**Statistical Analysis and Forecasting of Medical Cost Using Regression**

**Yogesh R. Tajave**

**University of Texas at Dallas**

**BUAN 6359.502: Advanced Statistics for Data Science**

**Dr Monica Brussolo**

**Statistical Analysis and Forecasting of Medical Cost Using Regression**

**Abstract:** Medical expense is one of the unexpected & substantial expenses in human life that generally occur due to a disease. A variety of factors determines the disease one can have that include healthy/unhealthy lifestyle, genetic makeup, gender, age, and mental health. According to several studies age, sedentary lifestyles, unhealthy eating, smoking, alcohol, and drugs are the leading cause behind the exorbitant medical expenses. In this work, the estimation of personal medical expenses based on various factors is studied statistically. The data is from the GitHub website that contains medical expenses for patients in the USA. Statistical methods like t-test, ANOVA, correlation, multiple linear regression is employed to study the association of personal medical expense with age, BMI, gender smoker/non-smoker, number of children, and region.

**Introduction**

The world has seen an exponential growth in medical expenses that has become a point of concern. This unexpected direct health care expense and related expenses often lead to a medical debt. A 2007 survey had found about 72 million (41%) Americans between the age of 19 to 64 either have difficulty paying for medical treatment or have medical debt (Heavey, 2008). This huge health expense has been majorly attributed to aging and lifestyle disease (An, 2015).

The diseases associated with the way people live their lives are known as a Lifestyle disease. These are commonly caused by sedentary lifestyles, unhealthy eating, smoking, alcohol, drugs (Wikipedia, n.d.). An unhealthy lifestyle is a leading cause of numerous diseases and premature deaths. An important factor associated with the medical expense is smoking. Smoking can cause cancer, CVD, stroke, lung disease, diabetes, and problems of the immune system. The health expense is strongly age-dependent. As per Bradford and Max, annual personal health expenditure for the elderly is approx. four to five times that of people in their early teens (Bradford & Max, 1996) indicating a strong positive relationship between age and health expenditure.

Obesity can be harmful and lead to many diseases. The Body Mass Index (BMI) is a prevalent way of getting the sense of how overweight or underweight an individual is relative to their height. A low value of BMI indicates underweight while a high value indicates overweight. It is common knowledge that excess weight increases the risk of developing conditions such as diabetes, CVD, and reduces life span.

This study aims to describe the relationship between the several predictor variables, and medical expense. Through this study, we will try to answer questions like “Does smoking cigarette affect the medical expense?”, “Does geographical location affect the medical expense?”, “Does medical expense is different for male smoker and female smoker?”. This can be helpful in determining how medical expenses differ in people with different age groups, BMI, smoker, and across different regions.

The paper is divided into four sections – 1. Dataset Description & Exploratory analysis, 2. Hypothesis Testing & Model creation, 3. Results and Findings, and 4. Conclusion and Future Improvement.

1. **Dataset Description & Exploratory Analysis**
2. **Dataset Description**

The dataset used here is obtained from the GitHub website. The data contains medical expenses for patients in the USA. The dataset has 1338 observations and seven variables. The description of variables used in this dataset is as follows

Table 1: Dataset description

|  |  |
| --- | --- |
| **Variables** | **Description** |
| **age** | An integer variable indicating the age of an individual. |
| **sex** | The individual’s gender, either male or female. |
| **bmi** | BMI stands for Body Mass Index. This is an integer variable that is commonly used to categorize a person as underweight, normal weight, overweight, or obese. |
| **children** | This is an integer variable indicating the number of children covered by the insurance. |
| **smoker** | This is a categorical variable indicating whether they smoke or not. |
| **region** | This is a categorical variable indicating the region of an individual in the USA. Four regions - northeast, southeast, southwest, or northwest. |
| **charges** | This is an integer variable indicating the medical cost charged for a year. |

The summary of the dataset is shown in figure 1. We can see that there are no missing values in the dataset.



Figure 1: Summary of Dataset

1. **Data Exploration**
2. ***Univariate data analysis:***

Here we are visually exploring each variable to understand the various characteristic and the data spread.

Only the density of the dependent variable i.e. charges (Figure 2-A) clearly showed that the data is heavily right-skewed (skewness: 1.51) due to the presence of outliers and departing from the normality assumption. We have used log transformation on the dependent variable (Figure 2-B) to ensure adherence to approx. normality assumption (skewness: -0.09).

