Factors Determining Personal Health Insurance charges*

Replicating 'Health Insurance Claim Prediction'

Zhihao Chen

April 26th, 2022

Abstract

The main aim of this paper was to determine relationship between personal health insurance charges and characteristics of the primary beneficiary of the insured. The demographic factors included; age, BMI, number of dependents and smoking status, while the geographic factor was the region of residence of the primary beneficiary.Cross-sectional data were obtained, and data description and visualizations were conducted to create a better insight into the data. The relationship between the variables was assessed using multiple regression techniques to achieve the aim of the study. Understanding the factors determining personal health insurance is vital for improving health and reducing longstanding differences in healthcare. There is no uniform amount of premium the insured is supposed to pay to the insurer, and the existence of these factors causes the variance in the monthly sum paid. It was found that the variables; age, BMI, smoking status and the number of children positively and significantly help in determining the amount of personal health insurance charges. These factors had a positive impact on the amount of payments. As the units of these variables increases, the amount of health insurance charges increases. Therefore, this study helped create a balanced criteria of determining the fair amount for both the insurer and the insured so that the money paid out will not exceed the money paid in.R-Studio software was used for analysis R Core Team (2022).

Keywords: linear regression, model, BMI, health.

Contents

*Code and data are available at: [github.com/ansonc](https://github.com/ansonc1437/sta304-final-project.git.

1 INTRODUCTION

Personal health insurance is the coverage purchased mainly to protect an individual or a family when they are sick. Personal health insurance cover can be obtained through an employer or government programs such as Medicaid or Medicare. On the other hand, private insurance cover refers to that coverage offered by a private entity instead of the federal government Rampal et al. (n.d.). There are social determinants affecting the health of the most vulnerable populations. The more an individual's health is risky, the more they are likely to pay more premiums than an individual whose health is considered stable. There are various factors determining personal health insurance charges, and they include; the age of the primary beneficiary, body mass index (BMI), the number of dependents covered by the insurance, region of the beneficiary such as the U.S, sex of the insurance contractor, if the primary beneficiary is smoking.

There are some considerations an individual ought to advocate for the future premiums to be paid because the insurer takes into account the total premiums paid out incase of the insurance claim. In the United States, there have been efforts to improve the health conditions of its citizens by reconsidering the traditional health care system as the vital driver of the outcomes of health status across all states. This has signified the increased recognition concerning improving and attaining healthy equity that demands border approaches that address economic, social, and environmental factors influencing health. Various researchers argue that health care is vital to health, and most health outcomes are motivated by the underlying factors related to genetics, social environment, and health behaviors of an individual Viganò et al. (2000). Similarly, other researchers argue that some other health behaviors, such as smoking, are the major drivers of health outcomes. In general, those individuals who did not complete their education are more likely to stay in those environments that are hazardous to their health due to staying in substandard houses.

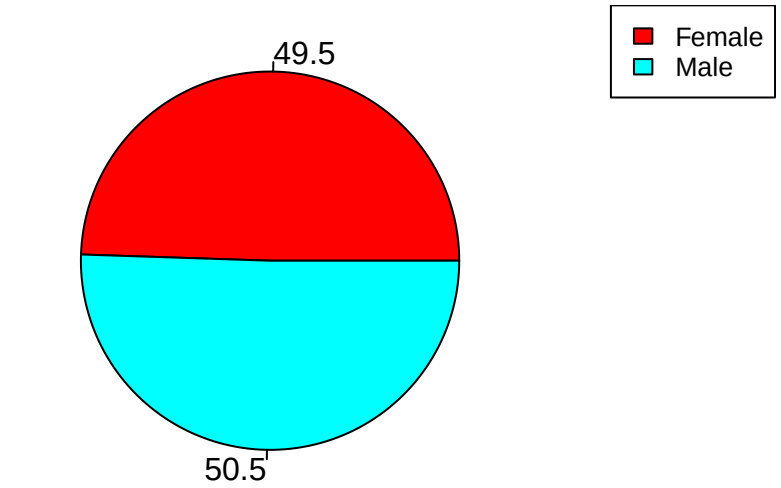
Understanding the factors determining personal health insurance is vital for improving health as well as reducing longstanding differences in healthcare. There is no uniform amount of premium the insured is supposed to pay to the insurer, and the existence of these factors causes the variance in the monthly sum paid. There is a maximum number the insured is supposed to indicate as the beneficiaries or the dependents of the policy, especially for the family but for the company, the number is dependent on the number of employees. Having a personal health insurance cover has numerous advantages, such as government income tax exemption, especially for the health benefits, as well as a wide range of options related to planning. Equally, an individual with personal health insurance cover is free from the high medical cost incurred by those individuals who don't have medical cover. The insured is likely to pay less for the covered health care before meeting the deductible amounts previously paid as premiums. The insured receives free preventive care such as screenings and routine checkups. Zhang, Fu, and Fang (2020)

2 DATA

2.1 DATA Source and introduction

The main aim of this study was to predict the amount of personal health charges based on the primary beneficiary demographic and geographical information. The dataset was obtained from the Kaggle website (https://www.kaggle.com/code/harshinisk/health-insurance-prediction/data). It contained seven variables and 1,338 observations. The individual characteristics contained in the datasheet include; their age, sex, BMI, number of children and smoking status, while the geographical feature was the region of residence of the primary beneficiary. The outcome variable "charges" and the predictor variables (age, BMI and number of children) were continuous, while the variables (sex, region and smoking status) were categorical. Based on these characteristics, an exploratory analysis was done. Finally, a multiple regression model was built to determine the relationship between the factors and predict the amount of health insurance charges. I analyzed it using R (R Core Team 2020), and packages tidyverse (Wickham et al. 2019),haven (Wickham and Miller 2020) . I used packages bookdown (Xie 2016), kableExtra (Zhu 2020), finalfit (Harrison, Drake, and Ots 2020), modelsummary (Arel-Bundock 2021), broom (Robinson, Hayes, and Couch 2020) to format the document and referenced Impact Evaluation in Practice (Gertler et al. 2016) to evaluate this experiment. I used Shiny (Chang et al. 2021) for enhancement to display interactive model results.

The dataset consisted of 49.5% females and 50.5% males. This was illustrated in the pie chart below.



The youngest respondent was 18 years of age, while the oldest was 64 years old.

