Linear Regression Analysis on Insurance Data

By: Kriti Srivastava

Background:

Medical costs incurred by insurance companies are constantly changing. Insurance companies have been charged billions of dollars for treatment by medical providers meanwhile no two individuals with the same health condition has ever been charged the same. Companies selling health insurance are businesses. They need to set premiums at a level that at least covers the costs of selling the policy, administering the policy and maintaining adequate funds to pay claims relating to the medical benefits provided to subscribers. The ACA has also restricted how premium rates can be set. Beginning from 2014, on the basis of pre-existing conditions such as diabetes, asthma, pregnancy or a disability policies cannot charge you more. Under the health care law, insurance companies are only allowed to adjust the price of the premium based on these factors: Age, Sex, BMI, Number of dependents, Region, Smoker or non-smoker.

The following dataset is a simulated data and not a real time data which shows the different charges of medical insurance on the basis of different values of the factors.

Problem Statement:

From this dataset I am trying to predict how premium charges are affected by a person's age, sex, BMI, gender, whether smoker or non-smoker, number of dependents and the region of his stay. We will also try to analyze the increase in charges when the person is a smoker as opposed to a non-smoker.

Number of dependent variables and description:

- Charges - Individual medical costs billed by health insurance. It is a quantitative variable

Number of independent variables and description:

- Age - Primary beneficiary age. It is a quantitative variable.
- Sex – Gender of Insurance contractor’s: female, male. It is a qualitative variable.
- BMI - Body mass index, providing an understanding of body, weights that are relatively high or low relative to height, objective index of body weight (kg/m^2) using the ratio of height to weight, ideally 18.5 to 24.9. It is a quantitative variable
- Children - Number of children covered by health insurance/Number of dependents. It is a quantitative variable.
- Smoker - Is the person a smoker or not. It is a qualitative variable.
- Region - The residential area of the beneficiary in the US: northeast, southeast, southwest, northwest. It is a qualitative variable

Cleaning dataset:

The dataset we used was already clean, therefore we need not to do anything in cleaning.

Import data:

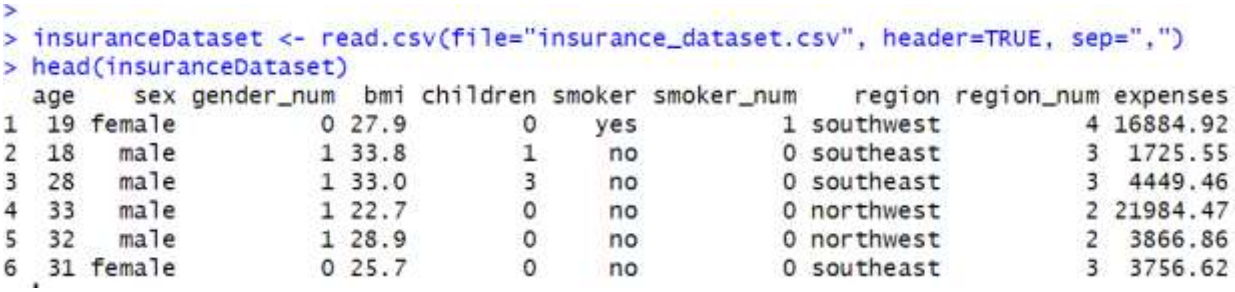
The url for the data set the dataset is www.kaggle.com/mirichoi0218/insurance . The data file name is insurance.cvs. This file contains simulated data on the basis of demographic statistics from the US Census Bureau, according to the book (Machine Learning with R by Brett Lantz) from which it is from. The book that provides an introduction to machine learning using R.

The insurance.cvs. is imported into the dataset named as “Insurance” using SAS 9.4 with 7 variables.

- **age**: age of primary beneficiary in years
- **sex**: insurance contractor gender, female, male
- **bmi**: Body mass index, providing an understanding of body, weights that are relatively high or low relative to height, objective index of body weight (kg / m ^ 2) using the ratio of height to weight, ideally 18.5 to 24.9

- **children**: Number of children covered by health insurance / Number of dependents
- **smoker**: Smoking
- **region**: the beneficiary's residential area in the US, northeast, southeast, southwest, northwest.
- **charges**: Individual medical costs billed by health insurance.

Type of data in data set



```
>
> insuranceDataset <- read.csv(file="insurance_dataset.csv", header=TRUE, sep=",")
> head(insuranceDataset)
  age    sex gender_num  bmi children smoker smoker_num    region region_num expenses
1  19 female          0 27.9        0    yes          1 southwest          4 16884.92
2  18   male          1 33.8        1     no          0 southeast          3  1725.55
3  28   male          1 33.0        3     no          0 southeast          3  4449.46
4  33   male          1 22.7        0     no          0 northwest          2 21984.47
5  32   male          1 28.9        0     no          0 northwest          2  3866.86
6  31 female          0 25.7        0     no          0 southeast          3  3756.62
```

Fig:A

Checking the data to make sure that the data set is as per the description

Number of samples = 1338

Number of columns = 10

1 Dependent variable: expenses’

5 Independent variable: Age, gender_num, bmi, children, smoker_num, region_num

Number of samples with missing values: 0

In ten columns, ‘expenses’ is the dependent variable. Age, gender_num, bmi, children, smoker_num, region_num are the independent variable or features. Since there is the numerical conversion of sex, smoker, and region, therefore, we can drop these three variables.

Pre Processing

Creating the dummy variables for region:

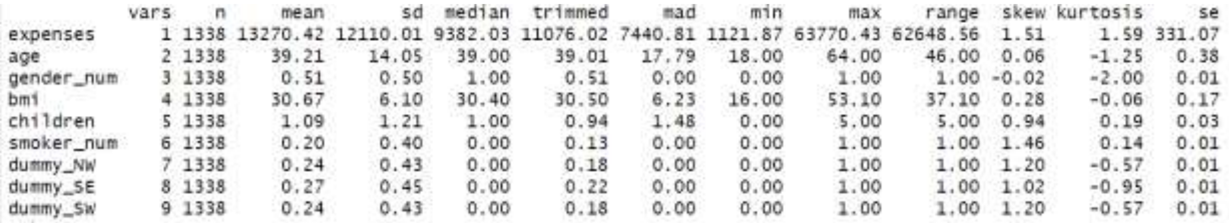
Keeping northeast as base value = 0. Then, creating a three binary dummy variables with two values as 0 and 1.

