Universidade Federal de São Carlos
Centro de Ciências Exatas e de Tecnologia
Departamento de Estatística

Introdução à Analise de Dados Longitudinais

Douglas de Paula Nestlehner

Agosto, 2022

Sumário

Capítulo 1

Problema Apresentado

Um dos maiores problemas relacionados a maternidade trata-se de recém nascidos que apresentam um peso abaixo do esperado, consequência de complicações na gravides, doenças, e outros fatores que podem afetar o peso da criança recém nascida tornando-a mais frágil e com nescidade de cuidados especiais.

O objetivo principal desse estudo será verificar o comportamento do peso de nascidos vivos da mesma mãe, verificando se existe uma possível relação entre as variáveis (peso nascimento e ordem gravidez), para posteriormente pode ajustar um modelo que realize a predição do peso nascimento.

1.1 Base Dados

A base de dados que sera utilizada trata-se dados de um estudo do Centro de Controle de Doenças dos EUA, realizado no estado da Geórgia de 1980 a 1992, dados vinculados em nascidos vivos da mesma mãe (Adams et al., 1997).

Esta base contém um subconjunto de dados restrito a 878 mães para as quais foram identificados cinco partos.

As principais variáveis de interesse são ordem gravidez (ordem em que a mãe teve o filho: 1^{o}, 2^{o}, ..., 5^{o}) e o peso do bebê ao nascer. Observe que, como cada mãe teve cinco nascimentos infantis, mas em idades maternas diferentes, estamos trabalhando com dados longitudinais.

Lista de Variáveis:

- ID da Mãe: variável identificadora da mãe no estudo (categórica);
- Ordem de Nascimento: variável indicativa da ordem de nascimento, podendo ter valores de 1 a 5;
- Peso de Nascimento: variável numérica do peso em gramas;
- Idade Materna: variável numérica da idade em anos;
- ID do Filho: variável identificadora do filho no estudo (categórica).

Na Tabela 1.4 estão representadas algumas observações da base de dados que iremos utilizar.

Tabela 1.1: Base de dados

	ID_Mãe	Ordem_Nascimento	Peso_Nascimento	Idade_Materna	ID_Filho
1	80	1	3175	18	1
2	80	2	3572	21	2
3	80	3	3317	24	3
4	80	4	4281	26	4
5	80	5	3827	28	5
6	84	1	2892	14	6
⋮	⋮	⋮	⋮	⋮	⋮
4390	370377	5	3487	29	4390

Apenas no intuito de verificar algum tipo de anomalia, calculamos algumas medidas descritivas da variável de interesse Peso_Nascimento, representada na Tabela 1.2.

Tabela 1.2: Base de dados

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
312	2850	3175	3156	3515	5528

Possivelmente teremos alguns outliers, porém esse tipo de evento podem ter acontecido no estudo, seguiremos então considerando todas as observações da base.

1.2 Analise Exploratória

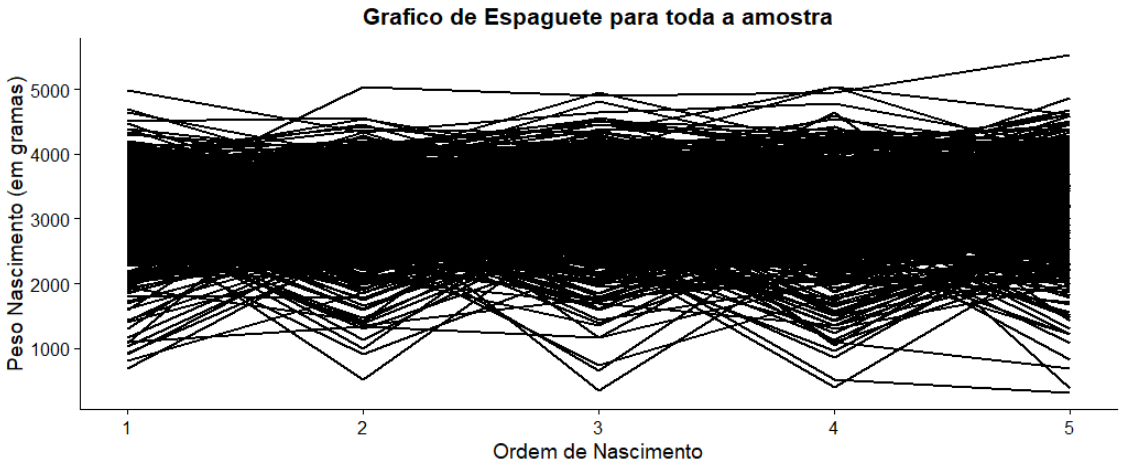
Nessa seção iremos realizar uma analise exploratória, no intuito de conhecer melhor os dados e verificar algum tipo de comportamento.

