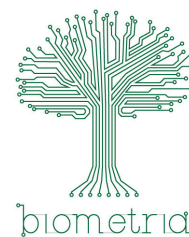




Universidade Estadual Paulista “Júlio de Mesquita Filho”
Instituto de Biociências – Câmpus de Botucatu
Programa de Pós-graduação em Biometria



Modelagem conjunta e univariada: Uma aplicação para dados longitudinais de crescimento infantil

Elizabete de Jesus Pinto

Botucatu
2022

Elizabete de Jesus Pinto

Modelagem conjunta e univariada: Uma aplicação para dados longitudinais de crescimento infantil

Tese de Doutorado apresentada ao Curso de Programa de Pós-graduação em Biometria da Universidade Estadual Paulista “Júlio de Mesquita Filho” como parte dos requisitos necessários para a obtenção do título de Doutora em Biometria.

Orientador: Profa. Dra. Miriam Harumi Tsunemi

Botucatu
2022

FICHA CATALOGRÁFICA ELABORADA PELA SEÇÃO TÉC. AQUIS. TRATAMENTO DA INFORM.
DIVISÃO TÉCNICA DE BIBLIOTECA E DOCUMENTAÇÃO - CÂMPUS DE BOTUCATU - UNESP
BIBLIOTECÁRIA RESPONSÁVEL: ROSANGELA APARECIDA LOBO-CRB 8/7500

Pinto, Elizabete de Jesus.

Modelagem conjunta e univariada : uma aplicação para dados longitudinais de crescimento infantil / Elizabete de Jesus Pinto. - Botucatu, 2022

Tese (doutorado) - Universidade Estadual Paulista "Júlio de Mesquita Filho", Instituto de Biociências de Botucatu

Orientador: Miriam Harumi Tsunemi

Capes: 10203001

1. Estudos longitudinais. 2. Modelos lineares (Estatística). 3. Crianças - Crescimento. 4. Cópulas (Estatística matemática)

Palavras-chave: Brokenstick; Cópula bivariada; Estudo longitudinal; Modelo linear misto; SITAR.

ELIZABETE DE JESUS PINTO

**MODELAGEM CONJUNTA E UNIVARIADA: UMA APLICAÇÃO PARA
DADOS LONGITUDINAIS DE CRESCIMENTO INFANTIL**

Tese de doutorado defendida e aprovada em Botucatu-SP, no dia 01 de agosto, pela banca examinadora constituída pelos professores:

Profa. Dra. Miriam Harumi Tsunemi
Orientador

Profa. Dra. Liciane Vaz de Arruda Silveira
Examinador

Profa. Dra. Lídia Raquel de Carvalho
Examinador

Profa. Dra. Rosemeire Leovigildo Fiaccone
Examinador

Prof. Dr. Maurício Lima Barreto
Examinador

Dedicatória

A meus pais, Augusto e Helena, pelo amor e apoio incondicional.

A minha irmã Tânia, pela amizade e incentivo em todos os momentos.

Agradecimentos

A redação desta tese é fruto de esforço que é, ao mesmo tempo, solitário e coletivo. Assumindo a total responsabilidade pelas falhas nele contidas, expresso aqui o meu reconhecimento por toda a colaboração recebida.

Primeiro agradeço a Deus, o Caminho, a Verdade e a Vida. Agradeço pela Sua presença constante em minha vida, por me guiar e orientar. A Ele a minha eterna gratidão e fé!

A meus pais, Augusto e Helena, agradeço pelo incentivo, apoio e orações. Vocês são meus exemplos de vida, que apesar da pouca instrução, sempre me ensinaram valores importantes na minha formação pessoal e profissional.

As minhas queridas irmãs Tânia e Sandra pela amizade, incentivo e apoio na concretização desse sonho e aos meus irmãos José, Carlos, Ana, Vera e Ana Lúcia pelo carinho e apoio. Aos meus sobrinhos Yuri, Marília, Victor, Vanessa e Maria Clara por aquecerem minha vida de alegrias e, nos momentos de ausência a lembrança de cada sorriso me deu força para continuar. Aos meus cunhados agradeço pelo apoio e compreensão.

A Profa. Dra. Miriam Harumi Tsunemi, agradeço por sua orientação, confiança, apoio, incentivo, compreensão, motivação, amizade e contribuições ao longo da caminhada.

A Profa. Dra. Ana Marlúcia Oliveira (*in memoriam*) pelo carinho, amizade, contribuições, sugestões, pelas palavras de incentivo em meus momentos de angústias, medos e inquietações e por ter sempre uma palavra de afeto, encorajamento e sabedoria. Meu sincero agradecimento pela preciosa contribuição na minha formação. Obrigado por tudo!

A Profa. Dra. Rosemeire Leogevido Fiaccone, por sempre contribuir na minha trajetória acadêmica, pela amizade, confiança e motivação fundamentais para construir o conhecimento indispensável na conclusão desse trabalho.

Agradeço aos amigos Juliana, Bethina, Gustavo Tomanik, Thiago, Felipe Teles, Renato, Carlos e Felipe Camargo, pelo companheirismo, momentos especiais dentro e fora da Universidade. Vocês fizeram toda diferença! Aos amigos Mariese, Michele e Gustavo pela cooperação, pela troca de experiências, por compartilharem comigo momentos de angustias e alegrias.

Ao Prof. Dr. Maurício Barreto, Dra. Maria Yury e toda a equipe do Centro de Integração de Dados e Conhecimentos para Saúde (Cidacs/IGM/Fiocruz) pelo acolhimento e pela confiança no desenvolvimento deste estudo.

Ao Grupo de Estudos em Nutrição do CIDACS (GeNut-CIDACS), em especial a Profa. Dra. Rita Ribeiro da Silva, por sua valiosa amizade, pela confiança, por acreditar que daria certo, sua ajuda foi simplesmente essencial. Aos colegas Natanael, Juliana e Carolina pela troca de experiências e valiosas contribuições para o avanço deste trabalho.

Aos professores e equipe técnica do Programa de Pós-Graduação em Biometria, pelos ensinamentos, convivência e todo suporte.

À banca examinadora que tão gentilmente aceitaram participar e colaborar com esta tese.

Ao Prof. Dr. Paulo Henrique Ferreira da Silva por sua disponibilidade, contribuições e apoio.

Ao Centro de Ciências da Saúde da Universidade Federal do Recôncavo da Bahia pela liberação das atividades docentes durante o doutorado e por todo apoio.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - CAPES.

A todos que ajudaram de forma direta ou indireta na minha formação e compartilham dessa etapa da minha vida.

Resumo

O crescimento infantil é um importante indicador de saúde desde a primeira infância até a puberdade, pois pode determinar o estado de saúde na idade adulta. O monitoramento das medidas de peso e altura pode auxiliar nas ações da gestão de políticas públicas para a prevenção da desnutrição e obesidade, voltada à promoção da saúde e nutrição adequada ao crescimento na infância. O acompanhamento dessas medidas antropométricas, através de medidas coletadas repetidamente no mesmo indivíduo, torna possível responder as questões relativas às mudanças que ocorrem na dinâmica do crescimento, como a trajetória do crescimento de cada indivíduo ao longo do tempo e as diferenças que ocorrem no crescimento médio entre o grupo de indivíduos. Entendendo que o monitoramento dessas medidas são importantes índices de saúde na infância e a forma mais adequada de avaliar seu estado nutricional e de saúde geral, o objetivo principal desta tese foi estimar trajetórias de crescimento com modelagem conjunta e univariada (marginal) para as medidas antropométricas de peso e altura, na população menor de dois anos de idade de uma coorte eletrônica de base populacional linkados com dados de rotina de nascidos vivos (SINASC) e do Sistema de Vigilância Alimentar e Nutricional (SISVAN), entre os anos 2008 a 2017. Foram utilizadas duas abordagens para lidar com dados longitudinais multivariados. Na primeira abordagem sugerimos uma comparação de modelos comumente usados para caracterizar as trajetórias de crescimento infantil como os Modelos de Efeitos Mistos, *Brokenstick* e o SITAR. Na segunda abordagem, o principal interesse foi estudar a evolução conjunta de duas variáveis respostas para o mesmo indivíduo, que exige métodos mais complexos e desafiadores, pois além da existência da dependência entre as medidas repetidas no mesmo indivíduo considera-se a correlação entre as variáveis respostas que pode mudar com o tempo e, além disso, a correlação natural existente em dados longitudinais. Dessa forma, explorou-se a modelagem via funções de cópulas bivariadas, com intuito de melhorar a modelagem de dados bivariados com formas mais flexíveis de se estimar a dependência no padrão de crescimento infantil.

Palavras-chave: Crescimento infantil, Estudo Longitudinal, Modelo Linear Misto, *Brokenstick*, SITAR, Cópula Bivariada.

Abstract

Child growth is an important health indicator from early childhood to puberty, as childhood can determine health status in adulthood. The monitoring of weight and height measurements can help public policies management actions to prevent malnutrition, obesity, and help promoting adequate nutrition for childhood growth. The monitoring of these anthropometric measurements repeatedly in the same individual along time makes it possible to answer questions regarding the changes that occur in the dynamics of growth, such as the trajectory of each individual's growth over time and the differences that can occur in the average growth between the group of individuals. Understanding that monitoring these measures are important indices of childhood health and the most appropriate way to assess their nutritional status and general health, the main objective of this thesis was to estimate growth trajectories with joint and univariate (marginal) modeling for anthropometric measurements of weight and height, in the population under two years of age of a population-based electronic cohort linked with routine live birth data (SINASC) and the Food and Nutrition Surveillance System (SISVAN), between the years 2008 to 2017. Two approaches were used to deal with multivariate longitudinal data. In the first approach, we suggest a comparison of models commonly used to characterize child growth trajectories, such as the Mixed Effects Models, *Brokenstick* and SITAR. In the second approach, the main interest was to study the joint evolution of two response variables for the same individual, which requires more complex and challenging methods, as the existence of dependence between repeated measurements in the same individual, the correlation between the variables is considered responses that can change over time and, in addition, the natural correlation existing in longitudinal data. Thus, modeling via bivariate copula functions was explored, in order to improve the modeling of bivariate data with more flexible ways of estimating dependence on the pattern of child growth.

Keywords: Child growth, Longitudinal Study, Mixed Linear Model, *Brokenstick*, SITAR, Bivariate Copula.

Lista de figuras

Figura 1 – Gráfico de Perfil de peso e altura por sexo de crianças menores de 2 anos, Brasil, 2008-2017	35
Figura 2 – Tempos de observação para um subconjunto aleatório de 100 crianças entre 0 e 2 anos do Banco Completo (Brasil, 2008-2017)	36
Figura 3 – Tempos de observação para a proposta de análise para um subconjunto aleatório de 100 crianças entre 0 e 2 anos, Brasil, 2008-2017	38
Figura 4 – Curvas estimadas pelo modelo <i>Brokenstick</i> por sexo, para dados antropométricos peso e altura de crianças a termo 0 a 2 anos de idade, Brasil, 2008-2017	39
Figura 5 – Curvas estimadas pelo modelo <i>Brokenstick</i> por sexo, para dados antropométricos peso e altura de crianças a termo com baixo peso ao nascer de 0 a 2 anos de idade, Brasil, 2008-2017	40
Figura 6 – Trajetória média ajustada para peso e altura, pelo modelo de efeitos mistos para crianças de 0 a 2 anos de idade por sexo e classificação de peso ao nascer.	43
Figura 7 – Trajetória média ajustada e observada para peso e altura, pelo modelo de efeitos mistos para crianças PAN de 0 a 2 anos de idade por sexo.	44
Figura 8 – Trajetória média ajustada e observada para peso e altura, pelo modelo de efeitos mistos para crianças BPN de 0 a 2 anos de idade por sexo.	45
Figura 9 – Gráficos de dispersão e histograma dos resíduos condicionais padronizados e quantis em relação aos pesos estimados pelo Modelo de efeitos mistos	46
Figura 10 – Gráficos de dispersão e histograma dos resíduos condicionais padronizados e quantis em relação as alturas estimadas pelo Modelo de efeitos misto	47
Figura 11 – Trajetória média ajustada para peso e altura, pelo modelo <i>Brokenstick</i> para crianças PAN de 0 a 2 anos de idade por sexo.	49
Figura 12 – Trajetória média ajustada para peso e altura, pelo modelo <i>Brokenstick</i> para crianças BPN de 0 a 2 anos de idade por sexo	49
Figura 13 – Gráficos de dispersão dos resíduos condicionais padronizados pelo ajuste do modelo <i>Brokenstick</i> para o peso de crianças de 0 a 2 anos por sexo e classificação de peso ao nascer idade	50
Figura 14 – Gráficos de dispersão dos resíduos condicionais padronizados pelo ajuste do modelo <i>Brokenstick</i> para a altura de crianças de 0 a 2 anos por sexo e classificação de peso ao nascer	50
Figura 15 – Histograma dos resíduos condicionais padronizados pelo ajuste do modelo <i>Brokenstick</i> para o peso de crianças de 0 a 2 anos idade por sexo e classificação de peso ao nascer	51

Figura 16 – Histograma dos resíduos condicionais padronizados pelo ajuste do modelo <i>Brokenstick</i> para a altura de crianças de 0 a 2 anos idade por sexo e classificação de peso ao nascer	51
Figura 17 – Valor ajustado pelo modelo <i>Brokenstick</i> versus observado dos dados antropométricos brutos de crianças PAN do sexo masculino de 0 a 2 anos idade . . .	52
Figura 18 – Valor ajustado pelo modelo <i>Brokenstick</i> versus observado dos dados antropométricos brutos de crianças PAN do sexo feminino de 0 a 2 anos de idade . .	52
Figura 19 – Valor ajustado pelo modelo <i>Brokenstick</i> versus observado dos dados antropométricos brutos de crianças BPN do sexo masculino de 0 a 2 anos idade. . .	53
Figura 20 – Valor ajustado pelo modelo <i>Brokenstick</i> versus observado dos dados antropométricos brutos de crianças BPN do sexo feminino de 0 a 2 anos idade. . . .	53
Figura 21 – Trajetória média do peso ajustado pelo modelo SITAR (esquerda), e curva de velocidade de crescimento (direito) para peso de crianças entre 0 e 2 anos. .	59
Figura 22 – Trajetória média da altura ajustada pelo modelo SITAR (esquerda), e curva de velocidade de crescimento (direito) para altura de crianças entre 0 e 2 anos. .	60
Figura 23 – Trajetória média, ajustada e observada, para peso e altura, pelo modelo SITAR para crianças PAN do sexo masculino de 0 a 2 anos	61
Figura 24 – Trajetória média, ajustada e observada, para peso e altura, pelo modelo SITAR para crianças PAN do sexo feminino de 0 a 2 anos	61
Figura 25 – Trajetória média ajustada e observada para peso e altura, pelo modelo SITAR para crianças BPN do sexo masculino de 0 a 2 anos	62
Figura 26 – Trajetória média ajustada e observada para peso e altura, pelo modelo SITAR para crianças BPN do sexo feminino de 0 a 2 anos	62
Figura 27 – Gráficos de dispersão dos resíduos padronizados em relação aos valores ajustados pelo modelo SITAR para variáveis antropométricas em crianças PAN de 0 a 2 anos por sexo.	63
Figura 28 – Gráficos de dispersão dos resíduos padronizados em relação aos valores ajustados pelo modelo SITAR para variáveis antropométricas em crianças BPN de 0 a 2 anos por sexo.	63
Figura 29 – Gráficos de probabilidade normal a partir dos modelos SITAR ajustado para variáveis antropométricas, para crianças PAN entre 0 e 2 anos por sexo . . .	64
Figura 30 – Gráficos de probabilidade normal a partir dos modelos SITAR ajustado para variáveis antropométricas, para crianças BPN entre 0 e 2 anos por sexo . . .	64
Figura 31 – Histograma dos resíduos padronizados pelo ajuste do modelo SITAR para o peso de crianças de 0 a 2 anos, por sexo e classificação de peso ao nascer . .	65
Figura 32 – Histograma dos resíduos padronizados pelo ajuste do modelo SITAR para a altura de crianças de 0 a 2 anos, por sexo e classificação de peso ao nascer .	65
Figura 33 – Histogramas do peso e altura para crianças PAN de 0 a 2 anos. Brasil 2008-2017	66
Figura 34 – Boxplot do peso e altura para crianças PAN de 0 a 2 anos. Brasil 2008-2017	67

Figura 35 – Gráficos de quantis para os resíduos de confundimento mínimo padronizados obtidos a partir dos modelos ajustados para peso e altura de crianças PAN de 0 a 2 anos de idade por sexo	70
Figura 36 – Gráficos de quantis para os resíduos de confundimento mínimo padronizados obtidos a partir dos modelos ajustados para peso e altura de crianças BPN de 0 a 2 anos de idade por sexo	71

Lista de tabelas

Tabela 1 – Característica sociodemográficas materna e do recém-nascido da população para crianças de 0 a 2 anos. Brasil 2008-2017	34
Tabela 2 – Proposta de análise para dados irregulares de crianças menores de 2 anos aninhada a coorte de 100 milhões de brasileiro. Brasil 2008 - 2017.	37
Tabela 3 – Comparação do modelo <i>Brokenstick</i> para variáveis antropométricas peso e altura em relação ao número de observações, tempo de processamento e variância residual para crianças a termo PAN de 0 a 2 anos e por sexo. . . .	39
Tabela 4 – Comparação do modelo <i>Brokenstick</i> para variáveis antropométricas peso e altura em relação ao número de observações, tempo de processamento e variância residual para crianças a termo BPN de 0 a 2 anos e por sexo. . . .	40
Tabela 5 – Número de medições das variáveis antropométricas brutas por intervalo de idade alvo para crianças de 0 a 2 anos, por sexo e classificação de peso ao nascer - Banco PropostaMed	41
Tabela 6 – Número de medições das variáveis antropométricas brutas por criança de 0 a 2 anos, por sexo e classificação de peso ao nascer - Banco PropostaMed . .	41
Tabela 7 – Modelos de efeitos mistos para variáveis antropométricas peso (kg) e altura (cm) de crianças menores de 2 anos de idade. Brasil, 2008-2017.	43
Tabela 8 – Estimativas e desvio padrão (DP) dos coeficientes do modelo <i>Brokenstick</i> para o peso das crianças menores de 2 anos de idade (nós nas idades alvo 0, 2,6,12 e 24 meses, por sexo e classificação de peso ao nascer)	48
Tabela 9 – Estimativas e desvio padrão (DP) dos coeficientes do modelo <i>Brokenstick</i> para a altura das crianças menores de 2 anos de idade (nós nas idades alvo 0, 2,6,12 e 24 meses, por sexo e classificação de peso ao nascer)	48
Tabela 10 – Comparação dos ajustes dos modelos SITAR aplicados aos dados antropométricos em crianças PAN de 0 a 2 anos, com diferentes parametrizações e transformações na variável idade em meses.	55
Tabela 11 – Comparação dos ajustes dos modelos SITAR aplicados aos dados antropométricos em crianças BPN de 0 a 2 anos, com diferentes parametrizações e transformações na variável idade em meses.	56
Tabela 12 – Modelo SITAR de melhor ajuste para o sexo masculino aplicados aos dados antropométricos de peso (kg) em crianças de 0 a 2 anos por classificação de peso ao nascer. Brasil 2008-2017.	57
Tabela 13 – Modelo SITAR de melhor ajuste para o sexo feminino aplicados aos dados antropométricos de peso (kg) em crianças de 0 a 2 anos por classificação de peso ao nascer. Brasil 2008-2017.	57

Tabela 14 – Modelo SITAR de melhor ajuste para o sexo masculino aplicados aos dados antropométricos de altura (cm) em crianças de 0 a 2 anos por classificação de peso ao nascer. Brasil 2008-2017.	58
Tabela 15 – Modelo SITAR de melhor ajuste para o sexo feminino aplicados aos dados antropométricos de altura (cm) em crianças de 0 a 2 anos por classificação de peso ao nascer. Brasil 2008-2017.	58
Tabela 16 – Distribuição em relação as características da base de dados e amostra de 10%. Brasil 2008-2017	59
Tabela 17 – Ajuste do modelo <i>Brokenstick</i> para uma amostra de crianças PAN, do sexo masculino entre de 0 a 2 anos de idade. Brasil 2008-2017	68
Tabela 18 – Ajuste do modelo <i>Brokenstick</i> para uma amostra de crianças PAN, do sexo feminino entre de 0 a 2 anos de idade. Brasil 2008-2017	68
Tabela 19 – Ajuste do modelo <i>Brokenstick</i> para crianças BPN, do sexo masculino de 0 a 2 anos de idade. Brasil 2008-2017	69
Tabela 20 – Ajuste do modelo <i>Brokenstick</i> para crianças com BPN, do sexo feminino de 0 a 2 anos de idade. Brasil 2008-2017	69
Tabela 21 – Principais resultados da estimativa IFM para os modelos <i>Brokenstick</i> baseados em cópulas bivariados ajustados aos dados de uma amostra de crianças PAN, e parâmetros de associação de cópulas por sexo	70
Tabela 22 – Principais resultados da estimativa para os modelos <i>Brokenstick</i> baseados em cópulas bivariados ajustados aos dados de uma amostra de crianças BPN, e parâmetros de associação de cópulas por sexo	71

Lista de abreviaturas e siglas

AIC	<i>Akaike information criteria</i>
BIC	<i>Bayesian information criteria</i>
BPN	Baixo peso ao nascer
CadÚnico	Cadastro Único para Programas Sociais do Governo Federal
CIDACS	Centro de Integração de Dados e Conhecimentos para Saúde
DP	Desvio Padrão
E/I	Estatura por idade
IFM	Método da inferência para as marginais
IMC	Índice de massa corpórea
MV	Máxima verossimilhança
MVR	Máxima verossimilhança restrita
PAN	Peso adequado ao nascer
P/E	Peso por estatura
P/I	Peso por idade
RSD	Desvio Padrão Residual
SINASC	Sistema de Informações sobre Nascidos Vivos
SISVAN	Sistema de Vigilância Alimentar e Nutricional
SITAR	<i>SuperImposition by Translation And Rotation</i>

Sumário

1	INTRODUÇÃO	1
1.1	Objetivos	5
2	MODELOS PARA TRAJETÓRIA DE CRESCIMENTO INFANTIL	6
2.1	Características dos estudos longitudinais	6
2.2	Abordagem Univariada	8
2.2.1	Modelo Linear de Efeitos Mistos - MLM	8
2.2.2	Especificação do modelo	9
2.2.3	Estimação Pontual	12
2.2.4	Modelo <i>Brokenstick</i> – <i>Splines</i> Lineares	13
2.2.5	Modelo SITAR	16
2.2.6	Análise de Resíduo	18
2.3	Abordagem Multivariada - Cópulas	19
2.3.1	Conceitos e Definições	19
2.3.2	Famílias Paramétricas de Cópulas	23
2.3.2.1	Clayton	23
2.3.2.2	Frank	24
2.3.2.3	Gumbel	24
2.3.2.4	Gaussiana	25
2.3.2.5	t-Student	25
2.3.3	Inferência	26
3	DESCRIÇÃO DO BANCO DE DADOS	28
3.1	Fonte de Dados	28
3.2	Integração e desidentificação das bases de dados	30
3.3	Descrição da abordagem metodológica e dos procedimentos	31
3.3.1	Critérios de exclusões	31
4	RESULTADOS	33
4.1	Estratégia de análise de dados para medições irregulares	35
4.2	Ajustes dos modelos univariados para as trajetórias de crescimento	42
4.2.1	Modelo Linear Misto	42
4.2.2	Modelo <i>Brokenstick</i>	44
4.2.3	Modelo SITAR	47

4.3	Ajuste do Modelo Bivariado - Cópula	57
5	DISCUSSÃO	72
6	CONCLUSÃO	75
6.1	Perspectivas Futuras	76
	Referências	78
	Apêndices	84
	APÊNDICE A – CÓDIGOS DO PROGRAMA R	85

1 Introdução

Há uma crescente ênfase nas pesquisas da área de saúde envolvendo a modelagem do crescimento ao longo da vida e na identificação de fatores associados ao crescimento (COLE, 2002; JOHNSON; BALAKRISHNA; GRIFFITHS, 2013; TILLING et al., 2014; HOWE et al., 2016; ANDERSON et al., 2019; DOBSON et al., 2020; CONFORTIN et al., 2021). O crescimento infantil é um importante indicador de saúde desde a primeira infância até a puberdade, pois pode determinar o estado de saúde na idade adulta. A investigação das curvas de crescimento através do peso e altura é a maneira mais apropriada de avaliar o estado nutricional e o crescimento infantil (COLE, 2002; HOSSEINI et al., 2014). Por sua própria natureza, o crescimento físico é um processo dinâmico longitudinal e, portanto, gera desafios conceituais e metodológicos (HOWE et al., 2016; ANDERSON et al., 2019).

Diversos fatores devem ser considerados na determinação dos padrões de crescimento infantil, como os aspectos nutricionais, culturais, ambientais, sociais, biológicos, genéticos, para além daqueles relacionados ao ambiente social e cultural em que vivem a criança e sua família (SPYRIDES et al., 2008; CONFORTIN et al., 2021). O crescimento e desenvolvimento adequados indicam condições plenas de saúde e vida da criança. Assim, o monitoramento do estado antropométrico pode garantir a promoção e manutenção da saúde, bem como indicar o momento de intervenção sobre fatores capazes de comprometê-la.

O crescimento é um dos principais indicadores de saúde na infância. A avaliação e monitoração do crescimento ponderal e linear são ações definidas nas políticas de atenção à saúde e nutrição para esta fase da vida (ANDERSON et al., 2019). Altura e índice de massa corporal (IMC) são medidas antropométricas da qualidade da nutrição e salubridade do ambiente de vida durante a infância e adolescência e preditivos de saúde e desenvolvimento ao longo da vida. (HORIMOTO et al., 2020). O acompanhamento dessas medidas vai ajudar numa ação precoce, prevenindo o desenvolvimento de morbidades futuras.

O acompanhamento das medidas antropométricas é fundamental desde o nascimento do indivíduo, pois existem diferenças no crescimento entre as nascidas a termo (idade gestacional igual ou maior que 37 semanas) com peso adequado ao nascer e aquelas nascidas prematuramente (idade gestacional inferior a 37 semanas), e ainda aquelas com baixo peso ao nascer, definido como inferior a 2500g, que pode ser uma consequência da prematuridade ou estar associada à restrição de crescimento intrauterino, ou uma combinação das duas (SULLIVAN et al., 2008). O

baixo peso ao nascer é um dos principais fatores que afetam a morbimortalidade infantil em todo o mundo, ((MANYEH et al., 2016) e, na vida adulta, está associado com ao aumento de várias doenças como a síndrome metabólica e diabetes mellitus.

O acompanhamento do peso e da altura nas escalas brutas ou padronizadas (altura ou peso relativo de uma criança em uma determinada idade na forma de escore Z), pode auxiliar as ações da gestão de política pública para a prevenção da desnutrição e obesidade voltadas à promoção da saúde e nutrição adequada ao crescimento na infância. O registro de tais medidas, do mesmo indivíduo, de forma repetida e periódica, durante um período de tempo, apresenta caráter longitudinal e melhora a consistência da informação, fornecendo dados sobre as variações global e individual ao longo do tempo. Esse registro torna possível responder as questões relativas às mudanças que ocorrem na dinâmica do crescimento, particularmente sobre duas questões que interessam diretamente a esse evento. A primeira delas diz respeito à trajetória do comportamento do crescimento de cada indivíduo ao longo do tempo. A segunda é a que busca prever as diferenças que ocorrem no crescimento médio entre o grupo de indivíduos (KOLETZKO; RZEHA; REGNAULT, 2014).

Uma característica importante em dados longitudinais é a sua estrutura balanceada ou desbalanceada, sendo a última mais comumente encontrada. O desbalanceamento é uma consequência natural em estudo de coorte, pois o tempo total de observações nem sempre é igual para todos os indivíduos, e ainda pode ocorrer a presença de padrões de observações desequilibradas entre os sujeitos devendo, portanto, ser considerado na análise dos dados (DIGGLE et al., 2002).

Existem muitos métodos para avaliar a trajetória de crescimento, mas há uma semelhança entre eles que leva a suposição de uma curva média geral para uma determinada população, e as diferenças entre as crianças podem ser explicadas como desvios dessa curva média (VERBEKE, 1997; ANDERSON et al., 2019). Esses métodos levam em consideração a correlação entre observações repetidas na mesma criança para obter erros padrões apropriados, por exemplo, os modelos multiníveis que são frequentemente usados para descrever trajetórias, os modelos não lineares, as análises de curva de crescimento latente e modelagem de mistura (ANDERSON et al., 2019; MCNEISH; MATTA, 2018; TILLING et al., 2014).

Na atualidade, muitos modelos estatísticos de crescimento têm sido explorados, especialmente por considerar diferenças inter e intra-individuais. O modelo linear de efeitos mistos é frequentemente usado, pois permite descrever a tendência temporal considerando a correlação existente entre as medidas repetidas do crescimento no mesmo indivíduo, e possui diferencial de se adequar quando o número de medições difere entre os indivíduos, ou quando as medições são feitas em tempos diferentes, tendo o tempo como uma covariável do modelo e atinge a convergência a grande volume de dados ((TILLING et al., 2014; MCNEISH; MATTA, 2018; ANDERSON et al., 2019). Uma abordagem é modelar as variáveis dependentes como uma simples função da idade (que pode ser linear ou pode incluir polinômios). Outro método denominado modelo *Brokenstick* é um caso especial do modelo linear de efeitos misto, que considera o tempo como

uma função B-spline linear e o indivíduo como o fator de agrupamento. Ele modela a trajetória de crescimento como uma série de linhas conectadas ("*splines* lineares", também chamados de modelos "quebrado" ou "por partes") unidos em nós. Este modelo também é flexível a dados desbalanceados e com falta de padrão regular entre as medidas (HOWE et al., 2016; BUUREN, 2020).

Outra abordagem que tem sido muito usada na atualidade para estimar trajetória de crescimento, foi proposto por Cole (2010), e é conhecido como *Superimposition by Translation and Rotation* (SITAR), que envolve cada indivíduo com uma curva que é uma versão deslocada e transformada da curva média de crescimento. Deslocar a curva verticalmente para cima ou para baixo corresponde às mudanças médias, deslocá-la horizontalmente para a esquerda ou direita corresponde a diferentes tempos de crescimento.

Os modelos descritos acima consideram apenas uma variável resposta antropométrica. Entretanto, é muito comum em estudos longitudinais a observação de um conjunto de variáveis respostas medidas repetidamente para um mesmo indivíduo, e desperta o interesse em estudar a relação de dependência entre as variáveis respostas no tempo nos padrões de crescimento infantil. A abordagem de cópula, que é a distribuição conjunta de duas ou mais variáveis respostas escrita como produto das distribuições marginais e condicionais das variáveis, possibilita estudar a evolução conjunta das variáveis peso e altura.

A investigação em grandes bases de dados com registros eletrônicos na área da saúde, por meio dos modelos descritos acima, permite auxiliar no monitoramento do crescimento e desenvolvimento infantil que são ações essenciais para a prevenção da desnutrição, obesidade e promoção da saúde na infância. Portanto, estimar as trajetórias de crescimento permite a caracterização de seus diferentes padrões, como, por exemplo, identificar as crianças que crescem com sucesso e aquelas cujo crescimento é menos saudável. Para as crianças que estão com crescimento menos saudável, é possível identificar a necessidade de intervenção e quando ocorre o momento de recuperação (HOSSEINI et al., 2014).

O papel das trajetórias de saúde ao longo da vida é um dos temas que destaca a necessidade de dados capazes de subsidiar as próximas etapas da pesquisa em saúde pública e saúde da população, tendo como ferramenta os estudos longitudinais (JUTTE; ROOS; BROWNELL, 2011; HARRON et al., 2017). No entanto, a coleta longitudinal de dados é difícil de coordenar, demorada e cara, além das questões de perda de seguimento, bem como limitações resultantes de critérios de exclusão e requisitos de consentimento do sujeito que reduzem a generalização para a população em geral (TU et al., 2004). Dessa forma, os dados administrativos vinculados compreendem informações já coletadas de forma ampla e diligente em grandes populações para outros fins, depois mescladas no nível individual usando identificadores únicos e anônimos e disponibilizadas para pesquisa acadêmica (HARRON et al., 2017). Esses ambientes ricos em informações podem, com custo relativamente baixo, fornecer amostras muito grandes e observações de longo prazo para futuras pesquisas em saúde pública que a coleta de dados

primários geralmente não consegue (JUTTE; ROOS; BROWNELL, 2011; HARRON et al., 2017).

A coorte eletrônica de base populacional linkados com dados de rotina de nascidos vivos (SINASC) e do Sistema de Vigilância Alimentar e Nutricional (SISVAN) do Ministério da Saúde, é um exemplo de grande base de dados administrativos que coleta informações contínuas sobre condições e monitoramento do estado nutricional da população. Essa grande base de dados, com a diversidade de populações, tem potencial para ser explorada para estimar trajetórias de crescimento na primeira infância. Esses registros são de demanda espontânea e são encontrados nos estabelecimentos de Assistência de Saúde. No entanto, as informações sobre as medições antropométricas das crianças, que muitas vezes permanecem nos serviços, são subutilizadas pela dificuldade metodológica não somente no desenho do estudo, mas, particularmente, na condução das análises estatísticas de medidas biológicas, especialmente aquelas computadas ao longo do tempo.

Este estudo tem o objetivo de modelar a trajetória do peso e altura de crianças menores de 2 anos da coorte eletrônica de base populacional linkados com dados de rotina de nascidos vivos (SINASC) e do Sistema de Vigilância Alimentar e Nutricional (SISVAN), uma base com grande volume de dados ainda pouco explorada na literatura. Mais especificamente, este estudo visa comparar três técnicas de modelagem marginal diferentes: os Modelos de Efeitos Mistos, *Brokenstick* e SITAR para as medidas de peso e altura na sua forma bruta longitudinalmente e em segundo explorar a relação de dependência dessas variáveis conjuntamente no padrão de crescimento infantil através da abordagem de Cópula.

1.1 Objetivos

Objetivo Geral:

- Estimar trajetórias de crescimento com modelagem conjunta e univariada (marginal) para as medidas antropométricas de peso e altura, na população menor de dois anos de idade aninhada a Coorte de 100 milhões de brasileiros.

Objetivos Específicos:

- Estimar as trajetórias de crescimento infantil na população geral e subgrupos de interesse como sexo e baixo peso ao nascer para crianças de 0 a 2 anos, através dos modelos estatísticos univariados, tais como, modelos lineares misto, *Brokenstick* e SITAR.
- Caracterizar a evolução temporal das variáveis antropométricas peso e altura em dois procedimentos, uma abordagem univariada, considerando os modelos lineares misto, *Brokenstick* e SITAR para cada variável resposta separadamente; e uma abordagem bivariada baseada em cópulas para caracterizar as relações entre as variáveis respostas conjuntamente.

Este estudo está organizado da seguinte forma. No capítulo 2, descrição das características dos estudos longitudinais e dos modelos de crescimento longitudinais univariados e conjunta bivariada abordados neste estudo. O capítulo 3, trata do detalhamento do banco de dados e da fase de pré-processamento dos dados. O capítulo 4, apresenta os resultados do pré-processamento dos dados e ajustes dos modelos de crescimento longitudinais. E, por fim, o capítulo 5 descreve conclusões e aponta perspectivas para trabalhos futuros.

2 Modelos para trajetória de crescimento infantil

Os modelos de curva de crescimento infantil são frequentemente definidos como representações matemáticas do processo do crescimento humano. Segundo Berkey (1982), um modelo de crescimento é considerado ideal quando os parâmetros possuem interpretações biológicas, seu procedimento de ajuste é simples e satisfatório aos dados, e quando fornece predições. Além disso, é desejável que seja robusto às observações ausentes, à variabilidade na idade sobre as observações e ao erro de medição, pois são ocorrências comuns em estudos de crescimento longitudinal.

Muitos modelos matemáticos de crescimento foram introduzidos na literatura e são classificados como modelos paramétricos ou estruturais e não paramétricos ou não estruturais (HAUSPIE; CAMERON; MOLINARI, 2004). Segundo Hauspie e colaboradores (2004) os modelos paramétricos exibem a forma funcional básica do modelo de crescimento, normalmente têm menos parâmetros que permitem alguma interpretação biológica e tendem a atingir uma assíntota superior. Por outro lado, os modelos não paramétricos são fáceis de ajustar, mas apresentam limitações por conter um grande número de coeficientes que precisam ser estimados e os parâmetros não têm interpretação biológica. Esses modelos fornecem aproximações boas para os valores dentro dos dados observados, mas geralmente têm um ajuste ruim em valores extremos observados na amostra. Como exemplos de modelos não paramétricos temos os de polinômios e splines cúbicos (TILLING et al., 2014; GOLDSTEIN et al., 2018; LU; PEARCE; LI, 2020) que podem ser usados para dados de crescimento de qualquer variável antropométrica em qualquer faixa etária (SILVERWOOD, 2008). Ambos os tipos de modelos não fornecem previsões confiáveis do resultado fora da faixa de dados observados.

Neste capítulo será exibida uma breve revisão bibliográfica das características dos estudos longitudinais e dos modelos estatísticos utilizados para avaliar o crescimento infantil.

2.1 Características dos estudos longitudinais

Em muitos conjuntos de dados agrupados de acordo com um ou mais fatores, o pressuposto da independência entre observações não pode ser verificado, como é o caso de dados com medidas

no mesmo indivíduo ao longo do tempo sendo o próprio tempo, em muitos casos, um fator de interesse no estudo. Essas medições realizadas no mesmo indivíduo tendem a estar relacionadas entre si e por isso é fundamental pensar em um modelo que leva em consideração essa dependência intra-grupo (FITZMAURICE et al., 2008; DIGGLE et al., 2002).

Num estudo longitudinal uma ou mais variáveis respostas são medidas repetidamente ao longo do tempo no mesmo indivíduo. Embora haja uma programação para que essa resposta seja avaliada em um conjunto de pontos no tempo, podem existir divergências nas reais medições, ou ainda, a saída permanente do indivíduo. Os dados longitudinais observacionais geralmente apresentam observações que variam no tempo entre os indivíduos, a exemplo de dados de registro de saúde que geralmente apresentam tempos de acompanhamento irregulares, ou seja, variação no tempo entre as visitas dos pacientes além de poucos indivíduos no mesmo tempo para a observação (LIN; SCHARFSTEIN; ROSENHECK, 2004). Para Lokku (2020), o problema da irregularidade entre as observações num mesmo indivíduo pode ser visto como dados ausentes, quando, por exemplo, o registro não ocorre na data programada, ou como dados irregulares, quando não há um padrão de visitas equilibrado entre os indivíduos.

Na década de 90, os métodos estatísticos para dados longitudinais sujeitos a acompanhamento irregular começaram a ser desenvolvidos (LIANG; LU; YING, 2009). Os principais desafios são a dificuldade de calcular a matriz de correlação quando os dados variam pouco ao longo do tempo, realizar previsões quando os tempos da variável resposta forem diferentes, ou ainda quando os tempos das observações estão relacionados ao processo de interesse. Por exemplo, indivíduos/pacientes que, por algum motivo, necessitem ir mais ao posto de saúde, resultando num maior número de medições. A variação resultante nos tempos de observação torna inviáveis métodos como a imputação múltipla, não ficando claro em quais pontos de tempo se deve imputar os dados, uma vez que normalmente dois indivíduos não terão uma medição no mesmo ponto de tempo (BUUREN, 2020; LOKKU et al., 2020).

