Basket Selection Using Stepwise Regression

[View all machine learning examples](https://www.mathworks.com/solutions/machine-learning/examples.html)

This example demonstrates how to use the stepwise regression technique to select a basket of securities (a subset of predictors) from the Dow Jones Industrial Average constituents. The purpose of this example is to outline how to identify a subset of stocks that will perform as well as the Dow Jones index itself.

This example uses functions from the [Financial Toolbox™](https://www.mathworks.com/products/finance.html) and [Datafeed Toolbox™.](https://www.mathworks.com/products/datafeed.html)

Get Data from the Federal Reserve and Yahoo Repositories

startDate = datenum('03-Jan-2006');

endDate = datenum('29-Dec-2006');

% FRED for the Dow Index

c = fred('http://research.stlouisfed.org/fred2/');

DowFred = fetch(c, 'DJIA', startDate, endDate);

DJIA = DowFred.Data(:,2);

% Yahoo for the constituents. GM has been excluded.

y = yahoo;

tickers = {'AA','AIG','AXP','BA','C','CAT',...

'DD','DIS','GE','HD','HON',...

'HPQ','IBM','INTC','JNJ','JPM',...

'KO','MCD','MMM','MO','MRK','MSFT',...

'PFE','PG','T','UTX','VZ','WMT','XOM'};

DowTickers = zeros(numel(DJIA),29);

for ii = 1:numel(tickers)

tickerValues = fetch(y,tickers{ii},'Close',startDate,endDate);

DowTickers(:,ii) = tickerValues(end:-1:1,2); % Reverse order to have oldest first

end

Dates = (startDate:endDate)';

% Remove non-business days

Dates(~isbusday(Dates)) = [];

clearvars DowFred c y ii tickerValues

Perform Linear Regression on Six Randomly Chosen Stocks

rng(10) % Set random seed for reproducibility

randStocks = randperm(numel(tickers),6);

disp('Randomly chosen stocks:')

disp(tickers(randStocks))

% Fit a linear model which are linear in the predictors (stocks). Remove

% the intercept term because we want to reproduce the DJIA benchmark using

% a linear combination of the predictors or stocks. Infact this is how DJIA

% is computed.

mdl = fitlm(DowTickers(:,randStocks),DJIA,'VarNames',[tickers(1:6),{'DJIA'}],...

'Intercept',false);

% Predict using the model

predDJIA = predict(mdl,DowTickers(:,randStocks));

Randomly chosen stocks:

'PFE' 'AA' 'MCD' 'MO' 'IBM' 'CAT'

Visualize the Predicted Response Along with the Original DJIA

figure(1), subplot(2,1,1)

plot(Dates,DJIA,'m',Dates,predDJIA,'g','LineWidth',2)

datetick('x','mmmm'), xlabel('Time'),

legend({'DJIA','predicted DJIA'},'Location','NorthWest')

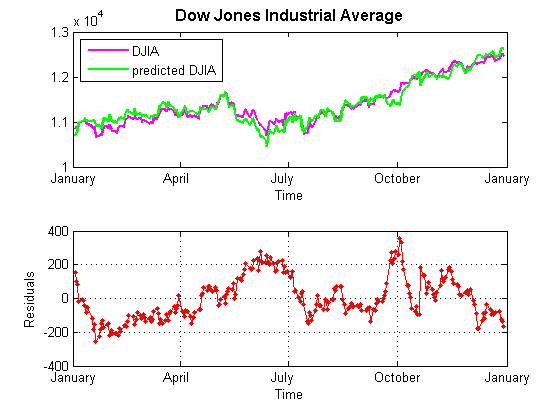
title('Dow Jones Industrial Average','FontSize',12,'FontWeight','Bold')

subplot(2,1,2)

plot(Dates,DJIA-predDJIA,'r.-')

datetick('x','mmmm'), xlabel('Time')

ylabel('Residuals'), grid on



We are now able to automate the process of choosing a subset of the predictors from the Dow Jones constituents such that the prediction error is minimum.

