ΣΕΜΙΝΑΡΙΑ ΕΚΠΟΝΗΣΗΣ ΔΙΠΛΩΜΑΤΙΚΗΣ ΕΡΓΑΣΙΑΣ – RESEARCH METHODS

Διδάσκων: Αναπλ. καθ. Λογιστικής Ορέστης Βλησμάς

Εφαρμογή 3η
Εκκαθάριση δεδομένων – Εισαγωγή στο Stata

Ακολουθήστε τα παρακάτω βήματα:

Βήμα 1ο:

1. Μεταβείτε στο e-class του σεμιναρίου. Στο folder με ονομασία Laboratory του e-class αναζητήστε το αρχείο με ονομασία Application3.xlsx. Αποθηκεύεται το τοπικά στον υπολογιστή σας.
2. Μεταβείτε στο e-class του σεμιναρίου. Στο folder με ονομασία Laboratory του e-class αναζητήστε το αρχείο με ονομασία Code_Application_3. Αποθηκεύεται το τοπικά στον υπολογιστή σας.
3. Μεταβείτε στο e-class του σεμιναρίου. Στο folder με ονομασία Library_Stata αναζητήστε και ανοίξτε το αρχείο με ονομασία Xristikos_Odigos_Stata. Αποθηκεύεται το τοπικά στον υπολογιστή σας. Το αρχείο αυτό θα το έχετε ανοικτό στην διάρκεια της εκτέλεσης της εφαρμογής προκειμένου να το συμβουλεύεστε αν χρειαστεί.

Βήμα 2ο: Εργασία στο spreadsheet Application3.xlsx

Στο πρόγραμμα λογισμικού MS excel ανοίξτε το αρχείο Application3.xlsx το οποίο εμπεριέχει δεδομένα όπως αυτά θα εμφανίζονται όταν τα λαμβάνεται από τη βάση δεδομένων. Ειδικότερα στο Sheet1 εμπεριέχονται δεδομένα που αφορούν τα χαρακτηριστικά των εταιρειών που συνθέτουν το δείγμα σας και στο Sheet2 εμπεριέχονται δεδομένα μεταβλητών. Εκτελέστε τις ακόλουθες ενέργειες:

1. Δημιουργήστε ένα νέο Sheet (Sheet3) στο οποίο να ενοποιήστε τα δεδομένα από το Sheet1 και το Sheet2 έτσι ώστε τα δεδομένα σας να εμπεριέχουν στις πρώτες στήλες τα ιδιαίτερα χαρακτηριστικά της εταιρείας (δηλαδή Name, Symbol, RIC), εν συνεχεία το κλάδο (Sector) που δραστηριοποιείται και, τέλος, να ακολουθούν οι στήλες του Sheet2 με τις μεταβλητές κατά τρόπο τέτοιο ώστε οι αριθμητικές τιμές τους να αντιστοιχούν στις εταιρείες που αναφέρονται.
2. Διαγράψτε όλες τις στήλες με περιττές και επαναλαμβανόμενες πληροφορίες (συνήθως είναι στήλες που παρεμβάλλονται μεταξύ στηλών που εμπεριέχουν μεταβλητές οικονομικών δεδομένου (π.χ. πωλήσεις, ενεργητικό, κ.λπ.) διαφορετικής φύσης.
3. Διαγράψετε (με αντικατάσταση με κενό) όλες τις περιπτώσεις που δεν υπάρχουν διαθέσιμα στοιχεία ή όποιο κελί εμφανίζει συμβολοχαρακτήρα ενώ θα έπρεπε να

εμφανίζει αριθμητική τιμή. Για παράδειγμα στο συγκεκριμένο αρχείο μπορεί να εμφανίζεται κάτι από τα ακόλουθα:

⇒ NA
⇒ $$ER: E100,NO WORLDSCOPE DATA FOR THIS CODE
⇒ $$ER: E100,INVALID CODE OR EXPRESSION ENTERED
⇒ $$ER: 4540,NO DATA VALUES FOUND

Εναλλακτικά να εκτελεστεί το 3 με το filter για κάθε μία μεταβλητή (στην πρώτη της χρονιά) για να εντοπίστε τους συμβολοχαρακτήρες και να τους διαγράψετε.

