ΕΙΣΑΓΩΓΗ ΣΤΗΝ ΟΙΚΟΝΟΜΕΤΡΙΑ

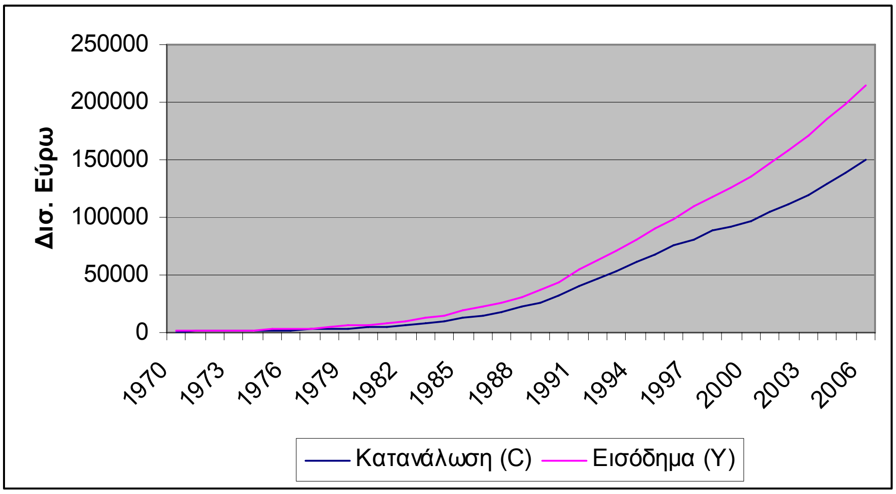
Το γραμμικό οικονομετρικό υπόδειγμα: Υποθέσεις, εκτίμηση, στατιστικές ιδιότητες των εκτιμητών του και έλεγχος στατιστικών υποθέσεων

Α. Υποθέσεις

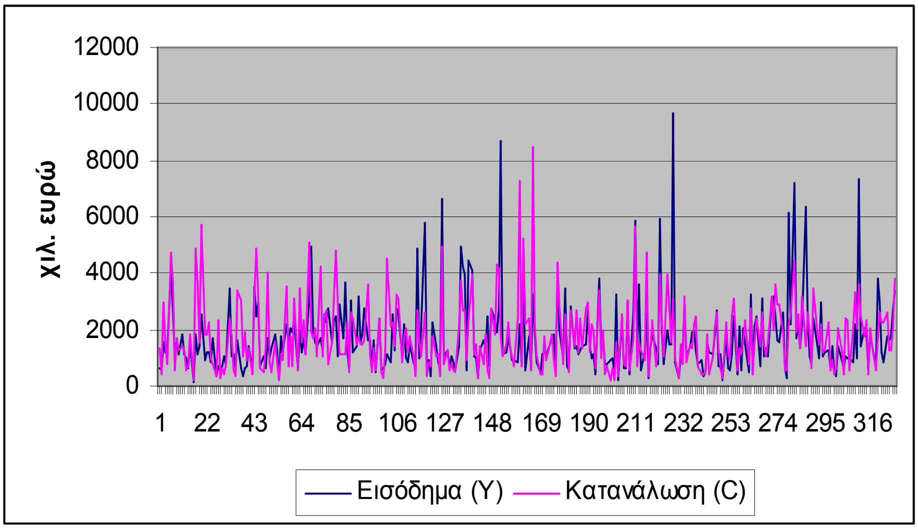
Σκοπός της Οικονομετρίας: Να εκτιμήσει τις παραμέτρους ενός οικονομικού υποδείγματος και να ελέγξει την ορθότητα και την επάρκειά του εμπειρικά, βάσει ενός δείγματος παρατηρήσεων.

Στοιχεία: Για την εκτίμηση ή τον έλεγχο των υποδειγμάτων χρησιμοποιούμε είτε διαστρωματικά στοιχεία (που συμβολίζονται ως $i = 1,2,\ldots,N$) ή χρονολογικές παρατηρήσεις (που συμβολίζονται ως $t = 1,2,\ldots,T$)

ΣΧΗΜΑ 1.1: Συνολική κατανάλωση και εισόδημα - χρονολογικά στοιχεία



ΣΧΗΜΑ 1.2: Κατανάλωση και εισόδημα νοικοκυριών – Διαστρωματικά Στοιχεία (Νοικοκυριά)



Πηγή: ΕΣΥΕ, Οικογενειακοί προϋπολογισμοί έτους 2004

Οικονομικό Υπόδειγμα (παραδείγματα):

C_i = a + bY_i (συνάρτηση κατανάλωσης-εισοδήματος νοικοκυριών-)

$Q_i = AK_i^{a}L_i^{b}$ (συνάρτηση παραγωγής Cobb-Douglas- επιχειρήσεων)

Οικονομετρικό Υπόδειγμα (ή Γραμμική Παλινδρόμηση κατά τη στατιστική θεωρία): Το υπόδειγμα αυτό προσθέτει ένα στοχαστικό όρο (τυχαία μεταβλητή) ε, που αναφέρεται ως διαταρακτικός όρος, στο οικονομικό υπόδειγμα που γίνεται

$C_i = a + bY_i + \varepsilon_i$, για κάθε οικονομική μονάδα (πχ νοικοκυριό) *i*=1,2,…,N

ή στην πιο γενική του μορφή

$y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$,

όπου

y_i αποτελεί την ερμηνευμένη μεταβλητή του οικονομικού ή οικονομετρικού υποδείγματος, που αναφέρεται ως εξαρτημένη,

x_i αποτελεί την ερμηνευτική μεταβλητή, που αναφέρεται ως ανεξάρτητη,

$\beta_1 + \beta_2 x_i$ αποτελεί το ερμηνευμένο μέρος από την οικονομική θεωρία της οικονομετρικού υποδείγματος (παλινδρόμησης) και

ε_i είναι ο διαταρακτικός όρος, που αποτελεί το ανερμήνευτο μέρος του οικ/μετρικού υποδ.

Ο διαταρακτικός όρος ε_i (ή σφάλμα όπως αναφέρεται διαφορετικά) αποτελεί μια τυχαία μεταβλητή που αντιπροσωπεύει τυχαίες (μη συστηματικές) αποκλίσεις των προβλέψεων του οικονομικού υποδείγματος για την παρατηρούμενη τιμή (πχ κατανάλωση) της μονάδας *i*, y_i (C_i).

Ιδιότητες του διαταρακτικού όρου ε_i (αναφέρονται και ως κλασσικές υποθέσεις)

- $E(\varepsilon_i) = 0$, $\forall i$ (μέσο σφάλμα μηδέν, διαφορετικά δεν θα ήταν τυχαία η απόκλιση)

- $Var(\varepsilon_i) = \sigma^2 = E(\varepsilon_i^2) = \sigma^2$, $\forall$ i (ομοσκεδαστικότητα, διαφορετικά θα υπήρχε τερογένεια –διαφορετική διακύμανση του διαταρακτικού όρου ανά διαφορετική μονάδα *i*)

- $Cov(\varepsilon_i, \varepsilon_j) = E(\varepsilon_i \varepsilon_j) - E(\varepsilon_i)E(\varepsilon_j) = 0$, $\forall\ i \neq j$ (μη συστηματικές, αλλά ανεξάρτητες τυχαίες αποκλίσεις μεταξύ των διαφορετικών μονάδων *i*)

Ιδιότητες της ανεξάρτητης μεταβλητής x_i

- Οι τιμές της θεωρούνται ως προκαθορισμένες (predetermined) (ή σταθερές σε επαναλαμβανόμενα δείγματα παρατηρήσεων-δηλ. μη τυχαίες μεταβλητές). Μια καλύτερη από στατιστικής άποψης υπόθεση είναι να τις θεωρήσουμε ως δεδομένες. Η υπόθεση αυτή σημαίνει ότι

$$E(x_i) = x_i \quad \text{και}$$

$$Cov(x_i, \varepsilon_i) = E(x_i \varepsilon_i) - E(x_i)E(\varepsilon_i) = E(x_i \varepsilon_i) = x_i E(\varepsilon_i) = 0,$$

Ιδιότητες της εξαρτημένης μεταβλητής y_i (ερμηνεία του οικον. Υποδ.)

- $E(y_i) = E(\beta_1 + \beta_2 x_i + \varepsilon_i) = \beta_1 + \beta_2 x_i$, καθώς $E(\varepsilon_i) = 0$ και $E(x_i) = x_i$, $\forall$ i.

που σημαίνει ότι

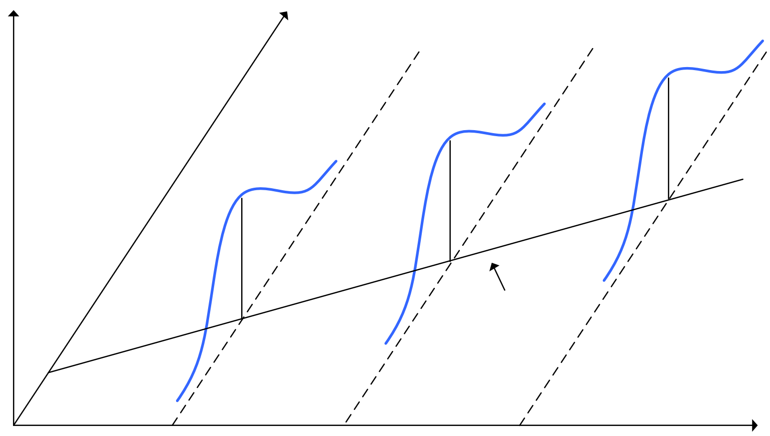
$$y_i = E(y_i) + \varepsilon_i, \text{ όπου } E(y_i) = \beta_1 + \beta_2 x_i \equiv \text{οικον. Υποδ.}$$

Δηλ.

Το Οικον. Υποδ. αποτελεί τη μέση τιμή (πρόβλεψη) του οικονομετρικού για το y_i

Διαγραμματικά

ΣΧΗΜΑ 2.1: Κατανομές των τιμών της εξαρτημένης μεταβλητής y_i γύρω από τη μέση τους τιμή $E(y_i) = \beta_1 + \beta_2 x_i$ - πρόβλεψη του οικον. Υποδ.



- $Var(y_i) = Var(\beta_1 + \beta_2 x_i + \varepsilon_i) = Var(\beta_1 + \beta_2 x_i) + Var(\varepsilon_i) + 2Cov(\beta_1 + \beta_2 x_i, \varepsilon_i)$

$$= Var(\beta_1) + Var(\beta_2 x_i) + \sigma^2 + 2Cov(\beta_1; \beta_2 x_i) + 2Cov(\beta_1; \varepsilon_i) + 2Cov(\beta_2 x_i; \varepsilon_i)$$

$$= 0 + 0 + \sigma^2 + 0 + 0 + 0 = \sigma^2, \quad \forall\, i.$$

(καθώς όλες οι τυχαίες αποκλίσεις της εξαρτημένης μεταβλητής y_i γύρω από τη μέση της τιμή $E(y_i) = \beta_1 + \beta_2 x_i$ προέρχονται από το διαταρακτικό όρο ε_i, για κάθε διαφορετική μονάδα)

- $Cov(y_i, y_j) = Cov(\beta_1 + \beta_2 x_i + \varepsilon_i,\ \beta_1 + \beta_2 x_j + \varepsilon_j) = \ldots = Cov(\varepsilon_i, \varepsilon_j) = 0$,

(καθώς $Cov(x_i ; x_j) = 0$ και $Cov(x_i ; \varepsilon_j) = 0$).

Επιπλέον (όχι απαραίτητες υποθέσεις για μεγάλα δείγματα):

- Η κατανομή του διαταρακτικού όρου ε_i είναι κανονική

$$\varepsilon_i \sim N\left(E(\varepsilon_i)=0, Var(\varepsilon_i)=\sigma^2\right),\ \forall\, i.,$$

που συνεπάγεται ότι

$$y_i \sim N\left(E(y_i)=\beta_1 + \beta_2 x_i,\ Var(y_i)\sigma^2\right)$$

B. Εκτίμηση των συντελεστών του οικονομετρικού υποδείγματος[1]

Η μέθοδος των ελαχίστων τετραγώνων-Least Squares (LS):

Έστω $\hat{\beta}_1$ και $\hat{\beta}_2$ ότι είναι *οι εκτιμητές (εκτιμήσεις)* της μεθόδου LS. Τότε,

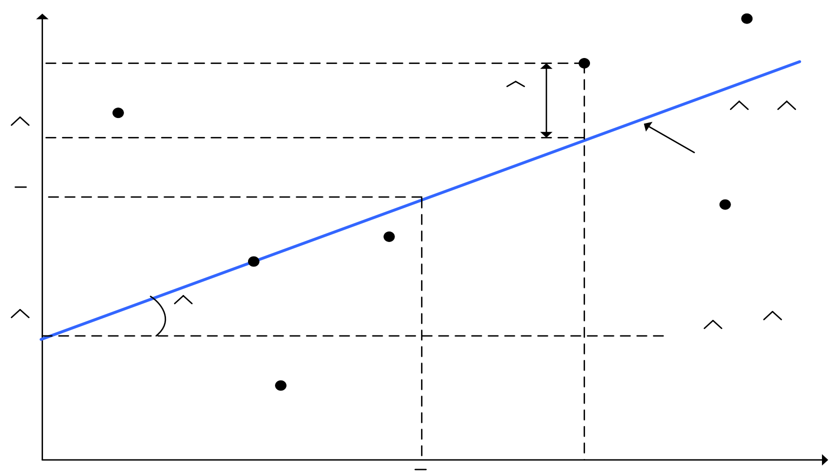
$$\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i, \ \forall\, i,$$

αποτελούν τις εκτιμήσεις (εκτιμημένες τιμές) για τις τιμές της εξαρτημένης μεταβλητής y_i της μεθόδου LS, και

$$\hat{\varepsilon}_i = y_i - \hat{y}_i$$

αποτελούν τις αποστάσεις ανάμεσα στις εκτιμημένες και παρατηρούμενες τιμές της y_i Οι αποστάσεις αυτές αναφέρονται ως *κατάλοιπα (residuals)*. Βλέπε ακόλουθο διάγραμμα:

ΣΧΗΜΑ 2.3: Σχέση ανάμεσα στις παρατηρούμενες (y_i), τις εκτιμημένες ($\hat{y}_i$) τιμές της εξαρτημένης μεταβλητής και τα κατάλοιπα $\hat{\varepsilon}_i = y_i - \hat{y}_i$



---

[1] Το υπόδειγμα αναφέρεται ως γραμμικό όταν οι πρώτοι παράγωγοι της y_i ως προς τις παραμέτρους β_1 και β_2 δίνουν τις ακόλουθες σχέσεις:

$$\frac{\partial y_i}{\partial \beta_1} = 1 \quad \text{και} \quad \frac{\partial y_i}{\partial \beta_2} = x_i,$$

οι οποίες είναι ανεξάρτητες των β_1 και β_2.

Η μέθοδος LS βρίσκει εκτιμητές (εκτιμήσεις) για τους συντελεστές β_1 και β_2 που ελαχιστοποιούν την ακόλουθη συνάρτηση του αθροίσματος των τετραγώνων των καταλοίπων (residual sum of squares -RSS), που ορίζεται ως εξής:

$$\mathrm{RSS}(\hat{\beta}_1,\hat{\beta}_2) \equiv \sum_{i=1}^{N} \hat{\varepsilon}_i^2 .$$

Μαθηματικά, το πρόβλημα της μεθόδου LS λύνεται ως ακολούθως:

$$\underset{\hat{\beta}_1,\hat{\beta}_2}{\varepsilon\lambda\alpha\chi}\ \mathrm{RSS}(\hat{\beta}_1,\hat{\beta}_2) \equiv \underset{\hat{\beta}_1,\hat{\beta}_2}{\varepsilon\lambda\alpha\chi} \sum_{i=1}^{N} \hat{\varepsilon}_i^2 \equiv \underset{\hat{\beta}_1,\hat{\beta}_2}{\varepsilon\lambda\alpha\chi} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

$$= \underset{\hat{\beta}_1,\hat{\beta}_2}{\varepsilon\lambda\alpha\chi} \sum_{i=1}^{N} \left[y_i - (\hat{\beta}_1 + \hat{\beta}_2 x_i) \right]^2$$

Συνθήκες πρώτης τάξης (first order conditions -FOC):

$$\frac{\partial}{\partial \hat{\beta}_1} \sum_{i=1}^{N} \hat{\varepsilon}_i^2 = 0 \qquad \text{και} \qquad \frac{\partial}{\partial \hat{\beta}_2} \sum_{i=1}^{N} \hat{\varepsilon}_i^2 = 0$$

Συνθήκες δεύτερης τάξης (second order conditions-SOC):

$$\frac{\partial^2}{\partial \hat{\beta}_1^2} \sum_{i=1}^{N} \hat{\varepsilon}_i^2 > 0 \qquad \text{και} \qquad \frac{\partial^2}{\partial \hat{\beta}_2^2} \sum_{i=1}^{N} \hat{\varepsilon}_i^2 > 0$$

Από τις συνθήκες πρώτης τάξης προκύπτουν οι εκτιμητές ελαχίστων τετραγώνων $\hat{\beta}_1$ και $\hat{\beta}_2$ ως

$$\frac{\partial}{\partial \hat{\beta}_1} \sum_{i=1}^{N} \hat{\varepsilon}_i^2 = \frac{\partial}{\partial \hat{\beta}_1} \sum_{i=1}^{N} (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i)^2 = 0 = \sum_{i=1}^{N} \frac{\partial}{\partial \hat{\beta}_1} (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i)^2 = 0$$

$$= \sum_{i=1}^{N} 2(y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i)\ \frac{\partial}{\partial \hat{\beta}_1} (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i) = 0$$

$$= 2 \sum_{i=1}^{N} (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i)(-1)\ = 0$$

$$\Rightarrow \qquad \sum_{i=1}^{N}(y_i-\hat{\beta}_1-\hat{\beta}_2 x_i)=0 \quad ή \quad \sum_{i=1}^{N} y_i \;=\; N\hat{\beta}_1+\hat{\beta}_2\sum_{i=1}^{N} x_i \qquad (4α)$$

και

$$\frac{\partial}{\partial\hat{\beta}_2}\sum_{i=1}^{N}\hat{\varepsilon}_i^2=\frac{\partial}{\partial\hat{\beta}_2}\sum_{i=1}^{N}(y_i-\hat{\beta}_1-\hat{\beta}_2 x_i)^2=0=\sum_{i=1}^{N}\frac{\partial}{\partial\hat{\beta}_2}(y_i-\hat{\beta}_1-\hat{\beta}_2 x_i)^2=0$$

$$=\sum_{i=1}^{N}2(y_i-\hat{\beta}_1-\hat{\beta}_2 x_i)\frac{\partial}{\partial\hat{\beta}_2}(y_i-\hat{\beta}_1-\hat{\beta}_2 x_i)=0$$

$$=\sum_{i=1}^{N}2(y_i-\hat{\beta}_1-\hat{\beta}_2 x_i)(-x_i)=0$$

$$\Rightarrow \quad \sum_{i=1}^{N}(y_i x_i-\hat{\beta}_1 x_i-\hat{\beta}_2 x_i{}^2)=0 \quad ή \quad \sum_{i=1}^{N} y_i x_i \;=\; \hat{\beta}_1\sum_{i=1}^{N} x_i+\hat{\beta}_2\sum_{i=1}^{N} x_i^2 \qquad (4β)$$

Το σύστημα αυτό των δύο εξισώσεων (4α)-(4β), που συνοψίζεται ως

$$\left.\begin{array}{ll} (4α): & \sum_{i=1}^{N} y_i = N\hat{\beta}_1+\hat{\beta}_2\sum_{i=1}^{N} x_i \\ (4β): & \sum_{i=1}^{N} y_i x_i = \hat{\beta}_1\sum_{i=1}^{N} x_i+\hat{\beta}_2\sum_{i=1}^{N} x_i^2 \end{array}\right\},$$

αναφέρεται στη βιβλιογραφία ως σύστημα κανονικών εξισώσεων. Λύνοντάς το με τη μέθοδο των οριζουσών, δίνει τους ακόλουθους γενικούς τύπους των εκτιμητών των ελαχίστων τετραγώνων:

$$\hat{\beta}_1 \;=\; \frac{\begin{vmatrix} \sum_{i=1}^{N} y_i & \sum_{i=1}^{N} x_i \\ \sum_{i=1}^{N} y_i x_i & \sum_{i=1}^{N} x_i^2 \end{vmatrix}}{\begin{vmatrix} N & \sum_{i=1}^{N} x_i \\ \sum_{i=1}^{N} x_i & \sum_{i=1}^{N} x_i^2 \end{vmatrix}} \;=\; \frac{\sum_{i=1}^{N} y_i\sum_{i=1}^{N} x_i^2-\sum_{i=1}^{N} x_i\sum_{i=1}^{N} y_i x_i}{N\sum_{i=1}^{N} x_i^2-\left(\sum_{i=1}^{N} x_i\right)^2} \qquad (5α)$$

και

$$\hat{\beta}_2 = \frac{\begin{vmatrix} N & \sum_{i=1}^{N} y_i \\ \sum_{i=1}^{N} x_i & \sum_{i=1}^{N} y_i x_i \end{vmatrix}}{\begin{vmatrix} N & \sum_{i=1}^{N} x_i \\ \sum_{i=1}^{N} x_i & \sum_{i=1}^{N} x_i^2 \end{vmatrix}} = \frac{N\sum_{i=1}^{N} y_i x_i - \sum_{i=1}^{N} x_i \sum_{i=1}^{N} y_i}{N\sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i\right)^2} \qquad (5\beta)$$

Οι δύο αυτοί τύποι των LS εκτιμητών $\hat{\beta}_1$ και $\hat{\beta}_2$ αποτελούν συναρτήσεις των παρατηρήσεων των μεταβλητών x_i και y_i του δείγματος, και έτσι αποτελούν <u>τυχαίες μεταβλητές</u>. Για να διαπιστώσουμε αν πράγματι οι παραπάνω λύσεις των LS εκτιμητών $\hat{\beta}_1$ και $\hat{\beta}_2$ δίδουν ελάχιστο για τη συνάρτηση $RSS(\hat{\beta}_1, \hat{\beta}_2)$, στη συνέχεια υπολογίζουμε τις δεύτερες παραγώγους της συνάρτησης αυτής:

$$\frac{\partial^2}{\partial \hat{\beta}_1^{\,2}} \sum_{i=1}^{N} \hat{\varepsilon}_i^2 = \frac{\partial}{\partial \hat{\beta}_1} \sum_{i=1}^{N} 2(y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i)(-1) = 2\sum_{i=1}^{N} 1 = 2N > 0$$

και

$$\frac{\partial^2}{\partial \hat{\beta}_2^{\,2}} \sum_{i=1}^{N} \hat{\varepsilon}_i^2 = \frac{\partial}{\partial \hat{\beta}_2} \sum_{i=1}^{N} 2(y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i)(-x_i) = 2\sum_{i=1}^{N} x_i^2 > 0$$

που έχουν θετικές τιμές και έτσι, επιβεβαιώνει ότι οι λύσεις των LS εκτιμητών πράγματι αποτελούν λύσεις στο ελάχιστο σημείο της συνάρτησης $RSS(\hat{\beta}_1, \hat{\beta}_2)$.

Ο εκτιμητής ελαχίστων τετραγώνων σε αποκλίσεις από τους μέσους $\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$ **και** $\overline{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$

Οι λύσεις αυτές βρίσκονται ως ακολούθως: Από την κανονική εξίσωση (4α) συνεπάγεται ότι ο εκτιμητής $\hat{\beta}_1$ γράφεται ως ακολούθως:

$$\sum_{i=1}^{N} y_i = N\hat{\beta}_1 + \hat{\beta}_2 \sum_{i=1}^{N} x_i \quad \Rightarrow \hat{\beta}_1 = \frac{1}{N}\sum_{i=1}^{N} y_i - \hat{\beta}_2 \frac{1}{N}\sum_{i=1}^{N} x_i \quad \Rightarrow$$

$$\hat{\beta}_1 = \overline{y} - \hat{\beta}_2 \overline{x}$$

Αντικαθιστώντας τη λύση αυτή του εκτιμητή $\hat{\beta}_1$ στην κανονική εξίσωση (4β) δίδει

$$\sum_{i=1}^{N} y_i x_i = (\overline{y} - \hat{\beta}_2 \overline{x}) \sum_{i=1}^{N} x_i + \hat{\beta}_2 \sum_{i=1}^{N} x_i^2$$

Διαιρώντας με Ν και τα δύο μέλη της τελευταίας σχέσης και κάνοντας πράξεις θα πάρουμε τον εκτιμητή $\hat{\beta}_2$ σε αποκλίσεις από τους μέσους ως εξής:

$$\frac{1}{N}\sum_{i=1}^{N} y_i x_i = (\overline{y} - \hat{\beta}_2 \overline{x}) \frac{1}{N}\sum_{i=1}^{N} x_i + \hat{\beta}_2 \frac{1}{N}\sum_{i=1}^{N} x_i^2 \quad \Rightarrow$$

$$\frac{1}{N}\sum_{i=1}^{N} y_i x_i - \overline{y}\frac{1}{N}\sum_{i=1}^{N} x_i = \hat{\beta}_2 \left(\frac{1}{N}\sum_{i=1}^{N} x_i^2 - \overline{x}\frac{1}{N}\sum_{i=1}^{N} x_i \right) \quad \Rightarrow$$

$$\frac{1}{N}\sum_{i=1}^{N} y_i x_i - \overline{y}\,\overline{x} = \hat{\beta}_2 \left(\frac{1}{N}\sum_{i=1}^{N} x_i^2 - \overline{x}^2 \right) \quad \Rightarrow$$

$$\frac{1}{N}\sum_{i=1}^{N} (y_i x_i - \overline{y}\,\overline{x}) = \hat{\beta}_2 \frac{1}{N}\sum_{i=1}^{N} (x_i^2 - \overline{x}^2) \quad \Rightarrow$$

$$\hat{\beta}_2 = \frac{\frac{1}{N}\sum_{i=1}^{N} (y_i x_i - \overline{y}\,\overline{x})}{\frac{1}{N}\sum_{i=1}^{N} (x_i^2 - \overline{x}^2)} = \frac{\frac{1}{N}\sum_{i=1}^{N} (y_i - \overline{y})(x_i - \overline{x})}{\frac{1}{N}\sum_{i=1}^{N} (x_i - \overline{x})^2} = \frac{\hat{\sigma}_{yx}}{\hat{\sigma}_x^2} = \hat{\beta}_2 = \frac{\hat{\sigma}_{yx}}{\hat{\sigma}_x^2} = \frac{\hat{r}_{yx}\hat{\sigma}_x\hat{\sigma}_y}{\hat{\sigma}_x^2} = \hat{r}_{yx}\frac{\hat{\sigma}_y}{\hat{\sigma}_x}$$

όπου

$$\hat{\sigma}_{yx} = \frac{1}{N}\sum_{i=1}^{N} (y_i - \overline{y})(x_i - \overline{x}) \text{ και } \hat{\sigma}_x^2 = \frac{1}{N}\sum_{i=1}^{N} (x_i - \overline{x})^2 \text{, και } \hat{r}_{yx} = \frac{\hat{\sigma}_{yx}}{\hat{\sigma}_x\hat{\sigma}_y}.$$

Ιδιότητες των καταλοίπων $\hat{\varepsilon}_i$

(i) Το άθροισμα των καταλοίπων είναι μηδέν, δηλαδή ισχύει

$$\sum_{i=1}^{N}\hat{\varepsilon}_i = 0, \qquad \text{και άρα} \quad \bar{\hat{\varepsilon}} = \frac{1}{N}\sum_{i=1}^{N}\hat{\varepsilon}_i = 0$$

επειδή έχουμε $\sum_{i=1}^{N}(y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i) = \sum_{i=1}^{N}\hat{\varepsilon}_i = 0$ (βλέπε κανονική εξίσωση 4α).

