Predicting Medical Expenses with Linear Regression

Abstract

For insurers, it’s important to develop models that accurately forecast medical expenses so that they can make money. Through this project, we aim at pretending to work for insurance companies and build different models to best predict the medical cost of individuals given their basic information. Also, through different methods, the article tries to detect inner relationship among different dependent variables to help better understand the impact of different variables on the medical charges. The final prediction result can be used as a benchmark for the insurance company to establish appropriate insurance claim coverage for their contractors.

We first applied linear regression models to find significant variables to help predict the medical charges. By adding interaction variables, changing the continuous variable to be categorical variable, we fit several different linear regression models. We detected the abnormality of residuals. We raised some hypothesis and did several tests to try to explain this issue.

We also used random forest method to better forecast the charges of individuals and find the importance of different variables.

After that, we applied K-Means to cluster the individuals and try to find some statistical differences among different groups, further confirming what key variables have a very large impact on charges.

Introduction

Usually, medical cost is hard to estimate because the cost is almost random. However, some of the situations are more common for some specific groups. For example, smokers are more likely to get lung cancer than non-smokers, fat people are more likely to get heart disease. The purpose of this project is to predict the health insurance contractors' medical cost with features indicating characteristics of the patient as well as the total medical expenses charged to the plan for the calendar year.

The first question before we start doing analysis is whether the data, we gained fulfills the reality. In the real world, the proportion of both genders are roughly equal to 50%, the number of people in different ages are uniformly distributed. We want to see if the data is scientifically and statistically reasonable and worth exploring.

After confirming the distribution of the data, we attend to use the data to solve several questions: whether certain variables are relevant to people's medical cost, whether the variables themselves are correlated, whether there is a tendency among certain group of people who shares the same features.

Materials

This dataset is obtained from kaggle's dataset and contains 1338 examples. The URL for this dataset is https://www.kaggle.com/mirichoi0218/insurance

Exploratory Data Analysis

Dependent variable

Charges: numeric, indicates individual medical cost billed by health insurance.

From the histogram in Figure 1.1, the distribution of charges is highly right skewed. After trying sqrt transformation and log transformation, we decided to use log transformation since it can remove right skewed distribution to a greater extent. Figure 1.2 reflects the distribution after removing skewness.

Independent variables

1. Age: integer, indicates the age of the beneficiary. (Smaller than 64 and greater than 18, because the people whose age beyond this age interval are covered by U.S. government)

Figure 1.3 reveals the uniform distribution of age variable.

2. Sex: factor, indicates the gender of the beneficiary. The proportion of female and male are nearly the same, which are both 50%.

3. Bmi: numeric, indicates the body mass index. The histogram in Figure 1.4 depicts bmi as a slightly right skewed normal distributed attribute with some of outliers. Therefore, most of the people have bmi around 30.

4. Children: integer, indicates the number of children in this family covered by health insurance.

5. Smoker: factor, indicates whether or not the beneficiary smoke, have values 274 “yes” and 1064 “no”.

6. Region: factor, indicates the beneficiary’s residential area in the U.S, have 324 “northeast”, 325 “northwest”, 364 “southeast” and 325 “southwest”, which seems a balance attribute.

Scatter Plot Matrix

The correlation ellipse in the scatter plot matrix (figure 1.5) provides a visualization of the correlation of variables. The more stretched the ellipse is, the stronger correlation exists between the variables. Therefore, from the matrix above, we can see that there is a relationship between variable smoker and charges. The boxplot (figure 1.7) and density plot (figure 1.8) of smoker and log(charges)s can reveal the point that smokers tend to have higher log(charges)s in average.

Methods

Linear Regression Model

According to the steps of multiple linear regression, we first look at the correlation matrix in Table 1.0 to explore the relationship between variables.

Secondly, we fit the simple linear regression using every variable. From the R^2 of every simple linear regression model, we can see that bmi and smoker have relatively higher R^2.

Next, we consider multiple linear regression model. We first put all of the variables into the multiple regression model, and then, improve the model by introducing new variables or interaction term.

(1). Create new variable bmi30:

A people should be classified as obesity of his/her body mass index is larger than 30 according to the definition of obesity, and obesity would lead to the increase of medical cost, so it is reasonable to create a new factor variable which indicates whether or not this contractor's bmi is larger than 30.

(2). Create interaction term by multiplying bmi30 with smoker:

Since we found variable bmi and smoker both have high R^2 in the simple linear regression model, so they have the same contribution to the charges. Therefore, we tried to include the interaction term in the model, which means add bmi30*smoker in the original model.

Check multicollinearity

Then check the multicollinearity. A VIF larger than 10 indicates multicollinearity.

Hypothesis testing

(1). Using F test to see if parameters equal to 0 in the final model.

$$H_0: \beta_k = 0, k = 1,2,3,4,5,6,7 \quad \text{vs.} \quad H_1: \beta_k \neq 0$$

Decision rule: Test statistic: $F^* = \frac{MSR}{MSE} \sim F_{df,n-df-1}$ Reject H_0if $F^* > F_{df,n-df-1}(1-\alpha)$

(2). Using Breusch-Pagan test to see if the residuals have constant variance.

$$H_0: Var(\epsilon_i) = \sigma^2 \text{ which means constant variance} \quad \text{vs.} \quad H_1: Var(\epsilon_i) \neq \sigma^2$$

Decision rule: if associated p-value is smaller than α=0.05, reject the null. Otherwise, fail to reject the null.

(3). Using Lilliefors test to see if the residuals have normal distribution.

$$H_0\text{: sample comes from } N(\mu, \sigma^2) \text{ distribution. vs. } H_1: o.w.$$

Decision rule: if associated p-value is smaller than α=0.05, reject the null. Otherwise, fail to reject the null.

Residual Analysis

With the abnormality of the QQ plot (figure 2.11), we selected the ‘right tail’ data and plotted them on the log-transformation normal plot to further explore the reason on non-normality.

By coloring this part of data using different variables, we want to figure out if any single variable influences the results.

Also, based on residual values, we classified our data into three class:

Class 1: residual<0

Class 2: 0<residual<20000

Class 3: residual>20000

We use decision tree to find out the relation between classes and other variables. Also, we color the three classes to see it contributes to the abnormality of the distribution.

Random Forest

Further, we applied Random Forest to better forecast the charges of individuals and find the importance of different variables. Random forest operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Also, it can provide the variable importance by calculating the increase of node purity.

K-Means Clustering

At last, we used the unsupervised learning method – K-Means clustering. K-Means is a simple unsupervised machine learning algorithm that groups a dataset into a user-specified number K of clusters. We used Elbow Method to select K, which is to run K-Means clustering on the dataset for a range of values of K (say, K from 1 to 10 in our project), and for each value of K calculate the sum of squared errors (SSE). Then, plot a line chart of the SSE for each value of K. If the line chart looks like an arm, then the "Elbow" on the arm is the value of K that is the best. We used K-Means to cluster the individuals and try to find some statistical differences among different groups, further confirming what key variables have a very large impact on charges.

Results

Linear Regression Model

1. From the correlation matrix in Table 1.0, we can see that although the correlation coefficient in the matrix does not reflect strong relationships between these variables, some correlations do exist. For example, age and bmi has moderate relationship, which means with the increase of age, body mass index will also increase. Besides of that, age and charges, bmi and charges all have moderate correlation.
2. When put the variables in simple linear regression model individually, we can see that bmi and smoker have relatively higher R^2, but all of the R^2 are below 0.5, which means simple linear regression model cannot address this problem.
3. When considering multiple linear regression model, we first put all of the variables in the model, and get R^2 0.7666. (Figure 2.1)
4. We introduced a new variable named bmi30, but this new variable does not improve the model so much.
5. The result shows that the introduction of the interaction term would successfully improve the adjusted R^2 from 0.7666 to 0.7869. (Figure 2.2), so we should add this interaction term in the model.
6. Then check the multicollinearity. A VIF larger than 10 indicates multicollinearity. We have no VIF larger than 10 (figure 2.4), so this indicates no multicollinearity.
7. We use 0.95 as confidence level for all of the test.

(1). Parameters test

$$H_0: \beta_k = 0, k = 1,2,3,4,5,6,7 \quad \text{vs.} \quad H_1: \beta_k \neq 0$$

Decision rule: Test statistic: $F^* = \frac{MSR}{MSE} \sim F_{df,n-df-1}$ Reject H_0 if $F^* > F_{df,n-df-1}(1-\alpha)$

Conclusion: Fail to reject $H_0: \beta_2 = 0$ since F^* associated with sex is 1.7380 according to the anova analysis table (figure 2.5), which is smaller than $F_{1,1336}(0.95)$ = 3.848429.

Therefore, we should remove sex from the original model. Thus, our final model would include variables: age, bmi, children, region, bmi30*smoker. (figure 2.6) The final model has adjusted R^2 0.7848.

(2). Constant variance test

From the plot (figure 2.7) we can conclude that the residuals do not have constant variance. After conducting the Breusch-Pagan test, we confirmed this conclusion.

$$H_0: Var(\epsilon_i) = \sigma^2 \ which\ means\ constant\ variance \quad vs. \quad H_1: Var(\epsilon_i) \neq \sigma^2$$

Decision rule: if associated p-value is smaller than α=0.05, reject the null. Otherwise, fail to reject the null.

Conclusion: the p-value of the Breusch-Pagan test is very small (figure 2.8), so we reject the null hypothesis, which means that the residuals do not have constant variance.

(3). Normality test

From the plot (figure 2.9) we can see that the residuals do not have normal distribution. Let's test it:

H_0: sample comes from $N(\mu, \sigma^2)$ distribution. Vs. $H_1: o.w.$

Decision rule: if associated p-value is smaller than α=0.05, reject the null. Otherwise, fail to reject the null.

Conclusion: the p-value of the Kolmogorov-Smirnov test is very small (figure 2.10), so we reject the null hypothesis, which means that the residuals are not normal distributed. The QQ plot (figure 2.11) also indicates the issue.

Residual Analysis

The colored QQ plot of smokers (figure 2.12) shows a difference. The distribution plot of three classes via charges (figure 2.13) shows that there are significant differences on charges among the three classes.

According to the decision tree (figure 2.14), the three classes are:
Class 1: non-smoker
Class 2: smoker but bmi<30
Class 3: smoker and bmi>30

As shown in the pair plot (figure 2.15), the relation between age and log of charges is different from that with class 2 or 3. By introducing the cross term “log of age” into class 1 only, we reached R-square of 0.965. (figure 2.16) The residual versus fitted plot (figure 2.17) of this model shows non-constant variance structure, so we took square root and added children as a factor and solved the problem. This time, the R-square value reaches 0.98 (figure 2.18) and the variance of the residual is constant within each class. (figure 2.19)

Random Forest

When applying Random Forest, we divide the data into training set and test set with 80 percent versus 20 percent of the original data. For the test dataset, we got the final R-square around 0.80 (0.7991286), which is quite ideal. The value is nearly the same as we gained from linear model.

Also, we plot the Age versus Charges (figure 3.1) to visualize the performance of the predicted values versus true values of test dataset. We can see that for most points, the predicted values and the true values are very close, excluding some points having very large charges, which means that Random Forest works pretty well for this dataset.

Additionally, we got the variable importance (figure 3.2) by calculating the increase of node purity in Random Forest. We can see that Smoker is the most important variable to predict the charges, which is very out-performance than other variables (the importance of Smoker is nearly 4 times than that of the second important variable). BMI and Age are also very important and nearly have the same importance. Children and Region have little importance, while Sex nearly has no importance to Charges by the result of the Variable Importance in Random Forest.

