Uma aplicação industrial de regressão binária com
erros na variável explicativa

Daniel Fernando de Favari

SERVIÇO DE PÓS-GRADUAÇÃO DO ICMC-USP

Data de Depósito: 12/05/2006

Assinatura: _____

Uma aplicação industrial de regressão binária com erros na variável explicativa

Daniel Fernando de Favari

Orientador: Prof. Dr. Dorival Leão Pinto Junior

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

USP - São Carlos
Maio/2006

Aos meus pais

Francisco Roberto de Favari

e

Maria Stella Ramin de Favari

Agradecimentos

Ao professor Dorival Leão Pinto Junior a excelente orientação dispendida, o contínuo apoio em todas as fases de realização deste trabalho, as discussões críticas que tanto contribuíram para o meu crescimento intelectual e pessoal, a amizade que pude desfrutar e por constituir um exemplo de pesquisador.

Ao professor Mário de Castro Andrade Filho a imprescindível co-orientação, o apoio promovendo discussões que muito enriqueceram a realização desta pesquisa e a amizade a qual me mostrou o respeito pela minha caminhada no mundo científico.

A Deus, aos meus pais Francisco Roberto de Favari e Maria Stella Ramin de Favari, à minha namorada Milena Garcia de Oliveira e ao meu irmão Marcos Roberto de Favari todo amor, compreensão, incentivo, suporte e carinho que me motivaram a prosseguir, mesmo nos momentos mais difíceis.

À equipe de qualidade da MWM-International, em especial ao funcionário Anderson de Oliveira o fornecimento dos dados e das fotos dos equipamentos de medição.

Aos meus amigos, em especial, Amanda, Amanda Manfrim, Cibele, Fábio, Paulo o apoio e a amizade que conquistamos.

A todos os colegas, professores e funcionários do Instituto de Ciências Matemáticas e de Computação da Universidade de São Paulo.

Resumo

Neste trabalho, aplicamos um modelo de regressão binária com erros de medição na variável explicativa para analisar sistemas de medição do tipo atributo. Para isto, utilizamos o modelo logístico com erros na variável, para o qual obtemos as estimativas de máxima verossimilhança via o algoritmo EM e a matriz de informação de Fisher observada. Além disso, fizemos um estudo de simulação para compararmos o método analítico e os modelos logístico sem erros na variável (ingênuo) e logístico com erros na variável. Finalmente, aplicamos nossa metodologia para avaliarmos um sistema de medição passa não passa da maior montadora de motores Diesel (MWM International).

Abstract

In this work, we apply a study of binary regression model with errors-in-variable to analyze attributive measurement systems. For this, we use the logistic model with errors-in-variable to obtain parameter estimates of maximum likelihood through EM algorithm and the observed Fisher information matrix. In addition we do a simulation study to compare analytic method and the logistic model with and without measurement errors-in-variable. Finally, we apply our methodology to evaluate a attributive measurement system for the largest Diesel motor company of the world (MWM International).

Conteúdo

1	Introdução	1
1.1	Motivação	3
1.2	Sistema de medição por atributo	5
1.3	Estudo do sistema de medição do tipo atributo	7
1.3.1	Descrição do estudo do sistema de medição do tipo atributo	8
1.4	Organização do Trabalho	9
2	Método Analítico	11
3	Modelo Estatístico e Estimação	17
3.1	Modelo proposto	18
3.2	Algoritmo EM	19
3.3	Estimação dos parâmetros	20
3.3.1	Modelo sem erros na variável explicativa	20
3.3.2	Modelo com erros na variável explicativa	20
3.4	Matriz de covariâncias para as estimativas dos parâmetros β_1 e β_2 do modelo ingênuo	26
3.5	Matriz de covariâncias para os estimadores dos parâmetros β_1, β_2, ω e Ω_w do modelo com erros	27
3.5.1	Cálculo das derivadas	28
3.6	Variância para a estimativa da tendência, utilizando o método Delta	29
3.6.1	Variância do estimador da tendência utilizando o teorema de Fieller	30
3.7	Repetitividade	32
4	Simulações	33
4.1	Comparação entre os modelos	34
4.2	Simulações para tendência	37
5	Análise do Sistema de Medição por Atributo	41
5.1	Modelo Ingênuo	42
5.1.1	Cálculo da variância para a estimativa da tendência	42
5.1.2	Cálculo da repetitividade	43

5.2	Modelo com erros na variável	43
5.2.1	Cálculo da variância para a estimativa da tendência	44
5.2.2	Cálculo da repetitividade	45
5.3	Propostas Futuras	48
Bibliografia		49

Lista de Tabelas

1.1	Dados do experimento	9
1.2	Dados Padronizados	9
2.1	Dados observados versus quantis da distribuição Normal	12
2.2	Dados observados versus probabilidade da peça ser aprovada	14
3.1	Apresentação dos dados	18
4.1	Valores Iniciais	34
4.2	Parâmetros das simulações	35
4.3	Resultados da simulação - Situação 1.	37
4.4	Resultados da simulação - Situação 2.	37
4.5	Resultados da simulação da tendência para os modelos Analítico, Ingênuo e Com erros - Situação 1	38
4.6	Resultados da simulação da tendência para os modelos Analítico, Ingênuo e Com erros - Situação 2	39
5.1	Estimativas e testes dos parâmetros - modelo ingênuo.	42
5.2	Estimativas e testes dos parâmetros - modelo com erros.	44
5.3	Probabilidades observadas e estimadas	46
5.4	Intervalo referente à tendência estimada - Método Delta, $\alpha = 0,05$	47
5.5	Intervalo referente à tendência estimada - Teorema de Fieller, $\alpha = 0,05$	47

Lista de Figuras

1.1	Sistema de medição inadequado	4
1.2	Relação entre o sistema de medição passa/não passa e os limites inferior e superior de engenharia (LIE e LSE).	5
1.3	Configuração do Sistema de medição passa não passa	6
1.4	Faixas de classificação	6
1.5	Peça Analisada - Carter de Automóvel	7
1.6	Zeragem do relógio comparador por um anel padrão	7
1.7	Relógio comparador	7
1.8	Calibrador tampão liso - Lado Passa	8
1.9	Calibrador tampão liso - Lado Não Passa	8
1.10	Classificação das peças no sistema de medição passa - não passa	9
1.11	Disposição das peças no sistema de medição passa - não passa em torno do LSE	9
1.12	Diferença entre o diâmetro e o LSE	9
2.1	Estimativa da Probabilidade de aceitação da peça versus valor da peça	12
2.2	Regressão ajustada = $\mathbf{x}(\pi) = -0,008117 - 0,02124 * \Phi^{-1}(\pi'_i)$	13
2.3	Probabilidade de a peça ser aprovada versus característica da qualidade	15
4.1	Impacto do erro de medição - situação 1	34
4.2	Impacto do erro de medição - situação 2	35
5.1	Probabilidade das aprovações obtidas pelo método analítico e dos modelos ingênuo e com erros versus valor padronizado da medição do diâmetro.	46

Capítulo 1

Introdução

Modelo de regressão constitui uma das técnicas mais utilizadas nas aplicações industriais, em particular, a regressão com variável resposta binária em que a resposta de um sistema é “zero” ou “um”, representa falha ou sucesso. Na maioria das aplicações industriais, estamos interessados em estabelecer uma relação entre a resposta binária do sistema e a variável explicativa que influencia no resultado do sistema. Em geral, os modelos probito e logito são utilizados para tratar tais aplicações, conforme McCullag & Nelder (1989), Collett (2003) e Hosmer & Lemeshow (1989).

As peças de uma empresa são manufaturadas com base em especificações técnicas para diferentes características da qualidade, como diâmetro, composição química, massa, entre outras. As especificações técnicas determinam intervalos para cada característica da qualidade. Por exemplo, o diâmetro de um furo em uma peça deve estar entre $9,4mm$ e $9,7mm$. Para avaliarmos nossa capacidade em atendermos estas especificações, precisamos medir (estimar) o valor destas características da qualidade para as peças que produzimos. O valor das características da qualidade das peças são medidos por sistemas de medição apropriados. Os sistemas de medição correspondem aos processos utilizados para obtermos o resultado da medição (estimativa), tais sistemas são compostos por pessoas, equipamentos, métodos, meio ambiente e a própria peça. Em geral, classificamos os sistemas de medição em dois grupos: variável e atributo. Os sistemas de medição por variável são aqueles que determinam um valor numérico para a característica da qualidade da peça, enquanto que os sistemas de medição por atributo classificam as peças em defeituosas ou não, conforme sua especificação. Devido ao baixo custo de fabricação e manutenção e à facilidade de manuseio, os sistemas de medição por atributo são um dos sistemas de medição mais utilizados na indústria automobilística mundial.

As montadoras americanas, General Motors Corporation, Ford Motor Company e Daimler-Chrysler Corporation, desenvolveram um manual para padronizar as técnicas de análise de sistemas de medição (MSA (2002)). Um dos capítulos deste manual (Capítulo III) é dedicado a sistemas de medição que classificam as peças em defeituosas ou não, também conhecido como sistemas de medição passa/não passa. O MSA (2002) discute três técnicas para análise:

i) Análise de concordância via a estatística kappa;

- ii) Método de Detecção de Sinal;
- iii) Método Analítico.

A estatística kappa determina o grau de concordância das medições do sistema de medição passa/não passa com um sistema de medição referência (por variável). Atualmente, é o método mais utilizado na indústria automobilística mundial. Porém, este método é empírico e é difícil definirmos um critério para avaliarmos o grau de concordância encontrado. Na prática, o manual MSA (2002) estabelece, como critério, um valor mínimo para a estatística kappa (0,75) baseado na "experiência" do grupo que elaborou o manual MSA (2002). O próprio manual comenta críticas em relação a esse método (MSA (2002), pg. 132).

O método de detecção de sinais avalia a região em que o sistema de medição passa/não passa comete erros de classificação. Como as peças foram medidas por um sistema de medição referência (por variável), ordenamos as peças em ordem decrescente e determinamos as regiões em que o sistema de medição passa não passa comete erros de classificação. Esse método também é empírico e não nos apresenta uma forma de avaliarmos o sistema de medição adequadamente.

O método analítico propõe estabelecermos uma relação entre o valor numérico da característica da qualidade da peça (exemplo, diâmetro) e a probabilidade do sistema de medição passa/não passa classificá-la como peça boa. Este é o método sugerido pelo manual MSA (2002). O manual MSA (2002) adotou o modelo probito com estimativa dos parâmetros de forma empírica, através do papel de probabilidade. Esse modelo será discutido no capítulo 2. Com este modelo, podemos avaliar duas propriedades do sistema de medição passa não passa:

- i) Tendência: diferença entre o sistema de medição e o Valor de Referência, para a característica da qualidade da peça;
- ii) Repetitividade: variação devido ao sistema de medição ao medir várias vezes a mesma grandeza de uma peça, utilizando o mesmo equipamento e método;

Essas propriedades são utilizadas para caracterizar sistemas de medição. A tendência representa a diferença entre as medições do sistema de medição passa/não passa e o sistema de medição referência. A repetitividade representa a variabilidade associada ao sistema de medição passa/não passa. Uma característica importante desse método consiste em obtermos essas características (tendência e repetitividade) na unidade da característica da qualidade da peça. Por exemplo, se a característica da qualidade corresponde ao diâmetro de um furo, expresso em milímetros (mm), obtemos a tendência e a repetitividade em milímetros, apesar do sistema de medição apresentar respostas binárias. Com isso, podemos testar se a tendência é nula e comparamos a repetitividade do sistema de medição com as especificações da peça.

A metodologia proposta pelas montadoras americanas no MSA (2002) para avaliar sistemas de medição por atributo via o método analítico, apresenta algumas deficiências.

Primeiro, o método de estimação via o papel de probabilidade. Segundo, para determinarmos o valor de referência de uma peça, devemos medi-las com sistemas apropriados. Porém, estes sistemas de medição apresentam erros de medição, que devem ser considerados no modelo. Neste trabalho, vamos aplicar modelos de regressão binária com erro na variável explicativa (vide Carroll *et al.*, 1984; Stefanski & Carroll, 1985; Schafer, 1987) para analisar sistemas de medição do tipo atributo. Para isto, utilizamos o modelo logístico com erro na variável, para o qual obtemos as estimativas de máxima verossimilhança via o algoritmo EM e a matriz de informação de Fisher observada. Além disso, fizemos um estudo de simulação para compararmos os modelos analítico, logístico sem erro na variável (ingênuo) e logístico com erro na variável. Finalmente, aplicamos nossa metodologia para avaliarmos um sistema de medição passa/não passa da maior montadora de motores Diesel para veículos leves do mundo, a MWM-International.

1.1 Motivação

O benefício da tomada de ação baseado em dados é determinado pela qualidade dos dados de medição utilizados. Se a qualidade for ruim, o benefício da ação será provavelmente baixo. De maneira similar, se a qualidade for boa, o benefício será provavelmente alto. Para assegurar que o benefício decorrente da utilização de dados de medição seja grande o suficiente para compensar o custo de obter esses dados, é necessário dedicar a devida atenção à qualidade deles.

A qualidade dos dados de medição está relacionada com as propriedades estatísticas de medições múltiplas obtidas a partir de um sistema de medição operando sob condições estabilizadas. Suponhamos que um sistema de medição, operando sob condições estabilizadas, seja utilizado para obter várias medições de uma certa característica. Se as medidas estiverem todas “próximas” do valor de referência, então a qualidade dos dados será “alta”. Da mesma forma, se algumas ou todas as medidas estiverem “afastadas” do valor referência, então a qualidade dos dados será considerada baixa.

As propriedades estatísticas mais utilizadas para caracterizar a qualidade dos dados são estabilidade, tendência, variabilidade e o grau de concordância. A propriedade chamada tendência refere-se à localização dos dados com relação ao valor característico, a propriedade chamada de variabilidade refere-se à dispersão dos dados. Porém, outras propriedades estatísticas, como a quantidade de aceitações de uma peça em relação ao valor característico, também são apropriadas para os sistemas passa/não passa.

Uma das razões mais comuns para dados de baixa qualidade é a variabilidade excessiva das medições. Grande parte da variação em um conjunto de medições é devido à interação entre o sistema de medição e o seu meio. Se esta interação vier a gerar muita variação, a qualidade dos dados poderá ser tão baixa que eles não terão utilidade.

Caso o sistema de medição não seja confiável, há um comprometimento na procedência dos produtos que são aprovados ou reprovados por este sistema e na tomada de decisões.

Um sistema de medição com variabilidade alta, pode aprovar produtos ruins, o que é um dos erros mais assustadores para uma empresa, pois põem em risco a reputação de seus produtos no mercado consumidor. E também um sistema que não reflete a realidade de um processo, pode rejeitar produtos bons, o que aumentaria o gasto com retrabalho desnecessário, entre outros prejuízos.

