

Description of the databases and examples from the Survey of Industrial and Service Firms

30 June 2023

For further information: statistiche@bancaditalia.it
www.bancaditalia.it/statistiche/index.html

1. General information

The data collected through the Survey of Industrial and Service Firms and covering all the editions of the survey from 1984 onwards are stored in a single database. Each edition of the survey is identified by the variable **annoril**, which indicates the year surveyed, i.e. the year prior to the one in which the data were actually collected (e.g. the data regarding the Survey for 2019, conducted in the early months of 2020, are identified by the variable **annoril=2019**).

Prior to 1998, the survey only covered manufacturing firms with 50 or more employees. Since 1999, the reference universe has been extended to include all industrial firms excluding construction and firms classified in subsection B (Mining and quarrying), D (Electricity, gas, steam and air-conditioning supply) and E (Water supply, sewerage, waste management and remediation) of the ATECO 2007 classification (derived from NACE Rev. 2).. In 2001, the survey was extended (with an abridged questionnaire) to include firms with 20 to 49 employees. In 2002, the reference population was expanded to include private non-financial service firms with 20 or more employees (while excluding from the market services category the following types of services: those provided by credit and insurance firms, public services and other social and personal services). In 2006, the survey was extended to construction firms with 20 or more employees. Starting from 2013, construction firms with 10-19 employees have also been included in the sample.

Since the questionnaires administered to construction firms are very different from those administered to firms in industry excluding construction and in services, their respective databases are also separate: the dataset containing only firms in industry excluding construction and in services is named 'indann_completo_csv.csv', while the dataset containing only construction firms is named 'costr.csv'. For a full list of all the variables observed in the individual years of the survey and available in the databases, see the variable description files published on the [Bank of Italy's website](http://www.bancaditalia.it).

In both datasets, each firm is assigned an identification code (the variable **ident**), which makes it possible to retrieve all the data available for firms that were included in the survey in multiple years. The code, generated randomly, is entirely unrelated to the variables that identify the firms and is used exclusively for longitudinal analyses. The **annoril/ident** pair identifies data on a given firm in a given year. Merely by way of example, the logical structure of the databases (which cannot be visualized by the user) is as follows:

ANNORIL	IDENT	VAR1	VAR2	VAR3
...
2013	1	31	25	400
2013	2	190	100	2000
...
2020	1	35	20	500
2020	7	240	100	7000
...

The databases contain some variables relating to the grossing-up of the sample estimates. Adopting a sampling weight allows for the alignment of the structure of the sample with that of the universe of firms according to the stratification variables;¹ its use in the analyses is recommended to obtain unbiased estimates that can be grossed up to the reference population.

¹ Given that the size of the population becomes known one or two years later, the weights are calculated provisionally using the most recent data available and are subsequently revised when the data on the actual population are published. Small deviations between the actual number of units in the reference population and the sum of the weights are due to post-stratification.

For the variables provided by the firms as ratios, or in any case without a scale factor, the weights applied should be such as to take account of the extent of the phenomenon. The databases also contain a number of variables relating to the classifications adopted for stratification on the basis of the sampling design. Please note that in the survey design, the geographical area is a post-stratification variable.

The financial data from the survey are expressed at **current prices and in thousands of euros** (except for the variables relating to wages, which are expressed in **euros**), also for the years prior to the introduction of the euro.

For several key variables, e.g. employment, investment and turnover, the survey gathers data referring to multiple subsequent years. For instance, the survey for 2021 inquires about average employment in 2021 and 2020 and about estimated employment for 2022 (the year in which the interview was conducted). By including multiple time horizons in the questionnaire, annual changes can be calculated without having to combine data from multiple runs of the survey. This choice, which is useful to stabilize the estimates for the rates of change, can lead to possible changes in values referring to the same year but observed in later editions of the survey. To better illustrate this phenomenon, let us consider the following example on average employment in the previous, current and following years (variables **v15**, **v24**, **v611m**) observed for the fictitious firm 'ident=999' in 2020 and 2021.

Reference year	Year for which the level declared refers				
	2018	2019	2020	2021	2022
2020		120	125	80	
2021			121	100	79

In the survey for 2021, the firm revised both the figures for average employment in 2020 and the expectations on 2021 it had reported a year earlier.

In this case, the database will show the following content:

Reference year	Firm	Average employment year t-1	Average employment year t	Average employment year t+1 (forecast)
<i>(annoril)</i>	<i>(ident)</i>	<i>(v15)</i>	<i>(v24)</i>	<i>(v611m)</i>
2019	999	115	120	125
2020	999	120	125	80
2021	999	121	100	79

During the interview, respondent firms that participated in the previous run of the survey are reminded of the figures provided the year before but they are free to revise them.

