Εισαγωγή στη Stata

Ευτυχία Σολέα
Κέντρο Διδασκαλίας και Μάθησης (ΚΕ.ΔΙ.ΜΑ)
Πανεπιστήμιο Κύπρου

Ιανουάριος 24, 2018

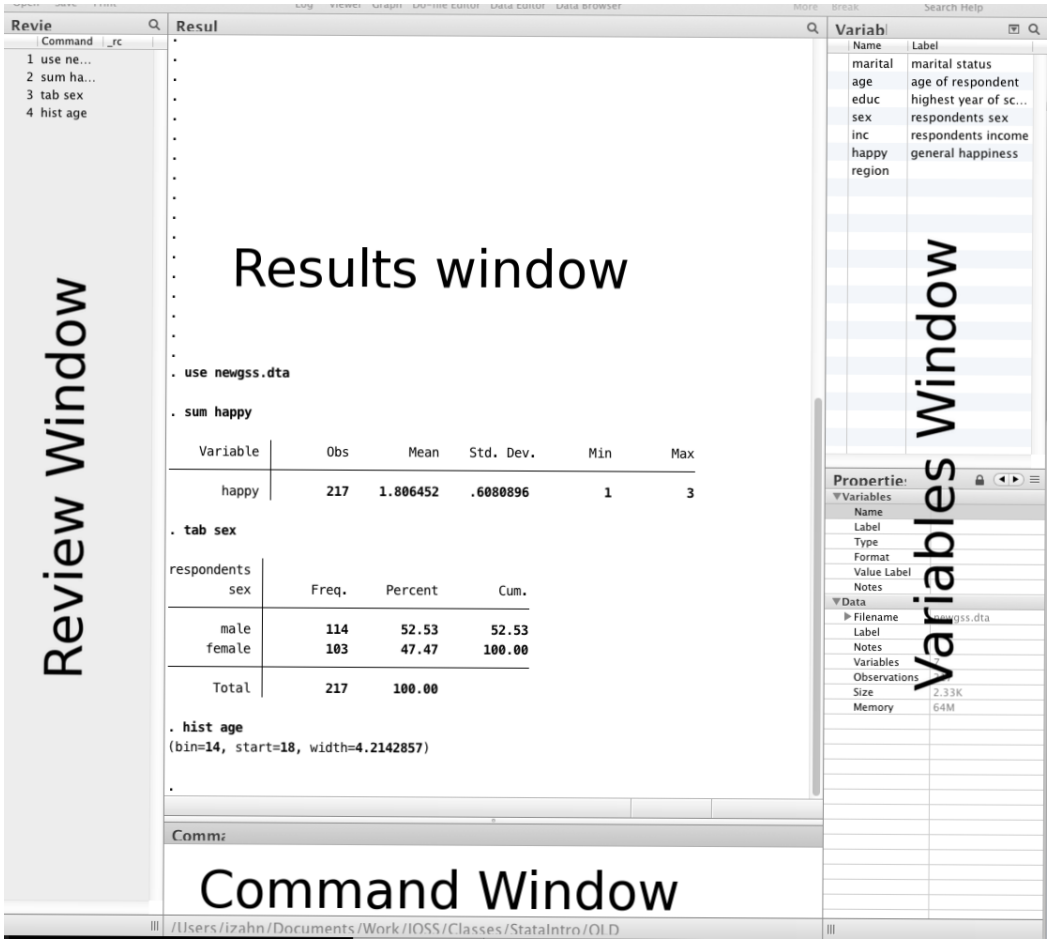
Outline

Outline

Why Stata

- Το Stata (http://www.stata.com/) είναι ένα λογισμικό πακέτο γενικής χρήσης που δημιουργήθηκε στις αρχές του 1985 από την Statacorp.
- Χρησιμοποιείται από πολλές επιχειρήσεις και ακαδημαϊκά ιδρύματα σε όλο τον κόσμο αλλά κυρίως χρησιμοποιείται σε κλάδους της οικονομικής επιστήμης.
- Περιλαμβάνει πλέον ένα ευρύ φάσμα από εντολές για την παλινδρόμηση, για την ανάλυση πάνελ δεδομένων, ιεραρχικά δεδομένων, χρονοσειρών, κ.ο.κ.
- Είναι πολύ γρήγορη και εύκολη στη χρήση της.

The Stata Interface



The Stata Interface

Τα παράθυρα της Stata:

1. Command window : Γράφω και εκτελώ τις εντολές.
2. Results window: Παρουσίαζει τα αποτέλεσματα των εντολών που εκτελούμε(output) -εκτός από τα γραφήματα.
3. Review window: Βλέπω ποιές εντολές εκτέλεσα.
4. Variable window: Παρουσιάζει τις μεταβλητές του δείγματος.

