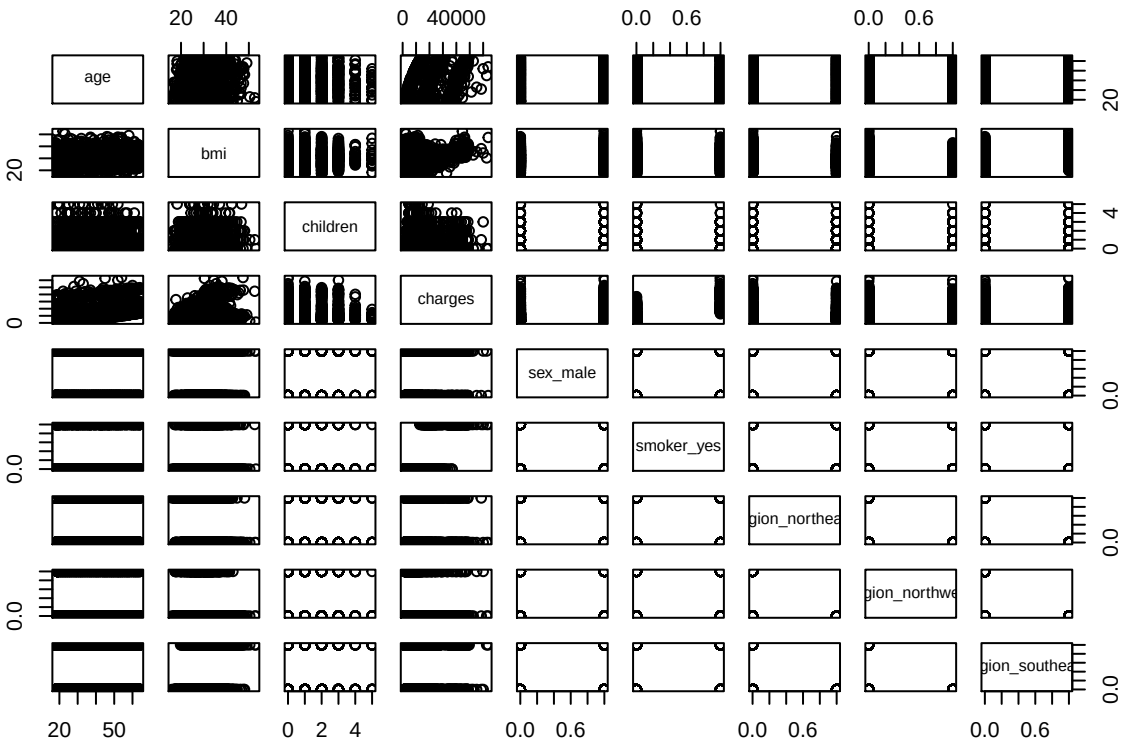
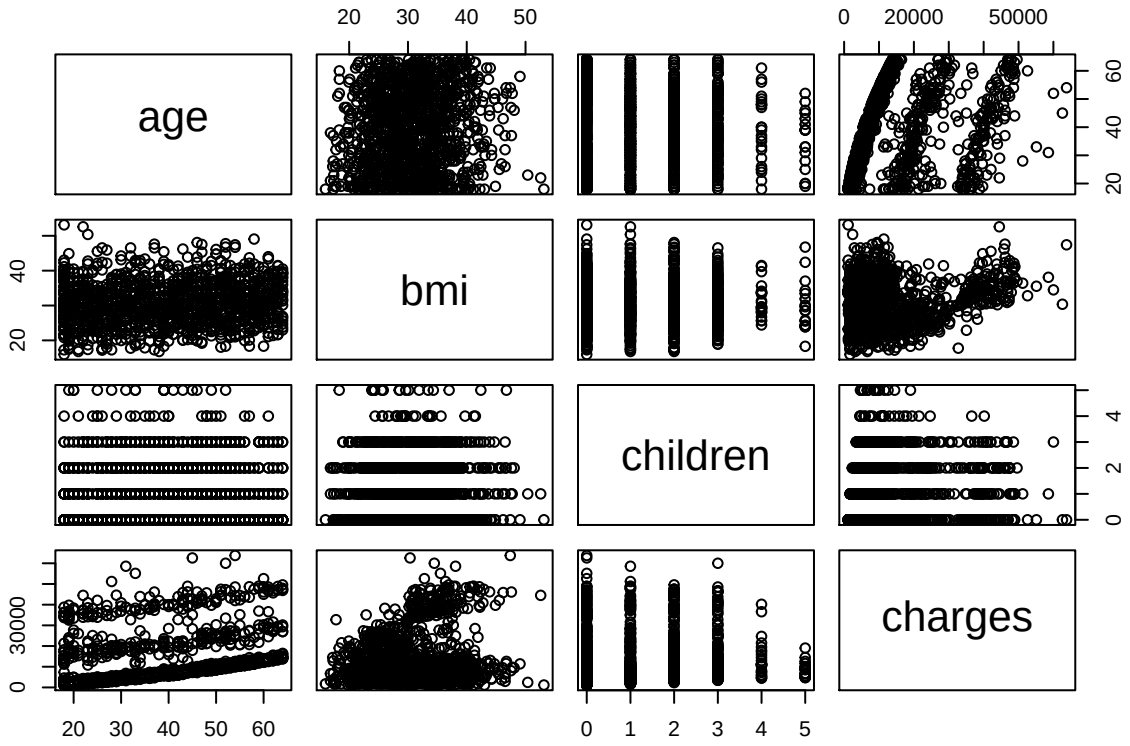
```
insurance <- read.csv("~/Downloads/insurance 2.csv")
library(fastDummies)
insurance_group19 <- dummy_cols(insurance, select_columns =c('sex','smoker', 'region'),remove_selected_
insurance_group19<-insurance_group19[,-5]
insurance_group19<-insurance_group19[,-6]
insurance_group19<-insurance_group19[,-10]
```

#Data Description

```
pairs(insurance_group19)
```

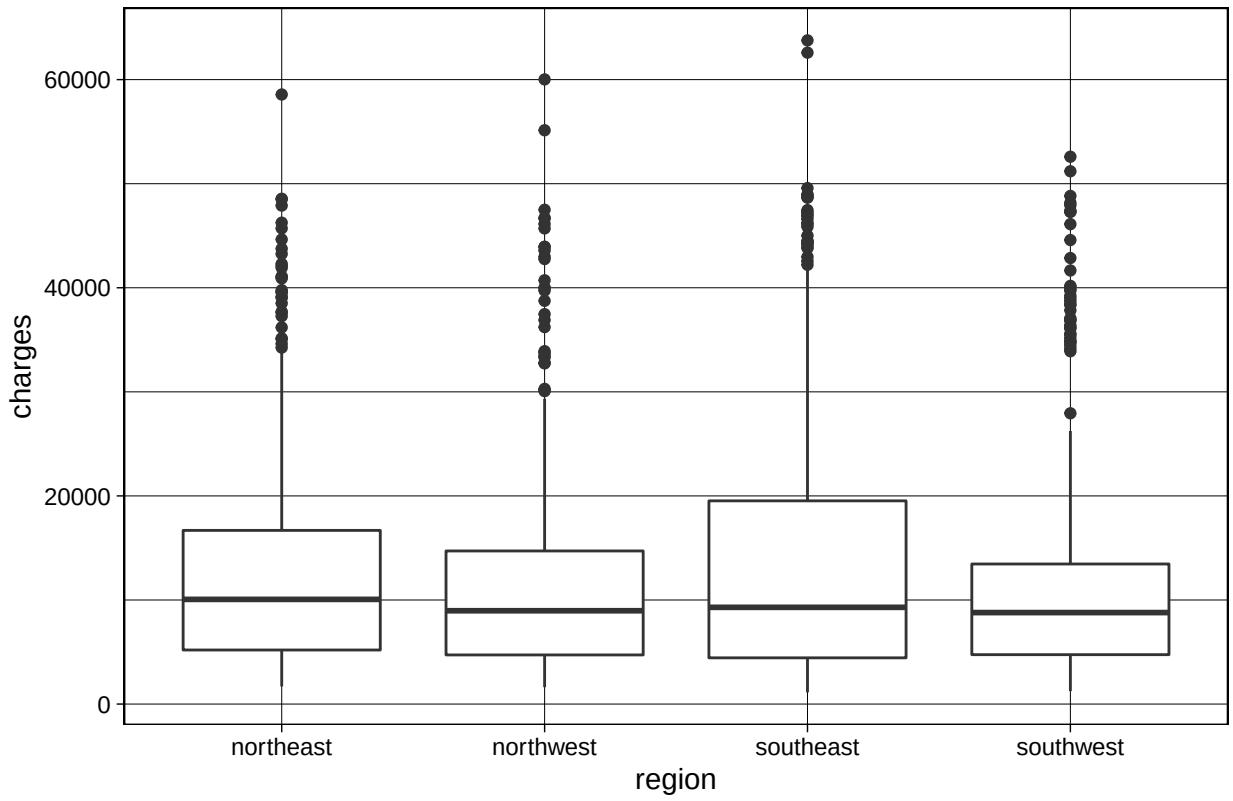



```
pairs(insurance_group19[c("age", "bmi", "children", "charges")])
```

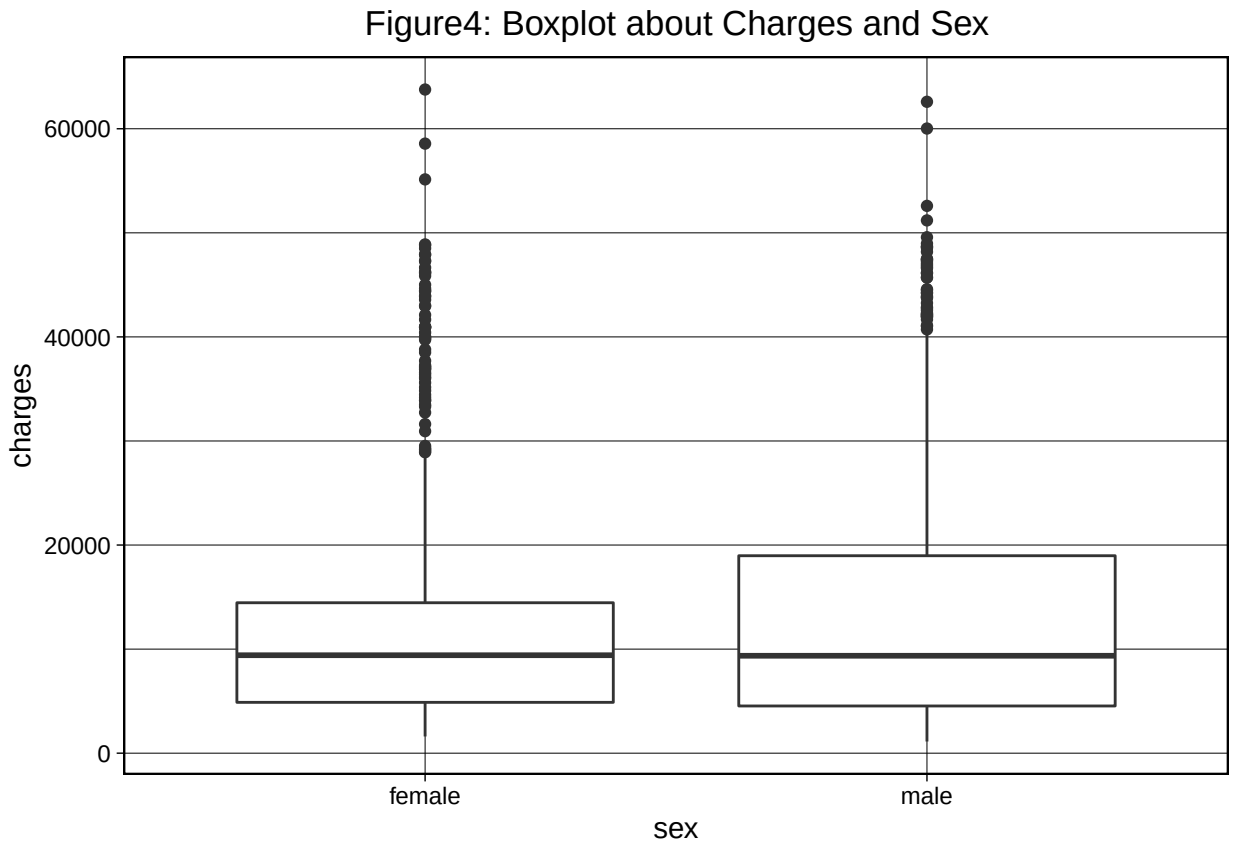



```
library(ggplot2)
ggplot(data = insurance,aes(region,charges))+geom_boxplot()+ggtitle("
```

Figure3: Boxplot about Charges and Region

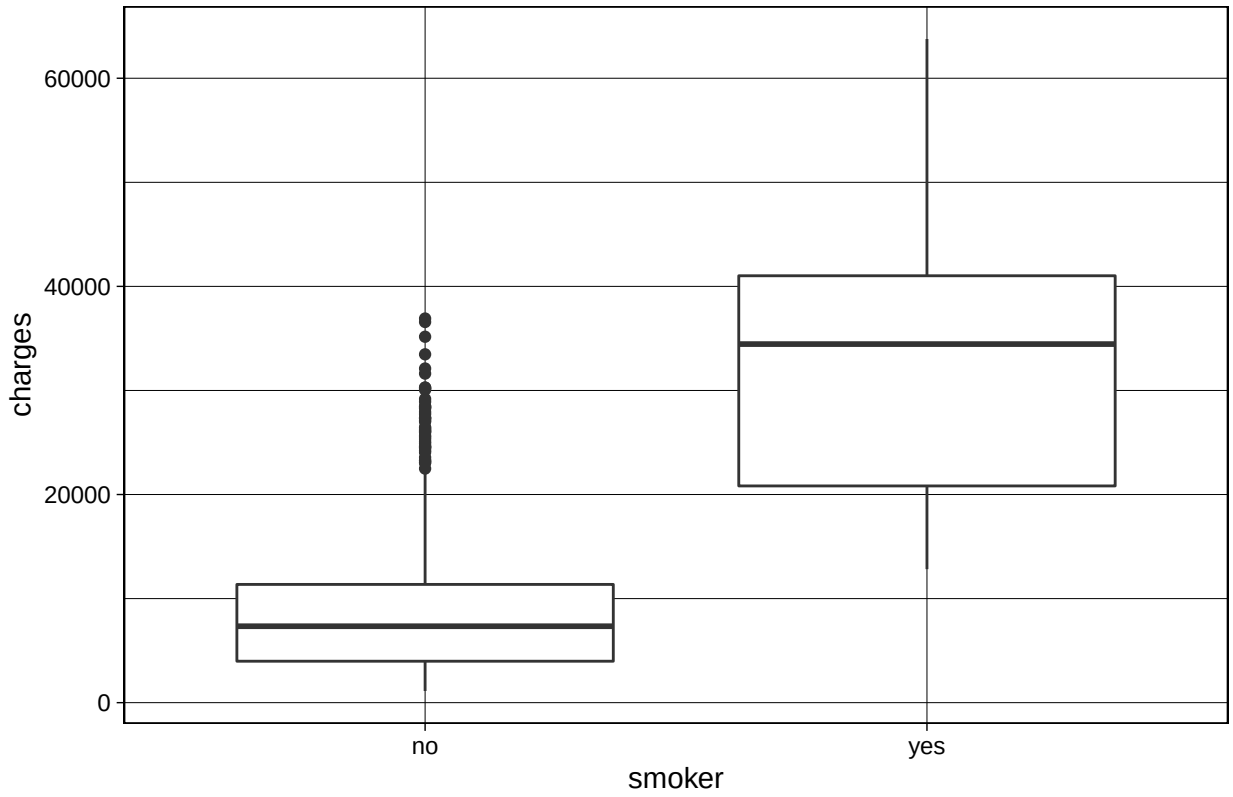


```
ggplot(data = insurance,aes(sex,charges))+geom_boxplot()+ggtitle("
```
Figure4: Boxp




```
ggplot(data = insurance,aes(smoker,charges))+geom_boxplot()+ggtitle("
```
Figure5: B

