UNIVERSIDADE ESTADUAL DE CAMPINAS

Instituto de Matemática, Estatística e Computação Científica

ALEXANDRE GARCIA DIAS

Correção de viés de estimadores em regressão linear com covariáveis categóricas com erro de classificação

Campinas

2025

Alexandre Garcia Dias

Correção de viés de estimadores em regressão linear com covariáveis categóricas com erro de classificação

Dissertação apresentada ao Instituto de Matemática, Estatística e Computação Científica da Universidade Estadual de Campinas como parte dos requisitos exigidos para a obtenção do título de Mestre em Estatística.

Orientadora: Mariana Rodrigues Motta

Este trabalho corresponde à versão final da Dissertação defendida pelo aluno Alexandre Garcia Dias e orientada pela Profa. Dra. Mariana Rodrigues Motta.

Campinas

2025

Ficha catalográfica
Universidade Estadual de Campinas (UNICAMP)
Biblioteca do Instituto de Matemática, Estatística e Computação Científica
Ana Regina Machado - CRB 8/5467

D543c

Dias, Alexandre Garcia, 1996-
Correção de viés de estimadores em regressão linear com covariáveis categóricas com erro de classificação / Alexandre Garcia Dias. – Campinas, SP : [s.n.], 2025.

Orientador: Mariana Rodrigues Motta.
Dissertação (mestrado) – Universidade Estadual de Campinas (UNICAMP), Instituto de Matemática, Estatística e Computação Científica.

1. Regressão linear. 2. Covariáveis com erro. 3. Covariáveis discretas. I. Motta, Mariana Rodrigues, 1975-. II. Universidade Estadual de Campinas (UNICAMP). Instituto de Matemática, Estatística e Computação Científica. III. Título.

Informações complementares

Título em outro idioma: Bias correction of estimators in linear regression with categorical covariates with misclassification error
Palavras-chave em inglês:
Linear regression
Covariates with error
Discrete covariates
Área de concentração: Estatística
Titulação: Mestre em Estatística
Banca examinadora:
Mariana Rodrigues Motta [Orientador]
Benilton de Sá Carvalho
Júlia Maria Pavan Soler
Data de defesa: 27-02-2025
Programa de Pós-Graduação: Estatística

Objetivos de Desenvolvimento Sustentável (ODS)
ODS: 12. Consumo e produção responsáveis

Identificação e informações acadêmicas do(a) aluno(a)
- ORCID do autor: https://orcid.org/0009-0009-0652-997X
- Currículo Lattes do autor: https://lattes.cnpq.br/3622269109833002

Dissertação de Mestrado defendida em 27 de fevereiro de 2025 e aprovada pela banca examinadora composta pelos Profs. Drs.

Profa. Dra. MARIANA RODRIGUES MOTTA

Prof. Dr. BENILTON DE SÁ CARVALHO

Profa. Dra. JÚLIA MARIA PAVAN SOLER

A Ata da Defesa, assinada pelos membros da Comissão Examinadora, consta no SIGA/Sistema de Fluxo de Dissertação/Tese e na Secretaria de Pós-Graduação do Instituto de Matemática, Estatística e Computação Científica.

Este trabalho é dedicado à Dona Nancy.

Agradecimentos

Esse trabalho não seria possível sem o apoio, orientação e confiança das pessoas próximas de mim ao longo dessa jornada.

À minha orientadora, Profa. Mariana Rodrigues Motta, sou eternamente grato pela paciência e dedicação nesses anos. Suas palavras encorajadoras, correções minuciosas e comprometimento com a excelência acadêmica foram imprescindíveis para conseguir completar esse objetivo.

Ao Prof. Alexandre Aono, agradeço pelo conhecimento técnico e pelo tempo dedicado a esse trabalho. Sua disposição em compartilhar seu expertise enriqueceu significativamente este trabalho.

O presente trabalho foi realizado com o apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

"Estatística é a arte de nunca dizer algo como 'tenho certeza'."– Bertrand Russell

Resumo

O objetivo deste trabalho é propor um método de correção assintótico para os estimadores dos parâmetros oriundos da regressão com covariáveis sujeitas a erro de classificação. Foi desenvolvido uma correção baseada nos estimadores de mínimos quadrados da regressão com covariáveis com erro, probabilidade marginal das covariáveis verdadeiras e probabilidade condicional das covariáveis com erro dado as covariáveis verdadeiras. Desta maneira, podemos corrigir esses estimadores sem a necessidade de corrigir as covariáveis observadas nem observar as covariáveis verdadeiras. Estudos de simulação foram utilizados para quantificar o desempenho das correções propostas. Os mesmos identificaram que corrigir o intercepto é crucial para uma melhora significativa da estimação.

Palavras-chave: Regressão Linear, Covariáveis com Erro, Covariáveis Discretas.

Abstract

The objective of this work is to propose an asymptotic correction method for the estimators of parameters from regression models with covariates subject to classification errors. A correction was developed based on the least squares estimators from regression with erroneous covariates, the marginal probability of the true covariates, and the conditional probability of the erroneous covariates given the true covariates. In this way, we can correct these estimators without the need to correct the erroneous covariates or observe the true covariates. Simulation studies were used to quantify the performance of the proposed corrections. These studies identified that correcting the intercept is crucial for a significant improvement in estimation.

Keywords: Linear Regression, Covariates with Error, Discrete Covariates.

Lista de ilustrações

Lista de abreviaturas e siglas

UNICAMP	Universidade Estadual de Campinas
IMECC	Instituto de Matemática, Estatística e Computação Científica
SIMEX	Simulation Extrapolation
DNA	Deoxyribonucleic Acid
SNP	Single Nucleotide Polimorphism
EQP	Erro Quadrático Ponderado

Lista de símbolos

$\mathbb{I}$	Matriz identidade de ordem n
$\mathbb{1}$	Vetor de 1s de tamanho n
L_k	Quantidade de categorias da k-ésima covariável
$\mathbf{W}$	Matriz de Desenho correspondente às covariáveis com erro
$\mathbf{X}$	Matriz de Desenho correspondente às covariáveis verdadeiras
$\mathbf{Y}$	Vetor de Variáveis Resposta
$\boldsymbol{\beta}$	Vetor de parâmetros correspondentes a regressão de $\mathbf{Y}$ em $\mathbf{X}$
$\boldsymbol{\gamma}$	Vetor de parâmetros correspondentes a regressão de $\mathbf{Y}$ em $\mathbf{W}$
θ	Matriz de probabilidades condicionais de $\mathbf{W}$ dado $\mathbf{X}$
π	Matriz de probabilidades condicionais de $\mathbf{X}$ dado $\mathbf{W}$
I	Função indicadora

Sumário

1 Introdução

Modelos lineares, em sua maioria, consideram covariáveis fixas e não sujeitas à aleatoriedade [Searle, 1997]. Esse fato causa um problema de importância vital, pois, principalmente na área médica e biológica, os erros nas covariáveis são reportadas em apenas 44% das publicações em periódicos de alto impacto, sendo que ainda menos utilizam algum método de correção [Brakenhoff et al., 2018]. Embora várias técnicas tenham sido desenvolvidas nos últimos anos para endereçar este problema, veja por exemplo [Blackwell et al., 2017] e [Keogh et al., 2020], no caso de regressão linear, o método de mínimos quadrados usuais continua sendo o método mais utilizado, mesmo na presença de covariáveis com erro, o que pode acarretar em conclusões leve até gravemente errôneas. Essa imprecisão é causada pelo viés que surge nos estimadores de mínimos quadrados dos parâmetros associados aos efeitos fixos [Keogh and Bartlett, 2021] e também pela não-consistência dos estimadores [Buonaccorsi, 2010]. Para o caso de covariáveis contínuas, [Davies and Mutton, 1975] estudaram o efeito da aleatoriedade na magnitude máxima dos viéses dos estimadores dos efeitos fixos ou aleatórios. Correção de viés para covariáveis contínuas foi assunto de décadas, e vários métodos foram desenvolvidos, desde simples correções baseados em momentos, calibração usando *quasi-likelihood* até extrapolação por simulação (SIMEX), para mais detalhes veja [Buonaccorsi, 2010]. Porém, como citado por [Christopher and Kupper, 1995], estes métodos não são adequados para erros em covariáveis categóricas por causa da dependência nas suposições que não são válidas neste caso, por exemplo a utilização de uma estrutura aditiva de erro normal com média zero e independência do erro e das covariáveis. Visto que uma das utilidades dos estimadores de parâmetros é a predição, é importante, notar que a acurácia preditiva dos modelos é negativamente impactada [Klein and Rossin, 1999]. Em comparação aos estudos em covariáveis contínuas, o desenvolvimento de métodos de correção para covariáveis discretas se restringe basicamente aos modelos de regressão. Alguns métodos foram propostos, por exemplo, [Kuha, 1997] utiliza método de *data augmentation* que necessita de dados de validação enquanto que [Chen et al., 2009] desenvolvem métodos não paramétricos de estimação que não exigem informação adicional. Alternativamente, [Zucker and Spiegelman, 2008] propõem um método de correção baseado em score. [Buonaccorsi et al., 2005] propôs um método de correção do viés no caso de uma covariável binária. A correção é baseada na matriz de covariância da variável resposta e as covariáveis observadas, bem como, entre as covariáveis observadas e as verdadeiras.

Neste trabalho, o principal foco é estender a abordagem de [Buonaccorsi et al., 2005] para o caso de múltiplas covariáveis multinomiais. O método proposto, quando aplicado a variáveis binárias, retorna exatamente o estimador proposto

por Buonaccorsi. Outrossim, cada covariável pode assumir níveis distintos. A motivação dessa extensão vem do fato de que, na área de genética quantitativa, bilhões de leituras são descartadas por terem níveis de certeza considerados insuficientes. Este descarte eleva o custo da genotipagem a ponto de ser uma análise proibitiva. Portanto, a possibilidade de usar leituras consideradas corrompidas, permite genotipar um número menor de sequências, consequentemente, reduzindo os custos e podendo acarretar na desmonopolização do uso da genética quantitativa.

1.1 O modelo de regressão linear com uma variável categórica com L níveis

Seja y_i, $i = 1, \ldots, n$ uma realização da variável aleatória resposta Y_i e x_i uma realização da covariável X_i com L categorias. Considere que o modelo linear foi formulado segundo a parametrização de casela de referência, onde a categoria L foi determinada como classe de referência e defina $X_{il} = 1 \iff X_i = l$ onde $l = 1, 2, \ldots, L-1$. Considere o seguinte modelo linear

$$Y_i = \beta_0 + \sum_{l=1}^{L-1} \beta_l X_{il} + \epsilon_{Xi}, \tag{1.1}$$

tal que na forma matricial temos

$$\mathbf{Y} = \mathbb{1}\beta_0 + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{1.2}$$

onde

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}, \mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_n \end{bmatrix}, \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix},$$

$\mathbf{X}_i = (X_{i1}, X_{i2}, \ldots, X_{iL-1})$ e $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_{L-1})^T$. Assume-se que $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \mathbb{I}\sigma^2)$ são erros aleatórios onde $\mathbb{I}$ é a matriz identidade de dimensão $n \times n$ e $\mathbb{1}$ é um vetor de 1s de tamanho n O parâmetro β_0 representa o efeito da classe de referência L e β_l representa o incremento da classe l com respeito à classe de referência. A interpretação do modelo não muda se a classe de referência mudar. Se $L = 2$ temos o modelo estudado por [Buonaccorsi et al., 2005] para uma covariável binária.

1.2 Discussão sobre a teoria genética

Após o advento do sequenciamento genético em grande escala, o estudo de melhoramento genético de plantas e animais obteve um significativo avanço utilizando

uma abundância de dados de sequenciamento e modelos lineares. O fato que hoje em dia é possível sequenciar bilhões de loci permite descartar grandes quantidades de dados, que apresentam incerteza e/ou redundância, utilizando somente dados altamente confiáveis [Mrode and Pocrnic, 2023]. As consequências disso são que poucas empresas ou institutos de pesquisa podem arcar com os custos da genotipagem em alta escala; ademais, erros de genotipagem impactam significativamente os estudos genéticos, podendo reduzir a eficiência e levando a conclusões falsas em análises de parentesco [Ward et al., 2021]. Neste trabalho, o modelo linear ainda não acomoda parentesco entre indivíduos, sendo esta uma suposição que precisa ser relaxada em trabalhos futuros.

Sequenciamento genético é a base de toda a pesquisa de empresas agropecuárias da modernidade. Seu objetivo é classificar cada posição do DNA (locus), de modo que haja informação sobre como é composto esse locus. Geralmente, os loci estudados são os SNPs (do inglês, *Single Nucleotide Polimorphisms*) que são "uma forma abundante de variação genômica, distinguindo-se das variações raras por exigirem que o alelo menos abundante tenha uma frequência de 1% ou mais"[Brookes, 1999]. Cada SNP pode ter valores diferentes entre organismos, causando diferenças físicas (fenótipos) observáveis ou não. As categorias dos SNPs são definidas pela ploidia do organismo ou da sua localização no DNA. Por exemplo, em animais, a sua ploidia é quase sempre dois, enquanto em plantas, é mais complicado, com sua ploidia podendo variar [Leitch and Leitch, 2008]. Quando o gene do organismo é sequenciado e genotipado, os dados são compostos de valores inteiros, variando de zero até o valor da ploidia para cada SNP. Após a obtenção do DNA, o mesmo é amplificado para identificar se o valor lido estava correto. Se todas as replicações resultarem num mesmo valor, existe uma alta certeza de que este valor está correto. Quanto mais incongruência, maior a probabilidade de erro. Esses erros podem resultar na ordenação incorreta dos SNPs de acordo com seu valor genético e na subestimação das correlações dos marcadores e dos traços genéticos que são desejados [Hackett and Broadfoot, 2003, Göring and Terwilliger, 2000].

Problemas de classificação em covariáveis categóricas também são presentes em outros contextos dentro da área de genética quantitativa. Por exemplo, em *linkage analysis*, um campo de estudo para determinar quais loci são responsáveis pela expressão de fenótipos específicos, os métodos de dois loci são mais robustos aos erros de genotipagem, definidos por erros de classificação, enquanto análises com múltiplos loci podem excluir falsamente locais verdadeiros de genes de doenças [Göring and Terwilliger, 2000]. Além disso, a estrutura dos dados permite a imputação de painéis de SNPs de baixa densidade para alta densidade, mitigando parcialmente a perda de dados, onde a magnitude dos erros e aumento da banda de confiabilidade dependem especialmente do tamanho da população de referência [Dassonneville et al., 2011].

