Nonparametric Project of Agricultural Productivity in the U.S.

GAM

Sofia Moroni*

2023-06-27

Contents

1 Load libraries and data

```
library(pbapply)
```

```
## Warning: il pacchetto 'pbapply' è stato creato con R versione 4.1.3
```

```
library(mgcv)
library(conformalInference)
library(ggplot2)
```

```
## Warning: il pacchetto 'ggplot2' è stato creato con R versione 4.1.3
```

```
library(progress)
library(parallel)
```

*sofia.moroni@mail.polimi.it

```r
data_path = file.path('data')
output_path = file.path('results')
data =
    read.table(
        file.path(data_path, 'agricultural_indices.csv'),
        header = T,
        sep = ';'
    )

# Sostituzione delle virgole con punti
data<- data.frame(lapply(data, function(x) gsub(",", ".", x)))
data <- as.data.frame(lapply(data, as.numeric))

data_test = data[69:72,]

set.seed(100)
B = 1000
n = nrow(data)
```

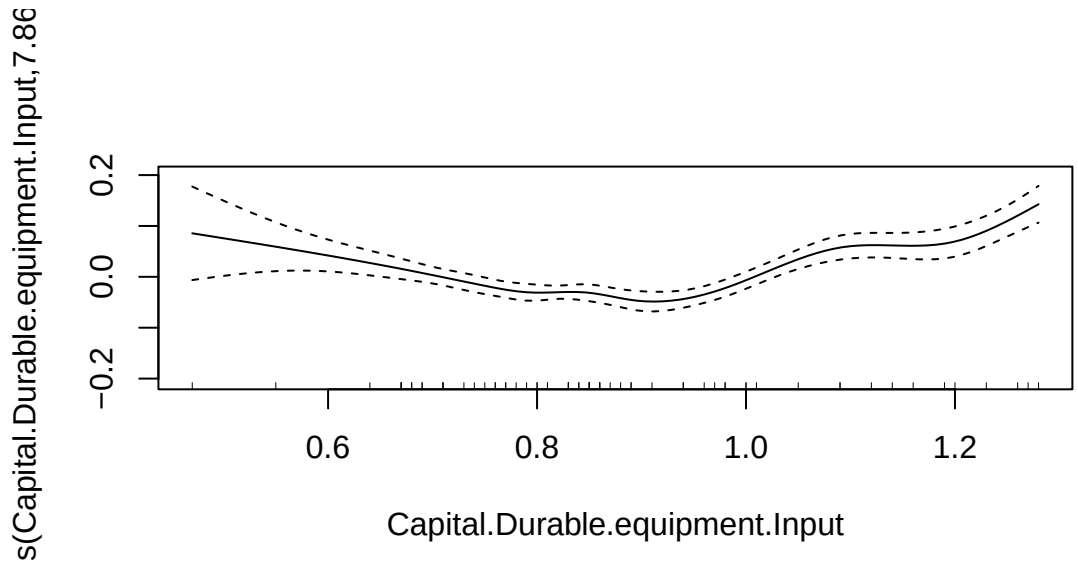
2 MODEL

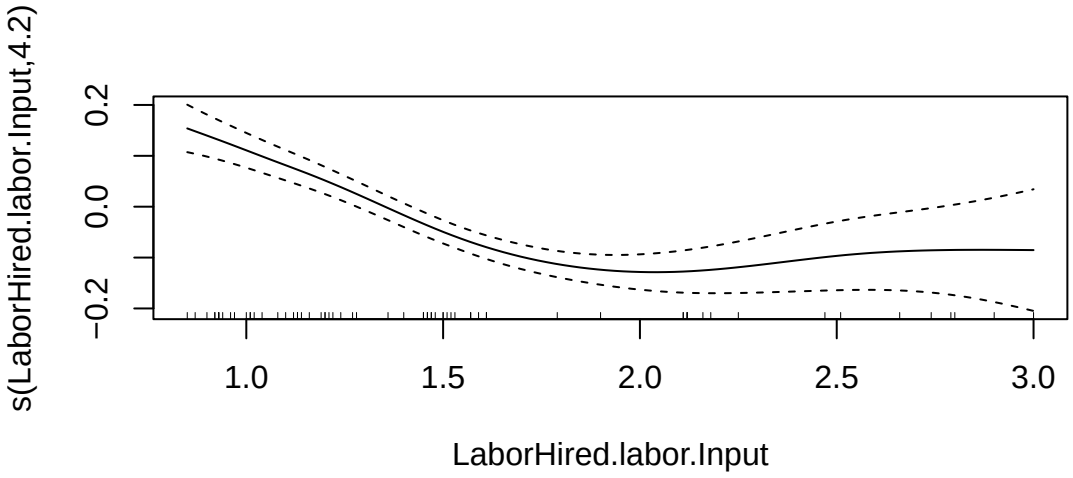
```r
data = data[1:68,]

model_gam = gam(Total.agricultural.output ~s(Capital.Durable.equipment.Input, bs = 'cr')
    + Capital.Service.buildings.Input
    + Labor.Self.employed.and.unpaid.family.Input
    + s(LaborHired.labor.Input, bs = 'cr')
    + s(Capital.Inventories.Input)
    + Intermediate.Energy.Input
    + Intermediate.Pesticides.Input,
    data = data
   )

summary(model_gam)
```

