Análise Multinível dos Fatores Determinantes nos Custos de Saúde: O Impacto do Tabagismo e da Obesidade

Daniel Jacob Tonn*

18 de junho de 2024

Resumo

O custo dos serviços de saúde varia amplamente devido a fatores comportamentais, biológicos e geográficos. Este estudo analisa um conjunto de dados ('*insurance*') contendo essas variáveis para prever os custos médicos individuais ('*charges*'). Utilizando regressão múltipla e modelos multiníveis, investigamos a significância das covariáveis sobre a variável resposta em cada divisão de grupos proposta. A validação dos modelos é feita com MSE, R^2 e AIC. Os resultados visam informar decisões, conscientizar sobre os impactos de certos comportamentos e ajudar na criação de planos de saúde mais justos.

Palavras-chave: Custos de Saúde, Modelo Multinível, Modelagem estatística, Análise estratificada.

*EMAp FGV, danieljt.djt@gmail.com

Introdução

O custo dos serviços de saúde é uma preocupação crescente em muitas sociedades modernas. As despesas médicas podem variar amplamente entre indivíduos devido a uma variedade de fatores, como fatores comportamentais, biológicos e até geográficos. Neste trabalho, vamos investigar a influência de idade, sexo, índice de massa corporal (IMC), número de filhos, hábitos de fumo e região geográfica sobre o valor anual médio gasto em serviços médicos. A capacidade de prever os custos médicos individuais é crucial não apenas para as seguradoras de saúde, que precisam estabelecer preços de seguro justos e competitivos, mas também para os formuladores de políticas que buscam entender os determinantes das despesas de saúde e formular intervenções adequadas. É de extrema importância que os órgãos responsáveis pelo sistema de saúde compreendam profundamente os fatores que moldam a saúde da população. Conhecimentos como o impacto negativo do tabagismo e da obesidade na qualidade de vida são essenciais para a elaboração de políticas eficazes e programas de prevenção que promovam uma sociedade mais saudável e consciente.

Este estudo visa explorar o conjunto de dados '*insurance*, usado como recurso didático no livro *Machine Learning With R* Lantz [2019] que contém informações sobre características demográficas, hábitos de vida e custos médicos individuais de 1338 estadunidenses que aderiram a planos de saúde. O conjunto de dados, denominado ' *insurance*', inclui as variáveis '*age*', '*sex*', '*bmi*', '*children*', '*smoker*', '*region*' e '*charges*'. A variável dependente '*charges*' representa os custos médicos individuais cobrados pelo seguro de saúde, enquanto as outras variáveis são consideradas como covariáveis que podem influenciar esse custo.

A análise será conduzida em várias etapas, começando pela exploração descritiva dos dados para compreender melhor as distribuições e possíveis relações entre as variáveis. Em seguida, construiremos diversos modelos estatísticos para quantificar o impacto de cada covariável sobre os custos médicos. Modelos de regressão linear e modelagem multinível serão utilizados. A validação dos modelos será realizada utilizando (MSE) (*Mean Square Error*), o coeficiente de determinação (R^2)e o AIC (*Akaike Information Criterion*).

Este estudo tem como objetivos principais responder as perguntas: (1) Quais variáveis têm maior influência sobre os custos médicos individuais? (2) Qual o impacto dessas variáveis? (3) Os dados disponíveis são suficientes pra explicar a variabilidade dos preços de saúde?.

Dessa forma, o presente trabalho se estrutura da seguinte maneira: após a introdução, será abordada a exploração e tratamento dos dados. Posteriormente, uma discussão dos métodos utilizados e a construção modelos, seguido dos resultados obtidos pelos mesmos, terminando com as conclusões obtidas.

1 Tratamento e Exploração dos dados

Nesta seção, detalhamos o tratamento e a exploração inicial dos dados. O objetivo é entender melhor as distribuições e relações entre as variáveis para informar a modelagem subsequente.

1.1 Descrição das Covariáveis

A tabela abaixo descreve cada covariável presente no conjunto de dados, incluindo seu nome, uma breve descrição, se é uma variável descritiva ou quantitativa, e seu domínio.

Covariável	**Descrição**	**Tipo**	**Domínio**
age	Idade do indivíduo	Q	Inteiros positivos (18 a 64 anos)
sex	Sexo do indivíduo	D	'male', 'female'
bmi	Índice de Massa Corporal (IMC)	Q	Reais positivos (16 a 53)
children	Número de filhos/dependentes	Q	Inteiros não negativos (0 a 5)
smoker	Se o indivíduo é fumante	D	'yes', 'no'
region	Região geográfica	D	'northeast', 'northwest' 'southeast', 'southwest'
charges	Custos médicos individuais	Q	Reais positivos

Tabela 1: Descrição das Covariáveis

1.2 Tratamento pré exploração dos dados

As variáveis categóricas no conjunto de dados, como *sex*, *smoker* e *region*, precisam ser transformadas em um formato que possa ser utilizado em modelos de regressão. Isso é feito através da codificação dessas variáveis em variáveis dummy (indicadoras).

- **Sexo (sex):** Esta variável possui duas categorias: ' *male*' e ' *female*'. Após a codificação, ela é transformada em uma variável *dummy* chamada *'sex_male'*, onde ' *male*' é representado por 1 e ' *female*' por 0.
- **Fumante (*smoker*):** Esta variável possui duas categorias: ' *yes*' e ' *no*'. Após a codificação, ela é transformada em uma variável *dummy* chamada *smoker_yes*, onde ' *yes*' é representado por 1 e ' *no*' por 0.
- **Região (region):** Esta variável possui quatro categorias: ' *northeast*', ' *northwest*', ' *southeast*' e ' *southwest*'. Após a codificação, ela é transformada em quatro variáveis dummy: *region_northeast*, *region_northwest*, *region_southeast*, e *region_southwest*, onde 1 indicando a presença na categoria e 0 indicando sua ausência.

Este processo de transformação é necessário para permitir que os modelos de regressão múltipla tratem adequadamente as variáveis categóricas.

1.3 Exploração dos dados

Efetuados os ajustes, podemos visualizar a matriz de correlação entre as variáveis.

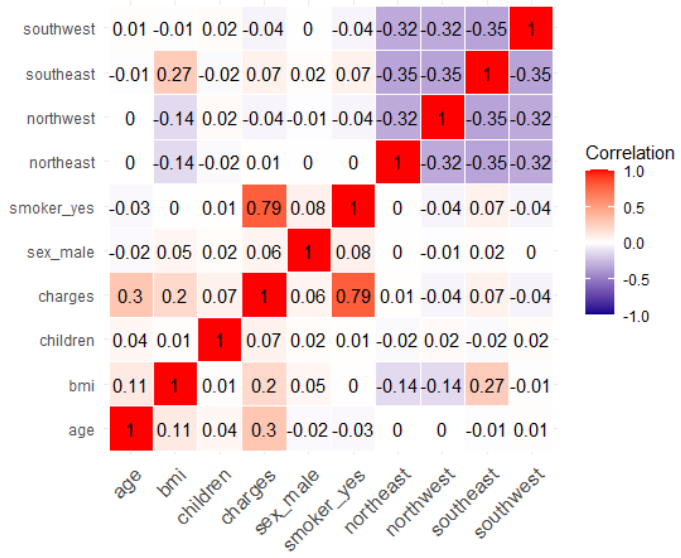


Figura 1: Matriz de correlação

A partir da Figura 1 podemos observar que a variável binária que descreve a prática de tabagismo ('smoke_yes') se destaca na correlação com os custos.

