Universidade Estadual de Maringá
Análise de Regressão

Regressão Linear Múltipla CORRIGIDO

Wesley Oliveira Furriel RA:61493

Professora Dr. Rosangela Gentirana Santa

Maringá
2016

Conteúdo

1 Introdução

O objetivo desse trabalho foi analisar os dados de um estudo transversal sobre baixo peso de crianças ao nascer. As informações que compõe o banco de dados foram coletadas, do recém nascido, e de seus pais, somando 680 casos. O intuito deste estudo foi identificar se algumas das características da mãe e do pai estão associados ao peso do recém nascido. Isto é, o que influencia no peso da criança em seu nascimento.

Para atingir os objetivos desejados, realizou-se uma modelagem por meio da regressão linear múltipla, uma vez que, a variável resposta, era quantitativa e apresentava distribuição normal. Para a construção do modelo final, foram selecionadas as variáveis que nos permitiram explicar da melhor forma possível o fenômeno, segundo algumas medidas que nos permitiram avaliar o ajuste e a fidedignidade do modelo.

O trabalho foi desenvolvido no ambiente SAS 9.4 e as macros utilizadas para a seleção de modelos, construção de gráficos específicos e também, avaliação de pontos influentes foram disponibilizadas nos anexos.

2 Regressão Linear

Antes de iniciar as investigações iremos apresentar de forma breve a regressão linear. Como expõe Paula (2013) na regressão linear simples, nos preocupamos em verificar e modelar a relação entre duas variáveis quantitativas, por exemplo, altura e peso, altitude e temperatura de ebulição da água, entre outros. Para a verificação da existência da relação linear, utilizamos um modelo da forma:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \tag{1}$$

em que
- y é variável resposta
- x é uma variável fixa conhecida, chamada de preditor ou covariável
- β_k são os parâmetros do modelo(coeficientes de regressão) sendo β_0 o intercepto e β_1 a inclinação, que são estimados a partir dos valores observados de xe y.
- ε é um erro aleatório não observado, em que $E[\varepsilon_i] = 0$ e $var[\varepsilon_i] = \sigma^2$. O erro ε_i é assumido com variância constante σ^2.

É preciso ressaltar que o erro não implica em engano ou equívoco na modelagem, mas sim, um termo estatístico que representa flutuações aleatórias, erros de medidas ou efeitos não controlados. Além disso, para realizar a regressão precisamos que o modelo(1) apresente linearidade em seus parâmetros, bem como a distribuição normal do erro e a independência dos valores observados de y. Quanto a regressão múltipla ela é uma extensão da simples, no caso deste trabalho nossa variável resposta pode ser influenciada por mais de uma covariável. Desse modo, deve-se utilizar o modelo de regressão múltipla que pode ter várias variáveis explicativas, ele é dado por:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + \varepsilon_i \tag{2}$$

cada variável preditora incluída tem seu coeficiente de regressão β_k e a resposta é explicada pela combinação linear destes.

Tendo em vista tais aspectos iremos iniciar a análise descritiva dos dados e seguir para construção e verificação da adequação do modelo.

3 Análise preliminar dos dados

O interesse desse trabalho foi verificar os fatores que explicam o peso do recém nascido (y), mais especificamente as características dos pais. Para tal o banco de dados conta com as variáveis preditoras: diâmetro da cabeça(cm); altura ao nascer(cm); peso ao nascer(kg); idade gestacional(semanas); idade da mãe(anos); número de cigarros que a mãe fuma por dia; altura da mãe(m); peso antes da gravidez(kg); idade do pai(anos); escolaridade do pai (anos); número de cigarros que o pai fuma por dia e altura do pai (m). Antes de iniciar a regressão é preciso conhecer as variáveis presentes em nosso banco de dados.
Desse modo, foram geradas medidas descritivas de posição e dispersão, bem como, gráficos que nos ajudarão a compreender o comportamento desses dados.

l m3cm p3cm

Tabela 1: Medidas de posição e dispersão

Variable	Label	Mean	Minimum	Maximum	Std Dev	Coeff of Variation	Skewness	Kurtosis
diacabeca	diâmetro da cabeça (cm)	33.5765588	27.9400000	38.1000000	1.5908395	4.7379470	0.0165888	0.2531777
altura	altura ao nascer (cm)	51.5097059	43.1800000	58.4200000	2.4945386	4.8428515	-0.2086449	0.0829440
pesonasc	peso ao nascer (kg)	0.0034125	0.0014982	0.0051756	0.000495925	14.5326921	-0.0260630	0.4154740
idadegest	idade gestacional (semanas)	39.7705882	29.0000000	48.0000000	1.8754329	4.7156278	-0.2185829	3.0116826
idademae	idade da mãe (anos)	25.8573529	15.0000000	42.0000000	5.4633824	21.1289314	0.6702995	-0.1556946
numcigmae	número de cigarros que fuma (número de cigarros/dia)	7.4308824	0	50.0000000	11.2720248	151.6916069	1.5065591	1.6087987
alturamae	altura da mãe (m)	1.6366191	1.4478000	1.8034000	0.0630742	3.8539301	-0.1194012	-0.0374088
pesomae	peso antes da gravidez (kg)	57.1030147	38.2500000	110.7000000	8.0449452	14.0884772	1.3430761	4.9988677
idadepai	idade do pai (anos)	28.8000000	18.0000000	52.0000000	6.1331327	21.2955996	0.7670672	0.3045250
escpai	escolaridade do pai (anos)	13.3794118	6.0000000	16.0000000	2.2025931	16.4625557	-0.3052998	-0.6211672
numcigpai	número de cigarros que fuma (número de cigarros/dia)	14.4382353	0	50.0000000	14.1702984	98.1442548	0.6332803	-0.4657740
alturapai	altura do pai (m)	1.7937256	1.5748000	2.0066000	0.0670134	3.7359911	-0.0853275	-0.1358455

Na tabela 1 temos as medidas de posição e dispersão das variáveis presentes em nosso banco de dados. Pelo coeficiente de variação verificamos que o número de cigarros que a mãe e o pai fumam, são as variáveis que apresentam a maior variação. As demais se mostram relativamente homogêneas, ou seja, não desviam de modo acentuado do valor médio, fato que indica que a média é um valor representativo dos dados. Quanto a assimetria, foi verificado que o número de cigarros fumados por dia e o peso da mãe, foram as variáveis que se mostraram mais assimétricas.

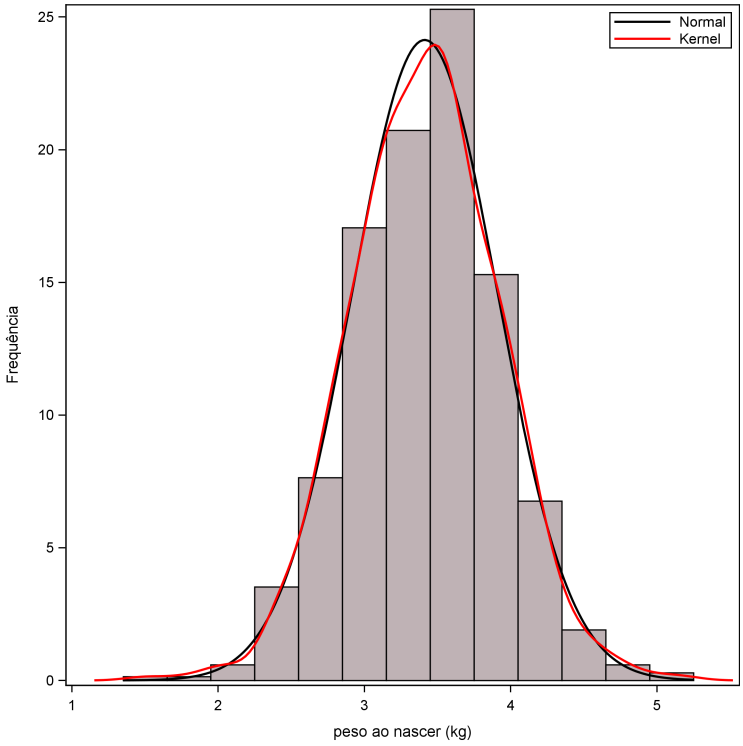

Figura 1: Histograma para a variável resposta

No que tange a distribuição da variável resposta observamos pelo histograma 1 que o peso do recém nascido apresenta uma distribuição aproximadamente normal. Verificando a estimativa de kernel constata-se que ela se ajusta muito bem a curva da normal. Além disso, o valor de Skewness está muito próxima de 1, o que indica ausência de problemas de assimetria. Portanto, não temos qualquer restrição em empregar o peso do bebê ao nascer como variável resposta.

