Linear Regression - US State Environmental Impact

Chinpei Tang

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?

Load the required library.

```
library(ggplot2)
```

Data Examination

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")
```

Examining 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 mil
##  - 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 _IregXpe
##   ..$ : chr  "_dta" "__xi__Vars__Prefix__" "_I _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
```

Now, we look into the energy consumed per capita (energy) from the percentage of residents living in metropolitan areas (metro).

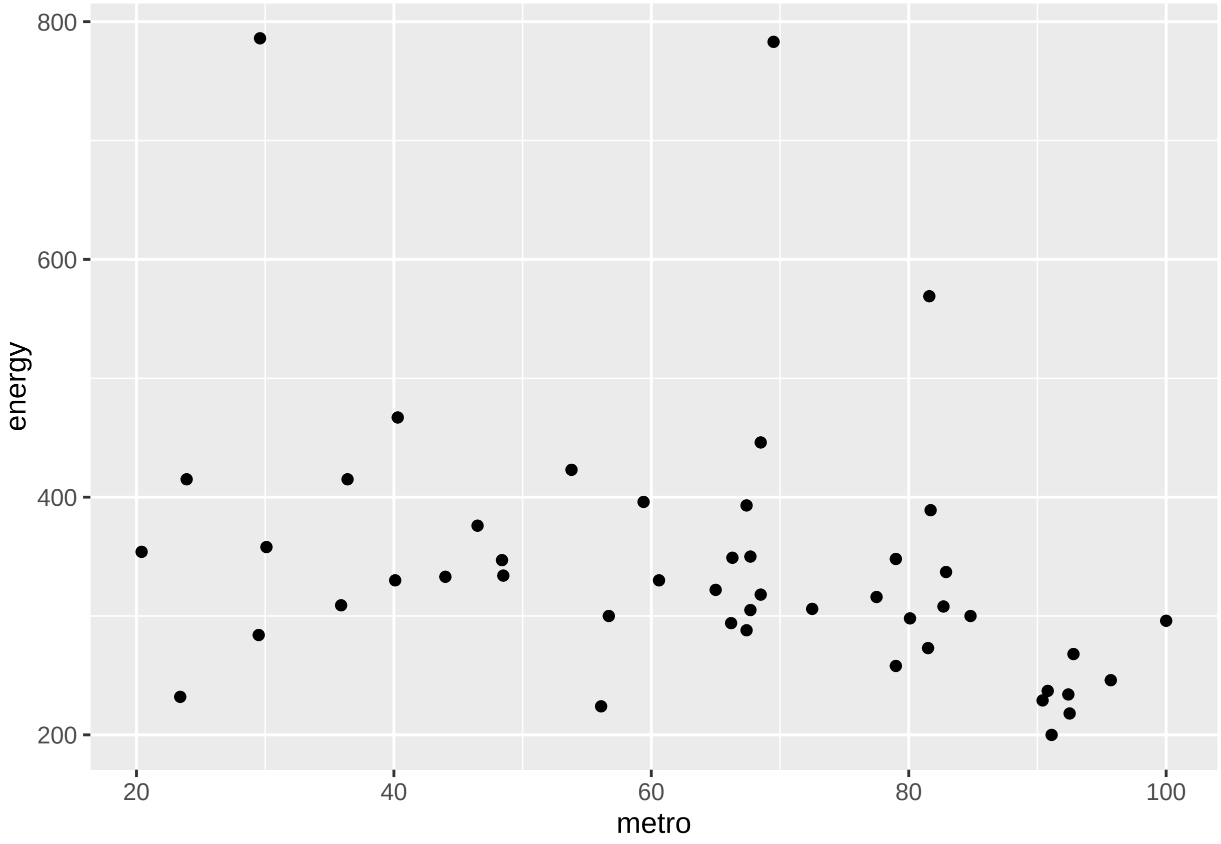
For analysis convenience, gather the states.data with only the numerical values into states.data.num, also removed all the NA's.

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

```
##       pop                area           density            metro
##  Min.   :  454000   Min.   :  1045   Min.   :   4.68   Min.   : 20.40
##  1st Qu.: 1562250   1st Qu.: 38666   1st Qu.:  32.13   1st Qu.: 47.92
##  Median : 3576000   Median : 54156   Median :  75.76   Median : 67.55
##  Mean   : 5134250   Mean   : 61691   Mean   : 169.35   Mean   : 64.31
##  3rd Qu.: 6058750   3rd Qu.: 80169   3rd Qu.: 170.41   3rd Qu.: 81.62
##  Max.   :29760000   Max.   :261914   Max.   :1041.92   Max.   :100.00
##      waste            energy          miles            toxic
##  Min.   :0.5400   Min.   :200.0   Min.   : 5.900   Min.   :  1.810
##  1st Qu.:0.8150   1st Qu.:287.0   1st Qu.: 8.500   1st Qu.:  7.232
##  Median :0.9600   Median :320.0   Median : 9.150   Median : 11.705
##  Mean   :0.9867   Mean   :343.6   Mean   : 9.121   Mean   : 17.544
##  3rd Qu.:1.1350   3rd Qu.:362.5   3rd Qu.: 9.725   3rd Qu.: 21.363
##  Max.   :1.5100   Max.   :786.0   Max.   :12.800   Max.   :101.280
##      green            house           senate           csat
##  Min.   : 11.76   Min.   : 0.00   Min.   :10.00   Min.   : 832.0
##  1st Qu.: 16.98   1st Qu.:31.00   1st Qu.:27.00   1st Qu.: 890.0
##  Median : 21.38   Median :44.50   Median :52.50   Median : 939.0
##  Mean   : 25.11   Mean   :44.92   Mean   :50.40   Mean   : 948.0
##  3rd Qu.: 26.34   3rd Qu.:57.75   3rd Qu.:67.75   3rd Qu.: 998.2
##  Max.   :114.40   Max.   :85.00   Max.   :97.00   Max.   :1093.0
##       vsat            msat          percent         expense
##  Min.   :395.0   Min.   :437.0   Min.   : 4.00   Min.   :2960
##  1st Qu.:423.2   1st Qu.:467.5   1st Qu.:11.00   1st Qu.:4340  
##  Median :446.0   Median :493.0   Median :23.50   Median :4920  
##  Mean   :449.8   Mean   :498.2   Mean   :34.52   Mean   :5089  
##  3rd Qu.:476.0   3rd Qu.:522.2   3rd Qu.:60.25   3rd Qu.:5693  
##  Max.   :515.0   Max.   :578.0   Max.   :81.00   Max.   :8645  
##      income           high          college     
##  Min.   :23.46   Min.   :64.30   Min.   :12.30  
##  1st Qu.:29.30   1st Qu.:73.45   1st Qu.:17.15  
##  Median :32.28   Median :76.70   Median :18.85  
##  Mean   :33.38   Mean   :76.03   Mean   :19.62  
##  3rd Qu.:36.20   3rd Qu.:80.03   3rd Qu.:21.92  
##  Max.   :48.62   Max.   :85.10   Max.   :27.20  
```