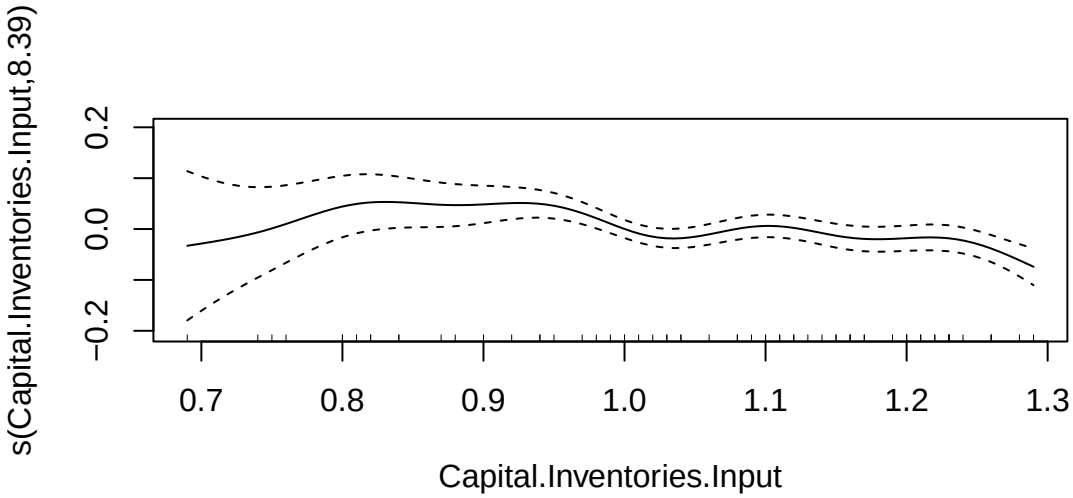
```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## Total.agricultural.output ~ s(Capital.Durable.equipment.Input,
##     bs = "cr") + Capital.Service.buildings.Input + Labor.Self.employed.and.unpaid.family.Input +
##     s(LaborHired.labor.Input, bs = "cr") + s(Capital.Inventories.Input) +
##     Intermediate.Energy.Input + Intermediate.Pesticides.Input
##
## Parametric coefficients:
##                                           Estimate Std. Error t value
## (Intercept)                                1.09040    0.07225  15.093
## Capital.Service.buildings.Input           -0.27896    0.04054  -6.881
## Labor.Self.employed.and.unpaid.family.Input -0.11459    0.02237  -5.123
```

```
## Intermediate.Energy.Input                    0.06787    0.02199   3.087
## Intermediate.Pesticides.Input                0.13931    0.02705   5.151
##                                             Pr(>|t|)
## (Intercept)                                  < 2e-16 ***
## Capital.Service.buildings.Input             2.01e-08 ***
## Labor.Self.employed.and.unpaid.family.Input 6.91e-06 ***
## Intermediate.Energy.Input                    0.00355 **
## Intermediate.Pesticides.Input               6.33e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                                     edf Ref.df     F p-value
## s(Capital.Durable.equipment.Input) 7.859  8.533 12.37 < 2e-16 ***
## s(LaborHired.labor.Input)          4.198  5.117 16.33 < 2e-16 ***
## s(Capital.Inventories.Input)       8.394  8.857  3.93 0.00119 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.994   Deviance explained = 99.6%
## GCV = 0.00041846  Scale est. = 0.00026184  n = 68
```

```
plot(model_gam)
```







