BANCA D'ITALIA
EUROSISTEMA

Descrizione archivi ed esempi di utilizzo dell'Indagine sulle Imprese Industriali e dei Servizi

30 giugno 2023

Per informazioni: statistiche@bancaditalia.it
www.bancaditalia.it/statistiche/index.html

1. Indicazioni generali

Le informazioni raccolte nell'ambito dell'Indagine sulle imprese industriali e dei servizi sono disponibili a partire dal 1984 in un unico archivio; ogni rilevazione è identificabile a partire dalla variabile **annoril,** che indica l'anno oggetto di rilevazione, ovvero l'anno precedente l'effettiva raccolta dei dati (ad esempio i dati relativi all'indagine sul 2019, condotta nei primi mesi del 2020, sono identificati dalla variabile annoril=2019).

Fino all'indagine sul 1998, la rilevazione si limitava alle imprese del settore manifatturiero con 50 addetti e oltre. A partire dall'indagine sul 1999, l'universo di riferimento è stato ampliato a tutta l'industria in senso stretto, integrando il campione con imprese della Sottosezione ATECO 2007 (derivata dalla Nace Rev. 2) B (estrazione di minerali da cave e miniere), D (fornitura di energia elettrica, gas, vapore e aria condizionata) ed E (fornitura di acqua, reti fognarie, attività di gestione dei rifiuti e risanamento). Nel 2001 l'indagine è stata estesa (con un questionario ridotto) anche alle imprese con 20-49 addetti. Dal 2002 nella popolazione di riferimento sono state inserite anche le imprese dei servizi privati non finanziari con 20 addetti e oltre (escludendo dai servizi destinabili alla vendita, le imprese del credito e assicurazioni, i servizi pubblici e gli altri servizi sociali e personali). Dal 2006 la rilevazione si è estesa anche al settore delle costruzioni con 20 addetti e oltre. L'indagine sul 2013 ha esteso il campione delle costruzioni alle imprese con 10-19 addetti.

Poiché i questionari rivolti alle imprese di costruzione differiscono significativamente da quelli indirizzati alle imprese dell'industria in senso stretto e dei servizi, anche i rispettivi archivi sono separati: il dataset con le sole imprese dell'industria in senso stretto e dei servizi viene chiamato "indann_completo_csv.csv", mentre quello con le sole imprese delle costruzioni viene chiamato "costr.csv". Per un elenco completo di tutte le variabili rilevate nei singoli anni di indagine e disponibili negli archivi si consultino i file di descrizione variabili disponibili sul sito internet dell'Istituto.

In entrambi i dataset, ogni impresa ha un codice identificativo (la variabile **ident**), che consente l'aggancio delle informazioni nel caso in cui sia stata oggetto di rilevazione in più anni. Questo codice, generato in modo casuale, è del tutto incorrelato con le variabili identificative delle imprese e serve esclusivamente per le analisi longitudinali. La coppia (**annoril, ident**) identifica le informazioni riguardanti una data impresa in un dato anno. A puro scopo esemplificativo, la struttura logica degli archivi (che l'utente non può visualizzare), è quindi la seguente:

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	…	…	…	**…**
…	**…**	**…**	**…**	**…**	**…**	**…**	**…**	**…**

Gli archivi contengono alcune variabili relative al riporto all'universo delle stime campionarie. L'adozione del peso campionario consente di allineare la struttura del campione a quella dell'universo secondo le variabili di stratificazione[1]; se ne raccomanda l'uso nelle analisi per ottenere stime non distorte che riportino alla popolazione di riferimento.

[1] Poiché la numerosità della popolazione è nota con uno o due anni di ritardo, i pesi sono provvisoriamente calcolati dapprima adoperando le informazioni più recenti disponibili e sono poi aggiornati quando i dati sulla popolazione effettiva vengono diffusi. I piccoli scostamenti tra l'effettiva numerosità della popolazione di riferimento e la somma dei pesi sono dovuti al processo di post-stratificazione.

Per le variabili che sono fornite dalle imprese sotto forma di rapporto, o comunque prive di fattore di scala, si dovrebbero in generale adottare pesi che tengano conto anche della dimensione del fenomeno. Gli archivi contengono anche alcune variabili relative alle classificazioni adottate per la stratificazione secondo il disegno campionario. Va tenuto presente che l'area geografica è considerata nel disegno dell'indagine come variabile di post-stratificazione.