Figure5: Boxplot about Charges and Smoker




```
cor(insurance_group19)
```

```
##                           age          bmi     children      charges
## age               1.0000000000  0.109520133  0.043353356  0.299667234
## bmi               0.1095201325  1.000000000  0.012972958  0.198494788
## children          0.0433533561  0.012972958  1.000000000  0.068597849
## charges           0.2996672342  0.198494788  0.068597849  1.000000000
## sex_male         -0.0218433314  0.046114466  0.016480358  0.056551166
## smoker_yes       -0.0244810170  0.003871969  0.008014232  0.787270688
## region_northeast  0.0030497862 -0.137998443 -0.022414438  0.006746237
## region_northwest  0.0001709373 -0.135837742  0.025179708 -0.039485507
## region_southeast -0.0110083540  0.270123539 -0.022640779  0.074379344
##                      sex_male   smoker_yes region_northeast region_northwest
## age              -0.021843331 -0.024481017      0.003049786     0.0001709373
## bmi               0.046114466  0.003871969     -0.137998443    -0.1358377418
## children          0.016480358  0.008014232     -0.022414438     0.0251797085
## charges           0.056551166  0.787270688      0.006746237    -0.0394855074
## sex_male          1.000000000  0.075774782     -0.002841816    -0.0115687243
## smoker_yes        0.075774782  1.000000000      0.003024712    -0.0367228785
## region_northeast -0.002841816  0.003024712      1.000000000    -0.3198616601
## region_northwest -0.011568724 -0.036722878     -0.319861660     1.0000000000
## region_southeast  0.016656660  0.068713897     -0.345213577    -0.3459163478
##                  region_southeast
## age                   -0.01100835
## bmi                    0.27012354
## children              -0.02264078
## charges                0.07437934
```

```
## sex_male               0.01665666
## smoker_yes             0.06871390
## region_northeast      -0.34521358
## region_northwest      -0.34591635
## region_southeast       1.00000000
```

#Model building

```
model<-lm(charges~age+bmi+children+sex_male+smoker_yes+region_northeast+region_northwest+region_southeas
summary(model)
```

```
##
## Call:
## lm(formula = charges ~ age + bmi + children + sex_male + smoker_yes +
##     region_northeast + region_northwest + region_southeast, data = insurance_group19)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11304.7  -2846.1   -978.6   1388.6  29992.8
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -12898.97    1019.78 -12.649  < 2e-16 ***
## age                256.86      11.89  21.609  < 2e-16 ***
## bmi                339.19      28.59  11.865  < 2e-16 ***
## children           475.53     137.71   3.453 0.000572 ***
## sex_male          -131.40     332.69  -0.395 0.692927
## smoker_yes       23848.59     412.95  57.751  < 2e-16 ***
## region_northeast   960.23     477.37   2.011 0.044475 *
## region_northwest   607.26     476.65   1.274 0.202879
## region_southeast   -74.80     470.10  -0.159 0.873607
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6060 on 1330 degrees of freedom
## Multiple R-squared:  0.7511, Adjusted R-squared:  0.7496
## F-statistic: 501.6 on 8 and 1330 DF,  p-value: < 2.2e-16
```

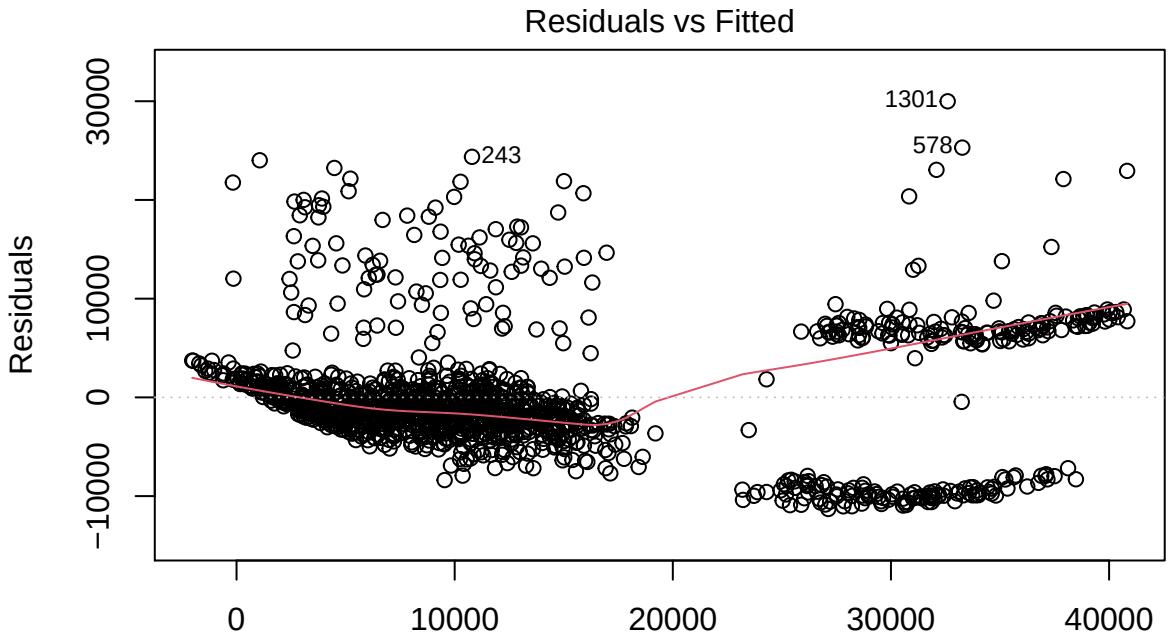
```
library(car)
```

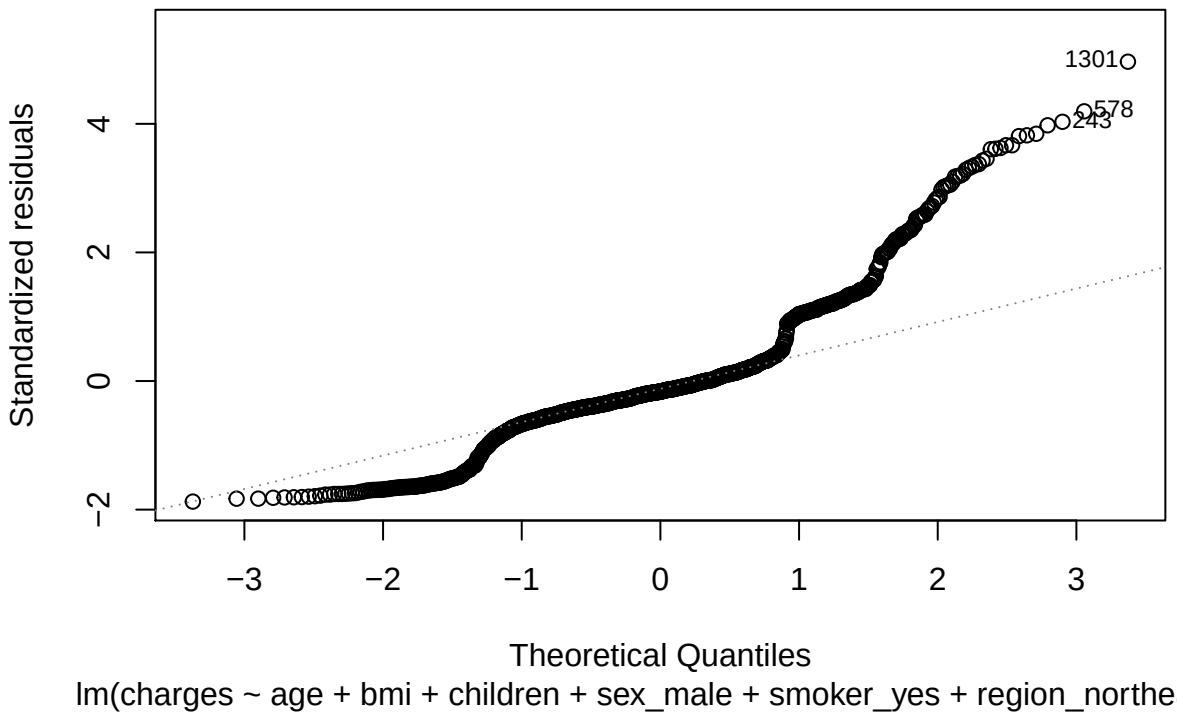
```
## Loading required package: carData
```

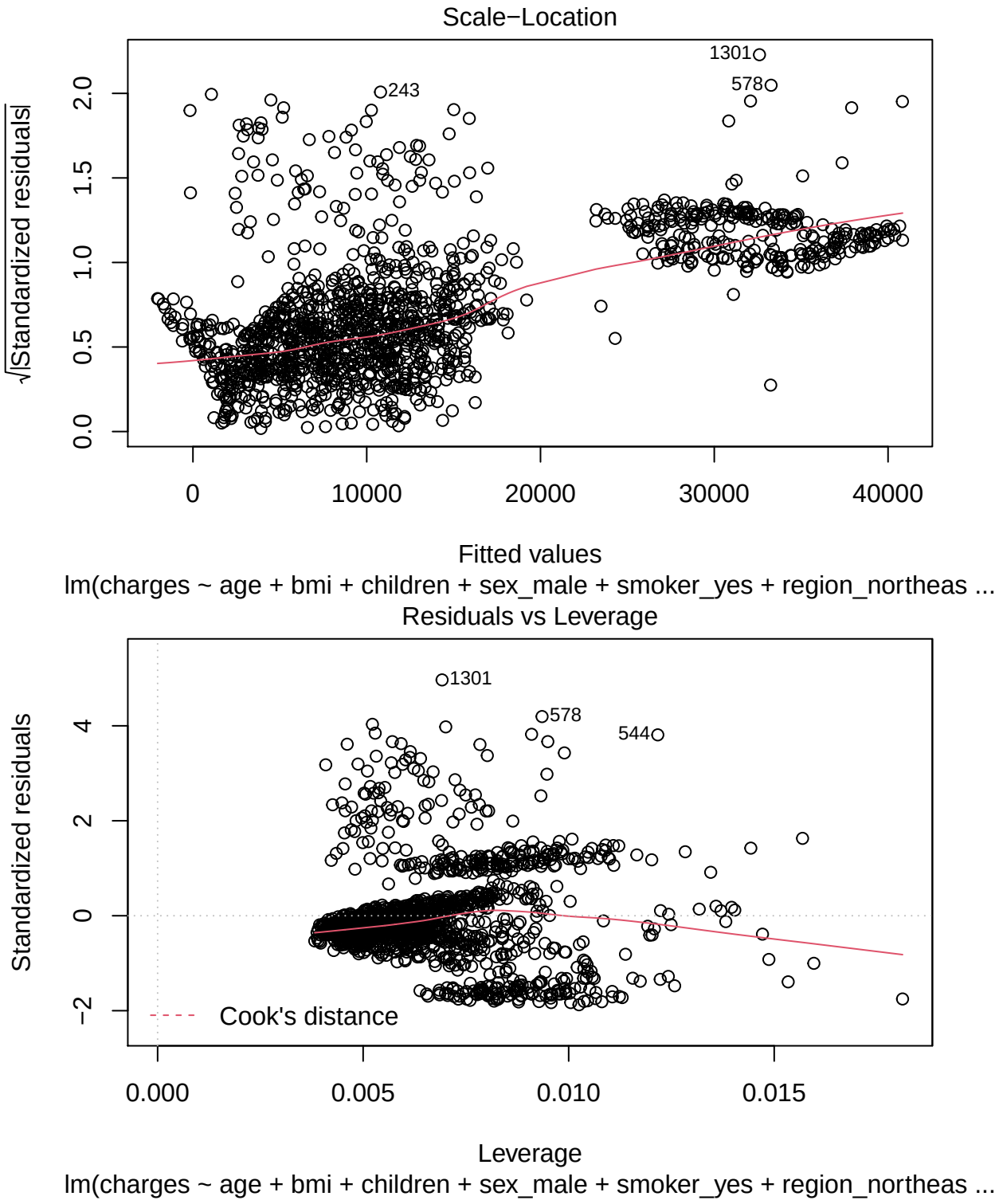
```
vif(model)
```

```
##              age              bmi         children         sex_male
##         1.016956         1.106681         1.004025         1.008840
##       smoker_yes region_northeast region_northwest region_southeast
##         1.012045         1.524145         1.522734         1.595104
```

```
plot(model)
```

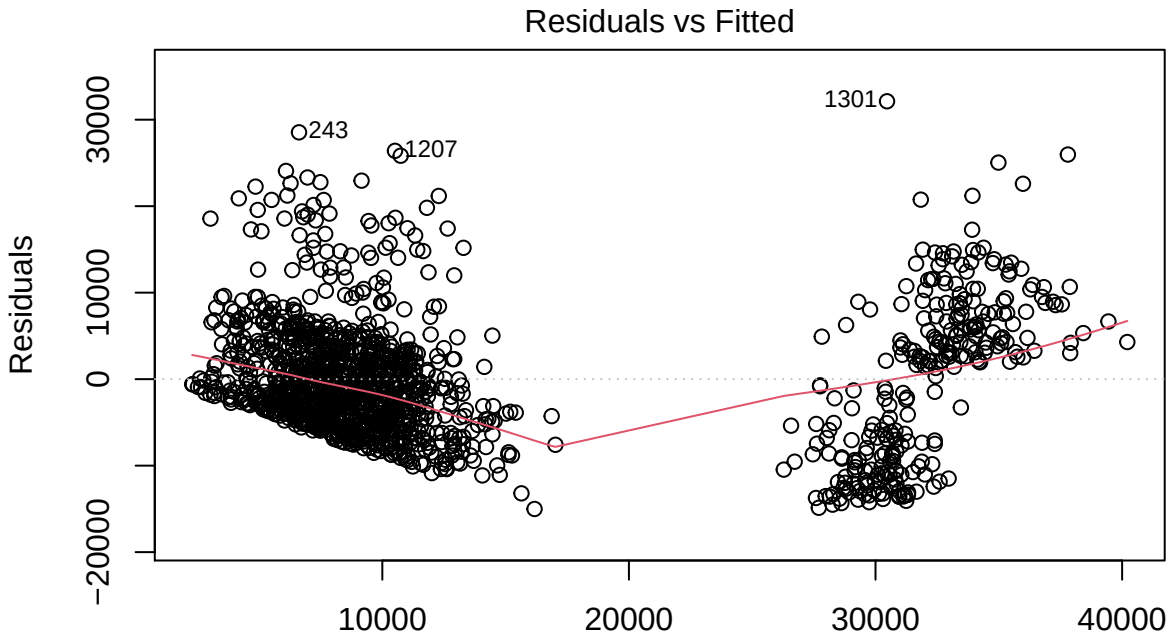

Residuals vs Fitted
Residuals
30000
10000
0
−10000
1301
578
243
0
10000
20000
30000
40000
Fitted values
lm(charges ~ age + bmi + children + sex_male + smoker_yes + region_northeas ...

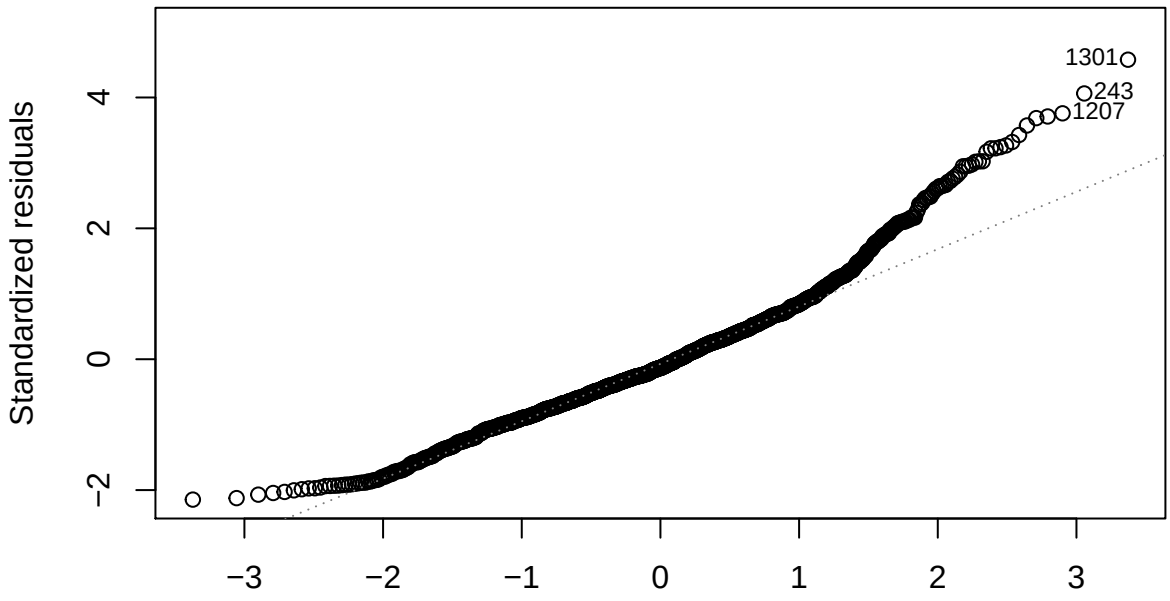

Normal Q−Q
Standardized residuals
4
2
0
−2
1301
578
243
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(charges ~ age + bmi + children + sex_male + smoker_yes + region_northeas ...

