ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

Ανάλυση Δεδομένων στη Λογιστική και Χρηματοοικονομική

Ενότητα 5η

Το Πολλαπλό Γραμμικό Υπόδειγμα Παλινδρόμησης

Περιγραφή Ενότητας

1. Παρουσίαση του Μοντέλου
2. Οι βασικές υποθέσεις του πολλαπλού γραμμικού υποδείγματος
3. Οι Ιδιότητες του εκτιμητή
4. Έλεγχοι Πολλαπλών Υποθέσεων
5. Καλή προσαρμογή της γραμμής παλινδρόμησης του δείγματος

Παρουσίαση του Πολλαπλόυ Γραμμικού Μοντέλου Παλινδρόμησης

- Μέχρι τώρα είχαμε δει το απλό γραμμικό υπόδειγμα παλινδρόμησης

$$y_t = \alpha + \beta x_t + u_t \quad t = 1,2,...,T$$

- Τι γίνεται όμως αν η εξαρτημένη (y) μεταβλητή μας εξαρτάται από περισσότερες από μία ανεξάρτητες μεταβλητές;

  Για παράδειγμα, ο αριθμός των αυτοκινήτων που πωλούνται μπορεί να εξαρτάται εύλογα από

  1. την τιμή των αυτοκινήτων.
  2. την τιμή των μέσων μαζικής μεταφοράς.
  3. την τιμή της βενζίνης.
  4. την έκταση της ανησυχίας του κοινού για την υπερθέρμανση του πλανήτη.

- Παρομοίως οι αποδόσεις μετοχών εξαρτώνται από διάφορους παράγοντες.

  ❖ Το να έχουμε μόνο μία ανεξάρτητη μεταβλητή δεν είναι καλό σε αυτή την περίπτωση - θέλουμε να έχουμε περισσότερες από μία μεταβλητές x. Είναι πολύ εύκολο να γενικεύσουμε το απλό μοντέλο σε ένα μοντέλο με περισσότερες ανεξάρτητες μεταβλητές.

Α. Το Πολλαπλό Γραμμικό Υπόδειγμα Παλινδρόμησης και η σταθερά

- Το πολλαπλό γραμμικό υπόδειγμα παλινδρόμησης είναι το ακόλουθο

$$y_t = \beta_1 + \beta_2 x_{2t} + \beta_3 x_{3t} + \ldots + \beta_k x_{kt} + u_t \ , \ t=1,2,\ldots,T$$

- Που είναι η x_1; Είναι ο σταθερός όρος β_1 . Στην πραγματικότητα ο σταθερός όρος αντιπροσωπεύεται από μια μοναδιαία στήλη με μήκος T:

$$x_1 = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$$

Β. Διαφορετικοί τρόποι έκφρασης του πολλαπλού μοντέλου γραμμικής παλινδρόμησης

- Θα μπορούσαμε να γράψουμε μια ξεχωριστή εξίσωση για κάθε τιμή του t

$$y_1 = \beta_1 + \beta_2 x_{21} + \beta_3 x_{31} + \ldots + \beta_k x_{k1} + u_1$$
$$y_2 = \beta_1 + \beta_2 x_{22} + \beta_3 x_{32} + \ldots + \beta_k x_{k2} + u_2$$
$$\vdots \qquad\qquad \vdots \qquad\qquad \vdots$$
$$y_T = \beta_1 + \beta_2 x_{2T} + \beta_3 x_{3T} + \ldots + \beta_k x_{kT} + u_T$$

- Μπορούμε να γράψουμε το παραπάνω σύστημα εξισώσεων σε μορφή μητρών

$$y = X\beta + u$$

όπου y: $T \times 1$ (διάνυσμα στήλης)
X: $T \times k$ (μήτρα)
β : $k \times 1$ (διάνυσμα στήλης)
u: $T \times 1$ (διάνυσμα στήλης)

Γ. «Μέσα» στους πίνακες του πολλαπλού μοντέλου γραμμικής παλινδρόμησης

- π.χ. εάν k=2, τότε:

$$\underset{T\times 1}{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_T \end{bmatrix}} = \underset{T\times 2}{\begin{bmatrix} 1 & x_{21} \\ 1 & x_{22} \\ \vdots & \vdots \\ 1 & x_{2T} \end{bmatrix}} \underset{2\times 1}{\begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}} + \underset{T\times 1}{\begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_T \end{bmatrix}}$$

Δ. Πώς υπολογίζουμε τις παραμέτρους (β) σε αυτήν τη γενικευμένη περίπτωση;

- Στο απλό γραμμικό μοντέλο παλινδρόμησης, παίρναμε το άθροισμα των τετραγώνων των καταλοίπων και το ελαχιστοποιούσαμε. Σε μορφή μητρών θα έχουμε:

$$\hat{\boldsymbol{u}} = \begin{bmatrix} \hat{\boldsymbol{u}}_1 \\ \hat{\boldsymbol{u}}_2 \\ \vdots \\ \hat{\boldsymbol{u}}_T \end{bmatrix}$$

- Το RSS (Residual Sum Squared) θα δίνεται από $\hat{u}'\hat{u} = [\hat{u}_1 \quad \hat{u}_2 \quad \dots \quad \hat{u}_T] \begin{bmatrix} \hat{u}_1 \\ \hat{u}_2 \\ \\ \hat{u}_T \end{bmatrix} = \hat{u}_1^2 + \hat{u}_2^2 + \dots + \hat{u}_T^2 = \sum \hat{u}_t^2$

