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 me in the Regression program. I received to uphold my practical and theoretical skills during the respective session. Firstly, I would like to thank **Pro. James C. Dickens** and secondly, I want to thank my 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 me to help me to 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.

**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   
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

## 1.1 About Data set

# 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 set source: https://www.kaggle.com/datasets/mirichoi0218/insurance  
  
# 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 ^ 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")  
  
# Let import the 300 rows   
  
# insurance\_new <- read\_csv("insurance\_300.csv")  
  
# Read in CSV file and specify column types  
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()  
))  
#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, 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.   
  
nrow(insurance\_new)

## [1] 300

ncol(insurance\_new)

## [1] 7

# 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$charges)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1136 5134 10097 13283 17154 51195

sd(insurance\_new$charges)

## [1] 11399.06

head(insurance\_new,10)

## # A tibble: 10 × 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.  
## 9 19 female 29.8 0 no southwest 1744.  
## 10 19 female 20.6 0 no southwest 1732.

## 1.2 cleaning the data and type of columns

# calling the data set using read csv file  
# insurance\_new <- read\_csv('insurance\_300.csv')  
# number of rows in data   
nrow(insurance\_new)

## [1] 300

# number of colums in data set   
ncol(insurance\_new)

## [1] 7

# colums names   
colnames(insurance\_new)

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

# visual data set look like   
head(insurance\_new,10)

## # A tibble: 10 × 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.  
## 9 19 female 29.8 0 no southwest 1744.  
## 10 19 female 20.6 0 no southwest 1732.

# type of columns used in data frame (double, charterer)  
str(insurance\_new)

## spc\_tbl\_ [300 × 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>

# summary of data colum wise   
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

# 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

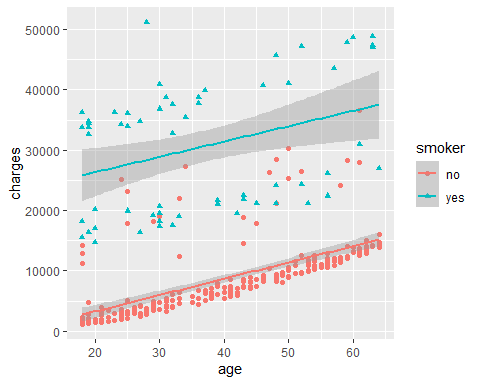
unique(insurance\_new$region)

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

## 1.3 visualization

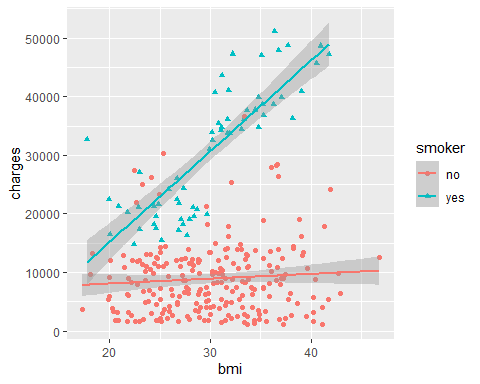
# Does age affect medical charges for smoker?   
ggplot(data = insurance\_new ,   
 aes(x=age, y=charges,shape=smoker,color = smoker)) +  
 geom\_point()+  
 geom\_smooth(method=lm)

## `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.  
  
# Does Body mass index (BMI) affect medical charges for smoker?   
ggplot(data = insurance\_new ,   
 aes(x=bmi, y=charges,shape=smoker,color = smoker)) +  
 geom\_point()+  
 geom\_smooth(method=lm)

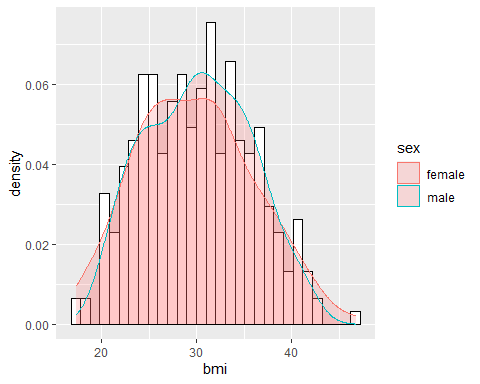
## `geom\_smooth()` using formula = 'y ~ x'



