STAT 462 Final Project

2023-11-10

Group Memeber

Eric Wu

Background

The Insurance dataset extracted from Kaggle.com describes the medical costs for over 1300 individuals based on their age, sex, BMI, how many children, Smoker or not, and which region in the U.S they lived. The goal of the project is to compare multiple ML method including multiple linear regression in order to find the best model with the lowest MSE value. In addition, there will be a detail diagnostic on the linear regression model for the assumptions, outliers, and other possible issues.

Exploratory Data Analysis

```
## [1] 0
```

Data Statistics

Table 1: Basic statistics of quantitative variables

	vars	n	mean	sd	min	max	range	se
age	1	1338	39.21	14.05	18.00	64.00	46.00	0.38
bmi	2	1338	30.66	6.10	15.96	53.13	37.17	0.17
children	3	1338	1.09	1.21	0.00	5.00	5.00	0.03
charges	4	1338	13270.42	12110.01	1121.87	63770.43	62648.55	331.07