Le informazioni monetarie provenienti dal questionario sono espresse a **prezzi correnti e in migliaia di euro** (ad eccezione delle variabili relative alle **retribuzioni**, espresse in **euro**), anche per quelle riferite agli anni in cui non era entrato in vigore l'euro.

Per diverse variabili chiave, quali ad esempio occupazione, investimenti e fatturato, l'indagine rileva valori riferiti a più anni successivi. Ad esempio, nella rilevazione sul 2021, vengono chieste informazioni sull'occupazione media nel 2021, nel 2020 e quella prevista per il 2022 (anno nel corso del quale si svolge l'intervista). La presenza di molteplici orizzonti temporali all'interno del questionario offre la possibilità di effettuare calcoli di variazioni annuali, senza la necessità di ricorrere all'utilizzo congiunto di più anni di indagine. Tale scelta, utile per la stabilizzazione della stima dei tassi di variazione, comporta possibili variazioni nei valori riferiti allo stesso anno ma rilevati in edizioni successive dell'indagine. Per illustrare meglio questo fenomeno si consideri il seguente esempio sull'occupazione media nell'anno precedente, nell'anno in corso e nell'anno successivo (variabili **v15**, **v24**, **v611m**) rilevate per l'impresa fittizia ident=999 nel 2020 e nel 2021.

anno indagine di riferimento	**riferimento temporale del dato**				
	2018	2019	2020	2021	**2022**
2020		120	125	80	
2021			**121**	**100**	**79**

Nella rilevazione sul 2021 l'impresa ha aggiornato sia il dato sull'occupazione media nel 2020 sia le attese sul 2021 fornite l'anno precedente.

In tal caso nel database si osserverà il seguente contenuto:

anno indagine di riferimento	**impresa**	**occupazione media anno t-1**	**occupazione media anno t**	**occupazione media anno t+1 (prev.)**
(annoril)	*(ident)*	*(v15)*	*(v24)*	*(v611m)*
2019	999	115	120	125
2020	999	120	125	80
2021	999	121	100	79

Nel corso dell'intervista, all'impresa rispondente nel caso avesse partecipato anche all'indagine precedente vengono proposti, per memoria, i valori forniti l'anno precedente, tuttavia l'impresa è libera di rivedere i valori già forniti.

Per alcune variabili quantitative, i dati mancanti sono imputati. Contestualmente al processo di imputazione, viene creata una nuova variabile *flag*, che permette di riconoscere se il valore nel database è stato fornito direttamente dall'impresa o è stato imputato: il *flag* assume valore 1 se il dato della relativa variabile di riferimento è frutto del processo di imputazione, altrimenti sono vuoti. Il nome delle variabili *flag* è sempre del tipo **f"X"**, dove X indica il nome della variabile imputata (ad esempio il *flag* d'imputazione della variabile occupazione attesa, **v611m**, sarà **fv611m**).

A partire dall'indagine sul 2010, alcune variabili monografiche, specificamente indicate nel database, sono state rilevate solo su metà del campione; le sezioni dei questionari che le contengono sono chiaramente contrassegnate dalla lettera "A" o "B", per indicare a quale metà del campione esse siano rivolte.

La suddivisione del campione totale in due sotto-campioni è effettuata in base a un meccanismo casuale, tale da mantenere la rappresentatività delle due metà del campione originario rispetto alla popolazione di riferimento. Nelle analisi relative a variabili rilevate su un sotto-campione è importante utilizzare i pesi di riporto appositamente creati a tal fine (si veda descrizione **pesoa** e **pesob** di seguito) per garantire il corretto riporto all'universo anche nel sottoinsieme in esame.

2. Aggiornamento periodico degli archivi

Gli archivi vengono aggiornati al termine di ogni rilevazione sulla base delle risposte fornite dalle imprese nel corso delle interviste. Le risposte sono sottoposte a un processo di controllo qualità che precede la formazione del dataset completo. Le stime pubblicate nel fascicolo della collana Statistiche fanno riferimento agli archivi come disponibili al momento delle elaborazioni e all'universo delle imprese più recente disponibile. Limitati scostamenti tra l'archivio utilizzato per le stime nel fascicolo e quello messo a disposizione per le elaborazioni a distanza possono essere dovuti sia a successive revisioni delle risposte fornite dalle imprese, sia a revisioni nel sistema di pesi dovute all'aggiornamento della popolazione di riferimento.

