Exercise 2-3 - Model Selection for mpg Data Set

Table of Contents

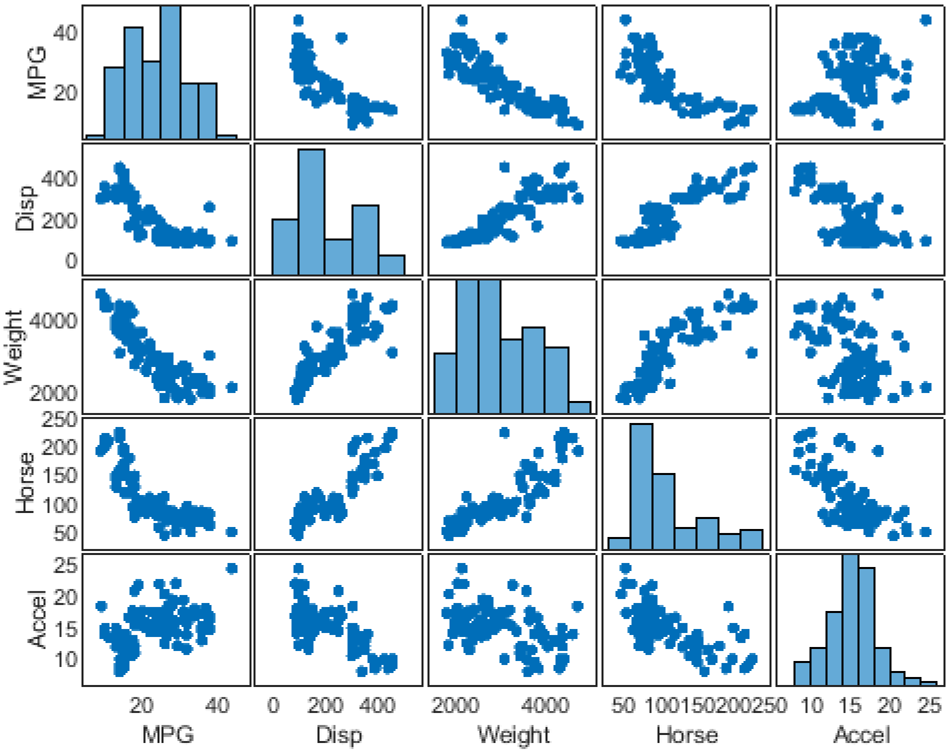
Felix Wittich, 01.06.2021

```
clear all
close all
clc
rng(1)
```

a)

```
%Load the carsmall data set and choose the displacement, weight, horsepower,
 and
%acceleration as potential input variables to predict mpg. Store all variables
 in a
%matrix and use data(any(isnan(data),2),:) = [] to get rid of NaN values. Plot
%the correlation between predictors and output and determine the corresponding
%correlation coefficients using corrcoef(). What can be concluded from the
 results?

load carsmall

data = [Displacement,Weight,Horsepower,Acceleration,MPG];
data(any(isnan(data),2),:) = [];

X = data(:,1:end-1);
Y = data(:,end);

figure
% plot correlation
[~,ax] = plotmatrix([Y X]);
% set labels (optional)
ylabel(ax(1,1),'MPG')
ylabel(ax(2,1),'Disp')
ylabel(ax(3,1),'Weight')
ylabel(ax(4,1),'Horse')
ylabel(ax(5,1),'Accel')
xlabel(ax(5,1),'MPG')
xlabel(ax(5,2),'Disp')
xlabel(ax(5,3),'Weight')
xlabel(ax(5,4),'Horse')
xlabel(ax(5,5),'Accel')
% determine correlation coefficients
R = corrcoef([Y X])
```

```
R =

    1.0000   -0.8048   -0.8591   -0.8028    0.4631
   -0.8048    1.0000    0.8860    0.9102   -0.6719
   -0.8591    0.8860    1.0000    0.8656   -0.4642
   -0.8028    0.9102    0.8656    1.0000   -0.6836
    0.4631   -0.6719   -0.4642   -0.6836    1.0000
```



b)

