Insurance Cost Prediction:

Problem Statement:-

Predicting medical insurance cost of an individual based on his/her age, sex, BMI, etc.

The dataset was downloaded from Kaggle. Here is the link to the dataset:- https://www.kaggle.com/mirichoi0218/insurance (Medical costs personal dataset).

Table of Content:

Describing the data : No. of rows, columns etc.

Looking for Missing Values.

EDA (Univariate and Multivariate analysis)

Feature engineering.

Feature Scaling.

Fitting an initial model.

Analysis of the fitted model (ANOVA).

Feature selection based on the initial fitted model.

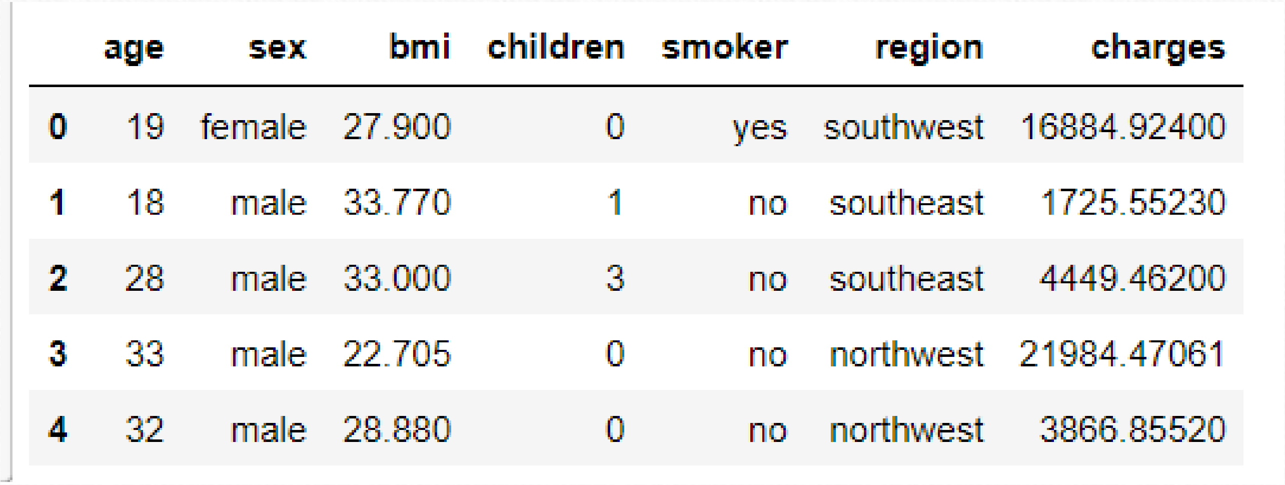
Fitting the final multiple linear regression model.

Checking important assumptions: Residual analysis, Homoscedasticity, Autocorrelation, Multi-Collinearity analysis .

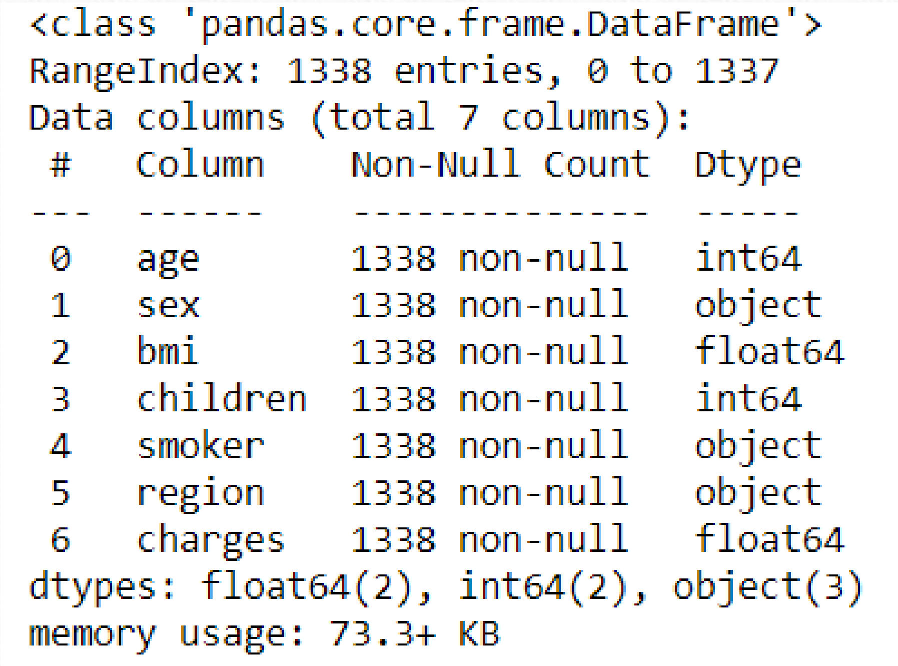
Real-time prediction from the model.

Bird’s eye view of dataset

There are 1338 rows and 7 columns in the dataset. The below figure shows the datatypes of all the variables.

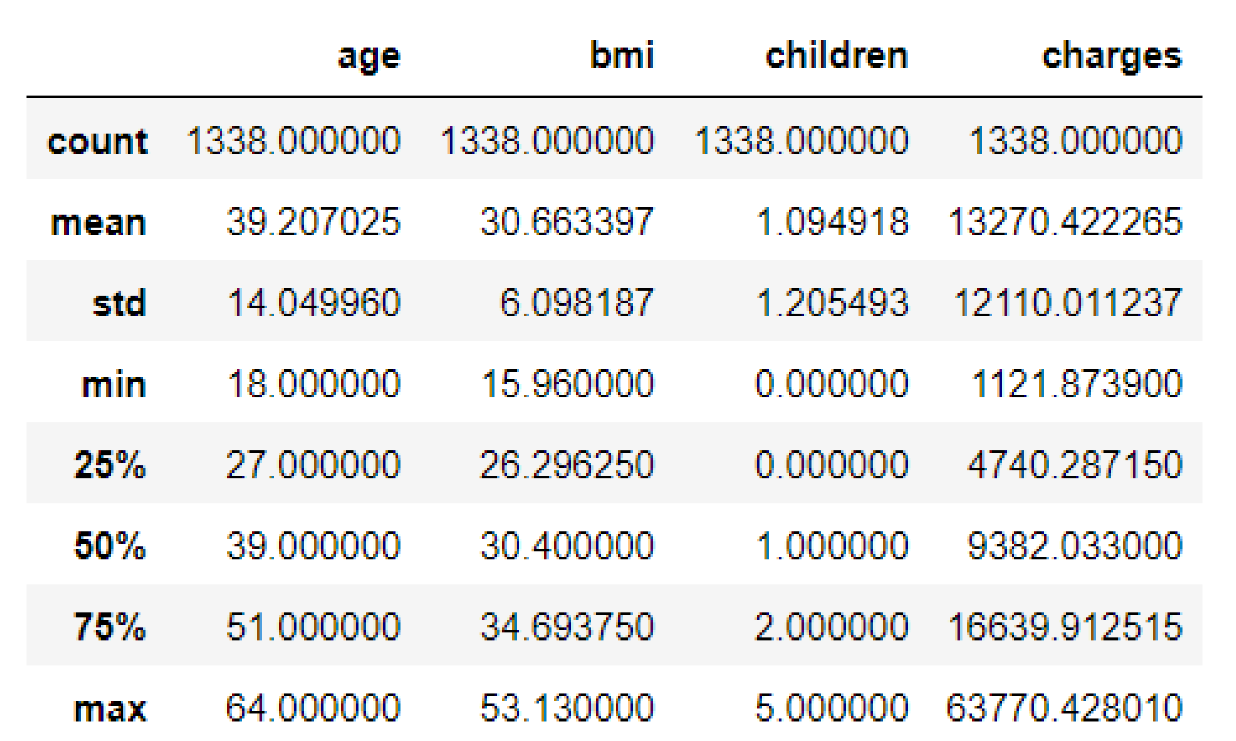


	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884.92400
1	18	male	33.770	1	no	southeast	1725.55230
2	28	male	33.000	3	no	southeast	4449.46200
3	33	male	22.705	0	no	northwest	21984.47061
4	32	male	28.880	0	no	northwest	3866.85520

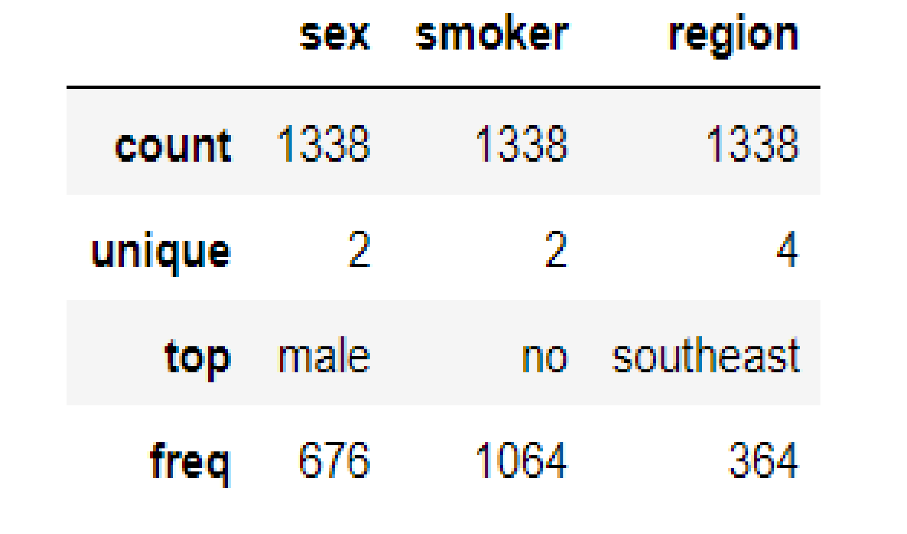


```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1338 entries, 0 to 1337
Data columns (total 7 columns):
 #   Column    Non-Null Count  Dtype  
---  ------    --------------  -----  
 0   age       1338 non-null   int64  
 1   sex       1338 non-null   object 
 2   bmi       1338 non-null   float64
 3   children  1338 non-null   int64  
 4   smoker    1338 non-null   object 
 5   region    1338 non-null   object 
 6   charges   1338 non-null   float64
dtypes: float64(2), int64(2), object(3)
memory usage: 73.3+ KB
```

Summary of numerical & categorical columns.

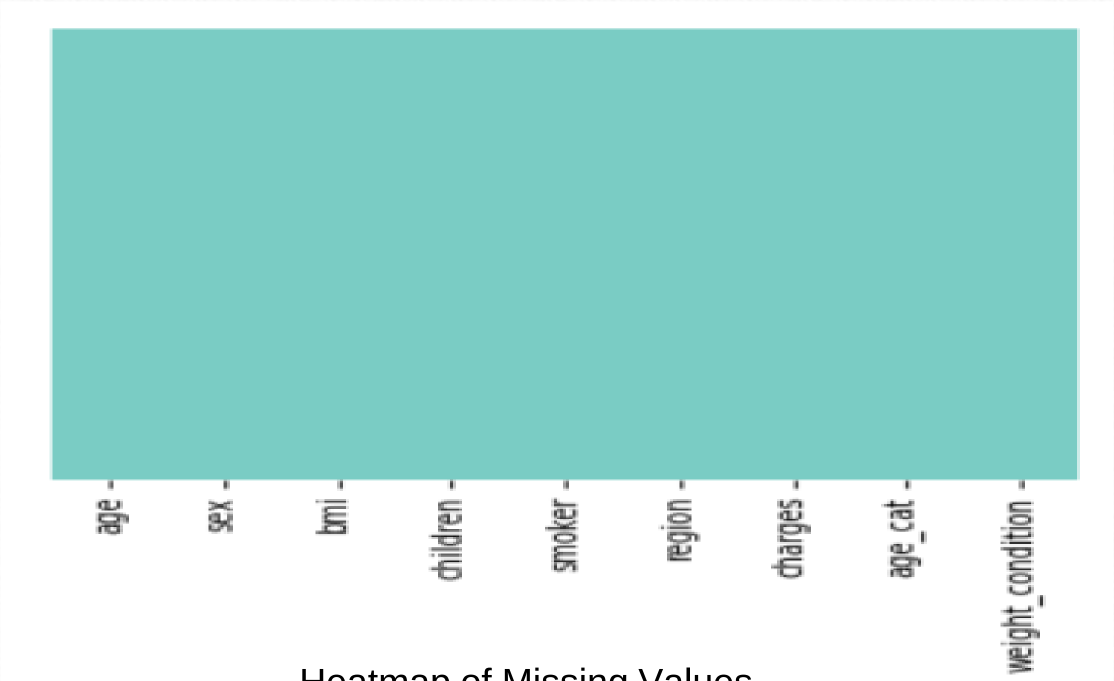


	age	bmi	children	charges
count	1338.000000	1338.000000	1338.000000	1338.000000
mean	39.207025	30.663397	1.094918	13270.422265
std	14.049960	6.098187	1.205493	12110.011237
min	18.000000	15.960000	0.000000	1121.873900
25%	27.000000	26.296250	0.000000	4740.287150
50%	39.000000	30.400000	1.000000	9382.033000
75%	51.000000	34.693750	2.000000	16639.912515
max	64.000000	53.130000	5.000000	63770.428010

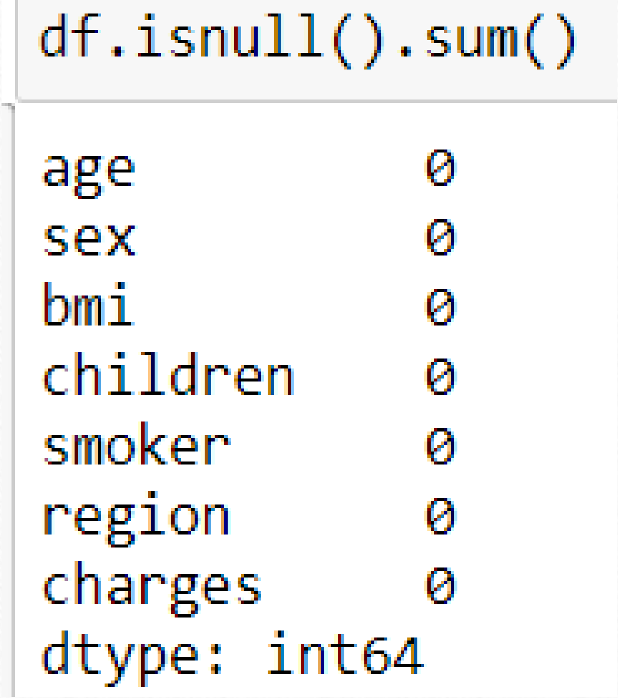


	sex	smoker	region
count	1338	1338	1338
unique	2	2	4
top	male	no	southeast
freq	676	1064	364

Missing Value Analysis:



Heatmap of Missing Values

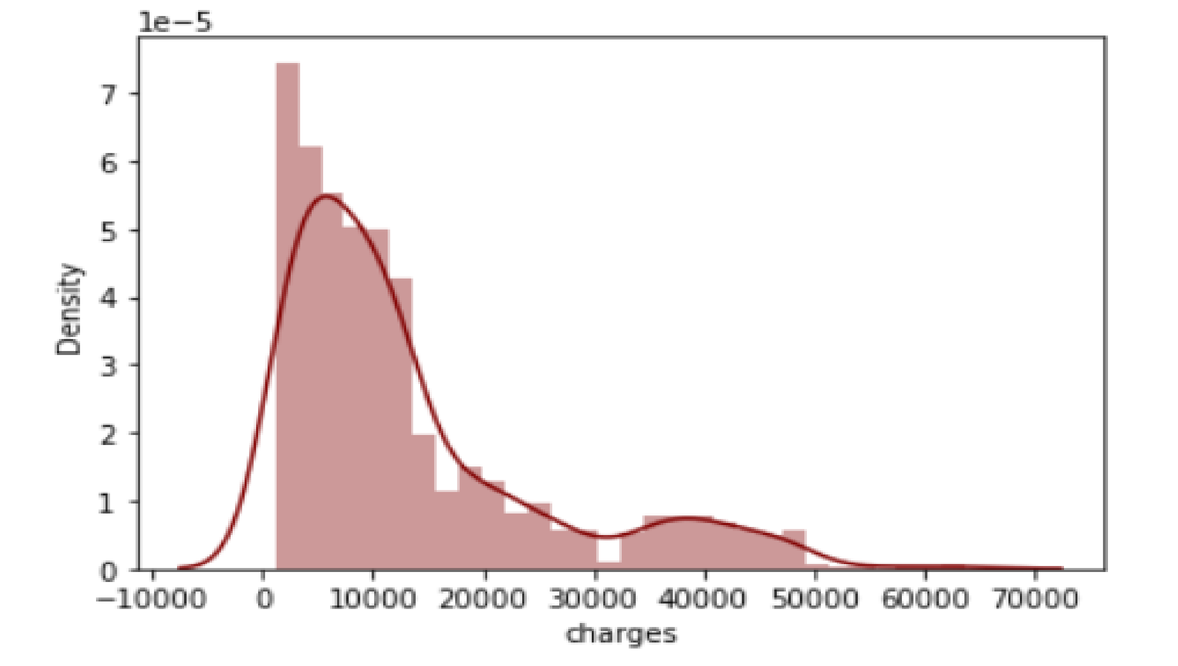


```
df.isnull().sum()

age         0
sex         0
bmi         0
children    0
smoker      0
region      0
charges     0
dtype: int64
```

The above heatmap signifies absence of missing values.

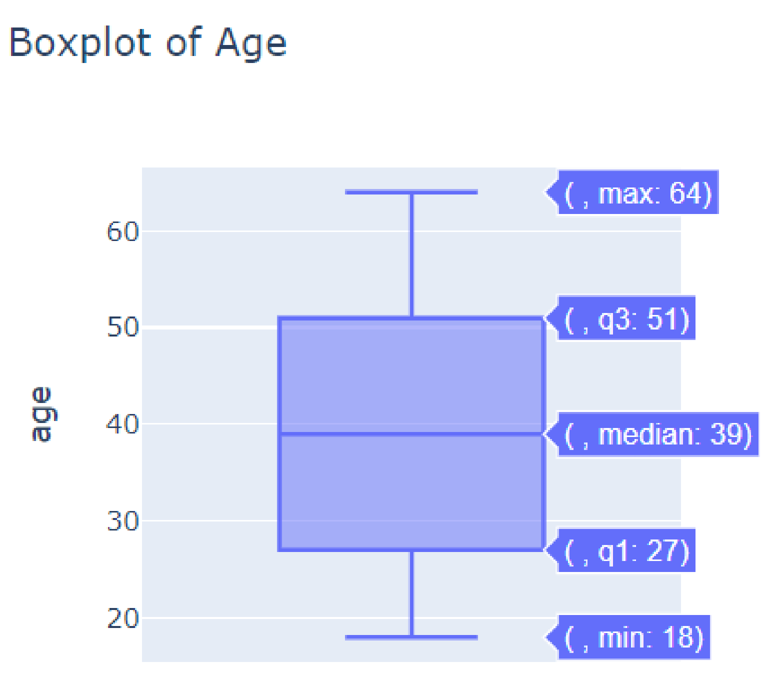
Exploratory Data Analysis (EDA)

Univariate Analysis:-



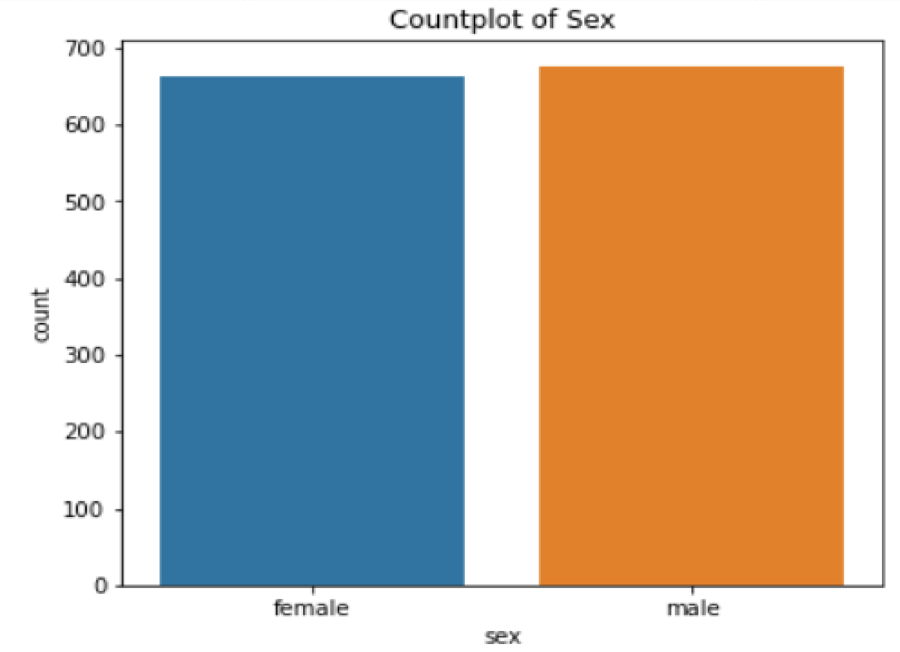
Plot of Charges

By just looking at the plot, we can figure out that “charges” which is our dependent variable has a right-skewed distribution with some extreme values of more than 60k.