A fim de minimizar os erros de genotipagem, [Ward et al., 2021] indica que

os pesquisadores devem otimizar os métodos experimentais, usar controles e réplicas apropriadas e desenvolver abordagens estatísticas para a identificação de erros. Idealmente, após a aplicação destas técnicas, o resultado deveria ser somente a exclusão de dados que são ruidosos ou que não contêm informação pertinente. O desafio é diminuir a quantidade de dados excluídos sem distorcer a informação gerada por dados suscetíveis a erro. Neste trabalho, propomos um método de correção assintótica do viés dos estimadores dos parâmetros de uma regressão linear cujas covariáveis apresentam nível de incerteza longe do perfeito.

O restante deste trabalho está organizado da seguinte maneira. O modelo de regressão linear com uma variável categórica com mais de dois níveis com erro de classificação é apresentado na Seção 2.1. Em seguida, na Seção 2.2 apresentamos o modelo de regressão linear com mais de uma variável categórica com mais de dois níveis com erro de classificação. A Seção 2.3 apresenta a correção do intercepto que deve ser feita devido à singularidade da matriz de covariância de variáveis multinomiais. Nas Seções 2.4 e 2.5 o viés e a variância, respectivamente, dos estimadores corrigidos são calculados. O Capítulo 3 apresenta o estudo de simulação utilizado para demonstrar o desempenho do método. Por fim, uma discussão geral acerca do problema estudado é apresentada no Capítulo 4.

2 O modelo de regressão linear com variáveis categóricas apresentando erro de classificação

Nesta seção será definido o modelo linear onde há somente covariáveis categóricas, com erro de classificação em K covariáveis onde a k-ésima covariável apresenta L_k níveis, $k = 1, \ldots, K$. O propósito é apresentar um método de correção de viés dos estimadores da regressão linear considerando o erro de classificação.

2.1 O modelo de regressão linear com $K = 1$ variável categórica com duas ou mais categorias apresentando erro de classificação

Considere X uma variável aleatória discreta tomando valores no conjunto $\{1, \ldots, L\}$ onde $p_l = P(X = l)$, $l = 1, \ldots, L$ e $\sum_{l=1}^{L} p_l = 1$. Suponha que X esteja sujeita a um erro de classificação, onde a variável aleatória W representa X. Assuma que as categorias de W sejam as mesmas de X. Seguindo [Buonaccorsi et al., 2005], seja

$$\theta_{l|m} = P(W = l|X = m), \tag{2.1}$$

tal que

$$\begin{aligned} \pi_{m|l} &= P(X = m|W = l) \\ &= \theta_{l|m} \frac{P(X = m)}{\sum_{m'=1}^{L} \theta_{l|m'} P(X = m')}. \end{aligned} \tag{2.2}$$

Note que, condicionado em $X = m$, $m = 1, \ldots, L$, W tem distribuição Multinomial com probabilidades $\theta_{1|m}, \ldots, \theta_{L|m}$ para as classes 1 a L, respectivamente.

Sejam $X_1, \ldots, X_n$ vetores aleatórios independentes com a mesma distribuição de X e $W_1, \ldots, W_n$ as respectivas variáveis com erro de classificação. Além disso, suponha que ao invés de observar X_i, observemos W_i. Os vetores auxiliares, $\mathbf{X}_i$ e $\mathbf{W}_i$, $i = 1, \ldots, n$ são construídos de modo que suas componentes, X_{il} e W_{il}, binárias, e definidas através das seguintes relações: $X_{il} = 1 \iff X_i = l$ e $W_{il} = 1 \iff W_i = l$ onde $l = 1, 2, \ldots, L-1$. Neste caso, o modelo linear contendo as variáveis sem erro é dado por

$$Y_i = \beta_0 + \sum_{l=1}^{L-1} \beta_l X_{il} + \epsilon_{Xi}. \tag{2.3}$$

Na forma matricial temos

$$\mathbf{Y} = \mathbb{1}\beta_0 + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{2.4}$$

onde

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}, \mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_n \end{bmatrix}, \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix},$$

$\mathbf{X}_i = (X_{i1}, X_{i2}, \ldots, X_{iL-1})$ e $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_{L-1})^T$. Assume-se que $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \mathrm{I}\sigma^2)$ são erros aleatórios. O parâmetro β_0 representa o efeito da classe de referência L e β_l representa o incremento da classe l com respeito à classe de referência. A interpretação do modelo não muda se a classe de referência mudar. Se $K = 1$ e $L = 2$ temos o modelo estudado por [Buonaccorsi et al., 2005].

Por outro lado, como as covariáveis $\mathbf{X}_i$ não são observáveis, a correção deve ser feita utilizando o ajuste do modelo linear que contêm as covariáveis $\mathbf{W}_i$, ou seja, modelamos

$$Y_i = \gamma_0 + \sum_{l=1}^{L-1} \gamma_l W_{il} + \epsilon_{Wi}, \tag{2.5}$$

cuja forma matricial é dada por

$$\mathbf{Y} = \mathbb{1}\gamma_0 + \mathbf{W}\boldsymbol{\gamma} + \boldsymbol{\epsilon}_W, \tag{2.6}$$

onde

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}, \mathbf{W} = \begin{bmatrix} \mathbf{W}_1 \\ \vdots \\ \mathbf{W}_n \end{bmatrix}, \boldsymbol{\epsilon}_W = \begin{bmatrix} \epsilon_{W1} \\ \vdots \\ \epsilon_{Wn} \end{bmatrix},$$

onde $\mathbf{W}_i = (W_{i1}, W_{i2}, \ldots, W_{iL-1})$ e $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_{L-1})$. Assume-se que $\boldsymbol{\epsilon}_W \sim N(\mathbf{0}, \mathbb{I}\sigma_W^2)$ são erros aleatórios com variância σ_W^2. Uma vez que o modelo (2.5) é linear em $\boldsymbol{\gamma}$ e γ_0, utilizamos o estimador de mínimos quadrados $\widehat{\boldsymbol{\gamma}} = \left(\mathbf{W}^T\mathbf{W}\right)^{-1}\mathbf{W}^T\mathbf{Y}$ para estimar $\boldsymbol{\gamma}$.

É sabido que $\widehat{\boldsymbol{\gamma}}$ é um estimador não viésado de $\boldsymbol{\gamma}$, porém, o interesse recai em obter um estimador de $\boldsymbol{\beta}$ que seja assintoticamente não-viésado. A fim de construir esse estimador $\widehat{\boldsymbol{\beta}}$, utilizaremos as matrizes de variância e covariância definidas a seguir.

Por (2.3) tem-se que, para todo i,

$$\begin{aligned} \mathrm{Cov}(X_{il}, Y_i) &= \mathrm{Cov}(X_{il}, \beta_0 + \sum_{l'=1}^{L-1} \beta_{l'} X_{il'}) = \sum_{l'=1}^{L-1} \beta_{l'} \mathrm{Cov}(X_{il}, X_{il'}) \\ &= \mathbf{C}_l\boldsymbol{\beta}, \text{ para } l = 1, \ldots, L-1, \end{aligned}$$

onde $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_{L-1})^T$ e $\mathbf{C}_l = [\text{Cov}(X_{il}, X_{i1}), \ldots, \text{Cov}(X_{il}, X_{iL-1})]$ é um vetor de dimensão $1 \times (L-1)$. Definindo

$$\boldsymbol{\Sigma}_X = \begin{bmatrix} \mathbf{C}_1 \\ \vdots \\ \mathbf{C}_{L-1} \end{bmatrix}$$

uma matriz de covariância de dimensão $(L-1) \times (L-1)$, e

$$\boldsymbol{\Sigma}_{XY} = \begin{bmatrix} \text{Cov}(Y_i, X_{i1}) \\ \vdots \\ \text{Cov}(Y_i, X_{iL-1}) \end{bmatrix}$$

um vetor de dimensão $(L-1) \times 1$. Note que

$$\boldsymbol{\Sigma}_{XY} = \boldsymbol{\Sigma}_X \boldsymbol{\beta}. \tag{2.7}$$

Analogamente, por (2.5), definindo $\mathbf{D}_l = [\text{Cov}(W_{il}, W_{i1}), \ldots, \text{Cov}(W_{il}, W_{iL-1})]$,

$$\boldsymbol{\Sigma}_W = \begin{bmatrix} \mathbf{D}_1 \\ \vdots \\ \mathbf{D}_{L-1} \end{bmatrix},$$

e

$$\boldsymbol{\Sigma}_{WY} = \begin{bmatrix} \text{Cov}(Y_i, W_{i1}) \\ \vdots \\ \text{Cov}(Y_i, W_{iL-1}) \end{bmatrix},$$

obtemos

$$\boldsymbol{\Sigma}_{WY} = \boldsymbol{\Sigma}_W \boldsymbol{\gamma}. \tag{2.8}$$

Finalmente, considerando Y_i em (2.3), obtemos

$$\begin{aligned} \text{Cov}(W_{il}, Y_i) &= \text{Cov}(W_{il}, \beta_0 + \sum_{l'=1}^{L-1} \beta_{l'} X_{il'}) = \sum_{l'=1}^{L-1} \beta_{l'} \text{Cov}(W_{il}, X_{il'}) \\ &= \begin{bmatrix} \text{Cov}(W_{il}, X_{i1}) & \ldots & \text{Cov}(W_{il}, X_{iL-1}) \end{bmatrix} \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_{L-1} \end{bmatrix}. \end{aligned}$$

e portanto

$$\begin{bmatrix} \text{Cov}(W_{i1}, Y_i) \\ \vdots \\ \text{Cov}(W_{il}, Y_i) \\ \vdots \\ \text{Cov}(W_{iL-1}, Y_i) \end{bmatrix} = \begin{bmatrix} \text{Cov}(W_{i1}, X_{i1}) & \dots & \text{Cov}(W_{iL-1}, X_{i1}) \\ \vdots & \ddots & \vdots \\ \text{Cov}(W_{i1}, X_{iL-1}) & \dots & \text{Cov}(W_{iL-1}, X_{iL-1}) \end{bmatrix} \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_{L-1} \end{bmatrix}.$$

Definindo

$$\boldsymbol{\Sigma}_{WX} = \begin{bmatrix} \text{Cov}(W_{i1}, X_{i1}) & \dots & \text{Cov}(W_{iL-1}, X_{i1}) \\ \vdots & \ddots & \vdots \\ \text{Cov}(W_{i1}, X_{iL-1}) & \dots & \text{Cov}(W_{iL-1}, X_{iL-1}) \end{bmatrix}$$

temos que

$$\boldsymbol{\Sigma}_{WY} = \boldsymbol{\Sigma}_{WX}\boldsymbol{\beta}. \tag{2.9}$$

Considerando o modelo linear em (2.5), e definindo $\widehat{\boldsymbol{\gamma}} = (\mathbf{W}^T\mathbf{W})^{-1}\mathbf{W}^T\mathbf{Y}$ como o estimador de mínimos quadrados de $\boldsymbol{\gamma}$, temos que

$$\widehat{\boldsymbol{\gamma}} \xrightarrow{P} \boldsymbol{\gamma} \tag{2.10}$$

ou seja $\widehat{\boldsymbol{\gamma}}$ converge em probabilidade para $\boldsymbol{\gamma}$. Usando a equação (2.8),

$$\boldsymbol{\gamma} = \boldsymbol{\Sigma}_W^{-1}\boldsymbol{\Sigma}_{WY}. \tag{2.11}$$

e substituindo (2.9) em (2.11), obtemos

$$\boldsymbol{\gamma} = \boldsymbol{\Sigma}_W^{-1}\boldsymbol{\Sigma}_{WX}\boldsymbol{\beta}.$$

Assim, o estimador corrigido $\widehat{\boldsymbol{\beta}}_C$ de $\boldsymbol{\beta}$ é dado pela correção de $\widehat{\boldsymbol{\gamma}}$ através da transformação

$$\widehat{\boldsymbol{\beta}}_C = \left(\boldsymbol{\Sigma}_W^{-1}\boldsymbol{\Sigma}_{WX}\right)^{-1}\widehat{\boldsymbol{\gamma}}. \tag{2.12}$$

2.1.1 Cálculo das matrizes de variância e covariância

A fim de obter a correção definida em (2.12) precisamos determinar as expressões analíticas das matrizes $\boldsymbol{\Sigma}_W$ e $\boldsymbol{\Sigma}_{WX}$. Lembrando que $\mathbf{W}_i = (W_{i1}, \dots, W_{i\,L-1})$, temos que

$$\begin{aligned} \text{Var}(W_{il}) &= E[W_{il}^2] - E^2[W_{il}] \\ &= E_X[E_{W|X}[(W_{il}|X_i]] - \left(E_X[E_{W|X}[(W_{il}|X_i]]\right)^2 \\ &= \sum_{x=1}^{L} P(X_i = x)E[W_{il}|X_i = x] - \left(\sum_{x=1}^{L} P(X_i = x)E[W_{il}|X_i = x]\right)^2. \end{aligned}$$

Considerando (2.1), temos que

$$\theta_{l|x} = P(W_i = l|X_i = x) = P(W_{il} = 1|X_i = x) = E[W_{il}|X_i = x],$$

de modo que

$$\text{Var}(W_{il}) = \sum_{x=1}^{L} P(X_i = x)\theta_{l|x} - \left(\sum_{x=1}^{L} P(X_i = x)\theta_{l|x}\right)^2. \tag{2.13}$$

Além disso, para $l \neq m$,

$$\begin{aligned}\text{Cov}(W_{il}, W_{im}) &= E[W_{il}W_{im}] - E[W_{il}]E[W_{im}] \\ &= E[W_{il}W_{im}] - \sum_{x=1}^{L} P(X_i = x)\theta_{l|x} \sum_{x=1}^{L} P(X_i = x)\theta_{m|x} \\ &= -\sum_{x=1}^{L} P(X_i = x)\theta_{l|x} \sum_{x=1}^{L} P(X_i = x)\theta_{m|x}, \end{aligned} \tag{2.14}$$

onde $E[W_{il}W_{im}] = 0$, uma vez que $W_{il}W_{im} = 0$ quando $l \neq m$. Lembrando que as Equações 2.13 e 2.14, são utilizadas para construir a matriz Σ_W. Observe que

$$\begin{aligned}\text{Cov}(W_{il'}, X_{il}) &= E[W_{il'}X_{il}] - E[W_{il'}]E[X_{il}] \\ &= E[W_{il'}X_{il}] - \left(\sum_{x=1}^{L} P(X_i = x)\theta_{l'|x}\right) \end{aligned} \tag{2.15}$$

$$\begin{aligned}E_X[E_{W|X}[W_{il'}, X_{il}|X_i]] &= \sum_{x=1}^{L} P(X_i = x)E[W_{il'}, X_{il}|X_i] \\ &= \sum_{x=1}^{L} P(X_i = x)\text{I}_{(x=l)}E[W_{il'}|X_i] \\ &= P(X_i = l')\theta_{l'|l}. \end{aligned} \tag{2.16}$$

Substituindo a equação (2.16) na equação (2.15) temos que

$$\text{Cov}(W_{il'}, X_{il}) = \left(\theta_{l'|l} - \sum_{x=1}^{L} P(X_i = x)\theta_{l'|x}\right) P(X = l),$$

que define uma coordenada da matriz Σ_{WX}.

