Python Companion Course

Kevin Sheppard
University of Oxford
www.kevinsheppard.com

September 2019

Contents

Topic 1

Data

1.1 Data Set Construction

Functions

`pd.read_csv`, `pd.read_excel`, `np.diff` or `DataFrame.diff`, `DataFrame.resample`

Exercise 1

1. Download all available daily data for the S&P 500 and the Hang Seng Index from Yahoo! Finance.
2. Import both data sets into Python. The final dataset should have a `DateTimeIndex`, and the date column should not be part of the `DataFrame`.
3. Construct weekly price series from each, using Tuesday prices (less likely to be a holiday).
4. Construct monthly price series from each using last day in the month.
5. Save the data to the HDF file "equity-indices.h5".

Exercise 2

Write a function that will correctly aggregate to weekly or monthly respecting the aggregation rules

- High: `max`
- Low: `min`
- Volume: `sum`

The signature should be:

```
def yahoo_agg(data, freq):
    <code here>
```

```
    return resampled_data
```

Exercise 3

1. Import the Fama-French benchmark portfolios as well as the 25 sorted portfolios at both the monthly and daily horizon from Ken French"s Data Library. **Note** It is much easier to clean to data file before importing than to find the precise command that will load the unmodified data.
2. Import daily FX rate data for USD against AUD, Euro, JPY and GBP from the Federal Reserve Economic Database (FRED). Use Excel rather than csv files.
3. Save the data to the HDF files "fama-french.h5" and "fx.h5"

Exercise 3 (Alternative method)

1. Install and use `pandas-datareader` to repeat the previous exercise.

Preliminary Step You must first install the module using

```
pip install pandas-datareader
```

from the command line. Then you can run this code. **Note**: Running this code requires access to the internet.

Exercise 4

Download data on 1 year and 10 year US government bond rates from FRED, and construct the term premium as the different in yields on 10 year and 1 year bonds. Combine the two yield series and the term premium into a `DataFrame` and save it as HDF.

1.2 Simulation

Functions

`np.random.standard_normal`, `np.random.standard_t`, `np.random.RandomState`

Exercise 5

Simulate 100 standard Normal random variables

Exercise 6

Simulate 100 random variables from a $N\left(.08, .2^2\right)$

Exercise 7

Simulate 100 random variables from a Students t with 8 degrees of freedom

Exercise 8

Simulate 100 random variables from a Students t with 8 degrees of freedom with a mean of 8% and a volatility of 20%. Note: $V[X] = \frac{\nu}{\nu-2}$ when $X \sim t_\nu$.

Exercise 9

Simulate two identical sets of 100 standard normal random variables by resetting the random number generator.

Exercise 10

Repeat exercise 7 using only `standard_normal`.

1.3 Expectations

Functions

`np.random.RandomState`, `RandomState.standard_normal`, `RandomState.standard_t`, `RandomState.chi2`, `np.exp`, `np.mean`, `np.std`, `scipy.integrate.quadrature`, `scipy.integrate.quad`

Exercise 11

Compute $E[X]$, $E\left[X^2\right]$, $V[X]$ and the kurtosis of X using Monte Carlo integration when X is distributed:

1. Standard Normal
2. $N\left(0.08, 0.2^2\right)$
3. Students t_8
4. χ^2_5

Exercise 12

1. Compute $E[\exp(X)]$ when $X \sim N\left(0.08, 0.2^2\right)$.
2. Compare this to the analytical result for a Log-Normal random variable.

Exercise 13

Explore the role of uncertainty in Monte Carlo integration by increasing the number of simulations 300% relative to the base case.

Exercise 14

Compute the expectation in exercise 11 using quadrature.

Note: This requires writing a function which will return $\exp(x) \times \phi(x)$ where $\phi(x)$ is the pdf evaluated at x.

Exercise 15

Optional (Much more challenging)

Suppose log stock market returns are distributed according to a Students t with 8 degrees of freedom, mean 8% and volatility 20%. Utility maximizers hold a portfolio consisting of a risk-free asset paying 1% and the stock market. Assume that they are myopic and only care about next period wealth, so that

$$U\left(W_{t+1}\right)=U\left(\exp\left(r_p\right)W_t\right)$$

and that $U(W)=\frac{W^{1-\gamma}}{1-\gamma}$ is CRRA with risk aversion γ. The portfolio return is $r_p=wr_s+(1-w)r_f$ where s is for stock market and f is for risk-free. A 4th order expansion of this utility around the expected wealth next period is

$$E_t\left[U\left(W_{t+1}\right)\right]\approx\phi_0+\phi_1\mu_1'+\phi_2\mu_2'+\phi_3\mu_3'+\phi_4\mu_4'$$

where

$$\phi_j=(j!)^{-1}U^{(j)}\left(E_t\left[W_{t+1}\right]\right),$$

$$U^{(j)}=\frac{\partial^j U}{\partial W^j},$$

$$\mu_k'=E_t\left[(r-\mu)_p^k\right],$$

and $\mu=E_t\left[r_p\right]$. Use Monte Carlo integration to examine how the weight in the stock market varies as the risk aversion varies from 1.5 to 10. Note that when $\gamma=1$, $U(W)=\ln(W)$. Use $W_t=1$ without loss of generality since the portfolio problem is homogeneous of degree 0 in wealth.

