**Identify factors that Have Effects on affect medical expenses**



**Student Name:**

Hao Wu

Xingyu Yi

Mengying Yuan

Shuai Feng

**Instructor Name:** Derek Bingham

**Course:** STAT350

**GitHub Link：** https://github.com/shuaif617/STAT350FInalProject.git

**Table of Content:**

[Abstract 3](#_Toc58287972)

[Introduction 4](#_Toc58287973)

[Dataset 4](#_Toc58287974)

[Response Variables 4](#_Toc58287975)

[Explanatory Variables 4](#_Toc58287976)

[Analysis to do 5](#_Toc58287977)

[Methods and models 6](#_Toc58287978)

[Results 9](#_Toc58287979)

[Scatterplots Matrix Study 9](#_Toc58287980)

[Analysis on the full model 9](#_Toc58287981)

[Model Selection 11](#_Toc58287982)

[Final model analysis 13](#_Toc58287983)

[Conclusion and discussion 15](#_Toc58287984)

[Reference 17](#_Toc58287985)

[Appendix 18](#_Toc58287986)

# Abstract

In this article, we will study to determine the factors affecting medical costs in the United States. For the data set under study, 1338 data and 7 variables are collected, which are the response variable expenses and the 6 explanatory variables we want to study. After research for scatter plot matrix and complete linear regression model analysis, we used the AIC and BIC model selection steps and selected two models. Then we build comparative tests on these two selected models. Through the use of ANOVA, the final model was determined and showed the influence of medical expenses on the following factors: 7 explanatory variables, 11 interaction variables, of which 7 influencing factors are more important, namely SEX, CHILDREN, AGE square, the interaction term BMI & smoking, BMI & living in SOUTHEAST, smoking & BMI over 30, smoking, and age square.

**Keywords:**

Medical expenses, AGE, SEX, Body Mass Index (BMI), CHILDREN, Smoking or not, Area of residence

# Introduction

## Dataset

The data set contains hypothetical medical expenses for patients in the United States. These data use demographic data from the US Census Bureau. It contains 1338 rows of data and the following columns: age, gender, BMI, children, smoker, region, charges, that is, the characteristics of the patient's characteristics and the total medical expenses calculated over the years.

## Response Variables

Charges

## Explanatory Variables

### Variable definitions

The third to fifth variables are considered important explanatory variables and must be studied in the regression model.

1. AGE: Indicates the age of the main beneficiary (not including those over 64, because they are generally paid by the government) (integer)
2. SEX: The gender of the policy holder (categorical variable: 1 = male, 0 = female)
3. BMI: Body Mass Index (BMI),which provides a way to judge whether a person’s weight is overweight or underweight relative to height. The BMI index is equal to the weight (kg) divided by the height (m) square, a ideal BMI is in the range of 18.5~24.9(float)
4. CHILDREN: Indicates the number of children/dependents included in the insurance plan (integer)
5. SMOKER: A categorical variable with a value of yes or no, indicating whether the insured person smokes frequently (categorical variable: 1 = yes, 0 = no)
6. REGION: According to the beneficiary's residence in the United States, it is divided into 4 geographic regions (categorical variable: northeast, northwest, southeast,

southwest), since there are multiple values, 3 variables are formed, namely

REGION(NORTHWEST), REGION(SOUTHEAST), REGION(SOUTHWEST). (categorical variable: 1 = this region, 0 = not this region)

### Variable statistics

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Table 1.1: Statistical description of continuous variables | | | | |
|  | AGE | BMI | CHILDREN | CHARGES |
| Min. | 18 | 15.96 | 0 | 1122 |
| 1st Qu. | 27 | 26.3 | 0 | 4740 |
| Median | 39 | 30.4 | 1 | 9382 |
| Mean | 39.21 | 30.66 | 1.095 | 13270 |
| 3rd Qu. | 51 | 34.69 | 2 | 16640 |
| Max. | 64 | 53.13 | 5 | 63770 |

From Table 1.1, the average value of Age is 39.21 and the 3rd Qu. is 51; the average BMI is 30.66 and the 3rd Qu. is 34.69; the average CHILDREN is 1.095, and the 3rd Qu. is 2; the average of CHARGES is 13270, and the 3rd Qu. is 16640.

|  |  |  |  |
| --- | --- | --- | --- |
| Table 1.2: Statistical description of categorical variables | | | |
| Variables | Value | Number of people | Percentage of total |
| SEX | FEMALE | 662 | 49% |
| MALE | 676 | 51% |
| SMOKER | NO | 1064 | 80% |
| YES | 274 | 20% |
| REGION | NORTHEAST | 324 | 24% |
| NORTHWEST | 325 | 24% |
| SOUTHEAST | 364 | 27% |
| SOUTHWEST | 325 | 24% |

From Table 1.2, SEX and REGION are evenly distributed, and 20% of SMOKER is smoking.

