Examples of data use: R platform[1]

To obtain the results of calculations more rapidly, you should limit the number of variables included in the datasets used. Please remember that permanent datasets cannot be saved.

The list headings contain an example of authentication data needed to submit the program to the BIRD system. **Obviously, users will have to substitute the content of the first three fields, which begin with the character "*", with their own personal authentication data.**

The statement ***package =** specifies the program language. R in this case.

The R commands are written in lower-case: please note that this language is case-sensitive.

All the examples assume that each row contains only one command and that a command can be extended on multiple rows, if too long.

1. Examples related to the Invind dataset

In each of these examples a file is opened in R format, containing the survey data. This file is identified with the conventional symbolic name: **indann.rdata**. You are shown how to restrict the analysis to a single sector (for example, the industrial sector, survey=1) or year (for example, 2005, annoril=2005). The first five examples show calculations exclusively for industrial firms in the year 2005.

Example n. 1

- We estimate a logit model in which the dichotomous dependent variable is membership of a group of firms. The explicative variables are the average number of workers (v24) and the variables relative to the geographical area of the headquarters and the sector of economic activity. These last two variables are devised so that they are suitable for treatment as dummies.

```
*user = user
*password = password
*project = project
*package = R

## loads data
load("indann.rdata")

## reads interest variables
annoril=dati$annoril
indagine=dati$indagine
peso=dati$peso
areag4=dati$areag4
settor11=dati$settor11
```

[1] R is an open-source package for the statistical analysis of data. For further information please visit http://cran.r-project.org/. The examples were prepared by Giuseppina Papadia.

```
v521=dati$v521
v24=dati$v24

## data frame is created containing the interest variables
oggetto<-data.frame("annoril"=annoril,"indagine"=indagine,"peso"=peso,
"areag4"=areag4,"settor11"=settor11,"v521"=v521,"v24"=v24)

## filters by anno=2005 and indagine=1
oggetto = oggetto[oggetto$annoril==2005 & oggetto$indagine==1,]

## create factor variables
oggetto$areag4<-factor(oggetto$areag4)
oggetto$settor11<-factor(oggetto$settor11)

## estimate logit model
fit <- glm(oggetto$v521 ~ oggetto$v24+oggetto$areag4+oggetto$settor11,
  weights = oggetto$peso, data = oggetto, family = binomial(logit))
summary(fit)
```

Example n. 2

- Only for industrial firms (indagine==1), the aim is to calculate the percentage change in the average number of workers and the number of firms belonging to a group as a proportion of the total and divided by geographical area. To obtain the weighted estimates correctly, you must perform the following steps (note that the creation of the variable var_occ serves merely to obtain estimates referred to a percentage change).

```
*user = user
*password = password
*project = project
*package = R

## loads appropriate package
library(survey)

## loads data
load("indann.rdata")

## reads interest variables
annoril=dati$annoril
indagine=dati$indagine
peso=dati$peso
popstr=dati$popstr
strato=dati$strato
areag4=dati$areag4
settor11=dati$settor11
```

```
v521=dati$v521
v15=dati$v15
v24=dati$v24
var_occ<-(v24-v15)*100

## creates data frame with interest variables and filters
##by anno = 2005 and indagine = 1
oggetto<-data.frame("annoril"=annoril,"indagine"=indagine,
  "peso"=peso,"popstr"=popstr,"strato"=strato,"areag4"=areag4,
"settor11"=settor11,"v521"=v521,"v15"=v15,"v24"=v24,
"var_occ"=var_occ)
oggetto = oggetto[oggetto$annoril==2005 & oggetto$indagine==1,]
##create factor variables
oggetto$areag4=factor(oggetto$areag4)

out_svy=svydesign(id=~1,      strata=~oggetto$strato,weights=~oggetto$peso,
  fpc=~oggetto$popstr, data=oggetto)
summary(out_svy)

##computes percentage change for average workforce
out_ratio= svyratio(~var_occ,~v15,out_svy)
print(out_ratio)

##computes percentage change for average workforce by geographical area
out_by_ratio<-svyby(~var_occ,by
  =~areag4,denominator=~v15,design=out_svy,svyratio)
print(out_by_ratio)

## share of firms belonging to a group
out_prop=svymean(~factor(v521),out_svy,na.rm=TRUE)
print(out_prop)
print(confint(out_prop))

##share of firms belonging to a group by geographical area
out_by_prop<-svyby(~factor(v521),by
  =~areag4,design=out_svy,svymean,na.rm=TRUE)
print(out_by_prop)
print(confint(out_by_prop))
```