3 MODEL WITH INTERACTION

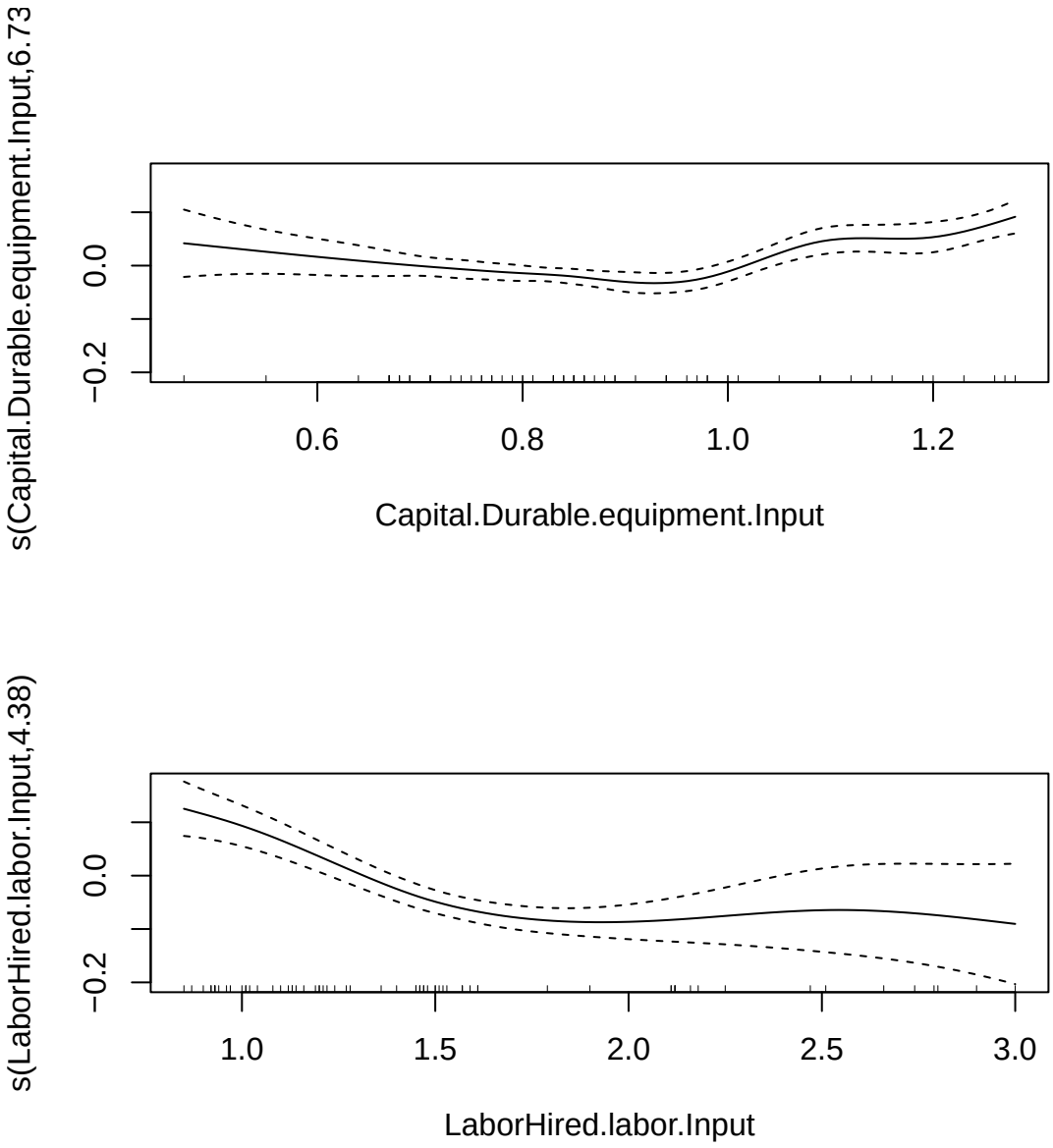
```
period = as.numeric(68)
period[1:47]= 'first'
period[48:68]= 'second'

data$period = as.factor(period)

model_gam = gam(Total.agricultural.output ~s(Capital.Durable.equipment.Input, bs = 'cr')
    + period:Capital.Service.buildings.Input
    + period:Labor.Self.employed.and.unpaid.family.Input
    + s(LaborHired.labor.Input, bs = 'cr')
   # + s(Capital.Inventories.Input)
```

```
    + Intermediate.Energy.Input
    + Intermediate.Pesticides.Input,
    data = data
   )
```

