Universidade Estadual de Campinas

Instituto de Matemática, Estatística e Computação Científica - IMECC

Trabalho Final

Aluno: Eduardo Vargas Ferreira

Professora: Dra. Samara Flamini Kiihl

1 Introdução

O peso ao nascer determina, em grande parte, a sobrevida da criança no primeiro ano de vida e, mais frequentemente, no primeiro mês (Victoria et al., 1985). Diante disto, não são escassos na literatura trabalhos que mostram a associação de diferentes variáveis com o peso da criança ao nascer. Um importante estudo sobre a prevalência de baixo peso nos recém-nascidos deve-se a Rush e Cassano (1983), que atribui à idade das mães como um fator determinante no peso do bebê. Autores como Van Den Berg e Yerushalmy (1966) referem-se ao hábito de fumar como um fator fortemente associado ao baixo peso do recém-nascido, além de aumentar a incidência de abortos e morte perinatal. Outros estudos (veja, por exemplo, Meyer e Tonascia, 1977) suportam a visão de que o tabagismo materno retarda o crescimento fetal, ou seja, filhos de mães fumantes tendem a pesar 150g a 350g a menos ao nascer, do que aqueles nascidos de não fumantes. Com isso em mente, o presente trabalho tem por objetivo verificar a validade dessa opinião. Os dados utilizados na análise provem do *Child Health and Development Studies* (CHDS), um estudo de todas as crianças nascidas no Kaiser Foundation Hospital em Oakland, entre 1960 e 1967. Considerar-se-á o peso do recém-nascido como a variável resposta, e as variáveis regressoras: peso dos pais, número médio de cigarros por dia, nível de educação etc. (para descrição de todas as variáveis, veja Tabela 2).

2 Metodologia

A base das metodologias estudadas encontra-se em Hastie, Tibshirani & Friedman (2001). Tal trabalho introduz as ferramentas utilizadas, relacionadas ao objeto de estudo. Com base nas discussões apresentadas nessa referência, iniciou-se a Seção 3 com uma análise exploratória das variáveis a fim de selecionar as mais relevantes para a construção dos modelos (a saber, Regressão Linear (através dos métodos automáticos de seleção de variáveis *Forward*, *Backward* e *Stepwise*), Regressão Ridge e Regressão Lasso). Segue-se-lhe a Seção 4 dedicada às questões inerentes a escolha do melhor modelo de acordo com o menor erro de previsão. Para tanto, 80% dos dados foram utilizados para o ajuste e os outros 20% para validação. Finalmente, a Seção 5 ressalta-se aspectos importantes e conclusões do trabalho.

3 Análise Exploratória

Em um primeiro momento, após uma observação particular do conjunto de dados, pode-se verificar a existência de caracteristicas constantes para todos os grupos, quais sejam, *pluralty*, *outcome* e *sex*. A variável *pluralty* assume valores iguais a 5 e designa que o feto é único. Se existissem casos de gêmeos trigêmes etc, indiscutivelmente, essa variável seria importante para explicar o peso do bebê, visto que o aumento do número de fetos pode acarretar na diminuição de seus pesos. A variável *outcome* indica que o bebê sobreviveu até 28 dias após o parto, houvesse casos na amostra em que o bebê não sobreviveu 28 dias esta variável seria incluida nas análises (considerando a hipótese de que o hábito do fumo altera a saúde do bebê, e filhos de mães fumantes têm maior chance de morrer antes dos 28 dias do que de mães que nunca fumaram). A variável *sex* indicou que todo o conjunto de dados é composto por bebês do sexo masculino. É sabido que meninas tendem a ser menos pesadas do que os meninos ao longo de toda a vida, todavia, como não se observa casos de meninas nos dados, não tem sentido acrescentar tal variável ao modelo. Adicionalmente, importa considerar que a variável *marital* revela que 98% das mulheres são casadas, portanto essa variável não terá importância na construção do modelo. Ademais, é escuso supor que o peso dos recém-nascidos pode ser explicado pela mãe ser ou não casada (possivelmente, por variáveis correlatas). Na Figura 1 encontra-se o *Box-Plot* do peso do bebê versus o hábito de fumar da mãe.