Pullenayegu e Lim (2014), apresentaram uma revisão de métodos estatísticos com foco nos processos de visita, suposições e desenho de estudo, para dados longitudinais sujeitos à observações irregulares. Nesta revisão, sugere-se métodos semiparamétricos, equações de estimativa generalizada ponderada de intensidade inversa (GEE), modelos mistos dentre outros modelos, fazendo suposições sobre a relação entre o resultado e a irregularidade. Segundo os autores o modelo de efeitos mistos pode lidar com visitas regulares e visitas irregulares quando consideradas completamente aleatórias, trazendo uma importante advertência em relação à correta especificação da estrutura de correlação do processo residual, pois a especificação incorreta geralmente resultará em estimativas de parâmetros tendenciosos. Farzanfar e colaboradores (2017), realizaram uma revisão sistemática de estudos longitudinais para explorar como a irregularidade é relatada e tratada na prática. Eles concluíram que a baixa proporção de estudos relatando informações importantes sobre como acontece as visitas sugere que os pesquisadores desconhecem o potencial de viés que poderá ser adicionado às análises. Uma das razões pelas

quais a irregularidade das visitas é ignorada na prática é que a maior parte da literatura sobre o tema é mais técnica. Lukko e colaboradores (2021), em seu trabalho com estratégia exploratória para sumarizar a extensão da irregularidade da visita em dados longitudinais, sugere um método que descreve a extensão da irregularidade e ajuda a distinguir entre visitas que podem ser guiadas por protocolo e visitas irregulares.

Até o momento, na literatura, não há uma abordagem padrão para resolver o problema da irregularidade nas observações. Uma estratégia é usar os dados disponíveis nas datas preestabelecidas, da criação de intervalos de tempo em torno das datas planejadas, ou ainda usar a abordagem *ad-hoc* que faz a previsão desse tempo a partir de dados vizinhos utilizando interpolação linear ou suavização. A sugestão de correções com a criação de um novo conjunto de dados, pode levar à idéia que o problema teve solução e levar o pesquisador a ignorar, nas suas conclusões, efeitos potenciais ao tentar consertar os dados ((TOWERS, 2014; BUUREN, 2020; REHFELD et al., 2011)). Nenhuma dessas propostas estão imunes a gerar vieses nas análises, e seus efeitos potenciais na conclusão não devem ser ignorados ((TOWERS, 2014; PULLENAYEGUM; LIM, 2016)).

Os dados provenientes de estudos longitudinais podem também ser caracterizados como dados hierárquicos, uma vez que as medidas repetidas podem ser aninhadas intra-indivíduo. A variável resposta pode apresentar uma estrutura desbalanceada, que é uma consequência natural em corte dinâmico, pelo fato de que o tempo total de observações nem sempre é igual para todos os indivíduos e as medidas não serem necessariamente tomadas em intervalos regulares. Muitos métodos estatísticos estão disponíveis para análise de dados longitudinais e recentemente, a modelagem para trajetória de crescimento tem se popularizado. Estes métodos são utilizados para descrever o crescimento médio de determinadas características da população ao longo do tempo em um determinado período, tendo como ênfase a explicação da variabilidade entre indivíduos, ou seja, a diferença inter-individual em mudanças intra-individuais (VERBEKE, 1997).

Por sua própria natureza, o crescimento físico é um processo dinâmico longitudinal e, portanto, gera desafios especiais em sua análise de uma perspectiva conceitual e metodológica. Existem muitos métodos para avaliar a trajetória de crescimento, e, em todos eles faz-se a suposição de uma curva média geral para uma determinada população, e as diferenças do crescimento entre crianças podem ser explicadas como desvios dessa curva média.

2.2 Abordagem Univariada

2.2.1 Modelo Linear de Efeitos Mistos - MLM

Na literatura, os modelos de crescimento paramétricos mais frequentemente aplicados foram desenvolvidos por Jenss-Bayley (1937), Count (1943) e Berkey e Reed (1987), e tem o objetivo de estimar uma curva de crescimento individual. No entanto, uma desvantagem desses

modelos é que os parâmetros estimados não têm quaisquer interpretações biológicas (JENSS; BAYLEY, 1937; BERKEY; REED, 1987; BOTTON et al., 2008). Laird e Ware (1982), propuseram uma abordagem alternativa e mais flexível para modelar dados longitudinais denominados modelos mistos (*mixed models*) ou modelo de efeitos mistos (*mixed effects models*) (também conhecidos como modelos de efeito aleatório, multinível ou hierárquico) (FITZMAURICE; LAIRD; WARE, 2012; PINHEIRO; BATES, 2000). A vantagem dos modelos mistos é que apresentam uma estrutura unificada que permite trabalhar tanto com dados balanceados quanto desbalanceados (PINHEIRO; BATES, 2000). Esta abordagem modela todas as crianças simultaneamente em vez de ajustá-las individualmente, por meio dos efeitos fixos e aleatórios dos parâmetros do modelo.

O modelo de efeitos mistos é frequentemente usado para descrever trajetórias, pois analisam medidas repetidas (nível 1) agrupadas dentro de indivíduos (nível 2) e possui diferencial de se adequar à estrutura de dados complexa, tendo o tempo como uma covariável do modelo e atinge a convergência a grande volume de dados (TILLING et al., 2014; MCNEISH; MATTA, 2018; ANDERSON et al., 2019).

Vários estudos utilizaram os modelos de efeitos mistos proposto por Laird e Ware (1982) envolvendo a questão do crescimento infantil. Spyrides e colaboradores (2005) avaliou o efeito da duração do aleitamento materno predominante no perfil antropométrico ao crescimento em lactentes brasileiros. Zhang e colaboradores (2012) investigou os fatores potenciais associados com elevados ganhos de peso e estado de IMC elevado nos primeiros três meses de vida. Jonhson e colaboradores (2013) caracterizou curvas de crescimento médio de indivíduos com base em dados antropométricos. Anderson e colaboradores (2018) comparou habilidades preditivas em modelos de crescimento infantil. Ferreira e colaboradores (2019) utilizou o modelo linear de efeitos misto univariado e bivariado para dados longitudinais de crescimento infantil. Em geral, através dos modelos mistos é possível estimar a variabilidade interindividual em padrões intra-individuais de mudança de crescimento ao longo do tempo (FITZMAURICE; LAIRD; WARE, 2012).

2.2.2 Especificação do modelo

O modelo de efeitos aleatórios para dados longitudinais de Laird e Ware (1982) é considerado um modelo base para análise de dados quando o número ou momento das medições difere entre os indivíduos, tendo como proposta o uso de efeitos aleatórios como forma de caracterizar os desvios específicos da criança de uma média global e formalmente, é dado por:

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\boldsymbol{\zeta}_i + \boldsymbol{\varepsilon}_i \quad (2.1)$$

Em que

$\mathbf{y}_i$ é o vetor ($n \times 1$) de n respostas do i -ésimo indivíduo;

$\mathbf{X}_i$ é uma matriz $(n \times p)$ de posto completo, conhecida e de especificação, associada ao vetor $(p \times 1)$ de parâmetros de efeito fixo β ;

$\mathbf{Z}_i$ é uma matriz $(n \times q)$ conhecida e de especificação, de posto completo e associada ao vetor $(q \times 1)$ de efeitos aleatórios ζ_i ;

ε_i é um vetor $(n \times 1)$ de erros aleatórios.

No contexto dos modelos mistos admite-se que $\zeta_i \sim N_q(\mathbf{0}, \mathbf{D})$, onde $\mathbf{D}$, é a matriz de covariância de dimensão $(q \times q)$ e $\varepsilon_i \sim N_n(\mathbf{0}, \mathbf{R}_i)$ para $i = 1, \dots, n$, onde $\mathbf{R}_i$ é matriz de covariância diagonal. Supõe-se ainda que ζ_i e ε_i são independentes entre os sujeitos dadas as covariáveis, e que a variância é a mesma em todos os pontos de tempo para todas as unidades, ou seja, $\mathbf{R}_i = \sigma^2 I_{n_i}$, em que I_{n_i} representa a matriz identidade de ordem n_i . Marginalmente, os $\mathbf{y}_i$ são normais e independentes com média $\mathbf{X}_i\beta$ e a matriz de covariância $\mathbf{R}_i + \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^T$ (LAIRD; WARE, 1982).

Em estudos longitudinais, espera-se uma diminuição da correlação entre as respostas à medida que a distância entre duas observações feitas no mesmo indivíduo aumenta e a existência de heterogeneidade das variâncias das respostas ao longo do tempo. Dessa forma é importante escolher uma estrutura de covariâncias adequada que represente bem essas características nos dados em estudo. Segundo Diggle (2002), a matriz de covariâncias deve apresentar boa flexibilidade em sua estrutura e apresentar a variação devida aos efeitos aleatórios, a variação relacionada à correlação existente na observação, onde haverá maior correlação para observações mais próximas do que observações mais distantes, e a variação que ocorre devido aos erros de mensuração.

A especificação funcional da estrutura de covariância para o modelo misto é feita através de $\mathbf{D}$ e de $\mathbf{R}_i$, muitas vezes apenas em termos de $\mathbf{R}_i$. A seguir, têm-se exemplos de estruturas amplamente utilizadas para essas matrizes (LITTELL; PENDERGAST; NATARAJAN, 2000).

Considere Y_{ij} e $Y_{ij'}$ a variável resposta do indivíduo i na observação j

- Não Estruturada (UN): A estrutura ‘não estruturada’ não especifica padrões na matriz de covariâncias, e é completamente geral, mas a generalidade traz a desvantagem de ter um número muito grande de parâmetros. Admite heterogeneidade de variâncias e covariâncias distintas entre os pares de observações (j, j') .

$$\text{cov}(Y_{ij}, Y_{ij'}) = \sigma_{\text{UN},kl},$$

Em termos de modelo, é dado com $\mathbf{D} = 0$ e um $\mathbf{R}_i$ completamente geral.

- Simetria Composta - assume homogeneidade de variâncias e mesma correlação entre as medidas feitas nas diferentes ocasiões, ou seja, as correlações entre duas observações são iguais para todos os pares de observações (j, j') sobre o mesmo indivíduo. Envolve

somente 2 parâmetros.

$$\text{cov}(Y_{ij}, Y_{ij'}) = \sigma^2 \rho^{1(j \neq j')}$$

onde $\rho \in [-1, 1]$, e $1(\cdot)$ é uma função indicadora, que neste caso assume valor 1 quando $j \neq j'$ e valor 0 quando $j = j'$;

- Autorregressiva de 1ª ordem - AR(1) - admite homogeneidade de variâncias e decréscimo da correlação à medida que as observações (j, j') ficam mais distantes. Envolve somente 2 parâmetros.

$$\text{cov}(Y_{ij}, Y_{ij'}) = \sigma_{\text{AR}(1)}^2 \rho_{\text{AR}(1)}^{|j-j'|}, \rho \in [-1, 1]$$

O número de parâmetros da matriz de covariância depende de sua estrutura, deve-se escolher a estrutura mais parcimoniosa, ou seja, aquela que possua o menor número possível de parâmetros e que represente as características de variabilidade e correlação dos dados longitudinais (LITTELL; PENDERGAST; NATARAJAN, 2000). A comparação entre essas estruturas pode ser realizada por meio dos *critérios de informação de Akaike* (AIC) ou *Bayesiano* (BIC).

Um exemplo de um dos modelos estatísticos em estudo, é o caso mais simples de modelos lineares mistos que é o modelo de intercepto aleatório, isso significa que cada indivíduo obtém sua própria estimativa de intercepto, mas apresenta uma inclinação comum. Há um único efeito aleatório normalmente distribuído, é dado por:

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + \zeta_{0i} + \varepsilon_{ij}, i = 1, \dots, n, j = 1, \dots, n_i \quad (2.2)$$

em que y_{ij} é o peso da criança i na medida j , β_0 é o intercepto, ζ_{0i} é o intercepto aleatório, β_1 é o efeito atribuído à idade e x_{1ij} é a idade da criança i no momento j e ε_{ij} é o erro aleatório.

O modelo de efeitos mistos com intercepto e inclinação aleatórios, permite a flexibilidade para que cada indivíduo tenha sua própria relação de regressão e pode ser expresso como:

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + \zeta_{0i} + \zeta_{1i} x_{1ij} + \varepsilon_{ij}, i = 1, \dots, n, j = 1, \dots, n_i \quad (2.3)$$

em que ζ_{0i} e ζ_{1i} são os efeitos aleatórios (não correlacionados) do intercepto e da inclinação, respectivamente e os demais termos seguem o modelo (2.2).

Os modelos (2.2) e (2.3) representam uma mudança linear no resultado ao longo do tempo, porém para a maioria dos processos biológicos, o crescimento não é linear. Essa não linearidade pode ser incorporada nos modelos multiníveis de várias maneiras, por exemplo, incluindo funções de idade não lineares no modelo. Isso requer que a função de tempo de melhor ajuste seja selecionada. Segundo Cameron (2002), a curva do crescimento infantil é um polinômio em decomposição, que com o aumento do tempo, ela tende a se afastar da linha reta de forma

gradual e negativa. Então uma sugestão é estender este modelo para que possa capturar tendências não lineares acrescentando, por exemplo, aos modelos (2.2) e (2.3) um efeito de tempo quadrático.

O modelo linear de efeitos misto para a variável resposta y_{ij} do individuo $i = 1, 2, \dots, n$ e pontos no tempo $j = 1, 2, \dots, n_i$ pode ser especificado como:

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \beta_2 x_{ij}^2 + \zeta_{0i} + \varepsilon_{ij}, i = 1, \dots, n, j = 1, \dots, n_i \quad (2.4)$$

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{ij}^2 + \zeta_{0i} + \zeta_{1i} x_{1ij} + \varepsilon_{ij}, i = 1, \dots, n, j = 1, \dots, n_i \quad (2.5)$$

2.2.3 Estimação Pontual

Vários métodos de estimação de parâmetros têm sido usados para modelos lineares de efeitos mistos. Os métodos utilizados na estimação dos parâmetros do modelo (2.1) foram máxima verossimilhança (*maximum likelihood* - ML) e máxima verossimilhança restrita (*restricted maximum likelihood* - REML)

A função de verossimilhança para o modelo linear de efeitos mistos (2.1) como $\mathbf{R}_i = \sigma^2 I_{n_i}$ é dada por

$$L(\boldsymbol{\delta} \mid \mathbf{y}) = L(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2 \mid \mathbf{y}) = \prod_{i=1}^n \int f(\mathbf{y}_i \mid \boldsymbol{\zeta}_i, \boldsymbol{\beta}, \sigma^2) f(\boldsymbol{\zeta}_i \mid \boldsymbol{\theta}, \sigma^2) d\boldsymbol{\zeta}_i \quad (2.6)$$

em que, $\boldsymbol{\theta}$ é conjunto irrestrito de parâmetros a serem estimados, f é a função densidade de probabilidade e $\mathbf{y}$ é vetor resposta N-dimensional, $N = \sum_{i=1}^n n_i$.

De acordo com Pinheiro (2000), o método usual de estimação por máxima verossimilhança fornece estimativas enviesadas para os modelos de interceptos aleatórios, na presença de efeitos aleatórios, nessas condições uma alternativa é usar o método de estimação de *máxima verossimilhança restrita* (MVR), que consiste em dividir as observações em duas partes independentes: uma se refere aos efeitos fixos e a outra aos efeitos aleatórios, onde a função densidade de probabilidade observada é obtida pelo produto dessas partes, mais detalhes em Mcculloch(2001). Neste trabalho, foi utilizado o método de estimação de máxima verossimilhança restrita para os modelos lineares de efeitos mistos propostos (2.2) a (2.5).

A escolha entre os modelos (2.2), (2.3), (2.4) e (2.5) foi feita com base nos valores do *Critério de Informação de Akaike* (AIC) e *Bayesian information criteria* (BIC), definidos como:

$$\begin{aligned} \text{AIC} &= -2l + 2d \\ \text{BIC} &= -2l + d \ln n \end{aligned} \quad (2.7)$$

em que l representa o valor máximo da função de verossimilhança (completa ou restrita), d o número de parâmetros do modelo e n o número de observações. Quanto menor for o valor dessas estatísticas, maior evidência favorável ao modelo em questão (AKAIKE, 1977).

Os modelos foram ajustados no *software* estatístico R usando o Pacote **nlme** e a função **lme(.)** (PINHEIRO et al., 2017).

2.2.4 Modelo *Brokenstick* – *Splines* Lineares

Como o modelo *Brokenstick* baseia-se na regressão por *splines*, uma introdução sobre este tema será apresentado.

A técnica de regressão via *splines* tem como objetivo interpolar polinômios interligados por nós que passe pelos pontos (x_i, y_i) no plano, onde $i = 1, 2, \dots, n$. Os nós definidos por ξ_j , $j = 0, 1, \dots, k$ são escolhidos através de um critério definido pelo pesquisador tais que $\Delta = \{\xi_1 < \xi_2 < \dots < \xi_k\}$ e podem coincidir com os valores assumidos por x , em que $\xi_1 > x_1$ e $\xi_k < x_n$ (WEGMAN; WRIGHT, 1983).

Esta técnica é muito utilizada atualmente devido à sua facilidade na implementação computacional. O B-*splines* é uma abordagem geral da regressão por *splines* e sua principal propriedade é que suas funções são contínuas nos pontos de encontro (nós) (EILERS; MARX, 1996). Outras propriedades do B-*splines*:

- A função consiste de $k + 1$ segmentos de polinômios de grau m ;
- Os segmentos se unem nesses k nós, cujas derivadas nos pontos serão existentes;
- O B-*splines* é positivo no domínio de $k + 2$ nós e nulo no restante;
- Exceto nos limites extremos, cada segmento do polinômio se sobrepõe com 2 x m segmentos de polinômio vizinhos;
- Os nós podem ser equidistantes.

A curva B-*splines* para os dados (x_i, y_i) $i = 1, \dots, m$ é dada pela combinação linear:

$$f(x) = \sum_{j=1}^{k+m-1} a_j B_j(x)$$

Onde $B_j(x)$ denota o valor do j -ésimo B-*splines* de grau m no ponto x para uma malha de k nós equidistantes e a_j é o coeficiente do j -ésimo B-*splines*.

Não existe um consenso entre os autores com respeito ao posicionamento dos nós. Alguns sugerem posicionar o primeiro nó no primeiro valor possível de x , ou definir o primeiro nó onde ocorre o primeiro “corte”. (EILERS; MARX, 1996; BUUREN, 2020).

A estimação dos parâmetros da regressão *splines* é realizada por meio do método dos mínimos quadrados, porém em cada segmento da função ajusta-se um polinômio com parâmetros com os mesmos nós em segmentos adjacentes. O grau da função *splines* m e o número de nós k geralmente são definidos pelo pesquisador, já que a posição dos nós ξ_j , pode ser livre ou fixa. A escolha dos nós e o grau do polinômio são as duas maiores dificuldades (WEGMAN; WRIGHT, 1983).

Uma abordagem alternativa aos Modelos Mistos apresentada na seção anterior para flexibilizar o ajuste da trajetória de crescimento por meio de modelos de efeitos mistos é o uso de uma série de linhas conectadas (*splines* lineares, também chamados de modelos bastão quebrado “*broken-stick*” ou por partes “*piecewise*”) unidas em nós (TILLING et al., 2014; HOWE et al., 2016; ANDERSON et al., 2019). Esse modelo é um caso especial do modelo de efeitos aleatórios de dois estágios (níveis) de Laird e Ware (1982), com o tempo sendo modelado como um conjunto de efeitos aleatórios codificados como um B-*splines* linear e os sujeitos como o fator de agrupamento. Ele resume trajetórias individuais irregulares, ou seja, medições não equidistantes ao longo do tempo, por estimativas feitas em um intervalo de tempo pré-especificado. (ANDERSON et al., 2019; BUUREN, 2020)

Na literatura o modelo de *Brokenstick* tem sido utilizado para modelar trajetória de crescimento infantil. Tilling et al (2014) modelou o crescimento estatural infantil em crianças do nascimento aos 10 anos modelos multiníveis de polinômios fracionários e *spline* linear, onde este último foi mais facilmente interpretável. Howe e colaboradores (2016) utilizou os modelos multiníveis de polinômios fracionários e *spline* linear para modelar trajetórias de altura e peso na infância em cinco coortes com idade variando entre 0 e 10 anos, onde o modelo de *spline* linear apresentou um bom ajuste na trajetória de crescimento. Anderson e colaboradores (2018) comparou modelos de crescimento infantil em 21 bancos de dados e destacou o melhor ajuste ao modelo *Brokenstick* para modelar a trajetória de crescimento tanto para variáveis brutas e padronizadas de peso e altura além da convergência a bases com maior volume de dados.

O modelo linear “*Brokenstick*”, foi proposto por Van Buuren (2014) para modelagem de crescimento através de uma combinação de segmentos lineares com diferentes inclinações. É uma abordagem de modelo *spline* linear, onde os nós são usados para representar pontos de mudança na trajetória de crescimento e considera a posição relativa de cada tempo dentro de um intervalo de tempo pré especificado. Dessa maneira, simplifica a modelagem do tempo contínuo por um conjunto de medidas em intervalos discretos, e t_{ij} (tempo do indivíduo i no momento j) pode ser codificado por uma B-*spline* linear de segunda ordem usando k nós internos colocados em $k + 1$ medidas ordenadas

$$\xi_0 < \xi_1 < \dots < \xi_k < \xi_{k+1}$$

Neste estudo, os nós internos $\xi_1 < \dots < \xi_k$ referem-se ao conjunto de idades e pode

ser especificada pelo pesquisador. Os nós externos, ξ_0 e ξ_{k+1} são definidos como o evento inicial do participante (nascimento ou primeira visita no estudo) e a medida máxima observada, respectivamente (TILLING et al., 2014; HOWE et al., 2016; BUUREN, 2020).

As localizações dos nós são importantes para a precisão geral das estimativas da trajetória de crescimento (BUUREN, 2020). Autores recomendam que os locais sejam selecionados para representar estágios específicos no desenvolvimento de uma criança, mas deve-se notar que isso ocorre mais precisamente no contexto de ajuste na escala bruta. Outra possibilidade é definir nós em visitas programadas como uma estratégia para obter previsões mais precisas nas datas programadas ou ainda começar com um grande número de pontos de nós, reduzindo gradativamente o número até que uma curva “suave” seja alcançada. É importante não colocar nós em áreas de dados pouco preenchidas, por exemplo, entre duas visitas com intervalos curtos, pois pode resultar em junções erráticas. Como uma regra prática pode se usar o limite do número de nós ao número médio de pontos de dados por criança (TILLING et al., 2014; HOWE et al., 2016; ANDERSON et al., 2019; BUUREN, 2020).

Segundo Boor (2001), a condição do B-spline de segunda ordem é descrita como:

$$s_{ijc} = \begin{cases} (t_{ij} - \xi_{c-1}) / (\xi_c - \xi_{c-1}), & \xi_{c-1} < t_{ij} \leq \xi_c \\ (\xi_{c+1} - t_{ij}) / (\xi_{c+1} - \xi_c), & \xi_c \leq t_{ij} < \xi_{c+1} \\ 0, & \text{cc} \end{cases} \quad (2.8)$$

com $i = 1, \dots, n, j = 1, \dots, n_i, t_{ij}$ é o tempo do indivíduo i no momento j e dessa forma, a função *spline* é igual a zero para todas as idades menores que o 1º nó e, aumenta linearmente até o próximo nó (BOOR; BOOR, 1978).

A condição s_{ijc} é aplicado aos tempos t_{ij} para obter $(k + 1)$ variáveis transformadas $x_{ij} = t_{ij}$ com $c = 1, \dots, k + 1$. Essas variáveis são agrupadas na matriz $n_i \times (k + 1)$ de covariáveis $X_i = (x_{i1}, \dots, x_{ik}, x_{i(k+1)})$. Cada linha em X_i tem apenas um ou dois elementos diferentes de zero, que somam 1.

Exemplificando a equação dada em (2.8): Suponha que $t_{ij} = 1,8$ anos e que o intervalo de tempo seja dado por $\xi_c = 1,5$ e $\xi_{c+1} = 2,0$.

A posição relativa a idade à direita é $s_{ijc} = (1,8 - 1,5) / (2,0 - 1,5) = 0,6$ e a idade à esquerda é $s_{ijc+1} = (2,0 - 1,8) / (2,0 - 1,5) = 0,4$. Aplicando a condição do B-spline linear, substitui-se o ponto de tempo $t_{ij} = 1,8$ anos por dois valores: 0,6 (para idade no tempo 1,5 anos) e 0,4 (para idade no tempo 2 anos), os quais somam 1.

Usando as variáveis transformada S_{ijc} o modelo *Brokenstick* é um caso especial do modelo de efeitos mistos de dois estágios (Laird e Ware 1982), com $Z_i = X_i$.

$$\mathbf{Y}_i = S_{ij}\boldsymbol{\beta} + S_{ij}\boldsymbol{\zeta}_i + \boldsymbol{\epsilon}_i \quad (2.9)$$

Em que

$\mathbf{Y}_i$ é o vetor $(n \times 1)$ de n respostas do i -ésimo indivíduo;

β é um vetor coluna $(k + 1)$ de parâmetros fixos;

ζ_i é um vetor coluna $(k + 1)$ de parâmetros aleatórios;

ϵ_i é um vetor $(n \times 1)$ de erros aleatórios.

Assume-se que os $\zeta_i \sim N(\mathbf{0}, \mathbf{D})$, onde $\mathbf{D}$, é a matriz de covariância de dimensão $(k + 1)$ e $\epsilon_i \sim N(0, \sigma^2)$, onde σ^2 é a variância intrapessoal que é a mesma para todos os indivíduo.

Dessa forma, como descrito na seção (2.2.3), a estimação dos parâmetros do modelo foi realizada utilizando o método de máxima verossimilhança restrita (MVR) (BUUREN, 2020). Foi utilizado o pacote **brokenstick** para ajustar os modelos, este pacote utiliza a função **lmer(.)** do pacote **lme4** do programa R para estimar os parâmetros do modelo de efeitos misto.

Para avaliar o crescimento, outra abordagem é usar modelos não paramétricos, como por exemplo, o modelo SITAR (*SuperImposition by Translation And Rotation*) proposto por Cole *et al* (2010). É um modelo de efeito aleatório semi-paramétrico não linear, em que a curva de cada indivíduo é uma versão deslocada e transformada da curva média de crescimento, ou seja, deslocar a curva para cima ou para baixo corresponde às mudanças médias, enquanto deslocando-a para a esquerda ou para a direita corresponde a diferentes tempos de crescimento. (ANDERSON *et al.*, 2019)

2.2.5 Modelo SITAR

O modelo SITAR é uma extensão da modelagem de crescimento invariante de forma, técnica proposta por Beath (2007) para modelar o peso em bebês. O conceito do modelo invariante foi originalmente introduzido em 1972 pelos autores Lawton, Sylvestre, e Maggio, mas só foi aplicada posteriormente em 1980 para modelar crescimento (STÜTZLE *et al.*, 1980)(STÜTZLE *et al.*, 1980). No modelo SITAR, é montada uma função *spline* cúbica natural que estima a curva de crescimento médio comum a todos os indivíduos da amostra. Além da função *spline* cúbica, o modelo também inclui três parâmetros que ajustam a curva de crescimento médio comum para se ajustar a cada trajetória de crescimento individual, e que possuem interpretações biológicas (BEATH, 2007; COLE; DONALDSON; BEN-SHLOMO, 2010).

Muitas das publicações com aplicação do modelo SITAR referem-se ao crescimento na fase da adolescência (COLE; DONALDSON; BEN-SHLOMO, 2010; SIMPKIN *et al.*, 2017; GOLDSTEIN *et al.*, 2018; COLE, 2018; WADHAWAN, 2018; OHUMA *et al.*, 2021), mas, recentemente, alguns estudos foram publicados na primeira infância (ANDERSON *et al.*, 2019; CROZIER *et al.*, 2019; FILTEAU *et al.*, 2019; OHUMA *et al.*, 2021). Os autores (ANDERSON *et al.*, 2019; CROZIER *et al.*, 2019; OHUMA *et al.*, 2021) relatam como pontos negativos a complexidade no ajuste do modelo, sendo necessário em alguns casos, a transformação da variável

resposta ou explicativa, para alcançar a convergência do modelo em grandes bases de dados para os efeitos fixos e aleatórios.

Após a introdução de Beath (2007), Cole et al. (2010) apresentaram sua extensão denominada de *SuperImposition by Translation And Rotation* (SITAR). Este modelo estima a curva de crescimento médio usando a função *spline* cúbica natural, cuja finalidade é capturar a não linearidade nos dados e identificar pontos de inflexão. Este modelo baseia-se na abordagem de efeitos mistos não linear e é definido por

$$y_{it} = \alpha_i + h\left(\frac{t - \beta_i}{\exp(-\gamma_i)}\right) + \varepsilon_{it} \quad (2.10)$$

em que y_{it} representa o peso ou altura da i -ésima criança na idade t para $i = 1, \dots, n, t = 1, \dots, n_i$, $h(t)$ é uma função que descreve a relação não linear entre peso ou altura com a idade baseada na curva *spline* cúbica natural, $\alpha_i = \alpha_0 + \alpha_{1i}, \beta_i = \beta_0 + \beta_{1i}, \gamma_i = \gamma_0 + \gamma_{1i}$, onde $\alpha_0, \beta_0, \gamma_0$ representa os efeitos fixos e $\alpha_{1i}, \beta_{1i}, \gamma_{1i}$ representa os efeitos aleatórios. Assume-se que os resíduos $\varepsilon_{ij} \sim N(0, \sigma^2)$.

Os parâmetros representam os deslocamentos do peso ou altura (eixo y) e idade (eixo x), respectivamente em relação a curva média, enquanto o parâmetro γ_i se refere a escala para o eixo x (idade). Os parâmetros α_i, β_i e γ_i do modelo SITAR são chamados de tamanho, tempo e velocidade/intensidade, respectivamente, onde os dois primeiros parâmetros levam a uma translação na curva *spline* e o terceiro uma rotação da curva de crescimento. O ajuste do modelo SITAR é baseado em um modelo de efeito misto não linear e o método de estimação por máxima verossimilhança restrita (MVR) (COLE; COLE, 2021).

Com base em sua interpretabilidade biológica, α_i se refere ao intercepto de peso ou altura aleatório que ajusta a diferença no peso ou altura médio do indivíduo, ou seja, ajusta a curva de crescimento médio deslocando-a verticalmente para cima ou para baixo. Dessa forma valores positivos ou negativos do tamanho indicam se a criança é mais pesada/maior ou mais leve/menor que o peso ou altura médio da amostra. Por sua vez, o parâmetro β_i refere-se ao intercepto aleatório da idade que se ajusta para os diferentes momentos de mudanças do crescimento. O parâmetro tempo (β_i) controla a mudança horizontal em relação à curva média. Os valores positivos ou negativos deste parâmetro indicam se o tempo do pico da velocidade do peso ou altura é anterior ou posterior ao pico da curva média. Por fim, o parâmetro γ_i se refere à escala de idade aleatória que é responsável por diferentes durações de saltos de crescimento. O parâmetro velocidade (γ_i) corresponde ao encolhimento ou alongamento da curva na escala de idade. Valor positivo da velocidade indica o encolhimento da escala de idade, o que torna a curva mais íngreme sugerindo um crescimento rápido. Por outro lado, o valor negativo da velocidade indica o alongamento da escala de idade que torna a curva mais rasa e um crescimento lento.

O comando *dfset* disponível no pacote *sitar* (COLE; COLE, 2021) no programa R foi utilizado para determinar o número apropriado de graus de liberdade (nós) para uma curva *spline*

natural que minimize o *Bayesian information criteria* (BIC). A avaliação dos diferentes modelos foi realizada comparando *Akaike information criteria* (AIC), o BIC e o desvio padrão residual (RSD), em que quanto menor o valor de BIC melhor é o ajuste do modelo (COLE; COLE, 2021; HAUSPIE; CAMERON; MOLINARI, 2004).

A variação percentual no peso ou altura foi explicada por cada modelo SITAR ajustado usando a percentagem de variância explicada igual a $100 \times \left(1 - (\sigma_2/\sigma_1)^2\right)$ onde σ_1 é o desvio padrão residual de efeitos fixos (RSD) e σ_2 o RSD de efeitos aleatórios SITAR, ou seja, é o desvio padrão da distribuição dos pontos de dados em torno da curva ajustada (HAUSPIE; CAMERON; MOLINARI, 2004).

2.2.6 Análise de Resíduo

Para examinar as premissas dos modelos ajustados, foi realizado a análise diagnóstica gráfica dos resíduos. Os modelos lineares mistos possuem duas fontes de variação (ζ e ε), e podem ser definidos três tipos de resíduos, a saber:

- Resíduos condicionais: $\hat{\varepsilon} = \mathbf{Y} - \mathbf{X}\hat{\beta} - \mathbf{Z}\hat{\zeta}$
- Resíduos aleatórios: $\mathbf{Z}\hat{\zeta}$
- Resíduos marginais: $\hat{\xi} = \mathbf{Y} - \mathbf{X}\hat{\beta}$

Segundo Nobre (2004), em modelos lineares mistos, os resíduos condicionais e aleatórios não são considerados resíduo puro, onde ele depende apenas dos componentes fixos do modelo e do respectivo erro do qual ele é preditor, ou seja, quando ele depende de dois ou mais erros é considerado resíduos confundidos. No caso do modelo mistos ele é confundido pela presença de ε e ζ . O gráfico dos resíduos condicionais versus os valores ajustados avalia as suposições de homocedasticidade do erro condicional e pode ser ainda utilizado para identificar observações discrepantes. Por sua vez o gráfico QQplot dos resíduos com confundimento mínimo é sugerido para avaliar a normalidade dos erros condicionais porque ele é confundido por ζ e quando ζ se afasta muito da normalidade $\hat{\varepsilon}$ pode não apresentar características de normalidade mesmo quando o ε tem uma distribuição normal. Uma discussão detalhada sobre esses resíduos pode ser encontrada em (NOBRE; SINGER, 2007). Os gráficos foram gerados mediante a função `residdiag_nlme_()` do *software* R.

Os parâmetros do modelo *Brokenstick* são estimados pela função `lmer(.)` do pacote **lme4** que tem a função `lmmresid(.)` implementada por Nobre e Singer (2007) para análise de resíduos em modelos de efeitos mistos. Não foi possível utilizar essa função pois não é possível extrair a matriz de covariância marginal. Dessa forma foi utilizado o pacote **robustlmm** do R para gerar os gráficos dos resíduos condicionais e probabilidade normal dos erros condicionais. Este último, pode não ser o adequado para verificar a normalidade dos resíduos condicionais pois é

um resíduo confundido, a proposta seria utilizar gráfico QQplot dos resíduos com confundimento mínimo (NOBRE; SINGER, 2007). Foi utilizado ainda, o gráfico de dispersão entre os valores observados versus valores ajustados, para avaliar a qualidade do ajuste.

Para avaliar a validade das suposições do modelo SITAR foram analisados os gráficos dos resíduos condicionais padronizados contra os valores ajustados para verificar a suposição da variância constante dos resíduos e, os gráficos de probabilidade normal foram usados para avaliar a normalidade dos erros residuais. Foi utilizado o pacote **robustlmm** do programa R.

2.3 Abordagem Multivariada - Cópulas

Com dados longitudinais multivariados surge a oportunidade de estudar a evolução conjunta de múltiplas variáveis respostas para o mesmo indivíduo. Segundo Cho (2016), o estudo de dados longitudinais multivariados exige métodos mais complexos e desafiadores, pois além da existência da dependência entre as medidas repetidas no mesmo indivíduo considera-se a correlação entre as diversas variáveis respostas que pode mudar com o tempo. Além disso, a correlação natural existente em dados longitudinais geralmente dificulta a especificação da função de verossimilhança completa. Uma alternativa de análise são as cópulas, que são funções de distribuição conjunta que vêm sendo utilizadas na modelagem multivariada da estrutura de dependência no contexto longitudinal. (NELSEN, 2007; CHO, 2016)

A abordagem de Cópula tem sido explorada em diversas áreas do conhecimento, pois podem ser empregadas mesmo quando há violação da suposição de normalidade multivariada dos dados, além de capturar dependência não linear e de cauda entre as variáveis. A aplicação de cópulas é bem abrangente, com áreas como finanças, economia, ciências ambientais, engenharia, Saúde Pública etc. Alguns exemplos de aplicação como Louzada *et al.* (2013), Cho (2016), Ferreira *et al.* (2019) e Liu *et al.* (2020). No presente trabalho, essa abordagem oferece a possibilidade de estudar relações de dependência entre as variáveis respostas nos padrões de crescimento infantil, os quais ainda não foram muito exploradas até o momento

Será apresentada uma breve revisão dos conceitos e definições fundamentais em Teoria das Cópulas e algumas famílias paramétricas de cópulas mais conhecidas.

2.3.1 Conceitos e Definições

Formalmente, uma cópula C é uma distribuição multivariada cujas marginais são uniformemente distribuídas com suporte no intervalo $(0, 1)$ (NELSEN, 2007). O primeiro estágio da modelagem de cópula consiste basicamente em ajustar os modelos das variáveis aleatórias univariadas pela distribuição marginal. Em seguida, estes ajustes são utilizados para estimar a dependência multivariada, o que torna uma boa alternativa quando a normalidade multivariada não é garantida (FERREIRA *et al.*, 2019).

Segundo Nelsen (2007), a definição de uma Cópula é dada por:

Definição 2.1. Dizemos que C é uma cópula com distribuição m -dimensional quando existe uma função $C : [0, 1]^m \rightarrow [0, 1]$, de modo que suas marginais univariadas U_j tenham distribuição uniforme no intervalo $(0, 1)$, ou seja, $U_j \sim U(0, 1)$, para $j = 1, 2, \dots, m$

Na abordagem bivariada ($m=2$), definimos uma cópula bivariada de uma função de distribuição bivariada $C : [0, 1]^2 \rightarrow [0, 1]$, com distribuições marginais $U(0, 1)$ isto é,

$$C(u; v) = P(U \leq u; V \leq v)$$

que satisfaz às seguintes condições.

$$C(u, 0) = C(0, v) = 0$$

e

$$C(u, 1) = u \text{ e } C(1, v) = v$$

Para todo $u_1, u_2, v_1, v_2 \in [0, 1]$ tal que $u_1 \leq u_2$ e $v_1 \leq v_2$

$$C(u_2, v_2) - C(u_2, v_1) - C(u_1, v_2) + C(u_1, v_1) \geq 0$$

As variáveis U e V são variáveis cujas distribuições marginais são originadas de uma transformação especial, chamada de transformação por integral de probabilidade (*probability integral transform*) que pode ser obtida de variáveis aleatórias contínuas quaisquer, por exemplo, sendo X e Y variáveis aleatórias contínuas com $U = F(x)$ e $V = G(y)$, onde $F(x)$ e $G(y)$ representam a função de distribuição acumulada (FDA).

O Teorema Sklar's é o teorema central para a teoria das cópulas, em especial nas aplicações na área de estatística, e esclarece o papel que as cópulas desempenham na relação entre as funções de distribuição multivariada e suas marginais univariadas (NELSEN, 2007).

