathbf{X_2}$. Logo, podemos considerar a variável W como uma substituta de $\mathbf{X_1}$, se for condicionalmente independente da variável resposta, dadas as verdadeiras covariáveis.

2.2 DISTRIBUIÇÃO NORMAL ASSIMÉTRICA

A distribuição normal assimétrica (ou *skew normal*) tem sido amplamente utilizada nos mais diversos problemas práticos em que, usualmente, se utiliza a distribuição normal (GONDIM, 2011). Em muitas situações, a suposição de normalidade não é satisfeita devido à falta de simetria da distribuição dos dados. Portanto, alguns autores propõem como alternativa a utilização de uma família mais geral de distribuições, a fim de modelar a assimetria e além de incluir a distribuição normal como um caso particular.

No trabalho de FREITAS (2005) flexibilizou-se a suposição de normalidade, dispondo de novas classes de distribuições assimétricas, em que a primeira classe considerada é a proposta por AZZALINI (1985), que inclui a distribuição normal como um caso particular. A segunda é a dos modelos flexíveis de MA e GENTON (2004) e a terceira é a classe de distribuições normais assimétricas estendidas, proposta por ARELLANO-VALLE e AZZALINI (2006). Com esse objetivo, métodos de inferência baseados na abordagem bayesiana, foram desenvolvidos para estimar os parâmetros envolvidos no modelo de regressão linear simples com

erros assimétricos.

SANTOS (2012) considerou uma extensão do modelo de resposta ao item (MRI) para o caso de grupos múltiplos. Um dos objetivos foi avaliar o impacto do número de respondentes por grupo, número de itens por grupo, número de itens comuns, assimetria da distribuição do grupo de referência e priori, na recuperação dos parâmetros. Foi analisado um conjunto de dados reais que apresentou indícios de assimetria na distribuição dos traços latentes de alguns grupos. Os resultados obtidos com os modelos confirmam a presença de assimetria na maioria dos grupos. Por fim, comparou-se os modelos simétrico e assimétrico, em que o modelo assimétrico se ajustou melhor aos dados segundo todos os critérios.

Já SANTOS, SCALON e OZAKI (2014) proporam a distribuição normal-assimétrica como uma alternativa à distribuição normal para modelar a distribuição da produtividade agrícola no Brasil. Foram analisadas séries de produtividade de milho, no período de 1981 a 2007, em 30 municípios do Paraná. A distribuição normal-assimétrica apresentou melhores ajustes do que a distribuição normal para a grande maioria dos municípios e, consequentemente, acarretou melhores estimativas para o pagamento esperado do seguro agrícola.

Essas aplicações da normal assimétrica vêm de AZZALINI (1985) que introduziu formalmente a distribuição, apresentando suas propriedades e mostrou os possíveis problemas na estimação do parâmetro de assimetria pelos métodos dos momentos e por máxima verossimilhança. A definição a seguir é o caso geral da distribuição e é uma extensão da distribuição normal com a adição de um parâmetro de assimetria.

Definição 1. *Seja Z uma variável aleatória com distribuição normal assimétrica, com parâmetro de locação* $\mu \in \mathbb{R}$*, parâmetro de escala* $\sigma^2 > 0$ *e parâmetro de assimetria* $\lambda \in \mathbb{R}$*. A função densidade de probabilidade (fdp) de Z é dada por*

$$f(z) = (2/\sigma)\phi\{(z-\mu)/\sigma\}\Phi\{(\lambda(z-\mu)/\sigma)\}, \quad z \in \mathbb{R} \tag{2.3}$$

em que $\phi(z)$ *e* $\Phi(z)$ *denotam a função densidade de probabilidade e a função de distribuição acumulada (fda) da distribuição normal padrão* $N(0,1)$*, respectivamente. A notação a ser utilizada é* $Z \sim SN(\mu, \sigma^2, \lambda)$*, e indica que Z segue distribuição normal assimétrica.*

Note que quando $\lambda < 0$ indica assimetria negativa, $\lambda > 0$ indica assimetria positiva e se $\lambda = 0$ a função de densidade coincide com a densidade da distribuição normal e portanto, é simétrica (BOLFARINE; LANCHOS, 2007).

A função de densidade na Equação (2.3), quando $\mu = 0$ e $\sigma = 1$, é dada por

$$f(z) = 2\phi(z)\Phi(\lambda z), \quad z \in \mathbb{R}. \tag{2.4}$$

em que é denotada por $SN(\lambda)$.

Na Figura 1 estão expostas as funções de densidade da distribuição normal assimétrica (2.4) para diferentes valores de λ.

A função de distribuição acumulada associada à densidade (2.4) é denotada por $F(z;\lambda)$, e é dada por:

$$\begin{aligned} F(z;\lambda) &= 2\int_{-\infty}^{z}\int_{-\infty}^{\lambda t}\phi(t)\Phi(u)dudt \\ &= 2\Phi_2(z,0|\Omega), \end{aligned} \tag{2.5}$$

com

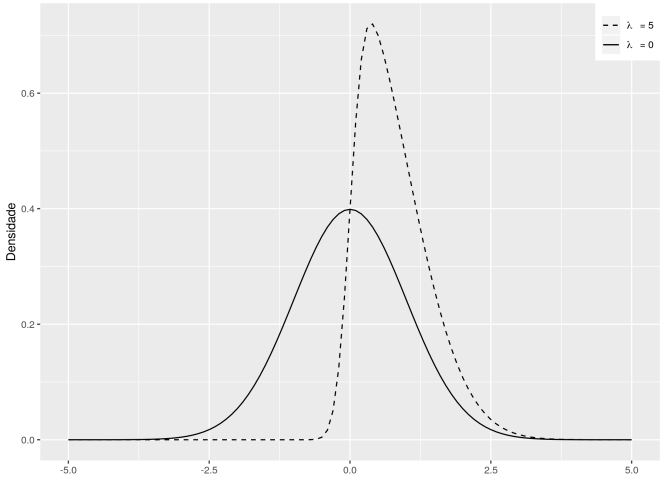
$$\Omega = \begin{bmatrix} 1 & -\omega \\ -\omega & 1 \end{bmatrix}, \quad \omega = \frac{\lambda}{\sqrt{1+\lambda^2}}, \quad z \in \mathbb{R}.$$

Neste caso, $\Phi_2(.|\Omega)$ é a função de distribuição de uma normal bivariada com média zero e matriz de covariâncias Ω (FERREIRA, 2008).

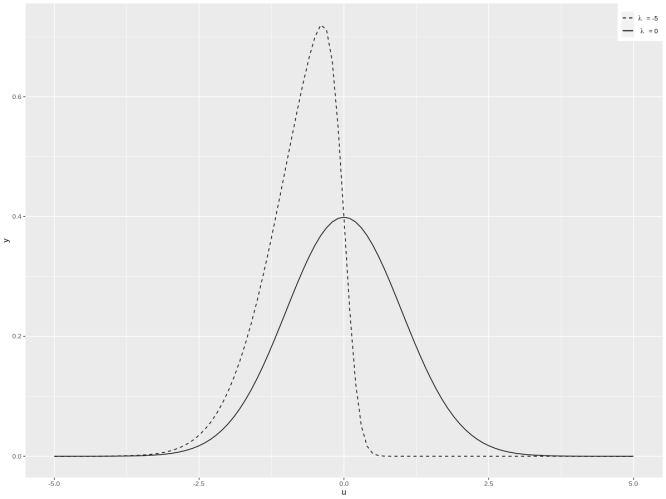
2.2.1 Propriedades da distribuição normal assimétrica

Se Z é uma variável aleatória com distribuição normal assimétrica $SN(0,1,\lambda)$, dada pela Equação (2.4), tem-se as seguintes propriedades:

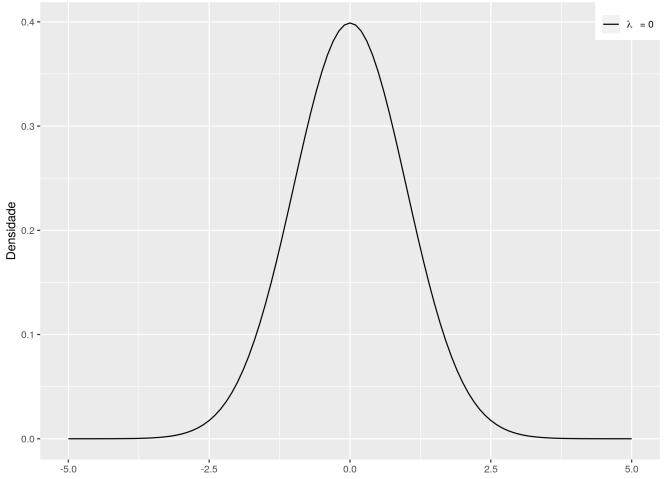
1. Se $\lambda = 0$, tem-se a função de densidade da distribuição $N(0,1)$;
2. Se $\lambda \to \infty$, a densidade da Equação (2.4) converge para a distribuição half-normal $HN(0,1)$;
3. Se $Z \sim SN(\lambda)$ então $-Z \sim SN(-\lambda)$;
4. A função de densidade da Equação (2.4) é unimodal, ou seja, $\log\{f(z)\}$ é uma função côncava;
5. $1 - F(-z;\lambda) = F(z;-\lambda)$;
6. $F(z;1) = \{\Phi(z)\}^2$;
7. $sup|\Phi(z) - F(z;\lambda)| = \pi^{-1}arctan|\lambda|$;



(a) Assimetria positiva ($\lambda = 5$)



(b) Assimetria negativa ($\lambda = -5$)



(c) Assimetria nula ($\lambda = 0$)

Figura 1 – Gráficos da função densidade de probabilidade da distribuição normal assimétrica para diferentes valores de λ.

8. Se $Z \sim SN(\lambda)$, então $Z^2 \sim \chi^2_1$. Onde, χ^2_1 representa a distribuição Qui-quadrado com um grau de liberdade.

As provas dessas propriedades podem ser encontradas em AZZALINI (1985) e AZZALINI (2005).

A função geradora de momentos de $Z \sim SN(\lambda)$ é dada por

$$M_Z(t) = \mathbb{E}(e^{tZ}) = 2\exp\left(t^2/2\right)\Phi(\omega t). \tag{2.6}$$

Logo, a partir da Equação (2.6), tem-se os momentos da normal assimétrica. Assim,

$$E(Z) = \sqrt{\frac{2}{\pi}}\frac{\lambda}{\sqrt{1+\lambda^2}} \text{ e } Var(Z) = 1 - \frac{2}{\pi}\frac{\lambda^2}{1+\lambda^2}.$$

Os coeficientes de assimetria e curtose são definidos, respectivamente, por

$$\gamma_1 = \frac{E\left((Z-E(Z))^3\right)}{E\left((Z-E(Z))^2\right)^{3/2}} \text{ e } \gamma_2 = \frac{E\left((Z-E(Z))^4\right)}{E\left((Z-E(Z))^2\right)^2} - 3.$$

O coeficiente γ_1 caracteriza como e quanto a distribuição se afasta da condição de simetria. É uma função crescente em $|\lambda|$ e se $\lambda = 0$, então $\gamma_1 = 0$. Já o coeficiente γ_2 caracteriza o formato da distribuição em relação ao seu achatamento. Se $\gamma_2 < 0$ é uma distribuição platicúrtica, $\gamma_2 = 0$ é mesocúrtica e $\gamma_2 > 0$ é leptocúrtica.

