

LABORATÓRIO 9

LABORATORY 9

Gabriel Penha*,

Moisés Augusto†,

Marla Lorrani‡

RESUMO

Em trabalhos anteriores, estudos para a compreensão de aspectos que influenciariam no preço dos seguros de saúde dos Estados Unidos foram realizados. Nesses estudos não foram consideradas no modelo ajustado as variáveis categóricas: sexo, região e o estado nutricional do paciente (uma classificação que considera o índice de massa corporal e a idade). Considerando, portanto, as discussões levantadas anteriormente; bem como uma base de dados mais completa, com mais características e cerca de 1200 observações, dois novos ajustes de regressão linear múltipla foram feitos; desta vez, considerando a influência de algumas variáveis categóricas, além das variáveis quantitativas que já vinham sendo utilizadas. O segundo modelo considerava as mesmas variáveis que foram utilizadas no primeiro modelo, com um acréscimo para a interação entre a idade e o sexo do cliente. Apesar disso, o poder preditivo e os pressupostos do Modelo 2 não foram superiores aos do Modelo 1. Ademais, a análise de resíduos de ambos os modelos apresentou algumas violações sérias de pressupostos, durante o texto, será possível se atentar a isso com mais detalhes, sobretudo considerando o Modelo 1.

Palavras-chave: Preço do seguro de saúde. Estado nutricional. Idade. Índice de massa corporal. Modelo de regressão linear múltiplo.

1 INTRODUÇÃO

No Relatório do Laboratório 2 e 7, disponíveis previamente para o leitor, as questões levantadas por Kanamura e Viana (2007) e por Damascena et al. (2008), foram discutidas. Nas ocasiões, apontou-se que algumas características influenciavam nos preços de planos de saúde dos Estados Unidos; entre elas, destacaram-se o índice de massa corporal (IMC) e o hábito de fumo do cliente. Agora, providos de uma base de dados mais completa, foi possível verificar a relação de fatores como idade e número de filhos dos clientes, além das características já consideradas anteriormente.


Originalmente, a base de dados possuía informações sobre 1338 clientes, no entanto, conforme a recomendação de classificação do IMC, que somente incluía no índice pessoas com idade igual ou superior a 20 anos, reduziu-se o número de observações para 1201.


Dessa maneira, o *estado nutricional* deles foi classificado da seguinte forma:

Estado Nutricional	IMC	Idade
Magreza	$\leq 18,5$	20 a 59 anos
	$< 22,0$	60 anos ou mais
Eutrofia	18,6 a 24,9	20 a 59 anos
	22,0 a 27,0	60 anos ou mais
Sobrepeso / Obesidade	≥ 25	20 a 59 anos
	$\geq 27,1$	60 anos ou mais

Figura 1 – Classificação - Estado Nutricional por idade e IMC

*  Instituto de Matemática e Estatística, Departamento de Estatística, Bacharelado em Estatística; penha.gabriel@ufba.br.

†  Instituto de Matemática e Estatística, Departamento de Estatística, Bacharelado em Estatística; moises.augusto@ufba.br.

‡  Instituto de Matemática e Estatística, Departamento de Estatística, Bacharelado em Estatística; marla.lorrani@ufba.br.

Levando isso em consideração, este trabalho objetivou ajustar modelos de regressão linear simples, se atentando para o uso de variáveis categóricas; com intuito de explicar a relação entre o preço dos planos de saúde (a variável resposta) e suas respectivas variáveis explicativas.

Na seguinte seção, será possível visualizar uma análise descritiva dos dados; em seguida, os modelos ajustados serão explicitados e analisados, com maior aprofundamento em um deles. Após isso, será possível apreciar as considerações finais.

2 ANÁLISE EXPLORATÓRIA

Os dados disponibilizados e já filtrados, incluíam algumas variáveis explicativas categóricas e outras contínuas. Na Tabela 1, é possível visualizar as estatísticas descritivas dos clientes observados, para as variáveis contínuas, incluindo o preço do seguro em dólares.

