Section I

Wenxiong Lu

9/15/2021

Section I

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

1. Read in Data and report summary statistics (mean + std / frequency) for age, sex, bmi, children, smoker, and charges) by region.

```r
insuData<- read.csv("/Users/wenxionglu/Downloads/jobappli/Insurify/data/insurance.csv",
                    header = TRUE)
# check null, check data structure
insuData$sex<-as.factor(insuData$sex)
insuData$smoker <- as.factor(insuData$smoker)
insuData$region <- as.factor(insuData$region)
```

```r
# quantitative data, sex and smoker not included.
quantByRegion <- sqldf('SELECT region, age,bmi,children,charges FROM insuData
                       ORDER BY region')
qualiByRegion <- sqldf('SELECT region, sex, smoker FROM insuData
                       ORDER BY region')

summary_avg <- sqldf('SELECT region,AVG(age), AVG(bmi), AVG(children),
                      AVG(charges) FROM insuData
                     GROUP BY region')

summary_std <- sqldf('SELECT region,STDEV(age), STDEV(bmi),
                      STDEV(children), STDEV(charges)
                     FROM insuData
                     GROUP BY region')

# qualitative data: sex, smoker
statistics_quali <- sqldf('SELECT region, COUNT(smoker) FROM insuData GROUP BY region')

# more in details:
freq_detail<-
    data.frame(v.name = character(),
             v.region=character(),v.freq_by_region=integer())
```

```
sex_freq_detail<- rbind(freq_detail,
                      sqldf('SELECT sex,region,COUNT(sex) AS sex_count
                            FROM insuData GROUP by region,sex'))
smoker_freq_detail<-rbind(freq_detail,
              sqldf('SELECT smoker,region,COUNT(smoker) AS smoker_count
                    FROM insuData
                    GROUP BY region,smoker'))


statistics_quant <- cbind(summary_avg,summary_std)
statistics_quant <- statistics_quant[-6]

qualitative_detail <- cbind(sex_freq_detail,smoker_freq_detail)
qualitative_detail <- qualitative_detail[-2]
```

2:

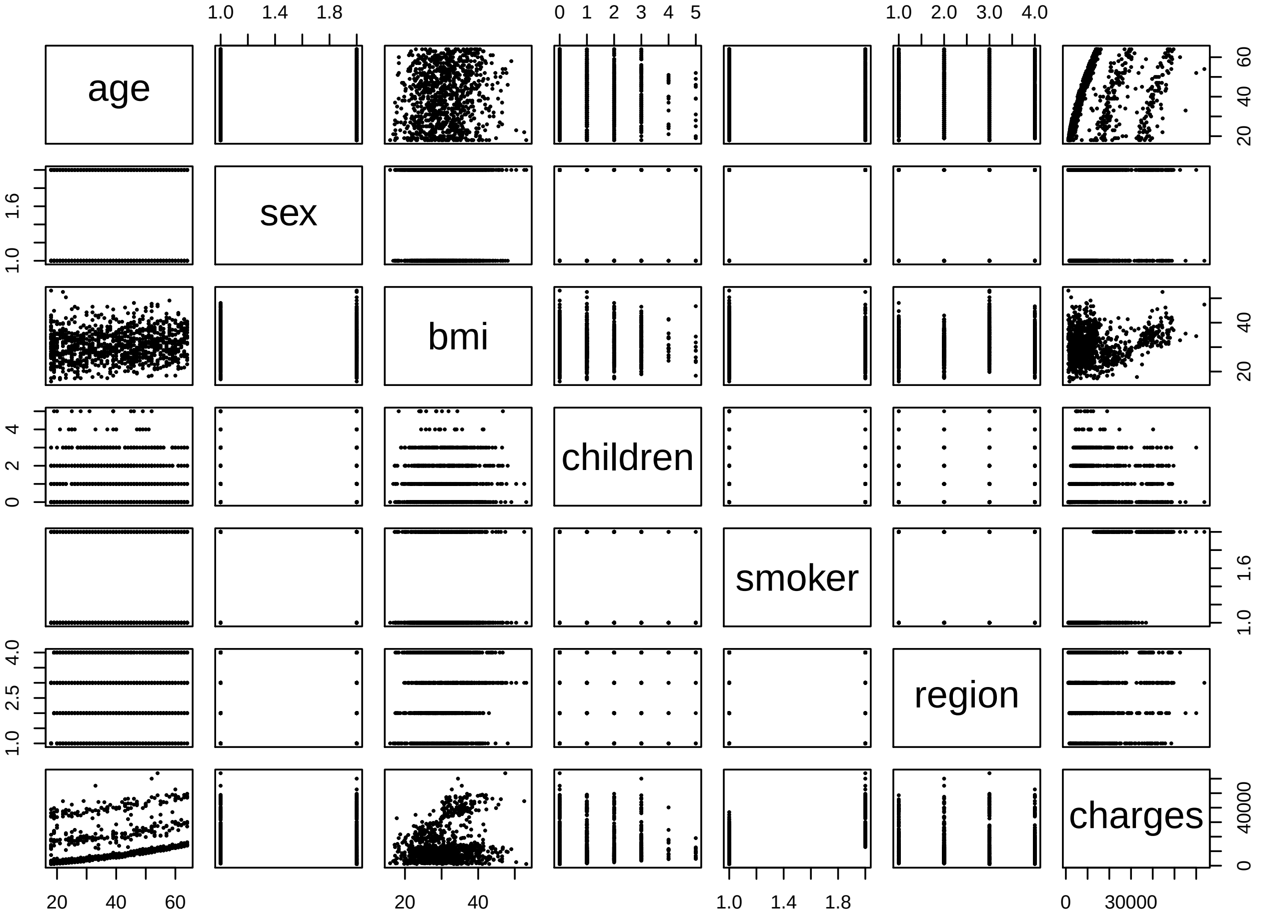
Characterize the population. We should start from a few aspects, data characteristics, central tendency, dispersion, and association.

Data Types and Overview.

```
str(insuData)
```

```
## 'data.frame': 1004 obs. of 7 variables:
## $ age : int 45 36 64 46 19 34 19 64 28 49 ...
## $ sex : Factor w/ 2 levels "female","male": 1 1 1 2 2 2
1 2 1 2 ...
## $ bmi : num 25.2 30 26.9 25.7 31.9 ...
## $ children: int 2 0 0 3 0 1 0 0 0 3 ...
## $ smoker : Factor w/ 2 levels "no","yes": 1 1 2 1 2 1 1
1 1 1 ...
## $ region : Factor w/ 4 levels
"northeast","northwest",..: 1 2 2 2 2 4 2 2 1 2 ...
## $ charges : num 9095 5272 29331 9302 33750 ...
```

```
pairs(insuData,cex=0.25)
```



The data set consists of quantitative data and qualitative data. It is worthy to point out that we can clearly see three lines in the scatterplot: "age vs charges", which may indicates a different linear relationships exist in different age periods(maybe a multinomial distribtion). We may categorize age into three categories for better fit, however, I do not recommend it for the following reasons: 1. It will decrease the power and precision for the linear model we fit; 2. We don't know the clear-cut or cause in reality for age (such as insurance age-policies), finding a best fit for the cut-point is possible but is also hazardous which may cause our model overfitting the current data. Let's try fit it with linear regression.

Qualitative Data (sex, region and smoker):

```
# result
statistics_quali
```

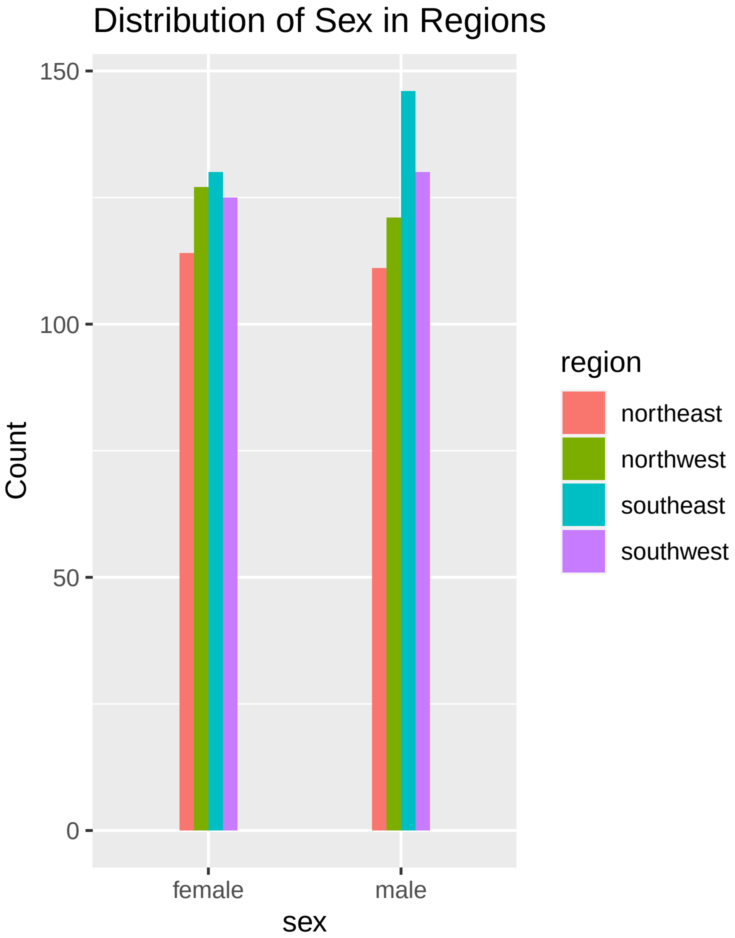
```
##      region COUNT(smoker)
## 1 northeast           225
## 2 northwest           248
## 3 southeast           276
## 4 southwest           255
```

```
qualitative_detail
```

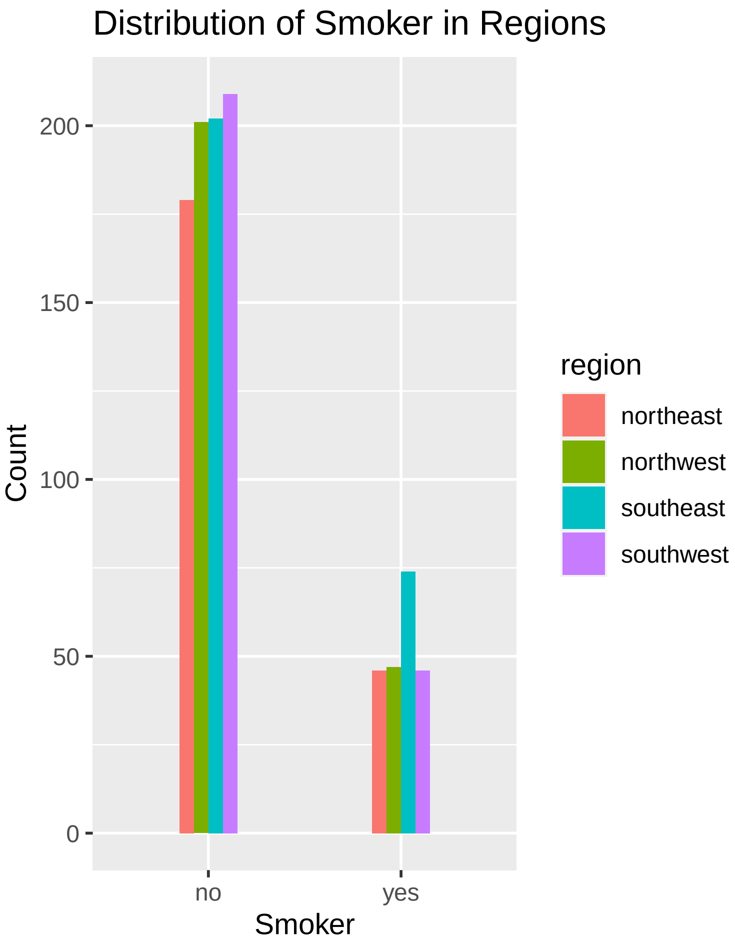
```
##      sex sex_count smoker    region smoker_count
## 1 female       114     no northeast          179
## 2   male       111    yes northeast           46
## 3 female       127     no northwest          201
## 4   male       121    yes northwest           47
## 5 female       130     no southeast          202
## 6   male       146     yes southeast          74
## 7 female       125      no southwest         209
## 8   male       130     yes southwest          46
```

In histogram:

```
# sex & smoker
par(mfrow=c(1,2))
ggplot(qualitative_detail, aes(fill=region, y=sex_count, x=sex)) +
    geom_bar(width = 0.3 ,position="dodge", stat="identity")+
    ylab("Count") +
    ggtitle("Distribution of Sex in Regions")
```



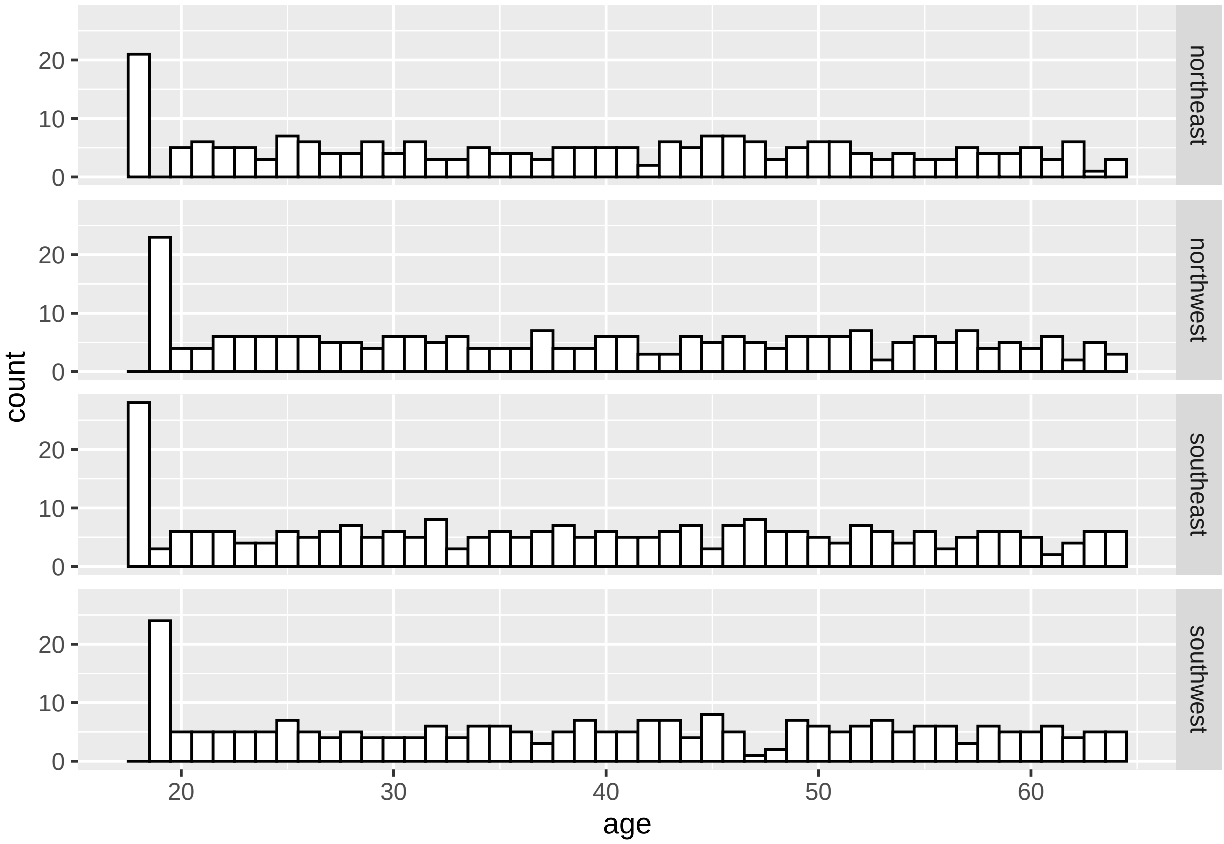
```
ggplot(qualitative_detail, aes(fill=region, y=smoker_count, x=smoker)) +
    geom_bar(width = 0.3 ,position="dodge", stat="identity")+
    ylab("Count")+
    labs(y="Count",x="Smoker",title = "Distribution of Smoker in Regions")
```



