Multiple Regression Insurance Data

Er

4/5/2021

```
#Read the dataset:
dataset=read.csv('insurance.csv')

#Data obtained from:
# https://www.kaggle.com/mirichoi0218/insurance
# https://github.com/stedy/Machine-Learning-with-R-datasets

#Exploration of dataset

typeof(dataset) #type
```

```
## [1] "list"
```

```
head(dataset) #look at first rows of dataset
```

```
##   age    sex    bmi children smoker    region   charges
## 1  19 female 27.900        0    yes southwest 16884.924
## 2  18   male 33.770        1     no southeast  1725.552
## 3  28   male 33.000        3     no southeast  4449.462
## 4  33   male 22.705        0     no northwest 21984.471
## 5  32   male 28.880        0     no northwest  3866.855
## 6  31 female 25.740        0     no southeast  3756.622
```

```
ncol(dataset) # Columns of dataset
```

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

```
nrow(dataset) # Rows in dataset
```

```
## [1] 1338
```

```
colnames(dataset) #column names in dataset
```

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

```
summary(dataset) #Summary of the dataset df
```

```
##       age            sex                 bmi           children
##  Min.   :18.00   Length:1338        Min.   :15.96   Min.   :0.000
##  1st Qu.:27.00   Class :character   1st Qu.:26.30   1st Qu.:0.000
##  Median :39.00   Mode  :character   Median :30.40   Median :1.000
##  Mean   :39.21                      Mean   :30.66   Mean   :1.095
##  3rd Qu.:51.00                      3rd Qu.:34.69   3rd Qu.:2.000
##  Max.   :64.00                      Max.   :53.13   Max.   :5.000
##     smoker             region            charges
##  Length:1338        Length:1338        Min.   : 1122
```

```
##  Class :character   Class :character   1st Qu.: 4740
##  Mode  :character   Mode  :character   Median : 9382
##                                        Mean   :13270
##                                        3rd Qu.:16640
##                                        Max.   :63770
```

```
str(dataset) # Description of data frame by type of data
```

```
## 'data.frame':    1338 obs. of  7 variables:
##  $ age     : int  19 18 28 33 32 31 46 37 37 60 ...
##  $ sex     : chr  "female" "male" "male" "male" ...
##  $ bmi     : num  27.9 33.8 33 22.7 28.9 ...
##  $ children: int  0 1 3 0 0 0 1 3 2 0 ...
##  $ smoker  : chr  "yes" "no" "no" "no" ...
##  $ region  : chr  "southwest" "southeast" "southeast" "northwest" ...
##  $ charges : num  16885 1726 4449 21984 3867 ...
```

Description of Data:

Columns

- age: age of primary beneficiary
- sex: insurance contractor gender, female, male
- bmi: Body mass index, providing an understanding of body, weights that are relatively high or low relative to height, objective index of body weight (kg / m ^ 2) using the ratio of height to weight, ideally 18.5 to 24.9 *c hildren: Number of children covered by health insurance / Number of dependents
- smoker: Smoking
- region: the beneficiary's residential area in the US, northeast, southeast, southwest, northwest.
- charges: Individual medical costs billed by health insurance. This is the TARGET

Visualization of data

```
#Summarizing categorical variables
table(dataset$sex)
```

```
##
## female   male
##    662    676
```

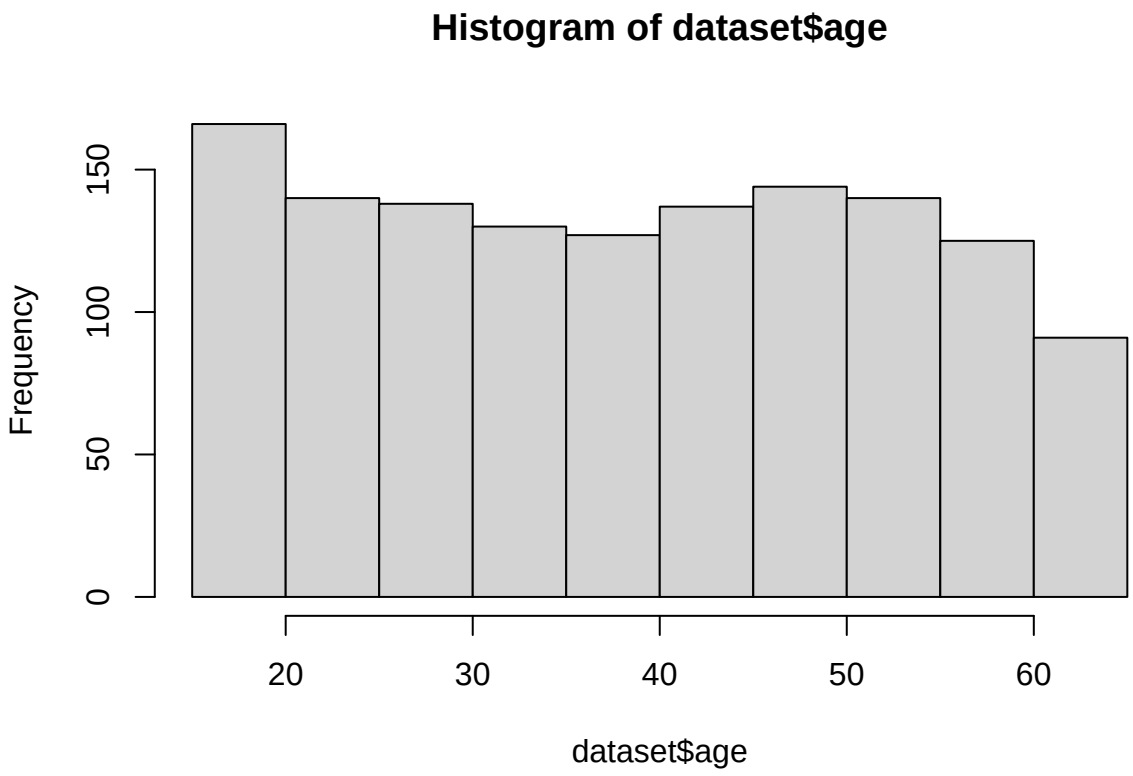
```
table(dataset$smoker)
```

```
##
##   no  yes
## 1064  274
```

