**predicting medical expenses using linear regression**

3251-014

CHAEWON KIM

AUGUST 2018

[Objective 3](#_Toc521359695)

[What are you setting out to prove or predict? 3](#_Toc521359696)

[What is your rationale for there being a correlation in the data that you’re looking to confirm and/or exploit? 3](#_Toc521359697)

[Data Preparation: 3](#_Toc521359698)

[What was your data source? How good was the data quality? 3](#_Toc521359699)

[What did you need to do to procure it? What tools or code did you need to use to prepare it for analysis? 3](#_Toc521359700)

[What challenges did you face? 3](#_Toc521359701)

[Analysis 4](#_Toc521359702)

[Conclusion 9](#_Toc521359703)

# Objective

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

Insurance companies invest a lot of time and money to develop models that accurately forecast people’s medical expenses because they make profit when they collect more in yearly premiums than they spend on medical care to their beneficiaries. The goal of this analysis is to use the dataset to understand the relationship between explanatory variables and the response variable and build a model that predicts medical care expenses for such population segments.

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

Although medical expenses are difficult to estimate, because certain conditions that are more costly are rare and seemingly random, 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 diseases may be more likely among the obese. Thus, the greater the risk associated with certain health conditions, the more expensive one’s medical expenses 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 charges (medical expenses). 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. After importing the Pandas library into Jupyter Notebook, the full dataset was loaded into a DataFrame using the *pandas.read\_csv* method. 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 needed to be converted to numerical values using the encoding process. There is a risk using the *LabelEncoder* class from Scikit-Learn library because the model could assume an ordinal relationship between categorical values and may result in less accurate performance. In this case, the *OneHotEncoder* class was applied where there is no relationship or order introduced between categories.

# Analysis

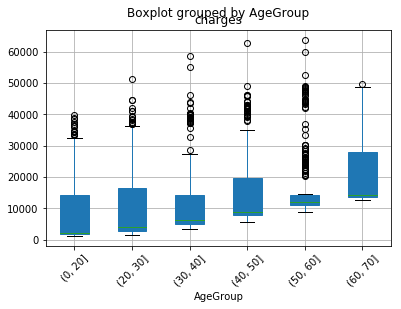
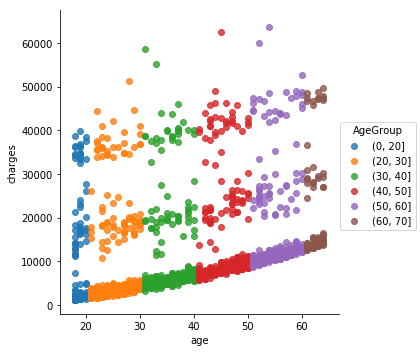
Multiple regression has one response variable but many predictors (denoted ,, ...) where they may be simultaneously connected to an output. As mentioned above, the goal of this analysis is to use multiple regression to estimate medical care expenses using 6 predictors in the dataset.

The following are the 7 columns in the dataset:

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

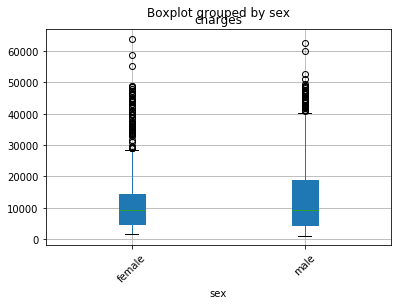
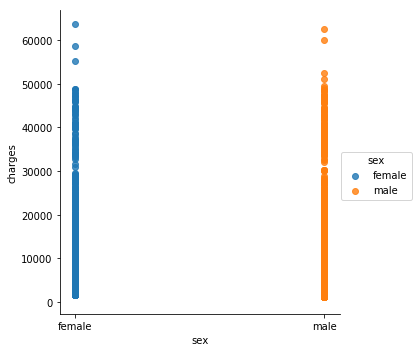
The following section will examine the relationship between each explanatory variables and the response variable to identify variables in the model that may not be helpful.

**Age vs charges**

****

The first explanatory variable to examine is the ***age*** variableand the values range from 18 to 64. ***age*** values were segmented into bins of 10 years to better identify any prevalent trends. The plot on the left indicates that although there is no clear linear relationship, as age increases, medical expenses increase as well. The minimum medical expenses for those in *AgeGroup* 60 to 70 is higher than that of those in *AgeGroup* 0 to 20. The boxplot on the right confirms these statements: the median medical expenses increase as age increases. Thus, ***age*** variable seems to have some relationship with ***charges***.