Combining Table 1.1 and Table 1.2, since our predicted target is CHARGES, we choose the larger value of Age, BMI, and CHILDREN, and take the corresponding 3rd Qu., Age=51, BMI=34.69, CHILDREN=2, SEX=MALE, SMOKER=YES, REGION=NORTHWEST as the new observation sample (Data Point).

## Analysis to do

1. Create features, build models, screen and compare the models, and determine a suitable linear regression model as the final model.

2. Analyze the goodness of fit of the final model.

3. Interpret the results of the final linear regression model to explain the variables and explain the meaning of the variables.

# Methods and models

In this paper, the multiple linear regression models were used to analyze the medical expenses data. Generally, when analyzing the data using the multiple linear regression models, the response variables would be considered as a linear function of the explanatory variables with an error variables term ε. If the data are appropriate to use the linear regression models, it should generally follow five assumptions:

1. The relationship between the response variable and the explanatory variable should be at least approximately linear.

2. The mean value of the error term ε is zero.

3. The variance of the error term should be constant.

4. The errors between different individuals are uncorrelated.

5. The error term should be normally distributed.

Therefore, check whether the relationship between the response variable and some explanatory variables is linear (first hypothesis). In the scatter plot matrix, outliers will be detected and whether the variables need to be transformed. The other party builds a more responsible model, including constructing nonlinear features, feature conversion, and adding variable interaction terms. Specifically include the following:

(1) Add a non-linear relationship

Age may not be constant for medical expenses: for older people, fees may be too expensive. Therefore, it is allowed to measure the influence of age by an age square. Variable name is age2.

(2) Conversion - convert a numeric variable into a binary indicator

Assume that the impact of a feature is not cumulative, but only when the feature reaches a given threshold. For example, for people in the normal weight range, the impact of BMI on medical expenses may be zero, but for obese people, it may be associated with higher expenses. Create a binary obesity indicator to establish this relationship, that is, if the BMI is greater than or equal to 30, then set to 1, otherwise set to 0.Variable name is bmi30.

(3) Model setting - adding the influence of interaction

Only consider the individual impact of each feature on the results. If certain features have a comprehensive impact on the corresponding variables, for example, smoking and obesity may have separate effects, the two joint marketing may get worse results than one alone. Therefore, construct bmi30\*smoker.

The original features, plus structural features and interactive features, the completed model is as follows:



For the full model, the residuals plots, Q-Q plots and the histogram of the residuals will be studied to see whether the assumptions of the linear regression model are violated, and whether the variables need further transformation. Checking the summary table of the full model, 10 explanatory variables and 17 interaction between the explanatory variables are involved, but some of them may have no significant effect on the response variables CHARGES.

Consequently, in order to construct a good linear regression model with appropriate explanatory variables and interaction terms, Akaike’s Information Criterion (AIC) and Bayesian’s Information Criterion (BIC) model selection methods are used. As the data description emphasized that the five explanatory variables were important explanatory variables, all of the linear models should have them.

Therefore, in the model selection step, we would set the two base models, one was the full model and the other one was the following one with the least number of explanatory variables:



Then use stepwise regression to screen the model, using forward, backward and both methods, respectively, using AIC and BIC as standards to measure the effect of the model, so a total of 6 models are obtained, and the models MAIC and MBIC with the smallest corresponding indicators are selected:





The t test and ANOVA test were used to select the final model in the above two models. Finally, the model MAIC was chosen, and its residuals plot, Q-Q plot and the histogram for the distribution of the residuals were constructed so as to check whether this MAIC model satisfied the five assumption of the multiple linear regression model.

# Results

## Scatterplots Matrix Study

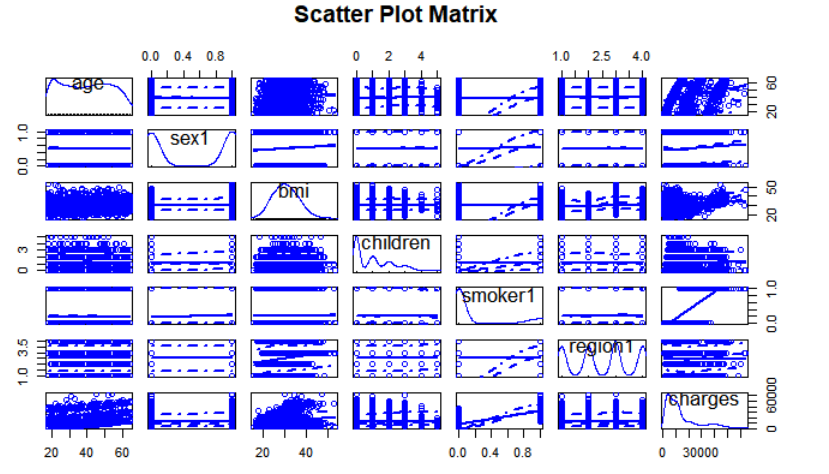


Figure 1: Scatterplots Matrix of the Response Variables and Explanatory Variables

Figure 1 is a scatter plot matrix of response variables. From the scatter plot, we observe that the explanatory variable AGE, BMI may be linearly related to the response variable CHARGES, but there seems to be no obvious relationship between the other variables (SEX, CHILDREN, SMOKE, and REGION) and the response variable.