2.2 O modelo de regressão linear com mais de uma variável categórica com duas ou mais categorias apresentando erro de classificação

Seja agora y_i, $i = 1, \ldots, n$, uma realização da variável aleatória resposta Y_i e X_{ik}, $k = 1, \ldots, K$, covariáveis categóricas onde a k-ésima covariável tem L_k classes.

Usando o mesmo critério da Seção 2.1, seja $X_{ikl} = 1$, se e somente se $X_{ik} = l$. Logo, o modelo tem forma matricial

$$\mathbf{Y} = \mathbb{1}\beta_0 + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

onde a matriz de desenho é dada por $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)^T$ com
$\mathbf{X}_i = (X_{i11}, \ldots, X_{i1L_1-1}, \ldots, X_{iK1}, \ldots, X_{iKL_K-1})$, β_0 é o intercepto e
$\boldsymbol{\beta} = (\beta_{11}, \beta_{12}, \ldots, \beta_{K1}, \ldots, \beta_{KL-1})$ é o vetor de parâmetros associados à $\mathbf{X}_i$ com dimensão $\sum_{k=1}^{K}(L_k - 1)$. Além disso, temos que $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)$ é um vetor de erros aleatórios, tal que $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \mathbb{I}\sigma^2)$, onde $\mathbb{I}$ representa a matriz identidade de ordem $n \times n$. Considerando agora que as covariáveis são medidas com erro, e $\mathbf{X}$ é não observável, o modelo linear que pode ser ajustado é dado por

$$\mathbf{Y} \quad = \quad \mathbb{1}\gamma_0 + \mathbf{W}\boldsymbol{\gamma} + \boldsymbol{\epsilon}_W, \tag{2.17}$$

onde

$$\boldsymbol{\gamma} = (\gamma_{11}, \gamma_{12}, \ldots, \gamma_{1L_1-1}, \gamma_{21}, \ldots, \gamma_{KL_K-1})$$

é um vetor de dimensão $\sum_{k=1}^{K}(L_k - 1)$ e $\boldsymbol{\epsilon}_W = (\epsilon_{W1}, \epsilon_{W2}, \ldots, \epsilon_{Wn})$, é um vetor de erros aleatórios com dimensão n, tal que $\boldsymbol{\epsilon}_W \sim N(\mathbf{0}, I\sigma_W^2)$. Além disso, a matriz de desenho $\mathbf{W} = (\mathbf{W_1}, \ldots, \mathbf{W_n})^T$ tem componentes
$\mathbf{W}_i = (W_{i11}, W_{i12}, \ldots, W_{i1L_1-1}, W_{i21}, \ldots, W_{iKL_K-1})$. A fim de construir um estimador $\widehat{\boldsymbol{\beta}}$ assintoticamente não-viesado, precisamos calcular as covariâncias a seguir, tal que, para Y_i definida em (2.17)

$$\begin{aligned}
\mathrm{Cov}(W_{ikl}, Y_i) &= \mathrm{Cov}(\gamma_0 + \sum_{k'=1}^{K}\sum_{l'=1}^{L_{k'}-1} W_{ik'l'}\gamma_{k'l'} + \epsilon_{Wi}, W_{ikl}) \\
&= \sum_{k'=1}^{K}\sum_{l'=1}^{L_{k'}-1} \mathrm{Cov}(W_{ik'l'}, W_{ikl})\gamma_{k'l'} \\
&= \sum_{l'=1}^{L_k-1} \mathrm{Cov}(W_{ikl'}, W_{ikl})\gamma_{kl'} \qquad (2.18)
\end{aligned}$$

pois $\mathrm{Cov}(W_{ikl}, W_{ik'l'}) = 0$ se $k \neq k'$. Portanto

$$\mathrm{Cov}(W_{ikl}, Y_i) \quad = \quad \Big[\mathrm{Cov}(W_{ikl}, W_{ik1}) \quad \ldots \quad \mathrm{Cov}(W_{ikl}, W_{ikL_k-1})\Big]\boldsymbol{\gamma}_k,$$

onde

$$\boldsymbol{\gamma}_k \quad = \quad \begin{bmatrix} \gamma_{k1} \\ \gamma_{k2} \\ \vdots \\ \gamma_{kL-1} \end{bmatrix}$$

e

$$\boldsymbol{\Sigma}_{W_k} = \begin{bmatrix} \mathrm{Cov}(W_{ik1}, W_{ik1}) & \dots & \mathrm{Cov}(W_{ik1}, W_{ikL-1}) \\ \mathrm{Cov}(W_{ik2}, W_{ik1}) & \dots & \mathrm{Cov}(W_{ik2}, W_{ikL-1}) \\ \vdots & \ddots & \vdots \\ \mathrm{Cov}(W_{ikL-1}, W_{ik1}) & \dots & \mathrm{Cov}(W_{ikL-1}, W_{ikL-1}) \end{bmatrix}. \tag{2.19}$$

Seja

$$\boldsymbol{\Sigma}_{WY} = \begin{bmatrix} \mathrm{Cov}(Y_i, W_{i11}) \\ \vdots \\ \mathrm{Cov}(Y_i, W_{i1L_1-1}) \\ \vdots \\ \mathrm{Cov}(Y_i, W_{iKL_K1}) \\ \vdots \\ \mathrm{Cov}(Y_i, W_{iKL_k-1}) \end{bmatrix},$$

$$\boldsymbol{\Sigma}_{W} = \begin{bmatrix} \boldsymbol{\Sigma}_{W_1} & \underset{\sim}{\mathbf{0}} & \dots & \underset{\sim}{\mathbf{0}} \\ \underset{\sim}{\mathbf{0}} & \boldsymbol{\Sigma}_{W_2} & \dots & \underset{\sim}{\mathbf{0}} \\ \vdots & & \ddots & \vdots \\ \underset{\sim}{\mathbf{0}} & \underset{\sim}{\mathbf{0}} & \dots & \boldsymbol{\Sigma}_{W_K} \end{bmatrix},$$

$$\boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_{11} \\ \boldsymbol{\beta}_{12} \\ \vdots \\ \boldsymbol{\beta}_{K1} \\ \vdots \\ \boldsymbol{\beta}_{KL_K-1} \end{bmatrix} \text{ e}$$

$$\boldsymbol{\gamma} = \begin{bmatrix} \boldsymbol{\gamma}_1 \\ \boldsymbol{\gamma}_2 \\ \vdots \\ \boldsymbol{\gamma}_k \end{bmatrix}.$$

Da mesma forma que na Equação (2.7) , temos que

$$\boldsymbol{\Sigma}_{WY} = \boldsymbol{\Sigma}_W \boldsymbol{\gamma}. \tag{2.20}$$

Definindo $\boldsymbol{\Sigma}_{W_k X}$ como

$$\boldsymbol{\Sigma}_{W_K X} = \begin{bmatrix} \text{Cov}(X_{ik1}, W_{ik1}) & \text{Cov}(X_{ik2}, W_{ik1}) & \dots & \text{Cov}(X_{ikL_k-1}, W_{ik1}) \\ \text{Cov}(X_{ik1}, W_{ik2}) & \text{Cov}(X_{ik2}, W_{ik2}) & \dots & \text{Cov}(X_{ikL_k-1}, W_{ik2}) \\ \vdots & & \ddots & \vdots \\ \text{Cov}(X_{ik1}, W_{ikL_k-1}) & \text{Cov}(X_{ik2}, W_{ikL_k-1}) & \dots & \text{Cov}(X_{ikL_k-1}, W_{ikL_k-1}) \end{bmatrix},$$

temos,

$$\boldsymbol{\Sigma}_{WX} = \begin{bmatrix} \boldsymbol{\Sigma}_{W_1 X} & \underset{\sim}{\mathbf{0}} & \dots & \underset{\sim}{\mathbf{0}} \\ \underset{\sim}{\mathbf{0}} & \boldsymbol{\Sigma}_{W_2 X} & \dots & \underset{\sim}{\mathbf{0}} \\ \vdots & & \ddots & \vdots \\ \underset{\sim}{\mathbf{0}} & \underset{\sim}{\mathbf{0}} & \dots & \boldsymbol{\Sigma}_{W_K X} \end{bmatrix}$$

e, da mesma forma que (2.9),

$$\boldsymbol{\Sigma}_{WY} = \boldsymbol{\Sigma}_{WX}\boldsymbol{\beta}. \tag{2.21}$$

Substituindo (2.21) em (2.20) obtemos

$$\boldsymbol{\Sigma}_{WY} = \boldsymbol{\Sigma}_{WX}\boldsymbol{\beta} = \boldsymbol{\Sigma}_{W}\boldsymbol{\gamma}, \tag{2.22}$$

tal que $\widehat{\boldsymbol{\beta}}_C = (\boldsymbol{\Sigma}_W^{-1}\boldsymbol{\Sigma}_{WX})^{-1}\widehat{\boldsymbol{\gamma}}$.

As componentes de $\boldsymbol{\Sigma}_{WX}$ e de $\boldsymbol{\Sigma}_W$ são calculadas da mesma forma que em (2.13), (2.14) e (2.17).

2.3 Correção para o intercepto β_0

Note que a regressão descrita em (2.17) possui $\sum_{k=1}^{K} k(L_k - 1) + 1$ parâmetros a serem estimados. Porém, devido à singularidade da matriz de covariância de variáveis multinomiais definida em (2.19), o método descrito na Seção 2.2 corrige somente o vetor de estimadores $\widehat{\gamma}$, dando origem ao vetor $\widehat{\boldsymbol{\beta}}_C$ definido em (2.12) e (2.22), de dimensão $\left(\sum_{k=1}^{K} k(L_k - 1)\right) \times 1$. Logo, o intercepto $\widehat{\gamma}_0$ ainda apresenta um viés.

O objetivo é corrigir o viés do estimador $\widehat{\gamma}_0$, obtido pelo método de mínimos quadrados através da regressão de $\mathbf{Y}$ em $\mathbf{W}$. Dada a variável aleatória $\mathbf{X}$, e

$$\mathbf{Y} = \beta_0 + \mathbf{X}\boldsymbol{\beta} + \epsilon,$$

tal que

$$E[\mathbf{Y}|\mathbf{X}] = \beta_0 + \mathbf{X}\boldsymbol{\beta}$$

e

$$E[\mathbf{Y}|\mathbf{W}] = \beta_0 + \boldsymbol{\beta} E[\mathbf{X}|\mathbf{W}] = \beta_0 + \boldsymbol{\beta}\boldsymbol{\pi},$$

onde

$$\boldsymbol{\pi} = \begin{bmatrix} \pi_{1|W_1} & \cdots & \pi_{L_k|W_1} \\ \pi_{1|W_2} & \cdots & \pi_{L_k|W_2} \\ \vdots & \ddots & \vdots \\ \pi_{1|W_n} & \cdots & \pi_{L_k|W_n} \end{bmatrix},$$

propomos o estimador corrigido $\widehat{\beta}_{0C}$ para β_0 dado por

$$\widehat{\beta}_{0C} = \frac{\sum_{i=1}^{n} \left(Y_i - \boldsymbol{\pi}_{(i)} \widehat{\boldsymbol{\beta}}_C \right)}{n}, \tag{2.23}$$

onde $\boldsymbol{\pi}_{(i)}$ é a i-ésima linha da matriz $\boldsymbol{\pi}$ e Y_i é a i-ésima observação do vetor $\mathbf{Y}$.

2.4 Cálculo de Viés dos Estimadores

Seja

$$\boldsymbol{\beta}^* = (\beta_0, \boldsymbol{\beta})^T,$$

com dimensão $M = \sum_{k=1}^{K} (L_k - 1) + 1$, o vetor de parâmetros da regressão de $\mathbf{Y}$ em $\mathbf{X}^* = (\mathbf{1}, \mathbf{X})$ dado por (2.3) e seja $\widehat{\boldsymbol{\beta}}^*_C = (\widehat{\gamma}_0, \widehat{\boldsymbol{\beta}}_C)^T$ onde $\widehat{\boldsymbol{\beta}}_C$ é vetor corrigido da forma descrita na Seção 2.2. Adicionalmente, seja $\boldsymbol{\gamma}* = (\gamma_0, \boldsymbol{\gamma})$ o vetor de parâmetros da regressão de $\mathbf{Y}$ em $\mathbf{W}^* = (\mathbf{1}, \mathbf{W})$ dado por (2.5). Considere

$$\widehat{\boldsymbol{\gamma}}^* = (\widehat{\gamma}_0, \widehat{\boldsymbol{\gamma}}) = \left(\mathbf{W}^{*T} \mathbf{W}^* \right)^{-1} \mathbf{W}^{*T} \mathbf{Y}$$

o vetor de estimadores de $\boldsymbol{\gamma}^*$.