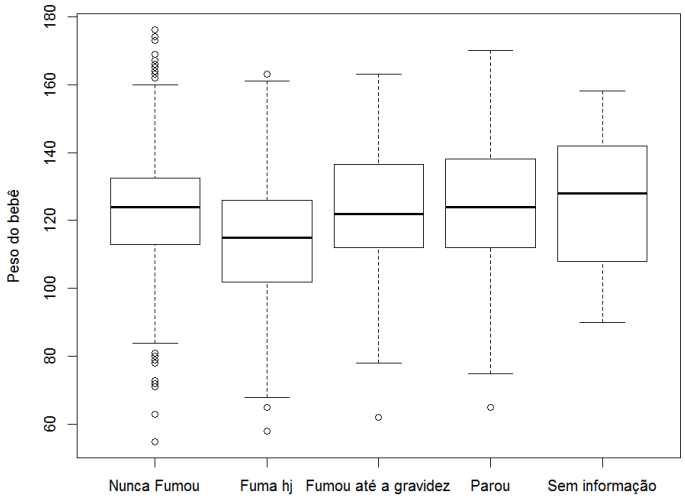

Figura 1: *Box-Plot do peso do bebê ao nascer versus o hábito de fumar*

Depois de breve comparação dos fatores, nota-se que bebês cujas mães fumaram durante a gravidez têm um peso menor do que aqueles com mães não fumantes ou que pararam algum dia. Além disso, considerando o fato de que o período em que a mãe fumou influencia no peso do seu filho, descreve-se na Figura 2 o *Box-Plot* do peso do bebê versus o tempo em que a mãe parou de fumar (no qual "até 1" significa que a mãe parou de fumar em até 1 ano antes da gravidez,"entre 1 e 2" significa que a mãe parou de fumar entre 1 e 2 anos antes da gravidez, e "?" denota sem informação disponível). Destaca-se que, em geral, o peso dos bebês é substancialmente menor quando a mãe fuma até hoje, e mães que pararam de fumar em até dois anos antes da gravidez apresentam recém-nascidos com maior peso.

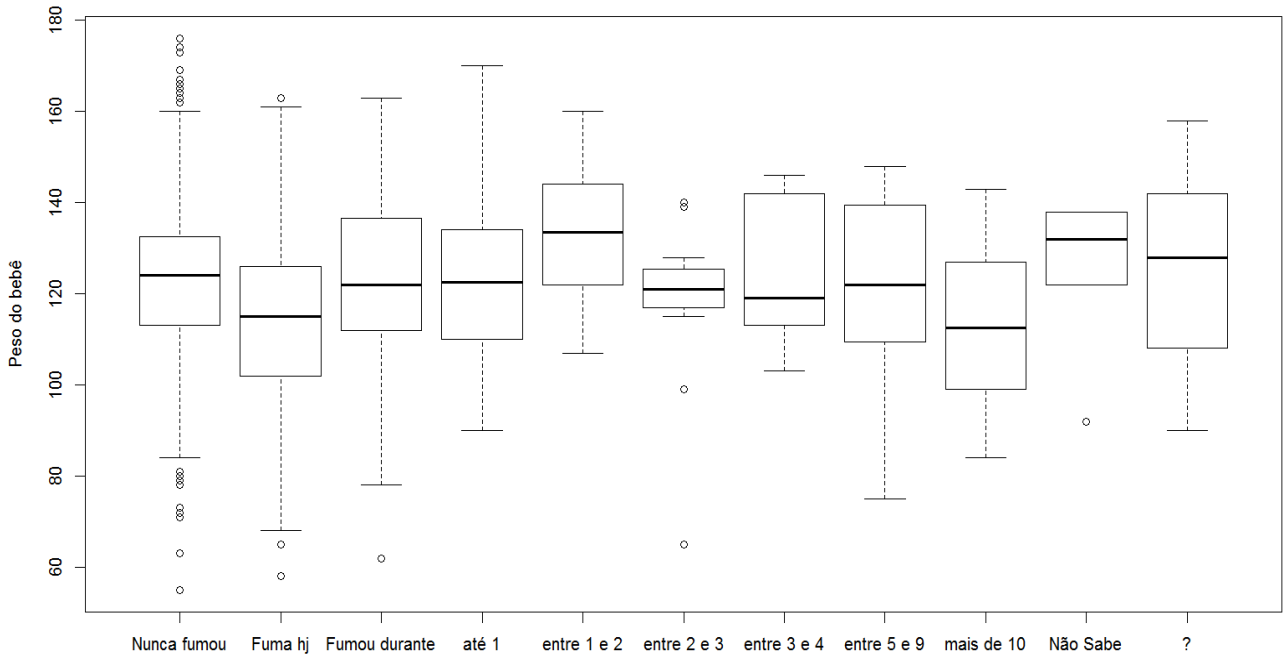

Figura 2: *Box-Plot do peso do bebê versus o tempo que a mãe parou de fumar*

Outra variável intimamente relacionada ao hábito de fumar é a quantidade de cigarros consumidos por dia, como ilustrado na Figura 3, apresentando o *Box-Plot* do peso dos bebês por números de cigarros consumidos ao dia (em que "1-4" representa que a mãe fuma ou fumava de 1 a 4 cigarros por dia, e "?" remete à mães na qual não se sabe o número de cigarro que consumiam). Destaca-se que o peso de bebês de mães que nunca fumaram é ligeiramente maior, exceto para as mães que não foi obtida essa informação (denotada por "?").

