Insurance charge estimation based on regression analysis

Hanzheng Qiu[2], Likun Lin[2], Tianxing Sun[2], Xiaoyang Li[1], and Zhenda Shen[2]

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

November 17, 2023

Abstract

A dataset from Kaggle is collected on the relationship between individual medical expenditure and several mutually independent factors. Based on this dataset, we analyzed the relationship between factors, such as age, BMI, and healthcare expenditures. We further developed a regression model based on simple linear regression approach. A simple linear regression is conducted on the nonsmoker category which achieves an accuracy of 0.9 on the test set to be in the prediction interval. The smoker category is separated into the high-charge group and low-charge group. Multiple linear regression is conducted on both groups. The R^2 scores are both over 0.9. Finally, it is concluded that the health expenditure of all populations is linearly related to age, and the effect of BMI on health expenditure varies with or without smoking.

1 Motivation and Background

Health and medical care are one of the most concerned topics in the modern society. Proper health care coverage can contribute to social stability, development and well being. For the authorities worldwide, within a limited budget, it is imperative for relevant departments to arrange medical insurance expenditures accurately and reasonably. Therefore, the analysis and management of public and personal health care expenditure is a crucial topic of administration.

2 Objectives

The main objective of this experiment is to assess whether several given factors have a linear or other relationship with medical expenditure under our data set using statistical and linear regression approaches and, if exist, give a reasonable and predictive linear regression model.

3 Data Collections and Data Prepossessing

3.1 Data Collection

The dataset is collected on Kaggle, where the dependent variable of this dataset is personal medical expenditure(represented as charges in the following reports) and the independent variables are six factors: age, sex, BMI, number of children, smoking or not, and residential area in the U.S. Among them, 'Smoking', and 'residential area' are categorical variables, and 'no. of children' is considered as small discrete integer variables. As a consequence, 'age' and 'BMI' are considered useful for our analysis. These data are considered quite complete and clean at acquisition.

3.2 Data Prepossessing

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.

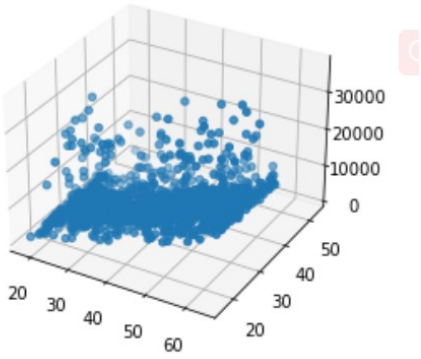

Figure 1: Nonsmoker's condition

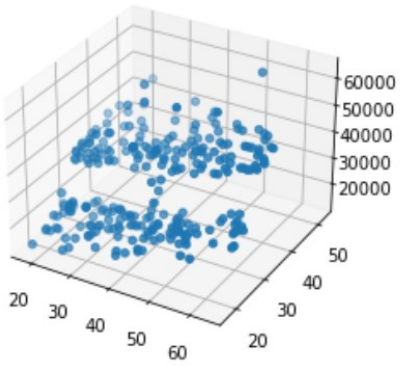

Figure 2: Smoker's condition

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. Since both population mean and variance are unknown, we should perform the following test:

$$H_0 : \mu_1 = \mu_2, H_1 : \mu_1 \neq \mu_2,$$

where $\bar{x_1}, s_2$ represents the sample(our data set) mean and variance charges for the smoker and $\bar{x_2}, s_2$ represents the sample mean and variance charges for the nonsmoker. Similarly, μ_1, σ_1 represents the population means and variance charges for the smoker and μ_2, σ_2 represents the sample mean and variance charges for the nonsmoker. Taking arbitrary sample from each of the two groups, the the random variable $\bar{X}_1 - \bar{X}_2$ of the difference between their averages satisfies the following model

$$\frac{\left(\bar{X}_1 - \bar{X}_2\right) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \overset{a}{\sim} t_\nu, \nu = \frac{\left(s_1^2/n_1 + s_2^2/n_2\right)^2}{\frac{\left(s_1^2/n_1\right)^2}{n_1 - 1} + \frac{\left(s_2^2/n_2\right)^2}{n_2 - 1}}$$

From R code, the p-value $< 2.2 \times 10^{-16}$, which leads to the rejection of H_0 and we need to classify the data into smoker and nonsmoker for analysis. Separately 20 data from

both categories are preserved to be test set and all remaining data form the training set.