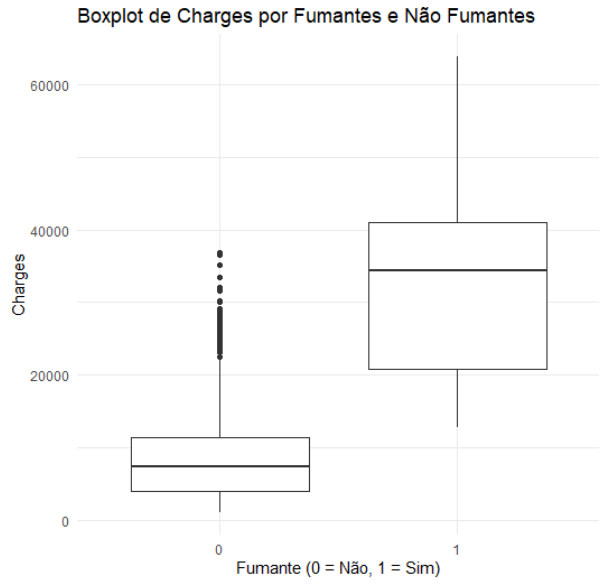


Figura 2: Boxplot de custos por categoria de fumantes

Notavelmente os custos de clientes que são fumantes tendem a ser em média maiores do que em não fumantes. Além disso os custos de maior amplitude são exclusivamente de

fumantes enquanto os de menor amplitude são exclusivamente de não fumantes.

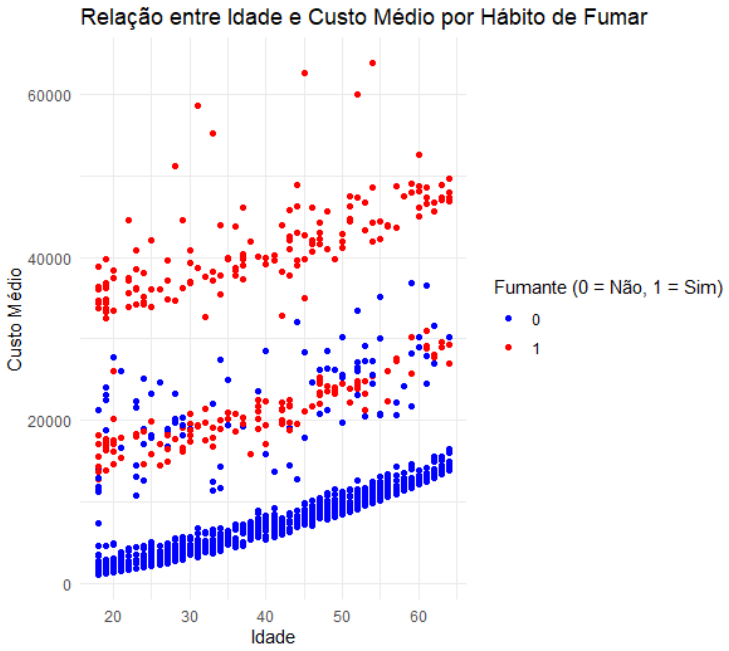


Figura 3: Idade x custos por categoria de fumantes

Outro destaque dar-se-à pela interferência não linear da idade sobre os custos. Pelo comportamento dos dados na Figura 3 a idade tem efeitos quadráticos sobre os custos. De fato, pela Figura 4 podemos ver que o quadrado da idade tem uma relação linear com os custos.