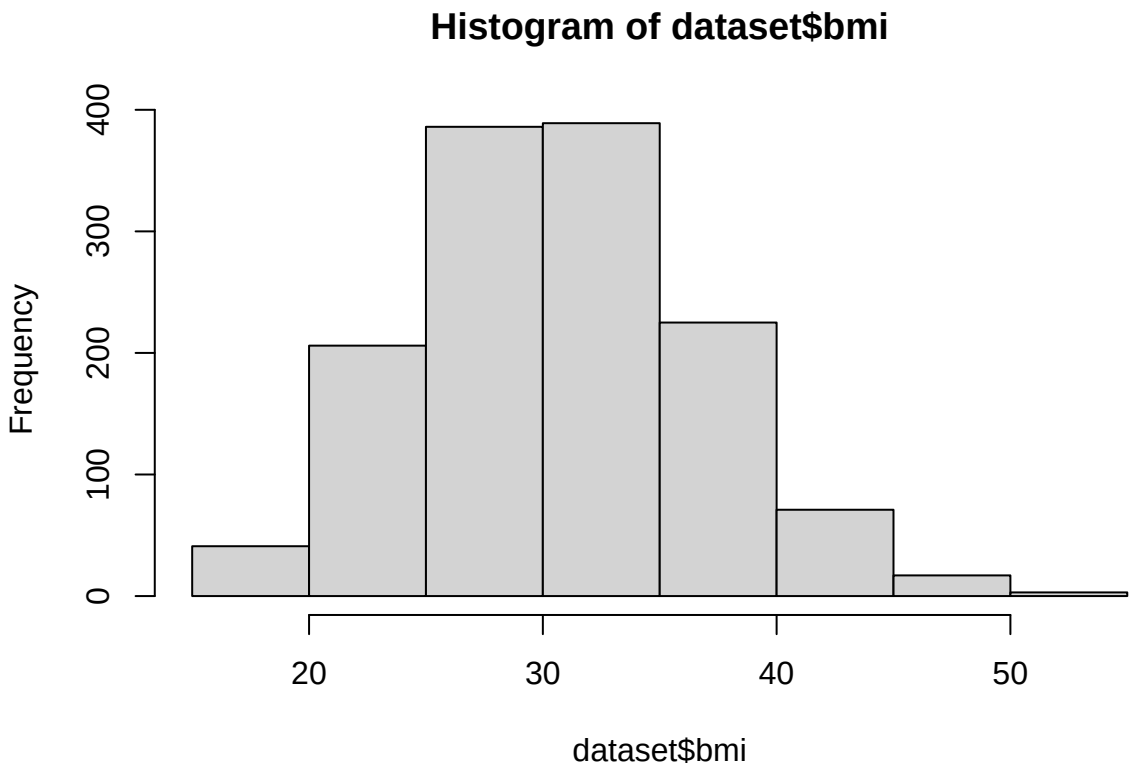
```
table(dataset$region)
```

```
##
## northeast northwest southeast southwest
##       324       325       364       325
```

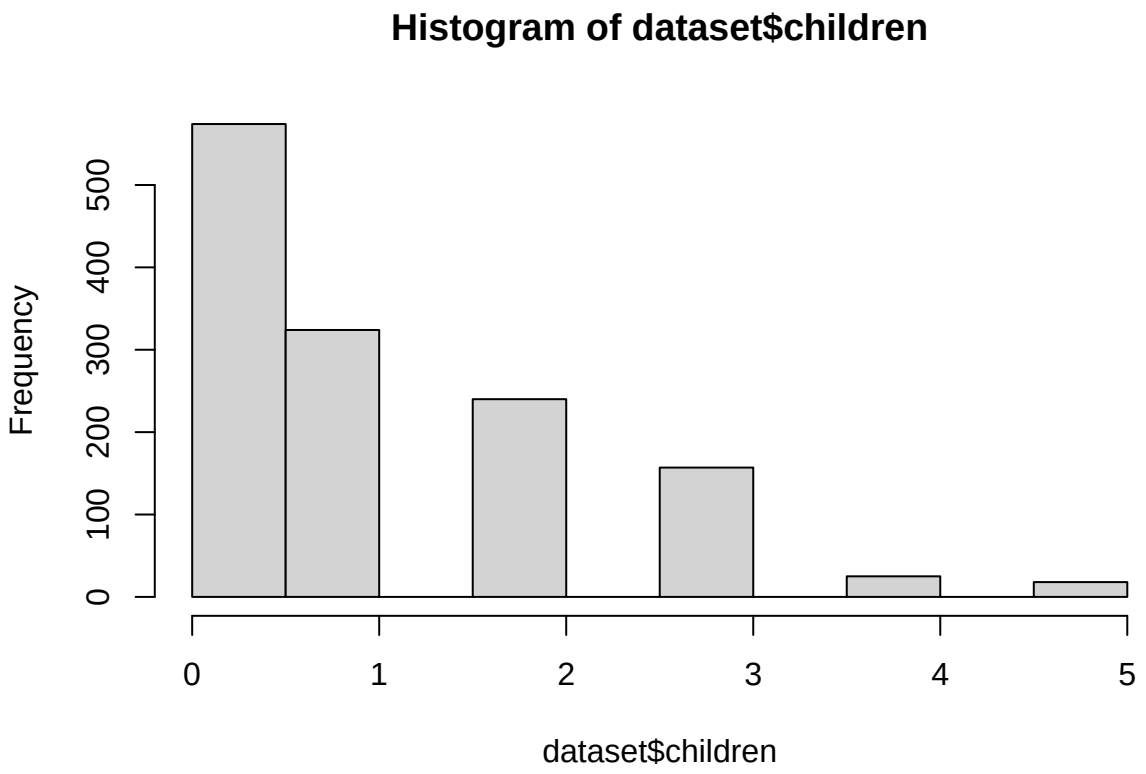
```
#Summary of integers

hist(dataset$age)
```




```
hist(dataset$bmi)
```