Look at the scattered plot and correlation between metro and energy data.

```
energy.metro.baseplot <- ggplot(states.data.num, aes(x = metro, y = energy)) + geom_point(size = 1.5)
energy.metro.baseplot
```



```
cor(subset(states.data.num, select = c("metro", "energy")))
```

```
##             metro     energy
## metro   1.0000000 -0.3116753
## energy -0.3116753  1.0000000
```

It can be noticed that there are not too much correlations.

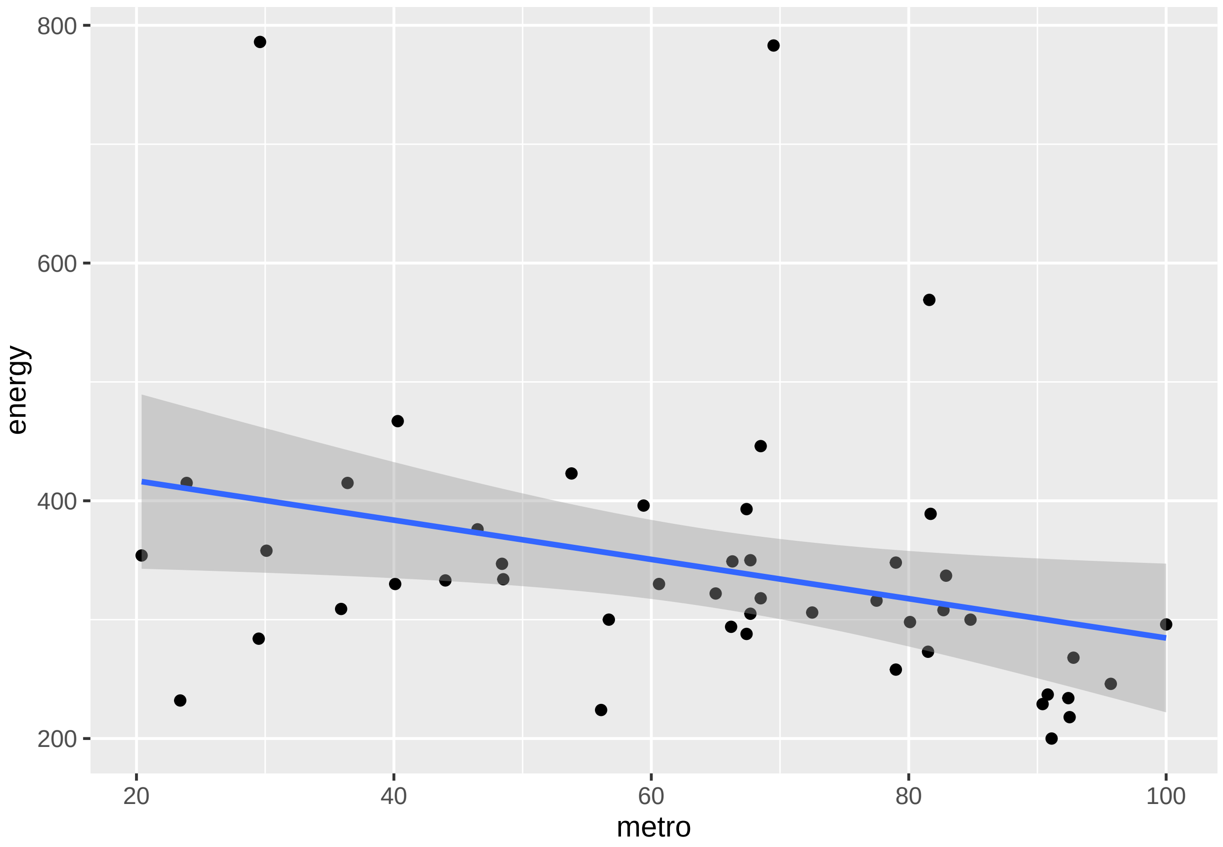
Linear Regression

Perform a linear regression between energy and metro:

```r
summary(lm(energy~metro, data = states.data.num))
```

```
##
## Call:
## lm(formula = energy ~ metro, data = states.data.num)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -179.17  -54.21  -21.64   15.07  448.02
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 449.8382    50.4472   8.917 1.37e-11 ***
## metro        -1.6526     0.7428  -2.225    0.031 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 112.3 on 46 degrees of freedom
## Multiple R-squared:  0.09714,    Adjusted R-squared:  0.07751
## F-statistic: 4.949 on 1 and 46 DF,  p-value: 0.03105
```

```r
energy.metro.baseplot + stat_smooth(method = lm)
```



Now, look the correlation between energy and all the other variables.

```r
summary(lm(energy ~ ., data = states.data.num))
```

```
##
## Call:
## lm(formula = energy ~ ., data = states.data.num)
##
## 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
```

