BANCA D'ITALIA
EUROSISTEMA

Indagine sulle imprese industriali e dei servizi

BIRD
Bank of Italy Remote access to micro Data

Esempi di utilizzo dati: Piattaforma R
Versione 3.3

Febbraio 2016

Sommario

Esempi di utilizzo dei dati: piattaforma R[1]

Per ottenere più rapidamente i risultati delle proprie elaborazioni si suggerisce di limitare il numero di variabili incluse nei dataset utilizzati nelle elaborazioni. Si ricorda che non si possono memorizzare dataset permanenti.

I listati contengono, in testa, un esempio di informazioni di autenticazione necessarie per la sottomissione del programma al sistema BIRD. Ciascun utente, ovviamente, dovrà sostituire il contenuto dei primi tre campi, che iniziano con il carattere "*", con le proprie informazioni di autenticazione.

Lo statement *package = dichiara invece il linguaggio utilizzato nel programma: R nel nostro caso.

Si ricorda che questo linguaggio è case-sensitive.

Tutti gli esempi presuppongono che su ogni riga ci sia un solo comando e che lo stesso comando possa essere esteso su più righe, se troppo lungo.

1. Esempi relativi ai dati dell'indagine sulle imprese industriali e dei servizi

Negli esempi che seguono viene aperto un file denominato **indann.rdata**. Vi si mostra come limitare l'analisi a un solo settore (ad esempio, il settore industriale, indagine==1) o a un solo anno (ad esempio al 2005, annoril==2005). I primi cinque esempi presentano delle elaborazioni sulle sole imprese industriali per l'anno 2005.

Esempio n.1

- Stimiamo, per le sole imprese industriali (indagine==1) un modello logit in cui la variabile dipendente dicotomica è l'appartenenza a un gruppo di imprese. Le variabili esplicative sono il numero medio di addetti (v24) e le variabili relative all'area geografica della sede amministrativa e al settore di attività economica.

[1] R è un ambiente opensource per l'analisi statistica dei dati; se si desiderano ulteriori informazioni sul linguaggio, si consiglia di visitare il sito http://cran.r-project.org/ . Gli esempi sono a cura di Giuseppina Papadia.

Queste ultime due variabili sono create in modo opportuno per essere trattate come *dummy*.

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

##lettura dei dati
load("indann.rdata")

##lettura delle variabili di interesse
annoril=dati$annoril
indagine=dati$indagine
peso=dati$peso
areag4=dati$areag4
settor11=dati$settor11
v521=dati$v521
v24=dati$v24

##creazione del data frame con le variabili di interesse
oggetto<-data.frame("annoril"=annoril,"indagine"=indagine,"peso"=peso,
"areag4"=areag4,"settor11"=settor11,"v521"=v521,"v24"=v24)

##filtro per anno=2005 e indagine=1
oggetto = oggetto[oggetto$annoril==2005 & oggetto$indagine==1,]

##trasformazione in factor delle variabili area geografica e settore
oggetto$areag4<-factor(oggetto$areag4)
oggetto$settor11<-factor(oggetto$settor11)

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

Esempio n.2

- Per le sole imprese industriali (indagine==1) si vuole calcolare la variazione percentuale degli addetti medi e la frazione di imprese appartenenti a un gruppo, sul totale e distintamente per area geografica. Per ottenere delle stime ponderate in modo corretto occorre eseguire le seguenti istruzioni (si noti che la creazione della variabile var_occ serve semplicemente a ottenere stime riferite a una variazione percentuale).

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

```
##caricamento del package survey
library(survey)

##lettura dei dati
load("indann.rdata")

##lettura delle variabili di interesse
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

##creazione del data frame con le variabili di interesse e filtro sui
## dati per anno = 2005 e 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,]
##Trasformazione in factor della variabile area geografica
oggetto$areag4=factor(oggetto$areag4)

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

##calcolo della variazione percentuale degli addetti medi sul totale
out_ratio= svyratio(~var_occ,~v15,out_svy)
print(out_ratio)

##calcolo della variazione percentuale degli addetti medi per area geografica
out_by_ratio<-svyby(~var_occ,by
  =~areag4,denominator=~v15,design=out_svy,svyratio)
print(out_by_ratio)

##calcolo della frazione di imprese appartenenti a un gruppo
out_prop=svymean(~factor(v521),out_svy,na.rm=TRUE)
print(out_prop)
print(confint(out_prop))
```

```
##calcolo della frazione di imprese appartenenti a un gruppo per

out_by_prop<-svyby(~factor(v521),by
  =~areag4,design=out_svy,svymean,na.rm=TRUE)
print(out_by_prop)
print(confint(out_by_prop))
```

Esempio n.3

- Analogamente al precedente esempio, si vuole calcolare la variazione percentuale degli investimenti a prezzi costanti. Essi sono precedentemente trattati per limitare l'effetto dei valori anomali (*outlier*) con un procedimento chiamato *winsorizzazione del secondo tipo*, utilizzato per il calcolo delle stime degli investimenti pubblicate nel Supplementi al Bollettino Statistico dedicati all'indagine.