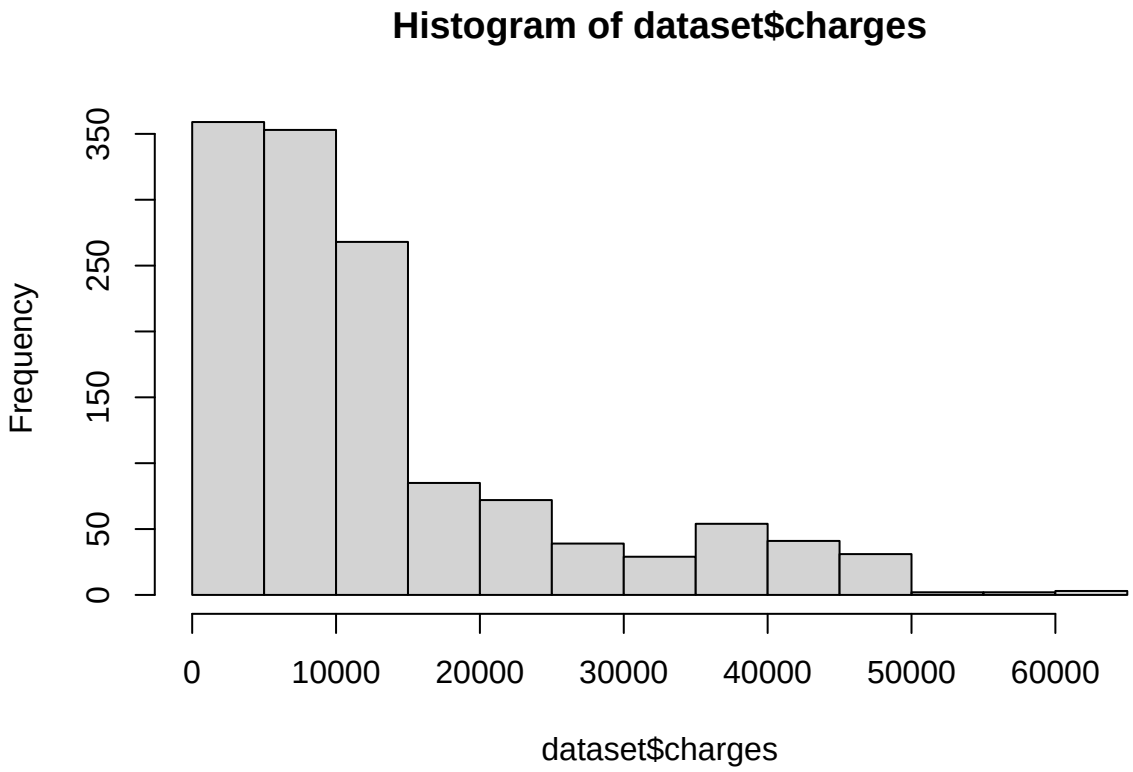

```
hist(dataset$children)
```



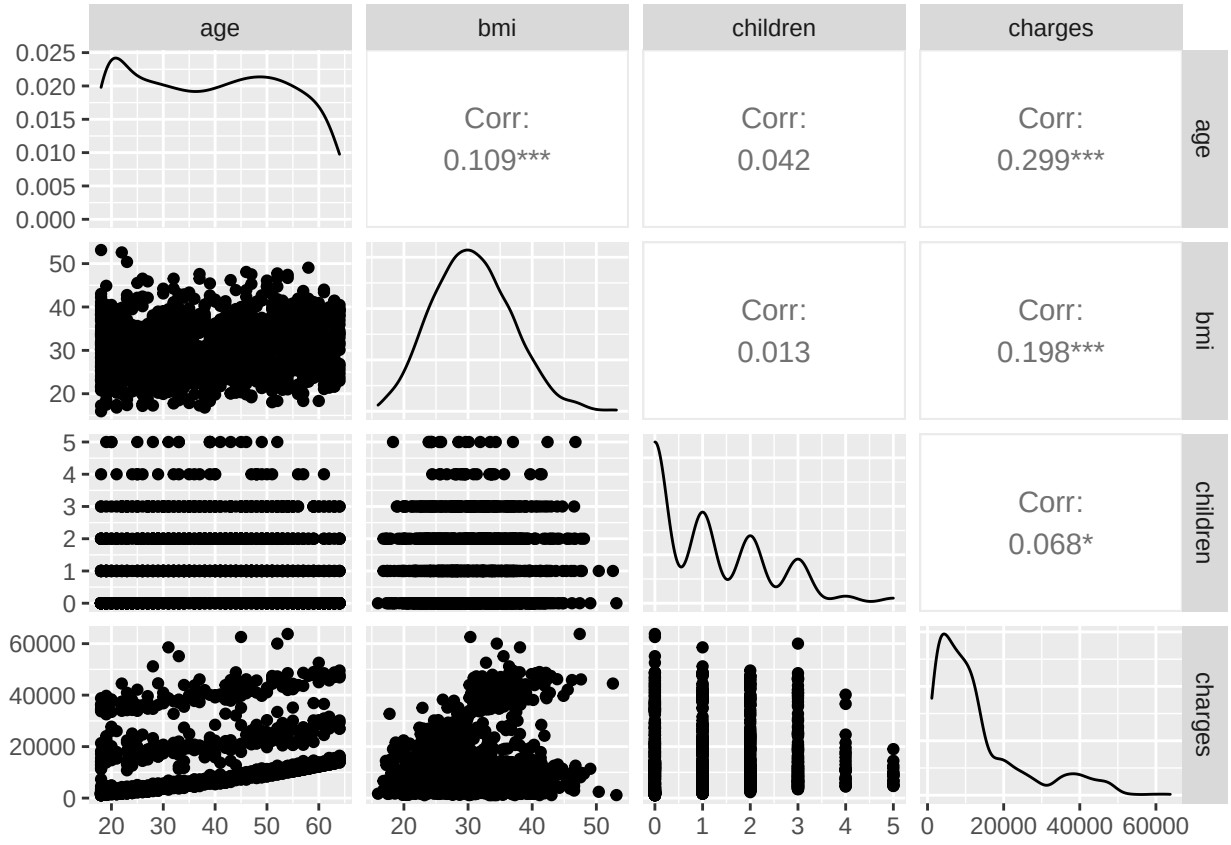

```
hist(dataset$charges)

#Scatter Plot Matrix of Numeric Variables
library(GGally)
```

```
## Loading required package: ggplot2
```

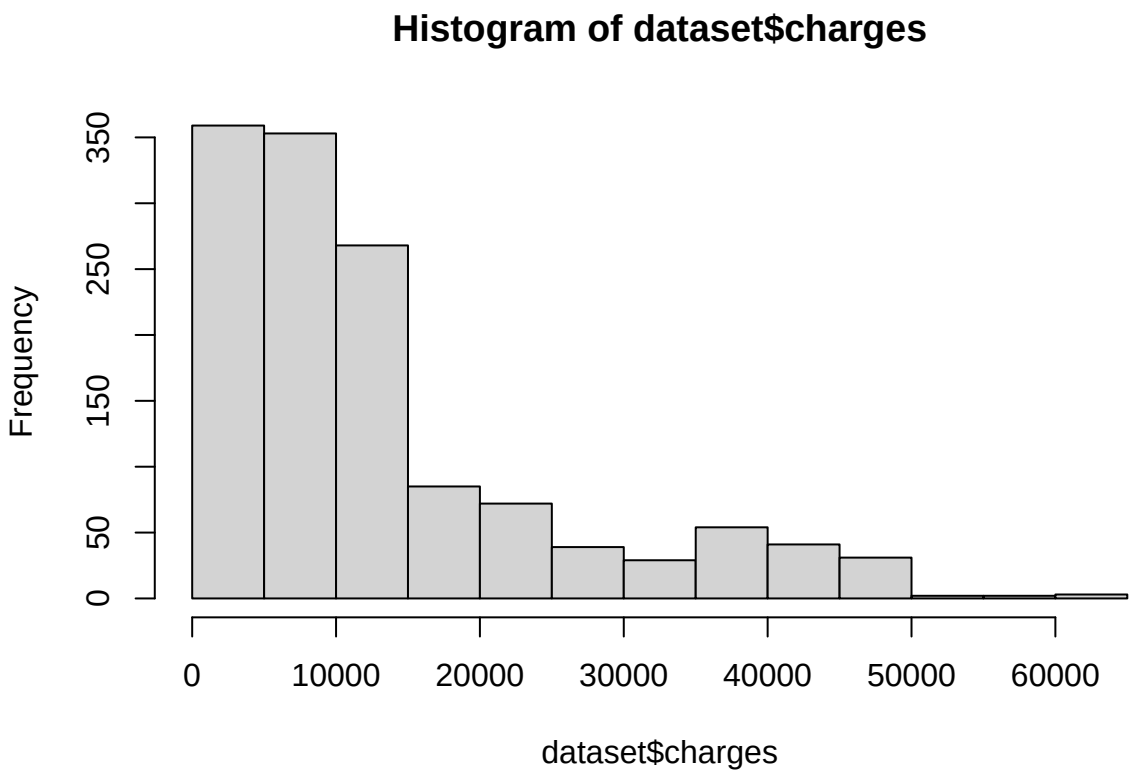
```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```




```
ggpairs(dataset[,c("age","bmi","children","charges")])
```



```r
hist(dataset$charges) #Histogram of Y (Target)
```



Encoding categorical variables as factors

```r
# Encoding categorical data using 'factor" function
dataset$sex=factor(dataset$sex,
                   levels=c('female', 'male'),
                   labels=c(1,0)) #Note: a factor in R is NOT a numeric value
dataset$smoker=factor(dataset$smoker,
                  levels=c('yes', 'no'),
                  labels=c(1,0))
dataset$region=factor(dataset$region,
                  levels=c('northeast', 'northwest','southeast','southwest'),
                  labels=c(1,2,3,4))
```

Splitting data into training and test sets