Para encontrar o viés condicional de $\widehat{\boldsymbol{\beta}}^*_C$ em $\mathbf{W}$ calculamos

$$\mathrm{E}_{\mathbf{Y}|\mathbf{W}} \left[\widehat{\boldsymbol{\beta}}^*_C | \mathbf{W} \right] = \mathrm{E}_{\mathbf{Y}|\mathbf{W}} \left[\mathbf{Z} \widehat{\boldsymbol{\gamma}}^* | \mathbf{W} \right],$$

onde $\mathbf{Z} = \begin{bmatrix} 1 & \mathbf{0}^T \\ \mathbf{0} & \left(\Sigma_W^{-1} \Sigma_{WX} \right)^{-1} \end{bmatrix}$ é uma matriz de dimensão $(M \times M)$. Assim,

$$\begin{aligned} \mathrm{E}_{\mathbf{Y}|\mathbf{W}} \left[\mathbf{Z} \widehat{\boldsymbol{\gamma}}^* | \mathbf{W} \right] &= \mathbf{Z} \mathrm{E}_{\mathbf{Y}|\mathbf{W}} \left[\widehat{\boldsymbol{\gamma}}^* | \mathbf{W} \right] \\ &= \mathbf{Z} \mathrm{E}_{\mathbf{Y}|\mathbf{W}} [(\mathbf{W}^{*T} \mathbf{W}^*)^{-1} \mathbf{W}^{*T} \mathbf{Y} | \mathbf{W}] \\ &= \mathbf{Z} (\mathbf{W}^{*T} \mathbf{W}^*)^{-1} \mathbf{W}^{*T} \mathrm{E}_{\mathbf{Y}|\mathbf{W}} [\mathbf{Y} | \mathbf{W}] \\ &= \mathbf{Z} (\mathbf{W}^{*T} \mathbf{W}^*)^{-1} \mathbf{W}^{*T} \mathrm{E}_{\mathbf{X}|\mathbf{W}} [\mathbf{X}^* \boldsymbol{\beta}^* | \mathbf{W}] \\ &= \mathbf{Z} (\mathbf{W}^{*T} \mathbf{W}^*)^{-1} \mathbf{W}^{*T} \mathrm{E}_{\mathbf{X}|\mathbf{W}} [\mathbf{X}^* | \mathbf{W}] \boldsymbol{\beta}^*, \end{aligned}$$

tal que

$$\mathrm{E}_{\mathbf{X}|\mathbf{W}}\left[\mathbf{X}^*|\mathbf{W}\right] = \begin{bmatrix} 1 & \pi_{1|w_1} & \dots & \pi_{M|w_1} \\ 1 & \pi_{1|w_2} & \dots & \pi_{M|w_2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \pi_{1|w_n} & \dots & \pi_{M|w_n} \end{bmatrix} = \boldsymbol{\pi}^*.$$

Logo,

$$\mathrm{E}_{\mathbf{Y}|\mathbf{W}}\left[\widehat{\boldsymbol{\beta}}_C^*|\mathbf{W}\right] = \mathbf{Z}(\mathbf{W}^{*T}\mathbf{W}^*)^{-1}\mathbf{W}^{*T}\boldsymbol{\pi}^*\boldsymbol{\beta}^*, \tag{2.24}$$

e portanto o viés $\mathbf{B}$ de $\widehat{\boldsymbol{\beta}}^*$ é dado por

$$\begin{aligned} \mathbf{B} &= \mathrm{E}_{\mathbf{Y}|\mathbf{W}}\left[\widehat{\boldsymbol{\beta}}_C^*|\mathbf{W}\right] - \boldsymbol{\beta}^* \\ &= \left(\mathbf{Z}(\mathbf{W}^{*T}\mathbf{W}^*)^{-1}\mathbf{W}^{*T}\boldsymbol{\pi}^* - \mathbb{I}\right)\boldsymbol{\beta}^*, \end{aligned}$$

onde $\mathbb{I}$ é a matriz identidade de dimensão $M \times M$. Note que quando as probabilidades condicionais

$$\pi_{j|w_k} \to \begin{cases} 1, & \text{se } j = w_k \\ 0 & \text{caso contrário} \end{cases}$$

temos que $\boldsymbol{\pi}^* \to \mathbf{W}^*$ e $\mathbf{Z} \to \mathbb{I}$ e portanto o viés $\mathbf{B} \to \mathbf{0}$.

O estimador $\widehat{\beta}_{0C}$ do intercepto β_0 na regressão de $\mathbf{Y}$ em $\mathbf{X}$, é determinado de forma independente dos outros estimadores a partir de (2.23). Consequentemente, calculamos seu viés de forma independente. Seja

$$\mathrm{E}_{\mathbf{Y}|\mathbf{W}}\left[\widehat{\beta}_{0C}|\mathbf{W}\right] = \beta_0 + B_0,$$

tal que

$$\begin{aligned} \mathrm{E}_{\mathbf{Y}|\mathbf{W}}\left[\widehat{\beta}_{0C}|\mathbf{W}\right] &= \mathrm{E}_{\mathbf{Y}|\mathbf{W}}\left[\frac{\sum_{i=1}^{n}\left(Y_i - \boldsymbol{\pi}_{(i)}\widehat{\boldsymbol{\beta}}_C\right)}{n}\Big|\mathbf{W}\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(\mathrm{E}_{\mathbf{Y}|\mathbf{W}}[Y_i|\mathbf{W}] - \boldsymbol{\pi}_{(i)}\mathrm{E}_{\mathbf{Y}|\mathbf{W}}\left[\widehat{\boldsymbol{\beta}}_C|\mathbf{W}\right]\right) \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(\mathrm{E}_{\mathbf{X}|\mathbf{W}}[(1, X_i)|\mathbf{W}]\boldsymbol{\beta}^* - \boldsymbol{\pi}_{(i)}\mathrm{E}_{\mathbf{Y}|\mathbf{W}}\left[\widehat{\boldsymbol{\beta}}_C|\mathbf{W}\right]\right) \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{\pi}_{(i)}^*\boldsymbol{\beta}^* - \boldsymbol{\pi}_{(i)}\mathrm{E}_{\mathbf{Y}|\mathbf{W}}\left[\widehat{\boldsymbol{\beta}}_C|\mathbf{W}\right]\right), \end{aligned} \tag{2.25}$$

onde $\boldsymbol{\pi}_{(i)}$ e $\boldsymbol{\pi}_{(i)}^*$ são a i-ésima linha da matriz $\boldsymbol{\pi}$ e $\boldsymbol{\pi}^*$, respectivamente. Note que

$\mathrm{E}_{\mathbf{Y}|\mathbf{W}}\left[\widehat{\boldsymbol{\beta}}_C|\mathbf{W}\right]$ é o vetor $\mathrm{E}_{\mathbf{Y}|\mathbf{W}}\left[\widehat{\boldsymbol{\beta}}^*_C|\mathbf{W}\right]$ sem o primeiro elemento. Então

$$\begin{aligned}
\mathrm{E}_{\mathbf{Y}|\mathbf{W}}\left[\widehat{\beta}_{0C}|\mathbf{W}\right] &= \frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{\pi}^*_{(i)}\boldsymbol{\beta}^* - \mathrm{E}_{\mathbf{Y}|\mathbf{W}}\left[\widehat{\boldsymbol{\beta}}_C|\mathbf{W}\right]\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(\beta_0 + \boldsymbol{\pi}_{(i)}\boldsymbol{\beta} - \mathrm{E}_{\mathbf{Y}|\mathbf{W}}\left[\widehat{\boldsymbol{\beta}}_C|\mathbf{W}\right]\right) \\
&= \beta_0 + \frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{\pi}_{(i)}\boldsymbol{\beta} - \mathrm{E}_{\mathbf{Y}|\mathbf{W}}\left[\widehat{\boldsymbol{\beta}}_C|\mathbf{W}\right]\right) \\
&= \beta_0 + B_0. \qquad (2.26)
\end{aligned}$$

Logo $B_0 = \dfrac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{\pi}_{(i)}\boldsymbol{\beta} - \mathrm{E}_{\mathbf{Y}|\mathbf{W}}\left[\widehat{\boldsymbol{\beta}}_C|\mathbf{W}\right]\right)$.

2.5 Cálculo de Variância de Estimadores

Seja $\boldsymbol{\gamma}_* = (\gamma_0, \boldsymbol{\gamma})$ o vetor de parâmetros da regressão de $\mathbf{Y}$ em $\mathbf{W}^* = (\mathbf{1}, \mathbf{W})$ dado por (2.5). E seja

$$\widehat{\boldsymbol{\gamma}}^* = (\widehat{\gamma}_0, \widehat{\boldsymbol{\gamma}}) = \left(\mathbf{W}^{*T}\mathbf{W}^*\right)^{-1}\mathbf{W}^{*T}\mathbf{Y}.$$

A variância de $\widehat{\boldsymbol{\gamma}}^*$ é calculada a seguir. Considere

$$\begin{aligned}
\mathrm{Var}(\widehat{\boldsymbol{\gamma}}^*) &= \mathrm{Var}\left((\mathbf{W}^{*T}\mathbf{W}^*)^{-1}\mathbf{W}^{*T}\mathbf{Y}\right) \\
&= (\mathbf{W}^{*T}\mathbf{W}^*)^{-1}\mathbf{W}^{*T}\mathrm{Var}\left(\mathbf{Y}\right)\mathbf{W}^*(\mathbf{W}^{*T}\mathbf{W}^*)^{-1} \\
&= (\mathbf{W}^{*T}\mathbf{W}^*)^{-1}\sigma^2_W \qquad (2.27)
\end{aligned}$$

$\widehat{\boldsymbol{\beta}}^*_C = (\widehat{\gamma}_0, \widehat{\boldsymbol{\beta}}_C)^T$ e $\mathbf{Z} = \begin{bmatrix} 1 & \mathbf{0}^T \\ \mathbf{0} & \left(\Sigma_W^{-1}\Sigma_{WX}\right)^{-1} \end{bmatrix}$, tal que

$$\begin{aligned}
\mathrm{Var}\left(\widehat{\boldsymbol{\beta}}^*_C\right) &= \mathrm{Var}\left(\mathbf{Z}\widehat{\boldsymbol{\gamma}}^*\right) \\
&= \mathrm{Var}\left(\mathbf{Z}\left(\mathbf{W}^{*T}\mathbf{W}^*\right)^{-1}\mathbf{W}^{*T}\mathbf{Y}\right) \\
&= \mathbf{Z}\left(\mathbf{W}^{*T}\mathbf{W}^*\right)^{-1}\mathbf{W}^{*T}\mathrm{Var}\left(\mathbf{Y}\right)\mathbf{W}^*\left(\mathbf{W}^{*T}\mathbf{W}^*\right)^{-1}\mathbf{Z}^T \\
&= \sigma^2_W\mathbf{Z}\left(\mathbf{W}^{*T}\mathbf{W}^*\right)^{-1}\mathbf{Z}^T. \qquad (2.28)
\end{aligned}$$

Portanto, pela construção de $\mathbf{Z}$, $\mathrm{Var}\left(\widehat{\boldsymbol{\beta}}_C\right)$ é igual a $\mathrm{Var}\left(\widehat{\boldsymbol{\beta}}^*_C\right)$ sem o primeiro elemento do vetor.

Utilizando a equação (2.23), obtemos

$$\begin{aligned}
\text{Var}\left(\widehat{\beta}_{0C}\right) &= \text{Var}\left(\frac{\sum_{i=1}^{n}\left(Y_i - \boldsymbol{\pi}_{(i)}\widehat{\boldsymbol{\beta}}_C\right)}{n}\right) \\
&= \text{Var}\left(\frac{\sum_{i=1}^{n}\left(Y_i - \boldsymbol{\pi}_{(i)}\left(\Sigma_W^{-1}\Sigma_{WX}\right)^{-1}\widehat{\boldsymbol{\gamma}}\right)}{n}\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\Big[\text{Var}\left(Y_i\right) + \text{Var}\left(\boldsymbol{\pi}_{(i)}\left(\Sigma_W^{-1}\Sigma_{WX}\right)^{-1}\widehat{\boldsymbol{\gamma}}\right) \\
&\quad - 2\text{Cov}\left(Y_i, \boldsymbol{\pi}_{(i)}\left(\Sigma_W^{-1}\Sigma_{WX}\right)^{-1}\widehat{\boldsymbol{\gamma}}\right)\Big]. \qquad (2.29)
\end{aligned}$$

A fim de calcular $\text{Cov}\left(Y_i, \boldsymbol{\pi}_{(i)}\left(\Sigma_W^{-1}\Sigma_{WX}\right)^{-1}\widehat{\boldsymbol{\gamma}}\right)$ é necessário definirmos $\widehat{\boldsymbol{\gamma}}$ em função de $\mathbf{Y}$. Então, utilizando (2.5) encontramos a solução

$$\min_{\gamma_0,\boldsymbol{\gamma}} Q(\gamma_0, \boldsymbol{\gamma}) = \sum_{i=1}^{n}\left(Y_i - \gamma_0 - \mathbf{w}_i\boldsymbol{\gamma}\right)^2, \qquad (2.30)$$

onde $\mathbf{w}_i$ é a i-ésima linha de $\mathbf{W}$. Para isso, temos

$$\frac{\partial Q}{\partial \gamma_0} = -2\sum_{i=1}^{n}\left(Y_i - \gamma_0 - \mathbf{w}_i\boldsymbol{\gamma}\right)$$

tal que para $\frac{\partial Q}{\partial \gamma_0} = 0$,

$$\begin{aligned}
\frac{\partial Q}{\partial \gamma_0} &= -2\sum_{i=1}^{n}\left(Y_i - \gamma_0 - \mathbf{w}_i\boldsymbol{\gamma}\right) = 0 \\
\gamma_0 &= \frac{\sum_{i=1}^{n}\left(Y_i - \mathbf{w}_i\boldsymbol{\gamma}\right)}{n}. \qquad (2.31)
\end{aligned}$$