From the CHARGES distribution chart, the data is not normally distributed, so other transformations such as logarithmic transformation, square transformation, etc. have been tried, but they have not been converted to normal distribution, so the original value is used for modeling, which may cause error terms Non-normal distribution.

## Analysis on the full model

After constructing the full model with 10 explanatory variables (5 numerical variables and 5 categorical variables) and 17 interactions, we got the following summary table:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Table 2: Summary of full model M1 | | | | | |
| Estimated σ | 4378 |  | Significant Predictors | Estimated Value | P-value |
| R2 | 0.8722 | CHILDREN | 938.4134 | 0.000092\*\*\* |
| F-statistic | 307.7 | AGE2 | 4.617 | 1.08E-07\*\*\* |
| F-stat P-value | < 2.2e-16\*\*\* | BMI:SMOKER | 518.313 | 2.15E-10\*\*\* |
|  |  | BMI:REGION(SOUTHEAST) | -141.3959 | 0.0109\* |
|  |  | SMOKER:BMI30 | 14038.5133 | < 2e-16\*\*\* |
|  |  | SMOKER:AGE2 | -3.7901 | 0.0364\* |

According to the R2 value, the full model could only explain 87.22% of the variability of the data. However, in all the 27 explanatory variables and interaction, the six terms, CHILDREN,AGE2,BMI:SMOKER,BMI:REGION(SOUTHEAST),SMOKER:BMI30,and SMOKER:AGE2 are significant with p-value<5%. Therefore, the model did follow the first assumption of the linear regression model.

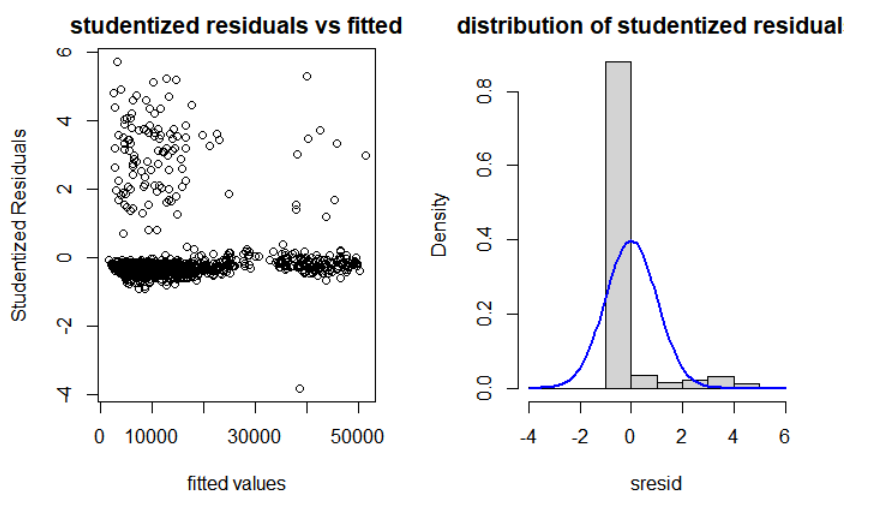


Figure 2: Studentized Residuals Plot (Studentized Residuals VS Fitted Values) and Histogram of residuals distribution

In addition, use studentized residual plots and residual distribution histograms to test the remaining four hypotheses. For the studentized residual plot in Figure 2 (the studentized residual VS the fitted value), the residuals are independent of each other. Most of them are randomly located on both sides of the y = 0 line, and the mean value of the difference = 0, but some of them are distributed away from each other. The far position of y=0 indicates that the variance is not necessarily constant. It follows the second and fourth assumptions of the multilinear model, but the third is not satisfied. In addition, in the histogram of residuals, the data does not appear to be normally distributed and does not follow the fifth hypothesis. Therefore, the complete model satisfies 3 assumptions of the linear regression model, and 2 of them are not satisfied. Since the predicted target charges also do not meet the normal distribution, the error term does not respect the normal distribution. Through the logarithmic sum of the target various transformations such as squaring still did not solve the problem, but from the perspective of R^2, the model explained 87.22% of the information, and the regression model made some strong assumptions about the data. These strong assumptions are not so important for numerical prediction, because the value of the model lies in whether it really captures the basic process-since the model explains 87.22% of the information, we believe that the model has a high accuracy rate, so no further transformation is required.

To check the outliners, the scale-location plot and the residuals VS leverage plots in figure 3 were applied. In the scale-location plot, three outliners are observed. However, the Cook’s distance for all the data points seemed small, which meant the outliners might not have significant effect on the model; therefore, we decided not to remove them.

