Ανάλυση Δεδομένων στη Λογιστική και Χρηματοοικονομική

Ενότητα 8η

Στασιμότητα, Ψευδομεταβλητές και Panel Data

Περιγραφή Ενότητας

1. Στασιμότητα (Stationarity)
2. Ψευδομεταβλητές
3. Panel data

Στασιμότητα χρονοσειρών (stationarity)

Γιατί πρέπει να ελέγχουμε τη Μη Στασιμότητα (Non-Stationarity);

- Η στασιμότητα ή μη μια σειράς μπορεί να επηρεάσει σημαντικά τη συμπεριφορά ή τις ιδιότητές της, π.χ. η επιμονή των σοκ θα είναι infinite για τις μη στατικές σειρές
- Παραπλανητικές παλινδρομήσεις. Εάν δύο μεταβλητές είναι σε τάση με την πάροδο του χρόνου (trending over time), μια παλινδρόμηση της μιας μεταβλητής στην άλλη μεταβλητή θα μπορούσε να έχει υψηλό R^2 ακόμα κι αν αυτές οι δύο μεταβλητές δεν έχουν καμία σχέση μεταξύ τους.
- Εάν οι μεταβλητές στο μοντέλο παλινδρόμησης είναι μη-στάσιμες, τότε μπορεί να αποδειχθεί ότι οι υποθέσεις/παραδοχές για την ασυμπτωτική ανάλυση δεν θα είναι έγκυρες. Με άλλα λόγια, οι στατιστικές t δεν θα ακολουθούν την κατανομή *t*, επομένως δεν μπορούμε να πραγματοποιήσουμε έγκυρους ελέγχους υποθέσεων σχετικά με τις παραμέτρους της παλινδρόμησης.

Δύο τύποι Μη-Στασιμότητας (Non-Stationarity)

- Υπάρχουν διάφοροι ορισμοί για τη μη-στασιμότητα.
- Σε αυτή τη διάλεξη, αναφερόμαστε στην αδύναμη μορφή της ή αλλιώς στη στασιμότητα συνδιακύμανσης (weak form or covariance stationarity).
- Υπάρχουν δύο μοντέλα που χρησιμοποιούνται συχνά για τον χαρακτηρισμό της μη στασιμότητας:

  - Το μοντέλο τυχαίου περίπατου με μετατόπιση (the random walk model with drift):

$$y_t = \mu + y_{t-1} + u_t$$

  - και το μοντέλο ντετερμινιστικής τάσης (deterministic trend process):

$$y_t = \alpha + \beta t + u_t$$

όπου ο u_t είναι iid και στις δύο περιπτώσεις.

Στοχαστική Μη Στασιμότητα (Stochastic Non-Stationarity)

- Σημειώστε ότι το μοντέλο τυχαίου περίπατου με μετατόπιση (the random walk model with drift) μπορεί να γενικευτεί στην περίπτωση που η y_t είναι μια «εκρηκτική» (explosive) διαδικασία:

$$y_t = \mu + \phi y_{t-1} + u_t$$

όπου $\phi > 1$.

- Τυπικά, η περίπτωση της «εκρηκτικής» (explosive) διαδικασίας αγνοείται και χρησιμοποιούμε $\phi = 1$ για να χαρακτηρίσουμε μια μη στάσιμη (non-stationarity) διαδικασία επειδή
  - $\phi > 1$ δεν περιγράφει πολλές σειρές δεδομένων στις επιστήμες των οικονομικών και χρηματοοικονομικών.
  - $\phi > 1$ έχει μια διαισθητικά μη ελκυστική ιδιότητα: τα σοκ στο σύστημα δεν είναι μόνο επίμονα στο χρόνο, αλλά διαδίδονται έτσι ώστε ένα δεδομένο σοκ να έχει μια όλο και μεγαλύτερη επιρροή.

Detrending μια στοχαστική μη-στάσιμη σειρά(Detrending Stochastically Non-stationary Series

- Θυμηθείτε τις δύο μορφές μια μη στάσιμης (non-stationarity) διαδικασίας, ο τυχαίος περίπατος με μετατόπιση (random walk with drift):

$$y_t = \mu + y_{t-1} + u_t \qquad (1)$$

και η trend-stationary διαδικασία

$$y_t = \alpha + \beta t + u_t$$

- Και οι δύο θα απαιτήσουν διαφορετικές «θεραπείες» για να γίνουν στάσιμες. Η δεύτερη περίπτωση είναι γνωστή ως ντετερμινιστική μη-στασιμότητα (deterministic non-stationarity) και το ζητούμενο είναι η αφαίρεση της τάσης (detrending):
  - Αφαιρέστε από την εξίσωση (1) την y_{t-1} και από τις δύο πλευρλες

$$y_t - y_{t-1} = \mu + u_t$$

$$\Delta y_t = \mu + u_t$$

  - Λέμε ότι έχουμε δημιουργήσει στασιμότητα παίρνοντας τις πρώτες διαφορές μια φορά “differencing once”.

Detrending μια σειρά: Χρησιμοποιώντας τη σωστή μέθοδο

- Παρόλο που οι trend-stationary και difference-stationary σειρές έχουν και οι δύο «τάση» με την πάροδο του χρόνου (“trending” over time), η σωστή προσέγγιση πρέπει να χρησιμοποιείται σε κάθε περίπτωση.

- Αν πάρουμε τις πρώτες διαφορές (first difference) στην trend-stationary σειρά, αυτό θα «αφαιρούσε» τη μη στασιμότητα, αλλά σε βάρος της εισαγωγής ενός MA (1) στα σφάλματα.

- Αντίθετα, αν προσπαθήσουμε να κάνουμε detrend μια σειρά που έχει στοχαστική τάση, τότε δεν θα αφαιρέσουμε τη μη στασιμότητα.

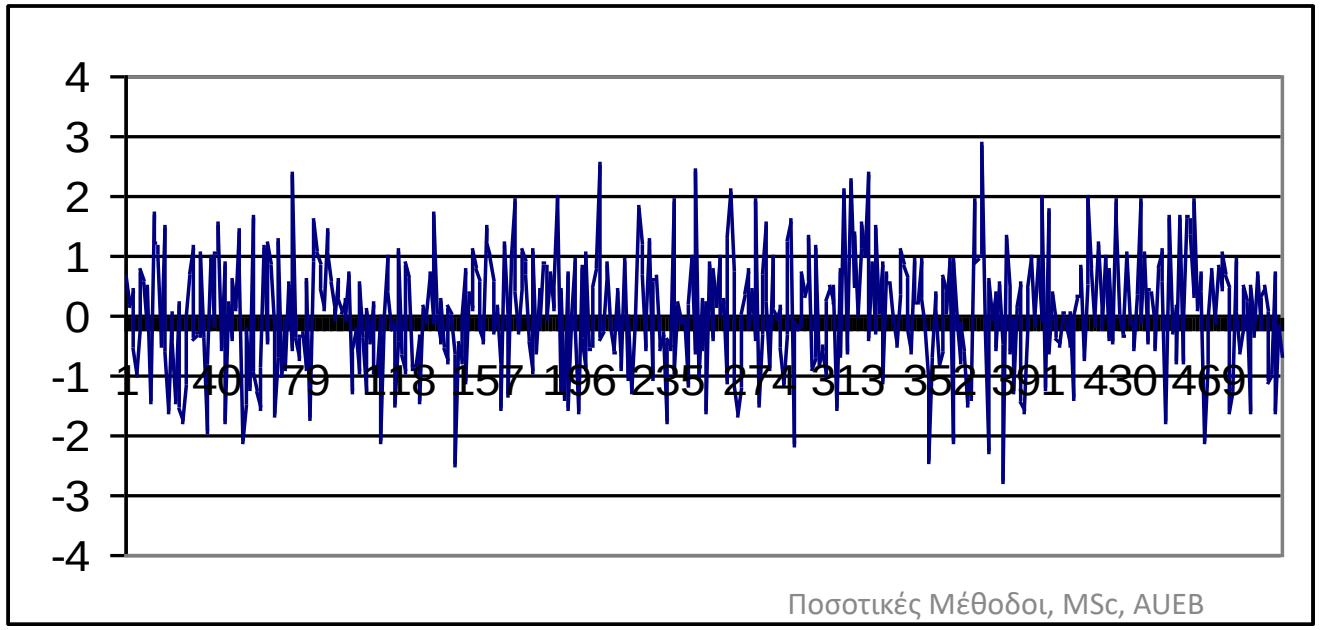
- Θα επικεντρωθούμε τώρα στο στοχαστικό μοντέλο μη στασιμότητας, καθώς η ντετερμινιστική μη στασιμότητα δεν περιγράφει επαρκώς τις περισσότερες σειρές στην επιστήμη των οικονομικών ή χρηματοοικονομικών.

