

Το Απλό Γραμμικό Μοντέλο
Μαθηματικά της Αγοράς και της Παραγωγής

Διδάσκουσα: Λουκία Μελιγκοτσίδου
Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών
Τμήμα Μαθηματικών

April 26, 2020

Γραμμική Παλινδρόμηση

Έστω δύο μεταβλητές X και Y . Μια συναρτησιακή σχέση μεταξύ των μεταβλητών της μορφής

$$Y = f(X)$$

είναι μια ντετερμινιστική σχέση (*deterministic relationship*), που σημαίνει ότι η τιμή της X καθορίζει πλήρως την τιμή της Y .

Για παράδειγμα, $Y = \beta_0 + \beta_1 X$, γραμμική σχέση.

Η στατιστική σχέση μεταξύ δυο μεταβλητών είναι της μορφής

$$Y = f(X) + \varepsilon,$$

όπου ε τυχαίος (στοχαστικός) όρος. Η σχέση αυτή είναι στοχαστική (*stochastic relationship*). Η τυχαία μεταβλητή Y εξαρτάται από την μεταβλητή X (η οποία έχει προκαθορισμένες τιμές), αλλά και από κάποιους μη μετρήσιμους παράγοντες που συνοψίζονται στον στοχαστικό όρο ε .

Για παράδειγμα, $Y = \beta_0 + \beta_1 X + \varepsilon$, απλή γραμμική παλινδρόμηση (*regression*) η απλό γραμμικό μοντέλο (*simple linear model*).

Έχοντας παρατηρήσει δείγμα ζευγών $(X_i, Y_i), i = 1, \dots, n$, για τα οποία υποθέτουμε ότι ακολουθούν το μοντέλο γραμμικής παλινδρόμησης,

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, \dots, n,$$

σκοπός μας είναι η εκτίμηση και γενικά η στατιστική συμπερασματολογία για τις άγνωστες παραμέτρους β_0, β_1 χρησιμοποιώντας το δείγμα (X_i, Y_i) .

Στο απλό γραμμικό μοντέλο η Y είναι η εξαρτημένη μεταβλητή (*dependent or response*) και η X είναι η ανεξάρτητη μεταβλητή (*independent or predictor*)

Τα ε_i ονομάζονται τυχαία σφάλματα.

Σύμφωνα με τη μοντελοποίηση, η Y είναι τ.μ. ενώ η X όχι.

Υποθέσεις για τα τυχαία σφάλματα:

- $E(\varepsilon_i) = 0$, σφάλματα με μηδενική μέση τιμή.
- $V(\varepsilon_i) = \sigma^2$, ομοσκεδαστικότητα (ίση διασπορά).
- $Cov(\varepsilon_i, \varepsilon_j) = 0$, ασυσχέτιστα τυχαία σφάλματα (το σφάλμα σε οποιαδήποτε δοκιμή δεν επηρεάζει τα σφάλματα άλλων δοκιμών).

Οι υποθέσεις για τους τυχαίους όρους οδηγούν σε υποθέσεις για τα Y_i . Έχουμε, λοιπόν,

- $E(Y_i) = \beta_0 + \beta_1 X_i$, $V(Y_i) = \sigma^2$, $Cov(Y_i, Y_j) = 0$

Η γραμμή παλινδρόμησης δίνει την αναμενόμενη τιμή της Y για κάθε τιμή της X .

Απλό γραμμικό μοντέλο

Απλό σημαίνει ότι υπάρχει μια μόνο ανεξάρτητη μεταβλητή.

Γραμμικό σημαίνει γραμμικό ως προς τις παραμέτρους.

Το υπόδειγμα $Y_i = \beta_0 + \beta_1 X_i^2 + \varepsilon_i$ είναι γραμμικό, ενώ το $Y_i = \beta_1^{\beta_1} X_i + \varepsilon_i$ όχι.

Ερμηνεία των Παραμέτρων της Παλινδρόμησης

β_0 : είναι το σημείο όπου η ευθεία τέμνει τον άξονα των Y , δηλαδή αντιστοιχεί στην αναμενόμενη τιμή του Y για $X = 0$

β_1 : είναι η κλίση της ευθείας και αντιπροσωπεύει την μεταβολή (αύξηση ή μείωση) στην αναμενόμενη τιμή του Y που αντιστοιχεί σε αύξηση του X κατά μια μονάδα.

