Α. Πρόβλεψη μέσω ενός απλού γραμμικού υποδείγματος

Θεωρητική πρόβλεψη του οικονομετρικού υποδείγματος (αληθινή τιμή) για κάποια γνωστή τιμή (αναφέρεται ως **σημείο "0"**) της ανεξάρτητης μεταβλητής x_0.

$$y_0 = \beta_1 + \beta_2 x_0 + \varepsilon_0, \qquad (1)$$

Πρόβλεψη του οικονομικού υποδείγματος στον πληθυσμό για την τιμή x_0:

$$E(y_0) = \beta_1 + \beta_2 x_0, \quad \text{καθώς } E(\varepsilon_0)=0.$$

Πρόβλεψη με βάση τις LS εκτιμήσεις $\hat{\beta}_1$ και $\hat{\beta}_2$ του δείγματος:[1]

$$\hat{y}_0 = \hat{\beta}_1 + \hat{\beta}_2 x_0, \qquad (2)$$

Σφάλμα πρόβλεψης:

$$\hat{y}_0 - y_0 = (\hat{\beta}_1 - \beta_1) + (\hat{\beta}_2 - \beta_2)x_0 - \varepsilon_0. \qquad (3)$$

- Μέση τιμή του σφάλματος $\hat{y}_0 - y_0$

$$E(\hat{y}_0 - y_0) = E(\hat{\beta}_1 - \beta_1) + x_0 E(\hat{\beta}_2 - \beta_2) - E(\varepsilon_0) = 0,$$

που σημαίνει αμεροληψία προβλέψεων.

- Διακύμανση του σφάλματος $\hat{y}_0 - y_0$:

$$\mathrm{Var}(\hat{y}_0 - y_0) = \mathrm{Var}\left[(\hat{\beta}_1 - \beta_1) + (\hat{\beta}_2 - \beta_2)x_0 - \varepsilon_0\right]$$

[1] Η δε σχέση (2) με βάση την οποία υπολογίζουμε την πρόβλεψη αναφέρεται ως εκτιμητής της πρόβλεψης, καθώς βασίζεται στους εκτιμητές $\hat{\beta}_1$ και $\hat{\beta}_2$.

$$= \mathrm{Var}(\hat{\beta}_1 - \beta_1) + x_0^2 \mathrm{Var}(\hat{\beta}_2 - \beta_2) + \mathrm{Var}(\varepsilon_0) + 2x_0 \mathrm{Cov}(\hat{\beta}_1 - \beta_1; \hat{\beta}_2 - \beta_2)$$

$$= \mathrm{Var}(\hat{\beta}_1) + x_0^2 \mathrm{Var}(\hat{\beta}_2) + \sigma^2 + 2x_0 \mathrm{Cov}(\hat{\beta}_1, \hat{\beta}_2)\,. \qquad (4)$$

Χρησιμοποιώντας τους τύπους διακυμάνσεων-συνδιακυμάνσεων των $\hat{\beta}_1$ και $\hat{\beta}_2$

$$\mathrm{Var}(\hat{\beta}_1) = \frac{\sigma^2 \sum_{i=1}^{N} x_i^2}{N \sum_{i=1}^{N} (x_i - \bar{x})^2},\ \mathrm{Var}(\hat{\beta}_2) = \frac{\sigma^2}{\sum_{i=1}^{N} (x_i - \bar{x})^2},\ \mathrm{Cov}(\hat{\beta}_1, \hat{\beta}_2) = -\frac{\sigma^2 \bar{x}}{\sum_{i=1}^{N} (x_i - \bar{x})^2}$$

