Trabalho Análise de Regressão

André Felipe B. Menezes*
Prof. Dra. Rosangela Getirana Santana

31 de janeiro de 2017

Resumo

Neste trabalho foi analisado dados proveniente de um estudo transversal, tendo como intuito identificar quais características da mãe e do pai estão associadas com o peso do recém nascido. Utilizamos o modelo de regressão linear múltipla para explicar o peso de recém nascido. Foi realizado uma breve revisão sobre a analise de regressão e seus pressupostos. No que tange os resultados, verificou-se que somente a covariável número de cigarros que a mãe fuma por dia exerce efeito negativo no peso do bebê. Além disso, o peso da mãe e sua idade gestacional exercem efeito positivo sobre o peso do recém nascido, sendo esperado um aumento de 6,28 gramas para cada kg a mais no peso da mãe e 71,51 gramas para cada semana a mais na idade gestacional.

Sumário

*Graduando em Estatística pela Universidade Estadual de Maringá.

1 Introdução

O presente trabalho analisará os dados provenientes de um estudo transversal (cross-sectional) sobre peso de recém nascidos. As informações foram coletadas: da criança, da mãe e do pai, a partir de 680 nascimentos consecutivos. O objetivo principal desta análise é identificar se alguma característica da mãe e do pai está associada ao peso do recém nascido.

Para tais propósitos, primeiramente foi realizada uma análise exploratória das variáveis a qual resultou em medidas de resumo, tabelas e gráficos de frequências. Em seguida, buscou-se ajustar um modelo de regressão para descrever o peso do recém nascido em função das variáveis da mãe e do pai. Por fim, foi realizado a análise de resíduos e diagnóstico de pontos influentes e outliers. Ressalta-se que todas análises foram conduzidas no software SAS (INSTITUTE, 2014).

Nesse sentido, este trabalho esta organizado da seguinte forma. Uma revisão da análise de regressão, bem como avaliação dos pressupostos, medidas para disgnósticos de influencia e seleção de variáveis são apresentadas na Seção 2. Em seguida, na Seção 3, apresentamos os resultados da análise descritiva, ajuste do modelo de regressão com todas covariáveis, a seleção das variáveis explicativas, e finalizando, avaliamos e interpretamos o modelo de regressão selecionado. Algumas considerações na Seção 4 finaliza este artigo.

2 Metodologia

Uma regressão linear múltipla é caracterizada quando se admite que a variável resposta (Y_i) é função de duas ou mais variáveis explicativas (regressoras). O modelo estatístico de uma regressão linear múltipla com k covariáveis pode ser expresso como

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik} + \varepsilon_i \tag{1}$$

Em notação matricial temos

$$\boldsymbol{Y} = \boldsymbol{X}\,\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{2}$$

em que $\boldsymbol{Y}$ é o vetor, de dimensões $n \times 1$, da variável aleatória Y, $\boldsymbol{X}$ é a matriz de dimensões $n \times p$, conhecida do estudo/delineamento, $\boldsymbol{\beta}$ é o vetor, de dimensões $p \times 1$, de parâmetros desconhecidos e $\boldsymbol{\varepsilon}$ são variáveis aleatórias não observáveis.

Ao estabelecer este modelo, pressupõe-se que

(i) a variável resposta Y é função linear das variáveis explicativas X_j, $j = 1, 2, \ldots, k$;

(ii) as covariáveis X_j são fixas;

(iii) $\mathbb{E}(\varepsilon_i) = 0$

(iv) os erros são homocedásticos, isto é, $\mathrm{Var}(\varepsilon_i) = \sigma^2$;

(v) os erros são independentes, isto é, $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0$, $i \neq j$;

(vi) os erros tem distribuição normal.

A suposição de normalidade é necessária para elaboração dos testes de hipóteses e obtenção de intervalos de confiança. Pode-se mostrar facilmente que o estimador de mínimos quadrados e verossimilhança para o vetor de parâmetros $\boldsymbol{\beta}$ são iguais a

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^T\,\boldsymbol{X}\right)^{-1}\,\boldsymbol{X}^T\,\boldsymbol{Y} \tag{3}$$

No que se refere a interpretação dos parâmetros em um modelo de regressão linear múltipla têm-se que (NETER et al., 1996)

- O parâmetro β_0 é o intercepto do modelo de regressão. Se o conjunto de variáveis preditoras incluir o ponto $\boldsymbol{x} = (x_1, x_2, \ldots, x_p) = (0, 0, \ldots, 0)$, o parâmetro β_0 fornece a resposta esperada nesse ponto. Caso contrário, não possui qualquer significado, como um termo isolado no modelo de regressão.
- Já o parâmetro β_i indica a mudança ocorrida em $\mathbb{E}[Y \mid X = x_i]$ a cada unidade de mudança em uma das variáveis preditoras, quando as outras são mantidas fixas.

Se as suposições são violadas, têm-se as falhas sistemáticas, tais como: não linearidade, multicolinearidade, não normalidade, heterocedasticidade e não-independência dos erros, assim a análise resultante pode levar a conclusões duvidosas. Outro fato bastante comum é a presença de pontos atípicos, que podem influenciar, ou não, o ajuste do modelo. Desse modo, nas próximas subseções mostraremos como avaliar imprecisões do modelo de regressão ajustado.

2.1 Multicolinearidade

Um problema usual em estudos observacionais é a multicolinearidade, isto é, quando duas variáveis estão moderadamente ou altamente correlacionadas. As principais medidas utilizadas para avaliar a multicolinearidade são VIF e TOL, definidos respectivamente por

$$VIF_j = \frac{1}{1 - R_j^2} \tag{4}$$

e

$$TOL_j = \frac{1}{VIF_j} \tag{5}$$

em que R_j^2 é o coeficiente de determinação da regressão de X_j sobre as outras variáveis explicativas. Nota-se que o fator de inflação da variância mede o quanto variância do coeficiente β_j é inflacionada por sua colinearidade. De modo geral, uma variável cujo *VIF* é maior que 10, possui indicativo de problemas de multicolinearidade.

Além disso, o coeficiente de correlação amostral de Pearson também pode ser utilizado para identificar altas correlações lineares o que é um indicio de multicolinearidade. O coeficiente de correlação de Pearson entre duas variáveis X e Y é definido por

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{6}$$

em que $\bar{x}$ e $\bar{y}$ são as médias amostrais.

2.2 Normalidade

Uma importante suposição que deve ser atendida pelo modelo de regressão linear é a normalidade dos resíduos, isto é, os resíduos devem seguir uma distribuição normal com média zero e variância constante. Dentre os métodos disponíveis para avaliar tal suposição têm-se os teste de normalidade, histograma dos resíduos e os gráficos normal e semi-normal de probabilidade com envelopes simulados.

Os testes de aderência sugeridos são: Shapiro-Wilk, Kolmogorov-Smirnov, Cramér-von Mises e Anderson-Darling, lembrando que a hipótese nula dos testes indicam que os resíduos são provenientes de uma distribuição normal especificada.

Segundo Weisberg (1985) o gráfico normal de probabilidades destaca-se por dois aspectos:

- identificação da distribuição originária dos dados;
- identificação de valores que se destacam no conjunto.

Consideremos $d_{(1)}, d_{(2)}, \ldots, d_{(n)}$ as estatísticas de ordem correspondentes aos resíduos obtidos a partir do ajuste de um determinado modelo. O fundamento geral para a construção do gráfico normal de probabilidades é que se os valores de uma dada amostra provêm de uma distribuição normal, então os valores das estatísticas de ordem e os Z_i correspondentes, obtidos da distribuição normal padrão são linearmente relacionados. Portanto, o gráfico

dos valores, $d_{(i)}$ versus Z_i deve ser uma reta, aproximadamente. Formatos aproximados comuns que indicam ausência de normalidade são (DEMÉTRIO; ZOCCHI, 2011):

- **S**, indica distribuições com caudas muito curtas, isto é, distribuições cujos valores estão muito próximos da média;
- **S invertido**, indica distribuições com cudas muito longas e, portanto, presença de muitos valores extremos;
- **J e J invertido**, indicam distribuições assimétricas, positivas e negativas, respectivamente.