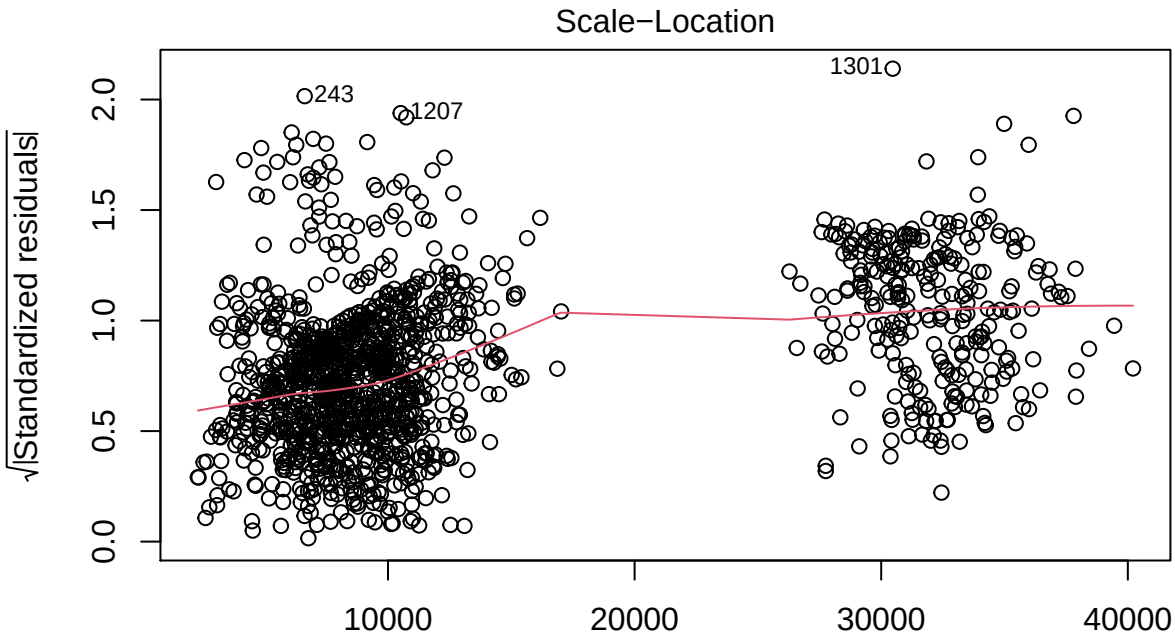


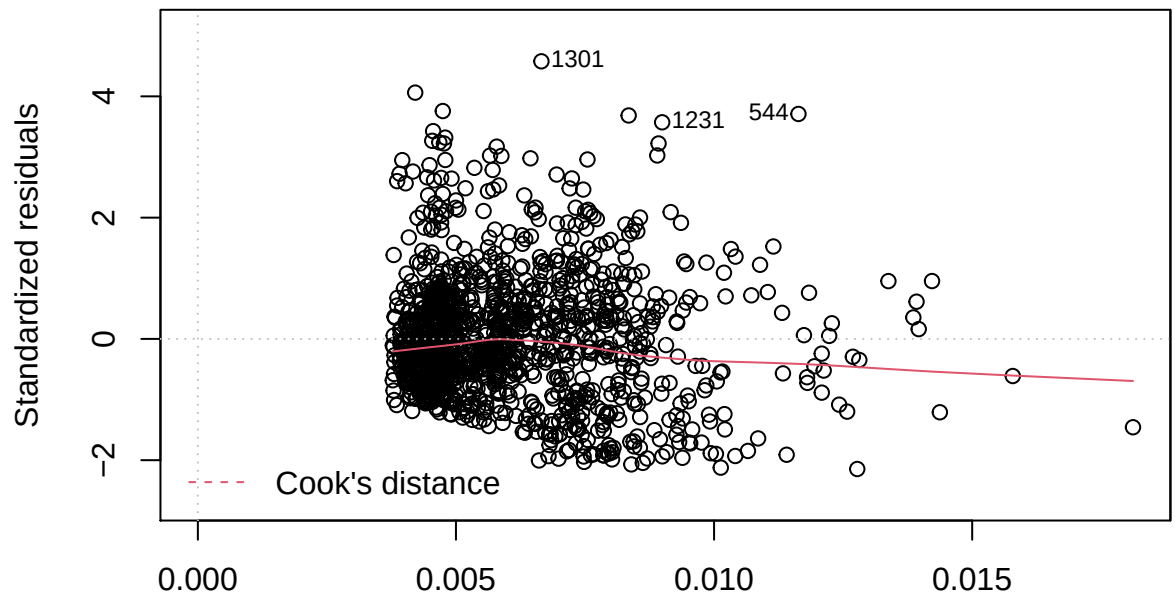
#(a) full-model transformation

```
plot(lm(charges~1/age+bmi+children+sex_male+smoker_yes+region_northeast+region_northwest+region_southeas
```


Residuals vs Fitted
Residuals
30000
10000
0
−20000
243
1207
1301
10000
20000
30000
40000
Fitted values
lm(charges ~ 1/age + bmi + children + sex_male + smoker_yes + region_northe ...


Normal Q−Q
Standardized residuals
4
2
0
−2
1301
243
1207
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(charges ~ 1/age + bmi + children + sex_male + smoker_yes + region_northe ...






```
summary(lm(charges~1/age+bmi+children+sex_male+smoker_yes+region_northeast+region_northwest+region_soutl
```

```
##
## Call:
## lm(formula = charges ~ 1/age + bmi + children + sex_male + smoker_yes +
##     region_northeast + region_northwest + region_southeast, data = insurance_group19)
```

```
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -15010  -4649   -943   3653  32126
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -5006.19    1106.29  -4.525 6.57e-06 ***
## bmi                411.49      32.99  12.473  < 2e-16 ***
## children           600.25     159.87   3.755 0.000181 ***
## sex_male          -318.36     386.43  -0.824 0.410169
## smoker_yes       23663.46     479.72  49.328  < 2e-16 ***
## region_northeast  1046.81     554.65   1.887 0.059333 .
## region_northwest   654.23     553.83   1.181 0.237703
## region_southeast  -364.45     546.00  -0.667 0.504576
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7041 on 1331 degrees of freedom
## Multiple R-squared:  0.6637, Adjusted R-squared:  0.6619
## F-statistic: 375.2 on 7 and 1331 DF,  p-value: < 2.2e-16
```

#(b) leaps method

```
library(leaps)
data<-as.matrix(insurance_group19)
leaps(x=data[,-4],y=data[,4],method="adjr2")
```

```
## $which
##       1     2     3     4     5     6     7     8
## 1 FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE
## 1  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## 1 FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE
## 1 FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE
## 1 FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE
## 1 FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE
## 1 FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE
## 1 FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE
## 2  TRUE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE
## 2 FALSE  TRUE FALSE FALSE  TRUE FALSE FALSE FALSE
## 2 FALSE FALSE  TRUE FALSE  TRUE FALSE FALSE FALSE
## 2 FALSE FALSE FALSE FALSE  TRUE FALSE FALSE  TRUE
## 2 FALSE FALSE FALSE FALSE  TRUE FALSE  TRUE FALSE
## 2 FALSE FALSE FALSE FALSE  TRUE  TRUE FALSE FALSE
## 2 FALSE FALSE FALSE  TRUE  TRUE FALSE FALSE FALSE
## 2  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE
## 2  TRUE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE
## 2  TRUE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE
## 3  TRUE  TRUE FALSE FALSE  TRUE FALSE FALSE FALSE
## 3  TRUE FALSE  TRUE FALSE  TRUE FALSE FALSE FALSE
## 3  TRUE FALSE FALSE FALSE  TRUE FALSE FALSE  TRUE
## 3  TRUE FALSE FALSE FALSE  TRUE FALSE  TRUE FALSE
## 3  TRUE FALSE FALSE FALSE  TRUE  TRUE FALSE FALSE
## 3  TRUE FALSE FALSE  TRUE  TRUE FALSE FALSE FALSE
## 3 FALSE  TRUE  TRUE FALSE  TRUE FALSE FALSE FALSE
```

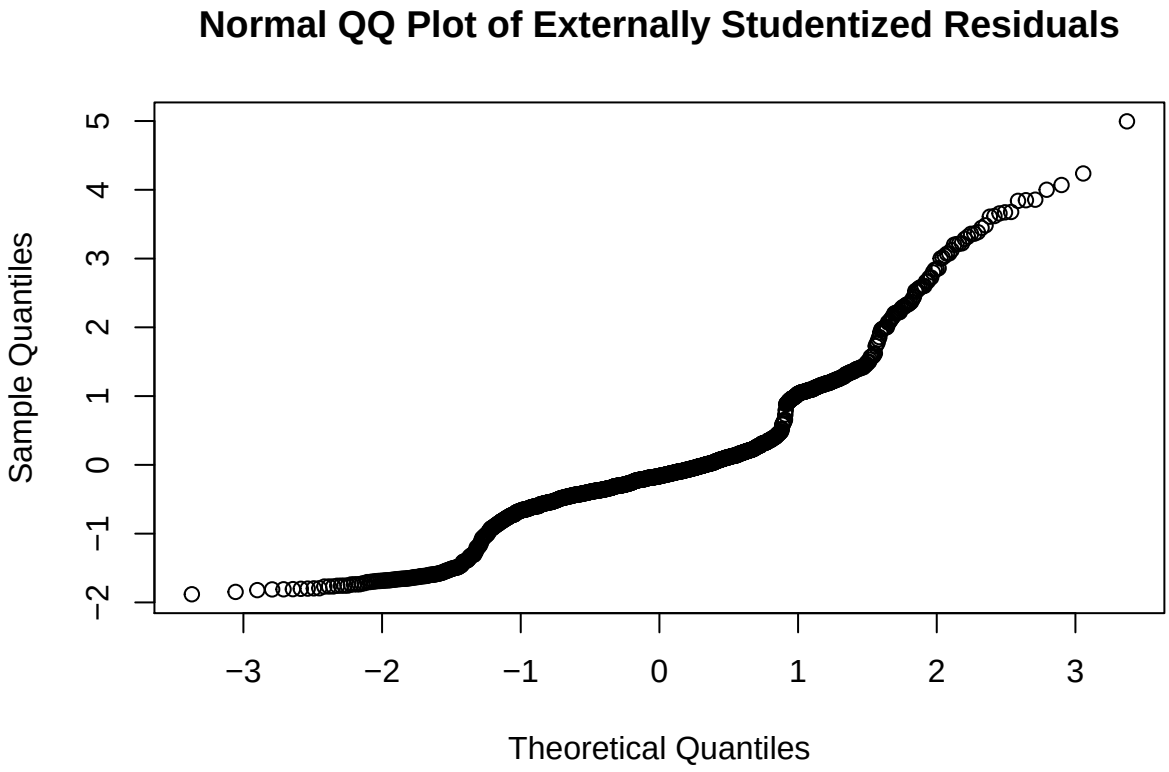
```
## 3 FALSE  TRUE FALSE FALSE  TRUE FALSE FALSE  TRUE
## 3 FALSE  TRUE FALSE FALSE  TRUE  TRUE FALSE FALSE
## 3 FALSE  TRUE FALSE FALSE  TRUE FALSE  TRUE FALSE
## 4  TRUE  TRUE  TRUE FALSE  TRUE FALSE FALSE FALSE
## 4  TRUE  TRUE FALSE FALSE  TRUE  TRUE FALSE FALSE
## 4  TRUE  TRUE FALSE FALSE  TRUE FALSE FALSE  TRUE
## 4  TRUE  TRUE FALSE FALSE  TRUE FALSE  TRUE FALSE
## 4  TRUE  TRUE FALSE  TRUE  TRUE FALSE FALSE FALSE
## 4  TRUE FALSE  TRUE FALSE  TRUE FALSE FALSE  TRUE
## 4  TRUE FALSE  TRUE FALSE  TRUE FALSE  TRUE FALSE
## 4  TRUE FALSE  TRUE FALSE  TRUE  TRUE FALSE FALSE
## 4  TRUE FALSE  TRUE  TRUE  TRUE FALSE FALSE FALSE
## 4  TRUE FALSE FALSE FALSE  TRUE  TRUE FALSE  TRUE
## 5  TRUE  TRUE  TRUE FALSE  TRUE  TRUE FALSE FALSE
## 5  TRUE  TRUE  TRUE FALSE  TRUE FALSE FALSE  TRUE
## 5  TRUE  TRUE  TRUE FALSE  TRUE FALSE  TRUE FALSE
## 5  TRUE  TRUE  TRUE  TRUE  TRUE FALSE FALSE FALSE
## 5  TRUE  TRUE FALSE FALSE  TRUE  TRUE  TRUE FALSE
## 5  TRUE  TRUE FALSE FALSE  TRUE  TRUE FALSE  TRUE
## 5  TRUE  TRUE FALSE  TRUE  TRUE  TRUE FALSE FALSE
## 5  TRUE  TRUE FALSE FALSE  TRUE FALSE  TRUE  TRUE
## 5  TRUE  TRUE FALSE  TRUE  TRUE FALSE FALSE  TRUE
## 5  TRUE  TRUE FALSE  TRUE  TRUE FALSE  TRUE FALSE
## 6  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE FALSE
## 6  TRUE  TRUE  TRUE FALSE  TRUE  TRUE FALSE  TRUE
## 6  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE FALSE
## 6  TRUE  TRUE  TRUE  TRUE  TRUE FALSE FALSE  TRUE
## 6  TRUE  TRUE  TRUE FALSE  TRUE FALSE  TRUE  TRUE
## 6  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE FALSE
## 6  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE FALSE
## 6  TRUE  TRUE FALSE FALSE  TRUE  TRUE  TRUE  TRUE
## 6  TRUE  TRUE FALSE  TRUE  TRUE  TRUE FALSE  TRUE
## 6  TRUE  TRUE FALSE  TRUE  TRUE FALSE  TRUE  TRUE
## 7  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE
## 7  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE
## 7  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE
## 7  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE
## 7  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE
## 7  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## 7 FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## 7  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE
## 8  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
##
## $label
## [1] "(Intercept)" "1"           "2"           "3"           "4"
## [6] "5"           "6"           "7"           "8"
##
## $size
##  [1] 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 4 4 4 4 4 4 4 4 4 4 5 5 5 5 5 5 5 5 5
## [39] 6 6 6 6 6 6 6 6 6 6 7 7 7 7 7 7 7 7 7 8 8 8 8 8 8 8 8 9
##
## $adjr2
##  [1]  0.6195107643  0.0891196737  0.0386817069  0.0047884815  0.0039612412
##  [6]  0.0024524831  0.0008123283 -0.0007023974  0.7211623318  0.6574830596
```

```
## [11]  0.6231118793  0.6196399338  0.6193381091  0.6192450487  0.6192356726
## [16]  0.1162609653  0.0944815430  0.0924269586  0.7470747518  0.7233049746
## [21]  0.7214982897  0.7210608131  0.7209648341  0.7209642794  0.6608064837
## [26]  0.6583713270  0.6582299533  0.6574865080  0.7491113526  0.7475680695
## [31]  0.7473535119  0.7470248768  0.7469058518  0.7236961277  0.7232323012
## [36]  0.7231175422  0.7231036806  0.7214386762  0.7496614247  0.7493403047
## [41]  0.7490346093  0.7489515239  0.7478727512  0.7475730617  0.7474003195
## [46]  0.7471901930  0.7471850883  0.7468555722  0.7499250429  0.7496243115
## [51]  0.7495030973  0.7491808689  0.7491698438  0.7488743016  0.7477048973
## [56]  0.7476958950  0.7474052954  0.7470213774  0.7497664843  0.7497418967
## [61]  0.7494658688  0.7490100086  0.7475278311  0.7232857186  0.6619205006
## [66]  0.1222751890  0.7495831053
```

