Linear Regression - US State Environmental Impact

Chinpei Tang

August 30, 2016

Introduction

The dataset is some general US states statistical data from 1990-1991 that appears to be used to examine each of the state's environmental impact based on some of the potential key statistical characteristics. The data includes geographical information such as general population (in square miles, % in metropolitan area), and land area; environmental impact information such as per capita solid waste, energy consumed, toxics released, greenhouse gas; political voting performance in both house and senate (see this link); people's educational level such as mean SAT scores, % of adult high school and college graduates; and some financial details such as per pupil expenditures (primary & secondary schools) (see this link) and household incomes.

This exercise can be found on this website.

Loading the Data

```
# Set working directory
setwd("C:/Users/Chinpei/Documents/GitHub/Springboard_FDS/linear_regression")
# Read the states data. Note that the data is is RDS format
states.data <- readRDS("dataSets/states.rds")
```

Examine the Data

```
str(states.data)
```

```
## 'data.frame':    51 obs. of  21 variables:
##  $ state  : chr  "Alabama" "Alaska" "Arizona" "Arkansas" ...
##  $ region : Factor w/ 4 levels "West","N. East",..: 3 1 1 3 1 1 2 3 NA 3 ...
##  $ pop    : num  4041000 550000 3665000 2351000 29760000 ...
##  $ area   : num  52423 570374 113642 52075 155973 ...
##  $ density: num  77.08 0.96 32.25 45.15 190.8 ...
##  $ metro  : num  67.4 41.1 79 40.1 95.7 ...
##  $ waste  : num  1.11 0.91 0.79 0.85 1.51 ...
##  $ energy : int  393 991 258 330 246 273 234 349 NA 237 ...
##  $ miles  : num  10.5 7.2 9.7 8.9 8.7 ...
##  $ toxic  : num  27.86 37.41 19.65 24.6 3.26 ...
##  $ green  : num  29.2 NA 18.4 26 15.6 ...
##  $ house  : int  30 0 13 25 50 36 64 69 NA 45 ...
##  $ senate : int  10 20 33 37 47 58 87 83 NA 47 ...
##  $ csat   : int  991 920 932 1005 897 959 897 892 840 882 ...
##  $ vsat   : int  476 439 442 482 415 453 429 428 405 416 ...
##  $ msat   : int  515 481 490 523 482 506 468 464 435 466 ...
##  $ percent: int  8 41 26 6 47 29 81 61 71 48 ...
```

```
##  $ expense: int  3627 8330 4309 3700 4491 5064 7602 5865 9259 5276 ...
##  $ income : num  27.5 48.3 32.1 24.6 41.7 ...
##  $ high   : num  66.9 86.6 78.7 66.3 76.2 ...
##  $ college: num  15.7 23 20.3 13.3 23.4 ...
##  - attr(*, "datalabel")= chr "U.S. states data 1990-91"
##  - attr(*, "time.stamp")= chr " 6 Apr 2012 08:40"
##  - attr(*, "formats")= chr  "%20s" "%9.0g" "%9.0g" "%9.0g" ...
##  - attr(*, "types")= int  20 251 254 254 254 254 254 252 254 254 ...
##  - attr(*, "val.labels")= chr  "" "region" "" "" ...
##  - attr(*, "var.labels")= chr  "State" "Geographical region" "1990 population" "Land area, square mi
##  - attr(*, "expansion.fields")=List of 4
##   ..$ : chr  "_dta" "_lang_c" "default"
##   ..$ : chr  "_dta" "_lang_list" "default"
##   ..$ : chr  "_dta" "__xi__Vars__To__Drop__" "_Iregion_2 _Iregion_3 _Iregion_4 _IregXperce_2 _IregXp
##   ..$ : chr  "_dta" "__xi__Vars__Prefix__" "_I _I _I _I _I"
##  - attr(*, "version")= int 12
##  - attr(*, "label.table")=List of 1
##   ..$ region: Named int  1 2 3 4
##   .. ..- attr(*, "names")= chr  "West" "N. East" "South" "Midwest"
```

```
summary(states.data)
```

```
##     state               region        pop               area        
##  Length:51          West   :13   Min.   :  454000   Min.   :  1045  
##  Class :character   N. East: 9   1st Qu.: 1299750   1st Qu.: 36802  
##  Mode  :character   South  :16   Median : 3390500   Median : 54156  
##                     Midwest:12   Mean   : 4962040   Mean   : 70759  
##                     NA's   : 1   3rd Qu.: 5898000   3rd Qu.: 81272  
##                                  Max.   :29760000   Max.   :570374  
##                                  NA's   :1          NA's   :1       
##     density             metro             waste            energy      
##  Min.   :   0.96   Min.   : 20.40   Min.   :0.5400   Min.   :200.0  
##  1st Qu.:  31.88   1st Qu.: 46.98   1st Qu.:0.8225   1st Qu.:285.0  
##  Median :  75.76   Median : 67.55   Median :0.9600   Median :320.0  
##  Mean   : 166.04   Mean   : 64.07   Mean   :0.9888   Mean   :354.5  
##  3rd Qu.: 170.29   3rd Qu.: 81.58   3rd Qu.:1.1450   3rd Qu.:371.5  
##  Max.   :1041.92   Max.   :100.00   Max.   :1.5100   Max.   :991.0  
##  NA's   :1         NA's   :1        NA's   :1        NA's   :1      
##      miles             toxic             green            house       
##  Min.   : 5.900   Min.   :  0.770   Min.   : 11.76   Min.   : 0.00  
##  1st Qu.: 8.500   1st Qu.:  6.737   1st Qu.: 16.98   1st Qu.:31.00  
##  Median : 9.100   Median : 11.705   Median : 21.38   Median :44.50  
##  Mean   : 9.046   Mean   : 17.606   Mean   : 25.11   Mean   :44.82  
##  3rd Qu.: 9.700   3rd Qu.: 21.488   3rd Qu.: 26.34   3rd Qu.:59.25  
##  Max.   :12.800   Max.   :101.280   Max.   :114.40   Max.   :85.00  
##  NA's   :1        NA's   :1         NA's   :3        NA's   :1      
##      senate            csat             vsat            msat       
##  Min.   :10.00   Min.   : 832.0   Min.   :395.0   Min.   :435.0  
##  1st Qu.:27.00   1st Qu.: 888.0   1st Qu.:421.0   1st Qu.:467.0  
##  Median :51.00   Median : 926.0   Median :441.0   Median :485.0  
##  Mean   :49.78   Mean   : 944.1   Mean   :447.8   Mean   :496.3  
##  3rd Qu.:67.00   3rd Qu.: 997.0   3rd Qu.:476.0   3rd Qu.:521.5  
##  Max.   :97.00   Max.   :1093.0   Max.   :515.0   Max.   :578.0  
##  NA's   :1
```

```
##     percent         expense         income           high
##  Min.   : 4.00   Min.   :2960   Min.   :23.46   Min.   :64.30
##  1st Qu.:11.00   1st Qu.:4352   1st Qu.:29.88   1st Qu.:73.50
##  Median :26.00   Median :5000   Median :33.45   Median :76.70
##  Mean   :35.76   Mean   :5236   Mean   :33.96   Mean   :76.26
##  3rd Qu.:60.50   3rd Qu.:5794   3rd Qu.:36.92   3rd Qu.:80.10
##  Max.   :81.00   Max.   :9259   Max.   :48.62   Max.   :86.60
##
##     college
##  Min.   :12.30
##  1st Qu.:17.30
##  Median :19.30
##  Mean   :20.02
##  3rd Qu.:22.90
##  Max.   :33.30
##
```

