Effetti del terremoto che ha colpito il Centro Italia nel 2016 sul Totale dei Ricavi delle imprese di Costruzioni di edifici

La nostra analisi è divisa nelle seguenti parti:

1. Osservazione del Dataset (Do File)
2. Regressione:
   - Stima del modello (Fixed o Random Effect)
   - Risultati della Regressione
3. Parallel Trend e Matching
4. Difference-in-Difference
5. Before and After

REGRESSIONE

La prima parte della nostra analisi è dedicata a stimare se la nostra regressione predilige effetti Fissi o effetti Casuali. Se il campione è perfettamente rappresentativo della popolazione saremmo interessati ad effetti fissi altrimenti saremmo interessati ad effetti random.
Possiamo verificare ciò tramite *Hausman Specification Test*.
La variabile dipendente che abbiamo preso in considerazione è il Totale dei ricavi, mentre le variabili indipendenti sono: EBITDA/Vendite, Utile o Perdita d'esercizio, costi del personale e totale costi della produzione.

Per semplificare l'interpretazione dei nostri risultati abbiamo deciso di trasformare il valore delle nostre variabili in logaritmi.

$$log_ric = \alpha + \beta_1 log_ebitdales + \beta_2 log_utile + \beta_3 log_costiper + \beta_4 log_costiprod$$

Hausman (1978) specification test

	Coef.
Chi-square test value	1355.368
P-value	0

H_0: il modello da applicare è con random effect
H_1: il modello da applicare è con fixed effect
Osservando il p-value del nostro test, che è uguale a 0 (<0,05), possiamo dire che rifiutiamo l'ipotesi nulla ed accettiamo l'ipotesi alternativa che suggerisce che dobbiamo usare un modello con Fixed Effect.

Abbiamo quindi inviato un'altra volta la regressione con fixed effect e salvato i risultati:

Regression results

log_ric	Coef.	St.Err.	t-value	p-value	[95% Conf	Interval]	Sig
log_ebitdasales	-.488	.005	-104.78	0	-.497	-.479	***
log_utile	.191	.003	72.85	0	.186	.196	***
log_costipers	.063	.004	14.09	0	.055	.072	***
log_costiprod	.554	.006	90.47	0	.542	.566	***
Constant	3.239	.039	83.19	0	3.163	3.316	***

Mean dependent var	7.152	SD dependent var	1.211
R-squared	0.714	Number of obs	30193
F-test	12518.675	Prob > F	0.000
Akaike crit. (AIC)	7469.839	Bayesian crit. (BIC)	7511.416

*** $p<.01$, ** $p<.05$, * $p<.1$

Tutte le variabili considerate nella nostra analisi sono statisticamente significative in quanto il loro p-value è minore del livello di significatività considerato nella regressione (5%).
log_ric = 3.24 – 0.49log_ebitdales + 0.19log_utile + 0.06log_costiper + 0.55log_costiprod

Notiamo che vi è una relazione negativa tra l'indice di redditività E/V e le vendite (un aumento unitario del logaritmo di E/V fa diminuire il valore del logaritmo delle vendite di 0.488), mentre vi è una relazione positiva tra tutte le altre variabili indipendenti e le vendite.
I risultati possono essere interpretati sia nel loro segno che nella loro entità (a differenza di quelli che avremmo ottenuto se avessimo applicato un modello probit o logit).

PARALLEL TREND

Abbiamo creato una DUMMY VARIABLE uguale a 1 quando considera le società delle 8 province più colpite dal nostro trattamento (che sono Perugia, Terni, Fermo, Macerata, Ascoli Piceno, Rieti, Teramo, L'Aquila) e uguale a 0 se considera le medesime società delle altre province italiane.
Effettuiamo un ttest: ne risulta un p-value piccolo e questo vuol dire che non c'era parallel trend (non è possibile considerare la media del gruppo dei trattati simile a quella del gruppo di controllo prima del trattamento, in altre parole, la differenza fra le medie dei due gruppi è statisticamente diversa da zero). Abbiamo quindi bisogno di forzare un minimo la nostra ipotesi facendo un matching uno a uno (in quanto il nostro campione è unbalanced) per rendere le nostre osservazioni più simili possibile. Dopo aver fatto ciò eliminiamo tutte le osservazioni non matchate (vedendo che ci cancella circa 80000 osservazioni) ed effettuiamo nuovamente il ttest:

Two-sample t test with equal variances

	obs1	obs2	Mean1	Mean2	dif	St Err	t value	p value
log ric by r: 0 1	947	948	6.899	6.942	-.043	.061	-.7	.485

Adesso il nostro p-value risulta grande (maggiore sicuramente del 5%) e possiamo dire che matchando le nostre osservazioni abbiamo osservato che prima del trattamento c'era PARALLEL TREND.