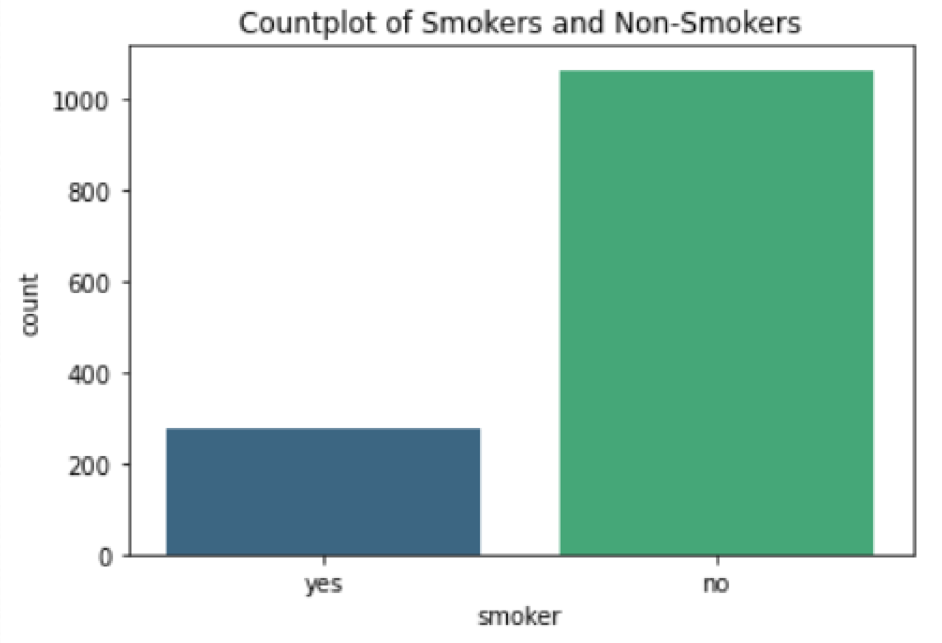
The figure on the left shows the boxplot of 'age'. Some important features are:-

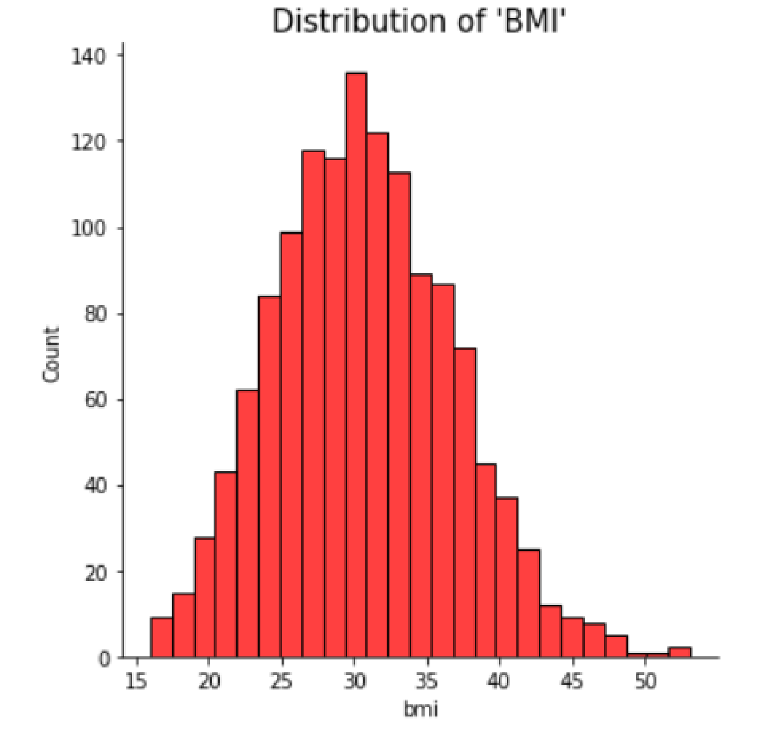
- Max age is 64 yrs.
- Median age is 39 yrs.
- Min age is 18 yrs.
- Standard deviation of age is 14.15 yrs.



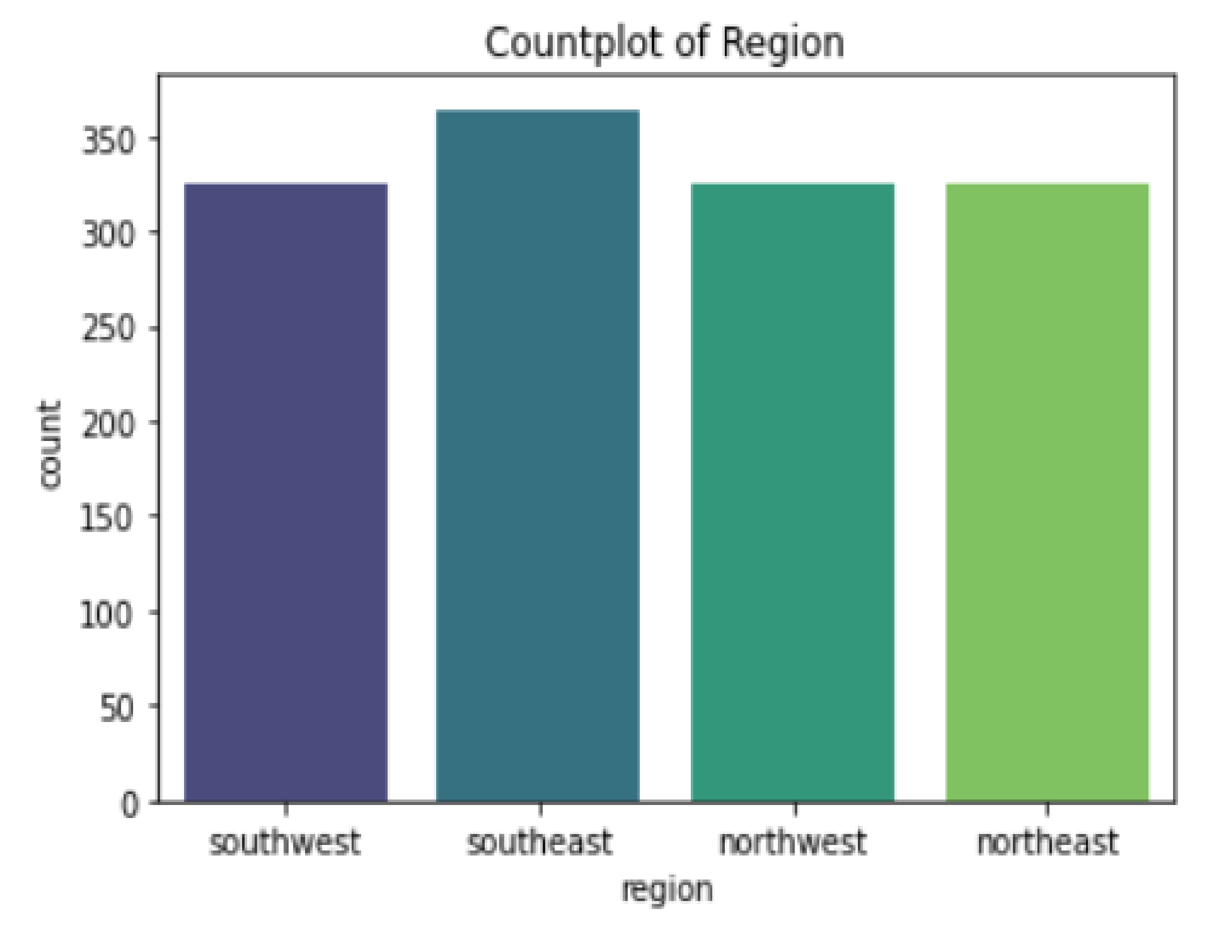
The figure on the right, shows the counts of male and female respondents in the dataset. The two classes are well balanced.

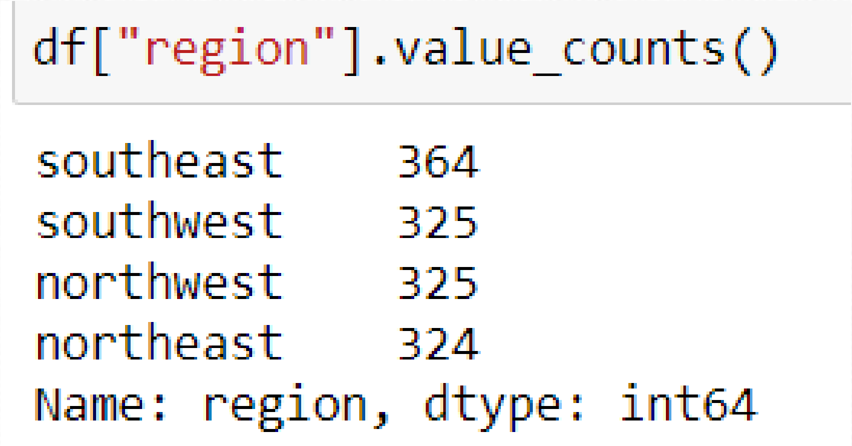
The figure on the right shows the count plot (Plot of counts) of smokers . There is a big imbalance among smokers and non-smokers in our dataset.





The figure on the left shows the distribution of ‘BMI’. By looking at the plot it seems like ‘BMI’ is normally distributed with mean somewhere around 32.





```
df["region"].value_counts()
```

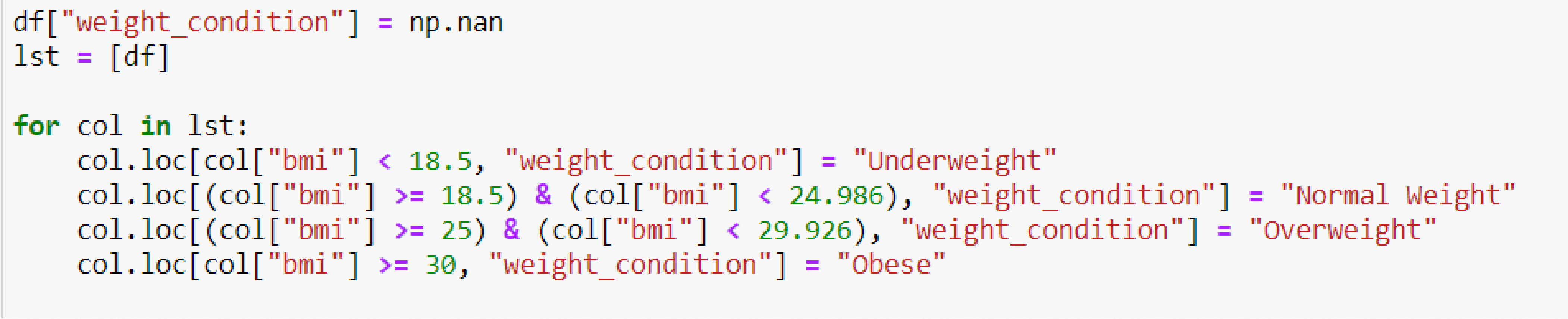
```
southeast    364
southwest    325
northwest    325
northeast    324
Name: region, dtype: int64
```

The counts for each class in the region column is almost the same, except for region southeast.

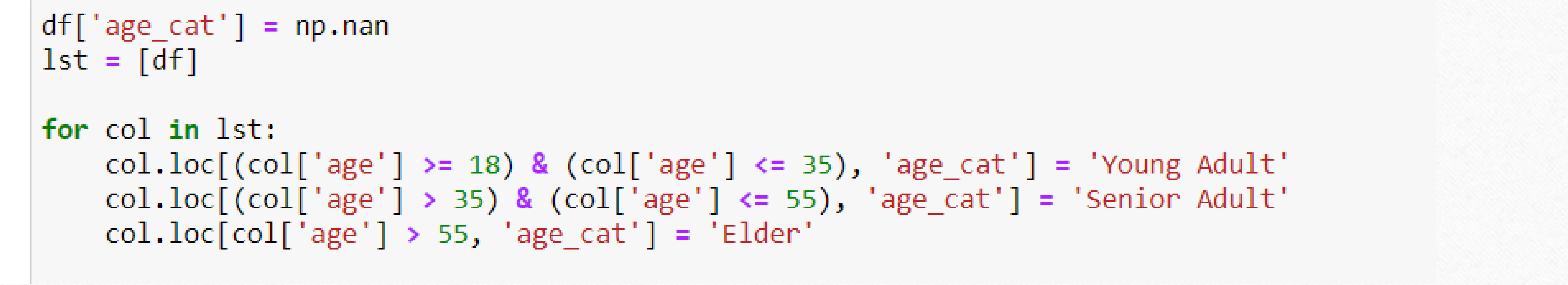
Creating additional variables from predefined variables:

Before moving onto the multivariate analysis, we will create some additional variables which will further help us in enhanced analysis.

Creating 2 new categorical features like ‘weight_condition’ and ‘age_cat’ from existing features like ‘age’ and ‘bmi’ with their respective code in python.

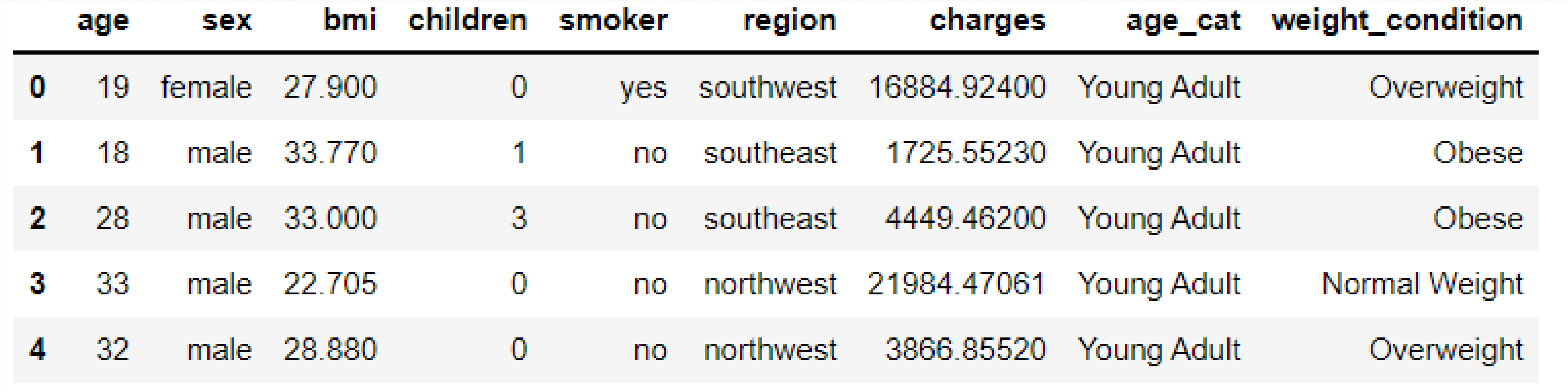


```
df["weight_condition"] = np.nan
lst = [df]

for col in lst:
    col.loc[col["bmi"] < 18.5, "weight_condition"] = "Underweight"
    col.loc[(col["bmi"] >= 18.5) & (col["bmi"] < 24.986), "weight_condition"] = "Normal Weight"
    col.loc[(col["bmi"] >= 25) & (col["bmi"] < 29.926), "weight_condition"] = "Overweight"
    col.loc[col["bmi"] >= 30, "weight_condition"] = "Obese"
```



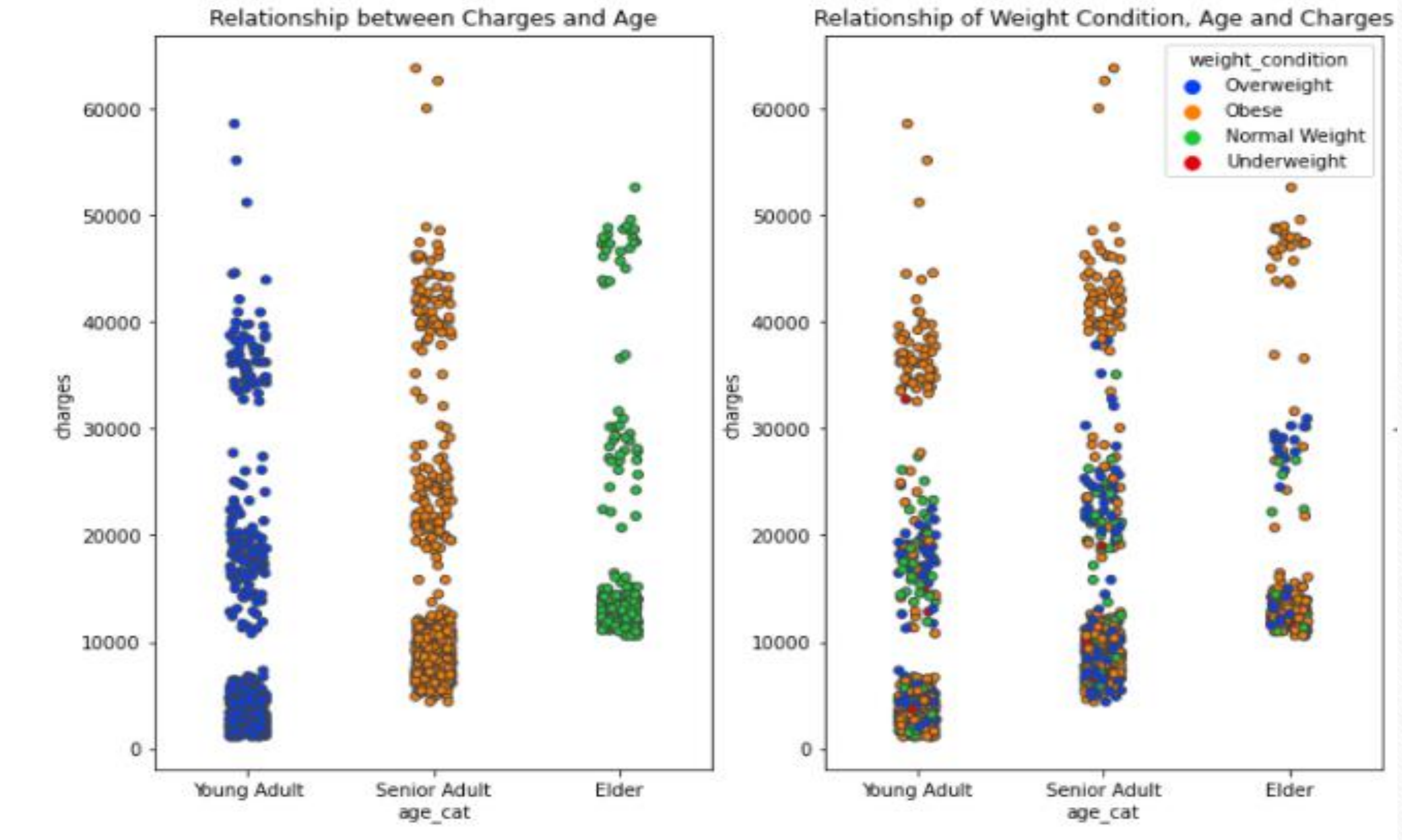
```
df['age_cat'] = np.nan
lst = [df]

for col in lst:
    col.loc[(col['age'] >= 18) & (col['age'] <= 35), 'age_cat'] = 'Young Adult'
    col.loc[(col['age'] > 35) & (col['age'] <= 55), 'age_cat'] = 'Senior Adult'
    col.loc[col['age'] > 55, 'age_cat'] = 'Elder'
```

Visualizing the dataset with the new added variables:

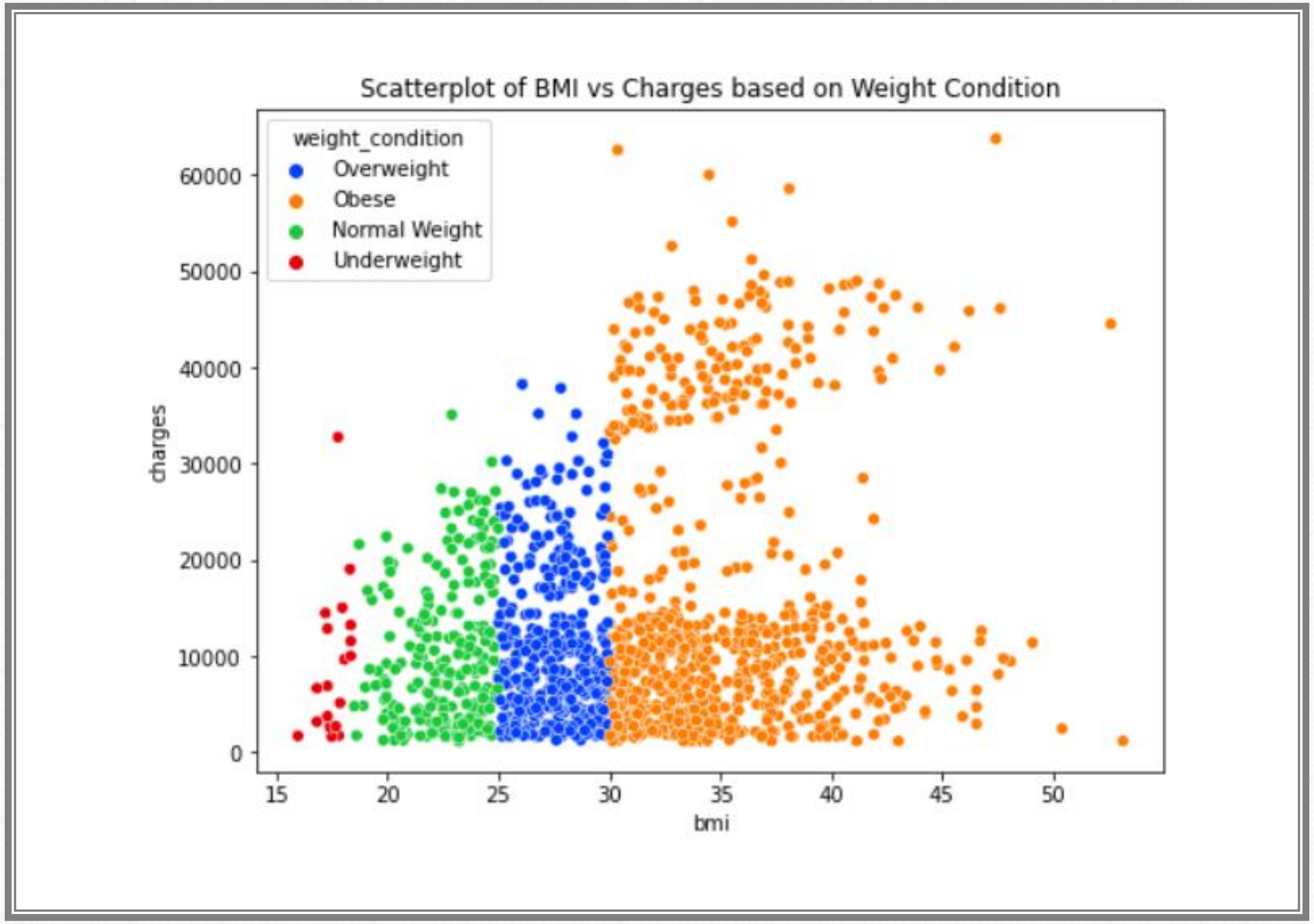


	age	sex	bmi	children	smoker	region	charges	age_cat	weight_condition
0	19	female	27.900	0	yes	southwest	16884.92400	Young Adult	Overweight
1	18	male	33.770	1	no	southeast	1725.55230	Young Adult	Obese
2	28	male	33.000	3	no	southeast	4449.46200	Young Adult	Obese
3	33	male	22.705	0	no	northwest	21984.47061	Young Adult	Normal Weight
4	32	male	28.880	0	no	northwest	3866.85520	Young Adult	Overweight

Multivariate Analysis:

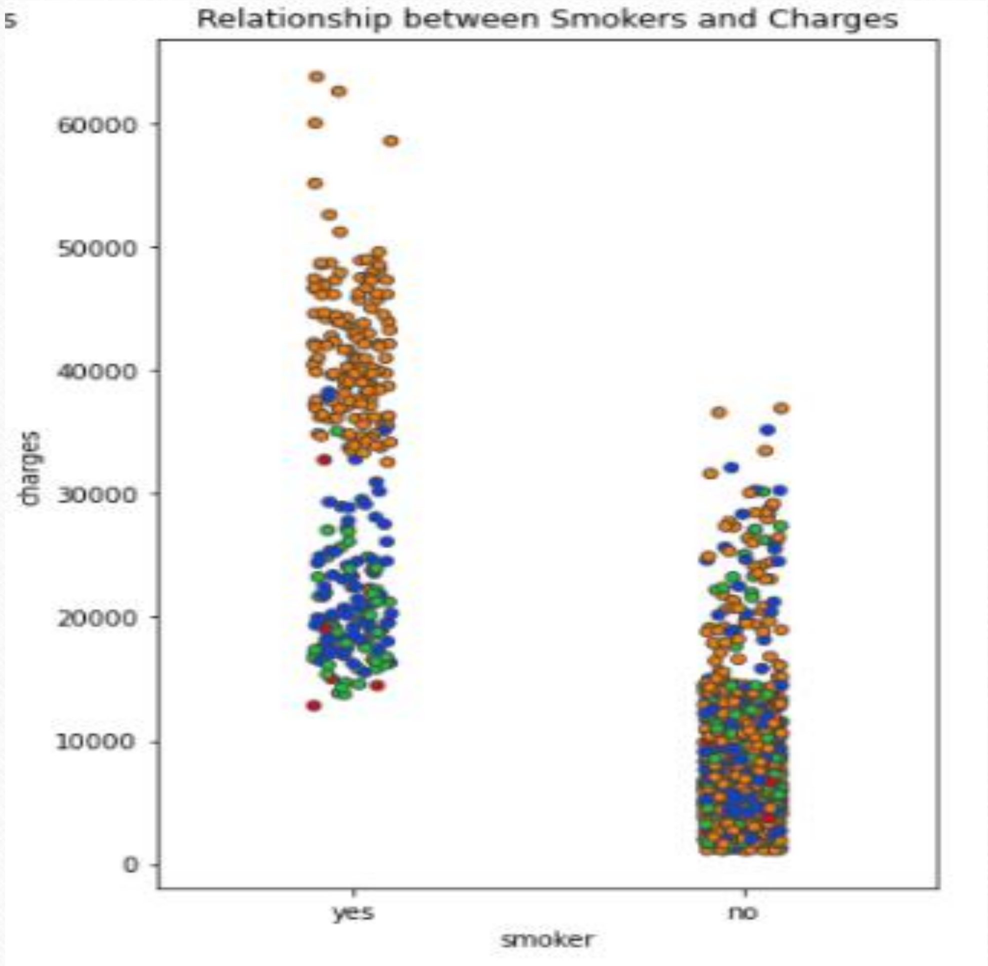


- From fig 1. we can observe that elder people have higher cost. Similarly in fig 2. it is clear that among all the age categories, obese people are paying much higher insurance costs then other weight conditions.



