Γραμμικά Μοντέλα
Το Γραμμικό Μοντέλο με Κανονικά Κατανεμημένα Σφάλματα

Διδάσκουσα: Λουκία Μελιγκοτσίδου
Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών
Τμήμα Μαθηματικών

April 2, 2020

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

Συνάρτηση πιθανοφάνειας ($likelihood\ function$):

$$L(\beta_0, \beta_1, \sigma^2) = \prod_1^n \frac{1}{\sqrt{2\pi\sigma^2}} exp\Big\{ -\frac{1}{2\sigma^2}(Y_i - \beta_0 - \beta_1 X_i)^2 \Big\}$$

$$= (2\pi\sigma^2)^{-n/2} exp\Big\{ -\frac{1}{2\sigma^2}(Y_i - \beta_0 - \beta_1 X_i)^2 \Big\}$$

.

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

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

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

$$\frac{\partial \log L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum(Y_i - \beta_0 - \beta_1 X_i)^2 = 0$$

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

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

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

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

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

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

Έχουμε δείξει ότι

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

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

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

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 εκτιμούμενη 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

$$E(\hat{\beta}_0) = \beta_0$$

$$\sigma^2(\hat{\beta}_0) = V(\bar{Y} - \hat{\beta}_1\bar{X}) = V(\bar{Y}) + V(\hat{\beta}_1\bar{X}) - 2Cov(\bar{Y}, \hat{\beta}_1\bar{X})$$

$$= \frac{\sigma^2}{n} + \bar{X}^2 V(\hat{\beta}_1) - 2\bar{X}\underbrace{Cov(\bar{Y}, \hat{\beta}_1)}_{0} = \frac{\sigma^2}{n} + \bar{X}^2 \frac{\sigma^2}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

$$= \sigma^2\Big[\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^n (X_i - \bar{X})^2}\Big].$$

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

Πρόταση. Ισχύει

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

Απόδειξη.

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

$$Cov(\hat{\beta}_1, \bar{Y}) = Cov(\sum_{i=1}^{n} k_i Y_i, \frac{\sum_{j=1}^{n} Y_j}{n}) = \sum_{i=1}^{n}\sum_{j=1}^{n} Cov(k_i Y_i, \frac{Y_j}{n}) = \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{k_i}{n} Cov(Y_i, Y_j)$$

$$= \sum_{i=1}^{n} \frac{k_i}{n} Cov(Y_i, Y_i) = \frac{1}{n}\sum_{i=1}^{n} k_i V(Y_i) = \frac{\sigma^2}{n} \underbrace{\sum_{i=1}^{n} k_i}_{0} = 0$$

αφού $Cov(Y_i, Y_j) = 0, i \neq j$.

Συμπερασματολογία για τους Συντελεστές του Γραμμικού Μοντέλου

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

Απόδειξη.

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

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

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

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

Τώρα

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

Άρα $\dfrac{S^2(\hat{\beta}_1)(n-2)}{\sigma^2(\hat{\beta}_1)} \sim X^2(n-2)$.

Επομένως

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

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.

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

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 $\dfrac{\hat{\beta}_0 - \beta_0}{s(\hat{\beta}_0)} \sim t(n-2)$, άρα το διάστημα εεμπιστοσύνης 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

$$H_0: \ \beta_0 = 0 \qquad\qquad H_0: \ \beta_1 = 0$$
$$H_1: \ \beta_0 \neq 0 \qquad\qquad H_0: \ \beta_1 \neq 0$$

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 $Y_i \sim N(\beta_0, \sigma^2)$, $i = 1, \ldots, n$, 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 $H_0 : \beta_1 = 3$ έναντί 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 $H_0 : \beta_1 = 3$ έναντί 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, p-value=$P(|T| > |t^*|)$, ενώ για τους μονόπλευρους ελέγχους p-value=$P(T > t^*)$ ή 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.

Άσκηση

Έστω το εναλλακτικό απλό γραμμικό μοντέλο

$$Y_i = \beta_0^* + \beta_1(X_i - \bar{X}) + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2), \quad \text{ανεξάρτητα.}$$

(a) Να βρεθούν οι εκτιμητές μέγιστης πιθανοφάνειας των παραμέτρων του μοντέλου.

(b) Να βρεθεί η κατανομή του $\hat{\beta}_0^*$.

(c) Να κατασκευαστεί διάστημα εμπιστοσύνης συντελεστή $1-\alpha$ για το β_0^*.

