Linear Models

Vasileios Katsianos

Washington University in St. Louis
College of Arts & Sciences
Department of Statistics and Data Science

August 2024

"*All models are wrong, but some are useful.*"

George E. P. Box

Περιεχόμενα

Πρόλογος

Στη Μαθηματική Στατιστική ασχοληθήκαμε σχεδόν αποκλειστικά με εκτιμητική βασισμένη πάνω σε δείγματα ανεξάρτητων και ισόνομων παρατηρήσεων. Για παράδειγμα, μπορεί να είχαμε παρατηρήσεις $Y_1, Y_2, \ldots, Y_n$ οι οποίες μπορούσαμε να θεωρήσουμε ότι ήταν ανεξάρτητες και ισόνομες πραγματοποιήσεις από την κανονική κατανομή με κοινή μέση τιμή μ και κοινή διασπορά σ^2. Βάσει αυτού του τυχαίου δείγματος, μάθαμε να υπολογίζουμε σημειακές εκτιμήσεις και διαστήματα εμπιστοσύνης για τις άγνωστες παραμέτρους μ και σ^2, καθώς και να πραγματοποιούμε διάφορα είδη ελέγχων υποθέσεων.

Μία φαινομενικά άμεση γενίκευση του παραπάνω παραδείγματος θα ήταν να θεωρήσουμε ότι τα $Y_1, Y_2, \ldots, Y_n$ είναι ανεξάρτητες, αλλά όχι ισόνομες, παρατηρήσεις από την κανονική κατανομή. Για παράδειγμα, θα είχε ενδιαφέρον να θεωρήσουμε ότι έχουν κοινή διασπορά σ^2, αλλά διαφορετικές μέσες τιμές $\mu_1, \mu_2, \ldots, \mu_n$ αντίστοιχα. Φυσικά, σε αυτό το πολύ γενικό πλαίσιο θα ήταν αδύνατο να εκτιμήσουμε τις $n+1$ στο πλήθος άγνωστες παραμέτρους $\mu_1, \mu_2, \ldots, \mu_n, \sigma^2$ βασιζόμενοι πάνω σε ένα δείγμα μεγέθους n. Είναι απαραίτητο να θέσουμε κάποιον επιθυμητό "περιορισμό" πάνω στις άγνωστες μέσες τιμές με σκοπό να ελαττώσουμε το πλήθος των αγνώστων παραμέτρων. Θα μπορούσαμε να υποθέσουμε ότι $\mu_i = c_i\mu$ για $i = 1, 2, \ldots, n$, όπου c_i γνωστές σταθερές και μ άγνωστη παράμετρος, δηλαδή ότι όλες οι μέσες τιμές είναι γνωστά πολλαπλάσια μίας κοινής, αλλά άγνωστης, ποσότητας μ, την οποία και θέλουμε να εκτιμήσουμε. Με αυτόν τον τρόπο, επιστρέφουμε στο πλαίσιο των δύο αγνώστων παραμέτρων μ και σ^2 και η διαδικασία εκτίμησης αλλάζει ελάχιστα σε σύγκριση με την περίπτωση των ανεξάρτητων και ισόνομων παρατηρήσεων.

Η ανάλυση παλινδρόμησης δανείζεται από αυτή την ιδέα. Έστω ότι ενδιαφερόμαστε να μελετήσουμε τα βάρη $Y_1, Y_2, \ldots, Y_n$ των αγοριών ενός σχολείου. Δε θα ήταν λογικό να υποθέσουμε ότι τα βάρη των αγοριών έχουν κοινή μέση τιμή και κοινή διασπορά, αφού γνωρίζουμε τουλάχιστον έναν παράγοντα που επιδρά σημαντικά πάνω στο βάρος ενός ατόμου - το ύψος του. Φυσικά θα μπορούσαμε να σκεφτούμε και πολλούς άλλους παράγοντες που επηρεάζουν το βάρος

ενός ατόμου, αλλά προς το παρόν περιοριζόμαστε στο να καταγράψουμε τα ύψη $X_1, X_2, \ldots, X_n$ των ίδιων αγοριών. Έχοντας αυτά τα δεδομένα στα χέρια μας, υποθέτουμε ότι οι μέσες τιμές $\mu_i = E(Y_i)$ ακολουθούν τη σχέση $\mu_i = \beta_0 + \beta_1 X_i$ για $i = 1, 2, \ldots, n$. Τα $X_1, X_2, \ldots, X_n$ είναι γνωστά και μπορούμε να τα θεωρήσουμε σταθερά, οπότε παίζουν ακριβώς τον ίδιο ρόλο όπως οι γνωστές σταθερές $c_1, c_2, \ldots, c_n$ του προηγούμενου παραδείγματος.

Υποθέτοντας, επιπλέον, ότι τα βάρη $Y_1, Y_2, \ldots, Y_n$ ακολουθούν την κανονική κατανομή με τις παραπάνω μέσες τιμές και κοινή διασπορά σ^2, παίρνουμε ότι $Y_i \sim N\left(\beta_0 + \beta_1 X_i, \sigma^2\right)$ για $i = 1, 2, \ldots, n$. Συνεπώς, έχουμε τρεις άγνωστες παραμέτρους προς εκτίμηση: β_0, β_1 και σ^2. Για να ξεχωρίσουμε καλύτερα το κομμάτι που αναφέρεται στη μέση τιμή των Y_i από το κομμάτι που αναφέρεται στην κατανομή των Y_i, θα μπορούσαμε ισοδύναμα να γράψουμε ότι $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, όπου $\varepsilon_i \sim N(0, \sigma^2)$ για $i = 1, 2, \ldots, n$. Προφανώς, τα βάρη $Y_1, Y_2, \ldots, Y_n$ είναι ανεξάρτητα μεταξύ τους, οπότε και οι τυχαίες μεταβλητές $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ είναι στοχαστικά ανεξάρτητες. Αυτό το μοντέλο που περιγράφει την κατανομή των Y_i συναρτήσει των X_i ονομάζεται απλό γραμμικό μοντέλο με κανονικά σφάλματα ή αλλιώς κανονικό απλό γραμμικό μοντέλο.

Οι παρούσες σημειώσεις καλύπτουν τη διδασκαλία των γραμμικών μοντέλων στο τμήμα Μαθηματικών του πανεπιστημίου Αθηνών και θίγουν αρκετά επιπλέον συναφή ζητήματα που παρουσιάζουν ενδιαφέρον για περαιτέρω εμβάθυνση στις μεθόδους που μελετώνται. Σε κάθε κεφάλαιο παρατίθεται η απαιτούμενη θεωρία και παρεμβάλλονται αρκετές παρατηρήσεις που στοχεύουν στην καλύτερη κατανόηση των εννοιών που παρουσιάζονται. Στο τέλος των σημειώσεων, παρατίθενται λυμένα θέματα εξετάσεων από τα ακαδημαϊκά έτη 2008 - 2017.

Σημειώνεται ότι έχει γίνει μία προσπάθεια να παρουσιαστούν αναλυτικά οι αποδείξεις των περισσότερων αποτελεσμάτων που χρησιμοποιούνται στα πλαίσια των γραμμικών μοντέλων. Πολλά από αυτά τα αποτελέσματα συνήθως θεωρούνται δεδομένα και δεν αποδεικνύονται στη διαθέσιμη βιβλιογραφία. Προτείνεται κάποιος να αποκτήσει μία οικειότητα με τις μεθόδους της μαθηματικής στατιστικής πριν ασχοληθεί περαιτέρω με τη μελέτη των γραμμικών μοντέλων.

Κατσιάνος Βασίλης

Κεφάλαιο 1

Simple Linear Regression

1.1 Introduction

Έστω μία μεταβλητή ενδιαφέροντος Y η οποία εξαρτάται από μία άλλη μεταβλητή X. Προφανώς, η μεταβλητή X δίνει πληροφορία για την Y και στόχος είναι να χρησιμοποιηθεί για την πρόβλεψη των τιμών της Y. Για αυτόν τον λόγο, η μεταβλητή Y καλείται **εξαρτημένη** ή **αποκριτική** και η μεταβλητή X καλείται **ανεξάρτητη**, **επεξηγηματική** ή **προβλεπτική**. Με βάση ένα δείγμα μεγέθους n από ζευγαρωτές παρατηρήσεις $(Y_1, X_1), (Y_2, X_2), \ldots, (Y_n, X_n)$, κατασκευάζουμε το λεγόμενο **απλό γραμμικό μοντέλο** ή μοντέλο **απλής γραμμικής παλινδρόμησης**

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, 2, \ldots, n,$$

όπου β_0, β_1 οι **συντελεστές παλινδρόμησης** και ε_i τα **τυχαία σφάλματα** με:

- $E(\varepsilon_i) = 0$,
- $\mathrm{Var}(\varepsilon_i) = \sigma^2$, όπου σ^2 άγνωστη παράμετρος,
- $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0$ για $i \neq j$.

Με άλλα λόγια, τα τυχαία σφάλματα έχουν **μέση τιμή 0**, είναι **ομοσκεδαστικά**, δηλαδή έχουν κοινή διασπορά σ^2, και είναι **ασυσχέτιστα**, αλλά όχι απαραίτητα ανεξάρτητα. Επιπλέον, τα τυχαία σφάλματα είναι προφανώς μη-παρατηρήσιμα και μη-υπολογίσιμα. Για να είναι καλά ορισμένο το απλό γραμμικό μοντέλο, θα πρέπει, προφανώς, να μην είναι όλα τα X_i ίσα μεταξύ τους.

Αφού η X δεν είναι η μεταβλητή ενδιαφέροντος, τις τιμές $X_1, X_2, \ldots, X_n$ του δείγματος τις θεωρούμε γνωστές και σταθερές, ενώ τις τιμές $Y_1, Y_2, \ldots, Y_n$ τις θεωρούμε τυχαίες μεταβλητές, κατά τα γνωστά. Συνοψίζοντας, τα β_0, β_1, σ^2 είναι σταθερές και άγνωστες παράμετροι προς εκτίμηση, τα X_i είναι γνωστά και

σταθερά, ενώ τα Y_i και ε_i είναι τυχαίες μεταβλητές.

Πρόταση 1.1. (Ιδιότητες των Y_i και $\overline{Y}$)

i. $E(Y_i) = \beta_0 + \beta_1 X_i$.

ii. $\text{Var}(Y_i) = \sigma^2$.

iii. $\text{Cov}(Y_i, Y_j) = 0$ για $i \neq j$.

iv. $E\left(\overline{Y}\right) = \beta_0 + \beta_1 \overline{X}$.

v. $\text{Var}\left(\overline{Y}\right) = \frac{\sigma^2}{n}$.

Απόδειξη. Χρησιμοποιώντας τις ιδιότητες της μέσης τιμής, της διασποράς και της συνδιακύμανσης, υπολογίζουμε τα εξής:

i. $E(Y_i) = E(\beta_0 + \beta_1 X_i + \varepsilon_i) = \beta_0 + \beta_1 X_i + E(\varepsilon_i) = \beta_0 + \beta_1 X_i$.

ii. $\text{Var}(Y_i) = \text{Var}(\beta_0 + \beta_1 X_i + \varepsilon_i) = \text{Var}(\varepsilon_i) = \sigma^2$.

iii. $\text{Cov}(Y_i, Y_j) = \text{Cov}\left(\beta_0 + \beta_1 X_i + \varepsilon_i, \beta_0 + \beta_1 X_j + \varepsilon_j\right) = \text{Cov}(\varepsilon_i, \varepsilon_j) = 0$ για $i \neq j$.

iv. $E\left(\overline{Y}\right) = \frac{1}{n}\sum_{i=1}^{n} E(Y_i) = \frac{1}{n}\sum_{i=1}^{n}\left(\beta_0 + \beta_1 X_i\right) = \beta_0 + \frac{\beta_1}{n}\sum_{i=1}^{n} X_i = \beta_0 + \beta_1 \overline{X}$.

v. $\text{Var}\left(\overline{Y}\right) = \frac{1}{n^2}\text{Var}\left(\sum_{i=1}^{n} Y_i\right) \overset{\text{ασυσχ.}}{=} \frac{1}{n^2}\sum_{i=1}^{n}\text{Var}(Y_i) = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{\sigma^2}{n}$. □

Ερμηνεία: Η σχέση $E(Y) = \beta_0 + \beta_1 X$ δίνει τη λεγόμενη **εξίσωση παλινδρόμησης**. Στην περίπτωση του απλού γραμμικού μοντέλου, η εξίσωση αυτή ορίζει μία ευθεία στον $\mathbb{R}^2$, με την ανεξάρτητη μεταβλητή στον άξονα των x και την αναμενόμενη τιμή της εξαρτημένης μεταβλητής στον άξονα των y. Οι συντελεστές παλινδρόμησης β_0 και β_1 έχουν τις εξής ερμηνείες:

- Για $X = 0$, έχουμε $E(Y) = \beta_0 + \beta_1 \cdot 0 = \beta_0$, δηλαδή ο συντελεστής β_0 εκφράζει την αναμενόμενη τιμή της εξαρτημένης μεταβλητής για τιμή της ανεξάρτητης μεταβλητής ίση με το 0. Γεωμετρικά, είναι το σημείο στο οποίο η ευθεία παλινδρόμησης τέμνει τον άξονα των y και καλείται **σταθερός όρος** του γραμμικού μοντέλου. Προφανώς, η παράμετρος β_0 μετριέται στην ίδια μονάδα με την εξαρτημένη μεταβλητή Y.

- Για $X = X_0 + 1$, έχουμε $E(Y) = \beta_0 + \beta_1 X = \beta_0 + \beta_1 X_0 + \beta_1 = E(Y_0) + \beta_1 \Rightarrow \beta_1 = E(Y) - E(Y_0)$, δηλαδή ο συντελεστής β_1 εκφράζει τη μεταβολή στην αναμενόμενη τιμή της εξαρτημένης μεταβλητής για αύξηση της ανεξάρτητης μεταβλητής κατά μία μονάδα. Γεωμετρικά, είναι ο **συντελεστής κλίσης** της ευθείας παλινδρόμησης, δηλαδή η εφαπτομένη της γωνίας που σχηματίζει με τον άξονα των x. Η παράμετρος β_1 μετριέται σε μονάδα της εξαρτημένης μεταβλητής Y ανά μονάδα της ανεξάρτητης μεταβλητής X.

Το μοντέλο καλείται απλό, επειδή κάνει χρήση μόνο μίας επεξηγηματικής μεταβλητής X. Σε αντίθετη περίπτωση, καλείται πολλαπλό και θα μελετηθεί στο επόμενο κεφάλαιο. Επιπλέον, καλείται γραμμικό, επειδή η εξίσωση παλινδρόμησης είναι γραμμική συνάρτηση των συντελεστών παλινδρόμησης. Για παράδειγμα, το μοντέλο $Y_i = \beta_0 + \beta_1 \log X_{1i} + \beta_2 X_{2i}^2 + \varepsilon_i$ αποτελεί ένα πολλαπλό γραμμικό μοντέλο με δύο επεξηγηματικές μεταβλητές, ενώ το μοντέλο $Y_i = \frac{\beta_0 X_i}{\beta_1 + X_i} + \varepsilon_i$ δεν είναι γραμμικό.

1.2 Μέθοδος Ελαχίστων Τετραγώνων

Η μέθοδος ελαχίστων τετραγώνων είναι μία μέθοδος σημειακής εκτίμησης, όπως η μέθοδος μέγιστης πιθανοφάνειας και η μέθοδος των ροπών, που χρησιμοποιείται στην ανάλυση παλινδρόμησης. Έχει ως στόχο να προσδιορίσει εκείνη την ευθεία για την οποία το άθροισμα των τετραγωνικών αποστάσεων των παρατηρήσεων Y_i από αυτή παίρνει την ελάχιστη δυνατή τιμή. Με άλλα λόγια, στόχος της είναι να προσδιορίσει τις τιμές των παραμέτρων β_0 και β_1 που ελαχιστοποιούν τη συνάρτηση:

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

Η μέθοδος ελαχίστων τετραγώνων δεν απαιτεί καμία γνώση για την κατανομή των παρατηρήσεων Y_i, σε αντίθεση με τη μέθοδο μέγιστης πιθανοφάνειας. Αρκεί να έχουμε κάνει υπόθεση για την μέση τιμή των παρατηρήσεων $Y_1, Y_2, \ldots, Y_n$.

Λήμμα 1.1. Ισχύει ότι $\sum_{i=1}^{n}(X_i - \overline{X}) = 0$. Ομοίως, $\sum_{i=1}^{n}(Y_i - \overline{Y}) = 0$.

Απόδειξη. Προφανώς, έχουμε ότι:

$$\sum_{i=1}^{n}(X_i - \overline{X}) = \sum_{i=1}^{n} X_i - n\overline{X} = \sum_{i=1}^{n} X_i - \sum_{i=1}^{n} X_i = 0. \quad \square$$

Ορισμός 1.1. (Δειγματικά Περιγραφικά Στοιχεία)

- Δειγματική διασπορά της X:

$$S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})X_i = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i - n\overline{X}^2\right).$$

- Δειγματική διασπορά της Y:

$$S_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})Y_i = \frac{1}{n-1}\left(\sum_{i=1}^{n} Y_i - n\overline{Y}^2\right).$$

- Δειγματική συνδιακύμανση μεταξύ X και Y:

$$S_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}) = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i Y_i - n\overline{X}\cdot\overline{Y}\right)$$
$$= \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})Y_i = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})X_i.$$

- Δειγματικός συντελεστής συσχέτισης του Pearson μεταξύ X και Y:

$$r_{XY} = \frac{S_{XY}}{S_X \cdot S_Y} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2 \cdot \sum_{i=1}^{n}(Y_i - \overline{Y})^2}}.$$

Λήμμα 1.2. (Θετικά Ορισμένοι Πίνακες)

i. Ένας συμμετρικός πίνακας $A = \begin{bmatrix} a & b \\ b & c \end{bmatrix} \in \mathbb{R}^{2\times 2}$ είναι θετικά (αρνητικά) ορισμένος αν και μόνο αν $a > 0$ $(a < 0)$ και $\det A > 0$.

ii. Έστω $f : \mathbb{R}^d \to \mathbb{R}$ μία συνάρτηση πολλών μεταβλητών. Ορίζουμε

$$H(\mathbf{x}) = \left[\frac{\partial^2 f}{\partial x_i \partial x_j}\right]_{i,j} \in \mathbb{R}^{d\times d}$$

τον **Εσσιανό πίνακα** της f για $\mathbf{x} = (x_1, x_2, \ldots, x_d) \in \mathbb{R}^d$. Η συνάρτηση f είναι γνησίως κυρτή αν και μόνο αν ο πίνακας H είναι θετικά ορισμένος $\forall \mathbf{x} \in \mathbb{R}^d$.

Πρόταση 1.2. Οι **εκτιμήτριες ελαχίστων τετραγώνων** των παραμέτρων β_0 και β_1 του απλού γραμμικού μοντέλου δίνονται από τις σχέσεις:

$$\widehat{\beta}_1 = \frac{1}{(n-1)S_X^2}\sum_{i=1}^{n}(X_i - \overline{X})Y_i = \frac{S_{XY}}{S_X^2} = r_{XY}\cdot\frac{S_Y}{S_X},$$

$$\widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1\overline{X}.$$

Απόδειξη. Σύμφωνα με τη μέθοδο ελαχίστων τετραγώνων, ελαχιστοποιούμε τη συνάρτηση $Q(\beta_0, \beta_1) = \sum_{i=1}^{n}\varepsilon_i^2 = \sum_{i=1}^{n}\left[Y_i - E(Y_i)\right]^2 = \sum_{i=1}^{n}\left(Y_i - \beta_0 - \beta_1 X_i\right)^2$. Για τον λόγο αυτό, μηδενίζουμε τις μερικές παραγώγους της συνάρτησης Q:

$$\frac{\partial Q}{\partial \beta_0} = -2\sum_{i=1}^{n}\left(Y_i - \beta_0 - \beta_1 X_i\right) = 0 \Rightarrow \sum_{i=1}^{n} Y_i = n\widehat{\beta}_0 + \widehat{\beta}_1\sum_{i=1}^{n} X_i,$$

$$\frac{\partial Q}{\partial \beta_1} = -2\sum_{i=1}^{n} X_i\left(Y_i - \beta_0 - \beta_1 X_i\right) = 0 \Rightarrow \sum_{i=1}^{n} X_i Y_i = \widehat{\beta}_0\sum_{i=1}^{n} X_i + \widehat{\beta}_1\sum_{i=1}^{n} X_i^2.$$

Οι δύο εξισώσεις στις οποίες καταλήξαμε λέγονται **κανονικές εξισώσεις**. Λύνουμε την πρώτη κανονική εξίσωση ως προς $\widehat{\beta}_0$ και αντικαθιστούμε στη δεύτερη για να υπολογίσουμε την εκτιμήτρια $\widehat{\beta}_1$. Επομένως, παίρνουμε $\widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1\overline{X}$ και

$$\sum_{i=1}^{n} X_i Y_i = \overline{Y}\sum_{i=1}^{n} X_i - \widehat{\beta}_1\overline{X}\sum_{i=1}^{n} X_i + \widehat{\beta}_1\sum_{i=1}^{n} X_i^2 \Rightarrow$$

$$\left(\sum_{i=1}^{n} X_i^2 - n\overline{X}^2\right)\widehat{\beta}_1 = \sum_{i=1}^{n} X_i Y_i - n\overline{X}\cdot\overline{Y} \overset{\text{Ορισμός 1.1}}{\Rightarrow}$$

$$\widehat{\beta}_1 = \frac{1}{(n-1)S_X^2}\sum_{i=1}^{n}(X_i - \overline{X})Y_i = \frac{S_{XY}}{S_X^2} = \frac{r_{XY}\cdot S_X\cdot S_Y}{S_X^2} = r_{XY}\cdot\frac{S_Y}{S_X}.$$

Τέλος, υπολογίζουμε τον Εσσιανό πίνακα της συνάρτησης $Q(\beta_0, \beta_1)$ για να επαληθεύσουμε ότι το σημείο $(\widehat{\beta}_0, \widehat{\beta}_1)$ είναι το μοναδικό ολικό ελάχιστό της:

$$\frac{\partial^2 Q}{\partial\beta_0^2} = 2n, \quad \frac{\partial^2 Q}{\partial\beta_1^2} = 2\sum_{i=1}^{n} X_i^2, \quad \frac{\partial^2 Q}{\partial\beta_0\partial\beta_1} = 2\sum_{i=1}^{n} X_i \Rightarrow$$

$$H(\beta_0,\beta_1) = \begin{bmatrix} 2n & 2\sum\limits_{i=1}^{n} X_i \\ 2\sum\limits_{i=1}^{n} X_i & 2\sum\limits_{i=1}^{n} X_i^2 \end{bmatrix} = 2n\begin{bmatrix} 1 & \overline{X} \\ \overline{X} & \frac{1}{n}\sum\limits_{i=1}^{n} X_i^2 \end{bmatrix} \Rightarrow$$

$$\det H = 4n^2\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2\right) = 4n\left(\sum_{i=1}^{n} X_i^2 - n\overline{X}^2\right) = 4n(n-1)S_X^2 > 0$$

$$\text{και } H_{11} = 2n > 0,\ \forall(\beta_0,\beta_1)\in\mathbb{R}^2 \overset{\text{Λήμμα 1.2}}{\Rightarrow} H \text{ θετικά ορισμένος.}$$

Αφού ο Εσσιανός πίνακας της $Q(\beta_0, \beta_1)$ είναι θετικά ορισμένος, η $Q(\beta_0, \beta_1)$ είναι γνησίως κυρτή και έχει μοναδικό ολικό ελάχιστο το $(\widehat{\beta}_0, \widehat{\beta}_1)$. □

Σημείωση 1.1. Παρατηρούμε τα εξής σχετικά με τις εκτιμήτριες $\widehat{\beta}_0$ και $\widehat{\beta}_1$:

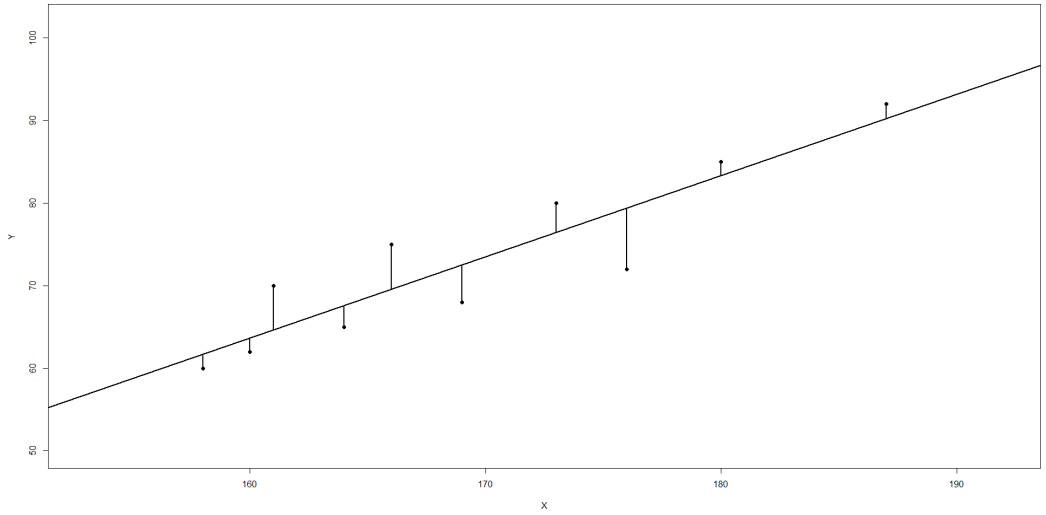
- Οι εκτιμήτριες $\widehat{\beta}_0$ και $\widehat{\beta}_1$ είναι **γραμμικοί συνδυασμοί** των Y_i. Συγκεκριμένα,

$$\widehat{\beta}_1 = \sum_{i=1}^{n} k_i Y_i \text{ και } \widehat{\beta}_0 = \frac{1}{n}\sum_{i=1}^{n} Y_i - \overline{X}\sum_{i=1}^{n} k_i Y_i = \sum_{i=1}^{n}\left(\frac{1}{n} - \overline{X}k_i\right)Y_i, \text{ όπου:}$$

$$k_i = \frac{X_i - \overline{X}}{(n-1)S_X^2}.$$

- Ο δειγματικός συντελεστής συσχέτισης r_{XY} είναι ελεύθερος μονάδων, η δειγματική τυπική απόκλιση S_X μετριέται στην ίδια μονάδα με το X και η δειγματική τυπική απόκλιση S_Y μετριέται στην ίδια μονάδα με το Y, οπότε επιβεβαιώνουμε ότι το $\widehat{\beta}_1$ μετριέται σε μονάδα της Y ανά μονάδα της X.

- Οι τιμές $\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i = \overline{Y} + \widehat{\beta}_1(X_i - \overline{X})$ καλούνται **προσαρμοσμένες τιμές** των Y_i. Γραφικά, είναι οι κατακόρυφες προβολές των Y_i πάνω στην εκτιμημένη ευθεία παλινδρόμησης.
- Η εξίσωση $\widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 X$ δίνει την **εκτιμημένη ευθεία παλινδρόμησης**.
- Οι τιμές $\widehat{\varepsilon}_i = Y_i - \widehat{Y}_i = Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i = Y_i - \overline{Y} - \widehat{\beta}_1(X_i - \overline{X})$ καλούνται **κατάλοιπα** ή **εκτιμημένα σφάλματα** της παλινδρόμησης. Γραφικά, είναι κατ' απόλυτη τιμή οι κατακόρυφες αποστάσεις των Y_i από την εκτιμημένη ευθεία παλινδρόμησης.
- Το παρακάτω γράφημα καλείται **σημειόγραμμα** ή **γράφημα διασποράς** των παρατηρήσεων (X_i, Y_i). Πάνω στο γράφημα έχουμε σχεδιάσει την εκτιμημένη ευθεία παλινδρόμησης που υπολογίζεται μέσω της μεθόδου ελαχίστων τετραγώνων με βάση αυτά τα δεδομένα. Οι κατακόρυφες γραμμές αντιπροσωπεύουν τα κατάλοιπα της παλινδρόμησης, δηλαδή τις αποστάσεις των Y_i από τα $\widehat{Y}_i$, τα οποία βρίσκονται πάνω στην ευθεία.



Σχημα 1.1: Σημειόγραμμα ή Διάγραμμα Διασποράς

- Εφόσον $S_X > 0$ και $S_Y > 0$, συμπεραίνουμε ότι το πρόσημο της εκτιμήτριας $\widehat{\beta}_1$ καθορίζεται από το πρόσημο του δειγματικού συντελεστή συσχέτισης του Pearson μεταξύ X και Y. Για παράδειγμα, αν υπάρχει θετική συσχέτιση μεταξύ X και Y, αυτό σημαίνει ότι όσο αυξάνεται το X τόσο αυξάνεται και το Y, οπότε αναμένουμε, προφανώς, ότι και η κλίση της εκτιμημένης ευθείας παλινδρόμησης θα είναι θετική.
- Το σημείο $\left(\overline{X}, \overline{Y}\right)$ καλείται το **κέντρο βάρους** του "σύννεφου" των παρατηρήσεων (X_i, Y_i) στον $\mathbb{R}^2$. Σύμφωνα με την προηγούμενη πρόταση, παρατηρούμε ότι $\overline{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 \overline{X}$, δηλαδή το κέντρο βάρους $\left(\overline{X}, \overline{Y}\right)$ ανήκει πάντα

στην εκτιμημένη ευθεία παλινδρόμησης.

Λήμμα 1.3. (Ιδιότητες των k_i)

i. $\sum_{i=1}^{n} k_i = 0.$

ii. $\sum_{i=1}^{n} k_i X_i = 1.$

iii. $\sum_{i=1}^{n} k_i^2 = \frac{1}{(n-1)S_X^2}.$

Απόδειξη. Με απλές πράξεις υπολογίζουμε τα εξής:

i. $\sum_{i=1}^{n} k_i = \frac{1}{(n-1)S_X^2} \sum_{i=1}^{n}(X_i - \overline{X}) \overset{\text{Λήμμα 1.1}}{=} 0.$

ii. $\sum_{i=1}^{n} k_i X_i = \frac{1}{(n-1)S_X^2} \sum_{i=1}^{n}(X_i - \overline{X})X_i = \frac{1}{(n-1)S_X^2}\left(\sum_{i=1}^{n} X_i^2 - \overline{X}\sum_{i=1}^{n} X_i\right) \Rightarrow$

$\sum_{i=1}^{n} k_i X_i = \frac{1}{(n-1)S_X^2}\left(\sum_{i=1}^{n} X_i^2 - n\overline{X}^2\right) = \frac{(n-1)S_X^2}{(n-1)S_X^2} = 1.$

iii. $\sum_{i=1}^{n} k_i^2 = \frac{1}{\left[(n-1)S_X^2\right]^2} \sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{(n-1)S_X^2}{\left[(n-1)S_X^2\right]^2} = \frac{1}{(n-1)S_X^2}.$ □

Πρόταση 1.3. (Εκτιμήτριες Ελαχίστων Τετραγώνων)

i. $E\left(\widehat{\beta}_1\right) = \beta_1$, δηλαδή η $\widehat{\beta}_1$ είναι **αμερόληπτη εκτιμήτρια** του β_1.

ii. $E\left(\widehat{\beta}_0\right) = \beta_0$, δηλαδή η $\widehat{\beta}_0$ είναι **αμερόληπτη εκτιμήτρια** του β_0.

iii. $\text{Cov}\left(\widehat{\beta}_1, Y_i\right) = k_i\sigma^2 = \sigma^2\frac{X_i - \overline{X}}{(n-1)S_X^2}.$

iv. $\text{Cov}\left(\widehat{\beta}_1, \overline{Y}\right) = 0.$

v. $\text{Cov}\left(\widehat{\beta}_0, Y_i\right) = \sigma^2\left(\frac{1}{n} - k_i\overline{X}\right) = \sigma^2\left[\frac{1}{n} - \frac{X_i - \overline{X}}{(n-1)S_X^2}\overline{X}\right].$

vi. $\text{Cov}\left(\widehat{\beta}_0, \overline{Y}\right) = \text{Var}\left(\overline{Y}\right) = \frac{\sigma^2}{n}.$

vii. $\text{Var}\left(\widehat{\beta}_1\right) = \sigma^2\sum_{i=1}^{n} k_i^2 = \frac{\sigma^2}{(n-1)S_X^2}.$

viii. $\text{Var}\left(\widehat{\beta}_0\right) = \sigma^2\left(\frac{1}{n} + \overline{X}^2\sum_{i=1}^{n} k_i^2\right) = \sigma^2\left[\frac{1}{n} + \frac{\overline{X}^2}{(n-1)S_X^2}\right] = \frac{\sigma^2}{n(n-1)S_X^2}\sum_{i=1}^{n} X_i^2.$

ix. $\text{Cov}\left(\widehat{\beta}_0, \widehat{\beta}_1\right) = -\overline{X}\sigma^2\sum_{i=1}^{n} k_i^2 = -\frac{\overline{X}\sigma^2}{(n-1)S_X^2}.$

Απόδειξη. Χρησιμοποιώντας τις ιδιότητες της μέσης τιμής, της διασποράς και της συνδιακύμανσης, υπολογίζουμε τα εξής:

i. $E\left(\widehat{\beta}_1\right) = E\left(\sum_{i=1}^{n} k_i Y_i\right) = \sum_{i=1}^{n} k_i E(Y_i) = \sum_{i=1}^{n} k_i\left(\beta_0 + \beta_1 X_i\right) \Rightarrow$

$$E\left(\widehat{\beta}_1\right) = \beta_0 \cancelto{0}{\sum_{i=1}^{n} k_i} + \beta_1 \cancelto{1}{\sum_{i=1}^{n} k_i X_i} = \beta_1.$$

ii. $E\left(\widehat{\beta}_0\right) = E\left(\overline{Y} - \widehat{\beta}_1\overline{X}\right) = E\left(\overline{Y}\right) - \overline{X}E\left(\widehat{\beta}_1\right) \overset{\text{Πρόταση 1.1}}{=} \beta_0 + \beta_1\overline{X} - \overline{X}\beta_1 = \beta_0.$

iii. $\text{Cov}\left(\widehat{\beta}_1, Y_i\right) = \text{Cov}\left(\sum_{j=1}^{n} k_j Y_j, Y_i\right) \overset{\text{ασυσχ.}}{=} \text{Cov}\left(k_i Y_i, Y_i\right) = k_i \text{Cov}\left(Y_i, Y_i\right) \Rightarrow$

$\text{Cov}\left(\widehat{\beta}_1, Y_i\right) = k_i \text{Var}\left(Y_i\right) = k_i\sigma^2 = \sigma^2 \frac{X_i - \overline{X}}{(n-1)S_X^2}.$

iv. $\text{Cov}\left(\widehat{\beta}_1, \overline{Y}\right) = \frac{1}{n} \cdot \text{Cov}\left(\widehat{\beta}_1, \sum_{i=1}^{n} Y_i\right) = \frac{1}{n}\sum_{i=1}^{n} \text{Cov}\left(\widehat{\beta}_1, Y_i\right) = \frac{1}{n}\sum_{i=1}^{n} k_i\sigma^2 \Rightarrow$

$\text{Cov}\left(\widehat{\beta}_1, \overline{Y}\right) = \frac{\sigma^2}{n} \cancelto{0}{\sum_{i=1}^{n} k_i} = 0.$

v. $\text{Cov}\left(\widehat{\beta}_0, Y_i\right) = \text{Cov}\left[\sum_{j=1}^{n} \left(\frac{1}{n} - \overline{X}k_j\right) Y_j, Y_i\right] \overset{\text{ασυσχ.}}{=} \text{Cov}\left[\left(\frac{1}{n} - \overline{X}k_i\right) Y_i, Y_i\right] \Rightarrow$

$\text{Cov}\left(\widehat{\beta}_0, Y_i\right) = \left(\frac{1}{n} - \overline{X}k_i\right) \text{Cov}(Y_i, Y_i) = \left(\frac{1}{n} - \overline{X}k_i\right) \text{Var}(Y_i) = \sigma^2\left(\frac{1}{n} - \overline{X}k_i\right) \Rightarrow$

$\text{Cov}\left(\widehat{\beta}_0, Y_i\right) = \sigma^2\left[\frac{1}{n} - \frac{X_i - \overline{X}}{(n-1)S_X^2}\overline{X}\right].$

vi. $\text{Cov}\left(\widehat{\beta}_0, \overline{Y}\right) = \text{Cov}\left(\overline{Y} - \widehat{\beta}_1\overline{X}, \overline{Y}\right) = \text{Cov}\left(\overline{Y}, \overline{Y}\right) - \overline{X}\cancelto{0}{\text{Cov}\left(\widehat{\beta}_1, \overline{Y}\right)} \Rightarrow$

$\text{Cov}\left(\widehat{\beta}_0, \overline{Y}\right) = \text{Var}\left(\overline{Y}\right) = \frac{\sigma^2}{n}.$

vii. $\text{Var}\left(\widehat{\beta}_1\right) = \text{Var}\left(\sum_{i=1}^{n} k_i Y_i\right) \overset{\text{ασυσχ.}}{=} \sum_{i=1}^{n} \text{Var}(k_i Y_i) = \sum_{i=1}^{n} k_i^2 \text{Var}(Y_i) \Rightarrow$

$\text{Var}\left(\widehat{\beta}_1\right) = \sum_{i=1}^{n} k_i^2\sigma^2 = \sigma^2 \sum_{i=1}^{n} k_i^2 \overset{\text{Λήμμα 1.3}}{=} \frac{\sigma^2}{(n-1)S_X^2}.$

viii. $\text{Var}\left(\widehat{\beta}_0\right) = \text{Var}\left(\overline{Y} - \widehat{\beta}_1\overline{X}\right) = \text{Var}\left(\overline{Y}\right) + \text{Var}\left(\widehat{\beta}_1\overline{X}\right) - 2\text{Cov}\left(\overline{Y}, \widehat{\beta}_1\overline{X}\right) \overset{\text{Πρόταση 1.1}}{\Rightarrow}$

$\text{Var}\left(\widehat{\beta}_0\right) = \frac{\sigma^2}{n} + \overline{X}^2\text{Var}\left(\widehat{\beta}_1\right) - 2\overline{X}\cancelto{0}{\text{Cov}\left(\overline{Y}, \widehat{\beta}_1\right)} = \sigma^2\left(\frac{1}{n} + \overline{X}^2\sum_{i=1}^{n} k_i^2\right) \Rightarrow$

$\text{Var}\left(\widehat{\beta}_0\right) = \sigma^2\left[\frac{1}{n} + \frac{\overline{X}^2}{(n-1)S_X^2}\right] = \frac{\sigma^2}{n(n-1)S_X^2}\left[(n-1)S_X^2 + n\overline{X}^2\right] \Rightarrow$

$\text{Var}\left(\widehat{\beta}_0\right) = \frac{\sigma^2}{n(n-1)S_X^2}\left(\sum_{i=1}^{n} X_i^2 - n\overline{X}^2 + n\overline{X}^2\right) = \frac{\sigma^2}{n(n-1)S_X^2}\sum_{i=1}^{n} X_i^2.$

ix. $\text{Cov}\left(\widehat{\beta}_0, \widehat{\beta}_1\right) = \text{Cov}\left(\overline{Y} - \widehat{\beta}_1\overline{X}, \widehat{\beta}_1\right) = \cancelto{0}{\text{Cov}\left(\overline{Y}, \widehat{\beta}_1\right)} - \overline{X}\text{Cov}\left(\widehat{\beta}_1, \widehat{\beta}_1\right) \Rightarrow$

$\text{Cov}\left(\widehat{\beta}_0, \widehat{\beta}_1\right) = -\overline{X}\text{Var}\left(\widehat{\beta}_1\right) = -\overline{X}\sigma^2\sum_{i=1}^{n} k_i^2 = -\frac{\overline{X}\sigma^2}{(n-1)S_X^2}.$ □

Παρατήρηση 1.1. Παρατηρούμε ότι το πρόσημο της συνδιακύμανσης μεταξύ των εκτιμητριών $\widehat{\beta}_0$ και $\widehat{\beta}_1$ καθορίζεται πλήρως από το πρόσημο του $\overline{X}$.

