Insurance Prediction Model

December 20, 2023

Department of Mathematics and Actuarial Science

MACT 4231- Applied Regression Methods

Dr. Noha Youssef

Name:	Email:
Sama Amr	samaamr@aucegypt.edu
Mona Ibrahim	monamahmoud@aucegypt.edu

Table of Contents

Section 1: Problem Definition

Making a decision on charging premium insurance to individuals by predicting their future medical expenses based on their existing data. Our dependent variable is the medical expenses and independent variables are age, sex, bmi, children, smoker, and region.

Section 2: Description of the data set:

Data obtained from: https://www.kaggle.com/datasets/noordeen/insurance-premium-prediction/data

The insurance.csv dataset contains 1338 observations (rows) and 7 features (columns). The dataset is observing people in regard to their medical expenses and personal conditions. The dataset contains 4 numerical quantitative variables: age in years, bmi is in kg/m^2, (number of) children, and expenses in dollars. And 3 nominal variables: sex (male or female), smoker, and region (northwest, northeast, southwest, southeast).

The purpose of this report is to look into different variables to observe their relationship, and plot a multiple linear regression based on several variables of individuals such as age, physical/family condition and location against their existing medical expense to be used for predicting future medical expenses of individuals

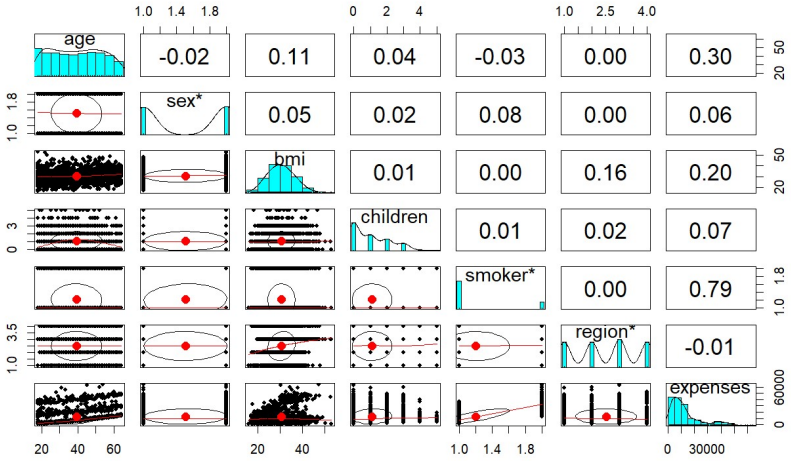
that help medical insurance to make decision on charging the premium. Therefore, the expense is used as the response variable.

We hypothesize that as the age, bmi and number of children increase, the predicted medical expenses will increase. We also hypothesize that smoking would increase medical expenses due to its negative effects on one's health. Being a female might also increase the medical expenses, especially if they have children, due to going into labor. Lastly, for the region, we believe that it will not play a big role in predicting medical expenses. Therefore, a smoking female with high age, bmi, and number of children is most likely to be charged with premium insurance.

Section 3: Data Analysis

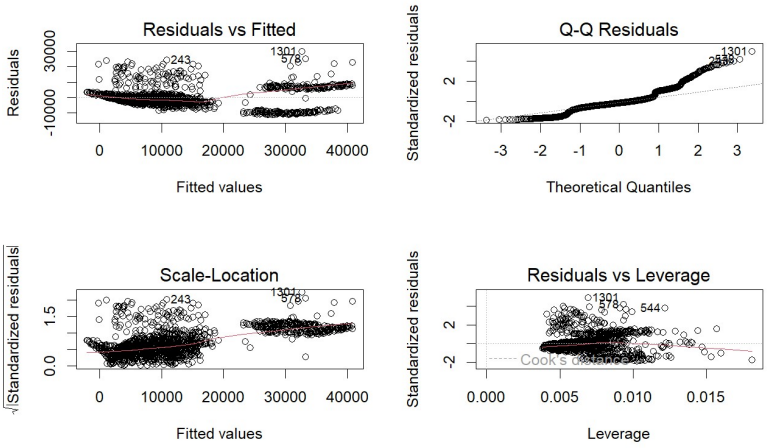
1. Observing our data

We started off our analysis by plotting a pairs panel or a plot matrix to observe the relationship and the correlation between our independent variables and our response. From the corresponding graph, we could observe that there is a linear relationship between our Y (expenses) and X's the independent variable where we could identify that expense is highly correlated with being a smoker or not. For our X's, they all happen to not be correlated, except for bmi and region; in addition to, children and region. Therefore, there is a possibility that there might be collinearity between these variables that we should keep an eye for. As for defining our linear model, we would need to use the commands "as.factor" and "as.numeric" to identify our categorical and integer variables respectively, since R doesn't recognize them automatically.



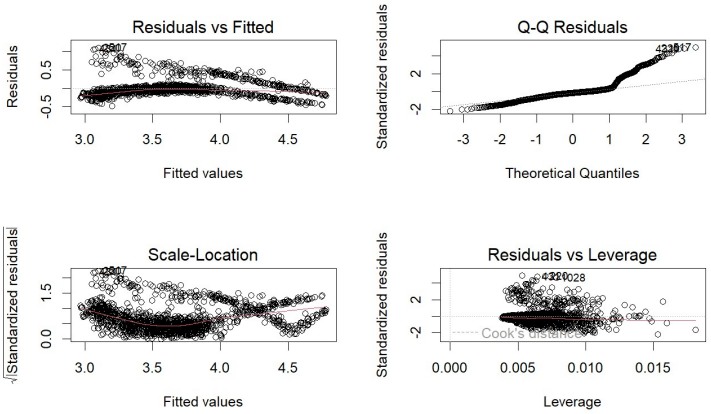
2. Running the Regression analysis

After we have identified the new variable for R to recognize them. We used the “lm()” function to run the regression analysis, and then we plotted the diagnostics plot to assess the accuracy of our model. According to the generated summary, the residual standard error was too high = 6062 which indicated the first problem of our model that we would need to solve later on. From the Residual vs fitted plot, it is clear that there is no constant variance, and the expectation could be around zero since the clusters are approximately equally distributed above and below the zero line. Moreover, the Q-Q plot has a heavy- tailed shape which indicates that the data has more outliers than expected and points do not lie on the theoretical normal line. That being said, we would need to keep an eye for influential points and outliers to remove them from our data. According to the residual vs leverage plot, the point 1301 happens to be an influential point, yet we need to make sure that it’s still one after we modify and transform our model to the better version.



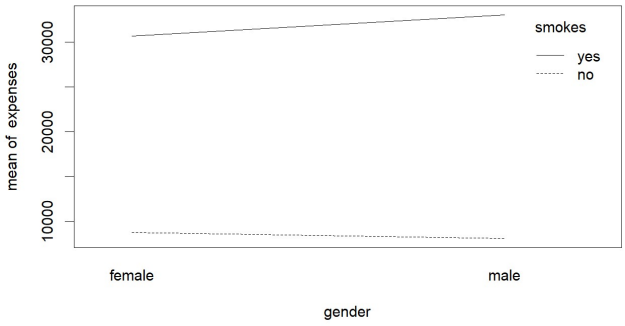
3. Transformation

Since we identified a couple of problems with our original model, we decided to apply a transformation on our response variable, and to identify such a suitable transformation, we used the “boxcox()” code/method. It turns out that our Y needs to be raised to the power of 0.1414. Therefore, we identified a new model after the transformation. The results show that our residual standard error has decreased much to be 0.2242, which is much lower than the previous value. As for the diagnostic plots, they haven’t changed much, except for the residual vs fitted plot where the clusters became more patterned which then shows that there happens to be a constant variance but we aren’t sure much of if the expectation is equal to zero. From here we decided to see if there happens to be any interaction between the variables to be able to then accurately identify the outliers and influential points.