Ε. Ο εκτιμητής β για το πολλαπλό γραμμικό μοντέλο παλινδρόμησης

- Στη συνέχεια ελαχιστοποιούμε το RSS σε σχέση με όλα τα β και με αυτόν τον τρόπο εκτιμούμε τα β_1, β_2,..., β_k,

- Αποδυκνύεται ότι:

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_k \end{bmatrix} = (X'X)^{-1} X' y$$

Οι βασικές υποθέσεις του πολλαπλού γραμμικού υποδείγματος

Η πολλαπλή γραμμική παλινδρόμηση έχει τις ίδιες βασικές υποθέσεις με το απλό γραμμικό υπόδειγμα παλινδρόμησης και μερικές πρόσθετες.

1. Η σχέση μεταξύ της εξαρτημένης μεταβλητής, Y, και των ανεξάρτητων (ερμηνευτικών) μεταβλητών (X_1, X_2, . . . , X_k) είναι γραμμική.
2. Οι ανεξάρτητες μεταβλητές (X_1, X_2, . . . , X_k) δεν είναι τυχαίες. Επίσης, δεν υπάρχει ακριβής γραμμική σχέση μεταξύ δύο ή περισσοτέρων ανεξάρτητων (ερμηνευτικών) μεταβλητών. Η υπόθεση αυτή αποκλείει την ύπαρξη πολυσυγγραμικότητας μεταξύ των ανεξάρτητων μεταβλητών
3. Η τυχαία μεταβλητή u_t (διαταρακτικός όρος υποδείγματος πολλαπλής παλινδρόμησης του πληθυσμού) είναι τυχαία μεταβλητή με μέσο το μηδέν: $E(u_t) = 0$
4. Η διακύμανση της τυχαίας μεταβλητής u_t είναι σταθερή (ομοσκεδαστικός όρος): $\text{var}(u_t) = \sigma^2$
5. Δεν υπάρχει αυτοσυσχέτιση (autocorrelation) στους διαταρακτικούς όρους, δηλαδή οι τιμές των διαταρακτικών όρων είναι ανεξάρτητες: $\text{cov}(u_i, u_j) = 0,\ j \neq i.$
6. Ο διαταρακτικός όρος δε συσχετίζεται με τις ανεξάρτητες μεταβλητές: $\text{cov}(u_t, X_t) = 0$
7. Η τυχαία μεταβλητή u_t ακολουθεί την κανονική κατανομή με μέσο μηδέν και σταθερή διακύμανση $u_t \sim N(0, \sigma^2)$

- **Παρατήρηση:**

Ο αριθμός των παρατηρήσεων του δείγματος πρέπει να είναι μεγαλύτερος από τον αριθμό των συντελεστών του υποδείγματος.

- **Αβεβαιότητα στη γραμμική παλινδρόμηση**

Υπάρχουν δύο πηγές αβεβαιότητας στα μοντέλα γραμμικής παλινδρόμησης:

1. Αβεβαιότητα που σχετίζεται με τον **διαταρακτικό όρο.**

   Ο ίδιος ο διαταρακτικός όρος περιέχει αβεβαιότητα, η οποία μπορεί να εκτιμηθεί από το τυπικό σφάλμα (standard error) της εκτίμησης για την εξίσωση παλινδρόμησης.

2. Αβεβαιότητα που σχετίζεται με τις **εκτιμήσεις παραμέτρων.**

   Οι εκτιμώμενες παράμετροι περιέχουν επίσης αβεβαιότητα επειδή είναι μόνο εκτιμήσεις των πραγματικών παραμέτρων του πληθυσμού.

Οι Ιδιότητες του εκτιμητή $\hat{\beta}$

Οι Ιδιότητες του εκτιμητή $\hat{\beta}$

α. Ο εκτιμητής $\hat{\beta}$ είναι γραμμική συνάρτηση των τιμών της εξαρτημένης μεταβλητής $\boldsymbol{Y}$.

β. Ο εκτιμητής $\hat{\beta}$ είναι αμερόληπτος, δηλαδή η μέση τιμή του ισούται με την μέση τιμή του συντελεστή β του πληθυσμού.

γ. Ο εκτιμητής $\hat{\beta}$ είναι συνεπής, δηλαδή όταν το μέγεθος του δείγματος τείνει στο άπειρο, τότε το $\hat{\beta}$ τείνει στο β του πληθυσμού.

δ. Ο εκτιμητής $\hat{\beta}$ είναι αποτελεσματικός, δηλαδή έχει την μικρότερη διακύμανση από κάθε άλλο εκτιμητή.

ε. Θεώρημα των GAUSS-MARKOV: Ο εκτιμητής ελαχίστων τετραγώνων είναι ο καλύτερος γραμμικός αμερόληπτος εκτιμητής (**BLUE = Best Linear Unbiased Estimator**).