Θεώρημα 1.1. (Gauss - Markov) Οι εκτιμήτριες ελαχίστων τετραγώνων $\widehat{\beta}_0$ και $\widehat{\beta}_1$ έχουν την ελάχιστη διασπορά ανάμεσα σε όλες τις αμερόληπτες εκτιμήτριες των β_0 και β_1 οι οποίες είναι γραμμικές συναρτήσεις των Y_i.

Απόδειξη. Βλέπε θεώρημα 2.1 (σελίδα 46).

Πρόταση 1.4. (Προσαρμοσμένες Τιμές και Κατάλοιπα)

i. $E(\widehat{Y}_i) = E(Y_i) = \beta_0 + \beta_1 X_i$, δηλαδή το $\widehat{Y}_i$ είναι μία αμερόληπτη εκτιμήτρια της $E(Y_i)$. Συμπεραίνουμε ότι $E(\widehat{\varepsilon}_i) = 0$.

ii. $\text{Cov}\left(\widehat{Y}_i, \widehat{Y}_j\right) = \text{Cov}\left(\widehat{Y}_i, Y_j\right) = \sigma^2\left[\frac{1}{n} + \frac{(X_i-\overline{X})(X_j-\overline{X})}{(n-1)S_X^2}\right]$. Ειδικότερα,
$\text{Var}\left(\widehat{Y}_i\right) = \sigma^2\left[\frac{1}{n} + \frac{(X_i-\overline{X})^2}{(n-1)S_X^2}\right]$.

iii. $\text{Cov}\left(\widehat{Y}_i, \widehat{\varepsilon}_j\right) = 0$.

iv. Για $i \neq j$, $\text{Cov}\left(Y_i, \widehat{\varepsilon}_j\right) = \text{Cov}\left(\widehat{\varepsilon}_i, \widehat{\varepsilon}_j\right) = -\sigma^2\left[\frac{1}{n} + \frac{(X_i-\overline{X})(X_j-\overline{X})}{(n-1)S_X^2}\right]$. Διαφορετικά,
$\text{Cov}\left(Y_i, \widehat{\varepsilon}_i\right) = \text{Var}\left(\widehat{\varepsilon}_i\right) = E\left(\widehat{\varepsilon}_i^2\right) = \sigma^2\left[1 - \frac{1}{n} - \frac{(X_i-\overline{X})^2}{(n-1)S_X^2}\right]$.

v. $\sum_{i=1}^{n} \widehat{Y}_i = \sum_{i=1}^{n} Y_i$. Συμπεραίνουμε ότι $\sum_{i=1}^{n} \widehat{\varepsilon}_i = 0$.

vi. $\sum_{i=1}^{n} X_i\widehat{\varepsilon}_i = 0$. Συμπεραίνουμε ότι $\sum_{i=1}^{n} \widehat{Y}_i\widehat{\varepsilon}_i = 0$.

Απόδειξη. Χρησιμοποιώντας τις ιδιότητες της μέσης τιμής, της διασποράς και της συνδιακύμανσης, υπολογίζουμε τα εξής:

i. $E\left(\widehat{Y}_i\right) = E\left(\widehat{\beta}_0 + \widehat{\beta}_1 X_i\right) = E\left(\widehat{\beta}_0\right) + X_i E\left(\widehat{\beta}_1\right) = \beta_0 + \beta_1 X_i = E(Y_i) \Rightarrow$
$E\left(\widehat{\varepsilon}_i\right) = E\left(Y_i - \widehat{Y}_i\right) = E\left(Y_i\right) - E\left(\widehat{Y}_i\right) = 0.$

ii. $\text{Cov}\left(\widehat{Y}_i, \widehat{Y}_j\right) = \text{Cov}\left(\widehat{\beta}_0 + \widehat{\beta}_1 X_i, \widehat{\beta}_0 + \widehat{\beta}_1 X_j\right) \Rightarrow$
$\text{Cov}\left(\widehat{Y}_i, \widehat{Y}_j\right) = \text{Var}\left(\widehat{\beta}_0\right) + X_i X_j \text{Var}\left(\widehat{\beta}_1\right) + (X_i + X_j)\text{Cov}\left(\widehat{\beta}_0, \widehat{\beta}_1\right) \overset{\text{Πρόταση 1.3}}{\Rightarrow}$
$\text{Cov}\left(\widehat{Y}_i, \widehat{Y}_j\right) = \sigma^2\left[\frac{1}{n} + \frac{\overline{X}^2}{(n-1)S_X^2} + \frac{X_i X_j}{(n-1)S_X^2} - \frac{(X_i+X_j)\overline{X}}{(n-1)S_X^2}\right] \Rightarrow$
$\text{Cov}\left(\widehat{Y}_i, \widehat{Y}_j\right) = \sigma^2\left[\frac{1}{n} + \frac{(X_i-\overline{X})(X_j-\overline{X})}{(n-1)S_X^2}\right]$. Επιπλέον,
$\text{Cov}\left(\widehat{Y}_i, Y_j\right) = \text{Cov}\left(\widehat{\beta}_0 + \widehat{\beta}_1 X_i, Y_j\right) = \text{Cov}\left(\widehat{\beta}_0, Y_j\right) + X_i \text{Cov}\left(\widehat{\beta}_1, Y_j\right) \overset{\text{Πρόταση 1.3}}{\Rightarrow}$
$\text{Cov}\left(\widehat{Y}_i, Y_j\right) = \sigma^2\left[\frac{1}{n} - \frac{X_j-\overline{X}}{(n-1)S_X^2}\overline{X} + X_i\frac{X_j-\overline{X}}{(n-1)S_X^2}\right] = \sigma^2\left[\frac{1}{n} + \frac{(X_i-\overline{X})(X_j-\overline{X})}{(n-1)S_X^2}\right]$.

iii. $\text{Cov}\left(\widehat{Y}_i, \widehat{\varepsilon}_j\right) = \text{Cov}\left(\widehat{Y}_i, Y_j - \widehat{Y}_j\right) = \cancel{\text{Cov}\left(\widehat{Y}_i, Y_j\right)} - \cancel{\text{Cov}\left(\widehat{Y}_i, \widehat{Y}_j\right)} = 0.$

iv. $\text{Cov}\left(Y_i, \widehat{\varepsilon}_j\right) = \text{Cov}\left(Y_i, Y_j - \widehat{Y}_j\right) = \text{Cov}\left(Y_i, Y_j\right) - \text{Cov}\left(Y_i, \widehat{Y}_j\right).$
Για $i \neq j$, $\text{Cov}\left(Y_i, Y_j\right) = 0$, οπότε $\text{Cov}\left(Y_i, \widehat{\varepsilon}_j\right) = -\text{Cov}\left(Y_i, \widehat{Y}_j\right) \Rightarrow$

$\mathrm{Cov}\left(Y_i, \widehat{\varepsilon}_j\right) = -\sigma^2\left[\frac{1}{n} + \frac{(X_i-\overline{X})(X_j-\overline{X})}{(n-1)S_X^2}\right]$. Επιπλέον,

$\mathrm{Cov}\left(\widehat{\varepsilon}_i, \widehat{\varepsilon}_j\right) = \mathrm{Cov}\left(Y_i - \widehat{Y}_i, \widehat{\varepsilon}_j\right) = \mathrm{Cov}\left(Y_i, \widehat{\varepsilon}_j\right) - \cancelto{0}{\mathrm{Cov}\left(\widehat{Y}_i, \widehat{\varepsilon}_j\right)} \Rightarrow$

$\mathrm{Cov}\left(\widehat{\varepsilon}_i, \widehat{\varepsilon}_j\right) = -\sigma^2\left[\frac{1}{n} + \frac{(X_i-\overline{X})(X_j-\overline{X})}{(n-1)S_X^2}\right]$.

Για $i = j$, $\mathrm{Cov}\left(Y_i, Y_i\right) = \mathrm{Var}\left(Y_i\right) = \sigma^2$, οπότε:

$\mathrm{Cov}\left(Y_i, \widehat{\varepsilon}_i\right) = \sigma^2 - \mathrm{Cov}\left(Y_i, \widehat{Y}_i\right) = \sigma^2\left[1 - \frac{1}{n} - \frac{(X_i-\overline{X})^2}{(n-1)S_X^2}\right]$. Επιπλέον,

$\mathrm{Var}\left(\widehat{\varepsilon}_i\right) = \mathrm{Cov}\left(\widehat{\varepsilon}_i, \widehat{\varepsilon}_i\right) = \mathrm{Cov}\left(Y_i - \widehat{Y}_i, \widehat{\varepsilon}_i\right) = \mathrm{Cov}\left(Y_i, \widehat{\varepsilon}_i\right) - \cancelto{0}{\mathrm{Cov}\left(\widehat{Y}_i, \widehat{\varepsilon}_i\right)} \Rightarrow$

$\mathrm{Var}\left(\widehat{\varepsilon}_i\right) = \sigma^2\left[1 - \frac{1}{n} - \frac{(X_i-\overline{X})^2}{(n-1)S_X^2}\right]$. Τέλος,

$E\left(\widehat{\varepsilon}_i^2\right) = \mathrm{Var}\left(\widehat{\varepsilon}_i\right) + \cancelto{0}{\left[E\left(\widehat{\varepsilon}_i\right)\right]^2} = \sigma^2\left[1 - \frac{1}{n} - \frac{(X_i-\overline{X})^2}{(n-1)S_X^2}\right]$.

v. $\overline{Y} = \widehat{\beta}_0 + \widehat{\beta}_1\overline{X} \Rightarrow \sum_{i=1}^{n} Y_i = n\widehat{\beta}_0 + \widehat{\beta}_1\sum_{i=1}^{n} X_i = \sum_{i=1}^{n}\left(\widehat{\beta}_0 + \widehat{\beta}_1 X_i\right) = \sum_{i=1}^{n}\widehat{Y}_i$. Άρα,

$\sum_{i=1}^{n}\widehat{\varepsilon}_i = \sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right) = \sum_{i=1}^{n} Y_i - \sum_{i=1}^{n}\widehat{Y}_i = 0.$

vi. $\sum_{i=1}^{n} X_i\widehat{\varepsilon}_i = \sum_{i=1}^{n} X_i\left[Y_i - \overline{Y} - \widehat{\beta}_1(X_i - \overline{X})\right] = \sum_{i=1}^{n} X_i(Y_i - \overline{Y}) - \widehat{\beta}_1\sum_{i=1}^{n} X_i(X_i - \overline{X}) \Rightarrow$

$\sum_{i=1}^{n} X_i\widehat{\varepsilon}_i \overset{\text{Ορισμός 1.1}}{=} (n-1)S_{XY} - (n-1)S_X^2\widehat{\beta}_1 \overset{\text{Πρόταση 1.2}}{=} 0$. Άρα,

$\sum_{i=1}^{n}\widehat{Y}_i\widehat{\varepsilon}_i = \sum_{i=1}^{n}\left(\widehat{\beta}_0 + \widehat{\beta}_1 X_i\right)\widehat{\varepsilon}_i = \widehat{\beta}_0\cancelto{0}{\sum_{i=1}^{n}\widehat{\varepsilon}_i} + \widehat{\beta}_1\cancelto{0}{\sum_{i=1}^{n} X_i\widehat{\varepsilon}_i} = 0.$ □

1.3 Εκτίμηση της Διασποράς

Αν $Y_1, Y_2, \ldots, Y_n$ τυχαίο δείγμα από κατανομή με **γνωστή** μέση τιμή μ και άγνωστη διασπορά σ^2, τότε μία αμερόληπτη εκτιμήτρια του σ^2 είναι η:

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left[Y_i - E(Y_i)\right]^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \mu)^2.$$

Αντιθέτως, αν $Y_1, Y_2, \ldots, Y_n$ τυχαίο δείγμα από κατανομή με **άγνωστη** μέση τιμή μ και άγνωστη διασπορά σ^2, τότε η αντίστοιχη αμερόληπτη εκτιμήτρια του σ^2 προκύπτει, αν αντικαταστήσουμε το άγνωστο, πλέον, μ από την αμερόληπτη εκτιμήτρια $\widehat{\mu} = \overline{Y}$ και αφαιρέσουμε **έναν** βαθμό ελευθερίας από τον παρονομαστή για τη μία παράμετρο που εκτιμήσαμε, δηλαδή το μ. Η αμερόληπτη εκτιμήτρια του σ^2 που προκύπτει είναι η λεγόμενη δειγματική διασπορά:

$$S_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2.$$

Επιστρέφοντας στο απλό γραμμικό μοντέλο, αν οι συντελεστές β_0 και β_1 της εξίσωσης παλινδρόμησης ήταν γνωστοί, τότε, με τον ίδιο τρόπο, ως αμερόληπτη εκτιμήτρια του σ^2 θα προέκυπτε η:

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left[Y_i - E(Y_i)\right]^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)^2.$$

Στη ρεαλιστική περίπτωση όπου οι συντελεστές β_0 και β_1 είναι άγνωστοι, τότε, με τον ίδιο τρόπο, η αντίστοιχη αμερόληπτη εκτιμήτρια του σ^2 προκύπτει, αν αντικαταστήσουμε τους δύο συντελεστές από τις αμερόληπτες εκτιμήτριες $\widehat{\beta}_0$ και $\widehat{\beta}_1$ και αφαιρέσουμε **δύο** βαθμούς ελευθερίας από τον παρονομαστή για τις δύο παραμέτρους που εκτιμήσαμε. Η αμερόληπτη εκτιμήτρια του σ^2 που προκύπτει είναι το λεγόμενο **μέσο τετραγωνικό σφάλμα** (mean squared error):

$$\text{MSE} = S^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left(Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i\right)^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2 = \frac{1}{n-2}\sum_{i=1}^{n}\widehat{\varepsilon}_i^2.$$

Πρόταση 1.5. Το μέσο τετραγωνικό σφάλμα MSE είναι μία αμερόλπητη εκτιμήτρια της διασποράς σ^2 στο απλό γραμμικό μοντέλο.

Απόδειξη. Χρησιμοποιώντας τις ιδιότητες της μέσης τιμής και της διασποράς, υπολογίζουμε ότι:

$$\begin{aligned} E\left(S^2\right) &= \frac{1}{n-2}\sum_{i=1}^{n}E\left(\widehat{\varepsilon}_i^2\right) \overset{\text{Πρόταση 1.4}}{=} \frac{1}{n-2}\sum_{i=1}^{n}\sigma^2\left[1 - \frac{1}{n} - \frac{(X_i - \overline{X})^2}{(n-1)S_X^2}\right] \\ &= \frac{\sigma^2}{n-2}\left[n - 1 - \frac{1}{\cancel{(n-1)S_X^2}}\cancel{\sum_{i=1}^{n}(X_i - \overline{X})^2}\right] = \frac{\sigma^2}{\cancel{n-2}}\cancel{(n-1-1)} = \sigma^2. \end{aligned}$$

Άρα, το S^2 είναι όντως μία αμερόληπτη εκτιμήτρια του σ^2. □

1.4 Συντελεστής Προσδιορισμού

Ένας από τους στόχους της γραμμικής παλινδρόμησης είναι να εξηγήσει ένα κομμάτι από τη συνολική μεταβλητότητα των δεδομένων, η οποία ποσοτικοποιείται μέσω του αθροίσματος τετραγώνων $\sum_{i=1}^{n}(Y_i - \overline{Y})^2$, το οποίο εμφανίζεται στη δειγματική διασπορά των παρατηρήσεων $Y_1, Y_2, \ldots, Y_n$.

Μπορούμε να δείξουμε ότι αυτή η συνολική μεταβλητότητα σπάει σε ένα κομμάτι που εξηγείται από το γραμμικό μοντέλο μέσω της απόκλισης των προσαρμοσμένων τιμών $\widehat{Y}_i$ από τον δειγματικό μέσο $\overline{Y}$ και σε ένα κομμάτι που παραμένει ανεξήγητο και ποσοτικοποιείται μέσω των καταλοίπων $\widehat{\varepsilon}_i$.

Η συνολική μεταβλητότητα των εκάστοτε δεδομένων που έχουμε στα χέρια μας είναι πάντα σταθερή και συγκεκριμένη, οπότε επιθυμία μας, όταν κατασκευάζουμε ένα γραμμικό μοντέλο, είναι να μεγιστοποιήσουμε το κομμάτι της μεταβλητότητας που εξηγείται από το μοντέλο ή, ισοδύναμα, να ελαχιστοποιήσουμε το κομμάτι της μεταβλητότητας που παραμένει ανεξήγητο.

Ορισμός 1.2. (Αθροίσματα Τετραγώνων)

- Ορίζουμε $\text{SST} = \sum_{i=1}^{n}(Y_i - \overline{Y})^2$ το συνολικό άθροισμα τετραγώνων (total sum of squares) των δεδομένων.
- Ορίζουμε $\text{SSR} = \sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2$ το άθροισμα τετραγώνων που οφείλεται στην παλινδρόμηση (sum of squares due to regression).
- Ορίζουμε $\text{SSE} = \sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2 = \sum_{i=1}^{n}\widehat{\varepsilon}_i^2$ το άθροισμα τετραγώνων των καταλοίπων (sum of squared errors).

Παρατήρηση 1.2. Παρατηρούμε ότι:

$$S^2 = \text{MSE} = \frac{\text{SSE}}{n-2}.$$

Πρόταση 1.6. (Ανάλυση Διασποράς)

$$\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2 + \sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2, \text{ δηλαδή } \text{SST} = \text{SSR} + \text{SSE}.$$

Απόδειξη. Έχοντας στο μυαλό μας μία τεχνική που εμφανίζεται και στον υπολογισμό του μέσου τετραγωνικού σφάλματος μίας εκτιμήτριας, σκεφτόμαστε να προσθαφαιρέσουμε το $\widehat{Y}_i$ στην ποσότητα $Y_i - \overline{Y}$ που εμφανίζεται στο SST:

$$\begin{aligned}
\text{SST} &= \sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2 = \sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i + \widehat{Y}_i - \overline{Y}\right)^2 \\
&= \sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2 + \sum_{i=1}^{n}\left(\widehat{Y}_i - \overline{Y}\right)^2 + 2\sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)\left(\widehat{Y}_i - \overline{Y}\right) \\
&= \text{SSE} + \text{SSR} + 2\sum_{i=1}^{n}\left(\widehat{Y}_i - \overline{Y}\right)\widehat{\varepsilon}_i = \text{SSE} + \text{SSR} + 2\cancelto{0}{\sum_{i=1}^{n}\widehat{Y}_i\widehat{\varepsilon}_i} - 2\overline{Y}\cancelto{0}{\sum_{i=1}^{n}\widehat{\varepsilon}_i} \\
&= \text{SSE} + \text{SSR}. \quad \square
\end{aligned}$$

Ορισμός 1.3. (Συντελεστής Προσδιορισμού)

$$R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}.$$

Ερμηνεία: Παρατηρούμε αρχικά ότι $0 \leqslant R^2 \leqslant 1$. Ο συντελεστής προσδιορισμού εκφράζει το ποσοστό της συνολικής μεταβλητότητας των δεδομένων που εξηγείται από το γραμμικό μοντέλο. Προφανώς, όσο πιο κοντά βρίσκεται στη μονάδα, τόσο καλύτερο είναι το γραμμικό μοντέλο που έχουμε κατασκευάσει. Αντιθέτως, όταν βρίσκεται κοντά στο μηδέν, αυτό σημαίνει ότι το γραμμικό μοντέλο εξηγεί ένα μικρό ποσοστό της μεταβλητότητας των δεδομένων, οπότε έχουμε εκτιμήσει χωρίς λόγο τον συντελεστή κλίσης της ευθείας παλινδρόμησης, ενώ θα μπορούσαμε να είχαμε σταθεί στην υπόθεση ότι τα $Y_1, Y_2, \ldots, Y_n$ είναι ισόνομα.

Λήμμα 1.4. Ισχύει ότι:

$$\text{SSR} = \widehat{\beta}_1^2 \sum_{i=1}^{n} (X_i - \overline{X})^2 = (n-1) S_X^2 \widehat{\beta}_1^2.$$

Απόδειξη. Χρησιμοποιώντας την έκφραση $\widehat{Y}_i = \overline{Y} + \widehat{\beta}_1 (X_i - \overline{X})$, βλέπουμε άμεσα το ζητούμενο:

$$\text{SSR} = \sum_{i=1}^{n} \left(\widehat{Y}_i - \overline{Y} \right)^2 = \sum_{i=1}^{n} \left(\overline{Y} + \widehat{\beta}_1 (X_i - \overline{X}) - \overline{Y} \right)^2 = \widehat{\beta}_1^2 \sum_{i=1}^{n} (X_i - \overline{X})^2. \quad \square$$

Πρόταση 1.7. (Σχέση μεταξύ R^2 και r_{XY})

$$R^2 = r_{XY}^2.$$

Απόδειξη. Κάνοντας χρήση του προηγούμενου λήμματος, παίρνουμε ότι:

$$R^2 = \frac{\text{SSR}}{\text{SST}} \overset{\text{Λήμμα 1.4}}{=} \widehat{\beta}_1^2 \cdot \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{\sum_{i=1}^{n} (Y_i - \overline{Y})^2} \overset{\text{Πρόταση 1.2}}{=} \left(r_{XY} \cdot \frac{S_Y}{S_X} \right)^2 \cdot \frac{(n-1) S_X^2}{(n-1) S_Y^2} = r_{XY}^2. \quad \square$$

Ερμηνεία: Σύμφωνα με την προηγούμενη πρόταση, όσο πιο ισχυρά συσχετισμένες μεταξύ τους είναι η εξαρτημένη και η ανεξάρτητη μεταβλητή σε ένα γραμμικό μοντέλο, τόσο μεγαλύτερος είναι ο συντελεστής προσδιορισμού. Επομένως, τόσο καλύτερο είναι και το γραμμικό μοντέλο το οποίο προκύπτει, αφού εξηγεί ένα μεγάλο ποσοστό από τη μεταβλητότητα της εξαρτημένης μεταβλητής.

1.5 Πρόβλεψη Καινούργιας Παρατήρησης

Έχοντας μία νέα παρατήρηση X_{n+1}, ενδιαφερόμαστε να κάνουμε μία πρόβλεψη για την τιμή $Y_{n+1} = \beta_0 + \beta_1 X_{n+1} + \varepsilon_{n+1}$ την οποία δεν έχουμε παρατηρήσει, όπου $E(\varepsilon_{n+1}) = 0$, $\text{Var}(\varepsilon_{n+1}) = \sigma^2$ και ε_{n+1} ασυσχέτιστο με τα $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$.

Με βάση τις προηγούμενες n παρατηρήσεις $Y_1, Y_2, \ldots, Y_n$, έχουμε υπολογίσει τις εκτιμήτριες ελαχίστων τετραγώνων $\widehat{\beta}_0$ και $\widehat{\beta}_1$, οπότε ορίζουμε φυσιολογικά $\widetilde{Y}_{n+1} = \widehat{\beta}_0 + \widehat{\beta}_1 X_{n+1}$. Παρατηρούμε ότι $E(\widetilde{Y}_{n+1}) = E(Y_{n+1}) = \beta_0 + \beta_1 X_{n+1}$, οπότε μπορούμε να χρησιμοποιήσουμε την τιμή $\widetilde{Y}_{n+1}$ για να προβλέψουμε την Y_{n+1}. Επιπλέον, ορίζουμε το **σφάλμα πρόβλεψης** $\widetilde{\varepsilon}_{n+1} = Y_{n+1} - \widetilde{Y}_{n+1}$.

Πρόταση 1.8. (Σφάλμα Πρόβλεψης)

i. $E\left(\widetilde{\varepsilon}_{n+1}\right) = 0$.

ii. $\mathrm{Var}\left(\widetilde{Y}_{n+1}\right) = \sigma^2\left[\frac{1}{n} + \frac{(X_{n+1}-\overline{X})^2}{(n-1)S_X^2}\right]$.

iii. $\mathrm{Var}\left(\widetilde{\varepsilon}_{n+1}\right) = \sigma^2\left[1 + \frac{1}{n} + \frac{(X_{n+1}-\overline{X})^2}{(n-1)S_X^2}\right]$.

Απόδειξη. Χρησιμοποιώντας τις ιδιότητες της μέσης τιμής, της διασποράς και της συνδιακύμανσης, υπολογίζουμε τα εξής:

i. $E\left(\widetilde{\varepsilon}_{n+1}\right) = E\left(Y_{n+1} - \widehat{Y}_{n+1}\right) = E\left(Y_{n+1}\right) - E\left(\widetilde{Y}_{n+1}\right) = 0$.

ii. $\mathrm{Var}\left(\widetilde{Y}_{n+1}\right) = \mathrm{Var}\left(\widehat{\beta}_0 + \widehat{\beta}_1 X_{n+1}\right) \Rightarrow$

$\mathrm{Var}\left(\widetilde{Y}_{n+1}\right) = \mathrm{Var}\left(\widehat{\beta}_0\right) + X_{n+1}^2 \mathrm{Var}\left(\widehat{\beta}_1\right) + 2X_{n+1}\mathrm{Cov}\left(\widehat{\beta}_0, \widehat{\beta}_1\right) \overset{\text{Πρόταση 1.3}}{\Rightarrow}$

$\mathrm{Var}\left(\widetilde{Y}_{n+1}\right) = \sigma^2\left[\frac{1}{n} + \frac{\overline{X}^2}{(n-1)S_X^2} + \frac{X_{n+1}^2}{(n-1)S_X^2} - \frac{2X_{n+1}\overline{X}}{(n-1)S_X^2}\right] = \sigma^2\left[\frac{1}{n} + \frac{(X_{n+1}-\overline{X})^2}{(n-1)S_X^2}\right]$.

iii. $\mathrm{Var}(Y_{n+1}) = \mathrm{Var}\left(\beta_0 + \beta_1 X_{n+1} + \varepsilon_{n+1}\right) = \mathrm{Var}(\varepsilon_{n+1}) = \sigma^2$ και

$\mathrm{Cov}\left(Y_{n+1}, \widetilde{Y}_{n+1}\right) = \mathrm{Cov}\left(Y_{n+1}, \widehat{\beta}_0 + \widehat{\beta}_1 X_{n+1}\right) = 0$, αφού τα $\widehat{\beta}_0$ και $\widehat{\beta}_1$ είναι συναρτήσεις μόνο των $Y_1, Y_2, \ldots, Y_n$, τα οποία είναι όλα ασυσχέτιστα με το Y_{n+1}. Επομένως,

$\mathrm{Var}\left(\widetilde{\varepsilon}_{n+1}\right) = \mathrm{Var}\left(Y_{n+1}\right) + \mathrm{Var}\left(\widetilde{Y}_{n+1}\right) - 2\cancelto{0}{\mathrm{Cov}\left(Y_{n+1}, \widetilde{Y}_{n+1}\right)} \Rightarrow$

$\mathrm{Var}\left(\widetilde{\varepsilon}_{n+1}\right) = \sigma^2 + \sigma^2\left[\frac{1}{n} + \frac{(X_{n+1}-\overline{X})^2}{(n-1)S_X^2}\right] = \sigma^2\left[1 + \frac{1}{n} + \frac{(X_{n+1}-\overline{X})^2}{(n-1)S_X^2}\right]$. □

1.6 Κανονικό Απλό Γραμμικό Μοντέλο

Ορισμός 1.4. Έστω $\mathbf{X} = (X_1, X_2, \ldots, X_d) \in \mathbb{R}^d$ τυχαίο διάνυσμα. Τότε, ορίζουμε:

- $E(\mathbf{X}) = (E(X_1), E(X_2), \ldots, E(X_d))^{\mathrm{T}} \in \mathbb{R}^d$ το διάνυσμα μέσων τιμών του $\mathbf{X}$,
- $\mathrm{Var}(\mathbf{X}) = E\left[(\mathbf{X} - E(\mathbf{X}))(\mathbf{X} - E(\mathbf{X}))^{\mathrm{T}}\right] = E\left(\mathbf{X}\mathbf{X}^{\mathrm{T}}\right) - E(\mathbf{X})\left[E(\mathbf{X})\right]^{\mathrm{T}} \in \mathbb{R}^{d\times d}$, δηλαδή $\mathrm{Var}(\mathbf{X}) = \left[\mathrm{Cov}(X_i, X_j)\right]_{i,j}$, τον πίνακα συνδιακύμανσης του $\mathbf{X}$.

Πρόταση 1.9. Έστω $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^d$ τυχαία διανύσματα, $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n\times d}$ σταθεροί πίνακες και $\mathbf{b} \in \mathbb{R}^n$ σταθερό διάνυσμα. Τότε, γνωρίζουμε ότι:

i. $E(\mathbf{A}\mathbf{X} + \mathbf{b}) = \mathbf{A}E(\mathbf{X}) + \mathbf{b}$,

ii. $E(\mathbf{AX}+\mathbf{BY}) = \mathbf{A}E(\mathbf{X}) + \mathbf{B}E(\mathbf{Y})$,

iii. $\mathrm{Var}(\mathbf{AX}+\mathbf{b}) = \mathbf{A}\mathrm{Var}(\mathbf{X})\mathbf{A}^{\mathrm{T}}$,

iv. $\mathrm{Var}(\mathbf{X})$ συμμετρικός και θετικά ημιορισμένος, δηλαδή ισχύει $\mathbf{u}^{\mathrm{T}}\mathrm{Var}(\mathbf{X})\mathbf{u} \geqslant 0$, $\forall \mathbf{u} \in \mathbb{R}^d$.

Ορισμός 1.5. Έστω διάνυσμα μέσων τιμών $\boldsymbol{\mu} \in \mathbb{R}^d$ και συμμετρικός, θετικά ορισμένος πίνακας συνδιακύμανσης $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$, δηλαδή $\mathbf{u}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{u} > 0$, $\forall \mathbf{u} \neq \mathbf{0}_d$. Ένα τυχαίο διάνυσμα $\mathbf{X} \in \mathbb{R}^d$ που ακολουθεί την **πολυδιάστατη κανονική κατανομή** έχει συνάρτηση πυκνότητας πιθανότητας:

$$f_{\mathbf{X}}(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) = (2\pi)^{-\frac{d}{2}}\left(\det\boldsymbol{\Sigma}\right)^{-\frac{1}{2}}\exp\left\{-\frac{(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}{2}\right\}, \quad \forall \mathbf{x} \in \mathbb{R}^d.$$

Συμβολικά, γράφουμε ότι $\mathbf{X} \sim N_d(\boldsymbol{\mu},\boldsymbol{\Sigma})$.