η σχέση (4) γράφεται ως

$$\mathrm{Var}(\hat{y}_0 - y_0) = \frac{\sigma^2 \sum_{i=1}^{N} x_i^2}{N \sum_{i=1}^{N} (x_i - \bar{x})^2} + \frac{\sigma^2}{\sum_{i=1}^{N} (x_i - \bar{x})^2} x_0^2 + \sigma^2 - \frac{2\bar{x}\sigma^2}{\sum_{i=1}^{N} (x_i - \bar{x})^2} x_0$$

$$= \sigma^2 \left[\frac{\sum_{i=1}^{N} x_i^2}{N \sum_{i=1}^{N} (x_i - \bar{x})^2} + \frac{N x_0^2}{N \sum_{i=1}^{N} (x_i - \bar{x})^2} - \frac{2N\bar{x}x_0}{N \sum_{i=1}^{N} (x_i - \bar{x})^2} + 1 \right]$$

$$= \sigma^2 \left[\frac{\sum_{i=1}^{N} x_i^2 + N x_0^2 - 2N\bar{x}x_0 + N\bar{x}^2 - N\bar{x}^2}{N \sum_{i=1}^{N} (x_i - \bar{x})^2} + 1 \right]$$

$$= \sigma^2 \left[\frac{\sum_{i=1}^{N} (x_i^2 - \bar{x}^2)}{N \sum_{i=1}^{N} (x_i - \bar{x})^2} + \frac{N(x_0 - \bar{x})^2}{N \sum_{i=1}^{N} (x_i - \bar{x})^2} + 1 \right], \quad \text{καθώς} \quad \sum_{i=1}^{N} (x_i - \bar{x})^2 = \sum_{i=1}^{N} (x_i^2 - \bar{x}^2)$$

$$= \sigma^2 \left[\frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{N \sum_{i=1}^{N} (x_i - \bar{x})^2} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{N} (x_i - \bar{x})^2} + 1 \right]$$

$$= \sigma^2 \left[\frac{1}{N} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{N} (x_i - \bar{x})^2} + 1 \right] \qquad (5)$$

Η τελευταία σχέση δείχνει:

i) $\mathrm{Var}(\hat{y}_0 - y_0)$ αυξάνεται όταν το μέγεθος του δείγματος N μειώνεται

ii) $\mathrm{Var}(\hat{y}_0 - y_0)$ αυξάνεται όταν η απόκλιση $x_0 - \bar{x}$ μεγαλώνει, δηλ. δεν προβλέπουμε για τη μέση τιμή της ανεξάρτητης μεταβλητής $\bar{x}$, και

iii) $\mathrm{Var}(\hat{y}_0 - y_0)$ αυξάνεται όταν η διακύμανση του διαταρακτικού όρου σ^2 μεγαλώνει.

Διάστημα εμπιστοσύνης της πρόβλεψης $\hat{y}_0$

Αν υποθέσουμε ότι $\varepsilon_i \sim N(0,\sigma^2)$, τότε το σφάλμα πρόβλεψης, που ορίζεται ως

$$\hat{y}_0 - y_0 = (\hat{\beta}_1 - \beta_1) + (\hat{\beta}_2 - \beta_2)x_0 - \varepsilon_0,$$

ακολουθεί και αυτό την κανονική κατανομή, δηλ.

$$\hat{y}_0 - y_0 \sim N\left[0,\ \mathrm{Var}(\hat{y}_0 - y_0)\right],$$

καθώς $E(\hat{y}_0 - y_0) = 0$ και $\mathrm{Var}(\hat{y}_0 - y_0)$ δίνεται από τη σχέση (5). Η τυποποιημένη κατανομή του $\hat{y}_0 - y_0$ είναι:

$$\frac{\hat{y}_0 - y_0}{\sqrt{\mathrm{Var}(\hat{y}_0 - y_0)}} \sim N(0,1).$$

Με βάση την κατανομή αυτή και θεωρώντας ότι η διακύμανση σ^2 εκτιμάται βρίσκουμε το διάστημα εμπιστοσύνης της αληθινής τιμής y_0 ορίζεται ως

