Prediction Of Medical Insurance Cost Through Linear Regression Model

Chengwei Cao [*]

Faculty of Natural, Mathematical and Engineering Sciences, Kings College London, London, E1 7BB, United Kingdom

* Corresponding author: K21018767@kcl.ac.uk

Abstract. Nowadays health care problem has become an increasingly important topic, which means that more and more people choose to pay for health care insurance. This article is mainly about constructing a model which is used to predict a person's health care insurance cost as a reason that it is usually hard to find the total amount of cost from people on health care insurance area. The author found a data set about health care insurance cost and used r to analyze it. In the end, 2 linear regression models which seem to fit the data set are built to predict the cost of heal care insurance. After the comparison of these 2 models, a better one is chosen. The model has also been checked to see if it is accurate and proper. The particular significance of this study is that it builds a model which can predict hard-to-collect data from some data which is easy to collect like BMI and age.

Keywords: Linear regression model, health care insurance cost, model comparison, model accuracy check.

1. Introduction

1.1. Background

At present, medical insurance is in a stage of a continuously rising market, which means that people tend to pay more and more attention to their health. The healthcare industry is one of the vital service providers which improve people's lives. With the increase in health care service cost, the only useful way to get quality care when an accident or critical illness happen is to depend on the person's health insurance [1]. Relatively speaking, people increase their costs of health care insurance. People are more conscious of buying medical insurance, because, in case of an accident, medical insurance can reimburse most of the medical expenses. Health insurance in the United States reported a 4.6 percent increase in 2018, compared to 4.2 percent in 2017 [2]. Hence an increasing trend can be found in the total amount of health care insurance cost for most people. An important reason to make this regression model is that it is usually hard to find out the total amount people spent on insurance. In this case, the model is needed because it only uses some data which is easy to collect.

1.2. Related research

At present, the linear regression model is already a relatively mature statistical method. An introduction to statistical learning is well exploited in this research, including some basic formulas about linear regression and how to apply linear regression in some data sets [3]. Linear regression is the cornerstone of many modern modeling tools [4]. Linear regression always provides an ideal approximation to the regression function and one particular case is that the quantity of samples is limited or the signal is comparatively weak [5]. The accuracy of the linear regression model always has to be guaranteed so it is important to use some method to make sure the regression model fits the data [6]. One another important things are that the research needs to include data visualization and the trend diagram of the model so that the characteristics of the linear regression model can be more prominently highlighted [7].

1.3. Objection

In summary, it is especially necessary for some insurance companies to predict a person's health insurance costs. In contrast, the linear analysis regression method is one of the important research methods for analyzing related factors. This research aims to create a simple and accurate linear regression model to fit the data as closely as possible.

2. Methodology

2.1. Source of data

This project uses the data set from Kaggle total of 1,338 pieces of data and 8 variables e.g., age and BMI. One of the columns is irrelevant which is the index of each interviewee. Each row provides each person’s related information. In this dataset, the dependent variables are region, age, BMI, number of children, smoking or not, and gender. The only independent variable is the charge of the health care insurance cost. The next step is to analyze the data set and build a linear regression model.

2.2. Models

Since it is known that the linear regression model is suitable for this data set, the next step is to plug the data into the model as

$$Y =\beta0 +\beta1X1 +\beta2X2 +\cdots+\beta pXp +\varepsilon \tag{1}$$

There are a total of 7 variables in the data set and 1 of them is charges which means that it can be seen as Y in the formula. The rest of the 6 variables can be seen as β1, β2, …, βn. Then R is used as a programming language to build this linear regression model.

2.3. Research Steps

First of all, a single-factor data analysis is conducted to explore the independent impact of each independent variable on the dependent variable.

The second step is that a univariate analysis is performed with the R language to explore the correlation between each independent variable and the dependent variable. The next step is to use the formula to construct the linear regression model.

After that, the regression model will be simplified as much as possible and the residue of the model will be tested to guarantee the accuracy of the model.

In the end, all of the models will be compared to get the most proper one.

