Linear Regression Mini Project

Anthony Stachowitz

May 30, 2018

Linear Regression Mini Project

Exercise 1: least squares regression

Use the /states.rds/ data set. Fit a model predicting energy consumed per capita (energy) from the percentage of residents living in metropolitan areas (metro). Be sure to

1. Examine/plot the data before fitting the model
2. Print and interpret the model `summary'
3. `plot' the model to look for deviations from modeling assumptions

Select one or more additional predictors to add to your model and repeat steps 1-3. Is this model significantly better than the model with /metro/ as the only predictor?

Examine/plot the data before fitting the model

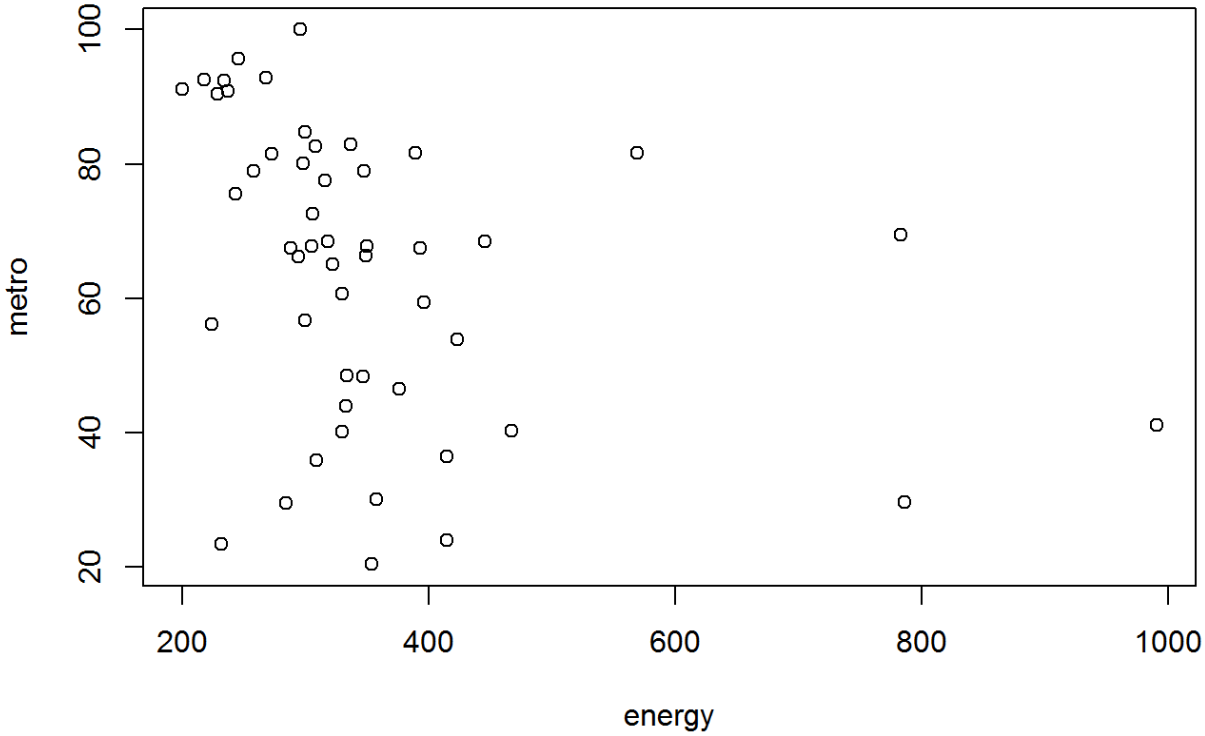
```
states.data <- readRDS("dataSets/states.rds")
states.info <- data.frame(attributes(states.data)[c("names", "var.labels")])
tail(states.info, 8)
```

```
##      names                      var.labels
## 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
```

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

```
##      energy          metro
##  Min.   :200.0   Min.   : 20.40
##  1st Qu.:285.0   1st Qu.: 46.98
##  Median :320.0   Median : 67.55
##  Mean   :354.5   Mean   : 64.07
##  3rd Qu.:371.5   3rd Qu.: 81.58
##  Max.   :991.0   Max.   :100.00
##  NA's   :1       NA's   :1
```

```
plot(sts.eng.mtr)
```



Print and interpret the model `summary’

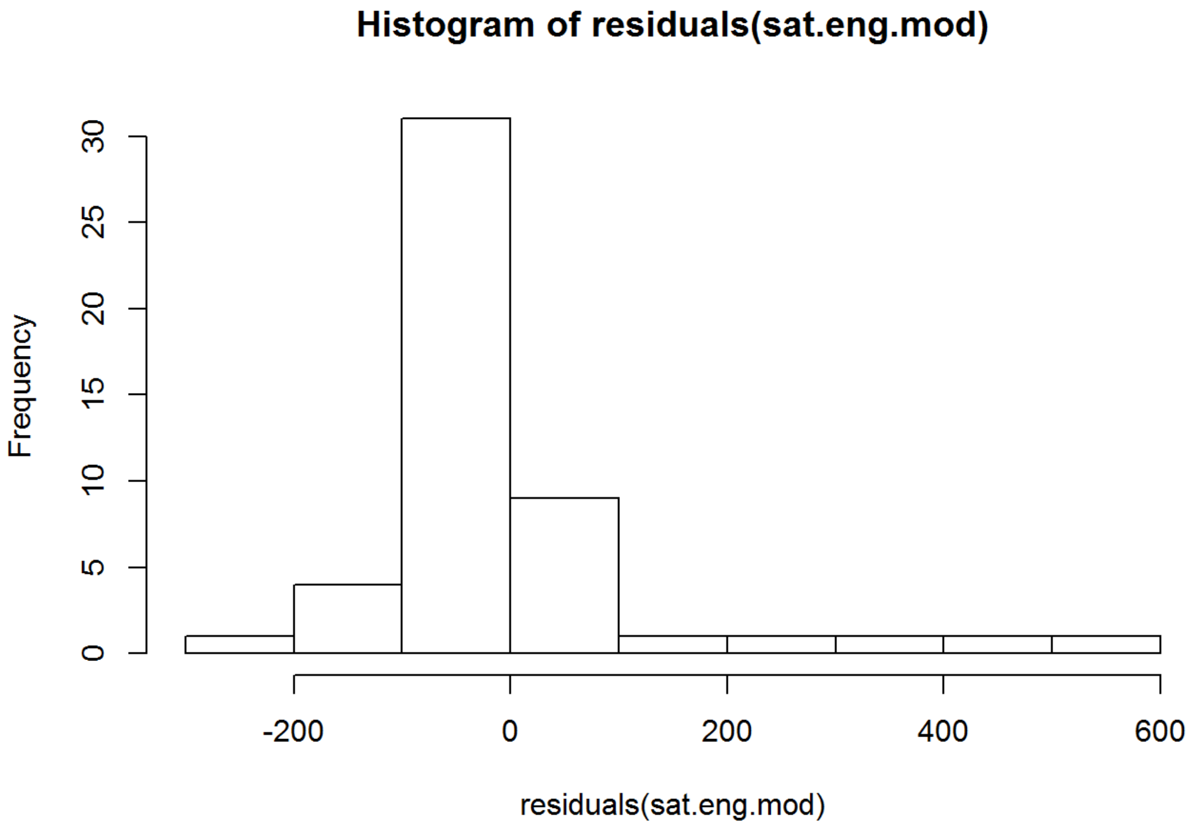
```
sat.eng.mod <- lm(energy ~ metro,
              data=states.data)
summary(sat.eng.mod)
```

```
##
## Call:
## lm(formula = energy ~ metro, data = states.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -215.51  -64.54  -30.87   18.71  583.97
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 501.0292    61.8136   8.105 1.53e-10 ***
## metro        -2.2871     0.9139  -2.503   0.0158 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 140.2 on 48 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.1154, Adjusted R-squared:  0.097
## F-statistic: 6.263 on 1 and 48 DF,  p-value: 0.01578
```

