

**Matheus Coutinho dos Santos**

**Analisando a obesidade na população  
residente nas capitais brasileiras via modelo  
de regressão logística**

Niterói - RJ, Brasil

13 de dezembro de 2023

**Matheus Coutinho dos Santos**

**Analisando a obesidade na população  
residente nas capitais brasileiras via  
modelo de regressão logística**

**Trabalho de Conclusão de Curso**

Monografia apresentada para obtenção do grau de Bacharel em  
Estatística pela Universidade Federal Fluminense.

Orientadora: Prof<sup>ª</sup> Dr<sup>ª</sup> Patrícia Lusié Velozo da Costa


Niterói - RJ, Brasil

13 de dezembro de 2023

**Matheus Coutinho dos Santos**


**Analisando a obesidade na população  
residente nas capitais brasileiras via modelo  
de regressão logística**

Monografia de Projeto Final de Graduação sob o título “*Analisando a obesidade na população residente nas capitais brasileiras via modelo de regressão logística*”, defendida por Matheus Coutinho dos Santos e aprovada em 13 de dezembro de 2023, na cidade de Niterói, no Estado do Rio de Janeiro, pela banca examinadora constituída pelos professores:

Documento assinado digitalmente  
 **PATRICIA LUSIE VELOZO DA COSTA**  
Data: 14/12/2023 11:45:17-0300  
Verifique em <https://validar.iti.gov.br>


---

**Prof<sup>a</sup>. Dr<sup>a</sup>. Patrícia Lusié Velozo da Costa**  
Departamento de Estatística – UFF

Documento assinado digitalmente  
 **MARIANA ALBI DE OLIVEIRA SOUZA**  
Data: 14/12/2023 11:25:03-0300  
Verifique em <https://validar.iti.gov.br>

---

**Prof<sup>a</sup>. Dr<sup>a</sup>. Mariana Albi de Oliveira Souza**  
Departamento de Estatística – UFF

Documento assinado digitalmente  
 **RAFAEL SANTOS ERBISTI**  
Data: 13/12/2023 19:48:55-0300  
Verifique em <https://validar.iti.gov.br>

---

**Prof. Dr. Rafael Santos Erbisti**  
Departamento de Estatística – UFF

Niterói, 13 de dezembro de 2023

Ficha catalográfica automática - SDC/BIME  
Gerada com informações fornecidas pelo autor

S237a Santos, Matheus Coutinho dos  
Analisando a obesidade na população residente nas capitais  
brasileiras via modelo de regressão logística / Matheus  
Coutinho dos Santos. - 2023.  
65 f.: il.

Orientador: Patrícia Lusié Velozo da Costa.  
Trabalho de Conclusão de Curso (graduação)-Universidade  
Federal Fluminense, Instituto de Matemática e Estatística,  
Niterói, 2023.

1. IMC. 2. Obesidade. 3. Regressão Logística Binomial. 4.  
VIGITEL. 5. Produção intelectual. I. Costa, Patrícia Lusié  
Velozo da, orientadora. II. Universidade Federal Fluminense.  
Instituto de Matemática e Estatística. III. Título.

CDD - XXX

# Resumo

A obesidade é uma condição de saúde caracterizada pelo acúmulo excessivo de gordura corporal e pode acarretar uma série de problemas de saúde como diabetes e doenças cardiovasculares. Além desses problemas de saúde, a obesidade também pode impactar a qualidade de vida geral, limitando a mobilidade, interferindo nas atividades diárias e reduzindo a expectativa de vida. No Brasil, a obesidade tem sido um problema crescente, afetando uma grande parcela da população. De acordo com o Ministério da Saúde, em 2021, quase seis a cada dez brasileiros se encontravam acima do peso, enquanto aproximadamente 22% da população se encontrava obesa. Desta maneira, este trabalho tem a intenção de avaliar potenciais características e hábitos relacionados a presença de obesidade em brasileiros adultos e idosos. Para isso, utilizou-se dados obtidos pela Vigilância de Fatores de Risco e Proteção para Doenças Crônicas por Inquérito Telefônico (VIGITEL), no ano de 2021 e no primeiro semestre de 2023, nas capitais de todos os estados brasileiros. Como variável resposta considerou-se a presença ou não de obesidade e o interesse estava em investigar a relação desta variável com covariáveis como estado civil, se é fumante ou não, entre outras. Os efeitos destas covariáveis são desconhecidos e foram estimados através de um modelo de regressão logística e usando 2 abordagens diferentes: Clássica e Bayesiana com diferentes distribuições *a priori*. Mesmo usando abordagens distintas, os ajustes apresentaram resultados semelhantes, sendo possível observar que indivíduos adultos de raça/cor preta ou com diagnóstico de depressão possuíram maior chance de estarem obesos. Já indivíduos que praticam atividade física, fumam ou estão solteiros possuíram menor chance em relação aos demais.

Palavras-chave: IMC. Obesidade. Regressão Logística Binomial. VIGITEL.

# Agradecimentos

Agradeço primeiramente a Deus, por todas as graças e oportunidades que me foram concedidas.

Agradeço aos meus pais, por todo apoio durante a realização deste trabalho.

Agradeço a minha orientadora pelos ensinamentos que me foram passados durante todo o desenvolvimento deste trabalho.

Por fim, agradeço a todos que contribuíram de alguma forma com meu desenvolvimento acadêmico.

# Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviações	p. 12
<b>1 Introdução</b>	p. 13
<b>2 Materiais e Métodos</b>	p. 17
2.1 Banco de dados . . . . .	p. 17
2.1.1 Amostragem . . . . .	p. 17
2.1.2 Ponderação . . . . .	p. 19
2.1.3 Estruturação dos dados . . . . .	p. 20
2.2 Modelo de Regressão Logística . . . . .	p. 21
2.2.1 Estimação pontual de parâmetros sob a abordagem clássica . . .	p. 23
2.2.1.1 Método da máxima Verossimilhança . . . . .	p. 23
2.2.1.2 Método dos mínimos quadrados ponderados iterativamente . . . . .	p. 25
2.2.2 Estimação de parâmetros sob a abordagem Bayesiana . . . . .	p. 25
2.2.3 Métodos de Monte Carlo via Cadeias de Markov . . . . .	p. 26
2.2.3.1 Algoritmo de Metropolis–Hastings . . . . .	p. 27
2.2.3.2 Amostrador de Gibbs . . . . .	p. 28
2.2.3.3 Monte Carlo Hamiltoniano . . . . .	p. 29
2.2.3.4 Convergência dos estimadores . . . . .	p. 29

2.2.4	Estimativa intervalar . . . . .	p. 30
2.2.5	Interpretação dos parâmetros . . . . .	p. 31
2.2.6	Testes de hipóteses clássicos . . . . .	p. 32
2.2.6.1	Teste de Wald . . . . .	p. 32
2.2.6.2	Teste da razão da verossimilhança . . . . .	p. 32
2.2.7	CrITÉRIOS de comparação entre modelos . . . . .	p. 33
2.2.7.1	CrITÉRIO de Informação de Akaike . . . . .	p. 33
2.2.7.2	CrITÉRIO de Informação de Desvio . . . . .	p. 33
<b>3</b>	<b>Análise de Resultados</b>	p. 35
3.1	Análise Descritiva . . . . .	p. 35
3.2	Ajuste do Modelo de Regressão Logística . . . . .	p. 38
3.2.1	Ajuste dos dados simulados . . . . .	p. 38
3.2.2	Modelo de Regressão Logística Clássico . . . . .	p. 43
3.2.3	Modelo de Regressão Logística Bayesiano . . . . .	p. 45
<b>4</b>	<b>Conclusões</b>	p. 62
	<b>Referências</b>	p. 64



# Lista de Figuras

1	Traço das Cadeias de Markov para o ajuste a <i>posteriori</i> dos dados simulados dos parâmetros com distribuição a <i>priori</i> Normal com média igual a 0 e variância igual a 0,01 . . . . .	p. 40
2	Traço das Cadeias de Markov para o ajuste a <i>posteriori</i> dos dados simulados dos parâmetros com distribuição a <i>priori</i> Normal com média igual a 0 e variância igual a 1 . . . . .	p. 41
3	Traço das Cadeias de Markov para o ajuste a <i>posteriori</i> dos dados simulados dos parâmetros com distribuição a <i>priori</i> Normal com média igual a 0 e variância igual a 1000 . . . . .	p. 41
4	Função de autocorrelação das amostras geradas para o ajuste a <i>posteriori</i> dos dados simulados dos parâmetros com distribuição a <i>priori</i> Normal com média 0 e variância 0,01 . . . . .	p. 42
5	Função de autocorrelação das amostras geradas para o ajuste a <i>posteriori</i> dos dados simulados dos parâmetros com distribuição a <i>priori</i> Normal com média 0 e variância 1 . . . . .	p. 42
6	Função de autocorrelação das amostras geradas para o ajuste a <i>posteriori</i> dos dados simulados dos parâmetros com distribuição a <i>priori</i> Normal com média 0 e variância 1000 . . . . .	p. 43
7	Traço das Cadeias de Markov para o ajuste a <i>posteriori</i> dos parâmetros com distribuição a <i>priori</i> Normal com média 0 e variância 0,01 . . . . .	p. 49
8	Traço das Cadeias de Markov para o ajuste a <i>posteriori</i> dos parâmetros com distribuição a <i>priori</i> Normal com média igual a 0 e variância igual a 1	p. 50
9	Traço das Cadeias de Markov para o ajuste a <i>posteriori</i> dos parâmetros com distribuição a <i>priori</i> Normal com média igual a 0 e variância igual a 1000 . . . . .	p. 51

10	Função de autocorrelação das amostras geradas para o ajuste a <i>posteriori</i> dos parâmetros com distribuição a <i>priori</i> Normal com média 0 e variância 0,01 . . . . .	p. 52
11	Função de autocorrelação das amostras geradas para o ajuste a <i>posteriori</i> dos parâmetros com distribuição a <i>priori</i> Normal com média 0 e variância 1 . . . . .	p. 53
12	Função de autocorrelação das amostras geradas para o ajuste a <i>posteriori</i> dos parâmetros com distribuição a <i>priori</i> Normal com média 0 e variância 1000 . . . . .	p. 54
13	Histograma dos parâmetros estimados para o ajuste a <i>posteriori</i> com <i>priori</i> Normal com média 0 e variância 0,01 . . . . .	p. 55
14	Histograma dos parâmetros estimados para o ajuste a <i>posteriori</i> com <i>priori</i> Normal com média 0 e variância 1 . . . . .	p. 56
15	Histograma dos parâmetros estimados para o ajuste a <i>posteriori</i> com <i>priori</i> Normal com média 0 e variância 1000 . . . . .	p. 57

# Lista de Tabelas

1	Volumetria dos dados . . . . .	p. 19
2	Distribuição dos dados ponderados do VIGITEL e proporção de obesos de acordo com o perfil sociodemográfico durante o ano de 2021 e primeiro semestre de 2023 . . . . .	p. 36
3	Distribuição dos dados ponderados do VIGITEL e proporção de obesos de acordo com potenciais hábitos e fatores relacionados a obesidade durante o ano de 2021 e primeiro semestre de 2023 . . . . .	p. 37
4	Valores reais, valores estimados, intervalo de confiança e p-valor da estatística de Wald para os parâmetros considerando o ajuste do modelo de regressão clássica para os dados simulados . . . . .	p. 38
5	Valores reais, valores estimados e intervalo de credibilidade para os parâmetros considerando o ajuste do modelo de regressão Bayesiana para os dados simulados com distribuição a priori $\beta_j \sim N(0; 0,01)$ . . . . .	p. 39
6	Valores reais, valores estimados e intervalo de credibilidade para os parâmetros considerando o ajuste do modelo de regressão Bayesiana para os dados simulados com distribuição a priori $\beta_j \sim N(0; 1)$ . . . . .	p. 40
7	Valores reais, valores estimados e intervalo de credibilidade para os parâmetros considerando o ajuste do modelo de regressão Bayesiana para os dados simulados com distribuição a priori $\beta_j \sim N(0; 1000)$ . . . . .	p. 40
8	Valores estimados, estimativa intervalar e p-valor da estatística para os parâmetros considerados no ajuste do modelo de regressão clássica . . .	p. 44
9	Razão de chances e estimativa intervalar da razão de chances para os parâmetros considerados no ajuste do modelo de regressão clássica . . .	p. 44
10	Estimação pontual, estimativa intervalar e valor da estatística $\hat{R}$ para os parâmetros com distribuição a priori Normal com média 0 e variância 0,001 . . . . .	p. 46

- 11 Estimação pontual, estimativa intervalar e valor da estatística  $\hat{R}$  para os parâmetros com distribuição a *priori* Normal com média 0 e variância 0,01 p. 47
- 12 Estimação pontual, estimativa intervalar e valor da estatística  $\hat{R}$  para os parâmetros com distribuição a *priori* Normal com média 0 e variância 1 p. 48
- 13 Razão de Chances e estimativa intervalar da razão de chances para os parâmetros com distribuição a *priori* Normal com média 0 e variância 0,01 p. 59
- 14 Razão de Chances e estimativa intervalar da razão de chances para os parâmetros com distribuição a *priori* Normal com média 0 e variância 1 p. 60
- 15 Razão de Chances e estimativa intervalar da razão de chances para os parâmetros com distribuição a *priori* Normal com média 0 e variância 1000 p. 61

# Lista de Abreviações

**AIC** Critério de informação de Akaike

**DIC** Critério de Informação de Desvio

**IMC** Índice de Massa Corporal

**IRLS** Mínimos Quadrados Ponderados Iterativamente

**MCMC** Monte Carlo via Cadeias de Markov

**MLG** Modelo Linear Generalizado

**VIGITEL** Vigilância de Fatores de Risco e Proteção para Doenças Crônicas por Inquérito Telefônico

# 1 Introdução

De acordo com World Health Organization (2021), a obesidade é uma doença crônica, caracterizada pelo acúmulo excessivo de gordura corporal e estando relacionada a potenciais doenças, como diabetes, hipertensão, dificuldades respiratórias e doenças do coração, além de ser responsável pelo agravamento de outras doenças, como a Covid-19. Essas condições podem levar a complicações graves e reduzir a qualidade de vida dos indivíduos afetados.

A obesidade representa uma questão importante para o sistema de saúde e para a economia como um todo. Os gastos com cuidados médicos relacionados à obesidade, incluindo tratamentos de doenças associadas, consultas médicas, medicamentos e hospitalizações, são elevados. Rezende et al. (2022) estimaram que o gasto anual direto com doenças relacionadas ao excesso de peso e obesidade no Brasil durante o ano de 2019 foi de 1,5 bilhão de reais, representando 22% do gasto com doenças crônicas não transmissíveis. Além disso, a obesidade pode resultar em perda de produtividade no trabalho, além de ausências prolongadas.

O diagnóstico para a doença possui viés clínico, apesar de ser avaliado em grande escala por meio do Índice de Massa Corporal (IMC) de cada indivíduo adulto. World Health Organization (2021) defende que valores maiores ou iguais a  $25 \text{ kg/m}^2$  e  $30 \text{ kg/m}^2$  indicam potencial sobrepeso e obesidade, respectivamente.

Segundo World Health Organization (2022), um pouco mais de 1 milhão de pessoas no mundo estão obesas, sendo aproximadamente 63% adultos. Também há uma estimativa de que, em 2035, 1,9 bilhão de pessoas se tornarão obesas, além de um aumento esperado de 60% para a obesidade em indivíduos adultos. De acordo com Ministério da Saúde (2022), aproximadamente 22,4% da população brasileira se encontra como obesa e existe um aumento médio de 0,66% por ano na proporção de obesos durante o período entre 2006 e 2021, com percentual inicial igual a 11,8% de obesos.

A obesidade é uma doença multicausal, estando relacionada principalmente à hábitos físicos e alimentares. Entre suas principais causas, estão dietas não saudáveis, como o consumo excessivo de alimentos ricos em açúcar e gordura, prática deficitária de exercícios físicos, além de horas inadequadas de sono. A propaganda possui grande influência no crescente aumento de indivíduos obesos. Em Thomas et al. (2018), realizou-se um estudo com residentes no Reino Unido com idade entre 11 e 19 anos mostrando existir relação entre o tempo de horas gasto assistindo televisão, e conseqüentemente propagandas de alimentos considerados prejudiciais à saúde, com os fatores obesidade e sobrepeso. Em média, participantes obesos relataram passar 6 horas semanais a mais assistindo televisão e serviços de *streaming* em relação aos participantes não obesos.

O aumento do consumo de substâncias alcoólicas é outro fator preocupante no combate à obesidade. A Pesquisa Nacional da Saúde realizada nos anos de 2013 e 2019 evidenciou um aumento semanal de aproximadamente 2% no consumo de bebidas alcoólicas. Este crescimento fica mais evidente ao considerar apenas mulheres, onde o crescimento foi de aproximadamente 4%. Shelton e Knott (2014) observou 8.864 britânicos e evidenciou a relação da obesidade com o consumo de bebidas alcoólicas. Neste estudo, foi observado que aproximadamente 27% da ingestão calórica média de indivíduos homens e 19% de mulheres eram calorias consumidas por meio de bebidas alcoólicas. Em Park et al. (2022), observou-se aproximadamente 27 milhões de adultos sul-coreanos, concluindo que indivíduos que consomem de 7,1 a 14 gramas de álcool por dia possuem maior probabilidade de serem obesos, ao se comparar com indivíduos que não consomem álcool. O estudo também mostra que quanto maior o consumo de álcool, maior a tendência para obesidade.

Em Ferreira, Szwarcwald e Damacena (2019), modelou-se a obesidade na população brasileira através de modelos de regressão logística aplicados em uma amostra de 59.402 indivíduos coletada a partir da Pesquisa Nacional de Saúde do ano de 2013, com interesse em avaliar fatores relacionados à obesidade. Como resultado, foi observado que indivíduos homens com idade entre 40 e 49 anos e mulheres com idade entre 50 e 59 anos possuem maior tendência a serem obesos enquanto indivíduos de ambos os sexos, com idade entre 18 e 29 anos possuem menor tendência a serem obesos. Também foi observada relação entre a obesidade e o grau de escolaridade, sendo indivíduos com menor grau de escolaridade mais propensos a serem obesos.

Silva et al. (2021) avaliaram a presença de excesso de peso e obesidade durante os anos de 2006 a 2019 na população adulta brasileira por meio de séries temporais, além de terem utilizado modelos de regressão de Prais-Winsten para controlar a autocorrelação

dos resíduos entre os períodos. Foi observado um aumento de indivíduos obesos durante os anos, passando de 11,8% em 2006 para 20,3% no último ano considerado, representando um crescimento de 3,8% ao ano. Também foi verificado que indivíduos maiores de idade com até 24 anos possuem menor prevalência a serem obesos. De maneira geral, foi identificado um aumento no percentual de obesos em todos os grupos analisados, sendo um aumento médio igual ao se considerar os estratos dos indivíduos masculinos. Já para as mulheres, o indicador de obesidade cresceu entre aquelas com maior escolaridade e idade superior a 45 anos.

