the Tarzan

[R] + applied economics.

« Calculate OLS regression manually using matrix algebra in R | TikZ diagrams for economists: An excise tax »

The Chow test in R: A case study of Yellowstone’s Old Faithful Geyser

Recently I took a road trip south to Yellowstone National Park, where the fascinating phenomenon that is Old Faithful is still spewing piping-hot water 120 feet into the air every hour or so. From fellow spectators, I was told that the time between eruptions was approximately 60 minutes, but could be longer depending on the length of the previous eruption. I wondered, would it be possible to get a better estimate of our waiting time, conditional on the previous eruption’s length?

Fortunately, there is plenty of data available on the geyser. See the site: http://www.geyserstudy.org/geyser.aspx?pGeyserNo=OLDFAITHFUL

Let’s look at some data showing length of the eruption and waiting time between eruptions captured by some park rangers in the 1980s.

A description of the data can be found here.

The Chow Test for Structural Breaks

The Chow test is used to see if it makes sense to run two separate regressions on two mutually exclusive subsets of your data (divided by a break point) by comparing the results of the two “unrestricted” regressions versus the “restricted” regression that pools all the data together.

$$H_0 : \beta_{ur_1} = \beta_{ur_2}$$

$$H_a : \beta_{ur_1} \neq \beta_{ur_2}$$

The procedure is as follows:

1. Run a “restricted” regression on all your data (pooled).
2. Divide your sample into to groups, determined by your breakpoint (e.g. a point in time, or a variable value).
3. Run an “unrestricted” regression on each of your subsamples. You will run two “unrestricted” regressions with a single breakpoint.
4. Calculate the Chow F-statistic as follows:

$$\frac{(SSR_r - SSR_u)/k}{SSR_u/(n-2k)} \sim F_{k,n-2k}$$

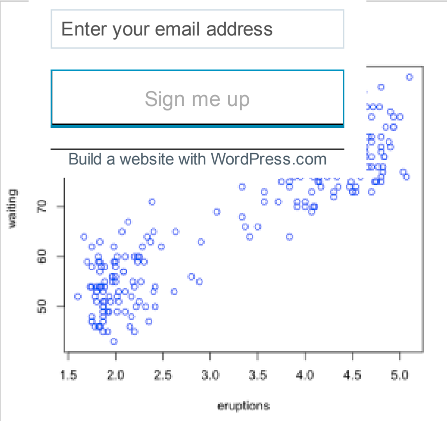
where $SSR_r =$ the sum of squared residuals of the restricted model, $SSR_u =$ the sum of squared residuals from the unrestricted models, $k =$ the number of regressors (including the intercept), and $n =$ the number of total observations.

For a much more thorough review of the topic, see page 113 of Dr. man’s class notes here.

Eruption Time versus Waiting

Following the relationship mentioned to me the length of the eruption and the subsequent waiting time between eruptions, let’s see if

```
of = faithful
plot(of, col="blue",main="Eruptic
```



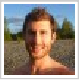
A scatterplot.

There certainly does. One thing that does pop out is the existence of two groups – there’s a cluster of data points in the upper

Search this blog

Search...

Contributors



Kevin Goulding

Categories

- Econometrics
- Econometrics with R
- Numpy
- Python
- R tips & tricks
- Surviving Graduate Econometrics with R
- TikZ for Economists
- Visualizing Data with R
- White Papers

Twitter feed

RT @gappy3000: This post, apparently about #julialang and #pydata, explains why #rstats has become the standard of data analysis http:// ... 2 years ago

RT @justinwolfers: "if prediction markets are really as valuable as economists think, then..more experimentation could prove worthwhile. ... 2 years ago

RT @vsbuffalo: For me the biggest victory is for statistics and empiricism. Go Nate Silver and @fivethirtyeight for a brilliant forecast ... 3 years ago

Follow @baha_kev

Tag Cloud

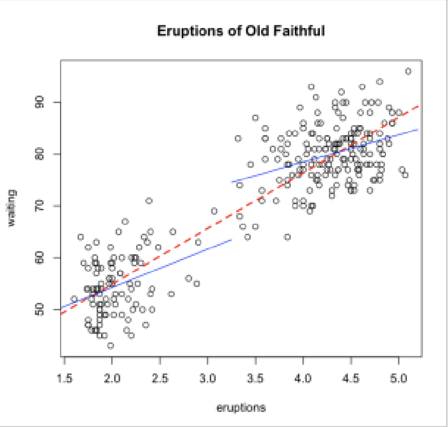
cluster-robust Econometrics heteroskedasticity LaTeX Numpy Parallel Computing plots Python R STATA tex TikZ right, and another in the lower left. What could explain this?