Por fim, como regra geral, a ocorrência de pontos próximos ou fora do envelope indicam que o modelo não esta apropriado.

2.3 Homoscedasticidade

Além dos resíduos serem normalmente distribuídos sua variância deve ser constante, essa suposição é denominada homoscedasticidade. Vale destacar que a ausência de homoscedasticidade é chamada de heteroscedasticidade.

Dentre os diagnósticos de heterocedasticidade temos os testes de White e Breusch-Pagan, ambos baseados na estatística Escore. Além disso, o mais usual é plotar os resíduos contra os valores ajustados, se algum padrão for observado temos indícios de heterocedasticidade.

Uma das tentativas para solucionar o problema de heterocedasticidade foi proposta por White em 1980, consiste em utilizar uma estimativa consistente para a matriz de covariância. No software SAS a opção `HCCMETHOD` fornece as seguintes matriz de covariâncias

$$\begin{aligned}
\mathrm{HC}_0 &= (\boldsymbol{X'X})^{-1}(\boldsymbol{X'}\mathrm{diag}(e_i^2)\boldsymbol{X})(\boldsymbol{X'X})^{-1}, && (7)\\
\mathrm{HC}_1 &= \frac{n}{n-p}\mathrm{HC}_0, && (8)\\
\mathrm{HC}_2 &= (\boldsymbol{X'X})^{-1}\boldsymbol{X'}\mathrm{diag}\left(\frac{e_i^2}{1-h_{ii}}\right)\boldsymbol{X}(\boldsymbol{X'X})^{-1}, && (9)\\
\mathrm{HC}_3 &= (\boldsymbol{X'X})^{-1}\boldsymbol{X'}\mathrm{diag}\left(\frac{e_i^2}{(1-h_{ii})^2}\right)\boldsymbol{X}(\boldsymbol{X'X})^{-1}. && (10)
\end{aligned}$$

em que n é o número de observações, p número de parâmetros e $h_{ii} = \mathbf{x}_i(\boldsymbol{X'X})^{-1}\mathbf{x}_i'$ é o leverage da i-ésima observação.

2.4 Medidas para diagnóstico de influência

Apresentaremos nas seguintes subseções as principais medidas para detectar observações que influenciam de alguma forma o modelo de regressão ajustado.

Matriz leverage

Um critério para detectar pontos influentes é através da matriz de projeção ou matriz *leverage*

$$\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T$$

Valores de $h_{ii} \geq \frac{2p}{n}$ indicam observações que merecem uma análise mais apurada (BELSLEY; KUH; WELSCH, 2005). Em outra palavras são consideradas possíveis pontos de alavancagem.

Distância de Cook

A distância de Cook é uma medida de afastamento do vetor de estimativas provocado pela retirada da observação i (COOK, 1979). Essa medida é expressa por:

$$D_{(i)} = \frac{(\widehat{\beta} - \widehat{\beta}_{(i)})^T \boldsymbol{X}^T\boldsymbol{X}\,(\widehat{\beta} - \widehat{\beta}_{(i)})}{p\,QME}$$

Pontos com grandes valores de D_i tem considerável influencia nas estimativas. Um critério válido para os $D_{(i)}$ é compara-los com a distribuição $F_{\alpha,p,n-p}$.

Se $D_i \approx F_{\alpha,p,n-p}$ então a retirada do ponto i deve deslocar $\widehat{\beta}$ para o limite de uma região de confiança de 50% de baseado nos dados completos. Como $F_{\alpha,p,n-p} \approx 1$, usualmente consideram-se os pontos em que $D_i > 1$ como sendo possivelmente influentes.

DFBeta

Esta medida indica o quanto cada coeficiente de regressão $\widehat{\beta}_i$ muda, em unidades de desvio-padrão, se a i-ésima observação for removida. Sua expressão é dada por

$$\text{DFBeta}_{j(i)} = \frac{\widehat{\beta}_j - \widehat{\beta}_{j(i)}}{\sqrt{S^2_{(i)}\left(\boldsymbol{X}^T\boldsymbol{X}\right)_{(jj)}}}$$

Um valor grande de $DFBeta_{j(i)}$ indica que a observação i tem considerável influência no j-ésimo coeficiente de regressão. Para amostras grandes, observações as quais $\text{DFBeta}_{j(i)} > \frac{2}{\sqrt{n}}$ merecem atenção. Já amostras pequenas ou moderadas, as observações que merecem atenção são aquelas em que $\text{DFBeta}_{j(i)} > 1$ (BELSLEY; KUH; WELSCH, 2005).

DFFit

Esta estatística investiga a influencia da i-ésima observação nos valores preditos. O DFFit é obtido por

$$\text{DFFit}_{(i)} = \frac{\widehat{y}_i - \widehat{y}_{(i)}}{\sqrt{S^2_{(i)}\left(1 - h_{ii}\right)}}$$

Dizemos que $\text{DFFit}_{(i)}$ representa o quanto valor ajustado muda, em unidades de desvio-padrão, se a i-ésima observação for removida. Valores absolutos excedendo $2\sqrt{\frac{p}{n}}$ podem identificar observações influentes (BELSLEY; KUH; WELSCH, 2005).

Covratio

As estatísticas D_i, $\text{DFBeta}_{j(i)}$ e $\text{DFFit}_{(i)}$ fornecem uma visão do efeito de cada observação nos coeficientes estimados e nos valores ajustados. Contudo, não fornecem qualquer informação sobre a precisão geral da estimação. Por outro lado a estatística *Covratio* mede a mudança no determinante da matriz de covariância das estimativas deletando a i-ésima observação:

$$\text{Covratio}_{(i)} = \frac{|(\boldsymbol{X}_{(i)}^T \boldsymbol{X}_{(i)})^{-1} S_{(i)}^2|}{|(\boldsymbol{X}_{(i)}^T \boldsymbol{X}_{(i)})^{-1} QME|}$$

Belsley, Kuh e Welsch (1980) sugerem que observações com

$$|\text{Covratio} - 1| \geq \frac{3\,p}{n}$$

são dignas de investigação, isto é, devem ser pontos influentes.

2.5 Medidas para seleção de variáveis

Na sequência é apresentado uma explicação teórica das medidas utilizadas na seleção de variáveis.

Coeficiente de determinação R^2

É utilizado como uma medida descritiva da qualidade do ajuste e indica a proporção de variação de Y que é explicada pela regressão, ele é definido por

$$R^2 = \frac{SQR}{SQT} = 1 - \frac{SQE}{SQT}$$

O valor do coeficiente de determinação deve ser usado com precaução, pois depende do número de observações da amostra, tendendo a crescer quando n diminui. Além disso, é sempre possível torná-lo maior, pela adição de um número suficiente de covariáveis. Portanto R^2 não deve ser utilizado sozinho, mas sempre aliado a outros diagnósticos do modelo.

Coeficiente de determinação ajustado R_a^2

Um critério para a seleção de um modelo ótimo é escolher o modelo que com maior R_a^2. O coeficiente de determinação ajustado penaliza R^2 pelo número de parâmetros, ele é definido como:

$$R_a^2 = 1 - \left(\frac{n-1}{n-p}\right)\frac{SQE}{SQT}.$$

em que p é o número de parâmetros. Diferentemente do R^2 o R_a^2 permite comparar modelos com diferentes números de parâmetros.

Quadrado médio dos resíduos QME

O quadrado médio dos resíduos de um modelo de regressão é obtido por meio de

$$QME = \frac{SQE}{n-p-1}$$

O QME sempre decresce conforme o número de parâmetros no modelo aumenta. Facilmente observamos que o modelo com menor QME possui o maior R_a^2.

C_p de Mallows

A estatística C_p de Mallows compara a capacidade de previsão de subconjuntos de variáveis preditoras com o modelo completo. Ela é definida por

$$C_p = \frac{SQE_p}{QME} - n + 2p,$$

em que SQE_p é a soma de quadrados do erro para o modelo com p covariáveis. Um pequeno valor de C_p significa que o modelo é relativamente preciso. Por outro lado, modelos com C_p próximo do número de parâmetros $(p+1)$ são preferíveis.