3. Results and Discussion

3.1. Single-factor analysis

This research uses R to analyze the data set and it shows a composite graph (Figure.1). This graph shows the correlation between each independent variable and the dependent variable. Through this graph, it can be found that age and BMI seem to have some relationship with charges and it seems to fit linear regression. Then 2 new graphs are plotted comparing age and BMI against charges separately, Figures 2 and 3, to check out the feasibility of the guess. By observing the two figures, it can be found that both age and BMI are positively correlated with the amount of insurance. These two graphs seem to be consistent with linear regression so this model can used with the data set.

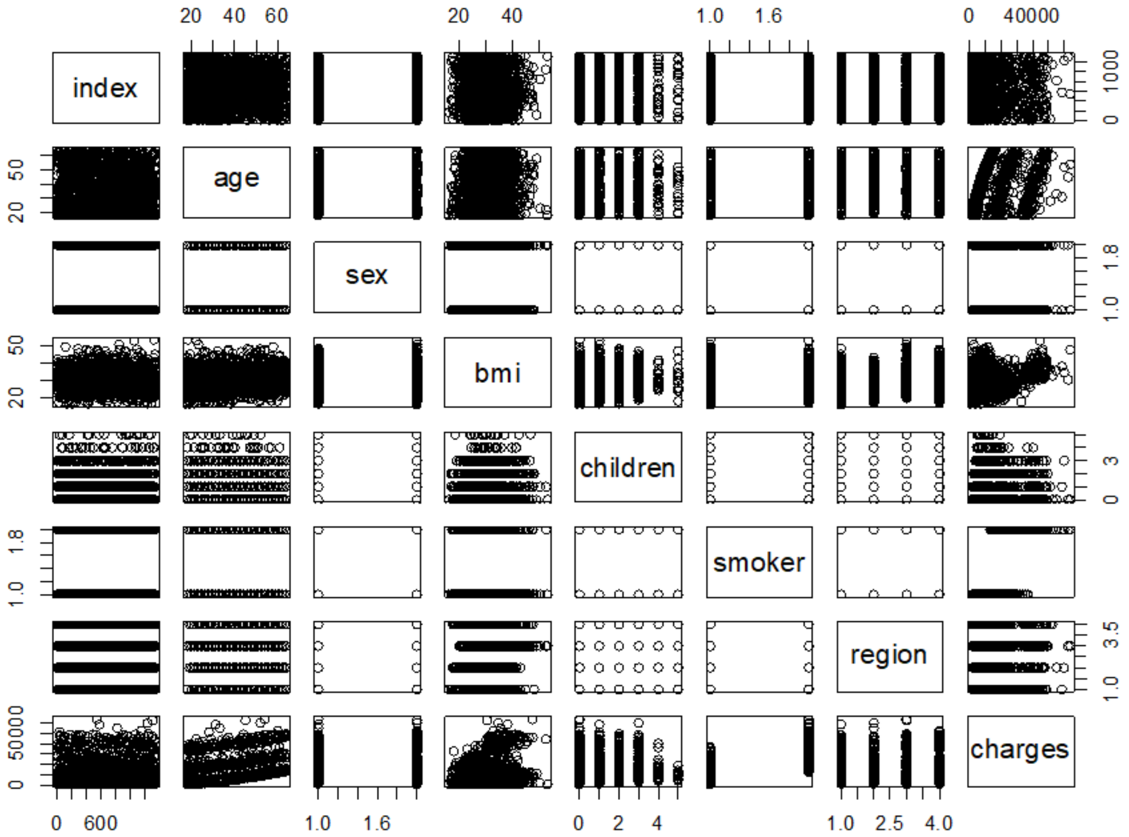


Figure 1. Composite Graph (Photo credit: Original)

By viewing the data set it can be seen that sex and region are factor variables which means that these two data have to be processed. While r can process the data automatically so the research doesn't have to deal with the data.