choose model type

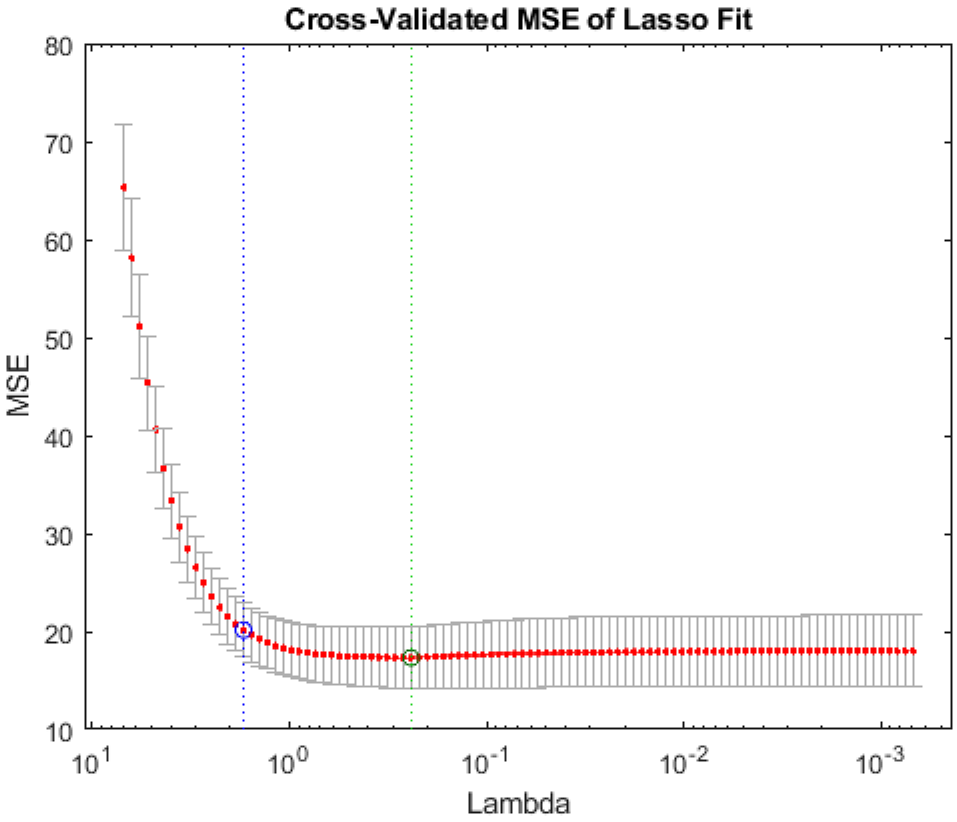
```
type = 'linear';

% perform stepwise selection using different criteria
myModelAIC = stepwiselm(X,Y,type,'Upper','linear','Criterion','AIC')
myModelBIC = stepwiselm(X,Y,type,'Upper','linear','Criterion','BIC')
myModelRsquared = stepwiselm(X,Y,type,'Upper','linear','Criterion','Rsquared')

% use lasso for selection
```