2.3 MODELOS LINEARES GENERALIZADOS

Para a resolução de problemas em diferentes campos da Ciência, a análise de regressão linear é frequentemente utilizada e tem como principal objetivo estudar a relação linear entre, no mínimo, duas variáveis quantitativas. Ou seja, entre a variável resposta, Y, e uma ou mais variáveis explicativas, X. O modelo linear descreve a variável resposta como soma de uma combinação linear de parâmetros desconhecidos com covariáveis independentes e um erro aleatório. Então, é possível expressar o valor esperado de Y como função de polinômio de maior grau de uma única covariável. O vetor aleatório erro (ε) segue uma distribuição normal $N(0, \sigma^2)$, fazendo com que a variável resposta também tenha distribuição normal (CHARNET *et al.*, 1999).

Os Modelos Lineares Generalizados (MLG) são uma extensão dos modelos normais lineares, e tem por objetivo principal abrir o leque de opções para a distribuição da variável resposta, permitindo que a mesma pertença à família exponencial de distribuições, bem como dar maior flexibilidade para a relação funcional entre a média da variável resposta e o preditor

linear (PAULA, 2004). Desta forma, permite que Y tenha qualquer distribuição, desde que esta pertença à família exponencial. Embora essa classe de modelos também exija independência das observações, problemas frequentes em modelos lineares como falta de normalidade e homoscedasticidade, não ocorrem. O que justifica este fato é que a variância é dada como uma função da média e a adição dos efeitos decorre naturalmente como propriedade das respostas esperadas (MCCULLOCH; NELDER, 1989). Segundo CORDEIRO e DEMÉTRIO (2008), os MLG's são constituídos por uma variável resposta univariada, variáveis explicativas e uma amostra aleatória de n observações independentes, em que

- a variável resposta, **componente aleatório** do modelo, tem uma distribuição oriunda da família exponencial que engloba as distribuições normal, gama e normal inversa para dados contínuos; binomial para proporções; Poisson e binomial negativa para contagens;
- as variáveis explicativas entram na forma de uma estrutura linear, constituindo o **componente sistemático** do modelo;
- a ligação entre os componentes aleatório e sistemático é feita através de uma função adequada como, por exemplo, a função complemento log-log, função *logit*, função *probit*, que são chamadas de **função de ligação**.

Nos modelos lineares, a média é tomada como uma combinação linear de parâmetros. Já no MLG, a relação entre as variáveis explicativas e a média da variável resposta é estabelecida por meio de uma função de ligação que não precisa ser, necessariamente, a função identidade (MCCULLOCH; SEARLE, 2001). Neste trabalho, utilizaremos a função probit que mapeia a média da variável resposta com o preditor linear sob a distribuição binomial.

2.3.1 Modelos para variáveis binárias

2.3.1.1 Distribuição Binomial

A estrutura probabilística da variável de interesse (Y) pode ser representada pela distribuição binomial (TIEPPO, 2007), logo pode ser definida por:

Definição 2. *Sejam $Y_t, t = 1, \ldots, n$ o número de sucessos obtidos em n_t ensaios de Bernoulli independentes. Dizemos que Y_t tem distribuição binomial, cada um com probabilidade de sucesso p_t. Dessa forma, $Y_t \sim Binomial(n_t, p_t)$ com função de probabilidade expressa da forma:*

$$p(y_t) = P[Y_t = y_t] = \binom{n_t}{y_t} p_t^{y_t} (1 - p_t)^{n_t - y_t},$$

com $y_t = 0, \ldots, n$. *A esperança e a variância de* Y_t *são, respectivamente,* $E(Y_t) = p_t$ *e* $Var(Y_t) = n_t p_t (1 - p_t)$.

A função geradora de momentos de Y_t, $M_{Y_t}(k)$ *é dada por:*

$$\begin{aligned} M_{Y_t}(k) &= \mathbb{E}\left(e^{kY_t}\right) \\ &= \sum_{t=0}^{n} e^{ky_t} \binom{n_t}{y_t} p^{y_t} (1-p)^{n_t - y_t} \\ &= \sum_{t=0}^{n} \binom{n_t}{y_t} (e^k p)^{y_t} (1-p)^{n_t - y_t} = (pe^k + 1 - p)^{n_t}. \end{aligned}$$

2.3.1.2 Modelo Probit

A seleção de modelos é uma parte importante da pesquisa. Envolve a procura de um modelo simples, que descreva bem os dados observados. Nesse contexto, há ensaios do tipo dose-resposta, em que uma determinada droga é administrada em t diferentes doses $(x_1, x_2, \ldots, x_t)$, a respectivamente, $n_1, n_2, \ldots, n_t$ indivíduos, obtendo-se como resposta, após um período especificado, $y_1, y_2, \ldots, y_t$ indivíduos que mudam de estado (ocorrência de um sucesso, por exemplo). Suponha que cada indivíduo responde, ou não, à droga, tal que a resposta é binária (isto é, 0 ou 1). Por exemplo, quando um inseticida é aplicado a um determinado número de insetos, eles morrem (1), ou não (0), à dose aplicada. Quando um remédio é administrado a um grupo de pacientes, eles podem melhorar (sucesso - 1), ou não (falha - 0). Dados resultantes desse tipo de ensaio podem ser considerados como provenientes de uma distribuição binomial com parâmetro (p_t), que representa a probabilidade de ocorrência (sucesso) do evento sob estudo, isto é, $Y_t \sim Binomial(n_t, p_t)$. Os objetivos desse tipo de experimento são, em geral, modelar a probabilidade de sucesso p_t como função de variáveis explicativas. Para todo indivíduo, haverá um certo nível de intensidade abaixo do qual a resposta não ocorre e acima do qual ela ocorre, o que é chamado de tolerância. Essa tolerância varia de um indivíduo para outro da população e então, há uma distribuição de tolerâncias, a qual pode-se associar uma variável aleatória U (DEMÉTRIO, 2001). Se a dose x_t é dada para toda a população, e $f(u)$ é a função de densidade para a distribuição de tolerâncias, todo indivíduo cuja tolerância é menor do que x_t responderá à droga, e a probabilidade de que um indivíduo escolhido ao acaso responda à dose é dada por

$$p_t = P(U \leq x_t) = F(x_t) = \int_{-\infty}^{x_t} f(u)du. \tag{2.7}$$

Assim, precisamos de uma função que seja estritamente crescente. Logo, o modelo

probit define-se por:

Seja U uma variável aleatória com distribuição normal de média μ e variância σ^2, isto é,

$$f_U(u;\mu,\sigma^2) = \frac{1}{\sqrt{(2\pi\sigma^2)}} \exp^{-\left(\frac{(u-\mu)^2}{2\sigma^2}\right)}, \quad \mu \in \mathbb{R} \quad \text{e} \quad \sigma^2 > 0.$$

E assim,

$$p_t = P(U \leq x_t) = P\left(\frac{U-\mu}{\sigma} \leq \frac{x_t-\mu}{\sigma}\right) = P\left(Z \leq -\frac{\mu}{\sigma} + \frac{1}{\sigma}x_t\right) = P(Z \leq \beta_0 + \beta_1 x_t),$$

em que $\beta_0 = -\mu/\sigma$ e $\beta_1 = 1/\sigma$.

Portanto, como $Z \sim N(0,1)$, tem-se que $p_t = \Phi(\beta_0 + \beta_1 x_t)$, onde $\Phi(.)$ representa a função de distribuição acumulada da normal padrão e é uma função não-linear em um conjunto linear de parâmetros. Para linearizar, tem-se

$$probit(p_t) = \Phi^{-1}(p_t) = \beta_0 + \beta_1 x_t. \tag{2.8}$$

2.4 ESTIMAÇÃO DOS PARÂMETROS

Para a estimação dos parâmetros dos modelos foi considerado o método da máxima verossimilhança onde as estimativas dos parâmetros são obtidas maximizando o logaritmo da função de verossimilhança. Nesta seção, serão apresentados conceitos preliminares sobre o princípio de máxima verossimilhança considerando conceitos de inferência estatística (BOLFARINE; SANDOVAL, 2010).

2.4.1 Amostras, Estatísticas e Estimadores

Definição 3. *Uma sequência $Y_1,\ldots,Y_n$ de n variáveis aleatórias independentes e identicamente distribuidas (i.i.d) com função de densidade (f.d.p.) ou, no caso discreto, função de probabilidade (f.p) $f(y|\theta)$ é dita ser uma amostra aleatória de tamanho n da distribuição de Y. Nesse caso, temos,*

$$f(y_1,\ldots,y_n|\theta) = \prod_{i=1}^{n} f(y_i|\theta) = f(y_1|\theta)\ldots f(y_n|\theta).$$

Então, a partir dessa definição, usamos a amostra $Y_1,\ldots,Y_n$ para obter informação sobre o parâmetro θ.

Para as definições a seguir, considere $Y_1,Y_2,\ldots,Y_n$ uma amostra aleatória com função de densidade $f(y;\theta)$, tal que θ é um parâmetro pertencente ao espaço paramétrico $\Theta \subset \mathbb{R}^r, r \in \mathbb{N}$. Assume-se que o valor de θ é desconhecido, dessa forma, o objetivo é estimá-lo.

Definição 4. *Qualquer função da amostra que não depende de parâmetros desconhecidos é denominada uma estatística. Logo, um estimador de θ é uma estatística $\hat{\theta}_n = W(Y_1,\ldots,Y_n)$, com $W : \mathbb{R}^n \to \Theta$.*

Definição 5. *A função de verossimilhança de um parâmetro θ baseada nas observações $y_1,\ldots,y_n$ da amostra $Y_1,\ldots,Y_n$ é dada por*

$$L(\theta;y_1,\ldots,y_n) = \prod_{i=1}^{n} f(y_i;\theta)$$

Definição 6. *O Estimador de máxima verossimilhança (EMV) de um parâmetro θ é o valor $\hat{\theta}$ que maximiza a função de verossimilhança $L(\theta) = L(\theta;y_1,\ldots,y_n)$, caso exista. Ou seja, é um valor $\hat{\theta} = \hat{\theta}_n = \hat{\theta}(y_1,\ldots,y_n) \in \Theta$ tal que*

$$L(\hat{\theta}) = \underset{\theta\in\Theta}{\operatorname{argmax}} L(\theta;y_1,\ldots,y_n) = \underset{\theta\in\Theta}{\operatorname{argmax}} \prod_{i=1}^{n} f(y_1;\theta),$$

em que $(y_1,\ldots,y_n)$ são observações da amostra $Y_1,\ldots,Y_n$.

Observações sobre os EMV's:

- É possível que $\hat{\theta}$ não exista ou pode existir mas não ser único.
- A função logarítmica é monótona estritamente crescente. Dessa forma, com o intuito de simplificar o processo para obter o EMV de θ, ele pode ser calculado maximizando a função de log-verossimilhança de θ, dada por

  $$l(\theta) = \ln L(\theta) = \sum_{i=1}^{n} \ln f(y_i;\theta)$$

  .