Tabela 1 – Estatísticas descritivas dos dados			
Descritiva	Idade	IMC	Preço do Seguro (\$)
Mínimo	20,00	16,82	1392,00
1° Quartil	30,00	26,40	5439,00
Mediana	42,00	30,50	9801,00
Média	41,57	30,74	13825,00
3° Quartil	52,00	34,77	17180,00
Máximo	64,00	52,58	63770,00

Pelas descritivas, é possível ver que a distribuição da idade e do IMC parecem ser simétricas. No entanto, a distribuição do preço do seguro de saúde parece ter uma assimetria à direita.

Para as variáveis categóricas, foi possível observar o seguinte:

- Fumantes: 20,3% dos clientes eram fumantes;
- Sexo: 50,4% dos clientes eram homens e 49,6% mulheres;
- Região: 24,3% dos clientes vieram do nordeste, 24,2% do noroeste, 27,0% do sudeste e o restante do sudoeste;
- N° Filhos: 39,4% dos clientes não possuíam filhos, 25,3% deles tinha um filho, 19,1% dois, 12,8% três, 2,0% quatro e o restante cinco;
- Estado Nutricional: Quanto ao estado nutricional, 17% dos clientes apresentavam bons índices de massa corporal, 1,7% deles estavam magros e o restante apresentava ou sobrepeso ou obesidade.

Realizaram-se análises visuais de tais variáveis, na Figura 2, é possível visualizar o comportamentos de cada uma dessas variáveis em relação ao preço do seguro de saúde.

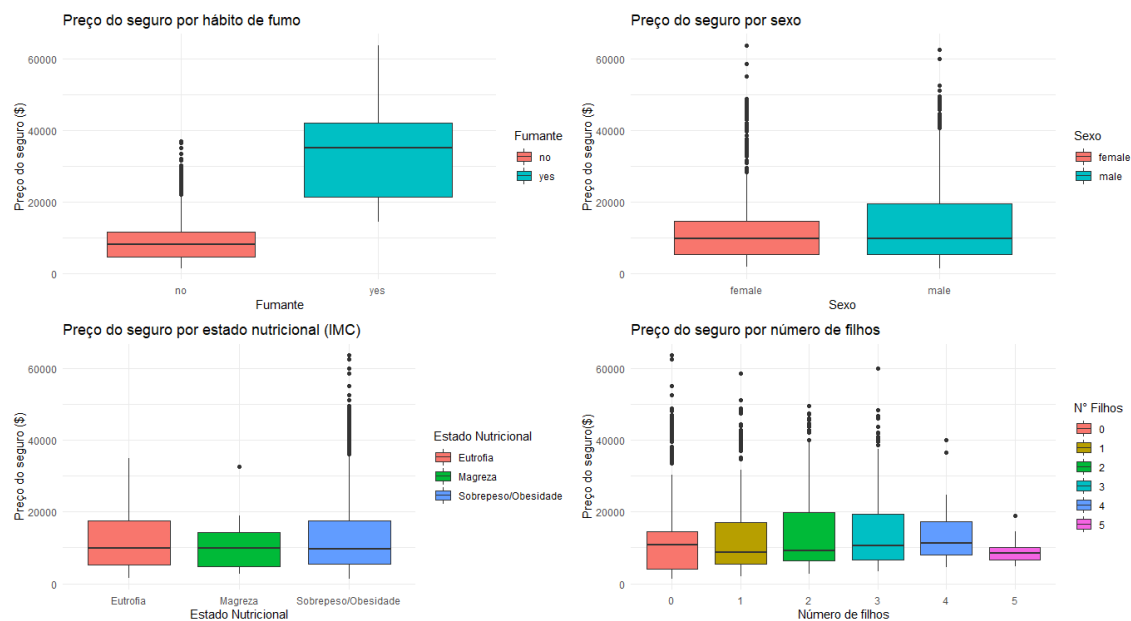


Figura 2 – Análise bi-variada

Pelos gráficos, é possível perceber que o hábito de fumo influencia bastante no preço do seguro de saúde. Não é possível dizer o mesmo do sexo; além disso, o estado nutricional também não parece apresentar grandes variações de acordo com o grupo, bem como o número de filhos. Vale salientar, entretanto, que alguns desses fatores precisam ser analisados em conjunto com outros; o que é possível observar na Figura 3.