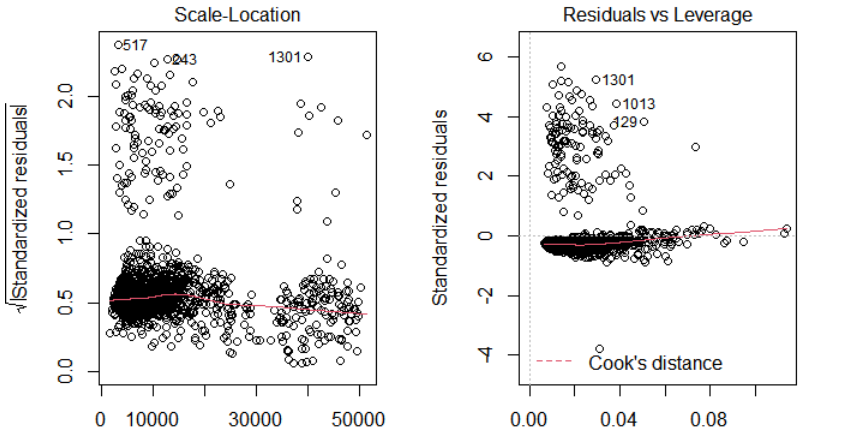


Figure 3: Scale-location plot and residuals VS leverage plot for full model M1

## Model Selection

The AIC and BIC model selection steps were used to find the final model with appropriate variables. Finally, for AIC, the final selected model was MAIC:



The summary table for the MAIC is:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Table 3: Summary of model MAIC | | | | | |
| Estimated σ | 4374 |  | Significant Predictors | Estimated Value | P-value |
| R2 | 0.8713 | SEX | -504.7322 | 0.0365\*\* |
| F-statistic | 496.1 | CHILDREN | 773.129 | 3.86E-11\*\*\* |
| F-stat P-value | < 2.2e-16\*\*\* | AGE2 | 4.4349 | 1.31E-07\*\*\* |
|  |  | BMI:SMOKER | 504.6852 | 4.88E-10\*\*\* |
|  |  | BMI:REGION(SOUTHEAST) | -129.0373 | 0.0183\* |
|  |  | SMOKER:BMI30 | 14853.6719 | < 2e-16\*\*\* |
|  |  | SMOKER:AGE2 | -3.4516 | 0.0496\* |

On the other hand, for BIC, the final selected model was MBIC:



The summary table for the MBIC is:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Table 4: Summary of model MBIC | | | | | |
| Estimated σ | 4409 |  | Significant Predictors | Estimated Value | P-value |
| R2 | 0.8682 | CHILDREN | 656.4832 | 7.57E-11\*\*\* |
| F-statistic | 1251 | AGE2 | 3.349 | < 2e-16\*\*\* |
| F-stat P-value | < 2.2e-16\*\*\* | BMI:SMOKER | 489.4235 | 1.56E-09\*\*\* |
|  |  | SMOKER:BMI30 | 14770.1305 | < 2e-16\*\*\* |

The difference between MAIC and MBIC was that the MAIC had more terms than the MBIC ; Therefore, the ANOVA was applied to test whether the MAIC and MBIC were significant different . The following table is the ANOVA table:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Table 5: ANOVA table for M**BIC VS MAIC | | | | | | |
|  | Res.Df | RSS | Df | Sum of Sq | F | Pr(>F) |
| MBIC | 1330 | 25850000000 |  |  |  |  |
| MAIC | 1319 | 25232000000 | 11 | 617882577 | 2.9363 | 0.00076\*\*\* |

The null hypothesis of the F test in the ANOVA table was that the model MAIC with three more interaction terms was not significantly better than the model MBIC, but the p-value was less than 1%; therefore, we had significant evidence to reject the null hypothesis, which meant the model MAIC was significantly better than the MBIC.

Two models are used to predict the new sample (Age=51, BMI=34.69, CHILDREN=2, SEX=MALE, SMOKER=YES, REGION=NORTHWEST) and the results are as follows:

|  |  |  |  |
| --- | --- | --- | --- |
| **Table** 6**: Forecast and 95% forecast interval table for MBIC VS MAIC** | | | |
|  | fit | lwr | upr |
| MBIC | 44656.14 | 35973.15 | 53339.12 |
| MAIC | 44831.02 | 36174.13 | 53487.9 |

Since the values of the new samples are relatively large, the result predicted by MBIC is 44656.14, and the result predicted by MAIC is 44831.02, both of which are greater than the 3rd Qu. of CHARGES, indicating that the selected new observation point corresponds to a higher cost, and MAIC is higher than MBIC 174.88.

## Final model analysis

Therefore, the final model chosen is MAIC . The R2 value indicates that MAIC can explain 87.16% of the variability, and the p value of some items in MAIC is much less than 1%. Although the model does not meet the five assumptions of linear regression, these strong assumptions are not so important for numerical prediction, because the value of the model lies in whether it really grasps the basic process-we care about the accuracy of the model prediction. So it is still concluded that MAIC is a good final model.

