Abstract

The aim of this report generated using R markdown is to predict the insurance charges of individuals based on their demographic and health-related characteristics using multiple linear regression analysis. The dataset contains 1338 observations and 7 variables. The dataset was examined and a prediction model was created using descriptive statistics, correlation analysis, and multiple linear regression analysis.

The developed model was able to explain 86% of the variation in insurance charges. This study can help insurance companies and policymakers set fair and affordable insurance premiums and provide insights into the relationship between health-related characteristics and insurance costs.

Introduction

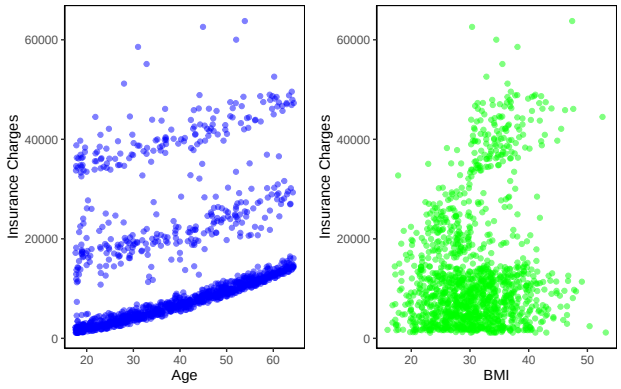
The cost of healthcare has been increasing rapidly in recent years, and medical insurance has become a crucial aspect of healthcare financing. Insurance companies need to set insurance premiums that are fair and affordable for their clients while still covering their costs.

Multiple linear regression analysis is used to develop a model to predict insurance charges for individuals based on their demographic and health-related characteristics using the Medical Cost Personal Dataset available on Kaggle. The findings can assist insurance companies in setting fair and affordable insurance premiums for clients and help policymakers design public health policies to address healthcare costs.

Data overview

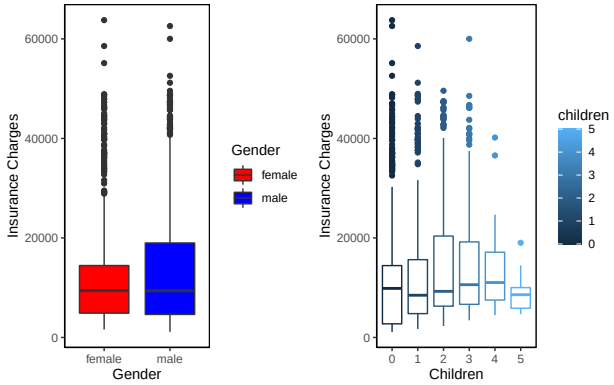
To assess the correlation between insurance charges and other variables including age, sex, BMI, number of children, smoking status, and region we used the cor() function in R.

Plot 1: Correlation between Insurance Charges and Age / BMI



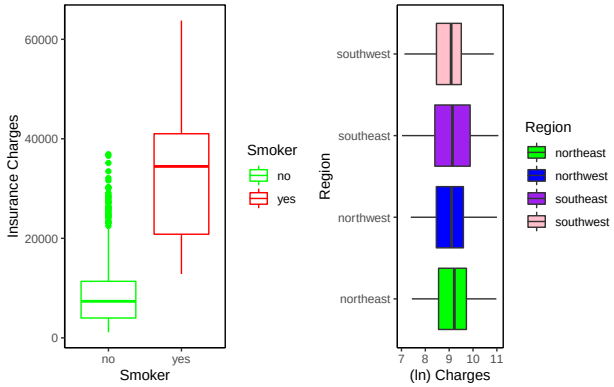
The result above depicts a positive correlation between insurance charges and age (correlation coefficient = 0.3) and BMI (correlation coefficient = 0.2), indicating that as age/BMI increases, insurance charges also tend to increase. These findings suggest that age and BMI are important predictors of insurance charges and should be considered when determining insurance premiums.

Plot 2: Correlation between Insurance Charges and Sex / Children



The results above shows that there is no significant correlation between insurance charges and sex (correlation coefficient = -0.06) or number of children (correlation coefficient = 0.07). These findings suggest that sex and number of children may not be strong predictors of insurance charges. However, further analysis is needed to explore the potential interactions between these variables and other demographic and health-related characteristics in predicting insurance charges.