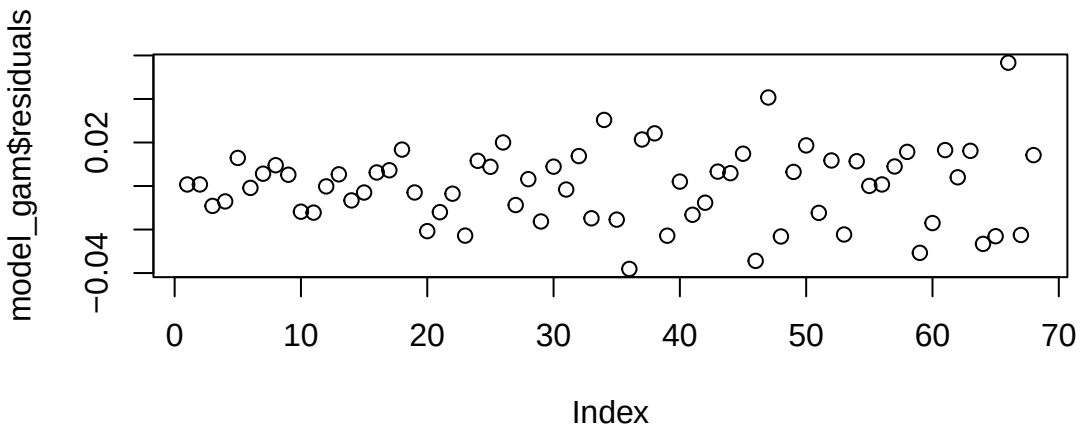
```
summary(model_gam)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## Total.agricultural.output ~ s(Capital.Durable.equipment.Input,
##     bs = "cr") + period:Capital.Service.buildings.Input + period:Labor.Self.employed.and.unpaid.famil
##     s(LaborHired.labor.Input, bs = "cr") + Intermediate.Energy.Input +
##     Intermediate.Pesticides.Input
##
## Parametric coefficients:
##                                                      Estimate Std. Error
## (Intercept)                                           0.92356    0.10504
## Intermediate.Energy.Input                             0.04836    0.02434
## Intermediate.Pesticides.Input                         0.18684    0.02976
## periodfirst:Capital.Service.buildings.Input          -0.18769    0.05614
## periodsecond:Capital.Service.buildings.Input          0.04182    0.15152
## periodfirst:Labor.Self.employed.and.unpaid.family.Input  -0.08791    0.02516
## periodsecond:Labor.Self.employed.and.unpaid.family.Input -0.29701    0.10028
##                                                      t value Pr(>|t|)
## (Intercept)                                            8.792 1.03e-11 ***
## Intermediate.Energy.Input                              1.987  0.05245 .
## Intermediate.Pesticides.Input                          6.278 8.15e-08 ***
## periodfirst:Capital.Service.buildings.Input           -3.343  0.00158 **
## periodsecond:Capital.Service.buildings.Input           0.276  0.78366
## periodfirst:Labor.Self.employed.and.unpaid.family.Input   -3.493  0.00101 **
## periodsecond:Labor.Self.employed.and.unpaid.family.Input  -2.962  0.00467 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                                       edf Ref.df      F  p-value
## s(Capital.Durable.equipment.Input) 6.730  7.731  7.178 3.24e-06 ***
## s(LaborHired.labor.Input)          4.379  5.297 11.181  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.991   Deviance explained = 99.3%
## GCV = 0.00055819  Scale est. = 0.00040954  n = 68
```

```
plot(model_gam)
```




```
plot(model_gam$residuals)
```




```
shapiro.test(model_gam$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  model_gam$residuals
## W = 0.97864, p-value = 0.2935
```

PREDICTION

```
durable_equipment.grid=seq(range(data$Capital.Durable.equipment.Input)[1],
                      range(data$Capital.Durable.equipment.Input)[2],length.out = 100)
#inventories.grid=seq(range(data$Capital.Inventories.Input)[1],
#                      range(data$Capital.Inventories.Input)[2],length.out = 10)
hired_labor.grid=seq(range(data$LaborHired.labor.Input)[1],
                      range(data$LaborHired.labor.Input)[2],length.out = 100)
grid = expand.grid(
    Capital.Durable.equipment.Input = durable_equipment.grid,
   # Capital.Inventories.Input = inventories.grid,
    LaborHired.labor.Input = hired_labor.grid,
    Capital.Service.buildings.Input = mean(data$Capital.Service.buildings.Input),
    Intermediate.Pesticides.Input = mean(data$Intermediate.Pesticides.Input),
    Intermediate.Energy.Input = mean(data$Intermediate.Energy.Input),
   Labor.Self.employed.and.unpaid.family.Input = mean(data$Labor.Self.employed.and.unpaid.family.Input)
   period = 'second'
)
pred_gam = predict(model_gam, newdata = grid)
```