Εκτίμηση παραμέτρων με τη Μέθοδο Ελαχίστων Τετραγώνων

Η Μ.Ε.Τ. στοχεύει στον προσδιορισμό της γραμμής παλινδρόμησης έτσι ώστε να ελαχιστοποιηθούν συνολικά οι αποκλίσεις των σημείων (που αντιστοιχούν στα ζεύγη (X_i, Y_i)) από την ευθεία (ελαχιστοποίηση των σφαλμάτων).

Έχουμε $\varepsilon_i = Y_i - E(Y_i) = Y_i - (\beta_0 + \beta_1 X_i)$.
Επειδή $E(\varepsilon_i) = 0$ δεν εξετάζουμε την ποσότητα $\sum_{i=1}^n \varepsilon_i$ (η οποία θα είναι ίση με 0), αλλά παίρνουμε το άθροισμα των τετραγώνων

$$Q = \sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_i)^2.$$

Οι εκτιμήτριες των β_0, β_1 προκύπτουν από την ελαχιστοποίηση του Q .

$$\begin{cases} \frac{dQ}{d\beta_0} = -2 \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_i) = 0 \\ \frac{dQ}{d\beta_1} = -2 \sum_{i=1}^n X_i (Y_i - \beta_0 - \beta_1 X_i) = 0 \end{cases} \Rightarrow$$

$$\begin{cases} \sum_{i=1}^n Y_i = n\beta_0 + \beta_1 \sum_{i=1}^n X_i \\ \sum_{i=1}^n X_i Y_i = \beta_0 \sum_{i=1}^n X_i + \beta_1 \sum_{i=1}^n X_i^2 \end{cases} \quad \text{Κανονικές Εξισώσεις}$$

Λύνοντας ως προς β_0 και β_1 έχουμε

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n X_i Y_i - \frac{\sum_{i=1}^n X_i \sum_{i=1}^n Y_i}{n}}{\sum_{i=1}^n X_i^2 - \frac{(\sum_{i=1}^n X_i)^2}{n}} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

$$\hat{\beta}_0 = \frac{1}{n} \left[\sum_{i=1}^n Y_i - \hat{\beta}_1 \sum_{i=1}^n X_i \right] = \bar{Y} - \hat{\beta}_1 \bar{X}$$

Εναλλακτική μορφή του απλού γραμμικού μοντέλου

$$Y_i = \beta_0^* + \beta_1 (X_i - \bar{X}) + \varepsilon_i, \text{ όπου } \beta_0^* = \beta_0 + \beta_1 \bar{X}$$

$$\text{ή } Y_i = \beta_0^* + \beta_1 \tilde{X}_i + \varepsilon_i \text{ όπου } \tilde{X}_i = X_i - \bar{X}$$

Η εκτιμήτρια του β_1 είναι η ίδια με αυτή της αρχικής εκδοχής του απλού γραμμικού μοντέλου.

Για το β_0^* είναι: $\hat{\beta}_0^* = \hat{\beta}_0 + \hat{\beta}_1 \bar{X} = \bar{Y} - \hat{\beta}_1 \bar{X} + \hat{\beta}_1 \bar{X} = \bar{Y}$

Θεώρημα. Τα $\hat{\beta}_0$ και $\hat{\beta}_1$ είναι γραμμικοί συνδυασμοί των Y_i .

Απόδειξη. Θα δείξουμε ότι η $\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}$ μπορεί να γραφτεί ως

$$\hat{\beta}_1 = \sum k_i Y_i, \text{ όπου } k_i = \frac{X_i - \bar{X}}{\sum_{i=1}^n (X_i - \bar{X})^2}.$$

Και επειδή τα X_i είναι γνωστές σταθερές και τα k_i θα είναι γνωστές σταθερές και άρα το $\hat{\beta}_1$ είναι γραμμικός συνδυασμός των Y_i .

Έχουμε

$$\begin{aligned}\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) &= \sum_{i=1}^n (X_i - \bar{X})Y_i - \sum_{i=1}^n (X_i - \bar{X})\bar{Y} \\ &= \sum_{i=1}^n (X_i - \bar{X})Y_i - \bar{Y} \sum_{i=1}^n (X_i - \bar{X}) = \sum_{i=1}^n (X_i - \bar{X})Y_i\end{aligned}$$

$$\text{Άρα } \hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} = \sum k_i Y_i$$

Προφανώς και $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} = \frac{1}{n} \sum Y_i - \bar{X} \sum k_i Y_i$ είναι γραμμικός συνδυασμός των Y_i .