Διαγράμματα για διάφορες στοχαστικές διαδικασίες:

Λευκός Θόρυβος (A White Noise Process)

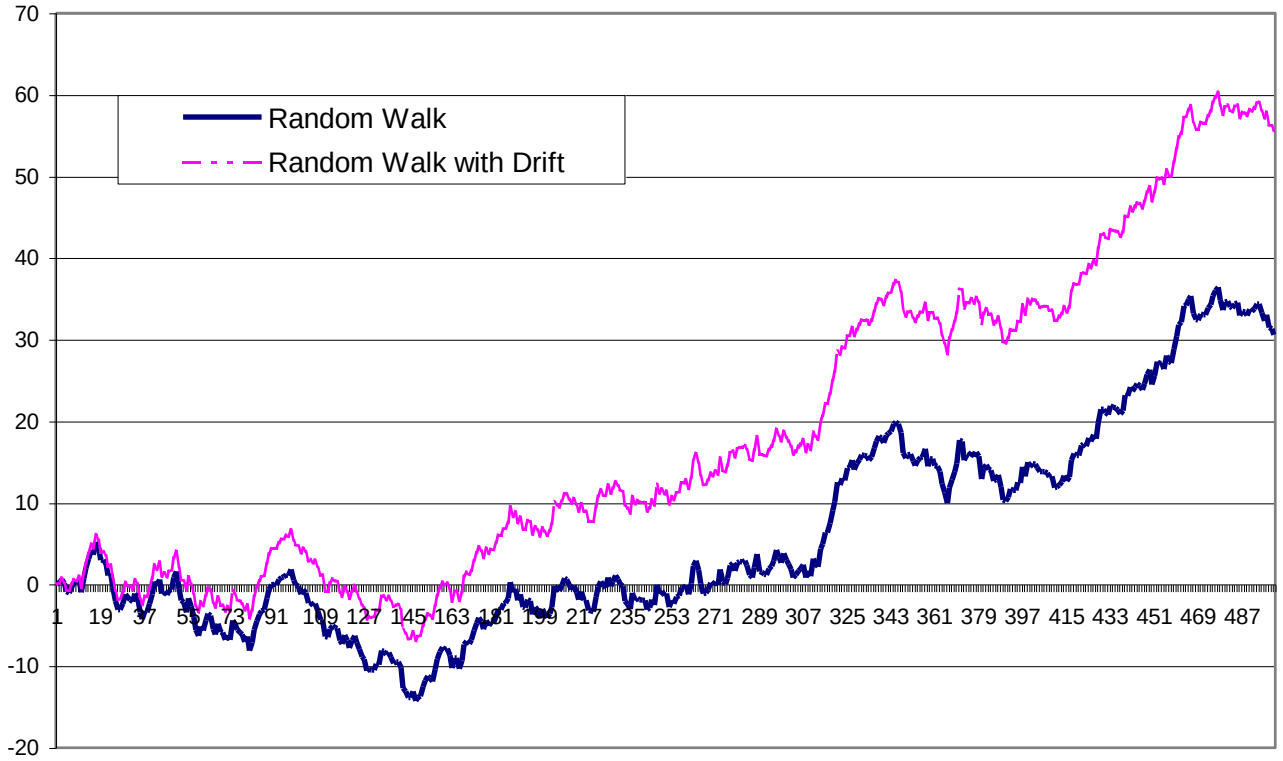
❖Η διαδικασία λευκού θορύβου είναι αυτή χωρίς (ουσιαστικά) καμία διακριτή δομή. Ο ορισμός της διαδικασίας λευκού θορύβου είναι ο ακόλουθος

$$E(y_t) = \mu$$
$$Var(y_t) = \sigma^2$$
$$\gamma_{t-r} = \begin{cases} \sigma^2 & if \quad t = r \\ 0 & otherwise \end{cases}$$

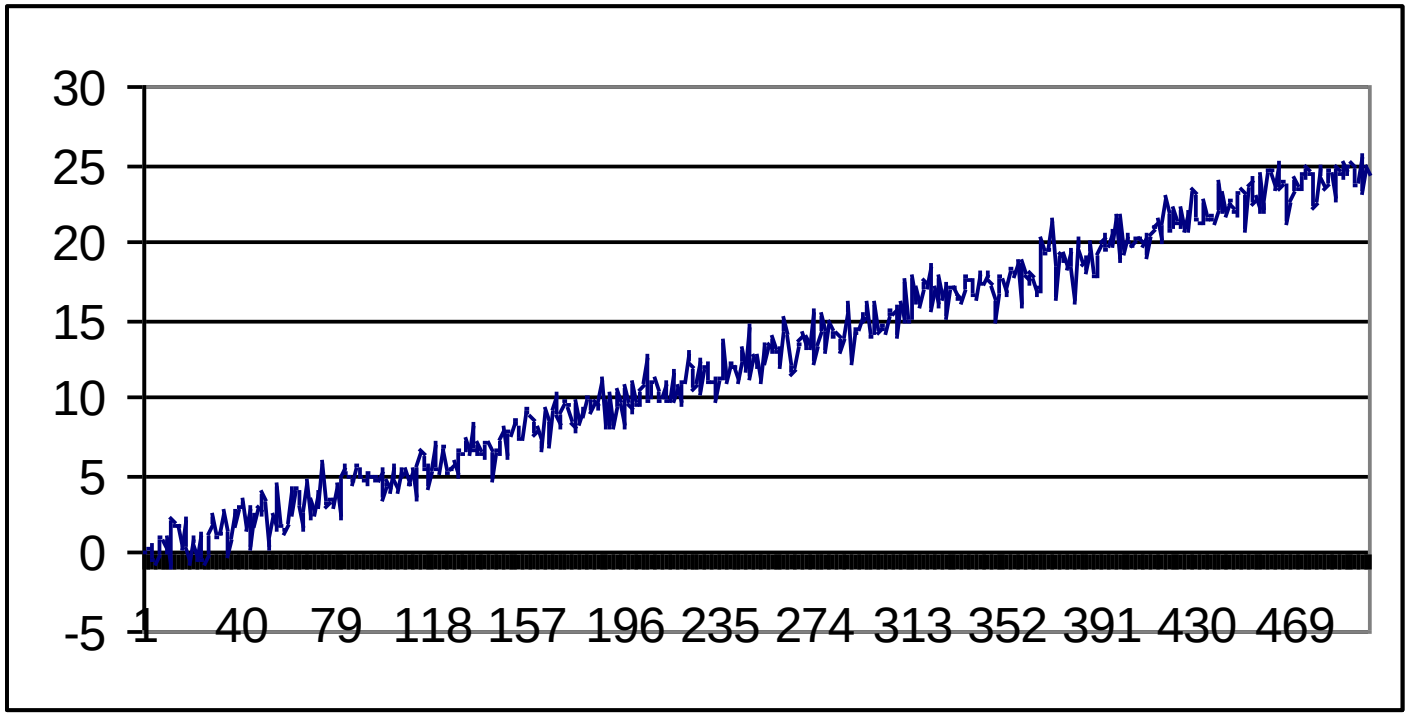


Διαγράμματα για διάφορες στοχαστικές διαδικασίες:

