UNIVERSITY OF EDINBURGH
Business School

2022-23

CMSE114592022-3SS1SB3
DATA MINING 1

PREDICTING INSURANCE CHARGES USING MULTIPLE LINEAR REGRESSION

B216927

WORD COUNT: 824

Introduction

This report uses multiple linear regression analysis to predict insurance charges based on demographic and health-related characteristics. The aim is to aid insurance companies in setting fair premiums and policymakers in designing effective public health policies to address healthcare costs. The model explains 86% of the variation in charges and is based on the Medical Cost Personal Dataset available on Kaggle.

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.

- There is a positive correlation between insurance charges and age (correlation coefficient = 0.3) and BMI (correlation coefficient = 0.2). This suggests that as age and BMI increase, insurance charges tend to increase as well.
- No significant correlation exists between insurance charges and sex (correlation coefficient = -0.06) or the number of children (correlation coefficient = 0.07). Thus, these variables may not strongly predict insurance charges.
- The correlation analysis reveals a strong 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. Additionally, there was a moderate correlation between insurance charges and region (correlation coefficient = 0.21), suggesting that individuals living in certain regions may have higher insurance charges.

The method used to analyze data

The method used for analyzing the data involved exploring the variables and their distributions, selecting relevant predictor variables, categorizing BMI into four common categories, splitting the dataset into 80% training and 20% testing data, fitting a multiple linear regression model to the data, investigating interactions between smoking and other predictor variables, evaluating the model's performance using various metrics, and making predictions based on the model.

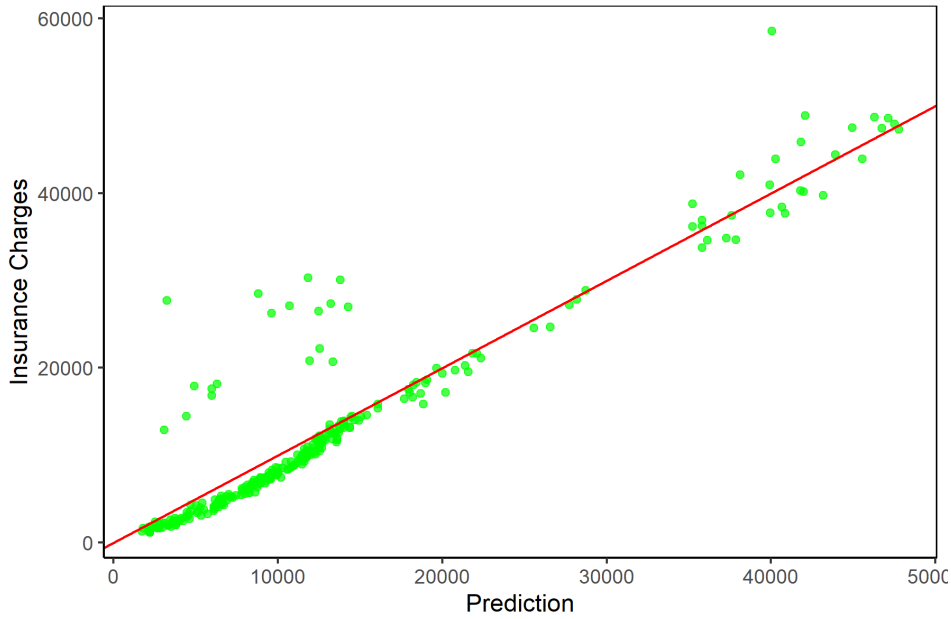
Training a model

Three multiple regression models were built to explain the variation in insurance charges using different predictor variables. The first model included all variables except gender, with an adjusted R-squared of 0.76 and an F-test value of 368.92. The second model added three interaction terms: smoker * BMI category, smoker * age, and smoker * region, resulting in an adjusted R-squared of 0.86 and an F-statistic of 416.06. The third model included only the smoker * BMI interaction term and all other main effects, with an adjusted R-squared of 0.86 and the F-test value of 551.87. Based on adjusted R-squared and F-test values, the third model was identified as the best fit for the data.

Testing a model

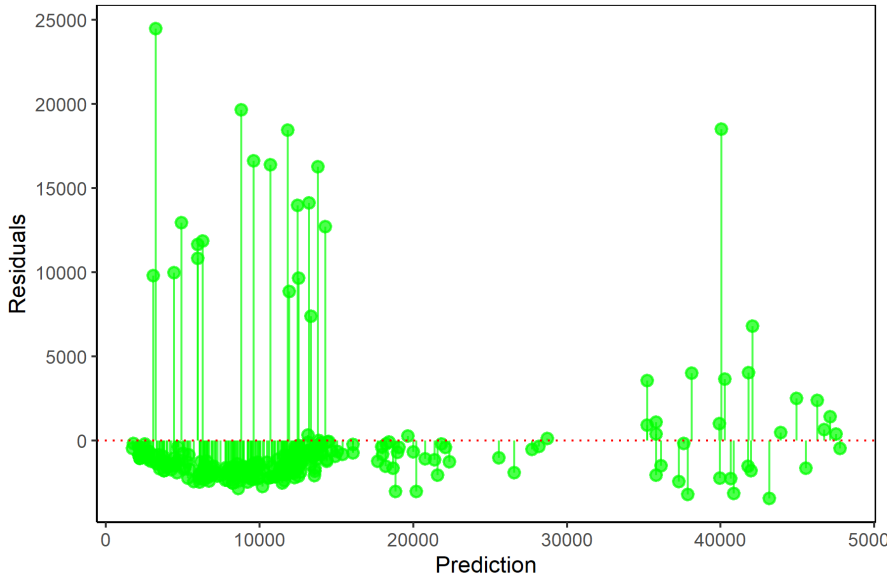
By testing a model, we can evaluate its accuracy and determine whether it can generalize to new data. This is important because a model that performs well on training data but poorly on new data is not useful in practice.

