2a1_estes

Andrew Estes

1/9/2022

1) Create a scatter plot and build a simple linear regression model using expenses as the dependent variable and bmi as the independent variable. Then answer the following questions:

A) Does there appear to be a linear relationship between BMI and expenses? What is its direction?

Yes, there appears to be a slight positive linear relationship

B) Is the relationship statistically significant?

Yes, the P-Value is less than .05 and the F-Statistic is above 50

C) Are the regression assumptions satisfied for this simple linear regression?

Linearity is not met. The Residuals vs Fitted graph shows an asymettrical spread above and below the zero line. Homoskedacity is not met. Heteroskedacity is clearly present in the Residuals vs Fitted grpah. Normality is not met. Expenses skew right, while BMI is normally distributed. The Q-Q Plot also indicates a lack of normality. Independence is not met. Cook's Distance for many observ ations is greater than 1.

D) What percentage of variation in expenses is explained by BMI?

BMI explains ~ 3.9% of the variation in expenses.

2) Build a multiple linear regression model using the backward elimination method. Continue to use expenses as the dependent variable.

A) Looking at the plot of residuals vs. fit for your stepwise model, what have you learned about the structure of the data?

There appears to be a grouping of data.

B) Are the regression assumptions satisfied?

Linearity is partially met. There is a somewhat symetrical breakdown of data in the Residuals vs Fitted graph. Homoskedacity is not met. There is a clear grouping in the Scale-Location graph. Normality is not met. The Q-Q Plot points are far from the projected line. Independence is not met. Cook's Distance for many observ ations is greater than 1.

C) Do there appear to be multiple groups in the data set?

Yes, there appears to be 2-3 groups.

3) There is at least one variable that's crucially important in making sense of the data.

A) What is it?

Smoking is the crucial variable. Smoking by itself is responsible for .6195 of the adjusted r-squared. The model with every single independent variable has an adjusted r-squared of .7494. Smoking explains 83% of the maximum (likely over-fitted) variance that the full-model provides.

B) Include at least one graph to illustrate the importance of this variable.

See bottom

C) Using this crucial variable, break the data set into two separate data sets, run a stepwise regression on each data set, and comment on which variables seem important in each case.

The only two independent variables that matter to smokers are BMI and Age. For non-smokers, the significant predictors include Age, Gender, Children, and Region. For non-smokers, the adjusted r-squared is a relatively paltry .4137 with a F-statistic of 126. And the Residual Standard Error is 4589 with 1057 degrees of freedom.

4) Extra

Males account for 58% of all respondents who smoke despite only being 50.5% of respondents. It would be interesting

```
library(tidyverse)
library(MASS)
library(leaps)
library(car)
```

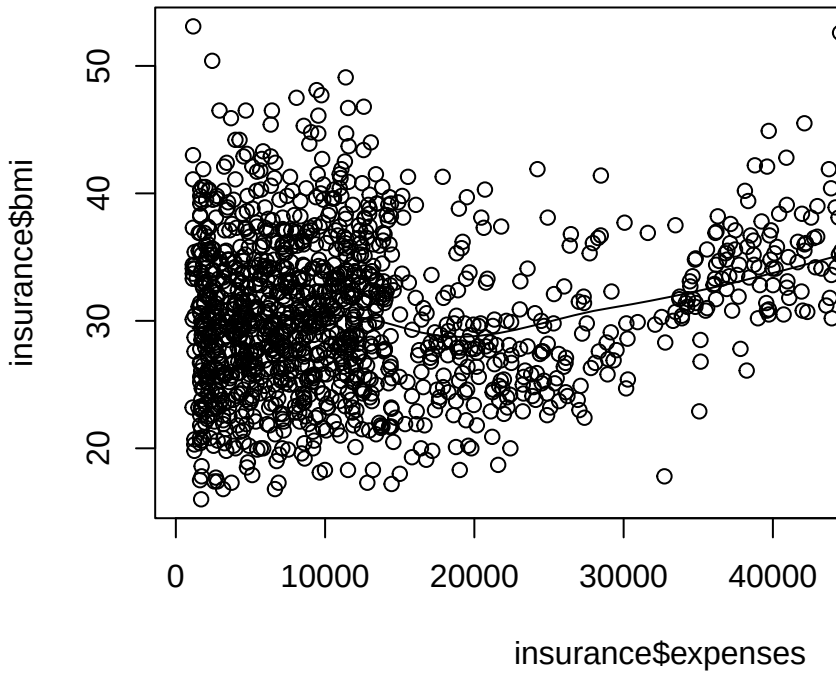
Installing the necessary packages

```
setwd("C:/Users/andre/OneDrive/Desktop/PDAT 613")

insurance <- read.csv("Insurance_A.csv", colClasses=c('numeric', 'factor', 'numeric', 'numeric', 'facto
```

Accessing the data

```
scatter.smooth(insurance$expenses, insurance$bmi)
```



Initial view of BMI's impact on EXPENSES

```
cor(insurance$expenses, insurance$bmi)
```

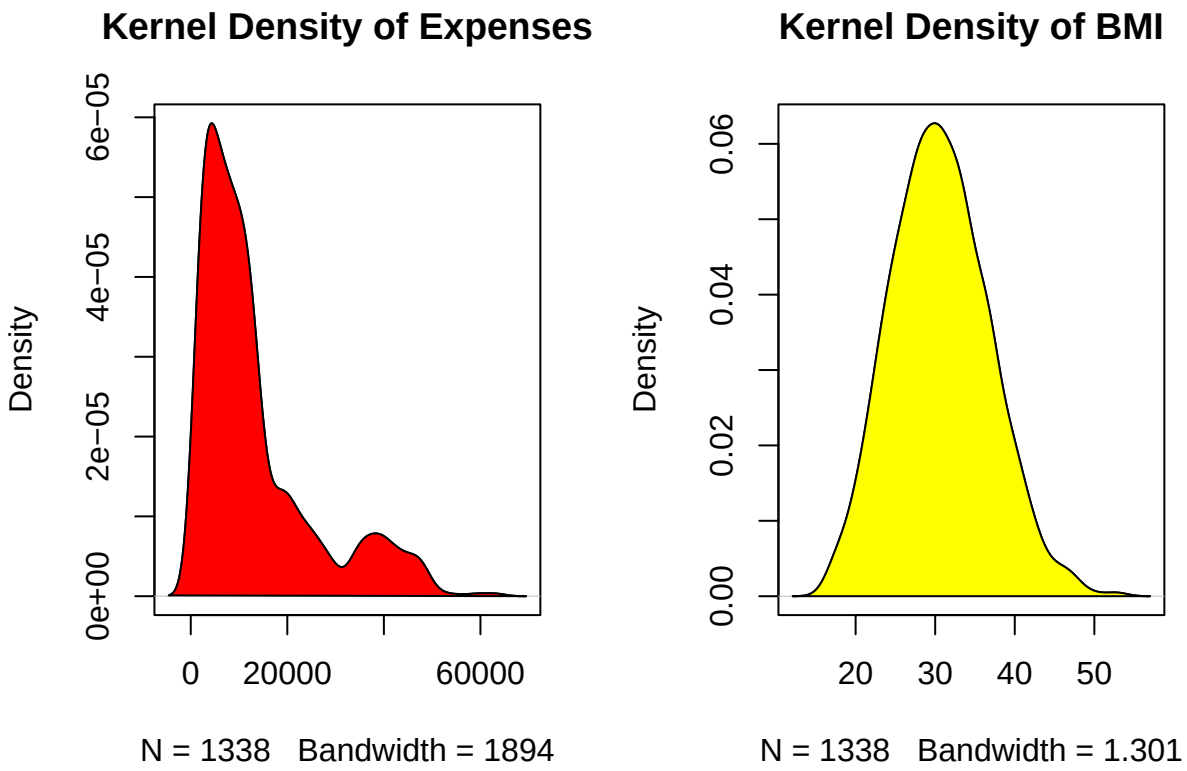
```
## [1] 0.1985763
```

```
#create 2-column graph area
par(mfrow=c(1, 2))

#density plot for expenses
e <- density(insurance$expenses)
plot(e, main="Kernel Density of Expenses")
polygon(e, col="red", border="black")

#density plot for bmi
d <- density(insurance$bmi)
plot(d, main="Kernel Density of BMI")
polygon(d, col="yellow", border="black")
```