(d) Να πραγματοποιηθεί ο έλεγχος στατιστικής σημαντικότητας του συντελεστή β_0^* σε ε.σ.σ α.

Λύση του (b)

Το εναλλακτικό γραμμικό μοντέλο της άσκησης είναι ισοδύναμο με το αρχικό απλό γραμμικό μοντέλο.

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

$$\text{Άρα} \quad \hat{\beta}_0^* = \hat{\beta}_0 + \hat{\beta}_1\bar{X} = \bar{Y} - \hat{\beta}_1\bar{X} + \hat{\beta}_1\bar{X} = \bar{Y}$$

Επειδή τα Y_i είναι κανονικές τ.μ., το $\hat{\beta}_0^* = \bar{Y}$ είναι επίσης κανονική τ.μ. ως γραμμικός συνδυασμός κανονικών.

$$E(\hat{\beta}_0^*) = E(\hat{\beta}_0 + \hat{\beta}_1\bar{X}) = E(\hat{\beta}_0) + \bar{X}E(\hat{\beta}_1) = \beta_0 + \beta_1\bar{X} = \beta_0^*.$$

Εναλλακτικά

$$E(\hat{\beta}_0^*) = E(\bar{Y}) = E\Big(\frac{\sum Y_i}{n}\Big) = \frac{1}{n}\sum E(Y_i) = \frac{1}{n}\sum(\beta_0^* + \beta_1(X_i - \bar{X}))$$
$$= \frac{1}{n}n\beta_0^* + \frac{1}{n}\underbrace{\sum(X_i - \bar{X})}_{0} = \beta_0^*.$$

Δηλαδή $\hat{\beta}_0^*$ αμερόληπτη εκτιμήτρια.

$$V(\hat{\beta}_0^*) = \sigma^2(\hat{\beta}_0^*) = V(\bar{Y}) = V\Big(\frac{\sum Y_i}{n}\Big) = \frac{1}{n^2}\sum V(Y_i) = \frac{1}{n^2}\sum\sigma^2 = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n}.$$

Άρα $\hat{\beta}_0^* \sim N(\beta_0^*, \frac{\sigma^2}{n})$.

Κατανομές Τετραγωνικών Μορφών

Ορισμός: Ένα ομογενές πολυώνυμο 2ου βαθμού σε n μεταβλητές καλείται τετραγωνική μορφή σε αυτές τις μεταβλητές. Εάν και οι συντελεστές και οι μεταβλητές είναι πραγματικοί αριθμοί τότε καλείται πραγματική τετραγωνική μορφή.

Σημείωση: Ομογενές πολυώνυμο είναι αυτό που έχει ίδια δύναμη σε όλους τους όρους του.

Παραδείγματα

$(1) X_1^2 + X_1X_2 + X_2^2$ τετραγωνική μορφή ως προς X_1, X_2

$(2) X_1^2 + X_2^2 + X_3^2 - 2X_1X_2$ τετραγωνική μορφή ως προς X_1, X_2, X_3

$(3) X_1^2 + X_2^2 - 2X_1 - 4X_2 + 5$ δεν είναι τετραγωνική μορφή ως προς X_1, X_2

$(X_1-1)^2 + (X_2-2)^2$ αλλά είναι ως προς $(X_1-1), (X_2-2)$

Πρόταση. Έστω $X_1, X_2, \ldots, X_n$ τ.δ. από την τ.μ. X και $\bar{X}$, $S^2 = \frac{1}{n-1}\sum(X_i - \bar{X})^2$ ο δειγματικός μέσος και η δειγματική διασπορά αντίστοιχα. Η τ.μ. $(n-1)S^2$ είναι μια τετραγωνική μορφή στις n τ.μ. $X_1, X_2, \ldots, X_n$.

Απόδειξη.

$$(n-1)S^2 = \sum(X_i - \bar{X})^2 = \sum(X_i - \tfrac{\sum X_i}{n})^2$$

$$= \sum X_i^2 + \frac{n(\sum X_i)^2}{n^2} - 2\sum X_i \frac{\sum X_i}{n} =$$
$$= \sum X_i^2 + \frac{(\sum X_i)^2}{n} - \frac{2}{n}(\sum X_i)^2 =$$
$$= \sum X_i^2 - \frac{(\sum X_i)^2}{n} = \sum X_i^2 - \frac{1}{n}\Big\{\sum X_i^2 + 2\sum_{\substack{i,j=1 \\ i<j}}^{n} X_iX_j\Big\} =$$

$$= \frac{n-1}{n}\sum X_i^2 - \frac{2}{n}\sum_{\substack{i,j=1\\ i<j}}^{n} X_iX_j \qquad \text{τετραγωνική μορφή}$$

Παρατήρηση: Γνωρίζουμε ότι εάν $X \sim N(\mu, \sigma^2)$ και $X_1, X_2, \ldots, X_n$ τ.δ. τότε $\dfrac{(n-1)S^2}{\sigma^2} \sim X^2_{(n-1)}$.