Relationship between ‘Weight Conditions’ and ‘Charges’.

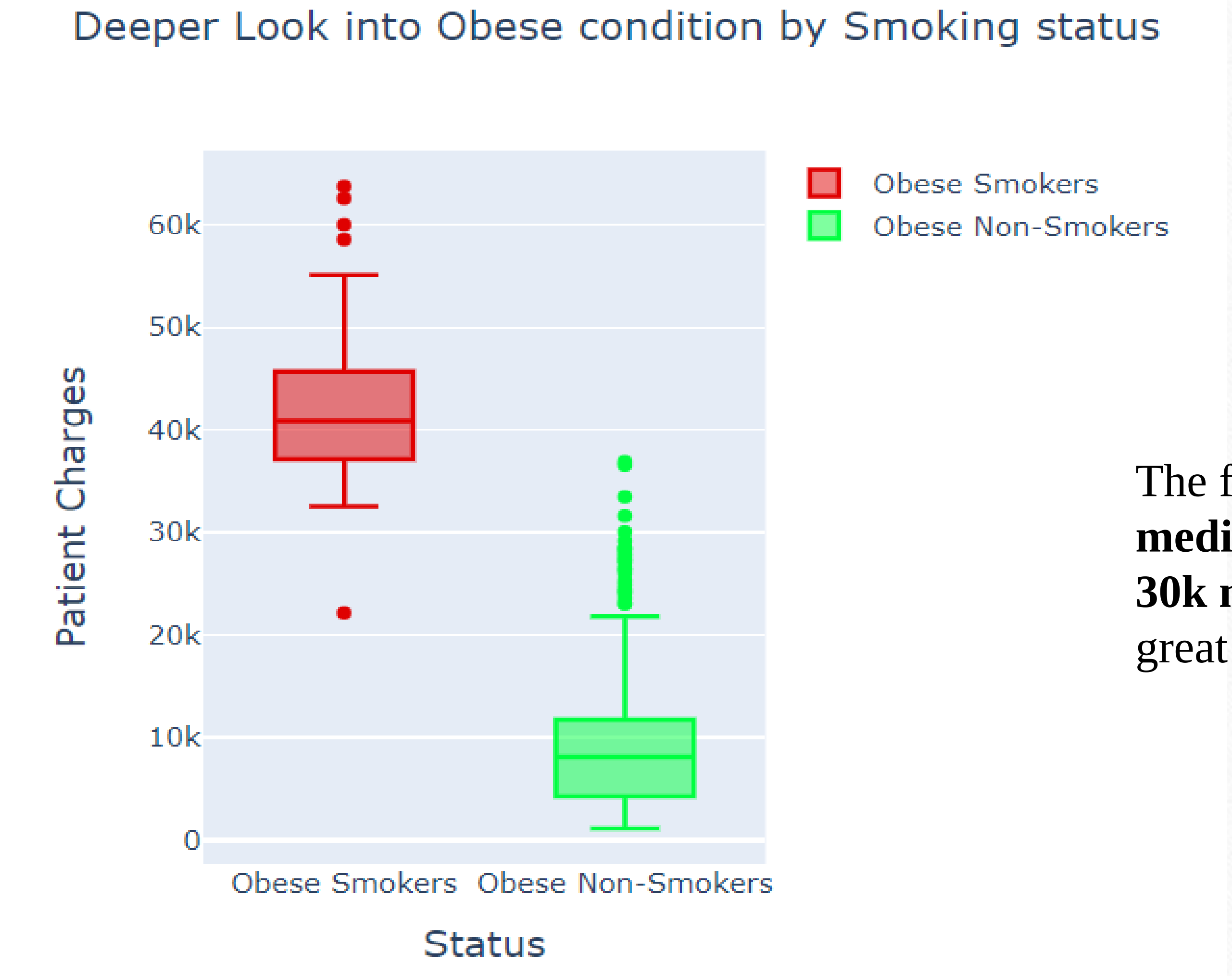
- The above plot confirms that, obese people pay much higher insurance cost then all the other categories.



Relationship between ‘Smokers’ and ‘charges’ paid.

- In the above figure, we can observe that smoker’s in general are paying much higher insurance costs and non-smokers.

For further analysis, let’s check the charges paid by an ‘obese smoker’ vs ‘obese non-smoker’.

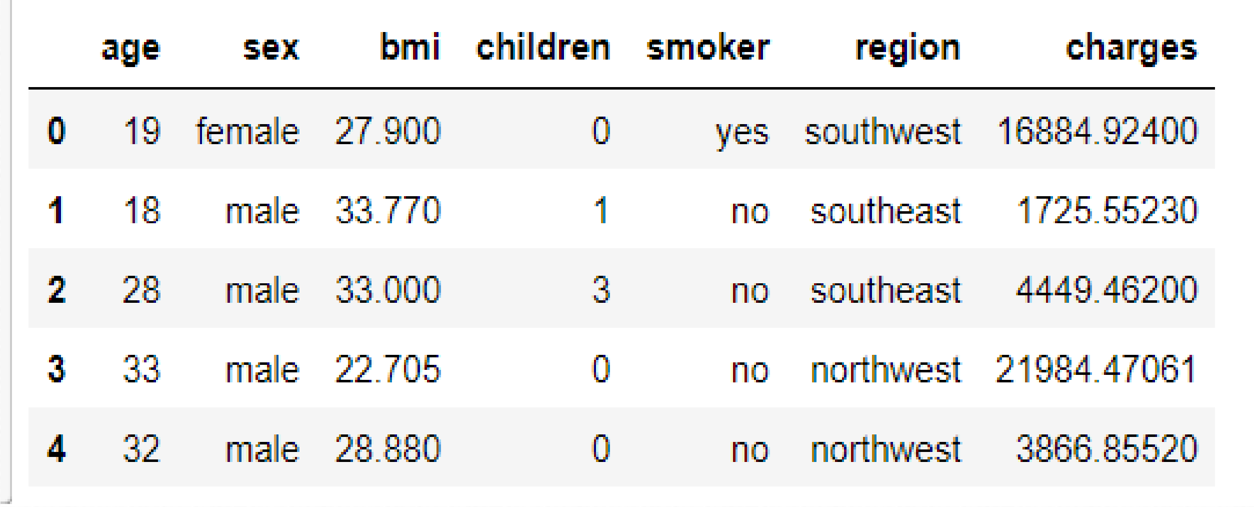


The figure on the left suggests that, on an average the **median** insurance **cost paid** by an **obese smoker** is **30k more** than that of an **obese non-smoker**. That’s great for an analysis!!

Feature Engineering:

Data Transformation:-

The table on the right shows the sample of our original dataset. We can figure out that the ‘sex’ and the “smoker” columns consists of **binary nominal object** type values. So, in this step we will convert these nominal values into **0s** and **1s**.

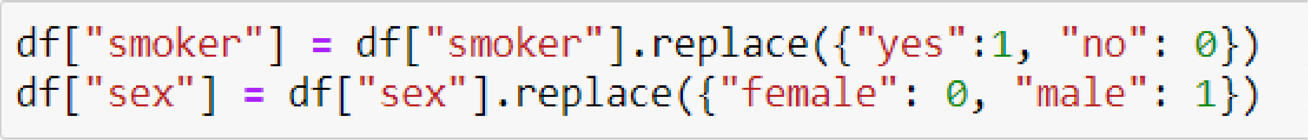


	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884.92400
1	18	male	33.770	1	no	southeast	1725.55230
2	28	male	33.000	3	no	southeast	4449.46200
3	33	male	22.705	0	no	northwest	21984.47061
4	32	male	28.880	0	no	northwest	3866.85520

But why are we converting the ‘text’ or ‘object’ type values present in ‘sex’ and ‘smoker’ column into 0s and 1s?

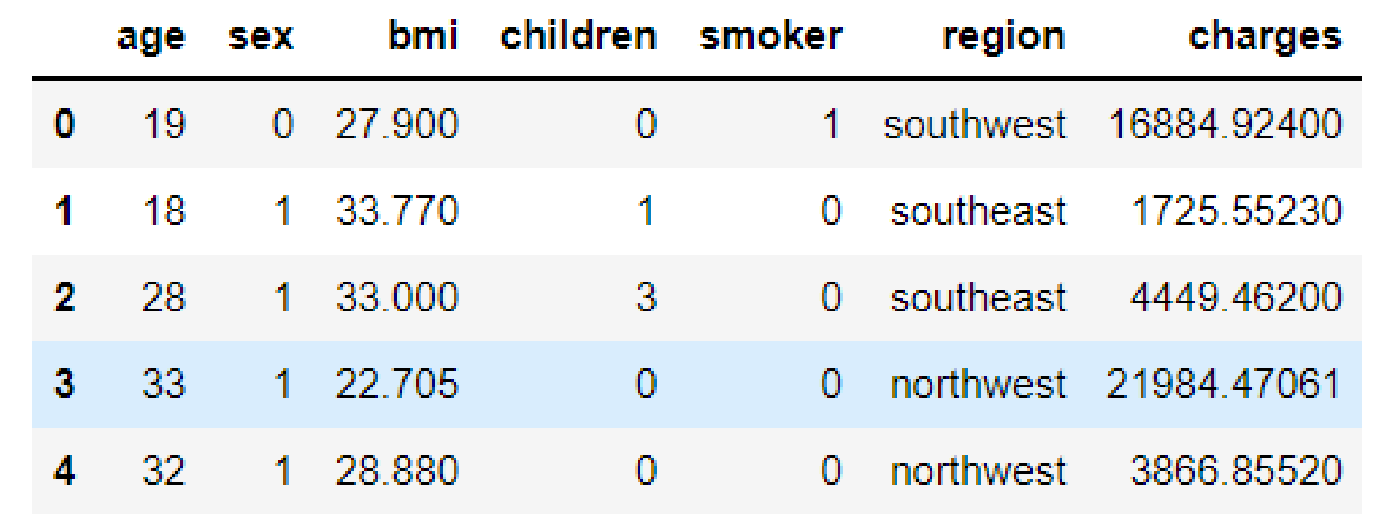
The simple answer is : Our regression model cannot interpret text data. And since the data values present in this columns are of nominal type we will replace them with 0s and 1s.

Python code for transforming ‘sex’ and ‘smoker’ columns.



```
df["smoker"] = df["smoker"].replace({"yes":1, "no": 0})
df["sex"] = df["sex"].replace({"female": 0, "male": 1})
```

Let’s take a look at the dataset after the transformation.

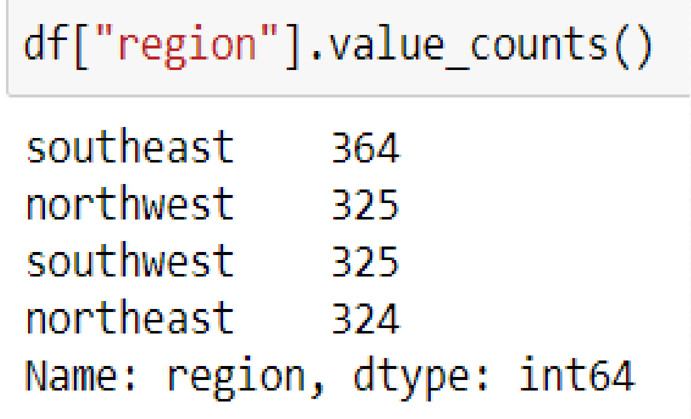


	age	sex	bmi	children	smoker	region	charges
0	19	0	27.900	0	1	southwest	16884.92400
1	18	1	33.770	1	0	southeast	1725.55230
2	28	1	33.000	3	0	southeast	4449.46200
3	33	1	22.705	0	0	northwest	21984.47061
4	32	1	28.880	0	0	northwest	3866.85520

One-hot Encoding the ‘region’ column:

We will now perform **One-hot Encoding** on the ‘**region**’ column.
So, what is **One-hot Encoding**?

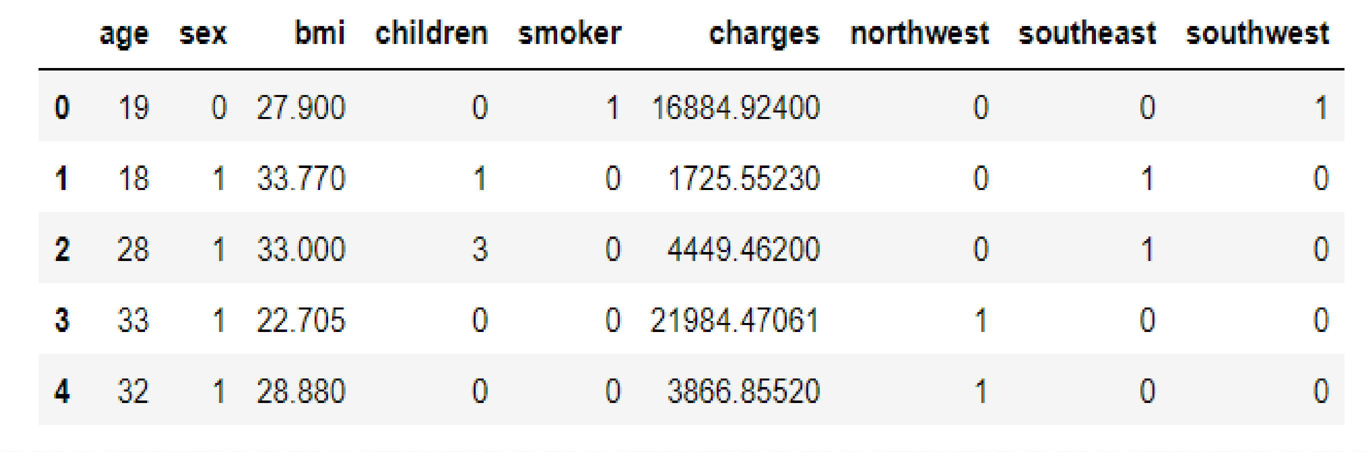
One hot encoding is a process by which **categorical** variables are **converted** into **numeric** form.



```
df["region"].value_counts()
```

```
southeast    364
northwest    325
southwest    325
northeast    324
Name: region, dtype: int64
```

When we apply one-hot encoding to the ‘region’ column, we will get additional columns known as dummy variables of 0s and 1s for each distinct value in the ‘region’ column.

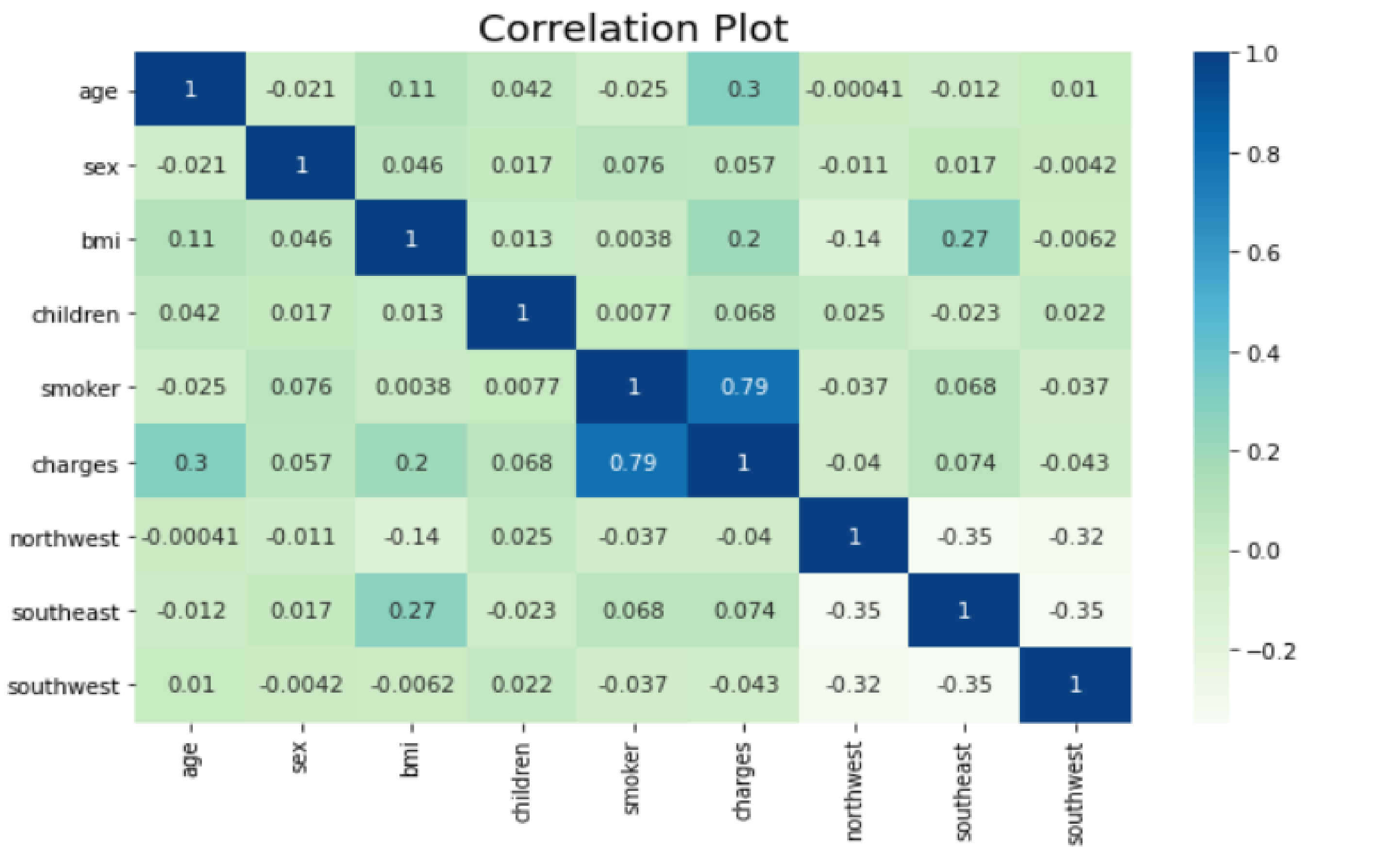


	age	sex	bmi	children	smoker	charges	northwest	southeast	southwest
0	19	0	27.900	0	1	16884.92400	0	0	1
1	18	1	33.770	1	0	1725.55230	0	1	0
2	28	1	33.000	3	0	4449.46200	0	1	0
3	33	1	22.705	0	0	21984.47061	1	0	0
4	32	1	28.880	0	0	3866.85520	1	0	0

Notice that we miss a column for ‘northeast’ in our dataset. So, ‘northeast’ column was dropped from our analysis to avoid the famous concept of dummy variable trap.

To know more about dummy variable trap refer this link:- https://towardsdatascience.com/one-hot-encoding-multicollinearity-and-the-dummy-variable-trap-b5840be3c41a

Correlation Plot of Variables:



- As we can see that among all the numeric variables, ‘smoker’ variable has the most correlation with our dependent variable ‘charges’.