Topic 2

Estimation

2.1 Method of Moments

Functions

`DataFrame.mean`, `DatFrame.sum`, `plt.subplots`, `plt.plot`, `stats.kurtosis`, `stats.skewness`

Exercise 16

Estimate the mean, variance, skewness and kurtosis of the S&P 500 and Hang Seng using the method of moments using monthly data.

Exercise 17

Estimate the asymptotic covariance of the mean and variance of the two series (separately, but not the skewness and kurtosis).

Exercise 18

Estimate the Sharpe ratio of the two series and compute the asymptotic variance of the Sharpe ratio. See Chapter 2 of the notes for more on this problem.

The asymptotic variance is computed as

$$D\Sigma D'$$

where

$$D = [\sigma^{-1}, -1/2\mu\sigma^{-3}]$$

and Σ is the asymptotic covariance of the mean and variance. Finally, we divide by n the sample size when computing the statistic variance.

Exercise 19

Plot rolling estimates of the four moments using 120 months of consecutive data using a 4 by 1 subplot against the dates.

The simple approach to this problem uses a loop accross 120, 121, ..., n and computes the statistics using 120 observations. The figure is then created with a call to `plt.subplots` and the series can be directly plotted by calling `ax.plot`.

The pandas-centric approach uses teh `rolling` method to compute rolling statistics and then uses `.plot.line` with `subplots=True` to produce the plot.

2.2 Maximum Likelihood

Functions

`np.log`, `scipy.special.gamma`, `scipy.special.gammaln`, `scipy.stats.norm.cdf`, `scipy.optimize.minimize`, `scipy.stats.t`, `np.var`, `np.std`, `scipy.stats.norm.pdf`

Exercise 20

Simulate a set of i.i.d. Student's t random variables with degree of freedom parameter $\nu = 10$. Standardize the residuals so that they have unit variance using the fact that $V[x] = \frac{\nu}{\nu-2}$. Use these to estimate the degree of freedom using maximum likelihood. Note that the likelihood of a standardized Student's t is

$$f(x;\nu,\mu,\sigma^2) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \frac{1}{\sqrt{\pi(\nu-2)}} \frac{1}{\sigma} \frac{1}{\left(1+\frac{(x-\mu)^2}{\sigma^2(\nu-2)}\right)^{\frac{\nu+1}{2}}}$$

where $\Gamma()$ is known as the gamma function.

Exercise 21

Repeat the previous exercise using daily, weekly and monthly S&P 500 and Hang Seng data. Note that it is necessary to remove the mean and standardize by the standard deviation error before estimating the degree of freedom. What happens over longer horizons?

Exercise 22

Repeat the previous problem by estimating the mean and variance simultaneously with the degree of freedom parameter.

Exercise 23

Simulate a set of Bernoulli random variables y_i where

$$p_i = \Phi(x_i)$$

where $X_i \sim N(0,1)$. (Note: p_i is the probability of success and $\Phi()$ is the standard Normal CDF). Use this simulated data to estimate the Probit model where $p_i = \Phi(\alpha_0 + \alpha_1 x_i)$ using maximum likelihood.

Exercise 24

Estimate the asymptotic covariance of the estimated parameters in the previous.
The derivative of the log-likelihood for a single observation is

$$\frac{\partial \left\{ y_i \ln \left(\Phi \left(\alpha_0 + \alpha_1 x_i \right) \right) + \left(1 - y_i \right) \ln \left(1 - \Phi \left(\alpha_0 + \alpha_1 x_i \right) \right) \right\}}{\partial \alpha_j}$$

which is

$$y_i \frac{\phi \left(\alpha_0 + \alpha_1 x_i \right)}{\Phi \left(\alpha_0 + \alpha_1 x_i \right)} - \left(1 - y_i \right) \frac{\phi \left(\alpha_0 + \alpha_1 x_i \right)}{1 - \Phi \left(\alpha_0 + \alpha_1 x_i \right)}$$

for α_0 and

$$y_i x_i \frac{\phi \left(\alpha_0 + \alpha_1 x_i \right)}{\Phi \left(\alpha_0 + \alpha_1 x_i \right)} - \left(1 - y_i \right) x_i \frac{\phi \left(\alpha_0 + \alpha_1 x_i \right)}{1 - \Phi \left(\alpha_0 + \alpha_1 x_i \right)}$$

for α_1 where $\Phi(\cdot)$ is the cdf of a standard Normal random variable and $\phi(\cdot)$ is the pdf of a standard Normal random variable.