Παρατήρηση: Γνωρίζουμε ότι εάν $Q_1, Q_2, \ldots, Q_k$ είναι ανεξάρτητες X^2 τ.μ. με $r_i, i = 1, \ldots, k$ β.ε. αντίστοιχα, τότε η τ.μ. $Q = \sum_{i=1}^{k} Q_i$ είναι $X^2_{(r_1+r_2+\ldots+r_k)}$.

Θεώρημα. Έστω $Q = \sum_{i=1}^{k} Q_i$ όπου $Q_1, Q_2, \ldots, Q_k$ είναι k πραγματικές τετραγωνικές μορφές σε n ανεξάρτητες τ.μ. που ακολουθούν την κανονική κατανομή με μέσους $\mu_1, \mu_2, \ldots, \mu_n$ αντίστοιχα και την ίδια διασπορά σ^2.
Έστω επίσης ότι (i) $\dfrac{Q}{\sigma^2}, \dfrac{Q_1}{\sigma^2}, \dfrac{Q_2}{\sigma^2}, \ldots, \dfrac{Q_{k-1}}{\sigma^2}$ έχουν την κατανομή X^2 με $r, r_1, r_2, \ldots, r_{k-1}$ β.ε. αντίστοιχα, και (ii) Q_k μη αρνητική. Τότε
$(1) Q_1, Q_2, \ldots, Q_k$ είναι ανεξάρτητες και

$(2) \dfrac{Q_k}{\sigma^2} \sim X^2_{(r_k)}$, όπου $r_k = r - (r_1 + r_2 + \ldots + r_{k-1})$.

Πρόταση. Έστω το γραμμικό μοντέλο

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i,$$

ή

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

Εάν $\varepsilon_i \sim N(0, \sigma^2)$, τότε $Y_i \overset{\text{ανεξ}}{\sim} N(\beta_0 + \beta_1 X_i, \sigma^2)$
ή $Y_i \sim N(\beta_0^* + \beta_1(X_i - \bar{X}), \sigma^2)$.

Ισχύει ότι $\dfrac{\sum \hat{\varepsilon}_i^2}{\sigma^2} = \dfrac{\sum (Y_i - \hat{Y}_i)^2}{\sigma^2} = \dfrac{\sum (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2}{\sigma^2} \sim X^2_{(n-2)}$
ή αλλιώς $\dfrac{\sum (Y_i - \hat{\beta}_0^* - \hat{\beta}_1(X_i - \bar{X}))^2}{\sigma^2} \sim X^2_{(n-2)}$.

Απόδειξη.

$$\sum(Y_i - \beta_0 - \beta_1 X_i)^2 = \sum(Y_i - \beta_0^* - \beta_1(X_i - \bar{X}))^2 =$$
$$= \sum \left[(\hat{\beta}_0^* - \beta_0^*) + (\hat{\beta}_1 - \beta_1)(X_i - \bar{X}) + [Y_i - \hat{\beta}_0^* - \hat{\beta}_1(X_i - \bar{X})]\right]^2$$
(όπου έχουμε προσθαφαιρέσει $\hat{Y}_i = \hat{\beta}_0^* + \hat{\beta}_1(X_i - \bar{X})$)
$$= n(\hat{\beta}_0^* - \beta_0^*)^2 + (\hat{\beta}_1 - \beta_1)^2 \sum(X_i - \bar{X})^2 + \sum \left[(Y_i - \hat{\beta}_0^* - \hat{\beta}_1(X_i - \bar{X})\right]^2 + 0$$
(αθροίσματα διπλασίων γινομένων $= 0$)

Συμβολίζουμε με

$$Q = \sum(Y_i - \beta_0 - \beta_1 X_i)^2 = \sum \left(Y_i - \beta_0^* - \beta_1(X_i - \bar{X})\right)^2$$
$$Q_1 = n(\hat{\beta}_0^* - \beta_0^*)^2$$
$$Q_2 = (\hat{\beta}_1 - \beta_1)^2 \sum(X_i - \bar{X})^2$$
$$Q_3 = \sum \left[Y_i - \hat{\beta}_0^* - \hat{\beta}_1(X_i - \bar{X})\right]^2$$

Είναι $Q = Q_1 + Q_2 + Q_3$.

