Insurance charge estimation based on regression analysis

Dake Li[1] and Likun Lin[2]

[1]Department of Mathematics, Fudan University
[2]Department of Mathematics, City University of Hong Kong

November 17, 2023

Abstract

In this article, we selected a set of data from Kaggle website about medical insurance charge, provided with BMI, age, whether smoking or not. By analysing this dataset, we aimed to predict the corresponding insurance charges for new BMI-age-smoking data. We mainly set up a linear regression model. However, in this process, we found that there was a significant difference between the data respect to smokers and non-smokers in the sample. Therefore, we classified the smokers and non-smokers into two categories and established prediction models separately.

1 Motivation and Background

The role of medical insurance in society is multifaceted. In terms of individuals, we want to build up a suitable expectation of the expenses on ourselves' medical insurance. Therefore, given personal information, a prediction model of the insurance charge is very useful. As taught in this course, linear regression model is a very powerful tool for people to estimate the relation between two variables. Since every function has a best linear approximation with its first-order derivatives around a fixed point, linear regression is especially useful to model some relations locally. In this article, we mainly built up a linear regression model to predict one's medical insurance charge in the future, based on his or her BMI, age, and whether smoking or not.

2 Objectives

The main objective of this article is to set up a prediction model of medical insurance charges, based on one's BMI, age, and whether smoking.

3 Data Collections and Data Prepossessing

3.1 Data Collection

We fetch the dataset from the website Kaggle (Data source: https://www.kaggle.com/datasets/mirichoi0218/insurance). The origin data has too many variables, but we only take BMI, age, whether smoking into consideration(See appendix). This allows us to get rid of disturbance and have a deeper insights into the most primary aspects.

3.2 Data Prepossessing

The following remains to edit
Firstly, It is observed that the data label of 'smoker' and 'non-smoker' led to a significant gap in the insurance charge, just as figure 1 and figure 2 show.

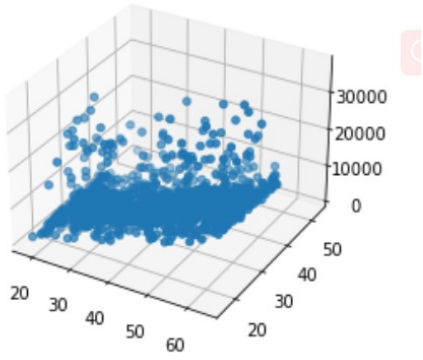

Figure 1: Nonsmoker's condition

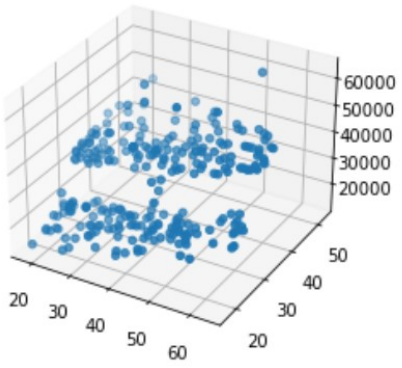

Figure 2: Smoker's condition

According to this observation, we are interested in whether it is reasonable to assume that the data in two labels are under different distributions. To examine this, we perform a hypothesis test on the mean of the two sets.

4 Clustering and Classification

4.1 Smokers' condition

By observation of the BMI-charge(figure 3) and the age-charge(figure 4) figure, it is possible to separate the data from both graphs into two parts.