- **Dummy_NW:** with 1 when region = “northwest ” otherwise 0.
- **Dummy_SE:** with 1 when region = “southeast ” otherwise 0.
- **Dummy_SW:** with 1 when region = “southwest” otherwise 0.

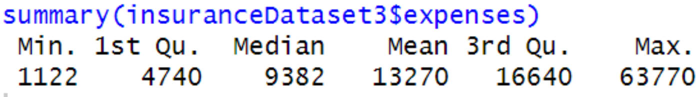
Since, I dummy variables are used for each region therefore, removing region_num from data set.

Data Description

Data description of all the variables can be seen in the below figure(fig:1):



```
           vars    n     mean      sd  median trimmed     mad     min      max    range  skew kurtosis     se
expenses      1 1338 13270.42 12110.01 9382.03 11076.02 7440.81 1121.87 63770.43 62648.56  1.51     1.59 331.07
age           2 1338    39.21    14.05   39.00    39.01   17.79   18.00    64.00    46.00  0.06    -1.25   0.38
gender_num    3 1338     0.51     0.50    1.00     0.51    0.00    0.00     1.00     1.00 -0.02    -2.00   0.01
bmi           4 1338    30.67     6.10   30.40    30.50    6.23   16.00    53.10    37.10  0.28    -0.06   0.17
children      5 1338     1.09     1.21    1.00     0.94    1.48    0.00     5.00     5.00  0.94     0.19   0.03
smoker_num    6 1338     0.20     0.40    0.00     0.13    0.00    0.00     1.00     1.00  1.46     0.14   0.01
dummy_NW      7 1338     0.24     0.43    0.00     0.18    0.00    0.00     1.00     1.00  1.20    -0.57   0.01
dummy_SE      8 1338     0.27     0.45    0.00     0.22    0.00    0.00     1.00     1.00  1.02    -0.95   0.01
dummy_SW      9 1338     0.24     0.43    0.00     0.18    0.00    0.00     1.00     1.00  1.20    -0.57   0.01
```



```
summary(insuranceDataset3$expenses)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 1122    4740    9382   13270   16640   63770
```

Fig:1

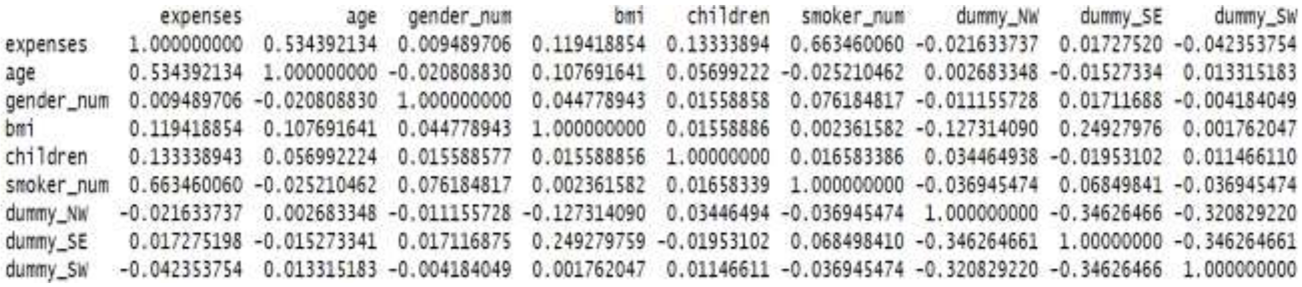
According to the samples we have, the median of expenses is 9382 dollars that mean most of the people pay $9382 for insurance where the middle 50% of the distribution of expenses ranges from $4740 to $16640. The people have to pay a minimum of $1121.87 and can go up to $63770.43 on insurance expenses.

a). Multicolliniarity Check

Multicollinearity can be checked in two way:

1) calculating the linear correlation strength between all the exploratory variables or features. If there are any two features with high linear relation with strength between (-1 to -0.7) or (0.7 to 1) then those two features are correlated to each other have the issue of multicollinearity. To remove multicollinearity, any one feature from the high linearly correlated feature pair has to be removed.

2) Using Variance inflation factor VIF: VIF measures) measures the severity of multicollinearity in regression analysis. Any exploratory variable has the VIF greater 10 which means that the variable is multicollinear with any other exploratory variable in the regression model.



```
               expenses          age   gender_num          bmi    children   smoker_num     dummy_NW    dummy_SE     dummy_SW
expenses    1.000000000  0.534392134  0.009489706  0.119418854  0.13333894  0.663460060 -0.021633737  0.01727520 -0.042353754
age         0.534392134  1.000000000 -0.020808830  0.107691641  0.05699222 -0.025210462  0.002683348 -0.01527334  0.013315183
gender_num  0.009489706 -0.020808830  1.000000000  0.044778943  0.01558858  0.076184817 -0.011155728  0.01711688 -0.004184049
bmi         0.119418854  0.107691641  0.044778943  1.000000000  0.01558886  0.002361582 -0.127314090  0.24927976  0.001762047
children    0.133338943  0.056992224  0.015588577  0.015588856  1.00000000  0.016583386  0.034464938 -0.01953102  0.011466110
smoker_num  0.663460060 -0.025210462  0.076184817  0.002361582  0.01658339  1.000000000 -0.036945474  0.06849841 -0.036945474
dummy_NW   -0.021633737  0.002683348 -0.011155728 -0.127314090  0.03446494 -0.036945474  1.000000000 -0.34626466 -0.320829220
dummy_SE    0.017275198 -0.015273341  0.017116875  0.249279759 -0.01953102  0.068498410 -0.346264661  1.00000000 -0.346264661
dummy_SW   -0.042353754  0.013315183 -0.004184049  0.001762047  0.01146611 -0.036945474 -0.320829220 -0.34626466  1.000000000
```