4 Coefficients

```r
tab = summary(model_gam)
format(as.data.frame(tab$p.coeff), scientific = FALSE)
```

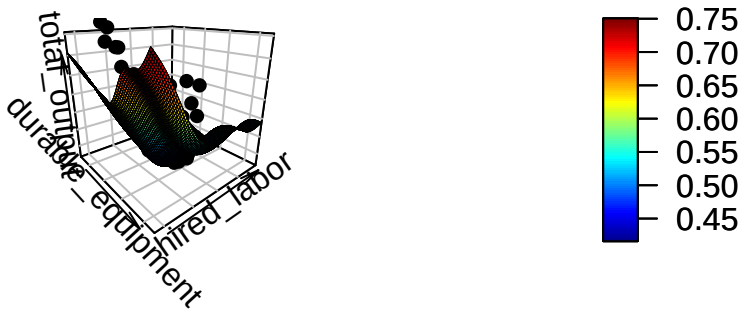
```
##                                                      tab$p.coeff
## (Intercept)                                           0.92356426
## Intermediate.Energy.Input                             0.04835645
## Intermediate.Pesticides.Input                         0.18684457
## periodfirst:Capital.Service.buildings.Input          -0.18769120
## periodsecond:Capital.Service.buildings.Input          0.04182428
## periodfirst:Labor.Self.employed.and.unpaid.family.Input  -0.08790924
## periodsecond:Labor.Self.employed.and.unpaid.family.Input -0.29701090
```

```r
as.data.frame(tab$s.table)
```

```
##                                        edf   Ref.df          F      p-value
## s(Capital.Durable.equipment.Input) 6.729754 7.730770  7.177641 3.240427e-06
## s(LaborHired.labor.Input)          4.378780 5.297005 11.180900 0.000000e+00
```

5 Prediction

```r
##da migliorare
plot3D::persp3D(
  x=durable_equipment.grid,
  y=hired_labor.grid,
  z=matrix(pred_gam, nrow=length(durable_equipment.grid), ncol=length(hired_labor.grid)),
  col.palette = heat.colors,

  xlab = 'durable_equipment',
  ylab = 'hired_labor',
  zlab = 'total_output',
  box = TRUE,
  #contour = TRUE,
  border='black',
  lwd=0.1,
  shade=0.1,
  bty="b2", # https://rdrr.io/cran/plot3D/man/perspbox.html
  phi = 20, theta = 50
)

with(
  data,
  plot3D::points3D(Capital.Durable.equipment.Input,
    LaborHired.labor.Input,
    Total.agricultural.output,
    col = 'black',
    size = 1,
    pch=16,
    add=TRUE
 )
)
```



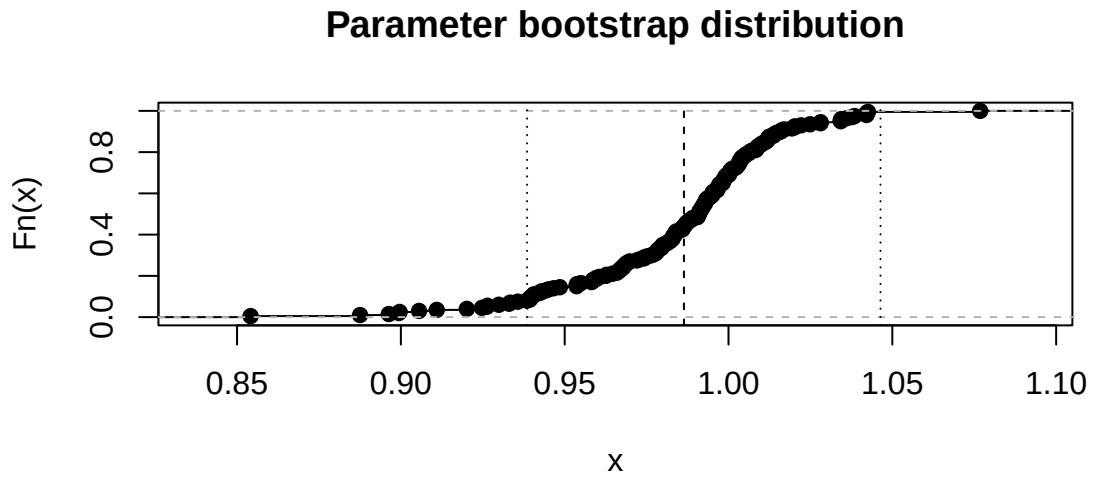
6 Bootstrap interval on response

```
period2 = as.numeric(4)
period2= 'second'

data_test$period = as.factor(period2)

data$period = as.factor(period)
service_buildings = data_test$Capital.Service.buildings.Input
pesticides = data_test$Intermediate.Pesticides.Input
durable_equipment = data_test$Capital.Durable.equipment.Input
hired_labor = data_test$LaborHired.labor.Input
self_employed= data_test$Labor.Self.employed.and.unpaid.family.Input
energy = data_test$Intermediate.Energy.Input
period2 = data_test$period
```

