Stat 615 Final_Project

Dhruv Jain & mekdim

2023-04-20

ACKNOWLEDGEMENTS

"I am not what happened to me, I am what I choose to become" by Christopher Gardner, The Pursuit of Happiness.

It is always a pleasure to remember the fine people who guided us in the Regression program. We received to uphold our practical and theoretical skills during the respective session. Firstly, we would like to thank **Pro. James C. Dickens** and secondly, we want to thank our family & friends for their love, motivation, and support during this semester in American university. Thanks for all the ideas, opinions, knowledge, and suggestions given to us to help us complete this report. We are very thankful to American University for giving us the opportunity to pursue this project.

Title Page with Executive Summary

Title: Estimating Medical Cost via random forest.

Type of analysis: Application analysis

Table 1:

Name	course
Dhruv Jain	STAT -615
Mekdim Ashebo	STAT -615

```
# calling all the libraries used in the code book
# Install this for latex pdf output.
#tinytex::install_tinytex()
library(olsrr)
library(tidyverse)
library(dbplyr)
library(dplyr)
library(Matrix)
library(MASS)
library(ggplot2)
library(tibble)
library(data.table)
library(ggmosaic)
library(ggforce)
library(ggmap)
library(ggthemes)
library(purrr)
library(keep)
library(readr)
library(gridExtra)
library(randomForest)
library(corrplot)
library(PerformanceAnalytics)
```

1 All About Data Set

1.1 Data Set Source

Summary: The type of analysis is **application based**. This data set is collected from different open-source databases manually. offer a preliminary description of the data set. For example, indicate the size of the data source, describe the variables, and include any other data profile information that would be of interest.

Data source: **website name**:Kaggle. Website link

Overall approach is Cleaning, analyzing, again cleaning, answering the five questions, playing with data to deliver more better output and finally graphical representation. **Defining** the issues and trying to resolve that by presenting the medical cost data set . **Measure** overall what can be done. **Analyze** the data to use for future capability. **Improving** You will use information gathered in the previous phases to design and implement improvements in processing with consistency. Overall Approach to this question is using various tools and coding sets with providing the statistical data with convincing evidence.

Clean –> Design –> Plot —> Types Regression model –> hypothesis –> statistical findings –> Random forest –> conclusion

Recommendations: Using R-studio markdown one can achieve high graphical results with better quality and one can also conclude that big data can be easily handled on this platform. Using this platform, we will continue to evaluate the medical cost for personal data using this tool. The issue is that what can be done with the data can we provided some explanation to justify the results. We would like to address those five questions with different visual ideas.

IMPORTANT KEYWORDS:

- charges
- Smoker
- sex
- BMI (body mass index)

- P-value
- Random forest

Columns Description

age: age of primary beneficiary **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 ˆ 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

We had to randomly sample 300 rows from our original data. We then saved these 300 rows into csv file so that we can import them later. going forward We will take that csv file. (which has only be run once)

```
# The preliminary steps we did
#insurance <- read_csv('Downloads/insurance.csv')
#insurance_300 <- sample_n(insurance, 300)
#write.csv(insurance_300 , file = "Desktop/insurance_300.csv")
```

```
insurance_new <- read_csv("insurance_300.csv",
col_types = cols(
  age = col_double(),
  sex = col_character(),
  bmi = col_double(),
  children = col_double(),
  smoker = col_character(),
  region = col_character(),
  charges = col_double()
))
```

- The data contains 300 rows and 7 columns. This project is about determining the factors that affect medical costs billed by health insurance.

- The independent variables include three categorical variables and three quantitative variables.Sex (male/female), region(Northeast, northwest etc), and smoker(whether a person smokes or not) are the categorical variables. While the quantitative variables include the BMI index, the age and the number of children the person have.

```
head(insurance_new,8)
```

```
## # A tibble: 8 x 7
##     age sex      bmi children smoker region    charges
##   <dbl> <chr>  <dbl>    <dbl> <chr>  <chr>       <dbl>
## 1    63 female  25.1        0 no     northwest  14255.
## 2    18 male    38.2        0 yes    southeast  36308.
## 3    48 male    29.6        0 no     southwest  21232.
## 4    46 female  33.4        1 no     southeast   8241.
## 5    52 male    30.2        1 no     southwest   9725.
## 6    36 female  19.9        0 no     northeast   5458.
## 7    19 male    20.9        1 no     southwest   1832.
## 8    48 male    36.7        1 no     northwest  28469.
```

```
nrow(insurance_new)
```

```
## [1] 300
```

```
ncol(insurance_new)
```

```
## [1] 7
```

```
colnames(insurance_new)
```

```
## [1] "age"      "sex"      "bmi"      "children" "smoker"   "region"   "charges"
```

- Let us quickly investigate the summary of our dependent variable. The median insurance charge is around 10097 and the mean of 13283. The standard deviation is 11399.

```
summary(insurance_new)
```

```
##       age            sex                 bmi           children   
##  Min.   :18.00   Length:300         Min.   :17.29   Min.   :0.00  
##  1st Qu.:27.00   Class :character   1st Qu.:25.25   1st Qu.:0.00  
##  Median :40.50   Mode  :character   Median :30.01   Median :1.00  
##  Mean   :39.88                      Mean   :30.02   Mean   :1.02  
##  3rd Qu.:53.00                      3rd Qu.:34.20   3rd Qu.:2.00  
##  Max.   :64.00                      Max.   :46.75   Max.   :5.00  
##     smoker             region             charges     
##  Length:300         Length:300         Min.   : 1136  
##  Class :character   Class :character   1st Qu.: 5134  
##  Mode  :character   Mode  :character   Median :10097  
##                                        Mean   :13283  
##                                        3rd Qu.:17154  
##                                        Max.   :51195  
```

- Type of columns used in data frame (double, charterer)

```
str(insurance_new)
```

```
## spc_tbl_ [300 x 7] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ age     : num [1:300] 63 18 48 46 52 36 19 48 19 19 ...
##  $ sex     : chr [1:300] "female" "male" "male" "female" ...
##  $ bmi     : num [1:300] 25.1 38.2 29.6 33.4 30.2 ...
##  $ children: num [1:300] 0 0 0 1 1 0 1 1 0 0 ...
##  $ smoker  : chr [1:300] "no" "yes" "no" "no" ...
##  $ region  : chr [1:300] "northwest" "southeast" "southwest" "southeast" ...
##  $ charges : num [1:300] 14255 36308 21232 8241 9725 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   age = col_double(),
##   ..   sex = col_character(),
##   ..   bmi = col_double(),
##   ..   children = col_double(),
##   ..   smoker = col_character(),
##   ..   region = col_character(),
##   ..   charges = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

1.2 Cleaning the data and type of columns

```
# calculating NA/missing data in columns
colSums(is.na(insurance_new))
```

```
##      age      sex      bmi children   smoker   region  charges
##        0        0        0        0        0        0        0
```

```
# converting to factor variable
insurance_new$sex = as.factor(insurance_new$sex)
insurance_new$smoker = as.factor(insurance_new$smoker)
# how many unique values
unique(insurance_new$sex)
```

```
## [1] female male
## Levels: female male
```

```
unique(insurance_new$children)
```

```
## [1] 0 1 2 5 3 4
```

```
unique(insurance_new$smoker)
```

```
## [1] no  yes
## Levels: no yes
```

```
unique(insurance_new$region)
```

```
## [1] "northwest" "southeast" "southwest" "northeast"
```

```
# Check levels of smoker variable
table(insurance_new$smoker)
```

```
## 
##  no yes 
## 239  61
```

```
# Check levels of region variable
table(insurance_new$region)
```

```
## 
## northeast northwest southeast southwest 
##        67        77        76        80
```

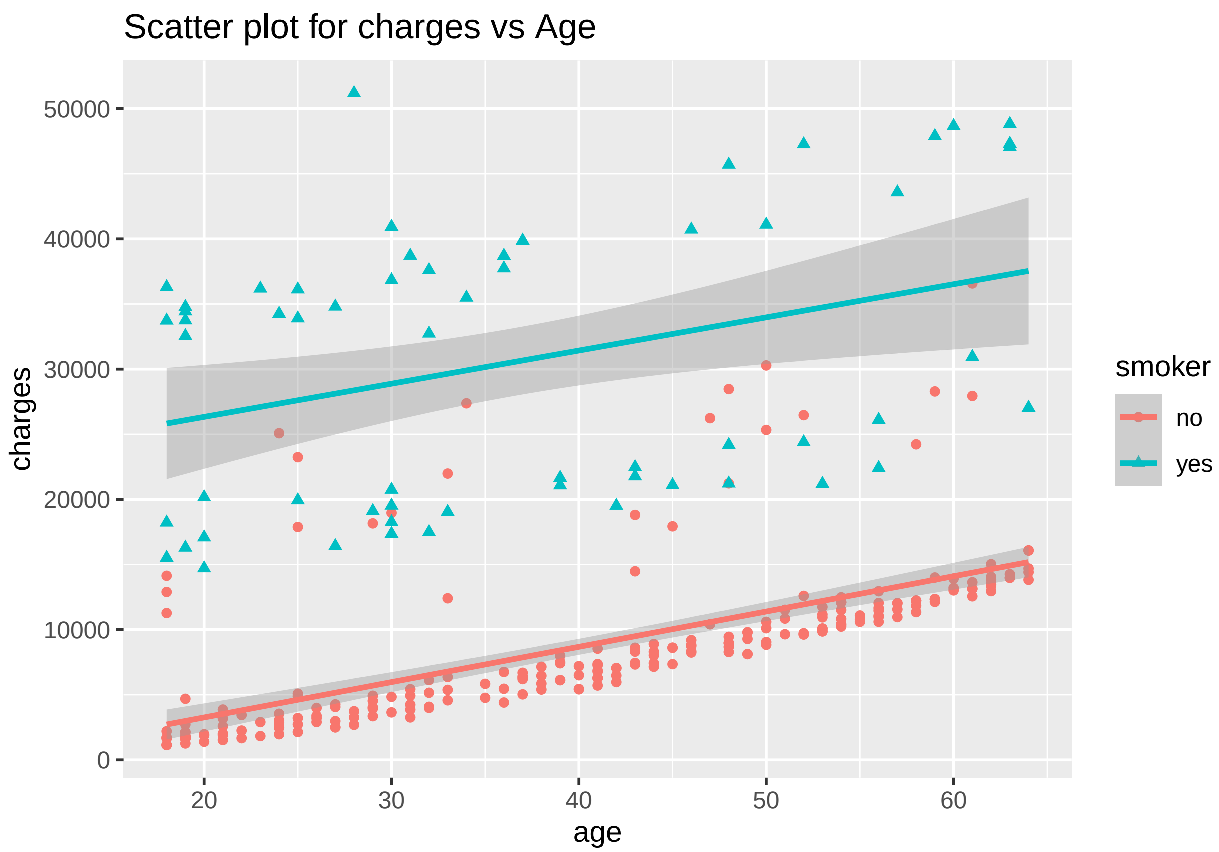
```
# Check levels of sex variable
table(insurance_new$sex)
```