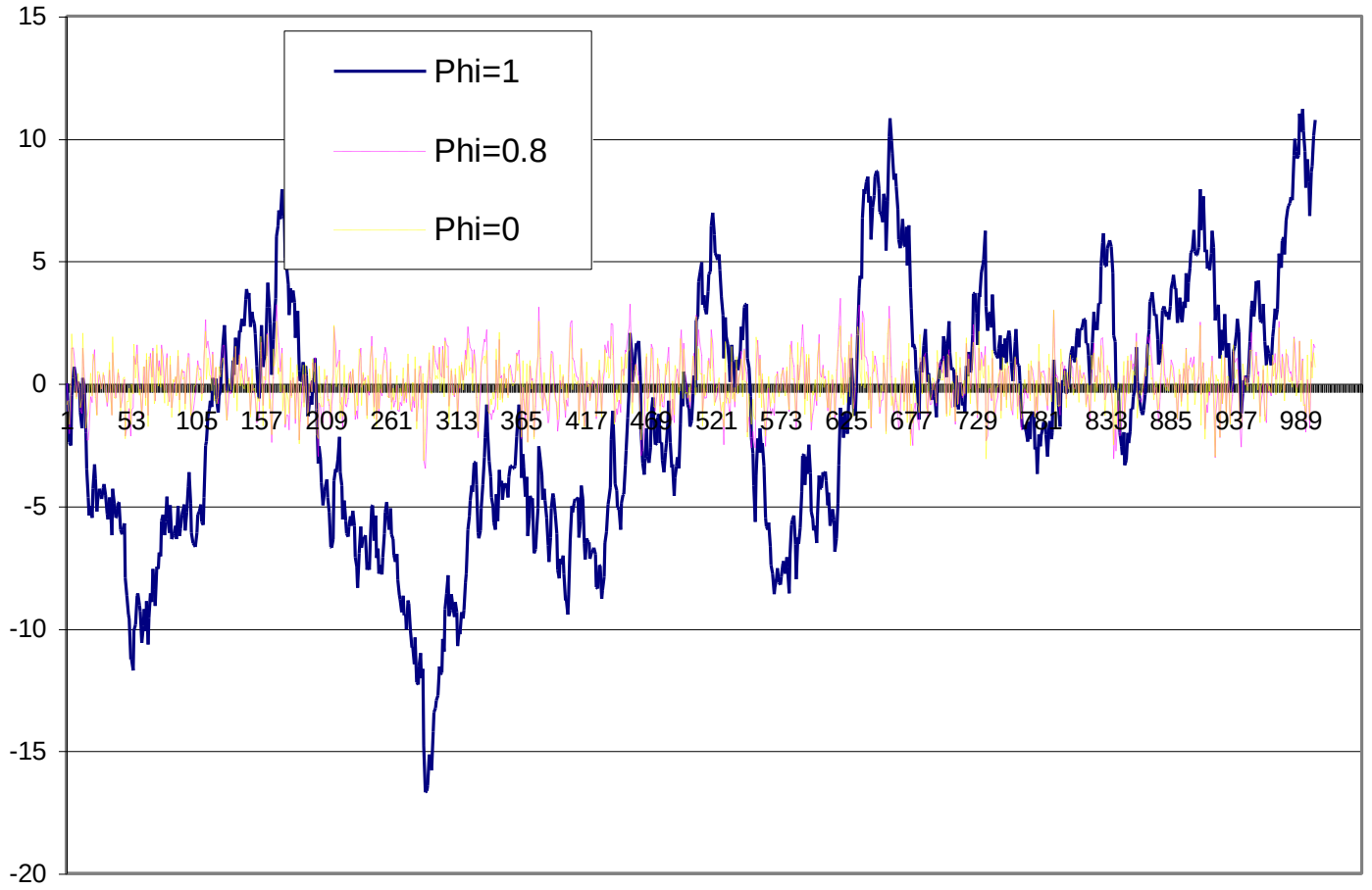
Ένας τυχαίος περίπατος και ένας τυχαίος περίπατος με μετατόπιση (A Random Walk and a Random Walk with Drift)



Διαγράμματα για διάφορες στοχαστικές διαδικασίες: A Deterministic Trend Process



Αυτοπαλίνδρομα μοντέλα με διαφορετικές τιμές του ϕ (0, 0.8, 1)



Ορισμός της μη-στασιμότητας (Non-Stationarity)

- Έστω το stochastic trend μοντέλο:

$$y_t = y_{t-1} + u_t$$

ή

$$\Delta y_t = u_t$$

- Μπορούμε να γενικεύσουμε αυτήν την έννοια για να εξετάσουμε την περίπτωση όπου η σειρά περιέχει περισσότερες από μία «μοναδιαίες ρίζες». Δηλαδή, θα πρέπει να πάρουμε την πρώτη διαφορά, Δ, περισσότερες από μία φορές για να έχουμε στασιμότητα.

  <u>Ορισμός</u>

  Εάν μια μη-στάσιμη σειρά y_t πρέπει να «υποστεί» d φορές τη διαφορά πριν γίνει στάσιμη, τότε λέμε ότι είναι ολοκληρωμένη (integrated) τάξης d. Γράφεται ως εξής: $y_t \sim \mathrm{I}(d)$.

  Άρα εάν $y_t \sim \mathrm{I}(d)$ τότε $\Delta^d y_t \sim \mathrm{I}(0)$.

  Μια I(0) σειρά είναι στάσιμη

  Μια I(1) σειρά έχει μοναδιαία ρίζα,

$$\pi.\chi.\ y_t = y_{t-1} + u_t$$

Χαρακτηριστικά των σειρών I(0), I(1) and I(2)

- Μια σειρά I(2) series έχει δύο μοναδιαίες ρίζες (unit roots), άρα θα χρειαστεί να παρούμε τις διαφορές δύο φορές (differencing twice) ώστε να έχουμε στασιμότητα (stationarity)

- Οι σειρές I(1) και I(2) μπορεί να απομακρυνθούν πολύ από τη μέση τιμή τους και να διασχίσει αυτή τη μέση τιμή σπάνια.

- Η σειρά I(0) θα πρέπει να διασχίσει τη μέση τιμή συχνά.

- Η πλειοψηφία των οικονομικών και χρηματοοικονομικών σειρών περιέχει μια μοναδιαία ρίζα (unit root), αν και ορισμένες είναι στάσιμες, ενώ οι τιμές καταναλωτή (consumer prices) έχουν υποστηριχθεί ότι έχουν 2 μονάδιαιες ρίζες.

Πώς ελέγχουμε για μοναδιαία ρίζα (unit root);

- Οι πρώτες και πρωτοποριακές εργασίες για τον έλεγχο μονάδας ρίζας σε χρονικές σειρές έγιναν από τους Dickey και Fuller (Dickey and Fuller 1979, Fuller 1976). Ο βασικός στόχος του ελέγχου είναι ο έλεγχος της μηδενικής υπόθεσης $\phi=1$ στο παρακάτω μοντέλο:

$$y_t = \phi y_{t-1} + u_t$$

  έναντι της μονόπλευρης (one-sided) εναλλακτικής $\phi<1$. Άρα έχουμε

  H_0: η σειρά έχει μοναδιαία ρίζα (unit root)

  έναντι H_1: η σειρά είναι στάσιμη (stationary)

- Συνήθως χρησιμοποιούμε την ακόλουθη παλινδρόμηση:

$$\Delta y_t = \psi y_{t-1} + u_t$$

  έτσι ώστε ένας έλεγχος του $\phi=1$ είναι ισοδύναμος ένος ελέγχου του $\psi=0$ (αφού $\phi-1=\psi$).

Διαφορετικές μορφές ελέγχου DF

- Οι έλεγχοι Dickey Fuller είναι γνωστοί και ως έλεγχοι τ : τ, τ_μ, τ_τ.
- Η μηδενική (H_0) και η εναλλακτική (H_1) υποθέσεις είναι

i) $H_0: y_t = y_{t-1} + u_t$

$H_1: y_t = \phi y_{t-1} + u_t,\ \phi < 1$

ο οποίος είναι ένας έλεγχος για τυχαίο περίπατο (random walk) έναντι μιας στάσιμης αυτοπαλίνδρομης χρονοσειράς τάξης 1 (stationary autoregressive process of order one) (AR(1))

ii) $H_0: y_t = y_{t-1} + u_t$

$H_1: y_t = \phi y_{t-1} + \mu + u_t,\ \phi < 1$

ο οποίος είναι ένας έλεγχος για τυχαίο περίπατο (random walk) έναντι μιας στάσιμης αυτοπαλίνδρομης χρονοσειράς τάξης 1 με μετατόπιση (stationary autoregressive process of order one with drift)

iii) $H_0: y_t = y_{t-1} + u_t$

$H_1: y_t = \phi y_{t-1} + \mu + \lambda t + u_t,\ \phi < 1$

ο οποίος είναι ένας έλεγχος για τυχαίο περίπατο (random walk) έναντι μιας στάσιμης αυτοπαλίνδρομης χρονοσειράς τάξης 1 με μετατόπιση και χρονική τάση (stationary autoregressive process of order one with drift and time trend)

Ο έλεγχος Dickey Fuller (ADF)

- Οι παραπάνω έλεγχοι ισχύουν μόνο εάν ο u_t είναι λευκός θόρυβος (white noise). Συγκεκριμένα, ο u_t θα χαρακτηρίζεται από αυτοσυσχέτιση εάν υπήρχε αυτοσυσχέτιση στην εξαρτημένη μεταβλητή της παλινδρόμησης (Δy_t). Η λύση είναι να "αυξήσουμε" τον έλεγχο χρησιμοποιώντας p χρονικές υστερήσεις της εξαρτημένης μεταβλητής. Το εναλλακτικό μοντέλο στην περίπτωση (i) θα μπορούσε να γραφτεί ως:

$$\Delta y_t = \psi y_{t-1} + \sum_{i=1}^{p} \alpha_i \Delta y_{t-i} + u_t$$

- Οι ίδιες κριτικές τιμές από τους DF πίνακες χρησιμοποιούνται όπως πριν. Ένα πρόβλημα προκύπτει τώρα στον προσδιορισμό του βέλτιστου αριθμού υστερήσεων της εξαρτημένης μεταβλητής.
- Υπάρχουν 2 τρόποι

❖χρησιμοποιήστε τη συχνότητα των δεδομένων για να αποφασίσετε.

