Ministério da Educação
Universidade Federal da Bahia
Campus Ondina
Departamento de Estatística - UFBA

LABORATÓRIO 7

LABORATORY 7

Gabriel Penha*, Moisés Augusto†

RESUMO

Em trabalhos anteriores, a relação entre o preço do plano de saúde e algumas características dos pacientes como idade, índice de massa corporal e hábitos de fumo foram analisadas através de dados de indivíduos dos Estados Unidos. Na ocasião, ajustaram-se modelos de regressão linear simples com intuito de se explicar o preço do plano de saúde através de uma das características dos pacientes. Neste trabalho, aquele conjunto de dados é retomado com algumas restrições e, além disso, utiliza-se do modelo de regressão linear múltiplo para explicar a variabilidade dos preços. Dois modelos foram ajustados considerando como variáveis explicativas a idade e o IMC dos pacientes, sendo que o segundo não continha os pontos atípicos identificados no primeiro modelo. Concluiu-se que, dentre os dois modelos, o primeiro parecia ser o mais adequado para descrever a realidade, por se adequar mais ao esperado de uma regressão linear múltipla.

Palavras-chave: Plano de Saúde. Idade. Indice de Massa Corporal. Modelo linear múltiplo. Regressão.

1 INTRODUÇÃO

No Relatório do Laboratório 2, disponível previamente para o leitor, as questões levantadas por Kanamura e Viana (2007) e Damascena et al. (2008), foram discutidas e uma base de dados que considerava o gasto dos pacientes com plano de saúde foi considerada; levando em consideração variáveis como idade, o hábito de fumo, ou mesmo o índice de massa corporal do paciente.

Naquela ocasião, verificou-se que havia uma diferença importante de preço entre os pacientes fumantes e os não fumantes. Além disso, concluiu-se que o modelo de regressão linear simples poderia não ser o mais adequado para analisar aqueles dados.

Dito isto, neste relatório 7, retomou-se o conjunto de dados do laboratório 2. Com as seguintes diferenças:

- Apenas fumantes foram analisados;
- Apenas pacientes do sudeste dos Estados Unidos foram analisados; desconsiderando aqueles que pertenciam a outras regiões;
- A variável sexo não foi considerada; no laboratório 2 sua não significância para explicar o preço dos planos de saúde (linearmente) fora apontada;

Assim, com este novo subconjunto, ajustaram-se dois modelos de regressão linear múltiplo, utilizando o IMC e a idade como variáveis explicativas e o preço do plano como variável resposta. Vale dizer que o segundo ocorreu em decorrência de algumas intempéries surgidos no primeiro.

Na seguinte seção, ter-se-á uma análise exploratória dos dados, incluindo uma análise visual. Em seguida, a seção de resultados tratará dos modelos ajustados e de seus pressupostos. A última seção conterá algumas considerações finais.

* Instituto de Matemática e Estatística, Departamento de Estatística, Bacharelando em Estatística; penha.gabriel@ufba.br.
† Instituto de Matemática e Estatística, Departamento de Estatística, Bacharelando em Estatística; moises.augusto@ufba.br.

2 ANÁLISE EXPLORATÓRIA

O novo conjunto de dados possuía 91 linhas, com nenhuma informação faltante ou duplicada. Na Tabela 1, é possível visualizar as estatísticas descritivas da idade, do IMC e do preço dos planos de saúde:

Tabela 1 – Estatísticas descritivas dos dados

Descritiva	IMC	Idade	Preço do Plano ($)
Mínimo	19,80	18,00	16578,00
1° Quartil	27,17	27,00	23155,00
Mediana	33,11	42,00	37484,00
Média	33,10	39,74	34845,00
3° Quartil	38,06	51,00	43392,00
Máximo	52,58	64,00	63770,00

É possível perceber que nenhuma das três variáveis parecem ter uma distribuição não simétrica; isto é, em todas elas a média e a mediana possuem valores parecidos.

Uma análise visual foi feita e corroborou esses indícios, com ressalvas para o preço do plano, que apresentou um valor máximo um pouco mais distante que o esperado.

Feito isso, seguiu-se para uma análise bi-variada.