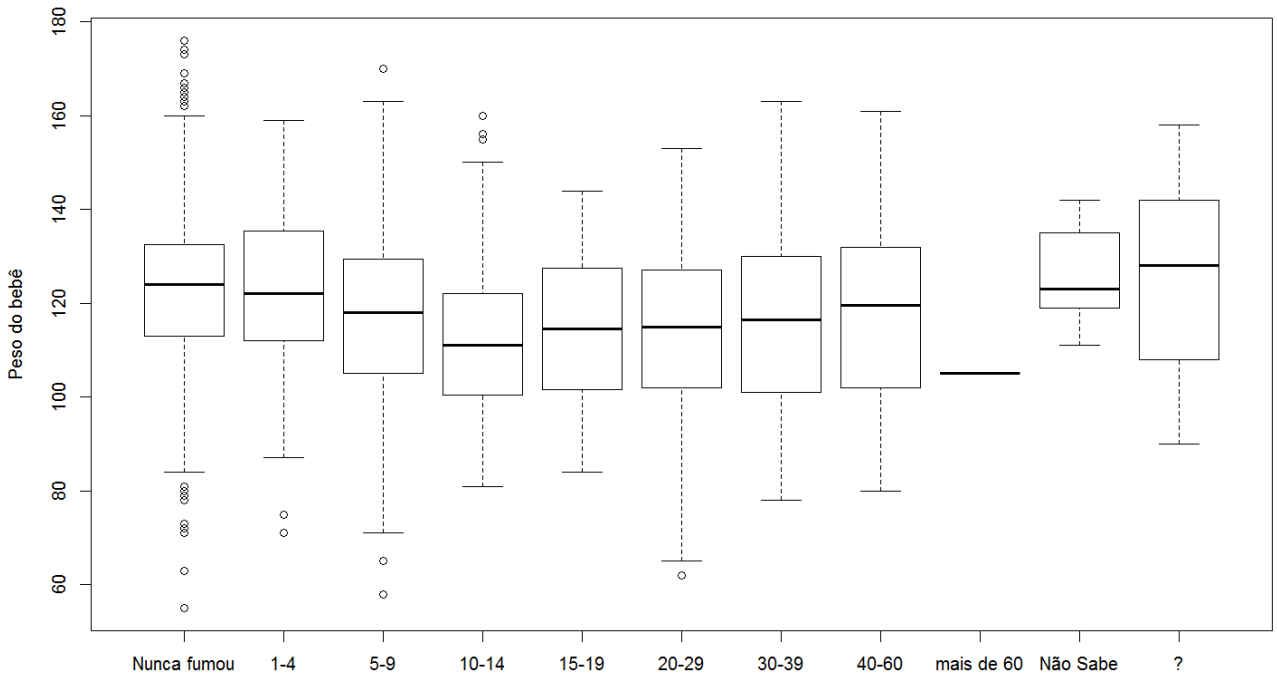

Figura 3: *Box-Plot do Peso do Bebê versus o número de cigarros que a mãe consome por dia*

Este é, com efeito, um ponto em que dificuldades surgem, não é simples tirar conclusões dos três *Box-Plots* supracitados, ao passo que não se sabe como se comporta a interação entre essas variáveis. Diante disto, considerou-se a possibilidade de inclusão no modelo a interação entre o hábito de fumar (*smoke*) e número de cigarros diários (*number*) e entre o tempo em que parou de fumar (*time*) e o hábito de fumar (*smoke*). Entretanto, é digno de nota que, após realizar a regressão da variável *smoke* e *time*, os resultados permitiram concluir a alta correlação entre elas. Por esse motivo incluir-se-á apenas uma delas no modelo. Optou-se pela variável *smoke*, haja vista que *time*, ainda que seja mais informativa, possui 10 categorias, podendo, dessa forma, contaminar o modelo com informações e parâmetros, possivelmente, desnecessários. Cumpre dizer que, aproximou-se a variável *number*, que contém o número de cigarros diários, pelo número médio de cigarros no dia (*number1*), que vive no conjunto $\{0, 2.5, 7, 12, 17, 19.5, 34.5, 50, 60\}$, facilitando assim a interpretação. O conjunto de dados apresenta outras variáveis, tais como: tempo de gestação do bebê, peso, altura, raça e educação dos pais, além de sua renda e idade. Neste momento, concentrou-se o foco na análise das interações entre cada uma dessas variáveis. Na Figura 4 exibe-se o gráfico do peso do bebê versus renda. Observa-se que casais de "baixa" renda tendem a ter filhos com peso menor do que casais com renda maior.