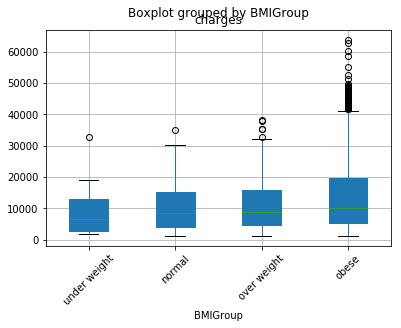
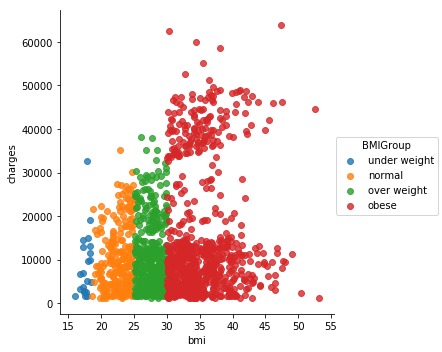
**Sex vs charges**

****

The ***sex*** variable is a two-level categorical variable indicating whether the person is a female or male. However, as mentioned above categorical variables were encoded using the *OneHotEncoder* class for better results (0=female, 1=male). By looking at the plot on the left, ***sex*** variable does not seem to influence medical expenses. Although the interquartile range is greater for male than that of female, the median medical expenses seem very similar for female and male. Thus, ***sex*** variable does not seem to be helpful when added into the model.

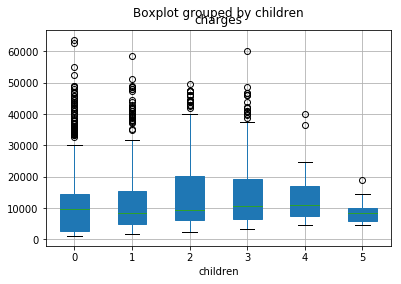
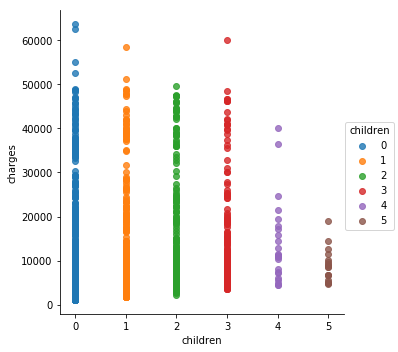
**BMI vs. charges**

Body mass index (BMI) is one of many factors that allows people to quickly assess one’s health risk based on one’s height and weight. Similar to ***age***values, ***bmi*** values were segmented into 4 groups representing whether a person is underweight (), normal over weight ( or obese to better identify any prevalent trends.

****

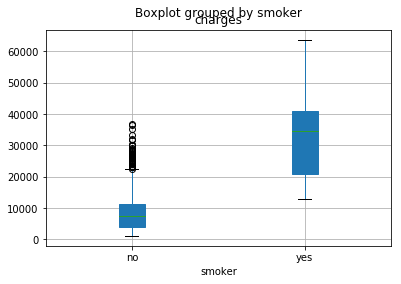
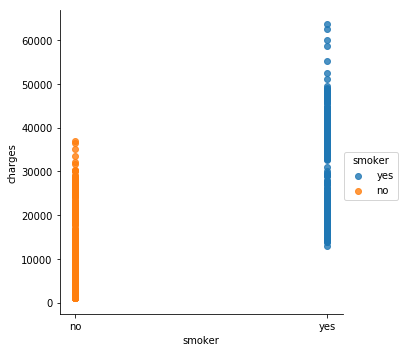
The plot on the left shows that majority of the people in the dataset were obese while the least number of people were underweight. Some of the health risks associated with being underweight include osteoporosis, infertility, and having eating disorder. The scatter plot as well as the box plot indicates that those who pay over $40,000 are the ones who are categorized as obese. Moreover, the statistics summary output showed that the mean medical expenses for obese people were $15,560.93 while for people in the normal group paid an average of $10,435.44. Obese people are at greater risk of developing conditions such as diabetes, heart disease and some types of cancer, which are also more costly to treat. Thus, ***bmi***variable seems to have some relationship with the ***charges***.

**Children vs. charges**



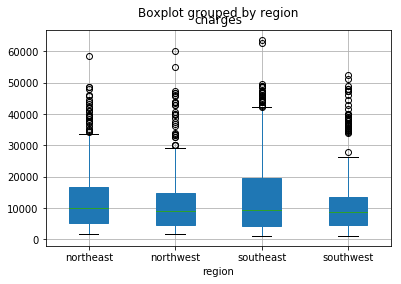
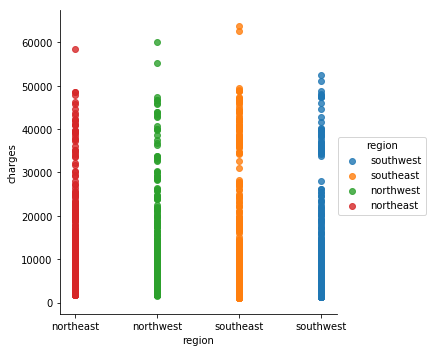
The **children** variable indicates how many children or number of dependents the person’s health insurance covers. The plot on the left shows that medical expenses are quite similar for the ***children*** value from 0 to 3. However, medical expenses decrease for those who have 4 and 5 children because according to Blue Cross, when a person has more than 3 children under the age of 21, he or she only pays health plan for the three oldest when all members are covered. The ***children*** variable does not seem to be a very important explanatory variable.