1.2.1 Gráfico Espaguete

Como dito anteriormente, estamos analisando dados longitudinais, e uma das melhores maneiras de observar o comportamento desse tipo de dados é utilizando o gráfico de espaguete.

Na Figura 1.1, temos o gráfico de espaguete do peso de nascimento em relação a ordem de nascimento.

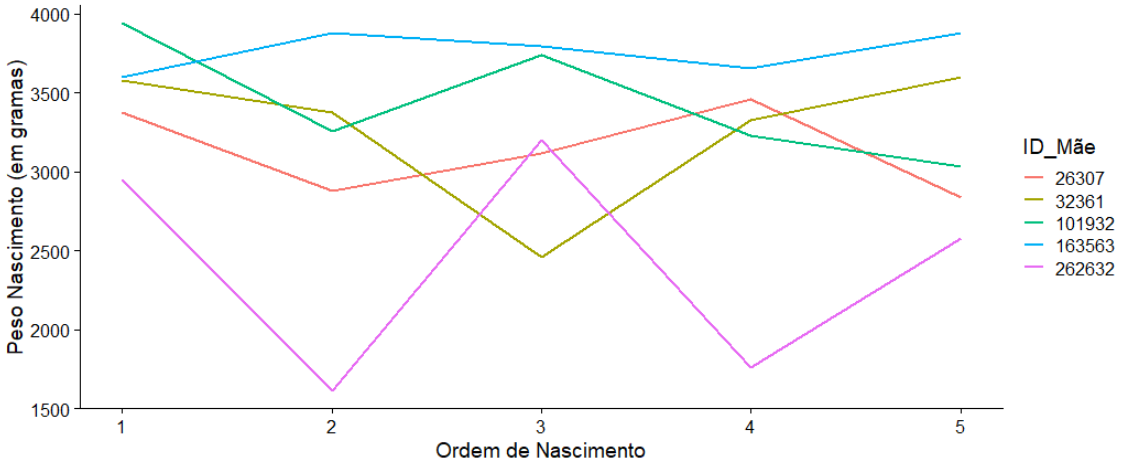
Figura 1.1



Como temos uma amostra relativamente grande, a visualização e interpretação do gráfico de espaguete para toda amostra se torna inviável.

Podemos então, realizar um sorteio dentro da amostra e observar o comportamento de apenas algumas observações. Na Figura 1.2 realizamos um sorteio aleatório de 5 mães, e observamos o peso de nascimento de seus filhos em suas respectivas ordem de nascimento.

