JSC370 Midterm Report

Xinpeng Shan

2022/2/26

Introduction

As medical treatment becomes more costly in the period of the global pandemic, it becomes more difficult for people to cover the medical bills. In this assignment, the primary research question is to construct a linear regression model to predict the insurance costs that will be covered by the insurance company so that it will give people insight into the overall medical charges they need to pay if they have medical insurance. I will conduct EDA and create linear regression models on the dataset regarding the medical health bills covered by health insurance companies in the US. There are seven columns and 1338 observations in the dataset. By the end of the analysis, I will produce a linear regression model that will best explain the variation in medical costs covered by insurance companies in this dataset.

Methods

The dataset is an open-source dataset acquired from Kaggle, it is a sample of medical insurance information from the whole population in the US. I took the following steps when analyzing it:

Data wrangling and visualization

- Check the variable types in the dataset and if there are missing values.
- Visualize the distribution of each variable to check if there are distinctive shapes in those variables.
- Create visualizations of selected variables versus insurance charges to check whether the potential predictor is associated with the response (charges).

Linear regression

- Variable Selection To create a linear regression model that is both easy to use and has a high adjusted R^2 value, I will aim to create a model that has 2 to 4 predictors. I will first use the partial F test to eliminate the variables that are not statistically significant and use backward selection on choosing the variables.

- Model Validation The dataset is separated into training and testing data before building the model so that we could create a model that is useful for future predictions on insurance charges. Afterward, we use the same model built using the training dataset on the test dataset to conduct model validation. If there does not appear much difference on the adjusted R^2, same or fewer model violations, we can confirm that the model validation is done.

Preliminary Results

Description of dataset

Table 1: Information of variables in the dataset

Variable	Description	type	missing values
age	age of beneficiary	numarical	0
sex	gender of insurance contractor	categorical (male, female)	0
bmi	body mass index of beneficiary	numarical	0
children	number of children covered by the insurance	categorical (0, 1, 2, 3, 4, 5)	0

Variable	Description	type	missing values
smoker	if the beneficiary smokes	categorical (yes, no)	0
region	the beneficiary's residances in US	categorical (northeast, northwest, southeast, southwest)	0
charges	the medical cost covered by the company	numarical	0

Visualizations

From Figure 1, we can see that the distribution of charges is extremely right-skewed, which means that there are more observations that have charges that close to zero.

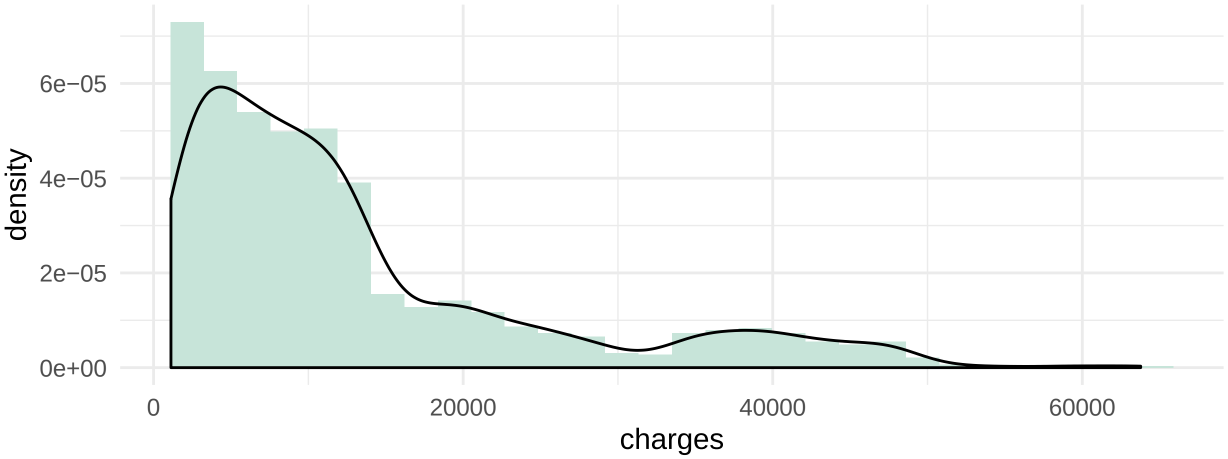
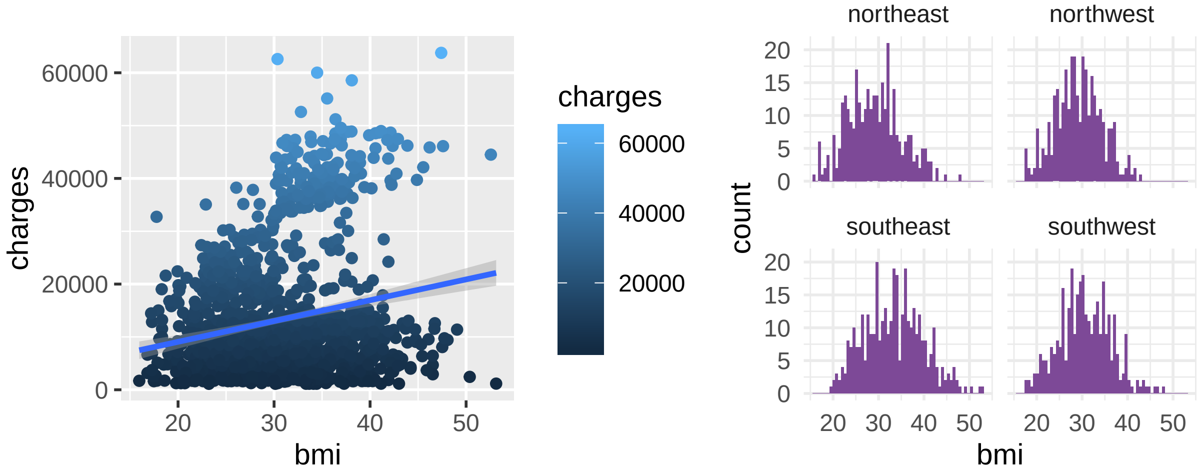