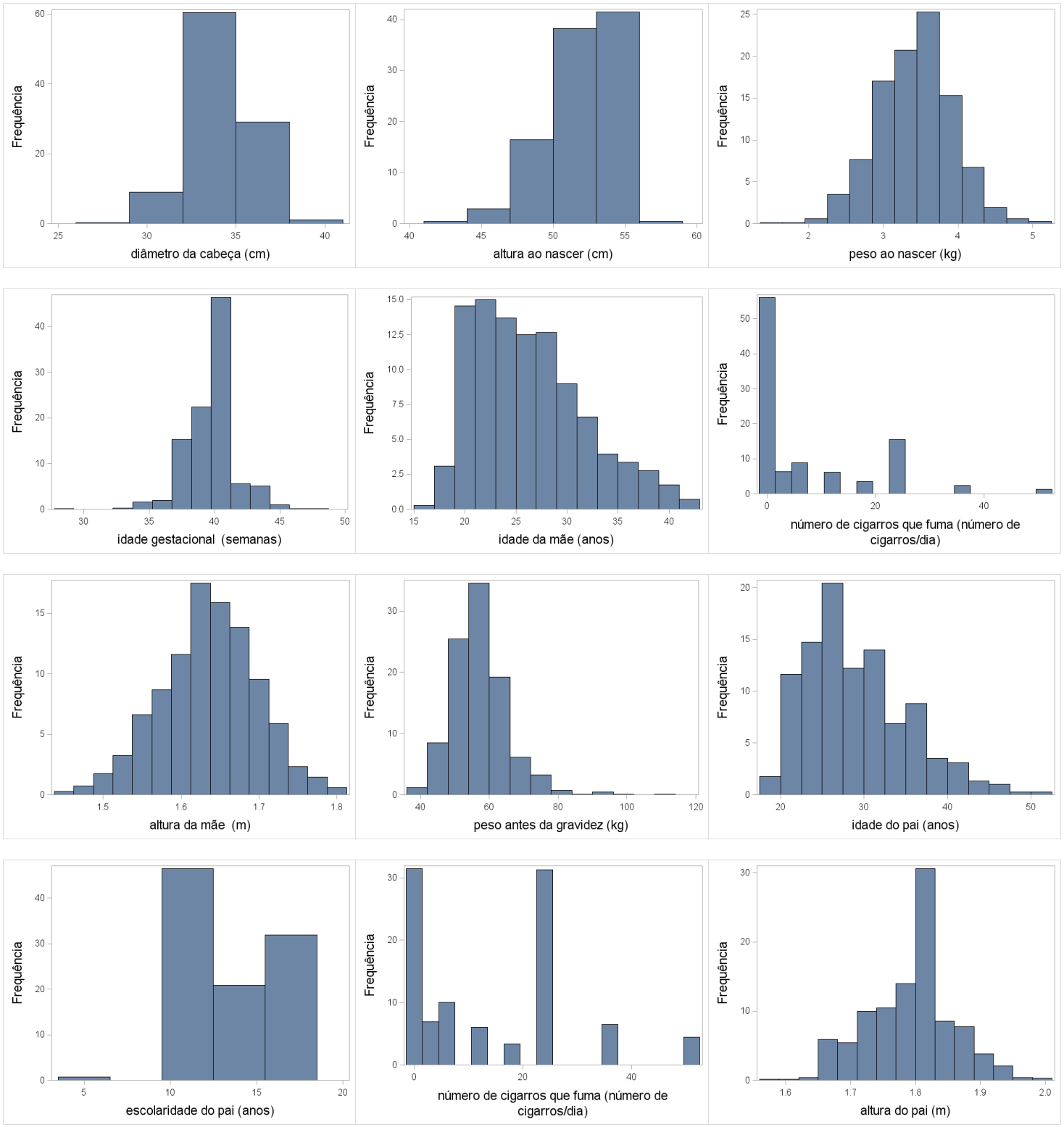

Figura 2: Histogramas

Na figura 2 apresentamos os histogramas das demais variáveis que fazem parte do banco de

dados. A exposição do comportamento dessas variáveis, nos ajuda a conhecê-las visualmente, identificando suas tendências e complementando os resultados das medidas descritivas. Nota-se que as variáveis número de cigarros que a mãe e o pai fumam e escolaridade do pai, são quantitativa discretas, assim, para estes casos o histograma foi interpretado como um gráfico de barras, já que para tais casos não foram construídos agrupamentos para a visualização gráfica.

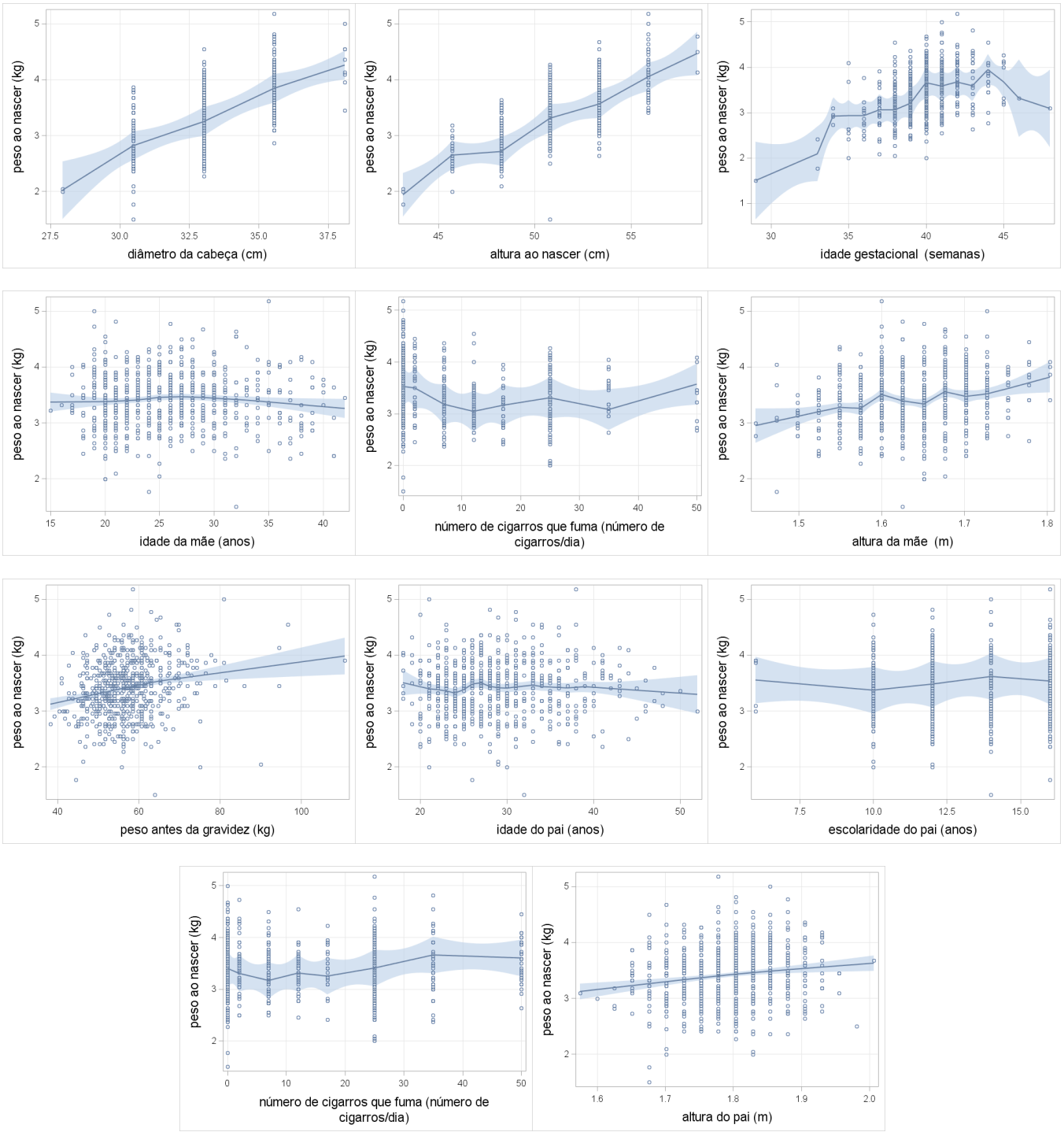

Figura 3: Scatter plots

Aqui verificamos por gráficos de dispersão apresentados em 3 as relações individuais entre a respostas e as covariáveis. Desse modo, constatamos que poucos gráficos apresentam uma relação linear com a resposta. Os casos em que tal efeito ocorre e é passível de visualização são: diâmetro da cabeça, altura ao nascer e altura da mãe e do pai. Nos demais, não foi

possível verificar tendências claras.