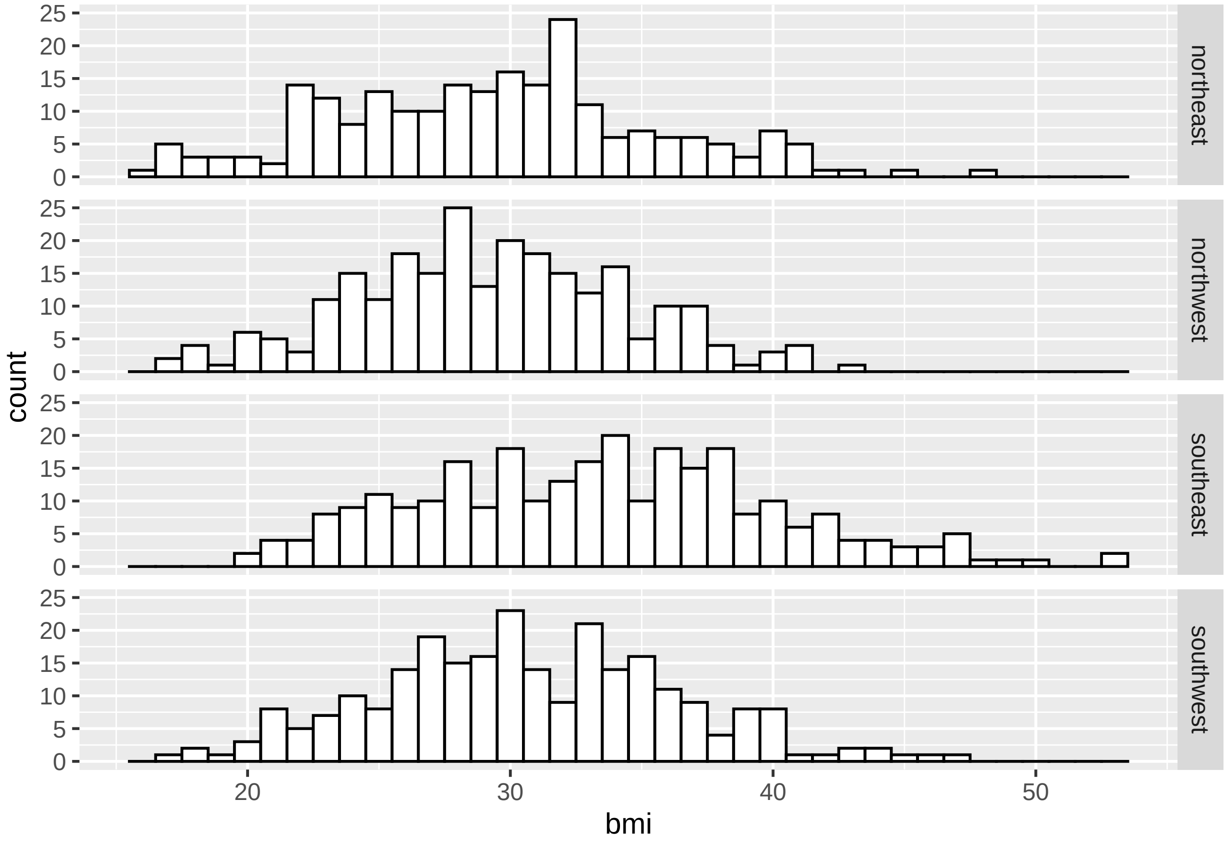
On a large scope, data of sex and smoker in different region are similar. They have close mean and dispersion. In detail, we can see that there are more males in southeast region than other regions, also, there are more smokers in southeast region. However, we should carefully look for confounding factors before concluding a causation statement.

Quantitative Data (age, bmi, children and charges)
Plot and check central tendency, dispersion, association, if applicable, check normality.

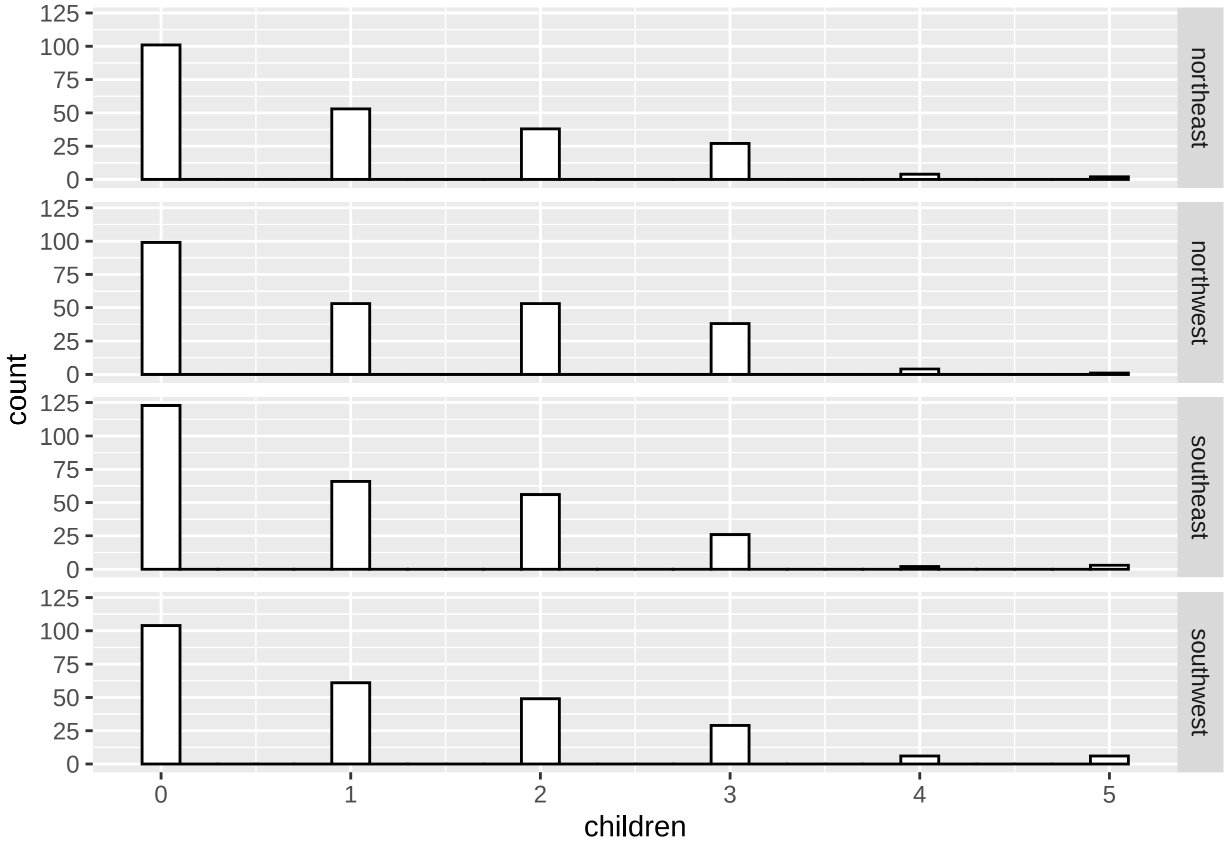
```
ggplot(insuData, aes(x = age)) +
  geom_histogram(binwidth = 1 ,fill = "white", colour = "black") +
  facet_grid(region ~ .)
```



```
ggplot(insuData, aes(x = bmi)) +
  geom_histogram(binwidth = 1 ,fill = "white", colour = "black") +
  facet_grid(region ~ .)
```



```
ggplot(insuData, aes(x = children)) +
  geom_histogram(binwidth = 0.2 ,fill = "white", colour = "black") +
  facet_grid(region ~ .)
```



bmi data's distribution looks similar to normal distribution across different regions, but not exactly normally distributed. Use Levene's test for homogeneity of variance.
The age data is not normal distributed, it is more similar(not equal) to an uniform distribution which has almost constant variation. Data for age at 18 and 19 are more frequently observed than other ages. Notice that, there is 0 observation for age 18 and large number in obs. for age 19 in both west-regions. In contrast, both east-regions has large number in obs. for age 18 and very small obs. numbers for age 19. The children data looks similar to a exponential distribution, where probability decrease dramatically in the beginning and keep decreasing slowly on the entire x-axis. A power transformation may be applied in order to obtain a normal distribution.

```
fligner.test(insuData$age ~ insuData$region)
```

```
##
## Fligner-Killeen test of homogeneity of variances
##
## data: insuData$age by insuData$region
## Fligner-Killeen:med chi-squared = 0.67808, df = 3,
p-value = 0.8783
```

```
kruskal.test(insuData$age ~ insuData$region)
```

```
##
## Kruskal-Wallis rank sum test
##
## data: insuData$age by insuData$region
## Kruskal-Wallis chi-squared = 1.3594, df = 3, p-value =
0.7151
```

```
leveneTest(insuData$bmi ~ insuData$region)
```

```
## Levene's Test for Homogeneity of Variance (center =
median)
## Df F value Pr(>F)
## group 3 5.7702 0.0006518 ***
## 1000
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
' ' 1
```

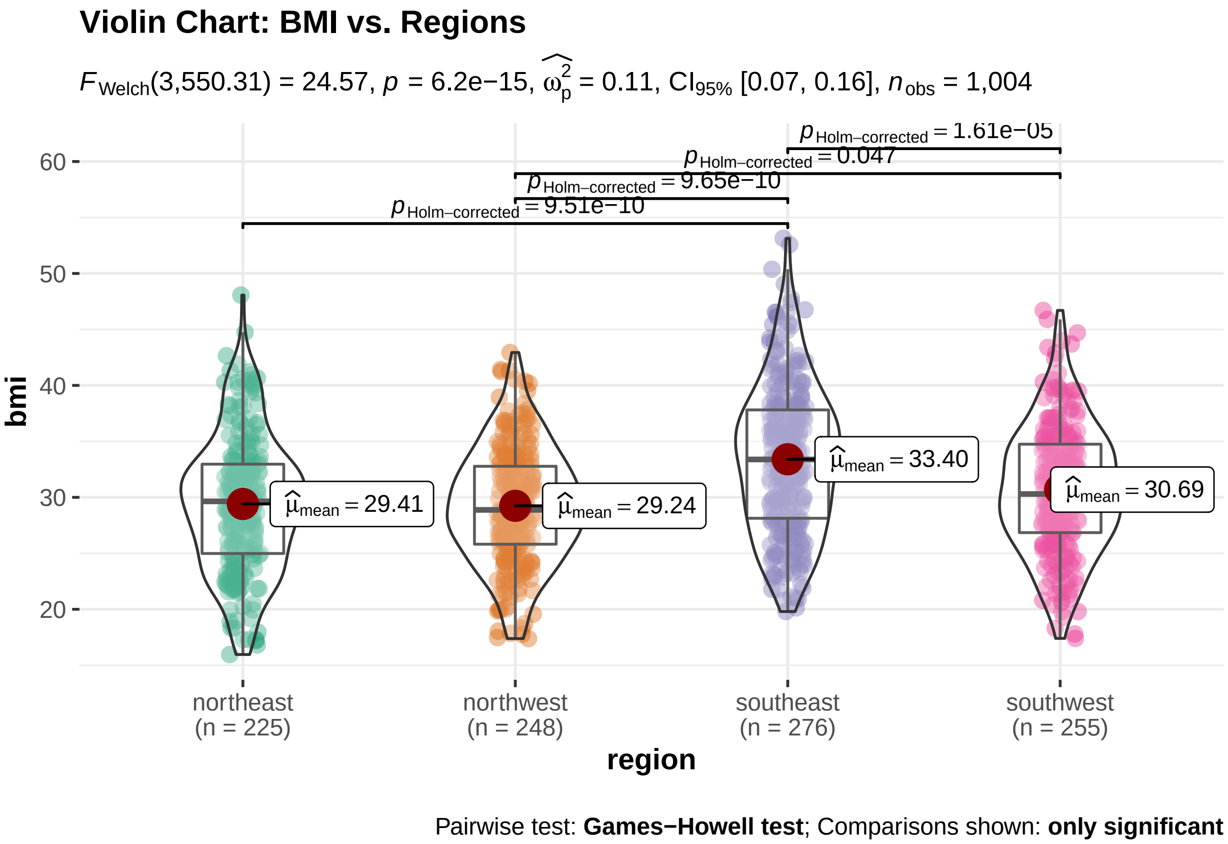
age H0: Homogeneity of variance for age hold. $\sigma_{northeast} = \sigma_{northwest} = \sigma_{southeast} = \sigma_{southwest}$
H1: Homogeneity of variance for age does not hold for at least 1 pair of variance comparison.

Krushkal: p-value = 0.7151 > 0.05 (default) Fligner: p-value = 0.8783 > 0.05 (reference) Conclusion: Since Krushkal p > 0.05 we fail to reject H0. The homogeneity of variance for age hold. For even more robust test, we can check Fligner test, which we fail to reject H0.

bmi H0: Homogeneity of variance for bmi hold. $\sigma_{northeast} = \sigma_{northwest} = \sigma_{southeast} = \sigma_{southwest}$
H1: Homogeneity of variance for bmi does not hold for at least 1 pair of variance comparison.

Since P = 0.0006518 « 0.05, reject H0. The homogeneity of variance for bmi does not hold. Since constant of variance assumption violated, we can look into Welch's Anova for mean comparison.

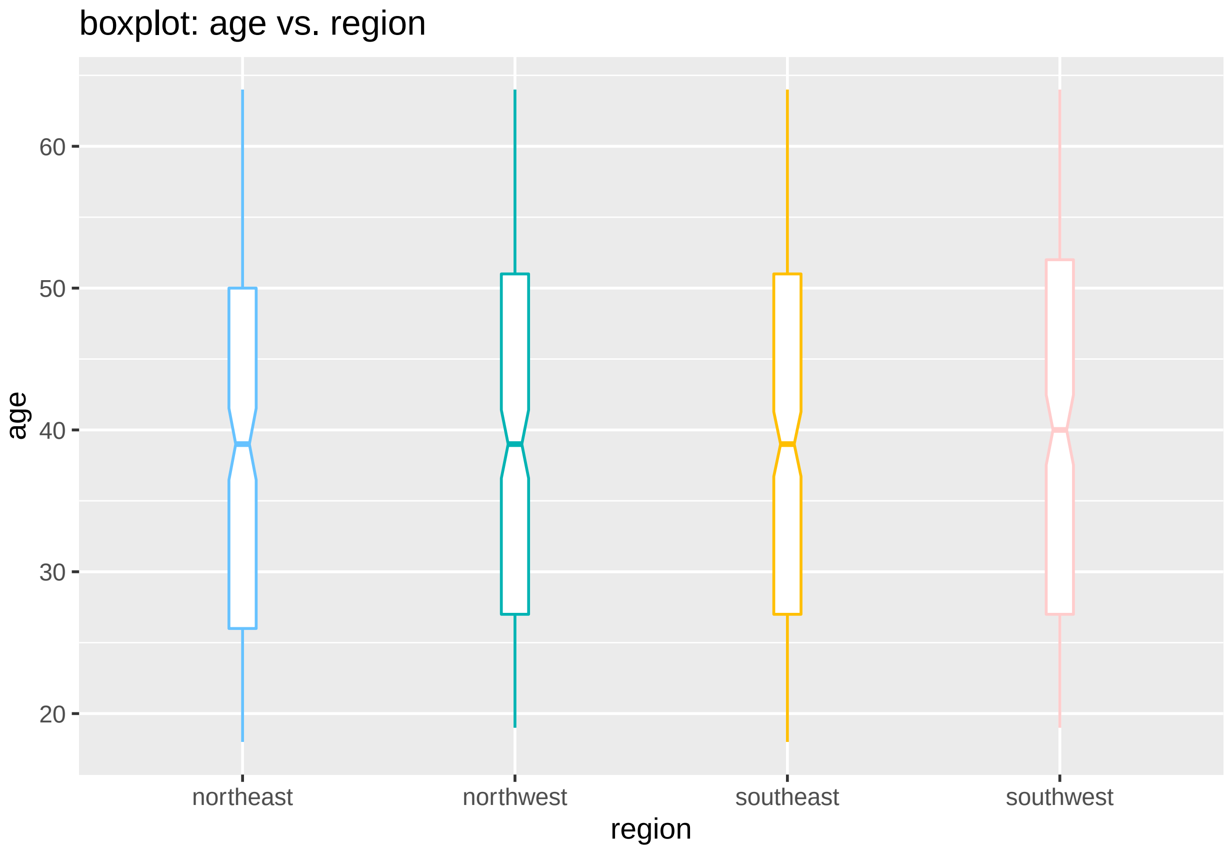
```
bmi_plt <- ggbetweenstats(
  data = insuData,
  x = region,
  y = bmi,
  title = "Violin Chart: BMI vs. Regions")

bmi_plt
```



bmi:
H0: $\mu_{n.east} = \mu_{n.west} = \mu_{s.east} = \mu_{s.west}$ H1: At least one of the mean comparison is not equal.
Welch's Anova has p-value= 6.2e-15 which is an extreme small number, we should reject H0. We in favor that the means of age distribution among different regions have significant differences.

boxplot and violin plot for age, children and charges.