Looking the distribution of the data (http://r-statistics.co/Linear-Regression.html)

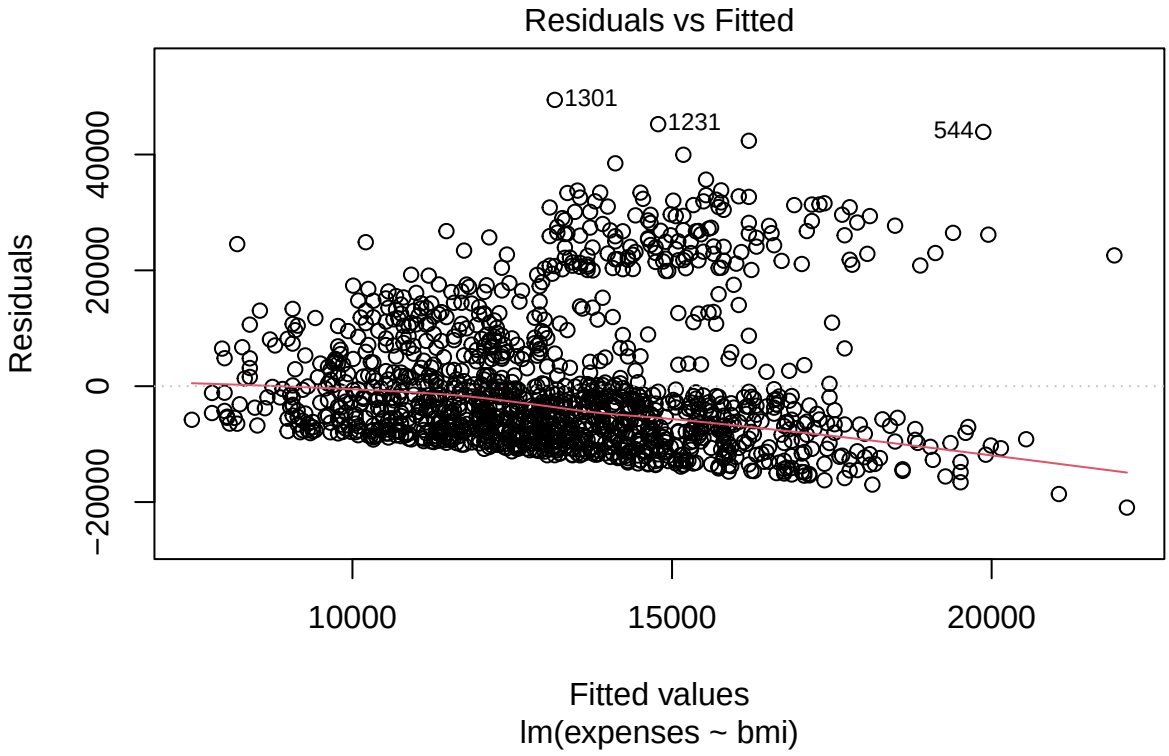


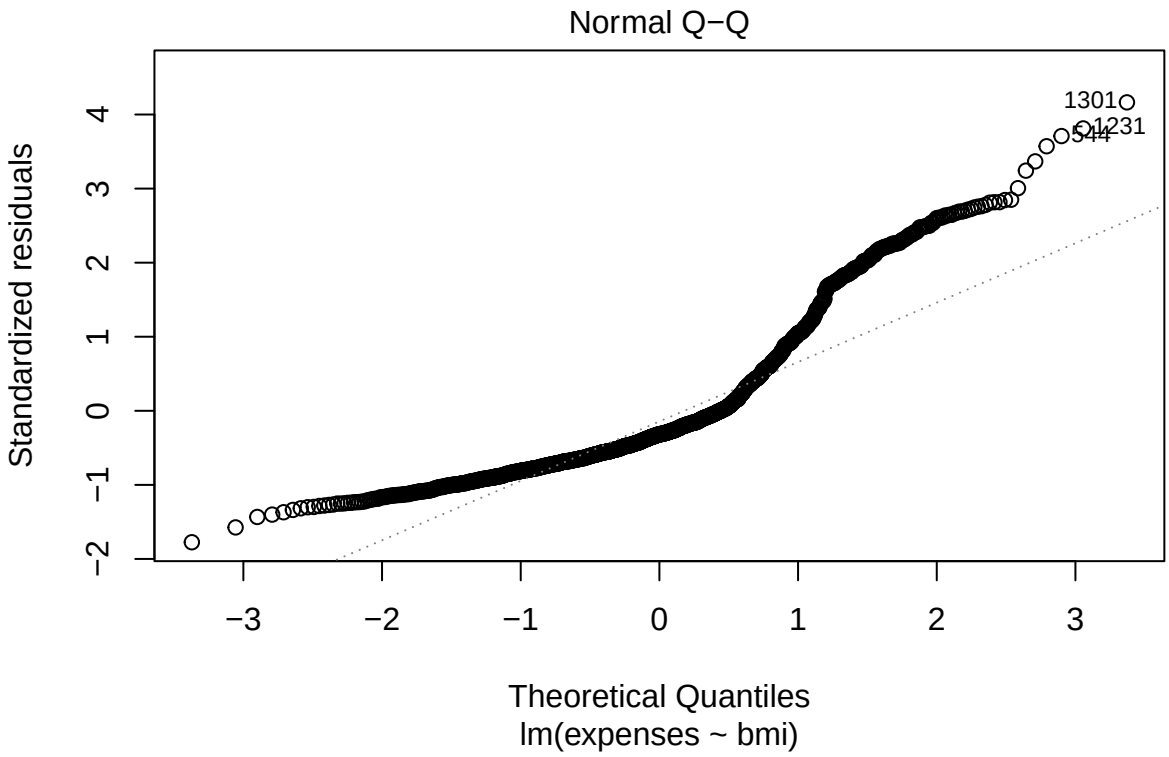

```
simpleLR <- lm(expenses ~ bmi, insurance)
summary(simpleLR)
```