Προσοχή στο , και το .

Όταν ολοκληρώσετε τη διαδικασία αποθηκεύστε τις αλλαγές στο αρχείο Applications3.xlsx.

Βήμα 3ο Μεταφορά δεδομένων στο STATA

Εκκινήστε την εφαρμογή STATA και μεταφέρεται τα δεδομένα από το σας Sheet3 του Applications3.xlsx. όπως το αποθηκεύσατε μετά την ολοκλήρωση του βήματος 2. Αποθηκεύεται το αρχείο των δεδομένων με μορφή αρχείου δεδομένων STATA (*.dta) με όποια ονομασία αρχείου κρίνεται δόκιμη (προτιμήστε λατινικούς χαρακτήρες).

Βήμα 4ο Δημιουργία Firm_id

Για να διακρίνεται τις εταιρείες του δείγματός σας δημιουργήστε μια νέα μεταβλητή με την ονομασία Firm_id η οποία να αποδίδει αύξων αριθμό σε κάθε εταιρεία. Η σύνταξη της σχετική εντολής είναι:

generate Firm_id = _n

Βήμα 5ο: Μετατροπή δομής δεδομένων (από wide σε long)

Έστω ότι έχουμε αποθηκεύσει δεδομένα για δύο εταιρείες (την Α και την Β) αναφορικά με την αξία των assets και των sales. Τα δεδομένα αφορούν τρία έτη: 2000, 2001 και 2002. Σε αυτή την περίπτωση, η δομή wide έχει ως εξής:

Firm_id	Assets2000	Assets2001	Assets2002	Sales200	Sales2001	Sales2002
A	XXXXX	XXXXX	XXXXX	XXXXX	XXXXX	XXXXX
B	XXXXX	XXXXX	XXXXX	XXXXX	XXXXX	XXXXX

Είναι εμφανές ότι οι εταιρείες θέτονται στην πρώτη στήλη, μετά ακολουθούν τρεις στήλες με την αξία των assets για διαδοχικά τρία έτη και έπειτα άλλες τρεις στήλες με την αξία των sales για διαδοχικά τρία έτη. Η πρώτη στήλη αντιπροσωπεύει τη μεταβλητή Firm_id και αφορά το μοναδικό κωδικό που διακρίνεται η κάθε εταιρεία στο δείγμα των δεδομένων.

Η στατιστική ανάλυση στο πλαίσιο του **STATA**, προϋποθέτει ότι η δομή των δεδομένων θα είναι long. Η δομή long στο προηγούμενο παράδειγμα έχει ως εξής:

Firm_id	Time_id	Assets	Sales
A	2000	XXXXX	XXXXX
A	2001	XXXXX	XXXXX
A	2002	XXXXX	XXXXX
B	2000	XXXXX	XXXXX
B	2001	XXXXX	XXXXX
B	2002	XXXXX	XXXXX

Το χαρακτηριστικό της δομής long σε σχέση με τη δομή wide είναι ότι εισάγεται ανεξάρτητη μεταβλητή δηλωτική του χρόνου (Time_id) και για κάθε μεταβλητή (assets, sales) αφιερώνεται μία στήλη.