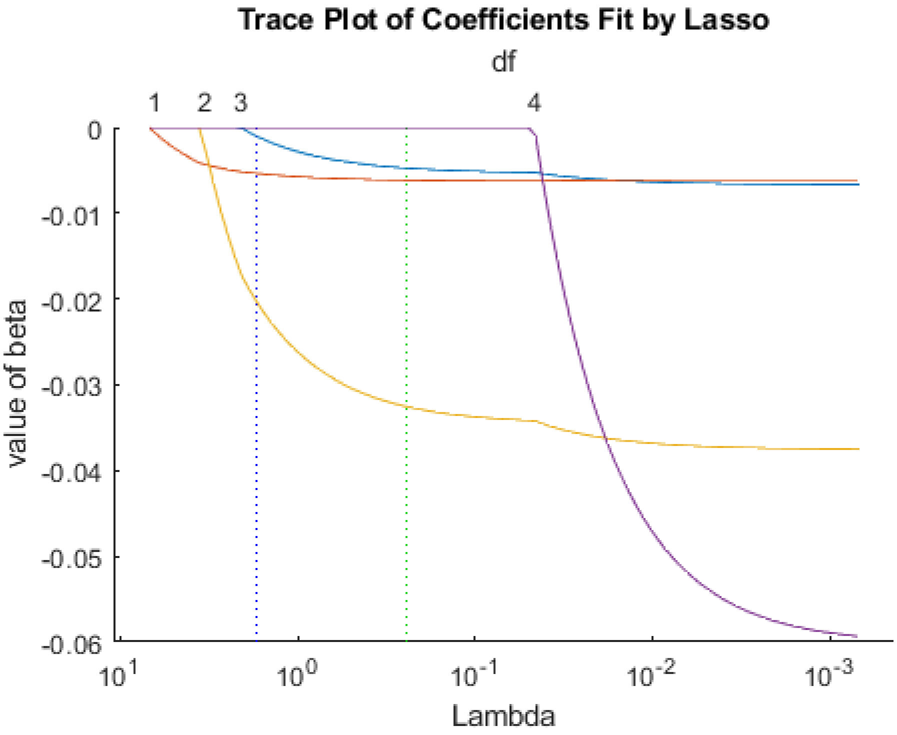
```
Phi = x2fx(X,'linear'); % generate regression matrix

[B, Stats] = lasso(Phi(:,2:end),Y,'CV',10); % estimate using lasso

lassoPlot(B, Stats, 'PlotType', 'CV') % plot CV results
lassoPlot(B, Stats, 'PlotType', 'Lambda','XScale','log') % plot coefficient
 path
ylabel('value of beta')


beta_0_Lasso = Stats.Intercept(Stats.IndexMinMSE);
BetaLasso = [beta_0_Lasso B(:,Stats.IndexMinMSE)']';

% re-estimate model selected by lasso in order to obtain unbiased estimate
myModelLasso = fitlm(X(:,BetaLasso(2:end)~=0),Y)

% estimate full model
myModel = fitlm(X,Y)

% evaluate models in CV
yFit = @(XTrain,yTrain,XTest)(XTest*regress(yTrain,XTrain));

Xtest = [ones(length(Y),1) X];
Ytest = Y;

cvMSEmyModelAIC = crossval('MSE',...
    Xtest(:,[true; myModelAIC.VariableInfo.InModel(1:end-1)]),...
    Ytest,'predfun',yFit);
cvRMSEmyModelAIC = sqrt(cvMSEmyModelAIC);

cvMSEmyModelBIC = crossval('MSE',...
    Xtest(:,[true; myModelBIC.VariableInfo.InModel(1:end-1)]),...
    Ytest,'predfun',yFit);
cvRMSEmyModelBIC = sqrt(cvMSEmyModelBIC);

cvMSEmyModelRsquared = crossval('MSE',...
    Xtest(:,[true; myModelRsquared.VariableInfo.InModel(1:end-1)]),...
    Ytest,'predfun',yFit);
cvRMSEmyModelRsquared = sqrt(cvMSEmyModelRsquared);

cvMSEmyModelLasso = crossval('MSE',...
    Xtest(:,BetaLasso~=0),Ytest,'predfun',yFit);
cvRMSEmyModelLasso = sqrt(cvMSEmyModelLasso);

cvMSEmyModel = crossval('MSE',...
    Xtest,Ytest,'predfun',yFit);
cvRMSEmyModel = sqrt(cvMSEmyModel);

% compare models
RowNames = {'myModelfull','myModelAIC','myModelBIC',...
    'myModelRsquared','myModelLasso'};
RMSE = [myModel.RMSE;myModelAIC.RMSE;myModelBIC.RMSE;...
    myModelRsquared.RMSE;myModelLasso.RMSE];
```

```
AIC = [myModel.ModelCriterion.AIC;...
    myModelAIC.ModelCriterion.AIC;myModelBIC.ModelCriterion.AIC;...
    myModelRsquared.ModelCriterion.AIC;myModelLasso.ModelCriterion.AIC];
BIC = [myModel.ModelCriterion.BIC;...
    myModelAIC.ModelCriterion.BIC;myModelBIC.ModelCriterion.BIC;...
    myModelRsquared.ModelCriterion.BIC;myModelLasso.ModelCriterion.BIC];
cvRMSE = [cvRMSEmyModel;cvRMSEmyModelAIC;cvRMSEmyModelBIC;...
    cvRMSEmyModelRsquared;cvRMSEmyModelLasso];
dimTheta = [1+size(X,2);...
    1+sum(myModelAIC.VariableInfo.InModel(1:end-1));...
    1+sum(myModelBIC.VariableInfo.InModel(1:end-1));...
    1+sum(myModelRsquared.VariableInfo.InModel(1:end-1));...
    1+sum(BetaLasso(2:end)~=0)];
Models = table(RMSE,AIC,BIC,cvRMSE,dimTheta,...
               'RowNames',RowNames)
```

```
1. Removing x4, AIC = 529.56
2. Removing x1, AIC = 527.83

myModelAIC = 


Linear regression model:
    y ~ 1 + x2 + x3

Estimated Coefficients:
                   Estimate         SE        tStat       pValue  
                  __________    _________    _______    __________

    (Intercept)       47.769       1.7417     27.427     1.751e-45
    x2            -0.0065651    0.0010507    -6.2484    1.3519e-08
    x3             -0.042018     0.018671    -2.2504       0.02686


Number of observations: 93, Error degrees of freedom: 90
Root Mean Squared Error: 4.07
R-squared: 0.752,  Adjusted R-Squared: 0.747
F-statistic vs. constant model: 136, p-value = 5.57e-28
1. Removing x4, BIC = 539.69
2. Removing x1, BIC = 535.42

myModelBIC = 


