# Inlamning_2_umut

Umut Arslan

2023-02-26

## Simuleringsstudie

Jag kommer att genomföra en simuleringsstudie bestående av 100 observationer som ska loopas 500 gånger, det kommer alltså resultera i 500 träningsdatamängder. Vi kommer att plocka ut en MSE för varje loop och modell, förutom på FSS, Forward Stepwise Selection, där kommer vi att plocka ut modellens BIC.

MSE är ett mått på hur bra modellen är, en MSE på noll är en optimal modell. På engelska kallas MSE för Mean Squared Error.

Vi kommer att beräkna prediktions-MSE, teoretiska definitionen ser ut såhär: $V(\varepsilon) + E[(f(x) - \hat{f}(x))]^2$

där $V(\varepsilon)$ är det icke-reducerbara felet, $E[(f(x)]$ är det "faktiska" medelfelet och $E[\hat{f}(x)]$ är det predikterade medelfelet.

BIC står för Bayesian Information Criterion på engelska. Den teoretiska definitionen av BIC är: $1/(n) * (RSS + log(n)d\hat{\sigma}^2$. Vi kommer dock att använda oss utav ett paket, leaps, som har inbyggda funktioner för att beräkna BIC.

Syftet är alltså att hitta den modell med lägst MSE/BIC.

Datat består av en X-vektor, förklaringsvariabler, som innehåller 40 stycken förklaringsvariabler som är normalfördelade med väntevärde 0 och standardavvikelse 1. En Y-vektor, $rowSums(X[, 1 : 20]) + V(\varepsilon)$ som genererats från en linjär regressionsmodell, där lutningsparametern, beta, är 1 för de första 20 förklaringsvariablerna och 0 för de sista 20. Alltså är de 20 första förklaringsvariablerna vi vill ha i modellen medan de 20 sista är de icke-önskvärda.

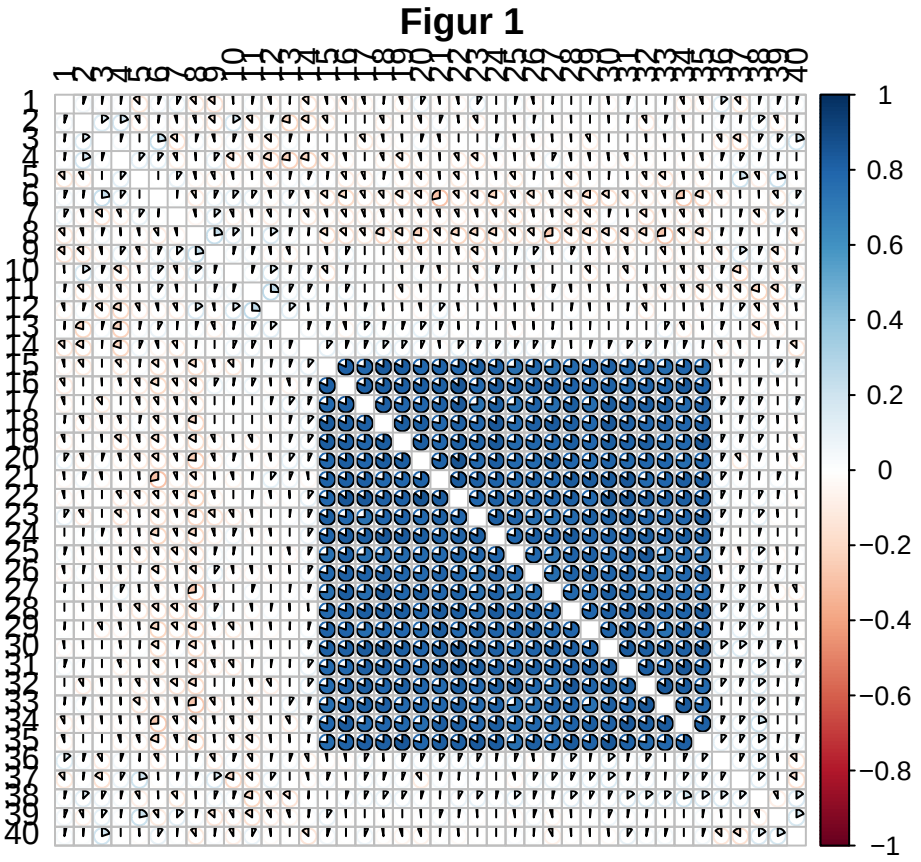
Värt att notera är att X-vektorn med de 40 förklaringsvariablerna inte är oberoende. Vi kommer senare se detta i en korrelationsplott.

## Modeller

Vi kommer att jämföra fem olika modeller

- OLS (En vanlig linjär regression med alla 40 förklaringsvariabler)
- Forward Stepwise Selection, FSS. Där den minsta BIC används som kriterium, det kan tilläggas att det även finns två andra metoder av samma modell som kallas för Best Subset selection och Backward Stepwise Selection.
- LASSO, där tuning-parametern $\lambda$, väljs genom korsvalidering på respektive tärningsdatamängd
- PCR, där antalet principalkomponenter, $M$, väljs genom korsvalidering på respektive träningsdatamängd
- PCR med två principalkomponenter, d.v.s $M = 2$

### Korrelationsplot




Figur 1 visar att det finns en hög korrelation mellan ett antal förklaringsvariabler, det finns en hög korrelation mellan 19 förklaringsvariabler. Med andra ord så har dessa förklaringsvariabler ett högt beroende mellan varandra.