Silva e Padilha (2023) consideraram uma amostra de indivíduos residentes na capital do Maranhão para avaliar o perfil nutricional e os fatores de risco relacionados a obesidade. A associação entre as variáveis categóricas foi avaliada por meio de um teste Qui-quadrado, além do ajuste de um modelo de regressão de Poisson com variância robusta. Em relação aos idosos, os adultos de ambos os sexos se mostraram com maior disposição a doença, sendo classificados como obesos, aproximadamente 20,03% dos homens e 14,91% das mulheres adultas. Também foi identificada uma relação entre o excesso de peso na idade adulta, 0 a 8 anos de escolaridade, consumo abusivo de álcool e ter um companheiro. Já entre os idosos, só houve associação entre o excesso de peso e o hábito de assistir televisão por mais de 3 horas diárias. Por fim, foi observada uma maior prevalência de obesidade em homens adultos e mulheres idosas.

Utilizando modelos de regressão Multinomial e de Poisson, Streb et al. (2020) se propuseram a avaliar a relação entre obesidade e fatores de risco, como inatividade física, tempo sentado em excesso, consumo de doces e carnes vermelhas com gordura, ou carnes de frango com pele entre pessoas com idade de 18 a 49 anos. Como resultado, foi observado que o fator de risco mais relacionado com a obesidade entre as mulheres foi a inatividade física. Já entre os homens, o consumo de carne vermelha com gordura ou carne de frango com pele se mostrou ligado à presença da doença.

Os fatores relacionados com a obesidade podem ser divididos em mutáveis e imutáveis. Os estudos citados anteriormente, obtiveram como exemplos de fatores mutáveis a alimentação não saudável e ausência ou prática irregular de exercícios físicos. Neste sentido, o indivíduo obeso pode buscar por mudanças para alterar a sua situação. Além disso, o governo pode desenvolver e implementar políticas eficazes para promover uma alimentação saudável, incentivar a prática regular de exercícios físicos e criar ambientes propícios para a adoção de hábitos saudáveis. Isso pode envolver a regulação da publicidade de alimentos não saudáveis, o estabelecimento de diretrizes nutricionais, a



promoção da atividade física nas escolas e a criação de espaços públicos adequados para exercícios. A identificação de fatores imutáveis também é muito importante uma vez que auxilia no direcionamento de políticas públicas para a população com tendência a doença.

Este trabalho teve como objetivo geral verificar quais fatores e hábitos tem relação com a obesidade para a população brasileira. Para isso, usou-se a amostra obtida pelo VIGITEL no ano de 2021 e no primeiro semestre de 2023. Não incluiu-se o ano de 2022 por não estar disponível. Inicialmente realizou-se uma análise descritiva dos dados. Posteriormente, ajustou-se modelos de regressão logística por meio da abordagem clássica e Bayesiana. Para o ajuste, realizou-se inicialmente um estudo simulado para avaliar a eficiência da modelagem e posteriormente aplicou-se o modelo proposto aos dados de interesse.

Os capítulos desse trabalho são divididos, de forma que o Capítulo 2 apresenta informações sobre a base de dados considerada para o estudo, descrevendo todo processo de amostragem, ponderação e estruturação, além de conter uma revisão em relação a metodologia aplicada. O Capítulo 3 contém todo o processo de análise e ajuste do modelo de regressão logística com abordagem clássica e Bayesiana, apresentando os resultados e conclusões obtidas por estes modelos. Por fim, o Capítulo 4 apresenta a conclusão do trabalho, se propondo a consolidar todo resultado obtido pelas diferentes abordagens estatísticas.

## 2 Materiais e Métodos

O presente Capítulo está dividido como descrito a seguir. A Seção 2.1 contém informações sobre o banco de dados utilizado neste trabalho, explicando todo processo de amostragem, ponderação e estruturação. A Seção 2.2 introduz o conceito de Modelo Linear Generalizado. São apontados métodos clássicos e Bayesianos de estimação e avaliação dos parâmetros estimados, além das interpretações dos parâmetros gerados.

### 2.1 Banco de dados

O VIGITEL compõe o sistema de vigilância de fatores de risco para doenças crônicas não transmissíveis, de responsabilidade do Ministério da Saúde, em conjunto a inquéritos domiciliares ou voltados para a população escolar. Este inquérito busca identificar a relação da população brasileira com o alcoolismo, tabagismo, prática regular de atividades físicas, consumo alimentar e potenciais diagnósticos de doenças crônicas, além de coletar dados que caracterizam o indivíduo, como idade, sexo, cor e indicadores antropométricos, como peso e altura. De acordo com o Ministério da Saúde (2023), desde sua primeira edição, foram entrevistados mais de 790 mil brasileiros, sendo aproximadamente 38% homens e 62% mulheres.

Os dados utilizados para este trabalho são de origem do sistema de VIGITEL e foram coletados em 2021 e no primeiro semestre de 2023, nas capitais de todos os estados brasileiros, além do Distrito Federal. As Subseções 2.1.1 e 2.1.2 explicam como os dados foram coletados e a Subseção 2.1.3 detalha como os dados foram utilizados neste trabalho.

#### 2.1.1 Amostragem

A coleta de amostras para realização do VIGITEL é feita por meio da telefonia fixa, tendo sido considerada a telefonia móvel a partir de 2023, com objetivo de obter dados referentes a população com idade igual ou superior a 18 anos completos, residentes em

domicílios com ao menos uma linha telefônica ativa. O processo de amostragem passou por duas variações durante os 16 anos de pesquisa. Para as edições realizadas durante os anos de 2006 a 2011, foi considerado um tamanho amostral maior ou igual a 2.000 entrevistados por cidade, possibilitando assim, uma estimativa com nível de confiança de 95% e erro máximo de dois pontos percentuais, com exceções em estimativas específicas, onde foram assumidas proporções semelhantes de acordo com o sexo do indivíduo, que originaram erros de até três pontos percentuais. As edições referentes aos anos de 2012 a 2019 passaram a considerar amostras de menor tamanho, com 1.000 a 1.500 indivíduos entrevistados por cidade. Esta exceção foi aceita apenas em regiões com menos de 50 mil domicílios com telefonia fixa cadastrada e uma abrangência de telefones fixos inferior a 40% do total de residências, o que originou estimativas com erro máximo de três pontos percentuais, com exceções das estimativas específicas por sexo, onde foram considerados erros máximos de até 4 pontos percentuais. Já nas duas últimas edições, referente aos anos de 2020 e 2021, o tamanho amostral foi reduzido para um mínimo de 1.000 indivíduos por cidade, gerando uma amostra que possibilite estimar a frequência dos fatores de risco e proteção na população adulta ao nível de confiança de 95% e erro máximo de quatro pontos percentuais, considerando erros de até 5 pontos percentuais em estimativas específicas que consideram proporções semelhantes de homens e mulheres na amostra. Esta redução gradativa resultou em uma leve diminuição da volumetria dos dados durante os anos de 2006 a 2019, porém gerou em uma redução de aproximadamente 50% nos dados durante as pesquisas realizadas nos anos de 2020 e 2021. A Tabela 1 contém a quantidade de entrevistados em cada edição do VIGITEL realizada. Devido a fatores externos, o VIGITEL não ocorreu durante o ano de 2022, tendo seu retorno durante o ano de 2023, de forma semestral.

A primeira etapa de amostragem se origina com o sorteio de pelo menos 10 mil linhas telefônicas para cada cidade de interesse, oriundas a partir do cadastro eletrônico de linhas residenciais fixas de empresas telefônicas. A partir da VIGITEL realizada no ano de 2021, as linhas telefônicas passaram a ser obtidas diretamente do cadastro eletrônico de linhas residenciais fixas da ANATEL. A primeira etapa tem como objetivo conhecer a quantidade de residentes com idade igual ou superior a 18 anos para composição da amostra. Assim, no primeiro contato é perguntado o sexo e idade de todos os residentes. As linhas coletadas passam por outro sorteio aleatório, onde são agrupadas e divididas em réplicas com tamanho de linhas igual a 200. Na segunda etapa, é realizada uma validação para verificar quais linhas telefônicas são elegíveis para o sistema VIGITEL. Telefones correspondentes a empresas, números que deixaram de existir ou se encontram

Tabela 1: Volumetria dos dados

Edição VIGITEL	Volumetria
2006	54.369
2007	54.251
2008	54.353
2009	54.367
2010	54.339
2011	54.144
2012	45.448
2013	52.929
2014	40.853
2015	54.174
2016	53.210
2017	53.034
2018	52.395
2019	52.443
2020	27.077
2021	27.093
2023 (Primeiro semestre)	21.690

como fora de serviço e linhas que não responderam a seis tentativas realizadas durante dias e horários variados, incluindo finais de semana e horários não comerciais são removidas da amostra. Após a identificação de linhas elegíveis, é realizado um novo sorteio, para definir qual residente com idade superior ou igual a 18 anos completos será o indivíduo a ser considerado para determinada residência. No relatório realizado no ano de 2021, por exemplo, foram realizadas 319.400 ligações telefônicas, que foram agrupadas em 1.597 réplicas com tamanho igual a 200. Destas ligações, 44.457 foram classificadas como elegíveis, que resultaram em 27.093 entrevistas completas, indicando uma taxa de sucesso de aproximadamente 61%.

### 2.1.2 Ponderação

Para obter estimativas confiáveis, o VIGITEL aplica peso aos indivíduos selecionados, uma vez que parte da população não tem a mesma probabilidade de ser selecionada para o estudo. Desde 2012, o peso pós-estratificação é calculado pelo método Rake, que depende do tamanho populacional em cada cidade, levando em conta a faixa etária, o nível de instrução e o sexo. A distribuição de cada variável sociodemográfica estimada para cada cidade é obtida a partir de projeções que consideram a distribuição da variável nos Censos Demográficos de 2000 e 2010 e sua variação anual média no período intercensitário. Mais detalhes sobre a ponderação e o método Rake podem ser vistos em Bernal et al. (2017).

### 2.1.3 Estruturação dos dados

Neste trabalho utilizou-se os dados obtidos pelo VIGITEL<sup>1</sup> em 2021 e no primeiro semestre de 2023, resultando em um banco de dados com informações de 48.783 indivíduos. Haviam mais de 200 colunas nos dados obtidos contendo informações de cada indivíduo como cidade, idade, sexo, estado conjugal, grau de escolaridade, escolaridade em anos, se dirige ou não, peso, altura, entre outros. Por questões computacionais e de tempo, o banco de dados utilizado neste trabalho considerou parte dessas informações.

A variável resposta considerada foi o indivíduo estar ou não obeso, resultando em uma variável binária. Esta variável não existia na base obtida, mas foi possível a sua construção através do peso e da altura, que existiam na base. Para isso, calculou-se então o IMC de cada indivíduo, utilizando o peso em quilos do indivíduo dividido pelo quadrado da altura em metros. Em seguida, classificou-se em obeso, todo indivíduo com IMC igual ou superior a 30 kg/m<sup>2</sup>. Foram excluídas da análise, participantes que relataram estar grávidas, uma vez que o indicador não considera o aumento de peso devido ao desenvolvimento do bebê. As variáveis peso e altura passaram por um processo de imputação de dados faltantes por parte do VIGITEL, utilizando a técnica *hot deck*. Nesta imputação, foram selecionadas as variáveis sexo, idade, raça/cor e escolaridade de cada indivíduo, tornando possível a criação de modelo que permite agrupar informações de acordo com as variáveis selecionadas.

Foram selecionadas variáveis referentes a características pessoais e sociais de cada indivíduo, como sexo, faixa etária, raça/cor, cidade de residência, estado conjugal e nível de escolaridade. Além disso, foram selecionadas variáveis e indicadores que representam hábitos e ações potencialmente relacionadas ao indicador obesidade.

O consumo de drogas lícitas com potencial relação a obesidade foi representado pelo indicador de indivíduos fumantes, que considera indivíduos que fumam, mesmo que esporadicamente. Já o consumo de bebidas alcoólicas foi categorizado ao selecionar indivíduos que alegaram costume de ingerir este tipo de substância. Outros indicadores com potencial relação com o IMC foram considerados, como o consumo regular de frutas, hortaliças e refrigerante, sendo classificado como regular, a ingestão desses alimentos em uma frequência superior ou igual a cinco vezes semanais. O indicador representativo do hábito de utilizar aparelho celular e computador durante o tempo livre também foi avaliado.

---

<sup>1</sup>Dados obtidos em <https://svs.aids.gov.br/download/Vigitel/>.

Após excluir os indivíduos que tinham algum dado faltante, a base de dados final ficou com 29.571 indivíduos e 17 colunas sendo uma com a variável resposta (se o indivíduo é ou não obeso), outra com o peso pós-estratificação e 15 variáveis explicativas (faixa etária, sexo, álcool, se faz exercício físico, raça/cor, fumante ou não, se usa computador/celular no tempo livre, anos de estudo, estado civil, se tem ou não companheiro, se teve ou não COVID-19, se tem ou não plano de saúde, se tem ou não depressão, se consome ou não legumes/verduras, se consome ou não frutas, se consome ou não refrigerante).

## 2.2 Modelo de Regressão Logística

Proposto por Nelder e Wedderburn (1972), o Modelo Linear Generalizado Modelo Linear Generalizado (MLG) se trata de uma extensão dos Modelos de Regressão Lineares, possibilitando trabalhar com uma variável resposta com distribuição diferente da normal. Agresti (2015) classifica as três componentes do modelo, sendo elas:

- Componente Aleatório do modelo, denotado por  $Y_i$ , que consiste na variável resposta com distribuição de probabilidade pertencente à família exponencial para a  $i$ -ésima unidade amostral. Sendo assim, a distribuição de  $\mathbf{Y} = (Y_1, \dots, Y_n)^T$ , quando condicionada a um vetor paramétrico  $\boldsymbol{\pi} = (\pi_1, \dots, \pi_n)^T$ , pode ser escrita da seguinte forma

$$f(\mathbf{Y}|\boldsymbol{\pi}) = a(\mathbf{Y}) \exp \left[ \sum_{i=1}^n Y_i b(\pi_i) + \sum_{j=1}^n c(\pi_i) \right] \quad (2.1)$$

sendo  $b(\cdot)$ ,  $c(\cdot)$  e  $a(\cdot)$  funções reais.

- Preditor linear, denotado por

$$\eta_i = \sum_{j=0}^p x_{ij} \beta_j, \quad i = 1, \dots, n,$$

onde  $x_{ij}$  representa o valor da  $j$ -ésima covariável (também chamada de variável explicativa) observada em relação a  $i$ -ésima unidade amostral, e  $\beta_j$  um parâmetro desconhecido, responsáveis pelo efeito da  $j$ -ésima covariável,  $j = 0, 1, \dots, p$ . Em forma matricial, o preditor pode ser reescrito como

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta},$$

sendo  $\boldsymbol{\eta} = (\eta_1, \dots, \eta_n)^T$ ,  $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)^T$  e  $\mathbf{X}$  uma matriz de dimensão  $n \times (p+1)$ , chamada de matriz desenho ou matriz modelo, formada pelos elementos  $x_{ij}$ ,  $i = 1, \dots, n$  e  $j = 0, \dots, p$ .

- Função de ligação, com objetivo de relacionar a esperança da componente aleatório ao preditor linear. Seja  $\mu_i = E(Y_i)$ , o modelo linear generalizado faz uso da função de ligação monótona e diferenciável  $g(\cdot)$ , de forma que o modelo possa ser reescrito como

$$g(\mu_i) = \eta_i, \quad i = 1, \dots, n.$$

Pode não existir uma única função de ligação possível. Um critério de escolha desta função pode ser usando a família exponencial canônica e, nesse caso, a função de ligação pode ser escolhida da seguinte forma

$$g(\mu_i) = b(\pi_i). \quad (2.2)$$

Para mais detalhes, veja Dobson e Barnett (2008).

O **modelo de Regressão Logística** consiste em um caso particular de MLG, onde a componente aleatória possui uma classificação binária, geralmente utilizado em estudos de caso-controle. De maneira geral, a componente aleatória pode ser denotada como  $Y_i \sim Bern(\pi_i)$ , ou seja,

$$Y_i = 1; \text{ em caso de sucesso}$$

$$Y_i = 0; \text{ em caso de fracasso}$$

de forma que  $P(Y_i = 1) = \pi_i$  e  $P(Y_i = 0) = 1 - \pi_i$ , para a  $i$ -ésima unidade amostral. Sendo assim, considerando uma amostra  $\mathbf{y} = (y_1, \dots, y_n)^T$ , tem-se que a função de verossimilhança é dada pela seguinte forma

$$l(\boldsymbol{\pi}; \mathbf{y}) = \prod_{i=1}^n \pi_i^{y_i} (1 - \pi_i)^{1-y_i} I(y_i \in \{0, 1\}). \quad (2.3)$$

Segundo Pessoa e Silva (1998), considerando pesos amostrais  $\boldsymbol{\omega} = (\omega_1, \dots, \omega_n)^T$ , a função de verossimilhança do modelo de regressão logística passa a ser dada pela seguinte forma

$$\begin{aligned} l(\boldsymbol{\pi}; \mathbf{y}) &= \prod_{i=1}^n \pi_i^{\omega_i y_i} (1 - \pi_i)^{\omega_i (1-y_i)} I(y_i \in \{0, 1\}) \\ &= \prod_{i=1}^n I(y_i \in \{0, 1\}) \exp \left\{ \sum_{i=1}^n \omega_i y_i \ln \left( \frac{\pi_i}{1 - \pi_i} \right) + \sum_{i=1}^n \omega_i \ln(1 - \pi_i) \right\}. \end{aligned} \quad (2.4)$$

Note que a função acima está escrita na forma da família exponencial, dada pela Equação em 2.1, sendo  $b(\pi_i) = \ln\left(\frac{\pi_i}{1-\pi_i}\right)$ ,  $c(\pi_i) = \omega_i \ln(1 - \pi_i)$  e  $a(\mathbf{y}) = \prod_{i=1}^n I(y_i \in \{0, 1\})$ .

A função de ligação canônica utilizada para o modelo é a função logito, possibilitando que o modelo seja escrito da forma

$$\ln\left(\frac{\pi_i}{1 - \pi_i}\right) = \eta_i = \mathbf{X}_i^T \boldsymbol{\beta}, \quad (2.5)$$

com  $\mathbf{X}_i^T = (1, x_{i1}, x_{i2}, \dots, x_{ip})$ , quando consideramos que há intercepto, ou seja, um efeito comum a todas as unidades amostrais. Sendo assim, a probabilidade de sucesso pode ser reescrita em função das covariáveis da seguinte forma

$$\pi_i = \frac{\exp(\mathbf{X}_i^T \boldsymbol{\beta})}{1 + \exp(\mathbf{X}_i^T \boldsymbol{\beta})}. \quad (2.6)$$

Note que no modelo de regressão logística, o vetor paramétrico  $\boldsymbol{\pi} = (\pi_1, \dots, \pi_n)^T$  é desconhecido e é uma função de  $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)$ . Logo, basta estimar  $\boldsymbol{\beta}$ , para ter  $\boldsymbol{\pi}$ .

## 2.2.1 Estimação pontual de parâmetros sob a abordagem clássica

Os métodos mais utilizados sob a abordagem clássica para estimar os parâmetros desconhecidos são: o de máxima verossimilhança e o de mínimos quadrados ponderados iterativamente.

### 2.2.1.1 Método da máxima Verossimilhança

Sob o método da máxima verossimilhança, o estimador pontual é obtido maximizando a função de verossimilhança. Maximizar o logaritmo da função de verossimilhança é equivalente a maximizar a própria função de verossimilhança, já que o logaritmo é uma função monótona crescente. Portanto, o ponto em que a função de verossimilhança atinge seu máximo é o mesmo ponto em que o logaritmo da função de verossimilhança atinge seu máximo.