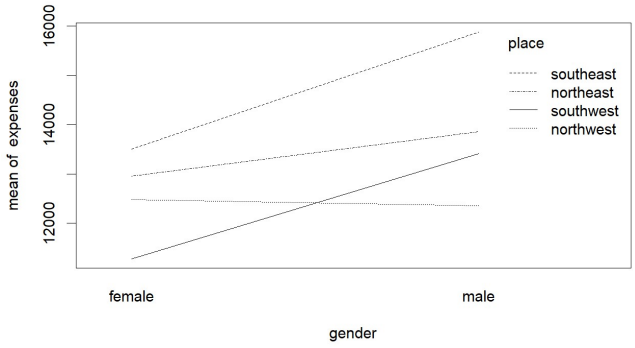


4. Interaction Terms

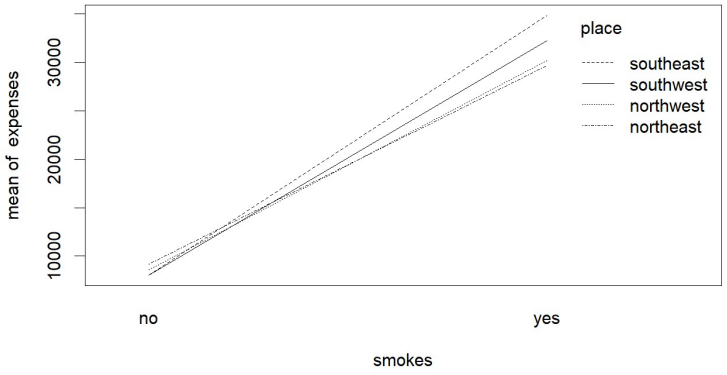
To identify which of our categorical variables have an interaction, we have plotted an interaction plot between our 3 variables and each other. We started by the gender against being a smoker and we found that they're related since we could observe 2 non-parallel lines



meaning that moving from a female to a male who smoke increases the mean expenses; while for those not smoking, the mean expenses decrease as we're moving from a female to a male. That means we would need to add (2-1)(2-1)=1 new term. In other words, it is the difference in the increase in the mean expenses is the same when moving from a female to a male in both smoking categories. Secondly, we looked at the interaction between the region and the gender, where we could observe that there is no interaction between southwest and southeast, yet northeast and northwest are somehow interacting with the gender. They show the difference in the increase in the mean expenses when being in the region of northeast and northwest with relation to being a female or a male. Therefore, we would need to add an interaction term symbolizing the interaction between gender and the region (gender * region) which would contain 3 terms since we have 4 categories in the



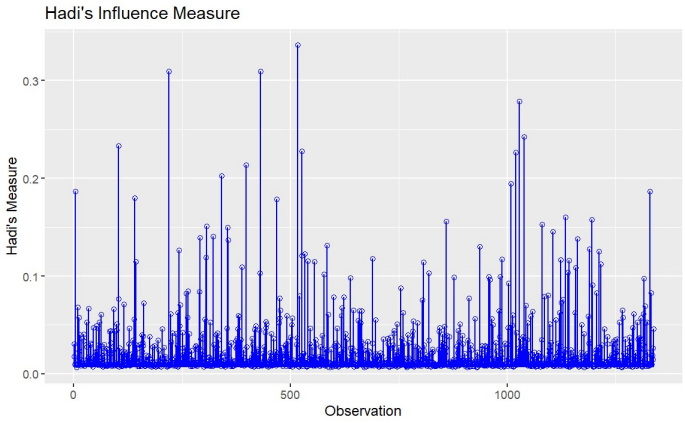
region against 2 in gender; therefore, a total of (4-1)*(2-1)= 3 new terms. As for our last 2 categorical variables smoking and region. All the lines are nearly interlinked which show the high level of interaction between all the categories; therefore, we would need to also add an interaction term for these 2 (smokes*region) which would contain 3 terms since we have 4 categories in the region against 2 in smoking; therefore, a total of (4-1)*(2-1)= 3 new terms. To sum it up, we would need to add 2 interaction terms of 6 new parameters to our model in order to account for the relations between our categorical variables. After adding our interaction terms, the residual standard error decreased a bit more to reach 0.2224, and the adjusted R^2 increased to reach 77% of the variation in y to be explained by that of x and the F-statistics value also decreased. We have also decided to run an analysis of variance test to ensure that the interaction terms have contributed to our model. From the output we can deduce that since the P-value < 0.05, we would reject the null hypothesis (H0), which means that our new model was adequate and more significant than the original model without the interaction terms.

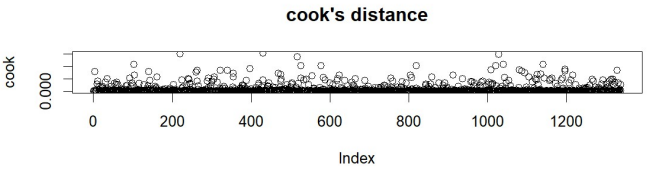


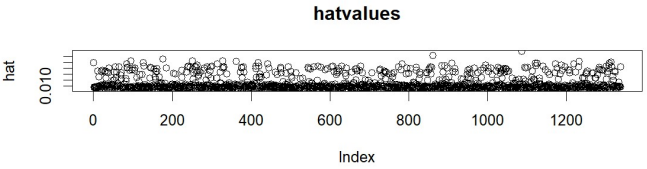

```
Analysis of Variance Table

Model 1: expenses^0.1414 ~ age + bmi + children + gender + smokes + place
Model 2: expenses^0.1414 ~ age + bmi + children + gender * smokes + gender *
    place + smokes * place
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1   1329 66.785
2   1322 65.155  7    1.6302 4.7253 2.97e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

5. Influential points and Outliers

After reaching that better model we decided to search for the influential points using hadi's measure and cook's distance. From hadi's measure, we identified that there are 4 points with high hadi's measure; thus, to identify the corresponding points we sorted the hadi values and used the "tail" command to extract the highest 4 values. As for cook's distance we used the same approach, and the extracted values turned out to be the same when using both measures; consequently, we removed these points from our data and defined a whole new model without these points and repeated the same steps again. After running the new model, the residual standard error decreased a bit to reach 0.214, as well as the adjusted R^2 increased to reach 79.7%; which in turn shows that our model is getting better.







6. Multicollinearity

For the collinearity problem we previously defined from the plot matrix, we decided to measure the variance inflation factor and condition index to identify the variables with the most linear association. The results of VIF don't show that there's multicollinearity, yet according to the condition index, there exist 2 sets of variables (14 and 15) that are suffering from multicollinearity since their kappa is greater than 10. By using the "step" command to apply the backward step function, R gave us the optimal output after dropping these 2 principal components which were identified by measuring the loss of information by the AIC. To reach in the end a residual standard error of 0.2138 with an F-statistic of 431.8 and an adjusted R^2 of 79.6%. To better verify if this model is more significant or not, we used the analysis of variance test to compare between the model we reached after dropping the influential points and adding the interaction variables and the other model we reached after dropping the principal components and measuring the AIC. The P-value identified was greater than 0.05 verifying that the model generated after dropping the PC's and (gender: place) interaction term made the model more adequate. That's because we failed to reject the null hypothesis (H0).