AIC e BIC

O AIC e BIC são definidos respectivamente por

$$AIC_p = -2log(L_p) + 2[(p+1)+1] = n\,\log\left(\frac{SQE_p}{n}\right) + 2\,p$$

e

$$BIC_p = -2log(L_p) + [(p+1)+1]\log(n) = n\,\log\left(\frac{SQE_p}{n}\right) + 2\,\log(p).$$

em que L_p é a função de máxima verossimilhança do modelo. Ambos critérios penalizam modelos com muitas variáveis sendo que valores menores de AIC e BIC são desejáveis.

Critério PRESS

A estatística $PRESS$ mede o quanto os valores ajustados predizem a resposta observada, ela é definida como sendo

$$PRESS_p = \sum_{i=1}^{n}(Y_i - \widehat{Y}_{(i)})^2,$$

em que $\widehat{Y}_{(i)}$ é o valor predito da regressão sem a i-ésima observação. Modelos com valores pequenos de $PRESS$ são considerados bons candidatos.

3 Resultados

3.1 Análise descritiva

Inicialmente procedeu-se com uma breve análise descritiva da variável resposta, peso do recém nascido (gramas), e das covariáveis presentes no estudo. Na Figura 1 é apresentado o comportamento do peso do recém nascido através do histograma com a densidade da distribuição normal e a curva kernel.

É possível notar que a variável tem comportamento simétrico, uma vez que a média, mediana e moda são próximas. Ressalta-se também, que os recém nascidos apresentaram um peso médio ao nascer de 2412,47 gramas, sendo o peso mais observado de 3314,2 gramas.

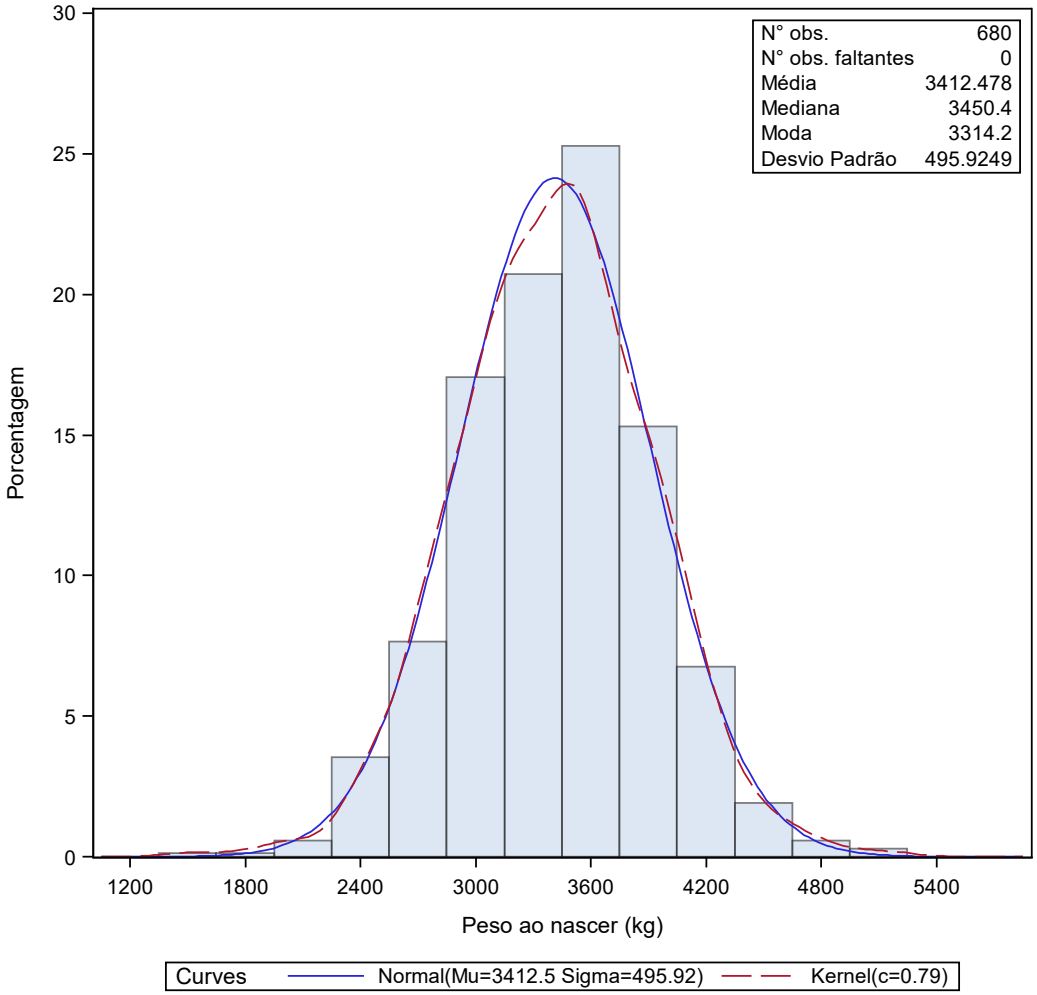

Figura 1 – Histograma do peso do recém nascido

Na Tabela 1 apresenta-se algumas medidas descritivas das covariáveis. Analisando o coeficiente de variação, observamos um comportamento homogêneo entre as alturas da mãe, sendo em média 1,64 metros. Em contrapartida, percebe-se uma alta variabilidade entre o número de cigarros por dia que os pais fumam. Enfim, com exceção das duas covariáveis enfatizadas as demais aparentam ter comportamento homogêneo. É importante enfatizar que variáveis com comportamento homogêneo, possuem maior representatividade na amostra coletada, ao passo que variáveis com alto coeficiente de variação não podem ser resumidas por medidas de posição tais como média e mediana.

Tabela 1 – Medidas descritivas das covariáveis.

Covariável	Média	Mediana	Desvio Padrão	CV(%)
diacabeça	33.58	33.02	1.59	4.74
altura	51.51	50.80	2.49	4.84
idadegest	39.77	40.00	1.88	4.72
idademae	25.86	25.00	5.46	21.13
numcigmae	7.43	0.00	11.27	151.69
alturamae	1.64	1.63	0.06	3.85
pesomae	57.10	56.25	8.04	14.09
idadepai	28.80	28.00	6.13	21.30
escpai	13.38	14.00	2.20	16.46
alturapai	1.79	1.8	0.07	3.74
numcigpai	14.44	12.00	14.17	98.14

Finalizando a análise descritiva é importante verificar se as covariáveis estão correlacionadas entre si, uma vez que alta correlação pode ser indício de multicolinearidade. O coeficiente de correlação de Pearson foi utilizado, pois todas covariáveis são contínuas.

Verifica-se na Figura 2 que a maior correlação ocorre entre as variáveis idadepai e idademae, uma alta correlação positiva de 0,81. Já as demais covariáveis apresentam coeficiente de correlação menores de 0,5 indicando baixa ou nenhuma correlação linear.

1

The CORR Procedure

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

	diacabeca	altura	idadegest	idademae	numcigmae	alturamae	pesomae	idadepai	escpai	alturapai	numcigpai
diacabeca	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	0.10762 0.0050	-0.01498 0.6965
altura	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	0.21222 <.0001	0.00040 0.9918
idadegest	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	0.02399 0.5324	-0.00247 0.9487
idademae	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	-0.07111 0.0639	0.01662 0.6653
numcigmae	-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	0.01078 0.7791	0.26171 <.0001
alturamae	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	0.30333 <.0001	-0.01470 0.7019
pesomae	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	0.16642 <.0001	-0.02747 0.4745
idadepai	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	-0.13441 0.0004	0.03968 0.3015
escpai	-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	0.10778 0.0049	-0.18228 <.0001
alturapai	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	1.00000	0.01365 0.7224
numcigpai	-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	0.01365 0.7224	1.00000

Figura 2 – Coeficiente de correlação de Pearson entre as covariáveis.

Na Figura 3 verifica-se qual o comportamento entre a variável resposta (peso do recém nascido) e as covariáveis, este gráfico é importante para determinar alguma transformação no parâmetro das variáveis explicativas.