## OLS

OLS är en grundläggande regression som försöker hitta en linje/kurva som bäst anpassar datat man har. OLS antar att alla förklaringsvariabler har ett linjärt samband samt en konstant varians.

Fördelar: Lätt att tolka, om antaganden är ok så finns det ingen bias i modellen, man kan kolla R-squared.

Nackdelar: Antar att datat har en konstant varians, antar linjärt samband, känsligt mot outliers.

## FSS

Forward Stepwise Selection, FSS, är en metod som väljer delmängder av förklaringsvariabler som inkluderas i modellen. Man börjar med 0 variabler och adderar en förklaringsvariabel i taget och modellen slutar addera fler förklaringsvariabler när den sista variabeln inte bidrar mer till modellen

Fördelar: Kan vara lättare att tolka resultatet eftersom man förhoppningsvis har reducerat antalet förklaringsvariabler.

Här kan man läsa en del problem med denna metod: https://www.stata.com/support/faqs/statistics/stepwise-regression-problems/

## LASSO

Lasso är en metod som likt FSS väljer delmängder av förklaringsvariabler där man på något sätt krymper koefficienter mot noll genom att straffa dessa. Man tvingar bort vissa koefficienter som inte bidrar så värst mycket till modellen.

Lasso står för Least Absolute Shrinkage and Selection Operator

Fördelar: Kan krympa och välja "bästa" förklaringsvariabler samtidigt.

Nackdelar: kan ta med många korrelerade variabler.

## PCR

Man minskar förklaringsvariablernas dimensioner genom att skapa nya variabler, principal components. Dessa nya variabler är linjära kombinationer av de gamla förklaringsvariablerna. Man använder de "principal components" som förklarar det mesta av variansen i datat.

Fördelar: Vid hög korrelation kan den minska bruset, hantera hög-dimensionell data med många förklaringsvariabler.

Nackdelar: Svår att tolka, lätt att överanpassa datat.

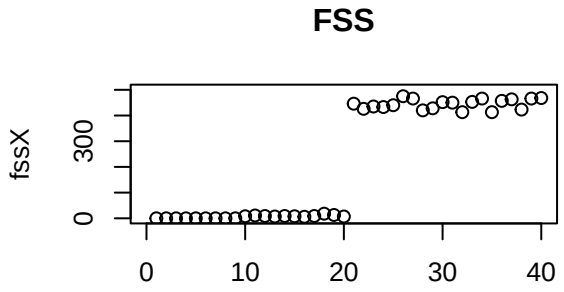
## Prediktions-MSE

```
##                                MSE
## OLS                       1.407472
## Forward Stepwise Selection  1.301227
## LASSO                      1.322857
## PCR, M                     3.520570
## PCR, M = 2                10.047218
```

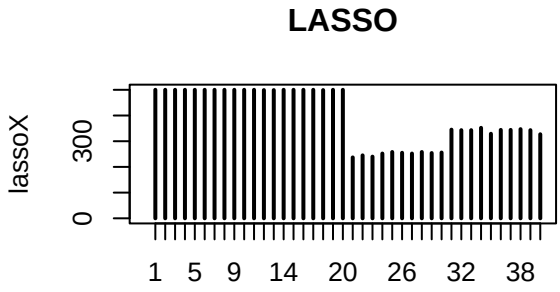
Ovan ser vi att FSS har den lägsta prediktions-MSE, vilket innebär att FSS är den bästa modellen att använda för detta data, om man endast kollar på prediktions-MSE.

## Jämför FSS och LASSO

Vi kommer nu att jämföra hur bra FSS och LASSO är på variabelselektion, alltså hur ofta modellerna har kvar de första 20 förklaringsvariablerna som ska vara med. Men även hur ofta modellerna tar med de sista 20 förklaringsvariablerna som ej ska vara med.




Figur 2




Figur 3

Figur 3 visar antalet förklaringsvariabler som togs med på dom 500 träningsdatamängderna medan FSS Figur 2 visar det motsatta, alltså hur många gånger FSS inte tog med de 40 förklaringsvariablerna. Y-led visar antalet loopar och X-led visar vilken X-variabel som har varit med i loopen på LASSO och dom som inte har varit med i FSS.

Vi ser en tydlig skillnad här, FSS tar med färre av de icke-önskade förklaringsvariablerna men den missar även att ta med de förklaringsvariabler som SKA vara med i modellen, vilket LASSO inte gör, den tar alltid med de 20 första förklaringsvariablerna, med nackdel att den även tar med de sista 20 förklaringsvariablerna mycket oftare.

Det kan vara lite svårt att se antalet X-variabler som inte blir vald på Figur 2.

FSS: X1-9 = 0 som inte blir vald X10 = 13 som inte blir vald X11 = 6 som inte blir vald X12 = 8 som inte blir vald X13 = 8 som inte blir vald X14 = 12 som inte blir vald X15 = 10 som inte blir vald X16 = 8 som inte blir vald X17 = 9 som inte blir vald X18 = 8 som inte blir vald X19 = 11 som inte blir vald X20 = 11 som inte blir vald

## Ändra förutsättningar?

Exempelvis skulle man kunna ändra antalet träningsdatamängder, minskar man antalet så kan man få bredare konfidensintervall och mest troligt högre sampling error. Ökar vi antalet träningsdata så får vi en mer "pricksäker" skattning av MSE.

Ändring av det icke-reducerbara felet, $V(\varepsilon)$, ändring av parametrarnas värde, reduktion av beroende, bias.

Det hade väl varit mest intressant att kanske ändra på beroendet, minska, för att se hur bra variabelselektion modellerna skulle ha då.