Apesar de ser um medidor da qualidade, o sistema de medição não está imune aos fatores que afetam um processo, sendo que podemos encontrar vários fatores que influenciam direta ou indiretamente os sistemas de medição. Entre eles, podemos citar o meio ambiente, as pessoas, os equipamentos, os padrões, os métodos e a própria peça.

O manual MSA (2002) propõe que para gerenciarmos efetivamente a variação de qualquer processo, precisamos conhecer três elementos básicos:

1. O que o processo de medição deveria fazer, isto é, qual o produto do processo;
2. O que pode sair errado;
3. O que o processo de medição está fazendo.

As especificações e os requerimentos de engenharia definem o que o processo deveria estar fazendo. Um sistema de medição ideal produziria somente medições “corretas” a cada vez que fosse utilizado. No entanto, sistemas de medições com tal propriedade não existem. Desta forma, um sistema de “má qualidade” poderá mascarar a variação real do processo ou produto conduzindo a conclusões erradas, conforme Figura 1.1.

Figura 1.1: Sistema de medição inadequado

O responsável pelo sistema de medição tem a obrigação de identificar quais as propriedades estatísticas para avaliar os resultados do sistema de medição. Apesar de cada sistema de medição requerer diferentes propriedades estatísticas, existem certas propriedades fundamentais que definem um bom sistema de medição. Algumas propriedades são listadas abaixo, conforme manual MSA (vide MSA, 2002, p.13).

- 1) Uma adequada discriminação ou sensibilidade. O incremento de medida deve ser pequeno o suficiente para detectar variações no processo ou nos limites de especificação.
- 2) O sistema de medição deve estar sob controle estatístico. Isso significa que as variações do sistema de medição são devidas a causas comuns e não devidas a causas especiais.
- 3) Para controle de produto, a variabilidade do sistema de medição deve ser pequena comparada com os limites de especificação.

4) Para controle do processo, a variabilidade do sistema de medição deve demonstrar uma resolução efetiva e pequena comparada com a variação do processo de manufatura.

Como já havíamos mencionado anteriormente, grande parte da variação em um conjunto de medições ocorre devido à interação entre o sistema de medição e seu meio. Os principais fatores que afetam a qualidade das medições, são padrão (valor conhecido dentro dos limites aceitáveis de incerteza), peça, equipamentos, pessoa, procedimento e meio ambiente.

1.2 Sistema de medição por atributo

Uma forma barata e rápida de as empresas controlarem seus processos e produtos consiste na utilização de sistemas de medição passa/não passa (go no go). Estes sistemas classificam as peças em defeituosas ou não, sem determinar o valor numérico da característica da qualidade em análise. Apesar de baratos e ágeis, esses sistemas de medição são passíveis de falhas, o que pode ser crucial para a confiabilidade das empresas perante seus clientes. O sistema de medição passa/não passa determina se a característica da peça pertence ao intervalo entre os limites de engenharia.

Figura 1.2: Relação entre o sistema de medição passa/não passa e os limites inferior e superior de engenharia (LIE e LSE).

A Figura 1.2 ilustra o sistema de medição passa não passa.

Para classificar a peça, o sistema de medição passa/não passa apresenta duas faces. A face passa é confeccionada com algumas frações da unidade da característica da qualidade acima do LIE, enquanto que a face não passa é confeccionada com algumas frações da unidade da característica da qualidade abaixo do LSE, conforme Figura 1.3.

Aqui, propomos um modelo de regressão logística, que associa o valor numérico da peça com a probabilidade de o sistema medição passa/não passa classificar a peça como aprovada (ou, não defeituosa). Apesar de o valor numérico ser determinado por um sistema de medição confiável, apresenta erros de medição. Por isso, é importante incorporarmos esses erros ao modelo logístico.

Quando uma peça está “próxima” aos limites de engenharia, o sistema de medição

Figura 1.3: Configuração do Sistema de medição passa não passa

Figura 1.4: Faixas de classificação

passa/não passa está suscetível a cometer erros de classificação. Assim, de acordo com a Figura 1.4, temos três faixas:

Faixa I. A peça considerada reprovada pelo sistema de medição referência e pelo sistema de medição passa/não passa.

Faixa II. Região próxima aos limites de engenharia que nos leva a erros de classificação pelo sistema de medição passa/não passa.

Faixa III. A peça considerada reprovada pelo sistema de medição referência e pelo sistema de medição passa/não passa.

Os principais objetivos deste trabalho consistem em incorporar erros de medição na análise de sistemas do tipo passa/não passa, para os quais vamos

- i) Determinar e testar a tendência;
- ii) Determinar a variabilidade;
- iii) Determinar a faixa II;
- iv) Realizar um estudo de simulação para avaliarmos o impacto dos erros de medição no modelo logístico.

1.3 Estudo do sistema de medição do tipo atributo

Na sequência, apresentamos o experimento proposto pelo manual MSA (2002) para avaliarmos um sistema de medição passa/não passa via o método analítico. Como os lados passa e não passa são confeccionados juntos, na prática avaliamos apenas um dos lados. Assim, vamos tomar o lado referente ao limites superior de engenharia (LSE) para análise. Aqui, realizamos o seguinte experimento:

- Seleccionamos k peças na produção que estejam em torno do LSE, vide (Figura 1.5);

Figura 1.5: Peça Analisada - Carter de Automóvel

- As peças são medidas em um sistema de medição confiável (referência). Nesse caso, determinamos o valor numérico da característica da qualidade, vide figuras 1.6 e 1.7;

Figura 1.6: Zeragem do relógio comparador por um anel padrão

Figura 1.7: Relógio comparador

- As mesmas peças são avaliadas pelo sistema passa/não passa (calibrador tampão liso) n vezes, vide figuras 1.8 e 1.9.
- Seleccionamos as peças de tal forma que a menor (maior) peça seja aprovada (reprovada) em todas as avaliações.

Figura 1.8: Calibrador tampão liso - Lado Passa

Figura 1.9: Calibrador tampão liso - Lado Não Passa

Para ilustrar, consideramos o sistema de medição passa/não passa calibrador tampão liso que classifica as peças em defeituosas ou não, vide figuras 1.8 e 1.9. Esse sistema de medição é aplicado na empresa MWM International para controlar a característica diâmetro dos furos de fixação no carter do automóvel (Figura 1.5), com especificação de $LIE = 9,3 \text{ mm}$, $LSE = 9,7 \text{ mm}$ e Valor Nominal $= 9,5 \text{ mm}$. Para um dispositivo de medição de limite duplo, por conveniência, somente o limite superior será analisado (não-passa), tomando as devidas suposições de linearidade e uniformidade do erro (Figura 1.10).

1.3.1 Descrição do estudo do sistema de medição do tipo atributo

- Seleccionamos 8 peças na linha de produção que foram enviadas ao laboratório de medidas dimensionais para determinarmos o valor numérico da característica da qualidade da peça, isto é, o valor do diâmetro, medido pelo relógio comparador (Figura 1.7) zerado por um anel padrão, com o auxílio de um súbito (Figura 1.6), cuja resolução é de $0,01 \text{ mm}$ e o desvio padrão associado é de $0,006 \text{ mm}$.
- Após medirmos as peças, estas retornaram à linha de produção onde foram medidas $n = 20$ vezes (cada peça) pelo sistema de medição passa/não passa (o tampão liso). As peças estão distribuídas conforme Figura 1.11.
- Após a realização do experimento resumimos os dados conforme Tabela 1.1.
- Para facilitar a análise da tendência, subtraímos o LSE do resultado da medição da peça pelo sistema de medição referência, conforme mostra a Tabela 1.2.

As figuras 1.10, 1.11 e 1.12 apresentam as medições das peças em relação ao LSE.

Tabela 1.1: Dados do experimento

Peça	Diâmetro	Aprovação	Total de Medições
1	9,64	20	20
2	9,65	20	20
3	9,67	20	20
4	9,68	8	20
5	9,7	5	20
6	9,71	3	20
7	9,73	0	20
8	9,75	0	20

Tabela 1.2: Dados Padronizados

Peça	Diâmetro	Aprovação	Total de Medições
1	-0,06	20	20
2	-0,05	20	20
3	-0,03	20	20
4	-0,02	8	20
5	0	5	20
6	0,01	3	20
7	0,03	0	20
8	0,05	0	20

Figura 1.10: Classificação das peças no sistema de medição passa - não passa

Figura 1.11: Disposição das peças no sistema de medição passa - não passa em torno do LSE

Figura 1.12: Diferença entre o diâmetro e o LSE

1.4 Organização do Trabalho

Em relação a sua organização, este trabalho foi estruturado em cinco capítulos.

Neste Capítulo abordamos, de uma forma sucinta, as definições relativas ao entendimento sobre Sistema de medição por Atributo e apresentamos os dados como demonstração do método.

No Capítulo 2 definimos o método analítico proposto pelo manual MSA (vide MSA, 2002).

No Capítulo 3 definimos o modelo utilizado para explicar os dados, apresentamos um algoritmo para determinarmos estimativas de máxima verossimilhança para os parâmetros em questão e seus respectivos intervalos de confiança.

No Capítulo 4 apresentamos uma análise gráfica e um estudo de simulação para comparar os modelos logísticos com e sem erros.

No Capítulo 5 apresentamos um exemplo de aplicação no qual utilizamos o nosso modelo para conduzir um estudo de Sistema de medição por Atributo na indústria automobilística MWM-International.

Capítulo 2

Método Analítico

Considere uma peça P_i onde controlamos uma característica da qualidade, por exemplo, o diâmetro. Denotamos por π_i a probabilidade de a peça ser aprovada pelo sistema de medição passa/não passa. Na prática, é importante relacionarmos o valor numérico da característica da qualidade da peça ($\mathbf{x}_i$) com a probabilidade de aceitação da mesma pelo sistema de medição passa/não passa. Assim, a peça P_i é enviada ao laboratório para medição da característica da Qualidade. Na nossa aplicação, a peça P_i é medida por um relógio comparador com resolução de 0,01 mm, sendo que subtraímos o LSE de todas as medições, conforme Tabela 1.2. De forma geral, tomamos

$$\pi_i = g(\mathbf{x}_i)$$

para todo $i = 1, 2, 3, \dots, k$ peças. Em particular, o MSA (2002) toma o modelo probito,

$$\pi_i = \Phi\left(\frac{\mathbf{x}_i - \mu}{b}\right). \quad (2.1)$$

sendo que Φ representa a função de distribuição acumulada da distribuição normal padrão e μ : representa a média das medições das peças, subtraída do LSE com a aplicação do sistema passa/não passa (tendência),

$\sigma^2 = b^2$: variância das medições das peças com aplicação do sistema de medição passa não passa.

Assim, podemos definir as grandezas de interesse como os parâmetros tendência (μ) e repetitividade ($|b| \sqrt{n}$), sendo n o número de aplicações do sistema de medição passa/não passa em cada peça. Através da expressão (2.1), temos que

$$\left(\frac{\mathbf{x}_i - \mu}{b}\right) = \Phi^{-1}(\pi_i), \quad (2.2)$$

e

$$\mathbf{x}_i = \mu + b\Phi^{-1}(\pi_i). \quad (2.3)$$

O manual MSA (2002) propõe que as probabilidades sejam estimadas segundo a expressão (2.4) e que o método de mínimos quadrados seja aplicado para estimarmos a tendência e a repetitividade.

Figura 2.1: Estimativa da Probabilidade de aceitação da peça versus valor da peça

Tabela 2.1: Dados observados versus quantis da distribuição Normal

Peça (i)	Valor (x_i)	Aprovação (y_i)	Total (n_i)	π'_i	$\Phi^{-1}(\pi'_i)$
1	-0,06	20	20	0,975	1,959964
2	-0,05	20	20	0,975	1,959964
3	-0,03	20	20	0,975	1,959964
4	-0,02	8	20	0,425	-0,189118
5	0	5	20	0,275	-0,597760
6	0,01	3	20	0,175	-0,934589
7	0,03	0	20	0,025	-1,959964
8	0,05	0	20	0,025	-1,959964

Na Figura 2.1, podemos visualizar o valor da medição da peça (x_i) com relação às estimativas da probabilidade de aceitação de cada peça. Com estas estimativas, aplicamos o modelos de regressão linear simples para estimarmos os parâmetros de interesse.

$$\pi'_i = \begin{cases} \frac{y_i + 0,5}{n} & , \text{ se } \frac{y_i}{n} < 0,5 ; \\ \frac{y_i - 0,5}{n} & , \text{ se } \frac{y_i}{n} > 0,5 ; \\ 0,5 & , \text{ se } \frac{y_i}{n} = 0,5 \end{cases} \quad (2.4)$$

A partir dos resultados de nosso experimento (Tabela 1.2), obtemos as estimativas para a probabilidade de aceitação da peça π'_i e para o quantil associado $\Phi^{-1}(\pi'_i)$ conforme Tabela 2.1.

As estimativas de mínimos quadrados são dadas por,

$$\mu' = -0,008117 \quad e \quad \sigma' = |-0,02124| = 0,02124.$$

Com isso, concluímos que a estimativa da tendência do sistema de medição passa não passa é dada por $-0,008117 \text{ mm}$. Além disso, a repetitividade do sistema de medição passa não

passa é dada por:

$$\sigma'_{rep} = |b'| \sqrt{n} = 0,02124 \times \sqrt{20} = 0,095 \text{ mm}.$$

Apesar de a estimativa da tendência ser pequena em relação à tolerância (0,4mm), obtemos que a repetitividade é grande. No capítulo 5 faremos uma discussão das causas de uma repetitividade alta.

Figura 2.2: Regressão ajustada = $x(\pi) = -0,008117 - 0,02124 * \Phi^{-1}(\pi'_i)$

Um ponto importante para análise consiste em testarmos se a tendência do sistema de medição passa não passa é significativa do ponto de vista estatístico. Para isto, o manual MSA (2002) propõe um teste para avaliarmos a tendência,

$$H_0 : \mu = 0$$

$$H_1 : \mu \neq 0.$$

A estatística do teste proposta pelo MSA (2002) é dada por, sob H_0 ,

$$t_0^* = \frac{(k-1) * \sqrt{n} * \mu'}{\sigma'_{rep}} \approx t_{(n-1)}.$$

Na aplicação, obtemos p-valor de 0,007472 e com isso, rejeitamos H_0 para $\alpha = 0,05$. Portanto o sistema de medição passa não passa apresenta uma tendência significativa de -0,008 mm. Para uma tolerância (Tol = LSE - LIE = 9,7 - 9,3 = 0,4 mm) de 0,4 mm concluímos que a tendência representa 2%. Em geral, uma tendência desta ordem (2% da tolerância) não incomoda. No nosso exemplo, esta tendência foi detectada devido às características metrológicas do sistema de medição que mediu as peças (resolução de 0,01 mm e incerteza e 0,006 mm).

