# practical_5

*Alexander Maslev, Hanxiong Wang, Minyoung Lee*

*2018 /03 / 23*

## EXERCISE 2

Simulations: Selection by Hypothesis Testing

```
## Loading required package: MASS
## Loading required package: glmnet
## Loading required package: Matrix
## Loading required package: foreach
## Loaded glmnet 2.0-13
## Loading required package: intervals
##
## Attaching package: 'intervals'
## The following object is masked from 'package:Matrix':
##
##     expand
## Loading required package: survival
```

### Simulation settings

### 1) Generate MVN

we have created the matrix $X_{n*p}$ with n = 1000 and p = 10.

### 2) Creating beta

$\beta$ = [3 1.5 0 2 rep(0,6)]

### 3) $\hat{y} = X_{n*p}\beta + \epsilon$

### SELEVTIVE INFERENCE

We have done the stepwise regression and model selection with ForwardStop rule, AIC and BIC

### a)Stepwise regression with forwardStop

```
fsfit1<-fs(X,Y_hat)
out1<-fsInf(fsfit1)
forwardStop(out1$pv,alpha=0.05)
```

```
## [1] 3
```

We did forward stepwise regression by fs(), then inference for forward stepwise regression by fsInf() and we selected model based on forward stop rule.
We tried with 7 different values for type one error($\alpha$=[0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4])

```
forwardStop(out1$pv,alpha=0.1)
```

```
## [1] 3
```

```
forwardStop(out1$pv,alpha=0.15)
```

```
## [1] 3
```

```
forwardStop(out1$pv,alpha=0.2)
```

```
## [1] 3
```

```
forwardStop(out1$pv,alpha=0.25)
```

```
## [1] 3
```

```
forwardStop(out1$pv,alpha=0.3)
```

```
## [1] 3
```

```
forwardStop(out1$pv,alpha=0.35)
```

```
## [1] 3
```

```
forwardStop(out1$pv,alpha=0.4)
```

```
## [1] 5
```

Until $\alpha$=0.35, the model selection procedure stopped in 3rd step which includes variable 1, 4 and 2 in each step. But if we increase the type one error to 0.4 then we includes two more variables which are variable 10 and 9.

### b) Stepwise regression with AIC and BIC

```
fsfit2<-fs(X,Y_hat)
out2<-fsInf(fsfit2,type="aic",mult = 2,alpha=0.05)
out2
```

```
##
## Call:
## fsInf(obj = fsfit2, alpha = 0.05, type = "aic", mult = 2)
##
## Standard deviation of noise (specified or estimated) sigma = 0.997
##
## Testing results at step = 5, with alpha = 0.050
##  Var   Coef Z-score P-value LowConfPt UpConfPt LowTailArea UpTailArea
##    1  2.964  94.680   0.000     2.903    3.103       0.025      0.025
##    4  2.026  63.996   0.000     1.964    2.091       0.024      0.025
##    2  1.487  47.540   0.000     1.425    1.563       0.025      0.024
##   10 -0.075  -2.373   0.839    -0.099    1.711       0.025      0.025
##    9  0.073   2.306   0.024     0.001    1.863       0.025      0.025
##
## Estimated stopping point from AIC rule = 5
```

```
fsfit3<-fs(X,Y_hat)
out3<-fsInf(fsfit2,type="aic",mult = log(n),alpha=0.05)
out3
```

```
##
## Call:
## fsInf(obj = fsfit2, alpha = 0.05, type = "aic", mult = log(n))
##
## Standard deviation of noise (specified or estimated) sigma = 0.997
##
## Testing results at step = 3, with alpha = 0.050
##  Var  Coef Z-score P-value LowConfPt UpConfPt LowTailArea UpTailArea
##    1 2.966  94.889       0     2.904    3.027       0.024      0.025
##    4 2.020  63.917       0     1.958    2.083       0.024      0.025
##    2 1.486  47.535       0     1.424    1.548       0.024      0.024
##
## Estimated stopping point from AIC rule = 3
```

We did the forward stepwise regression by fs(), then inference for forward stepwise regression by fsInf() and we selected model based on AIC and BIC.
First result is the model selection based on AIC. By this rule, the forward stepwise regression stops at 5th step with variable 1, 4, 2, 10 and 9 included. This is same as forwardstop rule with $\alpha = 0.4$. Second result is the model selection based on BIC. By this rule, the forward stepwise regression stops at 3rd step same as forwardstop rule with $\alpha = 0.05$. In this case, we includes variable 1, 4 and 2 in the model.