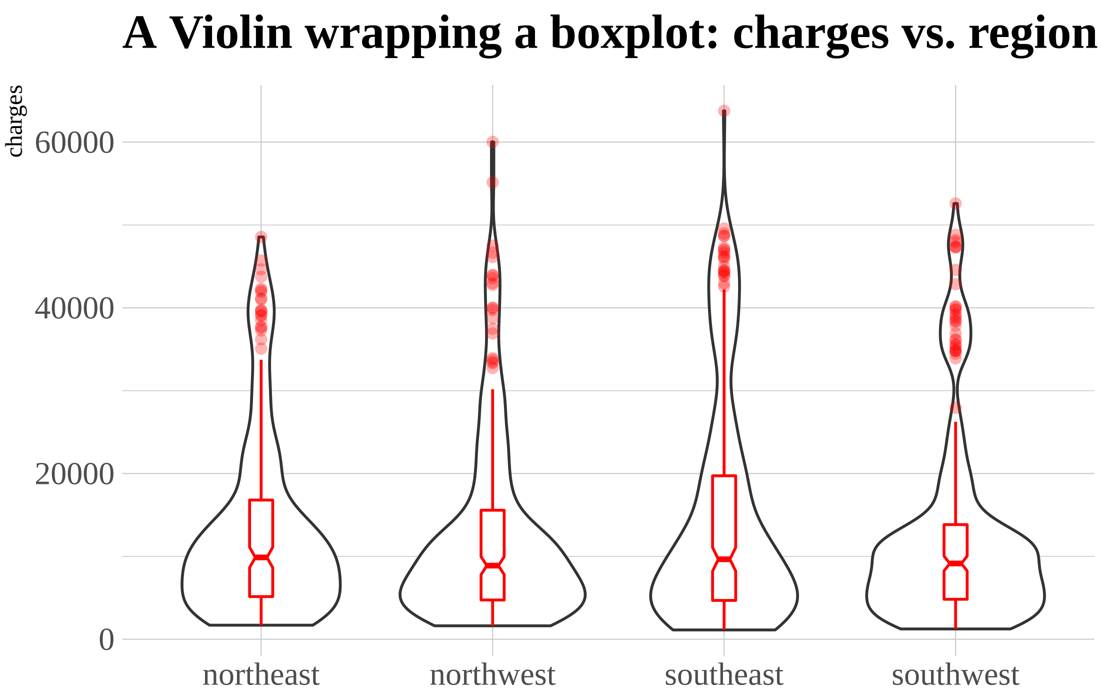
```
ggplot(insuData, aes(x=region, y=age))+
    geom_boxplot(color = c("#66c2ff","#00b3b3","#ffbf00","#ffcccc"),
                 width = 0.1, notch = TRUE) +
    ggtitle("boxplot: age vs. region")
```



Combined with the previous barplot for age data, we can see that the age distribution has similar mean value and variance.

violin chart (charges):

```
ggplot(insuData, aes(x=region, y=charges)) +
    geom_violin(width=0.8) +
    geom_boxplot(width=0.1, color="red", alpha=0.3, notch = TRUE) +
    scale_fill_viridis(discrete = TRUE) +
    theme_ipsum() +
    theme(plot.title = element_text(family = "serif"),
          axis.title.x = element_text(family = "serif"),
          axis.title.y  = element_text(family = "serif"),
          axis.text = element_text(family = "serif"),
      legend.position="none") +
    ggtitle("A Violin wrapping a boxplot: charges vs. region") +
    xlab("")
```



The violin charts are embedded with boxplots. From the boxplots, we can see that most of the data of charges concentrated below 20000, and a few observations that have extreme high charges around 40000 which are beyond 99.3% quantile. The overlapping notch indicates that with roughly 95% confidence there isn't significant differences between these means for different regions. In the case that extreme values are taking less weight(maybe transformations), the distribution of charges would be close to gamma distribution. We should avoid data snooping, in the opposite, I suggest further investigates these extreme values.

3:
Age: Female vs. Male

The sex data is nominal, and the age data is non-normal distributed. We can use non-parametric test: Kruskal-Wallis test in this scenario which we can compare the average age of male and female. Then we may infer whether male are female have the same age overall. H0: The average age difference between male and female is 0. H1: The average age difference between male and female is not 0.

```
kruskal.test(insuData$age ~ insuData$sex)
```

```
##
## Kruskal-Wallis rank sum test
##
## data: insuData$age by insuData$sex
## Kruskal-Wallis chi-squared = 2.5428, df = 1, p-value =
0.1108
```

The test result p-value = 0.1108 > 0.05, so we fail to reject the null hypothesis. Conclusion: The female and male in the data have the same average age overall. Then we may infer that male age is not different from female age in general.

4. correlation: smoker and children
   Since variables child__OR__not and smoker are a qualitative data, we can take chi-squared test for independence.

```
child_OR_not <- insuData %>%
  mutate(children = ifelse(children==0, "yes", "no"))
```

```
smoker_children <- table(child_OR_not$children,child_OR_not$smoker)
smoker_children
```

```
##
##        no yes
##   no  456 121
##   yes 335  92
```

```
chisq.test(smoker_children)
```

```
##
## Pearson's Chi-squared test with Yates' continuity
correction
##
## data: smoker_children
## X-squared = 0.02025, df = 1, p-value = 0.8868
```

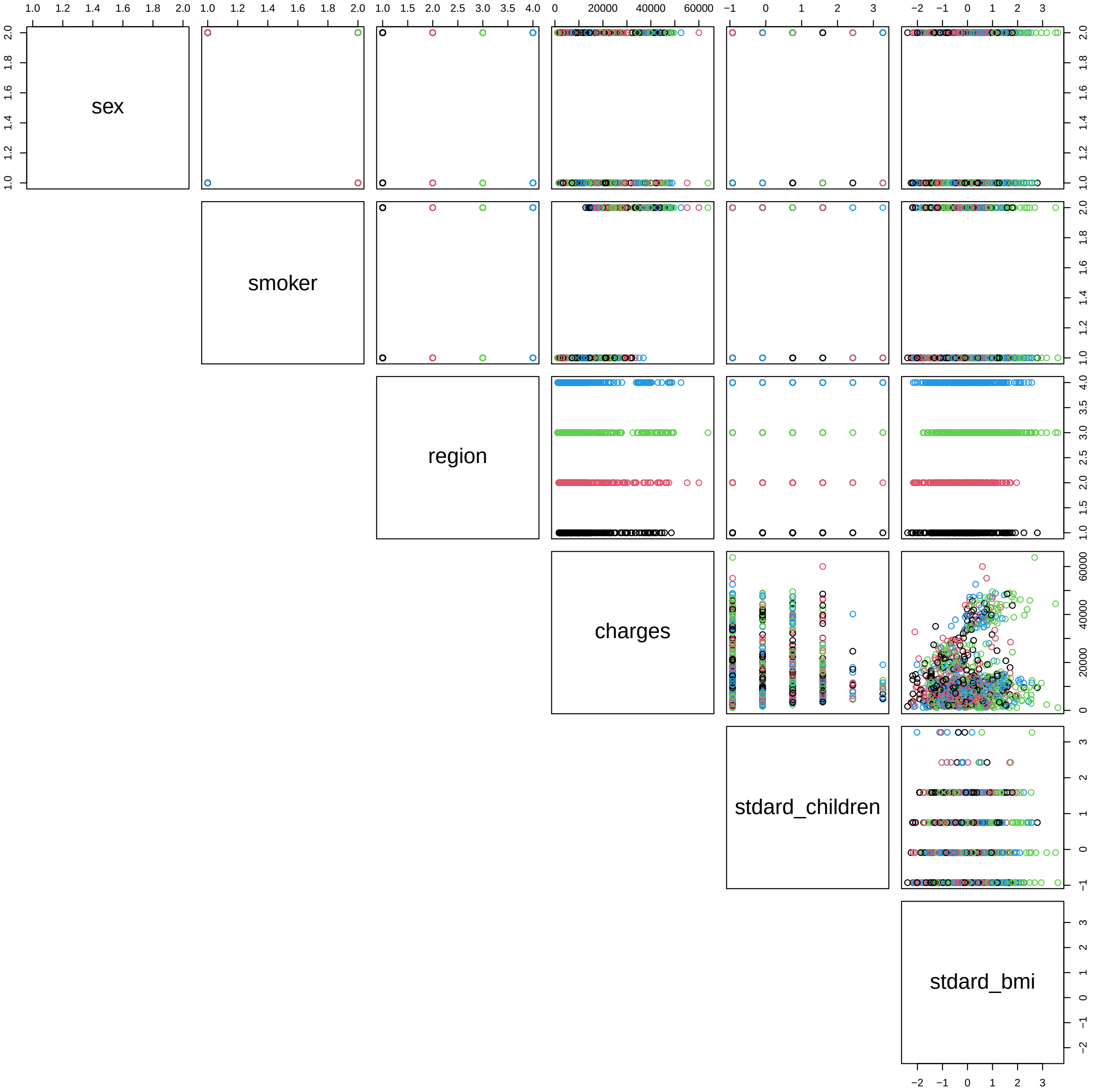
H0: Variable for have-children-or-not is independent from variable smoker-or-not (correlation = 0). H1: Variable smoker is not independent from variable children (correlations is not 0). The test p-value is 0.8868 > 0.05(by default), and we fail to reject H0.
Conclusion: the correlation between variable smoker and variable children is 0. Furthermore, there is no difference in smoking rates between those who have kids and those who do not.

5. Collinearity is a phenomenon in linear regression study. In order to find collinearity we first need to have a linear regression model.
   First check the overall data within matrix scatterplot.

standardize independent variables

```
insuData<-insuData%>%mutate(stdard_age = scale(age))
insuData<-insuData%>%mutate(stdard_children = scale(children))
insuData<-insuData%>%mutate(stdard_bmi = scale(bmi))
attach(insuData)
```

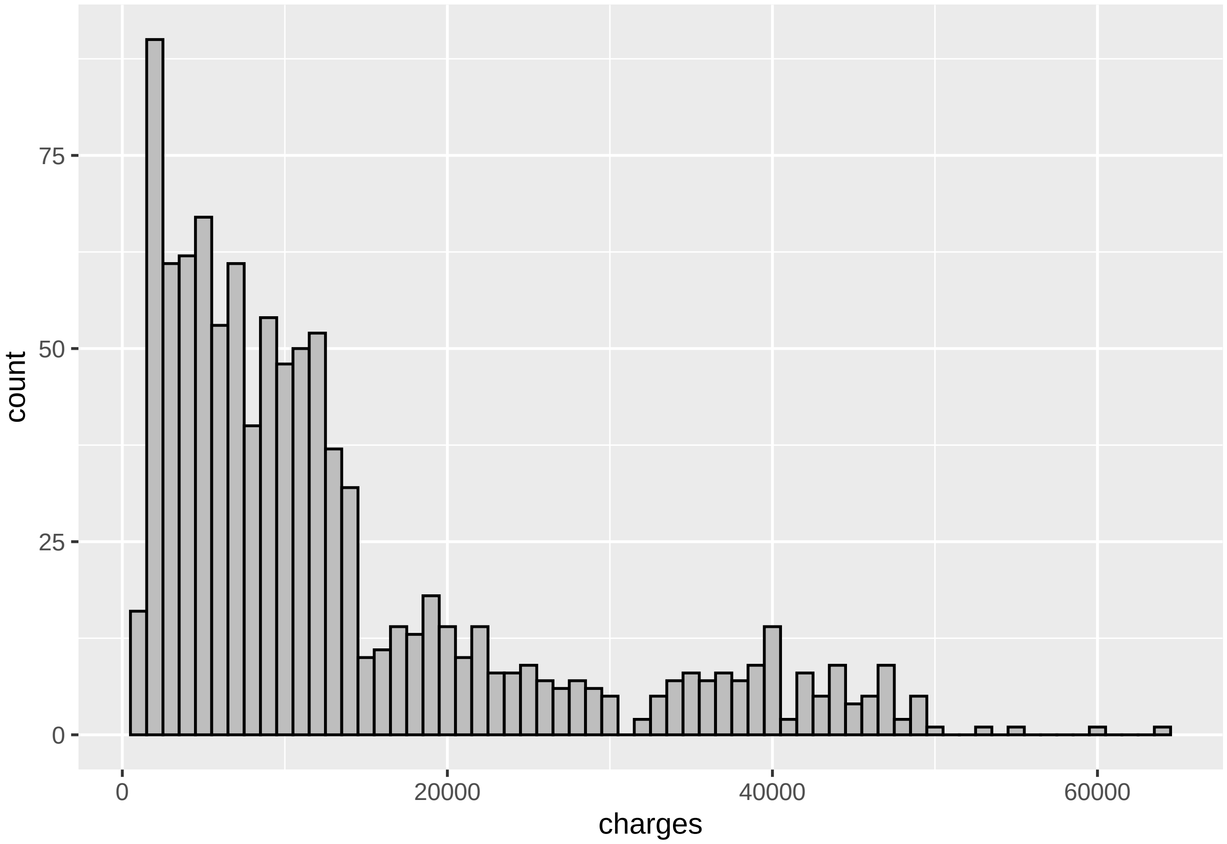
```
pairs(insuData[-c(1,3,4,8)], col=insuData$region, lower.panel = NULL, cex= 1.2)
```



We can see that seems age and charges may have a have linear relationship. Also, by looking at the figure, it seems that there are some relationships between 1.charges and bmi, 2.age and bmi, 3.age and children, 4.smoker and charges, 5.charges and children. When we put charges as a response variable with all other variables as predictor variables, there may be a multi/colinearity between exist between them. The relationships between other variables are not easily spotted from above figure. It makes most sense that we set charges as a response variable. In order to find instance of multi/collinearity we should start with building a linear regression model.