Similarmente, e utilizando (2.31)

$$\begin{aligned}
\frac{\partial Q}{\partial \boldsymbol{\gamma}} &= -2\sum_{i=1}^{n}\left(Y_i - \gamma_0 - \mathbf{w}_i\boldsymbol{\gamma}\right)\mathbf{w}_i^T \\
&= -2\sum_{i=1}^{n}\left(Y_i - \frac{\sum_{j=1}^{n}\left(Y_j - \mathbf{w}_j\boldsymbol{\gamma}\right)}{n} - \mathbf{w}_i\boldsymbol{\gamma}\right)\mathbf{w}_i^T
\end{aligned}$$

tal que para $\frac{\partial Q}{\partial \boldsymbol{\gamma}} = \mathbf{0}$,

$$\sum_{i=1}^{n} Y_i\mathbf{w}_i^T - \sum_i\sum_j \frac{\left(Y_j - \mathbf{w}_j\widehat{\boldsymbol{\gamma}}\right)}{n}\mathbf{w}_i^T - \sum_i \mathbf{w}_i\widehat{\boldsymbol{\gamma}}\mathbf{w}_i^T = \mathbf{0},$$

portanto,

$$-\sum_i \sum_j \frac{(Y_j - \mathbf{w}_j\widehat{\boldsymbol{\gamma}})}{n}\mathbf{w}_i^T = \frac{\sum_i \sum_j Y_j \mathbf{w}_i^T - \sum_i \sum_j \mathbf{w}_j \widehat{\boldsymbol{\gamma}} \mathbf{w}_i^T}{n}, \tag{2.32}$$

sendo assim,

$$\begin{aligned}\sum_i Y_i \mathbf{w}_i^T - \frac{\sum_i \sum_j Y_j \mathbf{w}_i^T}{n} &= -\frac{\sum_i \sum_j \mathbf{w}_j \widehat{\boldsymbol{\gamma}} \mathbf{w}_i^T}{n} + \sum_i \mathbf{w}_i \widehat{\boldsymbol{\gamma}} \mathbf{w}_i^T \\ &= -\frac{\sum_i \sum_j \mathbf{w}_j^T \mathbf{w}_i \widehat{\boldsymbol{\gamma}}}{n} + \sum_i \mathbf{w}_i^T \mathbf{w}_i \widehat{\boldsymbol{\gamma}},\end{aligned} \tag{2.33}$$

dessa maneira,

$$\sum_i (Y_i - \bar{Y}) w_i^T = \mathbf{A}\widehat{\boldsymbol{\gamma}}, \tag{2.34}$$

onde $\mathbf{A} = \sum_{i=1}^{n} \left(\mathbf{w}_i^T \left(\mathbf{w}_i - \frac{\sum_j \mathbf{w}_j}{n} \right) \right)$.
Consequentemente

$$\widehat{\boldsymbol{\gamma}} = A^{-1} \left(\sum_{k=1}^{n} \left(Y_K - \bar{Y} \right) \mathbf{w}_k^T \right) \tag{2.35}$$

Substituindo (2.35) em $\mathrm{Var}(\boldsymbol{\pi}_{(i)} \left(\Sigma_W^{-1} \Sigma_{WX} \right)^{-1} \widehat{\boldsymbol{\gamma}})$ obtemos

$$\begin{aligned}&\mathrm{Var} \left(\boldsymbol{\pi}_{(i)} \left(\Sigma_W^{-1} \Sigma_{WX} \right)^{-1} \widehat{\boldsymbol{\gamma}} \right) = \\ &= \boldsymbol{\pi}_{(i)} \left(\Sigma_W^{-1} \Sigma_{WX} \right)^{-1} \mathbf{A}^{-1} \mathrm{Var} \left(\sum_{k=1}^{n} \left(Y_K - \bar{Y} \right) \mathbf{w}_k^T \right) \left(\mathbf{A}^{-1} \right)^T \left(\left(\Sigma_W^{-1} \Sigma_{WX} \right)^{-1} \right)^T \boldsymbol{\pi}_{(i)}^T.\end{aligned} \tag{2.36}$$

Definindo $V_i = \boldsymbol{\pi}_{(i)} \left(\Sigma_W^{-1} \Sigma_{WX} \right)^{-1}$ e como

$$\sum_k \left(Y_k - \bar{Y} \right) \mathbf{w}_k^T = \sum_k \left(\mathbf{w}_k^T - \frac{1}{n} \sum_{l=1}^{n} \mathbf{w}_l^T \right) Y_k,$$

$$\begin{aligned}&\mathrm{Var} \left(\boldsymbol{\pi}_{(i)} \left(\Sigma_W^{-1} \Sigma_{WX} \right)^{-1} \widehat{\boldsymbol{\gamma}} \right) = \\ &= \sigma^2 \left[V_i \mathbf{A}^{-1} \left(\sum_k \left(\mathbf{w}_k^T - \frac{1}{n} \sum_l \mathbf{w}_l^T \right) \left(\mathbf{w}_k - \frac{1}{n} \sum_l \mathbf{w}_l \right) \right) \left(\mathbf{A}^{-1} \right)^T V_i^T \right]\end{aligned} \tag{2.37}$$

ademais, substituindo (2.35) em $\text{Cov}\left(Y_i, \boldsymbol{\pi}_{(i)} \left(\Sigma_W^{-1}\Sigma_{WX}\right)^{-1} \widehat{\boldsymbol{\gamma}}\right)$

$$\begin{aligned}
&\text{Cov}\left(Y_i, \boldsymbol{\pi}_{(i)} \left(\left(\Sigma_W^{-1}\Sigma_{WX}\right)^{-1}\right)^T \widehat{\boldsymbol{\gamma}}\right) = \qquad (2.38)\\
&= \text{Cov}\left(Y_i, V_i\mathbf{A}^{-1}\left(\sum_{k=1}^{n}\left(\mathbf{Y}_k - \bar{Y}\right)\mathbf{w}_k^T\right)\right)\\
&= \text{Cov}\left(Y_i, \mathbf{A}^{-1}\left(\sum_{k=1}^{n}\left(\mathbf{Y}_k - \bar{Y}\right)\mathbf{w}_k^T\right)\right)\left(\mathbf{A}^{-1}\right)^T V_i^T\\
&= \sigma^2\left(\mathbf{w}_i - \frac{1}{n}\sum_{l}\mathbf{w}_l\right)\left(\mathbf{A}^{-1}\right)^T V_i^T.
\end{aligned}$$

Concluindo, utilizando as equações (2.29), (2.37) e (2.38) podemos calcular a variância do estimador corrigido $\widehat{\beta}_{0C}$.

3 Estudo de simulação

Realizamos um estudo de simulação para avaliar o método de correção dos estimadores de mínimos quadrados do modelo de regressão utilizando as covariáveis observadas $\mathbf{W}$, assim como a precisão destes. Adicionalmente, o método de correção do intercepto definido em (2.23) é avaliado comparando seus resultados ao valor verdadeiro e à não-correção. A avaliação será feita comparando o erro quadrático ponderado definido em (3.5) de três métodos de estimação, sendo eles: regressão ingênua sem correção para $\boldsymbol{\beta}$ e β_0 (sem correção), correção para $\boldsymbol{\beta}$ e β_0 (correção total) e correção apenas para $\boldsymbol{\beta}$ (correção parcial). Além disso, neste estudo estamos interessados em verificar o efeito do número de observações, desvio padrão da variável resposta, número de variáveis categóricas, magnitude das componentes de $\boldsymbol{\theta}$ e número de categorias na bondade de correção em $\widehat{\boldsymbol{\beta}}_C$.

Definimos o número de variáveis $K = 1, 3, 10, 30, 50$, categorias $L_k = 2, 3, 4$, para todo $k = 1, \ldots, K$ e número de observações $n = 50, 75, 100, \ldots, 500$. Ademais, definimos $P(W = w|X = x)$ que compõem os elementos de $\boldsymbol{\theta}$ em três cenários (pouca distorção, média distorção e alta distorção).

1. **Pouca distorção:**

$$\boldsymbol{\theta} = \begin{cases} \begin{array}{|c|cc|}\hline & \text{W=0} & \text{W=1} \\ \hline \text{X=0} & 0.9 & 0.1 \\ \text{X=1} & 0.15 & 0.85 \\ \hline \end{array} & \text{se } L_K = 2, \\ \begin{array}{|c|ccc|}\hline & \text{W=0} & \text{W=1} & \text{W=2} \\ \hline \text{X=0} & 0.85 & 0.1 & 0.05 \\ \text{X=1} & 0.1 & 0.8 & 0.1 \\ \text{X=2} & 0.05 & 0.1 & 0.85 \\ \hline \end{array} & \text{se } L_K = 3 \text{ e} \\ \begin{array}{|c|cccc|}\hline & \text{W=0} & \text{W=1} & \text{W=2} & \text{W=3} \\ \hline \text{X=0} & 0.825 & 0.1 & 0.05 & 0.025 \\ \text{X=1} & 0.075 & 0.8 & 0.075 & 0.05 \\ \text{X=2} & 0.05 & 0.075 & 0.8 & 0.075 \\ \text{X=3} & 0.025 & 0.05 & 0.1 & 0.825 \\ \hline \end{array} & \text{se } L_K = 4. \end{cases} \quad . \tag{3.1}$$

2. **Média distorção:**

$$\boldsymbol{\theta} = \begin{cases} \begin{array}{|c|cc|}\hline & \text{W=0} & \text{W=1} \\ \hline \text{X=0} & 0.7 & 0.3 \\ \text{X=1} & 0.35 & 0.65 \\ \hline \end{array} & \text{se } L_K = 2, \\ \\ \begin{array}{|c|ccc|}\hline & \text{W=0} & \text{W=1} & \text{W=2} \\ \hline \text{X=0} & 0.7 & 0.2 & 0.1 \\ \text{X=1} & 0.15 & 0.7 & 0.15 \\ \text{X=2} & 0.1 & 0.2 & 0.7 \\ \hline \end{array} & \text{se } L_K = 3 \text{ e} \quad . \\ \\ \begin{array}{|c|cccc|}\hline & \text{W=0} & \text{W=1} & \text{W=2} & \text{W=3} \\ \hline \text{X=0} & 0.6 & 0.2 & 0.125 & 0.075 \\ \text{X=1} & 0.15 & 0.6 & 0.15 & 0.1 \\ \text{X=2} & 0.1 & 0.15 & 0.6 & 0.15 \\ \text{X=3} & 0.075 & 0.125 & 0.2 & 0.6 \\ \hline \end{array} & \text{se } L_K = 4. \end{cases} \tag{3.2}$$

3. **Alta distorção:**

$$\boldsymbol{\theta} = \begin{array}{|c|cccc|}\hline & \text{W=0} & \text{W=1} & \text{W=2} & \text{W=3} \\ \hline \text{X=0} & 0.3 & 0.25 & 0.25 & 0.2 \\ \text{X=1} & 0.25 & 0.3 & 0.25 & 0.2 \\ \text{X=2} & 0.2 & 0.25 & 0.3 & 0.25 \\ \text{X=3} & 0.2 & 0.25 & 0.25 & 0.3 \\ \hline \end{array} \quad \text{sendo } L_K = 4. \tag{3.3}$$

(3.4)

Em todos os cenários, com todos os parâmetros iniciais definidos, simulamos a matriz de desenho $\mathbf{X}$ e então, para simular o erro observacional, utilizando $\boldsymbol{\theta}$ geramos a matriz de desenho $\mathbf{W}$, tal que $W_i|X_i = x \sim \text{Mult}(1, \boldsymbol{\theta}_x)$ onde θ_x é a x-ésima linha da matriz $\boldsymbol{\theta}$. Adicionalmente, transformamos $\boldsymbol{\theta}$ em $\boldsymbol{\pi}$ como descrito em (2.2). Para finalizar, definimos $\beta_l = 0.5 + 0.2l$ para todo $l \in 0, 1, \ldots, 1 + \sum_{k=1}^{K}(L_k - 1)$. As matrizes de desenho $\mathbf{X}$ e $\mathbf{W}$ foram geradas inicialmente para tamanhos de amostra $n = 500$ e foram reduzidas para os tamanhos de amostra menores, de modo a termos dados encaixados. Isto é, por exemplo, a amostra de tamanho 400 é uma subamostra da amostra de tamanho 500. Após

todo o preparo, utilizamos $\boldsymbol{\beta}$ para simular o valor da variável resposta $\mathbf{Y}|\mathbf{W} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbb{I})$ onde $\sigma = 0.1, 0.2, 0.5, 1$. No caso de alta distorção, os resultados apresentados só contêm os casos onde σ assumiu os valores 0.1 e 1, pois não havia mais informação relevante nos outros gráficos. Realizamos a regressão de mínimos quadrados com a matriz de desenho $\mathbf{W}$ em (2.5) e (2.17) para obter a estimativa $\widehat{\boldsymbol{\gamma}}$ e $\widehat{\gamma}_0$. Em seguida, utilizamos o método de correção proposto para obter o estimador $\widehat{\boldsymbol{\beta}}_C$. Para finalizar, corrigimos $\widehat{\gamma}_0$ para obter $\widehat{\beta}_{C0}$. Repetimos esse processo $M = 300$ vezes e utilizamos a média do erro quadrático ponderado para comparar a efetividade das estimativas. O erro quadrático ponderado é dado pela equação

$$\text{EQP}_\alpha = \frac{\displaystyle\sum_{l=1}^{M} \frac{(\boldsymbol{\beta}_l - \widehat{\beta}_{\alpha l})^2}{\beta_l}}{1 + \displaystyle\sum_{k=1}^{K}(L_k - 1)}, \tag{3.5}$$

onde α é o método de correção.

3.1 Resultados da simulação

Nesta seção apresentamos os resultados de EQP para os diversos estudos de simulação, bem como a variância dos estimadores do intercepto pelo método sem correção e do método correção total.

1. **Pouca distorção:**

As Figuras 1 – 3 apresentam o EQP nos casos de pouca distorção, de números diferentes de categorias, número de observações, quantidade de variáveis categóricas e desvios padrões. Podemos observar que, neste caso, o modelo com correção parcial desempenha significativamente melhor que o modelo de não correção quando há baixa quantidade de covariáveis categóricas. Nos outros casos, tem um desempenho igual ou muito similar, exceto quando o número de observações é pequeno e a quantidade de categorias é igual a 3, onde o método sem correção apresenta um EQP menor, observado na Figura 2. Por outro lado, a performance do método de correção total é melhor em quase todos os casos de amostras pequenas e, nas simulações realizadas, com tamanho de amostra suficientemente grande, a magnitude de EQP calculada é menor do que os demais métodos. Na Figura 3 ,observamos que no caso de desvio padrão igual a 1 e 3 covariáveis categóricas, o método de correção total é inferior ao método sem correção com amostra menor que 100 indivíduos. Enquanto, independentemente do número de categorias, observável nos painéis superiores direito das Figuras 1 – 3, com desvio padrão igual a 1 e apenas uma covariável categórica, os métodos de correção são significativamente piores que não corrigir, até quando a amostra é suficientemente grande, seus EQPs são parecidos. Ademais, ao comparar os casos de mesma quantidade de covariáveis, independentemente

do número de categorias, percebemos que o aumento em desvio padrão não afeta a forma das figuras, exceto no caso de baixo número de covariáveis categóricas, enquanto com $K \geqslant 10$ as colunas das figuras se mantêm muito parecidas. No caso de pouca distorção, quanto maior o número de categorias, mais observações são necessárias para estabilizar o EQP do método de correção total em torno de seu mínimo assintótico, algo que pode ser facilmente observado quando comparado o caso de $K = 10$ pois na Figura 3 percebemos que um valor de EQP próximo do mínimo é observado quando o tamanho da amostra é 75 enquanto na Figura 1 este valor próximo do mínimo é só atingido quando o $n = 150$. Além do mais, quando o número de categorias e/ou número de covariáveis categóricas aumenta, a regressão não pode ser realizada pois não havia observações o suficiente para obter uma estimativa para todos os parâmetros. Como esperado, o EQP dos métodos de correção diminui conforme o número de observações aumenta, pois os métodos são assintóticos, e também os EQPs aumentam conforme o número de covariáveis categóricas aumenta, pois há mais parâmetros a serem estimados.