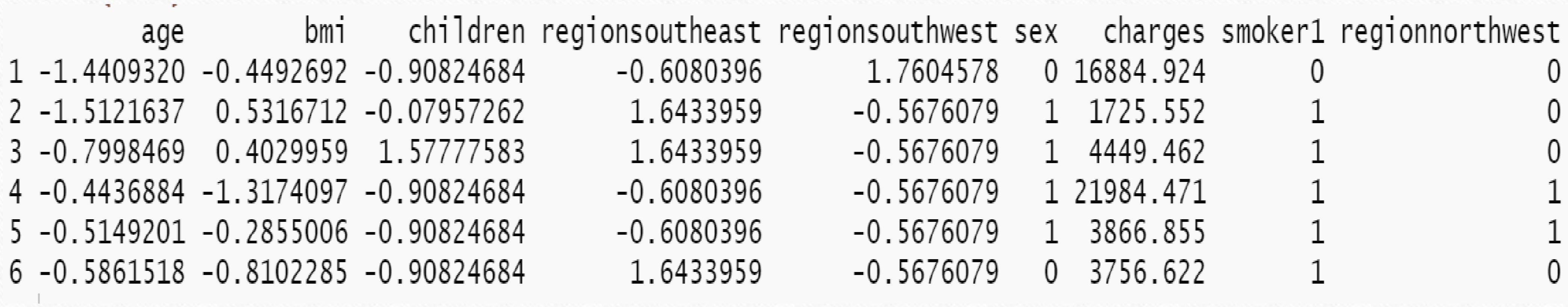
Feature Scaling:

Feature scaling is a very important step in which all the features (columns) in the dataset are brought under a common scale. The methods for scaling down the features include Standardization and Normalization.

Here, we will be perform standardization to scale our features.

$$x_{scaled} = \frac{x - mean}{sd}$$

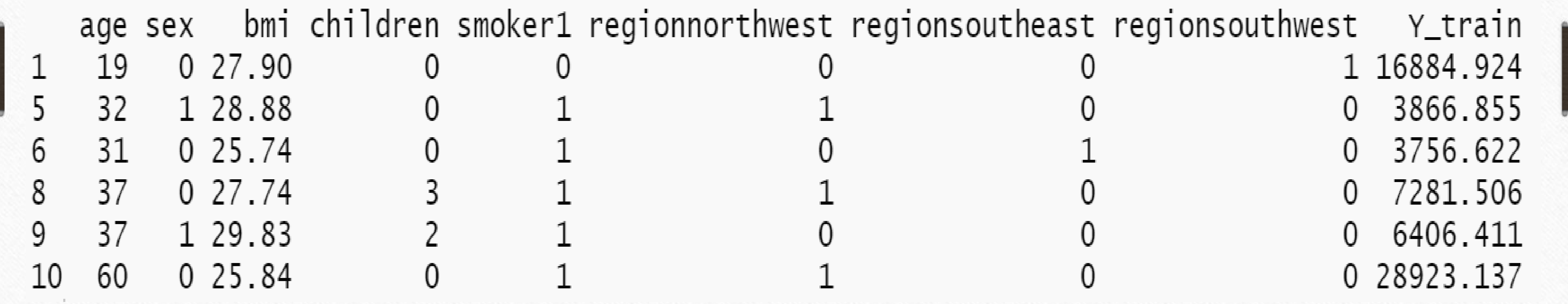
The data after standardization is as follows:



```
         age        bmi    children regionsoutheast regionsouthwest sex   charges smoker1 regionnorthwest
1 -1.4409320 -0.4492692 -0.90824684      -0.6080396       1.7604578   0 16884.924       0               0
2 -1.5121637  0.5316712 -0.07957262       1.6433959      -0.5676079   1  1725.552       1               0
3 -0.7998469  0.4029959  1.57777583       1.6433959      -0.5676079   1  4449.462       1               0
4 -0.4436884 -1.3174097 -0.90824684      -0.6080396      -0.5676079   1 21984.471       1               1
5 -0.5149201 -0.2855006 -0.90824684      -0.6080396      -0.5676079   1  3866.855       1               1
6 -0.5861518 -0.8102285 -0.90824684       1.6433959      -0.5676079   0  3756.622       1               0
```

Splitting the data into test and train datasets :

We have randomly assigned 70% of total data for fitting the model and 30% data for testing the model performance. First few rows of the train dataset after split are as follows: (Note here Y_train = Charges)



```
   age sex   bmi children smoker1 regionnorthwest regionsoutheast regionsouthwest   Y_train
1   19   0 27.90        0       0               0               0               1 16884.924
5   32   1 28.88        0       1               1               0               0  3866.855
6   31   0 25.74        0       1               0               1               0  3756.622
8   37   0 27.74        3       1               1               0               0  7281.506
9   37   1 29.83        2       1               0               0               0  6406.411
10  60   0 25.84        0       1               1               0               0 28923.137
```

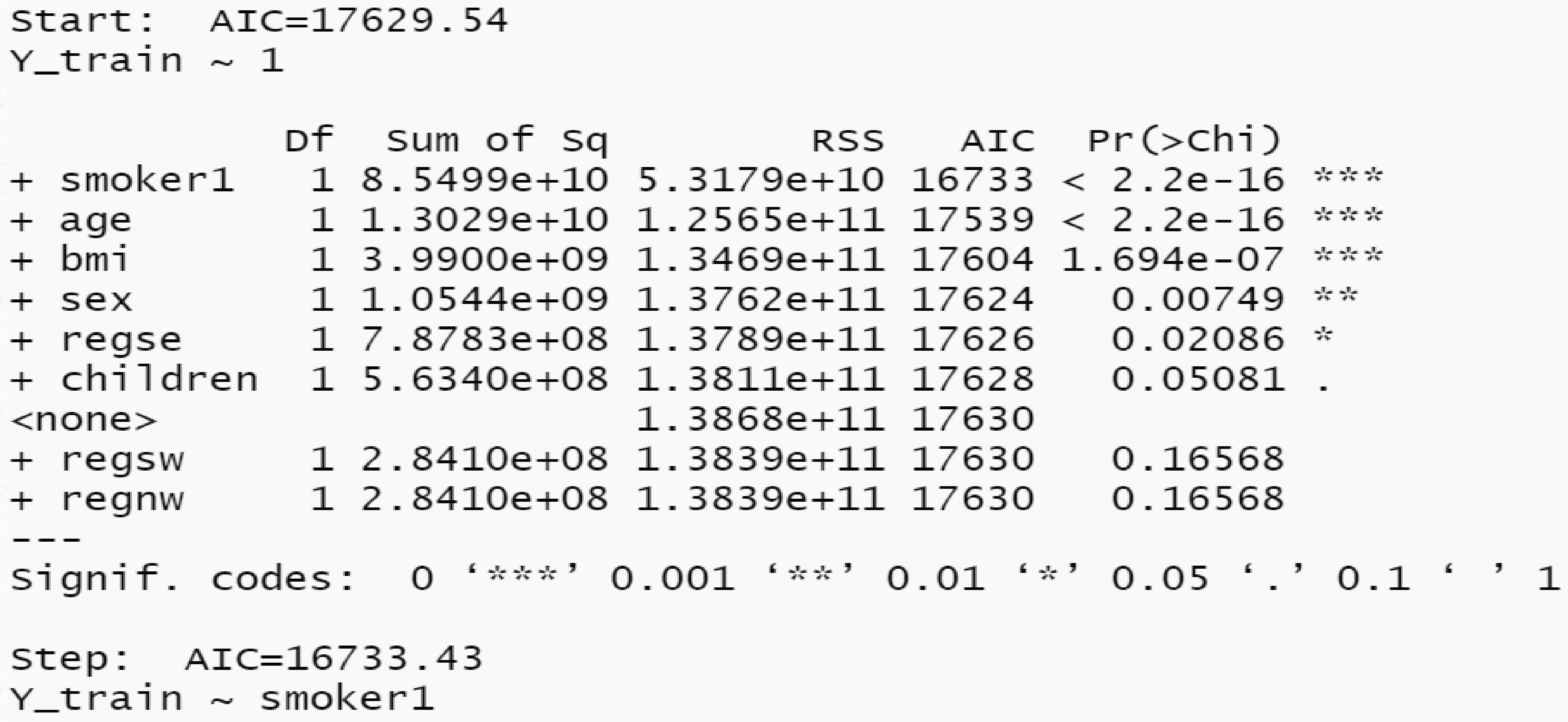
Fitting the model (Feature selection) :

We are now starting with the Forward selection method. This is done using step() function in R. The function uses AIC (Akaike's Information Criterion) to find the best model. The lower the AIC, the better the model.

Here we input the null model and the algorithm gives us the best model.

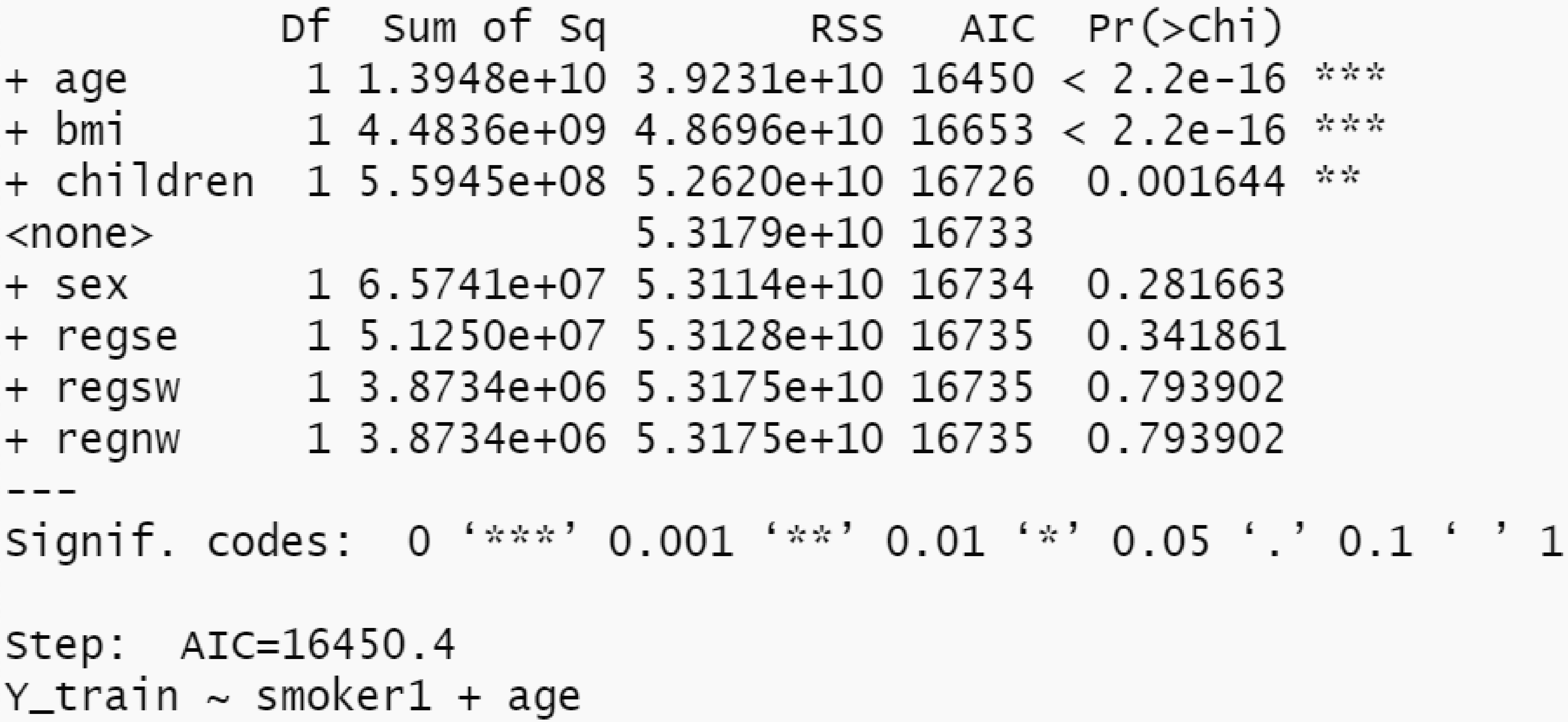
AIC = DEVIANCE + 2 x NO. OF PARAMETERS

Iteration 1 :



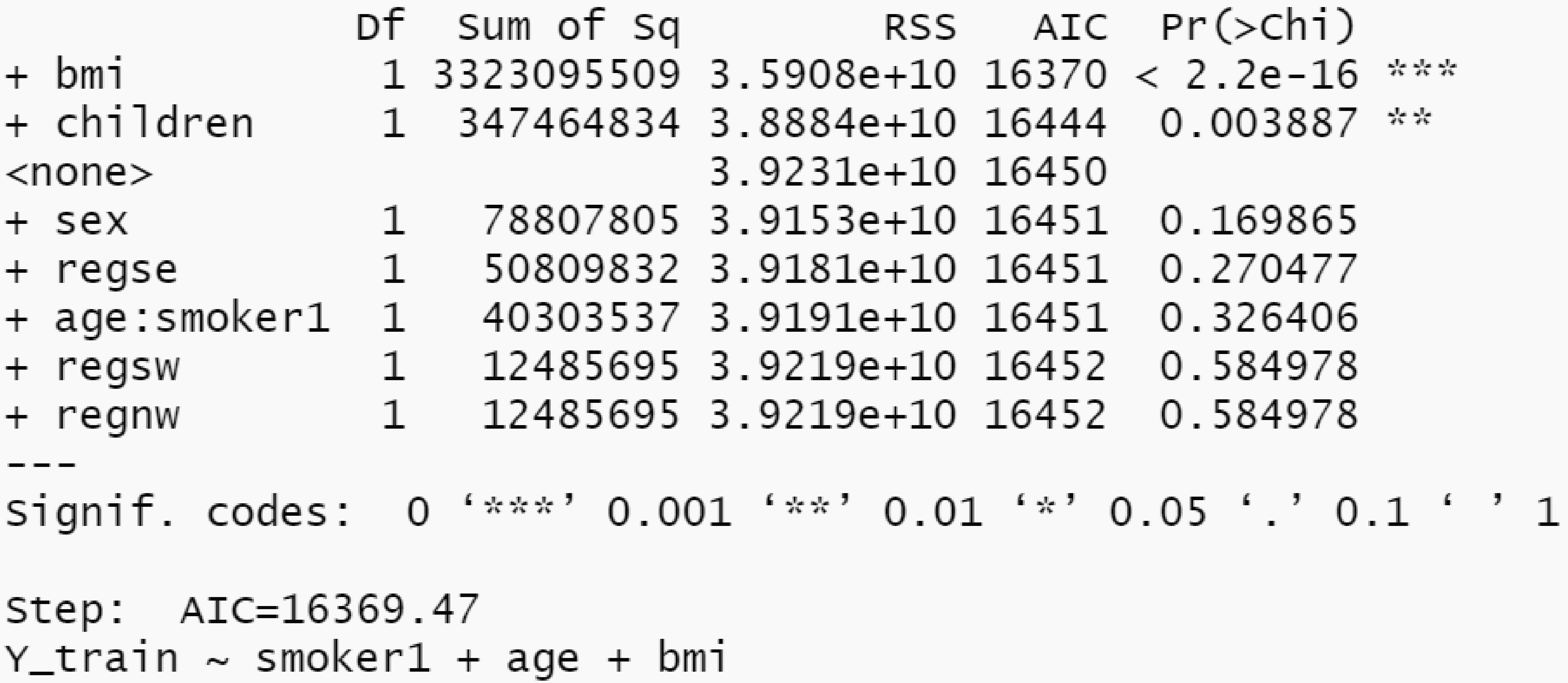
```
Start:  AIC=17629.54
Y_train ~ 1

           Df  Sum of Sq        RSS   AIC  Pr(>Chi)    
+ smoker1   1 8.5499e+10 5.3179e+10 16733 < 2.2e-16 ***
+ age       1 1.3029e+10 1.2565e+11 17539 < 2.2e-16 ***
+ bmi       1 3.9900e+09 1.3469e+11 17604 1.694e-07 ***
+ sex       1 1.0544e+09 1.3762e+11 17624   0.00749 ** 
+ regse     1 7.8783e+08 1.3789e+11 17626   0.02086 *  
+ children  1 5.6340e+08 1.3811e+11 17628   0.05081 .  
<none>                   1.3868e+11 17630              
+ regsw     1 2.8410e+08 1.3839e+11 17630   0.16568    
+ regnw     1 2.8410e+08 1.3839e+11 17630   0.16568    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Step:  AIC=16733.43
Y_train ~ smoker1
```

Iteration 2 :



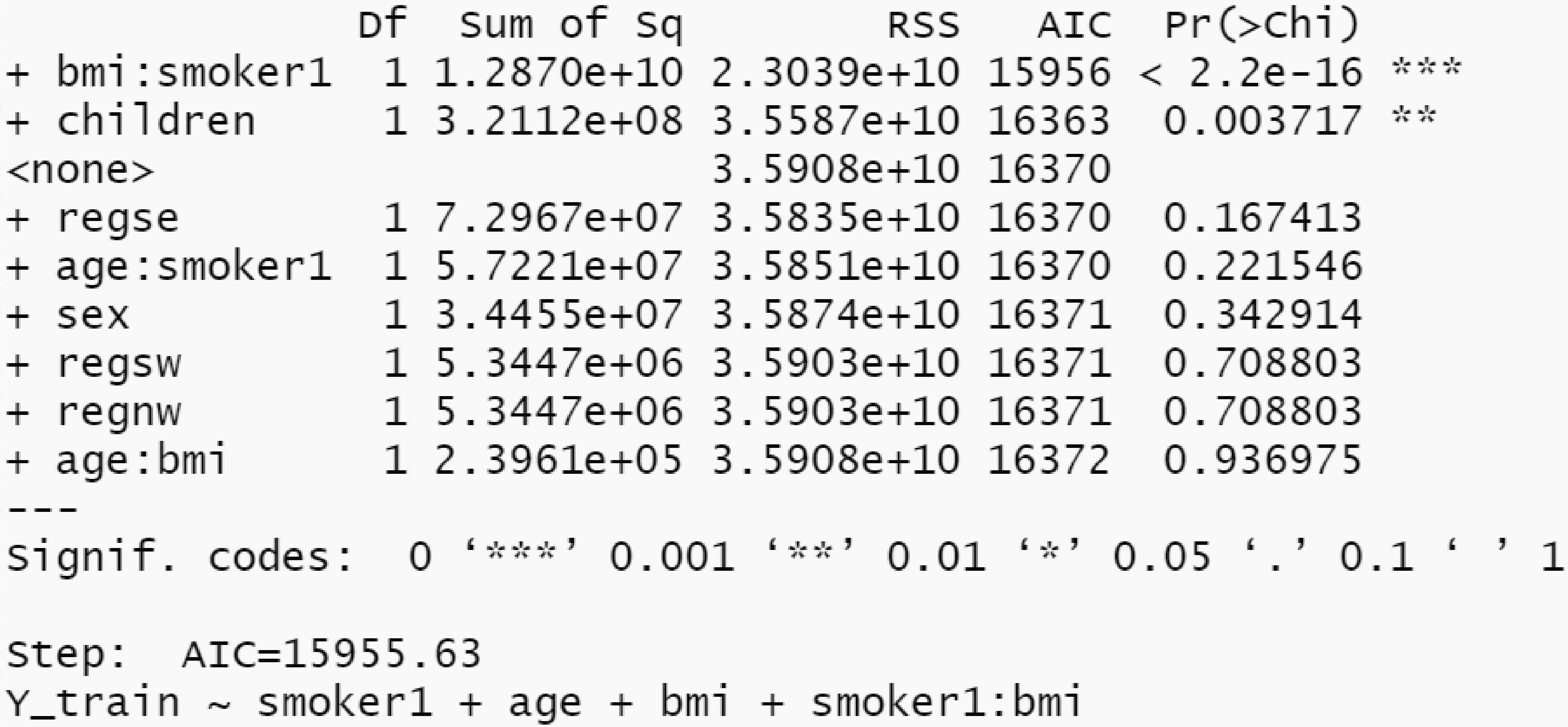
```
           Df  Sum of Sq        RSS   AIC  Pr(>Chi)
+ age       1 1.3948e+10 3.9231e+10 16450 < 2.2e-16 ***
+ bmi       1 4.4836e+09 4.8696e+10 16653 < 2.2e-16 ***
+ children  1 5.5945e+08 5.2620e+10 16726  0.001644 **
<none>                   5.3179e+10 16733
+ sex       1 6.5741e+07 5.3114e+10 16734  0.281663
+ regse     1 5.1250e+07 5.3128e+10 16735  0.341861
+ regsw     1 3.8734e+06 5.3175e+10 16735  0.793902
+ regnw     1 3.8734e+06 5.3175e+10 16735  0.793902
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Step:  AIC=16450.4
Y_train ~ smoker1 + age
```

Iteration 3 :



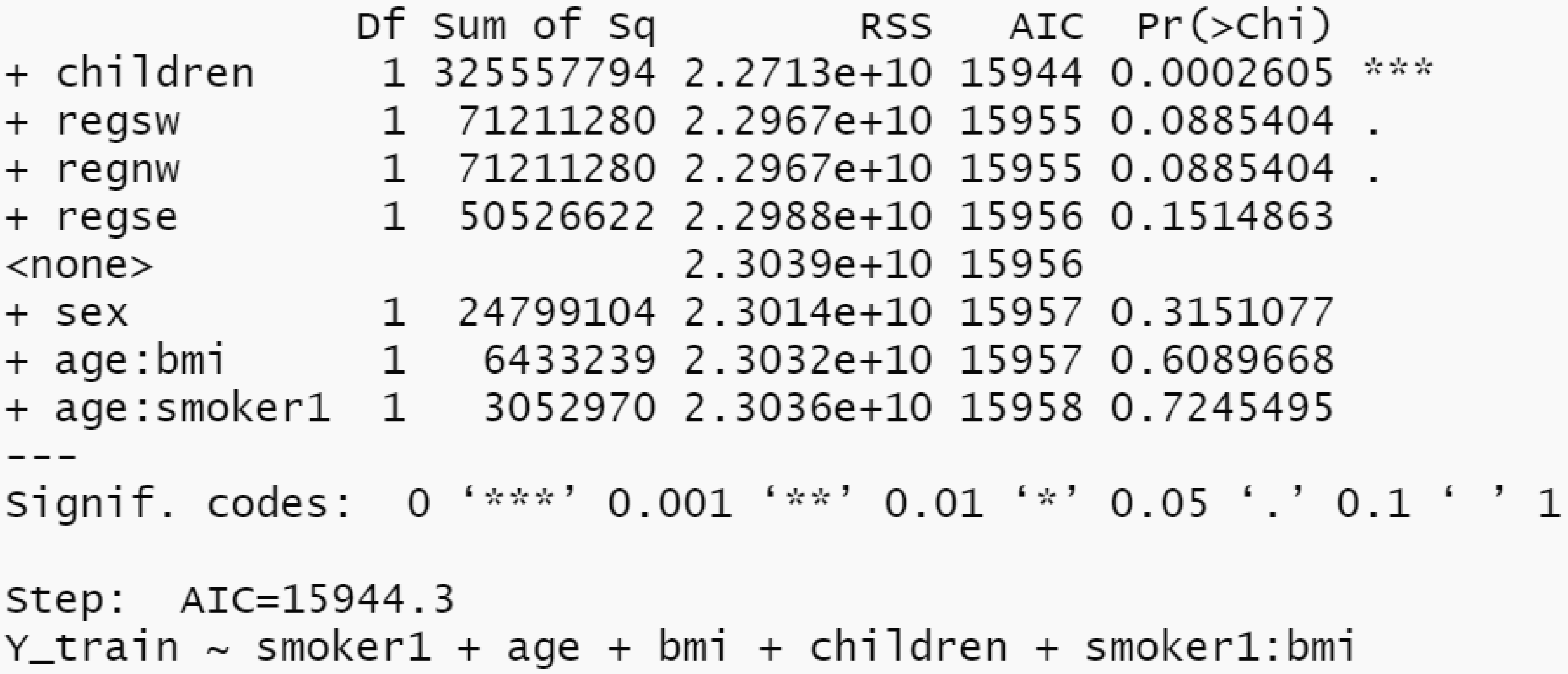
```
              Df  Sum of Sq        RSS   AIC  Pr(>Chi)
+ bmi          1 3323095509 3.5908e+10 16370 < 2.2e-16 ***
+ children     1  347464834 3.8884e+10 16444  0.003887 **
<none>                      3.9231e+10 16450
+ sex          1   78807805 3.9153e+10 16451  0.169865
+ regse        1   50809832 3.9181e+10 16451  0.270477
+ age:smoker1  1   40303537 3.9191e+10 16451  0.326406
+ regsw        1   12485695 3.9219e+10 16452  0.584978
+ regnw        1   12485695 3.9219e+10 16452  0.584978
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Step:  AIC=16369.47
Y_train ~ smoker1 + age + bmi
```

