Exploring the Relationships between Insurance Charges and Multiple Factors: A Regression Analysis

Group number: E5

Group members:

Kevin Nguyen

Yutong Li

Zhengling Jiang

Daniel Hong

1. Introduction

1.1 Background

In this analysis, we will examine a dataset that includes detailed information about American insurance customers, such as age, sex, body mass index (BMI), number of children, smoking status and region. This dataset is recorded and obtained from Kaggle[1]. Insurance charges are a significant public health issue, and understanding how insurance charges are determined can help clarify some of the underlying drivers of these charges. Analyzing the impact of various factors such as BMI and smoking status on insurance charges can provide valuable insights into the insurance industry.

In addition, exploring the factors that impact the insurance charges and building an appropriate model to predict insurance charges will allow the insurance companies better understand the risks of insuring different types of customers, and to adjust their pricing strategies.

Overall, the results of this analysis could have significant implications for both the insurance customers and the insurance industry.

1.2 Question

What are the most important factors that impact the insurance charges, and how would we build an appropriate regression model to predict the insurance charges? Based on the dataset, we will investigate the impact of age, sex, body mass index (BMI), number of children, smoking status and region on the insurance charges.

2. Analysis

2.1 Data summary

The data set contains 1338 observations, and there are no null values. From the summary table, we can see the ‘Age’ variable ranges from 18 to 64 and with a mean of 39.21 and a median of 39. In the ‘Sex’ variable, there are 662

female examples and 676 male examples - there's no imbalance in either class. The ‘Children’ variable spans from 1 to 5 children with a version having an average of 1.095 kids and a median of 1 kid. In the ‘Smoker’ variable, there are 1064 non-smokers and 274 smokers - it might be important to note the under-representation of smokers in this dataset. In the ‘Region’ variable, there are 4 regions - the northeast, northwest, southeast and southwest, all of which are equally represented in the dataset. The ‘bmi’ variable ranges from 15.96 to 53.13 and has a mean of 30.66. The ‘Charges’ variable is the response variable.

	age	bmi	children	charges
Min	18.00	15.96	0	1122
1st Qu	27.00	26.30	0	4740
Median	39.00	30.40	1	9382
Mean	39.21	30.66	1.095	13270
3rd Qu	51.00	34.69	2.00	16640
Max	64.00	53.13	5.00	63770

Table 1: Summary table of continuous variables

sex	smoker	region
female: 662	no: 1064	northeast: 324
male: 676	yes: 274	northwest: 325
		southeast: 364
		southwest: 325

Table 2: Summary table of categorical variables

2.2 Data visualization

Firstly, we have used scatter plots to explore the relationship between continuous features and the response variable. The continuous variables are ‘Age’ and ‘BMI’, although ‘Children’ was continuous, modeling ‘Children’ as a categorical variable made more sense. Based on the scatter plots, it appears that there may be a linear relationship between ‘Age’ and ‘Charges’, as there is a

general trend of higher charges associated with older ages. However, for ‘BMI’, it’s hard to figure out any clear relationship.