❖χρησιμοποιήστε κριτηρία πληροφόρησης.

Testing for Higher Orders of Integration

- Έστω η παρακάτω παλινδρόμηση:

$$\Delta y_t = \psi y_{t-1} + u_t$$

  Κάνουμε τον εξής έλεγχο H_0: $\psi=0$ vs. H_1: $\psi<0$.

- Αν H_0 απορριφθεί συμπεραίνουμε ότι η y_t δεν έχει μοναδιαία ρίζα (unit root).
- Αλλά τι συμπεράσματα μπορούμε να βγάλουμε εάν η H_0 δεν απορριφθεί; Η σειρά θα έχει μοναδιαία ρίζα (unit root), αλλά είναι μόνο αυτό; Όχι! Εάν $y_t \sim I(2)$ ακόμα δεν θα την είχαμε τη απορρίψει. Άρα τώρα χρειαζόμαστε τον παρακάτω έλεγχο υπόθεσης:

$$H_0: y_t \sim I(2) \text{ vs. } H_1: y_t \sim I(1)$$

  Θα συνεχίζουμε να ελέγχουμε για μοναδιαία ρίζα (unit root) μέχρι να απορρίψουμε την H_0

- Τώρα τρέχουμε την παλινδρόμηση της $\Delta^2 y_t$ πάνω στην Δy_{t-1} (+ χρονικές υστερήσεις της $\Delta^2 y_t$ εάν είναι απαραίτητο).
- Τώρα χρειαζόμαστε τον παρακάτω έλεγχο υπόθεσης: H_0: $\Delta y_t \sim I(1)$ που είναι ισοδύναμο του H_0: $y_t \sim I(2)$.
- Σε αυτήν την περίπτωση, εάν δεν απορρίψουμε (απίθανο) την H_0, συμπεραίνουμε ότι η y_t είναι τουλάχιστον I(2).

Στασιμότητα –περίπτωση 1

```
. dfuller PRICE_GREECEINDEX, regress lags(0)

Dickey-Fuller test for unit root                   Number of obs   =       416

                     ---------- Interpolated Dickey-Fuller ---------
            Test         1% Critical       5% Critical      10% Critical
          Statistic           Value             Value            Value
------------------------------------------------------------------------------
 Z(t)       -1.262            -3.447            -2.873           -2.570
------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.6462
```

D. PRICE_GREECEINDEX	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
PRICE_GREECEINDEX L1.	-.0074127	.0058724	-1.26	0.208	-.0189561	.0041307
_cons	6.471555	6.662104	0.97	0.332	-6.624215	19.56732

Μηδενική υπόθεση: Έχε μοναδιαία ρίζα (Unit Root)

Εναλλακτική υπόθεση : Δεν έχει μοναδιαία ρίζα (Not unit Root)

P-value = 0.6462> 0.05

$$\Delta y_t = \mu + y_{t-1} + u_t$$

Δεν απορρίπτουμε τη μηδενική υπόθεση,άρα η μεταβλητή PRICE_GREECEINDEX είναι μη-στάσιμη.

Στασιμότητα –περίπτωση 2

Μηδενική υπόθεση: Έχε μοναδιαία ρίζα (Unit Root)

Εναλλακτική υπόθεση : Δεν έχει μοναδιαία ρίζα (Not unit Root)

P-value = 0.000< 0.05

```
. dfuller d.PRICE_GREECEINDEX, regress lags(0)

Dickey-Fuller test for unit root                   Number of obs   =       415

                                 ------------ Interpolated Dickey-Fuller ------------
                 Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value            Value
------------------------------------------------------------------------------
 Z(t)           -18.089            -3.447            -2.873           -2.570
------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.0000
```

D2. PRICE_GREECEINDEX	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
PRICE_GREECEINDEX LD.	-.8841065	.0488762	-18.09	0.000	-.9801836	-.7880294
_cons	.1557546	4.427551	0.04	0.972	-8.547591	8.8591

Απορρίπτουμε τη μηδενική υπόθεση,άρα η μεταβλητή PRICE_GREECEINDEX είναι στάσιμη.

Ψευδομεταβλητές (dummies)

- Οι ψευδομεταβλητές (dummy variables) μας επιτρέπουν να κατασκευάσουμε μοντέλα στα οποία ορισμένες ή όλες οι παράμετροι του μοντέλου παλινδρόμησης, συμπεριλαμβανομένου του σταθερού όρου, αλλάζουν για ορισμένες παρατηρήσεις στο δείγμα.
- Εξετάστε ένα μοντέλο για να προβλέψετε την αξία ενός σπιτιού σε συνάρτηση με τα χαρακτηριστικά του:
- Μέγεθος, τοποθεσία, αριθμός υπνοδωματίων, ηλικία
  - Εξετάστε αρχικά τα τετραγωνικά μέτρα: $PRICE = \beta_1 + \beta_2 SQFT + e$
- Το β_2 είναι η αξία ενός επιπλέον τετραγωνικού μέτρου της επιφάνειας σπιτιού και το β_1 είναι η αξία της γης μόνη.

- Πώς λαμβάνουμε υπόψιν την τοποθεσία, η οποία είναι μια ποιοτική μεταβλητή;
- Οι ψευδομεταβλητές χρησιμοποιούνται για να λάβουν υπόψη τους ποιοτικούς παράγοντες στα οικονομετρικά μοντέλα.
- Συνήθως λαμβάνουν μόνο δύο τιμές, συνήθως ένα (1) ή μηδέν (0), για να υποδείξουν την παρουσία ή την απουσία ενός χαρακτηριστικού ή για να υποδείξουν εάν μια συνθήκη είναι αληθής ή ψευδής.
- Ουσιαστικά με τη χρήση ψευδομεταβλητών δημιουργούμε μια αριθμητική μεταβλητή για ένα ποιοτικό, μη αριθμητικό χαρακτηριστικό.

- Γενικά, ορίζουμε μια ψευδομεταβλητή D ως:

$$D = \begin{cases} 1 & \text{if characteristic is present} \\ 0 & \text{if characteristic is not present} \end{cases}$$

- Έτσι, για να λάβουμε υπόψιν την τοποθεσία, μια ποιοτική μεταβλητή, θα είχαμε:

$$D = \begin{cases} 1 & \text{if property is in the desirable neighborhood} \\ 0 & \text{if property is not in the desirable neighborhood} \end{cases}$$

- Προσθέτοντας τη ψευδομεταβλητή στο μοντέλο μας:

$$PRICE = \beta_1 + \delta D + \beta_2 SQFT + e$$

- Εάν το μοντέλο μας έχει καθοριστεί σωστά, τότε:

$$(PRICE|SQFT)\begin{cases}(\beta_1 + \delta) + \beta_2 SQFT\ when\ D = 1\\ \beta_1 + \beta_2 SQFT\ when\ D = 1\end{cases}$$

- Η προσθήκη της ψευδομεταβλητής D στο μοντέλο παλινδρόμησης προκαλεί μια παράλληλη μετατόπιση στη σχέση κατά το ποσό δ.
- Οι ιδιότητες του εκτιμητή ελαχίστων τετραγώνων δεν επηρεάζονται από το γεγονός ότι μία από τις επεξηγηματικές μεταβλητές αποτελείται μόνο από μηδενικά και ένα. Το D αντιμετωπίζεται ως οποιαδήποτε άλλη επεξηγηματική μεταβλητή.
- Μπορούμε να κατασκευάσουμε μια εκτίμηση διαστήματος για το D ή μπορούμε να ελέγξουμε τη σημασία της εκτίμησης των ελαχίστων τετραγώνων.
- Η τιμή D = 0 ορίζει την ομάδα αναφοράς ή την ομάδα βάσης.
- Μπορούμε να διαλέξουμε οποιαδήποτε βάση.
- Για παράδειγμα:

$$LD = \begin{cases} 1 & \text{if property is not in the desirable neighborhood} \\ 0 & \text{if property is in the desirable neighborhood} \end{cases}$$

- Τότε το μοντέλο μας θα ήταν: $PRICE = \beta_1 + \lambda LD + \beta_2 SQFT + e$
- Ας υποθέσουμε ότι συμπεριλαμβάνουμε και το D και το LD:

$$PRICE = \beta_1 + \delta D + \lambda LD + \beta_2 SQFT + e$$

- Οι μεταβλητές D και LD είναι τέτοιες ώστε D + LD = 1.
- Επειδή η ψευδομεταβλητή $x_1 = 1$, δημιουργήσαμε ένα μοντέλο με τέλεια συγγραμμικότητα.
- Έχουμε πέσει στην παγίδα της ψευδομεταβλητής (dummy variable trap).
- Συμπεριλαμβάνοντας μόνο μία από τις ψευδομεταβλητές, η μεταβλητή που παραλείπουμε ορίζει την ομάδα αναφοράς και αποφεύγουμε το πρόβλημα.

