Lab 10: Double LASSO

Monte Carlo simulations

```
library(hdm)
?rlassoEffect
n=100
R=300
rho=1
beta1=0
beta2=0.35
```

Write a function for generating data:

```
data_sim<-function(n,beta1,beta2,rho){
  X=matrix(rnorm(n*3),ncol=3)
  X[,2]<-rho*X[,1]+X[,2]
  Y=beta1*X[,1]+beta2*X[,2]+rnorm(n)
  data<-list(Y=Y,X=X)
}
```

Generate data on the main regressor (D), potential controls, and the dependent variable:

```
set.seed(5,sample.kind = "Rejection")
data<-data_sim(n,beta1,beta2,rho)
y=data$Y #dep. variable
Controls=data$X[,-1] # controls
D=data$X[,1] # the main regressor for which the effect is estimated
```

Run double LASSO:

```
Effect<-rlassoEffect(Controls,y,D,method="double selection")
summary(Effect)
```

```
## [1] "Estimates and significance testing of the effect of target variables"
##    Estimate. Std. Error t value Pr(>|t|)
## d1   -0.1102     0.1663  -0.663    0.508
```

Objects inside:

```
names(Effect)
```

```
##  [1] "alpha"            "se"               "t"                "pval"
##  [5] "no.selected"      "coefficients"     "coefficient"      "coefficients.reg"
##  [9] "selection.index"  "residuals"        "call"             "samplesize"
```

Included controls and t-statistic on D:

```
Effect$selection.index
```

```
##    x1    x2
##  TRUE FALSE
```

```
Effect$t
```

```
##         d1
## -0.6627201
```

We run the simulations using the setup from lab 9.

```
rho=1
set.seed(6064,sample.kind = "Rejection")
T_Beta1_post=rep(0,R) # Vector to store t-stats for the main regressor
for (r in 1:R){
  data<-data_sim(n,beta1,beta2,rho)
  Effect<-rlassoEffect(data$X[,-1],data$Y,data$X[,1],method="double selection")
  T_Beta1_post[r]=Effect$t
}
```

Plot of the distribution of the post-double-Lasso t-statistic:

```
low=min(T_Beta1_post)
high=max(T_Beta1_post)
step=(high-low)/20
hist(T_Beta1_post,breaks=seq(low-2*step,high+2*step,step),xlab="estimates",main="The exact distribution

# add a vertical line at the true value
abline(v=beta1,col="blue")

# add the plot of the N(0,1) pdf
x=seq(-4,4,0.01)
f=exp(-x^2/2)/sqrt(2*pi)
lines(x,f,col="red")
```