2.2 original data

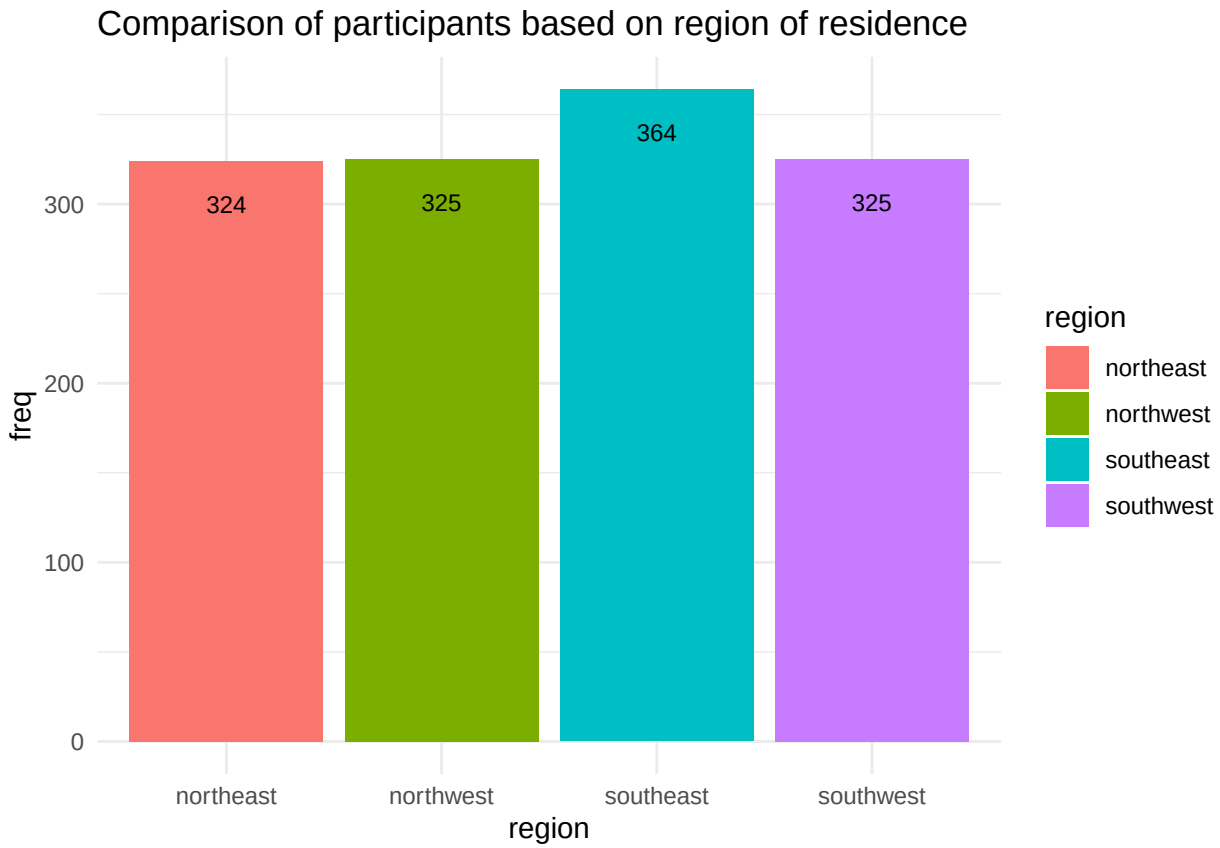
The important variables are in the following table.

2.3 Data EDA

For the body mass index of the participants varied between 15.96 kg/m2 and 53.13 kg/m2. Most of the respondents came from the south east while the participants from the other regions were equally sampled as indicated in the plot below;

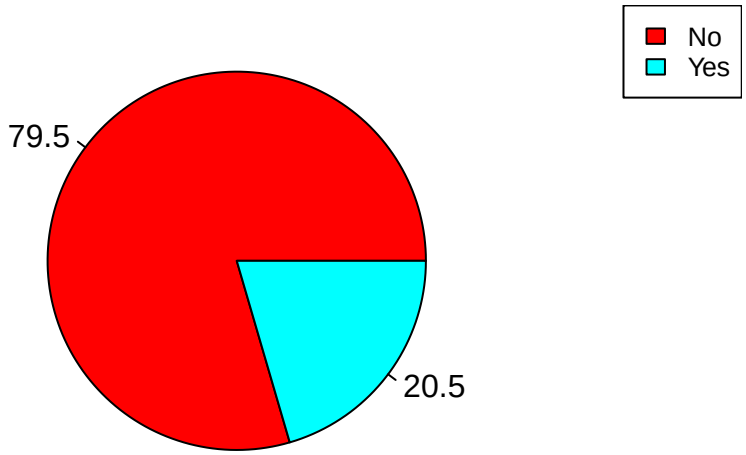
Table 1: Summary Table of the important variable

	age	sex	bmi	children	smoker
	Min. :18.00	Length:1338	Min. :15.96	Min. :0.000	Length:1338
	1st Qu.:27.00	Class :character	1st Qu.:26.30	1st Qu.:0.000	Class :character
	Median :39.00	Mode :character	Median :30.40	Median :1.000	Mode :character
	Mean :39.21	NA	Mean :30.66	Mean :1.095	NA
	3rd Qu.:51.00	NA	3rd Qu.:34.69	3rd Qu.:2.000	NA
	Max. :64.00	NA	Max. :53.13	Max. :5.000	NA