Figura 1.2: Gráfico de Espaguete de 5 mãe sorteadas

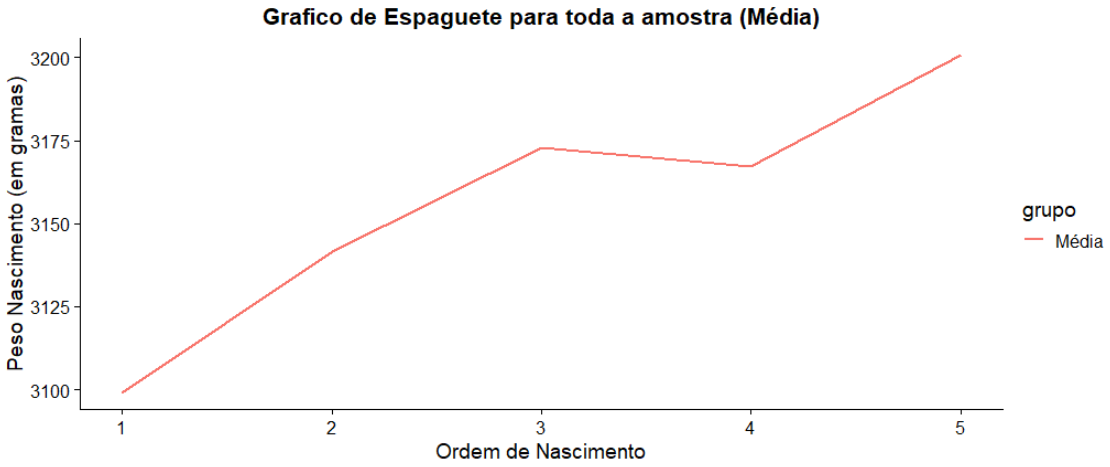


Mesmo com o sorteio de algumas observações, ainda fica dificil a identificação de algum tipo de padrão ou comportamento, pois em alguns casos a variabilidade é alta, outros baixa, alguns decrescem e outros não.

1.2.2 Espaguete Média

Uma alternativa bastante interessante para observar o comportamento ao longo do tempo do peso de nascimento, é calculando a média de peso em cada tempo (ordem de nascimento), representado na Figura 1.4.

Figura 1.3: Gráfico de Espaguete para a média

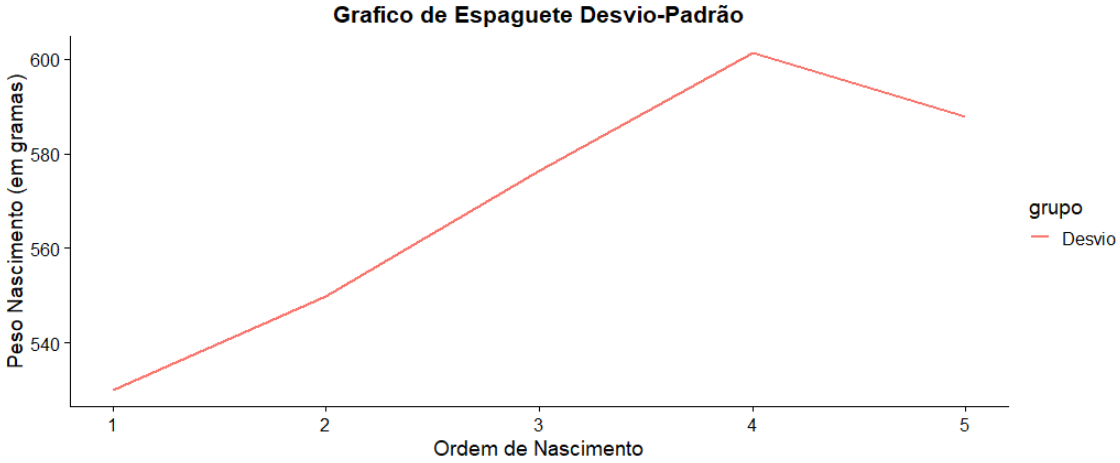


Agora podemos observar que em média existe um crescimento no peso de nascimento, entretanto essa escala é bem pequena (3100g para 3200g) e a amostra é grande, devemos então verificar a variabilidade para poder ver se essa afirmação faz sentido.

1.2.3 Espaguete Desvio Padrão

Podemos calcular o desvio padrão para cada ordem de nascimento, e também plotar em um gráfico de espaguete para ter uma visualização mais clara.

Figura 1.4: Gráfico de Espaguete para o desvio padrão



Observamos um comportamento parecido com o da média, o desvio aumenta ao longo do tempo, com desvios variando de aproximadamente 500g a 650g (valores bem elevados).