Ιδιότητες των ποσοτήτων k_i

- $\sum k_i = 0$, γιατί $\frac{\sum_{i=1}^n (X_i - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{0}{\sum_{i=1}^n (X_i - \bar{X})^2} = 0$
- $\sum k_i X_i = 1$, γιατί $\sum k_i X_i = \sum \frac{X_i - \bar{X}}{\sum_{i=1}^n (X_i - \bar{X})^2} X_i = \sum \frac{X_i - \bar{X}}{\sum_{i=1}^n (X_i - \bar{X})^2} X_i - \sum \frac{X_i - \bar{X}}{\sum_{i=1}^n (X_i - \bar{X})^2} \bar{X} = \frac{\sum (X_i - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2} = 1$
- $\sum k_i^2 = \frac{1}{\sum (X_i - \bar{X})^2}$,

$$\text{γιατί } \sum k_i^2 = \sum \left[\frac{X_i - \bar{X}}{\sum (X_i - \bar{X})^2} \right]^2 = \frac{\sum (X_i - \bar{X})^2}{(\sum (X_i - \bar{X})^2)^2} = \frac{1}{\sum (X_i - \bar{X})^2}$$

Θεώρημα των Gauss – Markov. Για το απλό γραμμικό μοντέλο οι εκτιμήτριες ελαχίστων τετραγώνων $\hat{\beta}_0, \hat{\beta}_1$

- 1) είναι αμερόληπτες
- 2) έχουν ελάχιστη διασπορά μεταξύ των αμερόληπτων εκτιμητριών που είναι γραμμικές συναρτήσεις των Y_i .

Απόδειξη

- Αμεροληψία της $\hat{\beta}_1$: Θέλουμε να δείξουμε ότι $E(\hat{\beta}_1) = \beta_1$.

$$E(\hat{\beta}_1) = E\left(\sum k_i Y_i\right) = \sum k_i E(Y_i) = \sum k_i (\beta_0 + \beta_1 X_i) = \beta_0 \underbrace{\sum k_i}_0 + \beta_1 \underbrace{\sum k_i X_i}_1 = \beta_1$$

- Έστω ότι όλες οι αμερόληπτες εκτιμήτριες του β_1 που είναι γραμμικές συναρτήσεις των Y_i είναι της μορφής

$$b_1 = \sum c_i Y_i,$$

όπου c_i αυθαίρετες σταθερές. Επειδή έχουμε αμεροληψία:

$$E(b_1) = \beta_1 \Rightarrow E\left(\sum c_i Y_i\right) = \sum c_i (\beta_0 + \beta_1 X_i) = \beta_0 \sum c_i + \beta_1 \sum c_i X_i = \beta_1.$$

Άρα πρέπει $\sum c_i = 0$ και $\sum c_i X_i = 1$.

Η διασπορά του b_1 είναι

$$V(b_1) = V\left(\sum c_i Y_i\right) = \sum c_i^2 V(Y_i) = \sum c_i^2 \sigma^2 = \sigma^2 \sum c_i^2,$$

αφού $Cov(Y_i, Y_j) = 0$.

Έστω ότι τα c_i έχουν τη μορφή $c_i = k_i + d_i$ όπου τα k_i είναι όπως ορίστηκαν στην εκτιμήτρια $\hat{\beta}_1 = \sum k_i Y_i$ και τα d_i είναι αυθαίρετες σταθερές.

Συνεπώς

$$\begin{aligned} V(b_1) &= \sigma^2 \sum c_i^2 = \sigma^2 \sum (k_i + d_i)^2 \\ &= \sigma^2 \left[\sum k_i^2 + \sum d_i^2 + 2 \sum k_i d_i \right] \end{aligned}$$

$$= \underbrace{\sigma^2 \sum k_i^2}_{V(\hat{\beta}_1)} + \sigma^2 \sum d_i^2 + 2\sigma^2 \sum k_i d_i$$

Έχουμε $\sum k_i = 0$ και $\sum c_i = \sum(k_i + d_i) = 0 \Rightarrow \sum d_i = 0$
 $\sum k_i X_i = 1$ και $\sum c_i X_i = \sum(k_i + d_i)X_i = 1 \Rightarrow \sum d_i X_i = 0$.