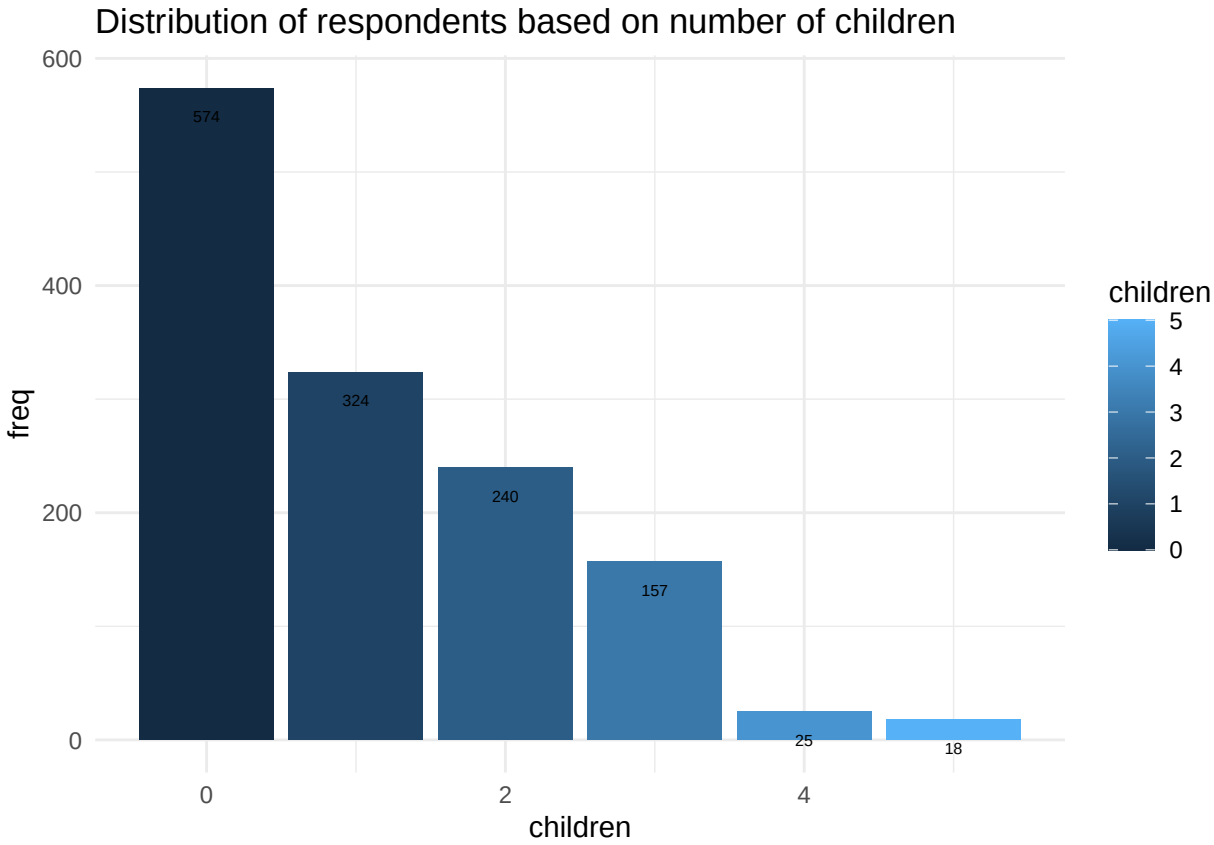


The dataset consisted of 79.5% non-smokers with only 20.5% of the respondents were smokers. The illustration was as below;

Distribution based on smoking status



From the bar chart below, most of the participants had no children.



As observed, the charges distributions in different regions and sex are different. So region and sex may be two potential predictors.

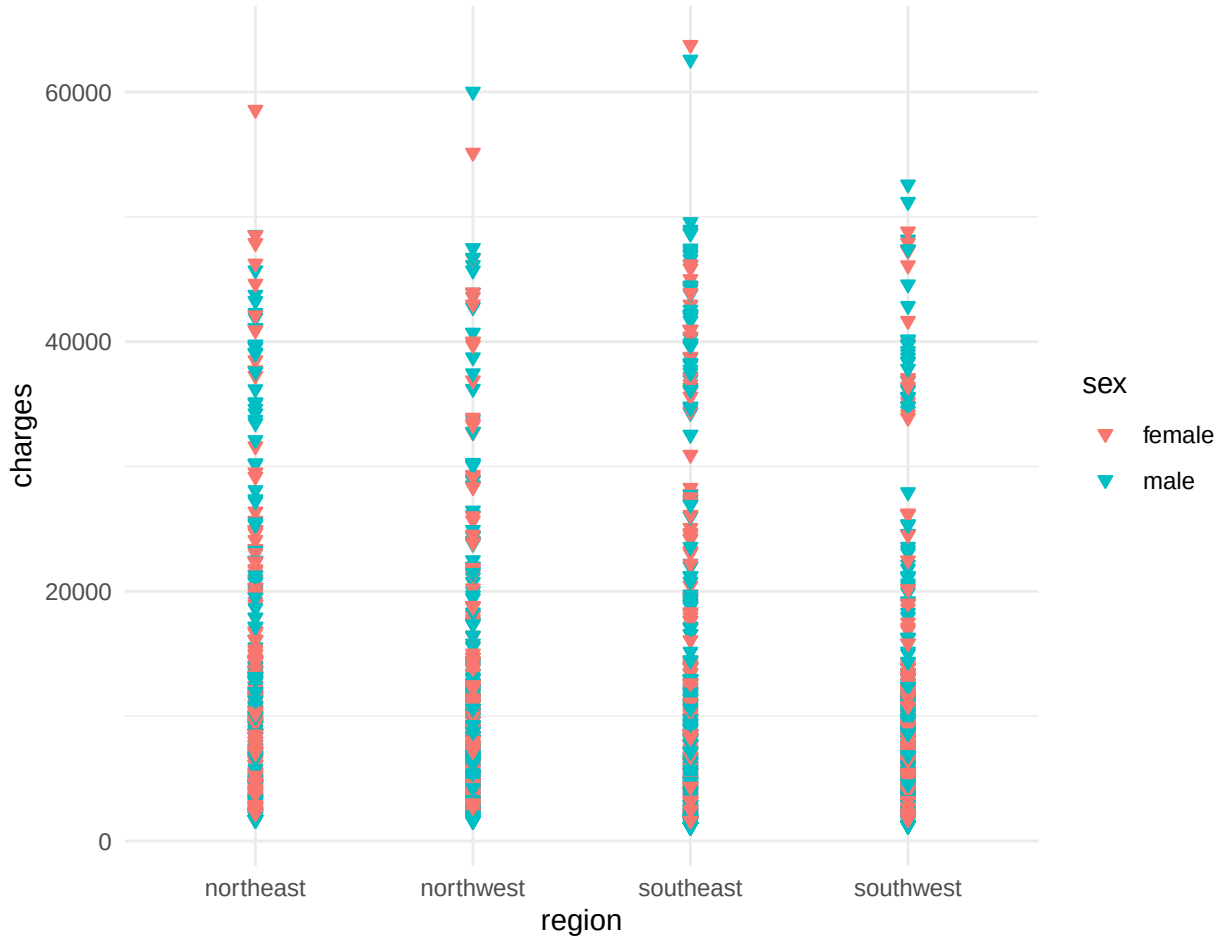


Figure 1: Charges distribution in different regions and sex

3 MODEL

Exploratory data analysis was conducted initially involving the investigations of the data to discover the patterns and checking for anomalies. These were achieved with the help of summary statistics and graphical representations. This process helps unearth the main characteristics of the data. (Weisberg 2005), (Su, Yan, and Tsai 2012), (Uyanık and Güler 2013) The dataset was first split into train and test sets. The training set was used to fit the model while the testing set was used to evaluate the model performance and make predictions. A multiple linear regression was considered as it analyzes the relationship between a single dependent variable and several independent variables.This aligned with this study varibales as there was one response variable and six independent variables.