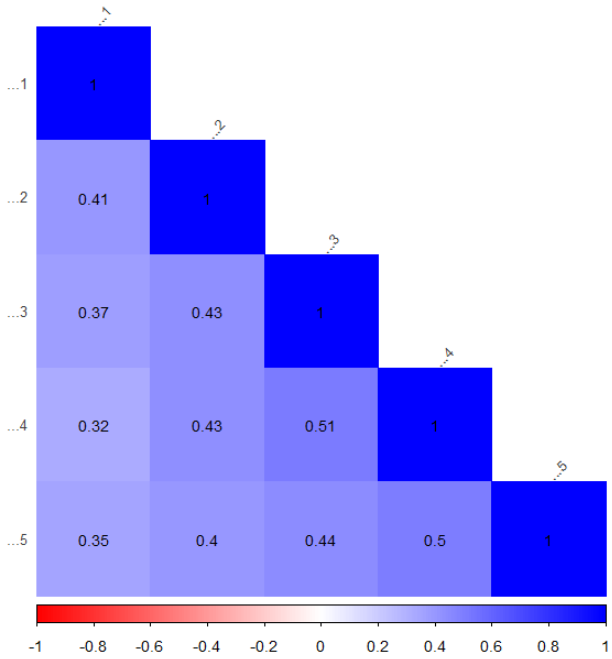
1.2.4 Dependência Temporal

No intuito de verificar a dependência das medidas dos elementos nos instantes de tempos (ordem de nascimento), calculamos a matriz de correlações de Pearson, representada na Tabela 1.3 e Figura 1.5.

Tabela 1.3: Matriz de correlações de Pearson

	...1	...2	...3	...4	...5
...1	1.00	0.41	0.37	0.32	0.35
...2	0.41	1.00	0.43	0.43	0.40
...3	0.37	0.43	1.00	0.51	0.44
...4	0.32	0.43	0.51	1.00	0.50
...5	0.35	0.40	0.44	0.50	1.00

Figura 1.5: Matriz de correlações de Pearson



Aparentemente, as correlações apresentam um estrutura autoregressiva, ou seja, as correlações diminuem ao longo do tempo (cada vez menos existe uma dependência entre elas)

Obs: A estrutura autorregressiva pode ser contestada, pois a correlação nos tempo 4 e 1 são menores do que no tempo 5 e 1, porem são bem próximas, e as únicas que apresentam comportamento distinto.

1.3 Modelo de Regressão com Erros Independentes

Apenas no intuito de observar os resultados obtidos ao se ajustar um modelo de regressão usual, porem para dados longitudinais, realizamos o ajuste do seguinte modelo:

$$\text{Peso\_Nascimento} = \beta_0 + \beta_1 * \text{Ordem\_Nascimento\_2}$$
$$+\beta_2 * \text{Ordem\_Nascimento\_3} + \beta_3 * \text{Ordem\_Nascimento\_4}$$
$$+\beta_4 * \text{Ordem\_Nascimento\_5} + \beta_5 * \text{Idade\_Materna}$$

$$Y_i = X_i\beta \tag{1.1}$$

Em que Ordem_Nascimento_i representa a variável indicadora da i-ésima ordem em que ocorreu o nascimento. Por exemplo, caso a ordem de nascimento da observação seja a primeira, o modelo sera dado por:

$$\text{Peso\_Nascimento} = \beta_0 + \beta_1 * 0 + \beta_2 * 0 + \beta_3 * 0 + \beta_4 * 0 + \beta_5 * \text{Idade\_Materna}$$

Desse modo, as estimativas encontradas para o modelo estão representadas na Tabela 1.4:

Tabela 1.4: Coeficientes estimados

—	Estimate Std.	Error	z-value	$Pr(>\|z\|)$
(Intercept)	2579.942	43.892	58.779	$< 2e-16$
Ordem_Nascimento_2	-14.451	27.028	-0.535	0.592917
Ordem_Nascimento_3	-36.331	27.971	-1.299	0.194053
Ordem_Nascimento_4	-95.670	29.466	-3.247	0.001176
Ordem_Nascimento_5	-115.983	31.428	-3.690	0.000227
Idade_Materna	29.046	2.218	13.097	$< 2e-16$

$$\text{Adjusted R-squared} = 0.04003$$

$$\text{Residual standard error} = 558.9$$

$$\text{p-value} < 2.2e-16$$

Já podemos observar que o modelo não aparenta ser adequado, para realmente poder verificar, realizamos a analise de diagnostico.