A partir da expressão (2.1) calculamos a estimativa da probabilidade do sistema de

Tabela 2.2: Dados observados versus probabilidade da peça ser aprovada

Peça	Valor	Total	Aprovação	π'_i	$\Phi^{-1}(\pi'_i)$	Prob. Peça aprovada (π_i)
1	-0,06	20	20	0,975	1,959964	0,992716
2	-0,05	20	20	0,975	1,959964	0,975700
3	-0,03	20	20	0,975	1,959964	0,848585
4	-0,02	20	8	0,425	-0,189118	0,712099
5	0	20	5	0,275	-0,597760	0,351160
6	0,01	20	3	0,175	-0,934589	0,196816
7	0,03	20	0	0,025	-1,959964	0,036345
8	0,05	20	0	0,025	-1,959964	0,003105

medição passa não passa aprovar a peça. Assim, considerando $x_T = -0,06$, temos que::

$$\begin{aligned}
 \pi'(-0,06) &= \Phi\left(\frac{(-0,06 - (-0,008117))}{-0,02124}\right) \\
 &= \Phi(2,44295) \\
 &= 0,9927163
 \end{aligned} \tag{2.5}$$

A Tabela 2.2 apresenta os resultados das estimativas da probabilidade do sistema de medição passa não passa aprovar peças para diferentes valores da característica da qualidade destas peças. Na sequência, apresentamos o gráfico (Figura 2.3) para que possamos visualizar a probabilidade de aceitação em função do valor da característica da qualidade da peça, em nosso exemplo do seu diâmetro. Este gráfico é importante para que o grupo de engenharia possa avaliar a faixa de valores do diâmetro da peça na qual o sistema de medição passa não passa possa cometer erros de classificação.

Neste trabalho, vamos manter a forma de análise proposta pelo manual MSA (2002), no qual avaliamos os parâmetros de tendência e repetitividade. Porém, como as peças foram medidas por um sistema de medição (que inclui relógio comparador, operador, método, meio ambiente e a peça) temos o erro de medição associado. Assim, vamos estender o método proposto pelo MSA (2002) em dois aspectos:

- i) Método de estimação;
- ii) Erro na variável explicativa.

Figura 2.3: Probabilidade de a peça ser aprovada versus característica da qualidade

Capítulo 3

Modelo Estatístico e Estimação

Neste capítulo, vamos apresentar um possível modelo para descrever os dados da análise do sistema de medição passa/não passa. Para isto, propomos o modelo de regressão logística com erros na variável, conforme Schafer (1987). Na seqüência, calculamos a matriz de informação de Fisher observada (conforme método Louis (1982)) para realizarmos inferência sobre os parâmetros do modelo. Finalmente, estimamos a tendência e a repetitividade do sistema de medição e desenvolvemos um teste para avaliarmos a tendência.

Na nossa aplicação, análise de sistemas de medição passa/não passa, a variável explicativa é medida na presença de erros de medição. No nosso caso, o diâmetro do furo da peça é medido por um relógio comparador de resolução 0,01 *mm*, correspondente ao sistema de medição referência. Como todos os equipamentos de medição de uma empresa, o relógio comparador é calibrado periodicamente, na MWM - International temos um período entre calibrações de 3 meses. A calibração consiste em comparar o equipamento a ser calibrado com um equipamento padrão (vide ISOGUM, 1998). Como resultado da calibração temos uma estimativa da média e uma estimativa do desvio padrão do erro de medição do equipamento (relógio comparador).

Com isso, na análise do sistema de medição passa/não passa, temos informação a priori sobre o sistema de medição (relógio comparador). Para incorporarmos essa informação no modelo, adotamos a seguinte estrutura para o valor observado da característica da qualidade da peça (diâmetro) medida no sistema referência (relógio comparador),

$$x_i = w_i + \epsilon_i,$$

sendo x_i corresponde às observações da característica da qualidade da peça medida na presença de erro de medição, w_i corresponde ao verdadeiro valor da característica da qualidade da peça e ϵ_i corresponde ao erro de medição, com $i = 1, \dots, k..$ Devido ao processo de calibração do equipamento de referência (relógio comparador), assumimos que o parâmetro dos erros de medição ϵ_i são conhecidos.

Tabela 3.1: Apresentação dos dados

Peça	Número de medições	Número de aceitações	Variáveis explicativas	
			Sem erros (intercepto)	Com erros
1	n	$y_1 = \sum_{j=1}^{n_1} y_{1j}$	1	x_1
2	n	$y_2 = \sum_{j=1}^{n_2} y_{2j}$	1	x_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	n	$y_i = \sum_{j=1}^{n_i} y_{ij}$	1	x_i
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
k	n	$y_k = \sum_{j=1}^{n_k} y_{kj}$	1	x_k

3.1 Modelo proposto

Nesta seção, apresentamos o modelo de regressão logística com erros na variável conforme Schafer (1987). Na sequência, obtemos aproximações para as estimativas de máxima verossimilhança via o algoritmo EM. Seja y uma variável aleatória com distribuição binomial com função de distribuição dada por:

$$f(y; \pi) = \binom{n}{y} \pi^y (1 - \pi)^{(n-y)} I_A(y), \quad \pi \in [0, 1], \quad A = \{0, 1, \dots, n\}.$$

Então,

$$\begin{aligned} f(y; \pi) &= \exp \left\{ \binom{n}{y} + y \log \pi + (n - y) \log(1 - \pi) \right\} I_A(y) \\ &= \exp \left\{ y \log \left(\frac{\pi}{1 - \pi} \right) + n \log(1 - \pi) + \log \binom{n}{y} \right\} I_A(y). \end{aligned}$$

Ao denotarmos

$$\begin{aligned} a(\phi) &= 1, \quad \theta = \log \left(\frac{\pi}{1 - \pi} \right) \Rightarrow \pi = \frac{\exp(\theta)}{1 + \exp(\theta)}, \\ b(\theta) &= -n \log(1 - \pi) = n \log(1 + \exp(\theta)), \quad c(y; \phi) = \log \binom{n}{y} \end{aligned}$$

obtemos que a distribuição binomial pertence à família exponencial na forma

$$f(y; \theta, \phi) = \exp \left\{ \frac{1}{a(\phi)} [y(\theta) - b(\theta)] + c(y; \phi) \right\} I_A(y), \quad (3.1)$$

conforme obtido em Demétrio (2002).

De acordo com a Tabela 3.1, denotamos por y_i o número de aceitações da i -ésima peça pelo sistema de medição passa/não passa ao ser aplicado n vezes. Aqui, assumimos que y_i , condicionado em w_i , tem a seguinte função de distribuição de probabilidade:

$$f(y_i | w_i; \beta) = \exp \{ y_i(\beta_1 + w_i \beta_2) - b(\beta_1 + w_i \beta_2) + h(y_i) \}, \quad (3.2)$$

sendo $\beta = (\beta_1, \beta_2)$ são os parâmetros do modelo logito e

$$a(\phi) = 1, \quad \theta_i = \log \left(\frac{\pi_i}{1 - \pi_i} \right) \Rightarrow \pi_i = \frac{\exp(\theta_i)}{1 + \exp(\theta_i)},$$

$$b(\theta_i) = n \log(1 + \exp(\theta_i)), \quad h(y_i) = \log \binom{n}{y_i}$$

sendo $\theta_i = \beta_1 + w_i \beta_2$ e $i = 1, \dots, k$. Para completar o modelo, temos a variável w_i que é observada indiretamente, somente através da medição da característica da qualidade na presença de erros de medição (x_i). Assim, temos

$$x_i = w_i + \epsilon_i, \quad w_i \sim N(\omega; \Omega_w), \quad \epsilon_i \sim N(0, \Omega_m) \quad \text{e} \quad x_i \sim N(\omega, \Omega_m + \Omega_w),$$

com

- i. ω : corresponde a média do verdadeiro valor da característica da qualidade da i -ésima peça;
- ii. Ω_w : corresponde à variância do verdadeiro valor da característica da qualidade da i -ésima peça;
- iii. Ω_m : corresponde à variância dos erros de medição da i -ésima peça. Assumimos que Ω_m foi determinado durante o processo de calibração do equipamento referência (relógio comparador). Portanto, admitimos Ω_m conhecido;

Consideramos que os erros ϵ_i são independentes entre si e de w_i . Assumimos que y_i e x_i são condicionalmente independentes, dado w_i , e que todas quantidades envolvendo i e i' são independentes, para $i \neq i'$. Denotamos por $\mathbf{y}' = (y_1, \dots, y_k)$, $\mathbf{x}' = (x_1, \dots, x_k)$, $\mathbf{w}' = (w_1, \dots, w_k)$ e $\boldsymbol{\xi}$ como o vetor contendo os parâmetros β , ω e Ω_w . Dessa forma, os parâmetros a serem estimados são

- ω : média da variável explicativa observada com erros de medição (x_i);
- Ω_w : variância da variável explicativa não observável (w_i);
- β_1 : parâmetro relativo à variável explicativa observada sem erros de medição (intercepto);
- β_2 : parâmetro relativo à variável explicativa observada com erros de medição (x_i).

3.2 Algoritmo EM

A metodologia proposta em Schafer (1987) é usada para obter estimativas de coeficientes de regressão para modelos lineares generalizados com função de ligação canônica quando temos uma variável medida na presença de erro de medição. Por tratar a variável não observada ($\mathbf{w}$) como “dado perdido”, Schafer (1987) sugere algoritmo EM, no qual cada ciclo consiste de um passo E de aproximações de primeiros e segundos momentos condicionais da variável não observada dada a observada, seguidos por um passo M no qual as estimativas dos parâmetros de regressão são atualizados iterativamente por mínimos quadrados ponderados, baseados nas aproximações obtidas no passo E.

3.3 Estimação dos parâmetros

Nesta seção, apresentamos um método para estimar os parâmetros nos modelos logísticos com e sem erro de medição.

3.3.1 Modelo sem erros na variável explicativa

Vamos descrever o método de mínimos quadrados ponderados para estimarmos os coeficientes β_1 e β_2 da regressão logística sem erros na variável (modelo ingênuo), no qual observamos a variável resposta $\mathbf{y}$ e a variável explicativa $\mathbf{w}$ sem erros de medição.

A estimativa de β no $(s + 1)$ ciclo é dado pela expressão (3.3) a seguir:

$$\beta^{(s+1)} = \beta^{(s)} + \left\{ \sum_{i=1}^k m_i^{(s)} \begin{bmatrix} 1 & w_i \\ w_i & w_i^2 \end{bmatrix} \right\}^{-1} \sum_{i=1}^k \left\{ (y_i - \dot{b}(\theta_i^{(s)})) \begin{pmatrix} 1 \\ w_i \end{pmatrix} \right\}, \quad (3.3)$$

Considerando a expressão

$$m_i^{(s)} = \ddot{b}(\theta_i^{(s)}) \quad z_i^{(s)} = \frac{y_i - \dot{b}(\theta_i^{(s)})}{\ddot{b}(\theta_i^{(s)})} + \theta_i^{(s)},$$

obtemos

$$y_i - \dot{b}(\theta_i^{(s)}) = m_i^{(s)} z_i^{(s)} - m_i^{(s)} \theta_i^{(s)}, \quad (3.4)$$

sendo

$$\theta_i^{(s)} = \beta_1^{(s)} + w_i \beta_2^{(s)}, \quad (3.5)$$

$\dot{b}(\theta_i^{(s)})$ e $\ddot{b}(\theta_i^{(s)})$ são as derivadas primeira e segunda de $b(\theta_i^{(s)})$.

Substituindo as expressões 3.4 e 3.5 em 3.3 e após algumas manipulações algébricas, resultamos na expressão (3.6) dada a seguir

$$\beta^{(s+1)} = \left\{ \sum_{i=1}^k m_i^{(s)} \begin{bmatrix} 1 & w_i \\ w_i & w_i^2 \end{bmatrix} \right\}^{-1} \sum_{i=1}^k \left\{ m_i^{(s)} z_i^{(s)} \begin{pmatrix} 1 \\ w_i \end{pmatrix} \right\}, \quad (3.6)$$

através da qual obtemos as estimativas usuais (ingênuas) de β_1 e β_2 do modelo linear generalizado sem erros na variável explicativa.

3.3.2 Modelo com erros na variável explicativa

No modelo com erros na variável, a variável $\mathbf{w}$ não é observada diretamente, mas medida na presença de erro de medição. Assim, vamos tratá-la como “dado perdido”, para que possamos utilizar o algoritmo EM (Dempster *et al.*, 1977) como forma de encontrarmos aproximações para as estimativas de máxima verossimilhança. Na aplicação do algoritmo EM, os “dados completos” são $\mathbf{y}$, $\mathbf{x}$ e $\mathbf{w}$, enquanto os “dados observados” são $\mathbf{y}$ e $\mathbf{x}$. Em sua forma mais geral, o algoritmo EM define um processo iterativo para estimação de ξ através

de dois passos (E e M). Sendo assim, descreveremos os passos e as iterações do algoritmo EM para o modelo com erros de medição na variável explicativa.

Passo E: Calculamos a esperança do logaritmo dos dados completos condicionado nos dados observados

$$Q(\boldsymbol{\xi} \mid \boldsymbol{\xi}^{(t)}) = E\{\log f(\mathbf{y}, \mathbf{x}, \mathbf{w}; \boldsymbol{\xi}) \mid \mathbf{y}, \mathbf{x}; \boldsymbol{\xi}^{(t)}\}.$$

Passo M: Encontramos os estimadores de máxima verossimilhança desta esperança condicional

$$\boldsymbol{\xi}^{(t+1)} = \operatorname{argmax} \left(Q(\boldsymbol{\xi} \mid \boldsymbol{\xi}^{(t)}) \right).$$

De acordo com as suposições das variáveis envolvidas no modelo, podemos escrever

$$\begin{aligned} f(y_i, x_i, w_i) &= f(y_i \mid x_i, w_i) f(x_i \mid w_i) f(w_i) \\ &= f(y_i \mid w_i) f(x_i \mid w_i) f(w_i). \end{aligned} \quad (3.7)$$

Com isso, a partir da expressão (3.7) obtemos a distribuição dos dados completos, expressa por

$$f(y_i, x_i, w_i) = f(y_i \mid w_i) f(x_i \mid w_i) f(w_i), \quad (3.8)$$

A partir da expressão (3.8) obtemos o logaritmo da função de verossimilhança do modelo com erros (levando em conta que Ω_m é conhecida) dada por

$$\log f(\mathbf{y}, \mathbf{x}, \mathbf{w}; \boldsymbol{\xi}) = \text{constante} + \sum_{i=1}^k \log f(y_i \mid w_i; \boldsymbol{\beta}) + \sum_{i=1}^k \log f(w_i; \omega, \Omega_w). \quad (3.9)$$

Uma vez obtida a expressão para o logaritmo da função de verossimilhança do modelo, descrevemos os passos.