- Δεδομένου ότι ισχύουν οι υποθέσεις του πολλαπλού γραμμικού υποδείγματος της παλινδρόμησης, οι εκτιμητές που προκύπτουν από τη μέθοδο ελαχίστων τετραγώνων είναι οι καλύτεροι γραμμικοί αμερόληπτοι εκτιμητές. Το τυπικό σφάλμα του εκτιμητή $\hat{\beta}$ (standard error of the estimate) ή το τυπικό σφάλμα της παλινδρόμησης ισούται με την εκτίμηση της τετραγωνικής ρίζας της διακύμανσης του διαταρακτικού όρου: $s^2 = \frac{u'u}{T-k}$

Παρατηρήσεις

- Ελέγξτε τις διαστάσεις: $\hat{\beta} : k \times 1$.

- Θυμηθείτε ότι στο απλό γραμμικό υπόδειγμα παλινδρόμησης, για να εκτιμήσουμε τη διακύμανση των σφαλμάτων, σ^2, χρησιμοποιήσαμε τον ακόλουθο τύπο:

$$s^2 = \frac{\sum \hat{u}_t^2}{T-2}$$

- Στο πολλαπλό γραμμικό υπόδειγμα παλινδρόμησης χρησιμοποιούμε μήτρες $s^2 = \frac{\hat{u}'\hat{u}}{T-k}$, όπου k = αριθμός των ανεξάρτητων μεταβλητών.

Μπορεί να αποδειχτεί ότι ο εκτιμητής ελαχίστων τετραγώνων της διακύμανσης του $\hat{\beta}$ δίνεται από τα **στοιχεία της διαγωνίου του** $s^2(X'X)^{-1}$,έτσι ώστε η διακύμανση του $\hat{\beta_1}$ είναι το πρώτο στοιχείο, η διακύμανση του $\hat{\beta_2}$ είναι το δεύτερο στοιχείο και …, η διακύμανση του $\hat{\beta_k}$ είναι το k στοπιχείο της διαγωνίου.

Εμπειρικό παράδειγμα: Το ακόλουθο μοντέλο με k=3 εκτιμάται βασιζόμενο σε 15 παρατηρήσεις:

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + u$$

και τα ακόλουθα δεδομένα υπολογίστηκαν απο την X.

$$(X'X)^{-1} = \begin{bmatrix} 2.0 & 3.5 & -1.0 \\ 3.5 & 1.0 & 6.5 \\ -1.0 & 6.5 & 4.3 \end{bmatrix}, (X'y) = \begin{bmatrix} -3.0 \\ 2.2 \\ 0.6 \end{bmatrix}, \hat{u}'\hat{u} = 10.96$$

Υπολογίστε τις εκτιμήσεις των συντελεστών και τα τυπικά σφάλματά τους.

Λύση

Για να υπολογίσετε τους συντελεστές, απλώς πολλαπλασιάστε τη μήτρα με το διάνυσμα για να λάβετε $(X'X)^{-1}X'y$.

Για να υπολογίσουμε τα τυπικά σφάλματα, χρειαζόμαστε μια εκτίμηση του σ^2.

$$s^2 = \frac{RSS}{T-k} = \frac{10.96}{15-3} = 0.91$$

Ο πίνακας διακύμανσης-συνδιακύμανσης του $\hat{\beta}$ δίνεται από:

$$s^2(X'X)^{-1} = 0.91(X'X)^{-1} = \begin{bmatrix} 1.83 & 3.20 & -0.91 \\ 3.20 & 0.91 & 5.94 \\ -0.91 & 5.94 & 3.93 \end{bmatrix}$$

- Οι διακυμάνσεις βρίσκονται στη διαγώνιο:

$$Var(\hat{\beta}_1) = 1.83 \qquad SE(\hat{\beta}_1) = 1.35$$
$$Var(\hat{\beta}_2) = 0.91 \Leftrightarrow \; SE(\hat{\beta}_2) = 0.96$$
$$Var(\hat{\beta}_3) = 3.93 \qquad SE(\hat{\beta}_3) = 1.98$$

- Άρα μπορούμε να γράψουμε

$$\hat{y} = 1.10 - 4.40x_{2t} + 19.88x_{3t}$$
$$(1.35) \quad (0.96) \quad (1.98)$$

Διάστημα εμπιστοσύνης των συντελεστών του πολλαπλού γραμμικού υποδείγματος

Σε προηγούμενο μάθημα δείξαμε πως μπορούμε να κατασκευάσουμε διαστήματα εμπιστοσύνης για τους συντελεστές (παραμέτρους) του πληθυσμού με βάση τις ιδιότητες των εκτιμητών ελαχίστων τετραγώνων. Ομοίως και στο πολλαπλό υπόδειγμα μπορούμε να κατασκευάσουμε διαστήματα εμπιστοσύνης :

$$\widehat{\beta_j} - t_{T-k,\frac{a}{2}} \times se(\widehat{\beta_j}) \leq \beta_j \leq \widehat{\beta_j} + t_{T-k,\frac{a}{2}} \times se\left(\widehat{\beta_j}\right), j = 1,2 \ldots . k$$

- $se(\widehat{\beta_j})$το τυπικό σφάλμα του εκτιμητή $\widehat{\beta_j}$
- $t_{T-k,\frac{a}{2}}$ η κριτική τιμή (τιμή των πινάκων) από κατανομή student με $T-k$ βαθμού ελευθερίας
- a το επίπεδο σημαντικότητας
- k το πλήθος των ανεξάρτητων μεταβλητών συμπεριλαμβανομένου του σταθερού όρου

Έλεγχος κοινών υποθέσεων

Η *F* στατιστική (*F*-test)

- Χρησιμοποιήσαμε το t-test για να ελέγξουμε μεμονωμένες υποθέσεις, δηλαδή υποθέσεις που περιλαμβάνουν μόνο έναν συντελεστή. Τι γίνεται όμως αν θέλουμε να ελέγξουμε περισσότερους από έναν συντελεστές ταυτόχρονα;

- Ο έλεγχος στην περίπτωση αυτή γίνεται με το *F*-test. Το *F*-test περιλαμβάνει την εκτίμηση δύο υποδειγμάτων παλινδρομήσης.