Πρόταση 1.10. Έστω $\mathbf{X} \sim N_d(\boldsymbol{\mu},\boldsymbol{\Sigma})$. Τότε,

i. $E(\mathbf{X}) = \boldsymbol{\mu}$ και $\mathrm{Var}(\mathbf{X}) = \boldsymbol{\Sigma}$.

ii. Για $\mathbf{A} \in \mathbb{R}^{n\times d}$ και $\mathbf{b} \in \mathbb{R}^n$, έχουμε ότι $\mathbf{AX}+\mathbf{b} \sim N_n\left(\mathbf{A}\boldsymbol{\mu}+\mathbf{b}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^{\mathrm{T}}\right)$.

iii. $X_1, X_2, \ldots, X_d$ ανεξάρτητες αν και μόνο αν $X_1, X_2, \ldots, X_d$ ασυσχέτιστες, δηλαδή $\boldsymbol{\Sigma}$ διαγώνιος.

Πρόταση 1.11. Έστω $\mathbf{X} \sim N_d(\boldsymbol{\mu},\boldsymbol{\Sigma})$ και $\boldsymbol{\Sigma} = \mathbf{A}\mathbf{A}^{\mathrm{T}}$ για κάποιον $\mathbf{A} \in \mathbb{R}^{d\times d}$. Τότε,

i. $\mathbf{Z} = \mathbf{A}^{-1}(\mathbf{X}-\boldsymbol{\mu}) \sim N_d(\mathbf{0}_d, \mathbf{I}_d)$.

ii. $\|\mathbf{Z}\|^2 = \mathbf{Z}^{\mathrm{T}}\mathbf{Z} = (\mathbf{X}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{X}-\boldsymbol{\mu}) \sim \chi^2_d$.

Ορισμός 1.6. Το μοντέλο $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ για $i = 1, 2, \ldots, n$, όπου ισχύει $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n) \sim N_n\left(\mathbf{0}_n, \sigma^2\mathbf{I}_n\right)$, καλείται **κανονικό απλό γραμμικό μοντέλο** ή απλό γραμμικό μοντέλο με κανονικά σφάλματα.

Παρατήρηση 1.3. (Κανονικό Απλό Γραμμικό Μοντέλο)

- Τα τυχαία σφάλματα στο κανονικό απλό γραμμικό μοντέλο είναι **κανονικά κατανεμημένα** με **μέση τιμή 0**, **ομοσκεδαστικά** και **ανεξάρτητα**, ή ισοδύναμα ασυσχέτιστα. Δηλαδή, $\varepsilon_i \sim N(0,\sigma^2)$ ανεξάρτητα για $i = 1, 2, \ldots, n$.
- Στο κανονικό απλό γραμμικό μοντέλο, ισχύει ότι $Y_i \sim N\left(\beta_0+\beta_1 X_i, \sigma^2\right)$ ανεξάρτητες για $i = 1, 2, \ldots, n$.
- Έχοντας, πλέον, κάνει υπόθεση για την κατανομή των τυχαίων σφαλμάτων ε_i, μπορούμε να κάνουμε χρήση άλλων γνωστών μεθόδων εκτίμησης εκτός από τη μέθοδο ελαχίστων τετραγώνων, όπως η μέθοδος μέγιστης πιθανοφάνειας που γνωρίζουμε από τη μαθηματική στατιστική.

Πρόταση 1.12. Οι **εκτιμήτριες μέγιστης πιθανοφάνειας** των συντελεστών παλινδρόμησης ταυτίζονται με τις εκτιμήτριες ελαχίστων τετραγώνων $\widehat{\beta}_0$ και $\widehat{\beta}_1$, ενώ η εκτιμήτρια μέγιστης πιθανόφανειας της διασποράς σ^2 δίνεται από τον τύπο:

$$\widehat{\sigma}^2 = \frac{\text{SSE}}{n} = \frac{(n-2)S^2}{n} = \frac{1}{n}\sum_{i=1}^{n}\widehat{\varepsilon}_i^2.$$

Απόδειξη. Αφού οι παρατηρήσεις Y_i είναι ανεξάρτητες, μπορούμε εύκολα να υπολογίσουμε τη συνάρτηση πιθανοφάνειας των δεδομένων ως εξής:

$$\begin{aligned}
L\left(\beta_0, \beta_1, \sigma^2 | y_1, y_2, \ldots, y_n\right) &= \prod_{i=1}^{n} f_{Y_i}\left(y_i; \beta_0, \beta_1, \sigma^2\right) \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{\left(y_i - \beta_0 - \beta_1 x_i\right)^2}{2\sigma^2}\right\} \\
&= \left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \beta_0 - \beta_1 x_i\right)^2\right\} \\
&= \left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\varepsilon_i^2\right\} \\
&= \left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left\{-\frac{Q(\beta_0, \beta_1)}{2\sigma^2}\right\}.
\end{aligned}$$

Λογαριθμίζουμε τη συνάρτηση πιθανοφάνειας για να υπολογίσουμε τη συνάρτηση λογαριθμοπιθανοφάνειας:

$$\begin{aligned}
\ell\left(\beta_0, \beta_1, \sigma^2 | y_1, y_2, \ldots, y_n\right) &= \log L\left(\beta_0, \beta_1, \sigma^2 | y_1, y_2, \ldots, y_n\right) \\
&= -\frac{n\log(2\pi)}{2} - \frac{n\log\sigma^2}{2} - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \beta_0 - \beta_1 x_i\right)^2.
\end{aligned}$$

Μεγιστοποιούμε πρώτα ως προς το διάνυσμα (β_0, β_1):

$$\frac{\partial\ell}{\partial\beta_0} = \frac{1}{\sigma^2}\sum_{i=1}^{n}\left(y_i - \beta_0 - \beta_1 x_i\right) = 0, \quad \frac{\partial\ell}{\partial\beta_1} = \frac{1}{\sigma^2}\sum_{i=1}^{n} x_i\left(y_i - \beta_0 - \beta_1 x_i\right) = 0.$$

Οι δύο παραπάνω εξισώσεις ταυτίζονται με τις κανονικές εξισώσεις στις οποίες καταλήξαμε στην πρόταση 1.2, οπότε μας οδηγούν στις εκτιμήτριες ελαχίστων τετραγώνων $\widehat{\beta}_0$, $\widehat{\beta}_1$. Στη συνέχεια, μεγιστοποιούμε τη λογαριθμοπιθανοφάνεια ως προς σ^2:

$$\frac{\partial\ell}{\partial\sigma^2}\left(\widehat{\beta}_0, \widehat{\beta}_1, \sigma^2\right) = -\frac{n}{2\sigma^2} + \frac{1}{2\left(\sigma^2\right)^2}\sum_{i=1}^{n}\left(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i\right)^2 = 0 \Rightarrow$$

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i\right)^2 = \frac{\text{SSE}}{n} = \frac{(n-2)S^2}{n}.$$

Υπολογίζουμε τη δεύτερη παράγωγο της συνάρτησης $\ell\left(\widehat{\beta}_0, \widehat{\beta}_1, \sigma^2\right)$ ως προς σ^2 για να επαληθεύσουμε ότι το $\widehat{\sigma}^2$ είναι σημείο μεγίστου:

$$\frac{\partial^2 \ell}{\partial \left(\sigma^2\right)^2}\left(\widehat{\beta}_0, \widehat{\beta}_1, \sigma^2\right) = \frac{n}{2\left(\sigma^2\right)^2} - \frac{1}{\left(\sigma^2\right)^3}\sum_{i=1}^{n}\left(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i\right)^2 \Rightarrow$$

$$\frac{\partial^2 \ell}{\partial \left(\sigma^2\right)^2}\left(\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\sigma}^2\right) = \frac{1}{\widehat{\sigma}^4}\left[\frac{n}{2} - \frac{1}{\widehat{\sigma}^2}\sum_{i=1}^{n}\left(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i\right)^2\right] = \frac{1}{\widehat{\sigma}^4}\left(\frac{n}{2} - n\right) \Rightarrow$$

$$\frac{\partial^2 \ell}{\partial \left(\sigma^2\right)^2}\left(\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\sigma}^2\right) = -\frac{n}{2\widehat{\sigma}^4} < 0.$$

Υπολογίζοντας τα όρια της συνάρτησης $L\left(\widehat{\beta}_0, \widehat{\beta}_1, \sigma^2\right)$ ως προς σ^2 στα 0^+ και ∞, επαληθεύουμε ότι το $\widehat{\sigma}^2$ είναι σημείο ολικού μεγίστου:

$$\lim_{\sigma^2 \to \infty} L\left(\widehat{\beta}_0, \widehat{\beta}_1, \sigma^2\right) = \lim_{\sigma^2 \to \infty}\left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left\{-\frac{Q(\widehat{\beta}_0, \widehat{\beta}_1)}{2\sigma^2}\right\} \overset{n>0}{=} 0,$$

$$\lim_{\sigma^2 \to 0^+} L\left(\widehat{\beta}_0, \widehat{\beta}_1, \sigma^2\right) \overset{\tau=\sigma^{-2}}{=} \lim_{\tau \to \infty}\left(\frac{\tau}{2\pi}\right)^{\frac{n}{2}} \exp\left\{-\frac{Q(\widehat{\beta}_0, \widehat{\beta}_1)}{2}\tau\right\} \overset{Q(\widehat{\beta}_0, \widehat{\beta}_1)>0}{=} 0. \quad \square$$

Τώρα, που έχουμε υποθέσει ότι τα τυχαία σφάλματα ε_i είναι κανονικά κατανεμημένα, μπορούμε να πάρουμε αποτελέσματα για τις κατανομές των εκτιμητριών $\widehat{\beta}_0$, $\widehat{\beta}_1$ και S^2. Οι κατανομές αυτές είναι πολύ σημαντικές, επειδή μας επιτρέπουν, πέρα από τις σημειακές εκτιμήσεις που έχουμε υπολογίσει, να κατασκευάσουμε διαστήματα εμπιστοσύνης και να πραγματοποιήσουμε ελέγχους υποθέσεων.

Πρόταση 1.13. Ισχύει ότι $\overline{Y} \sim N\left(\beta_0 + \beta_1 \overline{X}, \frac{\sigma^2}{n}\right)$.

Απόδειξη. Γνωρίζουμε ότι η τυχαία μεταβλητή $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ είναι, προφανώς, γραμμικός συνδυασμός των τυχαίων μεταβλητών Y_i, οι οποίες είναι ανεξάρτητες και κανονικά κατανεμημένες, σύμφωνα με την παρατήρηση 1.3. Συμπεραίνουμε ότι και η τυχαία μεταβλητή $\overline{Y}$ είναι κανονικά κατανεμημένη. Στην πρόταση 1.1, υπολογίσαμε τη μέση τιμή και τη διασπορά της τυχαίας μεταβλητής $\overline{Y}$. Συγκεκριμένα, αποδείξαμε ότι $E\left(\overline{Y}\right) = \beta_0 + \beta_1 \overline{X}$ και $\mathrm{Var}\left(\overline{Y}\right) = \frac{\sigma^2}{n}$. Επομένως, παίρνουμε άμεσα ότι $\overline{Y} \sim N\left(\beta_0 + \beta_1 \overline{X}, \frac{\sigma^2}{n}\right)$. $\square$

Πρόταση 1.14. (Κατανομές Εκτιμητριών)

i. $\widehat{\boldsymbol{\beta}} = \left(\widehat{\beta}_0, \widehat{\beta}_1\right)^{\mathrm{T}} \sim N_2\left(\boldsymbol{\beta}, \boldsymbol{\Sigma}_{\widehat{\boldsymbol{\beta}}}\right)$, όπου $\boldsymbol{\beta} = (\beta_0, \beta_1)^{\mathrm{T}}$ και

$$\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\beta}}} = \mathrm{Var}\left(\widehat{\boldsymbol{\beta}}\right) = \frac{\sigma^2}{(n-1)S_X^2}\begin{bmatrix} \frac{1}{n}\sum_{i=1}^{n} X_i^2 & -\overline{X} \\ -\overline{X} & 1 \end{bmatrix}.$$

Συμπεραίνουμε ότι $\widehat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{(n-1)S_X^2}\right)$ και $\widehat{\beta}_0 \sim N\left(\beta_0, \sigma^2\left[\frac{1}{n} + \frac{\overline{X}^2}{(n-1)S_X^2}\right]\right)$. Επιπλέον, σύμφωνα με την πρόταση 1.11, $W = \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)^{\mathrm{T}} \boldsymbol{\Sigma}_{\widehat{\boldsymbol{\beta}}}^{-1} \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \sim \chi_2^2$.

ii. $Q = \frac{(n-2)S^2}{\sigma^2} \sim \chi_{n-2}^2$.

iii. Η τυχαία μεταβλητή S^2 είναι ανεξάρτητη από το $\widehat{\boldsymbol{\beta}} = \left(\widehat{\beta}_0, \widehat{\beta}_1\right)^{\mathrm{T}}$.

Απόδειξη. Βλέπε πρόταση 2.11 (σελίδα 54).

Σημείωση 1.2. Την κατανομή που ακολουθούν οι εκτιμήτριες $\widehat{\beta}_0$ και $\widehat{\beta}_1$ μπορούμε να την προσδιορίσουμε δουλεύοντας όπως στην πρόταση 1.13. Σύμφωνα με τη σημείωση 1.1, οι τυχαίες μεταβλητές $\widehat{\beta}_0$ και $\widehat{\beta}_1$ είναι γραμμικοί συνδυασμοί των Y_i, επομένως είναι κανονικά κατανεμημένες. Επιπλέον, στην πρόταση 1.3, έχουμε υπολογίσει τις μέσες τιμές και τις διασπορές των $\widehat{\beta}_0$ και $\widehat{\beta}_1$, οπότε έχουμε άμεσα το ζητούμενο.

Πρόταση 1.15. (Κατανομές των $\widehat{\beta}_0$ και $\widehat{\beta}_1$ με χρήση του S^2)

i. $\frac{\widehat{\beta}_i - \beta_i}{S_{\widehat{\beta}_i}} \sim t_{n-2}$ για $i = 0, 1$, όπου $S_{\widehat{\beta}_1}^2 = \frac{S^2}{(n-1)S_X^2}$ και $S_{\widehat{\beta}_0}^2 = S^2\left[\frac{1}{n} + \frac{\overline{X}^2}{(n-1)S_X^2}\right]$.

ii. $\frac{1}{2}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)^{\mathrm{T}} \mathbf{S}_{\widehat{\boldsymbol{\beta}}}^{-1} \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \sim F_{2,n-2}$, όπου

$$\mathbf{S}_{\widehat{\boldsymbol{\beta}}} = \frac{S^2}{(n-1)S_X^2} \begin{bmatrix} \frac{1}{n}\sum_{i=1}^{n} X_i^2 & -\overline{X} \\ -\overline{X} & 1 \end{bmatrix}.$$

Απόδειξη. Για την απόδειξη αυτής της πρότασης, θα χρειαστεί να θυμόμαστε κάποια βασικά στοιχεία σχετικά με την κατανομή t του Student και την κατανομή F του Snedecor.

i. Σύμφωνα με την προηγούμενη πρόταση, παίρνουμε ότι:

$$Z = \frac{\widehat{\beta}_1 - \beta_1}{\sigma} \cdot S_X\sqrt{n-1} \sim N(0,1), \quad Q = \frac{(n-2)S^2}{\sigma^2} \sim \chi_{n-2}^2$$

και Q ανεξάρτητη από την Z, αφού S^2 ανεξάρτητη από τη $\widehat{\beta}_1$. Επομένως, συμπεραίνουμε ότι:

$$T = \frac{Z}{\sqrt{\frac{Q}{n-2}}} = \frac{\widehat{\beta}_1 - \beta_1}{\cancel{\sigma}} \cdot S_X\sqrt{n-1} \cdot \frac{\cancel{\sigma}}{S} = \frac{\widehat{\beta}_1 - \beta_1}{S} \cdot S_X\sqrt{n-1} = \frac{\widehat{\beta}_1 - \beta_1}{S_{\widehat{\beta}_1}} \sim t_{n-2},$$

όπου $S_{\widehat{\beta}_1}^2 = \frac{S^2}{(n-1)S_X^2}$. Με όμοιο τρόπο, υπολογίζουμε την κατανομή της $\widehat{\beta}_0$.

ii. Πάλι σύμφωνα με την προηγούμενη πρόταση, παίρνουμε ότι:

$$W = \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)^{\mathrm{T}} \boldsymbol{\Sigma}_{\widehat{\boldsymbol{\beta}}}^{-1} \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \sim \chi_2^2$$

και Q ανεξάρτητη από την W, αφού S^2 ανεξάρτητη από το $\widehat{\boldsymbol{\beta}}$. Επομένως,

$$\begin{aligned} F &= \frac{W}{2} \cdot \frac{n-2}{Q} = \frac{1}{2}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)^{\mathrm{T}} \boldsymbol{\Sigma}_{\widehat{\boldsymbol{\beta}}}^{-1} \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \cdot \frac{\sigma^2}{S^2} \\ &= \frac{1}{2}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)^{\mathrm{T}} \frac{(n-1)S_X^2}{\cancel{\sigma^2}} \begin{bmatrix} \frac{1}{n}\sum_{i=1}^{n} X_i^2 & -\overline{X} \\ -\overline{X} & 1 \end{bmatrix}^{-1} \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \cdot \frac{\cancel{\sigma^2}}{S^2} \\ &= \frac{1}{2}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)^{\mathrm{T}} \frac{(n-1)S_X^2}{S^2} \begin{bmatrix} \frac{1}{n}\sum_{i=1}^{n} X_i^2 & -\overline{X} \\ -\overline{X} & 1 \end{bmatrix}^{-1} \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \\ &= \frac{1}{2}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)^{\mathrm{T}} \mathbf{S}_{\widehat{\boldsymbol{\beta}}}^{-1} \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \sim F_{2,n-2}, \end{aligned}$$

όπου $\mathbf{S}_{\widehat{\boldsymbol{\beta}}} = \frac{S^2}{(n-1)S_X^2} \begin{bmatrix} \frac{1}{n}\sum_{i=1}^{n} X_i^2 & -\overline{X} \\ -\overline{X} & 1 \end{bmatrix}$. □

1.7 Διαστήματα και Περιοχές Εμπιστοσύνης

Χρησιμοποιώντας την πρόταση 1.15, μπορούμε άμεσα να κατασκευάσουμε διαστήματα και περιοχές εμπιστοσύνης για τις παραμέτρους του μοντέλου.

Πρόταση 1.16. (Διαστήματα και Περιοχές Εμπιστοσύνης)

i. Ένα $100(1-\alpha)\%$ διάστημα εμπιστοσύνης ίσων ουρών για το β_i για $i = 0, 1$ δίνεται από τη σχέση:

$$I_{1-\alpha}(\beta_i) = \left[\widehat{\beta}_i - t_{n-2;\frac{\alpha}{2}} \cdot S_{\widehat{\beta}_i}, \widehat{\beta}_i + t_{n-2;\frac{\alpha}{2}} \cdot S_{\widehat{\beta}_i}\right].$$

ii. Ένα $100(1-\alpha)\%$ διάστημα εμπιστοσύνης ίσων ουρών για το σ^2 δίνεται από τη σχέση:

$$I_{1-\alpha}\left(\sigma^2\right) = \left[\frac{(n-2)S^2}{\chi^2_{n-2;\frac{\alpha}{2}}}, \frac{(n-2)S^2}{\chi^2_{n-2;1-\frac{\alpha}{2}}}\right].$$

iii. Μία $100(1-\alpha)\%$ **περιοχή εμπιστοσύνης** για το $\boldsymbol{\beta} = (\beta_1, \beta_2)^{\mathrm{T}}$ δίνεται από τη σχέση:

$$R_{1-\alpha}(\boldsymbol{\beta}) = \left\{\boldsymbol{\beta} \in \mathbb{R}^2 : \frac{1}{2}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)^{\mathrm{T}} \mathbf{S}_{\widehat{\boldsymbol{\beta}}}^{-1} \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \leqslant F_{2,n-2;\alpha}\right\}.$$

Απόδειξη. Σύμφωνα με την προηγούμενη πρόταση, παίρνουμε τα παρακάτω αποτελέσματα:

i. Γνωρίζουμε ότι $T = \frac{\widehat{\beta}_i - \beta_i}{S_{\widehat{\beta}_i}} \sim t_{n-2}$. Ζητάμε σταθερές $c_1, c_2 \in \mathbb{R}$ τέτοιες, ώστε $P(c_1 \leqslant T \leqslant c_2) = 1 - \alpha$. Συγκεκριμένα, για το διάστημα εμπιστοσύνης ίσων

ουρών χρησιμοποιούμε τις σχέσεις:

$$P(T < c_1) = \frac{\alpha}{2} \Rightarrow P(T > c_1) = 1 - \frac{\alpha}{2} \Rightarrow c_1 = t_{n-2;1-\frac{\alpha}{2}} = -t_{n-2;\frac{\alpha}{2}},$$

$$P(T > c_2) = \frac{\alpha}{2} \Rightarrow c_2 = t_{n-2;\frac{\alpha}{2}}.$$

Επομένως, το διάστημα εμπιστοσύνης ίσων ουρών δίνεται από τη σχέση:

$$c_1 \leqslant \frac{\widehat{\beta}_i - \beta_i}{S_{\widehat{\beta}_i}} \leqslant c_2 \Leftrightarrow -c_2 \cdot S_{\widehat{\beta}_i} \leqslant \beta_i - \widehat{\beta}_i \leqslant -c_1 \cdot S_{\widehat{\beta}_i} \Leftrightarrow$$

$$\widehat{\beta}_i - t_{n-2;\frac{\alpha}{2}} \cdot S_{\widehat{\beta}_i} \leqslant \beta_i \leqslant \widehat{\beta}_i + t_{n-2;\frac{\alpha}{2}} \cdot S_{\widehat{\beta}_i}.$$

Τελικά, παίρνουμε ότι $I_{1-\alpha}(\beta_i) = \left[\widehat{\beta}_i - t_{n-2;\frac{\alpha}{2}} \cdot S_{\widehat{\beta}_i}, \widehat{\beta}_i + t_{n-2;\frac{\alpha}{2}} \cdot S_{\widehat{\beta}_i}\right]$.

ii. Γνωρίζουμε ότι $Q = \frac{(n-2)S^2}{\sigma^2} \sim \chi^2_{n-2}$. Ζητάμε σταθερές $c_1, c_2 \in \mathbb{R}$ τέτοιες, ώστε $P(c_1 \leqslant Q \leqslant c_2) = 1 - \alpha$. Συγκεκριμένα, για το διάστημα εμπιστοσύνης ίσων ουρών χρησιμοποιούμε τις σχέσεις:

$$P(Q < c_1) = \frac{\alpha}{2} \Rightarrow P(Q > c_1) = 1 - \frac{\alpha}{2} \Rightarrow c_1 = \chi^2_{n-2;1-\frac{\alpha}{2}},$$

$$P(Q > c_2) = \frac{\alpha}{2} \Rightarrow c_2 = \chi^2_{n-2;\frac{\alpha}{2}}.$$

Επομένως, το διάστημα εμπιστοσύνης ίσων ουρών δίνεται από τη σχέση:

$$c_1 \leqslant \frac{(n-2)S^2}{\sigma^2} \leqslant c_2 \Leftrightarrow \frac{(n-2)S^2}{\chi^2_{n-2;\frac{\alpha}{2}}} \leqslant \sigma^2 \leqslant \frac{(n-2)S^2}{\chi^2_{n-2;1-\frac{\alpha}{2}}}.$$

Τελικά, παίρνουμε ότι $I_{1-\alpha}\left(\sigma^2\right) = \left[\dfrac{(n-2)S^2}{\chi^2_{n-2;\frac{\alpha}{2}}}, \dfrac{(n-2)S^2}{\chi^2_{n-2;1-\frac{\alpha}{2}}}\right]$.

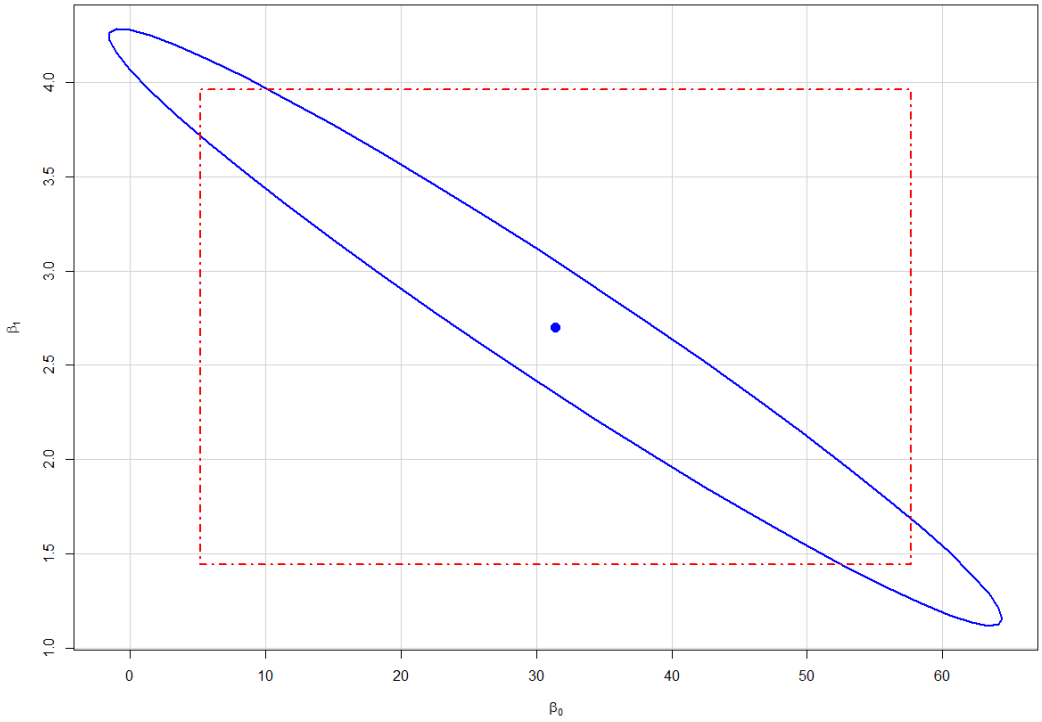
iii. Γνωρίζουμε ότι $F = \frac{1}{2}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)^{\mathrm{T}} \mathbf{S}^{-1}_{\widehat{\boldsymbol{\beta}}} \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \sim F_{2,n-2}$. Ζητάμε σταθερά $c \in \mathbb{R}$ τέτοια, ώστε $P(F \leqslant c) = 1 - \alpha$. Ισοδύναμα, ζητάμε $P(F > c) = \alpha$, δηλαδή $c = F_{2,n-2;\alpha}$. Επομένως, η ζητούμενη περιοχή εμπιστοσύνης δίνεται από τη σχέση $R_{1-\alpha}(\boldsymbol{\beta}) = \left\{\boldsymbol{\beta} \in \mathbb{R}^2 : \frac{1}{2}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)^{\mathrm{T}} \mathbf{S}^{-1}_{\widehat{\boldsymbol{\beta}}} \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \leqslant F_{2,n-2;\alpha}\right\}$. □

Σημείωση 1.3. Στο σχήμα 1.2 βλέπουμε σχεδιασμένη με μπλε χρώμα την 95% περιοχή εμπιστοσύνης για την παράμετρο $\boldsymbol{\beta} = (\beta_0, \beta_1)$. Βλέπουμε ότι η περιοχή εμπιστοσύνης παίρνει τη μορφή μίας έλλειψης στην περίπτωση του απλού γραμμικού μοντέλου. Παρατηρούμε ότι ο μεγάλος άξονας της έλλειψης έχει αρνητική κλίση και έχει μεγαλύτερο μήκος από τον μικρό άξονα, οπότε οι εκτιμήτριες $\widehat{\beta}_0$ και $\widehat{\beta}_1$ είναι ισχυρά αρνητικά συσχετισμένες στην περίπτωση αυτή.

Με κόκκινο χρώμα βλέπουμε σχεδιασμένα τα επί μέρους διαστήματα εμπιστο-

σύνης για τις δύο παραμέτρους β_0 και β_1, τα οποία σχηματίζουν ένα ορθογώνιο. Σε αντίθεση με την περιοχή εμπιστοσύνης, το ορθογώνιο αυτό δε λαμβάνει καθόλου υπόψη την πιθανή συσχέτιση μεταξύ $\widehat{\beta}_0$ και $\widehat{\beta}_1$.

Το κέντρο της έλλειψης ταυτίζεται με το κέντρο του ορθογωνίου και είναι το σημείο $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_0, \widehat{\beta}_1)$. Το εμβαδόν του ορθογωνίου είναι μεγαλύτερο από αυτό της έλλειψης, οπότε αποτελεί μία άλλη περιοχή εμπιστοσύνης για το ζεύγος των παραμέτρων $\boldsymbol{\beta} = (\beta_0, \beta_1)$, αλλά σε επίπεδο εμπιστοσύνης μεγαλύτερο από 95%. Από την άλλη μεριά, οι προβολές της έλλειψης πάνω στους άξονες βγαίνουν εκτός των ορίων που ορίζει το ορθογώνιο, οπότε δίνουν διαστήματα εμπιστοσύνης για τις παραμέτρους β_0 και β_1, αλλά σε επίπεδο εμπιστοσύνης μεγαλύτερο από 95%.



Σχημα 1.2: Σύγκριση Περιοχής και Διαστημάτων Εμπιστοσύνης

1.8 Στατιστικοί Έλεγχοι Υποθέσεων

Ορισμός 1.7. Σε έναν στατιστικό έλεγχο υποθέσεων καλούμε **παρατηρούμενο επίπεδο στατιστικής σημαντικότητας** ή **p-value** την πιθανότητα να παρατηρήσουμε τιμή της ελεγχοσυνάρτησης ίση ή ακόμα πιο ακραία από αυτήν που παρατηρήσαμε στο συγκεκριμένο δείγμα, υπό την ισχύ της μηδενικής υπόθεσης H_0. Αποφασίζουμε να απορρίψουμε την H_0 σε ε.σ.σ. α αν και μόνο αν p-value $< \alpha$.

Πρόταση 1.17. Υπό τη μηδενική υπόθεση $H_0 : \beta_i = \beta_{i,0}$ για $i = 0, 1$, γνωρίζουμε

ότι $T = \frac{\widehat{\beta}_i - \beta_{i,0}}{S_{\widehat{\beta}_i}} \sim t_{n-2}$. Αντικαθιστώντας τις τυχαίες μεταβλητές $Y_1, Y_2, \ldots, Y_n$ που εμφανίζονται στην ελεγχοσυνάρτηση T από τις παρατηρήσεις $y_1, y_2, \ldots, y_n$, υπολογίζουμε την παρατηρούμενη τιμή $t = \frac{\widehat{\beta}_i - \beta_{i,0}}{s_{\widehat{\beta}_i}}$ της ελεγχοσυνάρτησης.

i. Έστω ότι θέλουμε να πραγματοποιήσουμε τον έλεγχο με την αμφίπλευρη εναλλακτική υπόθεση $H_1 : \beta_i \neq \beta_{i,0}$ για $i = 0, 1$. Τότε, απορρίπτουμε την H_0 σε ε.σ.σ. α αν και μόνο αν $|t| > t_{n-2;\frac{\alpha}{2}}$ ή p-value$^{(\neq)} = P\left(|T| \geqslant |t|\right) < \alpha$.

ii. Έστω ότι θέλουμε να πραγματοποιήσουμε τον έλεγχο με τη μονόπλευρη εναλλακτική υπόθεση $H_1 : \beta_i > \beta_{i,0}$ για $i = 0, 1$. Τότε, απορρίπτουμε την H_0 σε ε.σ.σ. α αν και μόνο αν $t > t_{n-2;\alpha}$ ή p-value$^{(>)} = P\left(T \geqslant t\right) < \alpha$.

iii. Έστω ότι θέλουμε να πραγματοποιήσουμε τον έλεγχο με τη μονόπλευρη εναλλακτική υπόθεση $H_1 : \beta_i < \beta_{i,0}$ για $i = 0, 1$. Τότε, απορρίπτουμε την H_0 σε ε.σ.σ. α αν και μόνο αν $t < -t_{n-2;\alpha}$ ή p-value$^{(<)} = P\left(T \leqslant t\right) < \alpha$.

