**A Predictive Report on the Insurance Charges Based on Individual Characteristics**

Yagya Shekhar Vyas (2321197)

University Canada West

Business Analytics

BUSI 650 (HBD-SPRING24-06)

Prof. Amirhossein Zaji

May 26th, 2024

# **Introduction**

Insurance companies consider different factors to predict charges accurately to effectively manage risk, set appropriate premiums, and allocate their resources. Different factors like age, BMI, smoker/non-smoker, and geographic location are possible factors that influence the insurance charges. The relationship between the factors can be determined by using multiple regression analysis to explore the relationships between factors and insurance charges to enable better prediction and an understanding of dynamics. In this study, I conduct a predictive modeling of the insurance charges using piece-wise linear regression and multiple regression to develop a regression model that can accurately estimate insurance charges based on statistically significant predictors.

# **Background**

The importance of insurance companies in the modern economies globally cannot be overemphasized. They have proved their importance in safeguarding against various risks including health risks. Individuals pay premiums or insurance charges in exchange they get a cover and payment in case of any risk occurrence. The process of determining insurance charges is complex and is influenced by various demographic characteristics, health status, lifestyle choices, and geographic location. Different predictive model techniques are applied to offer understanding and determine insurance charges based on the historical patterns of the factors.

# **Literature Review**

Predictive models for evaluating insurance premiums are common and have been applied by many researchers. States that regression is used in prediction, classification, and explanation. Palmer and O’Connell (2009) use regression to predict cardiorespiratory status in a study of cardiorespiratory fitness. Kaushik et al., (2022) explore the use of AI and Machine learning in determining insurance premiums. The AI used regression based model to predict the health insurance cost that is incurred by individuals based on their features.

# **Methodology**

The dataset consists information on insured individuals that include age, sex, BMI, number of children, smoking status, region, and insurance charges. The data is first cleaned to remove data that is missing important information. The methodology involves conducting a descriptive analysis of the dataset to understand distribution of used variables and identify factors that contribute to high insurance charges. In descriptive analysis, there is summary statistics, frequency distributions, and visualizations. After carrying out descriptive analysis, the data was divided into two, training, and testing. Excel function that allocates random numbers to the data to ensure no bias was used to divide the data into two. There is predictive modeling using the linear regression that forecasts insurance charges based on the individual characteristics. The predictive regression modelling was trained on the training data which is 80% of the data or 1080 data sets and tested on the testing data that is 20% or 258 data sets.

# **Results and Analysis**

## **Descriptive Analysis**

The following summary indicates the summary statistics for the numerical variables. The descriptive statistics show the mean, median, minimum, maximum, and standard deviation for numerical variables that include BMI, charges, and age.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| ***Sex*** |  | ***Bmi*** |  | ***children*** |  | ***charges*** |  |
|  |  |  |  |  |  |  |  |
| **Mean** | 0.494768 | Mean | 30.6634 | Mean | 1.094918 | Mean | 13270.42 |
| **Standard Error** | 0.013674 | Standard Error | 0.166714 | Standard Error | 0.032956 | Standard Error | 331.0675 |
| **Median** | 0 | Median | 30.4 | Median | 1 | Median | 9382.033 |
| **Mode** | 0 | Mode | 32.3 | Mode | 0 | Mode | 1639.563 |
| **Standard Deviation** | 0.50016 | Standard Deviation | 6.098187 | Standard Deviation | 1.205493 | Standard Deviation | 12110.01 |
| **Sample Variance** | 0.25016 | Sample Variance | 37.18788 | Sample Variance | 1.453213 | Sample Variance | 1.47E+08 |
| **Kurtosis** | -2.00256 | Kurtosis | -0.05073 | Kurtosis | 0.202454 | Kurtosis | 1.606299 |
| **Skewness** | 0.020951 | Skewness | 0.284047 | Skewness | 0.93838 | Skewness | 1.51588 |
| **Range** | 1 | Range | 37.17 | Range | 5 | Range | 62648.55 |
| **Minimum** | 0 | Minimum | 15.96 | Minimum | 0 | Minimum | 1121.874 |
| **Maximum** | 1 | Maximum | 53.13 | Maximum | 5 | Maximum | 63770.43 |
| **Sum** | 662 | Sum | 41027.63 | Sum | 1465 | Sum | 17755825 |
| **Count** | 1338 | Count | 1338 | Count | 1338 | Count | 1338 |

Table 1

### **Scatter Plots**

Scatter Plots show a visual representation of the data. The following distributions show the distribution of the BMI and the age data respectively in relation to charges. It is clear that there is no significant relationship detected from the BMI distribution while age seems to have a positive relation and as age increases, the charges increase as well.

Figure 1

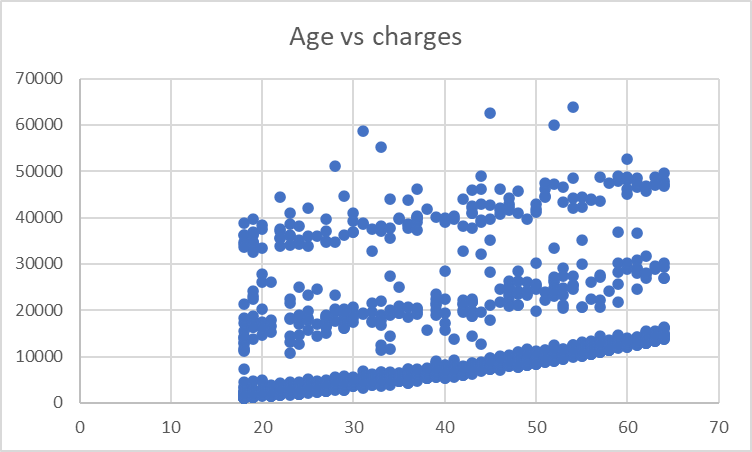


Figure 2

### **Categorical Data**

Analysis of the categorical data was done using pivot chart analysis.