Example n. 3

- In a similar way to the previous example, we want to calculate the percentage change in investment at constant prices. This variable is first treated to limit the effect of anomalous values (outliers), using the Type II Winsorization procedure, which is used to calculate the estimates of investment published in the Supplements to the Statistical Bulletin dedicated to surveys.

```
*user = user
*password = password
*project = project
*package = R

## loads appropriate package
library(survey)

## loads data
load("indann.rdata")

## reads interest variables
annoril=dati$annoril
indagine=dati$indagine
peso=dati$peso
strato=dati$strato
popstr=dati$popstr
areag4=dati$areag4
v200cos=dati$v200cos
v202cos=dati$v202cos
v810cos=dati$v810cos
v811cos=dati$v811cos
v24=dati$v24

## creates data frame with interest variables and filters
##by anno = 2005 and indagine = 1
oggetto<-data.frame("annoril"=annoril,"indagine"=indagine,"peso"=peso,
"strato"=strato,"popstr"=popstr,"areag4"=areag4,"v200cos"=v200cos,
"v202cos"=v202cos,"v810cos"=v810cos,"v811cos"=v811cos,"v24"=v24)
oggetto = oggetto[oggetto$annoril==2005 & oggetto$indagine==1,]
annoril=oggetto$annoril
indagine=oggetto$indagine
peso=oggetto$peso
strato=oggetto$strato
popstr=oggetto$popstr
areag4=oggetto$areag4
v200cos=oggetto$v200cos
v202cos=oggetto$v202cos
```

```
v810cos=oggetto$v810cos
v811cos=oggetto$v811cos
v24=oggetto$v24

##creates total investment at constant prices in 2004
i0tot=v200cos+v810cos

##creates total investment at constant prices in 2005
i1tot=v202cos+v811cos

## type II Winsorization
diffe=(i1tot-i0tot)/v24
f=1/peso

pp5<-quantile(diffe,0.05,na.rm=TRUE,weights=peso)
pp95<-quantile(diffe,0.95,na.rm=TRUE,weights=peso)

diffe_p5=pp5
diffe_p95=pp95
diffe=ifelse(is.na(diffe)==FALSE & diffe>diffe_p95,f*diffe+(1-f)*diffe_p95,diffe)
diffe=ifelse(is.na(diffe)==FALSE & diffe>diffe_p95 & f==1 &
 v24<5000,diffe_p95,diffe)
diffe=ifelse(is.na(diffe)==FALSE & diffe<diffe_p5,f*diffe+(1-f)*diffe_p5,diffe)
diffe=ifelse(is.na(diffe)==FALSE & diffe<diffe_p5 & f==1 &
 v24<5000,diffe_p5,diffe)

##creates i1totw containint total investment in 2005
## where the effect of outliers is attenuated
i1totw=i0tot+diffe*v24
var_inv=(i1totw-i0tot)*100
oggetto<-data.frame("annoril"=annoril,"indagine"=indagine,"peso"=peso,
"strato"=strato,"popstr"=popstr,"areag4"=areag4,"v200cos"=v200cos,
"v202cos"=v202cos,"v810cos"=v810cos,"v811cos"=v811cos,"v24"=v24,
"i0tot"=i0tot,"i1tot"=i1tot,"diffe"=diffe,"i1totw"=i1totw,"var_inv"=var_inv)

out_svy=svydesign(id=~1, strata=~oggetto$strato,weights=~oggetto$peso,
 fpc=~oggetto$popstr, data=oggetto)
summary(out_svy)

## computes percentage change of investments at constant prices, overall
out_ratio= svyratio(~var_inv,~i0tot,out_svy)
print(out_ratio)

##computes percentage change of investments at constant prices,
```

```
## by geographical area
out_by_ratio<-svyby(~var_inv,by
  =~areag4,denominator=~i0tot,design=out_svy,svyratio)
print(out_by_ratio)
```

