Multilinear Regression Model on Healthcare Charges Prediction

No. of Pages: 11

Contents

[Introduction 2](#_Toc90633125)

[Motivation and Research Questions 2](#_Toc90633126)

[Data Description 2](#_Toc90633127)

[Exploratory Data Analysis 6](#_Toc90633128)

[Data Analysis 8](#_Toc90633129)

[Resampling & Validation 10](#_Toc90633130)

[Discussion & Limitations 11](#_Toc90633131)

[Conclusion 12](#_Toc90633132)

[References 13](#_Toc90633133)

[Appendices 14](#_Toc90633134)

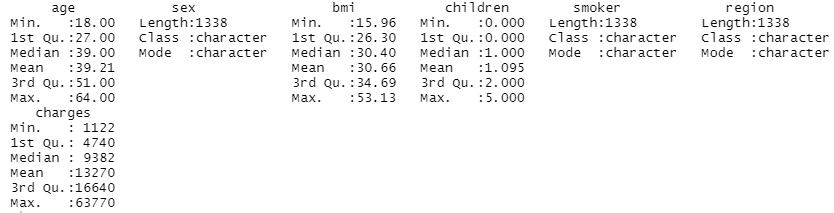
# Introduction

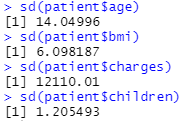
## Motivation and Research Questions

The recent and currently continuous Covid-19 pandemic has brought about tremendous catastrophe. Millions of lives have been taken away. During the pandemic, most countries experienced lack for beds in public hospitals while the expenses for private healthcare services is considerably higher so that they are usually covered by commercial insurances. Therefore, it is generally important for health insurance companies to set reasonable price by identifying key client features so as to keep the business profitable. Also, individuals can pay attention to the factors so that personal health conditions are under control based on the corresponding personal healthcare costs that may occur. This report analyses an insurance dataset, which contains a set of features of patients and the corresponding charges paid, by conducting exploratory data analysis, linear regression modelling, and cross validation approaches in order to construct a model that predicts personal healthcare costs by certain features. The dataset was formatted and uploaded by Caballero (2015) according to a work by Lantz (2013) via publicly available resources.