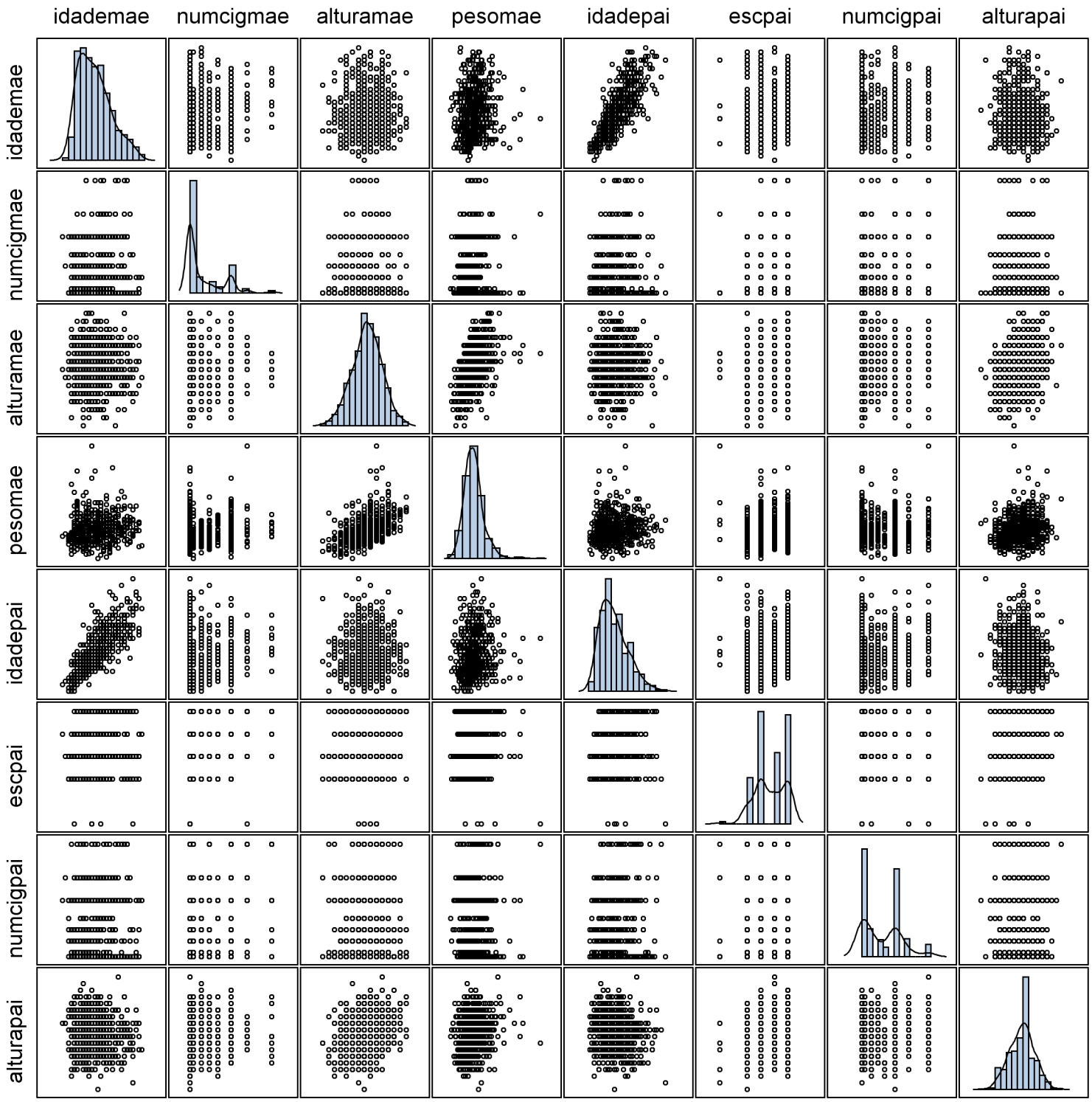

Figura 4: Matriz de scatter plots para os dados da mãe e do pai

No painel acima apresentamos a matriz de dispersão para as variáveis explicativas referentes

as características dos pais. Como já era desconfiado, há aparentemente uma relação linear entre a idade da mãe e do pai. Fato que pode causar problemas em nosso modelo, desse modo, este é um primeiro indício de problemas de colinearidade entre as covariáveis.

4 Modelo de regressão geral

Primeiramente realizamos um modelo incluindo todas as variáveis disponíveis em nosso banco de dados, a partir disso, foi iniciada a seleção do conjunto de covariáveis que permaneceram na análise final. Como o intuito é verificar as características dos pais que refletem no peso da criança ao nascer, iremos analisar com cautela as variáveis altura do bebê e diâmetro da cabeça. Já que por se tratar de uma característica da criança, ela provavelmente aparecerá muito mais relacionada com seu peso quando comparada as demais. Fato que pode causar problemas em nosso real interesse que é o de verificar a contribuição das características dos pais para o fenômeno.
Nas tabelas expostas abaixo temos os resultados do modelo geral, ou seja, o modelo em que todas as variáveis do banco foram inclusas. É válido ressaltar que as variáveis diâmetro da cabeça, altura do bebê ao nascer, idade gestacional e peso da mãe antes de nascer que se mostraram significativas, em sua maioria se referem as características da criança. Além disso, observando o $R^2_a just$ vemos que o modelos explicou apenas 64,62% da variação da variável respostas, valor relativamente baixo, dado a quantidade de variáveis empregadas.

Tabela 2: Regressão geral

Number of Observations Read	680
Number of Observations Used	680

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	109.36148	9.94195	115.23	<.0001
Error	668	57.63279	0.08628		
Corrected Total	679	166.99426			

Root MSE	0.29373	R-Square	0.6549
Dependent Mean	3.41248	Adj R-Sq	0.6492
Coeff Var	8.60749		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-7.35180	0.47042	-15.63	<.0001
diacabeca	diâmetro da cabeça (cm)	1	0.10871	0.00808	13.45	<.0001
altura	altura ao nascer (cm)	1	0.09456	0.00543	17.40	<.0001
idadegest	idade gestacional (semanas)	1	0.04386	0.00646	6.79	<.0001
idademae	idade da mãe (anos)	1	-0.00286	0.00362	-0.79	0.4293
numcigmae	número de cigarros que fuma (número de cigarros/dia)	1	-0.00131	0.00106	-1.23	0.2195
alturamae	altura da mãe (m)	1	0.28348	0.21492	1.32	0.1876
pesomae	peso antes da gravidez (kg)	1	0.00480	0.00165	2.92	0.0036
idadepai	idade do pai (anos)	1	-0.00013078	0.00325	-0.04	0.9679
escpai	escolaridade do pai (anos)	1	0.00558	0.00550	1.02	0.3103
numcigpai	número de cigarros que fuma (número de cigarros/dia)	1	-0.00007451	0.00084620	-0.09	0.9299
alturapai	altura do pai (m)	1	-0.12562	0.18339	-0.68	0.4936

Observando as estimativas e os erros, constatamos que alguns casos retornaram erros padrões bastante elevados, ultrapassando as estimativas em alguns casos. Dessa forma, precisamos realizar a análise dos resíduos e também, verificar as variáveis que contribuem de forma relevante para sua construção.

4.1 Diagnóstico de colinearidade e multicolinearidade

Tabela 3: Medidas de posição e dispersão

Pearson Correlation Coefficients, N = 680
Prob > |r| under H0: Rho=0

	pesonasc	diacabeca	altura	idadegest	idademae	numcigmae	alturamae	pesomae	idadepai	escpai	numcigpai	alturapai
pesonasc	1.00000	0.62459 <.0001	0.71136 <.0001	0.42585 <.0001	0.00131 0.9729	-0.17941 <.0001	0.20254 <.0001	0.22158 <.0001	0.01645 0.6685	0.03302 0.3899	-0.02337 0.5430	0.15416 <.0001
diacabeca	0.62459 <.0001	1.00000	0.45580 <.0001	0.27105 <.0001	0.04530 0.2381	-0.13105 0.0006	0.11587 0.0025	0.12016 0.0017	0.03980 0.3001	-0.00164 0.9660	-0.01498 0.6965	0.10762 0.0050
altura	0.71136 <.0001	0.45580 <.0001	1.00000	0.33070 <.0001	0.00497 0.8971	-0.18823 <.0001	0.17547 <.0001	0.17068 <.0001	0.01223 0.7503	0.01900 0.6208	0.00040 0.9918	0.21222 <.0001
idadegest	0.42585 <.0001	0.27105 <.0001	0.33070 <.0001	1.00000	0.00341 0.9292	-0.07084 0.0649	0.04765 0.2146	0.05173 0.1778	0.04223 0.2715	0.03536 0.3572	-0.00247 0.9487	0.02399 0.5324
idademae	0.00131 0.9729	0.04530 0.2381	0.00497 0.8971	0.00341 0.9292	1.00000	0.04500 0.2412	0.01749 0.6490	0.11573 0.0025	0.81711 <.0001	0.24059 <.0001	0.01662 0.6653	-0.07111 0.0639
numcigmae	-0.17941 <.0001	-0.13105 0.0006	-0.18823 <.0001	-0.07084 0.0649	0.04500 0.2412	1.00000	0.02593 0.4996	-0.02576 0.5024	0.02771 0.4707	0.02372 0.5370	0.26171 <.0001	0.01078 0.7791
alturamae	0.20254 <.0001	0.11587 0.0025	0.17547 <.0001	0.04765 0.2146	0.01749 0.6490	0.02593 0.4996	1.00000	0.49419 <.0001	0.01799 0.6396	0.10799 0.0048	-0.01470 0.7019	0.30333 <.0001
pesomae	0.22158 <.0001	0.12016 0.0017	0.17068 <.0001	0.05173 0.1778	0.11573 0.0025	-0.02576 0.5024	0.49419 <.0001	1.00000	0.12399 0.0012	0.00127 0.9736	-0.02747 0.4745	0.16642 <.0001
idadepai	0.01645 0.6685	0.03980 0.3001	0.01223 0.7503	0.04223 0.2715	0.81711 <.0001	0.02771 0.4707	0.01799 0.6396	0.12399 0.0012	1.00000	0.22040 <.0001	0.03968 0.3015	-0.13441 0.0004
escpai	0.03302 0.3899	-0.00164 0.9660	0.01900 0.6208	0.03536 0.3572	0.24059 <.0001	0.02372 0.5370	0.10799 0.0048	0.00127 0.9736	0.22040 <.0001	1.00000	-0.18228 <.0001	0.10778 0.0049
numcigpai	-0.02337 0.5430	-0.01498 0.6965	0.00040 0.9918	-0.00247 0.9487	0.01662 0.6653	0.26171 <.0001	-0.01470 0.7019	-0.02747 0.4745	0.03968 0.3015	-0.18228 <.0001	1.00000	0.01365 0.7224
alturapai	0.15416 <.0001	0.10762 0.0050	0.21222 <.0001	0.02399 0.5324	-0.07111 0.0639	0.01078 0.7791	0.30333 <.0001	0.16642 <.0001	-0.13441 0.0004	0.10778 0.0049	0.01365 0.7224	1.00000

