Predicting Medical Insurance Costs using Lifestyle and Demographic Factors

Paulina Skurzak, Gillian Dewsbury, Olivia Webster

Summary:

This project investigates how demographic and lifestyle characteristics influence medical insurance costs using the *Medical Insurance Cost dataset* (Kaggle, 1,338 observations, 7 variables).
We modeled total insurance charges based on age, sex, BMI, number of children, smoking status, and region.
Exploratory analysis showed that smoking and BMI are dominant cost drivers. Smokers have median charges around $35,000 compared to $8,000 for non-smokers. Costs increase with both age and BMI, while region and sex have minimal effects.
After testing multiple regression models, the best performing model included an interaction between BMI and smoking, achieving an adjusted R^2 = 0.84. This interaction reveals that BMI strongly amplifies costs for smokers, emphasizing how lifestyle factors compound financial risk in healthcare.

Final Model:

$$E(y) = -2223.454 + 263.620x_1 - 500.146x_2 + 23.533x_3 - 20415.611x_4 + 516.403x_5 - 585.478x_6 - 1210.131x_7 - 1231.108x_8 + 1443.096x_3x_4$$

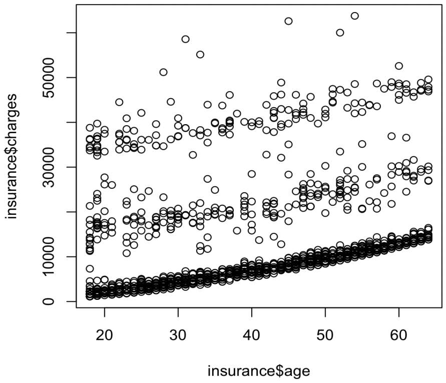
Where:

$y =$ medical insurance cost
$x_1 =$ age in years
$x_2 =$ sex (1 if male, 0 if female)
$x_3 =$ bmi (body mass index)
$x_4 =$ smoker (1 if smoker, 0 if not)

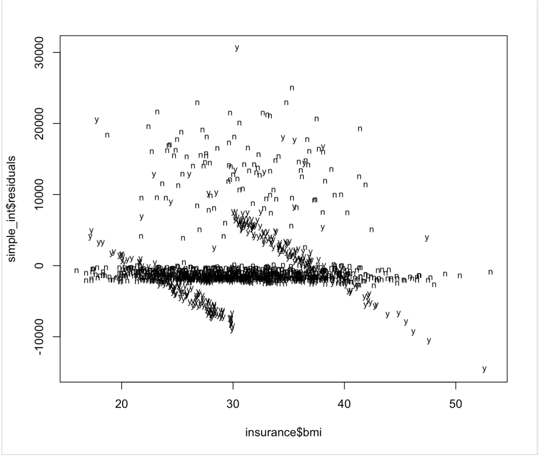
$x_5 =$ number of children
$x_6 =$ region (1 if northwest, 0 if not)
$x_7 =$ region (1 if southeast, 0 if not)
$x_8 =$ region (1 if southwest, 0 if not)

Interesting Plots:

The age vs charges plot was quite intriguing because, while it definitely showed a linear relationship, there were 3 different lines making up the graph. We were not able to determine what the cause of this relationship was, but suspect it may be from smoking and bmi factors.



The residual plots were also interesting in the sense that the residuals seemed to be smaller for non-smokers than smokers, as shown in the graph. This may be that our model was more accurate for non-smokers than smokers.


simple_int$residuals
insurance$bmi