A função de log-verossimilhança é dada por

$$\begin{aligned} L(\boldsymbol{\pi}; \mathbf{y}) &= \ln(l(\boldsymbol{\pi}; \mathbf{y})) \\ &= \sum_{i=1}^n \omega_i y_i \ln(\pi_i) + \sum_{i=1}^n \omega_i (1 - y_i) \ln(1 - \pi_i) \end{aligned} \quad (2.7)$$



que, em função dos parâmetros  $\boldsymbol{\beta}$ , é dada por

$$\begin{aligned}
L(\boldsymbol{\beta}; \mathbf{y}) &= \sum_{i=1}^n \omega_i y_i \ln \left( \frac{\exp(\mathbf{X}_i^T \boldsymbol{\beta})}{1 + \exp(\mathbf{X}_i^T \boldsymbol{\beta})} \right) + \sum_{i=1}^n \omega_i (1 - y_i) \ln \left( 1 - \frac{\exp(\mathbf{X}_i^T \boldsymbol{\beta})}{1 + \exp(\mathbf{X}_i^T \boldsymbol{\beta})} \right) \\
&= \sum_{i=1}^n \omega_i y_i \ln \left( \frac{\exp(\mathbf{X}_i^T \boldsymbol{\beta})}{1 + \exp(\mathbf{X}_i^T \boldsymbol{\beta})} \right) + \sum_{i=1}^n \omega_i (1 - y_i) \ln \left( \frac{1}{1 + \exp(\mathbf{X}_i^T \boldsymbol{\beta})} \right) \\
&= \sum_{i=1}^n \omega_i y_i \ln \left( \frac{\exp(\mathbf{X}_i^T \boldsymbol{\beta})}{1 + \exp(\mathbf{X}_i^T \boldsymbol{\beta})} \right) + \sum_{i=1}^n \omega_i (y_i - 1) \ln (1 + \exp(\mathbf{X}_i^T \boldsymbol{\beta})) \\
&= \sum_{i=1}^n \omega_i y_i \mathbf{X}_i^T \boldsymbol{\beta} - \sum_{i=1}^n \omega_i \ln (1 + \exp(\mathbf{X}_i^T \boldsymbol{\beta})) .
\end{aligned} \tag{2.8}$$

Definida a função, o método da máxima verossimilhança consiste em encontrar estimadores para cada  $\beta_j$  que maximizem a função de log-verossimilhança. Os candidatos à mínimos e máximos da função dos parâmetros desejados podem ser encontrados ao se derivar parcialmente a função de log-verossimilhança em relação a cada componente de  $\boldsymbol{\beta}$  e igualar o resultado a zero, da forma

$$\frac{\partial}{\partial \boldsymbol{\beta}} \left[ \sum_{i=1}^n \omega_i y_i \mathbf{X}_i^T \boldsymbol{\beta} - \sum_{i=1}^n \omega_i \ln (1 + \exp(\mathbf{X}_i^T \boldsymbol{\beta})) \right] = 0 \tag{2.9}$$

onde  $w_i$  representa o peso via metodologia Rake da  $i$ -ésima observação amostral, presente no conjunto de dados.

Para verificar se os candidatos encontrados acima maximizam a função de log-verossimilhança, é necessário aplicar a segunda derivada com objetivo de avaliar a concavidade da função. O estimador será considerado de máxima verossimilhança quando o valor da função gerado pela segunda derivada aplicada no candidato for menor que zero.

O sistema de equações obtido não possui solução explícita. Desta maneira, o resultado é dado por meio do uso de métodos iterativos, como o método de Newton-Raphson e o método de Fisher scoring. Segundo Morettin (2022), o método de Newton-Raphson consiste em aproximar a função de log-verossimilhança por uma função quadrática, fazendo uso da matriz hessiana. Já o método de Fisher scoring faz uso da matriz de informação de Fisher para maximização da função de log-verossimilhança.

### 2.2.1.2 Método dos mínimos quadrados ponderados iterativamente

O método dos Mínimos Quadrados Ponderados Iterativamente (IRLS) consiste em um algoritmo iterativo utilizado para estimação dos parâmetros ajustados, sendo uma aplicação do método de Newton-Raphson.

1. São definidos valores iniciais para os parâmetros  $\beta_j^{(t)}$ , com  $t$  inicial igual a 0.
2. São calculadas a probabilidade de cada observação, utilizando os valores definidos no item anterior, de forma

$$\pi_i = \frac{\exp(\mathbf{X}_i^T \boldsymbol{\beta})}{1 + \exp(\mathbf{X}_i^T \boldsymbol{\beta})}$$

3. É calculada a variância de cada observação, da forma  $V_i = \pi_i(1 - \pi_i)$ .
4. São definidas as variáveis respostas amostradas, denotadas como

$$z_i = \sum_{j=0}^p \beta_j x_{ij} + \omega_i \frac{y_i - \pi_i}{V_i}$$

5. Em caso de  $\beta_j^{(t+1)} - \beta_j^{(t)}$  pequeno, será aceita a convergência do algoritmo. Caso contrário, será realizada mais uma iteração, a partir do passo 2.

Para mais detalhes, consultar Wood (2006), Dobson e Barnett (2008) e Chartrand e Yin (2008).

## 2.2.2 Estimação de parâmetros sob a abordagem Bayesiana

Tendo sido definida a função de verossimilhança, a estimação dos parâmetros sob a abordagem Bayesiana é feita seguindo o teorema de Bayes, que implica na probabilidade da ocorrência de um evento dada a distribuição a *priori* dos parâmetros e a função de verossimilhança, da forma que:

$$p(\boldsymbol{\beta}|\mathbf{y}) = \frac{l(\boldsymbol{\beta}; \mathbf{y})p(\boldsymbol{\beta})}{p(\mathbf{y})} \propto l(\boldsymbol{\beta}; \mathbf{y})p(\boldsymbol{\beta}), \quad (2.10)$$

onde  $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_j)$  e distribuição a *priori*, denotada por  $p(\boldsymbol{\beta})$ , representa a adição de um conhecimento prévio sobre os parâmetros a serem estimados.

Quando não se possui informação prévia disponível em relação aos parâmetros a serem estimados, convém utilizar uma distribuição *a priori* não informativa. Neste caso, as estimativas pontuais e intervalares dos parâmetros desconhecidos sob a perspectiva Bayesiana costumam ser semelhantes às obtidas sob a perspectiva clássica.

Supondo que não há conhecimento prévio sobre os efeitos  $\boldsymbol{\beta}$ , uma distribuição *a priori* muito utilizada é a normal centrada em zero com uma variância grande:  $\beta_j \stackrel{iid}{\sim} N(0; \sigma^2)$ ,  $j = 0, \dots, p$ , de forma que as distribuições de todos os  $\beta_j$  são independentes e identicamente distribuídas.

Desta forma, a distribuição *a posteriori* é dada por

$$\begin{aligned} p(\boldsymbol{\beta}|\mathbf{y}) &\propto p(\boldsymbol{\beta})l(\boldsymbol{\beta}; \mathbf{y}) \\ &\propto \exp\left\{-\frac{\sum_{j=0}^p \beta_j^2}{2\sigma^2}\right\} \prod_{i=1}^n \left(\frac{\exp(\mathbf{X}_i^T \boldsymbol{\beta})}{1 + \exp(\mathbf{X}_i^T \boldsymbol{\beta})}\right)^{y_i} \left(1 - \frac{\exp(\mathbf{X}_i^T \boldsymbol{\beta})}{1 + \exp(\mathbf{X}_i^T \boldsymbol{\beta})}\right)^{1-y_i} \end{aligned} \quad (2.11)$$

A distribuição *a posteriori*, descrita acima é de difícil solução. Sendo assim, para inferir sobre o parâmetro  $\boldsymbol{\beta}$  pode-se recorrer aos Monte Carlo via Cadeias de Markov (MCMC). Desta forma, obtêm-se uma amostra da distribuição *a posteriori* para realizar a inferência Bayesiana.

### 2.2.3 Métodos de Monte Carlo via Cadeias de Markov

Os métodos de Monte Carlo via cadeias de Markov, também conhecidos por MCMC se tratam de métodos iterativos, que servem para amostrar de distribuições de probabilidade, quando é difícil ou impossível amostrar diretamente dessa distribuição, mas é possível construir uma cadeia de Markov que tem a distribuição desejada como sua distribuição estacionária.

Para explicar o método, considere que a distribuição de interesse é a distribuição *a posteriori*, dada pela Equação 2.11. O método deseja construir uma cadeia de Markov de primeira ordem, obtendo desta forma que

$$p(\boldsymbol{\beta}^{(t)}|\boldsymbol{\beta}^{(0)}, \dots, \boldsymbol{\beta}^{(t-1)}) = p(\boldsymbol{\beta}^{(t)}|\boldsymbol{\beta}^{(t-1)}), \quad (2.12)$$

sendo  $\boldsymbol{\beta}^{(t)}$  o valor amostrado na  $t$ -ésima iteração. Ou seja, a distribuição de  $\boldsymbol{\beta}^{(t)}$  depende apenas do valor anterior amostrado,  $\boldsymbol{\beta}^{(t-1)}$ .

Para ser válida, a cadeia deve possuir determinadas propriedades, como a homogeneidade, que consiste em dizer que as probabilidades se mantêm as mesmas durante

as iterações, irreduzibilidade, de forma a garantir que qualquer estado pode ser atingido com determinadas iterações, dado qualquer estado anterior e aperiodicidade, que garante a não existência de estados absorventes (EHLERS, 2003).

Neste trabalho, serão descritos 3 métodos muito utilizados: Metropolis-Hastings, Amostrador de Gibbs e Monte Carlo Hamiltoniano.

### 2.2.3.1 Algoritmo de Metropolis-Hastings

O Algoritmo de Metropolis-Hastings consiste em um método de amostragem usualmente utilizado para se obter uma amostra de variáveis cuja distribuição é difícil de se amostrar diretamente. De acordo com Hitchcock (2003), o método faz uso de uma distribuição simples de ser amostrada, denominada distribuição proposta, gerando amostras dos parâmetros desejados, que serão aceitas ou rejeitadas de acordo com uma regra de aceitação. Assumindo que o parâmetro de interesse seja  $\beta$ , a construção do algoritmo se dá de acordo com as etapas listadas:

1. Escolhe-se um valor inicial arbitrário para o parâmetro  $\beta$ , denotado por  $\beta^{(0)}$ , sendo este valor pertencente ao conjunto dos possíveis valores de  $\beta$ .
2. Inicializa-se o contador de iterações  $t=1$ .
3. Utilizando o valor  $\beta^{(t-1)}$ , gera-se um ponto  $\beta^*$ , candidato a novo ponto, de acordo com uma distribuição de probabilidade atribuída a ele, de forma que  $\beta^* \sim q(\beta^*|\beta^{(t-1)})$
4. Calcula-se a razão de densidades, denotada por  $\gamma$ , de forma que:

$$\gamma = \min \left( 1, \frac{f(\beta^*|\mathbf{y})q(\beta^{(t-1)}|\beta^*)}{f(\beta^{(t-1)}|\mathbf{y})q(\beta^*|\beta^{(t-1)})} \right)$$

onde  $f(\cdot)$  é a distribuição de interesse e pode ser resumida apenas pelo núcleo da distribuição.

5. Gera-se um valor aleatório da distribuição uniforme contínua no intervalo  $(0, 1)$ ,  $U \sim U(0, 1)$ .
6. Caso  $\gamma$  seja maior que  $U$ , assume-se que  $\beta^{(t)} = \beta^*$ . Caso contrário, assume-se  $\beta^{(t)} = \beta^{(t-1)}$ .
7. Faz-se  $t = t + 1$  e repete-se os passos de 3 a 7 até obter convergência.

Segundo Roberts e Rosenthal (2001), um algoritmo pode ser considerado eficaz para a correta estimação do valor do parâmetro quando são aceitos aproximadamente 44% dos valores gerados para o parâmetro, de acordo com o número de iterações geradas.

### 2.2.3.2 Amostrador de Gibbs

Segundo Casella e George (1992), amostrador de Gibbs se trata de um método de MCMC, utilizado para gerar amostras de uma distribuição multivariada, fazendo uso da amostragem condicional. O método consiste em amostrar valores de cada variável aleatória condicionada as demais variáveis previamente amostradas. O processo de amostragem é repetido finitas vezes, de forma a atingir convergência para distribuição de interesse, a partir de um determinado número de iterações.

Suponha que o vetor aleatório a ser amostrado seja dividido em  $K$  blocos, com cada bloco contendo um vetor, uma matriz ou um escalar. Dessa forma, tem-se que

$$\boldsymbol{\beta}^T = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \dots, \boldsymbol{\beta}_K)$$

com  $K \leq p$  blocos de variáveis. Seja  $\boldsymbol{\beta}_{-j}^T = (\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_{j-1}, \boldsymbol{\beta}_{j+1}, \dots, \boldsymbol{\beta}_K)$  a coleção de todas as variáveis de  $\boldsymbol{\beta}$ , excluindo o bloco  $\boldsymbol{\beta}_j$ .

O algoritmo do Amostrador de Gibbs é:

1. Defina valores iniciais para a variável  $\boldsymbol{\beta}$  e denote isso por  $\boldsymbol{\beta}^{(0)}$ .
2. Inicialize o contador de iterações da seguinte forma:  $t = 1$ .
3. Gere valores para  $\boldsymbol{\beta}$  das distribuições condicionais completas a posteriori:

$$\boldsymbol{\beta}_1^{(t)} \sim f(\boldsymbol{\beta}_1 | \boldsymbol{\beta}_1^{(t-1)}, \dots, \boldsymbol{\beta}_K^{(t-1)}, \mathbf{y})$$

$$\boldsymbol{\beta}_2^{(t)} \sim f(\boldsymbol{\beta}_2 | \boldsymbol{\beta}_1^{(t)}, \boldsymbol{\beta}_3^{(t-1)}, \boldsymbol{\beta}_K^{(t-1)}, \mathbf{y})$$

$$\vdots$$

$$\boldsymbol{\beta}_K^{(t)} \sim f(\boldsymbol{\beta}_K | \boldsymbol{\beta}_1^{(t)}, \boldsymbol{\beta}_3^{(t)}, \boldsymbol{\beta}_{K-1}^{(t)}, \mathbf{y})$$

4. Incremente o contador de iterações fazendo  $t = t + 1$  e repita os passos 3 e 4 até obter convergência.

As distribuições condicionais completas a posteriori (DCCP) são proporcionais a distribuição *a posteriori*. Este método é escolhido quando amostrar diretamente da

distribuição *a posteriori* não é possível, mas amostrar das DCCP's é mais simples. Quando parte das DCCP's são conhecidas e outras não, costuma-se combinar o amostrador de Gibbs com o Metropolis-Hastings.

### 2.2.3.3 Monte Carlo Hamiltoniano

Proposto por Duane et al. (1987), o método de Monte Carlo Hamiltoniano consiste em um algoritmo iterativo com objetivo de melhorar a eficiência de amostragem em relação aos demais algoritmos de MCMC, de forma que é utilizada a visão Hamiltoniana por meio do gradiente da função de log-verossimilhança a posteriori do modelo considerado.

Para o caso particular de um algoritmo de MCMC, a função Hamiltoniana pode ser escrita como

$$H(\boldsymbol{\beta}, \mathbf{p}) = -\log f(\boldsymbol{\beta}) + \frac{1}{2} \mathbf{p}^T \mathbf{M}^{-1} \mathbf{p} \quad (2.13)$$

onde  $\mathbf{p}$  representa o *momentum* da equação Hamiltoniana, com distribuição assumida igual a  $N_j(0, M)$  e  $M$  a matriz de covariância.

O algoritmo faz uso de equações Hamiltonianas para garantir a melhor aceitação das amostras, de forma que as equações podem ser definidas pelas derivadas parciais de primeira ordem iguais a

$$\frac{d\mathbf{p}}{dt} = -\frac{\partial H(\boldsymbol{\beta}, \mathbf{p})}{\partial \boldsymbol{\beta}} = \nabla_{\boldsymbol{\beta}} \log f(\boldsymbol{\beta}),$$

$$\frac{d\boldsymbol{\beta}}{dt} = -\frac{\partial H(\boldsymbol{\beta}, \mathbf{p})}{\partial \mathbf{p}} = \mathbf{M}^{-1} \mathbf{p}$$

onde  $\nabla_{\boldsymbol{\beta}} \log f(\boldsymbol{\beta})$  representa o gradiente do logaritmo de densidade a posteriori. Mais detalhes podem ser vistos em Thomas e Tu (2021) e Pérez (2021).

### 2.2.3.4 Convergência dos estimadores

A convergência de um algoritmo de MCMC pode ser avaliada por uma série de análises gráficas relacionadas as estimativas geradas pelo algoritmo realizado. É desejado que uma sequência de parâmetros amostrados não possua correlação entre si, garantindo independência entre as amostras geradas. Esta análise pode ser feita pelo gráfico de autocorrelação, de forma que a correlação entre o parâmetro é representada graficamente pelas defasagens desta função de autocorrelação, de forma que quanto maior a defasagem (também chamada de *lag*), maior o número de parâmetros consecutivos correlacionados.

Uma solução para a correlação se dá pela inclusão de intervalos entre os parâmetros amostrados, desconsiderando iterações que possuem correlação entre si.

Após a garantia da não correlação entre as amostras geradas, é possível visualizar a convergência dos parâmetros amostrados por meio do traço das cadeias de Markov geradas pelo ajuste a *posteriori* dos parâmetros, de forma que é esperado que as amostras estejam contidas em um intervalo mínimo. Casos de não convergência podem ser resolvidos ao se considerar um maior número de iterações. Por fim, pode-se observar o histograma dos parâmetros estimados pelo ajuste a *posteriori*, de forma que é esperado simetria entre a distribuição a *posteriori* amostrada, além de uma menor densidade entre as caudas da distribuição, indicando baixa aceitação de parâmetros considerados *outliers*.

Além das análises gráficas realizadas, é possível avaliar a convergência por meio de estimativas pontuais. Gelman e Rubin (1992) introduzem um diagnóstico de convergência para o algoritmo MCMC denominado  $\hat{R}$  e usualmente utilizado para avaliar a convergência de múltiplas cadeias. O  $\hat{R}$  compara as estimativas entre e dentro da cadeias, de forma que

$$\hat{R} = \sqrt{\frac{Var(T)}{W}}$$

onde  $Var(T)$  é dado pela variância das médias entre as cadeias geradas e  $W$  é a média das variâncias dentro das cadeias.

Valores iguais a 1 indicam convergências entre as cadeias executadas. Já valores superiores a 1 indicam potencial dependência entre as cadeias, sendo necessária a realização de mais iterações.

## 2.2.4 Estimativa intervalar

Na abordagem clássica, a estimativa intervalar é realizada por meio do intervalo de confiança para cada parâmetro estimado. O primeiro passo se dá pelo cálculo do erro padrão de cada parâmetro, definido como a raiz quadrada da variância do parâmetro  $\hat{\beta}_j$  estimado, ou seja,  $SE = \sqrt{var(\hat{\beta}_j)}$ . Encontrado o erro padrão, o intervalo de confiança de  $100(1 - \alpha)\%$  será definido como:

$$IC_{\beta_j; 100(1-\alpha)\%} = \left[ \hat{\beta}_j - z_{1-\frac{\alpha}{2}} SE; \hat{\beta}_j + z_{1-\frac{\alpha}{2}} SE \right]$$

onde  $z_{1-\frac{\alpha}{2}}$  é o quantil de uma distribuição normal padrão com probabilidade acumulada igual a  $1 - \frac{\alpha}{2}$  e SE é o erro padrão.