Upon examining the data frame, it is found that there are attributes that describing the dataset. The following command was used to examine what the attributes mean:

```
states.info <- data.frame(attributes(states.data)[c("datalabel", "time.stamp",
"formats", "types", "val.labels", "var.labels", "expansion.fields", "version", "names")])
```

The author is unable to examine the "label.table" attributes. However, the rest of the attributes look like the following:

- datalabel: basically just says that these are U.S. states data in 1990-1991.
- time.stamp: the time the data is downloaded. They are all on Apr 6, 2012.
- formats: the data format: string, number format with the number of digits and decimal points.
- types: not exactly sure, but it looks like the amount of memory required.
- region: not exactly sure either, but there is only one data entry for "Geograhical region".
- **var.labels**: these labels explains what each of the name of the variable means, which is important.
- expansion.field: not exactly sure, but it looks like more attributes to each of the variable.
- version: probably the version of this dataset, which is 12 for all of them.
- **names**: the variable names, which is important.

```
data.frame(attributes(states.data)[c("names", "var.labels")])
```

```
##      names                       var.labels
## 1    state                            State
## 2   region             Geographical region
## 3      pop                 1990 population
## 4     area          Land area, square miles
## 5  density           People per square mile
## 6    metro Metropolitan area population, %
## 7    waste    Per capita solid waste, tons
## 8   energy Per capita energy consumed, Btu
## 9    miles    Per capita miles/year, 1,000
## 10   toxic Per capita toxics released, lbs
## 11   green Per capita greenhouse gas, tons
## 12   house    House '91 environ. voting, %
## 13  senate   Senate '91 environ. voting, %
```

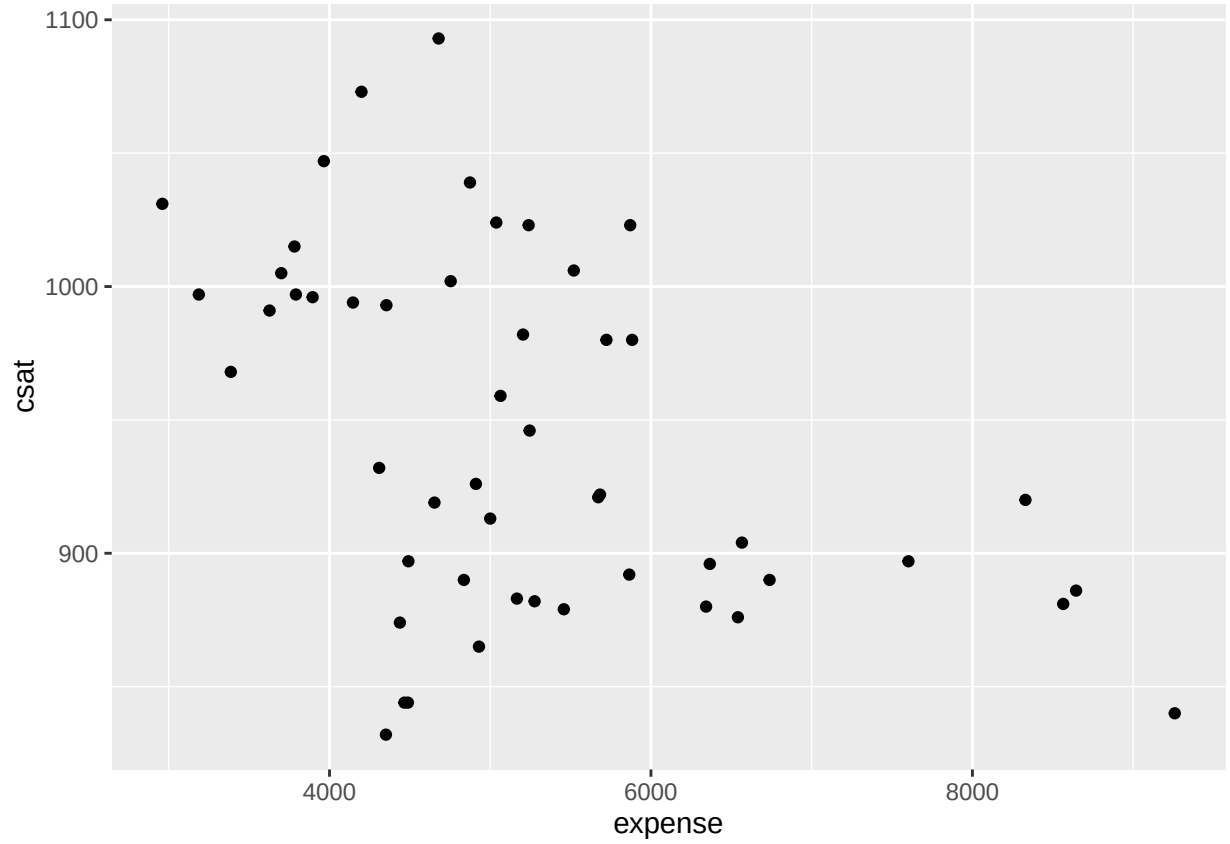
```
## 14    csat        Mean composite SAT score
## 15    vsat           Mean verbal SAT score
## 16    msat             Mean math SAT score
## 17 percent       % HS graduates taking SAT
## 18 expense Per pupil expenditures prim&sec
## 19  income Median household income, $1,000
## 20    high             % adults HS diploma
## 21 college         % adults college degree
```

Linear Regression

Pupil Expenditures and SAT Scores

First examine the correlation between per pupil expenditures (expense) and composite SAT score (csat).

```
library(ggplot2)
pl.sp.csat.exp <- ggplot(states.data, aes(x = expense, y = csat)) + geom_point(size = 1.5)
pl.sp.csat.exp
```