Perform Stepwise Regression

stepwise allows you to automatically remove statistically irrelevant predictors from model.

pRemove = 1e-4;

pEnter = 1e-3;

stepwisemdl = stepwiselm(DowTickers,DJIA,'constant','upper','linear',...

'premove',pRemove,'penter',pEnter,'VarNames',[tickers,{'DJIA'}],...

'Intercept',false,'Verbose',0,'Criterion','Rsquared');

predStepwiseDJIA = predict(stepwisemdl,DowTickers);

disp(stepwisemdl)

Linear regression model:

DJIA ~ AA + AXP + CAT + HPQ + IBM + JNJ + JPM

Estimated Coefficients:

Estimate SE tStat pValue

AA 30.871 2.3215 13.298 1.0325e-30

AXP 41.404 2.5743 16.084 3.5737e-40

CAT 11.605 0.75259 15.42 6.5099e-38

HPQ 32.547 2.9119 11.177 1.0833e-23

IBM 23.385 1.4718 15.889 1.6471e-39

JNJ 35.841 1.4867 24.107 1.6908e-66

JPM 48.493 2.3243 20.863 3.5882e-56

Number of observations: 251, Error degrees of freedom: 244

Root Mean Squared Error: 44.6

Visualize the Predicted Response Along with the Original DJIA

The index and the two predicted index values are visualized along with the error.

figure(2), subplot(2,3,1:2)

plot(Dates, DJIA,'m',Dates, predDJIA,'g',...

Dates, predStepwiseDJIA,'r','LineWidth',2)

datetick('x','mmmm'), xlabel('Time'),

legend({'DJIA','predicted DJIA', 'predicted Optimal DJIA'},'Location','NorthWest')

title('Dow Jones Industrial Average','FontSize',12,'FontWeight','Bold')

% Residuals

subplot(2,3,4:5)

plot(Dates,DJIA-predDJIA,'g.-',...

Dates,DJIA-predStepwiseDJIA,'r.-')

datetick('x','mmmm'), xlabel('Time')

ylabel('Residuals'), grid on

legend({['MSE: ',num2str(mdl.MSE)],['MSE: ',num2str(stepwisemdl.MSE)]},'Location','NorthWest')

title(['Basket: ' char(stepwisemdl.Formula)],'FontSize',12,'FontWeight','Bold')

% Pie Chart

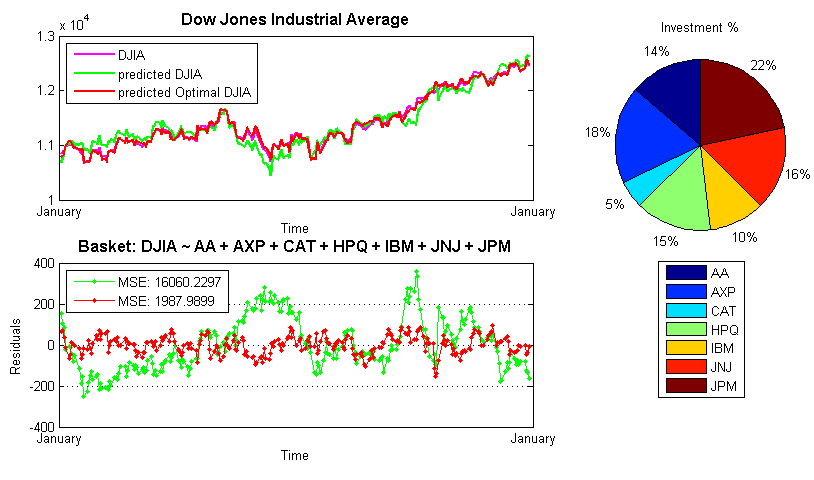
subplot(2,3,[3,6]),

pie(stepwisemdl.Coefficients.Estimate)

legend(stepwisemdl.CoefficientNames,'Location','SouthOutside')

title('Investment %')

set(2,'Units','Normalized','Position',[0.2 0.4 0.5 0.4])



From the figure you can see that the optimally chosen basket is much better at representing the DJIA index and has a much smaller mean square error.