Iteration 4 :



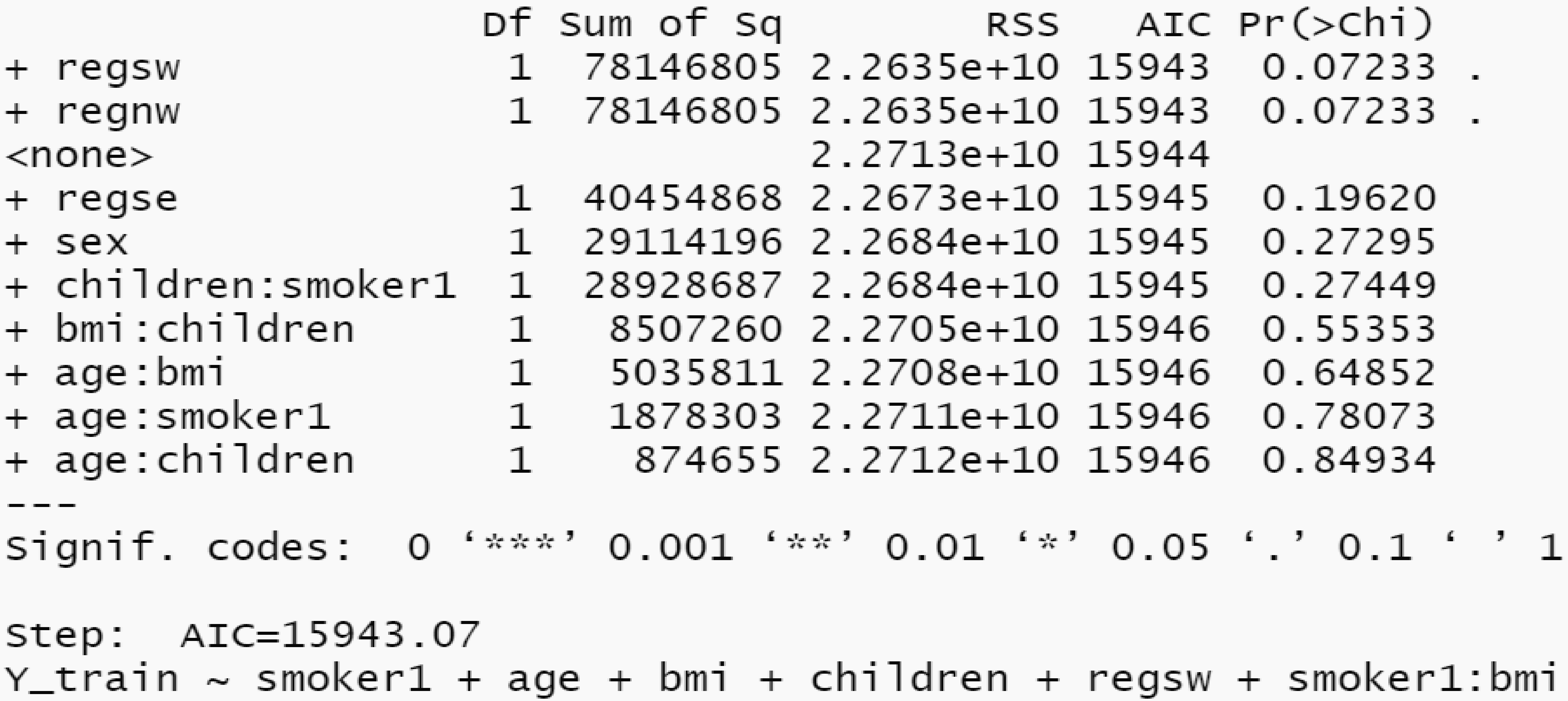
```
                  Df  Sum of Sq        RSS   AIC  Pr(>Chi)    
+ bmi:smoker1   1 1.2870e+10 2.3039e+10 15956 < 2.2e-16 ***
+ children      1 3.2112e+08 3.5587e+10 16363  0.003717 ** 
<none>                       3.5908e+10 16370              
+ regse         1 7.2967e+07 3.5835e+10 16370  0.167413    
+ age:smoker1   1 5.7221e+07 3.5851e+10 16370  0.221546    
+ sex           1 3.4455e+07 3.5874e+10 16371  0.342914    
+ regsw         1 5.3447e+06 3.5903e+10 16371  0.708803    
+ regnw         1 5.3447e+06 3.5903e+10 16371  0.708803    
+ age:bmi       1 2.3961e+05 3.5908e+10 16372  0.936975    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Step:  AIC=15955.63
Y_train ~ smoker1 + age + bmi + smoker1:bmi
```

Iteration 5 :



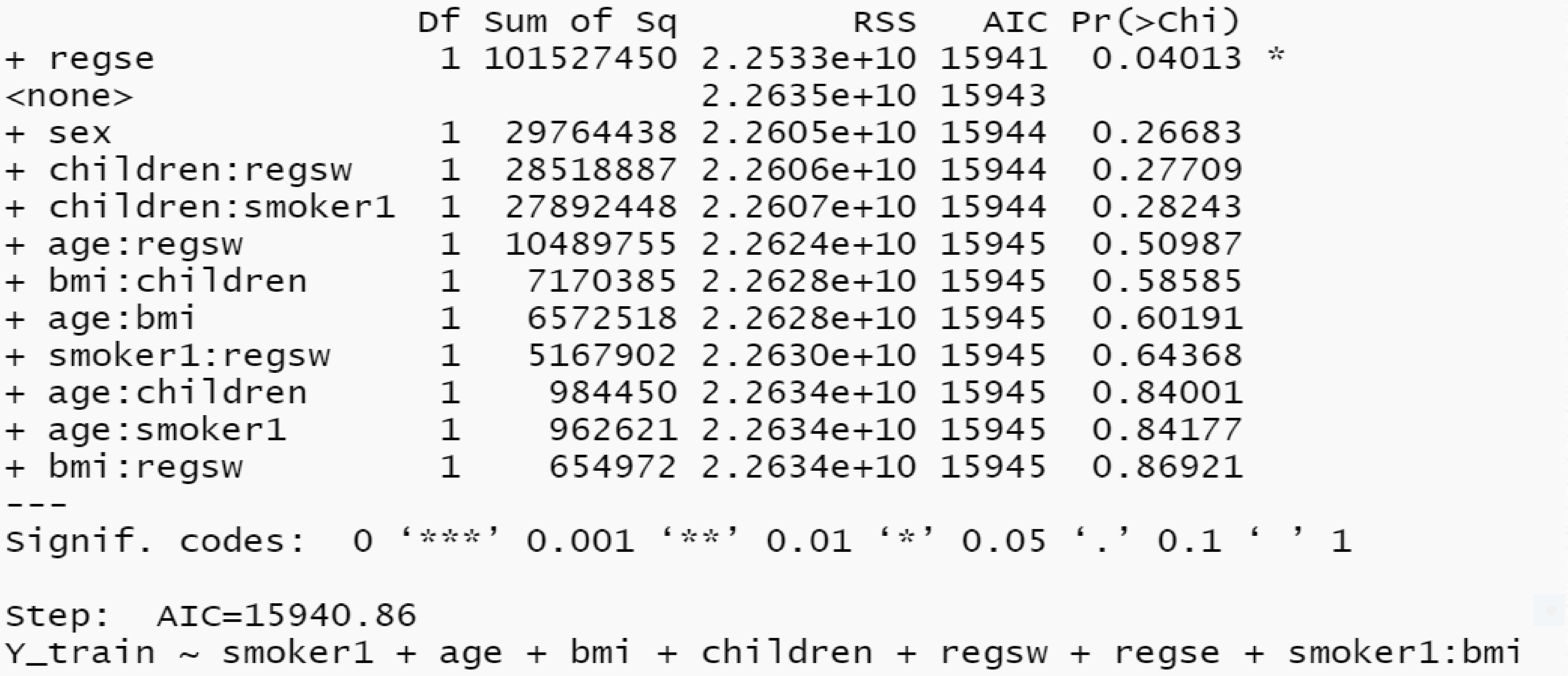
```
               Df Sum of Sq        RSS   AIC  Pr(>Chi)    
+ children      1 325557794 2.2713e+10 15944 0.0002605 ***
+ regsw         1  71211280 2.2967e+10 15955 0.0885404 .  
+ regnw         1  71211280 2.2967e+10 15955 0.0885404 .  
+ regse         1  50526622 2.2988e+10 15956 0.1514863    
<none>                      2.3039e+10 15956              
+ sex           1  24799104 2.3014e+10 15957 0.3151077    
+ age:bmi       1   6433239 2.3032e+10 15957 0.6089668    
+ age:smoker1   1   3052970 2.3036e+10 15958 0.7245495    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Step:  AIC=15944.3
Y_train ~ smoker1 + age + bmi + children + smoker1:bmi
```

Iteration 6 :



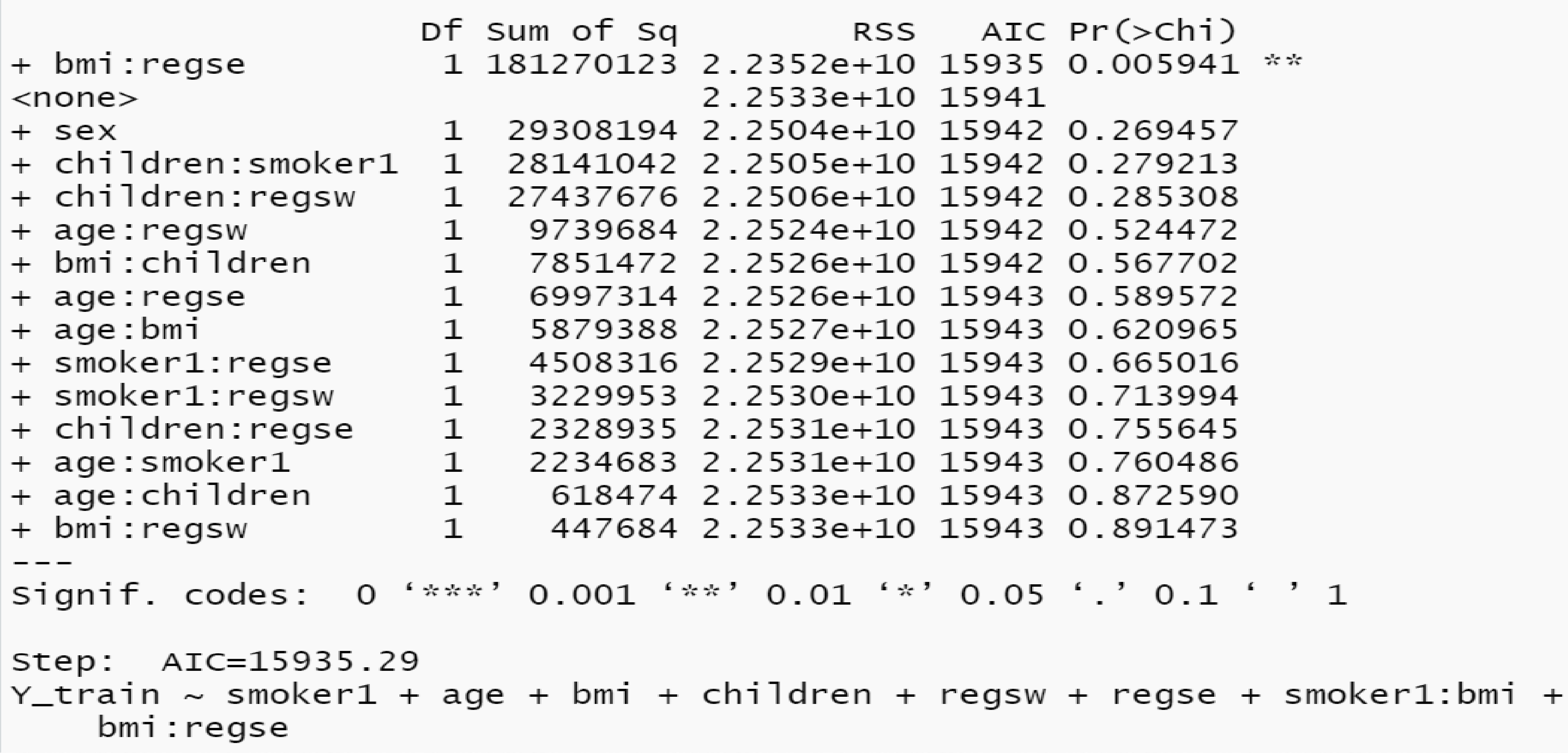
```
                   Df Sum of Sq        RSS   AIC Pr(>Chi)
+ regsw             1  78146805 2.2635e+10 15943  0.07233 .
+ regnw             1  78146805 2.2635e+10 15943  0.07233 .
<none>                          2.2713e+10 15944
+ regse             1  40454868 2.2673e+10 15945  0.19620
+ sex               1  29114196 2.2684e+10 15945  0.27295
+ children:smoker1  1  28928687 2.2684e+10 15945  0.27449
+ bmi:children      1   8507260 2.2705e+10 15946  0.55353
+ age:bmi           1   5035811 2.2708e+10 15946  0.64852
+ age:smoker1       1   1878303 2.2711e+10 15946  0.78073
+ age:children      1    874655 2.2712e+10 15946  0.84934
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Step:  AIC=15943.07
Y_train ~ smoker1 + age + bmi + children + regsw + smoker1:bmi
```

Iteration 7 :



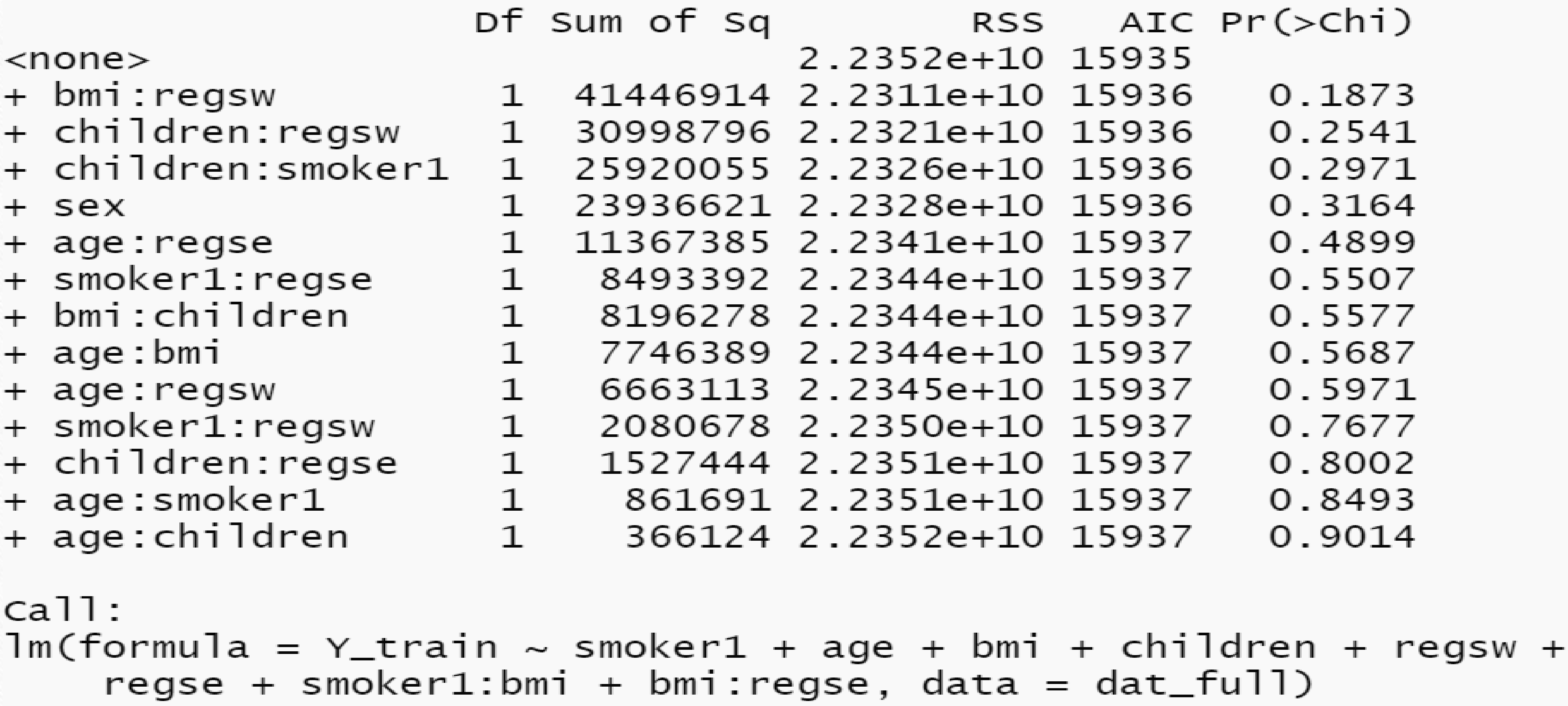
```
                  Df Sum of Sq        RSS   AIC Pr(>Chi)  
+ regse            1 101527450 2.2533e+10 15941  0.04013 *
<none>                         2.2635e+10 15943           
+ sex              1  29764438 2.2605e+10 15944  0.26683  
+ children:regsw   1  28518887 2.2606e+10 15944  0.27709  
+ children:smoker1 1  27892448 2.2607e+10 15944  0.28243  
+ age:regsw        1  10489755 2.2624e+10 15945  0.50987  
+ bmi:children     1   7170385 2.2628e+10 15945  0.58585  
+ age:bmi          1   6572518 2.2628e+10 15945  0.60191  
+ smoker1:regsw    1   5167902 2.2630e+10 15945  0.64368  
+ age:children     1    984450 2.2634e+10 15945  0.84001  
+ age:smoker1      1    962621 2.2634e+10 15945  0.84177  
+ bmi:regsw        1    654972 2.2634e+10 15945  0.86921  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Step:  AIC=15940.86
Y_train ~ smoker1 + age + bmi + children + regsw + regse + smoker1:bmi
```

Iteration 8 :



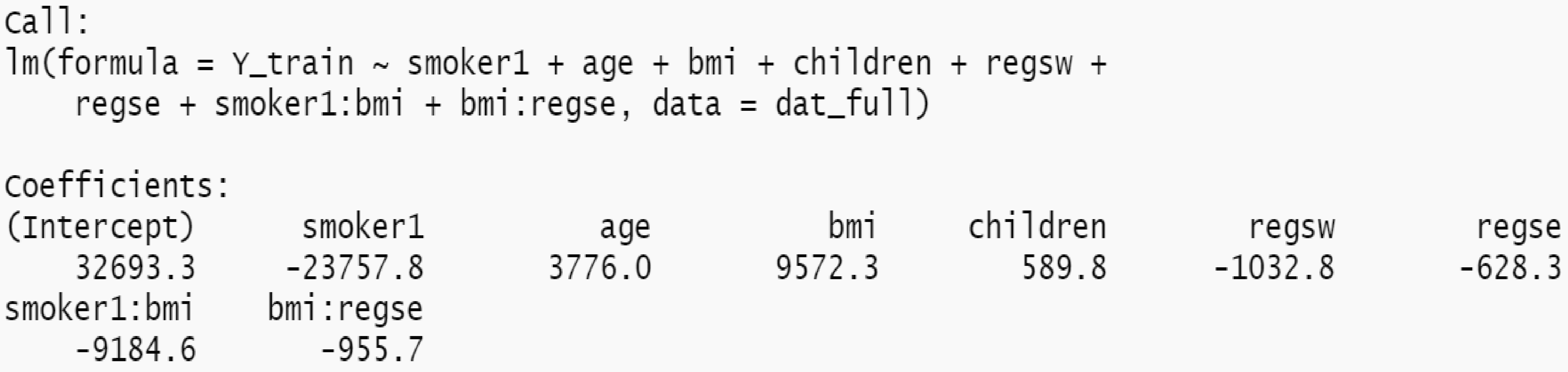
```
                   Df Sum of Sq        RSS   AIC Pr(>Chi)   
+ bmi:regse         1 181270123 2.2352e+10 15935 0.005941 **
<none>                          2.2533e+10 15941            
+ sex               1  29308194 2.2504e+10 15942 0.269457   
+ children:smoker1  1  28141042 2.2505e+10 15942 0.279213   
+ children:regsw    1  27437676 2.2506e+10 15942 0.285308   
+ age:regsw         1   9739684 2.2524e+10 15942 0.524472   
+ bmi:children      1   7851472 2.2526e+10 15942 0.567702   
+ age:regse         1   6997314 2.2526e+10 15943 0.589572   
+ age:bmi           1   5879388 2.2527e+10 15943 0.620965   
+ smoker1:regse     1   4508316 2.2529e+10 15943 0.665016   
+ smoker1:regsw     1   3229953 2.2530e+10 15943 0.713994   
+ children:regse    1   2328935 2.2531e+10 15943 0.755645   
+ age:smoker1       1   2234683 2.2531e+10 15943 0.760486   
+ age:children      1    618474 2.2533e+10 15943 0.872590   
+ bmi:regsw         1    447684 2.2533e+10 15943 0.891473   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Step:  AIC=15935.29
Y_train ~ smoker1 + age + bmi + children + regsw + regse + smoker1:bmi + 
    bmi:regse
```

