Insurance Analysis

Group 11: Isiah Montalvo, Tristan Dull, Winston Zheng, Trisha Agrawal

Topic & Motivation

We conducted our regression analysis on the Insurance dataset.

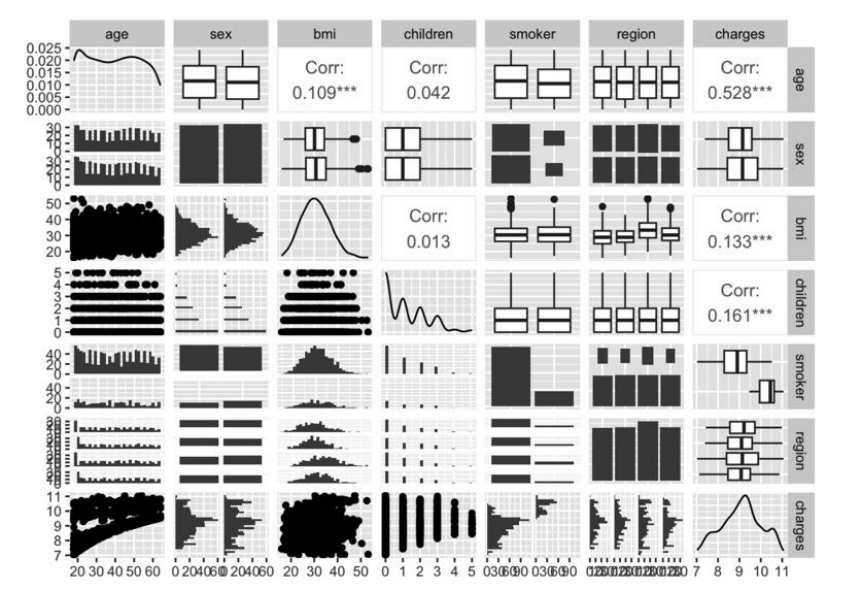
Motivated by generally high cost of medical bills.

Our goal is to predict how much a patient would be charged during a medical visit based on variables present in the dataset.

Null Hypothesis: $B_0 = \ldots B_n = 0$

Alternate Hypothesis: At-least one Beta is different.

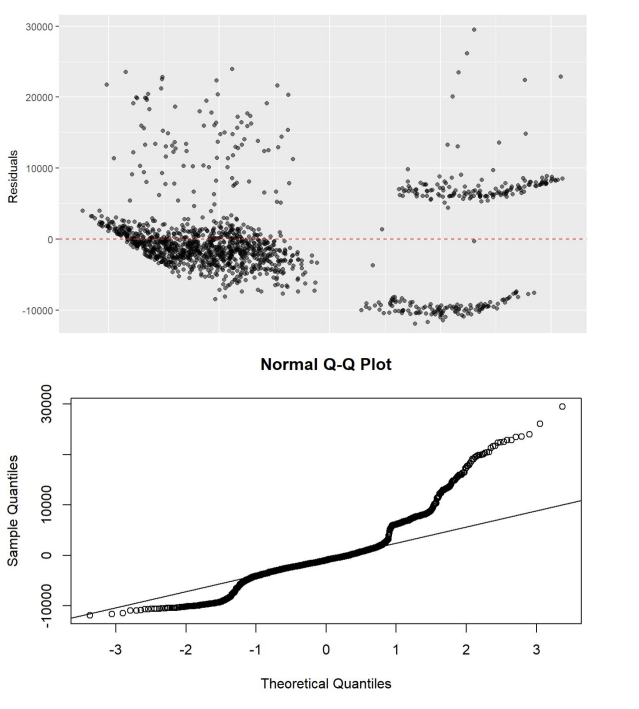
Data



- Predictor: Age of primary beneficiary
- Predictor: BMI
- Predictor: Children
- Predictor: Smoker
- Predictor: Sex
- Predictor: Region
- Response: Charges for medical costs billed by health insurance
- 1,338 rows of data

