Statistica Multivariata con Stata

Dott. Paolo Maranzano

Universita' Degli Studi di Bergamo
Dipartimento Scienze Aziendali, Economiche e Metodi Quantitativi
Corso di Laurea Triennale in Economia (L-33: Scienze economiche)

A.A. 2018/2019

Indice

1 Informazioni Generali

Cos'e' Stata?

Stata è un software statistico-econometrico che permette di organizzare ed elaborare dati, di produrre statistiche e grafici e di stimare una grande varietà di modelli econometrici.

Aprire e chiudere Stata

Si può trovare sul desktop o nella cartella di stata: ad esempio C: Per terminare la sessione si può eseguire il comando exit nella finestra Stata Command o cliccare sul simbolo a x che si usa per chiudere in generale le finestre di Windows. Se durante le operazioni eseguite il dataset è stato modificato, un avviso ci chiede se desideriamo chiudere il programma senza salvare le modifiche.

Versioni

Attualmente la versione disponibile in commercio di Stata è la 14.0. Rimangono ancora molto diffuse la 13.0 e la 12.0. Va precisato che le differenze fra queste versioni sono molto ridotte e limitate soprattutto a funzionalità complesse. La presente dispensa si riferisce alla versione per Windows ma va ricordato che il programma è disponibile anche per il sistema operativo MAC.

Per maggiori informazioni

I manuali del software sono voluminosi, completi e dettagliati. La guida in linea illustra in modo sintetico ma efficace i comandi. Per maggiori informazioni e curiosità si consiglia la consultazione del sito internet di Stata all'indirizzo: www.stata.com.

2 Stata per Windows

Le ultime versioni di Stata (12.0 e 13.0) sono caratterizzate da una struttura "a finestre". Una volta aperto il programma, ci si trova di fronte ad un desktop (di colore bianco) sul quale sono disposte le finestre. Ciascuna di esse ha una funzione specifica, solo le prime quattro descritte qui di seguito sono presenti al momento dell'apertura (default windowing) e possono tutte essere spostate, ridimensionate o chiuse dall'utente.

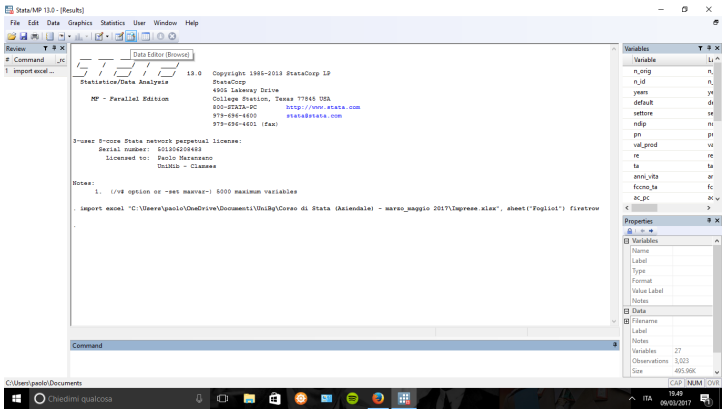


Figura 1: Schermata Principale

Stata Results

E' la finestra centrale, caratterizzata dallo sfondo bianco e dalla scritta Stata con le caratteristiche tecniche installate, dove vengono visualizzati i comandi eseguiti, i messaggi di errore e in generale tutte le operazioni eseguite dal programma.
È una finestra di output il cui funzionamento può essere ricondotto a quello di un "registro" che, se fatto scorrere, mostra all'utente ciò che fino a quel punto è stato fatto.

Stata Command

La sua posizione di default la colloca al di sotto della Stata Results. E' una finestra di input tramite la quale l'utente può immettere i tutti i comandi di Stata (uno alla volta) ed eseguirli premendo il tasto Enter. Il risultato dell'operazione o l'eventuale errore (con specificazione del tipo) vengono illustrati dopo l'esecuzione nella Stata Results.

Review

In alto a sinistra, contiene la sequenza dei comandi eseguiti nella Stata Command. Cliccando su una riga nella finestra review, il comando eseguito in precedenza ricompare nella finestra Stata Command. Funzione particolarmente utile per la correzione di errori o per procedure che richiedono ripetizioni frequenti di comandi articolati.

Variables e Properties

Variables, mostra l'elenco delle variabili contenute nel data set indicandone il nome, l'etichetta ed eventuali informazioni sulle modificazioni rilevanti intervenute nelle variabili. Properties indica le principali caratteristiche di ogni variabile: nome, etichetta, tipo di dato, formato visualizzazione.

Data Editor

Visualizza la matrice dei dati in memoria. Nelle colonne della matrice sono contenute variabili e nelle righe le singole osservazioni. Si apre con l'apposito pulsante sulla barra degli strumenti o digitando edit nella finestra Stata Command. Facendo doppio clic su un valore o una variabile ne vengono visualizzate nome, etichetta e formato. Il alto c'è una barra degli strumenti con la quale si possono eseguire alcune operazioni come mettere in ordine crescente i dati rispetto ad una variabile (sort) o cancellare dati o variabili. Una volta aperto l'editor, per accedere alle altre finestre è necessario chiuderlo.

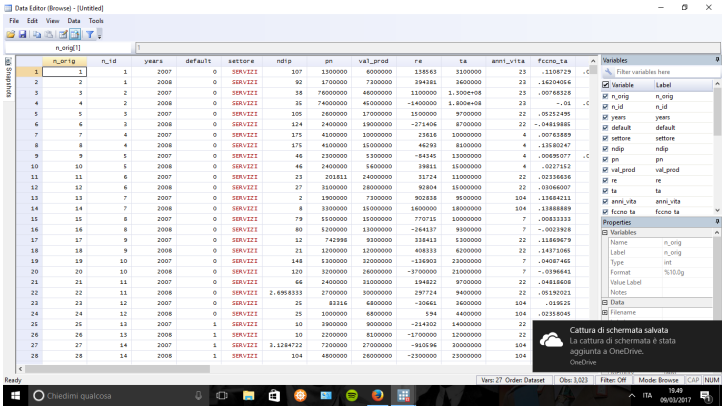


Figura 2: Data Editor

Do File Editor

Consiste in una finestra con la quale si può programmare Stata mediante la scrittura di una serie di comandi sul word editor di stata con funzione apposita per la scrittura dei do-files. Per eseguire un do file giaì elaborato, si clicca sul do-file editor al menu windows o sul pulsante dedicato nella barra degli strumenti.

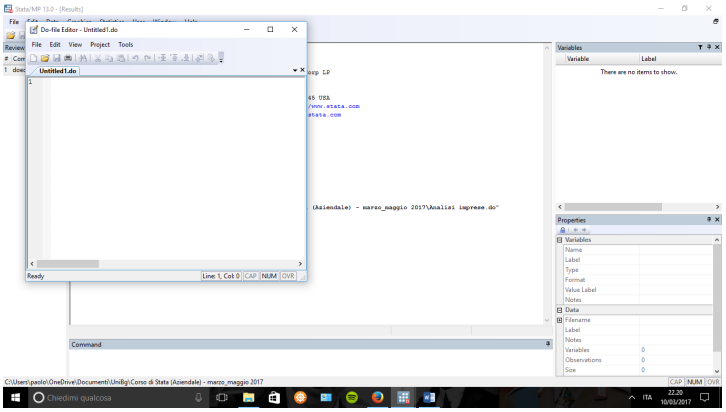


Figura 3: Do File Editor

Log File

La finestra Stata log file permette di visualizzare il log file, ossia il file che viene creato per la registrazione di tutti i comandi e le operazioni eseguite dal programma nella corrente sessione di utilizzo. Tale file può essere poi salvato o stampato per l'archiviazione dei risultati ottenuti. La finestra stata log si apre dal menu windows (se un log file è già in uso, altrimenti va prima creato) o utilizzando i comandi sulla barra degli strumenti.

3 Importazione di dataset da file esterni

3.1 File EXCEL

Per importare un dataset da file con estensione .XLS e .XLSX e' possibile utilizzare la barra dei comandi presente nella schermata principale oppure un comando apposito.
Manuale: FILE $\geq$ IMPORT $\geq$ EXCEL SPREADSHEET
Comando: import excel "Percorso File" firstrow, clear

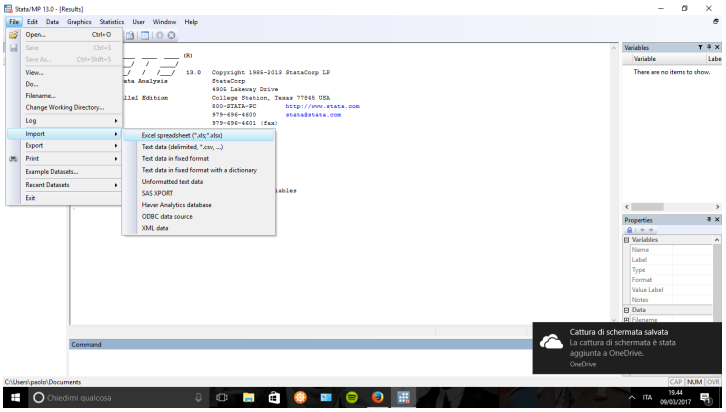


Figura 4: Import from Excel (1)

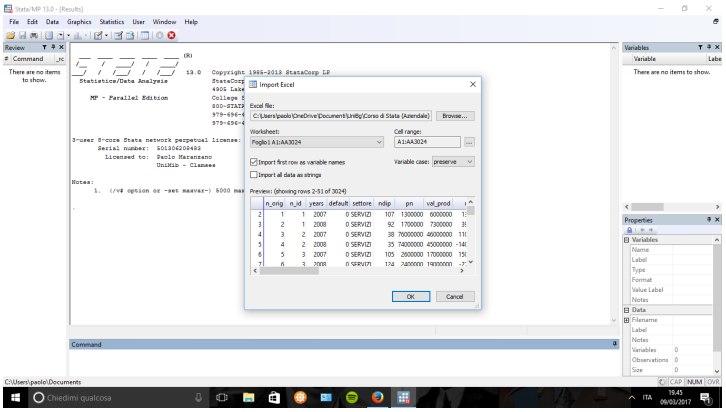


Figura 5: Import from Excel (2)

3.2 File CSV

Per importare un dataset da file con estensione .CSV ("Comma Separated Value") o formato .TXT e' possibile utilizzare la barra dei comandi presente nella schermata principale oppure un comando apposito.