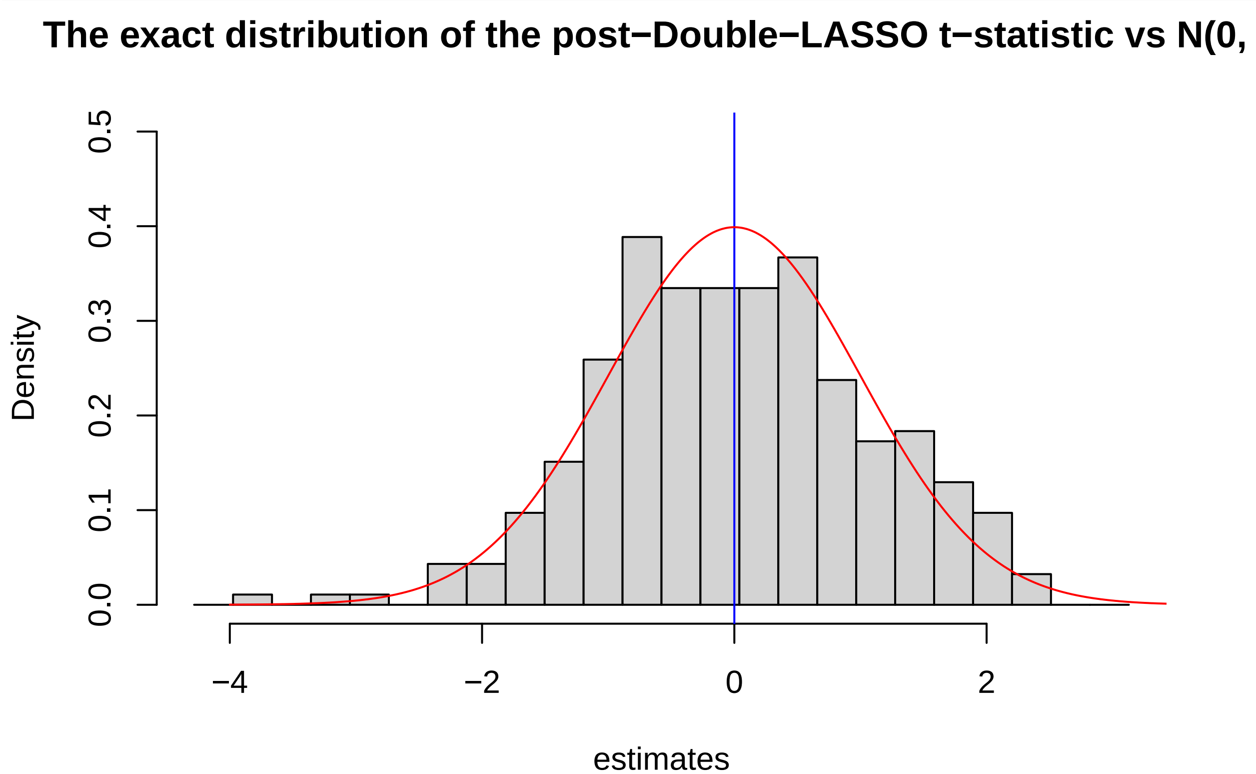

Illustration of double LASSO with cross country growth data

The model is $\Delta \log(GDP_{it}) = \alpha \cdot GDP_{i0} + U_i$. Hypothesis: $\alpha < 0$. Less developed countries catch up with more developed.

```
data("GrowthData")
?GrowthData
names(GrowthData)
```

```
##  [1] "Outcome"   "intercept" "gdpsh465"  "bmp1l"     "freeop"    "freetar"  
##  [7] "h65"       "hm65"      "hf65"      "p65"       "pm65"      "pf65"     
## [13] "s65"       "sm65"      "sf65"      "fert65"    "mort65"    "lifee065" 
## [19] "gpop1"     "fert1"     "mort1"     "invsh41"   "geetot1"   "geerec1"  
## [25] "gde1"      "govwb1"    "govsh41"   "gvxdxe41"  "high65"    "highm65"  
## [31] "highf65"   "highc65"   "highcm65"  "highcf65"  "human65"   "humanm65" 
## [37] "humanf65"  "hyr65"     "hyrm65"    "hyrf65"    "no65"      "nom65"    
## [43] "nof65"     "pinstab1"  "pop65"     "worker65"  "pop1565"   "pop6565"  
## [49] "sec65"     "secm65"    "secf65"    "secc65"    "seccm65"   "seccf65"  
## [55] "syr65"     "syrm65"    "syrf65"    "teapri65"  "teasec65"  "ex1"      
## [61] "im1"       "xr65"      "tot1"
```

The hypothesis fails:

```
summary(lm(Outcome~gdpsh465,data=GrowthData))
```

```
## 
## Call:
## lm(formula = Outcome ~ gdpsh465, data = GrowthData)
## 
## Residuals:
##       Min        1Q    Median        3Q       Max 
## -0.147387 -0.024088  0.001209  0.027721  0.139357 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.035207   0.047318   0.744    0.459
## gdpsh465    0.001317   0.006102   0.216    0.830
## 
## Residual standard error: 0.05159 on 88 degrees of freedom
## Multiple R-squared:  0.0005288,  Adjusted R-squared:  -0.01083 
## F-statistic: 0.04656 on 1 and 88 DF,  p-value: 0.8297
```

An alternative model controls for the institutional and technological characteristics: $\Delta \log(GDP_{it}) = \alpha \cdot GDP_{i0} + X_i'\beta + U_i$.

There are a lot of potential controls:

```
dim(GrowthData)
```

```
## [1] 90 63
```

Let's set up estimation

```
names(GrowthData)
```

```
##  [1] "Outcome"   "intercept" "gdpsh465"  "bmp1l"     "freeop"    "freetar"  
##  [7] "h65"       "hm65"      "hf65"      "p65"       "pm65"      "pf65"     
## [13] "s65"       "sm65"      "sf65"      "fert65"    "mort65"    "lifee065" 
## [19] "gpop1"     "fert1"     "mort1"     "invsh41"   "geetot1"   "geerec1"  
## [25] "gde1"      "govwb1"    "govsh41"   "gvxdxe41"  "high65"    "highm65"  
```

```
## [31] "highf65"   "highc65"   "highcm65"  "highcf65"  "human65"   "humanm65"
## [37] "humanf65"  "hyr65"     "hyrm65"    "hyrf65"    "no65"      "nom65"
## [43] "nof65"     "pinstab1"  "pop65"     "worker65"  "pop1565"   "pop6565"
## [49] "sec65"     "secm65"    "secf65"    "secc65"    "seccm65"   "seccf65"
## [55] "syr65"     "syrm65"    "syrf65"    "teapri65"  "teasec65"  "ex1"
## [61] "im1"       "xr65"      "tot1"
```

```
y=as.vector(GrowthData$Outcome)
D=as.vector(GrowthData$gdpsh465)
Controls=as.matrix(GrowthData)[,-c(1,2,3)]
```