Highlights from EDA

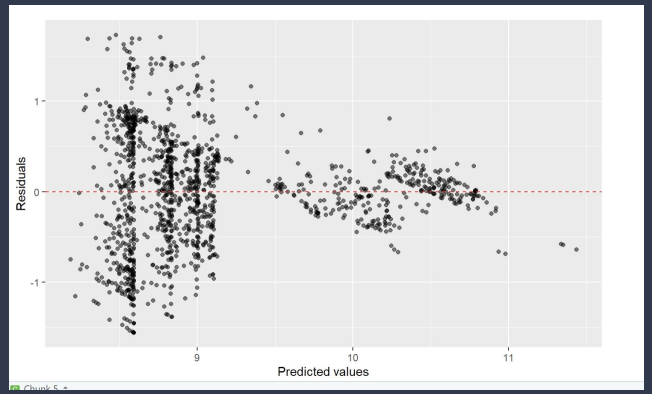
The plots below are the Residual and QQ plots BEFORE transformations (first order only)

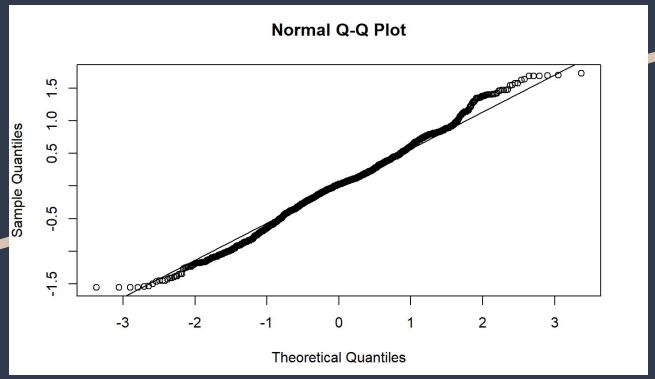


- The value between charges and age is the largest (0.528)
- Charges and bmi (corr = 0.133)
- Charges and children (corr = 0.161)
- Sex and Region do not significantly contribute to the charges
- Variance inflation factor (VIF) values are very close to 1

Final Model

Plots are Residual and QQ plots AFTER transformations (using final model)





