G+ More robert.l.mcpherson@gmail.com Dashboard Sign Out

Fear and Loathing in Data Science

A Savage Journey to the Heart of Big Data

Select Language ▼

Wednesday, February 5, 2014

An Inconvenient Statistic

As I sit here waiting on more frigid temperatures subsequent to another 10 inches of snow, suffering from metastatic cabin fever, I can't help but ponder what I can do examine global warming/climate change. Well, as luck would have it, R has the tools to explore this controversy. Using two packages, vars and forecast, I will see if I should be purchasing carbon offsets or continue with a life of conspicuous consumption, oblivious to the consequences of my actions.

The concept is to find data on man-made carbon emissions and global surface temperatures. Then, using vector autoregression to identify the proper number of lags to put into a granger causality model. I will not get into any theory here, but you can see a discussion of granger causality in my very first post where I showed how to solve the age-old mystery of what comes first, the chicken or the egg (tongue firmly planted in cheek).

It is important to point out that two prior papers have shown no causal linkage between CO2 emissions and surface temperatures (Triacca, 2005 & an unpublished manuscript from Bilancia/Vitale). In essence, past observations of CO2 concentrations do not improve the statistical predictions of current surface temperatures. Be that as it may, I will attempt to duplicate such an analysis, giving any adventurous data scientist the tools and techniques to dig into this conundrum on their own.

Where can we find the data? Global CO emission estimates can be found at the Carbon Dioxide Information Analysis Center (CDIAC) at the following website - http://cdiac.ornl.gov/. You can download data of total emissions of fossil fuel combustion and cement manufacture. Surface temperature takes some detective work, but a clever soul can find it at the website of the UK Met Office Hadley Centre, part of the climate research center at the University of East Anglia, website - http://www.metoffice.gov.uk/hadobs/hadcrut4/ . An anomaly is calculated as the difference between the average annual surface temperature versus the average of the reference years, 1961 - 1990.

The data have common years from 1850 until 2010 and I downloaded and put it into a .csv for import into R. Now, it's on to the code!

```
> require(forecast)
> require(vars)
> var.data = read.csv(file.choose())
> head(var.data)
  Year CO2   Temp
1 1850  54 -0.374
2 1851  54 -0.219
3 1852  57 -0.223
4 1853  59 -0.268
5 1854  69 -0.243
6 1855  71 -0.264

> #put data into a time series
> carbon.ts = ts(CO2, frequency=1, start=c(1850), end=c(2010))
> temp.ts = ts(Temp, frequency=1, start=c(1850), end=c(2010))
#subset the data from 1900 until 2010
> surfacetemp = window(temp.ts, start=c(1900), end=c(2010))
> co2 = window(carbon.ts, start=c(1900), end=c(2010))
> climate.ts = cbind(co2, surfacetemp)
> plot(climate.ts)
```

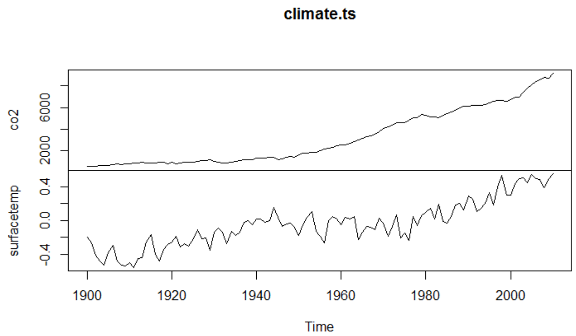
My Blog List

R bloggers
reticulate – another step towards a multilingual and collaborative way of working

Subscribe To

Posts
Comments

Blog Archive

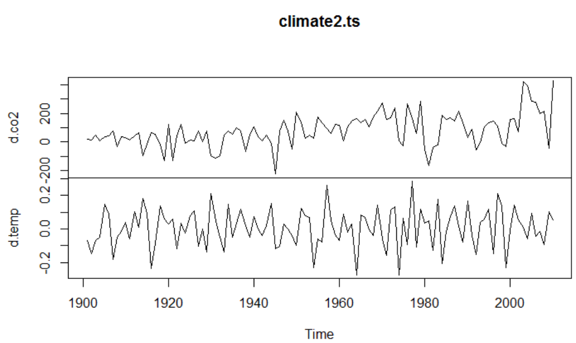


> #determine stationarity and number of lags to achieve stationarity
> ndiffs(co2, alpha = 0.05, test = c("adf"))
[1] 1
> ndiffs(surfacetemp, alpha = 0.05, test = c("adf"))
[1] 1

Using the adf test above in the ndiffs command of the forecast package, we can see that a 1st difference will allow us to achieve stationarity, which is necessary for vector autoregression and granger causality.