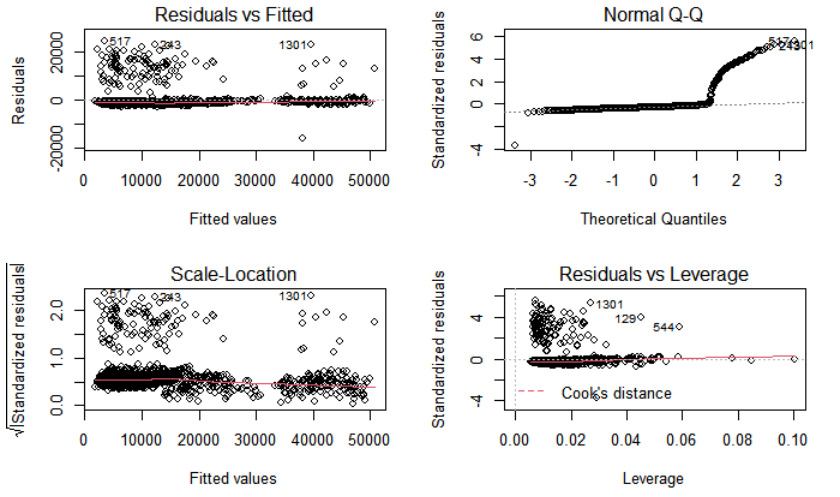


Figure 4.Linear regression model assumption diagnosis plot for MAIC

# Conclusion and discussion

According to the final model MAIC , we can conclude that the charges is linear related to:

1. Age of beneficiary (AGE)

2. The gender of the policy holder (SEX)

3. Body mass index BMI (BMI)

4. Number of children/dependents included in the insurance plan (CHILDREN)

5. Does the insured smoke regularly (SMOKER)

6. Whether the beneficiary lives in the Northwest (REGION(NORTHWEST))

7. Whether the beneficiary lives in the Southeast (REGION(SOUTHEAST))

8. Whether the beneficiary lives in the Southwest (REGION(SOUTHWEST))

9. Does the body mass index BMI exceed 30 (BMI30)

10. The square of the beneficiary's age (AGE2)

11. The interaction between the age of the beneficiary and the number of insured people who smoke regularly (AGE:SMOKER)

12. The interaction between BMI and the number of insured people who smoke regularly (BMI: SMOKER)

13. The interaction between the body mass index BMI and whether the beneficiary lives in the northwest (BMI: REGION (NORTHWEST))

14. The interaction between the body mass index BMI and whether the beneficiary lives in the Southeast (BMI: REGION (SOUTHEAST))

15. The interaction between the body mass index BMI and whether the beneficiary lives in the southwest (BMI: REGION (SOUTHWEST))

16. The interaction between the number of children/dependents included in the insurance plan and whether the number of insured people smoke regularly (CHILDREN: SMOKER)

17. The interaction between whether the insured person smokes frequently and whether the body mass index BMI exceeds 30 (SMOKER: BMI30)

18. Does the number of insured smoke regularly and the square of the beneficiary's age (SMOKER:AGE2)

According to the results of the t-test in the summary table, the seven important variables detected in the above eighteen interpretations are SEX, CHILDREN, AGE2, the interaction term BMI: SMOKER, BMI: REGION (SOUTHEAST), SMOKER: BMI30, SMOKER: AGE2 . For seven important variables, the estimated coefficients are

|  |  |
| --- | --- |
| Table 7: Estimated coefficients of significant explanatory variables in MAIC | |
| SEX | -504.7322 |
| CHILDREN | 773.1290 |
| AGE2 | 4.4349 |
| BMI:SMOKER | 504.6852 |
| BMI:REGION(SOUTHEAST) | -129.0373 |
| SMOKER:BMI30 | 14853.6719 |
| SMOKER:AGE2 | -3.4516 |

Based on the table 7, we made the following conclusion statement:

1. If the policy holder is male, the point estimation of charges will be 504.7322 lower than female, with other variables remaining constant.

2. The more children/dependents included in the insurance plan, the higher the cost, and the premium for one more child increases by 773.1290, while other variables remain unchanged.

3. The higher the square of the beneficiary’s age, the higher the premium. The square of the beneficiary’s age increases by 1 year and the cost increases by 4.4349, while other variables remain the same.

4. The BMI of smokers gains 1 unit, which is 504.6852 higher than the cost of non-smokers, while other variables remain unchanged.

5. The BMI of people living in SOUTHEAST increased by 1 unit, which is 129.0373 less than the cost of people living in other places, while other variables remain unchanged

6. People who smoke and have a BMI of more than 30 cost 14853.6719 more than others, while other variables remain the same.

7. The square of the age of the smoker increases by 1 unit, which is a decrease of 3.4516 compared to the cost of other people, while other variables remain unchanged.

Therefore, we know that the cost is related to the gender of the holder, the number of children/dependents, age, smoking, and BMI. Men need less expenses than women. The more children there are, the more expense is needed. Age is not linearly related to the cost. The older the age, the more the cost is. The cost of smokers is more than that of non-smokers, and the cost of smoking and overweight people is more than the cost of others.

