Multiple Linear Regression Analysis on Medical Insurance Cost

Lili Balazs, Desiree Gonzales, Daniela Nunez, Nils Kessler

Executive Summary

The cost of medical insurance can vary based on many aspects that could determine how much an individual has to pay in order to have access to the benefits that medical insurance coverage provides. In this paper, our goal is to determine what the top aspects are that could influence the cost of medical insurance in addition to investigating whether an individual has to pay a premium price when they identify as a smoker, along with predicting medical insurance cost based on customer characteristics. In order to answer the above-mentioned questions, we use multiple linear regression. To validate our findings, we analyze the R2 and the adjusted R2 as well as conduct an F-Test and a T-Test. During our analysis, we also address the model assumptions for multiple linear regression.

Introduction

In this paper, the questions that we are aiming to answer are the following: Are we able to accurately predict medical insurance charges given customer information? What is the medical insurance cost premium if the customer is a smoker? Which customer characteristic/s has/have the most impact on medical insurance costs?

To answer the above-mentioned questions, we used a multiple linear regression model. We chose this model since we are working with various numeric and continuous response variables, thus using a multiple linear regression model that takes multiple values helps us to answer the proposed questions. The explanatory variables are also numeric or can be turned into dummy variables. With this chosen regression model we were able to investigate the overall predictive power of all the variables indicated in the data frame.

To answer the first question, we needed to analyze the R2 and the adjusted R2 and conduct an F-Test. We chose to analyze the R2 and the adjusted R2 since these were able to tell us how much of the variations can be described with our chosen model and indicate how well our prediction regarding the medical insurance cost is. In addition, the F-Test showed results regarding the explanation power of our variables.

To address the second question, we analyzed the magnitude and the statistical significance of the coefficient of the explanatory variable of smoking. We determined the magnitude based on the analysis of the coefficient value, while we also conducted a T-Test to learn about the significance of said variables.

For the third question, we decided to analyze the rest of the explanatory variables in the same way that we addressed the second question. Additionally, we looked for the variable with the highest (relative)

magnitude and the biggest significance, while also considering the variables with different ranges in values.

In addition to answering these questions, our goal was to address the model assumptions for multiple linear regression to the best of our ability.

Exploratory Data Analysis

The dataset used for the analysis includes 1338 observations with 7 variables, 6 of which are predictors whose main effects can be observed in the response variable, insurance charges. The variables are comprised of various data types: age, BMI, charges, and [number of] children are numeric, while sex, region, and smoker are all of type character. In order to include the categorical variables (sex, region, and smoking status) in exploratory analyses, one-hot encoding was used to change the data type from non-numeric to numeric. There were no missing values, or NA's, within the dataset.

Histograms of each variable (refer to Figure 1) revealed that the distribution of the response variable, charges, was positively skewed, while the predictors produced various types of distributions. In order to alleviate the skewness of the response variable, a new variable was created consisting of the log of the values in the charges column. However, it is relevant to note that this new variable was not included in the final linear model, and that the untransformed response variable was used instead. Additionally, the BMI variable was standardized in order to achieve accuracy of descriptive statistics.

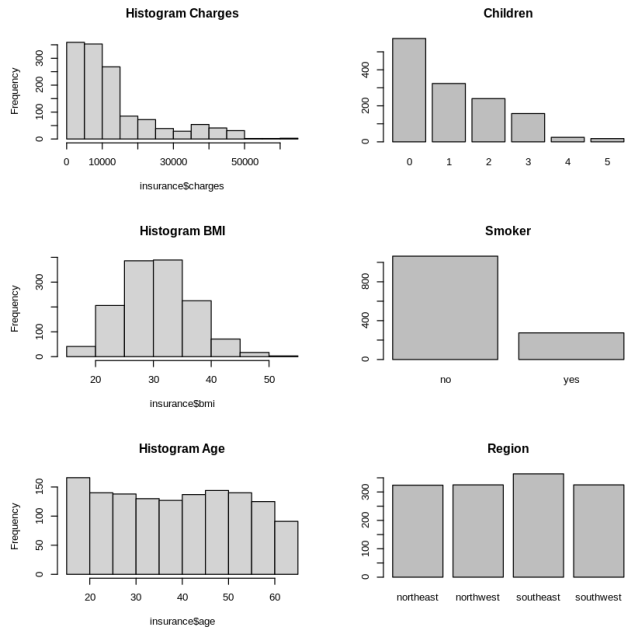

Figure 1: Histogram matrix of variables

The predictor variables comprised several types of customer attributes. For example, while age and sex are inherently uncontrollable, the majority of the predictor variables are largely associated with personal lifestyle choices. Exploratory data visualization revealed that BMI, age, and smoking status were all linked to higher cost of insurance, whereas other predictors appeared to have less observable effects (Figure 2).