Progressively removing the variable that has low correlations:

```r
summary(lm(energy ~ . - msat,
           data = states.data.num))
```

```
##
## Call:
## lm(formula = energy ~ . - msat, data = states.data.num)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -126.554  -27.297    0.968   21.840  159.899
##
```

```
## Coefficients:
##               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    
## 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
```

```r
summary(lm(energy ~ . - msat - csat,
           data = states.data.num))
```

```
## 
## Call:
## lm(formula = energy ~ . - msat - csat, data = states.data.num)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -125.408  -26.758    0.531   21.956  159.236 
## 
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -4.718e+01  4.998e+02  -0.094   0.9254    
## pop         -1.870e-06  3.133e-06  -0.597   0.5550    
## area         6.596e-04  3.590e-04   1.838   0.0757 .  
## density     -1.197e-02  8.703e-02  -0.138   0.8915    
## metro        6.107e-01  9.226e-01   0.662   0.5129    
## waste        1.410e+01  5.249e+01   0.269   0.7901    
## miles        1.213e+01  1.465e+01   0.828   0.4139    
## toxic        2.784e+00  6.130e-01   4.541 7.96e-05 ***
## green        4.405e+00  9.154e-01   4.812 3.68e-05 ***
## house        8.952e-02  9.586e-01   0.093   0.9262    
## senate       1.303e-01  6.267e-01   0.208   0.8366    
## vsat         3.466e-01  9.358e-01   0.370   0.7137    
## percent      9.070e-01  1.385e+00   0.655   0.5175    
## expense      1.221e-02  1.517e-02   0.805   0.4268    
## income       9.456e-01  4.590e+00   0.206   0.8381
## high        -1.640e+00  3.443e+00  -0.476   0.6373
## college     -6.548e+00  5.675e+00  -1.154   0.2574
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 62.11 on 31 degrees of freedom
## Multiple R-squared:  0.814,  Adjusted R-squared:  0.718
## F-statistic: 8.479 on 16 and 31 DF,  p-value: 2.551e-07
```

```
summary(lm(energy ~ . - msat - csat - house,
           data = states.data.num))
```

```
##
## Call:
## lm(formula = energy ~ . - msat - csat - house, data = states.data.num)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -124.337  -26.903    0.999   22.001  159.290
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.396e+01  4.908e+02  -0.090   0.9292
## pop         -1.843e-06  3.071e-06  -0.600   0.5527
## area         6.515e-04  3.429e-04   1.900   0.0665 .
## density     -9.338e-03  8.106e-02  -0.115   0.9090
## metro        5.834e-01  8.612e-01   0.677   0.5030
## waste        1.529e+01  5.011e+01   0.305   0.7622
## miles        1.239e+01  1.416e+01   0.875   0.3881
## toxic        2.783e+00  6.034e-01   4.612 6.11e-05 ***
## green        4.360e+00  7.637e-01   5.709 2.53e-06 ***
## senate       1.514e-01  5.754e-01   0.263   0.7941
## vsat         3.343e-01  9.121e-01   0.367   0.7164
## percent      9.077e-01  1.364e+00   0.666   0.5105
## expense      1.224e-02  1.493e-02   0.820   0.4182
## income       8.993e-01  4.492e+00   0.200   0.8426
## high        -1.560e+00  3.284e+00  -0.475   0.6380
## college     -6.554e+00  5.586e+00  -1.173   0.2494
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 61.14 on 32 degrees of freedom
## Multiple R-squared:  0.814,  Adjusted R-squared:  0.7267
## F-statistic: 9.333 on 15 and 32 DF,  p-value: 8.099e-08
```

```
summary(lm(energy ~ . - msat - csat - house - density,
           data = states.data.num))
```

```
##
## Call:
## lm(formula = energy ~ . - msat - csat - house - density, data = states.data.num)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -124.852  -27.320    0.665   22.770  159.910
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.811e+01  4.821e+02  -0.100   0.9211
## pop         -1.726e-06  2.854e-06  -0.605   0.5495
## area         6.491e-04  3.371e-04   1.925   0.0628 .
## metro        5.663e-01  8.356e-01   0.678   0.5027
## waste        1.452e+01  4.891e+01   0.297   0.7684
## miles        1.281e+01  1.348e+01   0.951   0.3487
## toxic        2.779e+00  5.933e-01   4.684 4.67e-05 ***
## green        4.344e+00  7.398e-01   5.871 1.41e-06 ***
## senate       1.454e-01  5.644e-01   0.258   0.7983
## vsat         3.235e-01  8.936e-01   0.362   0.7197
## percent      8.894e-01  1.334e+00   0.667   0.5096
## expense      1.174e-02  1.406e-02   0.835   0.4097
## income       7.519e-01  4.241e+00   0.177   0.8603
## high        -1.353e+00  2.704e+00  -0.500   0.6203
## college     -6.650e+00  5.440e+00  -1.222   0.2302
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 60.22 on 33 degrees of freedom
## Multiple R-squared:  0.8139, Adjusted R-squared:  0.7349
## F-statistic: 10.31 on 14 and 33 DF,  p-value: 2.451e-08
```

```
summary(lm(energy ~ . - msat - csat - house - density - income,
           data = states.data.num))
```

```
##
## Call:
## lm(formula = energy ~ . - msat - csat - house - density - income,
##     data = states.data.num)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -123.810  -26.606   -0.543   22.380  158.698
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.924e+01  4.712e+02  -0.126   0.9007
## pop         -1.616e-06  2.747e-06  -0.589   0.5601
## area         6.282e-04  3.115e-04   2.017   0.0516 .
## metro        6.490e-01  6.835e-01   0.949   0.3491
## waste        1.395e+01  4.810e+01   0.290   0.7736
## miles        1.338e+01  1.291e+01   1.036   0.3074
## toxic        2.762e+00  5.772e-01   4.785 3.26e-05 ***
## green        4.345e+00  7.292e-01   5.959 9.75e-07 ***
## senate       1.647e-01  5.458e-01   0.302   0.7647
## vsat         3.342e-01  8.787e-01   0.380   0.7061
## percent      9.221e-01  1.302e+00   0.708   0.4838
## expense      1.204e-02  1.375e-02   0.876   0.3873
## high        -1.195e+00  2.517e+00  -0.475   0.6381
## college     -6.301e+00  5.000e+00  -1.260   0.2162
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 59.35 on 34 degrees of freedom
## Multiple R-squared:  0.8137, Adjusted R-squared:  0.7425
## F-statistic: 11.42 on 13 and 34 DF,  p-value: 7.1e-09
```

```
summary(lm(energy ~ . - msat - csat - house - density - income - waste,
           data = states.data.num))
```

```
##
## Call:
## lm(formula = energy ~ . - msat - csat - house - density - income -
##     waste, data = states.data.num)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -125.889  -27.884   -0.994   22.113  158.780
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.380e+01  4.646e+02  -0.116   0.9085
## pop         -1.287e-06  2.468e-06  -0.522   0.6053
## area         6.064e-04  2.982e-04   2.033   0.0497 *
## metro        7.011e-01  6.508e-01   1.077   0.2887
## miles        1.405e+01  1.253e+01   1.121   0.2698
## toxic        2.725e+00  5.557e-01   4.904 2.14e-05 ***
## green        4.311e+00  7.105e-01   6.068 6.28e-07 ***
## senate       1.550e-01  5.376e-01   0.288   0.7748
## vsat         3.376e-01  8.671e-01   0.389   0.6993
## percent      9.025e-01  1.284e+00   0.703   0.4867
## expense      1.190e-02  1.356e-02   0.878   0.3861
## high        -1.183e+00  2.483e+00  -0.477   0.6367
## college     -6.330e+00  4.933e+00  -1.283   0.2079
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 58.57 on 35 degrees of freedom
## Multiple R-squared:  0.8132, Adjusted R-squared:  0.7492
## F-statistic:  12.7 on 12 and 35 DF,  p-value: 1.998e-09
```

```
summary(lm(energy ~ . - msat - csat - house - density - income - waste - senate,
           data = states.data.num))
```

```
##
## Call:
## lm(formula = energy ~ . - msat - csat - house - density - income -
##     waste - senate, data = states.data.num)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -127.815  -27.301    0.462   23.246  158.859
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.751e+01  4.497e+02  -0.061   0.9516
## pop         -1.344e-06  2.429e-06  -0.553   0.5834
## area         5.851e-04  2.853e-04   2.051   0.0476 *
## metro        6.848e-01  6.400e-01   1.070   0.2918
## miles        1.323e+01  1.205e+01   1.098   0.2794
## toxic        2.686e+00  5.314e-01   5.054 1.28e-05 ***
## green        4.283e+00  6.948e-01   6.165 4.19e-07 ***
## vsat         2.945e-01  8.431e-01   0.349   0.7289
## percent      8.412e-01  1.250e+00   0.673   0.5052
## expense      1.317e-02  1.267e-02   1.040   0.3055
## high        -1.156e+00  2.450e+00  -0.472   0.6398
## college     -6.020e+00  4.753e+00  -1.267   0.2134
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 57.82 on 36 degrees of freedom
## Multiple R-squared:  0.8128, Adjusted R-squared:  0.7556
## F-statistic: 14.21 on 11 and 36 DF,  p-value: 5.307e-10
```

```
summary(lm(energy ~ . - msat - csat - house - density - income - waste - senate
           - vsat, data = states.data.num))
```

```
##
## Call:
## lm(formula = energy ~ . - msat - csat - house - density - income -
##     waste - senate - vsat, data = states.data.num)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -128.589  -27.684   -1.736   21.869  158.678
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.138e+02  1.940e+02   0.587   0.5611