The multiple regression model was formulated as follows:
$Y_i = X_i\beta + \epsilon_i$ Where: Y_i:dependent variable.
X_i:explanatory variables
ϵ follows normal with mean 0 and variance $sigma^2$.

The detailed steps are as follows:

3.1 EDA

First of all, all the EDA are done to see the relationship between response and the predictors. And the numerical summary of the numerical variables is given. And from the scatter plot, we know which predictors should be added in the model at first. And also the common sense also could tell us which variables should be added.

3.2 Automated selection

Using stepwise selection by BIC and AIC, the system will choose a model for us. And this model may not satisfy the model assumption and we need to check if it pass all the model checking conditions.

3.3 Model checking

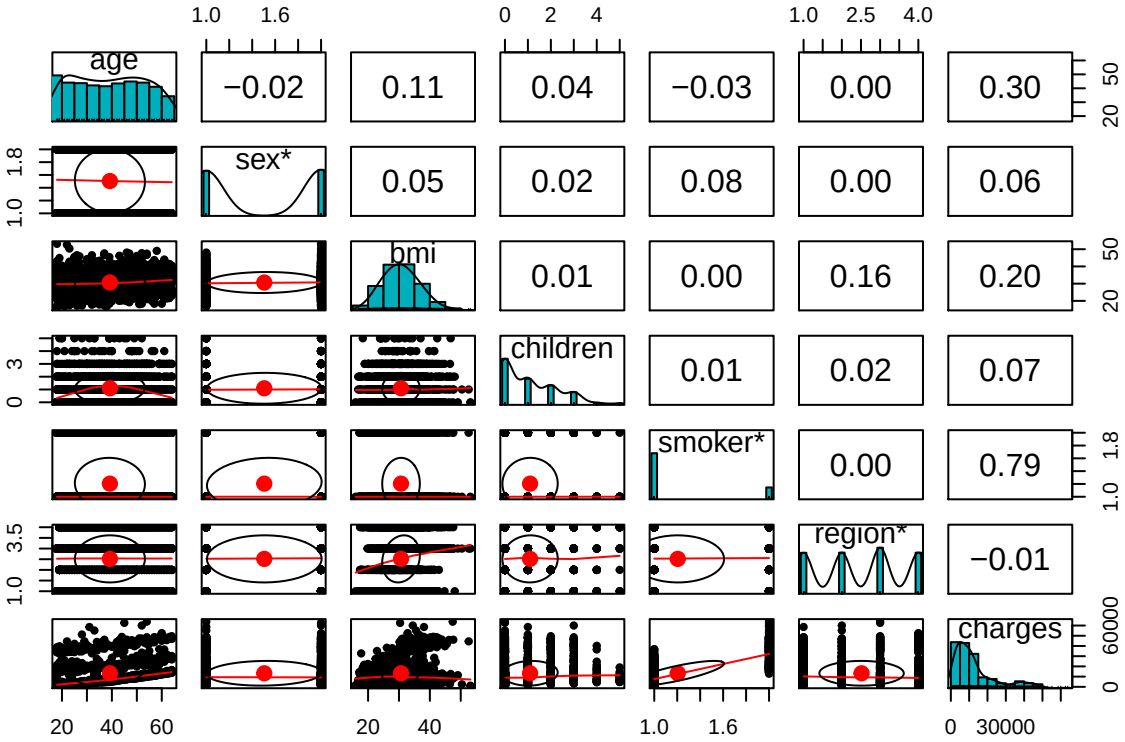
For residual plot, we need the residual plot to be no pattern. There are two kinds of residual plot. The first is between residuals and predictors. The second is between residual and fitted values. This will lead that the model is constant variance. Besides, the model is linear. If there is no cluster pattern in the residual plot, then the observations are independent.Then by normal QQ plot, if the QQ plot behaves like a straight line. This means the normality is ok for our model. And then leverage points, outlier and influential points are calculated and they are analysis in the limitation parts. At the final model, we check the VIF to see if there are correlations between predictors. (Yan and Su 2009), (Tranmer and Elliot 2008),(Olive 2017)

3.4 Model validation

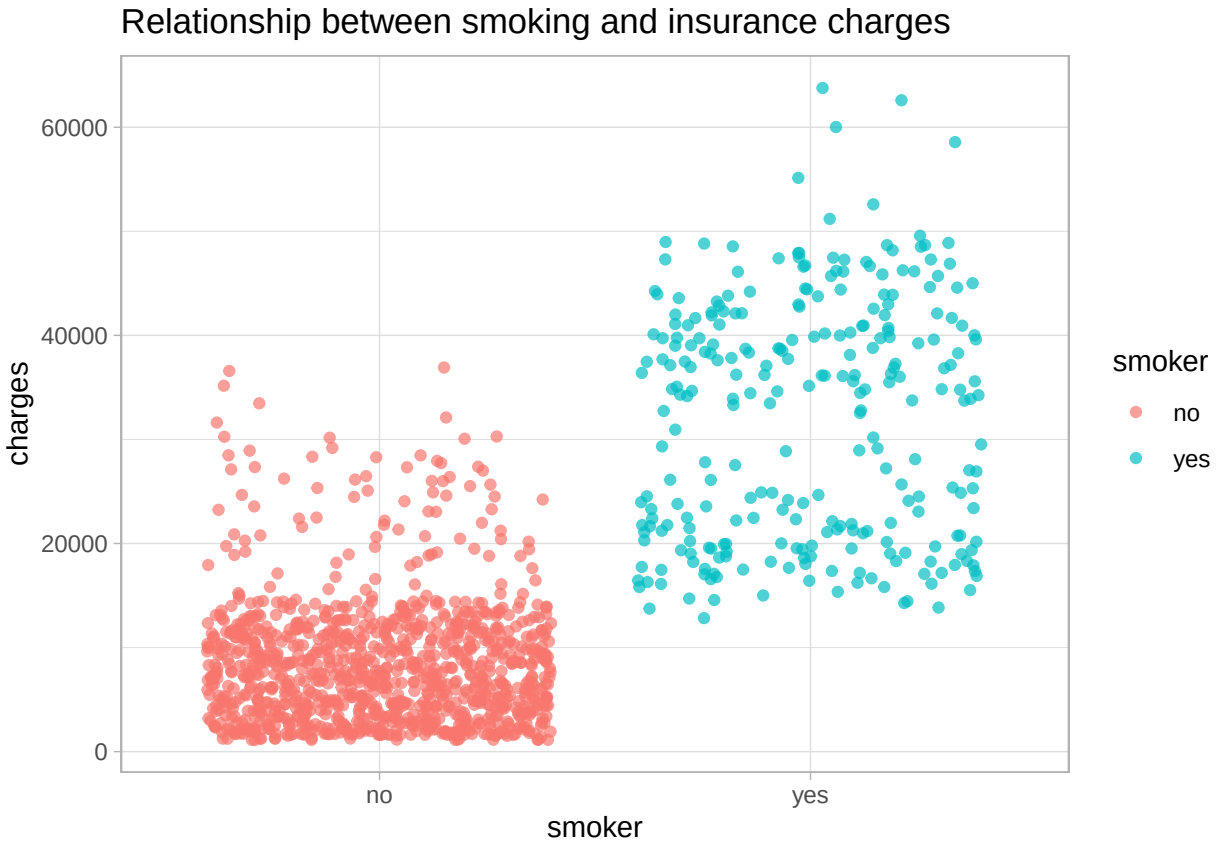
Further, we want to often build a model that could be used for predictions on data that is independent of the data we have already collected. We want to ensure that the model we land on as preferred is a good model not only on the sample we have collected, but on others from the same population. For the estimated coefficients, we can expect some differences, but we want to make sure the differences are not much bigger than the standard error of each coefficient in the training data. If we do not notice any substantial changes, then we can safely say your model has been validated. (Grégoire 2014)

