**predicting medical expenses using linear regression**

3251-014

CHAEWON KIM

AUGUST 2018

[Objective 3](#_Toc521250996)

[What are you setting out to prove or predict? 3](#_Toc521250997)

[What is your rationale for there being a correlation in the data that you’re looking to confirm and/or exploit? 3](#_Toc521250998)

[Data Preparation: 3](#_Toc521250999)

[What was your data source? How good was the data quality? 3](#_Toc521251000)

[What did you need to do to procure it? What tools or code did you need to use to prepare it for analysis? 3](#_Toc521251001)

[What challenges did you face? 4](#_Toc521251002)

[Analysis 4](#_Toc521251003)

[Conclusion 4](#_Toc521251004)

# Objective

## What are you setting out to prove or predict?

In order for an insurance company to make profit, it needs to collect more in yearly premiums than it spends on medical care to its beneficiaries. As a result, insurers invest a great deal of time and money to develop models that accurately forecast medical expenses.

The goal of this analysis is to use patient data to estimate the average medical care expenses for such population segments. These estimates could be used to create actuarial tables which set the price of yearly premiums higher or lower depending on the expected treatment costs.

## What is your rationale for there being a correlation in the data that you’re looking to confirm and/or exploit?

Medical expenses are difficult to estimate because the most costly conditions are rare and seemingly random. Still, some conditions are more prevalent for certain segments of the population. For instance, lung cancer is more likely among smokers than non-smokers, and heart disease may be more likely among the obese.

The price of medical expenses varies and depend on variety of factors. an insurance premium for a given insurance policy can vary and depends on a variety of factors. Among those factors are the type of insurance coverage, the likelihood of a claim being made, the area where the policyholder lives or operates a business, the behavior of the person or business being covered, and the amount of competition that the insurer faces. In general, the greater the risk associated with a policy, the more expensive the insurance policy will be.

# Data Preparation:

## What was your data source? How good was the data quality?

The dataset [“Medical Cost Personal Datasets”](https://www.kaggle.com/mirichoi0218/insurance) is a standard machine learning data set retrieved from Kaggle: an open platform for predictive modeling and analytics competition using real life datasets provided by companies and users. However, the original source of the dataset is from the book “Machine Learning with R” by Brett Lantz that introduces machine learning using R.

The overall quality of the data was very good as there were no missing values, unknown/inapplicable values. The full dataset consists of 1338 rows and 7 columns that represent information about person’s age, sex, BMI, number of children/dependents, smoking activity, region and medical charges. The data types vary from integer, float to object.

## What did you need to do to procure it? What tools or code did you need to use to prepare it for analysis?

The full dataset was procured by downloading the file (*insurance.csv*) [“Medical Cost Personal Datasets”](https://www.kaggle.com/mirichoi0218/insurance) from Kaggle as mentioned above. After importing the Pandas library into Jupyter Notebook, *pandas.read\_csv* method was used to load the full dataset into a DataFrame to proceed with the analysis. Additionally, other libraries such as Scikit-Learn, Matplotlib, Seaborn and the StatsModels module was imported for the analysis. Since the data did not have any missing unknown and inapplicable values there was not much cleaning to do when preparing for analysis.

## What challenges did you face?

Among 7 columns in the dataset, sex, smoker and region variables were categorical variables that are non-numerical. Since regression models and machine learning models are based on mathematical functions, having categorical data may not yield the best performance. Thus, categorical variables that are not quantifiable need to be converted to numerical predictors using the encoding process. There is a risk using the *LabelEncoder* class from Scikit-Learn library because the model could assume a natural ordering between categorical values and may result in poor performance. In this case, the *OneHotEncoder* class library can be applied where there is no relationship or order introduced between categories.

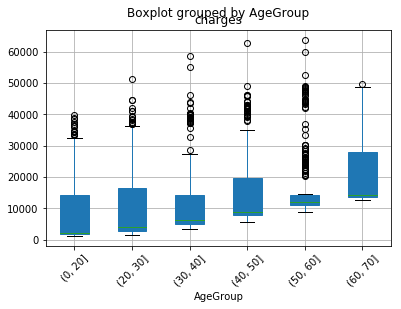
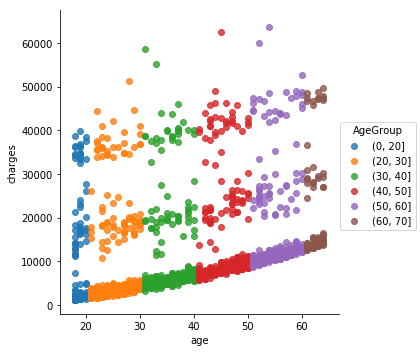
# Analysis