Como já adiantamos na análise gráfica, as variáveis referentes às idades dos pais apresentam uma correlação de Pearson bastante elevada representando $(0,8171)$. Tendo em vista tal aspecto, precisamos agora investigar se este fato influência negativamente nosso modelo, caso seja constatado que sim, será necessário cogitar a necessidade de exclusão de uma delas para prosseguir com a modelagem.

Tabela 4: VIF e Tolerância

Variable	Tolerance	VIF	Label
Intercept	.	0	Intercept
diacabeca	0.76876	1.30080	diâmetro da cabeça (cm)
altura	0.69156	1.44601	altura ao nascer (cm)
idadegest	0.86518	1.15583	idade gestacional (semanas)
idademae	0.32483	3.07853	idade da mãe (anos)
numcigmae	0.88239	1.13329	número de cigarros que fuma (número de cigarros/dia)
alturamae	0.69149	1.44615	altura da mãe (m)
pesomae	0.72486	1.37958	peso antes da gravidez (kg)
idadepai	0.32014	3.12359	idade do pai (anos)
escpai	0.86691	1.15353	escolaridade do pai (anos)

Na tabela 4 temos os valores de VIF ou fator de inflação da variância. Analisando os resultados, temos que todos os VIF são menores que 10, porém as idades dos pais apresentam valores superiores a 3, destacando mais uma vez a necessidade de trabalharmos essas variáveis para o modelo final. Assim, há indícios de multicolinearidade. Para resolver este problema, foram gerados e comparados modelos com a idade da mãe e do pai separadamente.

Partindo disso, constatamos que o modelo apenas com a idade da mãe se mostrou apenas um pouco mais expressivo, apresentando valores de AIC e BIC menores e um R_{ajust} um pouco maiores. Quanto aos erros do modelo verificamos que não ocorreram distorções significativas, assim, optamos por manter ambas as variáveis e seguir com o estudo.

4.2 Gráfico resíduos vs preditos

O gráfico dos resíduos versus valores ajustados (preditos) pode indicar a não existência de uma relação linear entre as variáveis explicativas e a variável resposta, por meio de alguma tendência nos pontos. Dessa forma, os resíduos devem estar aleatoriamente distribuídos em torno do 0, sem qualquer tendência.

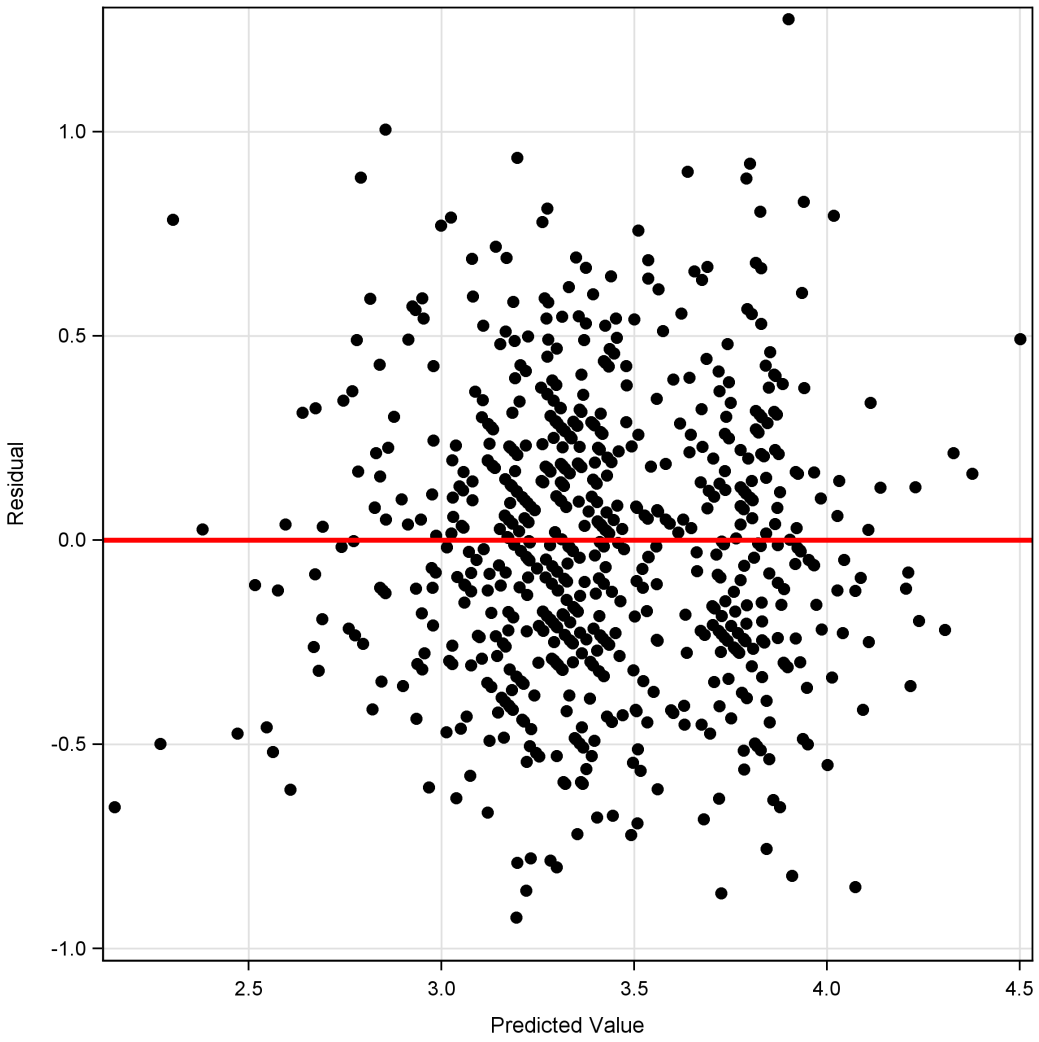

Figura 5: Gráfico dos resíduos vs preditos

Verificando a figura 5 constatamos que não há um padrão claro na distribuição dos valores preditos com os resíduos, apesar de ser possível visualizar uma clara concentração de pontos no centro do gráfico. Assim, pela análise gráfica não foi possível afirmar com certeza que não há problema de heterocedasticidade, desse modo, seguimos com os testes de White e Breusch-Pagan, para auxiliar em nossas conclusões.

4.3 Testes de homocedasticidade

Tabela 5: Testes de White e Breusch-Pagan

Test	Statistic	DF	ProbChiSq	Variables
White's Test	53.78	54	0.4829	Cross of all vars
Breusch-Pagan	9.39	9	0.4017	1, diacabeca, idadegest, idademae, numcigmae, alturamae, pesomae, escpai, numcigpai, alturapai

Pelos testes de Breusch-Pagan e White constatamos que as variâncias são igual, ou seja, não rejeitamos H_0, dessa forma, não há evidências de problemas de heterocedasticidade com o modelo. É preciso porém ressaltar que apesar de não conseguirmos constatar problemas a partir destes testes, a análise gráfica levantou duvida quanto a tal conclusão.

4.4 Diagnóstico de normalidade

A normalidade dos resíduos é uma suposição essencial para que os resultados do modelo de regressão linear sejam confiáveis. Podemos verificar essa suposição por meio do gráfico quantil-quantil e de testes como Shapiro-Wilk e Kolmogorov-Smirnov.

O gráfico quantil-quantil (q-q) é uma ferramenta muito útil para checar adequação de distribuição de frequência dos dados à uma distribuição de probabilidades(nesse caso a normal). Na análise de resíduos de modelos de regressão o gráfico q-q é usado para verificar se os resíduos apresentam distribuição normal.