Manuale: FILE $\geq$ IMPORT $\geq$ TEXT DATA
Comando: import delimited "Percorso File", clear

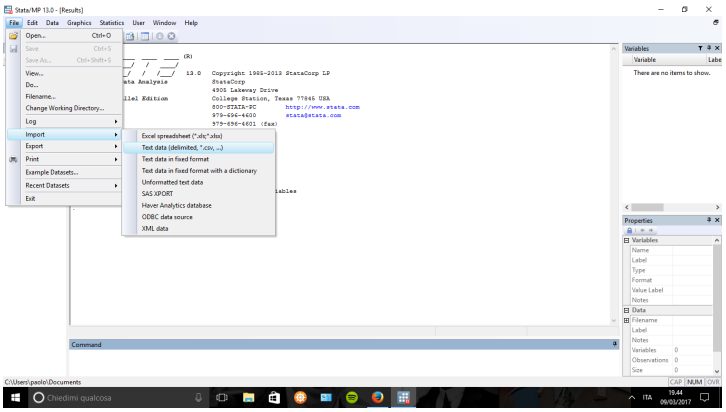


Figura 6: Import from CSV

3.3 File STATA

Per importare un dataset gia' predefinito in formato STATA .DTA e' possibile utilizzare la barra dei comandi presente nella schermata principale oppure un comando apposito.
Manuale: FILE $\geq$ OPEN
Comando: use "Percorso File", clear

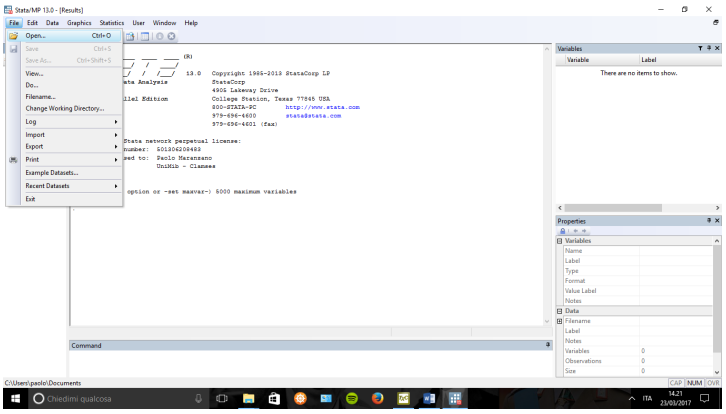


Figura 7: Import from DTA

4 Analisi preliminari

4.1 Descrizione delle variabili

L'analisi statistica deve innanzitutto partire dalla piena comprensione delle variabili presenti nel dataset.
Una descrizione della tipologia di variabili, del loro formato, le etichette che le rappresentano puo' aiutare ad identificare eventuali errori presenti nel dataset oppure capire quali analisi è importante effettuare.
Manuale: DATA $\geq$ DESCRIBE $\geq$ DESCRIBE IN MEMORY
Comando: describe

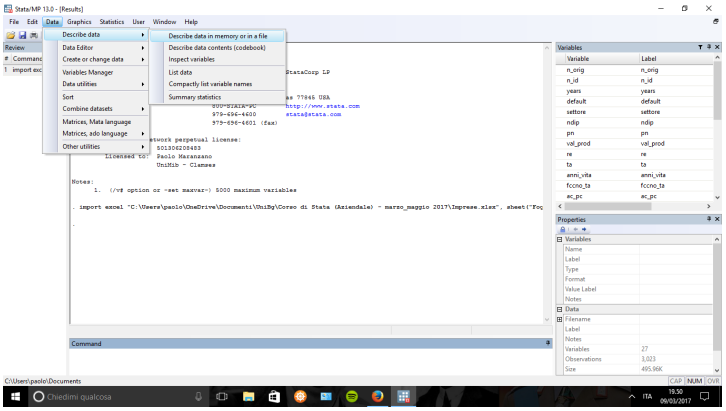


Figura 8: Descrizione delle variabili

Stata type	*Tipologia*
Byte	Variabile dicotomica (0/1)
Double	Numero decimale
Long	Numero decimale
Str	Stringhe e caratteri

Tabella 1: Tipologie di variabili

Missing values
Un numero puo' assumere un valore speciale se e' mancante (missing), denotato da un punto (.) Il valore missing non va confuso con lo zero poichè non viene utilizzato nelle espressioni aritmetiche.
Il valore missing viene considerato come il più elevato fra quelli che la variabile può assumere.
Nel produrre output statistici Stata ignora le osservazioni con missing values.

Stringhe
Una stringa è una serie di caratteri tipicamente racchiusa fra doppie virgo-

lette che la delimitano ma non ne fanno parte.
Esempi: “casa” “Casa” “ Casa” “Piazza Garibaldi” “” “x/y+k” “45.2”.
Si noti che “casa”, “Casa” e “ Casa” sono differenti e che “45.2” non è un numero perché è fra virgolette. La stringa “” è denominata null string e considerata da Stata come missing.

Formati numerici
Il formato numerico comincia con il simbolo %. Il primo numero indica l’ampiezza del risultato; il secondo, dopo il punto, la quantità di numeri dopo la virgola; ci può poi essere la lettera e per notazioni esponenziali; f per il formato fisso; g per il formato generale; l’opzione c indica il formato con la virgola.
Esempi: %9.0g %9.2fc %9.0gc %9.2e

Modifica del formato
Comando: format variabile nuovoformato

4.2 Gestione delle variabili

4.2.1 Eliminazione di una variabile

Manuale: DATA $\geq$ VAR. MANAG. $\geq$ TASTO DX VAR. $\geq$ DROP
Comando: drop variabile

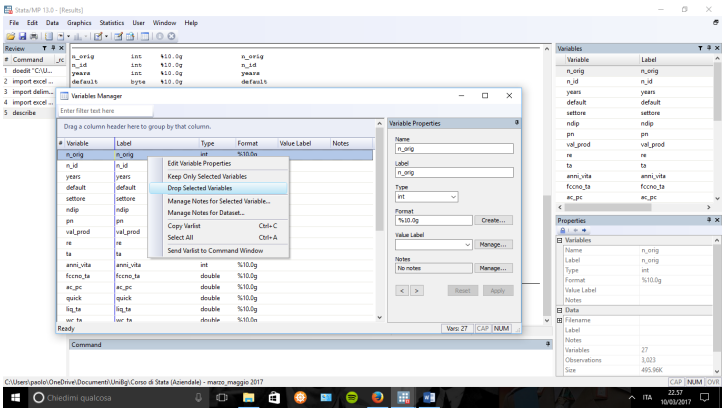


Figura 9: Eliminazione di una variabile

4.2.2 Creazione di una variabile

Manuale: DATA $\geq$ CREATE $\geq$ NEW VARIABILE
Comando: generate variabile= $f(...)$

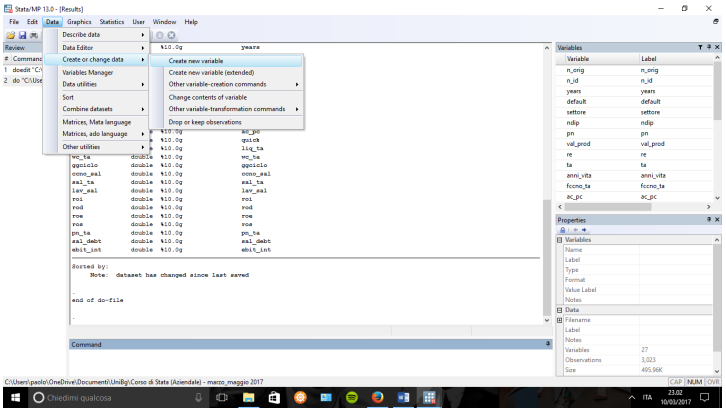


Figura 10: Creazione di una variabile

4.2.3 Rinominare una variabile

Manuale: DATA $\geq$ VAR. MANAG. $\geq$ TASTO DX VAR. $\geq$ NAME
Comando: rename vecchio.nome nuovo.nome

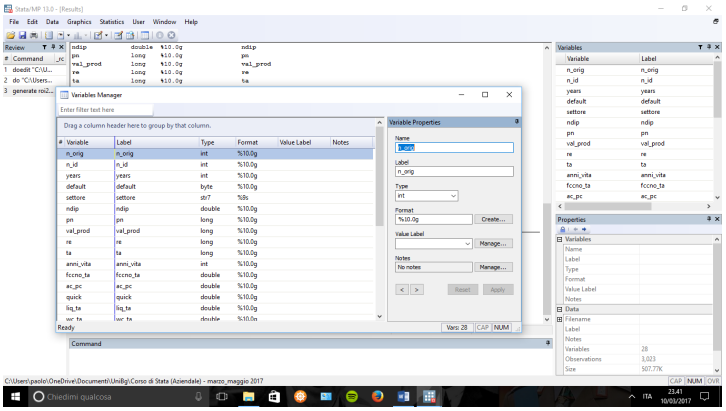


Figura 11: Rinominare una variabile

5 Analisi statistica descrittiva

5.1 Principali Statistiche Descrittive

5.1.1 Statistiche descrittive essenziali: Numero di osservazioni, Media, Deviazione standard, Range.

Manuale: STATS. $\geq$ SUMMAR. $\geq$ SUM.STATS. $\geq$ SUMMARY
Comando: SUMMARIZE variabili

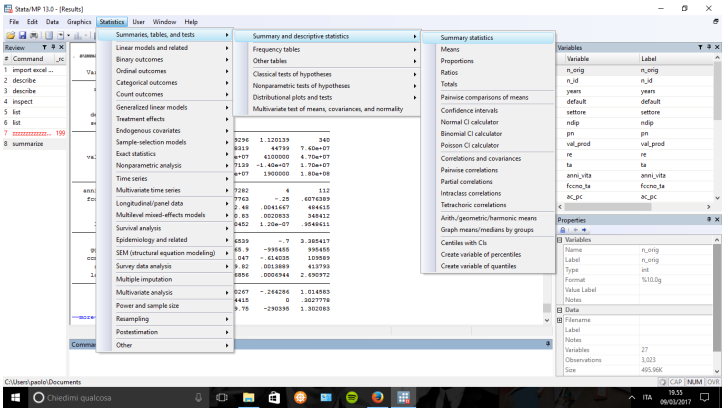


Figura 12: Summarize

```
. summarize default ndip pn val_prod re
```

Variable	Obs	Mean	Std. Dev.	Min	Max
default	300	.18	.3848294	0	1
ndip	300	60.34	55.11075	1	340
pn	300	5741623	9758319	44799	7.60e+07
val_prod	300	1.66e+07	1.06e+07	4100000	4.70e+07
re	300	232839	1857139	-1.40e+07	1.70e+07

Figura 13: Summarize

5.1.2 Statistiche descrittive dettagliate: Numero di osservazioni, Media, Deviazione standard, Range, Percentili, Varianza, Simmetria e Curtosi.

Manuale: STATS. $\geq$ SUMMAR. $\geq$ SUM.STATS. $\geq$ ADD DETAILS
Comando: SUMMARIZE variabili, DETAIL

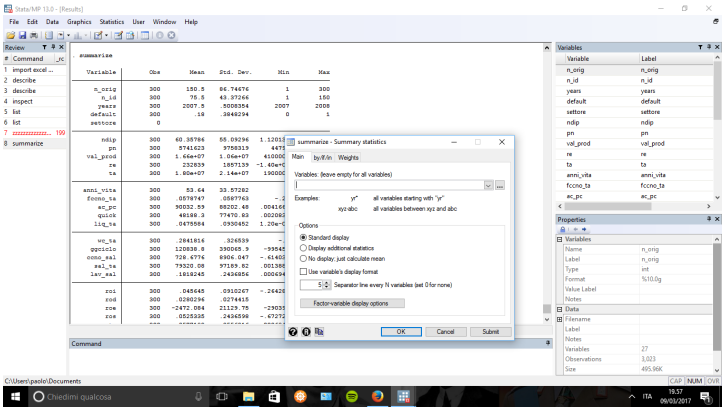


Figura 14: Summarize dettagliato (1)

```
. summarize ndip, detail

                            Num. Dipend.
-------------------------------------------------------------
      Percentiles      Smallest
 1%            1              1
 5%            7              1
10%         11.5              1       Obs                 300
25%           25              1       Sum of Wgt.         300

50%           40                      Mean              60.34
                      Largest         Std. Dev.       55.11075
75%         85.5            236
90%        126.5            237       Variance        3037.195
95%          177            340       Skewness        1.902112
99%        236.5            340       Kurtosis        7.662135
```

Figura 15: Summarize dettagliato (2)

5.1.3 Medie avanzate: media aritmetica, media geometrica e media armonica

Manuale: STATS. ≥ SUMMAR. ≥ SUM.STATS. ≥ GEOM/ARITH/...
Comando: AMEANS variabili