$x_1 = e^{age}$

x_2 = bmi

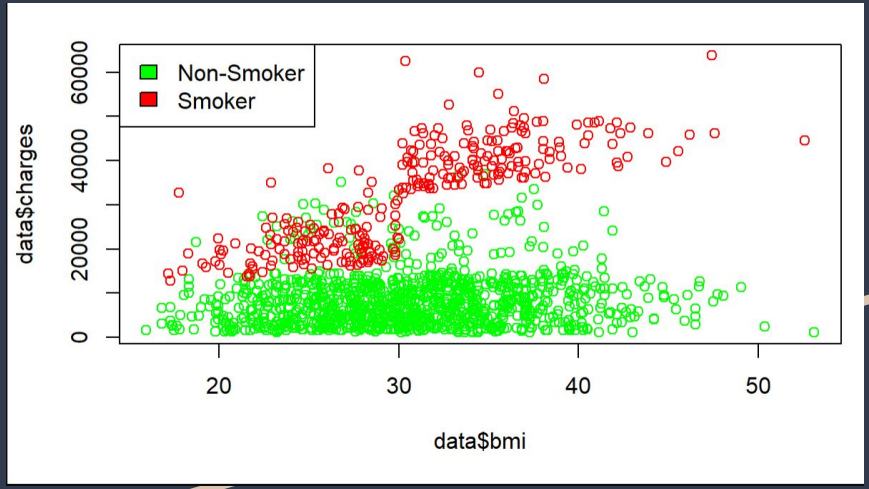
x_3 = children

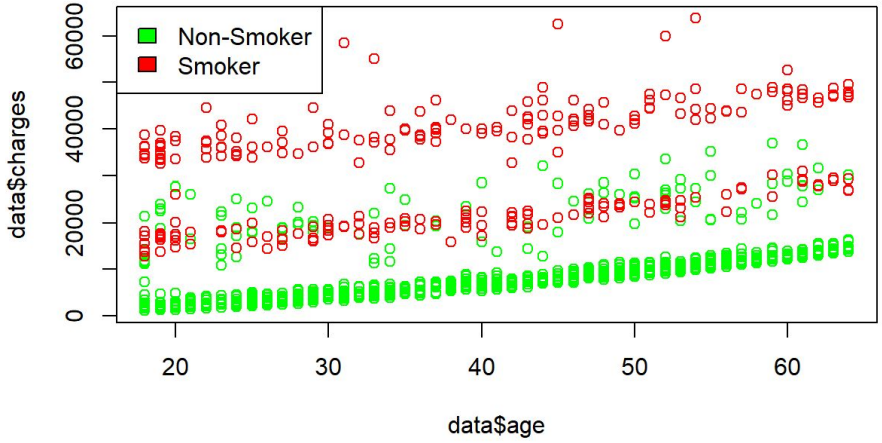
x_4 = smokeryes

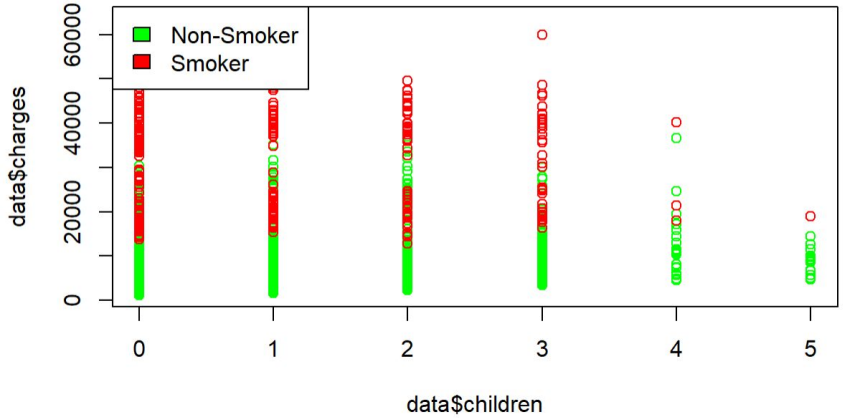
$\hat{y}$ = estimated log(Insurancecost)

$$\hat{y} = 7.209 + 7.312*10^{-28}x_1 + 7.312*10^{-28}x_2 + 2.813*10^{-01}x_3 + 4.246*10^{-01}x_4 - 9.285*10^{-56}x_1^2 - 1.135*10^{-03}x_2^2 - 3.667*10^{-02}x_3^2 - 1.316*10^{-28}x_1x_4 + 4.092*10^{-02}x_2x_4 - 1.386*10^{-01}x_3x_4$$

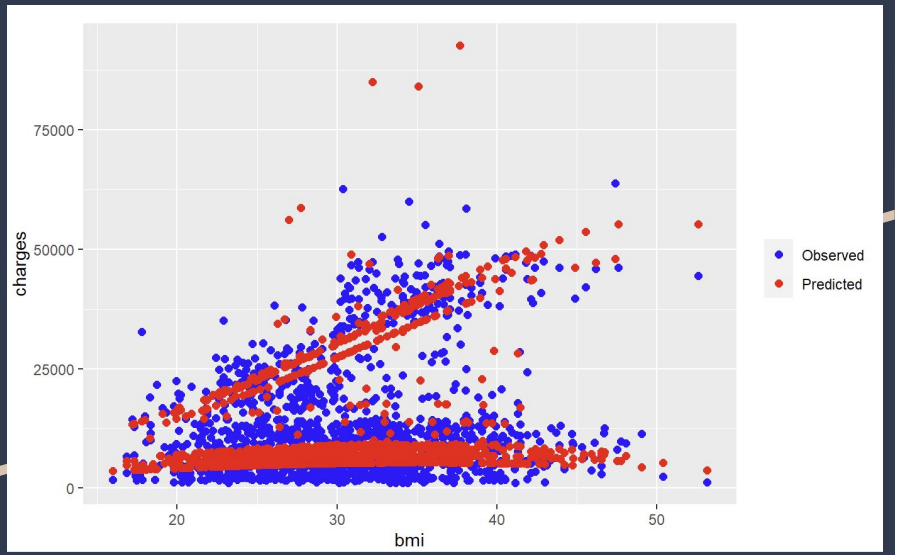
Interesting Findings







Limitations & Conclusion



- Transformations
  - log() on response variable (charges)
  - exp() on predictor variable (age)
- Adjusted R^2 = 0.5447
  - 54% of the variability in the dataset is explained by our model
  - This is a low R^2 value