DALHOUSIE UNIVERSITY

# STAT 3340 Regression Analysis
# Final Report

Professor: Dr. Joanna Mills Flemming

Prepared By

Xiaoyan Chang B00753321

Yansong Li B00755354

Yichen Yu B00750602

Dec 11, 2020

# Table of Contents

## Abstract

The project is designed to identify and establish a model to show the relationship between six personal factors and medical insurance premiums.  The paper content primarily consists of data description and model analysis; in the data description part, the paper introduces the new data point and provide the visualization of basic relationships between variables, along with the analysis of model selection and modification. By synthesizing these analyses, the paper finds that there is a possible linear relationship between insurances premiums and personal factors. The analysis in this paper is supported by Medical Cost Personal Datasets provided by Zach Stednick.

*Keywords: Medical insurance charges, age, age, sex, body mass index (BMI), children, smoking status, region, linear regression, model*

## Introduction

Medical insurance charges are determined by various factors. In this project, the influence of personal factors such as, age, sex, body mass index (BMI), children, smoking status, and region on medical insurance charges was evaluated. By visualizing these personal factors’ relationships with their insurance costs, it is possible to able to find a possible linear relationship between personal factors and insurance premiums. The stepwise procedures have been applied to find a linear model to fit the dataset, followed by the analysis of this model’s adequacy. The group revised the model through variable transformations, which enabled us to modify the model. As a result, group 18 believes that there is a linear relationship between personal factors and people’s medical insurance charges.

## Data Description

The data includes medical information and costs billed by health insurance companies, there are 1338 observations that are categorized in seven columns: age, sex, BMI, children, smoker, region, and charges (Scott, 2020). By exploring the relationship between these variables, we would be able to find how a person’s healthy status and relevant information background influence their medical insurance charges.

### Data preparation

Group 18 introduced a new additional data point of a 39-year-old male non-smoker with BMI of 30.66 and insurance premium of $6635.21 who has 1 child and lives in the northeast of the U.S. This data point was designed to fit into the original dataset. Therefore, we assigned the average age and BMI to this new data point. For its sex, smoking status, and region, we chose them randomly since they are dummy variables. Meanwhile, we assigned the charges to less than $10000 since this data point is not a smoker and a healthy man. After that, we introduced three sets of dummy variables by using the build in function of R. For variable sex, we replaced female with 1 and male with 2. For variable smoker, we replaced non-smoker with 1 and smoker with 2. For variable region, we set northeast equals to 1, northwest equals to 2, southeast equals to 3 and southwest equals to 4.

**Data visualization**

To get a better understanding of the data, we explored the data distribution of charges and age by using histogram plots. The histogram of charges indicates that most people have an insurance premium range between $0 and $20000, which is the first 25th quantile of all charges. According to the figure 1, the most insured people are at their twenties and least people are 65 years old, people of other ages were distributed evenly.