Perform a linear regression

```
lm.csat.exp <- lm(csat~expense, data=states.data)
summary(lm.csat.exp)
```

```
##
## Call:
## lm(formula = csat ~ expense, data = states.data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -131.811  -38.085    5.607   37.852  136.495
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.061e+03  3.270e+01   32.44  < 2e-16 ***
## expense     -2.228e-02  6.037e-03   -3.69 0.000563 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 59.81 on 49 degrees of freedom
## Multiple R-squared:  0.2174, Adjusted R-squared:  0.2015
## F-statistic: 13.61 on 1 and 49 DF,  p-value: 0.0005631
```

```
class(lm.csat.exp)
```

```
## [1] "lm"
```

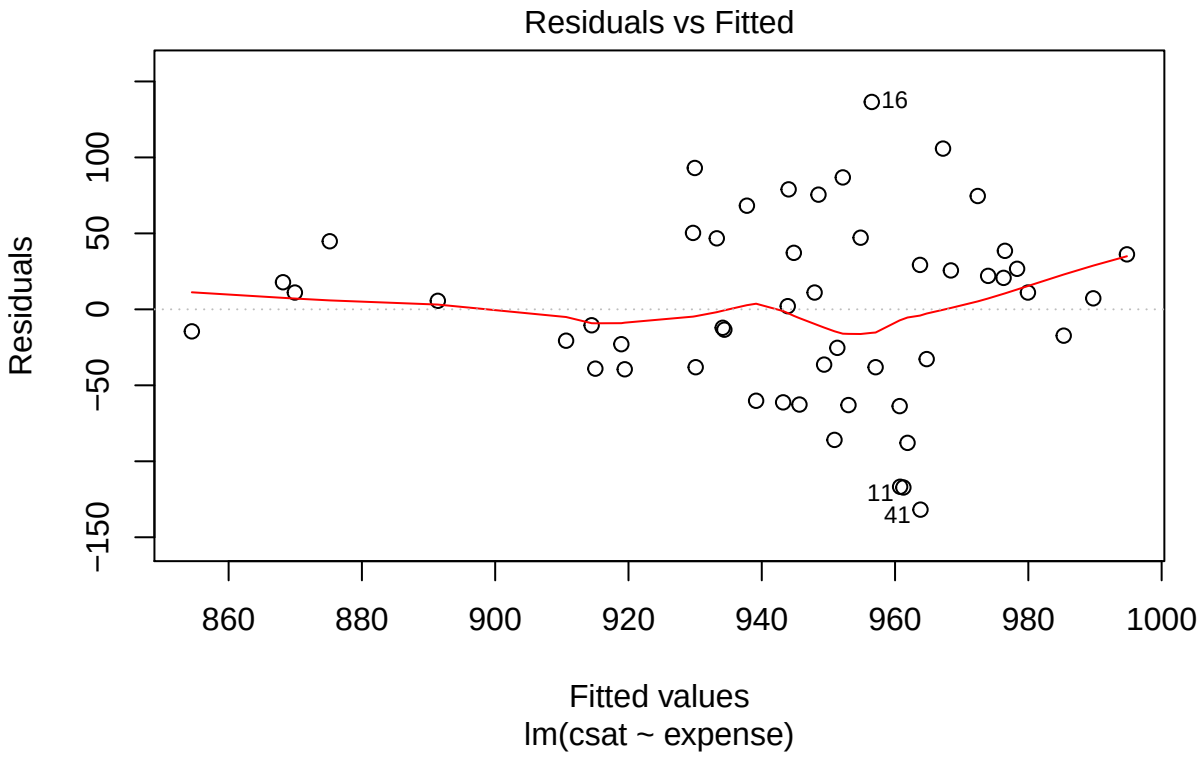
```
names(lm.csat.exp)
```

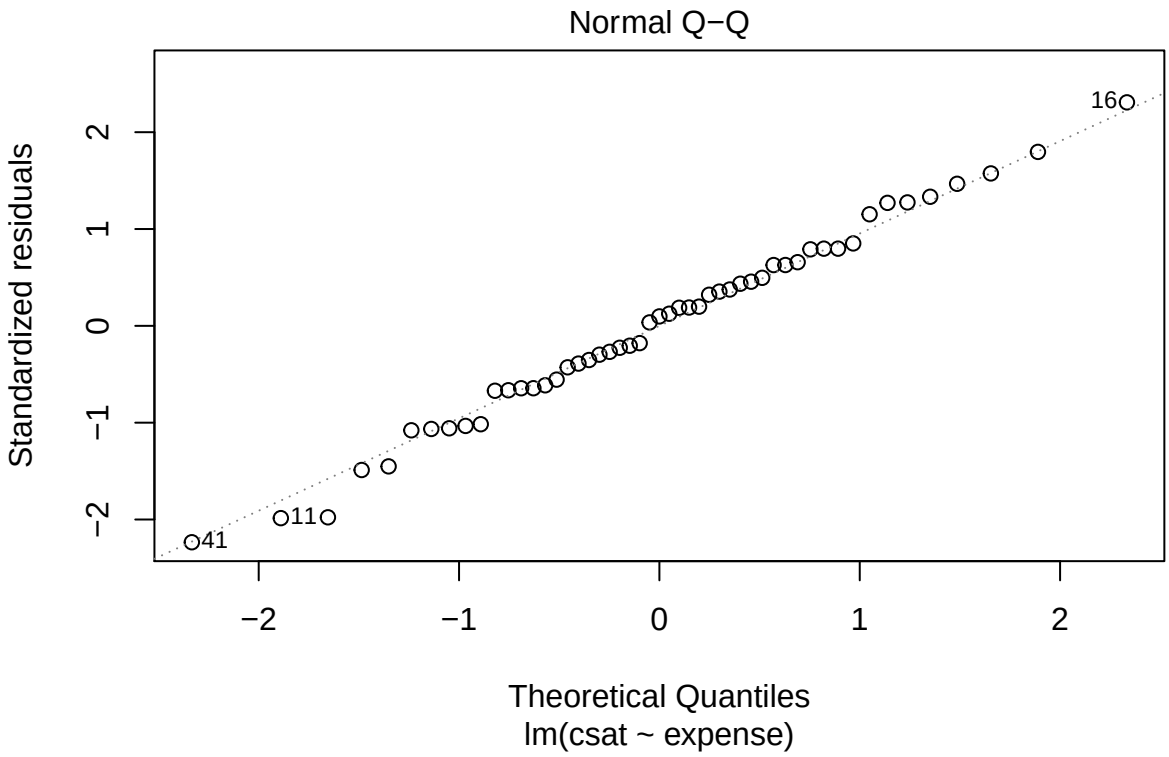
```
##  [1] "coefficients"  "residuals"     "effects"       "rank"
##  [5] "fitted.values" "assign"        "qr"            "df.residual"
##  [9] "xlevels"       "call"          "terms"         "model"
```