For some quantitative variables, missing data are imputed. At that same moment, a new flag variable is created, making it possible to know if the value in the database was provided directly by the firm or if it was imputed: if the figure of the reference variable is derived from imputation, the flag takes the value 1, otherwise it is blank. Flag variables follow the naming convention **f'X'**, where X is the name of the imputed variable (e.g. the imputation flag of the expected employment variable, **v611m**, is **fv611m**).

Since the Survey for 2010, some focus variables, specifically indicated in the database, have been recorded only for half of the sample; the sections of the questionnaire that contain these variables are clearly marked with the letters 'A' or 'B', to indicate to which half of the sample they pertain.

The whole sample was divided into two subsets using a random mechanism, so as to maintain the representativeness of the two halves of the original sample in respect of the reference population. When analysing the variables observed on a subset, it is important to use grossing-up weights created specifically for that purpose (see the descriptions for **pesoa** and **pesob** below) to make sure that the values are grossed up correctly in the subset examined.

2. Periodic updates of the databases

The databases are updated at the end of each run of the survey using the responses provided by the firms during the interviews. Before the complete dataset is created, the responses undergo quality checks. The estimates published in the Survey report, in the Statistics Series, refer to the databases as at the moment in which the calculations were performed and to the most recent data available regarding the population of

firms. Small differences between the database used to prepare the estimates presented in the report and the one made available to prepare the calculations remotely can be the result of both subsequent revisions of the responses provided by the firms and of revisions in the weighting scheme due to a change in the reference population.

The dataset containing the figures updated to the new reference year is made available to the users as soon as the Survey report is published on the [Bank of Italy's website](#). The weighting scheme is revised systematically, while any changes made to the responses provided in previous years are to be considered occasional.

3. Searchable variables in the databases

For confidentiality reasons, the databases are **not** searchable for questionnaire variables that would make it possible to identify the respondent firm. These variables are usually found in the first pages of the questionnaires and include: tax identification number, company name, subsidiary surveyed and group affiliation. Open text replies to questions of the type 'Other, specify' are also not searchable. Conversely, the databases contain some variables that are not included in the questionnaires but which are useful for calculation purposes. Besides the survey reference year (variable **annoril**), these include:

a) Classification variables for business sectors²

Name	Values	Description	ATECO 2002	ATECO 2007
settor11	SS1	Food industries, beverages and tobacco products	DA	10, 11, 12
	SS2	Textiles, clothing, and hide, leather and footwear products	DB, DC	13, 14, 15
	SS3	Coke manufacturing, chemical industry, rubber and plastics	DF, DG, DH	19, 20, 21, 22
	SS4	Processing of non-metallic minerals	DI	23
	SS5	Metal engineering industry	DJ, DK, DL, DM	24, 25, 26, 27, 28, 29, 30, 33
	SS6	Other manufacturing industries	DD, DE, DN	16, 17, 18, 31, 32
	SS7	Other industries excluding construction	CA, CB, CE	05, 06, 07, 08, 09, 35, 36, 37, 38, 39
	SS8	Wholesale and retail commerce	G	45, 46, 47
	SS9	Hotels and restaurants	H	55, 56
	SS10	Transport and communications	I	49, 50, 51, 52, 53, 58, 59, 60, 61, 62, 63
	SS11	Real estate activities, IT, etc.	K	68, 69, 70, 71, 72, 73, 74, 75, 77, 78, 79, 80, 81, 82
indag3	1	Manufacturing industry	D	C
	2	Extractive industries – energy	C, E	B, D, E
	3	Services	G, I, H, K	G, I, H, J, L, M, N
indagine	1	Industry excluding construction	C, D, E	C, B, D, E
	2	Services	G, I, H, K	G, I, H, J, L, M, N

b) Classification variables for size classes³

Name	Values	Description
cldimet	0	20 - 49 employees
	1	50 - 99 employees
	2	100 - 199 employees
	3	200 - 499 employees
	4	500 - 999 employees

² Up to 2009 these variables were obtained by aggregating some of the subsections of the ATECO 2002 code; as of 2010, they are based on the first two digits of the ATECO 2007 code. The variables in this table are contained only in the database 'indann_completo_csv.csv' and all the firms in the construction database have a single ATECO code (2002, 2007).

³ The size classes are based on the end-of-year workforce until 2003 and on the average annual workforce thereafter.