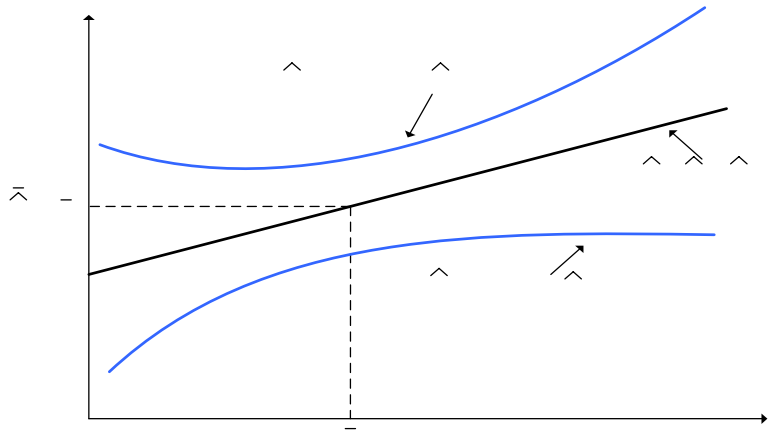
$$\Pr\left[\hat{y}_0 - t_c \cdot \sqrt{\mathrm{Var}(\hat{y}_0 - y_0)} \ \le\ y_0 \ \le\ \hat{y}_0 + t_c \cdot \sqrt{\mathrm{Var}(\hat{y}_0 - y_0)}\right] = 1 - \alpha$$

και δίνεται ως (βλέπε Διάγραμμα)

$$\hat{y}_0 - t_c \cdot \sqrt{\text{Var}(\hat{y}_0 - y_0)} \le y_0 \le \hat{y}_0 + t_c \cdot \sqrt{\text{Var}(\hat{y}_0 - y_0)},$$

όπου t_c αποτελεί την κριτική τιμή της t-student κατανομής με βαθμούς ελευθερίας N-K για επίπεδο σημαντικότητας α.

ΣΧΗΜΑ 5.1 Διάστημα πρόβλεψης



Β. Πρόβλεψη βάσει πολλαπλού υποδείγματος $y_i=\beta_1+\beta_2 x_{i2}+\beta_3 x_{i3}+...+\beta_K x_{iK}+\varepsilon_i$.

Πρόβλεψη με βάση τις LS εκτιμήσεις των συντελεστών του υποδείγματοςδείγματος:

$$\underset{(1X1)}{\hat{y}_0} = [1, x_{02}, ..., x_{0K}] \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_K \end{bmatrix} = \mathbf{x}_0' \hat{\boldsymbol{\beta}},$$

για το σημείο "0" του διανύσματος των ανεξάρτητων μεταβλητών: $\underset{(KX1)}{\mathbf{x}_0'} = [1, x_{02}, ..., x_{0K}]$.

Σφάλμα της πρόβλεψης $\hat{y}_0$:

$$\hat{y}_0 - y_0 = \mathbf{x}_0'\hat{\boldsymbol{\beta}} - (\mathbf{x}_0'\boldsymbol{\beta} + \varepsilon_0) = \mathbf{x}_0'(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) - \varepsilon_0 .$$

- Μέση τιμή σφάλματος πρόβλεψης: $\mathrm{E}(\hat{y}_0 - y_0) = E\left[\mathbf{x}_0'(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) - \varepsilon_0\right] = 0$

- Διακύμανσή του σφάλματος πρόβλεψης:

$$\mathrm{Var}(\hat{y}_0 - y_0) = \mathrm{E}(\hat{y}_0 - y_0)^2$$

$$= \mathrm{E}\left[\mathbf{x}_0'(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) - \varepsilon_0\right]^2 = \mathrm{E}\left[\mathbf{x}_0'(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\right]^2 + \mathrm{E}(\varepsilon_0)^2 - 2\mathrm{E}\left[\mathbf{x}_0'(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\varepsilon_0\right]$$

$$= \sigma^2\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0 + \sigma^2 = \sigma^2\left[\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0 + 1\right], \qquad (6)$$