Είναι $\sum k_i d_i = \frac{\sum (X_i - \bar{X})d_i}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{\sum X_i d_i}{\sum_{i=1}^n (X_i - \bar{X})^2} - \bar{X} \frac{\sum d_i}{\sum_{i=1}^n (X_i - \bar{X})^2} = 0$,
 οπότε $V(b_1) = V(\hat{\beta}_1) + \sigma^2 \sum d_i^2$.

Η ποσότητα $\sigma^2 \sum d_i^2$ ελαχιστοποιείται για $\sum d_i^2 = 0$. Άρα η διασπορά του b_1 είναι ελάχιστη όταν $\sum d_i^2 = 0 \Leftrightarrow d_i = 0 \forall i$, δηλαδή $c_i = k_i, \forall i$.
 Συνεπώς η εκτιμήτρια των ελαχίστων τετραγώνων (ε.ε.τ.), $\hat{\beta}_1$, έχει την ελάχιστη διασπορά μεταξύ των αμερόληπτων γραμμικών εκτιμητριών.

- Αμεροληψία της $\hat{\beta}_0$:

$$\begin{aligned} E(\hat{\beta}_0) &= E(\bar{Y} - \hat{\beta}_1 \bar{X}) = E\left[\frac{1}{n} \left(\sum_{i=1}^n Y_i - \hat{\beta}_1 \sum_{i=1}^n X_i\right)\right] = \\ &= \frac{1}{n} \left[\sum_{i=1}^n E(Y_i) - \left(\sum_{i=1}^n X_i\right) E(\hat{\beta}_1) \right] = \\ &= \frac{1}{n} \left[\sum_{i=1}^n (\beta_0 + \beta_1 X_i) - \beta_1 \sum_{i=1}^n X_i \right] = \\ &= \frac{1}{n} \left[n\beta_0 + \beta_1 \sum_{i=1}^n X_i - \beta_1 \sum_{i=1}^n X_i \right] = \beta_0. \end{aligned}$$

Παρατήρηση: Τα X_i δεν είναι τυχαίες μεταβλητές.

Τα Y_i είναι τυχαίες μεταβλητές, ανεξάρτητες αλλά όχι ισόνομες (έχουν διαφορετικές αναμενόμενες τιμές και κοινή διακύμανση).

Εκτίμηση του σ^2

- Αν Y_1, Y_2, \dots, Y_n τυχαίο δείγμα από κατανομή με γνωστό μέσο μ και διασπορά σ^2 , τότε η εκτιμήτρια του σ^2 είναι η $\hat{\sigma}^2 = \frac{1}{n} \sum (Y_i - \mu)^2$. Αν ο μέσος μ είναι άγνωστος θα εκτιμηθεί από το \bar{Y} και τότε το σ^2 εκτιμάται από το άθροισμα των τετραγωνικών αποκλίσεων των Y_i από τον κοινό τους μέσο, $S^2 = \frac{1}{n-1} \sum (Y_i - \bar{Y})^2$. Είναι $E(S^2) = \sigma^2$ [αμερόληπτη εκτιμήτρια].
- Στο γραμμικό μοντέλο τα Y_i έχουν διαφορετικές κατανομές που εξαρτώνται από τα X_i . Επομένως, η απόκλιση κάθε παρατήρησης πρέπει να υπολογιστεί από το μέσο της: $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$.
Άρα, αν συμβολίσουμε με $\hat{\varepsilon}_i$ τις εκτιμήσεις των σφαλμάτων (κατάλοιπα-*residuals*), υπολογίζουμε το άθροισμα

$$\sum \hat{\varepsilon}_i^2 = \sum (Y_i - \hat{Y}_i)^2 = \sum (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$$

error sum of squares or residuals sum of squares
 άθροισμα τετραγώνων καταλοίπων

Μια αμερόληπτη εκτιμήτρια του σ^2 είναι η

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$$

mean square error
 μέσο τετραγωνικό σφάλμα

Διαιρούμε με $n - 2$ (β.ε.) καθώς έχουν εκτιμηθεί 2 παράμετροι.

Εκτίμηση με τη Μέθοδο Μέγιστης Πιθανοφάνειας

Το απλό γραμμικό μοντέλο, κάτω από την υπόθεση της κανονικότητας για τους τυχαίους όρους, γράφεται ως

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2), \quad i = 1, \dots, n.$$

Παρατήρηση: Η υπόθεση των ασυσχέτιστων σφαλμάτων, $Cov(\varepsilon_i, \varepsilon_j) = 0, i \neq j$, κάτω από την υπόθεση της κανονικότητας αντιστοιχεί σε υπόθεση ανεξαρτησίας (ανεξάρτητα σφάλματα).