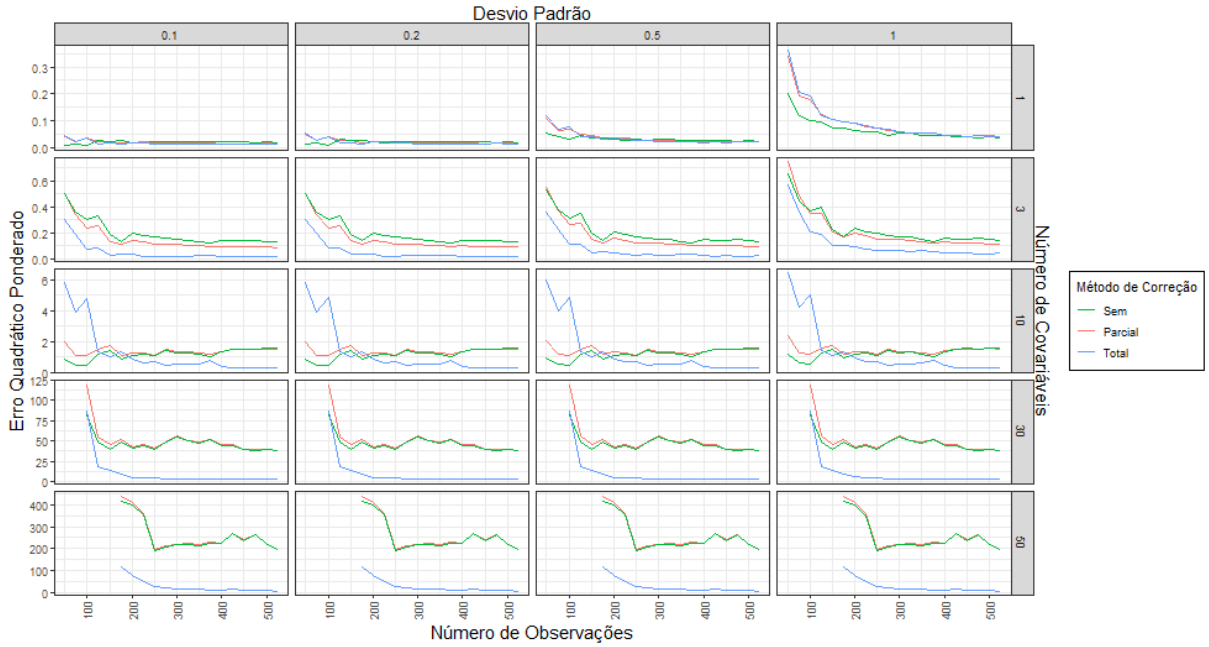

Figura 1 – EQP calculado para o caso de pouca distorção com $L_k = 4, k = 1, \ldots, K$, para diferentes número de observações, número de variáveis categóricas e desvios padrões

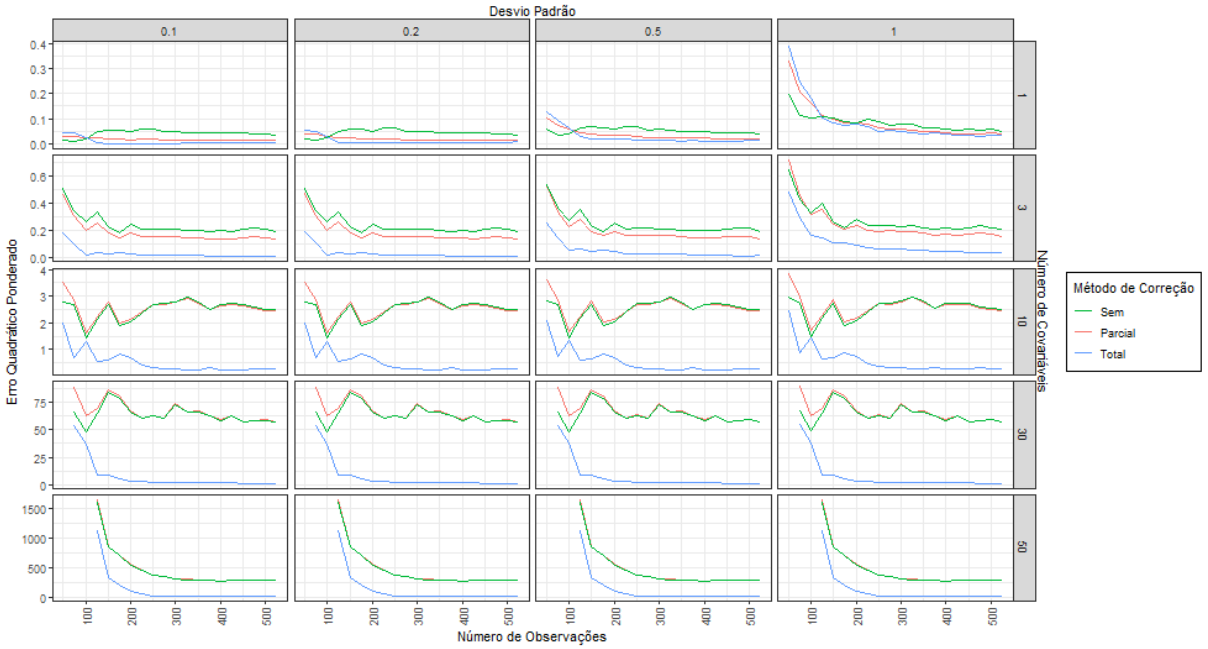

Figura 2 – EQP calculado para o caso de pouca distorção com $L_k = 3, k = 1, \ldots, K$, para diferentes número de observações, número de variáveis categóricas e desvios padrões

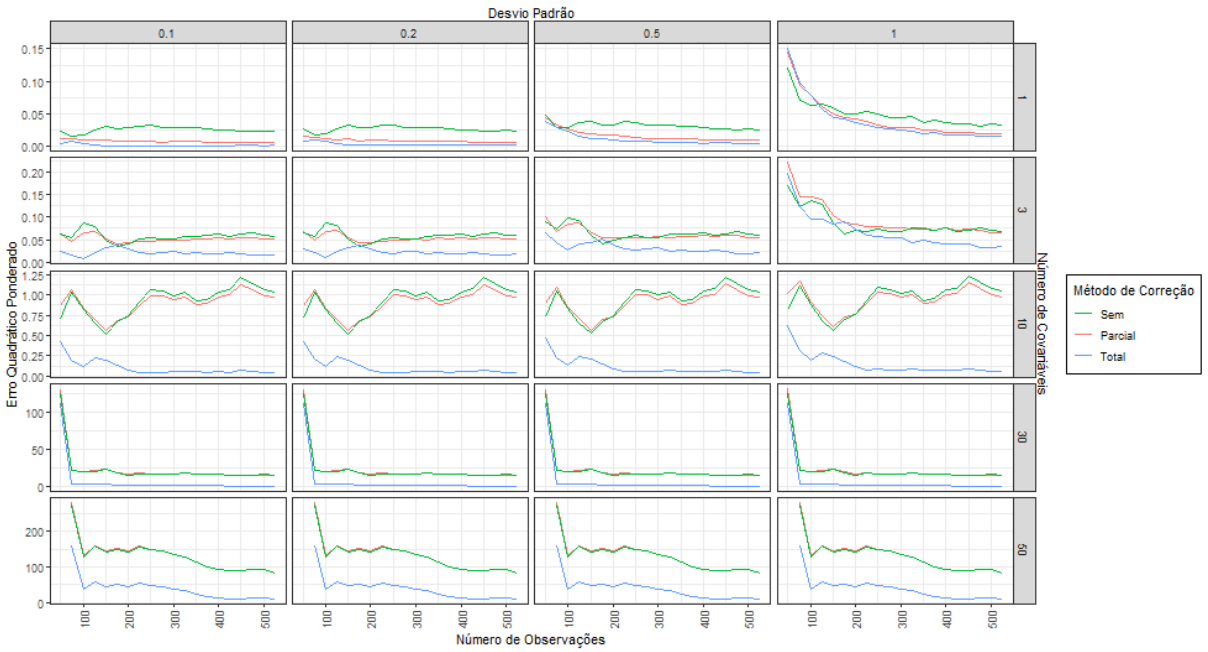

Figura 3 – EQP calculado para o caso de pouca distorção com $L_k = 2, k = 1, \ldots, K$, para diferentes número de observações, número de variáveis categóricas e desvios padrões

As Figuras 4 – 6 mostram a variância empírica para as estimativas dos in-

terceptos do método de correção total e sem correção calculadas após as 300 iterações. Independentemente do número de covariáveis e da quantidade de categorias, as figuras são similares e mostram que a variância do estimador do intercepto do método sem correção é estável e sempre menor que a variância do estimador proposto. Ademais, ao contrário da magnitude de EQP, a variância depende do desvio padrão atribuído à variável resposta, onde um alto desvio padrão implica em uma alta variância para o estimador proposto em amostras pequenas. Essa variância estabiliza e aproxima 0 conforme o tamanho da amostra aumenta.

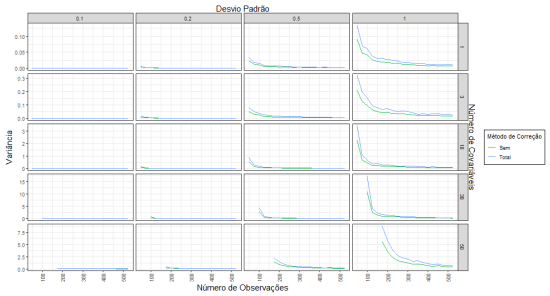

Figura 4 – Variância calculada para o caso de pouca distorção com $L_k = 4, k = 1, \ldots, K$, para diferentes número de observações, número de variáveis categóricas e desvios padrões

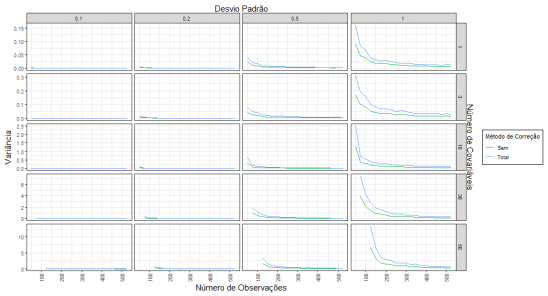

Figura 5 – Variância calculada para o caso de pouca distorção com $L_k = 3, k = 1, \ldots, K$, para diferentes número de observações, número de variáveis categóricas e desvios padrões

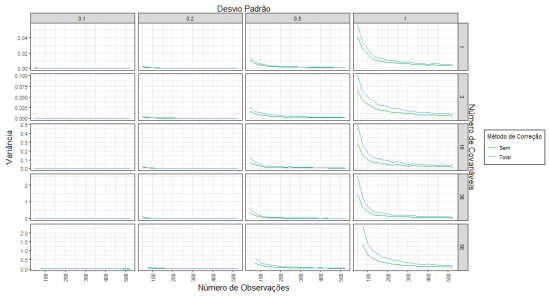

Figura 6 – Variância calculada para o caso de pouca distorção com $L_k = 2, k = 1, \ldots, K$, para diferentes número de observações, número de variáveis categóricas e desvios padrões

2. **Média distorção:**

As Figuras 7 – 9 apresentam o EQP nos casos de média distorção, de números diferentes de categorias, número de observações, quantidade de variáveis categóricas e desvios padrões. De forma geral, o comportamento dos estimadores corrigidos se mantém similares ao caso de pouca distorção, a saber, o desvio padrão continua afetando somente os casos de poucas covariáveis categóricas, independentemente do número de categorias,

continua havendo casos onde não é possível realizar a estimação pois o número de parâmetros a estimar é superior ao número de observações, o EQP do método de correção total se aproxima do 0 quando o número de observações aumenta e o EQP do método de correção parcial se aproxima do EQP do método de não correção enquanto o EQP do método de correção total é inferior a ambos na maioria dos casos estudados. Porém, existe uma quantidade maior de casos onde o método de não correção desempenha melhor que o método de correção total. Isso se deve a uma necessidade de uma amostra maior para estabilizar o EQP do método de correção total em torno do 0.

As Figuras 10 – 12 mostram a variância empírica para as estimativas dos interceptos do método de correção total e sem correção calculadas após as 300 iterações. Apesar do comportamento observado nessas figuras ser o mesmo que no caso de pouca distorção, a magnitude da variância de ambos métodos é maior que no caso anterior.