Iteration 9 :



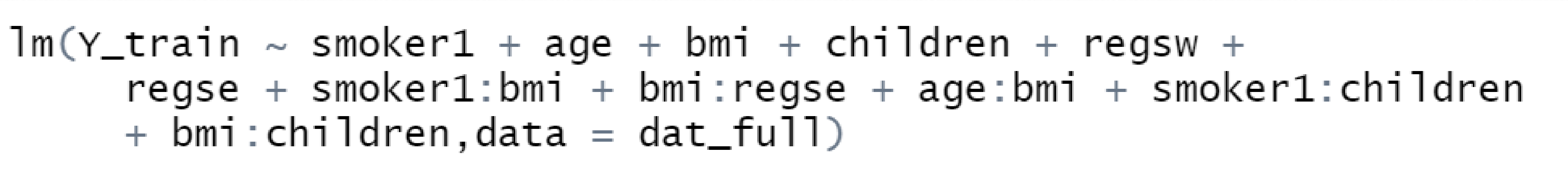
```
                  Df Sum of Sq        RSS   AIC Pr(>Chi)
<none>                         2.2352e+10 15935         
+ bmi:regsw        1  41446914 2.2311e+10 15936   0.1873
+ children:regsw   1  30998796 2.2321e+10 15936   0.2541
+ children:smoker1 1  25920055 2.2326e+10 15936   0.2971
+ sex              1  23936621 2.2328e+10 15936   0.3164
+ age:regse        1  11367385 2.2341e+10 15937   0.4899
+ smoker1:regse    1   8493392 2.2344e+10 15937   0.5507
+ bmi:children     1   8196278 2.2344e+10 15937   0.5577
+ age:bmi          1   7746389 2.2344e+10 15937   0.5687
+ age:regsw        1   6663113 2.2345e+10 15937   0.5971
+ smoker1:regsw    1   2080678 2.2350e+10 15937   0.7677
+ children:regse   1   1527444 2.2351e+10 15937   0.8002
+ age:smoker1      1    861691 2.2351e+10 15937   0.8493
+ age:children     1    366124 2.2352e+10 15937   0.9014

Call:
lm(formula = Y_train ~ smoker1 + age + bmi + children + regsw + 
    regse + smoker1:bmi + bmi:regse, data = dat_full)
```

Final model using forward selection :



```
Call:
lm(formula = Y_train ~ smoker1 + age + bmi + children + regsw + 
    regse + smoker1:bmi + bmi:regse, data = dat_full)

Coefficients:
(Intercept)      smoker1          age          bmi     children        regsw        regse  
    32693.3     -23757.8       3776.0       9572.3        589.8      -1032.8       -628.3  
smoker1:bmi    bmi:regse  
    -9184.6       -955.7  
```

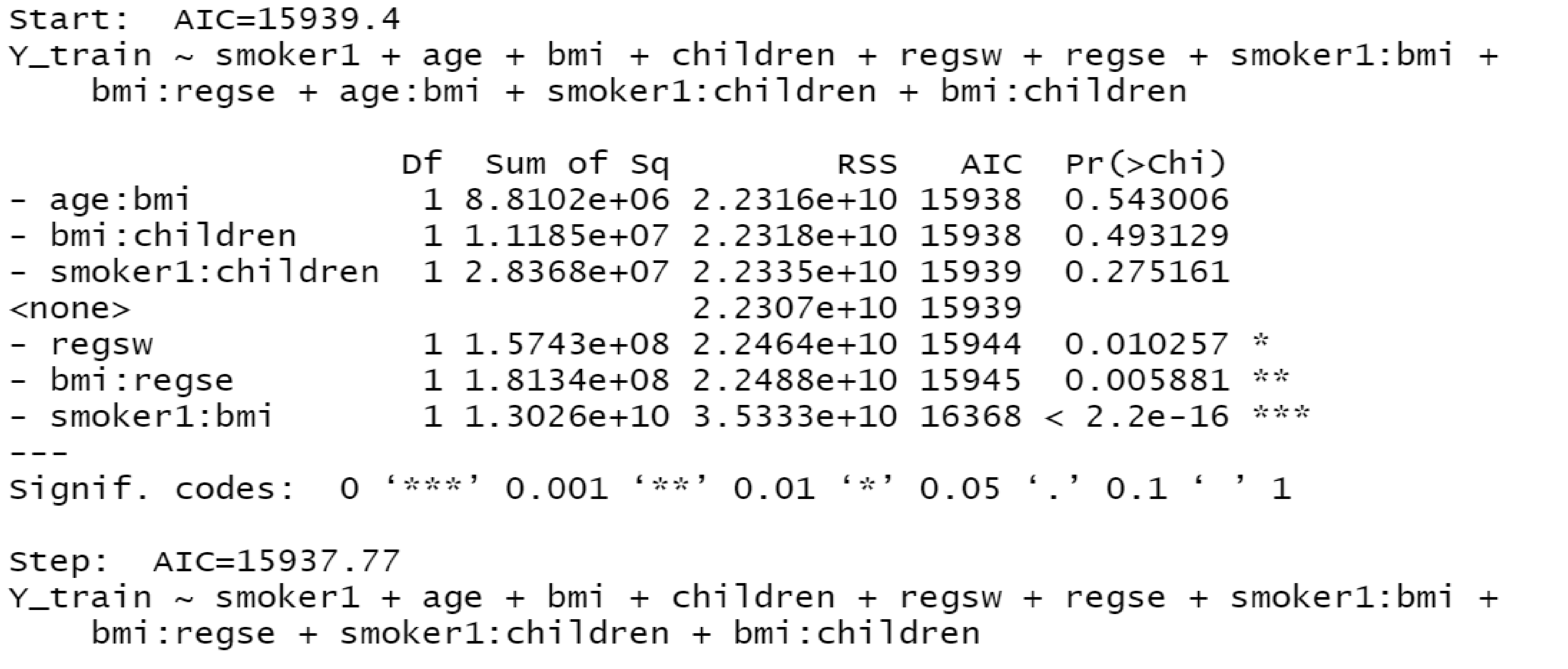
Backward elimination method in R:



```
lm(Y_train ~ smoker1 + age + bmi + children + regsw +
     regse + smoker1:bmi + bmi:regse + age:bmi + smoker1:children
     + bmi:children,data = dat_full)
```

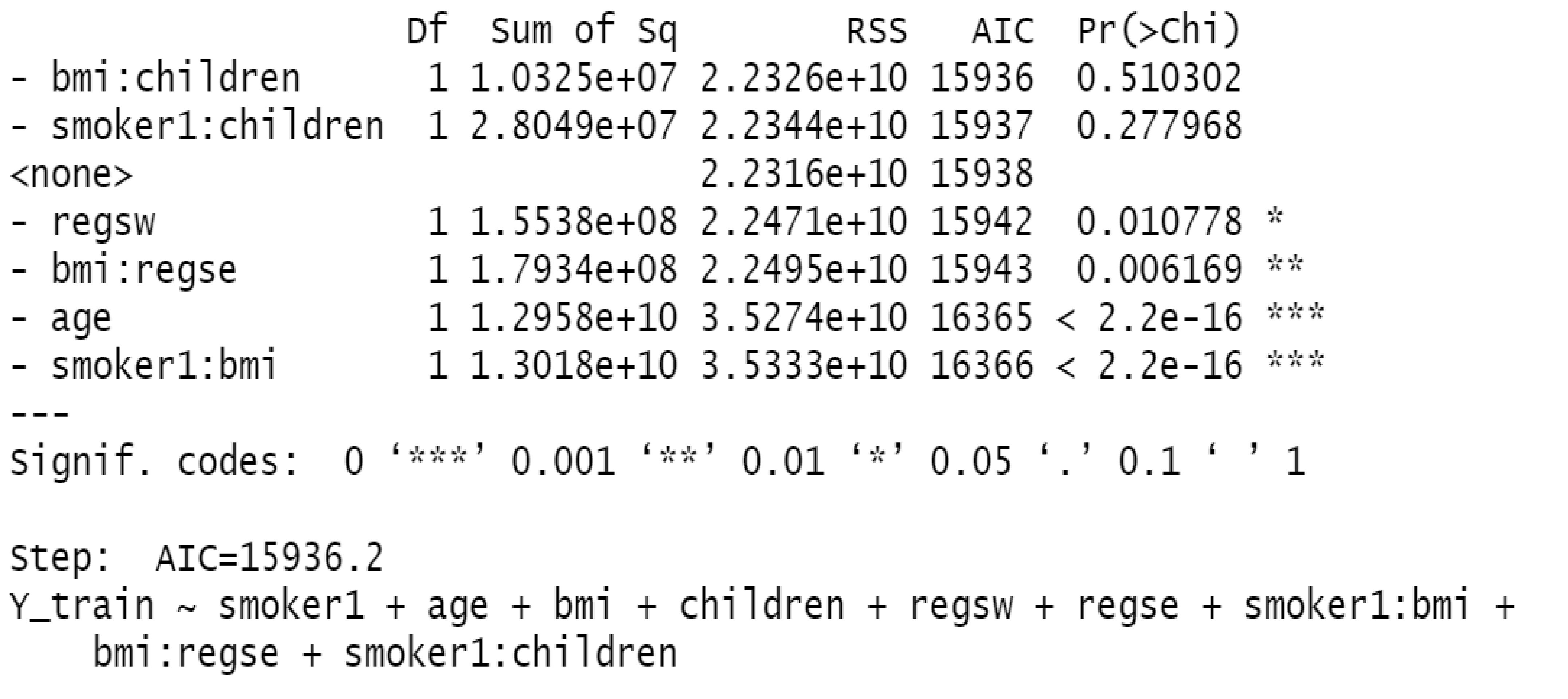
In R, to perform backward elimination, we have to feed in some model to the algorithm first. The algorithm then analyzes those terms and removes all insignificant terms, giving us the best model.

Iteration 1 :



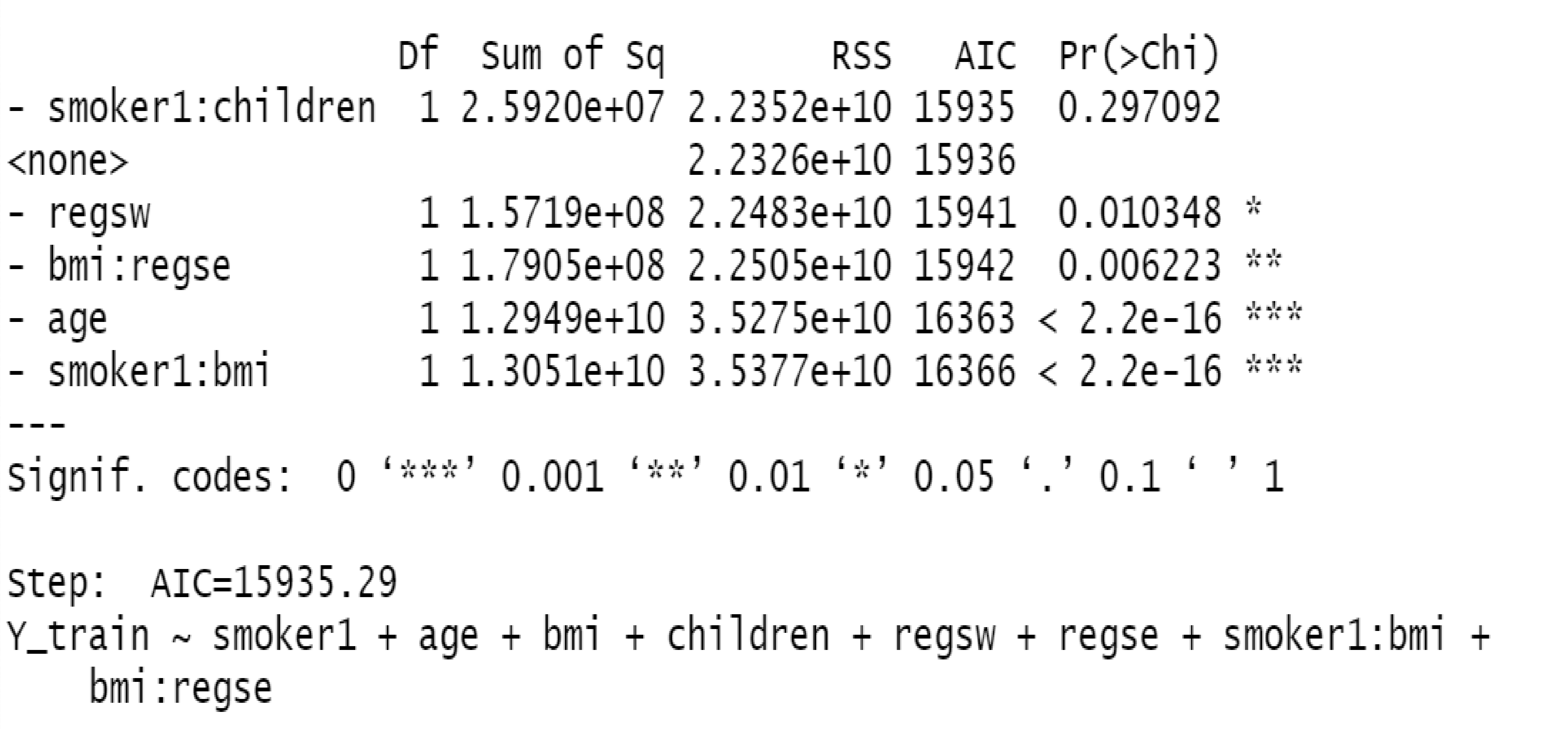
```
Start:  AIC=15939.4
Y_train ~ smoker1 + age + bmi + children + regsw + regse + smoker1:bmi +
    bmi:regse + age:bmi + smoker1:children + bmi:children

                   Df  Sum of Sq        RSS   AIC  Pr(>Chi)    
- age:bmi           1 8.8102e+06 2.2316e+10 15938  0.543006    
- bmi:children      1 1.1185e+07 2.2318e+10 15938  0.493129    
- smoker1:children  1 2.8368e+07 2.2335e+10 15939  0.275161    
<none>                           2.2307e+10 15939              
- regsw             1 1.5743e+08 2.2464e+10 15944  0.010257 *  
- bmi:regse         1 1.8134e+08 2.2488e+10 15945  0.005881 ** 
- smoker1:bmi       1 1.3026e+10 3.5333e+10 16368 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Step:  AIC=15937.77
Y_train ~ smoker1 + age + bmi + children + regsw + regse + smoker1:bmi +
    bmi:regse + smoker1:children + bmi:children
```

Iteration 2 :



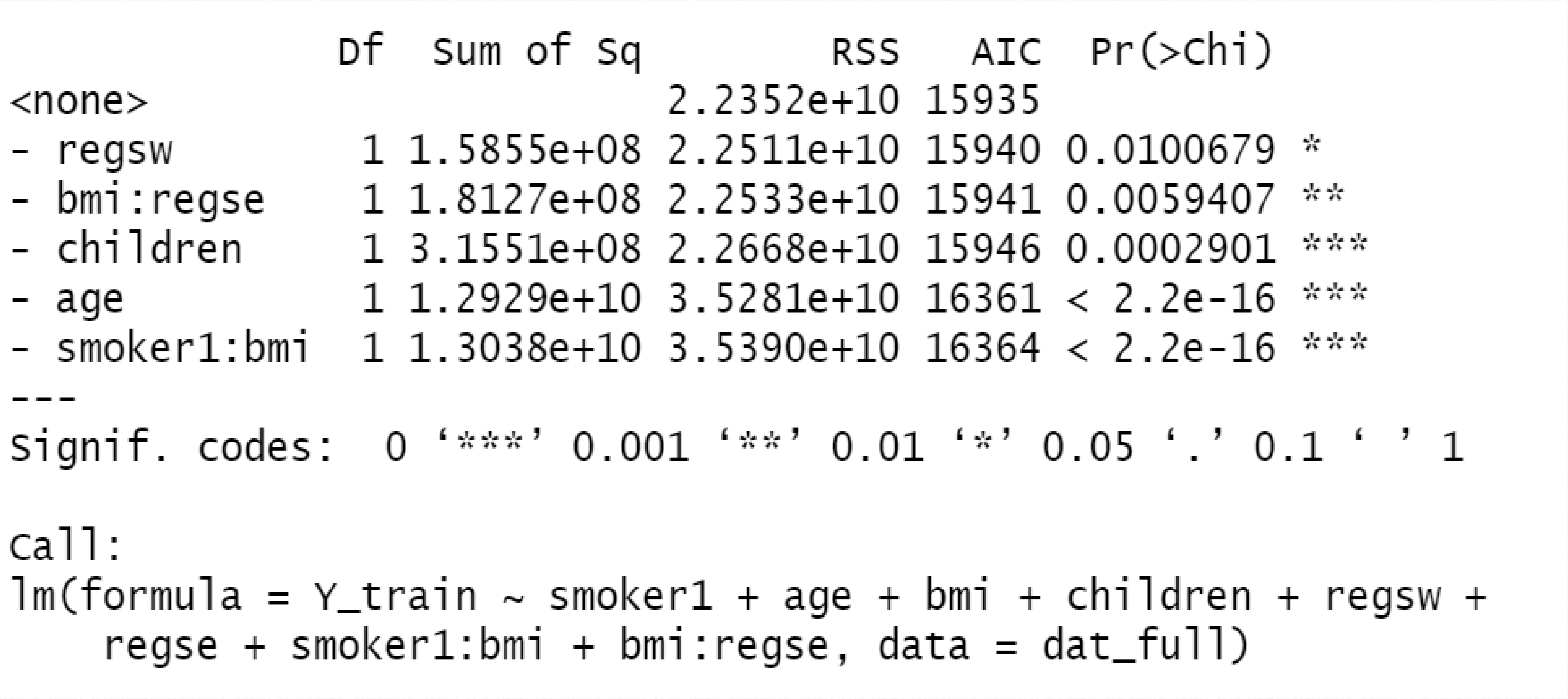
```
                    Df  Sum of Sq        RSS   AIC  Pr(>Chi)
- bmi:children       1 1.0325e+07 2.2326e+10 15936  0.510302
- smoker1:children   1 2.8049e+07 2.2344e+10 15937  0.277968
<none>                            2.2316e+10 15938
- regsw              1 1.5538e+08 2.2471e+10 15942  0.010778 *
- bmi:regse          1 1.7934e+08 2.2495e+10 15943  0.006169 **
- age                1 1.2958e+10 3.5274e+10 16365 < 2.2e-16 ***
- smoker1:bmi        1 1.3018e+10 3.5333e+10 16366 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Step:  AIC=15936.2
Y_train ~ smoker1 + age + bmi + children + regsw + regse + smoker1:bmi + 
    bmi:regse + smoker1:children
```

Iteration 3 :



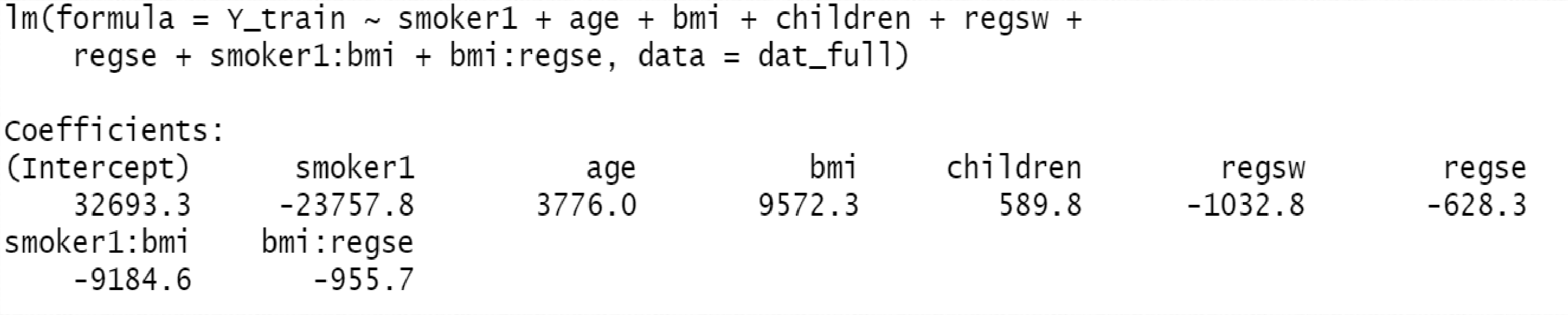
```
                   Df  Sum of Sq        RSS   AIC  Pr(>Chi)    
- smoker1:children  1 2.5920e+07 2.2352e+10 15935  0.297092    
<none>                           2.2326e+10 15936              
- regsw             1 1.5719e+08 2.2483e+10 15941  0.010348 *  
- bmi:regse         1 1.7905e+08 2.2505e+10 15942  0.006223 ** 
- age               1 1.2949e+10 3.5275e+10 16363 < 2.2e-16 ***
- smoker1:bmi       1 1.3051e+10 3.5377e+10 16366 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Step:  AIC=15935.29
Y_train ~ smoker1 + age + bmi + children + regsw + regse + smoker1:bmi + 
    bmi:regse
```

Iteration 4 :



```
              Df  Sum of Sq        RSS   AIC  Pr(>Chi)
<none>                      2.2352e+10 15935
- regsw        1 1.5855e+08 2.2511e+10 15940 0.0100679 *
- bmi:regse    1 1.8127e+08 2.2533e+10 15941 0.0059407 **
- children     1 3.1551e+08 2.2668e+10 15946 0.0002901 ***
- age          1 1.2929e+10 3.5281e+10 16361 < 2.2e-16 ***
- smoker1:bmi  1 1.3038e+10 3.5390e+10 16364 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Call:
lm(formula = Y_train ~ smoker1 + age + bmi + children + regsw +
    regse + smoker1:bmi + bmi:regse, data = dat_full)
```

Final model using Backward elimination:



```
lm(formula = Y_train ~ smoker1 + age + bmi + children + regsw +
    regse + smoker1:bmi + bmi:regse, data = dat_full)

Coefficients:
(Intercept)      smoker1          age          bmi     children        regsw        regse
    32693.3     -23757.8       3776.0       9572.3        589.8      -1032.8       -628.3
smoker1:bmi    bmi:regse
    -9184.6       -955.7
```

Analyzing the variables suggested by the algorithm:

The algorithm has suggested the following variables to be significant:

- Smoking status.
- Age
- BMI
- No. of children
- If region of residence was south west.
- If region of residence was south east.
- Interaction effect between smoking status and BMI
- Interaction effect between region south east and BMI

Interaction effect : Suppose Y is our response variable and A,B are explanatory variables. So, if effect of B on Y also depends on the value of A, then we say A interaction B is significant. It is denoted by A:B.