Stata help

1. Η εντολή help παρέχει βοήθεια και πληροφορίες για τη χρήση μιας εντολής. Γράφουμε:

   help command-name

2. Για παράδειγμα, αν θέλουμε βοήθεια για τη χρήση της εντολής summarize γράφουμε στο Comnand Window

   help summarize

3. Οι πληροφορίες για την εντολή δίνονται σε άλλο παράθυρο.

Do-file

1. Είναι ένα παράθυρο μέσα στο οποίο μπορούμε να γράψουμε μια σειρά εντολών. Δεν εκτελούνται.

2. Για να εκτελέσουμε τις εντολές, τις τονίζουμε (highlight) και πατάμε Ctrl+D ή τις αντιγράφουμε στο Comnand Window

3. Το Do-file αποθηκεύεται με την προέκταση .dta (Stata format .dta files). Έτσι μπορεί να χρησιμοποιηθεί σε μετέπειτα αναλύσεις.

4. Ξεκινώ ένα καινούριο Do-file πληκτρολογώντας Ctrl+9 ή πατούμε: Window → Do-file Editor → New Do-file Editor.

Use comments

1. Βάζοντας σχόλια στο Do file κάνουμε ευκολότερη τη ζωή μας για να μην ξεχάσουμε κάτι που κάναμε. Επίσης, είναι εύκολο για κάποιον άλλον να καταλάβει το Do file μας.

2. Στη Stata υπάρχουν δύο είδη comments: // και /* */.
   - Ό,τι ακολουθεί το σύμβολο // είναι σχόλιο.
   - Ό,τι βρίσκεται ανάμεσα σε /* και */ είναι σχόλιο.

Παράδειγμα: Do file

Αριθμητικές πράξεις με τη χρήση της εντολής display

Do-file Editor - Intro-Stata*

File Edit View Project Tools

Intro-Stata* × Untitled.do ×

```
1   /* January 2018
2   Introduction to Stata
3   University of Cyprus */
4   ////////////////////////////////////
5   //Stata can work as a calculator using the display command:
6   display 2+2^5*10
7   display log10(10)
8   display exp(0)
9   display sin(-2.3)
10  help mathfun // to see the complete list of mathematical functions
11  // Use /// to break long commands into lines
12  help summarize//summarize can be abbreviated to sum
13
14  /* Load data into Stata */
15  // change working directory
16  cd "C:\Users\user\Desktop\Stata Workshop\StataIntro\dataSets"
17  // dir or ls Show files in current directory
18  dir
19  //Read a Stata file
20  use gss.dta
21  //Read an excel file
22  /* the clear command :n Stata, we can only have one dataset loaded in memory at a time.
23  Before another dataset can be loaded, we must erase all data from memory using the clear command.
24  We can also clear memory as we load in another dataset using the clear option on one
25  of the data-loading commands (see below)*/
26  //the variable names are contained in the first row using the firstrow option
27  clear
```

General Stata command syntax

- Οι εντολές στη Stata έχουν την ακόλουθη γενική μορφή:

  command [*varlist*] [if *expression*] [in *range*] [, *options*]

- Για παράδειγμα,
  summarize var1 var2
  summarize var1 if var2>10
  summarize var1 in 1/10
  summarize var1 var2, detail

- Προσοχή: Σε κάποιες περιπτώσεις, αν γράψουμε μόνο command και παραλείψουμε τις μεταβλητές, η Stata θα εκτελέσει την εντολή για όλες τις μεταβλητές στο δείγμα μας.

Επιστροφή αποτελεσμάτων μιας εντολής-returned results in Stata

- Η Stata αποθηκεύει στη μνήμη της τα αποτελέσματα μιας τρέχουσας εντολής. Έτσι μπορούμε αυτά τα αποτελέσματα να τα ανακαλέσουμε αν θέλουμε να τα χρησιμοποιήσουμε σε επόμενες εντολές.
- Η Stata αποθηκεύει τα αποτελέσματα στη μορφή r(resultname) (r-class results) ή e(resultname) (e-class) και μπορεί να είναι αριθμός, πίνακας, συνάρτηση.
- Βλέπω τα αποθηκευμένα αποτελέσματα μιας εντολής με την εντολή help ή με τις εντολές return list (r-class) ή ereturn list (e-class).
- Παράδειγμα: summarize varname
  display $r(sd)^2$

r(sd): επιστρέφει την τυπική απόκλιση της μεταβλητής varname.