K-Means Clustering

When applying K-Means Clustering, we need to first to select the number of clusters, which is K. We applied Elbow Method to select K and found that the “Elbow” point of the plot is K = 3. Thus, we used K = 3 in our further analysis. (figure 3.3)

After applying K-Means with K = 3, we plot the Age versus Charges (figure 3.4) to visualize the clusters that we got in K-Means. We can see that the clustering results is kind of ideal since they have a clear dividing among three clusters with respect to the variable Charges. It divides the whole dataset into three groups with high-charges, medium-charges and low-charges.

We can further explore some descriptive statistical features among three clusters.

For continue variables, we calculate their mean values to find whether there are clear differences among three clusters.

We can see from Table 1.1 that for every group there are clear differences in Age and BMI. The group with mean charges of 40761.309 has the largest mean charges with the highest BMI and highest Children numbers. The group with mean charges of 6430.148 has the lowest mean charges with the lowest Age. The group with mean charges of 18897.644 has the medium mean charges with the highest Age and lowest BMI.

For categorical variables, we can calculate their proportions to find whether there are clear differences among three clusters.

For sex, we can see from Table 2.1 that the group with the highest charges has the obviously highest percentage of male.

For smoker, there exists very clearly differences. We can see from Table 2.2 that the group with the highest charges has the obviously highest percentage of smokers. For the group which has the lowest charges, there is none smoker.

For region, the difference is not that so obvious. (Table 2.3) For the group with the highest charges, it has the obviously highest percentage in southeast but also obviously lowest percentage in northwest.

Discussion and conclusions

For the linear regression model, our final model includes variables of age, bmi, children, region, smoker and bmi30*smoker with $R^2 = 0.7863$ and adjusted-$R^2 = 0.7848$. We applied 3 tests after building the linear regression model and found that although there is no multicollinearity problem, the residuals do not have constant variance or normal distribution.

To further explore our model, we ran residual analysis on our data. We found an abnormality that the left part of the line in QQ plot perfectly fits normal distribution while the right part of the line turns out to be abnormal. We tried to interpret the reason and reached out to the conclusion that with the given data, we could not distinguish the difference among non-smokers. If given more detailed and comprehensive information or variables of these individuals instead of dividing them using “smokers or not” only, or if given more observations in the dataset, we could better profile people in different payment range.

Finally, we applied Random Forest and K-Means Clustering as well. We got $R^2 = 0.7991286$, which performs a little bit better than linear regression model. At the same time, by Variable Importance provided by Random Forest, we can conclude that smokers (smoke or not), BMI and age are the main variables that influence the medical charges of an individual person most. Region, gender and children (the number of children covered in the insurance) show little influence on the individual medical charges. The results of descriptive statistical analysis comparing different clusters generated by K-Means also confirm this conclusion. It shows obvious differences in smoker, BMI and age variables, either on the mean values or on the proportion changes.

Bibliography

[1] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. An Introduction to Statistical Learning: with Applications in R. New York: Springer, 2013.

[2] Hastie, Trevor, Tibshirani, Robert and Friedman, Jerome. The Elements of Statistical Learning. New York, NY, USA: Springer New York Inc., 2001.

[3] Classification and Regression Trees by L. Breiman, J.H. Friedman, C.J. Stone, and R.A. Olshen (Chapman & Hall, 1984).

[4] Kutner, Michael H, Chris Nachtsheim, John Neter, and William Li. Applied Linear Statistical Models., 2005.

[5] Graybill, E A. Theory and Application of the Linear Model. Boston: Duxbury Press, 1976.

[6] Hocking, R. R. Methods and Applications of Linear Models: Regression and the Analysis of Variance. 2nd ed. New York: John Wiley & Sons, 2003.

[7] Littell, R. C.; W. W. Stroup; and R. J. Freund. SAS System for Linear Models. 4th ed. New York: John Wiley & Sons, 2002.

[8] Searle, S. R. Linear Modelsfor Unbalanced Data. New York: John Wiley & Sons, 1987.

Appendix

Figures and tables:

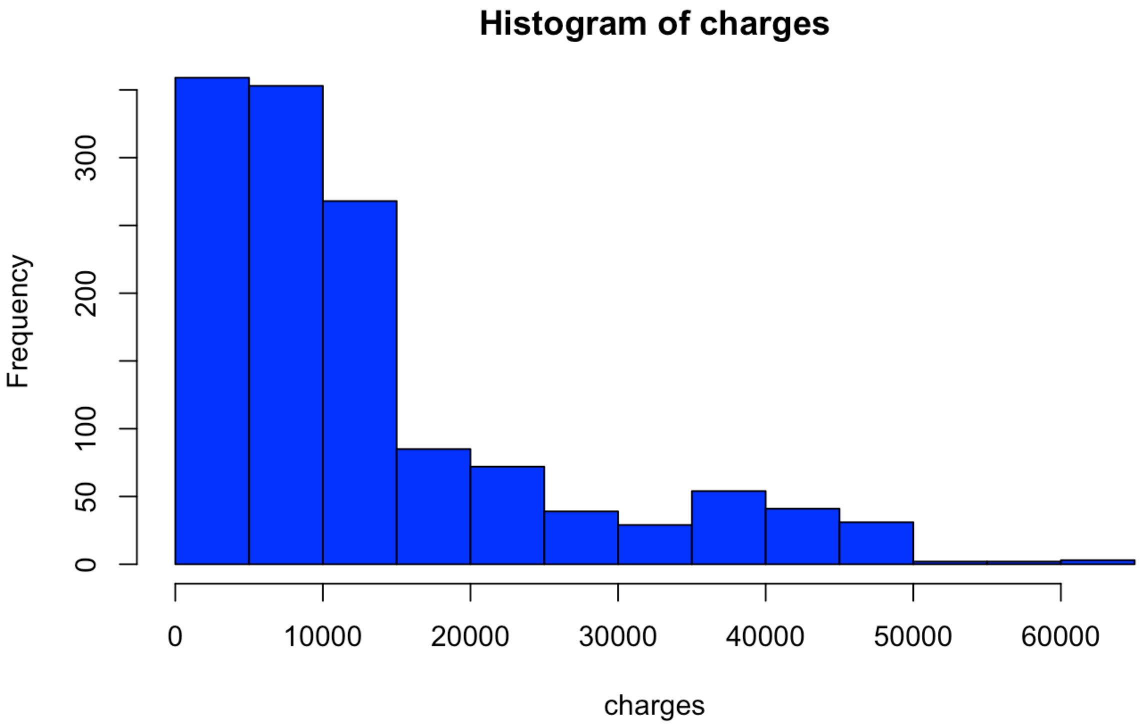


Figure 1.1 Histogram of charges

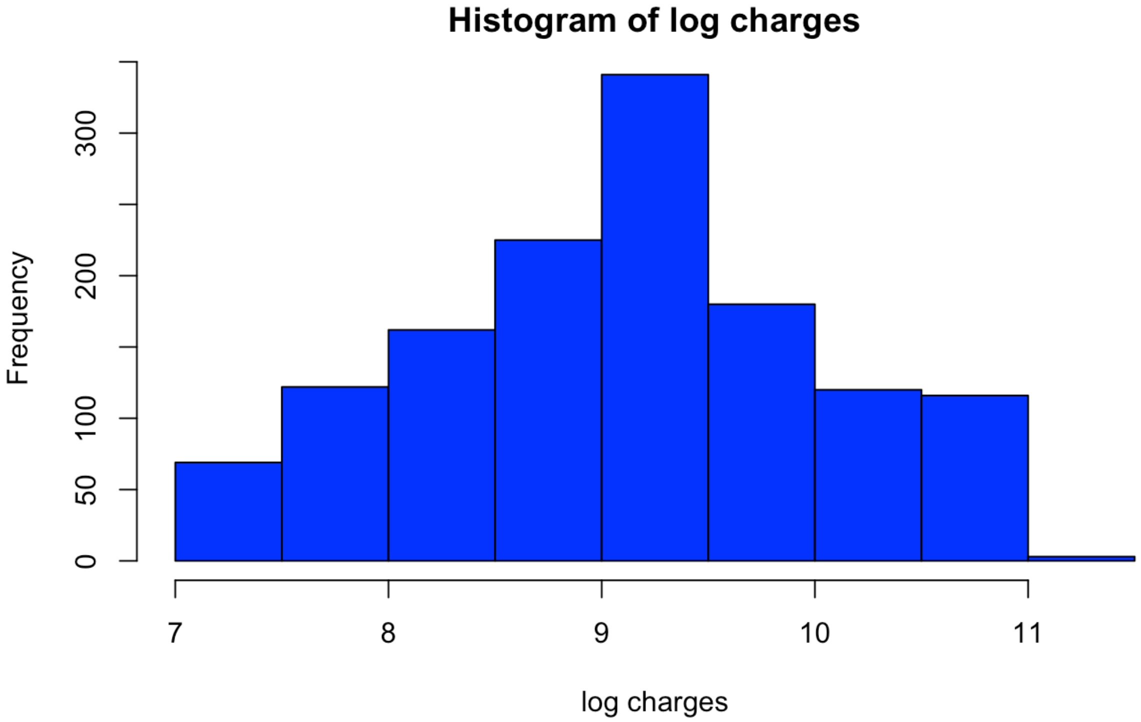


Figure 1.2 Histogram of log charges

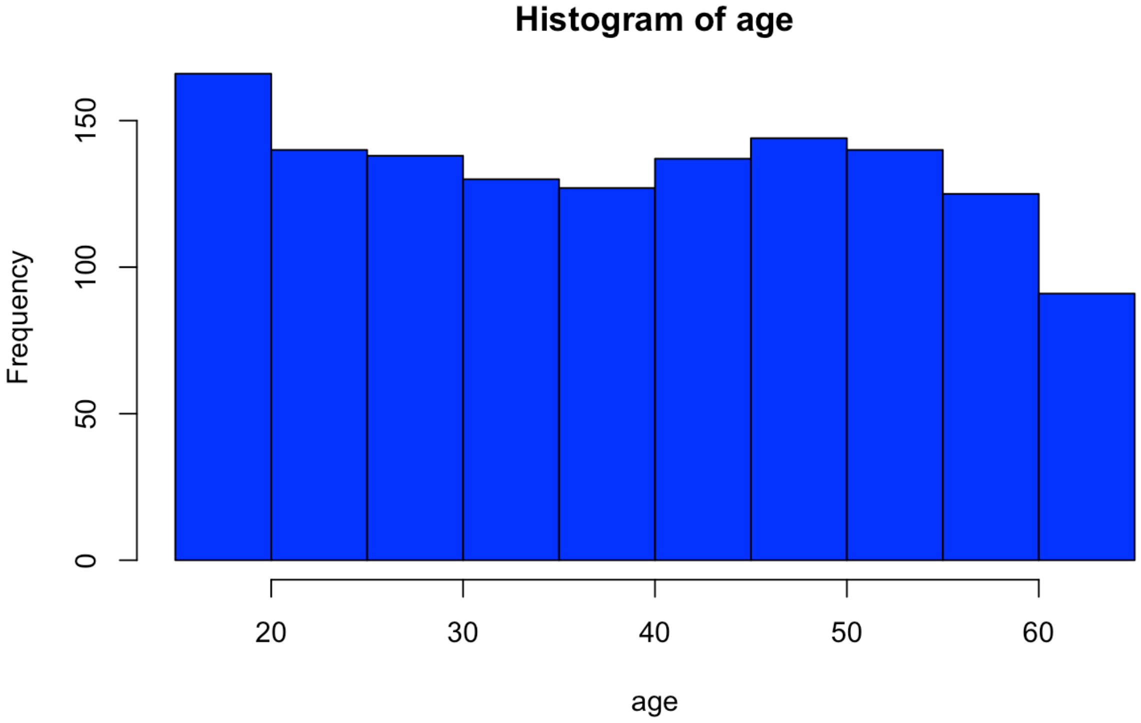
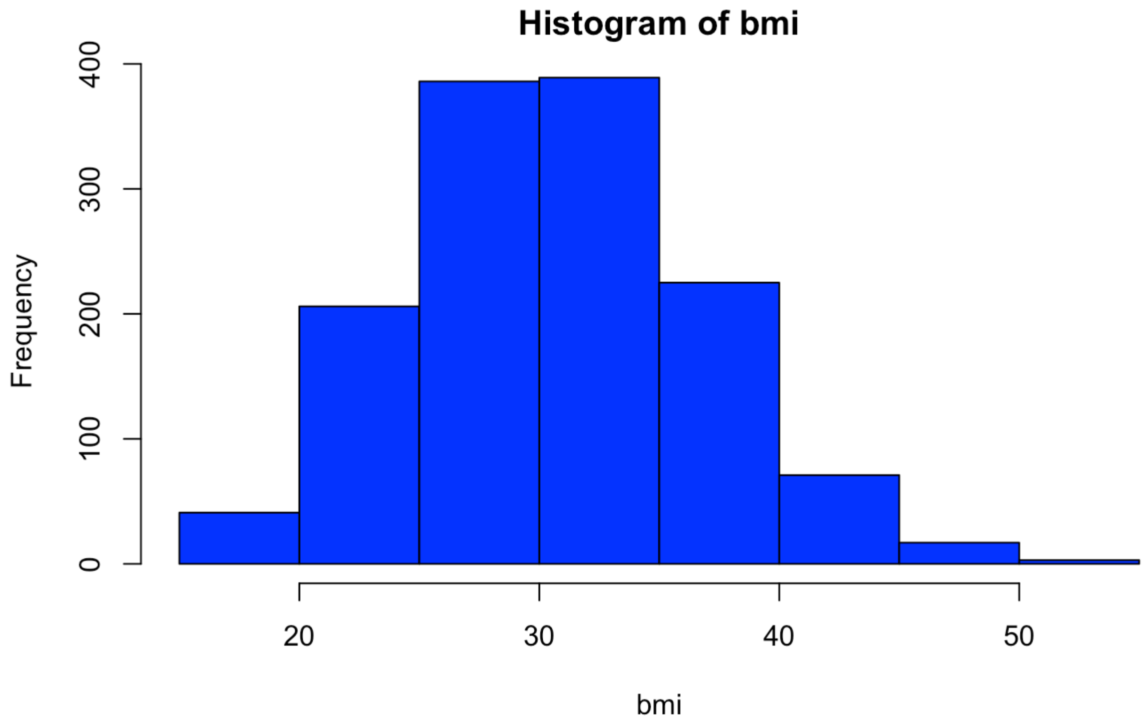


Figure 1.3 Histogram of age



1.

Figure 1.4 Histogram of bmi

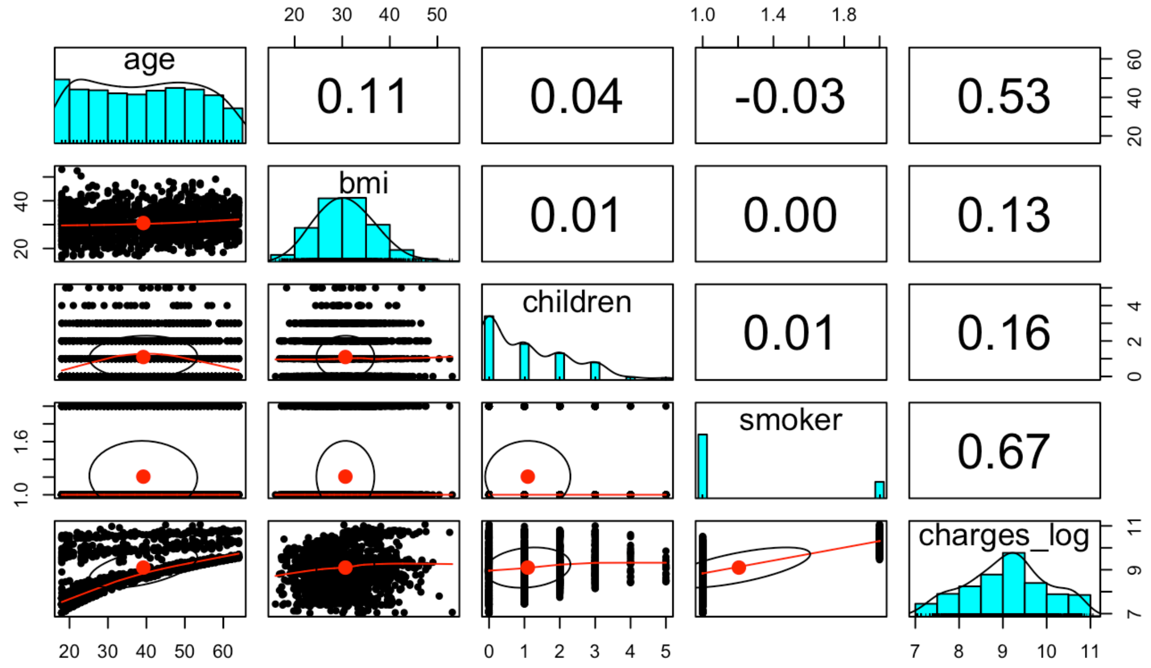


Figure 1.5 Scatter plot matrix

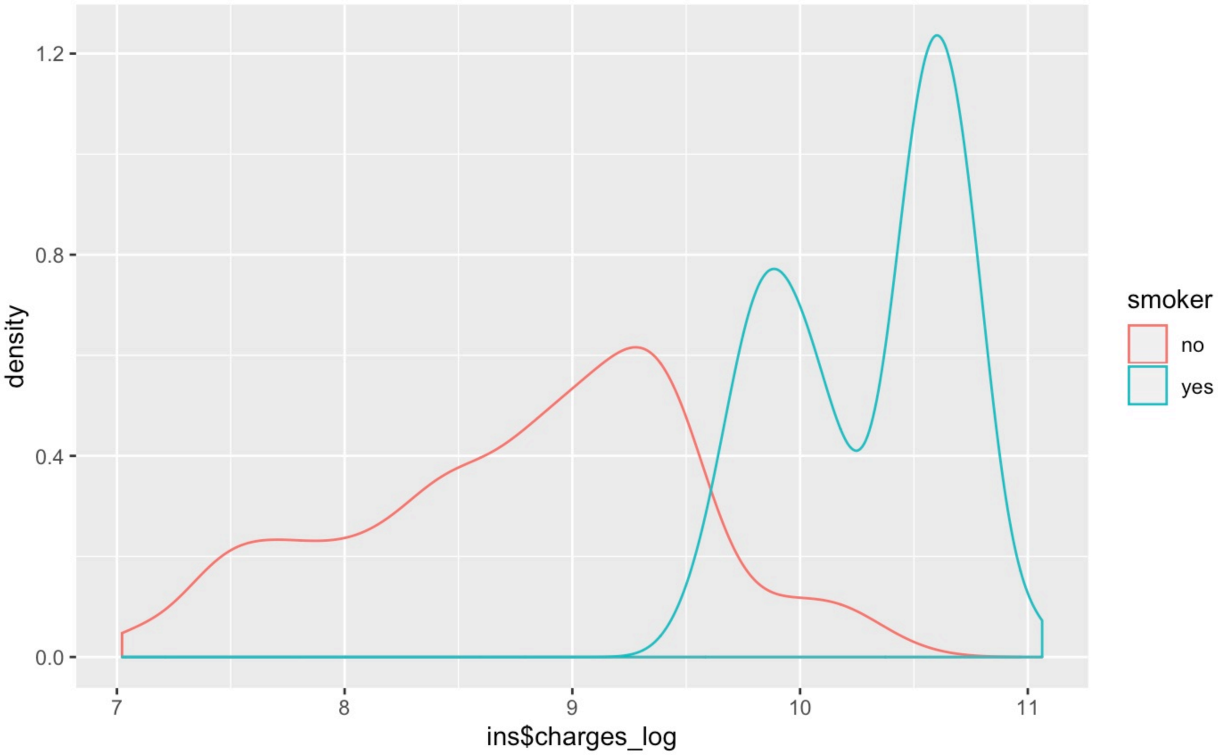
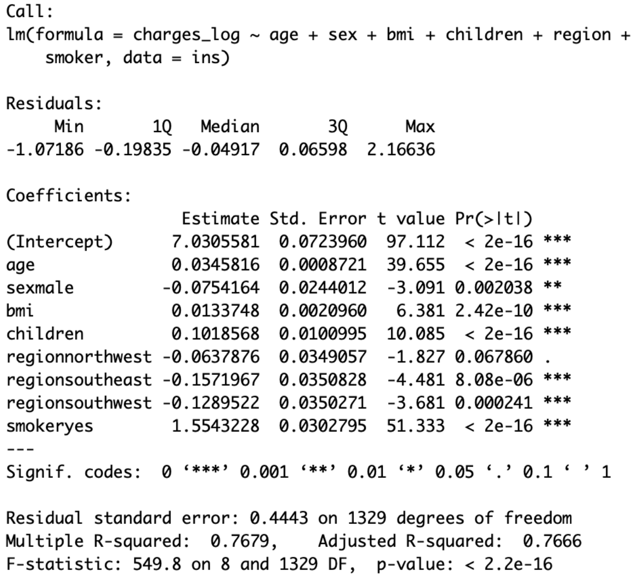
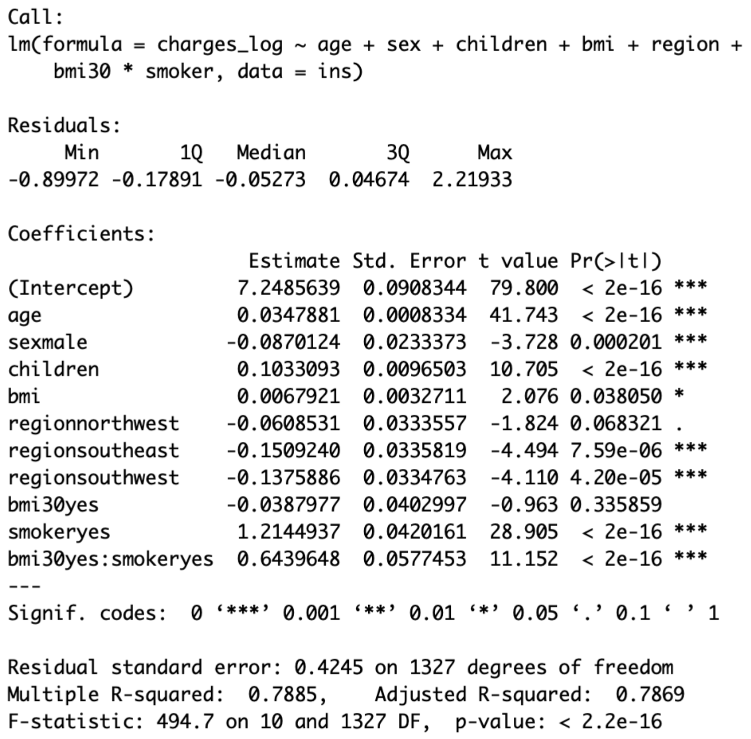