Looks at the charges histogram:

```
ggplot(insuData, aes(x = charges)) +
    geom_histogram(binwidth = 1000 ,fill="grey",colour="black")
```

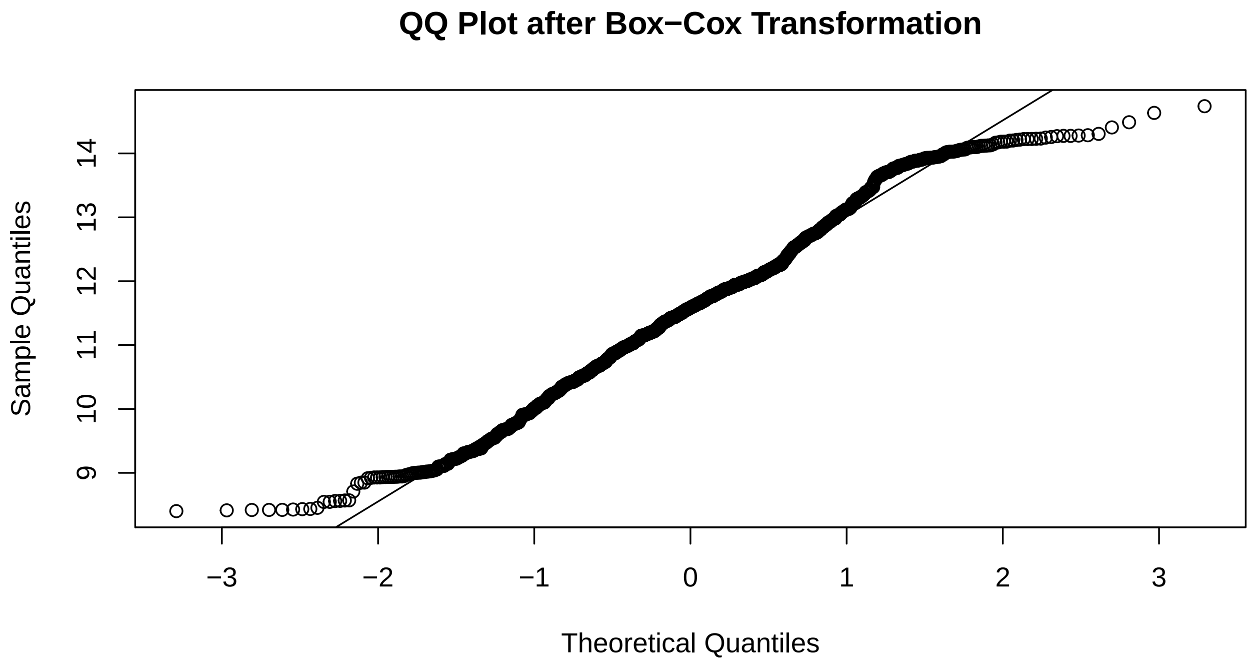


We will go through two multi-linear regression models: One model with a original dependent variable (charges), one with box-cox transformed dependent variable (bcCharge). Apply Box-Cox transformation on charges to make it "more" normal distributed. The reason to perform such transformation is that it is in generally easier to perform linear regression than non-linear.

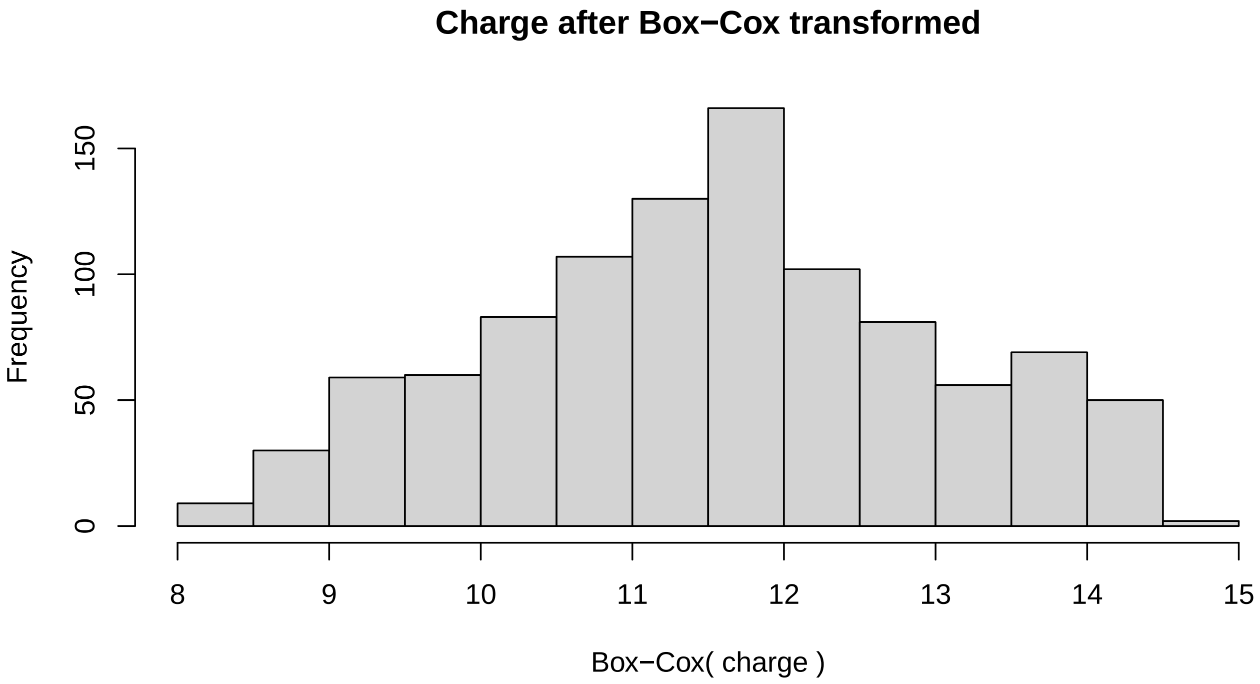
```
powerTransform(insuData$charges)
```

```
## Estimated transformation parameter
## insuData$charges
##       0.04959873
```

```
qqnorm(bcPower(insuData$charges,0.04959873), main="QQ Plot after Box-Cox Transformation")
qqline(bcPower(insuData$charges,0.04959873))
```



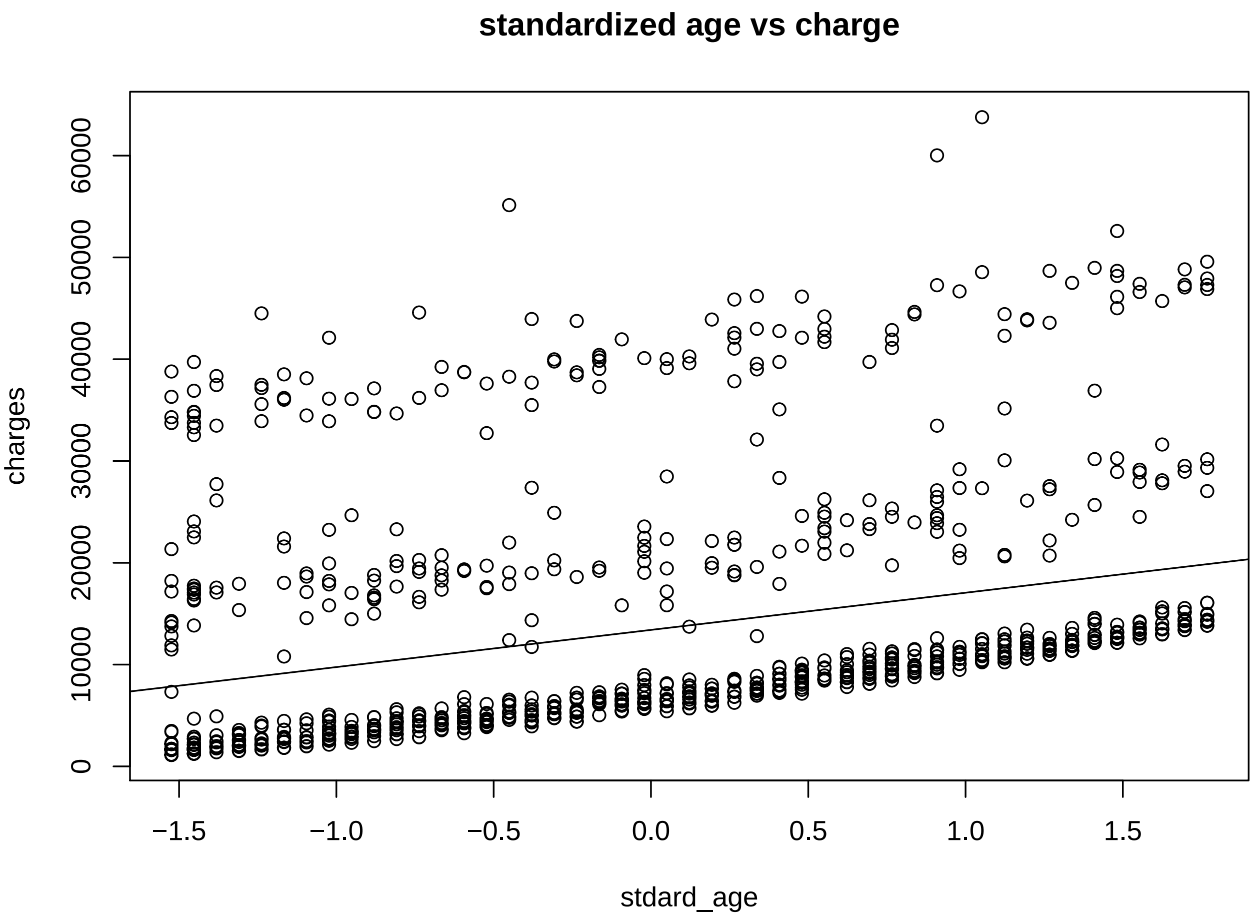
```r
hist(bcPower(insuData$charges,0.04959873),breaks = 22,
     main ="Charge after Box-Cox transformed",
     xlab= "Box-Cox( charge )")
```



```r
insuData<-cbind(insuData,"bcCharge"=bcPower(insuData$charges,0.04959873))
attach(insuData)
```

```
chg_age<- lm(charges~stdard_age,data=insuData)
plot(charges~stdard_age,data=insuData, main="standardized age vs charge")
abline(chg_age)
```

Original dependent variable: charges



```
# we can see three "lines" in the data.
```

We can see observations clustered into three groups and formed 3 lines.They are influential in the regression. Again we shouldn't categorize them into different age groups for the following reasons: 1.Loss of power and precision; 2.Don't know the "cut-off" point in reality; 3. Avoid overfitting.

```
# 1.charges and bmi, 2.age and bmi, 3.age and children, 4.smoker and charges, 5.charges and children.
# we can test 1, 2, 3, 5 in one regression below
cl.1<-lm(charges~stdard_bmi*stdard_age*stdard_children, data=insuData)
summary(cl.1)
```

```
##
## Call:
## lm(formula = charges ~ stdard_bmi * stdard_age *
stdard_children,
## data = insuData)
##
## Residuals:
## Min 1Q Median 3Q Max
## -13587 -7059 -5322 7597 42545
##
## Coefficients:
## Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13439.43 365.82 36.738 < 2e-16 ***
## stdard_bmi 1936.26 366.05 5.290 1.51e-07 ***
## stdard_age 3301.53 380.86 8.669 < 2e-16 ***
## stdard_children 529.93 370.76 1.429 0.153
## stdard_bmi:stdard_age -128.68 380.54 -0.338 0.735
## stdard_bmi:stdard_children 217.08 371.55 0.584 0.559
## stdard_age:stdard_children -342.37 393.78 -0.869 0.385
## stdard_bmi:stdard_age:stdard_children -66.99 408.91
-0.164 0.870
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
' ' 1
##
## Residual standard error: 11450 on 996 degrees of freedom
## Multiple R-squared: 0.1175, Adjusted R-squared: 0.1113
## F-statistic: 18.94 on 7 and 996 DF, p-value: < 2.2e-16
```

```r
vif(cl.1)
```

```
## stdard_bmi stdard_age
## 1.025087 1.109687
## stdard_children stdard_bmi:stdard_age
## 1.051616 1.145362
## stdard_bmi:stdard_children stdard_age:stdard_children
## 1.055660 1.129231
## stdard_bmi:stdard_age:stdard_children
## 1.190256
```

```r
pca<-princomp(insuData[c("stdard_bmi","stdard_age","stdard_children")])
# We can see VIFs > 1, but they are less than 5, which may indicate multi/collinearity exist, however,
# they are still in tolerant range that the regression model is not severely affected.
# The pca also support that there isn't any significant multicollinearity.
# We may still remove them based on their significant level #in the stepwise regression model.
# Lets remove the variate with highest p-value.
cl.2<-update(cl.1, charges~. -stdard_bmi:stdard_age:stdard_children)
summary(cl.2)
```

```
##
## Call:
## lm(formula = charges ~ stdard_bmi + stdard_age +
stdard_children +
## stdard_bmi:stdard_age + stdard_bmi:stdard_children +
stdard_age:stdard_children,
## data = insuData)
##
## Residuals:
## Min 1Q Median 3Q Max
## -13520 -7062 -5314 7576 42641
##
## Coefficients:
## Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13434.0 364.1 36.892 < 2e-16 ***
## stdard_bmi 1934.5 365.7 5.290 1.51e-07 ***
## stdard_age 3301.3 380.7 8.672 < 2e-16 ***
## stdard_children 517.0 362.1 1.428 0.154
## stdard_bmi:stdard_age -107.0 356.6 -0.300 0.764
## stdard_bmi:stdard_children 211.4 369.8 0.572 0.568
## stdard_age:stdard_children -345.0 393.3 -0.877 0.381
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
' ' 1
##
## Residual standard error: 11440 on 997 degrees of freedom
## Multiple R-squared: 0.1175, Adjusted R-squared: 0.1122
## F-statistic: 22.12 on 6 and 997 DF, p-value: < 2.2e-16
```

```
vif(cl.2)
```

```
##                 stdard_bmi                 stdard_age
##                   1.024168                   1.109672
##            stdard_children      stdard_bmi:stdard_age
##                   1.003964                   1.006768
## stdard_bmi:stdard_children stdard_age:stdard_children
##                   1.046585                   1.127346
```

```
AIC(cl.1)
```

```
## [1] 21625.49
```

```
AIC(cl.2)
```

```
## [1] 21623.52
```

```
# AIC dropped a small number, the model get a better fit to the data just for a little bit.
# lets move on and remove other non-significant variables.

cl.3<-update(cl.2, charges~. -stdard_bmi:stdard_age)
summary(cl.3)
```

```
##
## Call:
## lm(formula = charges ~ stdard_bmi + stdard_age +
stdard_children +
## stdard_bmi:stdard_children + stdard_age:stdard_children,
## data = insuData)
##
## Residuals:
## Min 1Q Median 3Q Max
## -13134 -7086 -5292 7572 42481
##
## Coefficients:
## Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13420.9 361.3 37.141 < 2e-16 ***
## stdard_bmi 1939.3 365.2 5.311 1.35e-07 ***
## stdard_age 3297.3 380.3 8.671 < 2e-16 ***
## stdard_children 512.4 361.6 1.417 0.157
## stdard_bmi:stdard_children 206.9 369.3 0.560 0.576
## stdard_age:stdard_children -344.5 393.1 -0.876 0.381
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
' ' 1
##
## Residual standard error: 11440 on 998 degrees of freedom
## Multiple R-squared: 0.1174, Adjusted R-squared: 0.113
## F-statistic: 26.55 on 5 and 998 DF, p-value: < 2.2e-16
```

```
vif(cl.3)
```

```
##                  stdard_bmi                  stdard_age
##                    1.022155                    1.108293
##             stdard_children stdard_bmi:stdard_children
##                    1.002197                    1.044798
## stdard_age:stdard_children
##                    1.127322
```

```
AIC(cl.2)
```

```
## [1] 21623.52
```

```
AIC(cl.3)
```

```
## [1] 21621.61
```

```
cl.4<-update(cl.3, charges~. -stdard_bmi:stdard_children)
summary(cl.4)
```

```
##
## Call:
## lm(formula = charges ~ stdard_bmi + stdard_age +
stdard_children +
## stdard_age:stdard_children, data = insuData)
##
## Residuals:
## Min 1Q Median 3Q Max
## -13287 -7082 -5279 7770 42325
##
## Coefficients:
## Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13419.9 361.2 37.152 < 2e-16 ***
## stdard_bmi 1935.4 365.0 5.303 1.41e-07 ***
## stdard_age 3318.3 378.3 8.772 < 2e-16 ***
## stdard_children 507.7 361.4 1.405 0.160
## stdard_age:stdard_children -300.1 384.9 -0.780 0.436
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
' ' 1
##
## Residual standard error: 11440 on 999 degrees of freedom
## Multiple R-squared: 0.1171, Adjusted R-squared: 0.1136
## F-statistic: 33.13 on 4 and 999 DF, p-value: < 2.2e-16
```

```
vif(cl.4)
```

```
##                  stdard_bmi                  stdard_age
##                    1.021777                    1.097531
##             stdard_children stdard_age:stdard_children
##                    1.001641                   1.081487
```

```
AIC(cl.3)
```

```
## [1] 21621.61
```

```
AIC(cl.4)
```

```
## [1] 21619.93
```

```
cl.5<-update(cl.4, charges~. -stdard_age:stdard_children)
summary(cl.5)
```

```
##
## Call:
## lm(formula = charges ~ stdard_bmi + stdard_age +
stdard_children,
## data = insuData)
##
## Residuals:
## Min 1Q Median 3Q Max
## -13497 -7087 -5323 7763 42263
##
## Coefficients:
## Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13408.1 360.8 37.159 < 2e-16 ***
## stdard_bmi 1912.7 363.8 5.258 1.78e-07 ***
## stdard_age 3398.2 364.0 9.335 < 2e-16 ***
## stdard_children 505.0 361.3 1.398 0.162
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
' ' 1
##
## Residual standard error: 11430 on 1000 degrees of
freedom
## Multiple R-squared: 0.1166, Adjusted R-squared: 0.1139
## F-statistic: 43.98 on 3 and 1000 DF, p-value: < 2.2e-16
```

```
vif(cl.5)
```

```
##      stdard_bmi      stdard_age stdard_children
##        1.015260        1.016822        1.001552
```

```
AIC(cl.4)
```

```
## [1] 21619.93
```

```
AIC(cl.5)
```