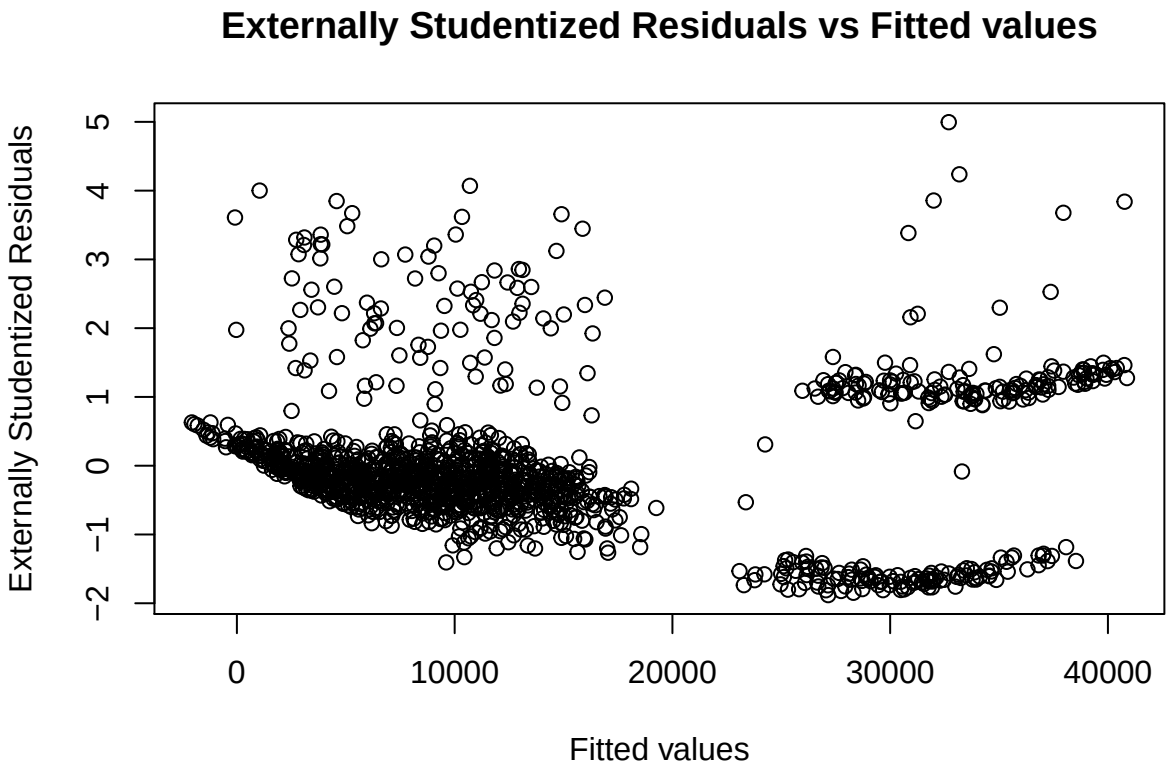
```
lmbest<-lm(charges~age+bmi+children+smoker_yes+region_northeast+region_northwest,data=insurance_group19)
summary(lmbest)
```

```
##
## Call:
## lm(formula = charges ~ age + bmi + children + smoker_yes + region_northeast +
##     region_northwest, data = insurance_group19)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11333.2  -2825.8   -999.4   1373.6  29902.6
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -12967.02    1005.35 -12.898  < 2e-16 ***
## age                257.04      11.87  21.654  < 2e-16 ***
## bmi                337.89      28.12  12.016  < 2e-16 ***
## children           475.26     137.54   3.455 0.000567 ***
## smoker_yes        23832.15     410.64  58.036  < 2e-16 ***
## region_northeast    996.89     415.86   2.397 0.016660 *
## region_northwest    644.54     415.60   1.551 0.121171
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6056 on 1332 degrees of freedom
## Multiple R-squared:  0.751,  Adjusted R-squared:  0.7499
## F-statistic: 669.7 on 6 and 1332 DF,  p-value: < 2.2e-16
```

```
qqnorm(rstudent(lmbest),main="Normal QQ Plot of Externally Studentized Residuals")
```




```
plot(fitted(lmbest),rstudent(lmbest),xlab="Fitted values",ylab="Externally Studentized Residuals", main=
```



#(c) stepwise procedure method

```r
nullmodel<-lm(charges~1,data=insurance_group19)
fullmodel1<-lm(charges~age+bmi+children+sex_male+smoker_yes+region_northeast+region_northwest+region_so
step1<-step(nullmodel,scope=list(lower=nullmodel,upper=fullmodel1),direction="forward")
```

```
## Start:  AIC=25178.89
## charges ~ 1
##
##                    Df  Sum of Sq        RSS   AIC
## + smoker_yes        1 1.2161e+11 7.4598e+10 23886
## + age               1 1.7619e+10 1.7859e+11 25055
## + bmi               1 7.7306e+09 1.8848e+11 25127
## + region_southeast  1 1.0855e+09 1.9512e+11 25174
## + children          1 9.2328e+08 1.9528e+11 25175
## + sex_male          1 6.2747e+08 1.9558e+11 25177
## + region_northwest  1 3.0591e+08 1.9590e+11 25179
## <none>                           1.9621e+11 25179
## + region_northeast  1 8.9297e+06 1.9620e+11 25181
##
## Step:  AIC=23886.01
## charges ~ smoker_yes
##
##                    Df  Sum of Sq        RSS   AIC
## + age               1 1.9971e+10 5.4628e+10 23471
## + bmi               1 7.4951e+09 6.7103e+10 23746
## + children          1 7.6130e+08 7.3837e+10 23874
## <none>                           7.4598e+10 23886
## + region_southeast  1 8.1101e+07 7.4517e+10 23887
## + region_northwest  1 2.1970e+07 7.4577e+10 23888
## + region_northeast  1 3.7383e+06 7.4595e+10 23888
## + sex_male          1 1.9014e+06 7.4597e+10 23888
##
## Step:  AIC=23470.81
## charges ~ smoker_yes + age
##
##                    Df  Sum of Sq        RSS   AIC
## + bmi               1 5113662989 4.9514e+10 23341
## + children          1  460345980 5.4168e+10 23462
## + region_southeast  1  106658306 5.4521e+10 23470
## <none>                           5.4628e+10 23471
## + region_northwest  1   21015210 5.4607e+10 23472
## + region_northeast  1    2225764 5.4626e+10 23473
## + sex_male          1    2117174 5.4626e+10 23473
##
## Step:  AIC=23341.21
## charges ~ smoker_yes + age + bmi
##
##                    Df Sum of Sq        RSS   AIC
## + children          1 435488078 4.9079e+10 23331
## + region_northeast  1 133591833 4.9381e+10 23340
## + region_southeast  1  91620188 4.9423e+10 23341
## <none>                          4.9514e+10 23341
## + region_northwest  1  27332762 4.9487e+10 23342
## + sex_male          1   4049169 4.9510e+10 23343
##
```

```
## Step:  AIC=23331.38
## charges ~ smoker_yes + age + bmi + children
##
##                    Df Sum of Sq        RSS   AIC
## + region_northeast  1 144314793 4.8934e+10 23329
## + region_southeast  1  81544539 4.8997e+10 23331
## <none>                          4.9079e+10 23331
## + region_northwest  1  21789402 4.9057e+10 23333
## + sex_male          1   5548457 4.9073e+10 23333
##
## Step:  AIC=23329.44
## charges ~ smoker_yes + age + bmi + children + region_northeast
##
##                    Df Sum of Sq        RSS   AIC
## + region_northwest  1  88201494 4.8846e+10 23329
## <none>                          4.8934e+10 23329
## + region_southeast  1  29460778 4.8905e+10 23331
## + sex_male          1   5784479 4.8929e+10 23331
##
## Step:  AIC=23329.02
## charges ~ smoker_yes + age + bmi + children + region_northeast +
##     region_northwest
##
##                    Df Sum of Sq        RSS   AIC
## <none>                          4.8846e+10 23329
## + sex_male          1   5723917 4.8840e+10 23331
## + region_southeast  1    924920 4.8845e+10 23331
```

```
fullmodel2=lm(charges~smoker_yes*age+smoker_yes*bmi+age*bmi+age*children+bmi*region_northeast+bmi*regio
step2<-step(fullmodel2,direction="backward")
```

```
## Start:  AIC=22736.54
## charges ~ smoker_yes * age + smoker_yes * bmi + age * bmi + age *
##     children + bmi * region_northeast + bmi * region_northwest +
##     bmi * region_southeast + children * region_northeast + children *
##     region_northwest + children * region_southeast
##
##                              Df  Sum of Sq        RSS   AIC
## - age:children                1 1.1200e+02 3.0869e+10 22734
## - smoker_yes:age              1 1.4726e+06 3.0871e+10 22735
## - children:region_southeast   1 9.1764e+06 3.0879e+10 22735
## - bmi:region_northwest        1 2.2214e+07 3.0892e+10 22736
## - bmi:region_northeast        1 2.6000e+07 3.0895e+10 22736
## - children:region_northeast   1 3.1603e+07 3.0901e+10 22736
## - age:bmi                     1 3.6174e+07 3.0906e+10 22736
## <none>                                     3.0869e+10 22736
## - bmi:region_southeast        1 6.9594e+07 3.0939e+10 22738
## - children:region_northwest   1 9.2895e+07 3.0962e+10 22739
## - smoker_yes:bmi              1 1.7553e+10 4.8423e+10 23337
##
## Step:  AIC=22734.54
## charges ~ smoker_yes + age + bmi + children + region_northeast +
##     region_northwest + region_southeast + smoker_yes:age + smoker_yes:bmi +
##     age:bmi + bmi:region_northeast + bmi:region_northwest + bmi:region_southeast +
##     children:region_northeast + children:region_northwest + children:region_southeast
```

```
##
##                             Df  Sum of Sq        RSS   AIC
## - smoker_yes:age             1 1.4744e+06 3.0871e+10 22733
## - children:region_southeast  1 9.2813e+06 3.0879e+10 22733
## - bmi:region_northwest       1 2.2249e+07 3.0892e+10 22734
## - bmi:region_northeast       1 2.6001e+07 3.0895e+10 22734
## - children:region_northeast  1 3.1606e+07 3.0901e+10 22734
## - age:bmi                    1 3.6186e+07 3.0906e+10 22734
## <none>                                    3.0869e+10 22734
## - bmi:region_southeast       1 6.9600e+07 3.0939e+10 22736
## - children:region_northwest  1 9.3417e+07 3.0963e+10 22737
## - smoker_yes:bmi             1 1.7553e+10 4.8423e+10 23335
##
## Step:  AIC=22732.6
## charges ~ smoker_yes + age + bmi + children + region_northeast +
##     region_northwest + region_southeast + smoker_yes:bmi + age:bmi +
##     bmi:region_northeast + bmi:region_northwest + bmi:region_southeast +
##     children:region_northeast + children:region_northwest + children:region_southeast
##
##                             Df  Sum of Sq        RSS   AIC
## - children:region_southeast  1 9.7434e+06 3.0881e+10 22731
## - bmi:region_northwest       1 2.2092e+07 3.0893e+10 22732
## - bmi:region_northeast       1 2.5665e+07 3.0897e+10 22732
## - children:region_northeast  1 3.1611e+07 3.0902e+10 22732
## - age:bmi                    1 3.5846e+07 3.0907e+10 22732
## <none>                                    3.0871e+10 22733
## - bmi:region_southeast       1 6.9604e+07 3.0940e+10 22734
## - children:region_northwest  1 9.3523e+07 3.0964e+10 22735
## - smoker_yes:bmi             1 1.7630e+10 4.8501e+10 23336
##
## Step:  AIC=22731.03
## charges ~ smoker_yes + age + bmi + children + region_northeast +
##     region_northwest + region_southeast + smoker_yes:bmi + age:bmi +
##     bmi:region_northeast + bmi:region_northwest + bmi:region_southeast +
##     children:region_northeast + children:region_northwest
##
##                             Df  Sum of Sq        RSS   AIC
## - children:region_northeast  1 2.2106e+07 3.0903e+10 22730
## - bmi:region_northwest       1 2.3134e+07 3.0904e+10 22730
## - bmi:region_northeast       1 2.6781e+07 3.0907e+10 22730
## - age:bmi                    1 3.7646e+07 3.0918e+10 22731
## <none>                                    3.0881e+10 22731
## - bmi:region_southeast       1 6.8567e+07 3.0949e+10 22732
## - children:region_northwest  1 8.6522e+07 3.0967e+10 22733
## - smoker_yes:bmi             1 1.7668e+10 4.8549e+10 23335
##
## Step:  AIC=22729.98
## charges ~ smoker_yes + age + bmi + children + region_northeast +
##     region_northwest + region_southeast + smoker_yes:bmi + age:bmi +
##     bmi:region_northeast + bmi:region_northwest + bmi:region_southeast +
##     children:region_northwest
##
##                             Df  Sum of Sq        RSS   AIC
## - bmi:region_northwest       1 2.3926e+07 3.0927e+10 22729
```

```
## - bmi:region_northeast       1 2.7022e+07 3.0930e+10 22729
## - age:bmi                    1 3.7535e+07 3.0940e+10 22730
## <none>                                    3.0903e+10 22730
## - bmi:region_southeast       1 6.6731e+07 3.0969e+10 22731
## - children:region_northwest  1 6.8243e+07 3.0971e+10 22731
## - smoker_yes:bmi             1 1.7716e+10 4.8619e+10 23335
##
## Step:  AIC=22729.02
## charges ~ smoker_yes + age + bmi + children + region_northeast +
##     region_northwest + region_southeast + smoker_yes:bmi + age:bmi +
##     bmi:region_northeast + bmi:region_southeast + children:region_northwest
##
##                              Df  Sum of Sq        RSS   AIC
## - bmi:region_northeast       1 1.0502e+07 3.0937e+10 22728
## - age:bmi                    1 3.5908e+07 3.0963e+10 22729
## <none>                                    3.0927e+10 22729
## - children:region_northwest  1 7.4205e+07 3.1001e+10 22730
## - bmi:region_southeast       1 1.5434e+08 3.1081e+10 22734
## - smoker_yes:bmi             1 1.7705e+10 4.8632e+10 23333
##
## Step:  AIC=22727.47
## charges ~ smoker_yes + age + bmi + children + region_northeast +
##     region_northwest + region_southeast + smoker_yes:bmi + age:bmi +
##     bmi:region_southeast + children:region_northwest
##
##                              Df  Sum of Sq        RSS   AIC
## - age:bmi                    1 3.5234e+07 3.0972e+10 22727
## <none>                                    3.0937e+10 22728
## - children:region_northwest  1 7.3370e+07 3.1011e+10 22729
## - bmi:region_southeast       1 2.2830e+08 3.1165e+10 22735
## - region_northeast           1 2.7162e+08 3.1209e+10 22737
## - smoker_yes:bmi             1 1.7848e+10 4.8785e+10 23335
##
## Step:  AIC=22727
## charges ~ smoker_yes + age + bmi + children + region_northeast +
##     region_northwest + region_southeast + smoker_yes:bmi + bmi:region_southeast +
##     children:region_northwest
##
##                              Df  Sum of Sq        RSS   AIC
## <none>                                    3.0972e+10 22727
## - children:region_northwest  1 7.6008e+07 3.1048e+10 22728
## - bmi:region_southeast       1 2.2419e+08 3.1197e+10 22735
## - region_northeast           1 2.6770e+08 3.1240e+10 22736
## - age                        1 1.7752e+10 4.8724e+10 23332
## - smoker_yes:bmi             1 1.7813e+10 4.8785e+10 23333
```

```
anova(step1,step2)
```

```
## Analysis of Variance Table
##
## Model 1: charges ~ smoker_yes + age + bmi + children + region_northeast +
##     region_northwest
## Model 2: charges ~ smoker_yes + age + bmi + children + region_northeast +
##     region_northwest + region_southeast + smoker_yes:bmi + bmi:region_southeast +
##     children:region_northwest
```

```
##   Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
## 1   1332 4.8846e+10
## 2   1328 3.0972e+10  4 1.7874e+10 191.59 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(step2)
```

```
##
## Call:
## lm(formula = charges ~ smoker_yes + age + bmi + children + region_northeast +
##     region_northwest + region_southeast + smoker_yes:bmi + bmi:region_southeast +
##     children:region_northwest, data = insurance_group19)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13071.0  -1987.1  -1279.5   -291.8  29941.1
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                -4848.301    979.627  -4.949 8.41e-07 ***
## smoker_yes                -20910.607   1651.674 -12.660  < 2e-16 ***
## age                          261.957      9.495  27.589  < 2e-16 ***
## bmi                           67.876     29.626   2.291 0.022112 *
## children                     394.841    124.951   3.160 0.001614 **
## region_northeast            1292.064    381.369   3.388 0.000725 ***
## region_northwest             176.545    484.650   0.364 0.715712
## region_southeast            4879.864   1612.669   3.026 0.002526 **
## smoker_yes:bmi              1457.045     52.722  27.636  < 2e-16 ***
## bmi:region_southeast        -149.996     48.379  -3.100 0.001973 **
## children:region_northwest    471.178    261.001   1.805 0.071258 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4829 on 1328 degrees of freedom
## Multiple R-squared:  0.8421, Adjusted R-squared:  0.841
## F-statistic: 708.5 on 10 and 1328 DF,  p-value: < 2.2e-16
```

```
step3<-step(nullmodel,scope=list(lower=nullmodel,upper=fullmodel2),direction="both")
```

```
## Start:  AIC=25178.89
## charges ~ 1
##
##                     Df  Sum of Sq        RSS   AIC
## + smoker_yes         1 1.2161e+11 7.4598e+10 23886
## + age                1 1.7619e+10 1.7859e+11 25055
## + bmi                1 7.7306e+09 1.8848e+11 25127
## + region_southeast   1 1.0855e+09 1.9512e+11 25174
## + children           1 9.2328e+08 1.9528e+11 25175
## + region_northwest   1 3.0591e+08 1.9590e+11 25179
## <none>                            1.9621e+11 25179
## + region_northeast   1 8.9297e+06 1.9620e+11 25181
##
## Step:  AIC=23886.01
## charges ~ smoker_yes
```

```
##
##                     Df  Sum of Sq        RSS   AIC
## + age                1 1.9971e+10 5.4628e+10 23471
## + bmi                1 7.4951e+09 6.7103e+10 23746
## + children           1 7.6130e+08 7.3837e+10 23874
## <none>                            7.4598e+10 23886
## + region_southeast   1 8.1101e+07 7.4517e+10 23887
## + region_northwest   1 2.1970e+07 7.4577e+10 23888
## + region_northeast   1 3.7383e+06 7.4595e+10 23888
## - smoker_yes         1 1.2161e+11 1.9621e+11 25179
##
## Step:  AIC=23470.81
## charges ~ smoker_yes + age
##
##                     Df  Sum of Sq        RSS   AIC
## + bmi                1 5.1137e+09 4.9514e+10 23341
## + children           1 4.6035e+08 5.4168e+10 23462
## + region_southeast   1 1.0666e+08 5.4521e+10 23470
## <none>                            5.4628e+10 23471
## + smoker_yes:age     1 6.0640e+07 5.4567e+10 23471
## + region_northwest   1 2.1015e+07 5.4607e+10 23472
## + region_northeast   1 2.2258e+06 5.4626e+10 23473
## - age                1 1.9971e+10 7.4598e+10 23886
## - smoker_yes         1 1.2396e+11 1.7859e+11 25055
##
## Step:  AIC=23341.21
## charges ~ smoker_yes + age + bmi
##
##                     Df  Sum of Sq        RSS   AIC
## + smoker_yes:bmi     1 1.7410e+10 3.2104e+10 22763
## + children           1 4.3549e+08 4.9079e+10 23331
## + region_northeast   1 1.3359e+08 4.9381e+10 23340
## + region_southeast   1 9.1620e+07 4.9423e+10 23341
## + smoker_yes:age     1 9.0075e+07 4.9424e+10 23341
## <none>                            4.9514e+10 23341
## + region_northwest   1 2.7333e+07 4.9487e+10 23342
## + age:bmi            1 9.6308e+05 4.9513e+10 23343
## - bmi                1 5.1137e+09 5.4628e+10 23471
## - age                1 1.7589e+10 6.7103e+10 23746
## - smoker_yes         1 1.2362e+11 1.7314e+11 25015
##
## Step:  AIC=22763.05
## charges ~ smoker_yes + age + bmi + smoker_yes:bmi
##
##                     Df  Sum of Sq        RSS   AIC
## + children           1 5.0342e+08 3.1601e+10 22744
## + region_northeast   1 2.2460e+08 3.1879e+10 22756
## + region_southeast   1 9.3264e+07 3.2011e+10 22761
## <none>                            3.2104e+10 22763
## + age:bmi            1 3.8050e+07 3.2066e+10 22764
## + region_northwest   1 1.5737e+07 3.2088e+10 22764
## + smoker_yes:age     1 2.1600e+04 3.2104e+10 22765
## - smoker_yes:bmi     1 1.7410e+10 4.9514e+10 23341
## - age                1 1.8570e+10 5.0674e+10 23372
```