  Se existir, o valor que maximiza a função $l(\theta)$ será o mesmo que maximiza $L(\theta)$, ou seja, ele será o estimador de máxima verossimilhança desejado.
- O valor do EMV de θ pode ser obtido resolvendo as equações de verossimilhança

  $$\frac{\partial l(\theta)}{\partial \theta_i} = 0, i = 1,\ldots,n$$

  desde que $l(\theta)$ seja diferenciável em Θ e $\hat{\theta}$ seja um máximo local.

Definição 7. *O erro quadrático médio (EQM) de um estimador $\hat{\theta}$ do parâmetro θ é dado por*

$$EQM = E[(\hat{\theta} - \theta)^2].$$

Tem-se que

$$EQM[\hat{\theta}] = Var[\hat{\theta}] + B^2(\hat{\theta}),$$

em que

$$B(\hat{\theta}) = E[\hat{\theta}] - \theta$$

*é denominado o viés do estimador $\hat{\theta}$. Um estimador é dito ser **não viesado** para θ se $E[\hat{\theta}] = \theta$, para todo $\theta \in \Theta$, ou seja, $B(\hat{\theta}) = 0$ para todo $\theta \in \Theta$.*

*Se $\lim_{n\to\infty} B(\hat{\theta}) = 0$ para todo $\theta \in \Theta$, tem-se que o estimador $\hat{\theta}$ é **assintoticamente não viesado** para θ. No caso em que $\hat{\theta}$ é um estimador não viesado para θ, então $EQM[\hat{\theta}] = Var[\hat{\theta}]$, isto é, o erro quadrático médio de $\hat{\theta}$ se reduz à variância.*

Definição 8. *Um estimador $\hat{\theta}_n$ é dito um estimador consistente de θ se a sequência $\{\hat{\theta}_n\}$ converge em probabilidade para θ, isto é, se para todo $\varepsilon > 0$ arbitrário, temos que*

$$\lim_{n\to\infty} P(|\hat{\theta}_n - \theta| < \varepsilon) = 1.$$

Definição 9. *Considere uma única observação Y com função densidade de probabilidade $f(y|\theta)$. A medida de informação esperada de Fisher de θ através de Y é definida como:*

$$I(\theta) = E\left[-\frac{\partial^2 \log f(y|\theta)}{\partial \theta^2}\right] \tag{2.9}$$

No caso de um vetor paramétrico $\Theta = (\theta_1, \ldots, \theta_k)$ define-se a matriz de informação esperada de Fisher de Θ através de Y como:

$$I(\Theta) = E\left[-\frac{\partial^2 \log f(y|\Theta)}{\partial \Theta \partial \Theta'}\right]$$

Para o modelo em estudo, de maneira geral, a função de verossimilhança é definida por:

$$L(\theta) = \prod_{t=1}^{n} p_t^{y_t} (1-p_t)^{1-y_t}. \tag{2.10}$$

Deste modo, aplicando o logaritmo na função de verossimilhança, obtemos a função de log-verossimilhança que é dada por

$$l(\theta) = \sum_{t=1}^{n} \{y_t \log(p_t)\} + \sum_{t=1}^{n} \{(1-y_t)\log(1-p_t)\}, \tag{2.11}$$

em que cada p_t é definido de acordo com o modelo correspondente. A seguir, serão apresentados os quatro modelos a serem estudados nos capítulos seguintes. Para cada modelo temos uma função de log-verossimilhança definida abaixo.

Modelo 1 (M1): Modelo sem erro de medida e sem erro de classificação (modelo Naive)

$$l(\theta) = \sum \left(y_t \log(\Phi(\beta_0 + \beta_1 W))\right) + \sum \left((1-y_t)\log(1-\Phi(\beta_0+\beta_1 W))\right), \tag{2.12}$$

em que $\theta = (\beta_0, \beta_1)^\top$.

Modelo 2 (M2): Modelo somente com erro de classificação, que também foi estudado em ROY, BANERJEE e MAITI (2005).

$$\begin{aligned} l(\theta) = &\sum \left(y_t \log(\pi_0 + (1-\pi_0-\pi_1)\Phi(\beta_0+\beta_1 W))\right) \\ &+ \sum \left((1-y_t)\log(1-\pi_0+(1-\pi_0-\pi_1)\Phi(\beta_0+\beta_1 W))\right), \end{aligned} \tag{2.13}$$

em que $\theta = (\beta_0, \beta_1, \pi_0, \pi_1)^\top$.

Modelo 3 (M3): Modelo somente com erro de medida, novo modelo que é a proposta deste trabalho.

$$\begin{aligned} l(\theta) = &\sum \left(y_t \log(E_{X|W}\{\Phi(\beta_0+\beta_1 W)\})\right) \\ &+ \sum \left((1-y_t)\log(1-E_{X|W}\{\Phi(\beta_0+\beta_1 W)\})\right), \end{aligned} \tag{2.14}$$

em que $\theta = (\beta_0, \beta_1)^\top$.

Modelo 4 (M4): Modelo com erro de medida e com erro de classificação, novo modelo que é a proposta deste trabalho.

$$\begin{aligned} l(\theta) = &\sum \left(y_t \log(\pi_0 + (1-\pi_0-\pi_1)E_{X|W}\{\Phi(\beta_0+\beta_1 W)\})\right) \\ &+ \sum \left((1-y_t)\log(1-(\pi_0+(1-\pi_0-\pi_1)E_{X|W}\{\Phi(\beta_0+\beta_1 W)\}))\right), \end{aligned} \tag{2.15}$$

em que $\theta = (\beta_0, \beta_1, \pi_0, \pi_1)^\top$.

2.5 CRITÉRIO DE SELEÇÃO DE MODELOS

Um tópico importante na análise de dados, do ponto de vista estatístico, é a escolha do modelo apropriado (BOZDOGAN, 1987). Logo, deve-se selecionar algum critério para decidir qual modelo se ajusta melhor ao banco de dados. Ao selecionarmos modelos é preciso ter em mente que não existem modelos verdadeiros. Há apenas modelos aproximados da realidade que, causam perda de informações. Deste modo, é necessário fazer a seleção do "melhor" modelo, dentre aqueles que foram ajustados, para explicar o fenômeno sob estudo (EMILIANO *et al.*, 2010).

Para a seleção de modelos podemos utilizar alguns critérios de seleção amplamente conhecidos e disponíveis na literatura, como AIC (AKAIKE, 1974) e BIC (SCHWARZ *et al.*, 1978). Eles são utilizadas quando os modelos são encaixados ou autoregressivos e são úteis quando comparamos dois ou mais modelos.

O critério de informação de Akaike (AIC) e o critério de informação bayesiano (BIC) são dados por:

$$AIC = -2\hat{l} + 2K, \tag{2.16}$$

$$BIC = -2\hat{l} + K\log(T), \tag{2.17}$$

em que $\hat{l}$ é a função de log-verossimilhança maximizada, K é o número de parâmetros do modelo e T é o tamanho da amostra.

2.6 MODELO DE REGRESSÃO BINÁRIA COM ERRO DE CLASSIFICAÇÃO E ERRO DE MEDIDA NORMAL ASSIMÉTRICO

Em estudos estatísticos, modelos de regressão são amplamente utilizados em áreas como economia, saúde e psicologia. O objetivo do estudo de modelos de regressão é estudar a relação entre variáveis explicativas e variáveis de interesse.

Em alguns tipos de experimentos, temos uma variável de interesse que tem comportamento binário (sucesso ou fracasso, sim ou não, 0 ou 1, A ou B). Esses tipos de dados podem ser encontrados em experimentos que tem por objetivo analisar

- Um tratamento médico, em que o paciente está: doente ou não doente;
- Em uma pesquisa de opinião, o cliente pode estar satisfeito ou não satisfeito;
- No esporte: vitória ou derrota.

Essas variáveis de interesse podem ser associadas a outras variáveis explicativas. Se modelarmos essa relação podemos ganhar conhecimentos importantes sobre os experimentos de interesse. Porém em alguns casos, como citados anteriormente, pode ocorrer erro na classificação da variável de interesse.

Devido à ocorrência da classificação incorreta de respostas binárias, o conceito de erro de classificação é conhecido por ser um grande problema em estudos epidemiológicos. Os métodos para análise desses dados consideram que a resposta binária é medida sem erro, porém na prática nem sempre isso acontece. Deste modo, a misclassificação é toda classificação incorreta de determinado indivíduo, segundo a medida da característica (KLEIN; COSTA, 1987). Alguns exemplos destas circunstâncias são:

- Em um estudo de caso-controle, um fumante, exposto ao fator de risco, pode ser classificado como não exposto por erro de registro do entrevistador;
- Quando a variável resposta indica ausência ou presença de uma condição médica, ela pode ser identificada por um teste de diagnóstico que, em geral, é imperfeito;
- Um indivíduo que venha a manifestar os sintomas de uma doença, pode ser erroneamente classificado como não doente, em um estudo de coorte, por defeito na técnica de diagnóstico.

Logo, suponha que a variável resposta binária verdadeira é denotada por Y_T, e a variável resposta observada Y. Esse erro está associado à variável resposta binária quando:

- Foi observada $Y = 0$, mas na verdade $Y_T = 1$;
- Foi observada $Y = 1$, mas na verdade $Y_T = 0$;

Sejam π_0 e π_1 as probabilidades de misclassificação e considerando erros de classificação não-diferenciais, tem-se:

$$\begin{aligned} P(Y = 1|Y_T = 0, X) &= P(Y = 1|Y_T = 0) = \pi_0 \\ P(Y = 0|Y_T = 1, X) &= P(Y = 0|Y_T = 1) = \pi_1. \end{aligned} \tag{2.18}$$

Assim, temos que

$$P(Y_T = 1|X) = H\{m(X, \beta)\} = \Phi(\beta_0 + \beta_1 X), \tag{2.19}$$

em que $H(\cdot)$ é uma função de ligação, que neste trabalho será utilizado a função probit; $m(\cdot)$ é uma função linear; X é o preditor que está sujeito ao erro de medida do tipo Berkson; e $\beta = (\beta_0, \beta_1)^\top$ é o vetor de parâmetros da regressão.

Tem-se que X é não observada e W é a variável substituta. Será considerado o erro não-diferencial

$$f(Y_T|X,W) = f(Y_T|X). \tag{2.20}$$

Pelas Equações (2.18) e (2.20) tem-se que

$$P(Y=1|X) = \pi_0 + (1-\pi_0-\pi_1)\Phi(\beta_0+\beta_1 X). \tag{2.21}$$

Considerando que a variável X está sujeita a erro de medida do tipo Berkson,

$$X|W \sim SN(l(W), v(W), \lambda), \tag{2.22}$$

onde $l(W)$ e $v(W)$ são parâmetros conhecidos de locação e escala, respectivamente. Quando $\lambda = 0$, tem-se o modelo desenvolvio por ROY, BANERJEE e MAITI (2005), em que a Equação (2.22) torna-se a distribuição normal com média $l(W)$ e variância $v(W)$.