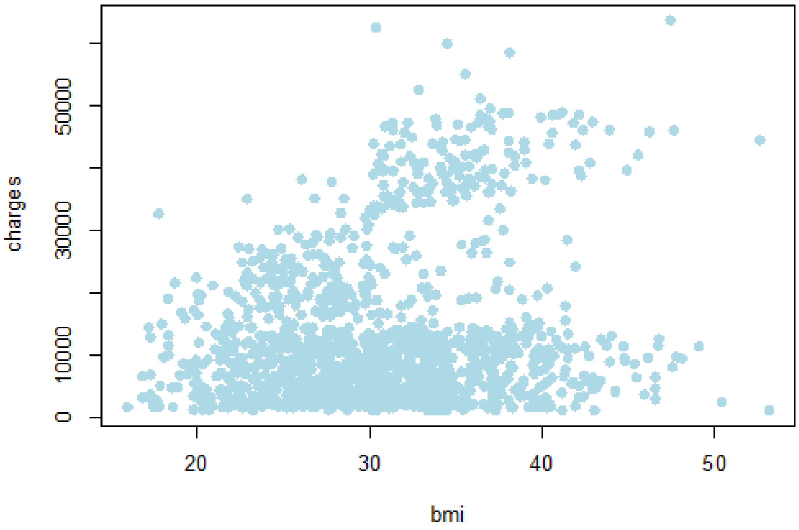


Figure 2. Bmi Against Charges (Photo credit: Original)

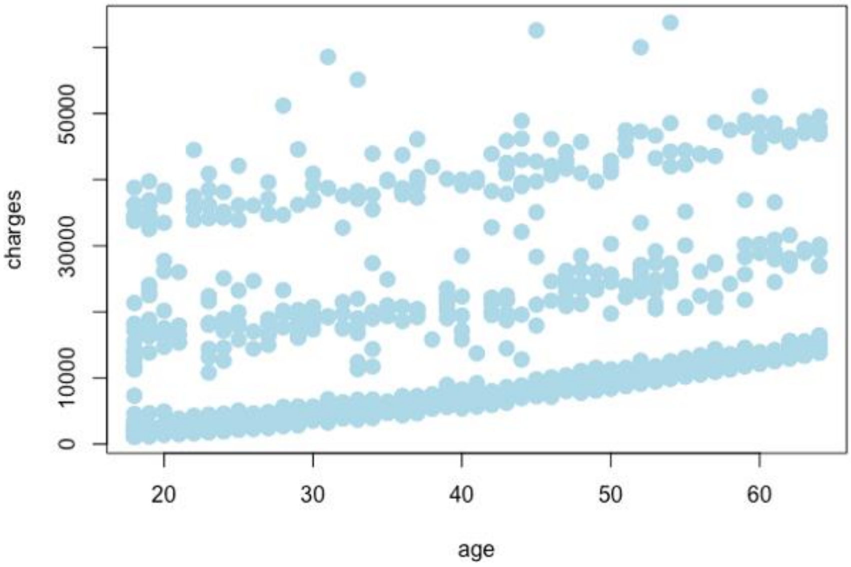


Figure 3. Age Against Charges Graph (Photo credit: Original)

3.2. Multifactor analysis

Table 1. Statistics of Coefficient of the model

	Estimate	Standard error	t value	Pr(>| t |)
Intercept	-11938.5	987.8	-12.086	< 2e^-16
age	256.9	11.9	21.587	< 2e^-16
bmi	339.2	28.6	11.860	< 2e^-16
children	475.5	137.8	3.451	0.000577
regionnorthwest	-353.0	476.3	-0.741	0.458769
regionsoutheast	-1035.0	478.7	-2.162	0.030782
regionsouthwest	-960.0	477.9	-2.009	0.044765
smokeryes	23848.5	413.1	57.723	< 2e^-16
sexmale	-131.3	332.9	-0.394	0.693348

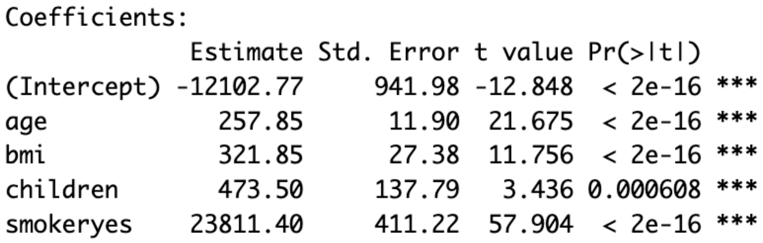
Table 1 shows the data on the coefficient. The estimate is the value of β of each variable. There are also standard error and t values in the table. The estimate in Table 1 is the β in the linear regression model which means that the formula of the model is