# One can clearly observe that smoking affect in BMI and increased with the medical expenses.   
  
  
# Histogram for density graph for Body mass index (BMI)   
ggplot(data = insurance\_new , aes(x=bmi,color=sex)) +  
 geom\_histogram(aes(y=..density..), colour="black", fill="white")+  
 geom\_density(alpha=.2, fill="#FF6666")

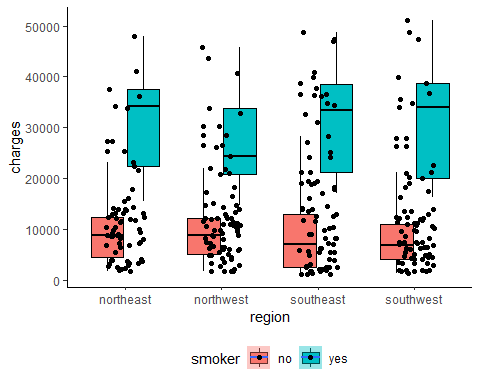
## Warning: The dot-dot notation (`..density..`) was deprecated in ggplot2 3.4.0.  
## ℹ Please use `after\_stat(density)` instead.

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



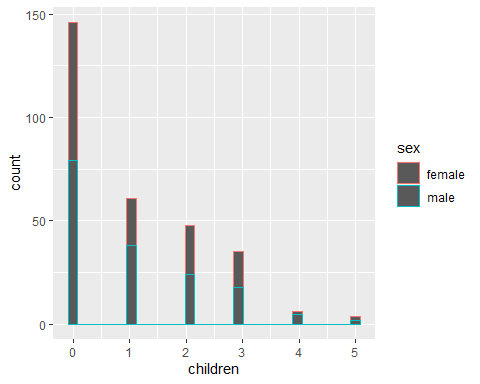
# box plot   
# comparing that does region of living affects the smokers or not. Now, looking at graph one cans say that the region does affect the charges on medical insurance.   
  
ggplot(data = insurance\_new ,   
 aes(x=region, y=charges,shape=smoker,fill = smoker)) +  
 geom\_boxplot(color="black")+  
 geom\_smooth(method=lm)+  
 theme\_classic()+  
 theme(legend.position="bottom")+  
 geom\_jitter(shape=16, position=position\_jitter(0.2))

## `geom\_smooth()` using formula = 'y ~ x'



# How many children per male/female.  
ggplot(data = insurance\_new , aes(x=children,color=sex)) +  
 geom\_histogram()

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



## ???? need to see what can be done.

insurance\_new%>%  
 filter(sex == "female")%>%  
 count(sex,children,smoker,region)%>%  
 arrange(sex, smoker)

## # A tibble: 29 × 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  
## 6 female 1 no northwest 4  
## 7 female 1 no southeast 4  
## 8 female 1 no southwest 6  
## 9 female 2 no northeast 3  
## 10 female 2 no northwest 6  
## # … with 19 more rows

insurance\_new

## # A tibble: 300 × 7  
## age sex bmi children smoker region charges  
## <dbl> <fct> <dbl> <dbl> <fct> <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.  
## 9 19 female 29.8 0 no southwest 1744.  
## 10 19 female 20.6 0 no southwest 1732.  
## # … with 290 more rows

ab <- insurance\_new%>%  
 dplyr::select(age,smoker,sex)  
ab

## # A tibble: 300 × 3  
## age smoker sex   
## <dbl> <fct> <fct>   
## 1 63 no female  
## 2 18 yes male   
## 3 48 no male   
## 4 46 no female  
## 5 52 no male   
## 6 36 no female  
## 7 19 no male   
## 8 48 no male   
## 9 19 no female  
## 10 19 no female  
## # … with 290 more rows

# 2 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.

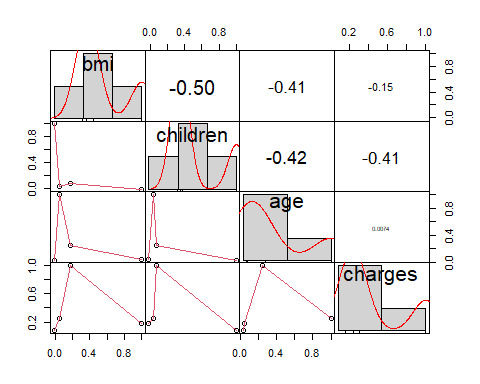
The age with bmi with only 0.04 which is nearly 0. So there exists no colinearity among the independent variables. This suggests that each of the variables might be useful if they are included in the regression model as they dont have any correlation with each other.