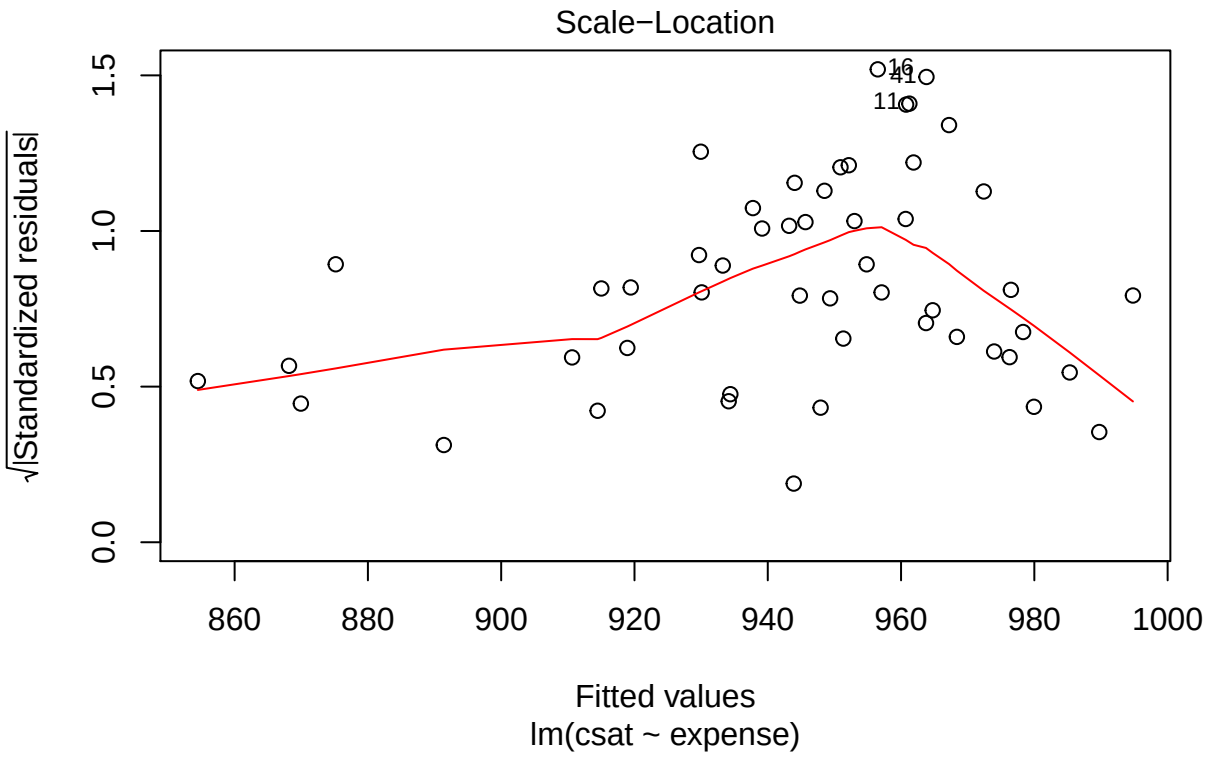
```
confint(lm.csat.exp)
```

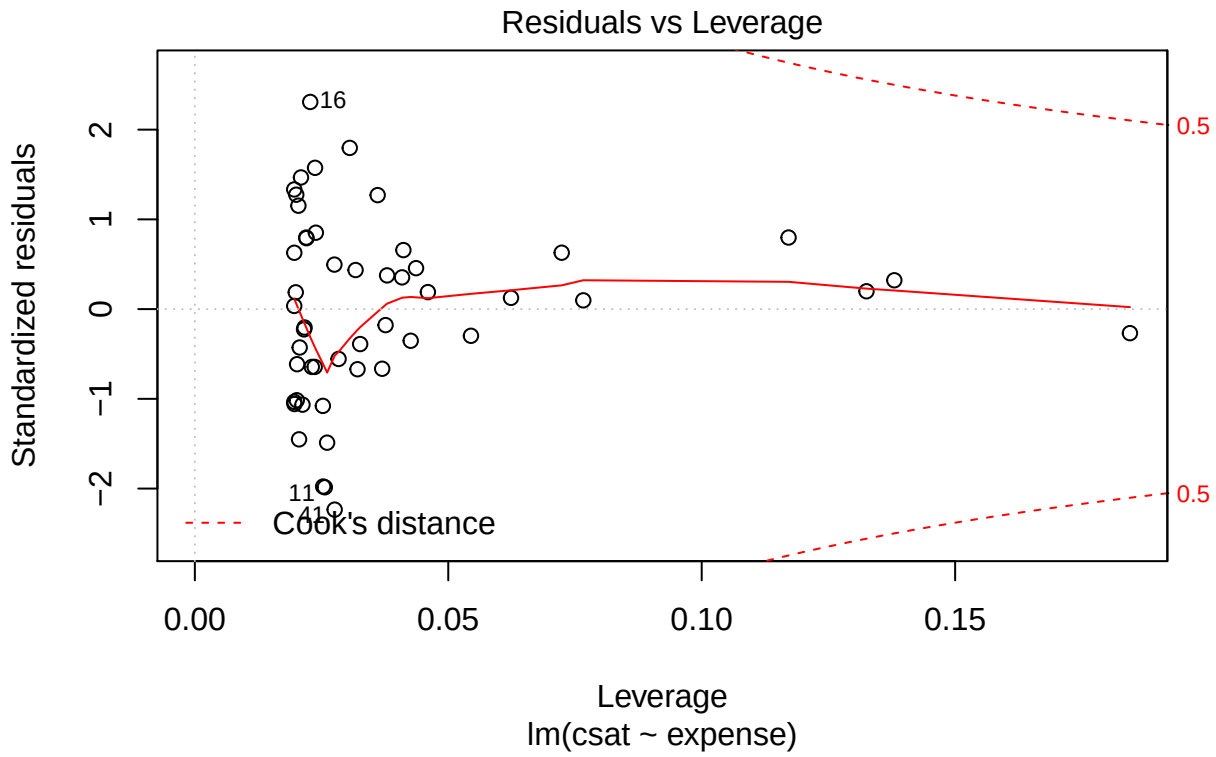
```
##                    2.5 %        97.5 %
## (Intercept) 995.01753164 1126.44735626
## expense      -0.03440768   -0.01014361
```

```
plot(lm.csat.exp)
```