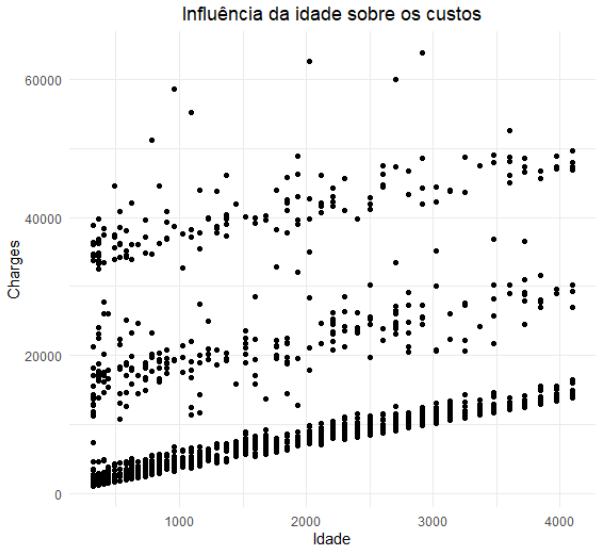
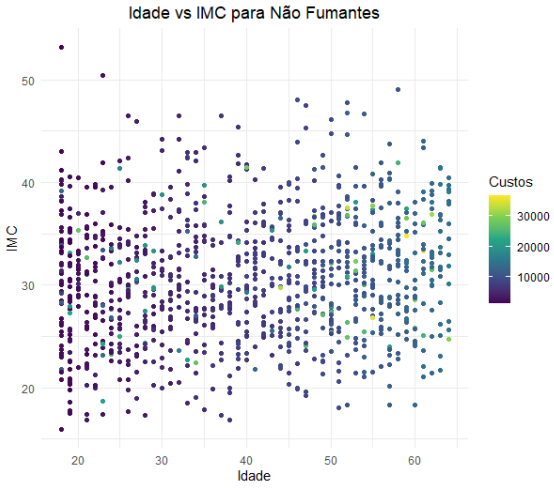
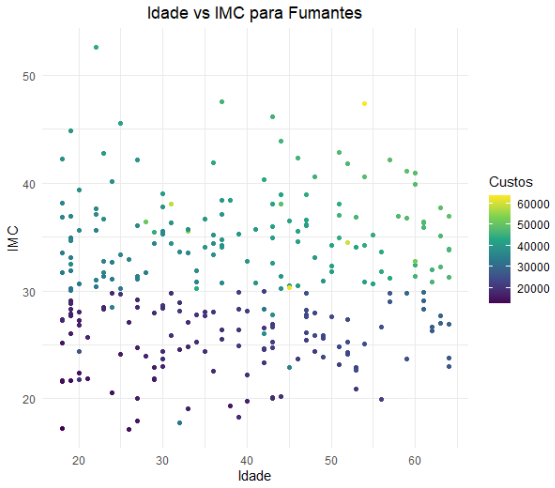


Figura 4: Quadrado da idade x custos

Ainda pela Figura 1 outras duas covariáveis se destacam: idade e IMC.



(a) Grupo de não fumantes



(b) Grupo de fumantes

Figura 5: Comparação entre os grupos de fumantes e não fumantes

Fazendo uma breve análise estratificada diante dos grupos da variável de tabagismo, é observável que em não fumantes a idade é um fator relevante para elevação dos custos, enquanto no grupo de fumantes o IMC exerce maior significância. Além disso, podemos notar que as escalas de ambas as imagens são diferentes, revelando uma diferença significativa entre os grupos.

É notável ainda que no grupo de fumantes há dois subgrupos, classificados quanto ao IMC. De fato, IMC$\geq$ 30 classifica a pessoa como obesa. Para tal análise, é plausível a adição de uma covariável ('*obese*'), que indica se a pessoa é obesa.

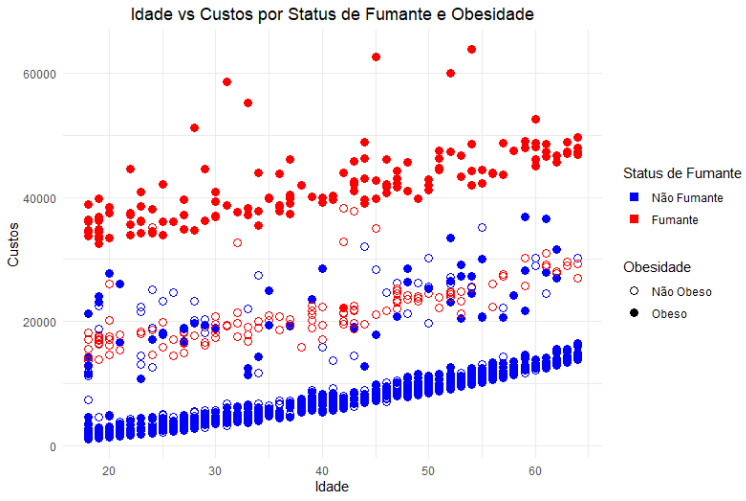


Figura 6: Idade x Custos por grupo

A clara influência diversifacada em cada grupo sugere uma abordagem de classificação hierárquica sobre os dados, isto é, utilização de modelos multínel.