Επομένως, για τις απαντητικές μεταβλητές μεταβλητές έχουμε υποθέσει $Y_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2), i = 1, \dots, n$. Η υπόθεση συγκεκριμένης κατανομής για της μεταβλητές ενδιαφέροντες επιτρέπει την εκτίμηση των παραμέτρων του μοντέλου με τη μέθοδο μέγιστης πιθανοφάνειας.

Συνάρτηση πιθανοφάνειας (*likelihood function*):

$$\begin{aligned} L(\beta_0, \beta_1, \sigma^2) &= \prod_1^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(Y_i - \beta_0 - \beta_1 X_i)^2\right\} \\ &= (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2} \sum (Y_i - \beta_0 - \beta_1 X_i)^2\right\} \end{aligned}$$

$$\log - likelihood: \log L = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum (Y_i - \beta_0 - \beta_1 X_i)^2$$

$$\left. \begin{aligned} \frac{\partial \log L}{\partial \beta_0} &= \frac{1}{\sigma^2} \sum (Y_i - \beta_0 - \beta_1 X_i) = 0 \\ \frac{\partial \log L}{\partial \beta_1} &= \frac{1}{\sigma^2} \sum X_i (Y_i - \beta_0 - \beta_1 X_i) = 0 \end{aligned} \right\} \text{κανονικές εξισώσεις}$$

$\Rightarrow \hat{\beta}_0, \hat{\beta}_1$ ίδιες με εκτιμήτριες ελαχίστων τετραγώνων.

$$\begin{aligned} \frac{\partial \log L}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum (Y_i - \beta_0 - \beta_1 X_i)^2 = 0 \\ \Rightarrow \hat{\sigma}^2 &= \frac{1}{n} \sum (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 \end{aligned}$$

Προσοχή: Η εκτιμήτρια μέγιστης πιθανοφάνειας της διασποράς δεν είναι αμερόληπτη.

Παρατηρήσεις

Οι εκτιμήτριες $\hat{\beta}_0$ και $\hat{\beta}_1$ ως εκτιμήτριες ελαχίστων τετραγώνων (EET)

- 1) είναι αμερόληπτες (Unbiased)
- 2) έχουν ελάχιστη διασπορά μεταξύ των α.ε. που είναι γραμμικοί συνδυασμοί των Y_i (Best Linear Unbiased Estimators - BLUE).

Επίσης ως εκτιμήτριες μέγιστης πιθανοφάνειας (EMΠ) είναι

- (1) συνεπείς
- (2) επαρκείς
- (3) αμερόληπτες εκτιμήτριες ελάχιστης διασποράς (έχουν ελάχιστη διασπορά μεταξύ όλων των α.ε.)

Κατανομές των Εκτιμητών των Συντελεστών της Παλινδρόμησης

Κατανομή της $\hat{\beta}_1$

$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, i = 1, \dots, \varepsilon_i \sim N(0, \sigma^2)$, ανεξάρτητα
 $\Rightarrow Y_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$.

$$\text{Είναι } \hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} = \sum k_i Y_i.$$

Άρα η $\hat{\beta}_1$ ακολουθεί την κανονική κατανομή ως γραμμικός συνδυασμός κανονικών τυχαίων μεταβλητών.

Έχουμε δείξει ότι

$$E(\hat{\beta}_1) = \beta_1$$

$$\sigma^2(\hat{\beta}_1) = V\left(\sum k_i Y_i\right) = \sum k_i^2 V(Y_i) = \sigma^2 \sum k_i^2 = \frac{\sigma^2}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

Άρα η $\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^n (X_i - \bar{X})^2}\right)$

Παρατήρηση: Η διασπορά σ^2 είναι άγνωστη. Μπορεί όμως να εκτιμηθεί από την α.ε. της

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2.$$

Οπότε η εκτιμώμενη διασπορά της $\hat{\beta}_1$ είναι

$$S^2(\hat{\beta}_1) = \frac{\hat{\sigma}^2}{\sum_{i=1}^n (X_i - \bar{X})^2}.$$

Κατανομή του $\hat{\beta}_0$

Είναι $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$, όπου τα Y_i είναι κανονικές τ.μ. και το $\hat{\beta}_1$ επίσης κανονική τ.μ. Άρα το $\hat{\beta}_0$ ακολουθεί την κανονική κατανομή ως γραμμικός συνδυασμός κανονικών τ.μ.