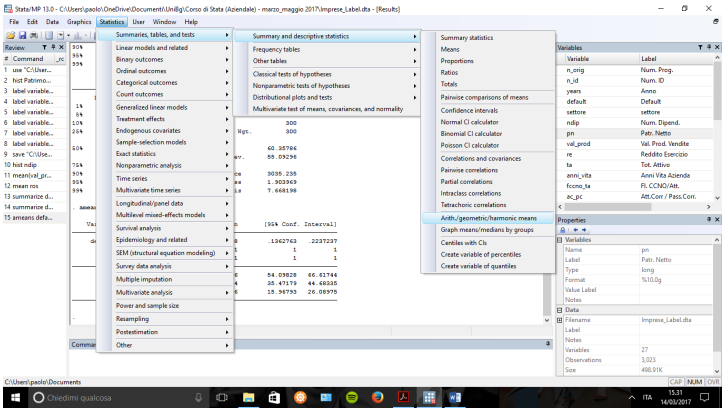


Figura 16: Medie avanzate

5.1.4 Descrittive con raggruppamento

Manuale: STATS. ≥ SUMMAR. ≥ SUM.STATS. ≥ IF/IN/
Comando: BY var.raggrupp., SORT: SUMMARIZE var.interesse, DETAIL

```
. by years, sort : summarize default ndip

-------------------------------------------------------------------------------
-> years = 2007

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     default |       150         .18    .3854745          0          1
       ndip  |       150     60.4052    54.45794    1.98125        340

-------------------------------------------------------------------------------
-> years = 2008

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     default |       150         .18    .3854745          0          1
       ndip  |       150    60.31052     55.9032   1.120139        340
```

Figura 17: Descrittive per gruppi

5.1.5 Tabella compatta di statistiche descrittive

Manuale: STATS. ≥ SUMMAR.TABS. ≥ OTHER TABS. ≥ COMPACT
Comando: TABSTAT variabili, STATISTICS(MEAN RANGE MAX MIN SD CV SKEWNESS KURTOSIS)

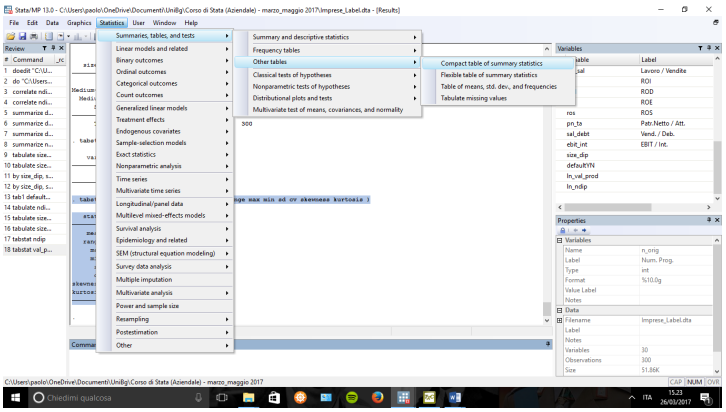


Figura 18: Descrittive compatte (1)

```
. tabstat val_prod ndip, statistics( mean range max min sd cv skewness kurtosis )

    stats |   val_prod       ndip
----------+--------------------
     mean |   1.66e+07      60.34
    range |   4.29e+07        339
      max |   4.70e+07        340
      min |    410000          1
       sd |   1.06e+07   55.11075
       cv |   .6387468   .913337
 skewness |   1.105914   1.902112
 kurtosis |   3.228197   7.662135
----------------------------------
```

Figura 19: Descrittive compatte (2)

5.1.6 Statistiche descrittive raggruppate

Manuale: STATS. $\geq$ SUMMAR.TABS. $\geq$ OTHER TABS. $\geq$ COMPACT
Comando: TABULATE var.raggrupp., SUMMARIZE(var.interesse)

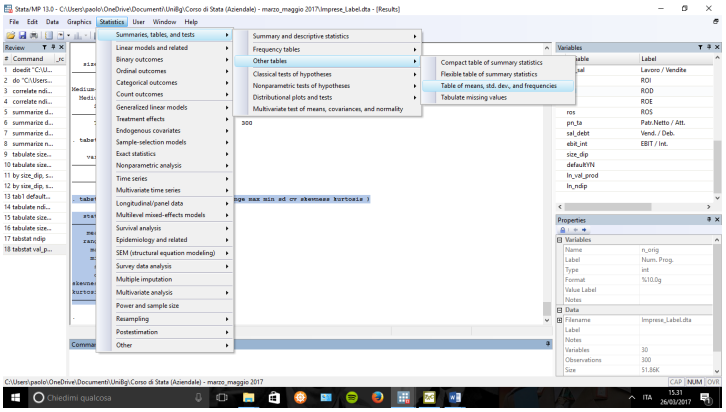


Figura 20: Descrittive per modalita' (1)

```
. tabulate size_dip, summarize(val_prod)
```

size_dip	Summary of Val. Prod. Vendite Mean	Std. Dev.	Freq.
Big	31888889	11952452	9
Medium-Down	17463636	9339307.1	66
Medium-Up	22419048	12885674	42
Small	14148634	9136793.1	183
Total	16568000	10582758	300

Figura 21: Descrittive per modalita' (2)

5.1.7 Covarianze e Correlazioni

Manuale: STATS. $\geq$ SUMMAR. $\geq$ CORR.COV.
Comando (CORR.): CORRELATE variabili
Comando (COV.): CORRELATE variabili, COVARIANCE

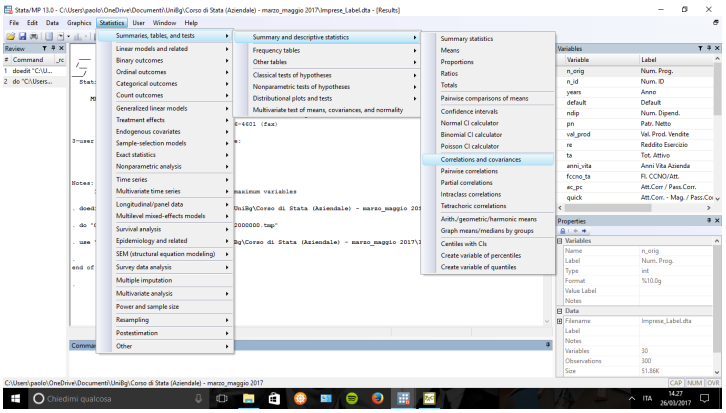


Figura 22: Procedimento correlazione

```
. correlate ndip pm val_prod re ta anni_vita fccno_ta quick ac_pc wc_ta liq_ta
(obs=300)
```

	ndip	pm	val_prod	re	ta	anni_v~a	fccno_ta	quick	ac_pc	wc_ta	liq_ta
ndip	1.0000										
pm	0.1996	1.0000									
val_prod	0.3665	0.4029	1.0000								
re	-0.0648	0.4221	0.0675	1.0000							
ta	0.2922	0.9190	0.5348	0.2796	1.0000						
anni_vita	-0.0213	-0.0717	0.0423	-0.0746	-0.0369	1.0000					
fccno_ta	0.0072	-0.0348	-0.0699	0.2802	-0.1021	0.0310	1.0000				
quick	-0.0293	0.1058	-0.0661	0.1429	-0.0040	-0.2107	0.1173	1.0000			
ac_pc	-0.0712	0.0036	-0.0295	0.0941	-0.0864	-0.0612	0.1350	0.5674	1.0000		
wc_ta	0.0155	-0.1674	0.0481	-0.1223	-0.1126	-0.0159	-0.0199	0.0894	0.1451	1.0000	
liq_ta	-0.0286	-0.0206	-0.0546	0.1451	-0.0823	-0.1477	0.0781	0.2340	0.2535	0.1297	1.0000

Figura 23: Matrice correlazione

```
. correlate ndip pm val_prod re ta anni_vita fccno_ta quick ac_pc wc_ta liq_ta, covariance
(obs=300)
```

	ndip	pm	val_prod	re	ta	anni_v~a	fccno_ta	quick	ac_pc	wc_ta	liq_ta
ndip	3037.2										
pm	1.1e+08	9.5e+13									
val_prod	2.1e+08	4.2e+13	1.1e+14								
re	-6.6e+06	7.7e+12	1.3e+12	3.4e+12							
ta	3.4e+08	1.9e+14	1.2e+14	1.1e+13	4.6e+14						
anni_vita	-39.3722	-2.4e+07	1.5e+07	-4.6e+06	-2.7e+07	1127.13					
fccno_ta	.023175	-19987.6	-43480.7	30582.8	-128578	.061206	.003455				
quick	-124919	8.0e+10	-5.4e+10	2.1e+10	-6.7e+09	-547950	533.97	6.0e+09			
ac_pc	-346275	3.1e+09	-2.8e+10	1.5e+10	-1.6e+11	-181079	699.804	3.9e+09	7.8e+09		
wc_ta	.278883	-533304	166164	-74163.6	-787791	-.174489	-.000381	2262.59	4178.58	.106628	
liq_ta	-.146723	-10735.4	-53803.9	25070.5	-164001	-.461366	.000427	1686.44	2080.08	.003942	.008657

Figura 24: Matrice varianza-covarianza

5.2 Tabelle di frequenza

Altro importante step nell'analisi descrittiva di un fenomeno e' la costruzione delle tabella di frequenza, che permettono di rappresentare le variabili associando ad ogni modalita' (tutti i modi differenti in cui si presenta un carattere) le rispettive FREQUENZE ASSOLUTE (conteggio fisico) e le FREQUENZE RELATIVE (% rispetto al totale delle osservazioni).

5.2.1 Tabelle frequenza assoluta e relativa singole

Manuale: STATS. $\geq$ SUMMAR. $\geq$ FREQ.TABLES $\geq$ ONE WAY TAB.
Comando: TABULATE variabile
Stata genera automaticamente in output una tabella con Frequenze assolute, Frequenze relative e Frequenze relative cumulate.
Come per le precedenti statistiche descrittive, le tabelle possono essere riprodotte usando variabili di raggruppamento oppure selezionando solo una certa parte dei dati.

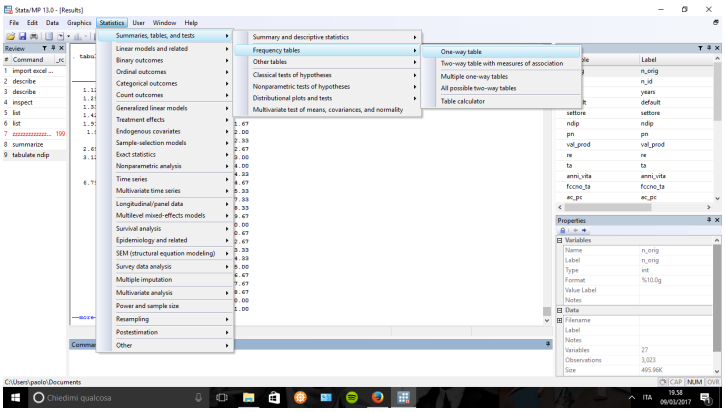


Figura 25: Tabella di frequenza (1)

```
. tabulate size_dip
```

size_dip	Freq.	Percent	Cum.
Big	9	3.00	3.00
Medium-Down	66	22.00	25.00
Medium-Up	42	14.00	39.00
Small	183	61.00	100.00
Total	300	100.00	

Figura 26: Tabella di frequenza (2)

5.2.2 Tabelle frequenza multiple

Manuale: STATS. ≥ SUMMAR. ≥ FREQ.TABLES ≥ MULTITPLE ONE WAY TAB.
Comando: TAB1 variabili
Stata genera automaticamente in output un numero di tabelle di frequenza pari al numero di variabili prese in considerazione.

