Medical Cost Analysis

Jon Duea & John Clark

2025-04-09

# Introduction

The goal of the STAT 301 project is to apply ANOVA methods to a dataset, so with that in mind we chose this dataset from Kaggle ([Choi, n.d.](#ref-choi_medical)), the original source can be found here as well ([Stednick 2022](#ref-stednick_2022_data)). This dataset contains information regarding medical costs billed to individuals by insurance as well as information about the individuals including their age, sex, bmi, number of children, smoking status, and region of the US where they reside.

The research questions we are aiming to address using this dataset are:

1. Do medical costs billed by insurance differ significantly based on smoking status (smoker vs. non-smoker) and BMI categories (underweight, normal weight, overweight, obese) independently?
   * Smoking Status:
   * BMI Category:
     + At least one BMI categories mean charges significantly differ from the others.
2. Is there a significant interaction between smoking status and BMI categories on medical costs billed by insurance, meaning that the effect of smoking on medical costs billed by insurance depends on the BMI category of the individual?
   * There is no interaction between smoking status and BMI categories on medical costs billed by insurance.
   * There is a significant interaction between smoking status and BMI categories on medical costs billed by insurance.

In order to answer these questions we have chosen to utilize three variable from our dataset:

* BMI (Body Mass Index):
  + Provides an understanding of body weights that are relatively high or low relative to height.
  + Uses the ratio of weight to height:
  + According to the CDC, BMI categories for adults 20 and older ([Disease Control and Prevention 2024](#Xc00f9e21965538f2ffb877e4296cba1129ffca6)) are as follows:
    - Underweight: , Healthy Weight: , Overweight: , Obesity:
  + The dataset records BMI as a numeric variable, but we will be encoding it as a categorical variable with 4 levels according to the CDC’s definitions of BMI categories.
* Smoker (Smoking Status):
  + Categorical variable (2 levels), True or False, Indicates whether an individual smokes or not.
* Charges (Medical Costs Billed by Health Insurance):
  + Numeric variable, Measured in USD

# Methods

Since we’re working with 2 categorical independent variables and one numeric dependent variable, we have chosen two-way ANOVA as the appropriate method for exploring the data and answering our research questions.

# Conclusion

# References

Choi, Miri. n.d. “Medical Cost Personal Datasets.” www.kaggle.com. <https://www.kaggle.com/datasets/mirichoi0218/insurance/data>.

Disease Control, Centers for, and Prevention. 2024. “Adult BMI Categories.” CDC. <https://www.cdc.gov/bmi/adult-calculator/bmi-categories.html>.

Stednick, Zach. 2022. “Data for Machine Learning with r.” GitHub. <https://github.com/stedy/Machine-Learning-with-R-datasets/blob/master/insurance.csv>.