1.4 Tratamento pós exploração dos dados

Como discutido acima, é viável a inclusão de uma variável binária que informa se a pessoa é classificada como obesa. Além disso, queremos aplicar modelos lineares então é conveniente reescalar a variável da idade para escala quadrática, afim de exercer influência linear sobre os custos.

```
> print(head(data))
     bmi children   charges sex_male smoker_yes northeast northwest southeast southwest obese square_age
1 27.900        0 16884.924        0          1         0         0         0         1     0        361
2 33.770        1  1725.552        1          0         0         0         1         0     1        324
3 33.000        3  4449.462        1          0         0         0         1         0     1        784
4 22.705        0 21984.471        1          0         0         1         0         0     0       1089
5 28.880        0  3866.855        1          0         0         1         0         0     0       1024
6 25.740        0  3756.622        0          0         0         0         1         0     0        961
```

Figura 7: Cabeçalho do *dataset* pós tratamento dos dados

2 Métodos

Nesta seção, descreveremos a metodologia adotada para modelar as despesas médicas, bem como comparar o desempenho de diferentes abordagens de regressão múltipla.

Os modelos serão ajustados utilizando R, devido a praticidade diante de modelos multiníveis. Antes dessa etapa, os dados serão divididos em conjuntos de treinamento e teste para avaliar o desempenho dos modelos de forma imparcial.

Ao separar os dados em conjuntos de treinamento e teste, garantimos uma avaliação imparcial do desempenho dos modelos em dados não vistos anteriormente, o que nos fornecerá uma estimativa confiável de sua capacidade preditiva e de generalização.

Após o ajuste dos modelos, aplicaremos cada um deles aos dados de teste para prever o turnover. Em seguida, avaliaremos o desempenho dos modelos utilizando duas métricas principais: o coeficiente de determinação (R^2) e o Critério de Informação de Akaike (AIC).

2.1 Modelos

2.1.1 Modelo Geral

Nesta abordagem, ajustamos um modelo de regressão múltipla que inclui todas as variáveis disponíveis. Este modelo serve como referência para comparação com os modelos subsequentes. Considerando as premissas de linearidade da variável dependente e homocedasticidade, define-se

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n + \epsilon, \tag{1}$$

onde:

$$\epsilon \sim \text{Distribuição Normal}(0, \sigma);$$

As predições $\bar{Y}$ são definidas como

$$\bar{Y} = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n. \tag{2}$$

Nesse modelo os coeficientes $\beta_0, \ldots, \beta_n$ serão estimados utilizando máxima verossimilhança, que procura o conjunto de parâmetros que maximiza a verossimilhança de se observar os dados em questão dado o modelo. Esse modelo básico nos permitirá entender a influência inicial das variáveis e obter uma ideia do poder preditivo. No entanto, nosso objetivo principal é explorar possíveis melhorias e refinamentos no modelo, a fim de aumentar sua capacidade de previsão e compreensão dos valores gastos em cuidados médicos.

2.1.2 Modelo Multinível

Como proposta alternativa, ajustamos um modelo que leva em consideração diferentes interceptos para os diferentes grupos presentes na base de dados. Como vimos na análise explanatória é factível a seguinte hierarquia dos dados:

Insurance			
Fumantes		**Não Fumantes**	
Obeso	Não obeso	Obeso	Não obeso

Tabela 2: Hierarquia dos dados

Para tal, como as covariáveis de prática de tabagismo e obesidade são binárias e considerando que as demais variáveis exercem a mesma influência sobre a variável resposta, definimos as predições $\bar{Y}$ dos dados:

$$\bar{Y}^{i,j} = \beta_0^{i,j} + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_n X_n; \tag{3}$$

onde:

$$i = \begin{cases} 0, \text{se não fumante} \\ 1, \text{se fumante} \end{cases};$$

$$j = \begin{cases} 0, \text{se não obeso} \\ 1, \text{se obeso} \end{cases}.$$

2.2 Métricas de avaliação dos modelos

As métricas de avaliação dos modelos desempenham um papel fundamental na seleção e validação dos modelos de regressão. Duas métricas amplamente utilizadas são o coeficiente de determinação ($\mathbf{R^2}$) e o Critério de Informação de Akaike (AIC).