Maybe there is an extra chamber within the geyser that, once it is breached, empties completely. Then, it takes longer for the next eruption to occur because that chamber has to fill back up with boiling hot water. *NOTE: This non-scientific explanation is not backed by any scientific studies.*

So, from the scatter plot, it appears that the breakpoint between the two groups occurs at the value of 3.25 minutes of eruption. For this example, this will be our break point.

```
## Run three regressions (1 restricted, 2 unrestricted)
r.reg = lm(waiting ~ eruptions, data = of)
ur.reg1 = lm(waiting ~ eruptions, data = of[of$eruptions > 3.25,])
ur.reg2 = lm(waiting ~ eruptions, data = of[of$eruptions
## review the regression results
summary(reg.r)
summary(ur.reg1)
summary(ur.reg2)

## Calculate sum of squared residuals for each regression
SSR = NULL
SSR$r = r.reg$residuals^2
SSR$ur1 = ur.reg1$residuals^2
SSR$ur2 = ur.reg2$residuals^2

## K is the number of regressors in our model
K = r.reg$rank

## Computing the Chow test statistic (F-test)
numerator = ( sum(SSR$r) - (sum(SSR$ur1) + sum(SSR$ur2)) ) / K
denominator = (sum(SSR$ur1) + sum(SSR$ur2)) / (nrow(of) - 2*K)
chow = numerator / denominator
chow

## Calculate P-value
1-pf(chow, K, (nrow(of) - 2*K))
```

Now, we can plot the results. In this figure, the red dotted line is the restricted regression line, and the blue lines are the two unrestricted regression lines.



```
## Plot the results
plot(of,main="Eruptions of Old Faithful")
# restricted model
abline(r.reg, col = "red",lwd = 2, lty = "dashed")
# unrestricted model 1
segments(0, ur.reg2$coefficients[1], 3.25,
ur.reg2$coefficients[1]+3.25*ur.reg2$coefficients[2], col= 'blue')
# unrestricted model 2
segments(3.25, ur.reg1$coefficients[1]+3.25*ur.reg1$coefficients[2],
5.2, ur.reg1$coefficients[1]+5.2*ur.reg1$coefficients[2], col= 'blue')
```

A Quicker Way

The CRAN package `strucchange` provides a more streamlined approach to calculating a Chow test statistic, but it requires that the data is ordered so that the breakpoint can be identified by a specific row number. In our example, we'll have to order our data by eruptions, and then point out that the 98th data point is the last data point where eruptions is less than 3.25. The code is below:

```
## Sort the data
sort.of = of[order(of$eruptions) , ]
sort.of = cbind(index(sort.of),sort.of)

## Identify the row number of our breakpoint
brk = max(sort.of[,1][sort.of$eruptions brk

## Using the CRAN package 'strucchange'
require(strucchange)
sctest(waiting ~ eruptions, type = "Chow", point = brk, data = sort.of)
```

The results above should mirror exactly the results we achieved via manual calculation in the section above.

A Single Regression

If you wanted to run a single regression that allowed the marginal effect of eruption time on waiting time to vary before and after the breakpoint, you add a dummy variable and an interaction term. In R, it is relatively straightforward:

```
of$dummy = as.numeric(of$eruptions >= 3.25)
summary(lm(eruptions ~ waiting + I(dummy*waiting) + dummy, data = of))
```

*Be careful as you interpret the coefficient estimates because you must add a few of them together.

Share this:

Share

★ Like

Be the first to like this.

Related

- Surviving Graduate Econometrics with R: Fixed Effects Estimation -- 3 of 8
  In "Surviving Graduate Econometrics with R"
- Surviving Graduate Econometrics with R: Advanced Panel Data Methods -- 4 of 8
  In "Surviving Graduate Econometrics with R"
- Surviving Graduate Econometrics with R: The Basics -- 1 of 8
  In "Surviving Graduate Econometrics with R"

Posted on June 16, 2011 at 11:24 am in Econometrics with R | RSS feed | Reply | Trackback URL

Tags: R

One Comment to “The Chow test in R: A case study of Yellowstone’s Old Faithful Geyser”

Louis
December 12, 2013 at 2:30 pm

Hello Kevin,

How would your proceed in the case of heteroskadasticity? Do you use some kind of Wald test?

Thanks

Reply

Leave a Reply

Enter your comment here...

Tags

cluster-robust econometrics heteroskedasticity latex numpy parallel computing plots python r stata tex tikz

Calendar

June 2011

M	T	W	T	F	S	S
		1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30			

« May Jul »

Archives

- October 2012
- February 2012
- July 2011
- June 2011
- May 2011

Blogroll

- Documentation
- Plugins
- Suggest Ideas
- Support Forum
- Themes
- WordPress Blog
- WordPress Planet