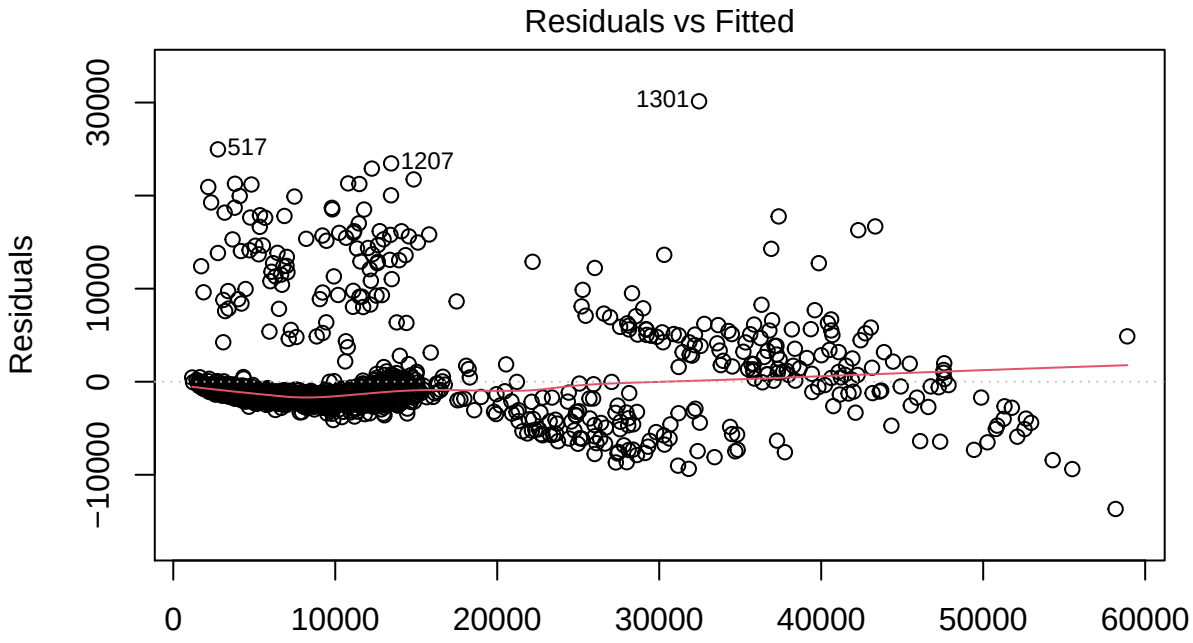
```
##
## Step:  AIC=22743.89
## charges ~ smoker_yes + age + bmi + children + smoker_yes:bmi
##
##                     Df  Sum of Sq        RSS   AIC
## + region_northeast   1 2.3974e+08 3.1361e+10 22736
## + region_southeast   1 8.2354e+07 3.1518e+10 22742
## <none>                             3.1601e+10 22744
## + age:bmi            1 3.4597e+07 3.1566e+10 22744
## + region_northwest   1 1.1300e+07 3.1589e+10 22745
## + age:children       1 5.8750e+05 3.1600e+10 22746
## + smoker_yes:age     1 3.2096e+05 3.1600e+10 22746
## - children           1 5.0342e+08 3.2104e+10 22763
## - smoker_yes:bmi     1 1.7478e+10 4.9079e+10 23331
## - age                1 1.8281e+10 4.9882e+10 23353
##
## Step:  AIC=22735.69
## charges ~ smoker_yes + age + bmi + children + region_northeast +
##     smoker_yes:bmi
##
##                               Df  Sum of Sq        RSS   AIC
## + bmi:region_northeast         1 9.6388e+07 3.1265e+10 22734
## + region_northwest             1 8.5999e+07 3.1275e+10 22734
## <none>                                      3.1361e+10 22736
## + age:bmi                      1 3.1596e+07 3.1329e+10 22736
## + region_southeast             1 1.8350e+07 3.1343e+10 22737
## + children:region_northeast    1 1.9690e+06 3.1359e+10 22738
## + age:children                 1 1.2955e+06 3.1360e+10 22738
## + smoker_yes:age               1 1.9000e+05 3.1361e+10 22738
## - region_northeast             1 2.3974e+08 3.1601e+10 22744
## - children                     1 5.1857e+08 3.1879e+10 22756
## - smoker_yes:bmi               1 1.7574e+10 4.8934e+10 23329
## - age                          1 1.8196e+10 4.9557e+10 23346
##
## Step:  AIC=22733.57
## charges ~ smoker_yes + age + bmi + children + region_northeast +
##     smoker_yes:bmi + bmi:region_northeast
##
##                               Df  Sum of Sq        RSS   AIC
## + region_northwest             1 6.8453e+07 3.1196e+10 22733
## <none>                                      3.1265e+10 22734
## + age:bmi                      1 3.4127e+07 3.1230e+10 22734
## + region_southeast             1 8.9806e+06 3.1256e+10 22735
## + children:region_northeast    1 2.3002e+06 3.1262e+10 22736
## + age:children                 1 1.4388e+06 3.1263e+10 22736
## + smoker_yes:age               1 4.5443e+05 3.1264e+10 22736
## - bmi:region_northeast         1 9.6388e+07 3.1361e+10 22736
## - children                     1 5.2651e+08 3.1791e+10 22754
## - smoker_yes:bmi               1 1.7520e+10 4.8785e+10 23327
## - age                          1 1.8106e+10 4.9371e+10 23343
##
## Step:  AIC=22732.63
## charges ~ smoker_yes + age + bmi + children + region_northeast +
##     region_northwest + smoker_yes:bmi + bmi:region_northeast
```

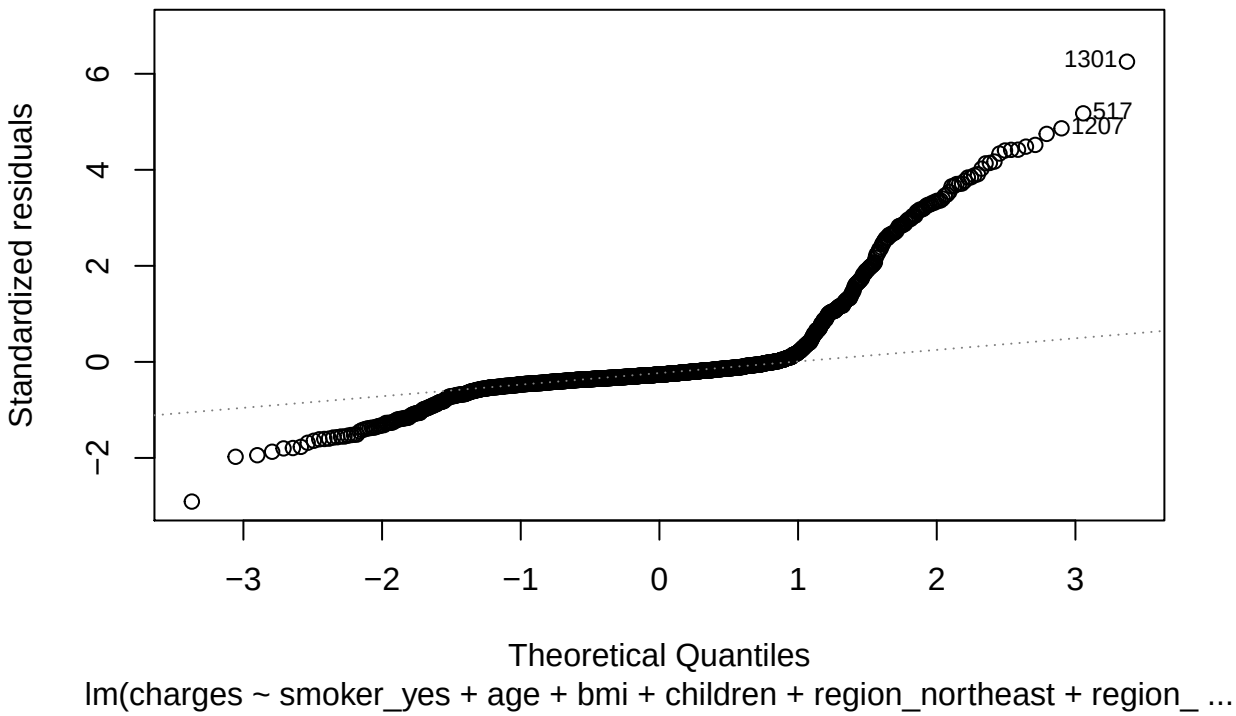
```
##
##                              Df  Sum of Sq        RSS   AIC
## + bmi:region_northwest        1 1.1548e+08 3.1081e+10 22730
## + children:region_northwest   1 8.0114e+07 3.1116e+10 22731
## <none>                                     3.1196e+10 22733
## + age:bmi                     1 3.5513e+07 3.1161e+10 22733
## - region_northwest            1 6.8453e+07 3.1265e+10 22734
## - bmi:region_northeast        1 7.8842e+07 3.1275e+10 22734
## + children:region_northeast   1 2.6319e+06 3.1193e+10 22734
## + region_southeast            1 1.2655e+06 3.1195e+10 22735
## + age:children                1 1.2281e+06 3.1195e+10 22735
## + smoker_yes:age              1 5.2062e+05 3.1196e+10 22735
## - children                    1 5.1762e+08 3.1714e+10 22753
## - smoker_yes:bmi              1 1.7523e+10 4.8719e+10 23328
## - age                         1 1.8059e+10 4.9255e+10 23342
##
## Step:  AIC=22729.67
## charges ~ smoker_yes + age + bmi + children + region_northeast +
##     region_northwest + smoker_yes:bmi + bmi:region_northeast +
##     bmi:region_northwest
##
##                              Df  Sum of Sq        RSS   AIC
## + children:region_northwest   1 6.8538e+07 3.1012e+10 22729
## <none>                                     3.1081e+10 22730
## + age:bmi                     1 3.8530e+07 3.1042e+10 22730
## + region_southeast            1 5.3617e+06 3.1075e+10 22731
## + children:region_northeast   1 3.2338e+06 3.1077e+10 22732
## + smoker_yes:age              1 6.6308e+05 3.1080e+10 22732
## + age:children                1 3.8953e+05 3.1080e+10 22732
## - bmi:region_northwest        1 1.1548e+08 3.1196e+10 22733
## - bmi:region_northeast        1 1.3036e+08 3.1211e+10 22733
## - children                    1 5.0145e+08 3.1582e+10 22749
## - smoker_yes:bmi              1 1.7637e+10 4.8717e+10 23330
## - age                         1 1.8041e+10 4.9122e+10 23341
##
## Step:  AIC=22728.71
## charges ~ smoker_yes + age + bmi + children + region_northeast +
##     region_northwest + smoker_yes:bmi + bmi:region_northeast +
##     bmi:region_northwest + children:region_northwest
##
##                              Df  Sum of Sq        RSS   AIC
## <none>                                     3.1012e+10 22729
## + age:bmi                     1 3.5942e+07 3.0976e+10 22729
## - children:region_northwest   1 6.8538e+07 3.1081e+10 22730
## + children:region_northeast   1 2.0568e+07 3.0991e+10 22730
## + region_southeast            1 4.8465e+06 3.1007e+10 22730
## + smoker_yes:age              1 9.3917e+05 3.1011e+10 22731
## + age:children                1 6.4392e+04 3.1012e+10 22731
## - bmi:region_northwest        1 1.0391e+08 3.1116e+10 22731
## - bmi:region_northeast        1 1.2923e+08 3.1141e+10 22732
## - smoker_yes:bmi              1 1.7644e+10 4.8656e+10 23330
## - age                         1 1.8020e+10 4.9032e+10 23340
```