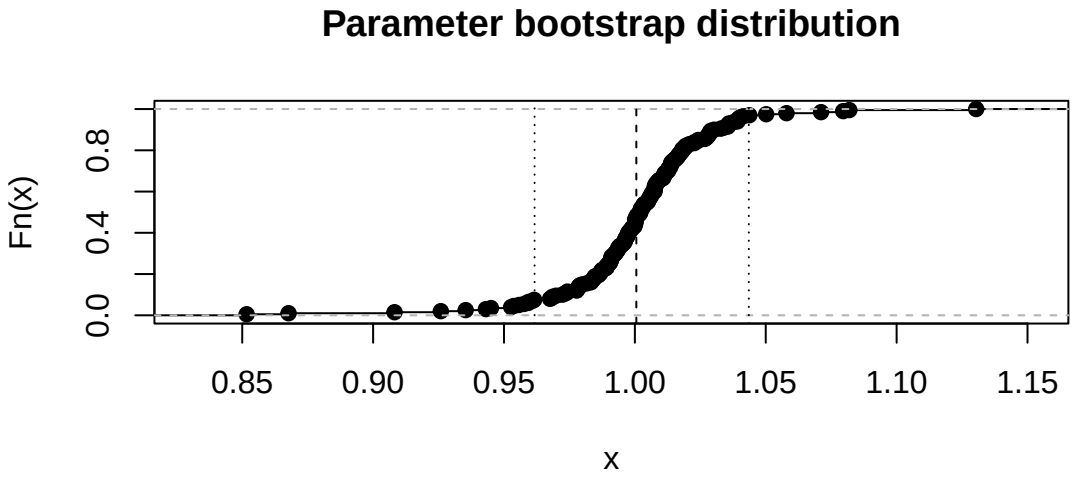
```
CI <- matrix(0,4,3)


set.seed(1)
for(i in 1:4){
  newdata <-data.frame(Capital.Service.buildings.Input=service_buildings[i],
                       Intermediate.Pesticides.Input=pesticides[i],
                       Capital.Durable.equipment.Input=durable_equipment[i],
                       LaborHired.labor.Input=hired_labor[i],
                       Intermediate.Energy.Input  =energy[i],
                       Labor.Self.employed.and.unpaid.family.Input= self_employed[i],
                       period = period2[i]
                       )
  B = 200
  fitted.obs <- predict(model_gam)
  res.obs <- data$Total.agricultural.output - fitted.obs
```

```r
  pred.obs = predict(model_gam, newdata = newdata)
  T.boot <- numeric(B)
  library(progress)
  pb <- progress_bar$new(
    format = "  processing [:bar] :percent eta: :eta",
    total = B, clear = FALSE)
  for (b in 1:B) {

    perm <- sample(1:nrow(data), replace = T)
    dataset.boot = data[perm,]

    model_gam_reduced.boot =
      mgcv::gam(Total.agricultural.output ~s(Capital.Durable.equipment.Input, bs = 'cr')
    + period:Capital.Service.buildings.Input
    + period:Labor.Self.employed.and.unpaid.family.Input
    + s(LaborHired.labor.Input, bs = 'cr')
   # + s(Capital.Inventories.Input)
    + Intermediate.Energy.Input
    + Intermediate.Pesticides.Input,
    data = dataset.boot
   )

    T.boot[b] <- predict(model_gam_reduced.boot, newdata = newdata)
    pb$tick()
  }
  inter <- diagnostic_bootstrap(distro = T.boot, obs = pred.obs)
  CI[i,] <- inter
}
```

