Predicting Medical Insurance Charges from Age, BMI and Smoking Status

DA Group 6

1 Introduction

~~In today's world, medical care is undoubtedly an expensive affair.~~ Consequently, numerous individuals enroll in a health insurance policy, where they agree to pay a sum of money (known as premiums) to a particular health insurance company, on a monthly or yearly basis. In return, the company will guarantee to reimburse a proportion of the medical costs in case the insured is injured or sick and needs medical treatment.

The ~~amount of~~ premium which an individual pays, depends on a number of factors. In particular, the higher the risk of having health problems, the higher the premium ~~would~~ be. For instance, overweight or obese persons are more likely to develop heart disease when compared with people of normal weight. Moreover, smoking is linked with an increased risk of lung cancer when compared to individuals who do not smoke. Apart from these, there is a tendency for health insurance rates to escalate with increasing age, since older people are more prone to health problems. In these situations, one might expect these individuals to have a higher risk of large medical expenses.

The ~~main~~ goal of this report is to predict the total yearly medical costs (`charges`) billed by health insurance from the age of the individual (`age`), the corresponding body mass index (`bmi`) and from whether the insured smokes tobacco or not (`smoker`). The `bmi` is a measure of body fat, which is defined as the body weight (in kg) divided by the square of the body height (in m^2). The data analysed here consists of a sample of 364 individuals living in the South East region of the United States, from 1338 observations, which contain hypothetical data on medical expenses for patients in the United States. The simulated data was based on the demographic statistics obtained from the US Census Bureau, and thus, it resembles real world data.

Section 2 consists of an exploratory data analysis to gain a better understanding of the distribution of the features used under this study. Section 3 provides the process of selecting the best regression model to predict the insurance charges after fitting a number of regression models to the data. ~~Moreover,~~ the model assumptions are also checked here. Finally, Section 4 sums up the results obtained after conducting this analysis.

2 Exploratory Data Analysis

To get an idea of the distribution of the data, the following summary statistics were obtained for the categorical variable `smoker` (Table 1) and for the continuous variables `charges`, `age` and `bmi` (Table 2).

Table 1: Numbers of smokers and non-smokers.

Smoker	n
no	273
yes	91

~~As can be clearly seen~~ from Table 1, out of the 364 individuals, the majority of them do not smoke (273 non smokers vs 91 smokers). In fact, only one fourth of the individuals in the study smokes.

~~If we look at~~ Table 2, the mean age of the individuals is 38.94 years, with a standard deviation (SD) of 14.16 years. Next, the middle 50% of the bmi lies between 28.57 and 37.81 kg/m^2, with an average bmi value of

Table 2: Summary statistics on insurance charges, age and bmi.

Variable	n	Mean	SD	Minimum	1st quartile	Median	3rd quartile	Maximum
age	364	38.94	14.16	18	26.75	39	51	64
bmi	364	33.36	6.48	19.8	28.57	33.33	37.81	53.13
charges	364	14735.41	13971.1	1121.87	4440.89	9294.13	19526.29	63770.43

33.36 kg/m^2 and standard deviation 6.48 kg/m^2. Finally, the middle 50% of the data for the medical costs (charges) lies between 4440.89 and 19526.29 dollars, with an average of 14735.41 dollars. The variation in the mean total charges seems to be quite high, with a value of 13971.1 dollars.

In order to measure the degree of association between the continuous variables in the study, the pairs scatterplot is plotted in Figure 1. The plot shows that there is no strong relationship between any two continuous variables. To confirm this, the correlation coefficients for each pair of variables was calculated. The correlation between the response variable charges and the explanatory variables age and bmi were found to be 0.311 and 0.143, respectively. Moreover, there does not seem to be ~~any~~ linear association between the two continuous explanatory variables age and bmi (0.02), implying that there is no evidence of multicollinearity in the data.

Figure 1: Pairs plot between charges, age and bmi

Figure 2 on the next page, shows two scatterplots of the insurance charges against each of the explanatory variables by smoking status of the individuals. From the left hand plot, which shows the relationship between charges and age by the smoking status, it is evident that as people ~~mature,~~ the health insurance charge increases. The plot indicates thata on average, people who do not smoke, pay less than ~~40000 dollar~~ whereas smokers pay up to ~~60000 dollars~~ on their health insurance. However, ~~the associated effect of age does not seem to change differently between the smokers and non smokers.~~ ~~On the other hand, the slopes of the lines of the right hand plot are distinct~~, and thus, ~~bmi seems to change differently with the smoking status.~~ In particular, it appears from the plot that as the bmi of a smoker individual increases, the corresponding insurance charges increase drastically. This is in contrast with those individuals who do not smoke, where a change in the bmi does not seem to make a significant change in the insurance charges. To sum up it is clear from both plots that ~~smoking people pay larger amounts on their health insurance~~ when compared with those who do not smoke.

To predict the medical insurance charges, a number of linear regression models will be fitted with age, bmi and smoking status as potential predictors.