

« Surviving Graduate Econometrics with R: Fixed Effects Estimation — 3 of 8 | Heteroskedasticity-Robust and Clustered Standard Errors in R »

## Surviving Graduate Econometrics with R: Advanced Panel Data Methods — 4 of 8

Some questions may arise when contemplating what model to use to empirically answer a question of interest, such as:

1. Is there unobserved-heterogeneity in my data sample? If so, is it time-invariant?
2. What variation in my data sample do I need to identify my coefficient of interest?
3. What is the data-generating process for my unobserved heterogeneity?

The questions above can be (loosely) translated into these more specific questions:

1. Should include fixed-effects (first-differenced, time-demeaned transformations, etc.) when I run my regression? Should I account for the unobserved heterogeneity using time dummy variables or individual dummy variables?
2. Is the variation I'm interested in between individuals or within individuals? This might conflict with your choice of time or individual dummy variables.
3. Can I use a random effects model?

That said, choosing a model for your panel data can be tricky. In what follows, I will offer some tools to help you answer some of these questions. The first part of this exercise will use the data `panel_hw.dta` (can be found [here](#)); the second part will use the data `wr-nevermar.dta` (can be found [here](#)).

### A Pooled OLS Regression

To review, let's load the data and run a model looking at voter participation rate as a function of a few explanatory variables and regional dummy variables (WNCentral, South, Border). `panel_hw.dta` is a panel data set where individual = "stcode" (state code) and time = "year". We are, then, pooling the data in the following regression.

STATA:

```
use panel_hw.dta
```

```
reg vaprte gsp midterm regdead WNCentral South Border
```

And then run an F-test on the joint significance of the included dummy variables:

```
test WNCentral South Border
```

R:

```
1 require(foreign)
2 voter = read.dta("/Users/kevingoulding/DATA/wr-nevermar.dta")
3
4 reg1 <- lm(vaprte ~ gsp + midterm + regdead + WNCentral + South + Border, data=voter)
```

Then run an F-test on the joint significance of the included regions:

```
1 require(car)
2 linearHypothesis(reg1, c("WNCentral", "South", "Border = 0"))
```

Similarly, this could be accomplished using the `plm` package (I recommend using this method).

```
1 reg1.pool <- plm(vaprte ~ gsp + midterm + regdead + WNCentral + South + Border,
2 data=voter, index = c("state", "year"), model = "pooling")
3 summary(reg1.pool)
4
5 # F-test
6 linearHypothesis(reg1.pool, c("WNCentral", "South", "Border = 0"), test="F")
```

### A Fixed Effects Regression

To review, let's load the data and run a model looking at voter participation rate as a function of a few explanatory variables and regional dummy variables (WNCentral, South, Border). `panel_hw.dta` is a panel data set where individual = "stcode" (state code) and time = "year". We are, then, pooling the data in the following regression.

STATA:

```
iis stcode
```

```
tis year
```

```
xtreg vaprte midterm gsp regdead WNCentral South Border, fe
```

In R, recall that we'll have to transform the data into a panel data form.

R:

```
1 require(plm)
2
3 # model is specified using "within" fixed effects.
4 reg1.fe <- plm(vaprte ~ gsp + mi 1 + Border,
5 data=voter, index = c("state", "ye
```

Follow "the Tarzan"

Get every new post delivered

Search this blog

Search...

Contributors



Goulding Kevin

Categories

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

Twitterfeed

RT @gappy3000: This post, apparently about #julialang and #pydata, explains why #rstats has become the standard of data analysis http:// ... 3 years ago

RT @justinwolffers: "If prediction markets are really as valuable as economists think, then..more experimentation could prove worthwhile. ... 3 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

```
6 | summary(reg1.fe)
```

to your Inbox.

Well, should we use the fixed effects model

Join 78 other followers

an run a test between the two:

```
1 | pFtest(reg1.fe, reg1.pool)
```

Or, we can test for individual fixed effects p

```
1 | plmtest(reg1.pool, effect = "indi
```

## The Random Effects Estimato

Build a website with WordPress.com

It could be, however, that the unobserved heterogeneity is uncorrelated with any of the regressors in all time periods — so called “random effects”. This would mean that if we did not account for these effects, we would still consistently estimate our coefficients, but their standard errors will be biased. To correct for this, we can use the random effects model, a form of Generalized Least Squares that accounts for the embedded serial correlation in the error terms caused by random effects.

STATA:

```
xtreg vaprte midterm gsp regdead WNCentral South Border, re
```

R:

```
1 | reg1.re <- plm(vaprte ~ gsp + midterm + regdead + WNCentral + South + Border,
2 | data=voter, index = c("state", "year"), model = "random")
3 | summary(reg1.re)
```

## Pooled OLS versus Random Effects

The Breush-Pagan LM test can be used to determine if you should use Random Effects model or pooled OLS. The null hypothesis is that the variance of the unobserved heterogeneity is zero, e.g.

$$H_0 = \sigma_\alpha^2 = 0$$

$$H_a = \sigma_\alpha^2 \neq 0$$

Failure to reject the null hypothesis implies that you will have more efficient estimates using OLS.

STATA:

```
xttest0
```

R:

```
1 | plmtest(reg1.pool, type="bp")
```

## Fixed Effects versus Random Effects

The Hausman test can help to determine if you should use Random Effects (RE) model or Fixed Effects (FE). Recall that a RE model is appropriate when the unobserved heterogeneity is uncorrelated with the regressors. The logic behind the Hausman test is that under the scenario that truth is RE, both the RE estimator and the FE estimator will be consistent (so you should opt to use the RE estimator because it is efficient). However, under the scenario that truth is FE, the RE estimator will be inconsistent — so you must use the FE estimator. The null hypothesis then, is that the unobserved heterogeneity  $\alpha_i$  and the regressors  $X_{it}$  are uncorrelated. Another way to think about it is that in the null hypothesis, the coefficient estimates of the two models are not statistically different. If you fail to reject the null hypothesis, this lends support for the use of the RE estimator. If the null is rejected, RE will produce biased coefficient estimates, so a FE model is preferred.

$$H_0 : \text{Corr}[X_{it}, \alpha_i] = 0$$

$$H_a : \text{Corr}[X_{it}, \alpha_i] \neq 0$$

STATA:

```
xtreg vaprte midterm gsp regdead WNCentral South Border, fe
estimates store fe
```

```
xtreg vaprte midterm gsp regdead WNCentral South Border, re
estimates store re
```

```
hausman fe re
```

R:

```
1 | phtest(reg1.fe, reg1.re)
```

## Some plots

The following examples use the data `wr-nevermar.dta`

Say we are interested in plotting the mean of the variable “nevermar” over time.

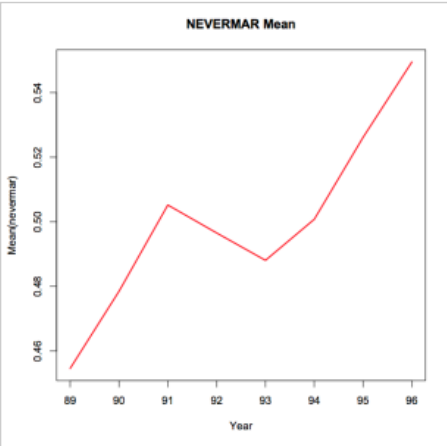
STATA:

```
egen meannevermar = mean(nevermar), by(year)
twoway (line meannevermar year, sort), ytitle(Mean--nevermar)
```

R:

```
1 | nmar <- read.dta(file="/Users/kevingoulding/DATA/wr-nevermar.dta")
2 |
3 | b1 <- as.matrix(tapply(nmar$nevermar, nmar$year, mean))
4 |
```

```
5 | plot(row.names(b1), b1, type="l", main="NEVERMAR Mean", xlab = "Year", ylab = "Mean(nevermar)", col="red", lwd=2)
```



About these ads



Share this:



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: The Basics -- 1 of 8](#)  
In "Surviving Graduate Econometrics with R"

[Surviving Graduate Econometrics with R: Difference-in-Differences Estimation -- 2 of 8](#)  
In "Surviving Graduate Econometrics with R"

Posted on May 27, 2011 at 10:58 am in [Surviving Graduate Econometrics with R](#) | [RSS feed](#) | [Reply](#) | [Trackback URL](#)

Tags: [R](#), [STATA](#)


3 Comments to “Surviving Graduate Econometrics with R: Advanced Panel Data Methods — 4 of 8”

- 

Bill Zhou

March 16, 2013 at 10:08 pm


Thank you!

Reply
- 

moein

July 3, 2013 at 11:23 am

Thank you for a great post! I'd like to go through these examples, but the links to the data files are not working. Could you update them?

Reply
- 

Kevin Goulding

July 3, 2013 at 12:44 pm

It's unlikely I'll have a chance to update the links. Google around, if you get lucky and find those files please post the link here. Thanks for reading!

Reply

Leave a Reply

Enter your comment here...

Tags

cluster-robust

econometrics

heteroskedasticity

latex

numpy

parallelcomputing

plots

python

r

stata

tex

tikz

Calendar

May 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 | 31 |    |    |    |    |    |

Jun »

Archives

October 2012

February 2012

July 2011

June 2011

May 2011

Blogroll

Documentation

Plugins

Suggest Ideas

Support Forum

Themes

WordPress Blog

WordPress Planet