	5	1.000 and over employees
cc2	1	20 - 49 employees
	2	50 and over employees

c) Classification variables for geographical areas⁴

Name	Values	Description
areag4	1	North-West
	2	North-East
	3	Centre
	4	South and Islands
areag2	1	North, Centre
	2	South and Islands

d) Variables relating to the sampling design and weighting scheme⁵

- strato:** Consisting of 66 combinations of **settor11** and **cldimet**. It should be noted that firms with at least 5,000 workers have a weight of one and can be considered, individually for industry and services, as being part of two separate strata.
- poststrato:** Consisting of 48 combinations of **areag4**, **cc2** and a re-aggregation of the sectors of economic activity into 6 groups: 1) indag3=1; 2) indag3=2; 3) settor11=ss8; 4) settor11=ss9; 5) settor11=ss10; 6) settor11=ss11.
- peso:** Sample expansion weight: at the stratum and poststratum level, the sum of the weights is equal to the number of the reference population, separately for each year and does not take account of the panel dimension of the sample.
- pesoadd:** Sample expansion weight: at the stratum and poststratum level, the sum of the weights is equal to the number of the reference population of employees, separately for each year and does not take account of the length of the sample (available from 2007).
- pesoa:** Equivalent to **peso** for firms belonging to rotation 'A'. It should be used for the variables collected over half the sample (subsample A) (available from 2010).
- pesoadda:** Equivalent to **pesoadd** for firms belonging to rotation 'A'. It should be used for the variables collected over half the sample (subsample A) (available from 2010).
- pesob:** Equivalent to **peso** for firms belonging to rotation 'B'. It should be used for the variables collected over half the sample (subsample B) (available from 2010).
- pesoaddb:** Equivalent to **pesoadd** for firms belonging to rotation 'B'. It should be used for the variables collected over half the sample (subsample B) (available from 2010).
- pesorisc:** The **pesorisc** variable is obtained as the product of **peso** and a suitable scale factor in such a way that, year by year, it sums to the sample number.
- popstr:** Size of the population at the stratum level.
- poppostr:** Size of the population at the post-stratum level.

e) Classification variables for the share of turnover from exports

Name	Values	Description
a6	0	non-exporting firm
	1	less than 1/3 of turnover exported
	2	between 1/3 and 2/3 of turnover exported
	3	more than 2/3 of turnover exported
qexp	1	less than 1/3 of turnover exported or non-exporting firm

⁴ For confidentiality reasons, only the classifications for the geographical macro-areas, and not those for the individual regions and provinces, are available.

⁵ For further details on the sampling design, on the construction of the weights and deflators and on all other methodological aspects, see the '[Methodological Notes](#)' available on the Bank of Italy's website.

2	between 1/3 and 2/3 of turnover exported
3	more than 2/3 of turnover exported

f) Variables whose levels are available at current and constant prices

The databases, limitedly to investment and turnover levels, contain variables expressed both at current and at constant prices. Constant prices refer both to the latest available reference year and to the reference year of each individual survey. The latter make it possible to calculate changes at constant prices for the year in which they were recorded, even after adding data from new runs of the survey. For firms in industry excluding construction and in services, the deflators are derived from those provided by the firms themselves, by aggregating the data at subsection and size class level.

Description	At current prices	At constant prices...	
		... referred to the most recent year	... referred to the survey reference year
Fixed investment t-1	v200	v200cos ^(a)	v200k ^(a)
Fixed investment t	v202	v202cos	v202k
Fixed investment t+1	v203	v203cos	v203k
Turnover t-1	v209	v209cos	v209k
Turnover t	v210	v210cos	v210k
Turnover t+1	v437	v437cos	v437k
Turnover exported t-1	v211	v211cos	v211k
Turnover exported t	v212	v212cos	v212k
Turnover exported t+1	v438	v438cos	v438k
Intangible investment t-1	v810	v810cos	v810k
Intangible investment t	v811	v811cos	v811k
Intangible investment t+1	v812	v812cos	v812k

Note: (a) Available since 1985; (b) According to the ESA 2010, the item 'fixed assets' consists of expenditure for computer software, databases and mineral exploration and evaluation but it excludes patents and trademarks.

4. Special survey on the impact of the coronavirus ('Iseco')

Between 16 March and 14 May 2020, the Bank of Italy conducted a special survey on 3,503 Italian firms (of which 2,391 in industry excluding construction and 1,112 in private non-financial services), with the aim of obtaining timely data on the economic fall-out from the spread of the COVID-19 epidemic under way in Italy at the time of the survey.