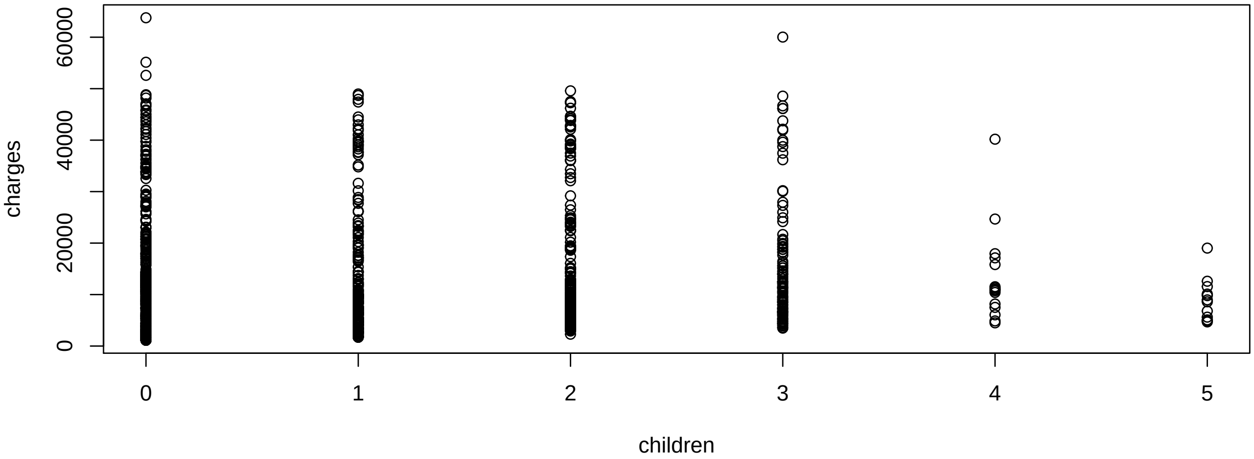
```
## [1] 21618.54
```

```
#so far we have removed all interaction variables, the only variables is insignificant is children.
#lets check charge and children scatterplot again.
chg_chld<- lm(charges~stdard_children,data = insuData)
summary(chg_chld)
```

```
##
## Call:
## lm(formula = charges ~ stdard_children, data = insuData)
##
```

```
## Residuals:
## Min 1Q Median 3Q Max
## -11689 -8795 -4040 3773 50960
##
## Coefficients:
## Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13408.1 383.0 35.010 <2e-16 ***
## stdard_children 645.3 383.2 1.684 0.0924 .
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
' ' 1
##
## Residual standard error: 12130 on 1002 degrees of
freedom
## Multiple R-squared: 0.002823, Adjusted R-squared:
0.001828
## F-statistic: 2.837 on 1 and 1002 DF, p-value: 0.09244
```

```
plot(charges~children,data=insuData)
```



```
#children is not significant to charges.
cl.6<-update(cl.5,charges~. -stdard_children)
summary(cl.6)
```

```
##
## Call:
## lm(formula = charges ~ stdard_bmi + stdard_age, data =
insuData)
##
## Residuals:
## Min 1Q Median 3Q Max
## -13932 -7133 -5266 7677 42365
##
## Coefficients:
## Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13408.1 361.0 37.141 < 2e-16 ***
## stdard_bmi 1912.0 363.9 5.254 1.82e-07 ***
## stdard_age 3418.2 363.9 9.393 < 2e-16 ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
```

```
' ' 1
##
## Residual standard error: 11440 on 1001 degrees of
freedom
## Multiple R-squared: 0.1148, Adjusted R-squared: 0.1131
## F-statistic: 64.94 on 2 and 1001 DF, p-value: < 2.2e-16
```

```
vif(cl.6)
```

```
## stdard_bmi stdard_age
##   1.015258   1.015258
```

```
AIC(cl.5)
```

```
## [1] 21618.54
```

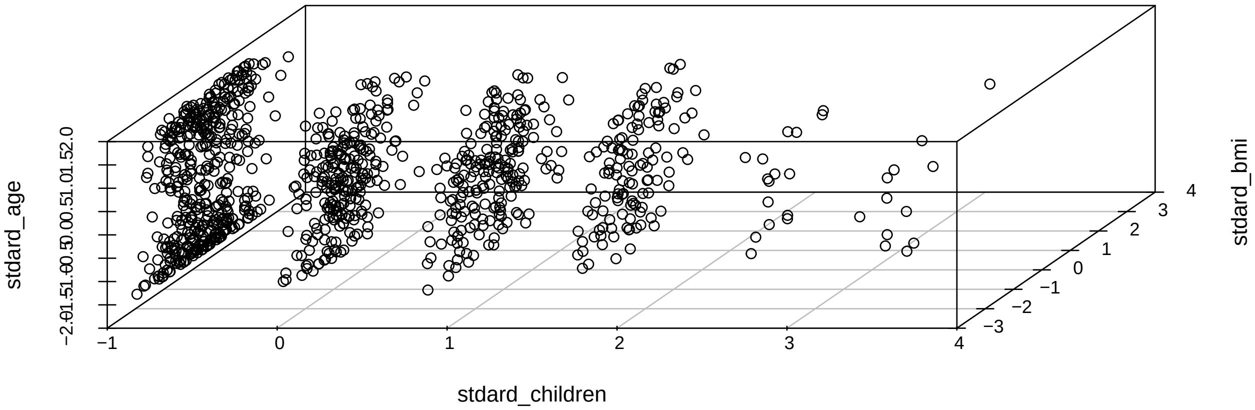
```
AIC(cl.6)
```

```
## [1] 21618.5
```

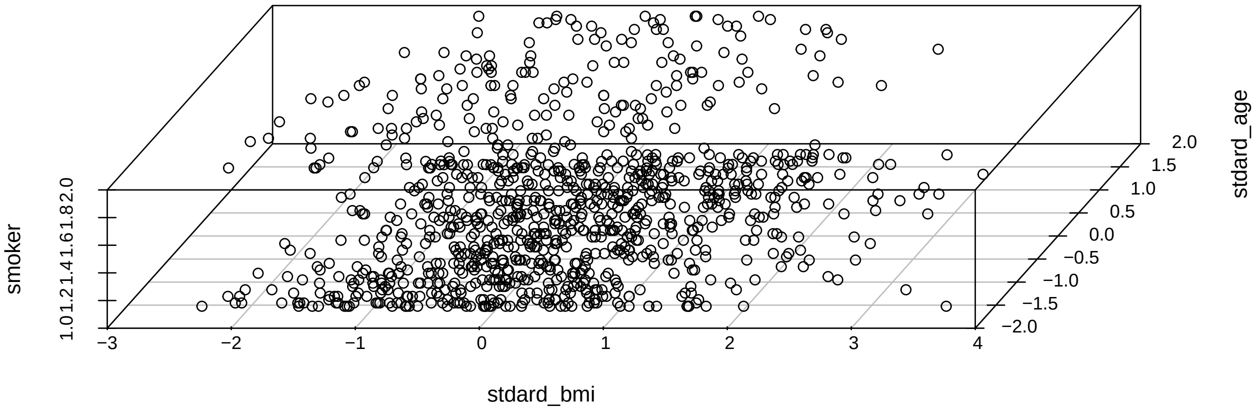
```
#That is all insignificant variable we can remove,
#and the AIC doesn't change much which mean this small linear model can't get better.
#Next we move on to model of smoker and charges.
e.1<- lm(charges~smoker,data=insuData)
summary(e.1)
```

```
##
## Call:
## lm(formula = charges ~ smoker, data = insuData)
##
## Residuals:
## Min 1Q Median 3Q Max
## -18986 -4968 -1031 3772 31955
##
## Coefficients:
## Estimate Std. Error t value Pr(>|t|)
## (Intercept) 8451.3 266.7 31.69 <2e-16 ***
## smokeryes 23364.4 579.0 40.36 <2e-16 ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
' ' 1
##
## Residual standard error: 7500 on 1002 degrees of freedom
## Multiple R-squared: 0.6191, Adjusted R-squared: 0.6187
## F-statistic: 1629 on 1 and 1002 DF, p-value: < 2.2e-16
```

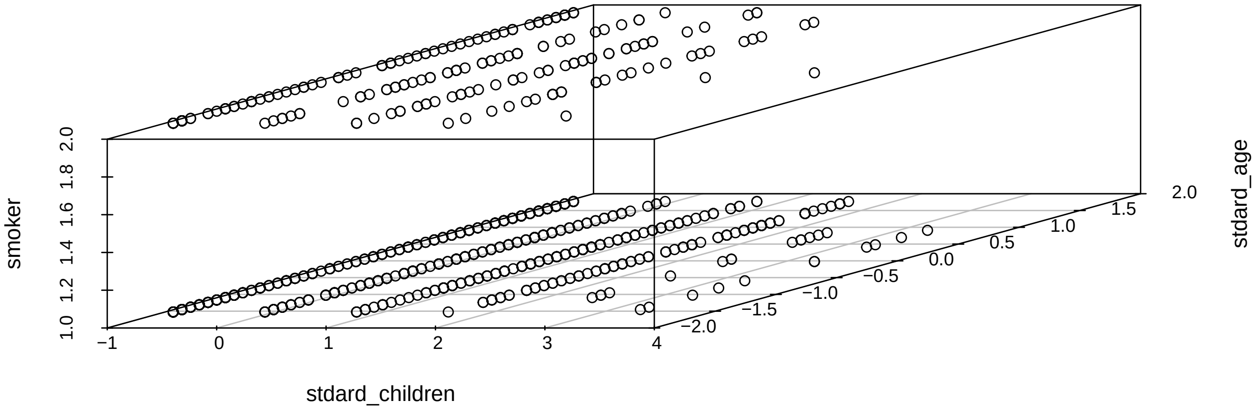
```
#smoker is significant is the simple model.
# We may combine looking for collinearity between smoker and other variabls.
scatterplot3d(stdard_children,stdard_bmi,stdard_age,angle = 75)
```



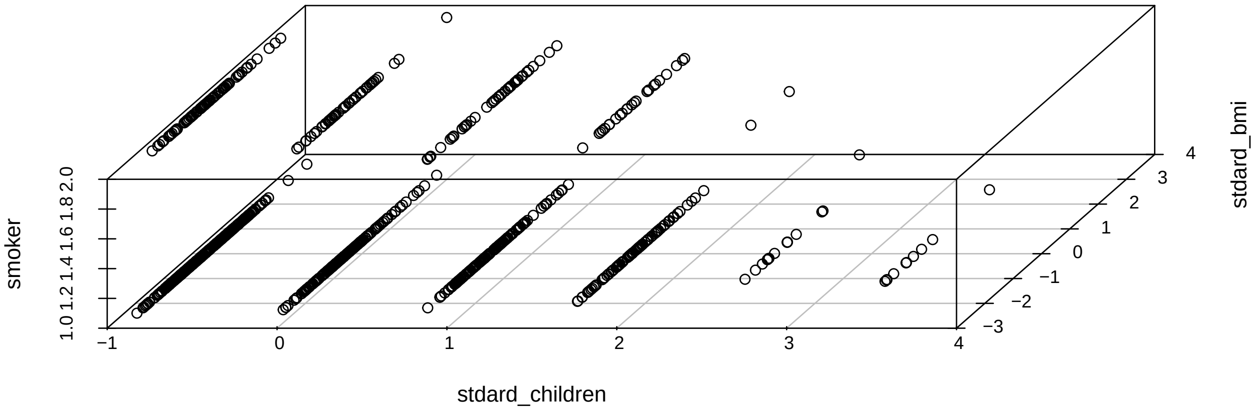
```
#this 3D-plot also proved what we find that there isn't much
#multi/collinearity between these three variables.
#We may plot other 3D-plot with different variables.
scatterplot3d(stdard_bmi,stdard_age,smoker,angle = 75)
```



```
scatterplot3d(stdard_children,stdard_age,smoker)
```



```
scatterplot3d(stdard_children,stdard_bmi,smoker, angle=75)
```



```
#that's all 3 varibale combos we can have,
#and there is no evidence of multi/collinearity. we can check 4-dimension-combo.
e.2<-lm(charges~smoker*stdard_age*stdard_bmi*stdard_children,data=insuData)

#notice that some of insignificant p-values of previous tests have turned into significant,
#this may indicate that in high dimension, multicollinearity exist among smoker:bmi and child.
# However, none of the variables have high vif(>5).
#so we can ignore them. Next we can fit our full model.
f.1<- lm(charges~ stdard_bmi + stdard_age + smoker + sex + region, data=insuData)

# sex is not significant (in this model !).
# So being male does not significantly affects the charges.
# remove sex
f.2<- lm(charges~ stdard_bmi + stdard_age + smoker + region, data=insuData)
summary(f.2)
```

```
##
## Call:
## lm(formula = charges ~ stdard_bmi + stdard_age + smoker
+ region,
## data = insuData)
##
## Residuals:
## Min 1Q Median 3Q Max
## -11861 -3002 -1013 1590 24216
##
## Coefficients:
## Estimate Std. Error t value Pr(>|t|)
## (Intercept) 9117.4 418.4 21.789 <2e-16 ***
## stdard_bmi 2045.4 201.3 10.162 <2e-16 ***
## stdard_age 3706.8 193.6 19.142 <2e-16 ***
## smokeryes 23725.3 471.1 50.359 <2e-16 ***
## regionnorthwest -517.1 559.5 -0.924 0.3556
## regionsoutheast -1293.8 561.9 -2.303 0.0215 *
## regionsouthwest -1020.9 557.5 -1.831 0.0674 .
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
' ' 1
##
## Residual standard error: 6076 on 997 degrees of freedom
## Multiple R-squared: 0.7513, Adjusted R-squared: 0.7498
## F-statistic: 501.9 on 6 and 997 DF,  p-value: < 2.2e-16
```

```
contrasts(region)
```

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

```
vif(f.2)
```

```
##                 GVIF Df GVIF^(1/(2*Df))
## stdard_bmi 1.100866  1        1.049222
## stdard_age 1.018850  1        1.009381
## smoker     1.009005  1        1.004492
## region     1.093601  3        1.015024
```