```r
# Splitting data into Training and Test sets (install.packages('caTools')
library(caTools)
set.seed(44) #Setting the seed for random split
split=sample.split(dataset$charges, SplitRatio=0.8) #sample.split {caTools}.
#cont..Also, split uses Split Ratio as fraction on TRAINING set
training_set=subset(dataset, split==TRUE)
test_set=subset(dataset, split==FALSE)
```

Fitting model

```r
#Fitting model into Training dataset
# Note: the regression model takes care of the Dummy trap
```

```r
# by eliminating one of the dummy columns in each case
library(car)
```

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

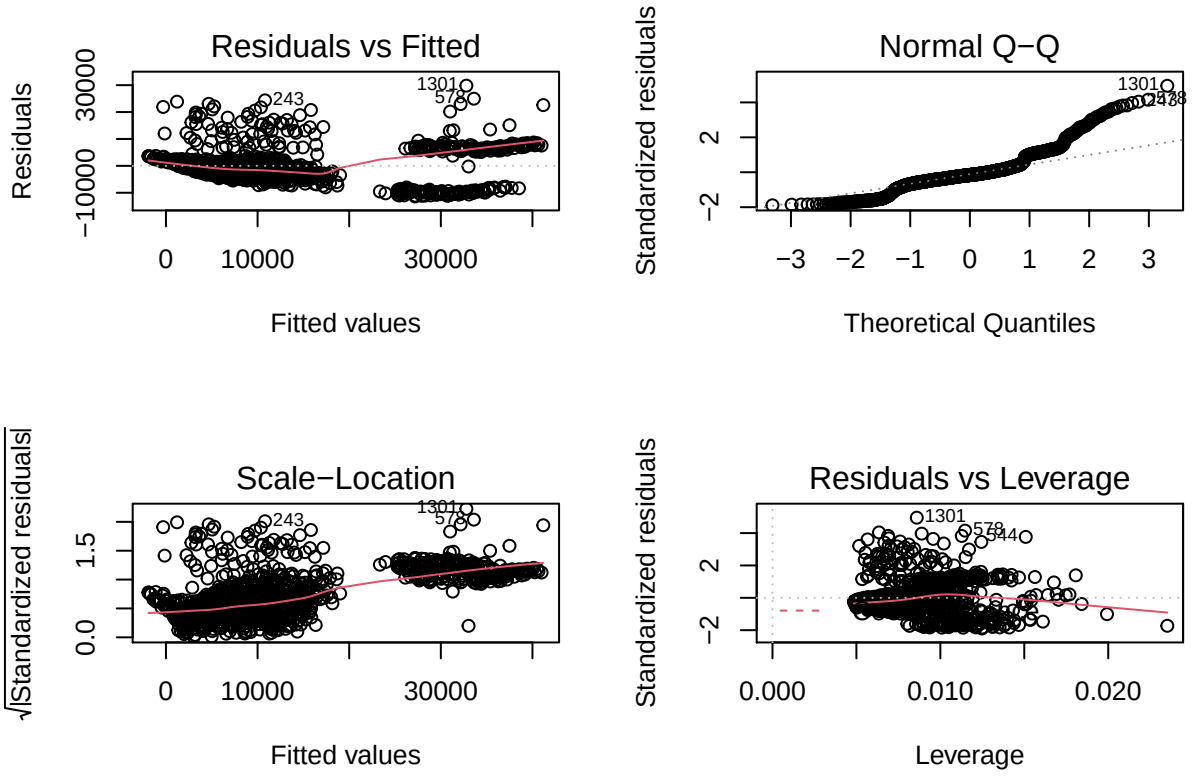
```r
regressor<-lm(formula=charges~., data<-training_set)
summary(regressor)
```

```
##
## Call:
## lm(formula = charges ~ ., data = data <- training_set)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11368.8  -2876.0   -932.7   1588.0  29805.2
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  12026.48    1170.17  10.278  < 2e-16 ***
## age            254.62      13.24  19.236  < 2e-16 ***
## sex0          -162.20     371.10  -0.437  0.66214
## bmi            348.42      32.03  10.877  < 2e-16 ***
## children       416.22     153.59   2.710  0.00684 **
## smoker0     -23898.46     464.05 -51.500  < 2e-16 ***
## region2       -645.34     530.01  -1.218  0.22365
## region3      -1112.70     533.34  -2.086  0.03719 *
## region4      -1075.01     526.53  -2.042  0.04143 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6044 on 1061 degrees of freedom
## Multiple R-squared:  0.7526, Adjusted R-squared:  0.7507
## F-statistic: 403.5 on 8 and 1061 DF,  p-value: < 2.2e-16
```

```r
vif(regressor)
```

```
##              GVIF Df GVIF^(1/(2*Df))
## age      1.014296  1        1.007123
## sex      1.008325  1        1.004154
## bmi      1.095830  1        1.046819
## children 1.003973  1        1.001985
## smoker   1.016200  1        1.008067
## region   1.093526  3        1.015013
```

```r
par(mfrow=c(2, 2))
plot(regressor)
```




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

durbinWatsonTest(regressor)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1     -0.07452378      2.145518    0.02
##  Alternative hypothesis: rho != 0
```