Já na abordagem Bayesiana, a estimativa intervalar é dada pelo intervalo de credibilidade a *posteriori*. Diferente da interpretação clássica, o intervalo de credibilidade se propõe a determinar a região a *posteriori* de maior densidade centrada à média. Desta forma, o intervalo de credibilidade é dado pelos limites da distribuição a *posteriori* centrada no parâmetro médio estimado, com probabilidade igual a  $100(1 - \alpha)\%$ .

### 2.2.5 Interpretação dos parâmetros

A interpretação de cada parâmetro  $\beta_j$  é realizada por meio da razão de chances e para calcular isso, é necessário falar antes do conceito de chance.

A chance ou *odd* representa a razão entre a probabilidade de um evento ocorrer e a probabilidade de não ocorrer. Para interpretar o efeito das covariáveis, incluiu-se nesta Subseção a dependência destes valores na distribuição dos dados. No modelo de regressão logística, a chance é dada por

$$ODD = \frac{P(Y_i = 1 | \boldsymbol{\beta}, \mathbf{X}_i)}{P(Y_i = 0 | \boldsymbol{\beta}, \mathbf{X}_i)} = \frac{\pi_i}{1 - \pi_i} = \exp(\mathbf{X}_i^T \boldsymbol{\beta}). \quad (2.14)$$

Uma chance igual a 1 indica que  $\pi_i = 1 - \pi_i$ , e neste caso, é dito que há uma chance de sucesso para uma chance de fracasso. Uma *odd* maior que 1 indica maior chance de sucesso enquanto uma *odd* inferior a 1 indica uma maior chance de fracasso.

Já a razão de chances, também denominada como *odds ratio* representa a razão entre a chance de dois grupos. No caso da regressão logística, o grupo sucesso, representado pelos indivíduos classificados como obesos e o grupo fracasso, representado pelos indivíduos classificados como não obesos. Dessa forma tem-se que a razão de chances é dada por

$$\begin{aligned} OR &= \frac{\frac{P(Y_i=1|\boldsymbol{\beta},x_{i1}=x_1,\dots,x_{i,j-1}=x_{j-1},x_{ij}=1,x_{i,j+1}=x_{j+1},\dots,x_{i,p}=x_p)}{P(Y_i=0|\boldsymbol{\beta},x_{i1}=x_1,\dots,x_{i,j-1}=x_{j-1},x_{ij}=1,x_{i,j+1}=x_{j+1},\dots,x_{i,p}=x_p)}}{\frac{P(Y_i=1|\boldsymbol{\beta},x_{i1}=x_1,\dots,x_{i,j-1}=x_{j-1},x_{ij}=0,x_{i,j+1}=x_{j+1},\dots,x_{i,p}=x_p)}{P(Y_i=0|\boldsymbol{\beta},x_{i1}=x_1,\dots,x_{i,j-1}=x_{j-1},x_{ij}=0,x_{i,j+1}=x_{j+1},\dots,x_{i,p}=x_p)}} \\ OR &= \frac{ODD(x_{i1} = x_1, \dots, x_{i,j-1} = x_{j-1}, x_{ij} = 1, x_{i,j+1} = x_{j+1}, \dots, x_{i,p} = x_p)}{ODD(x_{i1} = x_1, \dots, x_{i,j-1} = x_{j-1}, x_{ij} = 0, x_{i,j+1} = x_{j+1}, \dots, x_{i,p} = x_p)} \\ &= \exp\left(\beta_0 + \beta_j + \sum_{l=1;l \neq j}^p x_l \beta_l - \beta_0 - \sum_{l=1;l \neq j}^p x_l \beta_l\right) = \exp(\beta_j). \end{aligned} \quad (2.15)$$

Assim, a razão de chances será dada por  $\exp(\beta_j)$ , onde um valor igual a 1 indica que a variável independente  $x_j$  não possui efeito sobre a probabilidade do evento. Já uma razão de chances superior (ou inferior) a 1 indica que a variável independente  $x_j$  está associada ao aumento (à redução) de  $100 * [\exp(\beta_j) - 1]\%$  nas chances de ocorrência do evento.



### 2.2.6 Testes de hipóteses clássicos

Nesta Subseção, serão apresentados testes de hipóteses necessários para justificar a necessidade da remoção de parâmetros que não agregam valor ao modelo ajustado por meio da abordagem clássica, garantindo que todos os parâmetros sejam significativos ao modelo ajustado. Outro modo de avaliar a necessidade de remoção do parâmetro é por meio de sua significância, que pode ser avaliada pelo intervalo de confiança, no caso clássico e intervalo de credibilidade no caso Bayesiano, sendo intervalos que contenham o número zero como um critério para remoção ou tratamento da variável.

#### 2.2.6.1 Teste de Wald

O teste proposto por Wald (1943) consiste em avaliar o grau de significância do parâmetro  $\beta_j$ . É considerado significativo qualquer parâmetro não nulo, ou seja, diferente de zero, assim a hipótese nula pode ser denotada como  $H_0 : \beta_j = 0$  e a hipótese alternativa como  $H_1 : \beta_j \neq 0$ . Já a estatística de teste é dada por

$$W = \frac{\hat{\beta}_j}{SE}, \quad SE = \sqrt{Var(\hat{\beta}_j)} \quad (2.16)$$

com  $\hat{\beta}_j$  igual ao parâmetro amostrado e SE o erro padrão deste parâmetro. A estatística do teste segue uma distribuição qui-quadrado com 1 grau de liberdade. Para este trabalho, parâmetros testados que resultem em um p-valor menor que 0,05 serão considerados significativos, levando a rejeição da hipótese nula.

#### 2.2.6.2 Teste da razão da verossimilhança

O teste de razão de verossimilhança consiste em comparar dois modelos diferentes em um mesmo conjunto de dados. A ideia geral é verificar se a exclusão de um ou mais parâmetros  $\beta_j$  afeta na significância do modelo. O teste considera como hipótese nula o modelo sem os parâmetros que serão avaliados e como hipótese alternativa o modelo com todos os parâmetros  $\beta_j$  estimados. A estatística de teste é calculada com base na função de log-verossimilhança de ambas hipóteses, sendo

$$RV = -2 \ln \left( \frac{l_0(\hat{\beta}; \mathbf{y})}{l_1(\hat{\beta}; \mathbf{y})} \right) \quad (2.17)$$

onde  $l_0(\boldsymbol{\beta}; \mathbf{y})$  representação a função de log-verossimilhança sob a hipótese nula e  $l_1(\boldsymbol{\beta}; \mathbf{y})$  a função de verossimilhança sob a hipótese alternativa.

A estatística segue uma distribuição qui-quadrado com os graus de liberdade dados pela diferença entre as quantidades de parâmetros consideradas em ambas hipóteses. Para este trabalho, a hipótese nula será rejeitada nos casos em que o teste resulta em um p-valor menor que 0,05. Nesses casos, a exclusão de um ou mais parâmetros presentes interfere na eficiência do modelo, sugerindo seguir com o modelo com todos os parâmetros ajustados.

## 2.2.7 Critérios de comparação entre modelos

### 2.2.7.1 Critério de Informação de Akaike

O Critério de informação de Akaike, proposto por Akaike (1974) consiste em uma medida de comparação entre modelos de regressão onde os parâmetros são estimados por meio do método de máxima verossimilhança. O critério é definido como:

$$AIC = -2 \ln l(\hat{\boldsymbol{\beta}}; \mathbf{y}) + 2k \quad (2.18)$$

onde  $l(\hat{\boldsymbol{\beta}}; \mathbf{y})$  é a função de verossimilhança aplicada no EMV  $\hat{\boldsymbol{\beta}}$  e  $k$  é o número de parâmetros do modelo. Um valor de Critério de informação de Akaike (AIC) baixo indica melhor equilíbrio entre o ajuste e a complexidade do modelo, sendo um fator de escolha ao se comparar diferentes ajustes.

### 2.2.7.2 Critério de Informação de Desvio

Proposto por Spiegelhalter et al. (2002), o Critério de Informação de Desvio (DIC) corresponde a uma generalização do AIC, sendo uma medida de comparação entre ajustes de modelos de regressão Bayesianos onde os parâmetros da distribuição a *posteriori* amostrados são obtidos por MCMC.

O critério faz uso da estatística desvio, denotada como:

$$D(\boldsymbol{\beta}) = -2 \ln l(\hat{\boldsymbol{\beta}}; \mathbf{y}) + C \quad (2.19)$$

de forma que  $l(\hat{\boldsymbol{\beta}}; \mathbf{y})$  é a função de verossimilhança para um dado conjunto de dados observados  $\mathbf{y}$  e  $C$  uma constante dada.

Já o critério é definido como:

$$DIC = D(\bar{\boldsymbol{\beta}}) + p_D = D(\bar{\boldsymbol{\beta}}) + 2p_D \quad (2.20)$$

onde  $D(\bar{\boldsymbol{\beta}})$  é o desvio médio a *posteriori* do modelo,  $D(\bar{\boldsymbol{\beta}})$  é o desvio da média a *posteriori* do modelo e  $p_D$  é o número efetivo de parâmetros no modelo, calculado pela diferença entre o desvio médio e o desvio da média a *posteriori*.

Um valor de DIC baixo indica equilíbrio entre o ajuste dos dados e a complexidade do modelo, sendo um critério para escolha em detrimento de outros ajustes considerados.

## 3 Análise de Resultados

O presente capítulo está dividido de maneira que a Seção 3.1 contém informações relacionadas à análise descritiva dos dados, considerando a ponderação via método Rake realizada pelo VIGITEL a fim de aproximar os resultados amostrados com a população das capitais brasileiras no ano de 2021 e no primeiro semestre de 2023. Já a Seção 3.2 contém todo o processo de ajuste dos dados via regressão logística clássica e Bayesiana, incluindo a capacidade de estimação e qualidade de ajuste de cada modelo. Para verificar a capacidade de estimação de cada ajuste, foram realizadas simulações considerando parâmetros preestabelecidos, descritos na Subseção 3.2.1. Já as Subseções 3.2.2 e 3.2.3 descrevem todo o processo de ajuste dos dados na abordagem clássica e Bayesiana, respectivamente.

### 3.1 Análise Descritiva

A amostra final possui 29.571 indivíduos e considerando os pesos rakes é como se tivesse 49.950.734 indivíduos. Todas as análises foram feitas usando o peso rake. Aproximadamente 22,9% dos indivíduos amostrados eram obesos. Em relação ao perfil sociodemográfico, foram consideradas as variáveis Sexo, Faixa Etária, Raça/Cor, Estado Civil e Anos de Estudo. Foi possível identificar que a amostra é composta por sua maioria de adultos (73,2%), com raça/cor parda (44,1%) e branca (39,7%), solteiros (48,8%), casados ou em união estável (41,9%), além de indivíduos com pelo menos nove anos de estudo (77,5%).

A análise do percentual de obesidade de acordo com o perfil sociodemográfico presente na Tabela 2 sugere que a faixa etária do indivíduo possui relação com a obesidade, de forma que indivíduos adultos estão mais propensos a serem obesos. Ao se considerar a variável raça/cor, observa-se que indivíduos de raça/cor preta possuem uma maior proporção de obesos, seguidos de indivíduos com raça/cor diferente das três consideradas.

Ao se considerar o Estado Civil, foi possível verificar que indivíduos casados ou em união estável possuem maior probabilidade de serem obesos, contrapondo indivíduos solteiros, que possuem menor probabilidade. Por fim, observa-se que os anos de estudos se opõem a obesidade, sendo a proporção de indivíduos obesos com até 8 anos de estudo maior que indivíduos com mais anos de estudo.

Tabela 2: Distribuição dos dados ponderados do VIGITEL e proporção de obesos de acordo com o perfil sociodemográfico durante o ano de 2021 e primeiro semestre de 2023

<b>Perfil Sociodemográfico</b>	<b>Geral (%)</b>	<b>Obesos (%)</b>
<b>Sexo</b>		
Feminino	51,4	23,0
Masculino	48,6	22,8
<b>Faixa Etária</b>		
Jovem (18 a 24 anos)	14,6	12,7
Adulto (25 a 64 anos)	73,2	25,3
Idoso (65 em diante)	12,2	21,1
<b>Raça/Cor</b>		
Branca	39,7	21,8
Parda	44,1	22,6
Preta	14,2	26,6
Outros	2,0	25,7
<b>Estado Civil</b>		
Solteiro	48,8	20,1
Casado ou União Estável	41,9	26,1
Separado, Divorciado ou Viúvo	9,3	23,3
<b>Anos de Estudo</b>		
0 a 8 anos	22,5	24,8
9 a 11 anos	41,6	23,3
12 anos ou mais	35,9	21,3

Entre as variáveis que representam hábitos e fatores com potencial relação à obesidade, representadas na Tabela 3, observa-se que a proporção de obesos é consideravelmente maior em indivíduos que tiveram depressão. Por outro lado, a prática de exercícios físicos e o consumo de cigarros se mostraram contrárias à obesidade, sendo a proporção de obesos para estes grupos menor que a proporção geral. Verifica-se também que os hábitos alimentares também se mostraram relacionados a obesidade, de maneira que a proporção de obesos que alegaram consumir legumes, verduras ou frutas de forma regular é menor que a proporção de obesos que alegaram consumir refrigerante. Outros fatores também demonstraram relação com a presença de obesidade, como o diagnóstico positivo para Covid.

Tabela 3: Distribuição dos dados ponderados do VIGITEL e proporção de obesos de acordo com potenciais hábitos e fatores relacionados a obesidade durante o ano de 2021 e primeiro semestre de 2023

<b>Hábitos e Fatores</b>	<b>Geral (%)</b>	<b>Obesos (%)</b>
<b>Consumo de bebidas Alcoólicas</b>		
Sim	54,5	22,0
Não	45,5	24,0
<b>Consumo de cigarro</b>		
Sim	10,2	17,4
Não	89,8	23,5
<b>Prática de exercício físico</b>		
Sim	58,9	19,4
Não	41,1	28,0
<b>Uso de celular/computador no tempo livre</b>		
Sim	81,2	23,8
Não	18,7	22,7
<b>Teve COVID</b>		
Sim	29,6	25,1
Não	70,4	22,0
<b>Plano de Saúde</b>		
Sim	42,8	21,6
Não	57,2	23,9
<b>Depressão</b>		
Sim	11,5	28,3
Não	88,5	22,2
<b>Consumo de Legumes ou Verduras</b>		
Sim	74,4	22,4
Não	25,6	24,5
<b>Consumo de Frutas</b>		
Sim	69,3	22,2
Não	30,7	24,6
<b>Consumo de Refrigerante</b>		
Sim	29,1	24,9
Não	70,9	22,1

## 3.2 Ajuste do Modelo de Regressão Logística

Nesta Seção, serão descritos todos os processos para o ajuste dos dados reais e simulados, por meio da abordagem Clássica e Bayesiana. Os ajustes foram realizados por meio do ambiente de desenvolvimento integrado RStudio, que faz uso da linguagem de programação R Core Team (2023). Para o ajuste clássico, foi utilizado o pacote *survey*. Já para o ajuste Bayesiano, foi considerado o pacote *rstanarm*.

### 3.2.1 Ajuste dos dados simulados

Com intenção de garantir a eficiência de ambos os ajustes, foi realizada uma simulação dos conjuntos de dados. A base utilizada para simulação foi gerada de maneira aleatória com quatro covariáveis oriundas de uma distribuição de Bernoulli de tamanho 10.000 cada com probabilidade de sucesso igual a 0,6, 0,3, 0,5 e 0,2, respectivamente. Já os efeitos das covariáveis foram gerados de uma distribuição normal com média igual a zero e variância igual a um. As distribuições utilizadas para gerar os parâmetros a serem estimados resultaram no vetor de parâmetros  $\beta = [0, 2102; -0, 2909; 1, 4036; -0, 7508]^T$ . Já a variável que representa o peso de cada amostra foi gerada de uma distribuição exponencial com média 1.000, garantindo assim valores consideravelmente grandes e não negativos.

O primeiro ajuste realizado foi o ajuste pela abordagem clássica. Ao analisar os resultados obtidos por meio da função *svyglm* do pacote *survey*, descritos na Tabela 4, foi possível observar que os valores estimados se aproximaram do real valor de cada parâmetro, de forma que o valor a ser estimado esteve contido em todos os intervalos de confiança, ao nível de 95%. Em relação a significância dos parâmetros, todos os parâmetros se mostraram significantes (p-valor < 0,05).

Tabela 4: Valores reais, valores estimados, intervalo de confiança e p-valor da estatística de Wald para os parâmetros considerando o ajuste do modelo de regressão clássica para os dados simulados

Parâmetro	Valor real	Valor estimado	Intervalo de confiança	P-valor
$\beta_0$ (Intercepto)	0,2102	0,229	[0,1054; 0,3526]	0,0003
$\beta_1$	-0,2909	-0,264	[-0,3972; -0,1307]	0,0001
$\beta_2$	1,4036	1,2831	[1,1377; 1,4284]	0
$\beta_3$	-1,9892	-1,9858	[-2,1272; -1,8444]	0
$\beta_4$	-0,7508	0,7104	[-0,8804; -0,5403]	0

Para abordagem Bayesiana, foram realizados três ajustes distintos, com distribuições a priori independentes iguais a  $\beta_j \sim N(0; 0,01)$  para o primeiro ajuste,  $\beta_j \sim N(0; 1)$  para o segundo ajuste e  $\beta_j \sim N(0; 1000)$  para o terceiro ajuste, com  $j = 1, \dots, 4$ . Para o intercepto, foi atribuída a distribuição *a priori*  $\beta_0 \sim N(0; 6,25)$  em todos os ajustes. A estimação foi realizada por meio de um algoritmo de MCMC, com quadro cadeias de 5.000 iterações para cada ajuste, onde as 500 primeiras iterações foram consideradas iterações de aquecimento. A estimativa pontual foi dada pela média das iterações consideradas, enquanto a estimativa intervalar considerou o limite inferior e superior da distribuição a posteriori centrada à média, com probabilidade igual a 95%.

A convergência dos algoritmos realizados pode ser confirmada pela medida da estatística  $\hat{R}$ , presente nas Tabelas 5, 6 e 7, onde valores iguais a 1 indicam convergência. A análise gráfica presente nas Figuras 1, 2 e 3 confirma esta convergência, de forma que o traço de todas as cadeias se mostrou convergir para um determinado intervalo. Ao se analisar os gráficos de autocorrelação presentes nas Figuras 4, 5 e 6, foi possível verificar que de maneira geral, não existe dependência entre os valores seguidos das iterações realizadas. Alguns *lags* foram identificados, porém não foi necessário a inclusão de intervalos nas cadeias, uma vez que os ajustes se mostraram eficientes na estimação dos parâmetros  $\beta_j$  avaliados.

Diferente do ajuste clássico, os intervalos de credibilidade estimados não continham os valores reais dos parâmetros a serem estimados, devido a amplitude reduzida destes intervalos. Apesar disso, os ajustes Bayesianos também se mostraram eficientes para estimar os parâmetros definidos na simulação, uma vez que as estimativas paramétricas geradas ficaram próximas do real valor de cada parâmetro. Por conta do grande número de observações, as diferentes especificações *a priori* não possuíram um impacto considerável nos ajustes realizados.