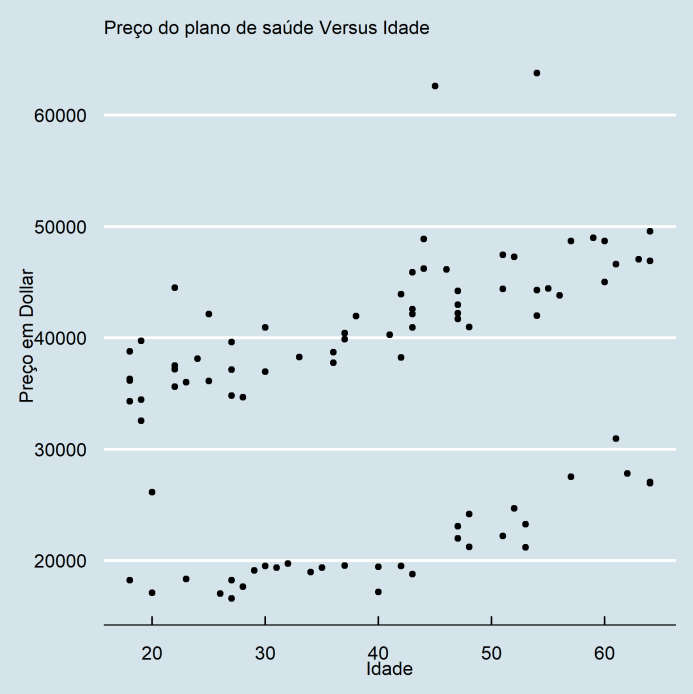

Figura 1 – Diagrama de Dispersão: Preço do Plano de Saúde ($) Vs. Idade

Como é possível visualizar pela 1, parece haver uma relação crescente entre a idade e o preço do plano, porém, dividida em dois grupos que, com bases nos dados utilizados, não foram identificados.

Na 2, é possível ver a relação entre o preço do plano de saúde em dólares e o índice de massa corporal.

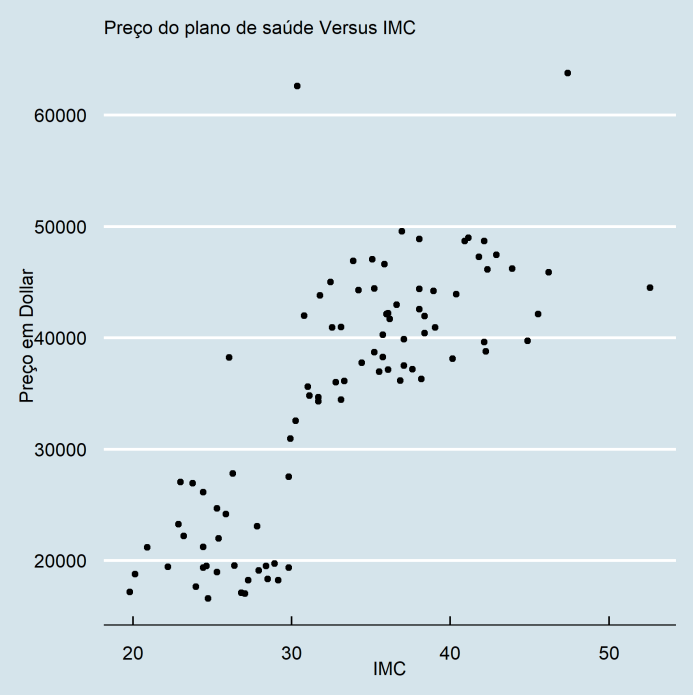

Figura 2 – Diagrama de Dispersão: Preço do Plano de Saúde ($) Vs. IMC

É possível perceber que, de fato, há uma relação crescente entre o IMC e o preço do plano de saúde; além disso, essa relação parece linear.

Finalmente, a 3 indica a relação entre o IMC e a idade:

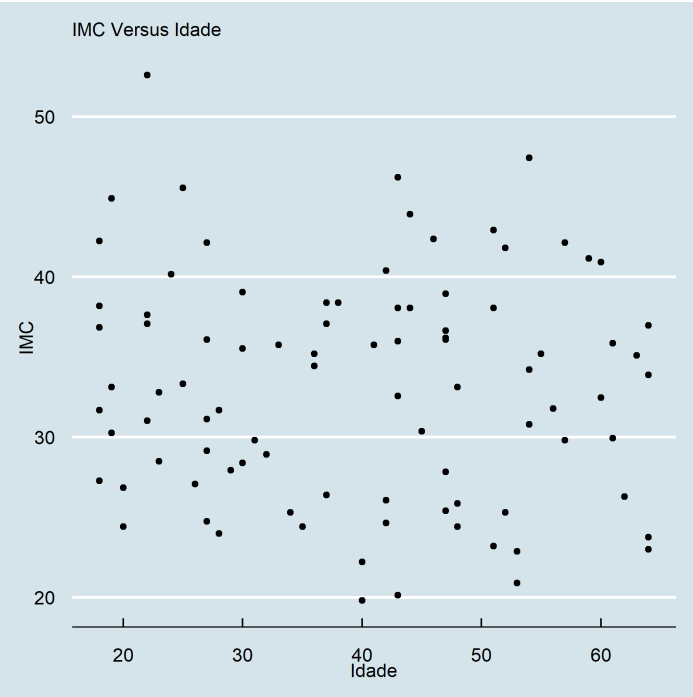

Figura 3 – Diagrama de Dispersão: IMC ($) Vs. Idade

Pela 3, é possível perceber que não parece haver relação entre as duas variáveis - que é o esperado, haja vista a suposição de não colinearidade entre as variáveis explicativas -.

Ao nível de 95%, realizaram-se testes de hipótese para o coeficiente de correlação entre cada uma das variáveis explicativas e a variável resposta. Em ambos os casos, o teste indicou favoravelmente pela relação entre as variáveis.

Um teste para a correlação entre as variáveis explicativas também foi utilizado. O resultado apontou para a não relação entre IMC e idade.

3 RESULTADOS

Considere que *age* indicará a idade em anos do paciente e que *imc* indicará o seu índice de massa corporal.

O Modelo 1, estimado ficou da seguinte forma:

$$\hat{Y} = -20175{,}61 + 285{,}83age + 1319{,}25imc \tag{1}$$

Em que $\hat{Y}$ indica o valor estimado para o preço do plano de saúde quando a idade do paciente é *age* e, quando seu índice de massa corporal é *imc*.

Este modelo indica que a cada acréscimo de idade (fixando o índice de massa corporal), o preço do plano aumentará em \$285,83. Além disso, caso se fixe a idade, a cada unidade acrescida no índice de massa corporal, o preço do plano aumenta em \$1319,25. Aponta-se que este modelo é interpretável somente para valores plausíveis, isto é, valores que podem, de fato, acontecer na realidade, tanto para o IMC quanto para a idade.

Com 95% de confiança, o modelo foi considerado significante por um teste F de Snedecor. Ao mesmo nível, as duas variáveis explicativas consideradas também foram consideradas significantes (através de um teste utilizando a distribuição *t-student*).

Uma análise de resíduos foi feita para este modelo, visual e quantitativamente. Os gráficos quantil quantil (Normal), resíduos *versus* ajustados, escala-locação e resíduos *versus leverage* podem ser visualizados na 4.