Passo E: Calcula-se a esperança de (3.9) condicionada nos dados observados. Assim,

$$Q(\boldsymbol{\xi} \mid \boldsymbol{\xi}^{(t)}) = Q_1(\boldsymbol{\beta} \mid \boldsymbol{\xi}^{(t)}) + Q_2(\omega, \Omega_w \mid \boldsymbol{\xi}^{(t)}), \quad (3.10)$$

Daí, temos

$$\begin{aligned} Q_2(\omega, \Omega_w \mid \boldsymbol{\xi}^{(t)}) &= E \left\{ \sum_{i=1}^k \log f(w_i; \omega, \Omega_w) \mid \mathbf{y}, \mathbf{x}; \boldsymbol{\xi}^{(t)} \right\} \\ &= -\frac{1}{2} \sum_{i=1}^k E \left\{ \frac{(w_i - \omega)^2}{\Omega_w} \mid y_i, x_i; \boldsymbol{\xi}^{(t)} \right\} - \frac{k}{2} \log(2\pi) - \frac{k}{2} \log(\Omega_w). \end{aligned} \quad (3.11)$$

A partir da expressão (3.11), calculamos a esperança condicional de Q_2 . Adotamos a seguinte notação

$$\begin{aligned} w_i^{(t)} &= E(w_i \mid y_i, x_i; \boldsymbol{\xi}^{(t)}) \text{ e} \\ V_i^{(t)} &= \operatorname{var}(w_i \mid y_i, x_i; \boldsymbol{\xi}^{(t)}) = E \left[(w_i - w_i^{(t)})^2 \mid y_i, x_i; \boldsymbol{\xi}^{(t)} \right]. \end{aligned} \quad (3.12)$$

Antes de calcularmos a esperança (3.12) é conveniente escrevermos a forma quadrática da expressão (3.11) como

$$\begin{aligned}(w_i - \omega)^2 &= (w_i - w_i^{(t)} + w_i^{(t)} - \omega)^2 \\ &= \left[(w_i - w_i^{(t)}) + (w_i^{(t)} - \omega) \right]^2 \\ &= \left[(w_i - w_i^{(t)})^2 + 2(w_i - w_i^{(t)})(w_i^{(t)} - \omega) + (w_i^{(t)} - \omega)^2 \right].\end{aligned}\tag{3.13}$$

Calculando a esperança condicional de cada parcela da expressão (3.13), temos

$$\begin{aligned}\mathbb{E} \left[(w_i - w_i^{(t)})^2 \mid y_i, x_i; \boldsymbol{\xi}^{(t)} \right] &= V_i; \\ 2\mathbb{E} \left[(w_i - w_i^{(t)})(w_i^{(t)} - \omega) \mid y_i, x_i; \boldsymbol{\xi}^{(t)} \right] &= 2(w_i^{(t)} - \omega)\mathbb{E} \left[(w_i - w_i^{(t)}) \mid y_i, x_i; \boldsymbol{\xi}^{(t)} \right] \\ &= 2(w_i^{(t)} - \omega)(w_i^{(t)} - w_i^{(t)}) = 0 \\ \mathbb{E} \left[(w_i^{(t)} - \omega)^2 \mid y_i, x_i; \boldsymbol{\xi}^{(t)} \right] &= (w_i^{(t)} - \omega)^2\end{aligned}\tag{3.14}$$

Assim, de (3.14), temos a expressão de Q_2 , dada por (3.15).

$$Q_2(\omega, \Omega_w \mid \boldsymbol{\xi}^{(t)}) = -\frac{1}{2} \sum_{i=1}^k \left[\frac{V_i^{(t)} + (w_i^{(t)} - \omega)^2}{\Omega_w} \right] - \frac{k}{2} \log(2\pi) - \frac{k}{2} \log(\Omega_w).\tag{3.15}$$

A primeira parte da expressão (3.10), que envolve $\boldsymbol{\beta}$, temos que

$$\log f(y_i \mid w_i; \boldsymbol{\beta}) = y_i(\beta_1 + w_i\beta_2) - b(\beta_1 + w_i\beta_2) + h(y_i)\tag{3.16}$$

e como

$$\mathbb{E} \left\{ h(y_i) \mid \mathbf{y}, \mathbf{x}; \boldsymbol{\xi}^{(t)} \right\} = h(y_i),$$

que não envolve $\boldsymbol{\beta}$. Assim, basta trabalhar com

$$\begin{aligned}Q_1(\boldsymbol{\beta} \mid \boldsymbol{\xi}^{(t)}) &= \mathbb{E} \left\{ \sum_{i=1}^k \log f(y_i \mid w_i; \boldsymbol{\beta}) \mid \mathbf{y}, \mathbf{x}; \boldsymbol{\xi}^{(t)} \right\} \\ &= \sum_{i=1}^k \left[y_i(\beta_1 + w_i^{(t)}\beta_2) - \mathbb{E} \{ b(\beta_1 + w_i\beta_2) \mid y_i, x_i; \boldsymbol{\xi}^{(t)} \} \right].\end{aligned}\tag{3.17}$$

A primeira parcela da expressão (3.17) se justifica, pois

$$\begin{aligned}\mathbb{E} \left\{ y_i(\beta_1 + w_i\beta_2) \mid y_i, x_i; \boldsymbol{\xi}^{(t)} \right\} &= y_i \mathbb{E} \left\{ (\beta_1 + w_i\beta_2) \mid y_i, x_i; \boldsymbol{\xi}^{(t)} \right\} \\ &= y_i(\beta_1 + w_i^{(t)}\beta_2),\end{aligned}$$

vide notação (3.12).

Passo M: Derivando a expressão (3.15) em relação a ω, Ω_w , e igualando a zero, temos que (3.15) é maximizada por (3.18) e (3.19).

$$\omega^{(t+1)} = \bar{w}^{(t)} = \frac{1}{k} \sum_{i=1}^k w_i^{(t)}\tag{3.18}$$

e

$$\Omega_w^{(t+1)} = \frac{1}{k} \sum_{i=1}^k \left(V_i^{(t)} + (w_i^{(t)} - \bar{w}^{(t)})^2 \right). \quad (3.19)$$

Utilizamos o método de Newton para maximizar $Q_1(\boldsymbol{\beta} \mid \boldsymbol{\xi}^{(t)})$, com $\boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$.

Assim, devemos resolver o sistema dado por

$$\dot{Q}_1(\boldsymbol{\beta}^{(s)} \mid \boldsymbol{\xi}^{(t)}) + \ddot{Q}_1(\boldsymbol{\beta}^{(s)} \mid \boldsymbol{\xi}^{(t)})(\boldsymbol{\beta} - \boldsymbol{\beta}^{(s)}) = 0, \quad (3.20)$$

inicialmente, supomos que integração e diferenciação podem ser permutados e usando a regra da cadeia, temos as derivadas primeira e segunda de (3.17)

$$\dot{Q}_1(\boldsymbol{\beta} \mid \boldsymbol{\xi}^{(t)}) = \sum_{i=1}^k \left[y_i \begin{pmatrix} 1 \\ w_i^{(t)} \end{pmatrix} - E \left\{ \dot{b}(\beta_1 + w_i \beta_2) \begin{pmatrix} 1 \\ w_i \end{pmatrix} \mid y_i, x_i; \boldsymbol{\xi}^{(t)} \right\} \right]$$

e

$$\ddot{Q}_1(\boldsymbol{\beta} \mid \boldsymbol{\xi}^{(t)}) = - \sum_{i=1}^k \left[E \left\{ \ddot{b}(\beta_1 + w_i \beta_2) \begin{pmatrix} 1 \\ w_i \end{pmatrix} \begin{pmatrix} 1 & w_i \end{pmatrix} \mid y_i, x_i; \boldsymbol{\xi}^{(t)} \right\} \right].$$

Sendo assim, se $\boldsymbol{\beta}^{(s)} = \begin{pmatrix} \beta_1^{(s)} \\ \beta_2^{(s)} \end{pmatrix}$ é a estimativa de $\boldsymbol{\beta}$ na iteração s , temos que $\boldsymbol{\beta}^{(s+1)}$ é solução de

$$\begin{aligned} & \sum_{i=1}^k \left\{ y_i \begin{pmatrix} 1 \\ w_i^{(t)} \end{pmatrix} - E \left\{ \dot{b}(\beta_1^{(s)} + w_i \beta_2^{(s)}) \begin{pmatrix} 1 \\ w_i \end{pmatrix} \mid y_i, x_i; \boldsymbol{\xi}^{(t)} \right\} \right. \\ & \quad \left. - E \left\{ \ddot{b}(\beta_1^{(s)} + w_i \beta_2^{(s)}) \begin{pmatrix} 1 \\ w_i \end{pmatrix} \begin{pmatrix} 1 & w_i \end{pmatrix} \mid y_i, x_i; \boldsymbol{\xi}^{(t)} \right\} (\boldsymbol{\beta} - \boldsymbol{\beta}^{(s)}) \right\} = 0. \end{aligned} \quad (3.21)$$

Antes de prosseguirmos, devemos considerar as seguintes aproximações

$$\dot{b}(\beta_1^{(s)} + w_i \beta_2^{(s)}) \cong \dot{b}(\beta_1^{(s)} + w_i^{(t)} \beta_2^{(s)})$$

e

$$\ddot{b}(\beta_1^{(s)} + w_i \beta_2^{(s)}) \cong \ddot{b}(\beta_1^{(s)} + w_i^{(t)} \beta_2^{(s)}). \quad (3.22)$$

São adequadas se

$$E(\dot{b}(w)) = \dot{b}(E(w)) \text{ e } E(\ddot{b}(w)) = \ddot{b}(E(w)).$$

Com a aproximação (3.22) e a notação (3.12) temos

$$\begin{aligned} E \left\{ \dot{b}(\beta_1^{(s)} + w_i^{(t)} \beta_2^{(s)}) \begin{pmatrix} 1 \\ w_i \end{pmatrix} \mid y_i, x_i; \boldsymbol{\xi}^{(t)} \right\} &= \dot{b}(\beta_1^{(s)} + w_i^{(t)} \beta_2^{(s)}) E \left\{ \begin{pmatrix} 1 \\ w_i \end{pmatrix} \mid y_i, x_i; \boldsymbol{\xi}^{(t)} \right\} \\ &= \dot{b}(\beta_1^{(s)} + w_i^{(t)} \beta_2^{(s)}) \begin{pmatrix} 1 \\ w_i^{(t)} \end{pmatrix} \end{aligned}$$

e

$$\begin{aligned}
\mathbb{E} \left\{ \ddot{b}(\beta_1^{(s)} + w_i^{(t)} \beta_2^{(s)}) \begin{pmatrix} 1 \\ w_i \end{pmatrix} \mid y_i, x_i, \boldsymbol{\xi}^{(t)} \right\} &= \ddot{b}(\beta_1^{(s)} + w_i^{(t)} \beta_2^{(s)}) \mathbb{E} \left\{ \begin{bmatrix} 1 & w_i \\ w_i & w_i^2 \end{bmatrix} \mid y_i, x_i, \boldsymbol{\xi}^{(t)} \right\} \\
&= \ddot{b}(\beta_1^{(s)} + w_i^{(t)} \beta_2^{(s)}) \begin{bmatrix} 1 & w_i^{(t)} \\ w_i^{(t)} & \mathbb{E}(w_i^2 \mid y_i, x_i, \boldsymbol{\xi}^{(t)}) \end{bmatrix} \\
&= \ddot{b}(\beta_1^{(s)} + w_i^{(t)} \beta_2^{(s)}) \begin{bmatrix} 1 & w_i^{(t)} \\ w_i^{(t)} & V_i^{(t)} + (w_i^{(t)})^2 \end{bmatrix},
\end{aligned}$$

pois

$$\begin{aligned}
\mathbb{E} \left\{ w_i^2 \mid y_i, x_i, \boldsymbol{\xi}^{(t)} \right\} &= \text{var}(w_i \mid y_i, x_i, \boldsymbol{\xi}^{(t)}) + \left[\mathbb{E}(w_i \mid y_i, x_i, \boldsymbol{\xi}^{(t)}) \right]^2 \\
&= V_i^{(t)} + (w_i^{(t)})^2,
\end{aligned}$$

de acordo com a notação (3.12).

Sendo assim, o sistema (3.21) passa a ser

$$\begin{aligned}
\sum_{i=1}^k \left\{ y_i \begin{pmatrix} 1 \\ w_i^{(t)} \end{pmatrix} - \dot{b}(\beta_1^{(s)} + w_i^{(t)} \beta_2^{(s)}) \begin{pmatrix} 1 \\ w_i^{(t)} \end{pmatrix} \right. \\
\left. - \ddot{b}(\beta_1^{(s)} + w_i^{(t)} \beta_2^{(s)}) \begin{bmatrix} 1 & w_i^{(t)} \\ w_i^{(t)} & V_i^{(t)} + (w_i^{(t)})^2 \end{bmatrix} (\boldsymbol{\beta} - \boldsymbol{\beta}^{(s)}) \right\} = 0,
\end{aligned}$$

que leva a

$$\begin{aligned}
&\left\{ \sum_{i=1}^k \ddot{b}(\beta_1^{(s)} + w_i^{(t)} \beta_2^{(s)}) \begin{bmatrix} 1 & w_i^{(t)} \\ w_i^{(t)} & V_i^{(t)} + (w_i^{(t)})^2 \end{bmatrix} \right\} (\boldsymbol{\beta} - \boldsymbol{\beta}^{(s)}) \\
&= \sum_{i=1}^k \left[y_i - \dot{b}(\beta_1^{(s)} + w_i^{(t)} \beta_2^{(s)}) \right] \begin{pmatrix} 1 \\ w_i^{(t)} \end{pmatrix},
\end{aligned}$$

resultando em

$$\boldsymbol{\beta}^{(t+1), (s+1)} = \left\{ \sum_{i=1}^k m_i^{(t,s)} \begin{bmatrix} 1 & w_i^{(t)} \\ w_i^{(t)} & V_i^{(t)} + (w_i^{(t)})^2 \end{bmatrix} \right\}^{-1} \times \sum_{i=1}^k m_i^{(t,s)} z_i^{(t,s)} \begin{pmatrix} 1 \\ w_i^{(t)} \end{pmatrix}, \quad (3.23)$$

s é o índice da iteração de mínimos quadrados ponderados, no qual está inserido no algoritmo EM, indicado pelo índice t . As expressões $m_i^{(t,s)}$ e $z_i^{(t,s)}$ são expressas como em (3.6), mas com w_i substituído por $w_i^{(t)}$.