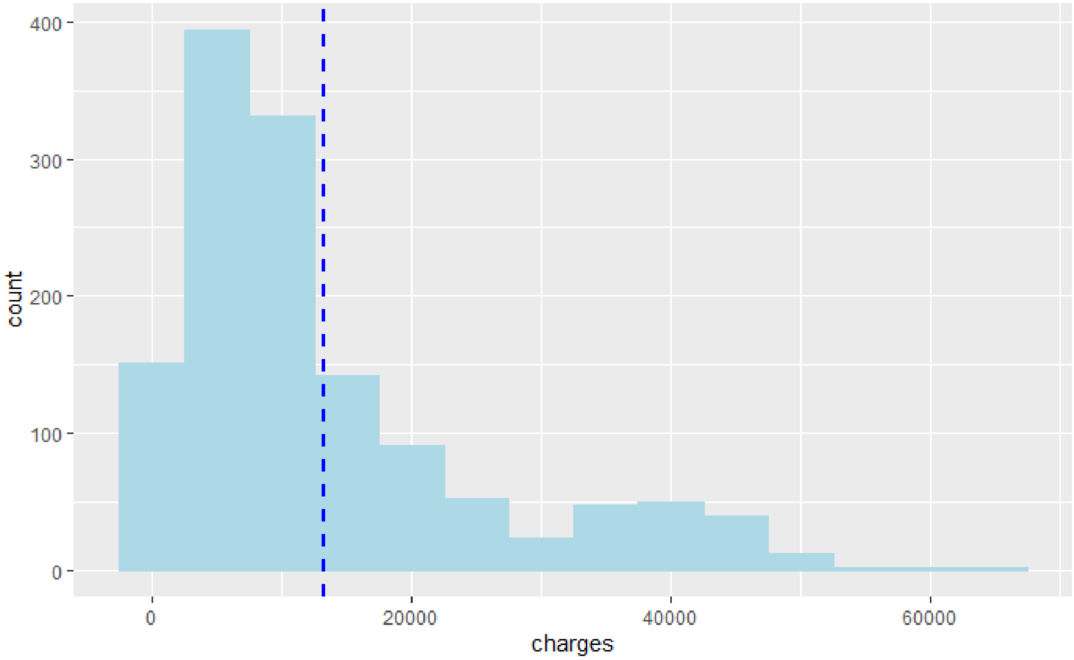


*Figure 1: The histogram plot of age*

Next, we derived out several ggplots in the figure 2 to demonstrate the relationship between some personal factors and charges. What interests us the most is that there is an obvious positive linear relationship between age and charges and an approximate positive linear relationship between BMI and charges. Through investigations, our group found that higher age is believed to be the cause of higher mortality, which means that senior people are charged by higher premium. According to t Schoenfeld, A. J., & Wahlquist, T. C., (2015): “age was the most important risk factor for mortality and the number of medical comorbidities strongly influenced total charges.”. Moreover, there is an evidence showed that higher BMI usually leads to higher insurance charges (Bhattacharya, J., & Bundorf, M. K., 2005). In addition, from the ggplot of Smoker vs. Charges, it is obvious that smokers are paying more money for their medical insurances compared to non-smokers. Kristin Madison (2013) did a relevant study to investigate how smoking affect insurance charges and she discovered that smokers are usually charged by higher premiums. On the other hand, we observed that people with more children in their family tend to have less insurance charges.

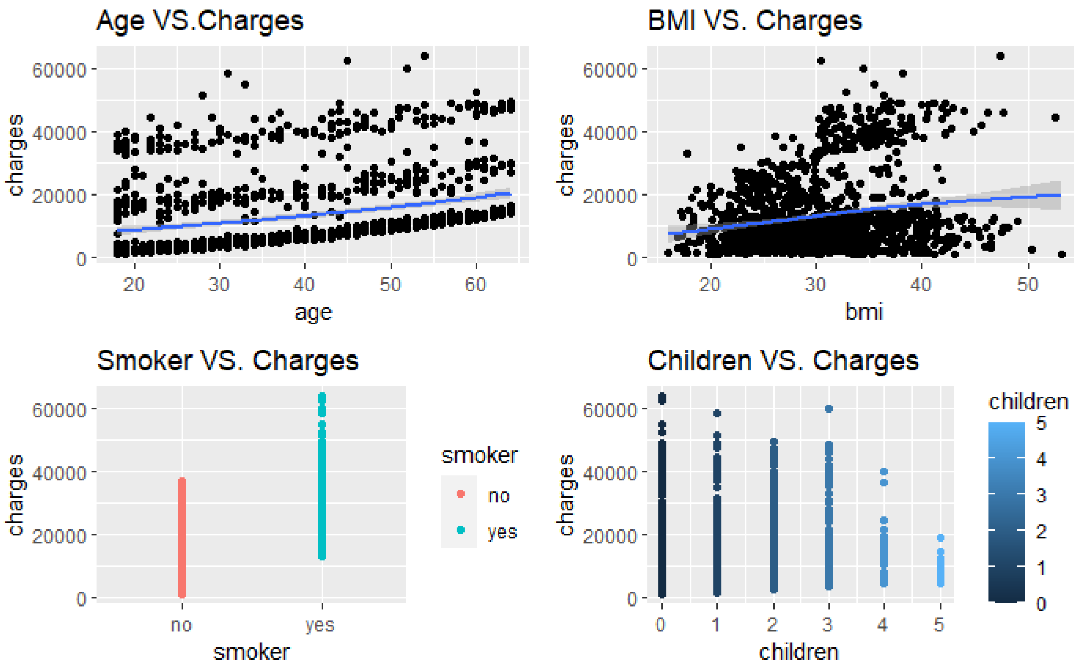


*Figure 2: Four ggplots showing the relationships between regressors and charges*