Y= -11938.5 + 256.9 age + 339.2 bmi + 475.5 children - 353 regionnorthwest – 1035 regionsouthwest - 960 regionsouthwest + 23848.5 smokeryes -131.3 sexmale

From the t value, an obvious truth can be found that age and BMI have some relationship to the charges to some extent. This can also be seen through the probability of t-test, the data of the number of age, BMI, and smoker is less than 2e^-16 so they are extremely small. The smaller the value of this data means the more relevant it is to the model. This says that age, BMI, and the smoker can truly affect the value of insurance charges. Whether or not the person is a smoker has a very strong relationship to the insurance charges because we can see that the t value of it is 57.723, which is much larger than the other t-value. The estimated value of smokers is also the largest, which means that people who are smokers are much more likely to spend money on health care insurance. Also, an elder relatively is more likely to spend money on health care insurance costs. Bmi is also an important factor that affects the cost of insurance which means that the larger BMI of a person, the more likely this person tends to cost on health care insurance.

The next step is to simplify the model while still maintaining its accuracy of the model. The research uses backward elimination [8] to select the ideal model. The principle of backward elimination is by eliminating the factor of the model which has very little impact on the model. After that, a conclusion can be given by comparing the adjusted r-square value of these 2 models. The adjusted r-squared value of the previous model is 0.7494 while this data of the new model is 0.7489

[9]. There is an extreme difference between the adjusted r-squared value of the 2 models which means that the new model is actually a more proper one.



```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -12102.77     941.98 -12.848  < 2e-16 ***
age            257.85      11.90  21.675  < 2e-16 ***
bmi            321.85      27.38  11.756  < 2e-16 ***
children       473.50     137.79   3.436 0.000608 ***
smokeryes    23811.40     411.22  57.904  < 2e-16 ***
```

Figure 4. Statistics of Coefficient of the new model (Photo credit: Original)

Figure 4 shows the coefficient data of the new model. Hence now the new formula of the model is Y= -12102.8 + 257.9 age + 321.9 bmi + 473.5 children + 23811.4 smokers

3.3. Residual analysis

Figures 4-7 are plotted by R and they are different graphs of the linear regression model.

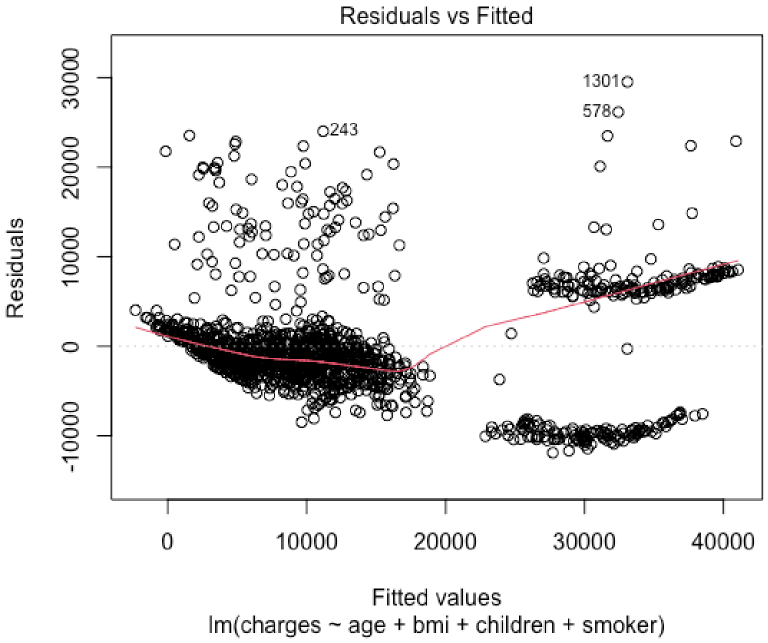


Figure 5. Residuals vs Fitted Graph (Photo credit: Original)

Figure 5 is Residuals vs Fitted Graph; the fitted value appears on the x-axis and the residuals act on the y-axis. This graph does not appear to have obvious abnormalities so this linear regression model does not have to be modified. However, there are some outliers in the graph so it may lead to some errors in the model.