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

##caricamento del package survey
library(survey)

##lettura dei dati
load("indann.rdata")

##lettura delle variabili di interesse
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

##creazione del data frame con variabili di interesse e filtro dati per anno=2005
##e 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

##creazione variabile investimenti totali a prezzi constanti per il 2004
i0tot=v200cos+v810cos

##creazione variabile investimenti totali a prezzi constanti per il 2005
i1tot=v202cos+v811cos

##procedimento di winsorizzazione del secondo tipo(sulla base del 5° e
##95° percentile
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)

##creazione della variabile i1totw contenente gli investimenti totali 2005 che
##attenua l'effetto dei dati anomali
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)

##calcolo della variazione percentuale degli investimenti a prezzi costanti sul
  totale
out_ratio= svyratio(~var_inv,~i0tot,out_svy)
print(out_ratio)

##calcolo della variazione percentuale degli investimenti a prezzi costanti per
##area geografica
out_by_ratio<-svyby(~var_inv,by
  =~areag4,denominator=~i0tot,design=out_svy,svyratio)
print(out_by_ratio)
```

Esempio n.4

- Si supponga di voler stimare un modello lineare dove il numero di addetti (variabile v24) è la dipendente e le covariate sono il fatturato (variabile v210) e l'area geografica dove è localizzata la sede amministrativa dell'impresa, quest'ultima utilizzata come variabile *dummy*.

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

##lettura dei dati
load("indann.rdata")

##lettura delle variabili di interesse
annoril=dati$annoril
indagine=dati$indagine
peso=dati$peso
areag4=dati$areag4
v210=dati$v210
v24=dati$v24

##creazione del data frame con variabili di interesse e filtro
##per anno=2005 e 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,]

##trasformazione in factor della variabile area geografica
oggetto$areag4<-factor(oggetto$areag4)

##stima del modello lineare per la variabile dipendente Numero di
## addetti
```

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

Esempio n.5

- Il seguente programma replica la stessa regressione dell'esempio precedente, ma la limita alle sole imprese con numero di addetti all'interno del primo e del 99-esimo percentile della distribuzione.

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

##lettura dei dati
load("indann.rdata")

##lettura delle variabili di interesse
annoril=dati$annoril
indagine=dati$indagine
peso=dati$peso
areag4=dati$areag4
v210=dati$v210
v24=dati$v24

##costruzione del data frame con variabili di interesse e filtro
##per anno=2005  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,]

##trasformazione in factor della variabile area geografica
oggetto$areag4<-factor(oggetto$areag4)

##creazione delle variabili pc1_v24 e pc99_v24 contenenti
##rispettivamente il 1° e il 99° percentile della variabile v24
pc1_v24<-quantile(oggetto$v24,0.01)
pc99_v24<-quantile(oggetto$v24,0.99)

##esclusione dei dati con v24 all'esterno dei percentili
oggetto<-oggetto[oggetto$v24<=pc99_v24 & oggetto$v24>=pc1_v24,]

##stima del modello di regressione lineare per la variabile dipendente
##v24 e limitatamente ai dati con numero di addetti all'interno
##del 1° e 99° percentile
```

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

Esempio n.6

- Il seguente programma presenta un esempio di stima panel ad effetti casuali su un gruppo di imprese sempre presenti negli anni considerati nel modello. L'analisi è limitata al solo settore industriale (indagine=1) per gli anni 2001-2006. Utilizziamo come variabile dipendente il fatturato (v210) e come covariate il numero medio di addetti (v24) e il risultato di esercizio (v545). La variabile v545 è prima ricodificata per essere utilizzata come *dummy*.

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

##caricamento del package plm
library(plm)

##lettura dei dati
load("indann.rdata")

##lettura delle variabili di interesse
annoril=dati$annoril
indagine=dati$indagine
areag4=dati$areag4
ident=dati$ident
v545=dati$v545
v210=dati$v210
v24=dati$v24
one=1

##creazione del data frame con le variabili di interesse e filtro
##per 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 ,]

##calcolo del numero di anni in cui un impresa è presente nell'indagine
filimp <- aggregate(oggetto$one, by=list(oggetto$ident),
FUN=sum,simplify=TRUE)

##filtro sulle imprese per escludere quelle presenti in meno
##di 6 indagini (6 anni)
filimp=filimp[filimp$x==6,]
```

```
##rinomina delle variabili del data frame contenenti le imprese di interesse
names(filimp) <- c("ident","numero")

##merge con il data frame di partenza per mantenere solo le imprese presenti
##in 6 anni
oggetto1=merge(oggetto, filimp)

##trasformazione in factor della variabile v545
oggetto1$v545<-factor(oggetto1$v545)

##indicizzazione delle variabili ident e annoril
E <- pdata.frame(oggetto1, index = c("ident", "annoril"), drop.index =
 TRUE,row.names = TRUE)

##stima del modello di regressione sul panel, a effetti casuali
outrandom <- plm(formula=v210 ~ v24 +v545, data=E, model="random")
summary(outrandom)
```

