Econometria | 2023/2024

Lezione 12: Panel Data

Giuseppe Ragusa
https://gragusa.org

Roma, aprile 2024

Sommario

Dati panel: cosa e perché (Paragrafo 10.1)

Un panel contiene osservazioni su più unità (individui, stati, imprese) in cui ogni entità è osservata in due o più istanti temporali diversi.

Esempi:

– Dati su 50 stati USA, ognuno è osservato per 3 anni, per un totale di 150 osservazioni.

– Dati su 1000 individuali, in quattro mesi diversi, per 4000 osservazioni in totale.

Notazione per dati panel

Un doppio pedice distingue unità (stati) e periodi temporali (anni)

- $i = 1, \ldots, n$

  i= unità (stati); n= numero di entità
- $t = 1, . . , T$

  t= periodo temporale (anno); T= numero di periodi temporali

Con un regressore i dati sono:

$$(X_{it}, Y_{it}),\ i = 1, \ldots, n,\ t = 1, \ldots, T$$

Notazione per dati panel (continua)

Dati panel con K regressori:

$$(X_{1it}, \ldots, X_{Kit}, Y_{it}),\ i = 1, \ldots, n,\ t = 1, \ldots, T$$

Un po' di gergo

- I dati panel sono chiamati anche dati longitudinali
- panel bilanciato: non ci sono osservazioni che mancano, cioè tutte le variabili sono osservate per tutte le unità (stati) e tutti i periodi temporali (anni)

Perché sono utili i dati panel?

Con i dati panel possiamo controllare per fattori che:

- Variano tra le unità ma non nel tempo
- Potrebbero causare distorsione da variabili omesse se fossero omessi
- Sono inosservati o non misurati, e perciò non possono essere inclusi in una regressione multipla

Idea chiave

Se una variabile omessa non varia nel tempo, allora qualsiasi variazione in Y nel tempo non può essere causata dalla variabile omessa.

Esempio di dati panel: morti sulle strade e imposte sugli alcolici

Unità di osservazione: un anno in uno stato USA

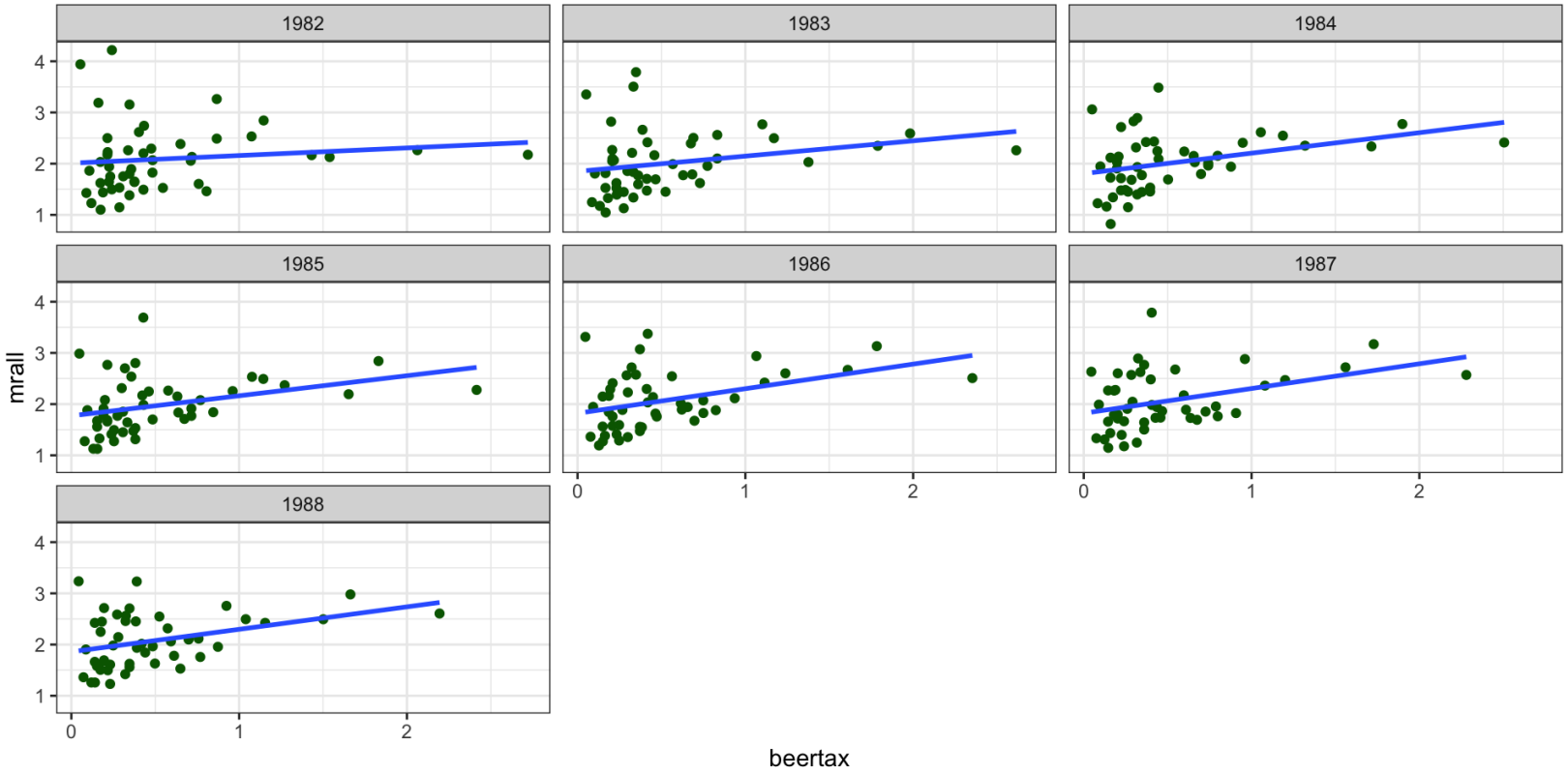
- 48 stati USA, perciò n= numero di unità = 48
- 7 anni (1982,…, 1988), perciò T= numeri di periodi temporali = 7
- Panel bilanciato, perciò numero totale di osservazioni = 7×48 = 336

Variabili:

- Tasso di mortalità stradale — morti per 10.000 residenti nello stato (`mrall`)
- Imposta su una cassa di birra (`beertax`)
- Altre (età minima per guidare, leggi sulla guida in stato di ebbrezza, ecc.)