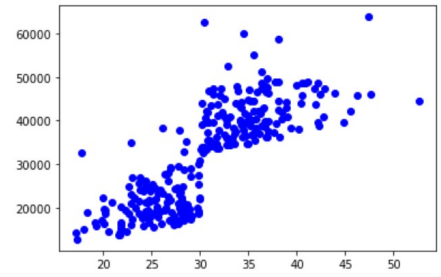

Figure 3: BMI-charges figure

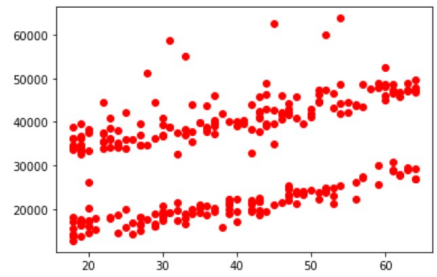

Figure 4: Age-charge figure

To do a more rigorous parting, Gaussian Mixture Model(GMM) is introduced to achieve this. For both graphs, the GMM model is set to do a two-cluster separation and achieves satisfying results as follows. As can be observed, even overlaying the two clustering together can almost separate the data points into two categories. One layer has obviously higher charges(over about 30,000) and higher BMI(over about 30) and the other's charges and BMI are lower.

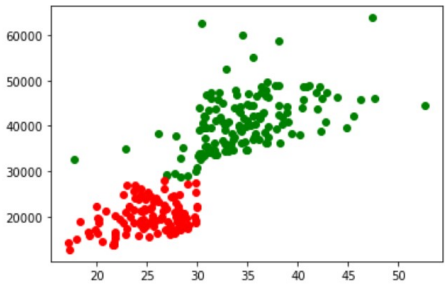

Figure 5: clustering result for bmi-charge

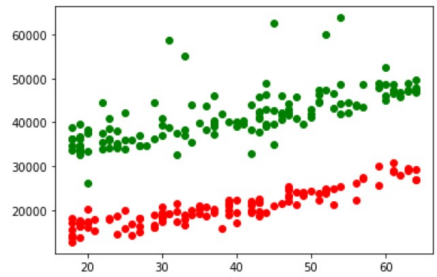

Figure 6: clustering result for age-charge

According to the above analysis, the smoking population data can be divided into two categories: high charge and low charge.

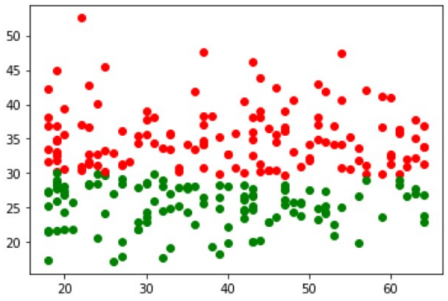

Figure 7: SVM Classification result

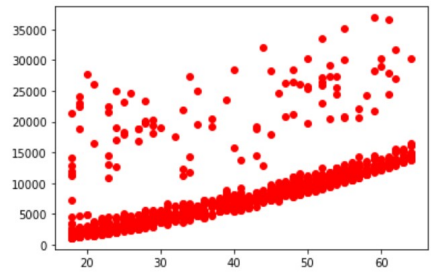

Figure 8: age-charge figure for non-smoker condition

4.2 Non-smokers' condition

By observation of the age-charges figure, The data are clustered densely in the low-charge area and scattered in the high-charge area. The GMM model is again used to

cluster the data into two sets, whose result is in figure 9. The above clustering results projected onto age-bmi were shown in Figure 10, which is apparently terrible. Thus, the speculation is that we do not acquire enough data to perform a good classification among these data points. It may be caused by the lacking of customers with charges over 30000 but we do have not enough data to show. Consequently, what to be considered is merely cleaning the data by dropping out outliers and do not apply classification to the 'non-smoker' category afterward.

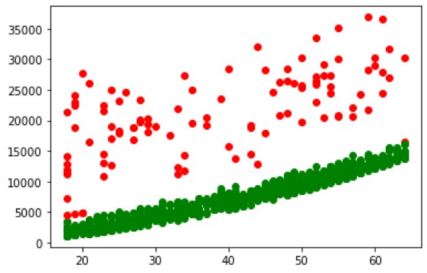

Figure 9: GMM result under nonsmoker condition

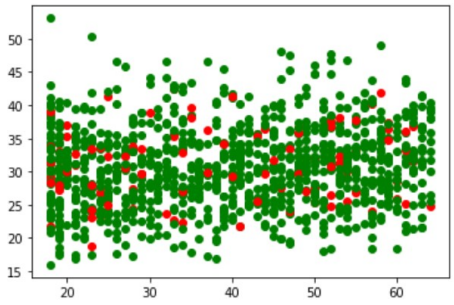

Figure 10: GMM cluster result on age bmi of non-smoker condition

In summary, currently, there are three categories, which are the nonsmoker category, the higher-charge smoker category, and the lower-charge smoker category.

