EXAMPLES OF ROBUST ANALYSIS IN ANTI-FRAUD RELEVANT DATASETS

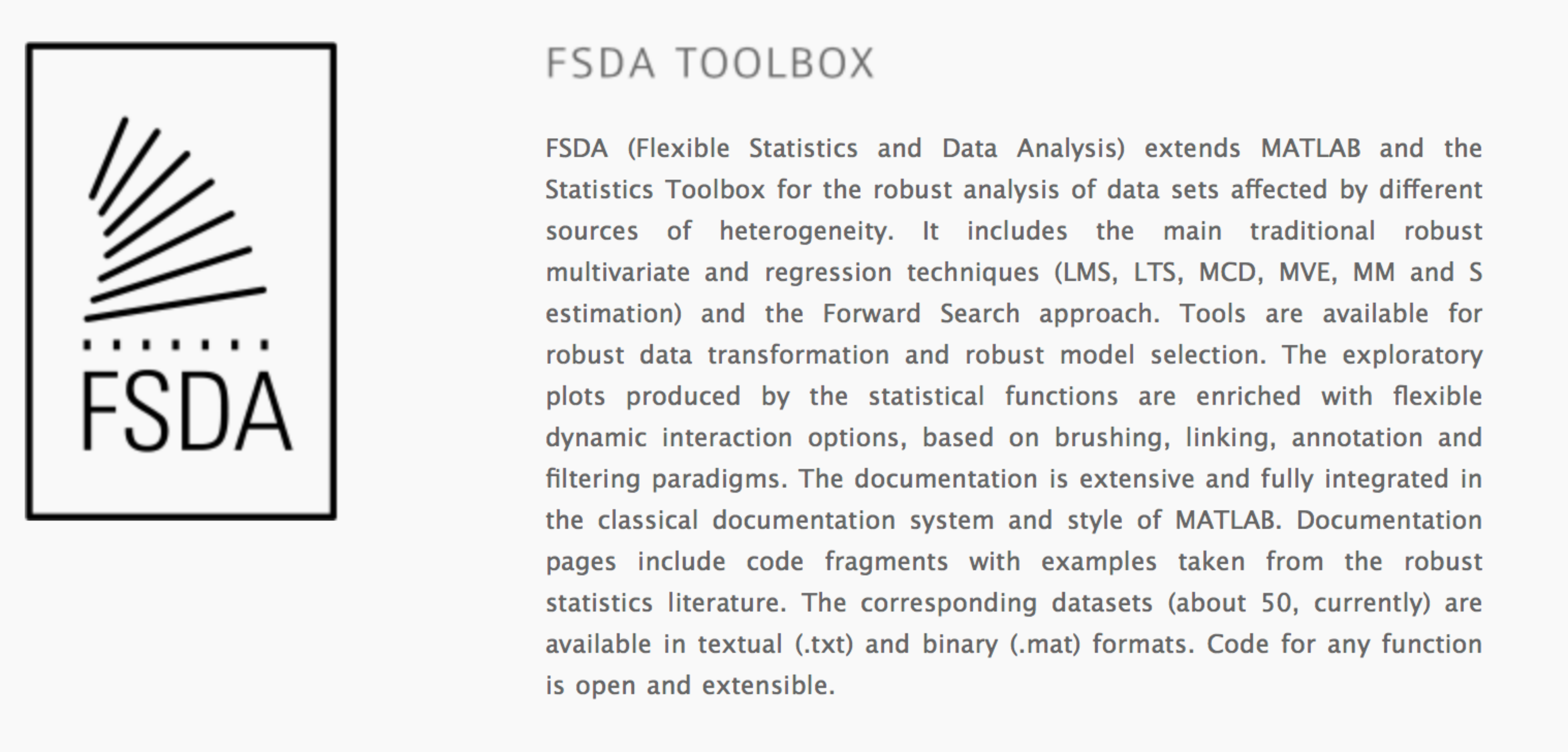
Domenico Perrotta
Joint Research Centre of the European Commission
domenico.perrotta@ec.europa.eu

23 November 2017
University of Parma, Department of Economics

Examples in this presentation are for illustrative purposes only and do not necessarily refer to ongoing or past investigations cases.

SOFTWARE USED FOR THE PRESENTATION

http://fsda.jrc.ec.europa.eu *or* *http://rosa.unipr.it/fsda.html*



FSDA TOOLBOX

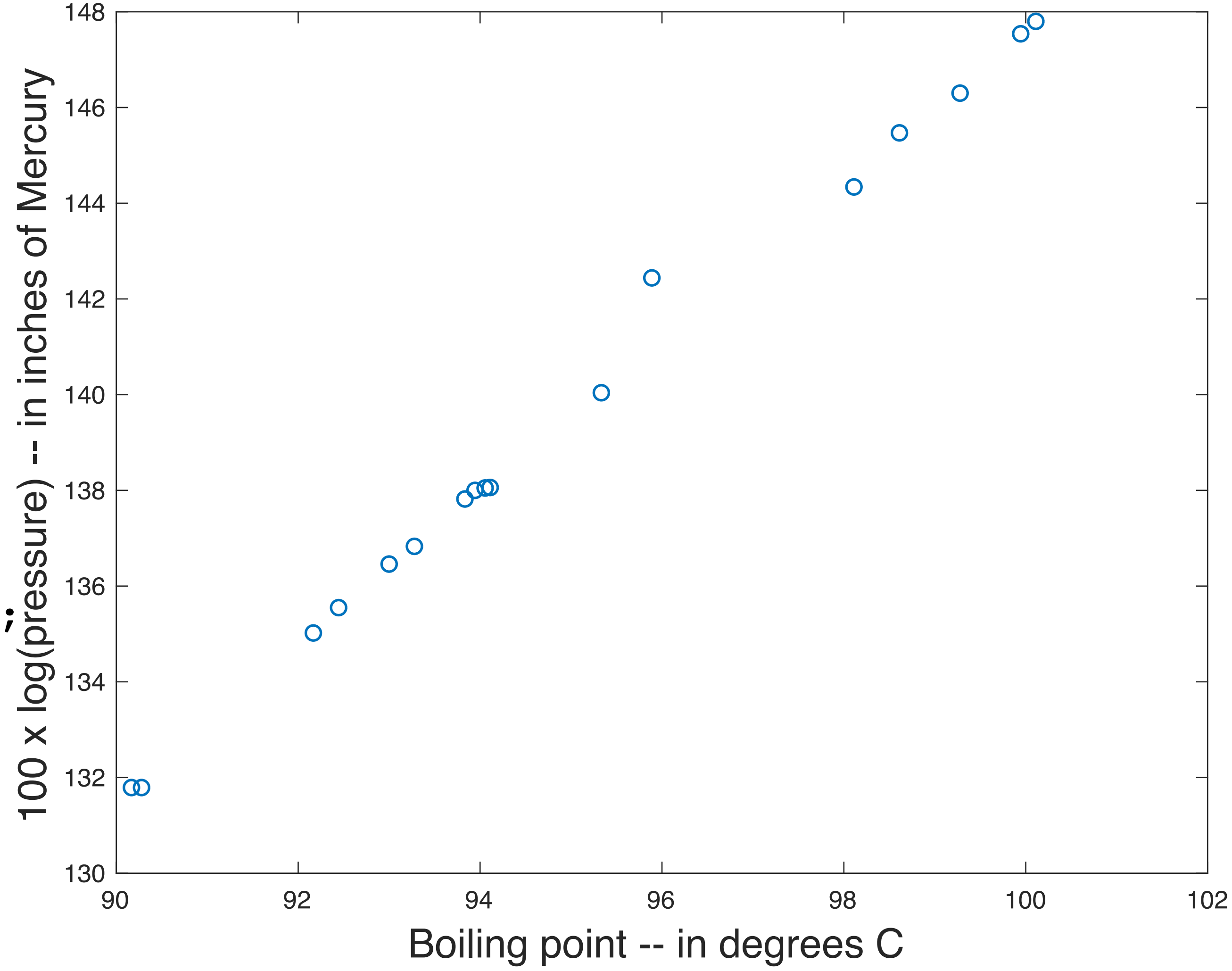
FSDA (Flexible Statistics and Data Analysis) extends MATLAB and the Statistics Toolbox for the robust analysis of data sets affected by different sources of heterogeneity. It includes the main traditional robust multivariate and regression techniques (LMS, LTS, MCD, MVE, MM and S estimation) and the Forward Search approach. Tools are available for robust data transformation and robust model selection. The exploratory plots produced by the statistical functions are enriched with flexible dynamic interaction options, based on brushing, linking, annotation and filtering paradigms. The documentation is extensive and fully integrated in the classical documentation system and style of MATLAB. Documentation pages include code fragments with examples taken from the robust statistics literature. The corresponding datasets (about 50, currently) are available in textual (.txt) and binary (.mat) formats. Code for any function is open and extensible.

OUTLIERS IN LINEAR REGRESSION ANALYSIS

Forbes dataset

17 observations on the boiling point of water at different pressures.
The purpose is to allow prediction of pressure from boiling point.

```
load('forbes.txt');
y=forbes(:,2);
X=forbes(:,1);
X = (X - 32) * 5/9; % Convert to Celsius
plot(X,y,'o');
xlabel('Boiling point','Fontsize',16);
ylabel('100 x log(pressure)','Fontsize',16);
f1 = gcf ; figure(f1);
```

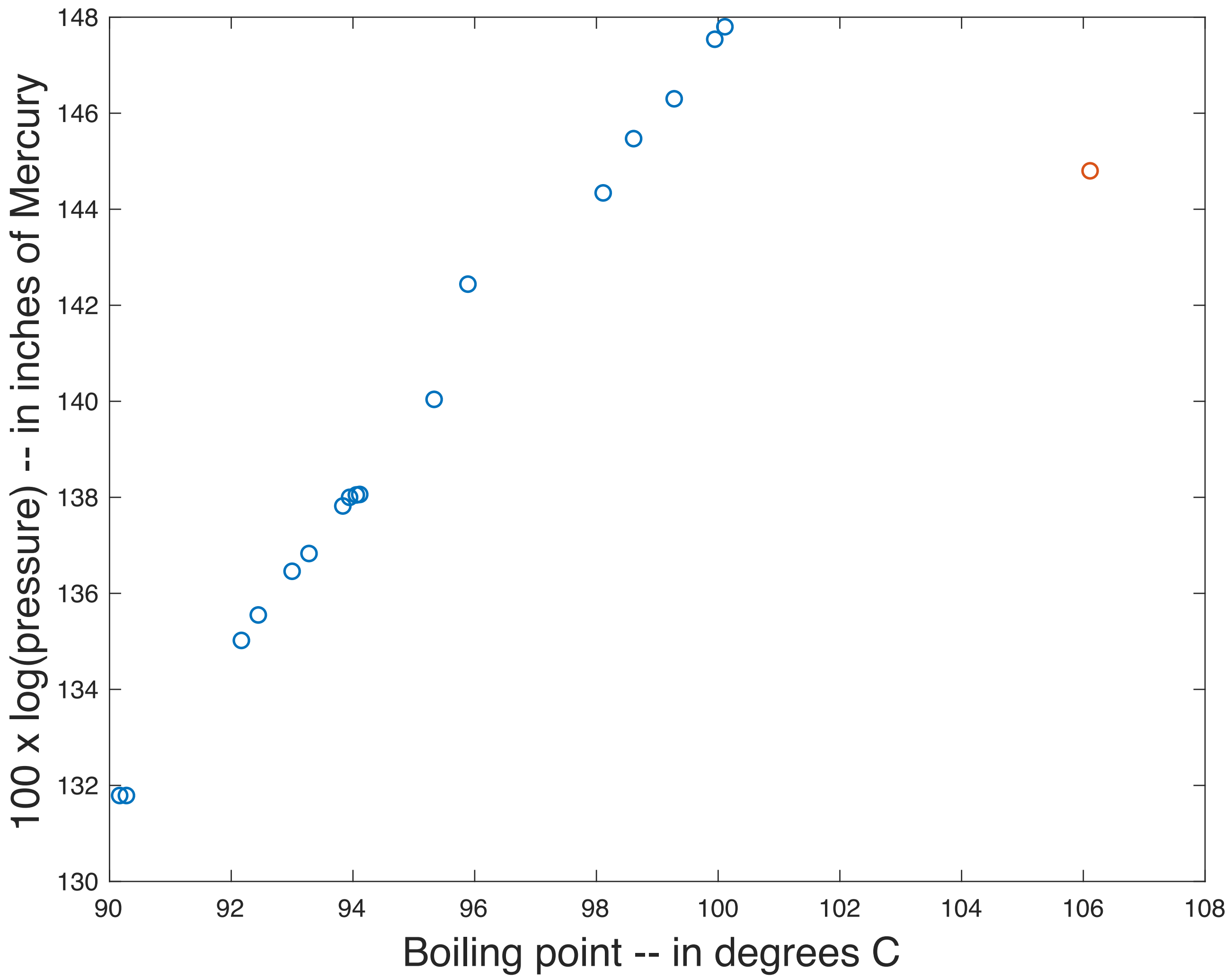


OUTLIERS IN LINEAR REGRESSION ANALYSIS

Contamination

Let’s add an influential outlier

```
yc = y; yc(end) = yc(end)-3;
Xc = X; Xc(end) = Xc(end)+6;
hold on
plot(Xc(end),yc(end),'o');
figure(f1);
```



OUTLIERS IN LINEAR REGRESSION ANALYSIS

Fit the data with a linear regression model

Let’s do it by Ordinary Least Square (OLS)

```
% let’s compute the OLS parameters
% with four equivalent codes

int = ones(size(Xc,1),1); % intercept
Xic = [int Xc];

beta0 = (Xic'*Xic)\(Xic'*yc);
beta1 = Xic\yc;
beta2 = regress(yc,Xic);
beta3 = fitlm(Xc,yc,'y ~ 1 + x1');
```

OLS objective: minimize the sum of the squares of the differences between the observed y-values (the log-pressures) and those predicted by a linear function of the x-values (the temperatures).

These are the squared vertical distances between each data point and the corresponding point on the regression line.

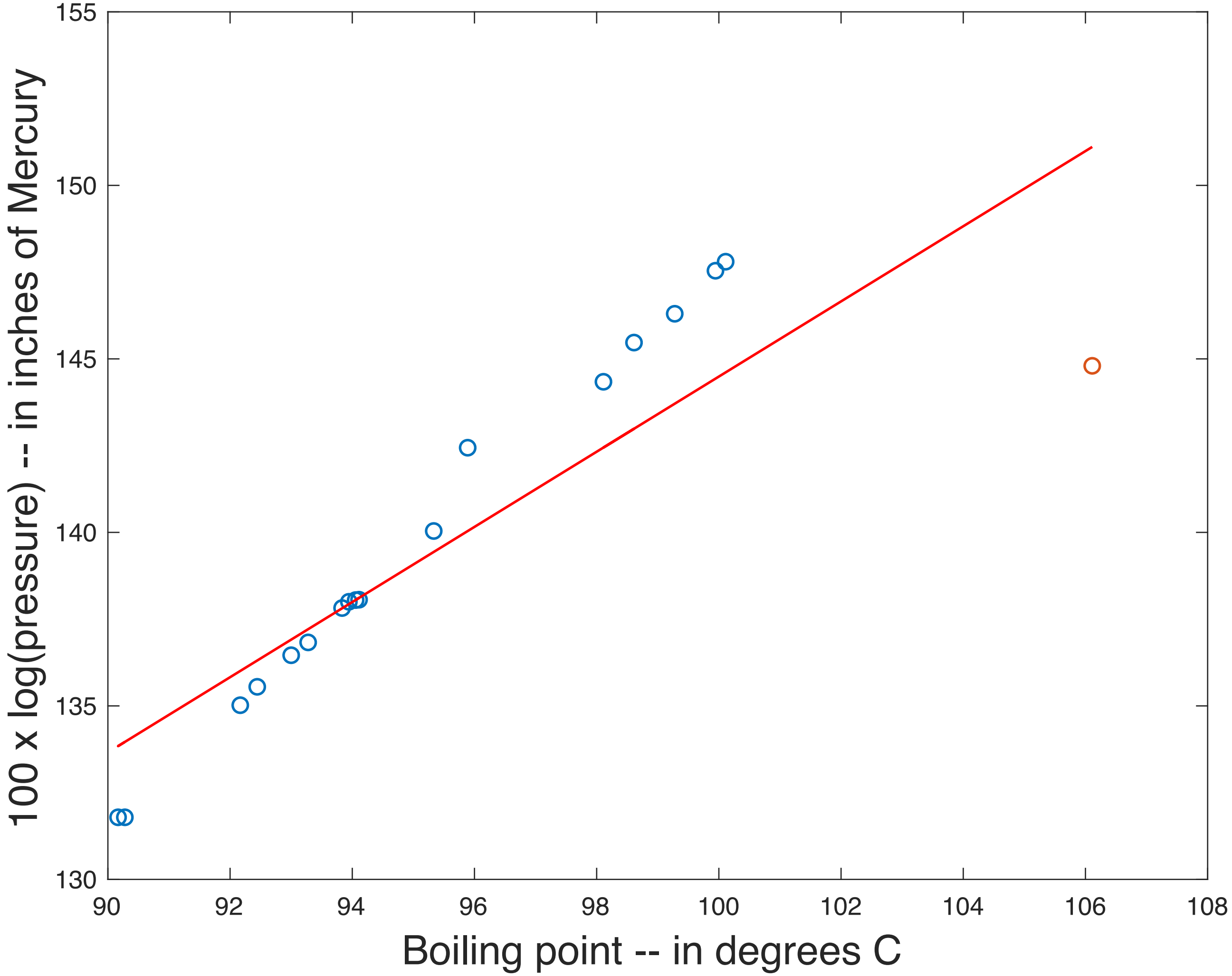
The smaller the differences, the better the model fits the data.

OUTLIERS IN LINEAR REGRESSION ANALYSIS

Plot the estimated OLS line

the outlier disrupts the fit !

```
b   = beta1;
fit = @(z) b(1) + b(2)*z;
hold all
plot(Xc,b(1)+b(2)*Xc,'r','LineWidth',1);
figure(f1);
```

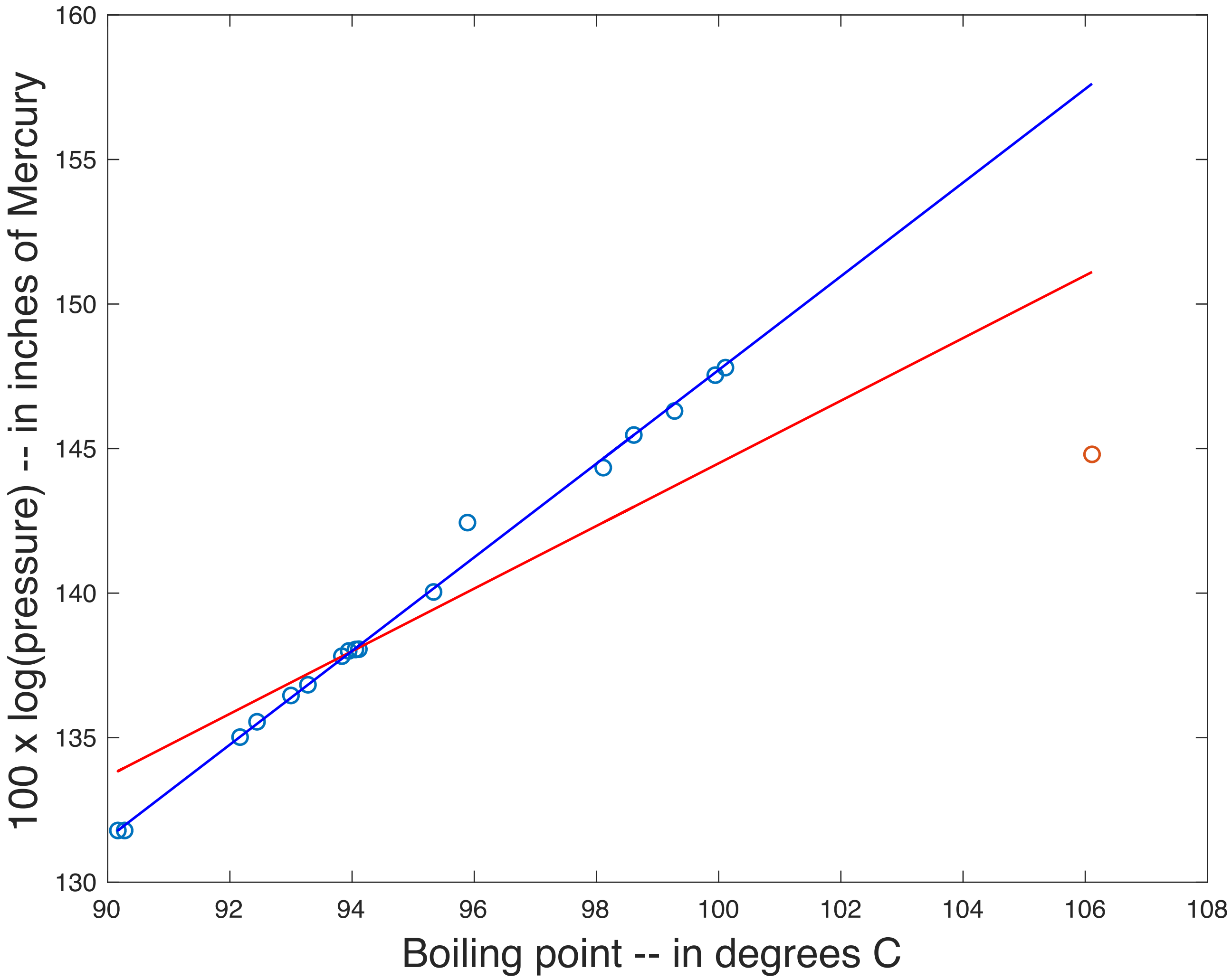


OUTLIERS IN LINEAR REGRESSION ANALYSIS

Now, let’s try a robust fit: the Least Trimmed Square (LTS) regression

LTS idea: Attempt to minimize the sum of squared residuals only over *a subset h of of the n points: typically, h = n/2.* The n – h points which are not used do not influence the fit.

```
[outLTS]=LXS(yc,Xc);
b = outLTS.beta;
plot(Xc,b(1)+b(2)*Xc,'b','LineWidth',1);
figure(f1);

% The LTS line fits
% very well the data!
```