Percebe-se que as covariáveis diâmetro da cabeça, altura do bebê, idade gestacional e peso da mãe estão relacionadas com o peso de forma linear positiva, enquanto que outras covariáveis não possuem qualquer tipo de relação com a variável resposta, por exemplo escolaridade do pai, numero de cigarros fumados pelo mãe e pai, idade do pai.

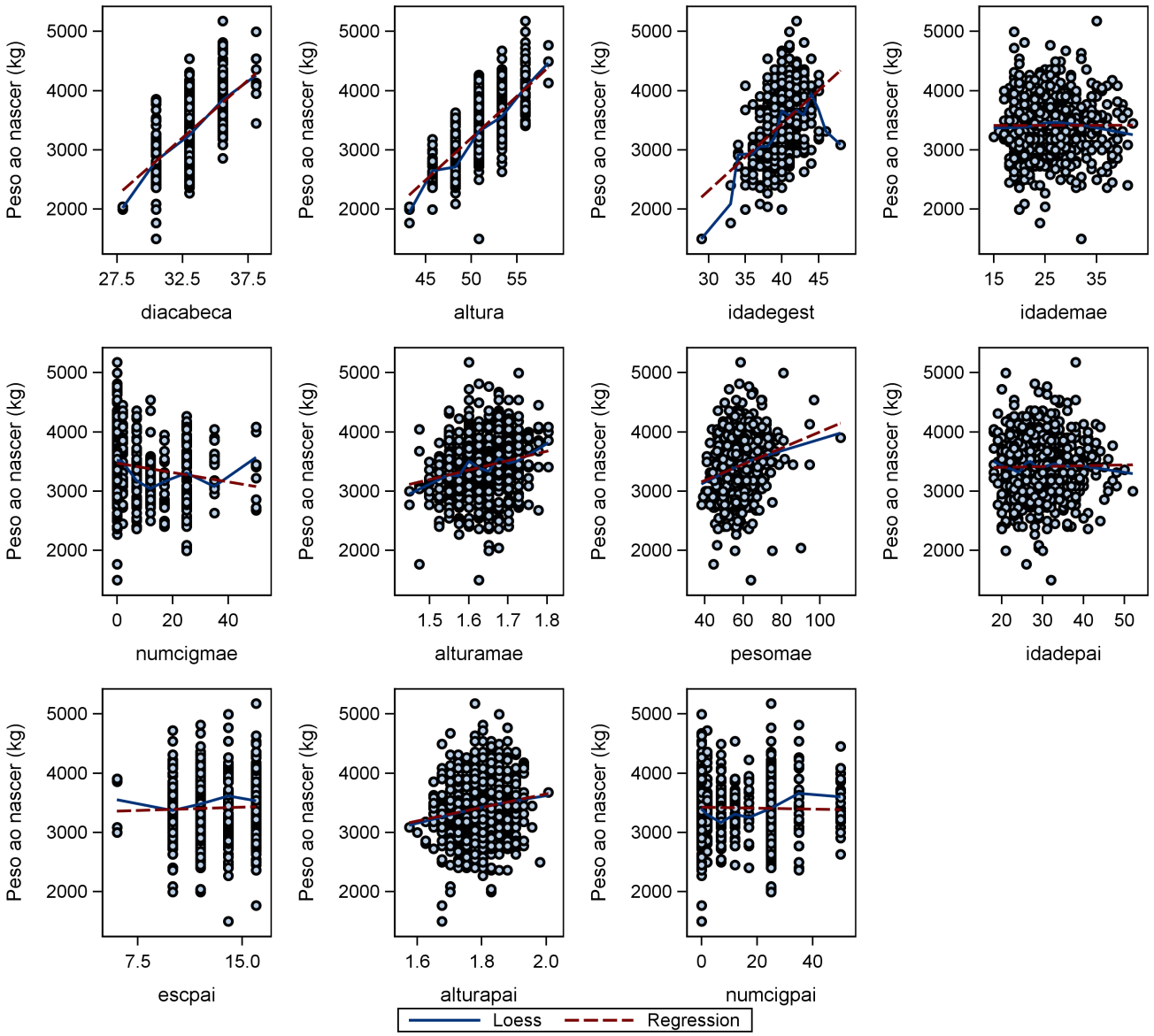

Figura 3 – Gráficos de dispersão do peso ao nascer versus as covariáveis.

3.2 Modelo de regressão completo

Antes da seleção de variáveis, avaliamos o modelo com todas as covariáveis no sentido de identificar eventuais multicolinearidade, heterocedasticidade e realizar, se necessário, alguma transformação na variável resposta.

Na Tabela 2 apresenta-se os resultados da análise de variância. Verifica-se que existem evidências que os efeitos das covariáveis são significativos, isto é, pelo menos uma covariável tem significância estatística. Além disso, foi observado um coeficiente de determinação $R^2 = 0,6492$, isto é, 64% da variabilidade do peso do recém nascido é explicada pelas variáveis explicativas.

Tabela 2 – Análise de variância.

Fonte de variação	Graus de liberdade	Soma de quadrados	Quadrados médios	Estatística F	valor-p
Modelo	11	109361476	9941952	115.23	$<.0001$
Erro	668	57632786	86277		
Total	679	166994262			

Por outro lado, na Tabela 3 podemos observar as estimativas dos parâmetros e seu respectivo erro padrão. Pelo teste t as variáveis que obtiveram significância estatística foram diacabeca, altura, idadegest e pesomae, as demais não foram significativas. Percebe-se que as variáveis diacabeca e altura são medidas do bebê, já as outras são características da mãe. É importante destacar os elevados erros padrão dos parâmetros alturapai e alturamae, sendo assim importante avaliar o motivo, uma vez que altos erros padrões caracterizam estimativas imprecisas.

Tabela 3 – Estimativas dos parâmetros.

Parâmetro	Estimativa	Erro padrão	Estatística t	valor-p	TOL	VIF
Intercepto	-7351.79629	470.41661	-15.63	$<.0001$	–	–
diacabeca	108.71003	8.08148	13.45	$<.0001$	0.76876	1.30080
altura	94.55806	5.43384	17.40	$<.0001$	0.69156	1.44601
idadegest	43.86487	6.46184	6.79	$<.0001$	0.86518	1.15583
idademae	-2.86327	3.62011	-0.79	0.4293	0.32483	3.07853
numcigmae	-1.30839	1.06458	-1.23	0.2195	0.88239	1.13329
alturamae	283.48061	214.91517	1.32	0.1876	0.69149	1.44615
pesomae	4.80317	1.64574	2.92	0.0036	0.72486	1.37958
idadepai	-0.13078	3.24830	-0.04	0.9679	0.32014	3.12359
escpai	5.58147	5.49656	1.02	0.3103	0.86691	1.15353
alturapai	-125.61842	183.38609	-0.68	0.4936	0.84133	1.18859
numcigpai	-0.07451	0.84620	-0.09	0.9299	0.88374	1.13156

Analisando a Figura 4 é possível verificar alguns pontos fora do envelope, o que indica uma má especificação para o modelo de regressão linear.

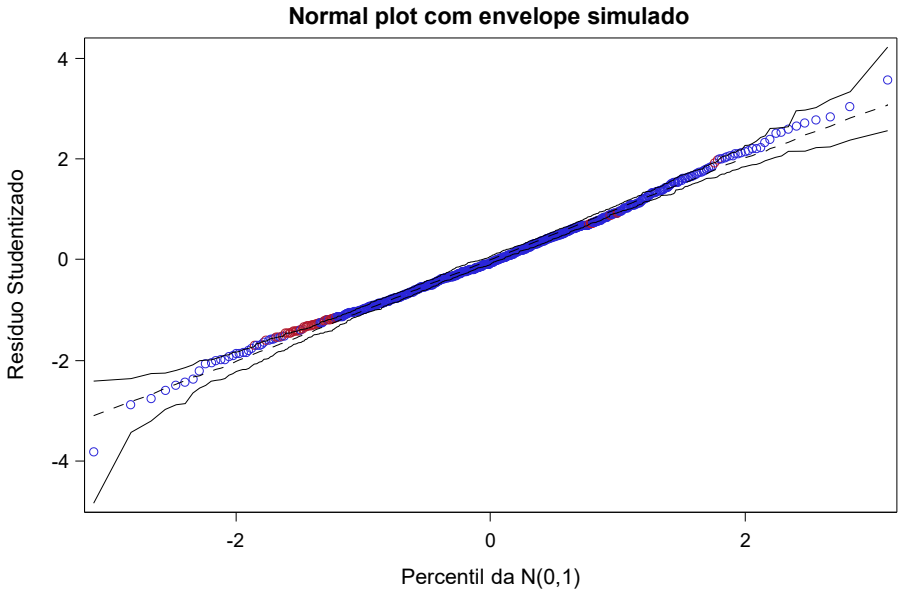

Figura 4 – Gráfico normal de probabilidade com envelope simulado para resíduo.