Απόδειξη.* Όλοι οι παραπάνω έλεγχοι υποθέσεων είναι της μορφής σύνθετη έναντι σύνθετης, οπότε θα κάνουμε χρήση του κριτηρίου γενικευμένου λόγου πιθανοφανειών. Θα πραγματοποιήσουμε τους ελέγχους μόνο για την παράμετρο β_1.

i. Για το απλό γραμμικό μοντέλο, έχουμε υπολογίσει ότι:

$$L\left(\beta_0, \beta_1, \sigma^2\right) = \left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \beta_0 - \beta_1 x_i\right)^2\right\} \Rightarrow$$

$$\ell\left(\beta_0, \beta_1, \sigma^2\right) = -\frac{n}{2}\log\left(2\pi\right) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \beta_0 - \beta_1 x_i\right)^2.$$

Έχουμε υπολογίσει τις εκτιμήτριες μέγιστης πιθανοφάνειας $\widehat{\beta}_0$, $\widehat{\beta}_1$ και $\widehat{\sigma}^2$ των παραμέτρων β_0, β_1, σ^2. Με όμοιο τρόπο υπολογίζουμε τις εκτιμήτριες μέγιστης πιθανοφάνειας των β_0, σ^2 υπό την $H_0 : \beta_1 = \beta_{1,0}$:

$$\widetilde{\beta}_0 = \overline{y} - \beta_{1,0}\overline{x}, \quad \widetilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \widetilde{\beta}_0 - \beta_{1,0}x_i\right)^2.$$

Υπολογίζουμε τον λογάριθμο του γενικευμένου λόγου πιθανοφανειών:

$$\begin{aligned}
\log\lambda^* &= \ell\left(\widetilde{\beta}_0, \beta_{1,0}, \widetilde{\sigma}^2\right) - \ell\left(\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\sigma}^2\right) \\
&= -\frac{n}{2}\log\frac{\widetilde{\sigma}^2}{\widehat{\sigma}^2} - \frac{1}{2\widetilde{\sigma}^2}\sum_{i=1}^{n}\left(y_i - \widetilde{\beta}_0 - \beta_{1,0}x_i\right)^2 + \frac{1}{2\widehat{\sigma}^2}\sum_{i=1}^{n}\left(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i\right)^2 \\
&= -\frac{n}{2}\log\frac{\widetilde{\sigma}^2}{\widehat{\sigma}^2} - \frac{\not{n}}{\not{2}} + \frac{\not{n}}{\not{2}} = -\frac{n}{2}\log\frac{\widetilde{\sigma}^2}{\widehat{\sigma}^2},
\end{aligned}$$

Προσπαθούμε να εκφράσουμε την εκτιμήτρια $\widetilde{\sigma}^2$ συναρτήσει της εκτιμήτριας

$\widehat{\sigma}^2$ για να απλοποιήσουμε την παράσταση:

$$\begin{aligned}
\widetilde{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}\left[y_i - \left(\widehat{\beta}_0 + \widehat{\beta}_1 x_i\right) + \left(\widehat{\beta}_0 + \widehat{\beta}_1 x_i\right) - \left(\widetilde{\beta}_0 + \beta_{1,0} x_i\right)\right]^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[\widehat{\varepsilon}_i + \left(\overline{y} - \widehat{\beta}_1\overline{x} + \widehat{\beta}_1 x_i\right) - \left(\overline{y} - \beta_{1,0}\overline{x} + \beta_{1,0} x_i\right)\right]^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[\widehat{\varepsilon}_i + \left(\widehat{\beta}_1 - \beta_{1,0}\right)(x_i - \overline{x})\right]^2 \\
&= \frac{1}{n}\left[\sum_{i=1}^{n}\widehat{\varepsilon}_i^2 + \left(\widehat{\beta}_1 - \beta_{1,0}\right)^2\sum_{i=1}^{n}(x_i - \overline{x})^2 + 2\left(\widehat{\beta}_1 - \beta_{1,0}\right)\sum_{i=1}^{n}(x_i - \overline{x})\,\widehat{\varepsilon}_i\right] \\
&= \widehat{\sigma}^2 + \frac{1}{n}\left[\left(\widehat{\beta}_1 - \beta_{1,0}\right)^2 (n-1)s_X^2 + 2\left(\widehat{\beta}_1 - \beta_{1,0}\right)\left(\sum_{i=1}^{n} x_i\widehat{\varepsilon}_i - \overline{x}\sum_{i=1}^{n}\widehat{\varepsilon}_i\right)\right],
\end{aligned}$$

όπου $\sum_{i=1}^{n} x_i\widehat{\varepsilon}_i = \sum_{i=1}^{n}\widehat{\varepsilon}_i = 0$, οπότε $\widetilde{\sigma}^2 = \widehat{\sigma}^2 + \frac{(n-1)s_X^2}{n}\left(\widehat{\beta}_1 - \beta_{1,0}\right)^2$. Επομένως,

$$\begin{aligned}
\log\lambda^* &= -\frac{n}{2}\log\left[1 + \frac{(n-1)s_X^2}{n\widehat{\sigma}^2}\left(\widehat{\beta}_1 - \beta_{1,0}\right)^2\right] \\
&= -\frac{n}{2}\log\left[1 + \frac{(n-1)s_X^2}{(n-2)s^2}\left(\widehat{\beta}_1 - \beta_{1,0}\right)^2\right] \\
&= -\frac{n}{2}\log\left[1 + \frac{1}{(n-2)s_{\widehat{\beta}_1}^2}\left(\widehat{\beta}_1 - \beta_{1,0}\right)^2\right].
\end{aligned}$$

Για τον υπολογισμό της κρίσιμης περιοχής του ελέγχου πρέπει να λύσουμε την ανισότητα $\lambda^* < c$ ως προς κάποια στατιστική συνάρτηση:

$$\lambda^* < c \Leftrightarrow \log\lambda^* < c^* = \log c \Leftrightarrow 1 + \frac{1}{(n-2)s_{\widehat{\beta}_1}^2}\left(\widehat{\beta}_1 - \beta_{1,0}\right)^2 > c^{**} = e^{-\frac{2c^*}{n}} \Leftrightarrow$$

$$\left[\frac{\widehat{\beta}_1 - \beta_{1,0}}{s_{\widehat{\beta}_1}}\right]^2 > c^{***} = (n-2)\left(c^{**} - 1\right) \Leftrightarrow |t| = \left|\frac{\widehat{\beta}_1 - \beta_{1,0}}{s_{\widehat{\beta}_1}}\right| > c_\alpha = \sqrt{|c^{***}|}.$$

Υπό τη μηδενική υπόθεση $H_0 : \beta_1 = \beta_{1,0}$, γνωρίζουμε ότι $T = \frac{\widehat{\beta}_1 - \beta_{1,0}}{S_{\widehat{\beta}_1}} \sim t_{n-2}$. Για τον υπολογισμό της σταθεράς c_α, απαιτούμε η πιθανότητα σφάλματος τύπου Ι του ελέγχου, δηλαδή η πιθανότητα λανθασμένης απόρριψης της H_0, να είναι ίση με α:

$$\begin{aligned}
P_{H_0}(|T| > c_\alpha) &= P_{H_0}(T > c_\alpha) + P_{H_0}(T < -c_\alpha) = 1 - F_T(c_\alpha) + F_T(-c_\alpha) \\
&= 1 - F_T(c_\alpha) + 1 - F_T(c_\alpha) = 2\left[1 - F_T(c_\alpha)\right] = \alpha \Rightarrow
\end{aligned}$$

$$F_T(c_\alpha) = 1 - \frac{\alpha}{2} \Rightarrow c_\alpha = F_T^{-1}\left(1 - \frac{\alpha}{2}\right) = t_{n-2;\frac{\alpha}{2}}.$$

Επομένως, απορρίπτουμε την H_0 αν και μόνο αν $|t| > t_{n-2;\frac{\alpha}{2}}$.

ii. Αρκεί να υπολογίσουμε τις εκτιμήτριες μέγιστης πιθανοφάνειας $\hat{\widehat{\beta}}_0$, $\hat{\widehat{\beta}}_1$ και $\hat{\widehat{\sigma}}^2$ των β_0, β_1, σ^2 υπό την $H_0 \cup H_1 : \beta_1 \geqslant \beta_{1,0}$. Ξεκινώντας από την εκτιμήτρια $\hat{\widehat{\beta}}_1$, υπολογίζουμε ότι:

$$\hat{\widehat{\beta}}_1 = \begin{cases} \widehat{\beta}_1, & \widehat{\beta}_1 \geqslant \beta_{1,0} \\ \beta_{1,0}, & \widehat{\beta}_1 < \beta_{1,0} \end{cases} \Rightarrow \hat{\widehat{\beta}}_0 = \begin{cases} \widehat{\beta}_0, & \widehat{\beta}_1 \geqslant \beta_{1,0} \\ \widetilde{\beta}_0, & \widehat{\beta}_1 < \beta_{1,0} \end{cases}, \quad \hat{\widehat{\sigma}}^2 = \begin{cases} \widehat{\sigma}^2, & \widehat{\beta}_1 \geqslant \beta_{1,0} \\ \widetilde{\sigma}^2, & \widehat{\beta}_1 < \beta_{1,0} \end{cases}.$$

Υπολογίζουμε τον γενικευμένο λόγο πιθανοφανειών λ^{**} συναρτήσει του λόγου λ^* που υπολογίσαμε για τον αμφίπλευρο έλεγχο:

$$\lambda^{**} = \frac{L\left(\widetilde{\beta}_0, \beta_{1,0}, \widetilde{\sigma}^2\right)}{L\left(\hat{\widehat{\beta}}_0, \hat{\widehat{\beta}}_1, \hat{\widehat{\sigma}}^2\right)} = \begin{cases} \lambda^*, & \widehat{\beta}_1 \geqslant \beta_{1,0} \\ 1, & \widehat{\beta}_1 < \beta_{1,0} \end{cases}.$$

Για τον υπολογισμό της κρίσιμης περιοχής του ελέγχου πρέπει να λύσουμε την ανισότητα $\lambda^{**} < c$. Αφού ισχύει πάντα ότι $0 \leqslant \lambda^{**} \leqslant 1$, αυτό ισοδυναμεί με το να λύσουμε την ανισότητα $\lambda^* < c$ υπό τον περιορισμό $\widehat{\beta}_1 \geqslant \beta_{1,0}$. Όμως, έχουμε δείξει ότι $\lambda^* < c \Leftrightarrow |t| > c_\alpha$. Επιπλέον, υπό τον περιορισμό $\widehat{\beta}_1 \geqslant \beta_{1,0}$, παίρνουμε ότι $t \geqslant 0$, οπότε $|t| = t$. Για τον υπολογισμό της σταθεράς c_α απαιτούμε $P_{H_0}(T > c_\alpha) = \alpha$, δηλαδή $c_\alpha = t_{n-2;\alpha}$. Επομένως, απορρίπτουμε την H_0 αν και μόνο αν $t > t_{n-2;\alpha}$.

iii. Αρκεί να υπολογίσουμε τις εκτιμήτριες μέγιστης πιθανοφάνειας $\hat{\widehat{\beta}}_0$, $\hat{\widehat{\beta}}_1$ και $\hat{\widehat{\sigma}}^2$ των β_0, β_1, σ^2 υπό την $H_0 \cup H_1 : \beta_1 \leqslant \beta_{1,0}$:

$$\hat{\widehat{\beta}}_1 = \begin{cases} \widehat{\beta}_1, & \widehat{\beta}_1 \leqslant \beta_{1,0} \\ \beta_{1,0}, & \widehat{\beta}_1 > \beta_{1,0} \end{cases} \Rightarrow \hat{\widehat{\beta}}_0 = \begin{cases} \widehat{\beta}_0, & \widehat{\beta}_1 \leqslant \beta_{1,0} \\ \widetilde{\beta}_0, & \widehat{\beta}_1 > \beta_{1,0} \end{cases}, \quad \hat{\widehat{\sigma}}^2 = \begin{cases} \widehat{\sigma}^2, & \widehat{\beta}_1 \leqslant \beta_{1,0} \\ \widetilde{\sigma}^2, & \widehat{\beta}_1 > \beta_{1,0} \end{cases}.$$

Υπολογίζουμε τον γενικευμένο λόγο πιθανοφανειών λ^{**} συναρτήσει του λόγου λ^* που υπολογίσαμε για τον αμφίπλευρο έλεγχο:

$$\lambda^{**} = \frac{L\left(\widetilde{\beta}_0, \beta_{1,0}, \widetilde{\sigma}^2\right)}{L\left(\hat{\widehat{\beta}}_0, \hat{\widehat{\beta}}_1, \hat{\widehat{\sigma}}^2\right)} = \begin{cases} \lambda^*, & \widehat{\beta}_1 \leqslant \beta_{1,0} \\ 1, & \widehat{\beta}_1 > \beta_{1,0} \end{cases}.$$

Για τον υπολογισμό της κρίσιμης περιοχής του ελέγχου πρέπει να λύσουμε την ανισότητα $\lambda^{**} < c$. Αφού ισχύει πάντα ότι $0 \leqslant \lambda^{**} \leqslant 1$, αυτό ισοδυναμεί με το να λύσουμε την ανισότητα $\lambda^* < c$ υπό τον περιορισμό $\widehat{\beta}_1 \leqslant \beta_{1,0}$. Όμως, έχουμε δείξει ότι $\lambda^* < c \Leftrightarrow |t| > c_\alpha$. Επιπλέον, υπό τον περιορισμό $\widehat{\beta}_1 \leqslant \beta_{1,0}$, παίρνουμε ότι $t \leqslant 0$, οπότε $|t| = -t$. Για τον υπολογισμό της σταθεράς c_α απαιτούμε $P_{H_0}(T < -c_\alpha) = \alpha$, δηλαδή $c_\alpha = -t_{n-2;1-\alpha} = t_{n-2;\alpha}$. Επομένως, απορρίπτουμε την H_0 αν και μόνο αν $t < -t_{n-2;\alpha}$. □

Σημείωση 1.4. (Σύνδεση Αμφίπλευρου και Μονόπλευρων Ελέγχων Υποθέσεων)

i. p-value$^{(\neq)} = 2P\left(T \geqslant |t|\right)$.

ii. p-value$^{(>)}$ + p-value$^{(<)} = 1$.

iii. p-value$^{(\neq)} = 2\min\left\{\text{p-value}^{(>)}, \text{p-value}^{(<)}\right\}$.

iv. Αν $t > 0$, τότε p-value$^{(>)} = \frac{1}{2} \cdot$ p-value$^{(\neq)}$. Διαφορετικά, αν $t < 0$, τότε p-value$^{(<)} = \frac{1}{2} \cdot$ p-value$^{(\neq)}$.

Σύμφωνα με τα παραπάνω, αν υπολογίσουμε το p-value οποιουδήποτε από αυτούς τους τρεις ελέγχους, τότε μπορούμε πάντα να υπολογίσουμε μέσω αυτού και τα p-value των άλλων δύο.

Σημείωση 1.5. Θεωρούμε τον έλεγχο υποθέσεων $H_0 : \beta_i = \beta_{i,0}$ vs. $H_1 : \beta_i \neq \beta_{i,0}$ για $i = 0, 1$. Αποδείξαμε ότι απορρίπτουμε την H_0 σε ε.σ.σ. α αν και μόνο αν $|t| > t_{n-2;\frac{\alpha}{2}}$. Ισοδύναμα, δεν απορρίπτουμε την H_0 αν και μόνο αν:

$$|t| \leqslant t_{n-2;\frac{\alpha}{2}} \Leftrightarrow -t_{n-2;\frac{\alpha}{2}} \leqslant \frac{\widehat{\beta}_i - \beta_{i,0}}{s_{\widehat{\beta}_i}} \leqslant t_{n-2;\frac{\alpha}{2}} \Leftrightarrow$$

$$-t_{n-2;\frac{\alpha}{2}} \cdot s_{\widehat{\beta}_i} \leqslant \beta_{i,0} - \widehat{\beta}_i \leqslant t_{n-2;\frac{\alpha}{2}} \cdot s_{\widehat{\beta}_i} \Leftrightarrow$$

$$\beta_{i,0} \in \left[\widehat{\beta}_i - t_{n-2;\frac{\alpha}{2}} \cdot s_{\widehat{\beta}_i}, \widehat{\beta}_i + t_{n-2;\frac{\alpha}{2}} \cdot s_{\widehat{\beta}_i}\right] = I_{1-\alpha}(\beta_i).$$

Με άλλα λόγια, απορρίπτουμε την $H_0 : \beta_i = \beta_{i,0}$ αν και μόνο αν $\beta_{i,0} \notin I_{1-\alpha}(\beta_i)$. Αυτή η **δυϊκότητα** μεταξύ διαστημάτων εμπιστοσύνης και ελέγχων υποθέσεων με αμφίπλευρη εναλλακτική μας δίνει έναν εναλλακτικό τρόπο απόφασης για την απόρριψη ή μη της μηδενικής υπόθεσης.

Ορισμός 1.8. Ένας έλεγχος της μορφής $H_0 : \beta_i = 0$ vs. $H_1 : \beta_i \neq 0$ για $i = 0, 1$ καλείται **έλεγχος στατιστικής σημαντικότητας** της παραμέτρου β_i.

Σημείωση 1.6. Σύμφωνα με την πρόταση 1.17, η ελεγχοσυνάρτηση του ελέγχου στατιστικής σημαντικότητας της παραμέτρου β_i προκύπτει θέτοντας $\beta_{i,0} = 0$. Υπό τη μηδενική υπόθεση $H_0 : \beta_i = 0$, γνωρίζουμε ότι $T = \frac{\widehat{\beta}_i}{S_{\widehat{\beta}_i}} \sim t_{n-2}$ με παρατηρούμενη τιμή $t = \frac{\widehat{\beta}_i}{s_{\widehat{\beta}_i}}$. Λέμε ότι η παράμετρος β_i είναι **στατιστικά σημαντική** ή στατιστικά διάφορη του μηδενός σε ε.σ.σ. α αν και μόνο αν μπορούμε να απορρίψουμε την H_0, δηλαδή αν και μόνο αν ισχύει κάποιο από τα εξής:

- $|t| > t_{n-2;\frac{\alpha}{2}}$,
- p-value $= P\left(|T| \geqslant |t|\right) < \alpha$,
- $0 \notin I_{1-\alpha}(\beta_i) = \left[\widehat{\beta}_i - t_{n-2;\frac{\alpha}{2}} \cdot s_{\widehat{\beta}_i}, \widehat{\beta}_i + t_{n-2;\frac{\alpha}{2}} \cdot s_{\widehat{\beta}_i}\right]$.

Αν μία παράμετρος δεν είναι στατιστικά σημαντική, τότε αυτό σημαίνει ότι δε συνεισφέρει σημαντικά στην ερμηνεία της αποκριτικής μεταβλητής, οπότε θα μπορούσε να παραλειφθεί χωρίς να χαθεί σημαντική πληροφορία από το μοντέλο.

1.9 Ανάλυση Διασποράς - ANOVA

Στο απλό γραμμικό μοντέλο μας ενδιαφέρει περισσότερο να μελετήσουμε τον συντελεστή κλίσης της ευθείας παλινδρόμησης, δηλαδή την παράμετρο β_1. Θέλουμε να προσεγγίσουμε με έναν εναλλακτικό τρόπο τον έλεγχο στατιστικής σημαντικότητας $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$. Αρχικά, είδαμε ότι:

$$Q = \frac{\text{SSE}}{\sigma^2} = \frac{(n-2)S^2}{\sigma^2} \sim \chi^2_{n-2}.$$

Υπό τη μηδενική υπόθεση $H_0 : \beta_1 = 0$, γνωρίζουμε ότι $Y_i \sim N\left(\beta_0, \sigma^2\right)$ ανεξάρτητες και ισόνομες για $i = 1, 2, \ldots, n$. Σύμφωνα με τα αποτελέσματα που έχουμε δει στη μαθηματική στατιστική, παίρνουμε ότι:

$$\frac{\text{SST}}{\sigma^2} = \frac{(n-1)S_Y^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2 \sim \chi^2_{n-1}.$$

Επιπλέον, υπό τη μηδενική υπόθεση, ισχύει ότι $\widehat{\beta}_1 \sim N\left(0, \frac{\sigma^2}{(n-1)S_X^2}\right)$, οπότε:

$$Z = \frac{\widehat{\beta}_1 S_X\sqrt{n-1}}{\sigma} \sim N(0,1) \Rightarrow Z^2 = \frac{(n-1)S_X^2\widehat{\beta}_1^2}{\sigma^2} \overset{\text{Λήμμα 1.4}}{=} \frac{\text{SSR}}{\sigma^2} \sim \chi^2_1.$$

Η κατανομή του SSR έχει **έναν** βαθμό ελευθερίας (d.f. - degree of freedom). Ορίζουμε **μέσο άθροισμα τετραγώνων που οφείλεται στην παλινδρόμηση** (mean sum of squares due to regression) το $\text{MSR} = \frac{\text{SSR}}{1} = \text{SSR}$. Σύμφωνα με την πρόταση 1.14 (σελίδα 25), οι τυχαίες μεταβλητές S^2 και $\widehat{\beta}_1$ είναι ανεξάρτητες, οπότε και οι τυχαίες μεταβλητές Q και Z^2 είναι ανεξάρτητες. Τελικά, παίρνουμε ότι:

$$F = \frac{Z^2}{1}\cdot\frac{n-2}{Q} = \frac{\text{SSR}}{\sigma^2}\cdot\frac{(n-2)\sigma^2}{\text{SSE}} = \frac{\text{MSR}}{\text{MSE}} \sim F_{1,n-2}, \text{ υπό την } H_0.$$

Προσπαθούμε να συνδέσουμε τον έλεγχο στατιστικής σημαντικότητας για την παράμετρο β_1 με την ελεγχοσυνάρτηση F. Θυμόμαστε ότι $T = \frac{\widehat{\beta}_1}{S_{\widehat{\beta}_1}} \sim t_{n-2}$ υπό την H_0, οπότε υπολογίζουμε ότι:

$$T^2 = \frac{\widehat{\beta}_1^2}{S^2_{\widehat{\beta}_1}} = \frac{(n-1)S_X^2\widehat{\beta}_1^2}{S^2} = \frac{\text{MSR}}{\text{MSE}} = F \sim F_{1,n-2}.$$

Δείξαμε ότι απορρίπτουμε την H_0 σε ε.σ.σ. α αν και μόνο αν $|t| > c_\alpha = t_{n-2;\frac{\alpha}{2}}$. Για τον υπολογισμό της κρίσιμης περιοχής του ελέγχου F, απαιτούμε:

$$P_{H_0}\left(|T| > c_\alpha\right) = P_{H_0}\left(|T|^2 > c_\alpha^2\right) = P_{H_0}\left(F > c_\alpha^2\right) = \alpha \Rightarrow c_\alpha^2 = F_{1,n-2;\alpha}.$$

Αν f είναι η παρατηρούμενη τιμή της ελεγχοσυνάρτησης F στο συγκεκριμένο δείγμα, τότε απορρίπτουμε την H_0 αν και μόνο αν $f > F_{1,n-2;\alpha}$ ή:

$$\text{p-value} = P\left(|T| \geqslant |t|\right) = P\left(|T|^2 \geqslant |t|^2\right) = P(F \geqslant f) < \alpha.$$

Βλέπουμε, λοιπόν, ότι ο έλεγχος στατιστικής σημαντικότητας της παραμέτρου β_1 με χρήση της κατανομής t είναι ισοδύναμος με τον έλεγχο F. Παρατηρούμε ότι οι βαθμοί ελευθερίας των κατανομών του SSR και του SSE αθροίζουν στους βαθμούς ελευθερίας της κατανομής του SST. Όλα τα παραπάνω τα συνοψίζουμε στον λεγόμενο πίνακα ανάλυσης διασποράς (ANOVA - analysis of variance). Καλείται έτσι επειδή βασίζεται στην ανάλυση της συνολικής μεταβλητότητας SST στις συνιστώσες SSR και SSE.

	Sum of Squares	d.f.	Mean Square	$F_{1,n-2}$	p-value
Regression	$\text{SSR} = \sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2$	1	$\text{MSR} = \frac{\text{SSR}}{1}$	$f = \frac{\text{MSR}}{\text{MSE}}$	$P(F \geqslant f)$
Error	$\text{SSE} = \sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2$	$n-2$	$\text{MSE} = \frac{\text{SSE}}{n-2}$		
Total	$\text{SST} = \sum_{i=1}^{n}(Y_i - \overline{Y})^2$	$n-1$			

Πινακας 1.1: Πίνακας ANOVA για το Κανονικό Απλό Γραμμικό Μοντέλο

Σημείωση 1.7. Μία ισοδύναμη γραφή της ελεγχοσυνάρτησης F δίνεται ως εξής:

$$F = \frac{\text{MSR}}{\text{MSE}} = (n-2)\cdot\frac{\text{SSR}}{\text{SSE}} = (n-2)\cdot\frac{\text{SSR}}{\text{SST}-\text{SSR}} = (n-2)\cdot\frac{R^2}{1-R^2}.$$

Σημείωση 1.8. Ελέγχουμε αν το MSR θα μπορούσε να χρησιμοποιηθεί ως εκτιμήτρια του σ^2, όπως το MSE. Υπολογίζουμε τη μέση τιμή του MSR για να ελέγξουμε αν είναι αμερόληπτη εκτιμήτρια του σ^2:

$$\begin{aligned} E(\text{MSR}) = E(\text{SSR}) &\overset{\text{Λήμμα 1.4}}{=} E\left[(n-1)S_X^2\widehat{\beta}_1^2\right] = (n-1)S_X^2\left[\text{Var}\left(\widehat{\beta}_1\right) + \left(E\left(\widehat{\beta}_1\right)\right)^2\right] \\ &= (n-1)S_X^2\left[\frac{\sigma^2}{(n-1)S_X^2} + \beta_1^2\right] = \sigma^2 + (n-1)S_X^2\beta_1^2 \geqslant \sigma^2. \end{aligned}$$

Επομένως, ισχύει ότι $E(\text{MSR}) = \sigma^2$, δηλαδή το MSR είναι αμερόληπτη εκτιμήτρια του σ^2 αν και μόνο αν ισχύει η μηδενική υπόθεση $H_0 : \beta_1 = 0$. Στη γενικότερη

περίπτωση το MSR υπερεκτιμά το σ^2.

Σημείωση 1.9. Σχετικά με το SST, υπολογίζουμε ότι:

$$E\left(\text{SST}\right) = E(\text{SSR}) + E(\text{SSE}) \overset{\text{Πρόταση } 1.5}{=} \sigma^2 + (n-1)S_X^2\beta_1^2 + (n-2)\sigma^2 \Rightarrow$$

$$E\left(\frac{\text{SST}}{n-1}\right) = E\left(S_Y^2\right) = \sigma^2 + S_X^2\beta_1^2 \geqslant \sigma^2.$$

Δηλαδή η στατιστική συνάρτηση S_Y^2 είναι αμερόληπτη εκτιμήτρια του σ^2 αν και μόνο αν ισχύει η μηδενική υπόθεση $H_0 : \beta_1 = 0$, δηλαδή οι παρατηρήσεις Y_i είναι ανεξάρτητες και ισόνομες. Στη γενικότερη περίπτωση το S_Y^2 υπερεκτιμά το σ^2.

1.10 Διαστήματα Μέσης και Ατομικής Πρόβλεψης

Για δεδομένο ζεύγος παρατηρήσεων (X_i, Y_i), θέλουμε αρχικά να κατασκευάσουμε διάστημα εμπιστοσύνης για τη $E(Y_i) = \beta_0 + \beta_1 X_i$. Αυτό το διάστημα εμπιστοσύνης καλείται και **διάστημα μέσης πρόβλεψης** για το Y_i. Γνωρίζουμε ότι $E(\widehat{Y}_i) = \beta_0 + \beta_1 X_i = E(Y_i)$, δηλαδή το $\widehat{Y}_i$ είναι μία αμερόληπτη εκτιμήτρια της $E(Y_i)$. Επιπλέον, γνωρίζουμε ότι το $\widehat{Y}_i$ είναι γραμμικός συνδυασμός των $\widehat{\beta}_0$ και $\widehat{\beta}_1$, τα οποία με τη σειρά τους είναι γραμμικοί συνδυασμοί των $Y_1, Y_2, \ldots, Y_n$. Εφόσον τα $Y_1, Y_2, \ldots, Y_n$ είναι ανεξάρτητα και κανονικά κατανεμημένα, συμπεραίνουμε ότι και το $\widehat{Y}_i$ θα είναι κανονικά κατανεμημένο με μέση τιμή και διασπορά που έχουμε υπολογίσει στην πρόταση 1.4, δηλαδή:

$$\widehat{Y}_i \sim N\left(\beta_0 + \beta_1 X_i, \sigma^2\left[\frac{1}{n} + \frac{(X_i - \overline{X})^2}{(n-1)S_X^2}\right]\right) \Rightarrow$$

$$Z = \frac{\widehat{Y}_i - E(Y_i)}{\sigma}\left[\frac{1}{n} + \frac{(X_i - \overline{X})^2}{(n-1)S_X^2}\right]^{-\frac{1}{2}} \sim N(0,1).$$

Επιπλέον, γνωρίζουμε ότι $Q = \frac{(n-2)S^2}{\sigma^2} \sim \chi^2_{n-2}$. Εφόσον η τυχαία μεταβλητή S^2 είναι ανεξάρτητη από το $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_0, \widehat{\beta}_1)$, θα είναι ανεξάρτητη και από το $\widehat{Y}_i$. Συμπεραίνουμε ότι η τυχαία μεταβλητή Z είναι ανεξάρτητη από την Q, οπότε:

$$T = \frac{Z}{\sqrt{\frac{Q}{n-2}}} = \frac{\widehat{Y}_i - E(Y_i)}{\sigma}\left[\frac{1}{n} + \frac{(X_i - \overline{X})^2}{(n-1)S_X^2}\right]^{-\frac{1}{2}}\frac{\sigma}{S} = \frac{\widehat{Y}_i - E(Y_i)}{S_{\widehat{Y}_i}} \sim t_{n-2}, \text{ όπου:}$$

$$S_{\widehat{Y}_i}^2 = S^2\left[\frac{1}{n} + \frac{(X_i - \overline{X})^2}{(n-1)S_X^2}\right].$$

Ζητάμε σταθερές $c_1, c_2 \in \mathbb{R}$ τέτοιες, ώστε $P(c_1 \leqslant T \leqslant c_2) = 1 - \alpha$. Συγκεκριμένα, για το διάστημα εμπιστοσύνης ίσων ουρών χρησιμοποιούμε τις σχέσεις:

$$P(T < c_1) = \frac{\alpha}{2} \Rightarrow P(T > c_1) = 1 - \frac{\alpha}{2} \Rightarrow c_1 = t_{n-2;1-\frac{\alpha}{2}} = -t_{n-2;\frac{\alpha}{2}},$$

$$P(T > c_2) = \frac{\alpha}{2} \Rightarrow c_2 = t_{n-2;\frac{\alpha}{2}}.$$

Επομένως, το διάστημα εμπιστοσύνης ίσων ουρών δίνεται από τη σχέση:

$$c_1 \leqslant \frac{\widehat{Y}_i - E(Y_i)}{S_{\widehat{Y}_i}} \leqslant c_2 \Leftrightarrow -c_2 \cdot S_{\widehat{Y}_i} \leqslant E(Y_i) - \widehat{Y}_i \leqslant -c_1 \cdot S_{\widehat{Y}_i} \Leftrightarrow$$

$$\widehat{Y}_i - t_{n-2;\frac{\alpha}{2}} \cdot S_{\widehat{Y}_i} \leqslant E(Y_i) \leqslant \widehat{Y}_i + t_{n-2;\frac{\alpha}{2}} \cdot S_{\widehat{Y}_i}.$$

Τελικά, παίρνουμε ότι $I_{1-\alpha}\left(E(Y_i)\right) = \left[\widehat{Y}_i - t_{n-2;\frac{\alpha}{2}} \cdot S_{\widehat{Y}_i}, \widehat{Y}_i + t_{n-2;\frac{\alpha}{2}} \cdot S_{\widehat{Y}_i}\right]$.

Τώρα, έχοντας μία νέα παρατήρηση X_{n+1}, ενδιαφερόμαστε να κατασκευάσουμε διάστημα πρόβλεψης για την τιμή $Y_{n+1} = \beta_0 + \beta_1 X_{n+1} + \varepsilon_{n+1}$ την οποία δεν έχουμε παρατηρήσει, όπου $\varepsilon_{n+1} \sim N\left(0, \sigma^2\right)$ ανεξάρτητο από τα $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$. Αυτό το διάστημα πρόβλεψης καλείται και **διάστημα ατομικής πρόβλεψης** για το Y_i. Με βάση τις προηγούμενες n παρατηρήσεις $Y_1, Y_2, \ldots, Y_n$, έχουμε υπολογίσει τις εκτιμήτριες ελαχίστων τετραγώνων $\widehat{\beta}_0$ και $\widehat{\beta}_1$. Με βάση αυτές τις εκτιμήτριες, έχουμε ορίσει πρόβλεψη $\widetilde{Y}_{n+1} = \widehat{\beta}_0 + \widehat{\beta}_1 X_{n+1}$ και σφάλμα σφάλμα πρόβλεψης $\widetilde{\varepsilon}_{n+1} = Y_{n+1} - \widetilde{Y}_{n+1} = Y_{n+1} - \widehat{\beta}_0 - \widehat{\beta}_1 X_{n+1}$.

Το σφάλμα πρόβλεψης $\widetilde{\varepsilon}_{n+1}$ είναι γραμμικός συνδυασμός των Y_{n+1}, $\widehat{\beta}_0$ και $\widehat{\beta}_1$. Τα $\widehat{\beta}_0$ και $\widehat{\beta}_1$ με τη σειρά τους είναι γραμμικοί συνδυασμοί των $Y_1, Y_2, \ldots, Y_n$. Εφόσον τα $Y_1, Y_2, \ldots, Y_n, Y_{n+1}$ είναι ανεξάρτητα και κανονικά κατανεμημένα, συμπεραίνουμε ότι και το $\widetilde{\varepsilon}_{n+1}$ θα είναι κανονικά κατανεμημένο με μέση τιμή και διασπορά που έχουμε υπολογίσει στην πρόταση 1.8, δηλαδή:

$$\widetilde{\varepsilon}_{n+1} = Y_{n+1} - \widehat{\beta}_0 - \widehat{\beta}_1 X_{n+1} \sim N\left(0, \sigma^2\left[1 + \frac{1}{n} + \frac{(X_{n+1} - \overline{X})^2}{(n-1)S_X^2}\right]\right) \Rightarrow$$

$$Z = \frac{Y_{n+1} - \widehat{\beta}_0 - \widehat{\beta}_1 X_{n+1}}{\sigma}\left[1 + \frac{1}{n} + \frac{(X_{n+1} - \overline{X})^2}{(n-1)S_X^2}\right]^{-\frac{1}{2}} \sim N(0,1).$$

Εφόσον το $\left(\widehat{\boldsymbol{\beta}}, S^2\right)$ είναι συνάρτηση των $Y_1, Y_2, \ldots, Y_n$, θα είναι ανεξάρτητο από το Y_{n+1}. Όμως, το $\widehat{\boldsymbol{\beta}}$ είναι ανεξάρτητο από το S^2, οπότε τα $\widehat{\boldsymbol{\beta}}$, S^2 και Y_{n+1} είναι αμοιβαία ανεξάρτητα μεταξύ τους. Επομένως, η τυχαία μεταβλητή S^2 είναι ανεξάρτητη από το $\left(\widehat{\boldsymbol{\beta}}, Y_{n+1}\right)$. Συμπεραίνουμε ότι η τυχαία μεταβλητή Z είναι

ανεξάρτητη από την τυχαία μεταβλητή $Q = \frac{(n-2)S^2}{\sigma^2} \sim \chi^2_{n-2}$, οπότε:

$$T = \frac{Z}{\sqrt{\frac{Q}{n-2}}} = \frac{Y_{n+1} - \widetilde{Y}_{n+1}}{\sigma} \left[1 + \frac{1}{n} + \frac{(X_i - \overline{X})^2}{(n-1)S_X^2}\right]^{-\frac{1}{2}} \frac{\sigma}{S} = \frac{Y_{n+1} - \widetilde{Y}_{n+1}}{S_{\widetilde{\varepsilon}_{n+1}}} \sim t_{n-2},$$

$$\text{όπου } S^2_{\widetilde{\varepsilon}_{n+1}} = S^2 \left[1 + \frac{1}{n} + \frac{(X_i - \overline{X})^2}{(n-1)S_X^2}\right].$$

Ζητάμε σταθερές $c_1, c_2 \in \mathbb{R}$ τέτοιες, ώστε $P(c_1 \leqslant T \leqslant c_2) = 1 - \alpha$. Όπως και πριν, παίρνουμε ότι $c_1 = -t_{n-2;\frac{\alpha}{2}}$ και $c_2 = t_{n-2;\frac{\alpha}{2}}$. Επομένως, το διάστημα εμπιστοσύνης ίσων ουρών δίνεται από τη σχέση:

$$c_1 \leqslant \frac{Y_{n+1} - \widetilde{Y}_{n+1}}{S_{\widetilde{\varepsilon}_{n+1}}} \leqslant c_2 \Leftrightarrow \widetilde{Y}_{n+1} + c_1 \cdot S_{\widetilde{\varepsilon}_{n+1}} \leqslant Y_{n+1} \leqslant \widetilde{Y}_{n+1} + c_2 \cdot S_{\widetilde{\varepsilon}_{n+1}}.$$

Τελικά, παίρνουμε ότι $I_{1-\alpha}(Y_{n+1}) = \left[\widetilde{Y}_{n+1} - t_{n-2;\frac{\alpha}{2}} \cdot S_{\widetilde{\varepsilon}_{n+1}}, \widetilde{Y}_{n+1} + t_{n-2;\frac{\alpha}{2}} \cdot S_{\widetilde{\varepsilon}_{n+1}}\right]$.