Mortalità e tassazione alcoloci (per anno)

```
1 library(Ecdat)
2 library(ggplot2)
3 data(Fatality)
4 ggplot(Fatality, aes(y=mrall, x=beertax)) +
5   geom_point(color = "darkgreen") +
6   geom_smooth(method="lm", se = FALSE) +
7   facet_wrap(~year) + theme_bw()
```



Perché correlazione positiva frac morti e tassazione alcol

Altri fattori che influenzano il tasso di mortalità stradale:

- Qualità (età) delle automobili
- Qualità delle strade
- “Cultura” sul bere e guidare
- Densità di auto sulle strade

Questi fattori omessi potrebbero causare distorsione da variabili omesse.

Esempio 1: densità del traffico

1. Elevata densità del traffico significa più morti sulle strade
2. Gli stati con minore densità di traffico (all'ovest) hanno imposte sugli alcolici minori

- Allora le due condizioni per la distorsione da variabili omesse sono soddisfatte. Nello specifico, "imposte elevate" potrebbero riflettere "alta densità di traffico" (perciò il coefficiente OLS sarebbe distorto positivamente - imposte elevate, più morti)
- I dati panel ci consentono di eliminare la distorsione da variabili omesse quando le variabili omesse sono costanti nel tempo in un dato stato.

Esempio 2: attitudini culturali verso il bere e la guida.

1. Sono presumibilmente un determinante della mortalità stradale; e
2. Sono potenzialmente correlate con le imposte sulla birra.

- Allora le due condizioni per la distorsione da variabili omesse sono soddisfatte. Nello specifico, “alte imposte” potrebbe captare l’effetto di “attitudini culturali verso il bere”, perciò il coefficiente OLS sarebbe distorto.
- I dati panel ci consentono di eliminare la distorsione da variabili omesse quando le variabili omesse sono costanti nel tempo in un dato stato.

Dati panel con due periodi temporali (Paragrafo 10.2)

Consideriamo il modello dei dati panel:

$$FatalityRate_{it} = \beta_0 + \beta_1 BeerTax_{it} + \beta_2 Z_i + u_{it}$$

Z_i è un fattore che non cambia nel tempo (e.g., densità) almeno durante gli anni per cui abbiamo dati.

- Supponiamo che Z_i non sia osservato — la sua omissione potrebbe comportare distorsione da variabili omesse.
- L'effetto di Z_i può essere eliminato usando T=2 anni.

L’idea chiave:

Qualsiasi variazione nel tasso di mortalità dal 1982 al 1988 non può essere causata da Z_i, perché Z_i (per ipotesi) non varia tra il 1982 e il 1988.

Consideriamo i tassi di mortalità nel 1988 e nel 1982:

$$FatalityRate_{i1988} = \beta_0 + \beta_1 BeerTax_{i1988} + \beta_2 Z_i + u_{i1988}$$
$$FatalityRate_{i1982} = \beta_0 + \beta_1 BeerTax_{i1982} + \beta_2 Z_i + u_{i1982}$$

Supponiamo $E(u_{it}|BeerTax_{i1988}, BeerTax_{i1982}, Z_i) = 0$.

Sottraendo 1988-1982 (ovvero calcolando la variazione) si elimina l’effetto di Z_i.

$$FatalityRate_{i1988} - FatalityRate_{i1982} = \beta_1(BeerTax_{i1988} - BeerTax_{i1982}) + (u_{i1988} - u_{i19}$$

$$FatalityRate_{i1988} - FatalityRate_{i1982} = \beta_1(BeerTax_{i1988} - BeerTax_{i1982}) + (u_{i1988} - u_{i19}$$

- Il nuovo termine d'errore, $(u_{i1988} - u_{i1982})$, non è correlato con $BeerTax_{i1988}$ o $BeerTax_{i1982}$.
- Questa equazione "alle differenze" può essere stimata con OLS, anche se Z_i non è osservata.
- La variabile omessa Z_i non cambia, perciò non può essere una determinante della variazione in Y.
- Questa regressione alle differenze non ha un'intercetta, che è stata eliminata dalla sottrazione

Esempio: mortalità stradale e imposte sulla birra

Dati del 1982 (n=48):

```
1 lm(mrall~beertax,
2         data=Fatality |> filter(year==1982))
```

```
Call:
lm(formula = mrall ~ beertax, data = filter(Fatality,
year ==
    1982))

Coefficients:
(Intercept)      beertax
     2.0104       0.1485
```

Dati del 1982 (n=48):

```
1 lm(mrall~beertax,
2         data=Fatality |> filter(year==1988))
```

```
Call:
lm(formula = mrall ~ beertax, data = filter(Fatality,
year ==
    1988))

Coefficients:
(Intercept)      beertax
     1.8591       0.4388
```

Regressione differenze $(n = 48)$

```
1 DFatality <- Fatality |>
2               arrange(state, year) |>
3               group_by(state) |>
4               filter(year==1982 | year==1988) |>
5               mutate(Dmrall = diff(mrall), Dbeertax=diff(beertax)) |>
6               filter(year==1988)
7 feols(Dmrall~Dbeertax-1, data=DFatality, vcov = "hetero")
```

```
OLS estimation, Dep. Var.: Dmrall
Observations: 48
Standard-errors: Heteroskedasticity-robust
          Estimate Std. Error  t value  Pr(>|t|)
Dbeertax -0.868922   0.268703 -3.23376 0.0022375 **
```

Se $T > 2$?