Considerando a suposição de não-diferenciabilidade (2.20) e a suposição (2.22), a partir da Equação (2.21) temos o modelo com erro de classificação e erro de medida do tipo Berkson com distribuição normal assimétrica

$$\begin{aligned} P(Y=1|W) &= \pi_0 + (1-\pi_0-\pi_1)\int_{-\infty}^{+\infty} \Phi(\beta_0+\beta_1 X) f(X|W)dx \\ &= \pi_0 + (1-\pi_0-\pi_1) E_{X|W}\{\Phi(\beta_0+\beta_1 X)\}, \end{aligned} \tag{2.23}$$

em que $f(X|W)$ é a densidade da distribuição normal assimétrica dada em (2.22) e $E_{X|W}$ é dada por

$$\begin{aligned} E_{X|W}\{\Phi(\beta_0+\beta_1 X)\} &= 2\int \Phi(\beta_0+\beta_1 X)\phi[\{x-l(W)\}/\sqrt{v(W)}]\Phi[\lambda\{x-l(W)\}/\sqrt{v(W)}]dx \\ &= 2\int \Phi_2(\Gamma x|-\mu(W), I_2)\phi[\{x-l(W)\}/\sqrt{v(W)}]dx \\ &= 2\Phi_2\{0|-\mu(W)-\Gamma l(W), I_2+v(W)\Gamma\Gamma^T\} \\ &= 2\Phi_2\{\mu(W)+\Gamma l(W)|0, I_2+v(W)\Gamma\Gamma^T\}, \end{aligned}$$

em que Φ_2 é a função de distribuição acumulada na distribuição normal bivariada,

$$\Gamma = \begin{pmatrix} \beta_1 \\ \lambda/\sqrt{v(W)} \end{pmatrix}, \quad \mu(W) = \begin{pmatrix} \beta_0 \\ -\{\lambda l(W)\}/\sqrt{v(W)} \end{pmatrix} \tag{2.24}$$

e I_2 é a matriz identidade 2×2.

A integração acima foi resolvida usando resultados que podem ser encontrados em ARELLANO-VALLE e GENTON (2005).

3 ESTUDO DE SIMULAÇÃO

Quando se utiliza um modelo matemático para descrever um sistema, é possível que o modelo seja complexo demais, ou então, não permita uma solução analítica. Nesse caso, a simulação computacional pode ser considerada uma ferramenta de grande valia na obtenção de uma resposta para um problema particular. Quando o modelo envolve amostragem aleatória de uma distribuição probabilística, o método designado é a simulação de Monte Carlo (DONATELLI; KONRATH, 2005).

A simulação de Monte Carlo é usada para avaliar desempenhos de procedimentos inferenciais. A importância dos estudos de simulação se deve ao fato de replicarmos os resultados obtidos computacionalmente.

O método de Monte Carlo é um método numérico estatístico usado para aproximar expressões matemáticas complexas e custosas de serem avaliadas com precisão. O método também fornece soluções aproximadas para uma grande variedade de problemas matemáticos que permitem a realização de experiências com amostras de números pseudo-aleatórios gerados em um computador.

Em estatística, é muito utilizado para a avaliação do desempenho de estimadores (pontuais e intervalares) e testes de hipóteses, usando como aporte teórico a Lei Fraca dos grandes números: a média amostral converge em probabilidade para a média populacional. Isso se dá pelo fato de que geramos de forma independente uma amostra finita de uma variável aleatória e para valores maiores de n a média dessa amostra tende a média da população, μ. Se $X_1, \ldots, X_n$ são variáveis aleatórias independentes tais que $\mathrm{E}(X_i) = \mu$ e $\mathrm{Var}(X_i) = \sigma^2 (< \infty)$ para $i = 1, \ldots, n$, então,

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i \xrightarrow{P} \mu,$$

isto é, a média amostral converge em probabilidade para a média populacional (JAMES, 1996).

A simulação de Monte Carlo foi utilizada como ferramenta para verificação de algumas propriedades assintóticas do modelo proposto apresentado neste trabalho.

3.1 SIMULAÇÃO DE MONTE CARLO: PARA DIFERENTES VALORES DE PARÂMETROS

Nesta seção, são realizadas simulações de Monte Carlo para estudar as propriedades dos estimadores dos parâmetros de regressão sujeitos aos erros de classificação e erros de medida. São considerados diferentes cenários de simulação, os quais são aplicados para cada modelo que foi descrito nos objetivos específicos e no Capítulo 2 deste trabalho.

Os cenários considerados são:

Cenário 1: $\pi_0 = 0,01, \pi_1 = 0,01, \beta_0 = 0, \beta_1 = 1, \lambda = 0,001, \sigma^2 = 0,01$

Cenário 2: $\pi_0 = 0,05, \pi_1 = 0,05, \beta_0 = 0, \beta_1 = 1, \lambda = 0,001, \sigma^2 = 0,5$

Cenário 3: $\pi_0 = 0,05, \pi_1 = 0,05, \beta_0 = 0, \beta_1 = 1, \lambda = 2, \sigma^2 = 0,5$

Nos dois primeiros cenários são avaliados os efeitos ao considerar a quase ausência de assimetria ($\lambda = 0,001$), e o aumento das probabilidades de misclassificação (de $0,01$ para $0,05$) e erro de medida ($0,01$ para $0,5$). Já o cenário 3 avalia os efeitos na estimação ao considerar a presença de assimetria nos dados ($\lambda = 2$), erro de medida e erro de classificação. Esses esquemas são aplicados para todos os modelos de estudo (Equações (2.12) - (2.15)).

Para as simulações foram consideradas 500 réplicas (R) e tamanho de amostra $n = 10.000$. A próxima seção (3.2), apresenta um novo estudo de simulação considerando diferentes tamanhos de amostra ($n = 50, 200, 500, 1.000, 10.000$) para um novo cenário descrito a seguir:

Cenário 4: $\pi_0 = 0,1, \pi_1 = 0,2, \beta_0 = 0, \beta_1 = 1, \lambda = 2, \sigma^2 = 0,02$.

Nos cenários 1-4, note que $(\sigma^2, \pi_0, \pi_1) = (\sigma^2, 0, 0), \sigma^2 > 0$ corresponde ao modelo 3 (ausência de erro de classificação); $(\sigma^2, \pi_0, \pi_1) = (0, \pi_0, \pi_1)$ com $(\pi_0, \pi_1) \neq (0,0)$ corresponde ao modelo 2 (ausência de erro de medida); e $(\sigma^2, \pi_0, \pi_1) = (0,0,0)$ corresponde ao modelo 1 (modelo Naive).

Para efeitos de comparação do modelo proposto com o modelo apresentado em ROY, BANERJEE e MAITI (2005), foi considerado similarmente o algoritmo da geração de dados aplicados aos quatro modelos (Equações (2.12) - (2.15)):

1. Gerar aleatoriamente a variável $W_t, t = 1, 2, \ldots, n$ da distribuição uniforme $U(-4,4)$ para representar os dados observados e fixar esses valores;
2. Gerar a variável $X_t, t = 1, 2, \cdots, n$ da distribuição $SN(W_t, v(W), \lambda)$, em que $v(W) = \sigma^2$ e λ assumem diferentes valores de acordo com o cenário de simulação;
3. Gerar a variável resposta $Y_{Tt}, t = 1, 2, \cdots, n$ da distribuição Bernoulli com probabilidade

de sucesso $\Phi(\beta_0+\beta_1 X_t)$. Neste caso iremos considerar $\beta_0=0$ e $\beta_1=1$;

4. Gerar a variável observada Y_t a partir das seguintes probabilidades

$$P(Y_t=1|Y_{Tt}=0)=\pi_0 \quad \text{e} \quad P(Y_t=0|Y_{Tt}=1)=\pi_1. \tag{3.1}$$

5. Ajustar os dados gerados $(Y_t,W_t), t=1,2,\ldots,n$, nos quatro modelos a fim de estimar os parâmetros pelo método da máxima verossimilhança.
6. Repetir os passos 2-5 para 500 réplicas, R, e encontrar a estimativa de $\hat{\theta}_{(l)}, l=1,2,\ldots,R$;
7. Calcular a média, o erro padrão de $\hat{\theta}$, o viés e o erro quadrático médio (EQM) empíricos de $\hat{\theta}_{(l)}, l=1,2,\ldots,R$;
8. Os passos 2-7 são repetidos para diferentes valores de $(\sigma^2,\pi_0,\pi_1,\lambda)$.

Foi considerado o método de otimização L-BFGS-B, mais conhecido como BFGS com restrição de caixa, para que o domínio de cada parâmetro a ser otimizado pudesse ser definido.

Para avaliar os resultados da simulação é considerado: as estimativas, o erro padrão (EP), o viés e o erro quadrático médio (EQM) empíricos dos estimadores dos parâmetros do modelo. O EP foi calculado a partir da informação de Fisher observada, isto é, a partir da Equação (2.9) sem a função esperança ($E(.)$). Os resultados são discutidos a partir de quatro abordagens diferentes: os efeitos ao não considerar nenhum tipo de erro, a partir do modelo 1 (modelo Naive - Eq. (2.12)); os efeitos dos erros de classificação, através do modelo 2 (Eq. (2.13)); os efeitos dos erros de medida, através do modelo 3 (Eq. (2.14)); e os efeitos de ambos os erros, através do modelo 4 (Eq. (2.15)).

3.1.1 Discussão dos resultados numéricos considerando n=10.000

Nesta seção, são discutidos os resultados do estudo de simulação de Monte Carlo para os cenários 1-3.

A Tabela 1 apresenta resultados de simulação utilizando dados que foram gerados a partir do modelo 2, e considera-se os cenários 1 e 2. As estatísticas consideradas são as estimativas dos parâmetros ($\hat{\theta}$), EP, viés e o EQM empírico para dados gerados do modelo 2 (M2), e ajustados também no modelo 1 (M1) para efeitos de comparação.

Comparando os resultados dos cenários 1 e 2, pode-se observar que com o aumento dos erros de classificação tem-se uma diminuição do EP, viés e EQM empírico do estimador de β_1 no M2. Para o modelo Naive (M1) mostra-se um aumento dos mesmos, o qual não acontece

Tabela 1 – Resultados de simulação para os cenários 1 e 2: estimativas dos parâmetros ($\hat{\theta}$), EP, viés e EQM empírico para dados gerados do M2, e ajustados também no M1

	M1				M2			
Parâmetros(θ)	$\hat{\theta}$	EP	Viés	EQM	$\hat{\theta}$	EP	Viés	EQM
Cenário 1: n = 10.000, (π_0; π_1) = (0,01; 0,01)								
π_0	-	-	-	-	0,0096	0,00003	$-0,0003$	0,00001
π_1	-	-	-	-	0,0094	0,00003	$-0,0006$	0,00001
β_0	0,0009	0,0001	0,0009	0,00001	0,0002	0,0002	$-0,0002$	0,0005
β_1	0,9880	0,0004	$-0,0120$	0,0021	0,9411	0,0023	$-0,0589$	0,0608
Cenário 2: n = 10.000, (π_0; π_1) = (0,05; 0,05)								
π_0	-	-	-	-	0,0503	0,00005	0,0002	0,00003
π_1	-	-	-	-	0,0496	0,00005	$-0,0004$	0,00003
β_0	0,0010	0,0000	0,0010	0,00001	$-0,0011$	0,0002	$-0,0011$	0,0008
β_1	0,9781	0,0011	$-0,0218$	0,0121	0,9971	0,0007	$-0,0030$	0,0054

com as estimativas de β_0, que parece não ser tão afetada com o aumento dos erros de classificação nesses dois modelos.

