Group: TUT22_Group9

1. Introduction

1.1. Background

In a purpose of balancing profit and risk, health insurance institutions rely on the clients' private information to adjust the medical charges to various individuals. This report will analyze how the clients' family and personal behaviors will affect the medical charges to them by using dataset from kaggle.com, which contains data for 1,338 individuals.

1.2. Data

The data was sourced from kaggle.com and originally published by GitHub https://github.com/stedy/Machine-Learning-with-R-datasets.

There are three ratio data (age, BMI, charges) and 4 categorical data (sex, children, smoker and region) in the dataset. The original dataset is pre-proceeded by python to translate categorical data into numerical expression, and the explanations are as follow:

age: Range from 18 to 64 years old

sex: 1 means female 0 means male

bmi: body mass index (will not be used in the report)

smoker: 1 means smoker 0 means non-smoker

children: range from 0~4, means the number of child in a family

region: southwest (USA) southeast (USA) northwest (USA) northeast (USA)

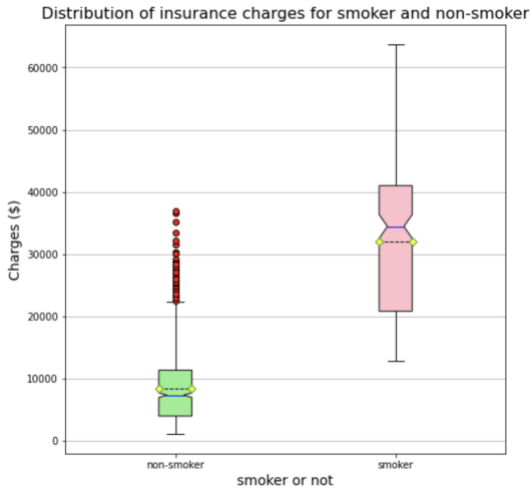
charges: medical cost billed by health insurance

1.3. Interest of direction

The main question covered in the report is the amount of extra dollars are charged to smokers by health insurance. Additionally, how health insurance according to the various of regions, age, and number of child adjust the insurance charges. This report discoveries the pricing method of medical insurance, thus can help clients understand deeper how they are charged by insurance institutions.

2. Preliminary analysis

As the boxplot of smoker and non-smoker's medical charges demonstrates clearly, the median and mean charges for smoker is much higher than non-smoker.



smoker	mean_charge	median_charges
1	32050.231832	34456.348450
0	8434.268298	7345.405300

Above table contains the exact value of mean charges and median charges for smoker and non-smoker. We can infer that, in accordance with the smoking behavior, there are an average of about $25,000 price discrimination to clients. But something interesting should be noticed that there are no outliers for smoker, and lots of upper outliers for non-smoker. Moreover, the lower fence of smoker charges distribution is less than the upper fence of non-smoker, therefore, there are might some other features also considered by insurance institution to charge the medical cost. But we will first consider the single variable and apply linear regression to explore the association between smoking behavior and medical charges.

3. Regression analysis

3.1. Single variable regression

By introducing OLS method, the beta0 and beta1 was calculated as follow. Smoker is a dummy variable and regards to 0 if do not smoke, 1 if smoke.

$$\widehat{charges} = 8434.2683 + 23615.9635 * smoker_i$$

The regression describes the information that the insurance institutions expect to charges 23,615.96 more dollars to smoker, comparing to non-smoker. To assess the regression model, the statistic describe of the regression model is presented as follow:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.78725143
R Square	0.61976481
Adjusted R S	0.61948021
Standard Err	7470.21621
Observations	1338

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	1.2152E+11	1.2152E+11	2177.614868	8.271E-283
Residual	1336	7.4554E+10	55804130.2		
Total	1337	1.9607E+11			

	Coefficients	*tandard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	8434.2683	229.014172	36.8285868	1.581E-205	7985.00176	8883.53484	7985.00176	8883.53484
X Variable 1	23615.9635	506.07529	46.6649212	8.271E-283	22623.1748	24608.7523	22623.1748	24608.7523

From the table, the value of r is approx. 0.70, which behalf of the evidence of strong association between variables. 0.62 of R^2 means the explainable error term (SSR) takes most percent than unexplainable error (SSE). Thus, the regression model performs well in predicting charges via dummy variable "smoker". Also, the values of R^2 and $adjusted$ R^2 are quite close, therefore, the model is valid and will not be significant affected by error terms. The p-value of coefficient is close to 0 gives evidence that the effect is statistically significant and not caused by random variation.