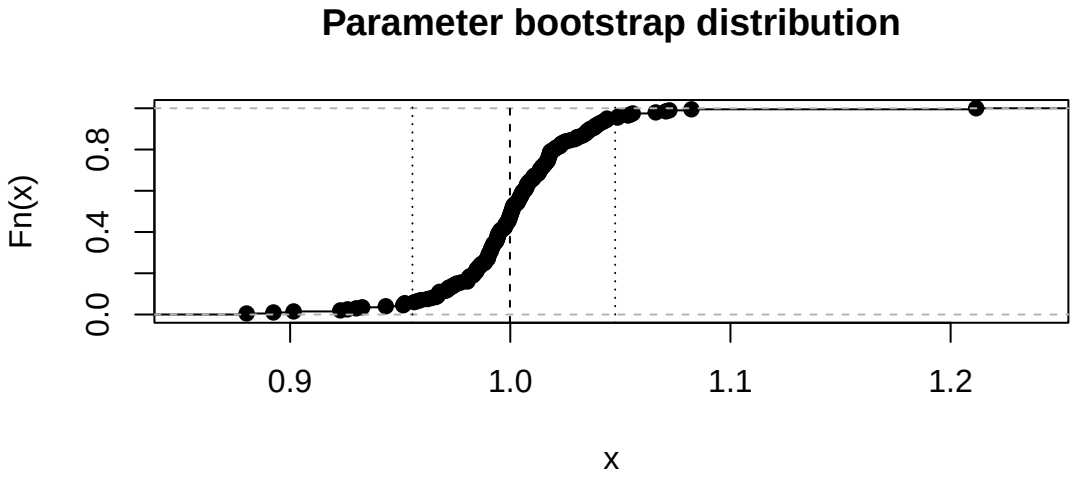
```
## [1] "Standard deviation:  0.0318047914484945"
## [1] "Bias:  -0.00181605159065634"
##       lwr       lvl       upr
## 0.9385257 0.9863989 1.0463802
```



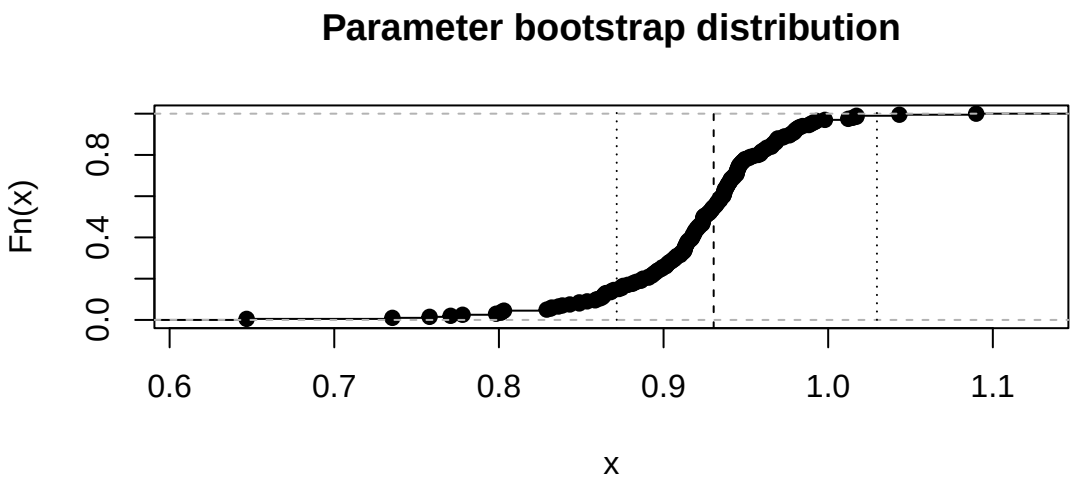
```
## [1] "Standard deviation:  0.0295436282143214"
## [1] "Bias:  0.00086442024096578"
##       lwr       lvl       upr
## 0.9616798 1.0005593 1.0435294
```