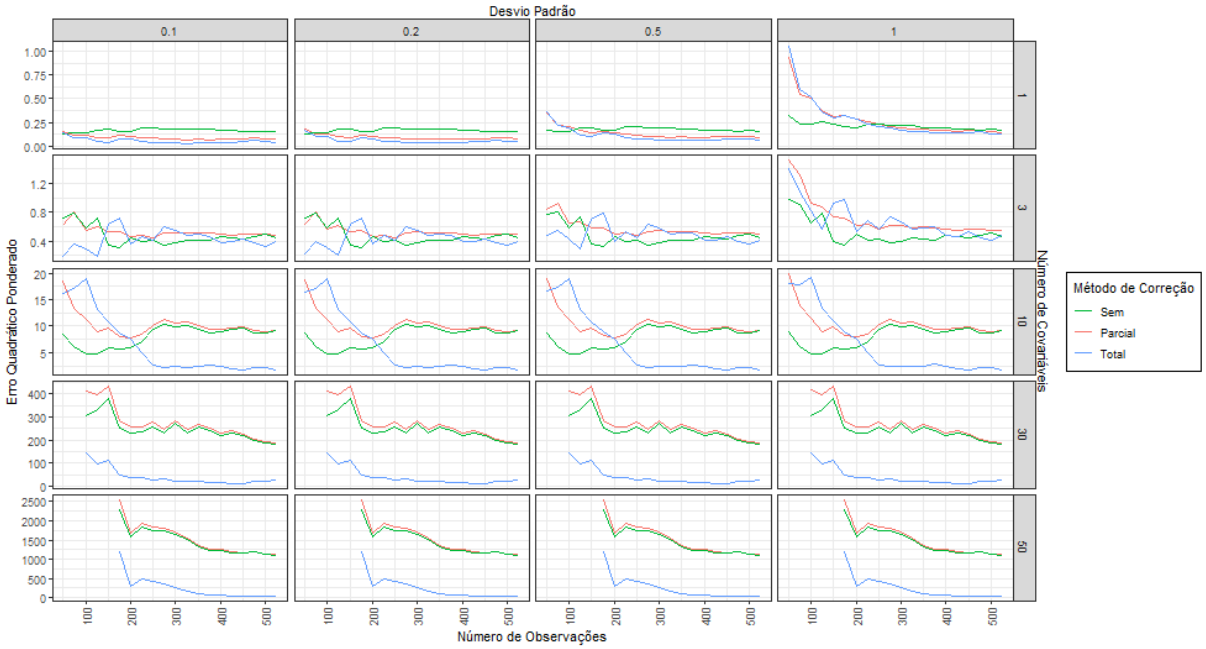

Figura 7 – EQP calculado para o caso de média distorção com $L_k = 4, k = 1, \ldots, K$, para diferentes número de observações, número de variáveis categóricas e desvios padrões

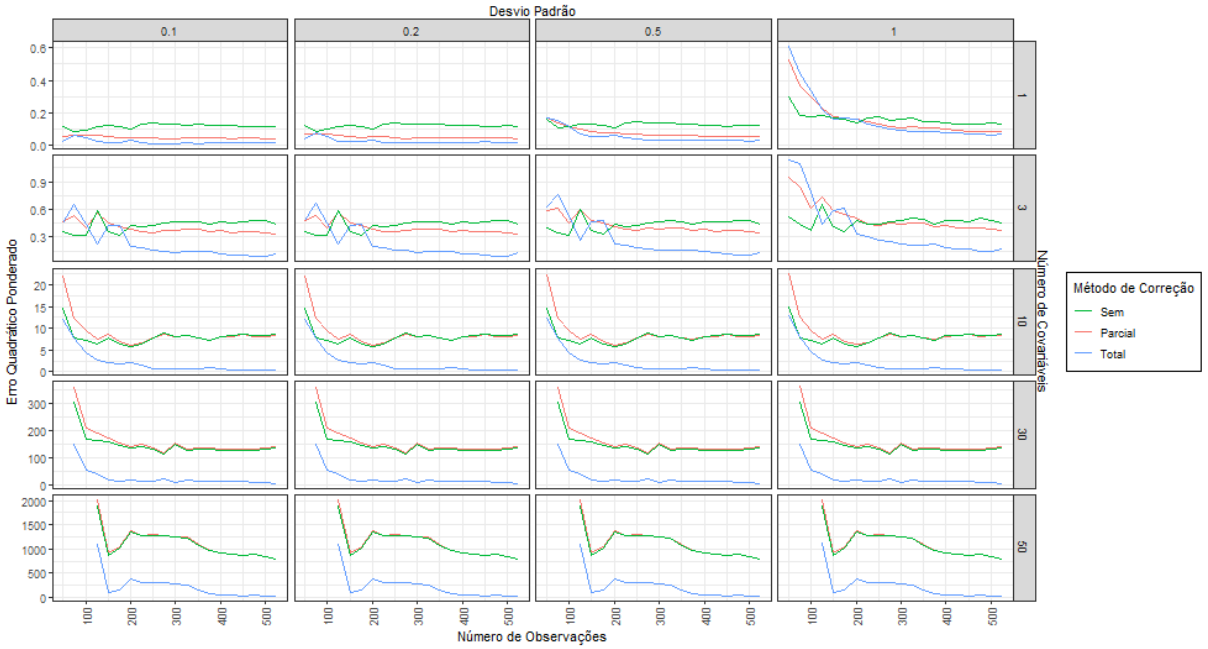

Figura 8 – EQP calculado para o caso de média distorção com $L_k = 3, k = 1, \ldots, K$, para diferentes número de observações, número de variáveis categóricas e desvios padrões

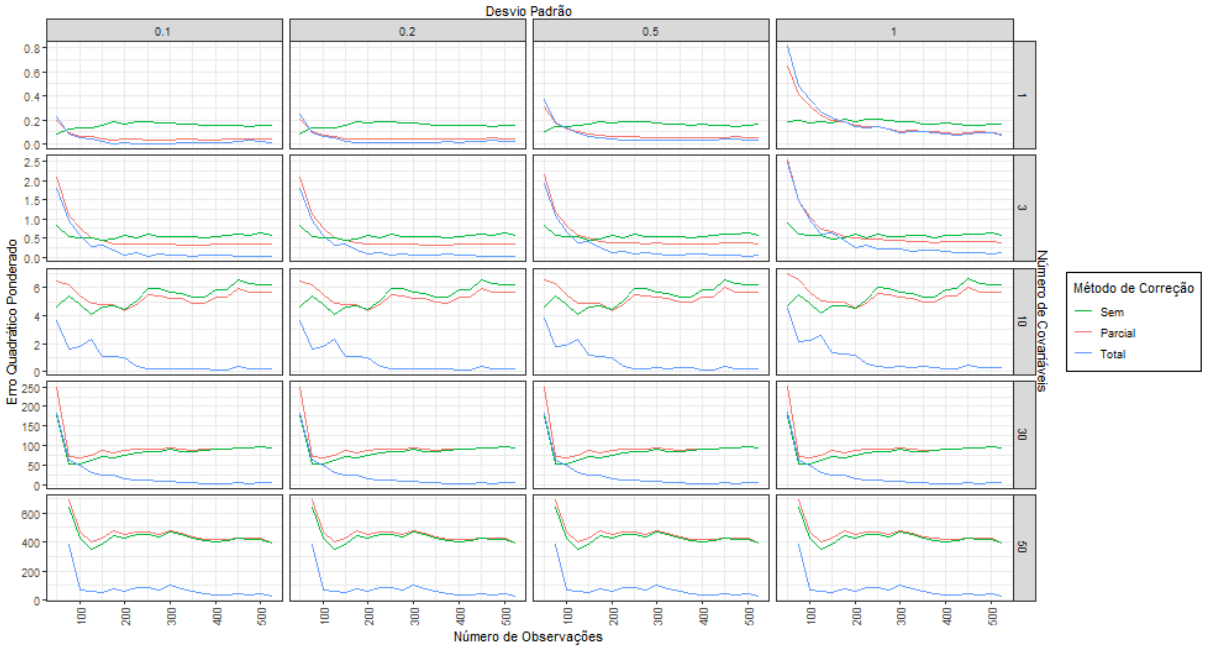

Figura 9 – EQP calculado para o caso de média distorção com $L_k = 2, k = 1, \ldots, K$, para diferentes número de observações, número de variáveis categóricas e desvios padrões

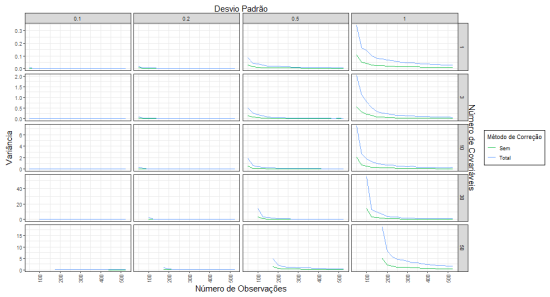

Figura 10 – Variância calculada para o caso de média distorção com $L_k = 4, k = 1, \ldots, K$, para diferentes número de observações, número de variáveis categóricas e desvios padrões

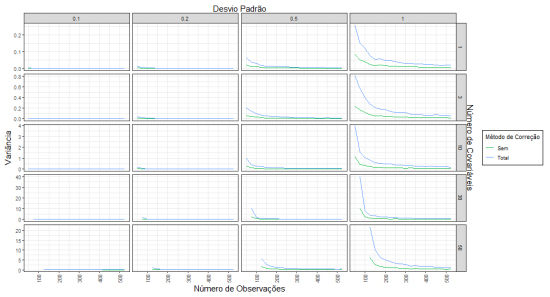

Figura 11 – Variância calculada para o caso de média distorção com $L_k = 3, k = 1, \ldots, K$, para diferentes número de observações, número de variáveis categóricas e desvios padrões

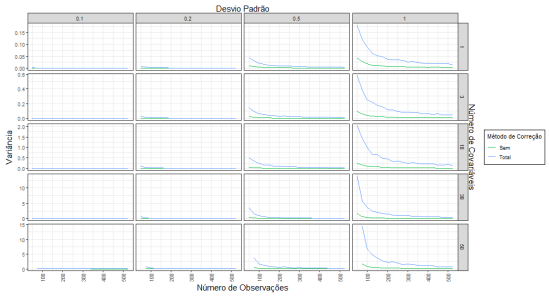

Figura 12 – Variância calculada para o caso de média distorção com $L_k = 2, k = 1, \ldots, K$, para diferentes número de observações, número de variáveis categóricas e desvios padrões

3. **Alta distorção:**

No caso de alta distorção, menos casos serão apresentados devido a precariedade dos ajustes. Assim, foram selecionados somente os casos com $L_k = 4$, $\sigma = 0.1$, $\sigma = 1$ e $K = 1, 3, 10, 30, 50$. A Figura 13 apresenta o EQP calculado para estes casos, onde percebemos que a magnitude do EQP quando comparada aos casos de média e baixa distorção, são maiores, seu máximo variando de aproximadamente 2500 na Figura 7 para mais de 10000 na Figura 13. Além do mais, o método de não correção desempenha melhor que os métodos corrigidos. Porém o método de não correção mantém um EQP estável independentemente do número de observações, enquanto os métodos corrigidos melhoram seu desempenho quanto maior a amostra.

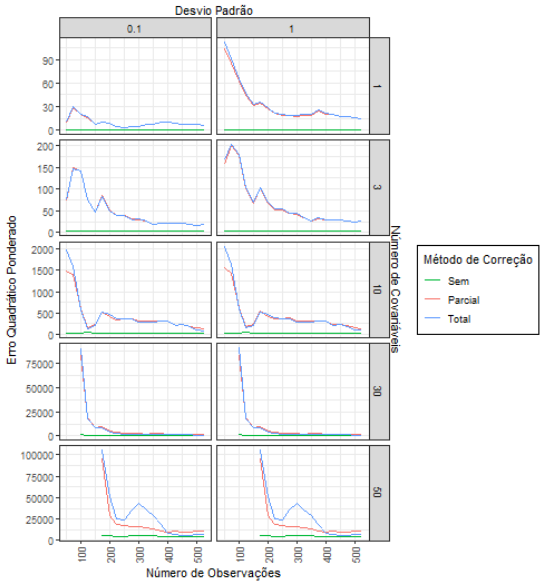

Figura 13 – EQP calculado para o caso de alta distorção com $L_k = 4, k = 1, \ldots, K$, para diferentes número de observações, número de variáveis categóricas e desvios padrões

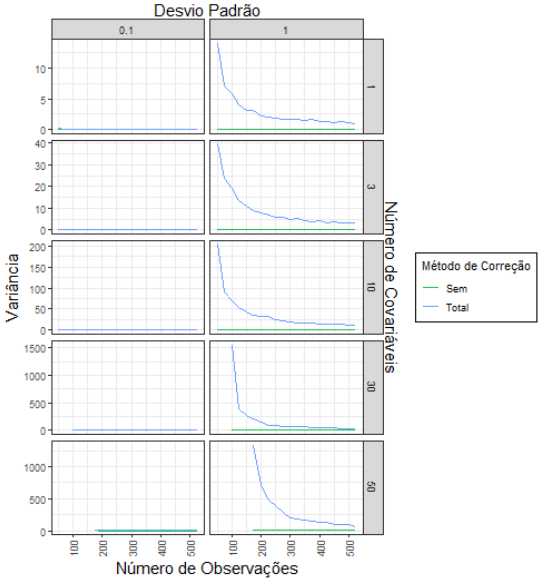

Figura 14 – Variância calculada para o caso de alta distorção com $L_k = 4, k = 1, \ldots, K$, para diferentes número de observações, número de variáveis categóricas e desvios padrões

4. L_k **aleatório:**

Simulamos um caso onde cada variável categórica foi atribuída um número de categorias possíveis aleatoriamente, sendo $L_k = 2$, $L_k = 3$ ou $L_k = 4$ com probabilidade igual a $\frac{1}{3}$. Pelas Figuras 15 e 16, observamos que quando L_k é aleatório a diferença entre o método de correção total e não correção é ainda mais clara onde seu desempenho é melhor em todos os casos exceto quando há pequenas amostras e alto número de covariáveis, ou com baixo número de covariáveis e desvio padrão alto. Novamente, pela Figura 17, podemos perceber que o método de correção tem pobre desempenho com alta distorção, entretanto, neste caso, quando há amostras grandes de uma quantidade média de covariáveis utilizar método de correção é melhor que não corrigir.