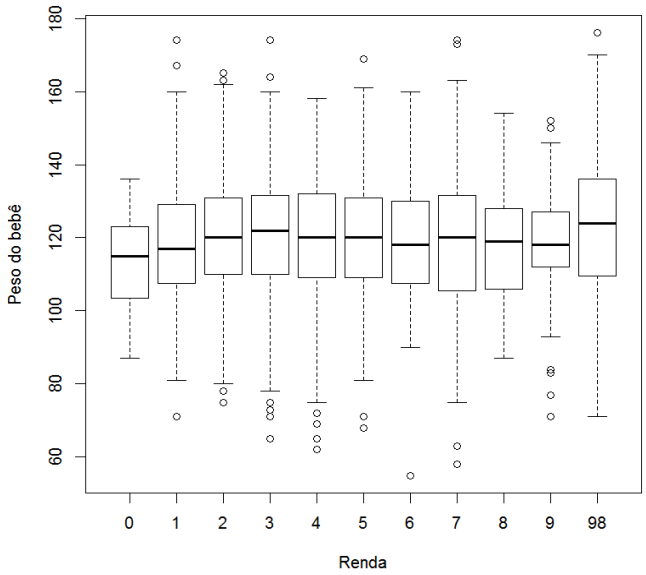

Figura 4: *Box-Plot do Peso do Bebê versus a Renda*

Da Figura 5 tem-se que a relação linear entre renda e educação é baixa.

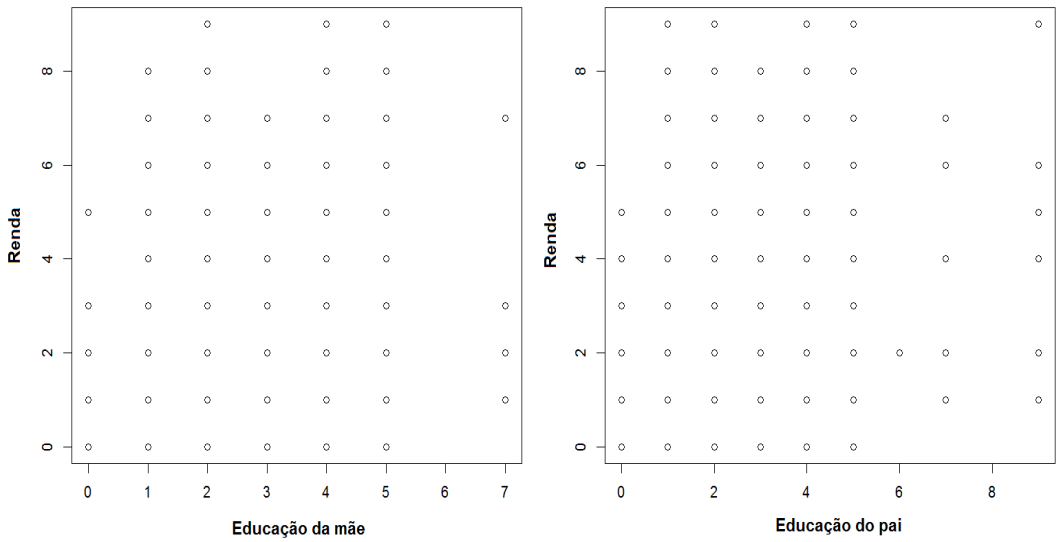

Figura 5: *Gráficos da Renda versus educação da mãe a do pai, respectivamente*

Além disso, através do *Box-Plots* da Figura 6, constata-se que há evidências de que o nível de educação, tanto do pai quanto da mãe, não ajudam muito a explicar o peso do recém-nascido. Concretamente, não era esperada tal relação, por mais que existam fatores (outros) relacionados ao grau de educação que influencie na variável resposta.

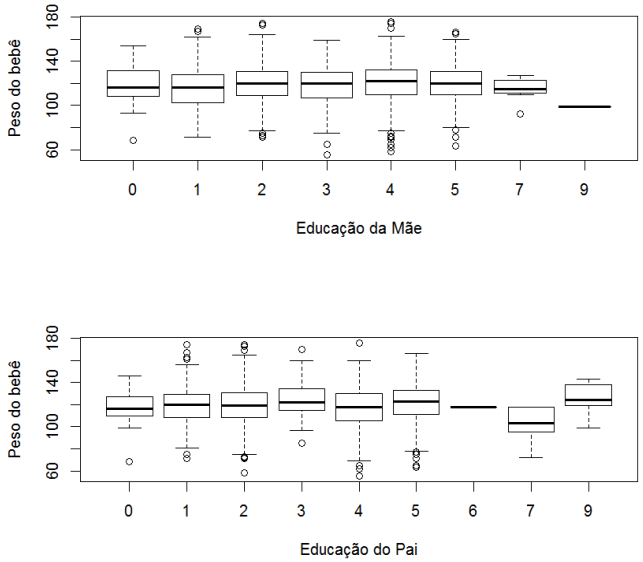

Figura 6: *Box-Plots do Peso do Bebê versus a educação da mãe e do pai, respectivamente*

Como ilustrado na Figura 7, verifica-se a existência de uma relação entre a raça dos pais e o peso do recém-nascido. Ressalta-se que bebês de mães e/ou pais negros ou asiáticos apresentam um peso menor em relação às outras raças. Importa dizer que, como a maior parte dos casais pertence à mesma raça, a inclusão de uma dessas variáveis (raça do pai ou raça da mãe) já é suficiente, sendo não informativa (e inútil) a permanência de ambas no modelo. Dito isto, utilizar-se-á a raça da mãe para construir os modelos.