Further, we apply the correlation plot as shown in figure 3 to illustrate the possible linear relationship between these variables. Age, BMI, children, and smoker are linear related to the dependent variable charges. Nevertheless, regressors like sex and region do not show an obvious linear relationship with charges, which implies they may not be necessary in the model. Hence, we would be able to confirm this conjecture in the model selection part.

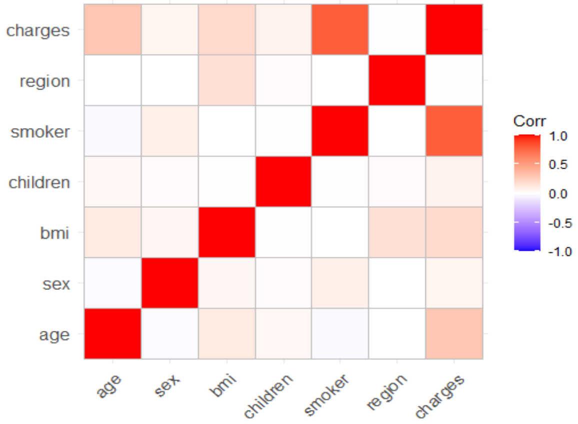


*Figure 3: Correlogram for correlation matrix*

# Methodology

## Analysis tools description

In the model selection part, we chose to utilize stepwise procedures (the combination of forward selection and backward elimination) to find a feasible model:

1. Forward selection: adding independent variables sequentially and deciding on whether to enter an independent variable by using Akaike Information Criterion (AIC).
2. Backward elimination: instead of fitting the largest model under consideration, the group introduced a reasonable interaction into the model, and then use AIC to remove regressors from the model.
3. Full stepwise: to check the inconsistencies issue of stepwise procedures, alternating between adding and removing variables, using AIC at each step.

Comparing the model from full stepwise procedure to backward elimination model was conducted to ensure that there are no inconsistencies and finalize the model for this section.

For the model revising, group 18 is intended to check whether it meets the Gauss Markov assumptions by analyzing the model's four plots:

1. Residual vs Fitted plot: detect non-linearity of residuals, unequal error variances, and outliers.
2. Normal Q-Q plot: check whether residuals of the regressors follow a normal distribution.
3. Scale – Location plot: check the assumption of equal variance.
4. Residual vs Leverage plot: find influential points if any.

By analyzing the plots of stepwise model, the group found that data is not normally distributed. Thus, we made variable transformations to the stepwise model in order to address this problem. Then, we plotted and compared the transformed model with the stepwise model, it is possible to make a more comprehensive analysis.

## Model selection

As described above, we chose stepwise procedures to conduct model selection. In the first step, by using forward selection to the model, the model adds smoker as the first regressor, as the model with this variable has the smallest AIC value. Similarly, applying the same selection procedures, the model adds regressors: age, BMI, children, and region, sequentially. Note that the variable of sex does not add into the model, since the model containing the variable of sex has a largest AIC value through the whole process of forward selection. Consequently, the model we obtained by using forward selection procedure is charges = smoker + age + BMI +children + region, with the AIC value of 23328.07.

After forward selection, we introduced a highly relevant interaction, BMI*smoker, into the model that we generated from the first step. Through investigation, we found out that people's BMI is highly related to their smoking status. Specifically, smoking is associated with lower BMI and smoking cessation with higher BMI (Piirtola et al., 2018). Thus, the model that we put into backward elimination procedure is charges = smoker + age + BMI + children + region + BMI * smoker. In this step, we will remove variables sequentially according to models' AIC values. However, the output model of this step remains the same as the input model with the AIC of 22733.64, which means it is unnecessary to remove these variables.

