scikits.statsmodels - a brief introduction

```
>>> import scikits.statsmodels.api as sm
```

http://scikits.appspot.com/statsmodels

Berian James <berian@berkeley.edu>

Astronomy Department, UC Berkeley

Statistical models and computations for SciPy

- statsmodels is a statistical modelling and computation toolbox for numpy/scipy, aimed at complementing scipy.stats with ‘frequentist’ modelling tools; cf. pymc, which is a ‘bayesian’ toolbox.

- It is built on numpy, i.e., numpy arrays are the most practical data type; they are generic, efficient and straight-forward to handle. Some of the time-series analysis is designed for use with Pandas (more on this later).

- statsmodels is available through PyPI, easy_install, github, ... http://pypi.python.org/pypi/scikits.statsmodels

- statsmodels is already compatible with Python 3 and is almost wholly pure python, with a handful of cython wrappings

scikits.statsmodels resources

- http://scikits.appspot.com/statsmodels
  statsmodels homepage, download, installation

- http://statsmodels.sourceforge.net/
  statsmodels documentation, API reference, examples; not complete

- http://www.github.com/statsmodels/statsmodels
  http://pypi.python.org/pypi/scikits.statsmodels/
  statsmodels repositories

- http://conference.scipy.org/scipy2010/slides/skipper_seabold_statsmodels.pdf
  http://conference.scipy.org/scipy2011/slides/mckinney_time_series.pdf
  SciPy 2010/2011 by Skipper Seabold and Wes McKinney

Inbuilt data sets and statsmodels io

- statsmodels contains a number of inbuilt data sets (sm.datasets)
  e.g. >>> data = sm.datasets.scotland.load()

- Variables are cast as either ‘endogenous’ or ‘exogenous’

- Particularly with the time series analysis module, the pandas TimeSeries data structure is available for use

- Ultimately, statsmodels is targetted at (in the words of its creator) “Statistical, Financial Econometric, and Econometric models”

Regression in statsmodels

- Implementation of least-squares routines: ordinary least squares, weight least squares and general least squares.

- Discuss notebook for examples

- Extensions of these methods: generalised linear models and robust linear models, which will not be covered here.

- There are also time-series specific regression methods.

Time-series analysis and regression

- statsmodels provides fundamental time-series analysis methods, including:
    - auto- and cross-correlation and -covariance
      `sm.tsa.acovf, sm.tsa.acf, sm.tsa.ccf, sm.tsa.ccovf`
    - periodogram for regularly-spaced data, i.e. $|\mathcal{F}(\mathbf{x})|^2/N$
      `sm.tsa.periodogram`
- Many of these are also available through numpy/scipy, so that the power of sm.tsa lies in its estimation methods, for univariate and vector autoregressive processes (AR, VAR) and auto-regressive moving-average processes (ARMA)
- Discuss example in notebook

Editorial: statsmodel “vs” pymc