DMwP-EnergyUsage

Wenze ma, Darren Zhu, Lawrence Ko, Shiloh Musser, Steven Feng

Jan 22, 2021

# Energy Usage in the U.S.(2007)

## “The U.S. Energy Information Adminstration collects and curates self-reported information from energy utilities about energy production and usage in the United States. This data set contains information from over 2,000 U.S. utilities in 2017. The information includes sources of energy, its uses in different economic sectors, and the revenues obtained from the sale of electrical energy.”

getwd()

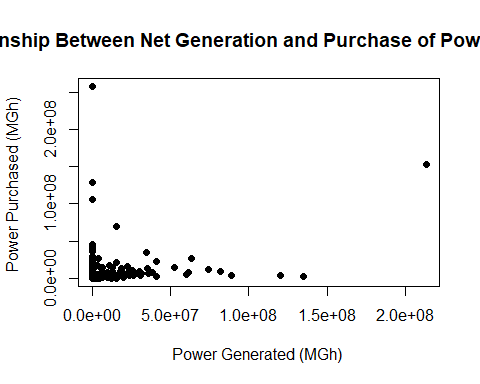
## [1] "C:/Users/fengr/Documents/GitHub/DMwP-EnergyUsage"

setwd('C:/Users/fengr/Desktop/CSSE386\_Data\_Mining/DMwP-EnergyUsage') #set to the folder where the data is saved  
electricity <- read.csv('electricity.csv', header = TRUE) #import the data to the R session

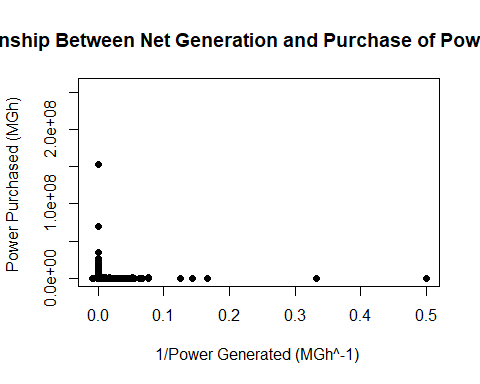
head(electricity)

## Utility.Number Utility.Name Utility.State  
## 1 55 City of Aberdeen - (MS) MS  
## 2 59 City of Abbeville - (LA) LA  
## 3 84 A & N Electric Coop VA  
## 4 97 Adams Electric Coop IL  
## 5 108 Adams-Columbia Electric Coop WI  
## 6 113 Agway Energy Services, LLC NY  
## Utility.Type Demand.Summer.Peak Demand.Winter.Peak  
## 1 Municipal 32.7 30.7  
## 2 Municipal 31.4 22.9  
## 3 Cooperative 152.6 156.1  
## 4 Cooperative 37.0 37.0  
## 5 Cooperative 109.1 83.4  
## 6 Retail Power Marketer 62.0 80.0  
## Sources.Generation Sources.Purchased Sources.Other Sources.Total Uses.Retail  
## 1 0 204875 0 204875 198442  
## 2 0 137537 0 137537 127849  
## 3 19 714961 0 714980 672572  
## 4 4677 170432 0 175109 164358  
## 5 0 524637 0 524637 493241  
## 6 0 423959 0 423959 422061  
## Uses.Resale Uses.No.Charge Uses.Consumed Uses.Losses Uses.Total  
## 1 0 0 70 6363 204875  
## 2 0 0 0 9688 137537  
## 3 0 0 1687 40721 714980  
## 4 0 0 318 10433 175109  
## 5 35 0 1122 30239 524637  
## 6 0 0 0 1898 423959  
## Revenues.Retail Revenue.Delivery Revenue.Resale Revenue.Adjustments  
## 1 14649.0 0 0 0  
## 2 12574.4 0 0 0  
## 3 68496.7 0 0 0  
## 4 22070.0 0 0 0  
## 5 67563.0 0 3 0  
## 6 34283.3 0 0 0  
## Revenue.Transmission Revenue.Other Revenue.Total Retail.Residential.Revenue  
## 1 0 300.5 14949.5 3644.0  
## 2 0 0.0 12574.4 5279.9  
## 3 0 1008.5 69505.2 38544.0  
## 4 0 241.0 22311.0 16341.0  
## 5 0 0.0 67566.0 42976.0  
## 6 0 0.0 34283.3 25624.0  
## Retail.Residential.Sales Retail.Residential.Customers  
## 1 32158 2595  
## 2 52746 4464  
## 3 346555 31583  
## 4 109301 8603  
## 5 260879 34138  
## 6 310489 39431  
## Retail.Commercial.Revenue Retail.Commercial.Sales Retail.Commercial.Customers  
## 1 4980.0 47295 661  
## 2 4082.7 37537 1032  
## 3 16000.0 156716 4019  
## 4 5729.0 55057 222  
## 5 9189.0 72087 2074  
## 6 8659.3 111572 6451  
## Retail.Industrial.Revenue Retail.Industrial.Sales Retail.Industrial.Customers  
## 1 6025.0 118989 2  
## 2 3211.8 37566 27  
## 3 13952.7 169301 8  
## 4 0.0 0 0  
## 5 15398.0 160275 926  
## 6 0.0 0 0  
## Retail.Transportation.Revenue Retail.Transportation.Sales  
## 1 0 0  
## 2 0 0  
## 3 0 0  
## 4 0 0  
## 5 0 0  
## 6 0 0  
## Retail.Transportation.Customers Retail.Total.Revenue Retail.Total.Sales  
## 1 0 14649.0 198442  
## 2 0 12574.4 127849  
## 3 0 68496.7 672572  
## 4 0 22070.0 164358  
## 5 0 67563.0 493241  
## 6 0 34283.3 422061  
## Retail.Total.Customers  
## 1 3258  
## 2 5523  
## 3 35610  
## 4 8825  
## 5 37138  
## 6 45882