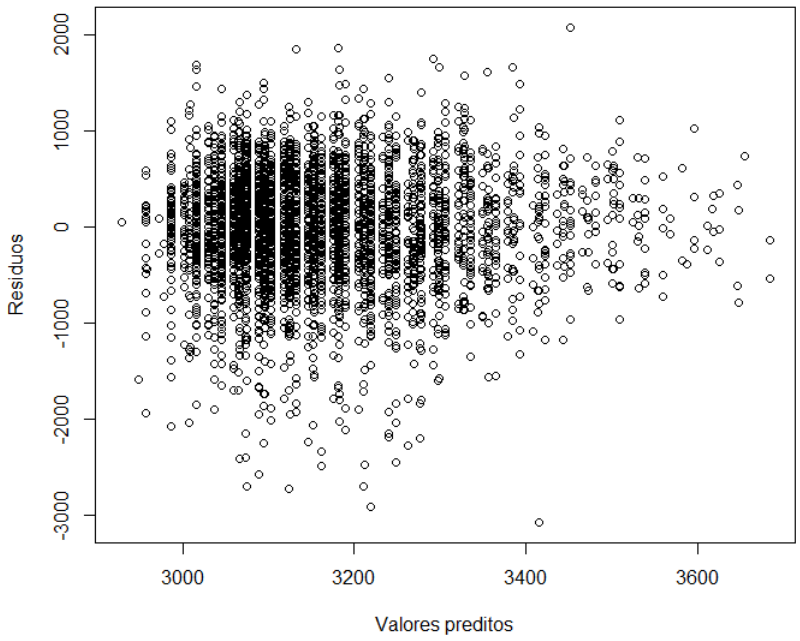
1.3.1 Analise de diagnostico

Quando ajustamos um modelo de regressão linear normal impomos uma séries de condições (linearidade, independência, normalidade, homocedasticidade, entre outros). A fim de verificar se essas condições iniciais foram de fato satisfeitas, fizemos a análise de resíduos, a qual estuda o comportamento dos dados observados através dos resíduos (erros) do modelo ajustado.

Para poder verificar independência devemos observar a ordem de coleta dos dados, temos que as observações são tomadas ao longo do tempo, ou seja, elas são dependentes, logo a suposição de independência já não é atendida. Entretanto, para fins didáticos, iremos admitir/supor que ela é satisfeita.

No intuito de checar a homocedasticidade, plotamos o gráfico de resíduos x valores preditos, representado na Figura 1.6

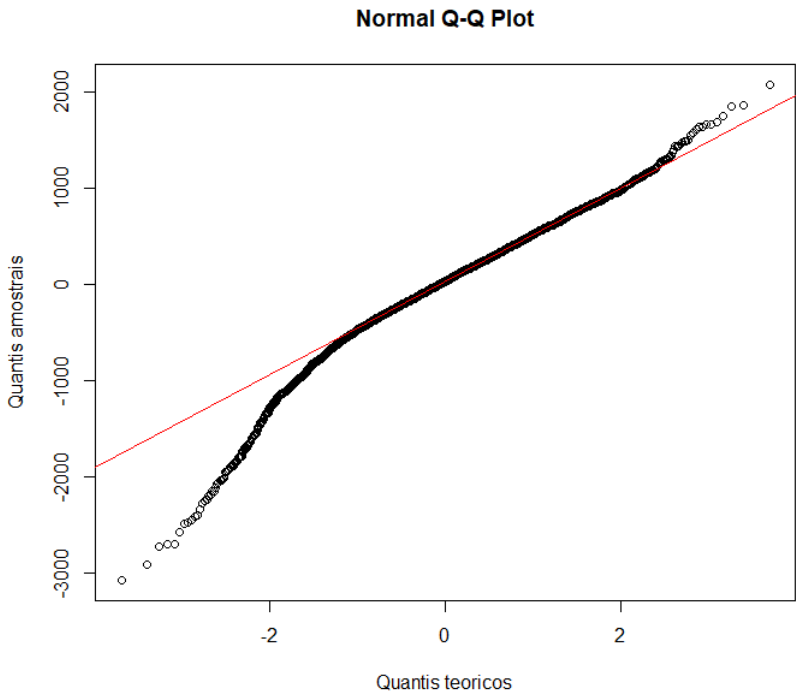
Figura 1.6: Gráfico resíduos x valores preditos



Apesar dos pontos estarem dispersos, existe muitas observações extremamente longes de 0, o que é um forte indicativo de que a homocedasticidade não será atendida.

