Group L2C5 Final Project:

Predicting Insurance Charges

## 

## Group Member：

Xinru Jiang 94227386

Haonan Su 15763659

Junhong Ma 13425160

Yitong Zhang 22072904

## Introduction

The rising cost of healthcare has made insurance a crucial component of safeguarding individuals' financial and personal well-being. According to a study by the National Bureau of Economic Research, BMI, smoking status, age, and children are the primary factors contributing to the cost of healthcare premiums. Our project will be using the [Prediction of Insurance Charges](https://www.kaggle.com/datasets/thedevastator/prediction-of-insurance-charges-using-age-gender) dataset to predict insurance charges based on input variables such as age, gender, BMI, number of children, smoking status, and region of residence.

Our goal is to develop a prediction model that accurately estimates insurance charges, enabling policyholders and insurers to manage risk and remain financially stable. In this project, we will explore different model selection techniques, including forward selection, to determine the most accurate prediction model. We will evaluate the performance of each model using the Root Mean Squared Error (RMSE) along with the Bayesian Information Criterion (BIC), and Adjusted R-squared and verify the model assumptions. RMSE is a widely used metric for assessing the accuracy of a predictive model, which quantifies the average of the squared differences between the predicted and actual values in a dataset. BIC is a statistical measure used for model selection that penalizes models with more parameters, favoring simpler models. Adjusted R-squared is a modified version of R-squared that adjusts for the number of predictors in a linear regression model and provides a better measure of the goodness of fit of the model.

The dataset we used is collected by Wakefield, B. (2020), and can be found on this website <https://data.world/bob-wakefield/insurance>. The dataset contains the following variables of insurance customers:

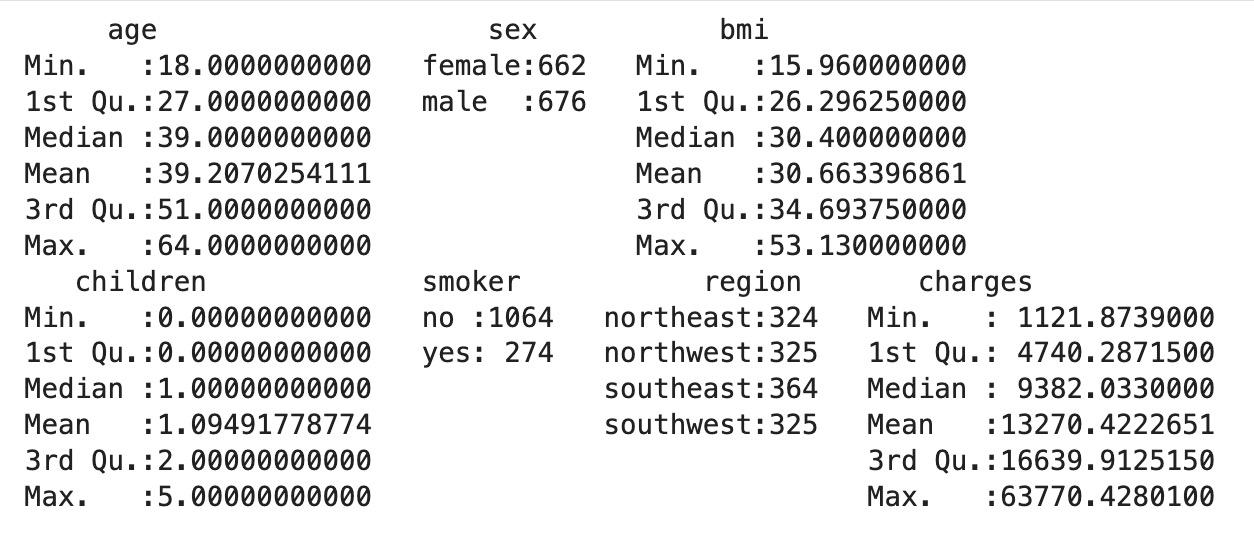
Table1. Variables in the Dataset

| **Variables** | **Units** | **Description** |
| --- | --- | --- |
| **Age** | years | The age of the customer. |
| **Sex** | male/ female | The gender of the customer. |
| **Bmi** | kg/m^2 | The body mass index of the customer. |
| **Children** | N/A | The number of children the customer has. |
| **Smoker** | false/ true | Whether or not the customer is a smoker. |
| **Region** | southeast/ southwest/ northwest/ northeast | The region the customer lives in. |
| **Charges** | dollar | The insurance charges for the customer. |