Outline

Εισαγωγή αρχείου Stata

1. Βεβαιωθείτε ότι είστε στο σωστό φάκελο (directory) του υπολογιστή (εκεί που είναι αποθηκευμένο το αρχείο). Μπορούμε να αλλάξουμε το directory με την εντολή:
cd "C://Users/dataclass/Desktop/StataIntro".

2. Για να εισάγωγουμε δεδομένα σε Stata format (dta file) γράφουμε: use filename.dta

3. Η εντολή clear σβήνει τα τρέχοντα δεδομένα από τη μνήμη της Stata (clear tells Stata to erase the previous dataset.)

What if my data is not a Stata file?

Εισαγωγή δεδομένων σε άλλες μορφές:

1. Excel file: import excel using filename.xlsx, firstrow clear
2. Delimited text files (.csv): import delimited using filename.csv, clear
3. SAS file (.xpt): import sasxport filename.xpt, clear
4. SPSS file: Η SPSS δίνει την επιλογή να αποθηκεύσετε ένα αρχείο δεδομένων σε .dta format (Go to: file > save as > Stata (use most recent version available)).
5. Από το διαδίκτυο: use https://stats.idre.ucla.edu/stat/data/hs0, clear

1. Ένας τρόπος να ελέγξουμε τα στοιχεία σχετικά με τα δεδομένα του αρχείου μας είναι μέσω της εντολής describe, όπου μας δίνεται πληροφορίες για το δείγμά μας (π.χ., αριθμός των παρατηρήσεων και μεταβλητών, ονόματα των μεταβλητών κ.τ.λ).

2. Η εντολή codebook μας δίνει αναλυτικές πληροφορίες για την κάθε μεταβλητή του δείγματος, όπως range, missing values, labeling information. Γράφοντας codebook varname παίρνουμε τις πληροφορίες μόνο για τη μεταβλητή varname.

3. Η εντολή list varname μας δίνει τις τιμές που παίρνουν οι συμμετέχοντες του δείγματός μας στη μεταβλητή varname. Ενώ με την εντολή *list varname in 1/10* παίρνω τις παρατηρήσεις για τους πρώτους 10 συμμετέχοντες.

Σώζωντας τα δεδομένα

1. Μπορούμε να σώσουμε το αρχείο μας σε μορφή .dta με την εντολή save NewFileName, replace

2. Αν θέλουμε να το σώσουμε σε άλλη μορφή:

   .xls format (Excel): export excel using newFileName.xls, replace
   .csv format: export delimited using newFileName.csv, replace
   .xpt format (SAS): export sasxport newFilename.xpt,replace

log files-Σώζωντας εντολές και αποτελέσματα

1. Μπορούμε να σώσουμε τις εντολές και τα αποτελέσματα (output) (εκτός από τα γραφήματα) σε .txt αρχείο με την εντολή

   log using filename.txt, text replace

2. Κλείνουμε log file με την εντολή log close.

3. Μπορούμε να δούμε το text αρχείο με την εντολή:

   view filename.txt

   ή ανοίγουμε το log file με NotePad.

Exercise 1

1. Ανοίξτε τη Stata.
2. Ανοίξτε και σώστε ένα καινούριο Do file.
3. Ανοίξτε ένα log file με το όνομα results.
4. Διαβάστε τα δεδομένα από αυτή την ιστοσελίδα https://stats.idre.ucla.edu/stat/data/hs0.
5. Εξέτασε τα δεδομένα με τις εντολές describe, codebook.
6. Σωστε τα δεδομένα σε Stata format με το όνομα hs0.
7. Close and view the log file.

Outline

Variable and value labels

1. Η εντολή label variable μας δίνει τη δυνατότητα να περιγράψουμε μια μεταβλητή. Για παράδειγμα, label variable schtyp "type of school".

2. Μπορούμε να αλλάξουμε το όνομα μιας μεταβλητής με την εντολή rename. Για παράδειγμα, rename gender sex αλλάζουμε το όνομα της μεταβλητής από gender σε sex.

3. Value labels: Σε περίπτωση που έχετε κατηγορική μεταβλητή, μπορεί να θέλουμε να καθορίσουμε τι αναπαριστά κάθε αριθμός π.χ., 1=δημόσιο σχολείο, 2=ιδιωτικό σχολείο. Στη Stata γίνεται με δύο βήματα:

   label define scl 1 public 2 private
   label values schtyp scl

Δημιουργία καινούριων μεταβλητών

1. Μπορούμε να αλλάξουμε τις κωδικοποιημένες τιμές μιας μεταβλητής με την εντολή recode. Για παράδειγμα, θέλουμε η μεταβλητή gender να παίρνει τιμές 0 και 1 αντί 1 και 2 (dummy variable). Τότε

   recode gender (1=0)(2=1).