Τούτων λεχθέντων να υλοποιηθούν οι κατάλληλες ενέργειες μετατροπής της δομής των δεδομένων σε long από wide. Δίδονται οδηγίες:

Έστω ότι στη δομή wide υπάρχει η αριθμητική μεταβλητή Firm_id δηλωτική της εταιρείας. Η σύνταξη της εντολής reshape έχει ως εξής:

reshape long vr1 vr2 vr3 . . vrX, i(Firm_id) j(Time_id)

και ειδικότερα στην περίπτωση της εφαρμογής:

reshape long sales sga goodwill intangibles cogs opinc empl fcfsh commondiv prefdiv cfo, i(Firm_id) j(Time_id)

Σημειώνονται τα εξής:

- Η έκφραση j(Time_id) δίδει εντολή στο STATA να δημιουργήσει μία αριθμητική τιμή δηλωτική του χρόνου με την ονομασία Time_id.
- Αν ο ερευνητής επιθυμεί να επαναφέρει τη δομή wide τότε εκτελεί την ακόλουθη εντολή:

  reshape wide

Βήμα 6ο Αντιμετώπιση πρόδηλων σφαλμάτων δεδομένων στο **STATA**

Ως πρόδηλα εσφαλμένο θεωρείται οτιδήποτε αντιβαίνει θεμελιώδεις αρχές. Παραδείγματα: πωλήσεις με αρνητικές τιμές, κυκλοφορούν ενεργητικό με αρνητικές τιμές, κ.λπ. Πραγματοποιήστε απολογιστικό ή προδραστικό έλεγχο στο πλαίσιο του STATA αξιοποιώντας μεταξύ άλλων εντολές όπως:

- drop (παράδειγμα: drop vr1, vr2, vr3)
- drop if (παράδειγμα: drop if vr1<0, drop if vr1<0 & vr1>10, drop if vr1==5 | vr1==10)
- keep (παράδειγμα: keep if vr1>=0)
- keep if (παράδειγμα: keep if vr1>=0, keep if vr1>=0 & vr1<=10)

Ωστόσο επειδή σε μια μεταβλητή μόνο μερικές τιμές μπορεί να είναι πρόδηλα εσφαλμένες τότε θα πρέπει να κάνετε generate και replace με if. Για παράδειγμα αν η μεταβλητή vr δεν πρέπει να έχει αρνητικές τιμές τότε ορίζουμε μία νέα μεταβλητή, έστω vr1 με τη σύνταξη της εντολής ν είναι:

generate vr1=.

replace vr1=vr if vr>=0

και ειδικότερα στην περίπτωση του βήματος 6:

generate sales1=.

replace sales1=sales if sales>=0

generate sga1=.

replace sga1=sga if sga>=0

generate goodwill1=.

replace goodwill1= goodwill if goodwill >=0

```
generate intangibles1=.
replace intangibles1= intangibles if intangibles >=0
generate cogs1=.
replace cogs1= cogs if cogs >=0
generate empl1=.
replace empl1=empl if empl >=0
generate fcfsh1=.
replace fcfsh1=fcfsh if fcfsh >=0
generate commondiv1=.
replace commondiv1=commondiv if commondiv>=0
generate prefdiv1=.
replace prefdiv1=prefdiv if prefdiv>=0
```

Μπορείτε να κάνετε drop όλες τις παλιές μεταβλητές (προσοχή στις μεταβλητές cfo και opinc οι οποίες μπορούν να λάβουν αρνητικές τιμές και για το λόγο αυτό δεν κάναμε μετασχηματισμό).

Βήμα 7ο Αντιμετώπιση ακραίων τιμών (outliers).