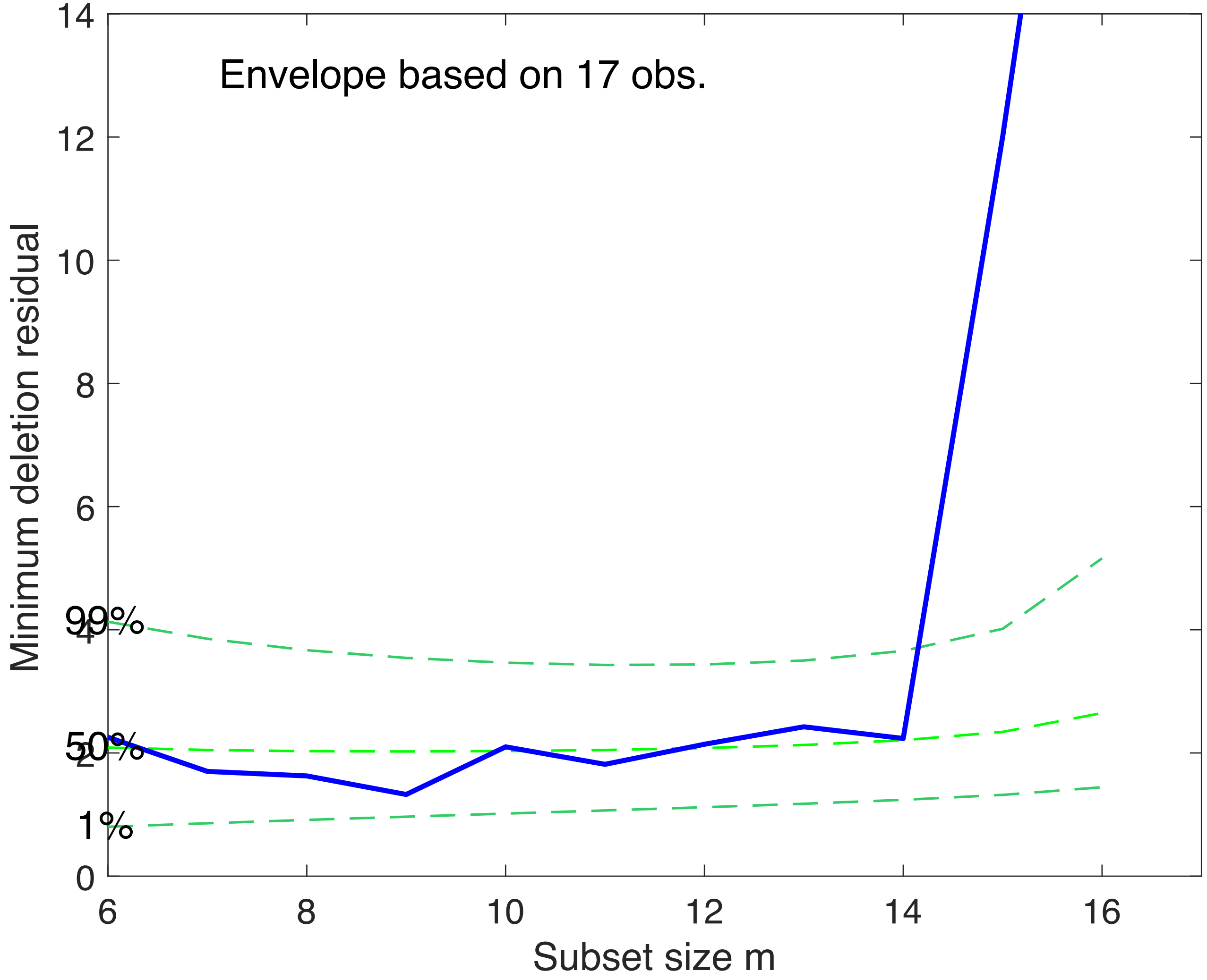


OUTLIERS IN LINEAR REGRESSION ANALYSIS

Now, let's try another robust fit: the Forward Search (FS) regression

FS idea: attempt to minimize the sum of squared residuals over *the largest possible outlier-free subset m of the n points.* The n – m points which are not used, are precisely the outliers.

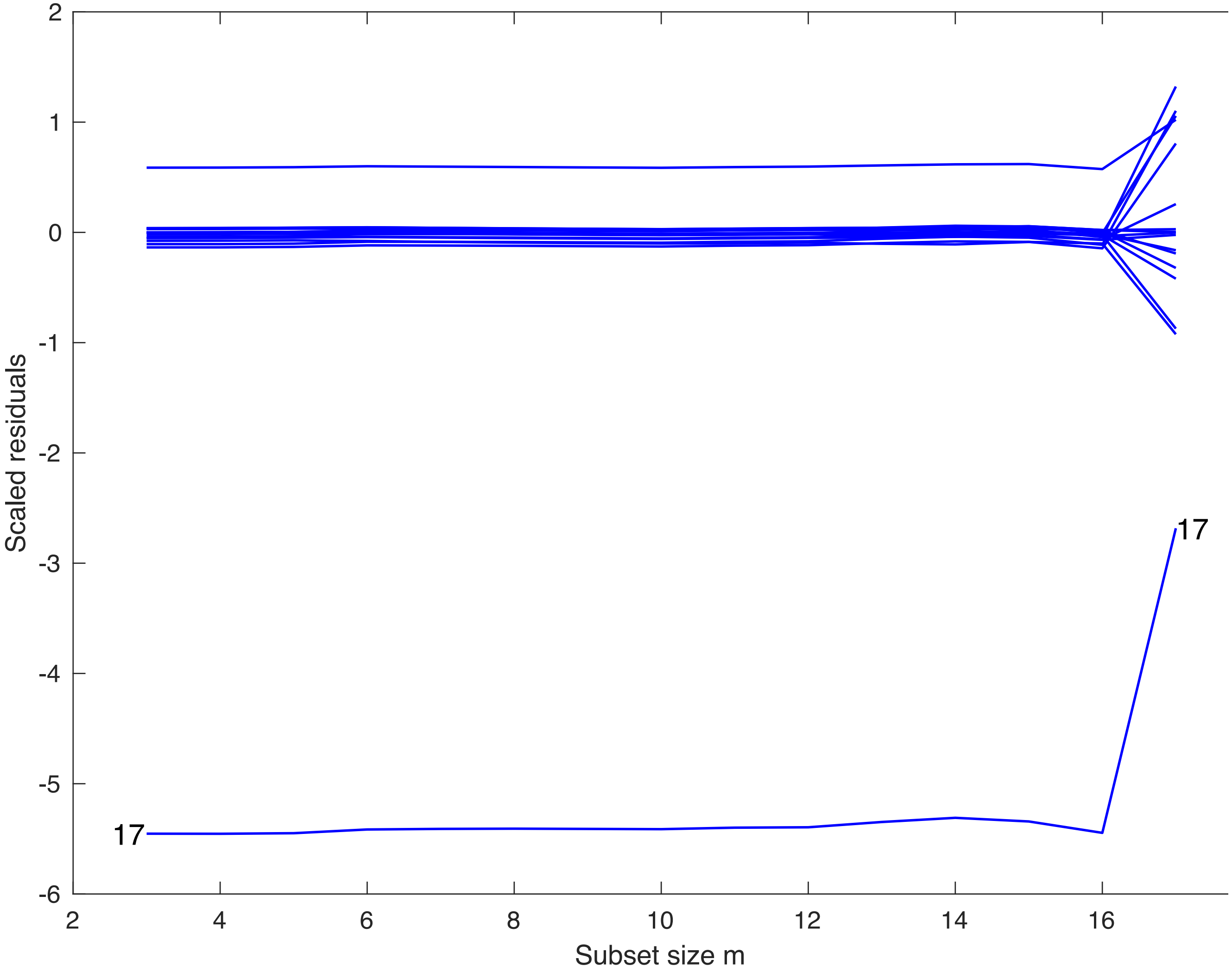
```
[out]=FSReda(yc,Xc,outLTS.bs);

% Plot minimum deletion residual (MDR)
close all;
mdrplot(out,'xlimx',[6 17],'ylimy',[0
14],'FontSize',16);
set(gca,'FontSize',14);

% The MDR plot shows a jump when the
% m-subset is formed by all the good
% units, and the outliers (i.e. the
% units outside the m-subset) have
% residuals very large with respect
% to those of the good units in the
% m-subset.
```