4 RESULTS

4.1 Exploratory data analysis



The above plot was used to access the association between the predictors and the outcome. As age and body mass index increase, the amount of health insurance charges also increases. Based on the correlation coefficients, there seemed to be a negligible relationship between the charges and the explanatory variables; sex, number of children and region residence.



4.2 Fitting Multiple regression model

Table 2: Summary Table of Initial Model

term	estimate	std.error	statistic	p.value
(Intercept)	-11965.683	1120.092	-10.683	0.000
age	264.416	13.597	19.447	0.000
sexmale	4.311	377.441	0.011	0.991
bmi	334.710	32.034	10.449	0.000
children	375.393	156.098	2.405	0.016
smokeryes	23559.682	471.169	50.003	0.000
regionnorthwest	-428.251	542.601	-0.789	0.430
regionsoutheast	-1217.144	542.219	-2.245	0.025
regionsouthwest	-1275.059	547.323	-2.330	0.020

From the initial model, the variable sex and region were found to be statistically insignificant. Therefore, the best model could be built by dropping the insignificant variables as conducted below;

Table 3: Summary Table of Final Model

term	estimate	std.error	statistic	p.value
(Intercept)	-12167.789	1059.853	-11.481	0.000
age	265.839	13.604	19.542	0.000
bmi	315.373	30.555	10.321	0.000

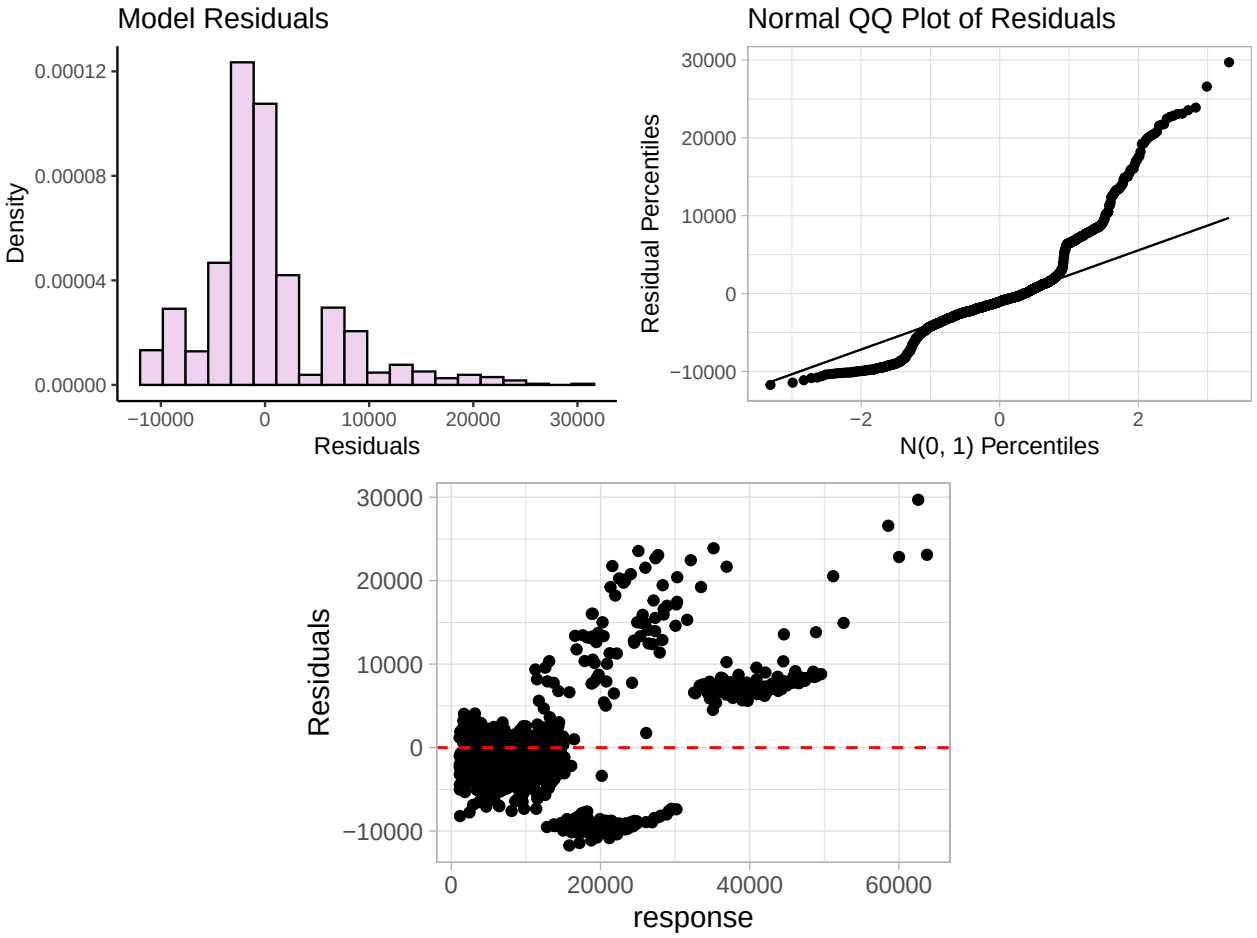
term	estimate	std.error	statistic	p.value
children	373.639	156.168	2.393	0.017
smokeryes	23530.944	468.150	50.264	0.000

The final model fitted above, had all the variables significant. The results indicate that for every additional year increase in the age of an individual, the health insurance charge increase by about 254.05 dollars. The insurance charges increases by approximately 295.17 dollars with a 1kg/mˆ2 increase in the individual's BMI. Smoking increases the amount of health insurance fee by about $24203.62. Finally, for an additional child as a dependent increases the health insurance charge by 540.34 dollars.