## pop         -1.487e-06  2.365e-06  -0.628   0.5336
## area         5.518e-04  2.656e-04   2.077   0.0448 *
## metro        6.104e-01  5.964e-01   1.024   0.3127
## miles        1.206e+01  1.144e+01   1.055   0.2983
## toxic        2.700e+00  5.234e-01   5.159 8.62e-06 ***
## green        4.274e+00  6.860e-01   6.230 3.07e-07 ***
## percent      4.519e-01  5.583e-01   0.809   0.4235
## expense      1.356e-02  1.247e-02   1.088   0.2838
## high        -1.096e+00  2.414e+00  -0.454   0.6526
## college     -5.194e+00  4.073e+00  -1.275   0.2102
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 57.13 on 37 degrees of freedom
## Multiple R-squared:  0.8122, Adjusted R-squared:  0.7614
## F-statistic:    16 on 10 and 37 DF,  p-value: 1.351e-10
```

```r
summary(lm(energy ~ . - msat - csat - house - density - income - waste - senate
           - vsat - high, data = states.data.num))
```

```
##
## Call:
## lm(formula = energy ~ . - msat - csat - house - density - income -
##     waste - senate - vsat - high, data = states.data.num)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -136.21  -26.01   -0.32   23.72  160.11
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.223e+01  1.373e+02   0.381   0.7057
## pop         -1.127e-06  2.205e-06  -0.511   0.6123
## area         5.086e-04  2.453e-04   2.073   0.0450 *
## metro        6.339e-01  5.879e-01   1.078   0.2877
## miles        1.250e+01  1.128e+01   1.109   0.2744
## toxic        2.738e+00  5.115e-01   5.353 4.39e-06 ***
## green        4.257e+00  6.778e-01   6.280 2.35e-07 ***
## percent      5.116e-01  5.369e-01   0.953   0.3466
## expense      1.236e-02  1.206e-02   1.025   0.3116
## college     -6.348e+00  3.148e+00  -2.016   0.0509 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 56.53 on 38 degrees of freedom
## Multiple R-squared:  0.8111, Adjusted R-squared:  0.7664
## F-statistic: 18.13 on 9 and 38 DF,  p-value: 3.358e-11
```

```r
summary(lm(energy ~ . - msat - csat - house - density - income - waste - senate
           - vsat - high - pop, data = states.data.num))
```

```
##
## Call:
## lm(formula = energy ~ . - msat - csat - house - density - income -
##     waste - senate - vsat - high - pop, data = states.data.num)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -133.431  -26.533    0.072   22.267  158.721
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.268e+01  1.347e+02   0.317   0.7530
## area         4.414e-04  2.052e-04   2.151   0.0377 *
## metro        5.088e-01  5.294e-01   0.961   0.3424
## miles        1.364e+01  1.095e+01   1.246   0.2201
## toxic        2.760e+00  5.047e-01   5.469 2.82e-06 ***
## green        4.251e+00  6.712e-01   6.333 1.79e-07 ***
## percent      4.473e-01  5.169e-01   0.865   0.3922
## expense      1.214e-02  1.193e-02   1.017   0.3153
## college     -5.906e+00  2.999e+00  -1.970   0.0560 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 55.99 on 39 degrees of freedom
## Multiple R-squared:  0.8098, Adjusted R-squared:  0.7708
## F-statistic: 20.76 on 8 and 39 DF,  p-value: 7.987e-12
```

```
summary(lm(energy ~ . - msat - csat - house - density - income - waste - senate
           - vsat - high - pop - percent, data = states.data.num))
```

```
##
## Call:
## lm(formula = energy ~ . - msat - csat - house - density - income -
##     waste - senate - vsat - high - pop - percent, data = states.data.num)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -134.599  -27.269   -1.304   21.873  159.145
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.792e+00  1.221e+02  -0.047   0.9624
## area         4.004e-04  1.990e-04   2.011   0.0511 .
## metro        6.108e-01  5.146e-01   1.187   0.2422
## miles        1.686e+01  1.026e+01   1.643   0.1083
## toxic        2.726e+00  5.016e-01   5.435 2.93e-06 ***
## green        4.045e+00  6.259e-01   6.463 1.06e-07 ***
## expense      1.695e-02  1.053e-02   1.610   0.1153
## college     -5.303e+00  2.907e+00  -1.824   0.0756 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 55.81 on 40 degrees of freedom
## Multiple R-squared:  0.8062, Adjusted R-squared:  0.7722
## F-statistic: 23.77 on 7 and 40 DF,  p-value: 2.233e-12
```

```
summary(lm(energy ~ . - msat - csat - house - density - income - waste - senate
           - vsat - high - pop - percent - metro, data = states.data.num))
```

```
##
## Call:
## lm(formula = energy ~ . - msat - csat - house - density - income -
##     waste - senate - vsat - high - pop - percent - metro, data = states.data.num)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -132.727  -30.155    1.581   21.121  164.617
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.549e+01  1.112e+02   0.499   0.6204
## area         4.446e-04  1.965e-04   2.263   0.0290 *
## miles        1.155e+01  9.284e+00   1.244   0.2205
## toxic        2.766e+00  5.029e-01   5.500 2.22e-06 ***
## green        3.946e+00  6.233e-01   6.330 1.47e-07 ***
## expense      1.825e-02  1.052e-02   1.734   0.0904 .
## college     -4.343e+00  2.807e+00  -1.547   0.1295
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 56.09 on 41 degrees of freedom
## Multiple R-squared:  0.7993, Adjusted R-squared:   0.77
## F-statistic: 27.22 on 6 and 41 DF,  p-value: 7.94e-13
```

```
summary(lm(energy ~ . - msat - csat - house - density - income - waste - senate
           - vsat - high - pop - percent - metro - miles, data = states.data.num))
```

```
##
## Call:
## lm(formula = energy ~ . - msat - csat - house - density - income -
##     waste - senate - vsat - high - pop - percent - metro - miles,
##     data = states.data.num)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -149.366  -25.937   -3.356   23.853  159.998
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.721e+02  6.020e+01   2.859  0.00658 **
## area         4.476e-04  1.978e-04   2.264  0.02883 *
## toxic        2.682e+00  5.016e-01   5.347 3.43e-06 ***
## green        4.334e+00  5.429e-01   7.983 5.90e-10 ***
## expense      1.234e-02  9.452e-03   1.306  0.19878
## college     -3.816e+00  2.792e+00  -1.366  0.17906
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 56.46 on 42 degrees of freedom
## Multiple R-squared:  0.7918, Adjusted R-squared:  0.767
## F-statistic: 31.94 on 5 and 42 DF,  p-value: 2.773e-13
```

```
summary(lm(energy ~ . - msat - csat - house - density - income - waste - senate
           - vsat - high - pop - percent - metro - miles - expense, data = states.data.num))
```

```
##
## Call:
## lm(formula = energy ~ . - msat - csat - house - density - income -
##     waste - senate - vsat - high - pop - percent - metro - miles -
##     expense, data = states.data.num)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -166.285  -24.407   -2.312   21.667  170.718
```

```
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.054e+02  5.499e+01   3.735 0.000547 ***
## area         3.409e-04  1.815e-04   1.878 0.067214 .
## toxic        2.458e+00  4.753e-01   5.172 5.75e-06 ***
## green        4.421e+00  5.432e-01   8.139 3.02e-10 ***
## college     -1.886e+00  2.388e+00  -0.790 0.434086
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 56.92 on 43 degrees of freedom
## Multiple R-squared:  0.7833, Adjusted R-squared:  0.7631
## F-statistic: 38.86 on 4 and 43 DF,  p-value: 9.379e-14
```

```
summary(lm(energy ~ . - msat - csat - house - density - income - waste - senate
           - vsat - high - pop - percent - metro - miles - expense - college, data = states.data.num))
```

```
##
## Call:
## lm(formula = energy ~ . - msat - csat - house - density - income -
##     waste - senate - vsat - high - pop - percent - metro - miles -
##     expense - college, data = states.data.num)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -176.556  -27.333   -5.631   24.405  167.922
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.643e+02  1.774e+01   9.262 6.72e-12 ***
## area        3.383e-04  1.807e-04   1.872   0.0678 .
## toxic       2.553e+00  4.578e-01   5.577 1.41e-06 ***
## green       4.523e+00  5.255e-01   8.608 5.50e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 56.67 on 44 degrees of freedom
## Multiple R-squared:  0.7802, Adjusted R-squared:  0.7652
## F-statistic: 52.05 on 3 and 44 DF,  p-value: 1.607e-14
```

```
summary(lm(energy ~ . - msat - csat - house - density - income - waste - senate
           - vsat - high - pop - percent - metro - miles - expense - college - area, data = states.data
```

```
##
## Call:
## lm(formula = energy ~ . - msat - csat - house - density - income -
##     waste - senate - vsat - high - pop - percent - metro - miles -
##     expense - college - area, data = states.data.num)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -174.763  -28.685   -3.589   17.280  196.598
```