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.

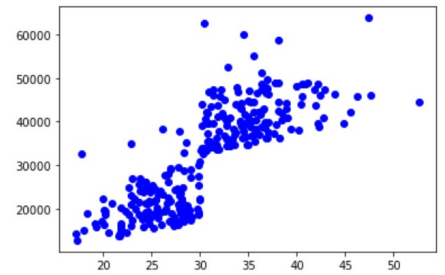

Figure 3: BMI-charges figure

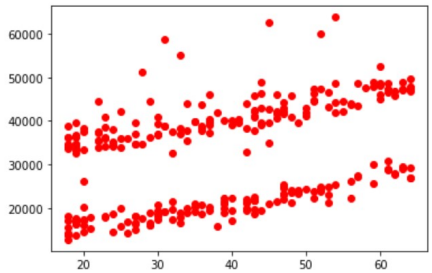

Figure 4: Age-charge figure

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.

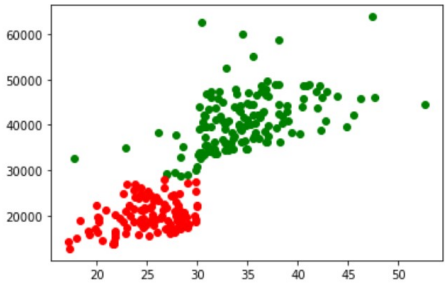

Figure 5: clustering result for bmi-charge

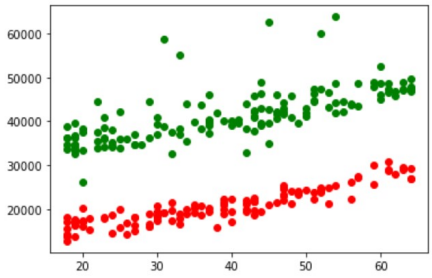

Figure 6: clustering result for age-charge

According to the above analysis, the smoking population data can be divided into two categories: high charge and low charge. However, given that the goal is to predict costs based on BMI and age, a cost-independent classification method is needed. Therefore, the Support Vector Machine (SVM) was introduced to divide the data into two categories: high-charge and low-charge. According to the GMM labels, the results of age bmi applied by the support vector machine are shown in Figure 7 and simple linear regression will be applied to these two distinct categories separately.

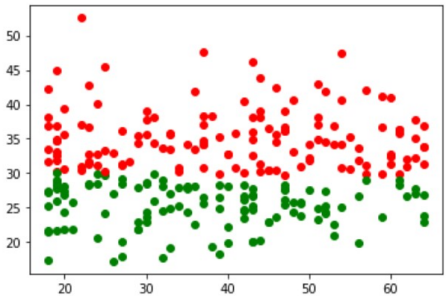

Figure 7: SVM Classification result

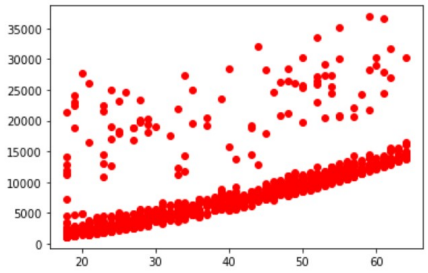

Figure 8: age-charge figure for non-smoker condition

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.

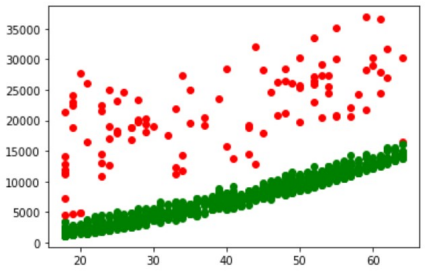

Figure 9: GMM result under nonsmoker condition

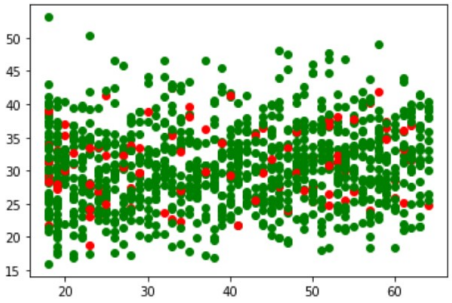

Figure 10: GMM cluster result on age bmi of non-smoker condition

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

Firstly, the training set is 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 11. From the result, it is obvious that

