ΚΕΦΑΛΑΙΟ 2

Απλό γραμμικό υπόδειγμα και η μέθοδος ελαχίστων τετραγώνων

2.1 Αιτιότητα και πλασματική συσχέτιση

Ένα χαρακτηριστικό λάθος στο οποίο μπορεί να υποπέσουμε στην εμπειρική οικονομετρική πρακτική, είναι η αναγνώριση σχέσεων αιτιότητας μεταξύ μεταβλητών (κυρίως χρονοσειρών) που εμφανίζουν μεν γραμμική συσχέτιση όμως ουσιαστικά είναι ανεξάρτητες. Σε αυτή την περίπτωση μπορεί να ανιχνεύσουμε στατιστικά σημαντική (δηλαδή μη-μηδενική) συσχέτιση μεταξύ των μεταβλητών η οποία όμως είναι **«πλασματική»**[1] και μπορεί να οδηγήσει σε αναξιόπιστα αποτελέσματα αλληλεξάρτησης ή αιτιότητας μεταξύ των μεταβλητών ενδιαφέροντος.

Είτε, η ύπαρξη συσχέτισης μεταξύ δύο χρονοσειρών (ή και μεταξύ δύο μεταβλητών διαστρωματικών δεδομένων) μπορεί να θεωρηθεί ως επιβεβαίωση μίας ήδη υπάρχουσας οικονομικής θεωρίας, η οποία αναπτύσσει και εξηγεί τη συγκεκριμένη θεωρητική σχέση αιτίας - αιτιατού δύο μεταβλητών **είτε**, αντίστροφα και υπό-προϋποθέσεις, μπορεί να δώσει το έναυσμα για την εξέλιξη μιας θεωρίας. **Σε καμμία όμως περίπτωση** δεν δύναται η ύπαρξη και μόνο γραμμικής συσχέτισης να αποτελέσει θεωρία ή αλλιώς επιβεβαίωση σχέσεων αιτίας-αιτιατού, αφού συχνά η εμφανιζόμενη δειγματική γραμμική συσχέτιση είναι «πλασματική».

Τα παραδείγματα πλασματικών συσχετίσεων είναι άφθονα ειδικότερα αν οι υπο-εξέταση μεταβλητές εμφανίζουν **«τάσεις»** (δεδομένα χρονοσειρών) και γε-

[1] Η έννοια της ψευδούς συσχέτισης εισήχθη για πρώτη φορά από τον Karl Pearson το 1897.

νικότερα αν οι υπο-εξέταση μεταβλητές εμπίπτουν στην κατηγορία των **μη στάσιμων χρονοσειρών** (θέμα στο οποίο θα επανέλθουμε).

Χαρακτηριστικά αναφέρουμε:

(α) την περίπτωση δύο ή περισσότερων χρονοσειρών που υπόκεινται σε **«στοχαστικές τάσεις»** και όπου δύο ή περισσότερες μεταβλητές - ανεξάρτητες μεταξύ τους - εμφανίζουν διαχρονικές σχέσεις φαινομενικά προφανείς και στατιστικά ισχυρές (ενώ όπως προείπαμε, η μεταξύ τους σχέση είναι ανύπαρκτη)

(β) την περίπτωση χρονοσειρών που εμφανίζουν **άλλου είδους τάσεις** και πιο συγκεκριμένα **προσδιοριστικές τάσεις** (π.χ. ο χρόνος εισέρχεται ως ερμηνευτική μεταβλητή)

(γ) την περίπτωση διαστρωματικών μεταβλητών ή χρονοσειρών που εμφανίζουν πλασματική συσχέτιση λόγω εξάρτησής τους από μία κοινή μεταβλητή δηλαδή λόγω παράλειψης από την ανάλυση μίας μεταβλητής που επιδρά και στις δύο (ή περισσότερες) εξεταζόμενες μεταβλητές

(δ) την περίπτωση όπου οι παρατηρούμενες συσχετίσεις είναι στατιστικά κατασκευάσματα (artifacts) παραγόμενα από διάφορους τύπους μετασχηματισμών των δεδομένων, π.χ. του τύπου X/Z με Y/Z, ή $X \times Z$ με $Y \times Z$, ή X με Y/X, και $X + Y$ με Y.

Παράδειγμα. Περίπτωση (γ)
Έστω ότι θέλουμε να διερευνήσουμε την επίδραση του ύψους ενός ατόμου X_i στην ταχύτητα τρεξίματός του σε συγκεκριμένο άθλημα (έστω αγώνας δρόμου ταχύτητας 100μ.). Ο υπολογισμός του δειγματικού συντελεστή συσχέτισης των μεταβλητών θα δείξει στατιστικά σημαντική (διάφορη του μηδενός) συσχέτιση (και μάλιστα έντονα θετική) συνεπώς μπορεί να διαπιστώσουμε ότι: *«ψηλότερα άτομα τρέχουν πιο γρήγορα, κατά μέσο όρο»*. Εάν μία αληθής αιτιώδης σχέση κρύβεται πίσω από τη συγκεκριμένη συσχέτιση (π.χ. κάποιος μπορεί να συμπεράνει ότι όσο μακρύτερα τα πόδια ενός ατόμου τόσο ψηλότερα άρα και γρηγορότερα) τότε η «πολιτική» που ενδείκνυται είναι να ενθαρρύνουμε ψηλότερα άτομα να ασχοληθούν με το στίβο (συγκεκριμένα με τα 100 μέτρα).
Ένας τέτοιος βιαστικός σχεδιασμός «πολιτικής» δεν θα ήταν συνετός χωρίς να μελετήσουμε βαθύτερα τη συγκεκριμένη σχέση (αιτιότητα ή πλα-

σματική συσχέτιση;). Υπάρχουν μεταβλητές που μπορούν να επηρεάσουν τόσο το ύψος όσο και την ταχύτητα τρεξίματος; Μία μάλλον προφανής μεταβλητή είναι το **φύλο** του ατόμου. Κατά μέσο όρο, οι άνδρες είναι ψηλότεροι από τις γυναίκες και έχουν επίσης άλλες φυσιολογικές ιδιότητες που τους κάνουν να τρέχουν πιο γρήγορα. «Έλεγχος» (control) ως προς τη μεταβλητή φύλο σημαίνει ότι συγκρίνουμε τους άνδρες με τους άνδρες και τις γυναίκες με τις γυναίκες. Αυτό που πρέπει να εξετάσουμε είναι αν οι ψηλές γυναίκες τρέχουν πιο γρήγορα από τις υπόλοιπες και αν οι ψηλοί άνδρες τρέχουν πιο γρήγορα από τους υπόλοιπους άνδρες. Όντως μία τέτοια ανάλυση θα έδειχνε ότι δεν υπάρχει σχέση μεταξύ ύψους και χρόνου. Η **συσχέτιση** ήταν **πλασματική**.

Το φαινόμενο της πλασματικής συσχέτισης μελετήθηκε συστηματικά τουλάχιστον από τη δεκαετία του 1920 και τα πειράματα του Yule. Για τις ανάγκες της διάλεξης μπορούμε να μιμηθούμε τα πειράματα ακολουθώντας την τακτική της δημιουργίας δύο ανεξάρτητων μεταξύ τους μεταβλητών οι οποίες παρουσιάζουν **στοχαστική τάση**.

ΣΗΜΕΙΩΣΗ: Η κυρίαρχη οικονομετρική ερμηνεία που δίδεται στις στοχαστικές τάσεις είναι χρονοσειρές που δημιουργούνται ως συσσωρευτικά αθροίσματα (cumulative sums) υποκείμενων χρονοσειρών που πληρούν τουλάχιστον τρεις υποθέσεις: (α) έχουν μέση τιμή και (β) διακύμανση, οι οποίες δεν μεταβάλλονται χρονικά δηλαδή δεν εξαρτώνται από τη διάσταση του χρόνου και (γ) έχουν συνδιακύμανση η οποία επίσης δεν εξαρτάται από τη διάσταση του χρόνου και φθίνει προς το μηδέν καθώς η χρονική απόσταση των παρατηρήσεων της υποκείμενης χρονοσειράς αυξάνει. Θα δούμε σε επόμενο κεφάλαιο ότι τέτοιες χρονοσειρές καλούνται ασθενώς στάσιμες ή στάσιμες και τα συσσωρευτικά αθροίσματά τους δημιουργούν στοχαστικές τάσεις ή χρονοσειρές που περιέχουν στοχαστικές τάσεις

Για το σκοπό αυτό θα χρησιμοποιήσουμε το gretl και θα ακολουθήσουμε το παράδειγμα του αρχείου «`ch2-correlation.inp`» όπου δημιουργούμε ένα τυχαίο δείγμα $T = 50$ παρατηρήσεων για δύο ανεξάρτητες τυχαίες μεταβλητές[2]

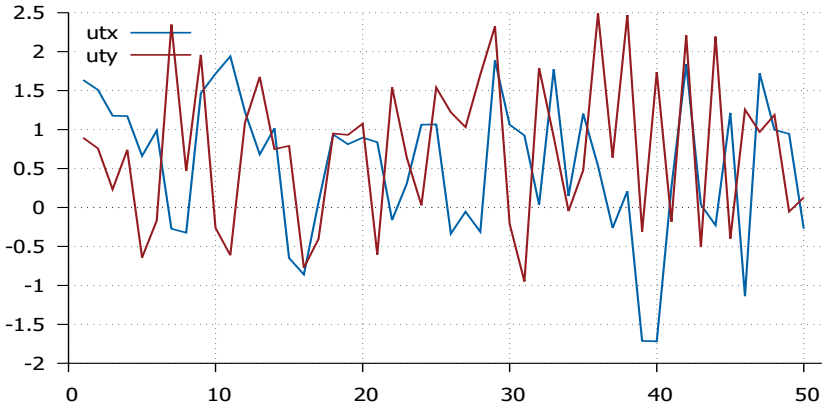
Ονομάζουμε τις μεταβλητές $u_{x,t}$ και $u_{y,t}$ και υποθέτουμε ότι κατανέμονται

[2]Μπορείτε να χρησιμοποιήσετε το Excel, ακολουθώντας το παράδειγμα του αρχείου «kefalaio2data1.xlsx».

κανονικά με ίδια αναμενόμενη τιμή 0.5 και διακύμανση 1

$$u_{x,t} \sim N\,(0.5, 1)$$
$$u_{y,t} \sim N\,(0.5, 1)$$

Το «παρατηρούμενο» δείγμα των τυχαίων και ανεξάρτητων μεταβλητών $u_{x,t}$ και $u_{y,t}$ εμφανίζεται στο παρακάτω γράφημα (2.1): Στη συνέχεια, δημιουργούμε τις



Γράφημα 2.1: Στο γράφημα εμφανίζονται οι μεταβλητές $u_{x,t}$ (μαύρη γραμμή) και $u_{y,t}$ (μπλε γραμμή) για $t = 1, 2, 3, ..., 50$. Αποτελούν πραγματοποιήσεις δύο τυχαίων μεταβλητών οι οποίες κατανέμονται κανονικά με αναμενόμενη τιμή 0.5 και διακύμανση 1 και είναι ανεξάρτητες άρα και γραμμικώς ασυσχέτιστες τόσο μεταξύ τους όσο και με τον «εαυτό τους».

μεταβλητές Y_t και X_t ως τα σωρευτικά αθροίσματα (cumulative sums) των $u_{y,t}$ και $u_{x,t}$ αντίστοιχα, δηλαδή

$$Y_t = \sum_{j=1}^{t} u_{y,j} =$$
$$= u_{y,1} + u_{y,2} + ... + u_{y,t}\ ,\ t = 1, 2, ..., 50$$

και

$$X_t = \sum_{j=1}^{t} u_{x,j} =$$
$$= u_{x,1} + u_{x,2} + ... + u_{x,t}\ ,\ t = 1, 2, ..., 50$$

ή, αναλυτικά για παράδειγμα - το σωρευτικό άθροισμα $Y_t = \sum_{j=1}^{t} u_{y,j}$ δίνεται από:

$$\begin{aligned} Y_1 &= u_{y,1} \quad , \quad t = 1 \\ Y_2 &= u_{y,1} + u_{y,2} \quad , \quad t = 2 \\ Y_3 &= u_{y,1} + u_{y,2} + u_{y,3} \quad , \quad t = 3 \\ &\vdots \\ Y_T &= u_{y,1} + u_{y,2} + ... + u_{y,T} \quad , \quad t = T = 50 \end{aligned}$$

Σχεδιάζουμε τις μεταβλητές Y_t και X_t. **Παρατηρήστε** ότι μοιάζουν αρκετά με συναθροιστικές μακροοικονομικές μεταβλητές (aggregate macroeconomic variables) όπως η κατανάλωση ή το πραγματικό Α.Ε.Π. Στο συγκεκριμένο παράδειγμα λαμβάνουμε το παρακάτω γράφημα των Y_t και X_t (γράφημα 2.2):

Στη συνέχεια υπολογίζουμε το δειγματικό συντελεστή (γραμμικής) συσχέτισης[3] των Y_t, X_t μέσω του εκτιμητή

$$\hat{\rho} = \frac{\sum_{t=1}^{T} \left(Y_t - \bar{Y}\right)\left(X_t - \bar{X}\right)}{\sqrt{\sum_{t=1}^{T} \left(Y_t - \bar{Y}\right)^2 \sum_{t=1}^{T} \left(X_t - \bar{X}\right)^2}}$$

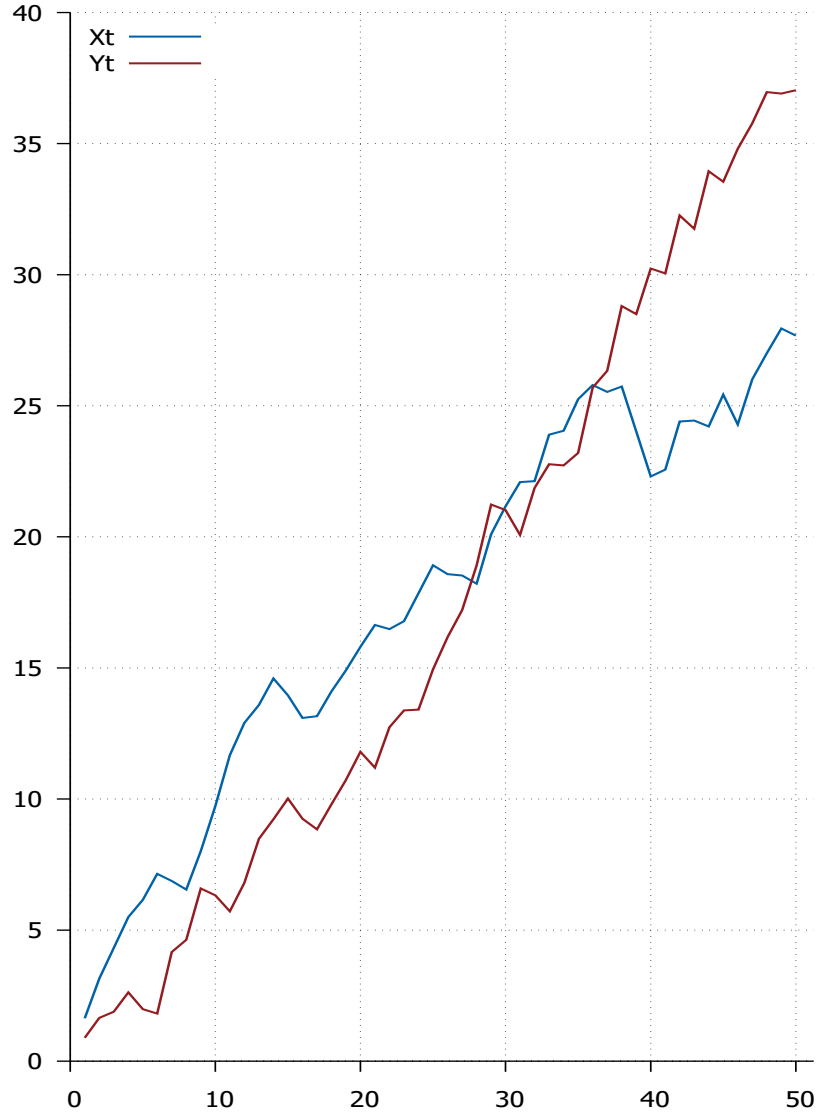
Θα παρατηρήσετε ότι ο συντελεστής συσχέτισης $\hat{\rho} = 0.9404$ είναι πολύ υψηλός υπονοώντας (ισχυρή) θετική συσχέτιση των Y_t, X_t. Ένας **δίπλευρος έλεγχος** της στατιστικής σημαντικότητας του συντελεστή $\hat{\rho}$, δηλαδή ο έλεγχος των υποθέσεων

$$\begin{aligned} H_0 &: \rho_{Y,X} = 0 \\ H_1 &: \rho_{Y,X} \neq 0 \end{aligned}$$

διεξάγεται με χρήση της t-student στατιστικής

$$t_{\hat{\rho}} = \hat{\rho} \cdot \sqrt{\frac{T-2}{1-\hat{\rho}^2}} \sim t_{T-2}$$

[3] Ο δειγματικός συντελεστής συσχέτισης (sample correlation coefficient) δύο μεταβλητών, έστω Y_t, X_t συμβολίζεται με $\hat{\rho}_{Y,X}$ ή $r_{Y,X}$ ή απλώς με $\hat{\rho}$ και r όταν δεν είναι απαραίτητο να τονίσουμε τις ονομασίες των μεταβλητών.



Γράφημα 2.2: Στο γράφημα εμφανίζονται οι X_t (μαύρη γραμμή) και Y_t (μπλε γραμμή) για $t = 1, 2, 3, ..., 50$. Δημιουργήθηκαν ως συσσωρευτικά αθροίσματα των πραγματοποιήσεων δύο τυχαίων μεταβλητών οι οποίες κατανέμονται κανονικά με μέσο 0.1 και διακύμανση 1 και είναι ανεξάρτητες άρα και γραμμικώς ασυσχέτιστες. Οι μεταβλητές Y_t, X_t είναι επίσης ανεξάρτητες.

Συγκρίνουμε δηλαδή την τιμή $t_{\hat{\rho}}$ με την κρίσιμη τιμή της t-student κατανομής με $T-2$ βαθμούς ελευθερίας t_{T-2} και επίπεδο σημαντικότητας[4], έστω 5%. Υπολογίζουμε την $t_{\hat{\rho}}$ και λαμβάνουμε την τιμή

$$t_{\hat{\rho}} = 0.9404 \cdot \sqrt{\frac{50-2}{1-0.9404^2}} = 19.1683$$

Προβείτε στον έλεγχο και θα διαπιστώσετε ότι ο εκτιμητής $\hat{\rho}$ είναι στατιστικά σημαντικός. Δηλαδή απορρίπτουμε τη μηδενική υπόθεση $H_0 : \rho_{Y,X} = 0$ ότι ο συντελεστής συσχέτισης του πληθυσμού είναι ίσος με το μηδέν. Παρόλα αυτά, γνωρίζουμε ότι οι δύο μεταβλητές Y_t, X_t είναι ανεξάρτητες αφού τις δημιουργήσαμε ανεξάρτητα (πρόκειται για σωρευτικά αθροίσματα ανεκξάτητων τυχαίων μεταβλητών) άρα είναι και γραμμικώς ασυσχέτιστες.

Έχουμε λοιπόν ένα χαρακτηριστικό παράδειγμα **πλασματικής συσχέτισης** (περίπτωση χρονοσειρών). Η «εξήγηση» βρίσκεται στο ότι και οι δύο μεταβλητές «κινούνται» ανοδικά άρα «ξεγελούν» τον εκτιμητή $\hat{\rho}$. Σε καμμία περίπτωση δεν μπορούμε να πούμε ότι η X_t αποτελεί **αιτιώδη παράγοντα** για την Y_t (ή το αντίστροφο).

Φανταστείτε όμως την ίδια κατάσταση σε ένα οικονομετρικό υπόδειγμα όπου δεν έχουμε ακολουθήσει την οικονομική θεωρία στην επιλογή των Y_t και X_t. Είναι πολύ εύκολο κανείς να ανακαλύψει καινούργιες (δυστυχώς πλασματικές) σχέσεις.

2.2 Απλό υπόδειγμα παλινδρόμησης και οι κλασσικές υποθέσεις

Το απλούστερο οικονομετρικό υπόδειγμα, το απλό (διμεταβλητό) γραμμικό υπόδειγμα ή υπόδειγμα παλινδρόμησης, λαμβάνει τη μορφή

$$Y_i = \alpha + \beta X_i + u_i$$

όπου u_i είναι ο διαταρακτικός όρος. Μία ισοδύναμη και πιο «διαισθητική» εξαγωγή του υποδείγματος πραγματώνεται μέσα από την ερμηνεία της από κοινού συνάρτησης πυκνότητας πιθανότητας $f(y,x)$ των Y_i, X_i και της παραγοντοποίη-

[4] Δείτε κεφάλαιο 3 για ερμηνεία του επιπέδου σημαντικότητας.

σης

$$f(y,x) = f(y\,|x)\, f(x)$$

σύμφωνα με την οποία η από κοινού συνάρτηση πυκνότητας πιθανότητας δίνεται ως το γινόμενο της δεσμευμένης σ.π.π $f(y|x)$ επί την οριακή σ.π.π $f(x)$ της μεταβλητής που δεσμεύσαμε[5] (της X_i).

Γενικεύουμε σημαντικά λοιπόν το υπόδειγμα με την παραδοχή ότι και η ερμηνευτική ή ανεξάρτητη μεταβλητή X_i είναι στοχαστική άρα περιγράφεται από κάποια κατανομή. Όμως, δεν είναι η σ.π.π $f(x)$ που μας ενδιαφέρει αλλά η $f(y|x)$. Σε ένα διαφορετικό επίπεδο ανάλυσης (π.χ., αν η οικονομετρία αποτελούσε πειραματική επιστήμη) θα μπορούσαμε να υποθέσουμε ότι η X_i είναι μη στοχαστική, δηλαδή σταθερή σε επαναλαμβανόμενα δείγματα (επαναλαμβανόμενες δειγματοληψίες θα παρήγαγαν τις ίδιες τιμές X_i). Όπως θα γίνει σαφές παρακάτω, σε αυτή την περίπτωση οι «περιοριστικές» υποθέσεις σχετικά με την τυχαιότητα του διαταρακτικού όρου θα αναφέρονταν στις μη δεσμευμένες ροπές του. Αντιθέτως, όταν η X_i είναι στοχαστική, οι ίδιες υποθέσεις θα αναφέρονται σε δεσμευμένες ή υπό συνθήκη ροπές του διαταρακτικού όρου θεωρώντας δεδομένη την πληροφόρησή μας σχετικά με τη X_i.

Θέλουμε λοιπόν να μελετήσουμε τη δεσμευμένη κατανομή $f(y|\cdot)$ με δεδομένη (ή κάνοντας χρήση) όλη τη διαθέσιμη πληροφόρηση από τη θεωρία σχετικά με μεταβλητές που επηρεάζουν τη δεσμευμένη κατανομή ή αλλιώς μελετούμε πως μεταβάλλεται η δεσμευμένη κατανομή της Y_i όταν μεταβάλλονται οι διαθέσιμες ερμηνευτικές μεταβλητές (στο απλουστευτικό παράδειγμά μας, ερμηνευτική μεταβλητή είναι μόνο η X_i).

Ένα πρώτο (και βασικό) μέτρο περίληψης ή περιγραφής μιας κατανομής είναι ο υπό συνθήκη ή δεσμευμένος μέσος[6] $E(Y\,|\mathbb{X})$ τον οποίο θεωρούμε συνάρτηση τουλάχιστον των τυχαίων μεταβλητών $X_1, ..., X_n$, δηλαδή το σύνολο πληροφόρησης $\mathbb{X}$ δίνεται τουλάχιστον ως

$$\mathbb{X} = \{X_1, ..., X_n\}$$

Μάλιστα - απλοποιώντας σημαντικά την ανάλυσή μας - υποθέτουμε ότι η υπό συνθήκη αναμενόμενη τιμή $E(Y_i|\mathbb{X})$ είναι γραμμική συνάρτηση της ερμηνευτικής μεταβλητής X_i με τις παραμέτρους του σταθερού όρου α και της κλίσης β να

[5] Στο παρόν πλαίσιο ανάλυσης, η συνάρτηση πυκνότητας πιθανότητας της ερμηνευτικής μεταβλητής ονομάζεται οριακή.

[6] Αποτελεί μέτρο της κεντρικής ροπής της δεσμευμένης κατανομής.

είναι σταθερές για κάθε i. Δηλαδή, υποθέτουμε ότι

$$E\left(Y_i \left|\mathbb{X}\right.\right) = \alpha + \beta X_i \tag{2.1}$$

με

$$\alpha \ , \ \beta \ \text{ αμετάβλητα } \ \forall i \tag{2.2}$$

Σύμφωνα με την υπόθεση (2.1), ο υπό συνθήκη μέσος $E(Y_i|\mathbb{X})$ περιγράφεται ακριβώς από μία συνάρτηση ευθείας. Το **γραμμικό σφάλμα παλινδρόμησης** (linear regression error) ή **διαταρακτικός όρος** ορίζεται ως η διαφορά της Y_i από τον υπό συνθήκη μέσο της

$$Y_i - E\left(Y_i \left|\mathbb{X}\right.\right) = u_i \tag{2.3}$$

και με συνδυασμό των (2.1), (2.2) και (2.3) έχουμε

$$\begin{aligned} Y_i &= E\left(Y_i \left|\mathbb{X}\right.\right) + u_i \\ &= \alpha + \beta X_i + u_i \end{aligned} \tag{2.4}$$

Παρατηρήστε ότι το **γραμμικό σφάλμα παλινδρόμησης** - εξ'ορισμού - θα ικανοποιεί την παρακάτω συνθήκη **ισχυρής εξωγένειας** της ερμηνευτικής μεταβλητής

$$E\left(u_i \left|\mathbb{X}\right.\right) = 0 \tag{2.5}$$

αφού

$$\begin{aligned} E\left(u_i \left|\mathbb{X}\right.\right) &= E\left(\left[Y_i - E\left(Y_i \left|\mathbb{X}\right.\right)\right] \left|\mathbb{X}\right.\right) \\ &= E\left(Y_i \left|\mathbb{X}\right.\right) - E\left(E\left(Y_i \left|\mathbb{X}\right.\right) \left|\mathbb{X}\right.\right) \\ &= E\left(Y_i \left|\mathbb{X}\right.\right) - E\left(Y_i \left|\mathbb{X}\right.\right) \\ &= 0 \end{aligned}$$

κάτι που συνεπάγεται και την ισότητα

$$E\left(u_i\right) = 0 \tag{2.6}$$

αφού από τον **νόμο των επαναλαμβανόμενων προσδοκιών**[7] **(ν.ε.π)**

$$E(u_i) = E(E(u_i|\mathbb{X})) = E(0) = 0$$

Ο συνδυασμός των (2.1), (2.2) και (2.3) ή απλώς η υπόθεση (2.5) της ισχυρής εξωγένειας, υπονοούν ότι οποιαδήποτε συνεχής και μετρήσιμη συνάρτηση των X_i δεν συσχετίζεται με τους διαταρακτικούς όρους

$$\begin{aligned} E(g(X_i)u_i) &= E(E(g(X_i)u_i|\mathbb{X})) \\ &= E(g(X_i)E(u_i|\mathbb{X})) \\ &= E(g(X_i) \times 0) \\ &= 0 \end{aligned}$$

άρα ισχύει και η

$$E(X_j u_i) = 0 \quad , \quad \forall i, j$$

που συνεπάγεται[8] την

$$E(X_i u_i) = 0$$

Σημείωση: Η υπόθεση $E(u_i|\mathbb{X})$=0 είναι εξαιρετικά αυστηρή (οικονομετρικά). Να σημειώσουμε όμως ότι δεν υπονοεί ανεξαρτησία των u_i και X_i, παρά μόνο ανεξαρτησία του δεσμευμένου μέσου των διαταρακτικών όρων από την ερμηνευτική μεταβλητή. Για παράδειγμα, αν οι μεταβλητές X_i, e_i είναι ανεξάρτητες με μέσο μηδέν, τότε η μεταβλητή $u_i = X_i e_i$ έχει δεσμευμένο μέσο (ως προς την X_i) ίσο με το μηδέν

$$\begin{aligned} E(u_i|X_i) &= E(X_i e_i|X_i) \\ &= X_i E(e_i|X_i) \\ &= X_i \times 0 \\ &= 0 \end{aligned}$$

[7] Law of iterated expectations. Δείτε το παράρτημα στατιστικής για κατανόηση του βασικού αυτού νόμου.

[8] Για παράδειγμα, θέστε $g(X_i) = X_i$.

όμως αυτό δεν σημαίνει (εμφανώς) ότι η u_i είναι ανεξάρτητη της X_i.

Πλέον της υπόθεσης (2.5), διατυπώνονται συνήθως και οι επόμενες **τρεις υποθέσεις** οι οποίες χαρακτηρίζουν τη στοχαστική συμπεριφορά - δηλαδή χαρακτηρίζουν περαιτέρω την κατανομή - του διαταρακτικού όρου u_i στο απλό γραμμικό υπόδειγμα παλινδρόμησης.

Πρώτον, υποθέτουμε ότι η διακύμανση του διαταρακτικού όρου είναι σταθερή, δηλαδή δεν εξαρτάται κατά οποιονδήποτε τρόπο από τον δείκτη i,

$$Var\left(u_i\right) = \sigma^2 \ , \ \forall i \ \ \text{ομοσκεδαστικότητα} \tag{2.7}$$

ή εναλλακτικά όταν η ερμηνευτική μεταβλητή είναι στοχαστική υποθέτουμε ότι η υπό συνθήκη (ή αλλιώς δεσμευμένη) διακύμανση είναι σταθερή, δηλαδή

$$Var\left(u_i\,|\mathbb{X}\right) = \sigma^2 \ , \ \forall i$$
$$\text{υπό συνθήκη ομοσκεδαστικότητα} \tag{2.8}$$

Οι παραπάνω δύο υποθέσεις μας διευκολύνουν σημαντικά αφού, αν για οποιοδήποτε λόγο, ο διαταρακτικός όρος είναι **ετεροσκεδαστικός**

$$Var\left(u_i\right) = \sigma_i^2$$

τότε τίθεται το ερώτημα γιατί υπάρχει ετεροσκεδαστικότητα, τι μορφή προσλαμβάνει καθώς και τι αντίκτυπο έχει στις ιδιότητες των εκτιμητών και τη στατιστική επαγωγή. Στην εισαγωγική οικονομετρία, η **ετεροσκεδαστικότητα αναφέρεται πάντα στη δεσμευμένη διακύμανση των διαταρακτικών όρων** $Var(u_i|\mathbb{X}) = \sigma_i^2$, η οποία συνήθως είναι κάποια συνάρτηση των X_i, ενώ η μη δεσμευμένη διακύμανση είναι σταθερή ως προς την X_i. Για παράδειγμα, έστω ότι

$$Var(u_i|\mathbb{X}) = E(u_i^2|\mathbb{X}) = \sigma_i^2 = \sigma^2 X_i$$

με την $X_i > 0, \ \forall i$ να έχει μέση τιμή $E(X_i) = \mu_X, \ \forall i$. Δηλαδή οι διαταρακτικοί όροι είναι ετεροσκεδαστικοί με τη δεσμευμένη διακύμανσή τους να εξαρτάται από το επίπεδο της ερμηνευτικής μεταβλητής. Τότε από τον ν.ε.π έχουμε

$$Var\left(u_i\right) = E\left(u_i^2\right) = E\left(E\left(u_i^2|\mathbb{X}\right)\right)$$

$$= E\left(\sigma^2 X_i\right) = \sigma^2 E\left(X_i\right) = \sigma^2 \mu_X$$

Άρα η (μη δεσμευμένη) διακύμανση $Var\left(u_i\right)$ είναι σταθερή (ομοσκεδαστικότητα) για κάθε i και λαμβάνει την τιμή $\sigma^2 \mu_X$ παρότι έχουμε ετεροσκεδαστικότητα αφού η δεσμευμένη διακύμανση είναι συνάρτηση της ερμηνευτικής μεταβλητής.

Δεύτερον, υποθέτουμε απουσία γραμμικής συσχέτισης στους διαταρακτικούς όρους[9], δηλαδή υποθέτουμε μηδενική συνδιακύμανση

$$Cov\left(u_i, u_j\right) = 0 \ , \forall i \neq j \tag{2.9}$$

ή μηδενική υπό συνθήκη συνδιακύμανση

$$Cov\left(u_i, u_j | \mathbb{X}\right) = 0 \ , \forall i \neq j \tag{2.10}$$

Σε αντίθετη περίπτωση (συσχετιζόμενοι[10] διαταρακτικοί όροι) έχουμε ότι

$$Cov\left(u_i, u_j\right) \neq 0$$

για τουλάχιστον κάποια $i \neq j$ και «πρέπει» να διατυπώσουμε ένα δεύτερο σαφές υπόδειγμα το οποίο να εξηγεί την ύπαρξη και τη μορφή της συσχέτισης ή κατά μία λιγότερο τεχνική και περισσότερο οικονομική θεώρηση της οικονομετρικής πρακτικής θα πρέπει να εξειδικεύσουμε ξανά το υπόδειγμα.

Τέλος, η **τρίτη υπόθεση** αφορά την κατανομή των διαταρακτικών όρων (δεσμευμένη ή μή) και υποθέτουμε ότι οι διαταρακτικοί όροι του υποδείγματος κατανέμονται **από κοινού κανονικά**.

ΣΗΜΕΙΩΣΗ: Η υπόθεση της **από κοινού κανονικότητας** είναι σημαντική αφού υπονοεί και ανεξαρτησία των διαταρακτικών όρων όταν αυτοί είναι ασυσχέτιστοι. Το συγκεκριμένο συμπέρασμα δεν θα ίσχυε αν υποθέταμε ότι οι διαταρακτικοί όροι κατανέμονται καθένας κανονικά αλλά δεν διατυπώναμε την από κοινού κατανομή τους. Τυχαίες μεταβλητές που κατανέμονται οριακά (δηλαδή η κάθε μία ξεχωριστά) κανονικά μπορεί να είναι ασυσχέτιστες παρότι ταυτόχρονα δεν είναι και ανεξάρτητες.

[9] Στην περίπτωση υποδειγμάτων χρονοσειρών, η συγκεκριμένη υπόθεση είναι μάλλον αυστηρή.

[10] Για χρονοσειρές αναφερόμαστε σε «αυτοσυσχέτιση».

Όταν θεωρούμε τις ερμηνευτικές μεταβλητές **μη τυχαίες**, γράφουμε

$$u_i \sim N\left(0, \sigma^2\right) \tag{2.11}$$

ενώ όταν, ρεαλιστικά, θεωρούμε τις ερμηνευτικές μεταβλητές **τυχαίες** γράφουμε ότι κατανέμονται από κοινού και υπό συνθήκη κανονικά,

$$u_i \,|\mathbb{X} \sim N\left(0, \sigma^2\right) \tag{2.12}$$

ΣΗΜΕΙΩΣΗ: Η υπόθεση της κανονικότητας διευκολύνει σημαντικά τις στατιστικές τεχνικές που αφορούν ιδιότητες των εκτιμητών σε πεπερασμένα ή μικρά δείγματα καθώς και στην εξαγωγή των κατανομών των ελέγχων υποθέσεων, για παράδειγμα στον t έλεγχο στατιστικής σημαντικότητας ή σε γενικότερους ελέγχους γραμμικών υποθέσεων (στατιστικές ελέγχου που κατανέμονται σύμφωνα με την F κατανομή). Σταδιακά, η υπόθεση (της κανονικότητας) μπορεί να (και θα) χαλαρώσει, με τίμημα όμως την καταφυγή μας στην **ασυμπτωτική θεωρία** και το κεντρικό οριακό θεώρημα για διατύπωση στατιστικών επαγωγής / συμπερασματολογίας. Παρότι υπάρχει πληθώρα φυσικών φαινομένων που ακολουθούν ακριβώς ή προσεγγιστικά την κανονική κατανομή, η υπόθεση της κανονικότητας στην οικονομετρία είναι, στην καλύτερη περίπτωση, αμφισβητήσιμη.

Άρα, συνοψίζοντας, στο απλούστερο δυνατό (γραμμικό) διμεταβλητό υπόδειγμα

$$Y_i = \alpha + \beta X_i + u_i$$

εμπεριέχονται οι παρακάτω υποθέσεις οι οποίες στο εξής θα ονομάζονται **«κλασσικές υποθέσεις»:**

Περίπτωση 1.

Η ερμηνευτική μεταβλητή του υποδείγματος, X_i, είναι σταθερή σε επαναλαμβανόμενα δείγματα

Υπόθεση 1. Ο μη δεσμευμένος μέσος της εξαρτημένης μεταβλητής $E(Y_i)$ είναι γραμμική συνάρτηση της ερμηνευτικής μεταβλητής

Υπόθεση 2. Οι παράμετροι α,β (συντελεστές υποδείγματος) είναι σταθερές δηλαδή αμετάβλητες ή αλλιώς δεν εξαρτώνται από τον υποδείκτη i ή t

Υπόθεση 3. $E(u_i) = 0$, μηδενική αναμενόμενη τιμή του διαταρακτικού όρου ή του σφάλματος παλινδρόμησης

Υπόθεση 4. $Var(u_i) = E(u_i^2) = \sigma^2$, $\forall i$, ομοσκεδαστικότητα

Υπόθεση 5. $Cov(u_i, u_j) = 0$, $\forall i \neq j$, απουσία συσχέτισης (ή αυτοσυσχέτισης όταν έχουμε δεδομένα χρονοσειρών) του διαταρακτικού όρου

Υπόθεση 6. από κοινού κανονικότητα των διαταρακτικών όρων

$$u_i \sim N(E(u_i), Var(u_i))$$

άρα

$$u_i \sim N(0, \sigma^2)$$

Οι τελευταίες τρεις υποθέσεις (υποθέσεις 4, 5, 6) **στην περίπτωση 1** συνοψίζονται στο συμβολισμό

$$u_i \sim N.i.d\left(0, \sigma^2\right)$$

δηλαδή οι διαταρακτικοί όροι κατανέμονται ως κανονικές και ανεξάρτητες τυχαίες μεταβλητές (normally and independently distributed) με μηδένική αναμενόμενη τιμή και διακύμανση σ^2.

Περίπτωση 2.
Η ερμηνευτική μεταβλητή του υποδείγματος, X_i, είναι στοχαστική (δηλαδή είναι τυχαία μεταβλητή)

Υπόθεση 1. Ο δεσμευμένος μέσος της εξαρτημένης μεταβλητής $E(Y_i|\mathbb{X})$ είναι γραμμική συνάρτηση της ερμηνευτικής μεταβλητής

Υπόθεση 2. Οι παράμετροι α,β (συντελεστές υποδείγματος) είναι σταθερές δηλαδή αμετάβλητες ή αλλιώς δεν εξαρτώνται από τον υποδείκτη i ή t

Υπόθεση 3. $E(u_i|\mathbb{X}) = 0$, μηδενική δεσμευμένη αναμενόμενη τιμή του διαταρακτικού όρου ή του σφάλματος παλινδρόμησης ως προς την X_i, $\forall i$.

Υπόθεση 4. $Var(u_i|\mathbb{X}) = E(u_i^2|\mathbb{X}) = \sigma^2$, $\forall i$, δεσμευμένη ομοσκεδαστικότητα ως προς την X_i, $\forall i$

Υπόθεση 5. $Cov(u_i, u_j|\mathbb{X}) = 0$, $\forall i \neq j$, απουσία δεσμευμένης συσχέτισης (ή δεσμευμένης αυτοσυσχέτισης όταν έχουμε δεδομένα χρονοσειρών) του διαταρακτικού όρου ως προς την X_i, $\forall i$

Υπόθεση 6. από κοινού δεσμευμένη κανονικότητα των διαταρακτικών όρων

$$u_i|\mathbb{X} \sim N\left(E\left(u_i|\mathbb{X}\right), Var\left(u_i|\mathbb{X}\right)\right)$$

άρα

$$u_i|\mathbb{X} \sim N(0, \sigma^2)$$

2.3 Μέθοδος (εκτίμησης παραμέτρων) ελαχίστων τετραγώνων

Έστω το απλό γραμμικό υπόδειγμα,

$$Y_i = \alpha + \beta X_i + u_i,\ i = 1, ..., n \tag{2.13}$$

όπου Y_i η **εξαρτημένη μεταβλητή**, X_i η **ερμηνευτική μεταβλητή** και u_i ο διαταρακτικός όρος.

Βασικός μας στόχος είναι να εκτιμήσουμε τις παραμέτρους α, β αφού σε αυτές συνοψίζονται οι επιδράσεις της ερμηνευτικής μεταβλητής X_i επί της Y_i. Οι τιμές των παραμέτρων α, β είναι άγνωστες. Με βάση λοιπόν ένα περιορισμένο δείγμα τιμών που έχουμε στη διάθεσή μας για τις μεταβλητές Y_i, X_i θα προσπαθήσουμε να εκτιμήσουμε τις άγνωστες παραμέτρους.

Οι **εκτιμητές** των α, β **συμβολίζονται** με $\hat{\alpha}, \hat{\beta}$ και αποτελούν μαθηματικούς τύπους (συναρτήσεις) που βασίζονται στα δεδομένα του δείγματος. Θεωρητικά υπάρχει ένας τεράστιος αριθμός δειγμάτων που θα μπορούσαμε να λάβουμε υπόψη. Επιλέγοντας ένα νέο δείγμα (ίδιου μεγέθους) θα άλλαζε και η τιμή του εκτιμητή. Ουσιαστικά λοιπόν, αντιμετωπίζουμε τους εκτιμητές ως τυχαίες μεταβλητές και τις κατανομές στις οποίες υπόκεινται τις ονομάζουμε **κατανομές δειγματοληψίας**.

Θεωρητικά υπάρχει ένας πολύ μεγάλος αριθμός (άπειρος;) εκτιμητών των συντελεστών α, β σε μία γραμμική σχέση όπως η (2.13). Θα αναφέρουμε ως παράδειγμα τρεις εκτιμητές. Σε εισαγωγικό επίπεδο στην οικονομετρία ασχολούμαστε σε βάθος μόνο με τον τρίτο εκτιμητή ο οποίος ονομάζεται **εκτιμητής**

ελαχίστων τετραγώνων[11]**, (συντομογραφία: ΕΤ)**. Οι λόγοι θα γίνουν κατανοητοί στη συνέχεια των διαλέξεων. Στο σημείο αυτό απλώς δηλώνουμε ότι ο τρίτος εκτιμητής είναι ο «καλύτερος» με βάση συγκεκριμένες **στατιστικές ιδιότητες** που έχει και οι οποίες απορρέουν από συγκεκριμένες **υποθέσεις του υποδείγματος** (τις «κλασσικές υποθέσεις» που αναφέρθησαν παραπάνω). Επιπλέον, ο εκτιμητής ΕΤ είναι μαθηματικά «ελκυστικός» διευκολύνοντας την - πάντα - επίπονη άλγεβρα.

«Εκτιμητής» 1. «Οπτική» εκτίμηση του γραμμικού υποδείγματος, δηλαδή της ευθείας $\hat{\alpha} + \hat{\beta}X_i$ άρα των $\hat{\alpha}, \hat{\beta}$. Προφανώς δεν πρόκειται για επιστημονική μέθοδο αλλά για αυθαίρετη άσκηση «ζωγραφικής» γραμμών οι οποίες «φαίνεται» να ταιριάζουν ή προσαρμόζονται στα δεδομένα. Φυσικά, ανάλογα με τον «καλλιτέχνη» μεταβάλλεται και η ευθεία που προσαρμόζεται στα δεδομένα. Επίσης, καταλαβαίνουμε ότι επειδή δεν είναι επιστημονική μέθοδος δεν υπάρχουν τρόποι να υποστηρίξει κανείς την εκτίμησή του αντικειμενικά. Για παράδειγμα, στο παρακάτω γράφημα διασποράς (2.3) για ένα δείγμα $n = 80$ παρατηρήσεων των Y_i, X_i έχουμε σχεδιάσει τρεις γραμμές οι οποίες φαίνεται ότι «προσαρμόζονται» στα δεδομένα. Ποια όμως είναι η «καλύτερη»; Με ποια κριτήρια θα επιλεγεί κάποια από τις τρεις γραμμές ως η «καλύτερη»;

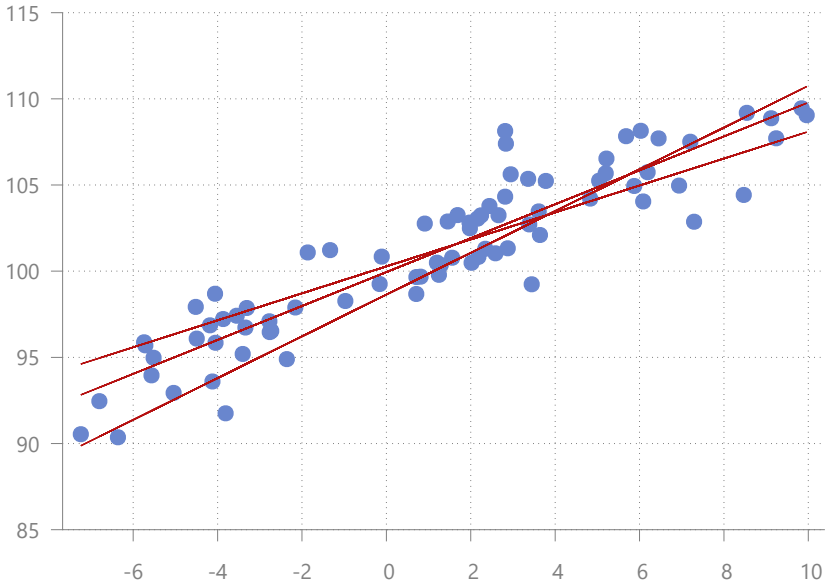
Εκτιμητής 2. Εφόσον θέλουμε να προσαρμόσουμε μία ευθεία γραμμή στα δεδομένα, οποιαδήποτε δύο σημεία είναι ικανά να μας δώσουν εκτιμήσεις για τα α, β. Για παράδειγμα, ας επιλέξουμε το σημείο που αντιστοιχεί στην υψηλότερη τιμή της Y_i μεταβλητής - έστω σημείο (Y_{max}, X_{ymax}) - και το σημείο που αντιστοιχεί στη χαμηλότερη τιμή της Y_i μεταβλητής - έστω σημείο (Y_{min}, X_{ymin}) - και ας τα ενώσουμε με μία ευθεία. Η κλίση της ευθείας και ο σταθερός όρος δίνονται από τους τύπους

$$\hat{\beta} = \frac{Y_{max} - Y_{min}}{X_{ymax} - X_{ymin}}$$

$$\hat{\alpha} = Y_{max} - \hat{\beta}X_{ymax} \text{ ή } \hat{\alpha} = Y_{min} - \hat{\beta}X_{ymin}$$

Η μέθοδος δεν είναι αξιόπιστη αφού χρησιμοποιεί μόνο δύο από τα διαθέσιμα σημεία, δηλαδή δεν αξιοποιεί όλη τη διαθέσιμη πληροφόρηση. Επιπλέον

[11] least squares estimator, LS ή ordinary least squares estimator, OLS

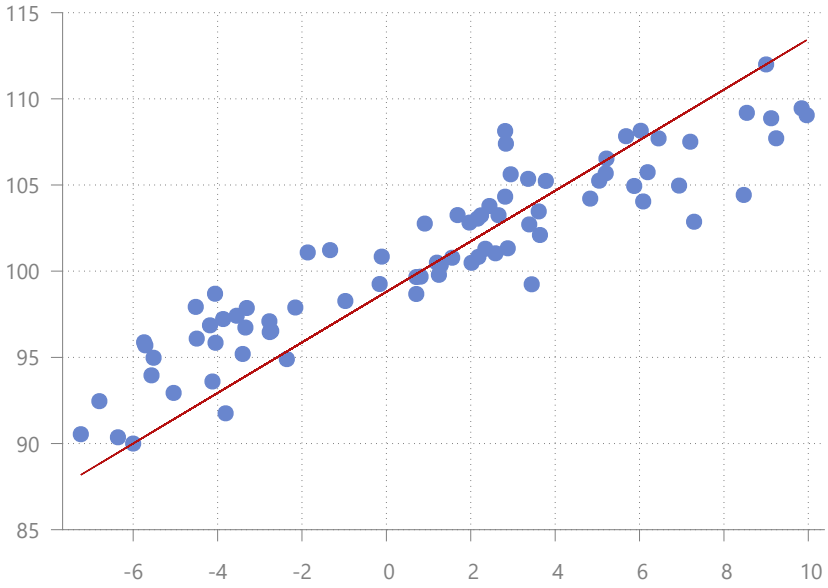


Γράφημα 2.3: Εκτιμητής 1 ή μέθοδος εκτίμησης 1: Στο γράφημα εμφανίζονται τρεις τυχαία σχεδιασμένες γραμμές που αντιστοιχούν σε τρεις διαφορετικές «εκτιμήσεις» των α, β. Οι «εκτιμήσεις» στη συγκεκριμένη περίπτωση δεν είναι επιστημονικές, αλλά βασίζονται στην οπτική ικανοποίησή μας ότι κάποια γραμμή «ταιριάζει» ή «προσεγγίζει» καλύτερα στα δεδομένα.

είναι εξαιρετικά «ευαίσθητη» σε ακραίες τιμές των μεταβλητών Y_i, X_i. Για παράδειγμα δείτε το παρακάτω γράφημα διασποράς (2.4) μαζί με την εκτιμημένη ευθεία $\hat{\alpha} + \hat{\beta} X_i$ για ένα δείγμα $n = 80$ παρατηρήσεων. Είναι εμφανές ότι η χρήση δύο μόνο σημείων (και μάλιστα ακραίων) μπορεί να δώσει εκτιμήσεις οι οποίες δεν προσαρμόζονται ή προσαρμόζονται «φτωχά» στα δεδομένα.

Εκτιμητής 3. Μέθοδος εκτίμησης ελαχίστων τετραγώνων (ΕΤ). Η μέθοδος θα δώσει εκτιμητές των παραμέτρων α, β που θα συμβολίζουμε με $\hat{\alpha}_{ET}, \hat{\beta}_{ET}$. Η διαδικασία εκτίμησης ονομάζεται «ελάχιστα τετράγωνα» αφού βασίζεται στην ελαχιστοποίηση του αθροίσματος των τετραγώνων των **καταλοίπων** (residuals) $\hat{u}_i$,

$$\sum_{i=1}^{n} \hat{u}_i^2 = \sum_{i=1}^{n} \left(Y_i - \hat{\alpha} - \hat{\beta} X_i \right)^2$$



Γράφημα 2.4: Εκτιμητής 2: Στο γράφημα εμφανίζεται η ευθεία που σχηματίζεται από τον εκτιμητή 2 για τις α, β. Ο συγκεκριμένος εκτιμητής υιοθετεί μόνο δύο ζεύγη του δείγματος, τα (Y_{max}, X_{ymax}) και (Y_{min}, X_{ymin}).

Η γραμμή προσαρμογής του **εκτιμητή ελαχίστων τετραγώνων** εμφανίζεται στο γράφημα (2.5) (γράφημα διασποράς των Y_i, X_i μαζί με την εκτιμημένη ευθεία $\hat{\alpha} + \hat{\beta} X_i$).

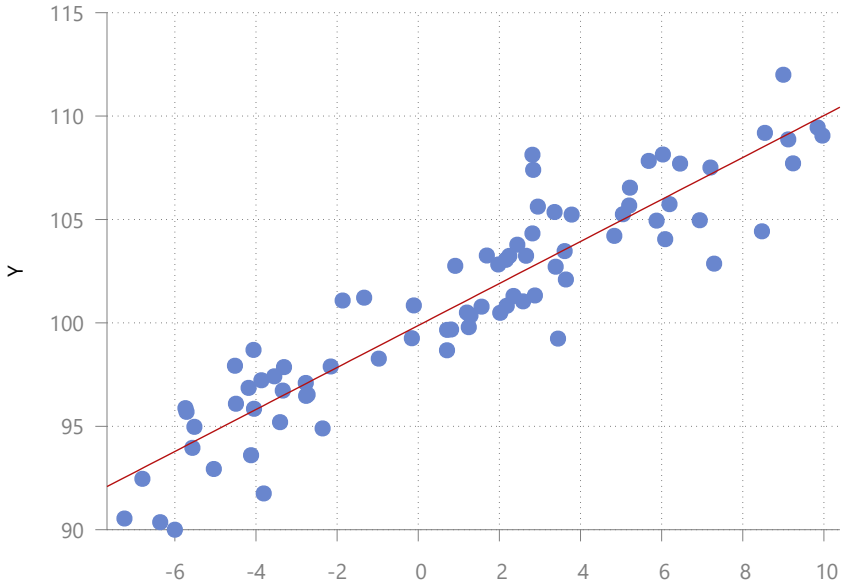
Σημείωση 1: Αν γνωρίζουμε ή έχουμε υπολογίσει τους δύο εκτιμητές $\hat{\alpha}, \hat{\beta}$ τότε οι όροι

$$\hat{u}_i = Y_i - \hat{\alpha} - \hat{\beta} X_i \, , \, \forall i$$

μπορούν να υπολογιστούν από τις ισότητες

$$\begin{aligned} \hat{u}_1 &= Y_1 - \hat{\alpha} - \hat{\beta} X_1 \\ \hat{u}_2 &= Y_2 - \hat{\alpha} - \hat{\beta} X_2 \\ &\vdots \\ \hat{u}_n &= Y_n - \hat{\alpha} - \hat{\beta} X_n \end{aligned}$$

και ονομάζονται **κατάλοιπα** (όχι διαταρακτικοί όροι). Στην ουσία, τα κατάλοιπα $\hat{u}_i$, $i = 1, ..., n$ αποτελούν εκτιμήσεις των διαταρακτικών όρων.



Γράφημα 2.5: Εκτιμητής 3: Στο γράφημα εμφανίζεται η ευθεία που σχηματίζεται από τον εκτιμητή 3 (**ελαχίστων τετραγώνων**) για τις α, β. Ο συγκεκριμένος εκτιμητής υιοθετεί όλα τα ζεύγη του δείγματος (αξιοποιεί **όλη** τη διαθέσιμη πληροφόρηση).

Σημείωση 2: Υπάρχουν βέβαια και άλλοι «παρόμοιοι» εκτιμητές όπως αυτός που βασίζεται στην ελαχιστοποίηση του αθροίσματος των απόλυτων αποκλίσεων $\left|Y_i - \tilde{\alpha} - \tilde{\beta}X_i\right|$ της εξαρτημένης μεταβλητής Y_i από τη γραμμή $\tilde{\alpha} + \tilde{\beta}X_i$, (Least Absolute Deviations ή εκτιμητής LAD)

$$S_1 = \sum_{i=1}^{n} |\hat{u}_i| = \sum_{i=1}^{n} \left|Y_i - \tilde{\alpha} - \tilde{\beta}X_i\right|$$

ενώ γενικότερα θα μπορούσαμε να ελαχιστοποιούμε

$$S_k = \sum_{i=1}^{n} |\hat{u}_i|^k = \sum_{i=1}^{n} \left|Y_i - \tilde{\alpha} - \tilde{\beta}X_i\right|^k$$

για κάποιο **ακέραιο μη-μηδενικό** k. Όμως, οι συναρτήσεις $S_k = \sum_{i=1}^{n} |\hat{u}_i|^k$ όταν $k \neq 2$ δεν προσφέρονται μαθηματικά για τον υπολογισμό των εκτιμητών. Ειδικότερα, όταν το k είναι μονός αριθμός, η απόλυτη τιμή παρουσιάζει τεχνι-

κά ζητήματα παραγωγισιμότητας ενώ για $k \geq 3$ δίνεται ολοένα και περισσότερη βαρύτητα (από τη μέθοδο) σε ακραίες τιμές/παρατηρήσεις των μεταβλητών του δείγματος. Δηλαδή ένα ακραίο ζεύγος τιμών των Y_i, X_i θα επηρεάσει άνισα τη συνολική εκτίμηση των υποκείμενων παραμέτρων.

Επίσης, δεν έχει νόημα να ελαχιστοποιήσουμε τη συνάρτηση του αθροίσματος των καταλοίπων $S_0 = \sum_{i=1}^{n} \hat{u}_i$ αφού σε αυτή την περίπτωση η συνάρτηση S_0 ελαχιστοποιείται (λαμβάνει τιμή μηδέν, $S_0 = \sum_{i=1}^{n} \hat{u}_i = 0$) για $\hat{\beta} = 0$ και $\hat{\alpha} = \bar{Y}$, δηλαδή παράγει μία ευθεία με μηδενική κλίση, το επίπεδο της οποίας (η τεταγμένη) αντιστοιχεί στη δειγματική μέση τιμή της εξαρτημένης μεταβλητής.

Κλείνοντας την **σημείωση 2**, να τονίσουμε ότι ο εκτιμητής LAD είναι λιγότερο ευαίσθητος σε μεγάλες ακραίες τιμές του δείγματος από ότι ο εκτιμητής ελαχίστων τετραγώνων, όμως στα περισσότερα σύνολα οικονομικών δεδομένων, οι σοβαρά ακραίες τιμές είναι μάλλον σπάνιες.

Επιστρέφοντας στη μέθοδο ΕΤ, οι εκτιμητές ΕΤ λύνουν το πρόβλημα ελαχιστοποίησης

$$\min_{\hat{\alpha},\hat{\beta}} S\left(\hat{\alpha},\hat{\beta}\right) \text{ όπου } S\left(\hat{\alpha},\hat{\beta}\right) = \sum_{i=1}^{n} \hat{u}_i^2 = \sum_{i=1}^{n} \left(Y_i - \hat{\alpha} - \hat{\beta}X_i\right)^2$$

Ακολουθώντας τη μαθηματική θεωρία θα πρέπει οι εκτιμητές (αν όντως ελαχιστοποιούν το άθροισμα $S\left(\hat{\alpha},\hat{\beta}\right)$ να «λύνουν» το σύστημα των εξισώσεων[12]

$$\frac{\partial S}{\partial \hat{\alpha}} = S_{\hat{\alpha}} = -2\sum_{i=1}^{n} \left(Y_i - \hat{\alpha} - \hat{\beta}X_i\right) = 0 \tag{2.14}$$

$$\frac{\partial S}{\partial \hat{\beta}} = S_{\hat{\beta}} = -2\sum_{i=1}^{n} \left(Y_i - \hat{\alpha} - \hat{\beta}X_i\right)(X_i) = 0 \tag{2.15}$$

οι οποίες αποτελούν τις συνθήκες πρώτης τάξης (**απαραίτητες συνθήκες**) για ελάχιστο. Τα $\hat{\alpha}$ και $\hat{\beta}$ που ικανοποιούν τις παραπάνω δύο εξισώσεις μαζί με την τιμή της $\min S\left(\hat{\alpha},\hat{\beta}\right)$ αποτελούν το λεγόμενο στάσιμο σημείο. Επιπλέον, πρέπει να ικανοποιούνται οι συνθήκες δεύτερης τάξης (**ικανές συνθήκες**) για

[12] Στην οικονομετρική ορολογία ονομάζονται **κανονικές εξισώσεις**.

ελάχιστο, δηλαδή η εσσιανή μήτρα (Hessian matrix)

$$\hat{H} = \begin{bmatrix} S_{\hat{\alpha}\hat{\alpha}} & S_{\hat{\beta}\hat{\alpha}} \\ S_{\hat{\alpha}\hat{\beta}} & S_{\hat{\beta}\hat{\beta}} \end{bmatrix}$$

των δεύτερων μερικών παραγώγων **υπολογισμένη στο στάσιμο σημείο** θα πρέπει να είναι θετικά ορισμένη.

ΣΗΜΕΙΩΣΗ: Χρησιμοποιούμε τους συμβολισμούς

$$\frac{\partial S}{\partial \hat{\alpha}} = S_{\hat{\alpha}} \ , \ \frac{\partial S}{\partial \hat{\beta}} = S_{\hat{\beta}}$$

για τις πρώτες μερικές παραγώγους της συνάρτησης S ως προς $\hat{\alpha}$ και $\hat{\beta}$ αντίστοιχα και τους συμβολισμούς

$$\frac{\partial^2 S}{\partial \hat{\alpha}^2} = S_{\hat{\alpha}\hat{\alpha}} \ , \ \frac{\partial^2 S}{\partial \hat{\beta}^2} = S_{\hat{\beta}\hat{\beta}}$$

και

$$\frac{\partial^2 S}{\partial \hat{\alpha}\partial \hat{\beta}} = S_{\hat{\alpha}\hat{\beta}}$$

για τις δεύτερες μερικές παραγώγους. Είναι γνωστό ότι για συνεχείς συναρτήσεις ισχύει $S_{\hat{\alpha}\hat{\beta}} = S_{\hat{\beta}\hat{\alpha}}$.