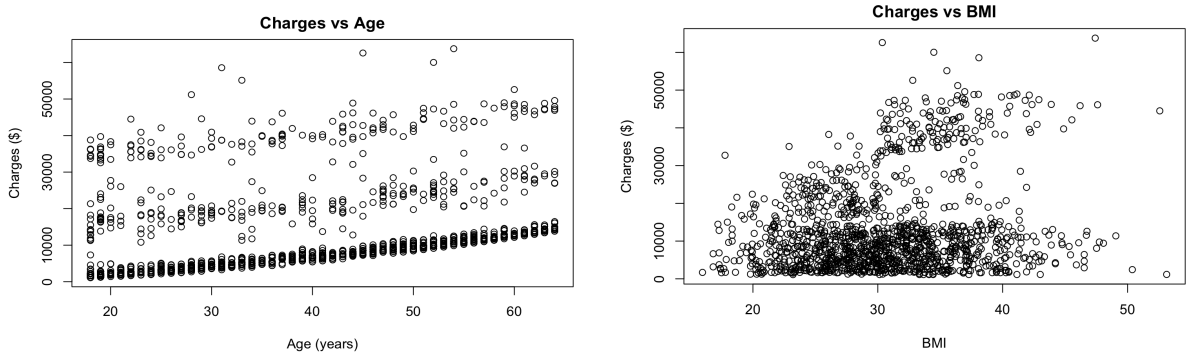

Figure 1, 2: Scatter plot of Charges by Age; Scatter plot of Charges by BMI

Secondly, we have used box plots to show the distribution of ‘Charges’ across different categories. Based on the boxplot, it appears that for two categories of ‘Smoker’ are significantly different, as the median charge for smoking status is much larger than the charge for non-smoking status. It suggests ‘Smoker’ may be a good predictor of ‘Charges’. For the rest categorical variables, the range and median of charges vary somewhat across categories, but we need to do more evaluations to ensure whether they are good predictors of charges.

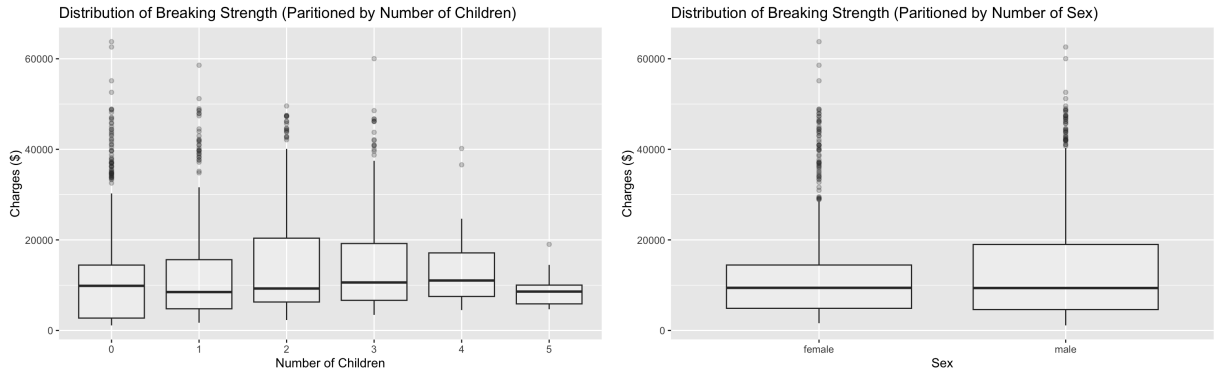

Figure 3, 4: Box plot of Charges partitioned by Number of Children; Box plot of Charges partitioned by Sex

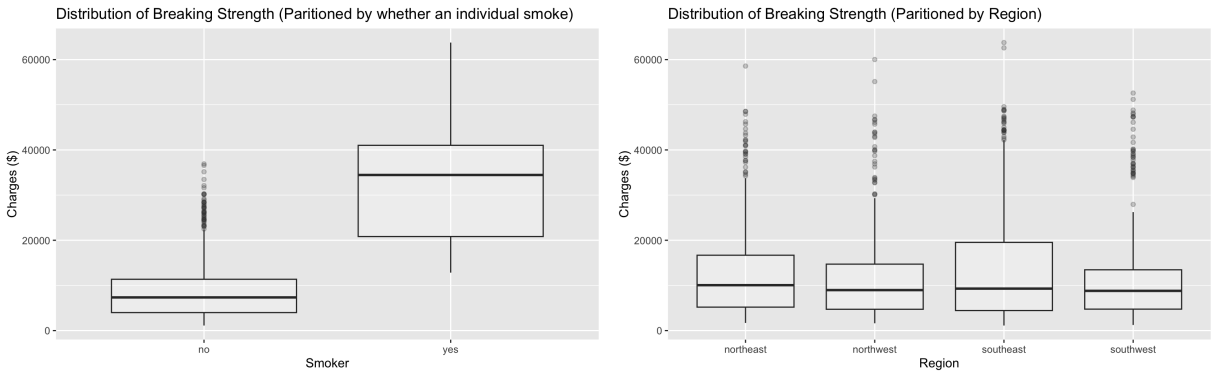

Figure 5, 6: Box plot of Charges partitioned by Smoker; Box plot of Charges partitioned by Region