Destacamos a dificuldade em obter os momentos condicionais de w_i . Schafer (1987) sugere substituir os dois primeiros momentos pelos momentos de uma distribuição normal que aproxima a densidade $f(w_i \mid y_i, x_i)$. Considere $Z \sim N(\mu_Z, \Omega_Z)$. Sabemos que μ_Z é a moda da densidade de Z e que

$$\frac{\partial^2 \log f(z)}{\partial z^2} = -\Omega_Z^{-1}.$$

Por analogia com a distribuição normal, a solução consiste em igualar $E(w_i | y_i, x_i)$ moda de $f(w_i | y_i, x_i)$ e

$$\text{var}(w_i | y_i, x_i) = \left(-\frac{\partial^2 \log f(w_i | y_i, x_i)}{\partial w_i^2} \right)^{-1}.$$

Temos que

$$\begin{aligned} f(w_i | y_i, x_i) &= \frac{f(y_i, x_i, w_i)}{f(x_i, y_i)} = \frac{f(y_i | x_i, w_i)f(x_i | w_i)f(w_i)}{f(x_i, y_i)} = \frac{f(y_i | w_i)f(x_i | w_i)f(w_i)}{f(x_i, y_i)} \\ &\propto f(y_i | w_i)f(x_i | w_i)f(w_i), \end{aligned}$$

sendo que na penúltima passagem usamos a independência condicional entre y_i e x_i . A constante de proporcionalidade da normal não é importante, pois não envolvendo w_i não interfere no cálculo da moda e das derivadas segundas.

De acordo com (3.2) temos $x_i | w_i \sim N(w_i, \Omega_m)$ e $w_i \sim N(\omega, \Omega_w)$, $i = 1, \dots, k$. Portanto,

$$f(w_i | y_i, x_i) \propto \exp[y_i(\beta_1 + w_i\beta_2) - b(\beta_1 + w_i\beta_2)] \exp\left[-\frac{1}{2\Omega_m}(x_i - w_i)^2\right] \exp\left[-\frac{1}{2\Omega_w}(w_i - \omega)^2\right]. \quad (3.24)$$

Notando que a constante de proporcionalidade não envolve w_i . Segue que,

$$\log f(w_i | y_i, x_i) = y_i(\beta_1 + w_i\beta_2) - b(\beta_1 + w_i\beta_2) - \frac{1}{2\Omega_m}(x_i - w_i)^2 - \frac{1}{2\Omega_w}(w_i - \omega)^2, \quad (3.25)$$

A derivada primeira da expressão (3.25) é dada por

$$\frac{\partial \log f(w_i | y_i, x_i)}{\partial w_i} = y_i\beta_2 - \dot{b}(\beta_1 + w_i\beta_2)\beta_2 + \frac{(x_i - w_i)}{\Omega_m} - \frac{(w_i - \omega)}{\Omega_w}, \quad (3.26)$$

da mesma forma, a derivada segunda de (3.25) é

$$\frac{\partial^2 \log f(w_i | y_i, x_i)}{\partial w_i^2} = -\ddot{b}(\beta_1 + w_i\beta_2)\beta_2^2 - \Omega_m^{-1} - \Omega_w^{-1}. \quad (3.27)$$

Usando a notação (3.12) chegamos a

$$V_i^{(t)} = \left\{ \ddot{b}(\theta_i^{(t)})(\beta_2^{(t)})^2 + \Omega_m^{(t)-1} + \Omega_w^{(t)-1} \right\}^{-1}, \quad (3.28)$$

sendo

$$\theta_i^{(t)} = \beta_1^{(t)} + w_i^{(t)}\beta_2^{(t)}. \quad (3.29)$$

A moda da densidade $f(w_i | y_i, x_i)$ é obtida igualando a expressão (3.26) a zero, que resulta nos seguintes passos

$$\begin{aligned} y_i\beta_2 - \dot{b}(\theta_i)\beta_2 + \Omega_m^{-1}(x_i - w_i) - \Omega_w^{-1}(w_i - \omega) &= 0 \\ y_i\beta_2 - \dot{b}(\theta_i)\beta_2 + \Omega_m^{-1}x_i + \Omega_w^{-1}\omega &= (\Omega_m^{-1} + \Omega_w^{-1})w_i. \end{aligned} \quad (3.30)$$

Da expressão (3.28) segue que

$$\Omega_m^{-1} + \Omega_w^{-1} = V_i^{-1} - \ddot{b}(\theta_i)\beta_2^2, \quad (3.31)$$

substituindo (3.31) no resultado final da expressão (3.30), temos

$$y_i\beta_2 - \dot{b}(\theta_i)\beta_2 + \Omega_m^{-1}x_i + \Omega_w^{-1}\omega + \ddot{b}(\theta_i)\beta_2^2w_i = V_i^{-1}w_i$$

e

$$y_i\beta_2 - \dot{b}(\theta_i)\beta_2 + \Omega_m^{-1}x_i + \Omega_w^{-1}\omega + \ddot{b}(\theta_i)(\theta_i - \beta_1)\beta_2 = V_i^{-1}w_i,$$

na qual usamos a expressão (3.29). Finalmente,

$$w_i^{(t)} = V_i^{(t)} \left[\left\{ y_i - \dot{b}(\theta_i^{(t)}) - \ddot{b}(\theta_i^{(t)})(\beta_1^{(t)} - \theta_i^{(t)}) \right\} \beta_2^{(t)} + \Omega_m^{(t)-1}x_i + \Omega_w^{(t)-1}\omega^{(t)} \right]. \quad (3.32)$$

Dessa forma utilizamos a expressão (3.32) no cálculo de (3.23).

A seguir, temos um resumo dos passos E e M do algoritmo EM

Passo E: Calculamos as expressões (3.17) e (3.15).

Passo M: Calculamos as expressões (3.18), (3.19) e (3.23).

A cada iteração do algoritmo EM, $w_i^{(t)}$ e $V_i^{(t)}$ são calculados. As atualizações das estimativas dos parâmetros $\omega^{(t+1)}$, $\Omega_w^{(t+1)}$ e $\beta^{(t+1)}$ são obtidas pelas expressões (3.18), (3.19) e (3.23).

Como $\ddot{b} > 0$, temos que o valor da expressão (3.27) é negativo, logo a expressão (3.32) de fato é a moda.

A expressão de $w_i^{(t)}$ de certa maneira oculta a dependência no próprio w_i , que ocorre através de $\theta_i^{(t)} = \beta_1^{(t)} + w_i^{(t)}\beta_2^{(t)}$. Sendo assim, w_i também requer valores iniciais no processo iterativo, podendo-se tomar $w_i^{(0)} = x_i$, $i = 1, \dots, k$.

A estimativa inicial de ω é a média amostral da variável medida com erros (x_i). As estimativas iniciais de β_1 e β_2 são tomadas da regressão logística usual (estimativas ingênuas).

A estimativa inicial de Ω_w é a variância amostral da variável medida com erros (x_i), subtraída da variância dos erros de medição (Ω_m).

3.4 Matriz de covariâncias para as estimativas dos parâmetros β_1 e β_2 do modelo ingênuo

A matriz de informação de Fisher, extraída em Hosmer & Lemeshow (1989) é dada por

$$\hat{I}(\hat{\beta}) = \mathbf{H}'\mathbf{V}\mathbf{H}, \quad (3.33)$$

sendo

$$\mathbf{H} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_k \end{bmatrix}.$$

Para as probabilidades de aprovação temos a seguinte expressão

$$\hat{\pi}_i = \frac{\exp(\hat{\beta}_1 + \hat{\beta}_2 x_i)}{1 + \exp(\hat{\beta}_1 + \hat{\beta}_2 x_i)}, \quad (3.34)$$

$\hat{\beta}_1$ e $\hat{\beta}_2$, são dadas pela expressão (3.6).

e a matriz $\mathbf{V}$ é dada por:

$$\mathbf{V} = \begin{bmatrix} n(\hat{\pi}_1(1 - \hat{\pi}_1)) & 0 & \dots & 0 \\ 0 & n(\hat{\pi}_2(1 - \hat{\pi}_2)) & \dots & 0 \\ & & \vdots & \\ 0 & 0 & \dots & n(\hat{\pi}_k(1 - \hat{\pi}_k)) \end{bmatrix}.$$

As variâncias e covariâncias assintóticas dos coeficientes estimados são obtidos, invertendo a matriz informação de Fisher (3.33), que é dada da seguinte forma:

$$\Sigma(\hat{\boldsymbol{\beta}}) = I^{-1}(\hat{\boldsymbol{\beta}}) = \begin{bmatrix} \hat{v}_{11} & \hat{v}_{12} \\ \hat{v}_{12} & \hat{v}_{22} \end{bmatrix}. \quad (3.35)$$

O desvio padrão para os elementos de $\boldsymbol{\beta}$ é definido por

$$\widehat{DP}(\hat{\beta}_j) = \sqrt{\hat{v}_{jj}}, \quad (3.36)$$

$\hat{v}_{jj}$ corresponde ao j-ésimo elemento da diagonal da matriz (3.35).

A estatística do teste Wald para os parâmetros β_1 e β_2 da regressão logística é definida pela expressão (3.37).

$$Z_j = \frac{\hat{\beta}_j}{\widehat{DP}(\hat{\beta}_j)}. \quad (3.37)$$

A partir daí podemos definir o p-valor como $P(|Z| > Z_j)$, quando Z denota a variável aleatória da distribuição normal padrão.

3.5 Matriz de covariâncias para os estimadores dos parâmetros β_1, β_2, ω e Ω_w do modelo com erros

Para calcularmos a matriz observada de Fisher para o modelo com erros na variável explicativa, utilizamos o método de Louis (1982). Aqui, seguimos a notação de Tanner (1996) para expressarmos a matriz de Fisher observada.

$$J(\boldsymbol{\xi}) = -\frac{\partial^2 \log f(\boldsymbol{\xi}; \mathbf{y}, \mathbf{x})}{\partial \boldsymbol{\xi} \partial \boldsymbol{\xi}'} = -\int \frac{\partial^2 \log f(\boldsymbol{\xi}; \mathbf{y}, \mathbf{x}, \mathbf{w})}{\partial \boldsymbol{\xi} \partial \boldsymbol{\xi}'} f(\mathbf{w}; \mathbf{y}, \mathbf{x}, \boldsymbol{\xi}) d\mathbf{w} - \text{var} \left\{ \frac{\partial \log f(\boldsymbol{\xi}; \mathbf{y}, \mathbf{x}, \mathbf{w})}{\partial \boldsymbol{\xi}} \right\}. \quad (3.38)$$

A partir das expressões (3.16) e (3.11) antes do cálculo da esperança condicional, temos o logaritmo da função de verossimilhança completa dada por

$$\begin{aligned} \log f(\boldsymbol{\xi}; \mathbf{y}, \mathbf{x}, \mathbf{w}) &= \sum_{i=1}^k [y_i(\theta_i) - b(\theta_i) + h(y_i)] - \frac{1}{2\Omega_w} \sum_{i=1}^k (w_i - \omega)^2 \\ &\quad - \frac{k}{2} \log(2\pi) - \frac{k}{2} \log(\Omega_w), \end{aligned} \quad (3.39)$$

e $\theta_i = \beta_1 + w_i \beta_2$.

Ao calcularmos a primeira derivada obtemos,

$$\mathbf{U} = \frac{\partial}{\partial \boldsymbol{\xi}} \log f(\boldsymbol{\xi}; \mathbf{y}, \mathbf{x}, \mathbf{w}). \quad (3.40)$$

Em algumas situações, torna-se difícil calcular a integral (primeira parcela) da expressão (3.38). Se amostramos de $f(\boldsymbol{\xi}; \mathbf{y}, \mathbf{x}, \mathbf{w})$, esta integral pode ser aproximada pela soma

$$\frac{1}{m} \sum_{j=1}^m \frac{\partial^2}{\partial \boldsymbol{\xi} \partial \boldsymbol{\xi}'} \log f(\hat{\boldsymbol{\xi}}; \mathbf{y}, \mathbf{x}, \mathbf{w}_j^*), \quad (3.41)$$

$\mathbf{w}_j^* = \mathbf{w}_1^*, \mathbf{w}_2^*, \dots, \mathbf{w}_m^*$ é uma amostra aleatória da distribuição $f(\mathbf{w}; \mathbf{y}, \mathbf{x}, \hat{\boldsymbol{\xi}})$ da variável não observada com parâmetro $\hat{\boldsymbol{\xi}}$.

Similarmente, podemos aproximar a variância (segunda parcela) da expressão (3.38) via

$$\frac{1}{m} \sum_{j=1}^m \mathbf{U}_j \mathbf{U}_j' - \left(\frac{1}{m} \sum_{j=1}^m \mathbf{U}_j \right) \left(\frac{1}{m} \sum_{j=1}^m \mathbf{U}_j \right)'. \quad (3.42)$$

Assim, calculamos a matriz de informação observada, dada por

$$J(\hat{\boldsymbol{\xi}}) = - \left[\frac{1}{m} \sum_{j=1}^m \frac{\partial^2}{\partial \boldsymbol{\xi} \partial \boldsymbol{\xi}'} \log f(\boldsymbol{\xi}; \mathbf{y}, \mathbf{x}, \mathbf{w}_j^*) \right] - \left[\frac{1}{m} \sum_{j=1}^m \mathbf{U}_j \mathbf{U}_j' - \left(\frac{1}{m} \sum_{j=1}^m \mathbf{U}_j \right) \left(\frac{1}{m} \sum_{j=1}^m \mathbf{U}_j \right)' \right]. \quad (3.43)$$

A matriz de covariâncias de $\hat{\boldsymbol{\xi}}$ é matriz inversa de (3.43). Assim,

$$\widehat{\text{var}}(\hat{\boldsymbol{\xi}}) = J(\hat{\boldsymbol{\xi}})^{-1} = \begin{bmatrix} \hat{z}_{11} & \hat{z}_{12} & \hat{z}_{13} & \hat{z}_{14} \\ \hat{z}_{12} & \hat{z}_{22} & \hat{z}_{23} & \hat{z}_{24} \\ \hat{z}_{13} & \hat{z}_{23} & \hat{z}_{33} & \hat{z}_{34} \\ \hat{z}_{14} & \hat{z}_{24} & \hat{z}_{34} & \hat{z}_{44} \end{bmatrix}^{-1}. \quad (3.44)$$

3.5.1 Cálculo das derivadas

Partindo da expressão (3.40), as derivadas primeiras da expressão (3.39) em relação ao vetor de parâmetros $\boldsymbol{\xi} = (\beta_1, \beta_2, \omega, \Omega_w)'$ são dadas respectivamente por

$$\begin{aligned} \frac{\partial}{\partial \beta_1} \log f(\boldsymbol{\xi}; \mathbf{y}, \mathbf{x}, \mathbf{w}) &= \sum_{i=1}^k \left[y_i - \dot{b}(\theta_i) \right], \\ \frac{\partial}{\partial \beta_2} \log f(\boldsymbol{\xi}; \mathbf{y}, \mathbf{x}, \mathbf{w}) &= \sum_{i=1}^k \left[y_i w_i - \dot{b}(\theta_i) w_i \right] \\ \frac{\partial}{\partial \omega} \log f(\boldsymbol{\xi}; \mathbf{y}, \mathbf{x}, \mathbf{w}) &= \frac{1}{\Omega_w} \sum_{i=1}^k (w_i - \omega), \\ \frac{\partial}{\partial \Omega_w} \log f(\boldsymbol{\xi}; \mathbf{y}, \mathbf{x}, \mathbf{w}) &= -\frac{k}{2\Omega_w} + \frac{1}{2\Omega_w^2} \sum_{i=1}^k (w_i - \omega)^2. \end{aligned}$$