5 Linear Regression model

5.1 Simple Linear Regression with non-smoker category

Consequently, Simple linear regression between age and charges is applied. For input feature age (represented as X), the estimator for charges (represented as $\widehat{y}(x)$) can be calculated as:

$$\widehat{y}(x) = \widehat{\mu}_{Y|X=x} = \widehat{\beta}_0 + \widehat{\beta}_1 X$$

, The simple linear regression model from the R code is $Y = -3298.954 + 267.753x$, and R^2 is 0.8536, which is shown in figure 13. Meanwhile, based on

$$\widehat{y}(x) \sim N\left(\mu_{Y|X=x}, \sigma^2\left(\frac{1}{n} + \frac{|x-\bar{x}|^2}{s_{xx}}\right)\right)$$

$$Y \mid (X=x) - \widehat{y}(x) \sim N\left(0, \sigma^2\left(1 + \frac{1}{n} + \frac{|x-\bar{x}|^2}{s_{xx}}\right)\right)$$

With the unbiased estimator s for σ, the 0.95 confidence interval and the prediction interval are calculated and shown in figure 14.

5.2 Multiple Linear Regression with higher-charges smokers category

Firstly, the training set is also cleaned by means of controlling the standardized residual between -2 and 2. Then, multiple linear regression is applied with the input features

such as BMI and age, whose result is shown in figure 15. From the result, it is obvious that charges have almost no linear relationship with BMI, which shares the same conclusion as the Added-variable plot in figure 16.

For input feature age (represented as X_1), the estimator for charges (represented as $\widehat{y}(x)$) can be calculated as:

$$\widehat{y}(x) = \widehat{\mu}_{Y|X=\vec{x}} = \widehat{\beta}_0 + \widehat{\beta}_1 X_1 + \widehat{\beta}_2 X_2$$

The simple linear regression model from the R code is $Y = 13838.395 + 274.530 * X_1 + 463.828 * X_2$, and R^2 is 0.9027, which is shown in figure 15. The result is satisfying.

5.3 Multiple Linear Regression with lower-charges smokers category

Firstly, the training set is also cleaned by means of controlling the standardized residual between -2 and 2. Then, multiple linear regression is applied with the input features such as BMI and age, whose result is shown in figure 17. From the result, it is obvious that charges have almost no linear relationship with BMI, which shares the same conclusion as the Added-variable plot in figure 18.

The simple linear regression model from the R code is $Y = 67.902 + 258.458 * X_1 + 423.518 * X_2$, and R^2 is 0.9325, giving a desirable result which is shown in figure 15.

6 Summary of techniques used that is not taught in class

Gaussian Mixture Model is used to cluster the data and the Support Vector Machine is used to classify the data.

7 Conclusion

This paper focused on the relationship between 'charge' and its independent variables 'smoker', 'BMI', and 'age'. Our discussion is based on the fact that different 'smoker' labels would lead to a significant difference in the value of charges. Further experiments proved that 'smokers' can be parted into different classes for their charges with respect to 'age' and 'BMI', while 'non-smokers' does not produce such a separating effect so only one class is adopted in the result. Generally speaking, the charge of the above three classes all possess a strong positive linear relationship with age. The result shows that our model may produce more errors when predicting the non-smoker category. It is suggested that relevant data, e.g. family status, is lacking in the present model, and thus further research can be conducted with aims of observing the relation between family and charge.

8 Appendix

Data source:
https://www.kaggle.com/datasets/mirichoi0218/insurance

.csv version:
https://drive.google.com/file/d/1_VQk7_KoigrBS3NEPzBX5rdbI6CE4w43/view?usp=sharing
Python Prepossessing code:
https://drive.google.com/file/d/1VziKPOPhUNzKtoHm06vYHopCrpNNXNq-/view?usp=sharing