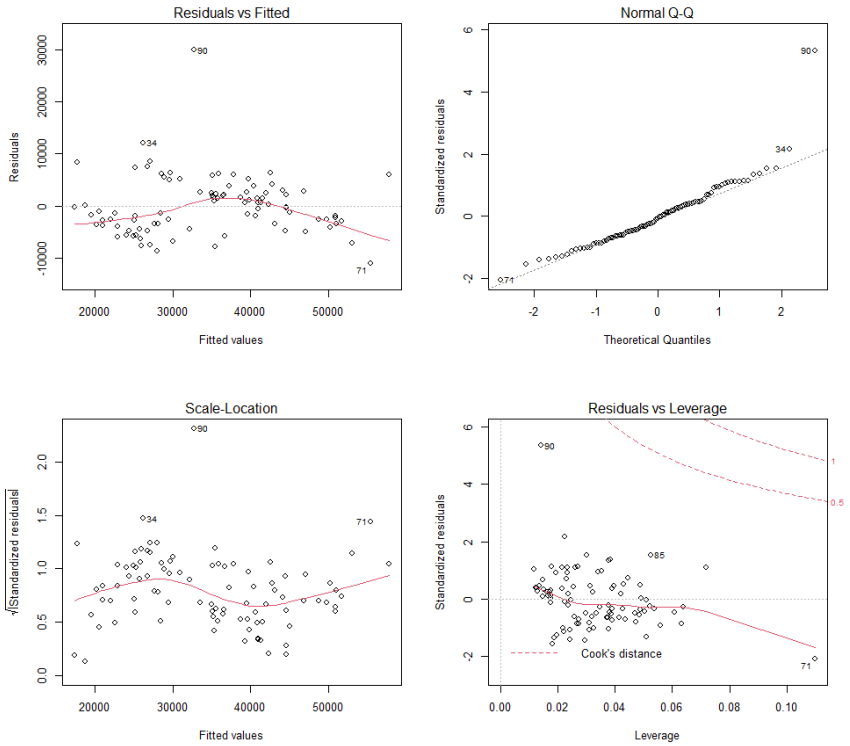

Figura 4 – Análise de resíduos, Modelo 1

Como é possível observar pelos gráficos *Residuals vs Fitted* e *Scale-Location* os resíduos não parecem ser *muito* heterocedásticos, exceto pelo ponto de $id = 90$ e, talvez, pelo de $id = 71$; os resíduos parecem seguir uma distribuição normal, também com ressalva ao ponto de $id = 90$ e, talvez os pontos com $id = 71; 85; 90$ sejam pontos atípicos influentes para o modelo.

Antes que a suspeita sobre os pontos atípicos fossem verificadas, através de testes de hipótese tanto a normalidade quanto a homocedasticidade do modelo foram verificadas; ambos os testes apontaram favoravelmente aos pressupostos do modelo, ao nível de 95%.

Sendo assim, obtiveram-se os os resíduos *Jacknife* e, através deles, para que se analisassem a influência dos pontos, verificaram-se o gráfico quantil-quantil para a distribuição *t-student*; a distribuição dos resíduos, o gráfico com as distâncias de Cook e com os DFFITS;

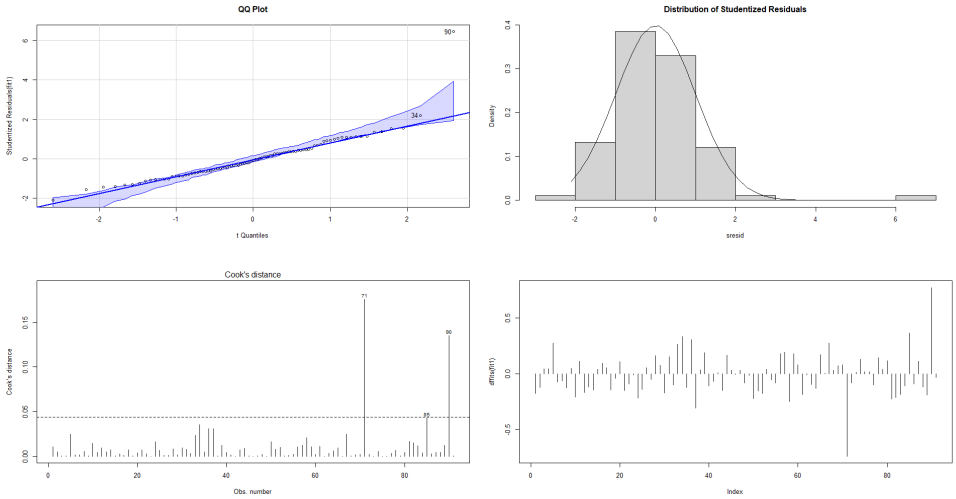

Figura 5 – Análise de resíduos - 2, Modelo 1

Como é possível visualizar na 5, exceto pela observação de $id = 90$, os resíduos *Jacknife* parecem seguir a distribuição esperada. A distribuição dos resíduos fortalece a hipótese de que trata-se de um *outlier* e a influência deste ponto é reforçada tanto pelo gráfico das distâncias de Cook, quanto para o gráfico dos DFFITS.