- Το «χωρίς περιορισμούς» υπόδειγμα παλινδρόμησης (unrestricted regression) είναι αυτό στο οποίο οι συντελεστές καθορίζονται ελεύθερα από τα δεδομένα.

- Το «με περιορισμούς» υπόδειγμα παλινδρόμησης (restricted regression) είναι αυτό στο οποίο οι συντελεστές είναι περιορισμένοι , δηλαδή οι περιορισμοί επιβάλονται σε κάποια β.

- **Παράδειγμα**

  Η γενική παλινδρόμηση είναι η

$$y_t = \beta_1 + \beta_2 x_{2t} + \beta_3 x_{3t} + \beta_4 x_{4t} + u_t$$

- Θέλουμε να ελέγξουμε τον περιορισμό ότι $\beta_3+\beta_4 = 1$ (έχουμε κάποια υπόθεση από τη θεωρία που υποδηλώνει ότι αυτή θα ήταν μια ενδιαφέρουσα υπόθεση για μελέτη). Το «χωρίς περιορισμούς» υπόδειγμα παλινδρόμησης (unrestricted regression) είναι η παράπάνω (γενική) παλινδρόμηση, άλλά ποιο είναι το «με περιορισμούς» υπόδειγμα παλινδρόμησης (restricted regression)

$$y_t = \beta_1 + \beta_2 x_{2t} + \beta_3 x_{3t} + \beta_4 x_{4t} + u_t \quad s.t.\ \beta_3+\beta_4 = 1$$

- Αντικαθιστούμε τον περιορισμό $\beta_3+\beta_4 = 1$ στην παλινδρόμηση έτσι ώστε να επιβληθεί αυτόματα στα δεδομένα.

$$\beta_3+\beta_4 = 1 \Rightarrow \beta_4 = 1- \beta_3$$

- Άρα θα έχουμε την ακόλουθη μορφή στην «με περιορισμούς» παλινδρόμηση (restricted regression):

$$y_t = \beta_1 + \beta_2 x_{2t} + \beta_3 x_{3t} + (1-\beta_3) x_{4t} + u_t$$

$$y_t = \beta_1 + \beta_2 x_{2t} + \beta_3 x_{3t} + x_{4t} - \beta_3 x_{4t} + u_t$$

- Συγκεντρώστε όρους στα β μαζί και αναδιατάξτε

$$(y_t - x_{4t}) = \beta_1 + \beta_2 x_{2t} + \beta_3 (x_{3t} - x_{4t}) + u_t$$

- Το παραπάνω είναι το «με περιορισμούς» υπόδειγμα παλινδρόμησης (restricted regression). Στην πραγματικότητα το εκτιμάμε δημιουργώντας δύο νέες μεταβλητές, ονομάστε τις, έστω P_t and $Q_{t,}$ όπου

$$P_t = y_t - x_{4t}$$

$$Q_t = x_{3t} - x_{4t} \qquad \text{έτσι}$$

$P_t = \beta_1 + \beta_2 x_{2t} + \beta_3 Q_t + u_t$ είναι το «με περιορισμούς» υπόδειγμα παλινδρόμησης (restricted regression) που ουσιαστικά εκτιμούμε.

Υπολογισμός της στατιστικής F (F-Test Statistic)

- Η F-statistic για ελέγχους πολλαπλών υποθέσεων των συντελεστών δίνεται από:

$$t-stat = \frac{RRSS-URSS}{URSS} \times \frac{T-k}{m}$$

όπου *URSS* = RSS από το «χωρίς περιορισμούς» υπόδειγμα παλινδρόμησης (unrestricted regression)

RRSS = RSS από το «με περιορισμούς» υπόδειγμα παλινδρόμησης (unrestricted regression)

m = αριθμός των περιορισμών

T = αριθμός των παρατηρήσεων

k = αριθμός των ανεξάρτητων μεταβλητών στο «χωρίς περιορισμούς» υπόδειγμα παλινδρόμησης (unrestricted regression) συμπεριλαμβανομένης της σταθεράς (ή ο συνολικός αριθμός παραμέτρων που πρέπει να εκτιμηθούν).

Η κατανομή της *F*

- Η *t-stat* ακολουθεί την F-κατανομή, η οποία έχει 2 βαθμούς ελευθερίας.
- Η τιμή των βαθμών ελευθερίας είναι *m* και *(T-k)* αντίστοιχα.
- Η κατάλληλη κριτική τιμή θα βρίσκεται στη στήλη *m*, σειρά (*T-k*).
- Η κατανομή F έχει μόνο θετικές τιμές και δεν είναι συμμετρική. Επομένως απορρίπτουμε τη μηδενική υπόθεση μόνο εάν το test statistic > κριτική τιμή F (test statistic > critical *F*-value.)