Διαφορετικά (εσσιανή μη-θετικά ορισμένη), η λύση των κανονικών εξισώσεων μπορεί να περιγράφει κάποιο μέγιστο σημείο ή κάποιο σαγματικό σημείο, δηλαδή ένα σημείο που δεν είναι ούτε μέγιστο ούτε ελάχιστο. Η 2×2 εσσιανή μήτρα $\hat{H}$ είναι θετικά ορισμένη αν ισχύει

$$S_{\hat{\alpha}\hat{\alpha}} > 0 \text{ και } \left|\hat{H}\right| > 0$$

δηλαδή όταν η ορίζουσα της $\hat{H}$ είναι θετική. Η ορίζουσα δίνεται από

$$\left|\hat{H}\right| = S_{\hat{\alpha}\hat{\alpha}} S_{\hat{\beta}\hat{\beta}} - \left(S_{\hat{\alpha}\hat{\beta}}\right)^2$$

αφού για το είδος των συναρτήσεων που εξετάζουμε $S_{\hat{\alpha}\hat{\beta}} = S_{\hat{\beta}\hat{\alpha}}$. Άρα για ελάχι-

στο θα πρέπει

$$S_{\hat{\alpha}\hat{\alpha}}S_{\hat{\beta}\hat{\beta}} - \left(S_{\hat{\alpha}\hat{\beta}}\right)^2 > 0$$

Λύνοντας τις εξισώσεις των συνθηκών πρώτης τάξης (2.14) και (2.15) για να βρούμε τους εκτιμητές $\hat{\alpha}_{ET}, \hat{\beta}_{ET}$ έχουμε:

από την (2.14)

$$\begin{aligned} -2\sum_{i=1}^{n}\left(Y_i - \hat{\alpha} - \hat{\beta}X_i\right) = 0 \Leftrightarrow \\ \sum_{i=1}^{n}\left(Y_i - \hat{\alpha} - \hat{\beta}X_i\right) = 0 \Leftrightarrow \\ \sum_{i=1}^{n}Y_i - n\hat{\alpha} - \hat{\beta}\sum_{i=1}^{n}X_i = 0 \Leftrightarrow \end{aligned}$$

$$\hat{\alpha}_{ET} = \bar{Y} - \hat{\beta}\bar{X}$$

και **αντικαθιστώντας στη (2.15)**

$$\begin{aligned} \sum_{i=1}^{n}\left(Y_i - \bar{Y} + \hat{\beta}\bar{X} - \hat{\beta}X_i\right)(X_i) = 0 \Leftrightarrow \\ \sum_{i=1}^{n}Y_iX_i - \bar{Y}\sum_{i=1}^{n}X_i + \hat{\beta}\bar{X}\sum_{i=1}^{n}X_i - \hat{\beta}\sum_{i=1}^{n}X_i^2 = 0 \Leftrightarrow \\ \sum_{i=1}^{n}Y_iX_i - \bar{Y}n\bar{X} + \hat{\beta}n\bar{X}^2 - \hat{\beta}\sum_{i=1}^{n}X_i^2 = 0 \end{aligned}$$

Η τελευταία εξίσωση, αν λυθεί ως προς $\hat{\beta}$ δίνει τον εκτιμητή κλίσης

$$\hat{\beta}_{ET} = \frac{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)\left(X_i - \bar{X}\right)}{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}$$

Ο υποδείκτης ΕΤ δηλώνει ότι ο εκτιμητής εξάγεται μέσω της μεθόδου των ελαχίστων τετραγώνων. Ο παραπάνω μαθηματικός τύπος για τον εκτιμητή της

κλίσης $\hat{\beta}_{ET}$ μπορεί να δοθεί σε πολλές εναλλακτικές μορφές, π.χ.,

$$\hat{\beta}_{ET} = \frac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} x_i^2} \text{ ή } \hat{\beta}_{ET} = \hat{\rho}\frac{s_y}{s_x}$$

όπου μικρά γράμματα συμβολίζουν αποκλίσεις από τον εκάστοτε δειγματικό μέσο, π.χ., $y_i = Y_i - \bar{Y}$ και $x_i = X_i - \bar{X}$ ενώ

$$\hat{\rho} = \frac{\sum_{i=1}^{n} y_i x_i}{\sqrt{\sum_{i=1}^{n} y_i^2}\sqrt{\sum_{i=1}^{n} x_i^2}}$$

συμβολίζει το δειγματικό συντελεστή συσχέτισης των μεταβλητών X_i, Y_i και

$$s_y = \sqrt{\frac{\sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2}{n-1}} \quad , \quad s_x = ... \text{ (αντίστοιχα για το x)}$$

είναι ένας δειγματικός εκτιμητής της τυπικής απόκλισης. Διαφορετικοί τύποι χρησιμοποιούνται ανάλογα με την «ερώτηση» που προσπαθούμε να απαντήσουμε. Στην προπτυχιακή οικονομετρία θα χρησιμοποιήσουμε σχεδόν αποκλειστικά τον τύπο των αποκλίσεων από το μέσο, δηλαδή

$$\hat{\beta}_{ET} = \frac{\sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)\left(X_i - \bar{X}\right)}{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2}$$

ή

$$\hat{\beta}_{ET} = \frac{\sum_{i=1}^{n} \left(y_i\right)\left(x_i\right)}{\sum_{i=1}^{n} \left(x_i\right)^2}$$

Μετά τη σχετικά επίπονη άλγεβρα φθάνουμε λοιπόν στο στάσιμο σημείο $\hat{\alpha}_{ET}$, $\hat{\beta}_{ET}$, $S\left(\hat{\alpha}_{ET}, \hat{\beta}_{ET}\right)$ όπου οι τιμές $\hat{\alpha}_{ET}, \hat{\beta}_{ET}$ καθιστούν τις κανονικές εξισώσεις

(2.14) και (2.15) ίσες με το μηδέν, δηλαδή στο σημείο που μηδενίζονται οι πρώτες μερικές παράγωγοι.

Τώρα, πρέπει να ελέγξουμε αν ικανοποιούνται οι συνθήκες δεύτερης τάξεως ώστε να βεβαιώσουμε ότι η συνάρτηση $S\left(\hat{\alpha},\hat{\beta}\right)$ **ελαχιστοποιείται** στις τιμές $\hat{\alpha}_{ET}$ και $\hat{\beta}_{ET}$.

Υπολογίζοντας τις δεύτερες μερικές παραγώγους $S_{\hat{\alpha}\hat{\alpha}}, S_{\hat{\beta}\hat{\beta}}, S_{\hat{\alpha}\hat{\beta}}$ στο στάσιμο σημείο έχουμε ότι

$$S_{\hat{\alpha}\hat{\alpha}} = 2n > 0 \ , \ S_{\hat{\alpha}\hat{\beta}} = 2\sum_{i=1}^{n} X_i$$

$$S_{\hat{\beta}\hat{\beta}} = \sum_{i=1}^{n} X_i^2$$

οπότε η ορίζουσα

$$\begin{aligned}\left|\hat{H}\right| &= S_{\hat{\alpha}\hat{\alpha}}S_{\hat{\beta}\hat{\beta}} - \left(S_{\hat{\alpha}\hat{\beta}}\right)^2 \\ &= 4n\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 > 0\end{aligned}$$

είναι θετική.

Άρα οι συνθήκες δεύτερης τάξης για ελάχιστο ικανοποιούνται και οι εκτιμητές

$$\hat{\alpha}_{ET} = \bar{Y} - \hat{\beta}\bar{X}$$

$$\hat{\beta}_{ET} = \frac{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)\left(X_i - \bar{X}\right)}{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}$$

ελαχιστοποιούν το άθροισμα των τετραγώνων των καταλοίπων

$$S\left(\hat{\alpha},\hat{\beta}\right) = \sum_{i=1}^{n}\hat{u}_i^2$$

2.4 Συντελεστής προσδιορισμού R^2

Παρατηρήστε ότι **μετά την εκτίμηση** το υπόδειγμα γράφεται ως (οι εκτιμημένες ποσότητες φέρουν «καπελάκι»)

$$Y_i = \hat{\alpha}_{ET} + \hat{\beta}_{ET} X_i + \hat{u}_i,\ i = 1, ..., n$$

και μπορούμε να γράψουμε ταυτοτικά ότι

$$Y_i = \hat{Y}_i + \hat{u}_i \tag{2.16}$$

όπου

$$\hat{Y}_i = \hat{\alpha}_{ET} + \hat{\beta}_{ET} X_i$$

συμβολίζει τις **προσαρμοσμένες ή εκτιμημένες τιμές** της Y_i (fitted values).

Η σχέση (2.16) διαχωρίζει την εξαρτημένη μεταβλητή σε δύο συστατικά, το **προσαρμοσμένο ή ερμηνευμένο** $\hat{Y}_i$ και το **ανερμήνευτο** $\hat{u}_i$ (τα κατάλοιπα).

Από τις κανονικές εξισώσεις και συγκεκριμένα την κανονική εξίσωση που αντιστοιχεί στο σταθερό όρο, ισχύει ότι $\sum_{i=1}^{n} \hat{u}_i = 0$, άρα οι Y_i και $\hat{Y}_i$ έχουν τον ίδιο δειγματικό μέσο

$$Y_i = \hat{Y}_i + \hat{u}_i \Rightarrow$$

$$\frac{1}{n}\sum_{i=1}^{n} Y_i = \frac{1}{n}\sum_{i=1}^{n} \hat{Y}_i + \frac{1}{n}\sum_{i=1}^{n} \hat{u}_i \Rightarrow$$

$$\bar{Y} = \overline{\hat{Y}} + \frac{1}{n} \times 0 \Rightarrow$$

$$\bar{Y} = \overline{\hat{Y}}$$

Από την κανονική εξίσωση που αντιστοιχεί στο συντελεστή κλίσης (παράμετρος β) έχουμε αντίστοιχα ότι

$$\sum_{i=1}^{n} X_i \hat{u}_i = 0$$

Μελετώντας λοιπόν την εξίσωση (2.16) σε αποκλίσεις από τους μέσους (ή απλά αφαιρώντας $\bar{Y}$ και από τα δύο σκέλη) έχουμε

$$Y_i - \bar{Y} = \hat{Y}_i - \bar{Y} + \hat{u}_i$$

και το άθροισμα των τετραγωνικών αποκλίσεων της Y_i από το μέσο της $\bar{Y}$ διαχωρίζεται σε

$$\begin{aligned}
\sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2 &= \\
&= \sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y} + \hat{u}_i\right)^2 \\
&= \sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y}\right)^2 + 2\sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y}\right) \hat{u}_i + \sum_{i=1}^{n} \hat{u}_i^2 \\
&= \sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y}\right)^2 + 2\sum_{i=1}^{n} \left(\hat{\alpha}_{ET} + \hat{\beta}_{ET} X_i - \bar{Y}\right) \hat{u}_i + \sum_{i=1}^{n} \hat{u}_i^2 \\
&= \sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y}\right)^2 + 2\sum_{i=1}^{n} \left(-\hat{\beta}_{ET} \bar{X} + \hat{\beta}_{ET} X_i\right) \hat{u}_i + \sum_{i=1}^{n} \hat{u}_i^2 \\
&= \sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y}\right)^2 + 2\hat{\beta}_{ET} \sum_{i=1}^{n} \left(X_i - \bar{X}\right) \hat{u}_i + \sum_{i=1}^{n} \hat{u}_i^2 \\
&= \sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y}\right)^2 + \sum_{i=1}^{n} \hat{u}_i^2
\end{aligned}$$

Συχνά, η παραπάνω ισότητα τετραγώνων, δηλαδή η

$$\sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2 = \sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y}\right)^2 + \sum_{i=1}^{n} \hat{u}_i^2$$

γράφεται και ως

$$TSS = ESS + RSS$$

δηλαδή το συνολικό άθροισμα τετραγώνων της εξαρτημένης μεταβλητής (total

sum of squares, TSS) είναι ίσο με το άθροισμα των προσαρμοσμένων «(επεξηγημένων» ή ερμηνευμένων) τετραγώνων (explained sum of squares, ESS) και το άθροισμα των τετραγώνων των καταλοίπων (residual sum of squares, RSS).

Ο **εκτιμητής της διακύμανσης του διαταρακτικού όρου** δίνεται από την

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\hat{u}_i^2$$

Ο παραπάνω εκτιμητής μπορεί να συμβολίζεται και με $\hat{\sigma}_u^2$ ενώ η τετραγωνική του ρίζα

$$\hat{\sigma}_u = \sqrt{\hat{\sigma}^2} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}\hat{u}_i^2}$$

καλείται **τυπικό σφάλμα της παλινδρόμησης (standard error of the regression)**.

> Γενικότερα, οι **«τυπικές αποκλίσεις»** εκτιμητών και στατιστικών που προκύπτουν από την εκτίμηση με βάση ένα και μόνο δειγμα καλούνται **«τυπικό σφάλμα»** και όχι τυπική απόκλιση ακριβώς επειδή εξαρτώνται από το παρατηρήσιμο δείγμα και μεταβάλλεται η τιμή τους καθώς μεταβάλλεται το δείγμα (υπόκεινται σε κατανομές δειγματοληψίας).

Επειδή αθροίσματα τετραγωνικών αποκλίσεων από τους εκάστοτε δειγματικούς μέσους «εκφράζουν» δειγματική **μεταβλητότητα** και το σύνηθες είναι - μετά από κατάλληλη τυποποίηση - να αποτελούν εκτιμητές διακύμανσης, ο παραπάνω διαχωρισμός υπονοεί ότι η συνολική μεταβλητότητα της εξαρτημένης μεταβλητής δίνεται από το άθροισμα της μεταβλητότητας που εξηγείται από το υπόδειγμα ESS (τη γραμμή παλινδρόμησης) και από την ανερμήνευτη μεταβλητότητα RSS.

Διαιρώντας και τα δύο σκέλη της εξίσωσης με TSS έχουμε

$$1 = \frac{ESS}{TSS} + \frac{RSS}{TSS} \Rightarrow \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

Ο λόγος $\frac{ESS}{TSS}$ συμβολίζεται με R^2 και ονομάζεται **συντελεστής προσδιορισμού**. Εκπροσωπεί το ποσοστό της μεταβλητότητας της εξαρτημένης μεταβλητής που «εξηγείται» από την παλινδρόμηση και ισχύει ότι $0 < R^2 < 1$. Ο **συντελεστής προσδιορισμού** είναι ένα μέτρο προσαρμογής του υποδείγματος στα δεδομένα.

Παράδειγμα. Έστω ότι με βάση ένα δείγμα και συγκεκριμένο υπόδειγμα υπολογίσαμε ότι $R^2 = 0.846$. Άρα το 84.6% της μεταβλητότητας της εξαρτημένης μεταβλητής εξηγείται από την παλινδρόμηση.

Παράδειγμα. Με ένα τυχαίο δείγμα 1000 εργαζομένων εκτιμήθηκε η σχέση $w_i = \hat{\alpha} + \hat{\beta}x_i + \hat{u}_i$ όπου w_i το ωρομίσθιο και x_i η μεταβλητή: «έτη εκπαίδευσης». Βρέθηκε ότι $R^2 = 0.1026$. Άρα το 10.26% της μεταβλητότητας του ωρομισθίου εξηγείται από τα έτη εκπαίδευσης (τα οποία προσεγγίζουν τη μεταβλητή «εκπαίδευση») του εργαζόμενου.

Παρατήρηση 1

Καθώς το R^2 απομακρύνεται από το 0 προς την τιμή 1 θεωρούμε ότι η προσαρμογή είναι ολοένα και καλύτερη. Ωστόσο, δεν πρέπει να είμαστε εξαιρετικά αυστηροί στην κρίση μας ειδικότερα όταν αντιμετωπίζουμε μικρο-οικονομικές μεταβλητές, για παράδειγμα ωρομίσθια και έτη εκπαίδευσης ή ωρομίσθια και ηλικία. Η ετερογένεια των μικροοικονομικών μονάδων είναι τέτοια που καθιστά εξαιρετικά δύσκολο για μία και μόνο ανεξάρτητη μεταβλητή να ερμηνεύει «μεγάλο» μέρος της μεταβλητότητας της εξαρτημένης μεταβλητής. Για παράδειγμα, τι θα λέγατε για μία οικονομία η οποία έδωσε $R^2 = 0.1424$ στην εκτίμηση $w_i = \hat{\alpha} + \hat{\beta}x_i + \hat{u}_i$ όπου w_i το ωρομίσθιο και x_i η μεταβλητή: «ηλικία»; Μία πρώτη εσφαλμένη ερμηνεία θα έλεγε ότι το R^2 είναι κοντά στο μηδέν άρα το υπόδειγμα δεν έχει καλή προσαρμογή. Μία άλλη ερμηνεία θα έλεγε ότι απλά και μόνο η ηλικία «εξηγεί» το 14.24% της μεταβλητότητας του ωρομισθίου. Φανταστείτε στο ίδιο παράδειγμα μία οικονομία που έδωσε $R^2 = 0.99$. Στην περίπτωση αυτή τα ωρομίσθια καθορίζονται σχεδόν απόλυτα από την ηλικία του εργαζομένου. Η εκπαίδευση, η ικανότητα, η εμπειρία και τόσοι άλλοι παράγοντες δεν έχουν καμμία θέση στη διαμόρφωση της μεταβλητότητας του ωρομισθίου. Μία τέτοια εκτίμηση φαντάζει τόσο «περίεργη» που είτε υπονοεί «πρόβλημα με τα δεδομένα ή την εμπειρική εφαρμογή» είτε αντιμετωπίζουμε μία πλήρως κατευθυνόμενη και ελεγχόμενη (από ποιόν;) αγορά εργασίας.

Παρατήρηση 2

Επιπλέον, πρέπει να είμαστε ιδιαίτερα προσεκτικοί με την ερμηνεία του συντελεστή προσδιορισμού R^2 όταν έχουμε στη διάθεσή μας δεδομένα χρονοσειρών. Αποδεικνύεται ότι στο διμεταβλητό υπόδειγμα $R^2 = \hat{\rho}^2$, δηλαδή ο συντελεστής προσδιορισμού R^2 είναι ίσος με το τετράγωνο της δειγματικής συσχέτισης εξαρτημένης και ανεξάρτητης μεταβλητής $\hat{\rho}^2$. Συνεπώς, σε περιπτώσεις πλασματικής

συσχέτισης ο συντελεστής R^2 θα εμφανίζεται εξαιρετικά υψηλός οδηγώντας σε εσφαλμένα συμπεράσματα.

Παρατήρηση 3

Επίσης, πρέπει να έχουμε υπόψιν ότι όταν η εξίσωση που εκτιμήθηκε με τη μέθοδο των ελαχίστων τετραγώνων δεν περιλαμβάνει σταθερό όρο, τότε το R^2 μπορεί να λάβει αρνητικές τιμές και δεν χρησιμοποιείται ως μέτρο προσαρμογής της εκτιμημένης γραμμής στα δεδομένα.

Παρατήρηση 4

Τέλος, ο συντελεστής προσδιορισμού R^2 μπορεί να χρησιμοποιηθεί για σύγκριση της προσαρμοστικότητας διαφορετικών υποδειγμάτων στα δεδομένα ή αλλιώς να συγκριθεί πόσο καλά εξηγούν διαφορετικά υποδείγματα τη μεταβλητότητα της **ίδιας** εξαρτημένης μεταβλητής. Φυσικά, τέτοιου είδους σύγκριση έχει νόημα μόνο όταν η εξαρτημένη μεταβλητή δεν υφίσταται αλλαγές στα υπο-εξέταση και υπό σύγκριση υποδείγματα. Για παράδειγμα, **δεν συγκρίνουμε** το R^2 δύο υποδειγμάτων με εξαρτημένη μεταβλητή την Y_i και $\ln(Y_i)$ αντίστοιχα.

Αλγοριθμική σύγκριση R^2 και επιλογή της Y_i ή της $\ln(Y_i)$ ως εξαρτημένης μεταβλητής.
Έστω ότι θέλουμε να αποφασίσουμε αν θα υιοθετήσουμε το υπόδειγμα

$$Y_i = \alpha + \beta X_i + u_i$$

ή το υπόδειγμα

$$\ln(Y_i) = \alpha + \beta X_i + u_i \text{ (μόνο αν } Y_i > 0\,,\ \forall i)$$

με την εξαρτημένη μεταβλητή να εισέρχεται στο υπόδειγμα λογαριθμικά. Στο κεφάλαιο 4 θα δούμε αναλυτικά τους **οικονομικούς λόγους** που οδηγούν στην υιοθέτηση της λογαριθμικής εξαρτημένης μεταβλητής $\ln(Y_i)$ σε ένα υπόδειγμα. Παρ'όλ'αυτά, αν θέλουμε να βασιστούμε σε ένα στατιστικό μέτρο ή να επιβεβαιώσουμε και στατιστικά την υιοθέτηση της $\ln(Y_i)$ αντί της Y_i ως εξαρτημένης μεταβλητής μπορούμε να προβούμε στην παρακάτω βηματική μεθοδολογία:

Βήμα 1. Εκτιμούμε με ΕΤ το υπόδειγμα $Y_i = \alpha + \beta X_i + u_i$, υπολογίζουμε και αποθηκεύουμε την τιμή του R^2

Βήμα 2. Εκτιμούμε με ΕΤ το υπόδειγμα $\ln(Y_i) = \alpha + \beta X_i + u_i$, υπο-

λογίζουμε τις προσαρμοσμένες τιμές $\hat{z}_i = \widehat{\ln}(Y_i)$ και στη συνέχεια υπολογίζουμε τις τιμές $\hat{m}_i = e^{\hat{z}_i}$

Βήμα 3. Εκτιμούμε με ΕΤ το υπόδειγμα $Y_i = \gamma\hat{m}_i + \varepsilon_i$, όπου ε_i το σφάλμα παλινδρόμησης και υπολογίζουμε τις προσαρμοσμένες τιμές $\widehat{\widehat{Y}}_i = \hat{\gamma}_1\hat{m}_i$

Βήμα 4. Υπολογίζουμε το τετράγωνο $\hat{\rho}^2$ του δειγματικού συντελεστή συσχέτισης των $\widehat{\widehat{Y}}_i$ και Y_i ή εναλλακτικά εκτιμούμε με ΕΤ τη σχέση $Y_i = \gamma_1 + \gamma_2\widehat{\widehat{Y}}_i + \eta_i$ και αποθηκεύουμε τον συντελεστή προσδιορισμού, έστω $R^2_{\ln}$.

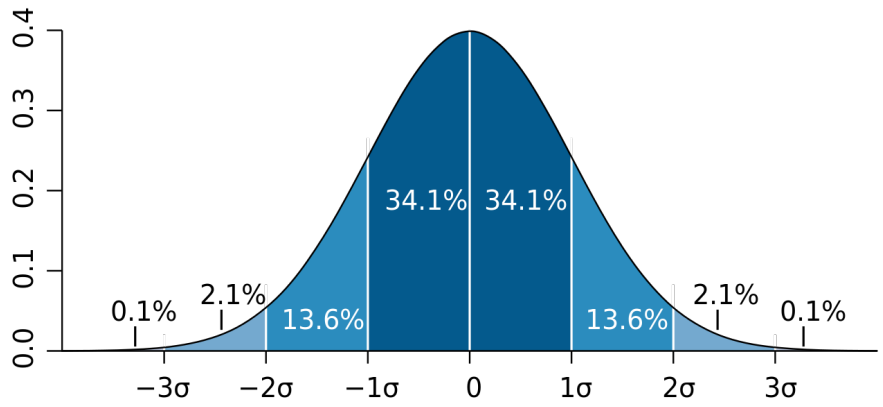
Βήμα 5. Συγκρίνουμε τους συντελεστές προσδιορισμού (όπου $\hat{\rho}^2 = R^2_{\ln}$) $R^2_{\ln}$ και R^2. Αν ο συντελεστής $R^2_{\ln}$ είναι μεγαλύτερος του R^2, δηλαδή αν $R^2_{\ln} > R^2$, τότε επιλέγουμε το υπόδειγμα με τη λογαριθμική εξαρτημένη μεταβλητή $\ln(Y_i)$ ως στατιστικά ανώτερο του υποδείγματος $Y_i = \alpha + \beta X_i + u_i$.

2.5 Τυπικό σφάλμα παλινδρόμησης ως μέτρο «προσαρμογής»

Ο συντελεστής προσδιορισμού R^2 μετρά το ποσοστό της μεταβλητότητας ή διακύμανσης της εξαρτημένης μεταβλητής που εξηγείται από την παλινδρόμηση ή αλλιώς από την ερμηνευτική μεταβλητή. Το **τυπικό σφάλμα της παλινδρόμησης** $\hat{\sigma}_u$ έχει τις **ίδιες μονάδες μέτρησης** με την εξαρτημένη μεταβλητή και μετρά τη μέση απόσταση των τιμών της εξαρτημένης, έστω Y_i, από την εκτιμημένη γραμμή παλνδρόμησης $\hat{Y}_i = \hat{\alpha} + \hat{\beta}X_i$.

Παρατηρήστε ότι ορισμένες τιμές της εξαρτημένης θα είναι κοντά ή πολύ κοντά στην εκτιμημένη γραμμή παλινδρόμησης, ενώ άλλες δεν είναι τόσο κοντά. **Κατά μέσο όρο**, οι παρατηρούμενες (πραγματικές) τιμές θα «πέφτουν» $\hat{\sigma}_u$ μονάδες από τη γραμμή παλινδρόμησης (πάνω ή κάτω). Η έκφραση «κατά μέσο όρο» προκύπτει από την τυποποιημένη κανονική κατανομή σύμφωνα με την οποία (δείτε την παρακάτω εικόνα[13]) το **68.27%** των τιμών της βρίσκονται στο διάστημα $(-\sigma, \sigma)$, το **95.45%** των τιμών της βρίσκονται στο διάστημα $(-2\sigma, 2\sigma)$ και το

[13]Πηγή: `https://en.wikipedia.org/wiki/Normal_distribution`



Γράφημα 2.6: Συνάρτηση πυκνότητας πιθανότητας κανονικής κατανομής.

99.73% των τιμών της βρίσκονται στο διάστημα $(-3\sigma, 3\sigma)$.

Το τυπικό σφάλμα της παλινδρόμησης είναι ιδιαίτερα χρήσιμο επειδή χρησιμοποιείται **για την αξιολόγηση της ακρίβειας των προβλέψεων** του υποδείγματος. Περίπου το 95% των παρατηρήσεων του δείγματος (της εξαρτημένης μεταβλητής) θα πέφτει μεταξύ συν ή πλήν δύο τυπικών σφαλμάτων της παλινδρόμησης $\pm 2\hat{\sigma}_u$, ουσιαστικά μια γρήγορη προσέγγιση ενός διαστήματος πρόβλεψης 95%.

Αν μας ενδιαφέρει λοιπόν η ακρίβεια πρόβλεψης τότε το τυπικό σφάλμα αποτελεί χρησιμότερη «μετρική» αφού υποδείγματα με ίδιες ή κοντικές τιμές του συντελεστή προσδιορισμού μπορεί να δίνουν αρκετά διαφορετικά τυπικά σφάλματα παλινδρόμησης και **θα επιλέγουμε το υπόδειγμα με το μικρότερο σφάλματα παλινδρόμησης**.

2.5.1 Εμπειρικό Παράδειγμα

Χρησιμοποιήστε τα δεδομένα του `kefalaio2data3.gdt` ώστε να εκτιμήσετε με τη μέθοδο ελαχίστων τετραγώνων τα παρακάτω δύο υποδείγματα

$$w_i = \alpha + \beta E_i + u_i \quad \textbf{(Υπόδειγμα 1)} \tag{2.17}$$

και

$$w_i = \alpha + \beta H_i + u_i \quad \textbf{(Υπόδειγμα 2)} \tag{2.18}$$

Τα αποτελέσματα σχετικά με τις εκτιμήσεις $\hat{\alpha}_{ET}$, $\hat{\beta}_{ET}$, R^2 και $\hat{\sigma}_u$ δίνονται αμέσως παρακάτω. Συγκεκριμένα, στο **υπόδειγμα 1** έχουμε $\hat{\alpha}_{ET} \approx 0.828$, $\hat{\beta}_{ET} \approx 0.480$, $R^2 \approx 0.1912$ και $\hat{\sigma}_u \approx 3.49$ (Τ.Σ. παλινδρόμησης).

Υπόδειγμα 1: OLS, χρήση των παρατηρήσεων 1–1000
Εξαρτημένη μεταβλητή: w

	Συντελεστής	Τυπ. Σφάλμα	t-λόγος	p-τιμή
const	0,828543	0,374794	2,211	0,0273
E	0,480149	0,0311679	15,41	0,0000

Μέσος εξαρτ. μτβλ	6,345931	Τ.Α. εξαρτ. μτβλ	3,883948
Άθρ. τετρ. καταλ	12174,83	Τ.Σ. παλινδρόμησης	3,492739
R^2	0,192113	Προσαρμ. R^2	0,191303
$F(1, 998)$	237,3212	P-τιμή(F)	3,36e–48
Λογ-πιθανοφάνεια	−2668,624	Akaike κριτήριο	5341,248
Schwarz κριτήριο	5351,063	Hannan–Quinn	5344,978

Αν παρακάμψουμε για λίγο το θέμα της στατιστικής σημαντικότητας των εκτιμήσεων των συντελεστών (δείτε αναλυτικά στο κεφάλαιο 3) παρατηρούμε στο **υπόδειγμα 1** ότι το 19.21% της μεταβλητότητας του πραγματικού ωρομισθίου w_i ερμηνεύεται από τη μεταβλητότητα της «εκπαίδευσης» E_i ενώ το αντίστοιχο τυπικό σφάλμα παλινδρόμησης είναι $\hat{\sigma}_u = 3.49$ **ευρώ** (έχει τις μονάδες μέτρησης της εξαρτημένης μεταβλητής, δηλαδή του ωρομισθίου). Άρα, περίπου (η εξαρτημένη δεν κατανέμεται κανονικά) το 68% των τιμών της w_i βρίσκονται στο διάστημα $\hat{w}_i \pm \hat{\sigma}_u$ και το 95% στο διάστημα $\hat{w}_i \pm 2 \cdot \hat{\sigma}_u = \hat{w}_i \pm 6.98$. Αν ζητούσαμε (αυθαίρετα ή μας το ζητούσαν κάποιοι που ασκούν πολιτική) το υπόδειγμα να δίνει προβλέψεις για το πραγματικό ωρομίσθιο με περιθώριο λάθους ±7 ευρώ τότε το υπόδειγμα 1 είναι ικανοποητικό αφού (για την ακρίβεια) $\pm 1.96 \cdot \hat{\sigma}_u = 6.94$ ευρώ. Στο τέλος του παραδείγματος θα δείτε και δύο διαγράμματα διασποράς με την εκτιμημένη γραμμή παλινδρόμησης και τα $\hat{w}_i \pm 2 \cdot \hat{\sigma}_u$ όρια έναντι της E_i.

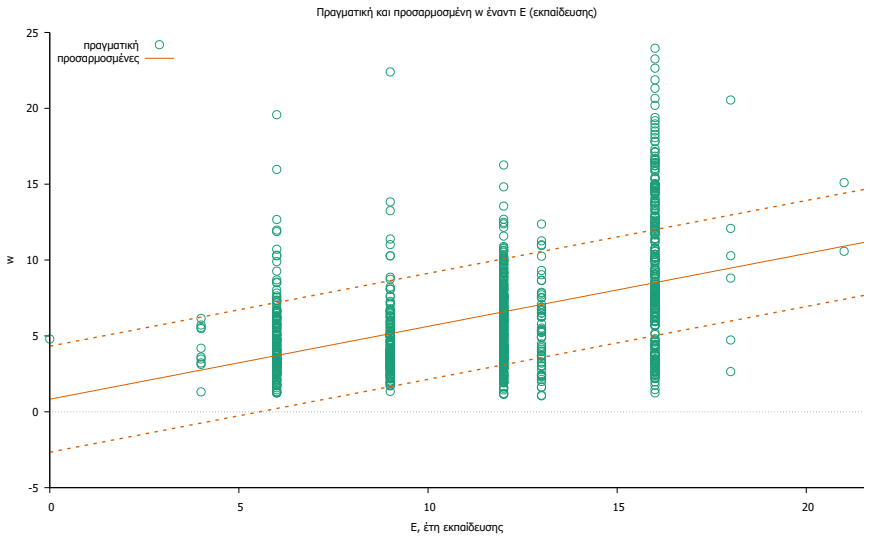
Υπόδειγμα 2: OLS, χρήση των παρατηρήσεων 1–1000

Εξαρτημένη μεταβλητή: w

	Συντελεστής	Τυπ. Σφάλμα	t-λόγος	p-τιμή
const	1,71865	0,455949	3,769	0,0002
H	0,115385	0,0109913	10,50	0,0000

Μέσος εξαρτ. μτβλ	6,345931	Τ.Α. εξαρτ. μτβλ	3,883948
Άθρ. τετρ. καταλ	13571,34	Τ.Σ. παλινδρόμησης	3,687620
R^2	0,099444	Προσαρμ. R^2	0,098542
$F(1, 998)$	110,2046	P-τιμή(F)	1,59e–24
Λογ-πιθανοφάνεια	−2722,919	Akaike κριτήριο	5449,838
Schwarz κριτήριο	5459,653	Hannan–Quinn	5453,568

Στο **υπόδειγμα 2** έχουμε $\hat{\alpha}_{ET} \approx 1.718$, $\hat{\beta}_{ET} \approx 0.115$, $R^2 \approx 0.1$ και $\hat{\sigma}_u \approx 3.68$ (Τ.Σ. παλινδρόμησης). Παρατηρούμε στο **υπόδειγμα 2** ότι το 9.94% της μεταβλητότητας του πραγματικού ωρομισθίου w_i ερμηνεύεται από τη μεταβλητότητα της «Ηλικίας» H_i ενώ το αντίστοιχο τυπικό σφάλμα παλινδρόμησης είναι $\hat{\sigma}_u = 3.68$ **ευρώ** (έχει τις μονάδες μέτρησης της εξαρτημένης μεταβλητής, δηλαδή του ωρομισθίου). Άρα, περίπου το 68% των τιμών της w_i βρίσκονται στο διάστημα $\hat{w}_i \pm \hat{\sigma}_u$ και το 95% στο διάστημα $\hat{w}_i \pm 2 \cdot \hat{\sigma}_u = \hat{w}_i \pm 7.37$. Αν ζητούσαμε (αυθαίρετα ή μας το ζητούσαν κάποιοι που ασκούν πολιτική) το υπόδειγμα να δίνει προβλέψεις για το πραγματικό ωρομίσθιο με περιθώριο λάθους ±7 ευρώ τότε το υπόδειγμα 2 **συγκριτικά** με το υπόδειγμα 1 **δεν** είναι ικανοποιητικό αφού (για την ακρίβεια) $\pm 1.96 \cdot \hat{\sigma}_u = 7,22$ ευρώ. Στο τέλος του παραδείγματος θα δείτε και δύο διαγράμματα διασποράς με την εκτιμημένη γραμμή παλινδρόμησης και τα $\hat{w}_i \pm 2 \cdot \hat{\sigma}_u$ όρια έναντι της E_i.



Γράφημα 2.7: Διάγραμμα διασποράς (**υπόδειγμα 1**) w_i μαζί με την εκτιμημένη γραμμή παλινδρόμησης $\hat{w}_i$ και τα $\hat{w}_i \pm 2 \cdot \hat{\sigma}_u$ όρια έναντι της E_i.

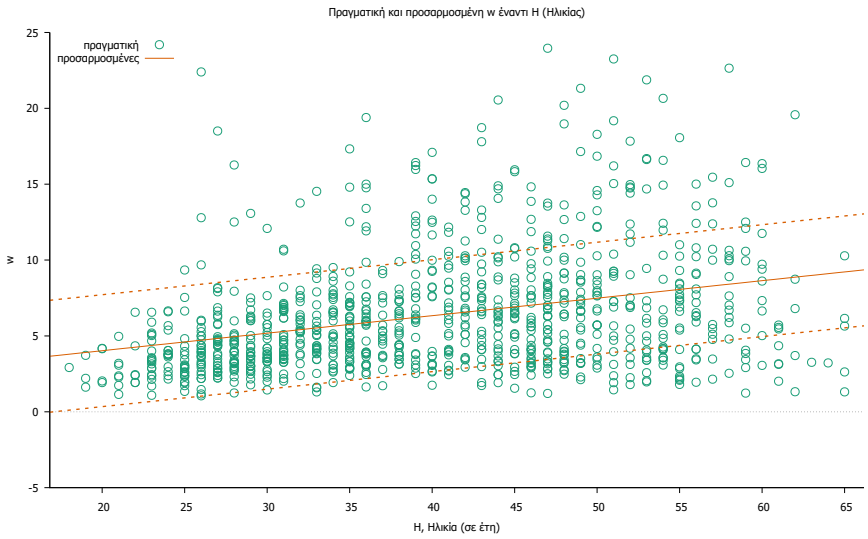
2.6 Ασκήσεις

1. Χρησιμοποιήστε το πρόγραμμα Excel ή gretl ώστε να αναπαράγετε τα αποτελέσματα του αρχείου kefalaio2data1.xlsx.

   Επιπλέον, χρησιμοποιήστε τυχαία δείγματα από μεταβλητές που κατανέμονται σύμφωνα με την ομοιόμορφη κατανομή $U\,(1, 10)$, την κατανομή χ^2 με 1 βαθμό ελευθερίας και την κανονική κατανομή με αναμενόμενη τιμή 0 και διακύμανση 1.

2. Η δεσμευμένη προσδοκία $E(Y_i|X_i)$ έχει την ιδιότητα να είναι ο καλύτερος «προλέγων» (συνάρτηση πρόβλεψης) της εξαρτημένης μεταβλητής Y_i στο γραμμικό υπόδειγμα $Y_i = f(X_i) + u_i$ όταν θεωρήσουμε δεδομένα τα X_i. Είναι ο καλύτερος προλέγων με την έννοια ότι επιτυγχάνει το μικρότερο **μέσο τετραγωνικό σφάλμα (Μ.Τ.Σ)** $E(u_i^2) = E[Y_i - f(X_i)]^2$ όπου $u_i = Y_i - f(X_i)$ θεωρούνται αποκλίσεις της Y_i από τη συνάρτηση πρόβλεψης $f(X_i)$.

   Δηλαδή η συνάρτηση $f(X_i)$ που ελαχιστοποιεί το Μ.Τ.Σ ή αλλιώς τη διακύμανση των u_i στο υπόδειγμα $Y_i = f(X_i) + u_i$ είναι η δεσμευμένη προσδοκία $E(Y_i|X_i)$. Αποδείξτε τον παραπάνω ισχυρισμό.



Γράφημα 2.8: Διάγραμμα διασποράς (**υπόδειγμα 2**) w_i μαζί με την εκτιμημένη γραμμή παλινδρόμησης $\hat{w}_i$ και τα $\hat{w}_i \pm 2 \cdot \hat{\sigma}_u$ όρια έναντι της H_i.

Απάντηση

Για ευκολία, συμβολίζουμε με ε_i την απόκλιση της Y_i από τη δεσμευμένη (ως προς X_i) προσδοκία $E(Y_i|X_i)$, δηλαδή $\varepsilon_i = Y_i - E(Y_i|X_i)$. Αναλύουμε το **μέσο τετραγωνικό σφάλμα (Μ.Τ.Σ)**

$$E\left(u_i^2\right) = E\left[Y_i - f\left(X_i\right)\right]^2$$

σε τρία επιμέρους στοιχεία, προσθαφαιρώντας τη δεσμευμένη προσδοκία $E(Y_i|X_i)$

$$\begin{aligned} E[Y_i - f\left(X_i\right)]^2 = \\ &= E\left[Y_i - f\left(X_i\right) + E\left(Y_i|X_i\right) - E\left(Y_i|X_i\right)\right]^2 \\ &= E\left[\varepsilon_i - f\left(X_i\right) + E\left(Y_i|X_i\right)\right]^2 \\ &= E\left[\varepsilon_i + \left[E\left(Y_i|X_i\right) - f\left(X_i\right)\right]\right]^2 \\ &= E\left(\varepsilon_i^2\right) + 2E\left[\varepsilon_i\left(E\left(Y_i|X_i\right) - f\left(X_i\right)\right)\right] \end{aligned}$$

$$+ E\left[\left(E\left(Y_i|X_i\right) - f\left(X_i\right)\right)^2\right]$$

Παρατηρήστε ότι με βάση τον νόμο των επαναλαμβανόμενων προσδοκιών (ν.ε.π) για τον κεντρικό όρο έχουμε το αποτέλεσμα

$$\begin{aligned} & E(\varepsilon_i \left[E\left(Y_i|X_i\right) - f\left(X_i\right)\right]) = \\ & = E\left[E\left(\varepsilon_i \left(E\left(Y_i|X_i\right) - f\left(X_i\right)\right)|X_i\right)\right] \\ & = \left[E\left(Y_i|X_i\right) - f\left(X_i\right)\right] E\left(\varepsilon_i|X_i\right) \\ & = \left[E\left(Y_i|X_i\right) - f\left(X_i\right)\right] \times 0 \\ & = 0 \end{aligned}$$

αφού

$$\begin{aligned} & E(\varepsilon_i|X_i) = \\ & = E\left[Y_i - E\left(Y_i|X_i\right)|X_i\right] \\ & = E\left(Y_i|X_i\right) - E\left(Y_i|X_i\right) \\ & = 0 \end{aligned}$$

Άρα το $\mathbf{M.T.\Sigma}$ απλοποιείται στην έκφραση

$$\begin{aligned} & E[Y_i - f\left(X_i\right)]^2 = \\ & = E\left(\varepsilon_i^2\right) + E\left[\left(E\left(Y_i|X_i\right) - f\left(X_i\right)\right)^2\right] \end{aligned}$$

Η «ποσότητα» δεξιά της ισότητας ελαχιστοποιείται (γίνεται ίση με τη θετική ποσότητα $E(\varepsilon_i^2)$) μόνο όταν ο θετικός όρος

$$E[(E(Y_i|X_i) - f(X_i))^2]$$

μηδενίζεται δηλαδή όταν

$$E[(E(Y_i|X_i) - f(X_i))^2] = 0$$

άρα μόνο όταν $E(Y_i|X_i) = f(X_i)$.

3. Χρησιμοποιώντας βασική άλγεβρα και μερικές «έξυπνες» αντικαταστάσεις αθροισμάτων όπως

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \Leftrightarrow n\bar{X} = \sum_{i=1}^{n} X_i$$

μπορούμε να δείξουμε ότι **(α)** όταν το γραμμικό υπόδειγμα περιλαμβάνει σταθερό όρο, τότε η μέθοδος ΕΤ δίνει κατάλοιπα με μηδενικό συνολικό άθροισμα άρα και με μηδενικό αριθμητικό μέσο ή αλλιώς

$$\sum_{i=1}^{n} \hat{u}_i = 0 \text{ άρα και } \frac{1}{n}\sum_{i=1}^{n} \hat{u}_i = 0$$

(β) η ερμηνευτική μεταβλητή X_i είναι «ορθογώνια» με τα κατάλοιπα δηλαδή $\sum_{i=1}^{n} X_i\hat{u}_i = 0$ ενώ ισχύει το ίδιο και για τις αποκλίσεις της ερμηνευτικής μεταβλητής από τον δειγματικό της μέσο, δηλαδή

$$\sum_{i=1}^{n} \left(X_i - \bar{X}\right)\hat{u}_i = \sum_{i=1}^{n} x_i\hat{u}_i = 0$$

Απάντηση

(α) Από τη συνθήκη πρώτης τάξης (δηλαδή από την κανονική εξίσωση) του σταθερού όρου, - σχέση (2.14) - έχουμε ότι

$$\sum_{i=1}^{n} \left(Y_i - \hat{\alpha}_{ET} - \hat{\beta}_{ET}X_i\right) = 0$$

$$\Leftrightarrow \sum_{i=1}^{n} \hat{u}_i = 0$$

$$\Leftrightarrow \frac{1}{n}\sum_{i=1}^{n} \hat{u}_i = 0$$

(β) Από τη συνθήκη πρώτης τάξης (2.15) για το συντελεστή κλίσης β

έχουμε

$$\sum_{i=1}^{n}\left(Y_i-\hat{\alpha}_{ET}-\hat{\beta}_{ET}X_i\right)(X_i)=0\Leftrightarrow\sum_{i=1}^{n}X_i\hat{u}_i=0$$

και

$$\begin{aligned}&\sum_{i=1}^{n}X_i\hat{u}_i=0\Leftrightarrow\\&\sum_{i=1}^{n}\left(X_i-\bar{X}+\bar{X}\right)\hat{u}_i=0\Leftrightarrow\\&\sum_{i=1}^{n}x_i\hat{u}_i+\bar{X}\sum_{i=1}^{n}\hat{u}_i=0\Leftrightarrow\\&\sum_{i=1}^{n}x_i\hat{u}_i=0\end{aligned}$$

αφού

$$\bar{X}\sum_{i=1}^{n}\hat{u}_i=\bar{X}\times 0=0$$

4. Δείξτε ότι στο απλό διμεταβλητό υπόδειγμα

$$Y_i=\alpha+\beta X_i+u_i$$

ο συντελεστής προσδιορισμού R^2 είναι ίσος με το δειγματικό συντελεστή συσχέτισης των Y_i, X_i στο τετράγωνο δηλαδή $R^2=\hat{\rho}^2$ όπου ο δειγματικός συντελεστής συσχέτισης $\hat{\rho}$ δίνεται από τον τύπο

$$\hat{\rho}=\frac{\sum\limits_{i=1}^{n}\left(Y_i-\bar{Y}\right)\left(X_i-\bar{X}\right)}{\sqrt{\sum\limits_{i=1}^{n}\left(Y_i-\bar{Y}\right)^2}\sqrt{\sum\limits_{i=1}^{n}\left(X_i-\bar{X}\right)^2}}$$

Ποιο μειονέκτημα, για την ερμηνεία του συντελεστή R^2, είναι συνέπεια αυτού του αποτελέσματος;

Υπόδειξη: σε περίπτωση πλασματικής συσχέτισης μία μεγάλη τιμή του $\hat{\rho}$

συνεπάγεται και μεγάλη τιμή του συντελεστή προσδιορισμού R^2.

5. **Αποδείξτε** ότι στο υπόδειγμα που λείπει ο σταθερός όρος

$$Y_i = \beta X_i + u_i \, , \, i = 1, ..., n$$

ο εκτιμητής ελαχίστων τετραγώνων δίνεται από

$$\hat{\beta}_{ET} = \frac{\sum_{i=1}^{n} Y_i X_i}{\sum_{i=1}^{n} X_i^2}$$

Απάντηση

Έχουμε

$$\begin{aligned} \sum_{i=1}^{n} \hat{u}_i^2 &= \sum_{i=1}^{n} \left(Y_i - \hat{\beta} X_i \right)^2 \\ &= \sum_{i=1}^{n} Y_i^2 - 2\hat{\beta} \sum_{i=1}^{n} Y_i X_i + \hat{\beta}^2 \sum_{i=1}^{n} X_i^2 \\ &= \hat{\beta}^2 K_1 - 2\hat{\beta} K_2 + K_3 \end{aligned}$$

Γνωρίζουμε ότι τετραγωνικές συναρτήσεις του τύπου

$$\hat{\beta}^2 K_1 - 2\hat{\beta} K_2 + K_3$$

με $K_1 > 0$ ελαχιστοποιούνται στο σημείο $\frac{K_2}{K_1}$ άρα ο εκτιμητής ελαχίστων τετραγώνων δίνεται από

$$\hat{\beta}_{ET} = \frac{\sum_{i=1}^{n} Y_i X_i}{\sum_{i=1}^{n} X_i^2}$$

Εναλλακτικά μπορείτε να ελαχιστοποιήσετε τη συνάρτηση

$$S\left(\hat{\beta}\right) = \sum_{i=1}^{n} \hat{u}_i^2 = \sum_{i=1}^{n} \left(Y_i - \hat{\beta} X_i \right)^2$$

ως προς $\hat{\beta}$ με συνθήκη πρώτης τάξης $\frac{dS}{d\hat{\beta}} = 0$ και συνθήκη δεύτερης τάξης $\frac{d^2S}{d\hat{\beta}^2} > 0$ υπολογισμένη στο στάσιμο σημείο.

6. Χρησιμοποιώντας τα αποτελέσματα της άσκησης 5, δείξτε ότι στο υπόδειγμα που περιλαμβάνει μόνο σταθερό όρο

$$Y_i = \alpha + u_i \, , \, i = 1, ..., n$$

ο εκτιμητής ελαχίστων τετραγώνων της σταθεράς δίνεται από

$$\hat{\alpha}_{ET} = \bar{Y}$$

Απάντηση

Η σταθερά α μπορεί να θεωρηθεί ως ο συντελεστής της «μεταβλητής» $X_i = 1$, $\forall i$. Άρα

$$\begin{aligned}\hat{\alpha}_{ET} &= \frac{\sum_{i=1}^{n} Y_i X_i}{\sum_{i=1}^{n} X_i^2} \\ &= \frac{\sum_{i=1}^{n} Y_i \times 1}{\sum_{i=1}^{n} 1^2} \\ &= \frac{\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} 1} = \frac{1}{n}\sum_{i=1}^{n} Y_i = \bar{Y}\end{aligned}$$

7. Δείξτε ότι αν δεν συμπεριληφθεί σταθερός όρος α στο απλό γραμμικό υπόδειγμα, τότε ο συντελεστής προσδιορισμού R^2 μπορεί να είναι αρνητικός (στην περίπτωση αυτή αποφεύγουμε την ερμηνεία του συντελεστή προσδιορισμού).

8. Δείξτε ότι σε μία παλινδρόμηση που περιλαμβάνει **μόνο** σταθερό όρο

$$Y_i = \alpha + u_i$$

ο συντελεστής προσδιορισμού R^2 είναι πάντα ίσος με 0. Σχολιάστε.

9. Δείξτε ότι στο απλό γραμμικό υπόδειγμα, η συνδιακύμανση του εκτιμητή ελαχίστων τετραγώνων $\hat{\beta}_{ET}$ με το διαταρακτικό όρο u_i είναι μη μηδενική όταν $1 \leq i \leq n$. Προβείτε στις αναγκαίες απλουστευτικές υποθέσεις.

Υπόδειξη: εκφράστε τον εκτιμητή $\hat{\beta}_{ET}$ ως συνάρτηση των διαταρακτικών όρων με βάση τον τύπο (σφάλμα δειγματοληψίας του εκτιμητή)

$$\begin{aligned}\hat{\beta}_{ET} &= \frac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} x_i^2} \\ &= \beta + \frac{\sum_{i=1}^{n} x_i \left(u_i - \bar{u}\right)}{\sum_{i=1}^{n} x_i^2} \\ &= \beta + \frac{\sum_{i=1}^{n} x_i u_i}{\sum_{i=1}^{n} x_i^2}\end{aligned}$$

Απάντηση

Έστω ότι η μεταβλητή X_i είναι μη στοχαστική (άρα και η $x_i = X_i - \bar{X}$) και ισχύουν οι κλασσικές υποθέσεις

$$E\left(u_i\right) = 0\,,\; Var\left(u_i\right) = \sigma^2\,,\; Cov\left(u_i, u_j\right) = 0$$

Τότε για $1 \leq i \leq n$ έχουμε

$$\begin{aligned}Cov\left(\hat{\beta}_{ET}, u_i\right) &= Cov\left(\hat{\beta}_{ET} - \beta, u_i\right) \\ [0.25cm] &= E\left[\left(\hat{\beta}_{ET} - \beta\right) u_i\right] \\ &= E\left[\left(\sum_{i=1}^{n} w_i u_i\right) u_i\right] \\ &= E\left[\left(w_1 u_1 + w_2 u_2 + ... + w_n u_n\right) u_i\right]\end{aligned}$$

$$= \sigma^2 w_i \neq 0$$

όπου για αλγεβρική ευκολία θέσαμε

$$w_i = \frac{x_i}{\sum_{i=1}^{n} x_i^2}$$

Όταν ο δείκτης i αντιστοιχεί σε παρατήρηση εκτός του δείγματος, δηλαδή είναι εκτός του προαναφερθέντος διαστήματος $1 \leq i \leq n$, τότε

$$Cov\left(\hat{\beta}_{ET}, u_i\right) = 0$$

Για παράδειγμα

$$Cov\left(\hat{\beta}_{ET}, u_{n+1}\right) = 0$$

ή

$$Cov\left(\hat{\beta}_{ET}, u_0\right) = 0$$

αφού

$$Cov\left(\hat{\beta}_{ET}, u_{n+1}\right)$$

$$\begin{aligned}
&= E\left[\left(w_1 u_1 + w_2 u_2 + ... + w_n u_n\right) u_{n+1}\right] \\
&= E\left(w_1 u_1 u_{n+1} + w_2 u_2 u_{n+1} + ... + w_n u_n u_{n+1}\right) \\
&= E\left(w_1 u_1 u_{n+1}\right) + E\left(w_2 u_2 u_{n+1}\right) + ... + E\left(w_n u_n u_{n+1}\right) \\
&= w_1 E\left(u_1 u_{n+1}\right) + w_2 E\left(u_2 u_{n+1}\right) + ... + w_n E\left(u_n u_{n+1}\right) \\
&= w_1 \times 0 + w_2 \times 0 + ... + w_n \times 0 \\
&= 0
\end{aligned}$$

10. Δείξτε ότι στο απλό γραμμικό υπόδειγμα, η συνδιακύμανση του εκτιμητή ελαχίστων τετραγώνων $\hat{\beta}_{ET}$ και των καταλοίπων της εκτίμησης $\hat{u}_i$ είναι μηδενική για κάθε κατάλοιπο του δείγματος, δηλαδή για κάθε $1 \leq i \leq n$.

Υπόδειξη 1: εκφράστε τα κατάλοιπα $\hat{u}_i$ ως συνάρτηση των διαταρακτικών

όρων u_i

$$\hat{u}_i = (u_i - \bar{u}) - \left(\hat{\beta}_{ET} - \beta\right) x_i$$

Υπόδειξη 2: Για αλγεβρική ευκολία, ορίστε τις σταθμίσεις

$$w_i = \frac{x_i}{\sum_{i=1}^{n} x_i^2}$$

οι οποίες ικανοποιούν την ισότητα

$$\sum_{i=1}^{n} w_i = 0$$

κάτι που συνεπάγεται ότι

$$\begin{aligned}
\frac{1}{n} E\left[\left(\sum_{i=1}^{n} w_i u_i\right)\left(\sum_{i=1}^{n} u_i\right)\right] &= \\
&= \frac{1}{n}\sum_{i=1}^{n} w_i E\left(u_i^2\right) \\
&= \frac{\sigma^2}{n}\sum_{i=1}^{n} w_i \\
&= \frac{\sigma^2}{n} \times 0 \\
&= 0
\end{aligned}$$

11. Έστω το απλό γραμμικό υπόδειγμα

$$Y_i = \alpha + \beta X_i + u_i$$

με τις κλασσικές υποθέσεις να ισχύουν και την ερμηνευτική μεταβλητή X_i να είναι σταθερή σε επαναλαμβανόμενα δείγματα (μη τυχαία μεταβλητή).

Στην προηγούμενη άσκηση είδαμε ότι μπορούμε να εκφράσουμε τα κατάλοιπα της ΕΤ εκτίμησης συναρτήσει των αποκλίσεων των διαταρακτικών όρων από τον αριθμητικό τους μέσο $(u_i - \bar{u})$, του σφάλματος δειγματοληψίας του

εκτιμητή ΕΤ

$$\left(\hat{\beta}_{ET} - \beta\right)$$

και των αποκλίσεων της ερμηνευτικής μεταβλητής από τον αριθμητικό (δειγματικό) μέσο, $x_i = X_i - \bar{X}$.

(α) Δείξτε ότι όπως και η αναμενόμενη τιμή των διαταρακτικών όρων έτσι και η αναμενόμενη τιμή των καταλοίπων είναι μηδέν, δηλαδή δείξτε ότι

$$E\left(\hat{u}_i\right) = 0$$

(β) Δείξτε ότι η διακύμανση των καταλοίπων δίνεται από τον τύπο

$$Var\left(\hat{u}_i\right) = \sigma^2 \left(1 - \frac{1}{n} - \frac{x_i^2}{\sum_{i=1}^{n} x_i^2} \right)$$

άρα είναι μικρότερη της διακύμανσης των διαταρακτικών όρων σ^2

(γ) Αν υποθέσουμε ότι η μεταβλητότητα της ερμηνευτικής μεταβλητής είναι «τυπική» δηλαδή

$$\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2 = \sum_{i=1}^{n} x_i^2 \to +\infty$$

καθώς το δείγμα τείνει στο άπειρο $n \to +\infty$, και $E\left|x_i\right|^2 < +\infty$, τι συμβαίνει με τη διακύμανση των καταλοίπων;

12. Σας δίνεται ένα δείγμα $n = 6$ παρατηρήσεων για τις μεταβλητές Y_i και X_i

Y_i	X_i
4	4
1.876	2
6.846	5
1.041	2
6.515	6
2.715	1

Χρησιμοποιήστε το πρόγραμμα Excel ή gretl και

(α) Σχεδιάστε το διάγραμμα διασποράς των Y_i (κάθετος άξονας) και X_i (οριζόντιος άξονας)

(β) Υπολογίστε τους εκτιμητές ελαχίστων τετραγώνων $\hat{\alpha}_{ET}$ και $\hat{\beta}_{ET}$ ενός απλού διμεταβλητού υποδείγματος $Y_i = \alpha + \beta X_i + u_i$

(γ) Υπολογίστε και σχεδιάστε τα κατάλοιπα $\hat{u}_i$, $i = 1, ..., 6$. Επιβεβαιώστε υπολογιστικά ότι

$$\sum_{i=1}^{n} \hat{u}_i = 0$$

(δ) Επιβεβαιώστε υπολογιστικά ότι $\sum_{i=1}^{n} X_i \hat{u}_i = 0$

(ε) Υπολογίστε τις **προσαρμοσμένες ή προβλεπόμενες τιμές**

$$\hat{Y}_i = \hat{\alpha}_{ET} + \hat{\beta}_{ET} X_i$$

και σχεδιάστε σε ένα διάγραμμα μαζί τις τιμές των Y_i και $\hat{Y}_i$

(στ) Εισάγετε στο διάγραμμα διασποράς των Y_i (κάθετος άξονας) και X_i (οριζόντιος άξονας) τις προσαρμοσμένες τιμές $\hat{Y}_i$ (επίσης οριζόντιος άξονας). Θα εμφανιστούν ως μία ευθεία γραμμή

(ζ) Υπολογίστε και σχολιάστε το συντελεστή προσδιορισμού R^2

13. Με βάση τα δεδομένα του αρχείου kefalaio2data2.xlsx υπολογίστε τον «συντελεστή προσδιορισμού» για την εκτίμηση που υιοθετεί μόνο δύο σημεία, συγκεκριμένα τα $(Y_{\max}, X_{y\max})$ και $(Y_{\min}, X_{y\min})$. **Υπόδειξη:** Οι προσαρμοσμένες τιμές δίνονται στη στήλη με τίτλο «Προσαρμοσμένες τιμές (γραμμή) ».

14. Δημιουργήστε σε ένα φύλλο εργασίας του Excel ή στο gretl ένα τυχαίο δείγμα 70 αριθμών από την κανονική κατανομή με αναμενόμενη τιμή 0 και διακύμανση 1. Ονομάστε τη συγκεκριμένη μεταβλητή u. Η u θα διαδραματίσει το ρόλο των διαταρακτικών όρων ενός γραμμικού υποδείγματος

$$Y_t = \alpha + \beta X_t + u_t \, , \, t = 1, ..., 70$$

με $\alpha = 1, \beta = 2$. Δημιουργήστε τη X_t (δώστε της τον τίτλο X) σύμφωνα με την κανονική κατανομή με αναμενόμενη τιμή 5 και διακύμανση 3. Στη

συνέχεια δημιουργήστε την εξαρτημένη μεταβλητή (δώστε της τον τίτλο Y) με βάση το παραπάνω υπόδειγμα.