Example n. 4

- We want to estimate a linear model where workforce (v24) is the LHS variable and turnover past year (v210) and geographical area (areag4, as dummies) are the RHS variables.

```
*user = user
*password = password
*project = project
*package = R

##loads data
load("indann.rdata")

## reads interest variables
annoril=dati$annoril
indagine=dati$indagine
peso=dati$peso
areag4=dati$areag4
v210=dati$v210
v24=dati$v24

## creates data frame with interest variables and filters
##by anno = 2005 and indagine = 1
oggetto<-data.frame("annoril"=annoril,"indagine"=indagine,
"peso"=peso,"areag4"=areag4,"v210"=v210,"v24"=v24)
oggetto = oggetto[oggetto$annoril==2005 & oggetto$indagine==1,]

##creates appropriate factor variables
oggetto$areag4<-factor(oggetto$areag4)

##estimates linear model
out_reg=lm(formula=oggetto$v24~oggetto$v210+oggetto$areag4,
  weights=oggetto$peso, data=oggetto)
summary(out_reg)
```

Example n. 5

- Same regression as above, but limited to firms whose workforce is between 1st and 95th percentile of the workforce distribution.

```
*user = user
*password = password
*project = project
*package = R

## loads data
load("indann.rdata")

## reads interest variables
annoril=dati$annoril
indagine=dati$indagine
peso=dati$peso
areag4=dati$areag4
v210=dati$v210
v24=dati$v24

## creates data frame with interest variables and filters
##by anno = 2005 and indagine = 1
oggetto<-data.frame("annoril"=annoril,"indagine"=indagine,"peso"=peso,
"areag4"=areag4,"v210"=v210,"v24"=v24)
oggetto = oggetto[oggetto$annoril==2005 & oggetto$indagine==1,]

## creates appropriate factor variables
oggetto$areag4<-factor(oggetto$areag4)

##creates pc1_v24 e pc99_v24 containing
##1° and 99° percentiles of variable v24
pc1_v24<-quantile(oggetto$v24,0.01)
pc99_v24<-quantile(oggetto$v24,0.99)

##excludes values outside percentile interval
oggetto<-oggetto[oggetto$v24<=pc99_v24 & oggetto$v24>=pc1_v24,]

##estimates linear model
out_reg=lm(formula=oggetto$v24~oggetto$v210+oggetto$areag4,
  weights=oggetto$peso, data=oggetto)
summary(out_reg)
```

Example n. 6

- The program below presents an example of panel assessment with random effects on a group of firms that have always been present in the years considered by the model. The analysis is restricted exclusively to the industrial sector (survey=1) in the years 2001-2006. We use turnover as a dependent variable (v210), and the average number of workers (v24) and operating result as co-variants (v545). Before being used as a dummy the variable v545 is re-codified.

```
*user = user
*password = password
*project = project
*package = R

##loads appropriate package
library(plm)

##reads data
load("indann.rdata")

##reads interest variables
annoril=dati$annoril
indagine=dati$indagine
areag4=dati$areag4
ident=dati$ident
v545=dati$v545
v210=dati$v210
v24=dati$v24
one=1

##creates data frame with interest variables
##by anno>=2001 e <=2006 e indagine=1
oggetto<-data.frame("ident"=ident,"annoril"=annoril,"indagine"=indagine,
"areag4"=areag4,"v545"=v545,"v210"=v210,"v24"=v24,"one"=one)
oggetto = oggetto[oggetto$annoril>=2001 & oggetto$annoril<=2006 &
 oggetto$indagine==1 ,]

## presence (years) of each firm in the sample
filimp <- aggregate(oggetto$one, by=list(oggetto$ident),
FUN=sum,simplify=TRUE)

##excludes firm present for less than 6 years
filimp=filimp[filimp$x==6,]

##renames variables accordingly
names(filimp) <- c("ident","numero")
```

```
## keeps only desired firms
oggetto1=merge(oggetto, filimp)

## creates appropriate factor variables
oggetto1$v545<-factor(oggetto1$v545)

##index variables ident and annoril
E <- pdata.frame(oggetto1, index = c("ident", "annoril"), drop.index =
 TRUE,row.names = TRUE)

##estimates panel regression model
outrandom <- plm(formula=v210 ~ v24 +v545, data=E, model="random")
summary(outrandom)
```