Plot 3: Correlation between Insurance Charges and Smoker / Region



The above results depict a significant positive correlation between insurance charges and smoking status (correlation coefficient = 0.79), indicating that smokers tend to have higher insurance charges compared to non-smokers. There was also a moderate correlation between insurance charges and region (correlation coefficient = 0.21), suggesting that individuals living in certain regions may have higher insurance charges. These findings highlight the importance of considering smoking status and region when setting insurance premiums, and further research may be needed to explore the underlying factors contributing to these correlations.

Method used to analyze data

The method used to analyze the data involves multiple linear regression analysis. A step-by-step illustration of the process is given below:

1. Data Exploration: The first step in analyzing the data is to explore the variables and their distributions. This involves using summary statistics, visualizations, and tests of normality to assess the data quality.

2. Variable Selection: Next, we select the variables to include in the regression model. In this case, we choose age, BMI, children, smoker, region, and sex as potential predictors of insurance charges.

3. Categorizing BMI: The BMI variable was originally a continuous variable. To incorporate it into the regression model, it was categorized into four common categories namely underweight, normal, overweight, and obese

4. Data splitting: Before fitting the regression model, the dataset was split into training and testing datasets. The training dataset was used to fit the regression model, while the testing dataset was used to evaluate the model's predictive accuracy. The sample() function was used to randomly split the dataset into 80% training and 20% testing data.

5. Model fitting: We then fit a multiple linear regression model to the data, using the selected variables as predictors and insurance charges as the outcome variable.

6. Variable Interactions: Another step in analyzing the data was to investigate the interactions between smoking and other predictor variables in the regression model as smoking is known to have a significant impact on health and can interact with other factors to affect insurance charges. For example, we added the interaction term "smoker:age" to the model to capture the potential interaction between smoking and age.

7. Model evaluation: We evaluate the performance of the regression model using various metrics such as R-squared, adjusted R-squared, residual plots, and hypothesis tests of the regression coefficients. These metrics help us assess the goodness of fit of the model and the statistical significance of each predictor variable.

8. Prediction: Finally, we use the fitted model to make predictions on new data. We can predict the insurance charges for individuals with specific values of age, BMI, children, smoker, region, and sex by plugging these values into the regression equation.

Training a model

Five multiple regression models were built using different sets of predictor variables to try to explain the variation in the outcome variable (charges).

1. The first model was built using age, BMI category, and smoking status as predictor variables. These variables were selected based on the significance of their correlation to insurance charges.

2. The second model was built by adding remainder variables to the first model. This model was created to determine if these variables are significant predictors of charges, and to assess if adding these variables improves the model fit.

3. The third model includes variables from the previous model apart from sex due to its significance level. This model is easy to interpret having an adjusted R-squared of 0.75 and an F-test value of 365.66. This indicates that the model explains 75% of the variation in insurance charges and that the model is statistically significant.

4. The fourth model includes all predictor variables from the third model as well as three interaction terms being smoker * BMI category, smoker * age, and smoker * region. This model is more complex and may be more difficult to interpret, but it has a higher adjusted R-squared of 0.86 and an F-statistic of 423.51, suggesting that it may provide a better fit to the data.

5. The fifth and final model includes all predictor variables apart from gender. It also includes only the interaction term between smoker and BMI category, in addition to the main effects. This model is less complex than the fourth model but more complex than the third. It has an adjusted R-squared of 0.86 and an F-test value of 564.39.

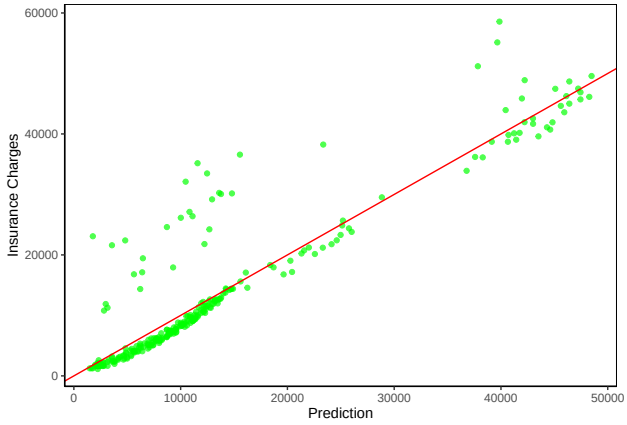
On comparing the third, fourth, and fifth model we can see that the last 2 models have a higher adjusted R-squared value, indicating that they explain more of the variation in the data. However, the final model has a higher F-test value compared to the fourth model, which suggests that it is a slightly better fit for the data.