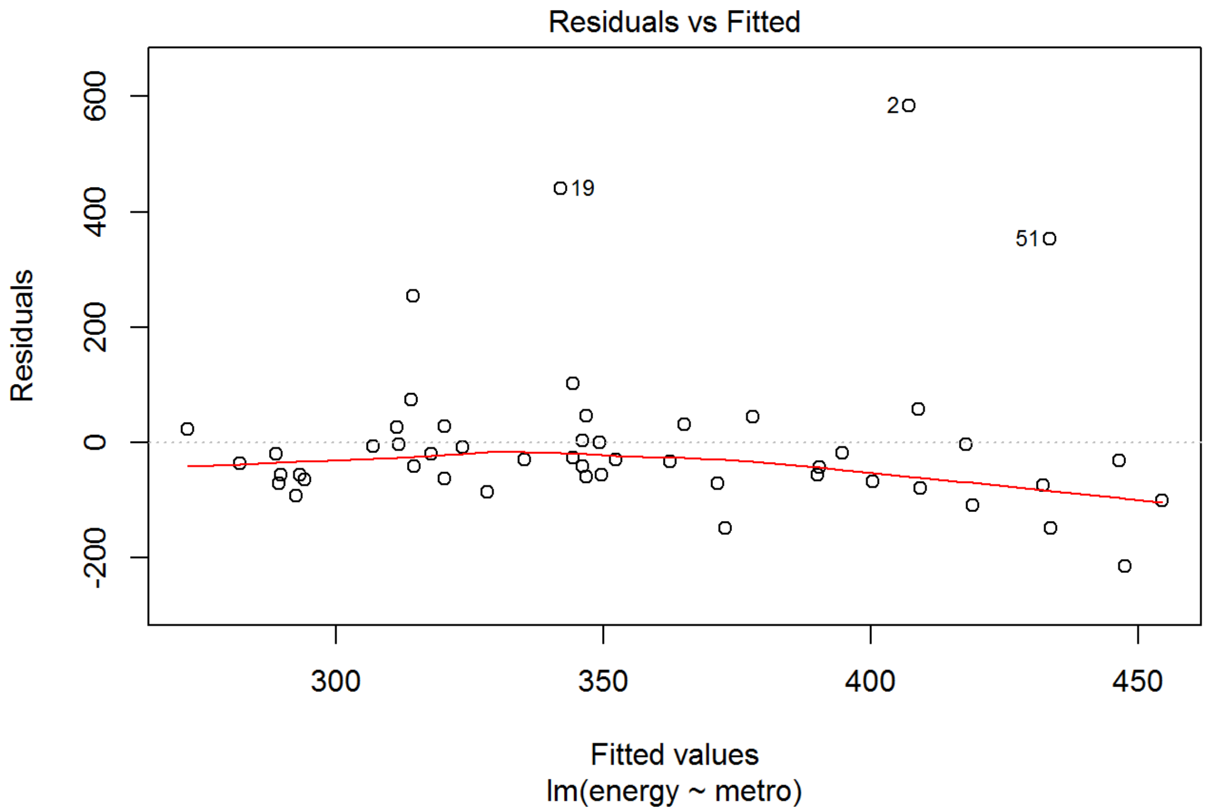
The R-squared is low at 0.1154. this does not look to be a very good regregression model with this data alone. There also seems to be some data points that skew the data a bit