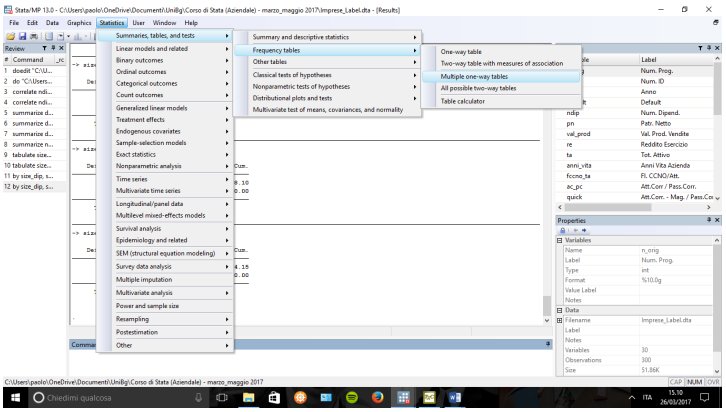


Figura 27: Tabelle di frequenza multiple (1)

```
. tabl defaultYN size_dip

-> tabulation of defaultYN

  defaultYN |      Freq.     Percent        Cum.
------------+-----------------------------------
         No |        246       82.00       82.00
        Yes |         54       18.00      100.00
------------+-----------------------------------
      Total |        300      100.00

-> tabulation of size_dip

   size_dip |      Freq.     Percent        Cum.
------------+-----------------------------------
        Big |          9        3.00        3.00
Medium-Down |         66       22.00       25.00
  Medium-Up |         42       14.00       39.00
      Small |        183       61.00      100.00
------------+-----------------------------------
      Total |        300      100.00
```

Figura 28: Tabelle di frequenza multiple (2)

5.2.3 Tabelle frequenza a doppia entrata (senza misure associazione)

Manuale: STATS. ≥ SUMMAR. ≥ FREQ.TABLES ≥ TWO WAY TAB.
Comando: TABULATE var1 var2

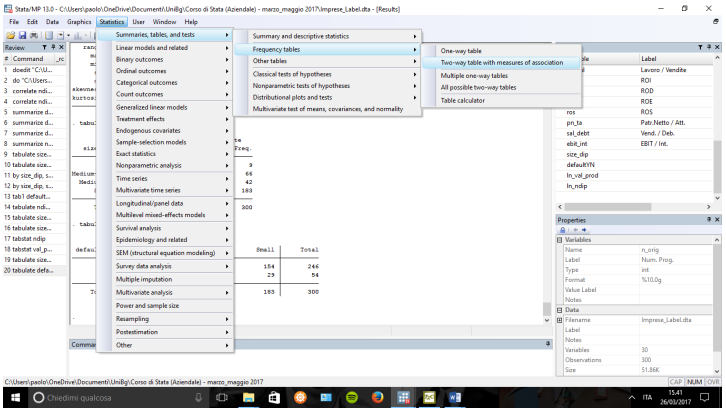


Figura 29: Tabelle di frequenza a doppia entrata (1)

```
. tabulate defaultYN size_dip
```

defaultYN	size_dip				
	Big	Medium-..	Medium-Op	Small	Total
No	7	48	37	154	246
Yes	2	18	5	29	54
Total	9	66	42	183	300

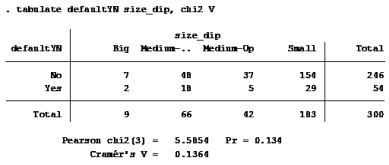
Figura 30: Tabelle di frequenza a doppia entrata (2)

5.2.4 Tabelle frequenza a doppia entrata con misure associazione

Manuale: STATS. $\geq$ SUMMAR. $\geq$ FREQ.TABLES $\geq$ TWO WAY TAB.
Comando: TABULATE var1 var2, CHI2 V
Questa aggiunta permette a Stata di calcolare direttamente due misure di associazione tra i caratteri X e Y selezionati: l'Indice di Chi Quadrato e la sua trasformazione, l'indice V di Cramèr.
Alla statistica di Chi Qaudrato e' associato una verifica/test di ipotesi, riportata nella seguente tabella:

Statistica	H_0	H_1	*Criterio*
Chi 2	X e Y indipend.	X e Y dipendenti	PV $\leq 0.05 \Rightarrow$ Rifiuto H_0

Tabella 2: Test ipotesi di indipendenza

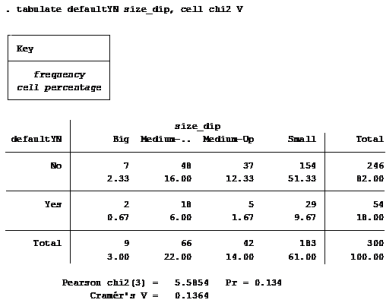


```
. tabulate defaultYN size_dip, chi2 V

           |                size_dip
defaultYN  |       Big  Medium-..  Medium-Up     Small |     Total
-----------+--------------------------------------------+----------
        No |         7        48        37       154 |       246
       Yes |         2        18         5        29 |        54
-----------+--------------------------------------------+----------
     Total |         9        66        42       183 |       300

          Pearson chi2(3) =   5.5854   Pr = 0.134
               Cramér's V =   0.1364
```

Figura 31: Tabelle di frequenza a doppia entrata assoc. (1)

5.2.5 Tabelle a doppia entrata con freq.relative totali

Manuale: STATS. $\geq$ SUMMAR. $\geq$ FREQ.TABLES $\geq$ TWO WAY TAB.
Comando: TABULATE var1 var2, CELL

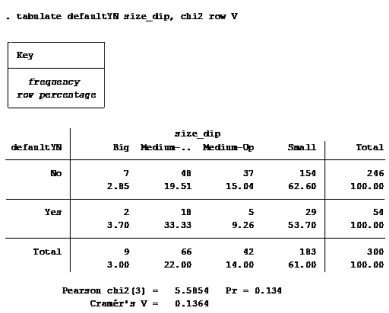


```
. tabulate defaultYN size_dip, cell chi2 V

+-----------------+
| Key             |
|-----------------|
|   frequency     |
| cell percentage |
+-----------------+

           |                size_dip
defaultYN  |       Big  Medium-..  Medium-Up     Small |     Total
-----------+--------------------------------------------+----------
        No |         7        48        37       154 |       246
           |      2.33     16.00     12.33     51.33 |     82.00
-----------+--------------------------------------------+----------
       Yes |         2        18         5        29 |        54
           |      0.67      6.00      1.67      9.67 |     18.00
-----------+--------------------------------------------+----------
     Total |         9        66        42       183 |       300
           |      3.00     22.00     14.00     61.00 |    100.00

          Pearson chi2(3) =   5.5854   Pr = 0.134
               Cramér's V =   0.1364
```

Figura 32: Tabelle di frequenza a doppia entrata relative totali

5.2.6 Tabelle a doppia entrata con freq.relative per riga

Manuale: STATS. ≥ SUMMAR. ≥ FREQ.TABLES ≥ TWO WAY TAB.
Comando: TABULATE var1 var2, ROW



```
. tabulate defaultYN size_dip, chi2 row V
```

Key
frequency *row percentage*

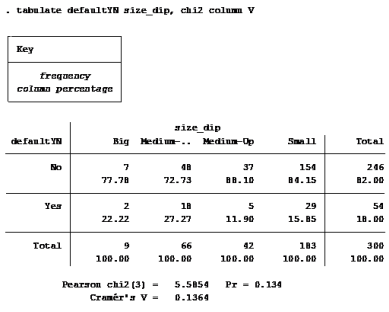
defaultYN	size_dip				Total
	Big	Medium-..	Medium-Up	Small	
No	7	48	37	154	246
	2.85	19.51	15.04	62.60	100.00
Yes	2	18	5	29	54
	3.70	33.33	9.26	53.70	100.00
Total	9	66	42	183	300
	3.00	22.00	14.00	61.00	100.00

```
          Pearson chi2(3) =   5.5854   Pr = 0.134
               Cramér's V =   0.1364
```

Figura 33: Tabelle di frequenza a doppia entrata relative per righe

5.2.7 Tabelle a doppia entrata con freq.relative per colonna

Manuale: STATS. ≥ SUMMAR. ≥ FREQ.TABLES ≥ TWO WAY TAB.
Comando: TABULATE var1 var2, COLUMN



```
. tabulate defaultYN size_dip, chi2 column V
```

Key
frequency *column percentage*

defaultYN	size_dip				Total
	Big	Medium-..	Medium-Up	Small	
No	7	48	37	154	246
	77.78	72.73	88.10	84.15	82.00
Yes	2	18	5	29	54
	22.22	27.27	11.90	15.85	18.00
Total	9	66	42	183	300
	100.00	100.00	100.00	100.00	100.00

```
          Pearson chi2(3) =   5.5854   Pr = 0.134
               Cramér's V =   0.1364
```

Figura 34: Tabelle di frequenza a doppia entrata relative per colonne

6 Rappresentazioni grafiche

Stata mette a disposizione dell'utilizzatore una notevole gamma di grafici, ognuno dei quali ha caratteristiche ed utilizzi specifici.
Ogni grafico puo' essere personalizzato applicando il titolo, intestazione degli assi e colori.

Parametro grafico	*Aggiunta*	*Note*
BIN(numero)	Definisce numero di classi	Istogramma
NOAXIS	Elimina gli assi cartesiani	
YTITLE(...)	Nome asse Y	
XTITLE(...)	Nome asse X	
XLABEL(...)	Etichette asse X	Scatter
YLABEL(...)	Etichette asse Y	Scatter
TITILE(...)	Titolo del grafico	
NORMAL	Aggiunge grafico di una Normale	Istogramma e Barre
FCOLOR(...)	Colore delle barre o dei puntini	
LCOLOR(...)	Colore dei contorni delle barre	
SYMBOL(o/t/s)	(o)pallini (t)triangoli (s)quadrati	Scatter
CONNECT(l/m/s)	Unisce i puntini (l)con un retta (m)unisce i punti medi (s)unisce con una curva	Scatterplot
BY(var.raggrup.)	Divide grafico in base ad una var. categorica	
PLABEL(...)	Etichette della torta	Grafico a torta

Tabella 3: Parametri grafici per istogramma

6.1 Grafici di distribuzioni

6.1.1 Caratteri quantitativi continui: Istogramma

Per rappresentare le distribuzioni di frequenza di caratteri quantitativi continui e' opportuno utilizzare un istogramma, un grafico che utilizza sull'asse X (ascisse) le classi di modalita' e sull'asse delle Y (ordinate) la densita' di frequenza.
La densita' di frequenza e' data dal rapporto tra frequenze assolute di ogni classe e ampiezza della classe.

$$Dens.Classe = \frac{Freq.Ass.Classe}{Ampiezza.Classe}$$

Manuale: GRAPHICS $\geq$ HISTOGRAM
Comando: HISTOGRAM variabile
E' possibile aggiungere una serie di parametri grafici per personalizzare l'aspetto del grafico a nostro piacere.

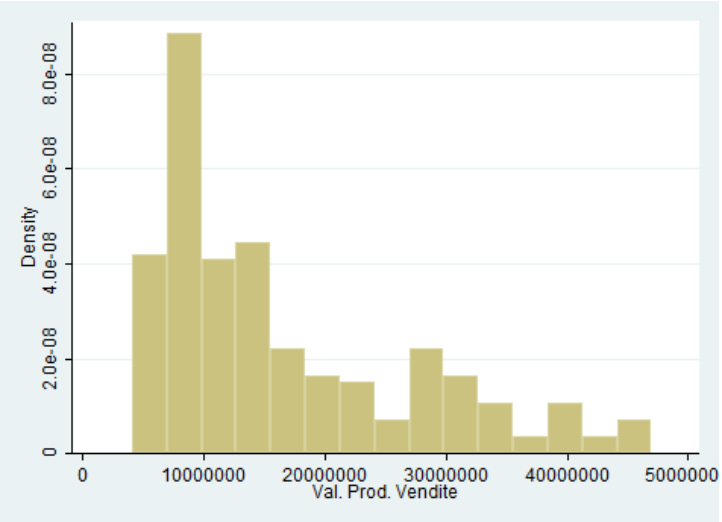


Figura 35: Istogramma

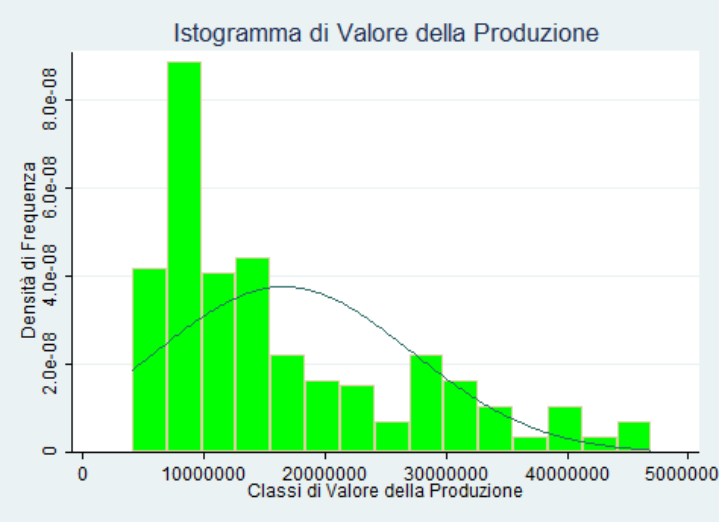


Figura 36: Istogramma personalizzato

Introducendo un altra parametro e' possibile creare piu' istogrammi usando una variabile categorica di raggruppamento.
Comando: HISTOGRAM variabile, BY(var.raggrupp.)