Selecting "best" model using best subset and stepwise regressiom

```
# Running stepwise and best subset methods
#install.packages("olsrr")
library(olsrr)
```

```
##
## Attaching package: 'olsrr'
```

```
## The following object is masked from 'package:datasets':
##
##     rivers
```

```
ols_step_both_p(regressor, pent = 0.1, prem = 0.1, details = FALSE)
```

```
##
##                                 Stepwise Selection Summary
## ----------------------------------------------------------------------------------------------
##                      Added/                   Adj.
## Step    Variable     Removed     R-Square    R-Square      C(p)         AIC          RMSE
## ----------------------------------------------------------------------------------------------
##    1      smoker     addition       0.622       0.622    553.0480    22118.3761    7441.9109
##    2       age       addition       0.722       0.721    129.3510    21793.9481    6392.0793
```

```
##    3      bmi       addition       0.750       0.749     12.0900    21683.2863    6067.1120
##    4    children    addition       0.751       0.750      6.8580    21678.0571    6049.4893
## ----------------------------------------------------------------------------------------
```

```
ols_step_both_aic(regressor, details = FALSE)
```

```
##
##
##                                        Stepwise Summary
## ------------------------------------------------------------------------------------------------
## Variable     Method        AIC             RSS                Sum Sq          R-Sq      Adj. R-Sq
## ------------------------------------------------------------------------------------------------
## smoker      addition    22118.376    59148017026.178     97527283216.230    0.62248       0.62213
## age         addition    21793.948    43596208954.615    113079091287.793    0.72174       0.72122
## bmi         addition    21683.286    39239297726.644         1.17436e+11    0.74955       0.74885
## children    addition    21678.057    38975081871.394    117700218371.013    0.75124       0.75030
## ------------------------------------------------------------------------------------------------
```

```
ols_step_all_possible(regressor) #Best subsets method
```

```
##    Index N                          Predictors    R-Square Adj. R-Square
## 5      1 1                              smoker 0.622480270   0.622126787
## 1      2 1                                 age 0.089260200   0.088407448
## 3      3 1                                 bmi 0.047054104   0.046161833
## 6      4 1                              region 0.010633635   0.007849302
## 2      5 1                                 sex 0.003329168   0.002395955
## 4      6 1                            children 0.002332608   0.001398462
## 10     7 2                          age smoker 0.721741660   0.721220089
## 17     8 2                          bmi smoker 0.661748377   0.661114354
## 19     9 2                     children smoker 0.625946596   0.625245465
## 21    10 2                       smoker region 0.623535299   0.622121347
## 14    11 2                          sex smoker 0.622578661   0.621871218
## 8     12 2                             age bmi 0.124396292   0.122755048
## 11    13 2                          age region 0.100291650   0.096912464
## 7     14 2                             age sex 0.093154579   0.091454775
## 9     15 2                        age children 0.090461341   0.088756489
## 18    16 2                          bmi region 0.051188841   0.047625231
## 12    17 2                             sex bmi 0.050386248   0.048606278
## 16    18 2                        bmi children 0.048887057   0.047104277
## 15    19 2                          sex region 0.013816923   0.010112949
## 20    20 2                     children region 0.013169678   0.009463273
## 13    21 2                        sex children 0.005536851   0.003672815
## 27    22 3                       age bmi smoker 0.749550199   0.748845369
## 29    23 3                 age children smoker 0.723729257   0.722951760
## 31    24 3                   age smoker region 0.722915541   0.721613452
## 24    25 3                      age sex smoker 0.721771944   0.720988939
## 38    26 3                 bmi children smoker 0.664643420   0.663699639
## 40    27 3                   bmi smoker region 0.663309749   0.661727558
## 33    28 3                      sex bmi smoker 0.661838423   0.660886749
## 41    29 3               children smoker region 0.627079158   0.625326711
## 35    30 3                 sex children smoker 0.626075326   0.625023005
## 37    31 3                   sex smoker region 0.623637775   0.621869156
## 28    32 3                      age bmi region 0.129469419   0.125378580
## 22    33 3                         age sex bmi 0.128255856   0.125802542
## 26    34 3                    age bmi children 0.125342680   0.122881167
```

```
## 25    35 3                    age sex region 0.104025194   0.099814786
## 30    36 3                age children region 0.101644610   0.097423015
## 23    37 3                   age sex children 0.094256072   0.091707074
## 34    38 3                     sex bmi region 0.054412126   0.049968574
## 39    39 3                bmi children region 0.053170853   0.048721468
## 32    40 3                   sex bmi children 0.052108438   0.049440826
## 36    41 3                sex children region 0.016223700   0.011600691
## 48    42 4            age bmi children smoker 0.751236591   0.750302268
## 50    43 4              age bmi smoker region 0.750858218   0.749451961
## 43    44 4                 age sex bmi smoker 0.749579145   0.748638597
## 51    45 4         age children smoker region 0.724968845   0.723416458
## 45    46 4            age sex children smoker 0.723772985   0.722735513
## 47    47 4              age sex smoker region 0.722948306   0.721384515
## 56    48 4         bmi children smoker region 0.666198313   0.664314202
## 52    49 4            sex bmi children smoker 0.664759987   0.663500870
## 54    50 4              sex bmi smoker region 0.663404416   0.661504535
## 55    51 4         sex children smoker region 0.627213185   0.625109026
## 44    52 4                 age sex bmi region 0.133205291   0.128312753
## 49    53 4            age bmi children region 0.130533052   0.125625430
## 42    54 4               age sex bmi children 0.129114452   0.125843520
## 46    55 4            age sex children region 0.105273063   0.100222864
## 53    56 4            sex bmi children region 0.056278914   0.050952172
## 61    57 5     age bmi children smoker region 0.752557929   0.750926955
## 57    58 5        age sex bmi children smoker 0.751277591   0.750108782
## 59    59 5          age sex bmi smoker region 0.750890023   0.749248055
## 60    60 5     age sex children smoker region 0.725015908   0.723203395
## 62    61 5     sex bmi children smoker region 0.666320330   0.664120935
## 58    62 5        age sex bmi children region 0.134175845   0.128468906
## 63    63 6 age sex bmi children smoker region 0.752602475   0.750737083
##    Mallow's Cp
## 5   553.047859
## 1  2839.839101
## 3  3020.846036
## 6  3177.040475
## 2  3208.366740
## 4  3212.640631
## 10  129.351059
## 17  386.640912
## 19  540.182020
## 21  550.523217
## 14  554.625893
## 8  2691.152901
## 11 2794.529135
## 7  2825.137495
## 9  2836.687835
## 18 3005.113621
## 12 3008.555658
## 16 3014.985155
## 15 3165.388486
## 20 3168.164288
## 13 3200.898771
## 27   12.090124
## 29  122.826962
## 31  126.316703
```

```
## 24  131.221180
## 38  376.225106
## 40  381.944744
## 33  388.254739
## 41  537.324863
## 35  541.629942
## 37  552.083731
## 28 2671.396062
## 22 2676.600604
## 26 2689.094180
## 25 2780.517296
## 30 2790.726772
## 23 2822.413584
## 34 2993.290099
## 39 2998.613478
## 32 3003.169795
## 36 3157.066678
## 48    6.857792
## 50    8.480498
## 43   13.965987
## 51  119.510813
## 45  124.639429
## 47  128.176182
## 56  371.556720
## 52  377.725189
## 54  383.538753
## 55  538.750069
## 44 2657.374236
## 49 2668.834520
## 42 2674.918392
## 46 2777.165629
## 53 2987.284108
## 61    3.191040
## 57    8.681957
## 59   10.344096
## 60  121.308975
## 62  373.033432
## 58 2655.211876
## 63    5.000000
```

