Anexo 2.2. Segmentacion de hogares ENIGH

Juan Carlos Martinez-Ovando
ITAM

Se presenta y comenta el codigo para segmentar los hogares de la ENIGH 2016

Preambulo

Leemos los paquetes necesarios para realizar la segmentacion de los datos generados de la ENIGH.

```
source("./Code/load.R.packages.R")
```

```
## Loading required package: parallel

## Loading required package: foreach

## Loading required package: digest

## Loading required package: devtools

## Loading required package: abind

## Loading required package: aplpack

## Loading required package: tcltk

## Loading required package: bayesQR

## Loading required package: car

## Loading required package: carData

## Loading required package: class

## Loading required package: colorspace

## Loading required package: data.table

## Loading required package: doBy

## Loading required package: DPpackage

##

## DPpackage 1.1-7.4
```

```
##
## Copyright (C) 2006 - 2012, Alejandro Jara
## Department of Statistics
## P.U. Catolica de Chile
##
## Support provided by Fondecyt
## 11100144 grant.
##
## Loading required package: GB2
## Loading required package: effects
## lattice theme set by effectsTheme()
## See ?effectsTheme for details.
## Loading required package: foreign
## Loading required package: GPDPQuantReg
## Loading required package: pscl
## Classes and Methods for R developed in the
## Political Science Computational Laboratory
## Department of Political Science
## Stanford University
## Simon Jackman
## hurdle and zeroinfl functions by Achim Zeileis
## Loading required package: tmvtnorm
## Loading required package: mvtnorm
## Loading required package: Matrix
## Loading required package: stats4
## Loading required package: gmm
## Loading required package: sandwich
## Loading required package: coda
## Loading required package: reshape
```

```
##
## Attaching package: 'reshape'

## The following object is masked from 'package:Matrix':
##
##     expand

## The following object is masked from 'package:data.table':
##
##     melt

## The following object is masked from 'package:class':
##
##     condense

## Loading required package: lattice

## Loading required package: doParallel

## Loading required package: iterators

## Loading required package: Hmisc

## Loading required package: survival

## Loading required package: Formula

## Loading required package: ggplot2

##
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:base':
##
##     format.pval, units

## Loading required package: kamila

## Loading required package: leaps

## Loading required package: lmtest

## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
## Loading required package: memisc

## Loading required package: MASS

##
## Attaching package: 'memisc'

## The following objects are masked from 'package:Hmisc':
##
##     %nin%, html

## The following object is masked from 'package:ggplot2':
##
##     syms

## The following object is masked from 'package:reshape':
##
##     rename

## The following object is masked from 'package:Matrix':
##
##     as.array

## The following object is masked from 'package:car':
##
##     recode

## The following object is masked from 'package:foreach':
##
##     foreach

## The following objects are masked from 'package:stats':
##
##     contr.sum, contr.treatment, contrasts

## The following object is masked from 'package:base':
##
##     as.array

## Loading required package: monomvn

## Loading required package: pls

##
## Attaching package: 'pls'

## The following object is masked from 'package:stats':
##
##     loadings
```

```
## Loading required package: lars

## Loaded lars 1.2

## Loading required package: multcomp

## Loading required package: TH.data

##
## Attaching package: 'TH.data'

## The following object is masked from 'package:MASS':
##
##     geyser

## Loading required package: quantreg

## Loading required package: SparseM

##
## Attaching package: 'SparseM'

## The following object is masked from 'package:base':
##
##     backsolve

##
## Attaching package: 'quantreg'

## The following object is masked from 'package:Hmisc':
##
##     latex

## The following object is masked from 'package:survival':
##
##     untangle.specials

## Loading required package: Rcpp

## Loading required package: relimp

## Loading required package: sampling

##
## Attaching package: 'sampling'

## The following objects are masked from 'package:survival':
##
##     cluster, strata

## Loading required package: survey
```

```
## Loading required package: grid

##
## Attaching package: 'survey'

## The following object is masked from 'package:Hmisc':
##
##     deff

## The following object is masked from 'package:graphics':
##
##     dotchart

## Loading required package: yaml
```

```
load("./Datos.Modelo/mdi_variables_modelo.RData")
```

Los datos con los que trabajaremos son:

- `hogares_enigh_agr.csv` – Tabla de hogares de la ENIGH 2016

Preparacion

```
hogares_agr <- read.csv("./Datos.Aux/hogares_enigh_agr.csv", header=TRUE)
colnames(hogares_agr)
```

```
##  [1] "X"           "FOLIOVIV"    "FOLIOHOG"    "tam_loc"     "sexo_jefe"
##  [6] "edad_jefe"   "educa_jefe"  "tot_integ.x" "p12_64"      "p65mas"
## [11] "remesas"     "ing_cor"     "int0a12"     "int12a64"    "int65a98"
## [16] "depdemog"    "muj12a49"    "tot_per"     "ltot_per"    "p_esc3"
## [21] "p_esc5b"     "trab_sub"    "trab_ind"    "trab_s_pago" "seg_alim2"
## [26] "seg_alim3"   "seg_alim_a"  "seg_pop"     "ss"          "jtrab_ind"
## [31] "ssjtrabind"  "con_remesas" "viv_prop"    "viv_rent"    "tot_cuar"
## [36] "bao13"       "piso_fir"    "piso_rec"    "combustible" "sin_refri"
## [41] "sin_vehi"    "sin_compu"   "sin_vidvd"   "sin_telef"   "sin_horno"
## [46] "tot_integ.y" "ing_mon"     "ali_nme"     "alta_nme"    "veca_nme"
## [51] "viv_nme"     "lim_nme"     "cris_nme"    "ens_nme"     "sal_nme"
## [56] "tpub_nme"    "tfor_nme"    "com_nme"     "edre_nme"    "edba_nme"
## [61] "cuip_nme"    "accp_nme"    "otr_nme"     "reda_nme"    "ali_nmr"
## [66] "alta_nmr"    "veca_nmr"    "viv_nmr"     "lim_nmr"     "cris_nmr"
## [71] "ens_nmr"     "sal_nmr"     "tpub_nmr"    "tfor_nmr"    "com_nmr"
## [76] "edre_nmr"    "edba_nmr"    "cuip_nmr"    "accp_nmr"    "otr_nmr"
## [81] "reda_nmr"    "rururb"      "pago_esp"    "reg_esp"     "nomon"
## [86] "ict"         "tamhogesc"   "ictpc"       "plb_m"       "plb"
```

```r
dim(hogares_agr)
```

```
## [1] 43578    90
```

Ingreso corriente (dos versiones)