Possiamo calcolare le differenze per ciascun periodo e calcolare la regressione

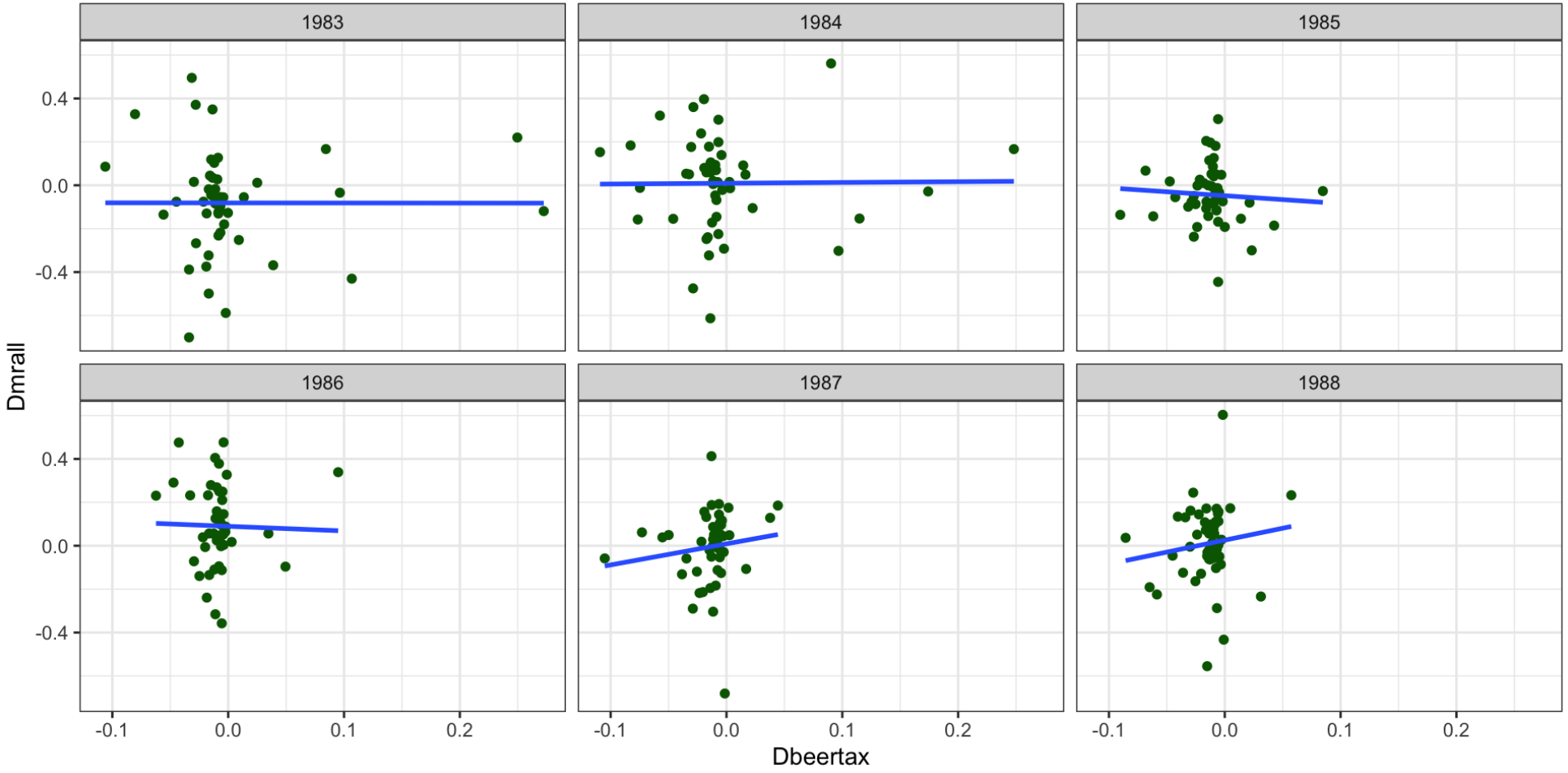
```r
DFatality <- Fatality |>
  arrange(state, year) |>
  group_by(state) |>
  mutate(Dmrall = mrall-lag(mrall), Dbeertax=beertax-lag(beertax)) |>
filter(year!=1982)   ## Rimuovi primo periodo

feols(Dmrall~Dbeertax-1, data=DFatality, vcov = "hetero")
```

```
OLS estimation, Dep. Var.: Dmrall
Observations: 288
Standard-errors: Heteroskedasticity-robust
         Estimate Std. Error  t value Pr(>|t|)
Dbeertax 0.028816   0.279001 0.103283  0.91781
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
RMSE: 0.197339   Adj. R2: -0.00372
```

La stima di β_1 e' imprecisa (vedremo perche')

```r
ggplot(DFatality, aes(y=Dmrall, x=Dbeertax)) +
  geom_point(color = "darkgreen") +
  geom_smooth(method="lm", se = FALSE) +
  facet_wrap(~year) + theme_bw()
```



Regressione con effetti fissi (Paragrafo 10.3)

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \beta_2 Z_i + u_{it}, \quad i = 1, \ldots, n, \quad t = 1, \ldots, T$$

1. modello di regressione “con effetti fissi”.
2. modello di regressione “$n - 1$ regressori binari”

Supponiamo di avere n=3 stati: California, Texas e Massachusetts.

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \beta_2 Z_i + u_{it}, \quad i = 1, \ldots, n, \quad t = 1, \ldots, T$$

Regressione per la California (i= CA):

$$\begin{aligned} Y_{CA,t} &= \beta_0 + \beta_1 X_{CA,t} + \beta_2 Z_{CA} + u_{CA,t} \\ &= (\beta_0 + \beta_2 Z_{CA}) + \beta_1 X_{CA,t} + u_{CA,t} \end{aligned}$$

Riscriviamo:

$$Y_{CA,t} = \alpha_{CA} + \beta_1 X_{CA,t} + u_{CA,t}$$

- $\alpha_{CA} = \beta_0 + \beta_2 Z_{CA}$ non cambia nel tempo
- α_{CA} è l'intercetta per CA, e β_1 è la pendenza
- L'intercetta è specifica per CA, ma la pendenza è la stessa in tutti gli stati: rette parallele.

Per Texas

$$\begin{aligned} Y_{TX,t} =& \beta_0 + \beta_1 X_{TX,t} + \beta_2 Z_{TX} + u_{TX,t} \\ =& (\beta_0 + \beta_2 Z_{TX}) + \beta_1 X_{TX,t} + u_{TX,t} \end{aligned}$$

Riscriviamo:

$$Y_{TX,t} = \alpha_{TX} + \beta_1 X_{TX,t} + u_{TX,t}$$

dove $\alpha_{TX} = \beta_0 + \beta_2 Z_{TX}$.