Multiple regression extends simple two-variable regression to the case that still has one response but many predictors (denoted ,, ...). The method is motivated by scenarios where many variables may be simultaneously connected to an output. The goal in this analysis is to use multiple regression to estimate the average medical care expenses, response variable, using 6 predictors in the dataset. The following are the columns in the dataset:

* age: age of primary beneficiary
* sex: gender of primary beneficiary (Female vs Male)
* bmi: Body Mass Index, a quick screening tool for assessing health risk based on your height and weight
* children: Number of children covered by health insurance / Number of dependents
* smoker: Smoking activity (Yes or No)
* region: the beneficiary's residential area in the US (northeast, southeast, southwest, northwest)
* charges: Individual medical expenses

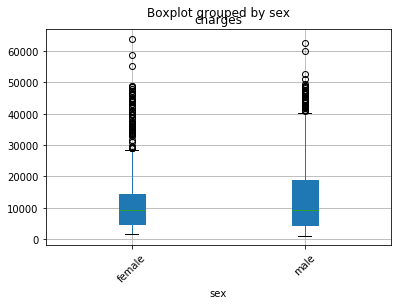
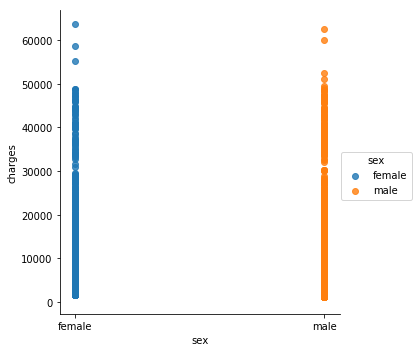
Let’s examine the relationship between each predictor variables to the response variable, *charges*.

**Age vs charges**

****

The first predictor variable is *age* and from the summary statistics, age ranges from 18 to 64. In order to identify any visible trend with the relationship between *age* and *charges*, *age* values were segmented and sorted into bins of 10 years. From the scatterplot on the left, there is no clear linear relationship between age and charges but a slight upward trend of charges increasing as age increasing was noticeable. From the box plot on the right, the median medical expenses increases as age increases.

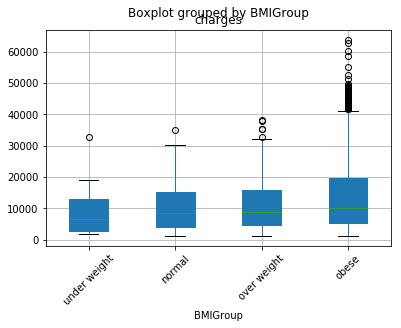
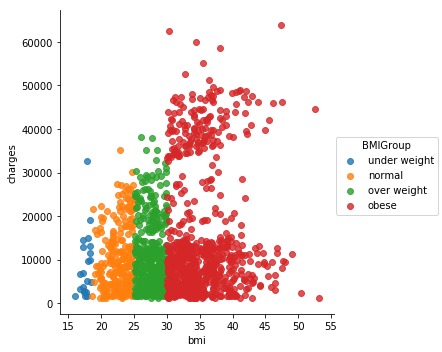
**Sex vs charges**

****

The sex variable is a categorical variable and represents whether the beneficiary is a female or male. However, as mentioned above categorical variables were encoded using the *OneHotEncoder* class and now the integer 0 is assigned to female and 1 is assigned to male. By looking at the scatterplot and the boxplot above, there seems to be no strong relationship between sex variable and charges. The median medical expenses seem very similar for female, although the interquartile range for male is greater than that of female.

**BMI vs. charges**

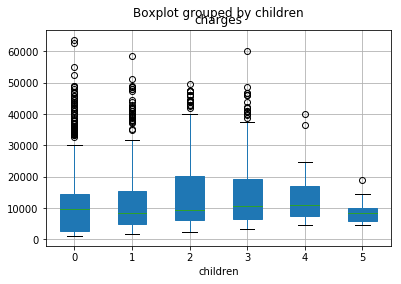
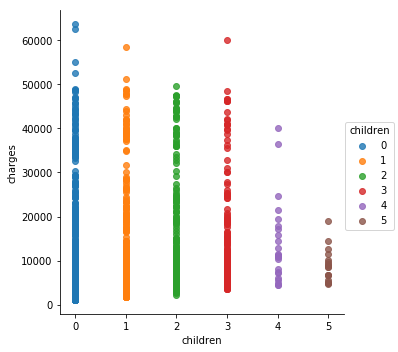
Body mass index (BMI) is one of many factors that can be used to assess one’s health. It is a quick screening tool for assessing health risk based on one’s height and weight.

****

If a person’s BMI is less than 18.5, he or she is underweight. Health risks associated with being underweight include osteoporosis, infertility, and impaired immune functioning. Underweight may also indicate an eating disorder or other underlying illness.If a person’s BMI is 18.5 – 24.9, it is a healthy range. This may lower the person’s risk for developing weight-related health problems.If a person’s BMI is 25 or higher, are at greater risk of developing diabetes, heart disease and some types of cancer. The higher your BMI number (over 25) is, the greater your risk of developing these conditions.