- Ας υποθέσουμε ότι προσδιορίζουμε το μοντέλο μας ως εξής:

$$PRICE = \beta_1 + \beta_2 SQFT + \gamma(SQFT \times D) + e$$

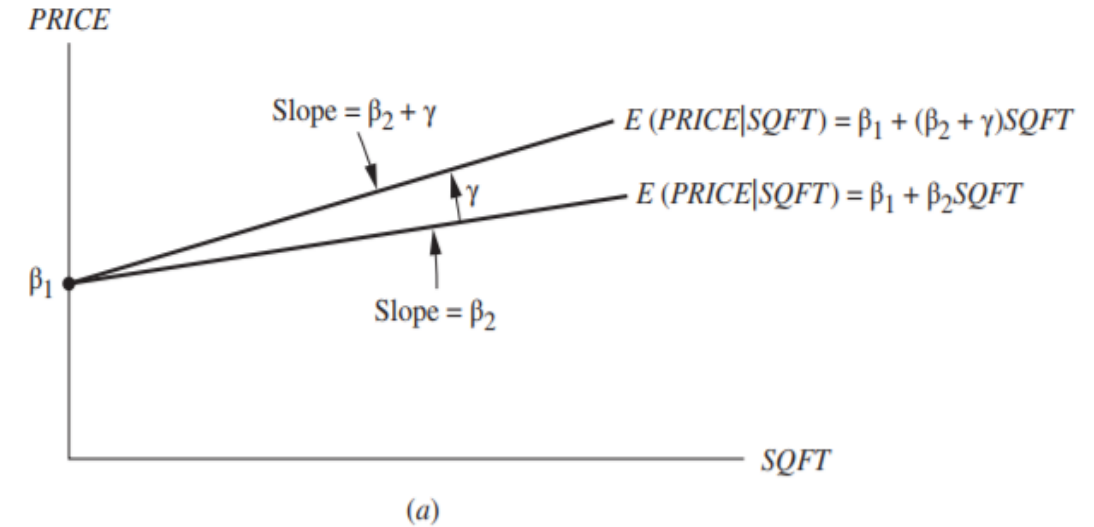
- Η νέα μεταβλητή (SQFT x D) είναι το γινόμενο του μεγέθους του σπιτιού και της ψευδομεταβλητής.
- Ονομάζεται μεταβλητή αλληλεπίδρασης (interaction variable), καθώς αποτυπώνει την επίδραση της αλληλεπίδρασης της τοποθεσίας και του μεγέθους στην τιμή του σπιτιού.
- Εναλλακτικά, ονομάζεται slope-indicator variable ή slope dummy variable, επειδή επιτρέπει μια αλλαγή στην κλίση της σχέσης.

- Τώρα μπορούμε να γράψουμε
  - $E(PRICE|SQFT, D) = \beta_1 + \beta_2 SQFT + \gamma(SQFT\ x\ D)$

$$= \begin{cases} \beta_1 + (\beta_2 + \gamma) SQFT\ when\ D = 1 \\ \beta_1 + \beta_2 SQFT\ when\ D = 0 \end{cases}$$

- Η κλίση μπορεί να εκφραστεί ως

$$\frac{\delta E(PRICE|SQFT)}{\delta SQFT} = \begin{cases} \beta_2 + \gamma\ when\ D = 1 \\ \beta_2\ when\ D = 0 \end{cases}$$



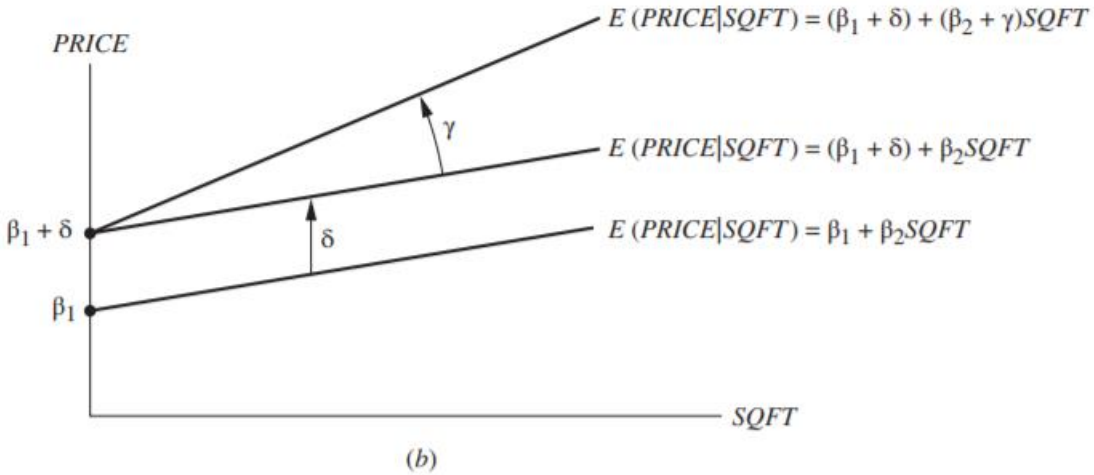


FIGURE 7.2 **(*a*) A slope-indicator variable. (*b*) Slope- and intercept-indicator variables.**

- Ας υποθέσουμε ότι η τοποθεσία του σπιτιού επηρεάζει τόσο τον σταθερό όρο όσο και την κλίση, τότε και οι δύο επιδράσεις μπορούν να ενσωματωθούν σε ένα ενιαίο μοντέλο:

$$PRICE = \beta_1 + \delta D + \beta_2 SQFT + \gamma(SQFT \times D) + e$$

- Η μεταβλητή (SQFT × D) είναι το γινόμενο του μεγέθους του σπιτιού και της ψευδομεταβλητής και ονομάζεται μεταβλητή αλληλεπίδρασης (interaction variable).

- Τώρα μπορούμε να δούμε ότι:

$$E(PRICE|SQFT) = \begin{cases} (\beta_1+\delta) + (\beta_2+\gamma)SQFT \ when \ D = 1 \\ \beta_1 + \beta_2 SQFT \ when \ D = 0 \end{cases}$$

Εφαρμογές με χρήση ψευδομεταβλητών

1. Αλληλεπιδράσεις Μεταξύ Ποιοτικών Παραγόντων

- Θεωρήστε την εξίσωση του μισθού:

$$WAGE = \beta_1 + \beta_2 EDUC + \delta_1 BLACK + \delta_2 FEMALE + \gamma\left(BLACK \times FEMALE\right) + e$$

- Η αναμενόμενη τιμή είναι

$$\mathrm{E}\,(\mathrm{WAGE}|\mathrm{EDUC}) = \begin{cases} \beta_1 + \beta_2 EDUC \;\; WHITE - MALE \\ (\beta_1 + \delta_1) + \beta_2 EDUC \;\; BLACK - MALE \\ (\beta_1 + \delta_2) + \beta_2 EDUC \;\; WHITE - FEMALE \\ (\beta_1 + \delta_1 + \delta_2 + \gamma) EDUC \;\; BLACK - FEMALE \end{cases}$$

2. Ποιοτικοί Παράγοντες με Διάφορες Κατηγορίες

- Εξετάστε το ενδεχόμενο να συμπεριλάβετε τις περιφέρειες στην εξίσωση μισθών:

$$WAGE = \beta_1 + \beta_2 EDUC + \delta_1 SOUTH + \delta_2 MIDWEST + \delta_3 WEST + e$$

- Το άθροισμα των μεταβλητών του περιφερειακού δείκτη είναι $NORTHEAST + SOUTH + MIDWEST + WEST = 1$.
- Έτσι, ο «σταθερός όρος» $x_1 = 1$ είναι ένας ακριβής γραμμικός συνδυασμός των δεικτών περιοχής.

- Η αποτυχία παράλειψης μιας ψευδομεταβλητής θα οδηγήσει στην εικονική παγίδα μεταβλητής.

- Η παράλειψη μιας ψευδομεταβλητής ορίζει μια ομάδα αναφοράς, οπότε η εξίσωσή μας είναι:

$$(\text{WAGE}|\text{EDUC}) = \begin{cases} (\beta_1 + \delta_3)EDUC\ \ WEST \\ (\beta_1+\delta_2) + \beta_2 EDUC\ \ MIDWEST \\ (\beta_1+\delta_1) + \beta_2 EDUC\ \ SOUTH \\ \beta_1 + \beta_2\ EDUC\ \ NORTHEAST \end{cases}$$

- Η παραλειπόμενη μεταβλητή δείκτη, NORTHEAST, προσδιορίζει την αναφορά.

