Selected Topic: Medical costs across different regions of the US

Title: Modeling medical cost differences across the US

Group: Kyle Smith and Jon Duea

April 12th, 2025

STAT 311 Regression Analysis, Spring 2025

Dataset: [Medical Cost Personal Datasets - Miri Choi - Kaggle](https://www.kaggle.com/datasets/mirichoi0218/insurance/data)

Abstract:

Our project, “Modeling medical cost differences across the US”, aims to explore possible differences in medical costs across the US as well as the impact the variables region, age, sex, BMI, number of children, and smoking status have on those costs. Through our analysis, we hope to uncover deeper insights into the effects of these variables on medical costs and determine whether their effects remain constant across different regions of the US.

To accomplish this, we have sourced a dataset of medical costs from [Kaggle](https://www.kaggle.com/datasets/mirichoi0218/insurance/data). This dataset provides information for 1338 individuals across four different regions of the US (Northeast, Northwest, Southeast, Southwest) and includes information about the age, sex, BMI, number of children, smoking status, region of residence, and medical costs billed by insurance for everyone. Using this dataset, we plan to construct several models, starting with a first-order model, to compare the effects of adding in new terms; taking this procedural approach to building on reduced models towards a complete model, we hope to find the answers to the questions most pressing to us and ultimately construct a useful model for predicting medical costs across the US.

References:

Choi, Miri. “Medical Cost Personal Datasets.” [*Www.kaggle.com*](https://usc-word-edit.officeapps.live.com/we/Www.kaggle.com), [www.kaggle.com/datasets/mirichoi0218/insurance/data](https://www.kaggle.com/datasets/mirichoi0218/insurance/data).