```
## 
## female   male 
##    134    166
```

1.3 Visualization

1.3.a Does age affect medical charges for smoker?

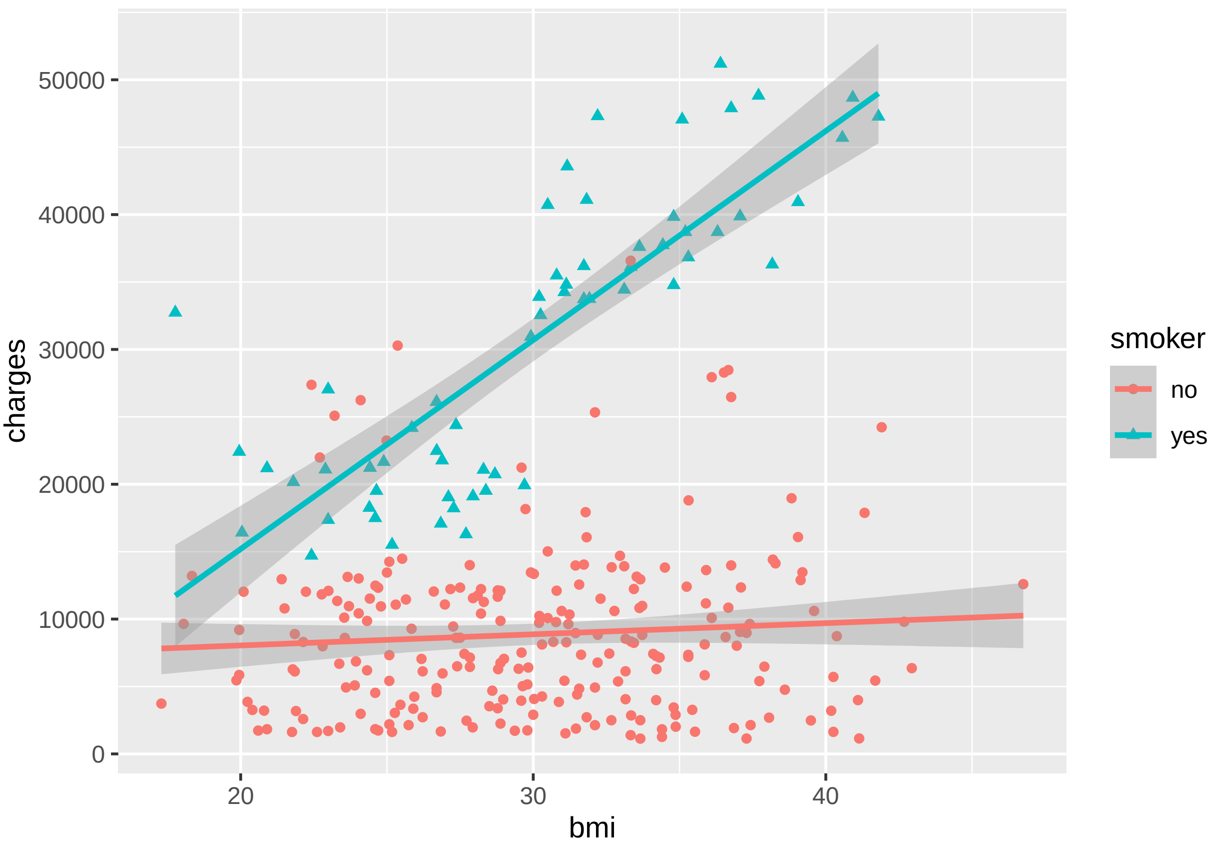
```
## ‘geom_smooth()‘ using formula = ’y ~ x’
```



- Yes the charges are increased as we increase the number of age. Now the fun part is if a person smokes he/she is paying more charges on medical then the person not smoking.

1.3.b Does Body mass index (BMI) affect medical charges for smoker?

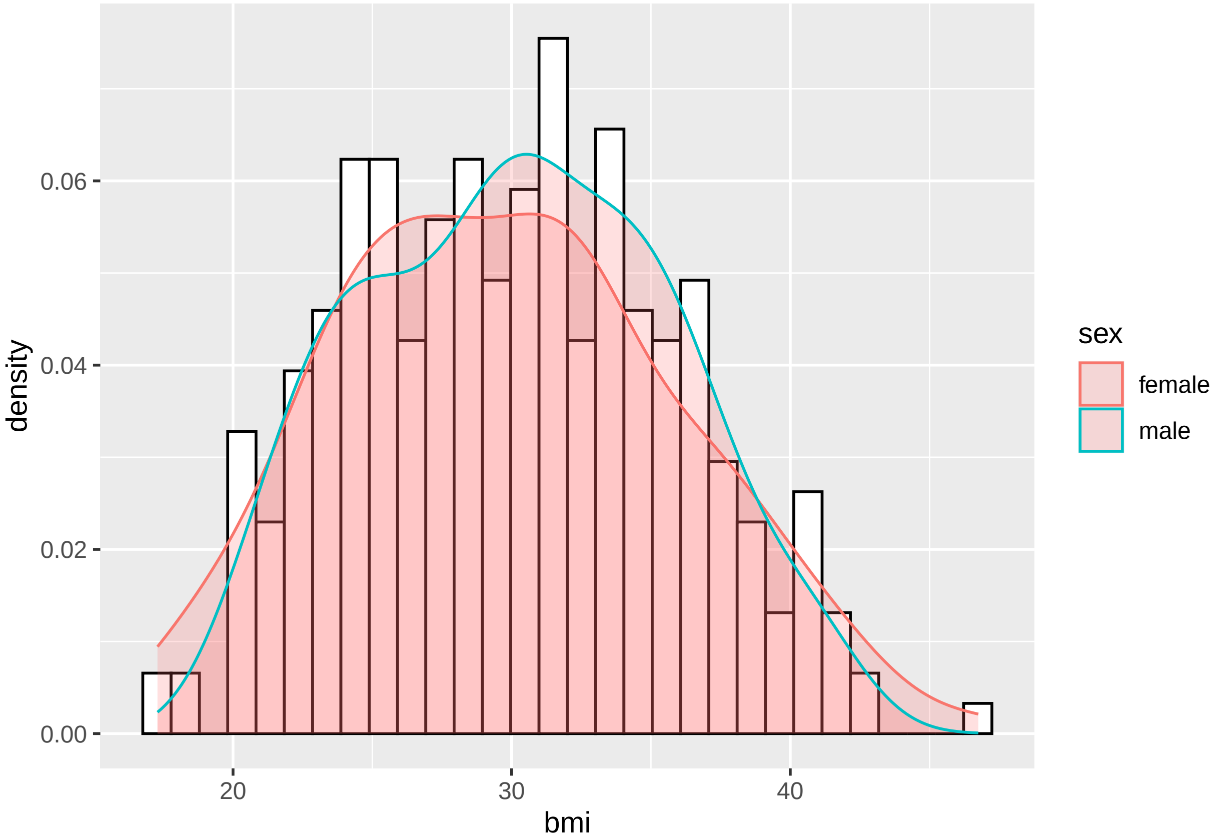
```
## ‘geom_smooth()‘ using formula = ’y ~ x’
```



- One can clearly observe that smoking affect in BMI and increased with the medical expenses.

1.3.c Histogram for density graph for Body mass index (BMI)

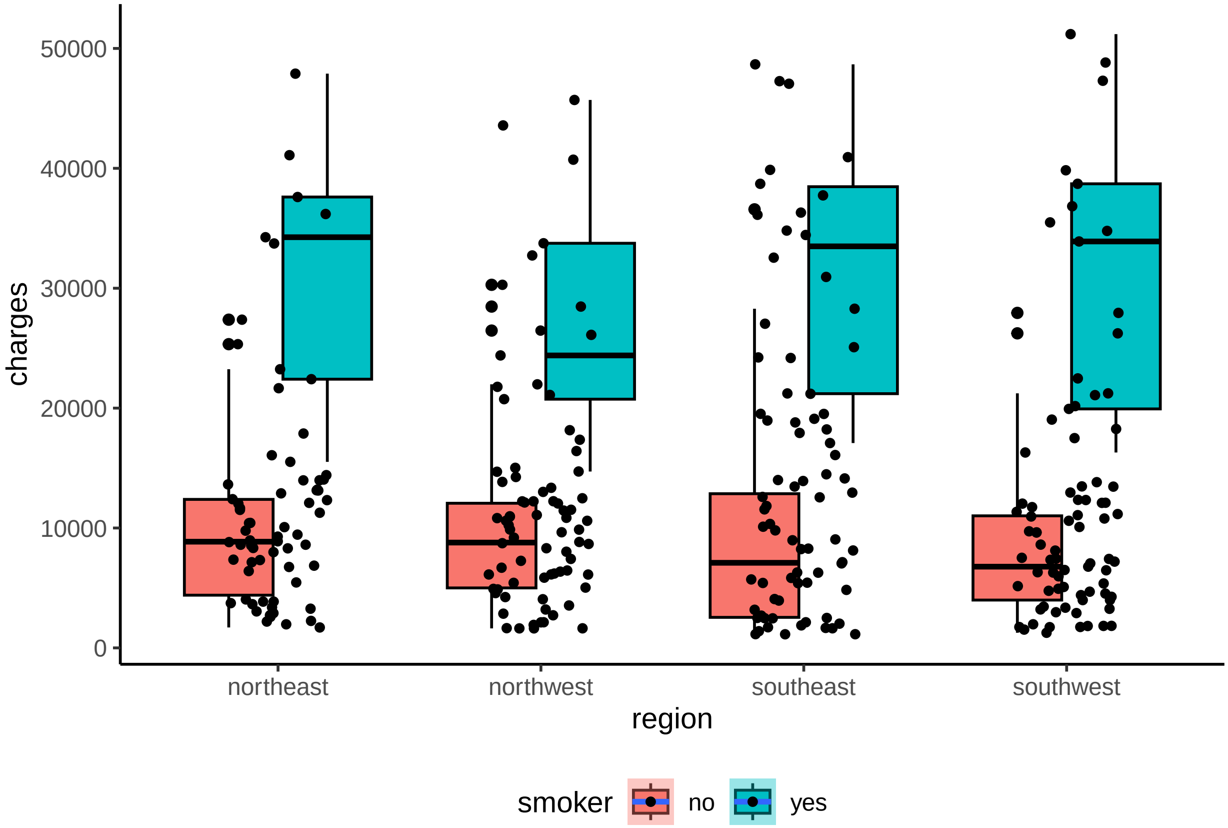
```
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
```



- The graph is density vs body mass index for male and female counts in the data set.

1.3.d High no. of smokers from which region ?

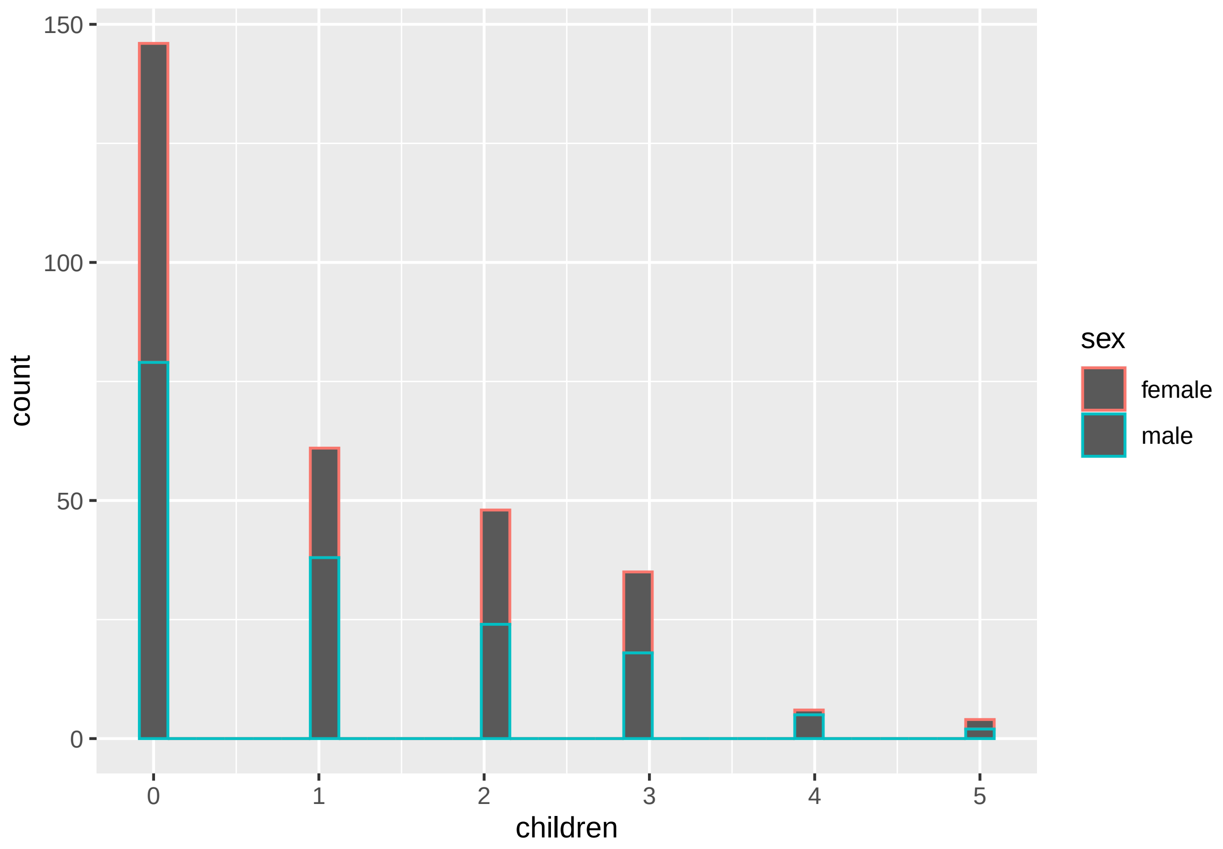
```
## ‘geom_smooth()‘ using formula = ’y ~ x’
```



- comparing that does region of living affects the smokers or not. Now, looking at graph one cans say that the region does affect the charges on medical insurance.

1.3.e