3. Λαμβάνοντας υπόψιν το χρόνο

- Οι ψευδομεταβλητές χρησιμοποιούνται επίσης σε παλινδρομήσεις χρησιμοποιώντας δεδομένα χρονοσειρών.
- Μπορεί να θέλουμε να συμπεριλάβουμε μια επίδραση για διαφορετικές εποχές του χρόνου.
- Με το ίδιο πνεύμα με τις εποχιακές ψευδομεταβλητές, οι ετήσιες ψευδομεταβλητές χρησιμοποιούνται για την καταγραφή των επιπτώσεων του έτους που δεν μετρώνται διαφορετικά σε ένα μοντέλο.
- Ένα οικονομικό regime είναι ένα σύνολο διαρθρωτικών οικονομικών συνθηκών που υπάρχουν για μια ορισμένη περίοδο.

- Η ιδέα είναι ότι οι οικονομικές σχέσεις μπορεί να συμπεριφέρονται με έναν τρόπο σε ένα regime, αλλά μπορεί να συμπεριφέρονται διαφορετικά σε ένα άλλο.

- Ένα παράδειγμα επίδρασης regime : η πίστωση φόρου επένδυσης (investment tax credit):

$$ITC_t = \begin{cases} 1 & \text{if } t = 1962\text{-}1965,\ 1970\text{-}1986 \\ 0 & otherwise \end{cases}$$

- Το μοντέλο είναι τότε: $INV_t = \beta_1 + \delta ITC_t + \beta_2 GNP_t + \beta_3 GNP_{t-1} + e_t$

- Εάν η πίστωση φόρου ήταν επιτυχής, τότε $\delta > 0$.

Panel Data:
Πως να διαμορφώσετε τα δεδομένα σας σε μορφή panel

Πως να διαμορφώσετε τα δεδομένα σας σε μορφή panel data (reshape)

- https://stats.oarc.ucla.edu/stata/modules/reshaping-data-wide-to-long/

RESHAPING DATA WIDE TO LONG | STATA LEARNING MODULES

This module illustrates the power (and simplicity) of Stata in its ability to reshape data files. These examples take **wide** data files and reshape them into **long** form. These show common examples of reshaping data, but do not exhaustively demonstrate the different kinds of data reshaping that you could encounter.

Example #1: Reshaping data wide to long

Consider the family income data file below.

```
use https://stats.idre.ucla.edu/stat/stata/modules/faminc, clear

list
```

```
        famid   faminc96   faminc97   faminc98
  1.        3      75000      76000      77000
  2.        1      40000      40500      41000
  3.        2      45000      45400      45800
```

This is called a **wide** format since the years of data are wide. We may want the data to be **long**, where each year of data is in a separate observation. The **reshape** command can accomplish this, as shown below.

```
reshape long faminc, i(famid) j(year)

(note:  j = 96 97 98)

Data                               wide   ->   long
-----------------------------------------------------------------------------
Number of obs.                        3   ->      9
Number of variables                   4   ->      3
j variable (3 values)                     ->   year
xij variables:
             faminc96 faminc97 faminc98   ->   faminc
-----------------------------------------------------------------------------
```

The **list** command shows that the data are now in **long** form, where each **year** is represented as its own observation.

```
list

        famid       year     faminc
  1.        1         96      40000
  2.        1         97      40500
  3.        1         98      41000
  4.        2         96      45000
  5.        2         97      45400
  6.        2         98      45800
  7.        3         96      75000
  8.        3         97      76000
  9.        3         98      77000
```

Δεδομένα διαμορφωμένα σε μορφή panel data

- Panel data (γνωστά και ως longitudinal or cross-sectional time-series data) είναι ένα σύνολο δεδομένων στο οποίο η συμπεριφορά των οντοτήτων παρατηρείται σε βάθος χρόνου.
- Αυτές οι οντότητες θα μπορούσαν να είναι κράτη, εταιρείες, ιδιώτες, χώρες κ.λπ.

country	year	Y	X1	X2	X3
1	2000	6.0	7.8	5.8	1.3
1	2001	4.6	0.6	7.9	7.8
1	2002	9.4	2.1	5.4	1.1
2	2000	9.1	1.3	6.7	4.1
2	2001	8.3	0.9	6.6	5.0
2	2002	0.6	9.8	0.4	7.2
3	2000	9.1	0.2	2.6	6.4
3	2001	4.8	5.9	3.2	6.4
3	2002	9.1	5.2	6.9	2.1

Fixed Effects (FE)

- Τα Panel data σας επιτρέπουν να ελέγχετε για μεταβλητές που δεν μπορείτε να παρατηρήσετε ή να μετρήσετε, όπως πολιτισμικούς παράγοντες ή διαφορές στις επιχειρηματικές πρακτικές μεταξύ των εταιρειών, ή μεταβλητές που αλλάζουν με την πάροδο του χρόνου αλλά όχι μεταξύ οντοτήτων (δηλαδή εθνικές πολιτικές, ομοσπονδιακοί κανονισμοί, διεθνείς συμφωνίες κ.λπ.).
- Αυτό σημαίνει ότι αντιπροσωπεύει την ατομική ετερογένεια (individual heterogeneity).
- Με τα Panel data μπορείτε να συμπεριλάβετε μεταβλητές σε διαφορετικά επίπεδα ανάλυσης (δηλαδή μαθητές, σχολεία, περιφέρειες, πολιτείες) κατάλληλες για πολυεπίπεδη ή ιεραρχική μοντελοποίηση.
- Ορισμένα μειονεκτήματα είναι ζητήματα συλλογής δεδομένων (δηλαδή σχεδιασμός δειγματοληψίας, κάλυψη), μη απόκριση στην περίπτωση μικροπλαισίων ή εξάρτηση μεταξύ των χωρών στην περίπτωση μακροπλαισίων (δηλαδή συσχέτιση μεταξύ χωρών)

Fixed Effects (FE)

- Τα FE διερευνούν τη σχέση μεταξύ των μεταβλητών πρόβλεψης και αποτελέσματος σε μια οντότητα (χώρα, άτομο, εταιρεία κ.λπ.). Κάθε οντότητα έχει τα δικά της μεμονωμένα χαρακτηριστικά που μπορεί να επηρεάσουν ή να μην επηρεάσουν τις προγνωστικές μεταβλητές (για παράδειγμα, το να είσαι άνδρας ή γυναίκα θα μπορούσε να επηρεάσει τη γνώμη για ένα συγκεκριμένο θέμα ή το πολιτικό σύστημα μιας συγκεκριμένης χώρας θα μπορούσε να έχει κάποια επίδραση στο εμπόριο ή στο ΑΕΠ, ή οι επιχειρηματικές πρακτικές μιας εταιρείας μπορεί να επηρεάσουν την τιμή της μετοχής της). Όταν χρησιμοποιούμε FE, υποθέτουμε ότι κάτι μέσα στο άτομο μπορεί να επηρεάσει ή να κάνει μεροληπτικές τις μεταβλητές πρόβλεψης ή αποτελέσματος και πρέπει να το ελέγξουμε.
- Αυτή είναι η λογική πίσω από την υπόθεση της συσχέτισης μεταξύ του όρου σφάλματος της οντότητας και των μεταβλητών πρόβλεψης. Το FE αφαιρεί την επίδραση αυτών των αμετάβλητων χρονικά χαρακτηριστικών, ώστε να μπορούμε να εκτιμήσουμε την καθαρή επίδραση των προγνωστικών παραγόντων στη μεταβλητή αποτελέσματος.
- Μια άλλη σημαντική υπόθεση του μοντέλου FE είναι ότι αυτά τα χρονικά αμετάβλητα χαρακτηριστικά είναι μοναδικά για το άτομο και δεν θα πρέπει να συσχετίζονται με άλλα μεμονωμένα χαρακτηριστικά. Κάθε οντότητα είναι διαφορετική, επομένως ο όρος σφάλματος της οντότητας και η σταθερά (η οποία καταγράφει μεμονωμένα χαρακτηριστικά) δεν πρέπει να συσχετίζονται με τους άλλους. Εάν οι όροι σφάλματος συσχετίζονται, τότε το FE δεν είναι κατάλληλο, καθώς τα συμπεράσματα μπορεί να μην είναι σωστά και πρέπει να μοντελοποιήσετε αυτή τη σχέση (πιθανώς χρησιμοποιώντας random effects), αυτή είναι η κύρια λογική για τη Hausman test.