OUTLIERS IN LINEAR REGRESSION ANALYSIS

Now, some interactive brushing starting from the monitoring residual plot

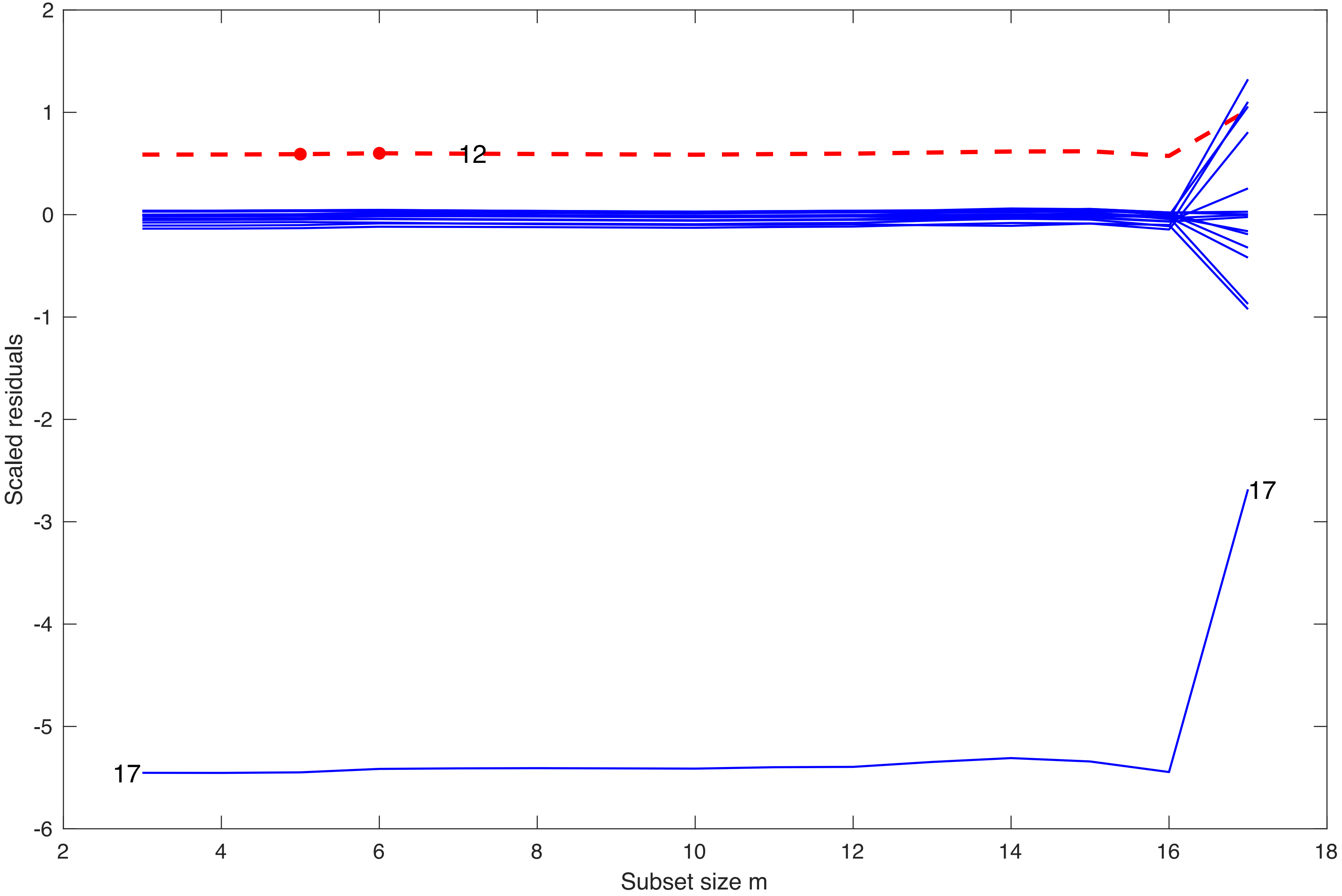
```
databrush=struct;
% Rectangular selection
databrush.selectionmode='Rect';
% Enable repeated mouse selections
databrush.persist='on';
% Write labels of trajectories while selecting
databrush.Label='on';
% Do not remove labels after selection
databrush.RemoveLabels='off';
%remove yellow
databrush.RemoveTool='on';

cascade;
resfwdplot(out,'databrush',databrush);

% Values of the residuals of all units
% while the m-subset increases.
```



Now, some interactive brushing starting from the monitoring residual plot

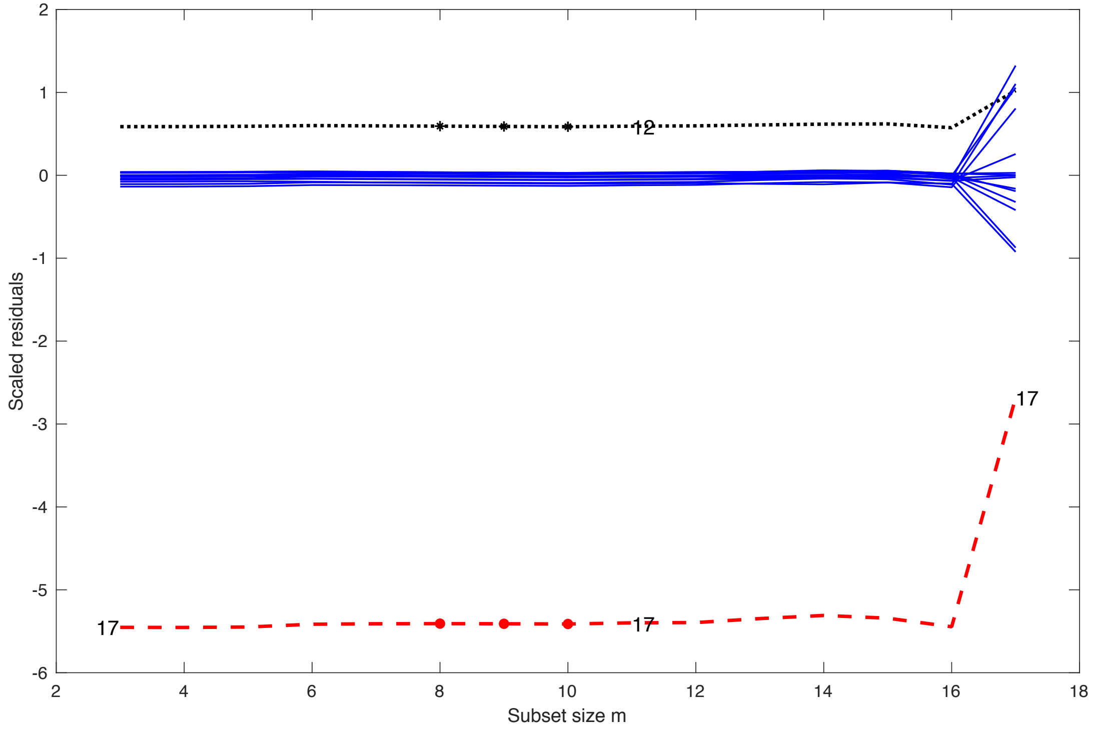
Note that, at the last step (rightmost side of the plot), only the residual of unit 17 (our contaminant) can be well separated from the others. Instead, the residual of unit 12 (red trajectory) at the last step enters in the fan of the other residuals.

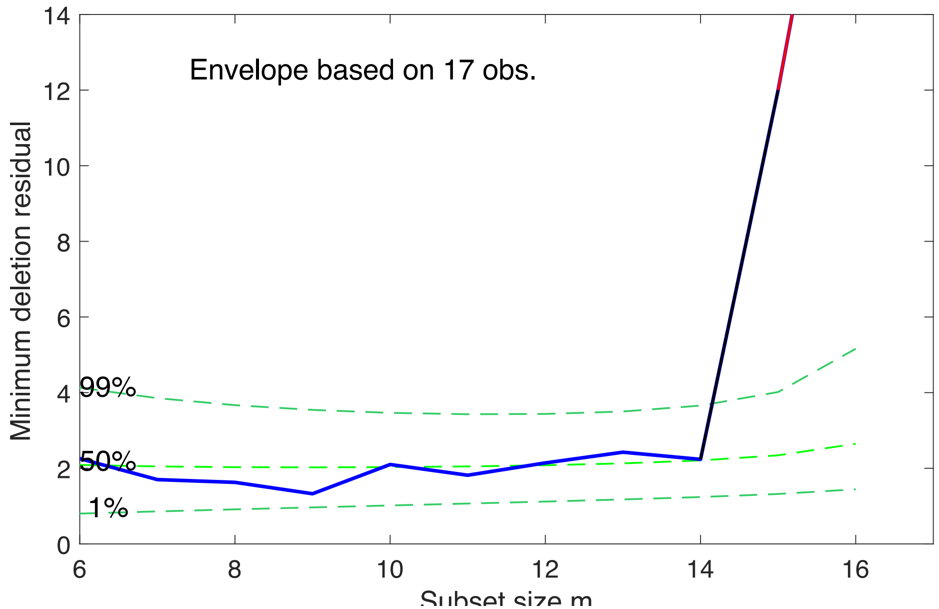
This means that unit 12 cannot be identified as outlying on the basis of the residuals estimated at the last step. Only if we look at the trajectories over several previous steps, we can realize that unit 12 is indeed an outlier.