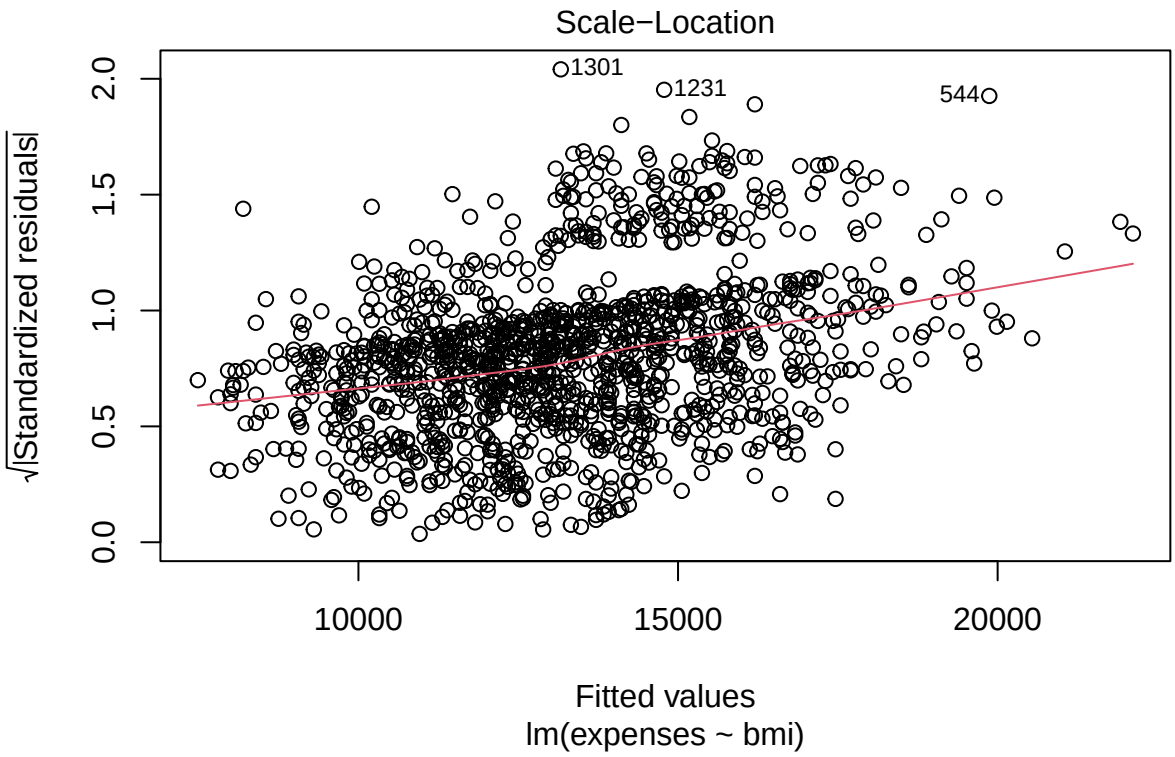
Creating Simple Linear Regression model

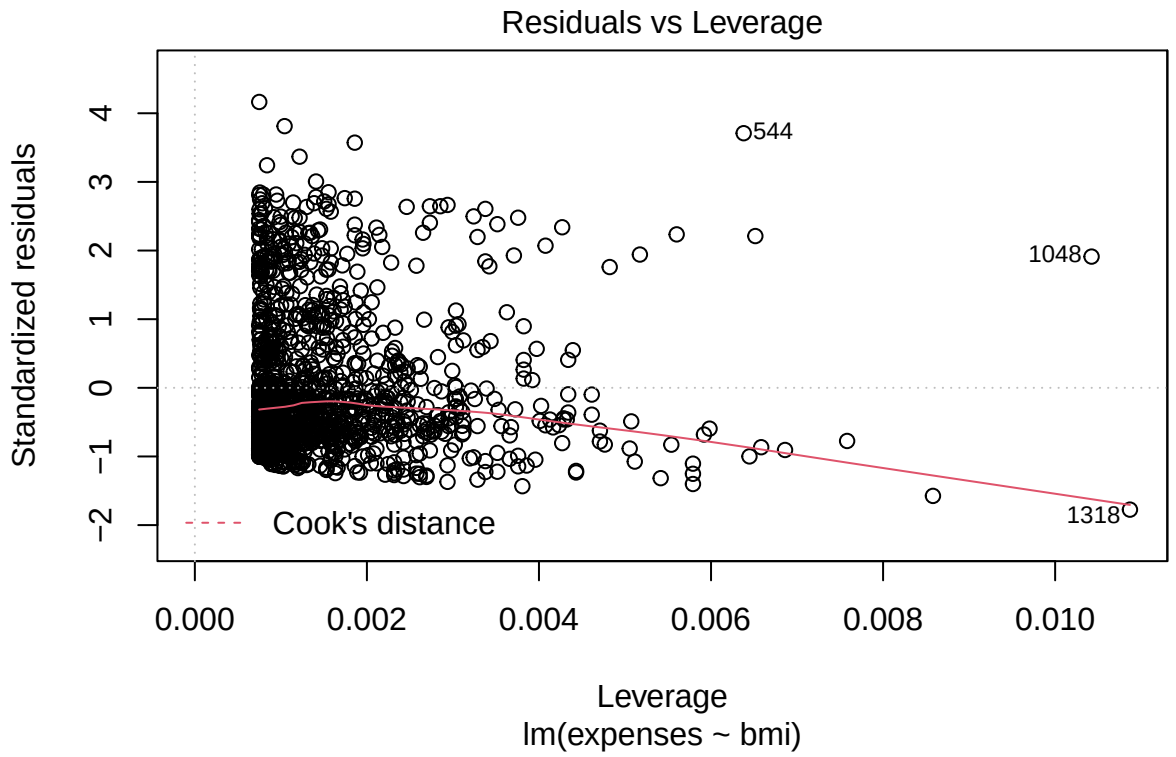
```
##
## Call:
## lm(formula = expenses ~ bmi, data = insurance)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -20954  -8125  -3750   4712  49427
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1178.18    1664.78   0.708    0.479
## bmi           394.33      53.25   7.406  2.3e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11870 on 1336 degrees of freedom
## Multiple R-squared:  0.03943,    Adjusted R-squared:  0.03871
## F-statistic: 54.84 on 1 and 1336 DF,  p-value: 2.302e-13
```

```
plot(simpleLR)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(expenses ~ bmi)
1301
1231
544
−3
−2
−1
0
1
2
3
−2
−1
0
1
2
3
4


Scale−Location
1301
1231
544
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
10000
15000
20000
Fitted values
lm(expenses ~ bmi)




```
#Entire model
full.model <- lm(expenses ~., data = insurance)
summary(full.model)
```

Creating Multiple Lineare Regression model

```
##
## Call:
## lm(formula = expenses ~ ., data = insurance)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11302.7  -2850.9   -979.6   1383.9  29981.7
##
## 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 ***
## sexmale           -131.3      332.9  -0.395 0.693255
## bmi                339.3       28.6  11.864  < 2e-16 ***
## children           475.7      137.8   3.452 0.000574 ***
## 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
```