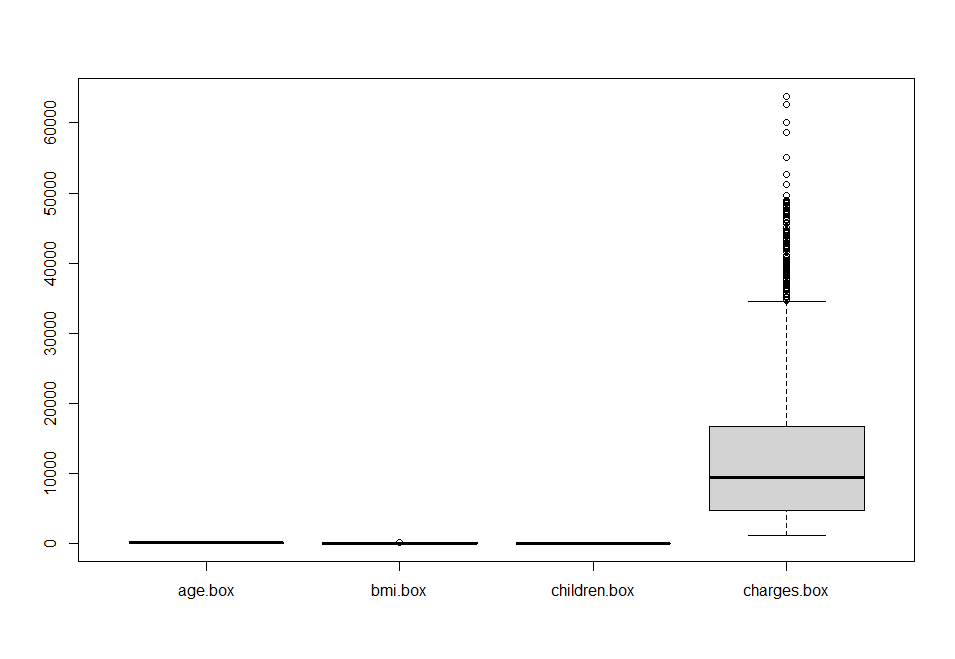
## Data Description

Overall, the dataset contains 9366 observations with 7 columns and 0 missing values. The head of the columns are age, the age of the insurance contractor (equivalently the patient), sex, patients’ sex, body mass index (BMI), a score indicating a patient’s thickness or thinness, children, the number of children a patient possess, smoker, indicating whether a patient smokes or not, region, the region of the patient’s residence in the US, and charges, the healthcare costs paid by the insurance company for the patient. It is not hard to determine that ‘charges’ should be the response variable and the others perform as predictors. A summary of the data is listed below:

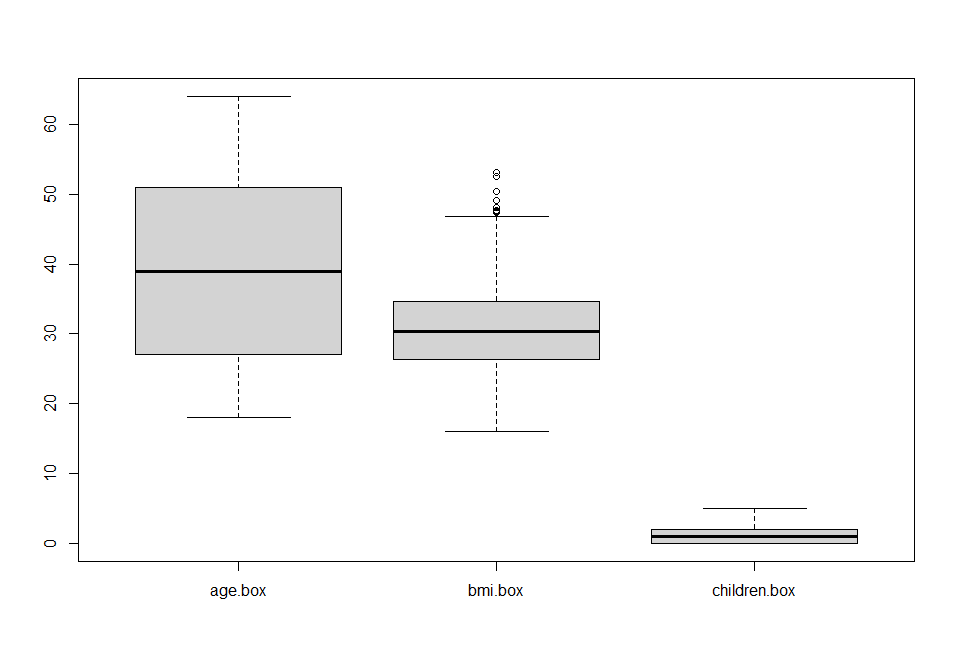




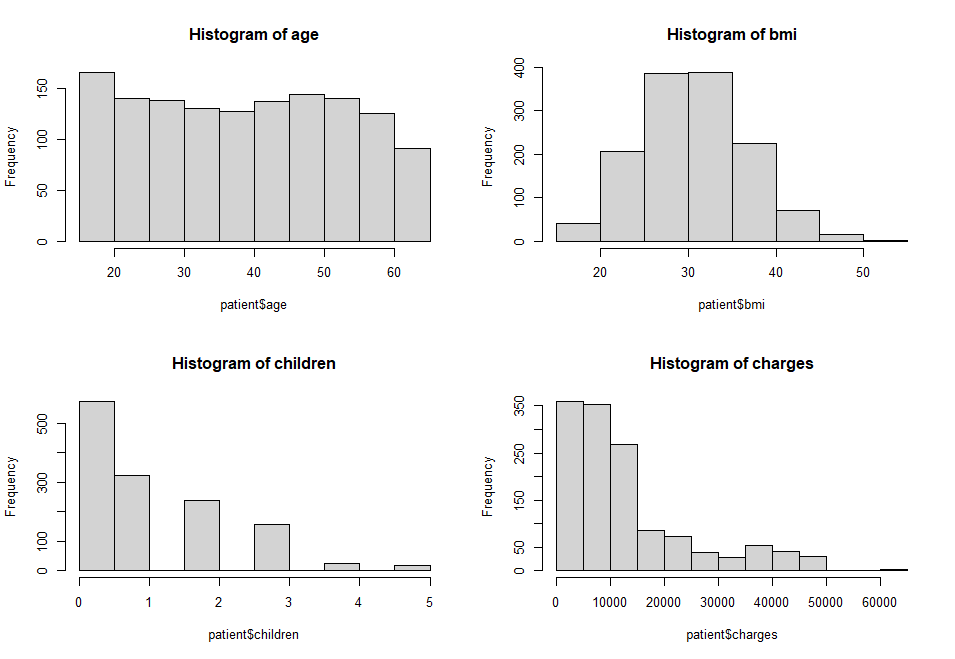
*Figure 1. Data Summary*



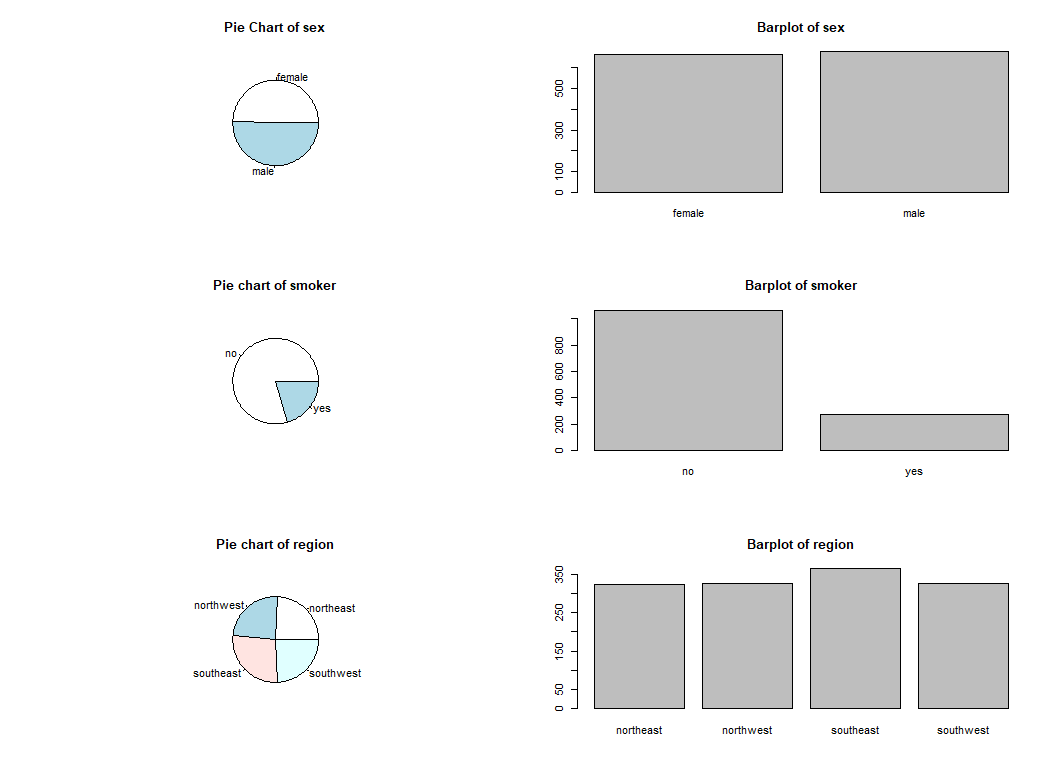
*Figure 2. Boxplots Overall*



*Figure 3. Boxplots in Specific*



*Figure 4. Histograms*

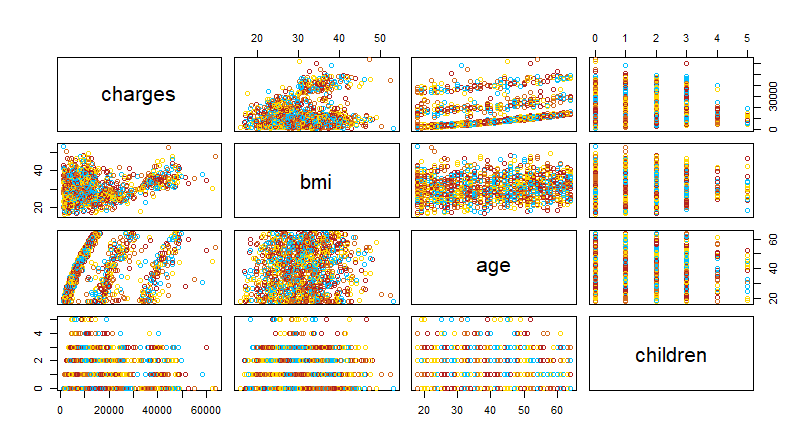


*Figure 5. Pie & Barplots*

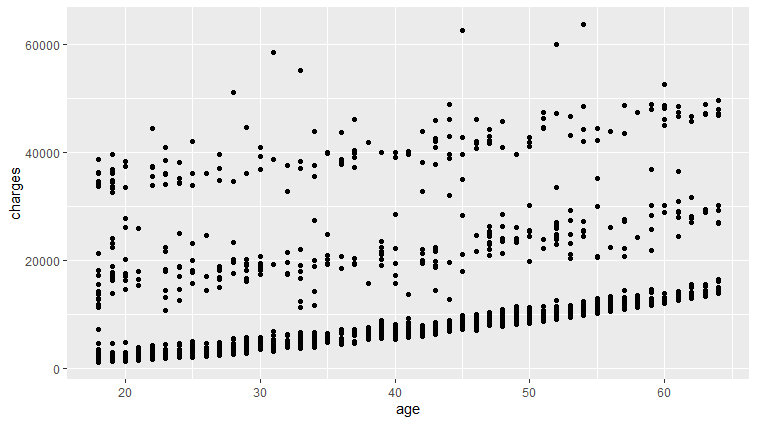
In terms of numeric data (e.g., age, bmi, children, charges), it can be seen that while bmi and age are centred with the mean around the median, charges and children appear a pattern of skewing to the right as indicated by both summary and histograms. Also, there seems to be some outliers in bmi and a significant number of outliers in charges according to figure 3 and 2 respectively. It is also the charges data that appears a large spread with standard deviation at 12110.01 while the others appear much lower scores according to figure 1. The corresponding pattern can be found in figure 4, where histograms indicate bmi to be in normal distribution.

In terms of categorical data (e.g., sex, smoker, region), it can be seen that while sex and region are roughly evenly distributed, the proportion of the smoking insurance contractors accounts to one fifth approximately according to figure 5.