Figure 1. Distribution of charges

From Figure 2, we can see the distribution of bmi has a normal distributed shape where it is similar for different regions. from the scatterplot of charges versus bmi, we can notice that as bmi increases, the value of charges increases.

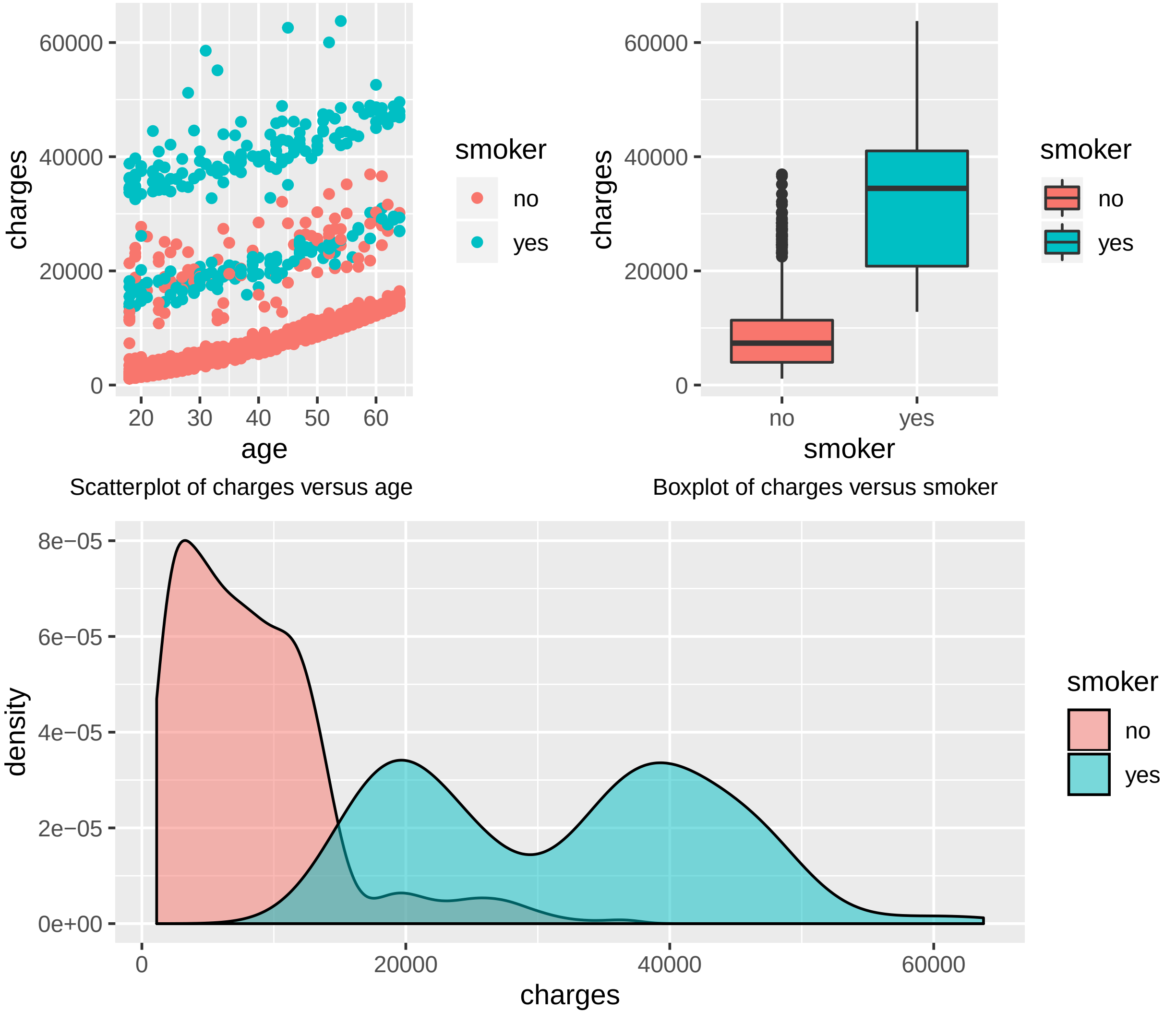


atterplot of charges versus bmi colored by charges

Histograms of bmi by region

Figure 2. Visualizations of bmi and region

From Figure 3, we can notice that smoker has an undeniable influence on the amount of charges. If the beneficiary smokes, the charges are higher which may indicate the worsening of health situation that they are in. From the leftmost plot, as age increases, the charges will become higher. Also, the overall charges at higher age for people who do not smoke are lower than the charges for people at lower ages who smoke.



Scatterplot of charges versus age

Boxplot of charges versus smoker

Distribution of charges by smoker

Figure3. Visualizations associated with age and smoker

From Figure 4, the stacked histogram indicates the distribution of age by region in the dataset. We can see that there are more observations for people who have age below 20, and a uniform distribution for people with an age higher than 20. This is caused by the selection bias when collecting the data. But there is a similar amount of data collected from different regions, which indicates that the dataset is random in location.

