RegCombin: Linear Regression under Data Combination

Note

This package is still a beta-version, please do not hesitate to email christophe.gaillac@economics.ox.ac.uk if you have any remark, question, or suggestion.

Introduction

The **RegCombin** package implements linear regression when the outcome of interest and some of the covariates are observed in two different datasets that cannot be linked, based on D'Haultfoeuille, Gaillac, and Maurel (2022), DGM hereafter.

The package allows for common regressors observed in both datasets, and for sign constraints on the effect of covariates on the outcome of interest.

We consider the following linear model:

$$E(Y|X) = \sum_{k=1}^{K} \gamma_{0,k} 1\{X_c = x_{c,k}\} + X_{nc}' \beta_0, \quad X = (X_{nc}, X_c),$$

where the support of X_c is $Supp(X_c) = \{x_{c,1}, ..., x_{c,K}\}$, and in a data combination environment where the distributions F_{Y,X_c} and F_{X_{nc},X_c} are supposed to be identified, but the joint distribution $F_{Y,X}$ is not. The variables X_c are thus common to both datasets, whereas the variables X_{nc} are only observed in one of the two datasets. In this setup, $\beta_0 = (\beta_{01}, ..., \beta_{0p})'$ is generally not point-identified, and as a result we focus on the identified set of either β_0 or β_{0k} for some $k \in \{1, ..., p\}$. Using that for all $x_{c,k} \in Supp(X_c)$,

$$E(Y|X_c = x_{c,k}) = \gamma_{0,k} + E(X_{nc}|X_c = x_{c,k})' \beta_0,$$

the identified set for each component $\gamma_{0,k}$ can be deduced from the one of β_0. A corollary is that imposing (weak) sign constraints on some of the $\gamma_{0,k}$ can shrink the identified set of β_0.

The package will soon be available on CRAN but for now, one could install it through Github and load it through:

```
# install.packages("devtools")
devtools::install_github("cgaillac/RegCombin")
library(RegCombin)
```

We detail below the different cases of interest described in Section 4 in DGM according to the dimension of X_{nc} and if there are common regressors X_c entering the regression. In the package **RegCombin**, the main function is **regCombin** and gathers all the different cases. Note that the minimal requirement for the function **regCombin** are the two datasets: "Ldata" containing the outcome Y as well as the variables X_c observed jointly, and "Rdata" containing the regressor X_{nc} as well as the variables X_c observed jointly, as well as the label (the "out_var" argument) of Y in the first dataset and the labels of the variables X_{nc} ("nc_var" argument) in the second dataset. If common regressors X_c are observed, their labels should be passed in the "c_var" argument.

You can also access the documentation of the main function using

```
## to obtain help on the main function
?regCombin
```

Univariate case without common regressors

We consider the example of the Normal/Normal case of Section 4.1 in DGM. Namely, we consider a sample of i.i.d. observations drawn from the model

$$Y = X_{nc}\beta_0 + U, \quad \beta_0 = 1, \;\; U|X_{nc} \sim \mathcal{N}(0,1).$$

Then, we assume that $X_{nc} \sim \mathcal{N}(0, 1.5)$ and $U \sim \mathcal{N}(0,1)$.

```
library(R.matlab)
library(pracma)
library(sfsmisc)
library(Hmisc)
library(RegCombin)
library(MASS)
library(snowfall)
```

```
### Simulating according to this DGP
n=800
Xnc_x =  rnorm(n,0,1.5)
Xnc_y =  rnorm(n,0,1.5)
epsilon =  rnorm(n,0,1)
## true value
beta0 =1
Y =  Xnc_y*beta0 + epsilon

out_var = "Y"
nc_var = "Xnc"

# create the datasets
Ldata<- as.data.frame(Y)
colnames(Ldata) <- c(out_var)
Rdata <- as.data.frame(Xnc_x)
colnames(Rdata) <- c(nc_var)

start <- Sys.time()
out <- regCombin(Ldata,Rdata,out_var,nc_var)
end <- Sys.time()

#### Time used
end-start
```

```
## Time difference of 1.670311 secs
```

```
## the confidence interval at 95% on beta_0
out$DGMCI_sign
```

```
##          [,1]     [,2]
## [1,] -1.30793 1.308749
```

```
## the point estimate for the DGM bounds
out$DGMpt_sign
```

```
##           [,1]     [,2]
## [1,] -1.216696 1.221357
```

```
## the value of the point estimates of the radial function in the directions -1 and 1
out$DGM_complete$point$upper
```

```
##           [,1]     [,2]
## [1,] 1.216696 1.221357
```