Afim de verificar a normalidade, plotamos o gráfico qqnorm, representado na figura 1.7

Figura 1.7: Gráfico resíduos x valores preditos



Nele observamos muitos pontos distantes da reta, indicativo de que a normalidade não é satisfeita, e presença de outliers. Para poder confirmar a não normalidade, realizamos os testes de shapiro wilk e liliefors, obtendo em ambos que a normalidade não é satisfeita.

Concluindo então que as condições necessárias para o ajuste de um modelo de regressão linear normal, não são satisfeitas. Poderíamos tentar realizar transformações na variável resposta no intuito de que as suposições fossem satisfeitas, entretanto a amostra em analise é muito grande, e dificilmente sera adequada para esse tipo de modelo.

1.3.2 Interpretação

Temos na seção anterior, que a analise de diagnostico falha em quase todas as suposições, porém assumindo que ela fosse satisfeita o modelo final que teríamos seria:

$$\text{Peso\_Nascimento} = 2579.942 - 14.451 * \text{Ordem\_Nascimento\_2}$$

$$-36.331 * \text{Ordem\_Nascimento\_3} - 95.670 * \text{Ordem\_Nascimento\_4}$$

$$-115.983 * \text{Ordem\_Nascimento\_5} + 29.046 * \text{Idade\_Materna}$$

Em que podemos observar que o Peso_Nascimento tem uma influencia negativa conforme a ordem de nascimento (quanto maior for a ordem de nascimento menor o peso) e uma contribuição positiva da variável Idade_Materna. O que pode acabar se equivalendo, pois quanto maior a ordem de nascimento, maior será a Idade_Materna.

1.4 Modelo Linear Misto

Nessa seção iremos ajustar modelos lineares mistos, buscando uma estrutura de covariância adequada aos dados para assim encontrar um preditor linear para a esperança.

Lembrando que o modelo linear misto é uma generalização do modelo linear padrão, sendo uma generalização em que os dados podem exibir correlação e variabilidade não contante, ou seja, com o modelo linear misto estamos modelando não apenas a média dos dados, e sim modelando a média, a variâncias e covariâncias.

1.4.1 Estrutura de Covariância

Para poder definir qual a estrutura de covariância mais adequada para os nossos dados, ajustamos o modelo 1.2 considerando diferentes tipos de estruturas de covariância.

$$Y_i = X_i\beta + Z_i b_i + e_i \tag{1.2}$$

Em que:

- Y_i: Peso Nascimento (cada observação);
- $X_i\beta$: Efeitos fixos (coeficientes de regressão apresentados em 1.1);
- b_i: São os coeficientes de efeito aleatório que são assumidos como multivariados normalmente distribuídos;
- Z_i: São as variáveis de efeito aleatório (preditores).
- $b_i \sim N(0, D)$; $e_i \sim N(0, \Sigma_i)$

Para poder est

Para realizar a estimação do modelo, utilizamos o método da Máxima Verossimilhança Restrita (REML), na Tabela 1.5 temos o AIC e BIC estimado para cada modelo.

Tabela 1.5: Medidas AIC e BIC estimadas para os modelos ajustados.

Estrutura de Covariância	AIC	BIC
Não estruturado	NA	NA
Toeplitz	NA	NA
Simetria Composta	67022.4	67036.7
AR(1)	67018.9	67033.3
Componentes de Variação (VC)	67022.4	67036.7

Para as estruturas Toeplitz e Não estruturado, o ajuste do modelo não convergiu.

Observamos que o modelo ajustado considerando a estrutura de covariância AR(1) foi o que apresentou menor AIC e BIC em relação aos demais (apesar de um valores bem próximos), desse modo, definimos que a estrutura de covariância para dar sequencia aos ajustes sera a AR(1).

1.4.2 Preditor Linear

Com a estrutura de covariância definida, utilizando o método da Máxima Verossimilhança Restrita (ML) para estimar o preditor linear. A seguir temos os resultados obtidos.

Na Figura 1.8 temos representado as informações gerais utilizadas no ajuste do modelo.