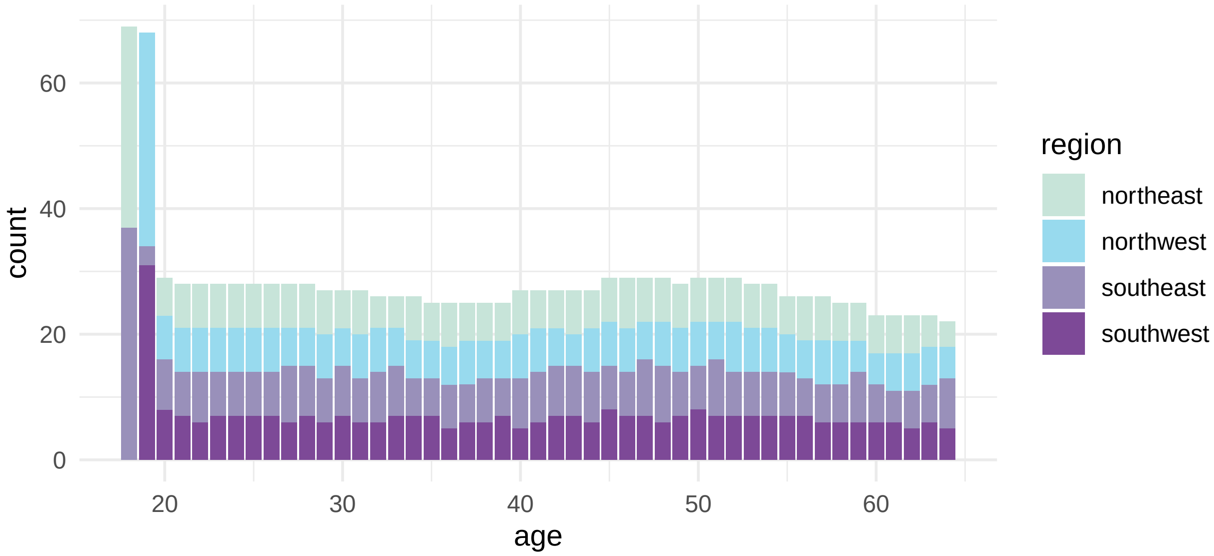
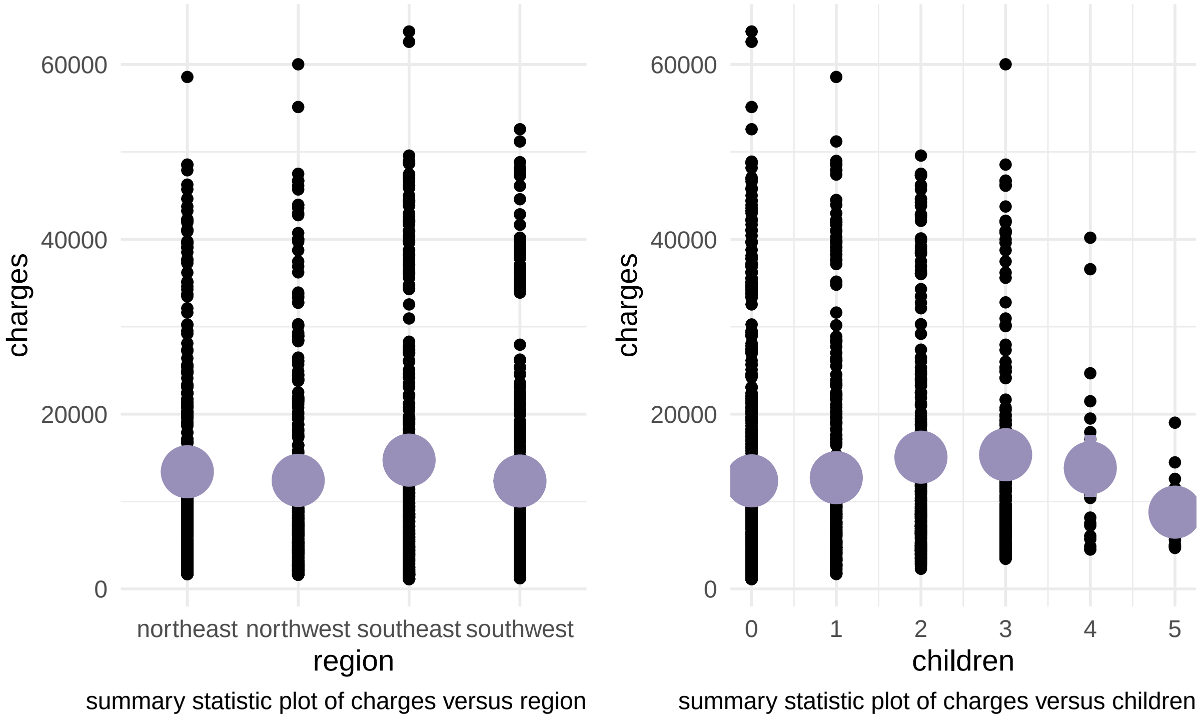


Figure 4. Stacked histogram of age by region

I created 2 statistic summary plots with mean as the purple point, to explore which variable affects the amount of insurance charges by a larger amount. From the statistic summary plots below, we can see that children can explain the variance in the amount of charges to a greater amount than region. From the statistic summary plot of children by charges, we can see that the mean charges are higher for 2 or 3 children and lower for other numbers of children.



summary statistic plot of charges versus region

summary statistic plot of charges versus children

Figure 5. Statistic summary plots of region and children

Linear Regression

- Choosing predictors

Before building the model, I did a train test split. The training dataset has 75% of data and the testing dataset has 25% of data.

Here are the number of rows in the two dataset:

Table 2: Number of rows in the two dataset

dataset	Number of Rows
train	993
test	345

Afterward, I built a linear regression model using all six potential predictors as predictors and charges as responses using the training dataset. From Table 3, we can see that age, bmi, smoker have a p-value that is extremely small, which means it is very unlikely to reject the null hypothesis for those variables that they do not explain the variation in charges.

Table 3: p-values of each predictor in the linear regression model built using training dataset

	Pr(>|t|)
age	< 2e-16
sexmale	0.48599
bmi	< 2e-16
children	0.000577
smokeryes	<2e-16
regionnorthwest	0.29430
regionsoutheast	0.33319
regionsouthwest	0.03229

Then, I builded two models and to decide which one to use as the final model.