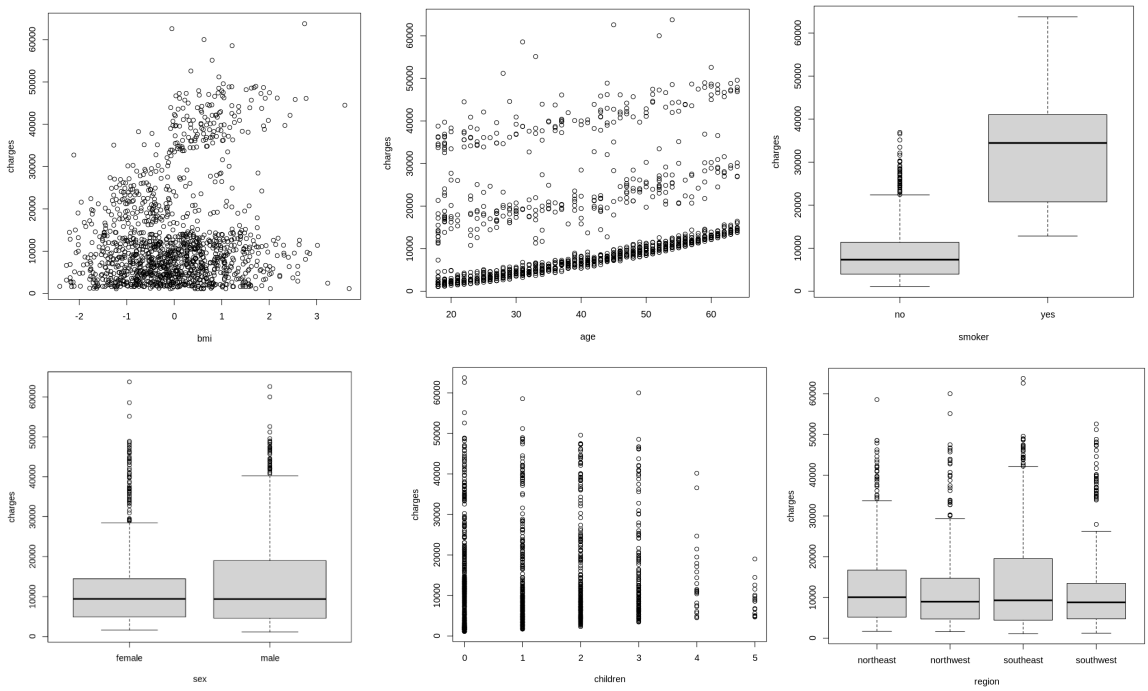

Figure 2: Predictors plotted against response variable

Because the analysis and research questions are largely focused on using customer attribute data to estimate resulting insurance charges, all predictors were included in the initial models. The specific interactions between the predictors and the response variable will be further investigated in the analysis section.

Statistical Analyses

The main objective of this analysis is to investigate how different lifestyle factors affect the cost of insurance coverage. Therefore, a multiple linear regression was performed in order to most accurately determine the effects of the predictor variables on insurance charges.

The initial linear model included all 6 predictor variables, while using the transformed log_charges column as the response variable. This resulted in a decent R-squared value of 0.77, and the model summary indicated significance across several predictors: age, BMI, number of children, and smoking status all had significant p-values. However, upon making a Q-Q plot and residual vs. fitted plot of the model, it was clear that several linear regression assumptions had been violated. The Q-Q plot indicated

non-normality of residuals, while the residual plot was highly heteroskedastic, indicating non-constant variance. As a result, some adjustments were made to the initial model.