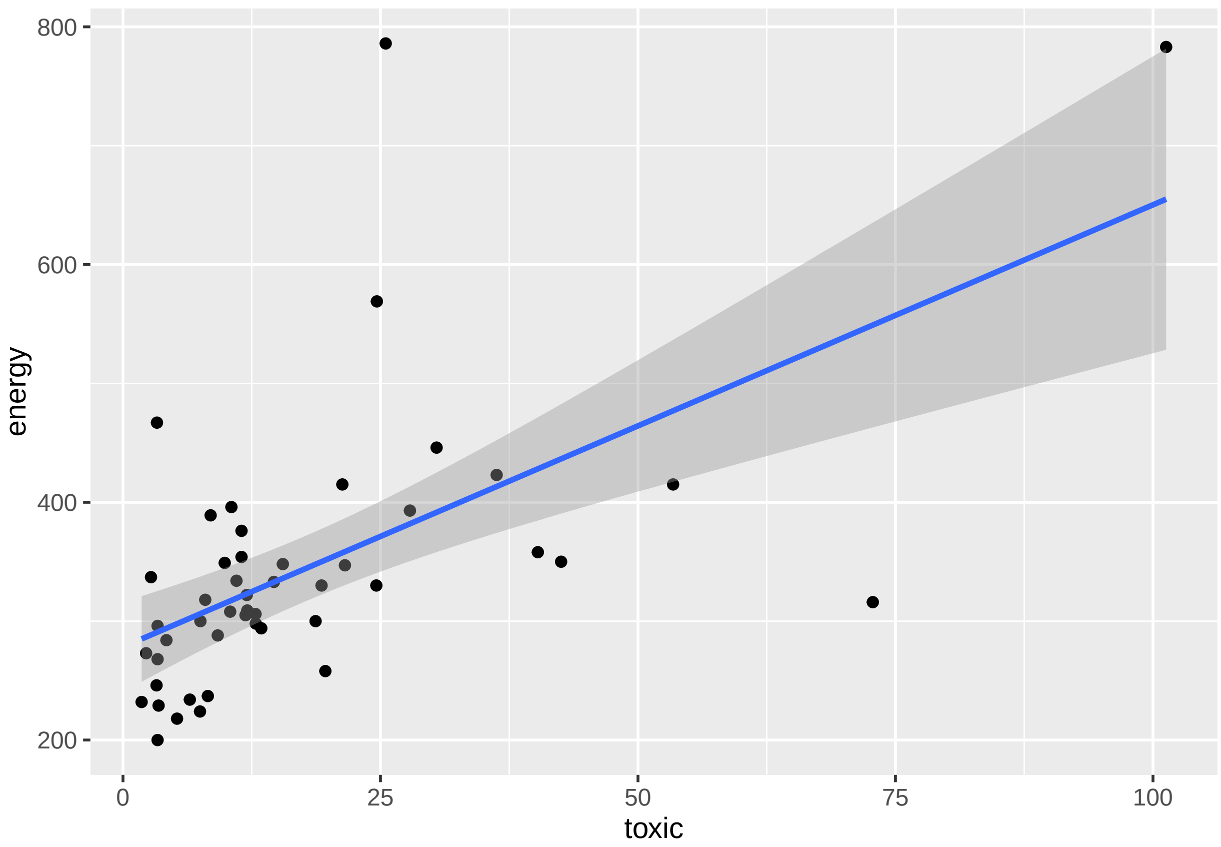
```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 179.8260    16.1194  11.156 1.51e-14 ***
## toxic         2.6455     0.4676   5.657 1.01e-06 ***
## green         4.6722     0.5336   8.756 2.81e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 58.23 on 45 degrees of freedom
## Multiple R-squared:  0.7627, Adjusted R-squared:  0.7521
## F-statistic:  72.3 on 2 and 45 DF,  p-value: 8.835e-15
```