καθώς ισχύουν τα ακόλουθα αποτελέσματα:

$$\mathrm{E}\left[\mathbf{x}_0'(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\right]^2 = \mathrm{E}\left[\mathbf{x}_0'(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\mathbf{x}_0\right]$$

$$= \mathbf{x}_0'\mathrm{Var}(\hat{\boldsymbol{\beta}})\mathbf{x}_0 = \sigma^2\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0, \quad \text{όπου } \mathrm{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1},$$

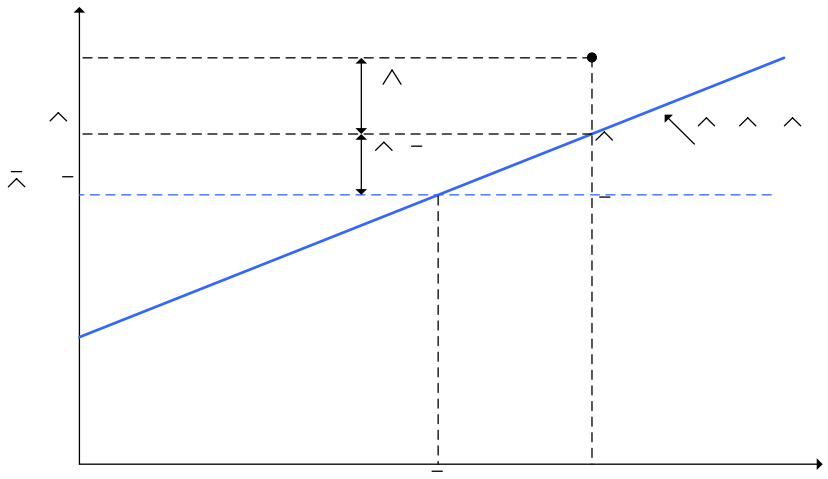
$\mathrm{E}[\varepsilon_0]^2 = \sigma^2$, καθώς $\mathrm{E}[\varepsilon_i]^2 = \sigma^2$ για όλα τα i, και

$$\mathrm{E}\left[\mathbf{x}_0'(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\varepsilon_0\right] = \mathrm{E}\left[\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}\varepsilon_0\right] = \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathrm{E}(\boldsymbol{\varepsilon}\varepsilon_0) = \mathbf{0},$$

όπου $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}$. Σημείωσε ότι $\mathrm{E}(\boldsymbol{\varepsilon}\varepsilon_0) = \mathbf{0}$, καθώς οι παρατηρήσεις του διανύσματος των τιμών του διαταρακτικού όρου **ε** θεωρούνται ως ανεξάρτητες μεταξύ τους και θεωρούμε ότι κάνουμε προβλέψεις σε κάποιο σημείο "0", που δεν συμπεριλαμβάνεται στις παρατηρήσεις του δείγματος. Το σημείο αυτό αναφέρεται και ως εκτός του δείγματος, και μας δίνει προβλέψεις για τιμές της μεταβλητής y_i εκτός του δείγματος. Αυτές αναφέρονται ρητά ως **εκτός του δείγματος προβλέψεις**.

Γ. Προβλεπτική (ερμηνευτική) ικανότητα υποδ. – ο συντελεστής R^2

ΣΧΗΜΑ 5.2: Ερμηνευμένο και ανερμήνευτο μέρος του υποδείγματος



Ο συντελεστής R^2, που θα ορίσουμε στη συνέχει, δείχνει το ποσοστό των μεταβολών της εξαρτημένης μεταβλητής y_i που μπορεί να προβλεφθεί από το υπόδειγμα (όλες δηλαδή τις ανεξάρτητες μεταβλητές του – βλέπε Σχήμα). Για να οριστεί γράψτε

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

ή

$$(y_i - \bar{y}) = (\hat{y}_i - \bar{y}) + \hat{\varepsilon}_i \qquad (7)$$

και υψώστε και τα δύο μέλη της σχέσης (7) στο τετράγωνο

$$(y_i - \bar{y})^2 = [(\hat{y}_i - \bar{y}) + \hat{\varepsilon}_i]^2 = (\hat{y}_i - \bar{y})^2 + \hat{\varepsilon}_i^2 + 2(\hat{y}_i - \bar{y})\hat{\varepsilon}_i$$