Προσδιορισμός του αριθμού των περιορισμών σε ένα *F*-test

- Παραδείγματα :

H_0: Μηδενική υπόθεση	Αριθμός περιορισμών, *m*
$\beta_1 + \beta_2 = 2$	1
$\beta_2 = 1$ και $\beta_3 = -1$	2
$\beta_2 = 0$, $\beta_3 = 0$ και $\beta_4 = 0$	3

- Εάν το μοντέλο είναι το ακόλουθο $y_t = \beta_1 + \beta_2 x_{2t} + \beta_3 x_{3t} + \beta_4 x_{4t} + u_t$,

  τότε η μηδενική υπόθεση H_0: $\beta_2 = 0$, και $\beta_3 = 0$ και $\beta_4 = 0$ ελέγχεται από την F-στατιστική. Ελέγχει την μηδενική υπόθεση ότι όλοι οι συντελεστές εκτός από τη σταθερά είναι μηδενικοί.
- Σημειώστε τη μορφή της εναλλακτικής υπόθεσης για όλους τους ελέγχους όταν περιλαμβάνονται περισσότεροι από ένας περιορισμοί: H_1: $\beta_2 \neq 0$, ή $\beta_3 \neq 0$ ή $\beta_4 \neq 0$

Τι δεν μπορούμε να ελέγξουμε είτε με F-test είτε με t-test;

Δεν μπορούμε να ελέγξουμε τις υποθέσεις που οι σχέσεις των συντελεστών δεν είναι γραμμικές ή που είναι πολλαπλασιαστικές,

π.χ. H_0: $\beta_2 \beta_3 = 2$ or H_0: $\beta_2^{\,2} = 1$

Η σχέση μεταξύ των *t* and the *F* κατανομών

- Οποιαδήποτε υπόθεση που θα μπορούσε να ελεγχθεί με t-test θα μπορούσε να ελεγχθεί χρησιμοποιώντας την F-test, αλλά όχι το αντίστροφο.

  Για παράδειγμα, λάβετε υπόψη την υπόθεση

  H_0: $\beta_2 = 0.5$

  H_1: $\beta_2 \neq 0.5$

  Θα μπορούσαμε να το ελέγξουμε χρησιμοποιώντας το συνηθισμένο t-test: $test\ stat = \dfrac{\hat{\beta}_2 - 0.5}{SE(\hat{\beta}_2)}$

  ή θα μπορούσε να ελεγχθεί με το *F*-test.

- Σημειώστε ότι οι δύο έλεγχοι δίνουν πάντα το ίδιο αποτέλεσμα αφού η κατανομή t είναι απλώς μια ειδική περίπτωση της κατανομής F.

- Για παράδειγμα, αν έχουμε κάποια τυχαία μεταβλητή Z, και $Z \sim t(T\text{-}k)$ τότε επίσης $Z^2 \sim F(1,T\text{-}k)$

F-test - Παράδειγμα

- Ερώτηση: Ας υποθέσουμε ότι ένας ερευνητής θέλει να ελέγξει αν οι αποδόσεις των μετοχών μιας εταιρείας (y) δείχνουν μοναδιαία ευαισθησία σε δύο παράγοντες (παράγοντας x_2 και παράγοντας x_3) μεταξύ τριών που εξετάστηκαν. Η παλινδρόμηση πραγματοποιείται χρησιμοποιώντας 144 μηνιαίες παρατηρήσεις. Η παλινδρόμηση είναι η εξής:

$$y_t = \beta_1 + \beta_2 x_{2t} + \beta_3 x_{3t} + \beta_4 x_{4t} + u_t$$

- Ποιες είναι οι «χωρίς περιορισμούς» παλινδρόμηση (unrestricted regression) και «με περιορισμούς» παλινδρόμηση (restricted regression);

- Εάν οι RSS είναι 436.1 and 397.2 αντίστοιχα, κάνετε τον έλεγχο.

- Λύση:

Η μοναδιαία ευαισθησία υποδηλώνει H_0:β_2=1 και β_3=1. Η «χωρίς περιορισμούς» παλινδόμηση (unrestricted regression) είναι αυτή στην ερώτηση. Η «με περιορισμούς» παλινδόμηση (restricted regression) είναι

$(y_t - x_{2t} - x_{3t}) = \beta_1 + \beta_4 x_{4t} + u_t$ ή θέτωντας

$z_t = y_t - x_{2t} - x_{3t}$, η «με περιορισμούς» παλινδόμηση (restricted regression) είναι $z_t = \beta_1 + \beta_4 x_{4t} + u_t$

Στο F-test, T=144, k=4, m=2, $RRSS$=436.1, $URSS$=397.2

F-test statistic = 6.68. Η κριτική τιμή $F(2,140)$ = 3.07 (5%) and 4.79 (1%).

Συμπέρασμα: Απορρίπτουμε τη μηδενική υπόθεση H_0.