Por outro lado, observando cada cenário, verifica-se que para erros de classificação menores (Cenário 1) o EP, o viés, e o EQM do estimador de β_1 do M2 são maiores quando comparados com os resultados do modelo Naive. Já quando tem-se um aumento dos erros de classificação (Cenário 2), o EP, o viés, e o EQM do estimador de β_1 do M2 são menores comparados aos valores correspondentes do modelo Naive.

Tabela 2 – Resultados de simulação para os três cenários: estimativas dos parâmetros ($\hat{\theta}$), EP, viés e EQM empírico para dados gerados do M3, e ajustados também no M1

	M1				M3			
Parâmetros (θ)	$\hat{\theta}$	EP	Viés	EQM	$\hat{\theta}$	EP	Viés	EQM
Cenário 1: n = 10.000, (σ^2; λ) = (0,01; 0,001)								
β_0	0,0006	0,0002	0,0006	0,0005	0,0005	0,0002	0,0006	0,0005
β_1	1,0021	0,00002	0,0021	0,0004	1,0072	0,0001	0,0072	0,0004
λ	-	-	-	-	0,0008	0,00001	$-0,0001$	0,00001
Cenário 2: n = 10.000, (σ^2; λ) = (0,5; 0,001)								
β_0	0,0013	0,0002	0,0013	0,0004	0,0009	0,0002	0,0009	0,0004
β_1	0,8948	0,0005	$-0,1052$	0,0145	1,1562	0,0003	0,1562	0,0256
λ	-	-	-	-	0,0010	0,00009	0,00001	0,0008
Cenário 3: n = 10.000, (σ^2; λ) = (0,5; 2)								
β_0	0,3319	0,0002	0,3319	0,1106	0,0744	0,0029	0,0744	0,0900
β_1	0,9471	0,0001	$-0,0529$	0,0031	1,1479	0,0010	0,1479	0,0337
λ	-	-	-	-	1,9800	0,0001	$-0,00001$	0,0007

A Tabela 2 apresenta resultados de simulação utilizando dados que foram gerados a

partir do modelo 3, e considera-se os cenários 1-3. As estatísticas consideradas são as estimativas dos parâmetros($\hat{\theta}$), erro padrão (EP), viés e erro quadrático médio (EQM) empírico para dados gerados do modelo 3 (M3), e ajustados também no modelo 1 (M1) para efeitos de comparação.

Analisando a Tabela 2, pode-se observar que com o aumento apenas do erro de medida ($\sigma^2 = 0,01$ do cenário 1 e $\sigma^2 = 0,5$ do cenário 2), e com o aumento apenas do valor do parâmetro de assimetria ($\lambda = 0,001$ do cenário 2 e $\lambda = 2$ do cenário 3) tem-se um aumento dos valores de EP, viés e EQM empíricos dos coeficientes nos dois modelos, M1 e M3, sendo que os valores correspondentes ao M3 são maiores na maioria dos casos quando comparados ao modelo Naive.

Tabela 3 – Resultados de simulação para os cenários 1, 2 e 3: estimativas dos parâmetros ($\hat{\theta}$), EP, viés e EQM empírico para dados gerados do M4, e ajustados também nos modelos M1, M2 e M3

	M1				M2				M3				M4			
Parâmetros (θ)	$\hat{\theta}$	EP	Viés	EQM	$\hat{\theta}$	EP	Viés	EQM	$\hat{\theta}$	EP	Viés	EQM	$\hat{\theta}$	EP	Viés	EQM
Cenário 1: n = 10.000, (π_0, π_1, σ^2, λ) = (0,01; 0,01; 0,01; 0,001)																
π_0	-	-	-	-	0,0096	0,00003	$-0,0004$	0,00001	-	-	-	-	0,0095	0,00003	$-0,0005$	0,00001
π_1	-	-	-	-	0,0097	0,00003	$-0,0003$	0,00001	-	-	-	-	0,0096	0,00003	$-0,0003$	0,00001
β_0	0,0009	0,00001	0,0009	0,000001	0,0017	0,0002	0,0017	0,0006	0,0005	0,0001	0,0005	0,0003	0,0016	0,0002	0,0017	0,0006
β_1	0,9821	0,0008	$-0,0179$	0,0081	0,9618	0,0019	$-0,0382$	0,0407	1,0522	0,0005	0,0522	0,0060	0,9627	0,0020	$-0,0373$	0,0447
λ	-	-	-	-	0,0009	0,00001	$-0,0001$	0,00000004	0,0006	0,0001	$-0,0003$	0,0001	0,0009	0,00001	$-0,0001$	0,000003
Cenário 2: n = 10.000, (π_0, π_1, σ^2, λ) = (0,05; 0,05; 0,5; 0,001)																
π_0	-	-	-	-	0,0502	0,00005	0,0002	0,00002	-	-	-	-	0,0503	0,00004	0,0003	0,00002
π_1	-	-	-	-	0,0499	0,00005	$-0,0001$	0,00002	-	-	-	-	0,0499	0,00004	$-0,00003$	0,00002
β_0	0,0010	0,00001	0,0010	0,000001	$-0,0012$	0,0002	$-0,0012$	0,0008	0,0001	0,0001	0,0001	0,0002	$-0,0013$	0,0002	$-0,0013$	0,0007
β_1	0,9900	0,00001	$-0,0100$	0,0001	0,8985	0,0005	$-0,1015$	0,0129	0,6250	0,0007	$-0,3750$	0,1455	1,1617	0,0006	0,1617	0,0308
λ	-	-	-	-	0,0009	0,00001	$-0,0001$	0,00000001	0,0004	0,00007	$-0,0006$	0,00005	$-0,0004$	0,0002	$-0,0014$	0,0002
Cenário 3: n = 10.000, (π_0, π_1, σ^2, λ) = (0,05; 0,05; 0,5; 2)																
π_0	-	-	-	-	0,0502	0,0001	0,0002	0,00004	-	-	-	-	0,0497	0,00005	$-0,0002$	0,00002
π_1	-	-	-	-	0,0495	0,0001	$-0,0005$	0,00003	-	-	-	-	0,0502	0,00004	0,0002	0,00001
β_0	0,0010	0,00001	0,0010	0,000001	0,3264	0,0004	0,3264	0,1083	0,0988	0,0014	0,0988	0,0292	$-0,1562$	0,0003	$-0,1562$	0,0255
β_1	0,9900	0,00001	$-0,0100$	0,0001	0,9407	0,0009	$-0,0593$	0,0118	0,6323	0,0002	$-0,3677$	0,1356	1,0693	0,0005	0,0694	0,0072
λ	-	-	-	-	1,9647	0,0018	0,0358	0,0324	0,6029	0,0089	$-1,3971$	2,7585	1,9775	0,0008	$-0,0224$	0,0064

A Tabela 3 apresenta resultados de simulação utilizando dados que foram gerados a partir do modelo 4, e considera-se os cenários 1-3. As estatísticas consideradas são as estimativas dos parâmetros($\hat{\theta}$), erro padrão (EP), viés e erro quadrático médio (EQM) empírico para dados gerados do modelo 4 (M4), e ajustados também nos modelos 3 (M3), 2 (M2) e 1 (M1) para efeitos de comparação.

Comparando os resultados dos cenários 1 e 2, pode-se observar que com o aumento dos erros de classificação e dos erros de medida tem-se uma diminuição do EP e EQM empírico do estimador de β_1 no M4. Para o modelo Naive (M1) mostra-se uma dimiuição dos mesmos. O mesmo comportamento se repete com o viés de β_0.

Comparando os resultados dos cenários 2 e 3, pode-se observar que com o aumento do parâmetro de assimetria, tem-se uma diminuição do EP, viés e do EQM empírico do estimador de β_1 no M4. Porém, o oposto ocorre para o parâmetro β_0. Para o modelo Naive (M1), o comportamento do EP, viés e do EQM não sofrem alteração de um cenário para outro.

Por outro lado, observando cada cenário, verifica-se que para erros de classificação e erros de medida maiores (Cenário 2) o EP, o viés, e o EQM do estimador de β_1 do M4 são maiores quando comparados com os resultados do modelo Naive. Já quando tem-se um aumento dos erros de classificação, dos erros de medida e do parâmetro de assimetria (Cenário 3), o EP, o viés, e o EQM do estimador de β_1 do M4 são maiores comparados aos valores correspondentes do modelo Naive.

Embora seja observado que pra alguns modelos o viés e o EQM aumentam, no geral para o M4 as estimativas dos parâmetros se aproximam do valor verdadeiro. Ao aumentar as probabilidades de misclassificação e erro de medida temos que no M2 ocorreu aumento do viés. Deste modo, pode-se obervar o impacto ao ajustar modelos com erro de classificação e erro de medida em modelos que não apresentam algum tipo de erro.

3.2 SIMULAÇÃO DE MONTE CARLO: CONSIDERANDO DIFERENTES TAMANHOS DE AMOSTRA

No presente estudo de simulação tem-se como objetivo avaliar as propriedades assintóticas dos estimadores de máxima verossimilhança dos quatro modelos (Equações (2.12) - (2.15)) considerando o cenário 4 descrito na seção anterior:

Cenário 4: $\pi_0 = 0,1, \pi_1 = 0,2, \beta_0 = 0, \beta_1 = 1, \lambda = 2, \sigma^2 = 0,02$.

Para esse estudo foi considerado diferentes tamanhos amostrais como: $n = 50, 200, 500$, 1.000 e 10.000. Os resultados são apresentados nas Tabelas 4 - 7.

Tabela 4 – Resultados de simulação: estimativas dos parâmetros ($\hat{\theta}$), EP, viés e EQM empírico para dados gerados do M2, e ajustados também no M1

	M1				M2			
Parâmetros(θ)	$\hat{\theta}$	EP	Viés	EQM	$\hat{\theta}$	EP	Viés	EQM
n = 50								
π_0	-	-	-	-	0,0748	0,0016	$-0,0250$	0,0073
π_1	-	-	-	-	0,1654	0,0025	$-0,0346$	0,0164
β_0	0,0009	0,000000008	0,0010	0,0000001	3,0436	0,8517	3,0436	1822,6540
β_1	0,9880	0,0009	$-0,0120$	0,0021	17,2501	1,4847	16,2501	5774,8660
n = 200								
π_0	-	-	-	-	0,0919	0,0002	$-0,0080$	0,0025
π_1	-	-	-	-	0,1856	0,0003	$-0,0143$	0,0047
β_0	0,0010	0,0000002	0,0010	0,000001	0,2115	0,0568	0,2115	128,9986
β_1	0,9880	0,0002	$-0,0120$	0,0021	6,4246	0,2136	5,4246	1854,9070
n = 500								
π_0	-	-	-	-	0,0962	0,00006	$-0,0038$	0,0009
π_1	-	-	-	-	0,1949	0,00008	$-0,0051$	0,0016
β_0	0,0010	0,0000	0,0010	0,000001	14,2615	0,6370	14,2615	101645,3
β_1	0,9900	0,0000	$-0,01$	0,0001	31,3801	1,3530	30,3801	458553,9
n = 1.000								
π_0	-	-	-	-	0,0981	0,00002	$-0,0019$	0,0004
π_1	-	-	-	-	0,1969	0,00003	$-0,0030$	0,0007
β_0	0,0010	0,0000	0,0010	0,000001	0,0019	0,0001	0,0019	0,0235
β_1	0,9900	0,0000	$-0,01$	0,0001	1,0411	0,0002	0,0410	0,0575