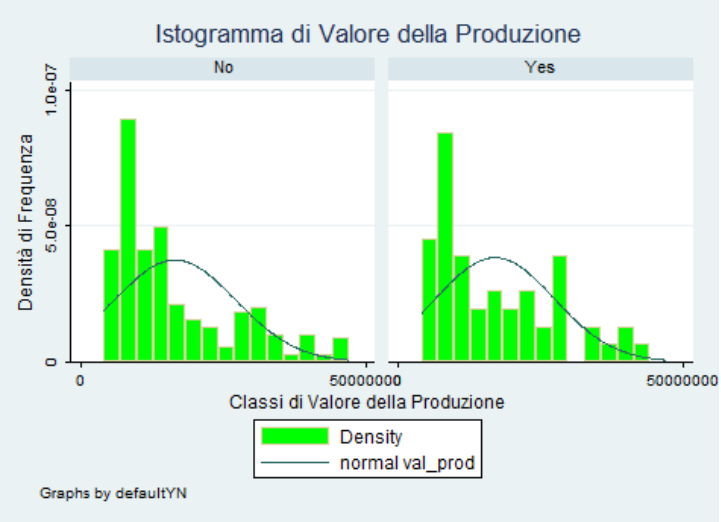


Figura 37: Istogramma personalizzato

6.1.2 Caratteri quantitativi discreti: Grafico a bastoncini

Per rappresentare le distribuzioni di frequenza di caratteri quantitativi discreti si utilizza il grafico a bastoncini, che utilizza sull'asse X (ascisse) le singole modalita' e sull'asse delle Y (ordinate) la frequenza assoluta o relativa.
Manuale: GRAPHICS $\geq$ HISTOGRAM
Comando: HISTOGRAM variabile, DISCRETE PERCENT

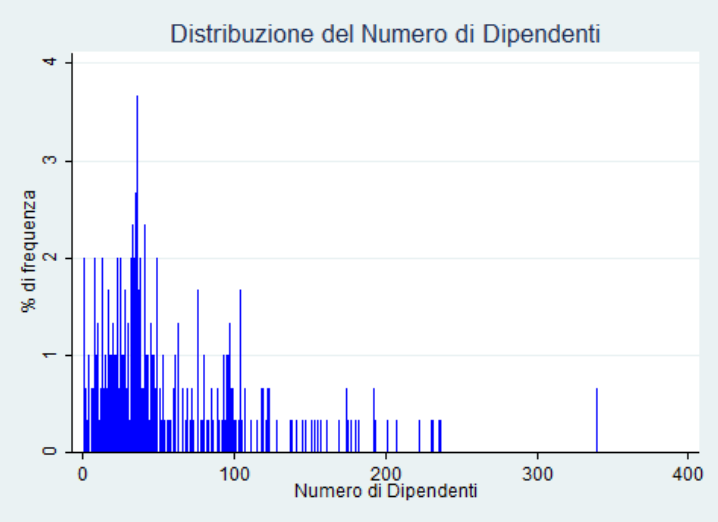


Figura 38: Grafico a bastoncini

6.1.3 Caratteri qualitativi (1): Grafico a barre

Per rappresentare le distribuzioni di frequenza di caratteri qualitativi si utilizza il grafico a barre, molto simile al grafico a bastoncini, che utilizza sull'asse X (ascisse) le singole modalita' qualitative e sull'asse delle Y (ordinate) la frequenza assoluta o relativa.
Manuale: GRAPHICS $\geq$ BAR CHART
Comando: GRAPH BAR (COUNT) var.qualsiasi, OVER(var.interesse)

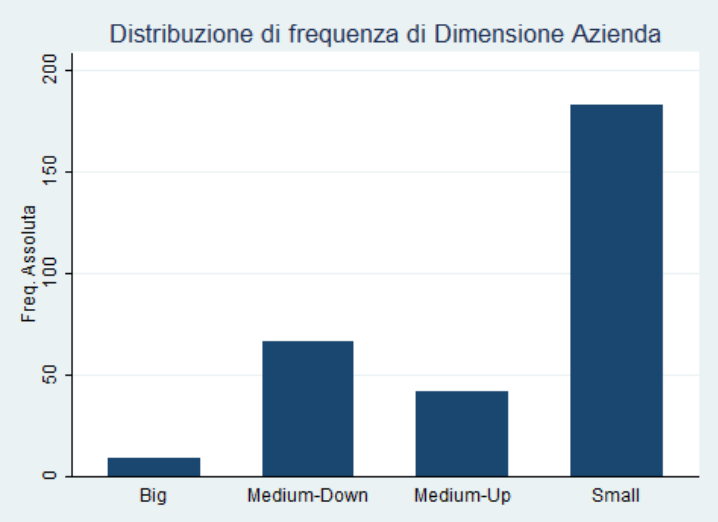


Figura 39: Grafico a barre

6.1.4 Caratteri qualitativi (2): Grafico a torta

Un secondo grafico utilie a rappresentare le distribuzioni di frequenza di caratteri qualitativi e' il grafico a barre, che divide un cerchio/torta (100%) in fette corrispondenti alle modalita' di una variabile qualitativa.
Manuale: GRAPHICS $\geq$ PIE CHART
Comando: GRAPH PIE, OVER(var.interesse)

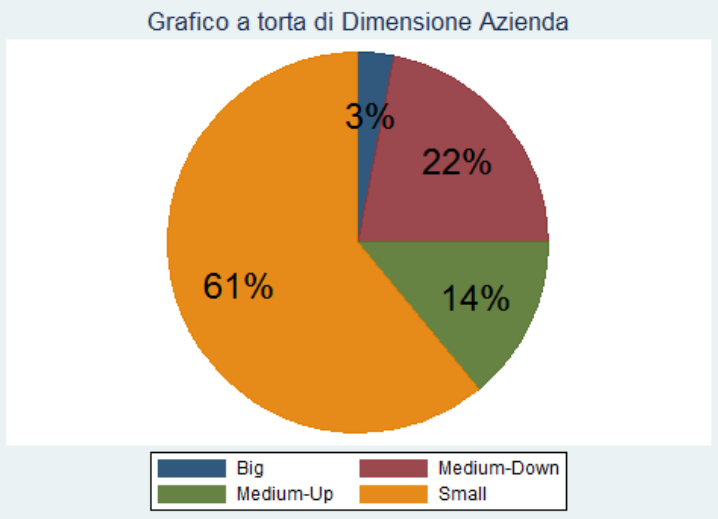


Figura 40: Grafico a torta

6.1.5 Box Plot

Il box plot e' la rappresentazione grafica dei percentili di una distribuzione di frequenza: e' composto dal valore minimo (percentile 0), dal 25 percentile, dalla mediana (50 percentile), dal 75 percentile e dal valore massimo (100 percentile).
Il 25 ed il 75 percentile sono uniti tramite una scatola, a sua volta divisa dalla mediana; mentre i due valori estremi sono uniti al resto tramite delle rette chiamate "baffi".
Grafico utile per identificare eventuali asimmetrie positive (baffo destro o superiore piu' lungo) e asimmetrie negative (baffo sinistro o inferiore piu' lungo) oppure per identificare valore anomali outlier.
E' possibile confrontare piu' variabili costruendo box-plots vicini nello stesso grafico.
Manuale: GRAPHICS $\geq$ BOX PLOT
Comando: GRAPH BOX variabili

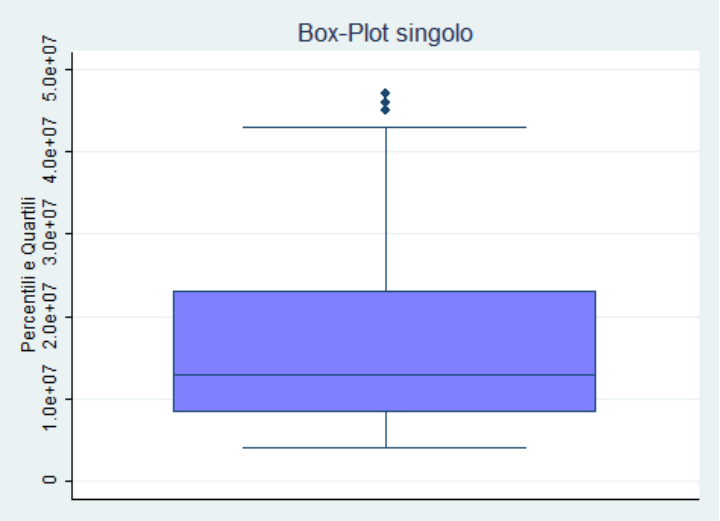


Figura 41: Box-plot singolo

6.1.6 Quantile-Quantile Plot

Il QQ Plot confronta le distribuzioni di due variabili verificando la distanza tra i percentili. Su un asse vengono proiettati i percentili di una delle due variabili e sull'altro asse si proiettano i percentili dell'altra.
Se i grafici presentano distribuzioni simili allora i puntini si distribuiranno vicini alla retta centrale, in caso opposto sara' chiaro che le variabili presentano distribuzioni molto diverse.
Manuale: GRAPHICS $\geq$ SYM. PLOTS $\geq$ QQ PLOT

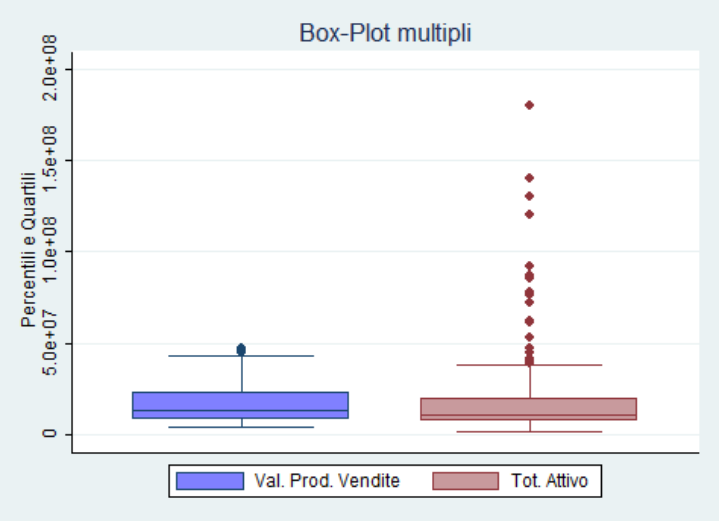


Figura 42: Box-plot multiplo

Comando: QQPLOT variabili

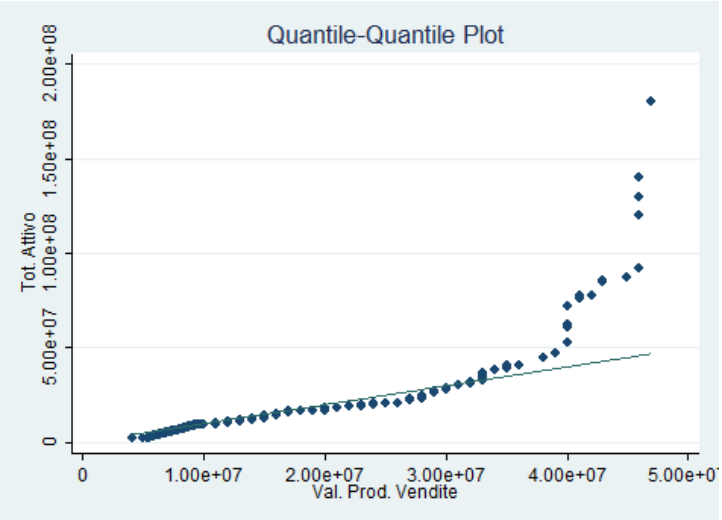


Figura 43: Quantile-Quantile plot

6.1.7 Normality Quantile-Quantile Plot

Il Normality Q-Q Plot e' un grafico che permette di confrontare la distribuzione di frequenza osservata nei dati con quella di una distribuzione teorica di tipo Normale, per definizione simmetrica e "Normale".
Sull'asse delle X sono rappresentanti i percentili della distribuzione teorica

