National Cheng Kung University
Department of Statistics

Regression Analysis on Personal Medical Cost Datasets

Professor: Ray-Bing Chen, PhD
Student: H24105014 Oh Wei Sheng

April 2024

1 Introduction

In this project, we are going to estimate personal insurance cost by using the personal medical cost dataset.

2 Data Preview

The dataset used in this project is from Kaggle, containing medical information and costs billed by health insurance companies. It contains 1338 rows of data and the following columns: age, gender, BMI, children, smoker, region and insurance charges. Table 1 displays the definition of each variable.

Table 1: Variable Table

Column Name	Description
age	Age of primary beneficiary
gender	Insurance contractor gender [female, male]
BMI	Body mass index
children	Number of children covered by health insurance / Number of dependents
smoker	Smoking [yes, no]
region	The beneficiary's residential area in the US [northeast, southeast, southwest, northwest]
insurance charges	Individual medical costs billed by health insurance

As mentioned above in the introduction, the insurance charges is set as the target (response variable) in this project, and the rest will be used as the explanatory variables.

Figure 1 shows the pairwise plot of the variable. From the diagonal plot from figure 1, we may observe the distribution of each variable, which ‘age’ is approximately uniformly distributed, ‘sex’ and ‘region’ are quite balanced, ‘BMI’ is approximately normally distributed, and ‘children’ and ‘smoker’ are not quite balanced.

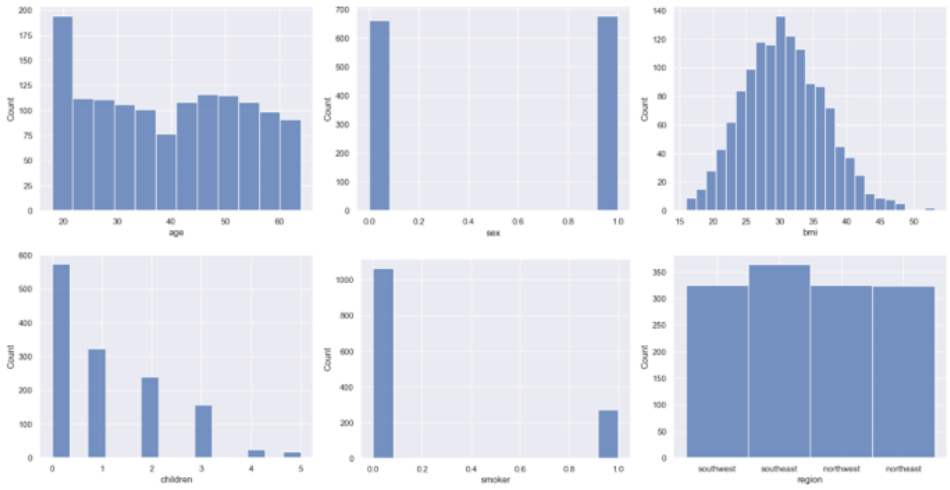

Figure 1: Distribution of Variable

Figure 2 shows the distribution of the response variable, the insurance cost. From figure 2, we may observe the distribution is slightly right-skewed, indicates that the data is concentrated at a lower amount of insurance cost.

