Análise Preditiva de Custos Médicos

Introdução à Ciência de Dados com R

Infnet

Resumo

A crescente variabilidade nos custos de seguros de saúde individuais representa um desafio significativo para seguradoras e segurados. Este estudo teve como objetivo desenvolver um modelo de regressão linear múltipla para identificar e quantificar o impacto de fatores demográficos e de estilo de vida nos custos médicos. Utilizando um conjunto de dados público com 1.338 observações, foi conduzida uma análise exploratória seguida pela construção de um modelo preditivo. Os resultados indicam que idade, Índice de Massa Corporal (IMC) e, predominantemente, o tabagismo são preditores estatisticamente significantes dos custos médicos. O modelo final demonstrou um forte poder explicativo, com um R^2 ajustado de 0.75, indicando que 75% da variabilidade nos custos pode ser explicada por essas variáveis. Notavelmente, o tabagismo foi identificado como o fator de maior impacto, associado a um aumento médio nos custos de aproximadamente 4,7 vezes em comparação a não fumantes, mantendo os demais fatores constantes. A validação dos pressupostos do modelo confirmou sua robustez, fornecendo uma ferramenta valiosa para a precificação de risco e a formulação de políticas de saúde.

Sumário

1. Introdução

1.1. Contextualização

Os custos no setor de saúde são um tema de intenso debate global. No âmbito individual, os prêmios de seguro de saúde são influenciados por uma complexa interação de fatores de risco, que vão desde características demográficas imutáveis até escolhas de estilo de vida. A capacidade de modelar e prever esses custos é fundamental para a sustentabilidade financeira das seguradoras e para a formulação de políticas de saúde pública mais eficazes.

1.2. Problema de Pesquisa e Objetivos

O problema central abordado neste estudo é a alta variabilidade nos custos médicos entre diferentes indivíduos. O objetivo principal é desenvolver um modelo de regressão linear múltipla para identificar e quantificar o impacto de fatores selecionados (idade, IMC, tabagismo) nos custos médicos anuais cobrados pelas seguradoras. Objetivos secundários incluem a validação estatística do modelo e a interpretação prática de seus resultados.

1.3. Hipóteses do Estudo

A análise foi guiada pelas seguintes hipóteses a priori:

- **H1:** A idade está positivamente correlacionada com os custos médicos, refletindo o aumento natural dos cuidados de saúde ao longo da vida.

- **H2:** Um Índice de Massa Corporal (IMC) mais elevado está associado a custos médicos maiores, devido a riscos de saúde relacionados.
- **H3:** O status de fumante tem um impacto positivo e estatisticamente significante nos custos médicos.

1.4. Relevância e Justificativa

Este estudo é relevante para múltiplos stakeholders. Para as seguradoras, oferece um modelo quantitativo para aprimorar a precificação de risco. Para gestores de saúde pública, fornece evidências quantificáveis do ônus financeiro de fatores de risco como o tabagismo, justificando investimentos em campanhas de prevenção. Para o indivíduo, elucida o impacto financeiro direto de escolhas de estilo de vida.

2. Metodologia

2.1. Descrição do Conjunto de Dados

O estudo utilizou o conjunto de dados "Medical Cost Personal Datasets", disponível publicamente na plataforma Kaggle (Link). O dataset contém 1.338 observações e 7 variáveis, representando uma amostra de segurados e seus respectivos custos anuais.

2.2. Dicionário de Variáveis

Variável	Descrição	Tipo	Papel no Modelo
age	Idade do segurado	Quantitativa	Independente
sex	Sexo biológico do segurado	Categórica	Controle (não usado no modelo final)
bmi	Índice de Massa Corporal (kg/m^2)	Quantitativa	Independente
children	Número de dependentes	Quantitativa	Controle (não usado no modelo final)
smoker	Status de fumante (sim/não)	Categórica	Independente
region	Região geográfica nos EUA	Categórica	Controle (não usado no modelo final)
charges	Custos médicos anuais faturados	Quantitativa	**Dependente**

2.3. Tratamento e Preparação dos Dados

A preparação dos dados iniciou-se com a verificação de valores ausentes, não sendo encontrado nenhum. As variáveis categóricas (sex, smoker, region) foram convertidas para o tipo factor para correta interpretação pelo software estatístico R. A análise da variável

dependente charges revelou uma forte assimetria positiva. Para atender aos pressupostos do modelo de regressão linear, foi aplicada uma transformação logarítmica (log(charges)), resultando em uma distribuição mais simétrica.