Thirdly, we have used a heatmap to show the correlation between variables after performing one-hot encoding for all non-numeric variables. Based on the heatmap, it revealed that there is a strong positive correlation between charges and smoking status, which supports our previous conclusion. In addition, we found that both BMI and age have a moderate positive correlation with charges.

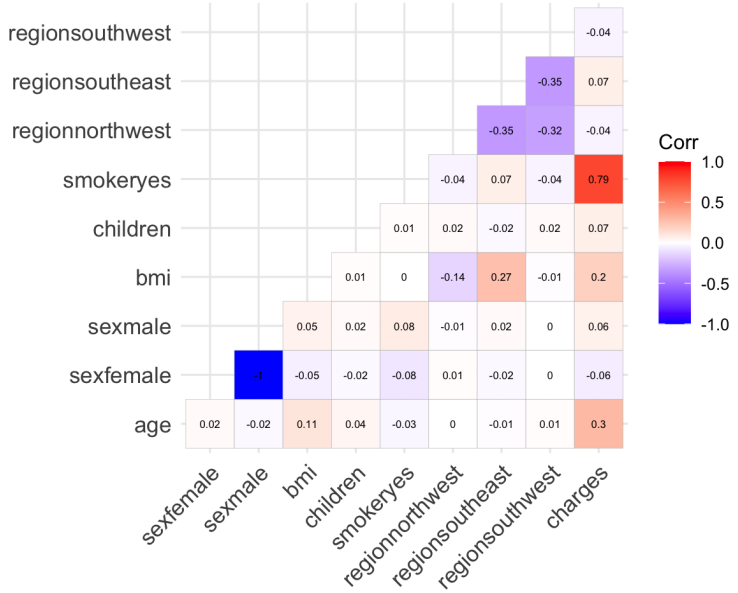

Figure 7: Correlation matrix heatmap for all variables

2.3 Model Selection and Evaluation

We see from the EDA that there appears to be a linear relationship between each age and smoking status with the insurance charge response variable. There does also seem to be a relationship between BMI and charges, but it is much less clear.

We used R to give us the best linear model for all linear combinations of variables using *regsubsets()*. From the result returned: we see that the “best” model for each number of parameters always includes age and smoker status, while sex is never included in any of the models.

For the following model evaluations, we will mostly be using AIC (Akaike Information Criterion) as an assessment of performance.

2.3.1 Age and Smoking Model

First, we tried fitting a linear model with just the age and smoking status and no transformations. This model yielded us an equation of *Charges = -2391.64 + 23855.30 * SmokerYes + 274.87 * Age.*

The Adjusted R-squared was 0.721, while the AIC (Akaike Information Criterion) was 27253.32. Both of the terms were significant, just as expected.

The interpretation of this model is that assuming all other factors are constant, a smoker would on average be charged $23855.3 more than non-smokers. Also, for every year increase in the age of the insuree, they are on average charged $274.87 more for their insurance.