```
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
```



1.4 No. of counts

```
# Female
insurance_new%>%
  filter(sex == "female")%>%
  count(sex,children,smoker,region)%>%
  arrange(sex, smoker) -> counts

head(counts,n=5)
```

```
## # A tibble: 5 x 5
##   sex    children smoker region        n
##   <fct>     <dbl> <fct>  <chr>     <int>
## 1 female        0 no     northeast    14
## 2 female        0 no     northwest    17
## 3 female        0 no     southeast    14
## 4 female        0 no     southwest    13
## 5 female        1 no     northeast     5
```

```
# Male
insurance_new%>%
  filter(sex == "male")%>%
  count(sex,children,smoker,region)%>%
  arrange(sex, smoker) -> counts_male

head(counts_male,n=5)
```

```
## # A tibble: 5 x 5
##   sex   children smoker region        n
##   <fct>    <dbl> <fct>  <chr>     <int>
## 1 male         0 no     northeast    17
## 2 male         0 no     northwest    15
## 3 male         0 no     southeast    14
## 4 male         0 no     southwest    14
## 5 male         1 no     northeast     6
```

2 QQ & Multicollinearity plot description

2.1 Multicollinearity plot

- The response variable is not dependent on explanatory variable interns of multicollinearity.
- The highest correlation is between charges and age with only 0.24. But if we exclude charges since charges is dependent variable, the highest correlation among the independent variables.
- The age with bmi with only 0.04 which is nearly 0. So there exists no colinearity among the independent variables. This suggests that each of the variables might be useful if they are included in the regression model as they dont have any correlation with each other.

```
numeric_insurance <- cor(insurance_new[,c("bmi", "children", "age", "charges")])

numeric_insurance
```

```
##                  bmi    children        age   charges
## bmi       1.00000000 -0.01371482 0.04733455 0.1785191
## children -0.01371482  1.00000000 0.03529611 0.0781793
## age       0.04733455  0.03529611 1.00000000 0.2461625
## charges   0.17851911  0.07817930 0.24616249 1.0000000
```

- we can also the scatter plots between the independent variables clearly there is no pattern that we can see verifying our output from the correlation matrix.
- One can say from the graph that the points are independently plotted and one cannot find any kind of pattern on left side of graph. On the other hand one can identify the

2.2 Normality plot