Τα Q, Q_1, Q_2, Q_3 είναι τετραγωνικές μορφές ως προς κανονικές τ.μ.

Επίσης έχουμε $Y_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2) \Rightarrow \dfrac{(Y_i - \beta_0 - \beta_1 X_i)^2}{\sigma^2} \sim X^2_{(1)}$.
Οι Y_i είναι ανεξάρτητες άρα $\sum_{i=1}^n \dfrac{(Y_i - \beta_0 - \beta_1 X_i)^2}{\sigma^2} \sim X^2_{(n)}$,

$$\text{δηλ. } \frac{Q}{\sigma^2} \sim X^2_{(n)}.$$

$$\hat{\beta}_0^* \sim N(\beta_0, \frac{\sigma^2}{n}) \Rightarrow \frac{n(\hat{\beta}_0^* - \beta_0^*)^2}{\sigma^2} \sim X^2_{(1)} \text{ , δηλ. } \frac{Q_1}{\sigma^2} \sim X^2_{(1)}.$$

$$\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{\sum_{i=1}^n (X_i - \bar{X})^2}) \Rightarrow \frac{\sum_{i=1}^n (X_i - \bar{X})^2 (\hat{\beta}_1 - \beta_1)^2}{\sigma^2} = \frac{Q_2}{\sigma^2} \sim X^2_{(1)}.$$

Το Q_3 είναι μη αρνητικό, άρα με βάση το θεώρημα

$$\frac{Q_3}{\sigma^2} = \frac{\sum \left[Y_i - \hat{\beta}_0^* - \hat{\beta}_1(X_i - \bar{X})\right]^2}{\sigma^2} = \frac{\sum \hat{\varepsilon}_i^2}{\sigma^2} \sim X^2_{(n-1-1)} \equiv X^2_{(n-2)}.$$

Μένει να δείξουμε ότι τα αθροίσματα διπλασίων γινομένων είναι μηδενικά.

$(i)\ 2\sum(\hat{\beta}_0^* - \beta_0^*)(\hat{\beta}_1 - \beta_1)(X_i - \bar{X}) = 2(\hat{\beta}_0^* - \beta_0^*)(\hat{\beta}_1 - \beta_1)\sum(X_i - \bar{X}) = 0$

$(ii)\ 2\sum(\hat{\beta}_0^* - \beta_0^*)\Big(Y_i - \hat{\beta}_0^* - \hat{\beta}_1(X_i - \bar{X})\Big) =$

$= 2(\hat{\beta}_0^* - \beta_0^*)\sum(Y_i - \bar{Y}) - 2(\hat{\beta}_0^* - \beta_0^*)\hat{\beta}_1\sum(X_i - \bar{X}) = 0$

$(iii)\ 2\sum(\hat{\beta}_1 - \beta_1)(X_i - \bar{X})\Big(Y_i - \hat{\beta}_0^* - \hat{\beta}_1(X_i - \bar{X})\Big) =$

$= 2(\hat{\beta}_1 - \beta_1)\sum(X_i - \bar{X})(Y_i - \bar{Y}) - 2(\hat{\beta}_1 - \beta_1)\hat{\beta}_1\sum(X_i - \bar{X})^2 =$

$= 2(\hat{\beta}_1 - \beta_1)\Big[\sum(X_i - \bar{X})(Y_i - \bar{Y}) - \dfrac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}\sum(X_i - \bar{X})^2\Big] = 0.$

Άσκηση

Δίνεται το γραμμικό μοντέλο

$$Y_i = \beta(X_i - \bar{X}) + \varepsilon_i, i = 1, \ldots, n, \varepsilon_i \sim N(0, \sigma^2) \text{ασυσχέτιστα.}$$

(i) Να βρεθεί η ε.ε.τ. $\hat{\beta}$ της β και η κατανομή της.

(ii) Να βρεθεί η κατανομή της στατιστικής συνάρτησης $\dfrac{\hat{\beta} - \beta}{s(\hat{\beta})}$ όπου $s^2(\hat{\beta})$ είναι α.ε. της διασποράς του $\hat{\beta}$.

