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

Survey of Industrial and service Firms (INVIND) Business Outlook Survey of Industrial and Service Firms (SONDTEL)

R Examples

BANCA D'ITALIA
EUROSISTEMA

July 2023

Table of contents

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 note that permanent datasets cannot be stored.

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 CSV file is imported, containing the survey data. This file is identified in the section "Available datasets". You are shown how to restrict the analysis to a single sector (for example, the industrial sector, indagine=1) or year (for example, 2021, annoril=2021). The first five examples show calculations exclusively for industrial firms in the year 2021.

Example 1: logistic regression

- We estimate, only for industrial firms (indagine==1), 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.

```
##functions for data manipulation
library(dplyr)
library(data.table)


##data loading
dati <- fread("indann_completo_csv.csv")

##filter by year and type of firm
oggetto <- dati %>% filter(annoril==2021, indagine==1)

##create data frame containing variables of interest
oggetto <- oggetto %>% select(annoril, indagine, peso, areag4, settor11, v521
, v24)

##create factor variables
oggetto <- oggetto %>% mutate(areag4   = as.factor(areag4),
                              settor11 = as.factor(settor11))


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

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

Example 2: frequency distribution

- 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).

```
##functions for data manipulation
library(dplyr)
library(data.table)

##functions for handling survey data
library(survey)

##data loading
dati <- fread("indann_completo_csv.csv")

##filter by year and type of firm
oggetto <- dati %>% filter(annoril==2021, indagine==1)

##create data frame containing variables of interest
oggetto <- oggetto %>% select(annoril, indagine, peso, areag4,
                              popstr, v521, v24, v15, strato)
##convert to factor the geographical area and create the
##new variable var_occ
oggetto <- oggetto %>% mutate(areag4   = as.factor(areag4),
                              var_occ  = (v24-v15)*100)
##convert data to survey object to use functions of the <survey> package
out_svy <- svydesign(id= ~1, strata= ~strato, weights= ~peso,
                     fpc= ~popstr, data=oggetto)
summary(out_svy)

##compute percentage change for average workforce
out_ratio <- svyratio(~var_occ,~v15, out_svy)
out_ratio

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

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


##compute 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)
```

```
out_by_prop
confint(out_by_prop)
```

Example 3: linear regression

- 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.

```
##functions for data manipulation
library(dplyr)
library(data.table)

##data loading
dati <- fread("indann_completo_csv.csv")

##filter by year and type of firm
oggetto <- dati %>% filter(annoril==2021, indagine==1)

##create data frame containing variables of interest
oggetto <- oggetto %>% select(annoril, indagine, peso, areag4, v24, v210)

##create factor variable
oggetto <- oggetto %>% mutate(areag4 = as.factor(areag4))

##fit linear regression model
out_reg <- lm(v24 ~ v210 + areag4, weights=peso, data=oggetto)
summary(out_reg)
```

Example 4: linear regression

- Same regression as above, but limited to firms whose workforce is between 1st and 99th percentile of the workforce distribution.

```
##functions for data manipulation
library(dplyr)
library(data.table)

##data loading
dati <- fread("indann_completo_csv.csv")

##filter by year and type of firm
oggetto <- dati %>% filter(annoril==2021, indagine==1)

##create data frame containing variables of interest
oggetto <- oggetto %>% select(annoril, indagine, peso, areag4, v24, v210)

##create factor variable
oggetto <- oggetto %>% mutate(areag4 = as.factor(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 %>% filter(v24<=pc99_v24 & v24>=pc1_v24)

##fit linear regression model
out_reg <- lm(v24 ~ v210 + areag4, weights=peso, data=oggetto)
summary(out_reg)
```

Example 5: panel regression with random effects

- 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 (indagine==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.

```
##functions for data manipulation
library(dplyr)
library(data.table)

##functions for panel data
library(plm)

##data loading
dati <- fread("indann_completo_csv.csv")

##filter by time interval and type of firm
oggetto <- dati %>% filter(annoril %in% 2016:2021, indagine==1)

##select variables of interest
oggetto <- oggetto %>% select(annoril, indagine, peso, areag4,
                              ident, v545, v24, v210)

oggetto <- oggetto %>% group_by(ident) %>%

##compute number of years of firm's partecipation to survey
mutate( num_anni = n()) %>%

##remove firms with less than 6 years of data
filter(num_anni == 6 ) %>%

##convert to factor variable v545
mutate(v545 = as.factor(v545))

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

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

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

Example 6: frequency distribution

- 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).

```
##functions for data manipulation
library(dplyr)
library(data.table)


##functions for handling survey data
library(survey)

##functions for data summary
library(gtsummary)

##data loading
dati <- fread("sondstor.csv")


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

##select variables of interest
oggetto <- oggetto %>% select(annoril, indag3, cc2, stg3, pesorisc)



##convert data to survey object to use functions of the <survey> package
out_svy <- svydesign(id= ~1, weights= ~pesorisc,
                     data=oggetto)
summary(out_svy)


##compute weighted relative frequencies
out_svy %>% tbl_svysummary(by = annoril,
                           percent = "column",
                           include = stg3,
                           statistic = list( stg3 ~ "{p}%")) %>%
  modify_footnote(c(all_stat_cols()) ~ NA) %>%
  modify_header(c(all_stat_cols()) ~ "**{level}**", label = "")
```

Example n. 7 merging the two surveys

- A merge between the two datasets (INVIND and the Short-term Outlook survey) is performed to compare investment plans for 2021 to realizations (surveyed as continuous variables) in 2021. Only co-present firms which provided valid responses are included in the analysis.

```r
##functions for data manipulation
library(dplyr)
library(data.table)


##data loading (short-term outlook survey)
dati <- fread("sond2021.csv")

##create data frame with data from the 2021 outlook survey
sond2021 <- dati %>% filter(annoril == 2021, stg3 !=9) %>%
  select(annoril, ident, stg3)

##data loading (INVIND)
dati <- fread("indann_completo_csv.csv")

##create data frame with data from the 2021 INVIND survey
invind2021 <- dati %>% filter(annoril == 2021) %>%
  select(annoril, ident, v200, v810, v202, v811, peso)%>%
  mutate(
    i0tot=v200+v810,
    i1tot=v202+v811,
    i0tot=if_else(i0tot==0, 0.1, i0tot),
    i1tot=if_else(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 data frames
out_merge=inner_join(sond2021,invind2021) %>% select(stg3, peso, varinvd)

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

##compute weighted relative frequencies
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()
```