charges have almost no linear relationship with BMI, which shares the same conclusion as the ANOVA analysis in figure 12.

```
Call:
lm(formula = charge ~ age + bmi, data = smoker)

Residuals:
    Min      1Q  Median      3Q     Max
-1999.1  -789.4  -243.4   349.9 10358.5

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -3280.5612   278.5200 -11.779   <2e-16 ***
age           267.7870     3.5925  74.540   <2e-16 ***
bmi            -0.6447     8.2407  -0.078    0.938
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1559 on 967 degrees of freedom
Multiple R-squared:  0.8536,    Adjusted R-squared:  0.8533
F-statistic:  2819 on 2 and 967 DF,  p-value: < 2.2e-16
```

Figure 11: Nonsmoker Multiple linear regression result

```
Analysis of Variance Table

Model 1: charge ~ age
Model 2: charge ~ age + bmi
  Res.Df        RSS Df Sum of Sq      F Pr(>F)
1    968 2351225674
2    967 2351210790  1     14883 0.0061 0.9377
```

Figure 12: Nonsmoker ANOVA analysis result

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. The red dashed lines are the boundaries for the prediction interval and the grey band around the blue line is the confidence interval, which is very small. The proportion for the test set to be in the prediction interval is 0.9, which is satisfying.

```
Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-1995.5  -792.0  -237.4   349.6 10363.4

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -3298.954    149.285  -22.10   <2e-16 ***
x             267.753      3.564   75.12   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1559 on 968 degrees of freedom
Multiple R-squared:  0.8536,    Adjusted R-squared:  0.8534
F-statistic:  5643 on 1 and 968 DF,  p-value: < 2.2e-16
```

Figure 13: Nonsmoker Multiple linear regression result

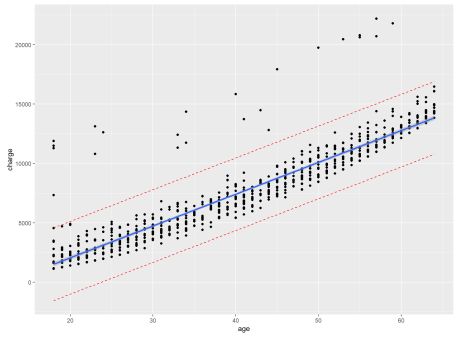

Figure 14: Nonsmoker ANOVA analysis result

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.

```
Call:
lm(formula = charge ~ age + bmi, data = nonsmokerII)

Residuals:
    Min      1Q  Median      3Q     Max
-1966.7  -840.5  -232.1   369.7  7067.1

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 13838.395   1174.748   11.78   <2e-16 ***
age           274.530      9.015   30.45   <2e-16 ***
bmi           463.828     31.144   14.89   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1459 on 122 degrees of freedom
Multiple R-squared:  0.9027,    Adjusted R-squared:  0.9011
F-statistic: 565.9 on 2 and 122 DF,  p-value: < 2.2e-16
```