```
# Light tailed at the end
qqnorm(insurance_new$bmi)
```

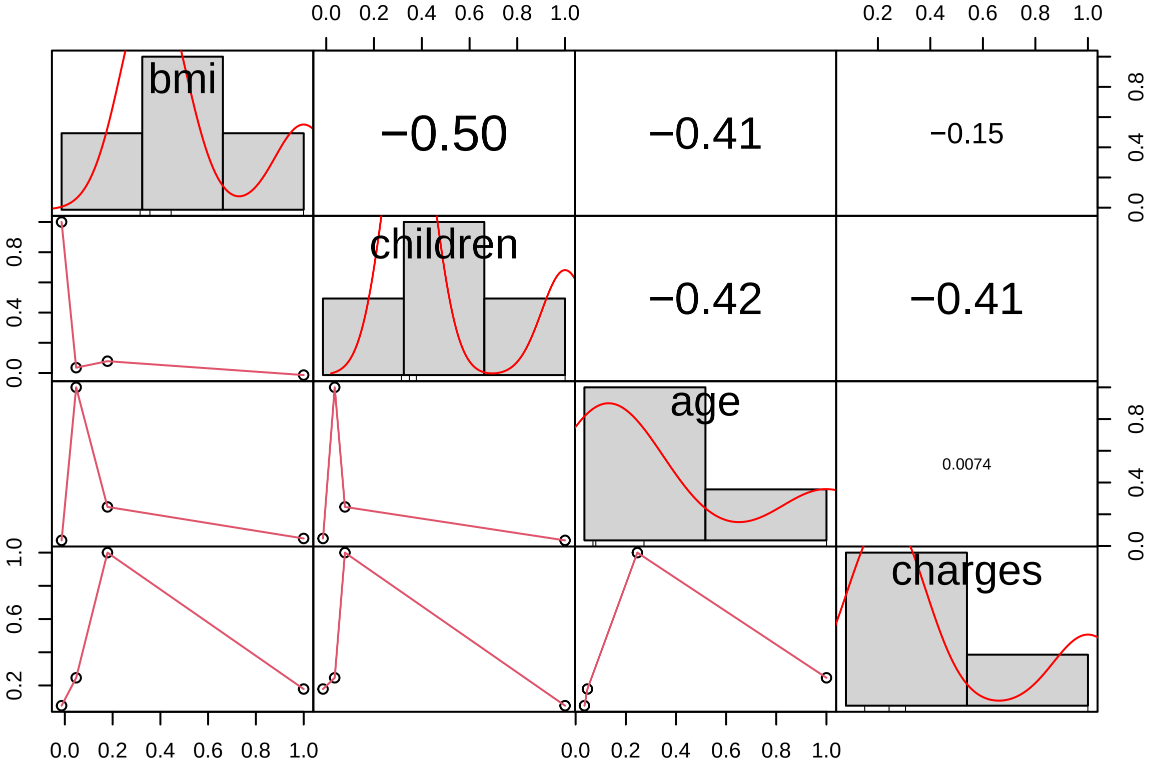
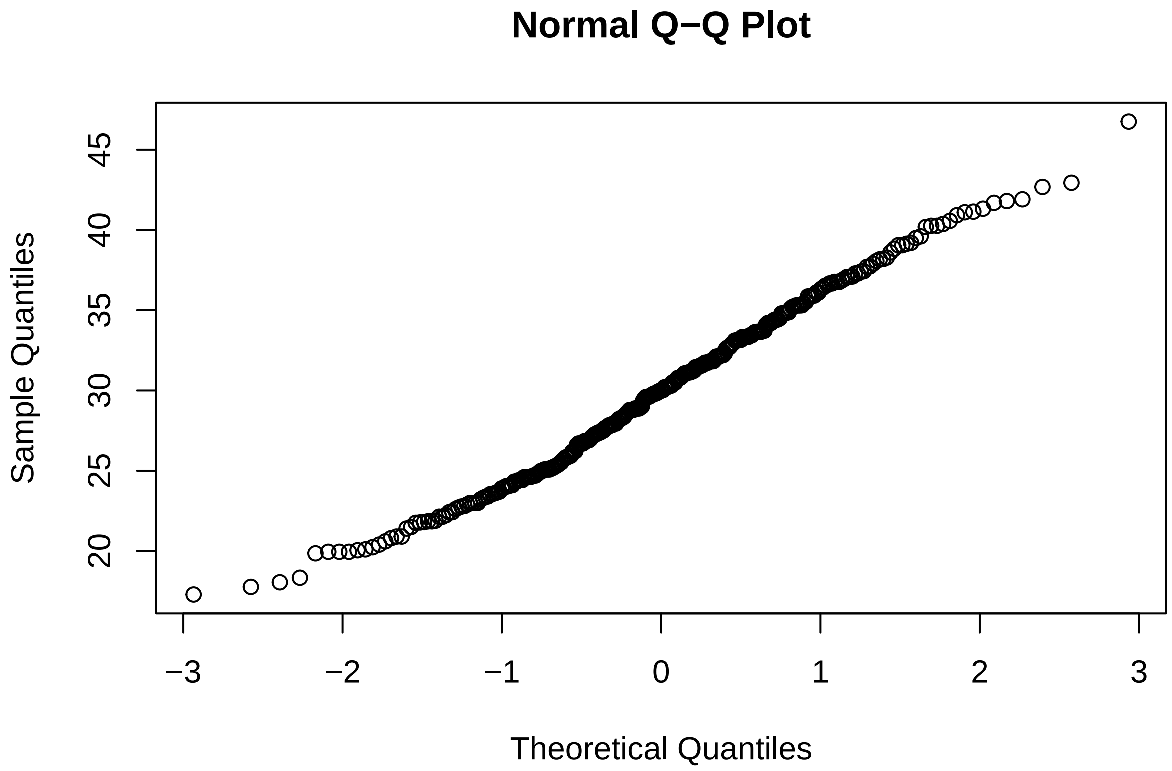
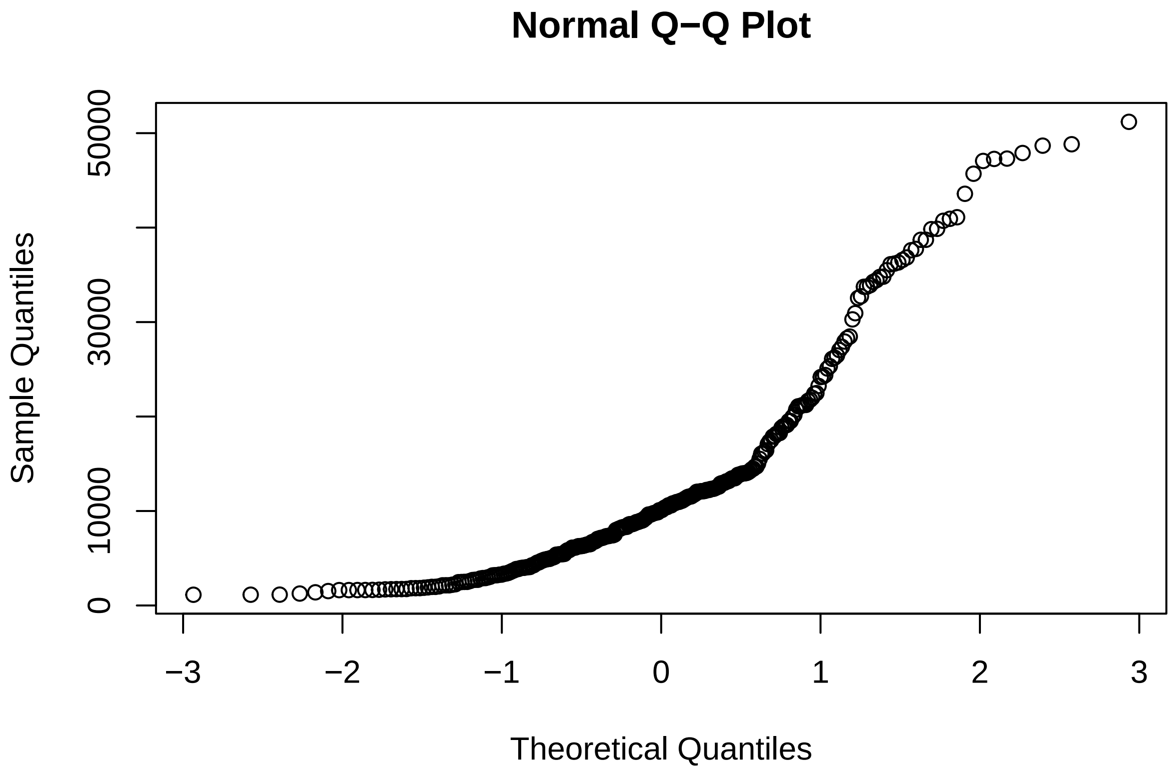


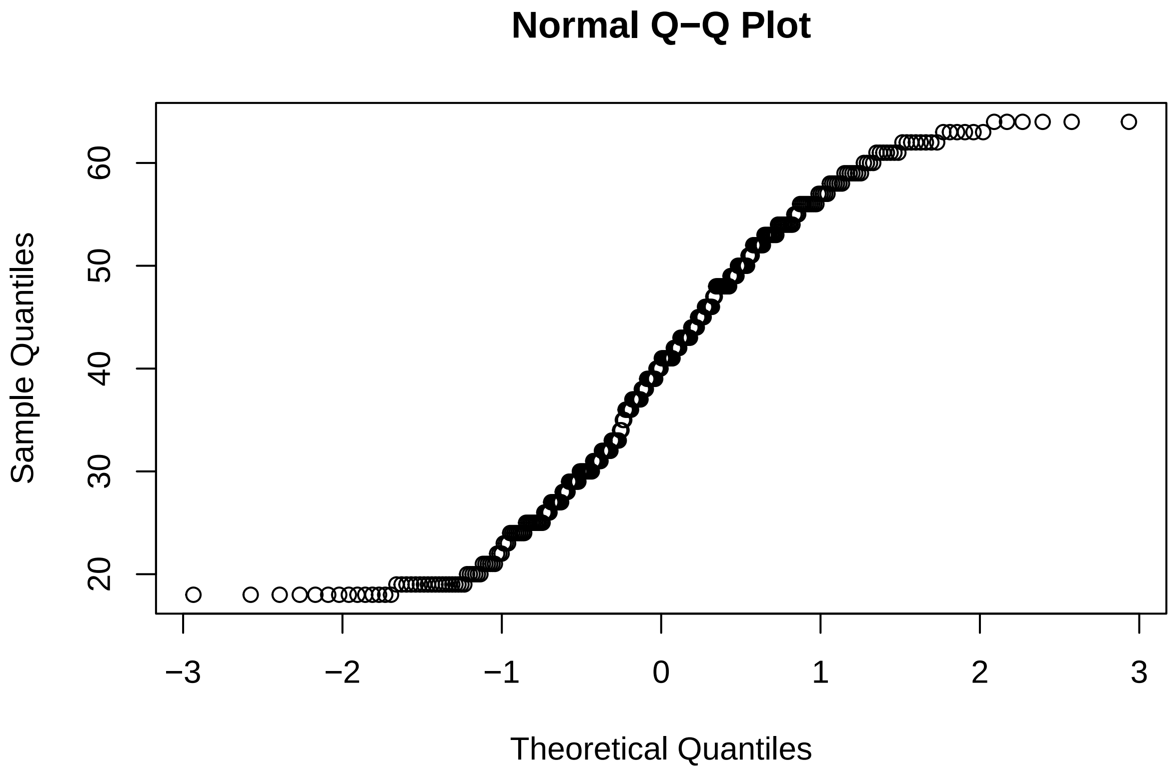
Figure 1: multicollinearity plot



```
# right skewed
qqnorm(insurance_new$charges)
```



```
# Heavy tailed at the end
qqnorm(insurance_new$age)
```



3 Full Regression Model

- full regression model including both categorical and quantitative variables

```
lm(charges ~ age + children + bmi + region + sex + smoker,data = insurance_new) -> x
x
```

```
##
## Call:
## lm(formula = charges ~ age + children + bmi + region + sex +
##     smoker, data = insurance_new)
##
## Coefficients:
##     (Intercept)              age         children              bmi
##        -12033.2            261.1            532.8            353.3
## regionnorthwest  regionsoutheast  regionsouthwest          sexmale
##         -1545.5          -1505.4          -1719.9            607.8
##       smokeryes
##         22876.4
```

4 Matrix

4.1 Fitted & Residual values using matrix for quantitative variables

```
Xm <-  model.matrix(~age + children + bmi   , data=insurance_new )
Ym <- as.matrix(insurance_new%>%dplyr::select(charges))
# Let's use R code to establish matrix X :
#-------------------------------------------------------------------
#  A = (X^T*X)^-1*X^TY
(solve(t(Xm)%*%Xm))%*%(t(Xm)%*%Ym)
```

```
##               charges
## (Intercept) -4825.6927
## age           185.6491
## children      669.1986
## bmi           333.8682
```

```
#-------------------------------------------------------------------
# fitted values
Xm%*%((solve(t(Xm)%*%Xm))%*%(t(Xm)%*%Ym)) -> fitted_values
#-------------------------------------------------------------------
# residual values
Ym-Xm%*%((solve(t(Xm)%*%Xm))%*%(t(Xm)%*%Ym)) -> residual_values
#-------------------------------------------------------------------
# producing the table of residula and fitted plot for the model.
matrix = data.frame(fitted_values, residual_values)
names(matrix)[1] <- "fitted_values"
names(matrix)[2] <- "residual_values"
head(matrix, 6)
```

```
##   fitted_values residual_values
## 1     15243.617       -989.0085
## 2     11259.742      25048.0563
## 3     13967.964       7264.2178
## 4     15547.919      -7307.3293
## 5     15580.081      -5855.5505
## 6      8486.629      -3028.5827
```

```
#-------------------------------------------------------------------
# Both fitted values and residual values match with matrix model
lm(charges ~ age + children + bmi, data= insurance_new) -> AB
#summary(AB)
#residuals(AB)
#fitted(AB)
```

4.2 Matrix method with both quantitative and qualitative variables(dummy variables included automatically)

```
# The results are the same using both the matrix method and lm method.