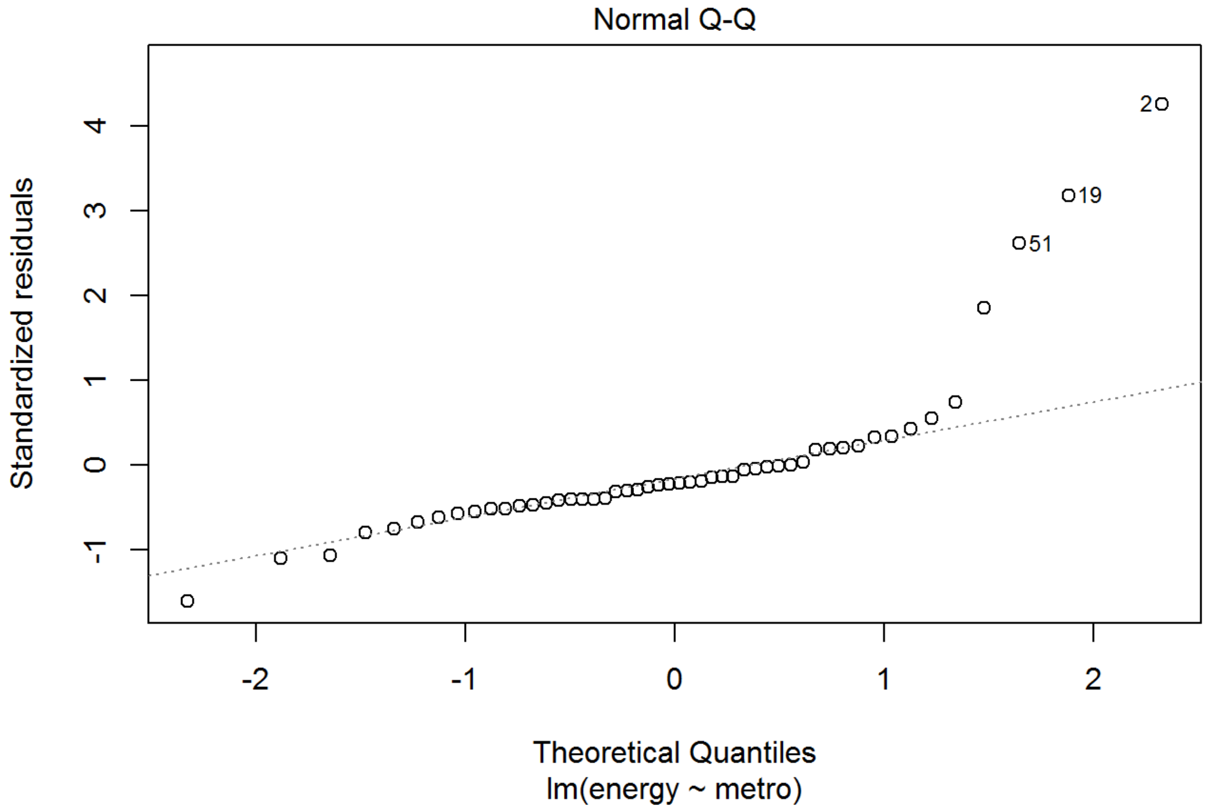
`plot’ the model to look for deviations from modeling assumptions

```
hist(residuals(sat.eng.mod))
```