Utilizando o fato de que

$$\frac{\partial^2}{\partial \boldsymbol{\xi} \partial \boldsymbol{\xi}'} \log f(\boldsymbol{\xi}; \mathbf{y}, \mathbf{x}, \mathbf{w}) = \frac{\partial}{\partial \boldsymbol{\xi}} \mathbf{U}.$$

Dessa forma as derivadas segundas da expressão (3.39) em relação a $\beta_1, \beta_2, \omega, \Omega_w$ e seus produtos são dadas respectivamente por

$$\begin{aligned} \frac{\partial^2}{\partial^2 \beta_1} \log f(\boldsymbol{\xi}; \mathbf{y}, \mathbf{x}, \mathbf{w}) &= \hat{z}_{11} = \sum_{i=1}^k \left[-\ddot{b}(\theta_i) \right], \\ \frac{\partial^2}{\partial^2 \beta_2} \log f(\boldsymbol{\xi}; \mathbf{y}, \mathbf{x}, \mathbf{w}) &= \hat{z}_{22} = \sum_{i=1}^k \left[-\ddot{b}(\theta_i) w_i^2 \right], \\ \frac{\partial^2}{\partial^2 \omega} \log f(\boldsymbol{\xi}; \mathbf{y}, \mathbf{x}, \mathbf{w}) &= \hat{z}_{33} = -\frac{k}{\Omega_w}, \\ \frac{\partial^2}{\partial^2 \Omega_w} \log f(\boldsymbol{\xi}; \mathbf{y}, \mathbf{x}, \mathbf{w}) &= \hat{z}_{44} = \frac{k}{2\Omega_w^2} - \frac{1}{\Omega_w^3} \sum_{i=1}^k (w_i - \omega)^2, \\ \frac{\partial^2}{\partial \beta_1 \partial \beta_2} \log f(\boldsymbol{\xi}; \mathbf{y}, \mathbf{x}, \mathbf{w}) &= \hat{z}_{12} = \sum_{i=1}^k \left[-\ddot{b}(\theta_i) w_i \right], \\ \frac{\partial^2}{\partial \omega \partial \Omega_w} \log f(\boldsymbol{\xi}; \mathbf{y}, \mathbf{x}, \mathbf{w}) &= \hat{z}_{34} = -\frac{1}{\Omega_w^2} \sum_{i=1}^k (w_i - \omega), \\ \frac{\partial^2}{\partial \beta_1 \partial \omega} \log f(\boldsymbol{\xi}; \mathbf{y}, \mathbf{x}, \mathbf{w}) &= \hat{z}_{13} = \frac{\partial^2}{\partial \beta_1 \partial \Omega_w} \log f(\boldsymbol{\xi}; \mathbf{y}, \mathbf{x}, \mathbf{w}) = \hat{z}_{14} = \\ \frac{\partial^2}{\partial \beta_2 \partial \omega} \log f(\boldsymbol{\xi}; \mathbf{y}, \mathbf{x}, \mathbf{w}) &= \hat{z}_{23} = \frac{\partial^2}{\partial \beta_2 \partial \Omega_w} \log f(\boldsymbol{\xi}; \mathbf{y}, \mathbf{x}, \mathbf{w}) = \hat{z}_{24} = 0. \end{aligned}$$

$$\dot{b}(\theta_i) = n \frac{\exp(\theta_i)}{1 + \exp(\theta_i)} \quad \text{e} \quad \ddot{b}(\theta_i) = n \frac{\exp(\theta_i)}{(1 + \exp(\theta_i))^2},$$

são as derivadas primeira e segunda de $b(\theta_i)$

3.6 Variância para a estimativa da tendência, utilizando o método Delta

Observamos que a estimativa para o valor de tendência (T), $\hat{T}$, é uma função da estimativa de dois parâmetros, $g(\hat{\beta}_1, \hat{\beta}_2)$, digamos, um resultado padrão para a função da variância aproximada pode ser usada para obter a estimativa do desvio padrão. Dessa forma, como definido em Collett (2003), escrevemos

$$\text{var}(\hat{T}) \approx \begin{pmatrix} \frac{\partial g}{\partial \hat{\beta}_1} & \frac{\partial g}{\partial \hat{\beta}_2} \end{pmatrix} \begin{bmatrix} v_{11} & v_{12} \\ v_{12} & v_{22} \end{bmatrix} \begin{pmatrix} \frac{\partial g}{\partial \hat{\beta}_1} \\ \frac{\partial g}{\partial \hat{\beta}_2} \end{pmatrix}.$$

Daí, temos

$$\text{var}(\hat{T}) \approx \left(\frac{\partial g}{\partial \hat{\beta}_1} \right)^2 v_{11} + \left(\frac{\partial g}{\partial \hat{\beta}_1} \frac{\partial g}{\partial \hat{\beta}_2} \right) v_{12} + \left(\frac{\partial g}{\partial \hat{\beta}_1} \frac{\partial g}{\partial \hat{\beta}_2} \right) v_{12} + \left(\frac{\partial g}{\partial \hat{\beta}_2} \right)^2 v_{22},$$

$v_{11} = \text{var}(\hat{\beta}_1)$, $v_{12} = \text{cov}(\hat{\beta}_1, \hat{\beta}_2)$ e $v_{22} = \text{var}(\hat{\beta}_2)$. Este resultado demonstra que a variância de $\hat{T}$ é aproximadamente

$$\text{var}(\hat{T}) \approx \left(\frac{\partial g}{\partial \hat{\beta}_1} \right)^2 v_{11} + \left(\frac{\partial g}{\partial \hat{\beta}_2} \right)^2 v_{22} + 2 \left(\frac{\partial g}{\partial \hat{\beta}_1} \frac{\partial g}{\partial \hat{\beta}_2} \right) v_{12}.$$

Para as estimativas dos limites da tendência escolhemos γ referindo-se à probabilidade de aceitação, $0 < \gamma < 1$, cujo cálculo definido em Collett (2003) é dado por

$$\log \left(\frac{\gamma}{1-\gamma} \right) = \hat{\beta}_1 + \hat{\beta}_2 \mathbf{x}.$$

Assim, temos

Para o limite superior, tomando-se as estimativas dos parâmetros (β_1 e β_2), temos

$$\hat{T} = \frac{\log \left(\frac{\gamma}{1-\gamma} \right) - \hat{\beta}_1}{\hat{\beta}_2}. \quad (3.45)$$

Quando escolhemos $\gamma = 0,5$, a expressão (3.45) para $\hat{T}$ é dada por

$$\hat{T} = \frac{-\hat{\beta}_1}{\hat{\beta}_2}. \quad (3.46)$$

Assim, temos

$$\text{var}(\hat{T}) \approx \left(-\frac{1}{\hat{\beta}_2} \right)^2 v_{11} + \left(\frac{\hat{\beta}_1}{\hat{\beta}_2^2} \right)^2 v_{22} + 2 \left(-\frac{\hat{\beta}_1}{\hat{\beta}_2^3} \right) v_{12}. \quad (3.47)$$

Escrevendo $\hat{\rho} = \frac{\hat{\beta}_1}{\hat{\beta}_2}$, e substituindo os estimadores de v_{11}, v_{12} e v_{22} , dados por $\hat{v}_{11}, \hat{v}_{12}$ e $\hat{v}_{22}$ tomados nas matrizes (3.35) (modelo ingênuo) e (3.44) (modelo com erros), a expressão (3.47) é dada por

$$\widehat{\text{var}}(\hat{T}) \approx \frac{\hat{v}_{11} - 2\hat{\rho}\hat{v}_{12} + \hat{\rho}^2\hat{v}_{22}}{\hat{\beta}_2^2}. \quad (3.48)$$

O desvio padrão para $\hat{T}$ é dado por

$$\widehat{\text{DP}}(\hat{T}) \approx \sqrt{\frac{\hat{v}_{11} - 2\hat{\rho}\hat{v}_{12} + \hat{\rho}^2\hat{v}_{22}}{\hat{\beta}_2^2}}. \quad (3.49)$$

Dessa forma utilizamos a expressão (3.49) para o intervalo de confiança aproximado de T . Assim, temos que o intervalo com 95% de confiança para a tendência é obtido pela expressão

$$\hat{T} \pm Z_{(1-\frac{\alpha}{2})} \widehat{\text{DP}}(\hat{T}).$$

3.6.1 Variância do estimador da tendência utilizando o teorema de Fieller

O teorema de Fieller é um resultado geral para intervalos de confiança da razão de duas variáveis aleatórias normalmente distribuídas, (vide Collett, 2003, p.109).

Suponha que $\hat{\rho} = \frac{\hat{\beta}_1}{\hat{\beta}_2}$, onde β_1, β_2 são estimados por $\hat{\beta}_1, \hat{\beta}_2$ e estes estimadores são normalmente distribuídas com médias β_1, β_2 , variâncias v_{11}, v_{22} e covariância v_{12} . Considere a

função $\psi = \hat{\beta}_1 - \rho\hat{\beta}_2$. Então, $E(\psi) = \beta_1 - \rho\beta_2 = 0$, desde que $\hat{\beta}_1, \hat{\beta}_2$ sejam estimadores não viciados de β_1 e β_2 , respectivamente, e a variância de ψ seja dada por

$$\begin{aligned}\text{var}(\psi) &= \text{var}(\hat{\beta}_1) + \text{var}(\rho\hat{\beta}_2) - 2\text{cov}(\hat{\beta}_1, \rho\hat{\beta}_2) \\ &= v_{11} + \rho^2 v_{22} - 2\rho v_{12}.\end{aligned}\tag{3.50}$$

Assumimos que $\hat{\beta}_1, \hat{\beta}_2$ e ψ são normalmente distribuídos e

$$\frac{\hat{\beta}_1 - \rho\hat{\beta}_2}{\sqrt{\text{var}(\psi)}}$$

tem uma distribuição normal padrão. Consequentemente, se $z_{\alpha/2}$ é o quantil $\alpha/2$ da distribuição normal padrão, um intervalo de confiança de $100(1 - \alpha)\%$ para ρ é obtido a partir da desigualdade

$$|\hat{\beta}_1 - \rho\hat{\beta}_2| \leq z_{\alpha/2} \sqrt{\text{var}(\psi)}.$$

Tomando o quadrado em ambos os lados e igualando a zero, temos

$$\begin{aligned}\hat{\beta}_1^2 + \rho^2 \hat{\beta}_2^2 - 2(\hat{\beta}_1)(\rho\hat{\beta}_2) - z_{\alpha/2}^2 \text{var}(\psi) &= 0 \\ \hat{\beta}_1^2 + \rho^2 \hat{\beta}_2^2 - 2\rho\hat{\beta}_1\hat{\beta}_2 - z_{\alpha/2}^2 \text{var}(\psi) &= 0\end{aligned}\tag{3.51}$$

Substituindo $\text{var}(\psi)$ da equação (3.50) e rearranjando a equação quadrática em ρ , obtemos

$$\begin{aligned}\hat{\beta}_1^2 + \rho^2 \hat{\beta}_2^2 - 2\rho\hat{\beta}_1\hat{\beta}_2 - z_{\alpha/2}^2 (v_{11} + \rho^2 v_{22} - 2\rho v_{12}) &= 0 \\ \hat{\beta}_1^2 + \rho^2 \hat{\beta}_2^2 - 2\rho\hat{\beta}_1\hat{\beta}_2 - z_{\alpha/2}^2 v_{11} - z_{\alpha/2}^2 \rho^2 v_{22} + 2z_{\alpha/2}^2 \rho v_{12} &= 0 \\ (\hat{\beta}_2^2 - z_{\alpha/2}^2 v_{22})\rho^2 + (2v_{12}z_{\alpha/2}^2 - 2\hat{\beta}_1\hat{\beta}_2)\rho + \hat{\beta}_1^2 - v_{11}z_{\alpha/2}^2 &= 0.\end{aligned}\tag{3.52}$$

As duas raízes da equação quadrática (3.52) constituirá os limites de confiança para ρ . Este é o resultado de Fieller. Para utilizarmos este resultado para obter o intervalo de confiança para $T = \frac{-\hat{\beta}_1}{\hat{\beta}_2}$, escrevemos $-T$ para ρ na equação (3.52). Utilizando os estimadores, $\hat{v}_{11}, \hat{v}_{22}$ e $\hat{v}_{12}$, tomadas nas matrizes (3.35) (modelo ingênuo) e (3.44) (modelo com erros) para v_{11}, v_{22} e v_{12} , o resultado na equação quadrática (3.52) em T é

$$(\hat{\beta}_2^2 - z_{\alpha/2}^2 \hat{v}_{22})T^2 - (2\hat{v}_{12}z_{\alpha/2}^2 - 2\hat{\beta}_1\hat{\beta}_2)T + \hat{\beta}_1^2 - \hat{v}_{11}z_{\alpha/2}^2 = 0,$$

resolvendo a equação quadrática (3.52), temos as raízes

$$\frac{-\left(\hat{\rho} - g \frac{\hat{v}_{12}}{\hat{v}_{22}}\right) \pm \frac{z_{\alpha/2}}{\hat{\beta}_2} \sqrt{\hat{v}_{11} - 2\hat{\rho}\hat{v}_{12} + \hat{\rho}^2 \hat{v}_{22} - g \left(\hat{v}_{11} - \frac{\hat{v}_{12}^2}{\hat{v}_{22}}\right)}}{1 - g},\tag{3.53}$$

para os limites de confiança de $100(1 - \alpha)\%$ para o valor de T , onde $\hat{\rho} = \frac{\hat{\beta}_1}{\hat{\beta}_2}$ e $g = \frac{z_{\alpha/2}^2 \hat{v}_{22}}{\hat{\beta}_2^2}$. Quando g é zero na expressão (3.53), os limites para T são exatamente baseados no estimador do desvio padrão de $\hat{T}$ dada em (3.49).