(ii) Οι τιμές της ανεξάρτητης μεταβλητής και των καταλοίπων είναι ορθογώνιες (μη συσχετιζόμενες) μεταξύ τους, δηλ.

$$\sum_{i=1}^{N} x_i\hat{\varepsilon}_i \qquad \text{(βλέπε κανονική εξίσωση 4β).}$$

Ιδιότητες των εκτιμημένων τιμών της εξαρτημένης μεταβλητής, $\hat{y}_i$ (αποτελούν συνέπεια των ιδιοτήτων των καταλοίπων)

(i) Η εκτιμημένη γραμμική σχέση $\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i$ περνάει από το σημείο των μέσων όρων των τιμών της εξαρτημένης και ανεξάρτητης μεταβλητής στο δείγμα. Το σημείο αυτό ορίζεται ως εξής:

$$\left(\bar{x} = \frac{1}{N}\sum_{i=1}^{N}x_i,\ \bar{y} = \frac{1}{N}\sum_{i=1}^{N}y_i\right).$$

Αποδ. Γράψτε το γραμμικό υπόδειγμα ως εξής:

$$y_i = \hat{y}_i + \hat{\varepsilon}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i + \hat{\varepsilon}_i,$$

Παίρνοντας τους μέσους όρους και των δύο μελών της παραπάνω εξίσωσης έχουμε

$$\frac{1}{N}\sum_{i=1}^{N} y_i = \frac{1}{N}\sum_{i=1}^{N}\hat{y}_i + \frac{1}{N}\sum_{i=1}^{N}\hat{\varepsilon}_i = \frac{1}{N}N\hat{\beta}_1 + \hat{\beta}_2\frac{1}{N}\sum_{i=1}^{N}x_i + \frac{1}{N}\sum_{i=1}^{N}\hat{\varepsilon}_i \quad \Rightarrow$$

$$\overline{y} = \overline{\hat{y}} = \hat{\beta}_1 + \hat{\beta}_2\overline{x},$$

λόγω της ιδιότητας των καταλοίπων: (i) $\sum_{i=1}^{N}\hat{\varepsilon}_i = 0$.

(ii) Οι τιμές $\hat{y}_i$ είναι ορθογώνιες με τα κατάλοιπα $\hat{\varepsilon}_i$, δηλαδή ισχύει

$$\sum_{i=1}^{N}\hat{y}_i\hat{\varepsilon}_i = 0,$$

Αποδ. Πολλαπλασιάστε και τα δύο μέλη σχέσης $\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i$ με $\hat{\varepsilon}_i$ και αθροίστε για όλες τις παρατηρήσεις i. Τότε, θα έχουμε

$$\sum_{i=1}^{N}\hat{y}_i\hat{\varepsilon}_i = \sum_{i=1}^{N}(\hat{\beta}_1 + \hat{\beta}_2 x_i)\hat{\varepsilon}_i = \hat{\beta}_1\sum_{i=1}^{N}\hat{\varepsilon}_i + \hat{\beta}_2\sum_{i=1}^{N}x_i\hat{\varepsilon}_i = 0,$$

λόγω των ιδιοτήτων των καταλοίπων: (i) $\sum_{i=1}^{N}\hat{\varepsilon}_i = 0$ και (ii): $\sum_{i=1}^{N}x_i\hat{\varepsilon}_i$.

Γ. Στατιστικές ιδιότητες του εκτιμητή ελαχίστων τετραγώνων

Οι εκτιμητές ελαχίστων τετραγώνων $\hat{\beta}_1$ και $\hat{\beta}_2$ αποτελούν τυχαίες μεταβλητές των οποίων οι τιμές αλλάζουν από δείγμα σε δείγμα. Οι κατανομές αυτών αναφέρονται ως κατανομές δειγματοληψίας και η γνώση τους είναι απαραίτητη προϋπόθεση για τη διεξαγωγή ελέγχων στατιστικών υποθέσεων για τους συντελεστές του υποδείγματος. Στη συνέχεια παρουσιάζουμε τις στατιστικές ιδιότητές τους

<u>Οι στατιστικές ιδιότητες των LS εκτιμητών</u> $\hat{\beta}_1$ και $\hat{\beta}_2$ έχουν ως εξής:

1. Γραμμικότητα (Linearity): Ο εκτιμητής LS $\hat{\beta}_2$ αποτελεί γραμμική συνάρτηση των τιμών της εξαρτημένης μεταβλητής του υποδείγματος y_i. για όλα τα δείγματα

Αποδ. Γράψτε τον εκτιμητή $\hat{\beta}_2$ σε αποκλίσεις από τους μέσους ως εξής:

$$\hat{\beta}_2 = \frac{\sum_{i=1}^{N}(y_i-\overline{y})(x_i-\overline{x})}{\sum_{i=1}^{N}(x_i-\overline{x})^2} = \frac{\sum_{i=1}^{N}(x_i-\overline{x})y_i - \overline{y}\sum_{i=1}^{N}(x_i-\overline{x})}{\sum_{i=1}^{N}(x_i-\overline{x})^2}$$

$$= \frac{\sum_{i=1}^{N}(x_i-\overline{x})y_i}{\sum_{i=1}^{N}(x_i-\overline{x})^2}, \qquad \text{καθώς } \sum_{i=1}^{N}(x_i-\overline{x}) = 0,$$

$$= \sum_{i=1}^{N}\left[\frac{(x_i-\overline{x})}{\sum_{i=1}^{N}(x_i-\overline{x})^2}\right]y_i = \sum_{i=1}^{N}c_i y_i\,, \tag{7}$$

όπου $c_i = \dfrac{(x_i-\overline{x})}{\sum_{i=1}^{N}(x_i-\overline{x})^2}$ θεωρούνται ως σταθερές τιμές. Η τελευταία σχέση δείχνει ότι $\hat{\beta}_2$ αποτελεί γραμμικό συνδυασμό των παρατηρήσεων της εξαρτημένης μεταβλητής y_i.

2. Αμεροληψία (unbiasedness): Η μέση τιμή της κατανομής δειγματοληψίας του εκτιμητή $\hat{\beta}_2$ ισούται με την αληθινή (θεωρητική) τιμή αυτού στον πληθυσμό β_2, δηλαδή ισχύει:

$$E(\hat{\beta}_2) = \beta_2$$

Αποδ: Πρώτα, γράψτε τον εκτιμητή $\hat{\beta}_2$ ως εξής:

$$\hat{\beta}_2 = \sum_{i=1}^{N}c_i y_i \qquad \text{(βλέπε σχέση (7))}$$

$$= \sum_{i=1}^{N}c_i\left(\beta_1+\beta_2 x_i+\varepsilon_i\right), \qquad \text{αντικαθιστώντας } y_i=\beta_1+\beta_2 x_i+\varepsilon_i,$$

$$= \beta_1 \sum_{i=1}^{N} c_i + \beta_2 \sum_{i=1}^{N} c_i x_i + \sum_{i=1}^{N} c_i \varepsilon_i \; = \; \beta_2 + \sum_{i=1}^{N} c_i \varepsilon_i \,, \tag{8}$$

καθώς ισχύουν τα ακόλουθα αποτελέσματα:

$$\sum_{i=1}^{N} c_i = \sum_{i=1}^{N} \left[\frac{(x_i - \overline{x})}{\sum_{i=1}^{N} (x_i - \overline{x})^2} \right] = \frac{1}{\sum_{i=1}^{N} (x_i - \overline{x})^2} \sum_{i=1}^{N} (x_i - \overline{x}) = 0$$

και

$$\sum_{i=1}^{N} c_i x_i = \sum_{i=1}^{N} \left[\frac{(x_i - \overline{x}) x_i}{\sum_{i=1}^{N} (x_i - \overline{x})^2} \right] = \frac{\sum_{i=1}^{N} (x_i - \overline{x}) x_i}{\sum_{i=1}^{N} (x_i - \overline{x}) x_i} = 1 \,.$$

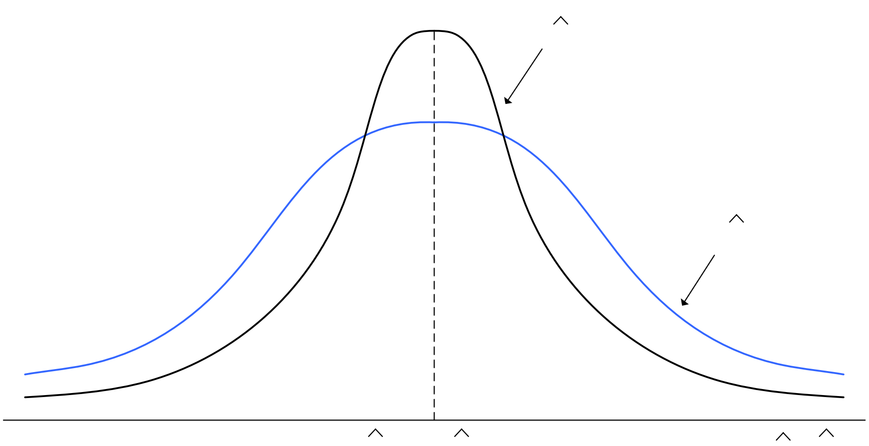
Παίρνοντας την αναμενόμενη τιμή της σχέσης (8) έχουμε την ακόλουθη σχέση:

$$E(\hat{\beta}_2) = \beta_2 + \sum_{i=1}^{N} c_i E(\varepsilon_i) \; = \; \beta_2 \,, \qquad \text{καθώς } E(\varepsilon_i) = 0 \,,$$

που αποδεικνύει ότι ο εκτιμητής $\hat{\beta}_2$ είναι αμερόληπτος.