Προβείτε σε εκτίμηση των α, β με τη μέθοδο των ελαχίστων τετραγώνων χρησιμοποιώντας τους τύπους που γνωρίζετε μέχρι τώρα.

Επαναλάβετε την άσκηση χρησιμοποιώντας την έτοιμη διαδικασία παλινδρόμησης που παρέχει το πρόγραμμα Excel (από το μενού «Εργαλεία» υιοθετούμε την επιλογή «Ανάλυση Δεδομένων» και χρησιμοποιούμε το εργαλείο ανάλυσης «Παλινδρόμηση»).

Δημιουργήστε το διάγραμμα διασποράς των Y_t, X_t, σχεδιάστε τα κατάλοιπα $\hat{u}_t$ και σχεδιάστε μαζί στον χρόνο τις Y_t και $\hat{Y}_t$.

Υπολογίστε τον συντελεστή προσδιορισμού.

Επαναλάβετε την άσκηση για

$$u_t \sim N.i.d\,(0\,,\,3) \text{ και } u_t \sim N.i.d\,(0\,,\,9)$$

Τι παρατηρείτε στα γραφήματα και στον συντελεστή προσδιορισμού;

15. Με βάση τα δεδομένα για τις μεταβλητές «πραγματικό ωρομίσθιο», έστω w_i, και «ηλικία», έστω x_i, του αρχείου kefalaio2data3.xlsx εκτιμήστε με τη μέθοδο ΕΤ (χρησιμοποιήστε το πρόγραμμα Excel ή **προτιμότερο το πρόγραμμα** gretl) τις παραμέτρους α, β καθώς και το συντελεστή προσδιορισμού R^2 στο υπόδειγμα $w_i = \alpha + \beta x_i + u_i$.

Δημιουργήστε ένα γράφημα για τα κατάλοιπα.

Στη συνέχεια εκτιμήστε και το υπόδειγμα $\ln(w_i) = \alpha + \beta x_i + u_i$. Επίσης, δημιουργήστε ένα γράφημα για τα κατάλοιπα.

(α) Ποιό από τα δύο θα επιλέγατε με βάση τον συντελεστή προσδιορισμού; **Προσοχή** στην παρατήρηση 5 του κειμένου. Μην συγκρίνετε κατευθείαν τους συντελεστές προσδιορισμού από τα δύο υποδείγματα.

Επίσης, προσοχή διότι ακόμα και αν ο συντελεστής προσδιορισμού R^2 δείχνει το πρώτο υπόδειγμα ως καλύτερο, το κέρδος από την οικονομική ερμηνεία των εκτιμημένων παραμέτρων μπορεί να είναι μεγαλύτερο στην περίπτωση του δεύτερου υποδείγματος.

Τέλος, παρατηρήστε ότι η βηματική μεθοδολογία του υπολογισμού ενός μέτρου σύγκρισης των δύο συντελεστών προσδιορισμού $R^2_{\ln}, R^2$ απλώς

υπολογίζει τη διαφορά τους, ενώ θα έπρεπε να προβούμε σε έλεγχο της στατιστικής σημαντικότητας της διαφοράς των $R^2_{\ln}$, R^2.

(β) Επαναλάβετε την άσκηση με ερμηνευτική μεταβλητή τα «έτη εκπαίδευσης».

16. Δημιουργία ετεροσκεδαστικού υποδείγματος: Στο Excel ή στο gretl δημιουργήστε για $i = 1, ..., 125$ δύο μεταβλητές, την $X_i \sim N.i.d\,(0, 4)$ και την $\eta_i \sim N.i.d\,(0, 1)$, οπότε $E\,(X_i) = 0$, $E\,(\eta_i) = 0$ και

$$Var\,(X_i) = E\left(X_i^2\right) = \sigma_X^2 = 4$$
$$Var\,(\eta_i) = E\left(\eta_i^2\right) = \sigma_\eta^2 = 1$$

Οι μεταβλητές X_i, η_i από κατασκευής είναι ανεξάρτητες μεταξύ τους οπότε $E\,(\eta_i|\mathbb{X}) = E\,(\eta_i)$. Δημιουργήστε την Y_i με βάση το υπόδειγμα

$$Y_i = \beta_i X_i$$

όπου

$$\beta_i = \beta + \eta_i\,,\ \beta = 2$$

Στη συνέχεια εκτιμήστε με τη μέθοδο ΕΤ τον συντελεστή κλίσης β από το (προς εκτίμηση) απλό γραμμικό υπόδειγμα

$$Y_i = \beta X_i + u_i$$

που προκύπτει μετά την αντικατάσταση

$$\begin{aligned} Y_i &= \beta_i X_i \\ &= (\beta + \eta_i)\, X_i \\ &= \beta X_i + \eta_i X_i \\ &= \beta X_i + u_i \end{aligned}$$

Παρατηρήστε ότι ο διαταρακτικός όρος του - προς εκτίμηση υποδείγματος - u_i πληροί την υπόθεση $E\,(u_i|\mathbb{X}) = 0$ αλλά είναι ετεροσκεδαστικός αφού η δεσμευμένη διακύμανσή του δίνεται από την

$$E\left(u_i^2|\mathbb{X}\right) = E\left(\eta_i^2 X_i^2|\mathbb{X}\right)$$

$$\begin{aligned} &= X_i^2 E\left(\eta_i^2|\mathbb{X}\right) \\ &= X_i^2 \sigma_\eta^2 \\ &= \sigma_i^2 \end{aligned}$$

Επίσης, παρατηρήστε ότι η μη δεσμευμένη διακύμανση του διαταρακτικού όρου είναι σταθερή και δίνεται από

$$\begin{aligned} E\left(u_i^2\right) &= E\left(E\left(u_i^2|\mathbb{X}\right)\right) \\ &= E\left(X_i^2\sigma_\eta^2\right) = \sigma_\eta^2 E\left(X_i^2\right) \\ &= \sigma_\eta^2 \sigma_X^2 = \sigma^2 \end{aligned}$$

(α) αφού εκτιμήσετε το β με τη μέθοδο ΕΤ δημιουργήστε το γράφημα διασποράς των καταλοίπων στο τετράγωνο $\hat{u}_i^2$ (οριζόντιος άξονας) ως προς την ερμηνευτική μεταβλητή στο τετράγωνο X_i^2 (κάθετος άξονας).

(β) επαναλάβετε την άσκηση με ένα παρόμοιο υπόδειγμα που δεν έχει πρόβλημα ετεροσκεδαστικότητας, π.χ., $Y_i = 2X_i + \eta_i$. Τι παρατηρείτε στην περίπτωση της ετεροσκεδαστικότητας;

17. Με τα δεδομένα[14] του αρχείου

`US_census_bureau_Communities.gdt`

εκτιμήστε το υπόδειγμα

$$Y_i = \alpha + \beta X_i + u_i \ , \ i = 1, 2, ..., 1994 \text{ (κοινότητες των Η.Π.Α)}$$

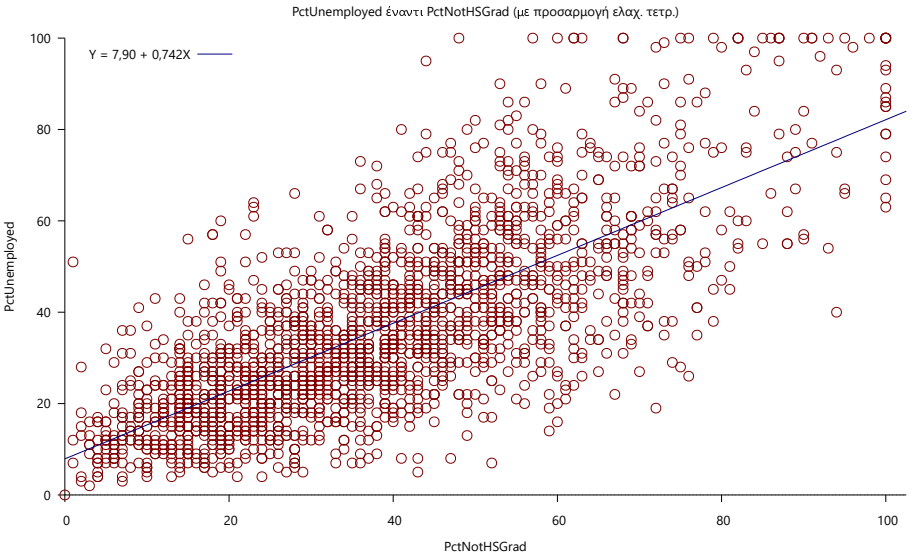
όπου Y_i η μεταβλητή PctUnemployed: **ποσοστό ανεργίας** και μετρά το ποσοστό των πολιτών σε κάθε κοινότητα που είναι άνεργοι (είναι μεταξύ του 0 και του 1, οπότε πολλαπλασιάστε αρχικά με 100 ώστε να διευκολύνουμε την οικονομική ερμηνεία) και X_i η μεταβλητή PctNotHSGrad: που μετρά το ποσοστό των πολιτών σε κάθε κοινότητα που δεν κατάφερε να τελειώσει τη δευτεροβάθμια εκπαίδευση (non High School graduates) (επίσης δίνεται ως αριθμός μεταξύ του 0 και του 1, οπότε πολλαπλασιάστε και αυτή τη μεταβλητή με 100 προς διευκόλυνση της ερμηνείας των αποτελεσμάτων).

[14] Πηγή δεδομένων, `https://raw.githubusercontent.com/QMUL-SPIR/Public_files/master/datasets/communities.csv`

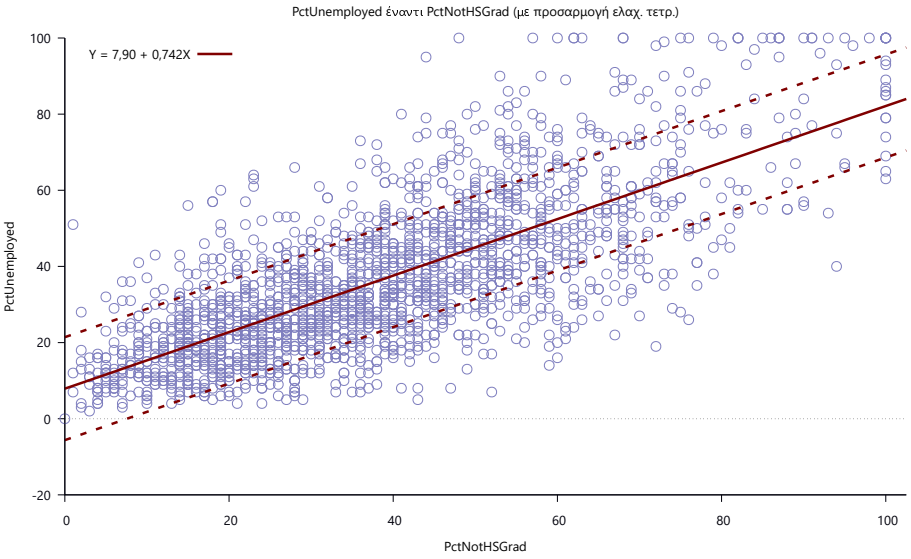
Κατασκευάστε τα δύο γραφήματα διασποράς που φαίνονται αμέσως παρακάτω και σχολιάστε. Επίσης, σχολιάστε κατάλληλα την εκτίμηση ελαχίστων τετραγώνων (υποθέστε προς στιγμήν ότι η εκτίμηση είναι στατιστικά σημαντική). **Υπόδειξη:** Ο εκτιμημένος συντελεστής για τη μεταβλητή PctNotHSGrad υπολογίστηκε σε 0.742, πράγμα που σημαίνει ότι μία αύξηση του ποσοστού των πολιτών χωρίς δευτεροβάθμια εκπαίδευση **κατά μία ποσοστιαία μονάδα** σχετίζεται/υποδηλώνει αύξηση στο ποσοστό ανεργίας των πολιτών κατά **0.742 ποσοστιαίες μονάδες**.

Η εκτιμημένη γραμμή παλινδρόμησης αντικατοπτρίζει τη θετική σχέση των μεταβλητών (θετική κλίση) άρα, υψηλότερα επίπεδα μη-αποφοίτησης συνδέονται με υψηλότερα επίπεδα του ποσοστού ανεργίας, όμως η σχέση δεν είναι εντελώς 1 προς 1. Δηλαδή, για κάθε επιπλέον ποσοστιαία μονάδα της PctNotHSGrad, το ποσοστό των πολιτών που είναι άνεργοι αυξάνεται λιγότερο από μία ποσοστιαία μονάδα.

Ο συντελεστής προσδιορισμού είναι αρκετά υψηλός, $R^2 = 55.29$, δηλαδή το 55.29% της μεταβλητότητας της ανεργίας εξηγείται από τη μη-αποφοίτηση από τη δευτεροβάθμια εκπαίδευση. Βέβαια, στο βαθμό που υπάρχουν μεταβλητές οι οποίες επηρεάζουν τις PctUnemployed και PctNotHSGrad το συγκεκριμένο ποσοστό είναι μάλλον υπερεκτιμημένο (και δεν βεβαιώνει - χωρίς περαιτέρω διερεύνηση - σχέση αιτίας αιτιατού).



Γράφημα 2.9: Διάγραμμα διασποράς των μεταβλητών Y_i (PctUnemployed) και X_i (PctNotHSGrad) μαζί με τις εκτιμημένες τιμές $\hat{Y}_i = 7.90 + 0.742 \cdot X_i$.



Γράφημα 2.10: Διάγραμμα διασποράς των μεταβλητών Y_i (PctUnemployed) και X_i (PctNotHSGrad) μαζί με τις εκτιμημένες τιμές $\hat{Y}_i = 7.90 + 0.742 \cdot X_i$ **και μαζί με τα όρια** $\hat{Y}_i \pm \hat{\sigma}_u$.

ΚΕΦΑΛΑΙΟ 3

Στατιστική επαγωγή στο απλό γραμμικό υπόδειγμα

3.1 Έλεγχος στατιστικής σημαντικότητας συντελεστών

Στο απλό γραμμικό υπόδειγμα της μορφής

$$Y_i = \alpha + \beta X_i + u_i, \quad i = 1, ..., n \tag{3.1}$$

υποθέτουμε ότι ισχύει το πρώτο σύνολο των κλασσικών υποθέσεων, δηλαδή

$$u_i \sim N.i.d\left(0, \sigma^2\right) \quad , \quad \forall i$$

και για αλγεβρική ευκολία υιοθετούμε τη **μη ρεαλιστική υπόθεση** ότι η ερμηνευτική μεταβλητή X_i, $\forall i$ είναι **μη στοχαστική** (**μη τυχαία**), άρα θεωρείται δεδομένη σε επαναλαμβανόμενα δείγματα. Οι διαταρακτικοί όροι u_i είναι τυχαίες μεταβλητές που κατανέμονται κανονικά και ανεξάρτητα με τον ίδιο μηδενικό μέσο και την ίδια (άγνωστη) διακύμανση σ^2 για κάθε i. Η εξαρτημένη μεταβλητή Y_i θεωρείται λοιπόν μία τυχαία μεταβλητή για κάθε i αφού είναι μία απλή γραμμική συνάρτηση των διαταρακτικών όρων u_i.

Με τη μέθοδο ελαχίστων τετραγώνων (ΕΤ) υπολογίζουμε τους εκτιμητές των παραμέτρων α, β, σ^2. Οι εκτιμητές θα συμβολίζονται:

- γενικά με $\hat{\alpha}, \hat{\beta}, \hat{\sigma}^2$

- ή με $\hat{\alpha}_{ET}, \hat{\beta}_{ET}, \hat{\sigma}^2_{ET}$ όταν θέλουμε να ξεχωρίσουμε τη μέθοδο ΕΤ από άλλες μεθόδους ή να δώσουμε έμφαση στη χρήση της μεθόδου
- ή με $\hat{\alpha}_n, \hat{\beta}_n, \hat{\sigma}^2_n$ όταν θέλουμε να δώσουμε έμφαση στην εξάρτηση των εκτιμητών από το μέγεθος του δείγματος n. Στην περίπτωση χρονοσειρών αντίστοιχα θα μπορούσαμε να γράψουμε $\hat{\alpha}_T, \hat{\beta}_T, \hat{\sigma}^2_T$

Μετά την εκτίμηση των παραμέτρων α, β, σ^2 γράφουμε το εκτιμημένο υπόδειγμα ως

$$Y_i = \hat{\alpha} + \hat{\beta} X_i + \hat{u}_i, \quad i = 1, ..., n$$

όπου $\hat{u}_i$ είναι τα κατάλοιπα (εκτίμηση των διαταρακτικών όρων u_i).

Οι εκτιμητές ΕΤ $\hat{\alpha}, \hat{\beta}, \hat{\sigma}^2$ δίνονται αναλυτικά από τις σχέσεις

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$$

$$\hat{\beta} = \frac{\sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)\left(X_i - \bar{X}\right)}{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2} = \frac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} x_i^2}$$

και

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} \hat{u}_i^2$$

Είναι εμφανές ότι οι εκτιμητές $\hat{\alpha}, \hat{\beta}, \hat{\sigma}^2$ αποτελούν συναρτήσεις τυχαίων μεταβλητών και η τιμή που θα λάβουμε σε κάποιο δείγμα για τις $\hat{\alpha}, \hat{\beta}, \hat{\sigma}^2$ εξαρτάται από την τιμή που έλαβε (τουλάχιστον[1]) η εξαρτημένη μεταβλητή Y_i στο συγκεκριμένο δείγμα. Ένα διαφορετικό δείγμα θα έδινε διαφορετικές εκτιμήσεις, άρα οι εκτιμητές είναι τυχαίες μεταβλητές που υπόκεινται σε κατανομές δειγματοληψίας[2]. Οπότε, σκοπός μας είναι, **πρώτα** η εκτίμηση των άγνωστων συντελεστών του υποδείγματος **και κατόπιν** η στατιστική επαγωγή για τους εκτιμημένους συντελεστές που θα επιτρέψει την ποιοτική διερεύνηση των χαρακτηριστικών του υποδείγματος.

Επιπλέον, οι εκτιμητές ΕΤ έχουν στατιστικές ιδιότητες οι οποίες τους καθιστούν ελκυστικούς έναντι άλλων εκτιμητών. Για παράδειγμα, κάτω από τις (αυστηρές) υποθέσεις του υποδείγματος (3.1) αποδεικνύεται ότι οι εκτιμητές ΕΤ

[1] Αν υιοθετήσουμε το δεύτερο σύνολο κλασσικών υποθέσεων (μία πιο ρεαλιστική κίνηση), τότε η τιμή του εκτιμητή εξαρτάται **και** από τις τιμές των τυχαίων μεταβλητών X_i.

[2] Κατανομές που μεταβάλλονται με το δείγμα.

είναι **αμερόληπτοι**, δηλαδή

$$E(\hat{\alpha}) = \alpha\,,\; E\left(\hat{\beta}\right) = \beta \text{ και } E\left(\hat{\sigma}^2\right) = \sigma^2$$

Το παραπάνω αποτέλεσμα προκύπτει αφού αναλύσουμε τον εκτιμητή ΕΤ $\hat{\beta}$ ως συνάρτηση του διαταρακτικού όρου (του τυχαίου στοιχείου του υποδείγματος) και εφαρμόσουμε τον τελεστή της αναμενόμενης τιμής. Για παράδειγμα, σχετικά με το $\hat{\beta}$, και υιοθετώντας την

$$\begin{aligned} y_i &= Y_i - \bar{Y} \\ &= \alpha + \beta X_i + u_i - \left(\alpha + \beta\bar{X} + \bar{u}\right) \\ &= \beta\left(X_i - \bar{X}\right) + u_i - \overline{u} \\ &= \beta x_i + u_i - \overline{u} \end{aligned}$$

έχουμε ότι

$$\begin{aligned} \hat{\beta} &= \frac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} x_i^2} = \frac{\sum_{i=1}^{n} \left(\beta x_i + u_i - \bar{u}\right) x_i}{\sum_{i=1}^{n} x_i^2} \\ &= \beta \cdot \frac{\sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n} x_i^2} + \frac{\sum_{i=1}^{n} x_i u_i}{\sum_{i=1}^{n} x_i^2} + \frac{\bar{u}\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2} \\ &= \beta + \frac{\sum_{i=1}^{n} x_i u_i}{\sum_{i=1}^{n} x_i^2} \\ &= \beta + \frac{1}{\left(\sum_{i=1}^{n} x_i^2\right)} \sum_{i=1}^{n} x_i u_i \end{aligned}$$

$$= \beta + \left(\sum_{i=1}^{n} x_i^2 \right)^{-1} \sum_{i=1}^{n} x_i u_i$$

Η παραπάνω σχέση είναι σημαντική και ονομάζεται **σφάλμα δειγματοληψίας** του εκτιμητή $\hat{\beta}$ και μπορεί να γραφεί αλγεβρικώς ευκολότερα αν αντικαταστήσουμε τους όρους

$$\frac{x_i}{\sum_{i=1}^{n} x_i^2}$$

με τα (για παράδειγμα)

$$w_i = \frac{x_i}{\sum_{i=1}^{n} x_i^2}$$

οπότε το **σφάλμα δειγματοληψίας** του εκτιμητή $\hat{\beta}$ λαμβάνει τη μορφή

$$\hat{\beta} = \beta + \sum_{i=1}^{n} w_i u_i \quad \text{ή} \quad \hat{\beta} - \beta = \sum_{i=1}^{n} w_i u_i$$

Κατόπιν εφαρμογής του τελεστή της αναμενόμενης τιμής στην παραπάνω σχέση, η αναμενόμενη τιμή του εκτιμητή δίνεται από

$$E\left(\hat{\beta}\right) = E\left(\beta\right) + E\left(\frac{\sum_{i=1}^{n} x_i u_i}{\sum_{i=1}^{n} x_i^2} \right) = \beta + \frac{E\left(\sum_{i=1}^{n} x_i u_i \right)}{\sum_{i=1}^{n} x_i^2}$$

$$= \beta + \frac{\sum_{i=1}^{n} x_i E\left(u_i\right)}{\sum_{i=1}^{n} x_i^2} = \beta + \frac{\sum_{i=1}^{n} x_i \times 0}{\sum_{i=1}^{n} x_i^2} = \beta$$

Παρατηρήστε ότι η δεύτερη ισότητα παραπάνω ισχύει μόνο όταν τα X_i, άρα και τα x_i, θεωρηθούν μη στοχαστικά.

Η έννοια της αμεροληψίας είναι σημαντική σε **πεπερασμένα δείγματα** (ή αλλιώς σε «μικρά» δείγματα) αφού υποδηλώνει πως ο μέσος της κατανομής του εκτιμητή (αναμενόμενη τιμή του εκτιμητή) συμπίπτει με την πληθυσμιακή τιμή της παραμέτρου που προσπαθούμε να εκτιμήσουμε.

Άρα παρότι δεν μπορούμε ποτέ να ισχυριστούμε ότι ο εκτιμητής συμπίπτει

με την τιμή της άγνωστης παραμέτρου $\hat{\beta} = \beta$, μπορούμε να ισχυριστούμε ότι η αναμενόμενη τιμή του εκτιμητή $E\left(\hat{\beta}\right)$ συμπίπτει με την πραγματική τιμή β.

Μία πιο διαισθητική προσέγγιση της έννοιας της αμεροληψίας υπονοεί πως αν χρησιμοποιούσαμε το ίδιο μέγεθος δείγματος n για να εκτιμήσουμε το $\hat{\beta}$ χρησιμοποιώντας ένα μεγάλο αριθμό δειγμάτων, τότε ο μέσος των εκτιμητών $\hat{\beta}$ από τα παραπάνω δείγματα θα έτεινε να εξισωθεί με τη τιμή β καθώς επαναλαμβάνουμε τη δειγματοληψία και την εκτίμηση.

Παρατηρούμε ότι η αμεροληψία είναι μία ελκυστική θεωρητική ιδιότητα όμως στην εφαρμοσμένη οικονομετρία διαδραματίζει μικρό ρόλο αφού ο εμπειρικός ερευνητής έχει στη διάθεσή του, συνήθως, μόνο ένα δείγμα δεδομένων, το οποίο αν είναι «ατυχές» τότε μπορεί να παράγει εκτίμηση αρκετά μακριά από την παράμετρο β, δηλαδή κοντά στις απολήξεις της κατανομής δειγματοληψίας του $\hat{\beta}$. Επιπλέον, η ιδιότητα της αμεροληψίας στο απλό γραμμικό υπόδειγμα βασίζεται σε υποθέσεις οι οποίες είναι αρκετά αυστηρές, όπως θα διαπιστώσουμε στη συνέχεια[3].

Επίσης, σημαντικό ρόλο στη στατιστική επαγωγή διαδραματίζει και η δεύτερη ροπή των εκτιμητών, δηλαδή η διακύμανσή τους, αφού σε αυτή βασίζεται η στατιστική ιδιότητα της **αποτελεσματικότητας** ή ακρίβειας των εκτιμητών, ενώ αποτελεί και ουσιαστικό μέρος της στατιστικής επαγωγής ως παράγοντας τυποποίησης των στατιστικών ελέγχου.

Η διακύμανση των εκτιμητών ΕΤ $\hat{\alpha}, \hat{\beta}, \hat{\sigma}^2$ των συντελεστών του απλού γραμμικού υποδείγματος δίνεται από τους παρακάτω τύπους

$$Var\left(\hat{\alpha}\right) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2} \right) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n} x_i^2} \right)$$

$$Var\left(\hat{\beta}\right) = \frac{\sigma^2}{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2} = \sigma^2 \left(\sum_{i=1}^{n} x_i^2 \right)^{-1}$$

και

$$Var\left(\hat{\sigma}^2\right) = \frac{2\sigma^4}{n-2}$$

[3] Ένα σχετικό ανέκδοτο αφορά τρεις «οικονομέτρες» οι οποίοι έχουν βγει για κυνήγι. Καθώς σημαδεύουν την (τυχερή...) πάπια ο πρώτος πυροβολεί λίγο δεξιά της, ο δεύτερος λίγο αριστερά της και ο τρίτος αναφωνεί «...την πετύχαμε!».

Για παράδειγμα, αν θέσουμε τώρα (ως αλγεβρική διευκόλυνση) ότι

$$w = \sum_{i=1}^{n} x_i^2$$

τότε η διακύμανση του εκτιμητή ελαχίστων τετραγώνων της κλίσης $\hat{\beta}$, υπό τις υποθέσεις ότι x_i είναι **μη στοχαστική** και οι διαταρακτικοί όροι u_i είναι ανεξάρτητες (άρα και ασυσχέτιστες) τυχαίες μεταβλητές με αναμενόμενη τιμή 0 και διακύμανση σ^2, δίνεται από

$$Var\left(\hat{\beta}\right) = Var\left(\beta + \frac{1}{w}\sum_{i=1}^{n} x_i u_i\right) = Var\left(\frac{1}{w}\sum_{i=1}^{n} x_i u_i\right)$$

$$= \frac{1}{w^2} Var\left(\sum_{i=1}^{n} x_i u_i\right) = \frac{1}{w^2}\sum_{i=1}^{n} Var\left(x_i u_i\right)$$

$$= \frac{1}{w^2}\sum_{i=1}^{n} x_i^2 Var\left(u_i\right) = \sigma^2 \frac{\sum_{i=1}^{n} x_i^2}{w^2}$$

$$= \sigma^2 \frac{\sum_{i=1}^{n} x_i^2}{\left(\sum_{i=1}^{n} x_i^2\right)^2}$$

$$= \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2}$$

Αφού έχουμε υπολογίσει την αναμενόμενη τιμή και τη διακύμανση των εκτιμητών μένει να διαπιστώσουμε αν μπορούμε να βρούμε την κατανομή τους.
Κάτω από την **υπόθεση της κανονικότητας και ανεξαρτησίας των διαταρακτικών όρων**,

$$u_i \sim N.i.d\left(0\,,\,\sigma^2\right)$$

αποδεικνύεται (δείτε άσκηση 1) ότι

$$\hat{\beta} \sim N.i.d\left(\beta, \sigma^2 \left(\sum_{i=1}^{n} x_i^2\right)^{-1}\right)$$

και

$$\frac{(n-2)\,\hat{\sigma}^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n} \hat{u}_i^2 \sim \chi^2_{n-2}$$

Τα παραπάνω δύο αποτελέσματα είναι ουσιώδη στη βασική στατιστική επαγωγή του απλού γραμμικού υποδείγματος. **Θεωρητικά**, ένας έλεγχος υπόθεσης σχετικά με την άγνωστη παράμετρο β θα μπορούσε να βασιστεί στην τυποποιημένη κανονική μεταβλητή

$$z = \frac{\hat{\beta} - E\left(\hat{\beta}\right)}{\sqrt{Var\left(\hat{\beta}\right)}} = \frac{\hat{\beta} - \beta}{se\left(\hat{\beta}\right)} \sim N\left(0,1\right) \tag{3.2}$$

όπου

$$se\left(\hat{\beta}\right) = \sqrt{Var\left(\hat{\beta}\right)} = \sigma\left(\sum_{i=1}^{n} x_i^2\right)^{-\frac{1}{2}} \tag{3.3}$$

Όμως, στην παραπάνω στατιστική, η παράμετρος της διακύμανσης του διαταρακτικού όρου σ^2, άρα και η τυπική απόκλιση σ, **είναι άγνωστη** οπότε **δεν μπορούμε να προβούμε σε εμπειρική χρήση της στατιστικής**. Καταφεύγουμε λοιπόν, στη χρήση ενός γνωστού θεωρήματος από τη στατιστική θεωρία και των (3.2) και (3.3) ώστε να «απαλλαγούμε» από την άγνωστη παράμετρο σ στη στατιστική ελέγχου.

Συγκεκριμένα,

Θεώρημα. «Μία τυχαία μεταβλητή στην παρακάτω μορφή κλάσματος

$$t = \frac{z}{\sqrt{\frac{y}{n}}}$$

κατανέμεται ως t-student με n βαθμούς ελευθερίας, δηλαδή

$$t \sim t_n$$

όταν

(α) ο αριθμητής του κλάσματος κατανέμεται σύμφωνα με την τυποποιημένη κανονική κατανομή $z \sim N(0,1)$

(β) ο παρανομαστής του κλάσματος δίνεται από την τετραγωνική ρίζα μίας $y \sim \chi^2_n$ τυχαίας μεταβλητής δια τους βαθμούς ελευθερίας n και

(γ) οι τυχαίες μεταβλητές z, y είναι ανεξάρτητες»

Συνεπάγεται ότι μπορούμε να «διώξουμε» την άγνωστη παράμετρο του πληθυσμού σ από τη στατιστική

$$z = \frac{\hat{\beta} - E\left(\hat{\beta}\right)}{\sigma\sqrt{\sum_{i=1}^{n} x_i^2}} = \frac{\hat{\beta} - \beta}{se\left(\hat{\beta}\right)}$$

μέσω της διαίρεσης των (3.2) και (3.3),

$$t_{\hat{\beta}} = \frac{\left(\frac{\hat{\beta}-\beta}{se\left(\hat{\beta}\right)}\right)}{\sqrt{\frac{(n-2)\hat{\sigma}^2}{\sigma^2}}} = \frac{\hat{\beta} - \beta}{\widehat{se}\left(\hat{\beta}\right)}$$

όπου

$$\widehat{se}\left(\hat{\beta}\right) = \frac{\hat{\sigma}}{\left(\sum_{i=1}^{n} x_i^2\right)^{\frac{1}{2}}} \quad \text{ή} \quad = \hat{\sigma}\left(\sum_{i=1}^{n} x_i^2\right)^{-\frac{1}{2}}$$

η εκτίμηση του τυπικού σφάλματος του εκτιμημένου συντελεστή.

Η νέα στατιστική $t_{\hat{\beta}}$ ονομάζεται **t-student στατιστική** και κατανέμεται σύμφωνα με μία *t-student* τυχαία μεταβλητή με $n-2$ βαθμούς ελευθερίας, δηλαδή

$$t_{\hat{\beta}} = \frac{\hat{\beta} - \beta}{\widehat{se}\left(\hat{\beta}\right)} \sim t_{n-2} \tag{3.4}$$

αν ισχύουν οι προϋποθέσεις του παραπάνω θεωρήματος. **Σημείωση:** έχουμε δείξει την **(α)**, έχουμε αναφέρει χωρίς απόδειξη την **(β)** ενώ αποδεικνύεται ότι αριθμητής και παρονομαστής είναι ανεξάρτητες τυχαίες μεταβλητές άρα ικανο-

ποιείται και η προϋπόθεση **(γ)**.

Η τετραγωνική ρίζα της διακύμανσης του εκτιμητή ονομάζεται τυπικό σφάλμα και η εκτίμηση του τυπικού σφάλματος του δίνεται από

$$\widehat{se}\left(\hat{\beta}\right) = \sqrt{\widehat{Var}\left(\hat{\beta}\right)} = \hat{\sigma}\left(\sum_{i=1}^{n} x_i^2\right)^{-1/2}$$

Δηλαδή, «υιοθετεί» τον αμερόληπτο εκτιμητή

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n} \hat{u}_i^2$$

της διακύμανσης σ^2 του διαταρακτικού όρου.

Σε αντίθεση, το τυπικό σφάλμα

$$se\left(\hat{\beta}\right) = \sqrt{Var\left(\hat{\beta}\right)}$$

που περιέχεται στη στατιστική z είναι συνάρτηση της διακύμανσης σ^2 ή τυπικής απόκλισης σ η οποία μας είναι άγνωστη.

Με βάση τη στατιστική t-student (3.4) μπορούμε να προβούμε σε μονόπλευρο ή δίπλευρο έλεγχο υπόθεσης. Συνήθως, για τους συντελεστές κλίσης οι έλεγχοι είναι δίπλευροι, δηλαδή ελέγχουμε την

Μηδενική υπόθεση	$H_0:$	β	$=$	αριθμός
έναντι της				
εναλλακτικής υπόθεσης	$H_1:$	β	$\neq$	αριθμός

Ο έλεγχος δίπλευρων υποθέσεων πραγματοποιείται ως εξής. Υπολογίζουμε την t-student στατιστική

$$t_{\hat{\beta}} = \frac{\hat{\beta} - \text{αριθμός}}{\widehat{se}\left(\hat{\beta}\right)} \sim t_{n-2}$$

και τη συγκρίνουμε με την κρίσιμη τιμή (έστω $t_{n-2}^{\alpha/2}$) από τους πίνακες της t-student κατανομής με $n-2$ βαθμούς ελευθερίας και δεδομένο α (επίπεδο σημαντικότητας).

Το επίπεδο σημαντικότητας α ορίζει τη πιθανότητα να απορρίψουμε τη μηδενική υπόθεση H_0 ενώ είναι σωστή ή εναλλακτικά είναι η πιθανότητα να προβούμε σε **σφάλμα τύπου I (type I error)** σύμφωνα με τη στατιστική ορολογία.

Αν ισχύει ότι $|t_{\hat{\beta}}| \leq t_{n-2}^{\alpha/2}$, τότε **δεν απορρίπτουμε**[4] τη μηδενική υπόθεση σε επίπεδο σημαντικότητας $(100\alpha)\%$, ενώ όταν $|t_{\hat{\beta}}| > t_{n-2}^{\alpha/2}$ τότε απορρίπτουμε τη μηδενική υπόθεση σε επίπεδο σημαντικότητας $(100\alpha)\%$.

Στην οικονομετρική - και γενικότερη στατιστική - πρακτική συνηθίζεται να θέτουμε $\alpha = 0.05$ ή αν δεν είμαστε «αυστηροί[5]» τότε $\alpha = 0.10$. Είναι σπάνιο να υιοθετήσουμε επίπεδο σημαντικότητας $\alpha = 0.01$, αν και απόρριψη της μηδενικής υπόθεσης σε τέτοια επίπεδα και για «συμβατικά μεγέθη δείγματος» παρέχει «σημαντικότατες ενδείξεις» εναντίον της μηδενικής υπόθεσης H_0.

Υπάρχει και δεύτερος τύπος σφάλματος στη στατιστική θεωρία. Το **σφάλμα τύπου II (type II error)**, που ορίζεται ως η πιθανότητα να μην απορρίψουμε μία λανθασμένη μηδενική υπόθεση. **Προσοχή:** αν το να υποπέσουμε σε σφάλμα τύπου II (δηλαδή να μην απορρίψουμε μία λανθασμένη μηδενική υπόθεση) είναι πολύ σημαντικό ή «κοστοβόρο», τότε επιλέγουμε μεγαλύτερα επίπεδα σημαντικότητας π.χ. στο $\alpha = 10\%$, δηλαδή είμαστε λιγότερο αυστηροί με τη μηδενική. Στην αντίθετη περίπτωση, αν το σφάλμα τύπου II θεωρούμε ότι δεν έχει ιδιαίτερο κόστος, ενώ ένα σφάλμα τύπου I οδηγεί στην κοστοβόρα απόρριψη μιας «καλής» μηδενικής υπόθεσης, τότε επιλέγουμε μικρότερα επίπεδα α π.χ. $\alpha = 1\%$.

Σχεδόν όλα τα συγγράμματα Στατιστικής και Οικονομετρίας συμπυκνώνουν στον παρακάτω πίνακα τα σφάλματα τύπου I και II, αλλά είναι γεγονός ότι η

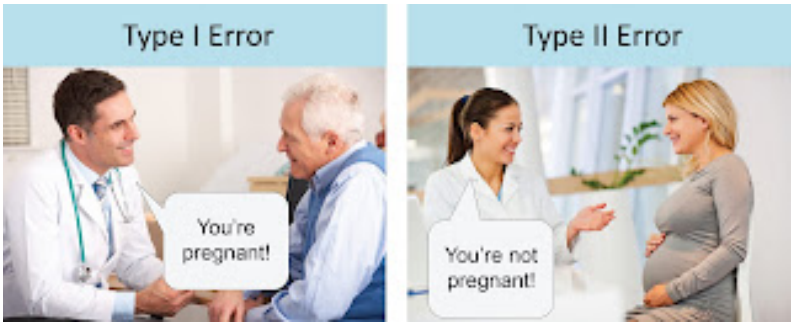
	Απόφαση: Δεν απορρίπτω	**Απόφαση: Απορρίπτω**
H0: Αληθής/Σωστή	OK	Σφάλμα τύπου I
H0: Εσφαλμένη	Σφάλμα τύπου II	OK

απομνημόνευση της διαφοράς των δύο τύπων είναι πιό σίγουρη μέσα από μία

[4] Θεωρούμε ότι δεν υπάρχουν ισχυρές ενδείξεις για την απόρριψη της υποθέσεως και όχι ότι η μηδενική υπόθεση είναι πραγματικά ορθή.

[5] Καθώς το δείγμα μεγαλώνει τα τυπικά σφάλματα μικραίνουν (η «ακρίβεια» των εκτιμητών μεγαλώνει), γι'αυτό συνηθίζεται να θέτουμε το $\alpha = 1\%$ για «μεγάλα» δείγματα και το $\alpha = 10\%$ για «μικρά» δείγματα. Για τις ανάγκες του μαθήματος, η επιλογή $\alpha = 5\%$ θα είναι αρκετή.

εικόνα, (μηδενική υπόθεση $H0$: Όχι εγκυμοσύνη ή «πιο» μαθηματικά $H0$: Εγκυμοσύνη=0)



Γράφημα 3.1: Πηγή: `http://unbiasedresearch.blogspot.com/2016/04/type-i-and-type-ii-errors.html`

Σχετικά με τους συμβολισμούς, συνηθίζεται να γράφουμε το α και επί τοις εκατό, π.χ., αν $\alpha = 0.05$ τότε το επίπεδο σημαντικότητας λέμε ότι είναι 5%. Όταν ο **δίπλευρος έλεγχος** είναι της μορφής

$$\begin{aligned} H_0: &\quad \beta = 0 \\ H_1: &\quad \beta \neq 0 \end{aligned}$$

δηλαδή όταν ελέγχουμε **αν η παράμετρος του πληθυσμού είναι μηδενική** και απορρίψουμε τη μηδενική υπόθεση λέμε ότι

- *«η εκτίμηση $\hat{\beta}$ διαφέρει σημαντικά από το μηδέν»*
- *«ή ότι η εκτίμηση είναι στατιστικά σημαντική»*
- *«ή ότι η μεταβλητή X_i έχει στατιστικά σημαντική επίδραση στην Y_i»*

Ο συγκεκριμένος έλεγχος ονομάζεται **έλεγχος σημαντικότητας** του εκτιμημένου συντελεστή και είναι ο βασικότερος έλεγχος που διεξάγουμε στα πρώτα στάδια της εμπειρικής ανάλυσης.

Το $(1-\alpha)\%$ διάστημα εμπιστοσύνης ορίζεται μέσω της παρακάτω πιθανότητας

$$P\left(-t_{n-2}^{\alpha/2} \leq \frac{\hat{\beta} - \beta}{\widehat{se}\left(\hat{\beta}\right)} \leq t_{n-2}^{\alpha/2}\right) = 1 - \alpha$$

όπου $t_{n-2}^{\alpha/2}$ αντιστοιχεί στην «κατάλληλη» κρίσιμη τιμή από τους πίνακες της t-student κατανομής με $n-2$ βαθμούς ελευθερίας. Αν αναδιατάξουμε τις ανισότητες μέσα στην πιθανότητα έχουμε

$$P\left(-t_{n-2}^{\alpha/2} \leq \frac{\hat{\beta}-\beta}{\widehat{se}\left(\hat{\beta}\right)} \leq t_{n-2}^{\alpha/2}\right) = 1-\alpha \Leftrightarrow$$

$$P\left(-t_{n-2}^{\alpha/2} \cdot \widehat{se}\left(\hat{\beta}\right) \leq \hat{\beta}-\beta \leq t_{n-2}^{\alpha/2} \cdot \widehat{se}\left(\hat{\beta}\right)\right) = 1-\alpha \Leftrightarrow$$

$$P\left(-t_{n-2}^{\alpha/2} \cdot \widehat{se}\left(\hat{\beta}\right) - \hat{\beta} \leq -\beta \leq t_{n-2}^{\alpha/2} \cdot \widehat{se}\left(\hat{\beta}\right) - \hat{\beta}\right) = 1-\alpha \Leftrightarrow$$

$$P\left(\hat{\beta} - t_{n-2}^{\alpha/2} \cdot \widehat{se}\left(\hat{\beta}\right) \leq \beta \leq \hat{\beta} + t_{n-2}^{\alpha/2} \cdot \widehat{se}\left(\hat{\beta}\right)\right) = 1-\alpha$$

δηλαδή η πιθανότητα οι τυχαίες μεταβλητές

$$\hat{\beta} - t_{n-2}^{\alpha/2} \cdot \widehat{se}\left(\hat{\beta}\right)$$

και

$$\hat{\beta} + t_{n-2}^{\alpha/2} \cdot \widehat{se}\left(\hat{\beta}\right)$$

να λαμβάνουν τιμές που περικλείουν την παράμετρο του πληθυσμού β είναι $(1-\alpha)\%$. Το διάστημα εμπιστοσύνης μπορεί επίσης να οριστεί ως το διάστημα των τιμών $\beta^{(0)}$ της παραμέτρου β για το οποίο η μηδενική υπόθεση $H_0 : \beta = \beta^{(0)}$ δεν απορρίπτεται από τους δίπλευρους ελέγχους.

Με παρόμοιο τρόπο μπορούμε να κατασκευάσουμε διαστήματα εμπιστοσύνης και να προβούμε σε έλεγχο υποθέσεων σχετικά με το σταθερό όρο του υποδείγματος α αλλά και τη διακύμανση σ^2 του διαταρακτικού όρου.

Παρατηρήστε όμως ότι $\sigma^2 > 0$, δηλαδή η διακύμανση λαμβάνει μόνο θετικές τιμές και η κατανομή χ^2 δεν είναι συμμετρική γύρω από τη μέση τιμή της. Για παράδειγμα το $100(1-\alpha)\%$ διάστημα εμπιστοσύνης για τη διακύμανση του διαταρακτικού όρου σ^2 βασίζεται στο αποτέλεσμα ότι η τυχαία μεταβλητή

$$\frac{(n-2)\,\hat{\sigma}^2}{\sigma^2} \sim \chi_{n-2}^2$$

ακολουθεί την κατανομή χ^2 με $n-2$ βαθμούς ελευθερίας και δίνεται από

$$P\left(\chi^2_{1-\frac{\alpha}{2},n-2} \leq \frac{(n-2)\,\hat{\sigma}^2}{\sigma^2} \leq \chi^2_{\frac{\alpha}{2},n-2}\right) = 1-\alpha \Leftrightarrow$$

$$P\left(\frac{\chi^2_{1-\frac{\alpha}{2},n-2}}{(n-2)\,\hat{\sigma}^2} \leq \frac{1}{\sigma^2} \leq \frac{\chi^2_{\frac{\alpha}{2},n-2}}{(n-2)\,\hat{\sigma}^2}\right) = 1-\alpha \Leftrightarrow$$

$$P\left(\frac{(n-2)\,\hat{\sigma}^2}{\chi^2_{\frac{\alpha}{2},n-2}} \leq \sigma^2 \leq \frac{(n-2)\,\hat{\sigma}^2}{\chi^2_{1-\frac{\alpha}{2},n-2}}\right) = 1-\alpha$$

όπου $\chi^2_{1-\frac{\alpha}{2},n-2}$ και $\chi^2_{\frac{\alpha}{2},n-2}$ δηλώνουν τις κρίσιμες τιμές που βρίσκονται στους πίνακες της χ^2 κατανομής με $n-2$ βαθμούς ελευθερίας και επίπεδο σημαντικότητας $1-(\alpha/2)$ και $(\alpha/2)$ αντίστοιχα.

Για παράδειγμα, αν $\alpha = 5\%$, τότε δεξιά της κρίσιμης τιμής $\chi^2_{0.975,n-2}$ βρίσκεται το 97.5% της κατανομής ενώ δεξιά της κρίσιμης τιμής $\chi^2_{0.025,n-2}$ βρίσκεται το 2.5% της κατανομής.

3.1.1 Η τιμή πιθανότητας, p-τιμή

Σε οποιοδήποτε έλεγχο υπόθεσης και όταν έχουμε υπολογίσει την στατιστική ελέγχου υπάρχει και ένας **εναλλακτικός τρόπος προσέγγισης** στην απόφασή μας σχετικά με την απόρριψη ή μη απόρριψη της μηδενικής υπόθεσης H_0.

Η τιμή πιθανότητας, (συντομογραφία p-τιμή ή στην Αγγλική p-value), υποδηλώνει την πιθανότητα, κάτω από τη μηδενική υπόθεση (δηλαδή όταν η μηδενική υπόθεση H_0 είναι αληθής), να λάβουμε/υπολογίσουμε μία στατιστική ελέγχου η οποία είναι ίση ή πιό ακραία από την ευρεθείσα τιμή της στατιστικής ελέγχου.

Προσοχή, διότι συχνά γίνεται το λάθος να ερμηνεύουμε την p-τιμή ως την πιθανότητα ότι η μηδενική υπόθεση είναι αληθής. Επαναλαμβάνουμε λοιπόν, ότι η p-τιμή δίνει την πιθανότητα να λάβουμε ορισμένα αποτελέσματα (τιμές στατιστικής ελέγχου) εάν η μηδενική είναι αληθής/σωστή, και όχι την πιθανότητα ότι η μηδενική υπόθεση είναι αληθής.

Εάν η p-τιμή είναι **μικρότερη** από το επίπεδο σημαντικότητας α, τότε η μηδενική υπόθεση H_0 απορρίπτεται. Ο έλεγχος των p-τιμών επιτρέπει στους ερευνητές να φτάσουν σε συμπεράσματα χωρίς να συμβουλευτούν τις κατάλληλες κρίσιμες τιμές από τους σχετικούς πίνακες κατανομών, καθιστώντας τις p-τιμές μία αρκετά βολική πηγή πληροφορίας. Επιπλέον, η p-τιμή δείχνει την «ευαισθησία» της απόφασης απόρριψης της μηδενικής υπόθεσης ως προς την επιλογή του επιπέδου σημαντικότητας. Για παράδειγμα, μια p-τιμή ίση με 0,0789 υποδηλώνει ότι η μηδενική υπόθεση **απορρίπτεται** στο επίπεδο σημαντικότητας 10%, αλλά όχι στο επίπεδο 5% (και εννοείται όχι στο 1%).

3.1.2 Σύνοψη βασικών σημείων

- Η **τιμή πιθανότητας** p-τιμή για κάθε έναν από τους εκτιμημένους συντελεστές του υποδείγματος μας λέει αν μπορούμε να απορρίψουμε τη μηδενική υπόθεση ή όχι
- Υπενθυμίζουμε ότι αν η μηδενική υπόθεση θέτει την τιμή του συντελεστή ενδιαφέροντος να είναι μηδέν, τότε λέμε ότι ελέγχουμε για τη **στατιστική σημαντικότητα** ή όχι του εκτιμητή.
- Το **τυπικό σφάλμα του εκτιμητή** εκτιμά την τυπική απόκλιση της δειγματοληπτικής κατανομής των εκτιμημένων συντελεστών στο υπόδειγμά μας. Μπορούμε να σκεφτούμε το τυπικό σφάλμα ως το μέτρο «ακρίβειας» για τους εκτιμημένους συντελεστές.
- Η **στατιστική t (t-λόγος στο gretl)** προκύπτει διαιρώντας τους εκτιμημένους συντελεστές με το τυπικό τους σφάλμα.

Παρότι, οι στατιστικές ελέγχου θεωρητικά κατανέμονται σύμφωνα με κάποια γνωστή (πινακοποιημένη) κατανομή πιθανότητας, είναι πλέον εξαιρετικά χρονοβόρο αλλά και σχετικά ανακριβές να καταφεύγουμε στους στατιστικούς πίνακες των κατανομών για την εύρεση της κατάλληλης κρίσιμης τιμής. Οι υπολογιστές και τα εξειδικευμένα λογισμικά καθιστούν εξαιρετικά πιο γρήγορο τον υπολογισμό **τιμών πιθανότητας** (p-values) απευθείας.

Οι πίνακες αναφοράς αποτελεσμάτων ή στατιστικών στα σύγχρονα λογισμικά πακέτα χρησιμοποιούν ένα κοινό εργαλείο αναφοράς για να δηλώσουν εάν μια στατιστική (π.χ. ένας εκτιμημένος συντελεστής) είναι στατιστικά σημαντική σε

ένα δεδομένο επίπεδο σημαντικότητας α. **Συγκεκριμένα:**

- Τρία αστέρια (ή αστερίσκοι) *** υποδηλώνουν ότι ο εκτιμημένος συντελεστής ή στατιστική είναι σημαντική στο **επίπεδο σημαντικότητας** $\alpha = 0.01$ **ή 1%**
- Δύο αστέρια (ή αστερίσκοι) ** υποδηλώνουν ότι ο εκτιμημένος συντελεστής ή στατιστική είναι σημαντική στο **επίπεδο σημαντικότητας** $\alpha = 0.05$ **ή 5%**
- Ένα αστέρι (ή αστερίσκος) * υποδηλώνει ότι ο εκτιμημένος συντελεστής ή στατιστική είναι σημαντική στο **επίπεδο σημαντικότητας** $\alpha = 0.10$ **ή 10%**

Τέλος να κλείσουμε την υποενότητα προτείνοντας τη μελέτη του άρθρου των Wasserstein, Schirm & Lazar (2019)[6] ώστε να κατατοπιστούμε στην τρέχουσα επιστημονική άποψη ή προβληματισμό σχετικά με το ρόλο των τιμών πιθανότητας στη στατιστική επαγωγή.

3.1.3 Παράδειγμα

Έστω ότι έχουμε στη διάθεσή μας 340 μηνιαίες παρατηρήσεις από τον Ιανουάριο του 1985 μέχρι τον Απρίλιο του 2013 για έναν δείκτη με έτος βάσης[7] το 2005=100, ο οποίος συμπυκνώνει πληροφόρηση για τις τιμές των μετοχών του χρηματιστηρίου αξιών Αθηνών καθώς και έναν αντίστοιχο δείκτη για τη χρηματαγορά (συγκεκριμένα τη χρηματιστηριακή αγορά) των Η.Π.Α. Τα στοιχεία είναι διαθέσιμα στο αρχείο kefalaio3data1.xlsx και καλείστε να επαναλάβετε την άσκηση είτε στο Excel είτε - **προτιμότερο** - στο gretl είτε σε οποιοδήποτε λογισμικό μπορεί να προβεί σε ανάλυση παλινδρόμησης. Θέλουμε να διερευνήσουμε εμπειρικά μία αιτιώδη σχέση σύμφωνα με την οποία οι μηνιαίες ποσοστιαίες (επί τοις εκατό) μεταβολές των τιμών των μετοχών στην Ελλάδα επηρεάζονται από τις αντίστοιχες μηνιαίες μεταβολές στην αγορά των Η.Π.Α. Οι μηνιαίες % μεταβολές

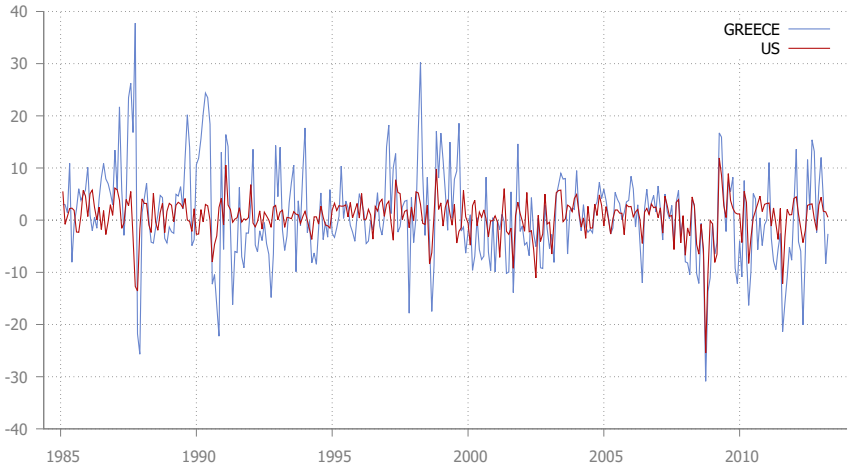
[6] Moving to a World Beyond "$p < 0.05$". Ronald L. Wasserstein, Allen L. Schirm & Nicole A. Lazar, Pages 1-19, `https://doi.org/10.1080/00031305.2019.1583913`

[7] Έτσι, μεταβολές από την τιμή του έτους βάσης αντιστοιχούν σε ποσοστιαίες μεταβολές. Για παράδειγμα, τον Απρίλιο του 2007 η τιμή του δείκτη p_t^{GR} είναι ίση με 151.25 που σημαίνει ότι οι τιμές των μετοχών στην Αθήνα έχουν αυξηθεί κατά περίπου 51.25% από το μέσο όρο του έτους 2005. Τα δεδομένα παρέχονται από τον Ο.Ο.Σ.Α (OECD) στην ιστοσελίδα της βάσης δεδομένων του οργανισμού.

του Ελληνικού δείκτη υπολογίζονται από τον τύπο

$$r_t^{GR} = 100 \times \ln\left(\frac{p_t^{GR}}{p_{t-1}^{GR}}\right) \approx 100 \times \left(\frac{p_t^{GR} - p_{t-1}^{GR}}{p_{t-1}^{GR}}\right)$$

όπου p_t^{GR} είναι το επίπεδο του δείκτη τη χρονική περίοδο t. Αντίστοιχα υπολογίστε τις αποδόσεις r_t^{USA}. Το παρακάτω γράφημα (3.2) απεικονίζει τις χρονοσειρές ενδιαφέροντος r_t^{GR}, r_t^{USA}.



Γράφημα 3.2: Μπλε απόχρωση: Ελληνικές αποδόσεις, Κόκκινη απόχρωση: Αποδόσεις Η.Π.Α

Θα διαπιστώσετε ότι το διαθέσιμο δείγμα αποδόσεων είναι $T = 339$ παρατηρήσεις αφού δεν μπορούμε να υπολογίσουμε την απόδοση από τον Δεκέμβριο του 1984 στον Ιανουάριο του 1985 καθώς δεν έχουμε στη διάθεσή μας την τιμή του δείκτη τον Δεκέμβριο του 1984 (και πιο πίσω βέβαια). Δηλαδή, στο θεωρητικό υπόδειγμα

$$r_t^{GR} = \alpha + \beta r_t^{USA} + u_t\,,\; t = 1, 2, ..., 339$$

$t = 1$ αντιστοιχεί στον μήνα Φεβρουάριο του 2013 και $t = 339$ αντιστοιχεί στον μήνα Απρίλιο του 2013. Η βασική υπόθεση προς διερεύνηση είναι

$$\begin{aligned} H_0: &\quad \beta = \quad 0 \\ H_1: &\quad \beta \neq \quad 0 \end{aligned}$$

δηλαδή ο **έλεγχος στατιστικής σημαντικότητας** του συντελεστή κλίσης.

Εκτιμούμε το υπόδειγμα με τη μέθοδο Ε.Τ, συγκεκριμένα εκτιμούμε τις παραμέτρους α, β, σ^2, όπου α: η σταθερά, β: ο συντελεστής της r_t^{USA} και σ^2: η διακύμανση του διαταρακτικού όρου και αναφέρουμε τα αποτελέσματα παρακάτω με έναν τυποποιημένο στην οικονομετρία τρόπο, δηλαδή αναφέρουμε τις εκτιμήσεις των συντελεστών του υποδείγματος, τις εκτιμήσεις των τυπικών τους σφαλμάτων και την εκτίμηση της τυπικής απόκλισης του διαταρακτικού όρου

$$r_t^{GR} = \underset{(0.439)}{0.233} + \underset{(0.115)}{1.072} r_t^{USA} + \hat{u}_t$$

$$\hat{\sigma} = 7.980$$

Προσοχή: σε παρενθέσεις αναφέρουμε τα εκτιμημένα τυπικά σφάλματα των εκτιμητών, άρα

$$\hat{\alpha} = 0.233\,,\ \widehat{se}\,(\hat{\alpha}) = 0.439$$

$$\hat{\beta} = 1.072\,,\ \widehat{se}\left(\hat{\beta}\right) = 0.115$$

$$\hat{\sigma} = 7.980\,,\ \widehat{se}\,(\hat{\sigma}) = 4.906$$

όπου για παράδειγμα

$$\widehat{se}\,(\hat{\sigma}) = \sqrt{\widehat{Var}\,(\hat{\sigma}^2)} = \sqrt{\frac{2\left(\hat{\sigma}^2\right)^2}{n-2}} = \sqrt{\frac{2\left(7.98^2\right)^2}{337}}$$

Η στατιστική t-student για **έλεγχο σημαντικότητας** του εκτιμημένου συντελεστή κλίσης δίνει

$$t_{\hat{\beta}} = \frac{1.072 - 0}{0.115} = \frac{1.072}{0.115} = 9.321$$

και επειδή $t_{\hat{\beta}} \sim t_{n-2}$ και το μέγεθος του δείγματος είναι 339 παρατηρήσεις βρίσκουμε την κρίσιμη τιμή της τυχαίας μεταβλητής t_{337} για επίπεδο σημαντικότητας 5% και συγκρίνουμε. Ευτυχώς, για δείγματα από περίπου 30 παρατηρήσεις και πάνω, η συγκεκριμένη κρίσιμη τιμή είναι περίπου 2 οπότε δεν χρειάζεται να

καταφύγουμε στους στατιστικούς πίνακες[8]. Προβαίνουμε στη σύγκριση

$$\left|t_{\hat{\beta}}\right| = |9.321| > 2$$

άρα **απορρίπτουμε** τη μηδενική υπόθεση $H_0 : \beta = 0$, δηλαδή ο συντελεστής $\hat{\beta}$ είναι στατιστικά σημαντικός ή στατιστικά διάφορος του μηδενός και εκτιμά μία παράμετρο που είναι διαφορετική του μηδενός. Από οικονομικής πλευράς συμπεραίνουμε ότι οι μηνιαίες αποδόσεις της χρηματιστηριακής αγοράς των Η.Π.Α όντως έχουν επίδραση στη χρηματιστηριακή αγορά της Αθήνας. Συγκεκριμένα, αν ποσοτικοποιήσουμε τα ευρήματά μας, μία μηνιαία αύξηση (μείωση) στις αποδόσεις των Η.Π.Α κατά μία μονάδα (δηλαδή κατά 1%) αντιστοιχεί κατά μέσο όρο σε περίπου 1.072 μονάδες (%) αύξηση (μείωση) της μηνιαίας απόδοσης στην Ελληνική χρηματαγορά.