Il dataset con i dati aggiornati al nuovo anno di riferimento viene messo a disposizione degli utenti in corrispondenza della pubblicazione del relativo fascicolo sulla pagina internet dell'Istituto. La revisione del sistema di ponderazione ha carattere sistematico, mentre eventuali rettifiche operate sulle risposte degli anni precedenti sono da considerarsi del tutto episodiche.

3. Variabili disponibili negli archivi

Per motivi di riservatezza, negli archivi **non** sono disponibili le variabili del questionario che permetterebbero l'identificazione dell'impresa rispondente, solitamente presenti nelle prime pagine dei questionari, tra le quali: Codice Fiscale, Ragione sociale, Filiale di rilevazione e Gruppo di appartenenza. Non sono inoltre rese disponibili le risposte testuali di tipo "Altro specificare". Al contrario, gli archivi contengono alcune variabili non presenti nei questionari ma utili ai fini delle elaborazioni. Tra queste, oltre l'anno di riferimento dell'indagine (variabile **annoril**) ci sono:

a)Variabili di classificazione rispetto al settore di attività economica[2]

Nome	***Valori***	***Descrizione***	***ATECO 2002***	***ATECO 2007***
settor11	*SS1*	Industrie alimentari, bevande e tabacco	*DA*	*10, 11, 12*
	SS2	Industrie tessili, dell'abbigliamento, pelli cuoio e calzature	*DB, DC*	*13, 14, 15*
	SS3	Fabbricazione di coke, industria chimica, gomma e plastica	*DF, DG, DH*	*19, 20, 21, 22*
	SS4	Industria della lavorazione dei minerali non metalliferi	*DI*	*23*
	SS5	Industria metalmeccanica	*DJ, DK, DL, DM*	*24, 25, 26, 27, 28, 29, 30, 33*
	SS6	Altre industrie manifatturiere	*DD, DE, DN*	*16, 17, 18, 31, 32*
	SS7	Altre industrie in senso stretto	*CA, CB, CE*	*05, 06, 07, 08, 09, 35, 36, 37, 38, 39*
	SS8	Commercio ingrosso e dettaglio, riparazione di autoveicoli e motocicli	*G*	*45, 46, 47*
	SS9	Servizi di alloggio e ristorazione	*H*	*55, 56*
	SS10	Trasporto e magazzinaggio, servizi di informazione e comunicazione	*I*	*49, 50, 51, 52, 53, 58, 59, 60, 61, 62, 63*
	SS11	Attività immobiliari, professionali, scientifiche e tecniche, amministrative e di servizi di supporto	*K*	*68, 69, 70, 71, 72, 73, 74, 75, 77, 78, 79, 80, 81, 82*
indag3	*1*	Industria manifatturiera	*D*	*C*
	2	Industria energetico – estrattiva	*C, E*	*B, D, E*
	3	Servizi	*G, I, H, K*	*G, I, H, J, L, M, N*
indagine	*1*	Industria in senso stretto	*C, D, E*	*C, B, D, E*
	2	Servizi	*G, I, H, K*	*G, I, H, J, L, M, N*

[2] Fino al 2009 tali variabili erano ottenute dall'aggregazione di alcune sottosezioni ATECO 2002; dal 2010 sono basate sulle prime due cifre della classificazione ATECO 2007. Le variabili di questa tabella sono presenti nel solo archivio "indann_completo_csv.csv", tutte le imprese dell'archivio costruzioni hanno un unico codice ateco (2002, 2007).

b)Variabili di classificazione rispetto alla classe dimensionale[3]

Nome	Valori	Descrizione
cldimet	0	20 - 49 addetti
	1	50 - 99 addetti
	2	100 - 199 addetti
	3	200 - 499 addetti
	4	500 - 999 addetti
	5	1.000 addetti e oltre
cc2	1	20 - 49 addetti
	2	50 addetti e oltre

c)Variabili di classificazione rispetto all'area geografica[4]