Fig:2

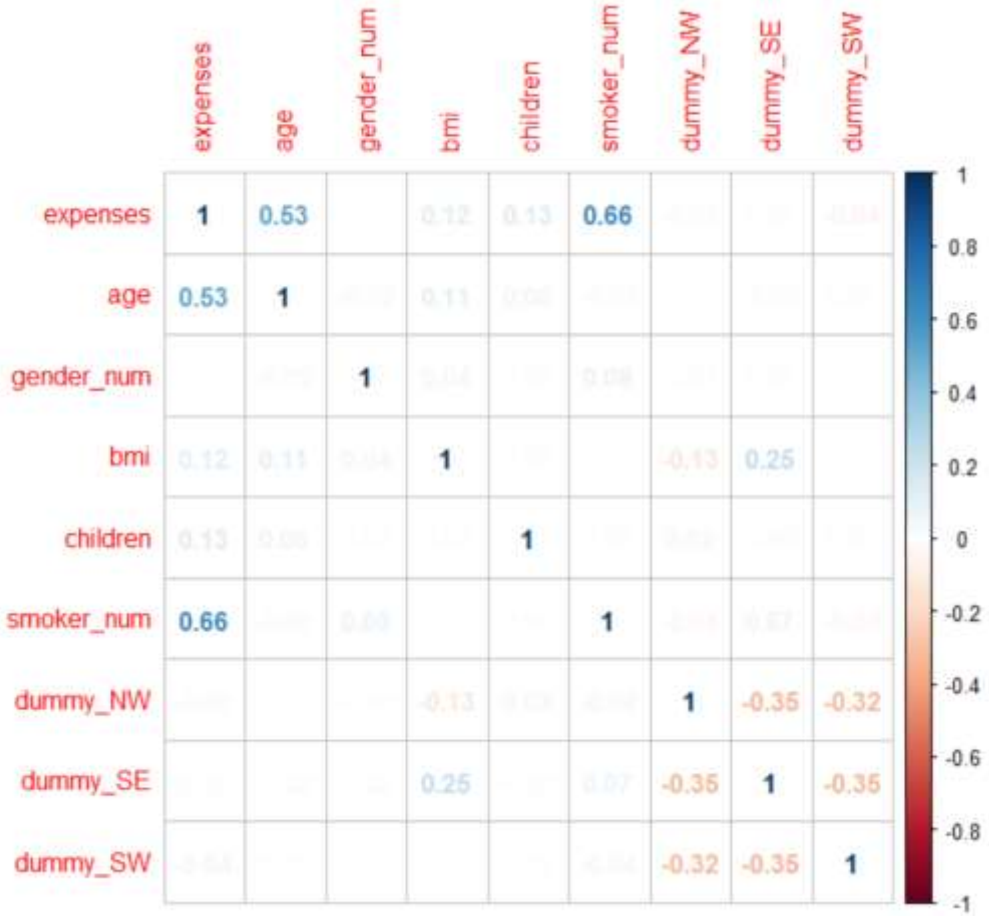


Fig:3

On looking at the figures (fig:2 and fig:3), we can see that there is no strong linear relation between two or more explanatory variables because none of the explanatory variables as strength between (-1 to -0.7) or (0.7 to 1). Therefore, we can see that there is no issue of multicollinearity hence no action is needed.

Creating the Initial Linear Regression Model with entering Method

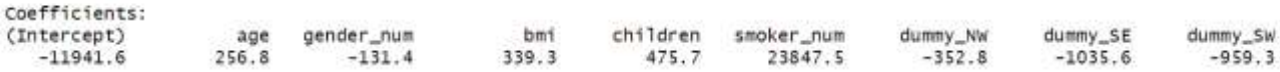
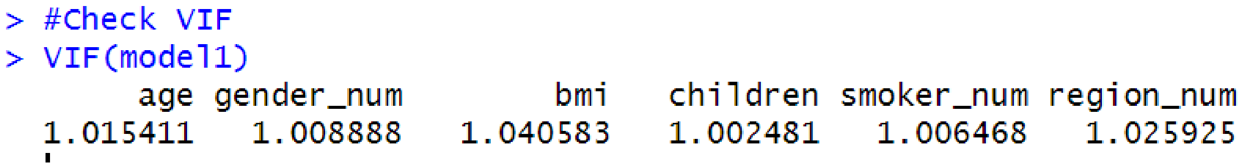

```
Coefficients:
(Intercept)          age   gender_num          bmi     children   smoker_num     dummy_NW     dummy_SE     dummy_SW
   -11941.6        256.8       -131.4        339.3        475.7      23847.5       -352.8      -1035.6       -959.3
```

Fig:4

expenses = -11941.6 + 256.8*age – 131.4*gender_num + 339.3*bmi + 475.7*children + 23847.5*smoker_num - 352.8*dummy_NW - 1035.6*dummy_SE – 959.3 dummy_SW

where assuming :

expenses are in dollars,
age is in years
gender_num is 1 when sex= female and 0 when sex= male,
bmi i is body mass index and it has 1 unit = 1(kg / m ^ 2).
children is the number of children
smoker_num = 1 when the smoker is yes and 0 when the smoker is no,
dummy_NW = 1 when the region is "northwest" and 0 otherwise,
dummy_SE = 1 when the region is "southeast" and 0 otherwise,
dumy_SW = 1 when the region is "southwest" and 0 otherwise



```
> #Check VIF
> VIF(model1)
       age gender_num        bmi   children smoker_num region_num 
  1.015411   1.008888   1.040583   1.002481   1.006468   1.025925 
```

Fig:5

We can observe in fig:5, there is no exploratory variable that has VIF more than 10 in the model (fig:4). Therefore, no issue of multicollinearity hence no action is needed.