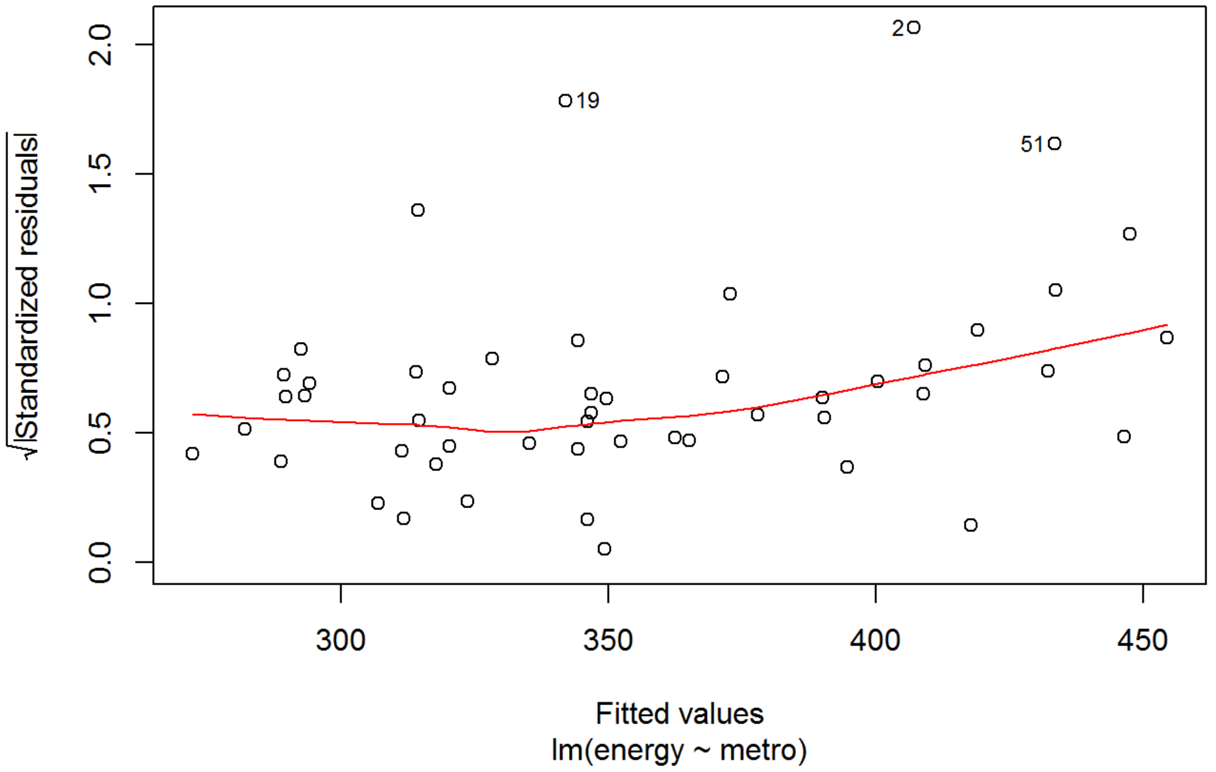


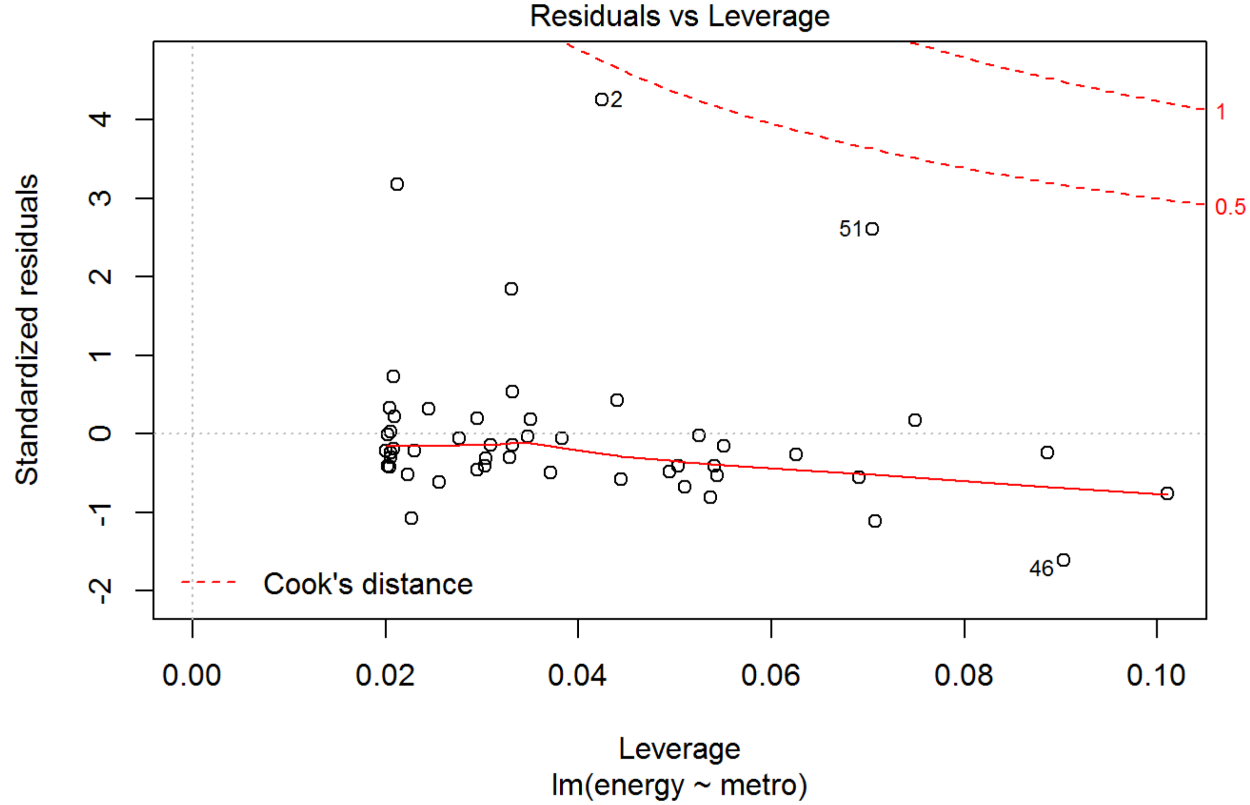
```
plot(sat.eng.mod)
```





Scale-Location





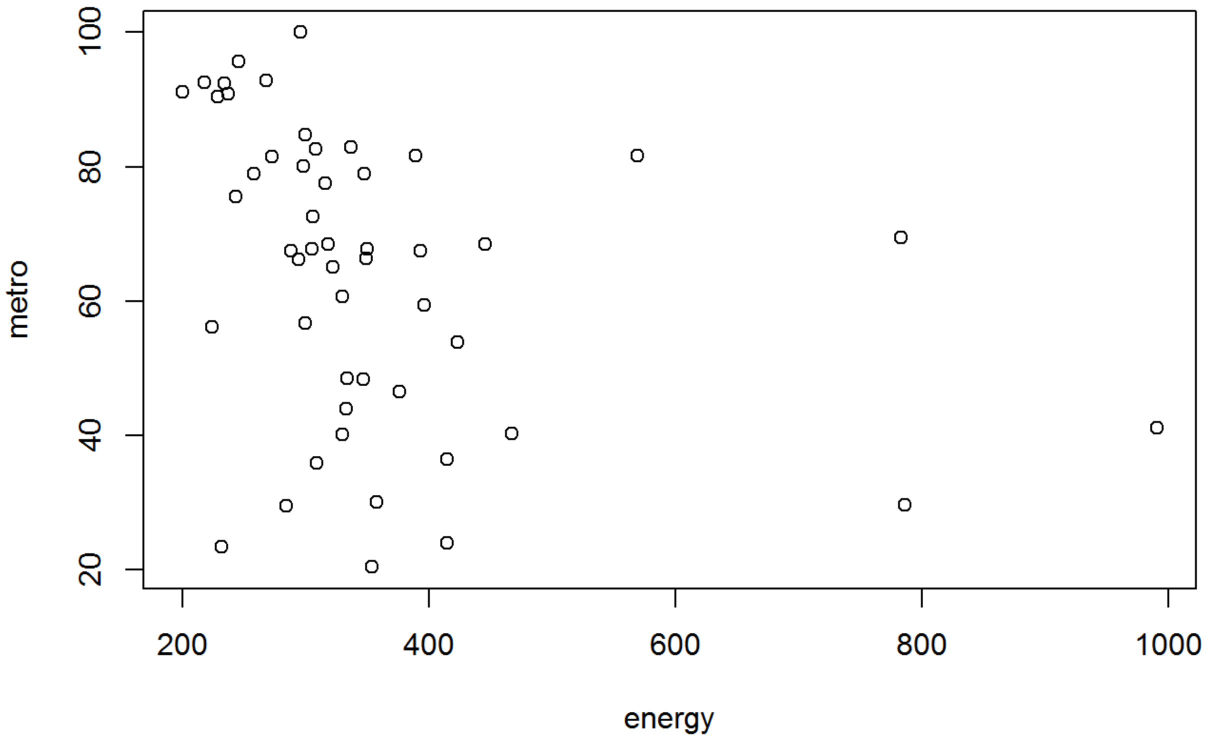
With additional predictors

Examine/plot the data before fitting the model

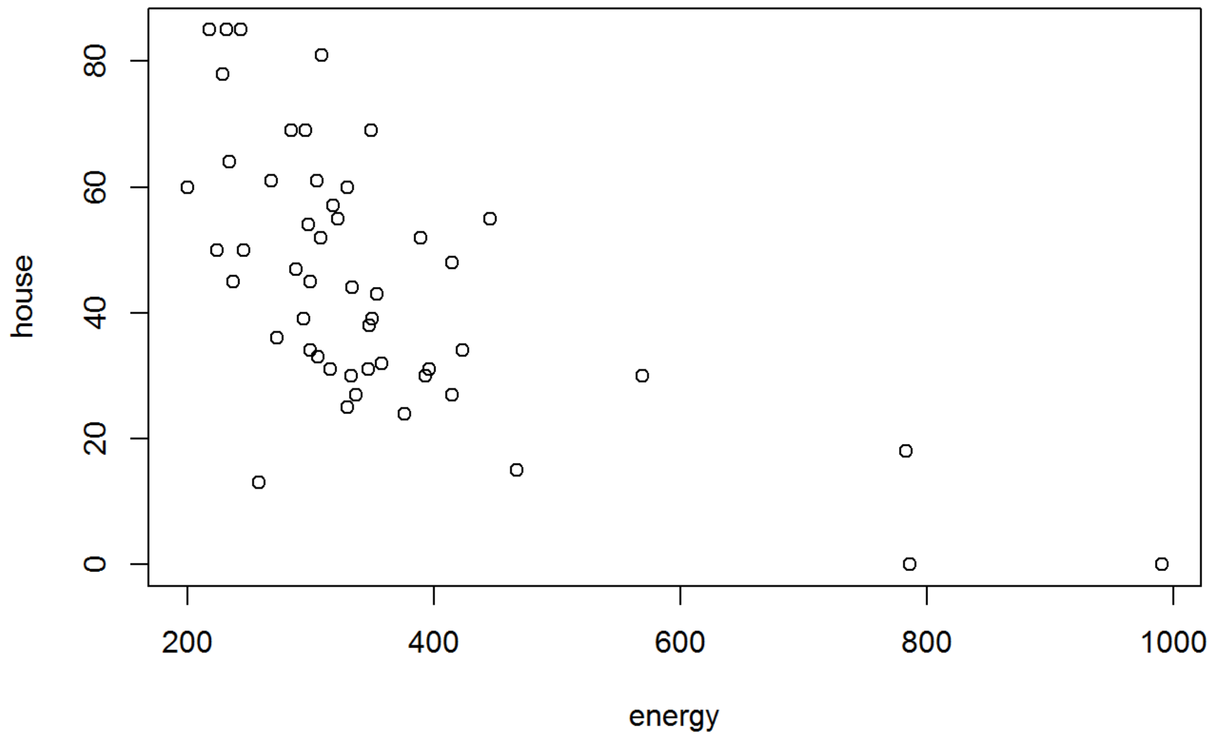
```
sts.eng.hse <- subset(states.data, select = c("energy", "house"))
sts.eng.snt <- subset(states.data, select = c("energy", "senate"))
sts.eng.total <- subset(states.data, select = c("energy", "metro", "house", "senate"))
summary(sts.eng.total)
```

```