DIFFERENCE-IN-DIFFERENCE

Dopo aver visto che avevamo parallel trend, e creato la DUMMY VARIABLE t uguale a 1 se prendiamo in considerazioni le osservazioni uguali o maggiori al 2016 e 0 altrimenti, possiamo effettuare la DID:

Regression results

log_ric	Coef.	St.Err.	t-value	p-value	[95% Conf	Interval]	Sig
o	0	.	.	.	.	.	
t	.254	.116	2.19	.029	.026	.482	**
rt	-.096	.121	-0.79	.427	-.334	.141	
Constant	6.844	.034	203.86	0	6.778	6.91	***

Mean dependent var	6.951	SD dependent var	1.358
R-squared	0.016	Number of obs	3934
F-test	12.899	Prob > F	0.000
Akaike crit. (AIC)	5094.250	Bayesian crit. (BIC)	5113.082

*** p<.01, ** p<.05, * p<.1

Dai risultati, possiamo notare che la variabile r esce dalla nostra regressione a causa della multicollinearità, e che il p-value del trattamento (rt) non è statisticamente significativo e di conseguenza non è possibile interpretare il suo coefficiente.
In altre parole, il terremoto del 2016 non ha prodotto differenze significative nei ricavi delle diverse imprese italiane appartenenti al settore di costruzione di edifici.

BEFORE AND AFTER

Possiamo, però, analizzare l'effetto che il nostro trattamento ha avuto sulle imprese delle province più colpite dal terremoto prima e dopo il suo verificarsi.
Abbiamo creato una DUMMY VARIABILE (trattamento) che assume il valore 1 quando considera le imprese delle province colpite dopo il trattamento e 0 quando considera le imprese delle principali province colpite prima del trattamento.

Regression results

log_ric	Coef.	St.Err.	t-value	p-value	[95% Conf	Interval]	Sig
trattamento	.192	.068	2.82	.005	.059	.326	***
Constant	6.924	.019	360.45	0	6.887	6.962	***

Mean dependent var	6.951	SD dependent var	1.358
R-squared	0.006	Number of obs	1967
F-test	7.975	Prob > F	0.000
Akaike crit. (AIC)	3781.973	Bayesian crit. (BIC)	3793.142

*** p<.01, ** p<.05, * p<.1

Osservando i risultati è possibile notare come questo evento abbia causato un aumento dei Ricavi delle Vendite delle imprese considerate in quanto, come suggerisce il p-value (minore del livello di confidenza del 5%), vi è una differenza statisticamente significativa.

Possiamo suppore che questo aumento dei ricavi delle vendite (calcolati come prezzo per quantità) delle imprese di costruzioni di edifici nelle province colpite possa essere dovuto principalmente da due fatti:

1. La maggiore necessità di dover fare affidamento a queste società per poter ricostruire quelle abitazioni distrutte dal terremoto (aumentando la variabile quantità)
2. Un aumento del livello generale dei prezzi nel tempo (aumentando la variabile prezzo);
   come esempio, a solo scopo dimostrativo, dell'inflazione in quegli anni abbiamo deciso di considerare il NIC:

Indice nazionale generale dei prezzi al consumo per l'intera collettività (NIC), al lordo dei tabacchi.
[1]**Base 2014 = 100**

2014	**2015**	**2016**	**2017**
100,00	100,09	99,99	101,29

La leggere diminuzione nel 2016 è stata più che compensata dall'aumento nel 2017

Fonte: ISTAT
https://www.istat.it/it/archivio/263418

[1]*I dati originariamente erano su base 2015 ma li abbiamo riponderati su base 2014*

In conclusione, dalla nostra analisi possiamo assumere che il terremoto del 2016 ha sicuramente impattato sui ricavi delle imprese di costruzioni delle principali province coinvolte, determinandone un aumento del 19.2% che tuttavia non è stato sufficiente a produrre differenze significative rispetto alle restanti imprese dello stesso settore.

DO File

******** RISK MANAGEMENT IN BANKING a.a. 20/21 ********

**** STATA PROJECT Lorenzo Guglielmetti 520303 Matteo Valentini 508274 ****

*** Analisi condotta sulle imprese italiane operanti nel settore "41 - Costruzione di edifici" ***
* Abbiamo usato il database di Aida per scaricare le seguenti variabili: i nomi delle società, i CCIAA (come numero identificativo delle società), le Province delle sedi legali, i valori dei ricavi dal 2014 al 2017, i costi del personale dal 2014 al 2017, gli utili o perdite dal 2014 al 2017, i costi della produzione dal 2014 al 2017, i valori degli EBITDA/Vendite dal 2014 al 2017.

