# Linear Regression

## Problem Statement

We will use health insurance dataset to predict insurance charges. The dataset has total six features to predict charges which are age, sex, BMI, children, smoker and region.

|  |  |
| --- | --- |
| **Variables** | **Definition** |
| Age | In years |
| Sex | Sex of the person |
| BMI | Body mass index that is calculated from weight and height of a person |
| Children | Number of children that the person has |
| Smoker | Whether the person is smoker or non-smoker |
| Region | The region that the person lives |
| Charges | Charges for health insurance of a year |

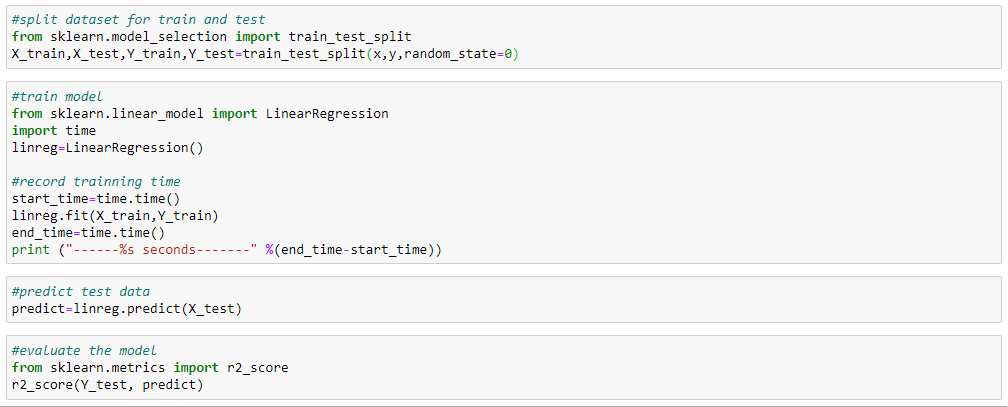
## Source of Dataset

<https://www.kaggle.com/mirichoi0218/insurance>

## Linear Regression Model for Insurance Dataset







Duration - 0.001960277557373047 seconds

## Sharing of the process

We will use insurance dataset which has 1338 rows and 7 columns in which 3 columns are string datatype, 2 columns are integer datatype and another 2 columns are decimal datatype. Firstly, we made categorical data encoding for string datatype columns, sex, smoker and region.

Data Mapping and Data Cleaning

* **Sex** - Male = 0, Female = 1
* **Smoker** – Non-smoker = 0, Smoker = 1
* **Region** – Northeast=0, Northwest=1, Southeast=2, Southwest=3

Then the dataset has no null values and unreasonable data. Therefore, we do not need data cleaning.

Feature Engineering

For feature engineering, we have tried feature selection method on the dataset. But it does not have highly correlated features with dependent variable to select and there is no high correlation between features so no independent variables to remove. After that, we also tried with Principal Component Analysis (PCA) technique in order to reduce dimensions for the dataset. We set component to six. In this case, although the sum of all explained variance ratios is 1.0, the accuracy score is 75%. Besides, we also trained the model without feature engineering whether the score is better or not, then the accuracy score we got is 79.6%.

Therefore, we decided to train the model without feature engineering because the accuracy of the model is the best and the total existing number of features in the dataset is not too many.