Teorema 2.1. Teorema de Sklar's. No contexto bivariado, esse teorema pode ser estabelecido como sendo X e Y duas variáveis aleatórias e $H(x, y)$ uma função de distribuição conjunta com as marginais $F(x)$ e $G(y)$. Então existe uma cópula C tal que satisfaz

$$H(x, y) = C(F(x), G(y))$$

para todo $x, y \in \mathbb{R}$,

Se $F(x)$ e $G(y)$ são contínuas, então C é única. Se C é uma cópula e $F(x)$ e $G(y)$ são funções de distribuição, então a função $H(x, y)$ é uma função de distribuição conjunta com as

marginais $F(x)$ e $G(y)$. Assim é possível construir uma função de distribuição bivariada em termos de duas funções de distribuição contínua univariada e uma cópula permitindo relações de dependência entre as duas variáveis aleatórias individuais (NELSEN, 2007; ESCARELA; HERNÁNDEZ, 2009).

O teorema a seguir, estabelece intervalo de variação de uma cópula.

Teorema 2.2. (*Limites de Fréchet-Hoeffding*):

As cópulas possuem limites de variação para as funções de distribuição, conhecidos como limites de Fréchet-Hoeffding (NELSEN, 2007), sendo sua versão bivariada escrita como:

$$\max \{0, F(x) + G(y) - 1\} \leq C(F(x), G(y)) \leq \min \{F(x), G(y)\}$$

A função $\max \{0, F(x) + G(y) - 1\}$ é denominada *Fréchet-Hoeffding upper bound*, ou seja, limite superior de *Fréchet-Hoeffding* que representa a dependência positiva perfeita, enquanto, analogamente, $\min \{F(x), G(y)\}$ é chamada de *Fréchet-Hoeffding lower bound*, ou seja, limite inferior de *Fréchet-Hoeffding* que representa dependência negativa perfeita.

A cópula captura toda informação de dependência entre as variáveis, porém sua distribuição conjunta não informa especificamente como se comporta essa dependência e nem sua mensuração, sendo necessário o uso de medidas capazes de quantificar essa associação (FERREIRA et al., 2019). O coeficiente de correlação linear de Pearson é a medida mais conhecida para quantificar o grau de dependências entre variáveis, entretanto em aplicações de cópulas, as medidas de associação mais conhecidas são o tau de Kendall e o rho de Spearman que possuem várias vantagens quando comparadas ao coeficiente de correlação de Pearson (IDA; ISHIMURA; NAKAMURA, 2014). Segundo Trivedi & Zimmer (2005), o coeficiente de correlação de Pearson é linear, não podendo ser utilizado na mensuração da dependência de funções não-lineares. Outra limitação está no fato de não ser definido para distribuições de caudas pesadas, além de não ser invariante.

Antes da definição do tau de Kendall (τ) e o rho de Spearman (ρ) será apresentado a definição de concordância onde tais medidas se baseiam.

Definição 2.2. (*Concordância*):

(i) *Duas observações (x_1, y_1) e (x_2, y_2) são concordantes se $x_1 \leq x_2$ e $y_1 \leq y_2$ ou se $x_1 \geq x_2$ e $y_1 \geq y_2$. De forma equivalente $(x_1 - x_2)(y_1 - y_2) > 0$. As observações (x_1, y_1) e (x_2, y_2) são discordantes se $(x_1 - x_2)(y_1 - y_2) < 0$.*

(ii) *Se C_1 e C_2 são cópulas, então C_1 é menos concordante que C_2 (ou C_2 é mais concordante que C_1) se $C_1(u, v) \leq C_2(u, v)$ para todo $(u, v) \in [0, 1]^2$.*

Definição 2.3. (*tau de Kendall*):

Seja (X_1, X_2) e (Y_1, Y_2) dois vetores aleatórios. A medida de associação tau de Kendall é definida como a diferença entre as probabilidades de concordância e discordância, ou seja,

$$\tau = \tau_{X,Y} = P[(X_1 - X_2)(Y_1 - Y_2) > 0] - P[(X_1 - X_2)(Y_1 - Y_2) < 0]$$

De acordo com (NELSEN, 2007), se (X,Y) é um vetor aleatório com variáveis contínuas e cuja cópula é C , então o tau de Kendall é dado por:

$$\tau_{X,Y} = 4 \int_0^1 \int_0^1 C(u, v) dC(u, v) - 1$$

Assim como no tau de Kendall, a versão populacional da medida de associação de rho de Spearman é baseada no conceito de concordância e discordância.

Definição 2.4. (*rho de Spearman*):

Seja (X_1, Y_1) , (X_2, Y_2) , e (X_3, Y_3) três vetores aleatórios independentes. A medida populacional de associação rho de Spearman é definido para ser proporcional à probabilidade de concordância menos a probabilidade de discordância para os dois vetores, ou seja,

$$\rho_{X,Y} = 3 (P[(X_1 - X_2)(Y_1 - Y_3) > 0] - P[(X_1 - X_2)(Y_1 - Y_3) < 0])$$

podendo ser usado o par (X_3, Y_2)

De acordo com (NELSEN, 2007), se X e Y são variáveis aleatórias contínuas e cuja cópula é C , então o rho de Spearman é dado por:

$$\rho_{X,Y} = 12 \iint_{[0,1]^2} C(u, v) dudv - 3$$

Outro conceito relevante em cópula é a dependência caudal, que mede a dependência (ou grau de associação) entre as variáveis do quadrante superior direito e do quadrante inferior esquerdo de uma distribuição bivariada (NELSEN, 2007).

Definição 2.5. (*Dependência caudal*):

Sejam X e Y variáveis aleatórias contínuas com função de distribuição F e G respectivamente. O coeficiente de dependência caudal inferior é definido como:

$$\lambda_L = \lambda_L(X, Y) = \lim_{t \rightarrow 1^-} P[Y > G^{(-1)}(t) \mid X > F^{(-1)}(t)] = \lim_{t \rightarrow 1^-} \frac{C(t, t)}{t}$$

O coeficiente de dependência caudal superior é dado por:

$$\lambda_U = \lambda_U(X, Y) = \lim_{t \rightarrow 1^-} P(Y > G^{-1}(t) \mid X > F^{-1}(t)) = \lim_{t \rightarrow 1^-} \frac{1 - 2t + C(t, t)}{1 - t}$$

em que G^{-1} e F^{-1} são as inversas de G e F respectivamente.

Outra definição importante para a composição da função de verossimilhança é a densidade de cópula.

Definição 2.6. (*Densidade de cópula*):

Se $C(u, v) = \int_0^u \int_0^v \frac{\partial^2 C(s, t)}{\partial s \partial t} dt ds$, para todo $(u, v) \in [0, 1]^2$ então C é absolutamente contínua e a densidade de cópula c pode ser definida por

$$c(u, v) = \frac{\partial^2 C(u, v)}{\partial u \partial v}$$

Dessa forma, a abordagem de cópula necessita da definição de uma família de cópula C e de suas características (limites de *Fréchet-Hoeffding*, medidas de associação e sua dependência caudal)

2.3.2 Famílias Paramétricas de Cópulas

Um problema essencial é a escolha adequada da família de cópula para modelar a estrutura de dependência dos dados. Essas famílias não são equivalentes em termos do tipo de dependência estocástica que representam, ou do grau de dependência que podem capturar (NELSEN, 2007). As cópulas mais utilizadas na literatura são as elípticas e as arquimedianas.

As cópulas elípticas são originárias de distribuições elípticas multivariadas, sendo a Gaussiana e a t-Student as mais utilizadas (FRAHM; JUNKER; SZIMAYER, 2003). As cópulas arquimedianas, possuem boas propriedades matemáticas e a sua construção é por meio de funções convexas, sendo Clayton, Frank e Gumbel as mais utilizadas (NELSEN, 2007).

A seguir, serão apresentados a função de distribuição, o espaço paramétrico, tau de Kendall, rho de spearman e coeficientes de dependência caudal das famílias paramétricas de cópulas bivariadas consideradas neste trabalho. Considere $u, v \in [0, 1]^2$.

2.3.2.1 Clayton

A cópula Clayton bivariada é uma cópula Arquimediana cuja função de distribuição é dada por:

$$C(u, v; \theta) = \left(u^{-\theta} + v^{-\theta} - 1 \right)^{-\frac{1}{\theta}} \quad (2.11)$$

em que $\theta \in (0, \infty)$. A dependência, apenas positiva, entre as marginais aumenta com o valor de θ . Quando $\theta \rightarrow \infty$, a cópula de Clayton atinge o limite superior de *Fréchet-Hoeffding* (dependência positiva perfeita). Quando $\theta \rightarrow 0^+$, a cópula de Clayton se reduz à cópula produto, que é definida como $C(u, v) = uv$ (independência).

A cópula de Clayton exhibe dependência na cauda inferior apenas, com $\lambda_L = 2^{-\frac{1}{\theta}}$ e $\lambda_U = 0$. A relação entre o tau de Kendall (τ) e o parâmetro de dependência da cópula de Clayton

(θ) é dada pela expressão:

$$\tau_K = \frac{\theta}{\theta + 2}$$

. Por fim, rho de Spearman não possui expressão fechada (FERREIRA et al., 2019).

2.3.2.2 Frank

A cópula de Frank é uma cópula Arquimediana que permite dependência negativa e é simétrica em ambas as caudas. Seu uso é atrativo, pois pode capturar toda a gama de dependência. É uma cópula bastante utilizada em aplicações empíricas. Sua função de distribuição bivariada é dada por:

$$C(u, v; \theta) = -\frac{1}{\theta} \log \left(1 + \frac{(e^{-\theta u} - 1)(e^{-\theta v} - 1)}{(e^{-\theta} - 1)} \right) \quad (2.12)$$

tal que $\theta \in \mathbb{R} - \{0\}$. A cópula de Frank não exibe dependência nas caudas, ou seja, $\lambda_L = \lambda_U = 0$. Esta cópula inclui o limite superior de *Fréchet* quando $\theta \rightarrow -\infty$, limite inferior de *Fréchet* quando $\theta \rightarrow \infty$, e de independência quando $\theta \rightarrow 0$.

As medidas de associação de tau de Kendall e rho de Spearman são respectivamente,

$$\tau_K = 1 - \frac{4[D_1(\theta) - 1]}{\theta}$$

e

$$\rho_S = 1 - \frac{12[4D_1(\theta) - 2D_2(\theta)]}{\theta}$$

em que $D_m(\theta) = \frac{m}{\theta^m} \int_0^\theta \frac{t^m}{\exp\{t\}-1} dt$, é chamada de função Debye de ordem m , $m = 1, 2$ (SUZUKI, 2012).

2.3.2.3 Gumbel

A cópula de Gumbel é uma cópula Arquimediana e sua forma bivariada é dada por:

$$C(u, v; \theta) = \exp \left\{ - \left([-\log(u)]^\theta + [-\log(v)]^\theta \right)^{1/\theta} \right\}. \quad (2.13)$$

em que $\theta \in (1, \infty)$. A dependência, apenas positiva, entre as marginais aumenta com o valor de θ . Quando $\theta \rightarrow \infty$ implica em dependência positiva perfeita, enquanto que para $\theta = 1$, tem-se a cópula produto.

A cópula de Gumbel exibe dependência na cauda superior somente, com $\lambda_U = 2 - 2^{\frac{1}{\theta}}$.

A medida de associação de tau de Kendall é dada por:

$$\frac{\theta - 1}{\theta}$$

Assim como a cópula de Clayton, não possuem forma fechada para o rho de Spearman (FERREIRA et al., 2019).

2.3.2.4 Gaussiana

A cópula Gaussiana bivariada é uma cópula elíptica representada pela função de distribuição da normal bivariada com vetor de médias $(0, 0)$ e matrix de correlação $\begin{bmatrix} 1 & \theta \\ \theta & 1 \end{bmatrix}$, com $\theta \in [-1, 1]$, dada pela expressão:

$$C(u, v; \theta) = \int_{-\infty}^{\Phi^{-1}(u)} \int_{-\infty}^{\Phi^{-1}(v)} \frac{1}{2\pi(1-\theta^2)^{\frac{1}{2}}} \exp \left\{ -\frac{s^2 - 2\theta st + t^2}{2(1-\theta^2)} \right\} dt ds \quad (2.14)$$

em que $\Phi^{-1}(\cdot)$ corresponde à função quantil da distribuição normal padrão.

Na cópula Gaussiana, a estrutura de dependência entre as marginais cresce com o valor do $|\theta|$, sendo que $\theta = -1$ implica em dependência negativa perfeita e $\theta = 1$ implica em dependência positiva perfeita, ao passo que $\theta = 0$ implica em independência.

A cópula Gaussiana não exibe dependência nas caudas, isto é, $\lambda_L = \lambda_U = 0$, as medidas associação tau de Kendall e rho de Spearman são dadas por: $\tau_K = \sin\left(\frac{\pi}{2}\tau\right)$ e $\rho_S = \frac{\arcsin(\theta/2)}{\pi/6}$, respectivamente (FERREIRA et al., 2019).

2.3.2.5 t-Student

A cópula t-Student (conhecida também como cópula t) bivariada é uma cópula elíptica definida por:

$$C(u, v; \theta) = \int_{-\infty}^{T_v^{-1}(u)} \int_{-\infty}^{T_v^{-1}(v)} \frac{1}{2\pi(1-\theta^2)^{\frac{1}{2}}} \times \left(1 + \frac{s^2 - 2\theta st + t^2}{v(1-\theta^2)} \right)^{-\frac{v+2}{2}} dt ds \quad (2.15)$$

em que $T_v^{-1}(\cdot)$ representa a função quantil da distribuição t-Student com ν graus de liberdade e $\theta \in [-1, 1]$.

Uma cópula Gaussiana bivariada com coeficiente de correlação θ pode ser considerada como o caso limite de uma cópula t bivariada com o mesmo parâmetro θ quando $\nu \rightarrow \infty$.

A dependência nas caudas é calculada por $\lambda_L = \lambda_U = 2T_{\nu+1}\left(-\sqrt{\frac{(\nu+1)(1-\theta)}{1+\theta}}\right)$, em que $T_{\nu+1}(\cdot)$ representa a função de distribuição acumulada da t-Student com $(\nu+1)$ graus de liberdade.

As medidas de associação tau de Kendall e rho de Spearman são dadas, respectivamente, por:

$$\tau_K = \frac{\arcsin(\theta)}{\pi/2}$$

e

$$\rho_S = \frac{\arcsin(\theta/2)}{\pi/6}$$

2.3.3 Inferência

Para estimação de parâmetros dos modelos baseados em cópulas, foi utilizado um método paramétrico de estimação clássica conhecido como método da inferência para as marginais (ou IFM, sigla do inglês "*Inference Function for Margins*"), proposto por Joe e Xu (1996). O método IFM tem por finalidade obter as estimativas dos parâmetros dos modelos derivados de cópulas em duas etapas. Na primeira etapa, estima-se (via método da MVR, por exemplo), e na segunda etapa, uma pseudo log-verossimilhança é utilizada para estimar o parâmetro da cópula θ (JOE; XU, 1996).

Joe e Xu (1996), resume algumas vantagens do método IFM em dois estágios.

1. O método IFM torna a inferência para muitos modelos multivariados computacionalmente viável.
2. Permite fazer inferência e modelagem, a partir de marginais univariadas e de menor dimensão.
3. Há alguma robustez contra a especificação incorreta da estrutura de dependência, bem como mais robustez contra outliers ou perturbações dos dados, em comparação com o método MV.
4. Dados multivariados esparsos podem criar problemas para o método de máxima verossimilhança, porém o método IFM evita o problema de esparsidade até certo ponto, especialmente se todos os parâmetros puderem ser estimados a partir de probabilidades univariadas e bivariadas (isso pode ser uma grande vantagem em uma situação de amostra menor).

Na primeira etapa, o método IFM estima os parâmetros marginais $\delta_j, j = 1, 2$ através

$$\hat{\delta}_j = \arg \max_{\delta_j} \log L_j(\delta_j | \mathbf{y}_j)$$

Então, obtém-se $\hat{u}_{it} = F(y_{1it} | \hat{\delta}_1)$ e $\hat{v}_{it} = G(y_{2it} | \hat{\delta}_2)$, para estimar o parâmetro de associação θ usando a pseudo log-verossimilhança a seguir:

$$\hat{\theta} = \arg \max_{\theta} \sum_{i=1}^m \sum_{t=1}^{n_i} \log c(\hat{u}_{it}, \hat{v}_{it} | \theta)$$

em que m é quantidade de variáveis para a t – ésima observação $t = 1, \dots, n_i$ e c é a densidade da cópula utilizada.

Para estimação dos parâmetros pelo método IFM dos modelos de crescimento derivados das cópulas bivariadas utiliza-se a função `optim(.)` (método “L-BFGS-B”) do programa R.

O procedimento de reamostragem bootstrap será adotado para obter os erros-padrões dos estimadores IFM para os parâmetros dos modelos de cópulas propostos (JOE; XU, 1996).

Será aplicado um procedimento bootstrap paramétrico, tal qual Ferreira et al. (2019), em que 500 novas amostras bootstrap são geradas dos modelos estimados. Em seguida, para cada uma das amostras geradas, são obtidas as estimativas IFM dos parâmetros e, por último, calcula-se os desvios-padrões amostrais dessas estimativas (que, neste caso, correspondem às estimativas bootstrap dos erros-padrões).

3 Descrição do banco de dados

As bases de dados utilizadas neste estudo foram disponibilizadas pelo Centro de Integração de Dados e Conhecimentos para Saúde (CIDACS), criado pela portaria 32-2016, vinculado à diretoria da Fundação Oswaldo Cruz (Fiocruz) no estado da Bahia. O CIDACS realiza estudos e pesquisas com base em projetos interdisciplinares originados na vinculação de grandes volumes de dados governamentais de seu domínio e são armazenados e preservados com base em procedimentos de Curadoria Digital.

O CIDACS tem em seu domínio grandes volumes de dados divididos em plataformas: Coorte de 100 Milhões de Brasileiros; Plataforma Zika; Tecnologias e Inovações para o SUS; Equidade e Sustentabilidade Urbana; Bioinformática e Epidemiologia Genética (Epigen). Este centro tem como missão contribuir para a produção de conhecimentos científicos inovadores para ampliar o entendimento dos determinantes e das políticas sociais e ambientais sobre a saúde da população, visando auxiliar e apoiar a tomada de decisões em políticas públicas em benefício da sociedade. No projeto de pesquisa serão utilizadas as bases de dados que pertencem à plataforma da Coorte de 100 milhões de brasileiros, veiculando as bases do Cadastro Único, Sistema de Vigilância Alimentar e Nutricional (SISVAN) e do Sistema de Informações sobre Nascidos Vivos (SINASC).

3.1 Fonte de Dados

Na realização deste estudo foi utilizado como fontes de dados os registros administrativos disponíveis no Cadastro Único para Programas Sociais do Governo Federal (CadÚnico), Sistema de Vigilância Alimentar e Nutricional (SISVAN) e o Sistema de Informação de Nascidos Vivos (SINASC). A seguir será apresentada uma síntese de cada uma destas fonte de dados.

- CadÚnico: O CadÚnico é um instrumento que identifica e caracteriza as famílias de baixa renda, entendidas como aquelas que têm renda mensal de até meio salário mínimo por pessoa ou renda mensal total de até três salários mínimos. O Cadastro Único permite conhecer a realidade socioeconômica dessas famílias, trazendo informações de todo o núcleo familiar, das características do domicílio, das formas de acesso a serviços públicos essenciais e, também, dados de cada um dos componentes da família. Sua coordenação está

a cargo do Ministério do Desenvolvimento Social e Combate à Fome (MDS), devendo ser obrigatoriamente utilizado para seleção de beneficiários de programas sociais do Governo Federal, como o Programa Bolsa Família, Programa Minha Casa Minha Vida, Bolsa Verde (Programa de Apoio à Conservação Ambiental), Programa de Erradicação do Trabalho Infantil (PETI), Fomento (Programa de Fomento às Atividades Produtivas Rurais), Carteira do Idoso, Aposentadoria para pessoa de baixa renda, Programa Brasil Carinhoso, Programa de Cisternas, Telefone Popular, Carta Social, Pro Jovem Adolescente, Tarifa Social de Energia Elétrica, Passe Livre para pessoas com deficiência, Isenção de Taxas em Concursos Públicos (CEF, 2018; MDS, 2018).

- **SISVAN:** A base de dados do SISVAN é composta pelos registros de acompanhamento provenientes do Sistema de Gestão do Programa Bolsa Família (SIGPBF), do e-SUS AB e dos registros inseridos no próprio SISVAN (BRASIL, 2017). As informações referentes ao acompanhamento do estado nutricional dos beneficiários do PBF, que ocorre duas vezes por ano, no mínimo, são incorporadas ao SISVAN no final de cada vigência (as primeiras quatro vigências de janeiro a junho e segunda vigência de julho a dezembro). Já os registros provenientes do e-SUS AB são incorporados gradativamente ao Sistema de Informação para a Atenção Básica (SISAB), respeitando o cronograma de envio de dados pelas equipes de Atenção Básica, e posteriormente exportados para o SISVAN após processamento e validação dos dados. Essa rotina de integração dos sistemas tem sido realizada pelo Núcleo de Tecnologia da Informação (NTI/DAB/SAS/MS), em conjunto com o Departamento de Informática do Sistema Único de Saúde (DATASUS), obtendo como produto final a consolidação dos dados na base do SISVAN (BRASIL, 2017). As medidas antropométricas coletadas e incorporadas às bases de dados do SISVAN abrangem o peso e o comprimento/altura para todas as fases de vida, peso ao nascer para menores de dois anos, circunferência da cintura para adultos, circunferência da panturrilha para idosos e peso pré-gestacional para gestantes.
- **SINASC:** É um sistema de informação que contém informações detalhadas das mães e dos recém-nascidos de todo território brasileiro. O SINASC, foi implantado oficialmente a partir de 1990, de forma gradual, com o objetivo de reunir informações epidemiológicas dos nascimentos informados em todo território nacional e fornecer dados sobre natalidade para todos os níveis do Sistema de Saúde (BRASIL, 2018). Desde o ano de 1994, alguns municípios já apresentavam um número maior de registros do que o publicado pelo IBGE com base nos dados de Cartório de Registro Civil, revelando a alta cobertura do SINASC (CGIAE/DASIS). Este sistema possibilita a construção de indicadores úteis para o planejamento de gestão dos serviços de saúde. Entre os benefícios que podem ser atribuídos à esta base de dados destaca-se: subsidiar as intervenções relacionadas à saúde materno-infantil para todos os níveis do Sistema Único de Saúde (SUS); subsidiar ações de atenção à gestante e ao recém-nascido; e identificar as prioridades de intervenção, com

acompanhamento da evolução das séries históricas do SINASC (BRASIL, 2018).

Declaração de Nascido Vivo (DN) é o instrumento que alimenta o SINASC e constitui-se em um formulário padronizado, impresso em três vias previamente numeradas, distribuído gratuitamente para as Secretarias Estaduais de Saúde que são repassadas às Secretarias Municipais de Saúde e, por sua vez, enviam aos estabelecimentos de saúde e cartórios. A sua distribuição e impressão, possíveis alterações no *layout* e os manuais do sistema são de responsabilidade da Secretaria de Vigilância em Saúde (SVS/MS), mais especificamente da Coordenação Geral de Informações e Análise Epidemiológica e Departamento de Análises e Situação em Saúde (CGIAE/DASIS).

3.2 Integração e desidentificação das bases de dados

Nesta seção será descrito, o processo de integração da base de dados da Coorte Brasileira de 100 milhões, o SISVAN e SINASC realizado pela equipe de produção de dados do CIDACS.

A primeira etapa do processo de integração foi a concatenação das bases de dados do SISVAN, separadas por ano ou meses. Em outras palavras, essa etapa consiste em reunir todos os dados, referentes aos anos de 2008 a 2017, em uma única base, adicionando as observações de uma base ao final de outra base. A segunda etapa de integração é o procedimento de *linkage* (relacionamento probabilístico), onde Coorte Brasileira de 100 milhões foi o banco chave para fazer a ligação entre as bases SISVAN e SINASC. Duas abordagens foram usadas para realizar a vinculação de registros entre as três fontes: a ligação determinística e a ligação não determinística baseada no índice de similaridade. A ligação determinística foi realizada entre a linha de base da Coorte Brasileira de 100 Milhões e o SISVAN porque ambos os conjuntos de dados contêm o Número de Informações Sociais (NIS), um identificador único atribuído a cada indivíduo. A ligação do SINASC com o SISVAN ocorreu usando uma abordagem não determinística. Este método também vinculou os dados da coorte com o subconjunto de indivíduos no SISVAN sem NIS. Para isso, criou-se uma ferramenta (CIDACS-RL) para vincular registros individuais com base em identificadores, como nome, sexo, idade ou data de nascimento, nome da mãe e município de residência (BARBOSA et al., 2020; ALI et al., 2019). A ferramenta CIDACS-RL aplica algoritmos de indexação e busca implementados na solução Apache Lucene como estratégia de bloqueio. A estratégia de indexação permite que o CIDACS-RL busque os registros mais semelhantes do banco de dados indexado para cada registro no banco de dados principal e os submeta à etapa de comparações par a par. Os registros de vinculação de candidatos são ordenados pelas pontuações de similaridade e apenas o par de comparação com a pontuação mais alta é retido como uma ligação potencial. Todos os registros de candidatos restantes são descartados (ALI et al., 2019; ALMEIDA et al., 2020).

As bases de dados identificadas foram processadas e armazenadas numa sala segura, que segue um protocolo interno de segurança de dados, onde somente os funcionários do Núcleo

de Produção de Dados do CIDACS têm acesso autorizado. Após o procedimento de *linkage*, a verificação de sua qualidade foi feita por meio da análise de acurácia. Por fim, a base passou pelo processo de desidentificação e dessa forma foi disponibilizada para a análise e construção das trajetórias de crescimento.

3.3 Descrição da abordagem metodológica e dos procedimentos

Trata-se de uma análise longitudinal dos dados antropométricos do SISVAN e do SINASC vinculados aos dados da linha de base da Coorte de 100 Milhões de Brasileiros, coordenada pelo CIDACS, Fundação Oswaldo Cruz (BARRETO et al., 2019).

A população de estudo é composta por crianças com idade entre 0 e 2 anos incompletos, com três ou mais medidas de peso e altura, avaliadas no período de 2008 a 2017. As trajetórias de crescimento foram avaliadas por meio dos dados brutos de peso e altura. Foram selecionadas e aplicadas técnicas de modelagem longitudinal, que são capazes de caracterizar os diferentes padrões de crescimento físico ao longo do tempo, que consideram diferenças inter e intra-individuais, e permitam analisar dados longitudinais desbalanceados e irregulares, incorporando a dependência e a estrutura de correlação dos erros.

Este trabalho pertence a um estudo maior, cujo projeto foi aprovado pelo comitê de ética do Instituto de Saúde Coletiva, Universidade Federal da Bahia (CAAE:41695415.0.0000.5030) com o objetivo de estudar o perfil de crescimento infantil entre os beneficiários do Programa Bolsa Família no período de 2008 a 2016.

3.3.1 Critérios de exclusões

A partir da veiculação das bases de dados, foram excluídos os prontuários com ausência de *idade gestacional* ao nascimento, *peso ao nascer* ou *sexo do recém-nascido*, para o qual não é possível avaliar o tamanho para a idade gestacional. Também foram excluídos pesos de nascimento implausíveis tais como menores 500g ou menores ou iguais a 6999g (VILLAR et al., 2014).

Crianças com idade gestacional abaixo de 37 semanas completas foram classificadas como prematuras e maior ou igual a 37 semanas, a termo. Neste estudo foram consideradas apenas a crianças classificadas como a termo (VILLAR et al., 2014).

A variável *peso ao nascer* foi categorizado como baixo peso (para valores abaixo de 2.500g) e peso normal ao nascer (entre 2500g e 6999g). Na tentativa de evitar algum tipo de viés na análise dos casos de baixo peso ao nascer, foram excluídos nascimentos múltiplos e nascidos vivos com anomalias congênitas que são condições associadas ao baixo peso ao nascer (CHEN et al., 2013)

Outros critérios de exclusão levaram em conta os escores z das variáveis antropométricas tais como: *peso por idade* (P/I) menor que -6 ou maior que 5, *estatura por idade* (E/I) menor que -6 ou maior que 6, *peso por estatura* (P/E) menor que -6 ou maior que 6 e *índice de massa corpórea* (IMC) por idade menor que -5 ou maior que 5, os quais são considerados implausíveis conforme recomendação da OMS (2011).

No banco de dados com estrutura longitudinal, a diferença da variável *altura* que apresentou valores negativos em medidas ordenadas(crescente) subsequentes, ou seja, se a diferença entre a medida atual e a medida posterior de uma mesma criança fosse menor que zero, esse valor era considerado implausível. De forma similar, foi tratado a diferença das datas de nascimento de uma mesma criança entre as medidas repetidas diferente de zero. Então, nestes caso, foi adotada para as variáveis *altura* e *data de nascimento*, a primeira medida como referência. Esse procedimento, foi repetido até que não houvesse mais diferença na medida de altura inferior a zero e data de nascimento diferente de zero totalizando 36 e 32 repetições, respectivamente. Além disso, crianças com menos de duas medidas repetidas de peso ou altura foram excluídas. Assim, a amostra final elegível totalizou 456.537 crianças.

Dessa forma, as variáveis selecionadas para a análise foram.

1. Variáveis dependentes

- Peso (Kg)
- Altura/Comprimento (cm)

2. Variáveis Independentes ou Preditoras

- Sexo (Masculino, Feminino)
- Idade (em anos ou em meses)
- Baixo peso ao nascer (Sim ou Não)
- Identificação da criança (ID)

Os modelos para as trajetórias de crescimento de cada variável dependente com as variáveis independentes acima foram ajustados no *software* estatístico R usando o pacote de nlme (PINHEIRO et al., 2017), *Brokenstick* (BUUREN, 2020), e *sitar* (COLE; COLE, 2021).

4 Resultados

Usando a teoria descrita no Capítulo 2, fase de processamento e metodologia proposta no Capítulo 3, vamos estimar trajetória de crescimento infantil para dados antropométricos de peso e altura. Inicialmente, nesta seção são apresentados os resultados obtidos de uma análise exploratória dos dados longitudinais sobre crescimento infantil de crianças menores de 2 anos; a descrição da estratégia de análise utilizada nos modelos marginais e bivariados (Subseção 4.1); os resultados obtidos a partir da abordagem univariada, pelo ajuste dos modelos univariados (Subseção 4.2); e, por último, os resultados obtidos a partir da abordagem bivariada (Subseção 4.3).

Foram incluídas no estudo 456.537 crianças do nascimento aos 2 anos de idade incompletos, registradas no período de 2008 a 2017 que totaliza 2.252,852 observações. A distribuição foi de aproximadamente 50% em cada um dos sexos (masculino $N=228683$, feminino $N=227854$) (após exclusões de dados especificadas anteriormente) e, dessas aproximadamente 3,6% foram classificados como baixo peso ao nascer. As características sociodemográficas maternas estão descritas na Tabela 1.

A caracterização do crescimento físico, individual e médio, das crianças do estudo foi realizada também por meio da análise gráfica, que descreve as trajetórias segundo o sexo. O gráfico de perfis é uma ferramenta descritiva importante na análise de dados longitudinais, e tem como objetivo verificar padrões individuais da variável resposta ao longo do tempo, bem como comparar a variabilidade inter e intra-crianças (SINGER; ANDRADE, 1986). Observa-se que a variabilidade no início é menor do que ao final do estudo, e essa variabilidade é maior no sexo feminino. A Figura 1 mostra os perfis individuais para o peso e altura por sexo. A linha sólida azul representa o perfil médio da variável resposta ao longo do tempo.

Dado que um dos objetivos é analisar a relação entre as variáveis, peso e altura, uma forma de descrever esta relação é descrevendo a dependência probabilística entre essas variáveis, sendo o motivo de estar considerando as cópulas como ferramenta de análise. A forma mais simples de dependência é a dependência linear, que pode ser medida a partir do coeficiente de correlação de Pearson, que revelou que o peso e altura apresenta uma correlação positiva e forte (0.871).

A Figura 2 exibe a idade registrada, em meses, para um subconjunto aleatório de 100

Tabela 1 – Característica sociodemográficas materna e do recém-nascido da população para crianças de 0 a 2 anos. Brasil 2008-2017

Variáveis	0 a 2 anos	
	<i>n</i> _{crianças}	%
Área de residência materna		
Urbana	323941	70.96
Rural	132585	29.04
Raça/cor da pele materna		
Branca/Amarela	89842	27.48
Parda	215385	65.89
Preta	19403	5.94
Indígena	2269	0.69
Escolaridade materna (anos)		
≤ 3	40859	9.13
4 - 7	153103	34.21
8 - 11	237388	53.05
≥ 12	16154	3.61
Estado civil materno		
Casada/união estável	232416	51.63
Solteira/divorciada/viúva	217735	48.37
Número de consultas pré natais		
Nenhuma	4092	0.90
1 - 3	18165	4.00
4 - 6	113658	25.00
≥ 7	318711	70.10
Sexo da criança		
Masculino	228683	50.09
Feminino	227854	49.91
Baixo Peso		
BPN	15758	3.45
PAN	440779	96.55
	<i>n</i> _{medidas}	MD (IIQ)
idade (meses)	2252852	16,312 (10,72-21,14)
medidas repetidas	2252852	3 (2 - 7)*

* número máximo de medidas repetidas = 40

IIQ - Intervalo Interquartil, MD - Mediana

crianças. Cada linha representa uma determinada criança e os pontos indicam os tempos das observações, sugerindo muita irregularidade nos espaçamentos entre as medidas e desbalanceamento do número de medidas. Nota-se ainda que os tempos de observação variam mais para as crianças mais velhas e que algumas crianças tiveram visitas razoavelmente próximas.

As análises de dados comparando as características de crescimento, para crianças a termo do nascimento até 2 anos de idade foram estratificadas por sexo e classificação de baixo peso ao nascer. Para o grupo das crianças a termo, que nasceram com peso adequado (PAN), a distribuição dos dados é aproximadamente igual entre os sexos com masculino N=222393(50,45%), e feminino

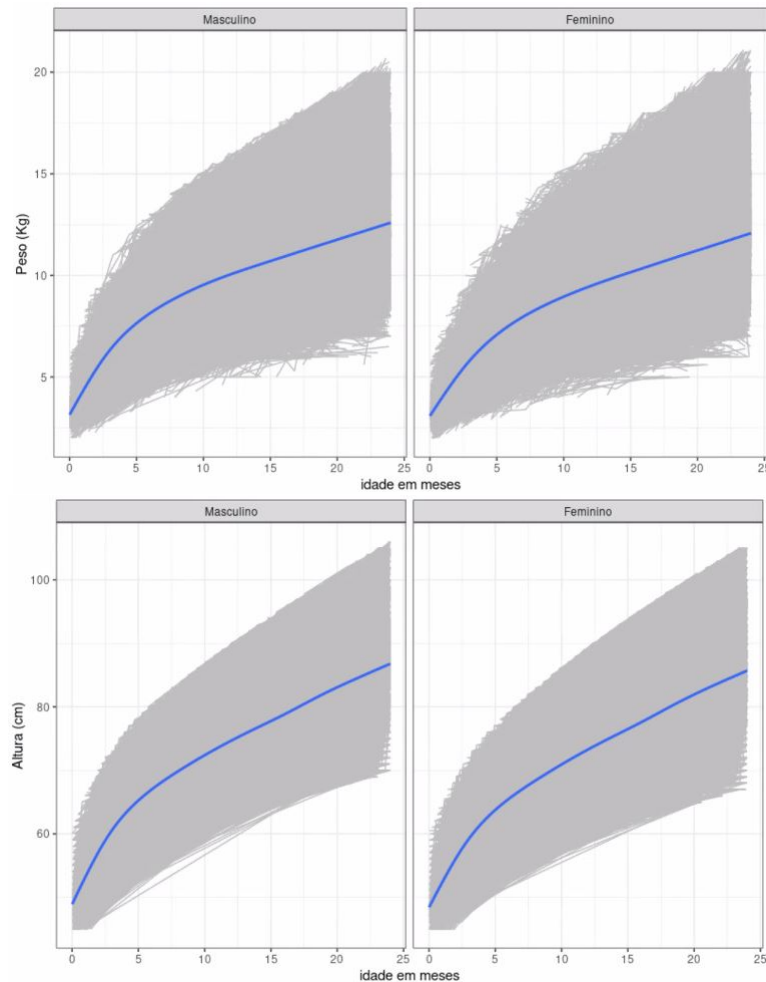


Figura 1 – Gráfico de Perfil de peso e altura por sexo de crianças menores de 2 anos, Brasil, 2008-2017

N=218386(49,55%), enquanto, que para o grupo com baixo peso ao nascer (BPN), a maior proporção é para o sexo feminino (masculino N=6290(39,92%), e feminino N=9468(60,08%)). Foram realizadas as etapas descritas abaixo para obter os modelos ajustados univariados e multivariados.

4.1 Estratégia de análise de dados para medições irregulares

Adotou-se como proposta de análise para dados irregulares a mesma abordagem usada por Buuren (2020) e Lukko *et al* (2020) que foi de criar uma grade de tempo pré-especificada, denominada de “idade alvo”. Ela deve ser idêntica para todos os indivíduos, mas não precisa ser equidistante. O comprimento do intervalo deve ser especificado segundo um contexto clínico ou um padrão de visita conhecido, dessa forma, neste estudo foi utilizado como referência, os dados de visitas previstas de puericultura e consultas de saúde (MD, 2002) por se tratar de dados administrativos do SISVAN. A seguir tem-se os passos utilizados para a criação dessa grade de

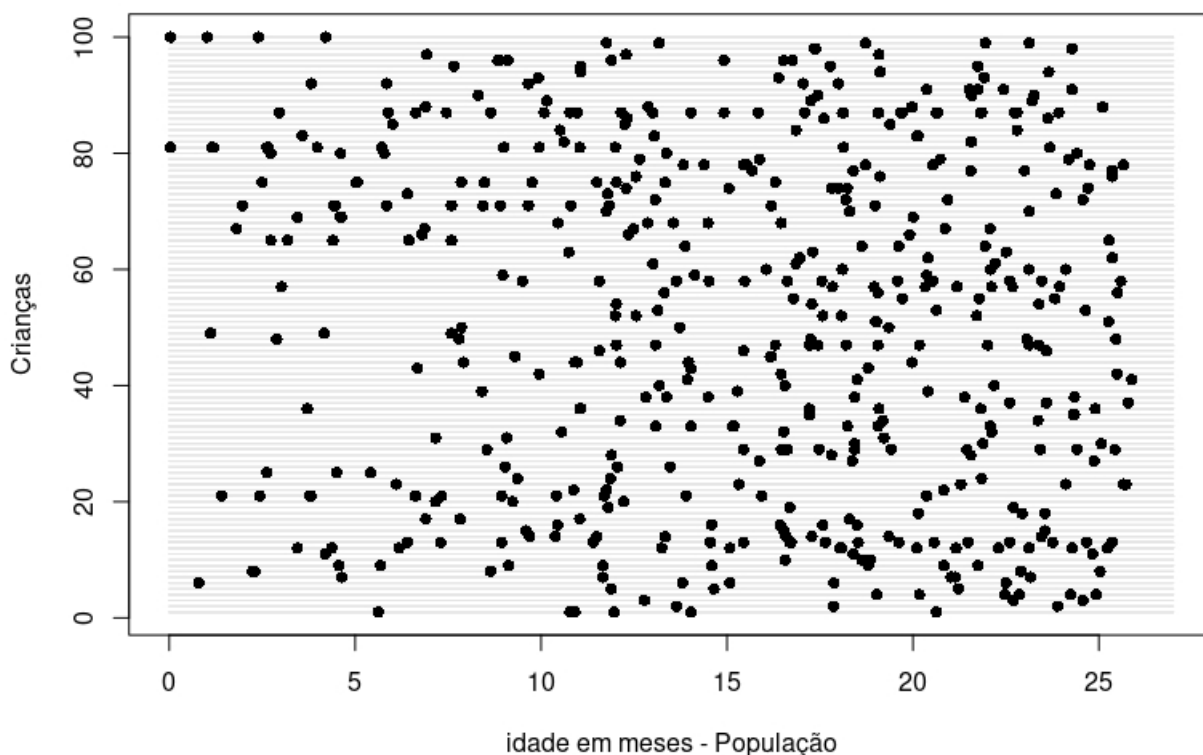


Figura 2 – Tempos de observação para um subconjunto aleatório de 100 crianças entre 0 e 2 anos do Banco Completo (Brasil, 2008-2017)

tempos de visita pré-especificados que foi denominada por - "idade alvo":

- a) Segundo a recomendação do Ministério da Saúde, o número mínimo de consultas de puericultura na Unidade Básica de Saúde (UBS), deve ocorrer aos 15 dias, 1, 2, 4, 6, 9, 12, 18, 24 meses e, após os 2 anos, a cada 12 meses até aos 6 anos de idade. Essas visitas também podem ocorrer quando necessário, caso a criança tenha algum problema de saúde por exemplo.
- b) Em seguida, definiram-se os intervalos em torno das idades recomendadas para visitas na Unidade Básica de Saúde com o objetivo de padronizar a idade, e na tentativa de colocar o espaçamento de maneira igualitária para todos. Consideraram-se os intervalos equidistantes em idades mais avançadas, (acima de 2 anos) aumentando a chance de ter pelo menos uma visita/medida na “idade alvo” (Tabela 2). Pode-se observar que os comprimentos dos intervalos não são iguais entre as idades-alvo. Por exemplo para 4 e 6 meses utilizaram-se intervalos com 30 dias, aos 18 meses, 60 dias e para 2 anos e mais 120 dias, dando mais chance de ter informações pertencentes a esses intervalos.