2. Μπορούμε να δημιουργήσουμε καινούριες μεταβλητές με την εντολή generate χρησιμοποιώντας μαθηματικές συναρτήσεις ή αριθμητικές πράξεις.

3. Για παράδειγμα, αν θέλουμε η μεταβλητή με το όνομα total να μας δίνει το συνολικό τεστ σκορ των μαθητών, θα γράψουμε: generate total = read + write + math + science

Δημιουργία καινούριων μεταβλητών

Δημιουργία κατηγορικών μεταβλητών- the 'generate and replace' strategy.

1. Θέλουμε να δημιουργήσουμε μια νέα dummy or indicator μεταβλητή με το όνομα totaldummy η οποία να διαχωρίζει τους μαθητές με βάση τις τιμές της μεταβλητής total έτσι ώστε οι μαθητές να παίρνουν την τιμή 1 αν έχουν total σκορ μεγαλύτερο ή ίσο από μέσο όρο της μεταβλητής total. Διαφορετικά, να παίρνουν την τιμή 0.

   generate totaldummy=0 if (total< *r*(*mean*) & !missing(total))
   replace totaldummy=1 if (total>= *r*(*mean*) & !missing(total))

Missing values

1. Στη Stata η τελεία . σημαίνει missing value.

2. Σημείωση: Η Stata θεωρεί τις missing values ως τεράστιες τιμές (infinity).Άρα για να τις αποκλείσουμε χρησιμοποιούμε την εντολή missing(varname).

Ταξινομώντας τα δεδομένα

1. Μπορείτε να ταξινομήσετε τα δεδομένα με βάση το επίπεδο του ses των συμμετεχόντων κατά κατιούσα φορά, γράφοντας:

   sort ses
   list ses gender race in 1/10

2. Μπορείτε να ταξινομήσετε τα δεδομένα με βάση ses και gender γράφοντας:

   sort ses gender

Exercise 2

1. Ανοίξτε ένα log file με το όνομα results.
2. Διαβάστε τα δεδομένα hs0 που αποθηκεύσατε.
3. Δώστε την περιγραφή "Type of program" στη μεταβλητή prgtype.
4. Αλλάξτε το όνομα της μεταβλητής gender σε female.
5. Αλλάξτε τις τιμές της μεταβλητής schtyp έτσι ώστε το 1 να γίνει 0 και το 2 να γίνει 1.
6. Δημιουργείστε τα value labels της μεταβλητής schtyp έτσι ώστε 1=public 0=private.
7. Αφαιρέστε από τη μεταβλητή math τη μέση της τιμή και αποθηκεύστε την με νέο όνομα.
8. Σώστε τα δεδομένα σε Stata format με ένα νέο όνομα.
9. Close and view the log file.

Outline

Περιγραφικά μέτρα για συνεχείς (ποσοτικές) μεταβλητές

Υπολογίσμός μέτρα περιγραφικής στατιστικής για συνεχείς (continuous) μεταβλητές όπως η μέση τιμή, η τυπική απόκλιση και άλλα.

Συνεχείς (ποσοτικές) μεταβλητές: Το σύνολο των δυνατών τιμών είναι ένα συνεχές υποσύνολο των πραγματικών αριθμών όπως το βάρος, το ύψος, η ηλικία κ.ο.κ.

- Η εντολή summarize μας υπολογίζει τη μέση τιμή, την τυπική απόκλιση, την ελάχιστη (minimum) και τη μέγιστη τιμή (maximum). Η επιλογή detail μας δίνει περισσότερα μέτρα περιγραφικής στατιστικής όπως τη διάμεσο, τη διασπορά, τα ποσοστιαία σημεία κ.λ.π.