Αν θελήσουμε να ελέγξουμε την υπόθεση ότι οι μεταβολές των αποδόσεων στην Ελληνική αγορά είναι 1-προς-1 ως προς τις μεταβολές της αγοράς των Η.Π.Α, δηλαδή να ελέγξουμε την υπόθεση

$$\begin{aligned} H_0 : &\quad \beta = \quad 1 \\ H_1 : &\quad \beta \neq \quad 1 \end{aligned}$$

τότε

$$t_{\hat{\beta}} = \frac{1.072 - 1}{0.115} = \frac{0.072}{0.115} = 0.626$$

οπότε $\left|t_{\hat{\beta}}\right| = |0.626| < 2$ και **δεν απορρίπτουμε** τη μηδενική υπόθεση $H_0 : \beta = 1$ σε επίπεδο σημαντικότητας $\alpha = 5\%$. Δηλαδή, όντως κάποιος θα μπορούσε να ισχυριστεί ότι οι μεταβολές στις αποδόσεις είναι 1-προς-1.

Το 95% διάστημα εμπιστοσύνης του συντελεστή β δίνεται από

$$P\left(-t_{n-2}^{\alpha/2} \leq \frac{\hat{\beta} - \beta}{\widehat{se}\left(\hat{\beta}\right)} \leq t_{n-2}^{\alpha/2}\right) = 0.95 \Rightarrow$$

$$P\left(\hat{\beta} - t_{n-2}^{\alpha/2} \cdot \widehat{se}\left(\hat{\beta}\right) \leq \beta \leq \hat{\beta} + t_{n-2}^{\alpha/2} \cdot \widehat{se}\left(\hat{\beta}\right)\right) = 0.95 \Rightarrow$$

$$P\left(1.072 - 2 \cdot 0.115 \leq \beta \leq 1.072 + 2 \cdot 0.115\right) = 0.95 \Rightarrow$$

[8]Επιπλέον, τα περισσότερα εξειδικευμένα λογισμικά παρέχουν είτε τις κρίσιμες τιμές είτε τις αντίστοιχες τιμές πιθανότητας άμεσα.

$$P\left(0.842 \leq \beta \leq 1.302\right) = 0.95$$

δηλαδή με πιθανότητα 95%, η άγνωστη τιμή του συντελεστή β βρίσκεται στο διάστημα (0.842 , 1.302).

Η εκτίμηση του τυπικού σφάλματος της παλινδρόμησης είναι $\hat{\sigma} = 7.980$, δηλαδή το μεγαλύτερο μέρος της μεταβλητότητας στη χρηματιστηριακή αγορά της Αθήνας δεν μπορεί να ερμηνευτεί με βάση τη σύγχρονη μηνιαία μεταβλητότητα της χρηματαγοράς των Η.Π.Α. Αυτό διότι η εκτίμηση της τυπικής απόκλισης της εξαρτημένης μεταβλητής είναι ίση με περίπου $\hat{\sigma}_y = 8.931$, άρα η χρήση των αποδόσεων της αγοράς των Η.Π.Α ναι μεν δείχνει κάποια ταυτόχρονη συσχέτιση και επιβεβαιώνει τυχόν θεωρία που θα ισχυριζόταν την αιτιώδη σχέση

$$r^{GR} = f\left(r^{USA}\right) \quad , \quad f' > 0 \quad , \quad f' \approx 1$$

όμως - από την άλλη - η αυτόνομη ή ιδιοσυγκρασιακή μεταβλητότητα του χρηματιστηρίου αξιών της Αθήνας τεκμηριώνεται εμπειρικά μιάς και μεγάλο μέρος της μεταβλητότητάς της - περίπου το 80% αφού $\frac{\hat{\sigma}^2}{\hat{\sigma}_y^2} = \frac{(7.980)^2}{(8.931)^2} = 0.798$ - δεν μπορεί να αποδοθεί στη μεταβλητότητα της αγοράς των Η.Π.Α.

3.1.4 Παράδειγμα

Ένα μαθηματικά εκφρασμένο υπόδειγμα που υπονοεί αιτιώδη σχέση και θετική εξάρτηση μισθού από την εκπαίδευση είναι το παρακάτω

$$\text{μισθός} = f\left(\text{εκπαίδευση}\right) \quad , \quad f' > 0$$

Έστω ότι έχουμε στη διάθεσή μας ένα τυχαίο δείγμα 1500 εργαζομένων από την Ελλάδα για τους οποίους γνωρίζουμε το πραγματικό τους ωρομίσθιο w_i και τα έτη εκπαίδευσης τους E_i. Ένα απλό γραμμικό υπόδειγμα που μπορούμε να εκτιμήσουμε λαμβάνει τη μορφή

$$w_i = \alpha + \beta E_i + u_i$$

όπου όλοι οι υπόλοιποι παράγοντες που μπορεί να επηρεάζουν τον μισθό (όπως ηλικία, εμπειρία ή προϋπηρεσία, μονιμότητα, ιδιοσυγκρασιακά χαρακτηριστικά του φορέα απασχόλησης, ιδιοσυγκρασιακά (ατομικά) χαρακτηριστικά ετερογένειας του εργαζόμενου κτλ.) εσωκλείονται στον διαταρακτικό όρο u_i, $i = 1, ..., n$ όπου

$n = 1500$. Για το παράδειγμά μας, θεωρουμε ότι όλες οι κλασσικές υποθέσεις ισχύουν, άρα και ότι τα παρατηρούμενα έτη εκπαίδευσης E_i (η εκπαίδευση δηλαδή) είναι τουλάχιστον ασυσχέτιστα με τους διαταρακτικούς όρους[9]. Τα δεδομένα είναι διαθέσιμα στο αρχείο

kefalaio3data2.xlsx

το οποίο καλείστε να μεταφέρετε είτε στο gretl είτε σε οποιοδήποτε λογισμικό μπορεί να προβεί σε ανάλυση παλινδρόμησης (είτε να εργαστείτε στο Excel) και να επαναλάβετε την άσκηση (**χρησιμοποιήστε κατά προτίμηση το** gretl).

Θέλουμε να απαντήσουμε στα παρακάτω ερωτήματα

- έχει επίδραση η εκπαίδευση στο πραγματικό ωρομίσθιο;
- αν ναι, είναι θετική και τι έντασης; (ποσοτικοποίηση επίδρασης)
- πόσο παραπάνω κερδίζει κατά μέσο όρο ένας εργαζόμενος με 4 επιπλέον έτη εκπαίδευσης; (αν όντως υπάρχει θετική επίδραση)
- ποια η σημασία του σταθερού όρου α στην παραπάνω παλινδρόμηση;
- το **τυπικό σφάλμα της παλινδρόμησης** εκφράζει τη μεταβλητότητα στην εξαρτημένη μεταβλητή του πραγματικού ωρομισθίου που δεν εξηγείται από την εκπαίδευση. Πόση είναι αυτή η μη εξηγημένη μεταβλητότητα;

Τα αποτελέσματα της εκτίμησης ΕΤ του γραμμικού υποδείγματος δίνονται συνοπτικά παρακάτω.

$$w_i = \underset{(0.348)}{2.013} + \underset{(0.027)}{0.383} E_i + \hat{u}_i$$
$$\hat{\sigma} = 3.363$$

όπου σε παρενθέσεις αναφέρονται τα **τυπικά σφάλματα** των εκτιμητών.

Ο έλεγχος της μηδενικής υπόθεσης $H_0 : \beta = 0$ σε επίπεδο σημαντικότητας 5% δίνει

$$\left|t_{\hat{\beta}}\right| = \left|\frac{0.383}{0.027}\right| = |14.185| > 2$$

[9] Η συγκεκριμένη υπόθεση είναι μη ρεαλιστική καθώς ένας αριθμός ατομικών χαρακτηριστικών που βρίσκονται στον διαταρακτικό όρο u_i συσχετίζονται ή προκαλούν αιτιωδώς την παρατηρούμενη εκπαίδευση.

άρα **απορρίπτουμε** τη μηδενική υπόθεση $H_0 : \beta = 0$, δηλαδή ο συντελεστής $\hat{\beta}$ είναι **στατιστικά σημαντικός ή στατιστικά διάφορος του μηδενός**.

ΣΗΜΕΙΩΣΗ: στην παρουσίαση των αποτελεσμάτων έχουμε προβεί σε περικοπή δεκαδικών ψηφίων πέραν του τρίτου, για συντομία και για λόγους παρουσίασης. Για παράδειγμα, αν επαναλάβετε την άσκηση με οποιοδήποτε λογισμικό τότε $\hat{\alpha} = 2.01317...$ με τυπικό σφάλμα $\widehat{se}(\hat{\alpha}) = 0.34817...$ και $\hat{\beta} = 0.38329...$ με τυπικό σφάλμα $\widehat{se}\left(\hat{\beta}\right) = 0.02704....$ Η στατιστική t-student για τη σημαντικότητα του $\hat{\beta}$ είναι ίση με

$$t_{\hat{\beta}} = \frac{0.38329...}{0.02704....} = 14.170542...$$

Όντως, με βάση το συγκεκριμένο δείγμα, η εκπαίδευση επηρεάζει το ύψος του πραγματικού ωρομισθίου. Μία σύντομη οικονομική ερμηνεία του αποτελέσματος είναι ότι παρατηρούμε μέση αύξηση του ωρομισθίου κατά 0.383€ ανά έτος εκπαίδευσης. Άρα, αν μεταφράσουμε το αποτέλεσμα σε ετήσια βάση, ένας πρόχειρος υπολογισμός με 8 ώρες την ημέρα, επί 22 εργάσιμες ανά μήνα επί 12 μήνες το έτος δίνει 2112 ώρες και αύξηση $0.383 \times 2112 = 808.89$€ το χρόνο για κάθε επιπλέον έτος εκπαίδευσης. Επίσης, η απάντηση στο ερώτημα, πόσο παραπάνω κερδίζει κατά μέσο όρο ένας εργαζόμενος με 4 επιπλέον έτη εκπαίδευσης είναι $0.383 \times 4 = 1.53$€ την ώρα ή 3235.6€ το έτος κ.ο.κ.

Παρατηρήστε πως το γραμμικό υπόδειγμα υπονοεί ότι μεταβολές στην εκπαίδευση κατά ένα έτος έχουν ίδια επίδραση στο ωρομίσθιο ανεξαρτήτως της αρχικής τιμής της εκπαίδευσης. Δηλαδή ένα επιπλέον έτος εκπαίδευσης υπονοεί μέση αύξηση του ωρομισθίου κατά 38.3 λεπτά είτε «ξεκινώντας» από 12 έτη εκπαίδευσης είτε από τα 16 έτη εκπαίδευσης κ.ο.κ. Αυτό είναι αποτέλεσμα της γραμμικότητας του υποδείγματος άρα είναι ένα αποτέλεσμα που έχουμε επιβάλλει. Σε επόμενες διαλέξεις, και κυρίως στο κεφάλαιο 4 θα δούμε πως μπορούμε να αντιμετωπίσουμε τέτοιες μη ρεαλιστικές επιβολές λόγω της γραμμικότητας του υποδείγματος. Επίσης, η ερμηνεία του σταθερού όρου στο συγκεκριμένο υπόδειγμα θα είναι ότι ένας εργαζόμενος χωρίς καθόλου εκπαίδευση, $E_i = 0$, κερδίζει κατά μέσο όρο 2.013€. Ο σταθερός όρος είναι επίσης στατιστικά σημαντικός αφού

$$|t_{\hat{\alpha}}| = \left|\frac{2.013}{0.348}\right| = |5.784| > 2$$

Να σημειώσουμε όμως ότι η ερμηνεία του σταθερού όρου **συνήθως δεν είναι**

άμεσης οικονομικής σημασίας και δεν χρειάζεται να μας προβληματίζει[10]. Ο κύριος λόγος εισαγωγής του σταθερού όρου α στα γραμμικά οικονομετρικά υποδείγματα θα γίνει πλήρως κατανοητός στο κεφάλαιο της πολλαπλής παλινδρόμησης και σχετίζεται με την ορθή μαθηματική διατύπωση και εκτίμηση του υποδείγματος και όχι με συγκεκριμένα οικονομικά ερωτήματα.

Σχετικά με το μέγεθος του τυπικού σφάλματος της παλινδρόμησης $\hat{\sigma} = 3.363€$ θα πρέπει να το συγκρίνουμε με την τυπική απόκλιση της εξαρτημένης μεταβλητής, η οποία εκτιμάται στην τιμή $\hat{\sigma}_w = 3.580€$. Παρατηρήστε ότι η θεωρητική μεταβλητότητα του ωρομισθίου, σύμφωνα με το υπόδειγμα είναι ίση με

$$Var\left(w\right) = \beta^2 Var\left(E\right) + Var\left(u\right)$$

όπου ο υποδείκτης i έχει παραληφθεί λόγω της υπόθεσης της σταθερής διακύμανσης των διαταρακτικών όρων και της ανεξάρτητης μεταβλητής, η οποία τώρα θεωρείται τυχαία μεταβλητή ασυσχέτιστη με το διαταρακτικό όρο. Το στοιχείο $\beta^2 Var\left(E\right)$ δείχνει τη μεταβλητότητα του πραγματικού ωρομισθίου που οφείλεται στην εκπαίδευση, ενώ η διακύμανση του διαταρακτικού όρου $Var\left(u\right)$ τη μεταβλητότητα που οφείλεται σε όλους τους άλλους παράγοντες. Καταφεύγοντας στις εκτιμήσεις του υποδείγματος, αφού

$$\frac{\widehat{Var}\left(u\right)}{\widehat{Var}\left(w\right)} = \frac{\hat{\sigma}^2}{\hat{\sigma}_w^2} = \frac{\left(3.363\right)^2}{\left(3.580\right)^2} = 0.882$$

συμπεραίνουμε ότι το 88.2% της μεταβλητότητας έχει μείνει ανερμήνευτη. Άρα το 88.2% της μεταβλητότητας του πραγματικού ωρομισθίου δεν ερμηνεύεται από την εκπαίδευση. Το αποτέλεσμα αυτό είναι παρόμοιο με το αποτέλεσμα που θα λαμβάναμε αν είχαμε υπολογίσει τον συντελεστή προσδιορισμού. Πράγματι, για το συγκεκριμένο δείγμα και υπόδειγμα έχουμε ότι $R^2 = 0.118$ δηλαδή το 11.8% της μεταβλητότητας του ωρομισθίου ερμηνεύεται από το υπόδειγμα (δηλαδή από την παρατηρούμενη εκπαίδευση).

[10] Για παράδειγμα, αν δεν είχαμε την εκπαίδευση ως ερμηνευτική μεταβλητή αλλά την ηλικία, τότε ποια είναι η οικονομική ερμηνεία της εκτιμημένης σταθεράς; Ότι για κάποια τιμή της ερμηνευτικής μεταβλητής (π.χ., ηλικία=0) που είναι εκτός του οικονομικά και λογικά εφικτού εύρους του δείγματος, το μέσο ωρομίσθιο είναι ίσο με κάποια συγκεκριμένη τιμή;

3.1.5 Παράδειγμα

Θα επαναλάβουμε την άσκηση του πρώτου παραδείγματος χρησιμοποιώντας δεδομένα για τη χρηματαγορά της Ιαπωνίας. Έστω λοιπόν ότι έχουμε στη διάθεσή μας τις μηνιαίες αποδόσεις ενός δείκτη της χρηματαγοράς της Ιαπωνίας. Έστω p_t^J το επίπεδο του δείκτη τη χρονική περίοδο t. Οι αποδόσεις κατασκευάζονται μέσω του τύπου

$$r_t^J = 100 \times \ln\left(\frac{p_t^J}{p_{t-1}^J}\right)$$

$$r_t^{USA} = 100 \times \ln\left(\frac{p_t^{USA}}{p_{t-1}^{USA}}\right)$$

Εκτιμήστε το απλό υπόδειγμα παλινδρόμησης

$$r_t^J = \alpha + \beta r_t^{USA} + u_t$$

το οποίο υποθέτει μία αιτιώδη σχέση από τις αποδόσεις της χρηματαγοράς των Η.Π.Α σε αυτές της Ιαπωνίας.

Θέλουμε να

- εκτιμήσουμε το β,
- να ελέγξουμε αν το $\hat{\beta}$ είναι διαφορετικό του μηδενός (δηλαδή αν υπάρχει επίδραση από την r_t^{USA} στην r_t^J)
- και στη συνέχεια, αν **απορρίψουμε** την υπόθεση $H_0 : \beta = 0$, να ποσοτικοποιήσουμε την επίδραση ερμηνεύοντας την εκτίμηση $\hat{\beta}$
- επίσης θέλουμε να ελέγξουμε την υπόθεση «οι αποδόσεις στην Ιαπωνία κινούνται ένα-προς-ένα με τις αποδόσεις των Η.Π.Α» ή αλλιώς ότι $\beta = 1$
- και να εκτιμήσουμε το 95% διάστημα εμπιστοσύνης για τη διακύμανση του διαταρακτικού όρου σ^2.

Αναλυτικά, η εκτίμηση ελαχίστων τετραγώνων δίνει

$$r_t^J = \underset{(0.2240)}{-0.4185} + \underset{(0.0588)}{0.7410} r_t^{USA} + \hat{u}_t \ , \ t = 1, ..., T = 339$$

$$\hat{\sigma} = 4.0671 \ , \ \hat{\sigma}^2 = 16.5417$$

$$\chi^2_{0.975,337} = 288.03 \ , \ \chi^2_{0.025,337} = 389.7$$

όπου παρουσιάζουμε το εκτιμημένο τυπικό σφάλμα της παλινδρόμησης $\hat{\sigma}$, την εκτίμηση Ε.Τ της διακύμανσης των διαταρακτικών όρων $\hat{\sigma}^2 = (\hat{\sigma})^2$ (ή διακύμανση των καταλοίπων) και τις κρίσιμες τιμές της κατανομής χ^2 που θα βοηθήσουν στην κατασκευή του διαστήματος εμπιστοσύνης για τη διακύμανση του διαταρακτικού όρου σ^2.

Ο έλεγχος

$$\begin{aligned} H_0: &\quad \beta = 0 \\ H_1: &\quad \beta \neq 0 \end{aligned}$$

στηρίζεται στην t-student στατιστική

$$t_{\hat{\beta}} = \frac{0.7410}{0.0588} = 12.602$$

Αφού $\left|t_{\hat{\beta}}\right| > 2$ **απορρίπτουμε** τη μηδενική υπόθεση περί στατιστικής μη - σημαντικότητας του εκτιμημένου συντελεστή κλίσης. Δηλαδή, σωστά εισέρχεται η μεταβλητή r_t^{USA} στο υπόδειγμα. Ποσοτικοποιώντας την επίδραση της r_t^{USA} στην r_t^J, μία αύξηση της r_t^{USA} κατά μία μονάδα (που αντιστοιχεί σε 1% αφού οι αποδόσεις μετρούνται επί τοις εκατό) αντιστοιχεί κατά μέσο όρο σε μία αύξηση των αποδόσεων στην Ιαπωνία κατά 0.741 μονάδες (δηλαδή 0.741%).

Ο έλεγχος για την υπόθεση

$$\begin{aligned} H_0: &\quad \beta = 1 \\ H_1: &\quad \beta \neq 1 \end{aligned}$$

στηρίζεται στη στατιστική t-student

$$t_{\hat{\beta}} = \frac{0.7410 - 1}{0.0588} = -4.408$$

Αφού $\left|t_{\hat{\beta}}\right| = 4.408 > 2$ απορρίπτουμε τη μηδενική υπόθεση $\beta = 1$ ή αλλιώς την υπόθεση ότι «οι αποδόσεις στην Ιαπωνία κινούνται ένα-προς-ένα με τις αποδόσεις των Η.Π.Α».

Το 95% διάστημα εμπιστοσύνης της διακύμανσης σ^2 δίνεται μέσω του τύπου

$$P\left(\frac{(n-2)\,\hat{\sigma}^2}{\chi^2_{\frac{a}{2},n-2}} \leq \sigma^2 \leq \frac{(n-2)\,\hat{\sigma}^2}{\chi^2_{1-\frac{a}{2},n-2}}\right) = 1 - a \Rightarrow$$

$$P\left(\frac{337\times 16.5417}{\chi^2_{0.025,n-2}}\leq\sigma^2\leq\frac{337\times 16.5417}{\chi^2_{0.975,n-2}}\right)=0.95\Rightarrow$$

$$P\left(\frac{337\times 16.5417}{389.7}\leq\sigma^2\leq\frac{337\times 16.5417}{288.03}\right)=0.95\Rightarrow$$

$$P\left(14.305\leq\sigma^2\leq 19.354\right)=0.95$$

Άρα με πιθανότητα 95%, η άγνωστη τιμή της διακύμανσης σ^2 βρίσκεται στο διάστημα $\sigma^2\in(14.305\ ,\ 19.354)$.

Το διάστημα εμπιστοσύνης της τυπικής απόκλισης σ δίνεται από την τετραγωνική ρίζα των ορίων του παραπάνω διαστήματος, συγκεκριμένα με πιθανότητα 95% η άγνωστη τιμή σ βρίσκεται στο διάστημα $(3.782\ ,\ 4.399)$.

3.2 Πρόβλεψη με το απλό γραμμικό υπόδειγμα

Μία από τις βασικότερες εμπειρικές χρήσεις της Οικονομετρίας εδράζεται στη χρήση του υποδείγματος για την πρόβλεψη της εξαρτημένης μεταβλητής. Ουσιαστικά, οι προσαρμοσμένες τιμές $\hat{Y}_i$ του υποδείγματος

$$\hat{Y}_i=\hat{\alpha}+\hat{\beta}X_i$$

αποτελούν την πρόβλεψη του υποδείγματος και της μεθόδου εκτίμησης για την εξαρτημένη μεταβλητή και για δεδομένες (δειγματικές) τιμές της ανεξάρτητης μεταβλητής X_i. Οι συγκεκριμένες τιμές, $\hat{Y}_i$, $i=1,...,n$ ονομάζονται **προσαρμοσμένες τιμές ή προβλέψεις**. Όταν όμως θεωρήσουμε τιμές της X οι οποίες είναι εκτός των τιμών του δείγματος ή ειδικότερα με δεδομένα χρονοσειρών, εκτός του χρονικού εύρους που καλύπτει το δείγμα, τότε η παραγόμενη τιμή $\hat{Y}$ της Y ονομάζεται μόνο **πρόβλεψη** με βάση το εκτιμημένο υπόδειγμα.

Έστω μία νέα (δοσμένη ή υποτιθέμενη) τιμή X_0 ή X_f της μεταβλητής X_i, $i=1,...,n$ όπου ο υποδείκτης «μηδέν» ή «f» συμβολίζει ότι η τιμή είναι δοσμένη και μπορεί να βρίσκεται εντός ή εκτός του εύρους του δείγματος.

Στην ενδιαφέρουσα περίπτωση των χρονοσειρών, που υπονοείται αυστηρή χρονική διάταξη των παρατηρήσεων, αντί του υποδείκτη i χρησιμοποιούμε τον χρόνο $t=1,...,T$ και όταν η δοσμένη ή υποτιθέμενη τιμή X_f βρίσκεται εκτός

του χρονικού εύρους του δείγματος, τότε δίνουμε το συμβολισμό X_{T+1} ή X_{T+2} κ.ο.κ. για να συμβολίσουμε τιμές της ερμηνευτικής μεταβλητής πέραν του ορίου του δείγματος.

Με βάση την εκτίμηση του υποδείγματος και τη δοσμένη τιμή X_0 είμαστε σε θέση να «προβλέψουμε» την εξαρτημένη μεταβλητή Y. Συγκεκριμένα, η **σημειακή πρόβλεψη** (point forecast) δίνεται από

$$\hat{Y}_0 = \hat{\alpha} + \hat{\beta} X_0 \tag{3.5}$$

ή όταν έχουμε δεδομένα χρονοσειρών και προβλέπουμε εκτός του δείγματος (out-of-sample forecasting) από την ή τις

$$\begin{aligned} \hat{Y}_{T+1} &= \hat{\alpha} + \hat{\beta} X_{T+1} \\ \hat{Y}_{T+2} &= \hat{\alpha} + \hat{\beta} X_{T+2} \\ &\vdots \\ \hat{Y}_{T+k} &= \hat{\alpha} + \hat{\beta} X_{T+k} \end{aligned}$$

όπου k ονομάζεται **ορίζοντας πρόβλεψης**. Για παράδειγμα, η τιμή $\hat{Y}_{T+1}$ αντιστοιχεί στην πρόβλεψη μία περιόδο εμπρός, η τιμή $\hat{Y}_{T+2}$ στην πρόβλεψη δύο περιόδους εμπρός κτλ. Γίνεται κατανοητό ότι η πρόβλεψη **υποθέτει σταθερότητα του υποδείγματος την περίοδο πρόβλεψης**, δηλαδή οι παράμετροι α, β, σ^2 θεωρούνται αμετάβλητες στον χρόνο $T+1, T+2, ...$

Το σφάλμα πρόβλεψης

$$\begin{aligned} e_0 &= Y_0 - \hat{Y}_0 \\ &\text{ή} \\ e_{T+k} &= Y_{T+k} - \hat{Y}_{T+k} \end{aligned}$$

ορίζεται ως η απόκλιση της «πραγματικής» τιμής Y_0 που θα λάβει η εξαρτημένη μεταβλητή από την πρόβλεψή της $\hat{Y}_0$ με βάση το υπόδειγμα αναφοράς. Παρακάτω, θα προβούμε σε μία ανάλυση του σφάλματος πρόβλεψης e_0 ώστε να υπολογίσουμε εύκολα τις δύο πρώτες ροπές του e_0 και κατόπιν την κατανομή του. Έτσι θα προβούμε σε στατιστική επαγωγή για το σφάλμα πρόβλεψης, η οποία στην προκειμένη περίπτωση αναφέρεται στην κατασκευή ενός **διαστήματος εμπιστοσύνης της πρόβλεψης**.

Το σφάλμα πρόβλεψης e_0 αναλύεται σε δύο στοχαστικούς παράγοντες, (α) το σφάλμα δειγματοληψίας του εκτιμητή ΕΤ και (β) μία συνάρτηση που εξαρτάται αποκλειστικά από τους διαταρακτικούς όρους, ως εξής

$$\begin{aligned} e_0 &= Y_0 - \hat{Y}_0 \\ &= \alpha + \beta X_0 + u_0 - \left(\hat{\alpha} + \hat{\beta} X_0\right) \\ &= \alpha + \beta X_0 + u_0 - \hat{\alpha} - \hat{\beta} X_0 \\ &= \alpha + \beta X_0 + u_0 - \left(\bar{Y} - \hat{\beta}\bar{X}\right) - \hat{\beta} X_0 \\ &= \alpha + \beta X_0 + u_0 - \bar{Y} + \hat{\beta}\bar{X} - \hat{\beta} X_0 \\ &= \alpha + \beta X_0 + u_0 - \left(\alpha + \beta\bar{X} + \bar{u}\right) + \hat{\beta}\bar{X} - \hat{\beta} X_0 \\ &= \beta X_0 - \beta\bar{X} + \hat{\beta}\bar{X} - \hat{\beta} X_0 + u_0 - \bar{u} \\ &= \beta\left(X_0 - \bar{X}\right) - \hat{\beta}\left(X_0 - \bar{X}\right) + u_0 - \bar{u} \\ &= \beta x_0 - \hat{\beta} x_0 + (u_0 - \bar{u}) \\ &= -\left(\hat{\beta} - \beta\right) x_0 + (u_0 - \bar{u}) \end{aligned}$$

όπου προβήκαμε κατά σειρά στην αντικατάσταση της $\hat{Y}_0$, της $\hat{\alpha}$ και της $\bar{Y}$. Με βάση την υπόθεση ότι x_0 μη στοχαστική μεταβλητή, το μέσο σφάλμα πρόβλεψης είναι μηδενικό $E(e_0) = 0$ όταν ο εκτιμητής ΕΤ είναι αμερόληπτος, δηλαδή όταν $E\left(\hat{\beta} - \beta\right) = 0$ αποτέλεσμα που βασίζεται στην υπόθεση $E(u_i) = 0$, $\forall i$, ενώ αποδεικνύεται με χρήση των υποθέσεων $Var(u_i) = \sigma^2$, $Cov(u_i, u_j) = 0$, $\forall i \neq j$ και τα αποτελέσματα[11] της άσκησης 9 του κεφαλαίου 2 ότι

$$\begin{aligned} Var(e_0) &= Var\left(-\left(\hat{\beta} - \beta\right) x_0 + (u_0 - \bar{u})\right) \\ &= x_0^2 Var\left(\hat{\beta} - \beta\right) + Var(u_0 - \bar{u}) \\ &= \frac{x_0^2 \sigma^2}{\sum_{i=1}^{n} x_i^2} + \sigma^2\left(1 + \frac{1}{n}\right) \end{aligned}$$

[11] $Cov\left(\hat{\beta} - \beta, u_0\right) = 0$ και $Cov\left(\hat{\beta} - \beta, \bar{u}\right) = 0$

$$= \sigma^2 \left(1 + \frac{1}{n} + \frac{x_0^2}{\sum_{i=1}^{n} x_i^2} \right)$$

Έχοντας βρει την αναμενόμενη τιμή $E(e_0) = 0$ και διακύμανση $Var(e_0)$ του **σφάλματος πρόβλεψης** e_0, μένει να βρούμε την κατανομή του, ώστε να προβούμε σε στατιστική επαγωγή.

Είναι «εμφανές» ότι το σφάλμα πρόβλεψης e_0 ακολουθεί την κανονική κατανομή αφού με βάση τις κλασσικές υποθέσεις ο διαταρακτικός όρος και ο εκτιμητής ΕΤ κατανέμονται κανονικά και το σφάλμα πρόβλεψης είναι γραμμικός συνδυασμός των $\left(\hat{\beta} - \beta\right)$, u_i και u_0. Άρα

$$e_0 \sim N\left(E\left(e_0\right), Var\left(e_0\right)\right)$$

ή αναλυτικά

$$e_0 \sim N\left(0, \sigma^2 \left(1 + \frac{1}{n} + \frac{x_0^2}{\sum_{i=1}^{n} x_i^2} \right)\right)$$

Αντικαθιστώντας τη διακύμανση $Var\left(e_0\right)$ με την εκτιμημένη διακύμανση $\widehat{Var}\left(e_0\right)$ (η οποία χρησιμοποιεί $\hat{\sigma}^2$ αντί σ^2), το σφάλμα πρόβλεψης κατανέμεται ως μία t-student τυχαία μεταβλητή με $n-2$ βαθμούς ελευθερίας, δηλαδή

$$t = \frac{e_0}{\widehat{se}\left(e_0\right)} = \frac{Y_0 - \hat{Y}_0}{\widehat{se}\left(e_0\right)} \sim t_{n-2}$$

όπου $\widehat{se}\left(e_0\right) = \sqrt{\widehat{Var}\left(e_0\right)}$. Συνεπώς, το $(1-\alpha)\%$ διάστημα εμπιστοσύνης πρόβλεψης για την «πραγματική» τιμή Y_0 δίνεται από τον τύπο

$$\hat{Y}_0 \pm t_{n-2}^{\frac{\alpha}{2}} \widehat{se}\left(e_0\right)$$

όπου $t_{n-2}^{\frac{\alpha}{2}}$ η κρίσιμη τιμή της t-student κατανομής με $n-2$ βαθμούς ελευθερίας (κοντά στην τιμή 2 για $n \geq 30$ και $\alpha = 5\%$). Στην εμπειρική ανάλυση είναι σχεδόν βέβαιο ότι πρέπει να υιοθετούμε διαστήματα εμπιστοσύνης της πρόβλεψης ούτως ώστε να παρουσιάζουμε ένα εύρος πιθανών τιμών γύρω από τη σημειακή πρόβλεψη.

3.2.1 Παράδειγμα

Θα χρησιμοποιήσουμε το υπόδειγμα που σχετίζει αιτιωδώς τις αποδόσεις στην Ελληνική χρηματιστηριακή αγορά με τις αποδόσεις τις χρηματιστηριακής αγοράς των Η.Π.Α ώστε να προβούμε σε πρόβλεψη της απόδοσης στην Ελληνική αγορά τους επόμενους 2 μήνες. Έστω ότι βρισκόμαστε στις αρχές Μαΐου του 2013 και έχουμε στη διάθεσή μας 340 μηνιαίες παρατηρήσεις (τιμές) για έναν γενικό δείκτη της Ελληνικής αγοράς p_t^{GR} και της αγοράς των Η.Π.Α p_t^{USA} από τον Ιανουάριο του 1985 μέχρι και τον Απρίλιο του 2013[12]. Θέλουμε να προβλέψουμε την απόδοση της Ελληνικής χρηματιστηριακής αγοράς τους μήνες Μάϊο και Ιούνιο του 2013. Κατασκευάζουμε τις μηνιαίες επί τοις εκατό αποδόσεις r_t^{GR} και r_t^{USA} χρησιμοποιώντας τους τύπους

$$r_t^{GR} = 100 \times \ln\left(\frac{p_t^{GR}}{p_{t-1}^{GR}}\right)$$

$$r_t^{USA} = 100 \times \ln\left(\frac{p_t^{USA}}{p_{t-1}^{USA}}\right)$$

Το δείγμα των αποδόσεων έχει μέγεθος $T = 339$ παρατηρήσεων αφού «χάνουμε» μία παρατήρηση στην αρχή του δείγματος.

Εκτιμούμε όπως πριν το γραμμικό υπόδειγμα

$$r_t^{GR} = \alpha + \beta r_t^{USA} + u_t \quad , \quad t = 1, ..., 339$$

για το δείγμα των 339 παρατηρήσεων (Φεβ-1985 - Απρ-2013) και λαμβάνουμε τα αποτελέσματα

$$r_t^{GR} = \underset{(0.439)}{0.233} + \underset{(0.115)}{1.072} r_t^{USA} + \hat{u}_t$$

$$\hat{\sigma} = 7.980$$

όπου σε παρενθέσεις αναφέρονται τυπικά σφάλματα των εκτιμητών. Επίσης, η δειγματική μέση απόδοση για την αγορά των Η.Π.Α είναι ίση με $\bar{r}^{USA} = 0.6373$ ενώ το **άθροισμα των τετραγωνικών αποκλίσεων** των αποδόσεων r_t^{USA}

[12]Πρόκειται για στοιχεία που βρίσκονται στο αρχείο kefalaio3data1.xlsx

από τον δειγματικό τους μέσο $\bar{r}^{USA}$ είναι ίσο με

$$\sum_{t=1}^{339}\left(r_t^{USA}-\bar{r}^{USA}\right)^2 = 4778.5639$$

Επειδή τα δεδομένα μας αντιστοιχούν σε χρονοσειρές αντί του υποδείκτη f ή 0 θα υιοθετήσουμε τον υποδείκτη $T+1$ και $T+2$ για την πρόβλεψη ένα και δύο μήνες στο μέλλον αντίστοιχα (ορίζοντας πρόβλεψης $k=1$ και $k=2$). Σύμφωνα με το υπόδειγμα, και υποθέτοντας ότι δεν πρόκειται να επέλθει μεταβολή στην επικρατούσα σχέση έχουμε τις σημειακές προβλέψεις

$$\hat{r}_{T+1}^{GR} = 0.233 + 1.072 r_{T+1}^{USA}$$

και

$$\hat{r}_{T+2}^{GR} = 0.233 + 1.072 r_{T+2}^{USA}$$

Είναι εμφανές ότι όταν το υπόδειγμα δεν έχει δυναμική διάσταση (διαφορές στη χρονική αντίδραση των μεταβλητών) τότε πρέπει να προβούμε σε εικασίες σχετικά με την απόδοση στις Η.Π.Α τις περιόδους $T+1$ και $T+2$ ώστε να προβλέψουμε με βάση το εκτιμημένο υπόδειγμα (και το ιστορικό δείγμα που κατέχουμε) την τιμή $\hat{r}_{T+1}^{GR}$ και $\hat{r}_{T+2}^{GR}$.

Ας υποθέσουμε λοιπόν ότι τον μήνα Μάιο η απόδοση r_{T+1}^{USA} θα είναι αρνητική και ίση με -0.5 ενώ τον Ιούνιο θα εξακολουθήσει να είναι αρνητική η απόδοση αλλά η πτωτική πορεία παρουσιάζει ανάσχεση με $r_{T+2}^{USA} = -0.25$ (απόδοση μείον 0.25%).

Τότε, η **σημειακή πρόβλεψη** για τις Ελληνικές αποδόσεις έναν ορίζοντα εμπρός είναι

$$\hat{r}_{T+1}^{GR} = 0.233 + 1.072\,(-0.5) = -0.303$$

και δύο ορίζοντες εμπρός είναι

$$\hat{r}_{T+2}^{GR} = 0.233 + 1.072\,(-0.25) = -0.035$$

Στην οικονομική, οι **σημειακές προβλέψεις αποτελούν ενδείξεις και συνηθίζεται να συμβουλευόμαστε διαστήματα εμπιστοσύνης παρά τις σημειακές προβλέψεις**. Αυτό αφορά ειδικά περιπτώσεις όπου το τυπικό σφάλμα της παλινδρόμησης είναι μεγάλο σε σχέση με την εκτιμημένη τυπική απόκλιση της εξαρτημένης μεταβλητής του υποδείγματος (όπως συμβαίνει

στο παράδειγμά μας).

Το 95% διάστημα εμπιστοσύνης της πρόβλεψης ενός ορίζοντα εμπρός δίνεται από

$$\hat{Y}_{T+1} \pm t_{n-2}^{0.05/2} \widehat{se}\left(e_{T+1}\right)$$

Επειδή το τυπικό σφάλμα της πρόβλεψης δίνεται από

$$\widehat{se}\left(e_{T+1}\right) = \sqrt{\hat{\sigma}^2 \left(1 + \frac{1}{n} + \frac{x_{T+1}^2}{\sum_{i=1}^{n} x_i^2}\right)} =$$

$$= \sqrt{\hat{\sigma}^2 \left(1 + \frac{1}{339} + \frac{\left(-0.5 - \bar{r}^{USA}\right)^2}{\sum_{t=1}^{339} \left(r_t^{USA} - \bar{r}^{USA}\right)^2}\right)}$$

$$= 7.980 \sqrt{\left(1 + \frac{1}{339} + \frac{\left(-0.5 - 0.6373\right)^2}{4778.5639}\right)}$$

$$= 7.9928$$

λαμβάνουμε το διάστημα εμπιστοσύνης

$$\hat{r}_{T+1}^{GR} \pm 2 \cdot 7.9928 = -0.303 \pm 2 \cdot 7.9928$$

δηλαδή με πιθανότητα 95% η απόδοση r_{T+1}^{GR} του Ελληνικού δείκτη τον μήνα Μάιο θα βρίσκεται στο διάστημα $(-16.289\%\,,\, 15.683\%)$. Είναι εμφανές ότι το διάστημα είναι πολύ μεγάλο (έχει μεγάλο εύρος) για να είναι χρήσιμο σε βραχυχρόνιες επενδυτικές στρατηγικές. Για παράδειγμα, η μέση μηνιαία απόδοση του Ελληνικού δείκτη στο δείγμα είναι $\bar{r}^{GR} = 0.9169\%$.

Παρομοίως, για τον μήνα Ιούνιο και με βάση την υπόθεση ότι $r_{T+2}^{USA} = -0.25$ έχουμε

$$\widehat{se}(e_{T+2}) =$$

$$= 7.980\sqrt{\left(1+\frac{1}{339}+\frac{(-0.25-0.6373)^2}{4778.5639}\right)}$$

$$= 7.9924$$

δηλαδή με πιθανότητα 95% η απόδοση r_{T+2}^{GR} του Ελληνικού δείκτη τον μήνα Ιούνιο θα βρίσκεται στο διάστημα

$$\hat{r}_{T+2}^{GR} - 2 \cdot 7.9924 \leq r_{T+2}^{GR} \leq \hat{r}_{T+2}^{GR} + 2 \cdot 7.9924$$

ή $(-16.020\%\,,\, 15.950\%)$.

Παρατηρήστε ότι το σφάλμα πρόβλεψης

$$\widehat{se}\,(e_{T+2}) < \widehat{se}\,(e_{T+1})$$

αφού η προβλεπόμενη τιμή $r_{T+2}^{USA} = -0.25$ βρίσκεται πιο κοντά στη μέση τιμή $\bar{r}^{USA} = 0.6373$ από ότι η προβλεπόμενη τιμή $r_{T+1}^{USA} = -0.5$.

Επίσης, παρατηρήστε ότι το μικρότερο δυνατό σφάλμα πρόβλεψης το λαμβάνουμε αν υποθέσουμε ότι η μελλοντική απόδοση r_{T+1}^{USA} ή r_{T+2}^{USA} είναι ίση με τη δειγματική μέση τιμή $\bar{r}^{USA}$ αφού τότε

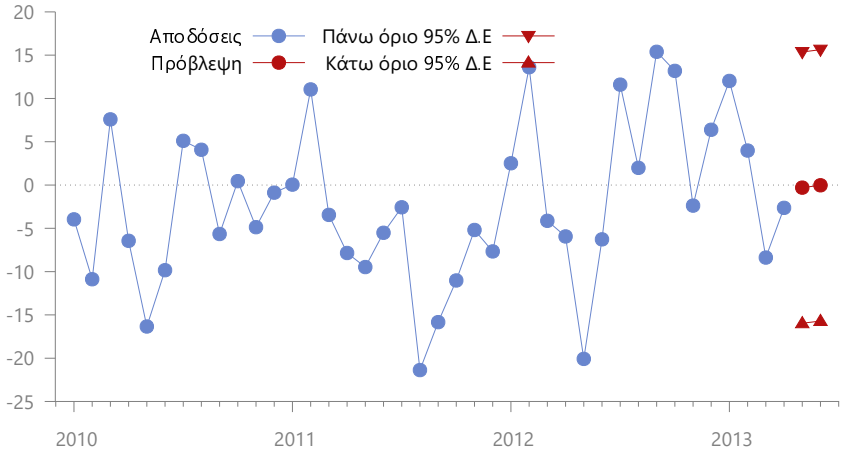
$$\widehat{se}\,(e_0) =$$

$$= 7.980\sqrt{\left(1+\frac{1}{339}\right)}$$

$$= 7.9918$$

Το παρακάτω γράφημα (3.3) παρουσιάζει τη χρονοσειρά r_t^{GR} για την περίοδο μετά τον Ιανουάριο του 2010 (μαύρη γραμμή με κύκλους), μαζί με τις προβέψεις $\hat{r}_{T+1}^{GR}$, $\hat{r}_{T+2}^{GR}$, (μπλε γραμμή με κύκλους) και τα κάτω όρια -16.289, -16.020 (μπλε γραμμή με τρίγωνα που δείχνουν προς τα επάνω) και πάνω όρια 15.683, 15.950 (μπλε τρίγωνα που δείχνουν προς τα κάτω) των διαστημάτων εμπιστοσύνης των προβλέψεων.

Σημείωση: Στο συγκεκριμένο παράδειγμα, η εκτίμηση της σταθεράς $\hat{\alpha} =$



Γράφημα 3.3: Μηνιαίες αποδόσεις της Ελληνικής χρηματαγοράς, r_t^{GR}, για την περίοδο από τον **Ιανουάριο του 2010 μέχρι και τον Απρίλιο του 2013** (μπλε γραμμή με κύκλους) μαζί με τις σημειακές προβλέψεις **Μαΐου, Ιουνίου 2013** (κόκκινη γραμμή με κύκλους) και το 95% διάστημα εμπιστοσύνης των προβλέψεων (κόκκινες γραμμές με τρίγωνα που δείχνουν προς τα επάνω και κάτω).

0.233 είναι **στατιστικά μη σημαντική** αφού

$$|t| = \left|\frac{0.233}{0.439}\right| < 2$$

δηλαδή η υπόθεση $\alpha = 0$ **δεν απορρίπτεται** τουλάχιστον σε επίπεδο σημαντικότητας 5%. Πρακτικά, θα ήταν ορθότερο και ρεαλιστικότερο να υιοθετήσουμε τον περιορισμό $\alpha = 0$ στο υπόδειγμα πρόβλεψης και να σχηματίσουμε προβλέψεις για την απόδοση της Ελληνικής χρηματαγοράς με βάση τους τύπους

$$\hat{r}_{T+1}^{GR} = 1.072 \cdot r_{T+1}^{USA} = 1.072 \times -0.5 = -0.536$$

$$\hat{r}_{T+2}^{GR} = 1.072 \cdot r_{T+2}^{USA} = 1.072 \times -0.25 = -0.268$$

$$\vdots$$

Όμως το διάστημα εμπιστοσύνης είναι τόσο μεγάλο, που πρακτικά δεν ενδιαφέρει η επιβολή του περιορισμού ή όχι.

3.3 Ασκήσεις

1. Στο απλό γραμμικό υπόδειγμα της μορφής

$$Y_i = \alpha + \beta X_i + u_i, \quad i = 1, ..., n$$

όπου υποθέτουμε ότι ισχύει το πρώτο σύνολο των κλασσικών υποθέσεων με $u_i \sim N.i.d\left(0, \sigma^2\right)$, $\forall i$ και X_i μη στοχαστική $\forall i$, δείξτε ότι ο εκτιμητής ελαχίστων τετραγώνων

$$\hat{\beta} = \frac{\sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)\left(X_i - \bar{X}\right)}{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2} = \frac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} x_i^2}$$

του β κατανέμεται σύμφωνα με την κανονική κατανομή.

Απάντηση

Αναλύουμε το $\hat{\beta}$ με βάση το σφάλμα δειγματοληψίας του εκτιμητή,

$$\hat{\beta} = \beta + \frac{\sum_{i=1}^{n} x_i u_i}{\sum_{i=1}^{n} x_i^2} = \beta + \sum_{i=1}^{n} w_i u_i$$

όπου

$$w_i = \frac{x_i}{\sum_{i=1}^{n} x_i^2}$$

Παρατηρήστε ότι

$$\sum_{i=1}^{n} w_i^2 = \frac{1}{\sum_{i=1}^{n} x_i^2}$$

Σημειώστε ότι αν δύο τυχαίες μεταβλητές, έστω z_1, z_2, ικανοποιούν την

$$z_1 \sim N.i.d\left(\mu_1, \sigma_1^2\right)$$
$$z_2 \sim N.i.d\left(\mu_2, \sigma_2^2\right)$$

και είναι ανεξάρτητες, τότε το άθροισμά τους κατανέμεται επίσης κανονικά

με

$$z_1 + z_2 \sim N.i.d\left(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2\right)$$

Γενικεύοντας σε n ανεξάρτητες κανονικά κατανεμόνενες μεταβλητές

$$\sum_{i=1}^{n} z_i \sim N.i.d\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right)$$

Σχετικά με την άσκησή μας, παρατηρήστε ότι επειδή

$$u_i \sim N.i.d\left(0, \sigma^2\right)$$

έχουμε

$$w_i u_i \sim N.i.d\left(0, \sigma^2 w_i^2\right)$$

και

$$\sum_{i=1}^{n} w_i u_i \sim N.i.d\left(0, \sigma^2 \sum_{i=1}^{n} w_i^2\right)$$

άρα

$$\sum_{i=1}^{n} w_i u_i \sim N.i.d\left(0, \frac{\sigma^2}{\sum\limits_{i=1}^{n} x_i^2}\right)$$

δηλαδή

$$\hat{\beta} = \beta + \sum_{i=1}^{n} w_i u_i \sim N.i.d\left(\beta, \frac{\sigma^2}{\sum\limits_{i=1}^{n} x_i^2}\right)$$

2. **(α)** εξηγήστε γιατί στο απλό γραμμικό υπόδειγμα το διάστημα εμπιστοσύνης της πρόβλεψης για την Y_0 αυξάνεται συμμετρικά καθώς απομακρύνεται η τιμή X_0 από τη μέση δειγματική τιμή $\bar{X}$.

   (β) Χρησιμοποιώντας αναλυτικούς τύπους, βρείτε το διάστημα εμπιστοσύνης της Y_0 όταν $X_0 = \bar{X}$.

   Απάντηση

   (α) Το εύρος του διαστήματος εμπιστοσύνης είναι άμεσα συνδεδεμένο με το

τυπικό σφάλμα της πρόβλεψης $\widehat{se}\,(e_0)$ το οποίο δίνεται αναλυτικά παρακάτω,

$$\widehat{se}\,(e_0) = \hat{\sigma}\left(1 + \frac{1}{n} + \frac{x_0^2}{\sum_{i=1}^{n} x_i^2}\right)^{1/2}$$
$$= \hat{\sigma}\left(1 + \frac{1}{n} + \frac{\left(X_0 - \bar{X}\right)^2}{\sum_{i=1}^{n} x_i^2}\right)^{1/2}$$

Το τυπικό σφάλμα πρόβλεψης είναι συνάρτηση της απόστασης της τιμής X_0 που υποθέτουμε για την ανεξάρτητη μεταβλητή από τον δειγματικό μέσο της μεταβλητής $\bar{X}$. Από την τελευταία σχέση είναι εμφανές ότι το τυπικό σφάλμα αυξάνει συμμετρικά όταν η απόκλιση $x_0 = \left(X_0 - \bar{X}\right)$ αυξάνει αφού η απόκλιση x_0 εισέρχεται στον τύπο του τυπικού σφάλματος πρόβλεψης τετραγωνικά x_0^2.

(β) Στην περίπτωση που η τιμή της ερμηνευτικής μεταβλητής αναμένεται να είναι ίση με το δειγματικό μέσο των παρατηρήσεων ή τη θέτουμε ίση, δηλαδή $X_0 = \bar{X}$, τότε η εκτίμηση του τυπικού σφάλματος απλοποιείται στην παρακάτω έκφραση,

$$\widehat{se}\,(e_0) = \sqrt{\hat{\sigma}^2\left(1 + \frac{1}{n}\right)}$$

αφού

$$\left(X_0 - \bar{X}\right) = \left(\bar{X} - \bar{X}\right) = 0$$

Άρα το 95% διάστημα εμπιστοσύνης για την «πραγματική» τιμή Y_0 δίνεται από

$$\hat{Y}_0 \pm t_{n-2}^{0.05/2}\hat{\sigma}\sqrt{\left(1 + \frac{1}{n}\right)}$$

3. Με βάση το υπόδειγμα (3.1) δείξτε ότι

$$Var\,(\hat{\alpha}) = \hat{\sigma}^2\left(\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n} x_i^2}\right)$$

4. Με βάση το υπόδειγμα (3.1) δείξτε ότι

$$\hat{\beta} \sim N\left(\beta, \sigma^2 \left(\sum_{i=1}^{n} x_i^2\right)^{-1}\right)$$

Υπόδειξη: Κάνετε χρήση, **πρώτον** του σφάλματος δειγματοληψίας

$$\hat{\beta} - \beta = \sum_{i=1}^{n} w_i u_i$$

όπου οι σταθμίσεις w_i δίνονται από

$$w_i = \frac{x_i}{\sum_{i=1}^{n} x_i^2}$$

και **δεύτερον** της υπόθεσης ότι $u_i \sim N.i.d\left(0, \sigma^2\right)$.

5. Για το υπόδειγμα (3.1) αποδείξτε ότι το σφάλμα πρόβλεψης

$$e_0 = -\left(\hat{\beta} - \beta\right) x_0 + (u_0 - \bar{u})$$

έχει διακύμανση (ή Μέσο Τετραγωνικό Σφάλμα)

$$Var\left(e_0\right) = \sigma^2 \left(1 + \frac{1}{n} + \frac{x_0^2}{\sum_{i=1}^{n} x_i^2}\right)$$

Υπόδειξη: Διαχωρίστε το σφάλμα πρόβλεψης σε δύο τυχαίους όρους, π.χ.,

$$e_0 = B + A$$

με

$$B = -\left(\hat{\beta} - \beta\right) x_0$$

και

$$A = u_0 - \bar{u}$$

οπότε η διακύμανση του σφάλματος πρόβλεψης αναλύεται σε τρία μέρη:

$$Var\left(e_0\right) = Var\left(B\right) + Var\left(A\right) + 2 \cdot Cov\left(B, A\right)$$

Κάνετε χρήση των υποθέσεων που αφορούν τις ροπές του διαταρακτικού όρου και του αποτελέσματος της άσκησης 9 του κεφαλαίου 2 για να δείξετε ότι $Cov(B, A) = 0$ και προβείτε σε υπολογισμό των διακυμάνσεων $Var(B)$, $Var(A)$.

Απάντηση

Η απόδειξη αρχίζει ως εξής:

$$\begin{aligned} Cov\left(\hat{\beta} - \beta, u_0 - \bar{u}\right) &= Cov\left(\hat{\beta} - \beta, u_0\right) - Cov\left(\hat{\beta} - \beta, \bar{u}\right) \\ [0.25cm] &= 0 - \frac{1}{n}\sum_{i=1}^{n} Cov\left(\hat{\beta} - \beta, u_i\right) \\ &= 0 - \frac{\sigma^2}{n}\sum_{i=1}^{n} w_i \\ &= 0 - \frac{\sigma^2}{n} \times 0 = 0 \end{aligned}$$

όπου η δεύτερη και τρίτη ισότητα προκύπτουν από τις υποθέσεις σχετικά με το διαταρακτικό όρο και τα αποτελέσματα της άσκησης 9 του κεφαλαίου 2. Οι σταθμίσεις w_i στην τρίτη ισότητα δίνονται από

$$w_i = \frac{x_i}{\sum\limits_{i=1}^{n} x_i^2}$$

οπότε

$$\begin{aligned} \sum_{i=1}^{n} w_i &= \sum_{i=1}^{n} \frac{x_i}{\left(\sum\limits_{i=1}^{n} x_i^2\right)} \\ &= \frac{1}{\left(\sum\limits_{i=1}^{n} x_i^2\right)} \cdot \sum_{i=1}^{n} x_i \end{aligned}$$

$$= \frac{1}{\left(\sum_{i=1}^{n} x_i^2\right)} \cdot 0 = 0$$

6. Δείξτε ότι **ο εκτιμητής ελαχίστων τετραγώνων**

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} \hat{u}_i^2$$

είναι **αμερόληπτος εκτιμητής της διακύμανσης του διαταρακτικού όρου** σ^2.

Υπόδειξη: χρησιμοποιήστε την άσκηση 10 του κεφαλαίου 2.

Απάντηση

Προτού εφαρμόσουμε τον τελεστή προσδοκιών στον εκτιμητή $\mathrm{E}(\hat{\sigma}^2)$ μπορούμε να απλουστεύσουμε την ανάλυσή μας με την εξής τεχνική: αναλύουμε (αναπτύσσουμε ή διαχωρίζουμε) τα κατάλοιπα $\hat{u}_i$ στα βασικά τους στοχαστικά συστατικά δηλαδή τον διαταρακτικό όρο ή συναρτήσεις του διαταρακτικού όρου για τις οποίες έχουμε ήδη υπολογίσει ή γνωρίζουμε σχετικές με την άσκηση ροπές (π.χ., το σφάλμα δειγματοληψίας του εκτιμητή $\hat{\beta} - \beta$ το οποίο είναι συνάρτηση των διαταρακτικών όρων). Στο συγκεκριμένο παράδειγμα, ο διαχωρισμός θα διευκολύνει την εφαρμογή του τελεστή προσδοκιών $E(\cdot)$. Διαχωρίζουμε λοιπόν τα τετραγωνικά κατάλοιπα σε όρους που θα βοηθήσουν στην εύρεση της ζητούμενης προσδοκίας. Από την άσκηση 10 του κεφαλαίου 2 έχουμε τον παρακάτω διαχωρισμό

$$\hat{u}_i = (u_i - \bar{u}) - \left(\hat{\beta} - \beta\right) x_i$$

άρα

$$\hat{u}_i^2 = \left[-\left(\hat{\beta} - \beta\right) x_i + (u_i - \bar{u})\right]^2$$

$$= \left(\hat{\beta} - \beta\right)^2 x_i^2 - 2\left(\hat{\beta} - \beta\right)(u_i - \bar{u})\, x_i + (u_i - \bar{u})^2$$

και

$$\begin{aligned}\sum_{i=1}^{n}\hat{u}_i^2 &= \left(\hat{\beta}-\beta\right)^2\sum_{i=1}^{n}x_i^2 \\ &\quad -2\left(\hat{\beta}-\beta\right)\sum_{i=1}^{n}u_i x_i \\ &\quad +2\left(\hat{\beta}-\beta\right)\bar{u}\sum_{i=1}^{n}x_i \\ &\quad +\sum_{i=1}^{n}\left(u_i-\bar{u}\right)^2 \\ &= A+B+C+D\end{aligned}$$

Σχετικά με τον όρο A παρατηρούμε ότι επειδή τα x_i θεωρούνται μη στοχαστικά και

$$E\left(\hat{\beta}-\beta\right)^2 = \sigma^2\left(\sum_{i=1}^{n}x_i^2\right)^{-1}$$

έχουμε

$$E\left(A\right) = E\left(\hat{\beta}-\beta\right)^2\sum_{i=1}^{n}x_i^2 = \sigma^2\left(\sum_{i=1}^{n}x_i^2\right)^{-1}\sum_{i=1}^{n}x_i^2 = \sigma^2$$

Σχετικά με τον όρο B παρατηρούμε ότι

$$\left(\hat{\beta}-\beta\right) = \frac{\sum_{i=1}^{n}x_i u_i}{\sum_{i=1}^{n}x_i^2} \Leftrightarrow \sum_{i=1}^{n}x_i^2\left(\hat{\beta}-\beta\right) = \sum_{i=1}^{n}x_i u_i$$

άρα

$$B = -2\left(\hat{\beta}-\beta\right)^2\left(\sum_{i=1}^{n}x_i^2\right)$$

και $E\left(B\right) = -2\sigma^2$.

Σχετικά με τον όρο C παρατηρούμε ότι $\sum_{i=1}^{n} x_i = 0$ άρα $C = 0$.

Τέλος σχετικά με τον όρο D έχουμε ότι

$$D = \sum_{i=1}^{n} (u_i - \bar{u})^2 = \sum_{i=1}^{n} u_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} u_i\right)^2$$

άρα

$$E(D) = n\sigma^2 - \frac{1}{n} n\sigma^2 = \sigma^2 (n-1)$$

Οπότε, η αναμενόμενη τιμή του εκτιμητή $\mathrm{E}(\hat{\sigma}^2)$ δίνεται από

$$E\left(\hat{\sigma}^2\right) = \frac{1}{n-2} E\left(\sum_{i=1}^{n} \hat{u}_i^2\right)$$

$$= \frac{1}{n-2}\left[E(A) + E(B) + E(C) + E(D)\right]$$

$$= \frac{1}{n-2}\left[\sigma^2 - 2\sigma^2 + 0 + \sigma^2 (n-1)\right]$$

$$= \frac{1}{n-2}\sigma^2 (n-2)$$

$$= \sigma^2$$

7. Παρακάτω, σας δίνονται εκτιμήσεις τεσσάρων υποδειγμάτων (με τη μέθοδο ελαχίστων τετραγώνων). Σε παρενθέσεις αναφέρονται αναφέρονται τα τυπικά σφάλματα $\widehat{se}(\cdot)$ των εκτιμημένων συντελεστών.

$$Y_i = \underset{(0.91)}{1.56} + \underset{(1.19)}{2.67} X_i + \hat{u}_i, \quad i = 1, ..., 12$$

$$Y_i = \underset{(0.092)}{0.35} + \underset{(0.28)}{1.55} X_i + \hat{u}_i, \quad i = 1, ..., 136$$

$$Y_i = \underset{(0.025)}{0.081} + \underset{(1.13)}{1.09} X_i + \hat{u}_i, \quad i = 1, ..., 120$$

$$Y_i = \underset{(7.88)}{10.56} + \underset{(0.356)}{0.875} X_i + \hat{u}_i, \quad i = 1, ..., 15$$

Προβείτε σε ελέγχους στατιστικής σημαντικότητας των συντελεστών κλίσης και των σταθερών όρων.

8. Σας δίνονται τα παρακάτω εκτιμημένα υποδείγματα. Σε παρενθέσεις αναφέρονται εκτιμήσεις τυπικών σφαλμάτων των εκτιμημένων συντελεστών του υποδείγματος

$$Y_i = \underset{(0.98)}{3.01} + \underset{(0.77)}{1.31} X_i + \hat{u}_i, \quad \hat{\sigma}^2 = 0.44 \quad , \quad i = 1, ..., 68$$

$$Y_i = \underset{(0.36)}{0.47} - \underset{(0.24)}{0.92} X_i + \hat{u}_i, \quad \hat{\sigma}^2 = 0.34 \quad , \quad i = 1, ..., 50$$

(α) Ελέγξτε αν ο συντελεστής κλίσης είναι ίσος με τη μονάδα στο πρώτο υπόδειγμα και ίσος με -1 στο δεύτερο υπόδειγμα.