The equation for the fixed effects model becomes:

$$Y_{it} = \beta_1 X_{it} + \alpha_i + u_{it} \qquad \text{[eq.1]}$$

Where

- α_i (i=1….n) is the unknown intercept for each entity (n entity-specific intercepts).
- Y_{it} is the dependent variable (DV) where i = entity and t = time.
- X_{it} represents one independent variable (IV),
- β_1 is the coefficient for that IV,
- u_{it} is the error term

Another way to see the fixed effects model is by using binary variables. So the equation for the fixed effects model becomes:

$$Y_{it} = \beta_0 + \beta_1 X_{1,it} + \ldots + \beta_k X_{k,it} + \gamma_2 E_2 + \ldots + \gamma_n E_n + u_{it} \qquad \text{[eq.2]}$$

Where

- Y_{it} is the dependent variable (DV) where i = entity and t = time.
- $X_{k,it}$ represents independent variables (IV),
- β_k is the coefficient for the IVs,
- u_{it} is the error term
- E_n is the entity n. Since they are binary (dummies) you have n-1 entities included in the model.
- γ_2 Is the coefficient for the binary repressors (entities)

Both eq.1 and eq.2 are equivalents:

- “The key insight is that if the unobserved variable does not change over time, then any changes in the dependent variable must be due to influences other than these fixed characteristics.” (Stock and Watson, 2003, p.289-290).
- **Fixed-effects will not work well with data for which within-cluster variation is minimal or for slow changing variables over time.**

Panel Data:
Εκτίμηση panel data regressions (fixed effects)

The Stata command to run fixed/random effecst is `xtreg`.

Before using `xtreg` you need to set Stata to handle panel data by using the command `xtset`. type:

```
xtset country year
```

```
. xtset country year
        panel variable:  country (strongly balanced)
         time variable:  year, 1990 to 1999
                 delta:  1 unit
```

In this case "`country`" represents the entities or panels (*i*) and "`year`" represents the time variable (*t*).

The note "`(strongly balanced)`" refers to the fact that all countries have data for all years. If, for example, one country does not have data for one year then the data is unbalanced. Ideally you would want to have a balanced dataset but this is not always the case, however you can still run the model.

NOTE: If you get the following error after using `xtset`:

```
varlist:  country:  string variable not allowed
```

You need to convert '`country`' to numeric, type:

```
encode country, gen(country1)
```

Use '`country1`' instead of '`country`' in the `xtset` command

OLS regression

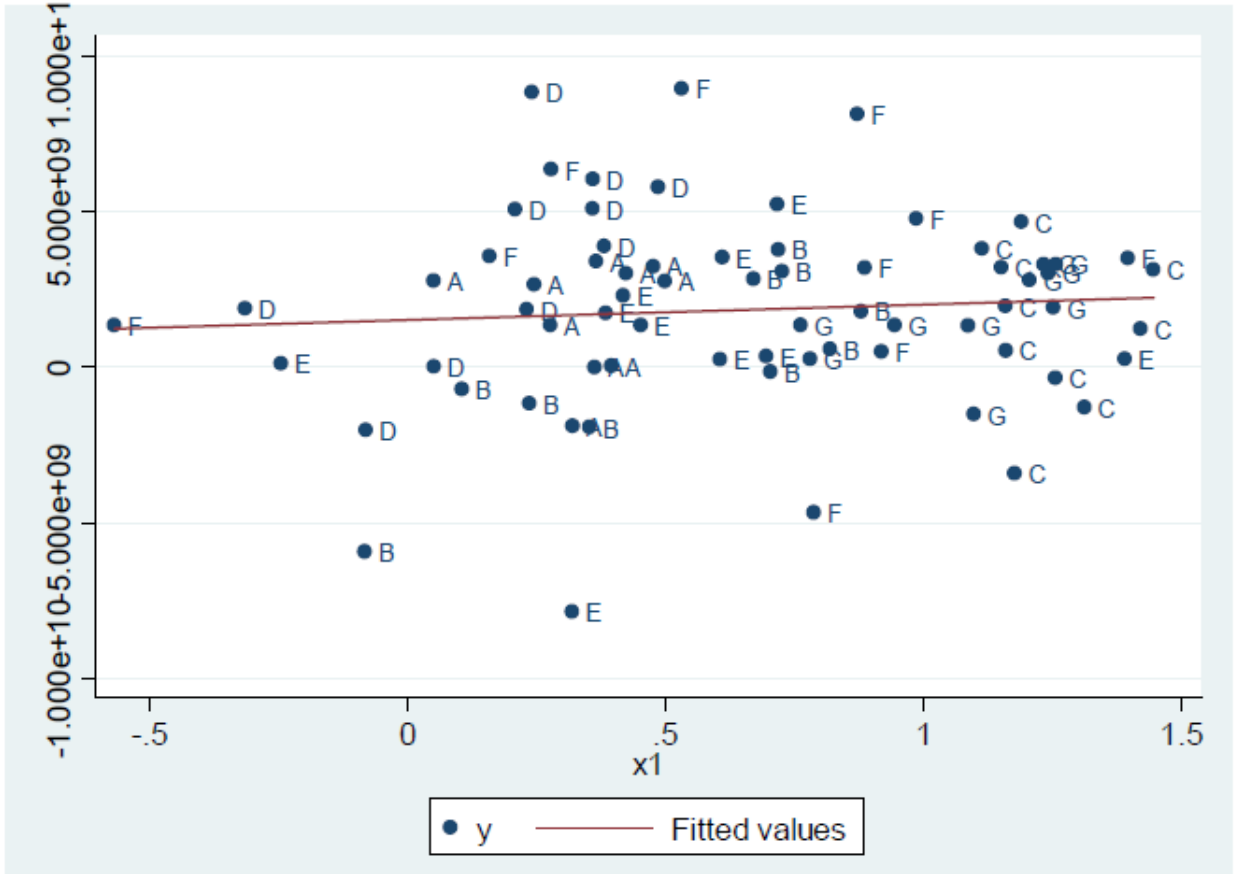
```
. regress y x1
```

Source	SS	df	MS
Model	3.7039e+18	1	3.7039e+18
Residual	6.2359e+20	68	9.1705e+18
Total	6.2729e+20	69	9.0912e+18

Number of obs	70
F(1, 68)	0.40
Prob > F	0.5272
R-squared	0.0059
Adj R-squared	-0.0087
Root MSE	3.0e+09

y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
x1	4.95e+08	7.79e+08	0.64	0.527	-1.06e+09	2.05e+09
_cons	1.52e+09	6.21e+08	2.45	0.017	2.85e+08	2.76e+09

```
twoway scatter y x1,
mlabel(country) || lfit y x1,
clstyle(p2)
```



Fixed Effects using least squares dummy variable model (LSDV)

```
. xi: regress y x1 i.country
i.country         _Icountry_1-7       (naturally coded; _Icountry_1 omitted)
```

Source	SS	df	MS
Model	1.4276e+20	7	2.0394e+19
Residual	4.8454e+20	62	7.8151e+18
Total	6.2729e+20	69	9.0912e+18

Number of obs	70
F(7, 62)	2.61
Prob > F	0.0199
R-squared	0.2276
Adj R-squared	0.1404
Root MSE	2.8e+09

y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
x1	2.48e+09	1.11e+09	2.24	0.029	2.63e+08	4.69e+09
_Icountry_2	-1.94e+09	1.26e+09	-1.53	0.130	-4.47e+09	5.89e+08
_Icountry_3	-2.60e+09	1.60e+09	-1.63	0.108	-5.79e+09	5.87e+08
_Icountry_4	2.28e+09	1.26e+09	1.81	0.075	-2.39e+08	4.80e+09
_Icountry_5	-1.48e+09	1.27e+09	-1.17	0.247	-4.02e+09	1.05e+09
_Icountry_6	1.13e+09	1.29e+09	0.88	0.384	-1.45e+09	3.71e+09
_Icountry_7	-1.87e+09	1.50e+09	-1.25	0.218	-4.86e+09	1.13e+09
_cons	8.81e+08	9.62e+08	0.92	0.363	-1.04e+09	2.80e+09