```
#Forward and Backward variable selection model
step.model <- stepAIC(full.model, direction = "both", trace = FALSE)
summary(step.model)
```

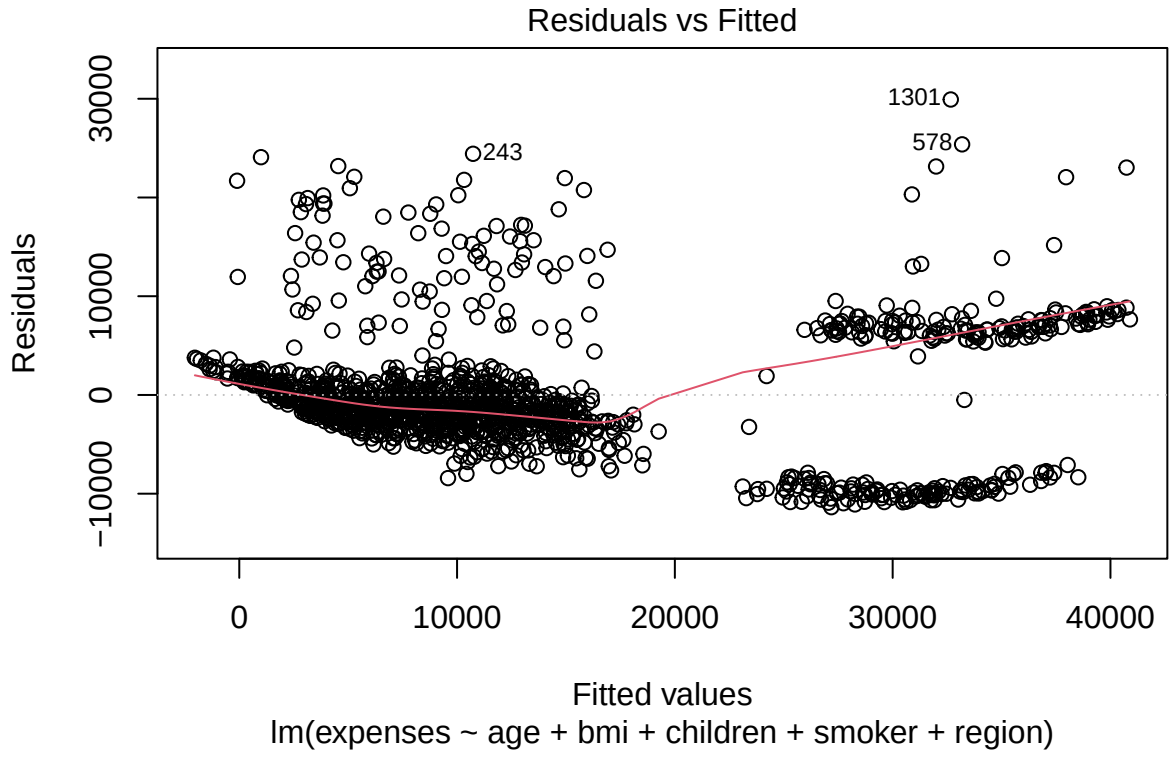
```
##
## Call:
## lm(formula = expenses ~ age + bmi + children + smoker + region,
##     data = insurance)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11365.0  -2839.4   -985.3   1375.5  29924.5
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -11993.31     978.75 -12.254  < 2e-16 ***
## age                256.96      11.89  21.609  < 2e-16 ***
## bmi                338.76      28.56  11.862  < 2e-16 ***
## children           474.75     137.74   3.447 0.000585 ***
## smokeryes        23835.24     411.84  57.875  < 2e-16 ***
## regionnorthwest   -352.01     476.11  -0.739 0.459825
## regionsoutheast  -1034.93     478.53  -2.163 0.030738 *
## regionsouthwest   -958.63     477.76  -2.007 0.045003 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6060 on 1330 degrees of freedom
## Multiple R-squared:  0.7509, Adjusted R-squared:  0.7496
## F-statistic: 572.7 on 7 and 1330 DF,  p-value: < 2.2e-16
```

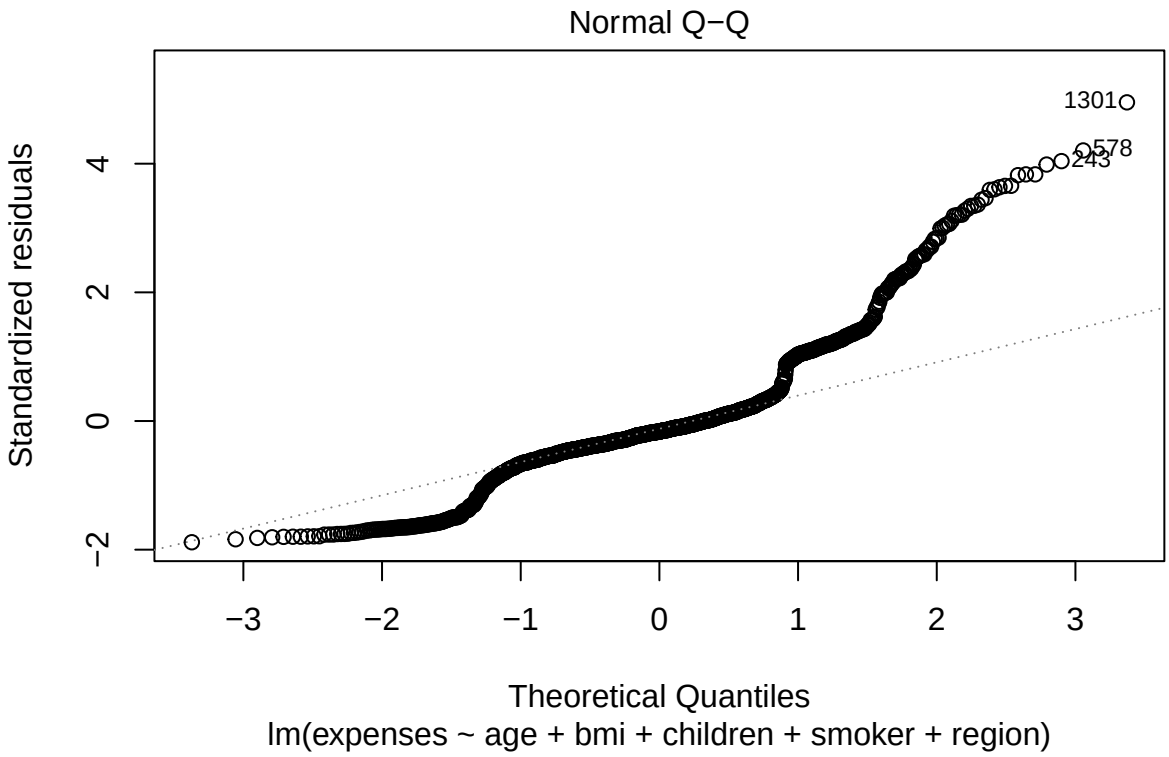
```