```
Analysis of Variance Table

Model 1: expenses^0.1414 ~ age + bmi + children + gender + place + smokes +
    place:smokes
Model 2: expenses^0.1414 ~ age + bmi + children + gender * place + smokes *
    place
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1   1322 60.716
2   1319 60.682  3  0.033859 0.2453 0.8647
```

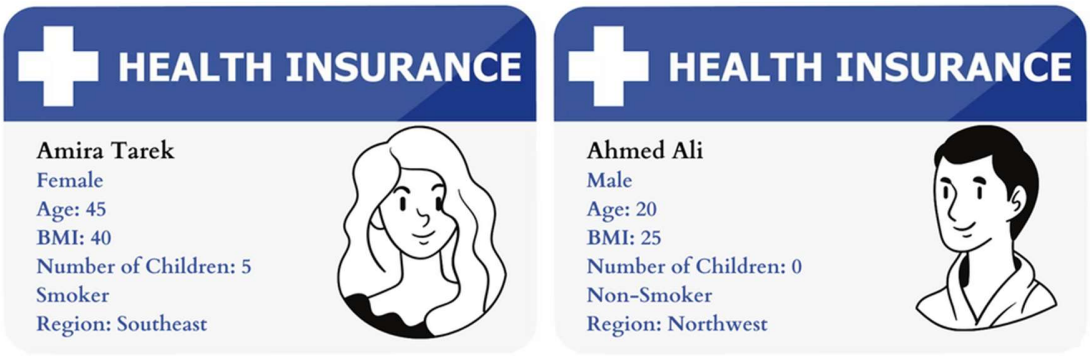
Section 4: Summary and Conclusion

Our best model:

$$Y^{0.1414} = 2.6 + 0.017\ (age) + 0.008\ (bmi) + 0.047\ (children) - 0.053\ (gendermale) + 0.7(smokesyes) - 0.039\ (placenorthwest) - 0.119\ (placesoutheast) - 0.098\ (placesouthwest) + 0.078\ (gendermale{:}\ smokesyes) + 0.022\ (smokesyes{:}\ placenorthwest) + 0.167\ (smokesyes{:}\ placesoutheast) + 0.163\ (smokesyes{:}\ placesouthwest)$$

From our model and analysis, we can deduce that the factor that highly affects the expenses is smoking; therefore, we could charge those smoking with a premium insurance plan to make sure that the company is profiting and at the same time is able to cover up the expenses of such an individual. Moreover, we can also observe that there are regional differences when it comes to expenses; thus, the company should also keep an eye for these differences and have different insurance plans for each region. The coefficients of the age, bmi and number of children show that they increase the predicted expenses while the coefficient of male shows that being a male reduces the predicted expenses. As for the interaction terms, they illustrate the difference in the increase of mean expenses as we moved from one category to another over the presence of another variable. For instance, the interaction term (smokesyes:placenorthwest) shows that the expenses increased by 0.022 when moving from being a non-smoker to a smoker when being in the northwest region. Moreover, the other 2 interaction terms (smokesyes:placesoutheast) and (smokesyes:placesouthwest) have the same interpretation, yet they are in southeast increasing by 0.167 and southwest increasing by 0.163, respectively. For the interaction term between being a male and smoking increase the expenses by 0.078. The coefficients of our interaction terms verify that being a smoker and a male in the southeast region would increase one's expenses more than in any other region.

To verify our initial hypothesis in *Section 2* we can substitute in our final equation for the model with two different individual profiles:



Amira's predicted medical expenses:

$$2.59 + 0.017(45) + 0.008(40) + 0.048(5) - 0.119(1) + 0.17(1) = 3.966$$

$$3.966^{1/0.1414} = 17046.35825\ dollars$$

Ahmed's predicted medical expenses:

$$2.59 + 0.017(20) + 0.008(25) + 0.048(0) - 0.038(1) - 0.039(1) = 3.053$$

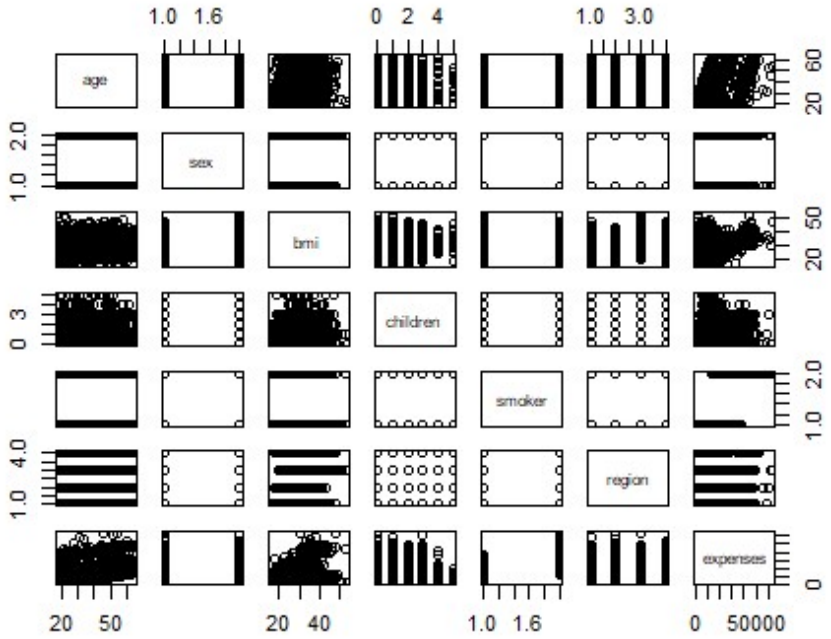
$$3.053^{1/0.1414} = 2679.5001\ dollars$$

This shows that Amira has a higher predicted medical expense; therefore, it would be recommended that she is charged a premium plan package. However, since Ahmed has a lower predicted medical expense, he should be charged a cheaper package. Therefore, it shows that our initial hypothesis was correct as it seems

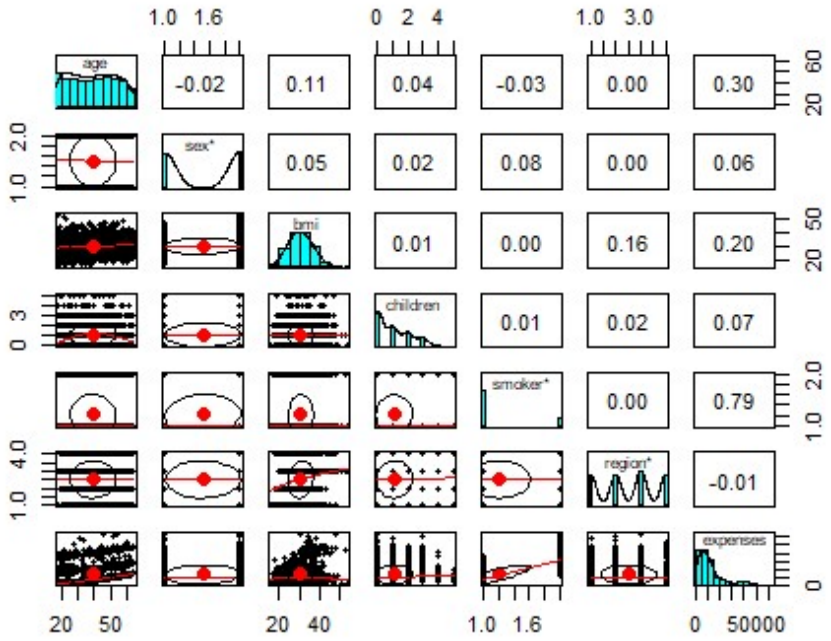
that a smoking female with high age, bmi, and number of children would have a higher predicted medical expense. That being said, health insurance providers might consider offering plans that provide incentives for adopting and maintaining healthy behaviors such as that of Ahmed. This could involve discounts for individuals who quit smoking or maintain a healthy BMI. It is also important to note that the region plays a key role in predicting the medical expenses and should be considered when recommending premium plans.

Appendix

```
my_data<-read.csv("insurance.csv",header=TRUE)
attach(my_data)
head(my_data,2)

##   age    sex  bmi children smoker    region expenses
## 1  19 female 27.9        0    yes southwest 16884.92
## 2  18   male 33.8        1     no southeast  1725.55

plot(my_data)
```




```
pairs.panels(my_data)
```




```
names(my_data)

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

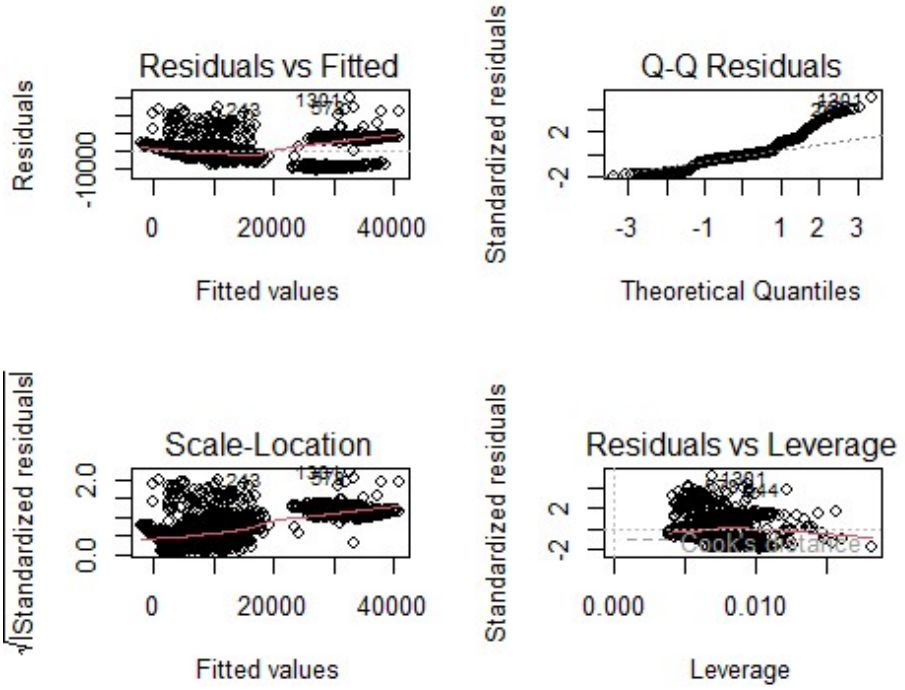
- Running the regression Analysis
- Modifying the model since R doesn't recognize categorical and integer variables

```
gender <- as.factor(sex)
smokes <- as.factor(smoker)
place <- as.factor(region)
my_data<-cbind(my_data,gender,smokes,place)
children <- as.numeric(children)

m1<-lm(expenses~age+bmi+children+sex+smoker+region)
summary(m1)

##
## Call:
## lm(formula = expenses ~ age + bmi + children + sex + smoker +
##     region)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11302.7  -2850.9   -979.6   1383.9  29981.7
##
```