Figure 1.8 Density plot of log_charge versus smoker



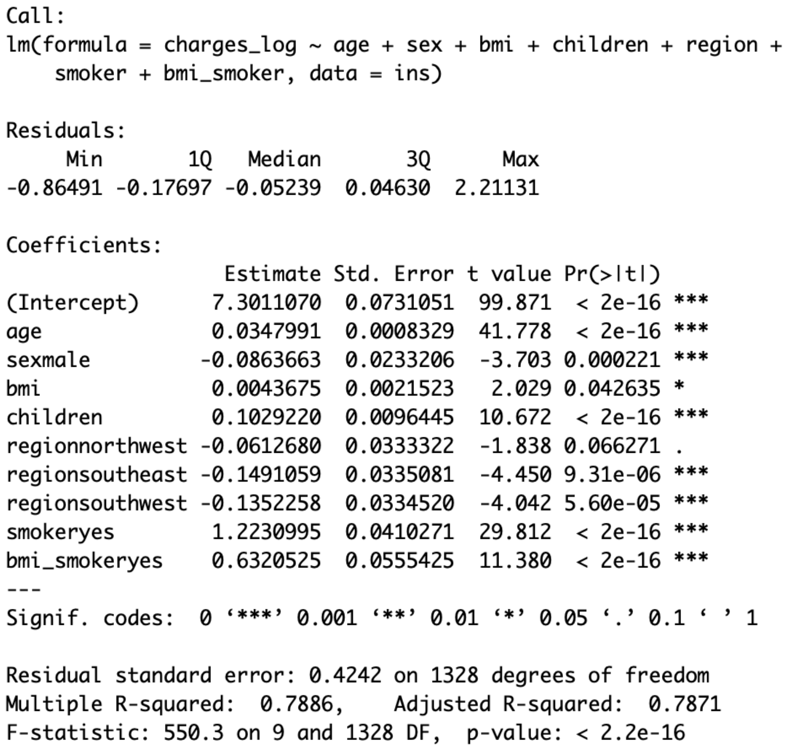
```
Call:
lm(formula = charges_log ~ age + sex + bmi + children + region +
    smoker, data = ins)

Residuals:
     Min       1Q   Median       3Q      Max
-1.07186 -0.19835 -0.04917  0.06598  2.16636

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      7.0305581  0.0723960  97.112  < 2e-16 ***
age              0.0345816  0.0008721  39.655  < 2e-16 ***
sexmale         -0.0754164  0.0244012  -3.091 0.002038 **
bmi              0.0133748  0.0020960   6.381 2.42e-10 ***
children         0.1018568  0.0100995  10.085  < 2e-16 ***
regionnorthwest -0.0637876  0.0349057  -1.827 0.067860 .
regionsoutheast -0.1571967  0.0350828  -4.481 8.08e-06 ***
regionsouthwest -0.1289522  0.0350271  -3.681 0.000241 ***
smokeryes        1.5543228  0.0302795  51.333  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4443 on 1329 degrees of freedom
Multiple R-squared:  0.7679,    Adjusted R-squared:  0.7666
F-statistic: 549.8 on 8 and 1329 DF,  p-value: < 2.2e-16
```

Figure 2.1



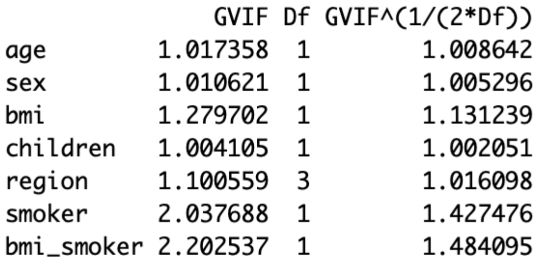
```
Call:
lm(formula = charges_log ~ age + sex + children + bmi + region + 
    bmi30 * smoker, data = ins)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.89972 -0.17891 -0.05273  0.04674  2.21933 

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)         7.2485639  0.0908344  79.800  < 2e-16 ***
age                 0.0347881  0.0008334  41.743  < 2e-16 ***
sexmale            -0.0870124  0.0233373  -3.728 0.000201 ***
children            0.1033093  0.0096503  10.705  < 2e-16 ***
bmi                 0.0067921  0.0032711   2.076 0.038050 *  
regionnorthwest    -0.0608531  0.0333557  -1.824 0.068321 .  
regionsoutheast    -0.1509240  0.0335819  -4.494 7.59e-06 ***
regionsouthwest    -0.1375886  0.0334763  -4.110 4.20e-05 ***
bmi30yes           -0.0387977  0.0402997  -0.963 0.335859    
smokeryes           1.2144937  0.0420161  28.905  < 2e-16 ***
bmi30yes:smokeryes  0.6439648  0.0577453  11.152  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4245 on 1327 degrees of freedom
Multiple R-squared:  0.7885,	Adjusted R-squared:  0.7869 
F-statistic: 494.7 on 10 and 1327 DF,  p-value: < 2.2e-16
```

Figure 2.2



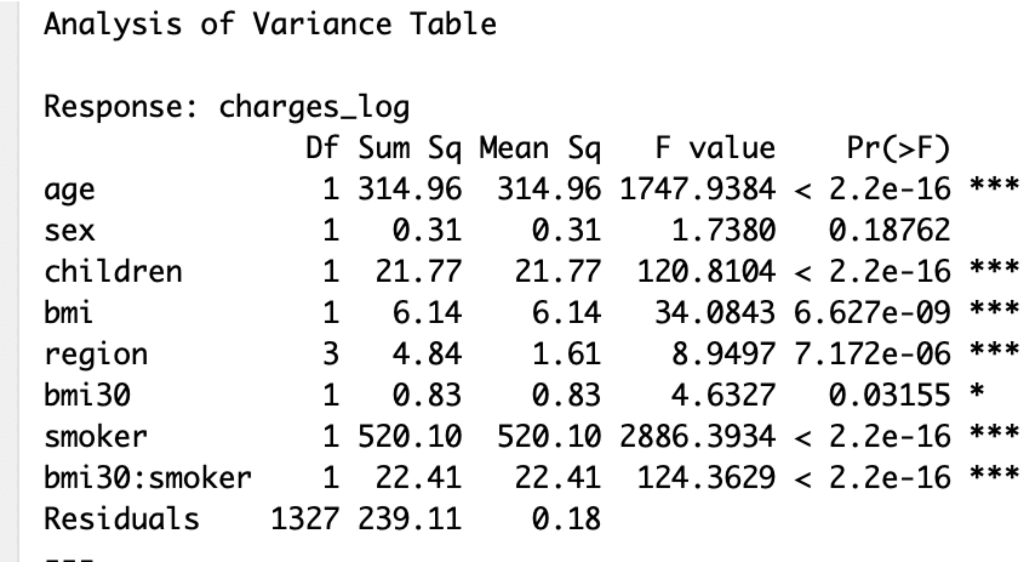
```
Call:
lm(formula = charges_log ~ age + sex + bmi + children + region + 
    smoker + bmi_smoker, data = ins)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.86491 -0.17697 -0.05239  0.04630  2.21131 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)      7.3011070  0.0731051  99.871  < 2e-16 ***
age              0.0347991  0.0008329  41.778  < 2e-16 ***
sexmale         -0.0863663  0.0233206  -3.703 0.000221 ***
bmi              0.0043675  0.0021523   2.029 0.042635 *  
children         0.1029220  0.0096445  10.672  < 2e-16 ***
regionnorthwest -0.0612680  0.0333322  -1.838 0.066271 .  
regionsoutheast -0.1491059  0.0335081  -4.450 9.31e-06 ***
regionsouthwest -0.1352258  0.0334520  -4.042 5.60e-05 ***
smokeryes        1.2230995  0.0410271  29.812  < 2e-16 ***
bmi_smokeryes    0.6320525  0.0555425  11.380  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4242 on 1328 degrees of freedom
Multiple R-squared:  0.7886,	Adjusted R-squared:  0.7871 
F-statistic: 550.3 on 9 and 1328 DF,  p-value: < 2.2e-16
```

Figure 2.3



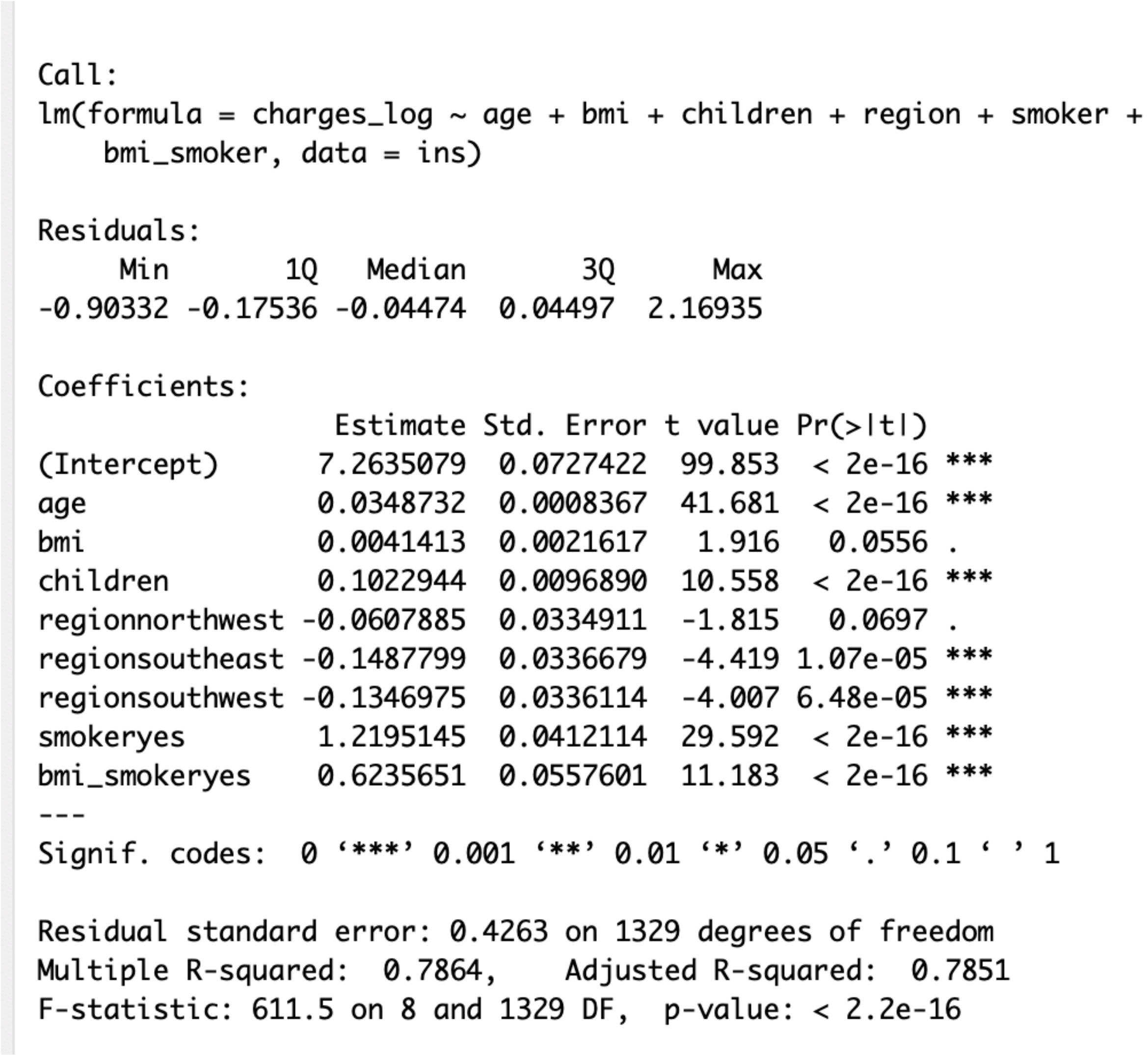
```
                GVIF Df GVIF^(1/(2*Df))
age         1.017358  1        1.008642
sex         1.010621  1        1.005296
bmi         1.279702  1        1.131239
children    1.004105  1        1.002051
region      1.100559  3        1.016098
smoker      2.037688  1        1.427476
bmi_smoker  2.202537  1        1.484095
```