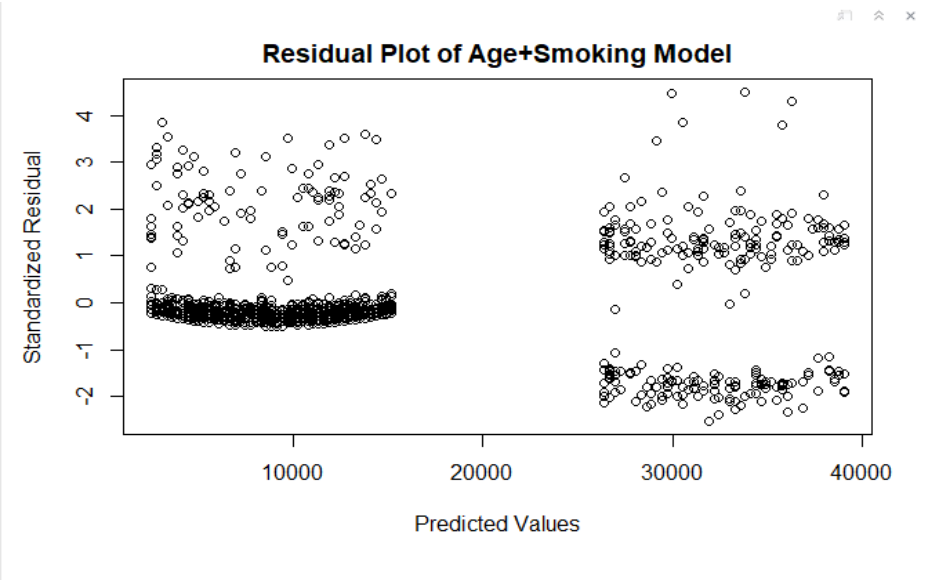

Figure 8: Residual plot of Age+Smoking model fitted

From the above residual plot, it actually seems that the prediction may not be as good as we imagined. There is a huge gap between 15000 and 27000 where there are no predicted values. Furthermore, we see that there does seem to be some sort of trend, in which the residuals trend downwards as the fitted values go up.

2.3.2 Model Using Age and Smoking and Their Interaction

We may suspect that there is an interaction between age and smoking due to the possible inadequacy of the previous model. Therefore, we fit a model the same as the previous one, but adding the interaction between age and smoking.

With this model, we find that the model was not made better, due to the AIC increasing, as well as the interaction term being insignificant. This suggests that maybe fitting another variable would be a good idea.

2.3.3 Model using Age, Smoking, and BMI

BMI is often seen as a good measure of a person's health and exercise status. The results from the *regsubsets()* also suggest that for a model with 3 explanatory variables, we should include the BMI variable. Thus, we have fitted a model including the BMI, smoking status, and age terms.

As expected, all three of the terms were incredibly significant, and led us to a model of *Charges = -11676.83 + 23823.68 * SmokerYes + 259.55 * Age + 322.62 * BMI.* This model gave us an adjusted R-squared of 0.7469, as well as an AIC of 27123.84, both significant improvements from the Smoking+Age model.

For the added BMI term, we interpret that for all other variables constant, for every one increase in BMI, the mean insurance charge is expected to increase by $322.62.

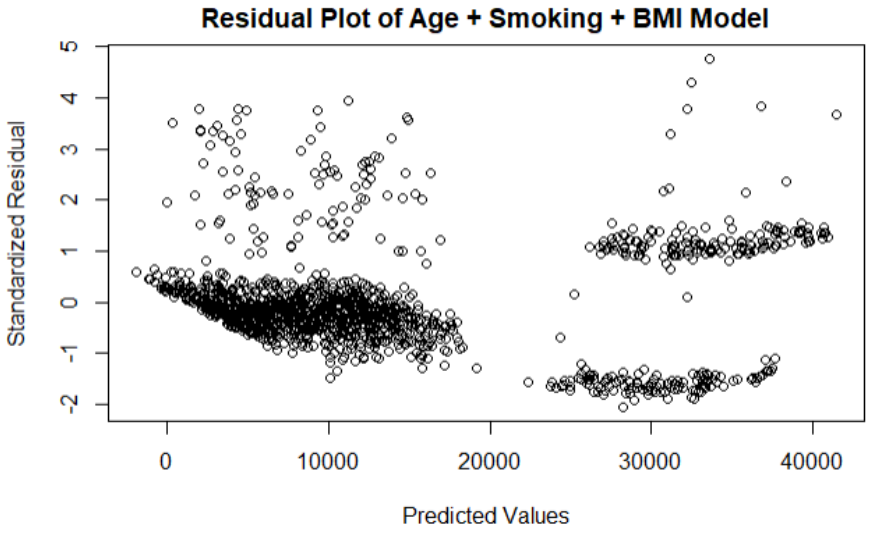

Figure 9: Residual plot of Age + Smoking + BMI Model

This residual plot seems better than the Age+Smoking residual plot, due to the fact that it covers most of the predicted values, not leaving as big a gap. Again, there is some sort of trend, but it is greatly improved over the Age+smoking model. The variance seems generally constant, so it is hard to decide on a transformation for the response variable.