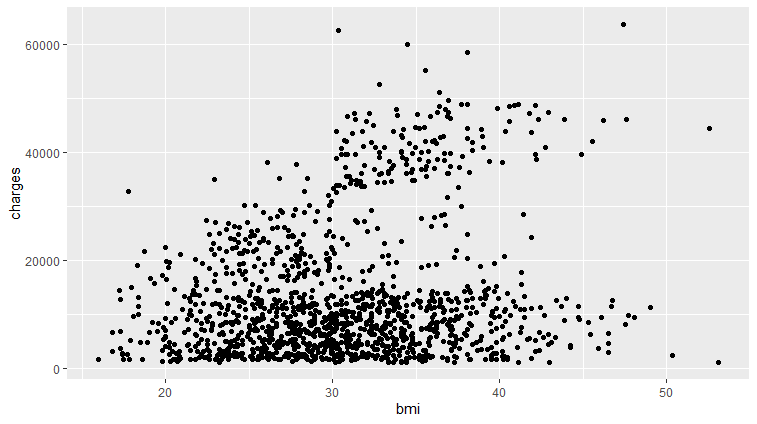
## Exploratory Data Analysis



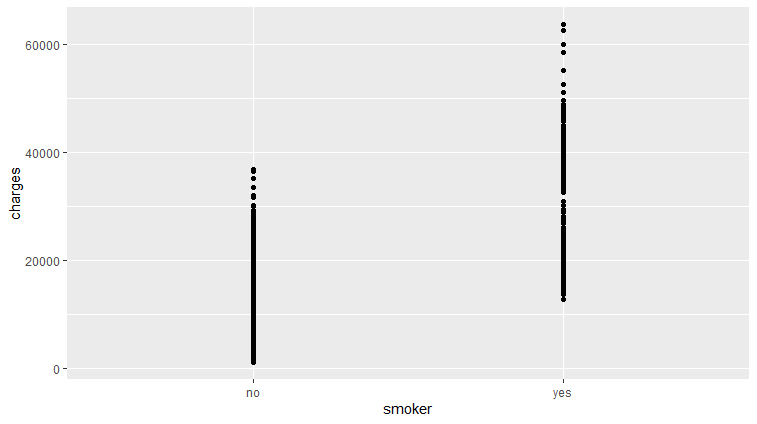
*Figure 6. Pairs Plot of Numeric Data*

**

*Figure 7. Scatter Points age vs charges*

**

*Figure 8. Scatter Points bmi vs charges*

**

*Figure 9. Scatter Point smoker vs charges*

For numerical data, the exploratory data analysis indicates that while age is somewhat correlated with charges, the corelation between charges and other numeric factors remain unclear. However, it is interesting to note that there seems to be a positive relation between charges and bmi in terms of charges as an independent variable and bmi as a dependent variable according to figure 6. Moreover, figure 6 also indicates that there is no obvious correlations between pairs of predictor variables.

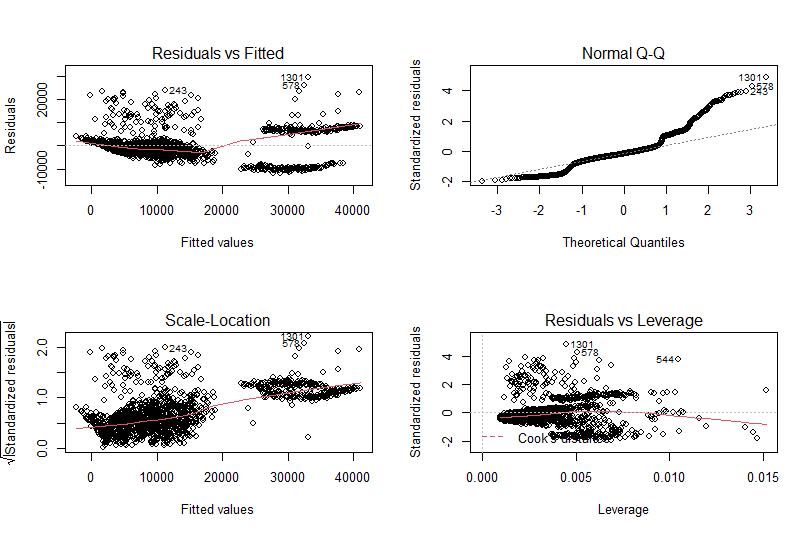
For categorical data, it is not surprising to notice that those who smoke generally cost higher healthcare expenses than people who do not, which corresponds with common sense. The scatter points of the other 2 categorical features are included in appendix 1 as they are not particularly interesting.