We can see that per capita toxics and greenhouse gas released have very strong correlation with the energy consumed per capita. The correlation makes sense, but the causality may be reversed since it is the energy consumption that cause the releases of toxic and greenhouse gas. We can plot the graphs to show the trends.

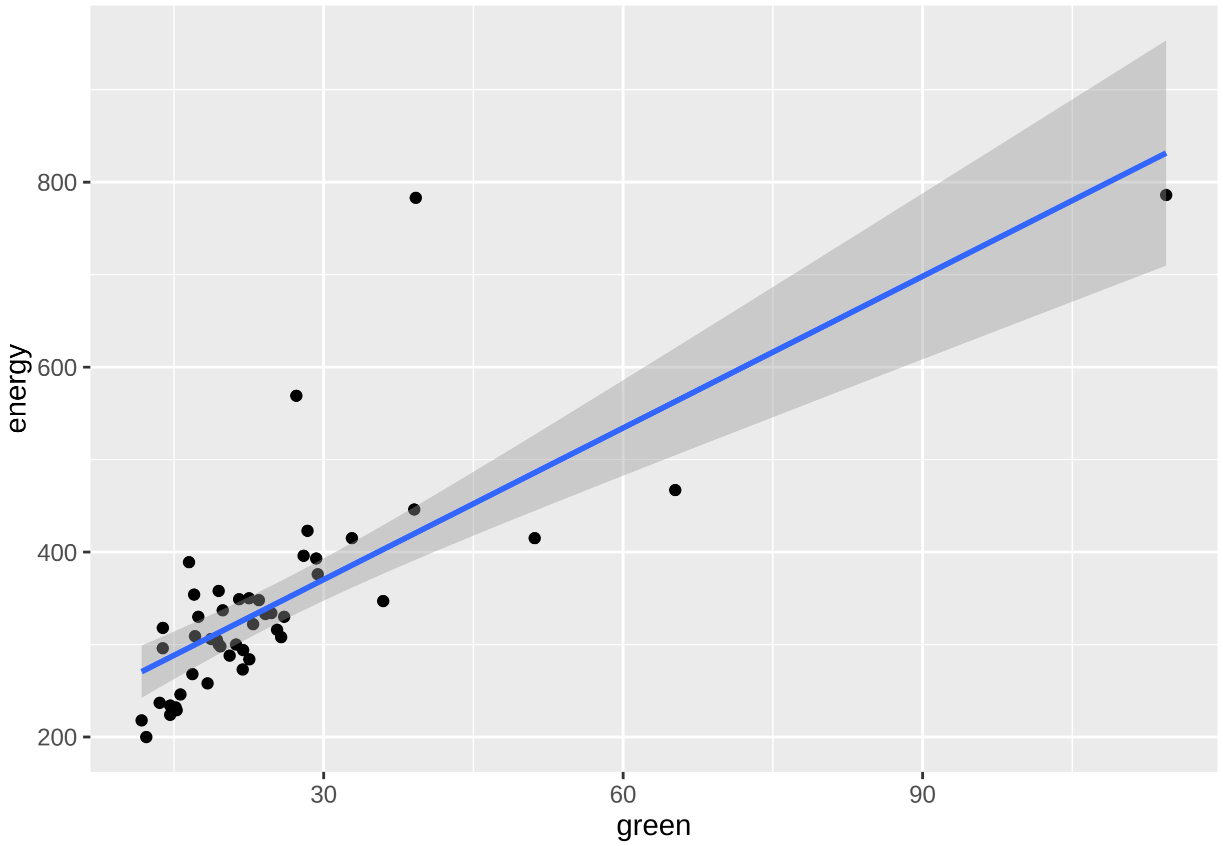
```
summary(lm(energy ~ toxic + green, data = states.data.num))
```

```
##
## Call:
## lm(formula = energy ~ toxic + green, data = states.data.num)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -174.763  -28.685   -3.589   17.280  196.598
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 179.8260    16.1194  11.156 1.51e-14 ***
## toxic         2.6455     0.4676   5.657 1.01e-06 ***
## green         4.6722     0.5336   8.756 2.81e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 58.23 on 45 degrees of freedom
## Multiple R-squared:  0.7627, Adjusted R-squared:  0.7521
## F-statistic:  72.3 on 2 and 45 DF,  p-value: 8.835e-15
```

```
ggplot(states.data.num, aes(x = toxic, y = energy)) + geom_point(size = 1.5) + stat_smooth(method = lm)
```



```
ggplot(states.data.num, aes(x = green, y = energy)) + geom_point(size = 1.5) + stat_smooth(method = lm)
```