Normale e sull'asse Y i percentili della nostra variabile.
Se i puntini si distribuiscono uniformemente vicino alla retta teorica significa che la nostra distribuzione e' simmetrica e a forma di campana, quindi Normale. Se invece la forma evidenziata e' una S, abbiamo a che fare con una distribuzione non simmetrica e iponormale. Se infine la forma si una S rovesciata potremmo avere a che fare con una distribuzione asimmetrica e ipernormale.
Manuale: GRAPHICS $\geq$ SYM. PLOTS $\geq$ NORMAL QQ PLOT
Comando: QNORM variabile

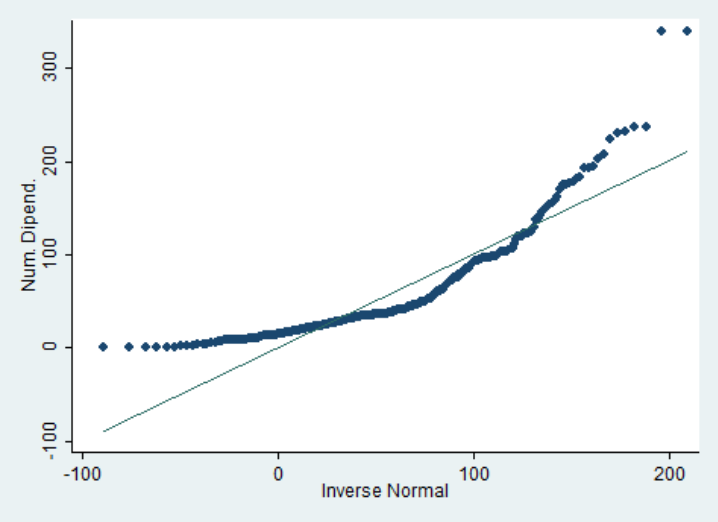


Figura 44: Normal Quantile-Quantile plot

6.1.8 Simmetry Plot

Il grafico di simmetria valuta le distanze tra le frequenze osservate nei nostri dati e la mediana della distribuzione.
Se i puntini si distribuiscono in modo evidente vicino alla retta teorica di simmetria significa che la nostra distribuzione e' molto vicina alla simmetria; se invece i punti giacciono al di sopra della retta abbiamo a che fare con una asimmetria positiva (di destra); mentre quando i puntini si distribuiscono al di sotto abbiamo una evidente asimmetria negativa (a sinistra).
Manuale: GRAPHICS $\geq$ DISTRIB. GRAPH. $\geq$ SYMMETRY
Comando: SYMPLOT variabile

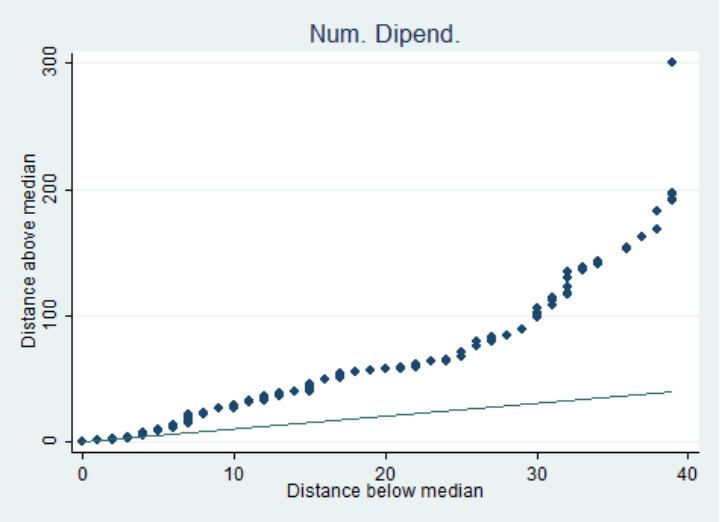


Figura 45: Grafico di simmetria

6.2 Grafici avanzati

6.2.1 Grafico a dispersione - Scatterplot

La rappresentazione congiunta di due variabili quantitative avviene tramite il grafico a dispersione (o scatterplot). Il grafico pone sull'asse delle ascisse le modalita' della variabile X, mentre sull'asse delle ordinate inserisce le modalita' della variabile Y.
I puntini possono essere interpretate come coordinate (X, Y) ed indicano le frequenze congiunte delle due variabili. **Manuale**: GRAPHICS $\geq$ TWO WAY $\geq$ SCATTER
Comando: TWOWAY (SCATTER x y)

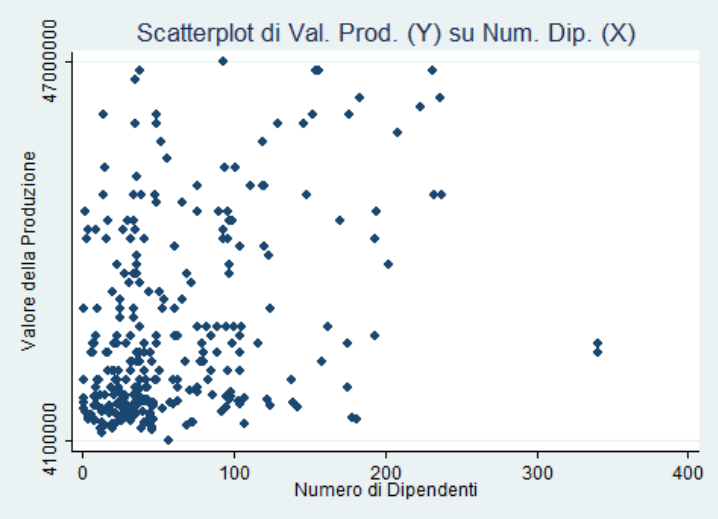


Figura 46: Grafico a dispersione

6.2.2 Matrice degli scatterplot

Un interessante grafico che permette di visualizzare la relazione lineare (o non lineare) tra diverse variabili e' la matrice degli scatterplot: un grafico che crea una matrice di ordine pari al numero di variabili e nelle cui celle sono presenti i grafici a dispersione tra tutte le coppie X e Y. **Manuale**: GRAPHICS $\geq$ SCATTERPLOT MATRIX
Comando: GRAPH MATRIX variabili

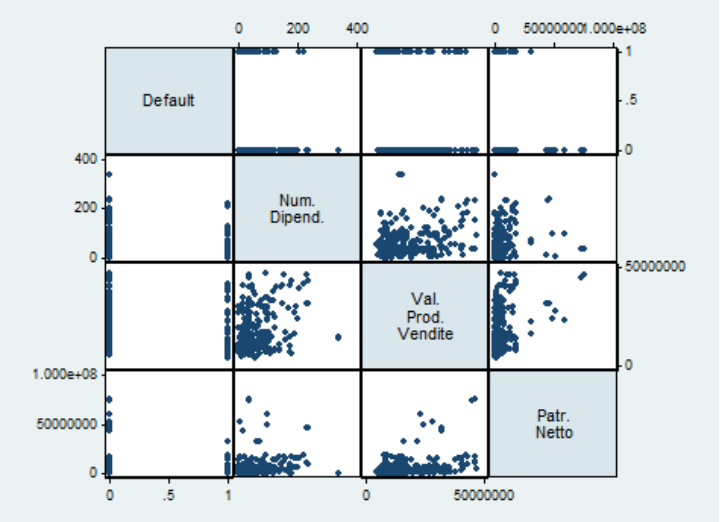


Figura 47: Matrice degli scatterplot

7 Introduzione alla Regressione Lineare semplice

7.1 Regressione lineare e applicazioni economiche

La regressione lineare e' una tecnica che ha lo scopo di studiare/investigare le relazioni empiriche (supportate dai dati) tra variabili diverse in modo da stabilire l'impatto di uno o piu' variabili indipendenti X_1, X_2, X_3, ... , X_k dette regressori su una variabile dipendente Y.

Quando nel modello viene inserita una sola variabile indipendente X, il modello prende il nome di regressione lineare semplice (lineare nei parametri) e si presenta a livello teorico nella seguente forma:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

Dove β_1 rappresenta un parametro coefficiente angolare ignoto da stimare, β_0 rappresenta l'intercetta ignota da stimare ed ϵ_i indica l'errore stocastico che si commette nella stima. Il pedice "i" indica il contatore delle osservazioni.
Il primo, e sicuramente tra i piu' complicati e importanti, passo da compiere per stimare un modello di regressione lineare semplice e' raccogliere quanti piu' possibili dati su entrambe le variabili considerate nell'analisi e costruire un dataset che raccolga le osservazioni.

Utilizzare Stata per compiere una regressione lineare e' molto semplice: dopo aver caricato il dataset nella console e aver verificato la correttezza dei dati (pulizia, formato, tipologia, ...) si utilizza il semplice comando REGRESS oppure si procede tramite il menu' in alto.

Manuale: STATISTICS $\geq$ LINEAR MODEL $\geq$ LINEAR REG.
Comando: REGRESS var.Y var.X

7.2 Le tabelle dei risultati

Le successive figure mostrano un esempio di risultato ottenuto tramite la procedura di regressione lineare semplice; possiamo suddividere l'output in 3 tabelle principali:

1. Tabella ANOVA - Analysis Of Variance (in alto a sinistra): una tabella in cui vengono mostrate la Varianza dei Residui (Residual SS), la Varianza del Modello (Model SS) e la Varianza totale data dalla loro somma e i rispettivi gradi di liberta'. Queste statistiche sono necessarie per il calcolo dell'indice di determinazione R^2 e del test F di significativita'.

2. Tabella dei COEFFICIENTI (centrale): tabella che mostra le stime dei coefficienti di regressione, i test-T di significativita' e i relativi P-Value.