Residuals vs Fitted
Residuals
100
50
0
−50
−150
16
11
41
860
880
900
920
940
960
980
1000
Fitted values
lm(csat ~ expense)

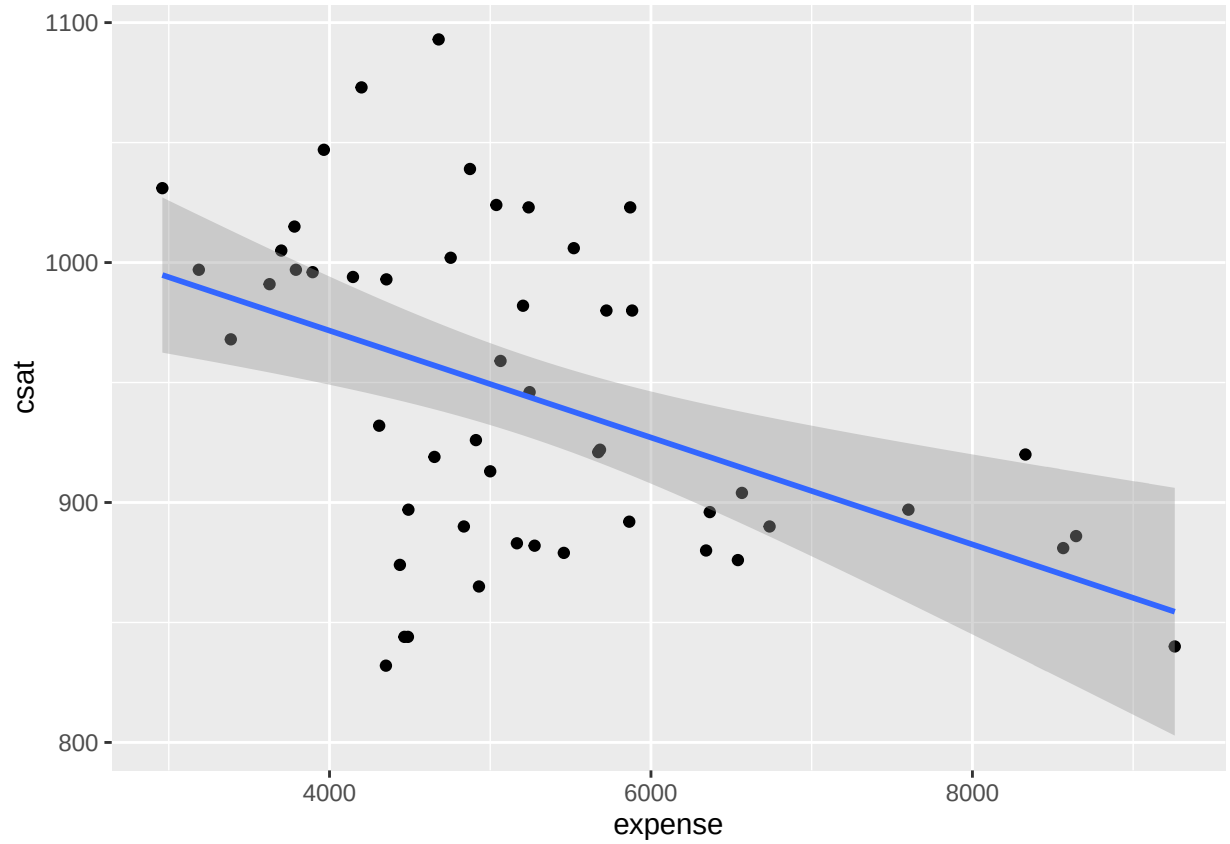

Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(csat ~ expense)
16
11
41
−2
−1
0
1
2


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
16
41
11
860
880
900
920
940
960
980
1000
Fitted values
lm(csat ~ expense)

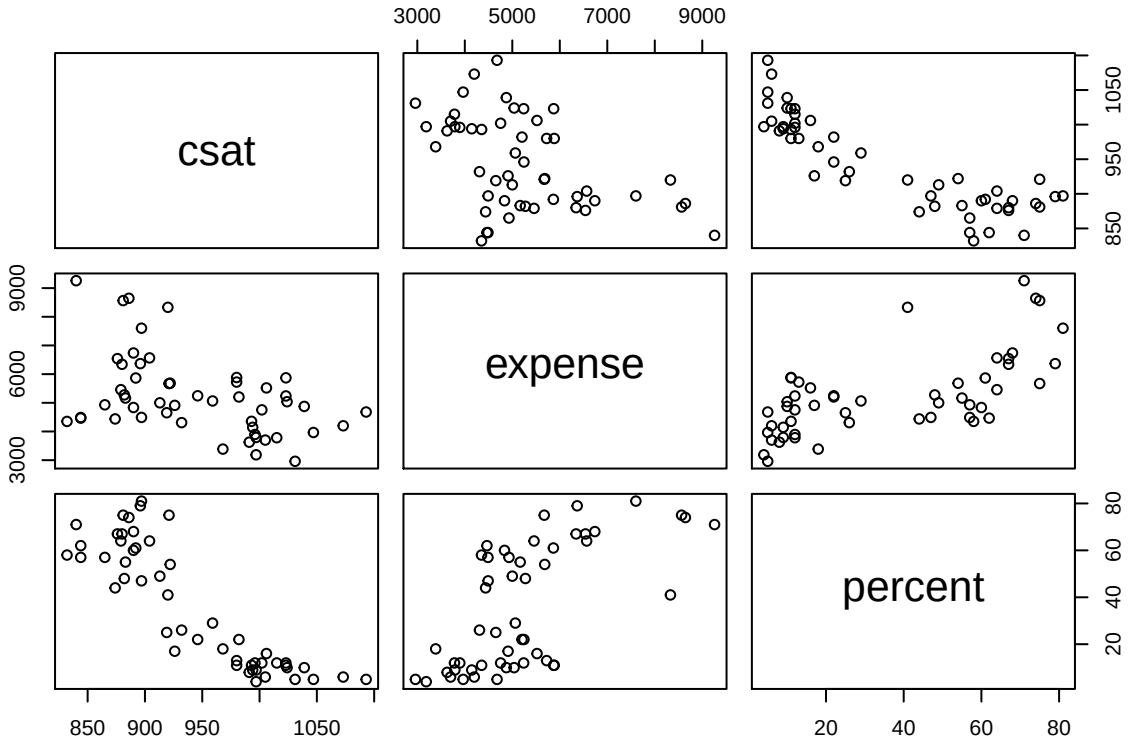


Plot it on the graph

```
pl.sp.lm.csat.exp <- pl.sp.csat.exp + stat_smooth(method = lm)
pl.sp.lm.csat.exp
```




```
df.csat.exp.percent <- states.data[c("csat", "expense", "percent")]
plot(df.csat.exp.percent)
```




```
summary(lm(csat ~ ., data = df.csat.exp.percent))
```