Na Figura 3, é apresentado o mesmo gráfico ábacus (Figura 2), com o resultado da aplicação da proposta de análise para tratar a irregularidade dos dados, neste caso excluíram-se as medições cujas idades encontraram-se fora do intervalo da idade alvo. Observa-se que as idades resultantes estão próximas das idades alvo. É observado, principalmente nas primeiras idades, que o banco de dados ainda apresenta desbalanceamento entre as medidas, mas a abordagem utilizada tem a tentativa de colocar o espaçamento entre as medições de maneira mais igualitária para todos.

- c) Adotou-se a mediana das medidas antropométricas peso e altura para as crianças que realizaram mais de uma visita à UBS em cada intervalo da “idade alvo”.

A escolha da mediana entre as mensurações foi baseada na redução do tempo computacional devido ao grande volume de dados e convergência dos modelos para ambos os sexos. Dessa forma tem-se três bancos de dados que foram estudados.

- “Completo” - base de dados completa;
- “Proposta” - base de dados podendo conter mais de uma mensuração em cada intervalo da "idade-alvo";
- “PropostaMed” - base de dados com as mensurações medianas das medidas antropométricas e idade em cada intervalo da "idade-alvo".

Tabela 2 – Proposta de análise para dados irregulares de crianças menores de 2 anos aninhada a coorte de 100 milhões de brasileiro. Brasil 2008 - 2017.

Idade - Alvo				
Meses	Dias	Intervalo (dias)	Amplitude em dias	Amplitude em meses
15 dias	15	10 a 17	7	0.2
1	30	18 a 45	27	0.9
2	60	46 a 75	29	1
4	120	105 a 135	30	1
6	180	165 a 195	30	1
9	270	255 a 285	30	1
12	360	350 a 380	30	1
18	540	510 a 570	60	2
24	720	670 a 790	120	4

Com objetivo de comparar os três bancos de dados e escolher aquele que viabilizaria os ajustes dos modelos de trajetória univariadas e bivariadas realizou-se o estudo do tempo de processamento e a estimativa das curvas de crescimento pelo modelo *Brokenstick*.

Nas Tabelas 3 e 4 são mostrados os números de observações e crianças, os tempos de processamentos e variância residual dos ajustes dos modelos nas variáveis antropométricas peso

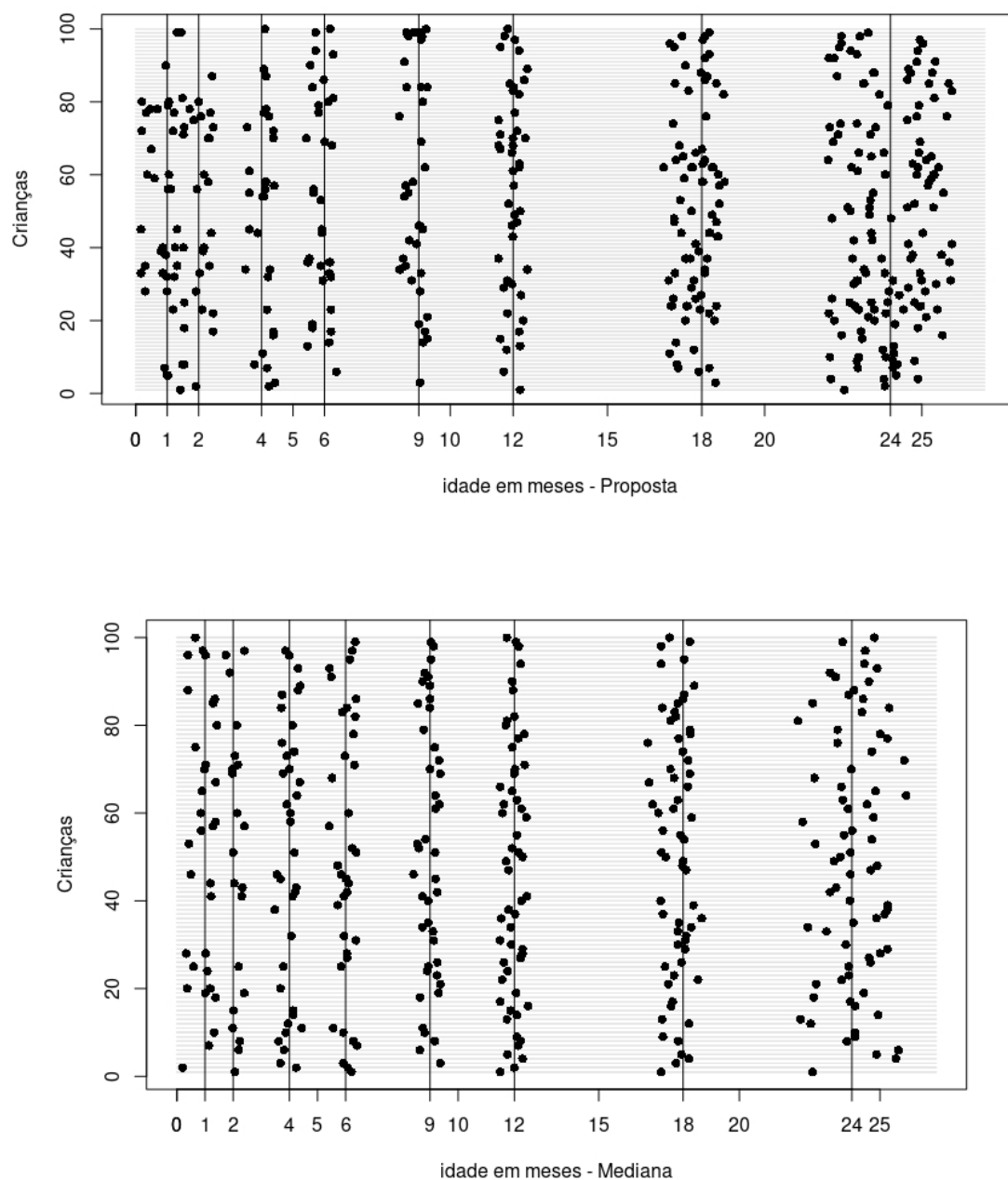


Figura 3 – Tempos de observação para a proposta de análise para um subconjunto aleatório de 100 crianças entre 0 e 2 anos, Brasil, 2008-2017

e altura. Dessa forma, adotou-se trabalhar com a mediana das variáveis dependentes em cada intervalo da "idade-alvo", ou seja, com o banco PropostaMed, pois o tempo de processamento foi otimizado, houve convergência dos modelos e não foi visto uma diferença significativa entre as curvas ajustadas e variâncias residuais (Figuras 4 e 5).

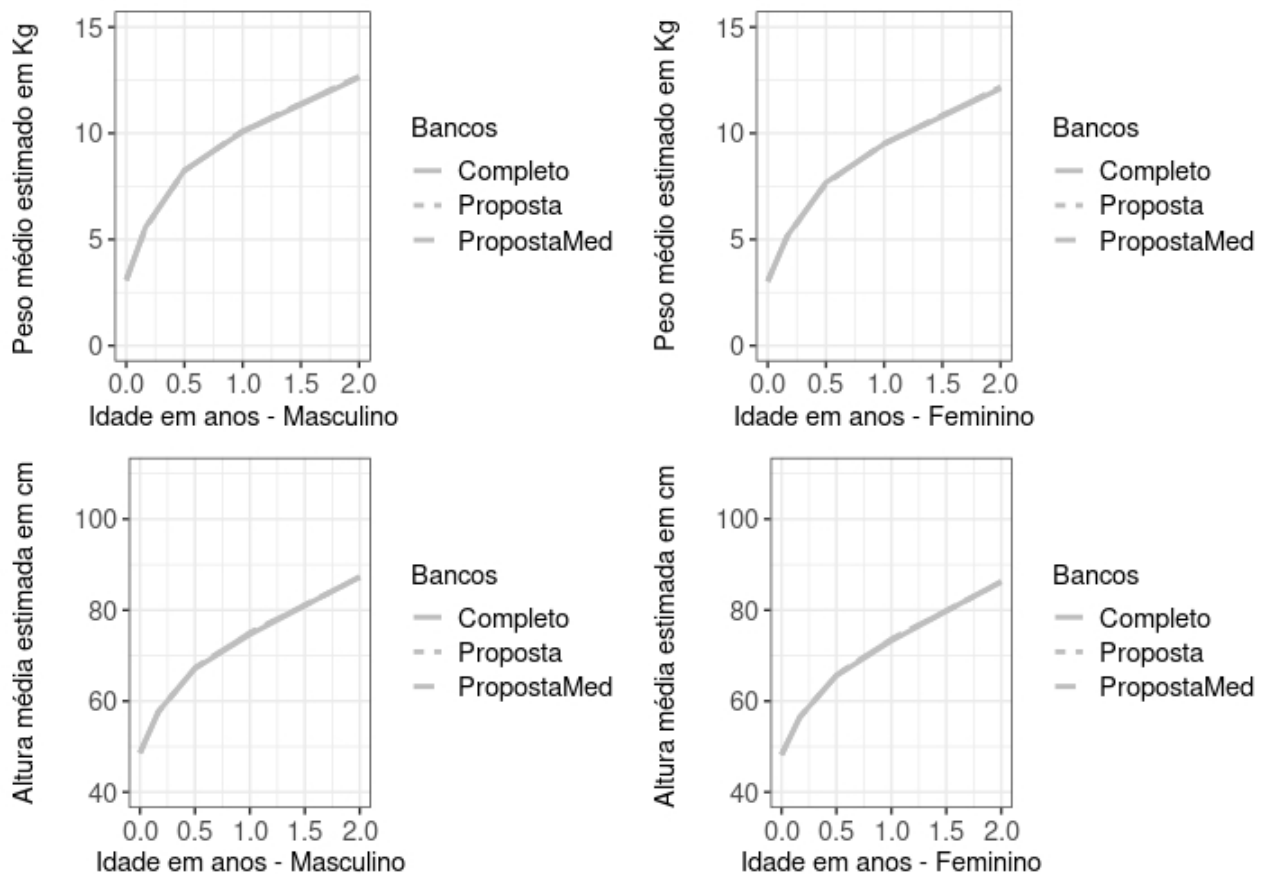


Figura 4 – Curvas estimadas pelo modelo *Brokenstick* por sexo, para dados antropométricos peso e altura de crianças a termo 0 a 2 anos de idade, Brasil, 2008-2017

Tabela 3 – Comparação do modelo *Brokenstick* para variáveis antropométricas peso e altura em relação ao número de observações, tempo de processamento e variância residual para crianças a termo PAN de 0 a 2 anos e por sexo.

Banco de dados	Número		Tempo de Processamento (horas)	Variância Residual	
	Crianças	Observações		Peso	Altura
Sexo: Masculino					
Completo	222393	1102556	00:52'	0.564	1.895
Proposta	79482	332402	00:15'	0.535	1.713
PropostaMed	52475	189714	00:09'	0.542	1.962
Sexo: Feminino					
Completo	218386	1070165	00:47'	0.542	1.901
Proposta	76536	318187	00:15'	0.540	1.764
PropostaMed	50637	182651	00:12'	0.516	1.945

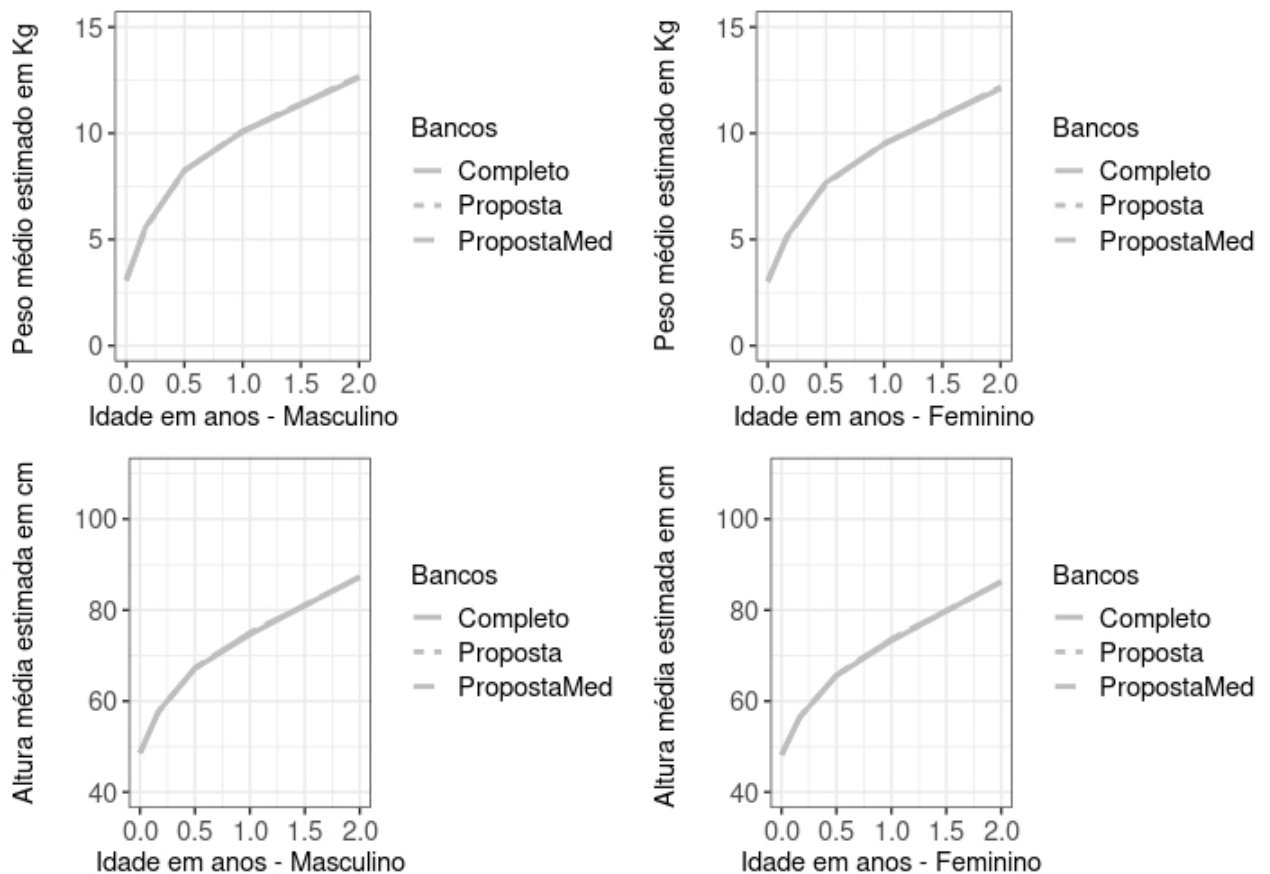


Figura 5 – Curvas estimadas pelo modelo *Brokenstick* por sexo, para dados antropométricos peso e altura de crianças a termo com baixo peso ao nascer de 0 a 2 anos de idade, Brasil, 2008-2017

Tabela 4 – Comparação do modelo *Brokenstick* para variáveis antropométricas peso e altura em relação ao número de observações, tempo de processamento e variância residual para crianças a termo BPN de 0 a 2 anos e por sexo.

Banco de dados	Número		Tempo de Processamento (horas)	Variância Residual	
	Crianças	Observações		Peso	Altura
Sexo: Masculino					
Completo	16920	32086	00:01’	0.553	1.913
Proposta	2303	9782	00:01’	0.529	1.732
PropostaMed	1502	5476	00:18’’	0.499	1.891
Sexo: Feminino					
Completo	9468	48045	00:03’	0.540	1.925
Proposta	3483	14761	00:01’	0.489	1.703
PropostaMed	2286	8353	00:19’’	0.489	1.842

Na Tabela 5, é apresentado o número de observações em relação a idade-alvo, tanto por sexo quanto por classificação de peso ao nascer e nota-se que aumentam as medições para as crianças com idade maior. Entretanto, na Tabela 6 enquanto aumenta o número de medidas

repetidas diminui consideravelmente a quantidade de crianças, lembrando que esperam-se 9 visitas na UBS por criança no período de 0 a 2 anos. O valor mediano do número de observações por criança é de 3, e para os grupos de crianças a termo com peso adequado ao nascer (PAN) e crianças a termo com baixo peso ao nascer (BPN), isso representa mais de 61% das observações (Tabela 6).

Tabela 5 – Número de medições das variáveis antropométricas brutas por intervalo de idade alvo para crianças de 0 a 2 anos, por sexo e classificação de peso ao nascer - Banco PropostaMed

Idade - Alvo (meses)	BPN				PAN			
	Masculino		Feminino		Masculino		Feminino	
	n	%	n	%	n	%	n	%
15 dias	158	2.89	211	2.53	6047	3.19	5613	3.07
1	324	5.92	521	6.24	11409	6.01	10800	5.91
2	399	7.29	617	7.39	13122	6.92	12437	6.81
4	454	8.29	702	8.40	15154	7.87	14374	7.87
6	584	10.66	891	10.67	19826	10.45	19305	10.57
9	609	11.12	849	10.16	29736	15.67	20862	15.80
12	849	15.50	1306	15.64	35029	18.46	34055	18.64
18	970	17.71	1522	18.22	39838	21.00	38617	21.14
24	1129	20.62	1734	20.76	82349	23.86	81318	24.10
	n	MD (IIQ)	n	MD (IIQ)	n	MD (IIQ)	n	MD (IIQ)
	5476	3 (3-5)	8353	3 (3-5)	189714	3 (3-4)	182651	3 (3-4)

IIQ - Intervalo Interquartil, MD - Mediana

Tabela 6 – Número de medições das variáveis antropométricas brutas por criança de 0 a 2 anos, por sexo e classificação de peso ao nascer - Banco PropostaMed

Medidas repetidas	BPN				PAN			
	Masculino		Feminino		Masculino		Feminino	
	n	%	n	%	n	%	n	%
3	923	61.45	1402	61.33	33277	63.38	32181	63.55
4	333	22.17	513	22.44	11029	21.02	10788	21.30
5	144	9.59	193	8.44	4773	9.10	4515	8.92
6	67	4.46	128	5.60	2215	4.22	2048	4.04
7	28	1.86	40	1.75	871	1.67	795	1.57
8	6	0.40	8	0.35	275	0.53	262	0.52
9	1	0.07	2	0.09	35	0.08	48	0.01
Total	1502	100.00	2286	100.00	52475	100.00	50637	100.00

4.2 Ajustes dos modelos univariados para as trajetórias de crescimento

Nesta seção, serão apresentados os ajustes dos modelos linear misto, *Brokenstick* e SITAR para o banco de dados escolhidos na seção anterior (PropostaMed) assim como a análise de resíduos para as variáveis antropométricas brutas por sexo e classificação de peso ao nascer. Estes modelos foram escolhidos por serem amplamente utilizados em outras aplicações na literatura.

4.2.1 Modelo Linear Misto

O padrão de crescimento ponderal e linear, ou seja, peso e altura ao longo dos dois primeiros anos de vida, segundo os modelos de efeitos mistos com efeitos fixos para idade, sexo, baixo peso ao nascer e efeito aleatório para o intercepto e a idade, estão apresentados na Tabela 7. Todas as variáveis foram significativamente associadas com as variáveis desfecho valor de $p < 0,001$. Utilizou-se o pacote **nlme** (PINHEIRO et al., 2017) do programa R (anexo).

Os coeficientes estimados para o efeito linear do peso e da altura variou positivamente com o avanço da idade. No entanto, o coeficiente negativo estimado para o termo quadrático da idade indica que a velocidade de crescimento é menos acentuada com o aumento da idade, ou seja, uma desaceleração na velocidade de crescimento com o aumento da idade (CAMERON, 2012). Crianças do sexo feminino tiveram menor incremento no peso (em média, -0.47kg) ou altura (em média, -1,28cm) quando comparadas àquelas do sexo masculino. Além disso, crianças com baixo peso ao nascer também tiveram em média, um menor incremento no peso (-0.92kg) ou altura (-2,52cm) quando comparadas com peso adequado no nascimento (Tabela 7).

Neste estudo observou-se que o incremento do ritmo do crescimento não eram constantes no tempo. Além disto, constatou-se que a velocidade do crescimento desacelerava à medida que aumentava a idade da criança (DIGGLE et al., 2002), assim, foi indicada a inclusão de um termo quadrático na modelagem estatística (CAMERON, 2012), que expressa a relação curvilínea entre o crescimento e a idade, para corrigir a inclinação não linear da reta, que se afasta negativamente da linha com o aumento do tempo. As trajetórias médias do peso e altura, em relação ao sexo e classificação de peso ao nascer estão na Figura 6.

Conforme apresentado na Subseção 2.2.1, os dois modelos finais ajustados consideram como efeitos aleatórios o intercepto e a idade, o que significa dizer que o desenvolvimento do peso/altura da criança varia não apenas em seu estágio inicial, mas sua taxa de mudança também varia porque existem diferenças intra e interindividuais. A convergência foi alcançada ao utilizar a matriz de covariâncias autoregressiva de ordem 1, na qual a correlação diminui ao longo do tempo para as variáveis peso e altura. Os valores do AIC e do BIC, para cada modelo ajustado, são exibidos na Tabela 7.

As trajetórias média, ajustada e observada, para peso e altura sugerem que o modelo de

Tabela 7 – Modelos de efeitos mistos para variáveis antropológicas peso (kg) e altura (cm) de crianças menores de 2 anos de idade. Brasil, 2008-2017.

	Peso (kg)			Altura (cm)		
	Estimativa	EP	valor de p	Estimativa	EP	valor de p
Intercepto	4,80	0,009	<0,001	54,01	0,029	<0,001
Idade (meses)	0,61	0,001	<0,001	2,32	0,002	<0,001
<i>idade</i> ²	-0,01	0,001	<0,001	-0,04	0,001	<0,001
Sexo						
Feminino	-0,47	0,006	<0,001	-1,28	0,017	<0,001
Baixo Peso						
BPN	-0,92	0,015	<0,001	-2,52	0,046	<0,001
$\hat{\sigma}_{intercepto}$		0,25			2,31	
$\hat{\sigma}_{idade}$		0,05			0,16	
$\hat{\sigma}_{residual}$		0,89			2,52	
$\hat{\rho}_{AR(1)}$		0.382			0.288	
AIC		1105878			2015664	
BIC		1105987			2015762	

efeitos mistos não fez uma boa predição, sugerindo que os estimativas de peso e altura, exceto na faixa etária dos 6 meses de vida são maiores do que as médias empíricas, onde esse resultado é concordante independente do sexo ou classificação do baixo peso ao nascer (Figuras 7 , 8)

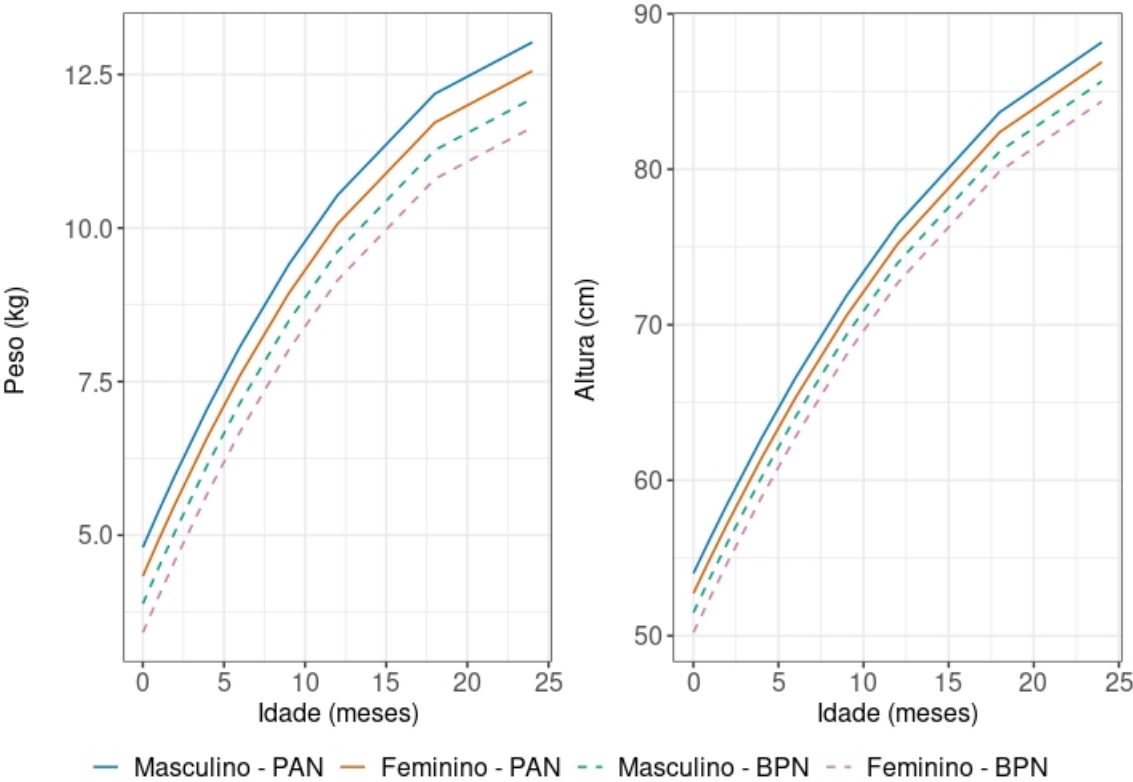


Figura 6 – Trajetória média ajustada para peso e altura, pelo modelo de efeitos mistos para crianças de 0 a 2 anos de idade por sexo e classificação de peso ao nascer.

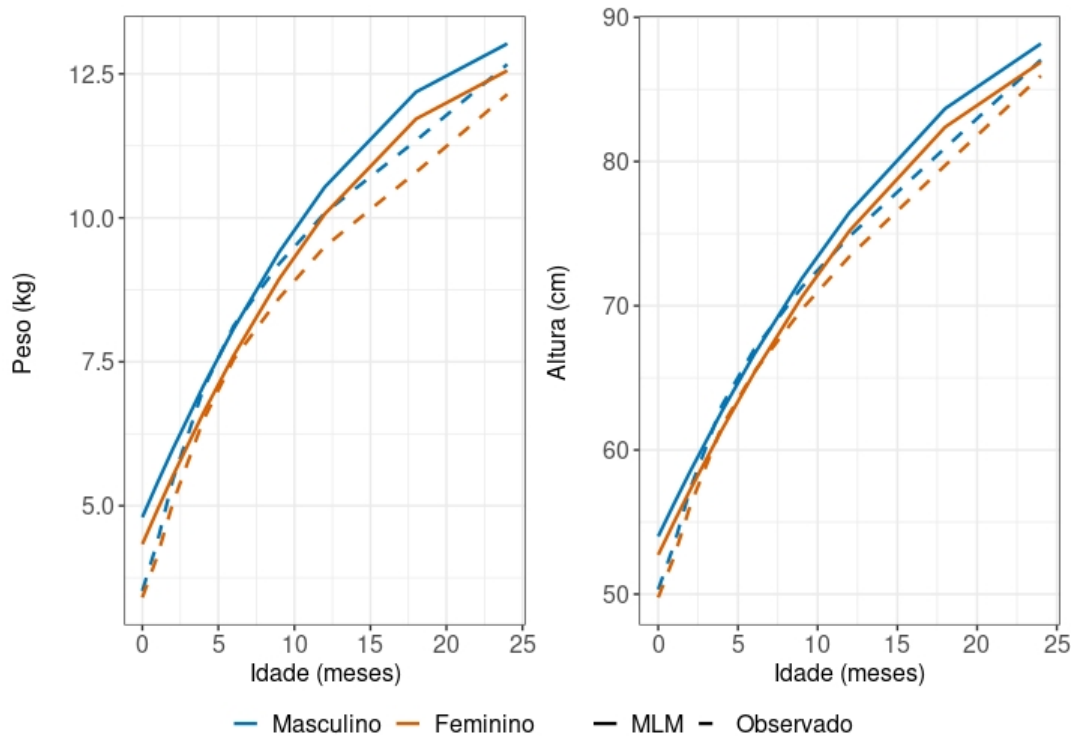


Figura 7 – Trajetória média ajustada e observada para peso e altura, pelo modelo de efeitos mistos para crianças PAN de 0 a 2 anos de idade por sexo.

Nas Figuras 9 e 10, para as variáveis antropométricas brutas, os gráficos de dispersão de resíduos condicionais em relação aos valores ajustados mostra que os modelos não forneceram um bom ajuste aos dados do estudo, indicando que as variâncias dos erros estão variando com a idade (erros heterocedásticos). Estudos futuros podem ser realizados incluindo uma matriz de variância e covariância que atenda à esta suposição. Os gráficos de probabilidade normal dos resíduos com confundimento mínimo mostraram a não rejeição da suposição de normalidade dos erros.

A seguir, o ajuste de outro modelo amplamente utilizado na área de estimação de trajetória de crescimento será apresentado.

4.2.2 Modelo *Brokenstick*

Como visto na metodologia, o modelo Brokenstick inclui termos de um modelo de spline linear cujos os nós referem-se à variável longitudinal (idade). Neste trabalho, a escolha dos nós das variáveis antropométricas brutas seguiu a orientação da literatura (TILLING et al., 2014; HOWE et al., 2016; ANDERSON et al., 2019; BUUREN, 2020). Assim, foram utilizados nós iguais a 0, 2, 6, 12 e 24 meses, onde ocorre mudanças específicas no desenvolvimento da criança (WHO, 2011). Foi retirado a idade de 3 meses pois não é uma visita pré especificada dentro do calendário do Ministério da Saúde (MD, 2002).

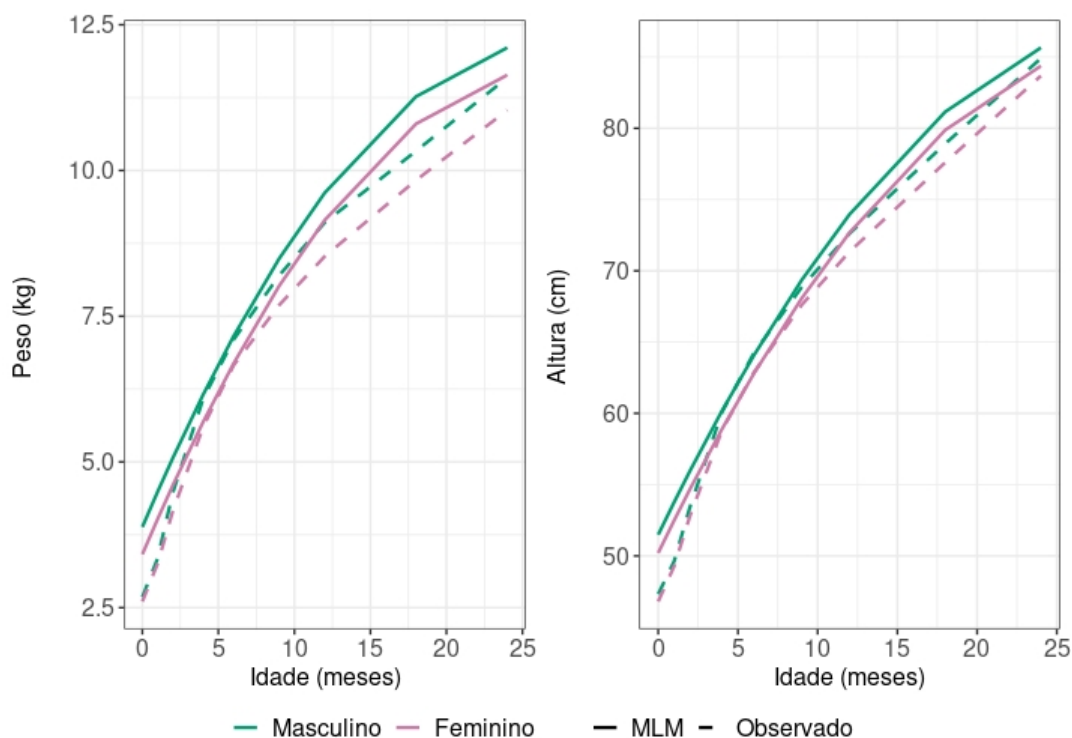


Figura 8 – Trajetória média ajustada e observada para peso e altura, pelo modelo de efeitos mistos para crianças BPN de 0 a 2 anos de idade por sexo.

Para o ajuste desse modelo a unidade de medida para os nós foi em anos, onde ocorreu convergência melhor do modelo, ou seja, os valores dos nós foram 0, 0.167, 0.5, 1 e 2 anos. Nas tabelas e gráficos a seguir manteve-se a unidade em meses para facilitar a sua interpretação em função da idade-alvo.

Diferentemente do modelo de efeitos mistos, o modelo *Brokenstick* (BUUREN, 2020) não admite covariáveis. Dessa forma, ajustaram-se os modelos de trajetória de crescimento para cada sexo e classificação de peso ao nascer (baixo peso e peso adequado), apresentadas nas Tabelas 8 e 9.

A média ajustada dos pesos e alturas em cada período de idade-alvo são mostradas nas Tabelas 8 e 9, neste caso manteve-se a idade em meses para uma melhor interpretação dos resultados. Nota-se, que crianças do sexo masculino apresentam coeficientes estimados de peso e altura um pouco maiores se comparados ao do sexo feminino, exceto no grupo de baixo peso ao nascer no primeiro nó, onde as meninas tiveram um coeficiente levemente superior.

As trajetórias médias ajustadas de peso e altura, para crianças, com peso adequado ao nascer ou não, teve uma curva bem similar em ambos os sexos (Figuras 11, 12), sugerindo que este modelo faz uma boa predição para os respectivos dados antropométricos em crianças menores de dois anos deste estudo.

Os resíduos condicionais padronizados versus valores ajustados sugerem que a suposição

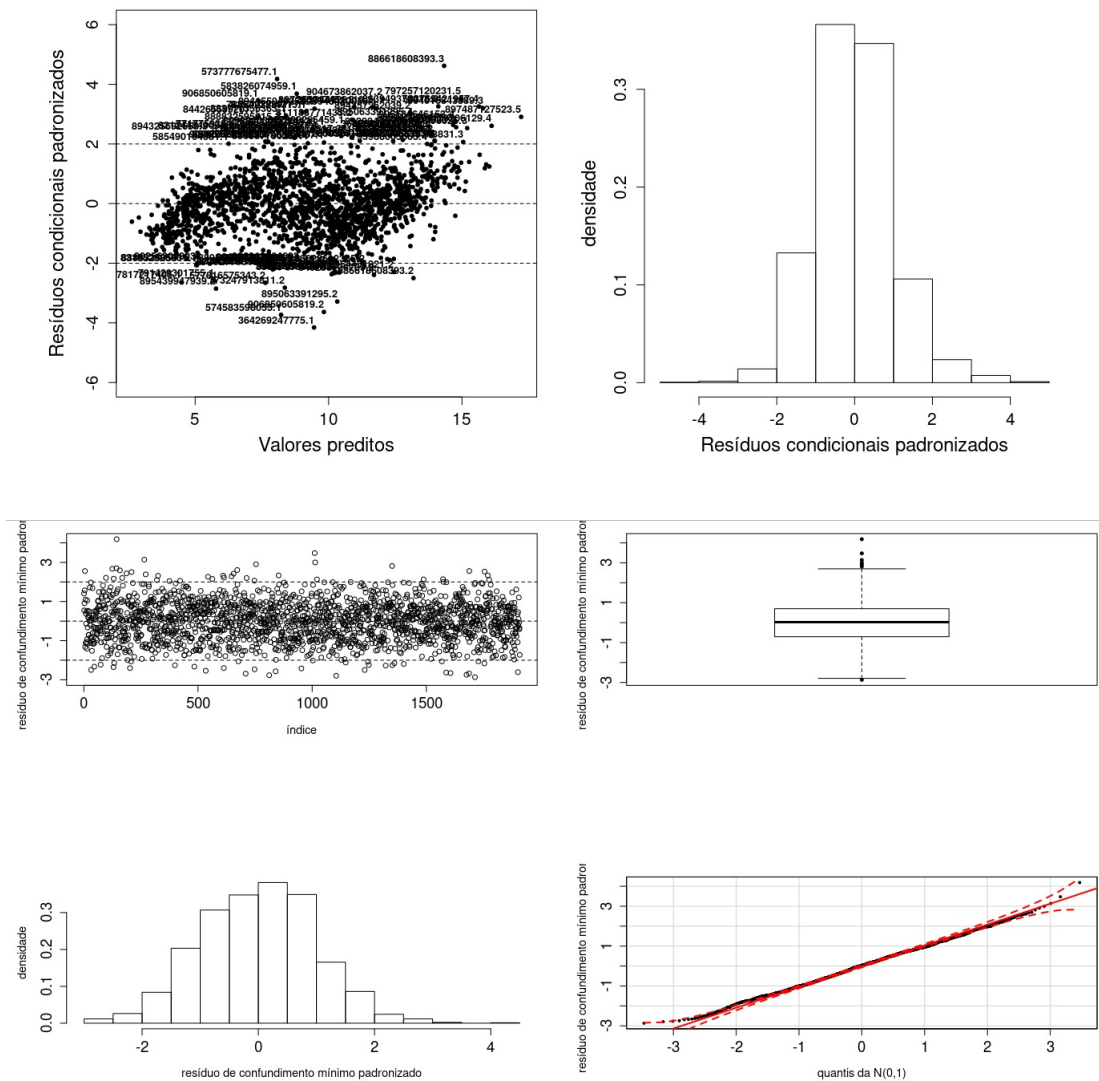


Figura 9 – Gráficos de dispersão e histograma dos resíduos condicionais padronizados e quantis em relação aos pesos estimados pelo Modelo de efeitos mistos

de homocedasticidade do erro condicional pode ser validada, entretanto na variável altura é observado muitos valores discrepantes e para variável peso, no geral há poucos registros de valores altos (Figuras 13 e 14). Segundo Nobre e Singer (2007), em modelos lineares mistos, o gráfico de normalidade de resíduos condicionais não deve ser utilizado para avaliar a suposição de normalidade dos erros condicionais, visto que não pode ser considerado como um erro puro, sendo sugerido utilizar o gráfico dos resíduos com confundimento mínimo. No modelo *Brokenstick*, caso especial do modelo linear misto, por questões de limitação do pacote, não foi possível gerar esse gráfico. Pelos histogramas, nota-se uma distribuição aproximadamente normal nos resíduos condicionais, e na variável altura a variabilidade nos resíduos é maior (Figuras 15 e 16).