Variables ENIGH-CUIS

```r
# Numeric
hogares_agr[,var_enighcuis_num] <- lapply(hogares_agr[,var_enighcuis_num],
                                          as.numeric)
summary(hogares_agr[,var_enighcuis_num])
```

```
##     int0a12          int12a64         int65a98       trab_sub
##  Min.   :0.0000   Min.   : 0.000   Min.   :0.00   Min.   : 0.000
##  1st Qu.:0.0000   1st Qu.: 2.000   1st Qu.:0.00   1st Qu.: 1.000
##  Median :0.0000   Median : 3.000   Median :0.00   Median : 1.000
##  Mean   :0.8127   Mean   : 2.753   Mean   :0.25   Mean   : 1.434
##  3rd Qu.:1.0000   3rd Qu.: 4.000   3rd Qu.:0.00   3rd Qu.: 2.000
##  Max.   :7.0000   Max.   :13.000   Max.   :4.00   Max.   :12.000
##     trab_ind       trab_s_pago        seg_pop            ss
##  Min.   :0.0000   Min.   :0.0000   Min.   : 0.000   Min.   : 0.000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.: 0.000   1st Qu.: 0.000
##  Median :0.0000   Median :0.0000   Median : 2.000   Median : 1.000
##  Mean   :0.4594   Mean   :0.1014   Mean   : 2.037   Mean   : 1.521
##  3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.: 4.000   3rd Qu.: 3.000
##  Max.   :6.0000   Max.   :7.0000   Max.   :15.000   Max.   :13.000
##     depdemog         muj12a49        tot_per          tot_cuar
##  Min.   :0.0000   Min.   :0.000   Min.   : 1.000   Min.   : 1.00
##  1st Qu.:0.0000   1st Qu.:1.000   1st Qu.: 3.000   1st Qu.: 3.00
##  Median :0.3333   Median :1.000   Median : 4.000   Median : 4.00
##  Mean   :0.4351   Mean   :1.144   Mean   : 3.815   Mean   : 3.67
##  3rd Qu.:0.6667   3rd Qu.:2.000   3rd Qu.: 5.000   3rd Qu.: 4.00
##  Max.   :5.0000   Max.   :7.000   Max.   :20.000   Max.   :23.00
```

```r
# Categoric
hogares_agr[,var_enighcuis_cat] <- lapply(hogares_agr[,var_enighcuis_cat],
                                          as.numeric)
summary(hogares_agr[,var_enighcuis_cat])
```

```
##      p_esc3          p_esc5b        seg_alim2        seg_alim3
##  Min.   :0.0000   Min.   :0.000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.0000   Median :0.000   Median :0.0000   Median :0.0000
##  Mean   :0.2387   Mean   :0.374   Mean   :0.1512   Mean   :0.1276
##  3rd Qu.:0.0000   3rd Qu.:1.000   3rd Qu.:0.0000   3rd Qu.:0.0000
##  Max.   :1.0000   Max.   :1.000   Max.   :1.0000   Max.   :1.0000
```

```
##    seg_alim_a       jtrab_ind        ssjtrabind       con_remesas
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.00000
##  Median :0.0000   Median :0.0000   Median :0.00000   Median :0.00000
##  Mean   :0.1146   Mean   :0.2629   Mean   :0.03903   Mean   :0.05888
##  3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:0.00000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.00000   Max.   :1.00000
##     viv_prop         viv_rent          bao13          piso_fir
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:0.0000
##  Median :1.0000   Median :0.0000   Median :1.000   Median :1.0000
##  Mean   :0.7284   Mean   :0.1129   Mean   :0.599   Mean   :0.5702
##  3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:1.000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.000   Max.   :1.0000
##     piso_rec       combustible       sin_refri         sin_vehi
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.0000   Median :0.0000   Median :0.0000   Median :1.0000
##  Mean   :0.3948   Mean   :0.1992   Mean   :0.1557   Mean   :0.5711
##  3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##    sin_compu        sin_vidvd        sin_telef        sin_horno
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:1.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :1.0000   Median :1.0000   Median :1.0000   Median :1.0000
##  Mean   :0.7606   Mean   :0.6391   Mean   :0.7248   Mean   :0.6176
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
```

```
hogares_agr <- as.data.frame(hogares_agr)

table(hogares_agr[,c("rururb","tam_loc")])
```

```
##       tam_loc
## rururb     1     2     3     4
##      0 14352  6015  6955     0
##      1     0     0     0 16256
```

```
table(hogares_agr$rururb)
```

```
##
##     0     1
## 27322 16256
```

Parte I - Segmentacion Rural

```
# Numerical
hogares_rur_num <- hogares_agr[which(hogares_agr$rururb==1),
                                          var_enighcuis_seg_num]
hogares_rur_num <- lapply(hogares_rur_num,as.numeric)
hogares_rur_num <- as.data.frame(hogares_rur_num)

# Categorical
hogares_rur_cat <- hogares_agr[which(hogares_agr$rururb==1),
                                          var_enighcuis_seg_cat]
hogares_rur_cat <- lapply(hogares_rur_cat,factor)
hogares_rur_cat <- as.data.frame(hogares_rur_cat)

# Segmentacion rural
modelo_seg_rur <- kamila(hogares_rur_num,
                         hogares_rur_cat,
                         numClust=5,
                         numInit=10)

summary(modelo_seg_rur)
```

```
##              Length Class  Mode
## finalMemb    16256  -none- numeric
## numIter         10  -none- numeric
## finalLogLik      1  -none- numeric
## finalObj         1  -none- numeric
## finalCenters    60  -none- numeric
## finalProbs       8  -none- list
## input           10  -none- list
## verbose          0  -none- list
## nClust           0  -none- list
```

```
table(modelo_seg_rur$finalMemb)
```

```
##
##    1    2    3    4    5
## 3430 1796 3596 2445 4989
```