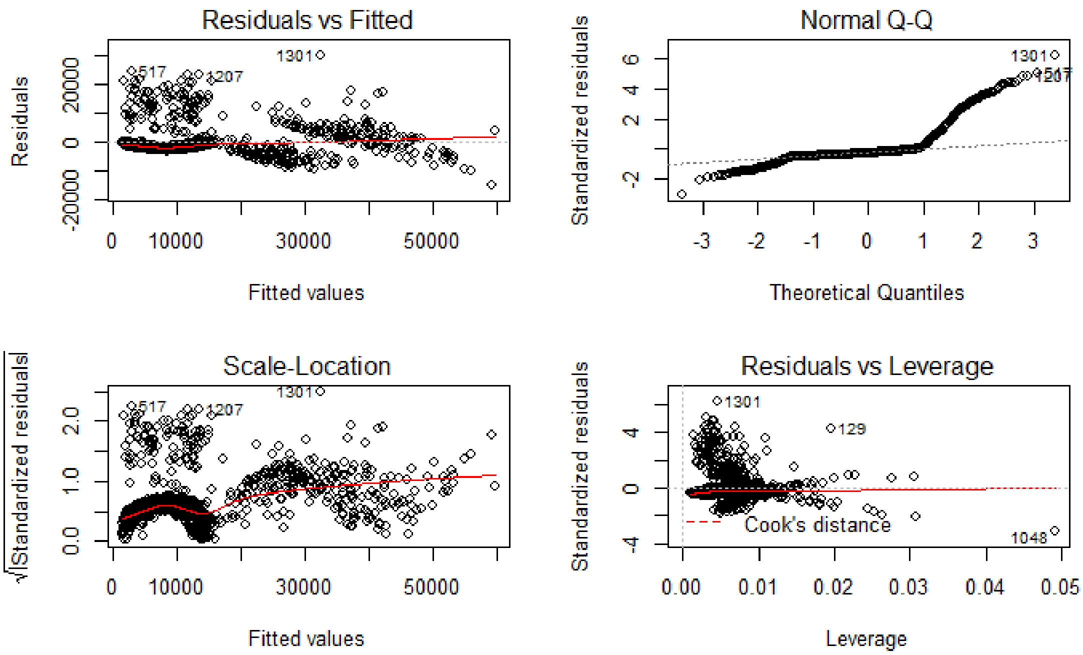
For the final step of model selection, despite that we have already obtained a model from backward elimination procedure, it is essential for us to check inconsistencies when using stepwise approach. By applying the full stepwise procedures, we alternated between adding and removing regressors at the same time by checking models' AIC values, and we figured out that the final model of this step is identical to the model that we obtained from backward elimination procedure. Therefore, we can select the model that charges = smoker + age + BMI + children + region + BMI * smoker to conduct further exploration.

# Results

## Diagnostic plots analysis and model summary

Once we got the model from the stepwise procedures, we chose four diagnostic plots to check if the model fits the data well.

*Figure 4: Diagnostic plots for the stepwise model*



## Residual vs Fitted plot

This plot helps us to determine if there is a liner relationship between predictor variables and an outcome variable by the distribution of residuals (Bommae, 2015). For our current model, the residuals are randomly distributed, which means there is no liner relationship between insurance charges and the personal factors that involved in the model such as age and BMI. However, it is worthwhile to notice that the fitted values are crowed together between 1000 and 15000 with very small residuals in the plot. These fitted value points that are distributed closely in this area represent people who do not have serious negative personal factors that can increase their insurance charges such as being smokers. Therefore, those people are just paying basic charges for their insurance with no significant differences. Hence, we reckon that the residuals are equally distributed.

### Normal Q-Q plot

For the Normal Q-Q plot, it helps us to show if residuals are normally distributed, which means the data points form as an increasing straight line (Bommae, 2015). Nevertheless, the Normal Q-Q plot displays a bimodal line rather than a straight line, which means it has a "spike" of two identical values. As a result, we believe that the residuals are not normally distributed, which triggers us to consider applying variable transformation in the stepwise model.