Figure 2.4 VIF



```
Analysis of Variance Table

Response: charges_log
               Df Sum Sq Mean Sq   F value    Pr(>F)    
age             1 314.96  314.96 1747.9384 < 2.2e-16 ***
sex             1   0.31    0.31    1.7380   0.18762    
children        1  21.77   21.77  120.8104 < 2.2e-16 ***
bmi             1   6.14    6.14   34.0843 6.627e-09 ***
region          3   4.84    1.61    8.9497 7.172e-06 ***
bmi30           1   0.83    0.83    4.6327   0.03155 *  
smoker          1 520.10  520.10 2886.3934 < 2.2e-16 ***
bmi30:smoker    1  22.41   22.41  124.3629 < 2.2e-16 ***
Residuals    1327 239.11    0.18                        
---
```

Figure 2.5



```
Call:
lm(formula = charges_log ~ age + bmi + children + region + smoker +
    bmi_smoker, data = ins)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.90332 -0.17536 -0.04474  0.04497  2.16935 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)      7.2635079  0.0727422  99.853  < 2e-16 ***
age              0.0348732  0.0008367  41.681  < 2e-16 ***
bmi              0.0041413  0.0021617   1.916   0.0556 .  
children         0.1022944  0.0096890  10.558  < 2e-16 ***
regionnorthwest -0.0607885  0.0334911  -1.815   0.0697 .  
regionsoutheast -0.1487799  0.0336679  -4.419 1.07e-05 ***
regionsouthwest -0.1346975  0.0336114  -4.007 6.48e-05 ***
smokeryes        1.2195145  0.0412114  29.592  < 2e-16 ***
bmi_smokeryes    0.6235651  0.0557601  11.183  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4263 on 1329 degrees of freedom
Multiple R-squared:  0.7864,	Adjusted R-squared:  0.7851 
F-statistic: 611.5 on 8 and 1329 DF,  p-value: < 2.2e-16
```

Figure 2.6 Final linear model

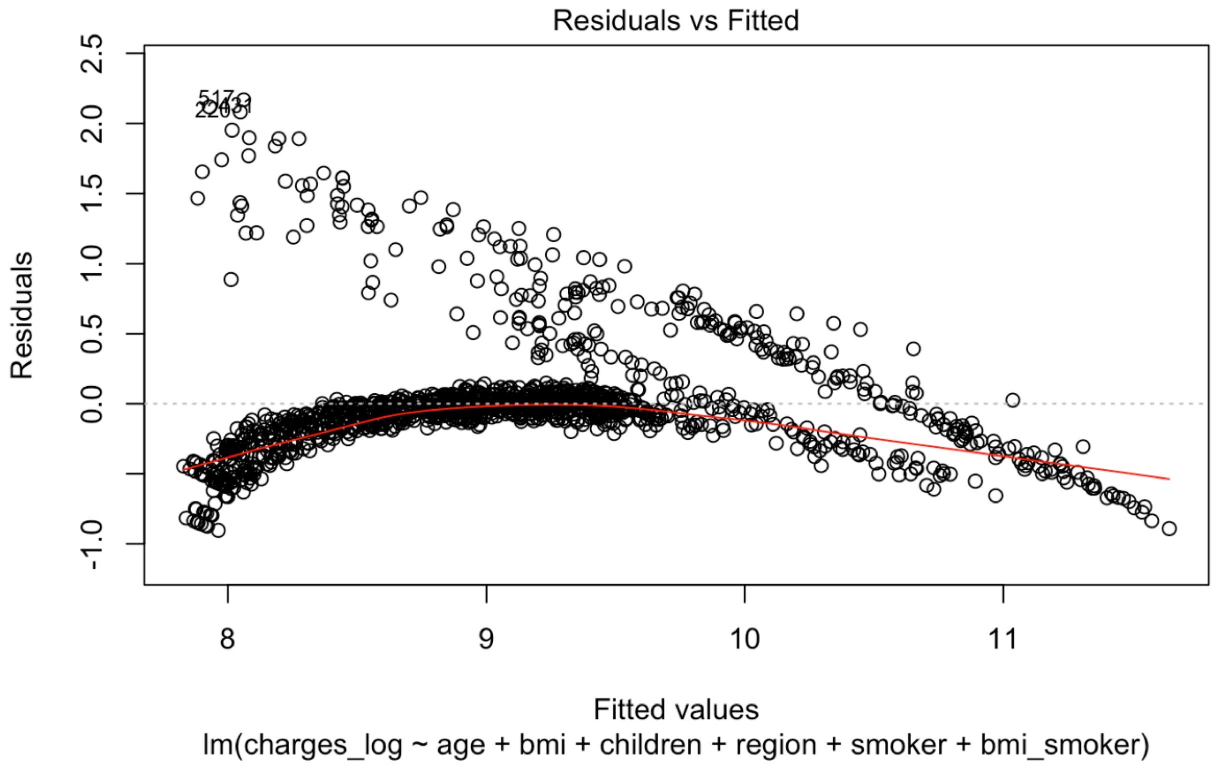
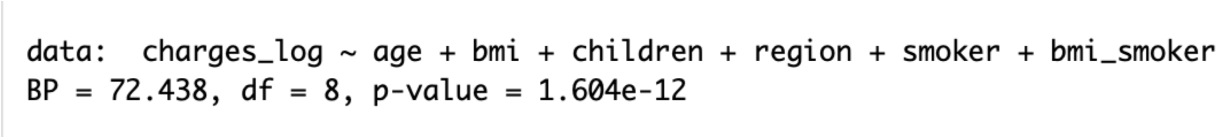


Figure 2.7



```
data:  charges_log ~ age + bmi + children + region + smoker + bmi_smoker
BP = 72.438, df = 8, p-value = 1.604e-12
```

Figure 2.8 BP test

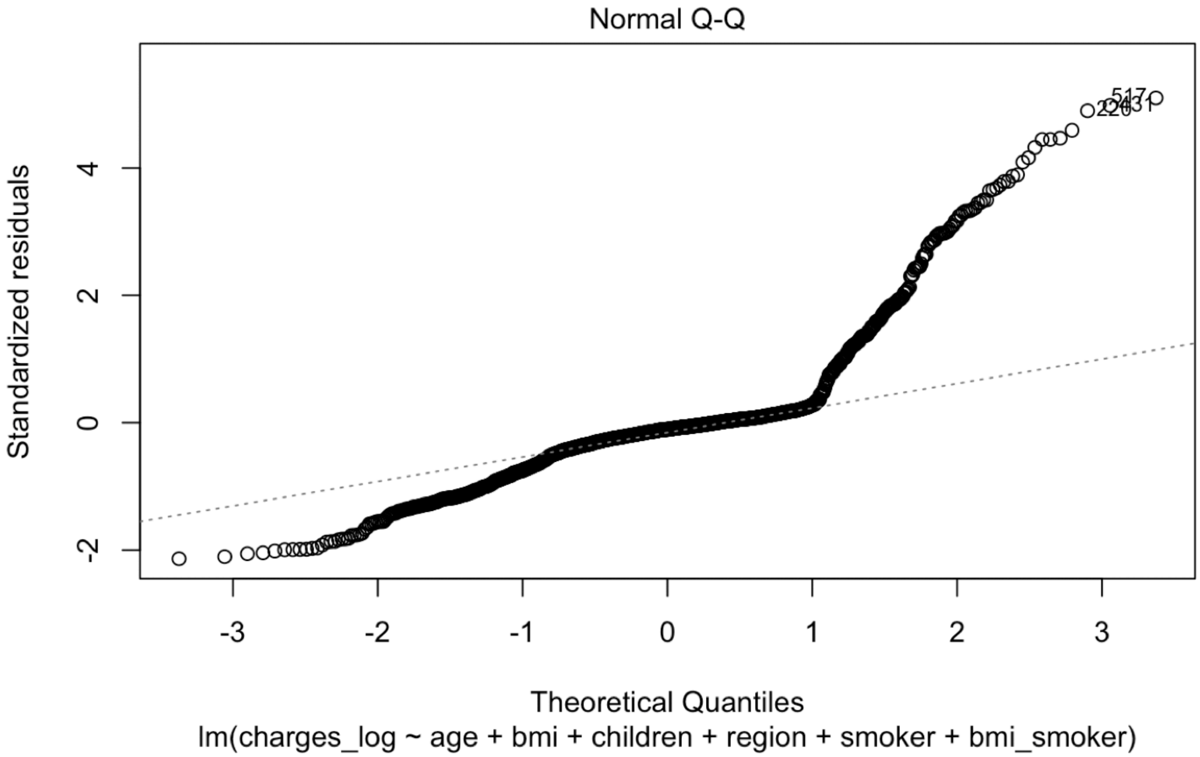
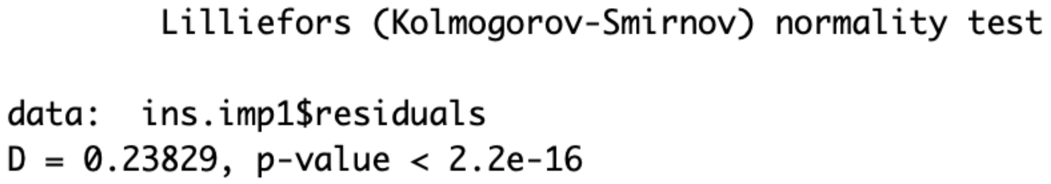
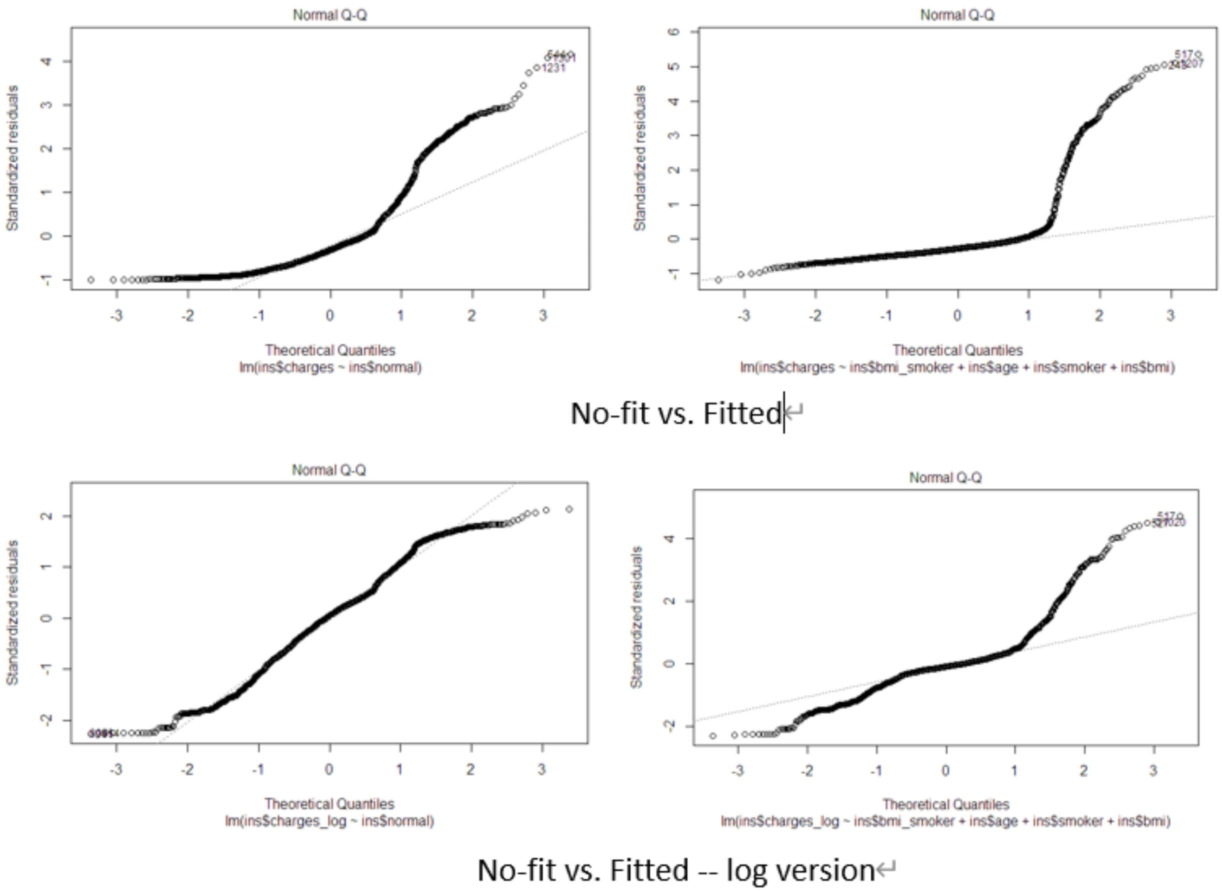