The adjusted linear model introduced a key aspect that increased model performance: an interaction term. While sex was not included as a predictor due to insignificant p-values, the model focused on the variables that demonstrated increased effect on the response. Special attention was paid to these variables, BMI and smoking status specifically, not only because of their statistical significance, but because of their relevance to the research questions at hand. It was thought that the best way to observe the effects of high BMI and positive smoking status was to include an interaction term of smoking status * BMI, effectively grouping these variables and allowing their combined effects to be studied more closely. This model, similar to the previous, used the transformed log_charges column as the response, but added an exponential transformation to the individual BMI variable. Upon investigating the summary of this first interaction model, it was clear that the interaction term had been effective: the R-squared value increased to 0.78. Nevertheless, the Q-Q and residual plot for this model, despite showing some improvement, still suggested that the assumptions of normality and constant variance were violated. Further adjustments were necessary to achieve the strongest linear model.

The final linear model took a somewhat different approach, using the raw, un-transformed charges variable as the response, while keeping the interaction variable smoking status * BMI intact alongside the other predictors age, BMI, children, smoker, and region, respectively. These adjustments resulted in a considerable 6% increase in R-squared value, which turned out to be 0.84. Essentially, the predictors included in this model accounted for 84% of the variability in insurance charges.

Checking the final model for violation of assumptions revealed an improvement in normality of residuals, despite the presence of several outliers whose values were checked individually; this is addressed more clearly in the Appendix section. However, the residuals vs fitted plot did remain somewhat heteroskedastic. With both t and F statistics being invalidated by heteroskedasticity, alternate methods of statistical analysis were used for testing on the final model. The standard F test was replaced by the Wald F test, whose detection of overall significance of a model remains unaffected by non-constant variances. Similarly, a robust t-test was performed to alleviate effects of outliers and heteroskedasticity. Both the Wald F and robust t were used due to their imperviousness to potential violation of assumptions.

The resulting significance values assisted in answering the third research question, which aims to detect the customer attributes that have the greatest impact on insurance costs. The Wald F test of overall significance provided a value of <0.001, indicating significance of the predictors and goodness of fit of the final model. The robust t-test supported this notion, with several predictors showing significance (Figure 3). Overall, these significance values and the high R squared value of 0.8405 demonstrate that the selected predictor variables explain a majority of the variance of insurance costs.

Coefficients	Values	p-Value (robust t-test)
Intercept	-1760	<0.001***
Age	264	<0.001***
BMI	138	0.378
Children	512	<0.001***
Smoking	23788	<0.001***
Region: North-west	-581	0.127
Region: South-east	-1207	0.002**
Region: South-west	-1228	0.001**
BMI * Smoking	8770	<0.001***

Figure 3: Significance values from robust t-test

Answering research questions and suggesting further research

Research question 1: Are we able to accurately predict medical insurance costs given customer information?

To answer the research question above we used our final model that was a very suitable model described in the section before. The high R^2 value of 0.84 and the very low Wald test p-value show that our model has predictive power. To further investigate this, we also added a 70/30 ratio train test split and used our model to predict unseen test samples, after being fitted on the train data.

Figure 4 shows the actual values vs the values that we predicted with our model. We can see that we are able to predict the values quite well, especially for values below $20,000. With increasing insurance costs, the residuals (vertical distance from black points to green line) increase on average, and appear to have a higher variance.

To answer the research question, we find our model to have good predictive power overall and good accuracy. However, and especially for predicted charges that are higher than $20,000, we find our model to deliver predictions with varying performance.

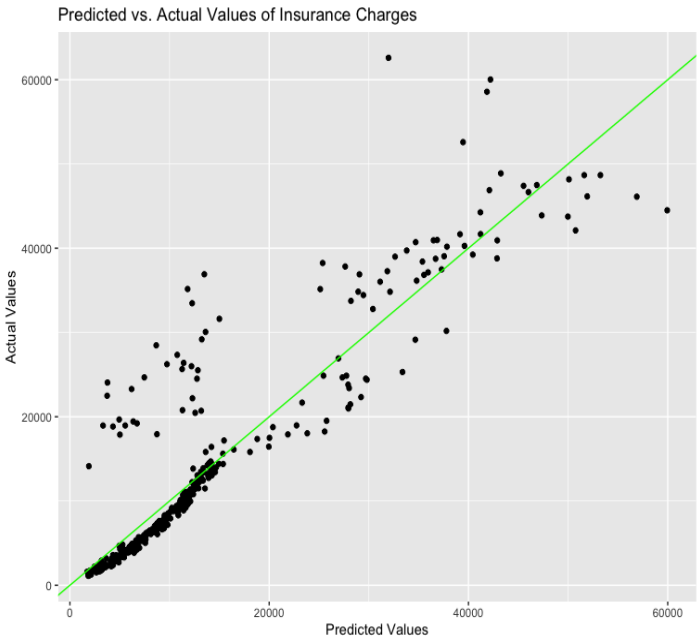

Figure 4: Predicted vs Actual Values of Insurance Charges

Research question 2: Does smoking status significantly affect insurance costs?