Tabela 5 – Resultados de simulação: estimativas dos parâmetros ($\hat{\theta}$), EP, viés e EQM empírico para dados gerados do M3, e ajustados também no M1

	M1				M3			
Parâmetros(θ)	$\hat{\theta}$	EP	Viés	EQM	$\hat{\theta}$	EP	Viés	EQM
n = 50								
β_0	0,0027	0,0073	0,0027	0,1338	$-0,1275$	0,0075	$-0,1275$	0,1576
β_1	1,0619	0,0096	0,0619	0,2334	1,0650	0,0101	0,0650	0,2587
λ	-	-	-	-	36,1777	1,7648	34,1777	8954,2490
n = 200								
β_0	0,0342	0,0008	0,0268	0,0294	$-0,1034$	0,0008	$-0,1034$	0,0401
β_1	1,0512	0,0008	0,0076	0,0279	1,0651	0,0008	0,0651	0,0321
λ	-	-	-	-	13,0367	0,2165	11,0367	1997,1810
n = 500								
β_0	0,0268	0,0002	0,0268	0,0089	$-0,1134$	0,0002	$-0,1134$	0,0220
β_1	1,0076	0,0002	0,0076	0,0123	1,0241	0,0002	0,0241	0,0098
λ	-	-	-	-	6,0769	0,0487	4,0769	609,4545
n = 1.000								
β_0	0,0268	0,00007	0,0268	0,0052	$-0,1162$	0,00007	$-0,1162$	0,0183
β_1	1,0099	0,00006	0,0099	0,0040	1,0178	0,00007	0,0178	0,0047
λ	-	-	-	-	2,6765	0,0076	0,6765	57,6361

Tabela 6 – Resultados de simulação: estimativas dos parâmetros ($\hat{\theta}$), EP, viés e o EQM empírico para dados gerados do M4, e ajustados também nos modelos M1, M2 e M3

	M1				M2				M3				M4			
Parâmetros (θ)	$\hat{\theta}$	EP	Viés	EQM	$\hat{\theta}$	EP	Viés	EQM	$\hat{\theta}$	EP	Viés	EQM	$\hat{\theta}$	EP	Viés	EQM
n = 50																
π_0	-	-	-	-	0,0683	0,0015	−0,0317	0,0065	-	-	-	-	0,0544	0,0014	−0,0456	0,0072
π_1	-	-	-	-	0,1674	0,0022	−0,0326	0,0140	-	-	-	-	0,1369	0,0024	−0,0630	0,0187
β_0	0,0009	0,00000008	0,0009	0,000001	9,5835	2,6895	9,5835	18174,86	−34,8952	5,6969	−34,8952	82356,32	−1000,06	197,9262	−1000,06	98.937.075
β_1	0,9880	0,00008	−0,0119	0,0021	44,5581	8,3736	43,5581	177192,8	86,6125	8,3785	85,6125	239580,8	2228,328	200,9394	2227,328	105.902.628
λ	-	-	-	-	1,9483	0,0050	−0,0517	0,0645	189,8656	7,6258	187,8656	180.678,2	81,8319	18,6061	79,8319	871.847,3
n = 200																
π_0	-	-	-	-	0,0867	0,0003	−0,0133	0,0028	-	-	-	-	0,0756	0,0003	−0,0243	0,0035
π_1	-	-	-	-	0,1858	0,0003	−0,0142	0,0048	-	-	-	-	0,1661	0,0004	−0,0339	0,0082
β_0	0,0009	0,0000002	0,0009	0,000001	3,0238	0,3804	3,0238	5798,112	−3,6331	0,2701	−3,6331	2931,495	−415,5589	15,6441	−415,5589	9.962.219
β_1	0,9880	0,0002	−0,0119	0,0021	25,4590	1,2082	24,4590	58987,57	41,6329	2,3747	40,6329	227211,4	796,7851	25,5142	795,7851	26.672.292
λ	-	-	-	-	1,9760	0,0004	−0,0239	0,0084	195,0637	1,7337	193,0637	157510,9	51,4041	2,9286	49,4041	345.499,1
n = 500																
π_0	-	-	-	-	0,0934	0,00006	−0,0067	0,0011	-	-	-	-	0,0859	0,00007	−0,0140	0,0017
π_1	-	-	-	-	0,1915	0,00008	−0,0084	0,0018	-	-	-	-	0,1799	0,0001	−0,0201	0,0042
β_0	0,0009	0,00000008	0,0009	0,000001	0,0402	0,0005	0,0402	0,0663	−0,3624	0,0001	−0,3624	0,1369	−98,6425	2,0872	−98,6425	1.098.839
β_1	0,9880	0,00008	−0,0119	0,0021	1,0959	0,0012	0,0959	0,3520	0,3880	0,00007	−0,6119	0,3758	225,1895	4,6593	224,1895	5.477.656
λ	-	-	-	-	1,9760	0,0002	−0,0239	0,0084	1,9162	0,0006	−0,0838	0,0957	14,5735	0,1500	12,5735	5783,781
n = 1.000																
π_0	-	-	-	-	0,0987	0,00002	−0,0012	0,0004	-	-	-	-	0,0954	0,00002	−0,0046	0,0007
π_1	-	-	-	-	0,1985	0,00003	−0,0014	0,0008	-	-	-	-	0,1933	0,00004	−0,0067	0,0019
β_0	0,0010	0,0000001	0,0010	0,000001	0,0352	0,0002	0,0352	0,0246	−0,3608	0,00005	−0,3608	0,1330	−4,7461	0,0677	−4,7461	4613,796
β_1	0,9900	0,000001	−0,0100	0,0001	1,0389	0,0002	0,0389	0,0494	0,3827	0,00002	−0,6172	0,3816	17,6541	0,2538	16,6541	64667,31
λ	-	-	-	-	1,9800	0,0000001	−0,0200	0,0004	1,9281	0,0002	−0,0718	0,0715	9,6553	0,0424	7,6553	1856,943

Tabela 7 – Resultado de simulação: estimativas dos parâmetros ($\hat{\theta}$), EP, viés e o EQM empírico para dados gerados do modelo M4

Parâmetros (θ)	$\hat{\theta}$	EP	Viés	EQM
		n = 10.000		
π_0	0.1002	0.0064	0.0002	-0.00004
π_1	0.1996	0.0081	-0.0004	0.00006
β_0	-0.1150	0.0453	-0.1150	0.0153
β_1	1.0142	0.0562	0.0142	0.0034
λ	1.9784	0.0007	-0.0216	0.0005

De maneira geral, ao analisar o comportamento dos modelos apresentados pelas Tabelas 4 - 7, observamos que conforme o tamanho da amostra aumenta, o viés e o EQM diminuem para M2, M3 e M4. Logo, as estimativas dos parâmetros se aproximam do valor verdadeiro.

Analisando a tabela 7, pode-se notar que ao considerar um tamanho de amostra $n = 10.000$ as estimativas dos parâmetros se aproximam do valor verdadeiro, e o viés e EQM do M4 diminuem. Desta forma, a propriedade de consistência é verificada.

Para amostras menores que $n = 10.000$, embora seja observado que à medida que a amostra cresce o viés diminui para os modelos M2, M3 e M4, para esses tamanhos de amostra o viés não está próximo de zero.

4 APLICAÇÃO

Em 1945, os Estados Unidos encerraram definitivamente a Segunda Guerra Mundial após lançarem duas bombas atômicas contra o Japão, nas cidades de Hiroshima e Nagasaki. A rendição japonesa se deu de forma incondicional poucos dias depois. As bombas, desenvolvidas durante anos pelos americanos, devastaram completamente as duas cidades em questão de segundos. Estas foram as únicas vezes em que armamentos nucleares foram utilizados em conflitos militares, e desde então foram responsáveis por modificar todos os paradigmas militares e pelo início de um período de medo, com a possibilidade de uma guerra nuclear. Muitas pessoas que sobreviveram aos ataques, ou que residiam em áreas próximas, foram afetadas pela radiação e desenvolveram uma série de problemas de saúde, como câncer.

Para aplicação foram utilizados os dados disponíveis no artigo SPOSTO *et al.* (1992). Este estudo iniciou 5 anos depois do bombardeio atômico das cidades de Hiroshima e Nagasaki, e tem como principal objetivo avaliar o efeito da exposição à radiação em relação as mortes por câncer naquele período de estudo. Foram considerados 86.520 sobreviventes ao ataque, membros do LSS (Life Span Study), diferenciados por grupos de expostos e não-expostos. Essa distinção teve como base as distâncias do local que foi atingido pela bomba ($< 2km$, 2 a $10km$). Esses sobreviventes foram acompanhados do dia 1 de Outubro de 1950 a 31 de Dezembro de 1985. A Tabela 8, contém informações sobre a dose de exposição à radiação, dose média de exposição à radiação, número de mortes por câncer, número de mortes por outras causas e a proporção de mortes por câncer . Assim, os erros de medida das doses de radiação dependem da localização de cada indivíduo e também porque por mais que os indivíduos tenham as mesmas condições de exposição, eles podem absorver diferentes quantidades de radiação, devido a razões biológicas. Logo, neste artigo foi considerado que as diferentes condições biológicas são as causadoras dos erros de medida. A dose de exposição a radiação é medida de acordo com a dosimetria[1] DS86 e tem por unidade de medida a sigla Gy (Gray). Foram disponibilizados dados sobre os mais variados tipos de câncer como: pulmão, boca, intestino, mama, próstata, entre outros. Porém, as doses de radiação (Gy) absorvidas pelo intestino no momento da exposição foram selecionadas como dose de referência.

Nessa aplicação, o interesse é avaliar se o modelo proposto (com erro de classificação e erro de medida) descreve bem os dados. Foi considerada que a variável substituta W representa a dose média de cada categoria, e a variável X representa a dose verdadeira. Logo, de acordo com

[1] Medida das doses de radiação que um indivíduo (ou um ser vivo) pode estar exposto.

o modelo proposto, assumimos que $X|W \sim SN(w, cw^2, \lambda)$. Em que c é conhecido e representa o coeficiente de variação igual a 0.5. Para as análises, foi utilizado os quatro modelos descritos no Capítulo 2 (Equações (2.12) - (2.15)).

Tabela 8 – Informações sobre os sobreviventes ao bombardeio em Hiroshima e Nagasaki por categoria da dose de absorção à radiação.

Dose	Dose média	Mortes por câncer	Mortes por outras causas	Proporção
0,000	0,000	2784	10201	0,2144
0,01 - 0,05	0,018	2105	7451	0,2203
0,06 - 0,09	0,072	439	1509	0,2253
0,10 - 0,19	0,137	523	1701	0,2352
0,20 - 0,49	0,324	586	1785	0,2471
0,50 - 0,99	0,693	339	826	0,2910
1,00 - 1,99	1,350	204	369	0,3560
2,00 - 2,99	2,350	57	86	0,3986
3,00 - 3,99	3,520	21	51	0,2917
4,00+	4,430	13	23	0,3611
Total/Média	0,113	7.071	24.002	-

Na Tabela 8 pode-se notar que a proporção de morte por câncer aumenta, a medida que a dose média aumenta. Porém a proporção diminui nas duas últimas categorias. Para explicar esse comportamento, SHIMIZU, KATO e SCHULL (1990) indicaram que isso ocorre devido à classificação incorreta de *mortes por câncer* como *mortes por outras causas* nos atestados de óbito.