Tabela 5: Valores reais, valores estimados e intervalo de credibilidade para os parâmetros considerando o ajuste do modelo de regressão Bayesiana para os dados simulados com distribuição a priori  $\beta_j \sim N(0; 0,01)$

Parâmetro	Valor real	Valor estimado	Estimativa intervalar
$\beta_0$ (Intercepto)	0,2102	0,229	[0,2281; 0,2299]
$\beta_1$	-0,2909	-0,264	[-0,2649; -0,263]
$\beta_2$	1,4036	1,283	[1,282; 1,284]
$\beta_3$	-1,9892	-1,9857	[-1,9867; -1,9847]
$\beta_4$	-0,7508	-0,7103	[-0,7115; -0,7091]



Tabela 6: Valores reais, valores estimados e intervalo de credibilidade para os parâmetros considerando o ajuste do modelo de regressão Bayesiana para os dados simulados com distribuição a priori  $\beta_j \sim N(0; 1)$

Parâmetro	Valor real	Valor estimado	Estimativa intervalar
$\beta_0$ (Intercepto)	0,2102	0,229	[0,2281; 0,2299]
$\beta_1$	-0,2909	-0,264	[-0,2649; -0,263]
$\beta_2$	1,4036	1,283	[1,282; 1,2841]
$\beta_3$	-1,9892	-1,9858	[-1,9868; -1,9848]
$\beta_4$	-0,7508	-0,7104	[-0,7116; -0,7092]

Tabela 7: Valores reais, valores estimados e intervalo de credibilidade para os parâmetros considerando o ajuste do modelo de regressão Bayesiana para os dados simulados com distribuição a priori  $\beta_j \sim N(0; 1000)$

Parâmetro	Valor real	Valor estimado	Estimativa intervalar
$\beta_0$ (Intercepto)	0,2102	0,229	[0,2281; 0,2299]
$\beta_1$	-0,2909	-0,264	[-0,2649; -0,263]
$\beta_2$	1,4036	1,283	[1,282; 1,2841]
$\beta_3$	-1,9892	-1,9858	[-1,9868; -1,9848]
$\beta_4$	-0,7508	-0,7103	[-0,7116; -0,7092]

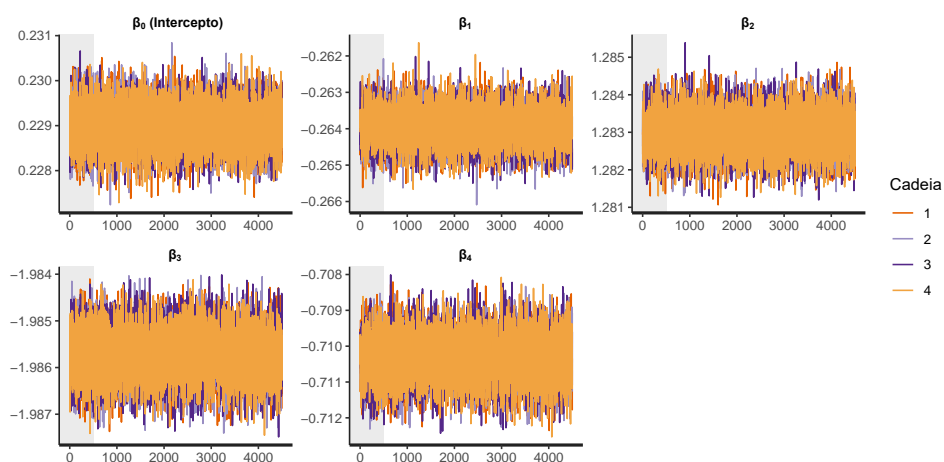


Figura 1: Traço das Cadeias de Markov para o ajuste a *posteriori* dos dados simulados dos parâmetros com distribuição a *priori* Normal com média igual a 0 e variância igual a 0,01

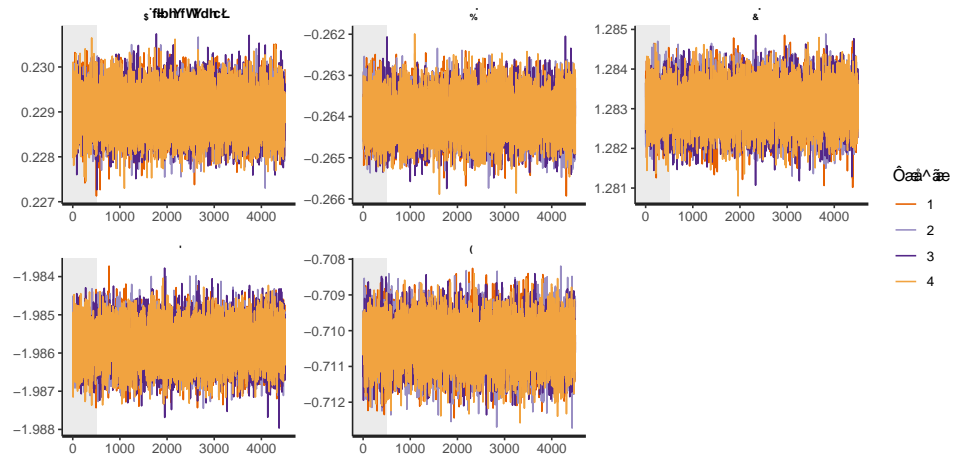


Figura 2: Traço das Cadeias de Markov para o ajuste *a posteriori* dos dados simulados dos parâmetros com distribuição *a priori* Normal com média igual a 0 e variância igual a 1

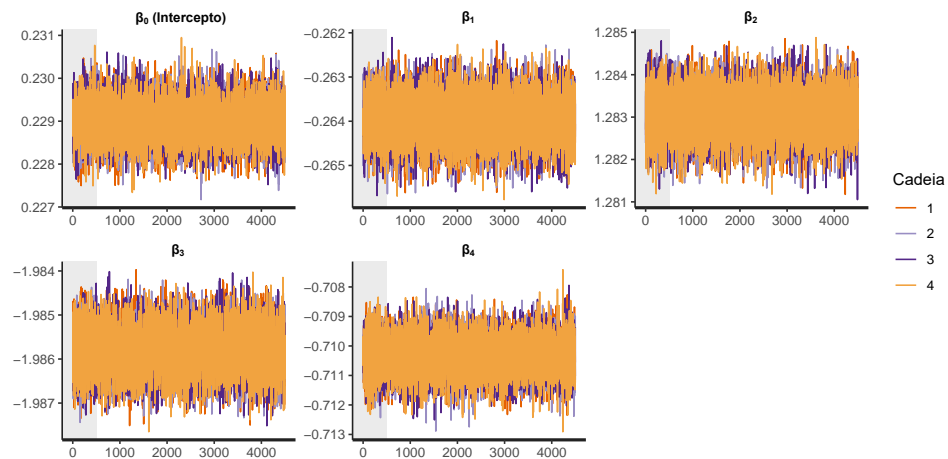


Figura 3: Traço das Cadeias de Markov para o ajuste *a posteriori* dos dados simulados dos parâmetros com distribuição *a priori* Normal com média igual a 0 e variância igual a 1000

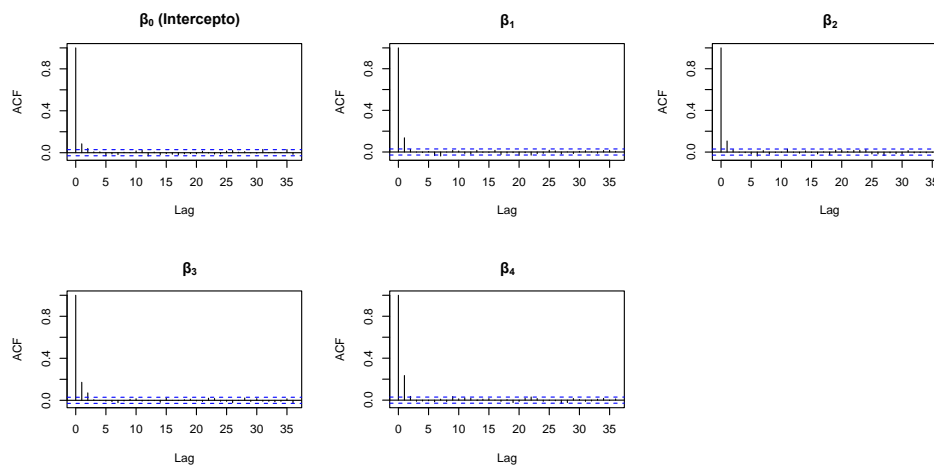


Figura 4: Função de autocorrelação das amostras geradas para o ajuste *a posteriori* dos dados simulados dos parâmetros com distribuição *a priori* Normal com média 0 e variância 0,01

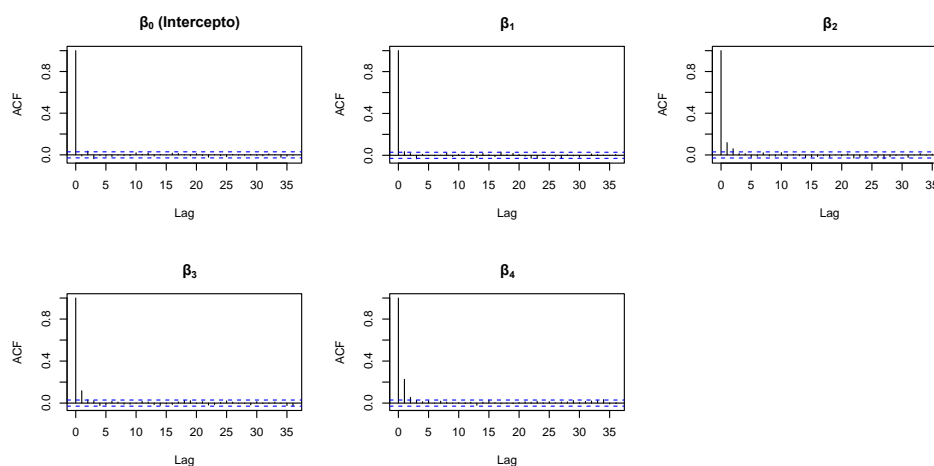


Figura 5: Função de autocorrelação das amostras geradas para o ajuste *a posteriori* dos dados simulados dos parâmetros com distribuição *a priori* Normal com média 0 e variância 1

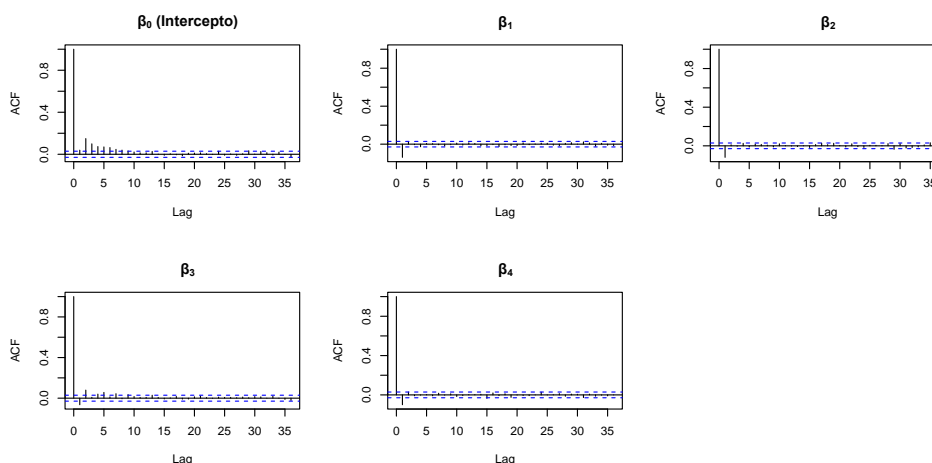


Figura 6: Função de autocorrelação das amostras geradas para o ajuste a *posteriori* dos dados simulados dos parâmetros com distribuição a *priori* Normal com média 0 e variância 1000

### 3.2.2 Modelo de Regressão Logística Clássico

Inicialmente, foi realizado um ajuste do modelo de Regressão Logística clássico que considerou todas as variáveis analisadas descritivamente, sendo o p-valor obtido pelo teste de Wald um fator de exclusão ou tratamento das variáveis, onde foram desconsideradas significativas aquelas variáveis cujo p-valor ficou abaixo de 0,05. As variáveis sexo, consumo de bebidas alcoólicas, uso de celular ou computador no tempo livre, escolaridade, plano de saúde, consumo de legumes e verduras, consumo de frutas e consumo de refrigerante foram desconsideradas por não possuírem significância no modelo ajustado. As variáveis Raça/Cor e Estado Civil foram tratadas, onde Raça/Cor passou a classificar os indivíduos como pretos ou não pretos e Estado Civil como solteiros ou não solteiros.

Um segundo ajuste foi realizado desconsiderando ou tratando todas as variáveis não significativas, resultando nas estimativas da tabela 8. Foi possível analisar que segundo o teste de Wald, todas as variáveis passaram a ser significativas (p-valor < 0,05), mostrando eficiência no tratamento e remoção das variáveis do primeiro ajuste. A qualidade do segundo ajuste também pode ser verificada ao se considerar o AIC obtido em ambos os ajustes, de forma que o ajuste com as variáveis não significativas removidas ou tratadas obteve um AIC igual a 31046 enquanto o modelo com todas as variáveis obteve um valor igual a 31100, justificando a escolha do segundo modelo realizado.

Ao se analisar os valores referentes as razões de chance, descritos na Tabela 9, foi possível observar que a faixa etária possui grande influência no diagnóstico de obesidade, de forma que indivíduos adultos possuem 101% de chance a mais de seres obesos, ao se

Tabela 8: Valores estimados, estimativa intervalar e p-valor da estatística para os parâmetros considerados no ajuste do modelo de regressão clássica

Variável	Parâmetro estimado	Estimativa Intervalar: 95%	p-valor
<b>Intercepto</b>	-1,5116	[-1,6258; -1,3989]	0
<b>Faixa Etária: Adulto</b>	0,6994	[0,6011; 0,7993]	0
<b>Faixa Etária: Idoso</b>	0,3320	[0,2014; 0,4631]	0,0166
<b>Exercício Físico</b>	-0,4611	[-0,517; -0,4053]	0
<b>Raça/Cor: Preta</b>	0,2609	[0,1846; 0,3365]	0,0077
<b>Fuma</b>	-0,465	[-0,5659; -0,366]	0
<b>Estado Civil: Solteiro</b>	-0,2064	[-0,2663; -0,1467]	0,0049
<b>Covid</b>	0,1397	[0,0797; 0,1995]	0,0392
<b>Depressão</b>	0,3308	[0,2487; 0,4122]	0,0003

comparar com indivíduos jovens. A raça/cor do indivíduo também possui relação parecida, onde indivíduos pretos possuem aproximadamente 30% a mais chances de serem obesos. O diagnóstico de depressão seguiu o mesmo comportamento, de forma que indivíduos deste grupo tiveram 39% a mais de chance de serem obesos.

Em relação as variáveis que reduziram a chance de obesidade, foi possível destacar a prática de exercícios físicos, hábito de fumar e o estado civil solteiro, de forma que indivíduos destas categorias possuíram uma chance de não serem obesos igual a aproximadamente 37% para os dois primeiros casos e 18% para o estado civil solteiro, em relação aos demais.

Tabela 9: Razão de chances e estimativa intervalar da razão de chances para os parâmetros considerados no ajuste do modelo de regressão clássica

Variável	Razão de chances %	Estimativa Intervalar: 95%
<b>Intercepto</b>	77,94	[-80,32; -75,31]
<b>Faixa Etária: Adulto</b>	101,25	[82,41; 122,39]
<b>Faixa Etária: Idoso</b>	39,38	[22,31; 58,90]
<b>Exercício Físico</b>	-36,94	[-40,37; -33,32]
<b>Raça/Cor: Preta</b>	29,8	[20,28; 40,01]
<b>Fuma</b>	-37,19	[-43,21; -30,65]
<b>Estado Civil: Solteiro</b>	-18,65	[-23,38; -13,64]
<b>Covid</b>	14,99	[8,30; 22,08]
<b>Depressão</b>	39,20	[28,24; 51,01]

### 3.2.3 Modelo de Regressão Logística Bayesiano

Foram realizados três ajustes considerando todas as variáveis presentes no modelo completo, onde  $\beta_j \sim N(0; 0,01)$  para o primeiro ajuste,  $\beta_j \sim N(0; 1)$  para o segundo ajuste e  $\beta_j \sim N(0; 1000)$  para o terceiro ajuste, com  $j = 1, \dots, p$ . Já para o parâmetro referente ao intercepto, foi considerada a distribuição a priori  $\beta_0 \sim N(0; 6, 25)$  em todos os ajustes. Observa-se que o primeiro ajuste tende a considerar uma maior informação obtida por meio da distribuição a priori, uma vez que possui menor variância em relação aos demais. Já a distribuição a priori do terceiro ajuste possui a maior variância dentre as priors consideradas, garantindo maior influência do conjunto de dados.

A estimação dos parâmetros foi realizada por meio do algoritmo de Monte Carlo Hamiltoniano, considerando 5000 iterações e 4 cadeias cada. As 500 primeiras iterações foram consideradas iterações de aquecimento, sendo descartadas do ajuste. A estimativa pontual de cada parâmetro  $\hat{\beta}_j$  foi dada pela média de cada distribuição a posteriori estimada, enquanto a estimativa intervalar considerou os quantis de menor densidade, com probabilidade igual a 95%.

Verificou-se convergência do algoritmo de MCMC, presente nas Figuras 7, 8 e 9, de forma que o traço de todas as cadeias apresentou convergência para o mesmo intervalo. Esta convergência se confirma ao se verificar que os valores de  $\hat{R}$  para todos os parâmetros estimados, presentes nas Tabelas 10, 11 e 12, de forma que todos resultaram em 1, diagnosticando convergência para todas as cadeias e descartando a necessidade da realização de mais iterações. As funções de autocorrelação, presentes nas Figuras 10, 11 e 12 indicam que, de forma geral, não existe dependência entre valores gerados em iterações consecutivas. Algumas exceções puderam ser verificadas, como por exemplo, a variável Raça/Cor:Outros na figura 12 que indica uma potencial relação entre uma sequência de até cinco iterações. Por serem casos isolados com autocorrelação apenas nos *lags* iniciais, não foi considerado nenhum tratamento, uma vez que o ajuste se mostrou eficiente na estimação dos parâmetros.

A análise da distribuição a posteriori, presente em forma de histogramas nas Figuras 13, 14 e 15 indicaram uma aproximação à distribuição unimodal e simétrica, indicando qualidade no ajuste e garantindo a não existência de viés ou subgrupos distintos. As distribuições a posteriori também se mostraram bem concentradas nas respectivas médias, gerando caudas curtas, o que indicou uma baixa aceitação a valores *outliers*. Esta evidência se confirmou ao se analisar a amplitude consideravelmente pequena dos intervalos de credibilidade.

Por fim, foi observado nas Tabelas 10, 11 e 12 que os parâmetros estimados de cada um dos ajustes ficaram próximos, indicando que as distribuições *a priori* consideradas trouxeram pouca ou nenhuma informação para a distribuição *a posteriori*. Por conta do grande número de observações, as diferentes especificações *a priori* não possuíram um impacto considerável nos ajustes dos dados. Devido aos intervalos de credibilidade não contemplarem o número zero, todas as variáveis ajustadas foram consideradas significativas, uma vez que ao nível de confiança de 95%, não há evidências para justificar que alguma das variáveis se mostrou não significativa nos ajustes realizados, de forma que todas variáveis possuem relação com a obesidade.