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



multicollinearity plot

Matrix Method with just the quantitative variables

# Both the matrix and lm method produced the same coefficients .   
lm(charges ~ age + children + bmi, data= insurance\_new)

##   
## Call:  
## lm(formula = charges ~ age + children + bmi, data = insurance\_new)  
##   
## Coefficients:  
## (Intercept) age children bmi   
## -4825.7 185.6 669.2 333.9

X <- model.matrix(~age + children + bmi , data=insurance\_new )  
  
Y <- as.matrix(insurance\_new%>%dplyr::select(charges) )  
  
Xm <- X  
Ym <- Y  
  
t(Xm) -> transposeXm  
transposeXm%\*%Xm-> ProDuct1  
solve(ProDuct1)%\*%transposeXm%\*%Ym -> interceptandslope  
interceptandslope

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

Fitted and residual values from the matrix result

Xm %\*% interceptandslope -> fitted\_values   
  
Ym - Xm %\*% interceptandslope -> residuals\_values  
  
data.frame(residuals\_values, fitted\_values)

## charges charges.1  
## 1 -989.0085 15243.617  
## 2 25048.0563 11259.742  
## 3 7264.2178 13967.964  
## 4 -7307.3293 15547.919  
## 5 -5855.5505 15580.081  
## 6 -3028.5827 8486.629  
## 7 -4516.5912 6348.685  
## 8 11471.3074 16997.612  
## 9 -6906.4489 8650.914  
## 10 -3847.6491 5579.326  
## 11 -5618.8301 15244.750  
## 12 6333.7752 11824.101  
## 13 -1710.3532 13544.136  
## 14 -1103.9309 13136.257  
## 15 -9102.4613 10929.304  
## 16 -4751.5016 18574.305  
## 17 -5580.2104 10109.687  
## 18 -11080.4895 16519.239  
## 19 -545.1995 15023.530  
## 20 -7512.8549 9390.784  
## 21 -4487.0540 6191.622  
## 22 9457.3790 14723.554  
## 23 -6272.3055 13322.948  
## 24 -5715.7305 12860.593  
## 25 -8454.0145 13291.597  
## 26 4586.0890 14377.083  
## 27 -3137.6972 15362.048  
## 28 -5854.1181 16830.364  
## 29 -2800.9658 13223.882  
## 30 -7291.4736 9422.150  
## 31 -6536.3015 19485.457  
## 32 -6918.8827 13230.835  
## 33 -5140.0970 16652.502  
## 34 24430.1401 14279.036  
## 35 -3547.1978 13654.418  
## 36 -4220.4061 7429.193  
## 37 -8388.6034 15807.125  
## 38 9204.2616 11991.556  
## 39 -8535.9195 12594.632  
## 40 24252.5506 10001.503  
## 41 -3975.1487 7836.358  
## 42 9298.2261 18989.672  
## 43 -7625.5042 16296.695  
## 44 -3355.1769 16707.277  
## 45 -7883.6191 13308.642  
## 46 28469.8943 18585.638  
## 47 -872.2145 13829.332  
## 48 24308.1376 12529.329  
## 49 -3702.9410 9558.844  
## 50 -7938.2199 15379.721  
## 51 -4261.3478 10378.842  
## 52 -8597.2703 11094.309  
## 53 -6245.2067 8212.229  
## 54 8587.7262 12510.828  
## 55 -5455.4043 19375.227  
## 56 -3754.3456 12645.485  
## 57 -7375.6232 18223.758  
## 58 -5384.7769 17490.097  
## 59 -10058.1192 19859.007  
## 60 -6102.1897 14717.490  
## 61 -4539.2015 7719.712  
## 62 -8119.2204 16243.629  
## 63 8415.9328 12330.056  
## 64 8549.8676 12532.292  
## 65 10181.3196 12231.329  
## 66 -4337.6615 5964.944  
## 67 -7181.7439 15516.201  
## 68 -9462.0035 11379.322  
## 69 -9146.4181 14982.938  
## 70 -7972.2918 12374.525  
## 71 -8586.9970 14876.752  
## 72 -2942.0742 15421.783  
## 73 -5693.2326 14299.450  
## 74 26734.7954 18967.227  
## 75 -7271.4638 8990.900  
## 76 -5725.9077 16327.540  
## 77 9068.7017 13409.898  
## 78 -2440.4142 13232.374  
## 79 8023.1201 11017.756  
## 80 -2946.9890 14017.524  
## 81 14001.7189 6165.617  
## 82 -10216.4342 13417.679  
## 83 -1148.9109 13178.198  
## 84 -7649.7062 17735.552  
## 85 -7965.9241 15314.066  
## 86 13103.2643 8881.206  
## 87 -6934.6005 13341.011  
## 88 -4621.9304 6250.401  
## 89 -8049.8928 12310.637  
## 90 -8454.4298 17502.457  
## 91 7533.6765 8886.818  
## 92 -5163.0167 19144.867  
## 93 -8924.8488 10186.708  
## 94 27147.7033 20122.151  
## 95 24459.3599 10320.255  
## 96 24002.2971 9898.356  
## 97 -4129.0268 12732.850  
## 98 -4497.7911 11184.222  
## 99 8025.9733 18441.124  
## 100 -9441.7343 12884.798  
## 101 -10518.2917 16227.456  
## 102 28825.3221 19071.469  
## 103 -6429.5241 14740.363  
## 104 -8063.2787 14192.076  
## 105 7603.9574 9757.809  
## 106 7382.1452 20559.142  
## 107 8317.4582 11198.083  
## 108 -11189.8598 23782.394  
## 109 -8190.4278 18791.840  
## 110 -5641.9275 10213.341  
## 111 8348.0554 7949.791  
## 112 17702.8001 7378.968  
## 113 -6758.7947 13507.386  
## 114 8552.1709 10555.609  
## 115 -6270.6414 14387.321  
## 116 24391.5772 9358.715  
## 117 -9428.2055 14825.822  
## 118 -2680.3563 14134.378  
## 119 -10466.2098 15228.539  
## 120 -5438.4295 15216.777  
## 121 -7694.1220 12617.038  
## 122 -3768.9032 12071.439  
## 123 -5474.7371 7106.773  
## 124 -6945.8250 10483.528  
## 125 -3310.1963 14392.773  
## 126 23722.1212 11769.519  
## 127 -8248.3523 13623.390  
## 128 -7929.9288 9456.241  
## 129 -3019.0397 13961.172  
## 130 -7907.2848 10762.722  
## 131 -3184.5013 18204.261  
## 132 -3522.6756 15181.055  
## 133 -1266.9528 14396.556  
## 134 -5558.0260 10973.687  
## 135 -8322.9520 15470.425  
## 136 -6908.4682 11596.265  
## 137 23842.8130 12281.761  
## 138 13745.8443 9495.630  
## 139 -7782.3420 10686.430  
## 140 -6554.6117 8219.611  
## 141 -9824.4928 10965.938  
## 142 8817.6636 16515.669  
## 143 -2375.5667 13721.086  
## 144 -4874.8960 17106.510  
## 145 -6203.7098 16804.258  
## 146 -7680.1952 17314.733  
## 147 27419.2985 16159.641  
## 148 -5848.2395 8569.560  
## 149 -6860.6041 10901.162  
## 150 28346.1492 18959.156  
## 151 -5714.2777 11837.846  
## 152 -3644.3520 15978.180  
## 153 17065.8045 10310.100  
## 154 -4048.5993 12034.414  
## 155 24595.2614 13012.266  
## 156 -3877.2614 6463.112  
## 157 -11104.5340 12251.331  
## 158 -7379.8524 11237.612  
## 159 -8623.5892 10015.118  
## 160 -11283.6324 17644.626  
## 161 -4083.2201 16179.871  
## 162 24143.6298 12045.472  
## 163 -8769.9492 12757.875  
## 164 -6017.6085 10257.501  
## 165 -2790.2575 15013.156  
## 166 -3215.9291 16673.890  
## 167 9641.0555 12133.267  
## 168 -2412.4392 6145.064  
## 169 7361.4545 10897.762  
## 170 -3269.9139 13133.386  
## 171 -9799.2645 18532.494  
## 172 -7250.6156 12402.750  
## 173 -5746.7607 7583.998  
## 174 -6658.8656 14017.041  
## 175 -6911.0324 11939.179  
## 176 -8185.0636 11084.553  
## 177 25143.4983 15788.931  
## 178 29367.4161 19457.034  
## 179 -6301.3271 19772.187  
## 180 -6060.8831 11140.979  
## 181 -8926.4238 11424.838  
## 182 -5022.5230 8066.736  
## 183 -2050.2143 14374.150  
## 184 -7429.7158 14205.677  
## 185 -4199.5602 15753.784  
## 186 23747.1859 8801.155  
## 187 -3608.3879 15738.002  
## 188 16020.9150 14263.728  
## 189 -3234.7147 17278.191  
## 190 -9196.5494 