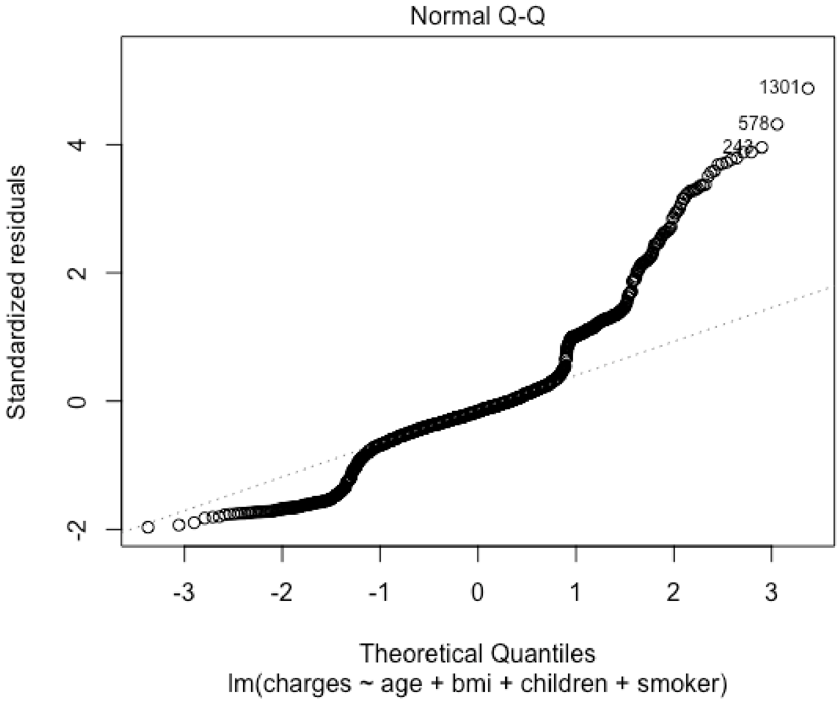


Figure 6. Normal Q-Q Graph (Photo credit: Original)

In Figure 6, the residuals tend to stray from the line quite a bit at the beginning, so it seems that the data does not obey normal distribution. However, on the whole, the image presents a 45-degree angle which means that this model is consistent with the trend of the normal distribution when the data set is large enough.

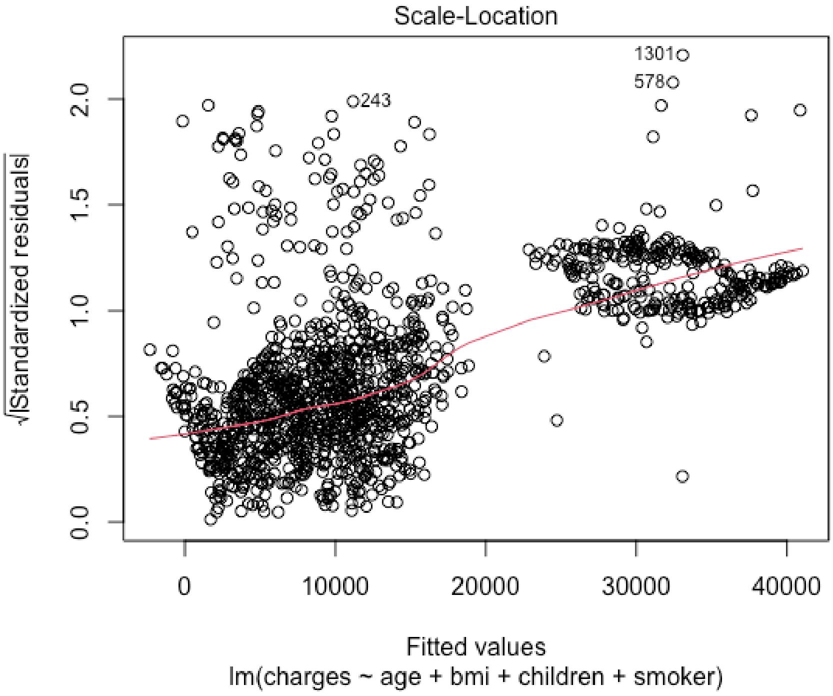


Figure 7. Scale-Location Graph (Photo credit: Original)

