Exercise 2-1 - Getting Familiar with Stepwise Model Selection in Matlab

Table of Contents

Felix Wittich, 01.06.2021

```
clear all
close all
clc
```

a)

Use the randn() function to generate a predictor Xtrain of length N = 1000, as well as a noise vector eps of length N = 1000 with a standard deviation of 0.8. Make sure to use rng(1000) prior to staring part a) to ensure consistent results.

```
rng(1000)

N = 1000;
X = randn(N,1);
eps = 0.8*randn(N,1);
```

b)

Generate a response vector Y of length n = 100 according to the model Y = beta_0 + beta_1X + beta_2X^2 + beta_3X^3 + epsilon, where beta_0=2, beta_1=3, beta_2=-1, and beta_3=0.5. Set seed to 1998.

```
beta_0 = 2;
beta_1 = 3;
beta_2 = -1;
beta_3 = 2;
Y = beta_0 + beta_1*X + beta_2*X.^2 + beta_3*X.^3 + eps;
```

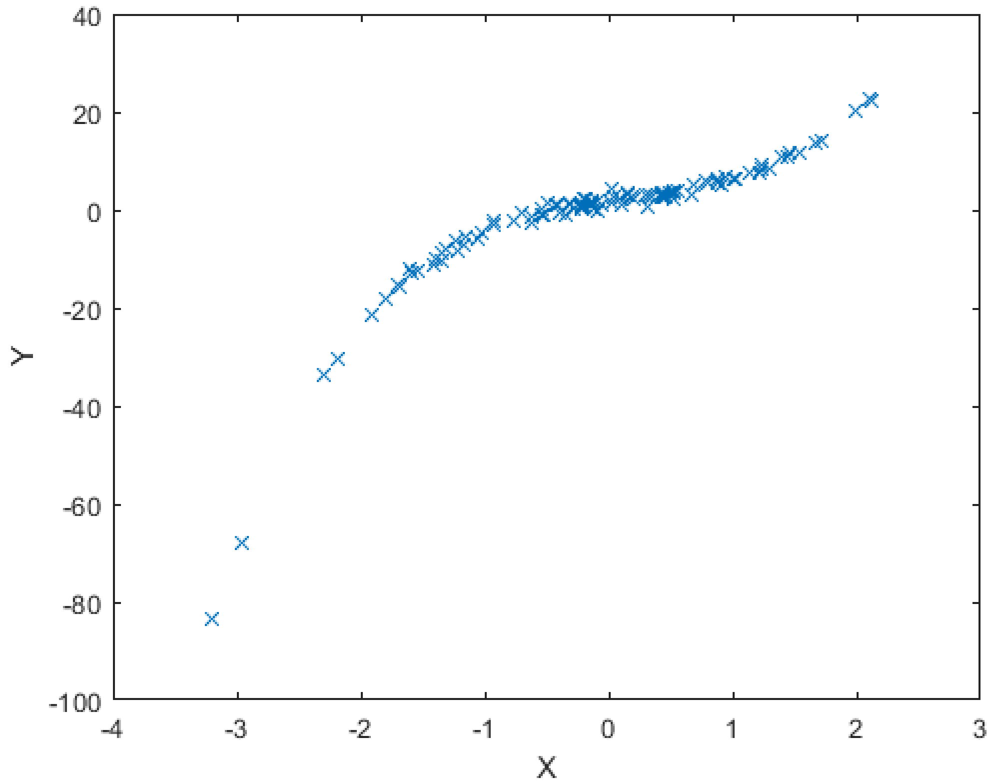
c)

```
Ytest = Y(101:end);
Ytrain = Y(1:100);
Xtest = X(101:end,:);
Xtrain = X(1:100,:);
```

d)

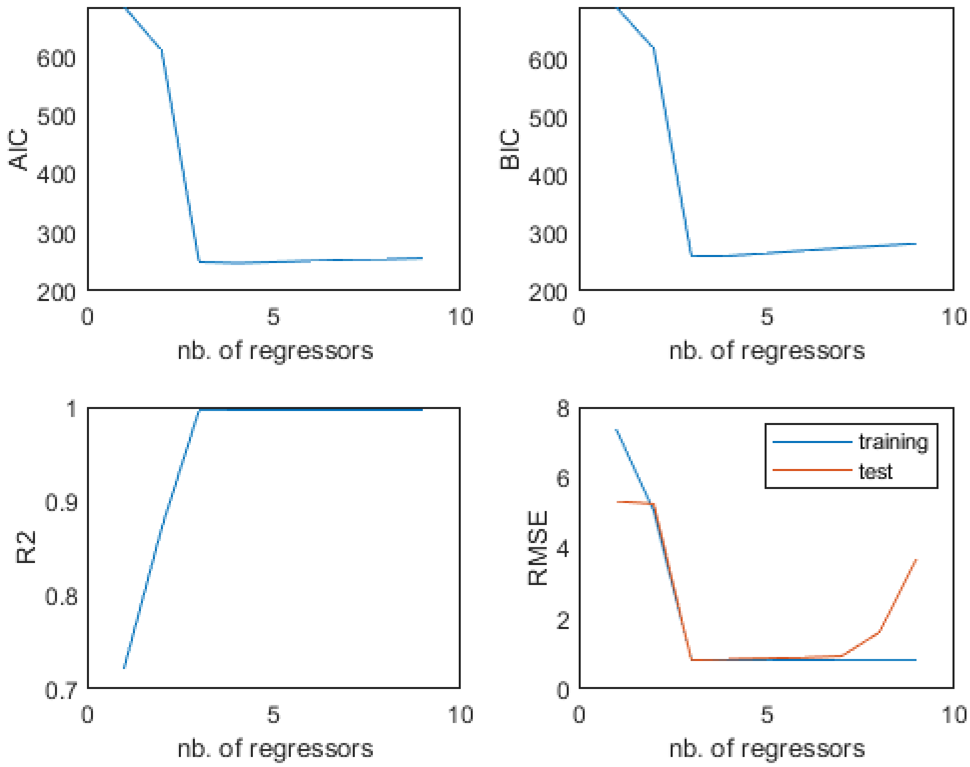
Use a for loop to generate a set of increasing nested models containing the predictors X, X^2,..., X^9. For each model, determine the AIC, BIC, and R2 and plot the results. What is the best model according to these criteria?

```
figure
plot(Xtrain,Ytrain,'x')
xlabel('X')
ylabel('Y')
hold all
for li = 1:9
    PhiTrain(:,li) = Xtrain.^li;
    PhiTest(:,li) = Xtest.^li;
    myModel = fitlm(PhiTrain,Ytrain);
    AIC(li) = myModel.ModelCriterion.AIC;
    BIC(li) = myModel.ModelCriterion.BIC;
    R2(li) = myModel.Rsquared.Ordinary;
    RMSE(li) = myModel.RMSE;

    YhatTrain = predict(myModel,PhiTrain);
    YhatTest = predict(myModel,PhiTest);

    RMSEtest(li) = sqrt(mean((Ytest-YhatTest).^2));

%     [~,idx] = sort(Xtrain);
%     plot(Xtrain(idx),YhatTrain(idx))

end
```