Assumptions:

- **Normality**:

Since bmi and expenses are the only quantitative variables, therefore, we check the normality of these two variables

Histogram of Expenses

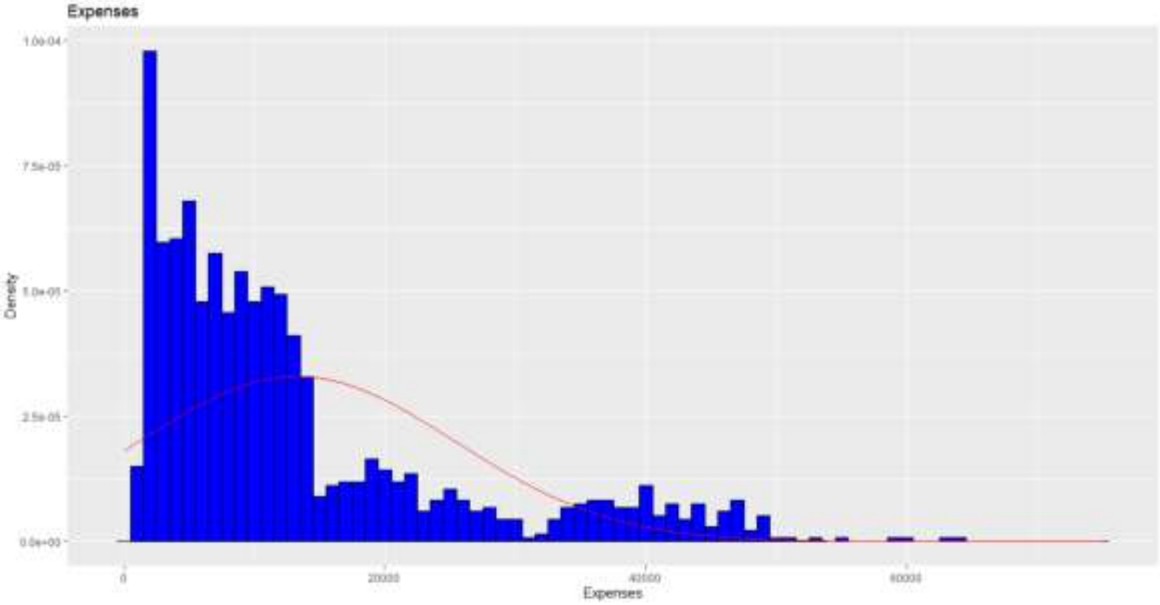


Fig:6

On looking into the histogram (fig:6) of expenses, we can see that expenses distribution is not symmetric but right-skewed. Therefore, the transformation of expenses is required.

Histogram after transforming expenses to log(expenses).

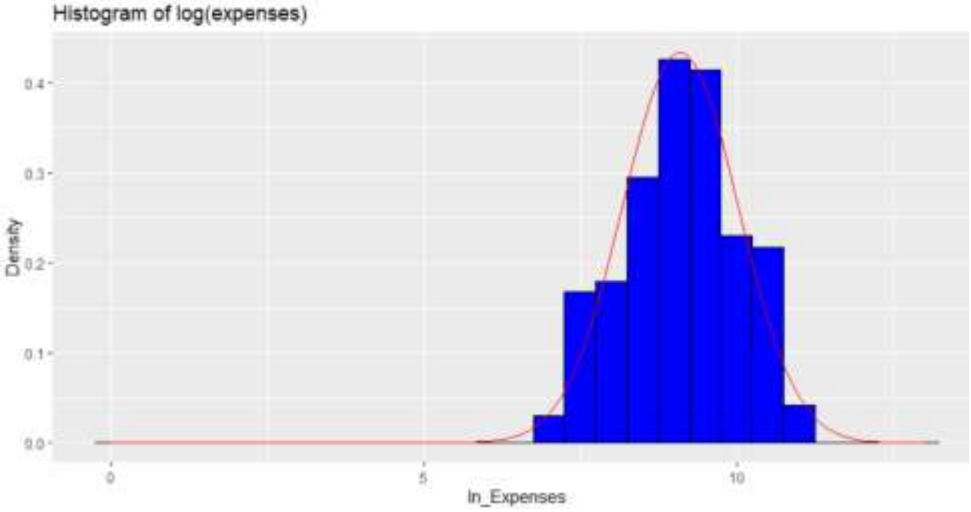


Fig:7

Histogram of log(expenses) is normal.

Histogram of BMI.

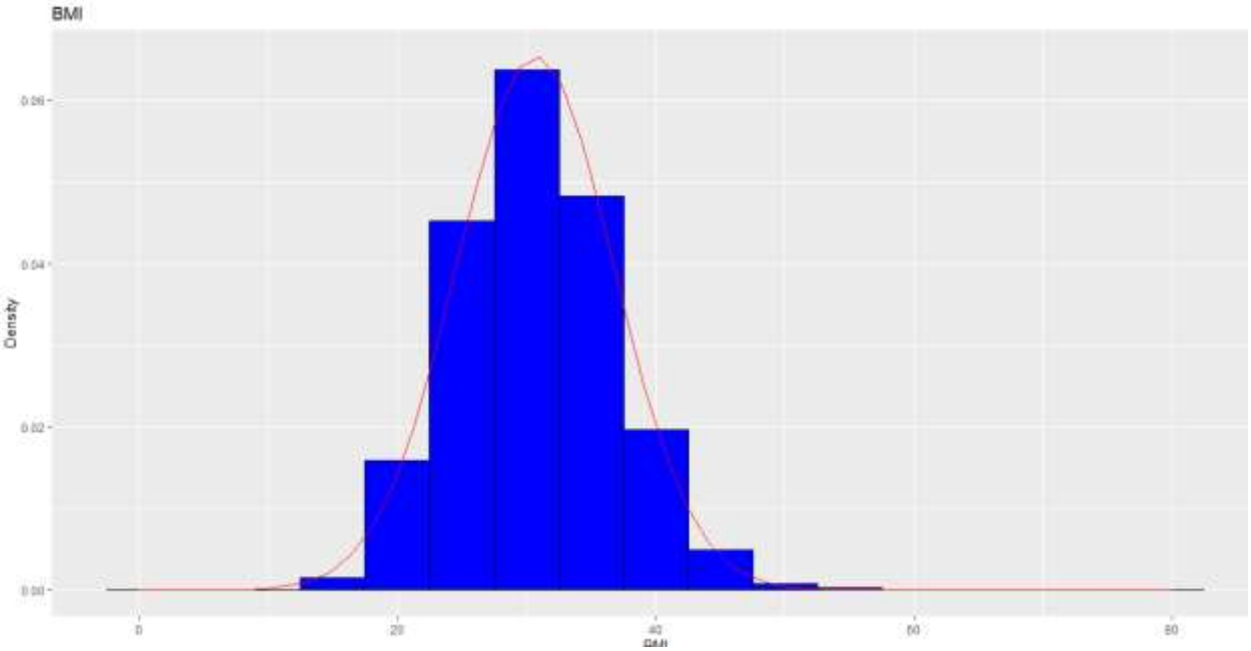


Fig :8

Distribution of the bmi looks normally distributed.

