

Modelling and selection

Dominik Klepl

12/4/2019

Now that we kinda understand our data we can start modelling. We'll use two approaches to identify the best model structure.

1. Forward subset selection to identify the best model by minimizing MSE. We'll split the data into training and testing set using 80:20 ratio.
2. Use AIC OR BIC to identify the best model.

Hopefully both methods will converge on the same model structure.

Of course, we need to load the data first.

```
data = read.csv("data/x_y.csv", header = F)
colnames(data) = c("x", "y")
```

Forward subset selection

First, create matrix X with all possible predictors, i.e. intercept and x^1 to x^5

```
## Is X matrix?
## [1] TRUE
## Show first 5 rows of X
##      intercept      x1      x2      x3      x4
x5
## [1,]      1 -0.18533753 0.034350000 -0.0063663442 1.179923e-03 -
2.186839e-04
## [2,]      1 -1.06786788 1.140341806 -1.2177343852 1.300379e+00 -
1.388633e+00
## [3,]      1  0.10666430 0.011377272  0.0012135487 1.294423e-04
1.380687e-05
## [4,]      1  1.81829611 3.306200727  6.0116519064 1.093096e+01
1.987573e+01
## [5,]      1  0.07756119 0.006015738  0.0004665877 3.618910e-05
2.806869e-06
```

Now we split the data into training (80%) and testing sets

Forward selection works as follows: fit models using all columns of X separately (fit 6 models) calculate MSE of all models find model with min(MSE) append that model predictor to final model formula and remove that predictor from X

Fit models using selected predictor + all columns of X individually (again 6 models) repeat the other steps

Do this until model has 3 terms

Model fitting function

Because we'll need to fit multiple models, generate predictions and calculate MSE it might be good to have all of this wrapped in one nice function.

Forward selection

```
## Best model with 1 parameter: y ~ x4 + error (MSE= 5.232054 )  
  
##  
## Best model with 2 parameters: y ~ x4 + x2 + error (MSE= 0.3079839 )  
  
##  
## Best model with 3 parameters: y ~ x4 + x2 + x1 + error (MSE= 0.007951127 )
```

According to forward selection the best model is: $y \sim b_1x + b_2x^2 + b_3x^4$

Information criterion selection