A qualidade do ajuste foi avaliada com a representação visual do gráfico de dispersão dos valores observados versus ajustados (Figuras 17, 18, 19, 20). Os gráficos dos dados antro-

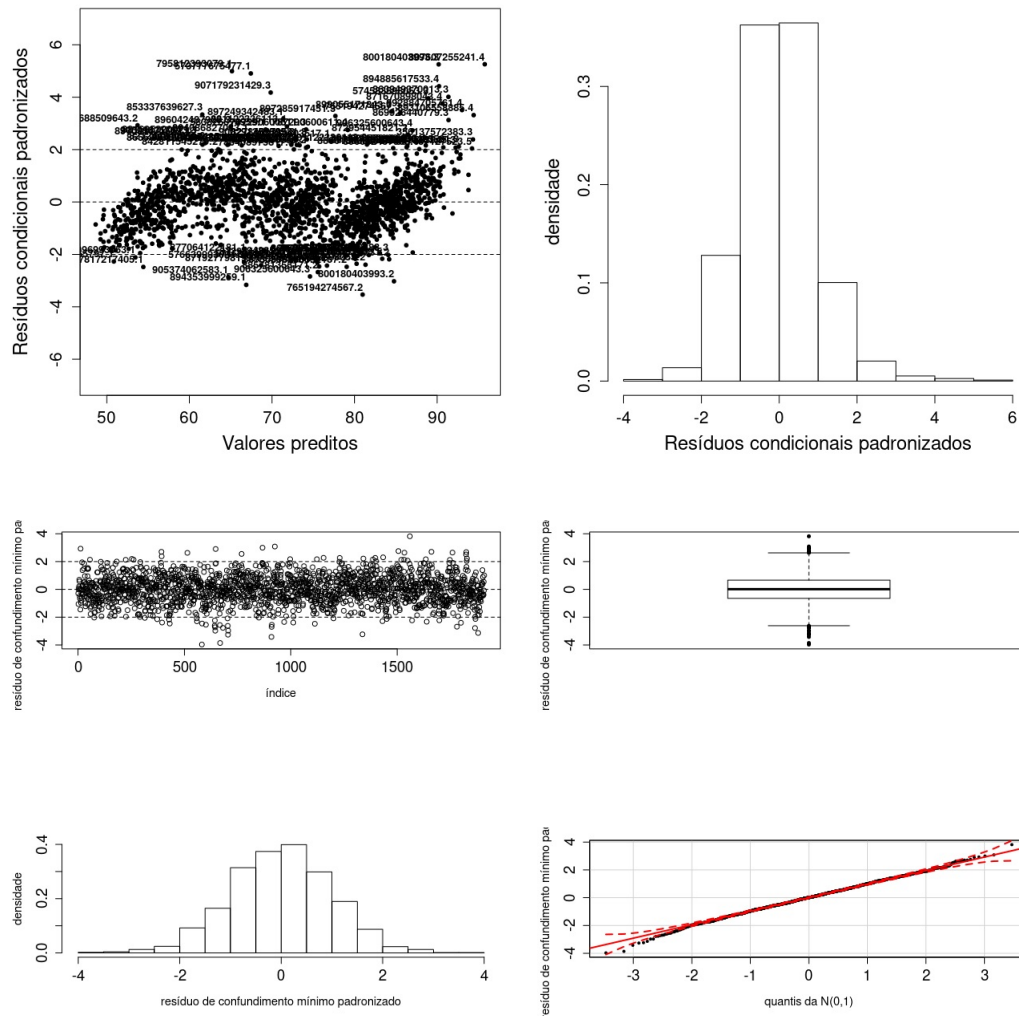


Figura 10 – Gráficos de dispersão e histograma dos resíduos condicionais padronizados e quantis em relação as alturas estimadas pelo Modelo de efeitos misto

métricos de peso e altura sugerem um bom ajuste. O percentual de explicação da variação tanto do peso médio quanto da altura média foi de 98% para ambos os sexos e grupo de classificação de peso ao nascer.

4.2.3 Modelo SITAR

Primeiramente, os modelos SITAR foram testados com todos os três parâmetros (tamanho, tempo e velocidade) como efeitos fixos e aleatórios para as variáveis dependentes peso e altura. Nestas tentativas, houve problemas de convergência e desse modo, foram adotadas estratégias como a transformação para o logaritmo da idade, da variável resposta, ou ainda a eliminação de um dos efeitos aleatórios (ANDERSON et al., 2019; RIDDELL et al., 2017a; COLE; DONALDSON; BEN-SHLOMO, 2010; COLE, 2018; OHUMA et al., 2021). Assim, dos modelos que alcançaram convergência, aqueles que contêm dois efeitos aleatórios (tamanho e velocidade) se ajustaram

Tabela 8 – Estimativas e desvio padrão (DP) dos coeficientes do modelo *Brokenstick* para o peso das crianças menores de 2 anos de idade (nós nas idades alvo 0, 2, 6, 12 e 24 meses, por sexo e classificação de peso ao nascer)

Nós (mês)	PAN		BPN	
	Estimativa	DP	Estimativa	DP
Sexo Masculino				
0	3.09	0.20	2.16	0.17
2	5.56	0.63	4.57	0.61
6	8.24	0.93	7.28	0.87
12	10.11	1.20	9.14	1.13
24	12.68	1.62	11.62	1.64
Sexo Feminino				
0	3.03	0.17	2.22	0.01
2	5.14	0.53	4.22	0.48
6	7.65	0.88	6.76	6.76
12	9.52	1.16	8.58	8.58
24	12.17	1.66	11.07	11.07

Tabela 9 – Estimativas e desvio padrão (DP) dos coeficientes do modelo *Brokenstick* para a altura das crianças menores de 2 anos de idade (nós nas idades alvo 0, 2, 6, 12 e 24 meses, por sexo e classificação de peso ao nascer)

Nós (mês)	PAN		BPN	
	Estimativa	DP	Estimativa	DP
Sexo Masculino				
0	48.64	1.35	45.35	1.21
2	57.61	2.46	53.87	2.64
6	67.29	2.51	64.84	2.85
12	75.03	3.00	72.82	2.91
24	87.23	3.95	85.16	4.53
Sexo Feminino				
0	48.23	1.23	45.21	0.96
2	56.44	2.27	56.09	2.23
6	65.72	2.51	63.36	2.59
12	73.63	3.03	71.50	3.08
24	86.16	2.74	83.83	4.29

melhor aos dados. A justificativa baseia-se no tamanho amostral e na escala original das variáveis.

Dessa forma, o modelo SITAR simplificado é dado por

$$y_{it} = \alpha_i + h\left(\frac{t}{\exp(-\gamma_i)}\right) + \varepsilon_{it} \quad (4.1)$$

em que y_{it} representa o peso ou altura da criança i com idade t em meses; $h(\cdot)$ é uma função B-spline cúbica natural na idade transformada que define a curva spline média; e α_i e γ_i são efeitos aleatórios específicos para tamanho e intensidade/velocidade, respectivamente, e ε_{it} são

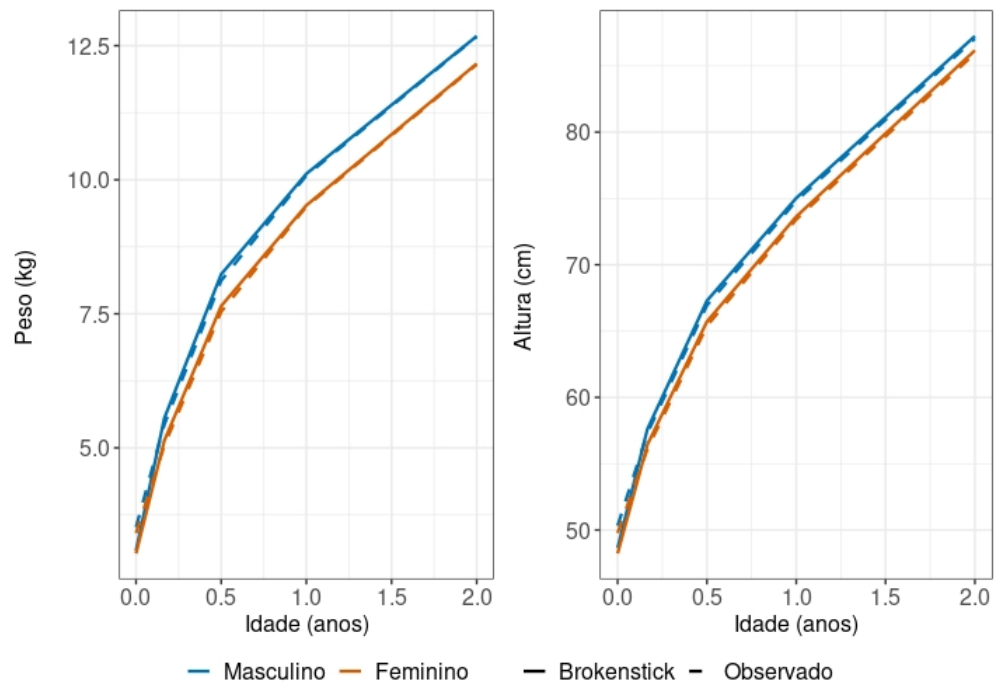


Figura 11 – Trajetória média ajustada para peso e altura, pelo modelo *Brokenstick* para crianças PAN de 0 a 2 anos de idade por sexo.

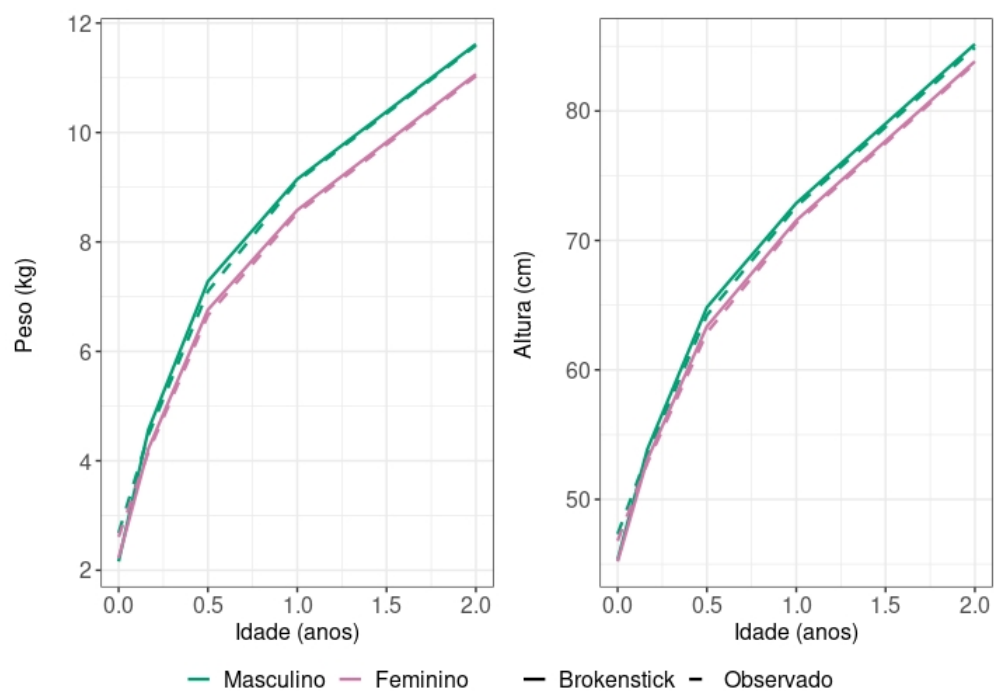


Figura 12 – Trajetória média ajustada para peso e altura, pelo modelo *Brokenstick* para crianças BPN de 0 a 2 anos de idade por sexo

resíduos normalmente distribuídos.

Realizou-se a escolha dos graus de liberdade para a curva B-spline cúbica pelo critério

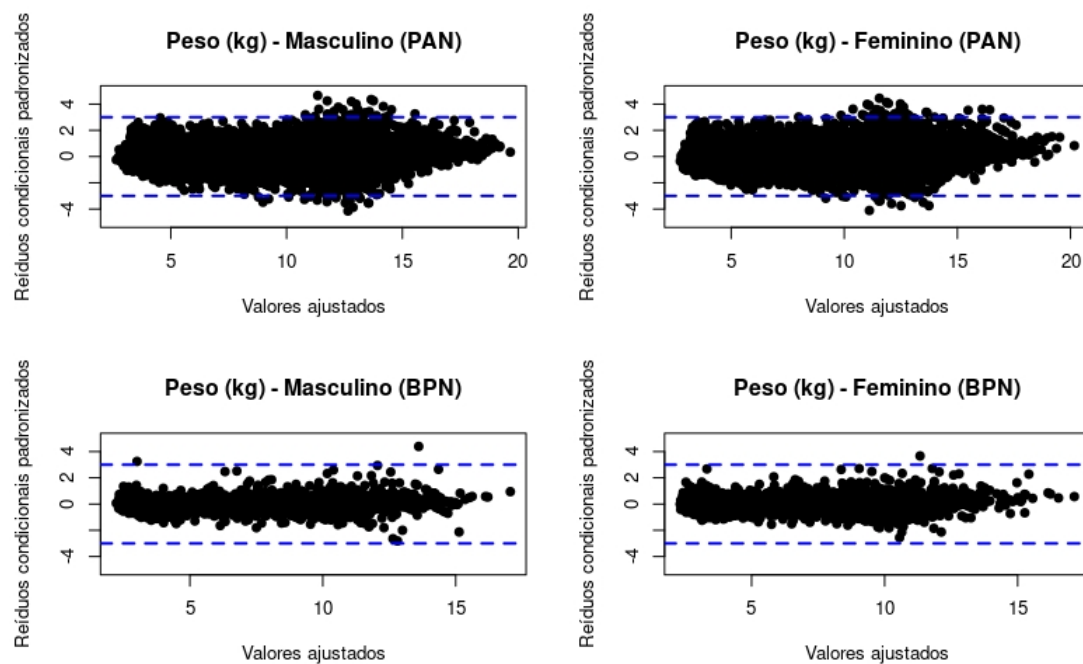


Figura 13 – Gráficos de dispersão dos resíduos condicionais padronizados pelo ajuste do modelo *Brokenstick* para o peso de crianças de 0 a 2 anos por sexo e classificação de peso ao nascer idade

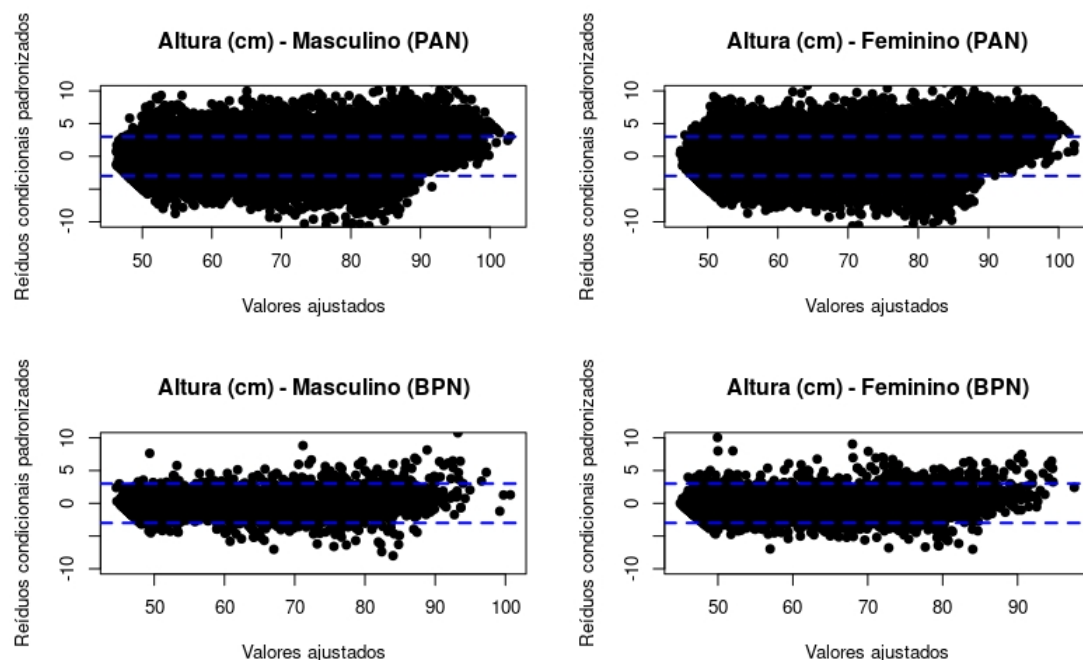


Figura 14 – Gráficos de dispersão dos resíduos condicionais padronizados pelo ajuste do modelo *Brokenstick* para a altura de crianças de 0 a 2 anos por sexo e classificação de peso ao nascer

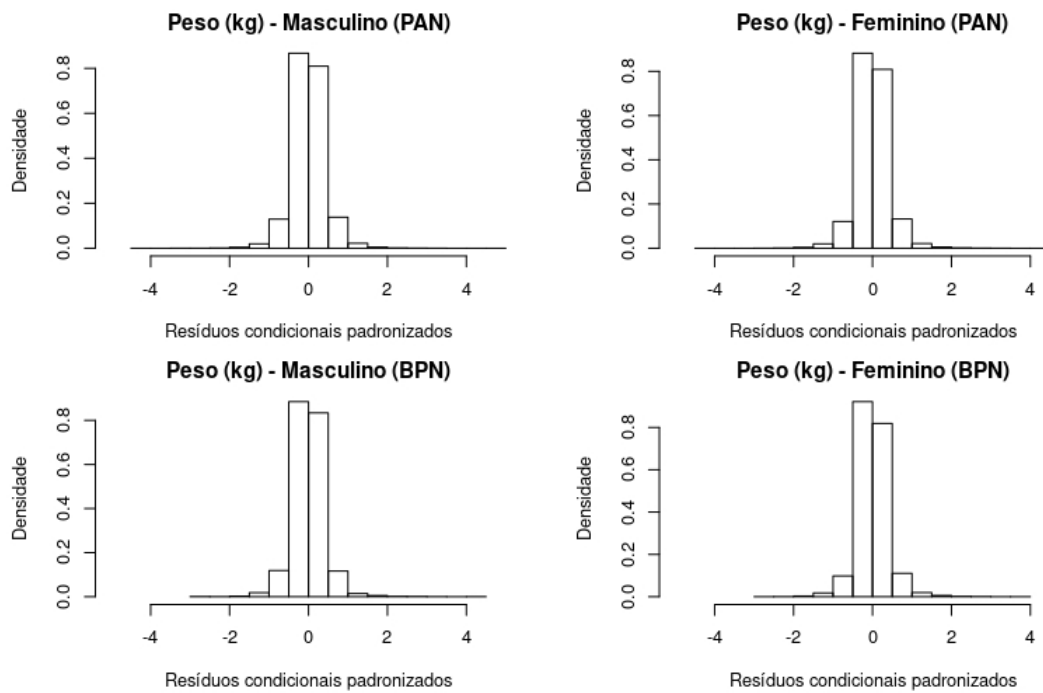


Figura 15 – Histograma dos resíduos condicionais padronizados pelo ajuste do modelo *Brokenstick* para o peso de crianças de 0 a 2 anos idade por sexo e classificação de peso ao nascer

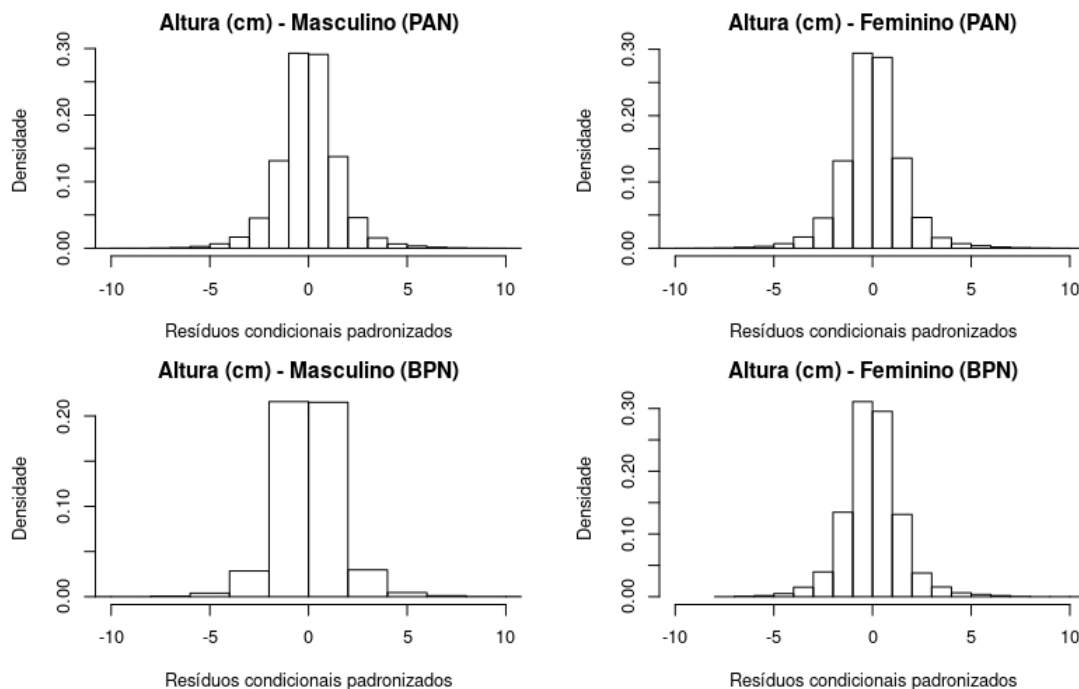


Figura 16 – Histograma dos resíduos condicionais padronizados pelo ajuste do modelo *Brokenstick* para a altura de crianças de 0 a 2 anos idade por sexo e classificação de peso ao nascer

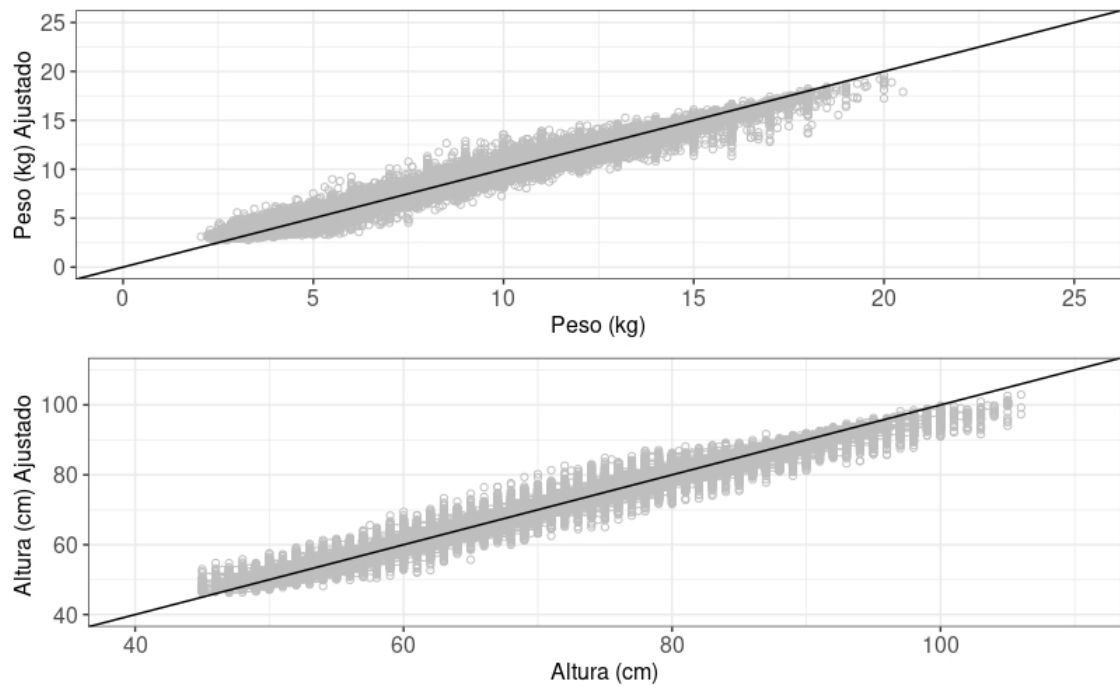


Figura 17 – Valor ajustado pelo modelo *Brokenstick* versus observado dos dados antropométricos brutos de crianças PAN do sexo masculino de 0 a 2 anos idade

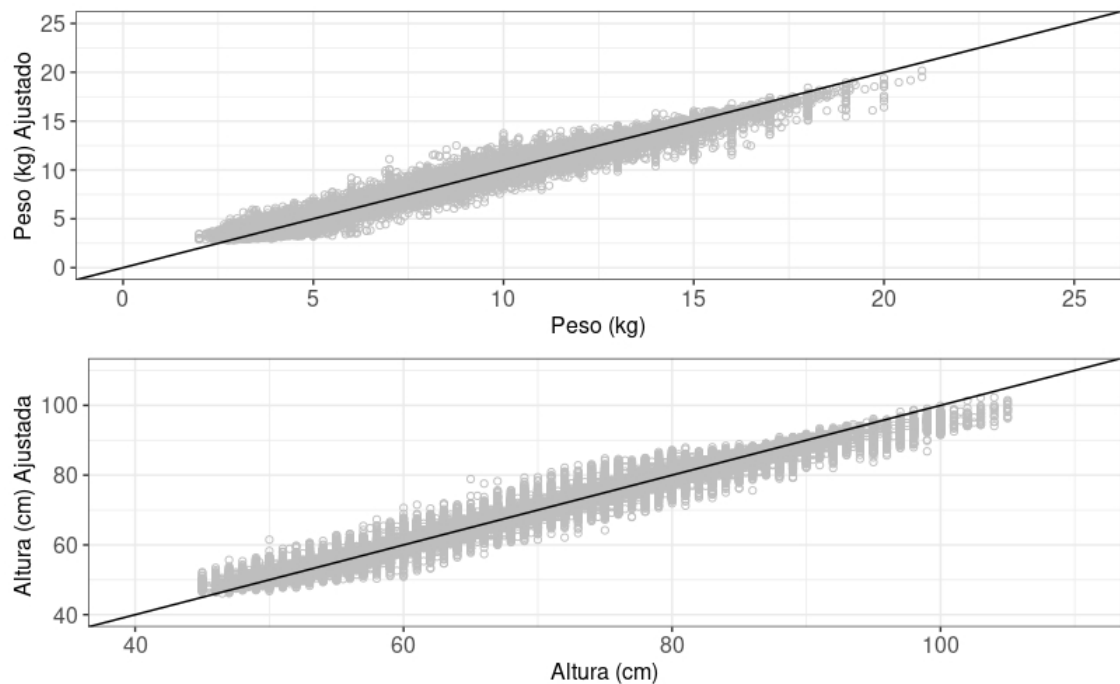


Figura 18 – Valor ajustado pelo modelo *Brokenstick* versus observado dos dados antropométricos brutos de crianças PAN do sexo feminino de 0 a 2 anos de idade

que resulta no menor BIC (COLE; DONALDSON; BEN-SHLOMO, 2010)(Tabelas 10, 11). Quando o efeito aleatório do parâmetro γ_i foi adicionado no modelo SITAR, houve um aumento

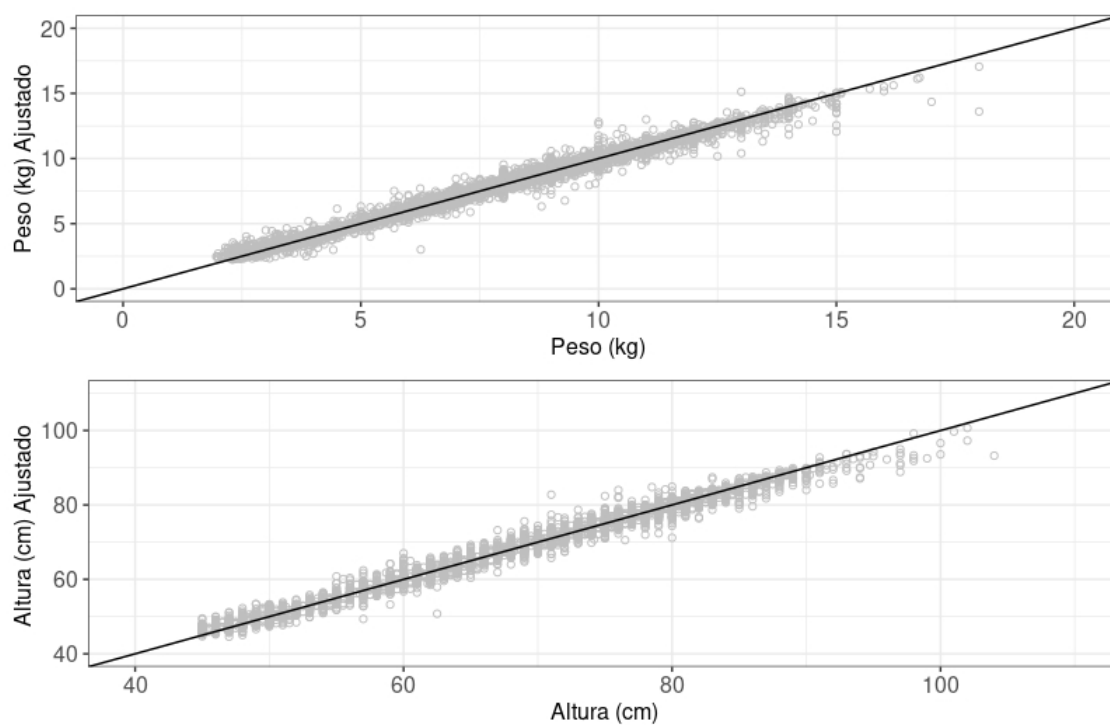


Figura 19 – Valor ajustado pelo modelo *Brokenstick* versus observado dos dados antropométricos brutos de crianças BPN do sexo masculino de 0 a 2 anos idade.

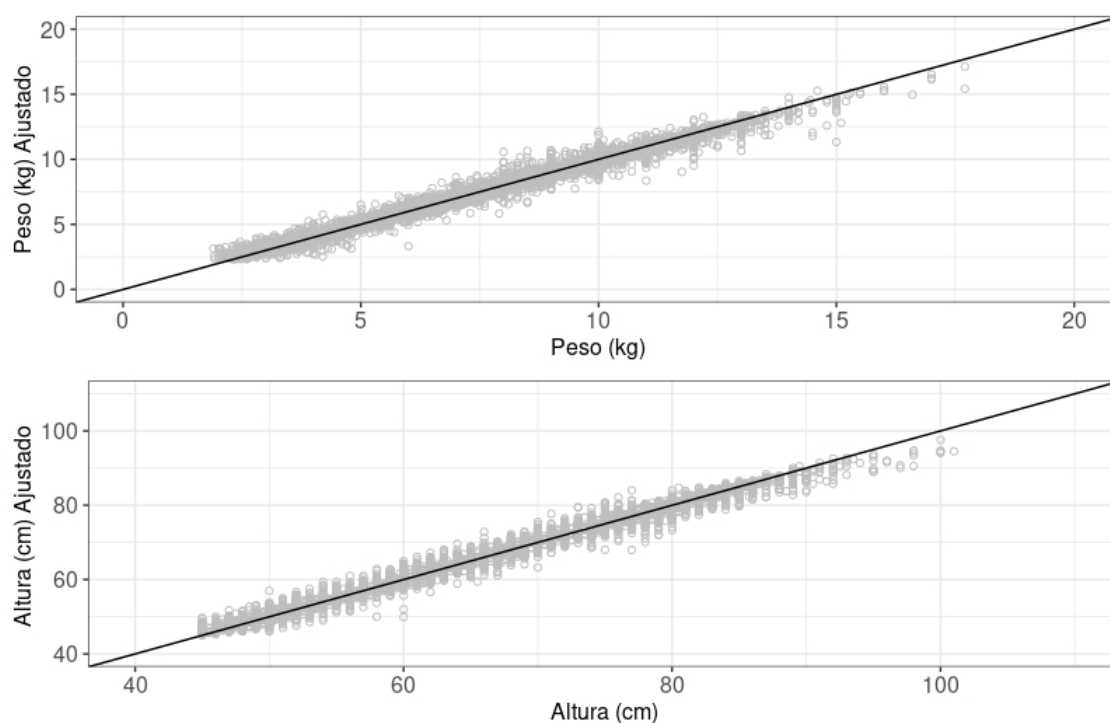


Figura 20 – Valor ajustado pelo modelo *Brokenstick* versus observado dos dados antropométricos brutos de crianças BPN do sexo feminino de 0 a 2 anos idade.

no percentual de explicação da variação tanto no peso quanto na altura. Assim, os modelos finais selecionados para crianças PAN e BPN, explicaram entre 63% e 74% da variação entre peso médio com graus de liberdade variando entre 6 e 8, e aproximadamente entre 60% e 66% da variação entre altura média para ambos os sexos, com graus de liberdade variando entre 7 e 10. Nas Tabelas de 12 a 15 apresentam-se os ajustes dos modelos selecionados nas Tabelas 10 e 11

O parâmetro α_i deve ser interpretado com a unidade de média da variável resposta, enquanto o parâmetro γ_i em unidades fracionárias. Assim, quando a velocidade é multiplicada por 100, ela pode ser interpretada como uma diferença percentual da criança na curva média. Para a variável peso das crianças, os meninos cujas curvas de crescimento diferem da curva de crescimento média em 2,95kg em média e uma desaceleração na velocidade de crescimento de 10% (Tabela 12). Da mesma forma para a altura sugere que os meninos cujas curvas de crescimento diferem da curva de crescimento média em 49,94cm em média e uma desaceleração na velocidade de crescimento de 33% (Tabela 14), enquanto que para meninos BPN foi de 18% essa desaceleração na velocidade de crescimento (Tabela 15).

Similarmente, para as meninas cuja curvas de crescimento diferem em média 21,9 kg e com velocidade do crescimento mais lenta em 23% e, em relação a altura (cm) a velocidade do crescimento foi mais lenta em 28% (Tabelas 12,14). Para as meninas BPN, o crescimento foi mais rápido em 24% e 10% para peso e altura, respectivamente.

Analisando as trajetórias de crescimento ponderal por sexo e classificação de peso ao nascer, os meninos PAN em todas as idades apresentaram em média, um maior ganho de peso, e os meninos BPN na faixa etária de 18 a 24 meses tiveram uma destacada recuperação no ganho de peso médio, podendo ser observado no gráfico de velocidade (à direita) (Figura 21). Para trajetória da altura, os grupos tinham formato semelhante, embora os meninos PAN fossem, mais altos em todo período estudado. Nota-se que depois dos 12 meses o grupo de baixo peso ao nascer teve variação no crescimento linear podendo ser observada essa variação no gráfico de velocidade (à direita) nesse intervalo (Figura 22).

As trajetórias médias, ajustadas e observadas, para peso e altura sugerem que o modelo SITAR fez uma boa predição para os respectivos dados antropométricos em crianças menores de anos deste estudo (Figuras 11, 12). Entretanto, em crianças BPN de ambos os sexos, depois dos 12 meses de vida, na curva média observada para variável altura não se observa uma variação no crescimento linear como apresentado na curva ajustada, o mesmo é observado para os meninos na curva média observada da variável peso, na faixa entre 18 e 24 meses.

O diagnóstico do modelo SITAR foi realizado através da análise de gráficos de resíduos. O gráfico dos resíduos padronizados versus valores ajustados foi usado para avaliar a suposição da variância constante dos resíduos e o gráfico de probabilidade normal foi sugerido para avaliar a normalidade dos erros residuais obtidos de cada modelo ajustado.

Para ambos os sexos e grupos de classificação de peso ao nascer, os gráficos de dispersão

Tabela 10 – Comparação dos ajustes dos modelos SITAR aplicados aos dados antropométricos em crianças PAN de 0 a 2 anos, com diferentes parametrizações e transformações na variável idade em meses.

Especificação do Modelo	Parâmetros (efeitos fixos e aleatório)	BIC	% Variância Explicada	Graus de Liberdade
Masculino - Peso				
Idade	α_i	554670,90	59,84	4
	$\alpha_i + \gamma_i$	Não convergiu	N/A	
	$\alpha_i + \beta_i + \gamma_i$	Não convergiu	N/A	
Logaritmo da Idade	α_i	553004,3	60,16	7
	$\alpha_i + \gamma_i$	505578,7	73,91	
	$\alpha_i + \beta_i + \gamma_i$	Não convergiu	N/A	
Masculino - Altura				
Idade	α_i	Não convergiu	N/A	
	$\alpha_i + \gamma_i$	Não convergiu	N/A	
	$\alpha_i + \beta_i + \gamma_i$	Não convergiu	N/A	
Logaritmo da Idade	α_i	972641,90	48,66	3
	$\alpha_i + \gamma_i$	952140,1	63,14	
	$\alpha_i + \beta_i + \gamma_i$	Não convergiu	N/A	
Feminino -Peso				
Idade	α_i	Não convergiu	N/A	
	$\alpha_i + \gamma_i$	Não convergiu	N/A	
	$\alpha_i + \beta_i + \gamma_i$	Não convergiu	N/A	
Logaritmo da Idade	α_i	532691,7	59,11	4
	$\alpha_i + \gamma_i$	484154,9	73,39	
	$\alpha_i + \beta_i + \gamma_i$	Não convergiu	N/A	
Feminino - Altura				
Idade	α_i	Não convergiu	N/A	
	$\alpha_i + \gamma_i$	Não convergiu	N/A	
	$\alpha_i + \beta_i + \gamma_i$	Não convergiu	N/A	
Logaritmo da Idade	α_i	Não convergiu	N/A	7*
	$\alpha_i + \gamma_i$	920861,9	61,92	
	$\alpha_i + \beta_i + \gamma_i$	Não convergiu	N/A	

* Modelo final selecionado.

N / A, não aplicável.

de resíduos padronizados em relação aos valores ajustados (Figuras 27) mostraram que não têm variância constante, e apresentam muita variabilidade com picos no meio da distribuição. Os gráficos de probabilidade normal dos resíduos padronizados concentraram-se na parte central próxima a linha de referência, apresentando caudas pesadas nas extremidades e com poucas observações remotas (Figura 30) mostrando falta de ajuste aos dados em estudo. Pelos histogramas notam-se muita variabilidade dos resíduos na variável altura (Figuras 31 e 32).