Check for multicollinearity after transforming expenses to log(expenses)

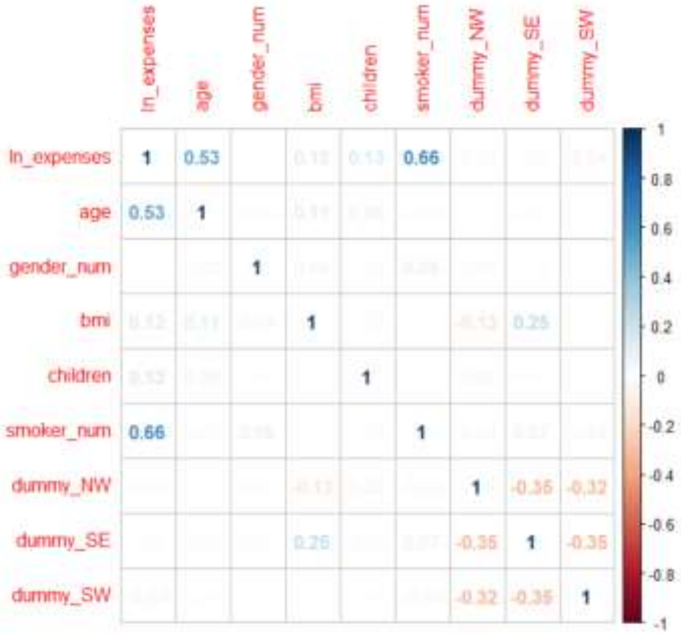
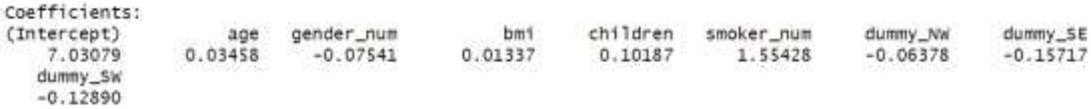


Fig:9

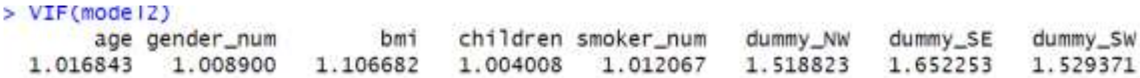
We can see fig:9, there are no exploratory variables that are strongly linear correlated.

Creating the Linear Regression Model after transforming expenses with Enter Method



```
Coefficients:
(Intercept)          age   gender_num          bmi     children   smoker_num     dummy_NW     dummy_SE
    7.03079      0.03458     -0.07541      0.01337      0.10187      1.55428     -0.06378     -0.15717
   dummy_SW
   -0.12890
```

Fig:10

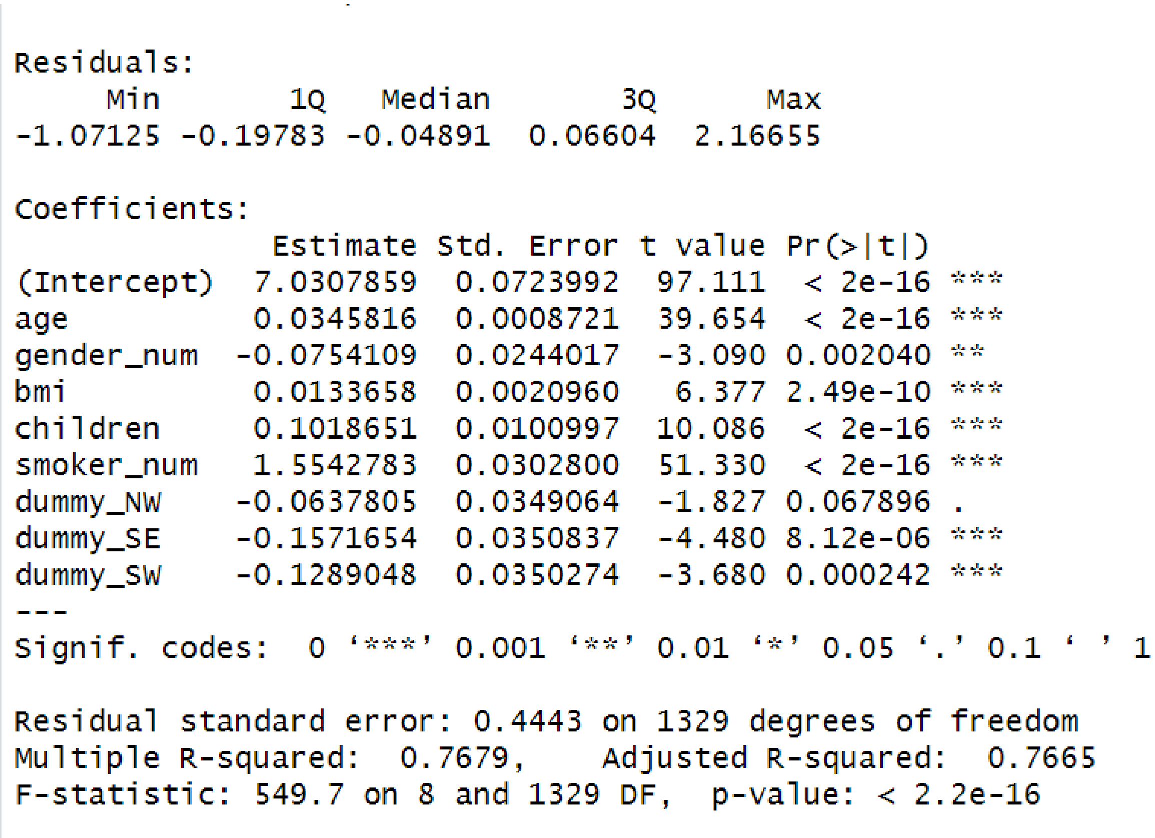


```
> VIF(model2)
       age gender_num        bmi   children smoker_num   dummy_NW   dummy_SE   dummy_SW
  1.016843   1.008900   1.106682   1.004008   1.012067   1.518823   1.652253   1.529371
```

Fig:11

We can observe in fig:11, there is no exploratory variable that has VIF more than 10 in the model (fig:10). Therefore, no issue of multicollinearity hence no action is needed.