Nome	Valori	Descrizione
areag4	1	Nord ovest
	2	Nord est
	3	Centro
	4	Sud e Isole
areag2	1	Nord, Centro
	2	Sud e isole

d)Variabili relative al disegno campionario e al sistema di ponderazione[5]

strato: Formato dalle 66 combinazioni di **settor11** e **cldimet** alle quali sono aggiunti due strati riferiti alle imprese con almeno 5.000 addetti, che hanno peso unitario e sono considerate, separatamente per industria e servizi (strati 67 e 68).

poststrato: Formato dalle 48 combinazioni di **areag4, cc2** e una riaggregazione dei settori di attività economica in 6 gruppi: 1) indag3=1; 2) indag3=2; 3) settor11=SS8; 4) settor11=SS9; 5) settor11=SS10; 6) settor11=SS11.

peso: peso di espansione all'universo: a livello di strato e post-strato, la somma dei pesi uguaglia la numerosità della popolazione di riferimento, separatamente per ciascun anno senza tener conto della dimensione longitudinale del campione. Alle imprese con più di 5000 addetti (c.d. auto-rappresentative) e ad un numero limitato di imprese che non si ritengono rappresentative dello strato a cui appartengono si attribuisce peso unitario.

pesoadd: Peso campionario (di espansione): a livello di strato e poststrato, la somma dei pesi equivale alla numerosità degli addetti della popolazione di riferimento separatamente per ciascun anno e non tiene conto della dimensione longitudinale del campione (disponibile dal 2007). Questo peso è particolarmente indicato per la ponderazione delle variabili di tipo categorico, perché consente di tener conto della diversa scala dimensionale delle imprese.

pesoa: Equivalente a **peso** per le imprese appartenenti alla prima metà del campione (presente dal 2010, da usare per i quesiti monografici rilevati sulla prima metà del campione).

pesoadda: Equivalente a **pesoadd** per le imprese appartenenti alla prima metà del campione (presente dal 2010, da usare per i quesiti monografici rilevati sulla prima metà del campione).

pesob: Equivalente a **peso** per le imprese appartenenti alla seconda metà del campione (presente dal 2010, da usare per i quesiti monografici rilevati sulla seconda metà del campione).

pesoaddb: Equivalente a **pesoadd** per le imprese appartenenti alla seconda metà del campione (presente dal 2010, da usare per i quesiti monografici rilevati sulla seconda metà del campione).

pesorisc: La variabile **pesorisc** è ottenuta come prodotto di **peso** e un opportuno fattore di scala, in modo che essa, anno per anno, sommi alla numerosità del campione.

popstr: Numerosità della popolazione a livello di strato.

poppost: Numerosità della popolazione a livello di post-strato.

[3] Fino all'anno di riferimento 2003 la classe dimensionale è riferita al numero di addetti a fine anno; dal 2004 in poi al numero di addetti medi nell'anno.

[4] Per motivi di riservatezza non sono disponibili le classificazioni relative alle singole regioni e province, ma solo quelle relative alle macro-aree geografiche.

[5] Per maggiori dettagli sul disegno campionario, la costruzione dei pesi, dei deflatori e per tutti gli altri aspetti riguardanti la metodologia utilizzata, si rimanda alla Nota metodologica reperibile sul sito internet della Banca d'Italia.

e) Variabili di classificazione della quota del fatturato esportato

Nome	Valori	Descrizione
a6	0	impresa non esportatrice
	1	meno di 1/3 di fatturato esportato
	2	tra 1/3 e 2/3 di fatturato esportato
	3	oltre 2/3 di fatturato esportato
qexp	1	meno di 1/3 di fatturato esportato o nessuna esportazione
	2	tra 1/3 e 2/3 di fatturato esportato
	3	oltre 2/3 di fatturato esportato

f) Variabili i cui livelli sono disponibili a prezzi correnti e costanti

Negli archivi, limitatamente ai livelli di investimenti e fatturato, sono presenti variabili espresse sia a prezzi correnti che a prezzi costanti. I prezzi costanti sono riferiti sia all'anno di riferimento più recente disponibile sia all'anno di riferimento di ogni singola indagine. Questi ultimi consentono di calcolare variazioni a prezzi costanti nell'anno in cui sono stati rilevati anche dopo l'aggiunta di nuove edizioni dell'indagine. Per le imprese dell'Industria in senso stretto e dei servizi, i deflatori sono ottenuti partendo da quelli forniti dalle stesse imprese e utilizzando una metodologia di aggregazione a livello di sottosezione e classe dimensionale.