## Exploratory Data Analysis

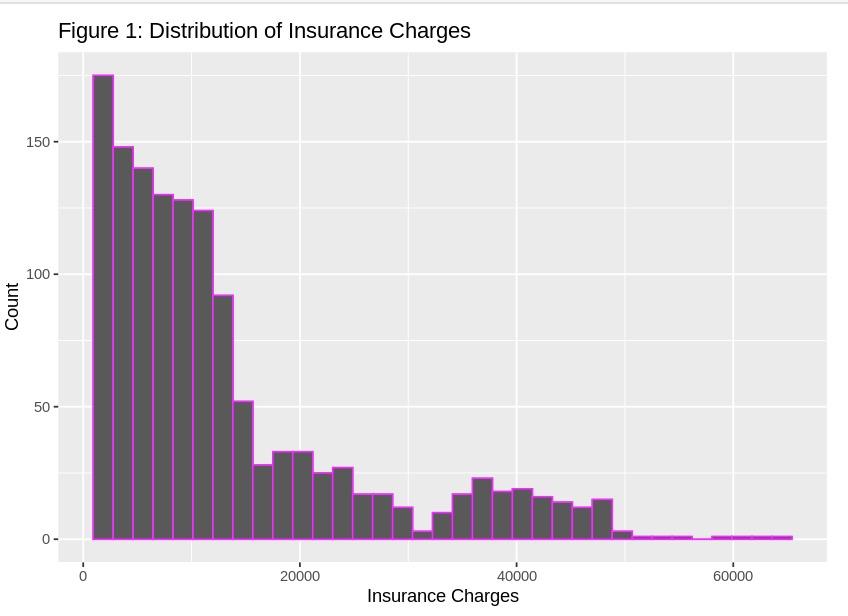
In this section, we will perform EDA on the insurance charges dataset to understand the characteristics of the data and identify any patterns or trends that could help us in predicting insurance charges. We first checked for data quality and observed no missing values. As shown in figure 1, we also explored the summarization of the dataset.

Figure 1. The Summarization of the Dataset



The summary of the dataset provides a quick overview of the variables in the dataset. Upon examining the distribution of our predicted variable, Insurance charges, according to figure 2, it was discovered that the 'charges' variable is not normally distributed. This finding prompted us to explore the possibility of using a transformation technique to improve the modeling process.