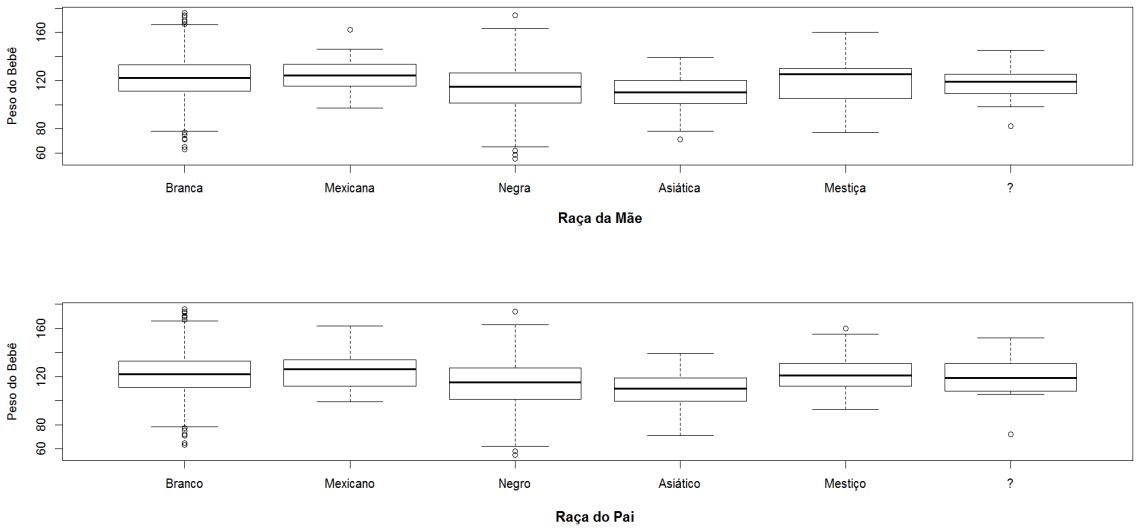

Figura 7: *Box-Plot do Peso do bebê versus a raça da mãe e do pai, respectivamente*

Outra importante variável considerada no estudo é o tempo de gestação. Após breve avaliação da Figura 8, verifica-se uma relação positiva e linear com o peso do bebê. Acresce notar, a existência de dois recém-nascidos (assinalados) pré-maturos, todavia, seus pesos apresentam-se próximos da média.

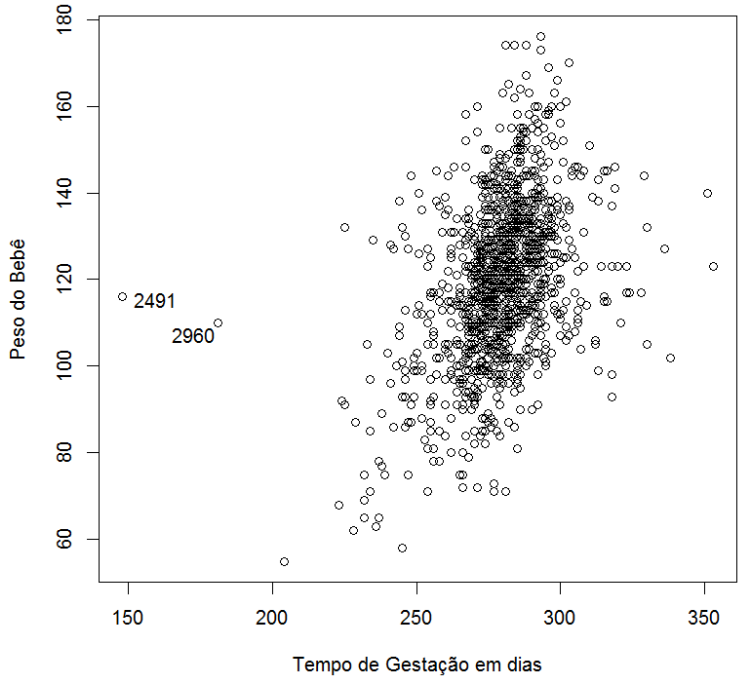

Figura 8: *Gráfico do Peso do Bebê versus o tempo de gestação em dias*

Diversos estudos ladeiam a questão da relação linear positiva entre peso e altura. Diante do apresentado, parece ser suficiente considerar apenas uma das variáveis para explicar o peso do bebê. Contudo, a observação da Figura 9 revela que tais medidas não parecem ser correlacionadas com a variável resposta, por outras palavras, provavelmente a altura e peso dos pais não serão úteis ao modelo, a não ser que na presença de outras regressoras elas tenham um efeito significativo. Ademais, nota-se uma elevada quantidade de dados faltantes, em torno de 40%, referente ao pai. Por este motivo, não será incluído seu peso e altura no modelo. Da exposição feita na análise exploratória chegou-se ao conjunto de regressoras candidatas a entrarem no modelo, nomeadamente

$$\boldsymbol{X} = \{gestation, race, ht, wt.1, smoke, number1\}$$

em que as variáveis *smoke* e *race* são consideradas fatores.

