**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 and 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*) from Kaggle mentioned above. and *pandas.read\_csv* method was used to load the full dataset into a DataFrame in Jupyter Notebook for the analysis.

## What challenges did you face?

# Analysis

If you’re conducting an inference test explain the analysis you performed clearly and include well-labelled diagrams to make your points. How did you confirm that the data met the requirements for the test or modeling technique to be valid?

# Conclusion

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