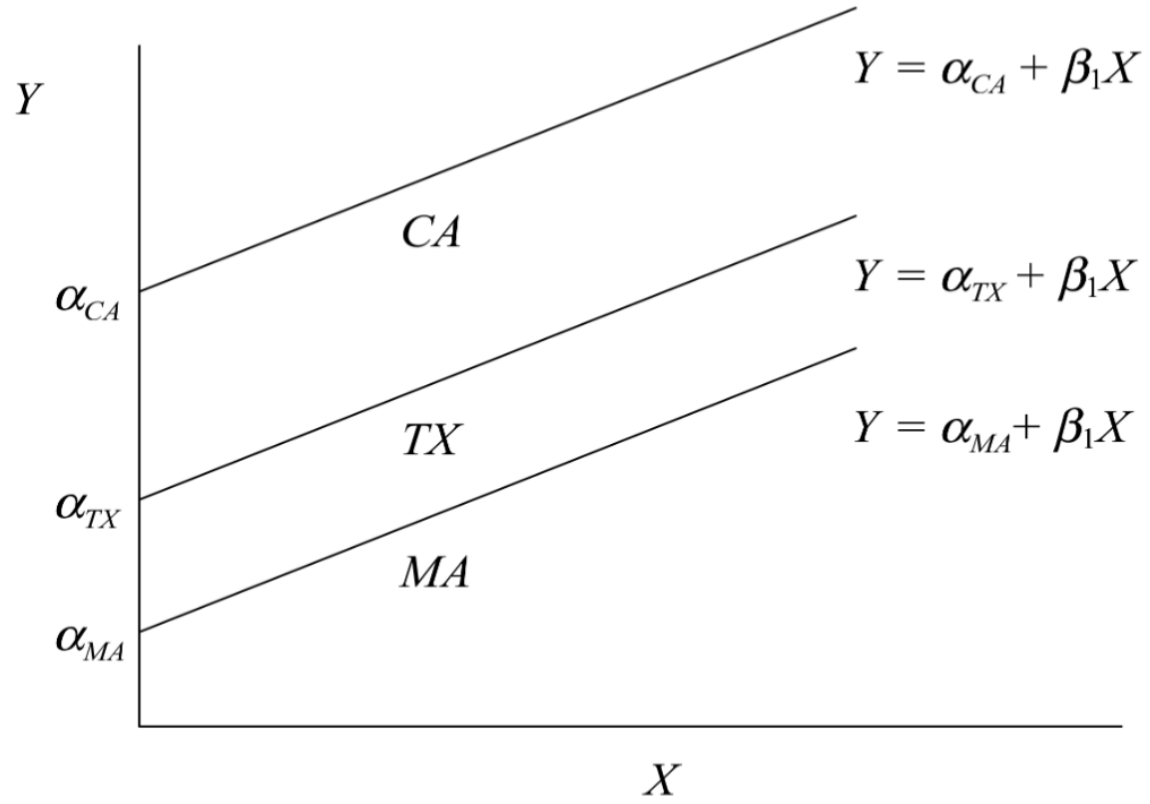
Mettendo insieme le rette dei tre stati (anche Massachusetts):

$$\begin{aligned} Y_{CA,t} &= \alpha_{CA} + \beta_1 X_{CA,t} + u_{CA,t} \\ Y_{TX,t} &= \alpha_{TX} + \beta_1 X_{TX,t} + u_{TX,t} \\ Y_{MA,t} &= \alpha_{MA} + \beta_1 X_{MA,t} + u_{MA,t} \end{aligned}$$

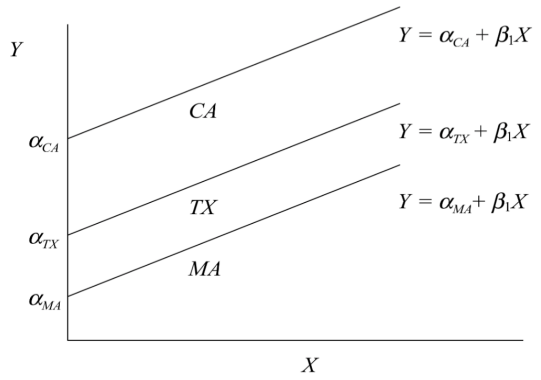
o

$$Y_{it} = \alpha_i + \beta_1 X_{it} + u_{it}, \quad i = CA, TX, MA, \quad t = 1, \ldots, T$$

Le rette di regressione per ciascuno stato



Si ricordi che gli spostamenti nell'intercetta possono essere rappresentati mediante regressori binari…



Nella forma con regressori binari:

$$Y_{it} = \beta_0 + \gamma_{CA} DCA_i + \gamma_{TX} DTX_i + \beta_1 X_{it} + u_{it}$$

$$DCA_i = \begin{cases} 1 & i = CA \\ 0 & altrimenti \end{cases}$$

$$DTX_i = \begin{cases} 1 & i = TX \\ 0 & altrimenti \end{cases}$$

- si lascia fuori DMA_i (perché?)

Riepilogo

1. Forma con "$n-1$ regressori binari"

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \gamma_2 D2_i + \ldots + \gamma_n Dn_i + u_{it}$$

dove:

$$D2_i \begin{cases} 1 & i = 2 \\ 0 & altrimenti \end{cases}$$

2. Forma con "effetti fissi"

$$Y_{it} = \beta_1 X_{it} + \alpha_i + u_{it}$$

dove: α_i e' "effetto fisso"

Regressione OLS con “n-1 regressori binari”

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \gamma_2 D2_i + \ldots + \gamma_n Dn_i + u_{it} \quad (1)$$

1. si creano le variabili binarie $D2_i, \ldots, Dn_i$
2. si stima (1) mediante OLS
3. L’inferenza (verifiche di ipotesi, intervalli di confidenza) è come di consueto (con errori standard robusti all’eteroschedasticità)

- Non è pratico quando n è molto grande (per esempio se n= 1000 lavoratori)

2. Regressione OLS con “unità in deviazioni dalle medie”

Modello di regressione con effetti fissi:

$$Y_{it} = \alpha_i + \beta_1 X_{it} + u_{it}$$

Le medie delle unità soddisfano:

$$\frac{1}{T}\sum_{t=1}^{T} Y_{it} = \alpha_i + \beta_1 \frac{1}{T}\sum_{t=1}^{T} X_{it} + \frac{1}{T}\sum_{t=1}^{T} u_{it}$$

Deviazioni dalle medie:

$$Y_{it} - \frac{1}{T}\sum_{t=1}^{T} Y_{it} = \beta_1 \left(X_{it} - \frac{1}{T}\sum_{t=1}^{T} X_{it} \right) + \left(u_{it} - \frac{1}{T}\sum_{t=1}^{T} u_{it} \right)$$

2. Regressione OLS con “unità in deviazioni dalle medie” (continua)

$$Y_{it} - \frac{1}{T}\sum_{t=1}^{T} Y_{it} = \beta_1 \left(X_{it} - \frac{1}{T}\sum_{t=1}^{T} X_{it} \right) + \left(u_{it} - \frac{1}{T}\sum_{t=1}^{T} u_{it} \right)$$

Riscriviamo:

$$\tilde{Y}_{it} = \beta_1 \tilde{X}_{it} + \tilde{u}_{it}$$

dove

$$\tilde{Y}_{it} = Y_{it} - \frac{1}{T}\sum_{t=1}^{T} Y_{it}$$

e

$$\tilde{X}_{it} = X_{it} - \frac{1}{T}\sum_{t=1}^{T} X_{it}.$$

2. Regressione OLS con “unità in deviazioni dalle medie” (continua)

$$\tilde{Y}_{it} = \beta_1 \tilde{X}_{it} + \tilde{u}_{it} \quad (2)$$

dove $\tilde{Y}_{it} = Y_{it} - \frac{1}{T}\sum_{t=1}^{T} Y_{it}$ e $\tilde{X}_{it} = X_{it} - \frac{1}{T}\sum_{t=1}^{T} X_{it}$.