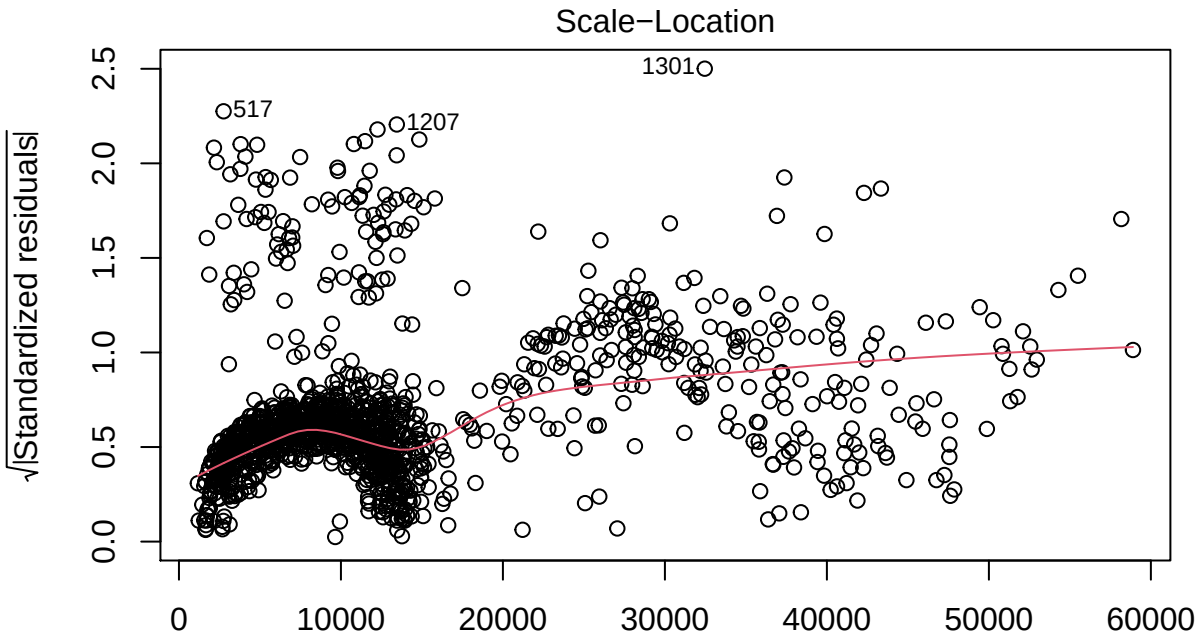
```r
summary(step3)
```

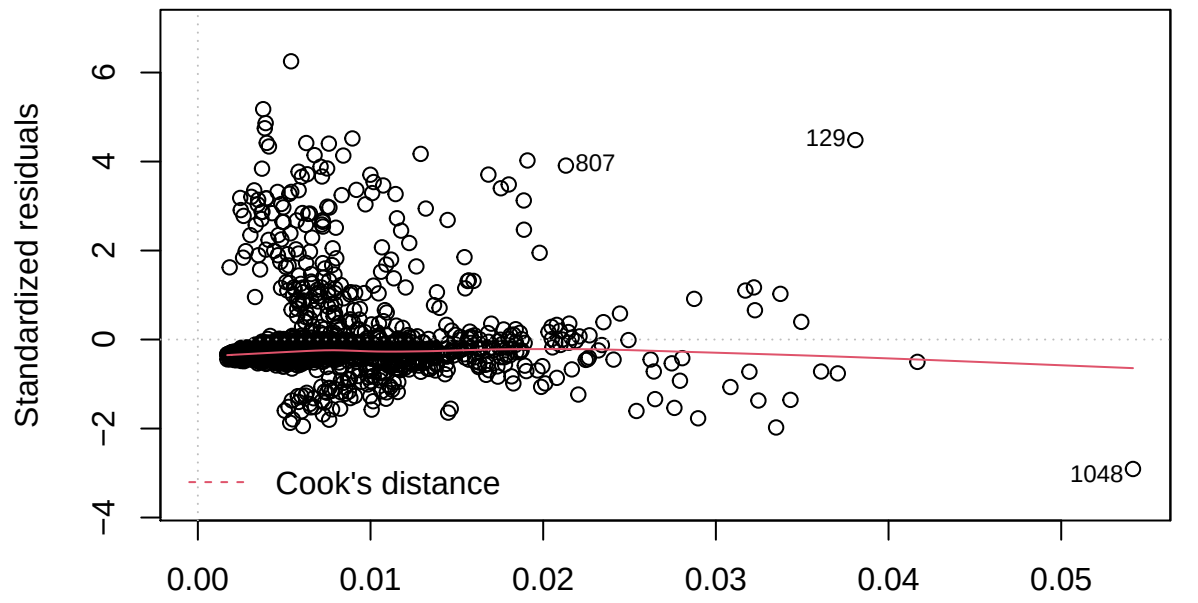
```
##
## Call:
## lm(formula = charges ~ smoker_yes + age + bmi + children + region_northeast +
##     region_northwest + smoker_yes:bmi + bmi:region_northeast +
##     bmi:region_northwest + children:region_northwest, data = insurance_group19)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13672.0  -1905.0  -1264.2   -337.3  30132.6
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                -1728.807   1080.207  -1.600  0.10974
## smoker_yes                -20512.753   1646.599 -12.458  < 2e-16 ***
## age                          263.329      9.479  27.779  < 2e-16 ***
## bmi                          -33.668     31.940  -1.054  0.29203
## children                     406.978    124.957   3.257  0.00115 **
## region_northeast           -2651.578   1655.790  -1.601  0.10953
## region_northwest           -3741.331   1837.213  -2.036  0.04191 *
## smoker_yes:bmi              1445.132     52.574  27.487  < 2e-16 ***
## bmi:region_northeast         127.103     54.030   2.352  0.01880 *
## bmi:region_northwest         127.154     60.280   2.109  0.03510 *
## children:region_northwest    448.347    261.706   1.713  0.08691 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4832 on 1328 degrees of freedom
## Multiple R-squared:  0.8419, Adjusted R-squared:  0.8408
## F-statistic: 707.4 on 10 and 1328 DF,  p-value: < 2.2e-16
```

```r
lmbest1<-lm(formula = charges ~ smoker_yes + age + bmi + children + region_northeast +
    region_northwest + smoker_yes:bmi + bmi:region_northeast +
    bmi:region_northwest + children:region_northwest, data = insurance_group19)
plot(lmbest1)
```

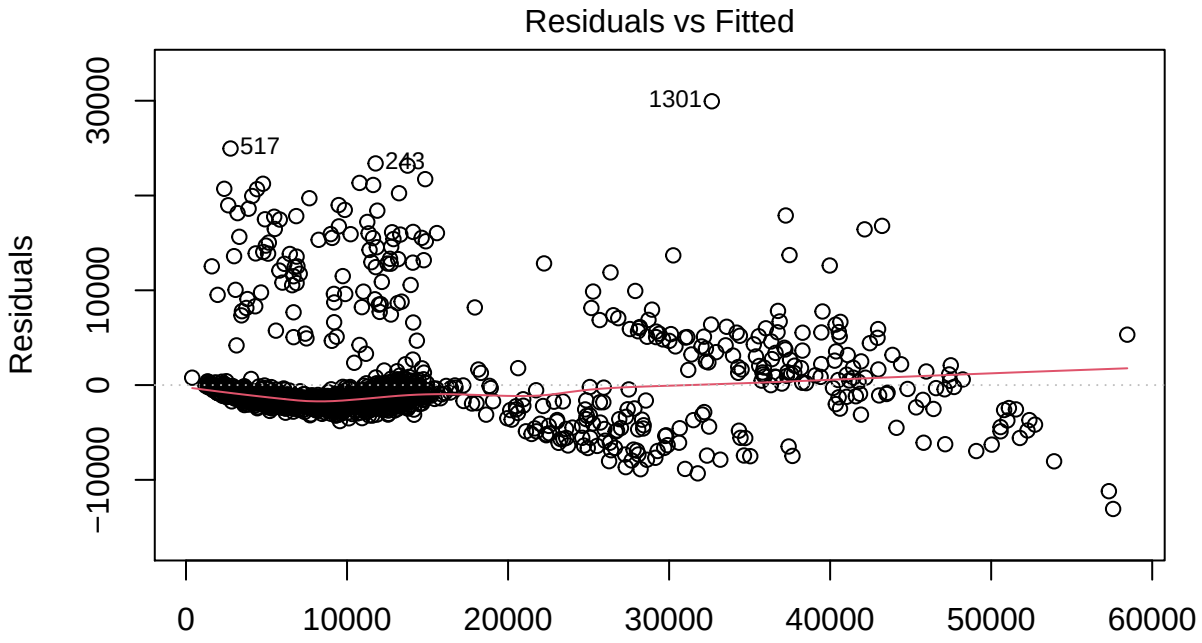

Residuals vs Fitted
Residuals
30000
10000
0
−10000
1301
517
1207
0
10000
20000
30000
40000
50000
60000
Fitted values
lm(charges ~ smoker_yes + age + bmi + children + region_northeast + region_ ...

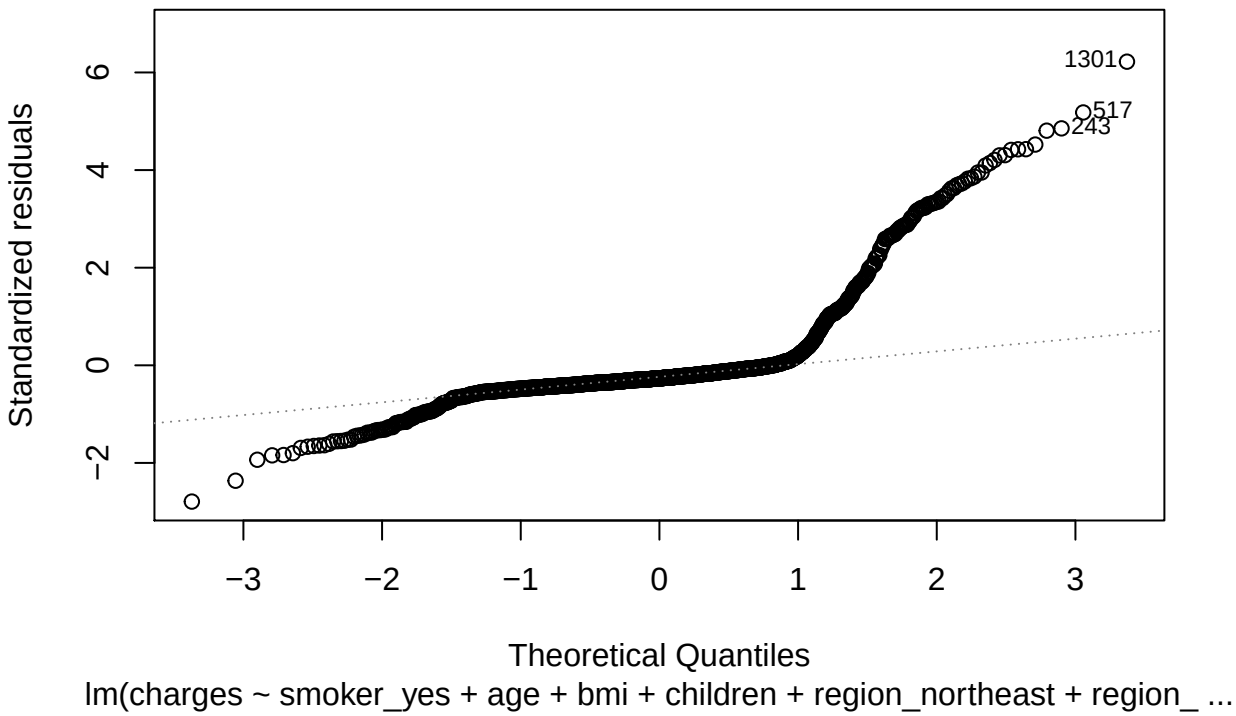

Normal Q−Q
Standardized residuals
6
4
2
0
−2
1301
517
1207
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(charges ~ smoker_yes + age + bmi + children + region_northeast + region_ ...

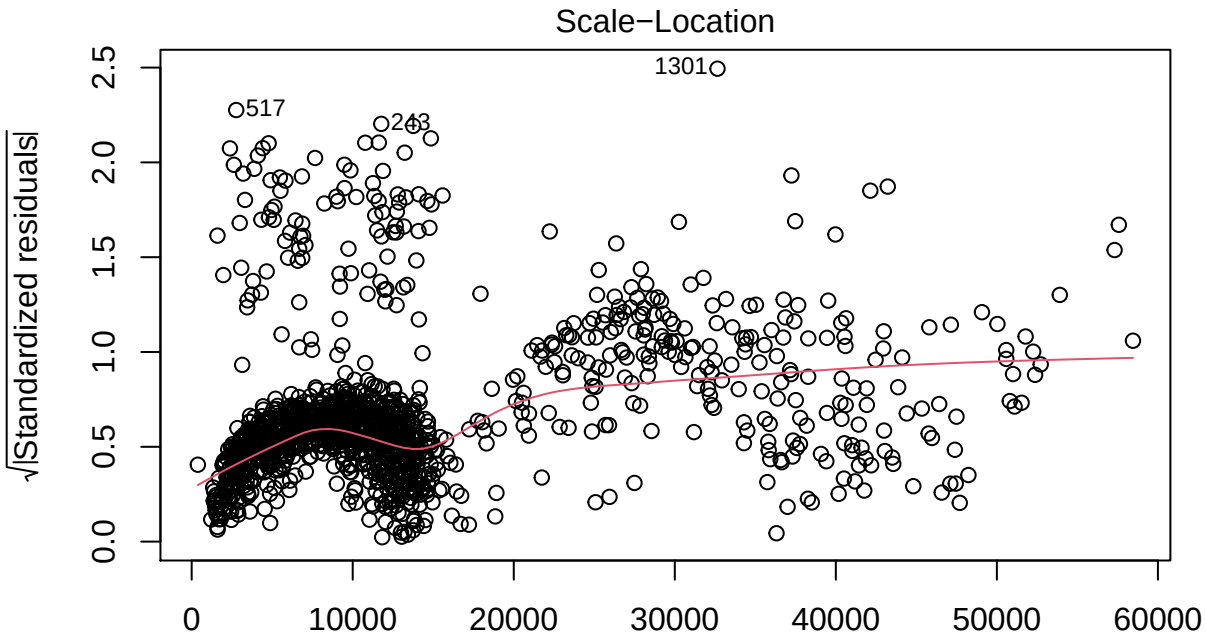


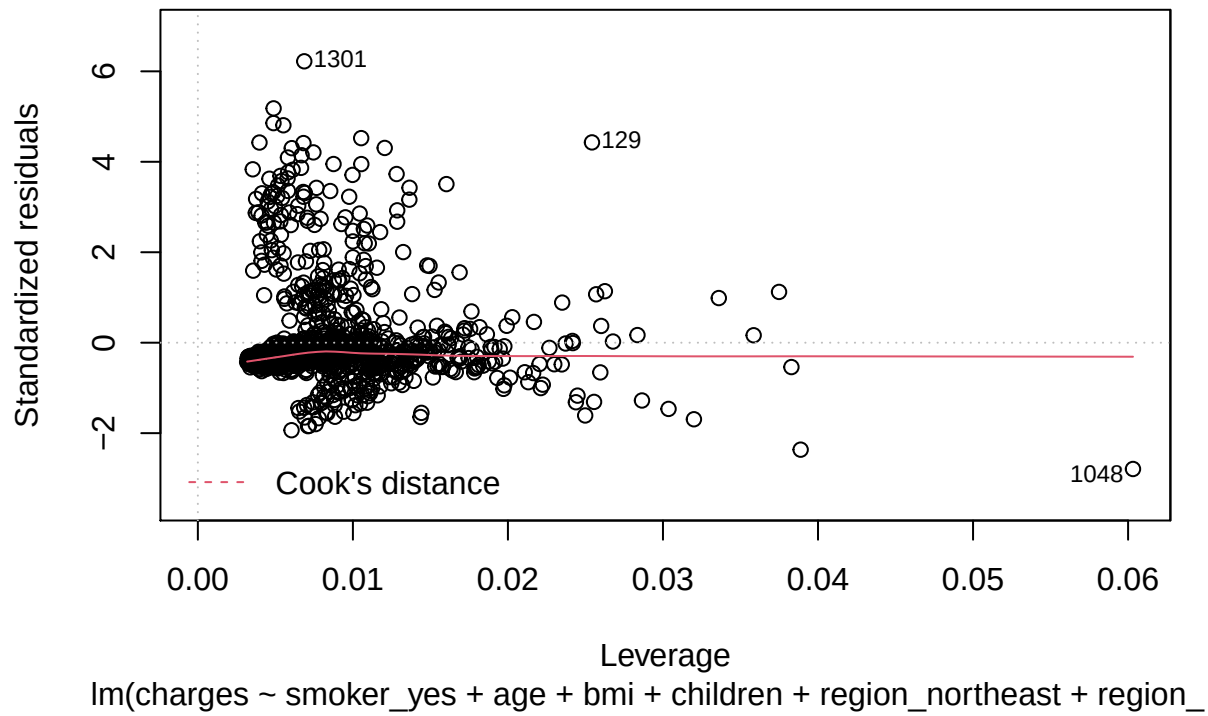



```
lmbest2<-lm(charges ~ smoker_yes + age + bmi + children + region_northeast +
    region_northwest + region_southeast + smoker_yes:bmi + bmi:region_southeast +
    children:region_northwest, data = insurance_group19)
plot(lmbest2)
```


Residuals vs Fitted
Residuals
30000
10000
0
−10000
1301
517
243
0
10000
20000
30000
40000
50000
60000
Fitted values
lm(charges ~ smoker_yes + age + bmi + children + region_northeast + region_ ...