4.3 modeling checking

The multiple linear regression should be under the following assumptions:

1. $\varepsilon_i \sim N(0, \sigma^2)$ where $\varepsilon_i = y_i|x - \mu_{Y|x}$
2. ε_i occur independently
3. As a result, the random observations Y_i can be modeled as $Y_i|x \sim N(x\beta, \sigma^2)$



By the model checking, we found that there is no pattern for the residual plot, so the constant variance is ok. And the mean of the residual is zero. The normal QQ plot behaves not like a straight line, so normality is a bit violated. By the above scatter plot, linear relationship is good and there is no curve pattern. So the linear model is good. And there are some outliers from the normal QQ plot. And there are some influential

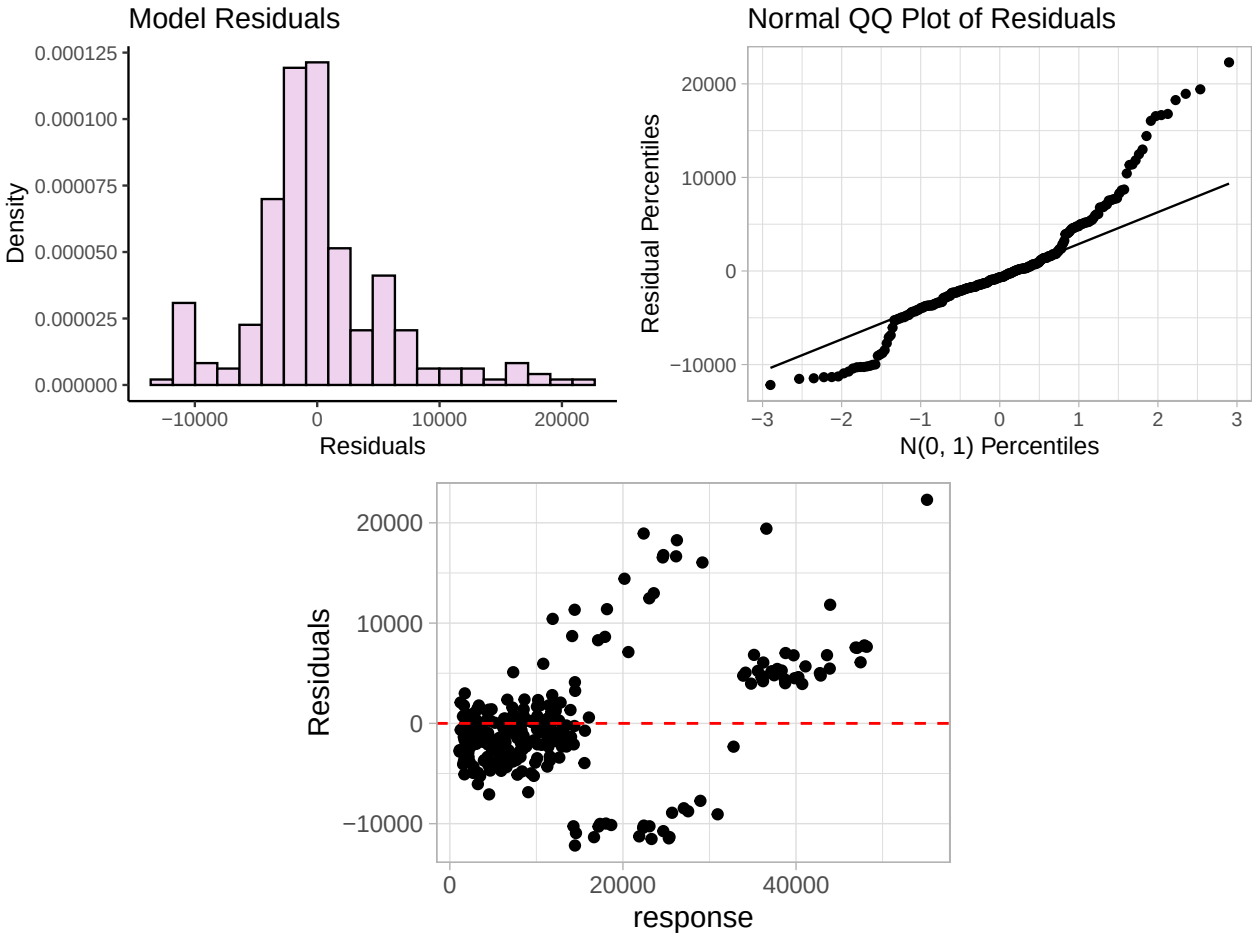
points. These points will be exlplained rather than deleting if these points are valid.

4.4 modeling validation

The coefficients for the testing model is very similar to the training model.

Table 4: Summary Table of Validation Model

term	estimate	std.error	statistic	p.value
(Intercept)	-12135.548	2071.182	-5.859	0.000
age	227.990	24.226	9.411	0.000
bmi	351.536	61.976	5.672	0.000
children	918.972	289.497	3.174	0.002
smokeryes	24961.537	852.991	29.264	0.000



The model checking plot is quite similar to the plot of the training model.

5 DISCUSSION

5.1 conclusion

Based on the results above, it was found that the variables; age, BMI, smoking status and the number of children significantly help in determining the amount of personal health insurance charges. These factors had a positive impact on the amount of payments. The study notes that the health insurance charges increase with an additional year in the age of a primary beneficiary of health insurance. This could be because as a person become older, they get closer to their life expectancy and the risk taken by the insurer also increases, thus an increase in the charges. As the body mass index increases, it results in higher insurance charges. The main reason is that BMI helps indicate an individual's health status. A person with a higher BMI stands at a higher risk for diseases; therefore has to make regular visits to the hospitals for weight-related issues.

Personal health insurance charges was much higher for individuals who smoke than for non-smoking beneficiaries. This could be attributed to the belief that smokers are more likely to develop health issues, thus resulting in more risk for the insurer. Furthermore, a higher number of dependents (children) leads to higher health insurance charges because the number of individuals to be covered increases hence greater risk for the insurance company. This study helped create balanced criteria for determining the fair amount for both the insurer and the insured. This is because when individuals are charged based on the risk helps ensure that the money paid out will not exceed the money paid in.