- si costruiscono le unità in deviazioni dalle medie $\tilde{Y}_{it}$ e $\tilde{X}_{it}$.
- si stima (2) con la regressione di $\tilde{Y}_{it}$ su $\tilde{X}_{it}$ usando OLS.
- È simile all’approccio “prima e dopo”, ma con Y_{it} deviato dalla media al posto di Y_{i1}.
- Gli errori standard vanno calcolati in un modo che tenga conto della natura “panel” dei dati (ne parleremo più avanti).
- Si può fare con un unico comando in R.

Esempio (continua)

Per $n = 48, T = 7$:

```r
library(plm)
plm(mrall~beertax, data=Fatality, model="within")
```

```
Model Formula: mrall ~ beertax

Coefficients:
 beertax
-0.65587
```

```r
feols(mrall~beertax|state, data=Fatality, vcov = "hetero")
```

```
OLS estimation, Dep. Var.: mrall
Observations: 336
Fixed-effects: state: 48
Standard-errors: Heteroskedasticity-robust
         Estimate Std. Error  t value  Pr(>|t|)
beertax -0.655874    0.20328 -3.22646 0.0013984 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
RMSE: 0.17547     Adj. R2: 0.889129
                Within R2: 0.040745
```

Regressione con effetti temporali (Paragrafo 10.4)

Una variabile omessa potrebbe variare nel tempo ma non tra gli stati:

- auto più sicure (air bag, ecc.); modifiche nelle leggi nazionali
- producono intercette che variano nel tempo
- Sia S_t l'effetto combinato di variabili che cambiano nel tempo ma non tra gli stati ("auto più sicure").
- Il modello di regressione risultante è:

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \beta_2 Z_i + \beta_3 S_t + u_{it}$$

Soli effetti fissi temporali

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \beta_3 S_t + u_{it}$$

Questo modello può essere ricomposto con un'intercetta che varia da un anno al successivo:

$$\begin{aligned} Y_{i,1982} &= \beta_0 + \beta_1 X_{i,1982} + \beta_2 S_{1982} + u_{i,1982} \\ &= (\beta_0 + \beta_3 S_{1982}) + \beta_1 X_{i,1982} + u_{i,1982} \\ &= \lambda_{1982} + \beta_1 X_{i,1982} + u_{i,1982} \end{aligned}$$

dove $\lambda_{1982} = \beta_0 + \beta_3 S_{1982}$.

Similmente:

$$Y_{i,1983} = \lambda_{1983} + \beta_1 X_{i,1983} + u_{i,1983}$$

dove $\lambda_{1983} = \beta_0 + \beta_3 S_{1983}$, ecc.

Due formulazioni di regressione con effetti temporali

1. Formulazione con "$T-1$ regressori binari":

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \delta_2 B2_i + \ldots + \delta n BT_i + u_{it} \quad (1)$$

dove:

$$B2_t = \begin{cases} 1 & \text{per } t = 2, \\ 0 & \text{altrimenti.} \end{cases}$$

2. Formulazione con "effetti fissi temporali":

$$Y_{it} = \beta_1 X_{it} + \lambda_t + u_{it}$$

Effetti temporali: metodi di stima

1. Regressione OLS con "$T-1$ regressori binari"

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \delta_2 B2_i + \ldots + \delta n BT_i + u_{it} \quad (1)$$

- Si creano variabili binarie $B2, \ldots, BT$
- $B2 = 1$ se t= anno 2, = 0 altrimenti
- Si esegue la regressione di Y su $X, B2, \ldots, BT$ con OLS
- Dov'è $B1$?

2. Regressione OLS "in deviazione dalle medie dell'anno"

- Si devia Y_{it}, X_{it} dalle medie dell'anno (non dello stato)
- Si stima con OLS usando dati "in deviazione dalle medie dell'anno"

Stima con effetti fissi ed effetti temporali

$$Y_{it} = \beta_1 X_{it} + \alpha_i + \lambda_t + u_{it}$$

- Quando $T = 2$, calcolare la differenza prima ed includere l'intercetta è equivalente a (fornisce esattamente la stessa regressione di) includere effetti individuali e temporali.
- Quando $T > 2$, esistono vari modi equivalenti di incorporare effetti individuali e temporali:
  - deviazione dalle medie e $T - 1$ indicatori temporali
  - deviazione dalle medie temporali e $n - 1$ indicatori individuali
  - T–1 indicatori temporali e $n - 1$ indicatori individuali
  - deviazione dalle medie individuali e temporali

Esempio (continua).

Per $n = 48, T = 7$:

```
## Stima della regressione usando le dummy tutto il campione
## equivalente a lm_robust(mrall ~ beertax, factor(state) + factor(year), data = Fatalities)
feols(mrall~beertax|state+year, data=Fatality, vcov = "hetero")
```

```
OLS estimation, Dep. Var.: mrall
Observations: 336
Fixed-effects: state: 48,  year: 7
Standard-errors: Heteroskedasticity-robust
        Estimate Std. Error  t value Pr(>|t|)
beertax -0.63998   0.254715 -2.51253 0.012547 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
RMSE: 0.171819     Adj. R2: 0.891425
                 Within R2: 0.036065
```

Esempio (continua).

Per $n = 48, T = 7$:

```r
## Stima della regressione usando trasformazioni
plm(mrall~beertax, data=Fatality, model = "within", effect="twoway")
```

```
Model Formula: mrall ~ beertax

Coefficients:
 beertax
-0.63998
```

Le assunzioni e gli errori standard della regressione con effetti fissi (Paragrafo 10.5 e Appendice 10.2)

- Sotto le assunzioni dei minimi quadrati nella versione per dati panel, lo stimatore OLS con effetti fissi di β_1 ha distribuzione normale.
- Tuttavia, è necessario introdurre una nuova formula dell'errore standard, quella per dati raggruppati, o "clustered".
- Questa nuova formula è necessaria perché le osservazioni per la stessa unità non sono indipendenti (è la stessa unità!), anche se le osservazioni di unità diverse sono indipendenti se tali unità sono ottenute mediante campionamento casuale semplice.

Qui consideriamo il caso di effetti fissi individuali. Gli effetti temporali possono semplicemente essere inclusi quali regressori binari aggiuntivi.

Assunzioni dei minimi quadrati per dati panel

Si consideri una singola X:

$$Y_{it} = \beta_1 X_{it} + \alpha_i + u_{it}, \quad i = 1, \ldots, n, \quad t = 1, \ldots, T$$

1. $E(u_{it}|X_{i1}, \ldots, X_{iT}, \alpha_i) = 0$.
2. $(X_{i1}, \ldots, X_{iT}, u_{i1}, \ldots, u_{iT}),\ i = 1, \ldots, n$ sono i.i.d. dalla distribuzione congiunta.
3. (X_{it}, u_{it}) hanno momenti quarti finiti.
4. Non vi è collinearità perfetta (molteplicità di X).