Therefore we recommend:

1. Appropriate weight control can reduce corresponding expenses;

2. Reducing smoking can also reduce costs;

3. For people who smoke and are overweight, more control is needed.

# Reference

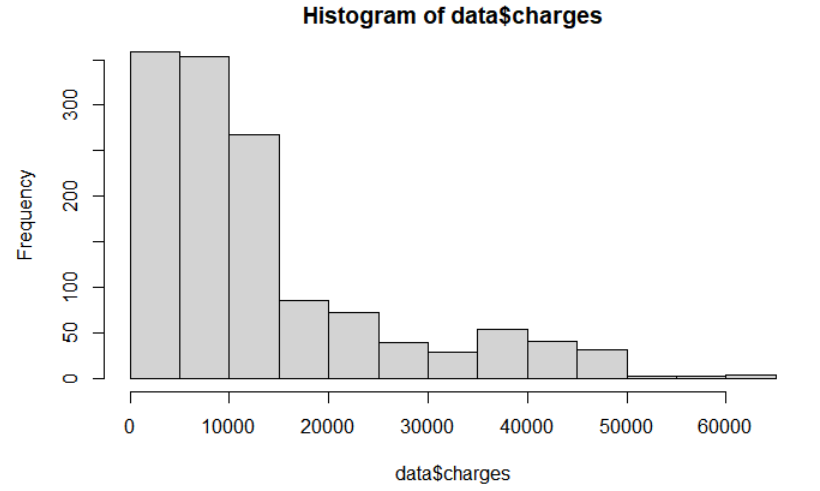
1.Machine learning and R language (2nd edition of the original book)[United States] Brett Lantz (Brett Lantz)

2.Introduction to Linear Regression Analysis (5th edition of the original book)

[America] Douglas C. Montgomery, [America] Elizabeth A. Parker, [America] G. Jeffrey Vaining

# Appendix

**Distribution Histogram of charges:**



Summary of Full Model:

> summary(full\_model)

Call:

lm(formula = charges ~ age + factor(sex) + bmi + children + factor(smoker) +

factor(region) + age \* factor(sex) + age \* factor(smoker) +

age \* factor(region) + bmi \* factor(sex) + bmi \* factor(smoker) +

bmi \* factor(region) + children \* factor(sex) + children \*

factor(smoker) + children \* factor(region) + bmi30 + age2 +

bmi30 \* factor(smoker) + bmi30 \* factor(smoker) \* age2, data = data)

Residuals:

Min 1Q Median 3Q Max

-16306.0 -1541.2 -1127.5 -788.3 24633.3

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 2099.7179 2034.4476 1.032 0.3022

age -119.8022 69.5935 -1.721 0.0854 .

factor(sex)male -655.7014 1370.7482 -0.478 0.6325

bmi 98.7780 55.7135 1.773 0.0765 .

children 938.4134 239.2032 3.923 9.20e-05 \*\*\*

factor(smoker)yes -3583.6372 3272.1619 -1.095 0.2736

factor(region)northwest 207.6148 2027.6963 0.102 0.9185

factor(region)southeast 2020.2367 1935.7204 1.044 0.2968

factor(region)southwest 678.4660 1949.5983 0.348 0.7279

bmi30 178.9560 622.9592 0.287 0.7740

age2 4.6170 0.8642 5.343 1.08e-07 \*\*\*

age:factor(sex)male 6.9244 17.3629 0.399 0.6901

age:factor(smoker)yes 276.5174 141.8573 1.949 0.0515 .

age:factor(region)northwest 9.7962 24.7905 0.395 0.6928

age:factor(region)southeast 44.6033 24.4673 1.823 0.0685 .

age:factor(region)southwest 32.3350 25.1526 1.286 0.1988

factor(sex)male:bmi -1.7713 40.2489 -0.044 0.9649

bmi:factor(smoker)yes 518.3130 80.9783 6.401 2.15e-10 \*\*\*

bmi:factor(region)northwest -38.6155 63.6535 -0.607 0.5442

bmi:factor(region)southeast -141.3959 55.4383 -2.551 0.0109 \*

bmi:factor(region)southwest -94.6688 60.7423 -1.559 0.1193

factor(sex)male:children -64.8700 201.3796 -0.322 0.7474

children:factor(smoker)yes -467.7699 268.2450 -1.744 0.0814 .