(β) Κατασκευάστε τα 90% και 95% διαστήματα εμπιστοσύνης για τους συντελεστές κλίσης και τη διακύμανση του διαταρακτικού όρου σε κάθε υπόδειγμα ξεχωριστά.

9. Σας δίνεται το παρακάτω υπόδειγμα αφού εκτιμήθηκε με τη μέθοδο των ελαχίστων τετραγώνων. Το μέγεθος του δείγματος είναι $T = 22$ παρατηρήσεις. Τα δεδομένα αντιστοιχούν σε χρονοσειρές. Αναλυτικά

$$Y_t = \underset{(0.209)}{0.614} - \underset{(0.027)}{0.450} X_t + \hat{u}_t \, , \, t = 1, ..., T$$

$$\hat{\sigma} = 0.389 \, , \, \bar{X} = 6.913 \, , \, \sum_{t=1}^{T} \left(X_i - \bar{X} \right)^2 = 195.146$$

Σε παρενθέσεις αναφέρονται τυπικά σφάλματα $\widehat{se}\,(\cdot)$ των εκτιμημένων συντελεστών του υποδείγματος. Το τυπικό σφάλμα της παλινδρόμησης ή αλλιώς η εκτίμηση της τυπικής απόκλισης των διαταρακτικών όρων είναι $\hat{\sigma} = 0.389$.

(α) Προβείτε σε πρόβλεψη των τιμών της μεταβλητής Y_t τις περιόδους $T + 1$, $T + 2$ και $T + 3$ αν $X_{T+1} = 6.43$, $X_{T+2} = 7.1$ και $X_{T+3} = 8.25$. Δηλαδή υπολογίστε τις $\hat{Y}_{T+1}$, $\hat{Y}_{T+2}$ και $\hat{Y}_{T+3}$.

(β) Κατασκευάστε το 90% διάστημα εμπιστοσύνης για κάθε ορίζοντα πρόβλεψης. Σε ποιόν ορίζοντα πρόβλεψης παρατηρείτε το μικρότερο σφάλμα πρόβλεψης και γιατί;

10. Έστω ότι i_t είναι το επιτόκιο δανεισμού της Ελλάδας (μέση ετήσια απόδοση 10-ετούς κρατικού ομολόγου) και E_t είναι το δημόσιο έλλειμα ως ποσοστό

του ΑΕΠ. Έχετε ετήσια δεδομένα για το διάστημα 1980 - 2009. Θέλετε να εκτιμήσετε το υπόδειγμα

$$i_t = a + \beta E_t + u_t$$

με βάση τα παρακάτω στοιχεία

$\sum (i_t - \bar{\imath})^2 = 107.206$ $\quad$ $\sum \left(E_t - \bar{E}\right)^2 = 319.577$

$\sum (i_t - \bar{\imath}) \left(E_t - \bar{E}\right) = 146.848$

$\bar{\imath} = 6.95$ $\quad$ $\bar{E} = 7.48$

(α) Εκτιμήστε με τη μέθοδο ελαχίστων τετραγώνων τις παραμέτρους α και β

(β) Εκτιμήστε το τυπικό σφάλμα του εκτιμητή $\hat{\beta}$ και προβείτε σε έλεγχο σημαντικότητας για το $\hat{\beta}$ όταν το άθροισμα των τετραγώνων των καταλοίπων είναι ίσο με $\sum \hat{u}_t^2 = 39.728$. Επίσης, προβείτε σε έλεγχο της υπόθεσης $\beta = 0.5$

(γ) Ερμηνεύστε οικονομικά την εκτίμηση $\hat{\beta}$ του συντελεστή β

(δ) Πόσο προβλέπεται να είναι το ετήσιο επιτόκιο δανεισμού με δημόσιο έλλειμα της τάξης του 8.1% του ΑΕΠ για το 2010 (όπως προϋποθέτει το αρχικό μνημόνιο); Υπολογίστε το τυπικό σφάλμα της πρόβλεψης και τα 99%, 95%, 90%, 85% διαστήματα εμπιστοσύνης της πρόβλεψης. Υιοθετήστε τη συνάρτηση $= tinv(prob, df)$ του Excel για να βρείτε τις κρίσιμες τιμές της t-student κατανομής. Παρομοίως, υιοθετήστε τη συνάρτηση $= norminv(prob, mean, sd)$ για την κανονική κατανομή. **Σημείωση:** Η συνάρτηση $= tinv(prob, df)$ επιστρέφει την κρίσιμη τιμή που αναφέρεται στη δίπλευρη κατανομή, δηλαδή αν $prob = 0.05$, τότε κάθε απόληξη της κατανομής έχει τιμή πιθανότητας ίση με 0.025 (ή p-value=0.025)

11. Θέλετε να εκτιμήσετε την επίδραση μίας συγκεκριμένης διαφοράς επιτοκίων (spread) επί του ρυθμού μεγέθυνσης του πραγματικού ΑΕΠ χρησιμοποιώντας το υπόδειγμα

$$z_t = a + \beta s_{t-1} + u_t$$

όπου

$$z_t = 100 \times \left(\frac{y_t - y_{t-1}}{y_{t-1}} \right)$$

με y_t το πραγματικό τριμηνιαίο ΑΕΠ και $s_t = r_t^{Gre} - r_t^{Ger}$ το μέσο τριμηνιαίο spread (**διαφορά**) Ελληνικών και Γερμανικών επιτοκίων (αποδόσεων) κρατικών μακροπρόθεσμων (για παράδειγμα 10-ετή) ομολόγων. Π.χ., $s_t = 1.26$ σημαίνει ότι η **διαφορά** είναι στο 1.26% ενώ αν $z_t = 0.85$, τότε η μεγέθυνση από το τρίμηνο $t-1$ στο τρίμηνο t ήταν 0.85%. Έχουμε στοιχεία για την περίοδο: πρώτο τρίμηνο 2000 μέχρι και το δεύτερο τρίμηνο του 2011 με βάση τα οποία υπολογίστηκαν τα παρακάτω αθροίσματα

$$\sum (z_t - \bar{z})^2 = 139.148 \qquad \sum (s_{t-1} - \bar{s})^2 = 210.316$$
$$\sum (z_t - \bar{z})(s_{t-1} - \bar{s}) = -76.837$$
$$\bar{z} = 0.503 \qquad \bar{s} = 1.264$$

Εξαιτίας των πρώτων διαφορών και της υστέρησης «χάνουμε» μία παρατήρηση στο δείγμα εκτίμησης, άρα $T = 45$.

(α) Εκτιμήστε με τη μέθοδο ελαχίστων τετραγώνων τις παραμέτρους α και β

(β) Εκτιμήστε το τυπικό σφάλμα του $\hat{\beta}$ και προβείτε σε έλεγχο σημαντικότητας για το $\hat{\beta}$ αν το άθροισμα των τετραγώνων των καταλοίπων είναι ίσο με $\sum \hat{u}_t^2 = 111.076$. Επίσης, προβείτε σε έλεγχο της υπόθεσης $\beta = -0.40$

(γ) Ερμηνεύστε τον εκτιμημένο συντελεστή $\hat{\beta}$ και προβείτε στις εξής προβλέψεις: «Πόσο θα αυξηθεί ή μειωθεί κατά μέσο όρο ο ρυθμός ανάπτυξης αν η διαφορά s_t αγγίξει το 12.4%;» και «Ποια είναι η μέση ανάπτυξη (ή μείωση) όταν η διαφορά s_t μηδενιστεί»; Υπολογίστε το 95% διάστημα εμπιστοσύνης των προβλέψεων. Επειδή $T - 2 = 43 > 30$ υιοθετήστε την κρίσιμη τιμή 2 στο διάστημα εμπιστοσύνης. **Σημείωση:** η διαφορά όντως άγγιξε το 12.4% το τρίτο τρίμηνο του 2011.

12. Σας δίνονται τα παρακάτω 5 ζεύγη τιμών για τις μεταβλητές Y_i, X_i

X_i	Y_i
1.7640	0.6633
1.2013	2.4432
2.1929	−0.0578
2.7982	−1.1955
2.7692	−2.1030

και η γραμμική σχέση $Y_i = \beta_1 + \beta_2 X_i + u_i$ που τις συνδέει με βάση την οικονομική θεωρία.

(**α**) Προβείτε σε εκτίμηση των $\hat{\beta}_1$, $\hat{\beta}_2$ με τη μέθοδο των ελαχίστων τετραγώνων

(**β**) Προβείτε σε έλεγχο στατιστικής σημαντικότητας της επίδρασης της X_i στην Y_i. Σας δίνονται οι κρίσιμες τιμές (δικατάληκτου ελέγχου) $t_2^{0.05/2} = 4.302$, $t_3^{0.05/2} = 3.182$, $t_4^{0.05/2} = 2.776$, $t_5^{0.05/2} = 2.570$ για επίπεδο σημαντικότητας 5% της κατανομής t-student.

(**γ**) Προβείτε στον έλεγχο της υπόθεσης $H_0 : \beta_2 = -2.5$ έναντι της $H_1 : \beta_2 \neq -2.5$

(**δ**) Πόσο αναμένεται να μεταβληθεί η Y_i αν αυξηθεί η X_i κατά 2.5 μονάδες;

(**ε**) Υπολογίστε τον συντελεστή προσδιορισμού R^2 και προβείτε σε ερμηνεία του

(**στ**) Ποια η πρόβλεψή σας για τη μεταβλητή Y_i όταν η $X_i = \bar{X}$;

13. (βασισμένη σε άσκηση και δεδομένα του Wooldridge, 2002). Σας δίνονται τα παρακάτω στοιχεία από ένα δείγμα 935 εργαζομένων με σκοπό να εκτιμήσετε ένα απλό γραμμικό υπόδειγμα που εξηγεί τον πραγματικό μηνιαίο μισθό W (ευρώ) των εργαζομένων σε όρους του δείκτη IQ των εργαζομένων (μονάδες IQ, ελάχιστο: 50, μέγιστο: 145)

$$\bar{W} = 957.95 \ , \ \overline{IQ} = 101.28 \ , \ \overline{\ln W} = 6.779$$

$$\sum_i \left(W_i - \bar{W}\right)^2 = 152716168$$

$$\sum_i \left(IQ_i - \overline{IQ}\right)^2 = 211627$$

$$\sum_i \left(W_i - \bar{W}\right)\left(IQ_i - \overline{IQ}\right) = 1757156$$

$$\sum_i \left(\ln W_i - \overline{\ln W}\right)^2 = 165.6562$$

$$\sum_i \left(\ln W_i - \overline{\ln W}\right)\left(IQ_i - \overline{IQ}\right) = 1863.8361$$

Το τυπικό σφάλμα της σχετικής παλινδρόμησης εκτιμήθηκε ίσο με $\hat{\sigma} = 384.7667$.

(α) εκτιμήστε ένα υπόδειγμα όπου μία μοναδιαία μεταβολή στον δείκτη IQ υπονοεί σταθερή μεταβολή στον μηνιαίο μισθό. Χρησιμοποιήστε το υπόδειγμα για να προβλέψετε τη μεταβολή στον μισθό από μία αύξηση του δείκτη IQ κατά 15 μονάδες. Εξηγεί ο δείκτης IQ το μεγαλύτερο ποσοστό της μεταβλητότητας στον μηνιαίο μισθό; Αναφέρετε τουλάχιστον δύο επιπλέον ερμηνευτικές μεταβλητές που θα μπορούσαν να προστεθούν στο συγκεκριμένο υπόδειγμα.

(β) εκτιμήστε ένα υπόδειγμα όπου μία μοναδιαία μεταβολή στον δείκτη IQ υπονοεί σταθερή ποσοστιαία μεταβολή στον μηνιαίο μισθό. Αν ο δείκτης IQ αυξηθεί κατά 15 μονάδες, ποια είναι η προβλεπόμενη ποσοστιαία αύξηση στον μηνιαίο μισθό; Πόσο πρέπει να αυξηθεί ο δείκτης IQ ώστε να δώσει μία αύξηση στον μηνιαίο μισθό της τάξης του 25%;

14. Σας δίνονται στοιχεία[13] για τις μεταβλητές, **wdi_lifexp** και **p_polity2**. Τα οικονομικά δεδομένα βρίσκονται συνολικά στο αρχείο δεδομένων

```
qog_std_cs_jan21.gdt
```

και η εμπειρική άσκηση αναπτύσσεται στο αρχείο εντολών

```
qog_std_cs_jan21.inp
```

Το δείγμα μας έχει 194 παρατηρήσεις ($n = 194$ χώρες!!!) ενώ υπάρχουν διαθέσιμες (με απούσες τιμές σε κάποιες μεταβλητές και/ή χώρες) 1743

[13] Βλ. Teorell, Jan, Aksel Sundström, Sören Holmberg, Bo Rothstein, Natalia Alvarado Pachon & Cem Mert Dalli. 2021. The Quality of Government Standard Dataset, version Jan21. University of Gothenburg: The Quality of Government Institute, http://www.qog.pol.gu.se doi:10.18157/qogstdjan21

μεταβλητές!!!.

(**α**) wdi_lifexp: το προσδόκιμο ζωής κατά τη γέννηση, (έτη). Το προσδόκιμο ζωής κατά τη γέννηση υποδεικνύει τον αριθμό των ετών που θα ζούσε ένα νεογέννητο εάν τα επικρατούντα πρότυπα θνησιμότητας κατά τη γέννησή του παρέμεναν ίδια καθ' όλη τη διάρκεια της ζωής του.

(**β**) p_polity2: Αναθεωρημένος (συνδυαστικός) δείκτης «δημοκρατικότητας» ή βαθμού δημοκρατίας. Ο βαθμός δημοκρατίας, όπως μετράται από το Polity project. Η κλίμακα της μεταβλητής κυμαίνεται από +10 (έντονα δημοκρατική χώρα) έως -10 (έντονα απολυταρχική).

Σκοπός μας είναι να εξετάσουμε τη σχέση μεταξύ βαθμού δημοκρατίας και προσδόκιμου ζωής. Υποθέτουμε ότι οι άνθρωποι σε πιο δημοκρατικές χώρες ζουν περισσότερο και το ελέγχουμε στατιστικά. Βέβαια, δεν ξεχνάμε ότι η στατιστική επιβεβαίωση μίας σχέσης των δύο μεταβλητών μπορεί να μην είναι αξιόπιστη, είτε (**α**) επειδή είναι τελείως πλασματική είτε (**β**) επειδή βλέπουμε ένα «πληθωρισμένο» αποτέλεσμα λόγω παράλειψης σημαντικών μεταβλητών που επιδρούν **και στις δύο** (ή περισσότερες) υπο μελέτη μεταβλητές μας, για παράδειγμα, μία τέτοια μεταβλητή στη συγκεκριμένη άσκηση θα μπορούσε να είναι ο βαθμός οικονομικής ευμάρειας ή επίπεδο οικονομικής ανάπτυξης κάθε χώρας;

Αρχικά, εξετάζουμε ορισμένα περιγραφικά/περιληπτικά στατιστικά μέτρα:

Περιληπτικά στατιστικά, χρησιμοποιώντας τις παρατηρήσεις 1 - 194
(απούσες τιμές αγνοήθηκαν[α΄])

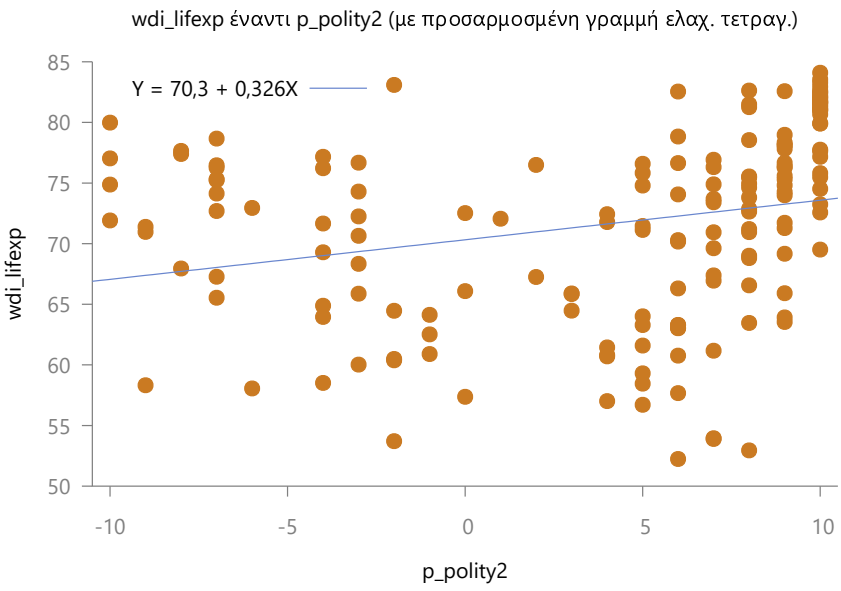
Μεταβλητή	Μέσος	Διάμεσος	Τ.Α.[β΄]	Ελάχ	Μέγ
p_polity2	4,10	6,00	6,17	−10,00	10,00
wdi_lifexp	72,00	73,33	7,65	52,24	84,10

[α΄]Να τονίσουμε ότι έχουμε πληροφορίες για 184 χώρες αναφορικά με το προσδόκιμο ζωής και για 165 χώρες αναφορικά με το βαθμό δημοκρατίας.
[β΄]Τ.Α.: Τυπική απόκλιση

Βλέπουμε ότι το προσδόκιμο ζωής (κατά τη γέννηση) είναι κατά μέσο όρο 72 έτη. Η μεταβλητή του βαθμού δημοκρατίας κυμαίνεται από −10 (μέγιστου

βαθμού ανελεύθερο καθεστώς) έως +10 (μέγιστου βαθμού δημοκρατία), με μέση τιμή[14] 4.10 βαθμούς (πρόκειται για τον αριθμητικό μέσο της p_polity2 χωρίς να υπολογίζονται οι 10 απούσες τιμές).

Το **γράφημα διασποράς (3.4)** απεικονίζει μία πιθανή **θετική** σχέση των δύο μεταβλητών. Η μπλε γραμμή της εκτιμημένης (με τη μέθοδο ελαχίστων τετραγώνων) γραμμικής σχέσης έχει «ελαφρώς» θετική κλίση κάτι που αναδεικνύεται και στατιστικά πιο κάτω στα αναλυτικά αποτελέσματα της εκτίμησης (στατιστικά σημαντικός και θετικός συντελεστής κλίσης). Βέβαια, η προσαρμογή της γραμμής δεν είναι αρκετά καλή καθώς υπάρχει σημαντική διασπορά των παρατηρούμενων τιμών γύρω από τη γραμμή (έτσι εξηγείται και η χαμηλή τιμή του συντελεστή προσδιορισμού R^2 που αναφέρεται παρακάτω).



Γράφημα 3.4: Διάγραμμα διασποράς του προσδόκιμου ζωής κατά τη γέννηση έναντι της μεταβλητής p_polity2.

Συνοπτικά, η εκτιμημένη (με ελάχιστα τετράγωνα) εξίσωση έδωσε τα παρακάτω αποτελέσματα

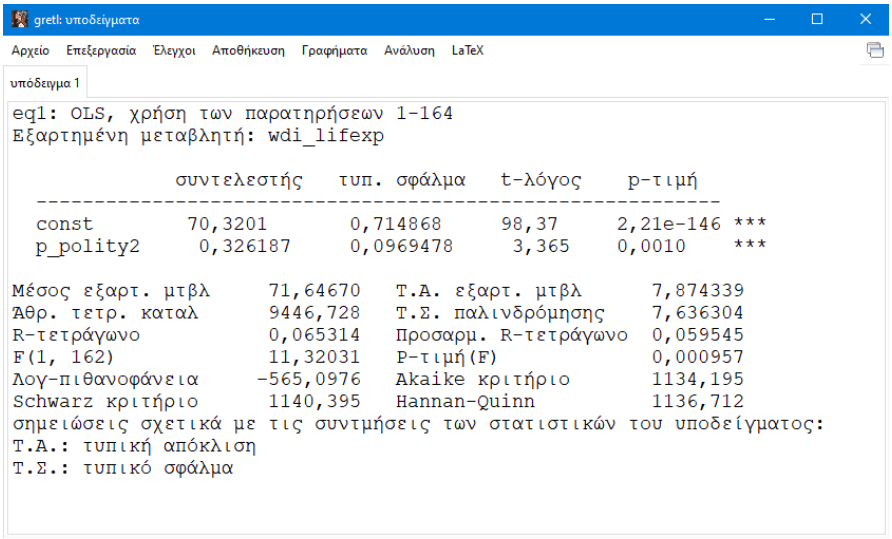
[14] **Η Ελλάδα** είναι η χώρα με τον αριθμό $ccode = 300$ ή είναι η $68η$ παρατήρηση του δείγματος, $i = 68$.

$$\widehat{\text{wdi\_lifexp}} = \underset{(0.71487)}{70.3201} + \underset{(0.096948)}{0.326187}\ \text{p\_polity2}$$

$$T = 164\ ,\ R^2 = 0.0653\ ,\ \hat{\sigma} = 7.6363$$

(τυπικά σφάλματα σε παρενθέσεις)

ενώ η επόμενη εικόνα δείχνει αναλυτικά αποτελέσματα από την εκτίμηση μέσω του gretl,



Παρατηρούμε ότι τόσο η εκτίμηση του σταθερού όρου όσο και του συντελεστή κλίσης είναι στατιστικά σημαντικές. Συγκεκριμένα, ο εκτιμημένος συντελεστής κλίσης $\hat{\beta} = 0.326$ είναι στατιστικά σημαντικός αφού $t_{\hat{\beta}} = \frac{0.3261}{0.0969} = 3.365$. Η τιμή της t-student στατιστικής 3.365 είναι πολύ μεγαλύτερη της κρίσιμης τιμής 1.96 για επίπεδο σημαντικότητας 5% άρα η μηδενική υπόθεση $H_0 : \beta = 0$ του δίπλευρου ελέγχου απορρίπτεται. Εναλλακτικά, **η τιμή πιθανότητας** p-τιμή είναι ίση με 0.0010. Δηλαδή η μηδενική υπόθεση απορρίπτεται όχι μόνο σε 5% αλλά και σε 1% επίπεδο σημαντικότητας αφού p-τιμή $= 0.0010 < 0.01$.

Μάλιστα, **(ερμηνεία p-τιμής)** η πιθανότητα να παρατηρήσουμε μια στατιστική ελέγχου $t_{\hat{\beta}}$ μεγαλύτερη ή ίση από την τιμή 3.365 ενώ

η μηδενική υπόθεση $H_0 : \beta = 0$ είναι αληθής είναι μόλις 0.0010 ή μία στις χίλιες.

Αφού απορρίπτεται η μηδενική υπόθεση $H_0 : \beta = 0$ (δηλαδή απορρίπτεται η υπόθεση ότι ο βαθμός δημοκρατίας δεν εισέρχεται ως ερμηνευτικός παράγοντας του προσδόκιμου ζωής) προβαίνουμε σε ερμηνεία του. Μία αύξηση του βαθμού δημοκρατίας p_polity2 κατά μία μονάδα αυξάνει (ceteris paribus) το προσδόκιμο ζωής wdi_lifexp κατά 0.39 έτη. Στη χειρότερη περίπτωση, p_polity2=-10, έχουμε προσδόκιμο ζωής $70.32 + 0.326 \cdot (-10) = 67.06$ έτη ενώ στην καλύτερη p_polity2=10 έχουμε προσδόκιμο ζωής $70.32 + 0.326 \cdot 10 = 73.58$ έτη (μία διαφορά 6.52 ετών).

Ο συντελεστής προσδιορισμού είναι ίσος με $R^2 = 0.0653$. Αυτό σημαίνει ότι ο βαθμός δημοκρατίας μίας χώρας επεξηγεί το 6.53% της μεταβλητότητας του προσδόκιμου ζωής. Ίσως όχι πολύ όμως ούτε και ανύπαρκτος επεξηγηματικός παράγοντας. Προσοχή, δεν υπάρχουν σαφή όρια βάσει των οποίων να κρίνουμε εάν η R^2 τιμή είναι «μεγάλη ή μικρή» - εξαρτάται πλήρως από το πλαίσιο ανάλυσης.

Σημαίνει όμως η παραπάνω θετική σχέση ότι η δημοκρατία αποτελεί αιτία του προσδόκιμου ζωής με το τελευταίο να είναι το αιτιατό ή αποτέλεσμα; Όχι απαραίτητα, σύμφωνα με την προηγούμενη ανάλυσή μας σε αυτό το κεφάλαιο σχετικά με την πλασματική συσχέτιση. Μπορεί να υπάρχουν άλλοι παράγοντες που **οδηγούν** τόσο σε δημοκρατία όσο και σε υψηλό προσδόκιμο ζωής. Άρα, ο βαθμός δημοκρατίας και το προσδόκιμο ζωής μπορεί να είναι δύο συμπτώματα, παρά η αιτία και το αποτέλεσμα. Επίσης, το αποτέλεσμα μπορεί να είναι «πληθωρισμένο» εξαιτίας της παράλειψης σημαντικών μεταβλητών από το υπόδειγμα που συσχετίζονται τόσο με την εξαρτημένη μεταβλητή όσο και με το βαθμό δημοκρατίας.

Αμέσως παρακάτω στην άσκηση 15, μελετάμε ξεχωριστά την επίδραση του επιπέδου της πραγματικής οικονομίας στο προσδόκιμο ζωής. Στο κεφάλαιο 5 θα δοκιμάσουμε πολυμεταβλητές αναλύσεις και θα ξεκαθαρίσουμε ακόμα περισσότερο το ρόλο του βαθμού δημοκρατίας και του οικονομικού επιπέδου ανάπτυξης μίας χώρας στο προσδόκιμο ζωής των κατοίκων της.

15. Έστω η μεταβλητή wdi_gdpcappppcon2017: το κατά κεφαλήν πραγματικό ΑΕΠ με βάση την ισοτιμία αγοραστικής δύναμης (PPP, purchasing power parity). Με αυτό το μέτρο του πραγματικού ΑΕΠ, το **Α**καθάριστο

Εγχώριο **Π**ροϊόν μετατρέπεται σε «διεθνή» δολάρια (international dollars) χρησιμοποιώντας συντελεστές ισοτιμίας αγοραστικής δύναμης. Ένα διεθνές δολάριο έχει την ίδια αγοραστική δύναμη έναντι του ΑΕΠ με το δολάριο ΗΠΑ στις Ηνωμένες Πολιτείες. Το μέτρο είναι πραγματικό αφού δίνεται σε «σταθερά» διεθνή δολάρια του 2017 (έτος βάσης το 2017).[15]

Σκοπός μας είναι να εξετάσουμε τη σχέση μεταξύ επιπέδου οικονομικής ανάπτυξης ή ευμάρειας (όπως προσεγγίζεται από την wdi_gdpcappppcon2017) και προσδόκιμου ζωής. Υποθέτουμε ανάλογη - θετική σχέση, χαμηλότερα (υψηλότερα) επίπεδα οικονομικής ευμάρειας συνδέονται με χαμηλότερο (υψηλότερο) προσδόκιμο ζωής.

Αρχικά, εξετάζουμε ορισμένα περιγραφικά/περιληπτικά στατιστικά μέτρα:

Περιληπτικά στατιστικά, χρησιμοποιώντας τις παρατηρήσεις 1–194
(απούσες τιμές αγνοήθηκαν[α΄])

Μεταβλητή	Μέσος	Διάμεσος	Τ.Α.[β΄]	Ελάχ	Μέγ
wdi_gdpcap pppcon2017	19919	12649	20291	773,6	112800
wdi_lifexp	72,00	73,33	7,651	52,24	84,10

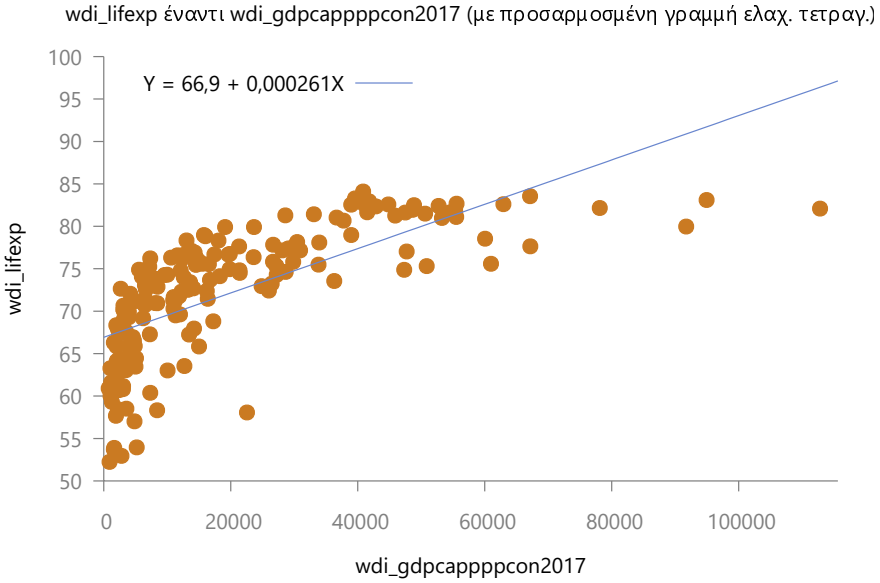
[α΄] Να τονίσουμε ότι έχουμε πληροφορίες για 184 χώρες αναφορικά με το προσδόκιμο ζωής και για 182 χώρες αναφορικά με το βαθμό δημοκρατίας.
[β΄] Τ.Α.: Τυπική απόκλιση

Βλέπουμε ότι το πραγματικό ΑΕΠ κυμαίνεται από 773.60 δολλάρια (ετήσιο εισόδημα, χώρα: Burundi) έως $1,128e+005 = 1,128 \cdot 10^5 = 112800$ (ετήσιο εισόδημα εκατόν δώδεκα χιλιάδες δολλάρια περίπου, χώρα: Luxembourg), με μέση τιμή τα 19 χιλιάδες εννιακόσια δεκαεννιά δολλάρια (19919).

Το παρακάτω **γράφημα διασποράς** (3.5) απεικονίζει την **θετική** σχέση των δύο μεταβλητών. Η μπλε γραμμή της εκτιμημένης (με τη μέθοδο ελαχίστων τετραγώνων) γραμμικής σχέσης έχει θετική κλίση κάτι που αναδει-

[15] *«Η οικονομική ανάπτυξη εκφράζεται σε ισοτιμίες αγοραστικής δύναμης (ΙΑΔ) – οι οποίες λαμβάνουν υπόψη τα διάφορα επίπεδα τιμών στα κράτη καθιστώντας έτσι δυνατή τη δικαιότερη σύγκριση. Με τη χρήση των ΙΑΔ ως συντελεστή μετατροπής, το ΑΕΠ μετατρέπεται σε τεχνητή κοινή νομισματική μονάδα, τη λεγόμενη μονάδα αγοραστικής δύναμης (ΜΑΔ), η οποία καθιστά δυνατή τη σύγκριση της αγοραστικής δύναμης χωρών με διαφορετικά εθνικά νομίσματα.»* Πηγή: `https://ec.europa.eu/eurostat/statistics-explained/index.php`

κνύεται και στατιστικά πιο κάτω στα αναλυτικά αποτελέσματα της εκτίμησης (στατιστικά σημαντικός και θετικός συντελεστής κλίσης). Η προσαρμογή της γραμμής είναι πάρα πολύ καλή όπως θα δούμε και από τον συντελεστή προσδιορισμού R^2 παρακάτω, όμως είναι εμφανές ότι η υποκείμενη σχέση είναι **μη-γραμμική** με το αποτέλεσμα μίας αύξησης του επιπέδου οικονομικής ευμάρειας στο προσδόκιμο ζωής να μειώνεται καθώς το επίπεδο του κατά κεφαλήν πραγματικού ΑΕΠ αυξάνεται.



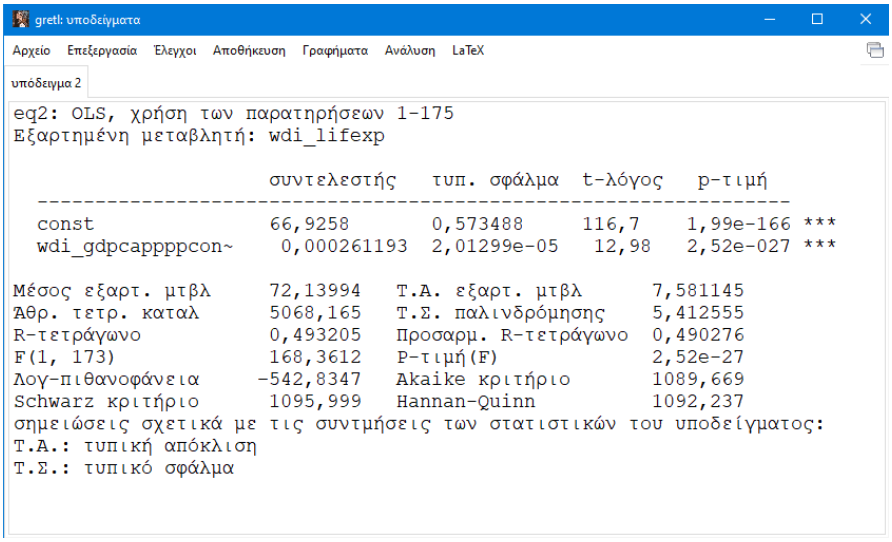
Γράφημα 3.5: Διάγραμμα διασποράς του προσδόκιμου ζωής κατά τη γέννηση έναντι της μεταβλητής wdi_gdpcappppcon2017.

Συνοπτικά, η εκτιμημένη (με ελάχιστα τετράγωνα) εξίσωση παρουσιάζεται παρακάτω

$$\widehat{\text{wdi\_lifexp}} = \underset{(0.57349)}{66.9258} + \underset{(2.0130\text{e}-005)}{0.000261193}\ \text{wdi\_gdpcappppcon2017}$$
$$T = 175\ ,\ R^2 = 0.4932\ ,\ \hat{\sigma} = 5.4126$$
(τυπικά σφάλματα σε παρενθέσεις)

ενώ η επόμενη εικόνα δείχνει αναλυτικά αποτελέσματα από την εκτίμηση

μέσω του gretl,



Παρατηρούμε ότι τόσο η εκτίμηση του σταθερού όρου όσο και του συντελεστή κλίσης είναι στατιστικά σημαντικές. Συγκεκριμένα, ο εκτιμημένος συντελεστής κλίσης $\hat{\beta} = 0.000261193$ είναι στατιστικά σημαντικός αφού

$$t_{\hat{\beta}} = \frac{0.000261193}{0.00002013} = 12.98$$

Η τιμή της t-student στατιστικής 12.98 είναι **πολύ μεγαλύτερη** της κρίσιμης τιμής 1.96 για επίπεδο σημαντικότητας 5% άρα η μηδενική υπόθεση $H_0 : \beta = 0$ του δίπλευρου ελέγχου απορρίπτεται. Εναλλακτικά, **η τιμή πιθανότητας** p-τιμή είναι ίση[16] με 0.0000. Δηλαδή η μηδενική υπόθεση απορρίπτεται όχι μόνο σε 5% αλλά και σε 1% επίπεδο σημαντικότητας αφού p-τιμή $= 0.0000 < 0.01$.

Μάλιστα, **(ερμηνεία p-τιμής)** η πιθανότητα να παρατηρήσουμε μια στατιστική ελέγχου $t_{\hat{\beta}}$ μεγαλύτερη ή ίση από την τιμή 12.98 ενώ η μηδενική υπόθεση $H_0 : \beta = 0$ είναι αληθής είναι πρακτικά μηδενική (2.52×10^{-27}). Το δείγμα λοιπόν παρέχει **ισχυρές ενδείξεις ενάντια** στη μηδενική υπόθεση.

[16] 2.52e-027 αντιστοιχεί σε μία εξαιρετικά μικρή ποσότητα, μηδέν κόμμα 26 μηδενικά και μετά ο αριθμός 252. Πιό αναλυτικά, 0.000...000252 ή 2.52×10^{-27}.

Αφού **απορρίπτεται** η μηδενική υπόθεση $H_0 : \beta = 0$ (δηλαδή απορρίπτεται η υπόθεση ότι το πραγματικό κατά κεφαλήν ΑΕΠ δεν εισέρχεται ως ερμηνευτικός παράγοντας του προσδόκιμου ζωής) προβαίνουμε σε ερμηνεία του. Εδώ τώρα θα παρατηρήσουμε την ανάγκη για πιθανή μετατροπή στις μονάδες μέτρησης[17] αφού η ερμηνεία των εκτιμημένων συντελεστών και ειδικότερα του συντελεστή κλίσης, με τις μεταβλητές (**ειδικά με την ερμηνευτική μεταβλητή**) στις τρέχουσες μονάδες μέτρησης είναι **οικονομικά «μη-διαισθητική»**. Συγκεκριμένα: μία αύξηση του πραγματικού ΑΕΠ wdi_gdpcappppcon2017 κατά μία μονάδα (**δηλαδή κατά ένα δολλάριο** - σε τιμές του 2017) αυξάνει το προσδόκιμο ζωής wdi_lifexp κατά 0.000261193 **έτη** ή $0.000261193 \cdot 365 = 0.095335445$ **ημέρες** ή $0.000261193 \cdot 365 \cdot 24 = 2.2880507$ **ώρες**. Είναι πολύ πιο «διαισθητική» η ερμηνεία του συντελεστή αν **διαιρέσουμε** την wdi_gdpcappppcon2017 με 1000 (σε χιλιάδες δολλάρια αντί απλώς για δολλάρια). Έστω λοιπόν ότι

$$\text{rgdpcap} = \frac{\text{wdi\_gdpcappppcon2017}}{1000}$$

Εκτιμούμε ξανά την εξίσωση με ελάχιστα τετράγωνα και το αποτέλεσμα που λαμβάνουμε (συνοπτικά) δίνεται παρακάτω. **Προσοχή, μόνο ο συντελεστης κλίσης και το τυπικό του σφάλμα πολ/νται με 1000**. **Όλες** οι υπόλοιπες στατιστικές τιμές παραμένουν ίδιες:

$$\widehat{\text{wdi\_lifexp}} = \underset{(0.57349)}{66.9258} + \underset{(0.0201299)}{0.261193}\ \text{rgdpcap}$$

$$T = 175\ ,\ R^2 = 0.4932\ ,\ \hat{\sigma} = 5.4126$$

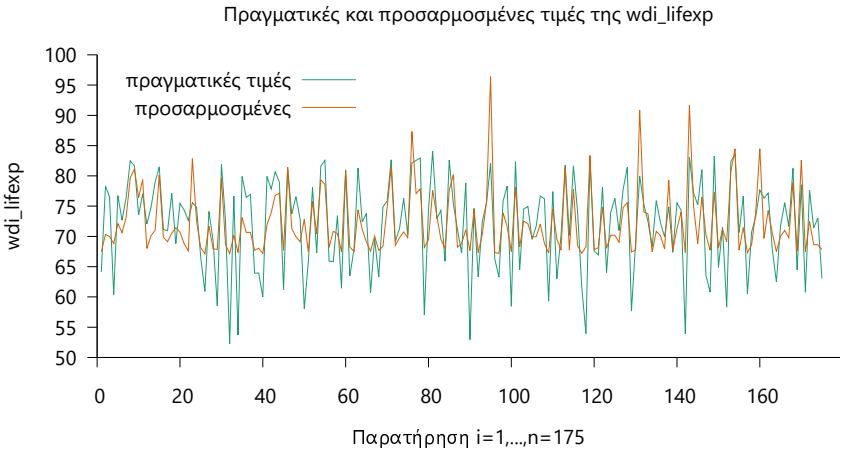
(τυπικά σφάλματα σε παρενθέσεις)

Ερμηνεία: μία αύξηση του πραγματικού ΑΕΠ rgdpcap κατά μία μονάδα (**δηλαδή κατά χίλια δολλάρια** - σε τιμές του 2017) αυξάνει το προσδόκιμο ζωής wdi_lifexp κατά 0.261193 **έτη** ή $0.261193 \cdot 365 = 95.33$ **ημέρες** ή περίπου 3.2 **μήνες**.

Ο **συντελεστής προσδιορισμού** είναι ίσος με $R^2 = 0.4932$. Αυτό

[17] ή στη λογαριθμοποίηση συγκεκριμένων μεταβλητών, βλ. διάγραμμα διασποράς και μη-γραμμικότητα συγκεκριμένου τύπου. Περισσότερα στο κεφάλαιο 4.

σημαίνει ότι το ύψος του πραγματικού κατά κεφαλήν ΑΕΠ μίας χώρας επεξηγεί το 49.32% της μεταβλητότητας του προσδόκιμου ζωής (**σχεδόν το ήμισυ!!!**). Βέβαια, το παρακάτω γράφημα (3.6) δείχνει ότι το συγκεκριμένο απλό γραμμικό υπόδειγμα **υπερεκτιμά** τα χαμηλά προσδόκιμα ζωής, οπότε διακρίνουμε θέματα εξειδίκευσης που δεν φαίνονται άμεσα από την όποια τιμή του συνελεστή προσδιορισμού.



Γράφημα 3.6: Διάγραμμα πραγματικών και εκτιμημένων τιμών του προσδόκιμου ζωής με ερμηνευτική μεταβλητή την rgdpcap και συντελεστή προσδιορισμού $R^2 = 0.4932$.

ΚΕΦΑΛΑΙΟ 4

Περαιτέρω εξειδίκευση του υποδείγματος

4.1 Ο χρόνος ως ερμηνευτική μεταβλητή

Έστω ότι η ερμηνευτική μεταβλητή είναι ο ίδιος ο χρόνος, δηλαδή θέτουμε

$$X_t = t \quad , \quad t = 1, ..., T$$

όπου T το μέγεθος του δείγματος και γράφουμε το απλό γραμμικό υπόδειγμα ως

$$Y_t = \alpha + \beta \cdot t + u_t \quad , \quad E(u_t) = 0 \quad , \quad t = 1, ..., T \tag{4.1}$$

Το υπόδειγμα (4.1) είναι «μη θεωρητικό» αφού καμμία μεταβλητή **πλην του χρόνου** δεν εισέρχεται ως ερμηνευτική, ενώ η γραμμική συνάρτηση

$$\alpha + \beta \cdot t$$

απλώς «περιγράφει» την **ανοδική (όταν** $\beta > 0$**)** ή **καθοδική (όταν** $\beta < 0$**)** πορεία της αναμενόμενης τιμής της μεταβλητής y_t στο χρόνο. Άρα, ένα υπόδειγμα της μορφής (4.1) εφαρμόζεται (εκτιμάται) σε περιπτώσεις χρονοσειρών Y_t οι οποίες εμφανίζουν (ανοδική ή καθοδική) «τάση».

Μάλιστα, σύμφωνα με το παραπάνω υπόδειγμα, η «τάση» θα πρέπει να ικανοποιεί κάποια χαρακτηριστικά, συγκεκριμένα να είναι **γραμμική** τουλάχιστον μέσα στο υπο-εξέταση δείγμα.

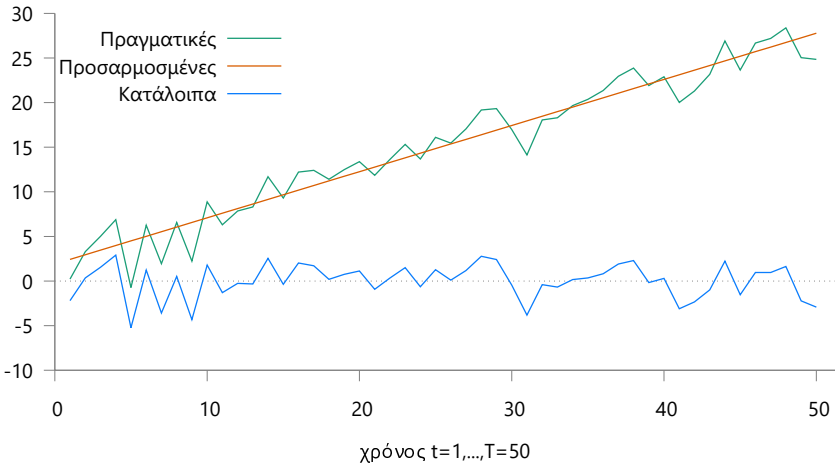
Για παράδειγμα, το παρακάτω γράφημα (4.1) παρουσιάζει μία χρονοσειρά Y_t (**πράσινη** γραμμή) και την εκτίμηση ελαχίστων τετραγώνων (**πορτοκαλί** γραμμή)

$$\hat{Y}_t = 1.908 + 0.517 \cdot t$$

με ένα δείγμα $T = 50$ παρατηρήσεων. Τα κατάλοιπα

$$\begin{aligned} \hat{u}_t &= Y_t - 1.908 - 0.517 \cdot t \\ &= Y_t - \hat{Y}_t \end{aligned}$$

παρουσιάζονται στο κάτω μέρος του γραφήματος (**μπλε** γραμμή, κενοί κύκλοι).



Γράφημα 4.1: Χρονοσειρά Y_t (**πράσινη** γραμμή) και οι προσαρμοσμένες τιμές $\hat{Y}_t = 1.908 + 0.517t$ (**πορτοκαλί** γραμμή) μετά την εκτίμηση ελαχίστων τετραγώνων. Στο κάτω πλαίσιο του γραφήματος εμφανίζονται τα κατάλοιπα της εκτίμησης $\hat{u}_t = Y_t - 2.16 - 0.96t$ (**μπλε** γραμμή)

Με βάση την υπόθεση ότι η αναμενόμενη τιμή του διαταρακτικού όρου είναι $E(u_t) = 0$, η αναμενόμενη τιμή της χρονοσειράς Y_t δίνεται από την εξίσωση

$$E\left(Y_t\right) = \alpha + \beta t$$

είναι δηλαδή μία απλή γραμμική συνάρτηση του χρόνου. Άρα για **μοναδιαίες μεταβολές του χρόνου**

$$\Delta t = t - (t-1) = 1$$

δηλαδή για μεταβολές από τη μία περίοδο στην άλλη ή από το χρόνο $t-1$ στο χρόνο t, η μέση μεταβολή της εξαρτημένης μεταβλητής είναι ίση με

$$\frac{E\left(Y_t - Y_{t-1}\right)}{\Delta t} = E\left(\Delta Y_t\right) = \alpha + \beta t - \alpha - \beta\left(t-1\right) = \beta$$

Κατά συνέπεια ο συντελεστής β μετρά τη μέση μεταβολή της μεταβλητής Y_t κάθε χρονική περίοδο. Παρομοίως, η μέση μεταβολή μετά από d περιόδους δίνεται από

$$E\left(Y_t\right) - E\left(Y_{t-d}\right) = \beta d$$

4.1.1 Εμπειρικό παράδειγμα I

Για παράδειγμα, έστω το εκτιμημένο υπόδειγμα

$$Y_t = \underset{(932.65)}{2074.67} + \underset{(305.9)}{7254.1}t + \hat{u}_t \quad , \quad t = 1, ..., T$$

όπου σε παρενθέσεις αναφέρονται τυπικά σφάλματα, με την εξαρτημένη μεταβλητή Y_t να αντιστοιχεί στο ετήσιο επίπεδο των πωλήσεων (σε ευρώ) μίας επιχείρησης για 10 έτη (έστω 1997 - 2006). Άρα $t = 1, 2, .., 10$ ή $T = 10$. Η εκτίμηση του συντελεστή κλίσης $\hat{\beta}$ είναι στατιστικά σημαντική καθώς

$$t = \frac{7254.1}{305.9} = 23.71 > t_8^{0.05/2} = 2.3$$

Άρα οι ετήσιες πωλήσεις αυξάνονται 7254.1 ευρώ κατά μέσο όρο ανά έτος.

4.1.2 Εμπειρικό παράδειγμα II

Σε ένα άλλο παράδειγμα, έστω Q_t οι μηνιαίες πωλήσεις καφέ (ποσότητα καφέ, έστω αριθμός συσκευασιών συγκεκριμένου βάρους) που πουλήθηκε κάθε μήνα από τα σημεία λιανικής πώλησης μίας εταιρείας στη διάρκεια 5 ετών και συγκεκριμένα από τον Οκτώβριο του 2002 μέχρι τον Σεπτέμβριο του 2007. Άρα έχουμε 60 μηνιαίες παρατηρήσεις, $t = 1, 2, ..., 60$.

Το εκτιμημένο υπόδειγμα δίνεται από την παρακάτω εξίσωση

$$Q_t = \underset{(455.1)}{9938.4} + \underset{(29.25)}{170.85}t + \hat{u}_t$$

Η εκτίμηση του συντελεστή κλίσης είναι στατιστικά σημαντική καθώς

$$t = \frac{170.85}{29.25} = 5.841 > t_{58}^{0.05/2} \approx 2$$

Άρα οι μηνιαίες πωλήσεις αυξάνονται κατά 171 περίπου συσκευασίες κατά μέσο όρο ανά μήνα. Στην ερώτηση *«πόσο αυξάνονται οι πωλήσεις ετήσια;»* θα απαντούσαμε $170.85 \times 12 \approx 2050$ συσκευασίες.

Τέλος (μία ερμηνεία της εκτίμησης του σταθερού όρου $\hat{\alpha}$), το μέσο επίπεδο πωλήσεων πριν την αρχή του δείγματος ή τον Σεπτέμβριο του 2002 ήταν περίπου 9938 συσκευασίες καφέ.

4.2 Λογαριθμικός - λογαριθμικός μετασχηματισμός

Συχνά θα παρατηρήσουμε στην εμπειρική οικονομετρία τη χρήση **λογαριθμισμένων μεταβλητών**

$$\text{εξαρτημένης και ανεξάρτητης}\,,\ \ln(Y_i)\ ,\ \ln(X_i)$$

αντί των αρχικών μεταβλητών Y_i και X_i. Ο εν λόγω μετασχηματισμός είναι χρήσιμος για πολλούς λόγους και φυσικά εφαρμόζεται όταν οι μεταβλητές Y_i και X_i **λαμβάνουν θετικές και μόνο τιμές**.

Εκτιμούμε λοιπόν ένα υπόδειγμα της μορφής

$$\ln(Y_i) = \alpha + \beta \ln(X_i) + u_i \tag{4.2}$$

Ο συντελεστής κλίσης β αντιστοιχεί (ceteris paribus) στην **ελαστικότητα** της Y_i ως προς τη X_i αφού

$$\beta = \frac{d\ln(Y_i)}{d\ln(X_i)} = \frac{dY_i/Y_i}{dX_i/X_i} = \varepsilon_{YX}$$

Άρα δεν χρειάζεται να προβούμε σε αλγεβρικούς υπολογισμούς για την εύρεση της ελαστικότητας, η οποία είναι **απαλλαγμένη** από τις μονάδες μέτρησης και συχνά παρέχει μία πιο άμεση και οικονομική ερμηνεία της σχέσης των μεταβλητών. Είναι περισσότερο εύλογο ότι οικονομικές μονάδες (άτομα, νοικοκυριά κτλ.) επιδεικνύουν σταθερές αντιδράσεις στις σχετικές αλλαγές μεταβλητών όπως οι τιμές και το εισόδημα παρά στις απόλυτες αλλαγές τους. Άρα το υπόδειγμα (4.2)

υποθέτει **σταθερή ελαστικότητα ή ισοελαστικότητα** της Y_i ως προς τη X_i ή ότι η αρχική εξάρτηση της Y_i πάνω στη X_i είναι πολλαπλασιαστικού τύπου $Y_i = AX_i^{\beta}e^{u_i}$.

Η ερμηνεία της παραμέτρου β είναι η εξής:

- όταν $\beta > 0$: μία **αύξηση** (μείωση) της μεταβλητής X_i κατά 1% οδηγεί σε μία β% **αύξηση** (μείωση) της μεταβλητής Y_i
- όταν $\beta < 0$: μία **αύξηση** (μείωση) της μεταβλητής X_i κατά 1% οδηγεί σε μία β% **μείωση** (αύξηση) της μεταβλητής Y_i

- Για παράδειγμα, έστω το εκτιμημένο υπόδειγμα

$$\ln(Y_i) = \underset{(0.22)}{1.76} + \underset{(0.18)}{0.75}\ln(X_i) + \hat{u}_i \quad , \quad i = 1, ..., n = 70$$

όπου σε παρενθέσεις αναφέρονται τυπικά σφάλματα. Αρχικά, ελέγχουμε τη **στατιστική σημαντικότητα** του συντελεστή κλίσης (της ελαστικότητας). Επειδή

$$t = \frac{0.75}{0.18} = 4.16 > t_{68}^{0.05/2} \approx 1.99$$

απορρίπτουμε τη μηδενική υπόθεση $H_0 : \beta = 0$ έναντι της εναλλακτικής $H_1 : \beta \neq 0$. Στη συνέχεια εξετάζουμε την **οικονομική σημασία - «οικονομική σημαντικότητα»** - της εκτιμημένης ελαστικότητας. Σύμφωνα με την εκτίμηση, κατά μέσο όρο, μία αύξηση 1% της X_i αυξάνει (θετικό πρόσημο) κατά 0.75% τη μεταβλητή Y_i. Άρα η Y_i είναι ανελαστική ως προς την X_i.

Ένα άλλο χαρακτηριστικό του λογαριθμικού μετασχηματισμού είναι η ικανότητά του **να μειώνει την ασυμμετρία και τη μεταβλητότητα** των υπο-εξέταση μεταβλητών, στοιχεία τα οποία είναι κοινά στα οικονομικά δεδομένα. Για παράδειγμα, αν ο διαταρακτικός όρος στο υπόδειγμα (4.2) κατανέμεται κανονικά, $u_i \sim N.i.d\left(0, \sigma^2\right)$, τότε η **λογαριθμική μεταβλητή** $\ln(Y_i)$ έχει:

- **αναμενόμενη τιμή**

$$E(\ln(Y_i)) = \mu_i = \alpha + \beta\ln(X_i)$$

- και **διακύμανση**

$$Var\left(\ln\left(Y_i\right)\right) = \sigma^2$$

ενώ η αρχική μεταβλητή Y_i έχει:

- **αναμενόμενη τιμή** (βλ. άσκηση 1(β) του παραρτήματος στατιστικής)

$$E\left(Y_i\right) = e^{\mu_i + \frac{1}{2}\sigma^2}$$

- **διάμεσο** ίση με

$$e^{\mu_i}$$

- και **διακύμανση**

$$Var\left(Y_i\right) = \left[E\left(Y_i\right)\right]^2 \left(e^{\sigma^2} - 1\right)$$

Δηλαδή, η αρχική μεταβλητή (η μεταβλητή την οποία λογαριθμίσαμε) είναι ασύμμετρη δεξιά (θετική ασυμμετρία) αφού η διάμεσος είναι μικρότερη του μέσου, $e^{\mu_i} < e^{\mu_i + \frac{1}{2}\sigma^2}$, ενώ η διακύμανσή της είναι ανάλογη του μέσου επιπέδου της μεταβλητής.

Συχνά, στα οικονομικά δεδομένα, μεταβλητές που λαμβάνουν μόνο θετικές τιμές παρουσιάζουν δεξιά ασυμμετρία και έχουν αυξημένη μεταβλητότητα. **Τέλος**, ο λογαριθμικός μετασχηματισμός, μέσω της μείωσης του εύρους μεταβλητότητας των δεδομένων, καθιστά τις μεταβλητές του υποδείγματος λιγότερο ευαίσθητες σε τυχόν ακραίες τιμές.

4.2.1 Εμπειρικό παράδειγμα ερμηνείας εκτιμημένων συντελεστών

Από το αρχείο δεδομένων `qog_ch4.gdt` του gretl ή χρησιμοποιώντας το script αρχείο `qog_ch4.inp` εκτιμούμε με τη μέθοδο ελαχίστων τετραγώνων **τρεις** εξισώσεις: **(υπόδειγμα 1:)** με εξαρτημένη μεταβλητή την undp_hdi: Δείκτης Ανθρώπινης Ανάπτυξης του ΟΗΕ[1] και ανεξάρτητη ή ερμηνευτική μεταβλητή την wdi_gdpcappppcon2017 (το κατά κεφαλήν πραγματικό ΑΕΠ με βάση την ισοτιμία αγοραστικής δύναμης σε δολλάρια), **(υπόδειγμα 2:)** με εξαρτημένη

[1] Βλ. ιστοσελίδα `http://hdr.undp.org/en/content/human-development -index-hdi`

την undp_hdi: Δείκτης Ανθρώπινης Ανάπτυξης και ανεξάρτητη ή ερμηνευτική μεταβλητή την

$$\text{rgdpcap} = \text{wdi\_gdpcappppcon2017}/1000$$

(το κατά κεφαλήν πραγματικό ΑΕΠ με βάση την ισοτιμία αγοραστικής δύναμης σε χιλιάδες δολλάρια), αντί της wdi_gdpcappppcon2017 ώστε να διευκολύνουμε την «οικονομική ανάγνωση» του εκτιμητή κλίσης, **(υπόδειγμα 3:)** με εξαρτημένη μεταβλητή τον λογάριθμο του δείκτη ανθρώπινης ανάπτυξης l_undp_hdi και ανεξάρτητη ή ερμηνευτική μεταβλητή τον λογάριθμο l_rgdpcap της rgdpcap.

Ο Δείκτης Ανθρώπινης Ανάπτυξης **undp_hdi** (Human Development Index, HDI) αποτελεί έναν σύνθετο (σύνθεση επιμέρους δεικτών με συγκεκριμένη μεθοδολογία στάθμισης) δείκτη κοινωνικοοικονομικής ανάπτυξης και αξιολόγησης της μίας χώρας. Για τον υπολογισμό του Δείκτη λαμβάνεται υπόψη τόσο το ακαθάριστο πραγματικό κατά κεφαλήν εθνικό εισόδημα της χώρας όσο και τα αναμενόμενα έτη εκπαίδευσης, ο μέσος αριθμός ετών σχολικής φοίτησης αλλά και το προσδόκιμο ζωής κατά τη γέννηση.

Παρακάτω συνοψίζονται τα αποτελέσματα. Σε όλες τις περιπτώσεις, οι **τιμές πιθανότητας (p-τιμές)** είναι πολύ μικρότερες του 0.01 άρα όλοι οι εκτιμημένοι συντελεστές είναι στατιστικά σημαντικοί. **Ποια όμως η ερμηνεία τους;**

Υπόδειγμα 1:

$$\widehat{\text{undp\_hdi}} = \underset{[0.0000]}{0.604361} + \underset{[0.0000]}{0.000005784} \cdot \text{wdi\_gdpcappppcon2017}$$

$$T = 178\ ,\ R = 0.6266\ ,\ \hat{\sigma} = 0.090520$$

Σημείωση: προσοχή σε αγκύλες ή ορθογώνιες παρενθέσεις αναφέρονται **p-τιμές**

Σύμφωνα με τις εκτιμήσεις του **υποδείγματος 1**, μία αύξηση στο πραγματικό κατά κεφαλήν ΑΕΠ κατά 1 μονάδα (1 «διεθνές» δολλάριο) αυξάνει τον δείκτη ανθρώπινης ανάπτυξης undp_hdi κατά 0.000005784 μονάδες. Είναι εμφανές ότι η χρήση των δεδομένων χωρίς προηγούμενη επεξεργασία (στην αρχική μορφή τους δηλαδή) δεν προσφέρεται για ευκολονόητη οικονομική ανάλυση. **Προσοχή**, διότι ο εκτιμημένος συντελεστής κλίσης $\hat{\beta} = 0.000005784$ είναι **στατιστικά σημαντικός** άρα αν και φαίνεται «πολύ μικρός» και κοντά στο μηδέν, δεν είναι

μηδενικός ή απορρίπτεται η μηδενική υπόθεση $H_0 : \beta = 0$ και είναι στατιστικά σημαντικός. Η μεταβλητή δίνει τα παρακάτω περιληπτικά στατιστικά

Περιληπτικά στατιστικά, χρησιμοποιώντας τις παρατηρήσεις 1 - 194
για τη μεταβλ. undp_hdi (186 έγκυρες παρατηρ.)