Normal Q−Q
Standardized residuals
6
4
2
0
−2
1301
517
243
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(charges ~ smoker_yes + age + bmi + children + region_northeast + region_ ...





#(d) k-fold validation

```
library(caret)
```

```
## Loading required package: lattice
```

```
train_control <- trainControl(method = "cv",number=10)
model1 <- train(charges~age+bmi+children+smoker_yes+region_northeast+region_northwest,data=insurance_gr
model1
```

```
## Linear Regression
##
## 1339 samples
##    6 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 1206, 1205, 1205, 1204, 1204, 1205, ...
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   6038.765  0.7510381  4184.313
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

```
train_control <- trainControl(method = "cv",number=10)
model2 <- train(charges ~ smoker_yes + age + bmi + children + region_northeast +
    region_northwest + region_southeast + smoker_yes:bmi + bmi:region_southeast +
    children:region_northwest, data=insurance_group19, method = "lm", trControl = train_control)
model2
```

```
## Linear Regression
##
## 1339 samples
##    7 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 1204, 1205, 1204, 1207, 1204, 1206, ...
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   4833.814  0.8380753  2913.155
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

```
train_control <- trainControl(method = "cv",number=10)
model3 <- train(charges ~ smoker_yes + age + bmi + children + region_northeast +
    region_northwest + smoker_yes:bmi + bmi:region_northeast +
    bmi:region_northwest + children:region_northwest, data = insurance_group19, method = "lm", trContro
model3
```

```
## Linear Regression
##
## 1339 samples
##    6 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 1207, 1205, 1205, 1206, 1205, 1205, ...
## Resampling results:
##
```

```
##   RMSE      Rsquared  MAE
##   4837.683  0.836877  2917.629
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

```
train_control <- trainControl(method = "cv",number=10)
model4<-train(charges~1/age+bmi+children+sex_male+smoker_yes+region_northeast+region_northwest+region_s
model4
```

```
## Linear Regression
##
## 1339 samples
##    8 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 1205, 1207, 1204, 1204, 1206, 1203, ...
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   7068.931  0.6578419  5388.254
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

#(e) Anova tests

```
lmbest1<-lm(formula = charges ~ smoker_yes + age + bmi + children + region_northeast +
    region_northwest + smoker_yes:bmi + bmi:region_northeast +
    bmi:region_northwest + children:region_northwest, data = insurance_group19)
lmbest2<-lm(charges ~ smoker_yes + age + bmi + children + region_northeast +
    region_northwest + region_southeast + smoker_yes:bmi + bmi:region_southeast +
    children:region_northwest, data = insurance_group19)
summary(lmbest1)
```

```
##
## Call:
## lm(formula = charges ~ smoker_yes + age + bmi + children + region_northeast +
##     region_northwest + smoker_yes:bmi + bmi:region_northeast +
##     bmi:region_northwest + children:region_northwest, data = insurance_group19)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13672.0  -1905.0  -1264.2   -337.3  30132.6
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                -1728.807   1080.207  -1.600  0.10974
## smoker_yes                 -20512.753   1646.599 -12.458  < 2e-16 ***
## age                           263.329      9.479  27.779  < 2e-16 ***
## bmi                           -33.668     31.940  -1.054  0.29203
## children                      406.978    124.957   3.257  0.00115 **
## region_northeast           -2651.578   1655.790  -1.601  0.10953
## region_northwest           -3741.331   1837.213  -2.036  0.04191 *
## smoker_yes:bmi              1445.132     52.574  27.487  < 2e-16 ***
## bmi:region_northeast         127.103     54.030   2.352  0.01880 *
## bmi:region_northwest         127.154     60.280   2.109  0.03510 *
```

```
## children:region_northwest    448.347    261.706   1.713  0.08691 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4832 on 1328 degrees of freedom
## Multiple R-squared:  0.8419, Adjusted R-squared:  0.8408
## F-statistic: 707.4 on 10 and 1328 DF,  p-value: < 2.2e-16
```

```
summary(lmbest2)
```

```
##
## Call:
## lm(formula = charges ~ smoker_yes + age + bmi + children + region_northeast +
##     region_northwest + region_southeast + smoker_yes:bmi + bmi:region_southeast +
##     children:region_northwest, data = insurance_group19)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13071.0  -1987.1  -1279.5   -291.8  29941.1
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                -4848.301    979.627  -4.949 8.41e-07 ***
## smoker_yes                -20910.607   1651.674 -12.660  < 2e-16 ***
## age                          261.957      9.495  27.589  < 2e-16 ***
## bmi                           67.876     29.626   2.291 0.022112 *
## children                     394.841    124.951   3.160 0.001614 **
## region_northeast            1292.064    381.369   3.388 0.000725 ***
## region_northwest             176.545    484.650   0.364 0.715712
## region_southeast            4879.864   1612.669   3.026 0.002526 **
## smoker_yes:bmi              1457.045     52.722  27.636  < 2e-16 ***
## bmi:region_southeast        -149.996     48.379  -3.100 0.001973 **
## children:region_northwest    471.178    261.001   1.805 0.071258 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4829 on 1328 degrees of freedom
## Multiple R-squared:  0.8421, Adjusted R-squared:  0.841
## F-statistic: 708.5 on 10 and 1328 DF,  p-value: < 2.2e-16
```

```
anova(lmbest1)
```

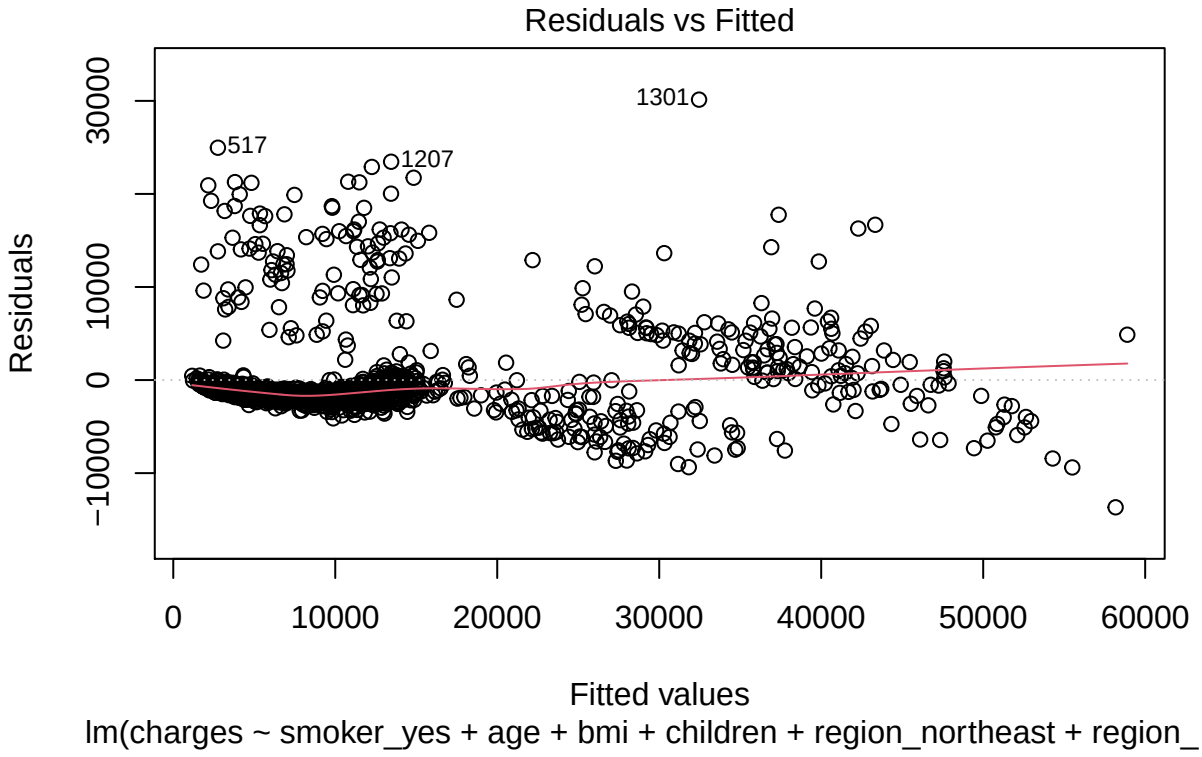
```
## Analysis of Variance Table
##
## Response: charges
##                             Df     Sum Sq    Mean Sq   F value    Pr(>F)
## smoker_yes                   1 1.2161e+11 1.2161e+11 5207.4904 < 2.2e-16 ***
## age                          1 1.9971e+10 1.9971e+10  855.1845 < 2.2e-16 ***
## bmi                          1 5.1137e+09 5.1137e+09  218.9778 < 2.2e-16 ***
## children                     1 4.3549e+08 4.3549e+08   18.6485 1.689e-05 ***
## region_northeast             1 1.4431e+08 1.4431e+08    6.1799   0.01304 *
## region_northwest             1 8.8201e+07 8.8201e+07    3.7770   0.05217 .
## smoker_yes:bmi               1 1.7571e+10 1.7571e+10  752.4399 < 2.2e-16 ***
## bmi:region_northeast         1 7.8842e+07 7.8842e+07    3.3762   0.06637 .
## bmi:region_northwest         1 1.1548e+08 1.1548e+08    4.9453   0.02633 *
```

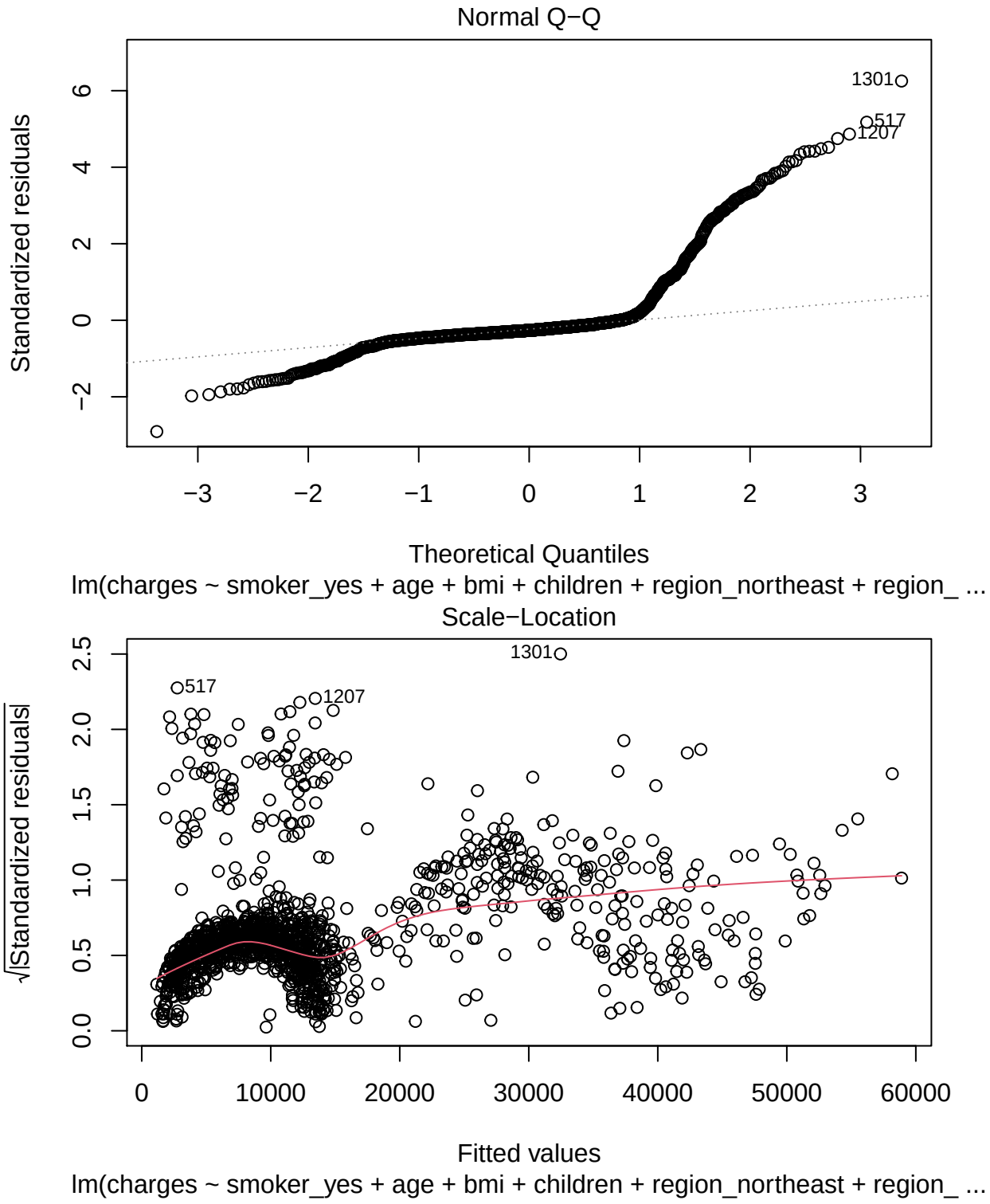
```
## children:region_northwest    1 6.8538e+07 6.8538e+07    2.9349   0.08691 .
## Residuals                 1328 3.1012e+10 2.3352e+07
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

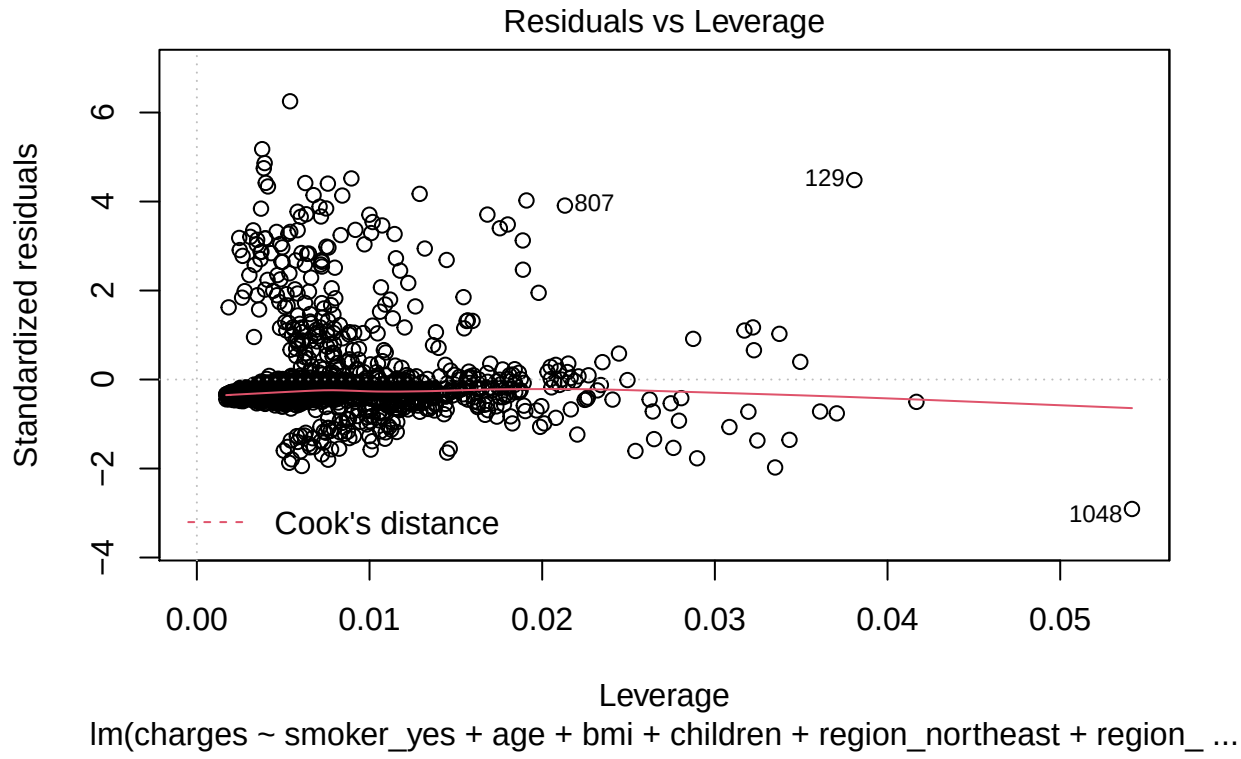
```
anova(lmbest2)
```