Παρατήρηση 1.4. Θέλουμε να συγκρίνουμε τα διαστήματα μέσης και ατομικής πρόβλεψης για το Y_{n+1}. Παρατηρούμε ότι και τα δύο διαστήματα πρόβλεψης είναι εστιασμένα γύρω από το $\widehat{Y}_{n+1} = \widetilde{Y}_{n+1} = \widehat{\beta}_0 + \widehat{\beta}_1 X_{n+1}$. Όμως, βλέπουμε ότι $S^2_{\widetilde{\varepsilon}_{n+1}} = S^2 + S^2_{\widehat{Y}_{n+1}}$. Το διάστημα μέσης πρόβλεψης για το Y_{n+1} έχει μήκος:

$$\lambda\left(I_{1-\alpha}\left(E(Y_{n+1})\right)\right) = 2t_{n-2;\frac{\alpha}{2}} S_{\widehat{Y}_{n+1}}.$$

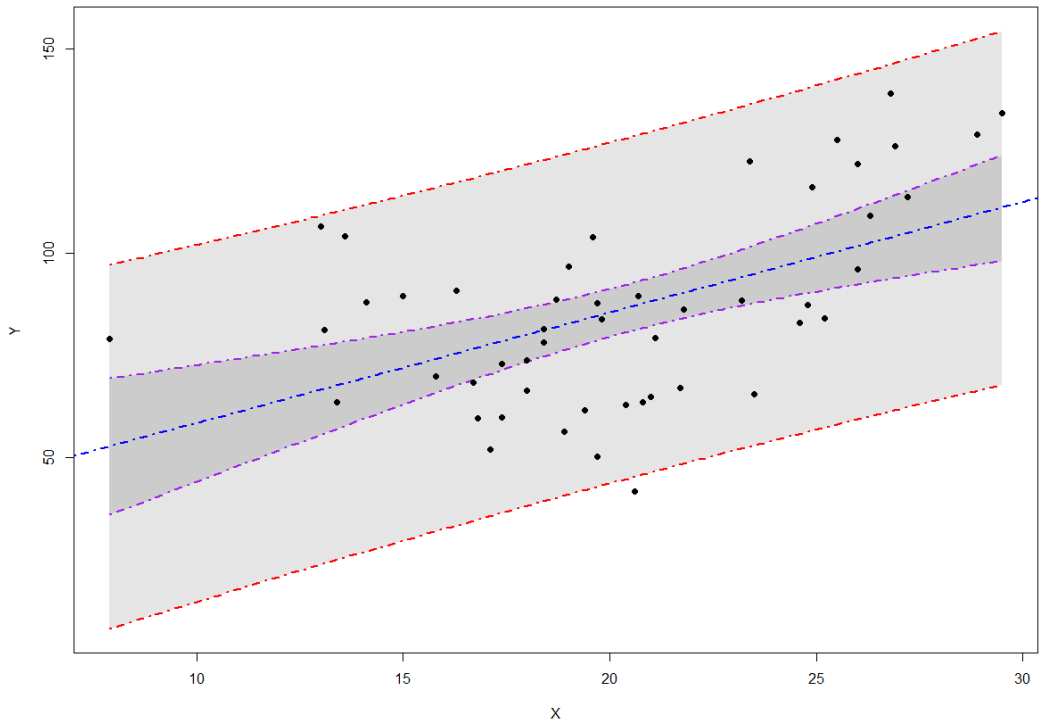
Το αντίστοιχο διάστημα ατομικής πρόβλεψης για το Y_{n+1} έχει μήκος:

$$\begin{aligned}\lambda\left(I_{1-\alpha}\left(Y_{n+1}\right)\right) &= 2t_{n-2;\frac{\alpha}{2}} S_{\widetilde{\varepsilon}_{n+1}} \\ &= 2t_{n-2;\frac{\alpha}{2}} \sqrt{S^2 + S^2_{\widehat{Y}_{n+1}}} > 2t_{n-2;\frac{\alpha}{2}} S_{\widehat{Y}_{n+1}} = \lambda\left(I_{1-\alpha}\left(E(Y_{n+1})\right)\right).\end{aligned}$$

Επομένως, το μήκος του διαστήματος ατομικής πρόβλεψης για το Y_{n+1} είναι πάντα μεγαλύτερο από αυτό του αντίστοιχου διαστήματος μέσης πρόβλεψης. Αυτή η διαφορά στα μήκη των διαστημάτων οφείλεται στις διαφορετικές πηγές αβεβαιότητας που λαμβάνουν υπόψη. Το διάστημα μέσης πρόβλεψης για το Y_{n+1} λαμβάνει υπόψη του μόνο την αβεβαιότητα για τη μέση τιμή $E(Y_{n+1})$. Από την άλλη μεριά, το διάστημα ατομικής πρόβλεψης για το Y_{n+1} συνυπολογίζει και την αβεβαιότητα για την ίδια την παρατήρηση Y_{n+1}, η οποία είναι άγνωστη.

Οι εκτιμήτριες $S^2_{\widehat{Y}_{n+1}}$ και $S^2_{\widetilde{\varepsilon}_{n+1}}$ των διασπορών $\mathrm{Var}(\widehat{Y}_{n+1})$ και $\mathrm{Var}(\widetilde{\varepsilon}_{n+1})$ δεν είναι σταθερές, αλλά εξαρτώνται από την τιμή της παρατήρησης X_{n+1}. Όσο μεγαλύτερη είναι η απόσταση της παρατήρησης X_{n+1} από τη δειγματική μέση τιμή $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, τόσο μεγαλύτερη είναι η τιμή της παράστασης $\left(X_{n+1} - \overline{X}\right)^2$, οπότε τόσο μεγαλύτερα είναι και τα μήκη των διαστημάτων πρόβλεψης. Για $X_{n+1} = \overline{X}$, τα μήκη των διαστημάτων πρόβλεψης παίρνουν την ελάχιστη τιμή τους. Μάλιστα,

αφού τα άκρα των διαστημάτων πρόβλεψης εξαρτώνται από το X_{n+1} μέσω της παράστασης $\left(X_{n+1} - \overline{X}\right)^2$, συμπεραίνουμε ότι έχουν παραβολικό σχήμα.



Σχημα 1.3: Διαστήματα Μέσης και Ατομικής Πρόβλεψης

Στο σχήμα 1.3, είναι σχεδιασμένη με μπλε χρώμα η εκτιμημένη ευθεία παλινδρόμησης των παρατηρήσεων (X_i, Y_i), οι οποίες φαίνονται με μαύρο χρώμα. Κοντά στην εκτιμημένη ευθεία παλινδρόμησης, βλέπουμε σχεδιασμένα με μωβ χρώμα τα άκρα των διαστημάτων μέσης πρόβλεψης για διάφορες τιμές του X_{n+1}. Επιβεβαιώνουμε ότι τα δύο αυτά άκρα σχηματίζουν δύο καμπύλες που φαίνονται να έχουν παραβολικό σχήμα.

Τα άκρα των διαστημάτων ατομικής πρόβλεψης, τα οποία είναι σχεδιασμένα με κόκκινο χρώμα, βρίσκονται πολύ πιο μακριά από την εκτιμημένη ευθεία παλινδρόμησης, όπως ακριβώς περιμέναμε. Τέλος, παρότι δεν απεικονίζεται καλά λόγω της κλίμακας του γραφήματος, τα άκρα των διαστημάτων ατομικής πρόβλεψης αποτελούν και αυτά τμήματα δύο παραβολικών καμπυλών.

Κεφάλαιο 2

Πολλαπλή Γραμμική Παλινδρόμηση

2.1 Εισαγωγή

Όταν δεν έχουμε μόνο μία επεξηγηματική μεταβλητή X, αλλά p απαντητικές μεταβλητές $X_1, X_2, \ldots, X_p$ και θέλουμε να τις χρησιμοποιήσουμε όλες για να προβλέψουμε τις τιμές της αποκριτικής μεταβλητής Y, τότε κατασκευάζουμε ένα **πολλαπλό γραμμικό μοντέλο** ή **μοντέλο πολλαπλής γραμμικής παλινδρόμησης.** Με βάση ένα δείγμα μεγέθους n από διανυσματικές παρατηρήσεις

$$(Y_1, X_{1,1}, X_{2,1}, \ldots, X_{p,1})$$

$$(Y_2, X_{1,2}, X_{2,2}, \ldots, X_{p,2})$$

$$\vdots$$

$$(Y_n, X_{1,n}, X_{2,n}, \ldots, X_{p,n})$$

ορίζουμε το πολλαπλό γραμμικό μοντέλο

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \cdots + \beta_p X_{p,i} + \varepsilon_i, \quad i = 1, 2, \ldots, n,$$

όπου $\beta_0, \beta_1, \ldots, \beta_p$ οι $p+1$ **συντελεστές παλινδρόμησης** και ε_i τα **τυχαία σφάλματα**, για τα οποία ισχύει ότι:

- $E(\varepsilon_i) = 0$,
- $\mathrm{Var}(\varepsilon_i) = \sigma^2$, όπου σ^2 άγνωστη παράμετρος,
- $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0$ για $i \neq j$.

Με άλλα λόγια, τα τυχαία σφάλματα έχουν **μέση τιμή 0**, είναι **ομοσκεδαστικά**, δηλαδή έχουν κοινή διασπορά σ^2, και είναι **ασυσχέτιστα**, αλλά όχι απαραίτητα ανεξάρτητα, ακριβώς όπως στο απλό γραμμικό μοντέλο. Μάλιστα, το απλό γραμμικό μοντέλο είναι ειδική περίπτωση του πολλαπλού γραμμικού μοντέλου για πλήθος επεξηγηματικών μεταβλητών $p = 1$. Για τον λόγο αυτό, όλα τα αποτελέσματα που θα δείξουμε στα πλαίσια αυτού του κεφαλαίου ισχύουν αυτομάτως και για το απλό γραμμικό μοντέλο.

Συνοψίζοντας, σε ένα πολλαπλό γραμμικό μοντέλο έχουμε p επεξηγηματικές μεταβλητές και $p + 2$ παραμέτρους προς εκτίμηση. Θα πρέπει, προφανώς, το μέγεθος του δείγματος να είναι επαρκές, ώστε να μπορούμε να εκτιμήσουμε όλες τις παραμέτρους του μοντέλου, δηλαδή να ισχύει $n \geqslant p + 2$.

Λόγω του μεγάλου πλήθους παραμέτρων στο πολλαπλό γραμμικό μοντέλο, είναι περισσότερο βολικό να το γράψουμε και να το επεξεργαστούμε περαιτέρω σε **πινακική μορφή**, δηλαδή $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, όπου:

$$\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)^{\mathrm{T}} \in \mathbb{R}^n,$$

$$\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^{\mathrm{T}} \in \mathbb{R}^{p+1},$$

$$\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)^{\mathrm{T}} \in \mathbb{R}^n \text{ και}$$

$$\mathbf{X} = \begin{bmatrix} 1 & X_{1,1} & X_{2,1} & \cdots & X_{p,1} \\ 1 & X_{1,2} & X_{2,2} & \cdots & X_{p,2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{1,n} & X_{2,n} & \cdots & X_{p,n} \end{bmatrix} \in \mathbb{R}^{n\times(p+1)}.$$

Ο πίνακας $\mathbf{X}$ καλείται **πίνακας σχεδιασμού** του γραμμικού μοντέλου. Η πρώτη στήλη με τις μονάδες αντιστοιχεί στον σταθερό όρο β_0, ενώ καθεμία από τις υπόλοιπες στήλες αποτελεί ένα δείγμα μεγέθους n για κάποια από τις p επεξηγηματικές μεταβλητές.

Οι 3 υποθέσεις που έχουμε κάνει για τα τυχαία σφάλματα του γραμμικού μοντέλου συνοψίζονται σε πινακική μορφή ως $E(\boldsymbol{\varepsilon}) = \mathbf{0}_n$ και $\mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}_n$. Εκτός από αυτές, πρέπει να κάνουμε άλλη μία υπόθεση σχετικά με τον πίνακα $\mathbf{X}$, ώστε το γραμμικό μοντέλο να είναι καλά ορισμένο. Θα πρέπει ο πίνακας σχεδιασμού $\mathbf{X}$ να είναι **πλήρους τάξης** (full rank), δηλαδή $\mathrm{rank}(\mathbf{X}) = p + 1$. Η σημασία αυτής της υπόθεσης θα φανεί στην εφαρμογή της μεθόδου ελαχίστων τετραγώνων.

Πρόταση 2.1. Ισχύει ότι $E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$ και $\mathrm{Var}(\mathbf{Y}) = \sigma^2\mathbf{I}_n$. Συμπεραίνουμε ότι:

i. $E(Y_i) = \beta_0 + \beta_1 X_{1,i} + \cdots + \beta_p X_{p,i} \Rightarrow E\left(\overline{Y}\right) = \beta_0 + \beta_1\overline{X}_1 + \cdots + \beta_p\overline{X}_p$.

ii. $\mathrm{Var}(Y_i) = \sigma^2$ και $\mathrm{Cov}(Y_i, Y_j) = 0$ για $i \neq j \Rightarrow \mathrm{Var}\left(\overline{Y}\right) = \frac{\sigma^2}{n}$.

Απόδειξη. Σύμφωνα με την πρόταση 1.9, έχουμε ότι:

$$E(\mathbf{Y}) = E(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = \mathbf{X}\boldsymbol{\beta} + E(\boldsymbol{\varepsilon}) = \mathbf{X}\boldsymbol{\beta},$$

$$\mathrm{Var}(\mathbf{Y}) = \mathrm{Var}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = \mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}_n. \quad \square$$

Ερμηνεία: Η σχέση $E(Y) = \beta_0 + \beta_1 X_1 + \beta_1 X_2 + \cdots + \beta_p X_p$ δίνει τη λεγόμενη **εξίσωση παλινδρόμησης**. Στην περίπτωση του πολλαπλού γραμμικού μοντέλου, η εξίσωση αυτή ορίζει ένα υπερεπίπεδο στον $\mathbb{R}^{p+1}$. Για παράδειγμα, θέτοντας $p = 2$, η σχέση $E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ ορίζει ένα επίπεδο στον $\mathbb{R}^3$. Επομένως, **μόνο** στην περίπτωση του απλού γραμμικού μοντέλου μπορούμε να κάνουμε λόγο για ευθεία παλινδρόμησης. Οι συντελεστές παλινδρόμησης $\beta_0, \beta_1, \ldots, \beta_p$ έχουν τις εξής ερμηνείες:

- Για $X_1 = X_2 = \cdots = X_p = 0$, έχουμε $E(Y) = \beta_0$, δηλαδή ο συντελεστής β_0 αποτελεί την αναμενόμενη τιμή της εξαρτημένης μεταβλητής για τιμή όλων των ανεξάρτητων μεταβλητών ίση με το 0. Προφανώς, η παράμετρος β_0 μετριέται στην ίδια μονάδα με την εξαρτημένη μεταβλητή Y.
- Έστω $j \in \{1, 2, \ldots, p\}$. Για $X_j = X_{j,0} + 1$, έχουμε ότι $E(Y) = E(Y_0) + \beta_j \Rightarrow \beta_j = E(Y) - E(Y_0)$, δηλαδή ο συντελεστής β_j εκφράζει τη μεταβολή στην αναμενόμενη τιμή της εξαρτημένης μεταβλητής για αύξηση της ανεξάρτητης μεταβλητής X_j κατά μία μονάδα, κρατώντας όλες τις υπόλοιπες ανεξάρτητες μεταβλητές σταθερές. Η παράμετρος β_j μετριέται σε μονάδα της εξαρτημένης μεταβλητής Y ανά μονάδα της ανεξάρτητης μεταβλητής X_j.

2.2 Μέθοδος Ελαχίστων Τετραγώνων

Σε πινακική μορφή, η συνάρτηση του αθροίσματος των τετραγωνικών αποστάσεων των παρατηρήσεων Y_i από τις μέσες τιμές τους γράφεται ως:

$$Q(\boldsymbol{\beta}) = \|\boldsymbol{\varepsilon}\|^2 = \|\mathbf{Y} - E(\mathbf{Y})\|^2 = \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}).$$

Τη συνάρτηση αυτή θέλουμε να την ελαχιστοποιήσουμε για να προσδιορίσουμε την εκτιμήτρια ελαχίστων τετραγώνων $\widehat{\boldsymbol{\beta}}$ της διανυσματικής παραμέτρου $\boldsymbol{\beta}$. Για να το επιτύχουμε αυτό, θα πρέπει να παραγωγίσουμε τη συνάρτηση $Q(\boldsymbol{\beta})$ ως προς τη διανυσματική παράμετρο $\boldsymbol{\beta}$, οπότε χρειάζεται να έχουμε κάποιες γνώσεις παραγώγισης ως προς διάνυσμα.

Λήμμα 2.1. (Διανυσματικός Διαφορικός Λογισμός)

i. Έστω $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^n$ μία διανυσματική συνάρτηση και $g(\mathbf{x}) = \|\mathbf{f}(\mathbf{x})\|^2$, $\forall \mathbf{x} \in \mathbb{R}^d$.

Τότε,

$$\frac{\partial g}{\partial \mathbf{x}}(\mathbf{x}) = 2 \cdot \frac{\partial \mathbf{f}}{\partial \mathbf{x}}(\mathbf{x}) \cdot \mathbf{f}(\mathbf{x}).$$

ii. Έστω $\mathbf{f}(\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{b}$, όπου $\mathbf{A} \in \mathbb{R}^{n \times d}$, $\mathbf{x} \in \mathbb{R}^d$ και $\mathbf{b} \in \mathbb{R}^n$. Τότε,

$$\frac{\partial \mathbf{f}}{\partial \mathbf{x}}(\mathbf{x}) = \mathbf{A}^{\mathrm{T}}.$$

Λήμμα 2.2. (Τάξη Πίνακα)

i. Για κάθε πίνακα $\mathbf{A} \in \mathbb{R}^{n \times d}$, ισχύει ότι:

$$\operatorname{rank}(\mathbf{A}^{\mathrm{T}}\mathbf{A}) = \operatorname{rank}(\mathbf{A}\mathbf{A}^{\mathrm{T}}) = \operatorname{rank}(\mathbf{A}) = \operatorname{rank}(\mathbf{A}^{\mathrm{T}}).$$

ii. Έστω $\mathbf{A} \in \mathbb{R}^{n \times n}$ ένας τετραγωνικός πίνακας. Ο πίνακας $\mathbf{A}$ είναι αντιστρέψιμος αν και μόνο αν είναι πλήρους τάξης, δηλαδή $\operatorname{rank}(\mathbf{A}) = n$.

Πόρισμα 2.1. Έστω $\mathbf{X} \in \mathbb{R}^{n \times (p+1)}$ ο πίνακας σχεδιασμού του γραμμικού μοντέλου. Τότε, ο πίνακας $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ είναι αντιστρέψιμος.

Απόδειξη. Γνωρίζουμε ότι ο πίνακας σχεδιασμού $\mathbf{X}$ είναι πλήρους τάξης, δηλαδή $\operatorname{rank}(\mathbf{X}) = p + 1$. Ο πίνακας $\mathbf{X}^{\mathrm{T}}\mathbf{X} \in \mathbb{R}^{(p+1)\times(p+1)}$ είναι τετραγωνικός και ισχύει ότι $\operatorname{rank}(\mathbf{X}^{\mathrm{T}}\mathbf{X}) = \operatorname{rank}(\mathbf{X}) = p + 1$, δηλαδή είναι πλήρους τάξης. Σύμφωνα με το προηγούμενο λήμμα, ο πίνακας $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ είναι αντιστρέψιμος. □

Λήμμα 2.3. Έστω $\mathbf{A} \in \mathbb{R}^{n \times d}$. Τότε, ο πίνακας $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ είναι θετικά ημιορισμένος. Αν, επιπλέον, ο πίνακας $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ είναι πλήρους τάξης, τότε είναι θετικά ορισμένος.

Πρόταση 2.2. Η εκτιμήτρια ελαχίστων τετραγώνων $\widehat{\boldsymbol{\beta}}$ της διανυσματικής παραμέτρου $\boldsymbol{\beta}$ δίνεται από τη σχέση $\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}$.

Απόδειξη. Έστω $\mathbf{f}(\boldsymbol{\beta}) = \mathbf{Y} - \mathbf{X}\boldsymbol{\beta}$. Σύμφωνα με το λήμμα 2.1, έχουμε ότι:

$$\frac{\partial \mathbf{f}}{\partial \boldsymbol{\beta}}(\boldsymbol{\beta}) = -\mathbf{X}^{\mathrm{T}} \Rightarrow \frac{\partial Q}{\partial \boldsymbol{\beta}}(\boldsymbol{\beta}) = 2 \cdot \frac{\partial \mathbf{f}}{\partial \boldsymbol{\beta}}(\boldsymbol{\beta}) \cdot \mathbf{f}(\boldsymbol{\beta}) = -2\mathbf{X}^{\mathrm{T}}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) = 0 \Rightarrow$$

$$\mathbf{X}^{\mathrm{T}}\mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}^{\mathrm{T}}\mathbf{Y} \overset{\text{Πόρισμα 2.1}}{\Rightarrow} \widehat{\boldsymbol{\beta}} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}.$$

Για να ελέγξουμε αν το $\widehat{\boldsymbol{\beta}}$ είναι σημείο μεγίστου, υπολογίζουμε τον Εσσιανό πίνακα της συνάρτησης $Q(\boldsymbol{\beta})$:

$$\frac{\partial Q}{\partial \boldsymbol{\beta}}(\boldsymbol{\beta}) = -2\mathbf{X}^{\mathrm{T}}\mathbf{Y} + 2\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta} \Rightarrow H(\boldsymbol{\beta}) = \frac{\partial^2 Q}{\partial \boldsymbol{\beta}^{\mathrm{T}} \partial \boldsymbol{\beta}}(\boldsymbol{\beta}) = 2\mathbf{X}^{\mathrm{T}}\mathbf{X}.$$

Εφόσον ο $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ είναι πλήρους τάξης, σύμφωνα με το λήμμα 2.3 είναι θετικά ορισμένος. Αφού ο Εσσιανός πίνακας της $Q(\boldsymbol{\beta})$ είναι θετικά ορισμένος, η $Q(\boldsymbol{\beta})$

είναι γνησίως κυρτή και έχει μοναδικό ολικό ελάχιστο το $\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}$. □

Ορισμός 2.1. Έστω $\mathbf{X} = (X_1, X_2, \ldots, X_d) \in \mathbb{R}^d$ και $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n) \in \mathbb{R}^n$ τυχαία διανύσματα. Τότε, ορίζουμε τον πίνακα διασυνδιακύμανσης μεταξύ $\mathbf{X}$ και $\mathbf{Y}$ ως $\mathrm{Cov}(\mathbf{X}, \mathbf{Y}) = E\left[(\mathbf{X} - E(\mathbf{X}))\,(\mathbf{Y} - E(\mathbf{Y}))^{\mathrm{T}}\right] = E\left(\mathbf{X}\mathbf{Y}^{\mathrm{T}}\right) - E(\mathbf{X})\left[E(\mathbf{Y})\right]^{\mathrm{T}} \in \mathbb{R}^{d\times n}$, δηλαδή $\mathrm{Cov}(\mathbf{X}, \mathbf{Y}) = \left[\mathrm{Cov}(X_i, Y_j)\right]_{i,j}$.

Πρόταση 2.3. Έστω $\mathbf{X} \in \mathbb{R}^d$, $\mathbf{Y} \in \mathbb{R}^k$ τυχαία διανύσματα, $\mathbf{A} \in \mathbb{R}^{n\times d}$, $\mathbf{B} \in \mathbb{R}^{n\times k}$ σταθεροί πίνακες και $\mathbf{b} \in \mathbb{R}^n$, $\mathbf{c} \in \mathbb{R}^n$ σταθερά διανύσματα. Τότε, γνωρίζουμε ότι:

i. $\mathrm{Cov}(\mathbf{X}, \mathbf{b}) = \mathbf{0}_{d\times n}$,

ii. $\mathrm{Cov}(\mathbf{X}, \mathbf{X}) = \mathrm{Var}(\mathbf{X})$,

iii. $\mathrm{Cov}(\mathbf{Y}, \mathbf{X}) = \left[\mathrm{Cov}(\mathbf{X}, \mathbf{Y})\right]^{\mathrm{T}}$,

iv. $\mathrm{Cov}(\mathbf{A}\mathbf{X} + \mathbf{b}, \mathbf{B}\mathbf{Y} + \mathbf{c}) = \mathbf{A}\mathrm{Cov}(\mathbf{X}, \mathbf{Y})\mathbf{B}^{\mathrm{T}}$,

v. $\mathrm{Var}(\mathbf{A}\mathbf{X} + \mathbf{B}\mathbf{Y}) = \mathbf{A}\mathrm{Var}(\mathbf{X})\mathbf{A}^{\mathrm{T}} + \mathbf{B}\mathrm{Var}(\mathbf{Y})\mathbf{B}^{\mathrm{T}} + \mathbf{A}\mathrm{Cov}(\mathbf{X}, \mathbf{Y})\mathbf{B}^{\mathrm{T}} + \mathbf{B}\mathrm{Cov}(\mathbf{Y}, \mathbf{X})\mathbf{A}^{\mathrm{T}}$,

vi. $\mathbf{X}, \mathbf{Y}$ ανεξάρτητα $\Rightarrow \mathrm{Cov}(\mathbf{X}, \mathbf{Y}) = \mathbf{0}_{d\times k} \Rightarrow E\left(\mathbf{X}\mathbf{Y}^{\mathrm{T}}\right) = E(\mathbf{X})\left[E(\mathbf{Y})\right]^{\mathrm{T}}$.

Πρόταση 2.4. (Εκτιμήτριες Ελαχίστων Τετραγώνων)

i. $E\left(\widehat{\boldsymbol{\beta}}\right) = \boldsymbol{\beta}$, δηλαδή η $\widehat{\boldsymbol{\beta}}$ είναι **αμερόληπτη εκτιμήτρια** του $\boldsymbol{\beta}$.

ii. $\mathrm{Var}\left(\widehat{\boldsymbol{\beta}}\right) = \sigma^2\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}$.

iii. $\mathrm{Cov}\left(\widehat{\boldsymbol{\beta}}, \mathbf{Y}\right) = \sigma^2\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}$.

Απόδειξη. Χρησιμοποιώντας τις ιδιότητες της μέσης τιμής, της διασποράς και της συνδιακύμανσης τυχαίων διανυσμάτων, υπολογίζουμε τα εξής:

i. $E\left(\widehat{\boldsymbol{\beta}}\right) = E\left[\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}\right] \overset{\text{Πρόταση 1.9}}{=} \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}E(\mathbf{Y}) \Rightarrow$

$E\left(\widehat{\boldsymbol{\beta}}\right) \overset{\text{Πρόταση 2.1}}{=} \cancel{\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}}\cancel{\mathbf{X}^{\mathrm{T}}\mathbf{X}}\boldsymbol{\beta} = \boldsymbol{\beta}$.

ii. $\mathrm{Var}\left(\widehat{\boldsymbol{\beta}}\right) = \mathrm{Var}\left[\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}\right] \Rightarrow$

$\mathrm{Var}\left(\widehat{\boldsymbol{\beta}}\right) \overset{\text{Πρόταση 1.9}}{=} \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathrm{Var}(\mathbf{Y})\left[\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\right]^{\mathrm{T}} \Rightarrow$

$\mathrm{Var}\left(\widehat{\boldsymbol{\beta}}\right) \overset{\text{Πρόταση 2.1}}{=} \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}} \cdot \sigma^2\mathbf{I}_n \cdot \mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1} \Rightarrow$

$\mathrm{Var}\left(\widehat{\boldsymbol{\beta}}\right) = \sigma^2\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\cancel{\mathbf{X}^{\mathrm{T}}\mathbf{X}}\cancel{\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}} = \sigma^2\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}$.

iii. $\mathrm{Cov}\left(\widehat{\boldsymbol{\beta}}, \mathbf{Y}\right) = \mathrm{Cov}\left[\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}, \mathbf{Y}\right] \overset{\text{Πρόταση 2.3}}{=} \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathrm{Cov}\left(\mathbf{Y}, \mathbf{Y}\right) \Rightarrow$

$\mathrm{Cov}\left(\widehat{\boldsymbol{\beta}}, \mathbf{Y}\right) = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathrm{Var}\left(\mathbf{Y}\right) = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}} \cdot \sigma^2\mathbf{I}_n = \sigma^2\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}$. □

Λήμμα 2.4. Έστω πίνακας $\mathbf{A} \in \mathbb{R}^{n\times d}$. Αν $\mathbf{A}\mathbf{x} = \mathbf{0}_n$, $\forall \mathbf{x} \in \mathbb{R}^d$, τότε $\mathbf{A} = \mathbf{0}_{n\times d}$.

Θεώρημα 2.1. (Gauss - Markov) Η εκτιμήτρια ελαχίστων τετραγώνων $\widehat{\boldsymbol{\beta}}$ έχει την "ελάχιστη" διασπορά ανάμεσα σε όλες τις αμερόληπτες εκτιμήτριες $\widetilde{\boldsymbol{\beta}}$ του διανύσματος παραμέτρων $\boldsymbol{\beta}$ οι οποίες είναι γραμμικοί συνδυασμοί του $\mathbf{Y}$.

Σημείωση 2.1. Ο πίνακας συνδιακύμανσης $\mathrm{Var}(\widehat{\boldsymbol{\beta}})$ είναι "ελάχιστος" ανάμεσα στους πίνακες συνδιακύμανσης όλων των αμερόληπτων εκτιμητριών $\widetilde{\boldsymbol{\beta}}$ του $\boldsymbol{\beta}$ που είναι γραμμικοί συνδυασμοί του $\mathbf{Y}$ αν και μόνο αν ο πίνακας $\mathrm{Var}(\widetilde{\boldsymbol{\beta}}) - \mathrm{Var}(\widehat{\boldsymbol{\beta}})$ είναι θετικά ημιορισμένος για κάθε τέτοια εκτιμήτρια $\widetilde{\boldsymbol{\beta}}$.

Απόδειξη. Έστω πίνακας $\mathbf{A} \in \mathbb{R}^{(p+1)\times n}$ και $\widetilde{\boldsymbol{\beta}} = \mathbf{AY}$ μία εκτιμήτρια του $\boldsymbol{\beta}$ που είναι γραμμικός συνδυασμός του $\mathbf{Y}$. Χωρίς βλάβη της γενικότητας, υποθέτουμε ότι $\mathbf{A} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}} + \mathbf{C}$, όπου $\mathbf{C} \in \mathbb{R}^{(p+1)\times n}$. Αρχικά, απαιτούμε η εκτιμήτρια $\widetilde{\boldsymbol{\beta}}$ να είναι αμερόληπτη για το $\boldsymbol{\beta}$, δηλαδή:

$$\begin{aligned} E\left(\widetilde{\boldsymbol{\beta}}\right) &= E\left(\mathbf{AY}\right) = E\left[\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}\right] + E\left(\mathbf{CY}\right) = E\left(\widehat{\boldsymbol{\beta}}\right) + \mathbf{C}E(\mathbf{Y}) \\ &= \boldsymbol{\beta} + \mathbf{CX}\boldsymbol{\beta} = \boldsymbol{\beta} \Rightarrow \mathbf{CX}\boldsymbol{\beta} = \mathbf{0}_{p+1},\ \forall \boldsymbol{\beta} \in \mathbb{R}^{p+1} \overset{\text{Λήμμα } 2.4}{\Rightarrow} \mathbf{CX} = \mathbf{0}_{(p+1)\times(p+1)}. \end{aligned}$$

Στη συνέχεια, υπολογίζουμε τον πίνακα συνδιακύμανσης του $\widetilde{\boldsymbol{\beta}}$:

$$\begin{aligned} \mathrm{Var}\left(\widetilde{\boldsymbol{\beta}}\right) &= \mathrm{Var}\left(\mathbf{AY}\right) = \mathbf{A}\mathrm{Var}(\mathbf{Y})\mathbf{A}^{\mathrm{T}} \\ &= \left[\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}} + \mathbf{C}\right]\sigma^2\mathbf{I}_n\left[\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}} + \mathbf{C}\right]^{\mathrm{T}} \\ &= \sigma^2\left[\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}} + \mathbf{C}\right]\left[\mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1} + \mathbf{C}^{\mathrm{T}}\right] \\ &= \sigma^2\left[\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1} + \mathbf{CC}^{\mathrm{T}} + \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{C}^{\mathrm{T}} + \cancel{\mathbf{CX}}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\right] \\ &= \sigma^2\left[\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1} + \mathbf{CC}^{\mathrm{T}} + \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\cancel{(\mathbf{CX})^{\mathrm{T}}}\right] \\ &= \sigma^2\left[\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1} + \mathbf{CC}^{\mathrm{T}}\right] = \mathrm{Var}\left(\widehat{\boldsymbol{\beta}}\right) + \sigma^2\mathbf{CC}^{\mathrm{T}}. \end{aligned}$$

Άρα, $\mathrm{Var}\left(\widetilde{\boldsymbol{\beta}}\right) - \mathrm{Var}\left(\widehat{\boldsymbol{\beta}}\right) = \sigma^2\mathbf{CC}^{\mathrm{T}}$. Όμως, σύμφωνα με το λήμμα 2.3, ο πίνακας $\mathbf{CC}^{\mathrm{T}}$ είναι θετικά ημιορισμένος. Αφού $\sigma^2 > 0$, συμπεραίνουμε ότι και ο πίνακας $\mathrm{Var}\left(\widetilde{\boldsymbol{\beta}}\right) - \mathrm{Var}\left(\widehat{\boldsymbol{\beta}}\right)$ είναι θετικά ημιορισμένος. Επομένως, έχουμε το ζητούμενο. □

Ορισμός 2.2. (Προσαρμοσμένες Τιμές και Κατάλοιπα)

- Οι τιμές $\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_{1,i} + \widehat{\beta}_1 X_{2,i} + \cdots + \widehat{\beta}_p X_{p,i}$ καλούνται **προσαρμοσμένες τιμές** των Y_i. Σε πινακική μορφή, γράφουμε $\widehat{\mathbf{Y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} \in \mathbb{R}^n$.
- Η σχέση $\widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 X_1 + \widehat{\beta}_2 X_2 + \cdots + \widehat{\beta}_p X_p$ ορίζει την **εκτιμημένη εξίσωση παλινδρόμησης**.
- Οι τιμές $\widehat{\varepsilon}_i = Y_i - \widehat{Y}_i = Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_{1,i} - \widehat{\beta}_1 X_{2,i} - \cdots - \widehat{\beta}_p X_{p,i}$ καλούνται **κα-**

τάλοιπα ή **εκτιμημένα σφάλματα** της παλινδρόμησης. Σε πινακική μορφή, γράφουμε $\widehat{\boldsymbol{\varepsilon}} = \mathbf{Y} - \widehat{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}} \in \mathbb{R}^n$.