### Scale-Location plot

The third diagnostic plot is Scale-Location plot. It shows how the standardized residuals are distributed along with the ranges of predictors (Bommae, 2015), which can help us check the assumption of equal variance (homoscedasticity). From the picture blow, we can say that the majority of standardized residuals are equally distributed. Besides, it shows a similar problem that the Residual vs Fitted plot has, which is some of the fitted values are concentratedly appears in the area of 1000 to 15000. Therefore, it should be caused by the same reason for the Residual vs Fitted plot. As a result, we say that the standardized residuals are equally distributed enough.

### Residual vs Leverage plot

The Residual vs Leverage plot is able to find influential cases, which means find out the data points that will affect the linear regression analysis if we remove them from the dataset (Bommae, 2015). Even though data have extreme values (outliers or leverage points), they might not be influential to determine a regression line. From the graph, we can say that we have data point 1048 as a leverage point and 1301, 129 as outliers, but none of them meet the two conditions of the influential point at the same time. Meanwhile, Cook's distance is also too small to be observed. Therefore, we conclude that there is no influential point in this dataset.

### Model summary

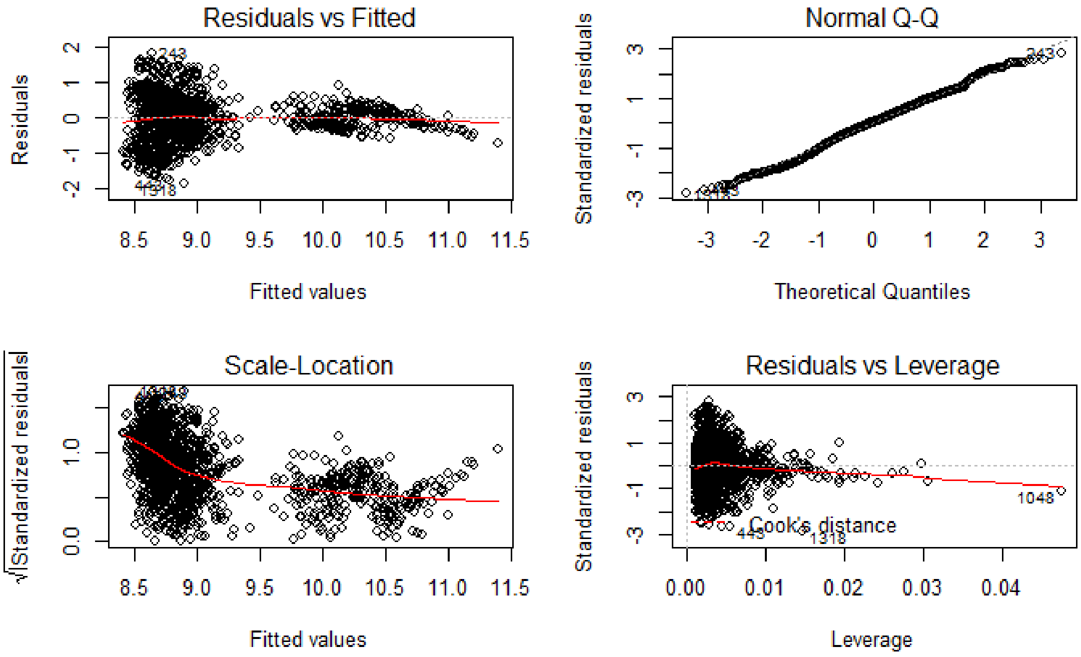
After we analyzed diagnostic plots of the stepwise model, we used summary command to check the adjusted coefficient of determination (R-square) of the stepwise model, as it can help us determine the percentage variation in the dependent variable (charges) explained by independent variables (personal factors). Specifically, the adjusted R-square of the stepwise model is 0.8396, which is a decent percentage for a model that is built from a real-world dataset.

In this case, we concluded that the model is reasonable in terms of adjusted R-square. Despite of that, the information that we interpreted from these diagnostic plots are not satisfying, which triggers us to apply variable transformation into the stepwise model in order to modify its adequacy.

## Model Revising – Variable transformation

To reduce the skewness, we performed transformation on dependent variable and independent variables. After several tries, we found that taking log of charges and inverse of age could meet our expectations.