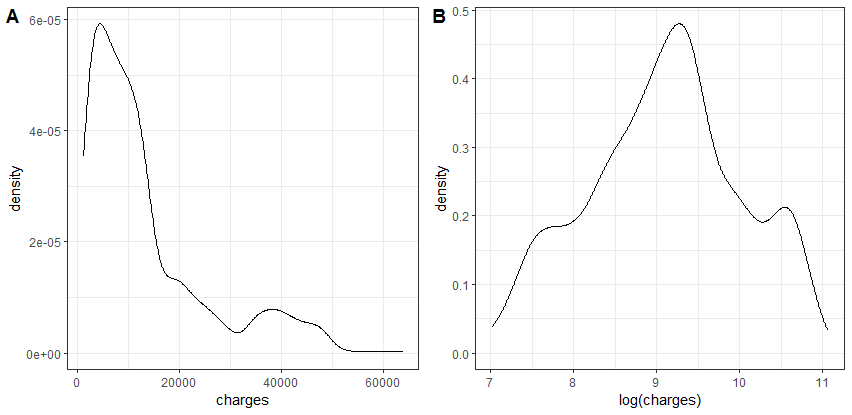


Figure 2: Density plot of charges, log(charges) [response variable]

1. ***Exploring data with respect to dependent variable***
2. **Numerical independent variable:**

Here we are visually exploring the numerical independent variable and how these variables affect the charges.

Figure 3-A clearly shows with the increase in age the charges also increase indicating a strong positive relationship. Furthermore, there are three distinct clusters formed within the graph, two out of three clusters have consistently higher charges that could be explained by person’s smoking habit and high bmi.

In figure 3-B, though it shows two distinct clusters, it is difficult to discern trends in the plot. The smoking habit may be able to explain the two clusters and the consistent high medical charges for one of the clusters.

Figure 3-C shows that individuals with four and five children are lesser individuals than those with zero to three children. Maximum variation in medical expense can be seen in individual with zero children.

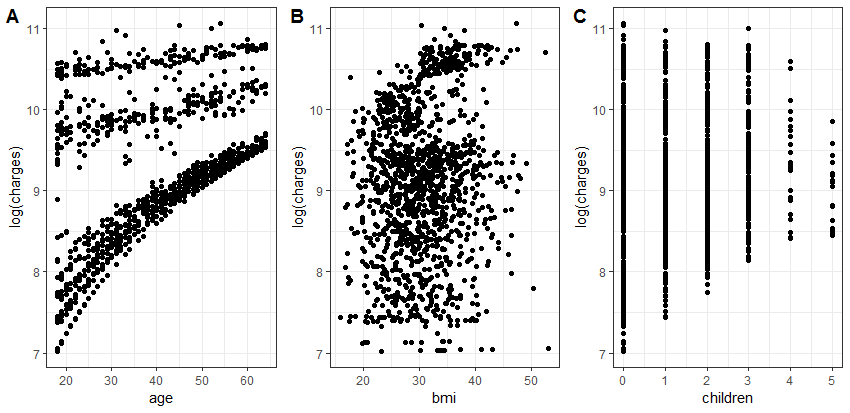


Figure 3: Scatter plot of log(charges) vs age, bmi, children

1. **Categorical independent variable:**

Here we are visually exploring the categorical independent variable and how these variables affect the charges.

Figure 4-A shows a strong relation between smoker and charges. It seems smoker pay significantly higher charges than the non-smoker.

Figure 4-B shows the charges for male and female are approx. equal indicating sex has little or no influence on the charges.

Similarly, in Figure 4-C, we can see that the charges for the northeast, northwest, southeast, and southwest region are approx. equal indicating little or no relation between region and charges.

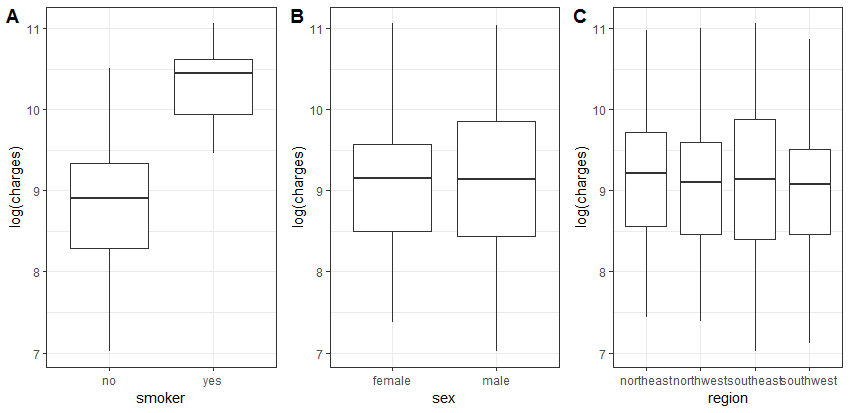


Figure 4: Box plot of log(charges) vs smoker, sex, region

1. **Hypothesis Testing & Model creation**

In this section, we will try to answer a few research questions using statistical tools such as t-statistic, ANOVA, and build a multiple linear regression model to predict the value of medical charges.