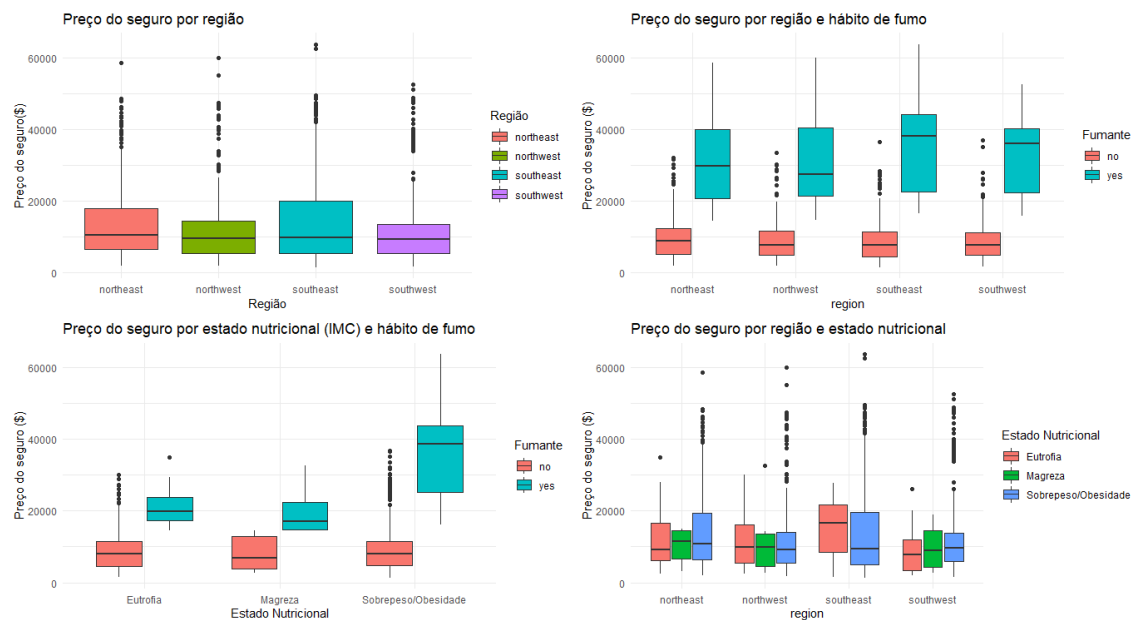


Figura 3 – Análise bi-variada - Parte 2

Pela Figura 3, é possível notar que a região em que o cliente mora também não parece influenciar tanto (diretamente) seu preço, com possível exceção para os clientes do sudeste, que parecem pagar um pouco mais caro.

Ressalta-se, além disso, que em relação ao estado nutricional, os clientes com sobrepeso e obesidade,

que são fumantes, parecem pagar mais caro que os demais grupos. Isto é, diferentemente do que acontece entre os não fumantes, os fumantes obesos pagam mais caro que os fumantes que eutróficos ou magros.

Finalmente, é possível perceber que nenhum dos clientes do sudeste estava com IMC abaixo da eutrofia. Uma análise visual também foi realizada para as variáveis contínuas, como é possível ver a seguir:

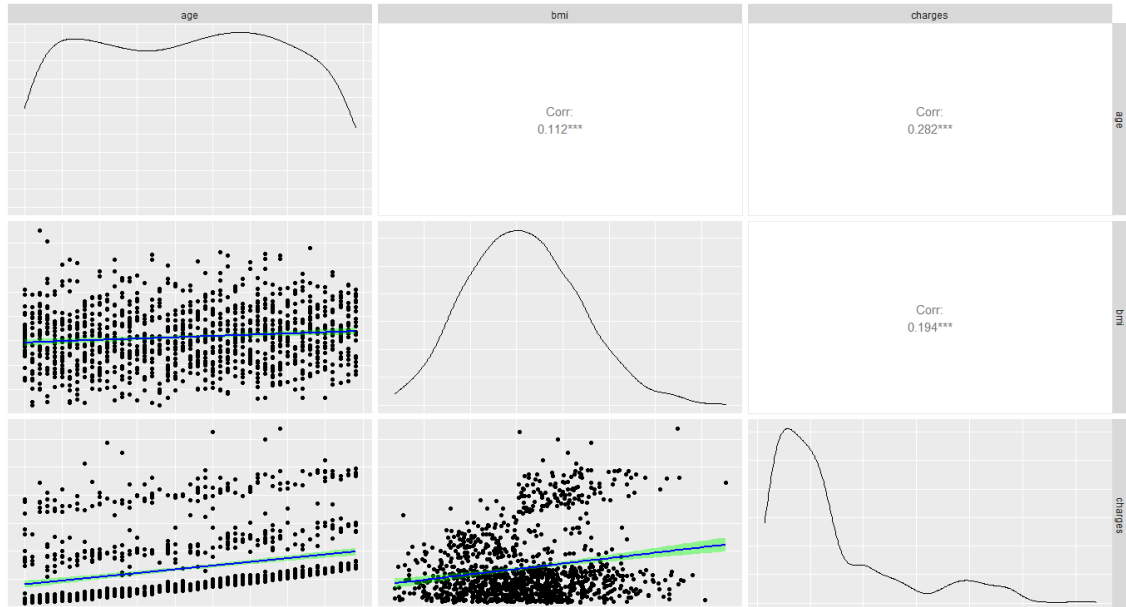


Figura 4 – Análise descritiva visual - variáveis contínuas

Na Figura 4, é possível observar as densidades e as análises bi-variadas das variáveis idade, IMC e preço do seguro. Nota-se uma forte relação entre idade e IMC, bem como entre idade e preço do seguro e preço do seguro e IMC. Vale salientar, no entanto, que como o estado nutricional foi composto através de uma regra entre as variáveis idade e IMC, esta última não foi utilizada nas modelagens.

A seguir, será possível observar alguns resultados obtidos com os ajustes de modelos de regressão linear.

3 RESULTADOS

Dois modelos de regressão linear foram ajustados, um deles não considerando interação entre as variáveis idade e sexo, enquanto o outro considerava.

3.1 Modelo 1

O Modelo 1 refere-se ao ajuste de uma regressão linear múltipla com todas as variáveis disponíveis na base em relação ao preço do seguro de saúde.

$$\begin{aligned}\hat{Y} = & -5632,19 + 24040,59 \text{fumante} + 281,59 \text{idade} - 63,55 \text{masculino} + 445,98 \text{filhos1} \\ & + 1794,55 \text{filhos2} + 1081,17 \text{filhos3} + 2992,84 \text{filhos4} + 1353,29 \text{filhos5} \\ & - 669,45 \text{noroeste} - 273,68 \text{sudeste} - 885,09 \text{sudoeste} \\ & - 152,16 \text{magreza} + 3300,87 \text{sobrepeso}\end{aligned}\quad (1)$$

Em que \hat{Y} indica o valor estimado para o preço do seguro de saúde do paciente. A interpretação do modelo presente na Equação 1 será feita considerando que todas as outras variáveis que não estão sendo interpretadas estão fixadas. Se as variáveis são como *filho1* e *filho2*, que indicam respectivamente que o

cliente possui um filho e que o cliente possui dois filhos, a interpretação acontece da seguinte maneira: considerando que o cliente tenha apenas um filho, os valores de *filho2*, *filho3* e afins serão iguais a zero.

Este modelo implica que, caso o cliente seja fumante, o preço médio do seguro aumenta em \$24040,59. Da mesma maneira, a cada acréscimo da idade do cliente, o preço médio do seguro aumenta em \$281,59. Se o cliente é homem, o preço médio é \$ – 63,55 menor do que quando os clientes são mulheres. Se o cliente possui um filho, o preço médio é \$445,98 mais caro do que quando não possui. Se são dois filhos, o valor é \$1794,55 mais caro, se três, \$1081,17 e assim por diante. Se o cliente é do noroeste, ele paga \$669,45 a menos do que se fosse do nordeste. A interpretação é análoga para quando o cliente é do sudeste ou do sudoeste.