Linear regression model:
    y ~ 1 + x2 + x3

Estimated Coefficients:
                   Estimate         SE        tStat       pValue  
                  __________    _________    _______    __________

    (Intercept)       47.769       1.7417     27.427     1.751e-45
    x2            -0.0065651    0.0010507    -6.2484    1.3519e-08
    x3             -0.042018     0.018671    -2.2504       0.02686
```

```
Number of observations: 93, Error degrees of freedom: 90
Root Mean Squared Error: 4.07
R-squared: 0.752,  Adjusted R-Squared: 0.747
F-statistic vs. constant model: 136, p-value = 5.57e-28
1. Removing x4, Rsquared = 0.75277
2. Removing x1, Rsquared = 0.75205
3. Removing x3, Rsquared = 0.7381

myModelRsquared = 


Linear regression model:
    y ~ 1 + x2

Estimated Coefficients:
                    Estimate         SE         tStat       pValue  
                   __________    __________    _______    __________

    (Intercept)        49.238        1.6504     29.834    9.0258e-49
    x2             -0.0086118    0.00053775    -16.014    3.2405e-28


Number of observations: 93, Error degrees of freedom: 91
Root Mean Squared Error: 4.16
R-squared: 0.738,  Adjusted R-Squared: 0.735
F-statistic vs. constant model: 256, p-value = 3.24e-28

myModelLasso = 


Linear regression model:
    y ~ 1 + x1 + x2 + x3

Estimated Coefficients:
                    Estimate        SE         tStat        pValue  
                   __________    _________    ________    __________

    (Intercept)        47.182       2.0973      22.497    1.7329e-38
    x1             -0.0053631     0.010574    -0.50719       0.61328
    x2             -0.0062775    0.0011978     -5.2409    1.0646e-06
    x3              -0.034562     0.023824     -1.4507       0.15037


Number of observations: 93, Error degrees of freedom: 89
Root Mean Squared Error: 4.08
R-squared: 0.753,  Adjusted R-Squared: 0.744
F-statistic vs. constant model: 90.3, p-value = 6.51e-27

myModel = 


Linear regression model:
    y ~ 1 + x1 + x2 + x3 + x4

Estimated Coefficients:
                   Estimate         SE         tStat       pValue  
                   __________    _________    ________    __________

    (Intercept)       48.117       3.9008       12.335    6.9194e-21
    x1            -0.0066826     0.011594     -0.57638       0.56583
    x2             -0.006084    0.0013823      -4.4014      3.01e-05
    x3             -0.037547     0.026139      -1.4364       0.15442
    x4             -0.060312      0.21167     -0.28493       0.77636


Number of observations: 93, Error degrees of freedom: 88
Root Mean Squared Error: 4.11
R-squared: 0.753,  Adjusted R-Squared: 0.742
F-statistic vs. constant model: 67.1, p-value = 6.49e-26

Models =

  5×5 table

                       RMSE      AIC       BIC      cvRMSE    dimTheta
                       ______    ______    ______    ______    ________

    myModelfull        4.1054    531.47    544.14    4.3775       5    
    myModelAIC         4.0673    527.83    535.42    4.1023       3    
    myModelBIC         4.0673    527.83    535.42    4.1464       3    
    myModelRsquared    4.1571    530.92    535.98    4.2076       2    
    myModelLasso       4.0842    529.56    539.69    4.2813       4    
```