Figure 2.9



```
        Lilliefors (Kolmogorov-Smirnov) normality test

data:  ins.imp1$residuals
D = 0.23829, p-value < 2.2e-16
```

Figure 2.10



No-fit vs. Fitted

No-fit vs. Fitted -- log version

Figure 2.11 QQ plot comparison

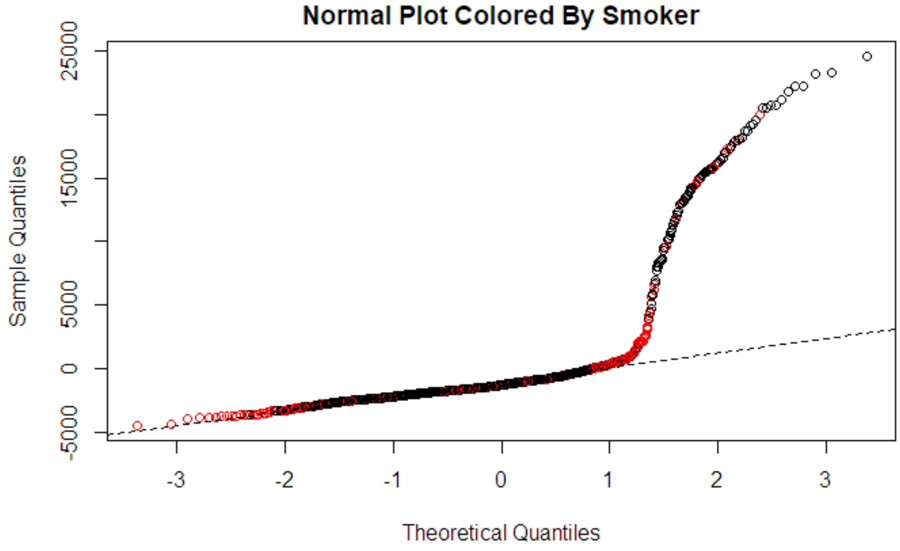


Figure 2.12 Colored QQ plot

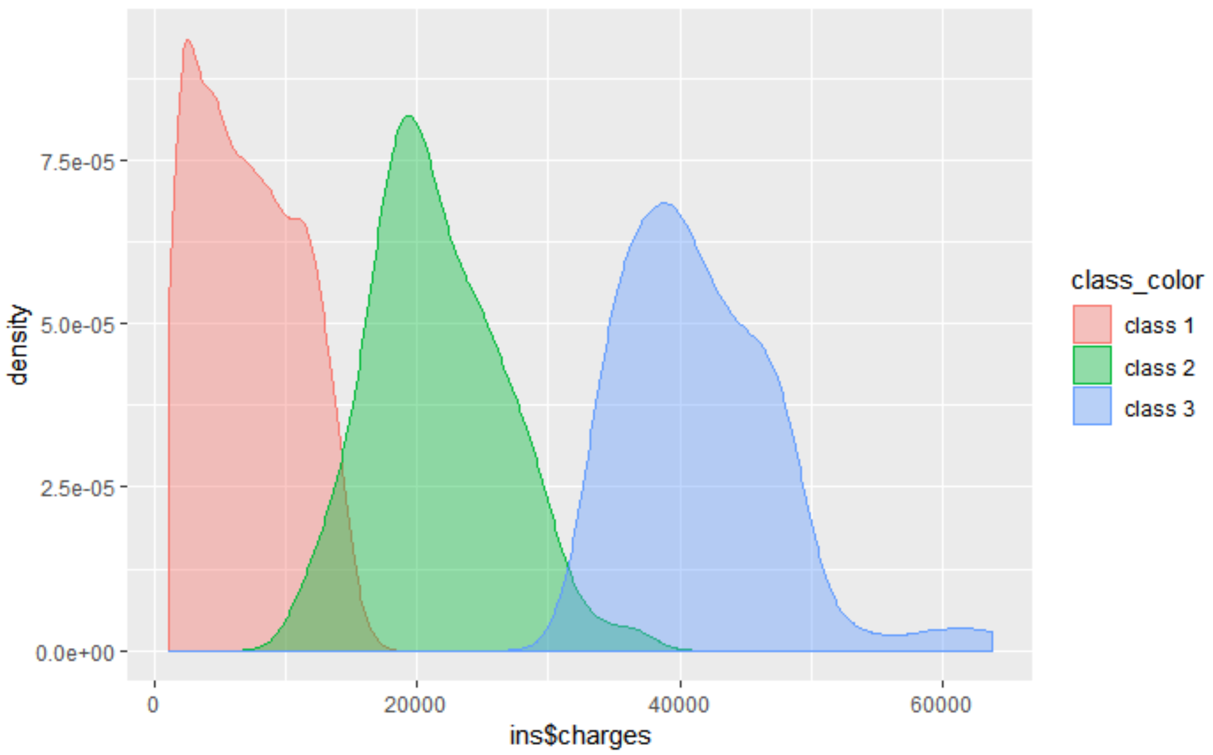


Figure 2.13 Density plot of three classes

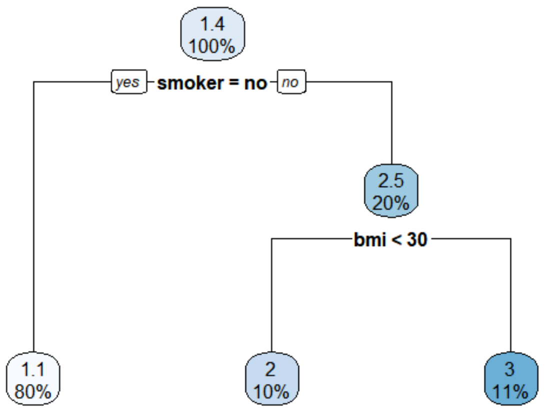


Figure 2.14 Decision tree

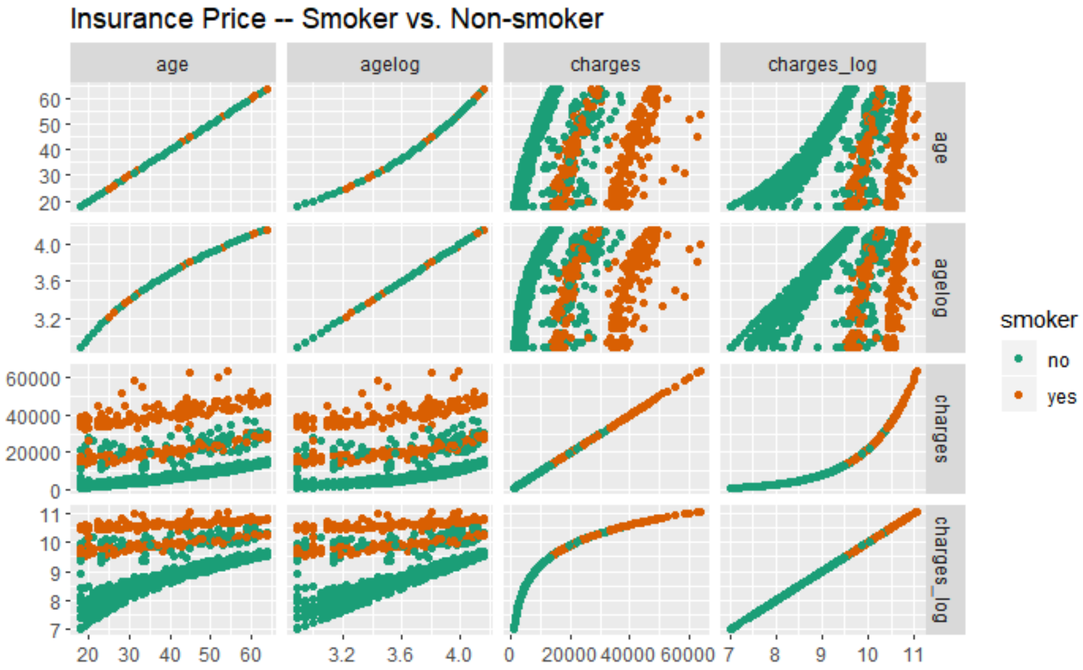
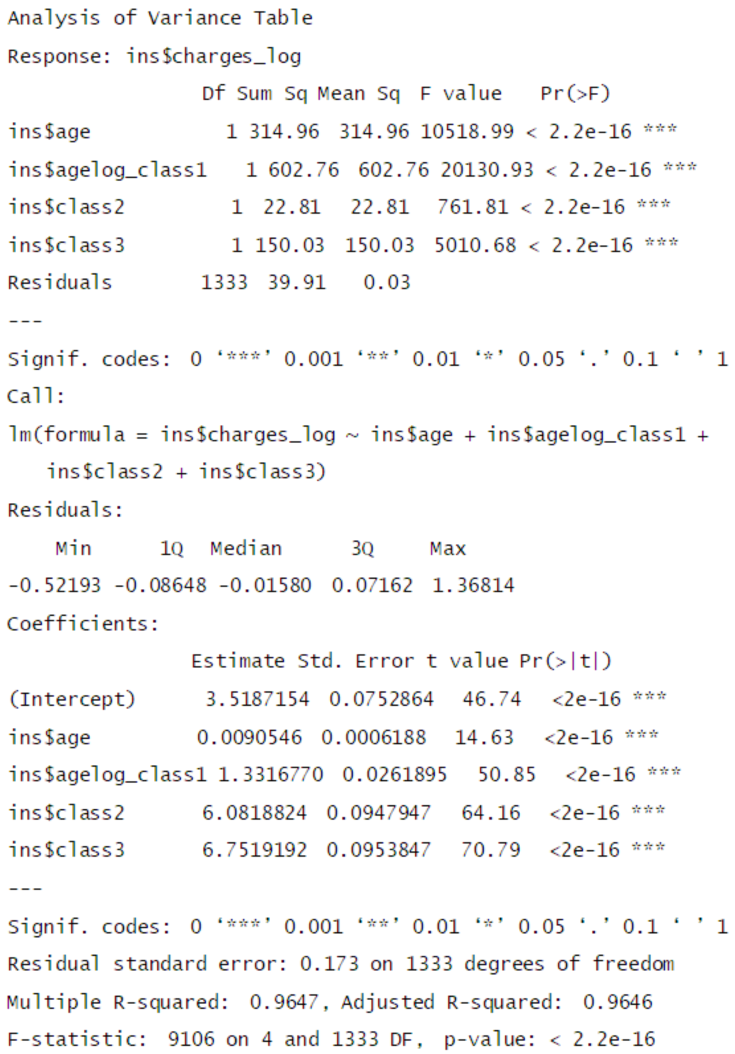


