Fuel Example: Multiple Regression and Outliers

In this observational study (from Weisberg), we consider data for 48 states relating various state road and income variables (predictors) to per capital fuel consumption (response). Of particular interest is the relationship between fuel tax (predictor) and fuel consumption (response). After model fitting, one outlier (WY) is seen in the diagnostic plots.

Multiple Regression and Diagnostic Plots

```
library(dplyr)
library(car)
#In original file, State names are in first column.
#Using row.names will help identify states in the diagnostic plots below.
FuelData <- read.csv("~/Dropbox/STAT512/Lectures/MultReg1/MR1_Fuel.csv", row.names = 1)
str(FuelData)
```

```
## 'data.frame':    48 obs. of  8 variables:
##  $ pop  : int  1029 771 462 5787 968 3082 18366 7367 11926 10783 ...
##  $ tax  : num  9 9 9 7.5 8 10 8 8 8 7 ...
##  $ nlic : int  540 441 268 3060 527 1760 8278 4074 6312 5948 ...
##  $ inc  : num  3.57 4.09 3.87 4.87 4.4 ...
##  $ road : num  1.976 1.25 1.586 2.351 0.431 ...
##  $ fuelc: int  557 404 259 2396 397 1408 6312 3439 5528 5375 ...
##  $ dlic : num  52.5 57.2 58 52.9 54.4 57.1 45.1 55.3 52.9 55.2 ...
##  $ fuel : int  541 524 561 414 410 457 344 467 464 498 ...
```

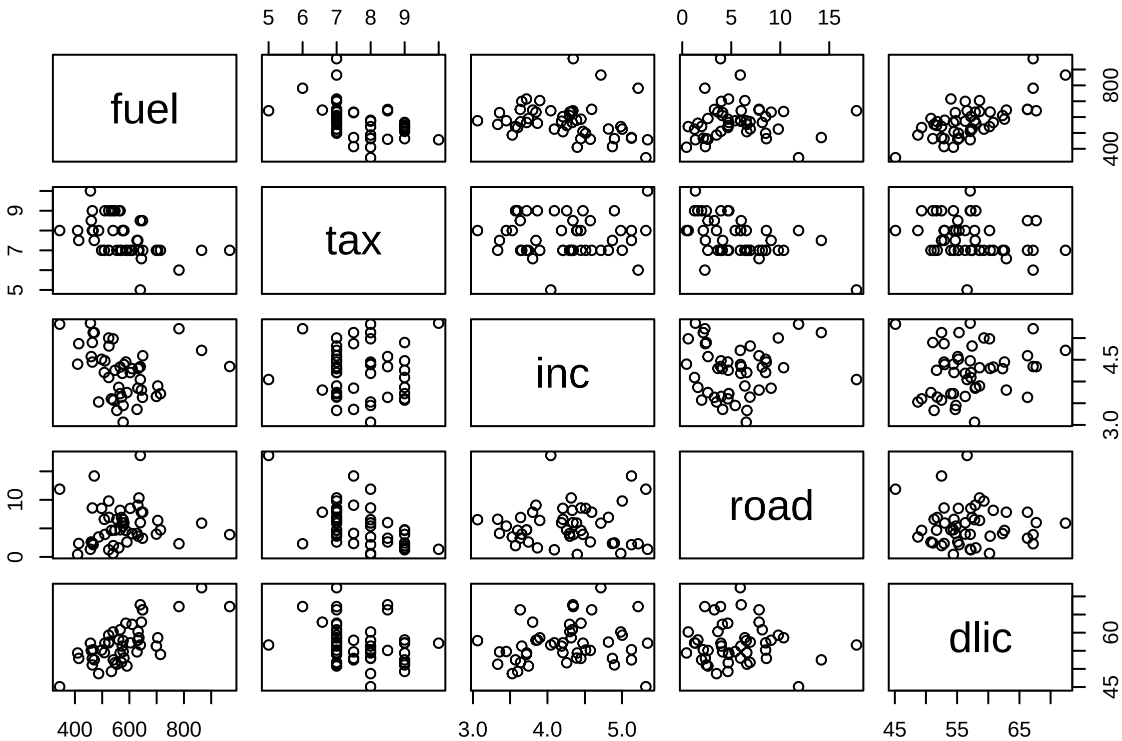
```
#Select columms of interest using select from dplyr.
FuelClean <- select(FuelData, fuel, tax, inc, road, dlic)
str(FuelClean)
```

```
## 'data.frame':    48 obs. of  5 variables:
##  $ fuel: int  541 524 561 414 410 457 344 467 464 498 ...
##  $ tax : num  9 9 9 7.5 8 10 8 8 8 7 ...
##  $ inc : num  3.57 4.09 3.87 4.87 4.4 ...
##  $ road: num  1.976 1.25 1.586 2.351 0.431 ...
##  $ dlic: num  52.5 57.2 58 52.9 54.4 57.1 45.1 55.3 52.9 55.2 ...
```

```
cor(FuelClean)
```

```
##             fuel         tax         inc        road       dlic
## fuel  1.00000000 -0.45128028 -0.24486207  0.01904194  0.6989654
## tax  -0.45128028  1.00000000  0.01266516 -0.52213014 -0.2880372
## inc  -0.24486207  0.01266516  1.00000000  0.05016279  0.1570701
## road  0.01904194 -0.52213014  0.05016279  1.00000000 -0.0641295
## dlic  0.69896542 -0.28803717  0.15707008 -0.06412950  1.0000000
```

```r
pairs(FuelClean)
```




```r
Model1 <- lm(fuel ~ tax, data = FuelData)
summary(Model1)
```