Παίρνοντας το άθροισμα της παραπάνω σχέσης για όλες τις παρατηρήσεις του δείγματος συνεπάγεται

$$\sum_{i=1}^{N}(y_i - \bar{y})^2 = \sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{N}\hat{\varepsilon}_i^2 + 2\sum_{i=1}^{N}(\hat{y}_i - \bar{y})\hat{\varepsilon}_i = \sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{N}\hat{\varepsilon}_i^2 ,$$

καθώς ισχύει $\sum_{i=1}^{N}(\hat{y}_i - \bar{y})\hat{\varepsilon}_i = 0$. Η τελευταία σχέση γράφεται ως

$$TSS = ESS + RSS, \tag{8}$$

όπου

$TSS = \sum_{i=1}^{N}(y_i - \bar{y})^2$ (Total Sum of Squares) αποτελεί το συνολικό άθροισμα των τετραγώνων των αποκλίσεων του y_i από το μέσο $\bar{y}$.

$ESS = \sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2$ (Explained Sum of Squares) αποτελεί το μέρος του TSS που ερμηνεύεται από το υπόδειγμα μέσω των μεταβολών της μεταβλητής x_i,

$RSS = \sum_{i=1}^{N}\hat{\varepsilon}_i^2$ (Residual Sum of Squares) αποτελεί το ανερμήνευτο άθροισμα του TSS, που οφείλεται στις μεταβολές του διαταρακτικού όρου. Αυτό αποτελεί το άθροισμα των τετραγώνων των καταλοίπων.

Διαιρώντας το αριστερό και το δεξιό σκέλος της σχέσης (8) με το άθροισμα TSS δίδει την ακόλουθη σχέση που ορίζει το συντελεστή προσδιορισμού R^2:

$$\frac{TSS}{TSS} = \frac{ESS}{TSS} + \frac{RSS}{TSS} \quad \Rightarrow \quad R^2 \equiv \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS},$$

ή

$$R^2 \equiv 1 - \frac{\sum_{i=1}^{N}\hat{\varepsilon}_i^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}, \tag{9}$$

με τιμές $0 < R^2 < 1$

[2] Μπορεί να γραφεί και ως $R^2 = 1 - \frac{\sum_{i=1}^{N}\hat{\varepsilon}_i^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^{N}\hat{\varepsilon}_i^2}{\sum_{i=1}^{N}y_i^2 - N\bar{y}^2} = 1 - \frac{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{\mathbf{y}'\mathbf{y} - N\bar{y}^2}$

Ο συντελεστής R^2 μετρά το ποσοστό του TSS που ερμηνεύεται με βάση τις ανεξάρτητες μεταβλητές του γραμμικού υποδείγματος, δηλαδή από το ESS.

Ιδιότητες του R^2:

1. Αποτελεί το τετράγωνο του συντελεστή γραμμικής συσχέτισης ανάμεσα στις τιμές της y_i και τις προβλέψεις της $\hat{y}_i$, $r_{\hat{y}y}$, δηλ.

$$r_{\hat{y}y}^2 = \frac{\left[\sum_{i=1}^{N}(\hat{y}_i - \bar{y})(y_i - \bar{y})\right]^2}{\sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2 \sum_{i=1}^{N}(y_i - \bar{y})^2} = R^2,$$

Αποδ.

$$r_{\hat{y}y}^2 = \frac{\left[\sum_{i=1}^{N}(\hat{y}_i - \bar{y})(y_i - \bar{y})\right]^2}{\sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2 \sum_{i=1}^{N}(y_i - \bar{y})^2}$$