Figure 3

The chart analyses the sum of the charges of the observations based on the region, sex, and whether smokers or not. Based on the analysis, males that are smokers and are from southeast had significantly high sum of charges. Females non-smokers from southeast had the least sum of charges.

### **Correlation**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | *age* | *Sex* | *bmi* | *children* | *Smoker* | *Region 1* | *Region 2* | *Region 3* | *charges* |
| age | 1 |  |  |  |  |  |  |  |  |
| Sex | 0.020856 | 1 |  |  |  |  |  |  |  |
| bmi | 0.109272 | -0.04637 | 1 |  |  |  |  |  |  |
| children | 0.042469 | -0.01716 | 0.012759 | 1 |  |  |  |  |  |
| Smoker | -0.02502 | -0.07618 | 0.00375 | 0.007673 | 1 |  |  |  |  |
| Region 1 | 0.002475 | 0.002425 | -0.13816 | -0.02281 | 0.002811 | 1 |  |  |  |
| Region 2 | 0.010016 | 0.004184 | -0.00621 | 0.021914 | -0.03695 | -0.32018 | 1 |  |  |
| Region 3 | -0.01164 | -0.01712 | 0.270025 | -0.02307 | 0.068498 | -0.34556 | -0.34626 | 1 |  |
| charges | 0.299008 | -0.05729 | 0.198341 | 0.067998 | 0.787251 | 0.006349 | -0.04321 | 0.073982 | 1 |

Table 2

Based on the correlation tests carried out on the data, there is a strong relationship between charges and smokers, a correlation coefficient of 0.7873. Meaning that more charges are associated to smokers as opposed to non-smokers. There is also a positive correlation between charges and children, BMI and age indicating that more charges are associated with increase in age, increase in the number of children and a high BMI. However, the correlation coefficients are low and would require more analysis to determine the relationship.

## Predictive Analysis

The regression analysis of the BMI and charges does not appear to be statistically significant as the P-value is 0.1005 which is more than the significance level of 0.05. The R Square and Adjusted R Square are low a 0.0025 and 0.0016 respectively.

The regression analysis of the number of children indicates a low multiple R, R square and Adjusted R square indicating a low positive correlation. The p-value of 0.02 shows that the number of children is statistically significant at 0.05 level of significance.

Multiple R of 0.311 in the regression analysis of age versus the charges suggest a moderate positive correlation. The p-value of 1.13\*10^-25 which is significantly less than 0.05 shows that the age data is significant and the increase in age causes an increase in charges.

Smoker variable indicates the strongest relationship of the variable and the charges. There is a strong positive correlation indicated by the Multiple R of 0.789. The coefficient of determination indicates that about 62.2% of the variance in charges is explained by smokers. There is a P-value of 7.5629\*10^230 which indicates the relationship between the smoker and charges is significantly significant at 0.05 level of significance.

In the regions, the reference category was Northwest and all regions were compared to Northwest region. Based on the analysis, the insurance charges in Northeast are higher that Northwest charges. However, p-value indicates that there is no strong evidence to conclude that the Northeast charges are higher than those in Northwest.

The coefficient for southwest is a negative number indicating that the average insurance charges for the southwest region are lower than the northwest region. A p-value of 0.046 is less than the level of significant indicating there is evidence that the southwest charges are lower than the northwest charges.

Southeast region has a coefficient of 2380 indicating that on average, the insurance charges in Southeast are higher than Northwest region. The p-value of 0.005 indicates there is evidence to conclude that charges in Southeast are significantly higher than the charges in Northwest region.

Based on the analysis of the region, there is evidence that shows that the average insurance charges rank, Southwest, Northwest, then Southeast from lowest to highest. There is no evidence to rank Northeast charges in relation to Northwest insurance charges.

### Multiple Regression

Multiple Regression was conducted on the variables, age, smoker, and children. The multiple R statistic is 0.8535, R Squared 0.7285, and a standard error of 6379.

A multiple regression equation from the modeling is:

Using the predictive modeling equation, the predictive charges were calculated from the testing data.

# Conclusion

In conclusion, I was able to identify statistically significant predictors that include age, number of children, smoker status, and geographic region that contribute to various insurance charges that enable making informed predictions. A change in any of the above factors leads to a significant change in charges. A well-fitted regression model was developed to achieve a better understanding of the relationships between variables and insurance charges. The predictive model showed that an increase in one child leads to an additional $575 in charges, smokers pay $24070 more and, increase in age each year is associated with $276 increase in charges.

## Implications

Insurance companies can set premiums, assess risks and optimize operations by considering predictive variables.

### Future Considerations

Future research in insurance analytics and provide framework for exploring additional factors.

Consider additional factors and refine the predictive models.

# References

Kaushik, K., Bhardwaj, A., Dwivedi, A. D., & Singh, R. (2022). Machine learning-based regression framework to predict health insurance premiums. *International Journal of Environmental Research and Public Health*, *19*(13), 7898.

Palmer, P. B., & O'Connell, D. G. (2009). Research corner: regression analysis for prediction: Understanding the process. *Cardiopulmonary physical therapy journal*, *20*(3), 23-26.