Figure 2. Distribution of Insurance Charges



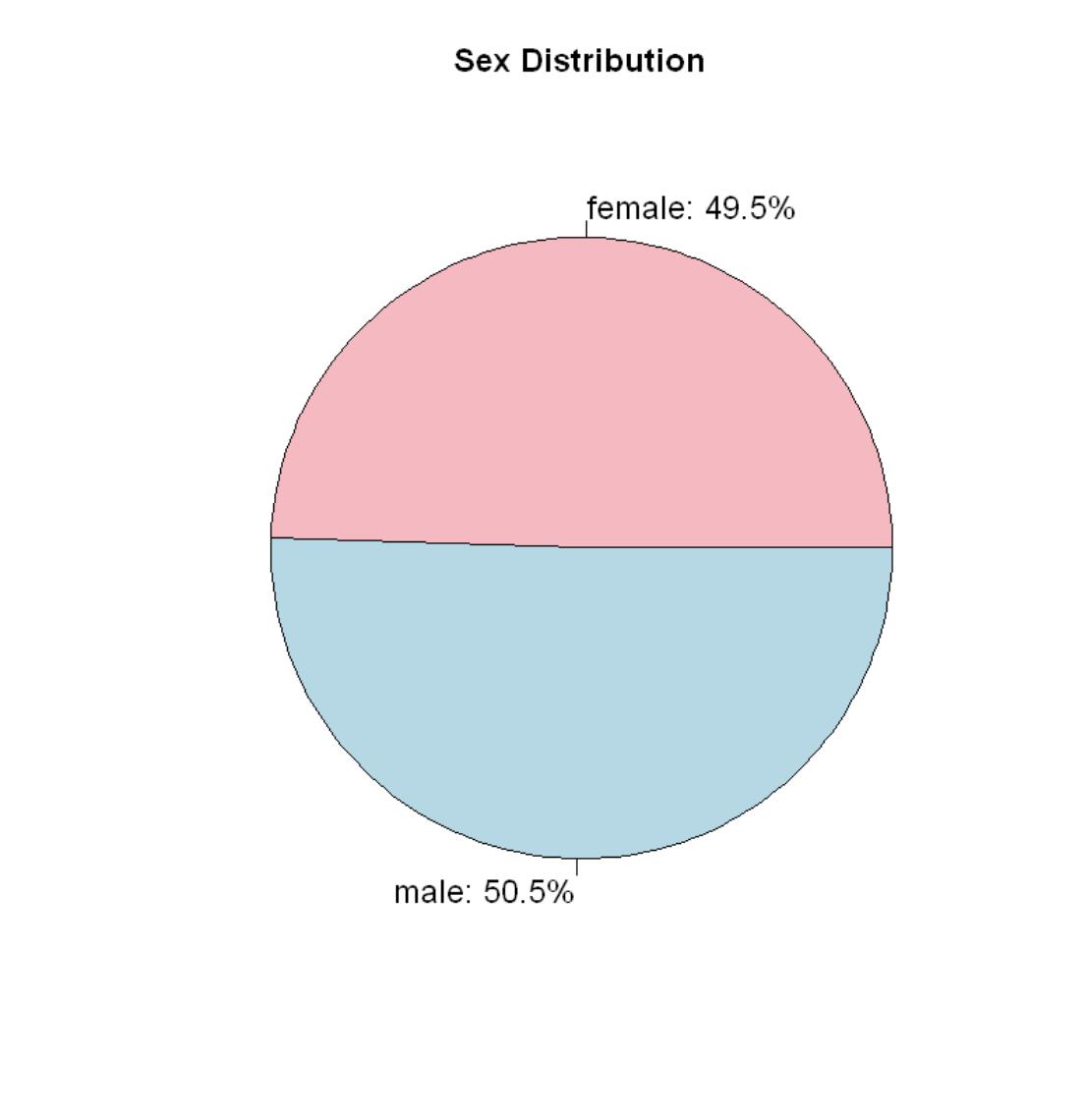
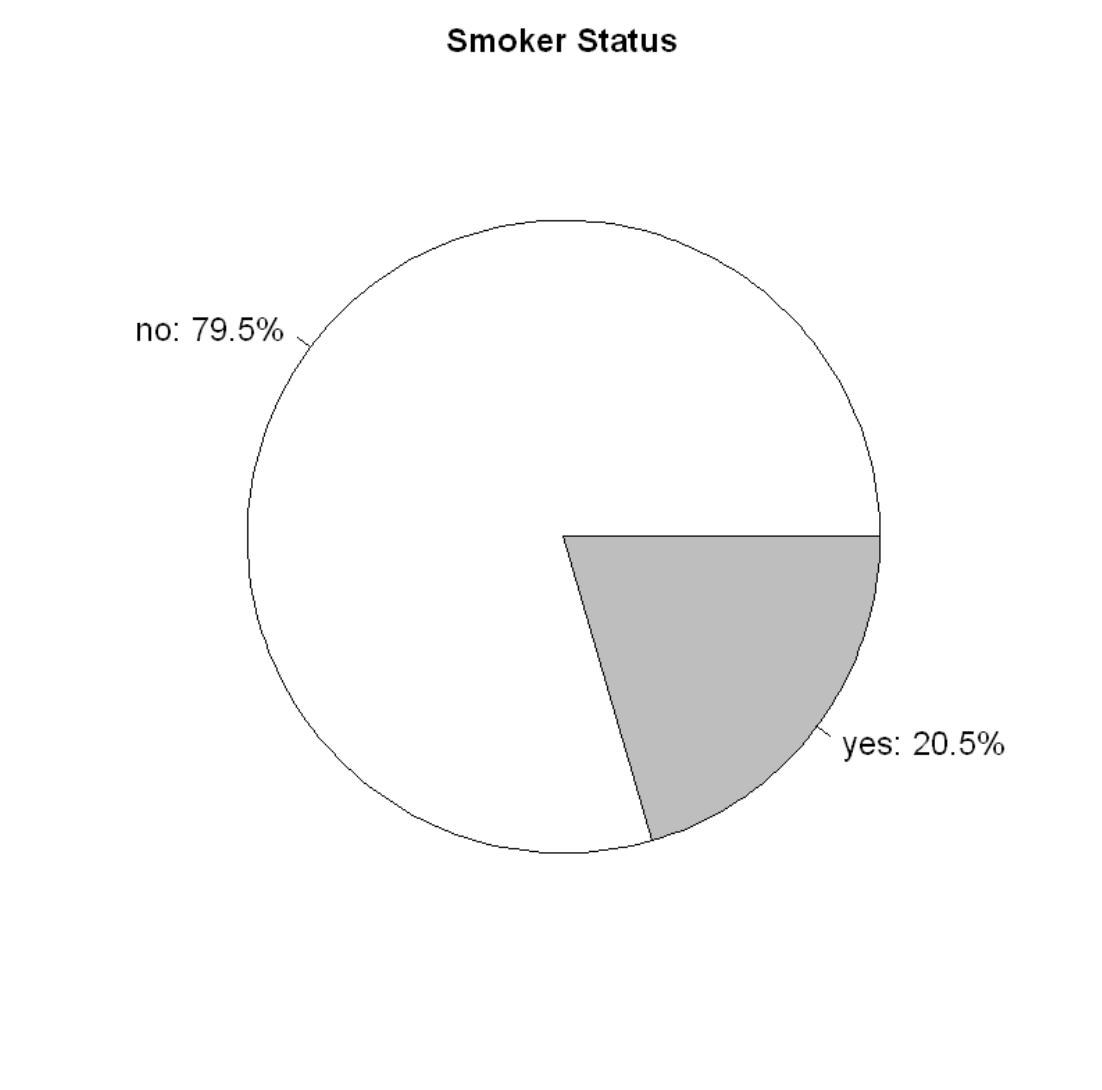
To further explore the data distribution and relationships between variables, we used the ggpairs() function in R. Our analysis revealed that the explanatory variables 'age', 'bmi', and 'children' exhibit a linear relationship with the response variable 'charges', as shown in figure 3.