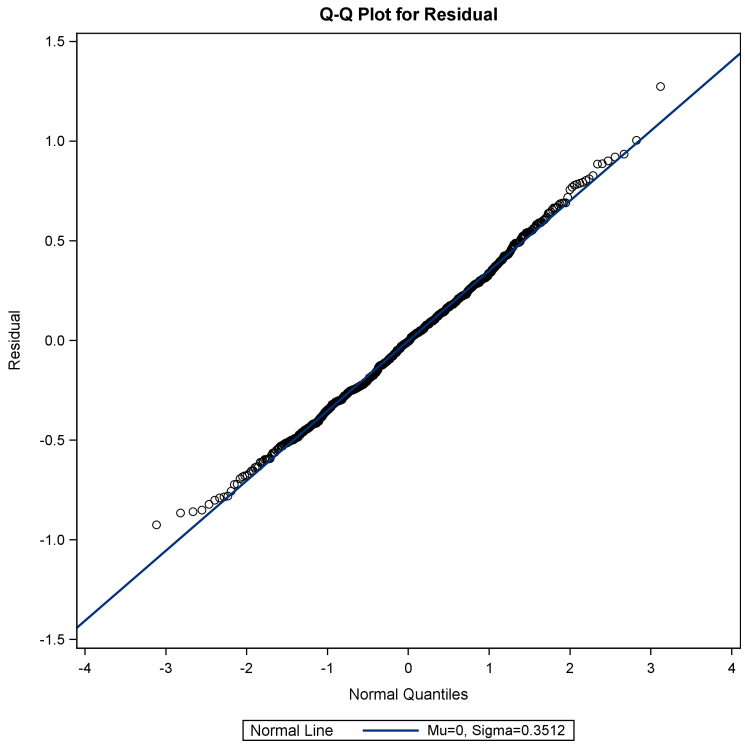

Figura 6: Q-Q Normal plot

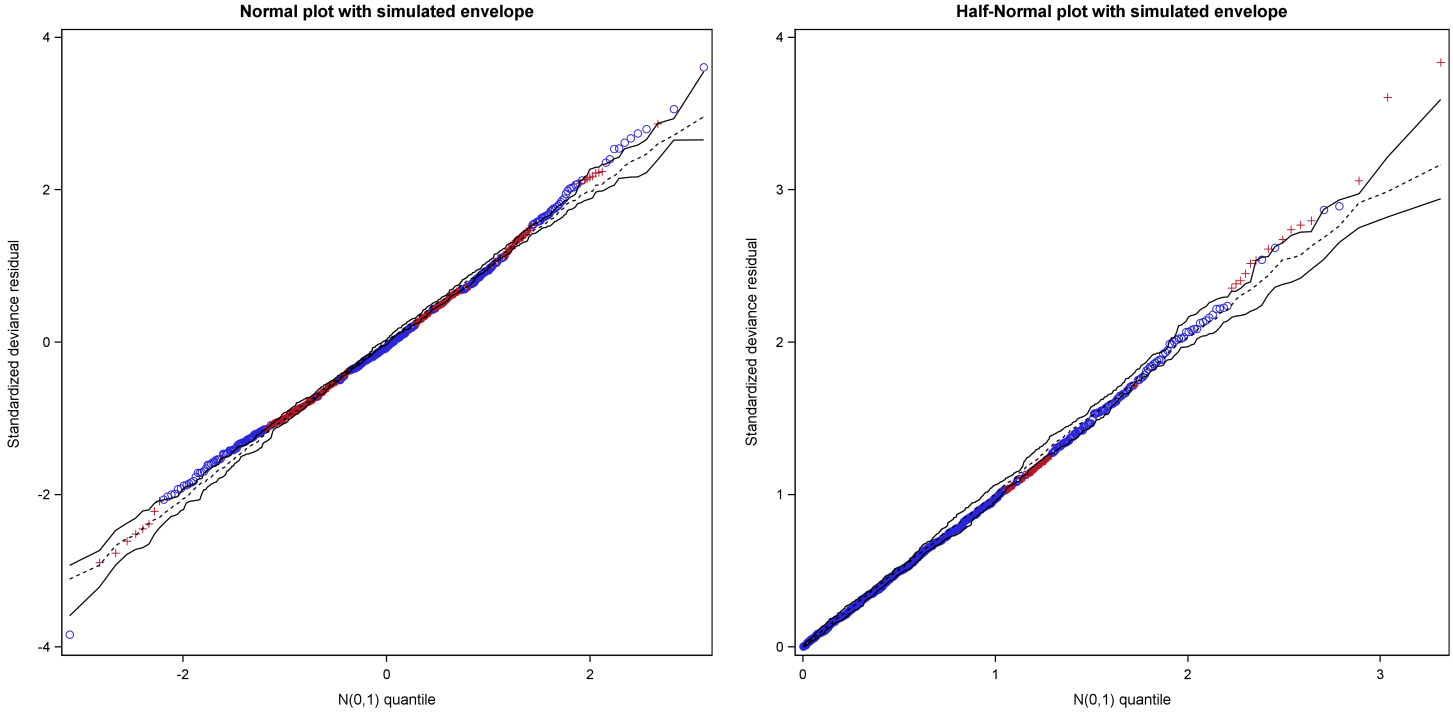

Figura 7: Normal plot com envelope

Verificamos no gráfico 6 que os resíduos parecem se ajustam bem a reta teórica e pelos gráficos de envelope, constatamos que a maior parte dos pontos encontram-se dentro ou muito próximo da banda de confiança. Entretanto, é possível observar que nas extremidades da reta existem 3 pontos distantes desta, e consequentemente, fora do intervalo, sendo possíveis casos de influência que exigem atenção. É preciso ressaltar que nesse modelo não iremos realizar a análise dos pontos de influência mesmo constatando sua necessidade, tal processo será realizado no modelo final com maior detalhamento.

4.5 Testes de Normalidade dos erros

Tabela 6: Testes de normalidade

Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	W	0.99521	Pr < W	0.0329
Kolmogorov-Smirnov	D	0.028561	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.12473	Pr > W-Sq	0.0527
Anderson-Darling	A-Sq	0.87803	Pr > A-Sq	0.0244

Como observamos os testes de normalidade não nos auxiliaram a obter conclusões definitivas, uma vez que 2 deles apontaram para a normalidade dos resíduos e outros 2 para a não normalidade. Apesar de amplamente utilizados estes testes contém sérias limitações, já que se baseiam em agrupamento dos dados em classes e posições para o cálculo de suas estatísticas, muitas vezes não os dados em si. Isto posto, decisões baseadas em seus resultados devem ser tomadas de modo parcimonioso e sempre em conjunto com a análise gráfica.

4.6 Estatísticas para seleção

Tabela 7: Os 5 melhores modelos segundo rank e Cp

NumInModel	Cp	VarsInModel	rank
6	6.0751	diacabeca idadegest numcigmae alturamae pesomae alturapai	2006
7	7.0291	diacabeca idadegest idademae numcigmae alturamae pesomae alturapai	2006
7	7.8355	diacabeca idadegest numcigmae alturamae pesomae escpai alturapai	1996
7	7.8605	diacabeca idadegest numcigmae alturamae pesomae numcigpai alturapai	1992
8	8.4182	diacabeca idadegest idademae numcigmae alturamae pesomae escpai alturapai	1992

Para a escolha no modelo com o melhor ajuste aplicamos um rank para cada umas das estatísticas utilizadas, dessa forma, o modelo que obteve o maior R^2_{ajust} recebeu o maior rank, o que apresentou o menor BIC e AIC também recebeu o maior e assim por diante. A

tabela 7 apresenta os 5 melhores modelos segundo esses critérios. Após a análise individual de cada um deles, foi escolhido o primeiro modelo da tabela, somando um rank de 2006 e $C_p = 6,0751$, muito próximo ao número de parâmetros que é 6.

5 Modelo final

O modelo final foi composto pelas seguintes variáveis explicativas: diâmetro da cabeça (cm); idade gestacional (semanas); número de cigarros que fuma (número de cigarros/dia); altura da mãe (m); peso antes da gravidez (kg); e altura do pai(m).

Antes de analisarmos suas estimativas, iremos verificar a qualidade de seu ajuste, fazendo as análises de normalidade dos resíduos, homoscedasticidade e multicolinearidade.

5.1 Multicolinearidade

Variable	Tolerance	VIF	Label
Intercept	.	0	Intercept
diacabeca	0.89387	1.11873	diâmetro da cabeça (cm)
idadegest	0.92470	1.08143	idade gestacional (semanas)
numcigmae	0.97837	1.02211	número de cigarros que fuma (número de cigarros/dia)
alturamae	0.70239	1.42371	altura da mãe (m)
pesomae	0.75056	1.33233	peso antes da gravidez (kg)
alturapai	0.90220	1.10840	altura do pai (m)

Figura 8: Valores VIF para o modelo final

Como é possível observar pela tabela acima, os valores VIF estão todos próximos a 1, valores bem menores que o obtido na versão inicial. Assim, concluímos que não há problema de multicolinearidade. Se voltarmos ao modelo anterior e verificarmos os erros padrões das estimativas notaremos que eles foram substancialmente maiores quando comparados aos valores de suas estimativas. No caso do modelo final, tal efeito não foi constatado. Como afirmam Agresti e Finlay (2012), o problema de multicolinearidade ocasiona erros padrões inflados para a estimativas dos parâmetros da regressão, fato que culmina em modelos com baixo valor explicativo.

