Summer Python '23
Correlation in Housing Data

Andrew Papanicolaou

August 3, 2023

The U.S. housing market is an example where time-series data have correlations giving overestimated/underestimated measure of dependence. In this exercise we compare time series data from the S&P/Case-Shiller U.S. National Home Price Index (CSUSHPINSA) and the Federal Home Loan Bank of San Francisco's 11th District Monthly Weighted Average Cost of Funds Index (COFI).

#1.) Import the `pandas` library and type `df = pd.read_csv('housingData.csv')`. Define `COFI = df['Index Value']` and `CSUSHPINSA = df['CSUSHPINSA']`. Draw time-series plots for CSUSHPINSA and COFI using `matplotlib`.

#2.) Compute the relative change in each times series, i.e.
`X[t] = (COFI[t+1]-COFI[t])/COFI[t]`
`Y[t] = (CSUSHPINSA [t+1]-CSUSHPINSA [t])/CSUSHPINSA[t]`
Compute correlations between `X[t+L]` and `Y[t]` for `L`=-2,-1,0,1,2. For which `L` is the correlation strongest?

#3.) Show histograms of `X` and `Y`, and identify potential outliers in each.

#4.) Write a function to perform winsorization. Winsorization is technique that reduces the effect of outliers. For example, suppose we want to winsorize `X` so that all outliers above the 97.5% quantile are reduced to the 97.5%. From scratch, start by sorting `X` and identifying the empirical 97.5% cutoff, call it `q`. Then for each `t` define `Xwin[t]=min(q,X[t])`. Repeat the correlations in question #2 with `X` winsorized from above at 97.5% and `Y` from below at 1%.