3. Αποτελεσματικότητα (efficiency) : Ένας εκτιμητής λέγεται αποτελεσματικός, αν αυτός έχει τη μικρότερη διακύμανση από κάποιους άλλους. Βλέπε παρακάτω διάγραμμα:

ΣΧΗΜΑ 2.3: Αποτελεσματικότητα του LS εκτιμητή $\hat{\beta}_2$



Gauss-Markov θεώρημα: Ο LS εκτιμητής $\hat{\beta}_2$ έχει τη μικρότερη διακύμανση από όλους τους γραμμικούς και αμερόληπτους εκτιμητές του συντελεστή β_2, πχ τον $\hat{\beta}_2^*$. Δηλαδή, έχουμε

$$\mathrm{Var}(\hat{\beta}_2^*) > \mathrm{Var}(\hat{\beta}_2).$$

Αποδ: Πρώτα, βρείτε τη διακύμανση του LS εκτιμητή $\hat{\beta}_2$ ως ακολούθως:

$$\mathrm{Var}(\hat{\beta}_2) = \sum_{i=1}^{N} c_i^2 \mathrm{Var}(y_i), \qquad \text{καθώς } y_i \text{ είναι μη συσχετιζόμενες μεταβλητές}$$

$$= \sum_{i=1}^{N} c_i^2 \sigma^2, \qquad \text{καθώς } \mathrm{Var}(y_i) = \sigma^2, \forall i$$

$$= \sigma^2 \sum_{i=1}^{N} c_i^2 = \sigma^2 \sum_{i=1}^{N} \left[\frac{(x_i - \bar{x})}{\sum_{i=}^{N} (x_i - \bar{x})^2} \right]^2 = \frac{\sigma^2}{\sum_{i=}^{N} (x_i - \bar{x})^2} \tag{9}$$

Τώρα, υποθέστε τώρα ότι υπάρχει κάποιος άλλος γραμμικός και αμερόληπτος εκτιμητής του β_2, που συμβολίζεται ως $\hat{\beta}_2^*$ και ορίζεται ως

$$\hat{\beta}_2^* = \sum_{i=1}^{N} d_i y_i, \qquad \text{όπου } d_i = \alpha_i + c_i$$

Λόγω αμεροληψίας του εκτιμητή $\hat{\beta}_2^*$ θα πρέπει να έχουμε:

$$E(\hat{\beta}_2^*) = E\left[\sum_{i=1}^{N} (\alpha_i + c_i)(\beta_1 + \beta_2 x_i + \varepsilon_i) \right]$$

$$= \sum_{i=1}^{N} (\alpha_i + c_i)\beta_1 + \sum_{i=1}^{N} (\alpha_i + c_i)\beta_2 x_i + \sum_{i=1}^{N} (\alpha_i + c_i) E(\varepsilon_i)$$

$$= \beta_1 \sum_{i=1}^{N} \alpha_i + \beta_2 \sum_{i=1}^{N} \alpha_i x_i + \beta_2, \text{ καθώς } \sum_{i=1}^{N} c_i = 0, \ \sum_{i=1}^{N} c_i x_i = 1 \ \ E(\varepsilon_i) = 0.$$

Η τελευταία σχέση δείχνει ότι, για να είναι αμερόληπτός ο εκτιμητής $\hat{\beta}_2^*$, θα πρέπει να ισχύουν οι ακόλουθες δύο συνθήκες:

$$\sum_{i=1}^{N}\alpha_i = 0 \qquad \text{και} \qquad \sum_{i=1}^{N}\alpha_i x_i = 0\,, \tag{10}$$

καθώς μόνο τότε θα ισχύει $E(\hat{\beta}_2^*) = \beta_2$.

Με βάση τις συνθήκες της σχέσης (10), η διακύμανση του εκτιμητή $\hat{\beta}_2^*$ βρίσκεται ως εξής: Πρώτα, παρατηρήστε ότι ο εκτιμητής $\hat{\beta}_2^*$ γράφεται:

$$\hat{\beta}_2^* = \sum_{i=1}^{N} d_i y_i = \sum_{i=1}^{N}(\alpha_i + c_i)(\beta_1 + \beta_2 x_i + \varepsilon_i) = \beta_2 + \sum_{i=1}^{N}(\alpha_i + c_i)\varepsilon_i$$

Παίρνοντας τη διακύμανση της τελευταίας σχέσης έχουμε:

$$\begin{aligned}
\operatorname{Var}(\hat{\beta}_2^*) &= \operatorname{Var}\left[\beta_2 + \sum_{i=1}^{N}(\alpha_i + c_i)\varepsilon_i\right] = \sum_{i=1}^{N}(\alpha_i + c_i)^2 \operatorname{Var}(\varepsilon_i) \\
&= \sigma^2\sum_{i=1}^{N}(\alpha_i + c_i)^2 = \sigma^2\sum_{i=1}^{N}\left(c_i^2 + \alpha_i^2 + 2\alpha_i c_i\right) \\
&= \sigma^2\sum_{i=1}^{N}c_i^2 + \sigma^2\sum_{i=1}^{N}\alpha_i^2 + 2\sigma^2\sum_{i=1}^{N}\alpha_i c_i = \operatorname{Var}(\hat{\beta}_2) + \sigma^2\sum_{i=1}^{N}\alpha_i^2 \geq \operatorname{Var}(\hat{\beta}_2)\,,
\end{aligned}$$

καθώς $\sum_{i=1}^{N}\alpha_i c_i = 0$ και $\sum_{i=1}^{N}\alpha_i^2 \geq 0$.[2] Η τελευταία σχέση δείχνει ότι, αν $\alpha_i \neq 0$, τότε

$$\operatorname{Var}(\hat{\beta}_2^*) > \operatorname{Var}(\hat{\beta}_2)\,,$$

πράγμα που αποδεικνύει το Gauss-Markov θεώρημα.

---

[2] Η σχέση $\sum_{i=1}^{N}\alpha_i c_i = 0$ αποδεικνύεται εύκολα αντικαθιστώντας σε αυτήν $c_i = \dfrac{(x_i - \bar{x})}{\sum_{i=1}^{N}(x_i - \bar{x})^2}$. Τότε, θα έχουμε

$$\sum_{i=1}^{N}\alpha_i c_i = \frac{\sum_{i=1}^{N}\alpha_i(x_i - \bar{x})}{\sum_{i=1}^{N}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{N}\alpha_i x_i}{\sum_{i=1}^{N}(x_i - \bar{x})^2} - \frac{\bar{x}\sum_{i=1}^{N}\alpha_i}{\sum_{i=1}^{N}(x_i - \bar{x})^2} = 0\,,$$

λόγω των συνθηκών της σχέσης (10).

4. Η κατανομή του εκτιμητή $\hat{\beta}_2$. Όταν $\varepsilon_i \sim N(0,\sigma^2)$, $\forall$ i, η **κατανομή του** $\hat{\beta}_2$ είναι κανονική και δίδεται ως εξής

$$\hat{\beta}_2 \sim N\left[E(\hat{\beta}_2),\ Var(\hat{\beta}_2)\right] \sim N\left[\beta_2,\ \frac{\sigma^2}{\sum_{i=}^{N}(x_i-\bar{x})^2}\right], \qquad (11)$$

Αποδ: Προκύπτει κατευθείαν από τη γραμμική σχέση του LS εκτιμητή $\hat{\beta}_2$ ως προς τις παρατηρήσεις της εξαρτημένης μεταβλητής ή του διαταρακτικού όρου, δηλ.

$$\hat{\beta}_2 = \beta_2 + \sum_{i=1}^{N} c_i\varepsilon_i \quad (\text{ή } \hat{\beta}_2 = \sum_{i=1}^{N} c_i y_i)$$

Όταν $\varepsilon_i \sim N(0,\sigma^2)$ (οπότε $y_i \sim N\left(E(y_i)=\beta_1+\beta_2 x_i,\ Var(y_i)=\sigma^2\right)$), τότε ισχύει το αποτέλεσμα της σχέσης (11) καθώς ο εκτιμητής $\hat{\beta}_2$ αποτελεί γραμμικό συνδυασμό κανονικά κατανεμημένων μεταβλητών και έτσι, και αυτός (που αποτελεί τυχαία μεταβλητή) κατανέμεται κανονικά. Η μέση τιμή και η διακύμανσης της κατανομής του $\hat{\beta}_2$ βρέθηκαν παραπάνω.

Εκτίμηση της διακύμανσης του διαταρακτικού όρου σ^2**:** Η σχέση της διακύμανσης του LS εκτιμητή $\hat{\beta}_2$, $N\left[E(\hat{\beta}_2),\ Var(\hat{\beta}_2)\right] \sim N\left[\beta_2,\ \frac{\sigma^2}{\sum_{i=}^{N}(x_i-\bar{x})^2}\right]$, δείχνει ότι για να υπολογιστεί χρειάζεται να εκτιμηθεί η διακύμανση σ^2. Ένας αμερόληπτος εκτιμητής αυτής είναι ο ακόλουθος:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{N}\hat{\varepsilon}_i^2}{N-K}, \qquad (12)$$

όπου $\hat{\varepsilon}_i = y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i$ είναι τα κατάλοιπα, N είναι ο αριθμός των παρατηρήσεων του δείγματος και K ισούται με τον αριθμό των συντελεστών του υποδείγματος που εκτιμώνται (δηλ. η σταθερά και ο συντελεστής κλίσης, δηλ. K=2). N-K, που αποτελούν τους **βαθμούς ελευθερίας** που διαθέτουμε στην εκτίμηση της σ^2. Για να τους υπολογίσουμε, θα πρέπει από τον αριθμό των παρατηρήσεων του δείγματος N να αφαιρέσουμε τον αριθμό των συντελεστών του υποδείγματος K που εκτιμώνται σε πρώτο στάδιο με τη LS μέθοδο.

Δ. Υπολογισμός Διαστήματος εμπιστοσύνης

Στο παρόν τμήμα θα ασχοληθούμε με τον υπολογισμό διαστημάτων εμπιστοσύνης (confidence intervals) και τον έλεγχο υποθέσεων (hypothesis testing) για τους συντελεστές του γραμμικού υποδείγματος $y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$.