17

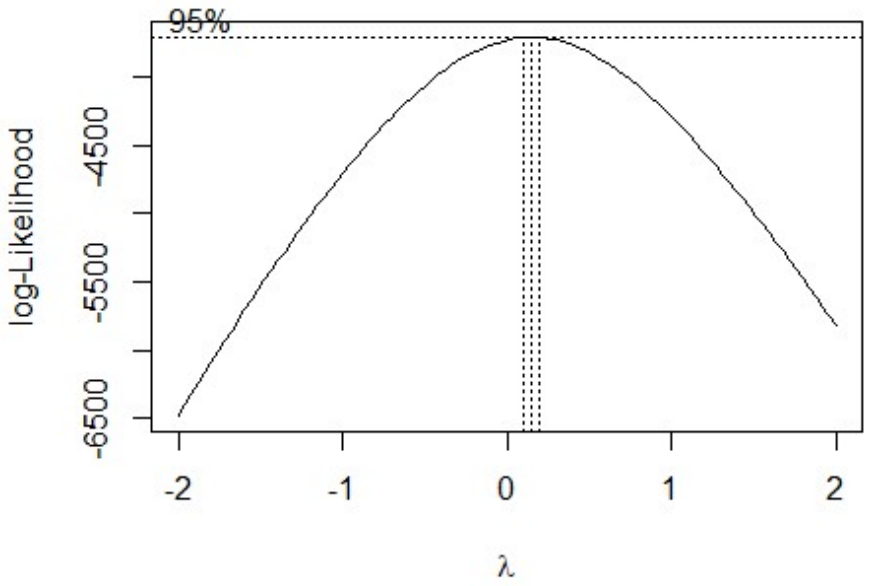
```
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -11941.6      987.8 -12.089  < 2e-16 ***
## age                256.8       11.9  21.586  < 2e-16 ***
## bmi                339.3       28.6  11.864  < 2e-16 ***
## children           475.7      137.8   3.452 0.000574 ***
## sexmale           -131.3      332.9  -0.395 0.693255
## smokeryes        23847.5      413.1  57.723  < 2e-16 ***
## regionnorthwest   -352.8      476.3  -0.741 0.458976
## regionsoutheast  -1035.6      478.7  -2.163 0.030685 *
## regionsouthwest   -959.3      477.9  -2.007 0.044921 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6062 on 1329 degrees of freedom
## Multiple R-squared:  0.7509, Adjusted R-squared:  0.7494
## F-statistic: 500.9 on 8 and 1329 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(m1)
```



Residual standard error is very high –> modify the model

```
xx<-boxcox(m1)
```

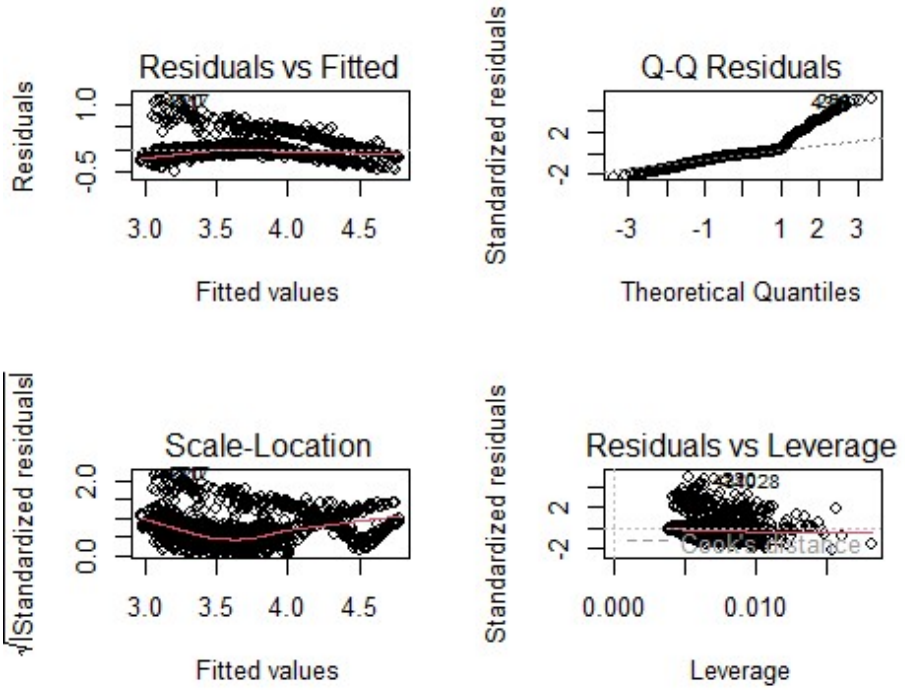



```
xx$x[which.max(xx$y)]

## [1] 0.1414141
```

TRANSFORMATION (POWER 0.1414)

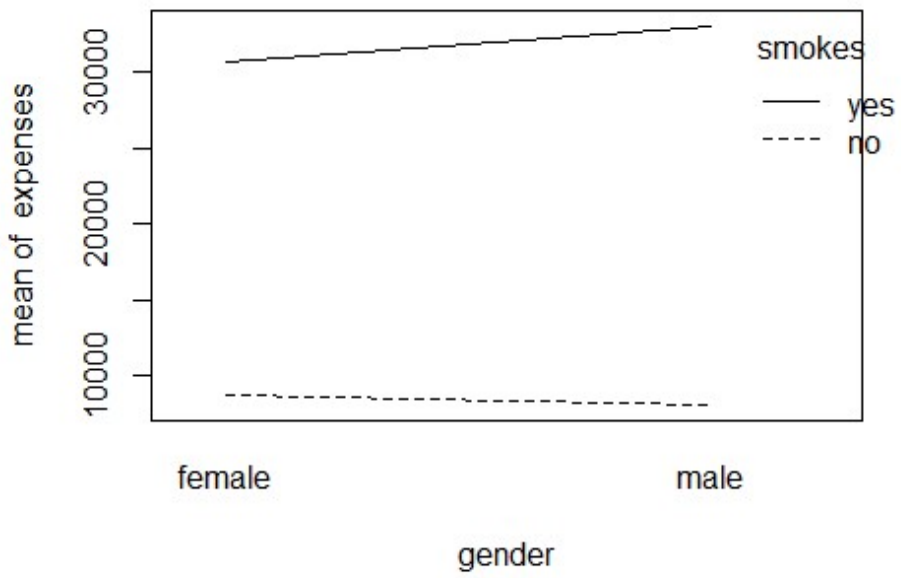
```
m2<-lm(expenses^0.1414~age+bmi+children+gender+smokes+place)
summary(m2)

##
## Call:
## lm(formula = expenses^0.1414 ~ age + bmi + children + gender +
##     smokes + place)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.48822 -0.10777 -0.03368  0.02447  1.10570
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     2.593483   0.036529  70.997  < 2e-16 ***
## age             0.016738   0.000440  38.039  < 2e-16 ***
## bmi             0.007789   0.001058   7.365 3.11e-13 ***
## children        0.046299   0.005096   9.086  < 2e-16 ***
## gendermale     -0.032644   0.012312  -2.651 0.008110 **
## smokesyes       0.833747   0.015278  54.572  < 2e-16 ***
```

```
## placenorthwest -0.030414   0.017612  -1.727 0.084423 .
## placesoutheast -0.074172   0.017702  -4.190 2.97e-05 ***
## placesouthwest -0.062143   0.017673  -3.516 0.000453 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2242 on 1329 degrees of freedom
## Multiple R-squared:  0.7762, Adjusted R-squared:  0.7748
## F-statistic: 576.1 on 8 and 1329 DF,  p-value: < 2.2e-16

par(mfrow=c(2,2))
plot(m2)
```