Περιγραφικά μέτρα για συνεχείς μεταβλητές

- Παράδειγμα: summarize write, detail
- Αν θέλουμε να υπολογίσουμε περιγραφικά μέτρα για κάθε κατηγορία μια κατηγορικής μεταβλητής που μας ενδιαφέρει, χρησιμοποιούμε την εντολή tabstat με την επιλογή by:

  tabstat write, by(gender) stat(mean sd)

Γραφήματα για συνεχείς μεταβλητές

1. Ιστόγραμμα: histogram math, normal
2. Θηκόγραμμα: graph box math
3. Θηκόγραμμα/Ιστόγραμμα για κάθε κατηγορία μιας άλλης κατηγορικής μεταβλητής χρησιμοποιώντας τις επιλογές by or over:

   histogram math, normal by(prgtype) /* densities by prgtype */
   graph box write, over(race) /* box plots by race */

Συσχέτιση μεταξύ συνεχών μεταβλητών

Μπορούμε επίσης να εξετάσουμε τη σχέση μεταξύ δύο ή περισσότερων συνεχών μεταβλητών:

1. Συντελεστής γραμμικής συσχέτισης: pwcorr read math write science, sig. Η επιλογή sig μου δίνει τα p-values.

2. Διάγραμμα διασποράς (scatter plot):
   - Μεταξύ δύο συνεχών μεταβλητών: twoway (scatter write read)
   - Μεταξύ περισσότερων από δύο συνεχών μεταβλητών: graph matrix read science write, half

Περιγραφικά μέτρα για ποιοτικές μεταβλητές

1. Ποιοτικές μεταβλητές: Οι τιμές δε δίδονται με αριθμούς αλλά με διακριτικό είδος, για παράδειγμα το 'φύλο' παίρνει τιμές άρρεν και θήλυ.

2. Μπορούμε να δημιουργήσουμε τον πίνακα συχνοτήτων (frequency table) και να κατασκευάσουμε γραφήματα όπως το τομεόγραμμα (pie chart), το ραβδόγραμμα (bar graph)

   - Πίνακας συχνοτήτων (frequency table): tab gender
   - Τομεόγραμμα (pie chart): graph pie, over(race) plabel(_all name)
   - Ραβδόγραμμα (bar graph): graph bar, over(race)

Exercise 3

1. Ανοίξτε το log file με το όνομα results.
2. Διαβάστε τα δεδομένα hs0 που αποθηκεύσατε.
3. Υπολογίστε τον πίνακα συχνοτήτων για τη μεταβλητή ses.
4. Υπολογίστε το δειγματικό μέσο (mean), το εύρος (range) και τη διασπορά (variance) της μεταβλητής write, για τα αγόρια και τα κορίτσια. (Στο output να φαίνονται οι κατηγορίες 'αγόρια', 'κορίτσια' και όχι οι τιμές).
5. Κατασκευάστε το θηκόγραμμα (boxplot) της μεταβλητής write, για την κάθε κατηγορία της μεταβλητής gender
6. Close and view the log file.

Outline

Συγκρίσεις διαφοράς μέσων όρων (t-tests)

Υπάρχουν τρείς t έλεγχοι γνωστοί:

1. Έλεγχος υποθέσεων για τη μέση τιμή, μ, ενός πληθυσμού (one-sample t-test).
2. Έλεγχος υποθέσεων για τη διαφορά των μέσων δύο ανεξάρτητων δειγμάτων (Independent two sample t-test).
3. Έλεγχος υποθέσεων για τη διαφορά των μέσων δύο εξαρτημένων δειγμάτων (paired t-test).

Σημείωση: Οι τρείς έλεγχοι υποθέσεων βασίζονται στην υπόθεση ότι το δείγμα μας ακολουθεί την κανονική κατανομή, εκτός και αν το μέγεθος του δείγματος είναι πολύ μεγάλο ($n \geq 30$).

Έλεγχος υποθέσεων για τη μέση τιμή, μ, ενός πληθυσμού (one-sample t-test)

Ενδιαφερόμαστε να ελέγξουμε την υπόθεση ότι η μέση τιμή της βαθμολογίας στα μαθηματικά (math) του πληθυσμού των μαθητών δεν είναι ίση με 50. Δηλαδή οι υποθέσεις (μηδενική και εναλλακτική) είναι της μορφής:

$$H_0 : \mu = 50 \quad H_1 : \mu \neq 50.$$

(αμφίπλευρος έλεγχος)

Για να διεξάγουμε το t τεστ στη Stata γράφουμε:

```
ttest math=50
```

Σύγκριση δύο μέσων όρων για ανεξάρτητα δείγματα (Independent two sample t-test)

Θέλουμε να ελέγξουμε αν υπάρχει διαφορά στη μέση βαθμολογία των μαθηματικών (math) μεταξύ αγοριών και κοριτσιών. Οι υποθέσεις διαμορφώνονται ως εξής:

$$H_0 : \mu_1 = \mu_2 \quad H_1 : \mu_1 \neq \mu_2.$$

(αμφίπλευρος έλεγχος)

Για να διεξάγουμε τον έλεγχο στη Stata γράφουμε:

Ίσες διασπορές: ttest math, by(gender)
Άνισες διασπορές: ttest math, by(gender) unequal

Μπορώ να ελέγξω την ισότητα των διασπορών με την εντολή: sdtest math, by(gender)

Σύγκριση δύο μέσων όρων για εξαρτημένα δείγματα (Paired t-test)

1. Τι γίνεται όμως όταν τα δύο δείγματα δεν είναι, ή δεν μπορούμε να υποθέσουμε ότι είναι ανεξάρτητα; Στην περίπτωση αυτή κάνουμε έλεγχο για ζεύγη παρατηρήσεων (paired data t-test). Αυτός ο έλεγχος εφαρμόζεται όταν έχουμε μετρήσεις από δύο μεταβλητές για το ίδιο άτομο (βάρος πριν και μετά από μια δίαιτα) ή μετρήσεις από αδέρφια.

2. Παράδειγμα, θέλουμε να ελέγξουμε αν οι μέσες βαθμολογίες των μαθητών στα μαθηματικά (math) και στην επιστήμη (science) διαφέρουν. Στη Stata το εξέταζουμε εκτελώντας:

   ttest math=science

Ανάλυση διασποράς κατά ένα παράγοντα (one-way ANOVA)

1. Η ανάλυση διασποράς χρησιμοποιείται όταν θέλουμε να ελέγξουμε την ισότητα μέσων όρων ανάμεσα σε περισσότερες από δύο ανεξάρτητες ομάδες.
2. Για παράδειγμα, θέλουμε να ελέγξουμε αν υπάρχουν διαφορές στη μέση βαθμολογία των μαθηματικών (math) με βάση το κοινωνικο-οικονομικό επίπεδο των μαθητών (ses). Οι υποθέσεις διαμορφώνονται ως εξής:

$$H_0 : \mu_1 = \mu_2 = \mu_3 \quad H_1 : \mu_i \neq \mu_j, i \neq j.$$

3. Στη Stata γράφουμε: anova math i.ses
   Με το γράμμα i λέμε στη Stata ότι η μεταβλητή ses είναι κατηγορική μεταβλητή.

Ανάλυση διασποράς κατά ένα παράγοντα (one-way ANOVA)

1. Το F-test του ANOVA δε μου λέει ποιες ακριβώς ομάδες διαφέρουν μεταξύ τους.

   Για να δούμε ποια ζεύγη μέσων διαφέρουν θα εφαρμόσουμε τον έλεγχο των πολλαπλών συγκρίσεων του Bonferroni χρησιμοποιώντας την εντολή pwcompare και την επιλογή mcompare(bonferroni)

   pwcompare ses, mcompare(bonferroni) effects

Έλεγχος συσχέτισης δύο κατηγορικών μεταβλητών (Chi-square test)

1. Ο χ^2 (Chi-squared) έλεγχος χρησιμοποιείται για τον έλεγχο της υπόθεσης ότι δύο κατηγορικές μεταβλητές είναι ανεξάρτητες μεταξύ τους. Οι υποθέσεις είναι:

   H_0: Οι δύο μεταβλητές είναι ανεξάρτητες H_1: Οι δύο μεταβλητές δεν είναι ανεξάρτητες

2. Η εντολή tabulate με τις επιλογές (options) row column chi2 υπολογίζει τον χ^2 έλεγχο ανεξαρτησίας και τον πίνακα συχνοτήτων με τα ποσοστά γραμμών και στηλών.

Έλεγχος συσχέτισης δύο κατηγορικών μεταβλητών (Chi-square test)

1. Παράδειγμα: Υπάρχει εξάρτηση μεταξύ του κοινωνικο-οικονομικού επιπέδου των μαθητών (ses) και της φυλής (race);

   tabulate ses race, row column chi2

2. Αν θέλω και τα expected frequencies, χρησιμοποιώ την επιλογή exp

   tabulate ses race, row column chi2 exp

Exercise 4

1. Ανοίξτε το log file με το όνομα results.
2. Διαβάστε τα δεδομένα hs0 που αποθηκεύσατε.
3. Να ελεγχθεί σε επίπεδο σημαντικότητας 5% η υπόθεση ότι η μέση τιμή της τ.μ. socst (social studies score) είναι 55 με εναλλακτική ότι είναι χαμηλότερη από 55.
4. Να ελεγχθεί σε ε.σ. 5% η υπόθεση ότι η μέση τιμή της τ.μ. socst δε διαφέρει ανάμεσα σε αγόρια και κορίτσια και με εναλλακτική ότι διαφέρει.
5. Επειδή η τ.μ prgtype είναι τύπου string, κατασκευάστε μια νέα κατηγορική μεταβλητή με το όνομα prog έτσι ώστε (1=Academic, 2=general, 3=vocati). Να ελεγθεί η υπόθεση ότι το είδος της εκπαίδευσης ενός μαθητή (prog) δεν εξαρτάται από τη μεταβλητή (ses) με εναλλακτική ότι εξαρτάται.
6. Close and view the log file.