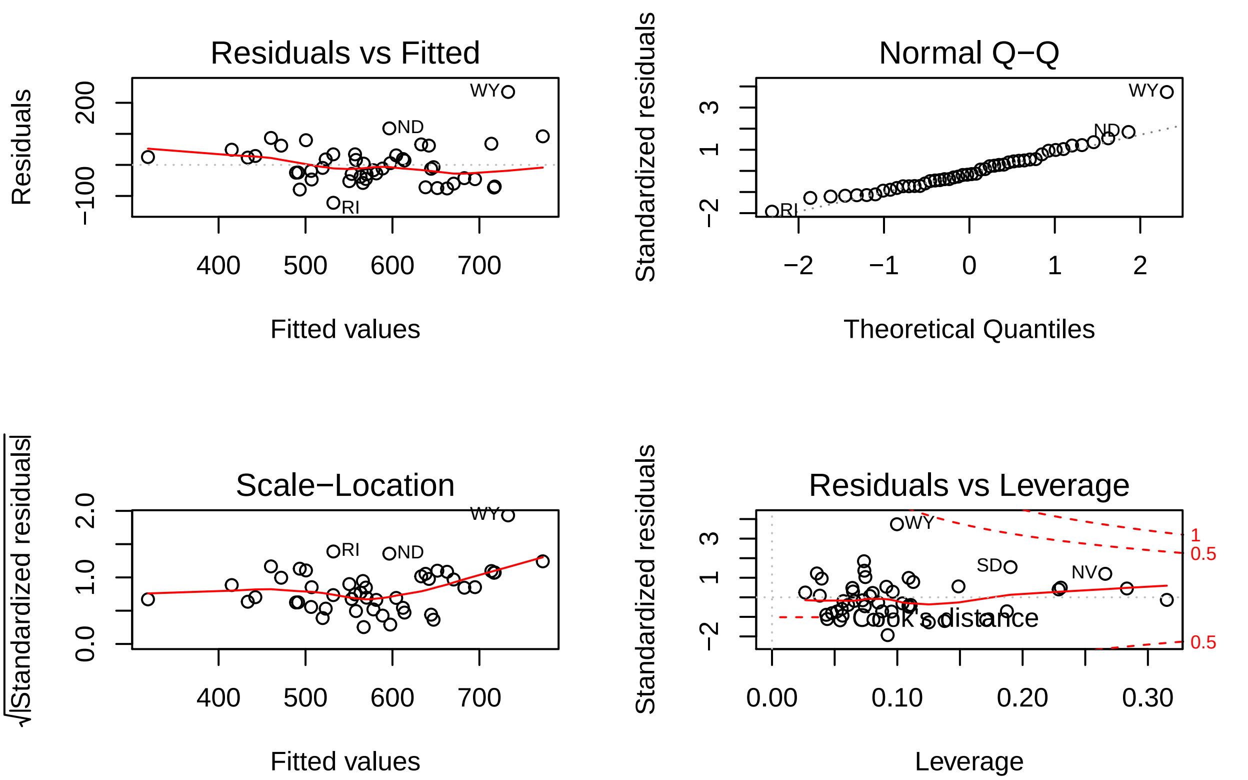
```
##
## Call:
## lm(formula = fuel ~ tax, data = FuelData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -215.16  -72.27    6.74   41.28  355.74
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   984.01     119.62   8.226 1.38e-10 ***
## tax           -53.11      15.48  -3.430  0.00128 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 100.9 on 46 degrees of freedom
## Multiple R-squared:  0.2037, Adjusted R-squared:  0.1863
## F-statistic: 11.76 on 1 and 46 DF,  p-value: 0.001285
```

```r
Model2 <- lm(fuel ~ tax + inc + road + dlic, data = FuelData)
summary(Model2)
```

```
##
## Call:
## lm(formula = fuel ~ tax + inc + road + dlic, data = FuelData)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -122.03  -45.57  -10.66   31.53  234.95
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  377.291    185.541   2.033 0.048207 *
## tax          -34.790     12.970  -2.682 0.010332 *
## inc          -66.589     17.222  -3.867 0.000368 ***
## road          -2.426      3.389  -0.716 0.477999
## dlic          13.364      1.923   6.950 1.52e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 66.31 on 43 degrees of freedom
## Multiple R-squared:  0.6787, Adjusted R-squared:  0.6488
## F-statistic: 22.71 on 4 and 43 DF,  p-value: 3.907e-10
```

```r
par(mfrow = c(2, 2))
plot(Model2)
```