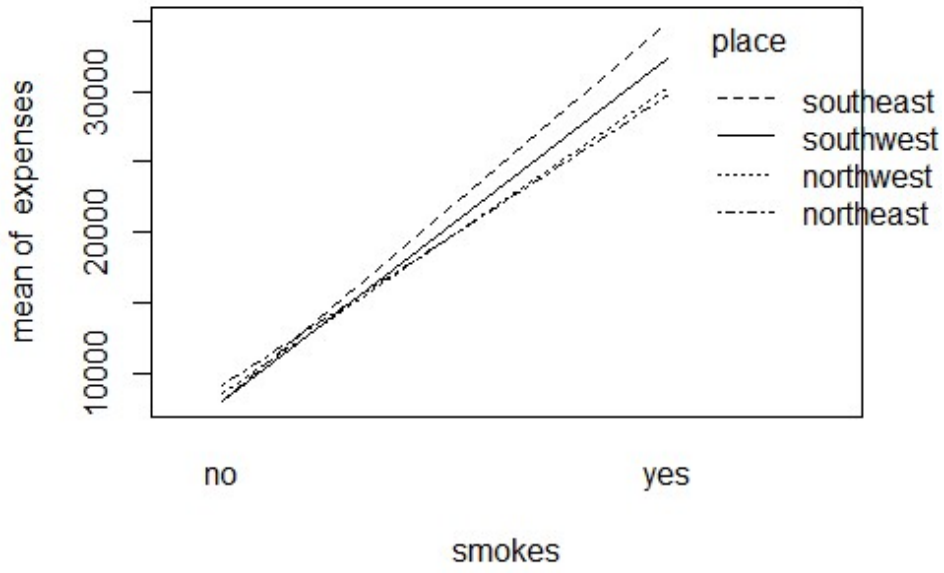
Plotting the Interaction plots to identify the interaction variables

```
interaction.plot(gender,smokes,expenses)
```




```
interaction.plot(gender,place,expenses)
```

```
interaction.plot(smokes,place,expenses)
```

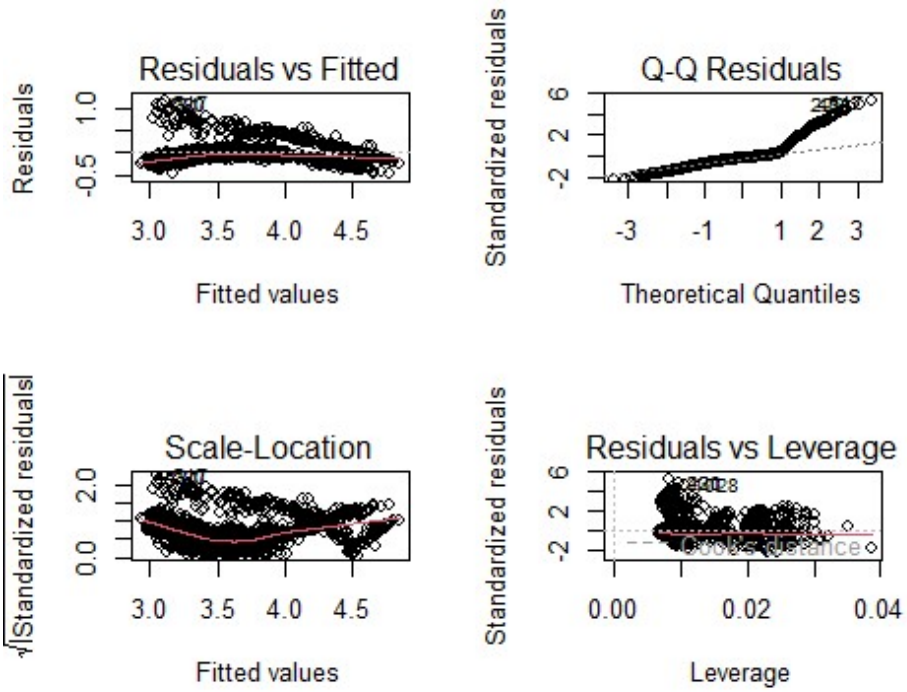


Adding the interaction terms

```
m3<-lm(expenses^0.1414~age+bmi+children+ gender*smokes+gender*place+ smokes*p
lace)
summary(m3)

##
## Call:
## lm(formula = expenses^0.1414 ~ age + bmi + children + gender *
##     smokes + gender * place + smokes * place)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.45756 -0.10707 -0.03410  0.02533  1.13946
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 2.6188603  0.0388020  67.493  < 2e-16 ***
## age                         0.0167496  0.0004365  38.373  < 2e-16 ***
## bmi                         0.0076262  0.0010511   7.256 6.79e-13 ***
## children                    0.0461921  0.0050516   9.144  < 2e-16 ***
## gendermale                 -0.0373750  0.0255048  -1.465 0.143046
## smokesyes                   0.7004938  0.0349032  20.070  < 2e-16 ***
```

22

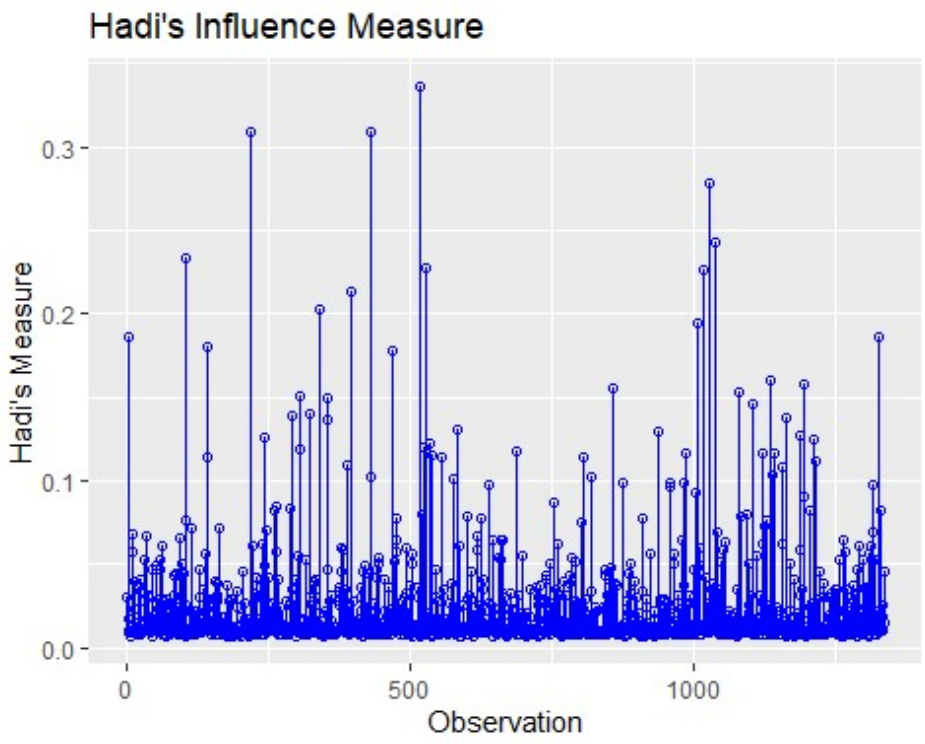
```
## placenorthwest             -0.0319561  0.0258752  -1.235 0.217047
## placesoutheast             -0.0977375  0.0257313  -3.798 0.000152 ***
## placesouthwest             -0.0861256  0.0256744  -3.355 0.000818 ***
## gendermale:smokesyes        0.0762229  0.0306437   2.487 0.012991 *
## gendermale:placenorthwest  -0.0077241  0.0349054  -0.221 0.824903
## gendermale:placesoutheast  -0.0252424  0.0341039  -0.740 0.459333
## gendermale:placesouthwest  -0.0157476  0.0350852  -0.449 0.653621
## smokesyes:placenorthwest    0.0207961  0.0443826   0.469 0.639458
## smokesyes:placesoutheast    0.1624497  0.0407720   3.984 7.14e-05 ***
## smokesyes:placesouthwest    0.1602559  0.0445990   3.593 0.000339 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.222 on 1322 degrees of freedom
## Multiple R-squared:  0.7816, Adjusted R-squared:  0.7792
## F-statistic: 315.5 on 15 and 1322 DF,  p-value: < 2.2e-16

par(mfrow=c(2,2))
plot(m3)
```



