RESEARCH DATA CENTER
REX - Remote Execution / Sistema di elaborazione a distanza

Indagine sulle Indagine sulle imprese industriali e dei servizi (INVIND)
Sondaggio congiunturale sulle Imprese Industriali e dei Servizi (SONDTEL)

Esempi di utilizzo dati: Piattaforma R

BANCA D'ITALIA
EUROSISTEMA

Luglio 2023

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.

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 importato il file denominato **indann_completo_csv.csv**. 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 2021.

Esempio 1: regressione logistica

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

```
##funzioni per la manipolazione dei dati
library(dplyr)
library(data.table)


##lettura dei dati
dati <- fread("indann_completo_csv.csv")

##filtro per anno=2021 e indagine=1
oggetto <- dati %>% filter(annoril==2021, indagine==1)

##creazione del data frame con le variabili di interesse
oggetto <- oggetto %>% select(annoril, indagine, peso, areag4, settor11, v521
, v24)

##trasformazione in factor delle variabili area geografica e settore
oggetto <- oggetto %>% mutate(areag4   = as.factor(areag4),
                              settor11 = as.factor(settor11))

##stima del modello logit
fit <- glm(v521 ~ v24 + areag4+ settor11,
```

[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/ .

```
          weights = peso, data = oggetto, family = "quasibinomial")
summary(fit)$coefficients
```

Esempio 2: distribuzioni di frequenza

- 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).

```
##funzioni per la manipolazione dei dati
library(dplyr)
library(data.table)


##funzioni per la gestione del disegno campionario
library(survey)

##lettura dei dati
dati <- fread("indann_completo_csv.csv")

##filtro per anno=2021 e indagine=1
oggetto <- dati %>% filter(annoril==2021, indagine==1)

##creazione del data frame con le variabili di interesse
oggetto <- oggetto %>% select(annoril, indagine, peso, areag4,
                              popstr, v521, v24, v15, strato)


##trasformazione in factor delle variabili area geografica e
##creazione della variabile var_occ
oggetto <- oggetto %>% mutate(areag4   = as.factor(areag4),
                              var_occ  = (v24-v15)*100)

##trasformazione in oggetto di tipo "survey" per utilizzare le funzioni del
##pacchetto survey precedentemente caricato
out_svy <- svydesign(id= ~1, strata= ~strato, weights= ~peso,
                     fpc= ~popstr, data=oggetto)

summary(out_svy)

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

##calcolo della variazione percentuale degli addetti medi per area geografica
##si riportano le numerosità campionarie per il controllo degli output
out_by_ratio <- svyby(~var_occ,by = ~areag4,
                      denominator = ~v15,
                      design=out_svy,
                      svyratio)
```

```
oggetto%>%group_by(areag4)%>%summarise(n_obs=n())
out_by_ratio

##calcolo della frazione di imprese appartenenti a un gruppo
out_prop <- svymean(~factor(v521),out_svy,na.rm=TRUE)
oggetto%>%group_by(v521)%>%summarise(n_obs=n())
out_prop
confint(out_prop)

##calcolo della frazione di imprese appartenenti a un gruppo per area
##geografica
out_by_prop <- svyby(~factor(v521), by =~areag4,
                     design = out_svy,
                     svymean, na.rm=TRUE)
oggetto%>%group_by(areag4,v521)%>%summarise(n_obs=n())
out_by_prop
confint(out_by_prop)
```

Esempio 3: regressione lineare

- 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*.

```
##funzioni per la manipolazione dei dati
library(dplyr)
library(data.table)
library(stargazer)


##lettura dei dati
dati <- fread("indann_completo_csv.csv")

##filtro per anno=2021 e indagine=1
oggetto <- dati %>% filter(annoril==2021, indagine==1)

##lettura delle variabili di interesse
oggetto <- oggetto %>% select(annoril, indagine, peso, areag4,
                              v24, v210)

##trasformazione in factor della variabile area geografica
oggetto <- oggetto %>% mutate(areag4 = as.factor(areag4))

##stima del modello lineare per la variabile dipendente Numero di
## addetti
out_reg <- lm(v24 ~ v210 + areag4,
              weights=peso, data=oggetto)
stargazer(out_reg, type="text")
```

Esempio 4: regressione lineare

- 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.

```
##funzioni per la manipolazione dei dati
library(dplyr)
library(data.table)
library(stargazer)


##lettura dei dati
dati <- fread("indann_completo_csv.csv")

##filtro per anno=2021 e indagine=1
oggetto <- dati %>% filter(annoril==2021, indagine==1)

##lettura delle variabili di interesse
oggetto <- oggetto %>% select(annoril, indagine, peso, areag4,
                              v24, v210)

##trasformazione in factor della variabile area geografica
oggetto <- oggetto %>% mutate(areag4 = as.factor(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 %>% filter(v24<=pc99_v24 & 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(v24 ~ v210 + areag4,
              weights=peso, data=oggetto)
stargazer(out_reg, type="text")
```

Esempio 5: regressione con effetti casuali per dati panel

- 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 2016-2021. 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*.

```
##funzioni per la manipolazione dei dati
library(dplyr)
library(data.table)
library(stargazer)
```

```
##funzioni per i dati di tipo panel
library(plm)

##lettura dei dati
dati <- fread("indann_completo_csv.csv")


##creazione del data frame con le variabili di interesse e filtro
oggetto <- dati %>% filter(annoril %in% 2016:2021, indagine==1)

##lettura delle variabili di interesse
oggetto <- oggetto %>% select(annoril, indagine, peso, areag4,
                              ident, v545,
                              v24, v210)
oggetto <- oggetto %>% group_by(ident) %>%

  ##calcolo del numero di anni in cui un impresa è presente nell'indagine
  mutate( num_anni = n()) %>%

  ##filtro escludere le imprese presenti in meno di 6 indagini (6 anni)
  filter(num_anni == 6 ) %>%

  ##trasformazione in factor della variabile v545
  mutate(v545 = as.factor(v545))

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

##stima del modello di regressione sul panel, a effetti casuali
out_random <- plm(formula=v210 ~ v24 +v545, data=oggetto_panel, model="random
")
stargazer(out_random, type="text")
```

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

Esempio 6: distribuzione di frequenza

- 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).

```
##funzioni per la manipolazione dei dati
library(dplyr)
library(data.table)


##funzioni per la gestione del disegno campionario
library(survey)
```

```
##funzioni per statistiche descrittive di dati
library(gtsummary)

##lettura dei dati
dati <- fread("sondstor.csv")


oggetto <- dati %>% filter(cc2>=2, indag3 == 1)

##lettura delle variabili di interesse
oggetto <- oggetto %>% select(annoril, indag3, cc2, stg3, pesorisc)

out_svy <- svydesign(id= ~1, weights= ~pesorisc,
                     data=oggetto)

##calcolo delle frequenze relative pesate
options(tibble.width = Inf)
options(tibble.print_min = Inf)
out_svy %>% tbl_svysummary(by = annoril,
                           percent = "column",
                           include = stg3,
                           statistic = list( stg3 ~ "{p}%")) %>%
                           as_tibble()
```

Esempio 7: integrazione delle due indagini

- 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 2021 e le corrispondenti realizzazioni (rilevate in forma continua nell'indagine annuale sul 2021 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").

```
##funzioni per la manipolazione dei dati
library(dplyr)
library(data.table)

##funzioni per statistiche descrittive di dati
library(gtsummary)

##lettura dei dati storici del sondaggio
dati <- fread("sondstor.csv")

##lettura delle variabili di interesse
##creazione del data frame con le variabili di interesse e filtro
##per anno=2006 e stg3<>9
sond2021 <- dati %>% filter(annoril == 2021, stg3 !=9) %>%
  select(annoril, ident, stg3)

##lettura dei dati sull'indagine INVIND 2021
```

```
dati <- fread("indann_completo_csv.csv")

invind2021 <- dati %>% filter(annoril == 2021) %>%
  select(annoril, ident, v200, v810, v202, v811, peso)%>%
  mutate(
    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= cut(varinv,
                 breaks = c(-Inf, -10, -3, 3, 10,Inf),
                 labels = c(1,2,3,4,5)))

##merge dei due dataframe in modo da mantenere solo le imprese
##presenti in entrambi
out_merge=inner_join(sond2021,invind2021) %>% select(stg3, peso, varinvd)

out_svy <- svydesign(id= ~1, weights= ~peso,
                     data=out_merge)

##calcolo delle frequenze relative pesate
options(tibble.width = Inf)
options(tibble.print_min = Inf)
out_svy %>% tbl_svysummary(by = stg3,
                           percent = "cell",
                           include = varinvd,
                           statistic = list( stg3 ~ "{p}%", varinvd ~ "{p}%")
) %>% as_tibble()
```