Μέσος	Διάμεσος	Ελάχιστο	Μέγιστο
0,71677	0,73800	0,38600	0,95400
Τυπ.Αποκλ.	**Συντ.μτβλ.**	**Ασυμμετρία**	**Περ.κύρτωση**
0,15012	0,20944	−0,33615	−0,89347
5% Εκατοστ.	**95% Εκατοστ.**	**IQ Εύρος**	**Απούσες τιμ.**
0,44880	0,93565	0,23275	8

Έχει μέση τιμή 0.71 μονάδες (εύρος δείκτη από το 0 μέχρι το 1) με τυπική απόκλιση 0.15 μονάδες. Θα ήταν καλύτερα (από θέμα ερμηνείας) λοιπόν να μεταβάλλουμε τις μονάδες της wdi_gdpcappppcon2017 διαιρώντας με 1000 άρα εκφράζοντας τη μεταβλητή σε χιλιάδες διεθνή δολλάρια. Αυτό πράττουμε παρακάτω στο υπόδειγμα 2.

Υπόδειγμα 2:

$$\widehat{\text{undp\_hdi}} = \underset{[0.0000]}{0.604361} + \underset{[0.0000]}{0.005784} \cdot \text{rgdpcap}$$

$$T = 178 \ , \ R^2 = 0.6266 \ , \ \hat{\sigma} = 0.090520$$

Σημείωση: προσοχή σε αγκύλες ή ορθογώνιες παρενθέσεις αναφέρονται **p-τιμές**

Σύμφωνα με τις εκτιμήσεις του **υποδείγματος 2**, μία αύξηση στο πραγματικό κατά κεφαλήν ΑΕΠ κατά 1 μονάδα (1000 «διεθνή» δολλάρια) αυξάνει τον δείκτη ανθρώπινης ανάπτυξης undp_hdi κατά 0.005784 μονάδες.

Παρατηρούμε ότι η **διαίρεση** της ερμηνευτικής μεταβλητής με 1000 **πολλαπλασιάζει** τον εκτιμημένο συντελεστή με 1000 (και το τυπικό του σφάλμα. Όλα τα άλλα αποτελέσματα μένουν ίδια). Αντίστοιχα, ο **πολλαπλασιασμός** της εξαρτημένης (π.χ. με 100 ώστε να εκφραστεί σε εύρος 0 - 100 μονάδων), **πολλαπλασιάζει** την εκτίμηση του συντελεστή κλίσης, του τυπικού του σφάλ-

ματος αλλά και τον σταθερό όρο. Με έναν τέτοιο μετασχηματισμό θα καταλήγαμε στο ίσως πιο «διαισθητικό» αποτέλεσμα ότι 1000 επιπλέον δολλάρια αυξάνουν τον δείκτη ανθρώπινης ανάπτυξης κατά 0.57 μονάδες (σχεδόν μισή μονάδα).

Παρακάτω, λογαριθμίζουμε και τις δύο μεταβλητές ώστε (είναι και οι δύο μόνο θετικές άλλωστε) να «ξεφορτωθούμε» τις μονάδες μέτρησης.

Υπόδειγμα 3:

$$\widehat{\text{l\_undp\_hdi}} = \underset{[0.0000]}{-0.798207} + \underset{[0.0000]}{0.182514} \cdot \text{l\_rgdpcap}$$

$$T = 178 \ , \ R^2 = 08837 \ , \ \hat{\sigma} = 0.075712$$

Σημείωση: προσοχή σε αγκύλες ή ορθογώνιες παρενθέσεις αναφέρονται **p-τιμές**

Καλύτερα (ερμηνευτικά), σύμφωνα με τις εκτιμήσεις του **υποδείγματος 3**, μία αύξηση του πραγματικού κατά κεφαλήν ΑΕΠ κατά 1% υποδηλώνει αύξηση του δείκτη ανθρώπινης ανάπτυξης κατά 0.18%.

4.3 Λογαριθμικός - γραμμικός μετασχηματισμός

Σύμφωνα με τον ημιλογαριθμικό μετασχηματισμό, σε ορισμένες περιπτώσεις λογαριθμίζουμε μόνο την εξαρτημένη μεταβλητή

$$\ln(Y_i) = \alpha + \beta X_i + u_i$$

Στην περίπτωση αυτή η ελαστικότητα της Y ως προς τη X δίνεται, για δεδομένη παρατήρηση i, από τον τύπο

$$\varepsilon_{YX} = \beta X_i$$

αφού

$$\frac{d\ln(Y_i)}{dX_i} = \beta$$

και

$$\varepsilon_{YX} = \frac{d\ln(Y_i)}{d\ln(X_i)} = \frac{d\ln(Y_i)}{dX_i/X_i} = \left(\frac{d\ln(Y_i)}{dX_i}\right) X_i = \beta X_i$$

δηλαδή **η ελαστικότητα είναι ανάλογη του επιπέδου της μεταβλητής** X.

Η ερμηνεία της παραμέτρου β είναι η εξής:

- όταν $\beta > 0$: μία **αύξηση** (μείωση) της μεταβλητής X_i κατά μία μονάδα οδηγεί σε μία $(100 \times \beta)\%$ **αύξηση** (μείωση) της μεταβλητής Y_i
- όταν $\beta < 0$: μία **αύξηση** (μείωση) της μεταβλητής X_i κατά μία μονάδα οδηγεί σε μία $(100 \times \beta)\%$ **μείωση** (αύξηση) της μεταβλητής Y_i

Μετά την εφαρμογή της μεθόδου των ΕΤ, η **εκτιμημένη ελαστικότητα** υπολογίζεται με βάση το δειγματικό αριθμητικό μέσο της ερμηνευτικής μεταβλητής δηλαδή

$$\hat{\varepsilon}_{YX} = \hat{\beta}\bar{X}$$

Η χρήση του ημι-λογαριθμικού μετασχηματισμού είναι ευρέως διαδεδομένη ειδικά σε περιπτώσεις που η ανεξάρτητη μεταβλητή έχει μονάδα μέτρησης το χρόνο ή είναι **ψευδομεταβλητή (δηλαδή λαμβάνει μόνο τις τιμές 0 και 1)** ή λαμβάνει ακέραιες τιμές και η εξαρτημένη μεταβλητή έχει τα γνωστά χαρακτηριστικά που επιτρέπουν να λογαριθμίσουμε, δηλαδή θετικές τιμές μόνο και κατά περίπτωση θετική ασυμμετρία και/ή μεταβλητότητα που εξαρτάται από το μέσο επίπεδο της μεταβλητής.

4.3.1 Εμπειρικό παράδειγμα ερμηνείας εκτιμημένων συντελεστών

Για παράδειγμα, ο μισθός w_i (έστω ετήσιος πραγματικός μισθός σε ευρώ) υποδειγματοποιείται συχνά ως συνάρτηση της εκπαίδευσης (ερμηνευτική μεταβλητή η εκπαίδευση) με βάση τη σχέση

$$\ln(w_i) = \alpha + \beta \times \text{ΕΚΠ}_i + u_i$$

όπου η μεταβλητή της εκπαίδευσης ΕΚΠ_i μετριέται σε έτη φοίτησης των ατόμων του δείγματος στην πρωτοβάθμια, δευτεροβάθμια και τριτοβάθμια εκπαίδευση. Εδώ, η παράμετρος β αντιπροσωπεύει την ceteris paribus $(100 \times \beta)\%$ μεταβολή στον μισθό w_i από ένα επιπλέον έτος εκπαίδευσης. Αν το υπόδειγμα λάμβανε τη

μορφή

$$w_i = \alpha + \beta \times \text{ΕΚΠ}_i + u_i$$

τότε η παράμετρος β θα αντιστοιχούσε στα επιπλέον ευρώ που κερδίζει κατά μέσο όρο ο εργαζόμενος του δείγματος από ένα επιπλέον έτος εκπαίδευσης.

Σε ένα εμπειρικό παράδειγμα, έστω ότι εκτιμήσαμε το παρακάτω υπόδειγμα

$$\ln(w_i) = \underset{(2.568)}{7.912} + \underset{(0.0095)}{0.023} \times \text{ΕΚΠ}_i + \hat{u}_i \ , \ i = 1, ..., 580 \ , \ \overline{\text{ΕΚΠ}} = 11$$

υιοθετώντας ένα τυχαίο δείγμα 580 εργαζομένων. Η εκτίμηση του συντελεστή κλίσης $\hat{\beta} = 0.023$ είναι στατιστικά σημαντική αν κρίνουμε από τη στατιστική t-student

$$t_{\hat{\beta}} = \frac{0.023}{0.0095} = 2.421$$

οπότε ένα επιπλέον έτος εκπαίδευσης αυξάνει κατά μέσο όρο τον πραγματικό ετήσιο μισθό κατά 2.3%. Η εκτιμημένη ελαστικότητα του μισθού ως προς την εκπαίδευση είναι

$$\hat{\varepsilon}_{w,\text{ΕΚΠ}} = \hat{\beta} \times \overline{\text{ΕΚΠ}} = 0.023 \times 11 = 0.253$$

Δηλαδή μία αύξηση 1% στην εκπαίδευση (όπως αυτή μετριέται) οδηγεί κατά μέσο όρο σε αύξηση του ετήσιου μισθού κατά 0.253%. Άρα στη συγκεκριμένη περίπτωση ο μισθός είναι ανελαστικός ως προς την εκπαίδευση. Αν τεθεί το ερώτημα, ποια η ελαστικότητα για ένα άτομο με 8 έτη εκπαίδευσης και ποια για ένα άτομο με 16 έτη εκπαίδευσης, τότε στην πρώτη περίπτωση έχουμε

$$0.023 \times 8 = 0.184$$

ενώ στη δεύτερη

$$0.023 \times 16 = 0.368$$

Τέλος έστω ότι εκτιμήσαμε το υπόδειγμα (με ένα μέγαλο δείγμα άνω των 30 παρατηρήσεων)

$$\ln(w_i) = \underset{(2.938)}{6.334} + \underset{(0.0082)}{0.067} \times \text{ΦΥΛΟ}_i + \hat{u}_i$$

όπου η μεταβλητή ΦΥΛΟ_i είναι **ψευδομεταβλητή**, δηλαδή μετρά ένα «ποιοτικό» χαρακτηριστικό, έστω το φύλο του εργαζόμενου. Κατασκευάζουμε λοιπόν τη μεταβλητή ΦΥΛΟ_i θέτοντας

- $\Phi\Upsilon\Lambda\mathrm{O}_i = 1$ όταν το άτομο του δείγματος είναι **άνδρας** και
- $\Phi\Upsilon\Lambda\mathrm{O}_i = 0$ όταν το άτομο του δείγματος είναι **γυναίκα**

Παρατηρούμε ότι ο εκτιμημένος συντελεστής κλίσης $\hat{\beta} = 0.067$ είναι στατιστικά σημαντικός και η ερμηνεία του είναι ότι η διαφορά των μισθών ανδρών γυναικών είναι κατά μέσο όρο 6.7%, δηλαδή οι άνδρες κερδίζουν ceteris paribus 6.7% περισσότερα (πραγματικός ετήσιος μισθός) από τις γυναίκες.

4.4 Γραμμικός - λογαριθμικός μετασχηματισμός

Στην περίπτωση του γραμμικού - λογαριθμικού μετασχηματισμού

$$Y_i = \alpha + \beta \ln(X_i) + u_i$$

υποθέτουμε ότι η ελαστικότητα της Y ως προς τη X μεταβλητή είναι **αντιστρόφως ανάλογη** του επιπέδου της εξαρτημένης μεταβλητής Y. Δηλαδή καθώς αυξάνεται η εξαρτημένη μεταβλητή Y, η αντίδρασή της στην ανεξάρτητη μεταβλητή X μειώνεται.

Αναλυτικά η ελαστικότητα της Y ως προς την X δίνεται από τον τύπο

$$\varepsilon_{YX} = \frac{d\ln(Y_i)}{d\ln(X_i)} \text{ ή } = \frac{dY_i/Y_i}{d\ln(X_i)} \text{ ή } = \frac{dY_i}{d\ln(X_i)}\frac{1}{Y_i} = \beta\frac{1}{Y_i}$$

και υπολογίζεται με βάση το εκτιμημένο υπόδειγμα μέσω της εξίσωσης (αντικαθιστούμε τον αριθμητικό μέσο $\overline{\mathrm{Y}}$ όπου Y_i)

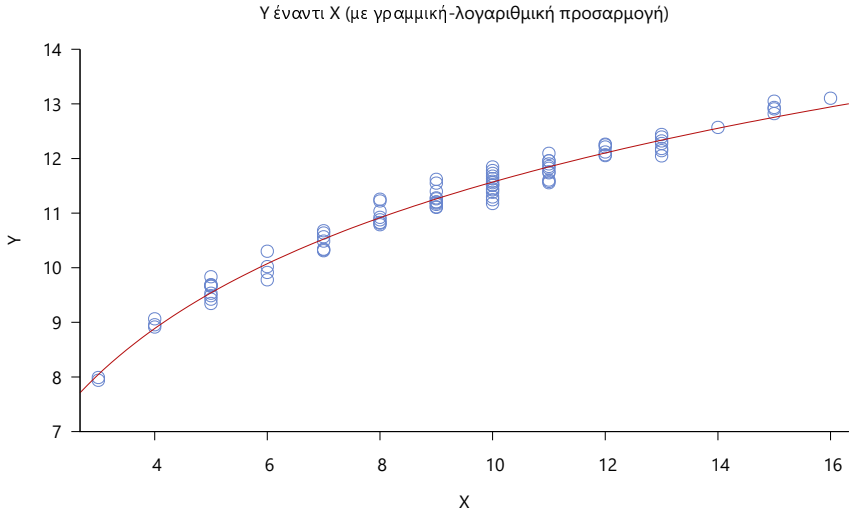
$$\hat{\varepsilon}_{YX} = \hat{\beta}\frac{1}{\overline{\mathrm{Y}}}$$

Βέβαια, με μία απλή αντικατάσταση της Y_i στον τύπο της ελαστικότητας (αφήνοντας κατά μέρος το διαταρακτικό όρο)

$$\varepsilon_{YX} = \beta\frac{1}{Y_i} = \beta\frac{1}{\alpha + \beta\ln(X_i)}$$

παρατηρούμε ότι και η σχέση U με Q είναι αντίστροφη, δηλαδή καθώς οι τιμές της ερμηνευτικής μεταβλητής X αυξάνονται, η αντίδρασή της εξαρτημένης βαίνει μειούμενη ή αλλιώς η επίδραση της X στην Y μειώνεται.

Τα παρακάτω δύο γραφήματα, (4.2) και (4.3), εμφανίζουν μία υποθετική γραμμική-λογαριθμική σχέση της εξαρτημένης μεταβλητής Y με την X. Συγκεκριμένα, $Y_i = \alpha + \beta \ln(X_i) + u_i$. Ειδικότερα, το γράφημα (4.2) εμφανίζει το διάγραμμα διασποράς των Y_i, $\hat{Y}_i$ (κάθετος άξονας) και X_i (οριζόντιος άξονας) όπου $\hat{Y}_i$ είναι οι προσαρμοσμένες τιμές $\hat{Y}_i = \hat{\alpha} + \hat{\beta} \ln(X_i)$.

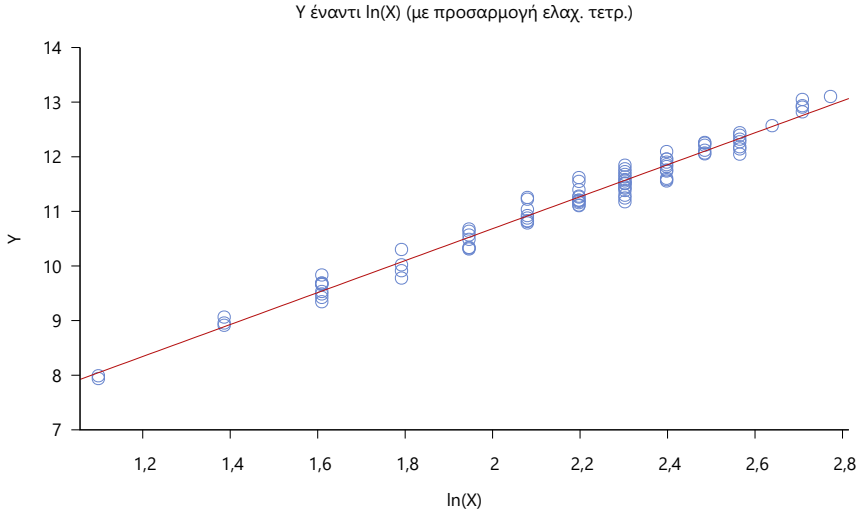


Γράφημα 4.2: Διάγραμμα διασποράς των πραγματικών τιμών Y_i, των προσαρμοσμένων ή εκτιμημένων τιμών $\hat{Y}_i = \hat{\alpha} + \hat{\beta} \ln(X_i)$ (και οι δύο μετρώνται στον κάθετο άξονα) και των τιμών της μεταβλητής X_i (οριζόντιος άξονας).

Η έντονα μη γραμμική σχέση των Y_i, X_i **«γραμμικοποιείται»** όταν εφαρμόσουμε φυσικούς λογάριθμους στην X_i, κάτι που γίνεται εμφανές στο διάγραμμα διασποράς (4.3) των Y_i, $\hat{Y}_i$ (κάθετος άξονας) και $\ln(X_i)$ (οριζόντιος άξονας). Πολλές μη-γραμμικές οικονομικές σχέσεις στην πραγματικότητα φαίνονται γραμμικές αν τις «σχεδιάσουμε» στη λογαριθμική κλίμακα.

Η ερμηνεία της παραμέτρου β είναι η εξής:

- όταν $\beta > 0$: μία **αύξηση** (μείωση) της μεταβλητής X_i κατά ένα τοις εκατό 1% οδηγεί σε μία **αύξηση** (μείωση) της μεταβλητής Y_i κατά $\beta/100$ μονάδες
- όταν $\beta < 0$: μία **αύξηση** (μείωση) της μεταβλητής X_i κατά ένα τοις εκατό 1% οδηγεί σε μία **μείωση** (αύξηση) της μεταβλητής Y_i



Γράφημα 4.3: Διάγραμμα διασποράς των Y_i, $\hat{Y}_i$ (κάθετος άξονας) και $\ln(X_i)$ (οριζόντιος άξονας) όπου $\hat{Y}_i$ είναι οι προσαρμοσμένες τιμές $\hat{Y}_i = \hat{\alpha} + \hat{\beta} \ln(X_i)$.

κατά $\beta/100$ μονάδες

4.4.1 Εμπειρικό παράδειγμα ερμηνείας εκτιμημένων συντελεστών

Επιστρέφουμε στο παράδειγμα μελέτης της σχέσης του **Δείκτη Ανθρώπινης Ανάπτυξης** (undp_hdi) και του **κατά κεφαλήν πραγματικού ΑΕΠ με βάση την ισοτιμία αγοραστικής δύναμης σε χιλιάδες δολλάρια**

$$\text{rgdpcap} = \text{wdi\_gdpcappppcon2017}/1000$$

Βασιζόμαστε στο αρχείο δεδομένων `qog_ch4.gdt` του gretl ή ακόμα καλύτερα χρησιμοποιούμε το script αρχείο `qog_ch4.inp` για να αναπαράγουμε την παρακάτω εμπειρική ανάλυση.

Η μεταβλητή του Δείκτη Ανθρώπινης ανάπτυξης undp_hdi πολλαπλασιάζεται με 100 ώστε να βρίσκεται στο διάστημα 0-100 (hdih = undp_hdi*100) διευκολύνοντας την οικονομική ερμηνεία των εκτιμημένων συντελεστών.

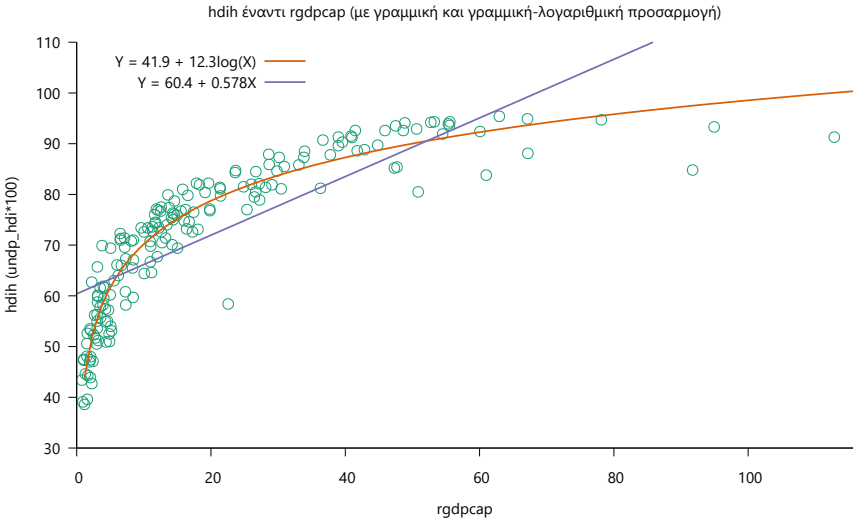
Στο **γράφημα (4.4)** με κάθετο άξονα να μετρά την εξαρτημένη μεταβλητή hdih και οριζόντιο άξονα να μετρά την ερμηνευτική μεταβλητή rgdpcap, εμφα-

νίζονται το διάγραμμα διασποράς των $Y_i = \text{hdih}$ και $X_i = \text{rgdpcap}$, η εκτίμηση ελαχίστων τετραγώνων του γραμμικού υποδείγματος

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$$
$$\hat{Y}_i = 60.4 + 0.578 \cdot X_i \ , \ R^2 = 0.6266 \ \ \text{(εκτίμηση ελαχ. τετρ.)}$$

και η εκτίμηση του γραμμικού λογαριθμικού υποδείγματος

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta} \ln(X_i)$$
$$\hat{Y}_i = 41.9 + 12.30 \cdot \ln(X_i) \ , \ R^2 = 0.9014 \ \ \text{(εκτίμηση ελαχ. τετρ.)}$$



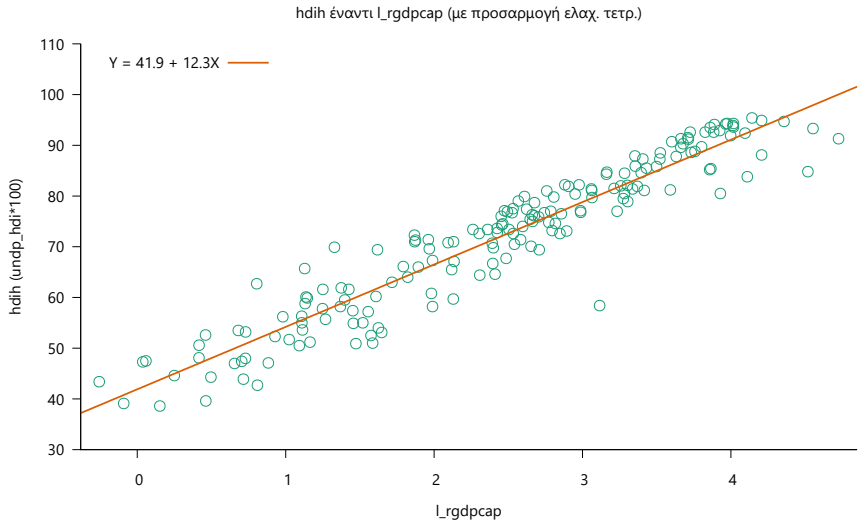
Γράφημα 4.4: Διάγραμμα διασποράς (**πράσινοι κύκλοι**) των Y_i (κάθετος άξονας) και X_i (οριζόντιος άξονας) μαζί με τις προσαρμοσμένες τιμές του γραμμικού υποδείγματος (**μπλε γραμμή**) $\hat{Y}_i = 60.4 + 0.578 \cdot X_i$ και τις προσαρμοσμένες τιμές του γραμμικού λογαριθμικού υποδείγματος (**πορτοκαλί καμπύλη**) $\hat{Y}_i = 41.9 + 12.30 \cdot \ln(X_i)$.

Το γραμμικό υπόδειγμα (μπλε γραμμή προσαρμοσμένων τιμών) φαίνεται να έχει **σημαντικά προβλήματα εξειδίκευσης** κάτι που δεν είναι εμφανές κοιτώντας - απλά - τον συντελεστή προσδιορισμού (62.66% της μεταβλητότητας του δείκτη ερμηνεύεται από τη μεταβλητότητα του κατά κεφαλήν ΑΕΠ). Συγκεκριμένα, **υπερεκτιμά** (προβλέπει υπερβολικά, overestimates ή overpredicts) τον δείκτη για τις «φτωχές» χώρες, δηλαδή χώρες με χαμηλό κατά κεφαλήν ε-

τήσιο πραγματικό ΑΕΠ. **Υποεκτιμά** (προβλέπει λιγότερο από το παρατηρήσιμο επίπεδο του δείκτη, underestimates ή underpredicts) για «μεσαία» επίπεδα ΑΕΠ/πλούτου και **υπερεκτιμά** και πάλι τον δείκτη στις «πλούσιες» χώρες.

Παρατηρούμε εύκολα (πράσινοι κύκλοι διαγράμματος διασποράς) ότι η σχέση μεταξύ του ετήσιου πραγματικού κατά κεφαλήν ΑΕΠ και του Δείκτη Ανθρώπινης Ανάπτυξης **δεν είναι γραμμική**. Αυξήσεις στο ΑΕΠ αυξάνουν με γρήγορο ρυθμό τον δείκτη ανθρώπινης ανάπτυξης σε χαμηλότερα επίπεδα κατά κεφαλήν ΑΕΠ. Όσο πιο πλούσια (σε όρους πραγματικού κατά κεφαλήν ΑΕΠ) είναι η χώρα, τόσο λιγότερο έντονο το αυξητικό αποτέλεσμα του πρόσθετου ΑΕΠ.

Στο **γράφημα (4.5)** απεικονίζονται το διάγραμμα διασποράς των Y_i και $\ln(X_i)$ με την εκτίμηση (προσαρμοσμένες τιμές) του γραμμικού λογαριθμικού υποδείγματος (πορτοκαλί γραμμή), $\hat{Y}_i = 41.9 + 12.30 \cdot \ln(X_i)$. Όπως ξεκάθαρα φαίνεται, η σχέση τώρα μετατρέπεται σε **γραμμική** και ο συντελεστής προσδιορισμού $R^2 = 0.9014$ υποδηλώνει πολύ καλύτερη προσαρμογή στα δεδομένα. Ο (στατιστικά σημαντικός συντελεστής) $\hat{\beta} = 12.30$ *«διαβάζεται»*: Μία αύξηση του πραγματικού κατά κεφαλήν ΑΕΠ 1% οδηγεί σε αύξηση του Δείκτη Ανθρώπινης Ανάπτυξης κατά 0.123 μονάδες.



Γράφημα 4.5: Διάγραμμα διασποράς των Y_i, $\hat{Y}_i$ (κάθετος άξονας) και $\ln(X_i)$ (οριζόντιος άξονας) όπου $\hat{Y}_i$ είναι οι προσαρμοσμένες τιμές $\hat{Y}_i = \hat{\alpha} + \hat{\beta}\ln(X_i)$.

Ερώτηση: Ποια η εκτιμημένη μεταβολή του Δείκτη Ανθρώπινης Ανάπτυξης

από μία αύξηση 1000 δολλαρίων στο κατά κεφαλήν ΑΕΠ (δηλαδή αύξηση μίας μονάδας στη μεταβλητή rgdpcap) όταν η χώρα είναι: (**i**) φτωχή με κατά κεφαλήν ΑΕΠ 1000 δολλάρια ετησίως, (**i**) μεσαίου επιπέδου (**ii**) με κατά κεφαλήν ΑΕΠ 20000 δολλάρια ετησίως και (**iii**) πλούσια με κατά κεφαλήν ΑΕΠ 50000 δολλάρια ετησίως ;

Απάντηση: Για κάθε μία από τις τρεις περιπτώσεις υπολογίζουμε

$$\Delta\hat{Y}_i = 12.30 \cdot (\ln(2) - \ln(1)) = 8.52$$

άρα 8.52 μονάδες στον δείκτη hdih

$$\Delta\hat{Y}_i = 12.30 \cdot (\ln(21) - \ln(20)) = 0.60$$

άρα 0.60 μονάδες στον δείκτη hdih

$$\Delta\hat{Y}_i = 12.30 \cdot (\ln(51) - \ln(50)) = 0.24$$

άρα 0.24 μονάδες στον δείκτη hdih.

Σημείωση: «Φτωχές» και «Πλούσιες» - οικονομικά - χώρες.

Το **γράφημα/ιστόγραμμα (4.7)** για τη μεταβλητή rgdpcap παράγεται στο gretl μέσω της εντολής

```
freq rgdpcap --min=0 --binwidth=5 --plot=display
```

ενώ το - **γράφημα (4.6)** - δείχνει τον πίνακα αποτελεσμάτων της κατανομής συχνοτήτων από τα οποία παράγεται το ιστόγραμμα.

Η κατανομή του δείκτη είναι εμφανώς ασύμμετρη δεξιά. Το «βάρος» της κατανομής συγκεντρώνεται αριστερά π.χ. του 20 (20000 δολλάρια ετησίως) ή του 66.29% των 178 χωρών με διαθέσιμες παρατηρήσεις. Η δεξιά ασυμμετρία φαίνεται και από την τοποθεσία της διαμέσου σε σχέση με τον αριθμητικό μέσο της μεταβλητής. Συγκεκριμένα, το μέσο κατά κεφαλήν ΑΕΠ είναι 19.94 ή περίπου 20 χιλ. δολλάρια ενώ η διάμεσος βρίσκεται αριστερά του μέσου αφού το 56.74% των παρατηρήσεων βρίσκεται κάτω από το επίπεδο των 15 χιλ. δολλαρίων. Αν υπολογίσουμε με ακρίβεια τη διάμεσο[2] λαμβάνουμε το αποτέλεσμα 12.87 χιλ. δολλάρια.

[2] Αφού έχετε εκτελέσει το script αρχείο `qog_ch4.inp`, πληκτρολογήστε στην κονσόλα του gretl: eval median(rgdpcap)

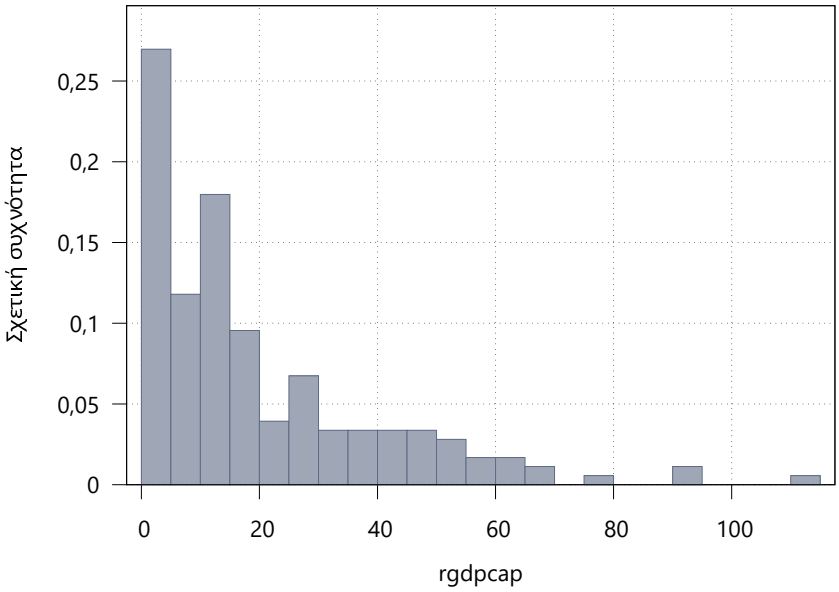
Κατανομή συχνότητας για rgdpcap, παρατηρήσεις 1-178
αριθμός κλάσεων = 23, μέσος = 19,9423, τυπ.αποκλ. = 20,2174

διάστημα	κεντρ. τιμή	συχν.	σχετ.	αθροιστ
< 5,0000	2,5000	48	26,97%	26,97%
5,0000 - 10,000	7,5000	21	11,80%	38,76%
10,000 - 15,000	12,500	32	17,98%	56,74%
15,000 - 20,000	17,500	17	9,55%	66,29%
20,000 - 25,000	22,500	7	3,93%	70,22%
25,000 - 30,000	27,500	12	6,74%	76,97%
30,000 - 35,000	32,500	6	3,37%	80,34%
35,000 - 40,000	37,500	6	3,37%	83,71%
40,000 - 45,000	42,500	6	3,37%	87,08%
45,000 - 50,000	47,500	6	3,37%	90,45%
50,000 - 55,000	52,500	5	2,81%	93,26%
55,000 - 60,000	57,500	3	1,69%	94,94%
60,000 - 65,000	62,500	3	1,69%	96,63%
65,000 - 70,000	67,500	2	1,12%	97,75%
70,000 - 75,000	72,500	0	0,00%	97,75%
75,000 - 80,000	77,500	1	0,56%	98,31%
80,000 - 85,000	82,500	0	0,00%	98,31%
85,000 - 90,000	87,500	0	0,00%	98,31%
90,000 - 95,000	92,500	2	1,12%	99,44%
95,000 - 100,00	97,500	0	0,00%	99,44%
100,00 - 105,00	102,50	0	0,00%	99,44%
105,00 - 110,00	107,50	0	0,00%	99,44%
>= 110,00	112,50	1	0,56%	100,00%

Γράφημα 4.6: Κατανομή συχνοτήτων μεταβλητής rgdpcap

4.5 Χρονοσειρές: Στασιμότητα και μη στασιμότητα

Η εκτίμηση **υποδειγμάτων χρονοσειρών** κρύβει τα δικά της μυστικά και έχει ιδιαιτερότητες που δεν απαντώνται στα υποδείγματα διαστρωματικών δεδομένων. Η έννοια της **στασιμότητας** είναι ειδοποιός αν θέλουμε να προχωρήσουμε σε εκτίμηση με **στατιστικά αξιόπιστα αποτελέσματα**. Ειδάλλως, τόσο οι εκτιμητές όσο και οι έλεγχοι υποθέσεων μπορεί να δίνουν είτε φτωχά είτε πλασματικά αποτελέσματα, άρα μη-χρήσιμα ή πλασματικά αποτελέσματα. Αν αντιμετωπίζουμε μη-στάσιμες χρονοσειρές, τότε **είτε** θα τις μεταμορφώσουμε σε στάσιμες είτε θα υιοθετήσουμε έναν διαφορετικό, αρκετά προχωρημένο, τρόπο στατιστικής επαγωγής που διαφέρει από τον «κλασσικό». Μαζί με την έννοια



Γράφημα 4.7: Ιστόγραμμα (histogram) μεταβλητής rgdpcap.

της στασιμότητας, θα αναφέρουμε και την **εργοδικότητα** αφού και αυτή συνεισφέρει στην αξιόπιστη στατιστική επαγωγή στα υποδείγματα χρονοσειρών.

4.5.1 Στασιμότητα

Χρονοσειρά ή χρονολογική σειρά είναι μία χρονικά διατεταγμένη ακολουθία παρατηρήσεων μίας ακολουθίας τυχαίων μεταβλητών (δείτε κεφάλαιο 1 του βιβλίου). Στο απλό γραμμικό υπόδειγμα

$$Y_t = \alpha + \beta X_t + u_t$$

οι Y_t και X_t αποτελούν παρατηρήσιμες χρονοσειρές ενώ ο διαταρακτικός όρος u_t και άλλες μη παρατηρήσιμες ή υποθετικές μεταβλητές, επίσης καλούνται χρονοσειρές.

Για δεδομένο χρόνο t, η μεταβλητή Y_t θεωρείται τυχαία. Η ιδιομορφία των χρονοσειρών έγκειται στο ότι αποτελούν μία οικογένεια τυχαίων μεταβλητών που εξελίσσονται στο χρόνο (**στοχαστική διαδικασία**[3]). Έτσι θεωρούμε ότι οι

[3]Stochastic process

«παρατηρήσεις»

$$\{..., Y_{t-2}, Y_{t-1}, Y_t, Y_{t+1}, Y_{t+2}, ...\} \quad \text{ή} \quad \{Y_t\}_{-\infty}^{+\infty} \quad \text{ή} \quad \{Y_t\}$$

προέρχονται από μία και μόνο «πραγματοποίηση» ενός τυχαίου πειράματος, ενώ συνήθως οι οικονομολόγοι παρατηρούν όχι απλώς μία πραγματοποίηση αλλά και ένα **πεπερασμένο τμήμα της ακολουθίας**

$$\text{π.χ.,} \quad \{Y_1, ..., Y_T\}$$

Για παράδειγμα, η χρονοσειρά του ετήσιου πληθωρισμού στην Ελλάδα από το 1950 μέχρι το 2011 είναι μία ακολουθία 62 τιμών (παρατηρήσεων) και αποτελούν τμήμα μίας και μόνο πιθανής πραγματοποίησης της ακολουθίας του πληθωρισμού. Αν παρατηρούσαμε τον κόσμο εξαρχής και συλλέγαμε τιμές για την ίδια χρονοσειρά τότε θα είχαμε στη διάθεσή μας και μία δεύτερη πραγματοποίηση. Άρα, έχοντας αρκετές πραγματοποιήσεις θα μπορούσαμε να εκτιμήσουμε το μέσο πληθωρισμό, π.χ., για το έτος 1992, αθροίζοντας τις παρατηρήσεις που αντιστοιχούν στο εν λόγω έτος και διαιρώντας με τον αριθμό των παρατηρήσεων. Ο συγκεκριμένος μέσος ονομάζεται «συνολικός μέσος» (ensemble average).

Αντιλαμβάνεστε βέβαια ότι κάτι τέτοιο είναι μη εφικτό. Αν όμως θεωρήσουμε ότι η κατανομή που παράγει τις παρατηρήσεις είναι αμετάβλητη στον χρόνο, τότε η μία και μόνο πραγματοποίηση που διαθέτουμε μπορεί να θεωρηθεί ότι προέρχεται από την ίδια κατανομή. Αν η διαδικασία **δεν έχει** «υπερβολική» **εμμονή (persistence)** δηλαδή συσχέτιση - κάτι που θα ονομάσουμε **εργοδικότητα** - τότε κάθε στοιχείο της πραγματοποίησης θα περιέχει μοναδική πληροφόρηση και ο χρονικός μέσος (άθροισμα παρατηρήσεων από το 1950 μέχρι το 2011 διά 62) παρέχει συνεπή[4] εκτίμηση του «συνολικού μέσου».

Η φύση λοιπόν των χρονοσειρών - (1) μία μόνο πραγματοποίηση, (2) η Y_{t-1} προηγείται της Y_t ενώ έπεται η Y_{t+1} κ.ο.κ - είναι τέτοια που αναμένουμε γενικά οι Y_t να μην είναι ανεξάρτητες και ειδικότερα να συσχετίζονται τουλάχιστον γραμμικά, $Cov(Y_t, Y_s) \neq 0$. Άρα η κλασσική υπόθεση της ανεξαρτησίας (π.χ., των διαταρακτικών όρων) είναι πολύ αυστηρή πόσο μάλλον όταν αναφέρεται σε οικονομικές χρονοσειρές. Στη συντριπτική τους πλειοψηφία (αν όχι όλες) οι μακροοικονομικές και χρηματοοικονομικές χρονοσειρές δεν μπορούν να χαρακτηριστούν

[4] Η συνέπεια αποτελεί ιδιότητα των εκτιμητών. Στην οικονομετρία, η συνέπεια αποτελεί την **ελάχιστη προϋπόθεση** που πρέπει να πληροί ένας εκτιμητής ώστε να θεωρηθεί στατιστικά ικανοποιητικός.

ανεξάρτητες, δηλαδή ως μία ακολουθία ανεξάρτητων τυχαίων μεταβλητών.

Για όλους τους παραπάνω λόγους, γενικεύουμε την τάξη των υπο-εξέταση χρονοσειρών και γνωρίζουμε ακολουθίες τυχαίων μεταβλητών με πλουσιότερες ιδιότητες από αυτές των ανεξάρτητων και ομοιογενώς κατανεμόμενων μεταβλητών (i.i.d)[5] ή απλώς ανεξάρτητων μεταβλητών (i.n.i.d)[6].

Αυστηρή στασιμότητα (strict stationarity).

Η Y_t είναι αυστηρώς στάσιμη αν η από κοινού συνάρτηση πυκνότητας πιθανότητας των

$$\{Y_{t-k}..., Y_{t-2}, Y_{t-1}, Y_t, Y_{t+1}, Y_{t+2}, ..., Y_{t+k}\}$$

είναι ανεξάρτητη του χρόνου t για όλα τα k. Είναι **η σχετική θέση και όχι η απόλυτη θέση** που είναι σημασίας για την κατανομή. Άρα, για παράδειγμα: **(α)** η από κοινού κατανομή των $\{Y_6, Y_{10}\}$ σε μία αυστηρώς στάσιμη διαδικασία θεωρείται ίδια με την από κοινού κατανομή των $\{Y_{21}, Y_{25}\}$ ή **(β)** η από κοινού κατανομή των $\{Y_1, Y_3, Y_5\}$ είναι ίδια με αυτή των $\{Y_2, Y_4, Y_6\}$ κ.ο.κ.

Στις **αυστηρώς στάσιμες χρονοσειρές**, η αναμενόμενη τιμή, η διακύμανση και αυτοσυνδιακύμανση και **όλες** οι κεντρικές ροπές υψηλότερης τάξης (π.χ., ασυμμετρία, κύρτωση και άλλες) δεν εξαρτώνται από την απόλυτη τιμή του χρόνου t, δηλαδή δεν είναι συναρτήσεις του χρόνου.

Παράδειγμα. Μία ακολουθία **i.i.d** μεταβλητών είναι **αυστηρώς στάσιμη**. Άρα ο διαταρακτικός όρος του απλού γραμμικού υποδείγματος με βάση τις κλασσικές υποθέσεις (κατανέμεται ως **N.i.d**) είναι μία **αυστηρώς στάσιμη διαδικασία**.

Η ιδιότητα της στασιμότητας μίας χρονοσειράς $\{Y_t\}$ επεκτείνεται και σε κατάλληλες συναρτήσεις της χρονοσειράς, π.χ., και η $\{Y_t^2\}$ είναι στάσιμη. Προσοχή, διότι ακόμα και αν όλα τα στοιχεία ενός διανύσματος είναι μεμονωμένα στάσιμα, το διάνυσμα ως σύνολο μπορεί να μην είναι στάσιμο.

[5] **I**ndependently and **i**dentically **d**istributed.
[6] **I**ndependently and **n**on **i**dentically **d**istributed.

Ασθενής ή κατά συνδιακύμανση στασιμότητα (weak or covariance stationarity).

Η $\{Y_t\}$ είναι κατά συνδιακύμανση (ασθενώς) στάσιμη αν η **αναμενόμενη τιμή**, η **διακύμανση** και η **αυτοσυνδιακύμανση** της χρονοσειράς Y_t είναι συναρτήσεις ανεξάρτητες του χρόνου.

Συγκεκριμένα:

- $E(Y_t) = \mu_Y$: δεν είναι συνάρτηση του χρόνου, π.χ., όπως όταν $\mu_Y(t)$
- $Var(Y_t) = \sigma_Y^2 = \gamma_Y(0)$: δεν είναι συνάρτηση του χρόνου, π.χ., όπως όταν $\gamma_Y(t, 0)$
- $Cov(Y_t, Y_{t-k}) = \gamma_Y(k)$, $\forall\, k = ..., -2, -1, 0, 1, 2, ...$: και όχι $\gamma_Y(t, k)$

Παρατηρήστε ότι η αυτοσυνδιακύμανση $Cov(Y_t, Y_{t-k})$ μπορεί να είναι συνάρτηση της χρονικής απόστασης $t - (t - k) = k$ αλλά όχι του χρόνου t. Επίσης, για διευκόλυνση ορίζουμε την αυτοσυνδιακύμανση ως μία συνάρτηση $\gamma()$. Λόγω στασιμότητας, έχουμε συμμετρία στη συνάρτηση αυτοσυνδιακύμανσης, δηλαδή $\gamma(k) = \gamma(-k)$. Επίσης, όταν $k = 0$ τότε η αυτοσυνδιακύμανση ταυτίζεται με τη διακύμανση της χρονοσειράς, αφού

$$Cov(Y_t, Y_t) = Var(Y_t) = \sigma_Y^2 = \gamma_Y(0)$$

Η έννοια της στασιμότητας δεν υπονοεί αυτόματα και ομοιογένεια στην εμφάνιση της χρονοσειράς (κατά την κίνησή της στον χρόνο) ή την έλλειψη περιοδικών σχηματισμών. Το βασικό χαρακτηριστικό των ασθενώς στάσιμων χρονοσειρών είναι ότι οι όποιοι σχηματισμοί (patterns) δεν εξαρτώνται συστηματικά από τον δείκτη του χρόνου.

Η συνάρτηση

$$\rho_Y(k) = \frac{\gamma_Y(k)}{\gamma_Y(0)}$$

ονομάζεται **συνάρτηση αυτοσυσχέτισης**[7] (autocorrelation function, ACF)

[7] Η **αυτοσυσχέτιση** καλείται και **σειριακή συσχέτιση** (serial correlation). Για παράδειγμα, αντί της έκφρασης *«η Y_t δεν αυτοσυσχετίζεται»* μπορεί να δείτε την έκφραση *«η Y_t δεν συσχετίζεται σειριακά»*. Αντίστοιχα η συνάρτηση $Cov(Y_t, Y_{t-k})$ καλείται **συνάρτηση αυτοσυνδιακύμανσης** (autocovariance function).

και αποδεικνύεται ότι

$$-1 \leq \rho_Y(k) \leq 1$$

Το «διάγραμμα» της $\rho_Y(k)$ ως προς το k ονομάζεται **κορρελόγραμμα** (correlogram). Όταν είναι σαφές ότι αναφερόμαστε στη χρονοσειρά Y_t μπορούμε να παραλείπουμε τον υποδείκτη Y από το συμβολισμό των συναρτήσεων/ροπών, π.χ., γράφουμε $\mu, \gamma(k), \rho(k)$ αντί $\mu_Y, \gamma_Y(k), \rho_Y(k)$.

Η δειγματική συνάρτηση αυτοσυσχέτισης εκτιμάται μέσω του

$$\hat{\rho}_Y(k) = \frac{\sum_{t=k+1}^{T} \left(Y_t - \overline{Y}\right)\left(Y_{t-k} - \overline{Y}\right)}{\sum_{t=1}^{T} \left(Y_t - \overline{Y}\right)^2}$$

ενώ ένα «προσεγγιστικό» τυπικό σφάλμα δίνεται από $1/\sqrt{T}$.

Σημείωση: Κάτω από συγκεκριμένες στατιστικές προϋποθέσεις και τη μηδενική υπόθεση της μη-αυτοσυσχέτισης $H_0 : \rho_Y(k) = 0$ σε οποιαδήποτε χρονική υστέρηση $k \geq 1$ αποδεικνύεται ότι

$$\hat{\rho}_Y(k) \overset{\alpha}{\sim} N\left(0, \frac{1}{T}\right)$$

Θα γίνει κατανοητό σε επόμενες διαλέξεις και ειδικότερα στο κεφάλαιο 8 γιατί χρησιμοποιούμε τον όρο «προσεγγιστικό».

Όταν λοιπόν η τιμή της $\hat{\rho}_Y(k)$ είναι μεγαλύτερη σε απόλυτους όρους από δύο φορές το τυπικό σφάλμα, δηλαδή $|\hat{\rho}_Y(k)| > \frac{2}{\sqrt{T}}$, θεωρούμε ότι η εκτίμηση $\hat{\rho}_Y(k)$ είναι στατιστικά σημαντική ή ότι απορρίπτουμε τη μηδενική υπόθεση

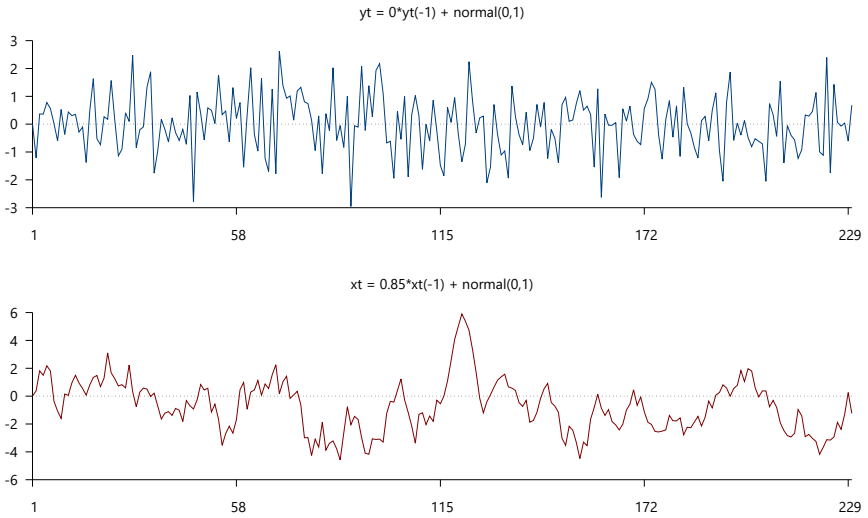
$$H_0 : \rho_Y(k) = 0 \ , \ k \geq 1$$
$$H_1 : \rho_Y(k) \neq 0 \ , \ k \geq 1$$

σε επίπεδο σημαντικότητας 5% που θέλει την υποκείμενη χρονοσειρά να μην αυτοσυσχετίζεται γραμμικά.

Για παράδειγμα το παρακάτω γράφημα (4.8) εμφανίζει στο άνω πλαίσιο μία χρονοσειρά[8] Y_t με $T = 230$ παρατηρήσεις, η οποία «φαίνεται» ότι δεν παρουσι-

[8] Η χρονοσειρά δημιουργήθηκε με βάση την $Y_t \sim N.i.d(0,1)$

άζει αυτοσυσχέτιση, δηλαδή δεν διακρίνουμε άμεσα «δομές» ή «σχηματισμούς» ή «τάσεις» στα δεδομένα. Σε αντίθεση, στο κάτω πλαίσιο του γραφήματος (4.8), η χρονοσειρά X_t παρουσιάζει «δομές» δηλαδή επίμονες αυξητικές ή μειωτικές τάσεις και υποπτευόμαστε ότι μπορεί να αυτοσυσχετίζεται.



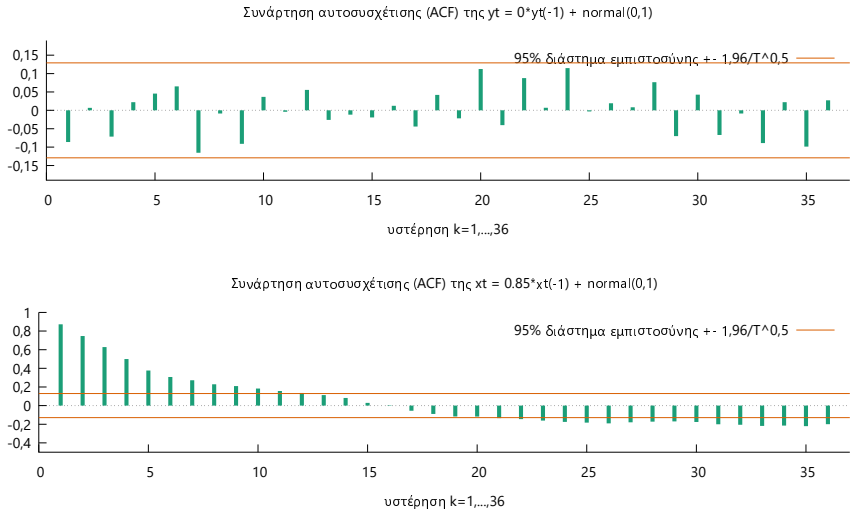
Γράφημα 4.8: Χρονοσειρά χωρίς αυτοσυσχέτιση Y_t (**μπλε γραμμή**), και μία χρονοσειρά με αυτοσυσχέτιση X_t (**κόκκινη γραμμή**).

Εκτιμούμε λοιπόν και στις δύο περιπτώσεις τη **δειγματική συνάρτηση αυτοσυσχέτισης** $\hat{\rho}(k)$ για υστερήσεις $k = 1, ..., 36$ όπου η τελική υστέρηση 36 επιλέχθηκε αυθαίρετα. Παρουσιάζουμε τις εκτιμημένες συναρτήσεις στο παρακάτω γράφημα (4.9) μαζί με τις **οριακές γραμμές** ± 2 φορές το τυπικό σφάλμα κάθε εκτιμημένου συντελεστή αυτοσυσχέτισης $\hat{\rho}(k)$, δηλαδή τις γραμμές

$$\pm\frac{2}{\sqrt{T}} = \pm\frac{2}{\sqrt{230}} = \pm 0.13187 \quad (95\% \text{ διάστημα εμπιστοσύνης})$$

Όταν οι εκτιμημένες αυτοσυσχετίσεις βρίσκονται **εντός των ορίων** τότε οι αντίστοιχες τιμές στον πληθυσμό θεωρούνται μηδενικές, δηλαδή οι εκτιμήσεις των $\hat{\rho}(k)$ είναι **στατιστικά μη σημαντικές**.

Πραγματικά λοιπόν και η στατιστική επαγωγή επιβεβαιώνει ότι η χρονοσειρά Y_t δεν αυτοσυσχετίζεται ενώ η χρονοσειρά X_t συσχετίζεται με τις 12 πρώτες αυτοσυσχετίσεις (τουλάχιστον) να εμφανίζονται στατιστικά σημαντικές.



Γράφημα 4.9: Γραφήματα εκτιμημένης συνάρτησης αυτοσυσχέτισης $\hat{\rho}(k)$ για $k = 1, ..., 36$ για τις χρονοσειρές Y_t (πάνω πλαίσιο) και X_t (κάτω πλαίσιο). Οι οριζόντιες γραμμές εκφράζουν ± 2 φορές το τυπικό σφάλμα $1/\sqrt{T}$ των εκτιμημένων αυτοσυσχετίσεων.

Διαχωρισμός Wold (Wold decomposition).

Εάν μία διαδικασία y_t είναι ασθενώς στάσιμη τότε μπορεί να γραφεί ως

$$y_t = d_t + \sum_{j=0}^{+\infty} \psi_j u_{t-j}$$

όπου

$$\psi_0 = 1 \quad \text{και} \quad \sum_{j=0}^{+\infty} \psi_j^2 < +\infty$$

ενώ τα u_t δεν αυτοσυσχετίζονται (βλ. υπόθεση **Υπ1** παρακάτω στον πίνακα της ενότητας 4.5.2), και ο όρος d_t αντιστοιχεί σε μία **προσδιοριστική διαδικασία**, δηλαδή η κατανομή του d_t είναι εκφυλισμένη και θεωρείται τελείως προβλέψιμο. Το κυριότερο παραδειγμα είναι αυτό με $d_t = \alpha$ (κάποια σταθερά που αντιστοιχεί στην αναμενόμενη τιμή της στάσιμης χρονοσειράς, $E(y_t) = \alpha$).

Στην περίπτωση δε που d_t=0 τότε η

$$y_t = \sum_{j=0}^{+\infty} \psi_j u_{t-j}$$

καλείται και **γραμμική διαδικασία** (linear process) με κύριο χαρακτηριστικό ότι η **αυτοσυσχέτιση** της y_t καθορίζεται απόλυτα από την ακολουθία των συντελεστών $\{\psi_j\}_{j=0}^{+\infty}$.

4.5.2 Ιεράρχηση στοχαστικών υποθέσεων χωρίς αυτοσυσχέτιση

Ο παρακάτω πίνακας «ταξινομεί» κάποιες υποθέσεις ή συνθήκες σχετικά με την κατανομή και/ή τις ροπές της κατανομής των διαταρακτικών όρων, οι οποίες είναι πολύ συχνές στη σχετική οικονομετρική βιβλιογραφία. Σε όλες αυτές τις υποθέσεις κοινός παρονομαστής είναι η **μηδενική αναμενόμενη τιμή** της διαδικασίας καθώς και η **μηδενική αυτοσυσχέτιση**. Επίσης είναι σύνηθες να υπάρχει κάποια επιπλέον συνθήκη που περιορίζει τη μη δεσμευμένη διακύμανση της χρονοσειράς, έτσι ώστε να είναι πεπερασμένη και να μην αποτελεί συνάρτηση του χρόνου, π.χ., $Var(u_t) = E(u_t^2) = \sigma_u^2 < +\infty$.

Θεωρήστε ότι το σύνολο $\mathbb{I}_t$ εκφράζει όλη την πληροφόρηση που έχουμε μέχρι και το χρόνο t. Συνήθως θα υποθέτουμε ότι περιλαμβάνει τουλάχιστον τις παρατηρηθείσες τιμές της υπο-εξέταση μεταβλητής, έστω u_t, μέχρι το χρόνο t δηλαδή

$$\mathbb{I}_t = (u_t, u_{t-1}, ...) \text{ και αντίστοιχα } \mathbb{I}_{t-1} = (u_{t-1}, u_{t-2}, ...)$$

Η πρώτη υπόθεση είναι και η γενικότερη ή «χαλαρότερη» όλων. Μία χρονοσειρά u_t που ικανοποιεί απλώς την **Υπ1** έχει μηνεδική αναμενόμενη τιμή, πεπερασμένη και σταθερή διακύμανση και δεν αυτοσυχετίζεται γραμμικά

$$Cov(u_t, u_s) = E(u_t \cdot u_s) - E(u_t) \cdot E(u_s) = E(u_t \cdot u_s) = 0$$

Αναλυτικά:

Πίνακας Υποθέσεων

- $E(u_t) = 0$, $E(u_t^2) = \sigma^2$, $E(u_t u_s) = 0$, $\forall t \neq s$ **λευκός θόρυ-**

βος, **Υπ1**

- $u_t \,|\mathbb{I}_t \sim \left(0, \sigma^2\right)$ περίπτωση «ακολουθίας διαφορών martingale», **Υπ2**
- $u_t \sim i.i.d\left(0\,,\,\sigma^2\right)$ ανεξάρτητος λευκός θόρυβος, **Υπ3**
- $u_t \sim N.i.d\left(0\,,\,\sigma^2\right)$ Gaussian ανεξάρτητος λευκός θόρυβος, **Υπ4**

Όπως προαναφέραμε, **κάθε υπόθεση/συνθήκη υπονοεί την ακριβώς προηγούμενή της** χωρίς να συμβαίνει και το αντίστροφο. Καθώς «κινούμαστε» από την **Υπ1** στην **Υπ4** οι ιδιότητες που προσδίδουμε στη χρονοσειρά u_t γίνονται ολοένα και πιο αυστηρές δηλαδή κατέχουμε (υποθέτουμε ότι κατέχουμε) ολοένα και περισσότερη πληροφόρηση σχετικά με την u_t. **Οπότε, καθώς κινούμαστε από την υπόθεση Υπ1 στην Υπ4, η στατιστική επαγωγή με βάση κατάλληλα τυποποιημένες συναρτήσεις των** u_t **διευκολύνεται**.

Παρατηρήστε ότι η συνθήκη **Υπ2** είναι πιο αυστηρή από ότι χρειάζεται για ορισμένα υποδείγματα χρονολογικών σειρών που υιοθετούνται στη σύγχρονη οικονομετρική πρακτική και επιτρέπουν δεσμευμένη ετεροσκεδαστικότητα, γι'αυτό υπάρχουν και περιπτώσεις (υποδειγμάτων) που απλώς υποθέτουμε $E\left(u_t \,|\mathbb{I}_{t-1}\right) = 0$ και ξαναγράφουμε την **Υπ2** ως

$$u_t \,|\mathbb{I}_{t-1} \sim \left(0, \sigma_t^2\right) \tag{4.3}$$

Στην περίπτωση αυτή, οι διαταρακτικοί όροι εξακολουθούν να αποτελούν μία «ακολουθία διαφορών martingale» (martingale difference sequence), αφού η συνθήκη **Υπ2** εγκλείεται στην (4.3). Η τελευταία λοιπόν, αποτελεί γενίκευση που επιτρέπει υπό συνθήκη ανομοιογένεια των όρων στο χρόνο (δεν υποθέτουμε $E\left(u_t^2 \,|\mathbb{I}_{t-1}\right) = \sigma^2$, άρα επιτρέπουμε την περίπτωση $E\left(u_t^2 \,|\mathbb{I}_{t-1}\right) = \sigma_t^2$).

Να σημειώσουμε ότι μία εφαρμογή του νόμου των επαναλαμβανόμενων προσδοκιών στις διαφορές martingale δείχνει εύκολα ότι οι διαφορές martingale δεν εμφανίζουν αυτοσυσχέτιση

$$Cov\left(u_t, u_{t-k}\right) = 0\,,\, \forall k \geq 1$$

ενώ επιπλέον ισχύει

$$Cov\left(u_t, g\left(u_{t-k}\right)\right) = 0\,,\, \forall k \geq 1$$

τουλάχιστον για συνεχείς και παραγωγίσιμες συναρτήσεις, $g(\cdot)$.