```
## Analysis of Variance Table
##
## Response: charges
##                              Df     Sum Sq    Mean Sq   F value    Pr(>F)
## smoker_yes                    1 1.2161e+11 1.2161e+11 5214.1581 < 2.2e-16 ***
## age                           1 1.9971e+10 1.9971e+10  856.2794 < 2.2e-16 ***
## bmi                           1 5.1137e+09 5.1137e+09  219.2582 < 2.2e-16 ***
## children                      1 4.3549e+08 4.3549e+08   18.6724 1.668e-05 ***
## region_northeast              1 1.4431e+08 1.4431e+08    6.1878  0.012986 *
## region_northwest              1 8.8201e+07 8.8201e+07    3.7818  0.052023 .
## region_southeast              1 9.2492e+05 9.2492e+05    0.0397  0.842182
## smoker_yes:bmi                1 1.7570e+10 1.7570e+10  753.3671 < 2.2e-16 ***
## bmi:region_southeast          1 2.2643e+08 2.2643e+08    9.7087  0.001873 **
## children:region_northwest     1 7.6008e+07 7.6008e+07    3.2590  0.071258 .
## Residuals                  1328 3.0972e+10 2.3323e+07
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
plot(lmbest1)
```

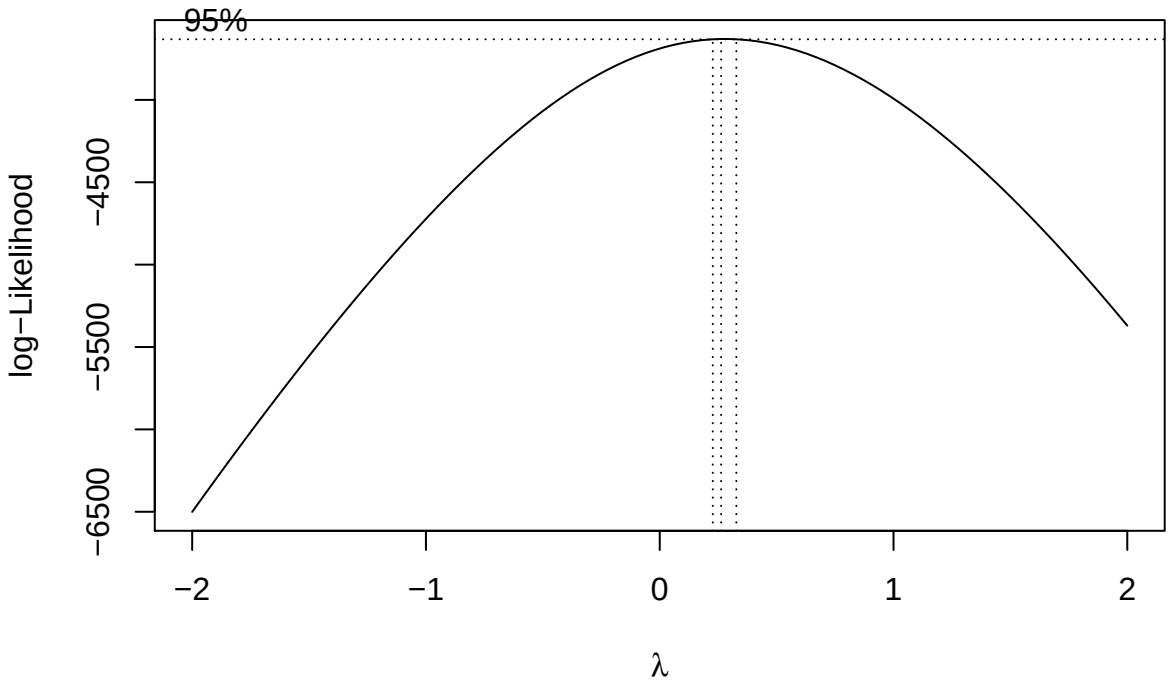



Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(charges ~ smoker_yes + age + bmi + children + region_northeast + region_ ...
1301
517
1207
Scale−Location
√|Standardized residuals|
Fitted values
lm(charges ~ smoker_yes + age + bmi + children + region_northeast + region_ ...



#(f) Model transformation

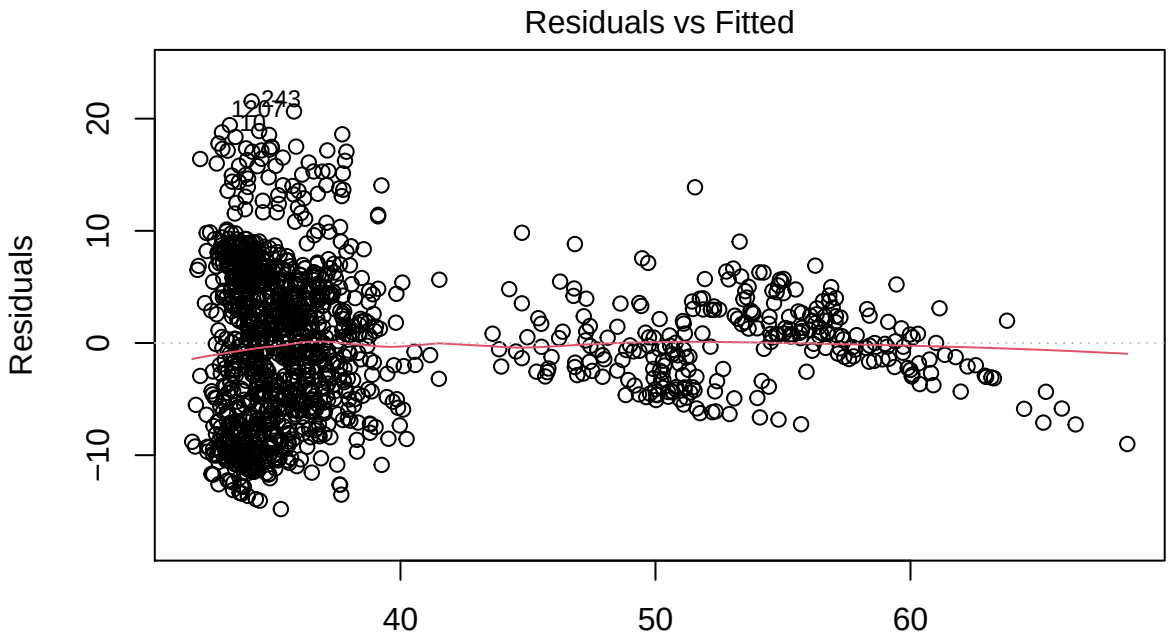
```
library(MASS)
bclm<- boxcox(charges ~ smoker_yes + age + bmi + children + region_northeast +
    region_northwest + smoker_yes:bmi + bmi:region_northeast +
    bmi:region_northwest + children:region_northwest, data = insurance_group19,plotit = T)
```

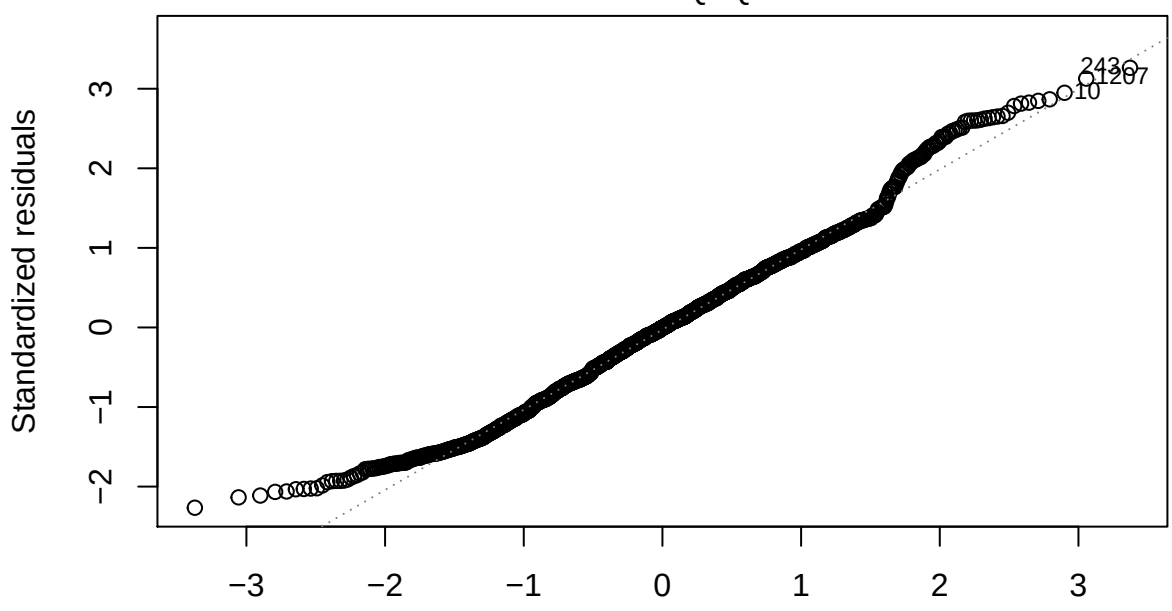


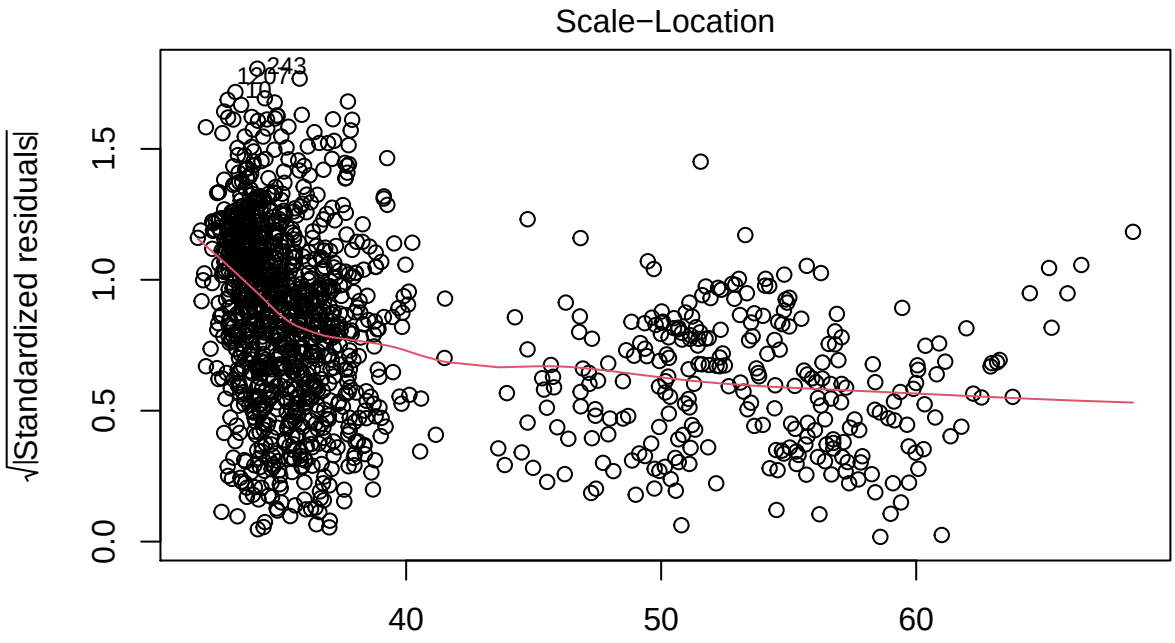

```
bclm$x[which(bclm$y==max(bclm$y))]
```

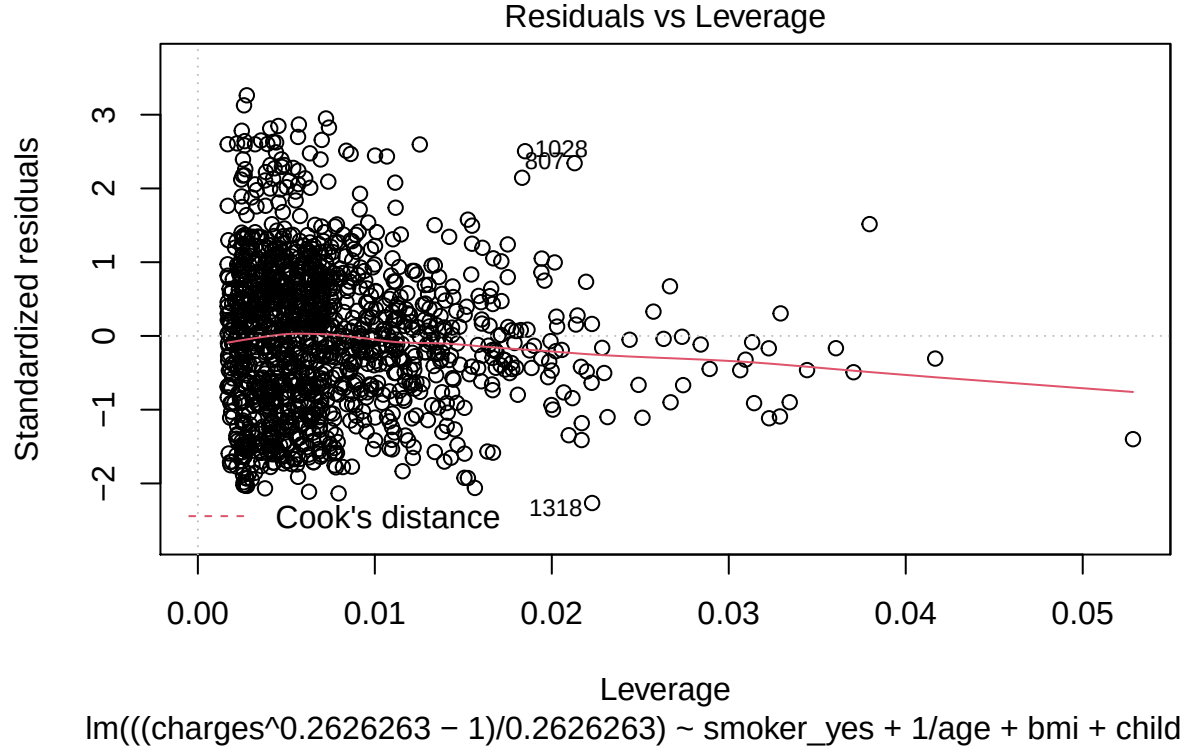
```
## [1] 0.2626263
```

```
lmbest1_bc<- lm(((charges^0.2626263-1)/0.2626263) ~ smoker_yes + 1/age + bmi + children + region_northea
plot(lmbest1_bc)
```










```
summary(lmbest1_bc)
```

```
##
## Call:
## lm(formula = ((charges^0.2626263 - 1)/0.2626263) ~ smoker_yes +
##     1/age + bmi + children + region_northeast + region_northwest +
```

```
##     smoker_yes:bmi + bmi:region_northeast + bmi:region_northwest +
##     children:region_northwest, data = insurance_group19)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.8060  -4.6573  -0.0204   4.2984  21.5357
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                30.96804    1.42272  21.767  < 2e-16 ***
## smoker_yes                 -1.20696    2.25160  -0.536   0.5920
## bmi                         0.08172    0.04356   1.876   0.0609 .
## children                    1.00638    0.17082   5.892 4.84e-09 ***
## region_northeast           -2.94419    2.26417  -1.300   0.1937
## region_northwest           -1.86101    2.51282  -0.741   0.4591
## smoker_yes:bmi              0.63595    0.07188   8.847  < 2e-16 ***
## bmi:region_northeast        0.15688    0.07387   2.124   0.0339 *
## bmi:region_northwest        0.08074    0.08245   0.979   0.3276
## children:region_northwest   0.33351    0.35793   0.932   0.3516
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.61 on 1329 degrees of freedom
## Multiple R-squared:  0.5873, Adjusted R-squared:  0.5845
## F-statistic: 210.1 on 9 and 1329 DF,  p-value: < 2.2e-16
```