Figure 2.15 Pair plot



```
Analysis of Variance Table
Response: ins$charges_log
                    Df Sum Sq Mean Sq  F value    Pr(>F)    
ins$age              1 314.96  314.96 10518.99 < 2.2e-16 ***
ins$agelog_class1    1 602.76  602.76 20130.93 < 2.2e-16 ***
ins$class2           1  22.81   22.81   761.81 < 2.2e-16 ***
ins$class3           1 150.03  150.03  5010.68 < 2.2e-16 ***
Residuals         1333  39.91    0.03                       
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Call:
lm(formula = ins$charges_log ~ ins$age + ins$agelog_class1 + 
    ins$class2 + ins$class3)
Residuals:
     Min       1Q   Median       3Q      Max 
-0.52193 -0.08648 -0.01580  0.07162  1.36814 
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)       3.5187154  0.0752864   46.74   <2e-16 ***
ins$age           0.0090546  0.0006188   14.63   <2e-16 ***
ins$agelog_class1 1.3316770  0.0261895   50.85   <2e-16 ***
ins$class2        6.0818824  0.0947947   64.16   <2e-16 ***
ins$class3        6.7519192  0.0953847   70.79   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 0.173 on 1333 degrees of freedom
Multiple R-squared:  0.9647, Adjusted R-squared:  0.9646 
F-statistic:  9106 on 4 and 1333 DF,  p-value: < 2.2e-16
```

Figure 2.16 Regression model after introducing log_age

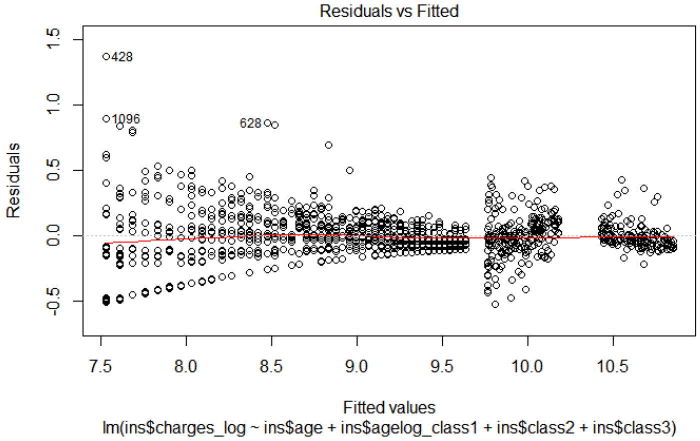
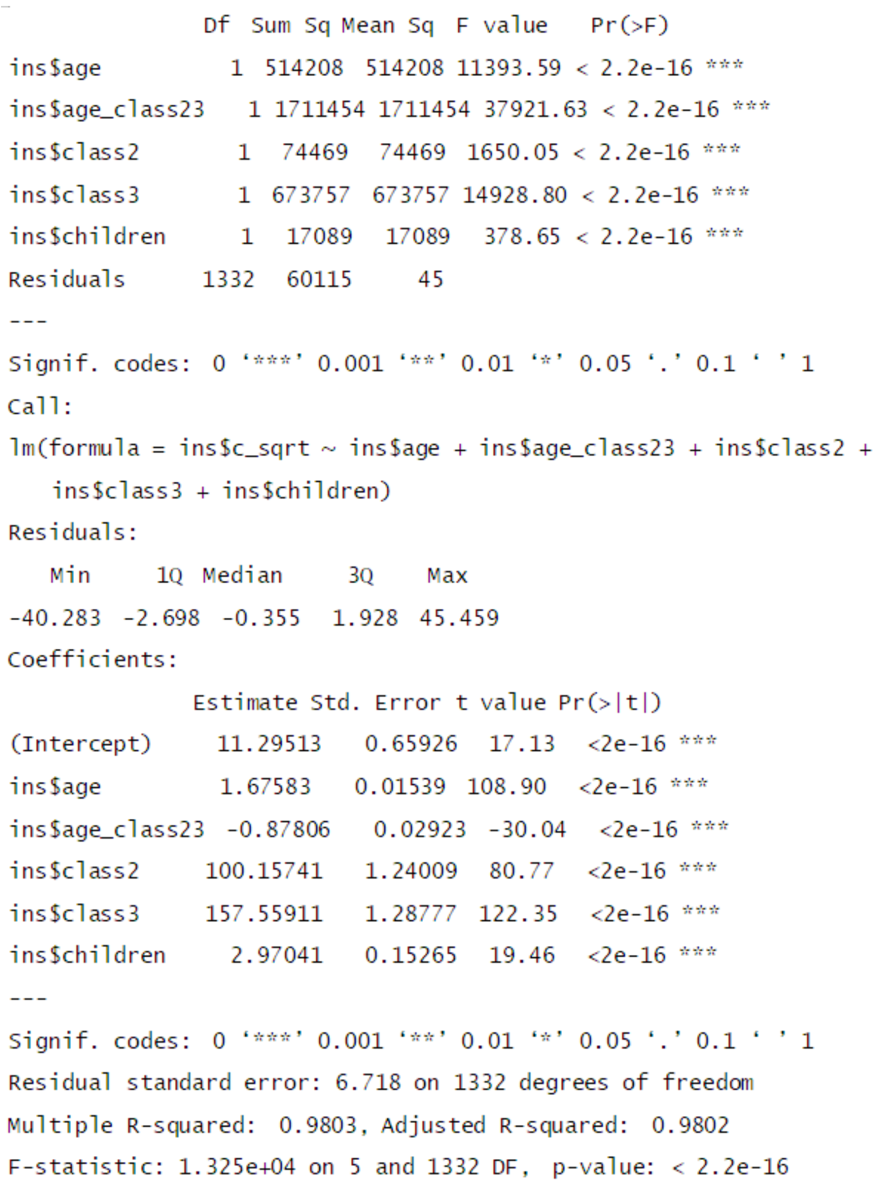


Figure 2.17 Residual versus fitted plot



```
                Df  Sum Sq Mean Sq  F value    Pr(>F)    
ins$age          1  514208  514208 11393.59 < 2.2e-16 ***
ins$age_class23   1 1711454 1711454 37921.63 < 2.2e-16 ***
ins$class2       1   74469   74469  1650.05 < 2.2e-16 ***
ins$class3       1  673757  673757 14928.80 < 2.2e-16 ***
ins$children     1   17089   17089   378.65 < 2.2e-16 ***
Residuals     1332   60115      45                       
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Call:
lm(formula = ins$c_sqrt ~ ins$age + ins$age_class23 + ins$class2 + 
    ins$class3 + ins$children)
Residuals:
    Min      1Q  Median      3Q     Max 
-40.283  -2.698  -0.355   1.928  45.459 
Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)     11.29513    0.65926   17.13   <2e-16 ***
ins$age          1.67583    0.01539  108.90   <2e-16 ***
ins$age_class23  -0.87806    0.02923  -30.04   <2e-16 ***
ins$class2     100.15741    1.24009   80.77   <2e-16 ***
ins$class3     157.55911    1.28777  122.35   <2e-16 ***
ins$children     2.97041    0.15265   19.46   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 6.718 on 1332 degrees of freedom
Multiple R-squared:  0.9803, Adjusted R-squared:  0.9802 
F-statistic: 1.325e+04 on 5 and 1332 DF,  p-value: < 2.2e-16
```