Tabela 10: Estimação pontual, estimativa intervalar e valor da estatística  $\hat{R}$  para os parâmetros com distribuição *a priori* Normal com média 0 e variância 0,001

Variável	Parâmetro estimado	Estimativa Intervalar: 95%	$\hat{R}$
Intercepto	-1,5086	[-1,5125; -1,5047]	1,0
Faixa Etária: Adulto	0,722	[0,7196; 0,7244]	1,0
Faixa Etária: Idoso	0,3966	[0,3933; 0,4]	1,0
Sexo: Masculino	0,0786	[0,0772; 0,08]	1,0
Álcool	-0,0353	[-0,0367; -0,0339]	1,0
Exercício Físico	-0,416	[-0,4174; -0,4146]	1,0
Raça/Cor: Outros	0,1278	[0,1233; 0,1324]	1,0
Raça/Cor: Parda	-0,0223	[-0,0239; -0,0209]	1,0
Raça/Cor: Preta	0,2329	[0,2309; 0,2349]	1,0
Fuma	-0,4992	[-0,5016; -0,4967]	1,0
Celular ou Computador	0,0158	[0,0139; 0,0177]	1,0
Escolaridade: 12 anos ou mais	-0,0832	[-0,0853; -0,0811]	1,0
Escolaridade: 9 a 11 anos	0,0171	[0,0153; 0,019]	1,0
Estado Civil: Separado, divorciado ou Viúvo	-0,0892	[-0,0916; -0,0867]	1,0
Estado Civil: Solteiro	-0,2203	[-0,2218; -0,2188]	1,0
Covid	0,1673	[0,1658; 0,1689]	1,0
Plano de Saúde	-0,0611	[-0,0626; -0,0596]	1,0
Depressão	0,3516	[0,3496; 0,3536]	1,0
Legumes ou Verduras	-0,0489	[-0,0505; -0,0473]	1,0
Frutas	-0,0372	[-0,0388; -0,0357]	1,0

Tabela 11: Estimação pontual, estimativa intervalar e valor da estatística  $\hat{R}$  para os parâmetros com distribuição a *priori* Normal com média 0 e variância 0,01

Variável	Parâmetro estimado	Estimativa Intervalar: 95%	$\hat{R}$
Intercepto	-1,5287	[-1,5326; -1,5248]	1,0
Faixa Etária: Adulto	0,7386	[0,7361; 0,741]	1,0
Faixa Etária: Idoso	0,4191	[0,4157; 0,4226]	1,0
Sexo: Masculino	0,0797	[0,0784; 0,0811]	1,0
Álcool	-0,0346	[-0,036; -0,0332]	1,0
Exercício Físico	-0,4172	[-0,4186; -0,4158]	1,0
Raça/Cor: Outros	0,1352	[0,1305; 0,1399]	1,0
Raça/Cor: Parda	-0,0211	[-0,0227; -0,0196]	1,0
Raça/Cor: Preta	0,236	[0,2339; 0,238]	1,0
Fuma	-0,5082	[-0,5107; -0,5057]	1,0
Celular ou Computador	0,0183	[0,0163; 0,0202]	1,0
Escolaridade: 12 anos ou mais	-0,083	[-0,0851; -0,0809]	1,0
Escolaridade: 9 a 11 anos	0,0183	[0,0165; 0,0202]	1,0
Estado Civil: Separado, divorciado ou Viúvo	-0,0921	[-0,0945; -0,0896]	1,0
Estado Civil: Solteiro	-0,2186	[-0,2201; -0,2171]	1,0
Covid	0,168	[0,1666; 0,1695]	1,0
Plano de Saúde	-0,0616	[-0,0631; -0,0601]	1,0
Depressão	0,3559	[0,3539; 0,358]	1,0
Legumes ou Verduras	-0,0495	[-0,0511; -0,0479]	1,0
Frutas	-0,0372	[-0,0387; -0,0357]	1,0



Tabela 12: Estimação pontual, estimativa intervalar e valor da estatística  $\hat{R}$  para os parâmetros com distribuição a *priori* Normal com média 0 e variância 1

Variável	Parâmetro estimado	Estimativa Intervalar: 95%	$\hat{R}$
Intercepto	-1,5289	[-1,5329; -1,525]	1,0
Faixa Etária: Adulto	0,7387	[0,7363; 0,7412]	1,0
Faixa Etária: Idoso	0,4194	[0,4159; 0,4228]	1,0
Sexo: Masculino	0,0797	[0,0783; 0,0811]	1,0
Álcool	-0,0346	[-0,036; -0,0332]	1,0
Exercício Físico	-0,4172	[-0,4186; -0,4157]	1,0
Raça/Cor: Outros	0,1354	[0,1306; 0,14]	1,0
Raça/Cor: Parda	-0,0211	[-0,0226; -0,0196]	1,0
Raça/Cor: Preta	0,236	[0,234; 0,238]	1,0
Fuma	-0,5083	[-0,5107; -0,5057]	1,0
Celular ou Computador	0,0183	[0,0164; 0,0202]	1,0
Escolaridade: 12 anos ou mais	-0,083	[-0,0851; -0,0808]	1,0
Escolaridade: 9 a 11 anos	0,0184	[0,0165; 0,0202]	1,0
Estado Civil: Separado, divorciado ou Viúvo	-0,0921	[-0,0946; -0,0896]	1,0
Estado Civil: Solteiro	-0,2186	[-0,2202; -0,2171]	1,0
Covid	0,168	[0,1665; 0,1695]	1,0
Plano de Saúde	-0,0616	[-0,0631; -0,0601]	1,0
Depressão	0,356	[0,354; 0,358]	1,0
Legumes ou Verduras	-0,0495	[-0,0511; -0,0479]	1,0
Frutas	-0,0372	[-0,0388; -0,0357]	1,0

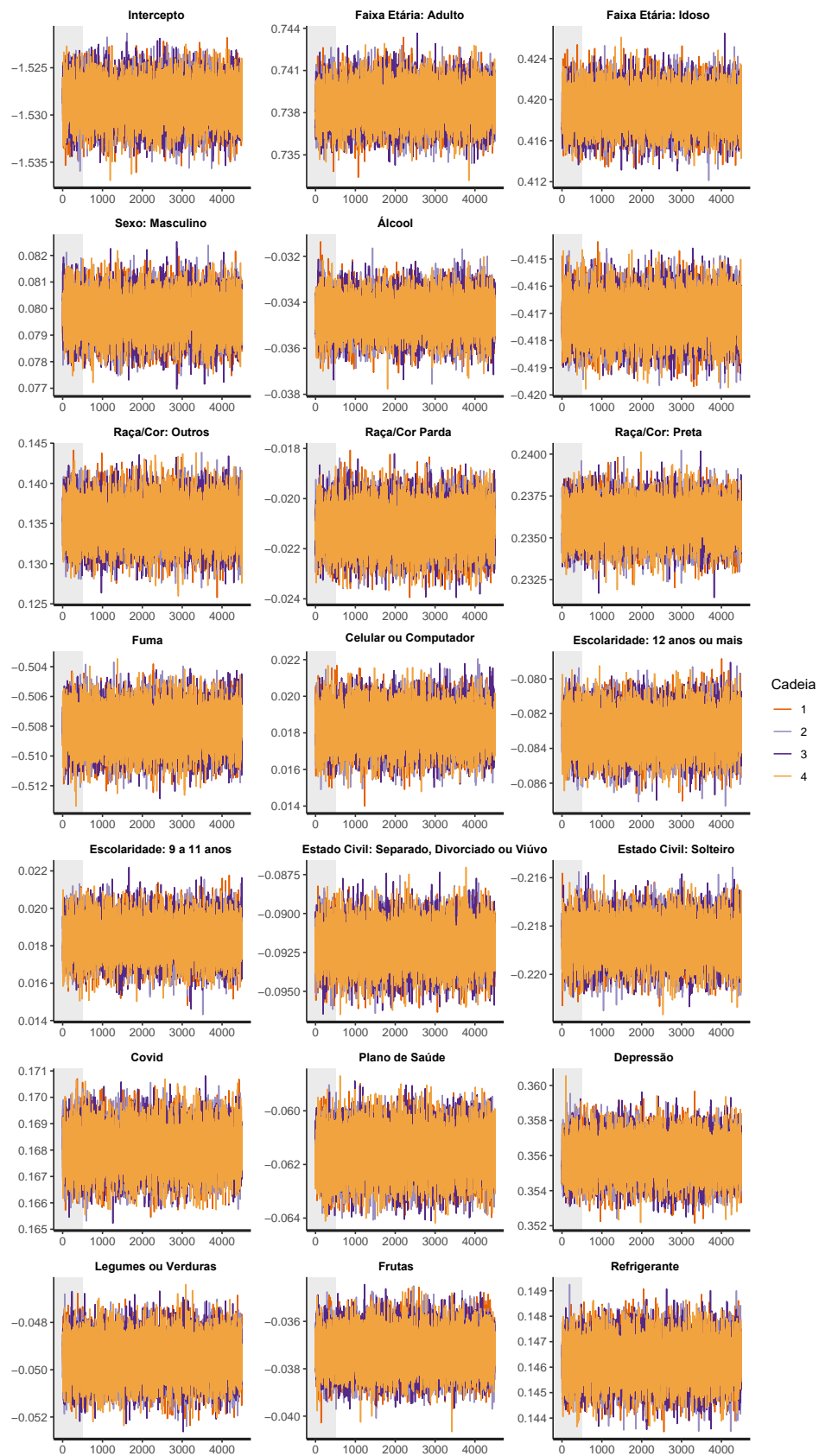


Figura 7: Traço das Cadeias de Markov para o ajuste a *posteriori* dos parâmetros com distribuição a *priori* Normal com média 0 e variância 0,01

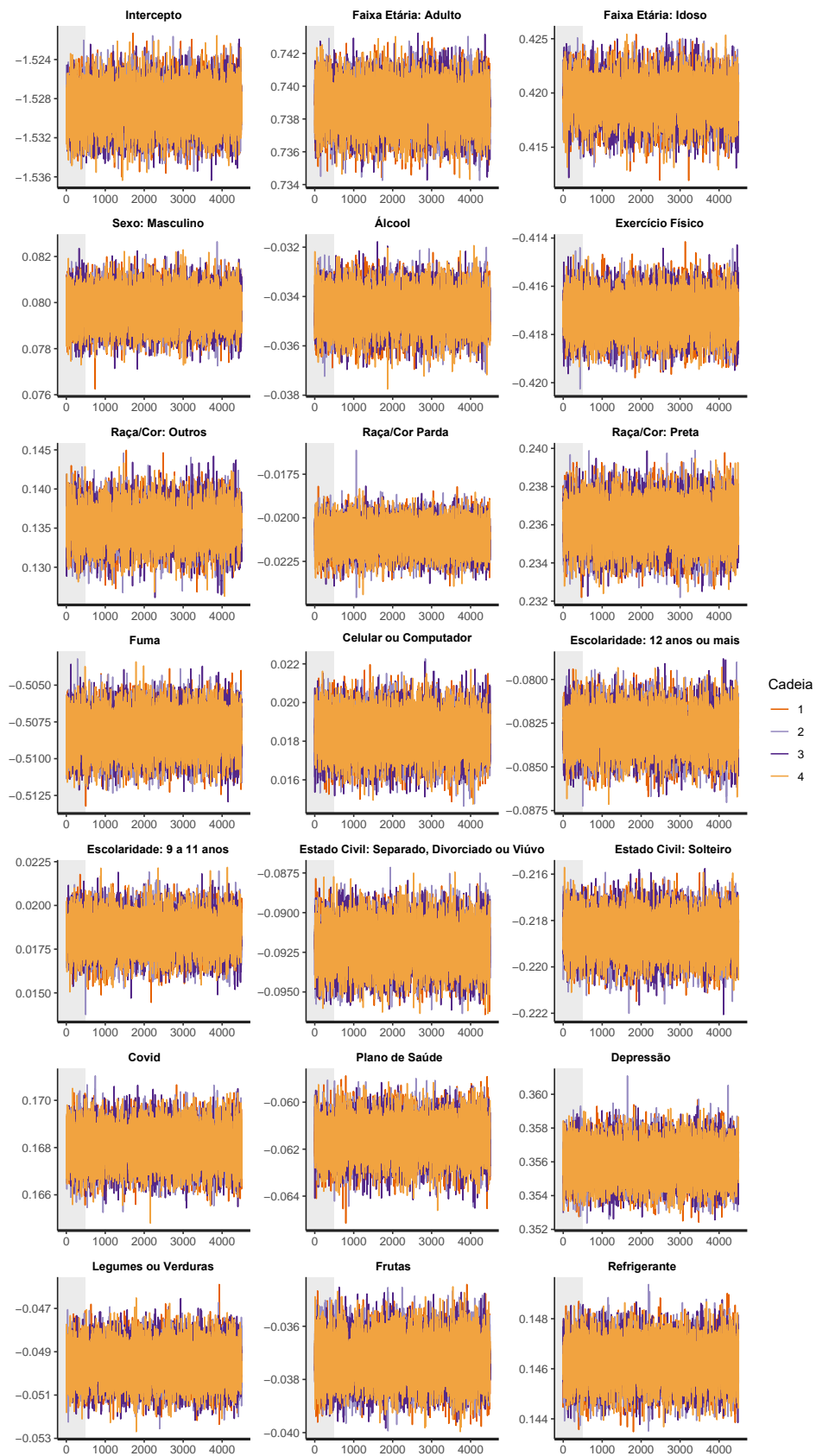


Figura 8: Traço das Cadeias de Markov para o ajuste a *posteriori* dos parâmetros com distribuição a *priori* Normal com média igual a 0 e variância igual a 1

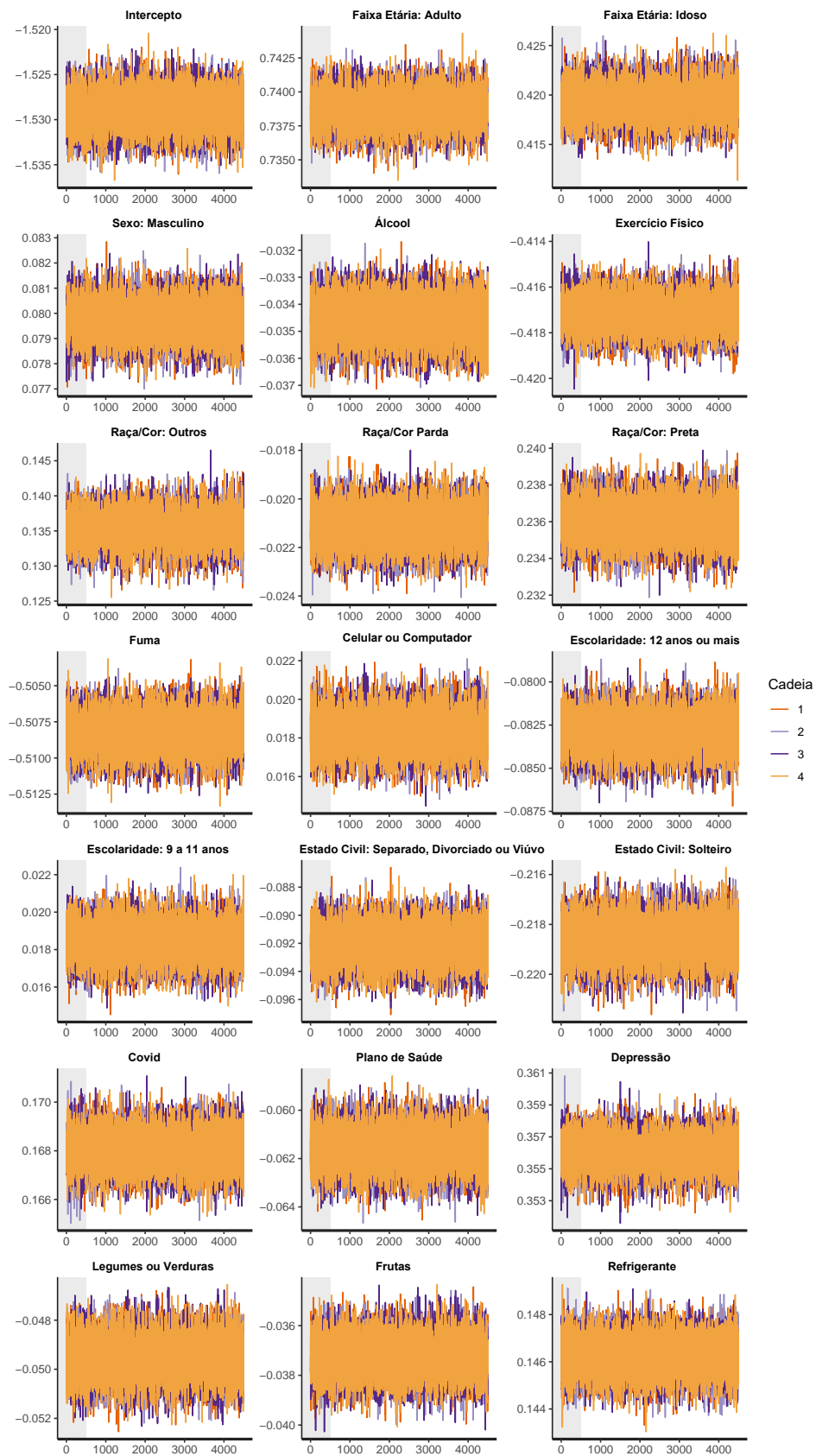


Figura 9: Traço das Cadeias de Markov para o ajuste a *posteriori* dos parâmetros com distribuição a *priori* Normal com média igual a 0 e variância igual a 1000