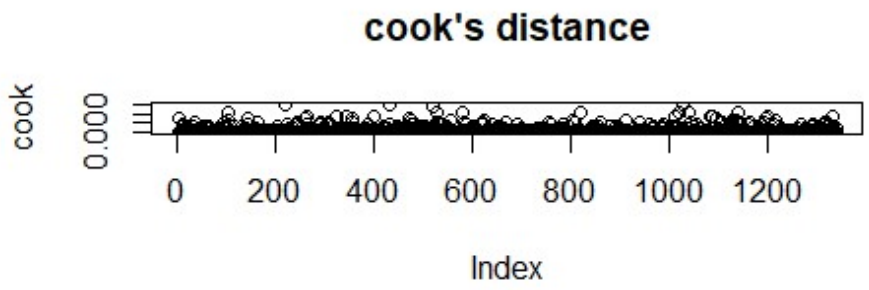
Calculating hadi's measure and cooks distance to identify the influential points

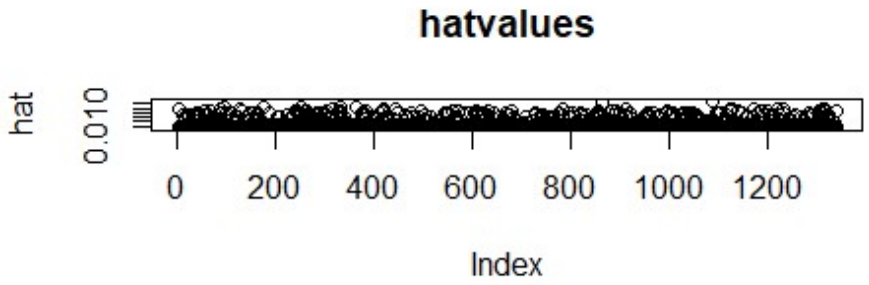
AFTER THE TRANSFORMATION AND INTERACTION TERMS

```
row.names(my_data)=1:nrow(my_data)

hadi<-ols_hadi(m3)
cook<-cooks.distance(m3)
hat<-hatvalues(m3)

ols_plot_hadi(m3)
```




```
par(mfrow=c(2,1))

plot(cook,main="cook's distance")

plot(hat,main="hatvalues")
```






```
#identify(row.names(my_data),hadi$hadi,my_data,labels = row.names(my_data)) # this command to be implemented in the console
```

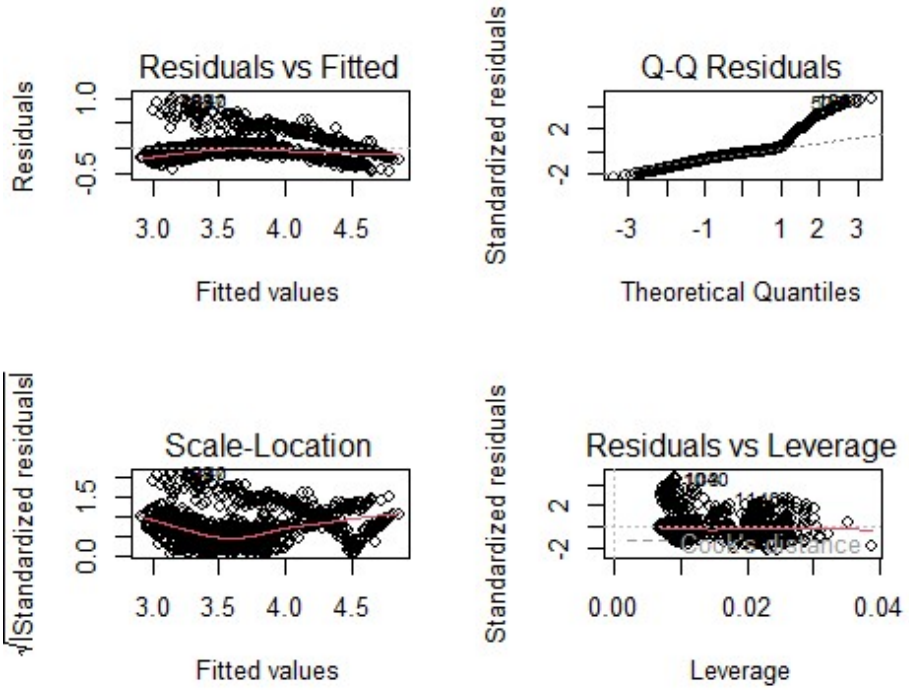
We Identified that the there are 4 influential points from hadi and so we need to identify the corresponding observation numbers

```
print("influentials from Hadi's measure")

## [1] "influentials from Hadi's measure"

tail(sort(hadi$hadi),4)

##      1028       220       431       517
## 0.2790428 0.3092192 0.3097992 0.3363182

print("influentials from cook's distance")

## [1] "influentials from cook's distance"

tail(sort(cook),4)

##        517       1028        220        431
## 0.01377723 0.01460463 0.01502904 0.01516790
```

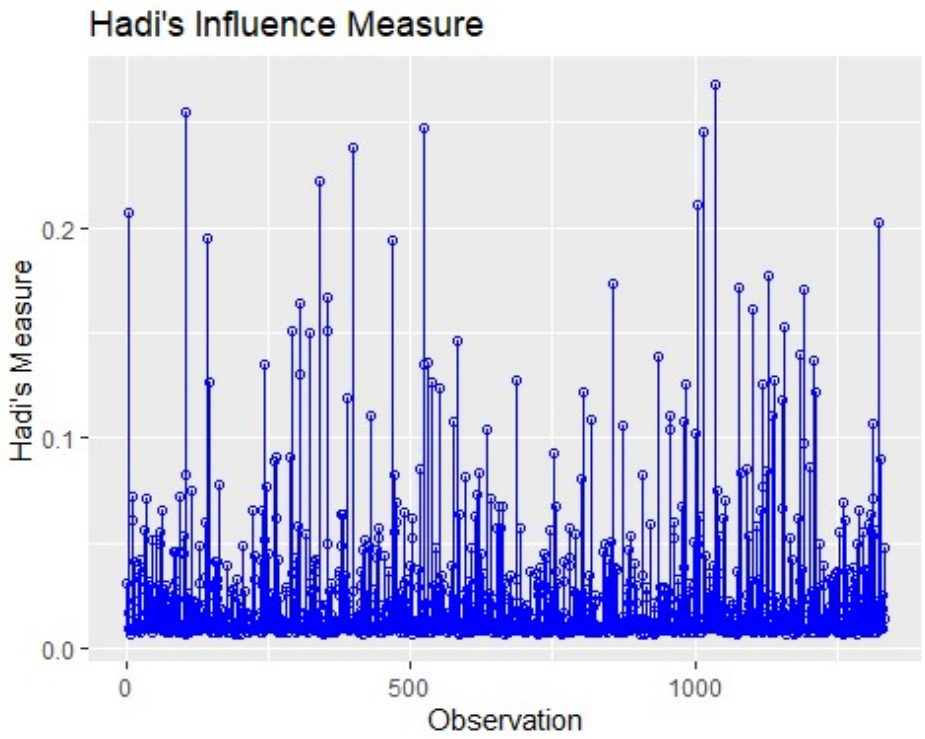
Removing the influential points identified from hadi's measure

25

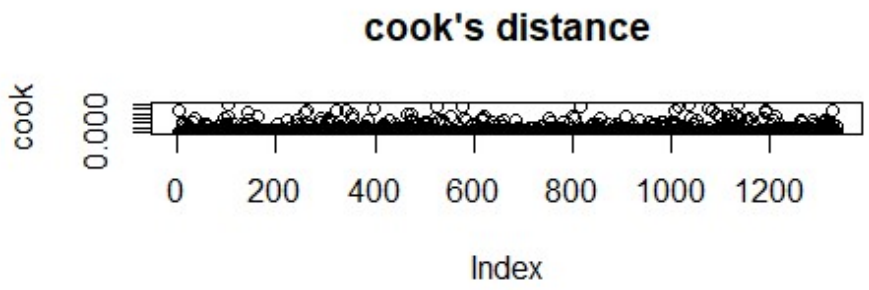
```
my_data1<-my_data[-c(1028,431,220,517),]
m4<-lm(expenses^0.1414~age+bmi+children+gender*smokes+gender*place+ smokes
*place,data=my_data1)
summary(m4)

##
## Call:
## lm(formula = expenses^0.1414 ~ age + bmi + children + gender *
##     smokes + gender * place + smokes * place, data = my_data1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.46502 -0.10540 -0.03159  0.02675  0.97595
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  2.5974804  0.0374942  69.277  < 2e-16 ***
## age                          0.0170230  0.0004216  40.374  < 2e-16 ***
## bmi                          0.0079322  0.0010153   7.813 1.13e-14 ***
## children                     0.0479342  0.0048734   9.836  < 2e-16 ***
## gendermale                  -0.0378477  0.0245849  -1.539 0.123930
## smokesyes                    0.6994577  0.0336471  20.788  < 2e-16 ***
## placenorthwest              -0.0326935  0.0249417  -1.311 0.190155
## placesoutheast              -0.1065527  0.0248497  -4.288 1.94e-05 ***
## placesouthwest              -0.0873327  0.0247482  -3.529 0.000432 ***
## gendermale:smokesyes         0.0791599  0.0295510   2.679 0.007482 **
## gendermale:placenorthwest   -0.0141461  0.0336720  -0.420 0.674471
## gendermale:placesoutheast   -0.0262825  0.0329151  -0.798 0.424729
## gendermale:placesouthwest   -0.0234065  0.0338476  -0.692 0.489357
## smokesyes:placenorthwest     0.0240344  0.0427876   0.562 0.574406
## smokesyes:placesoutheast     0.1700834  0.0393156   4.326 1.63e-05 ***
## smokesyes:placesouthwest     0.1657308  0.0430028   3.854 0.000122 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.214 on 1318 degrees of freedom
## Multiple R-squared:  0.797,  Adjusted R-squared:  0.7947
## F-statistic: 344.9 on 15 and 1318 DF,  p-value: < 2.2e-16

par(mfrow=c(2,2))
plot(m4)
```