Therefore, based on the adjusted R-squared and F-test values, we can conclude that the fifth and final model is the best model amongst the lot.

Testing a model

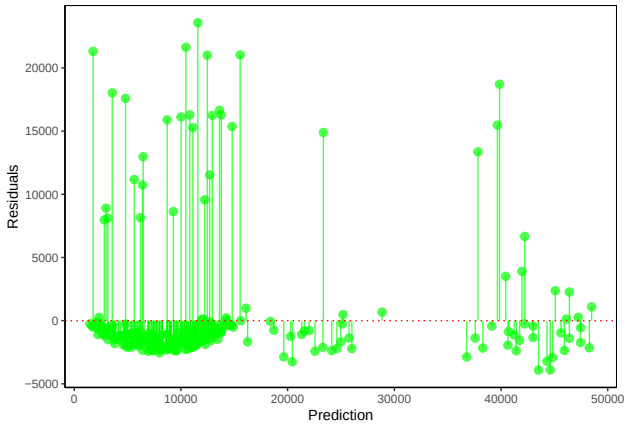
We test models to assess how well they can predict outcomes using a set of input variables or predictors. By testing a model, we can evaluate its accuracy and determine whether it can generalize to new data. This is important because a model that performs well on training data but poorly on new data is not useful in practice.

Plot 1: Prediction vs Actual Values



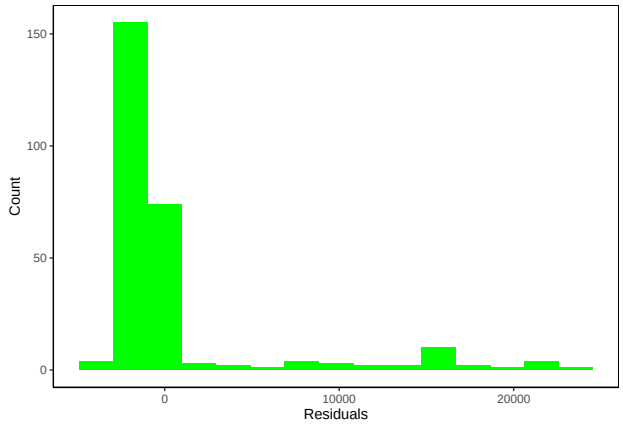
Looking at the prediction vs. actual values plot, we can see that the predicted charges are generally in line with the actual charges, as most of the points fall close to the red line representing perfect predictions.

Plot 2: Residuals vs Linear Model Prediction



The residuals vs. linear model prediction plot shows that the residuals are randomly scattered around zero, which is a good indication that the model is a good fit for the data.

Plot 3: Residual Histogram



The residual histogram shows that the residuals are approximately normally distributed with a mean of zero and a standard deviation of around 10, which is another good indication that the model is accurate.

Conclusion

1. Age has a positive impact on charges with an estimate of 262.53. This means that for every one year increase in age, the charges increase by $262.53 on average.
2. BMI category variables (Obese, Overweight, and Underweight) do not have a statistically significant impact on charges as their p-values are greater than 0.05.
3. The number of children has a positive impact on charges with an estimate of 514.98. This means that for every one additional child, the charges increase by $514.98 on average.
4. Smoking has a strong positive impact on charges with an estimate of 11645.15. Smokers pay significantly higher charges compared to non-smokers.
5. Region variables (Northwest, Southeast, and Southwest) also have a statistically significant impact on charges. Among these, the Southwest region has the largest negative impact with an estimate of -1018.58, indicating that individuals in the Southwest region pay lower charges compared to those in other regions.
6. The interaction terms between BMI categories and smoking status have statistically significant impacts on charges. The BMI category of Obese and smoking status have the largest impact on charges with an estimate of 21495.57. This means that smokers in the Obese category pay the highest charges on average.
7. The adjusted R-squared value of 0.86 indicates that this model explains 86% of the variability in charges. The F-test value of 564.39 and a p-value of 0.00 indicate that the overall model is statistically significant.

In conclusion, this model suggests that age, number of children, smoking status, and region are important predictors of charges. Additionally, the interaction between smoking status and BMI category is also a significant factor in determining charges. However, BMI category alone does not have a significant impact on charges.

References

https://www.kaggle.com/datasets/mirichoi0218/insurance (2018), Updated 21 February 2018