* Apriamo il log file
log using "/Users/lorenzoguglielmetti/Desktop/PROGETTO STATA LG MV/Log File Lorenzo Guglielmetti Matteo Valentini.smcl"

* Dopo aver scaricato i dati gli abbiamo importati su Stata chiedendo di importare la prima riga come nome delle variabili:
import excel "/Users/lorenzoguglielmetti/Desktop/PROGETTO STATA LG MV/aida province.xls", sheet("Risultati") firstrow

********** OSSERVIAMO IL DATASET **********

* Osserviamo la lista delle variabili del nostro dataset:
describe

* Osservando le variabili (both variable name and variable label) vediamo che i nostri dati non sono organizzati come PANEL DATA.
* Abbiamo bisogno di fare un reshape:
reshape long ric costipers utile costiprod ebitdasales, i(CCIAA) j(year)

* Osserviamo di nuovo le variabili:
describe
* I nostri dati adesso sono organizzati come PANEL DATA (classificati in base agli anni).

* Possiamo vedere anche le statistiche descrittive:
sum
* In questo modo possiamo vedere il numero delle osservazioni, la media, la standard deviation, ed i valori minimi e massimi delle nostre variabili.

* Inoltre, inviando il seguente codice potremmo vedere il nostro database ordinato in base al CCIAA ovvero all'identificativo della società:
br

********** REGRESSIONE **********

* Prima di fare la nostra REGRESSIONE dobbiamo dichiarare il PANEL
xtset CCIAA year

* STATA ci da il seguente errore: "string variables are not allowed in var list". Infatti anche guardando il risultato di "describe" vediamo che "CCIAA" è una stringa e non possiamo usare panelvar con una variabile non-numerica, quindi dobbiamo trasformare la stringa in numero prima di dichiarare il panel:
encode CCIAA, generate(CCIAA_n)
describe CCIAA_n
* Abbiamo trasformato la nostra stringa in numero.

*DICHIARIAMO DI NUOVO IL PANEL:
xtset CCIAA_n year

* Controlliamo che non ci sia duplicati:
duplicates report CCIAA_n year

***** FIXED O RANDOM EFFECT - HAUSMAN SPECIFICATION TEST *****
* Quando il campione è perfettamente rappresentativo della popolazione saremmo interessati a fixed effect; se invece il campione non è totalmente rappresentativo della popolazione dovremmo applicare dei random effect.
* Ci servono nella regressione per controllare tutte quelle caratteristiche che sono differenti da unità a unità.
* Metodo pratico per capire quale effetto usare: HAUSMAN SPECFICATION TEST
* Ipotesi nulla (H0): modello da applicare è con random effect
* Ipotesi alternativa (H1): modello da applicare è con fixed effect

* Prima di effettuare il test abbiamo deciso di trasformare il valore delle nostre variabili in logaritmi in modo che possiamo semplificare l'interpretazione dei nostri risultati:
gen log_ric = log(ric)
gen log_costipers = log(costipers)
gen log_utile = log(utile)
gen log_costiprod = log(costiprod)
gen log_ebitdasales = log(ebitdasales)

xtreg log_ric log_ebitdasales log_utile log_costipers log_costiprod, fe
estimates store fixed
xtreg log_ric log_ebitdasales log_utile log_costipers log_costiprod, re
estimates store random
hausman fixed random

* Guardando il p-value risultante dal "Hausman Test" (che è più piccolo di un qualsiasi normale livello di confidenza usato in un test) non acccettiamo l'ipotesi nulla quindi il modello preferito è un modello con fixed effect.

* Salviamo la nostra tabella come documento sul nostro computer in modo tale da poterla commentare meglio in un file word appropriato:
asdoc hausman fixed random, save(/Users/lorenzoguglielmetti/Documents/TabRisultati.doc)

* Ora che sappiamo che il modello da applicare è con FIXED EFFECT, Mandiamo di nuovo la regressione (in cui la variabile dipendente sono i logaritmi delle vendite) e Salviamo i risultati.
* Abbiamo usato "append" per salvare la tabella nello stesso file usato precedentemente: useremo questo comando anche in seguito per le prossime tabelle.
xtreg log_ric log_ebitdasales log_utile log_costipers log_costiprod, fe

asdoc xtreg log_ric log_ebitdasales log_utile log_costipers log_costiprod, fe,
save(/Users/lorenzoguglielmetti/Documents/TabRisultati.doc), append