To answer this question we can simply analyze the coefficient value of the smoking variable and the p-value of the robust t-test. The coefficient value is 23,788. Interpreting the coefficient, yields that according to our model, the fact that someone is a smoker, increases the medical insurance costs by nearly $24,000. This is a really high value, indicating that smoking is a very unhealthy habit that can lead to diseases that are very costly to cure or to treat. With a p-value of less than 0.001, the coefficient was highly significantly different from 0. To conclude, it is reasonable to infer that smoking status does significantly affect insurance costs.

Research question 3: Which customer characteristics have the most impact on medical insurance costs?

Looking at the other coefficient values and t-test p-values, we can answer the final research question. While we already identified the smoking habit to be an important variable, the age and the interaction term were variables with much importance as well. Both variables being highly significant according to their t-test p-values, the age coefficient shows that aging one decade will increase someone's insurance costs by around 2640$. The coefficient of the interaction term shows that while already having a smoking habit, increasing the BMI by one standard deviation, will increase insurance cost by $8,770.

Besides these most important variables due to high magnitude and significance and for the sake of completeness the other variables can be analyzed as well. Adding a child to the insurance plan adds $512 to the costs. Also, compared to insurance costs in the northern part of the US, the insurance costs in the south are a bit cheaper, which is indicated with the significant coefficients of the southern regions. While

the interaction term of the smoking variable and the BMI was very important, the BMI variable alone is not significantly different from zero and has a low magnitude.

Further research on this topic could be done to analyze where exactly these costs come from. For example it would be interesting to know which diseases (obese) smoking people get more frequently due to their lifestyle, compared to the average person and what treatments of these deseases will cost. The described analysis will enlighten our analysis and deliver more arguments to underline our research findings. Two concrete questions could be: “Which diseases are smoking people more likely to get that soar medical insurance cost?” and “Which possible treatments cost the most and explain why smoking customers pay a premium for their medical insurance?”

Conclusions

Using multivariate regression as a statistical tool to answer our research questions proved to be very effective. With an R^2 and adjusted R^2 of around 0.84 our model was able to explain a great amount of the variation of the medical insurance charges.

With our final model we were able to answer all of our research questions. By conducting the train test split, we tested the predictability of the medical charges with our model and found that overall our predicted values are close to the actual values in the test set, while we observed that with increasing value of the predicted value, the variance of the residuals increased and predictions became more uncertain. Furthermore, low p-values for the Wald test indicate that the variables in our model have predictive power and with our adjusted R^2 to be 0.84 we found overall that our model was able to predict the charges of medical insurance with good accuracy.

The coefficient of the smoking variable was highly significant and indicated that the habit of smoking increased the medical insurance cost by almost $24,000. Therefore we found the smoking habit to be a variable that significantly affects the insurance costs.

Finally, to answer the last research question that looked for further significant variables we found the age and the interaction term of smoking habit and BMI to be the most important ones, due to the low p-values and the relatively high magnitudes of the coefficients. Aging one decade will increase the costs by around $2,600 while having a BMI that is one standard deviation higher than the average additionally to being a smoker will add another $8,770 to your bill according to our model.

With our project three key guidelines for every person in the US can be derived that are relevant with regard to medical insurance and its costs. First, do not start smoking or try to quit the unhealthy smoking habit, because it will cost you a lot of money for your insurance. Secondly, if you happen to smoke, try to stay in shape, as increasing weight will increase your insurance bill further. Finally, be aware of the fact that with increasing age, your insurance bill will increase as well.

References

Astivia, Oscar L. Olvera and Zumbo, Bruno D. (2019) "Heteroskedasticity in Multiple Regression Analysis: What it is, How to Detect it and How to Solve it with Applications in R and SPSS," Practical Assessment, Research, and Evaluation: Vol. 24 , Article 1. DOI: https://doi.org/10.7275/q5xr-fr95

Hothorn, Torsten, and Brian Everitt. "A Handbook of Statistical Analyses Using R (3rd Edition)." *CRAN*, Comprehensive R Archive Network (CRAN), https://cran.r-project.org/web/packages/HSAUR3/.