Unimos los clasificaciones

```
hogares_agr_rur <- hogares_agr[which(hogares_agr$rururb==1),]
dim(hogares_agr_rur)
```

```
## [1] 16256    90
```

```r
finalMemb_rur <- as.data.frame(modelo_seg_rur$finalMemb)
dim(finalMemb_rur)
```

```
## [1] 16256     1
```

```r
colnames(finalMemb_rur) <- c("finalMemb_rur")

hogares_agr_rur <- cbind(hogares_agr_rur,finalMemb_rur)
colnames(hogares_agr_rur)
```

```
##  [1] "X"             "FOLIOVIV"      "FOLIOHOG"      "tam_loc"
##  [5] "sexo_jefe"     "edad_jefe"     "educa_jefe"    "tot_integ.x"
##  [9] "p12_64"        "p65mas"        "remesas"       "ing_cor"
## [13] "int0a12"       "int12a64"      "int65a98"      "depdemog"
## [17] "muj12a49"      "tot_per"       "ltot_per"      "p_esc3"
## [21] "p_esc5b"       "trab_sub"      "trab_ind"      "trab_s_pago"
## [25] "seg_alim2"     "seg_alim3"     "seg_alim_a"    "seg_pop"
## [29] "ss"            "jtrab_ind"     "ssjtrabind"    "con_remesas"
## [33] "viv_prop"      "viv_rent"      "tot_cuar"      "bao13"
## [37] "piso_fir"      "piso_rec"      "combustible"   "sin_refri"
## [41] "sin_vehi"      "sin_compu"     "sin_vidvd"     "sin_telef"
## [45] "sin_horno"     "tot_integ.y"   "ing_mon"       "ali_nme"
## [49] "alta_nme"      "veca_nme"      "viv_nme"       "lim_nme"
## [53] "cris_nme"      "ens_nme"       "sal_nme"       "tpub_nme"
## [57] "tfor_nme"      "com_nme"       "edre_nme"      "edba_nme"
## [61] "cuip_nme"      "accp_nme"      "otr_nme"       "reda_nme"
## [65] "ali_nmr"       "alta_nmr"      "veca_nmr"      "viv_nmr"
## [69] "lim_nmr"       "cris_nmr"      "ens_nmr"       "sal_nmr"
## [73] "tpub_nmr"      "tfor_nmr"      "com_nmr"       "edre_nmr"
## [77] "edba_nmr"      "cuip_nmr"      "accp_nmr"      "otr_nmr"
## [81] "reda_nmr"      "rururb"        "pago_esp"      "reg_esp"
## [85] "nomon"         "ict"           "tamhogesc"     "ictpc"
## [89] "plb_m"         "plb"           "finalMemb_rur"
```

```r
100 * table(hogares_agr_rur$finalMemb_rur) /
  sum(table(hogares_agr_rur$finalMemb_rur))
```

```
##
##        1        2        3        4        5
## 21.09990 11.04823 22.12106 15.04060 30.69021
```

```r
write.csv(hogares_agr_rur,file="./Datos.Modelo/Tablas/hogares_agr_rur.csv")
```

Parte II - Segmentacion Urbano

```r
# Numerical
hogares_urb_num <- hogares_agr[which(hogares_agr$rururb==0),
                               var_enighcuis_seg_num]
hogares_urb_num <- lapply(hogares_urb_num,as.numeric)
hogares_urb_num <- as.data.frame(hogares_urb_num)

# Categorical
hogares_urb_cat <- hogares_agr[which(hogares_agr$rururb==0),
                               var_enighcuis_seg_cat]
hogares_urb_cat <- lapply(hogares_urb_cat,factor)
hogares_urb_cat <- as.data.frame(hogares_urb_cat)

# Segmentacion urbano
modelo_seg_urb <- kamila(hogares_urb_num,
                         hogares_urb_cat,
                         numClust=7,
                         numInit=10,
                         verbose=TRUE)

summary(modelo_seg_urb)
```

```
##              Length Class  Mode
## finalMemb    27322  -none- numeric
## numIter         10  -none- numeric
## finalLogLik      1  -none- numeric
## finalObj         1  -none- numeric
## finalCenters    84  -none- numeric
## finalProbs       8  -none- list
## input           10  -none- list
## verbose          6  -none- list
## nClust           0  -none- list
```

```r
table(modelo_seg_urb$finalMemb)
```

```
##
##    1    2    3    4    5    6    7
## 1691 2763 2998 5514 3030 4183 7143
```