Figure 15: Higher-charges smoker Multiple linear regression result

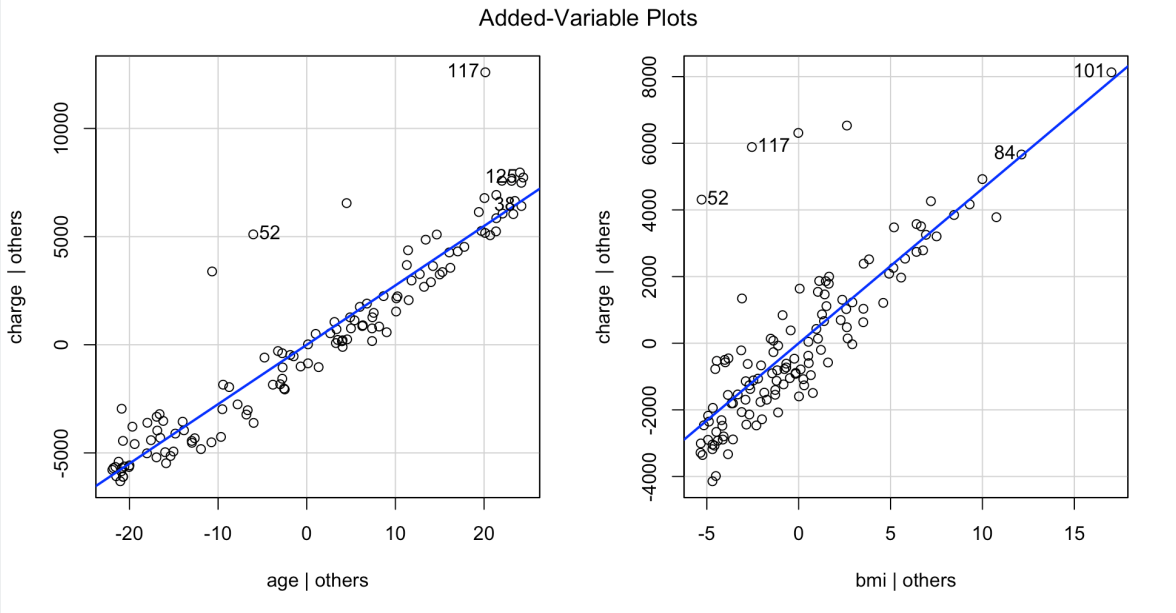

Figure 16: Higher-charge smoker Added variable Plot result

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.

```
Call:
lm(formula = charge ~ age + bmi, data = nonsmokerI)

Residuals:
    Min      1Q  Median      3Q     Max
-2242.5  -684.8    23.9   529.2  5697.6

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   67.902    828.296   0.082    0.935
age          258.458      7.184  35.979   <2e-16 ***
bmi          423.518     30.910  13.702   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1000 on 107 degrees of freedom
Multiple R-squared:  0.9325,    Adjusted R-squared:  0.9313
F-statistic: 739.4 on 2 and 107 DF,  p-value: < 2.2e-16
```

Figure 17: Lower-charges smoker Multiple linear regression result

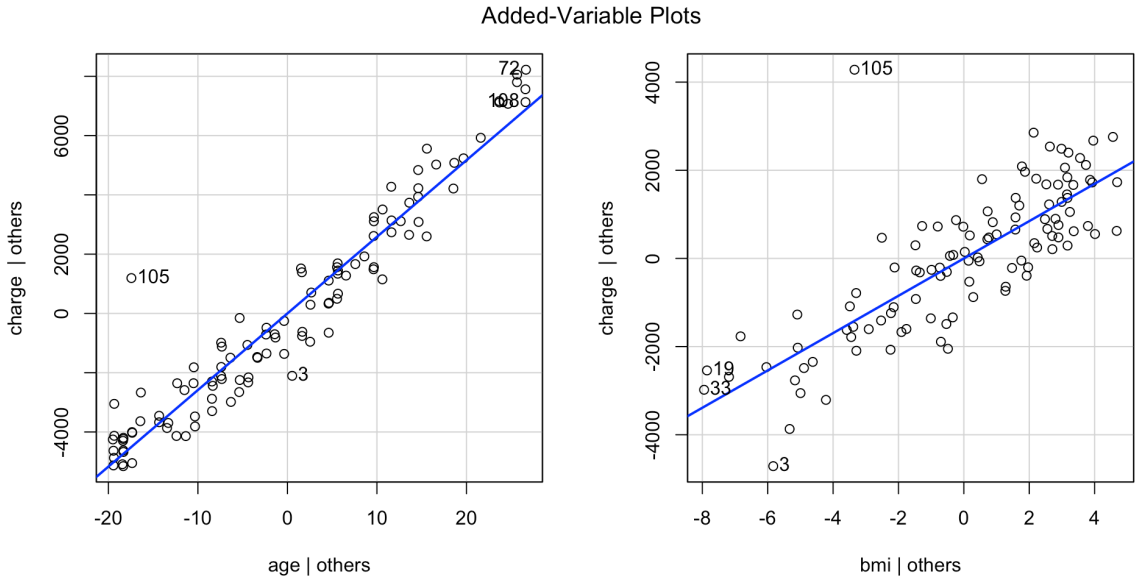

Figure 18: Lower-charge smoker Added variable Plot result

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
Our source code of R:
https://github.com/AharenDaisuki/insurance_charges_estimation
Python Prepossessing code:
https://drive.google.com/file/d/1VziKPOPhUNzKtoHm06vYHopCrpNNXNq-/view?usp=sharing