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 sex and region independently?
   * Sex:
   * Region:
     + At least one regions mean charges are significantly different from the others.
2. Is there a significant interaction between sex of the primary beneficiary and region on medical costs billed by insurance, meaning that the effect of sex on medical costs billed by insurance depends on the region of the US where the individual resides?
   * There is no interaction between sex and region on medical costs billed by insurance.
   * There is a significant interaction between sex and region on medical costs billed by insurance.

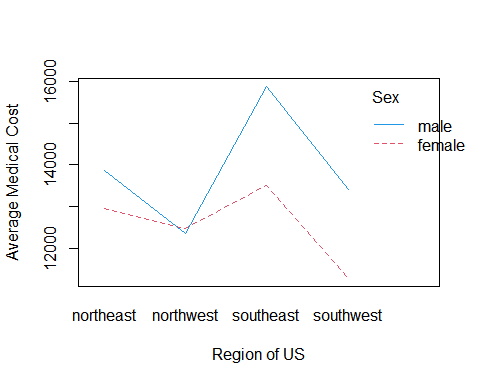
In order to answer these questions we have chosen to utilize three variable from our dataset:

* Sex:
  + Sex of the primary beneficiary
  + Categorical variable with 2 levels (Male or Female)
* Region:
  + Residential area of the primary beneficiary in the US
  + Categorical variable with 4 levels (Northeast, Northwest, Southeast, Southwest)
* Charges:
  + Medical Costs Billed by Health Insurance
  + Numeric variable, Measured in USD

# Methods

Since we’re working with two categorical independent variables and one numeric dependent variable, a two-way ANOVA was performed to analyze the effect of sex and region on medical costs billed by insurance. A two-way ANOVA is the appropriate method for our analysis because our research questions are interested in investigating the main effect of sex and region (both categorical variables with two or more levels) as well as the interaction between the two.

|  | Df | Sum.Sq | Mean.Sq | F.value | Pr..F. |
| --- | --- | --- | --- | --- | --- |
| region | 3 | 1300759681 | 433586560 | 2.9752836 | 0.0306591 |
| sex | 1 | 613069167 | 613069167 | 4.2068985 | 0.0404554 |
| region:sex | 3 | 340173346 | 113391115 | 0.7780931 | 0.5062076 |
| Residuals | 1330 | 193820219374 | 145729488 | NA | NA |



# References

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

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