R/Python for Economic Data Analysis

- Using R
  - R 3.5.x
  - RStudio IDE
    - R Notebook

- Using Python
  - Python 3.7.x
  - Jupyter Notebook
  - Spyder IDE

Introduction to Python

- Installation
  - Anaconda
  - Miniconda (Python only)
    - Jupyter Notebook
    - Spyder IDE
  - Essential Libraries
    - Numpy, Matplotlib, PyLib, …,
    - Pandas, Statsmodels, Scikit-Learn

Introduction to Python

- Reference
  - William McKinney, Python for Data Analysis
  - Kevin Sheppard, Python for Econometrics, 2017.
  - Thomas J. Sargent and John Stachurski, Lectures in Quantitative Economics, 2017.

Introduction to R

- Installation
  - R-Projects (CRAN)
  - Rstudio
  - Useful Packages
    - AER, MASS, lmtest, maxLik, gmm, systemfit
    - plm, glmnet, caret, ...

Economic Data Analysis Using R

- Introduction to R
  - Getting Started
  - Using RStudio IDE
  - R Basics
    - ceR eBook Project: R/Python for Econometric Analysis by Example (WIP, 2019)
- Economic Data
  - Cross Sections
  - Time Series
  - Panel Data

Economic Data

- Data Generating Process (DGP)
  - Economic Theory
  - Probability Distribution

$$uniform \quad f(x;\min,\max)=\frac{1}{\max-\min},\ \min\le x\le\max$$

$$normal \quad f(x;\mu,\sigma^2)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right),\ -\infty<x<\infty$$

$$binomial \quad f(x;n,p)=\binom{n}{x}p^x(1-p)^{n-x},\ x=0,1,\ldots,n$$

$$possion \quad f(x;\lambda)=\lambda^x\frac{\exp(\lambda)}{x!},\ x=0,1,2,\ldots$$

$\ldots$

Economic Data

- Data Structure
  - Cross Sections
  - Time Series
  - Panel Data
- Random Experiments

$$x_i \sim iid\ (normal)$$

$$x_t \sim arima$$

$$x_{it} \sim hetero. + arima$$

- Data Size
  - High Frequency Financial Data
  - Massive Datasets (Big Data)
    - Government and Company Data
    - Internet Search and Scan Data
    - High Dimensional Data

Economic Data

- High Frequency Financial Data
  - Stock Market Daily Time Series
  - S&P 500, VIX (Fear Index), China SSEC
  - Exchange Rates
  - Ticker Data
    - Apple Co. (AAPL)
    - Tesla Motors, Inc. (TSLA)
    - …

Economic Data

- Useful On Line Data Resources
  - Quandl
  - quantmod
  - Kaggle Datasets
  - Google Public Data
  - Amazon AWS Public Data
  - Intraday Stock Quotes
  - …

Economic Data Analysis

- Data Preparation
  - Getting Data
  - Tidying Data
- Data Visualization
  - Using Graphs
- Data Analysis
  - Regression
  - Classification

Data Preparation Using R

- Getting Data
  - Using R: read.table(), read.csv(), read.dta(), read_excel()
- Tidying Data
  - Structuring datasets to facilitate visualization and analysis:
    - Each variable forms a column; Each observation forms a row; Each type of observational unit forms a table.
  - Variable Transformation
  - Using R: reshape(), subset(), order(), transform(), aggregate()
- Using package tidyverse

Data Visualization Using R

- Simple Graphs
  - Scatter, Line, Bar, and Box Plots
  - Histogram and Distribution
- Using ggplot2 Package
  - Grammar of Graphics (Part1, Part2)
- References
  - Data Visualization with ggplot2

R Package ggplot2: qplot

- qplot(data, x, y, …, geom(…)) + …
  - Data Frame
  - Aesthetics: x, y, alpha, color, size, shape, fill, …
  - Geometry: point, line, bar, jitter, boxplot, histogram, density, smooth, …
- Adding facets
  - facet_grid

R Package ggplot2: ggplot

- ggplot(data, aes(x,y,…), geom_...) + …
  - Data
  - Aesthetics
  - Geometry
  - Stats, Facets
  - Scales, Colors, Coordinates
  - Labels, Legends
  - Themes

Data Analysis

- Descriptive Statistics
- Correlation and Covariance
- Analysis of Variances (AOV, ANOVA)
- Hypothesis Testing
  - DGP ~ Non IID
  - One-Variable Testing (t-test)
  - Two-Variable Testing (paired t-test)

Data Analysis

- Cross Sections Data
  - Hypothesis Testing
    - Normality
    - Independence
    - Homoscedasticity
  - Linear Regression
    - Least Squares Regression
    - Maximum Likelihood
  - Regression Diagonstics

Data Analysis

- Time Series Data
  - Hypothesis Testing
    - Durbin-Watson
    - Box-Pierce / Ljung-Box
    - ACF/PACF
  - Transformation: Lag, Difference
  - Model Estimation
  - Forecasting

Data Analysis

- Panel Data
  - Hypothesis Testing
    - Unobserved Hetrogeneity
    - Time Series Correlation
    - Spatial Correlation
  - Model Estimation
    - Fixed Effects
    - Random Effects
  - Multilevel Analysis

Example: Big Mac Price Index

- The Big Mac price index was invented by *The Economist* in 1986 as a lighthearted guide to whether currencies are at their "correct" level. It is based on the theory of purchasing-power parity (PPP), the notion that in the long run exchange rates should move towards the rate that would equalize the prices of an identical basket of goods and services (in this case, a burger) in any two countries.
- Data Source: The Economist-GitHub

Big Mac Price Index

- Cross Sectional Data
  - Find the latest data (January 2019) over 57 countries
  - Read the data (from Excel spreadsheet)
  - Summarize the data

Big Mac Price Index

- Time Series Data
  - Find the data for US, from 1986 to the latest year, irregularly unequal-spaced time series
  - Find the data for China, from 1996 to the latest year.
  - Read the data (from multiple Excel spreadsheets)
  - Summarize and compare
  - Time series analysis

Big Mac Price Index

- Panel Data
  - Find the data for all available countries, from all available years until the latest year, unequal-spaced time series and unbalanced panels
  - Read the data (from multiple Excel spreadsheets)
  - Summarize the data
  - Panel data analysis