#Backward selection only model. It has the same output as step.model
back.model <- stepAIC(full.model, direction = "backward", trace = FALSE)
summary(back.model)
```

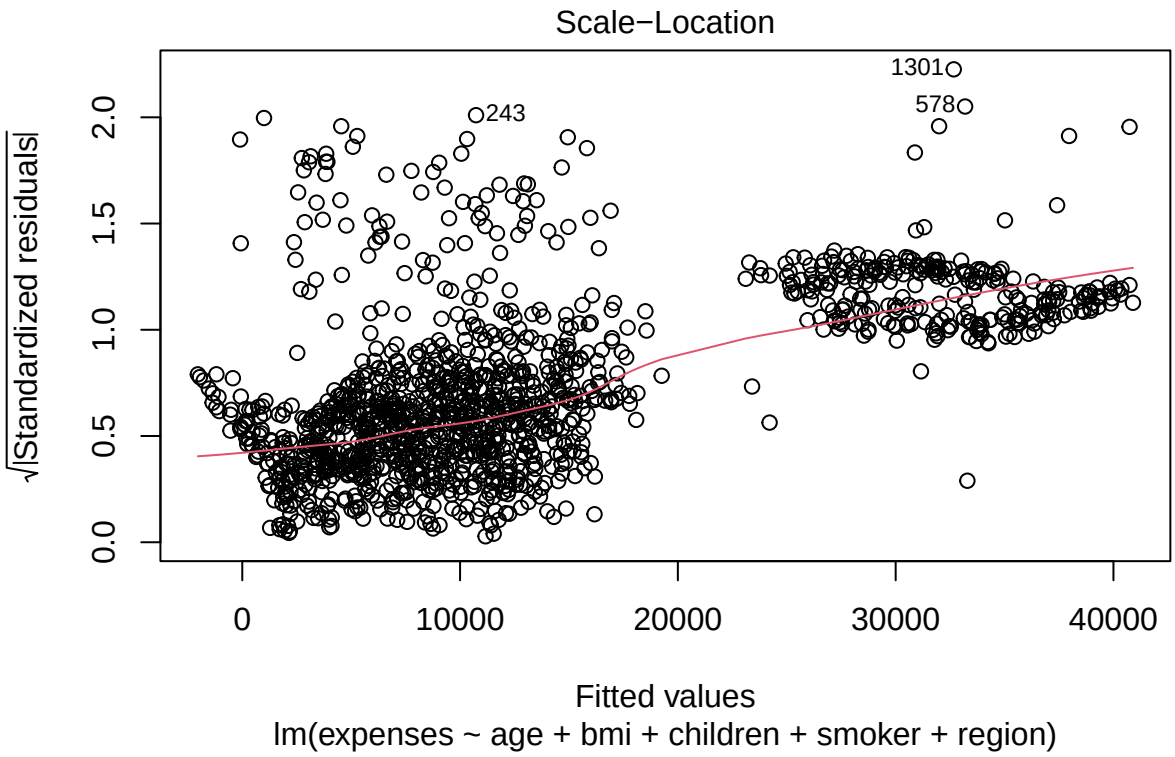
```
##
## Call:
## lm(formula = expenses ~ age + bmi + children + smoker + region,
##     data = insurance)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11365.0  -2839.4   -985.3   1375.5  29924.5
##
## Coefficients:
```

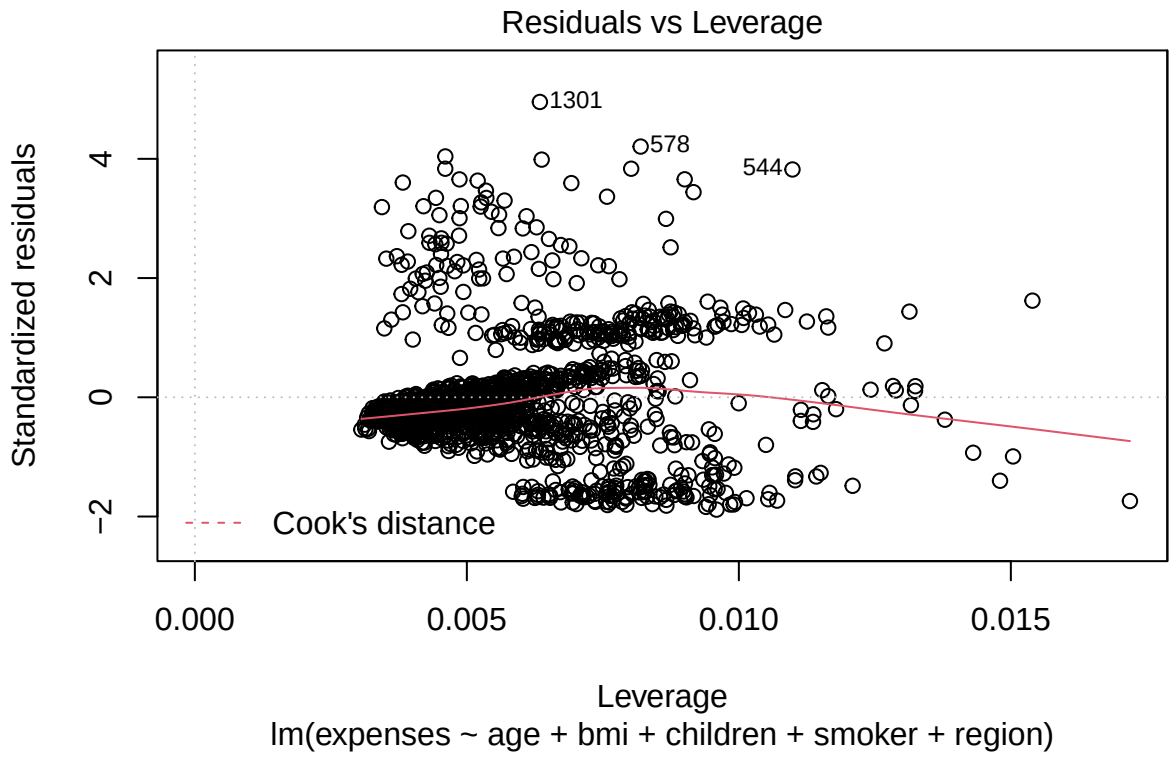
```
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -11993.31     978.75 -12.254  < 2e-16 ***
## age                256.96      11.89  21.609  < 2e-16 ***
## bmi                338.76      28.56  11.862  < 2e-16 ***
## children           474.75     137.74   3.447 0.000585 ***
## smokeryes        23835.24     411.84  57.875  < 2e-16 ***
## regionnorthwest   -352.01     476.11  -0.739 0.459825
## regionsoutheast  -1034.93     478.53  -2.163 0.030738 *
## regionsouthwest   -958.63     477.76  -2.007 0.045003 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6060 on 1330 degrees of freedom
## Multiple R-squared:  0.7509, Adjusted R-squared:  0.7496
## F-statistic: 572.7 on 7 and 1330 DF,  p-value: < 2.2e-16
```