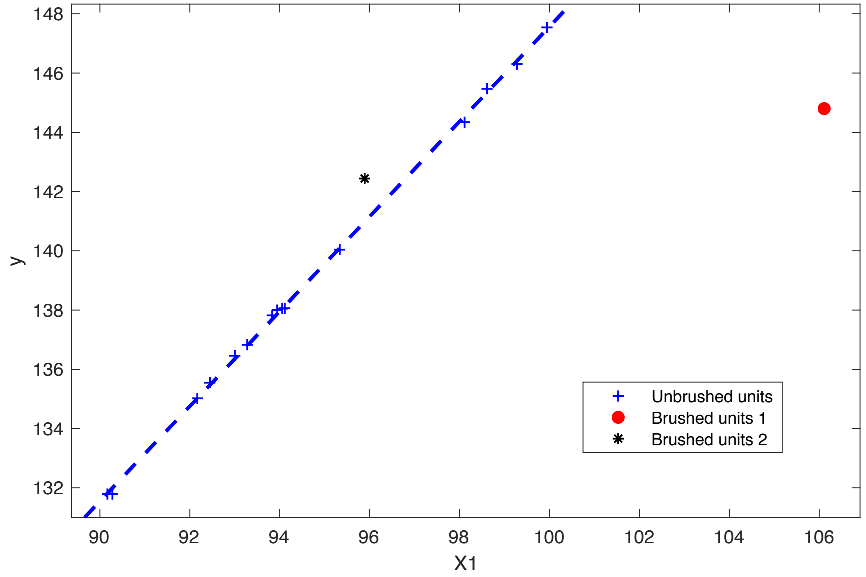


OUTLIERS IN LINEAR REGRESSION ANALYSIS

Once a set of trajectories is highlighted in the monitoring residual plot, the corresponding units are highlighted in the other plots.