Summary of the model after transformation



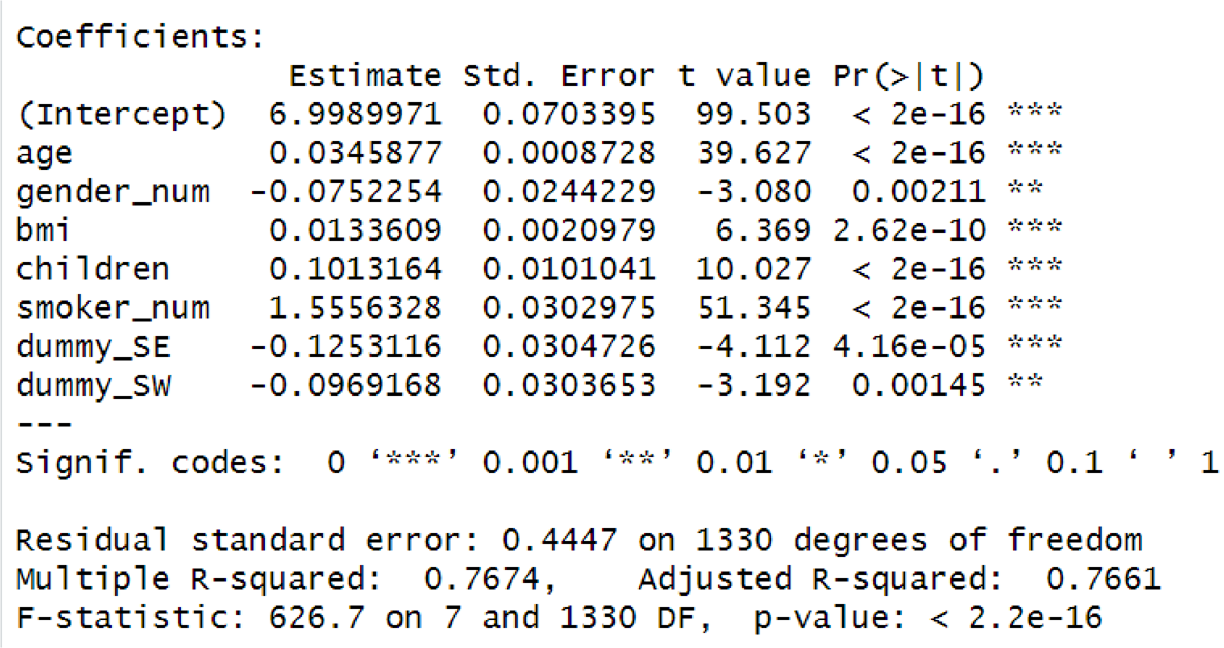
```
Residuals:
     Min       1Q   Median       3Q      Max 
-1.07125 -0.19783 -0.04891  0.06604  2.16655 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  7.0307859  0.0723992  97.111  < 2e-16 ***
age          0.0345816  0.0008721  39.654  < 2e-16 ***
gender_num  -0.0754109  0.0244017  -3.090 0.002040 ** 
bmi          0.0133658  0.0020960   6.377 2.49e-10 ***
children     0.1018651  0.0100997  10.086  < 2e-16 ***
smoker_num   1.5542783  0.0302800  51.330  < 2e-16 ***
dummy_NW    -0.0637805  0.0349064  -1.827 0.067896 .  
dummy_SE    -0.1571654  0.0350837  -4.480 8.12e-06 ***
dummy_SW    -0.1289048  0.0350274  -3.680 0.000242 ***
---
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:  0.7665 
F-statistic: 549.7 on 8 and 1329 DF,  p-value: < 2.2e-16
```

Fig:12

The F- statistics: 549.7 and its p-value is 2.2e^(-16) which is less than 0.05 significant level that means F-statistic is significant and there are at least one the beta values mentioned in ‘Coefficients’ is not zero.

Adjusted R-squared is 0.7665 and R-squared is 0.7679. These two values are approximately the same, hence we are confident that there is no issue of multicollinearity with this model.

However, t- value of dummy_NW is 0.068 more than 0.05 there it is not significant therefore at 0.05, this predictor variable needs to be removed and re-reun the regression.



```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  6.9989971  0.0703395  99.503  < 2e-16 ***
age          0.0345877  0.0008728  39.627  < 2e-16 ***
gender_num  -0.0752254  0.0244229  -3.080  0.00211 ** 
bmi          0.0133609  0.0020979   6.369 2.62e-10 ***
children     0.1013164  0.0101041  10.027  < 2e-16 ***
smoker_num   1.5556328  0.0302975  51.345  < 2e-16 ***
dummy_SE    -0.1253116  0.0304726  -4.112 4.16e-05 ***
dummy_SW    -0.0969168  0.0303653  -3.192  0.00145 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4447 on 1330 degrees of freedom
Multiple R-squared:  0.7674,	Adjusted R-squared:  0.7661 
F-statistic: 626.7 on 7 and 1330 DF,  p-value: < 2.2e-16
```

Fig:13

The F- statistics: 626.7 and its p-value is 2.2e^(-16) which is less than 0.05 significant level that means F-statistic is significant and there are at least one the beta values mentioned in ‘Coefficients’ is not zero.