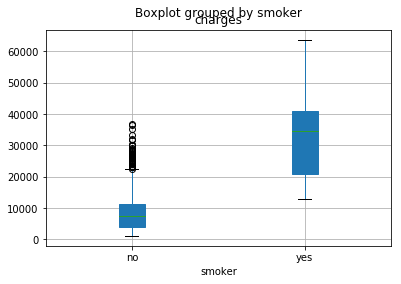
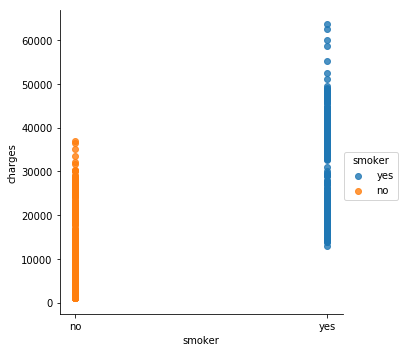
The mean medical expenses for people who are obese was USD 15,560.93 while those who are normal pay an average of USD 10,435.44. There seems to be a slight relationship between the two factors.

**Children vs. charges**



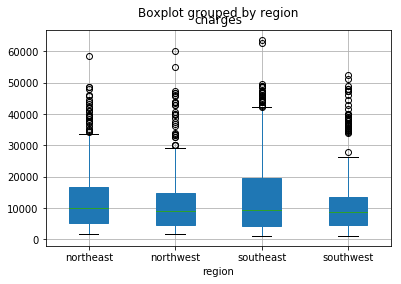
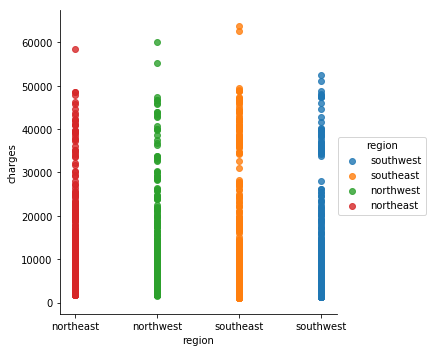
There is no strong relationship between number of children/dependents and medical charges. However, as the number of children/dependents increase, medical expenses seem to decrease, especially at 4 and 5 children/dependents.

**Smoker vs charges**

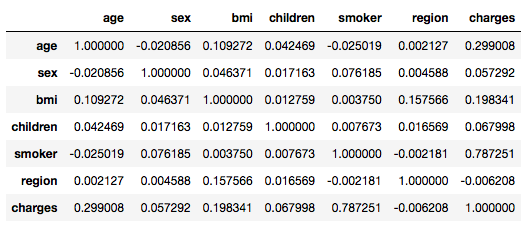
****

Smoker is a two-level categorical variable that takes value 1 when the person is a smoker and value 0 when the person is a non-smoker. The regression output indicates that the value is 0.62, there was an increase of 62% in the data’s variation by using information about smoking activity for predicting medical expenses using a linear model. The coefficient for the linear model for predicting medical expenses based on smoking activity is 23620. Thus 23620 means that the model predicts an extra $23,620 for those people who smoke versus those that don’t. Examining the regression output, the p-value for smoker is zero, indicating there is strong evidence that the coefficient is different from zero when using this simple one variable model.

**Region vs charges**

****

By looking at the plot on the left, there seems to be no strong relationship between region and charges.

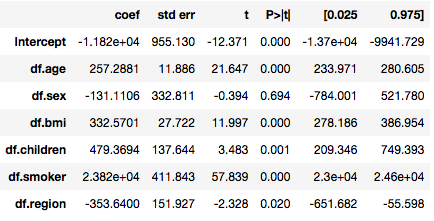


After analyzing the relationship between each factors and charges, smoker variable seemed to have the strongest correlation. By using Pandas df.corr() we can verify that the highest correlation value is 0.787 between smoker variable and charges variable followed by age variable 0.299 and bmi 0.198. Another observation is that no numeric values are highly correlated with each other, so multicollinearity isn’t an issue.

**Including and assessing many variables in a model**

We would like to fit a model that includes all potentially important variables simultaneously. This would help us evaluate the relationship between a predictor variable and the outcome while controlling for the potential influence of other variables.

Model 1



The first model includes all predictor variables in the dataset and got a decent r-squared of 0.7509 which implies that 75.09% of the variation of charges could be explained by the set of independent variables we have included. We could also observe that all of the independent variables we have included with the exception of gender is a statistically significant predictor of medical charges (p-value less than 0.05 <- level of significance). Sex variable can be eliminated.

Model 2

# Conclusion

Did you prove/disprove your hypothesis or create a useful model? What did you learn about your data set?