Um dos objetivos deste trabalho foi estimar e selecionar o melhor modelo marginal para descrever a trajetória de crescimento infantil em crianças menores de 2 anos de vida na base de dados em estudo e, a partir desta escolha, estimar uma cópula para o conjunto de dados com

Tabela 11 – Comparação dos ajustes dos modelos SITAR aplicados aos dados antropométricos em crianças BPN de 0 a 2 anos, com diferentes parametrizações e transformações na variável idade em meses.

Especificação do Modelo	Parâmetros (efeitos fixos e aleatório)	BIC	% Variância Explicada	Graus de Liberdade
Masculino - Peso				
Idade	α_i	Não convergiu	N/A	
	$\alpha_i + \gamma_i$	Não convergiu	N/A	
	$\alpha_i + \beta_i + \gamma_i$	Não convergiu	N/A	
Logaritmo da Idade	α_i	15799.65	59.58	5
	$\alpha_i + \gamma_i$	15063.24	72.05	7*
	$\alpha_i + \beta_i + \gamma_i$	Não convergiu	N/A	
Masculino - Altura				
Idade	α_i	Não convergiu	N/A	
	$\alpha_i + \gamma_i$	Não convergiu	N/A	
	$\alpha_i + \beta_i + \gamma_i$	Não convergiu	N/A	
Logaritmo da Idade	α_i	28366.43	50.76	5
	$\alpha_i + \gamma_i$	27295.33	66.14	10*
	$\alpha_i + \beta_i + \gamma_i$	Não convergiu	N/A	
Feminino - Peso				
Idade	α_i	Não convergiu	N/A	
	$\alpha_i + \gamma_i$	Não convergiu	N/A	
	$\alpha_i + \beta_i + \gamma_i$	Não convergiu	N/A	
Logaritmo da Idade	α_i	28357.95	50.89	4
	$\alpha_i + \gamma_i$	27755.66	63.35	7*
	$\alpha_i + \beta_i + \gamma_i$	Não convergiu	N/A	
Feminino - Altura				
Idade	α_i	Não convergiu	N/A	
	$\alpha_i + \gamma_i$	Não convergiu	N/A	
	$\alpha_i + \beta_i + \gamma_i$	Não convergiu	N/A	
Logaritmo da Idade	α_i	42914.53	46.4	3
	$\alpha_i + \gamma_i$	41581.56	61.15	8*
	$\alpha_i + \beta_i + \gamma_i$	Não convergiu	N/A	

* Modelo final selecionado.

N / A, não aplicável.

objetivo de descrever a relação de dependência conjunta entre as variáveis respostas peso e altura. Como resultado, recomendamos o uso do modelo *Brokenstick* para variáveis antropométricas brutas que mostrou ter boa precisão do ajuste na trajetória de crescimento infantil, e além disso é fácil de ajustar e fornece estimativas interpretáveis dessas trajetórias. Decerto haverá outras abordagens de modelagem em potencial na literatura, então esta recomendação é baseada apenas nos métodos que estudamos.

Tabela 12 – Modelo SITAR de melhor ajuste para o sexo masculino aplicados aos dados antropométricos de peso (kg) em crianças de 0 a 2 anos por classificação de peso ao nascer. Brasil 2008-2017.

Efeitos	Estimativas (EP)	
	PAN	BPN
Fixos		
α	2,95(0,02)	2,12(0,22)
γ	-0,10(0,01)	-0,18(0,03)
Aleatórios		
$\hat{\sigma}_{Tamanho}$	0,85	0,89
$\hat{\sigma}_{Velocidade}$	0,12	0,05
$\hat{\sigma}_{Residual}$	0,65	0,66
Graus de liberdade	8	7
N de crianças	52475	1502
N de observações	189714	5476

Todos os modelos foram ajustados para idade em meses (log transformado).
EP - erro padrão.

Tabela 13 – Modelo SITAR de melhor ajuste para o sexo feminino aplicados aos dados antropométricos de peso (kg) em crianças de 0 a 2 anos por classificação de peso ao nascer. Brasil 2008-2017.

Efeitos	Estimativas (EP)	
	PAN	BPN
Fixos		
α	2,91(0,03)	2,38(0,10)
γ	-0,23(0,01)	0,24(0,05)
Aleatórios		
$\hat{\sigma}_{Tamanho}$	0,82	0,75
$\hat{\sigma}_{Velocidade}$	0,11	0,17
$\hat{\sigma}_{Residual}$	0,65	0,60
Graus de liberdade	6	7
N de crianças	50637	2286
N de observações	182651	8353

Todos os modelos foram ajustados para idade em meses (log transformado).
EP - erro padrão.

4.3 Ajuste do Modelo Bivariado - Cópula

Nesta seção, são apresentados os principais resultados dos modelos de crescimento infantil bivariados derivados das cópulas, que melhor se ajustou aos dados em estudo, considerando os modelos *Brokenstick*.

Por limitação computacional e tempo de processamento, optou-se neste trabalho na análise bivariada de cópula por retirar uma amostra aleatória para posteriormente poder adaptar essa abordagem em situações envolvendo dados massivos. Foi retirada uma amostra de 10%, a partir do ID das crianças usando a função **sample(.)** do *software* R com semente, apenas para crianças do subgrupo de peso adequado ao nascer. Para o subgrupo de crianças com baixo peso ao nascer, por representar apenas 3,56% da base em estudo, foi possível fazer uma análise completa desse grupo. A seguir será apresentada uma análise descritiva da amostra das crianças com peso adequado ao nascer com o objetivo de verificar a preservação das características em

Tabela 14 – Modelo SITAR de melhor ajuste para o sexo masculino aplicados aos dados antropométricos de altura (cm) em crianças de 0 a 2 anos por classificação de peso ao nascer. Brasil 2008-2017.

Efeitos	Estimativas (EP)	
	PAN	BPN
Fixos		
α	49,94(0,87)	46,75(0,51)
γ	-0,33(0,05)	-0,05(0,01)
Aleatórios		
$\hat{\sigma}_{Tamanho}$	2,32	2,43
$\hat{\sigma}_{Velocidade}$	0,11	0,07
$\hat{\sigma}_{Residual}$	2,16	2,23
Graus de liberdade	7	10
N de crianças	52475	1502
N de observações	189714	5476

Todos os modelos foram ajustados para idade em meses (log transformado).
EP - erro padrão.

Tabela 15 – Modelo SITAR de melhor ajuste para o sexo feminino aplicados aos dados antropométricos de altura (cm) em crianças de 0 a 2 anos por classificação de peso ao nascer. Brasil 2008-2017.

Efeitos	Estimativas (EP)	
	PAN	BPN
Fixos		
α	49,09(0,73)	46,65(0,26)
γ	-0,28(0,05)	0,10(0,02)
Aleatórios		
$\hat{\sigma}_{Tamanho}$	2,24	2,21
$\hat{\sigma}_{Velocidade}$	0,11	0,06
$\hat{\sigma}_{Residual}$	2,19	2,22
Graus de liberdade	7	8
N de crianças	50637	2286
N de observações	182651	8353

Todos os modelos foram ajustados para idade em meses (log transformado).
EP - erro padrão.

relação a base PropostaMed (Tabela 16 e Figuras 33 e 34). O número de crianças selecionadas nessa amostra é de 10312 (masculino=5248 e feminino=5064) e contempla 37312 observações (masculino=18957 e feminino=18355).

Os gráficos histogramas e boxplots, comparativo das medidas antropométricas, da população em estudo e da amostra selecionada, são apresentados nas Figuras 33 e 34, respectivamente. Observa-se que o peso e altura apresentam similaridade em termos de variabilidade, por exemplo há valores discrepantes (outliers) de peso.

O pacote *Brokenstick*, utilizado na Subseção(4.2.2) para estimar as trajetórias de crescimento antropométricas de peso e altura marginais, não foi viável para o ajuste dos respectivos modelos bivariados derivados de cópulas, visto que este pacote não exibe informações importantes como a matriz de variância-covariância marginal do modelo ajustado, informações para o cálculo da distribuições conjuntas. Diante disso, foi utilizado o pacote **nlme**, na função **lme(.)** do

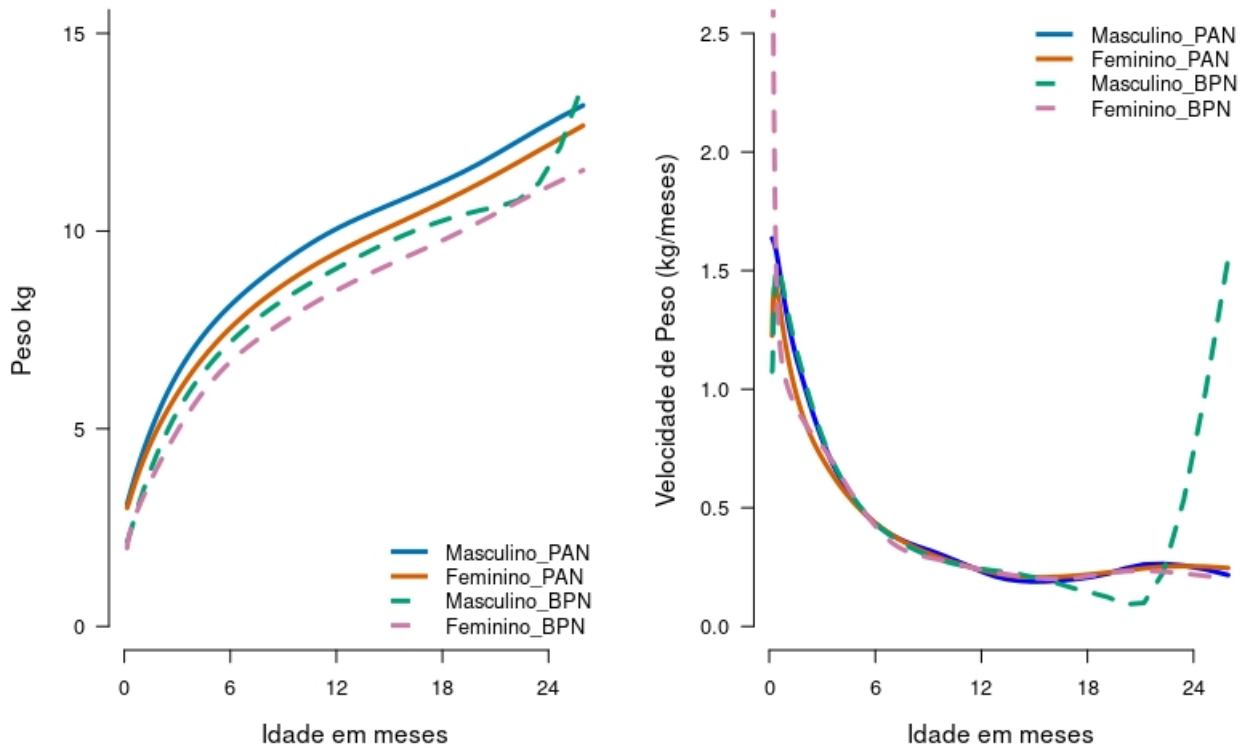


Figura 21 – Trajetória média do peso ajustado pelo modelo SITAR (esquerda), e curva de velocidade de crescimento (direito) para peso de crianças entre 0 e 2 anos.

Tabela 16 – Distribuição em relação as características da base de dados e amostra de 10%. Brasil 2008-2017

	População		Amostra 10%	
	n_{obs}	%	n_{obs}	%
Sexo				
Masculino	195190	50,54	18957	50,81
Feminino	191004	49,46	18312	49,19
Medidas Repetidas				
3	203349	52,65	20181	52,10
4	90652	23,47	9156	23,64
5	48125	12,46	4835	12,48
6	26748	6,93	2766	7,14
7	12138	3,14	1260	3,25
8	4408	1,14	440	1,14
9	774	0,20	99	0,26

software R, para fazer a estimação dos parâmetros do modelo marginal *Brokenstick*. Foi utilizado o modelo com os nós nos efeitos fixos e apenas com intercepto aleatório, visto que com $(k + 1)$

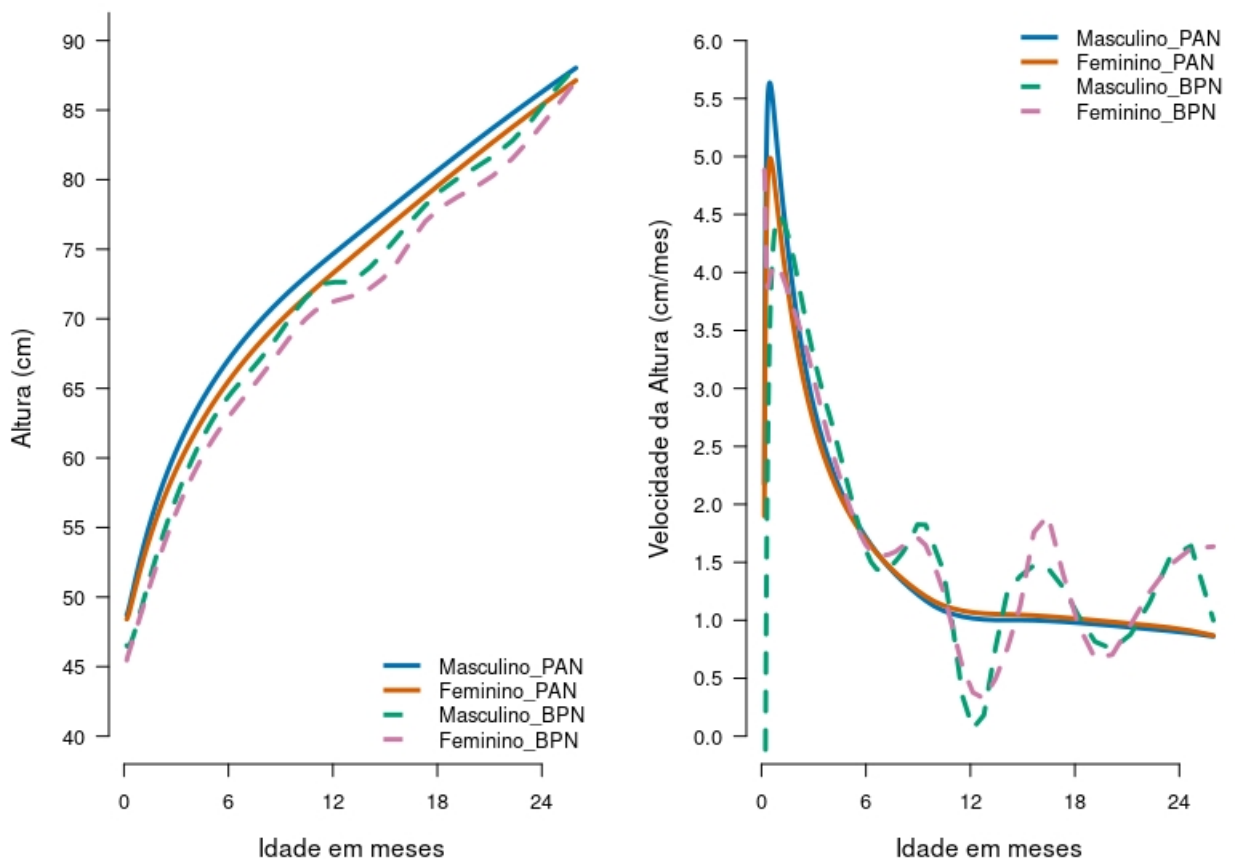


Figura 22 – Trajetória média da altura ajustada pelo modelo SITAR (esquerda), e curva de velocidade de crescimento (direito) para altura de crianças entre 0 e 2 anos.

parâmetros aleatórios referente as B-splines não houve convergência do modelo.

Nas Tabelas 17 a 20 são apresentadas as médias ajustadas dos pesos e alturas em cada período de idade-alvo para os modelos marginais *Brokenstick* provenientes da amostra em estudo, considerando o sexo e classificação de peso ao nascer. As Figuras 35 e 36 do gráfico de quantis normais dos resíduos de confundimento mínimo padronizado indicam a adequabilidade dos modelos *Brokenstick* para peso e altura, uma vez que o pressuposto de normalidade dos resíduos condicionais padronizados foi validado.

Para a modelagem conjunta das distribuições marginais (pelos modelos *Brokenstick* bivariados) referentes ao crescimento infantil, foram utilizados os modelos de cópulas bivariadas que permitem associação positiva entre as marginais, destacamos as cópulas elípticas Gaussianas, t de Student com graus de liberdade arbitrários (por exemplo, $\nu = 2$), e cópulas arquimedianas de Clayton, Frank e Gumbel, onde é possível explorar o cenário de dependências assimétrica e simétrica.

Os resultados das estimativas pelo método da inferência para as marginais - IFM dos

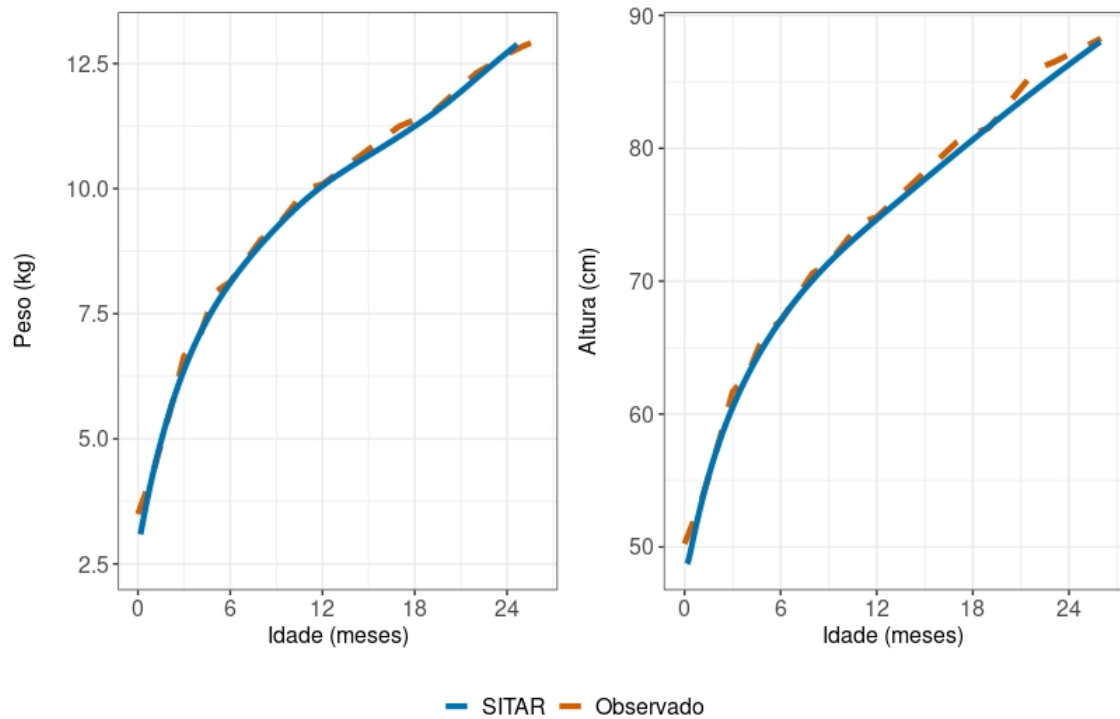


Figura 23 – Trajetória média, ajustada e observada, para peso e altura, pelo modelo SITAR para crianças PAN do sexo masculino de 0 a 2 anos

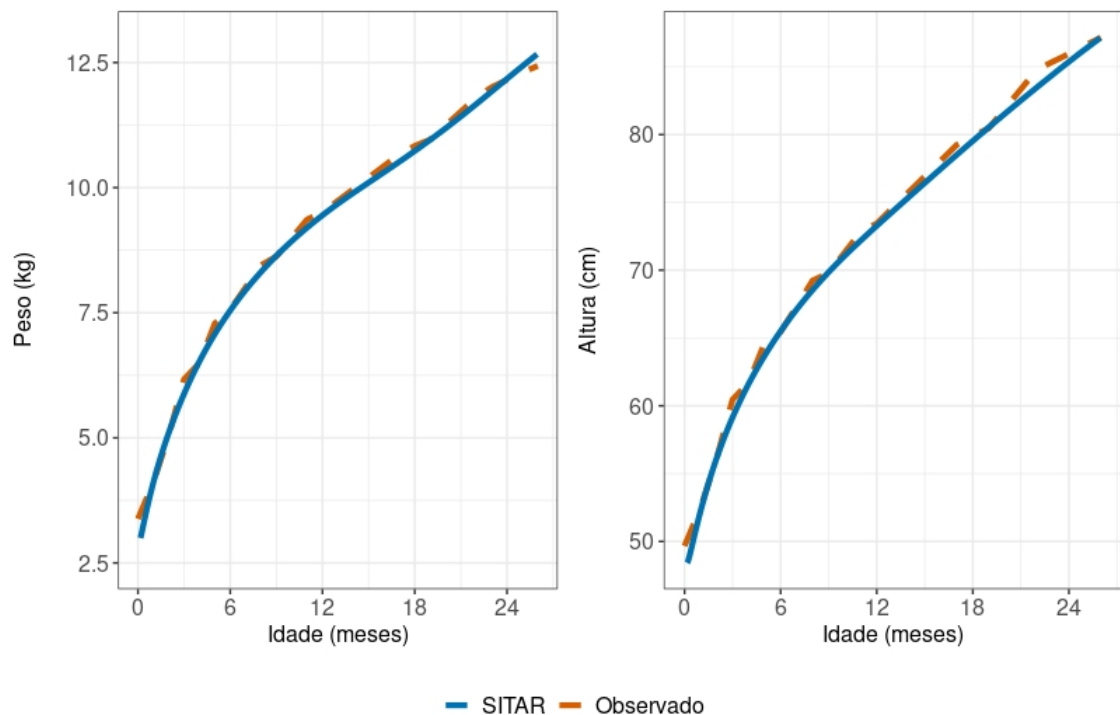


Figura 24 – Trajetória média, ajustada e observada, para peso e altura, pelo modelo SITAR para crianças PAN do sexo feminino de 0 a 2 anos

modelos *Brokenstick* baseados em cópulas bivariados ajustados aos dados longitudinais de crescimento infantil estão descritos nas Tabela 21 e 22, onde incluem os resultados das medi-

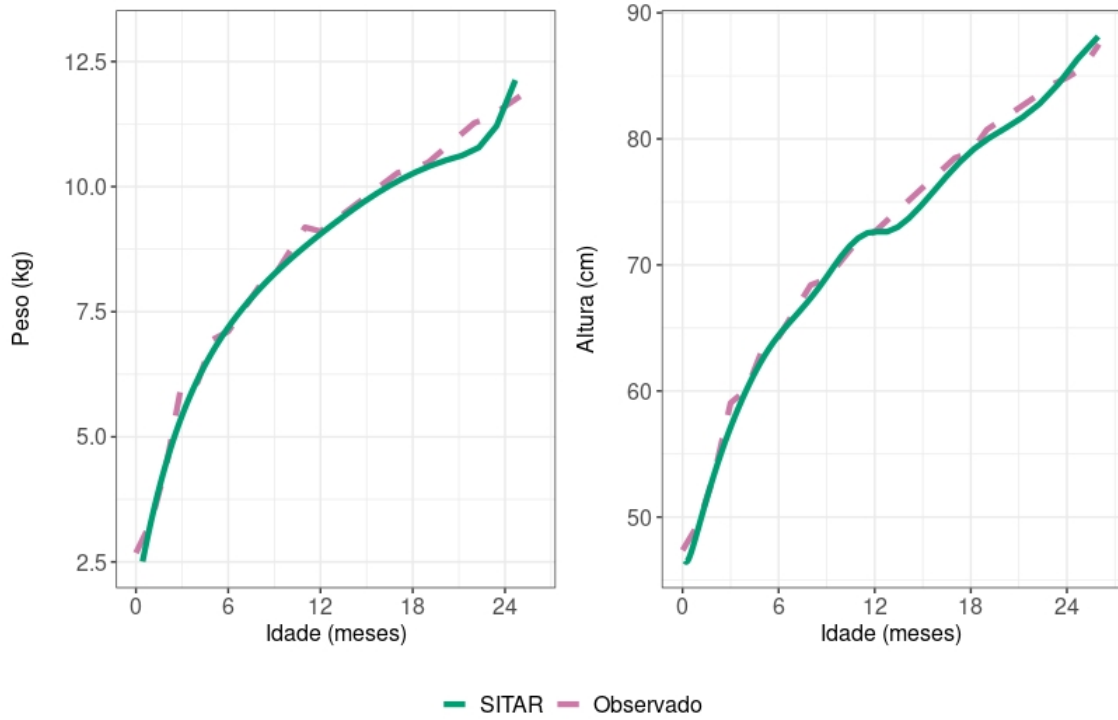


Figura 25 – Trajetória média ajustada e observada para peso e altura, pelo modelo SITAR para crianças BPN do sexo masculino de 0 a 2 anos

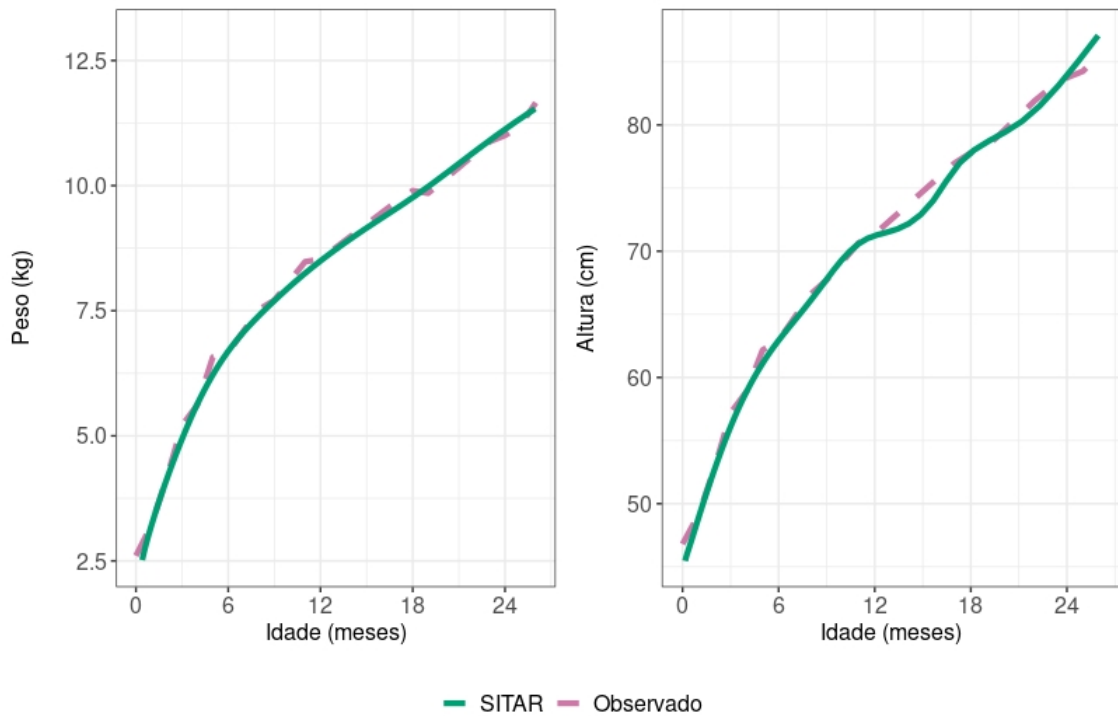


Figura 26 – Trajetória média ajustada e observada para peso e altura, pelo modelo SITAR para crianças BPN do sexo feminino de 0 a 2 anos

das correspondentes de dependência (tau de Kendall τ_K , rho de Spearman ρ_S , coeficientes de dependência das caudas inferior e superior (λ_L , λ_U)) e os valores de AIC e BIC.

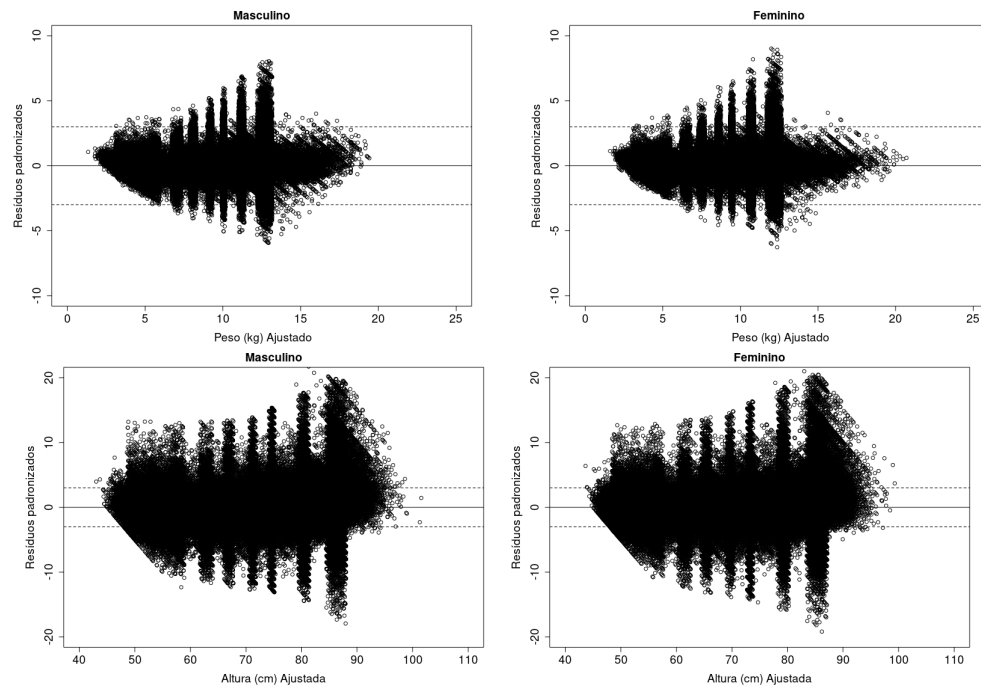


Figura 27 – Gráficos de dispersão dos resíduos padronizados em relação aos valores ajustados pelo modelo SITAR para variáveis antropométricas em crianças PAN de 0 a 2 anos por sexo.

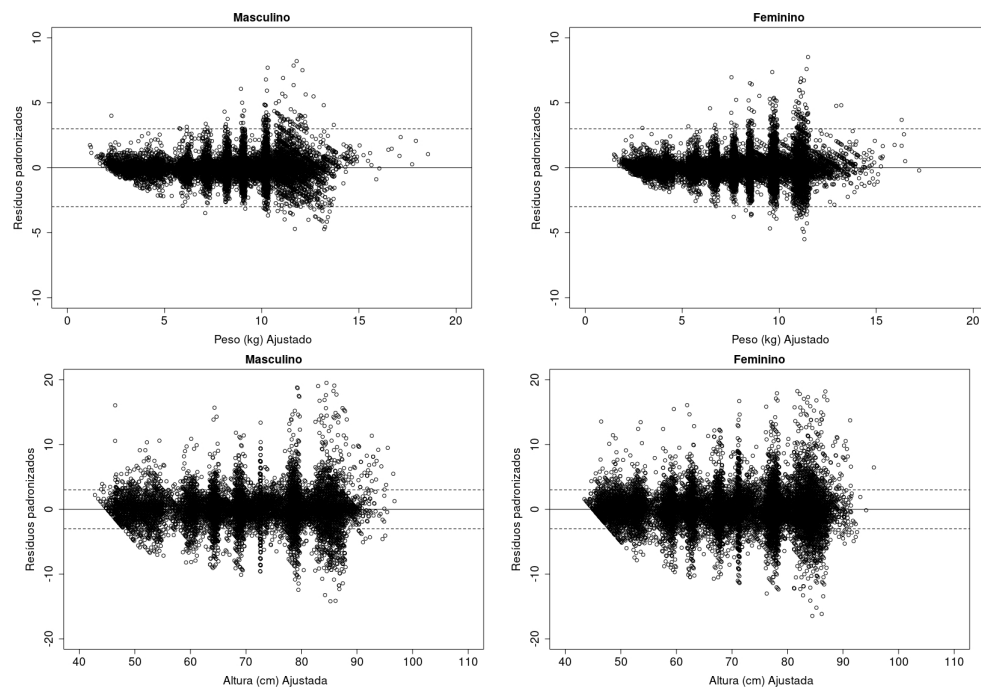


Figura 28 – Gráficos de dispersão dos resíduos padronizados em relação aos valores ajustados pelo modelo SITAR para variáveis antropométricas em crianças BPN de 0 a 2 anos por sexo.

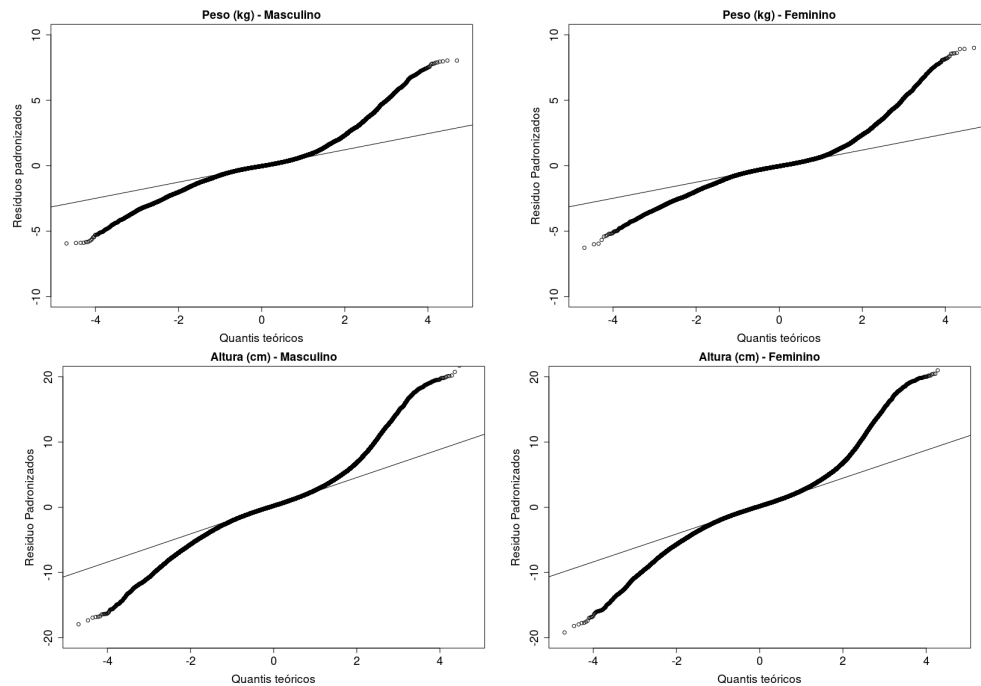


Figura 29 – Gráficos de probabilidade normal a partir dos modelos SITAR ajustado para variáveis antropométricas, para crianças PAN entre 0 e 2 anos por sexo

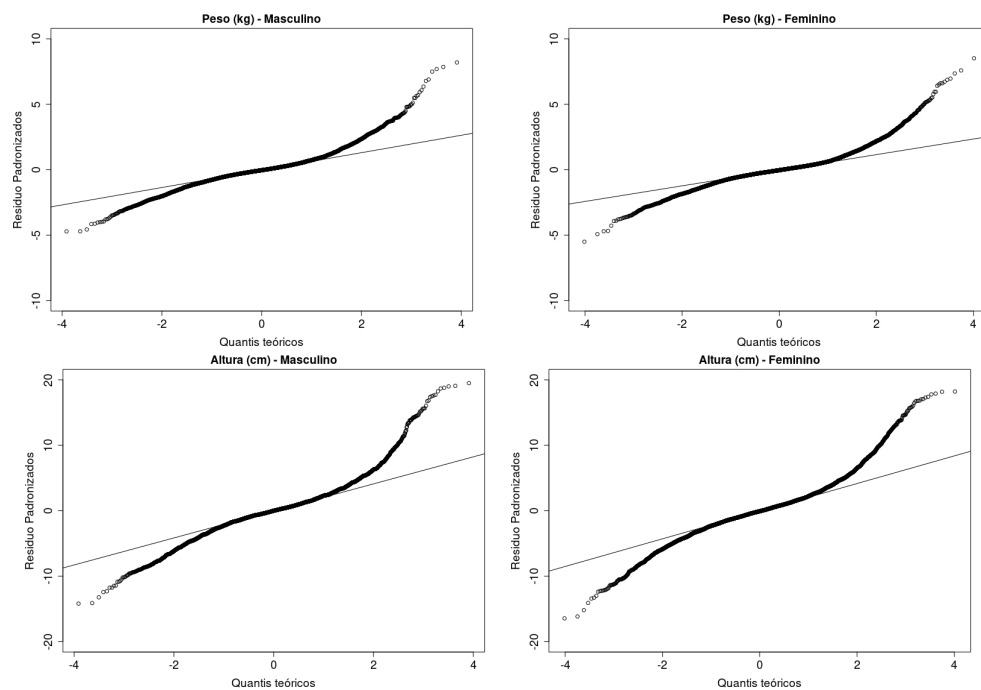


Figura 30 – Gráficos de probabilidade normal a partir dos modelos SITAR ajustado para variáveis antropométricas, para crianças BPN entre 0 e 2 anos por sexo

Nota-se que a abordagem conjunta por meio das cópulas bivariadas (Tabela 21) apresenta-se como uma melhor alternativa de modelagem quando comparada com a abordagem univariada

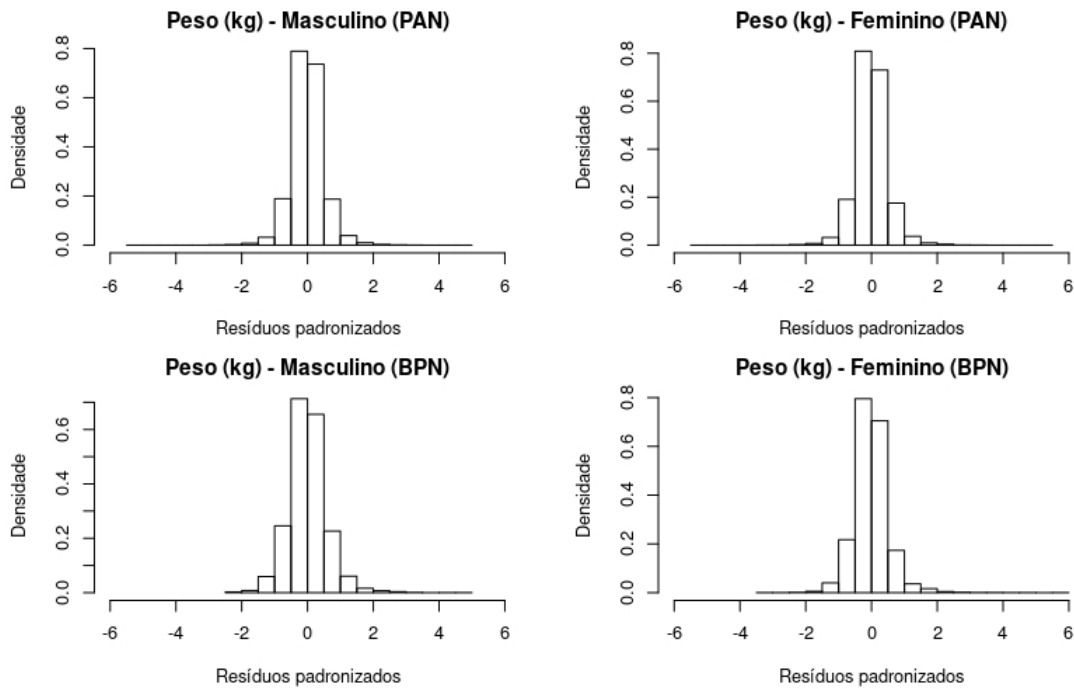


Figura 31 – Histograma dos resíduos padronizados pelo ajuste do modelo SITAR para o peso de crianças de 0 a 2 anos, por sexo e classificação de peso ao nascer