In practical meaning: Insurance costs will be lower for young people because they are less likely to get sick and have accidents. As you get older, the cost of insurance increases because insurance companies think older people are more likely to get sick and have accidents. Insurance costs for smokers will also increase because smoking is prone to cardiovascular and cerebrovascular diseases. This will increase the probability of insurance compensation, and the corresponding insurance amount will also increase. People with a higher BMI also lead to larger insurance amounts. Because BMI is an important indicator of physical health. The higher the BMI, the more obese the person will be. In this way, there will be more diseases caused by obesity. This will undoubtedly increase the amount of insurance.

5.2 limitation

The first limitation is that there are some outliers which are observed from the normal QQ plot. In the future, the model improve the performance of the model by replacing or eliminating the outliers for non-smokers. And after the outliers are removed, the QQ plot will be better. The second limitation is that our model may not be the best models. In the future, the transformation may be considered and more models could be compared.

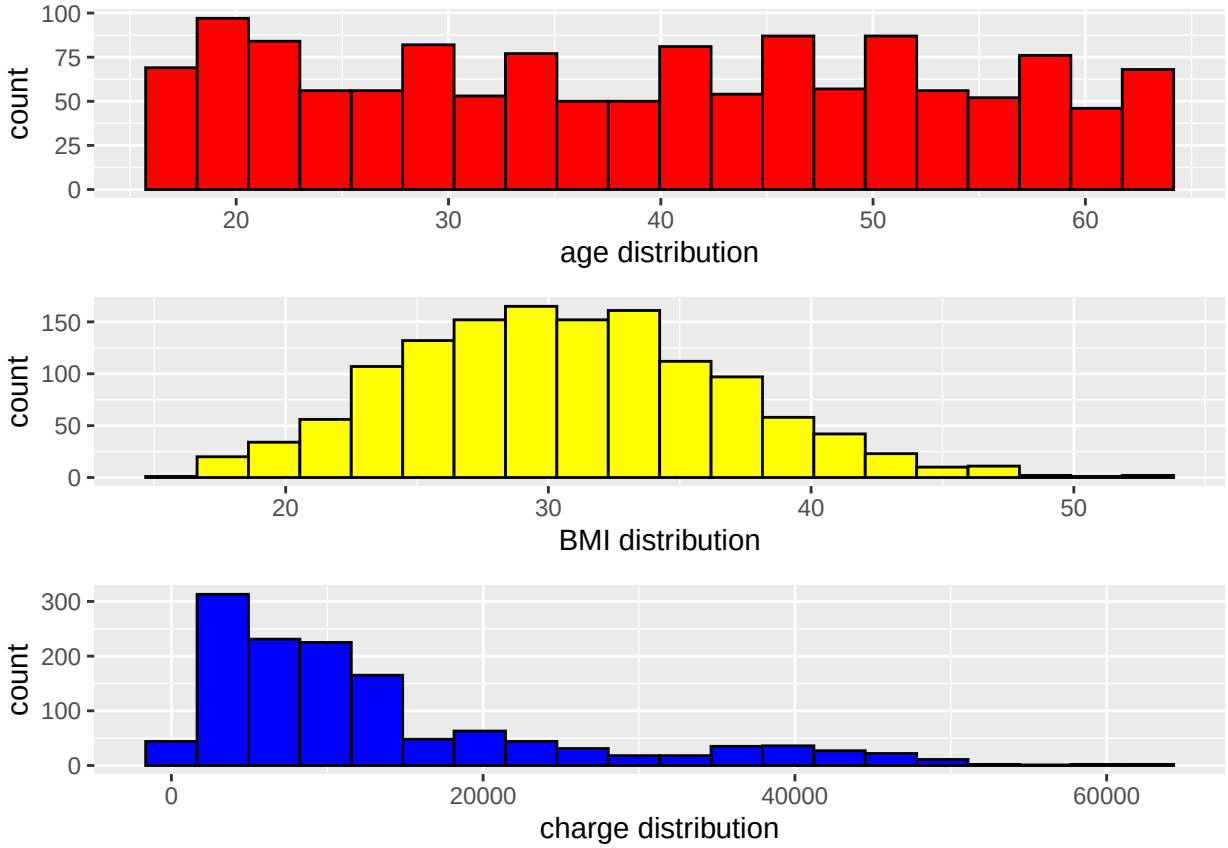
6 Appendix A

All the data statistics table.

```
##
##                         Overall
##   n                        1338
##   age (mean (SD))         39.21 (14.05)
##   sex = male (%)            676 (50.5)
##   bmi (mean (SD))         30.66 (6.10)
##   children (mean (SD))     1.09 (1.21)
##   smoker = yes (%)          274 (20.5)
##   region (%)
##      northeast              324 (24.2)
##      northwest              325 (24.3)
##      southeast              364 (27.2)
##      southwest              325 (24.3)
##   charges (mean (SD))  13270.42 (12110.01)
```

7 Appendix B

More EDA results.