Εχουμε δείξει ότι

$$\begin{aligned} E(\hat{\beta}_0) &= \beta_0 \\ \sigma^2(\hat{\beta}_0) &= V(\bar{Y} - \hat{\beta}_1 \bar{X}) = V(\bar{Y}) + V(\hat{\beta}_1 \bar{X}) - 2Cov(\bar{Y}, \hat{\beta}_1 \bar{X}) \\ &= \frac{\sigma^2}{n} + \bar{X}^2 V(\hat{\beta}_1) - 2\bar{X} \underbrace{Cov(\bar{Y}, \hat{\beta}_1)}_0 = \frac{\sigma^2}{n} + \bar{X}^2 \frac{\sigma^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \\ &= \sigma^2 \left[\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \right]. \end{aligned}$$

$$\text{Εκτιμώμενη διασπορά: } S^2(\hat{\beta}_0) = \hat{\sigma}^2 \left[\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \right].$$

Πρόταση. Ισχύει

$$Cov(\hat{\beta}_1, \bar{Y}) = 0$$

Απόδειξη.

Ιδιότητες συνδιακύμανσης: $Cov(X + Y, Z) = Cov(X, Z) + Cov(Y, Z)$
 $Cov(X + Y, Z + W) = Cov(X, Z) + Cov(X, W) + Cov(Y, Z) + Cov(Y, W)$

$$\begin{aligned}
Cov(\hat{\beta}_1, \bar{Y}) &= Cov\left(\sum_{i=1}^n k_i Y_i, \frac{\sum_{j=1}^n Y_j}{n}\right) = \sum_{i=1}^n \sum_{j=1}^n Cov\left(k_i Y_i, \frac{Y_j}{n}\right) = \sum_{i=1}^n \sum_{j=1}^n \frac{k_i}{n} Cov(Y_i, Y_j) \\
&= \sum_{i=1}^n \frac{k_i}{n} Cov(Y_i, Y_i) = \frac{1}{n} \sum_{i=1}^n k_i V(Y_i) = \frac{\sigma^2}{n} \underbrace{\sum_{i=1}^n k_i}_0 = 0
\end{aligned}$$

αφού $Cov(Y_i, Y_j) = 0, i \neq j$.

Συμπερασματολογία για τους Συντελεστές του Γραμμικού Μοντέλου

Πρόταση. Η τυχαία μεταβλητή $\frac{\hat{\beta}_1 - \beta_1}{s(\hat{\beta}_1)} \sim t_{(n-2)}$

Απόδειξη.

Ισχύει: Αν $Z \sim N(0, 1), U \sim X_{(r)}^2$ και Z, U ανεξάρτητες, τότε

$$T = \frac{Z}{\sqrt{\frac{U}{r}}} \sim t_{(r)}$$

Έχουμε $\hat{\beta}_1 \sim N(\beta_1, \sigma^2(\hat{\beta}_1)) \Rightarrow \frac{\hat{\beta}_1 - \beta_1}{\sigma(\hat{\beta}_1)} \sim N(0, 1)$.

Επίσης ισχύει ότι (θα αποδειχθεί στη συνέχεια):

$$\frac{\sum \hat{\varepsilon}_i^2}{\sigma^2} = \frac{\sum (Y_i - \hat{Y}_i)^2}{\sigma^2} = \frac{\sum (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2}{\sigma^2} \sim X^2(n-2).$$

Και επειδή $\hat{\sigma}^2 = \frac{1}{n-2} \sum \hat{\varepsilon}_i^2$, έχουμε $\frac{\hat{\sigma}^2(n-2)}{\sigma^2} \sim X^2(n-2)$.