1. **Does smoking cigarette affect medical expense?**

**𝐻0**: The true difference in charges for smoker & non-smoker is zero

**𝐻1**: The true difference in charges for smoke & non-smoker is zero



Figure 5: Two sample t-test summary of smoker & non-smoker population

The two-sided test of the difference in average of log-transformed data has a p-value less than 0.05 suggest that the true difference in the population mean value is not equal to zero. Therefore, we reject the Null hypothesis in the favour of the alternative hypothesis. It is estimated that the median medical expense for a smoker is 4.55 () times as large as the median medical expense for a non-smoker. (95% CI: 4.27 to 4.85)

1. **Does medical expense is different for male smoker and female smoker?**

**𝐻0**: The true difference in charges for male-smoke & female-smoker is zero

**𝐻1**: The true difference in charges for male-smoke & female-smoker is zero



Figure 6: Two sample t-test summary of male smoker & female smoker population

The two-sided test of the difference of average log-transformed data has a p-value greater than 0.05 suggests that the true difference in the mean of male smoker and female smoker population is equal to zero. Therefore, we failed to reject the Null hypothesis in the favour of the alternative hypothesis

1. **Does geographical location affect medical expense?**

**𝐻0**:

**𝐻1**: 𝐴𝑡 𝑙𝑒𝑎𝑠𝑡 𝑜𝑛𝑒 𝑚𝑒𝑎𝑛 𝑖𝑠 𝑑𝑖𝑓𝑓𝑒𝑟𝑒𝑛𝑡



Figure 7: Summary of ANOVA log(charges) ~ region

The analysis of variance test of log-transformed populations has a p-value greater than 0.05 that provides no evidence that the distribution of medical charges differs in four populations belonging to four regions (see appendix figure 16 for detailed comparison).

1. **Multiple regression model**

In the data exploration section Part-I and II, we identified that smoker, age, BMI, children significantly influence the charges while region and sex have no significant impact on charges. We considered these influential variables as a predictor variable for the initial multiple regression model.

Equation 1: The initial regression model equation



Figure 8: Summary of initial model (log(charges) ~ age + bmi + smoker + children)

In the above model summary, the Adjusted R-squared value is 0.7614 which represents the amount of variation in the log(charges) that the initial model can explain i.e. 76.14%. The F-statistic is 1068 with 4 and 1333 degrees of freedom has p-value less than 0.05 (<2.2e-16) indicating that the addition of predictor variables is significant than the single mean model.

Equation 2: The initial regression model equation with coefficient

Now we will try to improve this initial model by adding other predictor variables like sex, region, and interaction terms of indicator variables. We will use Akaike’s Information Criterion (AIC) model selection statistic because it is more appropriate if there are not too many redundant and unnecessary variables. We further improved this model by considering only those variables whose p-value is less than 0.05 and by checking the Variance Inflation Factor (VIF) for the existence of multicollinearity among predictor variable.

Equation 3: The updated regression model equation



Figure 9: Summary of model (log(charges) ~ age + sex + bmi + children + smokeryes + interaction)

The updated model has an Adjusted R-squared value of 0.7684 which represents the amount of variation in the log(charges) that the updated model can explain i.e. 76.84%. In the model summary, we can see that all the predictor variables are significant including the interaction term children:smokeryes.

Equation 4: The updated regression model equation with coefficient

let us compare the initial model and updated model using partial F-test to ascertain if the addition of new variables and the corresponding reduction in unexplained variation in the response variable is statistically significant.



Figure 10: Summary of ANOVA initial model vs updated model

Since the p-value is less than 0.05, there is overwhelming evidence that the coefficient of sex () and interaction term () differs from 0 and are statistically significant.

Now let’s check for the existence of multicollinearity among the predictor variables using the Variance Inflation Factor. A general guideline is that a VIF larger than five or ten is large, indicating that the model has problems estimating the coefficient (Petrie, n.d.).



Figure 11: Variance inflation Factor

In figure 11, VIF value for all the predictor variable is less than 5 indicating there is no correlation among predictor variables.