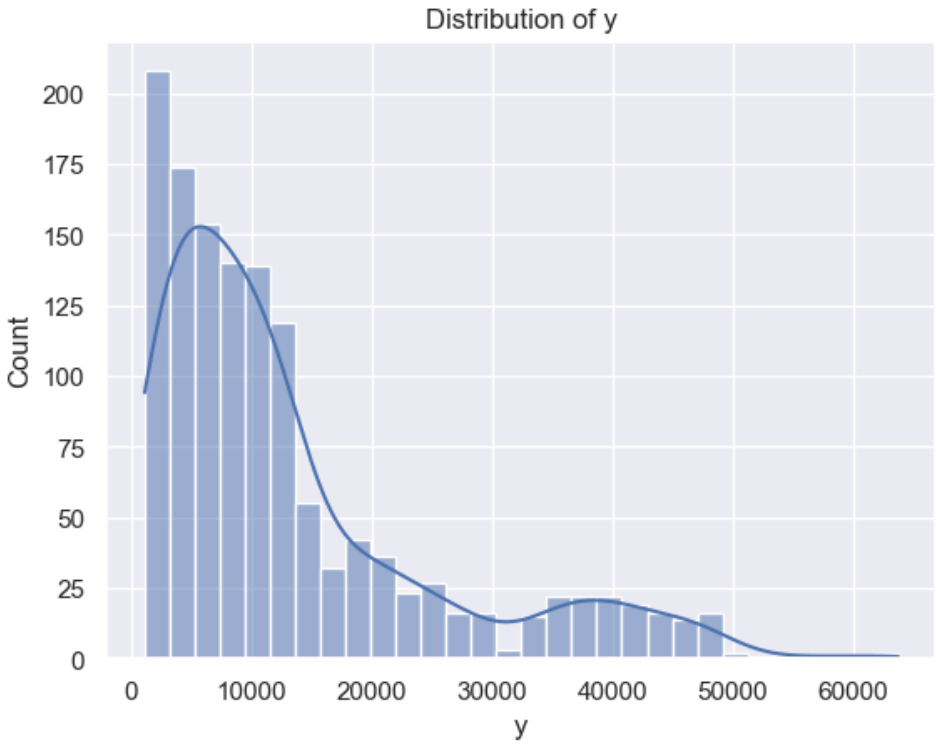

Figure 2: Distribution of Insurance Cost

3 Methodology

In the dataset, there are various categorical data. These include sex, smoker, and region. Hence, we transformed these variables into dummy variable, which takes value on 1 or -1. Especially the region variable is being encoded into 4 columns, which are northeast, northwest, southeast and southwest. However, the 'children' variable is a ordinal discrete variable as it only takes value on integer, we do not perform transformation on this variable since it is ordinal, although it is not continuous. Furthermore, the interaction terms are created by multiplying each two of the variables.

In this project, multiple linear regression model with second degree interaction effect is used to train the data. The general equation of the model is:

$$\mathrm{Y} = \beta_0 + \sum \beta_i X_i + \sum\sum \beta_{ij} X_i X_j + \varepsilon \quad , \quad i \neq j \quad \forall\, i, j$$

Y: the response variable, insurance cost

β_0: the intercept of the model

β_i: the coefficient of the i^{th} variables

X_i: the i^{th} explanatory variable

ϵ: the error terms

besides computing the coefficient of each variable, we would like to see which features and their corresponding interaction effect are significant in the model under 0.05 α level by using t-test.

The aim is to construct a model to precisely estimate the value of insurance cost. To measure the performance of the, root mean squared error (RMSE) and R^2 are introduced as below to evaluate the model:

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}{n}} \quad , \quad \mathrm{R}^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

RMSE represents the difference between the predicted value and the true value of burned area and R^2 represent how much of the total variation can be explained by the model. The lower the RMSE, the better the outcome as the predicted result is closer to the real value meanwhile the higher the R^2, the better the model.

The dataset will be split in two sets: training set and testing set, with a ratio of 7:3. The training set will be used to fit the model, and the testing set will be used to evaluate the performance of the model, based on RMSE.

4 Result

After fitting the model, its RMSE equals to 4419.23 meanwhile the R^2 is 0.857. In terms of RMSE, the performance is not as good as expected. However, in terms of R^2, the performance is acceptable. There are only 2 significant main effects, which are sex and the number of children. However, there are several interaction effects that are significant. In table 2, we have those terms which have their p-value smaller than 0.05 and their coefficient.

Table 2: Significant Variables & Coefficient

Variable	Coefficient	Variable	Coefficient
Sex	89.45	Age*Southeast	-47.3
BMI	295.74	BMI*Smoker	708.83
Smoker	-5093.69	BMI*Northeast	-115.75
Northeast	4366.57	BMI*Northwest	-125.62
Northwest	4206.8	BMI*Southeast	-188.04
Southeast	5833.39	BMI *Southwest	-162.07
Southwest	4341.47	Smoker*Northeast	2402.28
Age*Children	-20.99	Smoker*Northwest	2454.8
Age*Northeast	-53.78	Smoker*Southeast	2412.22
Age*Northwest	-52.69	Smoker*Southwest	2918.08

As shown, there are 4 main effects that are significant, which include sex, BMI, smoker and region. However, children variable is only significant in one of the interaction effects, thus we do not consider it as an important variable.

Furthermore, we analyze the error term of the model using training set, the residual has zero mean. However, the variance of residual has a relatively large value, which is approximately equals to $2.44*10^7$.