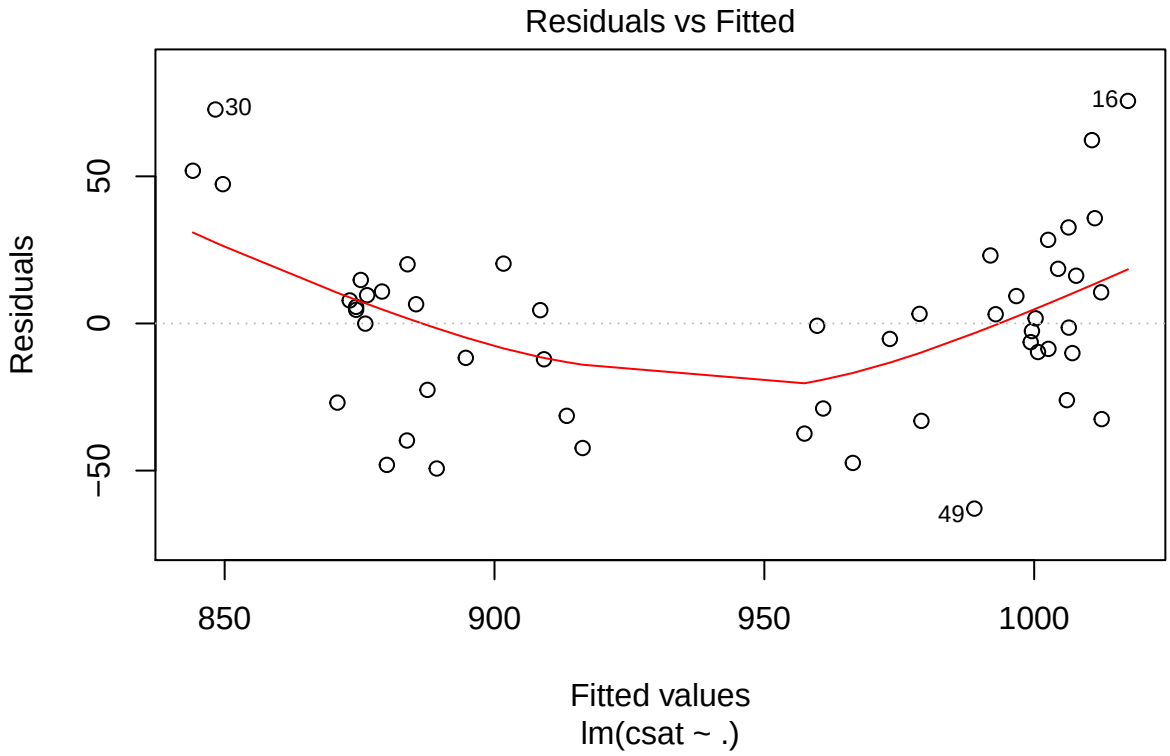
```
##
## Call:
## lm(formula = csat ~ ., data = df.csat.exp.percent)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -62.921 -24.318   1.741  15.502  75.623
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 989.807403  18.395770  53.806  < 2e-16 ***
## expense       0.008604   0.004204   2.046   0.0462 *
## percent      -2.537700   0.224912 -11.283 4.21e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 31.62 on 48 degrees of freedom
## Multiple R-squared:  0.7857, Adjusted R-squared:  0.7768
## F-statistic: 88.01 on 2 and 48 DF,  p-value: < 2.2e-16
```

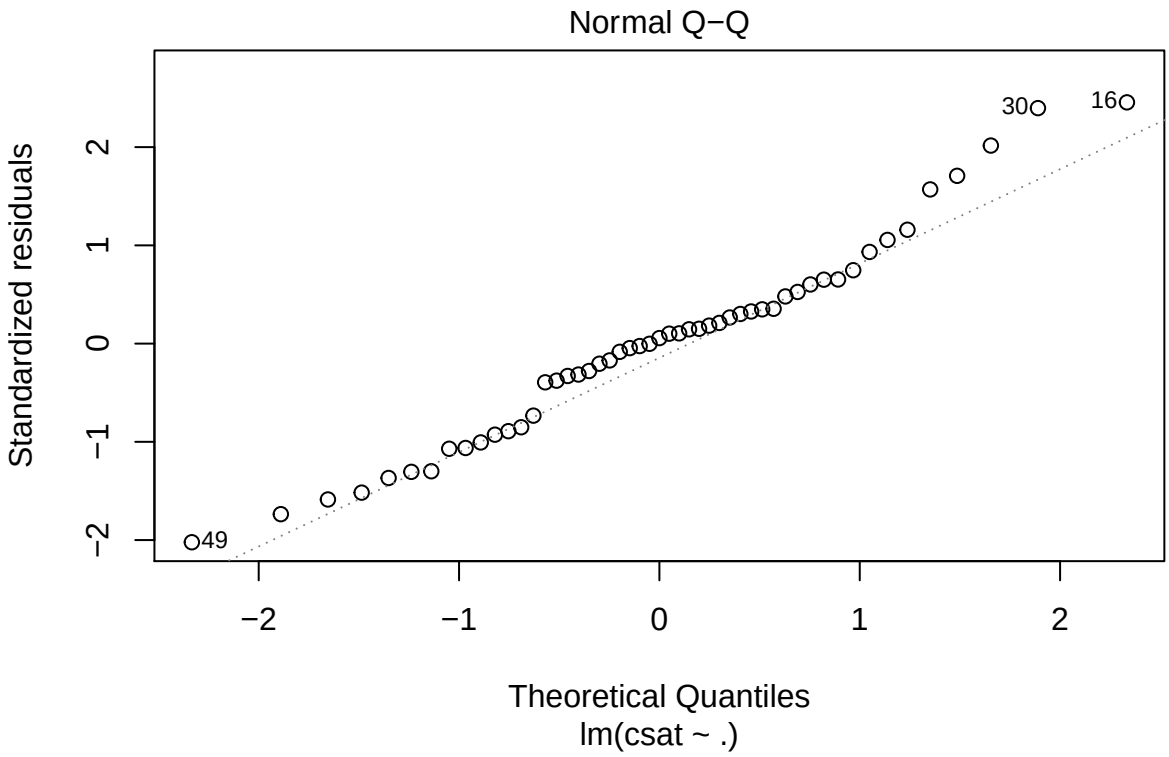
```
confint(lm(csat ~ ., data = df.csat.exp.percent))
```