Le assunzioni 3. e 4. sono identiche al caso dei minimi quadrati, le assunzioni 1. e 2. sono diverse.

Assunzione 1:

$$E(u_{it}|X_{i1}, \ldots, X_{iT}, \alpha_i) = 0$$

- u_{it} ha media zero, dato l'effetto fisso e l'intera storia delle X per l'unità corrispondente.
- Questa è un'estensione della precedente assunzione 1 della regressione multipla.
- Ciò significa che non vi sono effetti passati omessi (qualsiasi effetto passato di X deve essere incluso esplicitamente).
- Inoltre, non c'è feedback da u su X futuri:
  - il fatto che uno stato abbia un tasso di mortalità particolarmente alto quest'anno non influisce sull'aumento delle imposte sulla birra;
  - talvolta questa assunzione di "assenza di feedback" è plausibile, talvolta no.

Assunzione 2:

$$(X_{i1}, \ldots, X_{iT}, u_{i1}, \ldots, u_{iT}),\ i = 1, \ldots, n \text{ i.i.d.}$$

- È un'estensione dell'assunzione 2 per la regressione multipla con dati sezionali.
- È soddisfatta se le unità sono prese a caso dalla popolazione mediante campionamento casuale semplice.
- Non richiede che le osservazioni siano i.i.d. nel tempo per la stessa unità - sarebbe irrealistico. Il fatto che uno stato abbia un'imposta sulla birra elevata quest'anno è un buon predittore del (è correlato con) fatto che avrà un'imposta sulla birra elevata l'anno seguente. Similmente, il termine d'errore per un'unità in un anno è plausibilmente correlato con il suo valore l'anno dopo, cioè $corr(u_{it}, u_{it+1})$ è plausibilmente diverso da zero.

Autocorrelazione (o correlazione seriale)

Supponiamo che una variabile Z sia osservata in diverse date t, perciò le osservazioni sono su $Z_t, t = 1, \ldots, T$ (consideriamo che vi sia una sola unità).

Allora Z_t è detta autocorrelata o serialmente correlata se $corr(Z_t, Z_{t+j}) \neq 0$ per date $j \neq 0$.

- "Autocorrelazione" significa correlazione con se stesso.
- $cov(Z_t, Z_{t+j}) \neq 0$ è detta j-esima autocovarianza di Z_t.
- Nell'esempio della guida in stato di ebbrezza, u_{it} include la variabile omessa delle condizioni meteo dell'anno per lo stato i. Se gli inverni nevosi si presentano in gruppi (uno segue l'altro), allora u_{it} sarà autocorrelata (perché?)
- In molte applicazioni con dati panel, u_{it} è plausibilmente autocorrelata.

Indipendenza e autocorrelazione in dati panel, un'immagine:

	$i=1$	$i=2$	$i=3$	L	$i=n$
$t=1$	u_{11}	u_{21}	u_{31}	L	u_{n1}
M	M	M	M	L	M
$t=T$	u_{1T}	u_{2T}	u_{3T}	L	u_{nT}

← Campionamento i.i.d. tra le unità →

- Se le unità sono ottenute per campionamento casuale semplice, allora $(u_{i1}, \ldots, u_{iT})$ è indipendente da $(u_{j1}, \ldots, u_{jT})$ per unità diverse con $i \neq j$.
- Ma se i fattori omessi che comprendono u_{it} sono serialmente correlati, allora u_{it} è serialmente correlato.

Sotto le assunzioni dei minimi quadrati per dati panel:

- Lo stimatore OLS con effetto fisso $\hat{\beta}_1$ è non distorto, consistente e ha distribuzione asintotica normale.
- Tuttavia, i consueti errori standard OLS (sia di omoschedasticità pura sia robusti all'eteroschedasticità) saranno in generale sbagliati perché assumono che u_{it} non sia serialmente correlata.
  - In pratica, gli errori standard OLS spesso sottostimano l'incertezza del campionamento reale: se u_{it} è correlato nel tempo, non si hanno molte informazioni (molta variazione casuale) come si avrebbero se u_{it} fosse incorrelata.
  - Il problema si risolve usando errori standard "clustered".

Errori standard per dati raggruppati

- Gli errori standard per dati raggruppati stimano la varianza di $\hat{\beta}_1$ quando le variabili sono i.i.d. tra le unità ma sono potenzialmente autocorrelate in una unità.
- È più facile comprenderli se si considera prima il problema più semplice di stimare la media di Y usando dati panel…

Errori standard clustered per la media stimata con dati panel

$$Y_{it} = \mu + u_{it} \;\; i = 1, \ldots, n, \;\; t = 1, \ldots, T$$

Lo stimatore della media μ è $\bar{Y} = \frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} Y_{it}$.

E' utile scrivere $\bar{Y}$ come media tra le unità el valore medio per ciascuna unità:

$$\bar{Y} = \frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} Y_{it} = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{1}{T} \sum_{t=1}^{T} Y_{it} \right) = \frac{1}{n} \sum_{i=1}^{n} \bar{Y}_{i.}$$

dove $\bar{Y}_{i.} = \frac{1}{T} \sum_{t=1}^{T} Y_{it}$ è la media campionaria per l'unità i.

Poiché le osservazioni sono i.i.d. tra le entità, $(\bar{Y}_{1.}, \ldots, \bar{Y}_{n.})$ sono i.i.d. Quindi, se n è grande, vale il TLC:

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} \bar{Y}_{i.} \xrightarrow{p} N\left(0, \sigma^2_{\bar{Y}_{i.}}/n\right), \text{ dove } \sigma^2_{\bar{Y}_{i.}} = var(\bar{Y}_{i.})$$

- L'errore standard di $\bar{Y}_{i.}$ è la radice quadrata di uno stimatore di $\sigma^2_{\bar{Y}_{i.}}/n$.
- Lo stimatore naturale di $\sigma^2_{\bar{Y}_{i.}}$ è la varianza campionaria di $\bar{Y}_{i.}$, cioè $s^2_{\bar{Y}_{i.}}$

Quindi la formula dei clustered standard error (raggruppati) di $\bar{Y}$ è data da:

$$\text{se}(\bar{Y}) = \sqrt{\frac{s^2_{\bar{Y}_{i.}}}{n}}, \text{ dove } s^2_{\bar{Y}_{i.}} = \frac{1}{n-1}\sum_{i=1}^{n}(\bar{Y}_{i.} - \bar{Y})^2$$

Cos’hanno di speciale gli errori standard per dati raggruppati?