Xm <-  model.matrix(~age + children + bmi + region + sex+ smoker  , data=insurance_new )
Ym <- as.matrix(insurance_new%>%dplyr::select(charges))

# Let's use R code to establish matrix X :
#-------------------------------------------------------------------
#  A = (X^T*X)^-1*X^TY
(solve(t(Xm)%*%Xm))%*%(t(Xm)%*%Ym)
```

```
##                      charges
## (Intercept)     -12033.2491
## age                261.1443
## children           532.7552
## bmi                353.3493
## regionnorthwest  -1545.5302
## regionsoutheast  -1505.3521
## regionsouthwest  -1719.9011
## sexmale            607.7807
## smokeryes        22876.3789
```

```
#-------------------------------------------------------------------
# fitted values
Xm%*%((solve(t(Xm)%*%Xm))%*%(t(Xm)%*%Ym)) -> fitted_values
#-------------------------------------------------------------------
# residual values
Ym-Xm%*%((solve(t(Xm)%*%Xm))%*%(t(Xm)%*%Ym)) -> residual_values
#-------------------------------------------------------------------
# producing the table of residula and fitted plot for the model.
matrix = data.frame(fitted_values, residual_values)
names(matrix)[1] <- "fitted_values"
names(matrix)[2] <- "residual_values"
head(matrix,6)
```

```
##   fitted_values residual_values
## 1     11735.311        2519.298
## 2     28133.497        8174.301
## 3      9848.695       11383.488
## 4     10822.791       -2582.201
## 5     11638.037       -1913.507
## 6      4383.695        1074.351
```

```
#-------------------------------------------------------------------
# Both fitted values and residual values match with matrix model
lm(charges ~ age + children + bmi+ region + sex + smoker, data= insurance_new) -> AB
#summary(AB)
#residuals(AB)
#fitted(AB)
```

5 Analyze and Evaluate the full model

```
lm(charges ~ age + children + bmi+ region + sex + smoker, data= insurance_new) -> AB
summary(AB)
```

```
##
## Call:
## lm(formula = charges ~ age + children + bmi + region + sex +
##     smoker, data = insurance_new)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -10712  -3120  -1095   1496  24152
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -12033.25    2155.91  -5.582 5.46e-08 ***
## age                261.14      24.05  10.858  < 2e-16 ***
## children           532.76     279.27   1.908   0.0574 .
## bmi                353.35      62.03   5.696 2.99e-08 ***
## regionnorthwest  -1545.53     992.47  -1.557   0.1205
## regionsoutheast  -1505.35    1036.22  -1.453   0.1474
## regionsouthwest  -1719.90     990.21  -1.737   0.0835 .
## sexmale            607.78     694.59   0.875   0.3823
## smokeryes        22876.38     865.37  26.435  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5915 on 291 degrees of freedom
## Multiple R-squared:  0.7379, Adjusted R-squared:  0.7307
## F-statistic: 102.4 on 8 and 291 DF,  p-value: < 2.2e-16
```

5.1 Coefficients:

charges = -12033.25 + (age)*261.14 + (children)*532.76 + (bmi)*353.35 + (region_northwest)(-1545.53) + (region_southeast)*(-1505.35) + (regionsouthwest)*(-1719.90) + (sexmale)*(607.78) + (smokeryes)*(22876.38)

- The **Estimate column** lists the estimated coefficients of the predictor variables included in the model. For instance, the "age" coefficient has an estimated value of 261.14, which means that for every one-unit increase in age, the outcome variable (presumably a medical cost) is estimated to increase by $261.14, holding all other variables constant.

5.2 Standard Errors:

- The **Std Error** column lists the standard errors of the estimated coefficients. These are measures of the uncertainty or variability in the estimated coefficients. Smaller standard errors indicate more precise estimates.

5.3 T-values:

- The **t-value** column lists the t-statistics for each coefficient. These values represent the estimated coefficients divided by their standard errors. T-values are used to test the null hypothesis that the true coefficient is zero. Larger t-values indicate a stronger evidence against the null hypothesis.

5.4 P-values

- The pr column lists the p-values associated with the t-values. P-values represent the probability of observing the t-value or a more extreme value if the true coefficient is zero. Smaller p-values indicate stronger evidence against the null hypothesis.
- The significance codes provided in the table help to quickly identify significant coefficients; for instance, **age,bmi,smokers** represents p-value less than two decimal places and **children** represents p-value less than 0.05.One can say that **region** represents bigger p-value than significant level.

5.5 Residual Standard Error:

- The **Residual standard error** provides an estimate of the variability of the errors or unexplained variance in the model. It measures the average distance that the observed values fall from the predicted values.

5.6 R-squared:

- The "Multiple R-squared" and "Adjusted R-squared" measures how well the model fits the data. R-squared ranges from 0 to 1 and represents the proportion of variance in the outcome variable that is explained by the predictor variables. Adjusted R-squared is a corrected version of R-squared that takes into account the number of predictor variables in the model. In full model one can see 0.7379.

5.7 summary

- This data summary provides information on the estimated coefficients, their standard errors, t-values, and corresponding p-values, as well as model diagnostics such as residual standard error, multiple R-squared, adjusted R-squared, and F-statistic. These measures can be used to interpret the strength and significance of the relationship between the predictor variables and the outcome variable, as well as the overall goodness-of-fit of the model. Still working to find the best parameters to be included in the model.

6 Confidence intervals for all variables used in full model

- These intervals provide a range of plausible values for the true population coefficients based on the sample data. The first column, "2.5%", represents the lower bound of the interval, and the second column, "97.5%", represents the upper bound of the interval.
- The 95% confidence interval for the "age" variable is [213.81, 308.48]. This means that we are 95% confident that the true population coefficient for age falls within this range. Similarly, the confidence interval for the intercept is [-16276.39, -7790.11], which suggests that the expected value of the response variable (when all predictor variables are 0) is likely to be within this range.

```
confint(AB,level = 0.95)
```

```
##                        2.5 %     97.5 %
## (Intercept)     -16276.39288 -7790.1053
## age                213.80684   308.4817
## children           -16.88546  1082.3959
## bmi                231.26578   475.4327
## regionnorthwest  -3498.86525   407.8049
## regionsoutheast  -3544.78722   534.0830
## regionsouthwest  -3668.78747   228.9853
## sexmale           -759.27612  1974.8375
## smokeryes        21173.19777 24579.5601
```

7 Full and Reduced Model for transformed variables with interaction terms

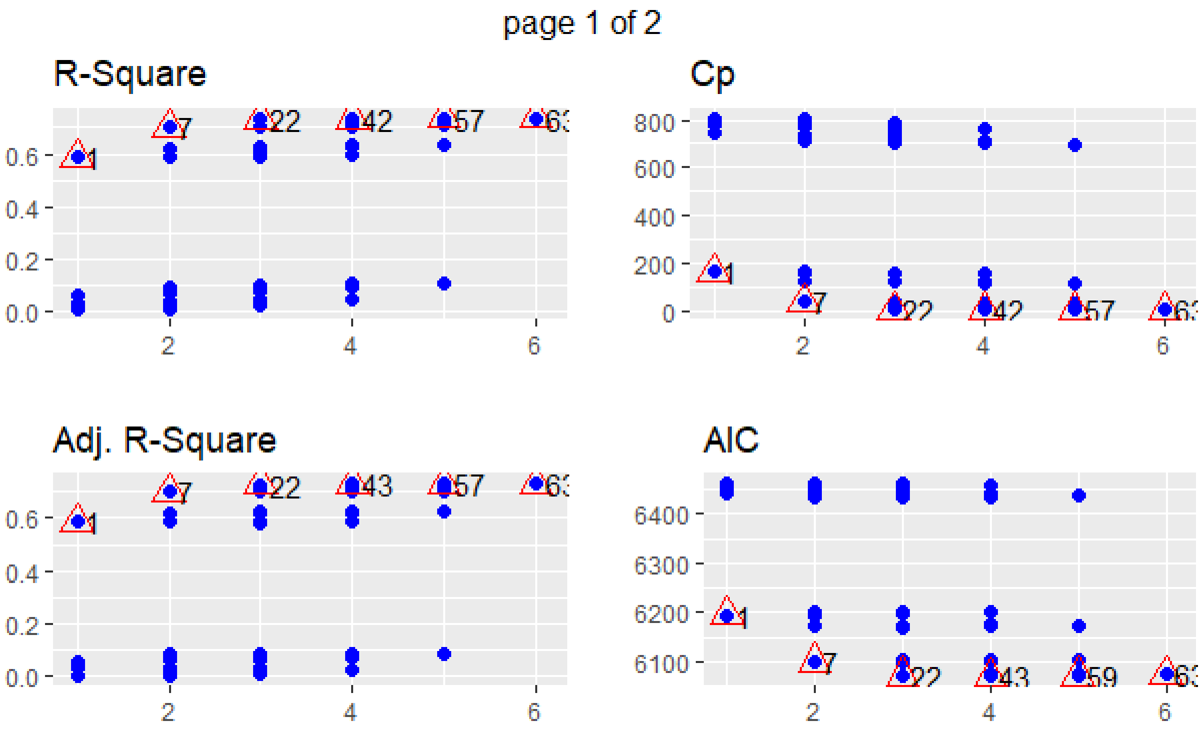
7.1 Evaluating Various regression models.

- Let us begin by using a multiple linear regression model that uses all the six variables. From the summary table we see that our R squared and Adjusted r square are around 0.73 and the residual standard error is 5915. The r squared value is high enough to be considered good but let us continue finding better fits.