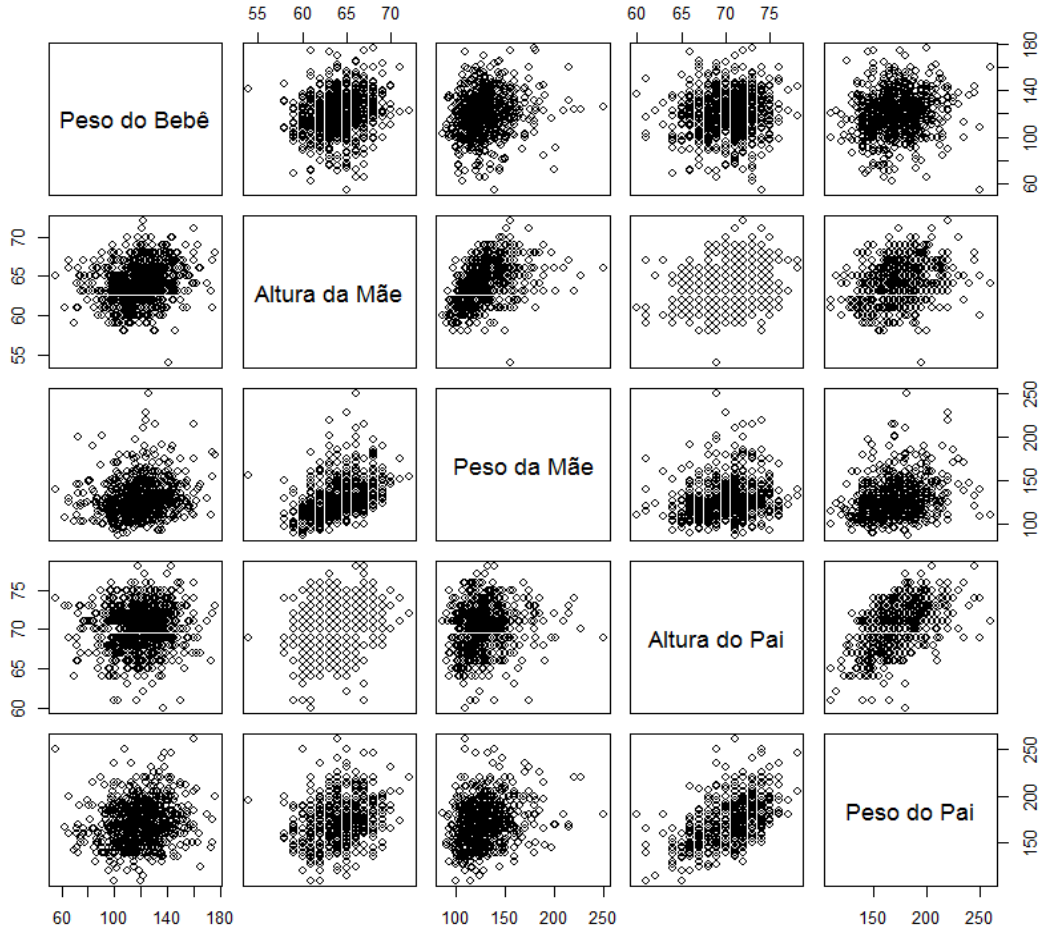

Figura 9: *Matriz de gráficos com as variáveis: peso do bebê, altura da mãe, peso da mãe, altura do pai e peso do pai*

A quantidade de dados faltantes para cada uma das variáveis gira em torno de 2%, o que não acarreta em uma perda significativa dos dados. Como mencionado, as variáveis altura e peso do pai possuem uma porcentagem de dados faltantes de, aproximadamente, 40%, em vista disto não as utilizou. À semelhança da variável anterior, renda, por apresentar 10% de dados faltantes, o que conduz a uma perda relativamente grande no conjunto de dados, foi igualmente excluída do modelo.

Com um novo conjunto de dados, sem as informações faltantes, o número de observações diminuiu de 1236 para 1120, redução de 9%. Destas, 224 observações foram destinadas a validação e 896 para o ajuste dos modelos propostos. Em princípio, não considerou-se nenhuma transformação nas regressoras, e a necessidade de tais modificações será verificada depois do ajuste alguns modelos.

4 Escolha dos modelos

4.1 Modelo 1: Modelo Total

Primeiramente, considerou-se o modelo com todas as regressoras, incluindo a interação entre *number1* e *smoke*, obtendo o seguinte modelo ajustado

$$\begin{aligned}\mathbb{E}(Y|\boldsymbol{X}=\boldsymbol{x}) &= 0.01gestation + 39.82I_{\{smoke=0\}} + 32.50I_{\{smoke=1\}} + 39.10I_{\{smoke=2\}} + 37.63I_{\{smoke=3\}} \\ &+ 4.40I_{\{race=6\}} - 8.40I_{\{race=7\}} - 9.71I_{\{race=8\}} - 2.98I_{\{race=9\}} + 0.07wt.1 + 1.11ht \\ &+ 0.07number1 - 0.14number1I_{\{smoke=1\}} - 0.05number1I_{\{smoke=2\}}\end{aligned}$$

Importante sublinhar que atribuiu-se *race=0* e a interação *smoke=0:number1* como referência. E como *smoke=3:number1* é combinação linear das outras interações, não foi incluído no modelo.