We had observed in EDA part that, higher BMI i.e. higher weight and a presence of smoking habit lead to a significant increase in charges of insurance, as compared to higher BMI and absence of smoking habit. Thus, the effect of BMI on charges also depends on whether the person smokes or not. Thus, interaction between bmi and smoking status is actually significant.

Further, let's analyze whether the other variables suggested by the algorithm, are significant or not.

We'll continue with the interaction term between BMI and region south east.

From this plot, we can see that,
after BMI = 40, there are
more individuals from region south
east as compared to other regions.

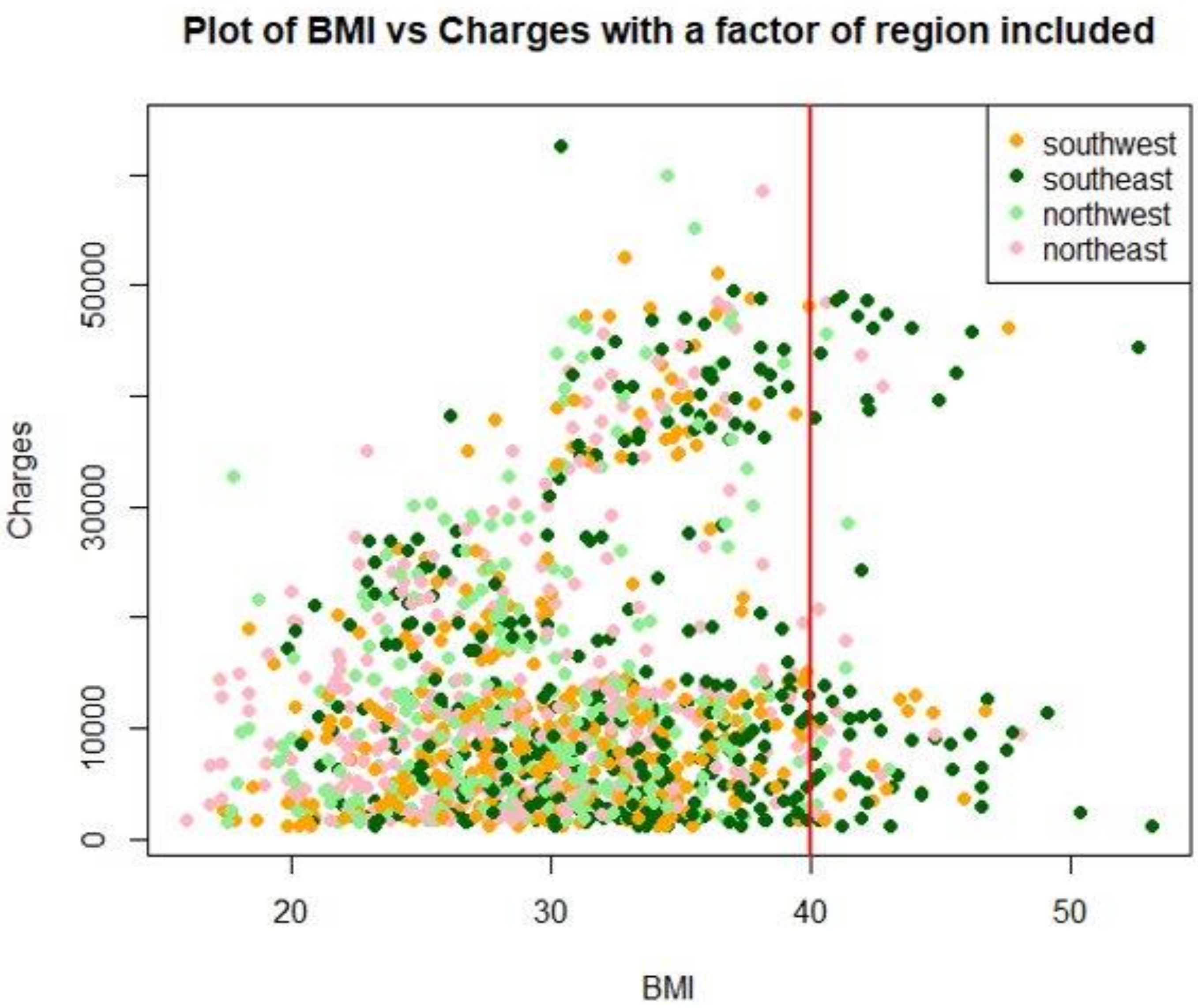
This actually means,

P(Person being from SE | Large BMI) is greater than,

P(Person being from SW | Large BMI)

P(Person being from NE | Large BMI)

P(Person being from NW | Large BMI)



$$P(Person\ being\ from\ south\ east|\ Large\ BMI)$$

$$= \frac{P(Large\ BMI\ |Person\ being\ from\ south\ east)*P(Person\ being\ from\ south\ east)}{P(Large\ BMI)}$$

We’ve split this using Bayes’ theorem.
Here, the numerator of this expression will change for,

P(Person being from SW | Large BMI)

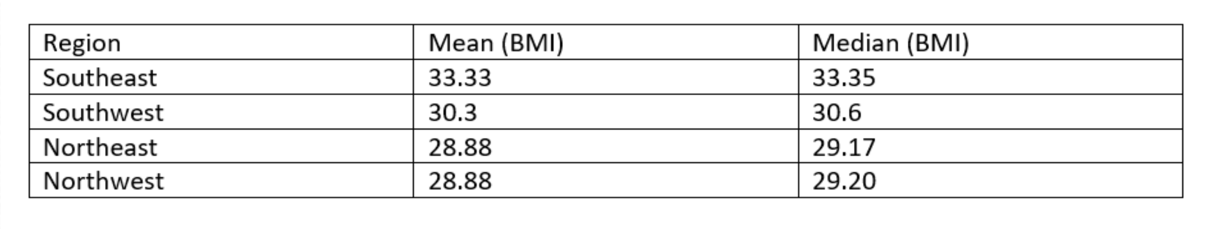
P(Person being from NE | Large BMI)

P(Person being from NW | Large BMI)

while the denominator remains the same for all these probabilities.

There can be 3 reasons for P(Person being from region SW | large BMI) being the largest:

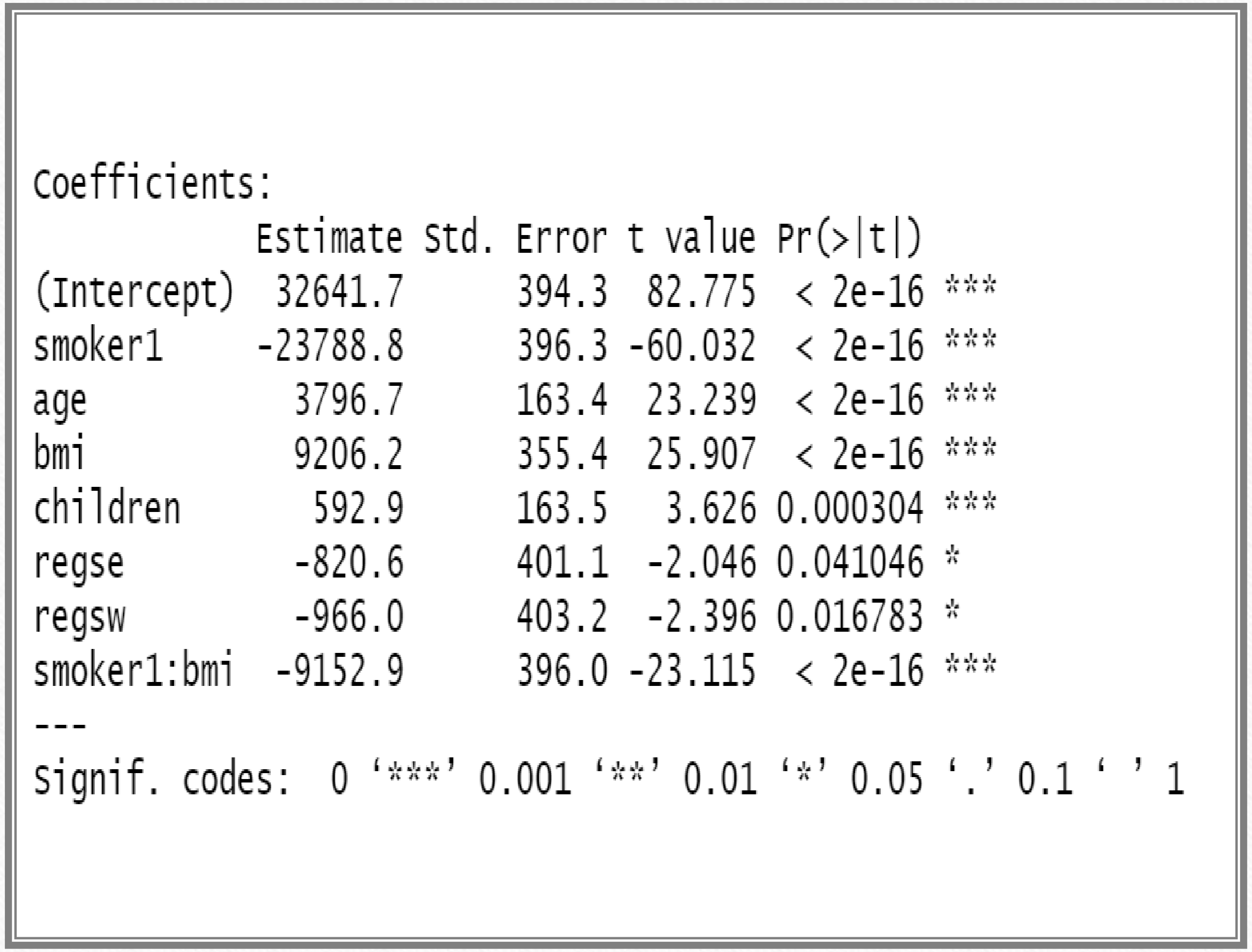
- The no. of people sampled from region southeast was a bit higher as compared to other regions.
- The BMI of individuals of southeast is higher on average as compared to other regions.
- Both the above reasons.
- But, on analyzing we have found that,



Region	Mean (BMI)	Median (BMI)
Southeast	33.33	33.35
Southwest	30.3	30.6
Northeast	28.88	29.17
Northwest	28.88	29.20

- Thus, we can observe that, the mean & median BMI is almost same for all regions. This means, our second reason and third reasons are eliminated. Thus, in reality we are getting bmi:region southeast as significant just because we have a higher no. of people from there.
- Thus, 2nd interaction effect isn't significant. Hence, we have removed that interaction from our model now.

ANOVA (For individual significance)

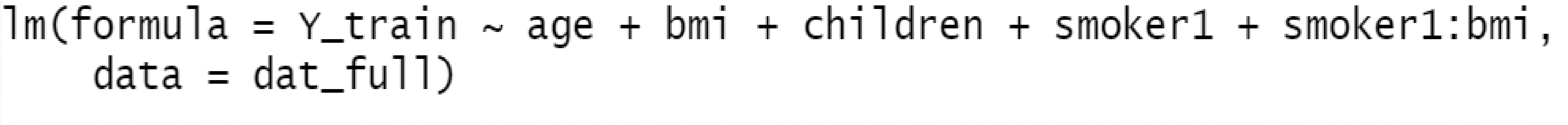


```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  32641.7      394.3  82.775  < 2e-16 ***
smoker1     -23788.8      396.3 -60.032  < 2e-16 ***
age           3796.7      163.4  23.239  < 2e-16 ***
bmi           9206.2      355.4  25.907  < 2e-16 ***
children       592.9      163.5   3.626 0.000304 ***
regse         -820.6      401.1  -2.046 0.041046 *  
regsw         -966.0      403.2  -2.396 0.016783 *  
smoker1:bmi  -9152.9      396.0 -23.115  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

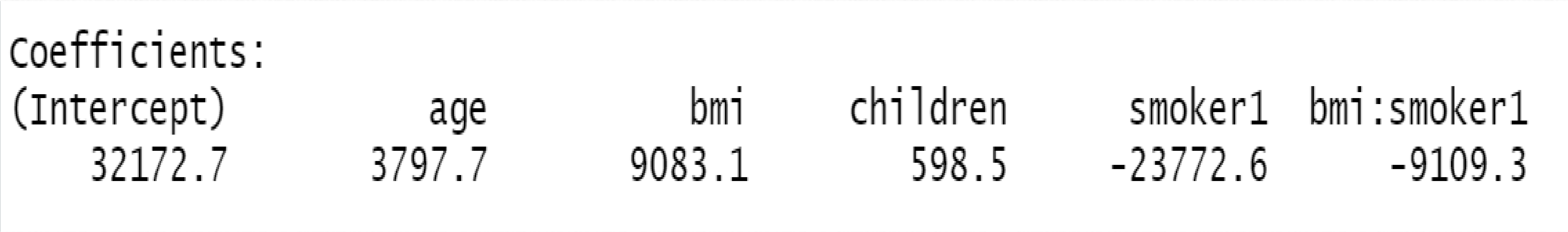
From this we can observe that, region SE and SW here are not that significant.

Logically, if we think about this, we understand that we should include these coefficients in the model, only when they have some significant impact on charges, as compared to other regions. But, on analyzing, we found that all regions have almost equal chance of paying higher charges. So, why include region SW and SE in our model?

So, we do not add the variables region SE and region SW in our model. Thus, our final model is then as follows:

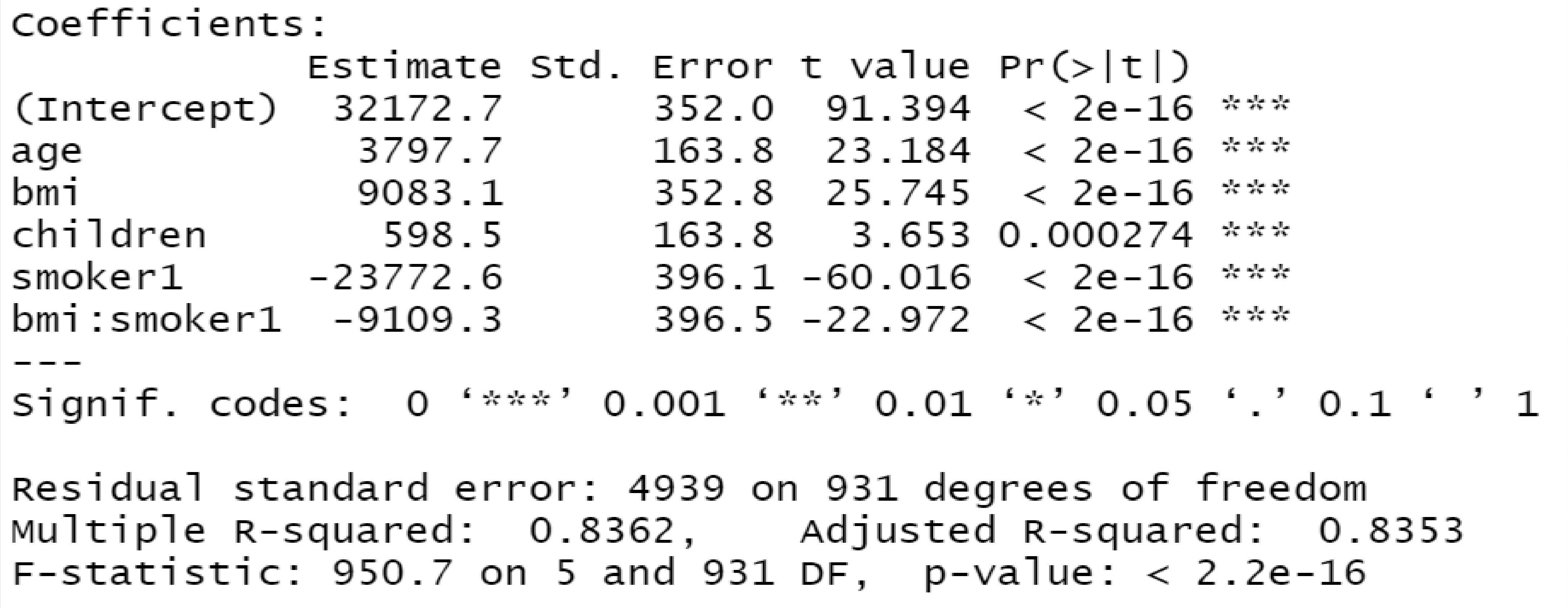


```
lm(formula = Y_train ~ age + bmi + children + smoker1 + smoker1:bmi,
    data = dat_full)
```



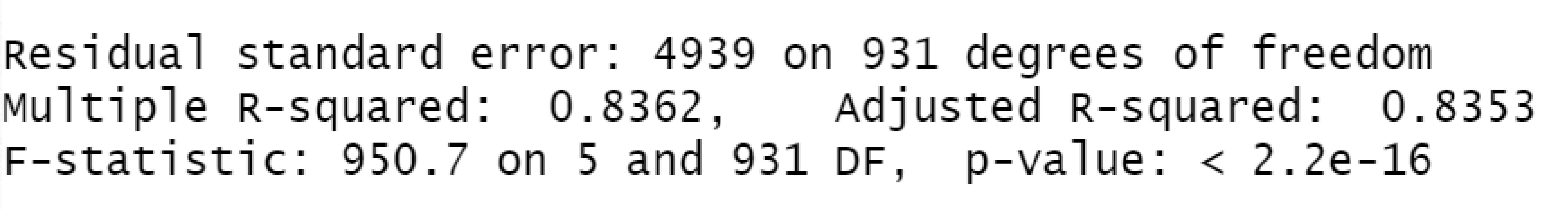
```
Coefficients:
(Intercept)          age          bmi     children      smoker1  bmi:smoker1
    32172.7       3797.7       9083.1        598.5     -23772.6      -9109.3
```

Summary of the fitted model & Individual significance.



```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)   32172.7      352.0  91.394  < 2e-16 ***
age            3797.7      163.8  23.184  < 2e-16 ***
bmi            9083.1      352.8  25.745  < 2e-16 ***
children        598.5      163.8   3.653 0.000274 ***
smoker1      -23772.6      396.1 -60.016  < 2e-16 ***
bmi:smoker1   -9109.3      396.5 -22.972  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4939 on 931 degrees of freedom
Multiple R-squared:  0.8362,    Adjusted R-squared:  0.8353 
F-statistic: 950.7 on 5 and 931 DF,  p-value: < 2.2e-16
```

ANOVA (For overall significance)



```
Residual standard error: 4939 on 931 degrees of freedom
Multiple R-squared:  0.8362,	Adjusted R-squared:  0.8353 
F-statistic: 950.7 on 5 and 931 DF,  p-value: < 2.2e-16
```

For testing the overall significance of our model:
Null : All variables have no significant linear relationship with the dependent variable charges.
Alternative : At least 1 variable has a significant linear relationship with the dependent variable charges.
From above output, we can see that, the p-value is very low indeed. Thus, we can reject our null hypothesis and may conclude that our model is overall significant at 5% level of significance.

Checking the assumptions of Classical Linear Regression Model:

Following assumptions will be checked in the subsequent slides:

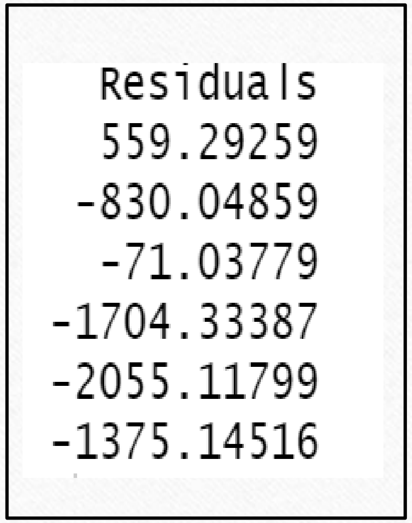
- Error distribution for the fitted model.
- Auto-correlation between residuals.
- Homoscedasticity of residuals.
- Multicollinearity between independent variables.

Error distribution :

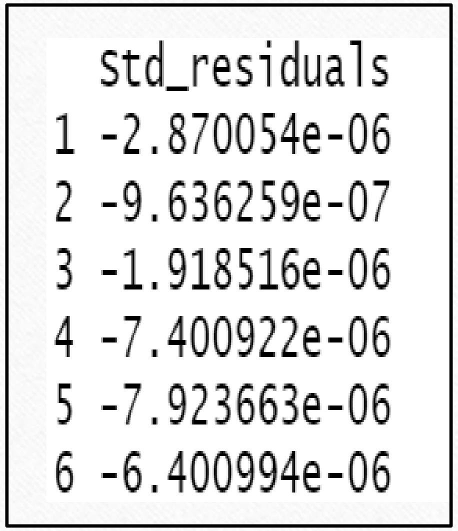
Checking if residuals follow a normal distribution. For checking the error distribution, calculating the errors first.

Errors = Y actual – Y predicted

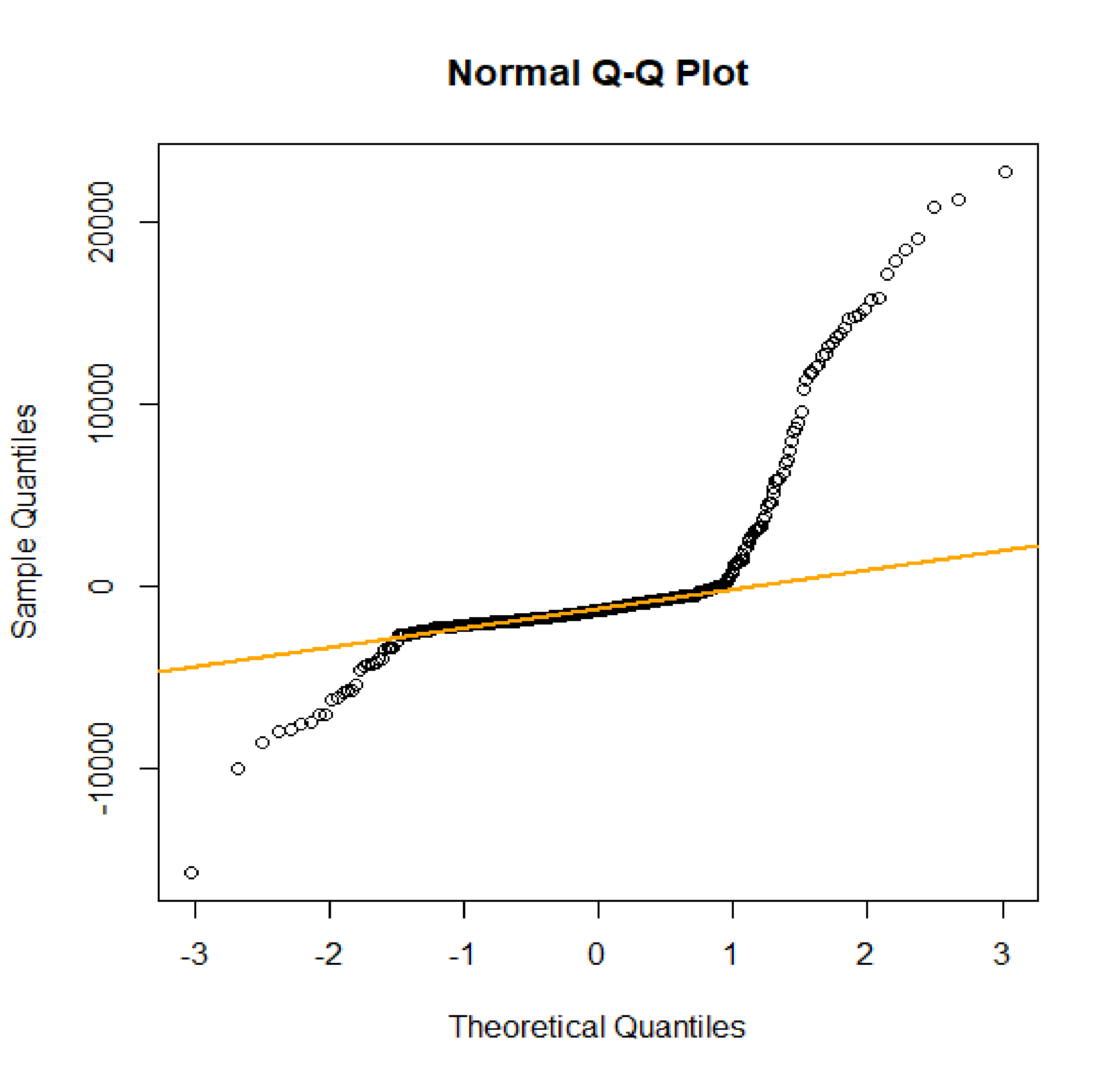
We've used the test data, consisting 30% of total data for the prediction. Then using these predicted, residuals have been calculated. First few residuals and standardized residuals calculated are as follows:



```
   Residuals
   559.29259
  -830.04859
   -71.03779
 -1704.33387
 -2055.11799
 -1375.14516
```



```
  Std_residuals
1 -2.870054e-06
2 -9.636259e-07
3 -1.918516e-06
4 -7.400922e-06
5 -7.923663e-06
6 -6.400994e-06
```



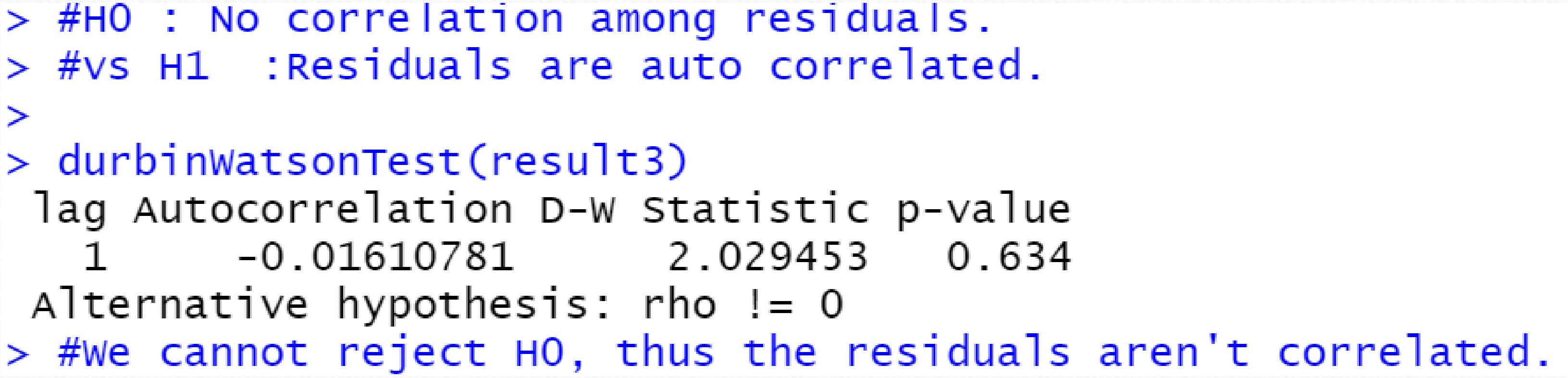
From this plot we can see the distribution of residuals is heavy tailed i.e. it has fatter tails as compared to normal distribution. Thus, our regression model is violating one of the assumptions of CLRM.

Autocorrelation:

Our assumption was that there is no autocorrelation between the residuals.

Checking whether that assumption holds true. This is an important assumption, since presence of autocorrelation makes our OLS estimation invalid. We would have to use WLS (Weighted Least Squares) method in that case.

We've used Durbin Watson test in R, to check whether autocorrelation exists. The code and output for same is as follows:



```
> #H0 : No correlation among residuals.
> #vs H1  :Residuals are auto correlated.
>
> durbinWatsonTest(result3)
 lag Autocorrelation D-W Statistic p-value
   1     -0.01610781      2.029453   0.634
 Alternative hypothesis: rho != 0
> #We cannot reject H0, thus the residuals aren't correlated.
```

Homoscedasticity:

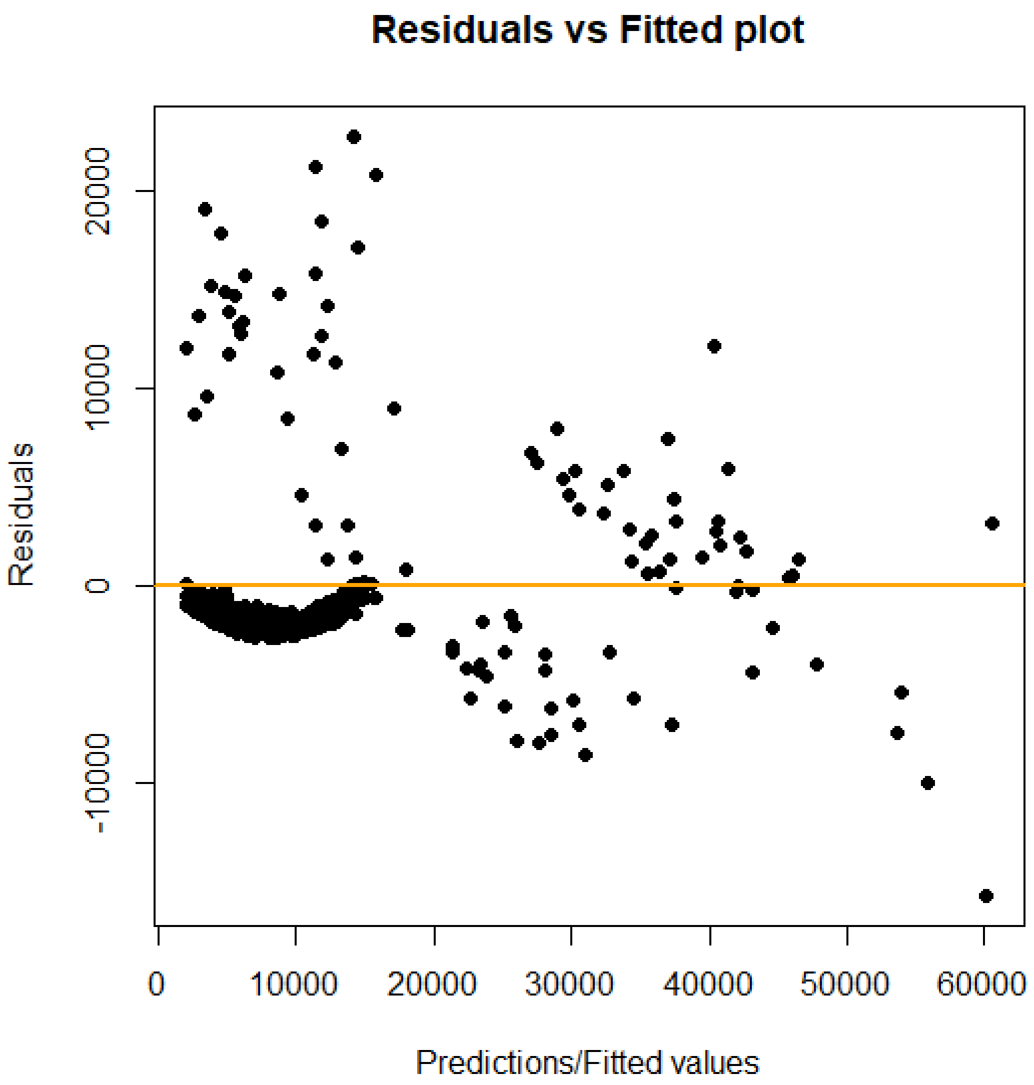
Checking if errors have equal variance.

To do this, we will first plot residuals vs fitted values.

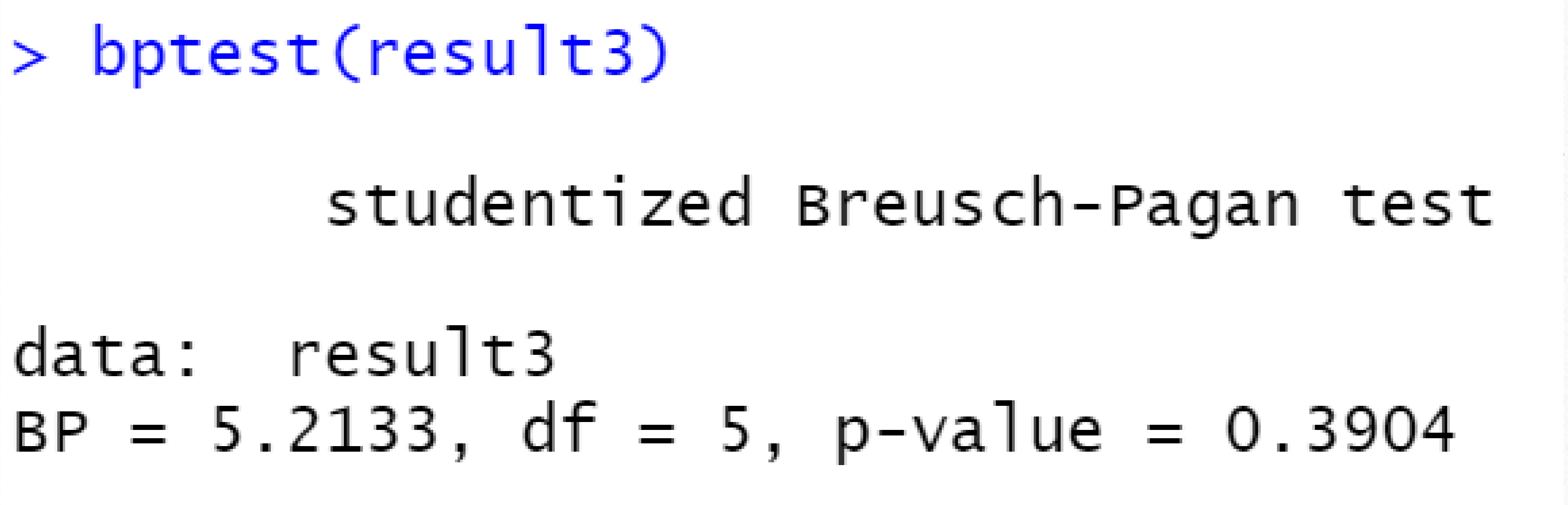
From this we have 2 conclusions:

- For some part of residuals, it seems that the response variable is non-linearly correlated with the independent variables.
- And for the remaining part, the residuals have a negative correlation with fitted values. This implies, there is some variable outside our dataset, which has some significant relationship with the response variable 'Charges'.

Thus, now it becomes difficult to judge homoscedasticity using this plot. Thus, we go for an alternative way to check this, i.e. to use the BPG test.



Now, we formally checked this using BPG test, the output is as follows:



```
> bptest(result3)

        studentized Breusch-Pagan test

data:  result3
BP = 5.2133, df = 5, p-value = 0.3904
```

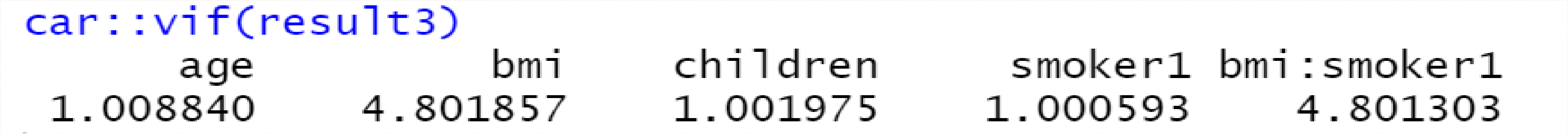
The null hypothesis in this case was that heteroscedasticity is absent, alternative being heteroscedasticity is present.

Thus, we conclude that since the p-value is not significant, we can't reject null. And hence, the residuals have equal variances.

Multicollinearity :

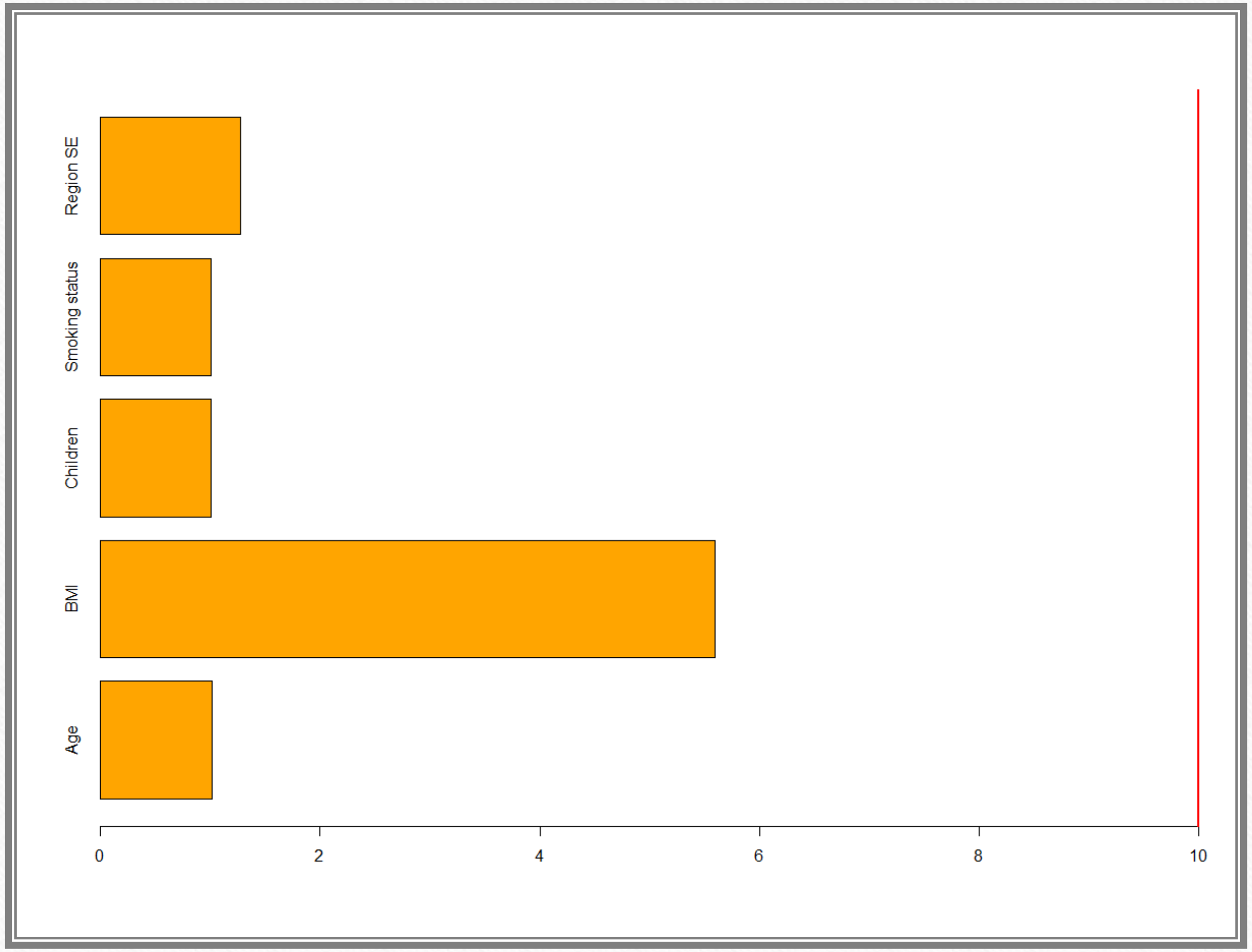
Checking if the independent variables have any correlation between them.

To check this, we had already plotted a heatmap of correlation between all the variables in our data. From that, we could see that there was not much correlation between any 2 independent variables. To formally check this, we use the Variance Inflation Factor (VIF). The output for the same has been pasted below.



```
car::vif(result3)
      age         bmi     children      smoker1 bmi:smoker1
 1.008840    4.801857     1.001975     1.000593    4.801303
```

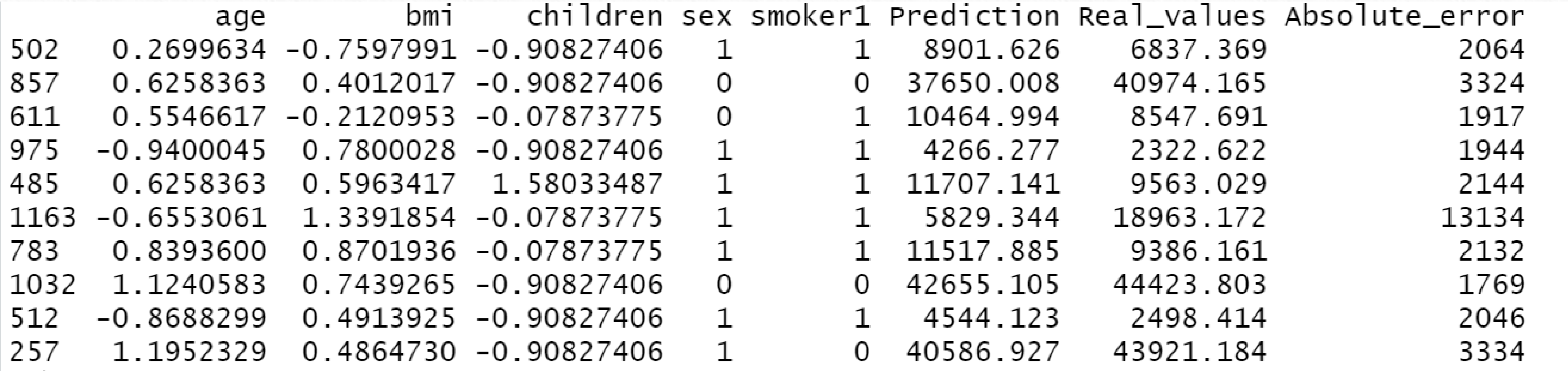
Each value of VIF can be examined for understanding whether the corresponding independent variable is highly correlated with other independent variables or not.



To easily understand and better represent whether the VIF for our independent variables is larger than a value of 10 or not, we've drawn this bar plot.

We conclude that, multicollinearity is not present.

Real-Time Prediction from the model



```
                age         bmi    children sex smoker1 Prediction Real_values Absolute_error
502   0.2699634 -0.7597991 -0.90827406   1       1   8901.626    6837.369           2064
857   0.6258363  0.4012017 -0.90827406   0       0  37650.008   40974.165           3324
611   0.5546617 -0.2120953 -0.07873775   0       1  10464.994    8547.691           1917
975  -0.9400045  0.7800028 -0.90827406   1       1   4266.277    2322.622           1944
485   0.6258363  0.5963417  1.58033487   1       1  11707.141    9563.029           2144
1163 -0.6553061  1.3391854 -0.07873775   1       1   5829.344   18963.172          13134
783   0.8393600  0.8701936 -0.07873775   1       1  11517.885    9386.161           2132
1032  1.1240583  0.7439265 -0.90827406   0       0  42655.105   44423.803           1769
512  -0.8688299  0.4913925 -0.90827406   1       1   4544.123    2498.414           2046
257   1.1952329  0.4864730 -0.90827406   1       0  40586.927   43921.184           3334
```

The absolute error column is actually the absolute value of residuals. The error is very large in terms of dollars. So, we can see that the predictions given by our model are not good. This might be because the distribution of errors was not normal for our model.

Conclusions:

- We had taken secondary data for predicting medical insurance costs of individuals based on their various features.
- We had equal no. of male and female respondents in our dataset.
- Charges have a right skewed distribution.
- BMI has an almost normal distribution.
- There are more non-smokers in our dataset as compared to smokers.
- The smoking status, is most correlated with our dependent variable.
- As age increases, the charges for medical insurance increase too.
- As BMI increases, the charges for medical insurance increase too.
- Medical insurance cost for smokers was higher as compared to non-smokers.
- Medical insurance costs for obese smokers were significantly higher than obese non-smokers.
- We arrived at the same model using both Forward selection & Backward elimination method.
- The adjusted R squared for our fitted model was about 84% .
- The assumption of error distribution being normal was violated by our model.
- While the assumptions of Homoscedasticity, absence of autocorrelation and absence of multicollinearity were satisfied.

Bibliography :

- https://www.kaggle.com/
- https://www.r-bloggers.com/
- http://www.sthda.com/
- https://www.statology.org/
- https://stackoverflow.com/
- https://www.geeksforgeeks.org/
- https://towardsdatascience.com/
- https://en.wikipedia.org/

Thank you.