5.2 Diagnóstico de homoscedasticidade

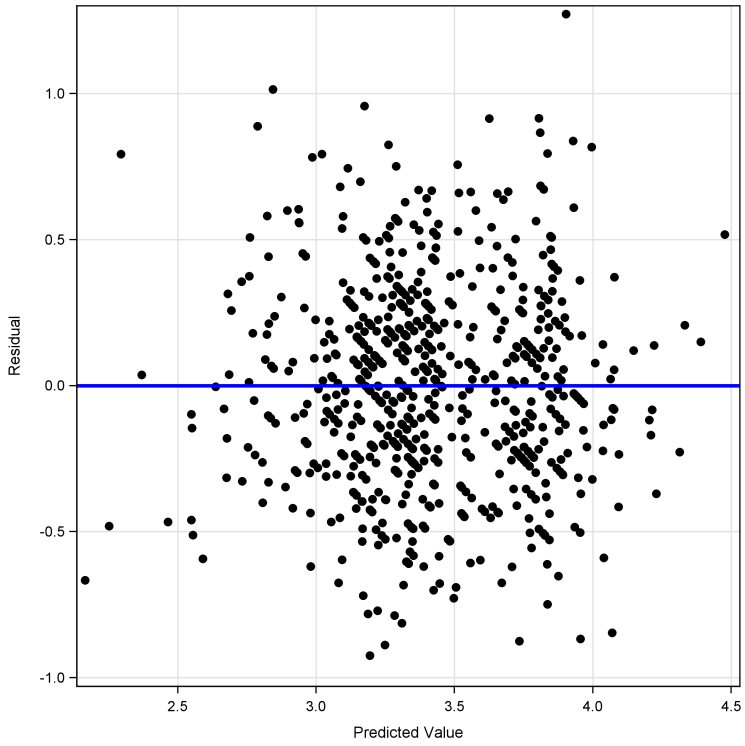

Figura 9: resíduo vs predito para o modelo final

A figura 9 mostra o gráficos da distribuição dos resíduos pelos valores preditos. Observando a distribuição dos resíduos em torno de 0 constatamos que, apesar de uma concentração dos dados no centro do gráfico, não há um padrão claro em seu comportamento. Desse modo, podemos concluir que eles se organizam de forma aleatória, apontando para a não existência de homoscedasticidade. Para reforçar essa hipótese seguiremos com os testes White e Breusch-Pagan expostos abaixo.

Tabela 8: Teste de White e Breusch-Pagan para o modelo final

Test	Statistic	DF	ProbChiSq	Variables
White's Test	34.00	27	0.1660	Cross of all vars
Breusch-Pagan	8.94	6	0.1768	1, diacabeca, idadegest, numcigmae, alturamae, pesomae, alturapai

Partindo dos resultados observados em 8 verificamos a homogeneidade da variância pelos testes de White e Breusch-Pagan. A partir do valor-p de ambos os testes, constatamos que não é possível rejeitar H_0, ou seja, as variâncias são homogêneas, assim, não constatamos problemas de heterocedasticidade.

5.3 Diagnóstico de normalidade

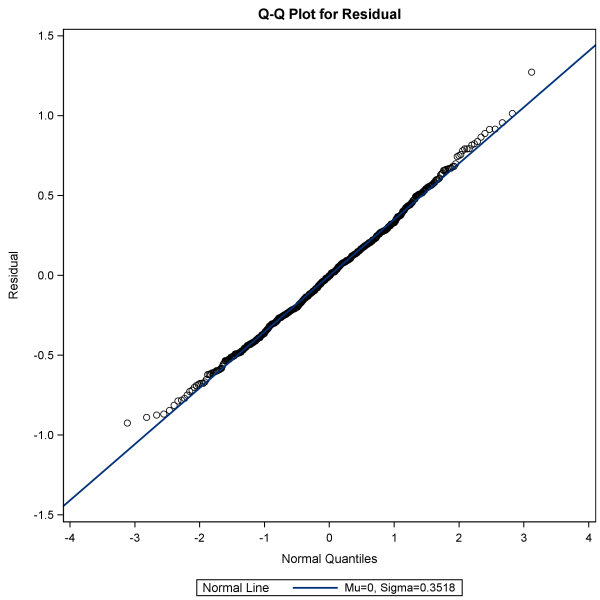

Figura 10: Q-Q plot para o resíduo

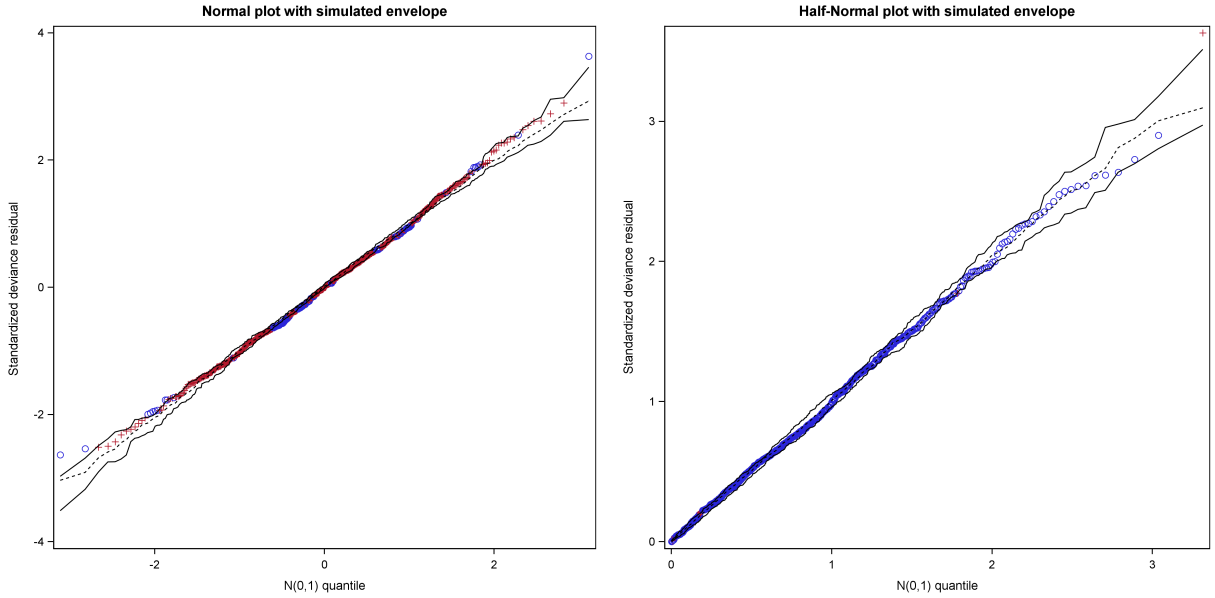

Figura 11: Normal plot com envelope

No que diz respeito a normalidade dos resíduos, verificamos pelo gráfico quantil-quantil 10 que os pontos parecem se ajustam bem a reta teórica, ou seja, há uma adequação de distribuição de frequência dos dados à uma distribuição de probabilidade normal. Além

disso, pelos gráficos de envelope constatamos que a maior parte dos pontos no decorrer da reta encontram-se dentro ou muito próximo da banda de confiança, sendo que apenas 3 pontos se distanciam de modo acentuado dela. Esse pontos serão investigados de modo mais detalhado na análise de influência.

Tabela 9: Testes de normalidade

Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	W	0.99745	Pr < W	0.3809
Kolmogorov-Smirnov	D	0.024888	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.046139	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.324647	Pr > A-Sq	>0.2500

Por fim foram realizados testes de normalidade para os erros, analisando os resultados verificamos que para todos os testes é possível assumir a normalidade dos erros. Diferente do modelos inicial, estes testes nos conduziram aos mesmos resultados, fato que nos dá mais confiança em assumir a normalidade dos resíduos.

5.4 Pontos influentes

Como as estimativas dos mínimos quadrados dos parâmetros podem ser fortemente influenciadas por valores atípicos é preciso identificar a existência desses pontos e o tipo de influência que eles podem causar no modelo. Para isso, analisaremos o leverage (pontos de alavancagem) e os outliers a partir do gráfico exposto na figura 13.

Por default o SAS utiliza 2 desvios padrão para considerar uma observação outlier, entretanto, se considerarmos 3 desvios da média como é mostrado no gráfico a direita da figura 13 teremos apenas um, sendo ele o caso 442. No que diz respeito ao Leverage, Belsley, Kuh e Welsch (1980) apontam que seu ponto de corte, para considerarmos casos problemáticos deve ser $\frac{2p}{n}$, sendo p o número de parâmetros do modelo. Assim, verificamos a existência de vários pontos de alavancagem. Porém, nenhum aparece como outlier e alavancagem de modo simultâneo.