Além disso, os testes descritos na Tabela 4 a suposição de normalidade não é satisfeita por dois testes. Vale ressaltar que estes testes são baseados na distância entre as distribuições acumulada empírica e teórica.

Tabela 4 – Testes de normalidade.

Teste	Estatística	valor-p
Shapiro−Wilk	0.995125	0.0299
Kolmogorov−Smirnov	0.028417	0.1500
Cramér−von Mises	0.124739	0.0527
Anderson−Darling	0.881033	0.0242

Pela Figura 5 observamos três observações fora do intervalo (-3,3). Percebe-se também, uma certa tendência no centro do gráfico, o que pode ser indicio de heterocedasticidade.

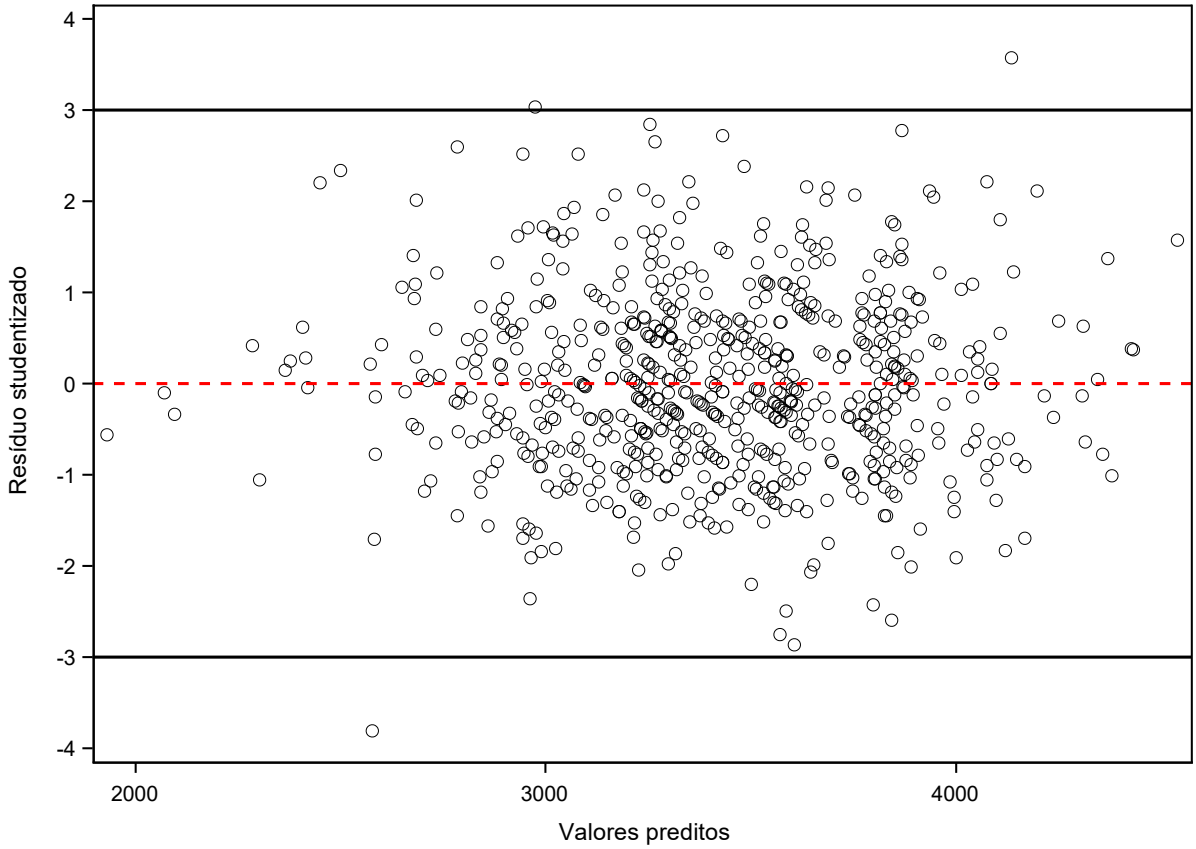

Figura 5 – Gráfico dos resíduos versus os valores preditos.

Os testes de White e Bresch-Pagan indicam que não há heterocedasticidade em nível de sginificância 5%, ver Tabela 5. Todavia, se considerarmos nível de significância 10% o teste de White rejeita a hipótese nula e a tendencia exibida pela Figura 5 é indicio de heterocedasticidade. Conclui-se então que existe indícios de homoscedasticidade.

Tabela 5 – Testes de heterocedasticidade.

Teste	Estatística	valor-p
White	94.33	0.0875
Breusch-Pagan	4.79	0.9407

Na Tabela 6 têm-se os erros padrão estimados utilizando as matrizes de covariância consistente especificadas na subseção 2.5. Ao contrário do que se esperava não há de fato uma diminuição no erro padrão das estimativas.

Tabela 6 – Erros padrão das estimativas conforme as matrizes de covariância consistentes.

Parâmetro	Estimativa	Erro Padrão	HC_0	HC_1	HC_2	HC_3
Intercepto	−7351.79629	470.41661	509.38987	513.94486	518.46947	527.90395
diacabeca	108.71003	8.08148	8.41310	8.48833	8.50279	8.59418
altura	94.55806	5.43384	5.19293	5.23937	5.25846	5.32571
idadegest	43.86487	6.46184	8.29898	8.37319	8.49767	8.70386
idademae	−2.86327	3.62011	3.45224	3.48311	3.49622	3.54133
numcigmae	−1.30839	1.06458	1.03066	1.03987	1.04436	1.05848
alturamae	283.48061	214.91517	211.27981	213.16909	213.96811	216.74281
pesomae	4.80317	1.64574	1.72430	1.73972	1.77104	1.82074
idadepai	−0.13078	3.24830	2.97572	3.00233	3.01669	3.05880
escpai	5.58147	5.49656	5.66516	5.71581	5.74676	5.83134
numcigpai	−125.61842	0.84620	0.81125	0.81851	0.82047	0.82993
alturapai	−0.07451	183.38609	184.62641	186.27735	186.99821	189.43097

Por fim, é importante ressaltarmos que o objetivo desta análise é identificar se alguma característica da mãe e do pai está associada com o peso do recém nascido. Dessa forma, na seleção de variáveis não será considerado a altura do bebê, uma vez que é uma medida dele.

3.3 Seleção de variáveis

Prosseguindo com a análise, nesta seção iremos descrever os critérios e a forma que procedeu-se para selecionar o modelo mais parcimonioso. Tendo em vista a obtenção de um modelo parcimonioso, avaliamos os critérios descritos para todos os 2^p modelos possíveis, nesta análise $p = 11$, através do software SAS.

A Tabela 7 apresenta os as medidas de comparação para os seis "melhores" modelos. A coluna Rank, indica qual modelo foi melhor conforme todos os critérios. Ressalta-se que com exceção do critério C_p de Mallows do primeiro modelo, todas as outras medidas dos outros modelos são muito próximas.

Tabela 7 – Medidas de comparação para principais modelos.