Διάστημα εμπιστοσύνης: Ο υπολογισμός του διαστήματος εμπιστοσύνης ενός συντελεστή του γραμμικού υποδείγματος (πχ του β_2) αφορά στην εκτίμηση δύο τιμών αυτού, την πάνω $\hat{\beta}_2^U$ και κάτω $\hat{\beta}_2^L$, με βάση τις LS εκτιμήσεις του $\hat{\beta}_2$ στο δείγμα. Εντός των τιμών $\hat{\beta}_2^U$ και $\hat{\beta}_2^L$ αναμένεται να βρίσκεται η αληθινή τιμή του στον πληθυσμό με κάποια πιθανότητα, π.χ. 95%, που γράφεται 1-*α*, δηλαδή:

$$\Pr\left(\hat{\beta}_2^L \le \beta_2 \le \hat{\beta}_2^U\right) = 1 - \alpha$$

Για τον υπολογισμό του διαστήματος εμπιστοσύνης (πχ του β_2), χρειαζόμαστε την κατανομή του $\hat{\beta}_2$:

$$\hat{\beta}_2 \sim N\left[\beta_2, Var(\hat{\beta}_2)\right], \quad \text{όπου } Var(\hat{\beta}_2) = \frac{\sigma^2}{\sum_{i=1}^{N}(x_i - \overline{x})^2}.$$

που συνεπάγεται

$$\hat{\beta}_2 - \beta_2 \sim N\left[0, Var(\hat{\beta}_2)\right].$$

Τυποποιώντας την μεταβλητή $\hat{\beta}_2 - \beta_2$, δηλ. διαιρώντας με το τυπική σφάλμα (τυπική απόκλιση) του $\hat{\beta}_2$:

$$SE(\hat{\beta}_2) \equiv \sqrt{Var(\hat{\beta}_2)} = \sqrt{\frac{\sigma^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2}},$$

συνεπάγεται την ακόλουθη τυποποιημένη μεταβλητή z

$$z \equiv \frac{\hat{\beta}_2 - \beta_2}{SE(\hat{\beta}_2)} = \frac{\hat{\beta}_2 - \beta_2}{\sqrt{\sigma^2 \Big/ \sum_{i=1}^{N}(x_i - \bar{x})^2}} \sim N(0,1),$$

που **ακολουθεί την κανονική κατανομή N(0,1)** **<u>όταν η διακύμανση του διαταρακτικού όρου σ² είναι γνωστή</u>**. Η τυχαία μεταβλητή *z* αναφέρεται ως **στατιστικό κριτήριο (ή πιο απλά ως στατιστική)**.

<u>Όταν η διακύμανση σ² αποτελεί άγνωστη παράμετρο και εκτιμάται</u> με βάση τα κατάλοιπα του γραμμικού υποδείγματος ως $\hat{\sigma}^2 = \frac{\sum_{i=1}^{N}\hat{\varepsilon}_i^2}{N-K}$, τότε η στατιστική z δεν ακολουθεί πλέον την τυποποιημένη κανονική κατανομή, **αλλά την t-student κατανομή με βαθμούς ελευθερίας N-K και ορίζεται ως**[3]

[3] Το αποτέλεσμα αυτό αποδεικνύεται εύκολα διαιρώντας και τον αριθμητή και τον παρονομαστή του στατιστικού κριτηρίου t με τον όρο $\sqrt{\frac{\sigma^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2}}$. Τότε, συνεπάγεται

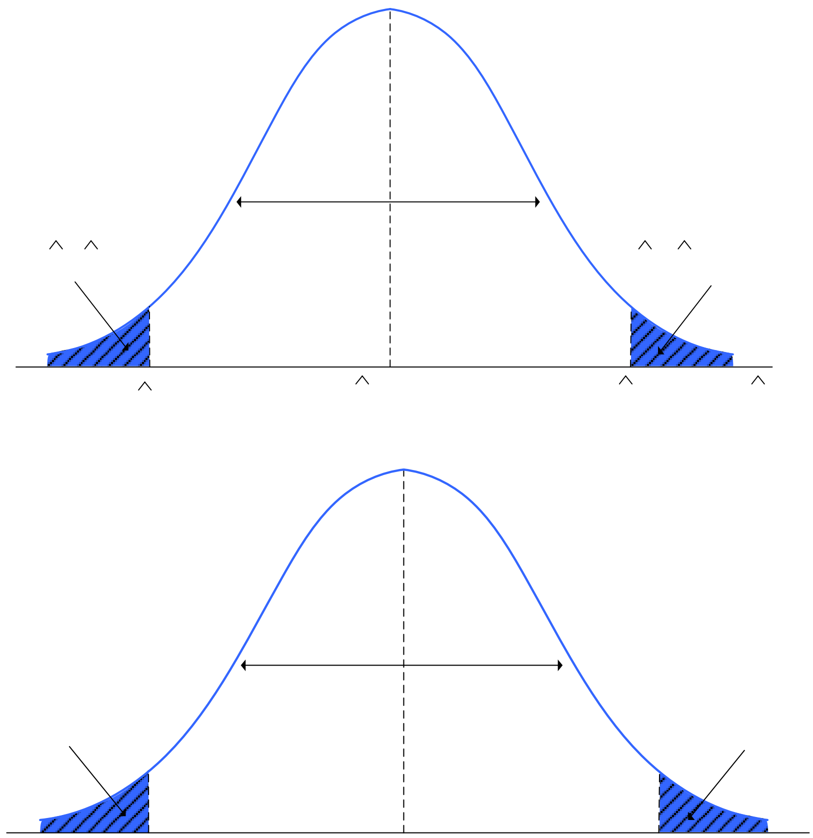
$$t = \frac{\dfrac{\hat{\beta}_2 - \beta_2}{\sqrt{\sigma^2 \Big/ \sum_{i=1}^{N}(x_i - \bar{x})^2}}}{\sqrt{\dfrac{\hat{\sigma}^2}{\sigma^2}}} = \frac{\dfrac{\hat{\beta}_2 - \beta_2}{\sqrt{\sigma^2 \Big/ \sum_{i=1}^{N}(x_i - \bar{x})^2}}}{\sqrt{\dfrac{(N-K)\hat{\sigma}^2}{\sigma^2(N-K)}}},$$

που αποτελεί το πηλίκο δύο τυχαίων μεταβλητών, δηλαδή της $\frac{\hat{\beta}_2 - \beta_2}{\sqrt{\sigma^2 \Big/ \sum_{i=1}^{N}(x_i - \bar{x})^2}}$, που ακολουθεί την τυποποιημένη κανονική κατανομή, ως προς $\frac{(N-K)\hat{\sigma}^2}{\sigma^2}$, που ακολουθεί την χ^2 κατανομή με N-K

$$t \equiv \frac{\hat{\beta}_2 - \beta_2}{SE(\hat{\beta}_2)} = \frac{\hat{\beta}_2 - \beta_2}{\sqrt{\hat{\sigma}^2 \Big/ \sum_{i=1}^{N} (x_i - \bar{x})^2}} \sim t_{(N-K)} \, .$$

Η κατανομή της μεταβλητής $\hat{\beta}_2 - \beta_2$ και της τυποποιημένης της z δίνονται στα ακόλουθα διαγράμματα:

ΣΧΗΜΑ 4.1: Διάστημα εμπιστοσύνης του συντελεστή $\hat{\beta}_2$ και της στατιστικής z



βαθμούς ελευθερίας. δηλαδή $\frac{(N-K)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{(N-K)}$. Επειδή οι δύο αυτές μεταβλητές είναι και ανεξάρτητες μεταξύ τους, το πηλίκο τους που σε όρους κατανομών ορίζεται ως $\frac{N(0,1)}{\sqrt{\chi^2_{(N-K)} \big/ (N-K)}}$ ακολουθεί την t-Student κατανομή με N-K βαθμούς ελευθερίας (βλέπε Παράρτημα Β), που αποδεικνύει την κατανομή του στατιστικού κριτηρίου t.

Στηριζόμενοι στα διαγράμματα του σχήματος 4.1, το διάστημα εμπιστοσύνης του συντελεστή β_2 **<u>όταν η διακύμανση σ² είναι γνωστή</u>** βρίσκεται μέσω του αντιστοίχου του για την τυποποιημένη μεταβλητής *z* ως εξής:

$$\Pr(-z_c \le z \le z_c) = 1-\alpha,$$

όπου z_c και $-z_c$ αποτελούν τις κριτικές τιμές της τυποποιημένης κανονικής κατανομής που ορίζονται ως $\Pr(z < -z_c) = \frac{\alpha}{2}$ και $\Pr(z > z_c) = \frac{\alpha}{2}$, και συμβολίζονται ως $z_c = z_{(\alpha/2)}$ και $-z_c = -z_{(\alpha/2)}$.

Αντικαθιστώντας $z = \frac{\hat{\beta}_2 - \beta_2}{SE(\hat{\beta}_2)}$ στην παραπάνω σχέση, όπου $SE(\hat{\beta}_2) = \sqrt{Var(\hat{\beta}_2)}$, έχουμε:

$$\Pr(-z_c \le \frac{\hat{\beta}_2 - \beta_2}{SE(\hat{\beta}_2)} \le z_c) = 1-\alpha \quad \Rightarrow$$

$$\Pr\left(z_c \ge -\frac{\hat{\beta}_2 - \beta_2}{SE(\hat{\beta}_2)} \ge -z_c\right) = 1-\alpha \quad \Rightarrow$$

$$\Pr\left(\hat{\beta}_2 + z_c \cdot SE(\hat{\beta}_2) \ \ge\ \beta_2 \ \ge\ \hat{\beta}_2 - z_c \cdot SE(\hat{\beta}_2)\right) = 1-\alpha \quad \Rightarrow$$

$$\Pr\left(\hat{\beta}_2 - z_c \cdot SE(\hat{\beta}_2) \ \le\ \beta_2 \ \le\ \hat{\beta}_2 + z_c \cdot SE(\hat{\beta}_2)\right) = 1-\alpha$$

ή ως

$$\Pr\left(\hat{\beta}_2^L \ \le\ \beta_2 \ \le\ \hat{\beta}_2^U\right) = 1-\alpha,$$

όπου

$$\hat{\beta}_2^L = \hat{\beta}_2 - z_c \cdot SE(\hat{\beta}_2) \quad \text{και} \quad \hat{\beta}_2^U = \hat{\beta}_2 + z_c \cdot SE(\hat{\beta}_2)$$

αποτελούν αντίστοιχα το άνω και το κάτω όριο του διαστήματος εμπιστοσύνης β_2.