Considerando, agora, os estados nutricionais, clientes que possuem sobrepeso ou obesidade pagam \$3300,87 mais caro que os clientes eutróficos, sendo que os clientes em estado de magreza pagam \$152,16 mais barato que os eutróficos.

Ao nível de 95% de confiança, o cliente ser fumante, ter dois filhos, ter quatro filhos, e possuir sobrepeso ou obesidade foram consideradas mudanças significantes em relação a sua categoria de referência (isto é, não fumar, não ter filhos e ser eutrófico). Além disso, a idade também foi significativa ao mesmo nível. As demais variáveis não foram consideradas influentes para o modelo.

Vale dizer que o Modelo 1 é interpretável somente para valores razoáveis das características (idades menores que 20, por exemplo, não são razoáveis para o modelo).

3.2 Modelo 2

O segundo modelo foi ajustado acrescentando o termo de interação entre as características Idade e Sexo ao Modelo 1.

$$\begin{aligned} \hat{Y} = & -5536,33 + 24040,81 \textit{fumante} + 279,22 \textit{idade} - 258,50 \textit{masculino} + 446,03 \textit{filhos1} \\ & + 1797,58 \textit{filhos2} + 1083,44 \textit{filhos3} + 3002,47 \textit{filhos4} + 1362,47 \textit{filhos5} \\ & - 669,32 \textit{noroeeste} - 274,48 \textit{sudeste} - 885,81 \textit{sudoeste} \\ & - 155,88 \textit{magreza} + 3303,52 \textit{sobrepeso} + 4,68 \textit{interação_idade_sexo} \end{aligned} \quad (2)$$

Analisando o modelo 2, houveram pequenas mudanças nos coeficientes das características, já interpretadas no modelo 1 (a interpretação aqui é análoga), além de um acréscimo de \$4,68 no preço do seguro de saúde, se considerada a relação entre a idade e o sexo.

Assim como o sexo, a interação entre idade e sexo não foi considerada significativa ao nível de 95%. A significância das demais variáveis se manteve como no Modelo 1.

Considerando isso, concluiu-se que o uso da interação entre as duas variáveis não valia a pena. O acréscimo da interação entre estas variáveis não aumentou o poder preditivo do Modelo 2 em relação ao primeiro, além de não garantir as suposições esperadas de um modelo de regressão linear.

3.3 Análise de resíduos

Levando o mencionado em conta, realizaram-se análises de resíduos para ambos os ajustes. No entanto, somente a do Modelo 1 será exibida neste relatório.

Na Figura 5, é possível visualizar graficamente pressupostos como normalidade, homocedasticidade, alguns pontos influentes, entre outros.