Ακραίες τιμές μίας ή περισσότερων μεταβλητών είναι δυνατόν να επηρεάσουν τα αποτελέσματα της ανάλυσης. Έχουν διαμορφωθεί δύο αντιλήψεις επί του τρόπου αντιμετώπισης των ακραίων τιμών. Η πρώτη διατείνεται ότι ο ερευνητής δεν πρέπει να πειράξει τις ακραίες τιμές διότι εμπεριέχουν πληροφόρηση που πρέπει να ληφθεί υπόψη από την εμπειρική ανάλυση. Η δεύτερη θεωρεί ότι οι ακραίες τιμές αποτελούν μία πηγή σφάλματος και άρα πρέπει να αντιμετωπισθούν με έναν από τους ακόλουθους τρόπους: (α) Διαγραφή όλων των τιμών που ευρίσκονται στο χαμηλότερο ή υψηλότερο 5% (ή 1%) του εύρους τιμών της μεταβλητής (trimming), (β) αντικατάσταση όλων των τιμών που ευρίσκονται στο χαμηλότερο 5% (ή 1%) του εύρους τιμών της μεταβλητής με την τιμή της που αντιστοιχεί στο 5% (ή 1%) του εύρους τιμών της ΚΑΙ αντικατάσταση όλων των τιμών που ευρίσκονται στο υψηλότερο 5% (ή 1%) του εύρους τιμών της μεταβλητής με την τιμή της που αντιστοιχεί στο 95% (ή 99%) του εύρους τιμών της (winsorization), και (γ) Οποιοδήποτε άλλο τρόπο υποδεικνύει η εξειδικευμένη βιβλιογραφία.

ΓΙΑ ΕΚΠΑΥΔΕΥΤΙΚΟΥΣ ΚΑΙ ΜΟΝΟ ΣΚΟΠΟΥΣ ΘΑ ΠΡΑΓΜΑΤΟΠΟΙΗΣΟΥΜΕ TRIMMING ΚΑΙ WINSORIZING ΣΤΟ ΙΔΙΟ ΔΕΙΓΜΑ. ΥΠΟ ΣΥΝΗΘΕΙΣ ΣΥΝΘΗΚΕΣ ΕΚΤΕΛΕΙΤΕ ΕΙΤΕ ΤΟ ΕΝΑ ΕΙΤΕ ΤΟ ΑΛΛΟ.

Πραγματοποιήστε trimming στο δείγμα σας (για 1% και 99%) όπως έχει διαμορφωθεί ως τώρα στο πλαίσιο του STATA αξιοποιώντας μεταξύ άλλων εντολές όπως:

- ssc install winsor2 (εγκατάσταση εντολής winsor2 – ενδέχεται να υπάρχει ήδη).
- winsor2 vr1, cuts (1 99) suffix(_new) trim

Η επιλογή suffix(_new) κατευθύνει το **STATA** να δημιουργήσει μία νέα αριθμητική μεταβλητή με προέκταση ονόματος το _new (στο παράδειγμα μας θα δημιουργηθεί η μεταβλητή vr1_new από την αρχική μεταβλητή vr1 και την προέκταση _new).

Πραγματοποιήστε winsorization στο δείγμα σας (για 5% και 5%) όπως έχει διαμορφωθεί ως τώρα στο πλαίσιο του STATA αξιοποιώντας μεταξύ άλλων εντολές όπως:

- winsor2 vr1_new, cuts (5 95) suffix(_new) trim

Βήμα 8ο Αποθήκευση του τελικού αρχείου δεδομένων του STATA

Αποθήκευση του τελικού αρχείου δεδομένων του **STATA**. Το τελικά διαμορφωμένο δείγμα παρατηρήσεων θα πρέπει να αποθηκευτεί με τη μορφή αρχείου δεδομένων **STATA** δηλαδή αποθηκευτεί με την προέκταση ονόματος .dta. Συνίσταται, οι περαιτέρω εργασίες να εκτελούνται με αντίγραφο του αρχείου αυτού για λόγους ασφάλειας και ακεραιότητας των δεδομένων.

ΓΕΝΙΚΗ ΠΑΡΑΤΗΡΗΣΗ: Θα πρέπει να συνταχθεί ένας (έστω πρόχειρος) πίνακας που περιγράφει τον αρχικό αριθμό των παρατηρήσεων, τις αιτίες διαγραφών παρατηρήσεων, τον αριθμό των παρατηρήσεων που διαγράφθηκαν για κάθε αιτία διαγραφής και ο τελικός αριθμός των παρατηρήσεων που χρησιμοποιήθηκε για εμπειρική ανάλυση.