<u>Όταν η διακύμανση σ² είναι άγνωστη και εκτιμάται από το δείγμα</u>, το διάστημα εμπιστοσύνης του β_2 υπολογίζεται μέσω της t-student και δίδεται ως:

$$\Pr\left(\hat{\beta}_2 - t_c \cdot SE(\hat{\beta}_2) \leq \beta_2 \leq \hat{\beta}_2 + t_c \cdot SE(\hat{\beta}_2)\right) = 1 - \alpha,$$

όπου $-t_c$ και t_c αποτελούν κριτικές τιμές της t-Student με T–K βαθμούς ελευθερίας που ορίζονται ως $\Pr(t < -t_c) = \frac{\alpha}{2}$ και $\Pr(t > t_c) = \frac{\alpha}{2}$, αντίστοιχα, και συμβολίζονται ως $t_c = t_{(\alpha/2,N\text{-}K)}$ και $-t_c = -t_{(\alpha/2,N\text{-}K)}$. Το κάτω και το άνω όριο εμπιστοσύνης του διαστήματος του συντελεστή β_2 που συνεπάγονται από την παραπάνω σχέση δίνονται αντίστοιχα ως εξής: $\hat{\beta}_2^L = \hat{\beta}_2 - t_c \cdot SE(\hat{\beta}_2)$ και $\hat{\beta}_2^U = \hat{\beta}_2 + t_c \cdot SE(\hat{\beta}_2)$.

Ε. Έλεγχοι Στατιστικών Υποθέσεων

Οι έλεγχοι στατιστικών υποθέσεων **ενδιαφέρονται να ελέγξουν** κάποια στατιστική υπόθεση για τους συντελεστές του γραμμικού υποδείγματος (εδώ το συντελεστή β_2) που αναφέρεται ως **Μηδενική (συμβολίζεται ως H_0) έναντι κάποιας Εναλλακτικής της (που συμβολίζεται ως H_α).** Παραδείγματα H_0 και H_α υποθέσεων:

$$\begin{array}{l} H_0: \ \beta_2 \leq \beta_2^0 \\ H_\alpha: \ \beta_2 > \beta_2^0 \end{array} \qquad \begin{array}{l} H_0: \ \beta_2 \geq \beta_2^0 \\ H_\alpha: \ \beta_2 < \beta_2^0 \end{array} \qquad \text{και} \qquad \begin{array}{l} H_0: \ \beta_2 = \beta_2^0 \\ H_\alpha: \ \beta_2 \neq \beta_2^0 \end{array},$$

όπου β_2^0 συμβολίζει μια υποθετική (θεωρητική) τιμή του β_2 στον πληθυσμό.

Για την απόρριψη της H_0 και την αποδοχή της H_α βασιζόμαστε στην κατανομή κάποιου στατιστικού κριτηρίου (όπως τα κριτήρια z και t) κάτω από τη μηδενική υπόθεση, η οποία υποθέτουμε αρχικά ότι είναι αληθής. Στηριζόμενοι στην κατανομή αυτή και την τιμή του στατιστικού κριτηρίου z ή t, απορρίπτουμε (ή αποδεχόμαστε τη μηδενική υπόθεση H_0) σύμφωνα με κάποιον κανόνα (**Κανόνα απόφασης**).

Ο κανόνας απόφασης ορίζει την περιοχή της κατανομής του στατιστικού κριτηρίου *z* ή *t* για τις οποίες απορρίπτεται η μηδενική υπόθεση ως περιοχή απόρριψης, ενώ εκείνη για τις οποίες γίνεται δεκτή ορίζεται ως περιοχή αποδοχής. Οι περιοχές αυτές μας επιτρέπουν να ταξινομήσουμε τις δυνατές αποφάσεις ενός στατιστικού ελέγχου με τα σφάλματά τους στις ακόλουθες τρεις κατηγορίες:

(1.) η σωστή υπόθεση (H_0) να απορριφθεί: Το σφάλμα απόρριψης της H_0 όταν είναι αληθής ονομάζεται **σφάλμα τύπου Ι** και ορίζεται ως

α = P(Σφάλμα τύπου Ι) = P(Απόρριψη H_0 | H_0 αληθής) - **επίπεδο σημαντικότητας**

(2.) η λανθασμένη υπόθεση να γίνει δεκτή: Το σφάλμα αποδοχής της H_0 όταν είναι λανθασμένη ονομάζεται **σφάλμα τύπου ΙΙ**

β = P(Σφάλμα τύπου ΙΙ) = P(Αποδοχή H_0 | H_α αληθής)

.

(3.) να γίνει αποδεκτή η υπόθεση H_0 όταν είναι σωστή, ή να απορριφθεί όταν είναι λανθασμένη: Στη δεύτερη περίπτωση η πιθανότητα απόρριψής της H_0 δίνεται ως

P(Απόρριψη H_0 | H_α αληθής) = 1- P(Αποδοχή H_0 | H_α αληθής)

= 1- P(Σφάλμα τύπου ΙΙ) = 1-β.

δύναμη του κριτηρίου

Από τους παραπάνω ορισμούς των σφαλμάτων Ι και ΙΙ φαίνεται ότι θα ήταν ιδανικό στη διεξαγωγή στατιστικών ελέγχων να διατηρούμε και τα δύο αυτά σφάλματα Ι και ΙΙ όσο το δυνατό μικρότερα. Δυστυχώς όμως αυτό δεν είναι εφικτό. Για το λόγο αυτό στον ***κλασικό έλεγχο υποθέσεων*** *(κλασική στατιστική), θέτουμε τη μέγιστη πιθανότητα του σφάλματος τύπου Ι σε κάποιο επίπεδο πιθανότητας που είναι γενικά αποδεκτό (έστω α=5% ή α=10%) και, με βάση αυτό, προσπαθούμε να επιλέξουμε εκείνο το στατιστικό κριτήριο που ελαχιστοποιεί την πιθανότητα σφάλματος τύπου ΙΙ. Η επιλογή αυτή* *τεκμηριώνεται από τη γενική αρχή του δικαίου ότι είναι προτιμότερο να μην καταδικάσουμε κατά λάθος έναν αθώο (δηλαδή, να μην απορρίψουμε την μηδενική υπόθεση όταν είναι αληθής) παρά να αθωώσουμε έναν ένοχο (δηλαδή, να κάνουμε δεκτή τη μηδενική υπόθεση παρόλο που είναι λανθασμένη).*

Ατομικοί έλεγχοι στατιστικών υποθέσεων

(i) Μονοκατάληκτοι έλεγχοι:

$$\begin{aligned} H_0 &: \ \beta_2 \geq \beta_2^0 \\ H_\alpha &: \ \beta_2 < \beta_2^0 \end{aligned}$$

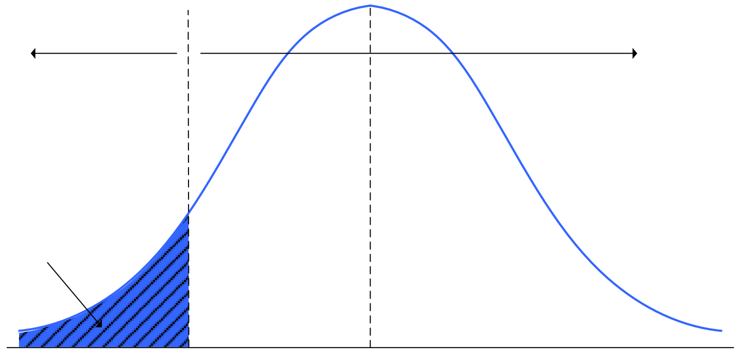
(προς τα αριστερά)

Με βάση κάποιο επίπεδο σημαντικότητας α (σφάλμα τύπου I), απορρίπτουμε τη H_0 προς τα αριστερά, δηλαδή για το κριτήριο z απορρίπτουμε την H_0 και δεχόμαστε την H_α όταν ή τιμή του (ζ ή z):

$$z \,(\text{ή } \zeta) \ < -z_{(\alpha=0.05)} = -1.65\,,$$

για α=0.05. Η περιοχή απόρριψης του ελέγχου αυτού ορίζεται ως $\Pr(z < -z_c) = \alpha$

ΣΧΗΜΑ 4.3: Μονοκατάληκτος έλεγχος προς τα αριστερά



Εναλλακτικά, ο έλεγχος μπορεί να γίνει με βάση την **p-τιμή μιας τιμής** ζ του στατιστικού κριτηρίου z που υπολογίζεται ως εξής:

$$\text{p-τιμή} = P(z \leq \zeta) = \Phi(\zeta) = \int_{-\infty}^{\zeta} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz\,,$$

Αν p-τιμή < α, απορρίπτουμε την H_0 .

