Data Quality
Imputing missing values

Anàlisi de Dades i Explotació de la Informació

Grau d'Enginyeria Informatica.

Information System tracking

Prof. Mónica Bécue Bertaut & Lidia Montero

Monica.becue@upc.edu lidia.montero@upc.edu

1. Introducción

- En las encuestas, los missing values suelen provenir de las no-respuestas
- Las no respuestas son una de las fuentes de error en la estimación (aquí de los ejes y valores propios, entre otros parámetros)
- Las no-respuestas no son al azar, responden a un mecanismo subyacente complejo y sutil.
- Es complejo, pero OBLIGATORIO, hacer una corrección de las no-respuestas
- Para las no respuestas parciales, se puede utilizar la información conocida sobre los respondientes para corregir los datos faltantes.

- se diferencian no respuestas parciales/ no respuestas totales
- se presume que existe una verdadero valor para la variable
- aquí, nos limitamos al caso en el cual la no-respuesta depende de las variables auxiliares pero no de la variable de interés

- no respuestas parciales: faltan datos para una parte del cuestionario
- no respuestas totales: por razones de rechazo, de imposibilidad de acceder a la unidad a encuestar, ...) . En la aplicación, no tenemos información sobre las no-respuestas totales, lo que no deja de ser un problema

<u>Soluciones (aproximadas)</u>

- no respuesta total: modelización o reponderación
- no respuesta parcial: imputación

Mecanismo de no-respuesta

Se supone que el mecanismo de no-respuesta se puede modelizar mediante una variable aleatoria. Responder o no responder se considera como el resultado de una experiencia aleatoria cuya modelización determinará el tratamiento de la no-respuesta

La no-repuesta puede ser

- Completamente aleatoria o uniforme
- Aleatoria (missing at random)
- Confundida, si depende de la variable de interés

El término no-confundida reagrupa completamente aleatoria u aleatoria, que es lo que se verá aquí

2. Imputation methods

Con la ayuda de la información auxiliar, o con la ayuda de las respuestas dadas en caso de "no-respuestas parciales", se reemplazan los datos faltantes de una unidad con la ayuda de uno de los métodos de "imputación" siguientes (que también suponen el uso de hipótesis de comportamiento):

deductivo, con la ayuda de una regla determinista (por ejemplo, un individuo de menos de 14 años pertenece a la población no activa)

predictivo (en función de las características observadas en la unidad, se hace referencia de los datos constatados en las unidades semejantes que han respondido) del tipo (por ejemplo)

Se supone que la variable Y (que se debe imputar) está ligada a la(s) variable(s) **X** por unas relaciones que pueden ser complejas. Varias soluciones.

solución mediante métodos multivariados como missMDA
que vemos ahora

3. Viewpoint implemented in missMDA

Starting point: to take into account the relationships among the variables and the similarities among the individuals

The process is iterative

It is described through examples, first in the case of quantitative variables and then for categorical variables.

orange {missMDA} R Documentation

Sensory description of 12 orange juices by 8 attributes.

Description

Sensory description of 12 orange juices by 8 attributes. Some values are missing.

Usage

```
data(orange)
```

Format

A data frame with 12 rows and 8 columns. Rows represent the different orange juices, columns represent the attributes.

Details

A sensory data frame.

Source

Francois Husson, Agrocampus Rennes

Examples

```
data(orange)
## Not run:
nb <- estim_ncpPCA(orange,ncp.min=0,ncp.max=5,method.cv="Kfold",nbsim=20,pNA=0.05)
res.comp <- imputePCA(orange,ncp=nb$ncp)
res.pca <- PCA(res.comp$completeObs)
resMI <- MIPCA(orange,ncp=nb$ncp)
plot(resMI)

## End(Not run)
```

Two words about method.CV=Kfold

- Model validation: dividing all the sample into training-set and test-set

- Cross-validation

  exhaustive cross-validation
    leave-p-out cross-validation
    leave-one-out cross validation

  non-exhaustive cross-valitadion
    k-fold cross-validation
    2-fold cross-validation