# Data Analysis

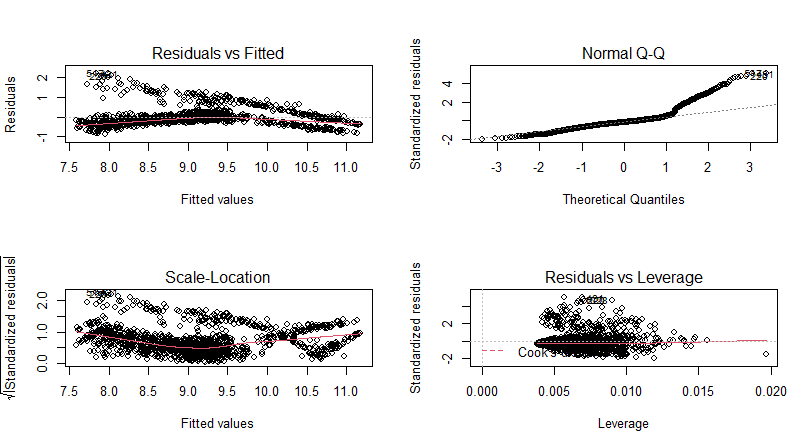
Multilinear regression is used in the modelling process. From the exploratory data analysis, it is indicated that there is no obvious correlation between charges and a specific independent variable. Therefore, instead of simple linear regression, multilinear regression is applied in this part. Considering that cross-validation is carried out, the whole dataset is used for training.

Overall, 3 models in total are built sequentially when the previous model presents obvious problems. For variable selection, the Akaike Information Criterion (AIC) is firstly used. Essentially, AIC returns a score indicating the relative information lost of a function, and thus indicates a desirable model when it returns a small value. In this report specifically, both a multilinear regression fit with no variable (e.g., containing intercept only), namely ‘null’, and a model with all variables, namely ‘full’ are created. Then, forward, backward, and stepwise selection are conducted, all of which returns the same result with AIC score at 23314.58. However, region is dropped for variable selection as the trained data summary indicates insignificance for its coefficients (e.g., lower than 90% for region northwest). The training model—model2 is therefore *lm(formula = charges ~ smoker + age + bmi + children, data = patient)* with adjusted R2 =0.7489 as opposed to 0.7496 for the original model1. However, as indicated by the diagnostic plots in figure 10, the model is obviously not a good one as non-linear patterns are shown by the residual plot, indicating a wrong modelling method for this data. Model3 is then constructed by taking log values at both sides of ‘full’ and repeating the variable selection process. The formula of model3 is *lm(formula = log(charges) ~ smoker + log(age) + log(bmi) + children + region + sex, data = patient)* with adjusted R2 at 0.7676. Most importantly, the residuals pattern, though not good enough, now appears to be linear as shown in figure 11 and model3 is therefore the final model. However, it should be noted that log transformation is not applied to ‘children’ variable as it contains values of ‘0’. The model summary is presented in figure 12.

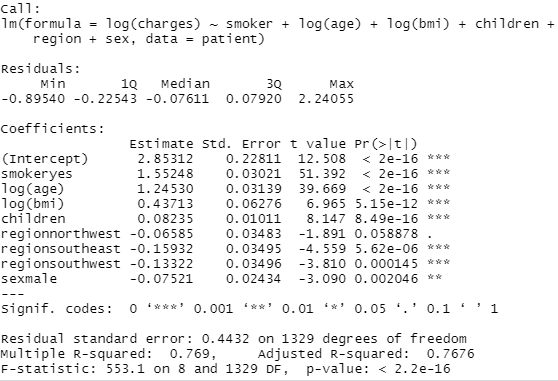
Despite the solution for linearity, the diagnostic plot for model3 indicates a series of problems. First, the residuals are not evenly spread around the fitted line, indicating a space for improvement. Second, the qqplot indicates a right-skewed residual distribution though log transformation is applied, suggesting that outliers may be present. Finally, the unequally spread residual points in scale-location plot implies the heteroscedasticity problem. In this case, ordinary least square errors method may not produce the best results as standard errors calculated via least squares could be inaccurate. Fortunately, the residual vs leverage plot indicates no influential cases as Cook’s distance line is not observed.