4.5.3 Εργοδικές χρονοσειρές (ergodic time series)

Θυμηθείτε ότι η χροοσειρά αποτελείται από παρατηρήσεις μίας και μόνο πραγματοποίησης της υποκείμενης στοχαστικής διαδικασίας. Η εκτίμηση των στατιστικών χαρακτηριστικών από μία και μόνο πραγματοποίηση της στοχαστικής διαδικασίας είναι ένα σημαντικό πρόβλημα και η εργοδικότητα υποθέτει ότι όλα τα στατιστικά χαρακτηριστικά παραμένουν αναλλοίωτα σε όλες τις πραγματοποιήσεις της διαδικασίας. Η έννοια της εργοδικότητας απαιτεί ουσιαστικά **ασυμπτωτική ανεξαρτησία «κατά μέσο όρο»** της χρονοσειράς.

Μία απόρροια της εργοδικότητας είναι ότι αν η χρονοσειρά Y_t είναι αυστηρώς στάσιμη και εργοδική τότε

$$\lim_{n\to+\infty}\frac{1}{n}\sum_{k=1}^{n}\gamma(k)=\lim_{n\to+\infty}\frac{1}{n}\sum_{k=1}^{n}Cov(Y_t,Y_{t-k})=0 \tag{4.4}$$

Το αποτέλεσμα (4.4) μπορεί να ερμηνευτεί ως ότι: οι αυτοσυνδιακυμάνσεις «κατά μέσο όρο» τείνουν στο μηδέν. Μία ικανή συνθήκη για το παραπάνω αποτέλεσμα είναι η απόλυτη αθροισιμότητα των αυτοσυνδιακυμάνσεων $\sum_{k=1}^{+\infty}|\gamma(k)|<+\infty$, δηλαδή ότι η σειρά $\sum_{k=1}^{+\infty}\gamma(k)$ συγκλίνει. Το τελευταίο υπονοεί ότι $\gamma(k)\to 0$ καθώς $k\to+\infty$ που ισχύει για οικονομικές χρονοσειρές αφού δηλώνει ότι καθώς η απόσταση k μεταξύ των παρατηρήσεων αυξάνεται ($k\to+\infty$), οι (γραμμικές) παρελθοντικές επιδράσεις στη χρονοσειρά μειώνονται μέχρι που μηδενίζονται (ασυμπτωτική μη-συσχέτιση).

Η εργοδικότητα είναι αυστηρότερη της απόλυτης αθροισιμότητας των συνδιακυμάνσεων καθώς επιτρέπει ιδιόμορφες περιπτώσεις όπου η σειρά δεν συγκλίνει όμως ικανοποιεί την (4.4). Παράδειγμα, η περίπτωση να έχουμε $\gamma(k)=(-1)^k$ (δεν απαντάται τέτοια μορφή αυτοσυνδιακύμανσης και συνεπώς αυτοσυσχέτισης στις οικονομικές χρονοσειρές), όπου $\lim_{n\to+\infty}\frac{1}{n}\sum_{k=1}^{n}(-1)^k=0$.

Ένα τυπικό παράδειγμα στάσιμης και εργοδικής (οικονομικής) χρονοσειράς είναι το παρακάτω

$$E(Y_t)=\mu\ \ ,\ \ Var(Y_t)=\sigma^2\ \ \text{και}\ \ Cov(Y_t,Y_{t-k})=0$$

Δείτε ότι η Y_t είναι (ασθενώς) στάσιμη **και εργοδική** αφού η αυτοσυνδιακύμανση είναι μηδενική για κάθε χρονική απόσταση k των παρατηρήσεων της Y_t. Γενικότερα, οι περισσότερες στάσιμες οικονομικές χρονοσειρές ικανοποιούν την παρακάτω οριακή συνθήκη

$$E(Y_t) = \mu \ \ , \ \ Var(Y_t) = \sigma^2 \ \ \text{και} \ \ Cov(Y_t, Y_{t-k}) \to 0 \ \text{καθώς} \ k \to +\infty$$

Ένα αντίθετο παράδειγμα (μη-εργοδικότητας) έχουμε αν

$$E(Y_t) = \mu \ \ , \ \ Var(Y_t) = \sigma^2 \ \ \text{και} \ \ Cov(Y_t, Y_{t-k}) = 2$$

Τότε η Y_t είναι (ασθενώς) στάσιμη όμως **δεν είναι εργοδική** αφού η (4.4) δεν ικανοποιείται, και πόσο μάλλον (διαισθητικά) η αυτοσυνδιακύμανση δεν τείνει στο μηδέν καθώς αυξάνει η χρονική απόσταση των παρατηρήσεων της Y_t.

Τέτοιες περιπτώσεις (μη-εργοδικές) είναι εξαιρετικά σπάνιες στην οικονομετρία αφού δεν αντιστοιχούν σε παρατηρούμενες ή θεωρητικές οικονομικές χρονοσειρές. Αν υποθέσουμε ότι η μόνη εξάρτηση που υπάρχει στη χρονοσειρά είναι γραμμική σειριακή εξάρτηση (γραμμική συσχέτιση) και η χρονοσειρά είναι στάσιμη, τότε είναι λογικό (για οικονομικές χρονοσειρές) να υποθέσουμε ότι η εξάρτηση φθίνει στο μηδέν καθώς οι παρατηρήσεις απομακρύνονται χρονικά.

4.5.4 Παραδείγματα ασθενώς στάσιμων χρονοσειρών

Το υπόδειγμα AR(1)

Το υπόδειγμα που θα παρουσιάσουμε παρακάτω είναι από τα πλέον συνηθισμένα στη σύγχρονη οικονομετρία λόγω της απλότητάς του, της διαισθητικής του ερμηνείας καθώς και της «επιτυχίας» του στην υποδειγματοποίηση της εξάρτησης οικονομικών χρονοσειρών.

Ονομάζεται «αυτοπαλίνδρομο υπόδειγμα[9] πρώτης τάξης» ή αλλιώς AR(1) υπόδειγμα και λαμβάνει τη μορφή,

$$Y_t = \alpha + \varphi Y_{t-1} + u_t \ , \ |\varphi| < 1 \qquad (4.5)$$

όπου u_t είναι λευκός θόρυβος

Το παραπάνω υπόδειγμα συνδέει τη μεταβλητή Y_t με το «παρελθόν» της. Το

[9] Η αυτοπαλίνδρομο σχήμα (autoregressive model or scheme).

υπόδειγμα όπως δίνεται στη σχέση (4.5) είναι ίσως παραπλανητικό ως προς την έκταση της σειριακής συσχέτισης της χρονοσειράς αφού μία πρώτη «αφελής» διατύπωση θα ήθελε την Y_t να εξαρτάται άμεσα μόνο από το πρόσφατο παρελθόν της, και ειδικότερα από την τιμή της μεταβλητής μία χρονική περίοδο πριν, δηλαδή την Y_{t-1}. Όμως, μία τέτοια θεώρηση θα ήταν εσφαλμένη.

Λύνοντας προς τα πίσω (backwards solution) εκφράζουμε τη χρονοσειρά Y_t ως μία συνάρτηση των τυχαίων όρων u_t και προχωρούμε στον υπολογισμό των βασικών ροπών της χρονοσειράς, δηλαδή του μέσου, της διακύμανσης και της αυτοσυνδιακύμανσης (και αυτοσυσχέτισης).

Έχουμε λοιπόν ότι,

$$\begin{aligned}
Y_t &= \alpha + \varphi Y_{t-1} + u_t \\
&= \alpha + \varphi\left(\alpha + \varphi Y_{t-2} + u_{t-1}\right) + u_t \\
&= \alpha + \varphi\alpha + \varphi^2 Y_{t-2} + u_t + \varphi u_{t-1} \\
&= \alpha + \varphi\alpha + \varphi^2\left(\alpha + \varphi Y_{t-3} + u_{t-2}\right) + u_t + \varphi u_{t-1} \\
&= \alpha + \varphi\alpha + \varphi^2\alpha + \varphi^3 Y_{t-3} + u_t + \varphi u_{t-1} + \varphi^2 u_{t-2} \\
&= \alpha + \varphi\alpha + \varphi^2\alpha + \varphi^3\left(\alpha + \varphi Y_{t-4} + u_{t-3}\right) + u_t + \varphi u_{t-1} + \varphi^2 u_{t-2} \\
&= \alpha + \varphi\alpha + \varphi^2\alpha + \varphi^3\alpha + \varphi^4 Y_{t-4} + u_t + \varphi u_{t-1} + \varphi^2 u_{t-2} + \varphi^3 u_{t-3} \\
&= ... \\
&= \alpha \sum_{j=0}^{t-1} \varphi^j \quad + \quad \varphi^t Y_0 \quad + \quad \sum_{j=0}^{t-1} \varphi^j u_{t-j}
\end{aligned}$$

όπου Y_0 συμβολίζει κάποια **αρχική τιμή** της Y στο χρόνο $t = 0$ ενώ είναι εμφανές ότι η πορεία της Y_t στο χρόνο σχετίζεται με τις τιμές που πήρε το σφάλμα u_t από το χρόνο 1 μέχρι και το χρόνο t.

Έστω ότι $|\varphi| < 1$ (συνθήκη στασιμότητας) και έστω ότι $t \to +\infty$, δηλαδή επιμηκύνουμε τη διαδικασία άπειρα προς το παρελθόν. Τότε,

επειδή

$$\text{γιά } |\varphi| < 1 \text{ ισχύει } \sum_{j=0}^{+\infty} \varphi^j = \frac{1}{1-\varphi} \quad (\text{η σειρά συγκλίνει})$$

και

$$\lim_{t \to +\infty} \varphi^t = 0 \text{ όταν } |\varphi| < 1$$

η σχέση

$$Y_t = \alpha \sum_{j=0}^{t-1} \varphi^j + \varphi^t Y_0 + \sum_{j=0}^{t-1} \varphi^j u_{t-j}$$

γράφεται ως

$$Y_t = \alpha \sum_{j=0}^{+\infty} \varphi^j + \sum_{j=0}^{+\infty} \varphi^j u_{t-j}$$

ή ισοδύναμα ως

$$Y_t = \frac{\alpha}{1-\varphi} + \sum_{j=0}^{+\infty} \varphi^j u_{t-j}$$

Εξετάζοντας τις ροπές της AR(1) διαδικασίας έχουμε σχετικά με την **αναμενόμενη τιμή** ότι

$$\begin{aligned} E(Y_t) &= \frac{\alpha}{1-\varphi} + \sum_{j=0}^{+\infty} \varphi^j E(u_{t-j}) \\ &= \frac{\alpha}{1-\varphi} + \sum_{j=0}^{+\infty} \varphi^j \times 0 \\ &= \frac{\alpha}{1-\varphi} \end{aligned}$$

Άρα η ύπαρξη σταθερού όρου στο AR(1) υπόδειγμα ισοδυναμεί με την ύπαρξη **μη μηδενικής** αναμενόμενης τιμής.

Σχετικά με τη **διακύμανση και αυτοσυνδιακύμανση θα θεωρήσουμε για ευκολία ότι** $\alpha = 0$, άρα και ότι $E(Y_t) = 0$. Ούτως ή άλλως, οι ροπές αυτές «δίνονται» ως συναρτήσεις αποκλίσεων από την αναμενόμενη τιμή.

Θα διαπιστώσουμε ότι η διακύμανση και η συνδιακύμανση της Y_t δίνονται από συνδυασμούς παραμέτρων που δεν αποτελούν συναρτήσεις του χρόνου, άρα το AR(1) υπόδειγμα «δημιουργεί» (γεννά) χρονοσειρές οι οποίες είναι **στάσιμες**, **όταν** $|\varphi| < 1$, και βέβαια όταν η $\{u_t\}$ είναι μία στάσιμη χρονοσειρά.

Αναλυτικά, η **διακύμανση** σ_Y^2 ή $\gamma_Y(0)$ ή $Var(Y_t)$ της χρονοσειράς Y_t δίνε-

ται από

$$\begin{aligned} Var(Y_t) &= E\left(\sum_{j=0}^{+\infty}\varphi^j u_{t-j}\right)\left(\sum_{i=0}^{+\infty}\varphi^i u_{t-i}\right) \\ &= E\left(u_t+\varphi u_{t-1}+\varphi^2 u_{t-2}+...\right)\left(u_t+\varphi u_{t-1}+\varphi^2 u_{t-2}+...\right) \\ &= \sigma_u^2+\varphi^2\sigma_u^2+\varphi^4\sigma_u^2+\varphi^6\sigma_u^2+... \\ &= \sigma_u^2\left(1+\varphi^2+\varphi^4+\varphi^6+...\right) \\ &= \sigma_u^2\sum_{j=0}^{+\infty}\left(\varphi^2\right)^j \\ &= \frac{\sigma_u^2}{1-\varphi^2} \end{aligned}$$

Η τελευταία ισότητα προκύπτει μόνο όταν $|\varphi|<1$ αφού τότε $\left|\varphi^2\right|<1$ και η σειρά $\sum_{j=0}^{+\infty}\left(\varphi^2\right)^j$ συγκλίνει, με

$$\sum_{j=0}^{+\infty}\left(\varphi^2\right)^j=\frac{1}{1-\varphi^2}$$

Η **αυτοσυνδιακύμανση** σε χρονική υστέρηση k δίνεται ως εξής. Αφού στον χρόνο t ισχύει

$$\begin{aligned} Y_t &= \sum_{j=0}^{+\infty}\varphi^j u_{t-j} \\ &= u_t+\varphi u_{t-1}+\varphi^2 u_{t-2}+\varphi^3 u_{t-3}+... \\ &...+\varphi^k u_{t-k}+\varphi^{k+1}u_{t-k-1}+... \end{aligned}$$

τότε και αντίστοιχα στον χρόνο $t-k$ ισχύει ότι

$$Y_{t-k}=\sum_{j=0}^{+\infty}\varphi^j u_{t-k-j}$$

$$= u_{t-k} + \varphi u_{t-k-1} + \varphi^2 u_{t-k-2} + \varphi^3 u_{t-k-3} + ...$$

Οπότε, οι συνδιακυμάνσεις (συνάρτηση αυτοσυνδιακύμανσης) δίνονται από

$$\begin{aligned}
\gamma_Y(k) &= E(Y_t Y_{t-k}) \\
&= E(u_t + \varphi u_{t-1} + ...)(u_{t-k} + \varphi u_{t-k-1} + ...) \\
&= \sigma_u^2 \left(\varphi^k + \varphi^{k+2} + \varphi^{k+4} + ...\right) \\
&= \sigma_u^2 \varphi^k \left(1 + \varphi^2 + \varphi^4 + ...\right) \\
&= \sigma_u^2 \varphi^k \sum_{j=0}^{+\infty} \left(\varphi^2\right)^j = \frac{\sigma_u^2}{1-\varphi^2}\varphi^k \\
&= \sigma_Y^2 \varphi^k
\end{aligned}$$

Επομένως, η **συνάρτηση αυτοσυσχέτισης** για τη στάσιμη AR(1) χρονοσειρά Y_t του υποδείγματος (4.5) δίνεται από

$$\rho_Y(k) = \frac{\gamma_Y(k)}{\gamma_Y(0)} = \frac{\sigma_Y^2 \varphi^k}{\sigma_Y^2} = \varphi^k$$

Η παραπάνω συνάρτηση σχηματοποιείται στο αμέσως επόμενο διάγραμμα (4.10) για $\varphi = 0.85$ και $k = 1, 2, ..., 12$. Συμπερασματικά, ενώ το αρχικό υπόδειγμα περιλαμβάνει μόνο την πρώτη χρονική υστέρηση της Y_t, η αυτοσυνδιακύμανση

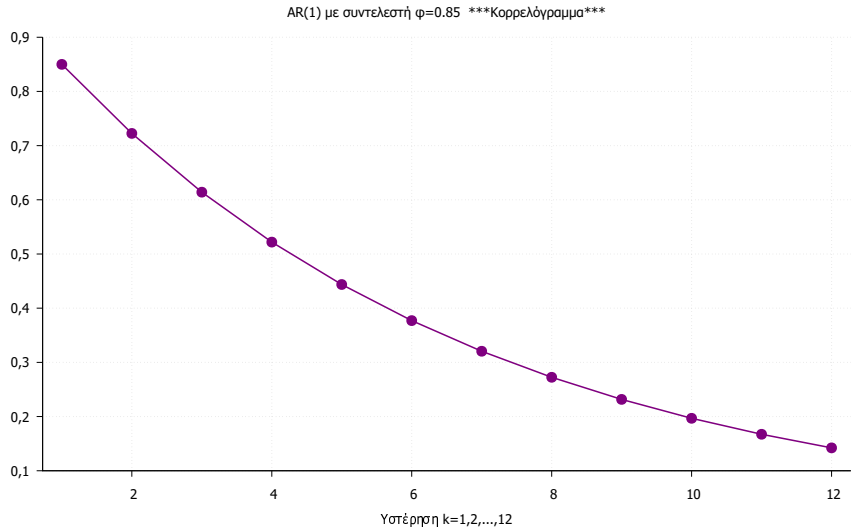
$$Cov(Y_t, Y_{t-k}) = \sigma_Y^2 \varphi^k$$

υπονοεί ότι η σειρά συσχετίζεται ή συνδέεται με το άπειρο παρελθόν της, δηλαδή με όλες τις παρελθοντικές τιμές της. Αυτό είναι απόρροια του ότι η Y_t δίνεται ως ένας σταθμικός μέσος (weighted average) των σφαλμάτων από το άπειρο παρελθόν μέχρι το χρόνο t

$$Y_t = \sum_{j=0}^{+\infty} \varphi^j u_{t-j}$$

με σταθμίσεις φ^j.

Η συνθήκη στασιμότητας $|\varphi| < 1$ επιβάλλει στις σταθμίσεις φ^j να μειώνονται γρήγορα προς το μηδέν όσο προχωράμε προς τα πίσω άρα και η αυτοσυσχέτιση - παρά το ότι θεωρητικά εκτείνεται στο άπειρο παρελθόν - πρακτικά φθίνει εκθετικά προς το μηδέν.



Γράφημα 4.10: Θεωρητικό κορρελόγραμμα μίας AR(1) χρονοσειράς με $\varphi = 0.85$ για υστερήσεις $k = 1, 2, ..., 12$. Απεικονίζονται οι τιμές της $(0.85)^k$ για $k = 1, 2, ..., 12$.

Ο διαχωρισμός Wold

$$Y_t = d_t + \sum_{j=0}^{+\infty} \psi_j u_{t-j}$$

για τη χρονοσειρά της σχέσης (4.5) προκύπτει αν θέσουμε

$$d_t = E\left(Y_t\right) = \frac{\alpha}{1-\varphi}$$

με σταθμίσεις

$$\psi_j = \varphi^j$$

Το υπόδειγμα AR(p)

Έστω το αυτοπαλίνδρομο υπόδειγμα[10] τάξεως p ή αλλιώς υπόδειγμα AR(p)

$$y_t = \alpha + \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + ... + \varphi_p y_{t-p} + u_t \quad , \quad t = 1, ..., T$$

[10] Θα υιοθετήσουμε μικρά γράμματα για να συμβολίσουμε τις χρονοσειρές. Επίσης, κάνουμε μία υπέρβαση εδώ καθώς για $p > 1$ τα αυτοπαλίνδρομα υποδείγματα περιλαμβάνουν πάνω από μία ερμηνευτική μεταβλητή!.

όπου u_t είναι λευκός θόρυβος

Μία AR(p) διαδικασία ή χρονοσειρά y_t μπορεί να ξαναγραφεί ως

$$\varphi(L)\, y_t = \alpha + u_t$$

όπου

$$\varphi(L) = 1 - \varphi_1 L - \varphi_2 L^2 - ... - \varphi_p L^p$$

ένα πολυώνυμο p τάξεως του τελεστή υστέρησης[11] L. Επειδή το L είναι ένας τελεστής συνηθίζεται, όταν προβαίνουμε σε αλγεβρικές πράξεις, να γράφουμε $\varphi(z)$ αντί $\varphi(L)$.

Η **συνθήκη στασιμότητας** γενικά για AR(p) χρονοσειρές δίνεται ως εξής:

> «όλες οι ρίζες r_j, για $j = 1, ..., p$ του πολυωνύμου $\varphi(z)$ πρέπει να βρίσκονται εκτός του μοναδιαίου κύκλου»

Δηλαδή αν $\varphi(r_j) = 0$ (που σημαίνει ότι η r_j είναι μία ρίζα του πολυωνύμου) θα πρέπει η απόλυτη τιμή της ρίζας να είναι μεγαλύτερη του ένα, $|r_j| > 1$ για κάθε j, ώστε η y_t να είναι **ασθενώς στάσιμη**. Αν κάποια ρίζα r_j του πολυωνύμου είναι μιγαδική, δηλαδή μπορεί να εκφραστεί ως $r_j = \alpha \pm bi$ όπου α, b είναι πραγματικοί αριθμοί και $i = \sqrt{-1}$, τότε η απόλυτη τιμή της ρίζας αντιστοιχεί στο μέτρο (modulus) του μιγαδικού αριθμού

$$|r_j| = \sqrt{a^2 + b^2}$$

Για παράδειγμα, έστω ένα $AR(2)$ υπόδειγμα

$$y_t = 0.85 y_{t-1} - 0.23 y_{t-2} + u_t$$

Η συνθήκη στασιμότητας επιβάλλει οι ρίζες του πολυωνύμου

$$\varphi(z) = 1 - 0.85z + 0.23z^2$$

να βρίσκονται εκτός του μοναδιαίου κύκλου.

[11] Έστω ότι j ακέραιος. Τότε ο τελεστής υστέρησης δίνει $L^j y_t = y_{t-j}$. Για σταθερούς όρους, π.χ., α, ισχύει ότι $L^j \alpha = \alpha$.

Έχουμε ότι

$$\varphi(z) = 0 \Rightarrow r_{1,2} = 1.847 \pm 0.966i$$

και

$$|r_{1,2}| = \sqrt{(1.847)^2 + (0.966)^2} = 2.084 > 1$$

οπότε η χρονοσειρά y_t είναι στάσιμη.

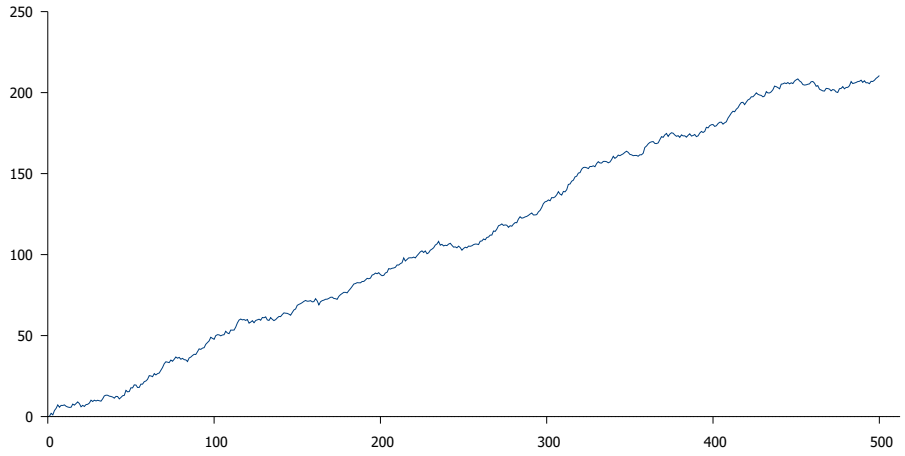
4.5.5 Μη-στασιμότητα

Μη-στάσιμες (nonstationary) καλούνται οι χρονοσειρές για τις οποίες τουλάχιστον μία ροπή τους εξαρτάται άμεσα από τον χρόνο. Συνήθως, η οικονομετρική θεωρία περιορίζεται στις δύο πρώτες ροπές ή από κοινού ροπές δηλαδή το μέσο, τη διακύμανση και τη αυτοσυνδιακύμανση ή αυτοσυσχέτιση.

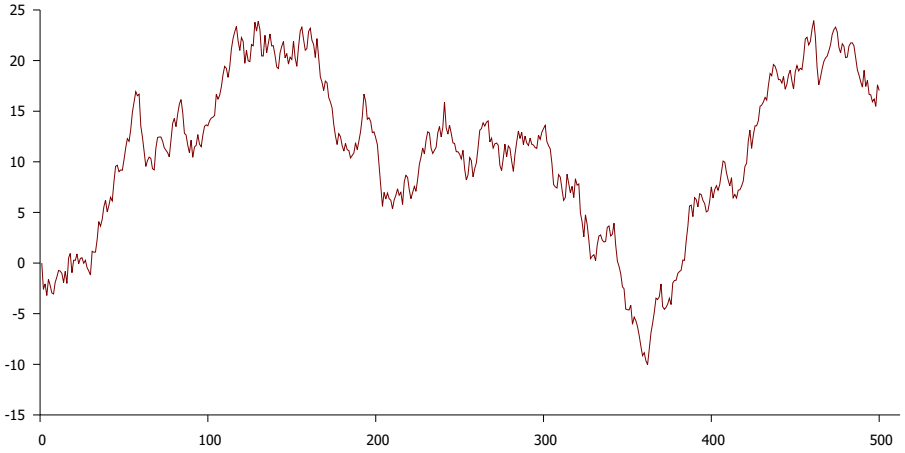
Η μη στασιμότητα αναφέρεται λοιπόν στην άμεση εξάρτηση από το χρόνο της αναμενόμενης τιμής και/ή της διακύμανσης και/ή της συνάρτησης αυτοσυνδιακύμανσης μίας διαδικασίας. Οπτικά, οι μη στάσιμες χρονοσειρές εμφανίζουν ορισμένα πρόδηλα χαρακτηριστικά. **Δύο** από αυτά είναι η **εμφάνιση έντονων «δομών» ή «σχηματισμών» (structures or patterns)** και η δεύτερη είναι η **εμφάνιση «τάσεων» (trends)**.

Ως «οπτικό» παράδειγμα/βοήθημα, μεταφέρουμε τη συζήτηση στα παρακάτω δύο διαγράμματα τα οποία παρουσιάζουν πραγματοποιήσεις μη στάσιμων χρονοσειρών. Η μεταβλητή στο επόμενο διάγραμμα (4.11) παρουσιάζει εμφανώς θετική τάση δηλαδή κινείται συνεχώς ανοδικά παρόλο που η κίνηση δεν είναι μονότονη και εμφανίζει τυχαίες αυξομειώσεις. Στην οικονομική απαντώνται συχνά χρονοσειρές με παρόμοια χρονικά μονοπάτια, και ειδικότερα μακροοικονομικές μεταβλητές όπως το πραγματικό ΑΕΠ, η κατανάλωση κ.α. Η μεταβλητή στο παρακάτω διάγραμμα (4.12) μπορεί να μην καθιστά πρόδηλη την ύπαρξη τάσης όμως σαφέστατα αυτοσυσχετίζεται έντονα καθώς παρουσιάζει μεγάλα διαστήματα που κινείται ανοδικά ή καθοδικά ενώ η διακύμανσή της φαίνεται να εξαρτάται από το χρόνο καθώς «περιφέρεται» έντονα προς τα πάνω ή προς τα κάτω στον χρόνο. Λανθασμένα θα μπορούσαμε να υποθέσουμε την ύπαρξη τάσεων οι οποίες «αλλάζουν» κατεύθυνση κατά την εξέλιξη της χρονοσειράς στο χρόνο. Παρόμοιες χρονοσειρές απαντώνται επίσης συχνά στην οικονομική όπως στις χρονοσειρές τιμών χρηματοοικονομικών στοιχείων, π.χ., οι τιμές των μετοχών.

Ένα δημοφιλές υπόδειγμα «γέννησης» μίας μη στάσιμης χρονοσειράς $\{y_t\}$



Γράφημα 4.11: Γράφημα **μη στάσιμης χρονοσειράς** με έντονο το στοιχείο της τάσης



Γράφημα 4.12: Γράφημα **μη στάσιμης** χρονοσειράς

είναι το παρακάτω

$$y_t = \alpha + bt + u_t, \quad t = 1, 2, ..., T$$

όπου υποθέτουμε ότι η χρονοσειρά u_t των διαταρακτικών όρων του υποδείγματος είναι στάσιμη. Για παράδειγμα, μπορούμε να υιοθετήσουμε για την u_t κάποια από τις υποθέσεις **Υπ1, Υπ2, Υπ2α, Υπ3, Υπ4** ή γενικότερα μπορούμε να υποθέσουμε ότι η χρονοσειρά u_t αυτοσυσχετίζεται αλλά είναι στάσιμη με μηδενικό

μέσο.

Είναι εμφανές ότι ο μέσος της σειράς $\{y_t\}$ (δεσμευμένος ή μη) εξαρτάται γραμμικά από το χρόνο. Στο συγκεκριμένο υπόδειγμα ο μη δεσμευμένος μέσος δίνεται από τη συνάρτηση

$$E(y_t) = \alpha + \beta t$$

ενώ διακύμανση και συνδιακύμανση δεν εξαρτώνται άμεσα από το χρόνο αφού

$$Var(y_t) = \sigma_y^2 = \sigma_u^2$$

και

$$Cov(y_t, y_{t-k}) = \gamma_y(k) = \gamma_u(k)$$

Βέβαια, αφού ο μέσος εξαρτάται άμεσα από το χρόνο, η χρονοσειρά $\{y_t\}$ **δεν είναι ασθενώς στάσιμη**. Παρατηρούμε όμως ότι οι αποκλίσεις από την «τάση» $u_t = y_t - \alpha - \beta t$ είναι ασθενώς στάσιμες αφού ουσιαστικά οι αποκλίσεις από την τάση ταυτίζονται με τους διαταρακτικούς όρους του υποδείγματος.

Χρονοσειρές με μη στασιμότητα αυτού του τύπου ονομάζονται και **«στάσιμες γύρω από τάση» (trend stationary)** και μπορούν να γενικευτούν ως προς την υποδειγματοποίηση της μέσης τιμής της y_t χρησιμοποιώντας πολυώνυμα του χρόνου μεγαλύτερης τάξης, για παράδειγμα

$$y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + ... + \beta_k t^k + u_t$$

ή και πιο σύνθετες μη γραμμικές συναρτήσεις του χρόνου,

$$y_t = f(t) + u_t$$

με την $\{u_t\}$ να θεωρείται πάντα μία ασθενώς στάσιμη χρονοσειρά μηδενικού μέσου.

Ένα άλλο δημοφιλές παράδειγμα μη στάσιμων χρονοσειρών, ίσως το σημαντικότερο στη σύγχρονη οικονομετρία χρονοσειρών, είναι αυτές οι οποίες καθίστανται στάσιμες μετά την εφαρμογή του τελεστή πρώτων διαφορών. Οι χρονοσειρές αυτού του τύπου ονομάζονται **«στάσιμες μέσω πρώτων διαφορών» (difference stationary)**.

Έστω ότι η χρονοσειρά $\{y_t\}$ δημιουργείται από το υπόδειγμα **τυχαίας ε-**

ξέλιξης ή τυχαίου περιπάτου (random walk)

$$y_t = y_{t-1} + u_t, \quad u_t \sim i.i.d\left(0, \sigma_u^2\right) \tag{4.6}$$

ή από το υπόδειγμα **τυχαίας εξέλιξης με μετατόπιση (random walk with drift)**

$$y_t = a + y_{t-1} + u_t, \quad u_t \sim i.i.d\left(0, \sigma_u^2\right) \tag{4.7}$$

Αποδεικνύεται εύκολα όταν λύσουμε προς τα πίσω (backwards solution) ότι χρονοσειρές που «δημιουργούνται» σύμφωνα με τα υποδείγματα (4.6) και (4.7) μπορούν να γραφούν αντίστοιχα ως

$$y_t = y_0 + \sum_{j=1}^{t} u_j \quad \text{ή} \quad y_t = \sum_{j=1}^{t} u_j$$

με την δεύτερη ισότητα να προκύπτει αν θέσουμε την αρχική τιμή $y_0 = 0$ **και**

$$y_t = y_0 + at + \sum_{j=1}^{t} u_j \quad \text{ή} \quad y_t = at + \sum_{j=1}^{t} u_j$$

όπου επίσης η δεύτερη ισότητα προκύπτει όταν θέσουμε την αρχική τιμή $y_0 = 0$.

Ο όρος

$$\sum_{j=1}^{t} u_j$$

καλείται **«στοχαστική τάση» (stochastic trend)** σε αντίθεση με τον όρο αt ο οποίος αντιστοιχεί στην **προσδιοριστική τάση**.

Συνεπώς, όταν η χρονοσειρά y_t «δημιουργείται» ή προκύπτει από ένα υπόδειγμα **τυχαίου περιπάτου** με μηδενική αρχική τιμή $y_0 = 0$, τότε η αναμενόμενη τιμή, η διακύμανση, η συνάρτηση αυτοσυνδιακύμανσης και η συνάρτηση αυτοσυσχέτισης δίνονται αντίστοιχα από τις παρακάτω σχέσεις:

$$\begin{aligned} \mu_y &= E\left(y_t\right) = 0 \\ \sigma_y^2 &= \gamma_y\left(0\right) = Var\left(y_t\right) = t\sigma_u^2 \\ \gamma_y\left(k\right) &= Cov\left(y_t, y_{t-k}\right) = \left(t-k\right)\sigma_u^2, \; k \geq 1 \end{aligned} \tag{4.8}$$

$$\rho_y(k) = \frac{\gamma_y(k)}{\gamma_y(0)} = \frac{(t-k)\,\sigma_u^2}{t\sigma_u^2} = 1 - \frac{k}{t},\ k \geq 1$$

Όταν δημιουργείται από ένα υπόδειγμα **τυχαίου περιπάτου με μετατόπιση** οι αντίστοιχες ροπές δίνονται από τις σχέσεις:

$$\begin{aligned} \mu_y &= E(y_t) = at \\ \sigma_Y^2 &= \gamma_y(0) = Var(y_t) = t\sigma_u^2 \\ \gamma_y(k) &= Cov(y_t, y_{t-k}) = (t-k)\,\sigma_u^2,\ k \geq 1 \\ \rho_y(k) &= \frac{\gamma_y(k)}{\gamma_y(0)} = \frac{(t-k)\,\sigma_u^2}{t\sigma_u^2} = 1 - \frac{k}{t},\ k \geq 1 \end{aligned} \tag{4.9}$$

Είναι εμφανές ότι, σε κάθε περίπτωση, η χρονοσειρά είναι μη στάσιμη αφού τουλάχιστον η διακύμανση και η αυτοσυνδιακύμανση αποτελούν συναρτήσεις του χρόνου.

Εφαρμόζοντας τον τελεστή πρώτων διαφορών $\Delta = (1 - L)$ στη χρονοσειρά y_t στο υπόδειγμα τυχαίας εξέλιξης χωρίς μετατόπιση καταλήγουμε σε μία στάσιμη χρονοσειρά, αφού

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

Παρομοίως, στο υπόδειγμα τυχαίας εξέλιξης με μετατόπιση

$$\Delta y_t = \alpha + u_t$$

Αυτού του τύπου οι μη στάσιμες χρονοσειρές καλούνται και **ολοκληρώσιμες πρώτης τάξης**, $I(1)$, **(integrated of order 1)** ενώ αντίστοιχα η σειρά Δy_t καλείται **ολοκληρώσιμη μηδενικής τάξης**, $I(0)$, αφού δεν χρειάζεται να εφαρμόσουμε πρώτες διαφορές για να έχουμε στασιμότητα.

Σπανιότερα σε οικονομικές χρονοσειρές μπορεί να συναντήσουμε μη στασιμότητα τύπου $I(2)$, δηλαδή πρέπει να εφαρμόσουμε τον τελεστή διαφορών δύο φορές ώστε να τις καταστήσουμε στάσιμες. Για παράδειγμα, η χρονοσειρά y_t που δημιουργείται με βάση το παρακάτω αυτοπαλίνδρομο υπόδειγμα

$$y_t = 2y_{t-1} - y_{t-2} + u_t$$

είναι μη στάσιμη αλλά στάσιμη κατόπιν δεύτερων διαφορών, δηλαδή είναι $I(2)$ αφού

$$\begin{aligned} y_t &= 2y_{t-1} - y_{t-2} + u_t \\ &\Leftrightarrow y_t - 2y_{t-1} + y_{t-2} = u_t \\ &\Leftrightarrow (1-L)^2 y_t = u_t \end{aligned}$$

Μεγαλύτερες τάξεις ολοκλήρωσης, π.χ., $I(3)$, $I(4)$ κ.ο.κ, δεν απαντώνται στην οικονομική. Μολονότι χρονοσειρές με τάσεις οι οποίες είναι υπερβολικά «ομαλές» με «αργές» μεταβολές στα επίπεδά τους μπορεί να υπονοούν συμπεριφορά τύπου $I(2)$ η οικονομική δικαιολόγηση τέτοιας μη στασιμότητας είναι δύσκολη. Για παράδειγμα, οι $I(2)$ χρονοσειρές έχουν διαταράξεις με «επιδράσεις» που μεγενθύνονται στο χρόνο.

Το ζήτημα της μη στασιμότητας των οικονομικών χρονοσειρών αποτέλεσε τον κυρίαρχο άξονα έρευνας τα τελευταία - τουλάχιστον - 20 χρόνια στην οικονομετρία χρονοσειρών αφού, όπως έχουμε ήδη αναφέρει, το ζήτημα των πλασματικών συσχετίσεων μεταξύ μη στάσιμων χρονοσειρών είναι άμεσο και έντονο.

Η σύγχρονη οικονομετρική πρακτική αφαιρεί τις «τάσεις» είτε αυτές είναι προσδιοριστικές είτε στοχαστικές και κατόπιν προβαίνει σε εκτίμηση και επαγωγή των παραμέτρων ενδιαφέροντος, αφού σε αντίθετη περίπτωση οποιαδήποτε ευρήματα κινδυνεύουν να χαρακτηριστούν «πλασματικά».

Το ερώτημα κατά πόσο πρέπει να «αφαιρούμε» μία προσδιοριστική χρονική τάση ή κατά πόσο πρέπει να εφαρμόζουμε τον τελεστή των πρώτων διαφορών για να καταστήσουμε μία - συνήθως μακροοικονομική - χρονοσειρά στάσιμη αποτέλεσε αντικείμενο έντονης ερευνητικής αντιπαράθεσης στο χώρο των μακροοικονομικών χρονοσειρών τις τελευταίες τρεις δεκαετίες μετά τα ευρήματα των Nelson και Plosser (1982)[12] και με δεδομένο ότι αρκετές οικονομικές χρονοσειρές «φαίνεται» να εμπίπτουν επιτυχώς και στις δύο κατηγορίες.

Πρέπει να αναφέρουμε ότι και οι δύο τρόποι υποδειγματοποίησης της μη στασιμότητας είναι «μη θεωρητικές» αφού υποδηλώνουν άγνοια σχετικά με τη δημιουργία της τάσης στη χρονοσειρά ενώ έχουν και γενικότερα σημαντικές επιπτώσεις στη διαδικασία της εμπειρικής έρευνας.

[12]Nelson, C.R. and Plosser, C.I. (1982) "Trends and Random Walks in Macroeconomic Time Series: some Evidence and Implications," Journal of Monetary Economics, 10, 139-162

Για παράδειγμα, αν η τάση είναι προσδιοριστική τότε

- η χρονοσειρά τείνει μακροχρόνια να επιστρέφει στην αναμενόμενη τιμή της (στο μέσο της, δηλαδή στη γραμμική ή άλλη προσδιοριστική τάση)
- οι διαταράξεις u_t έχουν αποτελέσματα τα οποία φθίνουν με το πέρασμα του χρόνου[α΄]
- ενώ η διακύμανση του σφάλματος πρόβλεψης $Var\left(y_{t+h} - \hat{y}_{t+h}\right)^2$, όπου $\hat{y}_{t+h}$ υποδηλώνει την πρόβλεψη της χρονοσειράς h περιόδους μπροστά, είναι σταθερή για κάθε ορίζοντα πρόβλεψης h.

[α΄] Τέτοιου είδους αποτελέσματα, ως συνάρτηση του χρονικού ορίζοντα στον οποίο εκδηλώνονται, ονομάζονται και «συναρτήσεις απόκρισης σε αιφνίδιες διαταραχές» (impulse response functions), ένα θέμα στο οποίο δεν θα επεκταθούμε περισσότερο.

Αν η τάση είναι στοχαστική τότε

- η υποκείμενη σειρά δεν τείνει μακροχρόνια προς κάποια σταθερή αναμενόμενη τιμή (μέσο) ή κάποια μεταβαλλόμενη προσδιοριστικά τιμή δηλαδή «τάση»
- οι διαταράξεις έχουν «μόνιμα» μη φθίνοντα αποτελέσματα στη χρονοσειρά
- ενώ η διακύμανση του σφάλματος πρόβλεψης αυξάνει καθώς αυξάνεται ο ορίζοντας της πρόβλεψης, h.

4.5.6 Παράδειγμα εκτίμησης ελαχίστων τετραγώνων ενός υποδείγματος AR(1)

Η παρακάτω εμπειρική άσκηση αναπτύσσεται επίσης με αρχείο εντολών

`spread.inp` του gretl

«Κατεβάζουμε» τη χρονοσειρά του πραγματικού ΑΕΠ των ΗΠΑ **gdpc1** (Δισεκατομμύρια αλυσωτά δολλάρια 2012, εποχικά προσαρμοσμένο) μέσω του gretl[13]

[13] Η χρονοσειρά παρέχεται από τη βάση δεδομένων FRED της ομοσπονδιακής τράπεζας του St. Louis (Η.Π.Α), `https://fred.stlouisfed.org/series/GDPC1`

και τη μετασχηματίζουμε σε ετησιοποιημένη τριμηνιαία ποσοστιαία μεταβολή (annualized q-o-q growth rate), (τριμηνιαίο ρυθμό μεγέθυνσης):

$$\mathrm{DY_t} = 4 \times 100 \times \ln\left(\frac{\mathrm{gdpc1_t}}{\mathrm{gdpc1_{t-1}}}\right)$$

Στο **γράφημα (4.13)** απεικονίζονται (πάνω) η χρονοσειρά του ΑΕΠ η οποία είναι εμφανώς μη-στάσιμη και (κάτω) ο ρυθμός μεγέθυνσης DY_t η οποία είναι μία στάσιμη χρονοσειρά, τουλάχιστον ως προς το αν περιέχει μοναδιαία ρίζα ή αν είναι ολοκληρώσιμη πρώτης τάξης, δηλαδή δεν περιέχει στοχαστική τάση.

Εκτιμούμε με τη μέθοδο ελαχίστων τετραγώνων το υπόδειγμα

$$\mathrm{DY_t} = \alpha + \phi \mathrm{DY_{t-1}} + u_t$$

και λαμβάνουμε τα αποτελέσματα

$$\widehat{\mathrm{DYt}} = \underset{(0.32082)}{2.67684} + \underset{(0.057852)}{0.122848}\,\mathrm{DY_{t-1}}$$

$$T = 296 \quad R^2 = 0.0151 \quad \hat{\sigma} = 4.6273$$

τυπικά σφάλματα σε παρενθέσεις

χρονική περίοδος εκτίμησης: 1947:3-2021:2

και

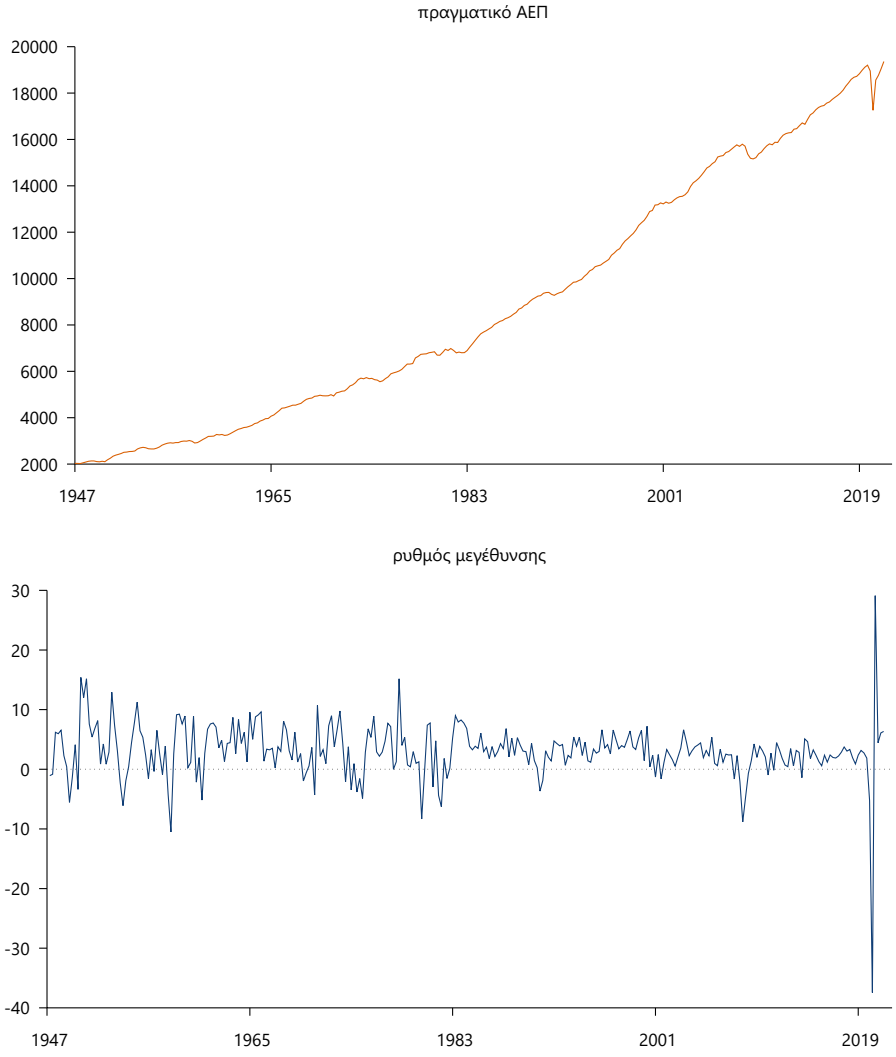
$$\widehat{\mathrm{DYt}} = \underset{(0.26477)}{1.98335} + \underset{(0.054818)}{0.361385}\,\mathrm{DY_{t-1}}$$

$$T = 290 \quad R^2 = 0.1311 \quad \hat{\sigma} = 3.4656$$

τυπικά σφάλματα σε παρενθέσεις

χρονική περίοδος εκτίμησης: 1947:3-2019:4

Η επίδραση της πανδημίας, δηλαδή ο ακραία αρνητικός ρυθμός μεγέθυνσης -37.44% το δεύτερο τρίμηνο του 2020 και στη συνέχεια ο ακραία θετικός ρυθμός 29.10% στο τρίτο τρίμηνο (αποτέλεσμα αναπήδησης ή rebound effect) «ξεγελούν» την εκτίμηση του συντελεστή $\hat{\phi}$ δείχνοντας ένα στατιστικά σημαντικό αποτέλεσμα αυτοσυσχέτισης $\hat{\phi} = 0.122$ το οποίο όμως είναι οικονομικά χαμηλό.



Γράφημα 4.13: Πραγματικό ΑΕΠ Η.Π.Α και τριμηνιαίος ρυθμός μεγέθυνσης (ετησιοποιημένος). Περίοδος:1947:1-1921:2, $T = 298$ παρατηρήσεις.

Στο δεύτερο υπόδειγμα, για την περίοδο 1947:3-2019:4 λαμβάνουμε $\hat{\phi} = 0.361$ και $R^2 = 0.1311$.

Να τονίσουμε ότι, προς το παρόν, δεν ανησυχούμε για το εάν το υπόδειγμα $AR(1)$ είναι το καταλληλότερο για αυτά τα δεδομένα ή όχι. Στόχος μας είναι να κατανοήσουμε τη διαδικασία εκτίμησης του υποδείγματος.

4.5.7 Παράδειγμα εκτίμησης υποδείγματος χρονοσειρών με υστερήσεις ερμηνευτικής μεταβλητής

Η εμπειρική άσκηση αναπτύσσεται επίσης στο αρχείο εντολών

`spread.inp` του gretl

«Κατεβάζουμε» δεδομένα χρονοσειρών για τις επιτοκιακές αποδόσεις δύο τύπων κρατικών ομολόγων: ενός μακροπρόθεσμου και ενός βραχυπρόθεσμου. Η χρονοσειρά **gs10** αντιστοιχεί στην επιτοκιακή απόδοση (yield) αγοράς των τίτλων (ομολόγων) του Αμερικανικού δημοσίου με σταθερή 10ετή λήξη (δεκαετή ομόλογα) ενώ η χρονοσειρά **tb3ms** αντιστοιχεί στην επιτοκιακή απόδοση δευτερογενούς αγοράς των τρίμηνων εντόκων γραμματίων του Αμερικανικού δημοσίου.

Η διαφορά τους καλείται «σπρέντ» (spread) ή επιτοκιακή διαφορά ή εύρος και αποτελεί έναν βασικό πρόδρομο δείκτη (leading indicator) της πραγματικής οικονομίας (είτε του μέσου ρυθμού μεγέθυνσης τα επόμενα 3-6 τρίμηνα είτε της πιθανότητας ύφεσης[14]). Θεωρούμε ότι η μεταβλητή του spread είναι στάσιμη (παρότι σίγουρα έχει αυξημένη εμμονή).

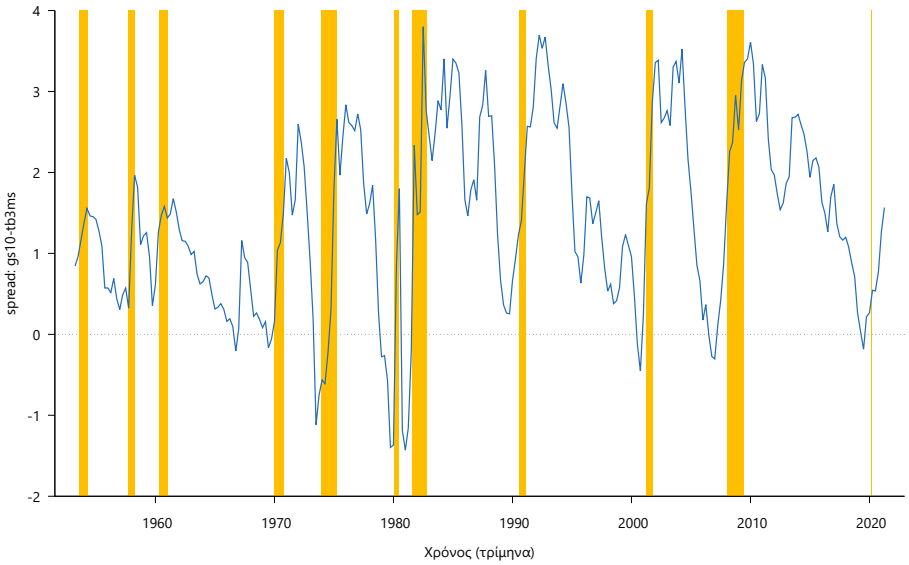
Το **γράφημα (4.14)** απεικονίζει τη μεταβλητή του spread

$$\text{spread}_\text{t} = \text{gs10}_\text{t} - \text{tb3ms}_\text{t}$$

την περίοδο 1953:2 - 2021:2. Παρατηρούμε ότι το spread γίνεται αρνητικό «πριν» από τις υφέσεις της Αμερικανικής οικονομίας, οι οποίες απεικονίζονται ως «σκιάσεις». **Σημείωση:** Περισσότερα για τις υφέσεις και πως καθορίζονται τα συγκεκριμένα χρονικά διαστήματα θα βρείτε στην ιστοσελίδα

`https://www.nber.org/research/business-cycle-dating` .

[14] Η ανάλυση της διαχρονικής πορείας του spread αποδόσεων δεκαετών ομολόγων του δημοσίου μιας χώρας με αυτά μίας άλλης (συνήθως χαμηλού κινδύνου, π.χ. εύρος επιτοκίων Ελληνικών-Γερμανικών ομολόγων) ανταναλκά τη διαφορά της επικινδυνότητας της χώρας αυτής με τη χώρα αναφοράς και τελικά την εμπιστοσύνη που εκφράζουν οι επενδυτές σε αυτήν



Γράφημα 4.14: Διαφορά (spread) μακροπρόθεσμου - βραχυπρόθεσμου επιτοκίου κρατικών ομολόγων του Αμερικανικού δημοσίου. Περίοδος 1953:2 - 2021:2 (δεύτερο τρίμηνο του 1953 μέχρι δεύτερο τρίμηνο του 2021). Περιοχές με σκίαση «δείχνουν» χρονικά διαστήματα ύφεσης της πραγματικής οικονομικής δραστηριότητας.

Στη συνέχεια προβαίνουμε σε τρεις εκτιμήσεις, χρησιμοποιώντας το spread σε υστέρηση $t-2$ (δηλαδή 2 τρίμηνα ή 6 μήνες υστέρηση) ως ερμηνευτική μεταβλητή. **Οι εκτιμήσεις (4.10), (4.11), (4.12) διαφέρουν** ως προς το χρονικό εύρος του δείγματος αλλά και την περίοδο εξέτασης της σχέσης και αποκαλύπτουν τον χρονικά μεταβαλλόμενο ρόλο της ερμηνευτικής μεταβλητής ως πρόδρομο δείκτη του ρυθμού μεγέθυνσης.

Θα παρατηρήσετε ότι και στις τρεις εκτιμήσεις αφήνουμε «εκτός» τα - μετά το 2019:4 - τρίμηνα που περιλαμβάνουν το μεγάλο «σοκ» στην οικονομία που οφείλεται στην πανδημία του κορωνοϊού λόγω του οικονομικά εξωγενούς χαρακτήρα του. Είναι γεγονός όμως ότι παρατηρήθηκε σημαντική μείωση του spread το 2018 και 2019 και «αναστρέφεται» (γίνεται αρνητικό) το 2019:3. Αυτή η κίνηση φαίνεται να συνδέεται με τις προσδοκίες των αγορών για ύπαρξη σημαντικών κινδύνων στις παγκόσμιες οικονομικές προοπτικές, συμπεριλαμβανομένης της απειλής κλιμάκωσης των διεθνών εμπορικών διαφορών. Το εάν αυτοί οι κίνδυνοι θα εκδηλωνόντουσαν σε μια ύφεση είναι άγνωστο καθώς συνέπεσαν με την έναρξη της παγκόσμιας πανδημίας.

Στην **εκτίμηση (4.10)** χρησιμοποιούμε το πλήρες δείγμα

$$\widehat{\mathrm{DYt}} = \underset{(0.34072)}{1.91791} + \underset{(0.18314)}{0.706152}\,\mathrm{spread}_{\mathrm{t}-2} \tag{4.10}$$

$$T = 265\ ,\ R^2 = 0.0499\ ,\ \hat{\sigma} = 3.3872$$

τυπικά σφάλματα σε παρενθέσεις.

Δείγμα εκτίμησης 1953:4-2019:4

Ο εκτιμητής $\hat{\beta} = 0.706$ είναι στατιστικά σημαντικός και έχει το σωστό θετικό πρόσημο, όμως η εξίσωση έχει αρκετά χαμηλό συντελεστή προσδιορισμού (περίπου 5%). Μία αύξηση του spread κατά **μία ποσοστιαία μονάδα δύο τρίμηνα πριν**, υπονοεί μία αύξηση του ετησιοποιημένου ρυθμού μεγέθυνσης κατά **0.70 ποσοστιαίες μονάδες**. Σε όρους πραγματικής οικονομικής δραστηριότητας, το συγκεκριμένο ποσοστό δεν είναι χαμηλό. Τα τελευταία 4 έτη, 2015:1 - 2019:4, ο ετησιοποιημένος ρυθμός μεγέθυνσης «έτρεχε» με μέση τιμή 2.27%.

Όμως, το 95% διάστημα εμπιστοσύνης του συντελεστή είναι μεγάλο, δείτε

προσεγγιστικά παρακάτω:

95% Διάστημα εμπιστοσύνης:

$$(0.703 - 2 \cdot 0.183\,,\ 0.703 + 2 \cdot 0.183) = (0.337\,,\ 1.069)$$

ενώ η χαμηλή ερμηνευτική ικανότητα του υποδείγματος φανερώνει έμμεσα τη σχετικά μεγάλη διακύμανση του ρυθμού μεγέθυνσης που δεν εξηγείται από το spread. Τέλος, σχετικά με την ερμηνεία του εκτιμημένου υποδείγματος, δείτε ότι ο ρυθμός μεγέθυνσης 2 τρίμηνα «εμπρός» γίνεται αρνητικός (κατ'εξοχήν σημάδι ύφεσης) όταν το spread βρίσκεται κάτω από το επίπεδο των -2.71 ποσοστιαίων μονάδων

$$\widehat{\mathrm{DYt}} < 0 \Leftrightarrow 1.917 + 0.706\,\mathrm{spread}_{t-2} < 0$$

ή

$$\mathrm{spread}_{t-2} < -\frac{1.917}{0.706} = \mathbf{-2.71}$$

Ποτέ στο δείγμα δεν έχει υπάρξει τόσο αρνητικό spread, δείτε και το **γράφημα (4.14)**.

Στην επόμενη εξίσωση χρησιμοποιούμε **το πρώτο μισό του δείγματος**, προ μίας σημαντικής μεταβολής στη νομισματική πολιτική των Η.Π.Α και κατόπιν διεθνώς. Παρατηρούμε ότι ο συντελεστής προσδιορισμού σχεδόν τετραπλασιάζεται (σχολιάστε και όλες τις υπόλοιπες μεταβολές).

$$\underset{}{\widehat{\mathrm{DYt}}} = \underset{(0.50217)}{1.46632} + \underset{(0.34269)}{1.85585}\,\mathrm{spread}_{t-2} \tag{4.11}$$

$$T = 125\ ,\ R^2 = 0.1860\ ,\ \hat{\sigma} = 4.0102$$

τυπικά σφάλματα σε παρενθέσεις.

Δείγμα εκτίμησης 1953:4-1984:4

Τέλος, παρακάτω υιοθετούμε την περίοδο 1970:1-1984:4 παρατηρώντας τη μεγάλη μεταβολή στις παραμέτρους καθώς και στον συντελεστή προσδιορισμού.

$$\widehat{\mathrm{DYt}} = \underset{(0.65124)}{0.639999} + \underset{(0.35216)}{1.92750}\,\mathrm{spread}_{t-2} \tag{4.12}$$

$$T = 60\ ,\ R^2 = 0.3292\ ,\ \hat{\sigma} = 3.7852$$

τυπικά σφάλματα σε παρενθέσεις.

Δείγμα εκτίμησης 1970:1-1984:4

4.6 Μέθοδος εκτίμησης μέγιστης πιθανοφάνειας

Σε αυτή την ενότητα θα γνωρίσουμε μία νέα μέθοδο εκτίμησης παραμέτρων ενδιαφέροντος. Η μέθοδος ονομάζεται **«μέθοδος μέγιστης πιθανοφάνειας» (maximum likelihood method, ML)** και προϋποθέτει γνώση της κατανομής από την οποία προέρχονται τα δεδομένα[15].

Έστω ένα τυχαίο δείγμα $\{z_1, z_2, ..., z_n\}$ από μία κατανομή με συνάρτηση πυκνότητας πιθανότητας

$$f\left(z|\theta\right)$$

όπου $\theta \in \Theta \subseteq \mathbb{R}^k$ είναι ένα k-διάστατο διάνυσμα άγνωστων παραμέτρων και Θ ένας δοσμένος παραμετρικός χώρος. Δηλαδή τα δεδομένα δημιουργήθηκαν με βάση μία συγκεκριμένη κατανομή $f\left(z|\cdot\right)$, η οποία χαρακτηρίζεται από την τιμή της παραμέτρου θ.

Η ανεξαρτησία[16] των $z_1, z_2, ..., z_n$ συνεπάγεται ότι η από κοινού συνάρτηση πυκνότητας πιθανότητας

$$f\left(z_1, z_2, ..., z_n|\theta\right)$$

γράφεται ως το γινόμενο των επιμέρους οριακών συναρτήσεων πυκνότητας πιθανότητας

$$f\left(z_1, z_2, ..., z_n|\theta\right) = \prod_{i=1}^{n} f\left(z_i|\theta\right)$$

[15] Στην οικονομετρία είναι συχνό φαινόμενο να μη γνωρίζουμε την κατανομή των δεδομένων ή του διαταρακτικού όρου. Στις περιπτώσεις αυτές συνήθως «επιβάλλουμε» την υπόθεση της κανονικής κατανομής παρότι κάτι τέτοιο μπορεί να μην ισχύει. Τότε, η μέθοδος εκτίμησης ονομάζεται **«μέθοδος οιονεί μέγιστης πιθανοφάνειας»** (quasi-maximum likelihood method).