2. Esempi relativi ai dati del sondaggio congiunturale sulle imprese industriali e dei servizi

Esempio n.7

- Si vuole tabulare nell'archivio storico, per tutti gli anni disponibili, la distribuzione di frequenza delle modalità di risposta alla variabile STG3 (investimenti programmati per l'anno successivo) per le sole imprese manifatturiere con 50 addetti e oltre (indag3==1).

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

##lettura dei dati
load("sondstor.rdata")

##lettura delle variabili di interesse
annoril=dati$annoril
indag3=dati$indag3
cc2=dati$cc2
stg3=dati$stg3
pesorisc=dati$pesorisc

##creazione del data frame con variabili di interesse e filtro
##per cc2>=2 e indag3=1
oggetto<-data.frame("annoril"=annoril,"indag3"=indag3,"cc2"=cc2,
"stg3"=stg3,"pesorisc"=pesorisc)
oggetto = oggetto[oggetto$cc2>=2 & oggetto$indag3==1,]
```

```
##calcolo delle frequenze pesate
out_tabs=xtabs(oggetto$pesorisc~oggetto$stg3 + oggetto$annoril, oggetto)

##trasformazione della tabella in data frame e rinomina delle variabili
df=data.frame(out_tabs)
names(df)<- c("stg3","annoril","fr")

##calcolo delle aggregazioni di pesi per anno e rinomina delle variabili
freq_par <- aggregate(df$fr, by=list(df$annoril),FUN=sum,simplify=TRUE)
names(freq_par)<- c("annoril","fr_tot")

##merge con data frame di partenza e calcolo
##della frequenza relativa (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)
```

Esempio n.8

- Il seguente programma presenta un esempio di *merge* tra l'archivio storico del sondaggio e quello con i dati dell'indagine annuale, al fine di confrontare i piani di investimento rilevati nel corso del sondaggio del 2006 e le corrispondenti realizzazioni (rilevate in forma continua nell'indagine annuale sul 2007 e successivamente discretizzate). Sono elaborati i dati solo per le imprese che hanno partecipato a entrambe le indagini e che nel sondaggio hanno fornito un dato valido (diverso dalla modalità 9="non so, non intendo rispondere").

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

##lettura dei dati storici del sondaggio
load("sondstor.rdata")

##lettura delle variabili di interesse
annoril=dati$annoril
ident=dati$ident
stg3=dati$stg3

##creazione del data frame con le variabili di interesse e filtro
##per anno=2006 e stg3<>9
sond2006<-data.frame("annoril"=annoril,"ident"=ident,"stg3"=stg3)
sond2006<-sond2006[annoril==2006 & stg3!=9,]

##lettura dei dati sull'indagine 2007
load("indann.rdata")
```

```
##lettura delle variabili di interesse
annoril=dati$annoril
ident=dati$ident
v200=dati$v200
v810=dati$v810
v202=dati$v202
v811=dati$v811
peso=dati$peso

##creazione del data frame con le variabili di interesse  e filtro per anno=2007
sond2007<-data.frame("annoril"=annoril,"ident"=ident,"v200"=v200,
"v810"=v810,"v202"=v202,"v811"=v811,"peso"=peso)
sond2007<-sond2007[annoril==2007,]

##selezione delle variabili di interesse
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)

##nuova creazione data frame per l'aggiunta delle nuove variabili create
sond2007<-data.frame(sond2007,"i0tot"=i0tot,"i1tot"=i1tot,
"varinv"=varinv,"varinvd"=varinvd)

##rinomina delle variabili del data frame sond2006, per poi permettere
##l'operazione di merge
names(sond2006)<-c("anno2006","ident","stg3")

##merge dei due dataframe in modo da mantenere solo le imprese
##presenti in entrambi
out_merge=merge(sond2006,sond2007)
```

```
##calcolo frequenze pesate assolute
out_tabs=xtabs(out_merge$peso~out_merge$stg3 + out_merge$varinvd,
 out_merge)

##calcolo somma dei pesi
tot_pesi=sum(out_tabs)


##trasformazione della tabella in data frame e rinomina delle variabili
df=data.frame(out_tabs)
names(df)<- c("stg3","varinvd","fr")

##aggiunta della variabile freq rel e stampa del data frame
df=data.frame(df,"freq_rel"=df$fr*100/tot_pesi)

print(df)
```