- model 1: Prodictor: age, bmi, smoker, children
- model 2: Prodictor: age, bmi, smoker

The informations are summarised in the Table 4.

Table 4: adjusted R squared of the two models

model	Number of predictors	Adjusted R squared
model 1	4	0.7657707
model 2	3	0.7640206

From the table above, we can notice that although model 2 has a slightly (0.0015) lower Adj R squared compared with model 1, it is 1 predictor fewer, which is more feasible when choosing linear regression models. Therefore, we choose model 2 to be the final model.

- Analyse performance of the final model

From the Actual versus Prediction Plot, we can see that the points are scattered around the fitted line and there are very few points that appear to be far away from the line vertically, which indicates there are few outliers in the dataset. Since there is no point that is horizontally further away from the mean, there appears to be no significant leverage that pulls the regression line by a great amount.

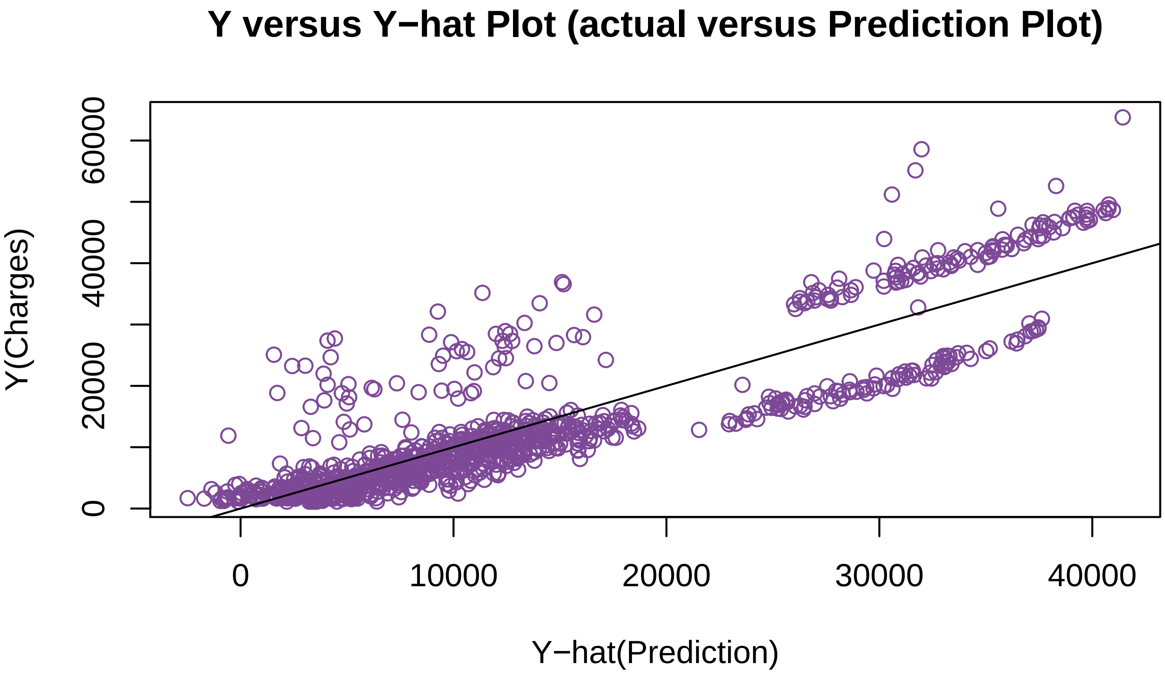


Figure 6. Actual versus Prediction plot of linear model built using training dataset

In the Residual vs. Fitted plot, there is a curved shape, which indicates a slight violation of the linearity assumption. From the Normal Q-Q plot, we can notice that there is some deviation at the right end of the Normal Q-Q plot, which means there is a slight violation of the normality assumption. There is an upward trend on the Scale-Location plot, which is caused by the unequal variance. And finally, in the Residual vs. Leverage plot, There are some amount of points that lie below Cook's distance, which demonstrates the influential points in the training dataset.