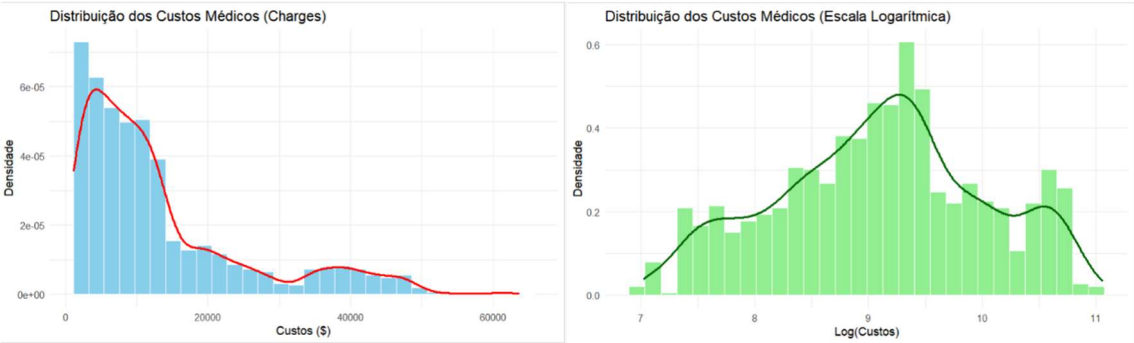
2.4. Análise Estatística

A principal técnica empregada foi a Regressão Linear Múltipla, ajustada pelo método de Mínimos Quadrados Ordinários. O software R e o ambiente RStudio foram utilizados para toda a análise. A avaliação do modelo foi baseada no R^2 ajustado, no teste F de significância global e nos testes t para os coeficientes individuais. A validação foi concluída com uma análise gráfica dos resíduos do modelo.