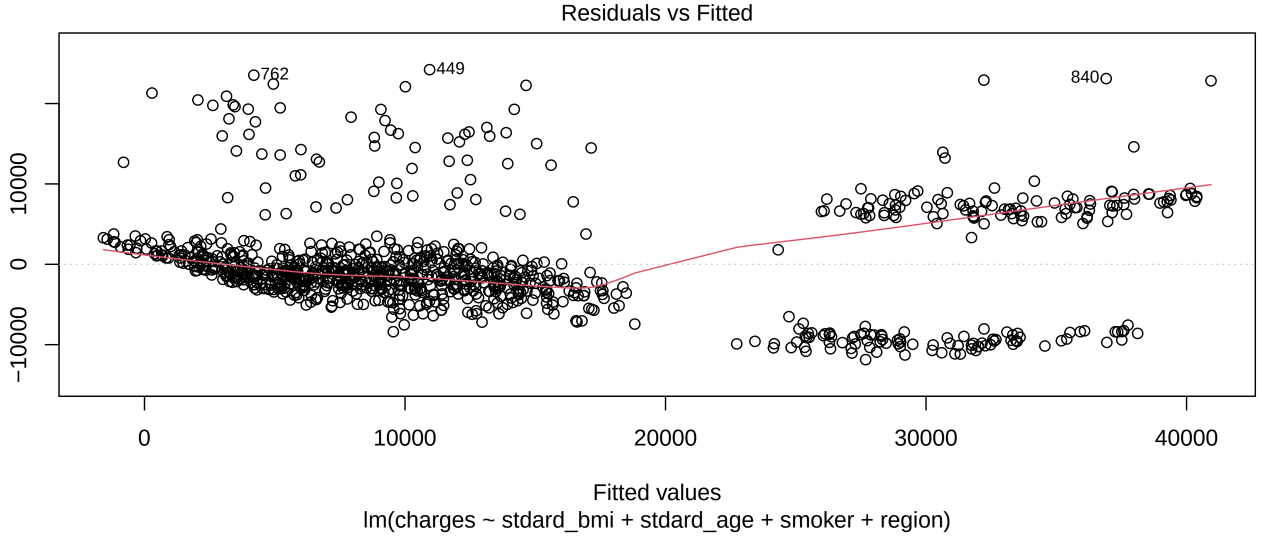
```
AIC(f.1)
```

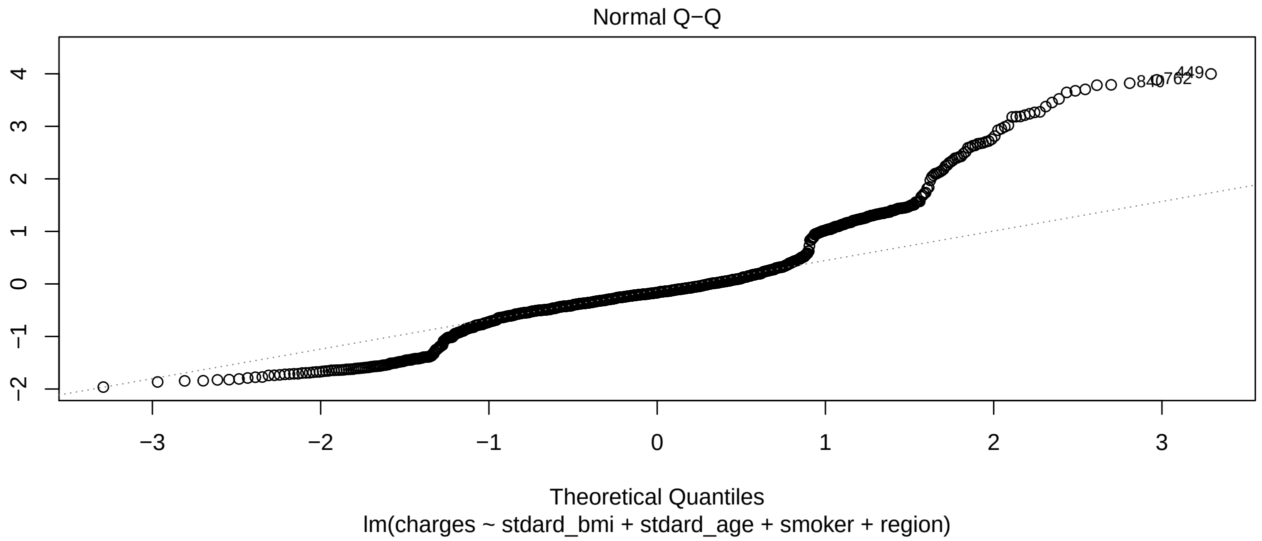
```
## [1] 20354
```

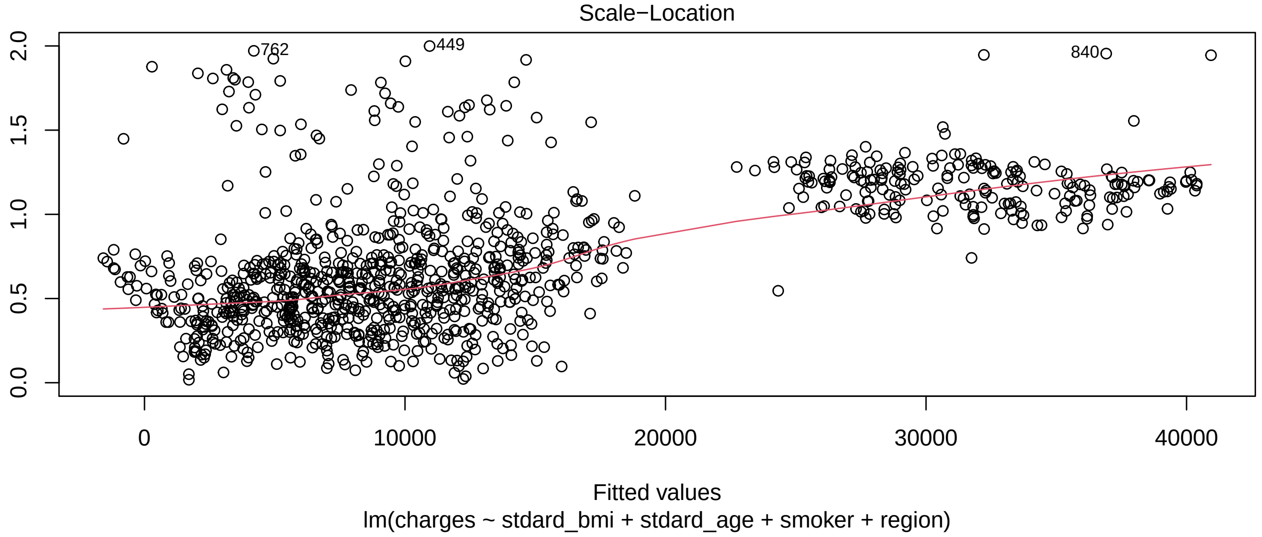
```
AIC(f.2)
```

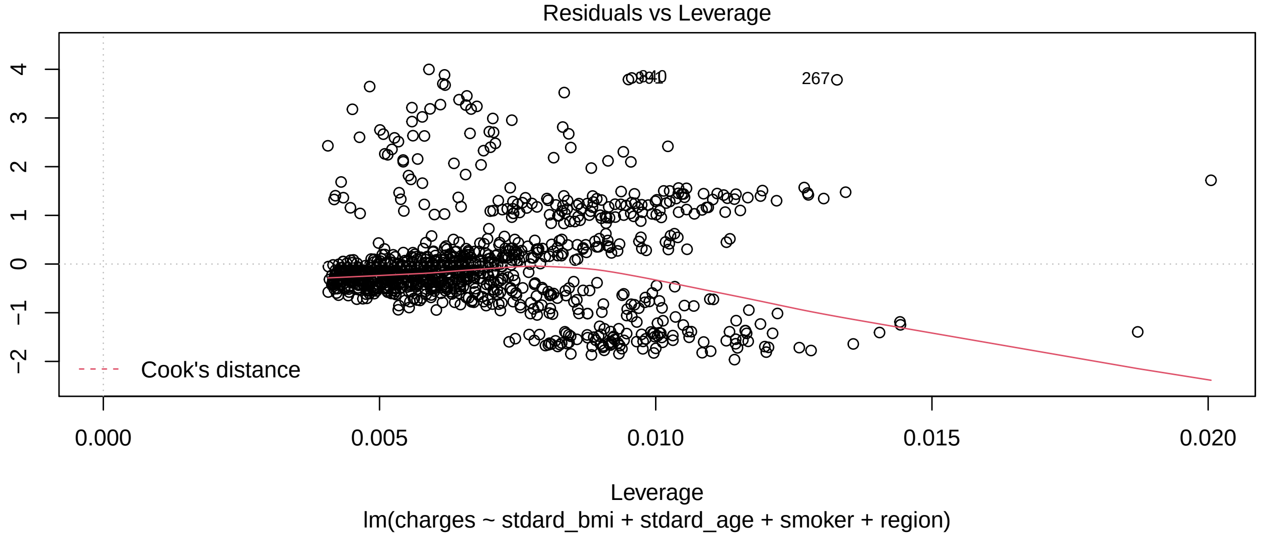
```
## [1] 20352.01
```

```
plot(f.2)
```









```
#region is a categorical data, the reference level is northeast.
#Since the reference level is significant, and we can
#not remove region. This is the best model so for we can find.
```

Diagnose: There is no strong evidence of significant multi/collinearity. The variance inflation factor stay under 5 consistently.
Assumptions and validations: 1. Linearity. The linear regression model requires four assumptions.
1.Linearity. We can check the "Residual vs Fitted" value, where data should spread within +/-2 standard deviation if linearity hold. We can roughly see that most of the data are within these range, with some points not.

2. Homoscedasticity. We can check figure "Residual vs Fitted", and find pattern that points are cluster into three major groups. Also, data are not equally spread around 0 line. This indicates that most of data are divided by three major types of spread with non-constant variances. Homoscedasticity does not hold.
3. Independence of observation. This has can be checked at the study design. We want the samples are randomly collected or generated to have this property hold, so that our conclusion truly describe population.
4. Normality. We may check it from a qq plot. If the distribution is perfectly normal, we ould see the data points follows a straight line. In the above "Normal Q-Q" graph we can see that data points are partially following a straight line, there is a large portion of data deviates from the dashed line. Overall, the normality assumption is violated.

We may remedy the violations on these assumptions by box-cox transform response variable. Let's try it.

```r
#check collinearity and multi-collinearity
lmFull <- lm(bcCharge ~ stdard_age +sex+smoker+stdard_bmi+stdard_children +region,data=insuData)
summary(lmFull)
```

Box-Cox transformed dependent (bcCharge)

```
##
## Call:
## lm(formula = bcCharge ~ stdard_age + sex + smoker +
stdard_bmi +
## stdard_children + region, data = insuData)
##
## Residuals:
## Min 1Q Median 3Q Max
## -1.5334 -0.3121 -0.0841 0.0937 3.4177
##
## Coefficients:
## Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.24504 0.05160 217.933 < 2e-16 ***
## stdard_age 0.74485 0.02188 34.042 < 2e-16 ***
## sexmale -0.12011 0.04356 -2.757 0.005938 **
## smokeryes 2.48680 0.05330 46.658 < 2e-16 ***
## stdard_bmi 0.12094 0.02273 5.320 1.28e-07 ***
## stdard_children 0.19549 0.02171 9.007 < 2e-16 ***
## regionnorthwest -0.13975 0.06313 -2.214 0.027072 *
## regionsoutheast -0.25875 0.06335 -4.084 4.77e-05 ***
## regionsouthwest -0.21754 0.06289 -3.459 0.000565 ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
' ' 1
##
## Residual standard error: 0.685 on 995 degrees of freedom
## Multiple R-squared: 0.7761, Adjusted R-squared: 0.7743
## F-statistic: 431.2 on 8 and 995 DF, p-value: < 2.2e-16
```

```
vif(lmFull)
```

```
##                       GVIF Df GVIF^(1/(2*Df))
## stdard_age      1.023414  1        1.011639
## sex             1.015066  1        1.007505
## smoker          1.016027  1        1.007982
## stdard_bmi      1.104918  1        1.051151
## stdard_children 1.007116  1        1.003552
## region          1.098325  3        1.015754
```

```
lmbcC_sex_age <-lm(bcCharge~ stdard_age*sex,data=insuData)
summary(lmbcC_sex_age)
```

```
##
## Call:
## lm(formula = bcCharge ~ stdard_age * sex, data =
insuData)
##
## Residuals:
## Min 1Q Median 3Q Max
## -2.0240 -0.7285 -0.4854 0.8388 3.4781
##
## Coefficients:
## Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.51601 0.05575 206.552 <2e-16 ***
## stdard_age 0.71571 0.05596 12.791 <2e-16 ***
## sexmale 0.07135 0.07838 0.910 0.363
## stdard_age:sexmale 0.04674 0.07842 0.596 0.551
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
' ' 1
##
## Residual standard error: 1.24 on 1000 degrees of freedom
## Multiple R-squared: 0.2626, Adjusted R-squared: 0.2604
## F-statistic: 118.7 on 3 and 1000 DF, p-value: < 2.2e-16
```

```
vif(lmbcC_sex_age)
```

```
##     stdard_age            sex stdard_age:sex
##       2.042134       1.002513       2.039472
```

```
#high VIF for interaction between sex and age: collinearity, Remove sex.

lmF_sex<-update(lmFull, bcCharge~.-sex)
summary(lmF_sex)
```

```
##
## Call:
## lm(formula = bcCharge ~ stdard_age + smoker + stdard_bmi
+ stdard_children +
## region, data = insuData)
##
## Residuals:
## Min 1Q Median 3Q Max
## -1.5780 -0.3032 -0.0880 0.0961 3.3636
```

```
##
## Coefficients:
## Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.18745 0.04734 236.331 < 2e-16 ***
## stdard_age 0.74826 0.02192 34.140 < 2e-16 ***
## smokeryes 2.47465 0.05329 46.436 < 2e-16 ***
## stdard_bmi 0.11720 0.02277 5.147 3.18e-07 ***
## stdard_children 0.19340 0.02176 8.886 < 2e-16 ***
## regionnorthwest -0.13928 0.06334 -2.199 0.028109 *
## regionsoutheast -0.25994 0.06356 -4.090 4.67e-05 ***
## regionsouthwest -0.21909 0.06309 -3.472 0.000538 ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
' ' 1
##
## Residual standard error: 0.6872 on 996 degrees of
freedom
## Multiple R-squared: 0.7744, Adjusted R-squared: 0.7729
## F-statistic: 488.5 on 7 and 996 DF, p-value: < 2.2e-16
```

```
vif(lmF_sex)
```

```
##                     GVIF Df GVIF^(1/(2*Df))
## stdard_age      1.020154  1        1.010027
## smoker          1.009082  1        1.004531
## stdard_bmi      1.100983  1        1.049277
## stdard_children 1.005877  1        1.002934
## region          1.098119  3        1.015722
```

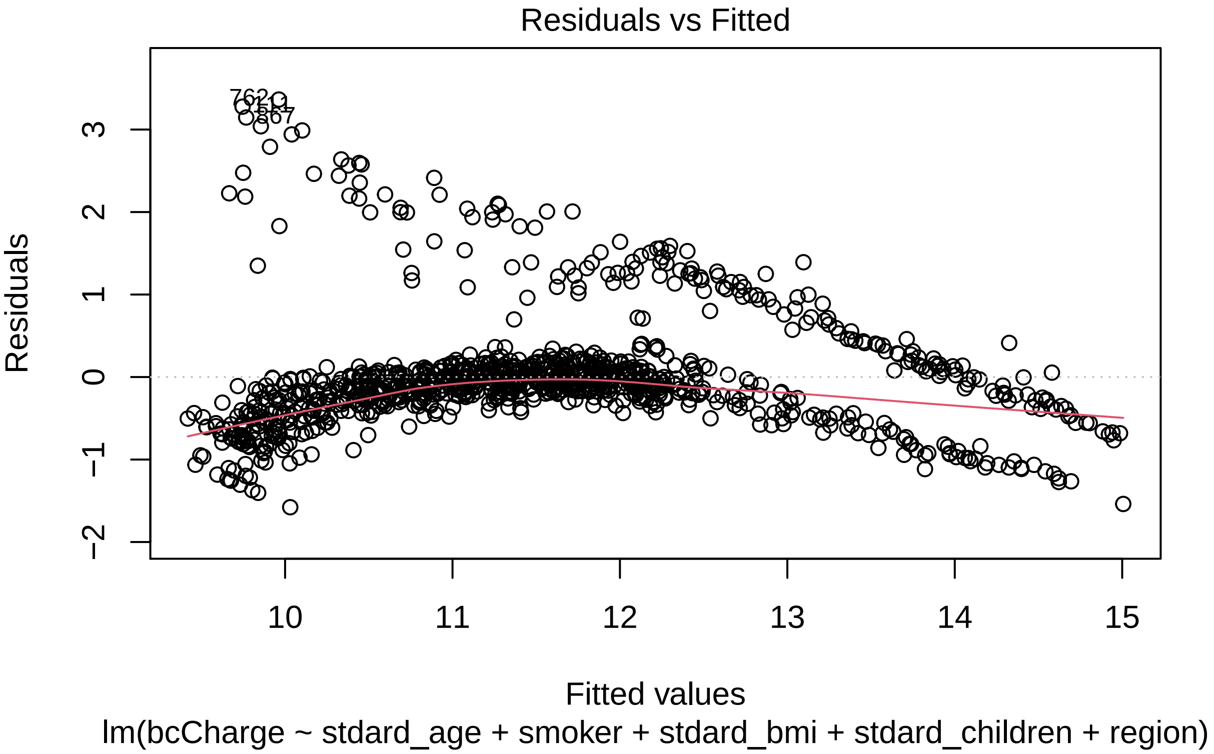
```
# Next we should check collinearity for bmi and age
# because we found that there might be linear relationship
# between 1.age and bmi, 2.bmi and charges,
# 3.age and charges from the previous matrix scatterplot.

temp<-lm(bcCharge ~ stdard_age + smoker + stdard_bmi + stdard_children +
    region,data=insuData)
summary(temp)
```