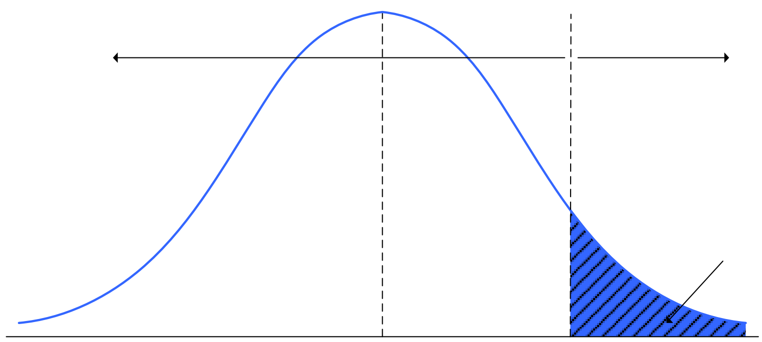
$$\begin{aligned} H_0 &: \ \beta_2 \leq \beta_2^0 \\ H_\alpha &: \ \beta_2 > \beta_2^0 \end{aligned}$$ **(προς τα δεξιά)**

Απορρίπτουμε τη H_0 προς τα δεξιά, δηλαδή για το κριτήριο z απορρίπτουμε την H_0 και δεχόμαστε την H_α όταν ή τιμή του (ζ ή z):

$$z \text{ (ή } \zeta) \ > -z_{(\alpha=0.05)} = 1.65 ,$$

για α=0.05. Η περιοχή απόρριψης του ελέγχου αυτού ορίζεται ως $\Pr(z > z_c) = \alpha$.

ΣΧΗΜΑ 4.2: Μονοκατάληκτος έλεγχος προς τα δεξιά



Εναλλακτικά ο έλεγχος μπορεί να γίνει με βάση την **p-τιμή μιας τιμής** ζ του στατιστικού κριτηρίου z υπολογίζεται ως εξής:

$$\text{p-τιμή} \equiv P(z \geq \zeta) = 1 - P(z \leq \zeta) \ = \ 1 - \Phi(\zeta) = 1 - \int_{-\infty}^{\zeta} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$$

Αν p-τιμή > α, απορρίπτουμε την H_0 .

(ii) Δικατάληκτοι έλεγχοι

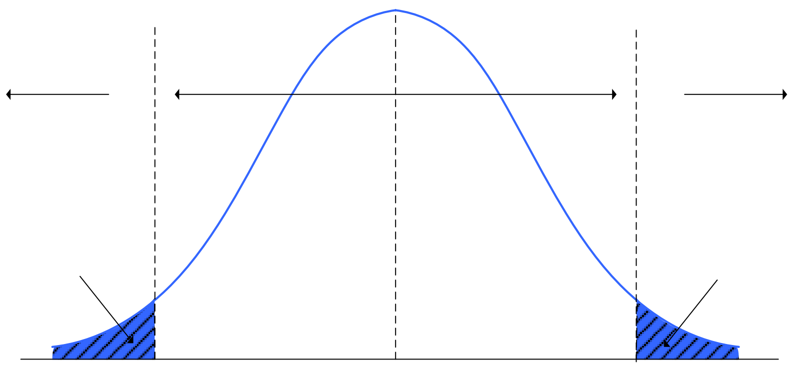
$$\begin{aligned} H_0 &: \quad \beta_2 = \beta_2^0 \\ H_\alpha &: \quad \beta_2 \neq \beta_2^0 \end{aligned}$$

Απορρίπτουμε τη H_0 και προς τα δεξιά και αριστερά, δηλαδή για το κριτήριο z απορρίπτουμε την H_0 και δεχόμαστε την H_α όταν ή τιμή του (ζ ή z):

$$z\ (\text{ή } \zeta) < -z_{(\alpha/2=0.025)} = -1.96 \qquad \text{ή} \qquad z\ (\text{ή } \zeta) < z_{(\alpha/2=0.025)} = 1.96\ ,$$

για α=0.05. Οι περιοχές απόρριψης του ελέγχου αυτού ορίζονται ως $\Pr(z < -z_c) = \dfrac{\alpha}{2}$ και $\Pr(z > z_c) = \dfrac{\alpha}{2}$, και ικανοποιούν τη σχέση: $\Pr(z < -z_c) + \Pr(z > z_c) = \dfrac{\alpha}{2} + \dfrac{\alpha}{2} = \alpha$

ΣΧΗΜΑ 4.4: Δικατάληκτος στατιστικός έλεγχος



Εναλλακτικά, ο έλεγχος μπορεί να γίνει με βάση την **p-τιμή μιας τιμής** ζ του στατιστικού κριτηρίου z που υπολογίζεται ως εξής:

$$\text{p-τιμή} = P(z \le -\zeta) + P(z \ge \zeta) = 2P(z \le -\zeta) = 2\int_{-\infty}^{-\zeta} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}\, dz$$

Αν p-τιμή < α, απορρίπτουμε την H_0 .

Παραδείγματα ελέγχων στατιστικών υποθέσεων

ΠΑΡΑΔΕΙΓΜΑ 1: Έστω η ακόλουθη εκτίμηση CAPM): $r_t^j = \beta_1 + \beta_2 (R_t^M - r_t^f) + \varepsilon_t$ με βάση ένα δείγμα T=20 χρονολογικών παρατηρήσεων:

$$r_t^j = 0.06 + 1.5(R_t^M - r_t^f) + \hat{\varepsilon}_t, \qquad t = 1, 2, \ldots, T = 20,$$

$$\text{SE}(\hat{\beta}_i): \qquad (0.03) \quad (0.25) \qquad\qquad \hat{\sigma}^2 = 0.05$$

όπου r_t^j συμβολίζει την απόδοση μιας μετοχής j την περίοδο t, R_t^M είναι η απόδοση του χαρτοφυλακίου της αγοράς M και $r^f{}_t$ συμβολίζει το επιτόκιο της αγοράς. Τυπικά σφάλματα (standard errors) βρίσκονται στις παρενθέσεις. Αν υποθέσουμε ότι οι τιμές του διαταρακτικού όρου κατανέμονται κανονικά, δηλαδή $\varepsilon_t \sim N(0,\sigma^2),\ \forall t$, τότε ελέγξατε τις ακόλουθες στατιστικές υποθέσεις για τους συντελεστές του παραπάνω υποδείγματος. Όλοι οι έλεγχοι να διεξαχθούν σε επίπεδο σημαντικότητας α=0.05.

(i) *Αν η αληθινή τιμή του συντελεστή β_1 είναι* $\beta_1^0 \geq 0.03$

Ο έλεγχος της υπόθεσης αυτής αποτελεί έναν μονοκατάληκτο έλεγχο:

$$H_0:\quad \beta_1^0 \geq 0.03$$

$$H_a:\quad \beta_1 < 0.03,$$

Επειδή η διακύμανση του διαταρακτικού όρου δεν είναι γνωστή αλλά εκτιμάται ως $\hat{\sigma}^2 = 0.05$, για τη διεξαγωγή του παραπάνω ελέγχου θα χρησιμοποιήσουμε το στατιστικό κριτήριο t. Αυτό υπολογίζεται ως

$$t = \frac{\hat{\beta}_1 - \beta_1^0}{SE(\hat{\beta}_1)} = \frac{0.06 - 0.03}{0.03} = \frac{0.03}{0.03} = 1.0$$

Η κριτική τιμή απόρριψης της μηδενικής υπόθεσης για επίπεδο σημαντικότητας α=0.05 είναι $-t_c = -t_{(\alpha=0.05,T\text{-}K=18)} = -1.73$. Η τιμή αυτή βρίσκεται με βάση τη συμμετρική της, που δίνεται από τους πίνακες της t-Student κατανομής για T-K = 20-2=18 βαθμούς ελευθερίας και α=0.05 ως $t_{(\alpha=0.05,T\text{-}K=18)} = 1.73$. Συγκρίνοντας την κριτική τιμή του στατιστικού $-t_c = -1.73$ με την τιμή του στατιστικού t=1.0 παρατηρούμε ότι:

$$t = 1.0 > -t_c = -1.73,$$

που σημαίνει ότι η μηδενική υπόθεση H_0 δεν απορρίπτεται. Ο έλεγχος αυτός απεικονίζεται στο ακόλουθο σχήμα:

(ii) Ελέγξατε αν ο συντελεστής βήτα του CAPM (β_2, εδώ) είναι διάφορος του μηδενός (δηλ. $\beta_2 \neq \beta_2^0 = 0$), σημαίνει ότι αυτός είναι στατιστικά σημαντικός. Για να ελεγχθεί η υπόθεση αυτή θα υπολογίσουμε το στατιστικό κριτήριο t:

$$t = \frac{\hat{\beta}_2 - \beta_2^0}{SE(\hat{\beta}_2)} = \frac{1.5 - 0.0}{0.25} = 6 \quad \Rightarrow \quad |t| = 6,$$

όπου $\beta_2^0 = 0.00$ αποτελεί την τιμή του συντελεστή β_2 που ελέγχεται κάτω από τη μηδενική υπόθεση. Επειδή ο έλεγχος αυτός είναι δικατάληκτος, θα θέσουμε τη μηδενική και εναλλακτική υπόθεση αντίστοιχα ως:

$$H_0: \quad \beta_2 = 0$$
$$H_\alpha: \quad \beta_2 \neq 0$$

Η κριτική τιμή για τον έλεγχο αυτό βρίσκεται από τους πίνακες της t-student κατανομής ως $t_c = t_{(\alpha/2=0.025, T-K=18)} = 2.10$. Συγκρίνοντας την τιμή του κριτηρίου $|t|=6$ με την κριτική τιμή $t_c=2.10$ έχουμε

$$|t| = 6 > t_c = 2.10,$$

που σημαίνει ότι η μηδενική υπόθεση H_0: $\beta_2 = 0$ απορρίπτεται.