Rank	N° de variáveis	Varíaveis no modelo	C_p	R_a^2	AIC	BIC
1	6	diacabeca idadegest numcigmae alturamae pesomae alturapai	5.0857	0.4924	7986.6428	7988.8284
2	7	diacabeca idadegest idademae numcigmae alturamae pesomae alturapai	6.0413	0.4924	7987.5837	7989.8210
3	7	diacabeca idadegest numcigmae alturamae pesomae idadepai alturapai	6.5813	0.4920	7988.1316	7990.3559
4	7	diacabeca idadegest numcigmae alturamae pesomae escpai alturapai	6.8465	0.4918	7988.4003	7990.6183
5	7	diacabeca idadegest numcigmae alturamae pesomae alturapai numcigpai	6.8715	0.4918	7988.4257	7990.6430
6	5	diacabeca idadegest numcigmae alturamae pesomae	6.9617	0.4902	7988.5589	7990.6484

Portanto o modelo selecionado é dado por

$$Y_i = \beta_0 + \beta_1 X_{i1} + +\beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} + \beta_6 X_{i6} + \varepsilon_i \tag{11}$$

sendo:

- X_{i1}: diâmetro da cabeça do recém nascido em centímetro;
- X_{i2}: idade gestacional em semanas;
- X_{i3}: número de cigarros que a mãe fuma por dia;
- X_{i4}: altura da mãe em metros;
- X_{i5}: peso da mãe antes da gravidez em kg;
- X_{i6}: altura do pai em metros;

3.4 Modelo de regressão selecionado

Tendo em vista o modelo de regressão selecionado, nesta seção iremos avaliar os pressupostos e analisar os resultados encontrados.

As estimativas dos parâmetros do modelo final encontram-se na Tabela 8. Podemos observar que todas as covariáveis são significativas em nível de significância $\alpha = 5\%$. Analisando as estimativas pode-se notar que somente a covariável número de cigarros que a mãe fuma por dia tem efeito negativo no peso do bebê, isto é, mantendo as demais covariáveis fixas a cada cigarro fumado por dia da mãe espera-se um decréscimo de 4 gramas no peso do bebê. Por outro lado, as demais variáveis tem efeito positivo no peso do recém nascido, como por exemplo o peso da mãe, em que peso esperado do recém nascido aumenta em 6,28 gramas, mantendo as outras covariáveis fixas. Além disso, o R_a^2 do modelo ajustado foi de 0.4924, indicando que 49,24% da variação do peso do recém nascido são explicadas pelas covariáveis consideradas.

No que tange a multicolinearidade, pode-se observar pelo fator de inflação da variância (VIF) que o modelo é bem comportado, ver Tabela 8.

Tabela 8 – Estimativas do parâmetros e VIF do modelo final.

Parâmetros	Estimativa	Erro padrão	Estatística t	valor-p	VIF
intercepto	−6722.02185	555.44428	−12.10	<.0001	–
diacabeca	160.00004	9.01540	17.75	<.0001	1.11873
idadegest	71.51283	7.51878	9.51	<.0001	1.08143
numcigmae	−4.07657	1.21618	−3.35	0.0008	1.02211
alturamae	510.87822	256.51248	1.99	0.0468	1.42371
pesomae	6.28630	1.94551	3.23	0.0013	1.33230
alturapai	419.99399	213.02777	1.97	0.0491	1.10840

Com o intuito de verificar possíveis afastamentos das suposições feitas para o modelo, a Figura 6 apresenta os gráficos de probabilidade normal com envelope simulado. Podemos notar que todos os resíduos permanecem dentro do envelope, indicando que o modelo esteja bem adequado.

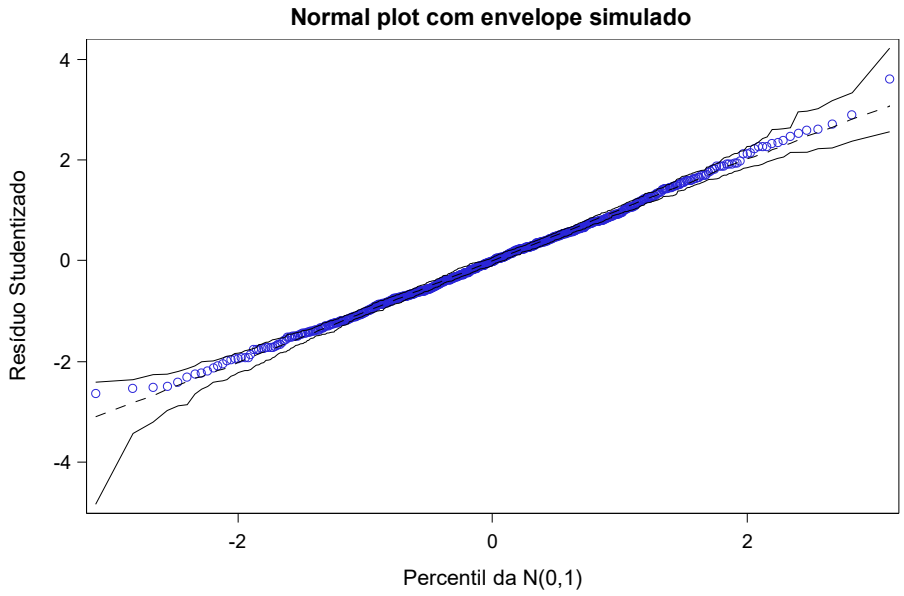

Figura 6 – Gráfico normal de probabilidades com envelope simulado.

Por outro lado, na Figura 7 podemos observar que somente um resíduo esta fora do intervalo $(-3, 3)$, além disso é possível verificar que os resíduos se distribuem de forma aleatória, sem nenhum padrão expressivo.

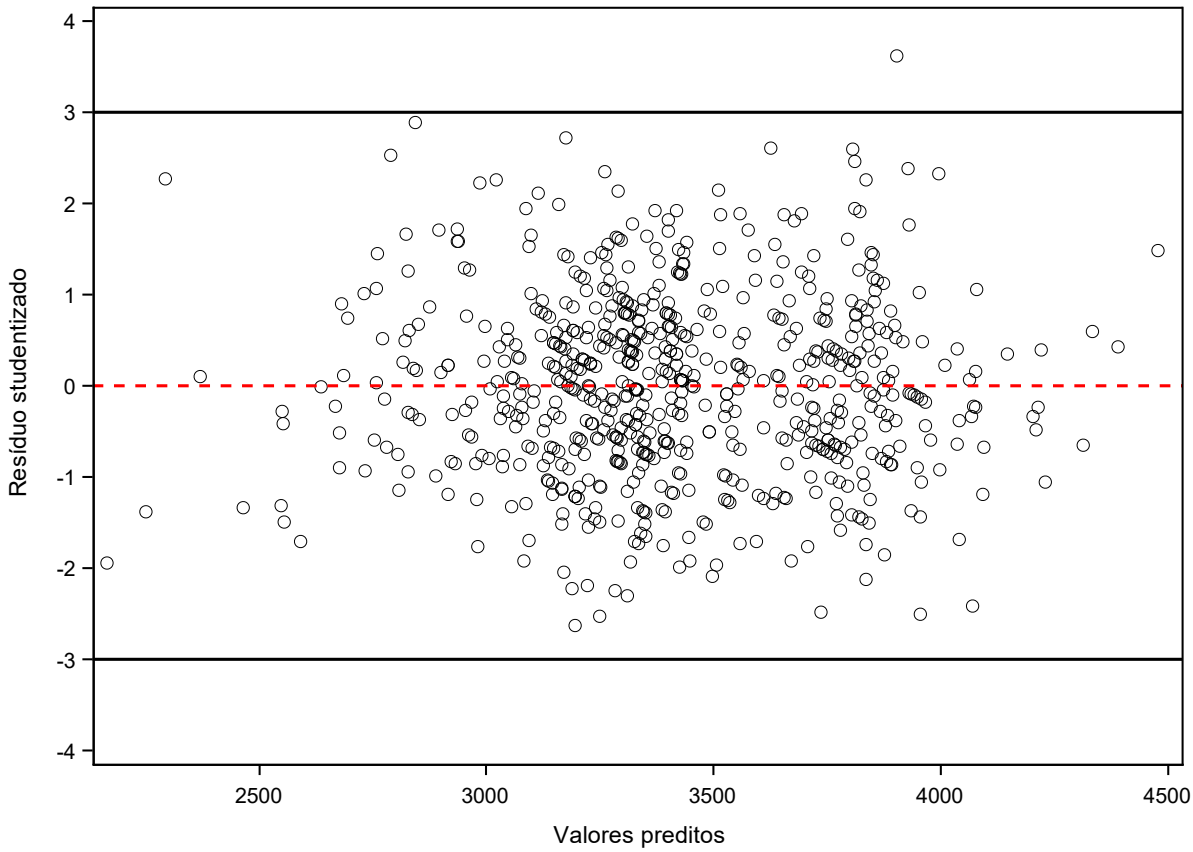

Figura 7 – Gráfico dos resíduos versus os valores preditos.