```
plot(step.model)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(expenses ~ age + bmi + children + smoker + region)
1301
578
243
−2
0
2
4
−3
−2
−1
0
1
2
3


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
243
1301
578
0
10000
20000
30000
40000
Fitted values
lm(expenses ~ age + bmi + children + smoker + region)




```r
#finding the one best predictor
models <- regsubsets(expenses ~., data = insurance, nvmax = 1, method = "seqrep")
summary(models)
```

Finding the singular best predictor

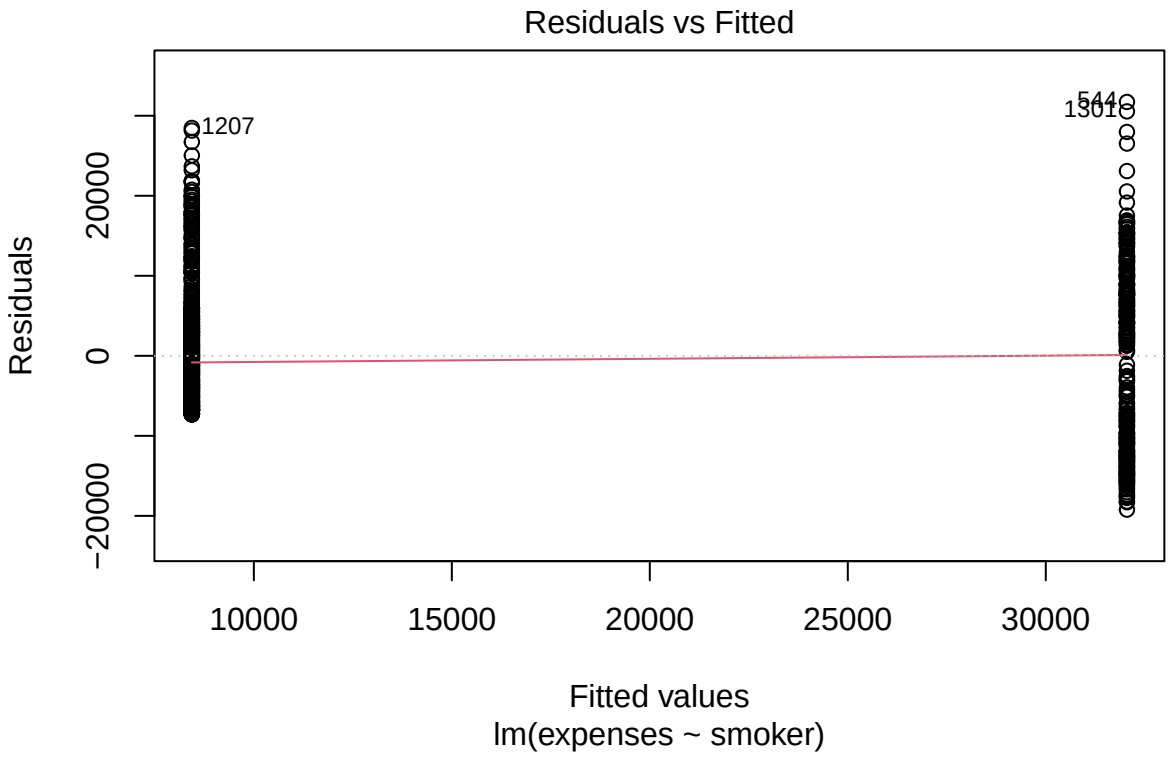
```
## Subset selection object
## Call: regsubsets.formula(expenses ~ ., data = insurance, nvmax = 1,
##     method = "seqrep")
## 8 Variables  (and intercept)
##                  Forced in Forced out
## age                  FALSE      FALSE
## sexmale              FALSE      FALSE
## bmi                  FALSE      FALSE
## children             FALSE      FALSE
## smokeryes            FALSE      FALSE
## regionnorthwest      FALSE      FALSE
## regionsoutheast      FALSE      FALSE
## regionsouthwest      FALSE      FALSE
## 1 subsets of each size up to 1
## Selection Algorithm: 'sequential replacement'
##          age sexmale bmi children smokeryes regionnorthwest regionsoutheast
## 1  ( 1 ) " " " "     " " " " "      "*"       " "              " "
```

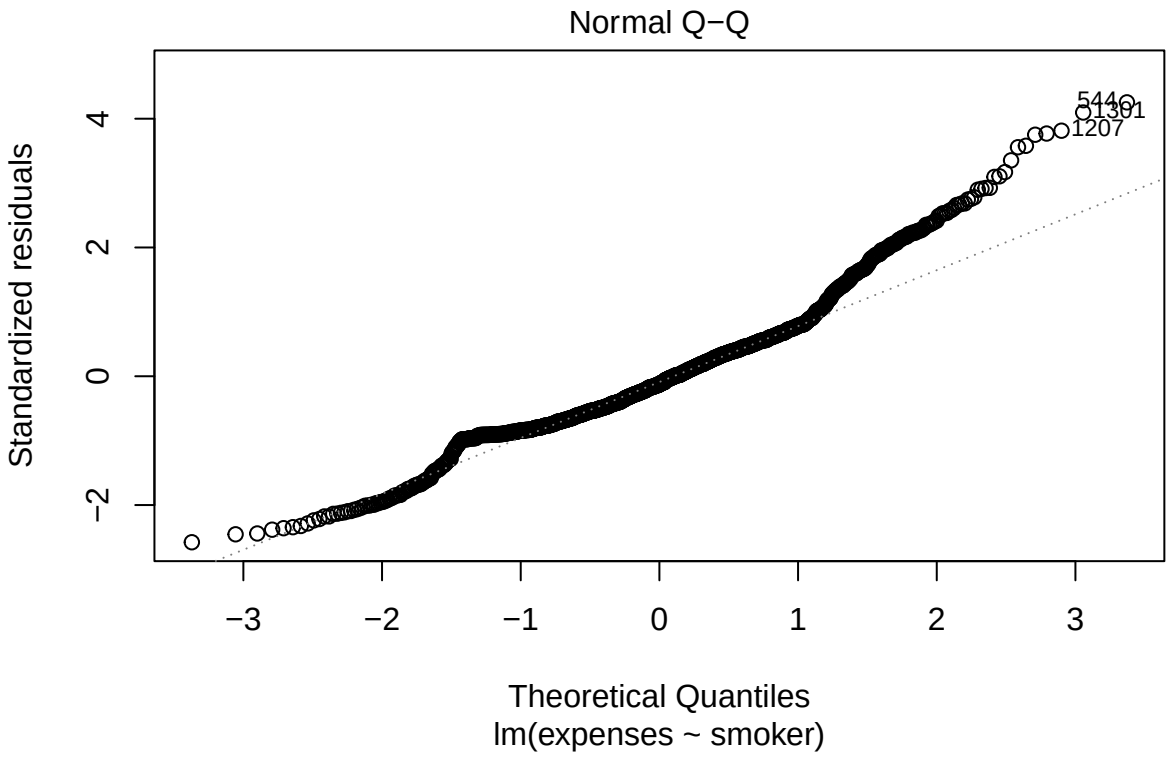
```
##          regionsouthwest
## 1  ( 1 ) " "
```