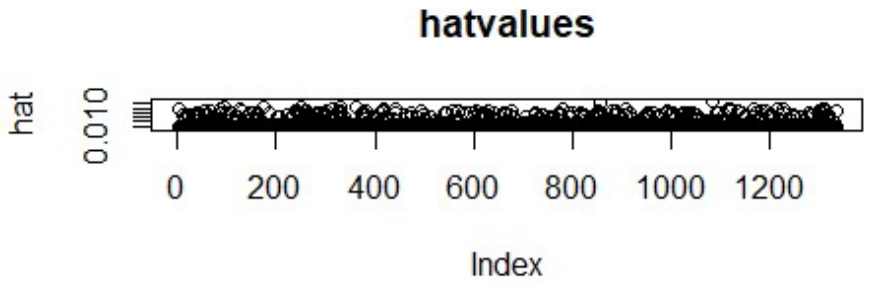



```
row.names(my_data)=1:nrow(my_data)

hadi<-ols_hadi(m4)
cook<-cooks.distance(m4)
hat<-hatvalues(m4)

ols_plot_hadi(m4)
```




```
par(mfrow=c(2,1))

plot(cook,main="cook's distance")

plot(hat,main="hatvalues")
```





Comparing the 2 models

```
anova(m2,m3)
```

```
## Analysis of Variance Table
##
## Model 1: expenses^0.1414 ~ age + bmi + children + gender + smokes + place
## Model 2: expenses^0.1414 ~ age + bmi + children + gender * smokes + gender
*
##     place + smokes * place
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1   1329 66.785
## 2   1322 65.155  7    1.6302 4.7253 2.97e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

P-value<0.05 Reject H0 and accept H1 = Model 2 is better

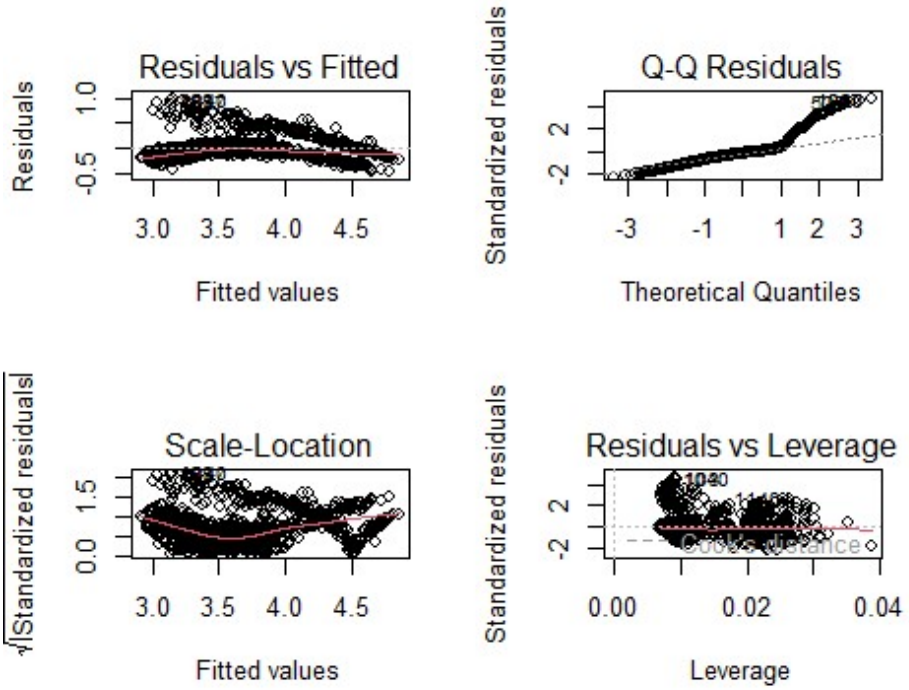
```
ols_coll_diag(m4)
```

```
## Tolerance and Variance Inflation Factor
## ---------------------------------------
##                    Variables Tolerance      VIF
## 1                        age 0.9807087 1.019671
## 2                        bmi 0.8975397 1.114157
## 3                   children 0.9942012 1.005833
```

```
## 4                  gendermale 0.2271897 4.401608
## 5                   smokesyes 0.1857769 5.382801
## 6              placenorthwest 0.3000703 3.332552
## 7              placesoutheast 0.2811421 3.556920
## 8              placesouthwest 0.3047806 3.281049
## 9        gendermale:smokesyes 0.3744234 2.670773
## 10 gendermale:placenorthwest 0.2868243 3.486455
## 11 gendermale:placesoutheast 0.2617031 3.821125
## 12 gendermale:placesouthwest 0.2808300 3.560873
## 13  smokesyes:placenorthwest 0.4508462 2.218051
## 14  smokesyes:placesoutheast 0.3493825 2.862193
## 15  smokesyes:placesouthwest 0.4463440 2.240424
##
##
## Eigenvalue and Condition Index
## ------------------------------
##    Eigenvalue Condition Index    intercept          age          bmi
## 1  6.13677545        1.000000 5.567414e-04 2.325992e-03 7.707044e-04
## 2  2.10682291        1.706696 6.367261e-05 2.992193e-04 3.431872e-05
## 3  2.00671654        1.748747 1.587137e-05 8.197854e-05 1.125982e-05
## 4  1.69577202        1.902332 4.556082e-04 2.311512e-03 7.243709e-04
## 5  0.85330100        2.681754 8.338014e-04 5.128788e-03 1.095332e-03
## 6  0.70315526        2.954231 1.460510e-04 6.372603e-04 1.438218e-04
## 7  0.65230223        3.067225 9.649404e-07 1.258025e-04 8.319596e-06
## 8  0.49073312        3.536286 9.290872e-04 4.948193e-03 1.559330e-03
## 9  0.36425123        4.104588 3.383754e-03 2.778189e-02 4.242333e-03
## 10 0.28815904        4.614812 3.377435e-05 2.181158e-04 7.218528e-05
## 11 0.27824970        4.696267 6.167755e-05 3.487489e-04 2.840950e-04
## 12 0.19554630        5.602028 3.523396e-04 5.846120e-03 2.093105e-05
## 13 0.09094789        8.214361 4.242879e-03 2.835841e-01 4.817358e-03
## 14 0.07796952        8.871717 2.181204e-02 5.714772e-01 5.057938e-02
## 15 0.04371001       11.848944 1.072804e-02 5.367598e-02 1.982512e-01
## 16 0.01558777       19.841664 9.563837e-01 4.120909e-02 7.373850e-01
##        children   gendermale    smokesyes placenorthwest placesoutheast
## 1  6.956458e-03 1.971379e-03 0.0016118873   0.0013493795   0.0016818816
## 2  2.168037e-03 6.779757e-05 0.0019023233   0.0090138642   0.0184284957
## 3  3.222894e-04 3.213800e-06 0.0003021362   0.0228132711   0.0001513376
## 4  7.600928e-03 4.234489e-04 0.0217360388   0.0006928121   0.0071736176
## 5  5.421800e-02 3.443491e-02 0.0059495871   0.0002063638   0.0011617194
## 6  1.477918e-02 5.796916e-03 0.0008588237   0.0104038056   0.0001585794
## 7  1.033138e-03 1.104469e-04 0.0005264058   0.0006970682   0.0131139921
## 8  7.545531e-01 5.140421e-03 0.0003775880   0.0183100407   0.0170608568
## 9  1.503901e-01 8.800271e-03 0.0011866255   0.0194493470   0.0470107701
## 10 3.590902e-05 1.069252e-03 0.0001697503   0.1077072390   0.0084114181
## 11 8.203300e-04 2.979164e-03 0.0024629235   0.2184040082   0.1658464306
## 12 6.357636e-05 1.062596e-01 0.0747227877   0.0774104356   0.1807672628
## 13 5.330645e-06 2.801961e-03 0.6333959581   0.0397024483   0.0292991953
## 14 2.895788e-03 1.421169e-01 0.1895009608   0.0060696909   0.0012659717
```