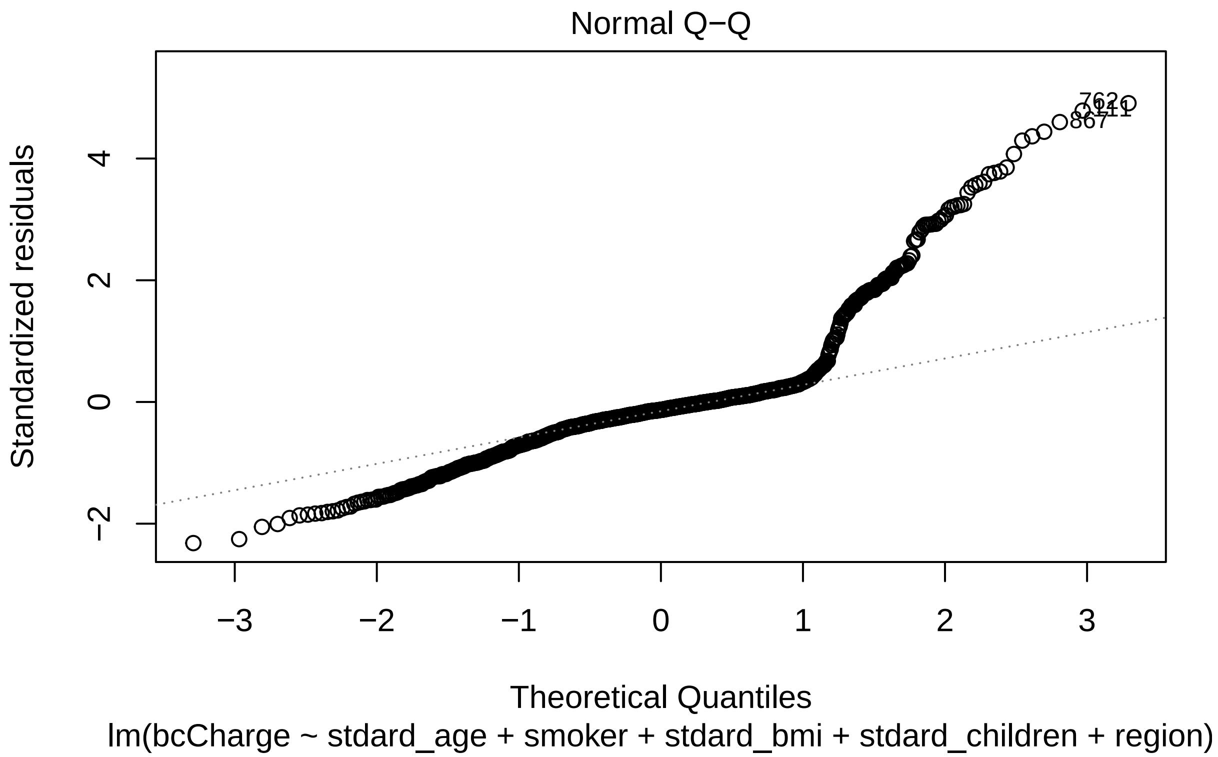
```
##
## Call:
## lm(formula = bcCharge ~ stdard_age + smoker + stdard_bmi
+ stdard_children +
## region, data = insuData)
##
## Residuals:
## Min 1Q Median 3Q Max
## -1.5780 -0.3032 -0.0880 0.0961 3.3636
##
## Coefficients:
## Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.18745 0.04734 236.331 < 2e-16 ***
## stdard_age 0.74826 0.02192 34.140 < 2e-16 ***
## smokeryes 2.47465 0.05329 46.436 < 2e-16 ***
## stdard_bmi 0.11720 0.02277 5.147 3.18e-07 ***
## stdard_children 0.19340 0.02176 8.886 < 2e-16 ***
## regionnorthwest -0.13928 0.06334 -2.199 0.028109 *
## regionsoutheast -0.25994 0.06356 -4.090 4.67e-05 ***
## regionsouthwest -0.21909 0.06309 -3.472 0.000538 ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
' ' 1
##
## Residual standard error: 0.6872 on 996 degrees of
freedom
## Multiple R-squared: 0.7744, Adjusted R-squared: 0.7729
## F-statistic: 488.5 on 7 and 996 DF, p-value: < 2.2e-16
```

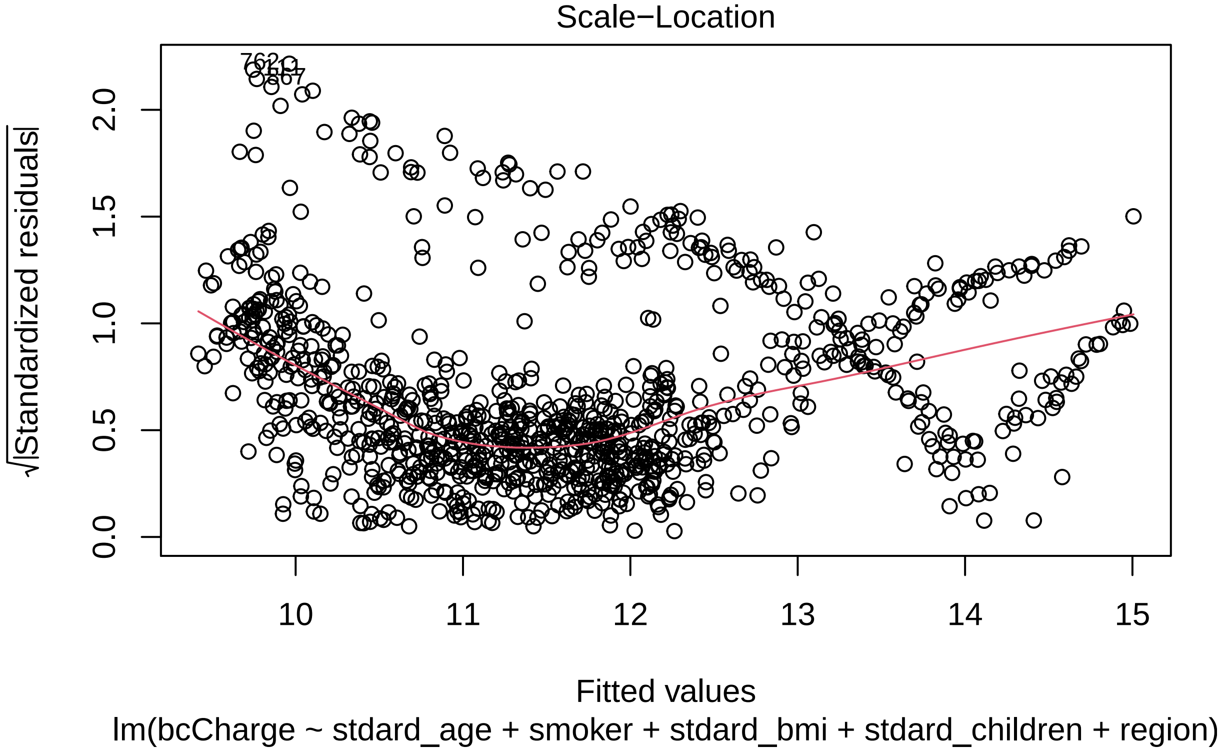
```
AIC(temp)
```

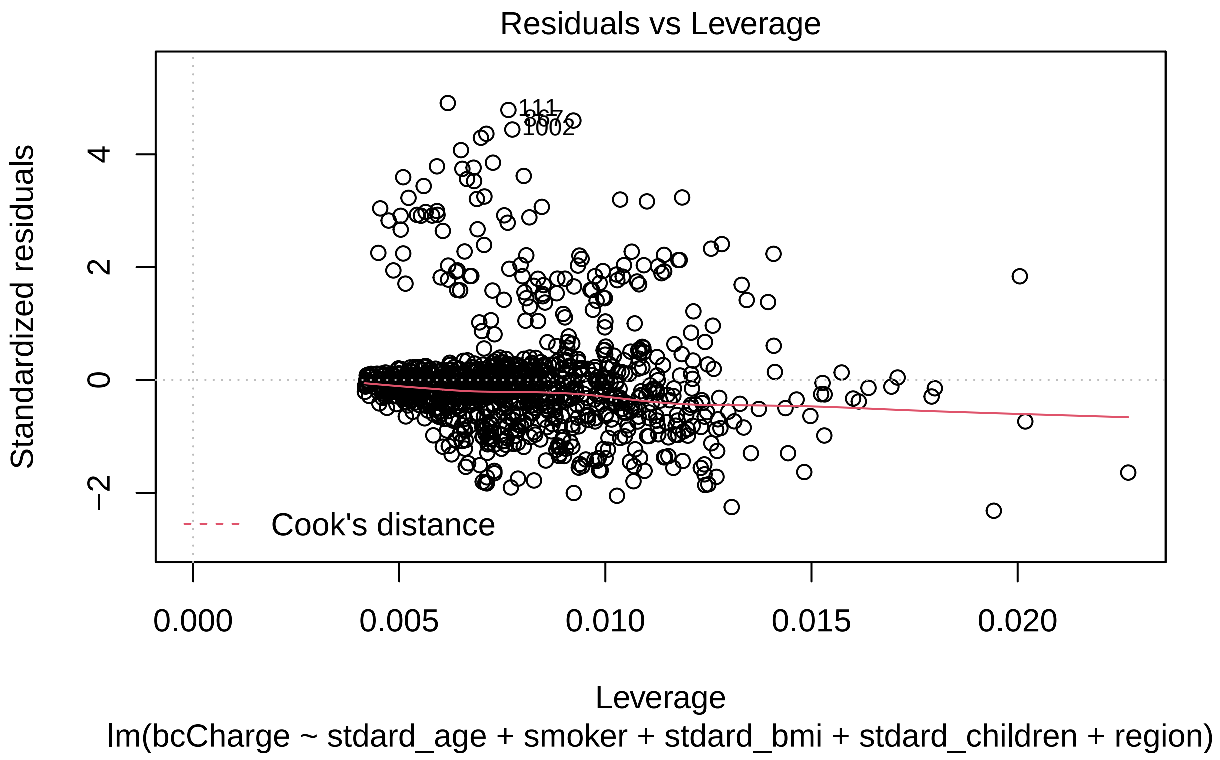
```
## [1] 2106.055
```

```
plot(temp)
```




Normal Q−Q
Standardized residuals
−2
0
2
4
−3
−2
−1
0
1
2
3
762
1211
867
Theoretical Quantiles
lm(bcCharge ~ stdard_age + smoker + stdard_bmi + stdard_children + region)


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
10
11
12
13
14
15
762
1211
867
Fitted values
lm(bcCharge ~ stdard_age + smoker + stdard_bmi + stdard_children + region)



Diagnose: In the tuning process, we identified the evidence of collinearity between sex and age from high VIF value (>2 after standardize). By checking the figure "residual vs fitted" we still can see two clusters of points. We may also spot that the "lower" cluster is slightly curved indicates there might be a quadratic relationship not yet addressed by the model. Moreover, it is likely a multivariate regression might fit the data better if the we know more about the dependent variable. There are pattern in this plot the homoscedasticity does not hold. From the same graph, we can find that most of the data points a within 2 standard deviation range, so the linearity after boxcox transformation hold. Again, we can only check the independence from the study design. From the Q-Q plot, the points are not following a straight line, so normality does not hold. In general, the data fit poorly with linear regression model. I recommend try multinomial regression.

Automatic model selection: Let's see what's might be the best model give by automatic model selection.

```
#We can also use automatic model selection form leaps to check what variables fit the data most.
data1<-
  insuData[c("stdard_age", "stdard_children", "stdard_bmi","smoker","region","sex","bcCharge")]

variableAutoSelect<-
  ols_step_best_subset(lm(bcCharge~stdard_age*stdard_children*stdard_bmi+smoker+region+sex,
                        data=data1))
variableAutoSelect
```

```
## Best Subsets Regression
##
---------------------------------------------------------------------------------------------------------
## Model Index Predictors
##
---------------------------------------------------------------------------------------------------------
## 1 smoker
## 2 stdard_age smoker
## 3 stdard_age stdard_children smoker
## 4 stdard_age stdard_children stdard_bmi smoker
## 5 stdard_age stdard_children stdard_bmi smoker region
## 6 stdard_age stdard_children stdard_bmi smoker region
stdard_age:stdard_children
## 7 stdard_age stdard_children stdard_bmi smoker region
sex stdard_age:stdard_children
## 8 stdard_age stdard_children stdard_bmi smoker region
sex stdard_age:stdard_children stdard_age:stdard_bmi
## 9 stdard_age stdard_children stdard_bmi smoker region
sex stdard_age:stdard_children stdard_age:stdard_bmi
stdard_children:stdard_bmi
## 10 stdard_age stdard_children stdard_bmi smoker region
sex stdard_age:stdard_children stdard_age:stdard_bmi
stdard_children:stdard_bmi
stdard_age:stdard_children:stdard_bmi
##
-------------------------------------------------------------------------------------------------------
##
## Subsets Regression Summary
##
-------------------------------------------------------------------------------------------------------
## Adj.  Pred
## Model R-Square R-Square R-Square C(p) AIC SBIC SBC MSEP
FPE HSP APC
##
-------------------------------------------------------------------------------------------------------
## 1 0.4638 0.4632 0.4621 1418.1440 2963.4769 74.3125
2978.2122 1120.5315 1.1183 0.0011 0.5384
## 2 0.7478 0.7473 0.746 139.4227 2208.2245 -900.4069
2227.8715 527.5916 0.5271 5e-04 0.2537
## 3 0.7657 0.7650 0.7636 60.6437 2136.2626 -995.6747
2160.8213 490.6130 0.4906 5e-04 0.2362
## 4 0.7701 0.7692 0.7676 42.7985 2119.2267 -998.0857
2148.6972 481.8810 0.4823 5e-04 0.2322
## 5 0.7744 0.7729 0.7706 25.1886 2106.0553 -1000.1668
2150.2611 473.2409 0.4751 5e-04 0.2283
## 6 0.7776 0.7758 0.7734 12.9276 2093.8796 -993.0711
2142.9971 467.0746 0.4694 5e-04 0.2255
## 7 0.7797 0.7777 0.775 5.6305 2086.5292 -977.1891
2140.5584 463.2099 0.4660 5e-04 0.2239
## 8 0.7800 0.7778 0.7748 5.9300 2086.8094 -948.0422
2145.7504 462.8823 0.4661 5e-04 0.2239
## 9 0.7802 0.7777 0.7746 7.3987 2088.2715 -916.8483
2152.1242 463.1003 0.4668 5e-04 0.2243
## 10 0.7802 0.7776 0.7742 9.0000 2089.8677 -885.4044
2158.6321 463.3807 0.4675 5e-04 0.2246
##
-------------------------------------------------------------------------------------------------------
## AIC: Akaike Information Criteria
## SBIC: Sawa's Bayesian Information Criteria
## SBC: Schwarz Bayesian Criteria
## MSEP: Estimated error of prediction, assuming
multivariate normality
## FPE: Final Prediction Error
## HSP: Hocking's Sp
## APC: Amemiya Prediction Criteria
```