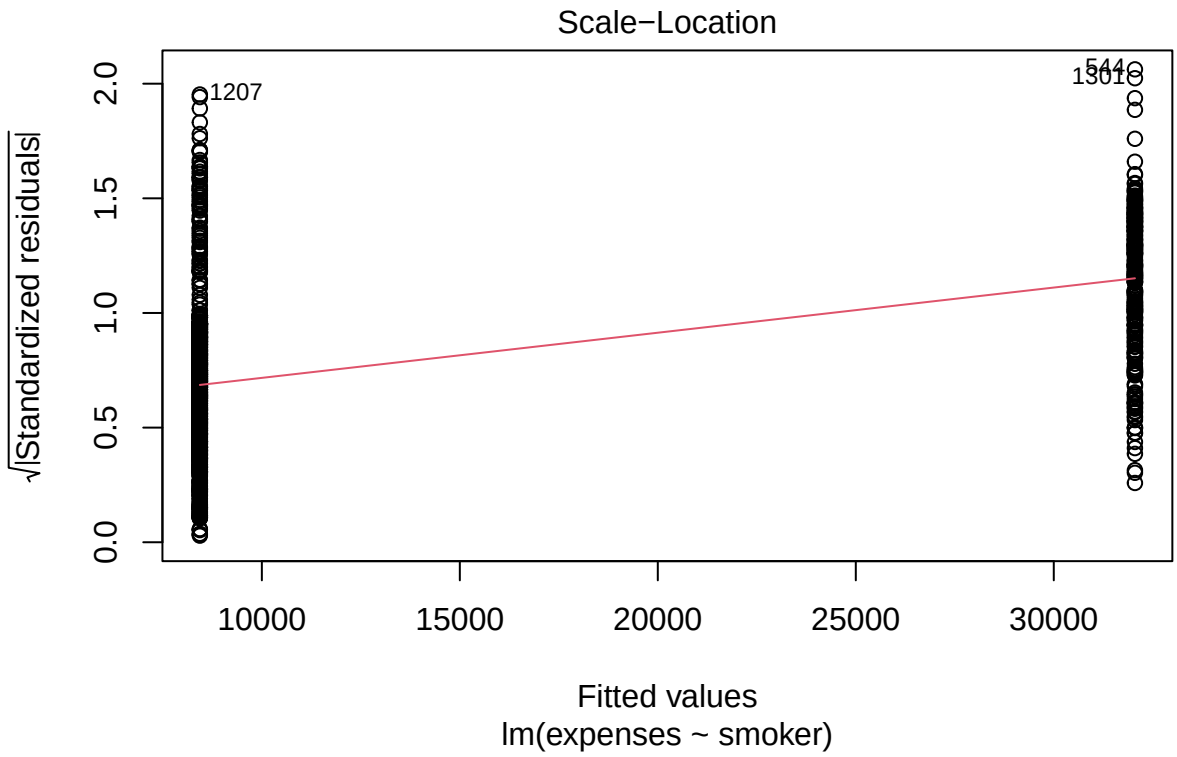
```
#indicates "smoker" is the most important factor
smoke <- lm(expenses ~ smoker, insurance)
summary(smoke)
```

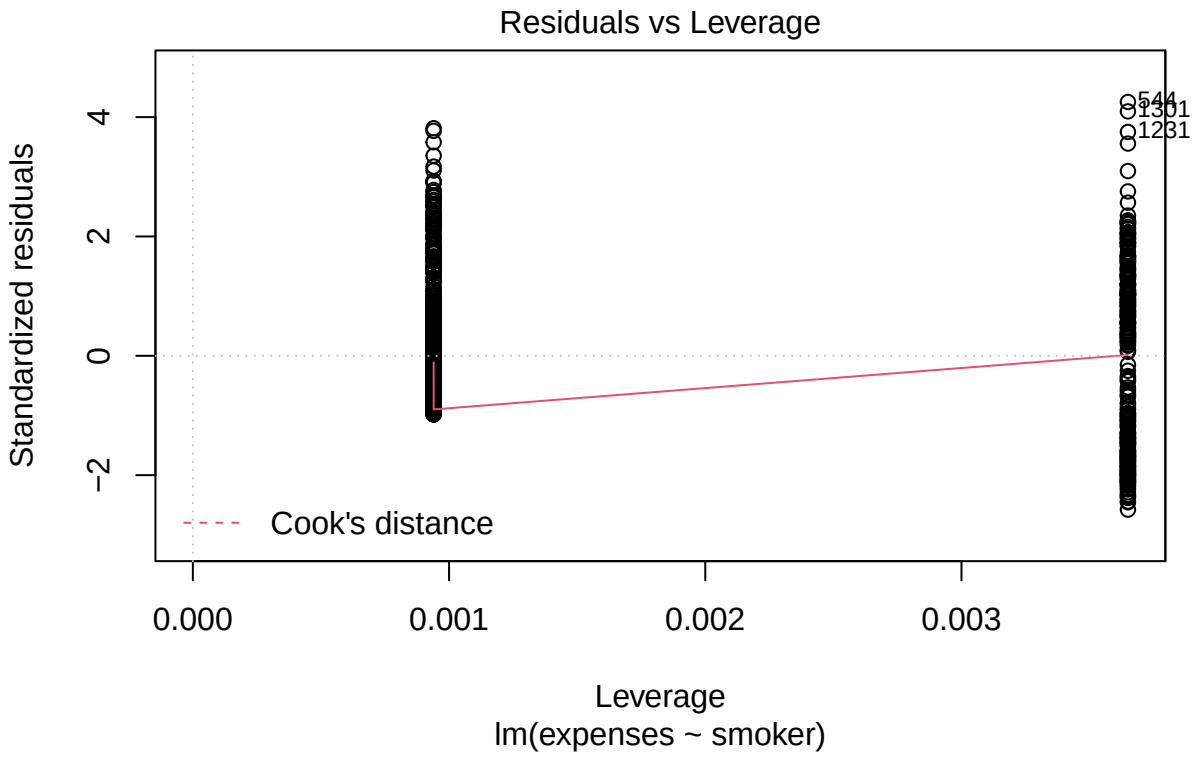
```
##
## Call:
## lm(formula = expenses ~ smoker, data = insurance)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -19221  -5042   -919   3705  31720
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   8434.3      229.0   36.83   <2e-16 ***
## smokeryes    23616.0      506.1   46.66   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7470 on 1336 degrees of freedom
## Multiple R-squared:  0.6198, Adjusted R-squared:  0.6195
## F-statistic:  2178 on 1 and 1336 DF,  p-value: < 2.2e-16
```

```
plot(smoke)
```


Residuals vs Fitted
Residuals
20000
0
−20000
1207
1301
10000
15000
20000
25000
30000
Fitted values
lm(expenses ~ smoker)


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(expenses ~ smoker)
5440
1301
1207
−3
−2
−1
0
1
2
3
−2
0
2
4


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
1207
544
13018
10000
15000
20000
25000
30000
Fitted values
lm(expenses ~ smoker)



Breaking the data into smokers and non-smokers

```
smoke <- insurance %>%
  filter(smoker=='yes') %>%
  dplyr::select(-smoker)

no.smoke <- insurance %>%
  filter(smoker=="no") %>%
  dplyr::select(-smoker)
```

```
smoke.lm <- lm(expenses ~., data=smoke)
smoke.step <- stepAIC(smoke.lm, direction = "both", trace = FALSE)
summary(smoke.step)
```

Backward regression on model on the smoker dataset

```
##
## Call:
## lm(formula = expenses ~ age + bmi, data = smoke)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14627.3  -4276.2   -221.9   3649.2  29266.8
```

```
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -22367.04    1930.02  -11.59   <2e-16 ***
## age            266.16      25.04   10.63   <2e-16 ***
## bmi           1438.03      55.16   26.07   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5750 on 271 degrees of freedom
## Multiple R-squared:  0.7536, Adjusted R-squared:  0.7518
## F-statistic: 414.4 on 2 and 271 DF,  p-value: < 2.2e-16
```