**Smoker vs charges**

****

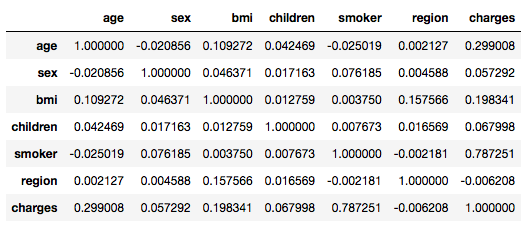
The ***smoker*** variable is a two-level categorical variable that indicates whether or not the person smokes. Similar to the ***sex*** variable, the values were encoded by using the *OneHotEncoder* class (0=non smoker, 1=smoker). The plot on the left suggests that medical expenses for smokers are generally higher than non-smokers. Moreover, the boxplot shows that median medical expenses for smokers are around 4 times higher than non-smokers. Since ***smoker*** variable seemed to have a strong relationship with ***charges*** the single-variable model was created, and the regression output indicates that the value was 0.62. The coefficient for the linear model predicting medical expenses based on smoking activity was 23620, which means that the model predicts smokers to spend an extra $23,620 compared to those who don’t smoke. The regression output shows that the p-value for ***smoker*** variable is 0, indicating there is strong evidence that ***smoker*** variable is related to **charges**.

**Region vs charges**

****

The last variable to examine is the ***region*** variable. The plot on the left shows that there seems to be no significant difference in medical expenses among 4 regions and by looking at the boxplot, the minimum as well as the median medical expenses values are similar among 4 regions indicating that ***region***variable seems to be not helpful when added in the model.

**Correlation between variables**

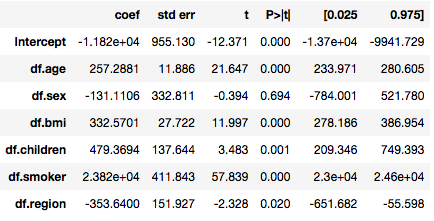


After examining the relationship between each explanatory variables and the ***charges***, ***smoker*** variable seemed to have the strongest correlation. This can be verified by using Pandas *df.corr()* method: ***smoker*** vs ***charges*** (0.787), followed by ***age*** vs ***charges*** (0.299) and ***bmi*** (0.198). Another thing to note is that predictor variables are not correlated among each other, so collinearity will not be an issue in this analysis.

**Building a model**

The goal is to fit a model that includes all potentially important variables simultaneously. This is helpful in evaluating the relationship between a predictor variable and the response variable while controlling for the potential influence of other variables.

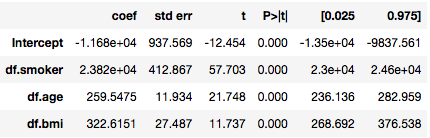
Model 1: Full model



First model is a full model that includes all explanatory variables with the adjusted R-squared value of 0.750. This means that 75.0% of the variation of charges could be explained by the set of explanatory variables that are included in the model. The p-value for the ***sex*** variable is greater than 0.05, which confirms the assumption earlier that the ***sex*** variable does not seem to be a significant variable predicting medical expenses. Among many coefficients from the regression output, the coefficient for ***smoker*** variable is worth mentioning because it means that those who smoke are likely to spend $23,823 more than those who don't smoke. This solidifies the finding from earlier that ***smoker*** variable is strongly correlated with ***charges***.

Model 2: Age & Smoker& Bmi

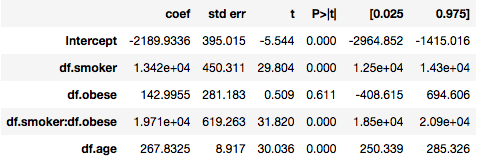
The second model will include explanatory variables that We identified earlier that smoker and age variables were the most



Since we observed from the scatter plot above that a good portion of obese people's charges are much higher than the rest of the group, we created a new categorical variable 'obese' to indicate whether the person is obese or not and then used OneHotEncoder to convert the values into numerical values. Then we added this variable to the model to see its correlation with charges.

We created a new variable obese to give a high penalty to those who smoke and are obese. Obesity increases health care costs by USD865, but if the person smokes and is obese, it can be expected that medical expenses will increase by USD 19,329. By adding the derived variable, we improved our model. We now have an adjusted R-squared of 0.762 which implies 76.2% of variation of charges can be explained by our independent variables in the model.

Model 3



From here, we will use the forward selection by adding variables one-at-a-time until we cannot find any variables that are likely to improve the accuracy in predicting future medical expenses. (as measured by adjusted R2).

Since *age* was another predictor variable that had a correlation with *charges*, we’ve added it back to our model to see if it improves our model's ability to predict charges. As a result, our model's adjusted R-squared value improved significantly to 0.858.

# Conclusion

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