3. Análise Exploratória de Dados (AED)

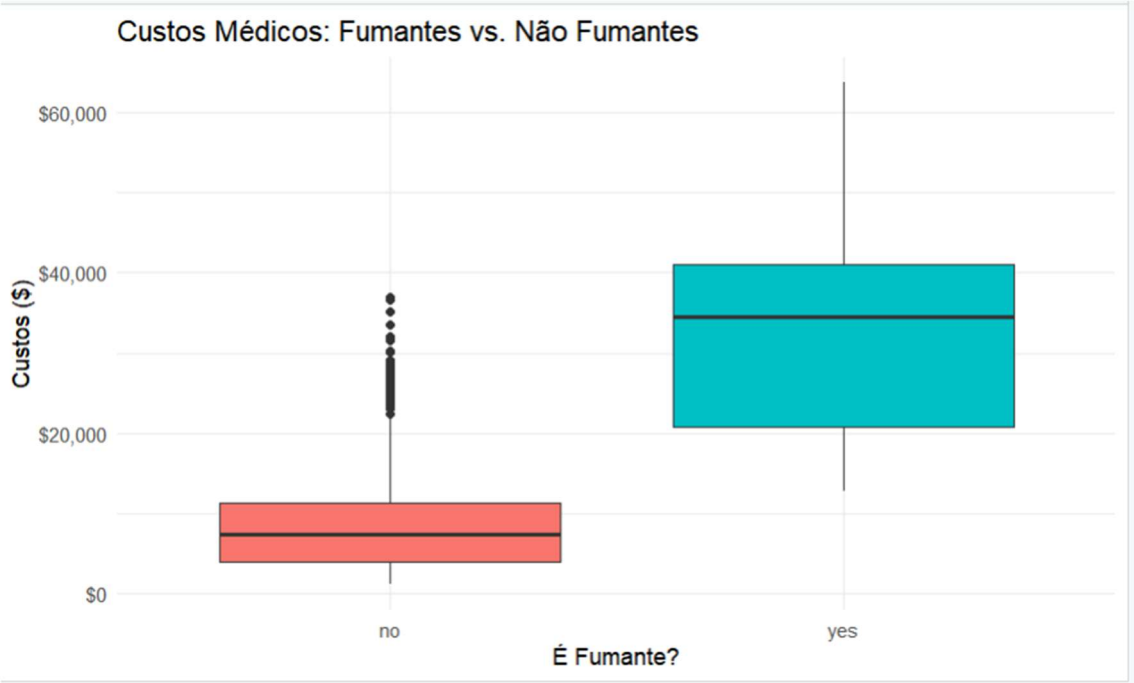
3.1. Análise da Variável Dependente

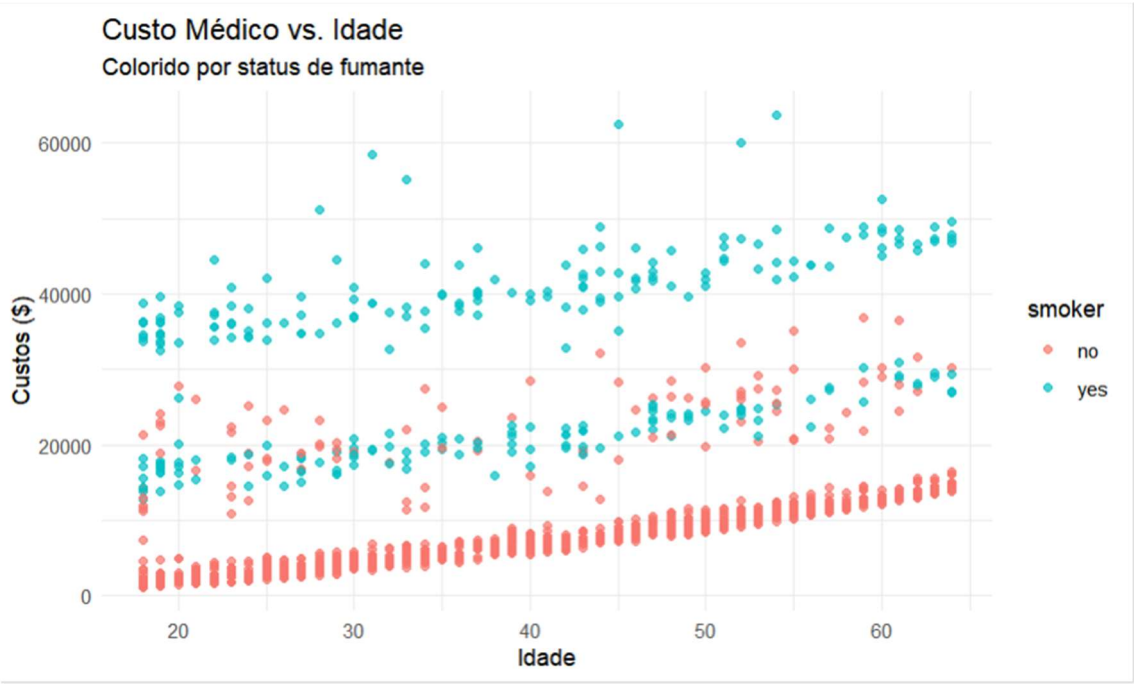
A distribuição da variável charges mostrou-se concentrada em valores baixos, com uma longa cauda à direita, indicando que poucos indivíduos possuem custos extremamente altos. Após a transformação logarítmica, a distribuição de log(charges) aproximou-se significativamente de uma distribuição normal, validando a abordagem para a modelagem.