Running "best" model

```
#Running best model after stepwise and Best Subsety:
regressor<-lm(formula=charges~age+bmi+children+smoker, data<-training_set)
summary(regressor)
```

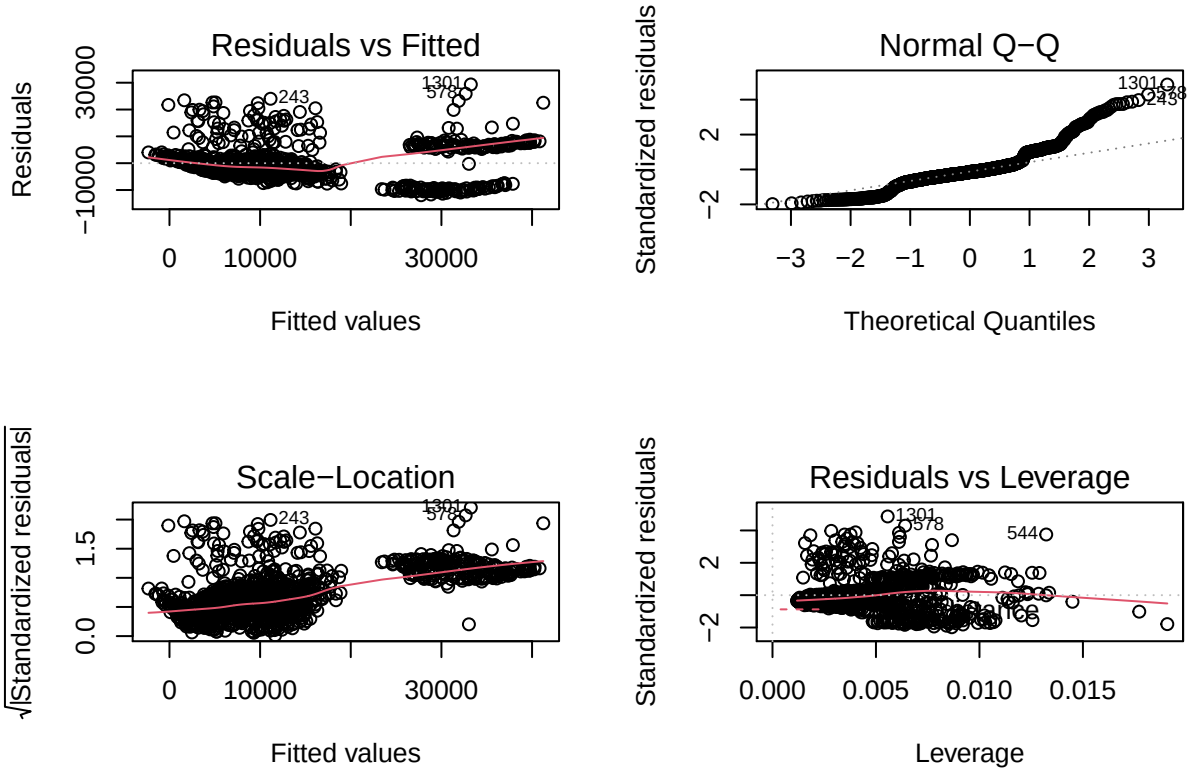
```
##
## Call:
## lm(formula = charges ~ age + bmi + children + smoker, data = data <- training_set)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11942.9  -2913.8   -893.5   1467.6  29347.7
##
```