Multivariate case without common regressor

We consider the example of Section 4.2 in DGM, i.e. a multivariate case ($p = 2$) with

$$Y = \gamma_0 + X'_{nc}\beta_0 + U,\ U|X \sim \mathcal{N}(0, 4).$$

We set the coefficients as follows: $\gamma_0 = -0.1$, $\beta_{0,1} = 1$, and $\beta_{0,2} = 1$. The variables X_{nc} follow a multivariate normal distribution with mean 0 and covariance matrix

$$\Sigma = \begin{pmatrix} 1 & -0.2 \\ -0.2 & 1 \end{pmatrix}.$$

```
### Simulating according to this DGP
n=800
norm = mvrnorm(n,mu=c(0,0,0),Sigma=matrix(c(1,0.8,-0.1,0.8,1,-0.2,-0.1,-0.2,1),3,3))
norm_y = mvrnorm(n,mu=c(0,0,0),Sigma=matrix(c(1,0.8,-0.1,0.8,1,-0.2,-0.1,-0.2,1),3,3))

## true value of beta0
beta0 = as.matrix(c(1,1));

Xnc_x=norm[,2:3];
Xnc_y=norm_y[,2:3]
Y =  Xnc_y%*%beta0 + 2*rnorm(n)

## creating the separate datasets
out_var = "Y"
nc_var = c("Xnc1","Xnc2")
Ldata <- as.data.frame(Y)
colnames(Ldata) <- c(out_var)
Rdata <- as.data.frame(Xnc_x)
colnames(Rdata) <- c(nc_var)
```

```
##### compute the bounds using 4 CPUs
start <- Sys.time()
out <- regCombin(Ldata,Rdata,out_var,nc_var,nbCores=4,projections= TRUE)
```

```
## R Version:  R version 4.1.0 (2021-05-18)
##
## Library R.matlab loaded.
## Library pracma loaded.
## Library Hmisc loaded.
```

```
end <- Sys.time()

#### Time used
end-start
```

```
## Time difference of 25.08313 secs
```

```
## the confidence interval at 95% on the components of beta_0
out$DGMCI
```

```
##           [,1]     [,2]
## [1,] -2.531793 2.533086
## [2,] -2.408513 2.461137
```

```
## the point estimate for the DGM bounds on the components of beta_0
out$DGMpt
```

```
##           [,1]     [,2]
## [1,] -2.393585 2.398571
## [2,] -2.274924 2.317879
```

```
## the value of the point estimates of the support function in the direction q
out$DGMsupport_pt
```

```
##      q_1 q_2  Support
## [1,]  -1   0 2.393585
## [2,]   0  -1 2.274924
## [3,]   1   0 2.398571
## [4,]   0   1 2.317879
```

Case with common regressor

Finally, we consider the example of Section 4.3 in DGM. Namely, we consider the data-generating process

$$Y = \gamma_0 1\{X_c = 1\} + X_{nc}\beta_0 + U,\ U|X \sim \mathcal{N}(0, 4).$$

Where we set the coefficients as follows: $\gamma_0 = 0.3$ and $\beta_0 = 1$. The covariates are transformations of $(N_1, N_2)'$, which follows a multivariate normal distribution with mean 0 and covariance matrix

$$\Sigma = \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1.5 \end{pmatrix}.$$

Specifically, the common regressor is given by $X_c = 1\{N_1 \leq 0.3\}$, and the regressors that are observed in one of the datasets only are such that $X_{nc} = N_2$.

```
### Simulating according to this DGP
n=800
norm = mvrnorm(n,mu=c(0,0),Sigma= matrix(c(1,0.8,0.8,1.5),2,2))
normy = mvrnorm(n,mu=c(0,0),Sigma= matrix(c(1,0.8,0.8,1.5),2,2))
Xc_x= 1*(norm[,1]<= 0.3)
Xc_y= 1*(normy[,1]<= 0.3)
Xnc_x=norm[,2];
Xnc_y=normy[,2];
epsilon = 2*rnorm(n)

beta0 = 1
Y = 0.3*Xc_y+   Xnc_y*beta0 + epsilon

### creating the datasets
out_var = "Y"
nc_var = "Xnc"
c_var =  "Xc"
Ldata<- as.data.frame(cbind(Xc_y,Y))
colnames(Ldata) <- c(c_var,out_var)
Rdata <- as.data.frame(cbind(Xc_x,Xnc_x))
colnames(Rdata) <- c(c_var,nc_var)

start <- Sys.time()
### computation
out <- regCombin(Ldata,Rdata,out_var,nc_var,c_var)
end <- Sys.time()
```

```
#### Time used
end-start
```

```
## Time difference of 3.935258 secs
```

```
## the confidence interval at 95% on beta_0
out$DGMCI_sign
```

```
##           [,1]     [,2]
## [1,] -2.549514 2.550623
```

```
## the point estimate for the DGM bounds
out$DGMpt_sign
```

```
##           [,1]     [,2]
## [1,] -2.284187 2.288953
```

We can also ad the sign constraint $\gamma_0 \geq 0$ which can help reducing the confidence intervals on β_0.

```
c_sign=c(1)
nc_sign=c(0)
### computation
out <- regCombin(Ldata,Rdata,out_var,nc_var,c_var, nc_sign,c_sign)

## the confidence interval at 95% on beta_0 with constraints
out$DGMCI_sign
```

```
##           [,1]     [,2]
## [1,] 0.3887466 2.558373
```

```
## the point estimate for the DGM bounds with constraints
out$DGMpt_sign
```

```
##           [,1]     [,2]
## [1,] 0.6485865 2.288953
```

References

- Xavier D'Haultfoeuille, Christophe Gaillac and Arnaud Maurel. "Partially Linear Models under Data Combination" working paper, (2022)