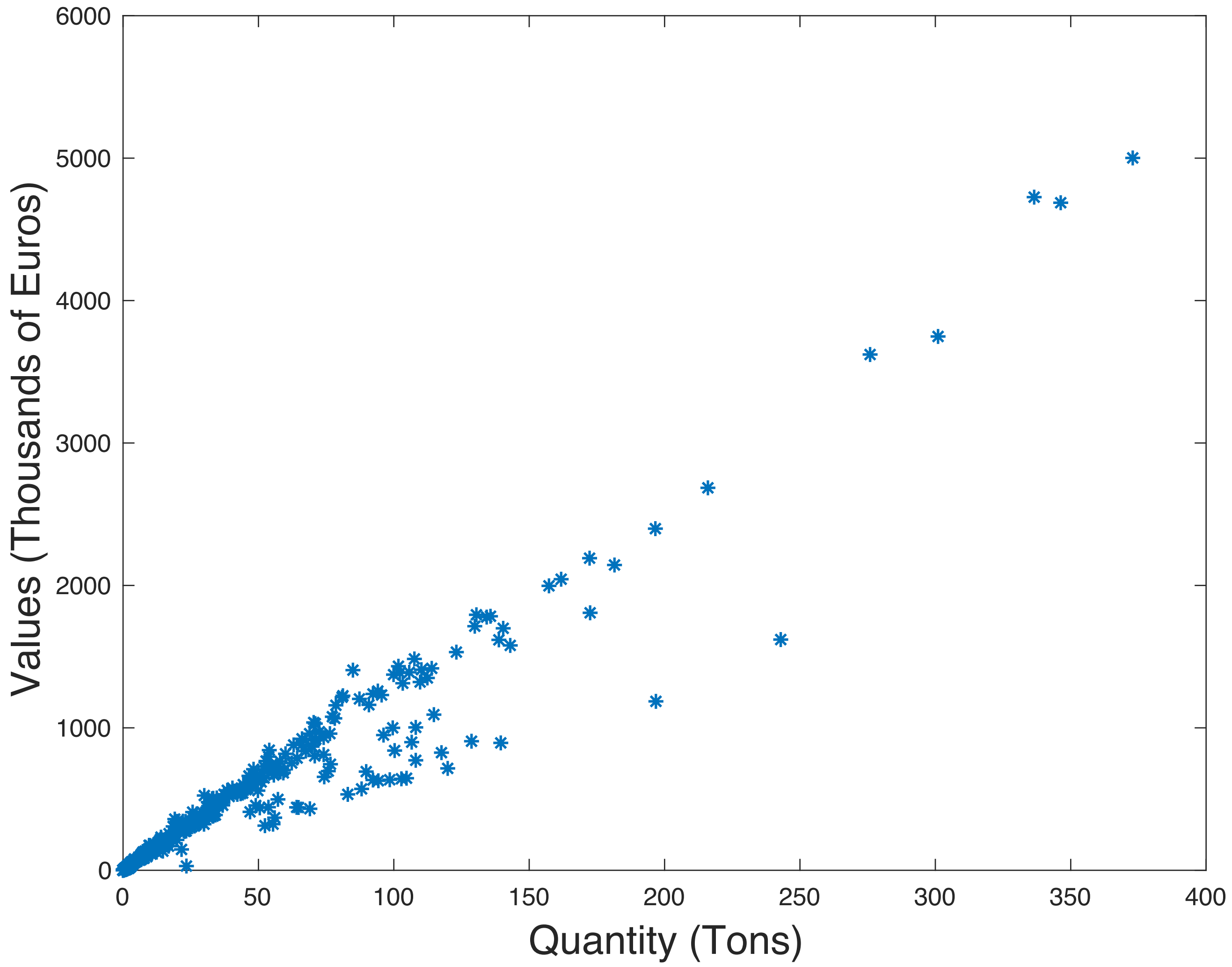






PRICE-OUTLIERS IN INTERNATIONAL TRADE DATA

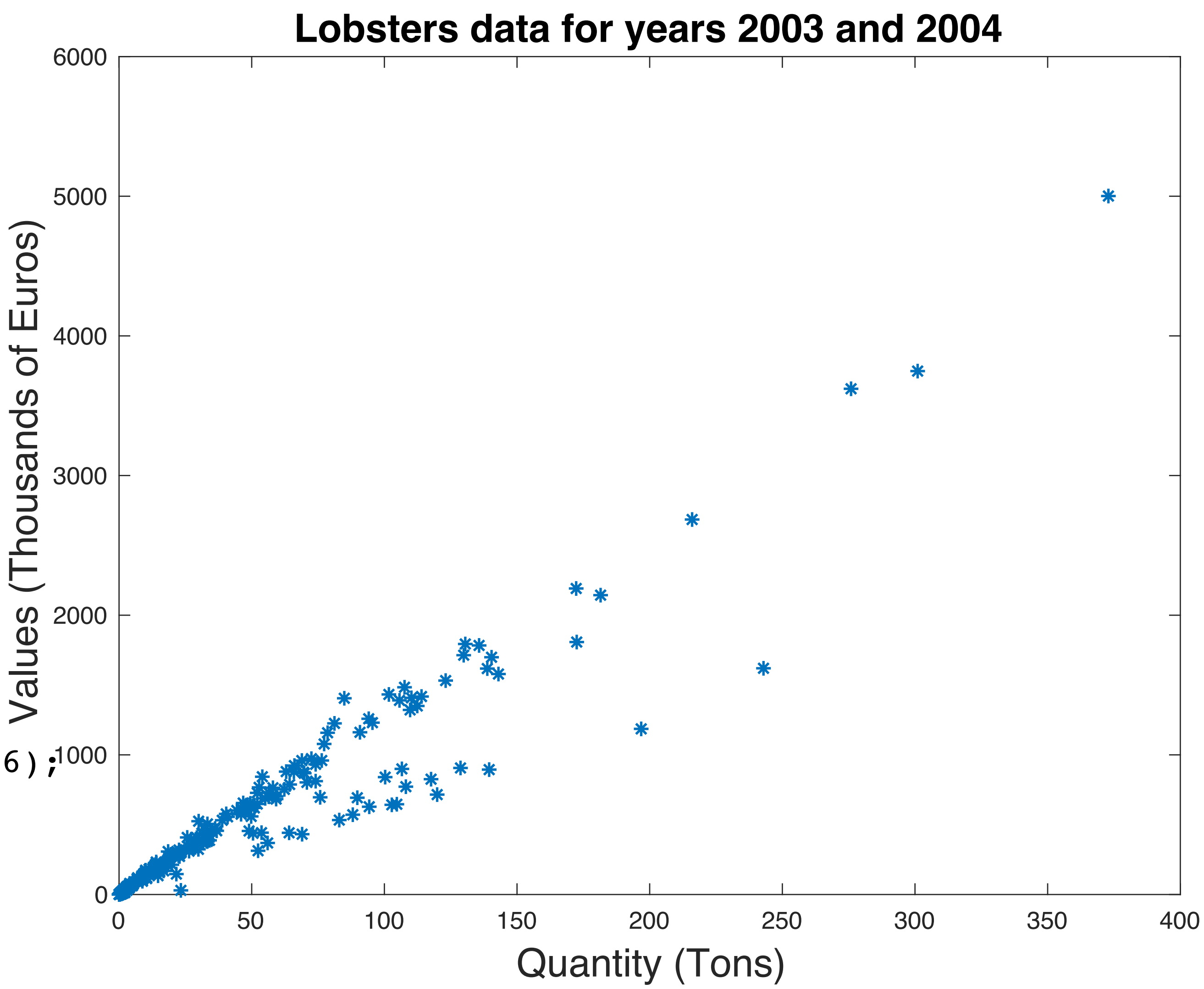
Fishery data: 677 transactions of a fishery product in Europe. For each transaction the Value in 1000 euro and the quantity in Tons are reported**.**



PRICE-OUTLIERS IN INTERNATIONAL TRADE DATA

Fishery data: A subset for years 2003 and 2004.

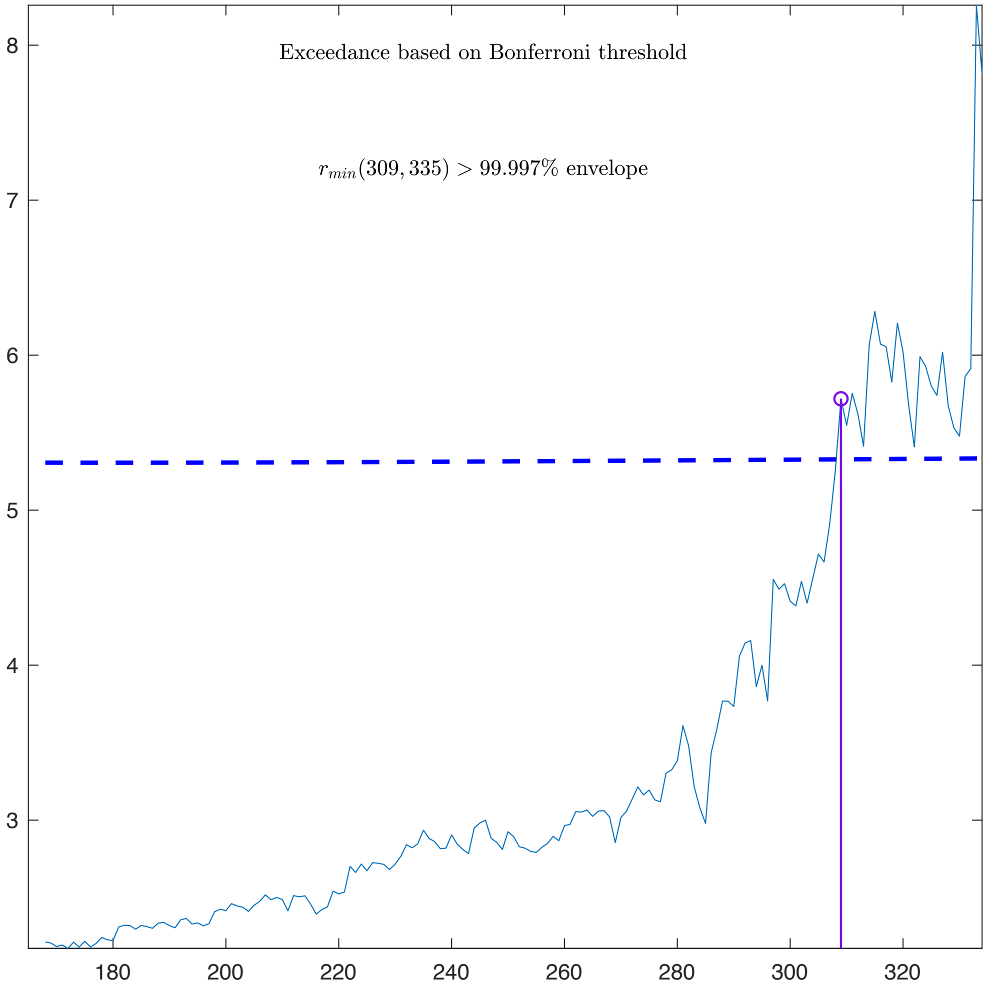
```
lobsters = readtable('lobsterIDyears.xlsx',
'ReadRowNames',true);
selected = find(or((lobsters.anno0 == 4) ,
(lobsters.anno0 == 3)));
lobstersS = lobsters(selected,:);
y = lobstersS.VALUE_1000EURO;
X = lobstersS.QUANTITY_TON;
close all;
figure;plot(X,y,'*');
xlabel('Quantity (Tons)','Fontsize',16);
ylabel('Values (Thousands of Euros)','Fontsize',16);
title('Lobsters data for years 2003 and
2004','FontSize',16);
f1 = gcf; figure(f1);
```