Wooldridge, Jeffrey et al. *Introductory Econometrics: A Modern Approach*. South-Western Publishing Co, 2nd, 2003.

Appendix

As previously mentioned in the analyses section, several outliers were present in the data and remained present throughout the process of selecting a final linear regression model. Initially, it was necessary to investigate these specific outliers to determine whether or not they should be included or removed. A plot of Cook's distances was made; subsequently, the outliers were investigated.

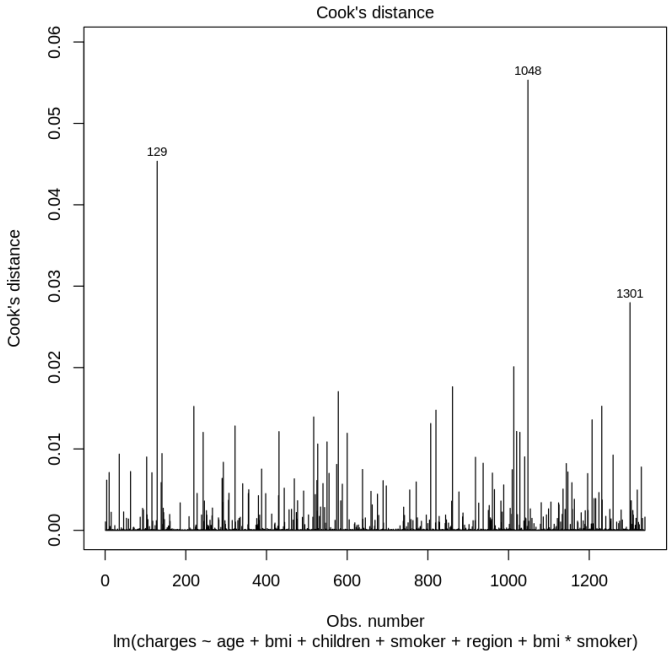

Figure 5: Cook's distances

According to the plot, observations 129, 1048, and 1301 are outliers with a combination of unusual explanatory and response variables. Examination of observation 129 revealed that the individual was a 32 year old female smoker whose BMI indicated almost severe thinness, with insurance charges more than double the average amount. We decided to keep this observation because extreme cases do exist when it comes to personal attributes. Not to mention, this individual is listed as having 2 children, and children are included in overall insurance charges, which could also be a contributing factor to their high insurance premium. As proven by our analysis, her status as a smoker also undoubtedly contributes to the high charges as well. Observation 1048 was a 22 year old male with a very high BMI indicating severe obesity; therefore, it was unsurprising that his charges were exceedingly high. We did not exclude this observation because the high charges lined up with the high BMI, and it is important to include extreme cases, as similar situations may be a reality for some customers. Finally, observation 1301 was an overweight, middle-aged smoker, which likely accounts for his high insurance premium. We did not decide to exclude this observation, either. Had any of these outliers possessed personal attributes that did not line up with inflated charges, removal would have been necessary. After inspection, however, it was determined that removal was not necessary for these three values.

It is also important to address potential collinearity present in the final model. According to the correlation matrix below (Figure 6), there was not much collinearity present among predictors besides active smokers and BMI. This was addressed through the creation of the interaction variable that grouped BMI and smoking status.

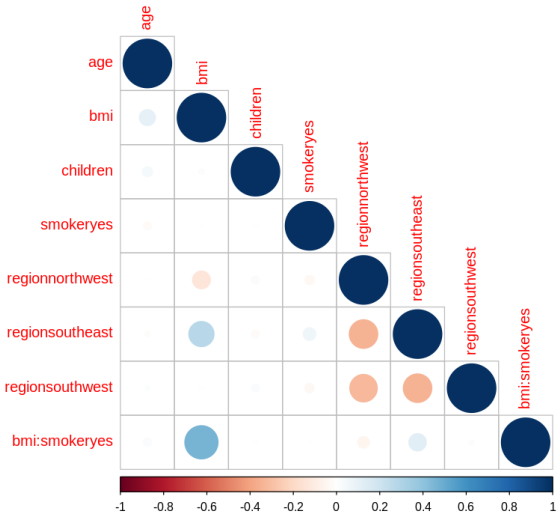

Figure 6: Correlation Matrix of final model predictors