[16] Ένα τυχαίο δείγμα $z_1, z_2, ..., z_n$ συνεπάγεται ότι οι $z_1, z_2, ..., z_n$ είναι ανεξάρτητες.

Η συνάρτηση πιθανοφάνειας του δείγματος (sample likelihood)

$$L_n(\theta|z_1, z_2, ..., z_n) \text{ ή } L_n(\theta)$$

ορίζεται σε αυτή την περίπτωση ως

$$L_n(\theta) = \prod_{i=1}^{n} f(z_i|\theta)$$

όπου τα τυχαία στοιχεία της συνάρτησης πυκνότητας αντικαθίστανται από τα αντίστοιχα παρατηρούμενα στοιχεία του δείγματος $z_1, z_2, ..., z_n$ και η από κοινού συνάρτηση πυκνότητας πιθανότητας μετατρέπεται σε συνάρτηση της παραμέτρου θ.

Η αρχή της μεγιστοποίησης της συνάρτησης $L_n(\theta)$ ως προς την παράμετρο θ ερμηνεύεται ως *«εύρεση (εκτίμηση) της τιμής της παραμέτρου θ που καθιστά την πιθανότητα παρατήρησης του συγκεκριμένου δείγματος $z_1, z_2, ..., z_n$ όσο το δυνατόν μεγαλύτερη»*.

Στην περίπτωση που δεν έχουμε στη διάθεσή μας ένα τυχαίο δείγμα ή στην περίπτωση που υποπτευόμαστε εξάρτηση των παρατηρήσεων του δείγματος, **π.χ., όταν μελετούμε χρονοσειρές**, τότε αναλύουμε την από κοινού συνάρτηση πυκνότητας πιθανότητας $f(z_1, z_2, ..., z_T|\theta)$ στο γινόμενο των **δεσμευμένων** συναρτήσεων πυκνότητας πιθανότητας

$$\begin{aligned} f(z_1, z_2, ..., z_T|\theta) = \\ &= f(z_T|z_{T-1}, z_{T-2}, ..., z_1; \theta) \\ &\times f(z_{T-1}|z_{T-2}, z_{T-3}, ..., z_1; \theta) \\ &\times ... \\ &\times f(z_2|z_1; \theta) \times f(z_1|\theta) \\ &= f(z_1|\theta) \times \prod_{t=2}^{T} f(z_t|\mathbb{I}_{t-1}; \theta) \end{aligned}$$

όπου $\mathbb{I}_{t-1}$ συμβολίζει το πληροφοριακό σύνολο

$$\mathbb{I}_{t-1} = \{z_{t-1}, z_{t-2}, ..., z_1\} \text{ γιά } t \geq 2$$

Ο εκτιμητής μέγιστης πιθανοφάνειας του θ θα συμβολίζεται με $\tilde{\theta}_{ML}$ και δίνεται από

$$\tilde{\theta}_{ML} = \arg\max_{\theta} L_n(\theta)$$

Ο συμβολισμός argmax (argument that maximizes) υποδηλώνει τη συχνή αδυναμία εύρεσης αναλυτικής λύσης (δηλαδή λύσης κλειστής μορφής) στο πρόβλημα μεγιστοποίησης, αφού στις περισσότερες των περιπτώσεων οι συνθήκες πρώτης τάξης

$$\frac{dL_n(\theta)}{d\theta} = 0$$

είναι μη γραμμικές εξισώσεις. Στις περιπτώσεις αυτές καταφεύγουμε σε αριθμητικές μεθόδους[17] μεγιστοποίησης της συνάρτησης πιθανοφάνειας.

Πριν προβούμε στην παρουσίαση δύο παραδειγμάτων εφαρμογής της μεθόδου, θα πρέπει να αναφέρουμε ότι το διάνυσμα των παραμέτρων θ θα πρέπει να **ταυτοποιείται** δηλαδή να είναι δυνατή η εκτίμησή του με βάση τη συνάρτηση πιθανοφάνειας. Γενικά, το διάνυσμα παραμέτρων θ ταυτοποιείται (είναι εκτιμήσιμο) όταν για οποιοδήποτε άλλο διάνυσμα παραμέτρων θ^* και για ένα σύνολο δεδομένων $z_1, ..., z_n$ έχουμε ότι $L_n(\theta^*) \neq L_n(\theta)$. Το ζήτημα της ταυτοποίησης μπορεί να συνδέεται τόσο με χαρακτηριστικά του δείγματος, π.χ., μη μεταβλητότητα των δεδομένων στο απλό υπόδειγμα παλινδρόμησης ή τέλεια γραμμική συσχέτιση των μεταβλητών ενός πολυμεταβλητού υποδείγματος που οδηγούν σε μη ταυτοποίηση, όσο και με την παραμετρική εξειδίκευσή του. Το ζήτημα της ταυτοποίησης δεν θα μας απασχολήσει περισσότερο στο παρόν επίπεδο ανάλυσης.

Τέλος, να αναφέρουμε, ότι η μέθοδος έχει και άλλες στατιστικές απαιτήσεις πέραν της γνώσης της υποκείμενης κατανομής και της ταυτοποίησης, όπως την **ομαλότητα** της συνάρτησης πιθανοφάνειας που βεβαιώνει την παραγωγισιμότητα της συνάρτησης πιθανοφάνειας κ.α.

Στη συνέχεια θα παρουσιάσουμε δύο παραδείγματα εκτίμησης παραμέτρων με άμεση εφαρμογή στην οικονομετρική πρακτική.

[17] Οι αριθμητικές μέθοδοι βασίζονται στη χρήση εξειδικευμένων υπολογιστικών προγραμμάτων τα οποία χρησιμοποιούν μαθηματικούς αλγόριθμους για την εύρεση (προσεγγιστικά) του μέγιστου της συνάρτησης. Δεν θα επεκταθούμε στο συγκεκριμένο ζήτημα.

4.6.1 Άσκηση μέγ. πιθανοφάνειας. Εκτίμηση διακύμανσης

Υποθέστε ότι

$$u_i \sim N.i.d\left(0, \sigma^2\right) \text{ με } i = 1, ..., n$$

Ζητούμενο είναι η εκτίμηση της άγνωστης παραμέτρου σ^2 αφού ο μέσος θεωρείται γνωστός και ίσος με το μηδέν. Λόγω ανεξαρτησίας των u_i , η από κοινού συνάρτηση πυκνότητας πιθανότητας δίνεται από τη σχέση

$$\begin{aligned} f\left(u_1, u_2, ..., u_n|\sigma^2\right) = & \\ & = f\left(u_1|\sigma^2\right) \times ... \times f\left(u_n|\sigma^2\right) \\ & = \prod_{i=1}^{n} f\left(u_i|\sigma^2\right) \end{aligned}$$

όπου

$$f\left(u_i|\sigma^2\right) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{u_i^2}{2\sigma^2}}$$

Εναλλακτικά αντί του συμβολισμού $e^{(\cdot)}$ χρησιμοποιούμε το συμβολισμό $\exp\{\cdot\}$ και γράφουμε

$$f\left(u_i|\sigma^2\right) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{u_i^2}{2\sigma^2}\right\}$$

Η απο-κοινού συνάρτηση πυκνότητας πιθανότητας με δεδομένη την παράμετρο σ^2 γράφεται ως

$$\begin{aligned} f\left(u_1, u_2, ..., u_n|\sigma^2\right) &= \prod_{i=1}^{n} f\left(u_i|\sigma^2\right) \\ &= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{u_i^2}{2\sigma^2}\right\} \\ &= \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{u_1^2}{2\sigma^2}\right\} \\ &\times \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{u_2^2}{2\sigma^2}\right\} \\ &\times ... \end{aligned} \tag{4.13}$$

$$\times \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{u_n^2}{2\sigma^2}\right\}$$

$$= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n} u_i^2\right\}$$

Η σχέση (4.13) δείχνει την πιθανότητα να παρατηρήσουμε το σύνολο των τιμών με δεδομένη την παράμετρο σ^2. Μεταβάλλοντας την παράμετρο σ^2 μεταβάλλουμε την αντίστοιχη πιθανότητα. Η συνάρτηση πιθανοφάνειας «αντιστρέφει» την ερμηνεία χρησιμοποιώντας τα δεδομένα του δείγματος για να εκτιμήσουμε την άγνωστη παράμετρο σ^2 με τη συνάρτηση πιθανοφάνειας

$$L_n\left(\sigma^2\right) = \prod_{i=1}^{n} f\left(u_i|\sigma^2\right)$$

να δίνει την πιθανότητα να παρατηρήσουμε τιμή σ^2 στην παράμετρο της διακύμανσης με βάση το συγκεκριμένο δείγμα.

Για να διευκολύνουμε την αλγεβρική μελέτη του προβλήματος μεγιστοποίησης λογαριθμίζουμε τη συνάρτηση πιθανοφάνειας $\ln L_n\left(\sigma^2\right)$ και καταλήγουμε στη λογαριθμική συνάρτηση πιθανοφάνειας. Ο λογαριθμικός μετασχηματισμός δεν επηρεάζει τη θέση του μέγιστου της συνάρτησης πιθανοφάνειας ως προς σ^2 και διευκολύνει σημαντικά την άλγεβρα. Δηλαδή αν το μέγιστο της $L_n\left(\sigma^2\right)$ βρίσκεται στην τιμή $\tilde{\sigma}^2_{ML}$ τότε και το μέγιστο της $\ln L_n\left(\sigma^2\right)$ βρίσκεται στην τιμή $\tilde{\sigma}^2_{ML}$. Στο συγκεκριμένο παράδειγμα, η λογαριθμική συνάρτηση πιθανοφάνειας δίνεται από την

$$\ln L_n\left(\sigma^2\right) = \ln\left[\prod_{i=1}^{n} f\left(u_i|\sigma^2\right)\right] = \sum_{i=1}^{n} \ln f\left(u_i|\sigma^2\right)$$

όπου ο όρος

$$\ln f\left(u_i|\sigma^2\right) = \ln\left[\frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{u_i^2}{2\sigma^2}\right\}\right]$$

$$= \ln\left[\left(\frac{1}{\sigma^2 2\pi}\right)^{1/2} \exp\left\{-\frac{u_i^2}{2\sigma^2}\right\}\right]$$

$$= -\frac{1}{2}\ln 2\pi - \frac{1}{2}\ln\sigma^2 - \frac{u_i^2}{2\sigma^2}$$

συμβολίζει τη «συνεισφορά» κάθε παρατήρησης στη λογαριθμική συνάρτηση πιθανοφάνειας. Μία σύντομη αλγεβρική απλοποίηση δίνει την τελική μορφή της συνάρτησης που θα μεγιστοποιήσουμε ως προς την παράμετρο ενδιαφέροντος,

$$\begin{aligned}\ln L_n\left(\sigma^2\right) &= \sum_{i=1}^{n}\ln f\left(u_i|\sigma^2\right)\\ &= \sum_{i=1}^{n}\left[-\frac{1}{2}\ln 2\pi - \frac{1}{2}\ln\sigma^2 - \frac{u_i^2}{2\sigma^2}\right]\\ &= -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}u_i^2\end{aligned}$$

Η μέθοδος μέγιστης λογαριθμικής πιθανοφάνειας (maximum log-likelihood) συνίσταται στη μεγιστοποίηση της συνάρτησης $\ln L_n\left(\sigma^2\right)$ ως προς σ^2 δηλαδή απαντά στο ερώτημα ποια συγκεκριμένη τιμή της παραμέτρου «υποστηρίζεται» περισσότερο από ένα δεδομένο δείγμα.

Στο συγκεκριμένο παράδειγμα, το πρόβλημα είναι αλγεβρικά απλό και δίνει λύση ανηγμένης ή αναλυτικής μορφής για τον εκτιμητή. Δηλαδή «μπορούμε» να λύσουμε την εξίσωση που παράγεται από τη συνθήκη πρώτης τάξεως $\frac{d\ln L_n}{d\sigma^2} = 0$.

Συγκεκριμένα, η **συνθήκη πρώτης τάξεως** δίνει την εξίσωση

$$\frac{d\ln L_n}{d\sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\left(\sigma^2\right)^2}\sum_{i=1}^{n}u_i^2 = 0$$

η οποία λύνεται ως προς σ^2 και

$$\frac{d\ln L_n}{d\sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\left(\sigma^2\right)^2}\sum_{i=1}^{n}u_i^2 = 0$$

$$\Rightarrow \tilde{\sigma}^2_{ML} = \frac{1}{n}\sum_{i=1}^{n}u_i^2$$

Η **συνθήκη δεύτερης τάξεως** για μέγιστο ικανοποιείται αφού

$$\left.\frac{d^2 \ln L_n}{d\left(\sigma^2\right)^2}\right|_{\sigma^2=\tilde{\sigma}^2_{ML}} = -\frac{n}{2\left(\tilde{\sigma}^2_{ML}\right)^2} < 0$$

Άρα ο εκτιμητής μέγιστης πιθανοφάνειας $\tilde{\sigma}^2_{ML}$ της διακύμανσης σ^2 δηλαδή ο εκτιμητής που μεγιστοποιεί τη συνάρτηση λογαριθμικής πιθανοφάνειας δίνεται από την

$$\tilde{\sigma}^2_{ML} = \frac{1}{n}\sum_{i=1}^{n} u_i^2$$

4.6.2 Άσκηση μέγ. πιθανοφάνειας. Οι εκτιμητές στο απλό γραμμικό υπόδειγμα

Στη δεύτερη άσκηση, σχετική με το κλασσικό γραμμικό υπόδειγμα

$$y_i = \alpha + \beta x_i + u_i\,,\; u_i \sim N.i.d(0, \sigma_u^2)\,,\; i = 1, ..., n$$

σκοπός μας είναι η εκτίμηση των παραμέτρων του πληθυσμού

$$\theta = \left(\alpha\;,\;\beta\;,\;\sigma_u^2\right)'$$

με τη μέθοδο μέγιστης πιθανοφάνειας και η σύγκριση των εκτιμητών

$$\tilde{\theta}_{ML} = \left(\tilde{\alpha}_{ML}\;,\;\tilde{\beta}_{ML}\;,\;\tilde{\sigma}^2_{u,ML}\right)'$$

με αυτών της μεθόδου ελαχίστων τετραγώνων (ΕΤ).

Για ευκολία, θεωρήστε την ερμηνευτική μεταβλητή x_i, $i = 1, ..., n$ ως μη στοχαστική. Η λογαριθμική συνάρτηση πιθανοφάνειας

$$\ln L_n\left(\alpha, \beta, \sigma_u^2\right)$$

δίνεται αναλυτικά από την

$$\ln L_n\left(\alpha, \beta, \sigma_u^2\right) = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln \sigma_u^2 - \frac{1}{2\sigma_u^2}\sum_{i=1}^{n} u_i^2$$

$$= -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\sigma_u^2 - \frac{1}{2\sigma_u^2}\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2$$

Οι **συνθήκες πρώτης τάξης** δίνονται από τις εξισώσεις

$$\frac{\partial \ln L_n}{\partial \alpha} = 0 \Rightarrow \sum_{i=1}^{n}(y_i - \alpha - \beta x_i) = 0$$

$$\frac{\partial \ln L_n}{\partial \beta} = 0 \Rightarrow \sum_{i=1}^{n}(y_i - \alpha - \beta x_i)\, x_i = 0$$

$$\frac{\partial \ln L_n}{\partial \sigma_u^2} = 0 \Rightarrow$$

$$-\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2 = 0$$

οι οποίες έχουν λύση

$$\tilde{\alpha}_{ML} = \bar{y} - \tilde{\beta}_{ML}\bar{x}$$

$$\tilde{\beta}_{ML} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\tilde{\sigma}_{u,ML}^2 = \frac{1}{n}\sum_{i=1}^{n}\hat{u}_i^2$$

με τους εκτιμημένους διαταρακτικούς όρους

$$\hat{u} = y_i - \tilde{\alpha}_{ML} - \tilde{\beta}_{ML}x_i$$

να ταυτίζονται με τα κατάλοιπα της μεθόδου των ΕΤ.

Ως άσκηση, μπορούμε να διαπιστώσουμε ότι η εσσιανή μήτρα των δεύτερων μερικών παραγώγων

$$\frac{\partial^2 \ln L_n}{\partial\theta\partial\theta'}$$

είναι αρνητικά ορισμένη όταν υπολογιστεί στο στάσιμο σημείο $\tilde{\theta}_{ML}$ οπότε **ικα-**

νοποιούνται οι συνθήκες δεύτερης τάξης για μεγιστοποίηση και οι εκτιμητές $\tilde{\alpha}_{ML}, \tilde{\beta}_{ML}, \tilde{\sigma}^2_{u,ML}$ όντως μεγιστοποιούν τη συνάρτηση λογαριθμικής πιθανοφάνειας $\ln L_n\left(\alpha, \beta, \sigma^2_u\right)$.

Παρατηρήστε ότι οι εκτιμητές μέγιστης πιθανοφάνειας $\tilde{\alpha}_{ML}$ και $\tilde{\beta}_{ML}$ είναι **ίδιοι ακριβώς**, στη συγκεκριμένη περίπτωση, με τους εκτιμητές ΕΤ $\hat{\alpha}_{ET}, \hat{\beta}_{ET}$ **ενώ η εκτίμηση μέγιστης πιθανοφάνειας της διακύμανσης του διαταρακτικού όρου**

$$\tilde{\sigma}^2_{u,ML} = \frac{1}{n}\sum_{i=1}^{n} \hat{u}_i^2$$

διαφέρει από τον αντίστοιχο εκτιμητή ελαχίστων τετραγώνων

$$\hat{\sigma}^2_u = \frac{1}{n-2}\sum_{i=1}^{n} \hat{u}_i^2$$

αφού

$$\tilde{\sigma}^2_{ML} = \frac{n-2}{n}\hat{\sigma}^2_u$$

Η παραπάνω διαφορά $\tilde{\sigma}^2_{ML} \neq \hat{\sigma}^2_u$ τονίζει ότι οι εκτιμητές μέγιστης πιθανοφάνειας μπορεί να μην έχουν βέλτιστες ιδιότητες σε μικρά (πεπερασμένα) δείγματα, δηλαδή σε δείγματα με καθορισμένο μέγεθος n, εκτός ίσως από συγκεκριμένες κατηγορίες κατανομών.

Για παράδειγμα, ο εκτιμητής μέγιστης πιθανοφάνειας της διακύμανσης των διαταρακτικών όρων του απλού γραμμικού υποδείγματος είναι **μεροληπτικός** αφού

$$E\left(\tilde{\sigma}^2_{ML}\right) = E\left(\frac{n-2}{n}\cdot\hat{\sigma}^2_u\right) = \frac{n-2}{n}\cdot E\left(\hat{\sigma}^2_u\right)$$

$$= \frac{n-2}{n}\cdot\sigma^2 = \sigma^2 - \frac{2}{n}\sigma^2 \neq \sigma^2$$

4.7 Ιδιότητες εκτιμητών μέγιστης πιθανοφάνειας

Κάτω από συγκεκριμένες συνθήκες «ομαλότητας» (regularity conditions) της συνάρτησης πιθανοφάνειας, οι εκτιμητές μέγιστης πιθανοφάνειας επιδεικνύουν έναν αριθμό ελκυστικών ασυμπτωτικών[18] ιδιοτήτων που τους καθιστούν δημο-

[18] Ασυμπτωτικά σημαίνει καθώς το δείγμα τείνει στο άπειρο (θεωρητικά).

φιλείς στην εφαρμοσμένη ανάλυση. Οι ιδιότητες αυτές συνοψίζονται παρακάτω και θα γίνουν κατανοητές όταν ασχοληθούμε με την ασυμπτωτική ανάλυση στο κεφάλαιο 8:

(α) καθώς το δείγμα τείνει στο άπειρο - δηλαδή καθώς το μέγεθος του δείγματος αυξάνει ή αλλιώς σε μεγάλα δείγματα - οι εκτιμητές συμπίπτουν με τις παραμέτρους που εκτιμούν[19]

(β) καθώς το δείγμα τείνει στο άπειρο, οι εκτιμητές είναι αποτελεσματικοί δηλαδή έχουν τη μικρότερη δυνατή διακύμανση ή αλλιώς επιτυγχάνουν τη μέγιστη δυνατή ακρίβεια,

(γ) καθώς το δείγμα τείνει στο άπειρο, κατάλληλα τυποποιημένες συναρτήσεις των εκτιμητών κατανέμονται κανονικά

Οι **(α)**, **(β)**, **(γ)** συνοψίζονται στις παρακάτω δύο σχέσεις στην περίπτωση που υπάρχει συνεπής[20] εκτιμητής $\hat{\mathbf{I}}_\theta$ για την $\mathbf{I}_\theta$ και η επαγωγή γίνεται προσεγγιστικά,

$$\tilde{\theta}_{ML} \overset{\alpha}{\sim} N\left(\theta\,,\, \hat{\mathbf{I}}_\theta^{-1}\right) \quad \text{όπου}$$

$$\mathbf{I}_\theta = -E\left(\frac{\partial^2 \ln L_n}{\partial\theta\partial\theta'}\right)$$

με τον συμβολισμό $\overset{\alpha}{\sim}$ να διαβάζεται «κατανέμεται προσεγγιστικά»[21].

Η μήτρα $\mathbf{I}_\theta$ ονομάζεται **μήτρα πληροφορίας** (information matrix), της οποίας η αντίστροφη μήτρα $\mathbf{I}_\theta^{-1}$ αποτελεί την προσεγγιστική μήτρα διακυμάνσεων - συνδιακυμάνσεων του εκτιμητή μέγιστης πιθανοφάνειας. Για τις ανάγκες του μαθήματος, ο συνεπής εκτιμητής $\hat{\mathbf{I}}_\theta$ θα λαμβάνεται αντικαθιστώντας τις παραμέτρους θ που εμφανίζονται στη μήτρα $\mathbf{I}_\theta$ με τις αντίστοιχες εκτιμήσεις μέγιστης πιθανοφάνειας $\tilde{\theta}_{ML}$.

Για παράδειγμα στην πρώτη άσκηση εκτίμησης που είδαμε, η μήτρα πληροφο-

[19] Ιδιότητα συνέπειας. Θα γίνει επίσης κατανοητή στο κεφάλαιο 8.

[20] Στο κεφάλαιο 8 θα αναφερθούμε αναλυτικά στην ιδιότητα της συνέπειας.

[21] Ο συμβολισμός $\overset{\alpha}{\sim}$ θα γίνει επίσης κατανοητός στο κεφάλαιο 8 που παρουσιάζεται η ασυμπτωτική ανάλυση.

ρίας υπολογίζεται ως εξής:

$$\begin{aligned}\mathbf{I}_\theta &= -E\left(\frac{\partial^2 \ln L_n}{\partial\theta\partial\theta'}\right)\\ &= -E\left(\frac{d^2 \ln L_n}{d\left(\sigma^2\right)^2}\right) = E\left(\frac{n}{2\left(\sigma^2\right)^2}\right)\\ &= \frac{n}{2\left(\sigma^2\right)^2}\end{aligned}$$

και ο εκτιμητής μέγιστης πιθανοφάνειας της διακύμανσης κατανέμεται προσεγγιστικά κανονικά

$$\tilde{\sigma}^2_{ML} \overset{\alpha}{\sim} N\left(\sigma^2,\, \frac{2\left(\tilde{\sigma}^2_{ML}\right)^2}{n}\right)$$

4.8 Ασκήσεις

1. Το αρχείο kefalaio4data1.xlsx περιέχει μία χρονοσειρά με $T = 24$. Θεωρήστε ότι τα δεδομένα είναι μηνιαίας συχνότητας.

   (α) Εκτιμήστε με τη μέθοδο ελαχίστων τετραγώνων (ΕΤ) το **γραμμικό υπόδειγμα τάσης**

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

   (β) Σχεδιάστε τη χρονοσειρά y_t, τη χρονοσειρά y_t μαζί με τις εκτιμημένες τιμές $\hat{y}_t$ καθώς και τη χρονοσειρά y_t μαζί με την πρόβλεψή της για τους επόμενους 12 μήνες

   (γ) Υπολογίστε το 95% διάστημα εμπιστοσύνης της πρόβλεψης $\hat{y}_{T+h}$ για ορίζοντα πρόβλεψης μέχρι και 12 μήνες μπροστά, δηλαδή για $h = 1, 2, ..., 12$

2. Έστω ότι A_t συμβολίζει τη μηνιαία χρονοσειρά του ποσοστού ανεργίας στην Ελλάδα. Έχετε στη διάθεσή σας 42 μηνιαίες παρατηρήσεις από τον Μάιο του 2008 μέχρι τον Οκτώβριο του 2011.

   (α) Εκτιμήστε με τη μέθοδο ελαχίστων τετραγώνων τους συντελεστές του παρακάτω απλού γραμμικού υποδείγματος (χρονικής τάσης)

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

   έχοντας στη διάθεσή σας τα δεδομένα

   - $\sum_{t=1}^{42} \left(A_t - \bar{A}\right)^2 = 488.385,$
   - $\sum_{t=1}^{42} \left(t - \bar{t}\right)^2 = 6170.5,$
   - $\sum_{t=1}^{42} \left(A_t - \bar{A}\right)\left(t - \bar{t}\right) = 1692.058,$
   - $\bar{A} = \frac{1}{42} \sum_{t=1}^{42} A_t = 11.6116,$

- $\bar{t} = \frac{1}{42} \sum_{t=1}^{42} t = 21.5$

(β) Σας δίνεται η πληροφορία ότι το άθροισμα των τετραγώνων των καταλοίπων είναι ίσο με

$$\hat{u}'\hat{u} = 24.3934$$

Προβείτε σε έλεγχο στατιστικής σημαντικότητας του εκτιμητή κλίσης. Αποφανθείτε αν το ποσοστό ανεργίας εμφανίζει θετική ανοδική τάση. **Υπόδειξη:** επιλέξτε μία κρίσιμη τιμή από τις $t_2^{0.05/2} = 4.302$, $t_{40}^{0.05/2} = 2.021$ (δικατάληκτος έλεγχος).

(γ) Προβείτε σε πρόβλεψη για το ποσοστό ανεργίας τους μήνες Νοέμβριο 2011, Δεκέμβριο 2011, Ιανουάριο 2012.

(δ) Υπολογίστε τα 95% διαστήματα εμπιστοσύνης των παραπάνω προβλέψεων.

3. Το αρχείο kefalaio4data2.xlsx (φύλλο εργασίας ΑΕΠ) περιέχει ετήσια δεδομένα για το πραγματικό (τιμές 2005) ΑΕΠ της Ελλάδος από τη βάση AMECO[22].

(α) Σχεδιάστε τα δεδομένα στον χρόνο, δηλαδή παρουσιάστε τα σε ένα χρονογράφημα

(β) Εκτιμήστε ένα υπόδειγμα γραμμικής τάσης για τα διαστήματα 1980 - 1993, 1994 - 2007 και ερμηνεύστε το συντελεστή κλίσης δηλαδή τον συντελεστή της τάσης

(γ) Εκτιμήστε ένα λογαριθμικό-γραμμικό υπόδειγμα για τα διαστήματα 1980 - 1993, 1994 - 2007 και ερμηνεύστε το συντελεστή κλίσης

4. Το αρχείο

kefalaio4data2.xlsx
φύλλο εργασίας ΔΒΠ_RT_ΔΤΚ_UN

περιέχει στοιχεία[23] για τον δείκτη βιομηχανικής παραγωγής (ΔΒΠ) στην Ελλάδα, για τον δείκτη λιανικών πωλήσεων RT, για τον δείκτη τιμών κατα-

[22] Κωδικός σειράς OVGD στην ιστοσελίδα http://ec.europa.eu/economy_finance/ameco/user/serie/SelectSerie.cfm

[23] Τα στοιχεία είναι από τον Ο.Ο.Σ.Α (OECD), http://stats.oecd.org/Index.aspx

ναλωτή (ΔΤΚ) και για το εναρμονισμένο ποσοστό ανεργίας (harmonized unemployment rate).

(α) Σχεδιάστε τις χρονοσειρές.

(β) Υπολογίστε τον % ετήσιο ρυθμό μεταβολής του ΔΒΠ

$$g_t = 100 \times \left(\frac{\Delta B\Pi_t - \Delta B\Pi_{t-12}}{\Delta B\Pi_{t-12}} \right)$$

του δείκτη λιανικών πωλήσεων

$$rt_t = 100 \times \left(\frac{RT_t - RT_{t-12}}{RT_{t-12}} \right)$$

και του ΔΤΚ (% ετήσιος πληθωρισμός)

$$\pi_t = 100 \times \left(\frac{\Delta TK_t - \Delta TK_{t-12}}{\Delta TK_{t-12}} \right)$$

Εκτιμήστε ένα γραμμικό - γραμμικό και ένα λογαριθμικό - γραμμικό υπόδειγμα τάσης για τη σειρά του ετήσιου πληθωρισμού το διάστημα Ιαν. 1994 - Δεκ. 2000.

(γ) Εκτιμήστε τη συνάρτηση αυτοσυσχέτισης των χρονοσειρών g_t, rt_t, π_t για τις πρώτες 24 υστερήσεις, $k = 1, 2, ..., 24$, και ελέγξτε τη στατιστική σημαντικότητα των εκτιμημένων συντελεστών αυτοσυσχέτισης. Σχεδιάστε τις εκτιμημένες συναρτήσεις αυτοσυσχέτισης (κορρελογράμματα). Τι παρατηρείτε;

5. Το αρχείο kefalaio4data3.xlsx περιέχει δεδομένα από ένα διαγωνισμό του ΑΣΕΠ με κλίμακα βαθμολογίας του εξεταζόμενου από (0-100). Η μεταβλητή «Βαθμός» περιέχει τα αποτελέσματα για 4986 άτομα που πέτυχαν βαθμολογία 25 και άνω. Στο αρχείο, επίσης, περιέχονται δεδομένα για τις μεταβλητές του φύλου (ψευδομεταβλητή, λαμβάνει την τιμή 1 όταν ο διαγωνιζόμενος είναι γυναίκα), της ηλικίας και τρεις μεταβλητές εκπαίδευσης (επίσης ψευδομεταβλητές, λαμβάνουν την τιμή 1 όταν ο διαγωνιζόμενος έχει μόνο απολυτήριο Λυκείου, 1 αν είναι μόνο πτυχιούχος ΑΤΕΙ και 1 αν είναι πτυχιούχος ΑΕΙ αντίστοιχα). Έστω ότι Y_i η εξαρτημένη μεταβλητή του βαθμού και X_i η ερμηνευτική μεταβλητή του φύλου, ή της ηλικίας ή κάθε μεταβλητής εκπαίδευσης.

(α) Εκτιμήστε ένα γραμμικό - γραμμικό $Y_i = \alpha + \beta X_i + u_i$ και ένα λογαριθμικό - γραμμικό $\ln(Y_i) = \alpha + \beta X_i + u_i$ υπόδειγμα για όλες τις περιπτώσεις και προβείτε σε ερμηνεία των εκτιμημένων συντελεστών κλίσης και των σταθερών όρων όπου κρίνετε εσείς απαραίτητο. Υπολογίστε τους συντελεστές προσδιορισμού σε κάθε περίπτωση.

(β) Εκτιμήστε επίσης ένα λογαριθμικό - λογαριθμικό υπόδειγμα $\ln(Y_i) = \alpha + \beta \ln(X_i) + u_i$ με X_i την ηλικία του διαγωνιζόμενου. Προβείτε σε ερμηνεία του εκτιμημένου συντελεστή κλίσης.

6. Δείξτε ότι αν η χρονοσειρά u_t, $t = 1, ..., T$ είναι λευκός θόρυβος, τότε η σύνθετη χρονοσειρά

$$z = u_t u_{t-1}$$

είναι μία σειρά διαφορών martingale.

7. Έστω ότι η χρονοσειρά x_t ακολουθεί ένα AR(1) υπόδειγμα

$$x_t = \varphi x_{t-1} + \nu_t \quad , \quad |\varphi| < 1$$

άρα συσχετίζεται σειριακά. Έστω ότι

$$\nu_t \sim i.i.d\left(0, \sigma_\nu^2\right)$$

Δείξτε ότι η σύνθετη χρονοσειρά

$$z_t = x_t \nu_{t-1} - \varphi x_{t-1} \nu_{t-1}$$

είναι μία σειρά διαφορών martingale ως προς το σύνολο πληροφορίας $\mathbb{I}_t = \{\nu_t, \nu_{t-1}, ...\}$

8. Έστω το AR(1) υπόδειγμα

$$y_t = 0.6 + \varphi y_{t-1} + u_t \ , \ t = 1, ..., 200 \ , \ y_0 = 0$$

$$u_t \sim N.i.d\,(0\,,\,0.25)$$

με $t = 1, ..., 200$, $y_0 = 0$, $u_t \sim N.i.d\,(0\,,\,0.25)$ και $\varphi = 0.75$

(α) Σε πρόγραμμα της επιλογής σας - π.χ., στο Excel ή στο gretl - δη-

μιουργήστε τη χρονοσειρά y_t και σχεδιάστε τη στο χρόνο.

(β) Υπολογίστε το μέσο και τη διακύμανση της y_t.

(γ) Υπολογίστε και σχεδιάστε τη συνάρτηση αυτοσυσχέτισης $\hat{\rho}(k)$ για $k = 0, 1, 2, 3, 4, 5, 6$ καθώς και τη θεωρητική συνάρτηση αυτοσυσχέτισης $\rho(k)$

(δ) Σχεδιάστε τη χρονοσειρά y_t σε ένα διάγραμμα όταν

$$\varphi = \left\{ \ 0, \quad 0.25, \quad 0.50, \quad 0.76, \quad 0.95, \quad 0.99 \ \right\}.$$

Τι παρατηρείτε;

9. Ποιές από τις παρακάτω χρονοσειρές $AR(2)$ ή διαδικασίες $AR(2)$

$$\begin{aligned}
(\alpha): \ y_t &= 5 + 0.65y_{t-1} + 0.15y_{t-2} + u_t \\
(\beta): \ y_t &= 5 + 1.25y_{t-1} - 0.75y_{t-2} + u_t \\
(\gamma): \ y_t &= 5 + 0.75y_{t-1} + 0.55y_{t-2} + u_t \\
(\delta): \ y_t &= 5 + 1.25y_{t-1} - 0.25y_{t-2} + u_t
\end{aligned}$$

είναι στάσιμες και ποιες μη στάσιμες; Η χρονοσειρά u_t είναι λευκός θόρυβος, δηλαδή ικανοποιεί τις συνθήκες

$$E(u_t) = 0 \ , \ Var(u_t) = \sigma_u^2 \ , \ Cov(u_t, u_{t-k}) = 0 \ , \ \forall k \neq 0$$

Απάντηση

Παρατηρήστε ότι $AR(2)$ διαδικασίες ή χρονοσειρές που ακολουθούν το υπόδειγμα $AR(2)$ μπορούν να γραφούν ως

$$\begin{aligned}
y_t &= a + \phi_1 y_{t-1} + \phi_2 y_{t-2} + u_t \Rightarrow \\
y_t - \phi_1 y_{t-1} - \phi_2 y_{t-2} &= a + u_t \Rightarrow \\
\left(1 - \phi_1 L - \phi_2 L^2\right) y_t &= a + u_t \Rightarrow \\
\phi(L)\, y_t &= a + u_t
\end{aligned}$$

όπου $\phi(L) = 1 - \phi_1 L - \phi_2 L^2$ ένα πολυώνυμο δεύτερης τάξης του τελεστή υστέρησης. Βρίσκουμε τις ρίζες r_1, r_2 του πολυωνύμου

$$1 - \phi_1 x - \phi_2 x^2 = 0$$

το οποίο έχει διακρίνουσα

$$\Delta = (-\phi_1)^2 - 4(-\phi_2) = \phi_1^2 + 4\phi_2$$

Το πολυώνυμο δεύτερης τάξης έχει το πολύ δύο ρίζες οι οποίες

(i) όταν $\Delta > 0$ τότε οι ρίζες είναι **πραγματικές** και δίνονται από

$$\Delta > 0 \Rightarrow \begin{cases} r_1 = -\frac{\phi_1}{2\phi_2} - \frac{\sqrt{\Delta}}{2\phi_2} \\ r_2 = -\frac{\phi_1}{2\phi_2} + \frac{\sqrt{\Delta}}{2\phi_2} \end{cases}$$

(ii) όταν $\Delta = 0$ τότε έχουμε **μία διπλή πραγματική ρίζα**, δηλαδή $r_1 = r_2$, που δίνεται από

$$\Delta = 0 \Rightarrow \begin{cases} r_1 = -\frac{\phi_1}{2\phi_2} \\ r_2 = -\frac{\phi_1}{2\phi_2} \end{cases}$$

(iii) όταν $\Delta < 0$ τότε οι ρίζες είναι **μιγαδικές** και δίνονται από

$$\Delta < 0 \Rightarrow \begin{cases} r_1 = -\frac{\phi_1}{2\phi_2} - \frac{i\sqrt{-\Delta}}{2\phi_2} \\ r_2 = -\frac{\phi_1}{2\phi_2} + \frac{i\sqrt{-\Delta}}{2\phi_2} \end{cases}$$

δηλαδή

$$r_{1,2} = \left(-\frac{\phi_1}{2\phi_2}\right) \pm \left(\frac{\sqrt{-\Delta}}{2\phi_2}\right)$$

ή $r_{1,2} = a \pm bi$ όπου $a = -\frac{\phi_1}{2\phi_2}$ το πραγματικό μέρος της ρίζας και $b = \left(\frac{\sqrt{-\Delta}}{2\phi_2}\right)$ το μιγαδικό μέρος της ρίζας.

Αν οι ρίζες r_i ικανοποιούν τις σχέσεις

(○) $|r_i| > 1$ όταν r_i πραγματικές

(○) και $\sqrt{a^2 + b^2} > 1$ όταν r_i μιγαδική

τότε η διαδικασία y_t είναι στάσιμη.

Αν μία από τις δύο ρίζες είναι ίση[24] με 1, δηλαδή $|r_i| = 1$ για $i = 1$ ή $i = 2$, τότε έχουμε **μοναδιαία ρίζα (unit root)** δηλαδή **μη στασι-**

[24] Αν έστω μία ρίζα είναι μεγαλύτερη της μονάδας, δηλαδή $|r_i| > 1$ για $i = 1$ ή $i = 2$, τότε έχουμε μία εκρηκτική χρονοσειρά (explosive time series) που δεν απαντάται στην οικονομι-

μότητα τύπου $I(1)$ και η χρονοσειρά y_t **ακολουθεί το υπόδειγμα του τυχαίου περιπάτου (έχει στοχαστική τάση)**.

ΣΗΜΕΙΩΣΗ: Στην περίπτωση που και οι δύο ρίζες του πολυωνύμου είναι μοναδιαίες, λέμε ότι έχουμε **δύο μοναδιαίες ρίζες** και η χρονοσειρά y_t είναι μη στάσιμη τύπου $I(2)$. Για παράδειγμα, αν η y_t δημιουργείται σύμφωνα με το υπόδειγμα

$$y_t = \alpha + 2y_{t-1} - y_{t-2} + u_t$$

τότε είναι $I(2)$ αφού

$$(1-L)(1-L)y_t = (1-L)^2 y_t = y_t - 2y_{t-1} + y_{t-2}$$

Το τελευταίο αποδεικνύεται ως εξής: κάθε πολυώνυμο της μορφής

$$1 - \phi_1 x - \phi_2 x^2$$

μπορεί να γραφεί ως

$$1 - \phi_1 x - \phi_2 x^2 = \left(1 - \frac{1}{r_1}x\right)\left(1 - \frac{1}{r_2}x\right)$$

όπου r_1, r_2 οι ρίζες του πολυωνύμου.

Σε περίπτωση μοναδιαίας ρίζας, έστω ότι $r_1 = 1$, έχουμε

$$\phi(L)\, y_t = a + u_t \Rightarrow$$

$$\left(1 - \frac{1}{r_1}x\right)\left(1 - \frac{1}{r_2}x\right) y_t = a + u_t \Rightarrow$$

$$(1-L)\left(1 - \frac{1}{r_2}x\right) y_t = a + u_t \Rightarrow$$

$$(1-L)(1-\rho L)\, y_t = a + u_t \Rightarrow$$

$$(1-L)\, y_t = \frac{a}{(1-\rho L)} + \frac{u_t}{(1-\rho L)}$$

κή. Μπορείτε να προβείτε σε προσομοίωση τέτοιων χρονοσειρών και θα διαπιστώσετε ότι οι γραφικές τους παραστάσεις είναι τελείως «ξένες» με παρατηρούμενες οικονομικές χρονοσειρές.

όπου $|\rho| = \left|\frac{1}{r_2}\right| < 1$ αφού η δεύτερη ρίζα είναι εκτός του μοναδιαίου κύκλου $|r_2| > 1$.

Άρα υιοθετώντας την ιδιότητα του τελεστή υστέρησης $L^j\alpha = \alpha$ για α σταθερά, (οπότε λαμβάνουμε και $\rho L = \rho$), και το μαθηματικό αποτέλεσμα

$$\sum_{j=0}^{+\infty} \rho^j = \frac{1}{1-\rho} \text{ όταν } |\rho| < 1$$

έχουμε

$$(1-L)\, y_t = \frac{a}{(1-\rho L)} + \frac{u_t}{(1-\rho L)} \Rightarrow$$

$$(1-L)\, y_t = \frac{a}{1-\rho} + \sum_{j=0}^{+\infty} \rho^j u_t$$

$$= \frac{a}{1-\rho} + \sum_{j=0}^{+\infty} (\rho L)^j\, u_t$$

$$= \frac{a}{1-\rho} + \sum_{j=0}^{+\infty} \rho^j L^j u_t$$

$$= \frac{a}{1-\rho} + \sum_{j=0}^{+\infty} \rho^j u_{t-j}$$

Όμως $\sum_{j=0}^{+\infty} \rho^j u_{t-j}$ αντιστοιχεί σε μία οποιαδήποτε AR(1) στάσιμη διαδικασία μηδενικού μέσου

$$e_t = \rho e_{t-1} + u_t$$

Συνεπώς, η περίπτωση μοναδιαίας ρίζας στα υποδείγματα $AR(2)$ με u_t λευκό θόρυβο οδηγεί σε μη στασιμότητα της y_t η οποία περιέχει μία στοχαστική τάση

$$(1-L)\, y_t = \frac{a}{1-\rho} + \sum_{j=0}^{+\infty} \rho^j u_{t-j} \Rightarrow$$

$$y_t = y_{t-1} + \frac{a}{1-\rho} + e_t \Rightarrow$$

$$y_t = y_0 + \left(\frac{a}{1-\rho}\right) t + \sum_{j=1}^{t} e_j$$

του τύπου $\sum_{j=1}^{t} e_j$ και αντίστοιχα σε στασιμότητα της $\Delta y_t = (1-L)\, y_t$, η οποία υποδειγματοποιείται ως μία στάσιμη AR(1) διαδικασία e_t.

Επιστρέφουμε στη λύση της άσκησης και έχουμε για κάθε διαδικασία ξεχωριστά

(α) Πολυώνυμο $1 - 0.65x - 0.15x^2$, ρίζες $r_1 = 1.204$, $r_2 = -5.5373$. Επειδή $|r_i| > 1$ γιά κάθε i έχουμε **στασιμότητα** της y_t

(β) Πολυώνυμο $1-1.25x+0.75x^2$, μιγαδικές ρίζες $r_1 = 0.83333-0.79931i$, $r_2 = 0.83333 + 0.79931i$. Επειδή

$$\sqrt{(0.83333)^2 + (0.79931)^2} = 1.1547 > 1$$

έχουμε **στασιμότητα** της y_t

(γ) Πολυώνυμο $1 - 0.75x - 0.55x^2$, ρίζες $r_1 = -2.1928$, $r_2 = 0.82916$. Επειδή $|r_2| < 1$ έχουμε **μη στασιμότητα**. Όμως η διαδικασία είναι **«εκρηκτική» στο χρόνο, δηλαδή δεν είναι μη στασιμότητα τύπου μοναδιαίας ρίζας**.

(δ) Πολυώνυμο $1-1.25x+0.25x^2$, ρίζες $r_1 = 1$, $r_2 = 4$ και επειδή $|r_1| = 1$ έχουμε **μη στασιμότητα τύπου** $I(1)$ δηλαδή **μοναδιαία ρίζα** και **υπόδειγμα τυχαίου περιπάτου** αφού μπορούμε να ξαναγράψουμε την y_t ως

$$(1-L)\left(1 - \frac{1}{4}L\right) y_t = 5 + u_t \Rightarrow$$

$$\Delta y_t = \frac{5}{1-0.25} + \frac{u_t}{1-0.25L} \Rightarrow$$

$$\Delta y_t = 6.6 + \sum_{j=0}^{+\infty} (0.25)^j\, u_{t-j} \Rightarrow$$

$$\Delta y_t = 6.6 + e_t \Rightarrow$$

$$y_t = 6.6 + y_{t-1} + e_t$$

όπου $e_t = \sum_{j=0}^{+\infty} (0.25)^j \, u_{t-j}$ μία στάσιμη AR(1) διαδικασία $e_t = 0.25e_{t-1} + u_t$

10. Δημιουργήστε στο Excel ή στο gretl και σχεδιάστε τις παρακάτω διαδικασίες $AR(2)$

$$\begin{aligned}
(i) &: y_t = 0.8y_{t-1} - 0.2y_{t-2} + u_t \\
(ii) &: y_t = 1.1y_{t-1} + 0.5y_{t-2} + u_t \\
(iii) &: y_t = 0.7y_{t-1} + 0.2y_{t-2} + u_t \\
(iv) &: y_t = 0.65y_{t-1} + 0.11y_{t-2} + u_t \\
(v) &: y_t = 1.24y_{t-1} - 0.6y_{t-2} + u_t
\end{aligned}$$

θέτοντας $y_0 = 0$, $u_t \sim N.i.d\,(0,1)$ και μέγεθος δείγματος της επιλογής σας. Ποιες από αυτές είναι στάσιμες;

11. Έστω ότι α αντιστοιχεί σε μία Bernoulli τυχαία μεταβλητή που λαμβάνει τις τιμές $1/t$ με πιθανότητα $p = 0.25$ και 0 με πιθανότητα $1 - p$ και $u_t \sim N.i.d\,(0,1)$. Είναι η χρονοσειρά $y_t = \alpha t + u_t$ (ασθενώς) στάσιμη;

Απάντηση

Έχουμε ότι

$$E(\alpha) = \frac{1}{t}p + 0\,(1-p) = \frac{p}{t}$$

και

$$\begin{aligned}
Var(\alpha) &= E\left(\alpha^2\right) - (E(\alpha))^2 \\
&= \frac{1}{t^2}p + 0\,(1-p) - \frac{p^2}{t^2} \\
&= \frac{p}{t^2} - \frac{p^2}{t^2} \\
&= \frac{p\,(1-p)}{t^2}
\end{aligned}$$

Οπότε

$$E(y_t) = E(\alpha t + u_t) = tE(\alpha) = \frac{pt}{t} = p$$

$$\begin{aligned} Var(y_t) &= Var(y_t) + Var(u_t) \\ &= t^2 Var(\alpha) + 1 \\ &= \frac{t^2 p(1-p)}{t^2} + 1 \\ &= p(1-p) + 1 \end{aligned}$$

ενώ

$$\begin{aligned} E(y_t y_{t-k}) &= E[(\alpha t + u_t)(\alpha(t-k) + u_t)] \\ &= p - k\frac{p}{t} = p\left(1 - \frac{k}{t}\right) \end{aligned}$$

Άρα

$$\begin{aligned} Cov(y_t, y_{t-k}) &= E(y_t - E(y_t))(y_{t-k} - E(y_{t-k})) \\ &= E(y_t - p)(y_{t-k} - p) \\ &= E(y_t y_{t-k}) - p^2 \\ &= p\left(1 - \frac{k}{t}\right) - p^2 \\ &= p\left(1 - \frac{k}{t} - p\right) \end{aligned}$$

η οποία εξαρτάται από τον χρόνο. Δηλαδή η y_t δεν είναι ασθενώς στάσιμη.

12. Χρησιμοποιήστε το Excel ή το gretl για να σχεδιάσετε (προσομοιώσετε) το χρονοδιάγραμμα ενός τυχαίου περιπάτου (τυχαία εξέλιξη)

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

και ενός τυχαίου περιπάτου με μετατόπιση (τυχαία εξέλιξη με μετατόπιση)

$$x_t = \alpha + x_{t-1} + e_t$$

με την πληροφόρηση ότι

$$y_0 = 0\,,\ x_0 = 0\,,\ \alpha = 0.01\,,\ u_t \sim N.i.d(0,1)\ ,\ e_t \sim N.i.d(0,1)$$

Υιοθετήστε δείγματα με $T = 50, 100, 250, 1000$ παρατηρήσεις. Τι παρατη-

ρείτε στο γράφημα της χρονοσειράς του τυχαίου περιπάτου με μετατόπιση; (δείτε και την επόμενη άσκηση)

13. Χρησιμοποιήστε το Excel ή το gretl για να προσομοιώσετε τις παρακάτω χρονοσειρές τυχαίας εξέλιξης με μετατόπιση και γραμμικής χρονικής τάσης

$$y_t = 0.02 + y_{t-1} + \frac{1}{3}u_t \ \ , \ \ t = 1, ..., T$$

$$x_t = 1 + 0.02t + u_t \, , \, t = 1, ..., T$$

με την πληροφόρηση

$$y_0 = 0 \, , \, u_t \sim N.i.d \, (0, 1)$$

Υιοθετήστε δείγματα με $T = 50, 100, 250, 1000$ παρατηρήσεις. **Θα διαπιστώσετε ότι οι δύο τύποι μη στασιμότητας είναι δύσκολο να διακριθούν οπτικά**.

14. Έστω ότι

$$\alpha = 15.347 \, , \, \beta = 0.087 \, , \, \gamma = 0.00244 \, , \, \delta = 0.0000083$$

Σχεδιάστε στο Excel ή το gretl τη χρονοσειρά

$$y_t = \alpha + \beta t + \gamma t^2 + \delta t^3 + u_t \ , \ t = 1, ..., T$$

γιά $T = 200$ με $u_t \sim N.i.d \, (0, 1)$. Τι παρατηρείτε καθώς αυξάνετε ή μειώνετε τη διακύμανση των διαταρακτικών όρων;

15. Εκτιμήστε τις παραμέτρους μ και σ^2 με τη μέθοδο εκτίμησης μέγιστης πιθανοφάνειας όταν έχετε ένα τυχαίο δείγμα n παρατηρήσεων από τις τυχαίες μεταβλητές

$$u_i \sim N.i.d \left(\mu, \sigma^2\right)$$

Βρείτε και επιλύστε τις συνθήκες πρώτης τάξης για μεγιστοποίηση. Επιπλέον, ελέγξτε αν οι συνθήκες δεύτερης τάξης ικανοποιούνται.

16. Έστω ότι έχετε στη διάθεσή σας ένα τυχαίο δείγμα x_i , $i = 1, ..., n$ από τις τυχαίες μεταβλητές X_i, $i = 1, ..., n$ που κατανέμονται ομοιογενώς σύμφωνα

με τη συνάρτηση πυκνότητας πιθανότητας

$$f(x) = (1+\theta)\, x^{\theta} \quad , \quad \theta > -1 \quad , \quad 0 < x < 1$$

ενώ $\theta > -1$ είναι μία παράμετρος που χαρακτηρίζει την κατανομή των δεδομένων. Υπολογίστε την αναμενόμενη τιμή $E(X_i) = \mu$ των τυχαίων μεταβλητών και στη συνέχεια βρείτε τον εκτιμητή μέγιστης πιθανοφάνειας $\tilde{\mu}_{ML}$ της αναμενόμενης τιμής μ.

Υπόδειξη: Θα βρείτε ότι η αναμενόμενη τιμή μ είναι συνάρτηση της παραμέτρου θ. Παρουσιάστε και λύστε αναλυτικά μόνο τη συνθήκη πρώτης τάξης ως προς θ. Με αυτό τον τρόπο θα βρείτε τον εκτιμητή $\tilde{\theta}_{ML}$. Κατόπιν, προβείτε σε αντικατάσταση για να βρείτε τον εκτιμητή $\tilde{\mu}_{ML}$.

Απάντηση

Η αναμενόμενη τιμή της X_i για κάθε i δίνεται από

$$E(X_i) = \mu =$$

$$= \int_0^1 x f(x)\, dx$$

$$= \int_0^1 x\,(1+\theta)\, x^{\theta} dx$$

$$= \frac{(1+\theta)}{(2+\theta)} x^{\theta+1} \Big|_0^1$$

$$= \frac{\theta+1}{\theta+2}$$

Επειδή η αναμενόμενη τιμή

$$\mu = \frac{\theta+1}{\theta+2}$$

είναι ένα-προς-ένα συνάρτηση της παραμέτρου θ και επειδή $\theta > -1$ έχουμε ότι

$$\theta = \frac{1-2\mu}{\mu-1}$$

και $0 < \mu < 1$. Η λογαριθμική συνάρτηση πιθανοφάνειας ως προς το θ δίνεται από τη σχέση

$$\begin{aligned}\ln L_n(\theta) = & \\ &= \ln\left(\prod_{i-1}^{n}(1+\theta)\, x_i^{\theta}\right) \\ &= n\ln(1+\theta) + \theta\sum_{i=1}^{n}\ln x_i\end{aligned}$$

η οποία δεν περιέχει την παράμετρο ενδιαφέροντος μ. Περιέχει όμως την παράμετρο θ, η οποία μπορεί να αντικατασταθεί αφού όπως διαπιστώσαμε παραπάνω $\theta = \frac{1-2\mu}{\mu-1}$. Η συνθήκη πρώτης τάξης για μεγιστοποίηση δίνει

$$\frac{d\ln L_n(\theta)}{d\theta} = \frac{n}{(1+\theta)} + \sum_{i=1}^{n}\ln x_i = 0 \Rightarrow$$

$$\tilde{\theta}_{ML} = -\left(\frac{n+\sum_{i=1}^{n}\ln x_i}{\sum_{i=1}^{n}\ln x_i}\right)$$

$$= -\left(\frac{n+n\overline{\ln x}}{n\overline{\ln x}}\right)$$

$$= -\left(\frac{1+\overline{\ln x}}{\overline{\ln x}}\right)$$

Προβαίνουμε λοιπόν στην αντικατάσταση

$$\tilde{\mu}_{ML} = \frac{\tilde{\theta}_{ML}+1}{\tilde{\theta}_{ML}+2}$$

και

$$\tilde{\mu}_{ML} = \frac{1}{1 - \overline{\ln x}}$$

17. Υποθέστε τη συνάρτηση πυκνότητας πιθανότητας

$$f(x) = \frac{1}{\theta} e^{-\frac{x}{\theta}} \quad , \quad x \geq 0 \quad , \quad \theta > 0$$

Υποθέστε ότι παρατηρούμε ένα τυχαίο δείγμα x_i για $i = 1, ..., n$. Βρείτε τον εκτιμητή μέγιστης πιθανοφάνειας $\tilde{\theta}_{ML}$ της παραμέτρου θ και βεβαιώστε ότι οι Σ.Δ.Τ ισχύουν στο στάσιμο σημείο (στην τιμή $\tilde{\theta}_{ML}$ που ικανοποιεί τις Σ.Π.Τ). Στη συνέχεια, παρουσιάστε τη μεθοδολογία ελέγχου της υπόθεσης $H_0 : \theta = 0.5$.

Σημείωση: Ισχύει ότι $E(x_i) = \theta$

Απάντηση

Η συνάρτηση πιθανοφάνειας για το τυχαίο δείγμα $x_1, ..., x_n$ δίνεται από

$$L(\theta) = f(x_1, x_2, ..., x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta) = \prod_{i=1}^{n} \frac{1}{\theta} e^{-\frac{x_i}{\theta}}$$

και η λογαριθμική συνάρτηση πιθανοφάνειας από

$$\ln L(\theta) = \ln \left(\prod_{i=1}^{n} \frac{1}{\theta} e^{-\frac{x_i}{\theta}} \right) = \sum_{i=1}^{n} \ln \left(\frac{1}{\theta} e^{-\frac{x_i}{\theta}} \right)$$

$$= -\sum_{i=1}^{n} \ln \theta - \frac{1}{\theta} \sum_{i=1}^{n} x_i$$

$$= -n \ln \theta - \frac{1}{\theta} \sum_{i=1}^{n} x_i$$

ή

$$\ln L(\theta) = -n \ln \theta - \frac{n\bar{x}}{\theta}$$

Σ.Π.Τ

$$\frac{\partial \ln L}{\partial \theta} = 0 \Rightarrow -\frac{n}{\theta} + \frac{n\bar{x}}{\theta^2} = 0 \Rightarrow \hat{\theta}_{ML} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

δηλαδή

$$\hat{\theta}_{ML} = \bar{x}$$

Σ.Δ.Τ

$$\frac{\partial^2 \ln L}{\partial \theta^2} = \frac{n}{\theta^2} - \frac{2n\theta\bar{x}}{(\theta^2)^2}$$

ενώ στο στάσιμο σημείο έχουμε $\hat{\theta}_{ML} = \bar{x}$ άρα

$$\begin{aligned}\left.\frac{\partial^2 \ln L}{\partial \theta^2}\right|_{\theta=\hat{\theta}_{ML}} &= \frac{n}{\hat{\theta}^2_{ML}} - \frac{2n\left(\hat{\theta}_{ML}\right)\left(\hat{\theta}_{ML}\right)}{\left(\hat{\theta}^2_{ML}\right)^2} \\ &= \frac{n}{\hat{\theta}^2_{ML}} - \frac{2n\hat{\theta}^2_{ML}}{\left(\hat{\theta}^2_{ML}\right)^2} \\ &= -\frac{n}{\hat{\theta}^2_{ML}} < 0\end{aligned}$$

Δηλαδή οι Σ.Δ.Τ ικανοποιούνται και ο εκτιμητής μέγιστης πιθανοφάνειας δίνεται όντως από την $\hat{\theta}_{ML} = \bar{x}$. Η «προσεγγιστική» κατανομή δίνεται από

$$\tilde{\theta}_{ML} \overset{\alpha}{\sim} N\left(\theta\,,\, \hat{\mathbf{I}}_{\theta}^{-1}\right)$$

όπου

$$\mathbf{I}_{\theta} = -E\left(\frac{\partial^2 \ln L}{\partial \theta^2}\right)$$

Επειδή

$$\begin{aligned}\mathbf{I}_{\theta} &= -E\left(\frac{\partial^2 \ln L}{\partial \theta^2}\right) \\ &= -E\left(\frac{n}{\theta^2} - \frac{2n\bar{x}\theta}{\theta^4}\right) \\ &= -\left(\frac{n}{\theta^2} - \frac{2n\theta E\left(\bar{x}\right)}{\theta^4}\right) \\ &= -\left(\frac{n}{\theta^2} - \frac{2n\theta^2}{\theta^4}\right)\end{aligned}$$

$$= \frac{n}{\theta^2}$$

έχουμε

$$\mathbf{I}_\theta^{-1} = \frac{\theta^2}{n}$$

και

$$\hat{\mathbf{I}}_\theta^{-1} = \frac{\tilde{\theta}_{ML}^2}{n}$$

Άρα για να προβούμε στον έλεγχο $H_0 : \theta = 0.5$ υιοθετούμε την προσεγγιστική κανονική κατανομή

$$\hat{\theta}_{ML} \overset{\alpha}{\sim} N\left(\theta\,,\, \frac{\tilde{\theta}_{ML}^2}{n}\right)$$

και τη στατιστική

$$z = \frac{\hat{\theta}_{ML} - 0.5}{\sqrt{\frac{\tilde{\theta}_{ML}^2}{n}}}$$

$$= \frac{\sqrt{n}\left(\hat{\theta}_{ML} - 0.5\right)}{\hat{\theta}_{ML}} \overset{\alpha}{\sim} N\left(0\,,\, 1\right)$$

Αν $|z| > 1.96$ τότε απορρίπτουμε τη μηδενική υπόθεση H_0 σε επίπεδο σημαντικότητας 5%. Παρατηρήστε ότι, εξ'ορισμού, η παράμετρος $\theta > 0$ άρα δεν έχει νόημα να ελέγξουμε την υπόθεση $\theta = 0$.