Αντικατέστησε $(y_i - \bar{y}) = (\hat{y}_i - \bar{y}) + \hat{\varepsilon}_i$, τότε

$$\hat{r}_{\hat{y}y}^2 = \frac{\left[\sum_{i=1}^{N}(\hat{y}_i - \bar{y})\left[(\hat{y}_i - \bar{y}) + \hat{\varepsilon}_i\right]\right]^2}{\sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2 \sum_{i=1}^{N}(y_i - \bar{y})^2}$$

$$= \frac{\left[\sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2\right]^2}{\sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2 \sum_{i=1}^{N}(y_i - \bar{y})^2}, \qquad \text{καθώς } \sum_{i=1}^{N}(\hat{y}_i - \bar{y})\hat{\varepsilon}_i = 0,$$

$$= \frac{\sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} = R^2.$$

2. R^2 γράφεται ως

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\frac{1}{N}\sum_{i=1}^{N}\hat{\varepsilon}_i^2}{\frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^2} = 1 - \frac{\hat{\sigma}^2}{\hat{\sigma}_y^2},$$

και για μικρά δείγματα διορθώνεται λαμβάνοντας υπόψη τους ΒΕ στις εκτιμήσεις των $\hat{\sigma}^2$ και $\hat{\sigma}_y^2$, δηλ. $\hat{\sigma}^2 = \frac{1}{(N-K)}\sum_{i=1}^{N}\hat{\varepsilon}_i^2$ και $\hat{\sigma}_y^2 = \frac{1}{(N-1)}\sum_{i=1}^{N}(y_i - \bar{y})^2$, ως

$$\begin{aligned} \bar{R}^2 = 1 - \frac{\hat{\sigma}^2}{\hat{\sigma}_y^2} &= 1 - \frac{\frac{1}{(N-K)}\sum_{i=1}^{N}\hat{\varepsilon}_i^2}{\frac{1}{(N-1)}\sum_{i=1}^{N}(y_i - \bar{y})^2} \\ &= 1 - \frac{RSS/(N-K)}{TSS/(N-1)} = 1 - \frac{(N-1)}{(N-K)}(1-R^2) \end{aligned} \qquad (10)$$

και τότε αναφέρεται ως προσαρμοσμένος (διορθωμένος) συντελεστής R^2.

Σημειώστε ότι

(i) Αν K = 1, τότε $\bar{R}^2 = R^2$.

(ii) Αν N είναι πολύ μεγάλο και K παραμένει σταθερό, τότε έχουμε

$$\frac{N-1}{N-K} \approx 1, \quad \text{και έτσι} \quad R^2 = \bar{R}^2.$$

(iii) Για πολύ μικρές τιμές του R^2, ο διορθωμένος συντελεστής προσδιορισμού $\bar{R}^2$ μπορεί να πάρει αρνητικές τιμές, καθώς N-1 είναι μεγαλύτερο του N-K.

Έλεγχος στατιστικής σημαντικότητας του υποδείγματος με βάση το συντελεστή R^2

Αν ισχύει η από κοινού υπόθεση $\beta_2=\beta_2=\ldots\beta_K=0$, τότε ο λόγος $\frac{ESS}{RSS}$ δεν θα πρέπει να είναι διάφορος από το μηδέν σε επίπεδο στατιστικής σημαντικότητας α. Διαιρώντας και

τον αριθμητή και τον παρονομαστή με K-1 και N-K αντίστοιχα, δίδει το ακόλουθο στατιστικό κριτήριο για τον έλεγχο της παραπάνω υπόθεσης:

$$F = \frac{ESS/(K-1)}{RSS/(N-K)} = \frac{\sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2/(K-1)}{\sum_{i=1}^{N}\hat{\varepsilon}_i^{\,2}/(N-K)}$$

ή ως

$$F = \frac{\frac{ESS/(K-1)}{TSS}}{\frac{RSS/(N-K)}{TSS}} = \frac{R^2/(K-1)}{(1-R^2)/(N-K)}, \quad \text{καθώς } R^2 = \frac{ESS}{TSS}.$$

Αν ο διαταρακτικός όρος ακολουθεί τις κλασικές υποθέσεις και κατανέμεται κανονικά, το κριτήριο αυτό κατανέμεται ως

$$F \sim F_{(K-1,N-K)}$$