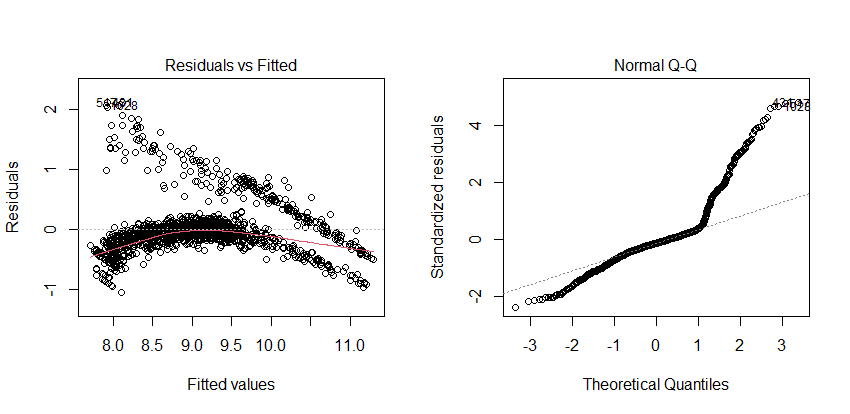


Figure 12: Residual vs Fitted and Normal Q-Q plot of updated model

A residual plot is examined for further exploration of model fit and outliers. Figure 12 (Residuals vs Fitted) shows the residual plot from the model fit with the residuals as log(charge). The residual must not follow any pattern, which is not the case here. From Residuals Vs Fitted plot it is clear that residuals are not exhibiting the random variation around the zero residual line suggesting the possibility of heteroscedasticity and serial correlation. There are two distinct group that exhibit completely different variation than one another. In the Q-Q plot, the points follow a line in the middle between -1 to 1 but curve off in both extreme quantiles. This Q-Q plot corroborates the findings of the residuals Vs fitted plot.

1. **Results and Findings**

In our study, we found that age, BMI, and smoking are the most significant predictor variable that influences medical charges. There is evidence that medical charges rise significantly for smoker when compared to non-smoker. Furthermore, the number of children coupled with smoker also significantly influence medical charges. Our final model is updated model that explains 76.84% variation in the medical charges. However, residual vs. fitted and Q-Q plot shows the variability pattern. We can increase the fit of the model either by removing influential observation, considering other transformations.

1. **Conclusion and Future Improvement**

Through research questions, multiple linear regression, and by comparing the fitted regression model using ANOVA, we predicted the medical charges with a comparatively high degree of accuracy. In real-world medical expense forecasting, several additional variables are considered such as type of disease, the number of times individuals used medical services annually, type of hospital (premium cost, moderate cost, low cost), etc. We were unable to study such variables due to their absence in the dataset. The improvement of this study can be done by gathering additional data on several variables mentioned above.

**References**

An, R. (2015). Health care expenses in relation to obesity and smoking among U.S. adults by gender, race/ethnicity, and age group: 1998-2011. *Public Health*, 129(1):29-36.

Bradford , D. F., & Max, D. A. (1996). Implicit Budget Deficits: The Case of a Mandated Shift to Community-Rated Health Insurance. *MA: National Bureau of Economic Research*.

Heavey, S. (2008). *Consumers Face Rising Medical debt: Survey.*

Nacho, C., & Roger, Z. (2015). *Machine-Learning-with-R-datasets*. Retrieved from github.com: https://github.com/stedy/Machine-Learning-with-R-datasets/blob/master/insurance.csv

Petrie, A. (n.d.). *Variance Inflation Factor*. Retrieved from RDocumentation: https://www.rdocumentation.org/packages/regclass/versions/1.6/topics/VIF

Ramsey, F. L., & Schafer, D. W. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis (3rd ed.).* Boston: MA: Brooks/Cole.

Wikipedia. (n.d.). *Lifestyle disease*. Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Lifestyle\_disease