2.3.4 Full Model

Suppose we believe that all of the variables in the data seem like they could affect charges, even though it is unclear how significant they are. We would fit a full model, using all of the variables we have in the data.

The fitted model yields us an equation of:

*Charges = -11927.17 + 23836.41 * SmokerYes + 257.19 * Age + 336.91*BMI - 128.17 * SexMale + 390.98 * OneChild + 1635.78 * TwoChildren + 964.34 * ThreeChildren + 2947.37 * FourChildren + 1116.04 * FiveChildren - 380.04 * RegionNW -1033.14 * RegionSE - 952.89 * RegionSW*

The AIC is 27118, and the adjusted R-squared is 0.7497.

However, many of the variables are not significant at the 5% level. These include the SexMale dummy variable, as well as some dummy variables for the region and children variables (OneChild, FiveChildren, and RegionNW).

We hypothesize that the region may not be significant enough to include, as from our exploratory data analysis, it seems that the region variables are highly correlated with the BMI variable.

This model is obviously inadequate because of the insignificant terms.

2.3.5 Reduced Model

Let us remove the sex and region variables from the full model and assess its performance.

The fitted model gives us an equation of
*Charges = -12093.32 + 23796.71 * SmokerYes + 258.08 * Age + 319.8 * BMI + 368.77 * OneChild + 1626.51 * TwoChildren + 996.95 * ThreeChildren + 2984.36 * FourChildren + 899.13 * FIveChildren*

This is the best model we have fitted so far, with an AIC of 27116.36, and an adjusted R-squared of 0.7393. All of the terms are significant at the 5% level save for OneChild and FiveChildren. However, it is hard to remove them from the model without removing the entire children variable.

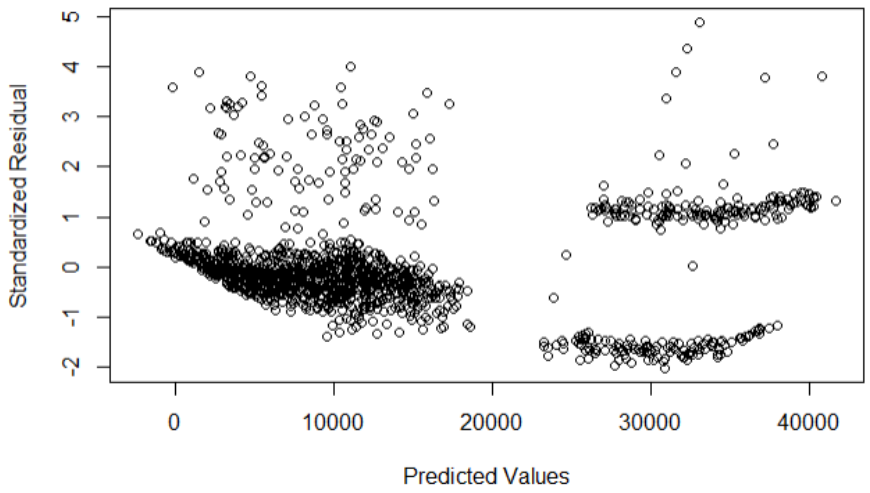

Figure 10: Residual plot of the reduced model

The residual plot looks slightly better than the Smoker+Age+BMI model.

2.3.6 Quadratic BMI Model

For our final model to evaluate, let us attempt to add a quadratic BMI term, as we may suspect a quadratic relationship between BMI and the charges, as shown by the exploratory data analysis.