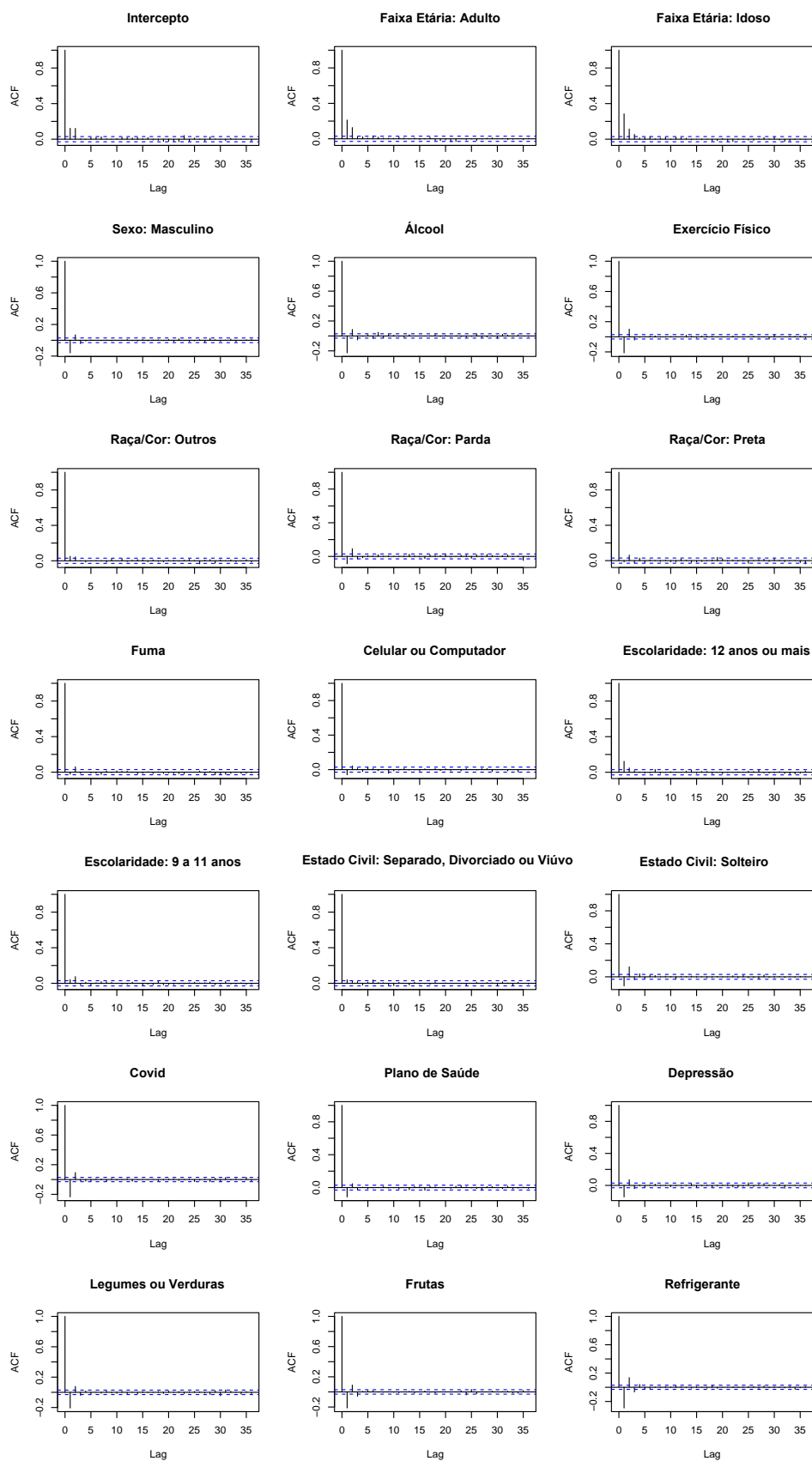


Figura 10: Função de autocorrelação das amostras geradas para o ajuste a *posteriori* dos parâmetros com distribuição a *priori* Normal com média 0 e variância 0,01

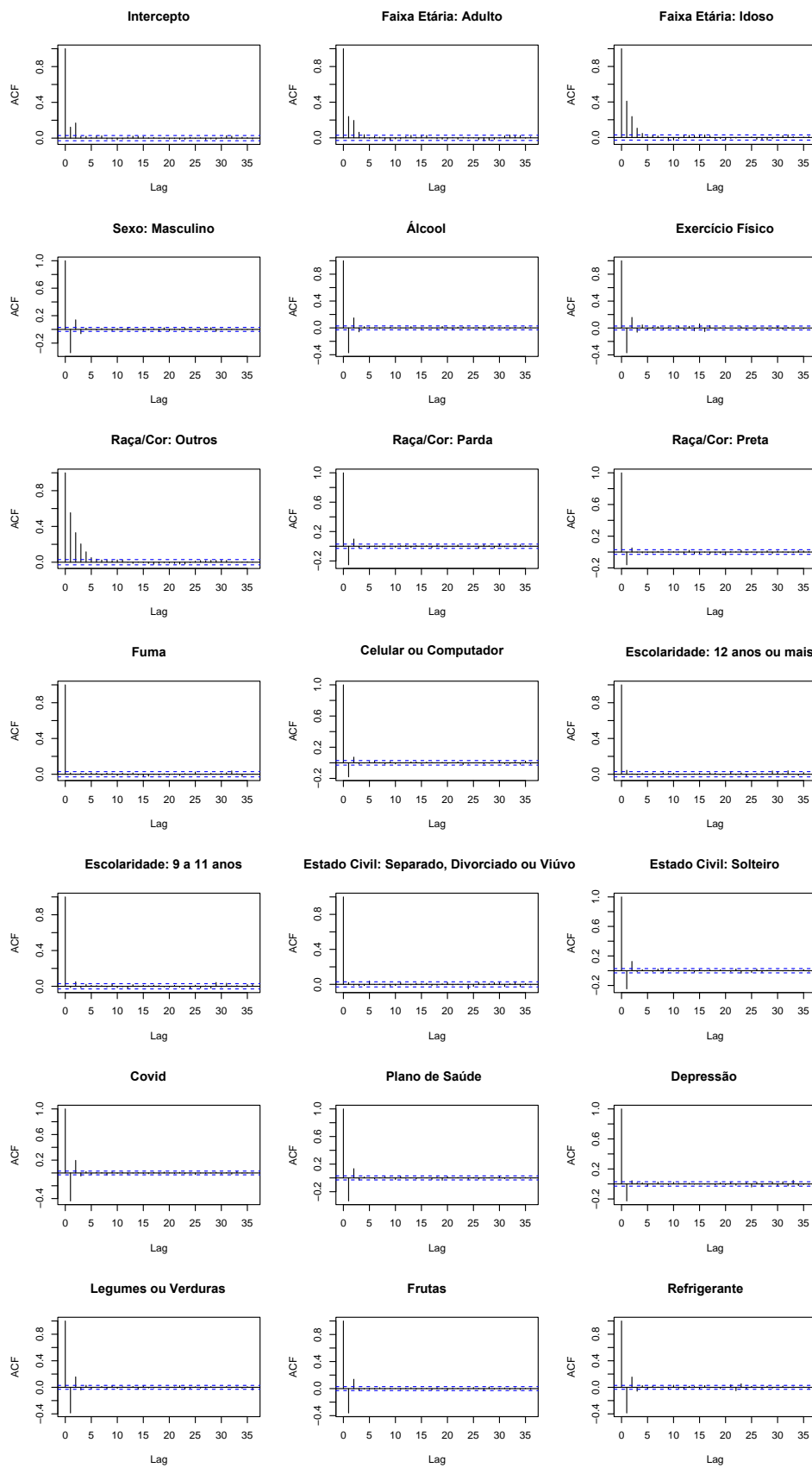


Figura 11: Função de autocorrelação das amostras geradas para o ajuste a *posteriori* dos parâmetros com distribuição a *priori* Normal com média 0 e variância 1

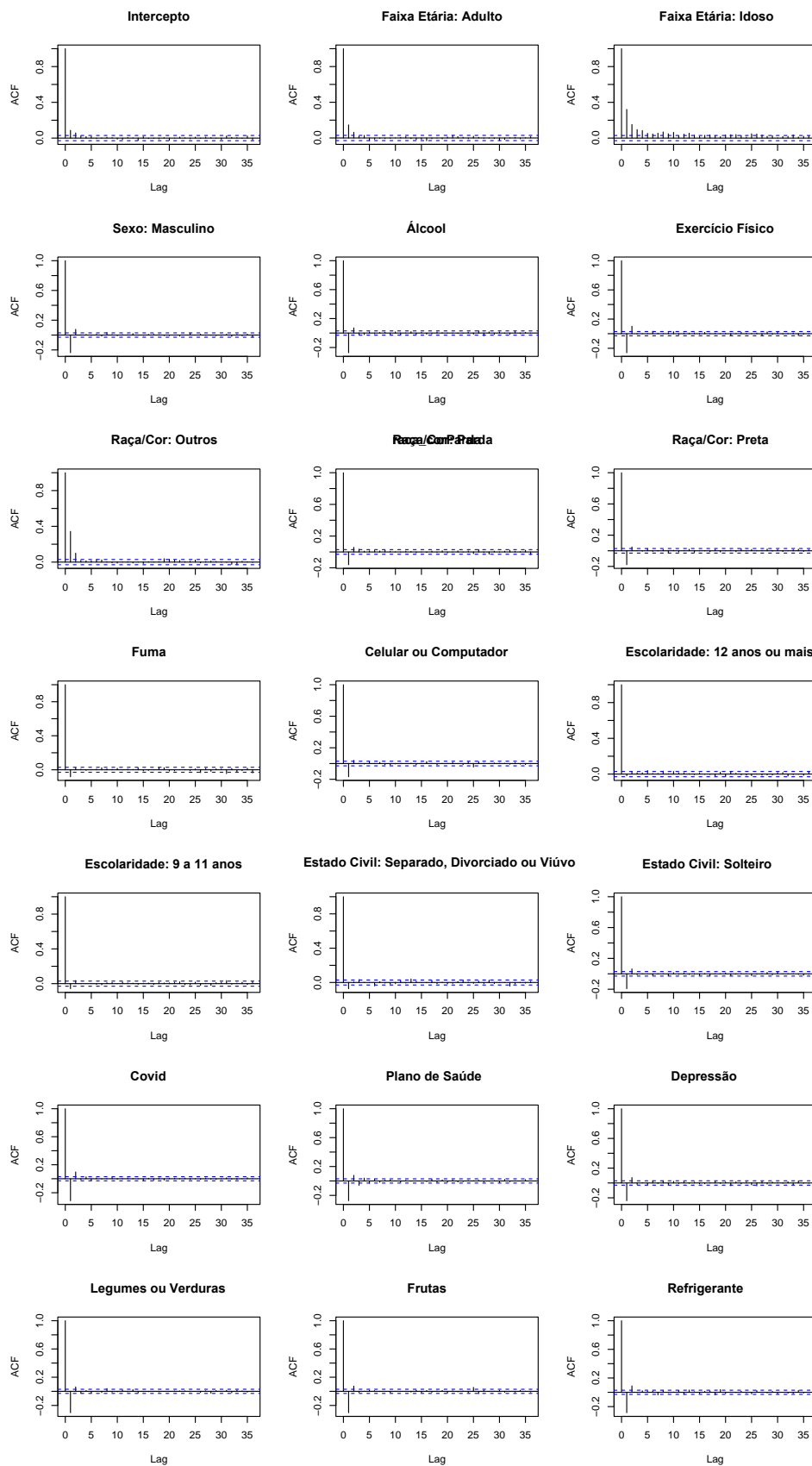


Figura 12: Função de autocorrelação das amostras geradas para o ajuste a *posteriori* dos parâmetros com distribuição a *priori* Normal com média 0 e variância 1000

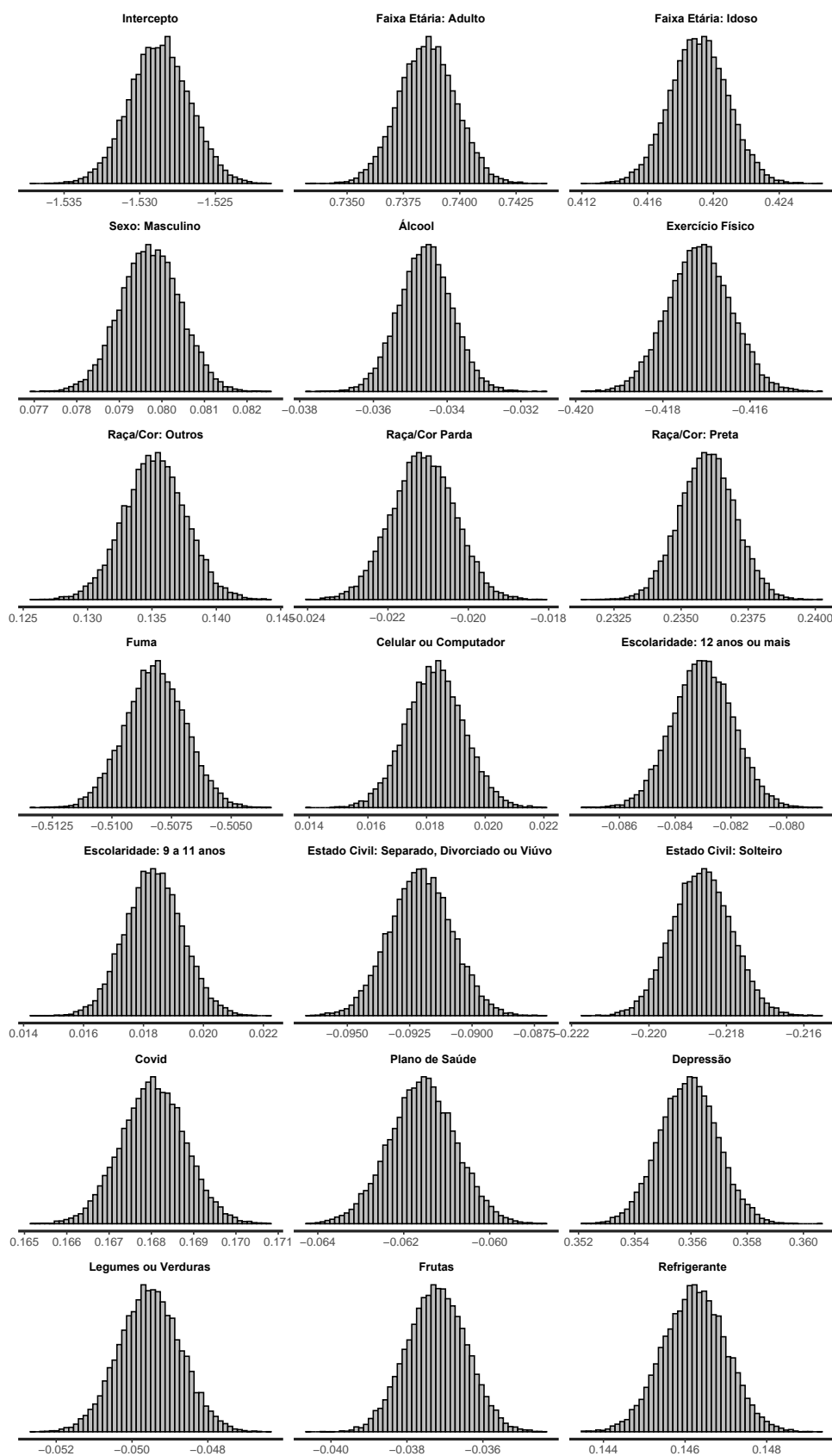


Figura 13: Histograma dos parâmetros estimados para o ajuste a *posteriori* com *priori* Normal com média 0 e variância 0,01



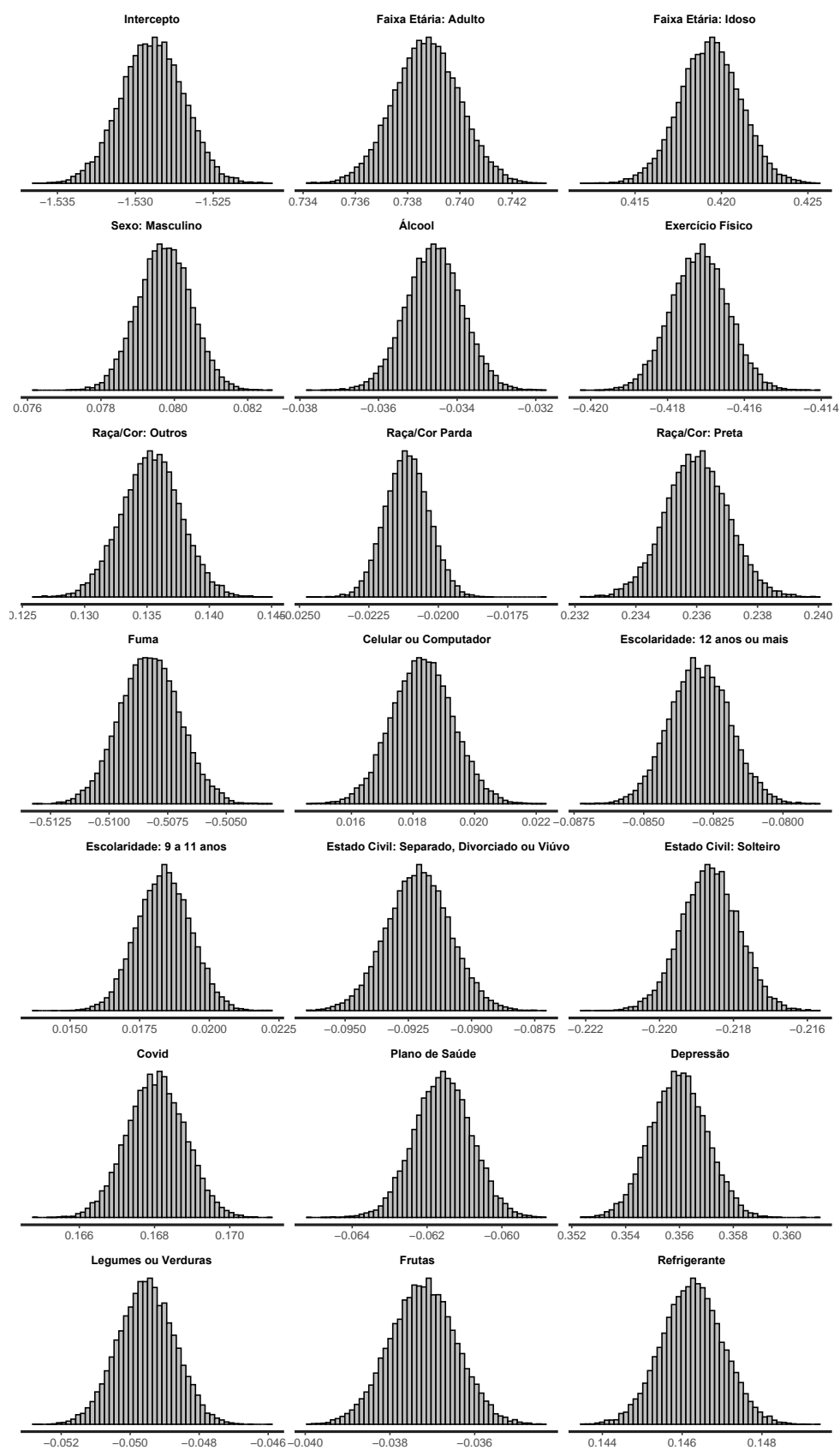


Figura 14: Histograma dos parâmetros estimados para o ajuste a *posteriori* com *priori* Normal com média 0 e variância 1