Appendix A

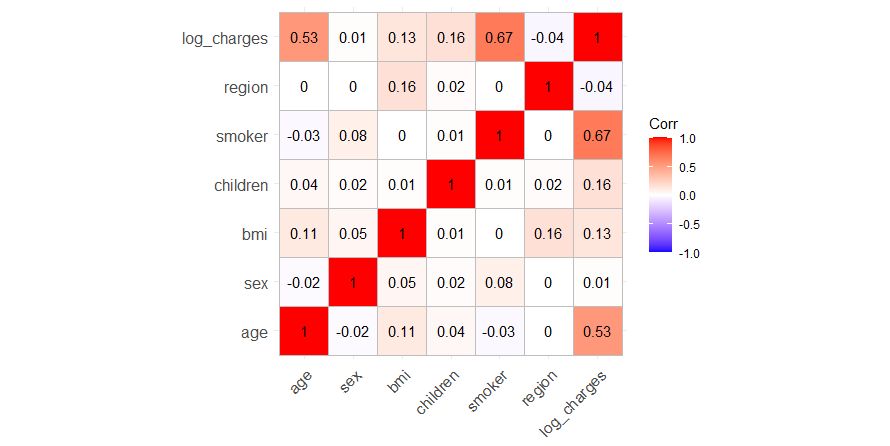
****

Figure 13: Correlation plot

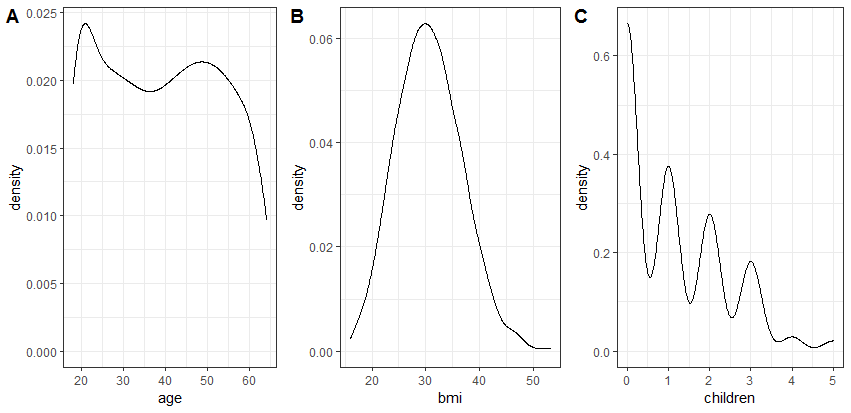


Figure 14: Density plot of age, bmi, children

Appendix A

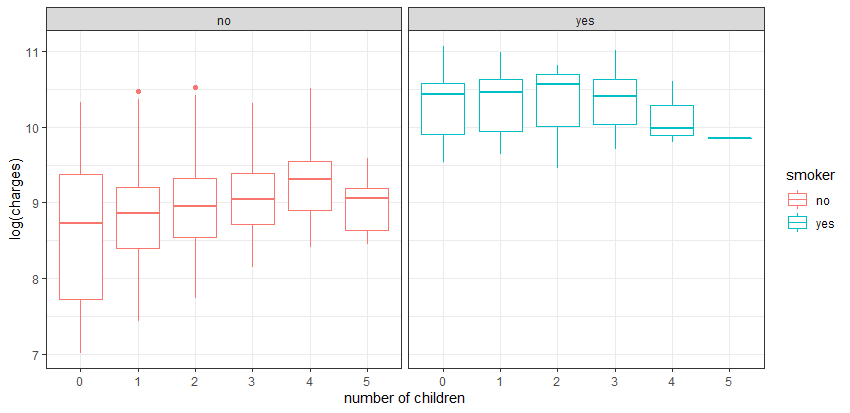


Figure 15: Boxplot of log(charges) vs children faceted by smoker

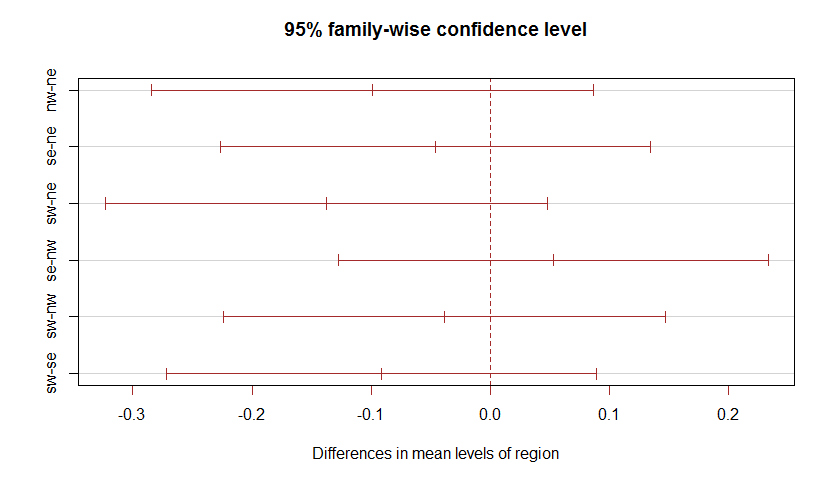
****

Figure 16: Tukey test plot for log(charges) ~ region

Appendix B



Figure 17: Variable selection using Akaike’s Information Criterion (AIC)