```
ols_step_all_possible(AB) ->  allmodels
as_tibble(allmodels) -> allmodels_1
tail(allmodels_1,8)
```

```
## # A tibble: 8 x 14
##   mindex     n predict~1 rsquare   adjr predrsq     cp   aic  sbic   sbc    msep
##    <int> <int> <chr>       <dbl>  <dbl>   <dbl>  <dbl> <dbl> <dbl> <dbl>   <dbl>
## 1     56     4 children~  0.0470 0.0275 0.00150 768.   6457. 5593. 6486. 3.77e10
## 2     57     5 age chil~  0.737  0.731  0.721     3.77 6072. 5217. 6105. 1.04e10
## 3     58     5 age bmi ~  0.735  0.728  0.718     6.64 6075. 5220. 6108. 1.05e10
## 4     59     5 age chil~  0.735  0.730  0.722     6.70 6071. 5220. 6097. 1.05e10
## 5     60     5 age chil~  0.709  0.702  0.690    35.4  6103. 5247. 6136. 1.15e10
## 6     61     5 children~  0.632  0.623  0.609   121.   6173. 5315. 6207. 1.46e10
## 7     62     5 age chil~  0.109  0.0873 0.0600  702.   6439. 5574. 6472. 3.53e10
## 8     63     6 age chil~  0.738  0.731  0.720     5    6073. 5218. 6110. 1.04e10
## # ... with 3 more variables: fpe <dbl>, apc <dbl>, hsp <dbl>, and abbreviated
## #   variable name 1: predictors
```

```
# the best model is represented by including all age, children,bmi, region, smoker full model.
knitr::include_graphics("./1.png")
```



```
# plot(allmodels)
```

7.2 Finding Reduced and full model using interaction terms

- Let us see how r will create dummy variables for us using Region variable. R will convert the categorical variables for us We can see R will do this automatically for us creating these dummy variables. Northeast is essentially the control variable. When the three variables are 0, it means that region is northeast! No need to bother here as R does this for us.

```
contrasts(as.factor(insurance_new$region))
```

```
##           northwest southeast southwest
## northeast         0         0         0
## northwest         1         0         0
## southeast         0         1         0
## southwest         0         0         1
```

```
# Next step - Let us include all the interaction terms as well. Our residual standard error reduced. si

lm(charges~ age + children + bmi +
     region + sex + smoker + age:children + age:bmi + age:region + age:sex + age:smoker+
     children:bmi + children:region + children:sex + children:smoker
     + bmi:region + bmi:sex+ bmi:smoker + region:sex + region + smoker
   + sex:smoker, insurance_new) -> interactionModel
summary(interactionModel)
```

```
##
## Call:
## lm(formula = charges ~ age + children + bmi + region + sex +
##     smoker + age:children + age:bmi + age:region + age:sex +
##     age:smoker + children:bmi + children:region + children:sex +
##     children:smoker + bmi:region + bmi:sex + bmi:smoker + region:sex +
##     region + smoker + sex:smoker, data = insurance_new)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8196.3 -2253.3 -1060.8   274.9 20420.1
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               -2834.483   5950.715  -0.476 0.634228
## age                         146.188    117.074   1.249 0.212868
## children                    -52.254   1533.089  -0.034 0.972835
## bmi                          98.895    196.858   0.502 0.615821
## regionnorthwest            4940.926   5420.759   0.911 0.362857
## regionsoutheast            1791.263   5455.771   0.328 0.742922
## regionsouthwest             485.686   5074.751   0.096 0.923825
## sexmale                   -1228.344   3662.175  -0.335 0.737574
## smokeryes                -17140.502   4375.582  -3.917 0.000114 ***
## age:children                 25.113     18.782   1.337 0.182338
## age:bmi                       1.850      3.667   0.504 0.614324
## age:regionnorthwest          60.038     59.631   1.007 0.314922
## age:regionsoutheast          69.054     61.319   1.126 0.261106
## age:regionsouthwest          98.009     58.561   1.674 0.095368 .
## age:sexmale                 -22.156     41.292  -0.537 0.592007
## age:smokeryes               -57.725     53.308  -1.083 0.279845
## children:bmi                -13.009     43.378  -0.300 0.764478
## children:regionnorthwest    105.184    740.993   0.142 0.887226
## children:regionsoutheast   -318.240    763.888  -0.417 0.677299
## children:regionsouthwest   -233.540    712.265  -0.328 0.743255
## children:sexmale            553.118    487.168   1.135 0.257230
## children:smokeryes        -1218.854    638.980  -1.908 0.057521 .
## bmi:regionnorthwest        -260.593    159.656  -1.632 0.103803
## bmi:regionsoutheast        -120.067    154.970  -0.775 0.439153
## bmi:regionsouthwest        -161.455    156.511  -1.032 0.303190
## bmi:sexmale                 105.175    108.360   0.971 0.332616
## bmi:smokeryes              1459.039    134.821  10.822  < 2e-16 ***
## regionnorthwest:sexmale    -915.301   1733.496  -0.528 0.597929
## regionsoutheast:sexmale   -3533.510   1791.412  -1.972 0.049580 *
## regionsouthwest:sexmale   -2069.808   1698.494  -1.219 0.224059
## sexmale:smokeryes          -380.218   1604.912  -0.237 0.812908
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 4910 on 269 degrees of freedom
## Multiple R-squared:  0.8331, Adjusted R-squared:  0.8144
## F-statistic: 44.75 on 30 and 269 DF,  p-value: < 2.2e-16
```

```
# Let us also see if transforming our dependent variable might help. Our R squared increased slightly
# but not much
```

```r
lm(log(charges)~ age + children + bmi +
     region + sex + smoker + age:children + age:bmi + age:region + age:sex + age:smoker+
     children:bmi + children:region + children:sex + children:smoker
   + bmi:region + bmi:sex+ bmi:smoker + region:sex + region + smoker
   + sex:smoker, insurance_new) -> interactionModel
summary(interactionModel)
```

```
##
## Call:
## lm(formula = log(charges) ~ age + children + bmi + region + sex +
##     smoker + age:children + age:bmi + age:region + age:sex +
##     age:smoker + children:bmi + children:region + children:sex +
##     children:smoker + bmi:region + bmi:sex + bmi:smoker + region:sex +
##     region + smoker + sex:smoker, data = insurance_new)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.56846 -0.16881 -0.07365  0.03661  2.07636
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               6.5081702  0.4688489  13.881  < 2e-16 ***
## age                       0.0404944  0.0092241   4.390 1.63e-05 ***
## children                  0.2323820  0.1207900   1.924 0.055428 .
## bmi                       0.0373892  0.0155102   2.411 0.016596 *
## regionnorthwest           0.1858655  0.4270944   0.435 0.663776
## regionsoutheast           0.0443852  0.4298529   0.103 0.917836
## regionsouthwest          -0.1660037  0.3998328  -0.415 0.678339
## sexmale                  -0.3419357  0.2885379  -1.185 0.237038
## smokeryes                 1.5843978  0.3447462   4.596 6.63e-06 ***
## age:children             -0.0027142  0.0014798  -1.834 0.067741 .
## age:bmi                  -0.0003435  0.0002889  -1.189 0.235491
## age:regionnorthwest       0.0125558  0.0046982   2.672 0.007990 **
## age:regionsoutheast       0.0145449  0.0048312   3.011 0.002855 **
## age:regionsouthwest       0.0165099  0.0046140   3.578 0.000410 ***
## age:sexmale               0.0040348  0.0032533   1.240 0.215983
## age:smokeryes            -0.0353317  0.0042001  -8.412 2.40e-15 ***
## children:bmi             -0.0004008  0.0034177  -0.117 0.906735
## children:regionnorthwest -0.0022434  0.0583819  -0.038 0.969377
## children:regionsoutheast -0.0215457  0.0601857  -0.358 0.720634
## children:regionsouthwest -0.0263279  0.0561184  -0.469 0.639343
## children:sexmale          0.0645546  0.0383833   1.682 0.093760 .
## children:smokeryes       -0.1738442  0.0503443  -3.453 0.000644 ***
## bmi:regionnorthwest      -0.0287379  0.0125790  -2.285 0.023116 *
## bmi:regionsoutheast      -0.0213882  0.0122099  -1.752 0.080963 .
## bmi:regionsouthwest      -0.0213983  0.0123313  -1.735 0.083835 .
## bmi:sexmale               0.0049372  0.0085376   0.578 0.563549
## bmi:smokeryes             0.0495150  0.0106224   4.661 4.95e-06 ***
## regionnorthwest:sexmale   0.0331110  0.1365798   0.242 0.808632
## regionsoutheast:sexmale  -0.2677983  0.1411430  -1.897 0.058852 .
## regionsouthwest:sexmale  -0.1418041  0.1338221  -1.060 0.290256
## sexmale:smokeryes         0.0337446  0.1264488   0.267 0.789778
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3869 on 269 degrees of freedom
## Multiple R-squared:  0.8363, Adjusted R-squared:  0.818
## F-statistic: 45.81 on 30 and 269 DF,  p-value: < 2.2e-16
```

7.3 Let us include interaction terms with transformed variables

- Let us add more interaction terms that include transformed x variables our y variable has been transformed here as well. Our r squared increased again to 0.8373 and the residual error is now 4839. It is better slightly but there are a lot of variables which are not significant. For example, sex:smoke interaction variable's p value is 0.7829 which is significantly above 0.05. In the next step, let us remove all those variables that are not significant