Figure 3. Plots between two variables



Moreover, from the two pie charts below, we observed that the categorical variable ‘sex’ is almost evenly split between two genders, while for variable ‘smoker’, 79.5% of the examples are not smokers, and 20.5% of the examples are smokers.

Figure 4. Pie Chart for Variable ‘sex’ Figure 5. Pie Chart for Variable ‘smoker’



## Model Selection and Results

Our selection approach involved examining the residual plots of each model derived from the training data. By analyzing the trends in the residual plots, we determined whether to introduce new terms to construct improved models or transform variables. The objective of this approach is to find the model with the minimum RMSE while ensuring that the residual plot is reasonable.

Firstly, the dataset was splitted into a training set (80% of the data) and a testing set (20% of the data). The training set was used to develop and refine our model, while the testing set was reserved to assess the model's predictive accuracy. The rationale behind splitting the data is to evaluate the model's performance on unseen data, and avoid double fitting the model. It provides the ability to apply statistical measures such as Bayesian Information Criterion (BIC) to assess the model's performance on the training set and estimate the root-mean-square error (RMSE) on the test set without direct access to the test data.

Secondly, the full model was fitted with all explanatory variables and checked the Variance Inflation Factor (VIF) for each explanatory variable in the multiple regression model. VIF is a statistical measure used to detect the presence and severity of multicollinearity in regression analysis. From Table 2, it can be seen that all of the explanatory variables have VIF (GVIF) values around 1.0, which indicates that there is no evidence of highly correlated variables in the model that need to be removed.

Table 2. The table of VIF for each explanatory variable in the dataset



When building a predictive model, using all available input variables may lead to overfitting, where the model fits the training data too closely but performs poorly on new data. A better approach is to use a smaller subset of input variables that are most relevant to the outcome (insurance charge) we are trying to predict.

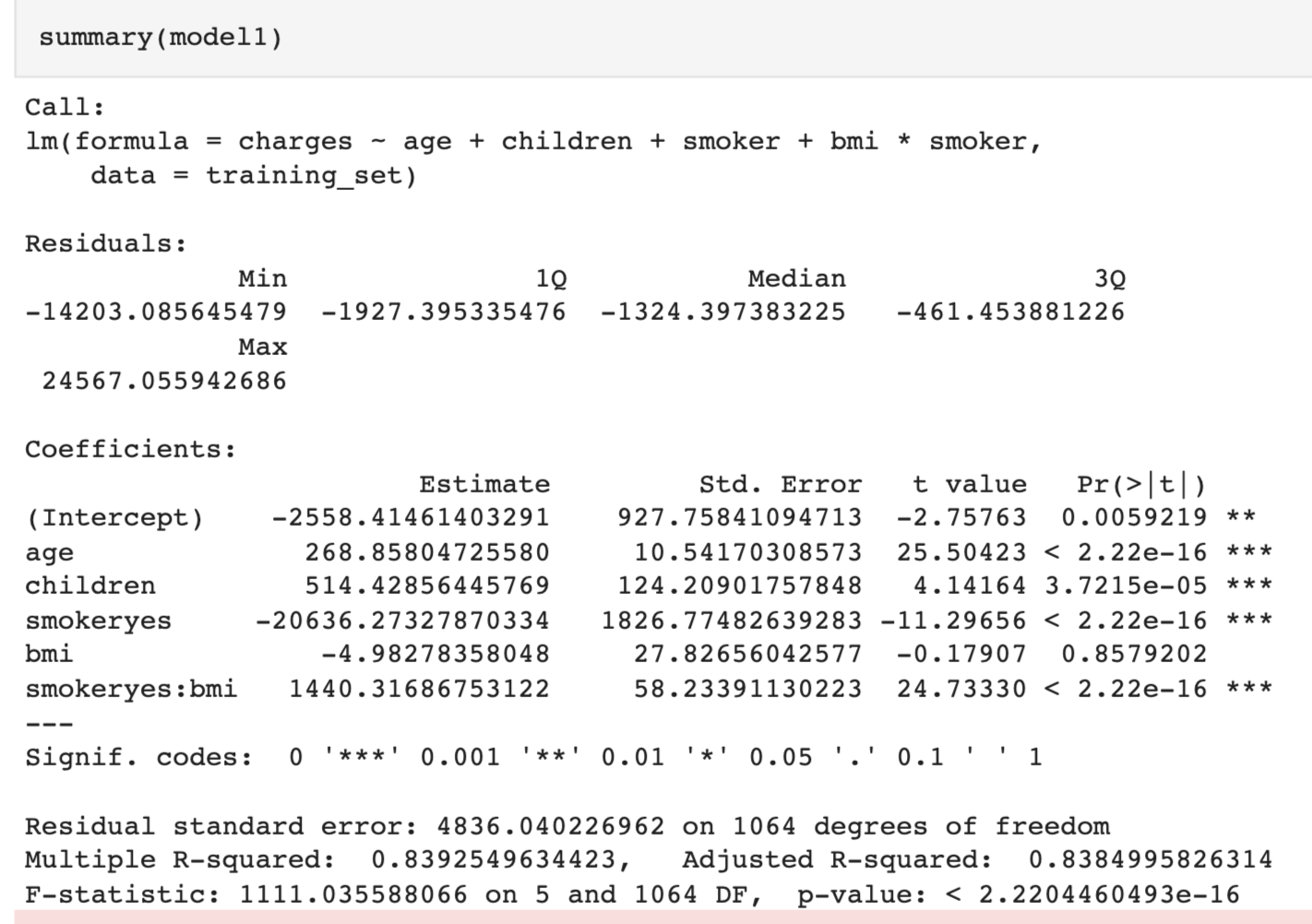
Therefore, a forward selection algorithm with a maximum of 8 parameters utilizing the regsubsets() function in R was employed to identify several subsets of input variables for models of varying sizes. Subsequently, the best model was selected by evaluating both of the Bayesian information criterion (BIC) and adjusted R-squared against a set of predefined criteria.