The Mixed Procedure

Model Information	
Data Set	WORK.PESONASC2
Dependent Variable	Peso_Nascimento
Covariance Structure	Autoregressive
Subject Effect	ID_Mãe
Estimation Method	ML
Residual Variance Method	Profile
Fixed Effects SE Method	Model-Based
Degrees of Freedom Method	Containment

Figura 1.8

Na Figura 1.9 temos representado as informações de dimensões das matrizes relevantes para o ajuste do modelo.

Dimensions	
Covariance Parameters	3
Columns in X	3
Columns in Z per Subject	1
Subjects	878
Max Obs per Subject	5

Figura 1.9

Na Figura 1.10 temos representado as informações sobre o tamanho da amostra utilizada no ajuste do modelo.

Number of Observations	
Number of Observations Read	4390
Number of Observations Used	4390
Number of Observations Not Used	0

Figura 1.10

Na Figura 1.11 temos representado as estimativas obtidas do parâmetro de covariância.

Covariance Parameter Estimates		
Cov Parm	Subject	Estimate
Variance	ID_Mãe	123672
AR(1)	ID_Mãe	0
Residual		188553

Figura 1.11

Na Figura 1.12 temos representado uma lista de informações sobre o modelo misto ajustado.

Fit Statistics	
-2 Log Likelihood	67060.7
AIC (Smaller is Better)	67072.7
AICC (Smaller is Better)	67072.7
BIC (Smaller is Better)	67101.3

Figura 1.12

Na Figura 1.13 temos representado as estimativas obtidadas e o teste de significância para os efeitos definidos no modelo (ordem_nascimento e idade). Em que, ao nível de significância de 5% notamos que os efeitos fixos (ordem nascimento e idade materna) são significativos.

Solution for Fixed Effects					
Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	2674.99	55.8792	877	47.87	<.0001
Ordemclass	-25.0870	7.6559	3510	-3.28	0.0011
Idade_Materna	25.7076	3.2646	3510	7.87	<.0001

Figura 1.13

1.4.3 Análise Diagnóstico

Depois de ter selecionado o modelo que aparenta ser o que melhor se ajusta aos dados, realizamos a analise de diagnóstico no intuito de avaliar o ajuste encontrado e suas premissas.

Na Figura 1.14 temos os gráficos obtidos na análise de resíduos de Pearson, com o objetivo de verificar normalidade e homocedasticidade.