children:factor(region)northwest 135.5779 293.2887 0.462 0.6440

children:factor(region)southeast -271.0908 284.2743 -0.954 0.3405

children:factor(region)southwest -356.9976 279.6649 -1.277 0.2020

factor(smoker)yes:bmi30 14038.5133 1370.9914 10.240 < 2e-16 \*\*\*

bmi30:age2 -0.1715 0.2474 -0.693 0.4885

factor(smoker)yes:age2 -3.7901 1.8099 -2.094 0.0364 \*

factor(smoker)yes:bmi30:age2 0.3896 0.5432 0.717 0.4734

---

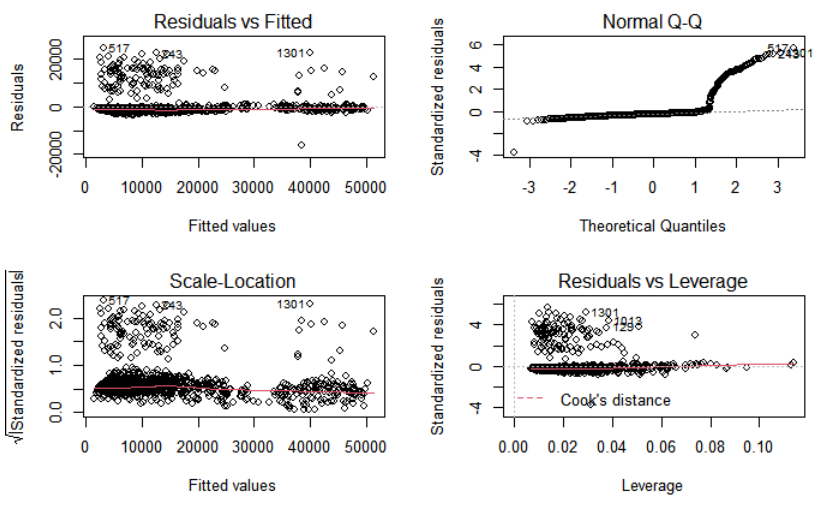
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4378 on 1308 degrees of freedom

Multiple R-squared: 0.8722, Adjusted R-squared: 0.8693

1. statistic: 307.7 on 29 and 1308 DF, p-value: < 2.2e-16

**Diagnosis Plots for full model:**

**Model Selection Step:**

|  |  |  |  |
| --- | --- | --- | --- |
| Table 2: Final selected model from three selecting direction (forward, backward and both) and two information criteria (AIC and BIC)Information | | | |
|
| information criteria | Selection  Direction | Final Selected Model | AIC/BIC value  for final model |
| AIC | Forward | CHARGES ~ AGE + SEX + BMI + CHILDREN + SMOKER + REGION(NORTHWEST) + REGION(SOUTHEAST) + REGION(SOUTHWEST) + BMI30 + AGE2 + AGE:SEX + AGE:SMOKER + AGE:REGION(NORTHWEST) + AGE:REGION(SOUTHEAST) + AGE:REGION(SOUTHWEST) + SEX:BMI + BMI:SMOKER + BMI:REGION(NORTHWEST) + BMI:REGION(SOUTHEAST) + BMI:REGION(SOUTHWEST) + SEX:CHILDREN + CHILDREN:SMOKER + CHILDREN:REGION(NORTHWEST) + CHILDREN:REGION(SOUTHEAST) + CHILDREN:REGION(SOUTHWEST) + SMOKER:BMI30 + BMI30:AGE2 + SMOKER:AGE2 + SMOKER:BMI30:AGE2 | 26264.98 |
| AIC | Backward | CHARGES ~ AGE + SEX + BMI + CHILDREN + SMOKER + REGION(NORTHWEST) + REGION(SOUTHEAST) + REGION(SOUTHWEST) + BMI30 + AGE2 + AGE:SMOKER + BMI:SMOKER + BMI:REGION(NORTHWEST) + BMI:REGION(SOUTHEAST) + BMI:REGION(SOUTHWEST) + CHILDREN:SMOKER + SMOKER:BMI30 + SMOKER:AGE2 | 26251.86 |
| AIC | Both | CHARGES ~ AGE + SEX + BMI + CHILDREN + SMOKER + REGION(NORTHWEST) + REGION(SOUTHEAST) + REGION(SOUTHWEST) + BMI30 + AGE2 + AGE:SMOKER + BMI:SMOKER + BMI:REGION(NORTHWEST) + BMI:REGION(SOUTHEAST) + BMI:REGION(SOUTHWEST) + CHILDREN:SMOKER + SMOKER:BMI30 + SMOKER:AGE2 | 26251.86 |
| BIC | Forward | CHARGES ~ AGE + SEX + BMI + CHILDREN + SMOKER + REGION(NORTHWEST) + REGION(SOUTHEAST) + REGION(SOUTHWEST) + BMI30 + AGE2 + AGE:SEX + AGE:SMOKER + AGE:REGION(NORTHWEST) + AGE:REGION(SOUTHEAST) + AGE:REGION(SOUTHWEST) + SEX:BMI + BMI:SMOKER + BMI:REGION(NORTHWEST) + BMI:REGION(SOUTHEAST) + BMI:REGION(SOUTHWEST) + SEX:CHILDREN + CHILDREN:SMOKER + CHILDREN:REGION(NORTHWEST) + CHILDREN:REGION(SOUTHEAST) + CHILDREN:REGION(SOUTHWEST) + SMOKER:BMI30 + BMI30:AGE2 + SMOKER:AGE2 + SMOKER:BMI30:AGE2 | 26426.14 |
| BIC | Backward | CHARGES ~ BMI + CHILDREN + SMOKER + BMI30 + AGE2 + BMI:SMOKER + SMOKER:BMI30 | 26309.02 |
| BIC | Both | CHARGES ~ BMI + CHILDREN + SMOKER + BMI30 + AGE2 + BMI:SMOKER + SMOKER:BMI30 | 26309.02 |