**c)Time takes for model selection procedure**

```
## [1] 3
```

```
## 0.66 sec elapsed
```

```
## 0.66 sec elapsed
```

```
## 0.61 sec elapsed
```

Forward Stepwise model selection in our case has 55 combination of the models to check. And It takes approximately 2sec for this process. If we wanted to do the exhaustive search, we need 2^p-1 = 1023 models to check. Then it will probably takes 40sec to do the exhaustive search. With just 10 variables, It is 40 sec, but if we have high dimension, time for the model selection process will grow rapidly.

## EXERCISE 2

Real Data Application : Zambia data set

```
load("malnutrion_zambia_cleaned.Rda")
#str(data_zambia)

Y_zambia<-data_zambia[,1]
X_zambia<-as.matrix(data_zambia[,c(2:7,21,22,24)])


fitzambia<-fs(X_zambia,Y_zambia)
out_zambia<-fsInf(fitzambia)
out_zambia
```

```
##
## Call:
## fsInf(obj = fitzambia)
##
## Standard deviation of noise (specified or estimated) sigma = 1.592
##
## Sequential testing results with alpha = 0.100
##  Step Var   Coef Z-score P-value LowConfPt UpConfPt LowTailArea UpTailArea
##     1   1 -0.062 -12.229   0.000    -0.070   -0.053       0.049      0.050
##     2   5  5.209   8.937   0.000     3.812    6.179       0.049      0.049
##     3   8  0.283   4.901   0.400    -0.662    0.337       0.050      0.050
##     4   7  0.000   4.719   0.126     0.000    0.000       0.050      0.050
##     5   6  0.011   2.902   0.369    -0.051    0.074       0.050      0.050
##     6   3  0.015   2.380   0.182    -0.029    0.131       0.050      0.050
##     7   2 -0.005  -1.865   0.422    -0.029    0.026       0.050      0.050
##     8   9  0.003   1.616   0.107    -0.002    0.025       0.050      0.050
##     9   4 -0.081  -0.683   0.478    -0.451    0.529       0.050      0.050
##
## Estimated stopping point from ForwardStop rule = 2
```

```
forwardStop(out_zambia$pv,alpha=0.05)
```

```
## [1] 2
```

```
out_zambia_aic<-fsInf(fitzambia,type="aic",mult = 2,alpha=0.05)
out_zambia_aic
```

```
##
## Call:
## fsInf(obj = fitzambia, alpha = 0.05, type = "aic", mult = 2)
##
## Standard deviation of noise (specified or estimated) sigma = 1.592
##
## Testing results at step = 9, with alpha = 0.050
##  Var   Coef Z-score P-value LowConfPt UpConfPt LowTailArea UpTailArea
##    1 -0.053  -7.681   0.016    -0.117   -0.006       0.025      0.025
##    5  1.348   0.394   0.552   -31.046   22.176       0.025      0.025
##    8  0.285   4.904   0.418    -0.948    0.372       0.025      0.024
##    7  0.000   3.178   0.330     0.000    0.000       0.025      0.025
##    6  0.043   0.909   0.401    -0.252    0.284       0.025      0.025
##    3  0.012   1.804   0.266    -0.065    0.154       0.025      0.025
##    2 -0.006  -1.904   0.414    -0.035    0.033       0.025      0.025
##    9  0.003   1.636   0.483    -0.033    0.029       0.025      0.025
##    4 -0.081  -0.683   0.436    -0.696    0.669       0.025      0.025
##
## Estimated stopping point from AIC rule = 9
```

```
out_zambia_bic<-fsInf(fitzambia,type="aic",mult = log(n),alpha=0.05)
out_zambia_bic
```

```
##
## Call:
## fsInf(obj = fitzambia, alpha = 0.05, type = "aic", mult = log(n))
##
## Standard deviation of noise (specified or estimated) sigma = 1.592
##
```

```
## Testing results at step = 5, with alpha = 0.050
##  Var   Coef Z-score P-value LowConfPt UpConfPt LowTailArea UpTailArea
##    1 -0.061 -12.032   0.000    -0.072   -0.051       0.025      0.024
##    5  3.735   5.775   0.060    -1.630    8.835       0.000      0.025
##    8  0.285   4.917   0.417    -0.948    0.378       0.025      0.025
##    7  0.000   3.292   0.340     0.000    0.000       0.025      0.025
##    6  0.011   2.902   0.369    -0.067    0.089       0.025      0.025
##
## Estimated stopping point from AIC rule = 5
```

First output shows the result based on Forwardstop rule. The process stopped at 2nd step and variable 1 and 5 is included. Second output shows the result based on AIC rule. The process stopped at 9th step and variable 1, 5, 8, 7, 6, 3, 2, 9 and 4 is included in a sequence. Last output with BIC rule stopped at 5th step. Thus variable 1, 5, 8, 7 and 6 is included sequentially.