(iii) Για δεδομένη τιμή X_0 να βρεθεί δ.ε. συντελεστή 1-α για το $\beta(X_0 - \bar{X})$.

Λύση

(i) Εκτίμηση ελαχίστων τετραγώνων: ελαχιστοποιούμε την ποσότητα $Q = \sum \varepsilon_i^2 = \sum(Y_i - \beta(X_i - \bar{X}))^2$ ως προς β.

$$\frac{\partial Q}{\partial \beta} = -2\sum(X_i - \bar{X})[Y_i - \beta(X_i - \bar{X})] = 0$$

$$\Rightarrow -\sum(X_i - \bar{X})Y_i + \beta\sum(X_i - \bar{X})^2 = 0 \Rightarrow \hat{\beta} = \frac{\sum(X_i - \bar{X})Y_i}{\sum(X_i - \bar{X})^2}$$

Το $\hat{\beta}$ είναι γραμμικός συνδυασμός των Y_i, αφού

$\hat{\beta} = \sum k_i Y_i$, όπου $k_i = \dfrac{X_i - \bar{X}}{\sum(X_i - \bar{X})^2}$ σταθερές.

Επειδή $\varepsilon_i \sim N(0,\sigma^2)$ ασυσχέτιστες άρα και ανεξάρτητες τ.μ., ισχύει ότι

$$Y_i \sim N(\beta(X_i-\bar{X}),\sigma^2).$$

Άρα το $\hat{\beta}$ ακολουθεί κανονική κατανομή σαν γραμμικός συνδυασμός ανεξάρτητων κανονικών τ.μ.
Είναι

$$E(\hat{\beta}) = E\Big[\frac{\sum(X_i-\bar{X})Y_i}{\sum(X_i-\bar{X})^2}\Big] = \frac{\sum(X_i-\bar{X})E(Y_i)}{\sum(X_i-\bar{X})^2}$$
$$= \frac{\sum(X_i-\bar{X})\beta(X_i-\bar{X})}{\sum(X_i-\bar{X})^2} = \frac{\beta\sum(X_i-\bar{X})^2}{\sum(X_i-\bar{X})^2} = \beta$$

Δηλαδή είναι α.ε. του β.

$$V(\hat{\beta}) = V\Big[\frac{\sum(X_i-\bar{X})Y_i}{\sum(X_i-\bar{X})^2}\Big] = \frac{\sum(X_i-\bar{X})^2V(Y_i)}{\Big(\sum(X_i-\bar{X})^2\Big)^2} = \frac{\sigma^2\sum(X_i-\bar{X})^2}{\Big(\sum(X_i-\bar{X})^2\Big)^2} = \frac{\sigma^2}{\sum(X_i-\bar{X})^2}$$

Άρα

$$\hat{\beta} \sim N(\beta, \frac{\sigma^2}{\sum(X_i-\bar{X})^2})$$

(ii) Θα δείξουμε πρώτα ότι το $\dfrac{\sum(Y_i-\hat{\beta}(X_i-\bar{X}))^2}{\sigma^2} \sim X^2_{(n-1)}$.

$$\sum(Y_i-\beta(X_i-\bar{X}))^2 = \sum(Y_i-\hat{\beta}(X_i-\bar{X})+\hat{\beta}(X_i-\bar{X})-\beta(X_i-\bar{X}))^2 =$$

$$= \sum\Big[(\hat{\beta}-\beta)(X_i-\bar{X})+(Y_i-\hat{\beta}(X_i-\bar{X}))\Big]^2 =$$

$$= (\hat{\beta}-\beta)^2\sum(X_i-\bar{X})^2+\sum\Big(Y_i-\hat{\beta}(X_i-\bar{X})\Big)^2+2(\hat{\beta}-\beta)\sum(X_i-\bar{X})[Y_i-\hat{\beta}(X_i-\bar{X})] =$$

$$= (\hat{\beta}-\beta)^2\sum(X_i-\bar{X})^2+\sum\Big[Y_i-\hat{\beta}(X_i-\bar{X})\Big]^2$$

γιατί $\quad \sum(X_i-\bar{X})[Y_i-\hat{\beta}(X_i-\bar{X})] = \sum(X_i-\bar{X})Y_i-\hat{\beta}\sum(X_i-\bar{X})^2$

$$= \sum(X_i-\bar{X})Y_i-\frac{\sum(X_i-\bar{X})Y_i}{\sum(X_i-\bar{X})^2}\sum(X_i-\bar{X})^2 = 0$$