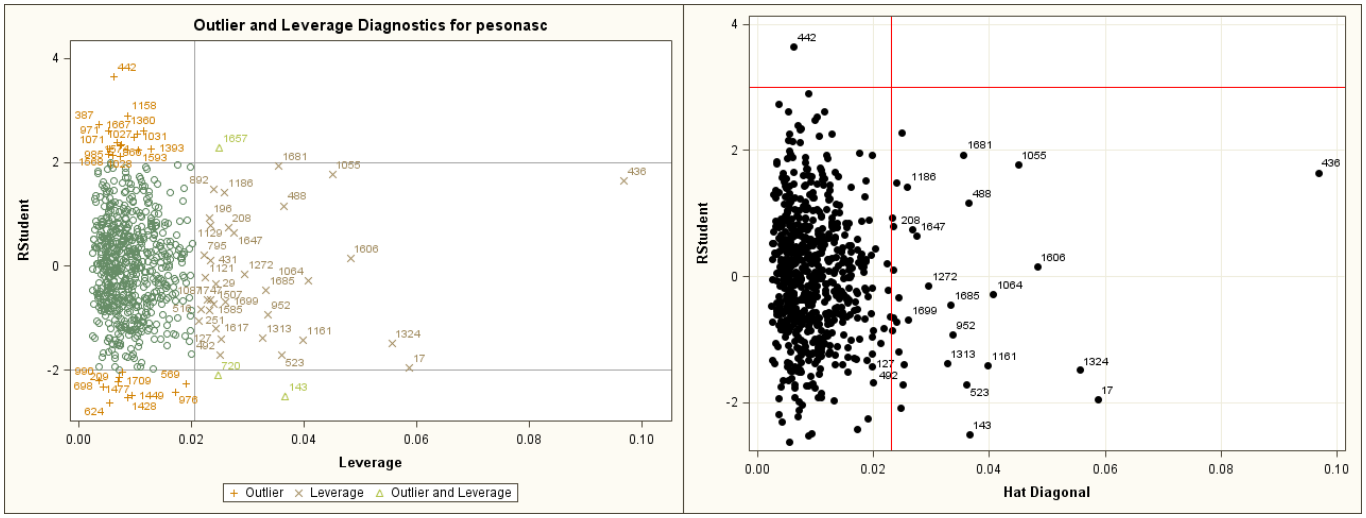

Figura 12: Leverage e Outlier para o modelo final

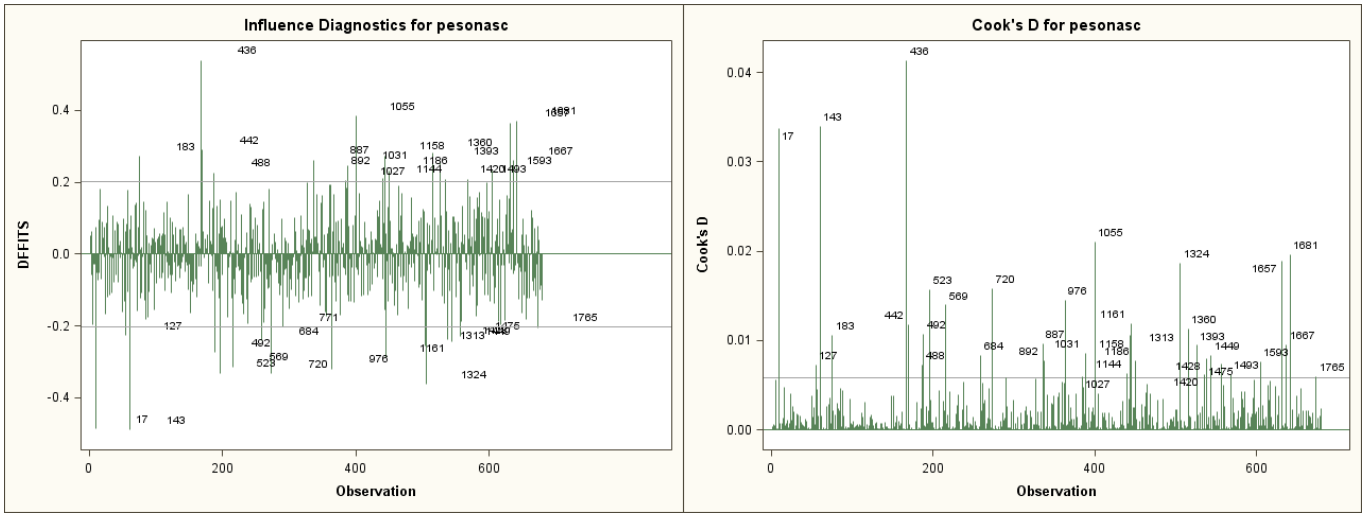

Figura 13: Cook's D e DFFITS para o modelo final

Os gráficos da distância de Cook e dos DFFits também apontam para a existência de pontos problemáticos no modelo empregado. No caso dos DFFits que resume o efeito no ajuste pela remoção da observação problemática, observamos que os pontos 436, 17,1055, 1606 e 143 são os mais destoantes, ultrapassando de forma acentuada os limites dados por $2\sqrt{\frac{p}{n}}$ como propõe Belsley, Kuh e Welsch (1980). Além disso, esses mesmo pontos são vistos no gráfico da distância de Cook e também, nos DfBetas como é possível verificar na figura abaixo.

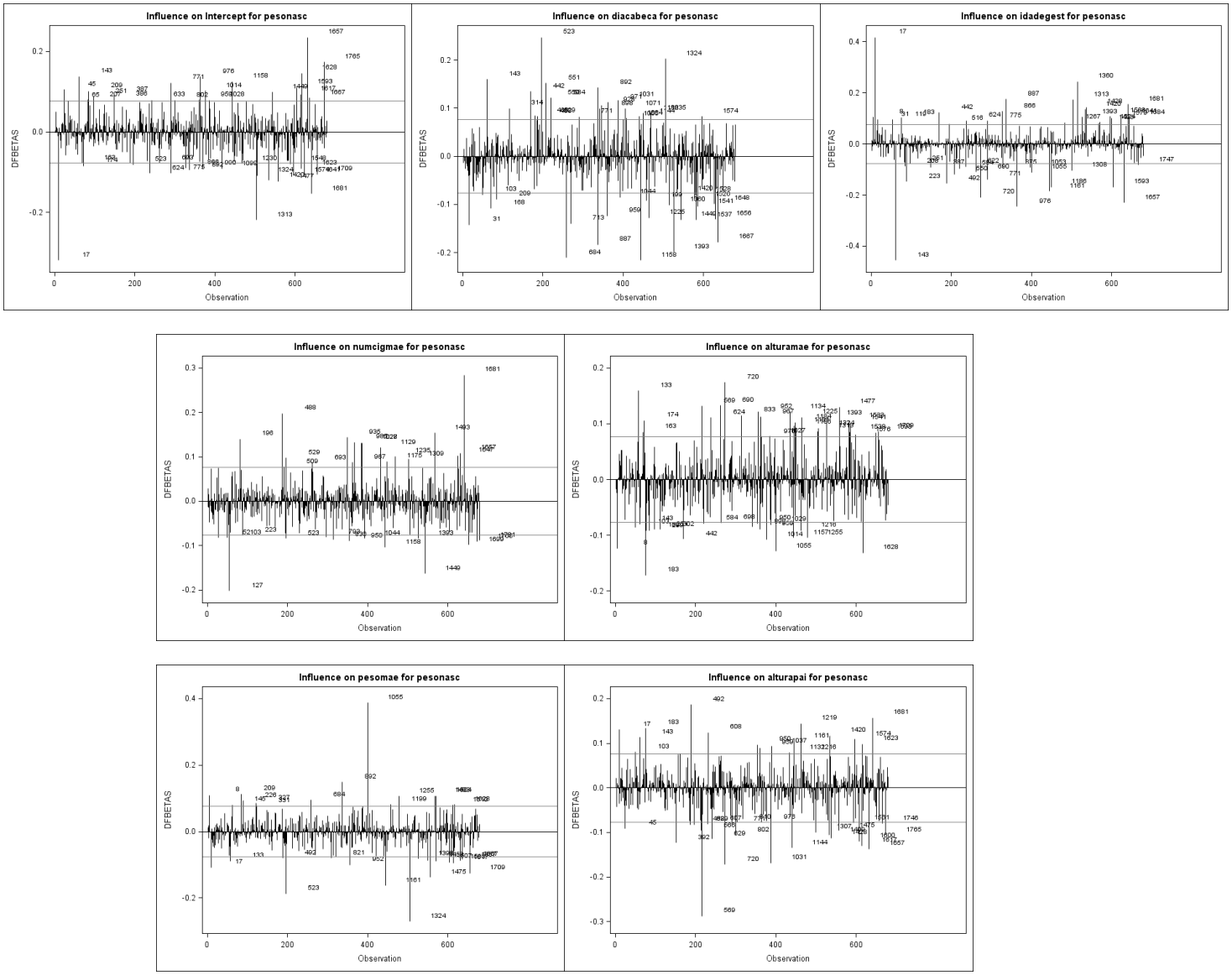

Figura 14: Gráficos dos DFBetas

Dado que muitos pontos foram constatados nas análises anteriores, adotamos o seguinte critério. Foram, verificados por meio de um algoritmos, os pontos que aparecem em comum na maior parte das medidas de influência, isto é, os pontos que além de aparecerem como alavancagem ou outlier, também ultrapassaram os limites desejados quando observamos os DFFits, CooksD e os DFBetas. A lista desses casos é composta por: 17, 127, 143, 208, 436, 488, 492, 523, 952, 1055, 1064, 1161, 1186, 1272, 1313, 1324, 1606, 1647, 1681, 1685, 1699.