The fitted model gives us an equation of

*Charges = -19152.163 + 23819.01 * SmokerYes + 256.56 * Age + 788.217 * BMI -7.422 * BMI^2 + 374.76 * OneChild + 1665.35 * TwoChildren + 975.53 * ThreeChildren + 2862.04 * FourChildren + 1005.89 * FIveChildren*

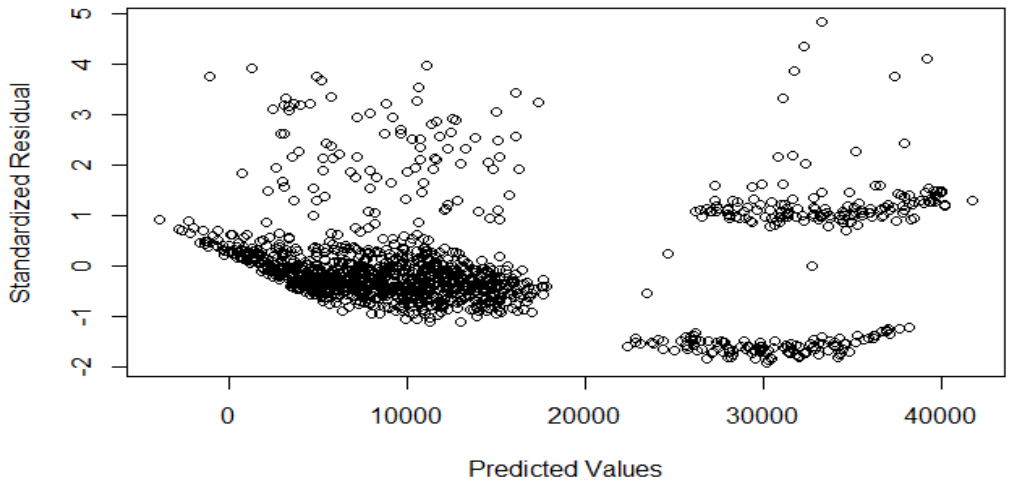

Figure 11: Residual plot of Quadratic BMI Model

The quadratic term is significant at the 5% level to indicate a possible quadratic relation. This also gives us the best AIC so far at 27113.2, as well as an adjusted R-squared of 0.75.

From all of the models we have fitted, we choose this model as the final model, as not only does it intuitively make sense, but it also has the best model statistics.

3. Conclusion

3.1 Findings

In our analysis of the influence of age, gender, body mass index (BMI), number of children, smoking status, and region on insurance charges, we aimed to identify the most significant factors and establish a suitable regression model for predicting insurance charges. The analysis revealed that age, smoking status, BMI, and the number of children are the critical factors affecting insurance fees. In contrast, gender and region were not significant predictors and were therefore excluded from the final model. Furthermore, we determined that incorporating a quadratic term for BMI was important, signifying a potential non-linear association between BMI and insurance fees, which enhanced the model's performance based on the Akaike Information Criterion (AIC).

3.2 Limitations

The final model, comprising age, smoking status, BMI, the quadratic term for BMI, and the number of children, exhibited the most reliable predictive performance, as indicated by the lowest AIC. Nonetheless, the study has limitations such as the dataset's representativeness, consideration of other potential factors influencing insurance fees, and the reliance on AIC as the main criterion for model selection, which might favour more complex models and result in overfitting.

3.3 Future Questions

Despite these limitations, the final model offers valuable insights into the factors influencing insurance fees. Insurance companies can use these findings to make well-informed decisions regarding their pricing strategies and gain a better understanding of the connections between these factors and insurance fees. We recommend exploring the impact of additional factors on insurance fees in future research, validating the model with an independent dataset, and testing more sophisticated machine learning techniques to potentially enhance predictive performance.

4. Appendix

4.1 Data set source

Devastator, T. (2023, January 7). *Prediction of insurance charges*. Kaggle. Retrieved April 10, 2023, from https://www.kaggle.com/datasets/thedevastator/prediction-of-insurance-charges-using-age-gender