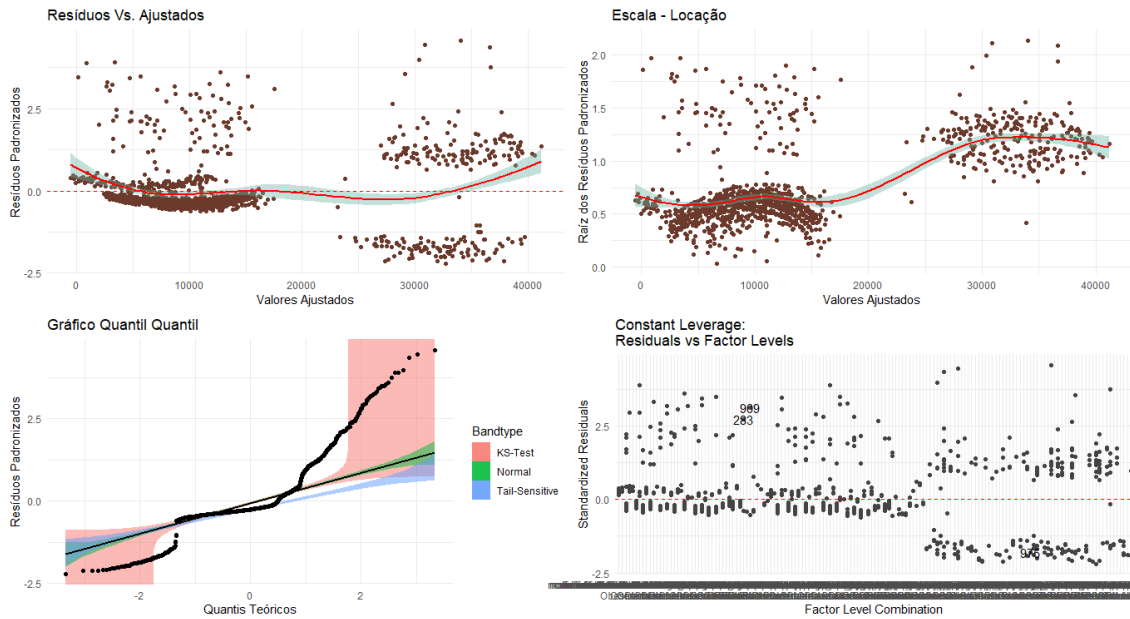


Figura 5 – Análise de Resíduos

Apesar do gráfico de resíduos e ajustados não apresentar uma tendência *muito forte*, há um vácuo de pontos entre dois polos, o que pode indicar que não há homocedasticidade. Uma interpretação similar pode ser feita para o gráfico de escala-localização. Dito isso, é possível ver que a normalidade do modelo está completamente comprometida. As legendas do gráfico quantil quantil indicam diferentes bandas de confiança para normalidade calculadas através de diferentes métodos. Notavelmente, boa parte dos resíduos foge do comportamento esperado. O gráfico inferior direito aponta para algumas possíveis observações influentes.

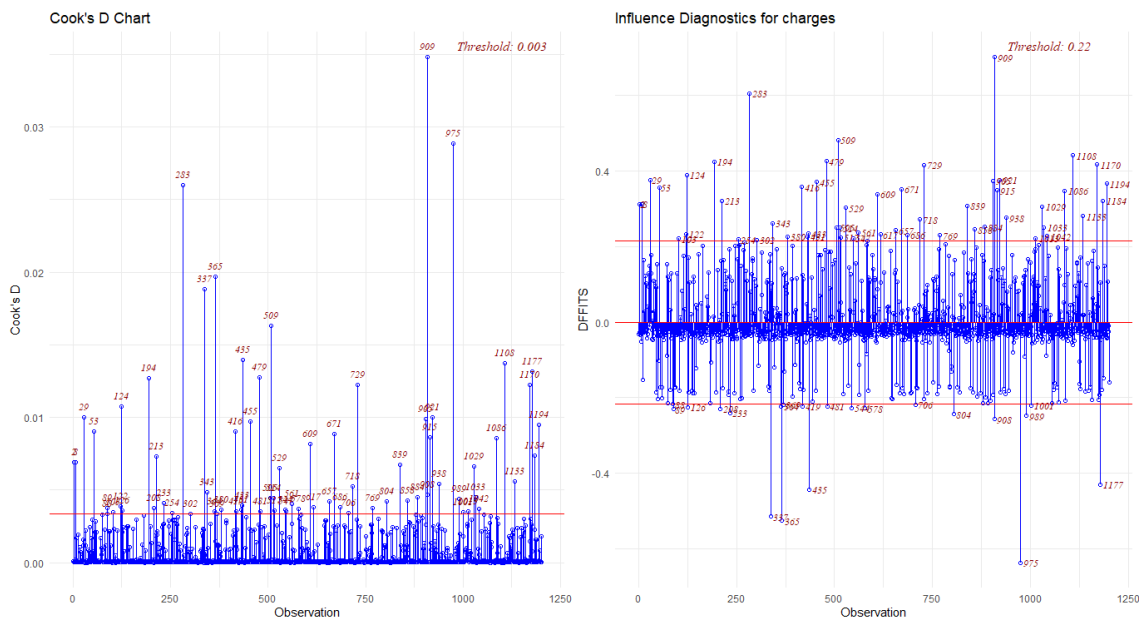


Figura 6 – Análise de Resíduos 2

Os gráficos da Figura 6 apresentam as distâncias de Cook e os DFFITS do modelo; é possível ver que vários pontos fogem do esperado (ou seja, estão acima da linha limite para o gráfico à esquerda, e fora das

duas linhas limites no gráfico à direita).