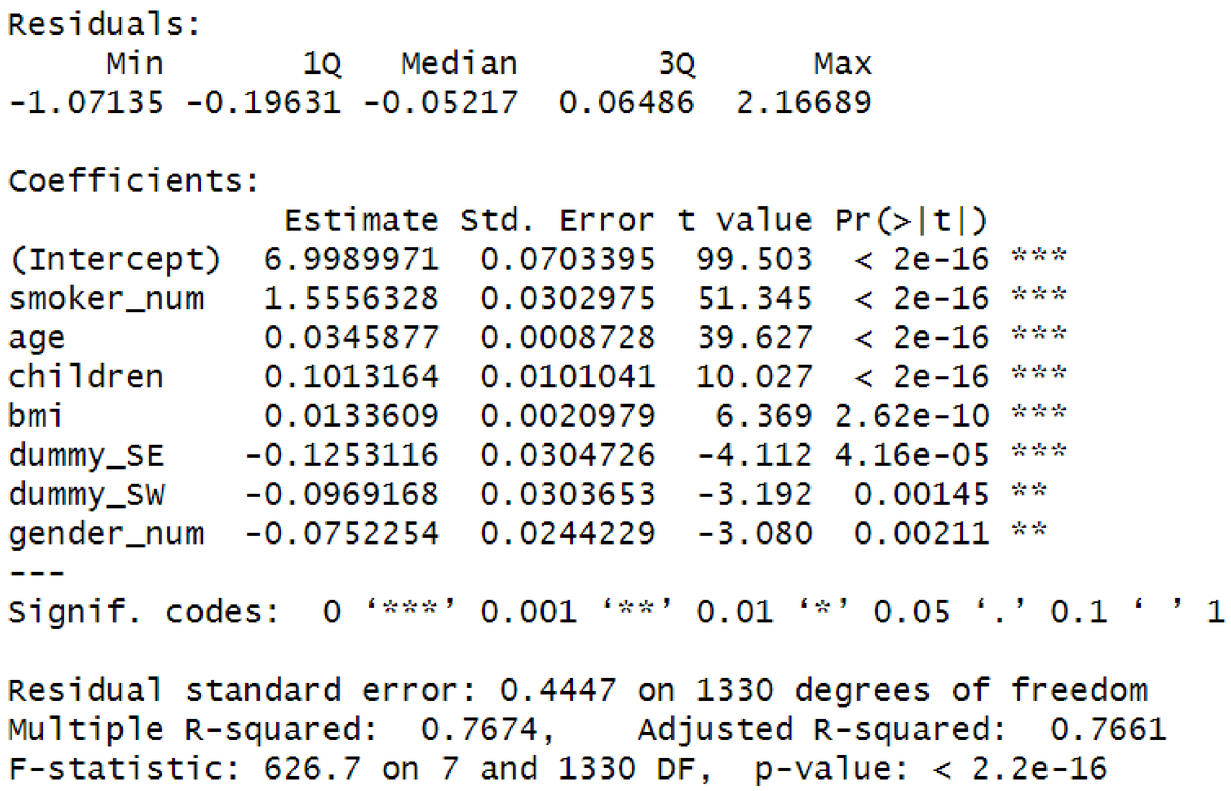
R-squared is 0.7674 and R-squared is 0.7661 These two values are approximately the same, hence we are confident that there is no issue of multicollinearity with this model.

b.

Running the ‘**forward’** regression model because of the following two reasons:

1. Since forward regression starts with choosing one subset of predictors at a time therefore, it requires less processor to compute.
2. Since we have already checked for the multicollinearity before computing forward regression, therefore, we are sure that forward regression will not face the issue of multicollinearity.

Summary of the Forward Regression model



```
Residuals:
     Min       1Q   Median       3Q      Max 
-1.07135 -0.19631 -0.05217  0.06486  2.16689 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  6.9989971  0.0703395  99.503  < 2e-16 ***
smoker_num   1.5556328  0.0302975  51.345  < 2e-16 ***
age          0.0345877  0.0008728  39.627  < 2e-16 ***
children     0.1013164  0.0101041  10.027  < 2e-16 ***
bmi          0.0133609  0.0020979   6.369 2.62e-10 ***
dummy_SE    -0.1253116  0.0304726  -4.112 4.16e-05 ***
dummy_SW    -0.0969168  0.0303653  -3.192  0.00145 ** 
gender_num  -0.0752254  0.0244229  -3.080  0.00211 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4447 on 1330 degrees of freedom
Multiple R-squared:  0.7674,	Adjusted R-squared:  0.7661 
F-statistic: 626.7 on 7 and 1330 DF,  p-value: < 2.2e-16
```

Fig:14

Model created by ‘forward regression’ (fig:14): The F- statistics: 626.7 and its p-value is 2.2e^(-16) which is less than 0.05 significant level that means F-statistic is significant and there are at least one the beta values mentioned in ‘Coefficients’ is not zero.

‘R-squared’ is 0.7674 and ‘Adjusted R-squared’ is 0.7661 which is approximately the same.

Global F-Test for overall model adequacy

The R- square of the model is 0.7674
The Adj R- square of the model is 0.7661
The F- statistics is 626.7
p-value is <2.2e-16

Ho: βj = 0
Ha: βj ≠ 0
F - value 626.7
P - value <2.2e-16
Conclusion:

We reject the null hupothesis as any one of the β is no zero therefore, ln_expenses can be explained by the predictors (smoker_num, age, children, bmi, dummy_SE, dummy_SW, and gender_num) in the model that has a significant effect on ln_Charge.

Predictors smofkr_num, age, children, bmi, dummy_SE, dummy_SW, and gender_num have coefficients that are significantly different from zero at the .05 level because there individual p-value is less than 0.05.