3.2. Multiple-variables regression

The simple regression model shows that the effect of smoking behavior on people's spending on medical treatment in the US, however, smoking behavior must not be the only factor that affects medical charges. Therefore, we extend the model 1 by including additional regressors: region, age (*Years*) as well as the number of children *(Person)*. The region is a categorical variable that takes one of four values, either Southwest, Southeast, Northwest or Northeast. Three dummies were defined as Southwest, which equals 1 if the region is Southwest and 0 otherwise, as Southeast, which equals 1 if the region is Southeast and 0 otherwise and Northwest, which equals 1 if the region is Northwest and 0 otherwise. The region of Northeast is the base case. Then the estimated multiple regression model is given by:

$$\widehat{Charges}_i = -2762.92 + 273.29 * Age_i + 496.86 * Children_i + 23780.51 * Smoker_i - 345.95 * Northwest_i + 389.81 * Southwest_i - 484.17 * Southwest_i$$

Assuming all other variables constant, when age increase by 1 year, the charge on medical in the US is expected to increase by $273 on average. For the number of children, assuming all other variables constant, when the number of children increase by 1person, the medical cost is expected to increase by $496.86 on average. The data output of the regression model as follow:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.8512018
R Square	0.72454451
Adjusted R S	0.72330279
Standard Err	6370.11006
Observations	1338

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	6	1.4206E+11	2.3677E+10	583.499447	0
Residual	1331	5.401E+10	40578302.2		
Total	1337	1.9607E+11			

	Coefficients	*tandard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-2762.9231	624.077645	-4.4272105	1.0327E-05	-3987.2061	-1538.6401	-3987.2061	-1538.6401
northwest	-345.95393	500.470119	-0.6912579	0.48952402	-1327.7501	635.842274	-1327.7501	635.842274
southeast	389.80921	486.912201	0.80057392	0.42352123	-565.38978	1345.0082	-565.38978	1345.0082
southwest	-484.17044	500.44336	-0.967483	0.3334784	-1465.9142	497.573268	-1465.9142	497.573268
smoker	23780.5104	432.891568	54.934104	0	22931.2863	24629.7346	22931.2863	24629.7346
age	273.294495	12.4155097	22.012346	7.8941E-92	248.938395	297.650595	248.938395	297.650595
children	496.860185	144.770871	3.43204528	0.00061742	212.856232	780.864138	212.856232	780.864138

Comparing the R-squared between Model 1 and Model 2, the adjusted R-squared increased from 0.6197 to 0.7245, and the Multiple R increased to 0.85. Thus, the multi-variable regression performs better in predicting the medical charges by considering more features. In addition, both the P-value associated with the coefficient for age, the number of children and smoker are close to 0 which means they all significant at the 5% level, while the coefficient of region is not significant at the 5% level since the P-value greater than 0.05. Therefore, our findings suggest that in addition to the impact of smoking behavior on health care costs, the number of children and age should also be considered.

4. Assumptions

For the appeal inference we made several statistical assumptions, and it must be met those assumptions. Firstly, for autocorrelation, since our data do not have dates on medical costs, if

these costs are intercepted within a similar time frame and therefore can be seen as cross-sectional data, it is unlikely to have this problem. Secondly, homoscedastic is normal issue for cross-sectional data, so there is no need to more future analyze. Finally, for normally distributed errors, the sample size is more than 30 (n=1338), so errors obey normal distribution according to the Central Limit Theorem.

5. Conclusion

The model represents the additional costs that health insurance does charge smokers, and in addition, to further confirm whether family and individual behavior can also influence health care expenditures, we added more variables such as age, number of children and region. Except for region, there was sufficient evidence to conclude that all the variables were statistically significant. Another usefulness of the model is to explore how health insurance is priced. Through this study, people can clearly understand how they are charged and can also better choose the right insurance package for themselves according to their individual and family behavior.