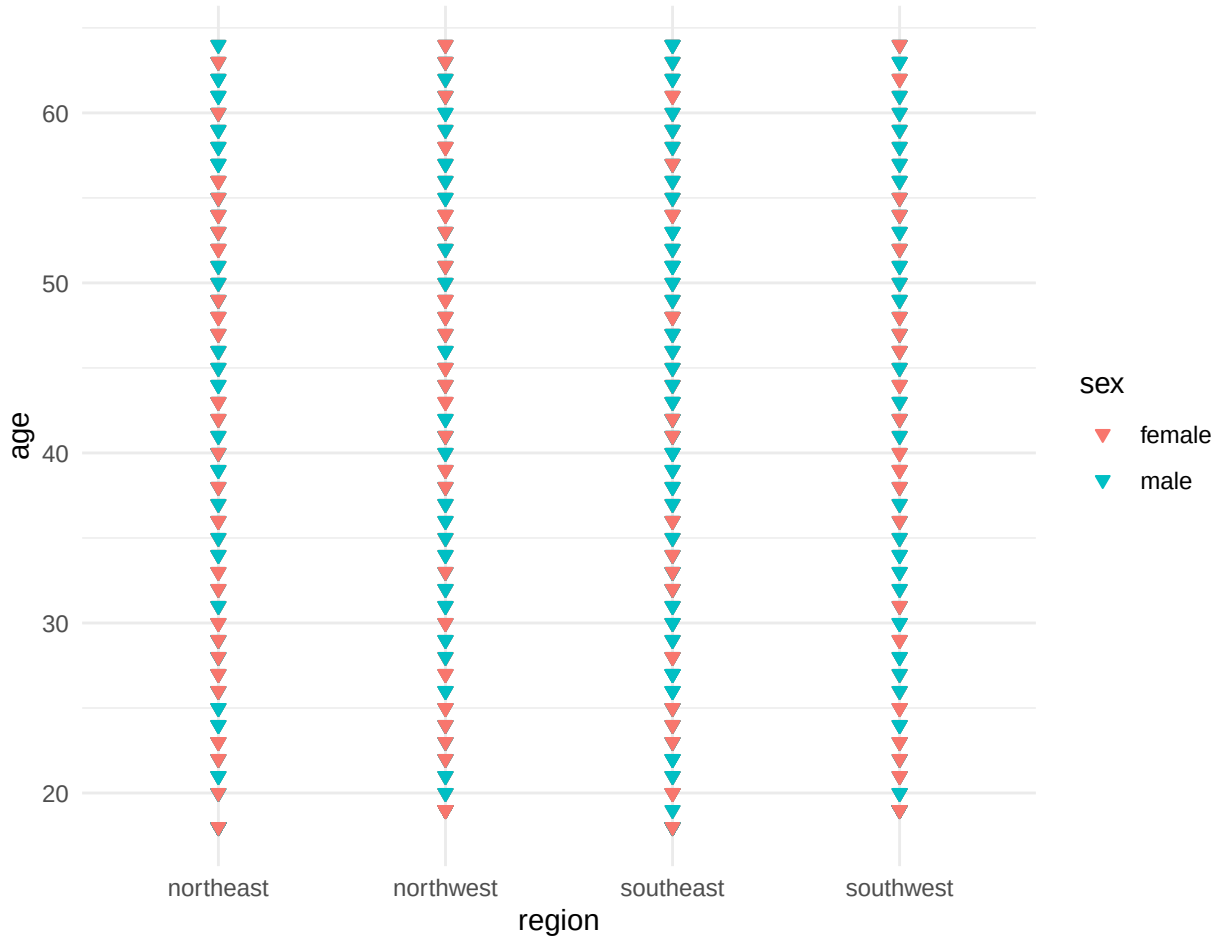
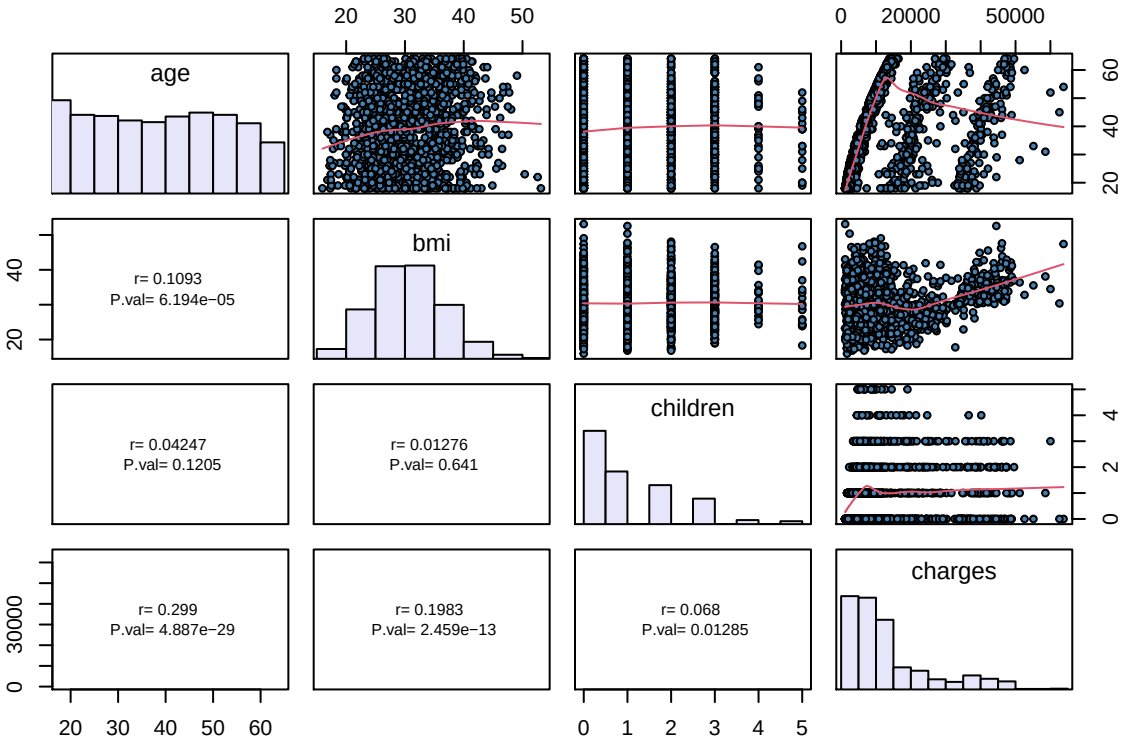


Figure 2: age distribution in different regions and sex

8 Appendix C

Correlation results.



References

Grégoire, Gérard. 2014. "Multiple Linear Regression." *European Astronomical Society Publications Series* 66. EDP Sciences: 45–72.

Olive, David J. 2017. "Multiple Linear Regression." In *Linear Regression*, 17–83. Springer.

Rampal, Jubeena, Prabhdeep Singh, Rajbir Kaur, and Kirandeep Singh. n.d. "An Ensemble Model to Predict Health Insurance Premium Using Machine Learning."

R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Su, Xiaogang, Xin Yan, and Chih-Ling Tsai. 2012. "Linear Regression." *Wiley Interdisciplinary Reviews: Computational Statistics* 4 (3). Wiley Online Library: 275–94.

Tranmer, Mark, and Mark Elliot. 2008. "Multiple Linear Regression." *The Cathie Marsh Centre for Census and Survey Research (CCSR)* 5 (5): 1–5.

Uyanık, Gülden Kaya, and Neşe Güler. 2013. "A Study on Multiple Linear Regression Analysis." *Procedia-Social and Behavioral Sciences* 106. Elsevier: 234–40.

Viganò, Antonio, Marlene Dorgan, Jeanette Buckingham, Eduardo Bruera, and Maria E Suarez-Almazor. 2000. "Survival Prediction in Terminal Cancer Patients: A Systematic Review of the Medical Literature." *Palliative Medicine* 14 (5). Sage Publications Sage CA: Thousand Oaks, CA: 363–74.

Weisberg, Sanford. 2005. *Applied Linear Regression*. Vol. 528. John Wiley & Sons.

Yan, Xin, and Xiaogang Su. 2009. *Linear Regression Analysis: Theory and Computing*. World Scientific.

Zhang, Liangwen, Sijia Fu, and Ya Fang. 2020. "Prediction the Contribution Rate of Long-Term Care Insurance for the Aged in China Based on the Balance of Supply and Demand." *Sustainability* 12 (8). Multidisciplinary Digital Publishing Institute: 3144.