2.3 Bias and Verification of Standard Errors

Methods/Functions

`mean`, `var`, `RandomState`, `RandomState.chisquare`, `array`, `DataFrame.plot.kde`, `stats.norm.ppf`

Exercise 25

Simulate a set of i.i.d. χ^2_5 random variables and use the method of moments to estimate the mean and variance.

Exercise 26

Compute the asymptotic variance of the method of moment estimator.

Exercise 27

Repeat Exercises 24 and 25 a total of 1000 times. Examine the finite sample bias of these estimators relative to the true values.

Exercise 28

Repeat Exercises 24 and 25 a total of 1000 times. Compare the covariance of the estimated means and variance (1000 of each) to the asymptotic covariance of the parameters (use the average of the 1000 estimated variance-covariances). Are these close? How does the sample size affect this?

Exercise 29

In the previous problem, for each parameter, form a standardized parameter estimate as

$$z_i = \frac{\sqrt{n}\left(\hat{\theta}_i - \theta_{i,0}\right)}{\sqrt{\hat{\Sigma}_{ii}}}$$

where

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \xrightarrow{d} N\left(0, \Sigma\right)$$

so that $\hat{\Sigma}$ is the estimated asymptotic covariance. What percent of these z_i are larger in absolute value than 10%, 5% and 1% 2-sided critical values from a normal?

Exercise 30

Produce a density plot of the z_i standardized parameters and compare to a standard normal.

Exercise 31

Repeat the same exercise for the Bernoulli problem from the previous question.

Topic 3

Linear Regression

3.1 Basic Linear Regression

Functions

`sm.OLS`

Exercise 32

Use the OLS function to estimate the coefficients of the Fama-French portfolios (monthly data) on the market, size and value factors. Include a constant in the regressions. Use only the four extremum portfolios – that is the 1-1, 1-5, 5-1 and 5-5 portfolios. Estimate the model with homoskedastic errors and with White's covariance estimator.

Exercise 33

Are the parameter standard errors similar using the two covariance estimators? If not, what does this mean?

Exercise 34

How much of the variation is explained by these three regressors?

3.2 Rolling and Recursive Regressions

Functions

`sm.OLS`, `plt.title`, `plt.legend`, `plt.subplots`, `plt.plot`

Exercise 35

For the same portfolios in the previous exercise, compute rolling βs using 60 consecutive observations.

Exercise 36

For each of the four βs, produce a plot containing four series:

- A line corresponding to the constant β (full sample)
- The β estimated on the rolling sample
- The constant β plus 1.96$\times$ the variance of a 60-observation estimate of β. The 60-month covariance can be estimated using a full sample VCV and rescaling it by T/60 where T is the length of the full sample used to estimate the VCV. Alternatively, the VCV could be estimated by first estimating the 60-month VCV for each sub-sample and then averaging these.
- The constant β minus 1.96$\times$ the variance of a 60-observation estimate of β.

Exercise 37

Do the factor exposures appear constant?

Exercise 38

What happens if only the market is used as a factor (repeat the exercise excluding SMB and HML).

3.3 Model Selection and Cross-Validation

Functions

`RandomState.permute`, `sm.OLS`, `set`, `scipy.random.norm.ppf`, `np.linspace`, `np.mean`

Exercise 39

For four portfolios we have been looking at, and considering all 8 sets of regressors which range from no factor to all 3 factors, which model is preferred by AIC, BIC, GtS and StG?

Explanation

For each of the portfolios, we start with a list of included variables that includes all three factors. We can then use a loop to see if any of the included variables have insignificant t-stats. We first create a temporary set of regressors that uses the factors are are in `included`. We can then check if any of the t-stats are less than our critical value that is defined above. If one is less than the value, we need to drop the variable. We sort the absolute t-stats so that the minimum is first, and then get the variable name from the index. Finally, we use `.remove` to remove this name from the list of included variables.

If no t-stat is less than our critical value, we can call `break` which terminates the loop early.

Exercise 40

Cross-validation is a method of analyzing the in-sample forecasting ability of a cross-sectional model by using α% of the data to estimate the model and then measuring the fit using the remaining $100 - \alpha$%. The most common forms are 5- and 10-fold cross-validation which use $\alpha = 20\%$ and 10%, respectively. k-fold cross validation is implemented by randomly grouping the data into k-equally-sized groups, using k-1 of the groups to estimate parameters, and then evaluating using the bin that was held out. This is then repeated so that each bin is held out.