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  11625.33    1114.77  10.428  < 2e-16 ***
## age            254.91      13.24  19.254  < 2e-16 ***
## bmi            334.29      30.80  10.852  < 2e-16 ***
## children       412.78     153.62   2.687  0.00732 **
## smoker0     -23865.10     461.03 -51.764  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6049 on 1065 degrees of freedom
## Multiple R-squared:  0.7512, Adjusted R-squared:  0.7503
## F-statistic:   804 on 4 and 1065 DF,  p-value: < 2.2e-16
```

```
vif(regressor)
```

```
##      age      bmi children   smoker
## 1.012921 1.011637 1.002704 1.001274
```

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




```
par(mfrow=c(1, 1))
durbinWatsonTest(regressor)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1     -0.07265679      2.141735   0.022
##  Alternative hypothesis: rho != 0
```

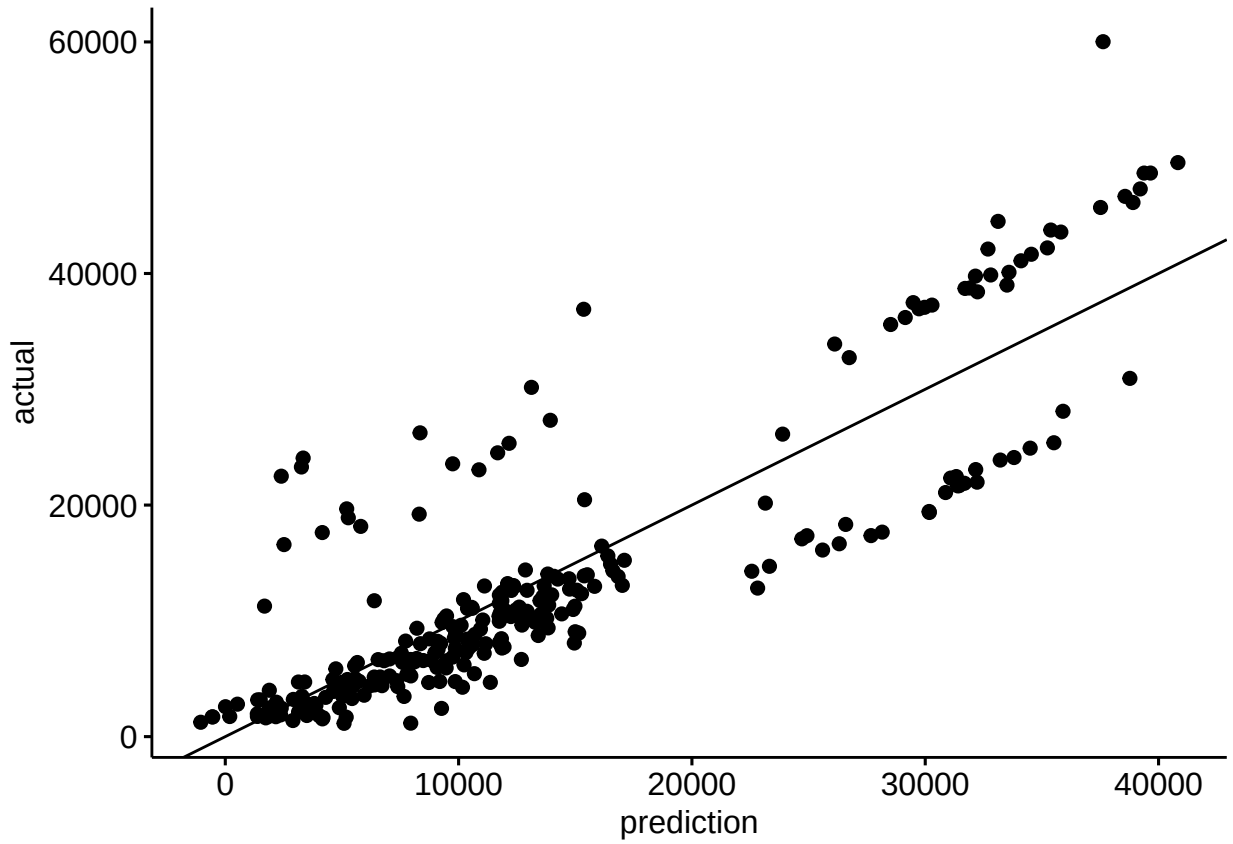
Running test data

```
regressor_test<-lm(formula=charges~age+bmi+children+smoker, data<-test_set)
summary(regressor_test)
```

```
##
## Call:
## lm(formula = charges ~ age + bmi + children + smoker, data = data <- test_set)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -10814  -2924  -1146   1116  22399
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  11574.88    2212.26   5.232 3.43e-07 ***
## age            270.86      27.24   9.945  < 2e-16 ***
## bmi            282.37      60.78   4.646 5.36e-06 ***
## children       741.76     316.23   2.346   0.0197 *
## smoker0     -23504.56     921.46 -25.508  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6168 on 263 degrees of freedom
## Multiple R-squared:  0.7457, Adjusted R-squared:  0.7419
## F-statistic: 192.8 on 4 and 263 DF,  p-value: < 2.2e-16
```

Plot of actual vs forecasted

```
library(ggpubr)
ggscatter(x = "prediction",
          y = "actual",
          data = data.frame(prediction = predict(regressor_test),
                            actual = test_set$charges)) +
  geom_abline(intercept = 0,
              slope = 1)
```




```
library(Metrics)
rmse(predict(regressor_test),test_set$charges)
```

```
## [1] 6110.588
```

```
y_hat<-predict(regressor_test, interval="prediction")
```

```
## Warning in predict.lm(regressor_test, interval = "prediction"): predictions on current data refer to
```

```
test_set<-cbind(y_hat, test_set)
```