Τώρα

$$\left. \begin{aligned} S^2(\hat{\beta}_1) &= \frac{\hat{\sigma}^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \\ \sigma^2(\hat{\beta}_1) &= \frac{\sigma^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \end{aligned} \right\} \Rightarrow \frac{S^2(\hat{\beta}_1)}{\sigma^2(\hat{\beta}_1)} = \frac{\hat{\sigma}^2}{\sigma^2}.$$

$$\text{Άρα } \frac{S^2(\hat{\beta}_1)(n-2)}{\sigma^2(\hat{\beta}_1)} \sim \chi^2(n-2).$$

Επομένως

$$\frac{\frac{\hat{\beta}_1 - \beta_1}{\sigma(\hat{\beta}_1)}}{\sqrt{\frac{S^2(\hat{\beta}_1)(n-2)}{\sigma^2(\hat{\beta}_1)(n-2)}}} = \frac{\hat{\beta}_1 - \beta_1}{s(\hat{\beta}_1)} \sim t_{(n-2)}.$$

Παρατήρηση: Η πρόταση αυτή είναι ιδιαίτερα χρήσιμη για στατιστική συμπερασματολογία, καθώς στο αποτέλεσμα αυτό στηρίζεται η κατασκευή διαστημάτων εμπιστοσύνης και η διενέργεια ελέγχων υποθέσεων για τους συντελεστές του γραμμικού μοντέλου.

Κατασκευή Διαστημάτων Εμπιστοσύνης

Θα κατασκευάσουμε διάστημα εμπιστοσύνης συντελεστή 1- α για το β_1 .

$$\text{Είναι } P\left(-t_{\frac{\alpha}{2}}(n-2) \leq \frac{\hat{\beta}_1 - \beta_1}{s(\hat{\beta}_1)} \leq t_{\frac{\alpha}{2}}(n-2)\right) = 1 - \alpha.$$

Άρα το ζητούμενο δ.ε. για το β_1 είναι

$$\hat{\beta}_1 \pm t_{\frac{\alpha}{2}}(n-2)s(\hat{\beta}_1).$$

Ομοίως με παραπάνω είναι $\frac{\hat{\beta}_0 - \beta_0}{s(\hat{\beta}_0)} \sim t(n-2)$, άρα το διάστημα εμπιστοσύνης με συντελεστή 1- α για το β_0 είναι

$$\hat{\beta}_0 \pm t_{\frac{\alpha}{2}}(n-2)s(\hat{\beta}_0).$$

Έλεγχοι Υποθέσεων

Χρησιμοποιώντας την κατανομή $t(n - 2)$ μπορούμε να κάνουμε και ελέγχους υποθέσεων για τις παραμέτρους β_0 και β_1 .

Η λογική των ελέγχων υποθέσεων είναι να βρεθεί στατιστική συνάρτηση (συνάρτηση του δείγματος που δεν εμπλέκει άγνωστες ποσότητες) με γνωστή κατανομή (κατανομή που δεν έχει άγνωστες παραμέτρους) και για την οποία ακραίες τιμές της αντιστοιχούν σε ένδειξη ότι η H_0 πρέπει να απορριφθεί έναντι της συγκεκριμένης εναλλακτικής ως προς την οποία ελέγχεται.

Οι έλεγχοι

$$H_0 : \beta_0 = 0$$

$$H_1 : \beta_0 \neq 0$$

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0$$

ονομάζονται έλεγχοι στατιστικής σημαντικότητας των αντίστοιχων παραμέτρων. Αν σε έναν έλεγχο στατιστικής σημαντικότητας δεν μπορεί να απορριφθεί η αρχική υπόθεση αυτό αντιστοιχεί στην παραδοχή ότι δεν υφίσταται η συγκεκριμένη παράμετρος στο μοντέλο. Στην περίπτωση που ο έλεγχος γίνεται για τον συντελεστή κλίσης, β_1 , μη απόρριψη της H_0 οδηγεί στο συμπέρασμα ότι δεν υφίσταται γραμμική σχέση ανάμεσα στην απαντητική μεταβλητή, Y , και την ερμηνευτική μεταβλητή, X . Πράγματι, αν $\beta_1 = 0$, το μοντέλο υποθέτει ότι $Y_i \sim N(\beta_0, \sigma^2)$, $i = 1, \dots, n$, δηλαδή τα Y_i είναι ανεξάρτητα και ισόνομα.

Για τον έλεγχο στατιστικής σημαντικότητας του β_1 χρησιμοποιείται η στατιστική συνάρτηση $T = \frac{\hat{\beta}_1 - \beta_1}{s(\hat{\beta}_1)}$ (η οποία όπως έχουμε πει ακολουθεί πάντα $t(n-2)$ κατανομή).

Κάτω από την ισχύ της H_0 , η στατιστική συνάρτηση ελέγχου (ελεγχοσυνάρτηση) $T = \frac{\hat{\beta}_1}{s(\hat{\beta}_1)} \sim t(n - 2)$.