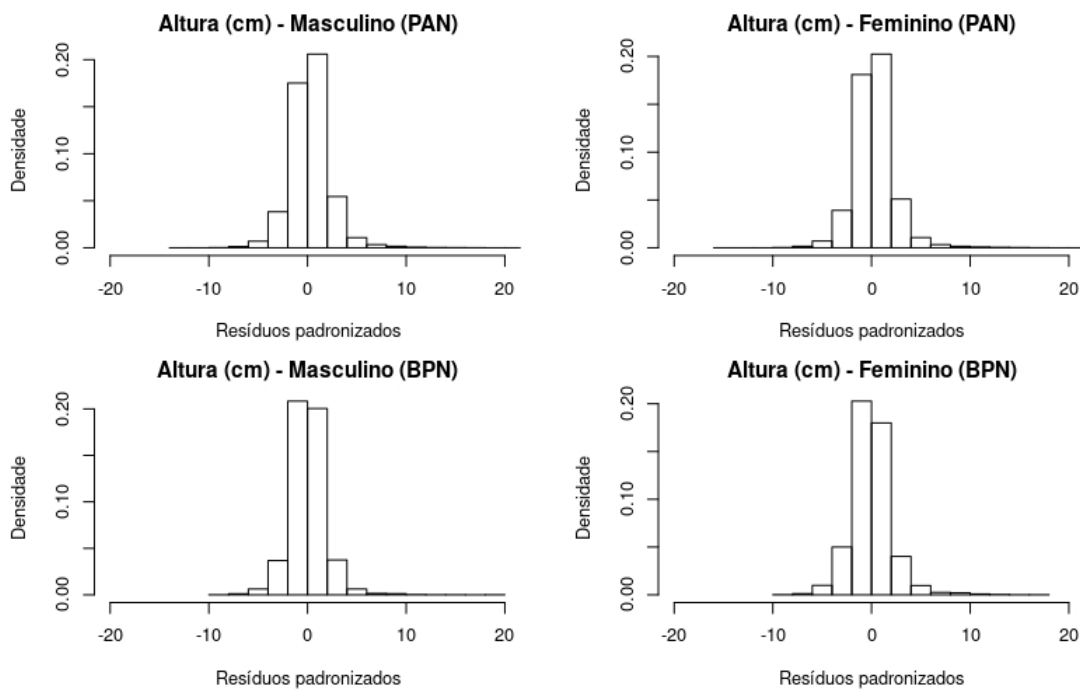


Figura 32 – Histograma dos resíduos padronizados pelo ajuste do modelo SITAR para a altura de crianças de 0 a 2 anos, por sexo e classificação de peso ao nascer

(Tabelas 17 e 18), isto pode ser observado comparando os valores AIC e BIC do modelos baseados em cópula que são inferiores ao AIC_{Total} e BIC_{Total} (exemplo: para crianças com

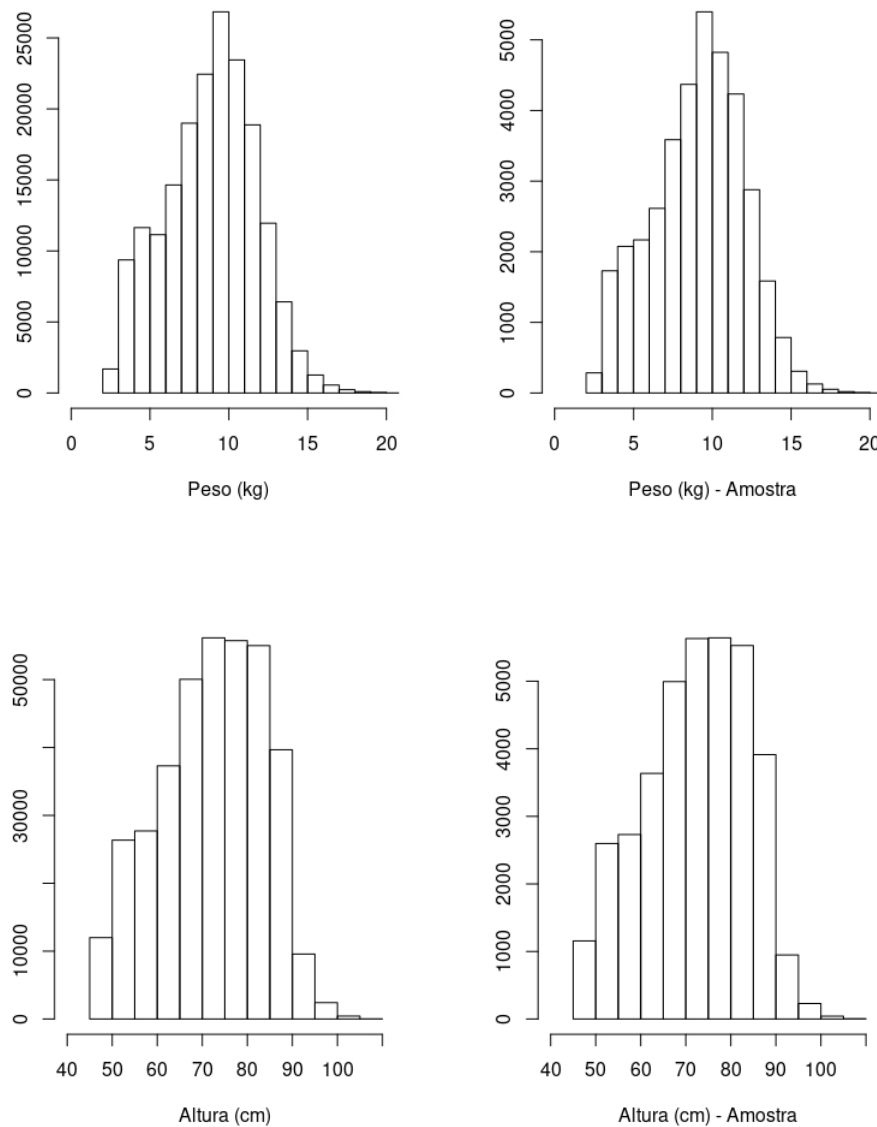


Figura 33 – Histogramas do peso e altura para crianças PAN de 0 a 2 anos. Brasil 2008-2017

peso adequado ao nascer, do sexo masculino $AIC_{\text{peso}} = 54007.58$ e $AIC_{\text{altura}} = 97880.06$, $AIC_{\text{total}} = 151887.64$) dos modelos univariados justificando a utilização das cópulas como uma ferramenta na estimação conjunta dos modelos *Brokenstick* para os dados antropométricos brutos, melhorando, assim, a eficiência estatística (FERREIRA et al., 2019).

Pelos resultados apresentados na Tabela 21 observamos que, para o sexo masculino, o modelo *Brokenstick* bivariado, baseados na cópula de Gumbel foi selecionado como o melhor, de acordo com ambos os critérios de AIC e BIC (tem os menores valores: $AIC = 137487.30$ e $BIC = 137636.40$). Neste melhor modelo, as medidas de dependência estimadas ($\hat{\tau}_K = 0,672$ e $\hat{\rho}_S = 0,853$) revelam uma moderada a forte relação positiva entre as marginais (peso e altura) e o

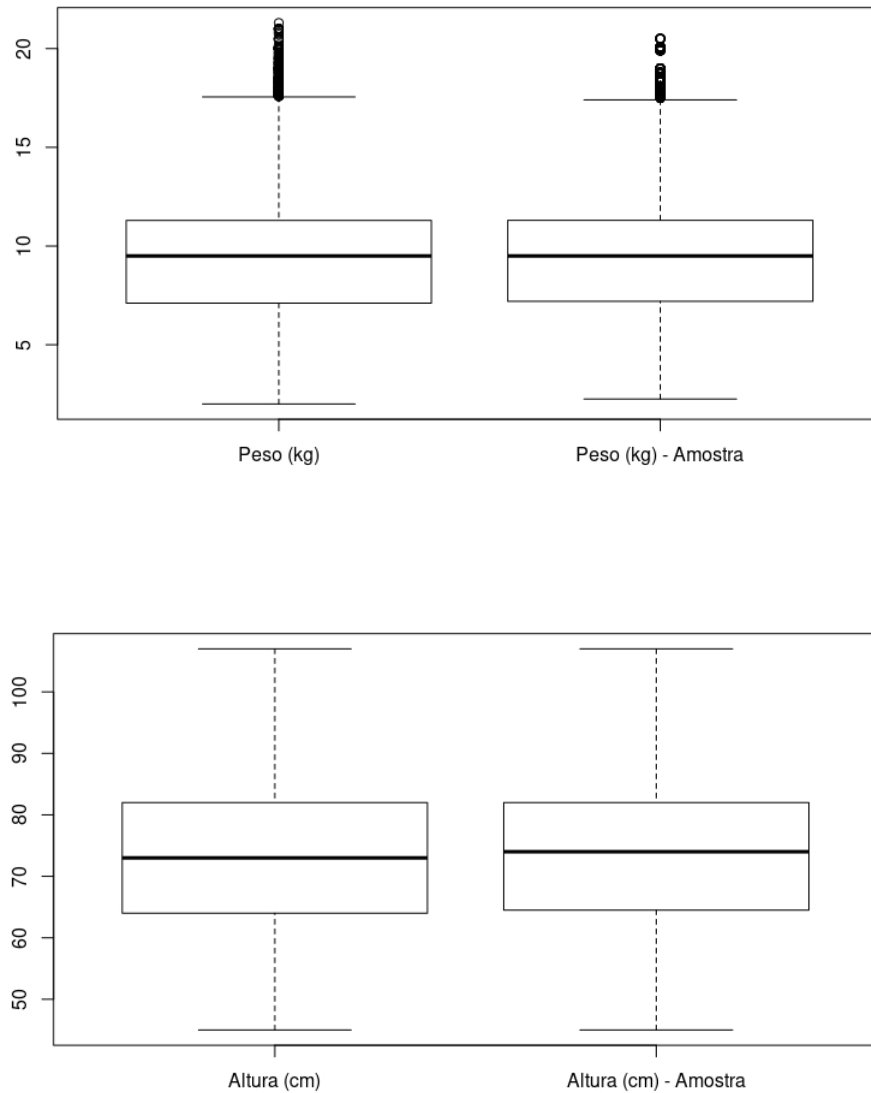


Figura 34 – Boxplot do peso e altura para crianças PAN de 0 a 2 anos. Brasil 2008-2017

coeficiente estimado de dependência caudal ($\lambda_U = 0.745$) mostra uma relação moderada positiva na cauda superior da distribuição conjunta, ou seja, para alturas e pesos elevados.

Observa-se ainda, na Tabela 21, para o sexo feminino, o modelo *Brokenstick* bivariado, o melhor ajuste foi derivado da cópula de Frank ($AIC = 134184.00$ e $BIC = 134332.60$). As medidas de dependência estimadas, ($\hat{\tau}_K = 0.539$ e $\hat{\rho}_S = 0,739$) mostram uma relação moderada positiva entre as marginais, e o coeficiente estimado de dependência da cauda como simétrico, ($\hat{\lambda}_L = \hat{\lambda}_U = 0$), ou seja, independentemente de quão alto seja o coeficiente de correlação, eventos extremos parecem ocorrer independentemente em cada margem. Nota-se na Tabela 22, que em ambos os sexo, o modelo *Brokenstick* bivariado, apresentou o melhor ajuste derivado da cópula

Tabela 17 – Ajuste do modelo *Brokenstick* para uma amostra de crianças PAN, do sexo masculino entre de 0 a 2 anos de idade. Brasil 2008-2017

Nós (mês)	Peso (kg)			Altura (cm)		
	Estimativa	EP	valor de p	Estimativa	EP	valor de p
0	3.08	0.04	<0.001	48.30	0.12	<0.001
2	5.53	0.03	<0.001	57.45	0.08	<0.001
6	8.21	0.02	<0.001	67.18	0.07	<0.001
12	10.04	0.02	<0.001	74.87	0.06	<0.001
24	12.68	0.02	<0.001	87.40	0.06	<0.001
$\hat{\sigma}_{intercepto}$		0.590			2.080	
$\hat{\sigma}_{residual}$		1.152			3.023	
$\hat{\rho}_{AR(1)}$		0.633			0.381	
AIC		54007.58			50860.12	
BIC		54078.23			50930.48	

Tabela 18 – Ajuste do modelo *Brokenstick* para uma amostra de crianças PAN, do sexo feminino entre de 0 a 2 anos de idade. Brasil 2008-2017

Nós (mês)	Peso (kg)			Altura (cm)		
	Estimativa	EP	valor de p	Estimativa	EP	valor de p
0	3.03	0.04	<0.001	48.18	0.12	<0.001
2	5.16	0.04	<0.001	56.33	0.08	<0.001
6	7.62	0.02	<0.001	65.33	0.07	<0.001
12	9.51	0.02	<0.001	73.59	0.06	<0.001
24	12.18	0.02	<0.001	86.28	0.06	<0.001
$\hat{\sigma}_{intercepto}$		0.263			1.988	
$\hat{\sigma}_{residual}$		1.235			2.950	
$\hat{\rho}_{AR(1)}$		0.711			0.379	
AIC		97880.06			93774.49	
BIC		97942.86			93837.03	

de Frank, de acordo com ambos os critérios de AIC e BIC. Para o sexo masculino, as medidas de dependência estimadas ($\hat{\tau}_K = 0.571$ e $\hat{\rho}_S = 0.772$) sugerem um relacionamento positivo e moderado, enquanto no sexo feminino é positivo de moderado a forte ($\hat{\tau}_K = 0.612$ e $\hat{\rho}_S = 0.812$).

Neste trabalho não foi possível estimar os erros padrões das estimativas para os modelos brokenstick derivados da cópula bivariada. Até o presente momento ainda não foi viável a sua estimação, via procedimento bootstrap, visto a limitação na instalação do pacote **Copula** do programa R no servidor de análises do CIDACS. Esforços tem sido empregado buscando novas estratégias que permitam a estimação desse parâmetro, inclusive levando em consideração o seu tempo de processamento.

Tabela 19 – Ajuste do modelo *Brokenstick* para crianças BPN, do sexo masculino de 0 a 2 anos de idade. Brasil 2008-2017

Nós (mês)	Peso (kg)			Altura (cm)		
	Estimativa	EP	valor de p	Estimativa	EP	valor de p
0	2.13	0.07	<0.001	45.13	0.23	<0.001
2	4.59	0.05	<0.001	53.93	0.15	<0.001
6	7.27	0.04	<0.001	64.85	0.12	<0.001
12	9.14	0.04	<0.001	72.85	0.11	<0.001
24	11.62	0.04	<0.001	85.16	0.12	<0.001
$\hat{\sigma}_{intercepto}$		0.420			2.100	
$\hat{\sigma}_{residual}$		1.190			3.095	
$\hat{\rho}_{AR(1)}$		0.680			0.421	
AIC		15278.36			23446.72	
BIC		15337.83			23502.96	

Tabela 20 – Ajuste do modelo *Brokenstick* para crianças com BPN, do sexo feminino de 0 a 2 anos de idade. Brasil 2008-2017

Nós (mês)	Peso (kg)			Altura (cm)		
	Estimativa	EP	valor de p	Estimativa	EP	valor de p
0	2.22	0.06	<0.001	45.09	0.19	<0,001
2	4.22	0.04	<0.001	53.08	0.12	<0,001
6	6.77	0.03	<0.001	63.37	0.09	<0,001
12	8.58	0.03	<0.001	71.50	0.08	<0,001
24	11.05	0.03	<0.001	83.89	0.09	<0,001
$\hat{\sigma}_{intercepto}$		0.465			1.638	
$\hat{\sigma}_{residual}$		1.105			3.168	
$\hat{\rho}_{AR(1)}$		0.642			0.379	
AIC		28360.50			42841.24	
BIC		28413.36			42897.47	

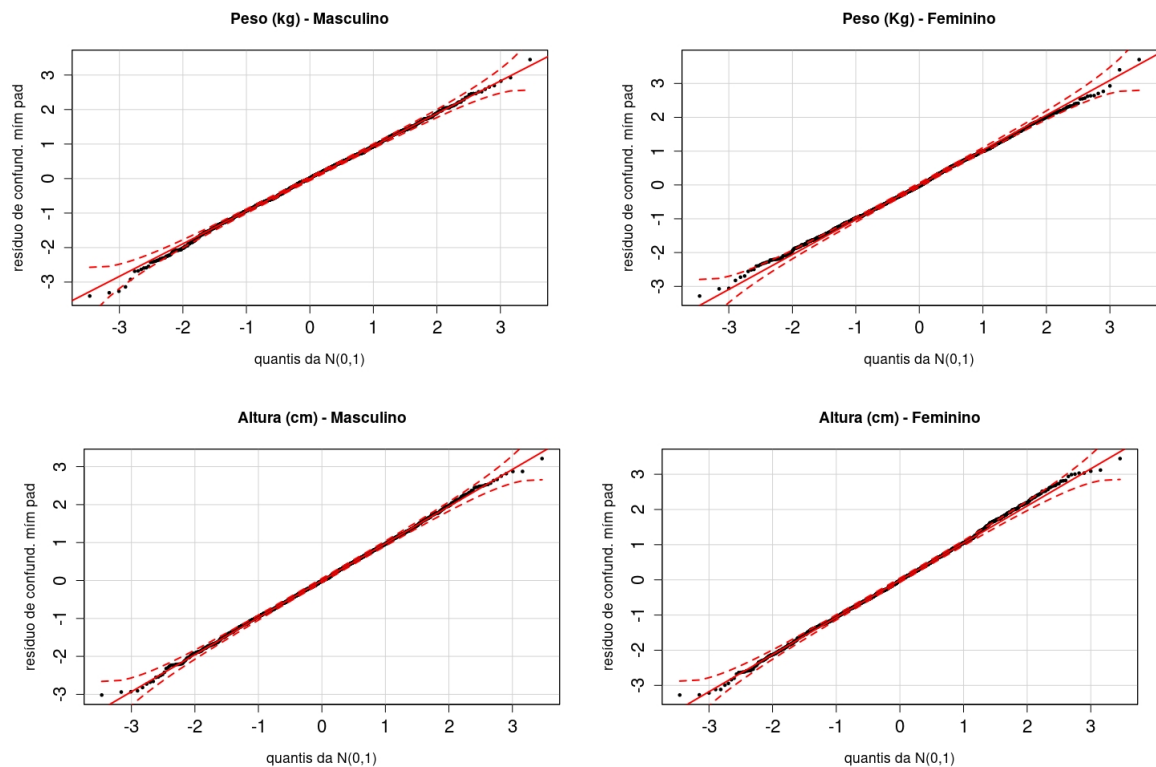


Figura 35 – Gráficos de quantis para os resíduos de confundimento mínimo padronizados obtidos a partir dos modelos ajustados para peso e altura de crianças PAN de 0 a 2 anos de idade por sexo

Tabela 21 – Principais resultados da estimativa IFM para os modelos *Brokenstick* baseados em cópulas bivariados ajustados aos dados de uma amostra de crianças PAN, e parâmetros de associação de cópulas por sexo

Cópula	$\hat{\theta}$	$\hat{\tau}_K$	$\hat{\rho}_S$	λ_L	λ_U	LL	AIC	BIC
Masculino								
Gaussiana	0.82	0.612	0.807	0	0	-68778.69	137595.40	137744.50
Student's t_2	0,78	0.570	0.765	0.521	0.521	-69034.89	138107.80	138256.90
Clayton	1.60	0.444	0.618	0.648	0	-69218.48	138475.00	138624.10
Frank	8.81	0.631	0.83	0	0	-68743.77	137525.50	137674.70
Gumbel	3.05	0.672	0.853	0	0.745	-68724.63	137487.30	137636.40
Feminino								
Gaussiana	0.72	0.512	0.703	0	0	-67162.16	134362.30	134510.90
Student's t_2	0.66	0.459	0.703	0.417	0.417	-67577.95	135193.90	135342.40
Clayton	1.08	0.351	0.501	0.526	0	-67542.93	135123.90	135272.40
Frank	6.50	0.539	0.739	0	0	-67073.02	134184.00	134332.60
Gumbel	2.31	0.567	0.755	0	0.65	-67180.73	134399.50	134548.00

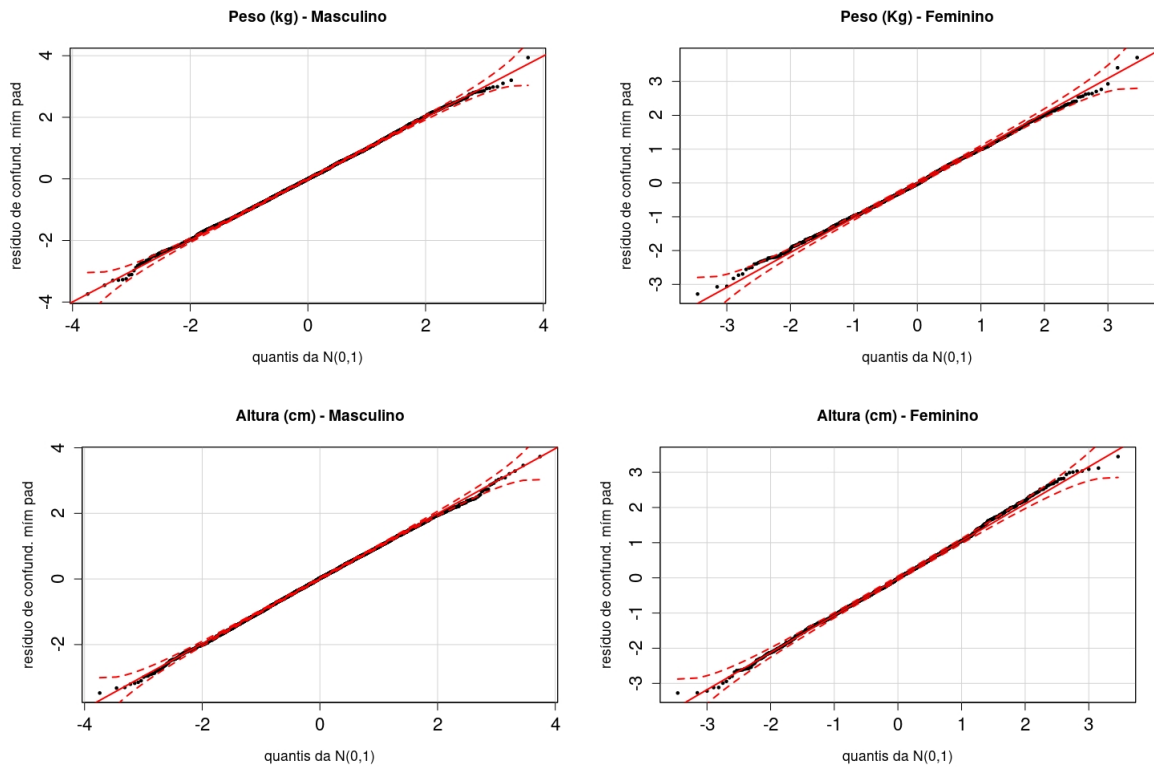


Figura 36 – Gráficos de quantis para os resíduos de confundimento mínimo padronizados obtidos a partir dos modelos ajustados para peso e altura de crianças BPN de 0 a 2 anos de idade por sexo

Tabela 22 – Principais resultados da estimativa para os modelos *Brokenstick* baseados em cópulas bivariados ajustados aos dados de uma amostra de crianças BPN, e parâmetros de associação de cópulas por sexo

Cópula	$\hat{\theta}$	$\hat{\tau}_K$	$\hat{\rho}_S$	λ_L	$\hat{\lambda}_U$	LL	AIC	BIC
Masculino								
Gaussiana	0.76	0.550	0.744	0.456	0.456	-20090.35	40218.69	40344.24
Student's t_2	0.71	0.503	0.693	0	0	-20178.07	40394.14	40519.69
Clayton	1.27	0.388	0.549	0.579	0	-20203.11	40444.21	40569.76
Frank	7.20	0.571	0.772	0	0	-20052.22	40142.44	40268.00
Gumbel	2.53	0.605	0.793	0	0.685	-20079.67	40197.34	40322.90
Feminino								
Gaussiana	0.80	0.590	0.786	0	0	-29828.09	59694.04	59827.62
Student's t_2	0.75	0.540	0.734	0.492	0.492	-29959.48	59956.96	60090.54
Clayton	1.46	0.422	0.591	0.622	0	-30010.44	60058.87	60192.45
Frank	8.25	0.612	0.812	0	0	-29809.16	59656.33	59786.9
Gumbel	2.86	0.650	0.835	0	0.726	-29815.06	59668.12	59801.7

5 Discussão

Neste estudo, foi sugerido estimar trajetórias de crescimento com modelagem conjunta e univariada (marginal) para as medidas antropométricas de peso e altura. Comparamos três modelos diferentes comumente usados para caracterizar essas trajetórias, em termos de ajuste univariado, para padrões de crescimento em uma base de dados (*linkage* das bases cadastro Único, SISVAN e SINASC), com crianças entre 0 e 2 anos de idade no período de 2008 a 2017.

Estudos longitudinais registram a tendência do crescimento no tempo, resguardando as características de cada criança e entre elas. É consensual que a análise estatística de dados longitudinais necessita de técnicas robustas que permitam levar em consideração a dependência entre as medidas, a mudança do crescimento no tempo e o desbalanceamento dos dados (DIGGLE et al., 2002; FAUSTO et al., 2008). Neste contexto, os métodos para avaliar a trajetória de crescimento é uma técnica estatística que atende à suposição de que as observações entre indivíduos não são independentes e que as medidas intraindividuais possuem relação de dependência com erros correlacionados.

Realizou-se uma comparação para identificar quais das modelagens propostas apresentou melhor desempenho em termos de verdadeira estimativa da trajetória de crescimento. A comparação entre os parâmetros estimados a partir dos modelos linear de efeitos mistos, *Brokenstick* e o SITAR, utilizados para análise de dados longitudinais univariados, permitiu sugerir que o modelo linear de efeitos mistos e o SITAR não se ajustaram na avaliação do crescimento da criança nos dois primeiros anos de vida, da base de dados em estudo, quando comparado ao modelo *Brokenstick*.

As crianças do sexo masculino, independente do grupo de peso ao nascer, apresentou ganho ponderal e linear mais elevado nos dois primeiros anos de vida quando comparado ao incremento proporcionado pelo sexo feminino, corroborando com outros achados descritos na literatura (COLE, 2002; BEATH, 2007; FAUSTO et al., 2008; HOWE et al., 2016; FILTEAU et al., 2019; OHUMA et al., 2021).

Ainda que a estimativa média do crescimento produzida por todos estes modelos não tenha apresentado diferenças expressivas, os erros produzidos pela diferença entre a trajetória estimada e observada pelo modelo *Brokenstick* mostraram menor dispersão. Neste estudo, observou-se que a velocidade ou incremento do ritmo do crescimento não eram constantes no tempo. Além disto,

constatou-se que a velocidade do crescimento desacelerava à medida que aumentava a idade da criança (CAMERON, 2012).

Na abordagem de Laird e Ware (1982), embora com a inclusão do termo quadrático no modelo tenha sido possível a correção da variabilidade da distância da velocidade do crescimento nos diferentes tempos (CAMERON, 2012), não foi flexível o suficiente para modelar a curva de crescimento com base nesses dados, este resultado também foi observado no estudo de Anderson e colaboradores (2019). O modelo SITAR não apresentou um bom ajuste aos dados, além de apresentar percentual de explicação na variação do peso ou altura média inferior ao encontrado na literatura. Em estudos de crescimento infantil, na mesma faixa etária, o percentual de explicação na variação do peso ou altura média foi superior a 80%, com graus de liberdade variando entre 4 e 6 (COLE, 2018; OHUMA et al., 2021). A falta de avaliação do diagnóstico do modelo na literatura é uma limitação do modelo SITAR, cito o estudo de Beath (2007) que encontrou resíduos similares, mas em outros artigos citados neste trabalho, a análise de resíduo não é relatada, o que dificulta uma comparação a respeito da adequação do ajuste deste modelo e ainda se outra abordagem mais específica deve ser realizada afim de validar às suposições do modelo. No entanto, a abordagem *Brokenstick* forneceu uma alternativa confiável e pode ser útil nos casos em que temos conjuntos de dados maiores (ANDERSON et al., 2019; BUUREN, 2020), ao contrário do SITAR que apresenta dificuldade de convergência em banco com grandes volumes de dados (ANDERSON et al., 2019; RIDDELL et al., 2017b).

Nestas circunstâncias, os parâmetros estimados pelo modelo linear de efeitos mistos por splines (*Brokenstick*), tanto para o crescimento ponderal quanto para o linear, expressaram de forma mais adequada o padrão de crescimento das crianças. É provável que essa abordagem funcione com sucesso em uma ampla gama de conjuntos de dados e, como resultado, recomendamos como abordagem ideal de modelagem para estes dados.

Neste estudo, foram usadas duas abordagens, univariada e bivariada, para lidar com dados longitudinais multivariados. A principal limitação na abordagem univariada é a impossibilidade de modelar conjuntamente algumas variáveis ao longo do tempo, levando também em consideração a correlação entre as variáveis respostas (CHO, 2016). Dessa forma, na segunda abordagem, o principal interesse foi a modelagem conjunta do peso e altura dada a correlação natural entre os mesmos, então explorou-se a modelagem via funções de cópulas bivariadas, com intuito de melhorar a modelagem de dados bivariados para estimar a dependência (NELSEN, 2007; CHO, 2016). Neste estudo, os resultados inéditos na literatura dos modelos *Brokenstick* bivariados ajustados foram satisfatórios, levando em conta o grau de associação (positiva) entre as variáveis.

A abordagem de cópula, possui alguma limitação no sentido de complexidade no procedimento de estimação do ponto de vista computacional, resultando numa maior dificuldade em obter estimativas (FERREIRA et al., 2019). Embora existam usos específicos para a escolha entre diferentes tipos de modelos de cópulas, uma seleção entre os modelos candidatos é necessário quando se utiliza a abordagem proposta (NELSEN, 2007; CHO, 2016).

A abordagem conjunta por meio das cópulas bivariadas apresentou-se como uma melhor alternativa de modelagem quando comparada com a abordagem univariada, justificando a utilização das cópulas como uma ferramenta na estimação conjunta dos modelos *Brokenstick* para os dados antropométricos brutos, melhorando, assim, a eficiência estatística (FERREIRA et al., 2019).

Pelos resultados apresentados neste estudo, para crianças do sexo masculino (PAN)), o modelo *Brokenstick* bivariado, baseados na cópula de Gumbel foi selecionado como o melhor modelo, onde as medidas de dependência estimadas revelam uma moderada a forte relação positiva entre as marginais (peso e altura) e o coeficiente estimado de dependência caudal mostra uma relação moderada positiva para alturas e pesos elevados. Este resultado também foi encontrado por Ferreira e colaboradores (2019). Para os demais grupos estudados a cópula de Frank foi escolhida devido à maior precisão quanto à representação da dependência entre as variáveis peso e altura.

Os resultados finais obtidos neste estudo indicam que os modelos *Brokenstick* bivariados ajustados foram satisfatórios, levando em conta a dependência moderada (positiva) entre peso e altura, ou seja, crianças mais magras estão associados com menor altura, assim como crianças mais pesadas estão associadas com maior altura.

6 Conclusão

Esta tese teve como objetivo geral estimar trajetórias de crescimento com modelagem conjunta e univariada (marginal) para as medidas antropométricas de peso e altura, na população menor de dois anos de idade, em uma coorte eletrônica de base populacional linkados com dados de rotina de nascidos vivos (SINASC) e do Sistema de Vigilância Alimentar e Nutricional (SISVAN) do Ministério da Saúde, no período de 2008 a 2017. O acompanhamento desta trajetória permite auxiliar no monitoramento do crescimento e desenvolvimento infantil que são ações essenciais para a prevenção da desnutrição, obesidade e promoção da saúde na infância, permitindo caracterizar diferentes padrões de crescimento, como, por exemplo, as crianças que crescem com sucesso e aquelas cujo crescimento é menos saudável.

Uma característica de dados longitudinais observacionais é a presença de observações que variam no tempo entre os indivíduos. Esses dados também podem apresentar tempos de acompanhamento irregulares, a exemplo do ocorrido na base de dados em estudo, que são desafios significativos para a análise quantitativa. Dentre as estratégias disponíveis na literatura para análise dos tempos de observações irregulares, adotamos a sugerida por Buuren (2020) e Lukko *et al* (2020) que foi a de criar uma grade de tempo pré-especificada, denominada de “idade alvo”, com o objetivo de padronizar a idade e de colocar o espaçamento de maneira igualitária para todos. No entanto, o banco de dados ainda apresentou desbalanceamento entre as medidas.

Para diminuir as incertezas em relação à estratégia sugerida foram criados três bancos de dados "Completo", "Proposta" e "PropostaMed", com o objetivo de compará-los e escolher aquele que viabilizaria os ajustes dos modelos de trajetória pelo modelo *Brokenstick*. Adotou-se o banco de dados PropostaMed para ajuste dos modelos univariados e bivariados, pois nele o tempo de processamento foi otimizado, houve convergência dos modelos e não foi observado uma diferença significativa entre as curvas ajustadas e variâncias residuais.

Neste estudo, foram utilizadas duas abordagens para lidar com dados longitudinais multivariados. Na primeira abordagem sugerimos uma comparação de modelos comumente usados para caracterizar as trajetórias de crescimento infantil, como os Modelos de Efeitos Mistos, *Brokenstick* e o SITAR, para modelar as medidas de peso e altura na sua forma bruta. Na segunda abordagem estudamos a relação de dependência dessas variáveis conjuntamente no padrão de crescimento infantil através da abordagem de Cópula.

A comparação entre os parâmetros estimados e análise de resíduos a partir de modelos

de regressão linear de efeitos mistos, *Brokenstick* e SITAR, utilizados para análise de dados longitudinais, permitiu concluir que o modelo linear de efeitos mistos não foi flexível o suficiente para modelar a curva de crescimento e SITAR não apresentou um bom ajuste aos dados. Dessa forma estes modelos não se ajustaram na avaliação do crescimento da criança nos dois primeiros anos de vida, quando comparado ao modelo *Brokenstick*, para a base de dados em estudo.

O modelo multinível de *spline* linear forneceu um método útil para resumir as trajetórias de crescimento, reduzindo a dimensionalidade dos dados em medições de crescimento e fornecendo resumos de crescimento interpretáveis. Os resultados sugerem que as incertezas associadas à transformação de vários tempos de observação em medidas repetidas menos irregulares foram pequenas. É provável que a abordagem *Brokenstick* funcione com sucesso em uma ampla gama de conjuntos de dados. A falta de avaliação do diagnóstico do modelo final com a análise de resíduos é uma limitação dessa abordagem, pois ainda não está implementada no pacote *Brokenstick* do R. O critério de escolha da posição dos nós foi respaldado pela literatura.

A abordagem univariada traz como limitação a impossibilidade de modelar conjuntamente algumas variáveis ao longo do tempo, levando também em consideração a correlação entre as variáveis respostas. Entretanto, na segunda abordagem, o principal interesse é a modelagem conjunta dessas variáveis dada a correlação natural entre as mesmas. Dessa forma, explorou-se a modelagem via funções de cópulas bivariadas, com intuito de melhorar a modelagem de dados bivariados com formas mais flexíveis de se estimar essa dependência. Levando em consideração os resultados dos critérios de seleção, as cópulas Gumbel (crianças do sexo masculino (PAN)) e Frank (demais grupos estudados), foram escolhidas devido à maior precisão quanto à representação da dependência entre as variáveis peso e altura.

Por fim, na prática, o entendimento destas trajetórias de crescimento, poderá apoiar as discussões acerca das ações de Vigilância Alimentar e Nutricional (VAN) no âmbito da Atenção Primária a Saúde no SUS. A médio prazo, subsidiar o planejamento de ações para prevenção e enfrentamento dos agravos relacionados aos desvios das trajetórias de crescimento infantil, onde é preciso intensificar ações de prevenção e promoção da saúde focalizando os fatores associados ao excesso de peso em idade mais tenras e assim reduzir a gravidade dessa condição e, com isso, reduzir o risco de desfechos adversos crônicos ao longo da vida. Ao longo prazo, assegurar a segurança alimentar e nutricional desta e de novas gerações.

6.1 Perspectivas Futuras

- Ressalta-se que no Brasil são escassos os estudos longitudinais que avaliam a dinâmica de crescimento físico e composição corporal na primeira infância, o que pode ser devido às dificuldades relacionadas a logística complexa e ao alto custo da pesquisa longitudinal. A base de dados em estudo apresenta a potencialidade de responder a essas questões sem limitações amostrais e estratificada por grupos de interesse, como características perinatal.

Destaca-se a necessidade de trabalhos futuros ampliando a faixa etária estudada, para crianças de 0 a 5 anos e considerando o sexo, peso ao nascer e prematuridade, entendendo que o monitoramento das medidas antropométricas são importantes índices de saúde na infância e a forma mais adequada de avaliar o estado nutricional e de saúde geral.

- Os pacotes do programa R, SITAR e *Brokenstick* foram desenvolvidos para admitir apenas a variável idade como covariável. Um desafio é poder implementar a inclusão de mais covariáveis no modelo permitindo que outras variáveis preditoras possam contribuir na explicação dessa trajetória, e ainda implementar a avaliação das suposições de validação do modelo através da análise gráfica dos resíduos.
- Uma limitação encontrada neste trabalho e que nos motiva a buscar soluções é a estimação dos erros padrões das estimativas para os modelos *Brokenstick* derivados da cópula bivariada. Esforços têm sido empregados buscando novas estratégias que permitam a estimação desse parâmetro, inclusive levando em consideração o seu tempo de processamento.
- Em estudos futuros pode ser interessante testar a relação de dependências em indicadores antropométricos padronizados, explorando novas famílias de cópula que possam se ajustar melhor aos dados e verificar se a relação de dependência entre as medidas antropométricas varia ao longo do tempo. Além disso, buscar estratégia computacional, como por exemplo, a paralelização como meio a viabilizar a análise de cópula envolvendo dados massivos, reduzindo o tempo computacional que foi um dos grandes desafios encontrados.