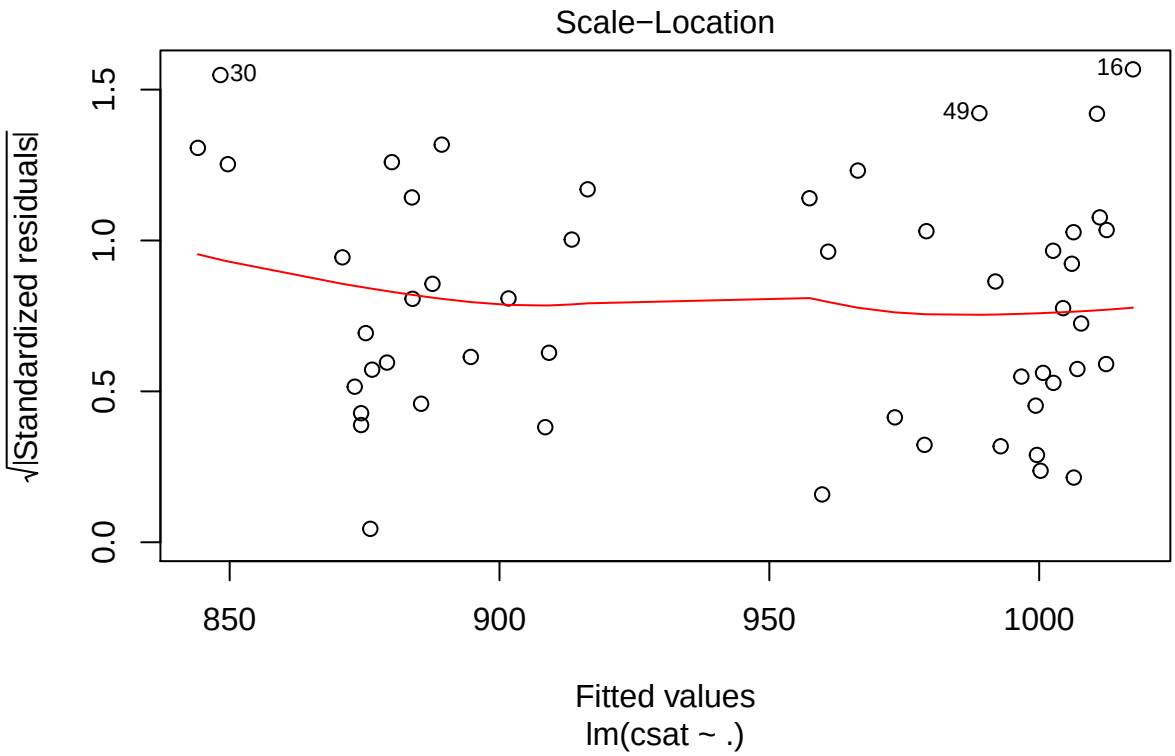
```
##                    2.5 %        97.5 %
```

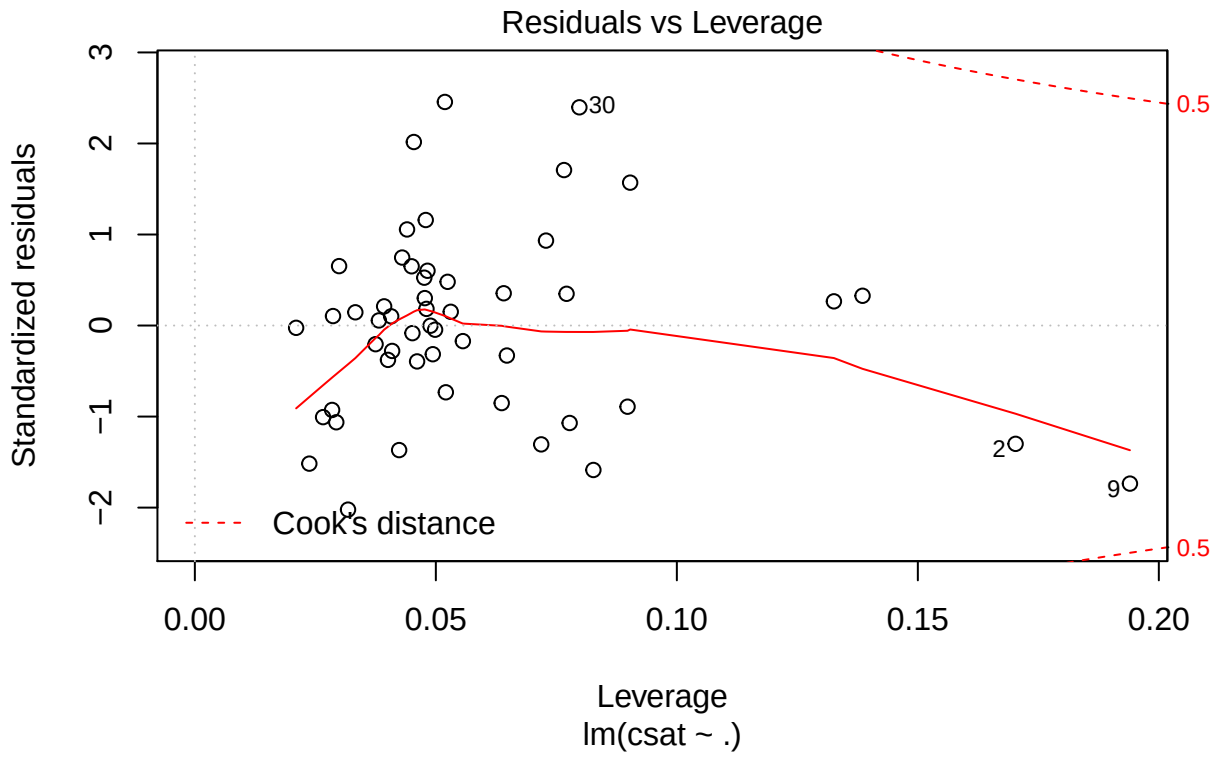
```
## (Intercept)  9.528202e+02 1026.79457731
## expense      1.505116e-04    0.01705769
## percent     -2.989915e+00   -2.08548496
```

```
plot(lm(csat ~ ., data = df.csat.exp.percent))
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(csat ~ .)
30
16
49
−2
−1
0
1
2


Scale−Location
30
16
49
√|Standardized residuals|
0.0
0.5
1.0
1.5
850
900
950
1000
Fitted values
lm(csat ~ .)



What if we do linear regression on all the variables:

```
states.data.num <- states.data[c("pop", "area", "density", "metro", "waste", "energy",
"miles", "toxic", "green", "house", "senate", "csat", "vsat", "msat", "percent",
"expense", "income", "high", "college")]
states.data.num.na.omit <- na.omit(states.data.num)
lm.csat.all <- lm(energy~., data=data.frame(states.data.num.na.omit))
summary(lm.csat.all)
```

```
##
## Call:
## lm(formula = energy ~ ., data = data.frame(states.data.num.na.omit))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -126.554  -27.297    0.968   21.840  159.899
##
## Coefficients: (1 not defined because of singularities)
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.536e+01  5.083e+02  -0.089 0.929477
## pop         -1.707e-06  3.568e-06  -0.478 0.635888
## area         6.663e-04  3.708e-04   1.797 0.082403 .
## density     -1.279e-02  8.882e-02  -0.144 0.886504
## metro        6.069e-01  9.384e-01   0.647 0.522692
## waste        1.316e+01  5.415e+01   0.243 0.809631
## miles        1.188e+01  1.511e+01   0.786 0.438014
```

```
## toxic        2.759e+00  6.682e-01   4.130 0.000267 ***
## green        4.426e+00  9.524e-01   4.647  6.3e-05 ***
## house        1.071e-01  9.896e-01   0.108 0.914513
## senate       1.348e-01  6.385e-01   0.211 0.834171
## csat        -2.084e-01  2.055e+00  -0.101 0.919910
## vsat         7.732e-01  4.314e+00   0.179 0.858958
## msat               NA         NA      NA       NA
## percent      8.668e-01  1.463e+00   0.593 0.557933
## expense      1.244e-02  1.558e-02   0.799 0.430823
## income       9.697e-01  4.671e+00   0.208 0.836945
## high        -1.582e+00  3.545e+00  -0.446 0.658682
## college     -6.541e+00  5.768e+00  -1.134 0.265771
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 63.12 on 30 degrees of freedom
## Multiple R-squared:  0.8141, Adjusted R-squared:  0.7087
## F-statistic: 7.726 on 17 and 30 DF,  p-value: 7.656e-07
```

For the exercise:

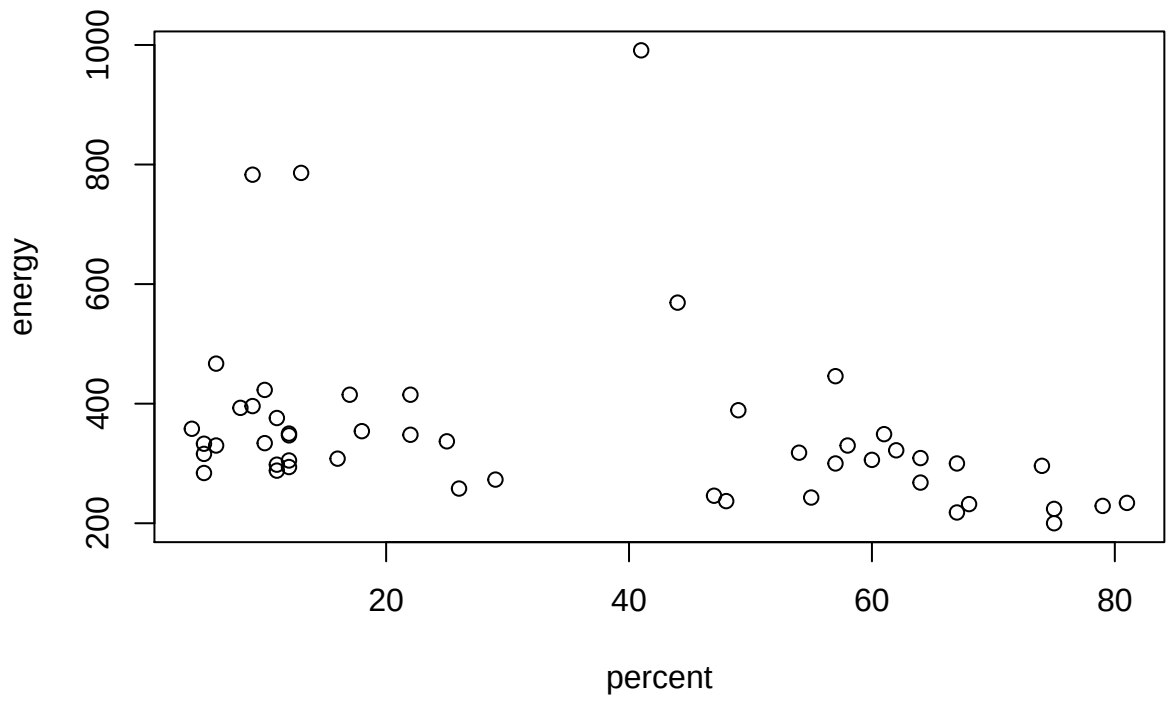
```
sts.percent.energy <- subset(states.data, select = c("percent", "energy"))
summary(sts.percent.energy)
```

```
##     percent          energy
##  Min.   : 4.00   Min.   :200.0
##  1st Qu.:11.00   1st Qu.:285.0
##  Median :26.00   Median :320.0
##  Mean   :35.76   Mean   :354.5
##  3rd Qu.:60.50   3rd Qu.:371.5
##  Max.   :81.00   Max.   :991.0
##                  NA's   :1
```

```
sts.percent.energy <- na.omit(sts.percent.energy)
summary(sts.percent.energy)
```

```
##     percent          energy
##  Min.   : 4.00   Min.   :200.0
##  1st Qu.:11.00   1st Qu.:285.0
##  Median :25.50   Median :320.0
##  Mean   :35.06   Mean   :354.5
##  3rd Qu.:59.50   3rd Qu.:371.5
##  Max.   :81.00   Max.   :991.0
```

```
plot(sts.percent.energy)
```




```
cor(sts.percent.energy)
```

```
##            percent     energy
## percent  1.0000000 -0.3040033
## energy  -0.3040033  1.0000000
```

```
per.energy.mod <- summary(lm(energy~percent, data = states.data))
per.energy.mod
```

```
##
## Call:
## lm(formula = energy ~ percent, data = states.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -122.41  -78.83  -35.70   12.71  646.76
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 415.0485    33.9594  12.222  2.4e-16 ***
## percent      -1.7270     0.7812  -2.211   0.0318 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 142 on 48 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.09242,    Adjusted R-squared:  0.07351
## F-statistic: 4.888 on 1 and 48 DF,  p-value: 0.03185
```

1. **do some exploratory analysis using a full plots**
2. **identify the major correlations using the full plots**
3. **use ggplot to plot the data with the linear lines**