Coal

James Woods

8/31/2016

This assignment is not so much about coal but finding out where the class is in econometrics and specific skills with R and RStudio, a statistical programming language and Integrated Development Environment (IDE) respectively. There is plenty of help on R but if you want a readable primer (http://users.stat.umn.edu/~sandy/alr4ed/links/alrprimer.pdf) is a good choice.

1. Get RStudio and R (https://cran.r-project.org/) working on a computer. Most PSU computers will have these installed but if you are working on your own computer you can find directions at https://www.rstudio.com.

The install process is different depending on if you are running Windows, or Linux. They key is to install R first and then install RStudio. Many people find git (https://git-scm.com/) useful for version control and collaboration.

2. I will walk you through a few steps on reading in the data. The biggest hurdle to doing stats on the data is reading it in. There is actually an R library for working with EIA data directly, EIAdata, but we will use the more general tools to read files.
3. Start with a new markdown file similar to this one through the menu File > New File > R Markdown…
4. Download into R data on coal prices and quantities. The assignment operator in R is the "<-" symbol.

```
Coal <- read.csv("https://www.eia.gov/totalenergy/data/browser/csv.cfm?tbl=T06.01")
```

There are many ways of loading data into R (http://www.r-tutor.com/r-introduction/data-frame/data-import). Some work some of the time. In most cases Comma Separated Values (CSV) is the safest format to work with.

4. Take a look at the summary of the data

```
summary(Coal)
```

```
##       MSN           YYYYMM                  Value       Column_Order
##  CLEXPUS: 591   Min.   :194913   Not Available: 244   Min.   :1.00
##  CLIMPUS: 591   1st Qu.:198207   816.667      :   6   1st Qu.:2.75
##  CLLUPUS: 591   Median :199312   2            :   3   Median :4.50
##  CLNIPUS: 591   Mean   :199301   3            :   3   Mean   :4.50
##  CLPRPUS: 591   3rd Qu.:200504   114          :   2   3rd Qu.:6.25
##  CLSCPUS: 591   Max.   :201608   129          :   2   Max.   :8.00
##  (Other):1182                    (Other)      :4468
##                           Description                  Unit
##  Coal Consumption               : 591   Thousand Short Tons:4728
##  Coal Exports                   : 591
##  Coal Imports                   : 591
##  Coal Losses and Unaccounted for: 591
##  Coal Net Imports               : 591
##  Coal Production                : 591
##  (Other)                        :1182
```

You will notice that for some variables they give counts, e.g., MSN, and for others you get numerical summaries, e.g., Column_Order. The difference has to do with the data types (http://www.statmethods.net/input/datatypes.html).

5. Since we are trying to make a simple supply model, i.e., trying to explain coal production, lets select just the production part of the data set and also get only the annual production values. This is a little primer on changing data types and taking part of data.

- Load the dplyr package. This is the normal way to load libraries of functions that you need.

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

- Select only the Coal Production Figures and save it as CoalProduction. There is a cheat sheet for dplyr built into R. Look under the help menu.

```
CoalProduction <- Coal %>% filter(MSN == "CLPRPUS")
```

- Note that you have monthly data but that the annual data is shown as month 13. Grab all of the observations with month equal to 13.

```
library(stringr)

CoalProduction <- CoalProduction %>% filter(str_sub(as.character(YYYYMM),5 ) == "13")
```

- At this point you will have noticed that we don't need all the columns so lets keep just the ones we need and then give the two columns we saved new names.

```
CoalProduction <- CoalProduction %>% select(YYYYMM, Value)

names(CoalProduction) <- c("RawYear", "ProductionKShortTon")
summary(CoalProduction)
```

```
##     RawYear         ProductionKShortTon
##  Min.   :194913   1000048.758: 1
##  1st Qu.:196563   1016458.418: 1
##  Median :198213   1029075.527: 1
##  Mean   :198213   1032973.77 : 1
##  3rd Qu.:199863   1033504.288: 1
##  Max.   :201513   1063855.51 : 1
##                   (Other)    :61
```

- Notice that the ProductionKShortTon variable shows a count rather than a numerical summary. This means that R thinks it is a factor rather than a number. Lets fix that. It requires to first convert the factor, which is an integer, to the real value as a character and then convert that to numeric.

```
CoalProduction$ProductionKShortTon <- as.numeric(as.character(CoalProduction$ProductionKShortTon))
```

Play around with this doing one function at a time to see what each does and what each does alone.