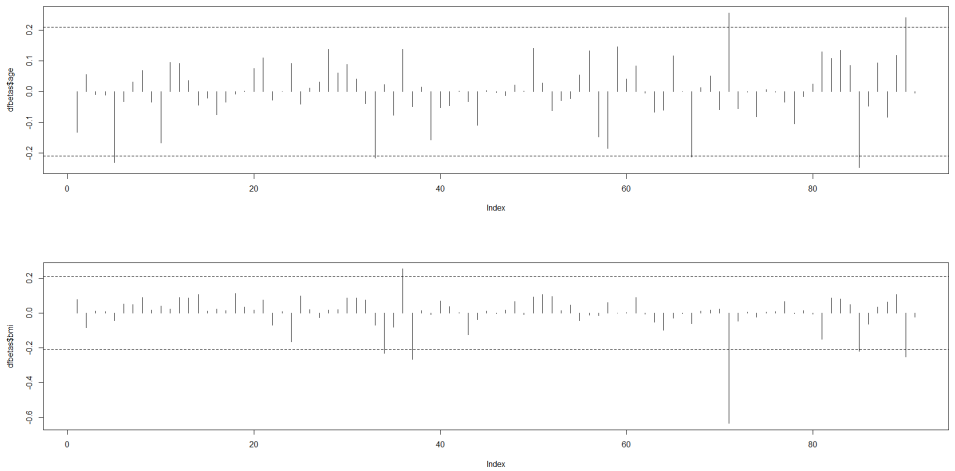

Figura 6 – DFBetas, Modelo 1

Na 6, é possível observar o *plot* dos DFBetas; também indicadores de influência, mas que aponta para a influência da observação sob cada uma das variáveis. Neste tipo de gráfico, em geral, os pontos que superam as linhas tracejadas são considerados pontos influentes.

Finalmente, a 7 é um gráfico que exibe os COVRATIO calculados para cada observação.

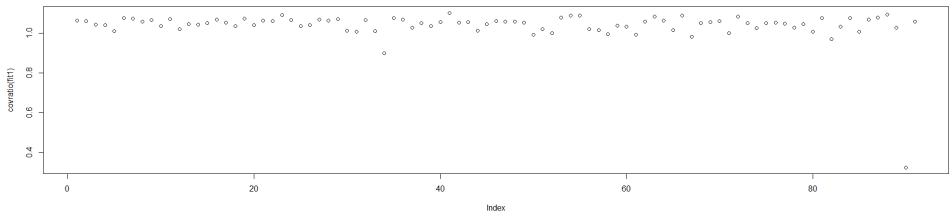

Figura 7 – COVRATIO, Modelo 1

Neste gráfico, se os valores estão muito pequenos, é sinal de que são pontos influentes. O ponto mais "baixo" da 7 é a observação 90.

Dito isso, considerando todas as análises realizadas, concluiu-se que os pontos de $id = 71; 90$ são pontos influentes. Dessa forma, decidiu-se realizar um novo ajuste, com as mesmas variáveis, mas os retirando; este foi chamado de Modelo 2.

O Modelo 2 foi da forma:

$$\hat{Y} = -21969{,}44 + 266{,}61age + 1390{,}60imc \tag{2}$$

Ao nível de 95%, tanto idade quanto o IMC foram consideradas significativas (pelo teste t). Este modelo indica que a cada acréscimo na idade, fixado o IMC, \$266,61 são acrescidos ao preço do plano de saúde. Além disso, fixada a idade, a cada acréscimo no índice de massa corporal, \$1390,60 são acrescidos ao preço do plano.

As ressalvas sobre as restrições para o índice de massa corporal e para a idade se mantêm aqui.

Assim como no Modelo 1, análises de resíduos e verificação de pressupostos do modelo de regressão linear foram feitas para o Modelo 2. No entanto, diferentemente do que se podia esperar, a retirada dos pontos influentes do Modelo 1 não melhorou o ajuste no Modelo 2. Isto é; apesar do poder preditivo ter aumentado, assim como a proporção da variabilidade do preço explicada pelo modelo, tanto a normalidade quanto a homocedasticidade dos resíduos ficaram comprometidas. Além disso, novos pontos considerados influentes através dos testes surgiram. De modo geral, o Modelo 2 parecia possuir maior poder preditivo que o 1; mas perdia na garantia de suposições (que foi justamente o motivo que fez com que ele fosse ajustado).

4 CONSIDERAÇÕES FINAIS

Considerando o mencionado, o Modelo 1 parece ser o mais consistente com a realidade, pois apesar de explicar um pouco menos a variabilidade do modelo, garante melhor a maioria das suposições do modelo de regressão linear múltiplo.

Dito isso, apesar dos pontos influentes não alterarem muito os coeficientes e estimadores dos parâmetros em si, retirá-los não parece adequado para se obter um melhor ajuste (em termos de não condizerem com a realidade muito menos).

Assim, conclui-se que seria adequado o recolhimento de um número maior de observações, para que talvez a influência das observações apontadas pelo texto fosse menor em relação ao modelo. Outra possibilidade é a inclusão de mais alguma variável explicativa, que possivelmente esteja faltando no ajuste para explicar os preços.

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.