Descrizione	Variabili a prezzi correnti	Variabili a prezzi costanti…	
		…dell'anno più recente disponibile	…dell'anno di riferimento della rilevazione
Investimenti fissi t-1	v200	v200cos[(a)]	v200k[(a)]
Investimenti fissi t	v202	v202cos	v202k
Investimenti fissi t+1	v203	v203cos	v203k
Fatturato t-1	v209	v209cos	v209k
Fatturato t	v210	v210cos	v210k
Fatturato t+1	v437	v437cos	v437k
Fatturato esportato t-1	v211	v211cos	v211k
Fatturato esportato t	v212	v212cos	v212k
Fatturato esportato t+1	v438	v438cos	v438k
Investimenti immateriali[(b)] t-1	v810	v810cos	v810k
Investimenti immateriali[(b)] t	v811	v811cos	v811k
Investimenti immateriali[(b)] t+1	v812	v812cos	v812k

Note: (a) Disponibile dal 1985; (b) Secondo il SEC2010, la voce "investimenti materiali" comprende la spesa per software, basi di dati e prospezioni minerarie, mentre sono esclusi i brevetti e i marchi.

4. L'Indagine straordinaria sugli effetti del Coronavirus

Tra il 16 marzo e il 14 maggio del 2020, la Banca d'Italia ha condotto un'indagine straordinaria su 3.503 imprese (2.391 dell'industria in senso stretto e 1.112 dei servizi privati non finanziari) con la finalità di ottenere informazioni tempestive sulle ricadute economiche derivanti dalla diffusione dell'epidemia di Covid-19 in corso in Italia al momento della rilevazione.

Il disegno campionario, il campione di riferimento, il sistema di ponderazione e le modalità di risposta al questionario coincidono con quelli dell'Indagine sulle imprese industriali e dei servizi (Invind) sul 2019. Le indagini Iseco e Invind si sono quindi svolte con le stesse modalità di rilevazione, tuttavia l'Indagine Iseco è stata predisposta e avviata in un secondo momento rispetto a Invind. Pertanto, il momento della compilazione delle due rilevazioni da parte delle imprese non ha necessariamente coinciso, nemmeno successivamente all'avvio dell'indagine Iseco, implicando possibili differenze nelle attese a breve termine espresse da una stessa impresa. La diversa numerosità dei campioni delle due indagini è invece spiegata dalla non obbligatorietà delle rilevazioni, che avrebbe portato alcune imprese a rispondere solo ad uno dei due questionari.

L'elenco completo delle variabili rilevate in Iseco e le relative modalità di risposta sono disponibili nel file di descrizione variabili sul sito internet dell'Istituto.

5. Esempi di utilizzo degli archivi

5.1. Esempi basati sul software R[6]

Per ottenere più rapidamente i risultati delle proprie elaborazioni si suggerisce di limitare il numero di variabili incluse nei dataset utilizzati per le stime. Si ricorda che R è un linguaggio case-sensitive.

Negli esempi che seguono viene importato il file denominato **indann_completo_csv.csv**, ovvero quello relativo alle imprese dell'industria in senso stretto e dei servizi che non include le imprese di costruzione. 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 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,
           weights = peso, data = oggetto, family = "quasibinomial")
summary(fit)
```

Esempio 2: distribuzioni di frequenza

- Calcoliamo 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** ha il solo scopo di ottenere stime riferite a una variazione percentuale). L'analisi è limitata alle sole imprese industriali (indagine==1).

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


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

##lettura dei dati
```

[6] 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/ .

```
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
out_by_ratio <- svyby(~var_occ,by = ~areag4,
                      denominator = ~v15,
                      design=out_svy,
                      svyratio)
out_by_ratio

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

Esempio 3: regressione lineare

- Stimiamo un modello di regressione lineare dove il numero di addetti (variabile **v24**) è la variabile dipendente e le covariate sono il fatturato (variabile **v210**) e l'area geografica (**areag4**) 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)


##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)
summary(out_reg)
```