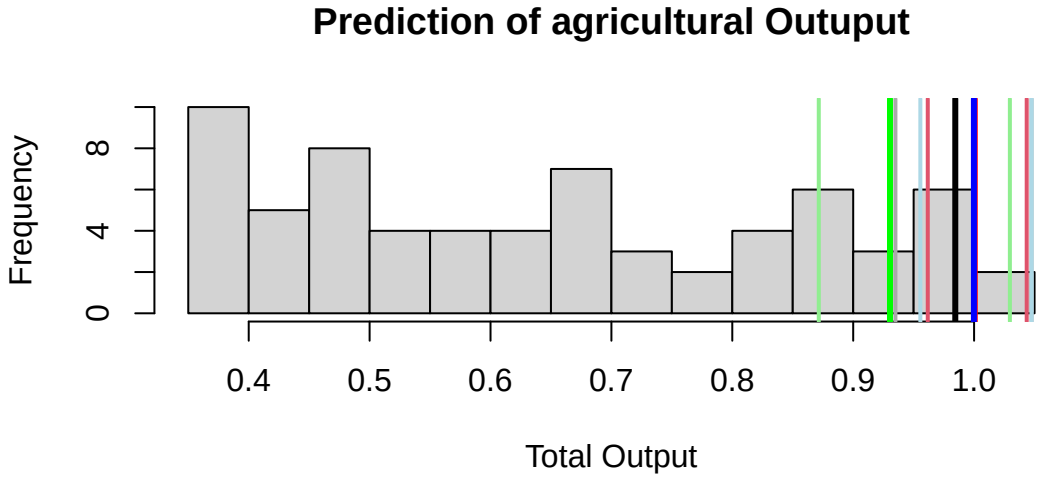

```
## [1] "Standard deviation:  0.0329424502253381"
## [1] "Bias:  0.00175601191604957"
##       lwr       lvl       upr
## 0.9555540 0.9998518 1.0476331
```



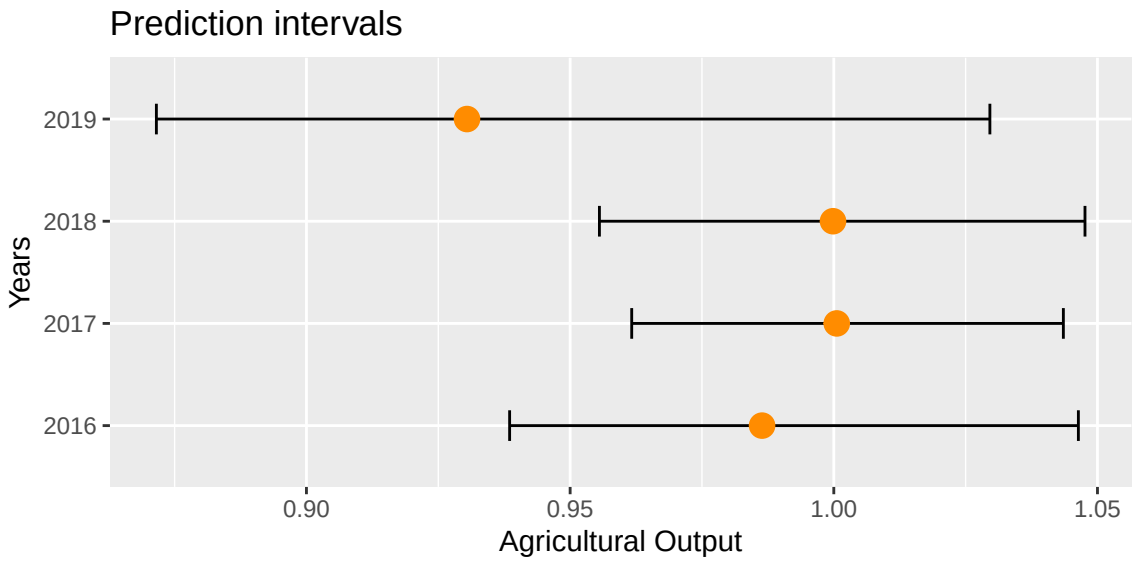

```
## [1] "Standard deviation:  0.0533039705422283"
## [1] "Bias:  -0.0105020796061384"
##       lwr       lvl       upr
## 0.8715392 0.9304462 1.0296020
```



and we compare them:




```
L = c(0.9385257, 0.9616798, 0.9555540, 0.8715392 )
U = c(1.0463802, 1.0435294 , 1.0476331 ,1.0296020  )
y = c(0.9863989, 1.0005593, 0.9998518, 0.9304462  )
x=c("2016","2017","2018", "2019")
df = data.frame(x=x, y =y)

ggplot(df, aes(x = x, y = y)) +
    geom_errorbar(aes(ymax = U, ymin = L), width = 0.3) +
    geom_point(size = 4, col = "darkorange") +
    coord_flip() +
    labs(x = "Years",
         y = "Agricultural Output",
         title = "Prediction intervals")
```


Prediction intervals
2019
2018
2017
2016
Years
0.90
0.95
1.00
1.05
Agricultural Output