Cross-Validated MSE of Lasso Fit
MSE
80
70
60
50
40
30
20
10
10^1
10^0
10^{-1}
10^{-2}
10^{-3}
Lambda


Trace Plot of Coefficients Fit by Lasso
df
1
2
3
4
value of beta
0
-0.01
-0.02
-0.03
-0.04
-0.05
-0.06
10^1
10^0
10^{-1}
10^{-2}
10^{-3}
Lambda

c)

```
clc, close all

X = data(:,3);

% choose model type
type = 'poly9';

% perform stepwise selection using differenkt criteria
myModelAIC = stepwiselm(X,Y,type,'Criterion','AIC')
myModelBIC = stepwiselm(X,Y,type,'Criterion','BIC')
myModelRsquared = stepwiselm(X,Y,type,'Criterion','Rsquared')

% use lasso for selection
for li = 1:9
    PhiPoly(:,li) = X.^li; % generate regression matrix
end

X = PhiPoly;

[B, Stats] = lasso(X,Y,'CV',10); % estimate using lasso

lassoPlot(B, Stats, 'PlotType', 'CV') % plot CV results
lassoPlot(B, Stats, 'PlotType', 'Lambda','XScale','log',...
    'PredictorNames',{'x1','x2','x3','x4','x5','x6','x7','x8','x9'}) % plot
 coefficient path
ylabel('value of beta')

beta_0_Lasso = Stats.Intercept(Stats.IndexMinMSE);
BetaLasso = [beta_0_Lasso B(:,Stats.IndexMinMSE)']';

% re-estimate model selected by lasso in order to obtain unbiased estimate
myModelLasso = fitlm(X(:,BetaLasso(2:end)~=0),Y);

% estimate full model
myModel = fitlm(X,Y)

% estimate linear model
myModelLinear = fitlm(X(:,1),Y)

% evaluate models in CV
yFit = @(XTrain,yTrain,XTest)(XTest*regress(yTrain,XTrain));

Xtest = [ones(length(Y),1) X];
Ytest = Y;

inModelAIC = false(9,1);
inModelAIC(myModelAIC.Formula.Terms(2:end,1)) = true;
inModelBIC = false(9,1);
inModelBIC(myModelBIC.Formula.Terms(2:end,1)) = true;
inModelRsquared = false(9,1);
inModelRsquared(myModelRsquared.Formula.Terms(2:end,1)) = true;
```