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)


##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)
summary(out_reg)
```

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*.

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


##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")
summary(out_random)
```

5.2. Esempi basati sul software Stata[7]

Negli esempi che seguono verrà importato il file CSV contenente i dati dell'indagine. Si mostra inoltre come limitare l'analisi a un solo settore (ad esempio, il settore industriale, indagine==1) o a un solo anno (ad esempio al 2021, annoril==2021). I primi cinque esempi quindi presenta delle elaborazioni sulle sole imprese industriali per l'anno 2021, mentre il sesto mostra un'analisi panel.

Tutti i comandi vanno scritti in minuscolo, poiché anche Stata è case-sensitive.

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

[7] Stata è un marchio registrato della StataCorp LP, 4905 Lakeway Drive, College Station, TX 77845 USA.

```
#delimit;

import delimited “indann_completo_csv.csv”;

keep annoril indagine peso areag4 settor11 v521 v24;

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

/* creo le dummy per l’area geografica e il settore di attività economica */

tabulate areag4, gen(areag4d);

tabulate settor11, gen(settor11d);

/* sono create in questo modo 4 dummy di area geografica e 7 di settore */

/* stimo il logit, in cui ometto una dummy ciascuna per area e settore, che funge da ri
ferimento per le altre */

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

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).

```
#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);
```

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 sul fascicolo.

```
#delimit;
```

```
import delimited “indann_completo_csv.csv”;

import delimited “indann_completo_csv.csv”;

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

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

/* creo la variabile investimenti totali a prezzi costanti per il 2004 */

generate i0tot=v200cos+v810cos;

/* creo la variabile investimenti totali a prezzi costanti per il 2021 */

generate i1tot=v202cos+v811cos;

/* procedimento di winsorizzazione del secondo tipo (sulla base del 5°e 95° percentile*
/

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;

/* creo una nuova variabile i1totw contenente gli investimenti totali 2021 che attenua
l’effetto dei dati anomali */

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);
```

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*.

```
#delimit;

import delimited “indann_completo_csv.csv”;

keep annoril indagine peso areag4 v210 v24;

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

/* creo le dummy per l’area geografica */

tabulate areag4, gen(areag4d);

/* sono create in questo modo 4 dummy di area geografica */

/* stimo la regressione, in cui ometto una dummy per l’area, che funge da riferimento p
er le altre */

regress v24 v210 areag4d1 areag4d2 areag4d3 [pweight=peso];
```

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.

```
#delimit;

import delimited “indann_completo_csv.csv”;

keep annoril indagine peso areag4 v210 v24;

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

/* creo le dummy per l’area geografica */

tabulate areag4, gen(areag4d);

/* sono create in questo modo 4 dummy di area geografica */

/* creo le due variabili pc1_v24 e pc99_v24 contenenti il primo e 99-esimo percentile d
ella variabile v24 */

egen pc1_v24=pctile(v24), p(1);

egen pc99_v24=pctile(v24), p(99);

/* stimo la regressione, in cui ometto una dummy per l’area, che funge da riferimento p
er le altre e escludo dalla regressione le unità con v24 all’esterno dei percentili */

regress v24 v210 areag4d1 areag4d2 areag4d3 [pweight=peso]

if v24>=pc1_v24 & v24<=pc99_v24;
```

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 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*.

```
#delimit;

import delimited “indann_completo_csv.csv”;
```

```
keep annoril indagine ident areag4 v545 v210 v24;

keep if inrange(annoril, 2016, 2021) & indagine == 1;

/* seleziono le sole imprese presenti nei 6 anni dal 2016 al 2021 */

generate one=1;

sort ident;

by ident: egen conta=sum(one);

keep if conta==6;

/* creo le dummy per il risultato di esercizio */

tabulate v545, gen(v545d);

/* stimo il modello di regressione sul panel, in cui ometto una dummy per il risultato
di esercizio, che funge da riferimento per le altre */

iis ident;

tis annoril;

xtreg v210 v24 v545d1 v545d2 v545d3 v545d4,re;
```