3.7 Repetitividade

Como já fora mencionada no capítulo 1, a repetitividade representa a variabilidade associada ao sistema de medição passa/não passa. Aqui, propomos dois modos de expressarmos o seu estimador ($\hat{\sigma}_{rep}$) para o modelo ingênuo e com erros tomando-se o método Delta e o teorema de Fieller, a partir da variância de cada um destes. Assim, temos

1 Método Delta

Da expressão (3.48), obtemos

$$\hat{\sigma}_{rep} \approx \sqrt{\widehat{\text{var}}(\hat{T}) \times n} \quad (3.54)$$

2 Teorema de Fieller

Da expressão (3.53), obtemos

$$\hat{\sigma}_{rep} \approx \frac{\text{Amplitude do intervalo 3.53}}{2 \times Z_{(1-\frac{\alpha}{2})}} \times \sqrt{n} \quad (3.55)$$

Capítulo 4

Simulações

Este capítulo pretende avaliar o impacto da presença de erros na variável no modelo de regressão binária, com ênfase no problema de análise de sistemas de medição passa não passa. Para isto, vamos simular observações sob o modelo com erros na variável e comparar as estimativas obtidas nos modelos analítico (MSA (2002)), regressão logística sem erros e logístico com erros na variável. Para todas as situações de comparação as observações foram geradas conforme algoritmo:

1. Valores iniciais: $\beta_1, \beta_2, \omega, \Omega_w, \Omega_m$ e n ;
2. Geramos a variável não observada $\mathbf{w}$ (verdadeiro valor da característica da qualidade da peça) com distribuição normal com média ω e variância Ω_w ;
3. Conforme modelo logístico, obtemos

$$\theta(\mathbf{w}) = \log \frac{\pi(\mathbf{w})}{1 - \pi(\mathbf{w})} = \beta_1 + \beta_2 \mathbf{w}$$

e

$$\pi(\mathbf{w}) = \frac{\exp \theta(\mathbf{w})}{1 + \exp \theta(\mathbf{w})}$$

4. Tomamos $\mathbf{y} \sim B(n, \pi(\mathbf{w}))$
5. Tomamos $\mathbf{x} \sim N(\mathbf{w}, \Omega_m + \Omega_w)$

Inicialmente, vamos avaliar o impacto do erro de medição na probabilidade teórica π . Para isto, vamos gerar os dados conforme algoritmo e comparar a curva $(\mathbf{w}, \pi(\mathbf{w}))$ com os pontos perturbados pelo erro de medição $(\mathbf{x}, \pi(\mathbf{w}))$. Para análise, tomamos duas situações bem próximas da realidade de um sistema de medição por atributo (conforme aplicação):

Nas Figuras 4.1 e 4.2, observamos que os erros de medição provocam um ruído em torno da curva teórica $(\mathbf{w}, \pi(\mathbf{w}))$ e, quanto maior a variância do erro de medição maior a variabilidade em torno da curva teórica.

Tabela 4.1: Valores Inciciais

Situação	β_1	β_2	ω	Ω_w	Ω_m	n
1	-1,4	-110	0	0,0015	0,002 ² ; 0,003 ² ; 0,004 ² ; 0,006 ²	20
2	1,4	-2,9	0	0,05	0,1 ² ; 0,14 ²	20

Figura 4.1: Impacto do erro de medição - situação 1

4.1 Comparação entre os modelos

Aqui, vamos fazer uma comparação entre os modelos de regressão logística sem erros e logístico com erros na variável explicativa. Como base para nossa comparação, vamos gerar os dados conforme algoritmo e considerar os seguintes elementos de comparação:

i. Vício:

$$\sum_{i=1}^{nrep} \frac{(\beta_2 - \hat{\beta}_2(i))}{nrep},$$

ii. Erro quadrático médio:

Figura 4.2: Impacto do erro de medição - situação 2

$$\sum_{i=1}^{nrep} \frac{(\hat{\beta}_2(i) - \beta_2)^2}{nrep},$$

$nrep$ corresponde ao número de simulações.

- iii. Proporção de intervalos corretos de β_2 : proporção observada de intervalos de confiança que contém o verdadeiro valor de β_2

Mais uma vez, vamos considerar duas situações para análise:

Tabela 4.2: Parâmetros das simulações

Situação	β_1	β_2	ω	Ω_w	Ω_m	n	k
1	-1,4	-110	0	0,0015	0,002 ² ; 0,003 ² ; 0,004 ² ; 0,006 ²	20	12,20,30 e 50
2	1,4	-2,9	0	0,05	0,1 ² ; 0,14 ²	20	12,20,30 e 50

O critério de convergência do algoritmo EM (Dempster *et al.*, 1977), implementado em linguagem de programação Ox (Doornik, 2001), é dado por

$$\max \left\{ \left| \frac{\xi_j^{(t+1)} - \xi_j^{(t)}}{\xi_j^{(t)}} \right| ; j = 1, 2, 3, 4 \right\} < 10^{-3}. \quad (4.1)$$

Conforme discutimos anteriormente, vamos considerar dois modelos:

1. Modelo ingênuo: modelos de regressão logística usual, que ignora erros de medição na variável explicativa;
2. Modelo com erro na variável explicativa: modelo de regressão logística com erro na variável explicativa.

Após a simulação dos dados, os parâmetros do modelo de regressão com erros na variável e a respectiva matriz de informação de Fisher observada foram obtidos da seguinte forma:

i. Modelo com erros - Aplicação do Algoritmo EM:

- i1. Calculamos as expressões (3.28), (3.29) e (3.32).
- i2. Calculamos as expressões (3.18), (3.19) e (3.23).

A distribuição de $\mathbf{w}^*$ será simulada de uma normal, lembrando que a média e variância dessa distribuição havia sido aproximada pela média e variância de uma distribuição normal, conforme expressões (3.32) e (3.28).

ii. Cálculo da matriz de informação de Fisher observada

- ii1. Gerar uma amostra aleatória $\mathbf{w}_1^*, \mathbf{w}_2^*, \dots, \mathbf{w}_m^*$, no qual $\mathbf{w}_j^* = (\mathbf{w}_{j1}^*, \dots, \mathbf{w}_{jk}^*)$, $\mathbf{w}_{ji}^* \stackrel{iid}{\sim} N(w^{(i)}, V^{(i)}), i = 1, \dots, k$ e $\mathbf{w}_j^* \stackrel{iid}{\sim} N_k(w^{(t)} \mathbf{1}_k, V^{(t)} I_k), j = 1, \dots, m$
- ii2. Calcular as derivadas conforme seção 3.5.1.
- ii3. Calcular as expressões (3.42) e (3.41)
- ii4. A matriz de informação observada é estimada pela expressão (3.43).
- ii5. A matriz de covariâncias é dada pela expressão (3.44).

iii. Calcular o intervalo de confiança

$$\hat{\beta}_2 \pm Z_{(1-\frac{\alpha}{2})} \sqrt{\text{var}(\hat{\beta}_2)}$$

A seguir, apresentamos os resultados das $nrep = 5000$ repetições e das $m = 1000$ amostras simuladas da variável não observada $\mathbf{w}$. Além disso, apresentamos os erros quadráticos e os vícios das estimativas em relação ao parâmetro β_2 .

Observamos nas Tabela 4.3 e 4.4, que o modelo com erros possui resultados melhores quando aumentamos o valor da variância do erro de medição e o tamanho da amostra, uma vez que apresenta os menores erros quadráticos e os menores vícios para β_2 , também há uma maior proporção de intervalos corretos em relação ao ingênuo. Os resultados são semelhantes quando a variância é pequena, independentemente do tamanho da amostra.

Ressaltamos a rápida convergência do algoritmo EM, embora existam alguns casos em que foram necessárias cerca de 1000 iterações para encontrarmos a solução, de acordo com o critério de convergência 4.1.

Tabela 4.3: Resultados da simulação - Situação 1.

Variância	k	Vício (Ingênuo)	Vício(Com erro)	EQM de β_2		% Int. corretos de β_2	
		β_2	β_2	Ingênuo	Com erro	Ingênuo	Com erro
0,002 ²	12	3,850	3,840	460,314	430,066	94,18	95,36
	20	1,530	1,620	184,863	182,132	94,00	94,94
	30	0,380	0,590	113,089	112,503	94,04	95,04
	50	-0,070	0,210	65,637	65,299	93,26	94,76
0,003 ²	12	2,900	2,870	484,757	423,075	92,06	95,16
	20	0,180	0,540	197,868	187,661	92,62	95,20
	30	-0,780	-0,170	118,566	115,021	92,16	94,96
	50	-1,340	-0,610	71,588	69,157	91,14	94,48
0,004 ²	12	1,710	1,810	692,168	409,814	90,68	94,92
	20	-1,770	-0,750	213,242	190,050	89,40	94,54
	30	-2,340	-0,790	129,537	122,487	89,66	94,84
	50	-3,240	-1,740	82,072	73,422	86,90	93,64
0,006 ²	12	-1,980	-0,550	1376,340	414,028	83,40	93,48
	20	-5,840	-2,540	288,181	235,742	79,24	91,84
	30	-6,880	-3,140	197,046	146,501	75,60	92,02
	50	-7,820	-3,840	143,453	92,064	68,20	90,84

Tabela 4.4: Resultados da simulação - Situação 2.

Variância	k	Vício (Ingênuo)	Vício(Com erro)	EQM de β_2		% Int. corretos de β_2	
		β_2	β_2	Ingênuo	Com erro	Ingênuo	Com erro
0,01	12	0,555	-0,257	1,320	26,227	84,88	95,80
	20	0,587	-0,027	0,839	0,857	80,00	95,88
	30	0,590	-0,009	0,661	0,499	74,06	95,64
	50	0,592	0,007	0,535	0,290	62,72	95,22
0,02	12	1,127	-0,706	2,269	309,247	66,00	90,10
	20	1,176	-0,119	1,900	33,774	48,50	92,86
	30	1,184	-0,159	1,728	17,318	34,52	94,20
	50	1,175	-0,008	1,567	0,541	15,70	95,72

4.2 Simulações para tendência

Para análise da tendência, vamos comparar três modelos (analítico, ingênuo e com erros na variável explicativa). A seguir, resumimos os passos para este tipo de simulação.

i. Método Analítico:

i1. Calculamos π'_i a partir da expressão (2.4).

i2. Calculamos $\Phi^{-1}(\pi'_i)$ de π'_i .

i3. Ajustamos a regressão linear dada pela expressão (2.3) na qual obtemos μ' e b' .

Assim, estimamos a tendência μ' .

i. Método Ingênuo e com erros:

O verdadeiro valor para a tendência do modelo ingênuo e com erros para as situações 1 e 2 (Tabela 4.2) é dado por

$$T = \frac{-(-1, 4)}{-110} = -0,012727 \quad \text{e} \quad T = \frac{-(1, 4)}{-2,9} = 0,48275.$$

ii. Calculamos $\hat{T}$ a partir da expressão (3.46) através das estimativas de β calculadas nas expressões (3.6) e (3.23).

A seguir, temos um estudo de simulação no qual verificaremos o verdadeiro valor (-0,012727 e 0,48275) para a tendência.

Tabela 4.5: Resultados da simulação da tendência para os modelos Analítico, Ingênuo e Com erros - Situação 1

Variância	k	Analítico	Ingênuo	Com erros
0,002 ²	12	-0,01117	-0,01272	-0,01268
	20	-0,01099	-0,01274	-0,01271
	30	-0,01095	-0,01275	-0,01271
	50	-0,01092	-0,01274	-0,01270
0,003 ²	12	-0,01106	-0,01273	-0,01268
	20	-0,01102	-0,01279	-0,01271
	30	-0,01089	-0,01273	-0,01265
	50	-0,01089	-0,01277	-0,01270
0,004 ²	12	-0,01095	-0,01271	-0,01262
	20	-0,01101	-0,01284	-0,01270
	30	-0,01097	-0,01282	-0,01269
	50	-0,01095	-0,01283	-0,01270
0,006 ²	12	-0,01116	-0,01289	-0,01261
	20	-0,01109	-0,01292	-0,01263
	30	-0,01088	-0,01294	-0,01264
	50	-0,01089	-0,01301	-0,01271

Observamos na Tabela 4.5 que as estimativas da tendência para o modelo ingênuo e com erros estão muito próximas do valor verdadeiro (-0,012727). A Tabela 4.6 mostra que as estimativas da tendência para o modelo com erros estão mais próximas do valor verdadeiro (0,48275) que os outros dois (ingênuo e analítico).

Tabela 4.6: Resultados da simulação da tendência para os modelos Analítico, Ingênuo e Com erros
- Situação 2

Variância	k	Analítico	Ingênuo	Com erros
0, 1 ²	12	0,24026	0,59674	0,44332
	20	0,24246	0,60168	0,47525
	30	0,24297	0,60120	0,47731
	50	0,24208	0,60050	0,47945
0, 14 ²	12	0,21031	0,78765	0,37796
	20	0,20993	0,80288	0,46169
	30	0,20893	0,80373	0,45167
	50	0,20834	0,79883	0,47497

Capítulo 5

Análise do Sistema de Medição por Atributo

A partir da descrição do estudo do sistema de medição do tipo atributo, mencionada no capítulo 1, reproduzimos o experimento seguindo as diretrizes

- Seleccionamos 8 peças na linha de produção que foram enviadas ao laboratório de medidas dimensionais para determinarmos o valor numérico da característica da qualidade da peça, isto é, o valor do diâmetro, medido pelo relógio comparador (Figura 1.7) zerado por um anel padrão, com o auxílio de um súbito (Figura 1.6). Com isso, temos que o conjunto (relógio comparador + súbito + anel padrão) possui um desvio padrão de $0,006\text{ mm}$ e daí a utilização do modelos com erros que nos mostrará a diferença com o método analítico e do modelo ingênuo.
- Após medirmos as peças, estas retornaram à linha de produção onde foram medidas $n = 20$ vezes (cada peça) pelo sistema de medição passa não passa (o tampão liso). As peças estão distribuídas conforme Figura 1.11.
- Após a realização do experimento resumimos os dados conforme Tabela 1.1.
- Para facilitar a análise da tendência, subtraímos o LSE do resultado da medição da peça pelo sistema de medição referência, conforme mostra a Tabela 1.2.

Assim, obtemos a probabilidade do sistema de medição passa/não passa classificar a peça como aprovada ou reprovada, de modo que tenhamos a expressão.

$$y_i = \begin{cases} 1; & \text{Aprovada} \\ 0; & \text{Reprovada.} \end{cases} \quad (5.1)$$

Dessa forma a variável resposta para cada peça (y_i) corresponde à quantidade de aprovações e, consideramos também o intercepto e o valor da medição do diâmetro de cada peça (x_i) a única variável explicativa observada com erros de medição.

Neste capítulo, calculamos as estimativas pontuais e seus respectivos intervalos de confiança de 95% para as estimativas de β_1 , β_2 tendência e repetitividade em relação à probabilidade de aceitação para os dados da Tabela 1.2 do capítulo 1.