4.2 Modelo 2: Modelo 1 Simplificado

Baseado no **Modelo 1**, observou-se que algumas varáveis são não significativas, a saber, *gestation*, *wt.1*, *number1* e *smoke:number1*. Resultando no seguinte modelo

$$\begin{aligned}\mathbb{E}(Y|\boldsymbol{X}=\boldsymbol{x}) &= 35.76I_{\{smoke=0\}} + 26.34I_{\{smoke=1\}} + 34.98I_{\{smoke=2\}} + 35.48I_{\{smoke=3\}} \\ &+ 4.51I_{\{race=6\}} - 7.70I_{\{race=7\}} - 10.71I_{\{race=8\}} - 3.25I_{\{race=9\}} + 1.40ht\end{aligned}$$

Constatou-se que as variáveis *smoke=1* e *race=9* não influenciam significativamente. Esta última permaneceu não significativa desde o **Modelo 1**, levando-se a concluir que a raça mista tem o mesmo efeito no peso do recém-nascido do que a raça de referência (raça branca).

4.3 Modelo 3: Modelo do Hábito de Fumar

Outra possibilidade considerada foi empregar apenas as variáveis relacionadas ao hábito de fumar, e o modelo, em termos compactos, se resume por

$$\begin{aligned}\mathbb{E}(Y|\boldsymbol{X}=\boldsymbol{x}) &= 123.41I_{\{smoke=0\}} + 117.29I_{\{smoke=1\}} + 121.95I_{\{smoke=2\}} + 123.59I_{\{smoke=3\}} \\ &+ 0.08number1 - 0.18number1I_{\{smoke=1\}} - 0.05number1I_{\{smoke=2\}}\end{aligned}$$

Neste caso, *number1* e suas interações foram não significativas. Visto que o modelo exclui variáveis sabidamente importantes, nomeadamente, peso e raça da mãe, por exemplo. Este foi desconsiderado no processo de validação.

4.4 Modelo 4: Modelo *Backward*

O modelo selecionado pelo critério *Backward* com $Fout = 4$, é escrito como

$$\begin{aligned}\mathbb{E}(Y|\boldsymbol{X}=\boldsymbol{x}) &= 26.01 + 9.39I_{\{smoke=0\}} + 8.65I_{\{smoke=2\}} + 9.07I_{\{smoke=3\}} \\ &+ 4.61I_{\{race=6\}} - 7.60I_{\{race=7\}} - 10.60I_{\{race=8\}} + 1.40ht\end{aligned}$$

4.5 Modelo 5: Modelo *Forward*

Já o modelo selecionado pelo critério *Forward* com $Fin = 4$, é dado é dado pela expressão

$$\begin{aligned}\mathbb{E}(Y|\boldsymbol{X}=\boldsymbol{x}) &= 42.91 - 9.04I_{\{smoke=1\}} + 4.56I_{\{race=6\}} - 8.58I_{\{race=7\}} \\ &- 9.94I_{\{race=8\}} + 1.13ht + 0.07wt.1\end{aligned}$$

4.6 Modelo 6: Modelo *Stepwise*

O modelo selecionado pelo método *Stepwise* com $Fout = Fin = 4$, é o mesmo apresentado no **Modelo 5**.

4.7 Modelo 7: Modelo *Ridge*

De acordo com o a regressão *Ridge*, o modelo final é caracterizado por

$$\begin{aligned}\mathbb{E}(Y|\boldsymbol{X}=\boldsymbol{x}) &= 92.14 + 0.005gestation + 1.22I_{\{smoke=0\}} - 1.93I_{\{smoke=1\}} + 0.78I_{\{smoke=2\}} \\ &+ 1.40I_{\{smoke=3\}} + 1.51I_{\{race=0\}} + 1.31I_{\{race=6\}} - 1.62I_{\{race=7\}} - 2.62I_{\{race=8\}} \\ &- 0.11I_{\{race=9\}} + 0.02wt.1 + 0.35ht - 0.02number1\end{aligned}$$

Vale notar, que para este caso, não mais utilizou-se variáveis de referência. Para a construção da matriz de planejamento, em cada observação, lhe era atribuído o valor 1 caso pertencesse a determinado fator e 0 caso contrário.