O coeficiente de determinação R^2 é uma medida que indica a proporção da variância da variável dependente que é explicada pelas variáveis independentes em um modelo de regressão. A fórmula para o R^2 em um modelo de regressão linear simples é:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \tag{4}$$

onde:

- y_i são os valores observados da variável dependente.

- $\hat{y}_i$ são os valores preditos pelo modelo.
- $\bar{y}$ é a média dos valores observados da variável dependente.
- $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ é a soma dos quadrados dos resíduos (SSE - Sum of Squared Errors).
- $\sum_{i=1}^{n}(y_i - \bar{y})^2$ é a soma total dos quadrados (TSS - Total Sum of Squares).

O R^2 mede a proporção da variabilidade da variável resposta que é explicada pelo modelo, variando de 0 a 1, onde valores mais próximos de 1 indicam um melhor ajuste do modelo aos dados.
O AIC oferece uma medida de ajuste do modelo que penaliza a complexidade, buscando um equilíbrio entre a precisão do ajuste e a simplicidade do modelo.

$$\text{AIC} = 2k - 2\ln(L); \tag{5}$$

onde:

- k é o número de parâmetros no modelo.
- L é a verossimilhança do modelo (ou seja, a probabilidade dos dados, dado o modelo).

Um valor mais baixo de AIC indica um modelo com melhor qualidade de ajuste. Ambas as métricas são essenciais para avaliar o desempenho dos modelos, permitindo comparar diferentes abordagens e selecionar o modelo mais apropriado para prever o turnover de funcionários com precisão e eficácia.

3 Resultados

Aplicando o método simples de regressão linear, obtem-se a seguinte previsão dos custos:

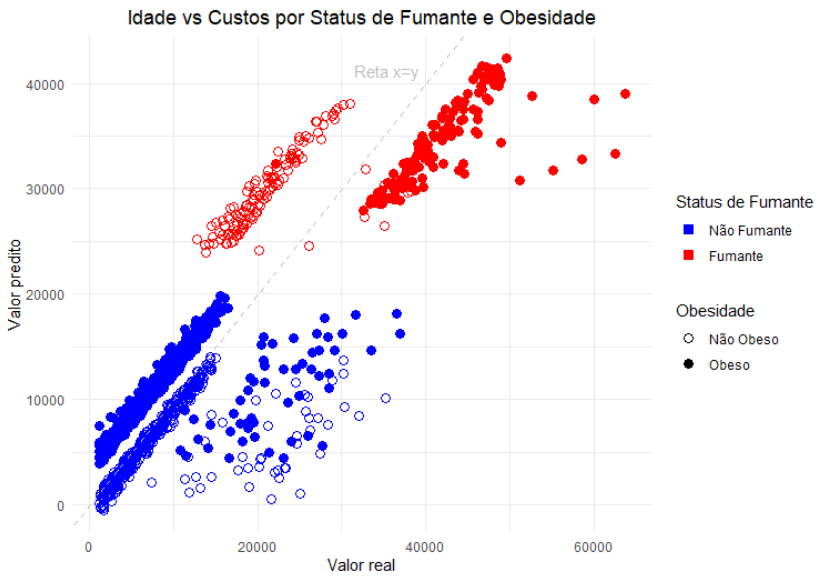


Figura 8: Comparação entre valor real e valor predito pelo primeiro modelo

Cada um dos grupos seguem uma tendência linear mas nem todas as predições estão condizentes com os valores reais. No grupo de fumantes não obesos por exemplo, o valor real é significativamente menor do que o valor predito, enquanto no grupo de fumantes obesos o valor real é maior do que o valor predito pelo modelo. Apesar disso, os resíduos seguem uma tendência normal centrada em 0:

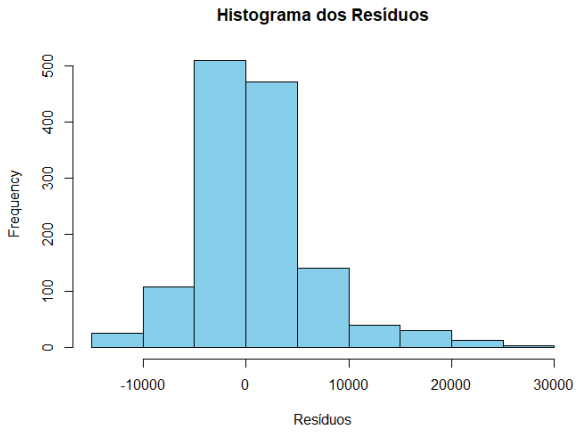


Figura 9: Resíduos do primeiro modelo

Aplicando o segundo modelo, de dois níveis, obtemos uma predição mais razoável para os valores:

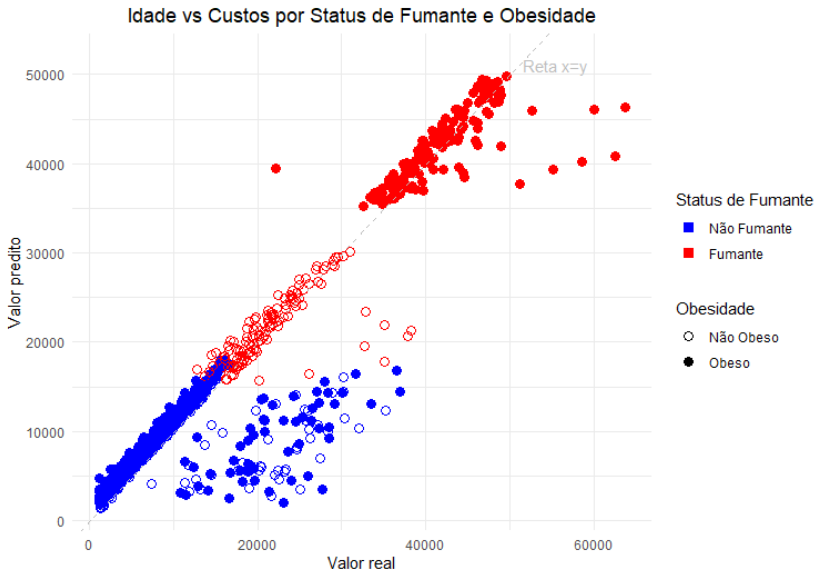


Figura 10: Comparação entre valor real e valor predito pelo segundo modelo

```
> print(coef(model))
$`smoker_yes:obese`
    (Intercept) square_age      bmi children  sex_male northeast northwest southeast
0:0   -1469.962   3.329511 114.3473 650.7454 -474.7824   1166.75  891.6763  341.7426
0:1   -2408.035   3.329511 114.3473 650.7454 -474.7824   1166.75  891.6763  341.7426
1:0   11934.186   3.329511 114.3473 650.7454 -474.7824   1166.75  891.6763  341.7426
1:1   30913.065   3.329511 114.3473 650.7454 -474.7824   1166.75  891.6763  341.7426
```

Figura 11: Coeficientes do segundo modelo

As predições e os valores reais estão mais próximos, no entanto ocorre uma discrepância na alálise dos resíduos. A maioria dos resíduos passa a ser ligeiramente negativa, ou seja, os valores preditos estão sendo ligeiramente maiores do que os valores reais. É perceptível ainda que alguns resíduos são relativamente altos (atingindo valores de até 25000).

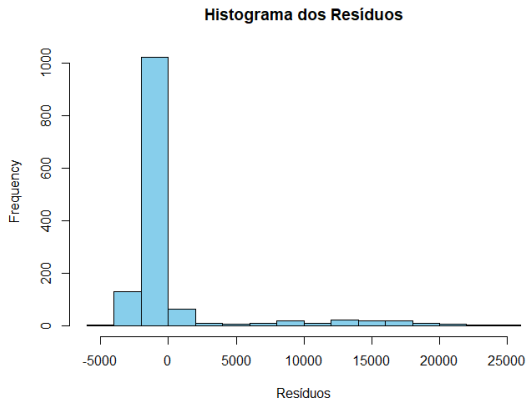


Figura 12: Resíduos do segundo modelo

Uma possível explicação para esse resultado é que algumas pessoas gastaram valores bem maiores que a média, o que é aceitável pois podem ter tido algum problema de saúde mais grave que precisou de tratamento e por consequência tiveram gastos maiores. No entanto, tais informações não são capturadas pelos dados disponíveis e portanto não pertencem ao escopo deste trabalho.