Πρόταση 2.5. (Ιδιότητες Προσαρμοσμένων Τιμών και Καταλοίπων)

i. $E(\widehat{\mathbf{Y}}) = E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$. Συμπεραίνουμε ότι $E(\widehat{Y}_i) = E(Y_i)$, δηλαδή το $\widehat{Y}_i$ είναι μία αμερόληπτη εκτιμήτρια της $E(Y_i)$. Επίσης, συμπεραίνουμε ότι $E\left(\widehat{\boldsymbol{\varepsilon}}\right) = \mathbf{0}_n$.

ii. $\text{Var}\left(\widehat{\mathbf{Y}}\right) = \text{Cov}\left(\widehat{\mathbf{Y}}, \mathbf{Y}\right) = \text{Cov}\left(\mathbf{Y}, \widehat{\mathbf{Y}}\right) = \sigma^2\mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}$.

iii. $\text{Cov}\left(\widehat{\mathbf{Y}}, \widehat{\boldsymbol{\varepsilon}}\right) = \text{Cov}\left(\widehat{\boldsymbol{\varepsilon}}, \widehat{\mathbf{Y}}\right) = \mathbf{0}_{n\times n}$.

iv. $\text{Cov}\left(\widehat{\boldsymbol{\varepsilon}}, \mathbf{Y}\right) = \text{Cov}\left(\mathbf{Y}, \widehat{\boldsymbol{\varepsilon}}\right) = \text{Var}\left(\widehat{\boldsymbol{\varepsilon}}\right) = E\left(\widehat{\boldsymbol{\varepsilon}}\widehat{\boldsymbol{\varepsilon}}^{\mathrm{T}}\right) = \sigma^2\left[\mathbf{I}_n - \mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\right]$.

v. $\mathbf{X}^{\mathrm{T}}\widehat{\boldsymbol{\varepsilon}} = \mathbf{0}_{p+1}$. Συμπεραίνουμε ότι $\widehat{\mathbf{Y}}^{\mathrm{T}}\widehat{\boldsymbol{\varepsilon}} = \mathbf{1}_n^{\mathrm{T}}\widehat{\boldsymbol{\varepsilon}} = 0$, δηλαδή $\sum\limits_{i=1}^{n}\widehat{Y}_i\widehat{\varepsilon}_i = \sum\limits_{i=1}^{n}\widehat{\varepsilon}_i = 0$.

Απόδειξη. Χρησιμοποιώντας τις ιδιότητες της μέσης τιμής, της διασποράς και της συνδιακύμανσης τυχαίων διανυσμάτων, υπολογίζουμε τα εξής:

i. $E\left(\widehat{\mathbf{Y}}\right) = E\left(\mathbf{X}\widehat{\boldsymbol{\beta}}\right) \overset{\text{Πρόταση 1.9}}{=} \mathbf{X}E\left(\widehat{\boldsymbol{\beta}}\right) \overset{\text{Πρόταση 2.4}}{=} \mathbf{X}\boldsymbol{\beta} = E(\mathbf{Y})$.

$E\left(\widehat{\boldsymbol{\varepsilon}}\right) = E\left(\mathbf{Y} - \widehat{\mathbf{Y}}\right) \overset{\text{Πρόταση 1.9}}{=} E\left(\mathbf{Y}\right) - E\left(\widehat{\mathbf{Y}}\right) = \mathbf{0}_n$.

ii. $\text{Var}\left(\widehat{\mathbf{Y}}\right) = \text{Var}\left(\mathbf{X}\widehat{\boldsymbol{\beta}}\right) \overset{\text{Πρόταση 1.9}}{=} \mathbf{X}\text{Var}\left(\widehat{\boldsymbol{\beta}}\right)\mathbf{X}^{\mathrm{T}} \overset{\text{Πρόταση 2.4}}{=} \sigma^2\mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}} \Rightarrow$

$\text{Cov}\left(\widehat{\mathbf{Y}}, \mathbf{Y}\right) = \text{Cov}\left(\mathbf{X}\widehat{\boldsymbol{\beta}}, \mathbf{Y}\right) \overset{\text{Πρόταση 2.3}}{=} \mathbf{X}\text{Cov}\left(\widehat{\boldsymbol{\beta}}, \mathbf{Y}\right) \Rightarrow$

$\text{Cov}\left(\widehat{\mathbf{Y}}, \mathbf{Y}\right) \overset{\text{Πρόταση 2.4}}{=} \sigma^2\mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}} = \text{Var}\left(\widehat{\mathbf{Y}}\right) \Rightarrow$

$\text{Cov}\left(\mathbf{Y}, \widehat{\mathbf{Y}}\right) \overset{\text{Πρόταση 2.3}}{=} \left[\text{Cov}\left(\widehat{\mathbf{Y}}, \mathbf{Y}\right)\right]^{\mathrm{T}} = \sigma^2\mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}} = \text{Cov}\left(\widehat{\mathbf{Y}}, \mathbf{Y}\right)$.

iii. $\text{Cov}\left(\widehat{\mathbf{Y}}, \widehat{\boldsymbol{\varepsilon}}\right) = \text{Cov}\left(\widehat{\mathbf{Y}}, \mathbf{Y} - \widehat{\mathbf{Y}}\right) \overset{\text{Πρόταση 2.3}}{=} \text{Cov}\left(\widehat{\mathbf{Y}}, \mathbf{Y}\right) - \text{Cov}\left(\widehat{\mathbf{Y}}, \widehat{\mathbf{Y}}\right) \Rightarrow$

$\text{Cov}\left(\widehat{\mathbf{Y}}, \widehat{\boldsymbol{\varepsilon}}\right) \overset{\text{Πρόταση 2.3}}{=} \cancel{\text{Cov}\left(\widehat{\mathbf{Y}}, \mathbf{Y}\right)} - \cancel{\text{Var}\left(\widehat{\mathbf{Y}}\right)} = \mathbf{0}_{n\times n} \Rightarrow$

$\text{Cov}\left(\widehat{\boldsymbol{\varepsilon}}, \widehat{\mathbf{Y}}\right) \overset{\text{Πρόταση 2.3}}{=} \left[\text{Cov}\left(\widehat{\mathbf{Y}}, \widehat{\boldsymbol{\varepsilon}}\right)\right]^{\mathrm{T}} = \mathbf{0}_{n\times n} = \text{Cov}\left(\widehat{\mathbf{Y}}, \widehat{\boldsymbol{\varepsilon}}\right)$.

iv. $\text{Cov}\left(\widehat{\boldsymbol{\varepsilon}}, \mathbf{Y}\right) = \text{Cov}\left(\mathbf{Y} - \widehat{\mathbf{Y}}, \mathbf{Y}\right) \overset{\text{Πρόταση 2.3}}{=} \text{Cov}\left(\mathbf{Y}, \mathbf{Y}\right) - \text{Cov}\left(\widehat{\mathbf{Y}}, \mathbf{Y}\right) \Rightarrow$

$\text{Cov}\left(\widehat{\boldsymbol{\varepsilon}}, \mathbf{Y}\right) \overset{\text{Πρόταση 2.3}}{=} \text{Var}\left(\mathbf{Y}\right) - \text{Cov}\left(\widehat{\mathbf{Y}}, \mathbf{Y}\right) \overset{\text{Πρόταση 2.4}}{=} \sigma^2\left[\mathbf{I}_n - \mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\right] \Rightarrow$

$\text{Cov}\left(\mathbf{Y}, \widehat{\boldsymbol{\varepsilon}}\right) \overset{\text{Πρόταση 2.3}}{=} \left[\text{Cov}\left(\widehat{\boldsymbol{\varepsilon}}, \mathbf{Y}\right)\right]^{\mathrm{T}} = \sigma^2\left[\mathbf{I}_n - \mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\right] = \text{Cov}\left(\widehat{\boldsymbol{\varepsilon}}, \mathbf{Y}\right) \Rightarrow$

$\text{Var}\left(\widehat{\boldsymbol{\varepsilon}}\right) = \text{Var}\left(\mathbf{Y} - \widehat{\mathbf{Y}}\right) \overset{\text{Πρόταση 2.3}}{\Rightarrow}$

$\text{Var}\left(\widehat{\boldsymbol{\varepsilon}}\right) = \text{Var}\left(\mathbf{Y}\right) + \cancel{\text{Var}\left(\widehat{\mathbf{Y}}\right)} - \cancel{\text{Cov}\left(\mathbf{Y}, \widehat{\mathbf{Y}}\right)} - \text{Cov}\left(\widehat{\mathbf{Y}}, \mathbf{Y}\right) \Rightarrow$

$\text{Var}\left(\widehat{\boldsymbol{\varepsilon}}\right) = \sigma^2\left[\mathbf{I}_n - \mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\right] \overset{\text{Ορισμός 1.4}}{\Rightarrow}$

$E\left(\widehat{\boldsymbol{\varepsilon}}\widehat{\boldsymbol{\varepsilon}}^{\mathrm{T}}\right) = \mathrm{Var}\left(\widehat{\boldsymbol{\varepsilon}}\right) + \cancel{E\left(\widehat{\boldsymbol{\varepsilon}}\right)\left[E\left(\widehat{\boldsymbol{\varepsilon}}\right)\right]^{\mathrm{T}}} = \sigma^2\left[\mathbf{I}_n - \mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\right].$

v. $\mathbf{X}^{\mathrm{T}}\widehat{\boldsymbol{\varepsilon}} = \mathbf{X}^{\mathrm{T}}\left(\mathbf{Y} - \widehat{\mathbf{Y}}\right) = \mathbf{X}^{\mathrm{T}}\left(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right) = \mathbf{X}^{\mathrm{T}}\mathbf{Y} - \cancel{\mathbf{X}^{\mathrm{T}}\mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}}\mathbf{X}^{\mathrm{T}}\mathbf{Y} = \mathbf{0}_{p+1} \Rightarrow$

$\widehat{\mathbf{Y}}^{\mathrm{T}}\widehat{\boldsymbol{\varepsilon}} = \left(\mathbf{X}\widehat{\boldsymbol{\beta}}\right)^{\mathrm{T}}\widehat{\boldsymbol{\varepsilon}} = \widehat{\boldsymbol{\beta}}^{\mathrm{T}} \cdot \mathbf{X}^{\mathrm{T}}\widehat{\boldsymbol{\varepsilon}} = \widehat{\boldsymbol{\beta}}^{\mathrm{T}} \cdot \mathbf{0}_{p+1} = 0.$

Το διάνυσμα $\mathbf{1}_n$ είναι η πρώτη στήλη του πίνακα σχεδιασμού $\mathbf{X}$ και δείξαμε ότι $\mathbf{X}^{\mathrm{T}}\widehat{\boldsymbol{\varepsilon}} = \mathbf{0}_{p+1}$, οπότε θα ισχύει ότι $\mathbf{1}_n^{\mathrm{T}}\widehat{\boldsymbol{\varepsilon}} = 0$. □

2.3 Εκτίμηση της Διασποράς

Όμοια με την περίπτωση του απλού γραμμικού μοντέλου, η αμερόληπτη εκτιμήτρια της διασποράς σ^2 προκύπτει, αν αντικαταστήσουμε τους $p+1$ άγνωστους συντελεστές παλινδρόμησης $\beta_0, \beta_1, \ldots, \beta_p$ από τις αμερόληπτες εκτιμήτριες $\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_p$ και αφαιρέσουμε $p+1$ βαθμούς ελευθερίας από τον παρονομαστή για τις $p+1$ παραμέτρους που εκτιμήσαμε. Η αμερόληπτη εκτιμήτρια του σ^2 που προκύπτει είναι το λεγόμενο **μέσο τετραγωνικό σφάλμα** (mean squared error):

$$\mathrm{MSE} = S^2 = \frac{\|\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\|^2}{n-p-1} = \frac{\|\mathbf{Y} - \widehat{\mathbf{Y}}\|^2}{n-p-1} = \frac{\|\widehat{\boldsymbol{\varepsilon}}\|^2}{n-p-1} = \frac{\widehat{\boldsymbol{\varepsilon}}^{\mathrm{T}}\widehat{\boldsymbol{\varepsilon}}}{n-p-1} = \frac{1}{n-p-1}\sum_{i=1}^{n}\widehat{\varepsilon}_i^2.$$

Λήμμα 2.5. (Ίχνος Πίνακα)

i. Έστω $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n\times d}$. Τότε, $\mathrm{tr}(\mathbf{A}+\mathbf{B}) = \mathrm{tr}(\mathbf{A}) + \mathrm{tr}(\mathbf{B})$.

ii. Έστω $\mathbf{A} \in \mathbb{R}^{n\times d}$ και $\mathbf{B} \in \mathbb{R}^{d\times n}$. Τότε, $\mathrm{tr}(\mathbf{AB}) = \mathrm{tr}(\mathbf{BA})$.

iii. Έστω $\mathbf{A} \in \mathbb{R}^{1\times 1}$. Τότε, $\mathrm{tr}(\mathbf{A}) = \mathbf{A} = \mathbf{A}^{\mathrm{T}}$.

iv. Έστω $\mathbf{X} \in \mathbb{R}^{n\times d}$ τυχαίος πίνακας. Τότε, $E\left[\mathrm{tr}(\mathbf{X})\right] = \mathrm{tr}\left[E(\mathbf{X})\right]$.

Πρόταση 2.6. Το μέσο τετραγωνικό σφάλμα MSE είναι μία αμερόληπτη εκτιμήτρια της διασποράς σ^2 στο πολλαπλό γραμμικό μοντέλο.

Απόδειξη. Παρατηρούμε ότι $\widehat{\boldsymbol{\varepsilon}}^{\mathrm{T}}\widehat{\boldsymbol{\varepsilon}} \in \mathbb{R}^{1\times 1}$, οπότε $\widehat{\boldsymbol{\varepsilon}}^{\mathrm{T}}\widehat{\boldsymbol{\varepsilon}} = \mathrm{tr}\left(\widehat{\boldsymbol{\varepsilon}}^{\mathrm{T}}\widehat{\boldsymbol{\varepsilon}}\right) = \mathrm{tr}\left(\widehat{\boldsymbol{\varepsilon}}\widehat{\boldsymbol{\varepsilon}}^{\mathrm{T}}\right)$, σύμφωνα με το προηγούμενο λήμμα. Επομένως,

$$E\left(\widehat{\boldsymbol{\varepsilon}}^{\mathrm{T}}\widehat{\boldsymbol{\varepsilon}}\right) = E\left[\mathrm{tr}\left(\widehat{\boldsymbol{\varepsilon}}\widehat{\boldsymbol{\varepsilon}}^{\mathrm{T}}\right)\right] \overset{\text{Λήμμα 2.5}}{=} \mathrm{tr}\left[E\left(\widehat{\boldsymbol{\varepsilon}}\widehat{\boldsymbol{\varepsilon}}^{\mathrm{T}}\right)\right] \overset{\text{Πρόταση 2.5}}{\Rightarrow}$$

$$E\left(\widehat{\boldsymbol{\varepsilon}}^{\mathrm{T}}\widehat{\boldsymbol{\varepsilon}}\right) = \sigma^2\mathrm{tr}\left[\mathbf{I}_n - \mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\right] = \sigma^2\mathrm{tr}(\mathbf{I}_n) - \sigma^2\mathrm{tr}\left[\mathbf{X}\cdot\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\right] \overset{\text{Λήμμα 2.5}}{\Rightarrow}$$

$$E\left(\widehat{\boldsymbol{\varepsilon}}^{\mathrm{T}}\widehat{\boldsymbol{\varepsilon}}\right) = n\sigma^2 - \sigma^2\mathrm{tr}\left[\cancel{\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\cdot\mathbf{X}}\right] = \left[n - \mathrm{tr}\left(\mathbf{I}_{p+1}\right)\right]\sigma^2 = \left[n-(p+1)\right]\sigma^2 \Rightarrow$$

$$E\left(S^2\right) = E\left(\frac{\widehat{\boldsymbol{\varepsilon}}^{\mathrm{T}}\widehat{\boldsymbol{\varepsilon}}}{n-p-1}\right) = \sigma^2. \quad \square$$

2.4 Συντελεστής Προσδιορισμού

Ορισμός 2.3. (Αθροίσματα Τετραγώνων)

- Ορίζουμε $\text{SST} = \|\mathbf{Y}-\overline{Y}\mathbf{1}_n\|^2 = \sum_{i=1}^{n}(Y_i-\overline{Y})^2$ το συνολικό άθροισμα τετραγώνων (total sum of squares) των δεδομένων.
- Ορίζουμε $\text{SSR} = \|\widehat{\mathbf{Y}} - \overline{Y}\mathbf{1}_n\|^2 = \sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2$ το άθροισμα τετραγώνων που οφείλεται στην παλινδρόμηση (sum of squares due to regression).
- Ορίζουμε $\text{SSE} = \|\mathbf{Y} - \widehat{\mathbf{Y}}\|^2 = \|\widehat{\boldsymbol{\varepsilon}}\|^2 = \sum_{i=1}^{n}\widehat{\varepsilon}_i^2$ το άθροισμα τετραγώνων των καταλοίπων (sum of squared errors).

Παρατήρηση 2.1. Παρατηρούμε ότι:

$$S^2 = \text{MSE} = \frac{\text{SSE}}{n-p-1}.$$

Πρόταση 2.7. (Ανάλυση Διασποράς)

$$\|\mathbf{Y} - \overline{Y}\mathbf{1}_n\|^2 = \|\widehat{\mathbf{Y}} - \overline{Y}\mathbf{1}_n\|^2 + \|\mathbf{Y} - \widehat{\mathbf{Y}}\|^2, \text{ δηλαδή } \text{SST} = \text{SSR} + \text{SSE}.$$

Απόδειξη. Έχοντας στο μυαλό μας μία τεχνική που εμφανίζεται και στον υπολογισμό του μέσου τετραγωνικού σφάλματος μίας εκτιμήτριας, σκεφτόμαστε να προσθαφαιρέσουμε το $\widehat{\mathbf{Y}}$ στην ποσότητα $\mathbf{Y} - \overline{Y}\mathbf{1}_n$ που εμφανίζεται στο SST:

$$\begin{aligned}
\text{SST} &= \|\mathbf{Y} - \overline{Y}\mathbf{1}_n\|^2 = \|\mathbf{Y} - \widehat{\mathbf{Y}} + \widehat{\mathbf{Y}} - \overline{Y}\mathbf{1}_n\|^2 \\
&= \left(\mathbf{Y} - \widehat{\mathbf{Y}} + \widehat{\mathbf{Y}} - \overline{Y}\mathbf{1}_n\right)^{\mathrm{T}}\left(\mathbf{Y} - \widehat{\mathbf{Y}} + \widehat{\mathbf{Y}} - \overline{Y}\mathbf{1}_n\right) \\
&= \|\mathbf{Y} - \widehat{\mathbf{Y}}\|^2 + \|\widehat{\mathbf{Y}} - \overline{Y}\mathbf{1}_n\|^2 + \left(\mathbf{Y} - \widehat{\mathbf{Y}}\right)^{\mathrm{T}}\left(\widehat{\mathbf{Y}} - \overline{Y}\mathbf{1}_n\right) + \left(\widehat{\mathbf{Y}} - \overline{Y}\mathbf{1}_n\right)^{\mathrm{T}}\left(\mathbf{Y} - \widehat{\mathbf{Y}}\right) \\
&= \text{SSE} + \text{SSR} + 2\left(\widehat{\mathbf{Y}} - \overline{Y}\mathbf{1}_n\right)^{\mathrm{T}}\left(\mathbf{Y} - \widehat{\mathbf{Y}}\right) = \text{SSE} + \text{SSR} + 2\cancelto{0}{\widehat{\mathbf{Y}}^{\mathrm{T}}\widehat{\boldsymbol{\varepsilon}}} - 2\overline{Y}\cancelto{0}{\mathbf{1}_n^{\mathrm{T}}\widehat{\boldsymbol{\varepsilon}}} \\
&= \text{SSE} + \text{SSR}. \quad \square
\end{aligned}$$

Ορισμός 2.4. (Συντελεστής Προσδιορισμού)

$$R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}.$$

Ορισμός 2.5. (Προσαρμοσμένος Συντελεστής Προσδιορισμού)

$$R^2_{\text{adj}} = \overline{R}^2 = 1 - \frac{n-1}{n-p-1}\cdot\frac{\text{SSE}}{\text{SST}} = 1 - \frac{n-1}{n-p-1}\left(1-R^2\right).$$

Πρόταση 2.8. (Ιδιότητες Προσαρμοσμένου Συντελεστή Προσδιορισμού)

i. $R^2_{\text{adj}} < R^2$.

ii. $-\frac{p}{n-p-1} \leqslant R^2_{\text{adj}} \leqslant 1$. Συμπεραίνουμε ότι ο προσαρμοσμένος συντελεστής προσδιορισμού μπορεί να πάρει και αρνητικές τιμές.

Απόδειξη. Υπολογίζουμε τα εξής:

i. $p > 0 \Rightarrow -p < 0 \Rightarrow n-p-1 < n-1 \Rightarrow \frac{n-1}{n-p-1} > 1 \Rightarrow \frac{n-1}{n-p-1} \cdot \frac{\text{SSE}}{\text{SST}} > \frac{\text{SSE}}{\text{SST}} \Rightarrow$
$1 - \frac{n-1}{n-p-1} \cdot \frac{\text{SSE}}{\text{SST}} < 1 - \frac{\text{SSE}}{\text{SST}} \Rightarrow R^2_{\text{adj}} < R^2$.

ii. $0 \leqslant R^2 \leqslant 1 \Rightarrow 0 \leqslant 1 - R^2 \leqslant 1 \Rightarrow 0 \leqslant \frac{\text{SSE}}{\text{SST}} \leqslant 1 \Rightarrow 0 \leqslant \frac{n-1}{n-p-1} \cdot \frac{\text{SSE}}{\text{SST}} \leqslant \frac{n-1}{n-p-1} \Rightarrow$
$1 - \frac{n-1}{n-p-1} \leqslant 1 - \frac{n-1}{n-p-1} \cdot \frac{\text{SSE}}{\text{SST}} \leqslant 1 \Rightarrow -\frac{p}{n-p-1} \leqslant R^2_{\text{adj}} \leqslant 1$.

Ερμηνεία: Η χρησιμότητα του προσαρμοσμένου συντελεστή προσδιορισμού έγκειται στο γεγονός ότι συνυπολογίζει και το πλήθος p των επεξηγηματικών μεταβλητών που χρησιμοποιεί το γραμμικό μοντέλο.

Έστω ότι έχουμε ένα σύνολο από k υποψήφιες επεξηγηματικές μεταβλητές για μία μεταβλητή ενδιαφέροντος Y και θέλουμε να επιλέξουμε μόνο p από αυτές για να κατασκευάσουμε ένα γραμμικό μοντέλο που να εξηγεί όσο το δυνατόν μεγαλύτερο ποσοστό της μεταβλητότητας των δεδομένων. Ξεκινώντας από ένα απλό γραμμικό μοντέλο που κάνει χρήση μόνο μίας εκ των k υποψήφιων επεξηγηματικών μεταβλητών και προσθέτοντας σταδιακά τις υπόλοιπες επεξηγηματικές μεταβλητές, θα παρατηρούσαμε ότι ο συντελεστής R^2 θα αυξανόταν συνεχώς. Συνεπώς, θα έπαιρνε τη μέγιστη δυνατή τιμή του όταν θα είχαμε προσθέσει όλες τις k υποψήφιες επεξηγηματικές μεταβλητές στο γραμμικό μοντέλο.

Αντιθέτως, βλέπουμε ότι ο προσαρμοσμένος συντελεστής προσδιορισμού είναι φθίνουσα συνάρτηση του p. Επομένως, ακολουθώντας την ίδια διαδικασία, θα καταλήγαμε σε ένα γραμμικό που εξηγεί ένα μεγάλο ποσοστό από τη μεταβλητότητα των δεδομένων με όσο το δυνατόν μικρότερο πλήθος επεξηγηματικών μεταβλητών.

2.5 Πρόβλεψη Καινούργιας Παρατήρησης

Έχοντας μία νέα παρατήρηση $\mathbf{X}_{n+1} = (1, X_{1,n+1}, X_{2,n+1}, \ldots, X_{p,n+1})^{\mathrm{T}} \in \mathbb{R}^{p+1}$, ενδιαφερόμαστε να κάνουμε μία πρόβλεψη για την τιμή:

$$Y_{n+1} = \beta_0 + \beta_1 X_{1,n+1} + \beta_2 X_{2,n+1} + \cdots + \beta_p X_{p,n+1} + \varepsilon_{n+1} = \mathbf{X}^{\mathrm{T}}_{n+1}\boldsymbol{\beta} + \varepsilon_{n+1},$$

την οποία δεν έχουμε παρατηρήσει, όπου $E(\varepsilon_{n+1}) = 0$, $\text{Var}(\varepsilon_{n+1}) = \sigma^2$ και ε_{n+1} ασυσχέτιστο με τα $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$. Με βάση τις προηγούμενες n παρατηρή-

σεις $Y_1, Y_2, \ldots, Y_n$, έχουμε υπολογίσει την εκτιμήτρια ελαχίστων τετραγώνων $\widehat{\boldsymbol{\beta}}$, οπότε ορίζουμε $\widetilde{Y}_{n+1} = \mathbf{X}_{n+1}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}$. Παρατηρούμε ότι $E(\widetilde{Y}_{n+1}) = E(Y_{n+1}) = \mathbf{X}_{n+1}^{\mathrm{T}}\boldsymbol{\beta}$, οπότε μπορούμε να χρησιμοποιήσουμε την τιμή $\widetilde{Y}_{n+1}$ για να προβλέψουμε την Y_{n+1}. Επιπλέον, ορίζουμε το **σφάλμα πρόβλεψης**:

$$\widetilde{\varepsilon}_{n+1} = Y_{n+1} - \widetilde{Y}_{n+1} = \mathbf{X}_{n+1}^{\mathrm{T}}\left(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\right) + \varepsilon_{n+1}.$$

Πρόταση 2.9. (Σφάλμα Πρόβλεψης)

i. $E\left(\widetilde{\varepsilon}_{n+1}\right) = 0.$

ii. $\mathrm{Var}\left(\widetilde{Y}_{n+1}\right) = \sigma^2\mathbf{X}_{n+1}^{\mathrm{T}}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}_{n+1}.$

iii. $\mathrm{Var}\left(\widetilde{\varepsilon}_{n+1}\right) = \sigma^2\left[1 + \mathbf{X}_{n+1}^{\mathrm{T}}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}_{n+1}\right].$

Απόδειξη. Χρησιμοποιώντας τις ιδιότητες της μέσης τιμής, της διασποράς και της συνδιακύμανσης τυχαίων διανυσμάτων, υπολογίζουμε τα εξής:

i. $E\left(\widetilde{\varepsilon}_{n+1}\right) = E\left(Y_{n+1} - \widehat{Y}_{n+1}\right) = E\left(Y_{n+1}\right) - E\left(\widetilde{Y}_{n+1}\right) = 0.$

ii. $\mathrm{Var}\left(\widetilde{Y}_{n+1}\right) = \mathrm{Var}\left(\mathbf{X}_{n+1}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}\right) \overset{\text{Πρόταση 1.9}}{=} \mathbf{X}_{n+1}^{\mathrm{T}}\mathrm{Var}\left(\widehat{\boldsymbol{\beta}}\right)\mathbf{X}_{n+1} \Rightarrow$

$\mathrm{Var}\left(\widetilde{Y}_{n+1}\right) \overset{\text{Πρόταση 2.4}}{=} \sigma^2\mathbf{X}_{n+1}^{\mathrm{T}}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}_{n+1}.$

iii. $\mathrm{Var}(Y_{n+1}) = \mathrm{Var}\left(\mathbf{X}_{n+1}^{\mathrm{T}}\boldsymbol{\beta} + \varepsilon_{n+1}\right) = \mathrm{Var}(\varepsilon_{n+1}) = \sigma^2$ και

$\mathrm{Cov}\left(Y_{n+1}, \widetilde{Y}_{n+1}\right) = \mathrm{Cov}\left(Y_{n+1}, \mathbf{X}_{n+1}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}\right) = 0$, αφού το $\widehat{\boldsymbol{\beta}}$ είναι γραμμικός συνδυασμός των $Y_1, Y_2, \ldots, Y_n$, τα οποία είναι όλα ασυσχέτιστα με το Y_{n+1}. Άρα,

$$\mathrm{Var}\left(\widetilde{\varepsilon}_{n+1}\right) = \mathrm{Var}\left(Y_{n+1}\right) + \mathrm{Var}\left(\widetilde{Y}_{n+1}\right) - 2\,\mathrm{Cov}\left(Y_{n+1}, \widetilde{Y}_{n+1}\right)^{\to 0} \Rightarrow$$

$$\mathrm{Var}\left(\widetilde{\varepsilon}_{n+1}\right) = \sigma^2 + \sigma^2\mathbf{X}_{n+1}^{\mathrm{T}}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}_{n+1} = \sigma^2\left[1 + \mathbf{X}_{n+1}^{\mathrm{T}}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}_{n+1}\right]. \quad \square$$

2.6 Κανονικό Πολλαπλό Γραμμικό Μοντέλο

Ορισμός 2.6. Το μοντέλο $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, όπου ισχύει $\boldsymbol{\varepsilon} \sim N_n\left(\mathbf{0}_n, \sigma^2\mathbf{I}_n\right)$, καλείται **κανονικό πολλαπλό γραμμικό μοντέλο** ή πολλαπλό γραμμικό μοντέλο με κανονικά σφάλματα.

Παρατήρηση 2.2. (Κανονικό Πολλαπλό Γραμμικό Μοντέλο)

- Τα τυχαία σφάλματα στο κανονικό απλό γραμμικό μοντέλο είναι **κανονικά κατανεμημένα** με **μέση τιμή 0**, **ομοσκεδαστικά** και **ανεξάρτητα**, ή ισοδύναμα ασυσχέτιστα. Δηλαδή, $\varepsilon_i \sim N(0, \sigma^2)$ ανεξάρτητα για $i = 1, 2, \ldots, n$.
- Σύμφωνα με την πρόταση 1.10, ισχύει ότι $\mathbf{Y} \sim N_n\left(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n\right)$.

- Έχοντας, πλέον, κάνει υπόθεση για την κατανομή των τυχαίων σφαλμάτων ε_i, μπορούμε να κάνουμε χρήση άλλων γνωστών μεθόδων εκτίμησης εκτός από τη μέθοδο ελαχίστων τετραγώνων, όπως η μέθοδος μέγιστης πιθανοφάνειας που γνωρίζουμε από τη μαθηματική στατιστική.

Λήμμα 2.6. Έστω $c \in \mathbb{R}$ και $\mathbf{A} \in \mathbb{R}^{n\times n}$. Τότε, $\det(c\mathbf{A}) = c^n \det \mathbf{A}$.

Πρόταση 2.10. Η **εκτιμήτρια μέγιστης πιθανοφάνειας** του $\boldsymbol{\beta}$ ταυτίζεται με την εκτιμήτρια ελαχίστων τετραγώνων $\widehat{\boldsymbol{\beta}}$, ενώ η εκτιμήτρια μέγιστης πιθανόφανειας της διασποράς σ^2 δίνεται από τον τύπο:

$$\widehat{\sigma}^2 = \frac{\text{SSE}}{n} = \frac{(n-p-1)S^2}{n} = \frac{\|\widehat{\boldsymbol{\varepsilon}}\|^2}{n}.$$

Απόδειξη. Σύμφωνα με τον ορισμό της πολυδιάστατης κανονικής κατανομής και το προηγούμενο λήμμα, έχουμε ότι:

$$\begin{aligned} L\left(\boldsymbol{\beta}, \sigma^2|\mathbf{y}\right) &= (2\pi)^{-\frac{n}{2}} \left[\det\left(\sigma^2\mathbf{I}_n\right)\right]^{-\frac{1}{2}} \exp\left\{-\frac{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}\left(\sigma^2\mathbf{I}_n\right)^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}{2}\right\} \\ &= \left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left\{-\frac{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}{2\sigma^2}\right\} = \left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left\{-\frac{Q(\boldsymbol{\beta})}{2\sigma^2}\right\}. \end{aligned}$$

Λογαριθμίζουμε τη συνάρτηση πιθανοφάνειας για να υπολογίσουμε τη συνάρτηση λογαριθμοπιθανοφάνειας:

$$\ell\left(\boldsymbol{\beta}, \sigma^2|\mathbf{y}\right) = \log L\left(\boldsymbol{\beta}, \sigma^2|\mathbf{y}\right) = -\frac{n\log(2\pi)}{2} - \frac{n\log\sigma^2}{2} - \frac{Q(\boldsymbol{\beta})}{2\sigma^2}.$$

Μεγιστοποιούμε πρώτα ως προς το διάνυσμα $\boldsymbol{\beta}$:

$$\frac{\partial\ell}{\partial\boldsymbol{\beta}}\left(\boldsymbol{\beta}, \sigma^2\right) = -\frac{1}{2\sigma^2}\cdot\frac{\partial Q}{\partial\boldsymbol{\beta}}(\boldsymbol{\beta}) = 0 \Rightarrow \frac{\partial Q}{\partial\boldsymbol{\beta}}(\boldsymbol{\beta}) = 0.$$

Η παραπάνω εξίσωση μας οδήγησε στην εκτιμήτρια ελαχίστων τετραγώνων $\widehat{\boldsymbol{\beta}}$ στην πρόταση 2.2. Στη συνέχεια, μεγιστοποιούμε ως προς σ^2:

$$\frac{\partial\ell}{\partial\sigma^2}\left(\widehat{\boldsymbol{\beta}}, \sigma^2\right) = -\frac{n}{2\sigma^2} + \frac{Q(\widehat{\boldsymbol{\beta}})}{2\left(\sigma^2\right)^2} = 0 \Rightarrow$$

$$\widehat{\sigma}^2 = \frac{Q(\widehat{\boldsymbol{\beta}})}{n} = \frac{(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}})^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}})}{n} = \frac{\|\widehat{\boldsymbol{\varepsilon}}\|^2}{n} = \frac{\text{SSE}}{n} = \frac{(n-p-1)S^2}{n}.$$

Υπολογίζουμε τη δεύτερη παράγωγο της συνάρτησης $\ell\left(\widehat{\boldsymbol{\beta}}, \sigma^2\right)$ ως προς σ^2 για να επαληθεύσουμε ότι το $\widehat{\sigma}^2$ είναι σημείο μεγίστου:

$$\frac{\partial^2\ell}{\partial\left(\sigma^2\right)^2}\left(\widehat{\boldsymbol{\beta}}, \sigma^2\right) = \frac{n}{2\left(\sigma^2\right)^2} - \frac{Q(\widehat{\boldsymbol{\beta}})}{\left(\sigma^2\right)^3} \Rightarrow$$

$$\frac{\partial^2 \ell}{\partial \left(\sigma^2\right)^2}\left(\widehat{\boldsymbol{\beta}}, \widehat{\sigma}^2\right) = \frac{1}{\widehat{\sigma}^4}\left[\frac{n}{2} - \frac{Q(\widehat{\boldsymbol{\beta}})}{\widehat{\sigma}^2}\right] = \frac{1}{\widehat{\sigma}^4}\left(\frac{n}{2} - n\right) = -\frac{n}{2\widehat{\sigma}^4} < 0.$$

Υπολογίζοντας τα όρια της συνάρτησης $L\left(\widehat{\boldsymbol{\beta}}, \sigma^2\right)$ ως προς σ^2 στα 0^+ και ∞, επαληθεύουμε ότι το $\widehat{\sigma}^2$ είναι σημείο ολικού μεγίστου:

$$\lim_{\sigma^2 \to \infty} L\left(\widehat{\boldsymbol{\beta}}, \sigma^2\right) = \lim_{\sigma^2 \to \infty} \left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left\{-\frac{Q(\widehat{\boldsymbol{\beta}})}{2\sigma^2}\right\} \overset{n>0}{=} 0,$$

$$\lim_{\sigma^2 \to 0^+} L\left(\widehat{\boldsymbol{\beta}}, \sigma^2\right) \overset{\tau=\sigma^{-2}}{=} \lim_{\tau \to \infty} \left(\frac{\tau}{2\pi}\right)^{\frac{n}{2}} \exp\left\{-\frac{Q(\widehat{\boldsymbol{\beta}})}{2}\tau\right\} \overset{Q(\widehat{\boldsymbol{\beta}})>0}{=} 0. \quad \square$$

Τώρα, που έχουμε υποθέσει ότι τα τυχαία σφάλματα ε_i είναι κανονικά κατανεμημένα, μπορούμε να πάρουμε αποτελέσματα για τις κατανομές των εκτιμητριών $\widehat{\boldsymbol{\beta}}$ και S^2. Οι κατανομές αυτές είναι πολύ σημαντικές, επειδή μας επιτρέπουν, πέρα από τις σημειακές εκτιμήσεις που έχουμε υπολογίσει, να κατασκευάσουμε διαστήματα εμπιστοσύνης και να πραγματοποιήσουμε ελέγχους υποθέσεων.