16393.416  
## 191 -4284.9871 13567.468  
## 192 -7735.8870 11810.341  
## 193 -5334.7169 12381.439  
## 194 -6526.9612 11404.942  
## 195 -5589.6165 15996.702  
## 196 -7174.4092 18337.977  
## 197 -4692.9335 6889.407  
## 198 -5290.4715 15629.403  
## 199 10430.9263 13962.696  
## 200 28698.4165 19975.142  
## 201 -8917.5492 10563.979  
## 202 2635.3316 8637.000  
## 203 -3750.5690 17595.366  
## 204 -7324.6297 15862.918  
## 205 -2519.3965 11713.235  
## 206 16276.7562 20303.526  
## 207 -663.0349 13067.914  
## 208 24683.8381 9756.018  
## 209 -5258.9318 8233.058  
## 210 -9060.2579 11198.329  
## 211 8633.8083 10888.160  
## 212 24323.7560 14387.244  
## 213 -5151.4611 12572.656  
## 214 8567.7559 8517.512  
## 215 -4454.2925 10726.770  
## 216 -3076.4093 15120.751  
## 217 -6649.8673 8900.702  
## 218 -10390.0122 13079.508  
## 219 8927.7311 12732.199  
## 220 1306.4635 11583.594  
## 221 -3369.1475 18061.817  
## 222 -4852.4736 18488.112  
## 223 23557.1175 11249.350  
## 224 -3394.2325 17368.688  
## 225 -1580.1362 15031.258  
## 226 -7159.0782 11106.491  
## 227 24389.8170 13352.759  
## 228 -8280.8709 15542.612  
## 229 -6404.7305 12860.593  
## 230 -2950.5196 19019.604  
## 231 -5122.9671 15195.022  
## 232 -8694.3718 10714.924  
## 233 -10333.3636 12814.343  
## 234 -6836.1431 19183.315  
## 235 8168.0685 9328.238  
## 236 -5210.8099 6948.186  
## 237 -7276.9160 9996.196  
## 238 8597.0561 6921.124  
## 239 -5920.5741 12417.460  
## 240 -1024.4812 10668.734  
## 241 4292.9636 19934.374  
## 242 -2709.6539 13668.984  
## 243 -4205.5863 11530.634  
## 244 24623.0562 9109.631  
## 245 -3950.4997 15470.600  
## 246 -6819.0856 15643.071  
## 247 12306.4329 14731.481  
## 248 24349.5409 8384.645  
## 249 -4786.0147 14655.825  
## 250 -4661.2776 16403.004  
## 251 -5852.0349 9205.319  
## 252 -6271.2931 17096.547  
## 253 -2791.4467 14885.925  
## 254 -8617.5968 9753.996  
## 255 -3425.4671 17426.754  
## 256 -5598.9766 9244.066  
## 257 -10322.9532 11957.527  
## 258 -6198.1602 14478.783  
## 259 13621.3413 12615.239  
## 260 24817.4768 15903.074  
## 261 -5302.6454 11501.397  
## 262 -6015.8735 22101.001  
## 263 -10847.2225 14837.063  
## 264 8985.1488 12238.527  
## 265 4266.2607 13612.640  
## 266 -5982.9026 11952.626  
## 267 -5215.4704 10147.117  
## 268 -6373.7245 15208.989  
## 269 -6226.1823 9611.581  
## 270 2836.5703 11296.467  
## 271 -6420.0947 8884.714  
## 272 2521.6416 16284.504  
## 273 -5472.7887 7442.403  
## 274 3787.1148 14142.189  
## 275 8591.9385 9631.513  
## 276 -9151.1632 15625.176  
## 277 10956.8603 15152.469  
## 278 -6271.6499 8409.304  
## 279 26015.0422 15082.120  
## 280 8194.4404 11739.018  
## 281 -1325.5684 14337.777  
## 282 -5054.0189 15285.519  
## 283 -5395.3463 19806.278  
## 284 -4551.8378 17695.174  
## 285 -4480.1428 7740.342  
## 286 24836.1826 15000.336  
## 287 8339.1284 6372.615  
## 288 -4481.5181 17039.123  
## 289 -6788.2897 13070.525  
## 290 -8327.0321 16350.168  
## 291 769.5575 12434.728  
## 292 12446.3552 18495.837  
## 293 -4410.3304 13857.581  
## 294 -6289.0895 15253.150  
## 295 24782.6857 15089.019  
## 296 -8934.2568 12203.103  
## 297 -4589.4455 11447.925  
## 298 -8895.6234 17873.808  
## 299 38000.0742 13194.485  
## 300 -7461.6500 14973.917