During the selection process, the Bayesian information criterion (BIC) was utilized as the primary determinant, while adjusted R-squared served as a secondary measure. Following our assessment, we concluded that the model with four parameters was the most optimal, as shown in Table 3, the model with 4 parameters has the lowest BIC value of approximately -1920.98 and a relatively high adjusted R-square value of about 0.839.

Table 3. The Adjusted R-squared and BIC Values of Each Model in Forward Model Selection

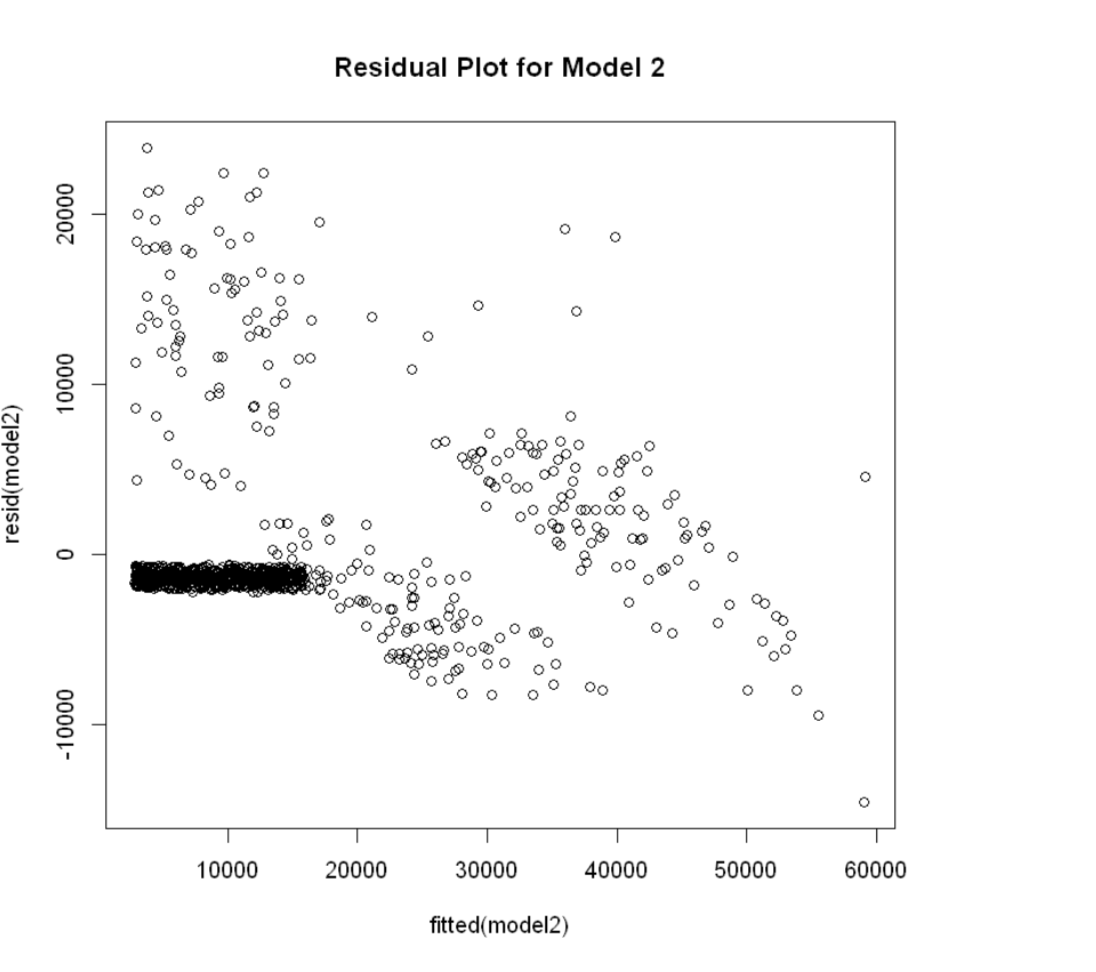
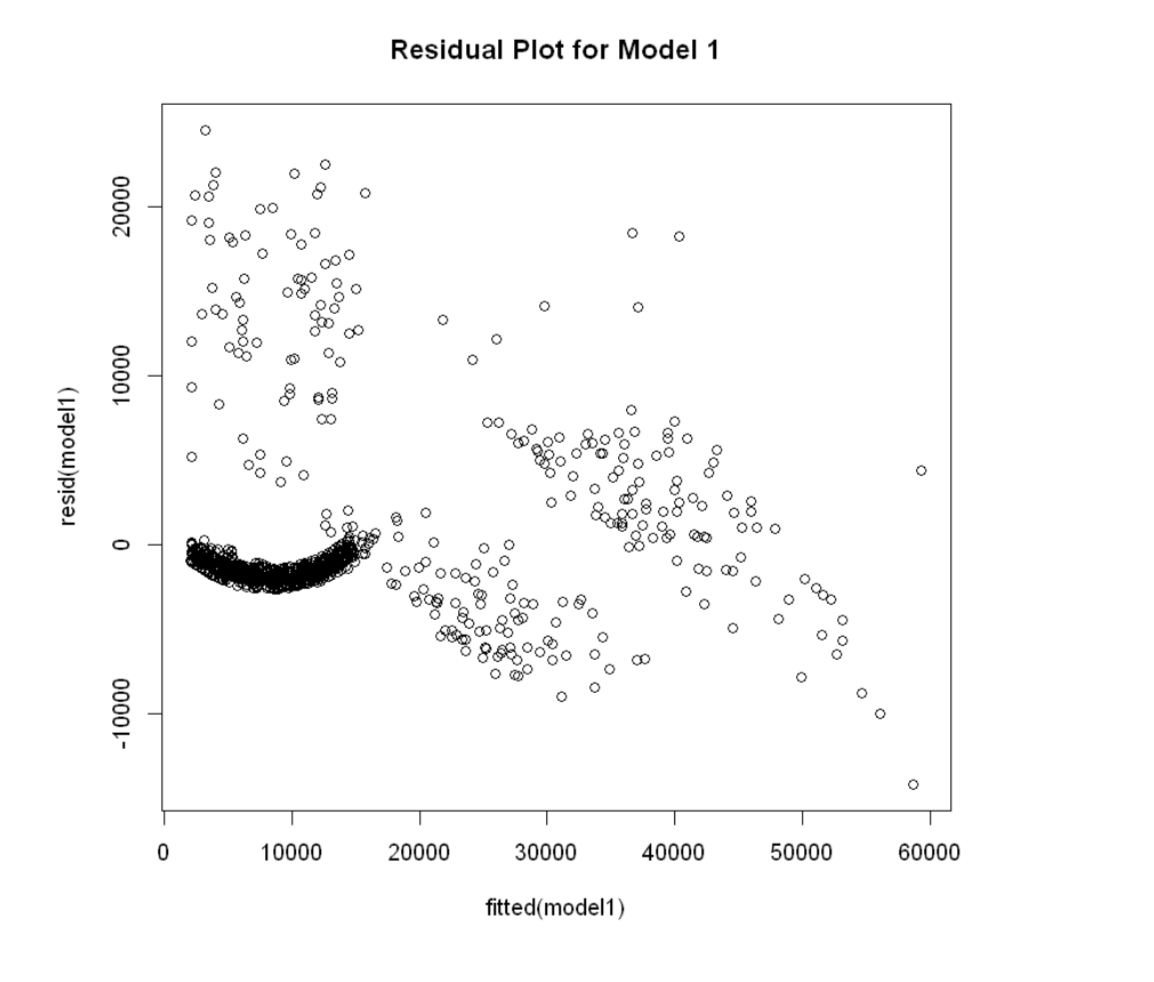
| **n** | **AdjR2** | **BIC** |
| --- | --- | --- |
| 1 | 0.709 | -1306.760 |
| 2 | 0.812 | -1770.673 |
| 3 | 0.836 | -1910.874 |
| **4** | **0.839** | **-1920.982** |
| 5 | 0.840 | -1920.962 |
| 6 | 0.840 | -1919.874 |
| 7 | 0.841 | -1917.974 |
| 8 | 0.841 | -1914.087 |

Based on the results of the forward model selection, the optimal model was identified to have four predictors, namely 'age', 'children', 'smoker', and 'bmi\*smoker'. Model1 was then fitted using these predictors, and the output is shown in figure 6. The Adjusted R-squared is around 0.8385.