Age and BMI are the only two significant factors. BMI is 7x as impactful. The model is has an adjusted R-Squared of .7518 with an F-Statistic of 414 Residual Standard Error is 5750 on 271 degrees of freedom

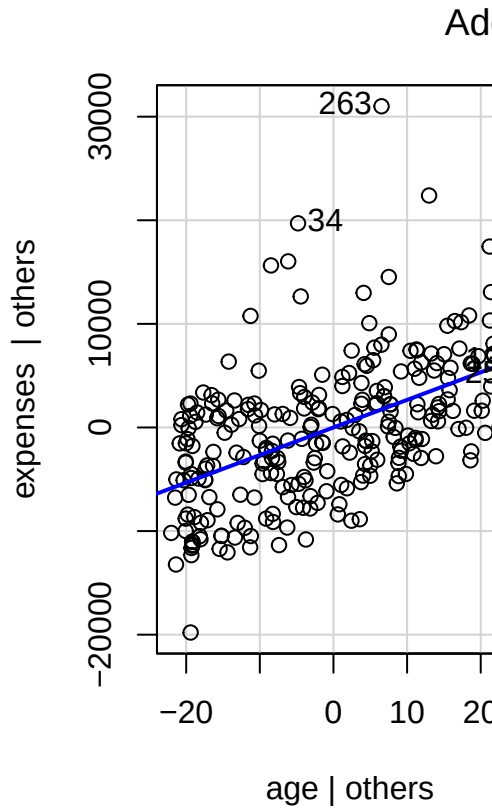
```
no.smoke.lm <- lm(expenses ~ ., data = no.smoke)
no.smoke.step <- stepAIC(no.smoke.lm, direction = "both", trace = FALSE)
summary(no.smoke.step)
```

Backward regression on model on the non-smoker datasets

```
##
## Call:
## lm(formula = expenses ~ age + sex + children + region, data = no.smoke)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2441.2 -1870.1 -1380.6  -673.9 24954.7
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -1695.87     520.71  -3.257 0.001162 **
## age               265.53      10.01  26.524  < 2e-16 ***
## sexmale          -521.01     281.62  -1.850 0.064586 .
## children          589.06     115.67   5.093 4.18e-07 ***
## regionnorthwest  -550.17     401.17  -1.371 0.170544
## regionsoutheast  -913.18     398.99  -2.289 0.022293 *
## regionsouthwest -1372.97     401.23  -3.422 0.000646 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4589 on 1057 degrees of freedom
## Multiple R-squared:  0.417,  Adjusted R-squared:  0.4137
## F-statistic:   126 on 6 and 1057 DF,  p-value: < 2.2e-16
```

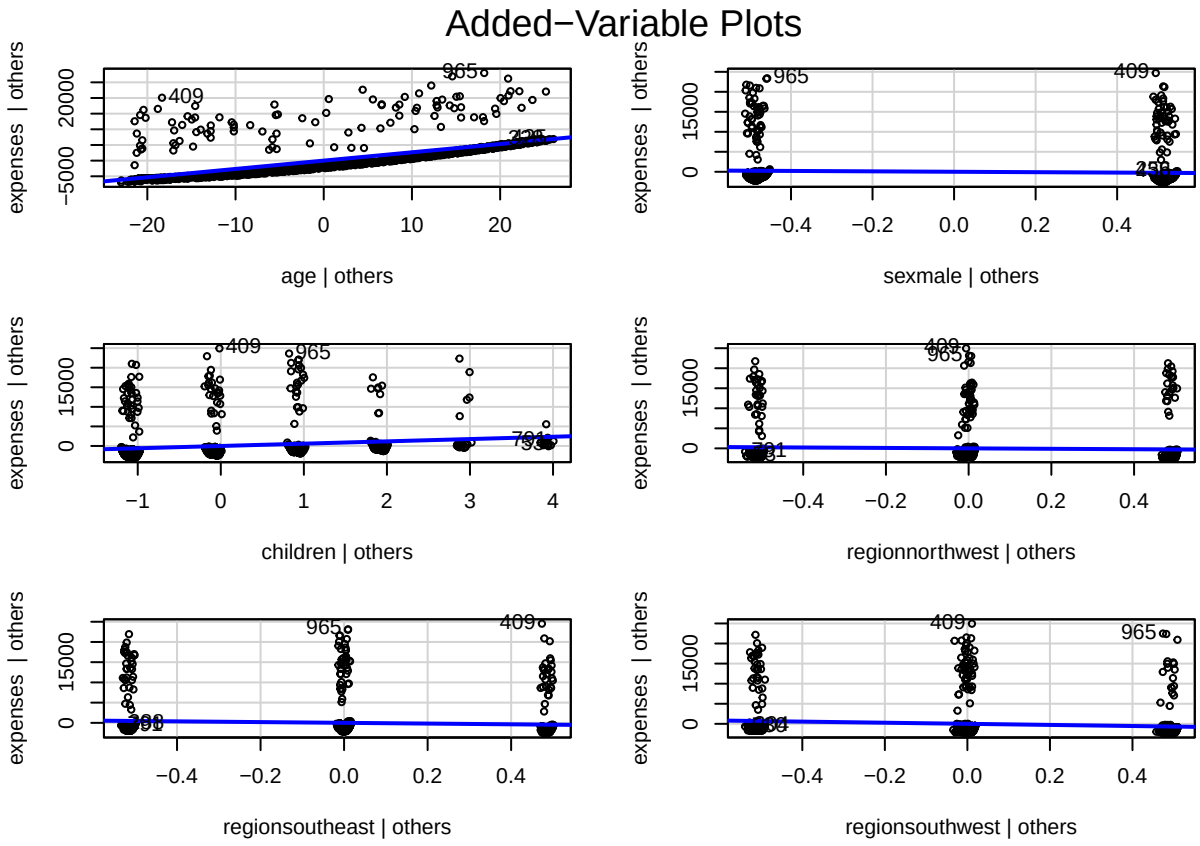
For non-smokers, age, gender, children, and region living all impact the calculation However, the adjusted r-squared is a relatively paltry .4137 with a F-statistic of 126 The Residual Standard Error is 4589 with 1057 degrees of freedom

```
avPlots(smoke.step)
```

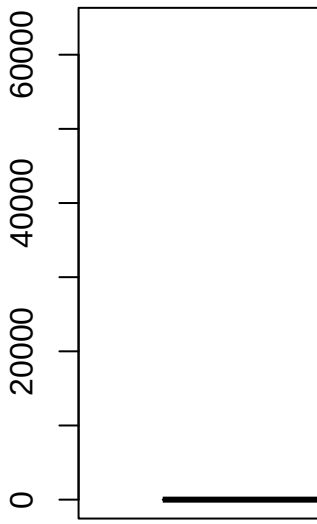


Graphically showing the affect on expense by each individual predictor

```
avPlots(no.smoke.step)
```



```
boxplot(insurance$smoker, insurance$expenses)
```



Graphically showing difference in expenses between non-smokers and smokers