Outline

Πολλαπλή Γραμμική Παλινδρόμηση

1. Η πολλαπλή γραμμική παλινδρόμηση εξετάζει τη γραμμική σχέση που μπορεί να έχουν κάποιες ανεξάρτητες μεταβλητές με μια εξαρτημένη ποσοτική μεταβλητή.

2. Οι ανεξάρτητες μεταβλητές μπορεί να είναι ποσοτικές ή κατηγορικές.

3. Στη Stata χρησιμοποιούμε την εντολή regress, όπου πρώτα βάζουμε την εξαρτημένη και μετά τις ανεξάρτητες:
   regress depvar [indepvars], options
   όπου,
   depvar: εξαρτημένη μεταβλητή
   indepvars: ανεξάρτητες μεταβλητές

Πολλαπλή Γραμμική Παλινδρόμηση

Παράδειγμα: Το πολλαπλό γραμμικό μοντέλο με εξαρτημένη μεταβλητή την write και ανεξάρτητες μεταβλητές τις read και gender

regress write c.read i.gender

όπου,

- write: εξαρτημένη μεταβλητή
- read, gender ανεξάρτητες μεταβλητές
- Βάζω c για να δηλώσω τις ποσοτικές μεταβλητές και i τις κατηγορικές μεταβλητές.

Πολλαπλή Γραμμική Παλινδρόμηση με αλληλεπιδράσεις

1. Οι αλληλεπιδράσεις (interactions) μας λένε αν η σχέση που υπάρχει ανάμεσα σε δύο μεταβλητές αλλάζει ή διαφέρει ως προς τις τιμές μιας άλλης μεταβλητής.
2. Για παράδειγμα, η θετική επίδραση που έχει η τ.μ read στη μεταβλητή write μπορεί να διαφέρει ανάμεσα σε αγόρια και κορίτσια. Με άλλα λόγια, υπάρχει αλληλεπίδραση μεταξύ των μεταβλητών gender και read;
3. Στη Stata δηλώνουμε την αλληλεπίδραση μεταξύ δύο μεταβλητών με το σύμβολο ##
   regress write c.read##i.gender

Interaction plot

Στη Stata σχεδιάζω το interaction plot εκτελώντας πρώτα την εντολή margins και μετά την εντολή marginsplot.

1. margins: Υπολογίζει τις προβλεπόμενες τιμές της εξαρτημένης (predicted values).

   margins gender, at(read=(28 (5) 76))

2. marginsplot: Κατασκευάζει το interaction plot.

   marginsplot

Εξέταση των υπολοίπων

- Οι υποθέσεις που πρέπει να ικανοποιούνται για την πολλαπλή γραμμική παλινδρόμηση είναι:
  1. Γραμμικότητα (**L**inearity)
  2. Ανεξαρτησία (**I**ndependence)
  3. Κανονικότητα (**N**ormality)
  4. Ομοσκεδαστικότητα (**E**qual Variance)

  "LINE" assumptions
- Οι υποθέσεις αυτές εξετάζονται γραφικά με τη χρήση των υπολοίπων (residuals).
- Στη Stata υπολογίζουμε τα υπόλοιπα με την εντολή predict και την επιλογή residuals, τα οποία αποθηκεύονται στη μεταβλητή με το όνομα res:
  predict res, residuals

Εξέταση των υπολοίπων

1. Για να ελέγξουμε τη γραμμικότητα και την ομοσκεδαστικότητα των υπολοίπων χρησιμοποιούμε ένα διάγραμμα διασποράς το οποίο θα περιέχει τις προβλεπόμενες τιμές της Y (fitted values) στον οριζόντιο άξονα και τα υπόλοιπα (residuals) στον κάθετο άξονα.