plot(electricity$Sources.Generation, electricity$Sources.Purchased, main="Relationship Between Net Generation and Purchase of Power (in MGh)",  
 xlab="Power Generated (MGh) ", ylab="Power Purchased (MGh) ", pch=19)

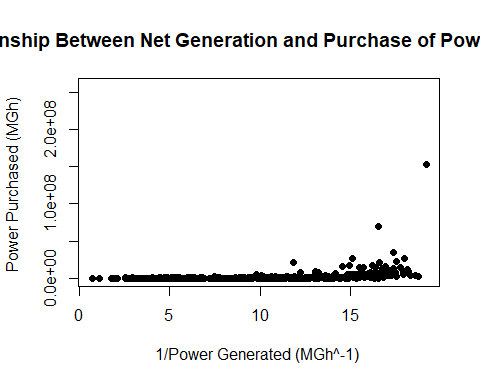
 It looks like they have inverse or exponential relationship, so just took reciprocal about the power generated.

plot(1/electricity$Sources.Generation, electricity$Sources.Purchased, main="Relationship Between Net Generation and Purchase of Power (in MGh)",  
 xlab="1/Power Generated (MGh^-1) ", ylab="Power Purchased (MGh) ", pch=19)



plot(log(electricity$Sources.Generation), electricity$Sources.Purchased, main="Relationship Between Net Generation and Purchase of Power (in MGh)",  
 xlab="1/Power Generated (MGh^-1) ", ylab="Power Purchased (MGh) ", pch=19)