Testes numéricos para verificação de pressupostos do modelo de regressão linear também foram considerados (inclui-se, aqui, os testes de hipótese). A normalidade foi rejeitada, como esperado. No entanto, o teste de Goldfeld-Quandt apontou para homocedasticidade dos dados, diferentemente do esperado e diferentemente também do de Breush-Pagan, que apontou para a homocedasticidade. Os resíduos são independentes, segundo o teste de Durbin-Watson; além disso, as co-variáveis não apresentaram multicolinearidade (lembrando que o IMC *puro* não foi incluído no modelo, mas sim a variável criada através dele, **Estado Nutricional**).

Mais algumas medidas foram consideradas, no entanto, dadas as falhas do modelo em alguns dos pressupostos mais importantes, elas serão aqui omitidas. Para verificar os gráficos de regressão e resíduos parciais, os gráficos dos DFBetas, e o gráfico que classifica as observações em *outliers*, *leverage* e não atípicos, acesse o endereço *web* disponibilizado nos anexos do documento.

4 CONSIDERAÇÕES FINAIS

Considerando o mencionado, é possível dizer que esta é uma base de dados promissoras, que pode, futuramente, render análises mais precisas da relação entre o preço do seguro do plano de saúde e algumas das variáveis explicativas.

Encontraram-se relações interessantes entre, por exemplo, o preço, o hábito de fumo e o estado nutricional; percebeu-se também, que a idade pode ser um fator encarecedor nas tarifas dos planos de saúde. Dito isso, algumas das variáveis ajustadas não parecem ser tão significantes, uma delas, por exemplo (e ao menos para os dados disponibilizados), é o número de filhos do cliente. O sexo também parece outra variável não tão importante para explicar a variabilidade nas taxas do seguro.

Dois modelos de regressão linear múltipla foram ajustados, sendo que o primeiro deles não considerava como co-variável explicativa a interação entre idade e sexo. O Modelo 2 não apresentou resultados melhores que o Modelo 1 nem em termos de predição, nem em termos de pressupostos da regressão linear. Dessa maneira, não parece muito adequado manter a interação supracitada.

Uma análise dos pressupostos foi performada para o Modelo 1, que apresentou violações graves de alguns deles, como normalidade e homocedasticidade, além de muitos pontos possivelmente influentes identificados pelo gráfico de distâncias de Cook.

Assim, recomenda-se um novo ajuste desconsiderando algumas das variáveis que, de fato, não pareciam ser significantes. Isso, possivelmente, poderia melhorar algumas das análises quanto aos pressupostos do modelo.

REFERÊNCIAS

- DAMASCENA, L. L. et al. Correlação entre obesidade abdominal, IMC e risco cardiovascular. **Centro de Ciências da Saúde, Departamento de Educação Física. 11º Encontro de Iniciação à Docência**, p. 9–11, 2008.
- KANAMURA, A. H.; VIANA, A. L. D. Gastos elevados em plano privado de saúde: com quem e em quê. **Revista de Saúde Pública, SciELO Brasil**, v. 41, p. 814–820, 2007.

ANEXO A – CÓDIGOS UTILIZADOS PARA ANÁLISE (NO R)

É possível todos os laboratórios feitos até agora através do *link* a seguir:
<https://github.com/Mkyou/labs-regressao>

DECLARAÇÃO DE RESPONSABILIDADE

O(s) autor(es) é(são) o(s) único(s) responsável(eis) pelas informações contidas neste documento.