- Next lets create a column for the year and make it a numeric value.

```
CoalProduction <- CoalProduction  %>% mutate(Year = as.numeric(str_sub(as.character(RawYear),0,4 )))


summary(CoalProduction)
```

```
##     RawYear       ProductionKShortTon      Year
##  Min.   :194913   Min.   : 420423     Min.   :1949
##  1st Qu.:196563   1st Qu.: 558547     1st Qu.:1966
##  Median :198213   Median : 829700     Median :1982
##  Mean   :198213   Mean   : 796953     Mean   :1982
##  3rd Qu.:199863   3rd Qu.:1033239     3rd Qu.:1998
##  Max.   :201513   Max.   :1171809     Max.   :2015
```

5. Now grab some price data. We are going to use Quandl (https://www.quandl.com/), which has a bunch of data on energy and a lot of other things (https://www.quandl.com/collections/markets/coal). Make sure you have the Quandl library installed. If you don’t install.packages(“Quandl”) should do it.

```
library(Quandl)
```

```
## Loading required package: xts
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
##
## Attaching package: 'xts'
```

```
## The following objects are masked from 'package:dplyr':
##
##     first, last
```

```
Prices <- Quandl("EPI/152")

summary(Prices)
```

```
##       Year               Price (U.S. Dollars)
##  Min.   :1949-01-01   Min.   :16.78
##  1st Qu.:1963-01-01   1st Qu.:20.19
##  Median :1977-01-01   Median :25.02
##  Mean   :1976-12-31   Mean   :27.85
##  3rd Qu.:1991-01-01   3rd Qu.:31.52
##  Max.   :2005-01-01   Max.   :50.92
```

- As before we will convert the year to a numeric value

```
Prices$Year <- as.numeric(str_sub(Prices$Year,0,4))
```

- And simplify the names

```
names(Prices) <- c("Year", "PriceMBTU")
```

Please note that we don't have the price per short ton of coal. What we have is the price per million BTUs, which is a measure of energy content. The BTUs per short ton of coal (2000 lbs) is about 20 MBTUs but varies from place-to-place and year-to-year.

- Merge the two data frames

```
summary(CoalProduction)
```

```
##     RawYear        ProductionKShortTon      Year
##  Min.   :194913   Min.   : 420423     Min.   :1949
##  1st Qu.:196563   1st Qu.: 558547     1st Qu.:1966
##  Median :198213   Median : 829700     Median :1982
##  Mean   :198213   Mean   : 796953     Mean   :1982
##  3rd Qu.:199863   3rd Qu.:1033239     3rd Qu.:1998
##  Max.   :201513   Max.   :1171809     Max.   :2015
```

```
summary(Prices)
```

```
##       Year        PriceMBTU
##  Min.   :1949   Min.   :16.78
##  1st Qu.:1963   1st Qu.:20.19
##  Median :1977   Median :25.02
##  Mean   :1977   Mean   :27.85
##  3rd Qu.:1991   3rd Qu.:31.52
##  Max.   :2005   Max.   :50.92
```

```
CoalMarket<-inner_join(Prices, CoalProduction, by ="Year")

summary(CoalMarket)
```

```
##       Year        PriceMBTU        RawYear       ProductionKShortTon
##  Min.   :1949   Min.   :16.78   Min.   :194913   Min.   : 420423
##  1st Qu.:1963   1st Qu.:20.19   1st Qu.:196313   1st Qu.: 529774
##  Median :1977   Median :25.02   Median :197713   Median : 684913
##  Mean   :1977   Mean   :27.85   Mean   :197713   Mean   : 750200
##  3rd Qu.:1991   3rd Qu.:31.52   3rd Qu.:199113   3rd Qu.: 995984
##  Max.   :2005   Max.   :50.92   Max.   :200513   Max.   :1131498
```

This is the data frame we all start with. Please note that we have not converted these to real prices or anything like that.

Undergraduate Students

Your group assignment is to estimate a supply or demand function, assuming no endogenaity, meaning that you can use OLS.

1. Look at Ch 2 of the primer (http://users.stat.umn.edu/~sandy/alr4ed/links/alrprimer.pdf)
2. Convert the prices to real with the CPI. Again quandl has what you need.
3. Create a simple supply model with the lm function. You can get as complicated as you like with the model, time trends, decades, log and square transformations.
4. Interpret each of the parameter in your model being very clear about units, e.g. a one dollar increase in the price per MBTU results in an . . .