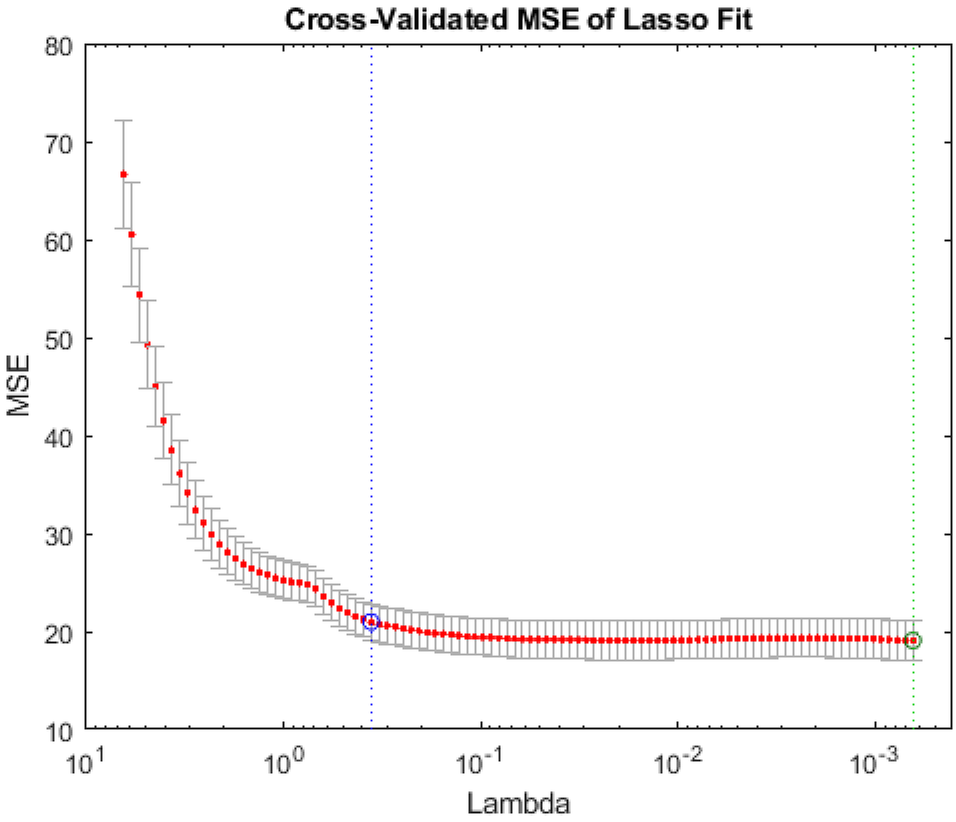
```
cvMSEmyModelAIC = crossval('MSE',...
    Xtest(:,[true; inModelAIC]),...
    Ytest,'predfun',yFit);
cvRMSEmyModelAIC = sqrt(cvMSEmyModelAIC);

cvMSEmyModelBIC = crossval('MSE',...
    Xtest(:,[true; inModelBIC]),...
    Ytest,'predfun',yFit);
cvRMSEmyModelBIC = sqrt(cvMSEmyModelBIC);

cvMSEmyModelRsquared = crossval('MSE',...
    Xtest(:,[true; inModelRsquared]),...
    Ytest,'predfun',yFit);
cvRMSEmyModelRsquared = sqrt(cvMSEmyModelRsquared);

cvMSEmyModelLasso = crossval('MSE',...
    Xtest(:,BetaLasso~=0),Ytest,'predfun',yFit);
cvRMSEmyModelLasso = sqrt(cvMSEmyModelLasso);

cvMSEmyModel = crossval('MSE',...
    Xtest,Ytest,'predfun',yFit);
cvRMSEmyModel = sqrt(cvMSEmyModel);

cvMSEmyModelLinear = crossval('MSE',...
    Xtest(:,[1,2]),Ytest,'predfun',yFit);
cvRMSEmyModelLinear = sqrt(cvMSEmyModelLinear);

% compare models
RowNames = {'myModelLinear','myModelfull','myModelAIC','myModelBIC',...
    'myModelRsquared','myModelLasso'};
RMSE = [myModelLinear.RMSE;myModel.RMSE;myModelAIC.RMSE;myModelBIC.RMSE;...
    myModelRsquared.RMSE;myModelLasso.RMSE];
AIC = [myModelLinear.ModelCriterion.AIC;myModel.ModelCriterion.AIC;...
    myModelAIC.ModelCriterion.AIC;myModelBIC.ModelCriterion.AIC;...
    myModelRsquared.ModelCriterion.AIC;myModelLasso.ModelCriterion.AIC];
BIC = [myModelLinear.ModelCriterion.BIC;myModel.ModelCriterion.BIC;...
    myModelAIC.ModelCriterion.BIC;myModelBIC.ModelCriterion.BIC;...
    myModelRsquared.ModelCriterion.BIC;myModelLasso.ModelCriterion.BIC];
cvRMSE = [cvRMSEmyModelLinear;cvRMSEmyModel;cvRMSEmyModelAIC;...
    cvRMSEmyModelBIC;cvRMSEmyModelRsquared;cvRMSEmyModelLasso];
dimTheta = [2;1+size(X,2);...
    1+sum(inModelAIC);...
    1+sum(inModelBIC);...
    1+sum(inModelRsquared);...
    1+sum(BetaLasso(2:end)~=0)];
Models = table(RMSE,AIC,BIC,cvRMSE,dimTheta,...
               'RowNames',RowNames)
```