4.1 ANÁLISE DESCRITIVA

No banco de dados utilizado, 49% dos indivíduos analisados são mulheres e 51% são homens. A partir da Tabela 9 e da Figura 2, nota-se que a variável que mede a dose de radiação absorvida, apresenta valores concentrados próximos a 0 Gy e o valor máximo observado é de 6.000 Gy. Temos que 75% das pessoas receberam doses de exposição superior a $71,62$ Gy, a mediana é igual a $324,14$ Gy e o valor médio é igual a $1252,19$ Gy o que sugere assimetria da distribuição, corroborando o histograma apresentado na Figura 4. Vale salientar que existem outliers no conjunto de dados. Ao analisar a Figura 3, pode-se notar que a distribuição da dose de radiação absorvida (por sexo) é igual e há presença de outlier para os dois sexos, nos quais os outliers indicam as maiores doses de radiação absorvidas.

Na Figura 4, pode-se notar que a dispersão da dose de radiação absorvida (por mil)

Tabela 9 – Medidas resumo da variável que mede a dose de radiação absorvida, com base na dosimetria DS86.

Min	1Q	2Q	Média	3Q	Max
0,00	71,62	324,14	1252,19	2287,78	6000,00

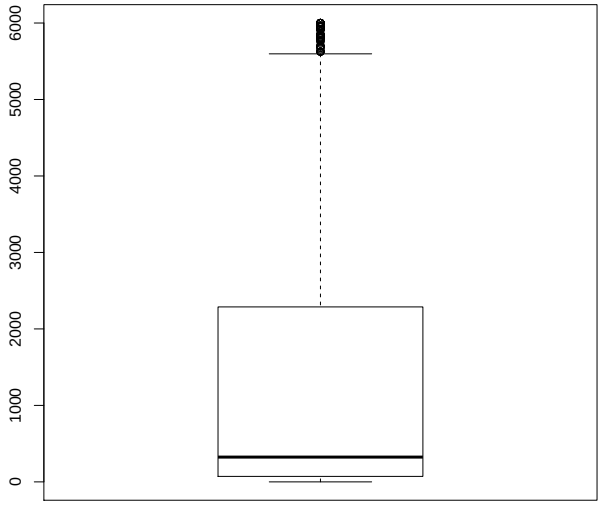


Figura 2 – Boxplot da variável dose de radiação absorvida.

vai de 0 a 6, em que a dose 0 é a que teve o maior número de observações. Também observa-se que a dispersão da dose de radiação absorvida segue uma distribuição assimétrica positiva.

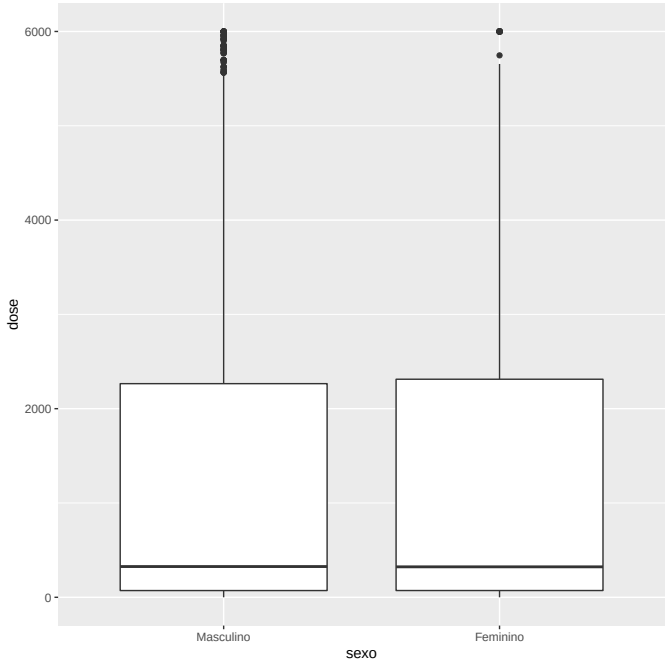


Figura 3 – Boxplot da variável dose de radiação absorvida por sexo.

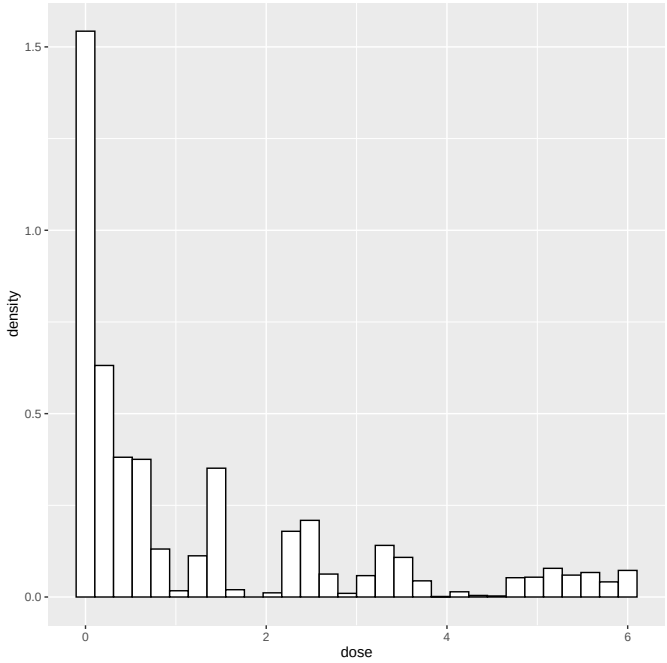


Figura 4 – Histograma da variável dose de radiação absorvida.

4.2 AJUSTES DOS MODELOS

A Tabela 10 apresenta as estimativas dos parâmetros ($\hat{\theta}$), erro padrão (EP), p-valor e os critérios de seleção de modelos AIC e BIC (Equações (2.16) e (2.17)) para os quatro modelos estudados no Capítulo 2 (Equações (2.12) - (2.15)).

Tabela 10 – Resultado da aplicação: estimativas dos parâmetros ($\hat{\theta}$), EP, p-valor e o AIC para os modelos M1, M2, M3 e M4.

Parâmetro	M1			M2			M3			M4		
	$\hat{\theta}$	EP	p-valor	$\hat{\theta}$	EP	p-valor	$\hat{\theta}$	EP	p-valor	$\hat{\theta}$	EP	p-valor
π_0	-	-	-	0,1718	0,0032	$< 0,001$	-	-	-	0,2276	0,0831	0,0093
π_1	-	-	-	0,6312	0,0079	$< 0,001$	-	-	-	0,00001	0,1271	0,9990
β_0	$-0,7727$	0,0083	$< 0,001$	$-0,7471$	0,0381	$< 0,001$	$-0,8016$	0,0079	$< 0,001$	$-0,6672$	0,0965	$< 0,001$
β_1	0,2012	0,0208	$< 0,001$	1,5092	0,1787	$< 0,001$	0,2013	0,0168	$< 0,001$	3,8509	0,8262	$< 0,001$
λ	-	-	-	-	-	-	1,9642	2,6890	0,4651	$-0,7262$	0,0661	$< 0,001$
AIC	33.242,05			33.220,44			33.244,05			33.222,46		
BIC	33.258,74			33.253,82			33.269,08			33.264,18		

Pode-se observar que os sinais das estimativas do parâmetro de assimetria mudam porque os dados tem erro de classificação e erro de medida, adequado para ajustar o M4. O M3 apenas considera o erro de medida e portanto para esses dados este modelo erra mais, o que leva a uma diferença grande entre as estimativas destes modelos. Para este conjunto de dados o erro de classificação não pode ser negligível.

A partir do critério AIC, temos que os modelos 2 e 4 apresentam valores próximos, sendo este valor o menor comparado com os outros valores de AIC para os modelos 1 e 3. Então pode-se concluir que a partir desse critério, os modelos que descrevem melhor os dados são o M2 e o M4. Isso faz sentido, uma vez que o M2 é um caso particular do modelo proposto (M4). Já segundo o critério BIC, o modelo que melhor descreve os dados é o M2, em que foi observado o menor valor. Porém nesta aplicação, propõe-se o uso do M4, uma vez que foi observado que existe erro de classificação e erro de medida.

5 CONCLUSÃO

5.1 CONSIDERAÇÕES FINAIS

Neste trabalho foi feito uma extensão do estudo realizado por ROY, BANERJEE e MAITI (2005), em que foi estudado o modelo de regressão binária, na qual a variável resposta está sujeita ao erro de classificação e as covariáveis ao erro de medida do tipo Berkson e seguem distribuição normal. Porém, no nosso trabalho foi considerado que o erro de medida segue uma distribuição normal assimétrica e quando o parâmetro de assimetria (λ) é zero, temos o modelo dado em ROY, BANERJEE e MAITI (2005).

O modelo proposto nesta dissertação (M4) tem 3 submodelos (M3, M2 e M1): o modelo com apenas erro de medida (M3), que é um modelo novo; o modelo com apenas erro de classificação (M2), que também foi estudado por ROY, BANERJEE e MAITI (2005) como caso particular; e por último o modelo Naive (M1), que é caso particular do M4 como também é caso particular do modelo proposto por ROY, BANERJEE e MAITI (2005).

No estudo de simulação verificou-se que para amostras grandes ($n = 10.000$), o viés e o EQM dos estimadores do modelo proposto se aproximam a zero. Quando o erro de classificação ou erro de medida aumenta, o viés e EQM dos modelos que não são apropriados (especialmente do modelo Naive) tendem a aumentar. Também foi feito um estudo de simulação para diferentes tamanhos de amostra ($n = 50, 200, 500, 1.000, 10.0000$), onde observa-se que para amostras pequenas o viés e o EQM empírico do modelo proposto são bem grandes, mas à medida que aumenta o tamanho da amostra, estes valores diminuem.

Por fim, uma aplicação foi realizada com dados provenientes dos efeitos causados na saúde da população que enfrentou o bombardeio atômico das cidades de Hiroshima e Nagasaki em 1945. O objetivo desse estudo foi avaliar o efeito da exposição à radiação em relação às mortes por câncer, no período de 5 anos após a explosão. Verificou-se que os dados tem distribuição assimétrica e ajustou-se os quatro modelos estudados nesta dissertação. Então, foi verificado que o modelo com apenas erro de classificação (M2) e o modelo com apenas erro de medida (M3) fornecem estimativas diferentes dos coeficientes do modelo, comparado com o modelo naive (M1). Isso sugere que estes dados possuem erro de classificação e erro de medida. Já o modelo proposto (M4) que considera os erros de classificação e erro de medida, forneceu estimativas dos coeficientes bastante diferentes dos outros modelos. Por outro lado, o modelo 4 foi o modelo selecionado pelo critério AIC.

5.2 TRABALHOS FUTUROS

Um estudo futuro, baseado no modelo estudado neste trabalho, seria estender o modelo proposto para o caso em que o erro de medida pertence à família de distribuições assimétricas (normal assimétrica, *slash* assimétrica, t-Student assimétrica, exponencial potencial assimétrica, etc).