```
## 15 1.741686e-04 5.821730e-01 0.0308216903   0.3778844042   0.4791019790
## 16 3.983633e-03 1.058514e-01 0.0344745134   0.0898858217   0.0293664922
##    placesouthwest gendermale:smokesyes gendermale:placenorthwest
## 1     0.001455961         0.0029677823              0.0011165527
## 2     0.009130966         0.0044135112              0.0083285066
## 3     0.021453666         0.0013703030              0.0220021669
## 4     0.001048601         0.0422584172              0.0012889583
## 5     0.002578172         0.0166991929              0.0619164297
## 6     0.005509081         0.0001915749              0.0136683883
## 7     0.011077779         0.0001424805              0.0039617824
## 8     0.012731176         0.0524116562              0.0006641031
## 9     0.012432803         0.1733992765              0.0365579804
## 10    0.306530352         0.0003439504              0.1355337364
## 11    0.039392732         0.0025773920              0.1234583941
## 12    0.060421101         0.5882135666              0.0037626191
## 13    0.040003786         0.0786753265              0.0040498867
## 14    0.006940062         0.0099687621              0.1303697267
## 15    0.393916043         0.0118467838              0.4008348425
## 16    0.075377718         0.0145200238              0.0524859263
##    gendermale:placesoutheast gendermale:placesouthwest smokesyes:placenort
hwest
## 1               1.468929e-03              1.228016e-03              0.0010
74558
## 2               2.033290e-02              8.170509e-03              0.0030
26651
## 3               7.962352e-05              2.326866e-02              0.0172
97211
## 4               4.386165e-03              8.126356e-05              0.0415
72853
## 5               2.297467e-02              1.717976e-02              0.0530
30241
## 6               1.650621e-02              4.201425e-02              0.2491
98657
## 7               4.292787e-02              2.006536e-02              0.0276
76394
## 8               2.263620e-05              2.220709e-04              0.0062
64319
## 9               1.610841e-02              4.861322e-02              0.0524
24675
## 10              2.266467e-02              2.303674e-01              0.0112
50689
## 11              2.216475e-01              1.043240e-02              0.0109
06697
## 12              5.363992e-03              1.088757e-02              0.0131
23006
## 13              9.787000e-03              5.591532e-03              0.3705
10799
## 14              1.506782e-01              1.334891e-01              0.1145
```

```
29928
## 15              3.961000e-01              3.811046e-01              0.0
16253722
## 16              6.895120e-02              6.728439e-02              0.0118
59600
##    smokesyes:placesoutheast smokesyes:placesouthwest
## 1              1.490689e-03             1.233610e-03
## 2              3.154763e-02             3.329946e-03
## 3              3.519978e-05             3.142848e-02
## 4              5.180981e-03             2.389395e-02
## 5              3.563245e-04             1.692212e-03
## 6              1.417978e-02             1.415008e-01
## 7              2.035442e-01             1.845518e-01
## 8              7.593247e-03             2.304696e-02
## 9              8.647721e-02             7.421264e-02
## 10             4.577858e-04             4.536400e-03
## 11             1.404031e-02             4.022275e-05
## 12             4.338494e-02             3.450409e-02
## 13             4.157804e-01             3.685579e-01
## 14             1.477104e-01             8.987941e-02
## 15             1.824597e-02             8.225578e-03
## 16             9.974808e-03             9.366031e-03

ols_step_backward_p(m4) #identifying model 4

##
##
##                               Elimination Summary
## ------------------------------------------------------------------------
---
##         Variable                       Adj.
## Step      Removed       R-Square    R-Square     C(p)        AIC        RM
SE
## ------------------------------------------------------------------------
---
##    1    gender:place      0.7969       0.795    2.7563    -315.1957    0.2
138
## ------------------------------------------------------------------------
---

par(mfrow=c(2,2))
plot(m4)
```




```
m5 <- step(m4,direction="backward",test="F") #identifying model 5

## Start:  AIC=-4097.69
## expenses^0.1414 ~ age + bmi + children + gender * smokes + gender *
##     place + smokes * place
##
##                 Df Sum of Sq     RSS     AIC   F value    Pr(>F)
## - gender:place   3     0.035  60.388 -4102.9    0.2521  0.859882
## <none>                       60.353 -4097.7
## - gender:smokes  1     0.329  60.682 -4092.4    7.1757  0.007482 **
## - smokes:place   3     1.327  61.680 -4074.7    9.6583 2.625e-06 ***
## - bmi            1     2.795  63.149 -4039.3   61.0432 1.134e-14 ***
## - children       1     4.430  64.784 -4005.2   96.7454 < 2.2e-16 ***
## - age            1    74.643 134.997 -3025.8 1630.0630 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=-4102.92
## expenses^0.1414 ~ age + bmi + children + gender + smokes + place +
##     gender:smokes + smokes:place
##
##                 Df Sum of Sq     RSS     AIC   F value    Pr(>F)
## <none>                       60.388 -4102.9
## - gender:smokes  1     0.328  60.716 -4097.7    7.1711  0.007501 **
## - smokes:place   3     1.303  61.691 -4080.5    9.4987 3.292e-06 ***
```

```
## - bmi            1     2.772  63.160 -4045.1   60.6285 1.384e-14 ***
## - children       1     4.442  64.830 -4010.2   97.1628 < 2.2e-16 ***
## - age            1    74.643 135.031 -3031.4 1632.8327 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(m5)
```

```
##
## Call:
## lm(formula = expenses^0.1414 ~ age + bmi + children + gender +
##     smokes + place + gender:smokes + smokes:place, data = my_data1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.46262 -0.10678 -0.03238  0.02653  0.97680
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                2.6066480  0.0357331  72.948  < 2e-16 ***
## age                        0.0170226  0.0004213  40.408  < 2e-16 ***
## bmi                        0.0078857  0.0010128   7.786 1.38e-14 ***
## children                   0.0479905  0.0048686   9.857  < 2e-16 ***
## gendermale                -0.0539766  0.0131466  -4.106 4.28e-05 ***
## smokesyes                  0.7008958  0.0335639  20.882  < 2e-16 ***
## placenorthwest            -0.0395700  0.0187058  -2.115  0.03458 *
## placesoutheast            -0.1191896  0.0190934  -6.242 5.79e-10 ***
## placesouthwest            -0.0985477  0.0187440  -5.258 1.70e-07 ***
## gendermale:smokesyes       0.0788583  0.0294480   2.678  0.00750 **
## smokesyes:placenorthwest   0.0227484  0.0427174   0.533  0.59445
## smokesyes:placesoutheast   0.1676609  0.0391569   4.282 1.99e-05 ***
## smokesyes:placesouthwest   0.1632800  0.0427428   3.820  0.00014 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2138 on 1321 degrees of freedom
## Multiple R-squared:  0.7969, Adjusted R-squared:  0.795
## F-statistic: 431.8 on 12 and 1321 DF,  p-value: < 2.2e-16
```

VIF < 10 –> no collinearity Condition Index for 2 sets of variables (14 & 15) was > 10 Step function removed the sets of variables with condition index > 10 (from the interaction term)

```
anova(m5,m4)
```

```
## Analysis of Variance Table
##
## Model 1: expenses^0.1414 ~ age + bmi + children + gender + smokes + place
+
```

34

```
##     gender:smokes + smokes:place
## Model 2: expenses^0.1414 ~ age + bmi + children + gender * smokes + gen
der *
##     place + smokes * place
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1   1321 60.388
## 2   1318 60.353  3   0.03463 0.2521 0.8599
```

P-value > 0.05 Fail to Reject H0; Model 1 is significant; after removing (gender:place) interaction term the model is more adequate = worth losing the degrees of freedom