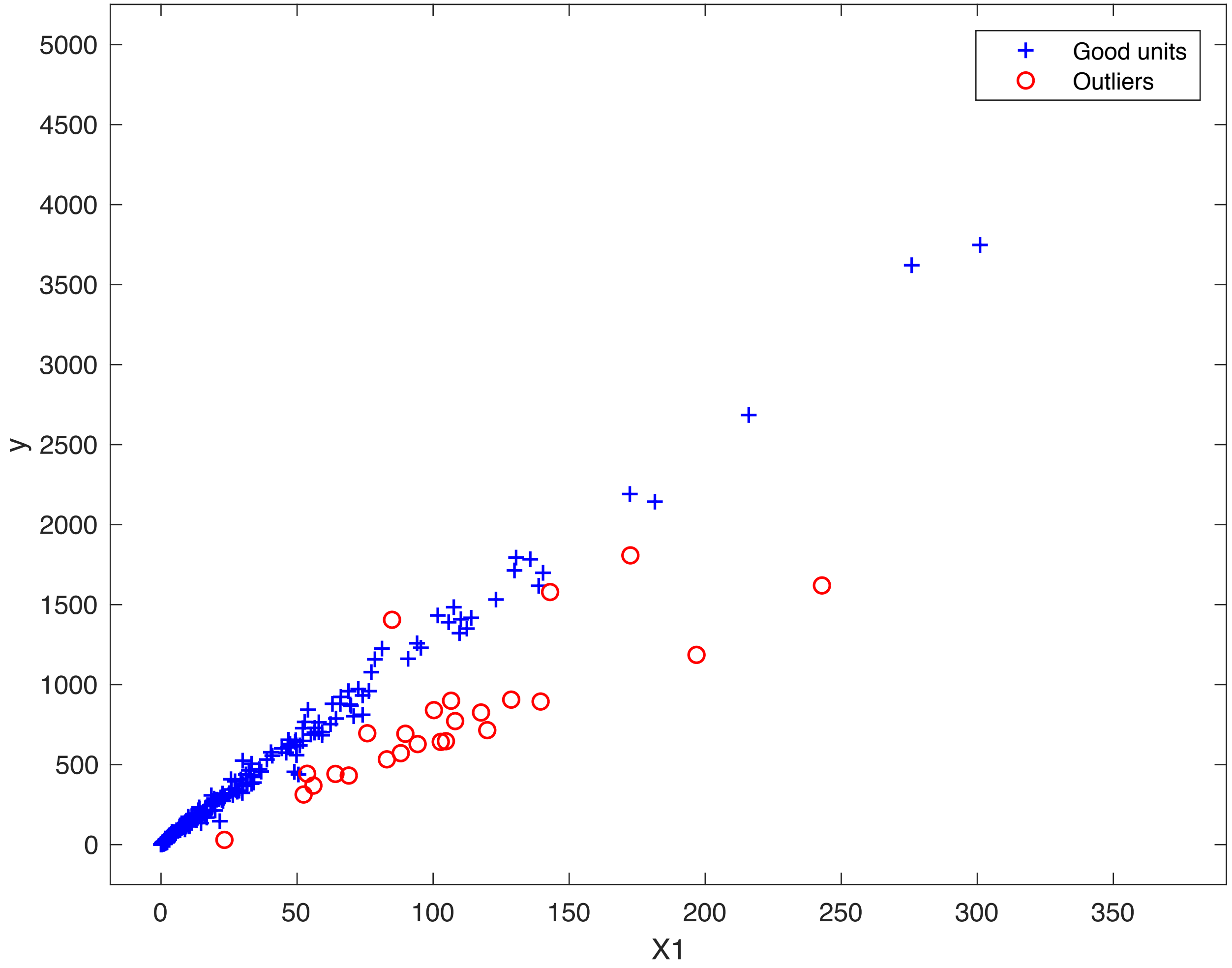


PRICE-OUTLIERS IN INTERNATIONAL TRADE DATA

Forward Search Regression applied to Fishery data

```
n = size(y,1);
[out]=FSRr(y,X,'plots',1,'bonflev',1-0.01/n);
pout = out.outliers;
```





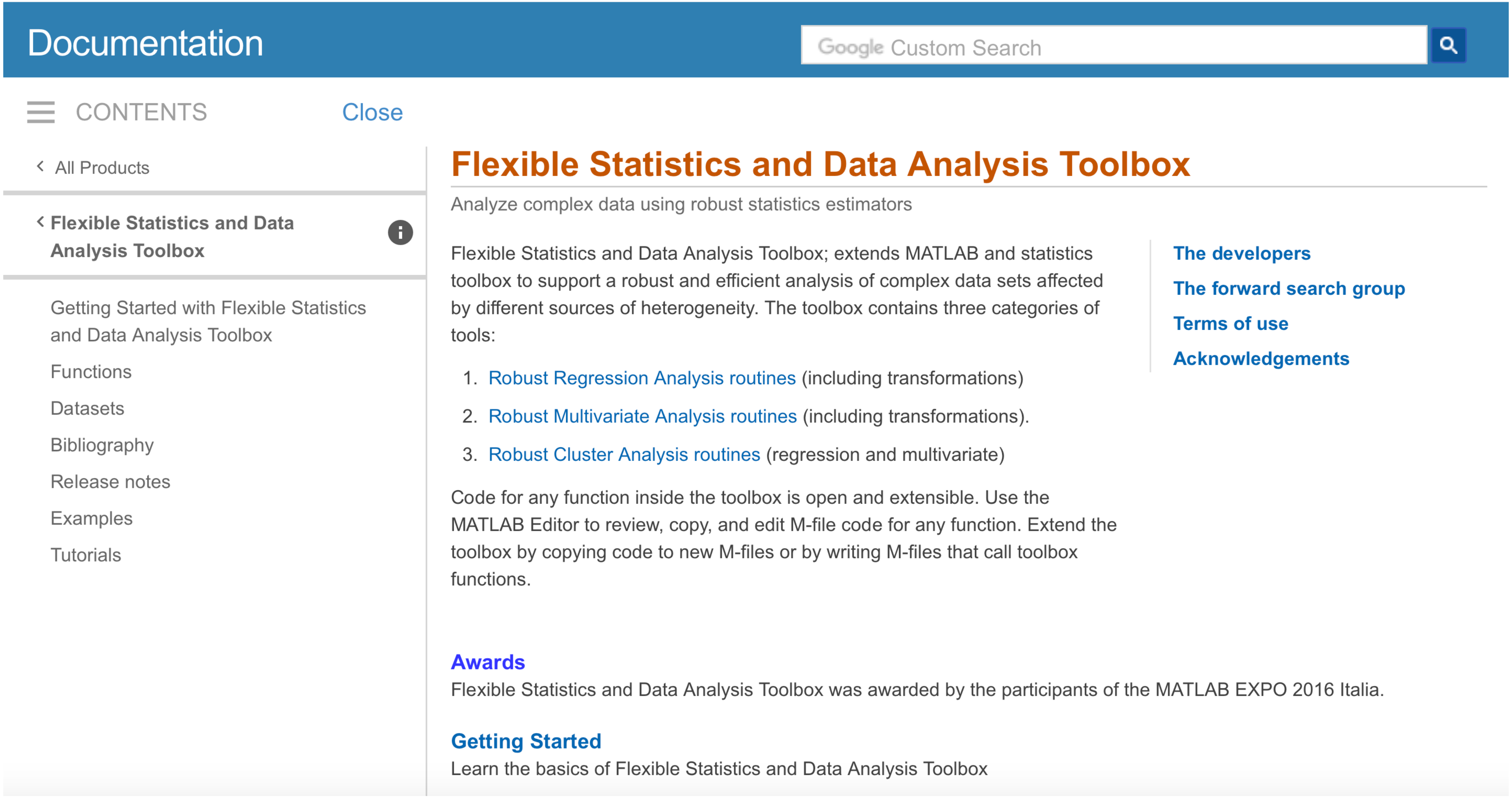
PRICE-OUTLIERS IN INTERNATIONAL TRADE DATA

Which Member States are associated to the price outliers?

```
lobstersS.DECLARANT_LAB(pout)
```

```
ans =  26×1 cell array

{'France' }    {'Greece' }    {'Spain'  }
{'Spain'  }    {'Spain'  }    {'Spain'  }
{'Spain'  }    {'Spain'  }    {'Spain'  }
{'Spain'  }    {'Spain'  }    {'Spain'  }
{'Spain'  }    {'Spain'  }    {'Spain'  }
{'Spain'  }    {'Spain'  }    {'Spain'  }
{'Spain'  }    {'Spain'  }    {'Spain'  }
{'Spain'  }    {'Spain'  }    {'Belgium'}
{'Belgium'}    {'Belgium'}
```

NOW, PLAY WITH FSDA BY YOUR OWN!



Flexible Statistics and Data Analysis Toolbox

Analyze complex data using robust statistics estimators

Flexible Statistics and Data Analysis Toolbox; extends MATLAB and statistics toolbox to support a robust and efficient analysis of complex data sets affected by different sources of heterogeneity. The toolbox contains three categories of tools:

1. Robust Regression Analysis routines (including transformations)
2. Robust Multivariate Analysis routines (including transformations).
3. Robust Cluster Analysis routines (regression and multivariate)

Code for any function inside the toolbox is open and extensible. Use the MATLAB Editor to review, copy, and edit M-file code for any function. Extend the toolbox by copying code to new M-files or by writing M-files that call toolbox functions.

The developers

The forward search group

Terms of use

Acknowledgements

Awards

Flexible Statistics and Data Analysis Toolbox was awarded by the participants of the MATLAB EXPO 2016 Italia.

Getting Started

Learn the basics of Flexible Statistics and Data Analysis Toolbox