Name: Mathieu Bernardi Cohort: DSTI Autumn 2018

Insurance dataset

Introduction:

To make their own profits, the insurance company(insurer) must collect more premiums than the amount paid to the insured person.

Try to create a model to predict health care cost?

Understanding the data:

Age: insurance contractor age, years

Sex: insurance contractor gender, [female, male]

BMI: Body mass index, providing an understanding of body, weights that are relatively high or low relative to height, objective index of body weight (kg / m ^ 2) using the ratio of height to weight, ideally 18.5 to 24.9

Children: number of children covered by health insurance / Number of dependents

Smoker: smoking, [yes, no]

Region: the beneficiary residential area in the US, [northeast, southeast, southwest, northwest]

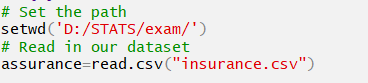
Charges: Individual medical costs billed by health insurance,

Load packages and dataset:

A screenshot of a cell phone

Description automatically generated

Set the correct path and load the data:



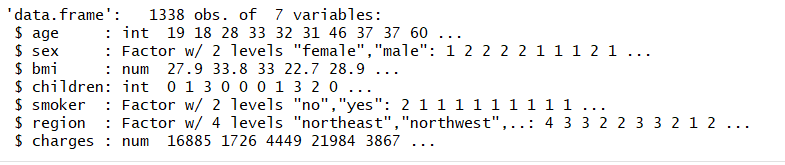
A picture containing person

Description automatically generated

A screenshot of a cell phone

Description automatically generated

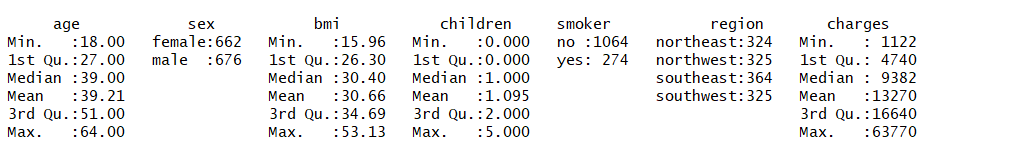




We are working with a dataset with 1338 observations and 7 variables.

A picture containing indoor

Description automatically generated



There are no missing data. We would try to predict the charges variable.

A picture containing person

Description automatically generated

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

A picture containing screenshot

Description automatically generated

Based from above plot, we can disclose that region of origin doesn’t have much impact with the amount of medical cost.



A screenshot of a social media post

Description automatically generated

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

As we can see that the smokers spend a lot more in terms of medical expenses compared to non-smokers by almost 4x.

A close up of a logo

Description automatically generated

A picture containing wall

Description automatically generated

We can see that smoker and age has the highest correlation with charges amongst our numeric variable.

A screenshot of a cell phone

Description automatically generated

We could see a trend here, as people get older the higher their medical expenses would be expected. But, regardless of age, smokers has higher medical expenses than non-smokers. It really seems like smoker is the single most important variable in predicting medical charges.

Model with all variables:

A close up of a mans face

Description automatically generated

A screenshot of a cell phone

Description automatically generated

We utilize all the variables on this model, and the r-squared of 0.7509 is good, which implies that 75.09% of the variation of charges could be explained by the set of independent variables. 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. The P-value of the model is very small 2.2 e-16 that we reject H0.

Model with age, smokeryes and bmi:

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

Using just age, smoker and bmi as independent variables, we have built a model with an r-squared of 74.75% which is comparable with our first model which use all of the original variables. In regression analysis, we would want to create a model that is accurate but at the same time as simple as possible. So if I would have to choose, I would select this second model than the first one. I will keep this model to predict the charges of a new person.

Create training set and test set:

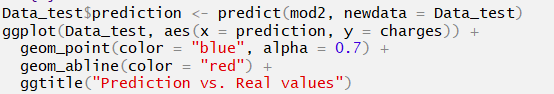
A picture containing indoor

Description automatically generated

A screen shot of a social media post

Description automatically generated

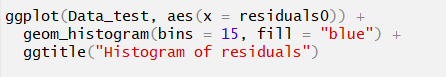




A picture containing indoor, showing, photo

Description automatically generated

Histogram of residuals:

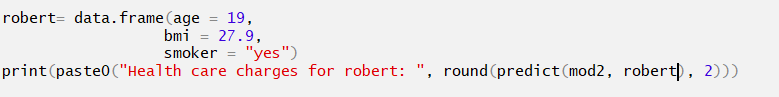


A screenshot of a social media post

Description automatically generated

The residuals follow a normal law, this second model is acceptable to make some predictions.

Predictions of charges on new people:





A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated