ADRC-S Jupyter Notebook

This notebook illustrates how Jupyter Notebooks could be used for the analysis of administrative social science data

The work is very exploratory.

Positive comments and brickbat are equally welcome.

or @profbigvern

Next Actions: Prepare for demonstration at ADRC-S Retreat 2016

Processing math: 100%

Author: Professor Vernon Gayle

Project: Reproducible Research with Administrative Social Science Data

Sub-project: Jupyter Notebooks / Social Science Data Analysis with Python

Latest Update: 22nd September, Stirling [Serious Windows 10 problems encountered]

Previous Updates:

21st September, ScotRail [Serious problems with Surface Pro]

19th September, ScotRail

12th September, University of Edinburgh

9th September, Stanstead Airport

2nd September 2016, University of Edinburgh

29th August 2016, UC Berkeley

My General Motivation

How can we draw on interdisciplinary expertise within ADRC-S (especially computer science and e-research) to enhance the analysis of administrative social science data?

How can we have better impact on research collaboration?

How can we improve computational reproducibility?

Are there tools we can leverage?

Can we learn some lessons from Open Science?

Overview

The Jupyter Notebook is an open source web application that facilitates the creation and sharing of documents that contain live code and supporting commentary in the form of explanatory text.

It is platform that can be used throughout the research process to organise and articulate elements of the social science workflow.

The Jupyter Notebook is open source and currently supports interactive data science and scientific computing across over 40 programming languages.

Obvious uses include data enabling and management, statistical analyses, simulations, and machine learning, but it is sufficiently flexible to be used more widely.

Jupyter Notebooks & Big Science

The detection of the gravitational waves is heralded as a major scientific discovery.

Here is a video of Fernando Perez demonstrating a Jupyter Notebook that includes data and analyses of the first observed gravitational waves which are from the LIGO collaborations

click on picture to run video

For more information on how this important work was rendered openly accessible by the LIGO collaboration see https://losc.ligo.org/s/events/GW150914/GW150914_tutorial.html .

Counting Galaxies in the Hubble Deep Field

This example uses the image-processing library scikit-image to identify galaxies in an image of the sky provided by the Hubble Space Telescope using a blob feature detection algorithm (an approach known as the determinant of Hessian).

After running the cell, we can explore the parameters of the detection algorithm to find galaxies of different sizes and prominences:

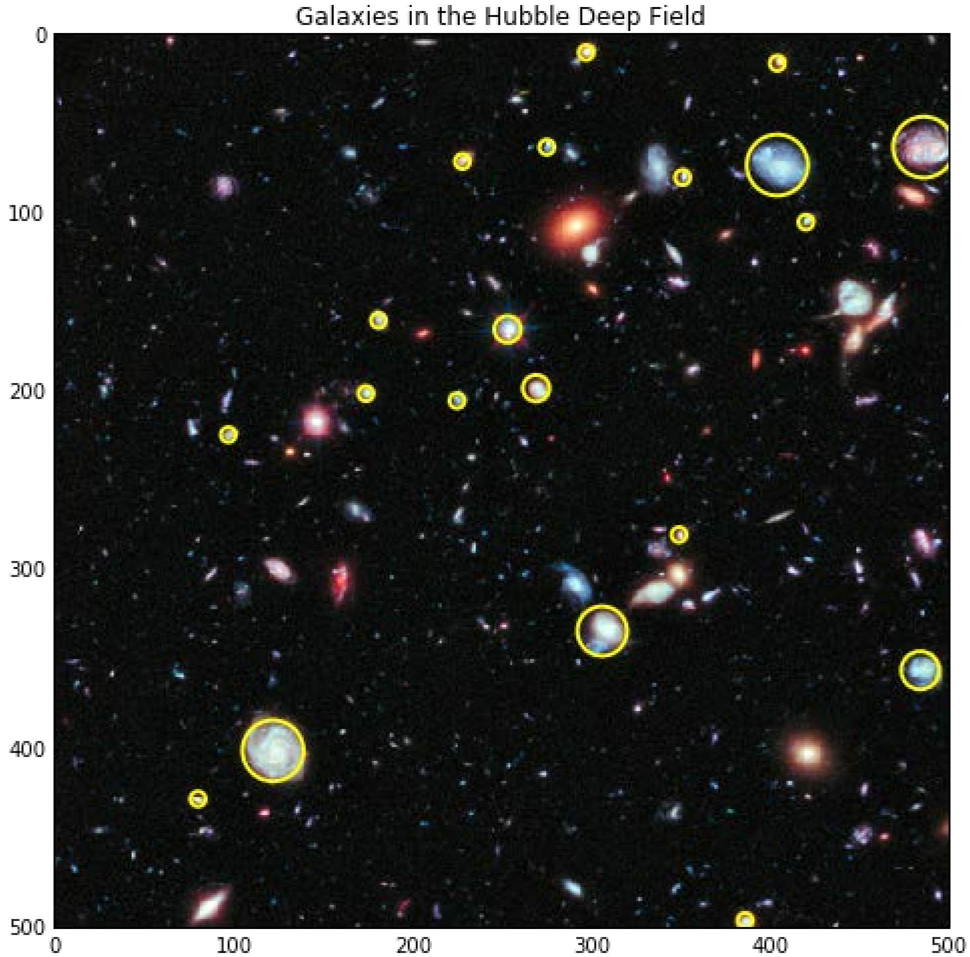
The max_sigma parameter determines the maximum size of the objects that will be identified

The threshold parameter can be reduced to detect less prominent objects

This example was provided in a live example hosted by the journal Nature

http://www.nature.com/news/ipython-interactive-demo-7.21492?article=1.16261

In [14]:
```
# Import matplotlib (plotting), skimage (image processing) and interact (us
er interfaces)
# This enables their use in the Notebook.
%matplotlib inline
from matplotlib import pyplot as plt

from skimage import data
from skimage.feature import blob_doh
from skimage.color import rgb2gray

from IPython.html.widgets import interact, fixed

# Extract the first 500px square of the Hubble Deep Field.
image = data.hubble_deep_field()[0:500, 0:500]
image_gray = rgb2gray(image)

def plot_blobs(max_sigma=30, threshold=0.1, gray=False):
    """
    Plot the image and the blobs that have been found.
    """
    blobs = blob_doh(image_gray, max_sigma=max_sigma, threshold=threshold)

    fig, ax = plt.subplots(figsize=(8,8))
    ax.set_title('Galaxies in the Hubble Deep Field')

    if gray:
        ax.imshow(image_gray, interpolation='nearest', cmap='gray_r')
        circle_color = 'red'
    else:
        ax.imshow(image, interpolation='nearest')
        circle_color = 'yellow'
    for blob in blobs:
        y, x, r = blob
        c = plt.Circle((x, y), r, color=circle_color, linewidth=2, fill=Fa
lse)
        ax.add_patch(c)

# Use interact to explore the galaxy detection algorithm.
interact(plot_blobs, max_sigma=(10, 40, 2), threshold=(0.005, 0.02, 0.001)
);

# The max_sigma parameter determines the maximum size of the objects that w
ill be identified

# The threshold parameter can be reduced to detect less prominent objects
```



Literate Computing

Fernando Perez says

Literate Computing is the weaving of a narrative directly into a live computation, interleaving text with code and results to construct a complete piece that relies equally on the textual explanations and the computational components, for the goals of communicating results in scientific computing and data analysis.

http://blog.fperez.org/

Why is it call Jupyter?

The computer languages Julia Python and R almost spell *JuPyteR*

The original authors Fernando Perez and Brian Granger are physicists.

Galileo's notebooks are an inportant early milestone (or deliverable) in open science.

Galileo's notes directly integrated his data (drawings of Jupiter and its moons), key metadata (timing of each observation, weather, telescope properties), and text (descriptions of methods, analysis, and conclusions) see https://www.authorea.com/users/3/articles/6316/_show_article .

Two obvious examples are *Sidereus Nuncius* and *la vacchetta* (which literally means the small cow or calf, because of the leather of its binding).

Translations of these notebooks can be viewed here

http://www.dioi.org/galileo/galileo.htm

http://homepages.wmich.edu/~mcgrew/Siderius.pdf

Three Galilean moons of Jupiter are visible on the logo.

Some people may have come accross IPython Notebooks which are from an earlier stage in the evolutionary process.

Jupyter Notebooks and Administrative Social Science Data Analysis

Jupyter Notebooks have a number of attractive features

1. They facilitate easy documentation alongside research code (using Markdown)
2. They have good portability (notebooks are easy to share)
3. They are 'language agnostic' (i.e. analyses can be undertaken in over 40 programming languages)
4. They can produce rich visual outputs
5. They can leverage 'big data' research tools (e.g. python)
6. They can be used as integrated tools in teaching, training, knowledge exchange and research capacity building
7. They might better facilitates collaboration

These features are demonstrated in this Jupyter Notebook.

Structure of a Jupyter Notebook

A Jupyter Notebook is made up of cells.

A cell can contain either

i. live research code (e.g. Stata syntax) that can be executed

ii. text comments that form the documentation of the research workflow

iii. cells that contain the results of data analyses

In addition to running your code the Notebook frontend stores code and output, together with markdown notes, in an editable document called a notebook. When you save it, this is sent from your browser to the notebook server, which saves it on disk as a JSON file with a .ipynb extension.

Documentation in the Jupyter Notebook

Writing comments and documenting research code in a Jupyter Notebook is relatively easy.

Notebooks use *Markdown*, which is explained below.

Comments can also be included within research code. For example a comment within Stata code can be written by starting with the familiar asterisk or star symbol.

Markdown

Markdown is an easy way to write documents.

It is written in what computer geeks like to call 'plaintext'. It is the sort of text that we used to writing and seeing.

Plaintext is just the regular alphabet plus a few other familiar symbols (for example the asterisk *).

Unlike cumbersome word processing applications, text written in Markdown can be easily shared between computers.

It’s quickly becoming the writing standard in some academic areas and in science.

Websites like GitHub and reddit use Markdown to style their comments.

Here is a summary of *Markdown* codes https://en.wikipedia.org/wiki/Markdown#Example

If you have half an hour you can learn *Markdown* here http://www.markdowntutorial.com/ .

Examples

Example 1 'Sitting in the Swivel Chair' (undertaking data analyses in different languages)

Example 2 'Rich Visual Output' (plots, pictures, videos and maps)

Example 3 'Widgets'

Example 4 'Using Magic Cells'

Example 5 'A Really Gnarly Example in Stata with a Complex Model'

Example 1 'Sitting in the Swivel Chair'

The Notebook can work with over 40 programming languages.

This includes data analysis software such as Stata and R and languages such as Python that are popular in some areas of data science.

Moving between languages with ease has been likened to sitting in an office chair and swivelling between three different computers.

The example below demonstrates some social science data analysis (estimating a logistic regression model) within the Jupyter Notebook.

To illustrate that the Jupyter Notebook is 'language agnostic', the analysis is undertaken using Stata, R and Python.

Jupyter Notebooks use different 'kernels' to work in different languages.

Stata Analysis

This is an example of running Stata from within the Jupyter notebook.

The live research code in each cell is Stata syntax.

Comments are written in cells using markdown.

WARNING

We would usually use a Stata Kernel to run Stata.

This is a new Windows 10 Machine and there is a conflict

Therefore in this current example we will use MAGIC - which will be explained later in the notebook

In [17]:
```
import ipystata
```

The cell above imports ipystata to run Stata

The cell below imports the data files

In [21]:
```
%%stata -o wemp_df
use "C:\Users\vgayle\wemp.dta", clear
codebook, compact
```

```
Variable    Obs Unique      Mean  Min   Max  Label
---------------------------------------------------------------------------
---------
case       1580    155  517.7411    1  1003
femp       1580      2  .6455696    0     1
mune       1580      2  .0740506    0     1
time       1580     14       7.2    0    13
und1       1580      2  .0746835    0     1
und5       1580      2  .2974684    0     1
age        1580     43  36.01013   18    60
---------------------------------------------------------------------------
---------
```

The data mirror a real example of data analysed in Davies et al. (1992).

The dataset is a panel of 155 married women.

Davies, Richard B., Peter Elias, and Roger Penn. "The relationship between a husband's unemployment and his wife's participation in the labour force." Oxford Bulletin of Economics and Statistics 54.2 (1992): 145-171.

Estimating a simple logit (logistic regression) model using Stata

The outcome variable is femp (female unemployed 0 or 1)

The explanatory variables are mune (husband unemployed 0,1) und5 (couple have a child under age 5)

In [22]:
```
%%stata -o wemp_df
logit femp mune und5
```

```
Iteration 0:   log likelihood = -1027.2309
Iteration 1:   log likelihood = -879.88806
Iteration 2:   log likelihood = -878.68101
Iteration 3:   log likelihood = -878.67998
```

```
Iteration 4:   log likelihood = -878.67998

Logistic regression                              Number of obs     =      1,58
0
                                                LR chi2(2)        =     297.10
                                                Prob > chi2       =     0.0000
Log likelihood = -878.67998                      Pseudo R2         =     0.14
46

----------------------------------------------------------------------------
--------
        femp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interva
l]
-------------+--------------------------------------------------------------
--------
        mune |  -1.703308   .2358489    -7.22   0.000    -2.165563   -1.2410
53
        und5 |  -1.733521   .1221909   -14.19   0.000    -1.973011   -1.494
031
       _cons |   1.306829   .0744154    17.56   0.000     1.160978    1.4526
81
----------------------------------------------------------------------------
--------
```

R Analysis

Switch Kernel to R Stata < Menu kernel - change kernel>

getting the libraries for R

In [1]:
```
library(foreign)
library(survey)
```

```
Loading required package: grid
Loading required package: Matrix
Loading required package: survival

Attaching package: 'survey'

The following object is masked from 'package:graphics':

    dotchart
```

getting the Stata data file (.dta) into an object called "mydata" then summarizing the dataset

In [2]:
```
mydata <- read.dta("C:/Users/vgayle/wemp.dta")
summary(mydata)
```

```
      case            femp             mune              time
 Min.   :   1.0   Min.   :0.0000   Min.   :0.00000   Min.   : 0.0
 1st Qu.: 274.0   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.: 4.0
 Median : 538.0   Median :1.0000   Median :0.00000   Median : 8.0
```

```
 Mean   : 517.7   Mean   :0.6456   Mean   :0.07405   Mean   : 7.2
 3rd Qu.: 753.0   3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:11.0
 Max.   :1003.0   Max.   :1.0000   Max.   :1.00000   Max.   :13.0
      und1              und5             age
 Min.   :0.00000   Min.   :0.0000   Min.   :18.00
 1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:29.00
 Median :0.00000   Median :0.0000   Median :35.00
 Mean   :0.07468   Mean   :0.2975   Mean   :36.01
 3rd Qu.:0.00000   3rd Qu.:1.0000   3rd Qu.:43.00
 Max.   :1.00000   Max.   :1.0000   Max.   :60.00
```

estimating the logit model and sending it to the object "mylogit"

In [3]:
```
mylogit <- glm(femp ~ mune + und5, data = mydata, family = "binomial")

summary(mylogit)
```

```
Call:
glm(formula = femp ~ mune + und5, family = "binomial", data = mydata)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-1.7586  -1.0024   0.6922   0.6922   2.1177  

Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept)  1.30683    0.07442  17.561  < 2e-16 ***
mune        -1.70331    0.23585  -7.222 5.12e-13 ***
und5        -1.73352    0.12219 -14.187  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2054.5  on 1579  degrees of freedom
Residual deviance: 1757.4  on 1577  degrees of freedom
AIC: 1763.4

Number of Fisher Scoring iterations: 4
```

the R results are the same as the Stata results - thankfully

Python Analysis

Switch Kernel to Stata < Menu kernel - change kernel>

In [1]:
```
import pandas as pd
```

construct the data frame "df" reading in the data from an Excel (xlsx) file. It could also be read in from a csv file.

In [2]:
```
df = pd.read_excel("C:/Users/vgayle/wemp.xlsx")
```

```
df.head()
```

Out[2]:

	case	femp	mune	time	und1	und5	age
0	1	1	0	10	1	1	23
1	1	0	0	11	0	1	24
2	1	0	0	12	0	1	25
3	1	0	0	13	0	1	26
4	6	1	0	0	0	0	42

Python is more general purpose and not primarily orientated towards social science data analysis. Therefore some things are a little more fiddly. For example we must set a constant for all case (int=1).

In [3]:
```
df['Int']=1
```

examining the data in the data frame "df"

In [ ]:
```
df.head()
```

import the package "statsmodels"

In [4]:
```
import statsmodels.api as sm
```

estimate a logistic regression model the independent variables are "mune" "und5" and "int" the outcome variable is "femp"

In [5]:
```
independentVar = ['mune', 'und5', 'Int']
logReg = sm.Logit(df['femp'] , df[independentVar])
answer = logReg.fit()
```

```
Optimization terminated successfully.
         Current function value: 0.556127
         Iterations 5
```

the results are in the oject "answer"

In [6]:
```
answer.summary()
```

Out[6]:

Logit Regression Results

Dep. Variable:	femp	**No. Observations:**	1580
Model:	Logit	**Df Residuals:**	1577
Method:	MLE	**Df Model:**	2
Date:	Thu, 22 Sep 2016	**Pseudo R-squ.:**	0.1446

Time:	17:29:18	**Log-Likelihood:**	-878.68
converged:	True	**LL-Null:**	-1027.2
		LLR p-value:	3.056e-65

	coef	std err	z	P>|z|	[95.0% Conf. Int.]
mune	-1.7033	0.236	-7.222	0.000	-2.166 -1.241
und5	-1.7335	0.122	-14.187	0.000	-1.973 -1.494
Int	1.3068	0.074	17.561	0.000	1.161 1.453

The result are the same in Python, R and Stata - Phew!

Summary

This example was designed to demonstrate that Jupyter Notebooks are language agnostic.

Stata is the primary data analysis software package at the ADRC-Scotland. From time to time there may be a requirement to used other data analysis tools. The language agnostic aspects of Jupyter Notebooks mean that they could be an appropriate unified environment in which to undertake research analyses using alternative software packages and languages.

Example 2 'Rich Visual Output'

There are several visual features of Jupyter Notebooks that make them attractive as environments in which to undertake analyses of administrative social science data and which might also enrich teaching, training and kwonledge exchange and research capacity buidling activities.

Here is a smart wee example that shows the flexibility of graphing within the Jupyter Notebook.

This wee example runs in Python. You MUST have the Python kernel running.

Make sure you have the wemp data in the the "df" data frame.

df.head()

if you can't see the wemp data you need to go back up and reload it in the Python part of the Example 1.

we will now construct a new variable called *age2* which is the woman's age squared

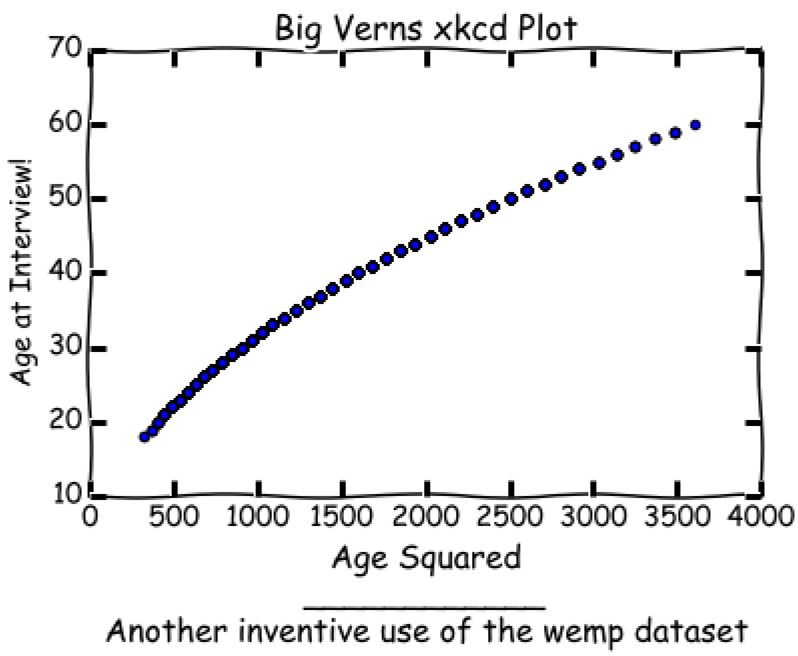
In [7]:
```
df['age2'] = df.age * df.age
```

this is an example of a graph in the style of a xkcd comic

In [8]:
```
%pylab inline
plt.xkcd()  # Yes...
fig1 = plt.figure(1)
ax1 = fig1.add_subplot(111)
ax1.scatter(df['age2'], df['age'],)
plt.title('Big Verns xkcd Plot')
plt.ylabel('Age at Interview!')
plt.xlabel("Age Squared\n"
          "____________\n"
           "Another inventive use of the wemp dataset", size=16)
```

```
Populating the interactive namespace from numpy and matplotlib
```

Out[8]:
```
<matplotlib.text.Text at 0x194e766de10>
```



see http://nbviewer.jupyter.org/url/jakevdp.github.com/downloads/notebooks/XKCD_plots.ipynb

Inserting a picture into the notebook (within a markdown cell)

A man and a bear...

Inserting a video into the notebook (within a markdown cell)

In [14]:
```
from IPython.display import YouTubeVideo
YouTubeVideo('p47tetYy7co')
```

Out[14]:

to render cell contents as LaTeX

In [9]:
```
%%latex
\begin{align}
a = \frac{1}{2} && b= \frac{1}{2} && c = \frac{1}{4} \\
\end{align}
```

a = 12 b = 12 c = 14

In [10]:
```
%%latex
$e^{i\pi} + 1 = 0
$
```

$e^{i\pi} + 1 = 0$

Code can produce rich output such as images, videos, LaTeX, and JavaScript.

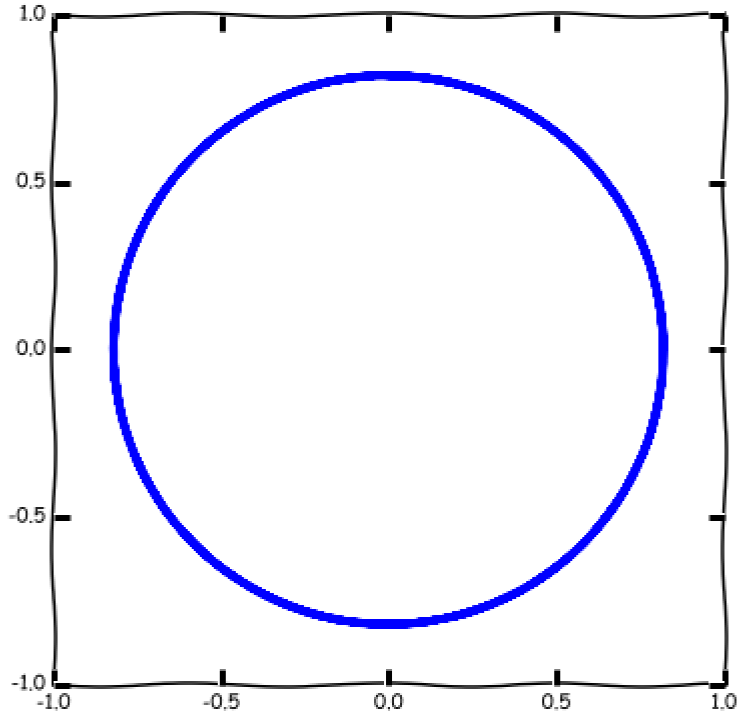
Example 3 'Widgets'

Widgets are clever devices that can be included in Notebooks to help users visualize and control changes in the data. Widgets may be useful in teaching and training because users can easily see how changing inputs to something impacts on the results.

An Interesting Wee Widget

In [15]:

```
%matplotlib inline
import matplotlib.pyplot as plt
from numpy import pi, exp, real, imag, linspace
from ipywidgets import interact

def f(t,a=1,b=6,c=-14,d=0):
    return exp(a*1j*t) - exp(b*1j*t)/2 + 1j*exp(c*1j*t)/3 + exp(d*1j*t)/
4

def plot_swirly(a=1,b=6,c=-14,d=0):
    t = linspace(0, 2*pi, 1000)
    ft = f(t,a,b,c,d)
    plt.plot(real(ft), imag(ft))

    # These two lines make the aspect ratio square
    fig = plt.gcf()
    fig.set_size_inches(6, 6, forward='True')

interact(plot_swirly,a=(-20,20),b=(-20,20),c=(-20,20),d=(-20,20));
```



Using an open street map

I've recently moved to a more commodious office in Buccleuch Place. Here is an example of an open source map on my new *hood*.

In [16]:
```
from ipyleaflet import Map
Map(center=[55.942535, -3.187269], zoom=20)
```

It is possible to zoom in and out on this map!

Example 4 'Using Magic Cells'

Any cell that starts with a % symbol or two symbols %% is a 'magic' cell (honestly that is what they are called)

Here are a list of some of the standard 'magics'

In [22]:
```
%lsmagic
```

Out[22]:
```
Available line magics:
%alias  %alias_magic  %autocall  %automagic  %autosave  %bookmark  %cd  %cl
ear  %cls  %colors  %config  %connect_info  %copy  %ddir  %debug  %dhist  %
dirs  %doctest_mode  %echo  %ed  %edit  %env  %gui  %hist  %history  %insta
```

```
ll_default_config  %install_ext  %install_profiles  %killbgscripts  %ldir  
%less  %load  %load_ext  %loadpy  %logoff  %logon  %logstart  %logstate  %l
ogstop  %ls  %lsmagic  %macro  %magic  %matplotlib  %mkdir  %more  %noteboo
k  %page  %pastebin  %pdb  %pdef  %pdoc  %pfile  %pinfo  %pinfo2  %popd  %p
print  %precision  %profile  %prun  %psearch  %psource  %pushd  %pwd  %pyca
t  %pylab  %qtconsole  %quickref  %recall  %rehashx  %reload_ext  %ren  %re
p  %rerun  %reset  %reset_selective  %rmdir  %run  %save  %sc  %set_env  %s
tore  %sx  %system  %tb  %time  %timeit  %unalias  %unload_ext  %who  %who_
ls  %whos  %xdel  %xmode

Available cell magics:
%%!  %%HTML  %%SVG  %%bash  %%capture  %%cmd  %%debug  %%file  %%html  %%ja
vascript  %%latex  %%perl  %%prun  %%pypy  %%python  %%python2  %%python3  
%%ruby  %%script  %%sh  %%stata  %%svg  %%sx  %%system  %%time  %%timeit  %
%writefile

Automagic is ON, % prefix IS NOT needed for line magics.
```

In [23]:
```
%time print ("How long does this cell take to run?")
```

```
How long does this cell take to run?
Wall time: 0 ns
```

In Example 1 we used a non-standard magic to run Stata which was written by Ties de Kok

---

Example 5 'A Really Gnarly Example in Stata with a Complex Model'

In the example we try to 'stress out' the Jupyter Notebook with by getting it to do something difficult.

I have chosen to estimate a dynamic probit random effects model see Stewart, M.B., 2006. redprob-: A Stata program for the Heckman estimator of the random effects dynamic probit model. *http://www2.warwick.ac.uk/fac/soc/economics/staff/mstewart/stata/redprobnote.pdf*

We used this model in a research paper some years ago see Boyle, P., Feng, Z. and Gayle, V., 2009. A new look at family migration and women's employment status. Journal of Marriage and Family, 71(2), pp.417-431. *http://www.jstor.org/stable/40262889?seq=1#page_scan_tab_contents*

The dynamic probit random effects model...

1. This is not a standard Stata model and the relevant program will have to be imported.
2. The relevant program is not available through the Statistical Software Components (SSC) archive, additional work will have to be done.
3. The model is estimated on panel data and is more computationally intense than most statistical models.

The first task is to import ipystata (the Magic)

In [47]:
```
import ipystata
```

The next task is to change the directory. We will need to import some Stata .ado files from the original author Professor Mark Stewart who was at Warwick University when he wrote the files.

In [48]:
```
import os
os.getcwd()
```

Out[48]:
```
'C:\\Program Files (x86)\\Stata14\\ado\\base\\r'
```

In Python the slashes are forward e.g. "/" and not "\"

In [49]:
```
os.chdir('C:/Program Files (x86)/Stata14/ado/base/r')
```

In [50]:
```
ls
```

```
 Volume in drive C is Windows
 Volume Serial Number is D096-0D61

 Directory of C:\Program Files (x86)\Stata14\ado\base\r

21/09/2016  17:58    <DIR>          .
21/09/2016  17:58    <DIR>          ..
22/09/2015  14:38             4,740 random_number_functions.sthlp
22/09/2015  14:38               376 range.ado
22/09/2015  14:38             1,793 range.dlg
22/09/2015  14:38               347 range.mata
22/09/2015  14:38             1,369 range.sthlp
22/09/2015  14:38               445 rangen.mata
22/09/2015  14:38               228 rank.mata
22/09/2015  14:38               671 rank_from_singular_values.mata
22/09/2015  14:38             3,477 ranksum.ado
22/09/2015  14:38             1,331 ranksum.dlg
22/09/2015  14:38             8,829 ranksum.sthlp
22/09/2015  14:38               152 ratio.ado
22/09/2015  14:38             6,792 ratio.dlg
22/09/2015  14:38             9,567 ratio.sthlp
22/09/2015  14:38             1,478 ratio_postestimation.sthlp
22/09/2015  14:38             2,683 rcap_options.sthlp
22/09/2015  14:38             5,475 rchart.ado
22/09/2015  14:38             3,110 rchart.dlg
22/09/2015  14:38             3,969 rchart_7.ado
22/09/2015  14:38               498 rchi2.mata
22/09/2015  14:38               329 rcof.ado
22/09/2015  14:38             2,986 rcof.sthlp
22/09/2015  14:38             1,639 recast.ado
22/09/2015  14:38             2,714 recast.sthlp
22/09/2015  14:38            12,871 recode.ado
22/09/2015  14:38             5,161 recode.dlg
22/09/2015  14:38             9,244 recode.sthlp
22/09/2015  14:38             3,112 recode_rules.sthlp
22/09/2015  14:38             3,257 record24.png
```

```
22/09/2015  14:38             3,808 recordToggled32.png
22/09/2015  14:38             5,771 rectangle_g.class
22/09/2015  14:38             3,714 redo24.png
22/09/2015  14:38             4,105 refresh24.png
22/09/2015  14:38               164 reg.ado
22/09/2015  14:38            29,062 reg3.ado
22/09/2015  14:38             5,948 reg3.dlg
22/09/2015  14:38            21,454 reg3.sthlp
22/09/2015  14:38            26,972 reg3_10.ado
22/09/2015  14:38             4,085 reg3_p.ado
22/09/2015  14:38             2,614 reg3_p.dlg
22/09/2015  14:38             6,599 reg3_postestimation.sthlp
22/09/2015  14:38             2,502 regdw.ado
22/09/2015  14:38            17,858 region_options.sthlp
22/09/2015  14:38               165 regr.ado
22/09/2015  14:38               166 regre.ado
22/09/2015  14:38             4,159 regre_p2.ado
22/09/2015  14:38               167 regres.ado
22/09/2015  14:38             6,624 regres_p.ado
22/09/2015  14:38             7,604 regres_p.dlg
22/09/2015  14:38             2,723 regress.ado
22/09/2015  14:38             5,173 regress.dlg
22/09/2015  14:38            15,385 regress.sthlp
22/09/2015  14:38             1,559 regress_estat.ado
22/09/2015  14:38            20,628 regress_estat.dlg
22/09/2015  14:38            31,118 regress_postestimation.sthlp
22/09/2015  14:38            18,184 regress_postestimation_plots.sthlp
22/09/2015  14:38            11,628 regress_postestimationts.sthlp
22/09/2015  14:38             3,515 regriv_p.ado
22/09/2015  14:38             5,071 regriv_p.dlg
22/09/2015  14:38               363 regstats.ihlp
22/09/2015  14:38               363 regstats_margins.ihlp
22/09/2015  14:38               642 relative_posn.class
22/09/2015  14:38             3,033 relativesize.sthlp
22/09/2015  14:38             2,994 relink16.png
22/09/2015  14:38               514 relsize.class
22/09/2015  14:38             1,054 remap.ado
22/09/2015  14:38               158 remap.sthlp
22/09/2015  14:38               722 remove16.png
22/09/2015  14:38               108 ren.ado
22/09/2015  14:38               109 rena.ado
22/09/2015  14:38               110 renam.ado
22/09/2015  14:38            98,695 rename.ado
22/09/2015  14:38             9,360 rename.dlg
22/09/2015  14:38             1,652 rename.sthlp
22/09/2015  14:38            21,827 rename_group.sthlp
22/09/2015  14:38             2,876 rename16.png
22/09/2015  14:38             3,827 renamegraph24.png
22/09/2015  14:38             1,266 renamevarno.ado
22/09/2015  14:38               508 renpfix.ado
22/09/2015  14:38               259 renpfix.sthlp
22/09/2015  14:38                69 repeat.ado
22/09/2015  14:38               195 repeat.sthlp
22/09/2015  14:38             5,980 repeated_options.sthlp
22/09/2015  14:38             1,051 replace.dlg
22/09/2015  14:38             3,277 replace_dta.dlg
22/09/2015  14:38            42,456 reshape.ado
```

```
22/09/2015  14:38             3,650 reshape.dlg
22/09/2015  14:38            10,751 reshape.sthlp
22/09/2015  14:38            30,191 reshape_10.ado
22/09/2015  14:38             1,176 reshape_example.sthlp
22/09/2015  14:38             4,439 resources.sthlp
22/09/2015  14:38               687 restore16.png
22/09/2015  14:38             3,400 results24.png
22/09/2015  14:38             1,347 reswords.sthlp
22/09/2015  14:38             9,795 return.sthlp
22/09/2015  14:38               875 return_list.dlg
22/09/2015  14:38             1,047 reventries.sthlp
22/09/2015  14:38             2,320 review.sthlp
22/09/2015  14:38               208 revorder.mata
22/09/2015  14:38             5,123 rhelp_alias.maint
22/09/2015  14:38             1,198 ringposstyle.sthlp
22/09/2015  14:38               250 rmdir.mata
22/09/2015  14:38             1,349 rmdir.sthlp
22/09/2015  14:38             1,477 rmsg.sthlp
22/09/2015  14:38            47,324 robust.mata
22/09/2015  14:38             4,701 robvar.ado
22/09/2015  14:38               911 robvar.dlg
22/09/2015  14:38             5,978 roc.sthlp
22/09/2015  14:38            13,193 roccomp.ado
22/09/2015  14:38             7,785 roccomp.dlg
22/09/2015  14:38            10,944 roccomp.sthlp
22/09/2015  14:38            27,265 roccomp_7.ado
22/09/2015  14:38            12,213 roccomp_8.ado
22/09/2015  14:38               786 rocf_lf.ado
22/09/2015  14:38            14,619 rocfit.ado
22/09/2015  14:38             3,721 rocfit.dlg
22/09/2015  14:38             7,734 rocfit.sthlp
22/09/2015  14:38            13,209 rocfit_8.ado
22/09/2015  14:38             5,554 rocfit_postestimation.sthlp
22/09/2015  14:38             4,240 rocgold.ado
22/09/2015  14:38             6,888 rocgold.dlg
22/09/2015  14:38             6,737 rocplot.ado
22/09/2015  14:38             2,446 rocplot.dlg
22/09/2015  14:38             5,869 rocplot_7.ado
22/09/2015  14:38            70,103 rocreg.ado
22/09/2015  14:38            16,111 rocreg.dlg
22/09/2015  14:38            23,693 rocreg.sthlp
22/09/2015  14:38             1,834 rocreg_estat.ado
22/09/2015  14:38             4,596 rocreg_estat.dlg
22/09/2015  14:38             2,841 rocreg_lf2.ado
22/09/2015  14:38            35,205 rocreg_p.ado
22/09/2015  14:38             6,927 rocreg_p.dlg
22/09/2015  14:38             9,990 rocreg_postestimation.sthlp
22/09/2015  14:38           102,037 rocregplot.ado
22/09/2015  14:38            57,058 rocregplot.dlg
22/09/2015  14:38             9,803 rocregplot.sthlp
22/09/2015  14:38             4,380 rocregplot_at.dlg
22/09/2015  14:38            18,010 rocregstat.ado
22/09/2015  14:38            20,387 roctab.ado
22/09/2015  14:38             5,479 roctab.dlg
22/09/2015  14:38             9,020 roctab.sthlp
22/09/2015  14:38            18,607 roctab_7.ado
22/09/2015  14:38            14,621 rolling.ado
```

```
22/09/2015  14:38             6,613 rolling.dlg
22/09/2015  14:38             8,050 rolling.sthlp
22/09/2015  14:38            12,858 rologit.ado
22/09/2015  14:38             4,655 rologit.dlg
22/09/2015  14:38            14,690 rologit.sthlp
22/09/2015  14:38             2,741 rologit_p.ado
22/09/2015  14:38             1,192 rologit_p.dlg
22/09/2015  14:38             5,501 rologit_postestimation.sthlp
22/09/2015  14:38                87 rot.ado
22/09/2015  14:38                88 rota.ado
22/09/2015  14:38                89 rotat.ado
22/09/2015  14:38               514 rotate.ado
22/09/2015  14:38            29,384 rotate.dlg
22/09/2015  14:38            15,884 rotate.sthlp
22/09/2015  14:38             1,476 rotate_criteria_table.ihlp
22/09/2015  14:38               576 rotate_left16.png
22/09/2015  14:38             7,563 rotate_opt_opts.idlg
22/09/2015  14:38               570 rotate_right16.png
22/09/2015  14:38            39,919 rotatemat.ado
22/09/2015  14:38            24,952 rotatemat.dlg
22/09/2015  14:38            18,615 rotatemat.sthlp
22/09/2015  14:38               128 rowmax.mata
22/09/2015  14:38               128 rowmin.mata
22/09/2015  14:38               194 rowscalefactors.mata
22/09/2015  14:38             1,398 rpt_xt.idlg
22/09/2015  14:38             1,610 rpt_xt_eform.idlg
22/09/2015  14:38             1,686 rpt_xt_irr.idlg
22/09/2015  14:38             1,680 rpt_xt_or.idlg
22/09/2015  14:38             6,149 rreg.ado
22/09/2015  14:38             3,802 rreg.dlg
22/09/2015  14:38             6,438 rreg.sthlp
22/09/2015  14:38             3,458 rreg_p.ado
22/09/2015  14:38             1,271 rreg_p.dlg
22/09/2015  14:38             3,942 rreg_postestimation.sthlp
22/09/2015  14:38             3,947 rspike_options.sthlp
22/09/2015  14:38             3,526 run24.png
22/09/2015  14:38             5,055 runtest.ado
22/09/2015  14:38             3,694 runtest.dlg
22/09/2015  14:38             4,825 runtest.sthlp
22/09/2015  14:38               650 rvfplot.ado
22/09/2015  14:38             1,040 rvfplot.dlg
22/09/2015  14:38               337 rvfplot_7.ado
22/09/2015  14:38             1,189 rvpplot.ado
22/09/2015  14:38             1,074 rvpplot.dlg
22/09/2015  14:38               803 rvpplot_7.ado
             188 File(s)      1,536,779 bytes
               2 Dir(s)  210,221,621,248 bytes free
```

In [55]:

```
%%stata
copy http://www2.warwick.ac.uk/fac/soc/economics/staff/mstewart/stata/redpr
ob.ado redprob.ado
```

```
> .ado redprob.ado
```

In [56]:

```
%%stata
copy http://www2.warwick.ac.uk/fac/soc/economics/staff/mstewart/stata/redpm
od ll.ado redpmod ll.ado
```

```
> _ll.ado redpmod_ll.ado
```

In [57]:
```
%%stata
copy http://www2.warwick.ac.uk/fac/soc/economics/staff/mstewart/stata/redpr
ob.hlp redprob.hlp
```

```
> .hlp redprob.hlp
```

In [78]:
```
%%stata
help redprob
```

The redprob program is now working.

The next step is to get the data that Stewart (2006) uses in his paper. The data are women in the US National Longitudinal Survey.

In [60]:
```
%%stata
use "http://www.stata-press.com/data/r8/union.dta", clear
codebook, compact
tab year
```

```
(NLS Women 14-24 in 1968)

Variable     Obs Unique      Mean  Min   Max  Label
--------------------------------------------------------------------------
---------
idcode     26200   4434  2611.582    1  5159  NLS id
year       26200     12  79.47137   70    88  interview year
age        26200     31  30.43221   16    46  age in current year
grade      26200     19  12.76145    0    18  current grade completed
not_smsa   26200      2  .2837023    0     1  1 if not SMSA
south      26200      2  .4130153    0     1  1 if south
union      26200      2  .2217939    0     1  1 if union
t0         26200     12  9.471374    0    18
southXt    26200     12   3.96874    0    18
black      26200      2   .274542    0     1  race black
--------------------------------------------------------------------------
---------

  interview |
       year |      Freq.     Percent        Cum.
------------+-----------------------------------
         70 |      1,658        6.33        6.33
         71 |      1,793        6.84       13.17
         72 |      1,958        7.47       20.65
         73 |      2,057        7.85       28.50
         77 |      2,670       10.19       38.69
         78 |      1,993        7.61       46.29
         80 |      2,028        7.74       54.03
         82 |      2,408        9.19       63.23
         83 |      2,194        8.37       71.60
         85 |      2,488        9.50       81.10
         87 |      2,536        9.68       90.77
```

```
        88 |      2,417        9.23      100.00
------------+-----------------------------------
      Total |     26,200      100.00
```

Stewart deletes the years 1978 and 1983

In [62]:
```
%%stata
drop if year<78
drop if year==83
```

```
(10,136 observations deleted)
(2,194 observations deleted)
```

In [63]:
```
%%stata
tab year, missing
```

```
  interview |
       year |      Freq.     Percent        Cum.
------------+-----------------------------------
         78 |      1,993       14.37       14.37
         80 |      2,028       14.62       28.99
         82 |      2,408       17.36       46.35
         85 |      2,488       17.94       64.29
         87 |      2,536       18.28       82.57
         88 |      2,417       17.43      100.00
------------+-----------------------------------
      Total |     13,870      100.00
```

Stewart (2006) treats this a panel with biennial data collection.

In [65]:
```
%%stata
sort idcode year
by idcode: gen nwav=_N
keep if nwav==6
```

```
(9,076 observations deleted)
```

Stewart (2006) constructs a balanced panel of women with 6 waves of data.

We now need to

1. Sort the data into 'long' format
2. Construct an indicator for the survey wave
3. Construct a lag of the dependent variable

In [68]:
```
%%stata
sort idcode year

by idcode: gen tper=_n

by idcode: gen Lunion= union[_n-1]
```

```
(799 missing values generated)
```

To make the analyses complete we will first estimate two probit models

1. A standard probit model of the initial conditions
2. A (pooled) probit model of the panel data

In [72]:
```
%%stata
probit union age grade south not_smsa if tper==1
probit union Lunion age grade south not_smsa if tper>1
```

```
Iteration 0:   log likelihood =  -429.9493
Iteration 1:   log likelihood = -408.59398
Iteration 2:   log likelihood = -408.41782
Iteration 3:   log likelihood = -408.41778
Iteration 4:   log likelihood = -408.41778

Probit regression                               Number of obs     =        79
9
                                                LR chi2(4)        =     43.06
                                                Prob > chi2       =    0.0000
Log likelihood = -408.41778                      Pseudo R2         =    0.05
01

---------------------------------------------------------------------------
--------
       union |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interva
l]
-------------+-------------------------------------------------------------
--------
         age |    .039145   .0167366     2.34   0.019     .0063419    .07194
81
       grade |   .0023396   .0212189     0.11   0.912    -.0392487     .0439
28
       south |  -.5412888   .1069614    -5.06   0.000    -.7509294   -.3316
482
    not_smsa |  -.2957917   .1160209    -2.55   0.011    -.5231885   -.0683
949
       _cons |  -1.604578   .5747252    -2.79   0.005    -2.731019   -.4781
374
---------------------------------------------------------------------------
--------

Iteration 0:   log likelihood = -2332.4529
Iteration 1:   log likelihood = -1575.6051
Iteration 2:   log likelihood = -1573.5206
Iteration 3:   log likelihood =   -1573.52
Iteration 4:   log likelihood =   -1573.52

Probit regression                               Number of obs     =      3,99
5
                                                LR chi2(5)        =   1517.87
                                                Prob > chi2       =    0.0000
Log likelihood =   -1573.52                      Pseudo R2         =    0.32
54

---------------------------------------------------------------------------
--------
```

```
       union |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interva
l]
-------------+----------------------------------------------------------------
--------
      Lunion |   1.883741   .0525663    35.84   0.000     1.780713    1.9867
69
         age |  -.0085892    .005816    -1.48   0.140    -.0199884    .00280
99
       grade |  -.0148996   .0102944    -1.45   0.148    -.0350762     .0052
77
       south |  -.1641046   .0526153    -3.12   0.002    -.2672287   -.0609
805
    not_smsa |  -.0276199   .0558868    -0.49   0.621     -.137156    .08191
61
       _cons |  -.6896341   .2480495    -2.78   0.005    -1.175802   -.2034
661
```

Running this model will take a little time as Stata has to sweat it out. The model is computationally demanding and fitted using an iterative process.

In [80]:

```
%%stata
redprob union Lunion age grade south (age grade south not_smsa), i(idcode)
t(tper) quadrat(24)
```

```
Dependent variable = union
x-variables = Lunion age grade south
z-variables = age grade south not_smsa
Assumed lagged dependent variable = Lunion

Pooled Probit Model for t>1

Iteration 0:   log likelihood = -2332.4529
Iteration 1:   log likelihood = -1583.2027
Iteration 2:   log likelihood = -1573.6572
Iteration 3:   log likelihood = -1573.6423

Probit regression                                 Number of obs   =       399
5
                                                 LR chi2(4)      =    1517.62
                                                 Prob > chi2     =     0.0000
Log likelihood = -1573.6423                       Pseudo R2       =     0.32
53

------------------------------------------------------------------------------
--------
       union |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interva
l]
-------------+----------------------------------------------------------------
--------
      Lunion |   1.884933   .0525191    35.89   0.000     1.781997    1.9878
68
         age |  -.0086679   .0058129    -1.49   0.136    -.0200609    .00272
52
       grade |  -.0145251   .0102686    -1.41   0.157    -.0346513    .00560
11
```

```
      south |  -.1684511   .0518759    -3.25   0.001     -.270126   -.06677
62
      _cons |  -.6985983   .2473588    -2.82   0.005    -1.183413    -.213
784
------------------------------------------------------------------------------
--------

Probit Model for t=1

Iteration 0:   log likelihood =  -429.9493
Iteration 1:   log likelihood = -408.59398
Iteration 2:   log likelihood = -408.41782
Iteration 3:   log likelihood = -408.41778

Probit regression                                 Number of obs   =        79
9
                                                 LR chi2(4)      =      43.06
                                                 Prob > chi2     =     0.0000
Log likelihood = -408.41778                       Pseudo R2       =     0.05
01

------------------------------------------------------------------------------
--------
      union |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interva
l]
-------------+----------------------------------------------------------------
--------
        age |    .039145   .0167366     2.34   0.019     .0063419    .07194
81
      grade |   .0023396   .0212189     0.11   0.912    -.0392487     .0439
28
      south |  -.5412888   .1069614    -5.06   0.000    -.7509293   -.3316
482
   not_smsa |  -.2957917   .1160209    -2.55   0.011    -.5231885   -.0683
949
      _cons |  -1.604578   .5747252    -2.79   0.005    -2.731019   -.4781
374
------------------------------------------------------------------------------
--------

Iterations for full ML estimation

------------------------------------------------------------------------------
--------
Iteration 0:
Coefficient vector:
       union:     union:     union:     union:     union:    rfper1:
      Lunion        age      grade      south      _cons        age
r1   1.884933  -.0086679  -.0145251  -.1684511  -.6985983    .039145

      rfper1:    rfper1:    rfper1:    rfper1:  logitrho:    ltheta:
       grade      south   not_smsa      _cons      _cons      _cons
r1   .0023396  -.5412888  -.2957917  -1.604578        -.5          0

                                                  log likelihood = -1965.437
8
Gradient vector (length = 4399.339):
```

```
        union:     union:     union:     union:     union:    rfper1:
       Lunion        age      grade      south      _cons        age
r1  -197.6275   -4063.55  -1454.954  -57.47154    -108.98  -755.4063

       rfper1:    rfper1:    rfper1:    rfper1:  logitrho:    ltheta:
        grade      south   not_smsa      _cons      _cons      _cons
r1  -313.5414  -14.47751  -10.95182  -25.29652  -25.79784   6.520916
--------------------------------------------------------------------------
--------
Iteration 1:
Coefficient vector:
        union:     union:     union:     union:     union:    rfper1:
       Lunion        age      grade      south      _cons        age
r1    .949095  -.0185269  -.0356356  -.3582592   .1128691   .0234799

       rfper1:    rfper1:    rfper1:    rfper1:  logitrho:    ltheta:
        grade      south   not_smsa      _cons      _cons      _cons
r1   .0020188  -.7014822  -.3991665  -1.350927   .1012893   .4679759

                                                 log likelihood = -1881.169
4
Gradient vector (length = 1572.243):
        union:     union:     union:     union:     union:    rfper1:
       Lunion        age      grade      south      _cons        age
r1   4.968151   385.2985   143.9159   4.607742   11.10839   -1387.83

       rfper1:    rfper1:    rfper1:    rfper1:  logitrho:    ltheta:
        grade      south   not_smsa      _cons      _cons      _cons
r1  -608.5585  -25.10177  -15.73838  -47.22767  -4.082508  -55.43869
--------------------------------------------------------------------------
--------
Iteration 2:
Coefficient vector:
        union:     union:     union:     union:     union:    rfper1:
       Lunion        age      grade      south      _cons        age
r1   .5796523    -.02723  -.0505079  -.4769203   .5486543   .0150768

       rfper1:    rfper1:    rfper1:    rfper1:  logitrho:    ltheta:
        grade      south   not_smsa      _cons      _cons      _cons
r1  -.0007599  -.6422034  -.3594996  -1.189416   1.014107  -.2380066

                                                 log likelihood =  -1862.54
3
Gradient vector (length = 897.3282):
        union:     union:     union:     union:     union:    rfper1:
       Lunion        age      grade      south      _cons        age
r1  -9.335107  -841.6972  -305.9961  -9.207022  -23.02098   -44.7157

       rfper1:    rfper1:    rfper1:    rfper1:  logitrho:    ltheta:
        grade      south   not_smsa      _cons      _cons      _cons
r1  -13.07108  -3.281669  -2.403796   -1.11905  -14.82983    .489322
--------------------------------------------------------------------------
--------
Iteration 3:
Coefficient vector:
        union:     union:     union:     union:     union:    rfper1:
       Lunion        age      grade      south      _cons        age
```

```
r1   .6347999  -.0284517  -.0550332  -.4889118   .5577935   .0084914

      rfper1:    rfper1:    rfper1:    rfper1:  logitrho:    ltheta:
        grade      south   not_smsa      _cons      _cons      _cons
r1  -.0068787  -.7221963  -.4129703  -.9702891   .8386983  -.1488076

                                                  log likelihood = -1860.254
8
Gradient vector (length = 171.0635):
       union:     union:     union:     union:     union:    rfper1:
       Lunion        age      grade      south      _cons        age
r1   1.935219   157.6291   59.70585   1.969945   4.495149  -26.19572

      rfper1:    rfper1:    rfper1:    rfper1:  logitrho:    ltheta:
        grade      south   not_smsa      _cons      _cons      _cons
r1  -11.50062  -.5304433  -.2830986  -.8937808   1.852525  -.3065477
---------------------------------------------------------------------
--------
Iteration 4:
Coefficient vector:
       union:     union:     union:     union:     union:    rfper1:
       Lunion        age      grade      south      _cons        age
r1   .6346962  -.0285705  -.0539338  -.4882501    .562863   .0081167

      rfper1:    rfper1:    rfper1:    rfper1:  logitrho:    ltheta:
        grade      south   not_smsa      _cons      _cons      _cons
r1  -.0064347  -.7261158  -.4152495  -.9597561   .8446825  -.1457469

                                                  log likelihood = -1860.215
2
Gradient vector (length = .5142436):
       union:     union:     union:     union:     union:    rfper1:
       Lunion        age      grade      south      _cons        age
r1   .0071309   .4225259   .2192779   .0050204   .0146686   .1723773

      rfper1:    rfper1:    rfper1:    rfper1:  logitrho:    ltheta:
        grade      south   not_smsa      _cons      _cons      _cons
r1   .0829091   .0013531   .0009419   .0059511   .0301568  -.0029917
---------------------------------------------------------------------
--------
Iteration 5:
Coefficient vector:
       union:     union:     union:     union:     union:    rfper1:
       Lunion        age      grade      south      _cons        age
r1   .6344452  -.0285863  -.0539267  -.4883257   .5632897   .0081099

      rfper1:    rfper1:    rfper1:    rfper1:  logitrho:    ltheta:
        grade      south   not_smsa      _cons      _cons      _cons
r1  -.0064163   -.726067  -.4152245  -.9597098   .8453735   -.146111

                                                  log likelihood = -1860.215
2
Gradient vector (length = .0002252):
       union:     union:     union:     union:     union:    rfper1:
       Lunion        age      grade      south      _cons        age
r1   8.20e-07  -.0002042  -.0000347  -3.77e-07  -4.84e-06   .0000825
```

```
      rfper1:    rfper1:    rfper1:    rfper1:  logitrho:    ltheta:
       grade      south   not_smsa      _cons      _cons      _cons
r1   .0000306   1.38e-06   7.08e-07   2.62e-06  -4.66e-06   4.88e-06
----------------------------------------------------------------------------
--------

Random-Effects Dynamic Probit Model              Number of obs   =      47
94
                                               Wald chi2(4)    =     82.04
Log likelihood = -1860.2152                      Prob > chi2     =    0.00
00

----------------------------------------------------------------------------
--------
      union |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interva
l]
-------------+--------------------------------------------------------------
--------
union        |
     Lunion |   .6344452    .098323     6.45   0.000     .4417357    .82715
47
        age |  -.0285863    .009217    -3.10   0.002    -.0466512   -.01052
13
      grade |  -.0539267   .0269226    -2.00   0.045    -.1066942   -.0011
593
      south |  -.4883257   .1238919    -3.94   0.000    -.7311494    -.245
502
      _cons |   .5632897   .4798556     1.17   0.240      -.37721    1.5037
89
-------------+--------------------------------------------------------------
--------
rfper1       |
        age |   .0081099   .0238048     0.34   0.733    -.0385467    .05476
64
      grade |  -.0064163   .0340851    -0.19   0.851    -.0732218    .06038
92
      south |   -.726067   .1650786    -4.40   0.000    -1.049615   -.40251
89
   not_smsa |  -.4152245   .1644004    -2.53   0.012    -.7374433   -.0930
056
      _cons |  -.9597098    .841375    -1.14   0.254    -2.608775    .68935
49
-------------+--------------------------------------------------------------
--------
  /logitrho |   .8453735    .163999     5.15   0.000     .5239414    1.1668
06
    /ltheta |   -.146111   .1267408    -1.15   0.249    -.3945184    .10229
63
-------------+--------------------------------------------------------------
--------
        rho |   .6995957   .0344663    20.30   0.000     .6280689    .76256
71
      theta |   .8640618   .1095119     7.89   0.000     .6740046    1.1077
12
----------------------------------------------------------------------------
--------
LR test of rho = 0:   chi2(1)     =  243.69
```

```
Prob > chi2 =  0.0000
```

Important Statistical Good News

The second part of the model (sometimes called vector z) additionally contains the variable *not_smsa* (1 if living outside a standard metropolitan statistical area).

This variable has a significant negative effect on the probability of union membership in the initial period, whether estimated as a separate probit or as part of the full model.

The model duplicates the results in Stewart (2006).

A Little Light Relief

My Jupyter Limerick

A researcher with time to fritter

Decided he didn't need Jupyter

His results he would show

Without a traceable workflow

Could a researcher be any stupider?

Remarks and Conclusion

Jupyter Notebooks have a number of attractive features

1. They facilitate easy documentation alongside research code (using Markdown)
2. They have good portability (notebooks are easy to share)
3. They are 'language agnostic' (i.e. analyses can be undertaken in over 40 programming languages)
4. They can produce rich visual outputs
5. They can leverage 'big data' research tools (e.g. python)
6. They can be used as integrated tools in teaching, training, knowledge exchange and research capacity building
7. They might better facilitates collaboration

Researchers such as Lorena Barba have demonstrated the value of using Jupyter Notebooks in research informed teaching by weaving executable code with multi-media content (see http://lorenabarba.com/).

The ADRC-S team and associates should therefore exploring the theoretical and practical possibilities for using Jupyter Notebooks for research using micro-level administrative social science data.

Finally converting this Jupyter Notebook

see http://nbconvert.readthedocs.io/en/latest/

1. At the cmd prompt *conda install nbconvert*
2. Change directory *C:\Users\vgayle\Documents*
3. Type *jupyter nbconvert --to html mynotebook.ipynb*

The work is very exploratory.

Positive comments and brickbat are equally welcome.

or @profbigvern