5.1 Modelo Ingênuo

Utilizando os dados da Tabela 1.2, calculamos a partir da expressão (3.34) as probabilidades das peças serem aprovadas para o modelo ingênuo. Dessa forma, temos

$$\hat{\pi} = \begin{pmatrix} 0,992785 \\ 0,979681 \\ 0,855514 \\ 0,67479 \\ 0,203065 \\ 0,081973 \\ 0,010846 \\ 0,001345 \end{pmatrix}$$

Para o cálculo da matriz de covariâncias de $\hat{\beta}$ do modelo ingênuo utilizamos a expressão (3.35), no qual temos o resultado a seguir

$$\Sigma(\hat{\beta}) = \left\{ \begin{pmatrix} 1 & 1 & \dots & 1 & 1 \\ -0,06 & -0,05 & \dots & 0,03 & 0,05 \end{pmatrix} \times \begin{pmatrix} 20 * 0,9927(1 - 0,9927) & 0 & \dots & 0 \\ 0 & 20 * 0,9796(1 - 0,9796) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \dots \\ 0 & 0 & \dots & 20 * 0,0013(1 - 0,0013) \end{pmatrix} \right. \\ \left. \times \begin{pmatrix} 1 & -0,06 \\ 1 & -0,05 \\ \vdots & \vdots \\ 1 & 0,03 \\ 1 & 0,05 \end{pmatrix} \right\}^{-1} = \begin{pmatrix} 0,13114 & 3,7240 \\ 3,7240 & 275,18 \end{pmatrix}. \quad (5.2)$$

As estimativas de β_1 e β_2 , são dadas pela expressão (3.6).

A Tabela 5.1 contém um resumo do modelo ingênuo.

Tabela 5.1: Estimativas e testes dos parâmetros - modelo ingênuo.

Parâmetro	Estimativa	$\widehat{DP}(\hat{\beta})$	Estatística Z	$Pr(> Z)$
β_1	-1,3672	0,3621	-3,7754	0,000159
β_2	-104,86	16,5885	-6,3212	2,60E-10

Dessa forma, rejeitamos H_0 ($\beta_j = 0$) a um nível de significância $\alpha = 0,05$ e concluimos então que os parâmetros β_1 e β_2 são significativos no modelo.

5.1.1 Cálculo da variância para a estimativa da tendência

A partir dos resultados da Tabela 5.1 e da matriz de variância 5.2, calculamos o intervalo de confiança para tendência através do método Delta, como a seguir.

Temos que $\hat{\rho} = \frac{-1,3672}{-104,86} = 0,013038$, $\hat{v}_{11} = 0,13114$, $\hat{v}_{22} = 275,18$, $\hat{v}_{12} = 3,7240$.

A partir da expressão (3.48), temos

$$\widehat{\text{var}}(\hat{T}) \approx \frac{0,13114 - 2(0,013038)(3,7240) + 0,013038^2(275,18)}{-104,86^2} \approx 0,000007349. \quad (5.3)$$

O intervalo pelo método Delta com 95% de confiança é dado por

$$\begin{aligned} -\hat{\rho} \pm Z_{(1-\frac{\alpha}{2})} \sqrt{\widehat{\text{var}}(\hat{T})} &\approx -0,013038 \pm 1,96(\sqrt{0,000007349}) \\ &\approx [-0,0183518 ; -0,0077248] \end{aligned} \quad (5.4)$$

Na sequência, calculamos o intervalo de confiança para tendência através do teorema de Fieller

Da expressão (3.53), temos

$$\frac{-\left(0,013038 - g \frac{3,7240}{275,18}\right) \pm \frac{z_{\alpha/2}}{-104,86} \sqrt{0,0730478}}{1 - g},$$

Onde: $g = \frac{z_{\alpha/2}^2 \hat{v}_{22}}{\hat{\beta}_2^2} = \frac{1,96^2 \times 275,18}{(-104,86)^2} = 0,096141$.

O intervalo pelo teorema de Fieller com 95% de confiança é dado por

$$\frac{-0,01173726 \pm (-0,0186915)(\sqrt{0,0730478})}{0,903858} \approx [-0,01857492 ; -0,00739653] \quad (5.5)$$

O sistema de medição passa não passa para este caso apresenta uma tendência de -0,013038 mm. Para uma tolerância de 0,4 mm concluímos que a tendência representa 3,25% da tolerância.

5.1.2 Cálculo da repetitividade

A partir das expressões 5.3 e 5.5, obtemos a repetitividade ($\hat{\sigma}_{rep}$) pelos dois métodos (Delta, teorema de Fieller). Dessa forma, temos

1 Método Delta

Da expressão 3.54, temos

$$\hat{\sigma}_{rep} \approx \sqrt{0,000007349 \times 20} = 0,01212 \quad (5.6)$$

2 Teorema de Fieller

Da expressão 3.55, temos

$$\hat{\sigma}_{rep} \approx \frac{[-0,00739653 - (-0,01857492)]}{2 \times 1,96} \times \sqrt{20} \approx 0,012752 \quad (5.7)$$

5.2 Modelo com erros na variável

Para o cálculo das estimativas de β_1 e β_2 utilizamos o procedimento do algoritmo EM (Passos E e M), mostrado no capítulo 3, subseção 3.3.2.

A Tabela 5.2 contém um resumo do modelo com erros.

Tabela 5.2: Estimativas e testes dos parâmetros - modelo com erros.

Parâmetro	Estimativa	$\widehat{DP}(\hat{\beta})$	Estatística Z	$Pr(> Z)$
β_1	-1,3391	0,5267	-2,5424	0,011009
β_2	-128,64	30,4373	-4,2263	2,37E-05

Dessa forma, rejeitamos H_0 ($\beta_j = 0$) a um nível de significância $\alpha = 0,05$ e concluímos então que os parâmetros β_1 e β_2 são significativos no modelo.

Para o cálculo da matriz de Informação Observada de Fisher, utilizamos o procedimento mostrado no capítulo 3, seção 3.5, no qual utilizamos a expressão (3.43) dada por,

$$\begin{pmatrix} 4,2543 & -0,028064 & -15,908 & 4,3925 \\ -0,028064 & 0,0012685 & 0,12823 & -3,0866 \\ -15,908 & 0,12823 & 6297,7 & 522,66 \\ 4,3925 & -3,0866 & 522,66 & 2414600 \end{pmatrix}. \quad (5.8)$$

Invertendo-se a matriz (5.8), temos a matriz de covariâncias utilizada no cálculo do intervalo de confiança para a tendência (T).

$$\begin{pmatrix} 0,27742 & 6,0968 & 0,000576 & 7,1643e-006 \\ 6,0968 & 926,43 & -0,0035598 & 0,0011740 \\ 0,000576 & -0,003559 & 0,00016032 & -4,0310e-008 \\ 7,1643e-006 & 0,0011740 & -4,0310e-008 & 4,1565e-007 \end{pmatrix}. \quad (5.9)$$

5.2.1 Cálculo da variância para a estimativa da tendência

A partir dos resultados da Tabela 5.1 e da matriz de variância 5.9, calculamos o intervalo de confiança para tendência através do método Delta, como a seguir.

Temos que $\hat{v}_{11} = 0,27742$, $\hat{\rho} = \frac{-1,3391}{-128,64} = 0,01041$, $\hat{v}_{22} = 926,43$, $\hat{v}_{12} = 6,0968$.

A partir da expressão (3.48), temos

$$\widehat{\text{var}}(\hat{T}) \approx \frac{0,27742 - 2(0,01041)(6,0968) + 0,01041^2(926,43)}{-128,64^2} \approx 0,00001516. \quad (5.10)$$

O intervalo pelo método Delta com 95% de confiança é dado por

$$\begin{aligned} -\hat{\rho} \pm Z_{(1-\frac{\alpha}{2})} \sqrt{\widehat{\text{var}}(\hat{T})} &\approx -0,01041 \pm 1,96(\sqrt{0,00001516}) \\ &\approx [-0,018041 ; -0,002778] \end{aligned} \quad (5.11)$$

Na sequência, calculamos o intervalo de confiança para tendência através do teorema de Fieller

Da expressão (3.53), temos

$$\frac{-\left(0,01041 - g \frac{6,0968}{926,43}\right) \pm \frac{z_{\alpha/2}}{-128,64} \sqrt{0,199845}}{1 - g},$$

Onde: $g = \frac{z_{\alpha/2}^2 \hat{v}_{22}}{\hat{\beta}_2^2} = \frac{1,96^2 \times 926,43}{(-128,64)^2} = 0,21505$

O intervalo pelo teorema de Fieller com 95% de confiança é dado por

$$\frac{-0,0089943 \pm (-0,015236)(\sqrt{0,199845})}{0,784941} \approx [-0,020135 ; -0,002781] \quad (5.12)$$

Portanto o sistema de medição passa não passa para este caso apresenta uma tendência de -0,01041 mm. Para uma tolerância de 0,4 mm concluímos que a tendência representa 2,6% da tolerância.

5.2.2 Cálculo da repetitividade

A partir das expressões 5.10 e 5.12, obtemos a repetitividade ($\hat{\sigma}_{rep}$) pelos dois métodos (Delta, teorema de Fieller). Dessa forma, temos

1 Método Delta

Da expressão 3.54, temos

$$\hat{\sigma}_{rep} \approx \sqrt{0,00001516 \times 20} = 0,017412 \quad (5.13)$$

2 Teorema de Fieller

Da expressão 3.55, temos

$$\hat{\sigma}_{rep} \approx \frac{[-0,002781 - (-0,020135)]}{2 \times 1,96} \times \sqrt{20} \approx 0,01979 \quad (5.14)$$

Para as probabilidades de aprovação temos as seguintes expressões

- modelo ingênuo:

$$\hat{\pi}_i = \frac{\exp(\hat{\beta}_1 + \hat{\beta}_2 x_i)}{1 + \exp(\hat{\beta}_1 + \hat{\beta}_2 x_i)}, \quad (5.15)$$

- modelo com erros:

$$\hat{\pi}_i = \frac{\exp(\hat{\beta}_1 + \hat{\beta}_2 \hat{w}_i)}{1 + \exp(\hat{\beta}_1 + \hat{\beta}_2 \hat{w}_i)}, \quad (5.16)$$

onde $\hat{w}_i$ é a estimativa de w_i obtida do algoritmo EM, após convergência.

Vale ressaltar que as estimativas de w_i são iterativamente calculadas pela expressão (3.32), mostrada no Capítulo 3.

Para visualizar os resultados, montamos a Tabela 5.3 com as probabilidades estimadas do método analítico e dos dois modelos logísticos (ingênuo e com erros).

Com os dados da Tabela 5.3, utilizamos o sistema estatístico R (R Development Core Team, 2004) para confeccionarmos a Figura 5.1, que mostra o comportamento dos três modelos (analítico, ingênuo e com erros), em função das probabilidades de aprovações observadas de cada peça.

De acordo com a Figura 5.1 observamos que, independente do modelo ao aumentarmos o valor padronizado do diâmetro de cada peça, menor a proporção de aprovações. Os modelos

Tabela 5.3: Probabilidades observadas e estimadas

Valor da medição	Prob. observada	Probabilidades estimadas		
		Analítico	Ingênuo	Com erros
-0,06	1,00	0,9927	0,9927	0,9983
-0,05	1,00	0,9757	0,9796	0,9938
-0,03	1,00	0,8485	0,8555	0,9255
-0,02	0,40	0,7120	0,6747	0,7744
0	0,25	0,3511	0,2030	0,2076
0,01	0,15	0,1968	0,0819	0,0675
0,03	0,00	0,0363	0,0108	0,0054
0,05	0,00	0,0031	0,0013	0,0004

Figura 5.1: Probabilidade das aprovações obtidas pelo método analítico e dos modelos ingênuo e com erros versus valor padronizado da medição do diâmetro.

ingênuo e com erros diferenciam-se em relação às maiores probabilidades (antes do LSE) e ficam próximos nas menores probabilidades (após o LSE). O método analítico apresenta a maior tendência (-0,0081) e distancia-se do modelo com erros.

Temos resumidos nas Tabelas 5.4 e 5.5 os intervalos (5.4, 5.5, 5.11 e 5.12) das estimativas para a tendência através dos dois modelos (ingênuo e com erros) para a região que delimita a faixa para o LSE.

A análise dos resultados das Tabelas 5.4 e 5.5, mostra que há diferença entre o modelo ingênuo em relação ao modelo com erros, embora a diferença seja muito pequena. Observa-

Tabela 5.4: Intervalo referente à tendência estimada - Método Delta, $\alpha = 0,05$

Modelo	Limite inferior	Limite superior
Ingênuo	-0,0183518	-0,0077248
Com erros	-0,018041	-0,002778

Tabela 5.5: Intervalo referente à tendência estimada - Teorema de Fieller, $\alpha = 0,05$

Modelo	Limite inferior	Limite superior
Ingênuo	-0,01857492	-0,00739653
Com erros	-0,020135	-0,002781

mos também que em ambos os métodos, a tendência é significativa (intervalos não contêm o zero), com repetitividades iguais a 0,01212 e 0,01275 (ingênuo) e 0,0174 e 0,0197 (com erros).

Ressaltamos que o modelo com erros possui uma repetitividade maior em função da inclusão dos erros de medição que tem como causa o uso do equipamento, rugosidade da peça, meio ambiente e operador.

5.3 Propostas Futuras

Como proposta de trabalho futuro, consideramos

- Modelos com sobre-dispersão;
- Modelos assimétricos.

Bibliografia

- Carroll, R., Spiegelman, C., Lan, K., Bailey, K. & Abbott, R. (1984). On errors-in-variables problem in the logistic regression model for binary regression models. *Biometrika*, **71**(1), 19–25.
- Collett, D. (2003). *Modelling binary data*. Chapman & Hall/CRC, Boca Raton, FL, second edition.
- Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B*, **39**(1), 1–38. With discussion.
- Demétrio, C. G. B. (2002). *Modelos Lineares Generalizados em Experimentação Agronômica*. ESALQ/USP, Piracicaba, first edition.
- Doornik, J. A. (2001). *Ox: An Object-Oriented Matrix Language*. Timberlake Consultants Press, London, fourth edition.
- Hosmer, D. W. & Lemeshow, S. (1989). *Applied Logistic Regression*. John Wiley & Sons, Inc., New York, second edition.
- ISOGUM (1998). *Guia para expressão da incerteza em medição*. Automotive Industry Action Group, Michigan, second edition.
- Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society B*, **44**, 226–233.
- McCullag, P. & Nelder, J. A. (1989). *Generalized Linear Models*. Chapman and Hall, London, second edition.
- MSA (2002). *MSA: Measurement Systems Analysis*. Automotive Industry Action Group, Michigan, third edition.
- R Development Core Team (2004). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Schafer, D. W. (1987). Covariate measurement error in generalized linear models. *Biometrika*, **74**(2), 385–391.

- Stefanski, L. A. & Carroll, R. J. (1985). Covariate measurement error in logistic regression. *Ann. Statist.*, **13**(1), 1335–1351.
- Tanner, M. A. (1996). *Tools for Statistical Inference*. Springer Series in Statistics, New York, third edition.