> #difference to achieve stationarity
> d.co2 = diff(co2)
> d.temp = diff(surfacetemp)

> #again, we need a mts class dataframe
> climate2.ts = cbind(d.co2, d.temp)
> plot(climate2.ts)



> #determine the optimal number of lags for vector autoregression
> VARselect(climate2.ts, lag.max=10) $selection

AIC(n) HQ(n) SC(n) FPE(n)
7 3 1 7

I find that the above divergence in the tests for optimal VAR modeling is quite common. Now, one can peruse the literature for what is the best statistical test to determine optimal lag length, but I like to use brute force and ignorance and try all of the above (i.e. lags 1, 3 and 7).

> #vector autoregression with lag1
> var = VAR(climate2.ts, p=1)

It is important now to test for serial autocorrelation in the model residuals and below is for the Portmanteau test (several options in the vars package are available).

> serial.test(var, lags.pt=10, type="PT.asymptotic")

Portmanteau Test (asymptotic)

About Me

Cory Lesmeister
G+ Follow 53

View my complete profile

```
data:  Residuals of VAR object var
Chi-squared = 55.4989, df = 36, p-value = 0.01996

#The null hypothesis is no serial correlation, so we can reject it with extreme prejudice...on to var3
> var3 = VAR(climate2.ts, p=3)
> serial.test(var3, lags.pt=10, type="PT.asymptotic")

 Portmanteau Test (asymptotic)

data:  Residuals of VAR object var3
Chi-squared = 36.1256, df = 28, p-value = 0.1394
```

That is more like it. You can review the details of the var model, in this case temperature, if you so choose:

```
> summary(var3, equation="d.temp")

VAR Estimation Results:
=========================
Endogenous variables: d.co2, d.temp
Deterministic variables: const
Sample size: 107
Log Likelihood: -548.435
Roots of the characteristic polynomial:
0.7812 0.7265 0.7265 0.6491 0.5846 0.5846
Call:
VAR(y = climate2.ts, p = 3)


Estimation results for equation d.temp:
=======================================
d.temp = d.co2.l1 + d.temp.l1 + d.co2.l2 + d.temp.l2 + d.co2.l3 + d.temp.l3 + const

            Estimate   Std. Error   t value  Pr(>|t|)
d.co2.l1     7.603e-05  1.014e-04    0.749   0.455372
d.temp.l1   -4.103e-01  9.448e-02   -4.343   3.37e-05 ***
d.co2.l2    -2.152e-05  1.115e-04   -0.193   0.847339
d.temp.l2   -3.922e-01  9.544e-02   -4.109   8.15e-05 ***
d.co2.l3     7.905e-05  1.041e-04    0.759   0.449465
d.temp.l3   -3.366e-01  9.263e-02   -3.633   0.000444 ***
const        7.539e-03  1.340e-02    0.563   0.574960
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


Residual standard error: 0.1014 on 100 degrees of freedom
Multiple R-Squared: 0.254, Adjusted R-squared: 0.2093
F-statistic: 5.676 on 6 and 100 DF,  p-value: 4.15e-05



Covariance matrix of residuals:
          d.co2       d.temp
d.co2   10972.588   -1.28920
d.temp     -1.289    0.01028

Correlation matrix of residuals:
          d.co2    d.temp
d.co2   1.0000   -0.1214
d.temp -0.1214    1.0000

> #does co2 granger cause temperature
> grangertest(d.temp ~ d.co2, order=3)

Granger causality test

Model 1: d.temp ~ Lags(d.temp, 1:3) + Lags(d.co2, 1:3)
Model 2: d.temp ~ Lags(d.temp, 1:3)
  Res.Df  Df      F       Pr(>F)
1    100
```

```
2    103   -3    0.5064     0.6787
```

```
> #Clearly the model is not significant, so we can say that carbon emissions do not granger-cause surface temperatures.

> #does temperature granger cause co2
> grangertest(d.co2 ~ d.temp, order =3)
```

```
Granger causality test

Model 1: d.co2 ~ Lags(d.co2, 1:3) + Lags(d.temp, 1:3)
Model 2: d.co2 ~ Lags(d.co2, 1:3)
  Res.Df    Df        F       Pr(>F)
1    100
2    103     -3    0.7799    0.5079
```

```
> #try again using lag 7
> grangertest(d.temp ~ d.co2, order=7)
```

```
Granger causality test

Model 1: d.temp ~ Lags(d.temp, 1:7) + Lags(d.co2, 1:7)
Model 2: d.temp ~ Lags(d.temp, 1:7)
  Res.Df    Df        F       Pr(>F)
1     88
2     95     -7    0.5817    0.7691
```

Again, nothing significant using lag 7. So, using this data and the econometric techniques spelled out above, it seems there is no causal effect (statistically speaking) between fossil fuel emissions and global surface temperatures. Certainly, this is not the final word on the matter as there is much measurement error in the data that the stewards have attempted to account for.