**

*Figure 10. Diagnostic Plots for Model2*

**

*Figure 11. Diagnostic Plots for Model3*

**

*Figure 12. Summary for Model3*

# Resampling & Validation

Leave one out cross-validation (LOOCV), K-fold cross-validation (K-fold CV) with 5 and 10 folds, and bootstrap are used to validate model3. For LOOCV and K-fold CV methods, considering that log transformation is applied in model building, the data in the column of the dependent variable (e.g., charges) of the training dataset are applied with log transformation while the others remain the same to create a dataset for resampling. Then, the formula for model3 is duplicated using *glm(*) function. Finally, validations are conducted by *cv.glm()* function, where adjusted mean square errors (MSE) are returned as output. Adjusted MSE from LOOCV, 5-fold CV, and 10-fold CV are 0.1954268, 0.1958661, and 0.1955007 respectively. For bootstrap method, the original training dataset is used for resampling and the resampling times are set to 1000. The method is realised by *boot()* function. Detailed results are presented below.

|  |  |
| --- | --- |
|  |  |

*Figure 13. Cross-validation Results*

The relatively small MSEs from LOOCV and K-fold CV indicate a good model fit. Similarly, the small bias presented in bootstrap statistics also indicate a good model selection. The difference between the standard errors from bootstrap (figure 13) and those in model3 summary (figure 12) corresponds with the heteroscedastic problem identified in the data analysis section. Overall, although the fitted model3 is not the best one due to the heteroscedastic problem, the results from cross-validation methods indicate model3 to be a good one.

# Discussion & Limitations

As shown in figure 12, the independent variables contain both numeric (e.g., age, bmi, children) and categorical (e.g., sex, smoker, region) variables. For numeric ones, the coefficient indicates changes contributed to the dependent variable when the independent variable changes marginally. Specifically, 1 percent increase in age and bmi will lead to 1.2453 percent increase and 0.43713 percent increase in charges respectively, whereas 1 percent increase in children will lead to 0.08235/100 = 0.00082 percent increase in charges. For categorical ones, the value of the coefficient(s) adds to the intercept when the corresponding variable(s) is (are) switched on. For example, if a female smoker resident from northeast were entered as an input, the intercept would be 2.85313 – 1.55248 = 1.30065. Also, to get the values for predicted charges, the exponential function should be applied to regression outputs.

Unsurprisingly, this model has some limitations. First, as mentioned in the data analysis section, log transformation is only applied to age and bmi variables, which shrinks the contribution from changes in children to changes in charges by 100 times. Second, the standard errors appeared in model summary is unreliable due to heteroscedastic problem. Standard deviations from bootstrap (figure 13) are recommended if they are needed. Finally, the training data only contains some healthcare insurance contractors from the United States. Therefore, the generalisation capability of the model to other places may be rather poor.

# Conclusion

To sum up, a multilinear regression model is built for predicting healthcare expenses given certain features of individuals based on the data from an insurance business in America, with no multicollinearity problem identified (figure 6). The modelling process adopted AIC method for variable selection and log transformation for linearity. Although the data is skewed to the right with heteroscedastic problem, cross-validation results suggest that it is a good model. However, it should be emphasised that the model will need extra modelling and training if it is to be applied to a different place other than the US.