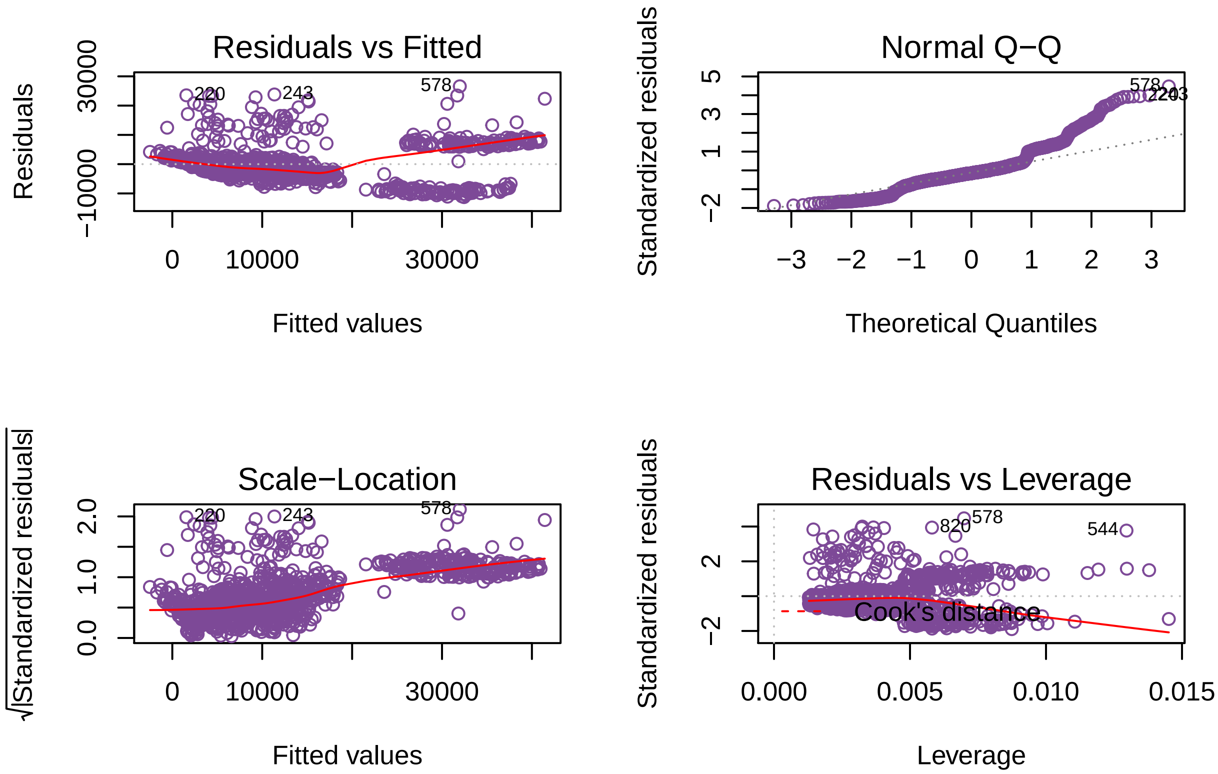


Figure 7. Plots to analysis the linear regression model built using the training dataset

- Validate the model using the test dataset

Afterward, I built another model using the same predictors as the final model but using the test dataset, to see the performance of the model. It achieves an adj R squared of 0.7048884, which means that our model explains around 70% of the variation in total charge in future data. There is no big difference in the violation of assumptions. The estimated coefficients do not differ by a great amount from the training dataset. Therefore, I can conclude that the model is validated.

Table 5: Estimated coefficients of the two models built using training and testing dataset respectively

dataset used to build the model	Coefficient of age	Coefficient of bmi	Coefficient of smokeryes
train	277.91	328.52	23581.99
test	205.65	289.05	24421.41

Summary

To summarize, I built a linear model to solve the question of predicting insurance charges paid by insurance companies in the US. The final model has 3 predictors including age, bmi, smoker. The linear model explains around 76% of the variation of the response variable (charges) in the training dataset and around 70% of the variation in the testing dataset. From the visualizations and the predictors in the linear model, we can notice whether the beneficiary smokes affect the amount of insurance bills that the insurance company pays by a large amount (more than 20000 higher on insurance bill for person who smokes), while variable bmi and age also have a positive association with charges (the coefficients are 289.05 and 205.65 respectively in test dataset). Since the higher insurance bills that the insurance company pays, the larger the bill from the hospital, we can see the drawbacks of smoking from the statistical results. To reduce the medical burden on individuals, we could start by quitting unhealthy habits like smoking and increasing the amount of time we spend on physical exercise to lower the bmi.

Appendix

Data source: https://www.kaggle.com/mirichoi0218/insurance?select=insurance.csv

Github repo: https://github.com/xinpeng13/JSC370/tree/main/midterm-report