Figure 2.18 Regression model adding children

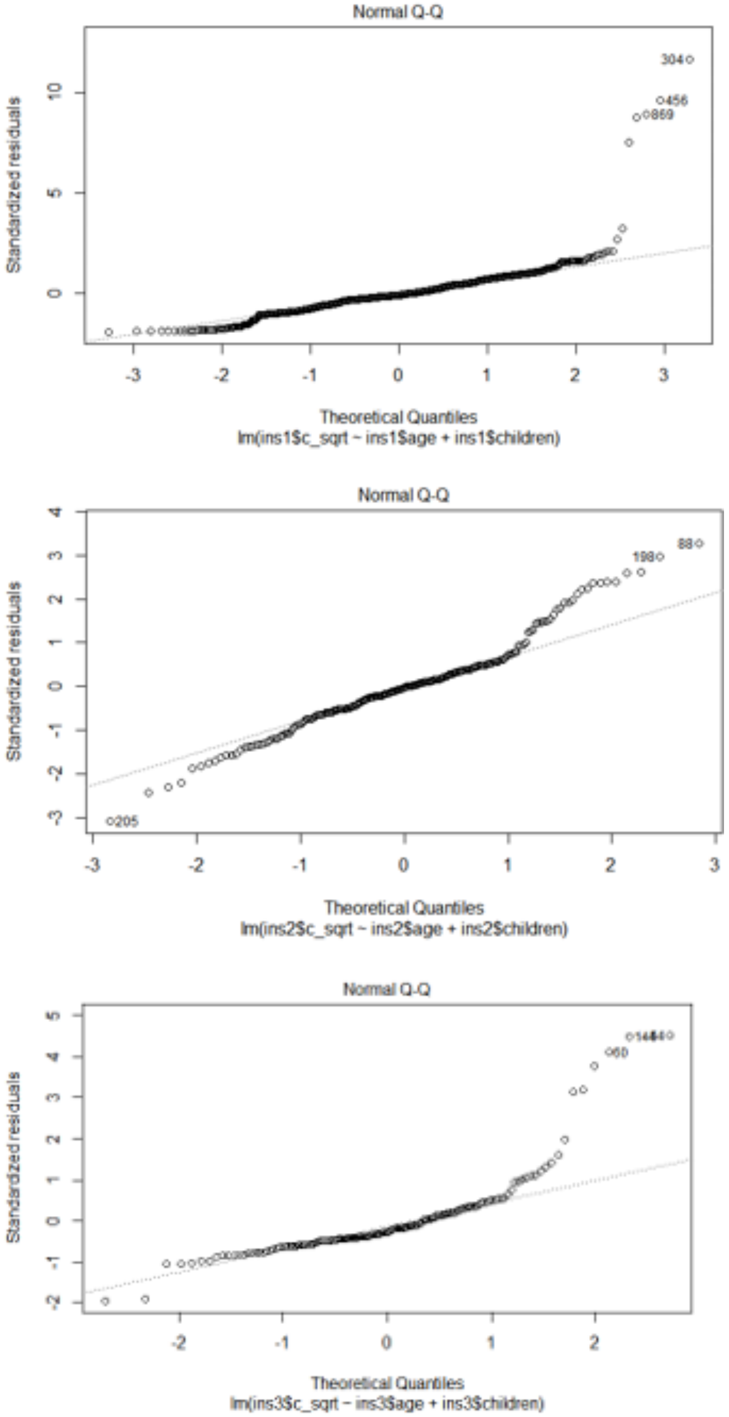


Figure 2.19 QQ plot of each class

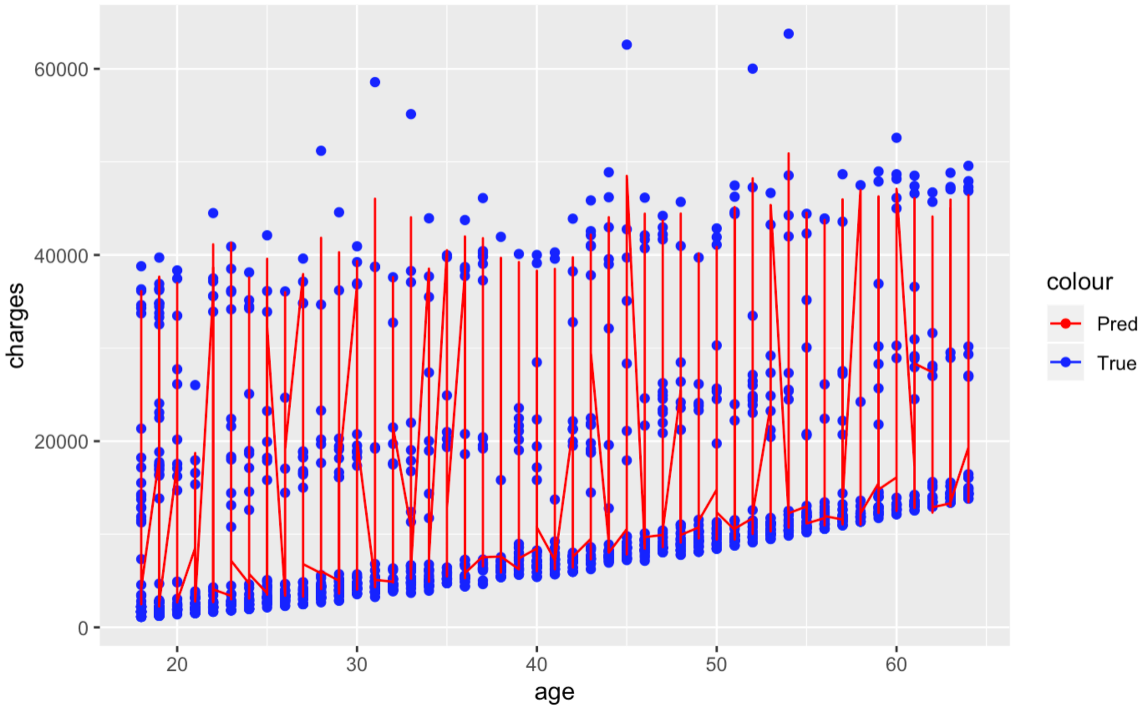


Figure 3.1 Age versus Charges

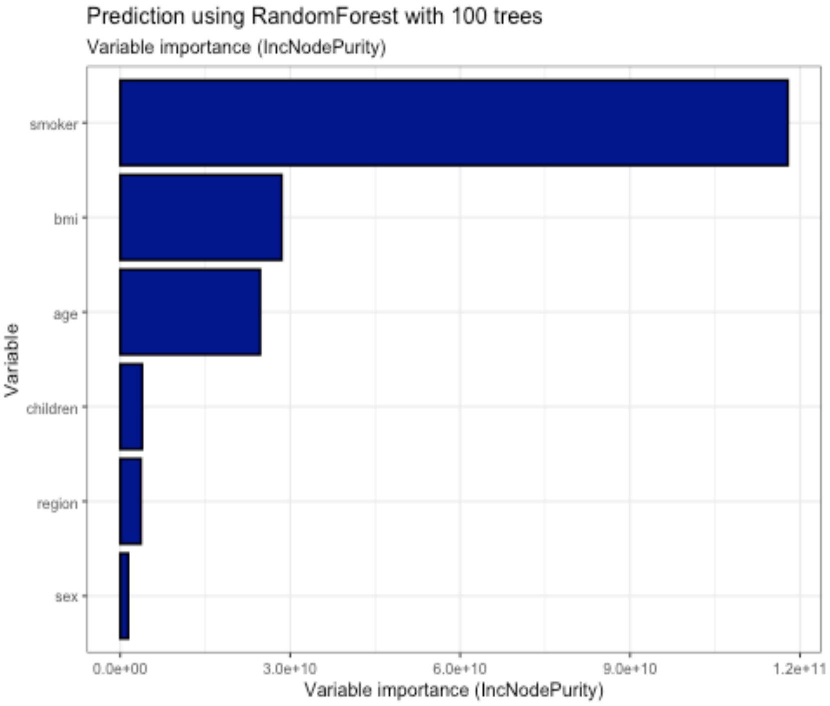


Figure 3.2 Variable Importance

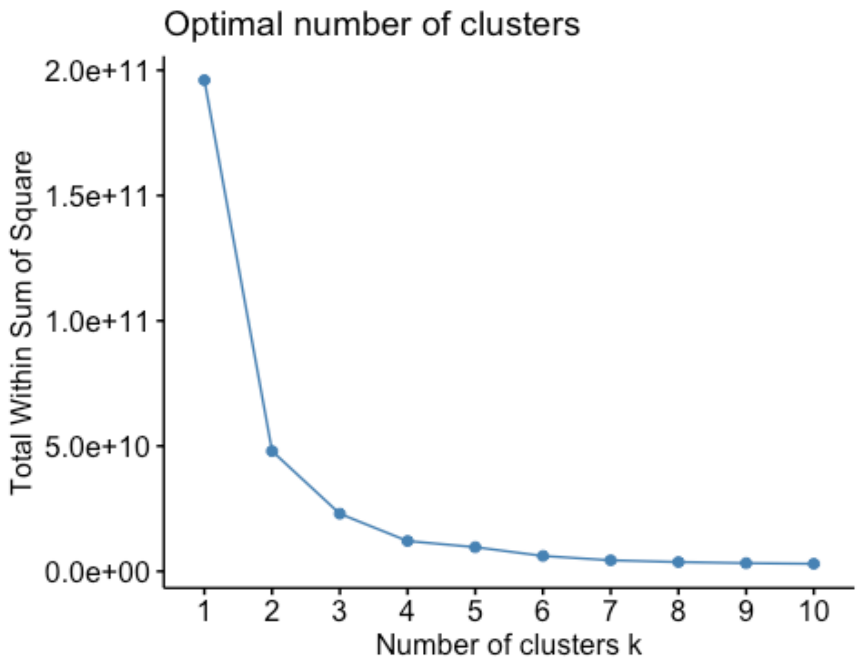


Figure 3.3 Choose K

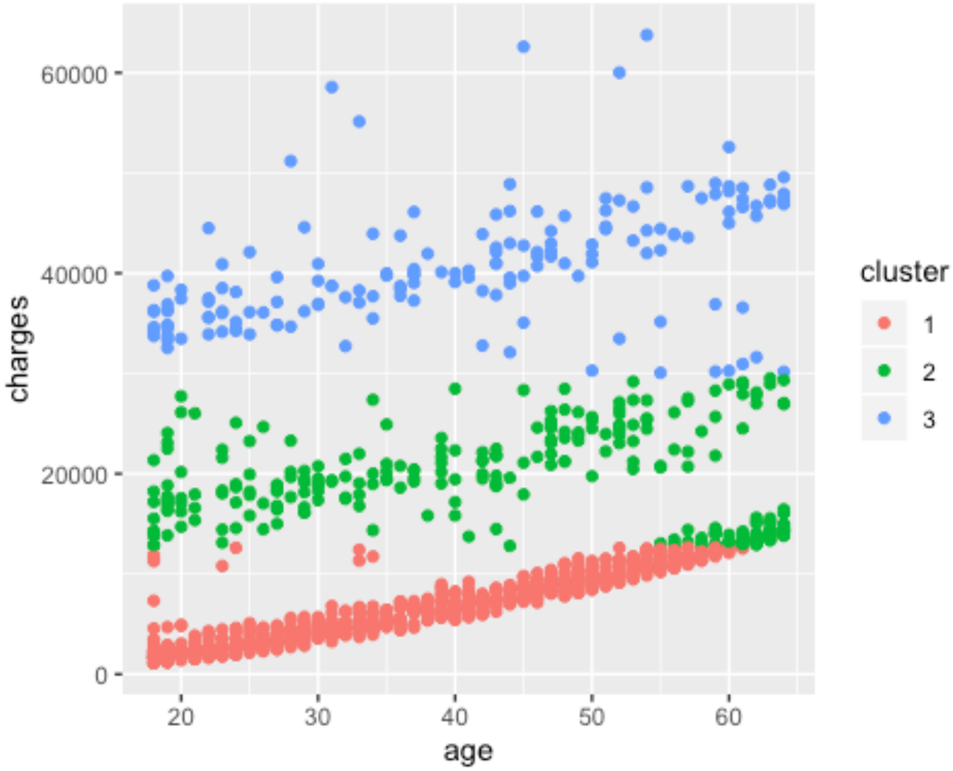


Figure 3.4 Age versus Charges

age bmi children charges_log

```
age         1.0000000 0.1092719 0.0424690   0.5278340
bmi         0.1092719 1.0000000 0.0127589   0.1326694
children    0.0424690 0.0127589 1.0000000   0.1613363
charges_log 0.5278340 0.1326694 0.1613363   1.0000000
```

Table 1.0 Correlation matrix of variables

Cluster	Age	BMI	Children	Charges
Lowest mean charges	37.12275	30.48004	1.090090	6430.148
Medium mean charges	44.97569	28.87635	1.069444	18897.644
Highest mean charges	40.37654	34.84543	1.166667	40761.309

Table 1.1 Mean values of the continuous variables

Sex		
Cluster	Female	Male
Lowest mean charges	0.5067568	0.4943432
Medium mean charges	0.53125	0.46875
Highest mean charges	0.3641975	0.6358025

Table 2.1 Proportions for categorical variables--Sex

Smoker		
Cluster	No	Yes
Lowest mean charges	1	0
Medium mean charges	0.5763889	0.4236111
Highest mean charges	0.0617284	0.9382716

Table 2.2 Proportions for categorical variables--Smoker

Region				
Cluster	Northeast	Northwest	Southeast	Southwest
Lowest mean charges	0.2240991	0.2545045	0.2578829	0.2635135
Medium mean charges	0.3125000	0.2430556	0.2564944	0.1875000
Highest mean charges	0.2160494	0.1790123	0.3765432	0.2283951

Table 2.3 Proportions for categorical variables--Region