```
lm(charges~ age + children + bmi +
     region + sex + smoker + age:children + age:bmi + age:region + age:sex + age:smoker+
      children:sex + children:smoker
   + bmi:region + bmi:sex+ bmi:smoker + region:sex + region + smoker
   + sex:smoker + log(bmi)+ bmi*bmi + log(age) + age*age + log(age)*log(bmi), insurance_new) -> interact
summary(interactionModel)
```

```
##
## Call:
## lm(formula = charges ~ age + children + bmi + region + sex +
##     smoker + age:children + age:bmi + age:region + age:sex +
##     age:smoker + children:sex + children:smoker + bmi:region +
##     bmi:sex + bmi:smoker + region:sex + region + smoker + sex:smoker +
##     log(bmi) + bmi * bmi + log(age) + age * age + log(age) *
##     log(bmi), data = insurance_new)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7396.2 -2274.3  -911.3   237.6 19910.0
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -496.094 225079.381  -0.002   0.9982
## age                        99.734    588.741   0.169   0.8656
## children                 -414.881    841.634  -0.493   0.6225
## bmi                      1080.305    931.509   1.160   0.2472
## regionnorthwest          5496.631   5208.089   1.055   0.2922
## regionsoutheast          6158.931   5612.118   1.097   0.2734
## regionsouthwest          1725.464   4974.472   0.347   0.7290
## sexmale                  -772.434   3586.073  -0.215   0.8296
## smokeryes              -17224.875   4309.003  -3.997 8.27e-05 ***
## log(bmi)                -5259.620  73591.622  -0.071   0.9431
## log(age)                22322.993  67679.455   0.330   0.7418
## age:children               22.167     17.580   1.261   0.2084
## age:bmi                     8.347     18.669   0.447   0.6551
## age:regionnorthwest        74.566     59.294   1.258   0.2096
## age:regionsoutheast        66.190     60.104   1.101   0.2718
## age:regionsouthwest        94.718     57.672   1.642   0.1017
## age:sexmale               -21.656     41.180  -0.526   0.5994
## age:smokeryes              -68.335     51.950  -1.315   0.1895    
## children:sexmale            691.627    473.110   1.462   0.1449    
## children:smokeryes        -1197.148    625.331  -1.914   0.0566 .  
## bmi:regionnorthwest        -301.338    155.290  -1.940   0.0534 .  
## bmi:regionsoutheast        -276.348    162.749  -1.698   0.0907 .  
## bmi:regionsouthwest        -201.368    155.430  -1.296   0.1962    
## bmi:sexmale                  91.781    105.495   0.870   0.3851    
## bmi:smokeryes              1473.976    132.003  11.166  < 2e-16 ***
## regionnorthwest:sexmale    -667.267   1702.405  -0.392   0.6954    
## regionsoutheast:sexmale   -3302.764   1780.210  -1.855   0.0646 .  
## regionsouthwest:sexmale   -2344.500   1670.199  -1.404   0.1615    
## sexmale:smokeryes          -430.457   1560.915  -0.276   0.7829    
## log(bmi):log(age)         -8223.943  19871.918  -0.414   0.6793    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 4839 on 270 degrees of freedom
## Multiple R-squared:  0.8373, Adjusted R-squared:  0.8198 
## F-statistic:  47.9 on 29 and 270 DF,  p-value: < 2.2e-16
```

7.4 Let Reduce the above models by eliminating some variables

- remove interaction Terms or the single variables whose p value is insignificant. For this case our starting model has a lot of variables so after removing many of the variables we will have a reduced model.
- Those variables whose p value is too high should be removed and a reduced model has to be produced.we run the reduced model below our r squared slightly decreased to 0.8332 and the adjusted r squared to 0.8244.
- It is very small change to our previous step so it is fine to take this. The p values are also significant so for that reason we chose this model. This could be one candidate model for our regression.

```
lm(charges~  children+bmi+smoker+age:children+age:region+
     age:smoker+children:smoker+bmi:region+bmi:smoker+smoker+log(bmi)+
     bmi*bmi + log(age),insurance_new) -> interactionModelSignificant
summary(interactionModelSignificant)
```

```
## 
## Call:
## lm(formula = charges ~ children + bmi + smoker + age:children + 
##     age:region + age:smoker + children:smoker + bmi:region + 
##     bmi:smoker + smoker + log(bmi) + bmi * bmi + log(age), data = insurance_new)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -7189.3 -2334.9 -1166.7   385.6 21603.4 
## 
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)    
## (Intercept)          55898.45   31237.64   1.789  0.07461 .  
## children              -319.02     745.49  -0.428  0.66902    
## bmi                   1285.83     454.43   2.830  0.00499 ** 
## smokeryes           -17169.12    4145.36  -4.142 4.55e-05 ***
## log(bmi)            -30700.93   13026.61  -2.357  0.01911 *  
## log(age)              5125.69    1406.95   3.643  0.00032 ***
## children:age            25.99      17.10   1.520  0.12966    
## age:regionnorthwest    136.02      51.05   2.664  0.00815 ** 
## age:regionsoutheast    122.86      51.49   2.386  0.01769 *  
## age:regionsouthwest    135.95      52.25   2.602  0.00976 ** 
## smokeryes:age          -71.61      50.30  -1.424  0.15563    
## children:smokeryes   -1125.32     607.17  -1.853  0.06487 .  
## bmi:regionnorthwest   -218.28      75.92  -2.875  0.00434 ** 
## bmi:regionsoutheast   -213.74      70.80  -3.019  0.00277 ** 
## bmi:regionsouthwest   -246.50      74.75  -3.297  0.00110 ** 
## bmi:smokeryes         1468.40     125.32  11.717  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 4858 on 284 degrees of freedom
## Multiple R-squared:  0.8275, Adjusted R-squared:  0.8184 
## F-statistic: 90.81 on 15 and 284 DF,  p-value: < 2.2e-16
```

- let also use the variables from the last step to establish model selection like we have covered in class R will use any combination of these variables to come up with r^2, cp values etc for each of them.we will order the result by r^2, then by adjusted then by cp and inspect if we should choose any other combination of the variables from above.
- As we can see the maximum r square values are around 0.833 (from 2047 models) which aligns with our finding from above. So need to change anything. Since there are a lot of models whose r squared is 0.833 we might choose the one with the lowest cp. And that is the 6th row with cp of 5.16. The variables used are smoker log(bmi) log(age) age:region smoker:age children:smoker bmi:region bmi:smoker with 8 variables. So that could be an alternative reduced version of the model we have above.
- So going forward let me choose the model from above. We felt it was good enough so now let us investigate more by analysing the residuals and the normality.

```
lm(charges~  children + bmi +
     smoker + age:children + age:region + age:smoker+
     + children:smoker + bmi:region +  bmi:smoker + smoker
   +  log(bmi)+ bmi*bmi + log(age)  , insurance_new) -> interactionModelSignificant

summary(interactionModelSignificant)
```

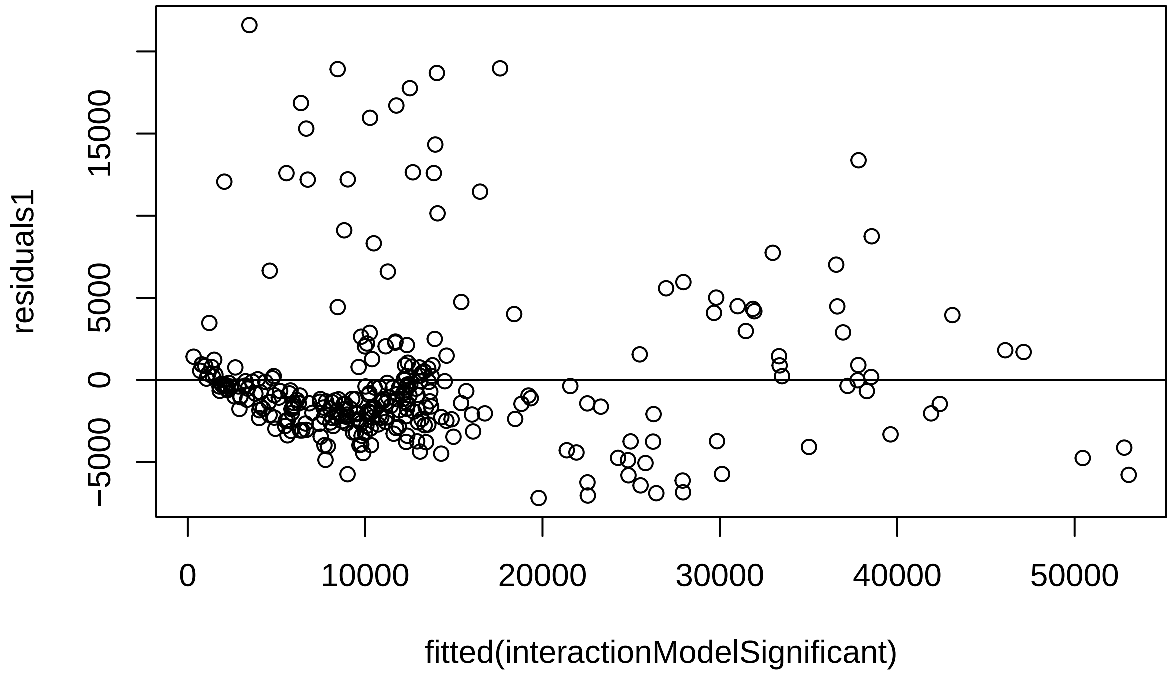
```
## 
## Call:
## lm(formula = charges ~ children + bmi + smoker + age:children + 
##     age:region + age:smoker + +children:smoker + bmi:region + 
##     bmi:smoker + smoker + log(bmi) + bmi * bmi + log(age), data = insurance_new)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -7189.3 -2334.9 -1166.7   385.6 21603.4 
## 
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)  
## (Intercept)          55898.45   31237.64   1.789  0.07461 .
## children              -319.02     745.49  -0.428  0.66902  
## bmi                     1285.83     454.43   2.830  0.00499 **
## smokeryes             -17169.12    4145.36  -4.142 4.55e-05 ***
## log(bmi)              -30700.93   13026.61  -2.357  0.01911 *
## log(age)                5125.69    1406.95   3.643  0.00032 ***
## children:age              25.99      17.10   1.520  0.12966
## age:regionnorthwest      136.02      51.05   2.664  0.00815 **
## age:regionsoutheast      122.86      51.49   2.386  0.01769 *
## age:regionsouthwest      135.95      52.25   2.602  0.00976 **
## smokeryes:age            -71.61      50.30  -1.424  0.15563
## children:smokeryes     -1125.32     607.17  -1.853  0.06487 .
## bmi:regionnorthwest     -218.28      75.92  -2.875  0.00434 **
## bmi:regionsoutheast     -213.74      70.80  -3.019  0.00277 **
## bmi:regionsouthwest     -246.50      74.75  -3.297  0.00110 **
## bmi:smokeryes           1468.40     125.32  11.717  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4858 on 284 degrees of freedom
## Multiple R-squared:  0.8275, Adjusted R-squared:  0.8184
## F-statistic: 90.81 on 15 and 284 DF,  p-value: < 2.2e-16
```