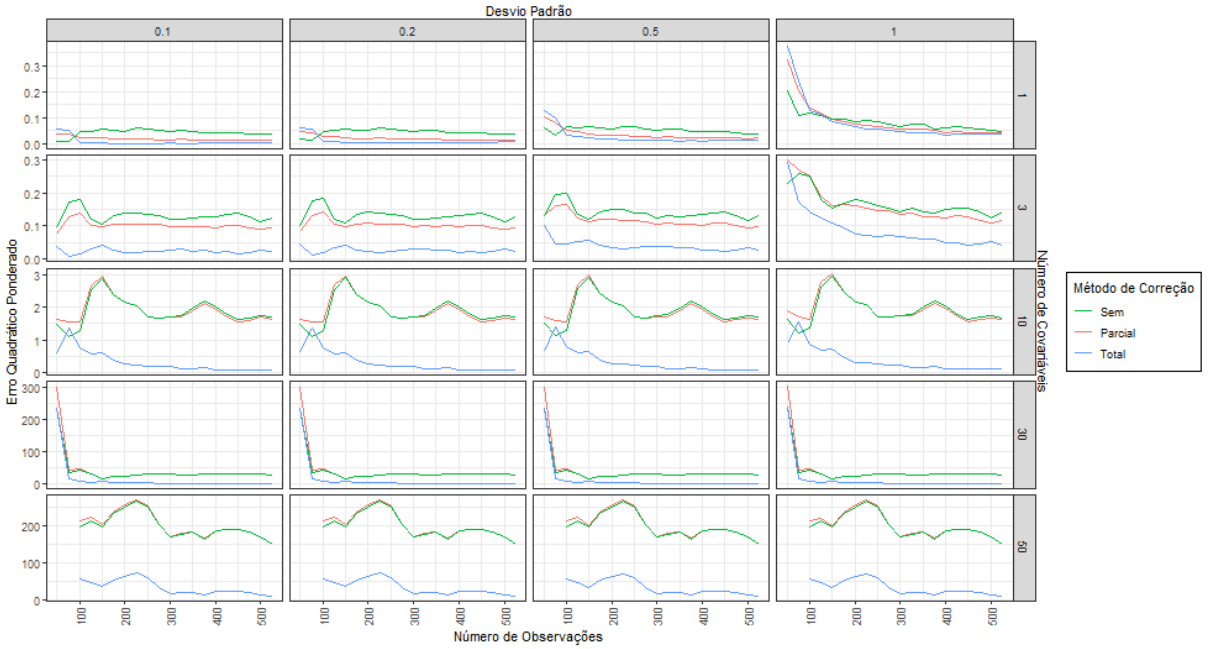

Figura 15 – EQP calculado para o caso de pouca distorção com L_k aleatório, para diferentes número de observações, número de variáveis categóricas e desvios padrões

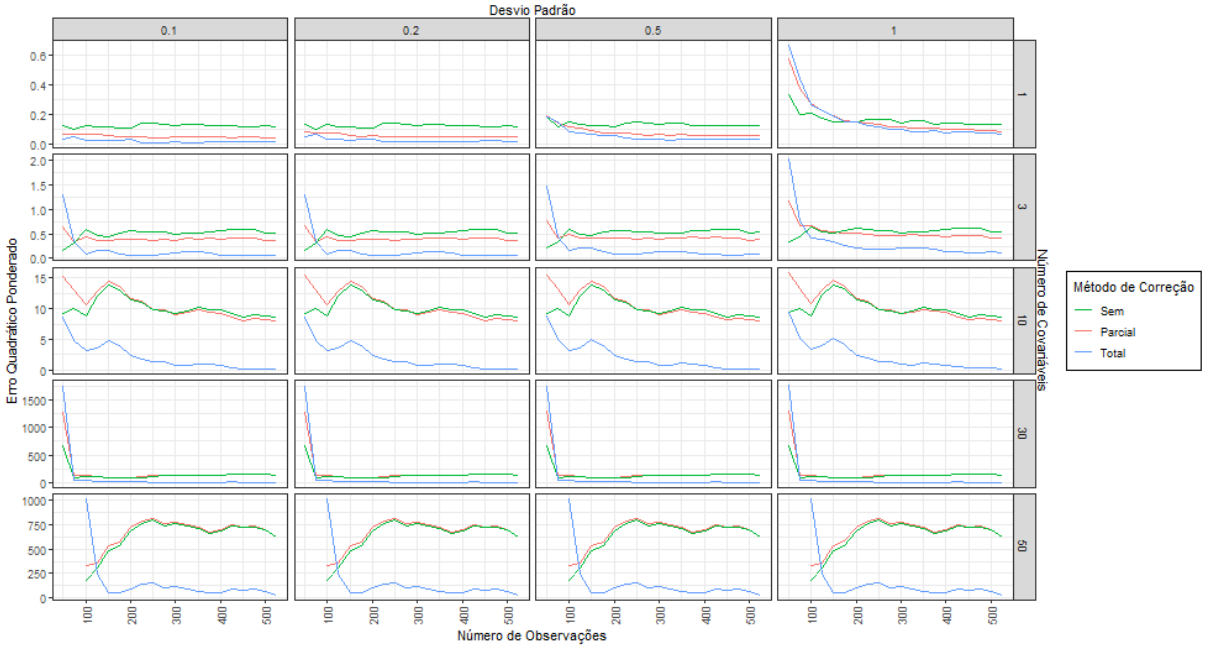

Figura 16 – EQP calculado para o caso de média distorção com L_k aleatório, para diferentes número de observações, número de variáveis categóricas e desvios padrões

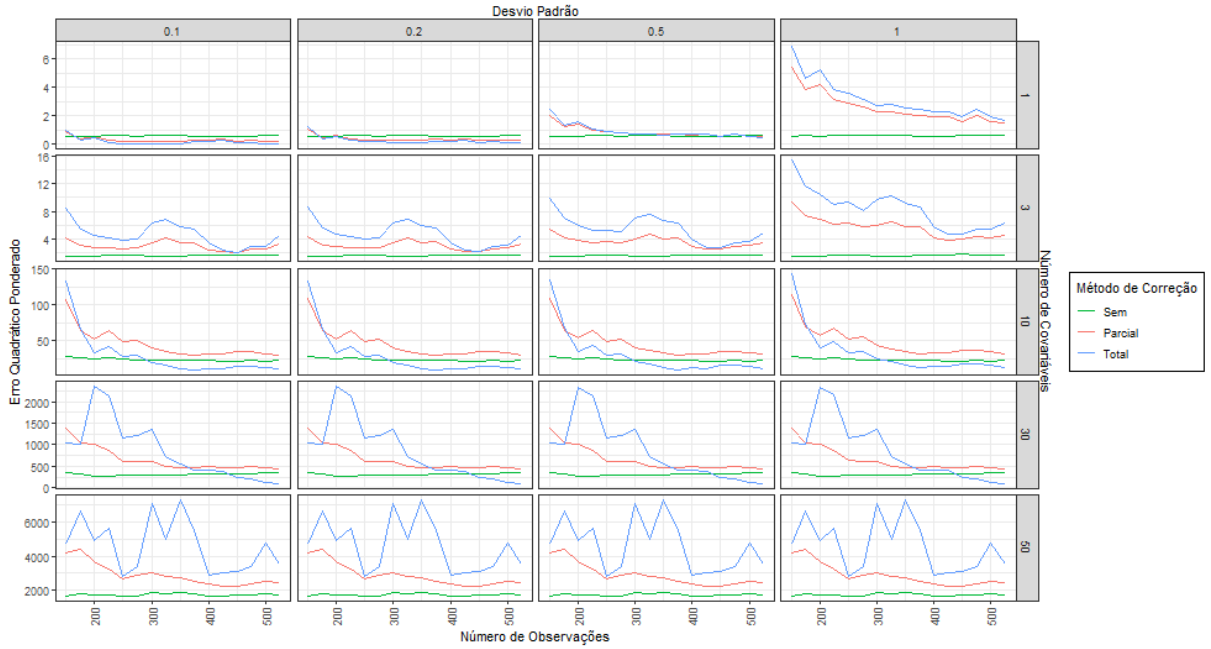

Figura 17 – EQP calculado para o caso de alta distorção com L_k aleatório, para diferentes número de observações, número de variáveis categóricas e desvios padrões

4 Considerações Finais

Neste trabalho propomos uma extensão do modelo de correção desenvolvido por [Buonaccorsi et al., 2005], onde expandimos suas ideias para um modelo multinomial. O problema descrito por esse autor reside no fato de que as covariáveis observadas $\mathbf{W}$ podem ter erros de classificação o que as torna diferente das covariáveis verdadeiras $\mathbf{X}$. Especificamente, nossa proposta de correção utiliza os estimadores de mínimos quadrados obtidos através da regressão em $\mathbf{W}$, as probabilidades marginais de $\mathbf{X}$ e as probabilidades condicionais de $\mathbf{W}$ dado $\mathbf{X}$. Dessa maneira é possível obter uma correção dos estimadores sem a necessidade de observar os valores verdadeiros.

Devido à singularidade da matriz de covariância de variáveis multinomiais, não foi possível corrigir o intercepto do modelo de regressão de forma conjunta com os demais coeficientes. Para contornar essa limitação, desenvolvemos um método no qual, baseado nos estimadores corrigidos, permite corrigir o intercepto. Os estudos de simulação demonstraram que a correção proposta para o intercepto é crucial na melhoria da estimação.

Algumas suposições foram necessárias para os cálculos desses estimadores corrigidos, especialmente, na independência entre as covariáveis e os indivíduos. Embora essas suposições sejam comumente utilizadas em modelos lineares, não são necessariamente razoáveis, especialmente no campo da genética onde essa correção poderia ser particularmente útil. Portanto é de suma importância que estudos futuros adaptem o método de correção, relaxando essas suposições.

Ademais, os estudos de simulação realizados, especialmente o caso de alta distorção, podem ser aprimorados expandindo o número de observações. Considerando que o método é assintótico, é possível que uma amostra grande o suficiente, apresente um melhor desempenho do estimador corrigido.

Referências

[Blackwell et al., 2017] Blackwell, M., Honaker, J., and King, G. (2017). A unified approach to measurement error and missing data: overview and applications. *Sociological Methods & Research*, 46(3):303–341. Citado na página 15.

[Brakenhoff et al., 2018] Brakenhoff, T. B., Mitroiu, M., Keogh, R. H., Moons, K. G., Groenwold, R. H., and van Smeden, M. (2018). Measurement error is often neglected in medical literature: a systematic review. *Journal of clinical epidemiology*, 98:89–97. Citado na página 15.

[Brookes, 1999] Brookes, A. J. (1999). The essence of snps. *Gene*, 234(2):177–186. Citado na página 17.

[Buonaccorsi, 2010] Buonaccorsi, J. P. (2010). *Measurement error: models, methods, and applications*. Chapman and Hall/CRC. Citado na página 15.

[Buonaccorsi et al., 2005] Buonaccorsi, J. P., Laake, P., and Veierød, M. B. (2005). On the effect of misclassification on bias of perfectly measured covariates in regression. *Biometrics*, 61(3):831–836. Citado 5 vezes nas páginas 15, 16, 19, 20 e 45.

[Chen et al., 2009] Chen, X., Hu, Y., and Lewbel, A. (2009). Nonparametric identification and estimation of nonclassical errors-in-variables models without additional information. *Statistica Sinica*, pages 949–968. Citado na página 15.

[Christopher and Kupper, 1995] Christopher, S. R. and Kupper, L. L. (1995). On the effects of predictor misclassification in multiple linear regression analysis. *Communications in Statistics-Theory and Methods*, 24(1):13–37. Citado na página 15.

[Dassonneville et al., 2011] Dassonneville, R., Brøndum, R. F., Druet, T., Fritz, S., Guillaume, F., Guldbrandtsen, B., Lund, M. S., Ducrocq, V., and Su, G. (2011). Effect of imputing markers from a low-density chip on the reliability of genomic breeding values in holstein populations. *Journal of Dairy Science*, 94(7):3679–3686. Citado na página 17.

[Davies and Mutton, 1975] Davies, B. and Mutton, B. (1975). The effect of errors in the independent variables in linear regression. *Biometrika*, 62(2):383–391. Citado na página 15.

[Göring and Terwilliger, 2000] Göring, H. H. and Terwilliger, J. D. (2000). Linkage analysis in the presence of errors ii: marker-locus genotyping errors modeled with hypercomplex

recombination fractions. *American journal of human genetics*, 66 3:1107–18. Citado na página 17.

[Hackett and Broadfoot, 2003] Hackett, C. C. and Broadfoot, L. (2003). Effects of genotyping errors, missing values and segregation distortion in molecular marker data on the construction of linkage maps. *Heredity*, 90:33–38. Citado na página 17.

[Keogh and Bartlett, 2021] Keogh, R. H. and Bartlett, J. W. (2021). Measurement error as a missing data problem. In *Handbook of measurement error models*, pages 429–452. Chapman and Hall/CRC. Citado na página 15.

[Keogh et al., 2020] Keogh, R. H., Shaw, P. A., Gustafson, P., Carroll, R. J., Deffner, V., Dodd, K. W., Küchenhoff, H., Tooze, J. A., Wallace, M. P., Kipnis, V., et al. (2020). Stratos guidance document on measurement error and misclassification of variables in observational epidemiology: part 1—basic theory and simple methods of adjustment. *Statistics in medicine*, 39(16):2197–2231. Citado na página 15.

[Klein and Rossin, 1999] Klein, B. and Rossin, D. (1999). Data quality in neural network models: effect of error rate and magnitude of error on predictive accuracy. *Omega*, 27(5):569–582. Citado na página 15.

[Kuha, 1997] Kuha, J. (1997). Estimation by data augmentation in regression models with continuous and discrete covariates measured with error. *Statistics in Medicine*, 16(2):189–201. Citado na página 15.

[Leitch and Leitch, 2008] Leitch, A. and Leitch, I. (2008). Genomic plasticity and the diversity of polyploid plants. *Science*, 320(5875):481–483. Citado na página 17.

[Mrode and Pocrnic, 2023] Mrode, R. A. and Pocrnic, I. (2023). *Linear models for the prediction of the genetic merit of animals*. CABI GB. Citado na página 17.

[Searle, 1997] Searle, S. R. (1997). *Linear models*, volume 65. John Wiley & Sons. Citado na página 15.

[Ward et al., 2021] Ward, A. M., Sweesi, M. E., Al-Mesilaty, L., Ahmed, A. A. M., Aswehli, A. A., Alkurdi, A. R. M., Elhafi, G. A., Hdud, I. M., and Benothman, M. A. (2021). Effects of molecular markers consequences on genotyping errors. Citado na página 17.

[Zucker and Spiegelman, 2008] Zucker, D. M. and Spiegelman, D. (2008). Corrected score estimation in the cox regression model with misclassified discrete covariates. *Statistical Models and Methods for Biomedical and Technical Systems*, pages 23–32. Citado na página 15.