Άρα έχουμε

$Q = Q_1 + Q_2$, όπου

$$Q = \sum \Big(Y_i - \beta(X_i - \bar{X})\Big)^2 \qquad \text{πραγματικές}$$
$$Q_1 = (\hat{\beta} - \beta)^2 \sum (X_i - \bar{X})^2 \qquad \text{τετραγωνικές μορφές}$$
$$Q_2 = \sum (Y_i - \hat{\beta}(X_i - \bar{X}))^2$$

Επειδή $Y_i \sim N(\beta(X_i - \bar{X}), \sigma^2)$ έχουμε

$$\frac{[Y_i - \beta(X_i - \bar{X})]^2}{\sigma^2} \sim X^2_{(1)}$$

$$\Rightarrow \frac{Q}{\sigma^2} = \frac{\sum [Y_i - \beta(X_i - \bar{X})]^2}{\sigma^2} \sim X^2_{(n)}$$

$$\hat{\beta} \sim N(\beta, \frac{\sigma^2}{\sum (X_i - \bar{X})^2}) \Rightarrow \frac{(\hat{\beta} - \beta)^2 \sum (X_i - \bar{X})^2}{\sigma^2} = \frac{Q_1}{\sigma^2} \sim X^2_{(1)}$$

Επειδή Q_2 είναι μη αρνητικό, από το θεώρημα τετραγωνικών μορφών έχουμε

$$\frac{Q_2}{\sigma^2} = \frac{\sum (Y_i - \hat{\beta}(X_i - \bar{X}))^2}{\sigma^2} \sim X^2_{(n-1)}$$

Ξέρουμε ότι αν $W \sim X^2_{(n-1)}$, τότε $E(W) = n - 1$.

Άρα $E\Big[\frac{\sum (Y_i - \hat{\beta}(X_i - \bar{X}))^2}{\sigma^2}\Big] = n - 1 \Rightarrow E\Big[\frac{\sum (Y_i - \hat{\beta}(X_i - \bar{X}))^2}{n-1}\Big] = \sigma^2$.

Άρα η $\hat{\sigma}^2 = \frac{\sum (Y_i - \hat{\beta}(X_i - \bar{X}))^2}{n-1}$ είναι α.ε. του σ^2

και η $S^2(\hat{\beta}) = \frac{\hat{\sigma}^2}{\sum (X_i - \bar{X})^2}$ α.ε του $\sigma^2(\hat{\beta}) = \frac{\sigma^2}{\sum (X_i - \bar{X})^2}$.

Έχουμε βρεί ότι $\frac{\sum [Y_i - \beta(X_i - \bar{X})]^2}{\sigma^2} = \frac{(n-1)\hat{\sigma}^2}{\sigma^2} \sim X^2_{(n-1)}$

Άρα

$$\frac{(n-1)S^2(\hat{\beta})}{\sigma^2(\hat{\beta})} \sim X^2_{(n-1)}$$

και η κατανομή της είναι ανεξάρτητη του $\hat{\beta}$.

Άρα

$$T = \frac{\dfrac{\hat{\beta} - \beta}{\sigma(\hat{\beta})}}{\sqrt{\frac{(n-1)S^2(\hat{\beta})}{\sigma^2(\hat{\beta})(n-1)}}} = \frac{\hat{\beta} - \beta}{s(\hat{\beta})} \sim t(n-1)$$

(iii) Ζητάμε Δ.Ε. για το $\beta(X_0 - \bar{X})$.

Είναι $P\Big(- t_{\frac{\alpha}{2}}(n-1) < \dfrac{\hat{\beta} - \beta}{s(\hat{\beta})} < t_{\frac{\alpha}{2}}(n-1)\Big) = 1 - \alpha$

Άρα το ζητούμενο Δ.Ε. για το β με συντελεστή 1-α είναι

$$\hat{\beta} \pm t_{\frac{\alpha}{2}}(n-1)s(\hat{\beta})$$

οπότε το Δ.Ε. για το $\beta(X_0 - \bar{X})$ είναι

$$\hat{\beta}(X_0 - \bar{X}) \pm t_{\frac{\alpha}{2}}(n-1)s(\hat{\beta})(X_0 - \bar{X})$$