********** PARALLEL TREND **********

* Osserviamo la variabile "Provincia" per vedere come sono distribuite le nostre osservazioni
tab Provincia
* Possiamo anche osservare la variabile "Provincia" in base agli anni. In questo modo avremo il numero di osservazioni per ogni Provincia in base ad ogni anno considerato:
tab Provincia year

* Creaiamo due DUMMY VARIABLES per "Provincia" e "year".
generate r = .
replace r = 1 if Provincia == "Perugia" | Provincia=="Terni" | Provincia=="Fermo" | Provincia=="Macerata" | Provincia == "Ascoli Piceno" | Provincia=="Rieti" | Provincia=="Teramo" | Provincia=="L'Aquila"
replace r = 0 if Provincia != "Perugia" & Provincia!="Terni" & Provincia!="Fermo" & Provincia!="Macerata" & Provincia != "Ascoli Piceno" & Provincia!="Rieti" & Provincia!="Teramo" & Provincia!="L'Aquila"

generate t = .
replace t = 1 if year >=2016
replace t = 0 if year <2016

* Controlliamo se le nostre DUMMY sono state create correttamente
tab r
tab t
* La DUMMY "r" assume valore 1 quando la sede legale della società è situata in Provincia di Perugia o Terni o Fermo o Macerata o Ascoli Piceno o Rieti o Teramo o L'Aquila e 0 altrimenti.
* La DUMMY "t" assume valore 1 quando consideriamo anni maggiori uguali al 2016 e 0 altrimenti. Ci servirà in seguito per la DIFFERENCE-IN-DIFFERENCE.

* Facciamo un ttest per controllare se prima del nostro trattamento c'era PARALLEL TREND:
ttest log_ric if year<2016 , by(r)

* Dal risultato del ttest possiamo vedere che NON c'era PARALLEL TREND: il p-value è piccolo e questo vuol dire che rifiutiamo l'ipotesi nulla.

* Abbiamo bisogno di fare un MATCHING:
* Il nostro campione è unbalanced quindi ci conviene fare un matching uno a uno per rendere le nostre osservazioni più simili possibili:
psmatch2 r (log_ric), neighbor(1) caliper (0.5) noreplacement
* Cancelliamo tutte le osservazioni non matchate:
keep if _weight!=.
tab r
* Vediamo che ci ha cancellato circa 80 mila osservazioni.

* Facendo il MATCHING sappiamo che stiamo forzando la nostra ipotesi ma in questa maniera, rifacendo il ttest, vediamo che c'era PARALLEL TREND in quanto il risultato del test ci dà p-value grande.
ttest log_ric if year<2016 , by(r)

* Salviamo i risultati del nostro ttest:
asdoc ttest log_ric if year<2016 , by(r), save(/Users/lorenzoguglielmetti/Documents/TabRisultati.doc), append

********** DIFFERENCE-IN-DIFFERENCE **********

* Come abbiamo visto, dopo il matching, avevamo PARALLEL TREND quindi possiamo procedere alla nostra analisi DIFFERENCE-IN-DIFFERENCE ed in seguito salviamo i nostri risultati.

* Dobbiamo creare una variabile interazione tra Regione e year che chiameremo "rt":
gen rt=r*t

xtreg log_ric r t rt, fe

asdoc xtreg log_ric r t rt, fe, save(/Users/lorenzoguglielmetti/Documents/TabRisultati.doc), append

********** BEFORE AND AFTER **********

* Abbiamo bisogno di creare una variabile (che chiameremo "trattamento") che è uguale a 1 quando contiene tutte le società delle province considerate DOPO il trattamento ed uguale a 0 quando contiene tutte le società delle province considerate PRIMA del trattamento:

gen trattamento=.
replace trattamento=1 if year>=2016 & Provincia == "Perugia" | Provincia=="Terni" | Provincia=="Fermo" | Provincia=="Macerata" | Provincia == "Ascoli Piceno" | Provincia=="Rieti" | Provincia=="Teramo" | Provincia=="L'Aquila"

```
replace trattamento=0 if year<2016 & Provincia == "Perugia" | Provincia=="Terni" |
Provincia=="Fermo" | Provincia=="Macerata" | Provincia == "Ascoli Piceno" | Provincia=="Rieti" |
Provincia=="Teramo" | Provincia=="L'Aquila"

* Mandiamo la nostra regressione e salviamo i risultati:
xtreg log_ric trattamento, fe
asdoc xtreg log_ric trattamento, fe, save(/Users/lorenzoguglielmetti/Documents/TabRisultati.doc),
append

* Chiudiamo il log file
log close
```