Plot 1: Prediction vs Actual Values



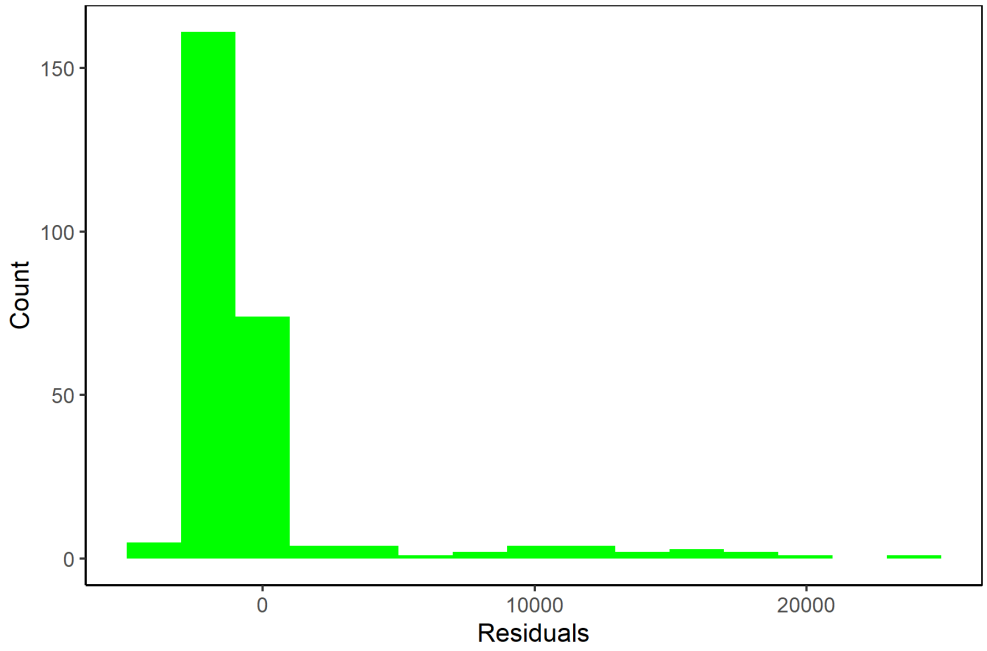
Looking at the prediction vs. actual values plot, we can see that the predicted charges are generally in line with the actual charges, as most of the points fall close to the red line representing perfect predictions.

Plot 2: Residuals vs Linear Model Prediction



The residuals vs. linear model prediction plot show 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

Variable	Estimate	Standard Error	t Value	Pr(>|t|)
Intercept	**-2072.60**	**588.50**	**-3.52**	**0.00**
Age	**263.02**	**9.99**	**26.32**	**0.00**
Obese	**416.31**	**459.85**	**0.91**	**0.37**
Overweight	**287.84**	**486.88**	**0.59**	**0.55**
Underweight	**-211.95**	**1067.02**	**-0.20**	**0.84**

Children	**513.96**	**116.10**	**4.43**	**0.00**
Smoker (Yes)	**12127.83**	**815.73**	**14.87**	**0.00**
Northwest	**-567.17**	**404.37**	**-1.40**	**0.16**
Southeast	**-884.40**	**398.15**	**-2.22**	**0.03**
Southwest	**-1204.33**	**399.98**	**-3.01**	**0.00**
Obese * Smoker (Yes)	**20919.97**	**942.79**	**22.19**	**0.00**
Overweight * Smoker (Yes)	**2405.20**	**1060.91**	**2.27**	**0.02**
Underweight * Smoker (Yes)	**1409.44**	**2059.28**	**0.68**	**0.49**

1. For every one-year increase in age, the charges increase by $263.02 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. For every additional child the charges increase by $513.96 on average.

4. Individuals who smoke pay charges on average $12,127.83 more than those who do not smoke

5. Individuals in the Southwest region pay lower charges compared to those in other regions.

6. Smokers in the Obese category pay the highest charges on average.

7. This model explains 86% of the variability in charges. The F-test value of 551.87 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, the BMI category alone does not have a significant impact on charges.

References

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