However, let's see if we take out toxic and green to see if there is any other factors that may have better correlation that cause higher energy consumption.

```r
summary(lm(energy ~ . - msat - toxic - green, data = states.data.num))
```

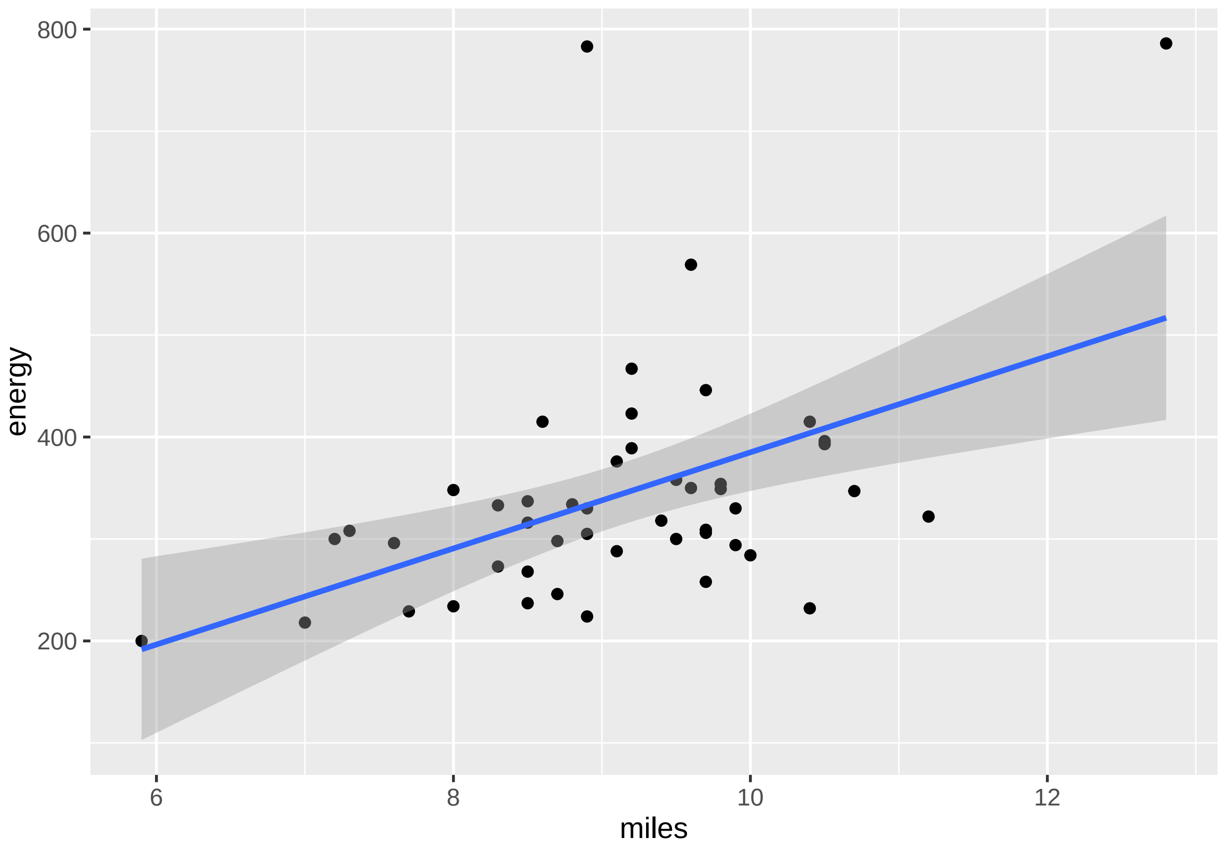
```
##
## Call:
## lm(formula = energy ~ . - msat - toxic - green, data = states.data.num)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -193.85  -45.33  -13.95   34.41  343.16
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.096e+02  7.800e+02   0.141   0.8891
## pop          2.415e-06  5.381e-06   0.449   0.6566
## area         1.860e-04  5.570e-04   0.334   0.7406
## density      1.551e-01  1.288e-01   1.205   0.2372
## metro        3.197e-01  1.405e+00   0.228   0.8215
## waste       -5.220e+01  8.034e+01  -0.650   0.5205
## miles        4.465e+01  2.059e+01   2.169   0.0377 *
## house       -2.588e+00  1.273e+00  -2.033   0.0504 .
## senate      -2.015e-01  9.695e-01  -0.208   0.8367
## csat        -1.521e+00  2.906e+00  -0.523   0.6043
## vsat         2.919e+00  6.072e+00   0.481   0.6340
## percent     -8.182e-01  2.219e+00  -0.369   0.7147
## expense      2.859e-02  2.316e-02   1.235   0.2259
## income      -4.906e+00  7.060e+00  -0.695   0.4921
## high         2.982e+00  5.326e+00   0.560   0.5794
## college     -5.976e+00  8.801e+00  -0.679   0.5020
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 97.16 on 32 degrees of freedom
## Multiple R-squared:  0.5301, Adjusted R-squared:  0.3099
## F-statistic: 2.407 on 15 and 32 DF,  p-value: 0.01815
```