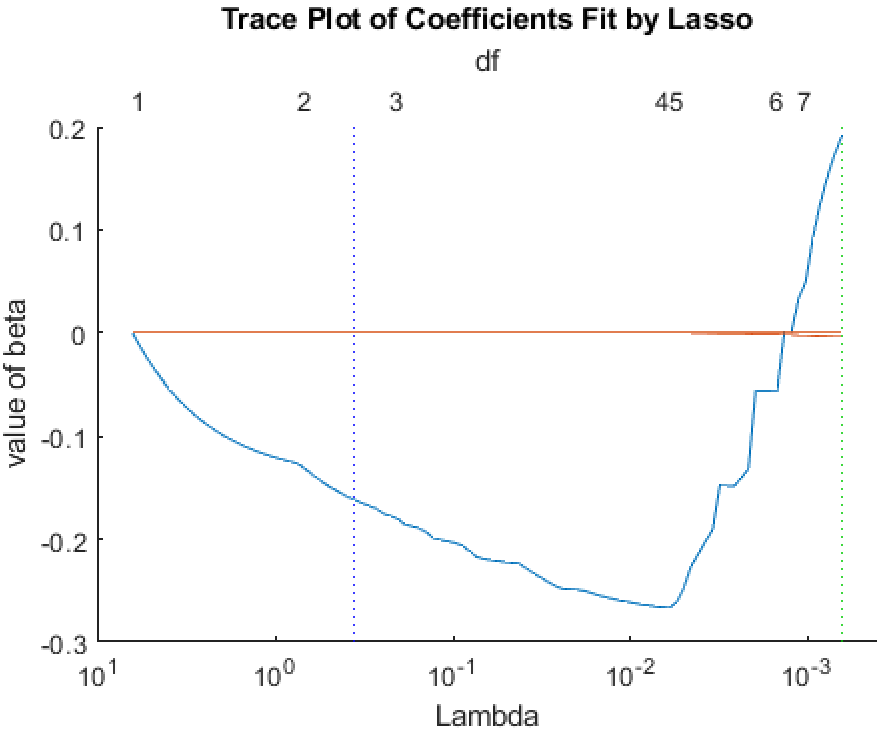
```
1. Removing x1^9, AIC = 653.37
2. Removing x1^8, AIC = 578.88
3. Removing x1^7, AIC = 538.73
4. Removing x1^6, AIC = 534.65
```

```
myModelAIC = 


Linear regression model:
    y ~ 1 + x1 + x1^2 + x1^3 + x1^4 + x1^5

Estimated Coefficients:
                    Estimate         SE          tStat       pValue  
                   __________    __________    _______    _________

    (Intercept)       -116.99          55.28    -2.1163     0.037179
    x1                 7.1711         2.5321     2.8321    0.0057444
    x1^2             -0.12384       0.043688    -2.8346    0.0057036
    x1^3           0.00096408     0.00035581     2.7096    0.0081136
    x1^4           -3.526e-06      1.376e-06    -2.5624     0.012115
    x1^5            4.929e-09     2.0343e-09      2.423     0.017466


Number of observations: 93, Error degrees of freedom: 87
Root Mean Squared Error: 4.16
R-squared: 0.75,  Adjusted R-Squared: 0.735
F-statistic vs. constant model: 52.2, p-value = 9.82e-25
1. Removing x1^9, BIC = 676.16
2. Removing x1^8, BIC = 599.14
3. Removing x1^7, BIC = 556.46
4. Removing x1^6, BIC = 549.85

myModelBIC = 


Linear regression model:
    y ~ 1 + x1 + x1^2 + x1^3 + x1^4 + x1^5

Estimated Coefficients:
                    Estimate         SE          tStat       pValue  
                   __________    __________    _______    _________

    (Intercept)       -116.99          55.28    -2.1163     0.037179
    x1                 7.1711         2.5321     2.8321    0.0057444
    x1^2             -0.12384       0.043688    -2.8346    0.0057036
    x1^3           0.00096408     0.00035581     2.7096    0.0081136
    x1^4           -3.526e-06      1.376e-06    -2.5624     0.012115
    x1^5            4.929e-09     2.0343e-09      2.423     0.017466


Number of observations: 93, Error degrees of freedom: 87
Root Mean Squared Error: 4.16
R-squared: 0.75,  Adjusted R-Squared: 0.735
F-statistic vs. constant model: 52.2, p-value = 9.82e-25
1. Removing x1^9, Rsquared = 0.15942
2. Removing x1^8, Rsquared = 0.61445
3. Removing x1^7, Rsquared = 0.74419
4. Removing x1^6, Rsquared = 0.74985
5. Removing x1^5, Rsquared = 0.73297
```

```
6. Removing x1^4, Rsquared = 0.72113
7. Removing x1^3, Rsquared = 0.71719

myModelRsquared = 


Linear regression model:
    y ~ 1 + x1 + x1^2

Estimated Coefficients:
                   Estimate         SE         tStat       pValue  
                   _________    __________    _______    __________

    (Intercept)       55.637         3.584     15.524    3.4677e-27
    x1              -0.42668      0.059821    -7.1327    2.3996e-10
    x1^2           0.0010538    0.00021911     4.8097    6.0387e-06


Number of observations: 93, Error degrees of freedom: 90
Root Mean Squared Error: 4.34
R-squared: 0.717,  Adjusted R-Squared: 0.711
F-statistic vs. constant model: 114, p-value = 2.08e-25
Warning: Regression design matrix is rank deficient to within machine
precision. 
Warning: Regression design matrix is rank deficient to within machine
precision. 

myModel = 


Linear regression model:
    y ~ 1 + x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9

Estimated Coefficients:
                    Estimate          SE         tStat       pValue  
                   ___________    __________    _______    __________

    (Intercept)              0             0        NaN           NaN