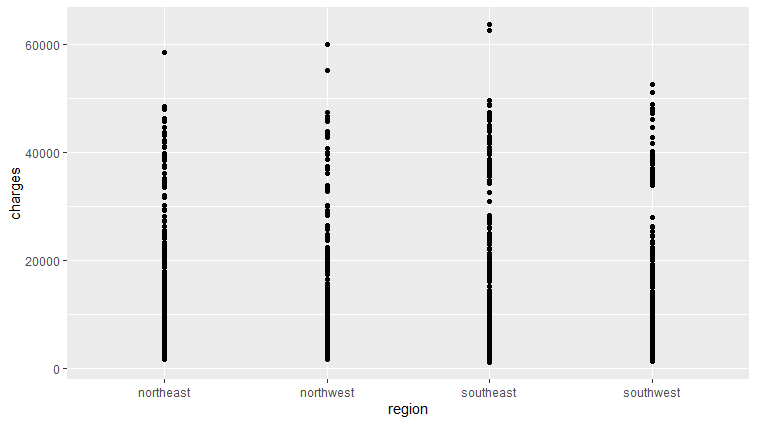
# References

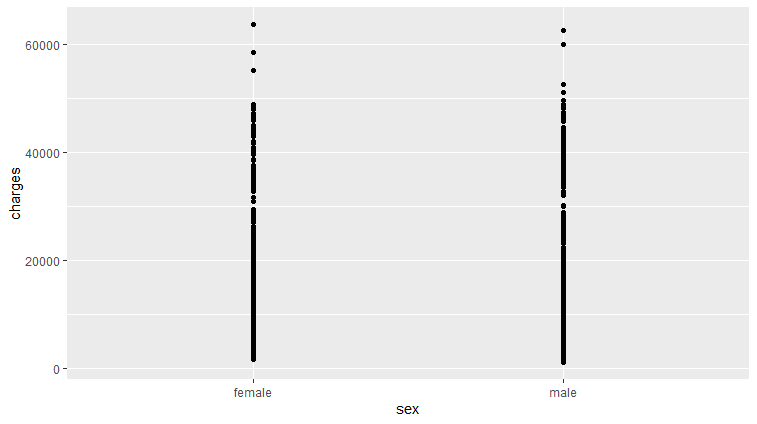
Caballero. (2015). Retrieved December, 2021, from https://github.com/stedy/Machine-Learning-with-R-datasets

Lantz, B. (2013). *Machine Learning with R*. Birmingham: Packt Publishing Ltd.

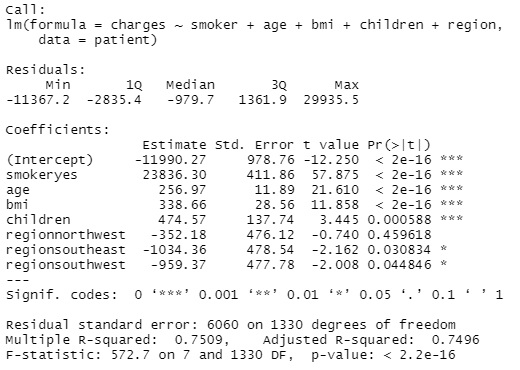
# Appendices

Appendix 1 Scatter point plots for region and sex against charges

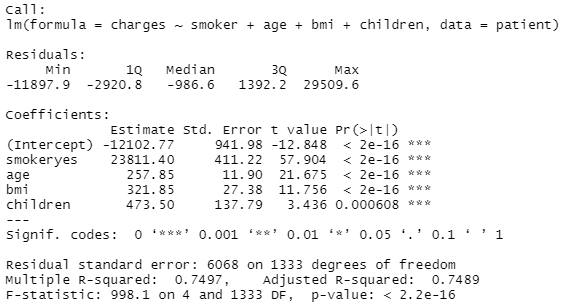




Appendix 2 Summary for Model1 and Model2



Summary for Model1



Summary for Model2