De acordo com os testes de heterocedasticidade apresentados na Tabela 9, não rejeitamos a hipótese nula, isto é, os resíduos possuem variância constante.

Tabela 9 – Testes de heterocedasticidade para o modelo final.

Teste	Estatística	valor-p
White	34.00	0.1660
Breusch-Pagan	8.94	0.1768

Do diagnóstico de influência, apresentado pelas Figuras 8, 9, 10 e 11, pode-se observar que:

(i) a observação 170 apresenta-se como possível outlier;

(ii) os elementos h_{ii} da diagonal da matriz leverage mostram que a observação 167 destoa das demais como sendo ponto influente, logo deve ser investigada;

(iii) o maior valor da distância de Cook é $D_{167} = 0,0413$, as observações 9 e 60 também possuem valores altos;

(iv) inspecionando os DFFit's temos que as observações 9, 60 e 167 destacam-se das demais, indicando-se serem possíveis pontos influentes;

(v) inspeção dos DFBeta's também mostram várias observações excedendo o ponto de corte 0,076, dar-se destaque as observações 9, 60 e 167 a quais apresentam grande efeito em pelo menos uma das três estimativas dos parâmetros;

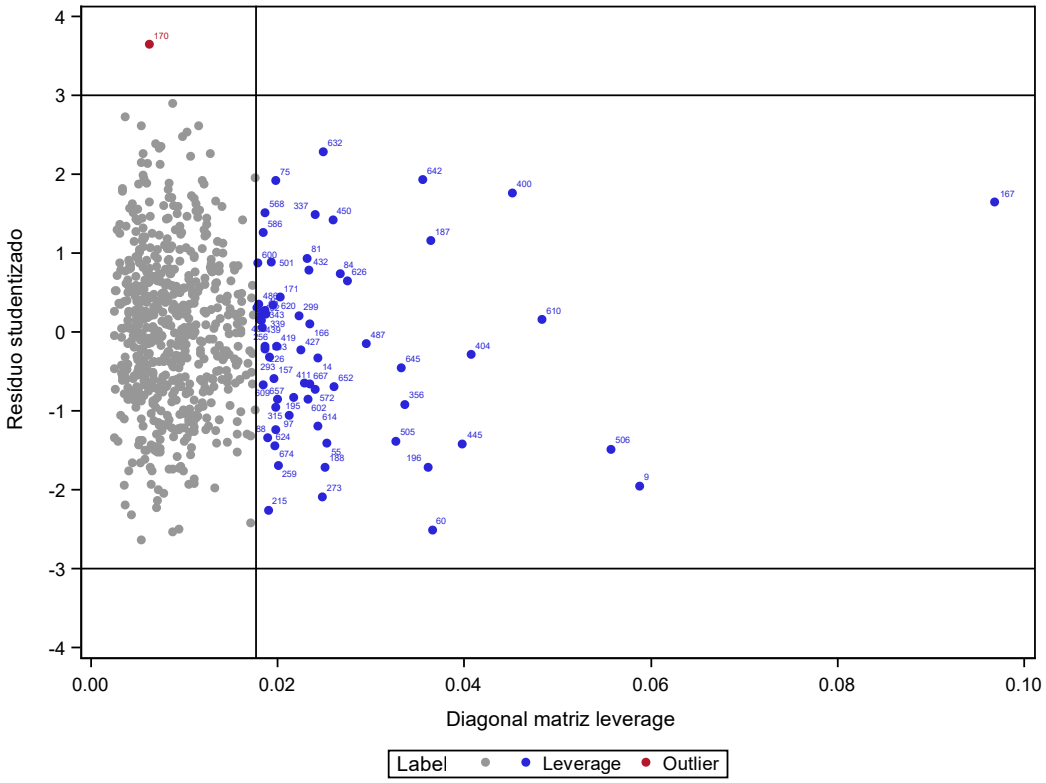

Figura 8 – Gráfico dos elementos da diagonal da matriz leverage.

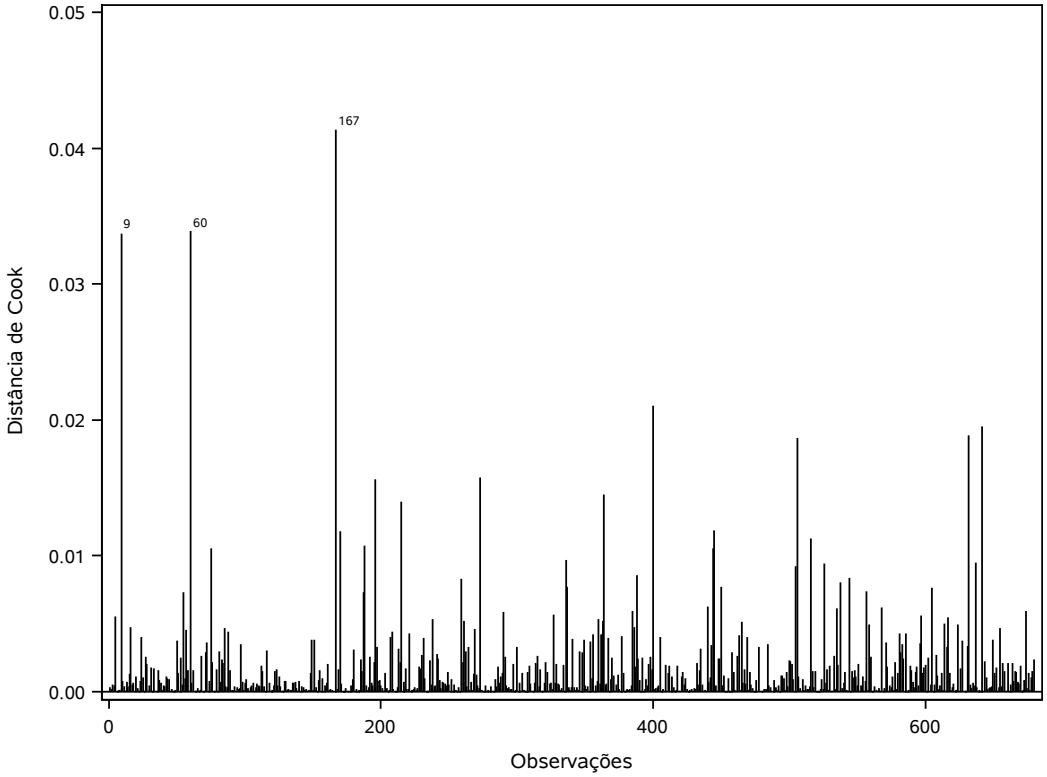

Figura 9 – Gráfico das distâncias de Cook.