The sampling design, the reference sample, the weighting scheme and the type of questionnaire responses were consistent with those of the Survey of Industrial and Service Firms ('Invind') for 2019. The two surveys were conducted using the same methodology, although the Iseco survey was prepared and launched later than the usual Invind survey. Therefore, firms did not necessarily respond to the two surveys at the same time, possibly not even after the Iseco survey was launched. This implies that a firm may have reported different short-term expectations in the two surveys. The difference in sample size for the two surveys can be explained by the fact that the surveys are compiled on a voluntary basis, which may have led some firms to reply to only one of them.

For the full list of variables recorded and the types of questionnaire responses used in the Iseco survey, see the variable description files published on the [Bank of Italy's website](#).

5. Examples

5.1. Examples based on the software 'R'⁶

To obtain more rapid computing results, limit the number of variables in the datasets used to calculate the estimates. Please note that 'R' is a case-sensitive programming language.

The examples below are constructed by importing the file **indann_completo_csv.csv**, i.e. the file that includes firms in industry excluding construction and in services but excludes construction firms. The examples show how to restrict the analysis to only one sector (e.g. the industrial sector, `indagine==1`) or to only one year (e.g. 2005, `annoril==2005`). The first five examples show calculations only for industrial firms for the year 2021.

Example 1: logistic regression

- This shows how to estimate a logit model where the dichotomous dependent variable is group affiliation. The explanatory variables are: average number of employees (**v24**), geographical area of the administrative offices, and sector of economic activity. The last two variables are designed to be treated as dummy 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, 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)
```

Example 2: frequency distribution

- This shows how to calculate the percentage change in the average number of employees and the subset of firms affiliated to a group, as a share in the total and separately by geographical area. To obtain correctly weighted estimates, please follow these instructions (note that the variable **var_occ** is created for the sole purpose of obtaining estimates for a percentage change). The analysis is restricted to industrial firms only (`indagine==1`).

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

##functions for handling survey data
library(survey)
```

⁶ 'R' is an opensource software environment for statistical data analysis; for more information on the programming language, visit the website <http://cran.r-project.org/>.


```

##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

- This shows how to estimate a linear regression model where the number of employees (variable **v24**) is the dependent variable and the covariates are the turnover (variable **v210**) and the geographical area (**areag4**) of the firm's administrative offices, the latter being used as a dummy variable.

```

##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

- This example replicates the same regression of Example 3, but limited to firms with a number of employees within the 1st and 99th percentile of the 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: random-effects regression for panel data

- This shows an example of random-effects panel estimation on a group of firms that are always included in the years considered in the model. The analysis is restricted to the industrial sector (indagine=1) for the years 2016-21. We use turnover as a dependent variable (v210) and the average number of employees (v24) and the net result for the year (v545) as covariates. v545 is first recoded to be used as a dummy variable.

```
##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 participation 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)

```

5.2. Examples based on the software ‘Stata’⁷

The examples below are constructed by importing the CSV file that contains the survey data. They show how to restrict the analysis to only one sector (e.g. the industrial sector, `indagine==1`) or to only one year (e.g. 2021, `annoril==2021`). The first five examples show calculations only for industrial firms for the year 2021, while the sixth example shows a panel analysis.

‘Stata’ is case-sensitive, so all commands must be lower case.

Example 1

- This shows how to estimate, only for industrial firms (`indagine==1`), a logit model where the dichotomous dependent variable is group affiliation. The explanatory variables are: average number of employees (`v24`), geographical area of the administrative offices, and sector of economic activity. The last two variables are designed to be treated as dummy variables.

```

#delimit;

import delimited "indann_completo_csv.csv";

keep annoril indagine peso areag4 settor11 v521 v24;

keep if annoril==2021 & indagine == 1;

/* creation of the geographical area and sector of economic activity dummies */

tabulate areag4, gen(areag4d);

```

⁷ ‘Stata’ is a registered trademark of StataCorp LP, 4905 Lakeway Drive, College Station, TX 77845 USA.

```

tabulate settor11, gen(settor11d);

/* this creates 4 geographical area dummies and 7 sector dummies */

/* estimation of the logit, in which one dummy is omitted for both area and sector, which acts as a reference for the others */

logit v521 v24 areag4d1-areag4d3 settor11d1-settor11d6 [pweight=peso];

```

Example 2

- Limitedly to industrial firms (indagine==1), this shows how to calculate the percentage change in the average number of employees and the subset of firms affiliated to a group, as a share in the total and separately by geographical area. To obtain correctly weighted estimates, follow these instructions (please note that the variable **var_occ** is created for the sole purpose of obtaining estimates for a percentage change).

```

#delimit;

import delimited "indann_completo_csv.csv";

keep annoril indagine peso popstr strato areag4 settor11 v521 v15 v24;

keep if annoril==2021 & indagine == 1;

svyset _n[pw=peso], strata(strato) fpc(popstr);

generate var_occ=(v24-v15)*100;

svy:ratio var_occ/v15;

svy:ratio var_occ/v15, over(areag4);

svy:proportion v521;

svy:proportion v521, over(areag4);