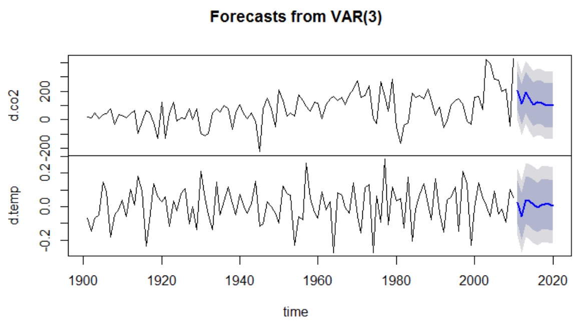
On a side note, we can use vars for predictions and forecast for time series plots of the predicted values.

```
> predict(var3, n.ahead=6, ci=0.95)
```

```
$d.co2
         fcst           lower         upper          CI
[1,] 202.5888      -2.717626     407.8953     205.3065
[2,] 110.3385  -105.847948     326.5249     216.1864
[3,] 192.1802     -26.160397     410.5207     218.3406
[4,] 152.5464     -74.948000     380.0408     227.4944
[5,] 108.4343  -122.198058     339.0666     230.6323
[6,] 123.9001  -107.882219     355.6824     231.7823

$d.temp
              fcst               lower           upper             CI
[1,]  0.026737000    -0.1719770    0.2254510    0.1987140
[2,] -0.057081637    -0.2731569    0.1589936    0.2160753
[3,]  0.040419451    -0.1803409    0.2611798    0.2207603
[4,]  0.032591047    -0.1893108    0.2544929    0.2219019
[5,]  0.013708836    -0.2143756    0.2417933    0.2280844
[6,] -0.004319714   -0.2324070    0.2237675    0.2280873
```

```
> fcst = forecast(var3)
> plot(fcst)
```



So what can we conclude from this exercise? Well, let's look to the good Doctor, Hunter S. Thompson for some philosophical insight. He would likely advise us...

"res ipsa locquitur"

References:

BILANCIA, MASSIMO, and DOMENICO VITALE. "GRANGER CAUSALITY ANALYSIS OF BIVARIATE CLIMATIC TIME SERIES: A NOTE ON THE ROLE OF CO2 EMISSIONS IN GLOBAL CLIMATE WARMING."

Morice, C. P., J. J. Kennedy, N. A. Rayner, and P. D. Jones (2012), Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: The HadCRUT4 dataset, J. Geophys. Res., 117, D08101, doi:10.1029/2011JD017187.

Triacca, U, Is Granger causality analysis appropriate to investigate the relationship between atmospheric concentration of carbon dioxide and global surface air temperature?, Theoretical and Applied Climatology, 81, 133-135

Posted by Cory Lesmeister at 12:08 AM

G+

9 comments:

Anonymous May 15, 2014 at 6:49 PM

Ooooo. Impressive single computer work. Probably better than all the server farms doing real climate modeling.

Reply

Cory Lesmeister May 15, 2014 at 7:13 PM

My data came from these server farms. Ethical server farms...free range, grass fed servers. At any rate, the point is to demonstrate Granger Causality applied to a controversial topic, not an attempt to rile the masses or sway individual's positions.

Reply

Anonymous March 16, 2015 at 11:52 AM

Hi there, I have a quick and possibly stupid question for you... when you were determining the best number of lags and put the code:

> VARselect(climate2.ts, lag.max=10) $selection

why did you have lag.max set to 10?

Reply

Cory Lesmeister March 16, 2015 at 6:58 PM

Actually, there was some trial and error involved when I was putting this together. If I recall, I tried up to max lag of 20, but always ended up with the results above. If you try and replicate the results and find something else, please let me know.

Reply

Replies

Anonymous March 16, 2015 at 7:39 PM

Thank you!

Anonymous March 16, 2015 at 8:09 PM

Is this also the reason you used 10 for the Portmanteau test? ie lags.pt=10?

Reply

Cory Lesmeister March 17, 2015 at 4:35 PM

Indeed.

Reply

Mehmet Ali Çakır February 18, 2018 at 7:59 AM

All codes work well except for fcst = forecast(var3). Can anyone explain why? Thanks

Reply

Cory Lesmeister February 18, 2018 at 10:44 AM

Odd. I will try and run the code what I put together in my book. Should have a copy around here somewhere. Perhaps, try doing forecast::forecast(var3). If you continue to get an error message, pass that along.

Reply

Enter your comment...

Comment as: Unknown (Goc ▼ Sign out

Publish Preview Notify me

Newer Post Home Older Post

Subscribe to: Post Comments (Atom)

Fear and Loathing in Data Science

Data You May Find Interesting - 4/20/2018
Simulation on Statistical Significance and Power - 10/1/2017
Exploration of Text-Mining Packages Using HST's 'Hey Rube' Columns - 7/23/2017
Mastering Machine Learning with R, 2nd Edition - 5/13/2017

Plotting Vietnam Airstrikes with Leaflet and R - 3/13/2017

My Blog List

R bloggers
reticulate – another step towards a multilingual and collaborative way of working
6 hours ago

My Blog List

R bloggers
reticulate – another step towards a multilingual and collaborative way of working

Subscribe To

Posts

Comments

My Blog List

R bloggers
reticulate – another step towards a multilingual and collaborative way of working

Awesome Inc. theme. Powered by Blogger.