- Non molto, in verità - la procedura di derivazione vista in precedenza è la stessa usata nel Capitolo 3 per derivare l’errore standard della media campionaria, con la differenza che qui i “dati” sono le medie di unità i.i.d. $(\bar{Y}_{1.}, \ldots, \bar{Y}_{n.})$ anziché una singola osservazione i.i.d. per ciascuna unità.
- C’è una caratteristica importante: nella derivazione dell’errore standard per dati raggruppati non abbiamo mai assunto che le osservazioni siano i.i.d. in una unità. Quindi abbiamo implicitamente consentito la correlazione seriale in una unità.
- E la correlazione seriale, dov’è finita? Determina $\sigma^2_{\bar{Y}_{i.}}$, la varianza di $\bar{Y}_i$ …

La correlazione seriale in Y_{it} si inserisce in $\sigma^2_{\bar{Y}_{i.}}$

$$
\begin{aligned}
\sigma^2_{\bar{Y}_{i.}} &= \text{var}(\bar{Y}_{i.}) \\
&= \text{var}\left(\frac{1}{T}\sum_{t=1}^{T} Y_{it}\right) = \frac{1}{T^2}\text{var}(Y_{i1}+\ldots+Y_{iT}) = \\
&= \frac{1}{T^2}[\text{var}(Y_{i1})+\ldots \text{var}(Y_{iT})+ \\
&+ \ 2\text{cov}(Y_{i1}, Y_{i2}) + 2\text{cov}(Y_{i1}, Y_{i3})+\ldots+2\text{cov}(Y_{iT-1}, Y_{iT})]
\end{aligned}
$$

- Se Y_{it} è serialmente incorrelata, tutte le autocovarianze=0 e abbiamo la consueta derivazione del Capitolo 3.
- Se queste autocovarianze non sono zero, la formula consueta (che le pone a 0) sarà errata.
- Se queste autocovarianze sono positive, la formula consueta sottostimerà la varianza di $\bar{Y}_i$.

Errore standard clustered di $\bar{Y}$ è dato da $\sqrt{s^2_{\bar{Y}_{i.}}/n}$, dove:

$$
\begin{aligned}
s^2_{\bar{Y}_{i.}} &= \frac{1}{n-1}\sum_{i=1}^{n}(\bar{Y}_{i.}-\bar{Y})^2 = \\
&= \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{1}{T}\sum_{t=1}^{T}Y_{it}-\bar{Y}\right)^2 = \\
&= \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{1}{T}\sum_{t=1}^{T}(Y_{it}-\bar{Y})\right)^2 = \\
&= \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{1}{T}\sum_{t=1}^{T}(Y_{it}-\bar{Y})\right)\left(\frac{1}{T}\sum_{s=1}^{T}(Y_{is}-\bar{Y})\right) = \\
&= \frac{1}{n-1}\sum_{i=1}^{n}\frac{1}{T^2}\left[\sum_{t=1}^{T}\sum_{s=1}^{T}(Y_{it}-\bar{Y})(Y_{is}-\bar{Y})\right] = \\
&= \frac{1}{T^2}\sum_{t=1}^{T}\sum_{s=1}^{T}\left[\frac{1}{n-1}\sum_{i=1}^{n}(Y_{it}-\bar{Y})(Y_{is}-\bar{Y})\right]
\end{aligned}
$$

Errore standard clustered di $\bar{Y}$

$$s^2_{\bar{Y}_{i.}} = \frac{1}{T^2} \sum_{t=1}^{T} \sum_{s=1}^{T} \left[\frac{1}{n-1} \sum_{i=1}^{n} (Y_{it} - \bar{Y})(Y_{is} - \bar{Y}) \right]$$

- Il termine tra le parentesi quadre $\frac{1}{n-1} \sum_{i=1}^{n} (Y_{it} - \bar{Y})(Y_{is} - \bar{Y})$ stima l'autocovarianza tra Y_{it} e Y_{is}. Quindi la formula dell'errore standard clustered implicitamente stima tutte le autocovarianze, usandole per stimare $\sigma^2_{\bar{Y}_{i.}}$!
- Per contrasto, la formula "consueta" pone a zero queste autocovarianze omettendo tutti i termini misti - il che è valido solo se queste autocovarianze sono tutte zero.

Errori standard clustered per lo stimatore con effetti fissi nella regressione con dati panel

- Il concetto di errori standard clustered per dati panel è del tutto analogo al precedente caso della media per dati panel - solo più complesso per notazione e formule. Si veda l'Appendice 10.2.
- Gli errori standard clustered per dati panel sono l'estensione logica di quelli robusti all'eteroschedasticità per dati sezionali. Nella regressione con dati sezionali, gli errori standard robusti all'eteroschedasticità sono validi indipendentemente dal fatto che vi sia eteroschedasticità. Nella regressione con dati panel, gli errori standard clustered sono validi indipendentemente dal fatto che vi sia eteroschedasticità e/o correlazione seriale.
- Tra l'altro…Il termine "clustered" deriva dal fatto che si consente correlazione in un "cluster" (o "gruppo") di osservazioni (in una entità) ma non tra cluster.
- `plm` calcola automaticamente gli errori standard clustered

Applicazione: leggi sulla guida in stato di ebbrezza e mortalità stradale (Paragrafo 10.6)

Alcuni fatti

- Circa 40.000 morti sulle strade ogni anno negli USA
- 1/3 degli incidenti mortali coinvolge un guidatore ubriaco
- 25% dei guidatori sulle strade tra l'1 e le 3 del mattino ha bevuto (stima)
- Un guidatore ubriaco ha 13 volte più probabilità di causare un incidente mortale rispetto a un guidatore sobrio (stima)

Leggi sulla guida in stato di ebbrezza e mortalità stradale (continua)

Aspetti di politica publica

- La guida in stato di ebbrezza causa importanti esternalità (guidatori sobri vengono uccisi, la società sostiene costi medici, ecc.) - vi è ampia giustificazione per un intervento del governo
- Esistono modi efficaci per ridurre la guida in stato di ebbrezza? Se sì, quali?
- Quali sono gli effetti di leggi specifiche:
  - pene obbligatorie
  - età minima legale per bere alcolici
  - interventi economici (imposte sugli alcolici)

Dati panel per la guida in stato di ebbrezza

$n = 48$ stati USA, $T = 7$ anni (1982,…,1988) (bilanciato)

Variabili

- Tasso di mortalità stradale (morti per 10.000 residenti)
- Imposta su una cassa di birra (Beertax)
- Età minima di legge per bere alcolici
- Pene minime per la prima violazione:
  - Pena obbligatoria
  - Servizio sociale obbligatorio
  - altrimenti, la sentenza sarà soltanto pecuniaria
- Miglia per veicolo per guidatore (US DOT)
- Dati economici sullo stato (reddito pro capite, ecc.)