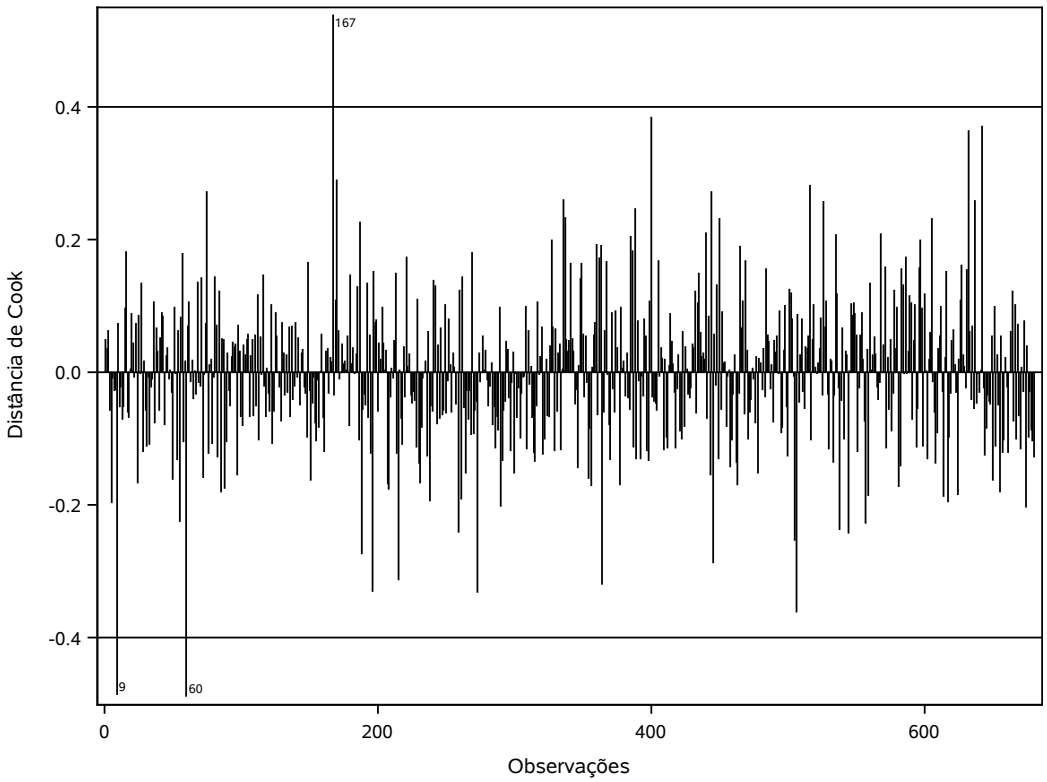

Figura 10 – Gráfico dos DFfit's.

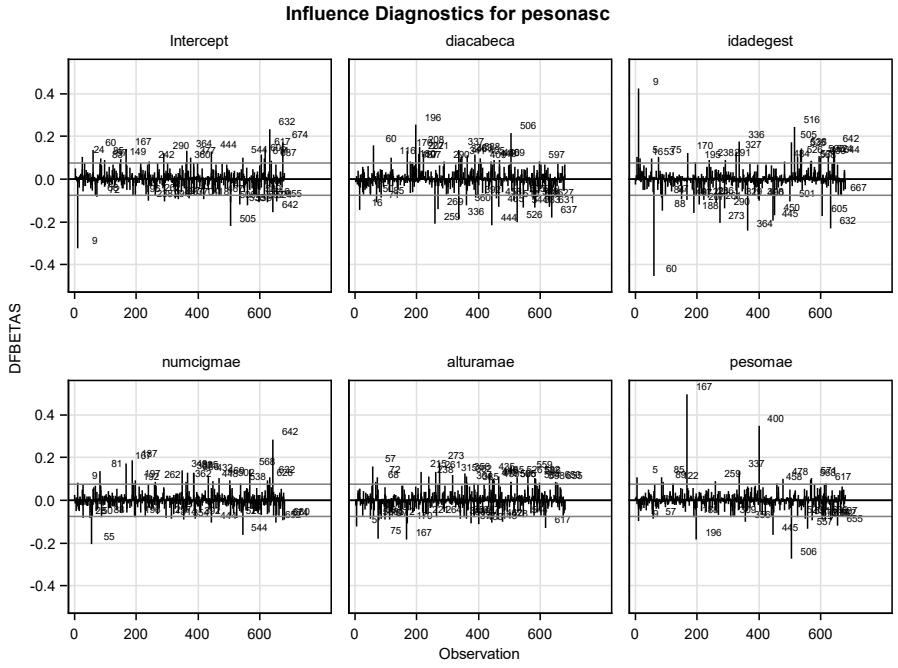

Figura 11 – Gráfico dos DFbeta's.

Claramente, as observações 9, 60 e 167 são as que merecem maior atenção em nossa análise. Para investigar o efeito dessas observações no modelo de regressão observe os resultados apresentados na tabela a seguir.

Tabela 10 – Diagnóstico das observações influentes

Modelo	β_0	β_1	β_2	β_3	β_4	β_5	β_6	QME	R_a^2
Todas	−6722.02	160	71.51	−4.07	510.87	6.28	419.99	13829014	0.4924
Sem obs 167	−6799.55	160.07	72.24	−4.28	557.96	5.32	433.75	13844755	0.4937
Sem obs 60	−6797.21	158.57	74.90	−4.14	530.78	6.29	395.90	13941542	0.4968
Sem obs 9	−6543.30	159.65	68.33	−4.17	510.90	6.47	392.10	13296080	0.4839
Sem obs 9, 60 e 167	−6704.60	158.35	72.58	−4.43	576.59	5.53	383.73	13416018	0.4894

Note, que retirar a observação 167 produz grandes mudanças no parâmetros β_4, β_5 e β_6, portanto esta observação exerce razoável influências nos coeficientes. Por outro lado, a retirada da observação 60 produz notáveis mudanças nos coeficientes β_4 e β_6. Por fim, a retirada da observação 9 mudanças significativas no parâmetro β_6. Conclui-se, assim, que as observações 9, 60 e 167, mais fortemente a 167, influenciam no ajuste do modelo, além disso, o coeficiente de determinação não sofre mudanças significativas com a retirada das observações. Neste contexto, investigações futuras, realizadas junto ao pesquisador, podem revelar razões para a retirada de uma, ou todas, as observações da análise.

4 Conclusão

Neste trabalho foi realizado uma análise de regressão linear múltipla de um estudo transversal, tendo em vista identificar se alguma característica do pai ou da mãe esta associada com o peso do recém nascido. O modelo ajustado apresentou seis variáveis explicando o peso de recém nascido, sendo somente uma variável do pai (altura), uma variável do próprio recém nascido (diâmetro da cabeça) e as demais, quatro variáveis da mãe. Foi observado que somente a covariável número de cigarros fumados pelo mãe por dia exerce efeito negativo no peso do bebê. Além disso, embora as covariáveis altura da mãe e do pai sejam significativas no modelo, elas apresentaram elevados erro padrão, caracterizando estimativas não precisas. É importante destacar que o peso da mãe e sua idade gestacional influenciam positivamente o peso do recém nascido, isto é, quanto maiores, maior é esperado o peso.

No que tange a análise de diagnóstico, foi visto pelos diversos gráficos e testes que o modelo de regressão ajustado demonstrou atender os pressupostos. Contudo, somente 50% da variação do peso do recém nascido esta sendo explicada pelas covariáveis. Além disso, foi possível observar claramente que as observações 9, 60 e 167, mais fortemente a 167, influenciam no ajuste do modelo, isto é, são pontos influentes. Dessa forma, investigações futuras, realizadas junto ao pesquisador, podem revelar razões para a retirada de uma, ou todas, as observações da análise.

Referências

BELSLEY, D. A.; KUH, E.; WELSCH, R. E. **Regression diagnostics: Identifying influential data and sources of collinearity**. [S.l.]: John Wiley & Sons, 2005. Citado na página 6.

COOK, R. D. Influential observations in linear regression. **Journal of the American Statistical Association**, Taylor & Francis, v. 74, n. 365, p. 169–174, 1979. Citado na página 6.

DEMÉTRIO, C. G. B.; ZOCCHI, S. S. **Modelos de Regressão**. [S.l.], 2011. Citado na página 5.

GIOLO, S. R. **Análise de Regressão Linear**. [S.l.], 2007.

INSTITUTE, S. **Base SAS 9.4 Procedures Guide**. [S.l.]: SAS Institute, 2014. Citado na página 2.

NETER, J. et al. **Applied linear statistical models**. [S.l.]: Irwin Chicago, 1996. Citado na página 3.

SILVA, P. H. D. da. A sas macro to generate normal and half-normal plots with simulated envelope. **Rev. Bras. Biom.**, v. 32, n. 3, p. 460–473, 2014.

WEISBERG, S. **Applied linear regression**. [S.l.]: John Wiley & Sons, 2005.