Θεώρημα 2.2. (Cochran) Έστω $\mathbf{Y} \sim N_n\left(\boldsymbol{\mu}, \sigma^2\mathbf{I}_n\right)$ και $\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_k \in \mathbb{R}^{n\times n}$ συμμετρικοί πίνακες με $\mathbf{A}_1 + \mathbf{A}_2 + \cdots + \mathbf{A}_k = \mathbf{I}_n$ και $\operatorname{rank}(\mathbf{A}_i) = r_i$ για $i = 1, 2, \ldots, k$. Αν $r_1 + r_2 + \cdots + r_k = n$, τότε $\frac{1}{\sigma^2}(\mathbf{Y}-\boldsymbol{\mu})^{\mathrm{T}}\mathbf{A}_i(\mathbf{Y}-\boldsymbol{\mu}) \sim \chi^2_{r_i}$ ανεξάρτητες για $i = 1, 2, \ldots, k$.

Ορισμός 2.7. Ένας πίνακας $\mathbf{A} \in \mathbb{R}^{n\times n}$ καλείται **πίνακας ορθογώνιας προβολής** αν και μόνο αν $\mathbf{A}^2 = \mathbf{A} = \mathbf{A}^{\mathrm{T}}$, δηλαδή είναι συμμετρικός και ταυτοδύναμος.

Λήμμα 2.7. Έστω $\mathbf{A} \in \mathbb{R}^{n\times n}$ ένας πίνακας ορθογώνιας προβολής. Τότε,

i. Ισχύει ότι $\mathbf{A}^{\mathrm{T}}\mathbf{A} = \mathbf{A}\mathbf{A}^{\mathrm{T}} = \mathbf{A}$.

ii. Ο πίνακας $\mathbf{I}_n - \mathbf{A}$ είναι και αυτός πίνακας ορθογώνιας προβολής.

iii. Ισχύει ότι $\operatorname{rank}(\mathbf{A}) = \operatorname{tr}(\mathbf{A})$.

Λήμμα 2.8. Ορίζουμε $\mathbf{P} = \mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}} \in \mathbb{R}^{n\times n}$. Τότε,

i. Ο πίνακας $\mathbf{P}$ είναι πίνακας ορθογώνιας προβολής.

ii. Ισχύει ότι $\mathbf{P}\mathbf{X} = \mathbf{X}$, δηλαδή $(\mathbf{I}_n - \mathbf{P})\,\mathbf{X} = \mathbf{0}_{n\times(p+1)}$. Εφόσον το διάνυσμα $\mathbf{1}_n$ αποτελεί την πρώτη στήλη του πίνακα $\mathbf{X}$, συμπεραίνουμε ότι $\mathbf{P}\mathbf{1}_n = \mathbf{1}_n$.

iii. Ισχύει ότι $\mathbf{P}\mathbf{Y} = \widehat{\mathbf{Y}}$ και $(\mathbf{I}_n - \mathbf{P})\,\mathbf{Y} = \widehat{\boldsymbol{\varepsilon}}$.

Απόδειξη. Βάσει του προηγούμενου ορισμού, υπολογίζουμε ότι:

i. $\mathbf{P}^2 = \mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\cancel{\mathbf{X}^{\mathrm{T}}\cdot\mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}}\mathbf{X}^{\mathrm{T}} = \mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}} = \mathbf{P}$ και

$\mathbf{P}^{\mathrm{T}} = \left[\mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\right]^{\mathrm{T}} = \mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}} = \mathbf{P}$.

ii. $\mathbf{P}\mathbf{X} = \mathbf{X}\cancel{\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\cdot\mathbf{X}} = \mathbf{X} \Rightarrow (\mathbf{I}_n - \mathbf{P})\,\mathbf{X} = \cancel{\mathbf{X}} - \cancel{\mathbf{P}\mathbf{X}} = \mathbf{0}_{n\times(p+1)}$.

iii. $\mathbf{PY} = \mathbf{X} \cdot \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y} \overset{\text{Πρόταση 2.2}}{=} \mathbf{X}\widehat{\boldsymbol{\beta}} \overset{\text{Ορισμός 2.2}}{=} \widehat{\mathbf{Y}}$ και

$(\mathbf{I}_n - \mathbf{P})\,\mathbf{Y} = \mathbf{Y} - \mathbf{PY} = \mathbf{Y} - \widehat{\mathbf{Y}} \overset{\text{Ορισμός 2.2}}{=} \widehat{\boldsymbol{\varepsilon}}$. □

Πρόταση 2.11. (Κατανομές Εκτιμητριών)

i. $\widehat{\boldsymbol{\beta}} \sim N_{p+1}\left(\boldsymbol{\beta}, \sigma^2\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\right)$. Σύμφωνα με την πρόταση 1.11, έπεται ότι:

$$W = \frac{(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})}{\sigma^2} = \frac{\|\mathbf{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|^2}{\sigma^2} = \frac{\|\widehat{\mathbf{Y}} - \mathbf{X}\boldsymbol{\beta}\|^2}{\sigma^2} \sim \chi^2_{p+1}.$$

ii. Το S^2 είναι ανεξάρτητο από το $\widehat{\boldsymbol{\beta}}$ και $Q = \frac{(n-p-1)S^2}{\sigma^2} \sim \chi^2_{n-p-1}$.

Απόδειξη. Σύμφωνα με την πρόταση 1.10 και το θεώρημα Cochran, έχουμε ότι:

i. Γνωρίζουμε ότι $\mathbf{Y} \sim N_n\left(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n\right)$ και $\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}$. Έχουμε υπολογίσει ότι $E(\widehat{\boldsymbol{\beta}}) = \beta$ και $\mathrm{Var}(\widehat{\boldsymbol{\beta}}) = \sigma^2\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}$, οπότε $\widehat{\boldsymbol{\beta}} \sim N_{p+1}\left(\boldsymbol{\beta}, \sigma^2\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\right)$. Σύμφωνα με την πρόταση 1.11, έπεται ότι:

$$\begin{aligned} W &= (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^{\mathrm{T}}\left[\sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\right]^{-1}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \frac{1}{\sigma^2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \\ &= \frac{[\mathbf{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})]^{\mathrm{T}} \cdot \mathbf{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})}{\sigma^2} = \frac{\|\mathbf{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|^2}{\sigma^2} = \frac{\|\widehat{\mathbf{Y}} - \mathbf{X}\boldsymbol{\beta}\|^2}{\sigma^2} \sim \chi^2_{p+1}. \end{aligned}$$

ii. Ορίζουμε $\mathbf{A}_1 = \mathbf{P}$ και $\mathbf{A}_2 = \mathbf{I}_n - \mathbf{P}$. Σύμφωνα με τις προηγούμενες προτάσεις οι πίνακες $\mathbf{A}_1$ και $\mathbf{A}_2$ είναι πίνακες ορθογώνιας προβολής, οπότε και συμμετρικοί. Προφανώς, ισχύει ότι $\mathbf{A}_1 + \mathbf{A}_2 = \mathbf{I}_n$. Υπολογίζουμε ότι:

$$\mathrm{rank}(\mathbf{A}_1) \overset{\text{Λήμμα 2.7}}{=} \mathrm{tr}(\mathbf{A}_1) = \mathrm{tr}(\mathbf{P}) = \mathrm{tr}\left[\mathbf{X} \cdot \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\right] \Rightarrow$$

$$\mathrm{rank}(\mathbf{A}_1) \overset{\text{Πρόταση 2.5}}{=} \mathrm{tr}\left[\cancel{\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}} \cdot \mathbf{X}}\right] = \mathrm{tr}\left(\mathbf{I}_{p+1}\right) = p + 1 \Rightarrow$$

$$\mathrm{rank}(\mathbf{A}_2) \overset{\text{Λήμμα 2.7}}{=} \mathrm{tr}\left(\mathbf{A}_2\right) = \mathrm{tr}\left(\mathbf{I}_n - \mathbf{P}\right) \overset{\text{Πρόταση 2.5}}{=} \mathrm{tr}\left(\mathbf{I}_n\right) - \mathrm{tr}\left(\mathbf{P}\right) = n - p - 1.$$

Επομένως, $\mathrm{rank}(\mathbf{A}_1) + \mathrm{rank}(\mathbf{A}_2) = n$, δηλαδή ισχύουν όλες οι προϋποθέσεις του θεωρήματος Cochran. Τέλος, υπολογίζουμε ότι:

$$\mathbf{A}_1(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{P}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{PY} - \mathbf{PX} \cdot \boldsymbol{\beta} \overset{\text{Λήμμα 2.8}}{=} \widehat{\mathbf{Y}} - \mathbf{X}\boldsymbol{\beta} \Rightarrow$$

$$\begin{aligned} W &= \frac{\|\widehat{\mathbf{Y}} - \mathbf{X}\boldsymbol{\beta}\|^2}{\sigma^2} = \frac{\|\mathbf{A}_1(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})\|^2}{\sigma^2} = \frac{[\mathbf{A}_1(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})]^{\mathrm{T}}\,\mathbf{A}_1(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})}{\sigma^2} \\ &= \frac{(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}\mathbf{A}_1^{\mathrm{T}}\mathbf{A}_1(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})}{\sigma^2} \overset{\text{Λήμμα 2.7}}{=} \frac{(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}\mathbf{A}_1(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})}{\sigma^2} \sim \chi^2_{p+1} \text{ και} \end{aligned}$$

$$\mathbf{A}_2(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) = (\mathbf{I}_n - \mathbf{P})\,(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{Y} - \mathbf{PY} - \cancel{(\mathbf{I}_n - \mathbf{P})\mathbf{X}} \cdot \boldsymbol{\beta} \overset{\text{Λήμμα 2.8}}{=} \mathbf{Y} - \widehat{\mathbf{Y}} = \widehat{\boldsymbol{\varepsilon}} \Rightarrow$$

$$Q = \frac{(n-p-1)S^2}{\sigma^2} = \frac{\text{SSE}}{\sigma^2} = \frac{\|\widehat{\boldsymbol{\varepsilon}}\|^2}{\sigma^2} = \frac{\|\mathbf{A}_2(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})\|^2}{\sigma^2}$$
$$= \frac{[\mathbf{A}_2(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})]^{\mathrm{T}}\,\mathbf{A}_2(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})}{\sigma^2} = \frac{(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}\mathbf{A}_2^{\mathrm{T}}\mathbf{A}_2(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})}{\sigma^2}$$
$$= \frac{(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}\mathbf{A}_2(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})}{\sigma^2} \sim \chi^2_{n-p-1}.$$

Σύμφωνα με το θεώρημα Cochran, οι τυχαίες μεταβλητές Q και W είναι ανεξάρτητες. Εφόσον η τυχαία μεταβλητή Q εξαρτάται μόνο από το S^2 και η τυχαία μεταβλητή W εξαρτάται μόνο από το $\widehat{\boldsymbol{\beta}}$, συμπεραίνουμε ότι το S^2 είναι ανεξάρτητο από το $\widehat{\boldsymbol{\beta}}$. □

Σημείωση 2.2. Σύμφωνα με την πρόταση 2.11, έπεται ότι $\widehat{\beta}_j \sim N\left(\beta_j, \sigma^2_{\widehat{\beta}_j}\right)$ για $j = 0, 1, \ldots, p$, όπου $\sigma^2_{\widehat{\beta}_j} = \text{Var}\left(\widehat{\beta}_j\right) = \sigma^2\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}_{j+1,j+1}$ το $(j+1)$-οστό διαγώνιο στοιχείο του πίνακα $\text{Var}\left(\widehat{\boldsymbol{\beta}}\right) = \sigma^2\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1} \in \mathbb{R}^{(p+1)\times(p+1)}$.

Σημείωση 2.3. Στην περίπτωση του απλού γραμμικού μοντέλου, έχουμε ότι:

$$\mathbf{X} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix} \mathbb{R}^{n\times 2} \Rightarrow \mathbf{X}^{\mathrm{T}}\mathbf{X} = \begin{bmatrix} n & \sum_{i=1}^n X_i \\ \sum_{i=1}^n X_i & \sum_{i=1}^n X_i^2 \end{bmatrix} = n\begin{bmatrix} 1 & \overline{X} \\ \overline{X} & \frac{1}{n}\sum_{i=1}^n X_i^2 \end{bmatrix} \text{ και}$$

$$\det\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right) = n^2\left(\frac{1}{n}\sum_{i=1}^n X_i^2 - \overline{X}^2\right) = n\left(\sum_{i=1}^n X_i^2 - n\overline{X}^2\right) = n(n-1)S_X^2 \Rightarrow$$

$$\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1} = \frac{n}{n(n-1)S_X^2}\begin{bmatrix} \frac{1}{n}\sum_{i=1}^n X_i^2 & -\overline{X} \\ -\overline{X} & 1 \end{bmatrix} = \frac{1}{(n-1)S_X^2}\begin{bmatrix} \frac{1}{n}\sum_{i=1}^n X_i^2 & -\overline{X} \\ -\overline{X} & 1 \end{bmatrix} \Rightarrow$$

$$\Sigma_{\widehat{\boldsymbol{\beta}}} = \text{Var}\left(\widehat{\boldsymbol{\beta}}\right) = \sigma^2\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1} = \frac{\sigma^2}{(n-1)S_X^2}\begin{bmatrix} \frac{1}{n}\sum_{i=1}^n X_i^2 & -\overline{X} \\ -\overline{X} & 1 \end{bmatrix}.$$

Λήμμα 2.9. Έστω $\mathbf{A} \in \mathbb{R}^{p\times p}$, $\mathbf{B} \in \mathbb{R}^{p\times q}$, $\mathbf{C} \in \mathbb{R}^{q\times p}$, $\mathbf{D} \in \mathbb{R}^{q\times q}$ με $\mathbf{A}$ αντιστρέψιμο. Ορίζουμε το συμπλήρωμα Schur του $\mathbf{A}$ ως $\mathbf{E} = \mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B} \in \mathbb{R}^{q\times q}$. Αν $\mathbf{E}$ αντιστρέψιμος, τότε:

$$\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{B}\mathbf{E}^{-1}\mathbf{C}\mathbf{A}^{-1} & -\mathbf{A}^{-1}\mathbf{B}\mathbf{E}^{-1} \\ -\mathbf{E}^{-1}\mathbf{C}\mathbf{A}^{-1} & \mathbf{E}^{-1} \end{bmatrix}.$$

Ορισμός 2.8. Για $j = 0, 1, \ldots, p$ και $i = 1, 2, \ldots, n$,, θεωρούμε το ακόλουθο πολλαπλό γραμμικό μοντέλο:

$$X_{j,i} = \gamma_0 + \gamma_1 X_{1,i} + \cdots + \gamma_{j-1}X_{j-1,i} + \gamma_{j+1}X_{j+1,i} + \cdots + \gamma_p X_{p,i} + \eta_i,$$

όπου $\eta_i \sim N\left(0, \sigma_j^2\right)$ ανεξάρτητα. Τότε, ορίζουμε:

$$\mathrm{SST}_j = \sum_{i=1}^{n} \left(X_{j,i} - \overline{X}_j\right)^2, \quad \mathrm{SSE}_j = \sum_{i=1}^{n} \left(X_{j,i} - \widehat{X}_{j,i}\right)^2 = \sum_{i=1}^{n} \widehat{\eta}_i^2, \quad R_j^2 = 1 - \frac{\mathrm{SSE}_j}{\mathrm{SST}_j}.$$

Πρόταση 2.12. Ισχύει ότι:

$$\mathrm{Var}\left(\widehat{\beta}_j\right) = \frac{\sigma^2}{\mathrm{SSE}_j} = \frac{\sigma^2}{\mathrm{SST}_j} \frac{1}{1 - R_j^2}.$$

Απόδειξη. Έστω $\mathbf{X} = \begin{bmatrix} \mathbf{1}_n & \mathbf{X}_1 & \cdots & \mathbf{X}_p \end{bmatrix} \in \mathbb{R}^{n \times (p+1)}$. Χωρίς βλάβη της γενικότητας, θεωρούμε ότι η στήλη $\mathbf{X}_j$ είναι η τελευταία στήλη του πίνακα $\mathbf{X}$, δηλαδή:

$$\mathbf{X} = \begin{bmatrix} \mathbf{1}_n & \mathbf{X}_1 & \cdots & \mathbf{X}_{j-1} & \mathbf{X}_{j+1} & \cdots & \mathbf{X}_p & \mathbf{X}_j \end{bmatrix} = \begin{bmatrix} \mathbf{X}_{-j} & \mathbf{X}_j \end{bmatrix}.$$

Τότε, παρατηρούμε ότι:

$$\mathbf{X}^{\mathrm{T}}\mathbf{X} = \begin{bmatrix} \mathbf{X}_{-j}^{\mathrm{T}} \\ \mathbf{X}_j^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} \mathbf{X}_{-j} & \mathbf{X}_j \end{bmatrix} = \begin{bmatrix} \mathbf{X}_{-j}^{\mathrm{T}}\mathbf{X}_{-j} & \mathbf{X}_{-j}^{\mathrm{T}}\mathbf{X}_j \\ \mathbf{X}_j^{\mathrm{T}}\mathbf{X}_{-j} & \mathbf{X}_j^{\mathrm{T}}\mathbf{X}_j \end{bmatrix}.$$

Ορίζουμε το συμπλήρωμα Schur του $\mathbf{A} = \mathbf{X}_{-j}^{\mathrm{T}}\mathbf{X}_{-j}$ ως:

$$\begin{aligned} \mathbf{E} &= \mathbf{X}_j^{\mathrm{T}}\mathbf{X}_j - \mathbf{X}_j^{\mathrm{T}}\mathbf{X}_{-j}\left(\mathbf{X}_{-j}^{\mathrm{T}}\mathbf{X}_{-j}\right)^{-1}\mathbf{X}_{-j}^{\mathrm{T}}\mathbf{X}_j = \mathbf{X}_j^{\mathrm{T}}\left(\mathbf{I}_n - \mathbf{P}_{-j}\right)\mathbf{X}_j \\ &= \mathbf{X}_j^{\mathrm{T}}\left(\mathbf{I}_n - \mathbf{P}_{-j}\right)^{\mathrm{T}}\left(\mathbf{I}_n - \mathbf{P}_{-j}\right)\mathbf{X}_j = \widehat{\boldsymbol{\eta}}^{\mathrm{T}}\widehat{\boldsymbol{\eta}} = \|\widehat{\boldsymbol{\eta}}\|^2 = \mathrm{SSE}_j \in \mathbb{R}. \end{aligned}$$

Σύμφωνα με το λήμμα 2.9, παίρνουμε ότι:

$$\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}_{j+1,j+1} = \mathbf{E}^{-1} = \frac{1}{\mathrm{SSE}_j} \Rightarrow$$

$$\mathrm{Var}\left(\widehat{\beta}_j\right) = \sigma^2\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}_{j+1,j+1} = \frac{\sigma^2}{\mathrm{SSE}_j} = \frac{\sigma^2}{\mathrm{SST}_j}\frac{1}{1 - R_j^2}. \quad \square$$

Παρατήρηση 2.3. Παρατηρούμε ότι όσο πιο κοντά στη μονάδα βρίσκεται ο συντελεστής προσδιορισμού R_j^2 τόσο μεγαλύτερη είναι η διασπορά της εκτιμήτριας $\widehat{\beta}_j$. Για τον λόγο αυτό, η ποσότητα $\mathrm{VIF}_j = \frac{1}{1-R_j^2}$ καλείται **variance inflation factor** της εκτιμήτριας $\widehat{\beta}_j$. Επιπλέον, παρατηρούμε ότι ο συντελεστής προσδιορισμού R_j^2 εκφράζει το ποσοστό της μεταβλητότητας της επεξηγηματικής μεταβλητής X_j η οποία μπορεί να εξηγηθεί από τις υπόλοιπες επεξηγηματικές μεταβλητές. Το φαινόμενο αυτό θα μελετηθεί περαιτέρω στην παράγραφο 2.10.

Πρόταση 2.13. (Κατανομή του $\widehat{\boldsymbol{\beta}}$ με χρήση του S^2)

i. Ισχύει ότι:

$$\frac{(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})}{(p+1)S^2}=\frac{\|\mathbf{X}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})\|^2}{(p+1)S^2}=\frac{\|\widehat{\mathbf{Y}}-\mathbf{X}\boldsymbol{\beta}\|^2}{(p+1)S^2}\sim F_{p+1,n-p-1}.$$

ii. $\frac{\widehat{\beta}_j-\beta_j}{S_{\widehat{\beta}_j}}\sim t_{n-p-1}$ για $j=0,1,\ldots,p$, όπου $S^2_{\widehat{\beta}_j}=S^2\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}_{j+1,j+1}$.

Απόδειξη. Για την απόδειξη αυτής της πρότασης, θα χρειαστεί να θυμόμαστε κάποια βασικά στοιχεία σχετικά με την κατανομή t του Student και την κατανομή F του Snedecor.

i. Σύμφωνα με την προηγούμενη πρόταση, οι τυχαίες μεταβλητές $W\sim\chi^2_{p+1}$ και $Q\sim\chi^2_{n-p-1}$ είναι ανεξάρτητες, οπότε παίρνουμε ότι:

$$F=\frac{W}{p+1}\cdot\frac{n-p-1}{Q}=\frac{1}{p+1}\cdot\frac{\|\widehat{\mathbf{Y}}-\mathbf{X}\boldsymbol{\beta}\|^2}{\cancel{\sigma^2}}\cdot\frac{\cancel{\sigma^2}}{S^2}=\frac{\|\widehat{\mathbf{Y}}-\mathbf{X}\boldsymbol{\beta}\|^2}{(p+1)S^2}\sim F_{p+1,n-p-1}.$$

ii. Σύμφωνα με την προηγούμενη σημείωση, παίρνουμε ότι:

$$Z=\frac{\widehat{\beta}_j-\beta_j}{\sigma_{\widehat{\beta}_j}}\sim N(0,1)\ \text{και}\ Q=\frac{(n-p-1)S^2}{\sigma^2}\sim\chi^2_{n-p-1}.$$

Το S^2 είναι ανεξάρτητο από το $\widehat{\boldsymbol{\beta}}$, οπότε η Q είναι ανεξάρτητη από το $\widehat{\beta}_j$. Επομένως, η Q είναι ανεξάρτητη από τη Z. Συμπεραίνουμε ότι:

$$T=\frac{Z}{\sqrt{\frac{Q}{n-p-1}}}=\frac{\widehat{\beta}_j-\beta_j}{\sigma_{\widehat{\beta}_j}}\cdot\frac{\sigma}{S}=\frac{\widehat{\beta}_j-\beta_j}{S_{\widehat{\beta}_j}}\sim t_{n-p-1}.\quad\square$$

2.7 Περιοχές Εμπιστοσύνης και Έλεγχοι Υποθέσεων

Χρησιμοποιώντας την πρόταση 2.13, μπορούμε άμεσα να κατασκευάσουμε διαστήματα και περιοχές εμπιστοσύνης για τις παραμέτρους του μοντέλου.

Πρόταση 2.14. (Διαστήματα και Περιοχές Εμπιστοσύνης)

i. Ένα $100(1-\alpha)\%$ διάστημα εμπιστοσύνης ίσων ουρών για το β_j δίνεται από τη σχέση:

$$I_{1-\alpha}(\beta_j)=\left[\widehat{\beta}_j-t_{n-p-1;\frac{\alpha}{2}}\cdot S_{\widehat{\beta}_j},\widehat{\beta}_j+t_{n-p-1;\frac{\alpha}{2}}\cdot S_{\widehat{\beta}_j}\right],\ j=0,1,\ldots,p.$$

ii. Ένα $100(1-\alpha)\%$ διάστημα εμπιστοσύνης ίσων ουρών για το σ^2 δίνεται από τη σχέση:

$$I_{1-\alpha}\left(\sigma^2\right)=\left[\frac{(n-p-1)S^2}{\chi^2_{n-p-1;\frac{\alpha}{2}}},\frac{(n-p-1)S^2}{\chi^2_{n-p-1;1-\frac{\alpha}{2}}}\right].$$

iii. Ένα $100(1-\alpha)\%$ **ελλειψοειδές εμπιστοσύνης** για το $\boldsymbol{\beta}$ δίνεται από τη σχέση:

$$R_{1-\alpha}(\boldsymbol{\beta}) = \left\{ \boldsymbol{\beta} \in \mathbb{R}^{p+1} : \frac{\|\mathbf{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|^2}{(p+1)S^2} \leqslant F_{p+1,n-p-1;\alpha} \right\}.$$

Απόδειξη. Ακριβώς όπως στην πρόταση 1.16 (σελίδα 27).

Πρόταση 2.15. Υπό τη μηδενική υπόθεση $H_0 : \beta_j = \beta_{j,0}$ για $j = 0, 1, \ldots, p$, έχουμε αποδείξει ότι $T = \frac{\widehat{\beta}_j - \beta_{j,0}}{S_{\widehat{\beta}_j}} \sim t_{n-p-1}$. Αντικαθιστώντας τις τυχαίες μεταβλητές $Y_1, Y_2, \ldots, Y_n$ που εμφανίζονται στην ελεγχοσυνάρτηση T από τις παρατηρήσεις $y_1, y_2, \ldots, y_n$, υπολογίζουμε την παρατηρούμενη τιμή $t = \frac{\widehat{\beta}_j - \beta_{j,0}}{s_{\widehat{\beta}_j}}$.

i. Έστω ότι θέλουμε να πραγματοποιήσουμε τον έλεγχο με την αμφίπλευρη εναλλακτική υπόθεση $H_1 : \beta_j \neq \beta_{j,0}$. Τότε, απορρίπτουμε την H_0 σε ε.σ.σ. α αν και μόνο αν $|t| > t_{n-p-1;\frac{\alpha}{2}}$ ή $\text{p-value}^{(\neq)} = P(|T| \geqslant |t|) < \alpha$ ή $\beta_{j,0} \notin I_{1-\alpha}(\beta_j)$.

ii. Έστω ότι θέλουμε να πραγματοποιήσουμε τον έλεγχο με τη μονόπλευρη εναλλακτική υπόθεση $H_1 : \beta_j > \beta_{j,0}$. Τότε, απορρίπτουμε την H_0 σε ε.σ.σ. α αν και μόνο αν $t > t_{n-p-1;\alpha}$ ή $\text{p-value}^{(>)} = P(T \geqslant t) < \alpha$.

iii. Έστω ότι θέλουμε να πραγματοποιήσουμε τον έλεγχο με τη μονόπλευρη εναλλακτική υπόθεση $H_1 : \beta_j < \beta_{j,0}$. Τότε, απορρίπτουμε την H_0 σε ε.σ.σ. α αν και μόνο αν $t < -t_{n-p-1;\alpha}$ ή $\text{p-value}^{(<)} = P(T \leqslant t) < \alpha$.

Απόδειξη. Ακριβώς όπως στην πρόταση 1.17 (σελίδα 29).

Σημείωση 2.4. Η ελεγχοσυνάρτηση του ελέγχου στατιστικής σημαντικότητας της παραμέτρου β_j προκύπτει θέτοντας $\beta_{j,0} = 0$. Υπό τη μηδενική υπόθεση $H_0 : \beta_j = 0$, γνωρίζουμε ότι $T = \frac{\widehat{\beta}_j}{S_{\widehat{\beta}_j}} \sim t_{n-p-1}$ με παρατηρούμενη τιμή $t = \frac{\widehat{\beta}_j}{s_{\widehat{\beta}_j}}$. Λέμε ότι η παράμετρος β_j είναι **στατιστικά σημαντική** ή στατιστικά διάφορη του μηδενός σε ε.σ.σ. α αν και μόνο αν μπορούμε να απορρίψουμε την H_0, δηλαδή αν και μόνο αν ισχύει κάποιο από τα εξής:

- $|t| > t_{n-p-1;\frac{\alpha}{2}}$,
- $\text{p-value} = P(|T| \geqslant |t|) < \alpha$,
- $0 \notin I_{1-\alpha}(\beta_j) = \left[\widehat{\beta}_j - t_{n-p-1;\frac{\alpha}{2}} \cdot s_{\widehat{\beta}_j}, \widehat{\beta}_j + t_{n-p-1;\frac{\alpha}{2}} \cdot s_{\widehat{\beta}_j}\right]$.

Γενικεύοντας, θέλουμε να πραγματοποιήσουμε τον έλεγχο των υποθέσεων:

$$\begin{cases} H_0 : \beta_{p_0+1} = \beta_{p_0+2} = \cdots = \beta_p = 0 \text{ vs.} \\ H_1 : \beta_j \neq 0 \text{ για κάποιο } j \in \{p_0+1, p_0+2, \ldots, p\}. \end{cases}$$

Υπό την H_0, προκύπτει ότι $\mathbf{Y} = \mathbf{X}_0\boldsymbol{\beta}_0 + \boldsymbol{\varepsilon} = \mathbf{X}\boldsymbol{\beta}_0^* + \boldsymbol{\varepsilon}$, όπου $\mathbf{X}_0 \in \mathbb{R}^{n\times(p_0+1)}$ ο πίνακας σχεδιασμού που αντιστοιχεί στις επεξηγηματικές μεταβλητές $X_1, X_2, \ldots, X_{p_0}$,

$\boldsymbol{\beta}_0 = (\beta_0, \beta_1, \ldots, \beta_{p_0})^{\mathrm{T}} \in \mathbb{R}^{p_0+1}$ και $\boldsymbol{\beta}_0^* = (\beta_0, \beta_1, \ldots, \beta_{p_0}, 0, 0, \ldots, 0)^{\mathrm{T}} \in \mathbb{R}^{p+1}$. Ορίζουμε $\mathbf{P}_0 = \mathbf{X}_0 \left(\mathbf{X}_0^{\mathrm{T}}\mathbf{X}_0\right)^{-1} \mathbf{X}_0^{\mathrm{T}}$ τον πίνακα ορθογώνιας προβολής που αντιστοιχεί στον πίνακα σχεδιασμού $\mathbf{X}_0$, $\widetilde{\boldsymbol{\beta}}_0 = \left(\widetilde{\beta}_0, \widetilde{\beta}_1, \ldots, \widetilde{\beta}_{p_0}\right)^{\mathrm{T}} = \left(\mathbf{X}_0^{\mathrm{T}}\mathbf{X}_0\right)^{-1} \mathbf{X}_0^{\mathrm{T}}\mathbf{Y} \in \mathbb{R}^{p_0+1}$ την εκτιμήτρια ελαχίστων τετραγώνων του $\boldsymbol{\beta}_0$ και $\widetilde{\boldsymbol{\beta}}_0^* = \left(\widetilde{\beta}_0, \ldots, \widetilde{\beta}_{p_0}, 0, \ldots, 0\right)^{\mathrm{T}} \in \mathbb{R}^{p+1}$. Τέλος, ορίζουμε $\widehat{\mathbf{Y}}_0 = \mathbf{X}\widetilde{\boldsymbol{\beta}}_0^* = \mathbf{X}_0\widetilde{\boldsymbol{\beta}}_0$. Προφανώς, ισχύει ότι $\mathbf{P}_0\mathbf{Y} = \widehat{\mathbf{Y}}_0$.

Λήμμα 2.10. Ισχύει ότι $\mathbf{P}\mathbf{P}_0 = \mathbf{P}_0\mathbf{P} = \mathbf{P}_0$.

Απόδειξη. Έστω $\mathbf{X} = \begin{bmatrix}\mathbf{X}_0 & \mathbf{X}_1\end{bmatrix} \in \mathbb{R}^{n\times(p+1)}$. Τότε, παρατηρούμε ότι:

$$\mathbf{X}^{\mathrm{T}}\mathbf{X} = \begin{bmatrix}\mathbf{X}_0^{\mathrm{T}} \\ \mathbf{X}_1^{\mathrm{T}}\end{bmatrix}\begin{bmatrix}\mathbf{X}_0 & \mathbf{X}_1\end{bmatrix} = \begin{bmatrix}\mathbf{X}_0^{\mathrm{T}}\mathbf{X}_0 & \mathbf{X}_0^{\mathrm{T}}\mathbf{X}_1 \\ \mathbf{X}_1^{\mathrm{T}}\mathbf{X}_0 & \mathbf{X}_1^{\mathrm{T}}\mathbf{X}_1\end{bmatrix}.$$

Ορίζουμε το συμπλήρωμα Schur του $\mathbf{A} = \mathbf{X}_0^{\mathrm{T}}\mathbf{X}_0$ ως:

$$\mathbf{E} = \mathbf{X}_1^{\mathrm{T}}\mathbf{X}_1 - \mathbf{X}_1^{\mathrm{T}}\mathbf{X}_0\left(\mathbf{X}_0^{\mathrm{T}}\mathbf{X}_0\right)^{-1}\mathbf{X}_0^{\mathrm{T}}\mathbf{X}_1 = \mathbf{X}_1^{\mathrm{T}}\left(\mathbf{I}_n - \mathbf{P}_0\right)\mathbf{X}_1 \in \mathbb{R}^{(p-p_0)\times(p-p_0)}.$$

Τότε, υπολογίζουμε ότι:

$$\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1} \overset{\text{Λήμμα 2.9}}{=} \begin{bmatrix}\mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{X}_0^{\mathrm{T}}\mathbf{X}_1\mathbf{E}^{-1}\mathbf{X}_1^{\mathrm{T}}\mathbf{X}_0\mathbf{A}^{-1} & -\mathbf{A}^{-1}\mathbf{X}_0^{\mathrm{T}}\mathbf{X}_1\mathbf{E}^{-1} \\ -\mathbf{E}^{-1}\mathbf{X}_1^{\mathrm{T}}\mathbf{X}_0\mathbf{A}^{-1} & \mathbf{E}^{-1}\end{bmatrix} \Rightarrow$$

$$\begin{aligned}\mathbf{P} &= \begin{bmatrix}\mathbf{X}_0 & \mathbf{X}_1\end{bmatrix}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\begin{bmatrix}\mathbf{X}_0^{\mathrm{T}} \\ \mathbf{X}_1^{\mathrm{T}}\end{bmatrix} \\ &= \begin{bmatrix}\mathbf{X}_0 & \mathbf{X}_1\end{bmatrix}\begin{bmatrix}\mathbf{A}^{-1}\mathbf{X}_0^{\mathrm{T}} + \mathbf{A}^{-1}\mathbf{X}_0^{\mathrm{T}}\mathbf{X}_1\mathbf{E}^{-1}\mathbf{X}_1^{\mathrm{T}}\mathbf{P}_0 - \mathbf{A}^{-1}\mathbf{X}_0^{\mathrm{T}}\mathbf{X}_1\mathbf{E}^{-1}\mathbf{X}_1^{\mathrm{T}} \\ -\mathbf{E}^{-1}\mathbf{X}_1^{\mathrm{T}}\mathbf{P}_0 + \mathbf{E}^{-1}\mathbf{X}_1^{\mathrm{T}}\end{bmatrix} \\ &= \mathbf{P}_0 + \mathbf{P}_0\mathbf{X}_1\mathbf{E}^{-1}\mathbf{X}_1^{\mathrm{T}}\mathbf{P}_0 - \mathbf{P}_0\mathbf{X}_1\mathbf{E}^{-1}\mathbf{X}_1^{\mathrm{T}} - \mathbf{X}_1\mathbf{E}^{-1}\mathbf{X}_1^{\mathrm{T}}\mathbf{P}_0 + \mathbf{X}_1\mathbf{E}^{-1}\mathbf{X}_1^{\mathrm{T}} \Rightarrow\end{aligned}$$

$$\mathbf{P}\mathbf{P}_0 = \mathbf{P}_0^2 + \cancel{\mathbf{P}_0\mathbf{X}_1\mathbf{E}^{-1}\mathbf{X}_1^{\mathrm{T}}\mathbf{P}_0^2} - \cancel{\mathbf{P}_0\mathbf{X}_1\mathbf{E}^{-1}\mathbf{X}_1^{\mathrm{T}}\mathbf{P}_0} - \cancel{\mathbf{X}_1\mathbf{E}^{-1}\mathbf{X}_1^{\mathrm{T}}\mathbf{P}_0^2} + \cancel{\mathbf{X}_1\mathbf{E}^{-1}\mathbf{X}_1^{\mathrm{T}}\mathbf{P}_0} = \mathbf{P}_0,$$

$$\mathbf{P}_0\mathbf{P} = \mathbf{P}_0^2 + \cancel{\mathbf{P}_0^2\mathbf{X}_1\mathbf{E}^{-1}\mathbf{X}_1^{\mathrm{T}}\mathbf{P}_0} - \cancel{\mathbf{P}_0^2\mathbf{X}_1\mathbf{E}^{-1}\mathbf{X}_1^{\mathrm{T}}} - \cancel{\mathbf{P}_0\mathbf{X}_1\mathbf{E}^{-1}\mathbf{X}_1^{\mathrm{T}}\mathbf{P}_0} + \cancel{\mathbf{P}_0\mathbf{X}_1\mathbf{E}^{-1}\mathbf{X}_1^{\mathrm{T}}} = \mathbf{P}_0. \quad \square$$

Πρόταση 2.16. Υπό τη μηδενική υπόθεση $H_0 : \beta_{p_0+1} = \beta_{p_0+2} = \cdots = \beta_p = 0$, ισχύει ότι:

$$F = \frac{n-p-1}{p-p_0} \cdot \frac{\|\widehat{\mathbf{Y}} - \widehat{\mathbf{Y}}_0\|^2}{\|\mathbf{Y} - \widehat{\mathbf{Y}}\|^2} \sim F_{p-p_0, n-p-1}.$$

Έστω f η παρατηρούμενη τιμή της ελεγχοσυνάρτησης F. Απορρίπτουμε την H_0 σε ε.σ.σ. α αν και μόνο αν $f > F_{p-p_0,n-p-1;\alpha}$ ή p-value $= P(F \geqslant f) < \alpha$.