We run OLS with all controls. The estimate is negative but the standard error is too large, since there are too many controls.

```
Full=lm(y~D+Controls)
head(coef(summary(Full)),2)
```

```
##                 Estimate Std. Error    t value  Pr(>|t|)
## (Intercept)  0.247160893 0.78450163  0.3150547 0.7550562
## D           -0.009377989 0.02988773 -0.3137739 0.7560185
```

Post-LASSO with Double LASSO

```
Effect<-rlassoEffect(Controls,y,D,method="double selection")
summary(Effect)
```

```
## [1] "Estimates and significance testing of the effect of target variables"
##    Estimate. Std. Error t value Pr(>|t|)
## d1  -0.05001    0.01579  -3.167  0.00154 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Included controls:

```
Effect$selection.index
```

```
##    bmp1l   freeop  freetar      h65     hm65     hf65      p65     pm65
##     TRUE    FALSE     TRUE    FALSE     TRUE    FALSE    FALSE    FALSE
##     pf65      s65     sm65     sf65   fert65   mort65  lifee065    gpop1
##    FALSE    FALSE    FALSE     TRUE    FALSE    FALSE     TRUE    FALSE
##    fert1    mort1  invsh41  geetot1  geerec1     gde1   govwb1  govsh41
##    FALSE    FALSE    FALSE    FALSE    FALSE    FALSE    FALSE    FALSE
## gvxdxe41   high65  highm65  highf65  highc65 highcm65 highcf65  human65
##    FALSE    FALSE    FALSE    FALSE    FALSE    FALSE    FALSE    FALSE
## humanm65 humanf65    hyr65   hyrm65   hyrf65     no65    nom65    nof65
##    FALSE     TRUE    FALSE    FALSE    FALSE    FALSE    FALSE    FALSE
## pinstab1    pop65 worker65  pop1565  pop6565    sec65   secm65   secf65
##    FALSE    FALSE    FALSE    FALSE     TRUE    FALSE    FALSE    FALSE
##   secc65  seccm65  seccf65    syr65   syrm65   syrf65 teapri65 teasec65
##    FALSE    FALSE    FALSE    FALSE    FALSE    FALSE    FALSE    FALSE
##      ex1      im1     xr65     tot1
##    FALSE    FALSE    FALSE    FALSE
```

```
sum(Effect$selection.index==TRUE)
```

```
## [1] 7
```

The partialling out approach

```
Effect_PO<-rlassoEffect(Controls,y,D,method="partialling out")
summary(Effect_PO)
```

```
## [1] "Estimates and significance testing of the effect of target variables"
##      Estimate. Std. Error t value Pr(>|t|)
## [1,]  -0.04981    0.01394  -3.574 0.000351 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Effect_PO$selection.index
```

```
##    bmp1l   freeop  freetar      h65     hm65     hf65      p65     pm65
##     TRUE    FALSE     TRUE    FALSE     TRUE    FALSE    FALSE    FALSE
##     pf65      s65     sm65     sf65   fert65   mort65 lifee065    gpop1
##    FALSE    FALSE    FALSE     TRUE    FALSE    FALSE     TRUE    FALSE
##    fert1    mort1  invsh41  geetot1  geerec1     gde1   govwb1  govsh41
##    FALSE    FALSE    FALSE    FALSE    FALSE    FALSE    FALSE    FALSE
## gvxdxe41   high65  highm65  highf65  highc65 highcm65 highcf65  human65
##    FALSE    FALSE    FALSE    FALSE    FALSE    FALSE    FALSE    FALSE
## humanm65 humanf65    hyr65   hyrm65   hyrf65     no65    nom65    nof65
##    FALSE     TRUE    FALSE    FALSE    FALSE    FALSE    FALSE    FALSE
## pinstab1    pop65 worker65  pop1565  pop6565    sec65   secm65   secf65
##    FALSE    FALSE    FALSE    FALSE     TRUE    FALSE    FALSE    FALSE
##   secc65  seccm65  seccf65    syr65   syrm65   syrf65 teapri65 teasec65
##    FALSE    FALSE    FALSE    FALSE    FALSE    FALSE    FALSE    FALSE
##      ex1      im1     xr65     tot1
##    FALSE    FALSE    FALSE    FALSE
```

```
sum(Effect_PO$selection.index==TRUE)
```

```
## [1] 7
```