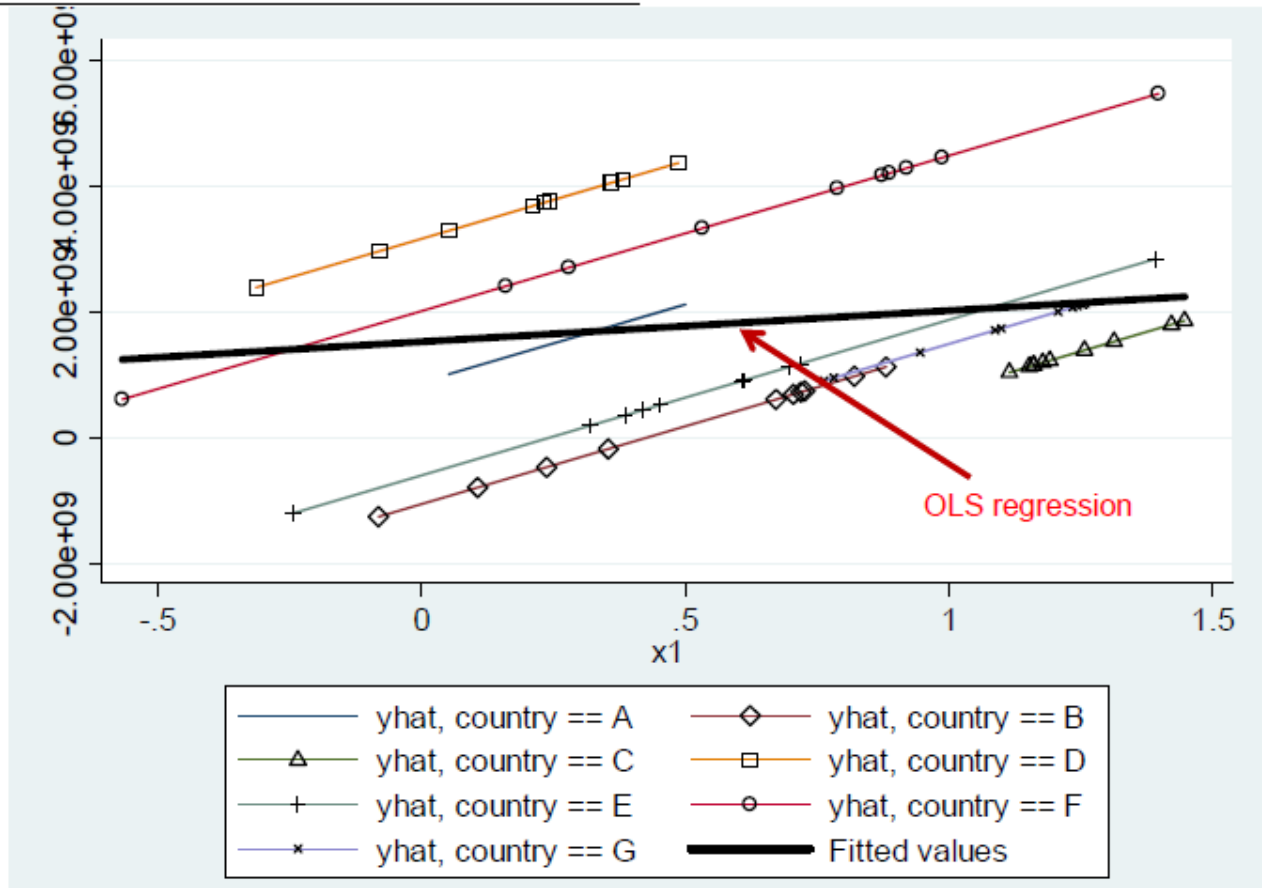
```
xi: regress y x1 i.country

predict yhat

separate y, by(country)

separate yhat, by(country)

twoway connected yhat1-yhat7
x1, msymbol(none
diamond_hollow triangle_hollow
square_hollow + circle_hollow
x) msize(medium) mcolor(black
black black black black black
black) || lfit y x1,
clwidth(thick) clcolor(black)
```



NOTE: In Stata 11 you do not need "xi:" when adding dummy variables

The least square dummy variable model (LSDV) provides a good way to understand fixed effects.

The effect of x1 is mediated by the differences across countries.

By adding the dummy for each country we are estimating the pure effect of x1 (by controlling for the unobserved heterogeneity).

Each dummy is absorbing the effects particular to each country.

```
regress y x1
estimates store ols
xi: regress y x1 i.country
estimates store ols_dum
estimates table ols ols_dum, star stats(N)
```

```
. estimates table ols ols_dum, star stats(N)
```

Variable	ols	ols_dum
x1	4.950e+08	2.476e+09*
_Icountry_2		-1.938e+09
_Icountry_3		-2.603e+09
_Icountry_4		2.282e+09
_Icountry_5		-1.483e+09
_Icountry_6		1.130e+09
_Icountry_7		-1.865e+09
_cons	1.524e+09*	8.805e+08
N	70	70

legend: * p<0.05; ** p<0.01; *** p<0.001

Fixed effects: n entity-specific intercepts using **xtreg**

Comparing the fixed effects using dummies with `xtreg` we get the same results.

```
. xtreg  y  x1, fe
```

Using `xtreg`

```
Fixed-effects (within) regression               Number of obs      =        70
Group variable: country                         Number of groups   =         7

R-sq:  within  = 0.0747                         Obs per group: min =        10
       between = 0.0763                                        avg =      10.0
       overall = 0.0059                                        max =        10

                                                F(1,62)            =      5.00
corr(u_i, Xb)  = -0.5468                        Prob > F           =    0.0289
```

y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
x1	2.48e+09	1.11e+09	2.24	0.029	2.63e+08	4.69e+09
_cons	2.41e+08	7.91e+08	0.30	0.762	-1.34e+09	1.82e+09
sigma_u	1.818e+09					
sigma_e	2.796e+09					
rho	.29726926	(fraction of variance due to u_i)				

OLS regression

```
. xi: regress y x1 i.country
i.country        _Icountry_1-7       (naturally coded; _Icountry_1 omitted)

      Source |       SS       df       MS              Number of obs =      70
-------------+------------------------------           F(  7,    62) =    2.61
       Model |  1.4276e+20     7  2.0394e+19           Prob > F      =  0.0199
    Residual |  4.8454e+20    62  7.8151e+18           R-squared     =  0.2276
-------------+------------------------------           Adj R-squared =  0.1404
       Total |  6.2729e+20    69  9.0912e+18           Root MSE      =  2.8e+09
```

y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
x1	2.48e+09	1.11e+09	2.24	0.029	2.63e+08	4.69e+09
_Icountry_2	-1.94e+09	1.26e+09	-1.53	0.130	-4.47e+09	5.89e+08
_Icountry_3	-2.60e+09	1.60e+09	-1.63	0.108	-5.79e+09	5.87e+08
_Icountry_4	2.28e+09	1.26e+09	1.81	0.075	-2.39e+08	4.80e+09
_Icountry_5	-1.48e+09	1.27e+09	-1.17	0.247	-4.02e+09	1.05e+09
_Icountry_6	1.13e+09	1.29e+09	0.88	0.384	-1.45e+09	3.71e+09
_Icountry_7	-1.87e+09	1.50e+09	-1.25	0.218	-4.86e+09	1.13e+09
_cons	8.81e+08	9.62e+08	0.92	0.363	-1.04e+09	2.80e+09

Fixed effects: *n* entity-specific intercepts (using `xtreg`)

$$Y_{it} = \beta_1 X_{it} + \ldots + \beta_k X_{kt} + \alpha_i + e_{it} \quad \text{[see eq.1]}$$

NOTE: Add the option 'robust' to control for heteroskedasticity

Outcome variable | Predictor variable(s)

Fixed effects option

Total number of cases (rows)

Total number of groups (entities)

```
. xtreg  y  x1, fe

Fixed-effects (within) regression               Number of obs      =        70
Group variable: country                         Number of groups   =         7

R-sq:  within  = 0.0747                         Obs per group: min =        10
       between = 0.0763                                        avg =      10.0
       overall = 0.0059                                        max =        10

                                                F(1,62)            =      5.00
corr(u_i, Xb)  = -0.5468                        Prob > F           =    0.0289
```

The errors u_i are correlated with the regressors in the fixed effects model

If this number is < 0.05 then your model is ok. This is a test (F) to see whether all the coefficients in the model are different than zero.

Coefficients of the regressors. Indicate how much *Y* changes when *X* increases by one unit.

y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
x1	2.48e+09	1.11e+09	2.24	0.029	2.63e+08	4.69e+09
_cons	2.41e+08	7.91e+08	0.30	0.762	-1.34e+09	1.82e+09
sigma_u	1.818e+09					
sigma_e	2.796e+09					
rho	.29726926	(fraction of variance due to u_i)				

Two-tail p-values test the hypothesis that each coefficient is different from 0. To reject this, the p-value has to be lower than 0.05 (95%, you could choose also an alpha of 0.10), if this is the case then you can say that the variable has a significant influence on your dependent variable (y)

29.7% of the variance is due to differences across panels.

'rho' is known as the intraclass correlation

$$rho = \frac{(sigma\_u)^2}{(sigma\_u)^2 + (sigma\_e)^2}$$

t-values test the hypothesis that each coefficient is different from 0. To reject this, the t-value has to be higher than 1.96 (for a 95% confidence). If this is the case then you can say that the variable has a significant influence on your dependent variable (y). The higher the t-value the higher the relevance of the variable.

sigma_u = sd of residuals within groups u_i

sigma_e = sd of residuals (overall error term) e_i

For more info see Hamilton, Lawrence, *Statistics with STATA*.

19