3. Tabella di SINTESI (in alto a destra): tabellina con riportate alcune importanti misure di diagnostica e di sintesi.

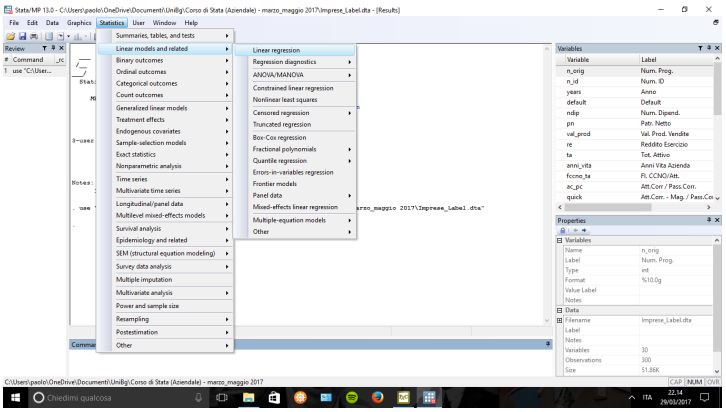


Figura 48: Regressione lineare semplice (1)

```
. regress ln_val_prod ln_ndip

      Source |       SS       df       MS              Number of obs =     300
-------------+------------------------------           F(  1,   298) =   30.10
       Model |  9.7865182     1  9.7865182              Prob > F      =  0.0000
    Residual |  96.8977516    298  .32516024            R-squared     =  0.0917
-------------+------------------------------           Adj R-squared =  0.0887
       Total |  106.68427     299  .356803578           Root MSE      =  .57023

------------------------------------------------------------------------------
 ln_val_prod |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     ln_ndip |   .1729737   .0315293     5.49   0.000     .1109254    .2350221
       _cons |   15.80355   .120448    131.21   0.000     15.56651    16.04059
------------------------------------------------------------------------------
```

Figura 49: Regressione lineare semplice - Risultati

Tabella Risultati - Coefficienti

Elemento	*Interpretazione*
Coef.	Valore stimato del coefficiente corrispondente alla variabile sulla riga
Std.Err.	Standard Error della stima calcolata
t	Statistica test t per signif. variabile $\Rightarrow$ t=$\frac{Coef.}{Std.Err.}$
	$H_0 : \beta_1 = 0$
P>t	P-Value della statistica t:
	$\Rightarrow$ PV$\leq$0.05 $\Rightarrow$ Variabile significativa nel modello
	$\Rightarrow$ PV$\geq$0.05 $\Rightarrow$ Variabile NON significativa nel modello
Conf. Interval	Intervallo confidenza 95% per coefficiente stimato

Tabella di sintesi

Elemento	*Interpretazione*
Number of obs.	Numero osservazioni campione (n)
F(k-1;n-k)	Statistica test F per "zero slopes" ($R^2 = 0$)
	H_0: $\beta_1 = 0$ $\Rightarrow$ (qui uguale al test-t)
Prob>F	P-Value della statistica F:
	$\Rightarrow$ PV$\leq$0.05 $\Rightarrow$ Modello significativo nel suo complesso
	$\Rightarrow$ PV$\geq$0.05 $\Rightarrow$ Modello NON significativo nel complesso
R^2	Indice di determinazione
R^2-adj	Indice di determinazione corretto per la dimensione campione.
Root MSE	Radice quadrata della stima della varianza dell'errore

Tabella ANOVA

Elemento	*Interpretazione*
SS Model	Varianza spiegata dal modello
df Model	Gradi liberta' SSM (k-1)
SS Residual	Varianza dei residui
df Residual	Gradi liberta' SSR (n-k)
SS Total	Varianza totale di Y
df Total	Gradi liberta' totali (n-1)
MSM	Stima della varianza spiegata: $MSM = \frac{SSM}{dfM}$
MSR	Stima della varianza dei residui: $MSR = \frac{SSR}{dfR}$
MST	Stima della varianza totale: $MSM = \frac{SST}{dfT}$

7.3 Valori stimati

Spesso, per effettuare calcoli successivi e grafici dei risultati, risulta utile salvare in memoria alcuni dei risultati ottenuti con i precedenti comandi.
Ad esempio, il salvataggio dei valori stimati (o valori previsti) permette sovrapporre il grafico dei dati originari osservati e quello dei valori stimati tramite la tecnica di regressione. Grafici di questo genere danno un immediato impatto visivo del lavoro svolto e identificano eventuali punti critici.

7.3.1 Valori previsti dal modello

Una volta stimati i coefficienti di regressione $\hat{\beta}_i$ è possibile calcolare i valori previsti dal modello. Con il termine 'valori previsti' intendiamo il valore medio previsto data una specifica combinazione di valori che la variabile indipendente X_i può assumere.
I valori stimati possono essere indicati come il principale output di ogni regressione lineare in quanto sono alla base di alcuni indici che permettono di valutare l'adeguatezza del modello ai dati che stiamo considerando. Più i valori stimati $\hat{Y}_i$ si avvicinano ai valori osservati Y_i, più il modello di regressione è adatto al tipo di dato.
In simboli, i valori stimati si indicano come $\hat{Y}_i$ e sono calcolati con la seguente formula:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

La procedura Stata per salvare questi valori stimati è la seguente:

Manuale: STATISTICS $\geq$ POSTESTIMATION $\geq$ ESTIMATES
Comando: PREDICT nome.variabile, XB

7.3.2 Residui della regressione

I residui $\hat{\varepsilon}_i$ hanno un'interpretazione antitetica a quella dei valori stimati: essi misurano la distanza tra i valori osservati e quelli previsti e più grandi sono, più grande sarà l'errore commesso complessivamente dal modello nel prevedere la variabile dipendente. I residui sono calcolati come differenza tra i valori stimati e quelli previsti:

$$\hat{\varepsilon}_i = Y_i - \hat{Y}_i$$

Data la loro strategica importanza, i residui di una regressione vengono analizzati in modo approfondito tramite alcune tecniche, che prendono il nome di 'Test di corretta specificazione del modello', in grado di fornire una valutazione delle sue caratteristiche statistiche e soluzioni ad eventuali violazioni.

La procedura Stata per salvare questi valori stimati è la seguente:

Manuale: STATISTICS ≥ POSTESTIMATION ≥ RESIDUALS
Comando: PREDICT nome.variabile, RESIDUALS

7.4 Analisi dei residui: test di (s)corretta specificazione

Come detto in precedenza, l'analisi statistica dei residui della regressione permette di stabilire se il modello rispetta le assunzioni teoriche iniziali e quindi permette di analizzare la relazione tra X e Y in modo adeguato.
Tre delle assunzioni principali che i modelli di regressione lineare introducono, riguardano la struttura dei residui:

- Omoschedasticità: la varianza dei residui deve essere costante per tutte le osservazioni di X (la variabilità di Y non deve variare in funzione della variabile indipedente);
- Normalità: la distribuzione (istogramma) dei residui deve essere simile a quello di una Normale (code leggere e forte maggioranza attorno alla media);
- Assenza di variabili omesse: la regressione deve tenere conto di tutte le variabili che spiegano Y (nella regressione devo includere ogni variabile considerata in un modello teorico. Ex: la curva Cobb-Douglas considera la produzione Y come funzione del fattore lavoro L e fattore capitale K).

Dalla console dei comandi è possibile testare numerose ipotesi, oltre a quelle citate, seguendo la breve procedura:
Manuale: STATS. ≥ LINEAR MODEL ≥ REGR. DIAGNOSTICS ≥ SPEC.TESTS

7.4.1 Test di Normalita' e di Simmetria di Bera-Jarque

Il test di Normalita' di B-J permette di costruire una verifica di ipotesi allo scopo di indagare la Normalita' di una distribuzione, intesa come combinazione di simmetria (skewness) e pesantezza delle code (kurtosis).
H_0: Normalita' della distribuzione (Simmetria + Campana)
H_1: NON normalita' della distribuzione
Comando: SKTEST variabile
P-Value $\leq 0.05 \Rightarrow$ Rifiuto H0 $\Rightarrow$ Normalita' della distribuzione.

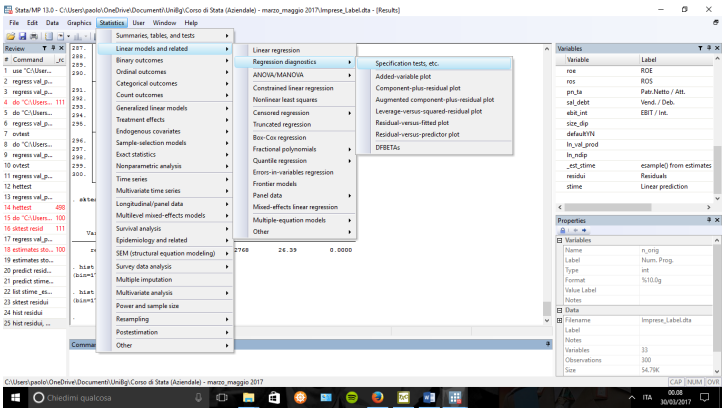


Figura 50: Diagnostiche modello regressione

7.4.2 Test di eteroschedasticita' di White

Il test di ipotesi sull'eteroschedasticita' permette di testare l'ipotesi nulla che il modello specificato abbia una varianza costante (omoschedasticita') contro l'ipotesi alternativa di varianza non costante (eteroschedasticita').
H_0: Omoschedasticita'
H_1: Eteroschedasticita'
Comando: HETTEST variabile
P-Value $\leq 0.05 \Rightarrow$ Rifiuto H0 $\Rightarrow$ Eteroschedasticita' del modello.

7.4.3 Test RESET di Ramsey

Il test di Ramsey (Regression Specification Error Test) permette di testare l'ipotesi nulla che il modello specificato NON abbia dimenticato variabili omesse.
H_0: Il modello NON ha variabili omesse
H_1: Il modello HA delle variabili omesse
Segnala una eventuale scorretta specificazione ma non indica alcun modo per sistemare il problema: test NON costruttivo.
Comando: OVTEST variabile
P-Value $\leq 0.05 \Rightarrow$ Rifiuto H0 $\Rightarrow$ Esistono delle variabili omesse

8 Introduzione alla Cluster analysis (Analisi dei gruppi)

8.1 Il concetto statistico di 'gruppo'

Numerose applicazioni economiche e del mondo del lavoro analizzano dati con il fine di identificare gruppi di osservazioni (ex. gruppi di individui che consumano beni simili) che siano il più possibili eterogenei tra di loro (alta variabilità tra gruppi) e il più possibile omogenei al loro interno (bassa variabilità nei gruppi).

Possiamo quindi definire un cluster come un gruppo di osservazioni statitiche tali per cui:

- la varianza tra i gruppi (σ^2_B) sia la più alta possibile: gruppi eterogenei tra di loro;
- la varianza nei gruppi (σ^2_W) sia la più bassa possibile: osservazioni omogenee all'interno dello stesso gruppo.

N.B.: i simboli 'B' e 'W' indicano rispettivamente 'Between' (tra) e 'Within' (nei).

Dal teorema di scomposizione della varianza (ANOVA) sappiamo che la varianza di una variabile Y (*Variabile clusterizzata*) a cui si applica una divisione in gruppi tramite un'altra variabile X (*Variabile clusterizzante*) può essere scomposta nella somma della varianza nei gruppi e la varianza tra i gruppi. Formalmente:

$$\sigma^2_Y = \sigma^2_B + \sigma^2_W$$

Un esempio grafico di identificazione statistics dei gruppi è ben rappresentato dal seguente grafico:

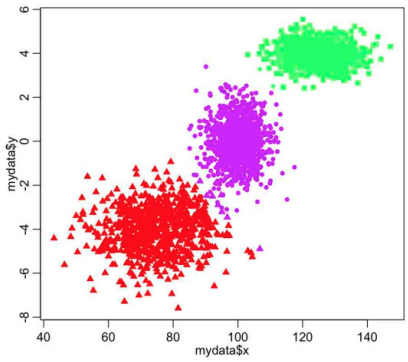


Figura 51: Clusterizzazione di Y in funzione di X

La cluster analysis può essere estesa al caso multivariato includendo più di una variabile di clsuterizzazione X_1, X_2 , ..., X_k in grado di suddividere la variabile di interesse in gruppi.

8.2 Il concetto matematico di distanza

Per raggruppare le osservazioni statistiche in gruppi è necessario definire un criterio comune che indiche quanto due unità x_i e x_j sono vicine (o lontane) e stabile poi se raggrupparle o tenerle separate.
La vicinanza o lontananza di due unità viene espresso matematicamente dal concetto di 'distanza' o 'metrica'.
Una metrica è un funzione $d(x_i, x_j)$ che rispetta alcune semplici proprietà:

1. Non negatività: $d(x_i, x_j) \geq 0 \quad \forall x_i, x_j$
2. Simmetria: $d(x_i, x_j) = d(x_j, x_i)0 \quad \forall x_i, x_j$
3. Nullità in un punto: $d(x_i, x_j) = 0 \iff x_i = x_j \quad \forall x_i, x_j$
4. Disuguaglianza triangolare: $d(x_i, x_j) \leq d(x_i, x_k)+d(x_k, x_j) \quad \forall x_i, x_j, x_k$

Esistono diverse funzioni metriche e ognuna di esse valuta la distanza tra unità in maniera differente. Esempi frequenti sono la *distanza Euclidea* (d_2), la *distanza di Manhattan* (d_1) e la *distanza di Minkowski* (d_∞).