Η Καλή προσαρμογή της γραμμής παλινδρόμησης του δείγματος

- Θα θέλαμε κάποιο μέτρο για το πόσο καλά το μοντέλο παλινδρόμησης πραγματικά αντιπροσωπεύει τα δεδομένα.
- Υπάρχουν διαθέσιμες στατιστικές για να ελέγξουμε πόσο καλά η συνάρτηση παλινδρόμησης δείγματος αντιπροσωπεύει τα δεδομένα.
- Ένα σημαντικό στατιστικό μέτρο είναι ο συντελεστής προσδιορισμού R^2. Ένας τρόπος να ορίσουμε τον R^2 είναι να υπολογίσουμε το τετράγωνο του συντελεστή συσχέτισης μεταξύ της y and $\hat{y}$.
- Μια άλλη εξήγηση: Θυμηθείτε ότι ενδιαφερόμαστε να εξηγήσουμε την μεταβλητότητα της y σε σχέση με τη μέση τιμή της, $\bar{y}$, δηλαδή τη συνολική μεταβλητότητα ή συνολικό άθροισμα τετραγώνων, TSS:

$$TSS = \sum_t \left(y_t - \bar{y}\right)^2$$

- Επίσης το TSS χωρίζεται σε δύο όρους:

$$TSS = ESS + RSS$$

$$\sum_t (y_t - \bar{y})^2 = \sum_t (\hat{y}_t - \bar{y})^2 + \sum_t \hat{u}_t^2$$

TSS: συνολική μεταβλητότητα ή το συνολικό άθροισμα τετραγώνων.
ESS: μεταβλητότητα που ερμηνεύεται από την παλινδρόμηση ή το άθροισμα τετραγώνων της παλινδρόμησης.
RSS: μεταβλητότητα που δεν ερμηνεύεται από την παλινδρόμηση ή το άθροισμα τετραγώνων των καταλοίπων.

Ο συντελεστής προσδιορισμού R^2

- Ο συντελεστής προσδιορισμού ορίζεται ως

$$R^2 = \frac{ESS}{TSS}$$

Ο συντελεστής προσδιορισμού αποτελεί ένα μέτρο της ικανότητας προσαρμογής του υποδείγματος και ορίζεται ως η αναλογία της μεταβλητότητας της εξαρτημένης μεταβλητής *y* που ερμηνεύεται από την παλινδρόμηση προς την συνολική της μεταβλητότητα.

- Εφόσον $TSS = ESS + RSS$, μπορούμε επίσης να γράψουμε

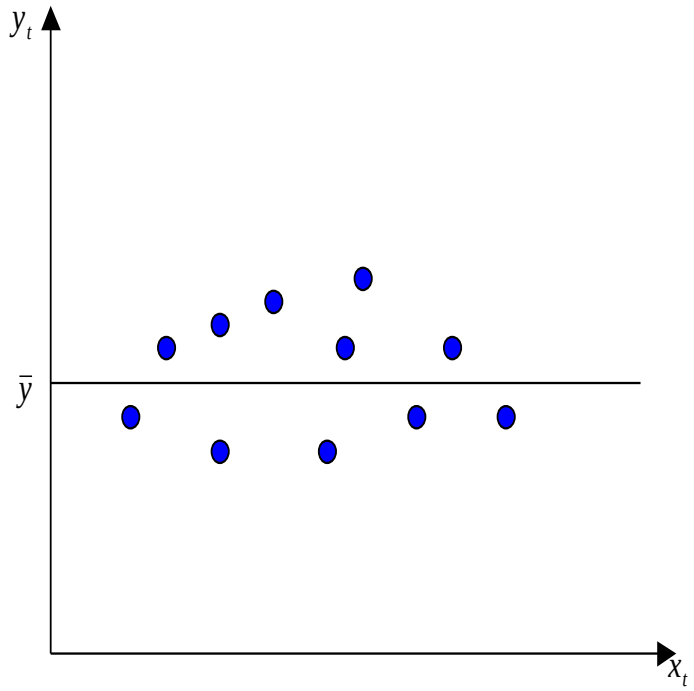
$$R^2 = \frac{ESS}{TSS} = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

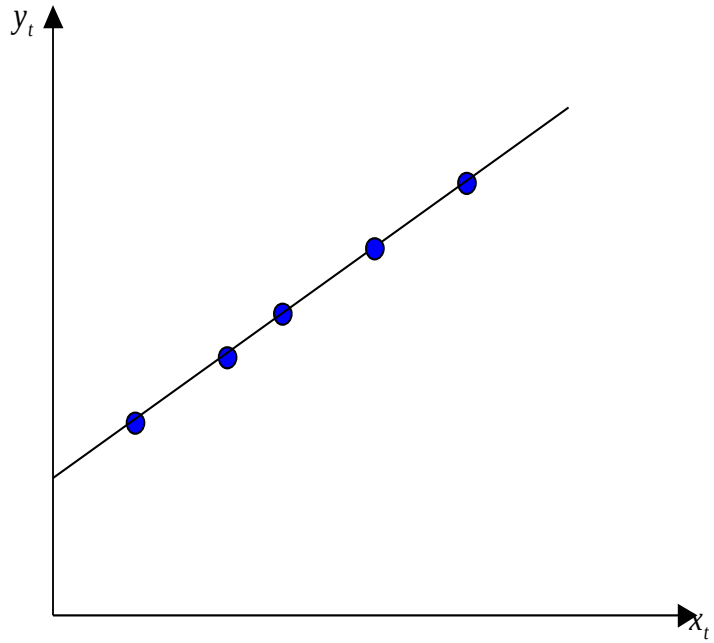
- *Ο* R^2 πρέπει να βρίσκεται πάντοτε μεταξύ του μηδενός και της μονάδας. Για περαιτέρω κατανόηση, θεωρείστε τις δύο παρακάτω ακραίες περιπτώσεις:

Όταν $RSS = TSS$ *τότε* $ESS = 0$, *άρα* $R^2 = ESS/TSS = 0$

Όταν $ESS = TSS$ *τότε* $RSS = 0$, *άρα* $R^2 = ESS/TSS = 1$

Οι περιπτώσεις του συντελεστή προσδιορισμού R^2 : R^2 = 0 and R^2 = 1





- **Προβλήματα του συντελεστή προσδιορισμού R^2 ως προς την καλή προσαρμογή της γραμμής παλινδρόμησης του δείγματος**

1. Ο R^2 ορίζεται ως προς τη διακύμανση του μέσου όρου της y έτσι ώστε εάν ένα μοντέλο επαναπαραμετροποιηθεί (αναδιαταχθεί) και η εξαρτημένη μεταβλητή αλλάξει, ο R^2 θα αλλάξει.

2. Ο R^2 δεν πέφτει ποτέ αν προστεθούν περισσότερες εξαρτημένες μεταβλητές στην παλινδρόμηση, π.χ.

Υπόδειγμα Παλινδρόμησης 1: $y_t = \beta_1 + \beta_2 x_{2t} + \beta_3 x_{3t} + u_t$

Υπόδειγμα Παλινδρόμησης 2: $y_t = \beta_1 + \beta_2 x_{2t} + \beta_3 x_{3t} + \beta_4 x_{4t} + u_t$

Ο R^2 θα είναι πάντα τουλάχιστον ο ίδιος υψηλός για την παλινδρόμηση 2 σε σχέση με την παλινδρόμηση 1.

3. Ο R^2 επηρεάζεται από το πλήθος των ανεξάρτητων μεταβλητών. Αν το δείγμα είναι μικρό και το πλήθος των ανεξάρτητων μεταβλητών μεγάλο, τότε η εκτίμηση του R^2 δεν είναι καλή.

- **Διορθωμένος (Προσαρμοσμένος) συντελεστής προσδιορισμού (Adjusted R^2)**

Για να ξεπεραστούν τα παραπάνω προβλήματα, γίνεται συχνά μια τροποποίηση η οποία λαμβάνει υπόψη την απώλεια βαθμών ελευθερίας που συνδέονται με την προσθήκη επιπλέον μεταβλητών. Αυτό είναι γνωστό ως $\overline{R}^2$ ή Διορθωμένος (Προσαρμοσμένος) συντελεστής προσδιοριμού (Adjusted R^2)

$$\overline{R}^2 = 1 - \left[\frac{T-1}{T-k} (1 - R^2) \right]$$

Παρατηρήσεις

1. Οι συντελεστές R^2 και $\overline{R}^2$ μετρούν την καλή προσαρμογή της γραμμής παλινδρόμησης του δείγματος.
2. Ο διορθωμένος συντελεστής $\overline{R}^2$ παίρνει τιμές μικρότερες από το R^2
3. Είναι $\mathbf{0} \leq R^2 \leq \mathbf{1}$, ενώ ο $\overline{R}^2$ παίρνει και αρνητικές τιμές.

Η σχέση μεταξύ της στατιστικής F και του συντελεστή πολλαπλού προσδιορισμού R^2

Υπάρχει μια ειδική σχέση μεταξύ του συντελεστή προσδιορισμού R^2 και την F-στατιστική. Θεωρείστε ότι έχουμε το παρακάτω υπόδειγμα της πολλαπλής παλινδρόμησης του πληθυσμού:

$y_t = \beta_1 + \beta_2 x_{2t} + \beta_3 x_{3t} + \ldots + \beta_k x_{kt} + u_t$, , t=1,2,...,T

Η (μηδενική) υπόθεση που θέλουμε είναι να ελέγξουμε οι συντελεστές (slope coefficients) του παραπάνω υποδείγματος να είναι ταυτόχρονα ίσοι με μηδέν: H_0: $\beta_2 = \beta_3 = \ldots \beta_k = 0$

Η εναλλακτική υπόθεση είναι τουλάχιστον ένας συντελεστής να είναι διάφορος του μηδενός H_1: $\beta_2 \neq 0$ ή $\beta_3 \neq 0 \ldots$ ή $\beta_k \neq 0$

Για τον υπολογισμό της F-statistic χρειαζόμαστε δύο υποδείγματα παλινδρόμησης, ένα «χωρίς περιορισμούς» υπόδειγμα (unrestricted regression) και ένα «με περιορισμούς» υπόδειγμα (restricted regression). Στην περίπτωση της παραπάνω υπόθεσης το «με περιορισμούς» υπόδειγμα παλινδρόμησης (restricted regression) περιλαμβάνει μόνο τον σταθερό όρο. Άρα το «με περιορισμούς» υπόδειγμα παλινδρόμησης είναι το ακόλουθο: $y_t = \beta_1 + u_t$

Θυμηθείτε ότι η F-στατιστική δίνεται από την στατιστική

$$t - stat = \frac{RRSS - URSS}{URSS} \times \frac{T - k}{m}$$

Αλλά στην περίπτωση $y_t = \beta_1 + u_t$, $RRSS = \text{TSS}$ γιατί $\text{ESS} = 0$, επειδή δεν υπάρχουν συντελέστες (slope parameters) (*TSS = ESS + RSS)*

Άρα η F-στατιστική μπορεί να γραφεί ως $F - stat = \frac{TSS - RSS}{RSS} \times \frac{T - k}{k - 1}$ ή

$$F - stat = \frac{ESS/TSS}{RSS/TSS} \times \frac{T - k}{k - 1}$$ και

$$F - stat = \frac{R^2}{1 - R^2} \times \frac{T - k}{k - 1}$$

- Αν $R^2 \to 0$ τότε $F \to 0$.
- Αν $R^2 \to 1$ τότε $F \to \infty$.