Em seguida, realizamos uma bateria de modelos retirando cada uma dessas observações por vez, e analisando os resultados. Para otimizar o processo foi construída uma macro que retornou as estimativas, os erros e os ajustes de cada um, permitindo assim a análise do panorama geral. Na tabela podemos ver os resultados da comparação das saídas dos modelos.

Tabela 10: Comparação entre os modelos

	Modelos					
	Completo	Sem 17	Sem 436	Sem 143	Sem 442	Sem todos os pontos
Intercept	-6.72202	-6.54330	-6.79956	-6.79722	-6.69434	-6.67371
diacabeca	0.16000	0.15966	0.16008	0.15858	0.15880	0.15717
idadegest	0.07151	0.06834	0.07224	0.07490	0.07060	0.07159
numcigmae	-0.00408	-0.00418	-0.00429	-0.00414	-0.00401	-0.00437
alturamae	0.51088	0.51090	0.55797	0.53078	0.53809	0.60372
pesomae	0.00629	0.00648	0.00532	0.00630	0.00618	0.00542
alturapai	0.41999	0.39211	0.43375	0.39590**	0.42458	0.38820**
R^2 ajust	0.4924	0.4839	0.4937	0.4968	0.4928	0.4897
Qmres	0.12484	0.12433	0.12453	0.12387	0.12261	0.12086

** valor-p > 0,05

Verificando as saídas constatamos que, os pontos 436,143,442 e 17 são os que mais influenciam nosso modelo, todavia, tal influência não se dá de modo acentuado, não prejudicando a maior parte das conclusões. O problema mais grave diz respeito a variável altura que pai, pois nos modelos em que a observação 17 é excluída, ela deixa de ser significativa. Agresti e Finlay(2012) expõe que dados individuais tendem a ter pouca influência em amostras grandes, neste caso nossa amostra pode ser considerada grande, sendo composta por 680 observações. Mesmo assim, o ponto 17 ainda é capaz de modificar algumas conclusões.

Para seguir com o trabalho todas as observações serão mantidas, principalmente por não termos conhecimento da área do fenômenos investigado, para justificar melhor a necessidade de exclusão. A apresentação dos pontos de influência serviu como forma de alerta acerca as alterações geradas por esses valores nas conclusões de um modelo.

5.5 O modelo

Tabela 11: Modelo de regressão múltipla final

Number of Observations Read	680
Number of Observations Used	680

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	82.97408	13.82901	110.77	<.0001
Error	673	84.02018	0.12484		
Corrected Total	679	166.99426			

Root MSE	0.35333	R-Square	0.4969
Dependent Mean	3.41248	Adj R-Sq	0.4924
Coeff Var	10.35415		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-6.72202	0.55544	-12.10	<.0001
diacabeca	diâmetro da cabeça (cm)	1	0.16000	0.00902	17.75	<.0001
idadegest	idade gestacional (semanas)	1	0.07151	0.00752	9.51	<.0001
numcigmae	número de cigarros que fuma (número de cigarros/dia)	1	-0.00408	0.00122	-3.35	0.0008
alturamae	altura da mãe (m)	1	0.51088	0.25651	1.99	0.0468
pesomae	peso antes da gravidez (kg)	1	0.00629	0.00195	3.23	0.0013
alturapai	altura do pai (m)	1	0.41999	0.21303	1.97	0.0491

A tabela 11 apresenta as estimativas do nosso modelo final. A primeira tabela nos mostra o número de observações presentes no banco e o número utilizado para a regressão, nesse caso todos os 680 casos foram empregados, dado que não tivemos dados faltantes ou observações retiradas por serem consideradas pontos influentes.

A segunda tabela é referente a análise de variância do modelo. A estatística F retornou um valor de 110,77, sendo seu valor-p $<$ 0,0001. Dessa forma, temos que nosso modelo está adequado para prever o peso do bebê ao nascer.

A terceira tabela contém os valores do R^2 e R^2_{ajust}, como estamos trabalhando com uma regressão múltipla utilizaremos os valores do R^2_{ajust}. Assim, constatamos que nosso modelo explica 49,69% da variação do peso da criança ao nascer. É preciso ressaltar que nosso modelo explica pouco o peso da criança, ou seja, menos de 50% está endo explicado pelas variáveis elencadas. Dessa forma, é válido ressaltar que outras variáveis não apresentadas no banco utilizado podem explicar os outros 50%.

Por fim a terceira tabela traz as estimativas individuais. Analisando os resultados constatamos que a cada centímetro elevado no diâmetro da cabeça das crianças aumenta em 0.16 kg seu peso ao nascer. No que se refere a idade gestacional contada em semanas, verificamos que quando mais longa ela é, maior o peso do bebê, ou seja, a cada semana adicionada na gestação, o peso do bebê aumenta em 0.07 kg. Quanto a variável número de cigarros fumados por dia pela mãe temos uma relação negativa, desse modo, observá-se que quanto mais cigarros a mãe fuma por dia, menor é o peso da criança ao nascer. Verificando as alturas da mãe e do pai constatamos que quanto mais elevada elas são, maior é o peso da criança. Por fim, o peso da mãe antes da gravidez também apresenta uma relação positivas com o peso da criança ao nascer, sendo que mães que pesavam mais tendem a ter bebê com maiores pesos.

É válido ressaltar que para as alturas da mãe e do pai os erros padrão das estimativas foram um pouco elevados, chegando a representar em torno de 50% do valor estimado em ambos os casos. Nos demais os erros foram relativamente baixos, o que permite concluir que temos estimativas mais precisas.

6 Conclusão

O presente trabalho buscou explicar as características gerais do pai e da mãe que influenciam no peso do bebê ao nascer. Para realizar essa investigação partimos de um banco de dados com 680 casos e 12 variáveis, procurando selecionar as que pudessem explicar esse fenômenos da forma mais robusta e adequada,tendo em vista que, o autor não detinha o conhecimento da área para auxiliar na seleção das variáveis, logo, este processo foi feito baseado apenas em métodos estatísticos.

Posto isto, inciamos as investigações realizando um modelo com todas as variáveis do banco, com o intuito de identificar possíveis candidatos para o modelos final. Partindo disso, foi feita a análise de resíduo, na qual constatamos que o modelo detinha problemas de colinearidade e multicolinearidade e um possível problema de normalidade dos resíduos. Tendo em vista tais problemas, foi implementando no software SAS 9.4 uma macro que auxiliou na escolha do modelo mais adequado, tal macro se baseou nas medidas R^2_{ajust}, AIC, BIC, EQM e Cp para retornar os modelos com melhor ajuste.

Assim sendo, o modelo final composto pelas covariáveis: diâmetro da cabeça (cm); idade gestacional (semanas); número de cigarros que fuma (número de cigarros/dia); altura da mãe (m); peso antes da gravidez (kg); e altura do pai(m). Mostrou-se significativamente melhor que o inicial, confirmando ausência de colinearidade e multicolinearidade, heterocedasticidade e normalidade dos resíduos. Contudo, identificamos a existência de pontos influentes que mesmo de modo superficial aumentaram os erros padrões das estimativas e diminuíram certa de 1% da capacidade explicativa das covariáveis sobre a resposta. Mesmo diante de tal situação consideramos que as alterações causadas por esse pontos dado o tamanho de nossa amostra, não resultaram em danos ao modelo, logo, optamos por mantê-las na análise.

Por fim, concluímos que o diâmetro da cabeça do bebê, a idade gestacional, o peso antes da gravidez e a altura do pai e da mãe contribuem em proporções distintas, mas todas positivamente para o peso da criança ao nascer. Enquanto, o número de cigarros que a mãe fuma por dia contribui negativamente para nossa resposta, ou seja, gera decréscimos no peso. No demais, como essas variáveis explicaram apenas cerca de 50% da resposta, cabe ressaltar

a necessidade de investigações com mais variáveis que possam causar influência no peso para investigações posteriores.

Referências

[1] Casella, George, and Roger L. Berger. Statistical inference. Vol. 2. Pacific Grove, CA: Duxbury, 2002.

[2] Paula, A. Gilberto. Modelos de regressão com apoio computacional, USP, 2013.

[3] Rencher, C. Alvin. Linear Models in Statistics, Department of Statistics Brigham Young University, Provo, Utah, 2000.

[4] Neter, John, et al. Applied linear statistical models. Vol. 4. Chicago: Irwin, 1996.

[5] Agresti, Alan, and Barbara Finlay. "Métodos estatísticos para as ciências sociais."Métodos estatísticos para as ciências sociais. 2012..

[6] Portal action disponível em: http://www.portalaction.com.br/

[7] Belsley, David A., Edwin Kuh, and Roy E. Welsch. "Regression diagnostics. J."(1980).