Για τα δεδομένα του δείγματός μας υπολογίζουμε την παρατηρούμενη τιμή της ελεγχοσυνάρτησης στο δείγμα, έστω t^* . Απορρίπτουμε την H_0 , σε επίπεδο στατιστικής σημαντικότητας α , αν $|t^*| > t_{\alpha/2}(n-2)$ (ισοδύναμα για $t^* > t_{\alpha/2}(n-2)$ ή $t^* < -t_{\alpha/2}(n-2)$). Δηλαδή, απορρίπτουμε την $H_0 : \beta_1 = 0$ έναντι της αμφίπλευρης εναλλακτικής, αν η παρατηρούμενη τιμή της ελεγχοσυνάρτησης υπερβαίνει το $\alpha/2$ άνω ποσοστιαίο σημείο της $t(n - 2)$ κατανομής.

Για τον έλεγχο της $H_0 : \beta_1 = 3$ έναντι της αμφίπλευρης εναλλακτικής $H_1 : \beta_1 \neq 3$ η διαδικασία είναι αντίστοιχη. Η ελεγχοσυνάρτηση που χρησιμοποιείται

τώρα είναι η $T = \frac{\hat{\beta}_1 - 3}{s(\hat{\beta}_1)}$, η οποία, κάτω από την ισχύ της H_0 ακολουθεί $t(n - 2)$ κατανομή. Για τα δεδομένα του δείγματός μας υπολογίζουμε την παρατηρούμενη τιμή αυτής της ελεγχοσυνάρτησης στο δείγμα, έστω t^* . Απορρίπτουμε την H_0 , σε επίπεδο στατιστικής σημαντικότητας α , αν $|t^*| > t_{\alpha/2}(n - 2)$.

Για τον έλεγχο της $H_0 : \beta_1 = 3$ έναντι της μονόπλευρης εναλλακτικής $H_1 : \beta_1 > 3$ η ελεγχοσυνάρτηση που χρησιμοποιείται είναι πάλι η $T = \frac{\hat{\beta}_1 - 3}{s(\hat{\beta}_1)}$, η οποία, κάτω από την ισχύ της H_0 ακολουθεί $t(n - 2)$ κατανομή. Απορρίπτουμε την H_0 , σε επίπεδο στατιστικής σημαντικότητας α , έναντι της συγκεκριμένης εναλλακτικής, αν $t^* > t_{\alpha}(n - 2)$.

Για τον έλεγχο της $H_0 : \beta_1 = 3$ έναντι της μονόπλευρης εναλλακτικής $H_1 : \beta_1 < 3$ η ελεγχοσυνάρτηση που χρησιμοποιείται είναι πάλι η $T = \frac{\hat{\beta}_1 - 3}{s(\hat{\beta}_1)}$, η οποία, κάτω από την ισχύ της H_0 ακολουθεί $t(n - 2)$ κατανομή. Απορρίπτουμε την H_0 , σε επίπεδο στατιστικής σημαντικότητας α , έναντι της συγκεκριμένης εναλλακτικής, αν $t^* < -t_{\alpha}(n - 2)$.

Παρατηρούμενο Επίπεδο Σημαντικότητας (p-value)

Το παρατηρούμενο επίπεδο σημαντικότητας (observed level of significance ή p-value) ορίζεται ως η πιθανότητα να πάρει η ελεγχοσυνάρτηση, T , τιμή τόσο ή περισσότερο ακραία από την παρατηρούμενη στο δείγμα μας, t^* .

Παρατήρηση: Όταν λέμε ακραία στον ορισμό του p-value, εννοούμε ακραία ως προς την συγκεκριμένη εναλλακτική του ελέγχου. Έτσι, για αμφίπλευρο έλεγχο, $\text{p-value} = P(|T| > |t^*|)$, ενώ για τους μονόπλευρους ελέγχους $\text{p-value} = P(T > t^*)$ ή $\text{p-value} = P(T < t^*)$, αντίστοιχα.

Σε οποιονδήποτε έλεγχο υποθέσεων, απορρίπτουμε την H_0 αν το παρατηρούμενο επίπεδο σημαντικότητας (p-value) είναι χαμηλότερο από το προκαθορισμένο επίπεδο σημαντικότητας, α . Μάλιστα, όσο μικρότερη είναι η τιμή του p-value τόσο ισχυρότερη ένδειξη αποτελεί αυτό εναντίον της H_0 .