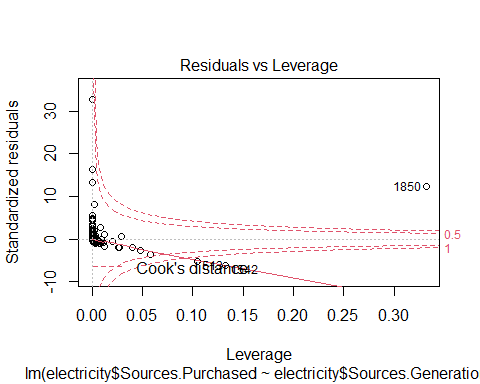
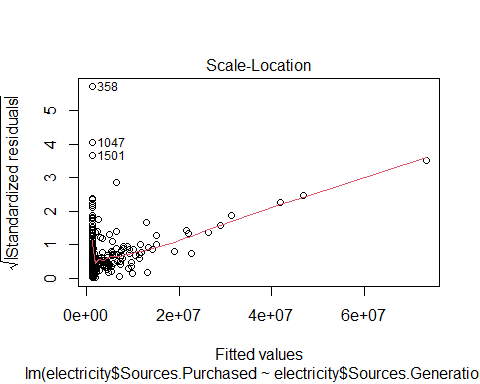
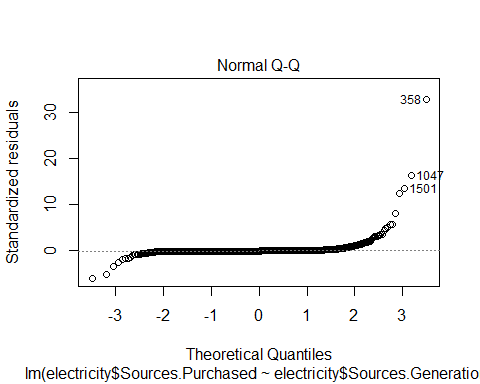
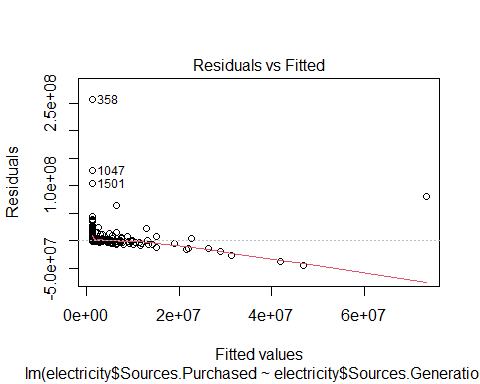
## Warning in log(electricity$Sources.Generation): 产生了NaNs



lm1=lm(electricity$Sources.Purchased~electricity$Sources.Generation)  
summary(lm1)

##   
## Call:  
## lm(formula = electricity$Sources.Purchased ~ electricity$Sources.Generation)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -44320855 -1030119 -879234 -502011 256843698   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.133e+06 1.722e+05 6.584 5.78e-11 \*\*\*  
## electricity$Sources.Generation 3.388e-01 2.130e-02 15.906 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7851000 on 2114 degrees of freedom  
## Multiple R-squared: 0.1069, Adjusted R-squared: 0.1065   
## F-statistic: 253 on 1 and 2114 DF, p-value: < 2.2e-16

plot(lm1)



# Create Training and Test data -  
set.seed(7) # setting seed to reproduce results of random sampling  
trainingRowIndex <- sample(1:nrow(electricity), 0.8\*nrow(electricity)) # row indices for training data  
trainingData <- electricity[trainingRowIndex, ] # model training data  
testData <- electricity[-trainingRowIndex, ] # test data

# Build the model on training data -  
lmMod <- lm(electricity$Revenue.Total ~ electricity$Sources.Total + electricity$Retail.Total.Revenue + electricity$Retail.Total.Sales + electricity$Retail.Total.Customers, data=trainingData) # build the model  
revPred <- predict(lmMod, testData) # predict distance

## Warning: 'newdata' had 424 rows but variables found have 2116 rows

summary (lmMod)

##   
## Call:  
## lm(formula = electricity$Revenue.Total ~ electricity$Sources.Total +   
## electricity$Retail.Total.Revenue + electricity$Retail.Total.Sales +   
## electricity$Retail.Total.Customers, data = trainingData)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2973143 -4600 -183 1257 3943771   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.312e+03 3.593e+03 -0.365 0.715   
## electricity$Sources.Total 4.366e-02 3.293e-04 132.606 <2e-16 \*\*\*  
## electricity$Retail.Total.Revenue 5.450e-01 1.911e-02 28.528 <2e-16 \*\*\*  
## electricity$Retail.Total.Sales -2.098e-02 1.352e-03 -15.514 <2e-16 \*\*\*  
## electricity$Retail.Total.Customers 7.454e-01 3.266e-02 22.821 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 159700 on 2111 degrees of freedom  
## Multiple R-squared: 0.9712, Adjusted R-squared: 0.9711   
## F-statistic: 1.779e+04 on 4 and 2111 DF, p-value: < 2.2e-16

library(MASS)

# Build the robust model on training data -  
rlmMod <- rlm(electricity$Revenue.Total ~ electricity$Sources.Total + electricity$Retail.Total.Revenue + electricity$Retail.Total.Sales + electricity$Retail.Total.Customers, data=trainingData, maxit=40, psi=psi.bisquare) # build the model  
revPred <- predict(lmMod, testData) # predict distance

## Warning: 'newdata' had 424 rows but variables found have 2116 rows

summary (rlmMod)