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.   
  
  
X <- model.matrix(~age + children + bmi + region + sex + smoker , data=insurance\_new )  
  
Y <- as.matrix(insurance\_new%>%dplyr::select(charges) )  
  
Xm <- X  
Ym <- Y  
  
t(Xm) -> transposeXm  
transposeXm%\*%Xm-> ProDuct1  
solve(ProDuct1)%\*%transposeXm%\*%Ym -> interceptandslope  
interceptandslope

## 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

#lm(charges ~ age + children + bmi + region + sex + smoker,insurance\_new) -> x  
  
#summary(x)

lm method including both categorical and quantitative variables

lm(charges ~ age + children + bmi + region + sex + smoker,insurance\_new) -> x  
  
summary(x)

##   
## 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

Evaluating Various regression models using summary tables and statistics

# 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.   
  
lm(charges ~ age + children + bmi + region + sex + smoker,insurance\_new) -> x  
summary(x)

##   
## 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

# Let us investigate the levels of the catagorical variables   
# There are two levels female and male   
levels(factor(insurance\_new$sex) )

## [1] "female" "male"

# Four regions : northeast northwest , southeast, southwest   
levels( factor( insurance\_new$region ) )

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

# Two levels - Smoker or not smoker   
levels ( factor(insurance\_new$smoker))

## [1] "no" "yes"

# 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   
# significantly to 4910. Our R squared also increased to 0.8331.   
  
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  
  
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

# 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 significanly 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) -> interactionModel  
summary(interactionModel)

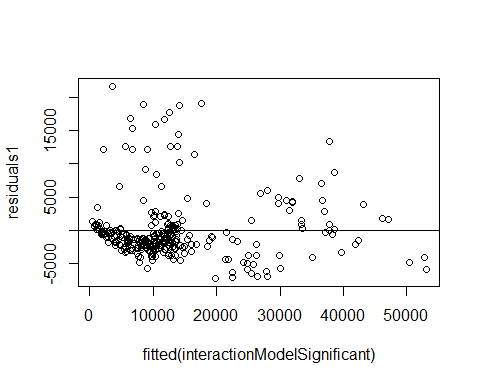
##   
## Call:  
## lm(formula = charges ~ age + children + bmi + region + sex +   
## smoker + age:children + age:bmi + age:region + age:sex +   
## age:smoker + children:sex + children:smoker + bmi:region +   
## bmi:sex + bmi:smoker + region:sex + region + smoker + sex:smoker +   
## log(bmi) + bmi \* 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

Producing a reduced model (removing variables of our choice with justification)

# # Let eliminate all interaction   
# terms or the single 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 estabilish 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.   
  
# not running it for now - slow  
#k <- ols\_step\_all\_possible(interactionModel)  
#as\_tibble(k) -> tk  
#arr <-arrange(tk, -rsquare, -cp, -adjr)  
# 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.   
  
#arr  
#arr[['predictors']][6:6]  
  
  
  
# 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  
  
# The residual plot is not perfect but there is no clear pattern. So it should be relatively fine.   
# It indicates relatively good amount of constant variance   
  
resid(interactionModelSignificant) -> residuals1  
plot(fitted(interactionModelSignificant), residuals1)  
abline(0,0)



Showing confidence intervals 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 the standard error for bmi is 454 and the coeffient is 1285.83 .   
  
# so to find the confidence interval, let us find the t statistics   
  
# 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

Researching and Applying a model analysis not discussed in the class

# 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.   
#install.packages("randomForest")  
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 37712690^(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  
# 37712690^(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.  
rf.fit

##   
## Call:  
## randomForest(formula = charges ~ ., data = insurance\_new, ntree = 3, keep.forest = FALSE, importance = TRUE)   
## 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

# 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 25902511^(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 <- 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, importance = TRUE)   
## 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 determing   
# 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()  
 )