Perché i dati panel potrebbero aiutare?

- Potenziale distorsione da variabili omesse per variabili che variano tra stati ma sono costanti nel tempo:
  - cultura del bere e del guidare
  - qualità delle strade
  - età delle automobili sulle strade

$\Rightarrow$ usa effetti fissi di stato

- Potenziale distorsione da variabili omesse per variabili che variano nel tempo ma sono costanti tra stati:
  - miglioramenti nella sicurezza delle auto nel tempo
  - mutamento atteggiamenti verso la guida in stato di ebbrezza a livello nazionale

$\Rightarrow$ usa effetti temporali

Tabella 10.1 Analisi degli effetti della legislazione in materia di guida in stato d'ebbrezza sulla mortalità stradale.

Variabile dipendente: tasso di mortalità stradale (morti su 10.000 abitanti)

Regressore	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Imposta sulla birra	0,36** (0,05)	– 0,66* (0,29)	– 0,64+ (0,36)	– 0,45 (0,30)	– 0,69* (0,35)	– 0,46 (0,31)	– 0,93** (0,34)
Età minima legale 18				0,028 (0,070)	– 0,010 (0,083)		0,037 (0,102)
Età minima legale 19				– 0,018 (0,050)	– 0,076 (0,068)		– 0,065 (0,099)
Età minima legale 20				0,032 (0,050)	– 0,100+ (0,056)		– 0,113 (0,125)
Età minima legale						– 0,002 (0,021)	
Pena detentiva o servizi per la comunità				0,038 (0,103)	0,085 (0,111)	0,039 (0,103)	0,089 (0,164)
Miglia medie per guidatore				0,008 (0,007)	0,017) (0,011)	0,009 (0,007)	0,124 (0,049)
Tasso di disoccupazione				– 0,063** (0,013)		– 0,063** (0,013)	– 0,091** (0,021)
Reddito reale pro capite (logaritmo)				1,82** (0,64)		1,79** (0,64)	1.00 (0,68)
Anni	1982-88	1982-88	1982-88	1982-88	1982-88	1982-88	solo 1982 e 1988
Effetti fissi di stato?	no	sì	sì	sì	sì	sì	sì
Effetti temporali?	no	no	sì	sì	sì	sì	sì
Errori standard per i dati raggruppati?	no	sì	sì	sì	sì	sì	sì

Statistiche *F* e valori-*p* per l'esclusione di gruppi di variabili							
Effetti temporali = 0			4,22 (0,002)	10,12 (< 0,001)	3,48 (0,006)	10,28 (< 0,001)	37,49 (< 0,001)
Coefficienti età minime legali = 0				0,35 (0,786)	1,41 (0,253)		0,42 (0,738)
Tasso di disoccupazione e reddito pro capite = 0				29,62 (< 0,001)	31,96 (< 0,001)		25,20 (< 0,001)
$\bar{R}^2$	0,091	0,889	0,891	0,926	0,893	0,926	0,899

Queste regressioni sono state stimate utilizzando i dati panel per 48 stati USA. Le regressioni da (1) a (6) utilizzano dati per tutti gli anni dal 1982 al 1988, mentre la regressione (7) utilizza solo dati del 1982 e del 1988. I dati sono descritti nell'Appendice 10.1. Gli errori standard sono riportati tra parentesi sotto i coefficienti, e i valori-*p* sono riportati tra parentesi sotto le statistiche *F*. I coefficienti sono statisticamente significativi al livello del +10%, *5% o **1%.

Analisi empirica: risultati principali

- Il segno del coefficiente dell'imposta sulla birra cambia quando sono inclusi gli effetti fissi dello stato
- Gli effetti temporali sono statisticamente significativi ma la loro inclusione non ha un grande impatto sui coefficienti stimati
- L'effetto stimato dell'imposta sulla birra cala quando si includono altre leggi
- L'unica variabile politica che sembra avere un impatto è l'imposta sulla birra - non l'età legale minima per bere alcolici, non la pena minima obbligatoria ecc. - tuttavia l'imposta sulla birra non è significativa anche al livello del 10% usando errori standard clustered nelle specifiche che controllano per le condizioni economiche dello stato (tasso di disoccupazione, reddito personale)

Risultati empirici (continua)

- In particolare, l'età legale minima per bere alcolici ha un coefficiente piccolo che è stimato con precisione - riducendola non pare si abbia un grande effetto sulla mortalità stradale complessiva.
- Quali sono le minacce alla validità interna? Cosa si può dire su:
  1. Distorsione da variabili omesse
  2. Errata forma funzionale
  3. Distorsione da errori nelle variabili
  4. Distorsione da selezione del campione
  5. Distorsione da causalità simultanea

Che cosa ne pensate?

Digressione: estensioni del concetto di “n-1 regressori binari”

L’idea di utilizzare molti indicatori binari per eliminare la distorsione da variabili omesse può essere estesa a dati non panel - la chiave è che la variabile omessa sia costante per un gruppo di osservazioni, il che in effetti significa che ciascun gruppo ha la propria intercetta.

Esempio: effetto della dimensione delle classi.

Supponiamo che livelli di finanziamento e di istruzione siano determinati a livello della contea, e che ogni contea abbia diversi distretti. Se si è preoccupati della distorsione da variabili omesse risultante da variabili non osservate a livello di contea, si possono includere gli effetti di contea (indicatori binari, uno per ciascuna contea, omettendo una sola contea per evitare la collinearità perfetta).

Vantaggi e limitazioni della regressione con effetti fissi

Vantaggi

- Si può controllare per variabili non osservate che:
  - variano tra stati ma non nel tempo e/o
  - variano nel tempo ma non tra stati
- La stima coinvolge estensioni relativamente semplici della regressione multipla
- La regressione con effetti fissi si può eseguire in tre modi:
  1. Metodo “prima e dopo”
  2. “n-1 regressori binari” quando n è piccolo
  3. Regressione “in deviazione dalle medie”
- Metodi simili si applicano alla regressione con effetti temporali e a quella con effetti fissi e temporali
- Inferenza statistica: come nella regressione multipla.

Limitazioni/problemi aperti**

- Necessaria la variazione in X nel tempo nelle entità
- Gli effetti di ritardo temporale possono essere importanti - anche se non ne abbiamo tenuto conto nel modello dell'imposta sulla birra
- È necessario usare errori standard clustered per evitare la possibilità che u_{it} sia autocorrelato