2. Υπολογίζουμε τα fitted values πάλι με την εντολή predict ως εξής: predict fit

3. Το διάγραμμα διασποράς κατασκευάζεται με την εντολή twoway: twoway (scatter res fit)

4. Με την εντολή rvfplot κατασκευάζω το διάγραμμα διασποράς, χωρίς να χρειάζεται να υπολογίσω τα residuals και τα fitted values

Εξέταση των υπολοίπων

1. Έλεγχος της κανονικότητας: Την υπόθεση της κανονικότητας των υπολοίπων μπορούμε να την ελέγξουμε γραφικά με διάφορους τρόπους όπως με ιστόγραμμα, διάγραμμα ποσοστιαίων σημείων (normal quantile plot).

   histogram res, normal
   qnorm res

2. Έλεγχος της ανεξαρτησίας: Για παράδειγμα, όταν έχουμε δεδομένα που παίρνονται με χρονική σειρά, μπορούμε να κάνουμε ένα διάγραμμα υπολοίπων ως προς το χρόνο ή εφαρμόζουμε τον έλεγχο Durbin-Watson (dwstat command in Stata).

Regression with Robust Standard Errors

Αν οι υποθέσεις της κανονικότητας και της σταθερής διασποράς δεν ισχύουν τότε μπορώ να κάνω χρηση της επιλογής robust έτσι ώστε να υπολογίσω σωστά τυπικά σφάλματα.

regress depvar [indepvar], robust

Regression with Clustered Standard Errors

- Μη ανεξαρτησίας των υπολοίπων συμβαίνει όταν έχω clustered data. Για παράδειγμα, όταν έχω μετρήσεις μαθητών από διάφορα σχολεία (cluster). Τότε οι μετρήσεις μαθητών που προέρχονται από το ίδιο σχολείο θα είναι συσχετισμένες μεταξύ τους, ενώ οι μετρήσεις από διαφορετικά σχολεία θα είναι ανεξάρτητες μεταξύ τους.

- Σε αυτή την περίπτωση χρησιμοποιούμε την επιλογή cluster:

  regress depvar [indepvars], cluster(varname)

Stored results for regress

1. Η Stata αποθηκεύει τα αποτελέσματα από την εντολή regress στη μορφή e(...).

2. Η εντολή ereturn list μου δίνει τα αποθηκευμένα αποτελέσματα της μορφής e(...).

Διωνυμική γραμμική παλινδρόμηση(Logistic regression)

1. Η διωνυμική γραμμική παλινδρόμηση (logistic regression) εξετάζει τη σχέση που έχουν κάποιες ανεξάρτητες μεταβλητές με μια κατηγορική μεταβλητή που έχει δύο κατηγορίες μόνο (δυαδική μεταβλητή).

2. Για παράδειγμα, θέλουμε να μελετήσουμε την επίδραση που έχουν οι ανεξάρτητες μεταβλητές gender, read, type of program στην πιθανότητα εγγραφής ενός μαθητή (δυαδική εξαρτημένη μεταβλητή) (honors).

   Η εξαρτημένη μεταβλητή ορίζει δύο κατηγορίες: εγγραφή και μη εγγραφή.

Διωνυμική γραμμική παλινδρόμηση (Logistic regression)

1. Στη Stata εφαρμόζουμε logistic regression με την εντολή logit. Για το παράδειγμα, γράφουμε

   logit honors c.read i.gender i.prog

2. Αν θέλουμε τους λόγους πιθανοτήτων (odds ratio) χρησιμοποιούμε την επιλογή or:

   logit , or

Ordinal Logistic regression with Stata

1. Η πολλαπλή τακτική παλινδρόμηση (Ordinal Logistic regression) επιλέγεται στις περιπτώσεις όπου η εξαρτημένη μεταβλητή διακρίνεται σε περισσότερες από δύο κατηγορίες οι οποίες αυξάνονται κατά κλίμακα, όπως η προτίμηση ενός προϊόντος.

2. Στη Stata εφαρμόζω ordinal logistic regression με την εντολή ologit

Ordinal Logistic regression with Stata

1. Παράδειγμα, το μοντέλο ordinal logistic regression για την εξαρτημένη μεταβλητή ses με την ανεξάρτητη μεταβλητή gender είναι

   ologit ses i.gender, or

Exercise 5

Χρησιμοποιώντας τα δεδομένα hs0 .

1. Κατασκευάστε το διάγραμμα διασποράς μεταξύ των μεταβλητών write και socst.

2. Προσαρμόστε το γραμμικό μοντέλο με αλληλεπιδράσεις με την εξαρτημένη τ.μ. write και τις ανεξάρτητες μεταβλητές socst και gender.

Resources

1. Stata user's guide (release 15): https://www.stata.com/manuals/u.pdf
2. UCLA website: https://stats.idre.ucla.edu/stata/
3. Princeton website: http://www.princeton.edu/∼ otorres/Stata/