It turned out that the miles of roads built and the house voting on environmental bill have pretty high impact on the energy consumption. The interesting part is the correlation between the house bill voted is negative to the energy consumption, which means that the more energy bill voted, the less energy consumption will be. Here are the correlation plots:

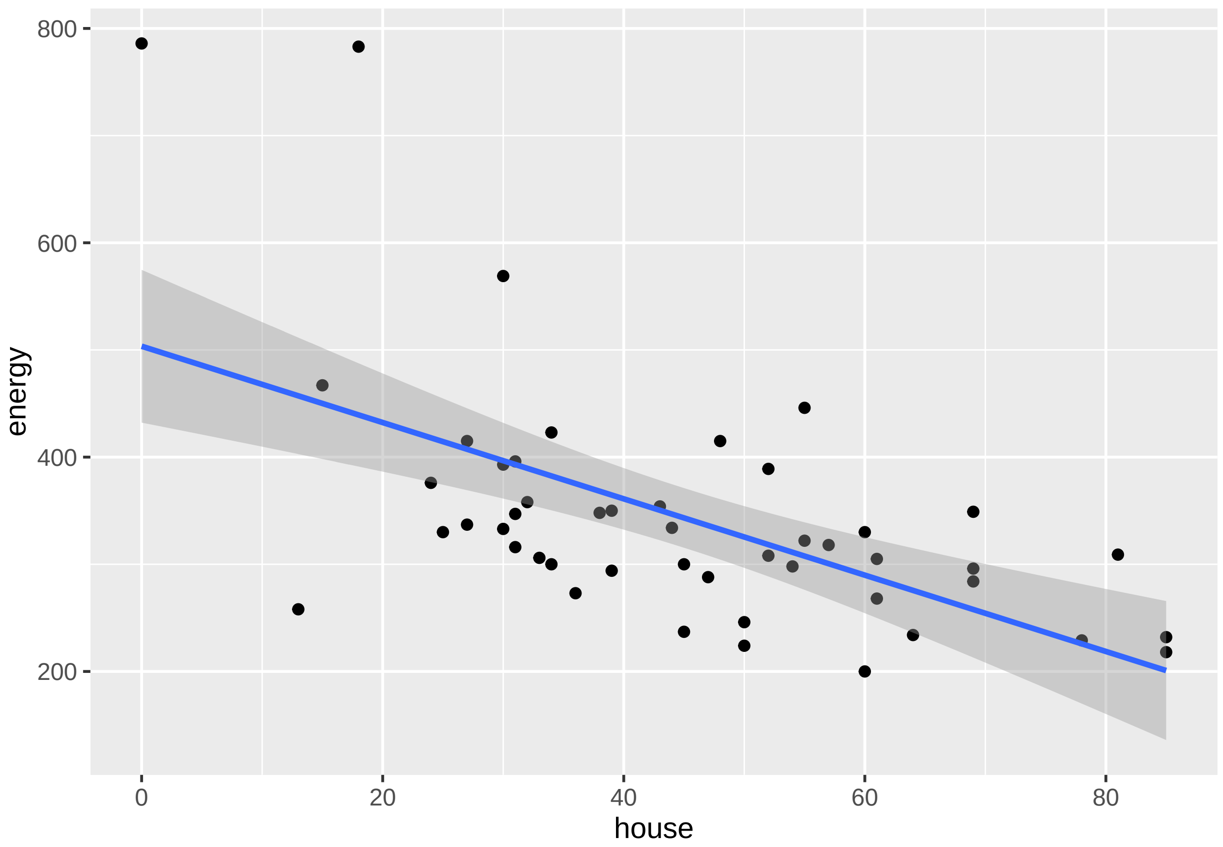
```
summary(lm(energy ~ miles + house, data = states.data.num))
```

```
##
## Call:
## lm(formula = energy ~ miles + house, data = states.data.num)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -197.69  -48.78  -16.69   35.51  367.35
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 186.5377   122.3020   1.525 0.134202
## miles        31.6854    11.7671   2.693 0.009920 **
## house        -2.9382     0.7192  -4.085 0.000179 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 89.89 on 45 degrees of freedom
## Multiple R-squared:  0.4344, Adjusted R-squared:  0.4093
## F-statistic: 17.28 on 2 and 45 DF,  p-value: 2.695e-06
```

```
ggplot(states.data.num, aes(x = miles, y = energy)) + geom_point(size = 1.5) + stat_smooth(method = lm)
```



```
ggplot(states.data.num, aes(x = house, y = energy)) + geom_point(size = 1.5) + stat_smooth(method = lm)
```



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?

Interactions

Let's say we looked at miles and house we explored in the previous exercise:

```
summary(lm(energy ~ miles*house, data = states.data.num))
```

```
##
## Call:
## lm(formula = energy ~ miles * house, data = states.data.num)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -198.80  -39.03  -11.31   32.91  383.33
##
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -137.3030   211.9374  -0.648  0.52045
## miles         65.4747    21.5839   3.033  0.00405 **
## house          4.2680     3.9627   1.077  0.28733
## miles:house   -0.7652     0.4141  -1.848  0.07139 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 87.57 on 44 degrees of freedom
## Multiple R-squared:  0.4752, Adjusted R-squared:  0.4394
## F-statistic: 13.28 on 3 and 44 DF,  p-value: 2.63e-06
```

We can see that independently miles of road have higher impact to the energy consumption than house bill voted. Additional of house bill voted to the miles built has negative impact to miles built factor, and vice versa.

Region

First, let's look at the region data:

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

```
##    West N. East   South Midwest    NA's
##      13       9      16      12       1
```

If we look at the correlation of the energy consumption to the region:

```
summary(lm(energy ~ region, data = states.data))
```

```
##
## Call:
## lm(formula = energy ~ region, data = states.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -162.62  -58.49  -30.62   12.00  585.38
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     405.62      39.23  10.339  1.4e-13 ***
## regionN. East  -156.50      61.34  -2.552   0.0141 *
## regionSouth     -25.49      52.82  -0.483   0.6317
## regionMidwest   -61.62      56.63  -1.088   0.2822
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 141.5 on 46 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.1367, Adjusted R-squared:  0.08041
## F-statistic: 2.428 on 3 and 46 DF,  p-value: 0.07737
```

We can see that the North East region has higher impact on the energy consumption per capita in the US.