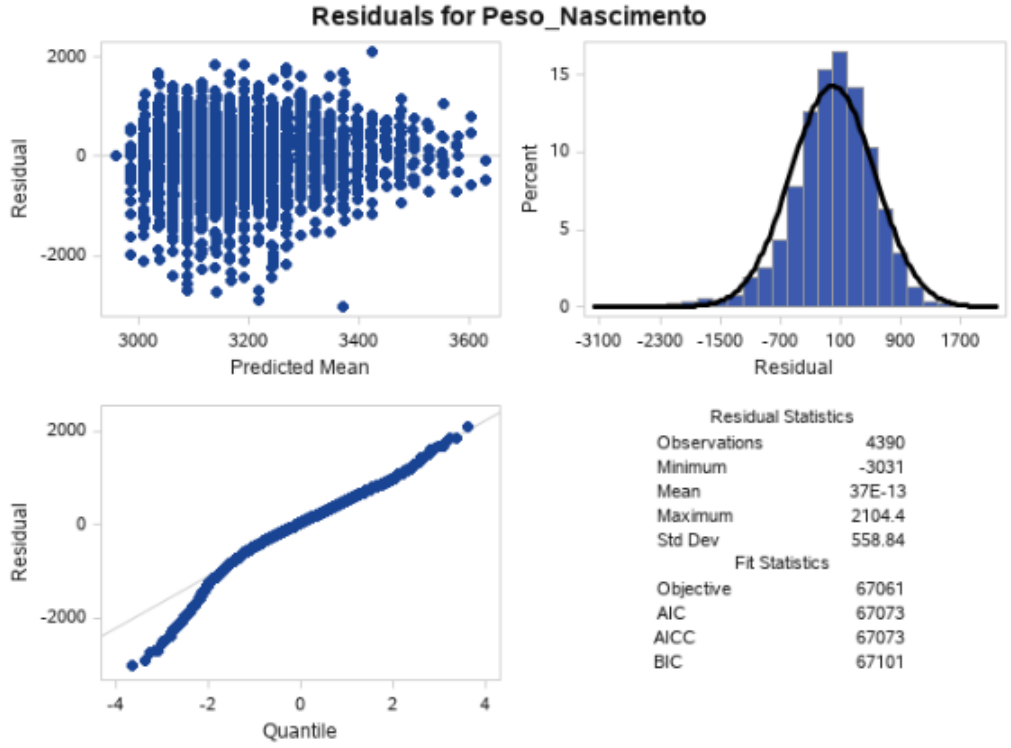


Figura 1.14: Resíduos de Pearson.

No gráfico de residual x predicted, aparentemente os pontos não estão distribuídos de forma aleatória em torno de zero, o que nos indica a não homocedasticidade.

No gráfico da normal, observamos muitos pontos distantes da reta, o que é um forte indicativo de que a normalidade não é satisfeita.

A reprovação na análise de diagnóstico era esperada, pois a amostra de dados era relativamente grande e não seguia distribuição normal.

O mais adequado nessa situação seria ajustar um modelo para dados longitudinais não-normais.

1.4.4 Modelo Marginal

Os modelos obtidos anteriormente foram ajustados no intuito de realizar inferências sobre os indivíduos em especifico (Mães que tiverem 5 filhos), agora iremos ajustar um modelo marginal, o qual tera como intenção trazer inferências sobre toda a população.

As estimativas obtidas estão representadas na Figura 1.15

Solution for Fixed Effects					
Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	2634.42	53.0691	1679	49.64	<.0001
Ordemclass	-27.9324	8.5407	2519	-3.27	0.0011
Idade_Materna	27.9141	2.9933	1503	9.33	<.0001

Figura 1.15: Estimativas para o modelo marginal.

Os gráficos da análise de diagnostico estão representado na Figura 1.16

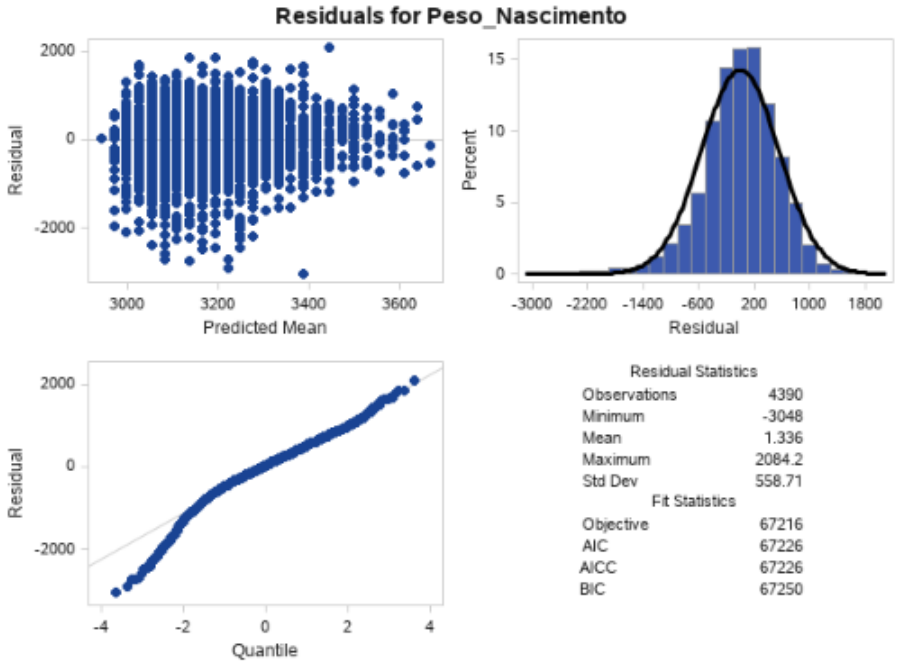


Figura 1.16: Análise de diagnóstico modelo marginal.

Nota-se que os valores obtidos para o modelo marginal são bem semelhantes ao obtido nos modelos anteriores, porém como as aplicações e objetivos dos dois modelos são distintos, a comparação entre eles se torna inadequada.