3.2. Relação entre Custos e Preditores

A análise bivariada revelou padrões claros. O boxplot comparando fumantes e não fumantes indicou uma diferença drástica, com a mediana e a dispersão dos custos sendo substancialmente maiores para o grupo de fumantes. O gráfico de dispersão entre idade e custos, quando segmentado por status de fumante, mostrou não apenas que os fumantes pagam mais, mas que o aumento dos custos com a idade é mais acentuado para este grupo. Uma tendência positiva, porém menos acentuada, também foi observada entre o IMC e os custos.





4. Resultados

4.1. Apresentação do Modelo Final

Com base na AED, foi ajustado um modelo de regressão linear múltipla para prever log(charges) a partir das variáveis age, bmi e smoker. A tabela a seguir apresenta os coeficientes estimados e as estatísticas associadas.

4.2. Interpretação dos Coeficientes

Tabela 1: Resumo do Modelo de Regressão Linear Final

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 7.0738088  0.0715965  98.801  < 2e-16 ***
age         0.0351455  0.0009113  38.565  < 2e-16 ***
bmi         0.0107729  0.0020990   5.132 3.29e-07 ***
smokeryes   1.5458693  0.0315282  49.031  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4652 on 1334 degrees of freedom
Multiple R-squared:  0.7446,    Adjusted R-squared:  0.744
F-statistic:  1296 on 3 and 1334 DF,  p-value: < 2.2e-16
```

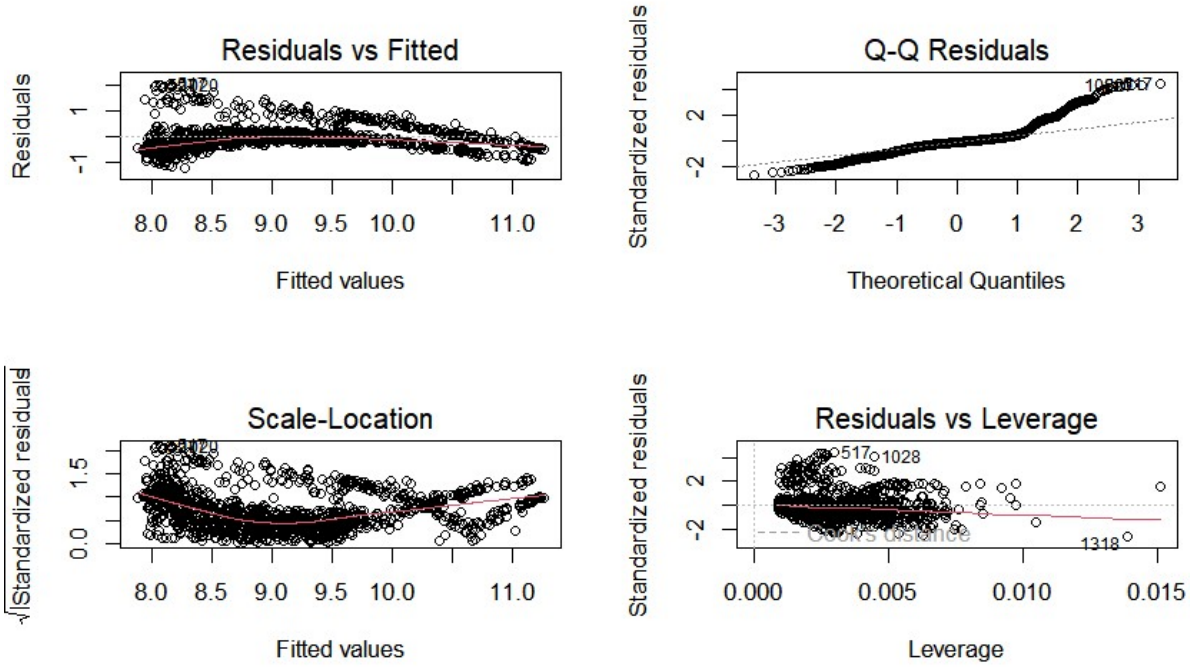
- **Idade (age):** Para cada ano adicional de idade, o logaritmo do custo médico aumenta em média 0.0351, mantendo as outras variáveis constantes.
- **IMC (bmi):** Para cada ponto adicional no IMC, o logaritmo do custo aumenta em média 0.0108.
- **Tabagismo (smokeryes):** Ser fumante aumenta o logaritmo do custo em 1.5459 em comparação a um não fumante de mesma idade e IMC. Em termos práticos, isso significa que os custos para um fumante são, em média, **exp(1.5459) ≈ 4.69 vezes maiores**. Todos os coeficientes foram altamente significantes ($p < 0.001$).

4.3. Qualidade do Ajuste do Modelo

O R^2 ajustado do modelo foi de 0.7446, indicando que aproximadamente **75% da variabilidade no logaritmo dos custos médicos é explicada pelo modelo**. O teste F global apresentou um p-valor extremamente baixo (< 2.2e-16), confirmando que o modelo como um todo é estatisticamente significante e possui um alto poder preditivo.

4.4. Diagnóstico e Validação dos Pressupostos

A análise dos resíduos, realizada por meio de gráficos de diagnóstico, validou a adequação do modelo. O gráfico de Resíduos vs. Valores Ajustados não mostrou padrões evidentes, confirmando a ***homoscedasticidade*** (É um pressuposto fundamental na análise de regressão onde a variância dos erros do modelo é constante em todos os níveis das variáveis independentes). O gráfico Q-Q Normal indicou que os resíduos seguem aproximadamente uma distribuição normal. Por fim, a análise de alavancagem não identificou pontos influentes que pudessem distorcer os resultados.



5. Conclusões

5.1. Síntese dos Achados

Este estudo demonstrou com sucesso a aplicação de um modelo de regressão linear para explicar os custos médicos. As hipóteses iniciais foram confirmadas: idade, IMC e, mais notavelmente, o tabagismo são preditores poderosos e estatisticamente significantes. O modelo final é robusto, válido e explica 75% da variação nos dados, com o tabagismo emergindo como o fator de maior impacto, ao quase quintuplicar os custos esperados.

5.2. Implicações Práticas

Os resultados oferecem insights valiosos. Seguradoras podem utilizar os coeficientes do modelo para refinar seus algoritmos de precificação. Para a saúde pública, o efeito multiplicador de 4.73x associado ao fumo é um argumento poderoso para justificar e intensificar políticas anti-tabagismo, enfatizando não apenas os benefícios para a saúde, mas também o enorme fardo financeiro.

5.3. Limitações do Estudo

O modelo, apesar de robusto, possui limitações. Fatores como condições pré-existentes, frequência de atividade física, hábitos alimentares ou histórico familiar não estavam disponíveis no dataset e podem responder por parte da variabilidade não explicada. Além disso, os dados são transversais e não capturam a evolução dos custos ao longo do tempo para um mesmo indivíduo.

5.4. Sugestões para Pesquisas Futuras

Pesquisas futuras poderiam enriquecer a análise incluindo um leque maior de variáveis de estilo de vida. A aplicação de modelos de machine learning mais complexos, como

Gradient Boosting ou Redes Neurais, poderia ser explorada para comparar o poder preditivo. Um estudo longitudinal, acompanhando os mesmos indivíduos por vários anos, ofereceria insights mais profundos sobre a dinâmica dos custos de saúde.

5. Código em R com comentários