##      energy          metro            house           senate     
##  Min.   :200.0   Min.   : 20.40   Min.   : 0.00   Min.   :10.00  
##  1st Qu.:285.0   1st Qu.: 46.98   1st Qu.:31.00   1st Qu.:27.00  
##  Median :320.0   Median : 67.55   Median :44.50   Median :51.00  
##  Mean   :354.5   Mean   : 64.07   Mean   :44.82   Mean   :49.78  
##  3rd Qu.:371.5   3rd Qu.: 81.58   3rd Qu.:59.25   3rd Qu.:67.00  
##  Max.   :991.0   Max.   :100.00   Max.   :85.00   Max.   :97.00  
##  NA's   :1       NA's   :1        NA's   :1       NA's   :1
```

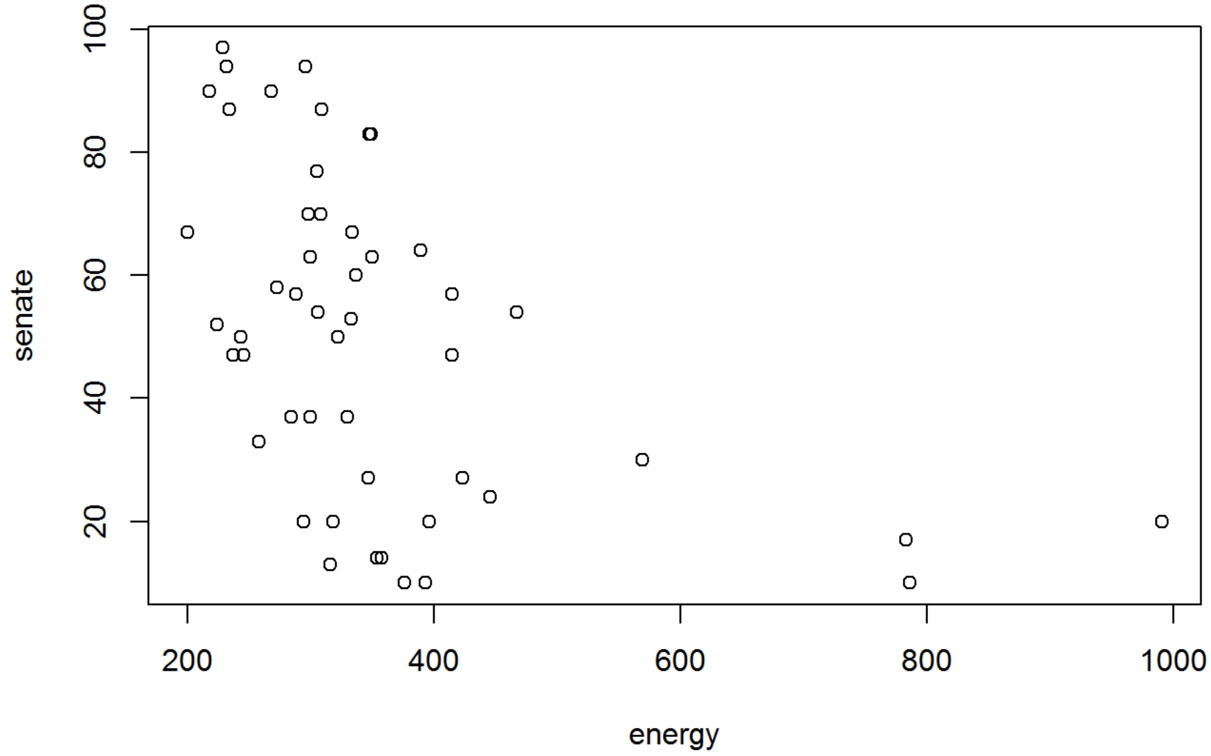
```
plot(sts.eng.mtr)
```



```
plot(sts.eng.hse)
```



```
plot(sts.eng.snt)
```



Print and interpret the model `summary’