1. Implement cross-validation using the 5- and 10-fold methods for all 8 models.
2. For each model, compute the full-sample sum of squared errors as well as the sum-of-squared errors using the held-out sample. Note that all data points will appear exactly once in both of these sum or squared errors. What happens to the cross-validated R^2 when computed on the held out sample when compared to the full-sample R^2? (k-fold cross validated SSE by the TSS).

Topic 4

ARMA Models

4.1 ARMA Estimation

Functions

`tsa.SARIMAX`

Exercise 41

Estimate an AR(1) on the term premium, and compute standard errors for the parameters.

Exercise 42

Estimate an MA(5) on the term premium, and compute standard errors for the parameters.

Exercise 43

Estimate an ARMA(1,1) on the term premium, and compute standard errors for the parameters.

4.2 ARMA Model Selection

Functions

`sm.tsa.SARIMAX`

Exercise 44

Perform a model selection exercise on the term premium using

1. General-to-Specific
2. Specific-to-General
3. Minimizing an Information Criteria

4.3 ARMA Residual Diagnostics

Functions

`tsa.SARIMAX`, `sm.stats.diagnostic.acorr_ljungbox`, `SARIMAXResults.test_serial_correlation`, `statsmodels.sandbox.stats.diagnostic.acorr_lm`

Exercise 45

Compute the residuals from your preferred model from the previous exercise, as well as a random-walk model.

1. Plot the residuals
2. Is there evidence of autocorrelation in the residuals?
3. Compute the Q statistic from both sets of residuals. Is there evidence of serial correlation?
4. Compute the LM test for serial correlation. Is there evidence of serial correlation?

4.4 ARMA Forecasting

Functions

`tsa.SARIMAX.forecast`

Exercise 46

Produce 1-step forecasts from your preferred model in the previous exercise, as well as a random-walk model.

1. Are the forecasts objectively accurate?
2. Compare these forecasts to the random walk models using MSE and MAE.

Note: Use 50% of the sample to estimate the model and 50% to evaluate it.

Exercise 47

Produce 3-step forecasts from the models selected in the previous exercises as well as a random walk model.

1. Compare these forecasts to the random walk models using MSE and MAE.

4.5 Unit Root Testing

Functions

`sm.tsa.stattools.adfuller`, `arch.unitroot.ADF`

Exercise 48

Download data on the AAA and BAA yields (Moodys) from FRED and construct the default premium as the difference between these two.

1. Test the default premium for a unit root.
2. If you find a unit root, test the change.

Exercise 49

Download data on consumer prices in the UK from the ONS.

1. Test the log of CPI for a unit root.
2. If you find a unit root, test inflation for one.

4.6 Model Selection and Cross-Validation

Functions

`RandomState.permute`, `sm.OLS`, `set`, `scipy.random.norm.ppf`, `np.linspace`, `np.mean`

Exercise 39

For four portfolios we have been looking at, and considering all 8 sets of regressors which range from no factor to all 3 factors, which model is preferred by AIC, BIC, GtS and StG?

Explanation

For each of the portfolios, we start with a list of included variables that includes all three factors. We can then use a loop to see if any of the included variables have insignificant t-stats. We first create a temporary set of regressors that uses the factors are are in `included`. We can then check if any of the t-stats are less than our critical value that is defined above. If one is less than the value, we need to drop the variable. We sort the absolute t-stats so that the minimum is first, and then get the variable name from the index. Finally, we use `.remove` to remove this name from the list of included variables.

If no t-stat is less than our critical value, we can call `break` which terminates the loop early.

Exercise 40

Cross-validation is a method of analyzing the in-sample forecasting ability of a cross-sectional model by using α% of the data to estimate the model and then measuring the fit using the remaining $100 - \alpha$%. The most common forms are 5- and 10-fold cross-validation which use $\alpha = 20\%$ and 10%, respectively. k-fold cross validation is implemented by randomly grouping the data into k-equally-sized groups, using k-1 of the groups to estimate parameters, and then evaluating using the bin that was held out. This is then repeated so that each bin is held out.

1. Implement cross-validation using the 5- and 10-fold methods for all 8 models.
2. For each model, compute the full-sample sum of squared errors as well as the sum-of-squared errors using the held-out sample. Note that all data points will appear exactly once in both of these sum or squared errors. What happens to the cross-validated R^2 when computed on the held out sample when compared to the full-sample R^2? (k-fold cross validated SSE by the TSS).