Κριτήρια επιλογής υποδειγμάτων της παλινδρόμησης

Ο συντελεστής προσδιορισμού R^2 μετράει την αναλογία της συνολικής μεταβλητότητας στην εξαρτημένη μεταβλητή που ερμηνεύεται από την παλινδρόμηση. Άρα όπως δείξαμε προηγουμένως ένας δείκτης που χρησιμοποιούμε για την σύγκριση υποδειγμάτων είναι το R^2 ή ο προσαρμοσμένος συντελεστής προσδιορισμού $\overline{R}^2$. Όταν όμως υπάρχουν δύο υποδείγματα των οποίων οι εξαρτημένες μεταβλητές τους αποτελούν διαφορετικές συναρτησιακές μορφές

Υπόδειγμα Παλινδρόμησης 1: $y_t = \beta_1 + \beta_2 x_{2t} + \beta_3 x_{3t} + u_t$

Υπόδειγμα Παλινδρόμησης 2: $\ln y_t = \beta_1 + \beta_2 \ln x_{2t} + \beta_3 \ln x_{3t} + u_t$

τότε δεν μπορούμε να συγκρίνουμε τα υποδείγματα με βάση το R^2 ή $\overline{R}^2$

Κριτήρια σύγκρισης υποδειγμάτων που λαμβάνουν υπόψιν το μέγεθος του δείγματος και τον αριθμό των ανεξάρτητων μεταβλητών:

1. Κριτήριο του Akaike $AIC = -\frac{2LL}{T} + \frac{2k}{T}$
2. Κριτήριο του Schwarz $BIC = -\frac{2LL}{T} + \frac{k \ln T}{T}$
3. Κριτήριο των Hannan και Quinn $HQ = -\frac{2LL}{T} + \frac{2k \ln(\ln T)}{T}$

όπου $LL = -\frac{T}{2}\left[1 + \ln(2\pi) + \ln\left(\frac{\sum u^2}{T}\right)\right]$

Με βάση τα παραπάνω κριτήρια, επιλέγουμε εκείνο το υπόδειγμα που έχει την **μικρότερη τιμή** στα κριτήρια αυτά. Οι τιμές των κριτηρίων μπορεί να είναι και αρνητικές. Όλα τα κριτήρια δεν υποδεικνύουν πάντα το ίδιο βέλτιστο υπόδειγμα.

Παράδειγμα - STATA

$$\left(R_{Microsoft} - r_f\right)_t = \gamma_0 + \gamma_1\left(R_M - r_f\right)_t + \varepsilon_t$$

$$\left(R_{Microsoft} - r_f\right)_t = \alpha + \beta_1\left(R_M - r_f\right)_t + \beta_2\left(d_{prod}\right)_t + \beta_3\left(d_{money}\right)_t + \beta_4\left(d_{\text{inflation}}\right)_t + \beta_5\left(d_{tspread}\right)_t + \beta_6\left(d_{dspread}\right)_t + u_t$$

```
. regress ermsoft ersand
```

Source	SS	df	MS		
Model	11968.8168	1	11968.8168	Number of obs	= 325
Residual	52117.0991	323	161.353248	F(1, 323)	= 74.18
				Prob > F	= 0.0000
				R-squared	= 0.1868
				Adj R-squared	= 0.1842
Total	64085.9159	324	197.796037	Root MSE	= 12.702

ermsoft	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
ersandp	1.325376	.1538871	8.61	0.000	1.022628	1.628124
_cons	-.6137005	.705782	-0.87	0.385	-2.002211	.7748094

```
.
```

```
. regress ermsoft ersand dprod dcredit dinflation dmoney dspread rterm
```

Source	SS	df	MS		
Model	13202.4359	7	1886.06227	Number of obs	= 324
Residual	50637.6544	316	160.245742	F(7, 316)	= 11.77
				Prob > F	= 0.0000
				R-squared	= 0.2068
				Adj R-squared	= 0.1892
Total	63840.0903	323	197.647338	Root MSE	= 12.659

ermsoft	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
ersandp	1.360448	.1566147	8.69	0.000	1.052308	1.668587
dprod	-1.425779	1.324467	-1.08	0.283	-4.031668	1.180109
dcredit	-.0000405	.0000764	-0.53	0.596	-.0001909	.0001098
dinflation	2.95991	2.166209	1.37	0.173	-1.302104	7.221925
dmoney	-.0110867	.0351754	-0.32	0.753	-.0802944	.0581209
dspread	5.366629	6.913915	0.78	0.438	-8.236496	18.96975
rterm	4.315813	2.515179	1.72	0.087	-.6327998	9.264426
_cons	-.1514086	.9047867	-0.17	0.867	-1.931576	1.628759