```

Example 3

- Similarly to example 2, this shows how to calculate the percentage change in investments at constant prices. The values are previously treated to limit the effects of outliers by applying Type II winsorization, a method used to calculate the investment estimates published in the survey report.

```

#delimit;

import delimited "indann_completo_csv.csv";

keep annoril indagine peso strato popstr areag4 v200cos v202cos v810cos v811cos v24;

keep if annoril ==2021 & indagine == 1;

/* Fixed investment at constant prices referred to the most recent year in 2020 */

generate i0tot=v200cos+v810cos;

/* Fixed investment at constant prices referred to the most recent year in 2021 */

generate i1tot=v202cos+v811cos;

/* Type II Winsorization */

```

```

generate diffe=(i1tot-i0tot)/v24;
generate f=1/peso;
su diffe [w=peso], de;
scalar pp5=r(p5);
scalar pp95=r(p95);
generate diffe_p5=pp5;
generate diffe_p95=pp95;
replace diffe=f*diffe+(1-f)*diffe_p95 if diffe !=. & diffe>diffe_p95;
replace diffe=diffe_p95 if diffe != . & diffe>diffe_p95 & f==1 & v24<5000;
replace diffe=f*diffe+(1-f)*diffe_p5 if diffe !=. & diffe<diffe_p5;
replace diffe=diffe_p5 if diffe != . & diffe<diffe_p5 & f==1 & v24<5000;
/* creation of a new variable i1totw*/
generate i1totw=i0tot+diffe*v24;
svyset _n[pw=peso], strata(strato) fpc(popstr);
generate var_inv=(i1totw-i0tot)*100;
svy:ratio var_inv/i0tot;
svy:ratio var_inv/i0tot, over(areag4);

```

Example 4

- This shows how to estimate a linear regression model where the number of employees (variable **v24**) is the dependent variable and the covariates are the turnover (variable **v210**) and the geographical area of the firm's administrative offices, the latter being used as a dummy variable.

```

#delimit;
import delimited "indann_completo_csv.csv";
keep annoril indagine peso areag4 v210 v24;
keep if annoril ==2021 & indagine == 1;
/* creation of the geographical area dummies */
tabulate areag4, gen(areag4d);
/* this creates 4 geographical area dummies */
/* estimation of the regression, in which one dummy is omitted for the area, which acts as a reference for the others */
regress v24 v210 areag4d1 areag4d2 areag4d3 [pweight=peso];

```

Example 5

- This example replicates the same regression of Example 3, but limited to firms with a number of employees within the 1st and 99th percentile of the distribution.

```
#delimit;
import delimited "indann_completo_csv.csv";
keep annoril indagine peso areag4 v210 v24;
keep if annoril ==2021 & indagine == 1;
/* creation of the geographical area dummies */

tab areag4,gen(areag4d)

/* this creates 4 geographical area dummies */
/* creation of the two variables pc1_v24 and pc99_v24 containing the first and 99th per
centiles of the variable v24 */
egen pc1_v24=pctile(v24), p(1);
egen pc99_v24=pctile(v24), p(99);
/* estimation of the regression, in which one dummy is omitted for the area, which acts
as a reference for the others, and the units with v24 outside the percentiles are omitt
ed */

regress v24 v210 areag4d1 areag4d2 areag4d3 [pweight=peso]
if v24>=pc1_v24 & v24<=pc99_v24;
```

Example 6

- This shows an example of random-effects panel estimation on a group of firms that are always included in the years considered in the model. The analysis is restricted to the industrial sector (indagine=1) for the years 2016-21. We use turnover as a dependent variable (**v210**) and the average number of employees (**v24**) and the net result for the year (**v545**) as covariates. **v545** is first recoded to be used as a dummy variable.

```
#delimit;
import delimited "indann_completo_csv.csv";
keep annoril indagine ident areag4 v545 v210 v24;
keep if inrange(annoril, 2016, 2021) & indagine == 1;
/* selection of the firms covered by the survey in the 6 years from 2016 to 2021 */

generate one=1;
sort ident;
by ident: egen conta=sum(one);
keep if conta==6;
/* creation of the result for the year dummies */

tabulate v545, gen(v545d);

/* estimation of the regression model on the panel, in which one dummy is omitted for t
he result for the year, which acts as a reference for the others */

iis ident;
tis annoril;
xtreg v210 v24 v545d1 v545d2 v545d3 v545d4,re;
```