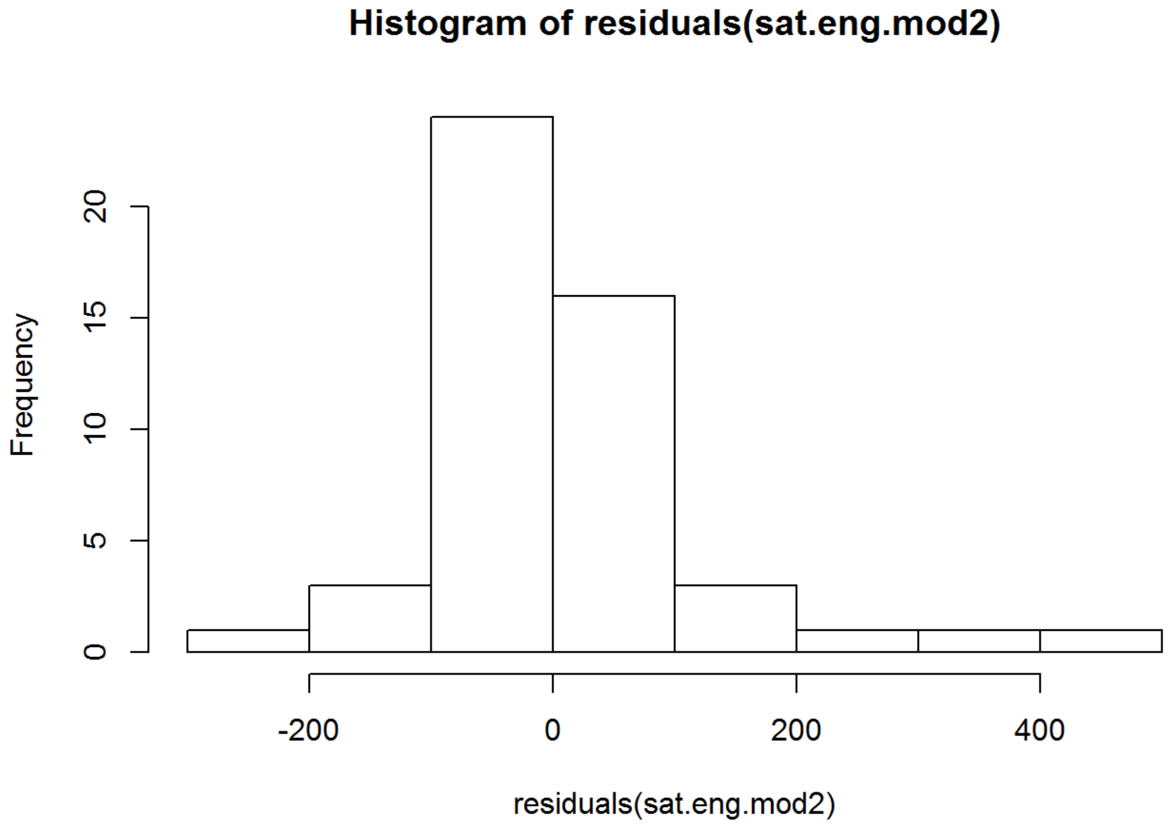
```
sat.eng.mod2 <- lm(energy ~ metro + house + senate,
              data=states.data)
summary(sat.eng.mod2)
```

```
##
## Call:
## lm(formula = energy ~ metro + house + senate, data = states.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -209.88  -69.43  -19.06   39.04  423.60
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 620.7157    55.6715  11.150 1.14e-14 ***
## metro        -1.1735     0.8085  -1.451 0.153461
## house        -3.9799     1.0407  -3.824 0.000393 ***
## senate       -0.2541     0.8595  -0.296 0.768797
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 114.4 on 46 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.4356, Adjusted R-squared:  0.3988
## F-statistic: 11.84 on 3 and 46 DF,  p-value: 7.197e-06
```

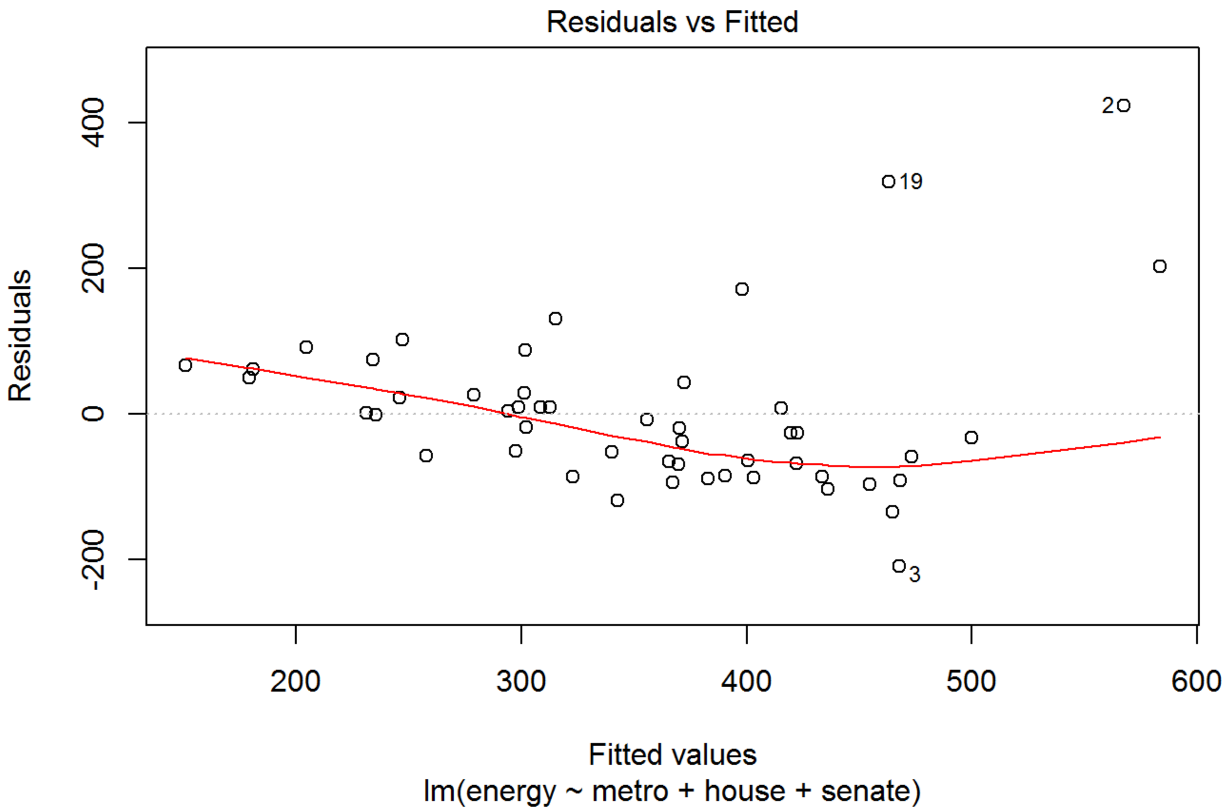
The R-squared is 0.4356. this is a little bit better but we should probably continue to work on it. There still seems to be some data points that skew the data