e)

```
figure
subplot 221
plot(AIC)
xlabel('nb. of regressors')
ylabel('AIC')
subplot 222
plot(BIC)
xlabel('nb. of regressors')
ylabel('BIC')
subplot 223
plot(R2)
xlabel('nb. of regressors')
ylabel('R2')
subplot 224
plot(RMSE)
hold on
plot(RMSEtest)
xlabel('nb. of regressors')
ylabel('RMSE')
legend('training','test')
```



f)

Now use the stepwiselm() function to perform a stepwise selection in order to choose the best model containing the predictors X,X2, . . .,X9. What is the best model obtained according to AIC, BIC, and R2?

```
myModel1 = stepwiselm(Xtrain,Ytrain,'poly9','Criterion','AIC')
myModel2 = stepwiselm(Xtrain,Ytrain,'poly9','Criterion','BIC')
myModel3 = stepwiselm(Xtrain,Ytrain,'poly9','Criterion','Rsquared')

% figure
% plot(Ytest,predict(myModel1,Xtest),'+')
% figure
% plot(Ytest,predict(myModel2,Xtest),'+')
% figure
% plot(Ytest,predict(myModel3,Xtest),'+')
```

```
1. Removing x1^9, AIC = 254.41
2. Removing x1^8, AIC = 253.48
3. Removing x1^7, AIC = 251.49
4. Removing x1^6, AIC = 249.54
5. Removing x1^5, AIC = 247.7

myModel1 =


Linear regression model:
```

```
    y ~ 1 + x1 + x1^2 + x1^3 + x1^4

Estimated Coefficients:
                   Estimate       SE        tStat       pValue  
                   ________    ________    _______    __________

    (Intercept)      1.7802     0.11463      15.53    8.1616e-28
    x1               3.4386     0.16323     21.066     1.425e-37
    x1^2           -0.73933     0.12515    -5.9077    5.3841e-08
    x1^3             1.8828    0.055678     33.815    9.0991e-55
    x1^4           -0.04669    0.023695    -1.9705      0.051691


Number of observations: 100, Error degrees of freedom: 95
Root Mean Squared Error: 0.815
R-squared: 0.997,  Adjusted R-Squared: 0.997
F-statistic vs. constant model: 7.2e+03, p-value = 5.5e-117
1. Removing x1^9, BIC = 277.85
2. Removing x1^8, BIC = 274.32
3. Removing x1^7, BIC = 269.73
4. Removing x1^6, BIC = 265.17
5. Removing x1^5, BIC = 260.73
6. Removing x1^4, BIC = 260.13

myModel2 = 


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

Estimated Coefficients:
                   Estimate       SE        tStat       pValue  
                   ________    ________    _______    __________

    (Intercept)      1.8734     0.10597     17.678    5.9831e-32
    x1               3.2357     0.12855      25.17     4.591e-44
    x1^2           -0.95107    0.065106    -14.608    3.9144e-26
    x1^3             1.9714     0.03327     59.256    2.0873e-77


Number of observations: 100, Error degrees of freedom: 96
Root Mean Squared Error: 0.827
R-squared: 0.997,  Adjusted R-Squared: 0.996
F-statistic vs. constant model: 9.32e+03, p-value = 3.51e-118
1. Removing x1^9, Rsquared = 0.99675
2. Removing x1^8, Rsquared = 0.99672
3. Removing x1^7, Rsquared = 0.99672
4. Removing x1^6, Rsquared = 0.99672
5. Removing x1^5, Rsquared = 0.99671
6. Removing x1^4, Rsquared = 0.99658

myModel3 = 
```

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

Estimated Coefficients:
                   Estimate        SE         tStat       pValue  
                   ________    ________    _______    __________

    (Intercept)      1.8734     0.10597     17.678    5.9831e-32
    x1               3.2357     0.12855      25.17     4.591e-44
    x1^2           -0.95107    0.065106    -14.608    3.9144e-26
    x1^3             1.9714     0.03327     59.256    2.0873e-77


Number of observations: 100, Error degrees of freedom: 96
Root Mean Squared Error: 0.827
R-squared: 0.997,  Adjusted R-Squared: 0.996
F-statistic vs. constant model: 9.32e+03, p-value = 3.51e-118
```

Published with MATLAB® R2021b