Figure 6. The Output of Fitting Model 1

For further exploration of the model, from figure 7, it can be observed that the residual plot of Model1 exhibited a quadratic trend for fitted values between 0 and 20,000.

Figure 7. Residual Plot for Model 1 Figure 8. Residual Plot for Model 2

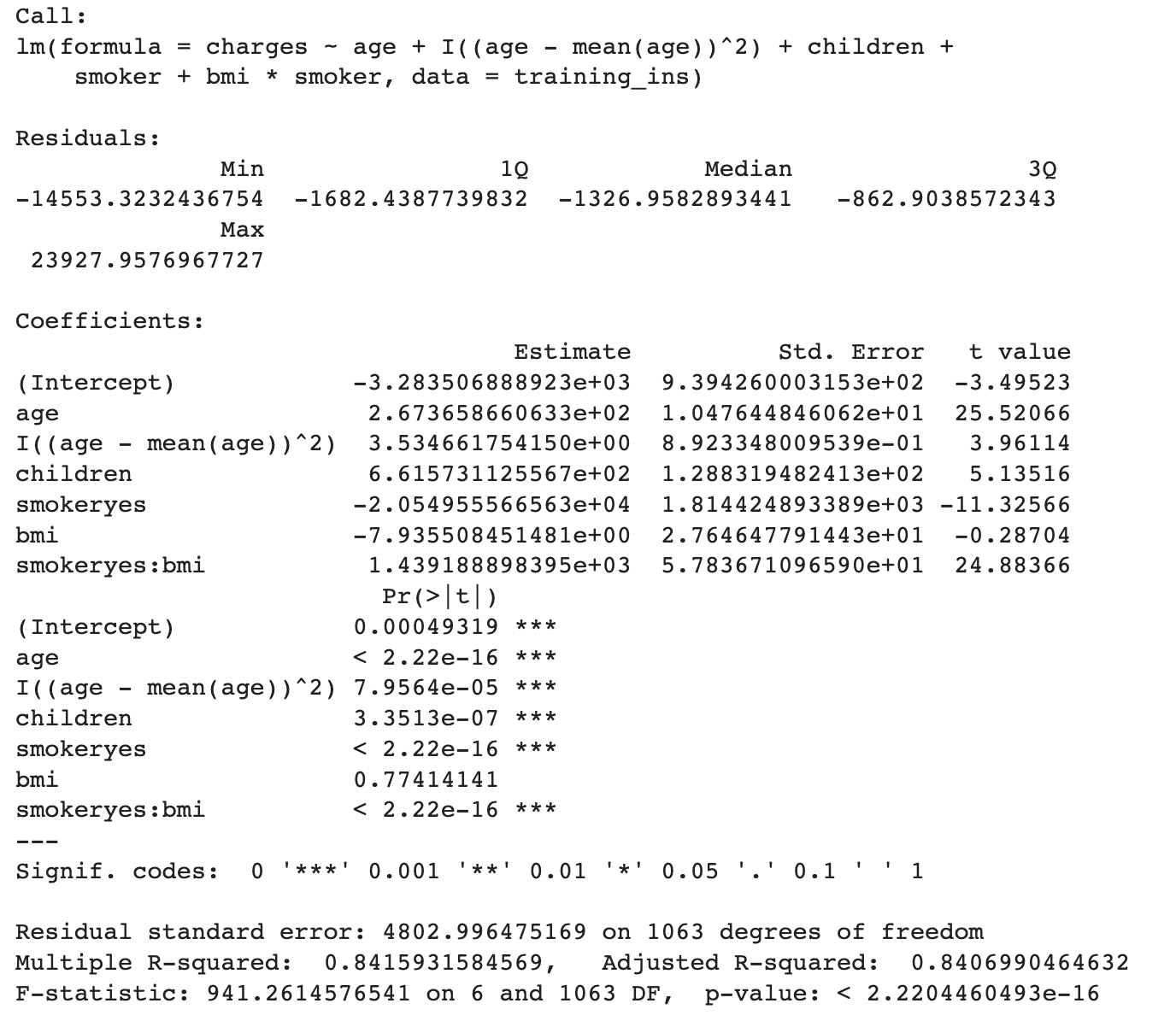


To address this issue, the term (age - mean(age))2 was added to Model1 to eliminate the trend while avoiding collinearity. The model with the new term was named Model2. By comparing figure 7 and figure 8, there is a change from a quadratic curve to a linear trend. Moreover, the residual plot for Model 2 still showed a pattern. Both residual plots seem to violate the homoscedasticity assumption associated with a linear regression model, in other way, the variance of residuals is not constant.