Forward Regression Model

Expenses_ln = 6.999 + 1.556*smoker_num + 0.0346*age + 0.101*children + 0.014*bmi – 0.125*dummy_SE – 0.097*dummy_SW - 0.075*gender_num

where assuming :

expenses are in dollars,
age is in years
gender_num is 1 when sex= female and 0 when sex= male,
bmi is body mass index and it has 1 unit = 1(kg / m ^ 2).
children is the number of children
smoker_num = 1 when the smoker is yes and 0 when the smoker is no,
dummy_SE = 1 when the region is "southeast" and 0 otherwise,
dumy_SW = 1 when the region is "southwest" and 0 otherwise

Interpretation

Interpretation of Smoker: The insurance expenses increases by [exp(1.556)-1]*100 = **$364.92** when a person goes from non-smoking to smoking, keeping all other independent variables as constant.
Interpretation of Age: The insurance expenses increases by [exp(0.0346)-1]*100 = **$3.52** when a person's age increase by one year, while all other independent variables were constant.

Interpretation of Children: The insurance expenses increases by [exp(0.101)-1]*100 = **$3.52** when the number of children increases by one, while all other independent variables were constant.

Interpretation of BMI: : If BMI of a person is increase by 1(kg/m^2), then insurance charges increases by [exp(0.014)-1]*100 = **$1.41** while other independent variables are constant.

Interpretation of Region: A person living in the Southeast region will have [exp(0.125)-1]*100 = **$13.31** fewer insurance expenses as compared to the person living in the Northeast region, keeping all other independent variables as constant. Likewise, a person living in the Southwest region will have [exp(0.097)-1]*100 = $**10.10** fewer insurance as compared to the person living in the Northeast region, keeping all other independent variables as constant.

Interpretation of sex: A female will have [exp(**0.075**)-1]*100 = **$7.79** lesser insurance expenses as compared to the male while other independent variables are constant

Application

How smoking effects, the insurance charge?

The insurance charges increase by [exp(1.556)-1]*100 = **$364.92** when a person goes from non-smoking to smoking, keeping all other independent variables as constant.

Calculating the predicted insurance charges for a non-smoker male beneficiary who is 37 years old with 2 children living in northeast region has 29.83 kg/m^2 BMI.

Expenses_ln = 6.999 + 1.556*smoker_num + 0.0346*age + 0.101*children + 0.014*bmi – 0.125*dummy_SE – 0.097*dummy_SW - 0.075*gender_num

Expenses_ln = 6.999 + 1.556*0 + 0.0346*37 + 0.101*2 + 0.014*29.83 – 0.125*0 – 0.097*0 - 0.075*0 =8.5

Visualization:

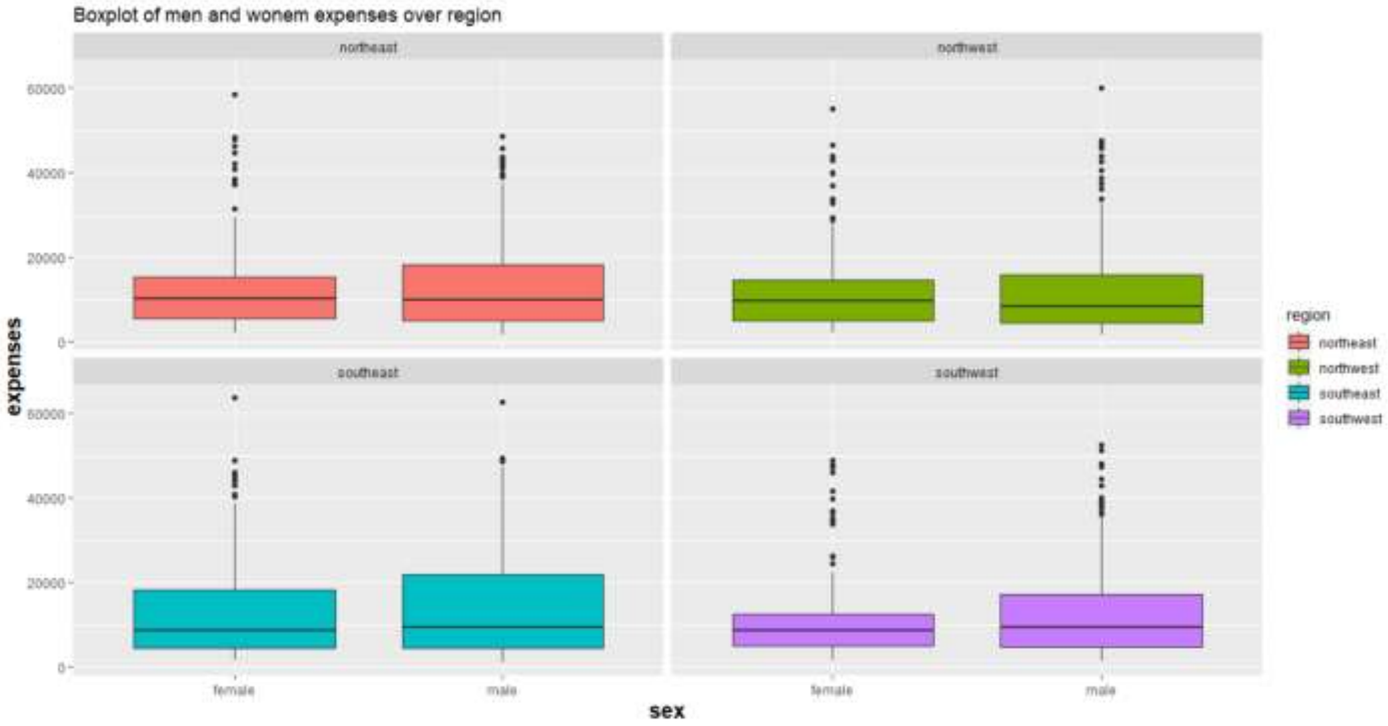


Fig:15

We can see in fig:15 that in general people from the Northeast and Southeast regions, especially southeast region pays more on insurance than people from the other regions. When we separate male and female distribution, we see males pay more as compare to the female. This ignite concerns for the health of males or may be the income of the male is more than female due to which expenses of males are more as compare to female. Therefore, in both the scenarios we need to have further investigations.

Second, thing that stands out is the distribution of expenses for both male and female of region southeast are way more right skewed as compare to other regions. This also needs further investigations.