```
# The residual plot is not perfect but there is no clear pattern. So it should be relatively fine. It i

# Observing the residual vs fitted plot one can say that more points are plotted in one area specifical

resid(interactionModelSignificant) -> residuals1

plot(fitted(interactionModelSignificant), residuals1)+
abline(0,0)
```



```
## integer(0)
```

7.5 CI for two of our chosen quantitative variables

- Using this regression model Let us take at least two variables and find confidence interval for the independent variables. Let us take two such as log(age) and bmi for example.
- The standard error for log(age) is 1406.95 and the coefficient is 5125.69 and standard error for bmi is 454 and the coefficient is 1285.83 .
- Degree of freedom is 284. (300-16). t is 1.9683. Which is closer to the z score actually.

```
qt(p=.025, df=284, lower.tail = FALSE) -> t
t
```

```
## [1] 1.968352
```

```
# so for bmi
#upper bound    2179.438
1285.83 + 1.9683 * 454
```

```
## [1] 2179.438
```

```
# lower bound   392.2218
1285.83 - 1.9683 * 454
```

```
## [1] 392.2218
```

```
# for log(age)
# upper bound ] 7894.99
5125.69 + 1.9683 * 1406.95
```

```
## [1] 7894.99
```

```
# lower bound
5125.69 - 1.9683 * 1406.95
```

```
## [1] 2356.39
```

8

8.1 Researching and Applying a model analysis.

- Here we will show how random forest model can be used on our full model or using all explanatory variables to predict the dependent variable charges.
- Random forest works by using if-else decision trees using all variables. For example one possible path could be if a person is a non smoker, male and if he has a bmi value above 90, the medical charge should approximately be 1000. This is just to. show how it works under the hood but the actual mechanisms and branching rules are not as trivial as my example. For a continuous dependent variable, we should use random forest regressor (it does regression but using random forest)
- We will compare the root mean squared error using our random forest model and the regression we had above Random Forest Model.

```
set.seed(42)
rf.fit <- randomForest(charges ~ ., data=insurance_new, ntree=3,
                       keep.forest=FALSE, importance=TRUE)
```

- from fitting the random forest model, we can see that the root mean squared is 37712690ˆ(0.5). Rf.fit gives us the squared value so we have to take the root of it to find the root mean square. so 37712690ˆ(0.5) = 6141.066. This random forest model produced worse result than our regression model.
- we got a value of 4858 as our best root mean squared error from our regressoin model.
- But also just like we can tune our regression model, we can also tune our random forest model. Let us increase the number of trees from 3 to 300 in the random forest model. we get 25902511ˆ(0.5) = 5089.451. So as we increase our number of trees the root mean squared approached our best regression model out put. Of course we can tune a lot of things in the decision trees of random forest as well so random might give us a root mean square value less than our regression model.
- The variability explained by this random forest model was also close to what we have in the regression model. It is 80% here.

```
rf.fit
```

```
##
## Call:
##  randomForest(formula = charges ~ ., data = insurance_new, ntree = 3,      keep.forest = FALSE, impo
##                Type of random forest: regression
##                      Number of trees: 3
## No. of variables tried at each split: 2
##
##           Mean of squared residuals: 41785075
##                     % Var explained: 67.73
```

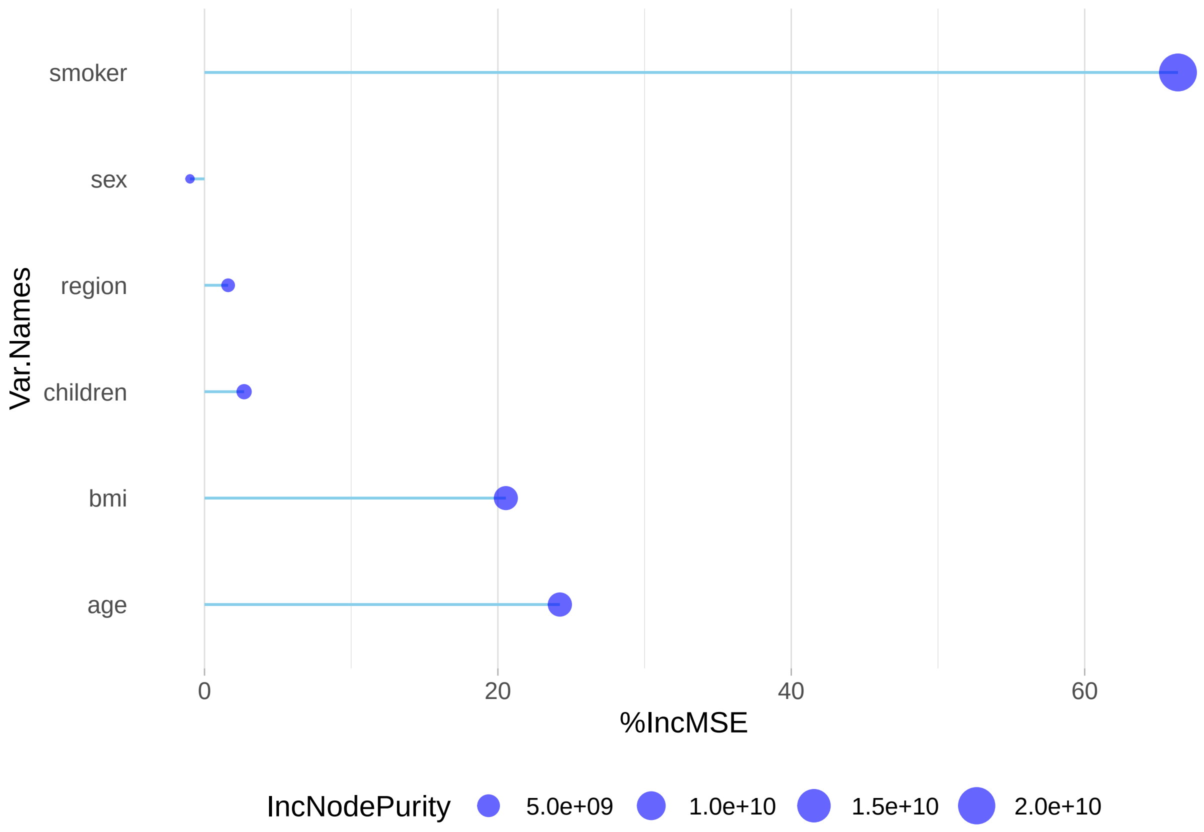
```
rf.fit <- randomForest(charges ~ ., data=insurance_new, ntree=300,
                       keep.forest=FALSE, importance=TRUE)
rf.fit
```

```
##
## Call:
##  randomForest(formula = charges ~ ., data = insurance_new, ntree = 300,      keep.forest = FALSE, imp
##                Type of random forest: regression
##                      Number of trees: 300
## No. of variables tried at each split: 2
##
##           Mean of squared residuals: 26152008
##                     % Var explained: 79.81
```

- Let also this which variables were important according to the latest random forest model.
- we had. We see that smoker variable was very important in terms of information gain(it is one of the best variable used in the decision tree and is found to be important interms of determining medical charges/bills. The other variables that are found important are age and BMI as we see from the diagram.) children, region and sex had minimal impact compared to the other variables.

```
ImpData <- as.data.frame(importance(rf.fit))
ImpData$Var.Names <- row.names(ImpData)

ggplot(ImpData, aes(x=Var.Names, y=`%IncMSE`)) +
  geom_segment( aes(x=Var.Names, xend=Var.Names, y=0, yend=`%IncMSE`), color="skyblue") +
  geom_point(aes(size = IncNodePurity), color="blue", alpha=0.6) +
  theme_light() +
  coord_flip() +
  theme(
    legend.position="bottom",
    panel.grid.major.y = element_blank(),
    panel.border = element_blank(),
    axis.ticks.y = element_blank()
  )
```



9 Final summary

- As we can see from most of our graphs, out of the 6 variables we have we found that age, BMI and smoker variables were the explanatory variables that affected the value of our dependent variable charge the most. From our final regression model we chose, we can also see that the p values for bmi, log(age), smoker, and the interaction term bmi*smoker etc were very small values showing that they were significant. Even from preliminary visual plots(1.3b), we have seen how the interaction of BMI with smoker could change the value of charges. A smoker with high BMI typically has higher charges. As we would expect, our random forest model also gave us similar conclusion when it comes to choosing the most important explanatory variables as depicted in the random forest model plot showing the important variables. So based on our regression models, preliminary visual plots and random forest model we conclude that an old smoker person who has a high BMI would likely pay the highest medical charges.

- It was also important to note that our random forest model performed very similar to our best regression model in terms of r squared (80-82 %) and in terms of residual error or root mean square(4900-5000). However, we have tried several regression model when we use the function ols_step_all_possible and insert several interaction terms and transformed functions. By using that function we have tried over 1000 different models and choose the best one. However, we only played with two or three random forest models since they are typically slower. If we have used different tuning for our random forest model, we might have got even better r squared and smaller root mean squared error

```
#
```