##   
## Call: rlm(formula = electricity$Revenue.Total ~ electricity$Sources.Total +   
## electricity$Retail.Total.Revenue + electricity$Retail.Total.Sales +   
## electricity$Retail.Total.Customers, data = trainingData,   
## maxit = 40, psi = psi.bisquare)  
## Residuals:  
## Min 1Q Median 3Q Max   
## -781059.17 -176.44 57.43 522.17 5478596.15   
##   
## Coefficients:  
## Value Std. Error t value   
## (Intercept) -60.9717 10.1313 -6.0181  
## electricity$Sources.Total 0.0328 0.0000 35354.8157  
## electricity$Retail.Total.Revenue 0.9996 0.0001 18558.5634  
## electricity$Retail.Total.Sales -0.0328 0.0000 -8601.5006  
## electricity$Retail.Total.Customers 0.0002 0.0001 1.6394  
##   
## Residual standard error: 416.8 on 2111 degrees of freedom

DMwR::regr.eval(electricity$Revenue.Total, lmMod$fitted.values) # error for the above lm model

## Registered S3 method overwritten by 'quantmod':  
## method from  
## as.zoo.data.frame zoo

## mae mse rmse mape   
## 2.905216e+04 2.545035e+10 1.595316e+05 Inf

DMwR::regr.eval(electricity$Revenue.Total, rlmMod$fitted.values) # error for the above rlm model

## mae mse rmse mape   
## 2.817872e+04 4.832209e+10 2.198229e+05 Inf

library(plyr)  
count(electricity$Revenue.Resale!=0)

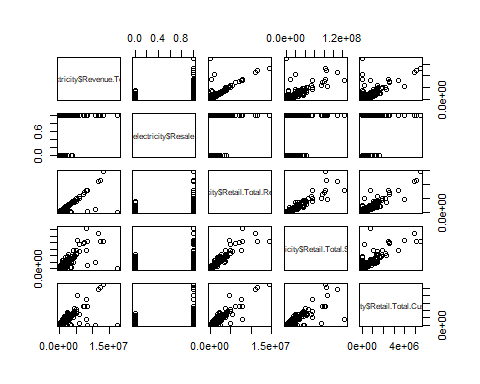
## x freq  
## 1 FALSE 1731  
## 2 TRUE 385

385/dim(electricity)[1]

## [1] 0.1819471

electricity$Resale[electricity$Revenue.Resale>0] <- 1  
electricity$Resale[electricity$Revenue.Resale==0] <- 0

pairs(~electricity$Revenue.Total + electricity$Resale + electricity$Retail.Total.Revenue + electricity$Retail.Total.Sales + electricity$Retail.Total.Customers)

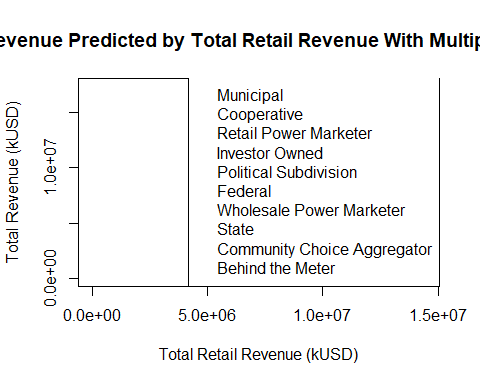


plot(electricity$Retail.Total.Revenue,electricity$Revenue.Total, pch=as.numeric(electricity$Utility.Type),  
 main="Total Revenue Predicted by Total Retail Revenue With Multiple Util Type",  
 ylab="Total Revenue (kUSD)", xlab="Total Retail Revenue (kUSD)")

## Warning in plot.xy(xy, type, ...): 强制改变过程中产生了NA

legend("bottomright", as.character(unique(electricity$Utility.Type)),   
 pch=as.numeric(unique(electricity$Utility.Type)))

## Warning in legend("bottomright",  
## as.character(unique(electricity$Utility.Type)), : 强制改变过程中产生了NA

 Add a new chunk by clicking the *Insert Chunk* button on the toolbar or by pressing *Ctrl+Alt+I*.

When you save the notebook, an HTML file containing the code and output will be saved alongside it (click the *Preview* button or press *Ctrl+Shift+K* to preview the HTML file).

The preview shows you a rendered HTML copy of the contents of the editor. Consequently, unlike *Knit*, *Preview* does not run any R code chunks. Instead, the output of the chunk when it was last run in the editor is displayed.