An analysis of the factors that determine the cost of health insurance for individuals

Objective of research

The main objective of this research is to determine the factors that explain the price of health insurance charged by insurance companies to individuals. The study will aim to determine the factors that have the most impact on the price of health insurance. This study therefore focuses on interpretation of the factors that determine the cost of health insurance.

Data description

The dataset consists of seven variables/columns. The variables are as follows:

- Age - age of primary beneficiary
- Sex – gender of the insurance customer, male/female
- BMI- Body mass index, providing an understanding of body, weights that are relatively high or low relative to height (ideal value 18.5 to 24.9)
- 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.
- Charges - Individual’s medical costs billed by health insurance, the units are in United States Dollars.

Data cleaning and feature engineering

The data contains 1338 rows for each column. All the variables were checked to ensure that there were no missing values. There were indeed no missing values in the data.

The categorical variables sex, smoker and region were converted to numerical using Pandas functions. This is so that we can fit them well in our regression models. The resulting dataframe now had nine variables in total.

Training the regression models

Description of models

For the linear regression models the dependent variable or the variable to be predicted is charges, which is the cost of health insurance as charged by the insurance company. The explanatory or independent variables are age, sex, bmi, children, smoker and region. The equation for the models is as follows:

Charges = $\beta_0 + \sum_{i=1}^{n}(\beta_i X_i)$

Where β_0 is the intercept and β_i and X_i are the coefficient and variable respectively.

Simple linear regression

Firstly, a simple linear regression model was fitted to the data. The coefficients for each variable for the model are as shown in figure 1 below. The R-squared score for the model is 0.7154803806270008, which means the models explains roughly 71.5% of the variations.

Variable	Coefficient
region_southeast	-1227.54
region_southwest	-1195.96
region_northwest	-877.21
sex_male	213.86
age	260.85
bmi	332.33
children	419.93
smoker_yes	24152.63

Figure 1: Linear regression coefficients

Ridge regression

Ridge regression was fitted to the data and trained using cross validation. The data was first scaled using standard scaling. The coefficients for the models are as shown in figure 2 below. The R-squared score for the model was 0.715853673400537. This means the model explains about 71.6% of the variation which is slightly better than simple linear regression.

Variable	Coefficient
region_southeast	-530.00
region_southwest	-508.94
region_northwest	-371.59
sex_male	109.11
children	504.11
bmi	2064.64
age	3614.54
smoker_yes	9774.31

Figure 2: Ridge regression coefficients

Linear regression with polynomial features

Polynomial features of order 2 were added to the variables age, children and bmi. A simple linear regression was then fitted to this data with polynomial features. The coefficients for the model are shown in figure 3 below. The R-squared score for the model was 0.7223876847255386. This means the model explains about 72.2% of the variation which is slightly better than both the simple linear regression without polynomial features and the Ridge regression.

Variable	Coefficient
region_southwest	-1243.69
region_southeast	-1165.27
region_northwest	-951.40
children	-520.97
children^2	-95.19
age	-35.77
bmi^2	-6.16
age x bmi	1.48
age^2	3.11
age x children	4.06
bmi x cildren	39.87
sex_male	189.65
bmi	613.44
smoker_yes	24212.68

Figure 3: Regression with polynomial features coefficients

Choice of final model

Adding polynomial features and regularization only increased slightly the percentage of the variations that the model is able to explain. Adding polynomial features increases the complexity of the model and regularization increases the computational time needed to fit the model. All these complexities made the models just slightly better than simple linear regression. My final choice for the model is therefore simple linear regression as adding polynomial features and regularization did not significantly improve the percentage of the total variation that the model was able to explain. Furthermore, simple linear regression is easier to interpret than the other models hence it is better.

Summary of key findings

Using the simple linear regression model chosen as the best model, I will highlight the key findings. The factor that has the most impact on the price of health insurance charged by insurance companies is smoking. If an individual is a smoker it increases the price charged by $24,152.63, holding all other variables constant.

Staying in the three regions, South East, South West and North West all decrease the cost of insurance by $1,227.54, $1,195.96, $877.21 respectively holding all other variables constant.

If an individual is a male it increases the cost of health insurance by $213.86, holding all other variables constant.

A unit increase in age, bmi and children will increase the cost of insurance by $260.85, $332.33 and $419.93 respectively, holding all other variables constant

Suggestion for next steps

All the three linear regression models failed to explained the variations in the model to a satisfactory extent. This is seen by the highest R-squared score achieved of 72.2% which was achieved by the linear regression model with polynomial features.

A possible reason for this is that the explanatory variables used were not enough to capture the whole underlying process. A suggestion would be to add more explanatory variables to the data and perform the linear regression again.

Another possible reason might be that the linear regression model is not very appropriate for this problem and maybe the use of another model will lead to better results.