Por fim, as métricas calculadas.

Modelo	**AIC**	$\mathbf{R^2}$
Simples	27077.71	0.758
Multinível	26227.92	0.929

Tabela 3: Resultados das métricas dos modelos

Para o modelo simples obteve-se um $R^2 \approx 0.75$, o que é bom dado que as únicas manipulações feitas nos dados foram a adição de uma variável *obese* e a transformação da variável de idade para escala quadrática. No segundo modelo, o AIC foi significativamente menor, enquanto o coeficiente de determinação foi capaz de atingir aproximadamente 0.93.

4 Conclusão

Os resultados obtidos neste estudo proporcionam informações significativas sobre a relação entre as variáveis status de fumante, obesidade e os custos médicos associados. Inicialmente, a aplicação de um modelo simples de regressão linear revelou predições que, embora mostrassem uma tendência linear, apresentaram discrepâncias consideráveis entre os valores preditos e os valores reais. Observou-se que, especialmente no grupo de fumantes não obesos, as predições tendiam a superestimar os custos reais, enquanto no grupo de fumantes obesos, ocorreu o oposto, com subestimação dos custos.

Ao adotar um modelo multinível mais complexo, foi possível obter uma melhoria significativa nas predições dos custos. Os valores preditos ficaram mais próximos dos valores reais, embora ainda tenham sido observadas algumas discrepâncias nos resíduos.

A análise das métricas dos modelos revelou que o modelo multinível apresentou um AIC substancialmente inferior em comparação ao modelo simples, sugerindo uma melhor adaptação aos dados observados. Além disso, o coeficiente de determinação (R^2) para o modelo multinível foi significativamente mais alto, indicando que este modelo foi capaz de explicar uma maior proporção da variabilidade nos custos médicos em comparação ao modelo simples.

No decorrer desta análise, ficou evidente que dois fatores emergiram como os mais significativos na determinação dos custos médicos: o status de fumante e a condição de obesidade. Ambas as variáveis demonstraram ter um impacto substancial nas predições de custos, influenciando diretamente os resultados dos modelos de regressão. A diferenciação entre fumantes e não fumantes revelou discrepâncias consideráveis nos custos médicos previstos, enquanto a presença de obesidade adicionou uma camada adicional de complexidade, afetando a magnitude e a direção das predições. Esses achados sublinham a importância de considerar não

apenas características individuais, mas também suas interações, para uma compreensão mais completa e precisa dos determinantes dos custos de saúde.

Esses resultados destacam a importância de considerar não apenas variáveis individuais, como idade e status de fumante, mas também suas interações e efeitos conjuntos, como no caso da obesidade. No entanto, é importante ressaltar que a presença de *outliers* nos dados pode ter influenciado as predições, especialmente em casos onde ocorreram custos médicos excepcionalmente altos.

Em suma, este estudo demonstrou que a abordagem multinível oferece uma melhor adequação para modelar e prever os custos médicos.

Referências

Maria Eugênia Ferrão Barbosa and Cristiano Fernandes. Modelo multinível: uma aplicação a dados de avaliação educacional. *Estudos em Avaliação Educacional*, (22):135–154, 2000.

Francisco De la Cruz. Modelos multinivel. *Revista peruana de epidemiología*, 12(3):1–8, 2008.

Andrew Gelman and Jennifer Hill. *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press, 2006.

Andrew Gelman, Jennifer Hill, and Aki Vehtari. *Regression and other stories*. Cambridge University Press, 2021.

Brett Lantz. *Machine learning with R: expert techniques for predictive modeling*. Packt publishing ltd, 2019.