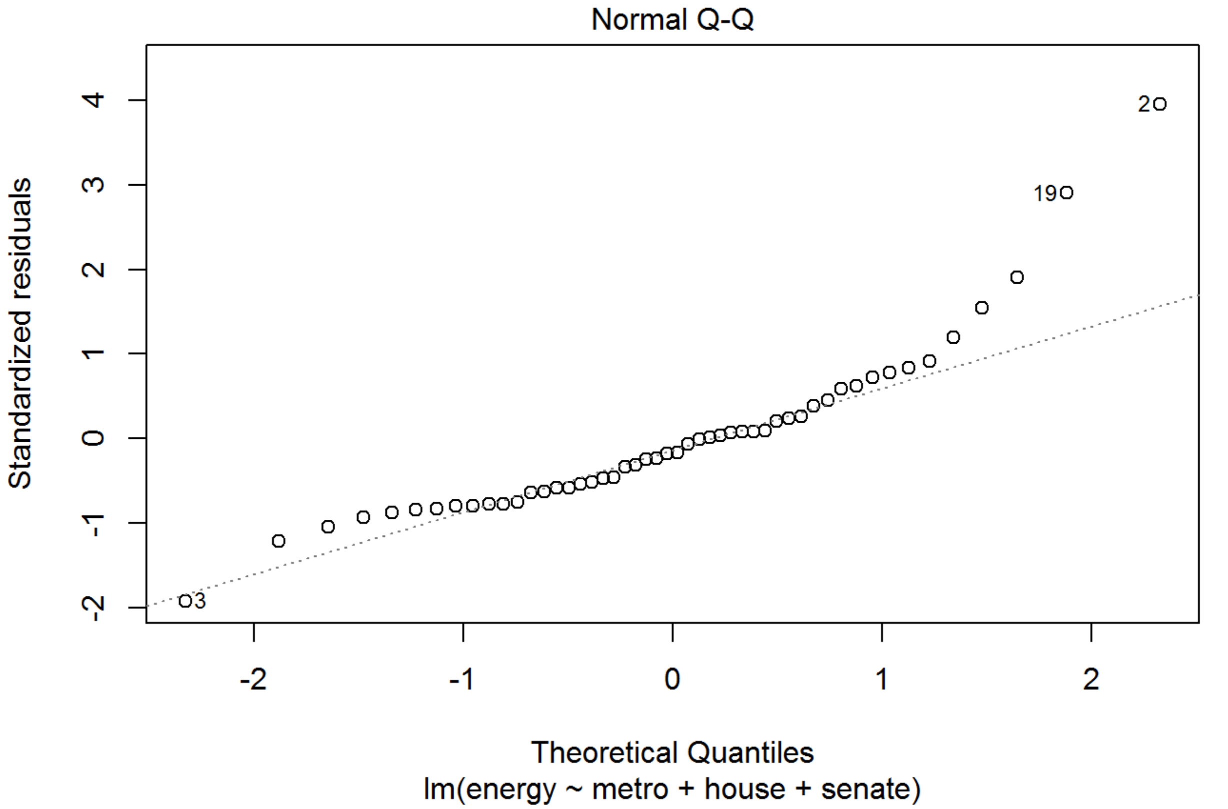
`plot’ the model to look for deviations from modeling assumptions

```
hist(residuals(sat.eng.mod2))
```