**Summary of** MAIC **:**

> summary(full\_model\_aic)

Call:

lm(formula = charges ~ age + factor(sex) + bmi + children + factor(smoker) +

factor(region) + bmi30 + age2 + age:factor(smoker) + bmi:factor(smoker) +

bmi:factor(region) + children:factor(smoker) + factor(smoker):bmi30 +

factor(smoker):age2, data = data)

Residuals:

Min 1Q Median 3Q Max

-16087.2 -1554.6 -1183.7 -793.3 24257.3

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1563.7263 1841.3542 0.849 0.3959

age -87.6725 66.9525 -1.309 0.1906

factor(sex)male -504.7322 241.1028 -2.093 0.0365 \*

bmi 90.1593 52.4342 1.719 0.0858 .

children 773.1290 115.9852 6.666 3.86e-11 \*\*\*

factor(smoker)yes -3560.7550 3245.9191 -1.097 0.2728

factor(region)northwest 429.0807 1869.2054 0.230 0.8185

factor(region)southeast 3115.6063 1725.7928 1.805 0.0713 .

factor(region)southwest 1292.1316 1813.7951 0.712 0.4763

bmi30 -128.5024 451.7184 -0.284 0.7761

age2 4.4349 0.8357 5.307 1.31e-07 \*\*\*

age:factor(smoker)yes 271.0136 141.1204 1.920 0.0550 .

bmi:factor(smoker)yes 504.6852 80.4905 6.270 4.88e-10 \*\*\*

bmi:factor(region)northwest -26.9653 62.9421 -0.428 0.6684

bmi:factor(region)southeast -129.0373 54.6011 -2.363 0.0183 \*

bmi:factor(region)southwest -85.4853 59.4985 -1.437 0.1510

children:factor(smoker)yes -487.7672 265.5281 -1.837 0.0664 .

factor(smoker)yes:bmi30 14853.6719 1002.9986 14.809 < 2e-16 \*\*\*

factor(smoker)yes:age2 -3.4516 1.7563 -1.965 0.0496 \*

---

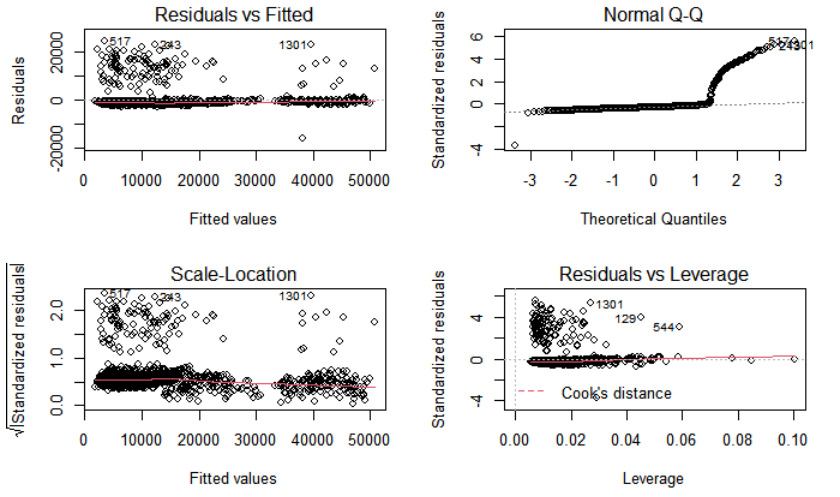
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4374 on 1319 degrees of freedom

Multiple R-squared: 0.8713, Adjusted R-squared: 0.8696

F-statistic: 496.1 on 18 and 1319 DF, p-value: < 2.2e-16

**Diagnosis Plots for** MAIC **:**



**Summary of** MBIC **:**

> summary(full\_model\_bic)

Call:

lm(formula = charges ~ bmi + children + factor(smoker) + bmi30 +

age2 + bmi:factor(smoker) + factor(smoker):bmi30, data = data)

Residuals:

Min 1Q Median 3Q Max

-16110.2 -1477.0 -1278.6 -838.7 23844.3

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1884.9675 979.6980 1.924 0.0546 .

bmi -2.0314 37.1540 -0.055 0.9564

children 656.4832 100.0385 6.562 7.57e-11 \*\*\*

factor(smoker)yes 999.1526 2087.5968 0.479 0.6323

bmi30 70.6508 448.6695 0.157 0.8749

age2 3.3490 0.1079 31.047 < 2e-16 \*\*\*

bmi:factor(smoker)yes 489.4235 80.4884 6.081 1.56e-09 \*\*\*

factor(smoker)yes:bmi30 14770.1305 1007.9438 14.654 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4409 on 1330 degrees of freedom

Multiple R-squared: 0.8682, Adjusted R-squared: 0.8675

F-statistic: 1251 on 7 and 1330 DF, p-value: < 2.2e-16

**Diagnosis Plots for** MBIC **:**