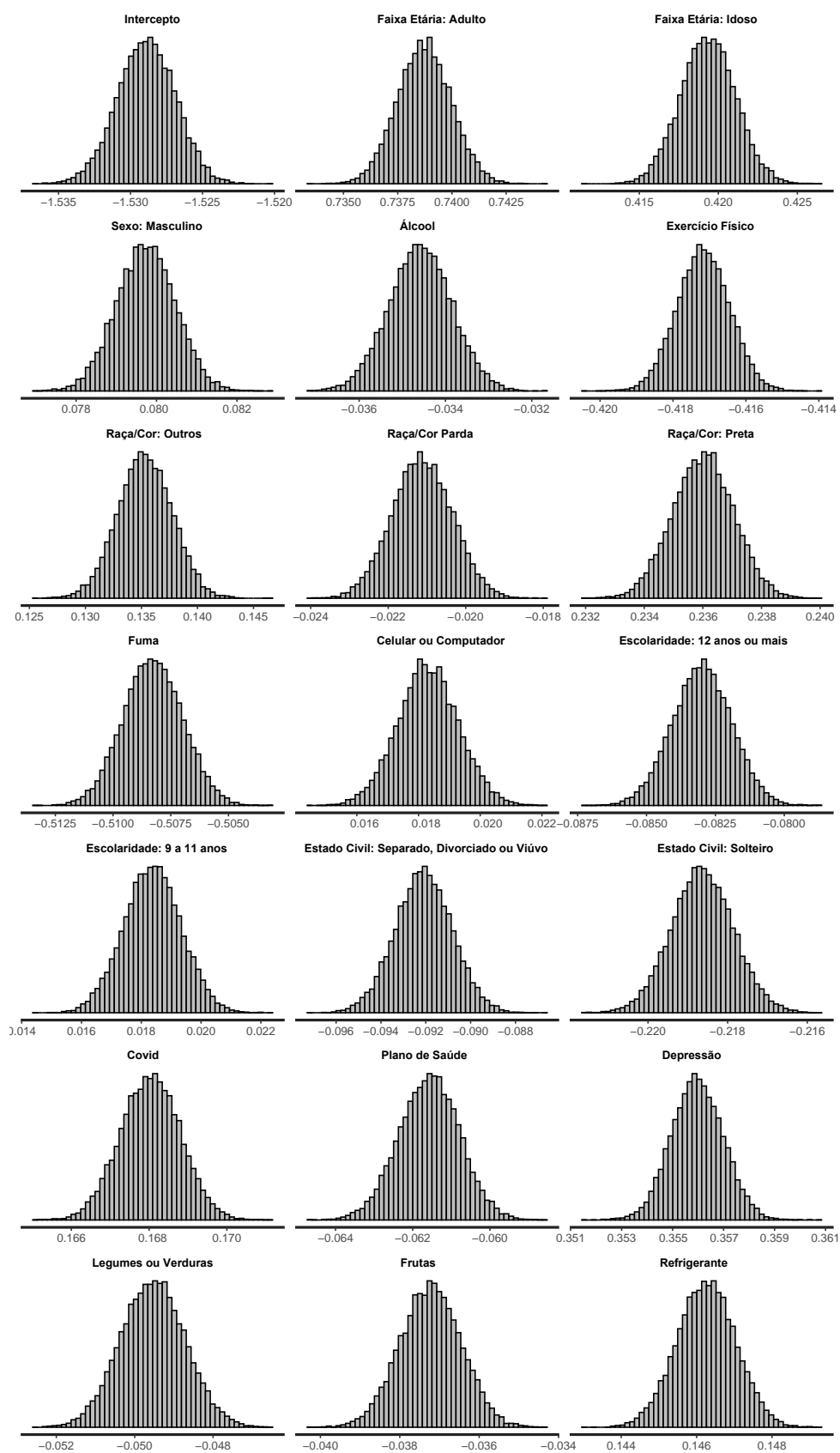


Figura 15: Histograma dos parâmetros estimados para o ajuste a *posteriori* com *priori* Normal com média 0 e variância 1000

A análise da influência dos fatores sociodemográficos e potenciais hábitos relacionados à obesidade, descrita nas Tabelas 13, 14 e 15, foi realizada considerando a razão de chances de cada parâmetro ajustado. Foi observado que a estimativa intervalar para todos os ajustes realizados não contemplou a razão de chances igual a 100%, indicando que todas as variáveis possuem relação positiva ou negativa com a obesidade. Ao se considerar a estimativa intervalar dos três modelos ajustados, foi possível concluir que entre as variáveis sociodemográficas, a idade do indivíduo possui influência positiva com a obesidade, sendo indivíduos adultos com 109% a mais de chance de serem obesos. Já para indivíduos idosos, a chance é aproximadamente 52%. Em relação a Raça/Cor, observa-se que indivíduos pretos possuem maior chance de serem obesos, sendo esta chance igual a aproximadamente 27%, ao se comparar com indivíduos brancos. O diagnóstico de depressão também se mostrou relacionado à obesidade, uma vez que indivíduos deste grupo possuem 43% a mais de chance de serem obesos.

Ao se considerar as variáveis que reduzem a chance de obesidade, foram destacadas a prática de atividade física regular, o hábito de fumar e o estado civil solteiro, de forma que essas variáveis reduzem a chance de obesidade em 34%, 40% e 20% respectivamente. As demais variáveis possuem uma razão de chances próximas de 100%, sendo pouco significativas na relação com a obesidade. Já o intercepto não possui interpretação, sendo desconsiderado na análise.

Tabela 13: Razão de Chances e estimativa intervalar da razão de chances para os parâmetros com distribuição a *priori* Normal com média 0 e variância 0,01

Variável	Razão de Chances %	Estimativa Intervalar: 95%
Intercepto	-78,32	[-78,40; -78,23]
Faixa Etária: Adulto	109,29	[108,78; 109,80]
Faixa Etária: Idoso	52,06	[51,54; 52,59]
Sexo: Masculino	8,30	[8,15; 8,45]
Álcool	-3,40	[-3,54; -3,26]
Exercício Físico	-34,11	[-34,20; -34,02]
Raça/Cor: Outros	14,48	[13,94; 15,01]
Raça/Cor: Parda	-2,09	[-2,24; -1,94]
Raça/Cor: Preta	26,61	[26,36; 26,87]
Fuma	-39,84	[-39,99; -39,69]
Celular ou Computador	1,84	[1,65; 2,04]
Escolaridade: 12 anos ou mais	-7,97	[-8,16; -7,77]
Escolaridade: 9 a 11 anos	1,85	[1,67; 2,04]
Estado Civil: Separado, divorciado ou Viúvo	-8,79	[-9,02; -8,57]
Estado Civil: Solteiro	-19,64	[-19,76; -19,52]
Covid	18,30	[18,12; 18,48]
Plano de Saúde	-5,97	[-6,12; -5,83]
Depressão	42,75	[42,46; 43,04]
Legumes ou Verduras	-4,83	[-4,98; -4,68]
Frutas	-3,65	[-3,80; -3,51]
Refrigerante	15,75	[15,57; 15,92]

Tabela 14: Razão de Chances e estimativa intervalar da razão de chances para os parâmetros com distribuição a *priori* Normal com média 0 e variância 1

Variável	Razão de Chances %	Estimativa Intervalar: 95%
Intercepto	-78,32	[-78,41; -78,24]
Faixa Etária: Adulto	109,33	[108,81; 109,85]
Faixa Etária: Idoso	52,10	[51,57; 52,62]
Sexo: Masculino	8,30	[8,15; 8,45]
Álcool	-3,40	[-3,54; -3,26]
Exercício Físico	-34,11	[-34,20; -34,01]
Raça/Cor: Outros	14,49	[13,95; 15,02]
Raça/Cor: Parda	-2,09	[-2,24; -1,94]
Raça/Cor: Preta	26,62	[26,36; 26,88]
Fuma	-39,85	-[39,99; -39,69]
Celular ou Computador	1,85	[1,65; 2,05]
Escolaridade: 12 anos ou mais	-7,97	[-8,16; -7,77]
Escolaridade: 9 a 11 anos	1,85	[1,67; 2,04]
Estado Civil: Separado, divorciado ou Viúvo	-8,80	[-9,02; -8,57]
Estado Civil: Solteiro	-19,64	[-19,76; -19,52]
Covid	18,30	[18,12; 18,47]
Plano de Saúde	-5,97	[-6,11; -5,83]
Depressão	42,76	[42,47; 43,05]
Legumes ou Verduras	-4,83	[-4,99; -4,68]
Frutas	-3,65	[-3,80; -3,51]
Refrigerante	15,75	[15,58; 15,92]

Tabela 15: Razão de Chances e estimativa intervalar da razão de chances para os parâmetros com distribuição a *priori* Normal com média 0 e variância 1000

Variável	Razão de Chances %	Estimativa Intervalar: 95%
Intercepto	-78,32	[-78,41; -78,24]
Faixa Etária: Adulto	109,33	[108,82; 109,83]
Faixa Etária: Idoso	52,10	[51,58; 52,62]
Sexo: Masculino	8,30	[8,15; 8,45]
Álcool	-3,40	[-3,54; -3,26]
Exercício Físico	-34,11	[-34,20; -34,02]
Raça/Cor: Outros	14,49	[13,95; 15,03]
Raça/Cor: Parda	-2,09	[-2,24; -1,94]
Raça/Cor: Preta	26,62	[26,36; 26,88]
Fuma	-39,85	[-40,00; -39,70]
Celular ou Computador	1,85	[1,65; 2,05]
Escolaridade: 12 anos ou mais	-7,96	[-8,16; -7,77]
Escolaridade: 9 a 11 anos	1,85	[1,66; 2,04]
Estado Civil: Separado, divorciado ou Viúvo	-8,80	[-9,03; -8,57]
Estado Civil: Solteiro	-19,64	[-19,76; -19,51]
Covid	18,30	[18,12; 18,48]
Plano de Saúde	-5,97	[-6,12; -5,83]
Depressão	42,76	[42,48; 43,04]
Legumes ou Verduras	-4,83	[-4,99; -4,68]
Frutas	-3,65	[-3,80; -3,51]
Refrigerante	15,75	[15,57; 15,92]

## 4 Conclusões

O presente trabalho teve como objetivo analisar e identificar potenciais fatores sociodemográficos e hábitos relacionados à presença de obesidade em indivíduos residentes nas capitais brasileiras e no Distrito Federal, fazendo uso da base de dados disponibilizada pelo VIGITEL no ano de 2021 e no primeiro semestre de 2023. Para isso, foram realizados ajustes de modelos de regressão logística sob as abordagens clássicas e Bayesianas. Para a abordagem Bayesiana, foram consideradas distribuições a priori com variâncias distintas, de forma a observar se a significância da distribuição a priori iria influenciar significativamente no ajuste.

De maneira geral, os ajustes refletiram o que havia sido analisado descritivamente, indicando que as variáveis com maior proporção de obesos foram significativas na equação dos modelos de regressão logística ajustados por ambas visões. Algumas variáveis com grande percentual de obesos não se mostraram significativas nos ajustes, como a Raça/cor: Outros, que pode ser justificada por conta do tamanho amostral pequeno. Já variáveis como Faixa etária, Estado civil e Prática de exercícios físicos se mostraram fortemente relacionadas à obesidade, sendo destas faixas as maiores proporções de obesos.

Em relação aos ajustes sob as abordagens clássica e Bayesiana, observou-se aproximação dos valores estimados de cada parâmetro, sendo um forte indicativo de que as distribuições possuíam pouca influência na distribuição a priori, mesmo no casos de *prioris* muito significativas, podendo ser justificado devido ao tamanho da amostra ponderada disponível. Diferente do ajuste clássico, o ajuste Bayesiano considerou todas as variáveis do modelo, por considerar intervalos de credibilidade com pouca amplitude. Apesar disso, as variáveis descartadas no modelo clássico apresentaram pouca contribuição com o diagnóstico dos modelos Bayesianos, uma vez que a razão de chances destas variáveis tiveram valores próximos de 1. Por conta da similaridade da estimação dos parâmetros, a razão de chances para cada variável se mostrou muito próxima.

Identificou-se que dentre as variáveis sociodemográficas, os indivíduos adultos são aqueles com maior chance de serem obesos, seguidos de indivíduos de cor preta. Entre as demais variáveis, observou-se que o diagnóstico de depressão possui influência no aumento da chance de obesidade. A prática de atividade física se mostrou eficaz no combate à obesidade, seguida do hábito de fumar e do estado civil solteiro. Já as outras variáveis consideradas tiveram pouca relação com a obesidade em si, não tendo sido consideradas relevantes para a análise.

A relação entre as variáveis analisadas e a presença de obesidade é bem ampla, podendo estar relacionada a diversos fatores sociais e econômicos para os grupos pertencentes às variáveis relacionadas. Uma revisão literária em torno destes fatores se faz necessária, para melhor entendimento da relação destes grupos com a obesidade, auxiliando na justificativa do resultado apresentado por este trabalho.



# Referências

- AGRESTI, A. *Foundations of Linear and Generalized Linear Models*. [S.l.]: John Wiley & Sons, Inc, 2015.
- AKAIKE, H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, v. 19, n. 6, p. 716–723, 1974.
- BERNAL, R. T. I. et al. Sistema de vigilância de fatores de risco e proteção para doenças crônicas por inquérito telefônico (vigitel): mudança na metodologia de ponderação. *Epidemiologia e Serviços de Saúde*, Secretaria de Vigilância em Saúde e Ambiente - Ministério da Saúde do Brasil, v. 26, n. 4, p. 701 – 712, 10 2017. Disponível em: <https://doi.org/10.5123/S1679-49742017000400003>.
- CASELLA, G.; GEORGE, E. I. Explaining the gibbs sampler. *The American Statistician*, [American Statistical Association, Taylor & Francis, Ltd.], v. 46, n. 3, p. 167–174, 1992. ISSN 00031305. Disponível em: <http://www.jstor.org/stable/2685208>.
- CHARTRAND, R.; YIN, W. Iteratively reweighted algorithms for compressive sensing. In: *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*. [S.l.: s.n.], 2008. p. 3869–3872.
- DOBSON, A. J.; BARNETT, A. G. *An Introduction to Generalized Linear Models*. [S.l.]: Chapman and Hall/CRC, 2008.
- DUANE, S. et al. Hybrid monte carlo. *Physics Letters B*, v. 195, n. 2, p. 216–222, 1987. ISSN 0370-2693. Disponível em: <https://www.sciencedirect.com/science/article/pii/037026938791197X>.
- EHLERS, R. S. *Introdução a Inferência Bayesiana*. [S.l.: s.n.], 2003.
- FERREIRA, A. P. d. S.; SZWARCOWALD, C. L.; DAMACENA, G. N. Prevalência e fatores associados da obesidade na população brasileira: estudo com dados aferidos da pesquisa nacional de saúde, 2013. *Revista Brasileira de Epidemiologia*, Associação Brasileira de Saúde Coletiva, v. 22, p. e190024, 2019. ISSN 1415-790X. Disponível em: <https://doi.org/10.1590/1980-549720190024>.
- GELMAN, A.; RUBIN, D. B. Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science*, Institute of Mathematical Statistics, v. 7, n. 4, p. 457 – 472, 1992. Disponível em: <https://doi.org/10.1214/ss/1177011136>.
- HITCHCOCK, D. B. A history of the metropolis-hastings algorithm. *The American Statistician*, [American Statistical Association, Taylor & Francis, Ltd.], v. 57, n. 4, p. 254–257, 2003. ISSN 00031305. Disponível em: <http://www.jstor.org/stable/30037292>.

- Ministério da Saúde. *Vigitel Brasil 2006-2021 : vigilância de fatores de risco e proteção para doenças crônicas por inquérito telefônico : estimativas sobre frequência e distribuição sociodemográfica do estado nutricional e consumo alimentar nas capitais dos 26 estados brasileiros e no Distrito Federal entre 2006 e 2021 : estado nutricional e consumo alimentar*. Ministério da Saúde, 2022. Disponível em: [http://bvsms.saude.gov.br/bvs/publicacoes/vigitel\\_brasil\\_2006-2021\\_estado\\_nutricional.pdf](http://bvsms.saude.gov.br/bvs/publicacoes/vigitel_brasil_2006-2021_estado_nutricional.pdf).
- Ministério da Saúde. *Vigitel - Ministério da Saúde*. 2023. Disponível em: <https://www.gov.br/saude/pt-br/assuntos/saude-de-a-a-z/v/vigitel#:~:text=O%20Vigitel%20faz%20parte%20das,brasileiros%20e%20no%20Distrito%20Federal.> Acesso em: 01/06/2023.
- MORETTIN, J. d. M. S. P. A. *Estatística e Ciência de Dados*. [S.l.]: Livros Técnicos e Científicos Editora Ltda, 2022.
- NELDER, J. A.; WEDDERBURN, R. W. M. *Generalized Linear Models*. [Royal Statistical Society, Wiley], 1972. v. 135. Disponível em: <http://www.jstor.org/stable/2344614>.
- PARK, E. J. et al. The effect of alcohol drinking on metabolic syndrome and obesity in koreans: Big data analysis. *International Journal of Environmental Research and Public Health*, 2022. Disponível em: <https://www.mdpi.com/1660-4601/19/9/4949>.
- PESSOA, D. G. C.; SILVA, P. L. d. N. *Análise de Dados Amostrais Complexos*. São Paulo: Associação Brasileira de Estatística (ABE), 1998.
- PÉREZ, F. L. *Machine Learning, Monte Carlo Methods*. 2021. [http://leg.ufpr.br/~lucambio/CE225/20212S/MLG\\_Monte\\_Carlo.html](http://leg.ufpr.br/~lucambio/CE225/20212S/MLG_Monte_Carlo.html).
- R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2023. Disponível em: <https://www.R-project.org/>.
- REZENDE, L. F. M. et al. A epidemia de obesidade e as dcnt – causas, custos e sobrecarga no sus. 2022. Acesso em: 02/07/2023.
- ROBERTS, G. O.; ROSENTHAL, J. S. Optimal scaling for various Metropolis-Hastings algorithms. *Statistical Science*, Institute of Mathematical Statistics, p. 351 – 367, 2001. Disponível em: <https://doi.org/10.1214/ss/1015346320>.
- SHELTON, N. J.; KNOTT, C. S. *Association between alcohol calorie intake and overweight and obesity in English adults*. *Am J Public Health*. 2014. Disponível em: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4025698/>.
- SILVA, K. S.; PADILHA, L. L. Perfil nutricional e fatores associados em adultos: análise de dados do inquérito telefônico vigitel (2020). *RBONE - Revista Brasileira de Obesidade, Nutrição e Emagrecimento*, v. 16, n. 105, p. 1060–1074, jan. 2023. Disponível em: <http://www.rbone.com.br/index.php/rbone/article/view/2181>.
- SILVA, L. E. S. d. et al. Tendência temporal da prevalência do excesso de peso e obesidade na população adulta brasileira, segundo características sociodemográficas, 2006-2019. *Epidemiologia e Serviços de Saúde*, Secretaria de Vigilância em Saúde e Ambiente - Ministério da Saúde do Brasil, v. 30, n. 1, p. e2020294, 2021. ISSN 2237-9622. Disponível em: <https://doi.org/10.1590/S1679-49742021000100008>.

- SPIEGELHALTER, D. J. et al. Bayesian Measures of Model Complexity and Fit. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, v. 64, n. 4, p. 583–639, 10 2002. Disponível em: <https://doi.org/10.1111/1467-9868.00353>.
- STREB, A. R. et al. Simultaneidade de comportamentos de risco para a obesidade em adultos das capitais do brasil. *Ciência & Saúde Coletiva*, ABRASCO - Associação Brasileira de Saúde Coletiva, v. 25, n. 8, p. 2999–3007, Aug 2020. ISSN 1413-8123. Disponível em: <https://doi.org/10.1590/1413-81232020258.27752018>.
- THOMAS, C. et al. *10 Years On: New evidence on TV marketing and junk food eating amongst 11-19 year olds 10 years after broadcast regulations*. 2018. Disponível em: [https://www.cancerresearchuk.org/sites/default/files/10\\_years\\_on\\_full\\_report.pdf](https://www.cancerresearchuk.org/sites/default/files/10_years_on_full_report.pdf).
- THOMAS, S.; TU, W. Learning hamiltonian monte carlo in r. *The American Statistician*, v. 75, n. 4, p. 403–413, 2021.
- WALD, A. *Tests of Statistical Hypotheses Concerning Several Parameters When the Number of Observations Is Large*. Transactions of the American Mathematical Societ, 1943. Disponível em: <https://doi.org/10.1090/S0002-9947-1943-0012401-3>.
- WOOD, S. N. *Generalized Additive Models: An Introduction with R*. [S.l.]: Chapman and Hall/CRC, 2006.
- World Health Organization. *Obesity and overweight*. 2021. Disponível em: <https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight>.
- World Health Organization. *World Obesity Day 2022 – Accelerating action to stop obesity*. 2022. Disponível em: <https://www.who.int/news/item/04-03-2022-world-obesity-day-2022-accelerating-action-to-stop-obesity#:~:text=More%20than%201%20billion%20people,they%20are%20overweight%20or%20obese>.