Measure of fit

in this case, mean-squared error (here, distance)

```
library(FactoMineR)
library(missMDA)
data(orange)
?estim_ncpPCA
nb <- estim_ncpPCA(orange,ncp.min=0,ncp.max=5,method.cv="Kfold",nbsim=20,pNA=0.05)
nb
res.comp <- imputePCA(orange,ncp=nb$ncp)
res.pca <- PCA(res.comp$completeObs)
resMI <- MIPCA(orange,ncp=2)
plot(resMI)

## Case MCA
?estim_ncpMCA
data(vnf)
vnf[1:20,]
result <- estim_ncpMCA(vnf[1:20,],ncp.min=0, ncp.max=5)
result

## Impute the indicator matrix and perform a MCA
?imputeMCA
tab.disj<-imputeMCA(vnf[1:20,], ncp=5)$tab.disj
tab.disj
res.impute <- imputeMCA(vnf[1:20,], ncp=5)
res.impute
res.impute$tab.disj
## The imputed indicator matrix can be used as an input of the MCA function of the
?MCA
res.mca <- MCA(vnf[1:20,],tab.disj=res.impute$tab.disj)
res.mca <- MCA(res.impute$completeObs)
```

Small examples

What happens if we consider that the null-values in the scores are actually “missing-values”

```
# Instalar el package que se va a utilizar; este package se tiene que instalar previamente
### CARGAR LOS PACKAGES FACTOMINER Y MASS

library(FactoMineR)
library(MASS)
library(missMDA)

rm(list = ls()) #eliminar objetos
```

Scores in Croatia survey

```
# Leer la base (que está en este mismo directorio)
base<-read.csv2("Croacia-Scores.csv",row.names=1,header=TRUE,dec=".")
colnames(base)
summary(base)

res.pca <- PCA(base[,6:13])

##
?estim_ncpPCA
nb <- estim_ncpPCA(base[,6:13],ncp.min=0,ncp.max=5,method.cv="Kfold",nbsim=20,pNA=0.05)
nb

res.comp <- imputePCA(base[,6:13],ncp=nb$ncp)
res.pca <- PCA(res.comp$completeObs)
```

Case MCA: MCA on life-style variables

```
> summary(base[,53:58])
                 C9                          C10                              C11
 smoking-no       :3288   cigarettes-20 more: 616   ever smoked-99      : 558
 smoking-sometimes: 421   cigarettes-99     : 535   ever smoked-no      :2167
 smoking-unknown  :  12   cigarettes-less 20:1128   ever smoked-sometime: 704
 smoking-yes      :1316   cigarettes-none   :2758   ever smoked-yes     :1608


                 C13                C14                C15
 alcohol-1ó2 week   : 536   drink-99    :1394   act-jogging: 836
 alcohol-99         :  27   drink-beer  :1094   act-reading:2260
 alcohol-every day  : 883   drink-spirit: 466   act-sport  : 184
 alcohol-never      :1004   drink-wine  :2083   act-unknown: 215
 alcohol-not now    : 386                       act-walking:1542
 alcohol-sometime   : 946
 alcohol-very rarely:1255
```

```
rm(list = ls()) #eliminar objet


library(FactoMineR)
library(Matrix)
library(missMDA)



###  lectura de los datos: tabla léxica y variables cerradas
# Leer la base (que está en este mismo directorio)
baseB<-read.csv2("Croacia-Scores-B-SocioeconoBIS.csv",row.names=1,header=TRUE,dec=".")
baseCD<-read.csv2("Croacia_Var_C_D_missing.csv",row.names=1,header=TRUE,dec=".",na.strings="      ")
base<-cbind(baseB,baseCD)
dim(base)
colnames(base)
summary(base)

res.mca<-MCA(base[,53:58])

result <- estim_ncpMCA(base[,53:58], ncp.min=0, ncp.max=5,method.cv="Kfold",nbsim=20,pNA=0.05)
result

## Impute the indicator matrix and perform a MCA
?imputeMCA
tab.disj<-imputeMCA(base[,53:58], ncp=5)$tab.disj
tab.disj[1:10,53:58]
res.impute <- imputeMCA(base[,53:58], ncp=5)
res.impute
res.impute$tab.disj[1:10,]
## The imputed indicator matrix can be used as an input of the MCA function of the
## FactoMineR package to perform the MCA on the incomplete data ozone
?MCA
res.mca <- MCA(base[,53:58],tab.disj=res.impute$tab.disj)
res.mca <- MCA(res.impute$completeObs)
```