```
# ------------------------------------------------------------------
# PROJETO FINAL: MODELO DE REGRESSÃO LINEAR
# GRUPO INFNET "R"
# Data: 08/09/2025
# Descrição: Análise dos fatores que influenciam os custos médicos.
# ------------------------------------------------------------------

# 1ª ETAPA: CONFIGURAÇÃO DO AMBIENTE
# ----------------------------------

# Instalar pacotes (quem ainda não tiver)
#install.packages("tidyverse")
#install.packages("corrplot")

# Carregar os pacotes que vamos utilizar
library(tidyverse)
library(corrplot)


# 2ª ETAPA: CARGA E VERIFICAÇÃO DOS DADOS
# ---------------------------------------

# Carregar o dataset a partir da pasta do projeto.
# Como o arquivo está na subpasta 'data', o caminho é
"data/insurance.csv".
dados <- read.csv("data/insurance.csv")

# Verificar se os dados foram carregados corretamente
head(dados)  # 6 primeiras linhas
str(dados)   # Estrutura do dataset

# PASSO 3: LIMPEZA E PREPARAÇÃO
# -----------------------------

# A função 'str()' nos mostrou que sex, smoker e region são 'chr'.
# Vamos convertê-los para o tipo 'factor', que é o tipo correto para
# variáveis categóricas no R.
dados <- dados %>%
  mutate(
    sex = as.factor(sex),
    smoker = as.factor(smoker),
    region = as.factor(region)
  )

# Vamos verificar a estrutura novamente para confirmar a mudança.
# Agora você verá 'Factor' ao lado dessas variáveis.
str(dados)

# Verificar se há valores ausentes (NA) no dataset.
```

```
# Este comando soma os NAs por coluna.
colSums(is.na(dados))
# Resultado: Zero para todos. Ótimo, nosso dataset é limpo!

# PASSO 4: ANÁLISE EXPLORATÓRIA
# -----------------------------

# a) Análise da Variável Dependente (Charges)
# Vamos criar um histograma para ver a distribuição dos custos.
ggplot(dados, aes(x = charges)) +
  geom_histogram(aes(y = ..density..), fill = "skyblue", color =
"white", bins = 30) +
  geom_density(col = "red", size = 1) +
  labs(
    title = "Distribuição dos Custos Médicos (Charges)",
    x = "Custos ($)",
    y = "Densidade"
  ) +
  theme_minimal()

# b) Visualizando a transformação logarítmica
ggplot(dados, aes(x = log(charges))) +
  geom_histogram(aes(y = ..density..), fill = "lightgreen", color =
"white", bins = 30) +
  geom_density(col = "darkgreen", size = 1) +
  labs(
    title = "Distribuição dos Custos Médicos (Escala Logarítmica)",
    x = "Log(Custos)",
    y = "Densidade"
  ) +
  theme_minimal()

# c) Relação de Charges com Preditores Categóricos

# Charges vs. Smoker
ggplot(dados, aes(x = smoker, y = charges, fill = smoker)) +
  geom_boxplot(show.legend = FALSE) +
  labs(
    title = "Custos Médicos: Fumantes vs. Não Fumantes",
    x = "É Fumante?",
    y = "Custos ($)"
  ) +
  scale_y_continuous(labels = scales::dollar) + # Formata o eixo Y
para dólar
  theme_minimal()

# d) Relação de Charges com Preditores Quantitativos

# Charges vs. Age (colorindo por smoker para mais insights)
ggplot(dados, aes(x = age, y = charges)) +
  geom_point(aes(color = smoker), alpha = 0.7) +
  labs(
    title = "Custo Médico vs. Idade",
    subtitle = "Colorido por status de fumante",
    x = "Idade",
    y = "Custos ($)"
  ) +
```

```
  theme_minimal()

# PASSO 5: MODELAGEM - CONSTRUÇÃO DO MODELO
# ----------------------------------------

# Vamos construir nosso primeiro modelo usando os preditores que a
Análise Exploratória sugeriu serem os mais importantes.
# Nossa variável dependente será log(charges) para satisfazer os
pressupostos do modelo.

modelo_1 <- lm(log(charges) ~ age + bmi + smoker, data = dados)

# A função summary() nos dá um relatório completo sobre o nosso
modelo.
# Este será o output mais importante da análise!

summary(modelo_1)

# PASSO 6: DIAGNÓSTICO DO MODELO
# ------------------------------

# A função plot() aplicada a um objeto 'lm' gera 4 gráficos de
diagnóstico.
# Usamos par(mfrow = c(2, 2)) para arranjar os 4 gráficos em uma grade
2x2.

par(mfrow = c(2, 2))
plot(modelo_1)

# Vamos resetar o layout gráfico para o padrão depois.
par(mfrow = c(1, 1))
```