4.8 Modelo 8: Modelo *Lasso*

Finalmente, o último modelo estudado foi obtido através da Regressão *Lasso*, cujo resultado é assim descrito

$$\mathbb{E}(Y|\boldsymbol{X}=\boldsymbol{x}) \quad = \quad 89.44 - 4.33I_{\{smoke=1\}} + 1.79I_{\{race=0\}} + 0.48ht$$

Conforme discutido, considerar-se-á o melhor modelo de acordo com o critério do menor erro de predição. Para tanto, considere a seguinte estatística

$$E_i = \sum_{j=1}^{224} \left| \frac{y_j - \hat{y}_j}{y_j} \right|, \qquad i = 1, 2, 4, 5, 7, 8.$$

em que y_j é a j-ésima observação na amostra de validação e $\hat{y}_j$ a respectiva previsão.

Na Tabela 1 encontra-se o resultado final com a medida sumária de cada modelo em competição. Um bom modelo deverá ter um valor baixo do erro de predição.

Tabela 1: Erro de predição dos modelos

	Erro de Predição
Modelo 1	26.35
Modelo 2	26.42
Modelo 4	26.18
Modelo 5	25.22
Modelo 7	26.04
Modelo 8	26.28

5 Conclusão

Considerando a estatística supracitada, verifica-se que o melhor modelo para os dados é o **Modelo 5**. A partir dele, ir-se-á fazer algumas afirmações sobre o peso do bebê e as variáveis consideradas no modelo. De acordo com a variável raça, não há diferença entre o peso de bebês de mães brancas ou mestiças. Por outras palavras, os coeficientes referentes à raça branca ($race = 0$) e mestiça ($race = 9$) são iguais a zero. Considerando que tais regressoras são ortogonais, se a

mãe pertence à raça mexicana ($race = 6$) o peso do bebê aumenta em média 4.56 unidades em relação à referência. E se a raça da mãe é negra ($race = 7$) ou asiática ($race = 8$) o peso do bebê diminui em média 8.58 unidades e 9.94 unidades em relação à referência, respectivamente. A cada unidade de altura da mãe, a variável resposta é, em média, 1.13 unidades maior. O fator $smoke = 1$ significa dizer que o vício de fumar, alimentado durante toda a gravidez, em média, causa um decréscimo de 9.04 no peso do recém-nascido. Concluindo, não se tem evidências para rejeitar a hipótese de que o fumo durante a gravidez influencia no peso do bebê. Além disso, nada pode-se dizer a respeito do fumo antes da gravidez. É importante constatar, de maneira convincente, que as afirmações foram feitas baseadas no resultado da regressão, sem nenhuma análise de diagnóstico do modelo.

Referências

[1] BUTLER, N.R. et al., *Cigarette smoking in pregnancy: its influence on birth weight and perinatal mortality.* Brit. med. J., 2: 127-30, 1972.

[2] HASTIE, T.; TIBSHIRANI, R. & FRIEDMAN, J., H., *The elements of statistical learning: data mining, inference, and prediction: with 200 full-color illustrations.* New York: Springer-Verlag, 2001.

[3] MEYER, M.B. & TONASCIA, J.A.,*Maternal smoking, pregnancy complications and perinatal mortality.* JAmer. J. Obstet. Gynec., 128: 494-502, 1977.

[4] RUSH, D. & CASSANO, P., *Relationship of cigarette smoking and social class to birth weight and perinatal mortality among all births in Britain, 5-11 April 1970. J.* Epidem. Community Hlth, 37: 249-55, 1983.

[5] VAN DEN BERG, B.J. & YERUSHALMY, J., *The relationship of the rate of intrauterine growth of infants of low birth weight to mortality, morbidity and congenital anomalies.* J. Pediat., 69:531-45, 1966.

[6] VICTORA, C.G.; BARROS, F.C.; MARTINES, J.C.; BORIA, J.U.; VAUGHAN, J.P., *Estudo longitudinal das crianças nascidas em 1982, em Pelotas, RS, Brasil: metodologia e resultados preliminares.* Rev. Saúde públ., S. Paulo, 19: 56-68, 1985.

Tabela 2: Descrição das covariáveis utilizadas no estudo

Covariável	Descrição
PLURALTY	Número de fetos na gravidez
OUTCOME	Variável dicotômica (=1 se o feto sobreviveu até 28 dias; 0 c.c.)
SEX	Variável dicotômica (=1 se bebê do sexo masculino; 0 c.c.)
MARITAL	Variável dicotômica (=1 se mulher é casada; 0 c.c.)
WT	Peso do bebê
AGE	Idade da mãe;
ED	Nível de educação da mãe;
HT	Altura da mãe;
WT.1	Peso da mãe
DAGE	Idade do pai;
DED	Nível de educação do pai;
DHT	Altura do pai;
DWT	Peso do pai;
INC	Renda dos pais;
TIME	Tempo, em anos, que parou de fumar antes da gravidez;
NUMBER	Número médio de cigarros por dia;
RACE	= 0 raça branca;
	= 6 raça mexicana;
	= 7 raça negra;
	= 8 raça asiática;
	= 9 raça mestiça;
	= 99 raça não declarada;
SMOKE	= 0 Nunca fumou;
	= 1 Fuma hoje;
	= 2 Fumou até a gravidez;
	= 3 Parou;
	= 4 Sem informações;