Additional Diagnostics

The outlierTest() function from the car package runs an outlier test based on the Rstudent residual. Note that p-values are calculated with and without Bonferonni adjustment.

```r
outlierTest(Model2)
```

```
##    rstudent unadjusted p-value Bonferonni p
```

```
## WY 4.490051          5.4704e-05    0.0026258
```

```r
Temp <- data.frame(State = row.names(FuelData),
                   Fuel = FuelData$fuel,
                   Pred = fitted(Model2),
                   ResidRaw = resid(Model2),
                   ResidStd = rstandard(Model2),
                   RStudent = rstudent(Model2))
#Reorder data by Rstudent using arrange from dplyr.
Temp <- arrange(Temp, desc(abs(RStudent)))
head(Temp)
```

```
##   State Fuel     Pred   ResidRaw  ResidStd  RStudent
## 1    WY  968 733.0528  234.94715  3.734462  4.490051
## 2    RI  410 532.0289 -122.02893 -1.931710 -1.997765
## 3    ND  714 596.4035  117.59655  1.842463  1.897347
## 4    SD  865 772.9678   92.03225  1.542568  1.568543
## 5    VA  547 460.2215   86.77850  1.359823  1.373781
## 6    MA  414 493.3563  -79.35625 -1.279561 -1.289381
```

```r
#Calculate Bonferonni adjusted p-value "By Hand".
#Matched outlierTest output.
2*48*(1-pt(4.490, 48-4-2))
```

```
## [1] 0.002626228
```