In Figure 7, the spread of magnitudes seems to be highest in the fitted values around 10000 and lowest in the fitted values close to 20000. This suggests heteroskedasticity [10]. There are also some outliers in the graph which means that it may impact the model’s accuracy.

3.4. Limitation

By looking at the comparison chart of the data and the model, a conclusion can be drawn that there are many extreme values in the data that greatly affect the model's accuracy. The standard error of the coefficient is relatively too large so it can lead error in the final result. The model deletes some variables and all of them are factor variable, so it may also lead to some error in the model. That is one of the reasons that the adjusted r squared of the new model is just about 0.75. This data illustrates that the model is not so accurate. Another limitation is that the algorithm of the model also needs to be improved, like using stepwise selection and Criterion-based best subsets methods [11]. The study used a relatively simple model which means it may not be as accurate in drawing conclusions. At the same time, the model has not been verified, so at the beginning of the research, some data should be set aside to specifically verify the accuracy of the model.

4. Conclusion

As shown in the research study, a person's specific spending on health insurance can be predicted by a constructed linear regression model. The data used in this model is also relatively easy to collect. At the same time, the accuracy of the model and specific charts is also reflected and verified in the research. Through data analysis, people who smoke seem to be more likely to spend money on health insurance. People also tend to spend more on health insurance as they age and a similar trend can also be seen in a person's BMI index. Therefore, health insurance companies can develop special programs for these aspects. They can also use the model to estimate the cost different people spend on health care insurance.

References

[1] The Private Market for Long-Term Care Insurance in the United States: A Review of the Evidence. (2009). The Journal of Risk and Insurance., 76 (1), 5 – 29. https://doi.org/10.1111/j.1539 - 6975.2009.01286.x

[2] Poisal, J. A., Truffer, C., Smith, S., Sisko, A., Cowan, C., Keehan, S., ... & National Health Expenditure Accounts Projections Team. (2007). Health Spending Projections Through 2016: Modest Changes Obscure Part D's Impact: As health care spending trends remain stable, the Medicare drug benefit changes who pays the bill. Health Affairs, 26(Suppl2), w242 - w253.

[3] Gareth James • Daniela Witten • Trevor Hastie • Robert Tibshirani (August 2021), An Introduction to Statistical Learning with Applications in R Second Edition.

[4] Linear regression. (2012). Wiley Interdisciplinary Reviews., 4 (3), 275 – 294. https://doi.org/10.1002/wics.1198.

[5] Zone-Ching Lin and Wen-Jang Wu, "Multiple linear regression analysis of the overlay accuracy model," in IEEE Transactions on Semiconductor Manufacturing, vol. 12, no. 2, pp. 229-237, May 1999, doi: 10.1109/66.762881.

[6] Zeileis, A., Leisch, F., Hornik, K., & Kleiber, C. (2002). strucchange: An R Package for Testing for Structural Change in Linear Regression Models. Journal of Statistical Software, 7 (2), 1 – 38.

[7] Ritov, Y. (1990). Estimation in a Linear Regression Model with Censored Data. The Annals of Statistics, 18 (1), 303 – 328.

[8] Sutter, J. M., & Kalivas, J. H. (1993). Comparison of forward selection, backward elimination, and generalized simulated annealing for variable selection. Microchemical journal, 47 (1-2), 60 - 66.

[9] Miles, J. (2005). R-squared, adjusted R-squared. Encyclopedia of statistics in behavioral science.

[10] Rigobon, R. (2003). Identification through heteroskedasticity. Review of Economics and Statistics, 85 (4), 777 - 792.

[11] Takano, Y., & Miyashiro, R. (2020). Best subset selection via cross-validation criterion. Top, 28 (2), 475 - 488.