2. Examples related to the dataset of the Short-term Outlook survey

Example n. 7

- Tabulate, for all reference years available in the dataset, the distribution of variable STG3 (investment planned for the following year) for manufacturing firms with 50+ employees (indag3==1).

```
*user = user
*password = password
*project = project
*package = R

##loads data
load("sondstor.rdata")

##read interest variables
annoril=dati$annoril
indag3=dati$indag3
cc2=dati$cc2
stg3=dati$stg3
pesorisc=dati$pesorisc

##creates data frame with interest variables and filter
##by cc2>=2 and indag3=1
oggetto<-data.frame("annoril"=annoril,"indag3"=indag3,"cc2"=cc2,"stg3"=stg3,
"pesorisc"=pesorisc)
oggetto = oggetto[oggetto$cc2>=2 & oggetto$indag3==1,]

##weighted frequencies
out_tabs=xtabs(oggetto$pesorisc~oggetto$stg3 + oggetto$annoril, oggetto)
```

```
##table becomes a data frame, and renaming
df=data.frame(out_tabs)
names(df)<- c("stg3","annoril","fr")

##aggregate weights by year, and renaming
freq_par <- aggregate(df$fr, by=list(df$annoril),FUN=sum,simplify=TRUE)
names(freq_par)<- c("annoril","fr_tot")

##merges and computes relative frequency (var freq_rel)
out_merge=merge(df,freq_par)
out_df=data.frame(out_merge,"freq_rel"=out_merge$fr/out_merge$fr_tot*100)
print(out_df)
```

Example n. 8

- A merge between the two datasets (Invind and the Short-term Outlook survey) is performed to compare investment plans in 2006 to realizations (surveyed as continuous variables) in 2007. Only co-present firms which provided valid responses are included in the analysis.

```
*user = user
*password = password
*project = project
*package = R

load("sondstor.rdata")

annoril=dati$annoril
ident=dati$ident
stg3=dati$stg3

sond2006<-data.frame("annoril"=annoril,"ident"=ident,"stg3"=stg3)
sond2006<-sond2006[annoril==2006 & stg3!=9,]

load("indann.rdata")

annoril=dati$annoril
ident=dati$ident
v200=dati$v200
v810=dati$v810
v202=dati$v202
v811=dati$v811
peso=dati$peso
```

```
sond2007<-data.frame("annoril"=annoril,"ident"=ident,"v200"=v200,
"v810"=v810,"v202"=v202,"v811"=v811,"peso"=peso)
sond2007<-sond2007[annoril==2007,]

annoril=sond2007$annoril
ident=sond2007$ident
v200=sond2007$v200
v810=sond2007$v810
v202=sond2007$v202
v811=sond2007$v811
peso=sond2007$peso

i0tot=v200+v810
i1tot=v202+v811
i0tot=ifelse(i0tot==0, 0.1, i0tot)
i1tot=ifelse(i1tot==0, 0.1, i1tot)
varinv=(i1tot/i0tot-1)*100
varinvd=0
varinvd=ifelse(varinv< -10,1,varinvd)
varinvd=ifelse(varinv>= -10 & varinv< -3,2, varinvd)
varinvd=ifelse(varinv>= -3 & varinv<= 3,3, varinvd)
varinvd=ifelse(varinv>3 & varinv<= 10,4, varinvd)
varinvd=ifelse(varinv>10,5, varinvd)

sond2007<-data.frame(sond2007,"i0tot"=i0tot,"i1tot"=i1tot,"varinv"=varinv,
"varinvd"=varinvd)

names(sond2006)<-c("anno2006","ident","stg3")

out_merge=merge(sond2006,sond2007)

out_tabs=xtabs(out_merge$peso~out_merge$stg3 + out_merge$varinvd,
 out_merge)

tot_pesi=sum(out_tabs)

df=data.frame(out_tabs)
names(df)<- c("stg3","varinvd","fr")

df=data.frame(df,"freq_rel"=df$fr*100/tot_pesi)

print(df)
```