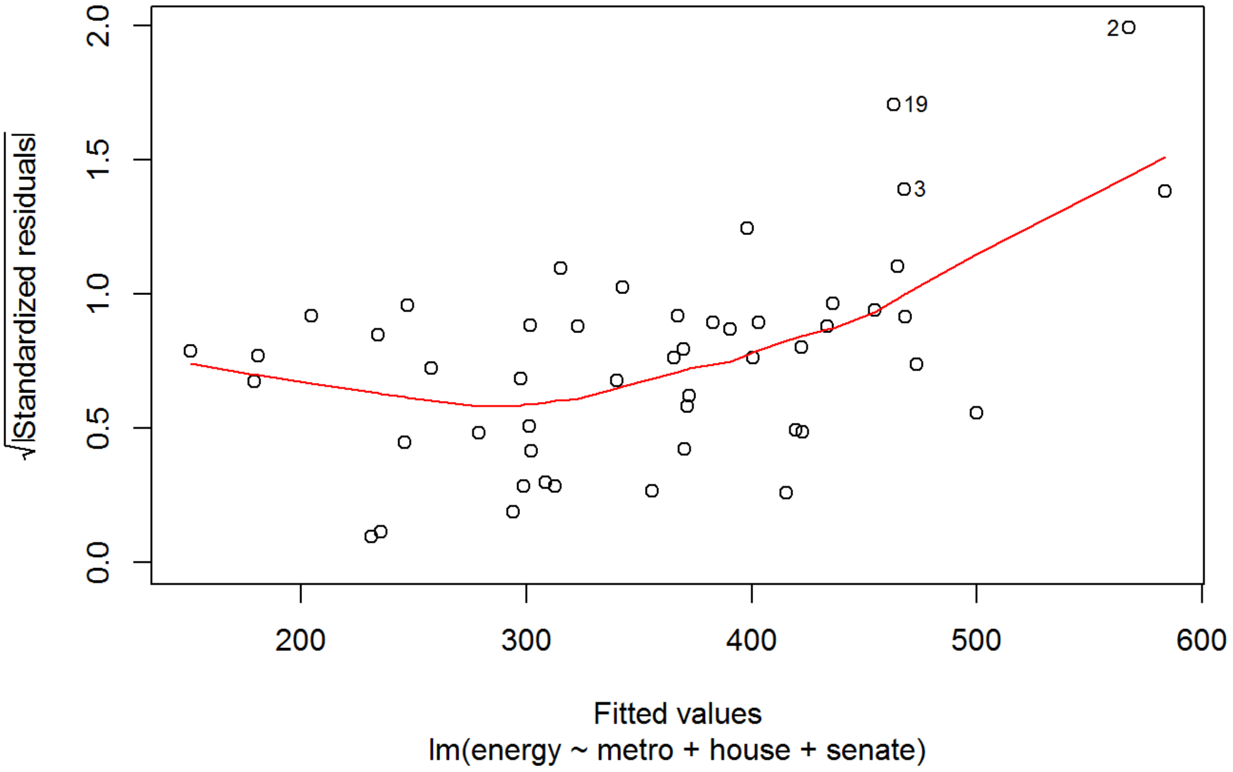


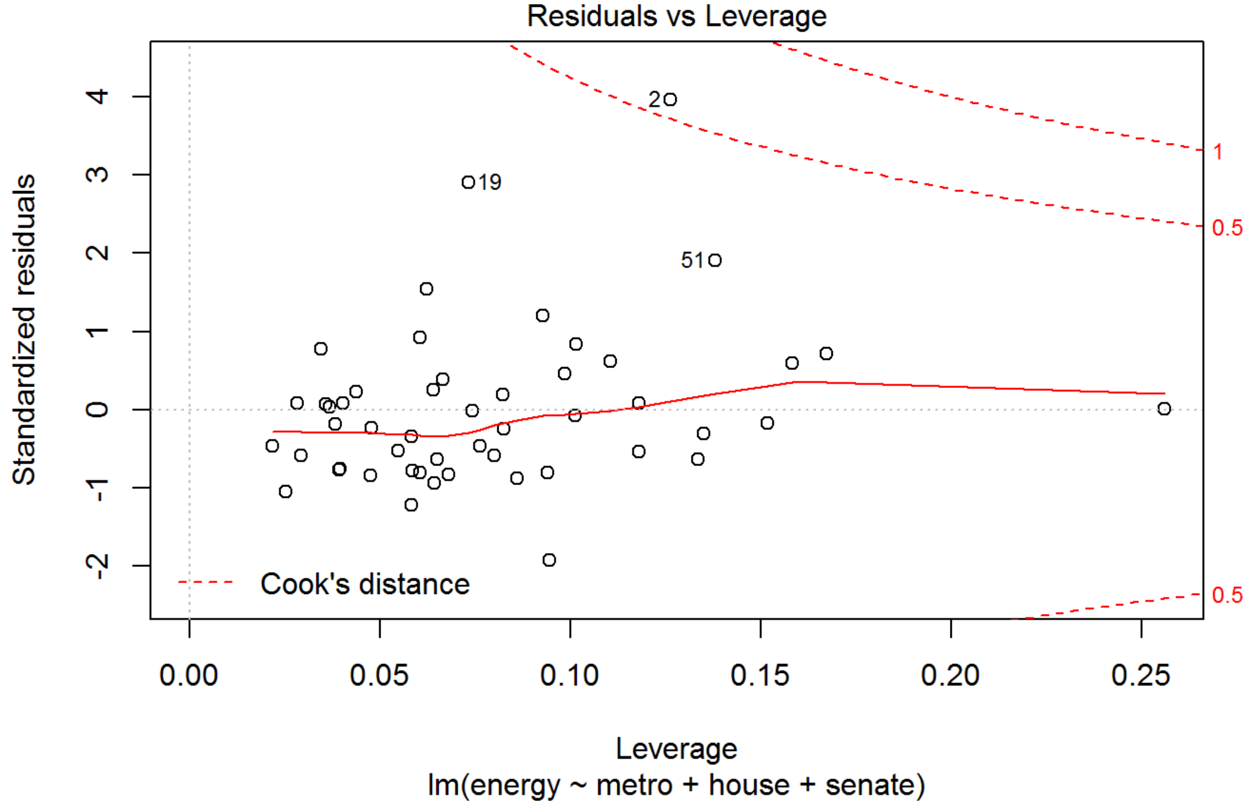
```
plot(sat.eng.mod2)
```





Scale-Location





Exercise 2: Interactions and factors

Use the states data set.

1. Add on to the regression equation that you created in exercise 1 by generating an interaction term and testing the interaction.

2. Try adding region to the model. Are there significant differences across the four regions?

Add on to the regression equation that you created in exercise 1 by generating an interaction term and testing the interaction.

```
sat.eng.mod2a <- lm(energy ~ metro + house*senate,
              data=states.data)
summary(sat.eng.mod2a)
```

```
##
## Call:
## lm(formula = energy ~ metro + house * senate, data = states.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -242.84  -54.50  -15.07   48.79  324.19
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  798.28128   71.08117  11.231 1.21e-14 ***
## metro         -0.74524    0.73469  -1.014  0.31584
## house         -8.66614    1.62959  -5.318 3.17e-06 ***
## senate        -5.04231    1.56768  -3.216  0.00241 **
## house:senate   0.09438    0.02692   3.507  0.00104 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 102.5 on 45 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.5567, Adjusted R-squared:  0.5173
## F-statistic: 14.13 on 4 and 45 DF,  p-value: 1.517e-07
```

The regression equation did improve to 0.5567 by making the house and energy dependant on the senate

Add region to the model. Are there differences across the four regions?

```
str(states.data$region)
```

```
##  Factor w/ 4 levels "West","N. East",..: 3 1 1 3 1 1 2 3 NA 3 ...
```

```
states.data$region <- factor(states.data$region)
sat.eng.mod2b <- lm(energy ~ metro + region + house*senate,
              data=states.data)
summary(sat.eng.mod2b)
```

```
## 
## Call:
## lm(formula = energy ~ metro + region + house * senate, data = states.data)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -243.29  -44.11   -6.97   38.34  316.97 
## 
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)    
## (Intercept)   818.01721   74.12523  11.036 5.46e-14 ***
## metro          -0.57826    0.75054  -0.770 0.445346    
## regionN. East -82.72805   60.29178  -1.372 0.177308    
## regionSouth    -1.45712   38.59053  -0.038 0.970059    
## regionMidwest  18.75761   44.47635   0.422 0.675363    
## house          -9.61870    1.73922  -5.530 1.88e-06 ***
## senate         -6.01098    1.72716  -3.480 0.001181 ** 
## house:senate    0.12205    0.03147   3.878 0.000364 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 102.4 on 42 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.5865, Adjusted R-squared:  0.5176 
## F-statistic: 8.512 on 7 and 42 DF,  p-value: 1.867e-06
```

There does not seem to be significant diferences with region added.