Le tecniche di clustering permettono l'utilizzo di ognuna di queste metriche, ma prediligono come default la versione Euclidea.
la distanza Euclidea sintetizza la distanza tra due osservazioni come la radice quadrata della somma degli scarti al quadrato ottenuti per ognuna delle variabili di clusterizzazione. Supponendo di avere k variabili di raggruppamento X_1, X_2 , ..., X_K essa è pari a:

$$d_2(x_i, x_j) = \sqrt{\sum_{k=1}^{K} (x_i - x_j)^2}$$

Proprio come per la varianza, la metrica euclidea segue il principio del dar maggiore peso a quegli scarti di elevato valore tramite l'elevamento al quadrato. Tanto più è grande questa distanza e tanto meno sarà ideale unire in unico cluster due osservazioni.

8.3 Tecniche di clustering

Le metodologie di cluster analysis sviluppate in letteratura sono molteplici e sempre più sofisticate. Principalmente possono essere individuate di grandi famiglie di tecniche con diversi punti in comune, punti di forza e di debolezza.

8.3.1 Clustering gerarchico

La prima famiglia prende il nome di *clustering gerarhico* e si tratta di tecniche che raggruppano unità statistiche basandosi esclusivamente sulla distanza tra le unità senza fissare un numero predefinito di gruppi. Esempi di clustering gerarchico sono gli algoritmi agglomerativi con utilizzo del legame singolo (ad ogni iterazione si uniscono le due unità con distanza minore) e legame completo (ad ogni iterazione si uniscono le due unità con distanza maggiore). Il numero di gruppi finale viene scelto dal ricercatore utilizzando regole di decisione basati sulla distanza massima di aggregazione (ex. dendrogrammi) o su criteri del settore di applicazione (ex. non ha senso dividere i consumatori di un prodotto in 10 categorie se posso offrire solamente 4 linee diverse). Il dendrogramma è un utile strumento grafico di valutazione della cluster che mostra a quale distanza sono state aggregate le unità e in che ordine: la sua forma ad albero roversciato viene 'tagliata' nel punto in cui la distanza di aggregazione supera una certa soglia oppure è la maggiore.

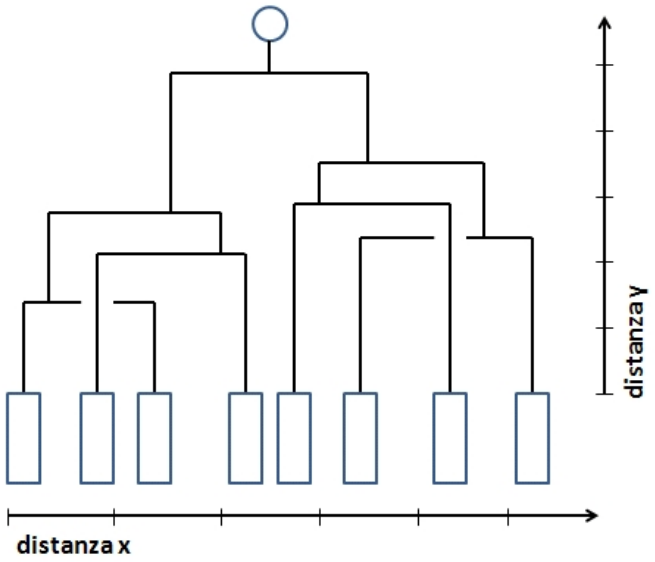


Figura 52: Esempio di dendrogramma

8.3.2 clustering non gerarchico basato sui centroidi

La seconda famiglia di metodi di clusterizzazione è denominata *clustering basato sui centroidi* in quanto utilizza come riferimento di aggregazione i valori centrali dei gruppi che si formano e di cui viene prefissato il numero.
Quando il 'centro' del gruppo è rappresentato dalla media aritmetica si parla

di centroide, mentre nel caso in cui si calcoli la mediana si parla di medioidi. La differenza tra queste versioni sta nel fatto che la media aritmetica può essere un numero reale non osservato nella realtà, mentre la mediana è sempre associata ad un valore osservato nei dati e quindi ad un membro del gruppo.

L'algoritmo non gerarchico più conosciuto prende il numero di *K-means algorithm* in quanto una volta prefissato il numero K di gruppi nella popolazione se ne calcola il valore medio centrale e lo si utilizza per valutare la distanza tra le unità.
L'algoritmo opera in modo iterativo dividendo le unità statistiche in K gruppi in modo da rispettare il criterio tale per cui la varianza interna ad ogni cluster σ^2_W sia la minore in assoluto. Questo principio, chiamato *'criterio di minima varianza di Ward'*, contemporaneamente minimizza la distanza interna e massima quella esterna, sfruttando così la proprietà di scomposizione della varianza.

Si supponga di voler dividere n osservazioni in K gruppi $c_1, ..., c_K$, ognuno di dimensione differente n_k; per ogni gruppo j viene calcolato la media $\overline{x_{\cdot j}}$ e la varianza (interna) come somma dei quadrati degli scarti delle unità i appartenenti al gruppo j (x_{ij}) dalla media di gruppo. L'obiettivo è minimizzare la media pesata delle varianze dei gruppi:

$$\min_{c_1,c_2,...,c_K} \sum_{j=1}^{K} \sum_{i=1}^{n_j} (x_{ij} - \overline{x_{\cdot j}})^2 n_j$$

Ricordiamo che:

- $n = \sum_{j=1}^{K} n_j = n_1 + n_2 + ... + n_K$
- $\overline{x_{\cdot j}} = \frac{\sum_{i=1}^{n_j} x_{ij}}{n_j}$

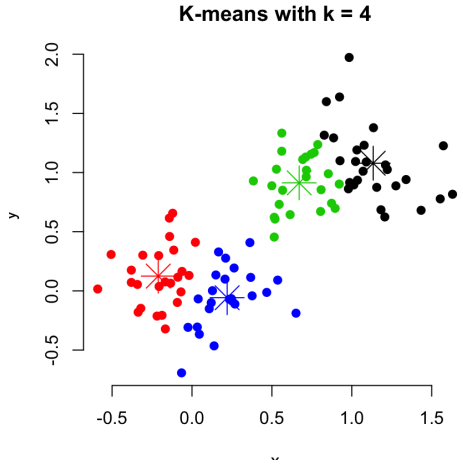


Figura 53: Esempio di K-means con 4 gruppi

Il grafico 53 evidenzia 4 gruppi distinti, al cui centro è evidenziato con una stella il valore del centroide.

8.4 Clustering con Stata

Per effettuare una clusterizzazione usando Stata si possono utilizzare sia i menù interattivi nella barra principale oppure utilizzare alcuni semplici comandi.

8.4.1 Clustering gerarchico con Stata

Utilizzando i menù è possibile effettuare tutte le tipologie classiche di clustering gerarchico attraverso la sezione *multivariate statistics*, dentro la quale troveremo l'elenco delle varie tipologie di cluster gerarchico, oppure tramite il comando *cluster* da console.
Oltre alla tipologia di cluster è possibile inserire specifiche condizioni e filtri sulle osservazioni da analizzare (ex. condizioni *if* per filtrare).

Gli algoritmi gerarchici ammessi dal software sono: *single linkage* (legame singolo, aggregazione della distanza minore), *complete linkage* (legame completo, aggregazione della distanza maggiore), *average linkage* (legame medio) e le versioni intermedie.

Manuale: STATISTICS $\geq$ MULTIVAR.STAT. $\geq$ CLUSTER
Comando: CLUSTER TipoCluster NomiVariabili

Al termine della procedura, Stata non mostrerà direttamente alcun output. Nel dataset iniziale verrà creata una nuova variabile con il gruppo di

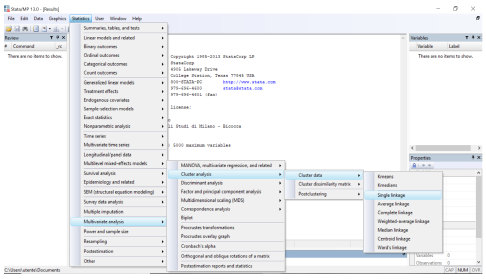


Figura 54: Clustering gerarchico con Stata

appartenenza di ogni osservazione. Per ottenere i grafici divisi per gruppi (ex. figure 51 e 53) si potranno utilizzare i classici comandi grafici visti nelle sezioni precedenti utilizzando come variabile di raggruppamento il numero del gruppo.

Per valutare le distanze di aggregazione delle unità e determinare un numero adeguato di gruppi, è possibile rappresentare graficamente il dendrogramma. Nei casi di elevato numero di osservazioni è consigliato effettuare modifiche grafiche ad etichette e colori in modo da rendere il grafico più immediato e interpretabile.

Manuale: STATISTICS $\geq$ MULTIVAR.STAT. $\geq$ POSTCLUSTERING
Comando: CLUSTER DENDROGRAM NomeClustering

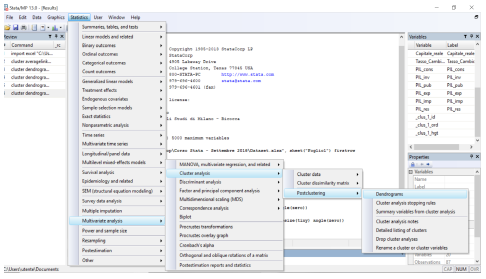


Figura 55: Dendrogramma con Stata

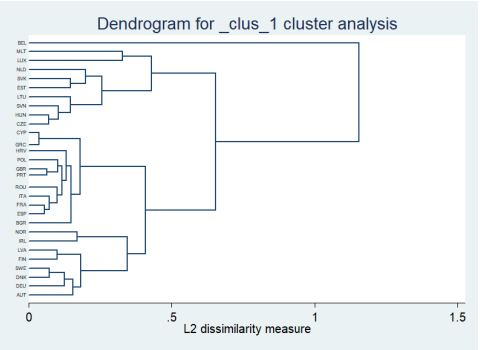


Figura 56: Esempio di dendrogramma

8.4.2 Clustering non gerarchico con Stata

Le metodologie di clustering non gerarchiche (o basate sui centroidi) in Stata seguono gli stessi comandi principali usati per quelle gerarchiche, con la differenza dei nomi.
Ad esempio, per applicare una clusterizzazione *K-means* è necessario utilizzare come TipoCluster "kmeans" (oppure cliccare K-means dal menù *cluster analysis*) e il numero di cluster che si vuole ottenere. Tutte le altre impostazioni sono comuni tra i metodi.

Manuale: STATISTICS $\geq$ MULTIVAR.STAT. $\geq$ CLUSTER
Comando: CLUSTER KMEANS NomiVariabili

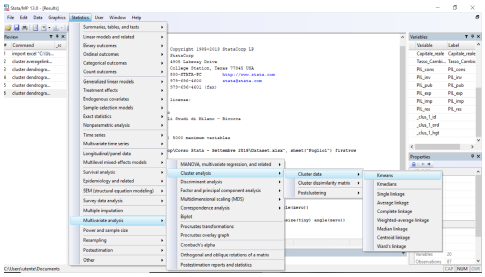


Figura 57: K-means con Stata

Le tecniche non gerarchiche non prevedono l'utilizzo di dendrogrammi per valutarne la bontà, perciò è utile descrivere i gruppi ottenuti tramite tabelle di frequenza, statistiche descrittive divise per gruppo oppure rappresentazioni grafiche come scatterplot colorati in base al gruppo di appartenenza.