Unimos los clasificaciones

```r
hogares_agr_urb <- hogares_agr[which(hogares_agr$rururb==0),]
dim(hogares_agr_urb)
```

```
## [1] 27322    90
```

```r
finalMemb_urb <- as.data.frame(modelo_seg_urb$finalMemb)
dim(finalMemb_urb)
```

```
## [1] 27322     1
```

```r
colnames(finalMemb_urb) <- c("finalMemb_urb")

hogares_agr_urb <- cbind(hogares_agr_urb,finalMemb_urb)
colnames(hogares_agr_urb)
```

```
##  [1] "X"             "FOLIOVIV"      "FOLIOHOG"      "tam_loc"
##  [5] "sexo_jefe"     "edad_jefe"     "educa_jefe"    "tot_integ.x"
##  [9] "p12_64"        "p65mas"        "remesas"       "ing_cor"
## [13] "int0a12"       "int12a64"      "int65a98"      "depdemog"
## [17] "muj12a49"      "tot_per"       "ltot_per"      "p_esc3"
## [21] "p_esc5b"       "trab_sub"      "trab_ind"      "trab_s_pago"
## [25] "seg_alim2"     "seg_alim3"     "seg_alim_a"    "seg_pop"
## [29] "ss"            "jtrab_ind"     "ssjtrabind"    "con_remesas"
## [33] "viv_prop"      "viv_rent"      "tot_cuar"      "bao13"
## [37] "piso_fir"      "piso_rec"      "combustible"   "sin_refri"
## [41] "sin_vehi"      "sin_compu"     "sin_vidvd"     "sin_telef"
## [45] "sin_horno"     "tot_integ.y"   "ing_mon"       "ali_nme"
## [49] "alta_nme"      "veca_nme"      "viv_nme"       "lim_nme"
## [53] "cris_nme"      "ens_nme"       "sal_nme"       "tpub_nme"
## [57] "tfor_nme"      "com_nme"       "edre_nme"      "edba_nme"
## [61] "cuip_nme"      "accp_nme"      "otr_nme"       "reda_nme"
## [65] "ali_nmr"       "alta_nmr"      "veca_nmr"      "viv_nmr"
## [69] "lim_nmr"       "cris_nmr"      "ens_nmr"       "sal_nmr"
## [73] "tpub_nmr"      "tfor_nmr"      "com_nmr"       "edre_nmr"
## [77] "edba_nmr"      "cuip_nmr"      "accp_nmr"      "otr_nmr"
## [81] "reda_nmr"      "rururb"        "pago_esp"      "reg_esp"
## [85] "nomon"         "ict"           "tamhogesc"     "ictpc"
## [89] "plb_m"         "plb"           "finalMemb_urb"
```

```r
100 * table(hogares_agr_urb$finalMemb_urb) /
  sum(table(hogares_agr_urb$finalMemb_urb))
```

```
##
##         1         2         3         4         5         6         7
##  6.189152 10.112730 10.972842 20.181539 11.089964 15.310007 26.143767
```

```r
write.csv(hogares_agr_urb,file="./Datos.Modelo/Tablas/hogares_agr_urb.csv")
```

Descripcion de salida

- `finalMemb` - Vector numerico con etiquetas numericas de asignacion

- `numIter`
- `finalLogLik` - Pseudo log-likelihood de la clasificacion final
- `finalObj`
- `finalCenters`
- `finalProbs`
- `input` - Vector con los parametros de inicio
- `nClust` - Descripcion de los resultados de la seleccion de segmentos
- `verbose` - Informacion compementaria

Exportacion

```
save( finalMemb_rur,finalMemb_urb,
      hogares_agr,
      hogares_agr_rur,hogares_agr_urb,
      hogares_rur_cat,hogares_rur_num,
      hogares_urb_cat,hogares_urb_num,
      modelo_seg_rur,modelo_seg_urb,
      var_enighcuis_cat,var_enighcuis_num,
      var_enighcuis_reg_cat,var_enighcuis_reg_num,
      var_enighcuis_seg_cat,var_enighcuis_seg_num,
     file = "./Datos.Modelo/mdi_segmentacion.RData")
ls()
```

```
##  [1] "finalMemb_rur"         "finalMemb_urb"
##  [3] "hogares_agr"           "hogares_agr_rur"
##  [5] "hogares_agr_urb"       "hogares_rur_cat"
##  [7] "hogares_rur_num"       "hogares_urb_cat"
##  [9] "hogares_urb_num"       "modelo_seg_rur"
## [11] "modelo_seg_urb"        "var_enighcuis_cat"
## [13] "var_enighcuis_num"     "var_enighcuis_reg_cat"
## [15] "var_enighcuis_reg_num" "var_enighcuis_seg_cat"
## [17] "var_enighcuis_seg_num"
```

```
gc()
```

```
##            used  (Mb) gc trigger  (Mb) max used  (Mb)
## Ncells  1987712 106.2    3886542 207.6  3264095 174.4
## Vcells 14197709 108.4   21535217 164.4 21534992 164.3
```