    x1                       0             0        NaN           NaN
    x2                       0             0        NaN           NaN
    x3                       0             0        NaN           NaN
    x4                       0             0        NaN           NaN
    x5               8.197e-08    6.6973e-09     12.239    2.7926e-20
    x6             -1.8226e-09     1.681e-10    -10.843    1.4192e-17
    x7              1.5108e-11    1.5433e-12     9.7896    1.7263e-15
    x8             -5.5226e-14    6.1514e-15    -8.9777    7.2365e-14
    x9              7.5061e-17    9.0032e-18     8.3371    1.3835e-12


Number of observations: 93, Error degrees of freedom: 88
Root Mean Squared Error: 10.5
R-squared: -0.615,  Adjusted R-Squared: -0.688
F-statistic vs. constant model: -8.38, p-value = 0
```

```
myModelLinear = 


Linear regression model:
    y ~ 1 + x1

Estimated Coefficients:
                   Estimate       SE        tStat       pValue  
                   ________    ________    _______    __________

    (Intercept)     39.362       1.3169     29.889    7.7492e-49
    x1              -0.143     0.011134    -12.844    3.7813e-22


Number of observations: 93, Error degrees of freedom: 91
Root Mean Squared Error: 4.84
R-squared: 0.644,  Adjusted R-Squared: 0.641
F-statistic vs. constant model: 165, p-value = 3.78e-22
Warning: X is rank deficient to within machine precision.
Warning: X is rank deficient to within machine precision.
Warning: X is rank deficient to within machine precision.
Warning: X is rank deficient to within machine precision.
Warning: X is rank deficient to within machine precision.
Warning: X is rank deficient to within machine precision.
Warning: X is rank deficient to within machine precision.
Warning: X is rank deficient to within machine precision.
Warning: X is rank deficient to within machine precision.
Warning: X is rank deficient to within machine precision.
Warning: X is rank deficient to within machine precision.
Warning: X is rank deficient to within machine precision.
Warning: X is rank deficient to within machine precision.
Warning: X is rank deficient to within machine precision.
Warning: X is rank deficient to within machine precision.
Warning: X is rank deficient to within machine precision.
Warning: X is rank deficient to within machine precision.
Warning: X is rank deficient to within machine precision.
Warning: X is rank deficient to within machine precision.
Warning: X is rank deficient to within machine precision.

Models =

  6×5 table

                        RMSE      AIC       BIC      cvRMSE    dimTheta
                       ______    ______    ______    ______    ________

    myModelLinear      4.8434    559.34     564.4    4.9176        2   
    myModelfull        10.498     716.1    741.42     10.69       10   
    myModelAIC         4.1552    534.65    549.85    4.2763        6   
    myModelBIC         4.1552    534.65    549.85    4.3851        6   
    myModelRsquared    4.3439    540.06    547.66     4.431        3   
    myModelLasso       13.134    754.82    775.08    13.287        8   
```


Cross-Validated MSE of Lasso Fit
MSE
80
70
60
50
40
30
20
10
10^1
10^0
10^{-1}
10^{-2}
10^{-3}
Lambda


Trace Plot of Coefficients Fit by Lasso
df
1
2
3
45
6 7
value of beta
0.2
0.1
0
-0.1
-0.2
-0.3
10^1
10^0
10^{-1}
10^{-2}
10^{-3}
Lambda

Published with MATLAB® R2021b