To identify the best model and correct the homoscedasticity problem, we applied a logarithmic transformation to the response variable, charges, of Model1 and Model2, creating two new models denoted Model1\_log and Model2\_log, respectively. However, the transformation to the response variable didn’t eliminate the violation of homoscedasticity, and it significantly increased the root-mean-square error (RMSE) for both models.

The Adjusted R-squared of model 2 is about 0.8407, which is a little bit higher than model 1’s. As a result, Model 2 seems to be the optimal model. An Adjusted R-squared value of 0.8407 indicates that approximately 84.07% of the variation in insurance charges can be explained by the independent variables included in the model, while the remaining 15.93% of the variation is not accounted for by the model. This suggests that the model is a good fit for the data and is able to explain a large proportion of the variation in insurance charges.

Figure 9. The Output of Fitting Model 2



By calculating the RMSE values for these models and comparing RMSE values in table 4, it can be concluded that Model 2 had the smallest value of about 4970.917 which aligns with our expectations. Therefore, Model2 was considered as the best predictive model we have discovered so far. The pattern in Model 2's residual plot may be because some parameters that affect the response variable were not collected in the dataset, such as annual income.

Table 4. The RMSE value of each model

| **Model** | **RMSE** |
| --- | --- |
| Full Model | 6097.208 |
| Model 1 | 5014.626 |
| Model 2 | 4970.917 |
| Log\_Model 1 | 18324.479 |
| Log\_Model 2 | 18324.484 |

The best prediction model that discovered so far is Model 2：

Charge = -328.35+26.74\*age +3.53\*(age-mean(age)^2)+ 661.57\*children -7.94\*bmi + z\*(-20594.56 + 1439.19\*bmi), where z is a dummy variable and defined as z = 0 if the ensured is not a smoker, otherwise z = 1. The coefficient of each explanatory variable represents the amount by which the charge is predicted to increase (or decrease) for a one-unit increase in that variable, holding all other variables constant.

## Conclusion

In this study, The forward selection algorithm was utilized to identify the most relevant predictor variables to include in the model. The optimal model was selected based on the lowest Bayesian Information Criterion (BIC) score, a reasonably high adjusted R-squared value, and the lowest RMSE value.

After evaluating multiple models, Model 2 was selected as the best-performing model, which includes the following predictor variables:

* Age: The age of the customer.
* (Age - mean(Age))2: The quadratic term of age, which allows for a non-linear relationship between age and insurance charges.
* Children: The number of children the customer has.
* Smoker: A binary variable indicating whether or not the customer is a smoker.
* Smoker\*BMI: The interaction term between smoker status and the customer's body mass index (BMI), which captures any non-additive effects of smoking and BMI on insurance charges.

By including these variables, we were able to capture their individual and combined effects on insurance charges, while accounting for any non-linear or non-additive relationships between these variables.

In particular, the quadratic term for age and the interaction term for smoker status and BMI allowed for more flexible relationships between these variables and insurance charges, resulting in a smaller Root Mean Squared Error (RMSE) and higher predictive accuracy compared to Model 1, which did not include the quadratic term for age.

Overall, our goal was to reduce the risk of overfitting the model while still capturing the most important predictors of insurance charges. With this optimized model, we can estimate insurance charges for new customers based on their age, smoker status, number of children, and BMI, while ensuring that our predictions are both accurate and reliable.

## Limitations：

Even so, the optimized model performs well and explains a large proportion of the variation in insurance charges. There are still some limitations of this study: First, the dataset has limited variables, so the model is based on a limited number of variables, which may not capture all of the factors that influence insurance charges. Other variables such as annual income, lifestyle, and pre-existing medical conditions could also have an impact on the insurance charges. Second, due to geographical differences, the prediction model may not be generalizable, since insurance charges may vary in different countries, which is not accounted for in the model. Lastly, the model is based on historical data and may not be accurate for predicting future charges, especially if there are changes in the healthcare industry or the emergence of a pandemic.

Last but not least, this study is still valuable since predicting insurance charges can be a meaningful tool for insurers, policymakers, and individuals seeking to understand the factors that influence the cost of healthcare. By analyzing historical data and building prediction models that take into account variables such as age, number of children, bmi, and smoking status, insurers can make informed decisions about the cost of providing medical coverage.