REFERÊNCIAS

ACHIC, B. G. B.; SANDOVAL, M. C.; YOSHIDA, O. S. Homoscedastic controlled calibration model. **Journal of Chemometrics: A Journal of the Chemometrics Society**, Wiley Online Library, v. 21, n. 3-4, p. 145–155, 2007.

AKAIKE, H. A new look at the statistical model identification. **IEEE transactions on automatic control**, Ieee, v. 19, n. 6, p. 716–723, 1974.

ARELLANO-VALLE, R. B.; AZZALINI, A. On the unification of families of skew-normal distributions. **Scandinavian Journal of Statistics**, Wiley Online Library, v. 33, n. 3, p. 561–574, 2006.

ARELLANO-VALLE, R. B.; GENTON, M. G. On fundamental skew distributions. **Journal of Multivariate Analysis**, Elsevier, v. 96, n. 1, p. 93–116, 2005.

AZZALINI, A. A class of distributions which includes the normal ones. **Scandinavian Journal of Statistics**, JSTOR, v. 12, n. 2, p. 171–178, 1985.

AZZALINI, A. The skew-normal distribution and related multivariate families. **Scandinavian Journal of Statistics**, Wiley Online Library, v. 32, n. 2, p. 159–188, 2005.

AZZALINI, A.; CAPITANIO, A. Statistical applications of the multivariate skew normal distribution. **Journal of the Royal Statistical Society: Series B (Statistical Methodology)**, Wiley Online Library, v. 61, n. 3, p. 579–602, 1999.

AZZALINI, A.; VALLE, A. D. The multivariate skew-normal distribution. **Biometrika**, Oxford University Press, v. 83, n. 4, p. 715–726, 1996.

BERKSON, J. Application of the logistic function to bio-assay. **Journal of the American statistical association**, Taylor & Francis Group, v. 39, n. 227, p. 357–365, 1944.

BERKSON, J. Are there two regressions? **Journal of the American Statistical Association**, JSTOR, v. 45, n. 250, p. 164–180, 1950.

BERKSON, J. Why i prefer logits to probits. **Biometrics**, JSTOR, v. 7, n. 4, p. 327–339, 1951.

BERKSON, J.; OTHERS. Minimum chi-square, not maximum likelihood! **The Annals of Statistics**, Institute of Mathematical Statistics, v. 8, n. 3, p. 457–487, 1980.

BICKEL, P.; RITOV, Y. Efficient estimation in the errors in variables model. **The Annals of Statistics**, JSTOR, p. 513–540, 1987.

BLISS, C. I. The method of probits. **Science**, American Assn for the Advancement of Science, 1934.

BLISS, C. I. The calculation of the dosage-mortality curve. **Annals of Applied Biology**, Wiley Online Library, v. 22, n. 1, p. 134–167, 1935.

BOLFARINE, H.; LANCHOS, V. H. Skew-probit measurement error models. **Statistical Methodology**, Elsevier, v. 4, n. 1, p. 1–12, 2007.

BOLFARINE, H.; SANDOVAL, M. C. **Introdução à inferência estatística**. [S.l.]: SBM, 2010. v. 2.

BOZDOGAN, H. Model selection and akaike's information criterion (aic): The general theory and its analytical extensions. **Psychometrika**, Springer, v. 52, n. 3, p. 345–370, 1987.

BURR, D. On errors-in-variables in binary regression—berkson case. **Journal of the American Statistical Association**, Taylor & Francis, v. 83, n. 403, p. 739–743, 1988.

CARRASCO, J. Modelos de regressão beta com erro nas variáveis. **Tese apresentada ao instituto de Matemática e Estatística, IME-USP**, 2012.

CARROL, R. J.; RUPPERT, D.; STEFANSKI, L. A.; CRAINICEANU, C. M. **Measurement error in nonlinear models: a modern perspective**. [S.l.]: Chapman and Hall/CRC, 2006.

CARROL, R. J.; SPIELGMAN, C. H.; LAN, K. G.; BAILE, K. T.; ABOOT, R. D. On errors-in-variables for binary regression models. **Biometrika**, Oxford University Press, v. 71, n. 1, p. 19–25, 1984.

CHARNET, R.; FREIRE, C. d. L.; CHARNET, E. M.; BONVINO, H. *et al.* Análise de modelos de regressão linear com aplicações. **Campinas: Unicamp**, 1999.

CHENG, C.-L.; NESS, J. W. V. **Statistical Regression with Measurement Error**. [S.l.]: London: Arnold and New York: Oxford University Press, 1999.

CORDEIRO, G. M.; DEMÉTRIO, C. G. Modelos lineares generalizados e extensões. **Piracicaba: USP**, 2008.

CUNHA, W. J. D.; COLOSIMO, E. A. Intervalos de confiança bootstrap para modelos de regressão com erros de medida. **Rev. Mat. Estat**, v. 21, n. 2, p. 25–41, 2003.

DEMÉTRIO, C. G. B. **Modelos lineares generalizados em experimentação agronômica**. [S.l.]: USP/ESALQ, 2001.

DINIZ, M. A. **Modelos bayesianos semi-paramétricos para dados binários**. Tese (Doutorado) — Universidade de São Paulo, 2015.

DOLBY, G. R. The ultrastructural relation: a synthesis of the functional and structural relations. **Biometrika**, Oxford University Press, v. 63, n. 1, p. 39–50, 1976.

DONATELLI, G. D.; KONRATH, A. C. Simulação de monte carlo na avaliação de incertezas de medição. **REVISTA DE CIÊNCIA E TECNOLOGIA**, v. 13, n. 25/26, p. 5–15, 2005.

EMILIANO, P. C.; VEIGA, E. P.; VIVANCO, M.; MENEZES, F. S. Critérios de informação de akaike versus bayesiano: análise comparativa. **19º Simpósio Nacional de Probabilidade e Estatística**, 2010.

EUGENIO, N. W.; OTHERS. Modelo de regressão para dados binários com mistura de funções de ligação. Universidade Federal de São Carlos, 2017.

FERREIRA, C. d. S. Inferência e diagnóstico em modelos assimétricos. **Tese apresentada ao IME USP**, 2008.

FREITAS, L. A. d. Modelo de regressão com erros normais assimétricos: Uma abordagem bayesiana. **Dissertação apresentada ao Departamento de Estatística da Universidade Federal de São Carlos - DEs/UFSCar**, 2005.

FULLER, W. A. **Measurement error models**. [S.l.]: John Wiley & Sons, 1987. v. 305.

GONDIM, N. R. G. **A DISTRIBUIÇÃO NORMAL ASSIMÉTRICA (SKEW NORMAL) E SUAS APLICAÇÕES**. Tese (Doutorado) — Universidade Federal da Paraíba, 2011.

HEID, I.; KÜCHENHOFF, H.; MILES, J.; KREINBROCK, L.; WICHMANN, H. Two dimensions of measurement error: classical and berkson error in residential radon exposure assessment. **Journal of Exposure Science and Environmental Epidemiology**, Nature Publishing Group, v. 14, n. 5, p. 365, 2004.

JAMES, B. R. **Probabilidade: um curso em nível intermediário**. [S.l.: s.n.], 1996.

KANNEL, W. B.; GORDON, T. **The Framingham study: an epidemiological investigation of cardiovascular disease**. [S.l.]: US Department of Health, Education, and Welfare, National Institutes of Health, 1968.

KIM, H. Y.; YASUI, Y.; BURSTYN, I. Attenuation in risk estimates in logistic and cox proportional-hazards models due to group-based exposure assessment strategy. **Annals of Occupational Hygiene**, v. 50, n. 6, p. 623–635, 2006.

KLEIN, C. H.; COSTA, E. d. A. Os erros de classificação e os resultados de estudos epidemiológicos. **Cadernos de saúde pública**, scielo, v. 3, p. 236–249, 1987.

MA, Y.; GENTON, M. G. Flexible class of skew-symmetric distributions. **Scandinavian Journal of Statistics**, v. 31, p. 459–468, 2004.

MCCULLOCH, C. E.; SEARLE, S. R. **Generalized, Linear, and Mixed Models**. [S.l.]: Wiley-Interscience, 2001.

MCCULLOCH, P.; NELDER, J. **Generalized linear models**. [S.l.]: Chapman and Hall, 1989.

MCGLOTHLIN, A.; STAMEY, J. D.; SEAMAN, J. W. Binary regression with misclassified response and covariate subject to measurement error: a bayesian approach. **Biometrical Journal**, v. 50, p. 123–134, 2008.

PAULA, G. A. **Modelos de regressão: com apoio computacional**. [S.l.]: IME-USP São Paulo, 2004.

REEVES, G. K.; COX, D. R.; DARBY, S. C. Some aspects of measurement error in explanatory variables for continuous and binary regression models. **Statistics in Medicine**, v. 17, p. 2157–2177, 1998.

ROY, S.; BANERJEE, T.; MAITI, T. Measurement error model for misclassified binary responses. **Statistics in medicine**, Wiley Online Library, v. 24, n. 2, p. 269–283, 2005.

SANTOS, C. O.; SCALON, J. D.; OZAKI, V. A. A distribuição normal-assimétrica como modelo para produtividade de milho aplicada ao seguro agrícola. **Revista de Economia e Sociologia Rural**, v. 52, n. 4, 2014.

SANTOS, J. R. S. d. Um modelo de resposta ao item para grupos múltiplos com distribuições normais assimétricas centralizadas. **Dissertação de mestrado apresentada ao Instituto de Matemática, Estatística e Computação Científica da UNICAMP**, 2012.

SCHWARZ, G. *et al.* Estimating the dimension of a model. **The annals of statistics**, Institute of Mathematical Statistics, v. 6, n. 2, p. 461–464, 1978.

SHIMIZU, Y.; KATO, H.; SCHULL, W. J. Studies of the mortality of a-bomb survivors: 9. mortality, 1950-1985: part 2. cancer mortality based on the recently revised doses (ds86). **Radiation research**, Academic Press, Inc., v. 121, n. 2, p. 120–141, 1990.

SPOSTO, R.; PRESTON, D. L.; SHIMIZU, Y.; MABUCHI, K. The effect of diagnostic misclassification on non-cancer and cancer mortality dose response in a-bomb survivors. **Biometrics**, JSTOR, p. 605–617, 1992.

STEFANSKI, L. A. Measurement error models. **Journal of the American Statistical Association**, Taylor & Francis, v. 95, n. 452, p. 1353–1358, 2000.

STEFANSKI, L. A.; CARROL, R. J. Covariate measurement error in logistic regression. **The Annals of Statistics**, JSTOR, p. 1335–1351, 1985.

STEFANSKI, L. A.; CARROL, R. J. Conditional scores and optimal scores for generalized linear measurement-error models. **Biometrika**, Oxford University Press, v. 74, n. 4, p. 703–716, 1987.

TIEPPO, S. M. Inferência em um modelo de regressão com resposta binária na presença de sobredispersão e erros de medição. **Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação - ICMC-USP**, 2007.

WANG, L.; OTHERS. Estimation of nonlinear models with berkson measurement errors. **The Annals of Statistics**, Institute of Mathematical Statistics, v. 32, n. 6, p. 2559–2579, 2004.

WHITTEMORE, A. S.; KELLER, J. B. Approximations for regression with covariate measurement error. **Journal of the American Statistical Association**, v. 83, n. 404, p. 1057–1066, 1988.