*Figure 5: Diagnostic plots of the modified stepwise model*



After the transformation, we observed a nearly horizontal line at 0 in the Residuals vs Fitted plot. The Normal Q-Q plot shows a pattern of a nearly straight line with positive slope, which means we succeeded in fixing the informality issue of the model. For Scale-Location plot, we observed a relative sharp decline at the beginning then it gradually became steady. According to our previous description on the distribution of charges, we believe that massive datapoint are located at the first twenty-five percentile, which could explain the sharp decline. Thus, we can still say that our transformed model has homoskedasticity. However, we observed that the line in Residual vs Leverage plot was clearly influenced by the 1048th data point. We labelled this point as our leverage point. Besides this leverage point, the line shows an overall horizontal pattern in Residual vs Leverage plot.

For the modified stepwise model, we apply summary command to analyze. According to the summary command, we observed a R-squared of 0.4997 and an adjusted R-squared of 0.4978. And the t tests for each independent variable have an extremely small p-value except variable smoker, we believe an interaction of BMI and smoker could explain the high p-value of variable smoker. According to the extremely small p-value we got, we can conclude that most of the independent variables are significant in this model. Interestingly, we found that even though the modified model has better performance in its diagnostic plots, the adjusted R-squared dramatically decreased compared to the original model. Nevertheless, as we took the log of variable charge and inverse of variable, the value of R-squared and adjusted R-squared cannot accurately describe the relationship between the dependent variable and independent variables. We made a conclusion that the transformed model could better describe our data.

## Conclusion

The group believes that this paper could be a valid explanation for those who want to gain a basic understanding of how medical insurance premium is calculated and affected by their personal factors. In summary, there is a linear relationship existing between these personal factors and medical insurance costs. Our group found that age, BMI, children, smoker, and region strongly influenced medical insurance charges trough data visualizations. We fitted the data into a linear regression model by applying stepwise procedures and examined the model's adequacy by interpreting its diagnostic plots. As we discovered that the original data points are not normally distributed, we transformed some variables of the model in order to reduce this skewness. Thereby, we could obtain a modified model with improved model adequacy. In practical, it is difficult to have a perfect normally distributed dataset. Thus, we will need to pay more attention to analyze the data before putting the data points into a model. In this way, we would greatly increase a model's accuracy and adequacy.

## References

Bhattacharya, J., & Bundorf, M. K. (2005). The Incidence of the Healthcare Costs of Obesity. doi:10.3386/w11303

Bommae, K. (2015, September 21). University of Virginia Library Research Data Services + Sciences. Retrieved December 09, 2020, from https://data.library.virginia.edu/diagnostic-plots/.

Madison, K., Schmidt, H., & Volpp, K. G. (2013). Smoking, Obesity, Health Insurance, and Health Incentives in the Affordable Care Act. *Jama, 310*(2), 143. doi:10.1001/jama.2013.7617

Piirtola, M., Jelenkovic, A., Latvala, A., Sund, R., Honda, C., Inui, F., . . . Silventoinen, K. (2018). Association of current and former smoking with body mass index: A study of smoking discordant twin pairs from 21 twin cohorts. *Plos One, 13*(7). doi:10.1371/journal.pone.0200140

Schoenfeld, A. J., & Wahlquist, T. C. (2015). Mortality, complication risk, and total charges after the treatment of epidural abscess. *The Spine Journal, 15*(2), 249-255. doi:10.1016/j.spinee.2014.09.003

Scott, L. (2020, October 27). 10 Open Datasets for Linear Regression. Retrieved December 04, 2020, from https://lionbridge.ai/datasets/10-open-datasets-for-linear-regression/

Stednick, Z. (2014). GitHub. *stedy/Machine-Learning-with-R-datasets*. https://github.com/stedy/Machine-Learning-with-R-datasets.

# Appendix

### GitHub Repository

https://github.com/second12138/STAT-3340-Final-Project-Group-18.git

### The source codes (R Markdown file)

https://github.com/second12138/STAT-3340-Final-Project-Group-18/blob/main/STAT%203340%20-%20Project.Rmd

### The dataset (insurance.csv)

https://github.com/second12138/STAT-3340-Final-Project-Group-18/blob/main/insurance.csv