Referências

- AKAIKE, H. On entropy maximization principle. *Application of statistics*, North-Holland Publishing Company, p. 27–41, 1977. 13
- ALI, M. S. et al. Administrative data linkage in brazil: potentials for health technology assessment. *Frontiers in pharmacology*, Frontiers, v. 10, p. 984, 2019. 30
- ALMEIDA, D. et al. Examining the quality of record linkage process using nationwide brazilian administrative databases to build a large birth cohort. *BMC Medical Informatics and Decision Making*, BioMed Central, v. 20, n. 1, p. 1–9, 2020. 30
- ANDERSON, C. et al. Comparing predictive abilities of longitudinal child growth models. *Statistics in medicine*, Wiley Online Library, v. 38, n. 19, p. 3555–3570, 2019. 1, 2, 9, 14, 15, 16, 44, 47, 73
- BARBOSA, G. C. et al. Cidacs-rl: a novel indexing search and scoring-based record linkage system for huge datasets with high accuracy and scalability. *BMC medical informatics and decision making*, BioMed Central, v. 20, n. 1, p. 1–13, 2020. 30
- BARRETO, M. L. et al. The centre for data and knowledge integration for health (cidacs): linking health and social data in brazil. *International Journal of Population Data Science*, Swansea University, v. 4, n. 2, 2019. 31
- BEATH, K. J. Infant growth modelling using a shape invariant model with random effects. *Statistics in medicine*, v. 26, n. 12, p. 2547–2564, 2007. 16, 72
- BERKEY, C. S.; REED, R. B. A model for describing normal and abnormal growth in early childhood. *Human biology*, JSTOR, p. 973–987, 1987. 9
- BOOR, C. D.; BOOR, C. D. *A practical guide to splines*. [S.l.]: springer-verlag New York, 1978. v. 27. 15
- BOTTON, J. et al. Postnatal weight and height growth velocities at different ages between birth and 5 y and body composition in adolescent boys and girls. *The American journal of clinical nutrition*, Oxford University Press, v. 87, n. 6, p. 1760–1768, 2008. 9
- BRASIL. *Manual Operacional para uso do Sistema de Vigilância Alimentar e Nutricional: SISVAN VERSÃO 3.0*. [S.l.]: Ministério da Saúde, 2017. Disponível em: <<http://sisaps.saude.gov.br/sisvan/documentos/index>>, acesso em 09/04/2021. 29
- BRASIL. *Secretaria de Vigilância da Saúde. Sistema de Informações sobre Nascidos Vivos (SINASC): Coordenação Geral de Informações e Análises Epidemiológicas - CGIAE*. 2018. Disponível em: <<http://svs.aids.gov.br/dantps/cgiae/sinasc/>>, acesso em 09/04/2021. 29, 30
- BUUREN, S. van. Broken stick model for irregular longitudinal data. *Journal of Statistical Software*, Foundation for Open Access Statistics, Submitted for publication, p. 1–47, 2020. 3, 7, 8, 13, 14, 15, 16, 32, 44, 45, 73
- CAMERON, N. The human growth curve, canalization and catch-up growth. In: *Human growth and development*. [S.l.]: Elsevier, 2012. p. 1–22. 42, 73

- CEF. *Caixa Econômica Federal. Cadastro Único*. 2018. Disponível em: <<http://www.caixa.gov.br/cadastros/cadastro-unico/Paginas/default.aspx>>, acesso em 09/04/2021. 29
- CHEN, Y. et al. An epidemiological survey on low birth weight infants in china and analysis of outcomes of full-term low birth weight infants. *BMC pregnancy and childbirth*, BioMed Central, v. 13, n. 1, p. 1–9, 2013. 31
- CHO, H. The analysis of multivariate longitudinal data using multivariate marginal models. *Journal of Multivariate Analysis*, Elsevier, v. 143, p. 481–491, 2016. 19, 73
- COLE, T.; COLE, M. T. Package ‘sitar’. 2021. 17, 18, 32
- COLE, T. J. Assessment of growth. *Best Practice & Research Clinical Endocrinology & Metabolism*, Elsevier, v. 16, n. 3, p. 383–398, 2002. 1, 72
- COLE, T. J. Optimal design for longitudinal studies to estimate pubertal height growth in individuals. *Annals of human biology*, v. 45, n. 4, p. 314–320, 2018. 16, 47, 73
- COLE, T. J.; DONALDSON, M. D.; BEN-SHLOMO, Y. Sitar—a useful instrument for growth curve analysis. *International journal of epidemiology*, Oxford University Press, v. 39, n. 6, p. 1558–1566, 2010. 16, 47, 52
- CONFORTIN, S. C. et al. Rps brazilian birth cohorts consortium (ribeirão preto, pelotas and são luís): history, objectives and methods. *Cadernos de Saúde Pública*, SciELO Brasil, v. 37, 2021. 1
- CROZIER, S. R. et al. A discussion of statistical methods to characterise early growth and its impact on bone mineral content later in childhood. *Annals of human biology*, Taylor & Francis, v. 46, n. 1, p. 17–26, 2019. 16
- DIGGLE, P. et al. *Analysis of longitudinal data*. [S.l.]: Oxford university press, 2002. 2, 7, 42, 72
- DOBSON, A. et al. Flexible age-period-cohort modelling illustrated using obesity prevalence data. *BMC medical research methodology*, Springer, v. 20, n. 1, p. 1–9, 2020. 1
- EILERS, P. H.; MARX, B. D. Flexible smoothing with b-splines and penalties. *Statistical science*, Institute of Mathematical Statistics, v. 11, n. 2, p. 89–121, 1996. 13
- ESCARELA, G.; HERNÁNDEZ, A. Modelado de parejas aleatorias usando cópulas. *Revista Colombiana de Estadística*, Universidad Nacional de Colombia, v. 32, n. 1, p. 33–58, 2009. 21
- FAUSTO, M. A. et al. O modelo de regressão linear misto para dados longitudinais: uma aplicação na análise de dados antropométricos desbalanceados. *Cadernos de Saúde Pública*, SciELO Public Health, v. 24, p. 513–524, 2008. 72
- FERREIRA, P. H. et al. Bivariate copula-based linear mixed-effects models: An application to longitudinal child growth data. *TEMA (São Carlos)*, SciELO Brasil, v. 20, p. 37–59, 2019. 19, 21, 24, 25, 66, 73, 74
- FILTEAU, S. et al. Steady growth in early infancy is associated with greater anthropometry in indian children born low birth weight at term. *The Journal of nutrition*, Oxford University Press, v. 149, n. 9, p. 1633–1641, 2019. 16, 72
- FITZMAURICE, G. et al. *Longitudinal data analysis*. [S.l.]: CRC press, 2008. 7

- FITZMAURICE, G. M.; LAIRD, N. M.; WARE, J. H. *Applied longitudinal analysis*. [S.l.]: John Wiley & Sons, 2012. v. 998. 9
- FRAHM, G.; JUNKER, M.; SZIMAYER, A. Elliptical copulas: applicability and limitations. *Statistics & Probability Letters*, Elsevier, v. 63, n. 3, p. 275–286, 2003. 23
- GOLDSTEIN, H. et al. Multilevel growth curve models that incorporate a random coefficient model for the level 1 variance function. *Statistical methods in medical research*, SAGE Publications Sage UK: London, England, v. 27, n. 11, p. 3478–3491, 2018. 6, 16
- HARRON, K. et al. Challenges in administrative data linkage for research. *Big data & society*, Sage Publications Sage UK: London, England, v. 4, n. 2, p. 2053951717745678, 2017. 3, 4
- HAUSPIE, R. C.; CAMERON, N.; MOLINARI, L. *Methods in human growth research*. [S.l.]: Cambridge University Press, 2004. v. 39. 6, 18
- HORIMOTO, A. et al. Height and body-mass index trajectories of school-aged children and adolescents from 1985 to 2019 in 200 countries and territories: a pooled analysis of 2181 population-based studies with 65 million participants. 2020. 1
- HOSSEINI, S.-M. et al. Child weight growth trajectory and its determinants in a sample of iranian children from birth until 2 years of age. *International Journal of Preventive Medicine*, Wolters Kluwer–Medknow Publications, v. 5, n. 3, p. 348, 2014. 1, 3
- HOWE, L. D. et al. Linear spline multilevel models for summarising childhood growth trajectories: a guide to their application using examples from five birth cohorts. *Statistical methods in medical research*, SAGE Publications Sage UK: London, England, v. 25, n. 5, p. 1854–1874, 2016. 1, 3, 14, 15, 44, 72
- IDA, A.; ISHIMURA, N.; NAKAMURA, M. Note on the measures of dependence in terms of copulas. *Procedia Economics and Finance*, Elsevier, v. 14, p. 273–279, 2014. 21
- JENSS, R. M.; BAYLEY, N. A mathematical method for studying the growth of a child. *Human Biology*, Johns Hopkins Press, v. 9, n. 4, p. 556, 1937. 9
- JOE, H.; XU, J. J. The estimation method of inference functions for margins for multivariate models. *Technical Report 166*, University of British Columbia, n. 1, 1996. 26
- JOHNSON, W.; BALAKRISHNA, N.; GRIFFITHS, P. L. Modeling physical growth using mixed effects models. *American journal of physical anthropology*, Wiley Online Library, v. 150, n. 1, p. 58–67, 2013. 1
- JUTTE, D. P.; ROOS, L. L.; BROWNELL, M. D. Administrative record linkage as a tool for public health research. *Annual review of public health*, v. 32, n. 1, p. 91–108, 2011. 3, 4
- KOLETZKO, B.; RZEHA, P.; REGNAULT, N. *Analysis of Growth Trajectories*. [S.l.]: Karger, 2014. 2
- LAIRD, N. M.; WARE, J. H. Random-effects models for longitudinal data. *Biometrics*, JSTOR, p. 963–974, 1982. 10
- LIANG, Y.; LU, W.; YING, Z. Joint modeling and analysis of longitudinal data with informative observation times. *Biometrics*, Wiley Online Library, v. 65, n. 2, p. 377–384, 2009. 7

- LIN, H.; SCHARFSTEIN, D. O.; ROSENHECK, R. A. Analysis of longitudinal data with irregular, outcome-dependent follow-up. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, Wiley Online Library, v. 66, n. 3, p. 791–813, 2004. 7
- LITTELL, R. C.; PENDERGAST, J.; NATARAJAN, R. Modelling covariance structure in the analysis of repeated measures data. *Statistics in medicine*, Wiley Online Library, v. 19, n. 13, p. 1793–1819, 2000. 10, 11
- LOKKU, A. et al. Summarizing the extent of visit irregularity in longitudinal data. *BMC Medical Research Methodology*, Springer, v. 20, p. 1–9, 2020. 7
- LU, Y.; PEARCE, A.; LI, L. Weight gain in early years and subsequent body mass index trajectories across birth weight groups: a prospective longitudinal study. *European journal of public health*, Oxford University Press, v. 30, n. 2, p. 316–322, 2020. 6
- MANYEH, A. K. et al. Socioeconomic and demographic determinants of birth weight in southern rural ghana: evidence from dodowa health and demographic surveillance system. *BMC pregnancy and childbirth*, BioMed Central, v. 16, n. 1, p. 1–9, 2016. 2
- MCNEISH, D.; MATTA, T. Differentiating between mixed-effects and latent-curve approaches to growth modeling. *Behavior research methods*, Springer, v. 50, n. 4, p. 1398–1414, 2018. 2, 9
- MD. Ministério da Saúde. Secretaria de Políticas Públicas. Saúde da Criança - Acompanhamento do crescimento e desenvolvimento infantil. 2002. Disponível em: <https://bvsmms.saude.gov.br/bvs/publicacoes/acompanhamento_crescimento_desenvolvimento_infantil_cab11.pdf> ,acesso em 22/11/2021. 35, 44
- MDS. Ministério do Desenvolvimento Social. Cadastro Único: O que é e para que serve. 2018. Disponível em: <<http://mds.gov.br/assuntos/cadastro-unico/o-que-e-e-para-que-serve>> , acesso em 09/04/2021. 29
- NELSEN, R. B. *An introduction to copulas*. [S.l.]: Springer Science & Business Media, 2007. 19, 20, 21, 22, 23, 73
- NOBRE, J. S.; SINGER, J. da M. Residual analysis for linear mixed models. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, Wiley Online Library, v. 49, n. 6, p. 863–875, 2007. 18, 19
- OHUMA, E. O. et al. A novel development indicator based on population-average height trajectories of children aged 0–5 years modelled using 145 surveys in 64 countries, 2000–2018. *BMJ global health*, BMJ Specialist Journals, v. 6, n. 3, p. 1–11, 2021. 16, 47, 72, 73
- PINHEIRO, J. et al. Package ‘nlme’. *Linear and nonlinear mixed effects models, version*, v. 3, n. 1, 2017. 13, 32, 42
- PINHEIRO, J. C.; BATES, D. M. Linear mixed-effects models: basic concepts and examples. *Mixed-effects models in S and S-Plus*, Springer, p. 3–56, 2000. 9
- PULLENAYEGUM, E. M.; LIM, L. S. Longitudinal data subject to irregular observation: A review of methods with a focus on visit processes, assumptions, and study design. *Statistical methods in medical research*, SAGE Publications Sage UK: London, England, v. 25, n. 6, p. 2992–3014, 2016. 8

- REHFELD, K. et al. Comparison of correlation analysis techniques for irregularly sampled time series. *Nonlinear Processes in Geophysics*, Copernicus GmbH, v. 18, n. 3, p. 389–404, 2011. 8
- RIDDELL, C. A. et al. Classifying gestational weight gain trajectories using the sitar growth model. *Paediatric and perinatal epidemiology*, Wiley Online Library, v. 31, n. 2, p. 116–125, 2017. 47
- RIDDELL, C. A. et al. Classifying gestational weight gain trajectories using the sitar growth model. *Paediatric and perinatal epidemiology*, Wiley Online Library, v. 31, n. 2, p. 116–125, 2017. 73
- SILVERWOOD, R. J. *Issues in modelling growth data within a life course framework*. Tese (Doutorado) — London School of Hygiene and Tropical Medicine, 2008. 6
- SIMPKIN, A. J. et al. Modelling height in adolescence: a comparison of methods for estimating the age at peak height velocity. *Annals of human biology*, Taylor & Francis, v. 44, n. 8, p. 715–722, 2017. 16
- SINGER, J. M.; ANDRADE, D. d. Análise de dados longitudinais. *Simpósio Nacional de Probabilidade e Estatística*, Embrapa São Paulo, v. 7, 1986. 33
- SPYRIDES, M. H. C. et al. Effect of predominant breastfeeding duration on infant growth: Prospective study using nonlinear mixed effect models. *Jornal de pediatria*, SciELO Brasil, v. 84, n. 3, p. 237–243, 2008. 1
- STÜTZLE, W. et al. Shape-invariant modelling of human growth. *Annals of Human Biology*, Taylor & Francis, v. 7, n. 6, p. 507–528, 1980. 16
- SULLIVAN, M. C. et al. Growth trajectories of preterm infants: birth to 12 years. *Journal of Pediatric Health Care*, Elsevier, v. 22, n. 2, p. 83–93, 2008. 1
- SUZUKI, A. K. Modelos de sobrevivência bivariados baseados na cópula fgm: uma abordagem bayesiana. Universidade Federal de São Carlos, 2012. 24
- TILLING, K. et al. Modelling childhood growth using fractional polynomials and linear splines. *Annals of Nutrition and Metabolism*, Karger Publishers, v. 65, n. 2-3, p. 129–138, 2014. 1, 2, 6, 9, 14, 15, 44
- TOWERS, S. Potential fitting biases resulting from grouping data into variable width bins. *Physics Letters B*, Elsevier, v. 735, p. 146–148, 2014. 8
- TU, J. V. et al. Impracticability of informed consent in the registry of the canadian stroke network. *New England Journal of Medicine*, Mass Medical Soc, v. 350, n. 14, p. 1414–1421, 2004. 3
- VERBEKE, G. Linear mixed models for longitudinal data. In: *Linear mixed models in practice*. [S.l.]: Springer, 1997. p. 63–153. 2, 8
- VILLAR, J. et al. International standards for newborn weight, length, and head circumference by gestational age and sex: the newborn cross-sectional study of the intergrowth-21st project. *The Lancet*, Elsevier, v. 384, n. 9946, p. 857–868, 2014. 31
- WADHAWAN, N. *Comparison of different weight growth models in a sample of children from 6 to 15 years*. Tese (Doutorado) — Concordia University, 2018. 16

WEGMAN, E. J.; WRIGHT, I. W. Splines in statistics. *Journal of the American Statistical Association*, Taylor & Francis, v. 78, n. 382, p. 351–365, 1983. 13, 14

WHO. *World Health Organization. WHO Anthro Survey Analyser. World Health Organization*. 2011. Disponível em: <https://www.who.int/childgrowth/software/anthro_pc_manual_v322.pdf>, acesso em 22/11/2021. 44

Apêndices

A Códigos do programa R

```
rm(list=ls(all=TRUE))
```

Bibliotecas necessárias (Required libraries)

```
library(nlme)
library(lme4)
library(splines)
library(brokenstick)
library(sitar)
library(IrregLong)
library(ggplot2)
library(parallel)
library(robustlmm)
library(mvtnorm)
```

```
#-----#
# ----- Gráfico Ábacus -----#
#-----#
```

```
# Figura 3
```

```
abacus.plot(n=100, time = "agemes_med0", id="id", data=atermo0a2MED,
tmin = 0, tmax = 27, xlab.abacus = "idade em meses Mediana",
ylab.abacus = "Crianças",
pch.abacus = 16, col.abacus=gray(0.9))
abline(v=c(1,2,4,6,9,12,18,24))
axis(1, c(0,1,2,4,6,9,12,18,24))
```

```

#-----#
# ----- Modelo Linear Mistos -----#
#-----#

#----- MLM - Variável Peso -----#

mlm0WM<-lme(pesomed0 ~ agemes_med0+gender+bxpeso, random = ~1|id,
data = atermo0a2MED, correlation = corAR1())

mlm1WM<-lme(pesomed0 ~ agemes_med0 +I(agemes_med0^2)+gender+bxpeso,
random = ~1|id, data = atermo0a2MED, correlation = corAR1())

mlm2WM<-lme(pesomed0 ~ agemes_med0 +I(agemes_med0^2)+gender+bxpeso,
random = ~agemes_med0|id, data = atermo0a2MED,
correlation = corAR1())

#-----Análise de Resíduo Modelo Linear Misto-----#

## FIGURA 8

residl<-residdiag.nlme(mlm2WM,limit=2,plotid=5:6)

#----- MLM - Variável Altura -----#

mlm0HM<-lme(altmed0 ~ agemes_med0+gender+bxpeso, random = ~1|id,
data = atermo0a2MED, correlation = corAR1())

mlm1HM<-lme(altmed0 ~ agemes_med0 +I(agemes_med0^2)+gender+bxpeso,
random = ~1|id, data = atermo0a2MED, correlation = corAR1())

mlm2HM<-lme(altmed0 ~ agemes_med0 +I(agemes_med0^2)+gender+bxpeso,
random = ~agemes_med0|id, data = atermo0a2MED,
correlation = corAR1())

#-----Análise de Resíduo Modelo Linear Misto-----#

# FIGURA 9

```

```
resid1<-residdiag.nlme(mlm2HM,limit=2,plotid=5:6)
```

```
#-----#
# ----- Modelo Brokenstick -----#
#-----#
```

```
# ----- SEXO MASCULINO (PAN) -----#
```

```
knots <- c(0, 0.1667, 0.5, 1, 2, 2.2)
{
  iMlmB<-Sys.time()
  yMlmB<-c("pesomed0", "altmed0")
  mod_MB<-mclapply(yMlmB, FUN=function(yMlmB){
    brokenstick(as.formula(paste(yMlmB," ~ ageanos0| id")),
    atermo0a2MMED,
    knots = knots, boundary = c(0,2.2), method = "lmer")
  })
  fMlmB<-Sys.time()
}
```

```
# ----- SEXO FEMININO (PAN) -----#
```

```
{
  iFlmB<-Sys.time()
  yFlmB<-c("pesomed0", "altmed0")
  mod_FB<-mclapply(yFlmB, FUN=function(yFlmB){
    brokenstick(as.formula(paste(yFlmB," ~ ageanos0| id")),
    atermo0a2FMED,
    knots = knots, boundary = c(0,2.2), method = "lmer")
  })
  fFlmB<-Sys.time()
}
```

```
# ----- SEXO MASCULINO (BPN) -----#
```

```
knots <- c(0, 0.1667, 0.5, 1, 2, 2.2)
{
```



```

iMBXlmB<-Sys.time()
yMBXlmB<-c("pesomed0", "altmed0")
mod_MBXB<-mclapply(yMBXlmB, FUN=function(yMBXlmB){
  brokenstick(as.formula(paste(yMBXlmB," ~ ageanos0| id")),
  atermo0a2MMEDBX,
  knots = knots, boundary = c(0,2.2), method = "lmer")
})
fMBXlmB<-Sys.time()
}

# ----- SEXO FEMININO (BPN) -----#

{
iFBXlmB<-Sys.time()
yFBXlmB<-c("pesomed0", "altmed0")
mod_FBXB<-mclapply(yFBXlmB, FUN=function(yFBXlmB){
  brokenstick(as.formula(paste(yFBXlmB," ~ ageanos0| id")),
  atermo0a2FMEDBX,
  knots = knots, boundary = c(0,2.2), method = "lmer")
})
fFBXlmB<-Sys.time()
}

#-----Análise de Resíduo Brokenstick -----#

##  FIGURA 12

fMPeso<-fitted(mod_MB[[1]], atermo0a2MMED)
rMPeso<-resid(mod_MB[[1]], atermo0a2MMED)

fFPeso<-fitted(mod_FB[[1]], atermo0a2FMED)
rFPeso<-resid(mod_FB[[1]], atermo0a2FMED)

fMBXPeso<-fitted(mod_MBXB[[1]], atermo0a2MMEDBX)
rMBXPeso<-resid(mod_MBXB[[1]], atermo0a2MMEDBX)

fFBXPeso<-fitted(mod_FBXB[[1]], atermo0a2FMEDBX)
rFBXPeso<-resid(mod_FBXB[[1]], atermo0a2FMEDBX)

```

```
par(mfrow=c(2,2))
```

```
plot(fMPeso, rMPeso, xlab = "Valores ajustados",
     ylab = "Resíduos condicionais padronizados",
     main="Peso (kg) - Masculino (PAN)",
     ylim = c(-10,10), pch=19)
abline(h=c(-3,3), lty=2, col="blue", lwd=2)
```

```
plot(fFPeso, rFPeso, xlab = "Valores ajustados",
     ylab = "Resíduos condicionais padronizados",
     main="Peso (kg) - Feminino (PAN)",
     ylim = c(-10,10), pch=19)
abline(h=c(-3,3), lty=2, col="blue", lwd=2)
```

```
plot(fMBXPeso, rMBXPeso, xlab = "Valores ajustados",
     ylab = "Resíduos condicionais padronizados",
     main="Peso (kg) - Masculino (BPN)",
     ylim = c(-10,10), pch=19)
abline(h=c(-3,3), lty=2, col="blue", lwd=2)
```

```
plot(fFBXPeso, rFBXPeso, xlab = "Valores ajustados",
     ylab = "Resíduos condicionais padronizados",
     main="Peso (kg) - Feminino (BPN)",
     ylim = c(-10,10), pch=19)
abline(h=c(-3,3), lty=2, col="blue", lwd=2)
```

```
#-----#
# ----- Modelo SITAR -----#
#-----#
```

```
# ----- SEXO MASCULINO (PAN) -----#
```

```
#----- PESO -----#
```

```
m1aWMM0<-sitar(x=agemes_med0, y = pesomed0, id = id,
data = atermo0a2MMED, df=4,
control= nlmeControl(pnlstol = .4), fixed= 'a', random = 'a')
```

```
summary(m1aWMM0)
varexp(m1aWMM0)

m1aWMM<-sitar(x=log(agemes_med0), y = pesomed0, id = id,
data = atermo0a2MMED, df=7,
control= nlmeControl(pnlstol = .4), fixed= 'a', random = 'a')
summary(m1aWMM)
varexp(m1aWMM)

m1acWMM<-sitar(x=log(agemes_med0), y = pesomed0, id = id,
data = atermo0a2MMED, df=8,
control= nlmeControl(pnlstol = .4), fixed= 'a+c', random = 'a+c')
summary(m1acWMM)
varexp(m1acWMM)

#----- ALTURA -----#

m1aHMM<-sitar(x=log(agemes_med0), y = altmed0, id = id,
data = atermo0a2MMED, df=7,
control= nlmeControl(pnlstol = .4), fixed= 'a', random = 'a')
summary(m1aHMM)
varexp(m1aHMM)

m1acHMM<-sitar(x=log(agemes_med0), y = altmed0, id = id,
data = atermo0a2MMED, df=7,
control= nlmeControl(pnlstol = .4), fixed= 'a+c', random = 'a+c')
summary(m1acHMM)
varexp(m1acHMM)

# ----- SEXO FEMININO (PAN) -----#

#----- PESO -----#

m1aWMF<-sitar(x=log(agemes_med0), y = pesomed0, id = id,
data = atermo0a2FMED3, df=4,
control= nlmeControl(pnlstol = .4), fixed= 'a', random = 'a')
summary(m1aWMF)
varexp(m1aWMF)
```

```
m1acWMF<-sitar(x=log(agemes_med0), y = pesomed0, id = id,
data = atermo0a2FMED3, df=6,
control= nlmeControl(pnlstol = .4), fixed= 'a+c', random = 'a+c')
summary(m1acWMF)
varexp(m1acWMF)

#----- ALTURA -----#

m1acHMF<-sitar(x=log(agemes_med0), y = altmed0, id = id,
data = atermo0a2FMED3, df=7,
control= nlmeControl(pnlstol = .4), fixed= 'a+c', random = 'a+c')
summary(m1acHMF)
varexp(m1acHMF)

# ----- SEXO MASCULINO (BPN) -----#

#----- PESO -----#

m1aWMMBX<-sitar(x=log(agemes_med0), y = pesomed0, id = id,
data = atermo0a2MMEDBX, df=5,
control= nlmeControl(pnlstol = .4), fixed= 'a', random = 'a')
summary(m1aWMMBX)
varexp(m1aWMMBX)

m1acWMMBX<-sitar(x=log(agemes_med0), y = pesomed0, id = id,
data = atermo0a2MMEDBX, df=7,
control= nlmeControl(pnlstol = .4), fixed= 'a+c', random = 'a+c')
summary(m1acWMMBX)
varexp(m1acWMMBX)

#----- ALTURA -----#

m1aHMMBX<-sitar(x=log(agemes_med0), y = altmed0, id = id,
data = atermo0a2MMEDBX, df=5,
control= nlmeControl(pnlstol = .4), fixed= 'a', random = 'a')
summary(m1aHMMBX)
varexp(m1aHMMBX)

m1acHMMBX<-sitar(x=log(agemes_med0), y = altmed0, id = id,
```

```

data = atermo0a2MMEDBX, df=10,
control= nlmeControl(pnlstol = .4), fixed= 'a+c', random = 'a+c')
summary(mlacHMMBX)
varexp(mlacHMMBX)

# ----- SEXO FEMININO (BPN) -----#

#----- PESO -----#

mlaWMFBX<-sitar(x=log(agemes_med0), y = pesomed0, id = id,
data = atermo0a2FMEDBX, df=4,
control= nlmeControl(pnlstol = .4), fixed= 'a', random = 'a')
summary(mlaWMFBX)
varexp(mlaWMFBX)

mlacWMFBX<-sitar(x=log(agemes_med0), y = pesomed0, id = id,
data = atermo0a2FMEDBX, df=7,
control= nlmeControl(pnlstol = .4), fixed= 'a+c', random = 'a+c')
summary(mlacWMFBX)
varexp(mlacWMFBX)

#----- ALTURA -----#

mlaHMFBX<-sitar(x=log(agemes_med0), y = altmed0, id = id,
data = atermo0a2FMEDBX, df=3,
control= nlmeControl(pnlstol = .4), fixed= 'a', random = 'a')
summary(mlaHMFBX)
varexp(mlaHMFBX)

mlacHMFBX<-sitar(x=log(agemes_med0), y = altmed0, id = id,
data = atermo0a2FMEDBX, df=8,
control= nlmeControl(pnlstol = .4), fixed= 'a+c', random = 'a+c')
summary(mlacHMFBX)
varexp(mlacHMFBX)

#-----Análise de Resíduo SITAR -----#

```

```
### FIGURA 22
```

```
par(mfrow=c(2,2))

par(mar =c(4,4,2,2)+0.2, cex=0.8)
plot(mlacWMM$fitted, mlacWMM$residuals, xlim = c(0,25),
ylim = c(-10,10), main = "Masculino", cex.main = 1.5,
xlab = " Peso (kg) Ajustado", ylab = "Resíduos padronizados",
cex.axis=1.5, cex.lab=1.5) #grafico residuo modelo
abline(0,0)
abline(h=c(-3,3), lty=2)
```

```
par(mar =c(4,4,2,2)+0.2, cex=0.8)
plot(mlacHMM$fitted, mlacHMM$residuals, xlim = c(40,110),
ylim = c(-20,20), main = "Masculino", cex.main = 1.5,
xlab = "Altura (cm) Ajustada", ylab = "Resíduos padronizados"
cex.axis=1.5, cex.lab=1.5) #grafico residuo modelo
abline(0,0)
abline(h=c(-3,3), lty=2)
```

```
par(mar =c(4,4,2,2)+0.2, cex=0.8)
plot(mlacWMF$fitted, mlacWMF$residuals, xlim = c(0,25),
ylim = c(-10,10), main = "Feminino", cex.main = 1.5,
xlab = " Peso (kg) Ajustado", ylab = "Resíduos padronizados",
cex.axis=1.5, cex.lab=1.5) #grafico residuo modelo
abline(0,0)
abline(h=c(-3,3), lty=2)
```

```
par(mar =c(4,4,2,2)+0.2, cex=0.8)
plot(mlacHMF$fitted, mlacHMF$residuals, xlim = c(40,110),
ylim = c(-20,20), main = "Feminino", cex.main = 1.5,
xlab = "Altura (cm) Ajustada", ylab = "Resíduos padronizados",
cex.axis=1.5, cex.lab=1.5) #grafico residuo modelo
abline(0,0)
abline(h=c(-3,3), lty=2)
```

```
## FIGURA 24
```

```
{
par(mfrow=c(1,1))
par(mar =c(4,4,2,2)+0.1, cex=0.8)
qqnorm(mlacWMM$residuals, ylim = c(-10,10),
main = "Peso (kg) - Masculino",
xlab = "Quantis teÃ³ricos", ylab = "Resíduos padronizados",
cex.axis=1.5,
cex.lab=1.5, cex.main = 1.5)
qqline(mlacWMM$residuals)
}
```

```
{
par(mfrow=c(1,1))
par(mar =c(4,4,2,2)+0.1, cex=0.8)
qqnorm(mlacHMM$residuals, ylim = c(-20,20),
main = "Altura (cm) - Masculino",
xlab = "Quantis teÃ³ricos", ylab = "Resíduos Padronizados",
cex.axis=1.5,
cex.lab=1.5,cex.main = 1.5)
qqline(mlacHMM$residuals)
}
```

```
{
par(mfrow=c(1,1))
par(mar =c(4,4,2,2)+0.1, cex=0.8)
qqnorm(mlacWMF$residuals, ylim = c(-10,10),
main = "Peso (kg) - Feminino",
xlab = "Quantis teÃ³ricos", ylab = "Resíduos Padronizados",
cex.axis=1.5,
cex.lab=1.5,cex.main = 1.5)
qqline(mlacWMF$residuals)
}
```

```
{
par(mfrow=c(1,1))
par(mar =c(4,4,2,2)+0.1, cex=0.8)
qqnorm(mlacHMF$residuals, ylim = c(-20,20),
main = "Altura (cm) - Feminino",
xlab = "Quantis teÃ³ricos", ylab = "Resíduos Padronizados",
```

```

cex.axis=1.5,
cex.lab=1.5,cex.main = 1.5)
qqline(mlacHMF$residuals)
}

#-----#
# -----  Análise Bivariada -----#
#-----#

# -----  Seleção da amostra -----#

cod<-unique(atermo0a2MMED$id)
print(Ncri<-length(cod))
dim(atermo0a2MMED)

set.seed(12124)
prop<-0.1
sorteio<-sample(cod,round(Ncri*prop))
length(sorteio)
dataM10<-atermo0a2MMED[atermo0a2MMED$id %in% sorteio,]
head(dataM10, 3)
dim(dataM10)

# -----  Modelo Bronkenstick univariado ----- ##

knots <- c(0, 0.1667, 0.5, 1, 2)
boundary_knots <- c(0, 2.2)

xMM10<-bs(dataM10$ageanos_med0, knots = knots,
Boundary.knots = boundary_knots, degree = 1)
colnames(xMM10)<-paste("ageanos0", c(knots, 2.2), sep = "_")
dataMM10<-cbind(dataM10[, c("id", "ageanos0", "altmed0", "pesomed0")]
, xMM10)
head(dataMM10)
names(dataMM10)

bkP10<- lme(pesomed0 ~ 0+ageanos0_0+ageanos0_0.1667+ageanos0_0.5+
ageanos0_1+ageanos0_2+ageanos0_2.2, random= ~1 | id,
data = dataMM10, correlation = corAR1())

```



```
bkc10<- lme(altmed0 ~ 0+ageanos0_0+ageanos0_0.1667+ageanos0_0.5+
ageanos0_1+ageanos0_2+ageanos0_2.2, random= ~1 | id,
data = dataMM10)
```

```
#### Estimação Cópula ####
```

```
#### Banco de dados ####
```

```
df <- dataMM10
```

```
#### Definindo Funções ####
```

```
acumMLM <- function(x, codgroup, fit){
```

```
# x <- df$pesomed
```

```
# codgroup <- df$id
```

```
# fit <- bkP10
```

```
u <- numeric(length(unique(codgroup)))
```

```
fittedValues <- fit$fitted[,2]
```

```
count <- 1
```

```
#O loop itera por cada código de criança e calcula as probabilidades
acumuladas
```

```
for(i in unique(codgroup)){
```

```
# i <- unique(codgroup)[2]
```

```
#Obtendo os índices de cada criança
```

```
index <- which(codgroup == i)
```

```
#Obtendo matrizes de covariância a partir de modelos multinivel
```

```
sigmaIter <- getVarCov(fit,type = "marginal",individual = count)[[1]]
```

```
#Calculando probabilidades acumuladas com base numa normal
```

```
multivariada por grupo
```

```
u[count] <- pmvnorm(upper = x[index], mean = fittedValues[index],
sigma = sigmaIter)
```

```
#Incremento do contador (para obter a matriz)
```

```
count <- count+1
```

```
}
```

```
return(u)
```

```
}
```

```
densMLM <- function(x, codgroup, fit){
  out <- numeric(length(unique(codgroup)))
  fittedValues <- fit$fitted[,2]
  count <- 1

  #O loop itera por cada código de criança e calcula as probabilidades
  acumuladas
  for(i in unique(codgroup)){
    #Obtendo os índices de cada criança
    index <- which(codgroup == i)
    #Obtendo matrizes de covariância a partir de modelos multinível
    sigmaIter <- getVarCov(fit,type = "marginal",individual = count)[[1]]
    #Calculando probabilidades acumuladas com base numa normal
    multivariada por grupo
    out[count] <- dmvnorm(x[index], fittedValues[index],
      sigma = sigmaIter)
    #Incremento do contador (para obter a matriz)
    count <- count+1
  }
  return(out)
}

fGaus <- function(u, v, rho){
  #Calculando o número de observações
  n <- length(u)
  #Calculando a matriz de dados para calcular densidades
  dataMatrix <- matrix(c(qnorm(u), qnorm(v)), ncol = 2, nrow = n)
  #Definindo a matriz de correlação da multivariada (rho)
  corMatrix <- matrix(c(1, rho, rho, 1), 2, 2)
  #Calculando as densidades de cada par
  out <- dmvnorm(dataMatrix, rep(0, 2), sigma =
    corMatrix)/(dnorm(qnorm(u))*dnorm(qnorm(v)))
  return(out)
}

#Densidade da cópula t
fT <- function(u, v, rho, ni){
```

```

#Calculando o número de observações
n <- length(u)
#Calculando a matriz de dados para calcular densidades
dataMatrix <- matrix(c(qt(u, ni), qt(v, ni)), ncol = 2, nrow = n)
#Definindo a matriz de correlação da multivariada (rho)
corMatrix <- matrix(c(1, rho, rho, 1), 2, 2)
#Calculando as densidades de cada par
out <- dmvtnorm(dataMatrix, df = ni, sigma = corMatrix,
log=FALSE)/(dt(qt(u, ni), ni)*dt(qt(v, ni), ni))

return(out)
}

#Densidade da cópula Frank
fFrank<- function(u, v, rho){
out <- (rho*(1-exp(-rho))*exp(-rho*(u+v)))/(((1-exp(-rho)) -
(1-exp(-rho*u))*(1-exp(-rho*v)))^2)
return(out)
}

#Densidade da cópula Clayton
fClayton<- function(u, v, rho){
out <- (1+rho)*((u*v)^((-1)*(rho +
1)))*((u^(-rho)+v^(-rho)-1)^((-1)*(2*rho + 1)/(rho)))
return(out)
}

#Densidade da cópula Gumbel
fGumbel<- function(u, v, rho){
out <- ((u*v)^(-1))*((log(u)*log(v))^(rho-1))*((( (-log(u))^rho)+
((-log(v))^rho))^(1/rho-2))*exp(-((( (-log(u))^rho)+((-log(v))^rho))^(
1/rho))*((rho-1)+((( (-log(u))^rho)+((-log(v))^rho))^(1/rho)))
return(out)
}

l <- function(u, v, uDens, vDens, params, type = "gauss", ni = NULL){
#Localizando parâmetros
rho <- params
#Calculando verossimilhança

```

```
if(type == "gauss"){
  L <- fGaus(u, v, rho)*uDens*vDens
} else if (type == "t"){
  L <- fT(u, v, rho, ni)*uDens*vDens
} else if (type == "clayton"){
  L <- fClayton(u, v, rho)*uDens*vDens
} else if (type == "frank"){
  L <- fFrank(u, v, rho)*uDens*vDens
} else if (type == "gumbel"){
  L <- fGumbel(u, v, rho)*uDens*vDens
}

#Calculando log-verossimilhança
out <- sum(log(L))
return(out)
}

estMLMCop <- function(x, y, codgroup, fitX, fitY, start, type,
  lower = -.99, upper = .99, ...){
  require(mvtnorm)

  # x = df$pesomed0; y = df$almed0; codgroup = df$id; fitX = bkP10;
  fitY = bkC10; start = 3; type = "frank"; lower = -.99; upper = .99

  #Calculando uniformes (marginais da cÃ³pula)
  u <- acumMLM(x, codgroup, fitX)
  v <- acumMLM(y, codgroup, fitY)

  #Calculando densidades
  uDens <- densMLM(x, codgroup, fitX)
  vDens <- densMLM(y, codgroup, fitY)

  #Otimizando o parâmetro da cópula
  out <- optim(start, l, u = u, v = v, type = type, ...,
    uDens = uDens, vDens = vDens,
    method = "L-BFGS-B", lower = lower, upper = upper,
    control = list(maxit=100000, fnscale = -1))

  return(out)
```

```
}
```

```
#### Iteração completa ####
```

```
estimateMLMCopClay <- estMLMCop(df$pesomed0, df$altmed0, df$id,  
bkP10, bkC10, start = 1, type = "clayton",  
lower = .01, upper = 100)  
estimateMLMCopClay$par
```

```
estimateMLMCopGumb <- estMLMCop(df$pesomed0, df$altmed0, df$id,  
bkP10, bkC10, start = 1, type = "gumbel",  
lower = 1.01, upper = 100)  
estimateMLMCopGumb$par
```

```
estimateMLMCopFrank <- estMLMCop(df$pesomed0, df$altmed0, df$id,  
bkP10, bkC10, start = 1, type = "frank",  
lower = .01, upper = 100)  
estimateMLMCopFrank$par
```

```
estimateMLMCopT <- estMLMCop(df$pesomed0, df$altmed0, df$id,  
bkP10, bkC10, start = 0, type = "t", ni = 2,  
lower = -.99, upper = .99)  
estimateMLMCopT$par
```

```
estimateMLMCopGauss <- estMLMCop(df$pesomed0, df$altmed0, df$id,  
bkP10, bkC10, start = 0, type = "gauss",  
lower = -.99, upper = .99)  
estimateMLMCopGauss$par
```