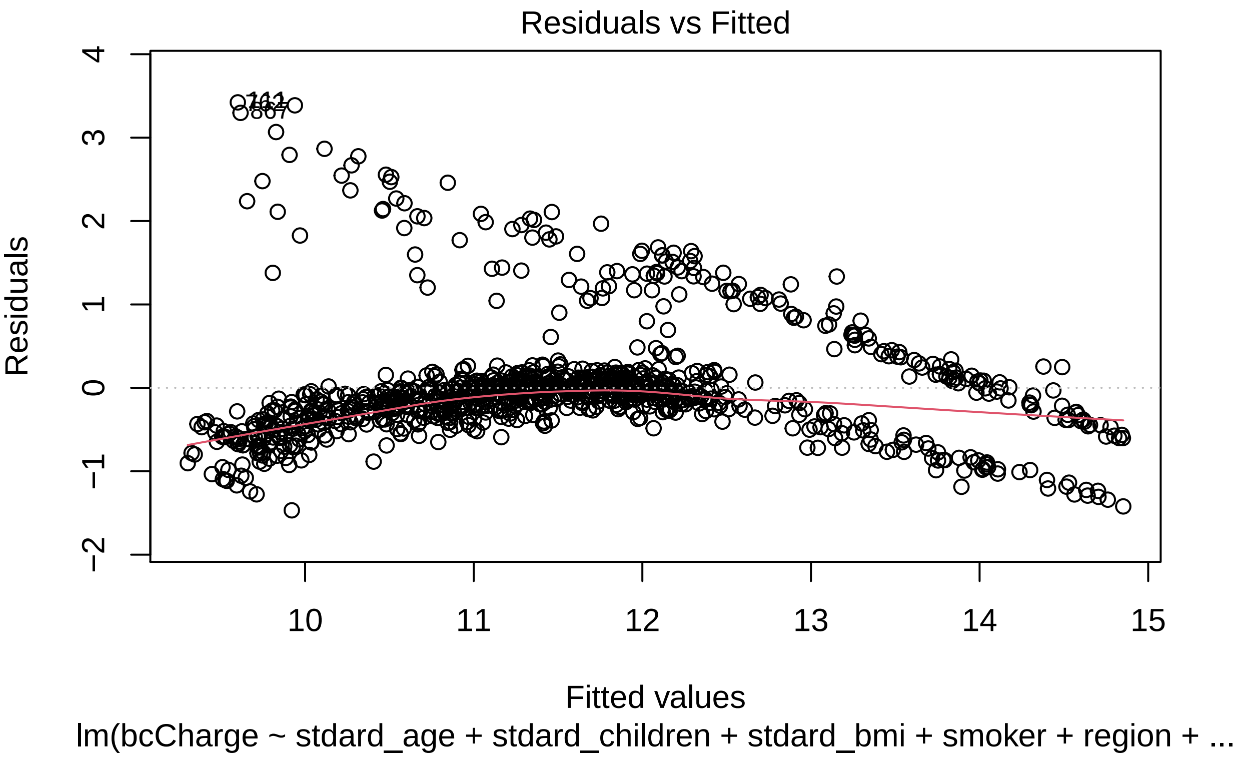
```
#select abased on best AIC
autofit <-
  lm(bcCharge~stdard_age+stdard_children+stdard_bmi+smoker+region+sex+stdard_age:stdard_children,
     data=data1)
vif(autofit)
```

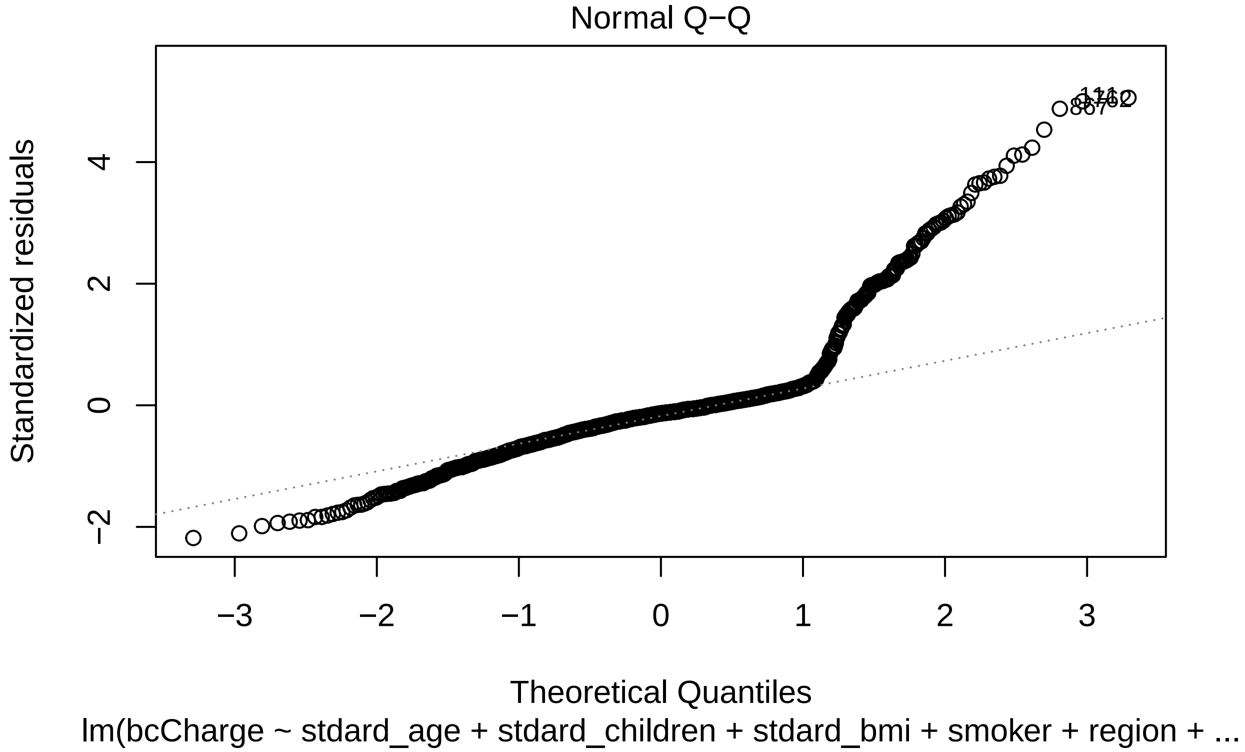
```
##                               GVIF Df GVIF^(1/(2*Df))
## stdard_age                1.107103  1        1.052190
## stdard_children           1.007251  1        1.003619
## stdard_bmi                1.111112  1        1.054093
## smoker                    1.016775  1        1.008353
## region                    1.100344  3        1.016065
## sex                       1.020070  1        1.009985
## stdard_age:stdard_children 1.089854  1        1.043961
```

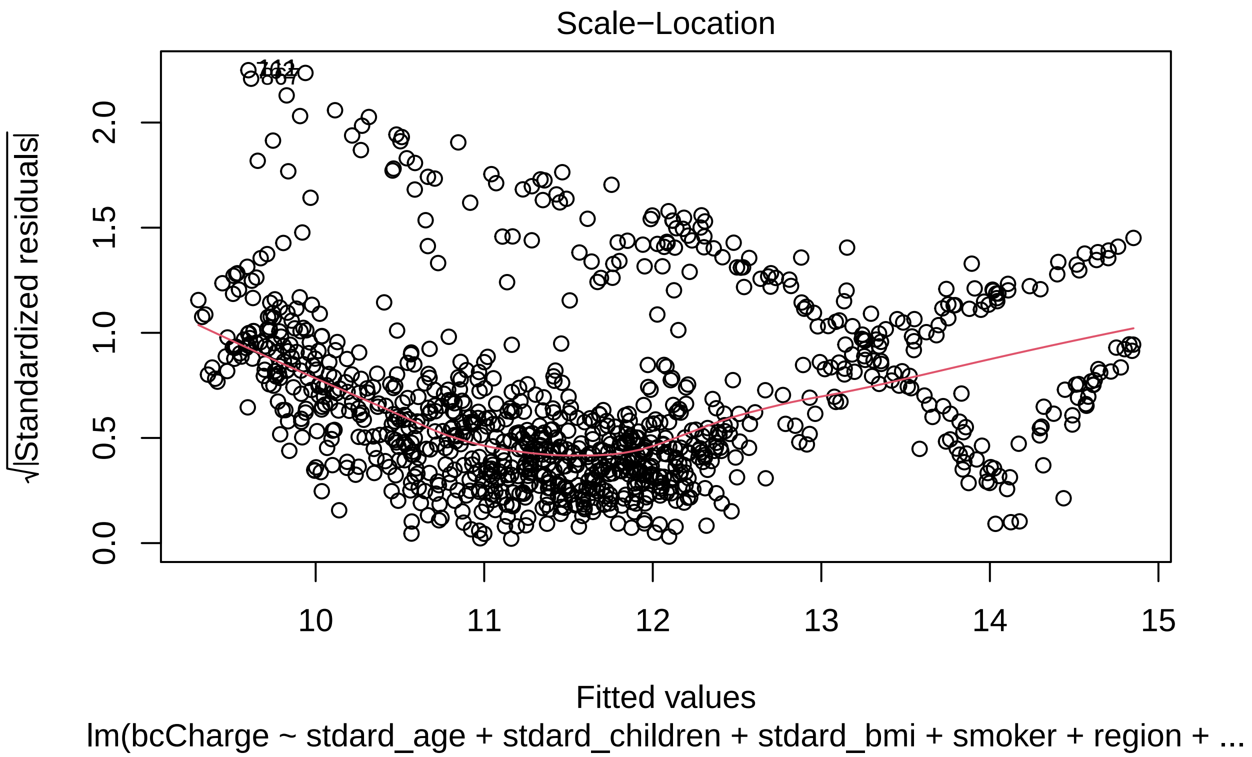
```
summary(autofit)
```

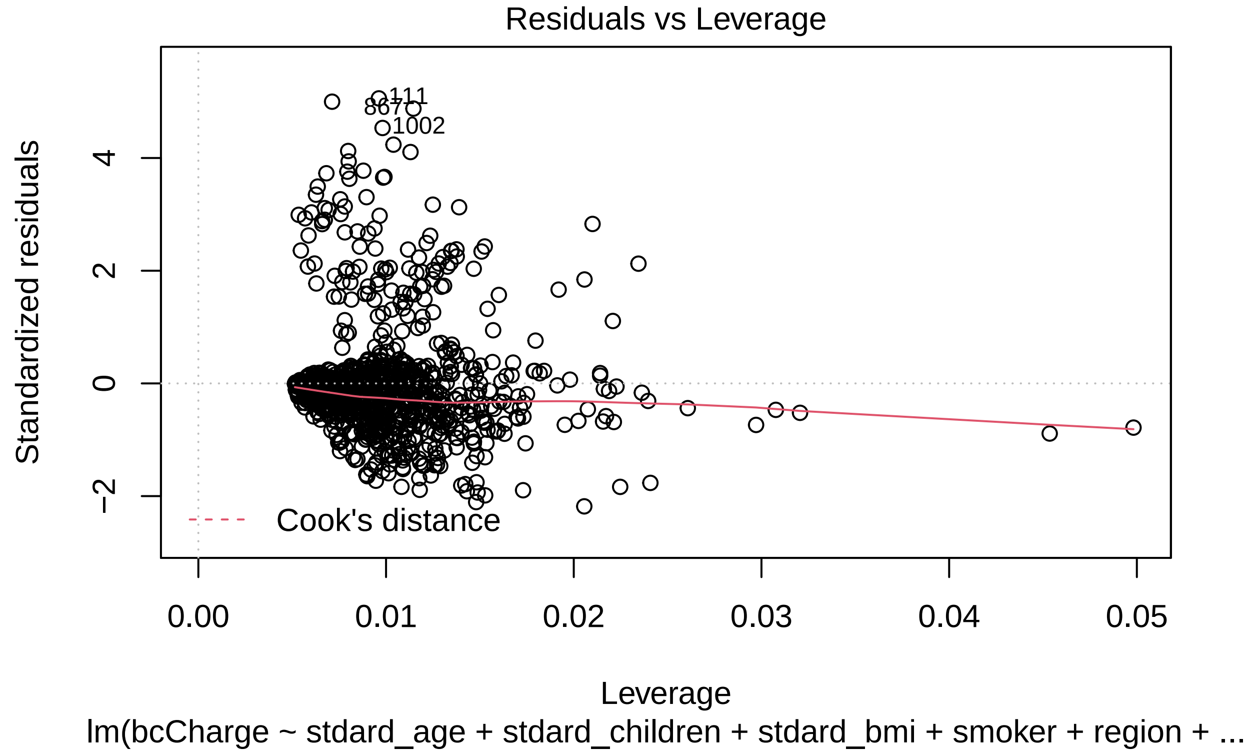
```
##
## Call:
## lm(formula = bcCharge ~ stdard_age + stdard_children +
stdard_bmi +
## smoker + region + sex + stdard_age:stdard_children, data
= data1)
##
## Residuals:
## Min 1Q Median 3Q Max
## -1.4681 -0.3272 -0.0897 0.0878 3.4225
##
## Coefficients:
## Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.25045 0.05124 219.582 < 2e-16 ***
## stdard_age 0.72013 0.02259 31.879 < 2e-16 ***
## stdard_children 0.19649 0.02155 9.119 < 2e-16 ***
## stdard_bmi 0.12767 0.02263 5.642 2.20e-08 ***
## smokeryes 2.48108 0.05292 46.879 < 2e-16 ***
## regionnorthwest -0.13160 0.06270 -2.099 0.036068 *
## regionsoutheast -0.24930 0.06293 -3.962 7.98e-05 ***
## regionsouthwest -0.21399 0.06243 -3.428 0.000634 ***
## sexmale -0.13220 0.04335 -3.050 0.002352 **
## stdard_age:stdard_children -0.09146 0.02297 -3.981
7.35e-05 ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
' ' 1
##
## Residual standard error: 0.6799 on 994 degrees of
freedom
## Multiple R-squared: 0.7797, Adjusted R-squared: 0.7777
## F-statistic: 390.8 on 9 and 994 DF, p-value: < 2.2e-16
```

```
plot(autofit)
```









```
autofit$coefficients
```

```
##             (Intercept)              stdard_age
##             11.25044941              0.72012638
##         stdard_children              stdard_bmi
##              0.19648564              0.12767156
##                  smokeryes            regionnorthwest 
##                 2.48108173                -0.13159760 
##            regionsoutheast            regionsouthwest 
##                -0.24929713                -0.21399270 
##                    sexmale stdard_age:stdard_children 
##                -0.13219642                -0.09145825 
```

We can see that the stepwise-auto-selection provide the best AIC is 2086.5292 with linear model: bcCharges = stdard_age+stdard_children+stdard_bmi+smoker+region+sex+stdard_age:stdard_children.

Based on this model we may conclude that generally male has lower medical charges than female, and smoking is positively related to medical charges. In each additional year charges will goes up.

7. We can fit a Regression Tree model in this case. First, we decide the data of characteristics (relative data, such as, income, age, smoker . . .)of patients to collect. Name the variable set we collected as predictor space. Use "patient cost" as response variable. Then select a predictor from predictor space and select a cut-point to minimize RSS = (compute mean, sum square the difference between mean in corresponding region divided by cut point). Repeat the above two steps many times, however, instead of split entire region we split on previously identified region, then we obtain our regression tree. For a better accuracy and performance, we can apply random forest method, that is, when in each split process, we choose one predictor from a random samples of m predictors from full n predictors.
After we build regression tree model, we collect and split the data into training and test sets. Train and prune the regression tree model with cross validation until we satisfy with the prediction accuracy rate. This rate is the type of probability we are looking for. We then can put test data into the tree regression model, check how good our model can predict. Evaluate the test output using using confusion matrix or AUC score. We can find the probability cost of beyond 50k by find out what proportion of data has cost more than 50k (sum all probabilities to the right of the tree-leave with approx. 50k cost). We can give this probability as the cut-off probability to the question.

package citation:

Patil, I. (2021). Visualizations with statistical details: The 'ggstatsplot' approach. Journal of Open Source Software, 6(61), 3167, doi:10.21105/joss.03167