Απόδειξη.* Ορίζουμε $\mathbf{A}_1 = \mathbf{I}_n - \mathbf{P}$, $\mathbf{A}_2 = \mathbf{P} - \mathbf{P}_0$ και $\mathbf{A}_3 = \mathbf{P}_0$. Αρκεί να δείξουμε ότι ο πίνακας $\mathbf{A}_2$ είναι πίνακας ορθογώνιας προβολής:

$$\mathbf{A}_2^2 = (\mathbf{P} - \mathbf{P}_0)^2 = \mathbf{P}^2 + \mathbf{P}_0^2 - \mathbf{P}\mathbf{P}_0 - \mathbf{P}_0\mathbf{P} \overset{\text{Λήμμα 2.10}}{=} \mathbf{P} + \cancel{\mathbf{P}_0} - \cancel{\mathbf{P}_0} - \mathbf{P}_0 = \mathbf{P} - \mathbf{P}_0 = \mathbf{A}_2,$$

$$\mathbf{A}_2^{\mathrm{T}} = (\mathbf{P} - \mathbf{P}_0)^{\mathrm{T}} = \mathbf{P}^{\mathrm{T}} - \mathbf{P}_0^{\mathrm{T}} = \mathbf{P} - \mathbf{P}_0 = \mathbf{A}_2.$$

Οι πίνακες $\mathbf{A}_1$, $\mathbf{A}_2$ και $\mathbf{A}_3$ είναι πίνακες ορθογώνιας προβολής, οπότε και συμμετρικοί. Προφανώς, ισχύει ότι $\mathbf{A}_1+\mathbf{A}_2+\mathbf{A}_3 = \mathbf{I}_n$. Έχουμε υπολογίσει ότι $\mathrm{rank}(\mathbf{P}) = p+1$ και $\mathrm{rank}(\mathbf{A}_1) = \mathrm{rank}\,(\mathbf{I}_n - \mathbf{P}) = n - p - 1$. Ομοίως, $\mathrm{rank}(\mathbf{A}_3) = \mathrm{rank}(\mathbf{P}_0) = p_0 + 1$, οπότε:

$$\mathrm{rank}(\mathbf{A}_2) \overset{\text{Λήμμα 2.7}}{=} \mathrm{tr}\,(\mathbf{A}_2) = \mathrm{tr}\,(\mathbf{P} - \mathbf{P}_0) \overset{\text{Πρόταση 2.5}}{=} \mathrm{tr}\,(\mathbf{P}) - \mathrm{tr}\,(\mathbf{P}_0) = p - p_0.$$

Επομένως, $\mathrm{rank}(\mathbf{A}_1) + \mathrm{rank}(\mathbf{A}_2) + \mathrm{rank}(\mathbf{A}_3) = n$, δηλαδή ισχύουν όλες οι προϋποθέσεις του θεωρήματος Cochran. Υπό την H_0, γνωρίζουμε ότι $E(\mathbf{Y}) = \mathbf{X}_0\boldsymbol{\beta}_0 = \mathbf{X}\boldsymbol{\beta}_0^*$, οπότε υπολογίζουμε ότι:

$$\mathbf{A}_3(\mathbf{Y} - \mathbf{X}_0\boldsymbol{\beta}_0) = \mathbf{P}_0(\mathbf{Y} - \mathbf{X}_0\boldsymbol{\beta}_0) = \mathbf{P}_0\mathbf{Y} - \mathbf{P}_0\mathbf{X}_0 \cdot \boldsymbol{\beta}_0 = \widehat{\mathbf{Y}}_0 - \mathbf{X}_0\boldsymbol{\beta}_0 \Rightarrow$$

$$\begin{aligned} V &= \frac{\|\widehat{\mathbf{Y}}_0 - \mathbf{X}_0\boldsymbol{\beta}_0\|^2}{\sigma^2} = \frac{\|\mathbf{A}_3(\mathbf{Y} - \mathbf{X}_0\boldsymbol{\beta}_0)\|^2}{\sigma^2} = \frac{[\mathbf{A}_3(\mathbf{Y} - \mathbf{X}_0\boldsymbol{\beta}_0)]^{\mathrm{T}}\,\mathbf{A}_3(\mathbf{Y} - \mathbf{X}_0\boldsymbol{\beta}_0)}{\sigma^2} \\ &= \frac{(\mathbf{Y} - \mathbf{X}_0\boldsymbol{\beta}_0)^{\mathrm{T}}\mathbf{A}_3^{\mathrm{T}}\mathbf{A}_3(\mathbf{Y} - \mathbf{X}_0\boldsymbol{\beta}_0)}{\sigma^2} \overset{\text{Λήμμα 2.7}}{=} \frac{(\mathbf{Y} - \mathbf{X}_0\boldsymbol{\beta}_0)^{\mathrm{T}}\mathbf{A}_3(\mathbf{Y} - \mathbf{X}_0\boldsymbol{\beta}_0)}{\sigma^2} \sim \chi^2_{p_0+1}, \end{aligned}$$

$$\begin{aligned} \mathbf{A}_2(\mathbf{Y} - \mathbf{X}_0\boldsymbol{\beta}_0) &= (\mathbf{P} - \mathbf{P}_0)\,(\mathbf{Y} - \mathbf{X}_0\boldsymbol{\beta}_0) = \mathbf{P}\mathbf{Y} - \mathbf{P}\cdot\mathbf{X}_0\boldsymbol{\beta}_0 - \mathbf{P}_0\mathbf{Y} + \mathbf{P}_0\mathbf{X}_0\cdot\boldsymbol{\beta}_0 \\ &= \widehat{\mathbf{Y}} - \mathbf{P}\cdot\mathbf{X}\boldsymbol{\beta}_0^* - \widehat{\mathbf{Y}}_0 + \mathbf{X}_0\boldsymbol{\beta}_0 = \widehat{\mathbf{Y}} - \cancel{\mathbf{X}\boldsymbol{\beta}_0^*} - \widehat{\mathbf{Y}}_0 + \cancel{\mathbf{X}_0\boldsymbol{\beta}_0} = \widehat{\mathbf{Y}} - \widehat{\mathbf{Y}}_0 \Rightarrow \end{aligned}$$

$$\begin{aligned} W &= \frac{\|\widehat{\mathbf{Y}} - \widehat{\mathbf{Y}}_0\|^2}{\sigma^2} = \frac{\|\mathbf{A}_2(\mathbf{Y} - \mathbf{X}_0\boldsymbol{\beta}_0)\|^2}{\sigma^2} = \frac{[\mathbf{A}_2(\mathbf{Y} - \mathbf{X}_0\boldsymbol{\beta}_0)]^{\mathrm{T}}\,\mathbf{A}_2(\mathbf{Y} - \mathbf{X}_0\boldsymbol{\beta}_0)}{\sigma^2} \\ &= \frac{(\mathbf{Y} - \mathbf{X}_0\boldsymbol{\beta}_0)^{\mathrm{T}}\mathbf{A}_2^{\mathrm{T}}\mathbf{A}_2(\mathbf{Y} - \mathbf{X}_0\boldsymbol{\beta}_0)}{\sigma^2} = \frac{(\mathbf{Y} - \mathbf{X}_0\boldsymbol{\beta}_0)^{\mathrm{T}}\mathbf{A}_2(\mathbf{Y} - \mathbf{X}_0\boldsymbol{\beta}_0)}{\sigma^2} \sim \chi^2_{p-p_0} \text{ και} \end{aligned}$$

$$\mathbf{A}_1(\mathbf{Y} - \mathbf{X}_0\boldsymbol{\beta}_0) = (\mathbf{I}_n - \mathbf{P})\,(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}_0^*) = \mathbf{Y} - \mathbf{P}\mathbf{Y} - \cancel{(\mathbf{I}_n - \mathbf{P})\mathbf{X}}\cdot\boldsymbol{\beta}_0^* \overset{\text{Λήμμα 2.8}}{=} \mathbf{Y} - \widehat{\mathbf{Y}} = \widehat{\boldsymbol{\varepsilon}} \Rightarrow$$

$$\begin{aligned} Q &= \frac{(n-p-1)S^2}{\sigma^2} = \frac{\|\widehat{\boldsymbol{\varepsilon}}\|^2}{\sigma^2} = \frac{\|\mathbf{A}_1(\mathbf{Y} - \mathbf{X}_0\boldsymbol{\beta}_0)\|^2}{\sigma^2} = \frac{[\mathbf{A}_1(\mathbf{Y} - \mathbf{X}_0\boldsymbol{\beta}_0)]^{\mathrm{T}}\,\mathbf{A}_1(\mathbf{Y} - \mathbf{X}_0\boldsymbol{\beta}_0)}{\sigma^2} \\ &= \frac{(\mathbf{Y} - \mathbf{X}_0\boldsymbol{\beta}_0)^{\mathrm{T}}\mathbf{A}_1^{\mathrm{T}}\mathbf{A}_1(\mathbf{Y} - \mathbf{X}_0\boldsymbol{\beta}_0)}{\sigma^2} = \frac{(\mathbf{Y} - \mathbf{X}_0\boldsymbol{\beta}_0)^{\mathrm{T}}\mathbf{A}_1(\mathbf{Y} - \mathbf{X}_0\boldsymbol{\beta}_0)}{\sigma^2} \sim \chi^2_{n-p-1}. \end{aligned}$$

Σύμφωνα με το θεώρημα Cochran, οι τυχαίες μεταβλητές Q και W είναι ανεξάρτητες. Επομένως, υπό την H_0, ισχύει ότι:

$$F = \frac{W}{p - p_0} \cdot \frac{n - p - 1}{Q} = \frac{n - p - 1}{p - p_0} \cdot \frac{\|\widehat{\mathbf{Y}} - \widehat{\mathbf{Y}}_0\|^2}{\|\mathbf{Y} - \widehat{\mathbf{Y}}\|^2} \sim F_{p-p_0,n-p-1}.$$

Για το πολλαπλό γραμμικό μοντέλο, έχουμε υπολογίσει ότι:

$$L\left(\boldsymbol{\beta}, \sigma^2|\mathbf{y}\right) = \left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left\{-\frac{\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2}{2\sigma^2}\right\} \Rightarrow$$

$$\ell\left(\boldsymbol{\beta}, \sigma^2|\mathbf{y}\right) = -\frac{n\log(2\pi)}{2} - \frac{n\log\sigma^2}{2} - \frac{\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2}{2\sigma^2}.$$

Έχουμε υπολογίσει τις εκτιμήτριες μέγιστης πιθανοφάνειας $\widehat{\boldsymbol{\beta}}$ και $\widehat{\sigma}^2 = \frac{1}{n}\|\widehat{\boldsymbol{\varepsilon}}\|^2$ των παραμέτρων $\boldsymbol{\beta}$ και σ^2. Υπό την H_0, έχουμε υπολογίσει την εκτιμήτρια ελαχίστων τετραγώνων $\widetilde{\boldsymbol{\beta}}_0$ του $\boldsymbol{\beta}_0$, η οποία ταυτίζεται με την εκτιμήτρια μέγιστης πιθανοφάνειας του $\boldsymbol{\beta}_0$. Με όμοιο τρόπο, προκύπτει η εκτιμήτρια μέγιστης πιθανοφάνειας $\widetilde{\sigma}^2 = \frac{1}{n}\|\mathbf{y} - \widehat{\mathbf{y}}_0\|^2 = \frac{1}{n}\|\mathbf{y} - \mathbf{X}\widetilde{\boldsymbol{\beta}}_0^*\|^2$ του σ^2. Υπολογίζουμε τον λογάριθμο του γενικευμένου λόγου πιθανοφανειών:

$$\begin{aligned}\log\lambda^* &= \ell\left(\widetilde{\boldsymbol{\beta}}_0, \widetilde{\sigma}^2|\mathbf{y}\right) - \ell\left(\widehat{\boldsymbol{\beta}}, \widehat{\sigma}^2|\mathbf{y}\right) = -\frac{n}{2}\log\frac{\widetilde{\sigma}^2}{\widehat{\sigma}^2} - \frac{1}{2\widetilde{\sigma}^2}\|\mathbf{y} - \mathbf{X}\widetilde{\boldsymbol{\beta}}_0^*\|^2 + \frac{1}{2\widehat{\sigma}^2}\|\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\|^2 \\ &= -\frac{n}{2}\log\frac{\widetilde{\sigma}^2}{\widehat{\sigma}^2} - \cancel{\frac{n}{2}} + \cancel{\frac{n}{2}} = -\frac{n}{2}\log\frac{\widetilde{\sigma}^2}{\widehat{\sigma}^2},\end{aligned}$$

Θα εκφράσουμε την εκτιμήτρια $\widetilde{\sigma}^2$ συναρτήσει της εκτιμήτριας $\widehat{\sigma}^2$:

$$\begin{aligned}\widetilde{\sigma}^2 &= \frac{\|\mathbf{y} - \widehat{\mathbf{y}} + \widehat{\mathbf{y}} - \widehat{\mathbf{y}}_0\|^2}{n} = \frac{\|\mathbf{y} - \widehat{\mathbf{y}}\|^2 + \|\widehat{\mathbf{y}} - \widehat{\mathbf{y}}_0\|^2 + 2\left(\widehat{\mathbf{y}} - \widehat{\mathbf{y}}_0\right)^{\mathrm{T}}(\mathbf{y} - \widehat{\mathbf{y}})}{n} \\ &= \widehat{\sigma}^2 + \frac{\|\widehat{\mathbf{y}} - \widehat{\mathbf{y}}_0\|^2 + 2\cancel{\widehat{\mathbf{y}}^{\mathrm{T}}\widehat{\boldsymbol{\varepsilon}}} - 2\widetilde{\boldsymbol{\beta}}_0^{*\mathrm{T}}\cancel{\mathbf{X}^{\mathrm{T}}\widehat{\boldsymbol{\varepsilon}}}}{n} \overset{\text{Πρόταση 2.5}}{=} \widehat{\sigma}^2 + \frac{\|\widehat{\mathbf{y}} - \widehat{\mathbf{y}}_0\|^2}{n} \Rightarrow\end{aligned}$$

$$\log\lambda^* = -\frac{n}{2}\log\left[1 + \frac{\|\widehat{\mathbf{y}} - \widehat{\mathbf{y}}_0\|^2}{n\widehat{\sigma}^2}\right] = -\frac{n}{2}\log\left[1 + \frac{\|\widehat{\mathbf{y}} - \widehat{\mathbf{y}}_0\|^2}{\|\mathbf{y} - \widehat{\mathbf{y}}\|^2}\right].$$

Για τον υπολογισμό της κρίσιμης περιοχής του ελέγχου πρέπει να λύσουμε την ανισότητα $\lambda^* < c$ ως προς κάποια στατιστική συνάρτηση:

$$\lambda^* < c \Leftrightarrow \log\lambda^* < c^* = \log c \Leftrightarrow 1 + \frac{\|\widehat{\mathbf{y}} - \widehat{\mathbf{y}}_0\|^2}{\|\mathbf{y} - \widehat{\mathbf{y}}\|^2} > c^{**} = e^{-\frac{2c^*}{n}} \Leftrightarrow$$

$$f = \frac{n-p-1}{p-p_0} \cdot \frac{\|\widehat{\mathbf{y}} - \widehat{\mathbf{y}}_0\|^2}{\|\mathbf{y} - \widehat{\mathbf{y}}\|^2} > c_\alpha = \frac{n-p-1}{p-p_0}(c^{**} - 1).$$

Υπό την H_0, δείξαμε ότι $F = \frac{n-p-1}{p-p_0} \cdot \frac{\|\widehat{\mathbf{Y}} - \widehat{\mathbf{Y}}_0\|^2}{\|\mathbf{Y} - \widehat{\mathbf{Y}}\|^2} \sim F_{p-p_0, n-p-1}$. Για τον υπολογισμό της σταθεράς c_α, απαιτούμε η πιθανότητα σφάλματος τύπου I του ελέγχου, δηλαδή η πιθανότητα λανθασμένης απόρριψης της H_0, να είναι ίση με α:

$$P_{H_0}(F > c_\alpha) = \alpha \Rightarrow c_\alpha = F_{p-p_0, n-p-1; \alpha}.$$

Επομένως, απορρίπτουμε την H_0 αν και μόνο αν $f > F_{p-p_0, n-p-1; \alpha}$. □

Σημείωση 2.5. Υπό την H_0, ορίζουμε $\mathrm{SSR}_0 = \|\widehat{\mathbf{Y}}_0 - \overline{Y}\mathbf{1}_n\|^2$ και $\mathrm{SSE}_0 = \|\mathbf{Y} - \widehat{\mathbf{Y}}_0\|^2$. Υπό την H_0, το γραμμικό μοντέλο χρησιμοποιεί μόνο ένα υποσύνολο του συνόλου των επεξηγηματικών μεταβλητών, οπότε εξηγεί αναγκαστικά μικρότερο ποσοστό από τη συνολική μεταβλητότητα των δεδομένων. Επομένως, θα ισχύει $\mathrm{SSR}_0 \leqslant \mathrm{SSR}$ και $\mathrm{SSE}_0 \geqslant \mathrm{SSE}$. Επιπλέον, γνωρίζουμε ότι $\mathrm{SST} = \|\mathbf{Y} - \overline{Y}\mathbf{1}_n\|^2 = \mathrm{SSR} + \mathrm{SSE}$.

Ορίζουμε τον συντελεστή προσδιορισμού του γραμμικού μοντέλου υπό την H_0:

$$R_0^2 = \frac{\mathrm{SSR}_0}{\mathrm{SST}} = 1 - \frac{\mathrm{SSE}_0}{\mathrm{SST}} \leqslant R^2.$$

Στην προηγούμενη απόδειξη, δείξαμε ότι $\|\mathbf{Y}-\widehat{\mathbf{Y}}_0\|^2 = \|\mathbf{Y}-\widehat{\mathbf{Y}}\|^2+\|\widehat{\mathbf{Y}}-\widehat{\mathbf{Y}}_0\|^2$, δηλαδή $\|\widehat{\mathbf{Y}} - \widehat{\mathbf{Y}}_0\|^2 = \mathrm{SSE}_0 - \mathrm{SSE} \geqslant 0$. Με αυτόν τον τρόπο, παίρνουμε μία εναλλακτική γραφή της ελεγχοσυνάρτησης F:

$$\begin{aligned} F &= \frac{n-p-1}{p-p_0} \cdot \frac{\|\widehat{\mathbf{Y}} - \widehat{\mathbf{Y}}_0\|^2}{\|\mathbf{Y} - \widehat{\mathbf{Y}}\|^2} = \frac{n-p-1}{p-p_0} \cdot \frac{\mathrm{SSE}_0 - \mathrm{SSE}}{\mathrm{SSE}} = \frac{n-p-1}{p-p_0} \cdot \frac{\mathrm{SSR} - \mathrm{SSR}_0}{\mathrm{SST} - \mathrm{SSR}} \\ &= \frac{n-p-1}{p-p_0} \cdot \frac{R^2 - R_0^2}{1-R^2} \sim F_{p-p_0, n-p-1}. \end{aligned}$$

2.8 Ανάλυση Διασποράς - ANOVA

Ενδιαφερόμαστε να πραγματοποιήσουμε τον έλεγχο των υποθέσεων:

$$\begin{cases} H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0 \text{ vs.} \\ H_1 : \beta_j \neq 0 \text{ για κάποιο } j \in \{1, 2, \ldots, p\}. \end{cases}$$

Ο έλεγχος αυτός μας επιτρέπει να αποφασίσουμε αν υπάρχει τουλάχιστον μία από τις p διαθέσιμες επεξηγηματικές μεταβλητές, η οποία να συνεισφέρει σημαντικά στην πρόβλεψη της αποκριτικής μεταβλητής Y. Μεγάλη προσοχή πρέπει να δοθεί στο γεγονός ότι ο σταθερός όρος β_0 δε μηδενίζεται στη μηδενική υπόθεση του ελέγχου, καθώς δεν εμπλέκεται με καμία από τις επεξηγηματικές μεταβλητές.

Παρατηρούμε ότι αυτός ο έλεγχος είναι ειδική περίπτωση του ελέγχου που πραγματοποιήσαμε στην πρόταση 2.16, αν θέσουμε $p_0 = 0$. Στη συγκεκριμένη περίπτωση, προκύπτει ότι $\mathbf{X}_0 = \mathbf{1}_n \in \mathbb{R}^n$, οπότε $\mathbf{Y} = \beta_0 \mathbf{1}_n + \boldsymbol{\varepsilon}$. Υπολογίζουμε ότι:

$$\mathbf{1}_n^{\mathrm{T}}\mathbf{1}_n = n \Rightarrow \left(\mathbf{1}_n^{\mathrm{T}}\mathbf{1}_n\right)^{-1} = \frac{1}{n} \Rightarrow \mathbf{P}_0 = \mathbf{X}_0 \left(\mathbf{X}_0^{\mathrm{T}}\mathbf{X}_0\right)^{-1} \mathbf{X}_0^{\mathrm{T}} = \mathbf{1}_n \left(\mathbf{1}_n^{\mathrm{T}}\mathbf{1}_n\right)^{-1} \mathbf{1}_n^{\mathrm{T}} = \frac{\mathbf{1}_n \mathbf{1}_n^{\mathrm{T}}}{n},$$

$$\widetilde{\beta}_0 = \left(\mathbf{X}_0^{\mathrm{T}}\mathbf{X}_0\right)^{-1} \mathbf{X}_0^{\mathrm{T}}\mathbf{Y} = \left(\mathbf{1}_n^{\mathrm{T}}\mathbf{1}_n\right)^{-1} \mathbf{1}_n^{\mathrm{T}}\mathbf{Y} = \frac{1}{n} \cdot \mathbf{1}_n^{\mathrm{T}}\mathbf{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i = \overline{Y},$$

$$\widehat{\mathbf{Y}}_0 = \mathbf{X}_0\widetilde{\beta}_0 = \overline{Y}\mathbf{1}_n \text{ και } \mathbf{P}_0\mathbf{Y} = \frac{\mathbf{1}_n\mathbf{1}_n^{\mathrm{T}}}{n} \cdot \mathbf{Y} = \mathbf{1}_n \cdot \frac{\mathbf{1}_n^{\mathrm{T}}\mathbf{Y}}{n} = \mathbf{1}_n\overline{Y} = \widehat{\mathbf{Y}}_0.$$

Υπό την H_0, σύμφωνα με το θεώρημα Cochran, συμπεραίνουμε ότι:

$$V = \frac{\|\widehat{\mathbf{Y}}_0 - \mathbf{X}_0\beta_0\|^2}{\sigma^2} = \frac{\|\overline{Y}\mathbf{1}_n - \beta_0\mathbf{1}_n\|^2}{\sigma^2} = \frac{\left(\overline{Y} - \beta_0\right)^2 \|\mathbf{1}_n\|^2}{\sigma^2} = \frac{n\left(\overline{Y} - \beta_0\right)^2}{\sigma^2} \sim \chi_1^2,$$

$$W = \frac{\|\widehat{\mathbf{Y}} - \widehat{\mathbf{Y}}_0\|^2}{\sigma^2} = \frac{\|\widehat{\mathbf{Y}} - \overline{Y}\mathbf{1}_n\|^2}{\sigma^2} = \frac{\mathrm{SSR}}{\sigma^2} \sim \chi_p^2 \text{ και}$$

$$Q = \frac{\|\mathbf{Y} - \widehat{\mathbf{Y}}\|^2}{\sigma^2} = \frac{\text{SSE}}{\sigma^2} \sim \chi^2_{n-p-1}.$$

Στο ίδιο αποτέλεσμα για την τυχαία μεταβλητή V θα καταλήγαμε, αν παρατηρούσαμε ότι $\overline{Y} \sim N\left(\beta_0, \frac{\sigma^2}{n}\right)$, υπό την H_0, σύμφωνα με την πρόταση 2.1. Επομένως,

$$Z = \frac{\overline{Y} - \beta_0}{\sigma}\sqrt{n} \sim N(0,1) \Rightarrow V = Z^2 = \frac{n\left(\overline{Y} - \beta_0\right)^2}{\sigma^2} \sim \chi^2_1.$$

Η κατανομή του SSR έχει p βαθμούς ελευθερίας, οπότε ορίζουμε **μέσο άθροισμα τετραγώνων που οφείλεται στην παλινδρόμηση** (mean sum of squares due to regression) το $\text{MSR} = \frac{\text{SSR}}{p}$. Πάλι σύμφωνα με το θεώρημα Cochran, οι τυχαίες μεταβλητές SSR και SSE είναι ανεξάρτητες. Τελικά, παίρνουμε ότι:

$$F = \frac{n-p-1}{p} \cdot \frac{\|\widehat{\mathbf{Y}} - \widehat{\mathbf{Y}}_0\|^2}{\|\mathbf{Y} - \widehat{\mathbf{Y}}\|^2} = \frac{n-p-1}{p} \cdot \frac{\text{SSR}}{\text{SSE}} = \frac{\text{MSR}}{\text{MSE}} \sim F_{p-p_0,n-p-1}.$$

Έστω f η παρατηρούμενη τιμή της ελεγχοσυνάρτησης F. Απορρίπτουμε την H_0 σε ε.σ.σ. α αν και μόνο αν $f > F_{p-p_0,n-p-1;\alpha}$ ή p-value $= P(F \geqslant f) < \alpha$.

Σημείωση 2.6. Εφόσον $\widehat{\mathbf{Y}}_0 = \overline{Y}\mathbf{1}_n$, παίρνουμε ότι $\text{SSR}_0 = \|\widehat{\mathbf{Y}}_0 - \overline{Y}\mathbf{1}_n\|^2 = 0$ και $\text{SSE}_0 = \|\mathbf{Y} - \widehat{\mathbf{Y}}_0\|^2 = \|\mathbf{Y} - \overline{Y}\mathbf{1}_n\|^2 = \text{SST}$, σύμφωνα με τη σημείωση 2.5. Επομένως, παίρνουμε ότι $R^2_0 = 0$ και η ελεγχοσυνάρτηση F γράφεται ισοδύναμα ως:

$$F = \frac{\text{MSR}}{\text{MSE}} = \frac{n-p-1}{p} \cdot \frac{\text{SSR}}{\text{SST} - \text{SSR}} = \frac{n-p-1}{p} \cdot \frac{R^2}{1-R^2} \sim F_{p,n-p-1}.$$

Παρατηρούμε ότι οι βαθμοί ελευθερίας των κατανομών του SSR και του SSE αθροίζουν στους βαθμούς ελευθερίας της κατανομής του SST. Όλα τα παραπάνω τα συνοψίζουμε στον λεγόμενο πίνακα ανάλυσης διασποράς (ANOVA - analysis of variance) για το πολλαπλό γραμμικό μοντέλο. Καλείται έτσι επειδή βασίζεται στην ανάλυση της συνολικής μεταβλητότητας SST στις συνιστώσες SSR και SSE.

	Sum of Squares	d.f.	Mean Square	$F_{p,n-p-1}$	p-value
R	$\text{SSR} = \|\widehat{\mathbf{Y}} - \overline{Y}\mathbf{1}_n\|^2$	p	$\text{MSR} = \frac{\text{SSR}}{p}$	$f = \frac{\text{MSR}}{\text{MSE}}$	$P(F \geqslant f)$
E	$\text{SSE} = \|\mathbf{Y} - \widehat{\mathbf{Y}}\|^2$	$n-p-1$	$\text{MSE} = \frac{\text{SSE}}{n-p-1}$		
T	$\text{SST} = \|\mathbf{Y} - \overline{Y}\mathbf{1}_n\|^2$	$n-1$			

Πινακας 2.1: Πίνακας ANOVA για το Κανονικό Πολλαπλό Γραμμικό Μοντέλο

2.9 Διαστήματα Μέσης και Ατομικής Πρόβλεψης

Για δεδομένο διάνυσμα παρατηρήσεων $\mathbf{X}_i = (1, X_{1,i}, X_{2,i}, \ldots, X_{p,i})^{\mathrm{T}} \in \mathbb{R}^{p+1}$, θέλουμε αρχικά να κατασκευάσουμε διάστημα εμπιστοσύνης για τη $E(Y_i) = \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}$. Αυτό το διάστημα εμπιστοσύνης καλείται και **διάστημα μέσης πρόβλεψης** για το Y_i. Γνωρίζουμε ότι $E(\widehat{Y}_i) = \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta} = E(Y_i)$, δηλαδή το $\widehat{Y}_i$ είναι μία αμερόληπτη εκτιμήτρια της $E(Y_i)$. Επιπλέον, γνωρίζουμε ότι το $\widehat{Y}_i$ είναι γραμμικός συνδυασμός του $\widehat{\boldsymbol{\beta}}$, το οποίο με τη σειρά του είναι γραμμικός συνδυασμός του $\mathbf{Y}$, οπότε θα είναι κανονικά κατανεμημένο. Υπολογίζουμε ότι:

$$\mathrm{Var}\left(\widehat{Y}_i\right) \overset{\text{Πρόταση 1.9}}{=} \mathbf{X}_i^{\mathrm{T}}\mathrm{Var}\left(\widehat{\boldsymbol{\beta}}\right)\mathbf{X}_i \overset{\text{Πρόταση 2.4}}{=} \sigma^2\mathbf{X}_i^{\mathrm{T}}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}_i = \sigma^2\mathbf{P}_{i,i},$$

όπου $\mathbf{P}_{i,i} = \mathbf{X}_i^{\mathrm{T}}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}_i$ το i-οστό διαγώνιο στοιχείο του πίνακα ορθογώνιας προβολής $\mathbf{P} = \mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}$. Επομένως,

$$\widehat{Y}_i \sim N\left(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}, \sigma^2\mathbf{P}_{i,i}\right) \Rightarrow Z = \frac{\widehat{Y}_i - E(Y_i)}{\sigma\sqrt{\mathbf{P}_{i,i}}} \sim N(0,1).$$

Επιπλέον, γνωρίζουμε ότι $Q = \frac{(n-p-1)S^2}{\sigma^2} \sim \chi^2_{n-p-1}$. Εφόσον η τυχαία μεταβλητή S^2 είναι ανεξάρτητη από το $\widehat{\boldsymbol{\beta}}$, θα είναι ανεξάρτητη και από το $\widehat{Y}_i$. Συμπεραίνουμε ότι η τυχαία μεταβλητή Z είναι ανεξάρτητη από την Q, οπότε:

$$T = \frac{Z}{\sqrt{\frac{Q}{n-p-1}}} = \frac{\widehat{Y}_i - E(Y_i)}{S_{\widehat{Y}_i}} \sim t_{n-p-1}, \text{ όπου } S^2_{\widehat{Y}_i} = S^2\mathbf{P}_{i,i}.$$

Ζητάμε σταθερές $c_1, c_2 \in \mathbb{R}$ τέτοιες, ώστε $P(c_1 \leqslant T \leqslant c_2) = 1 - \alpha$. Συγκεκριμένα, για το διάστημα εμπιστοσύνης ίσων ουρών χρησιμοποιούμε τις σχέσεις:

$$P(T < c_1) = \frac{\alpha}{2} \Rightarrow P(T > c_1) = 1 - \frac{\alpha}{2} \Rightarrow c_1 = t_{n-p-1;1-\frac{\alpha}{2}} = -t_{n-p-1;\frac{\alpha}{2}},$$

$$P(T > c_2) = \frac{\alpha}{2} \Rightarrow c_2 = t_{n-p-1;\frac{\alpha}{2}}.$$

Επομένως, το διάστημα εμπιστοσύνης ίσων ουρών δίνεται από τη σχέση:

$$c_1 \leqslant \frac{\widehat{Y}_i - E(Y_i)}{S_{\widehat{Y}_i}} \leqslant c_2 \Leftrightarrow -c_2 \cdot S_{\widehat{Y}_i} \leqslant E(Y_i) - \widehat{Y}_i \leqslant -c_1 \cdot S_{\widehat{Y}_i} \Leftrightarrow$$

$$\widehat{Y}_i - t_{n-p-1;\frac{\alpha}{2}} \cdot S_{\widehat{Y}_i} \leqslant E(Y_i) \leqslant \widehat{Y}_i + t_{n-p-1;\frac{\alpha}{2}} \cdot S_{\widehat{Y}_i}.$$

Τελικά, παίρνουμε ότι $I_{1-\alpha}\left(E(Y_i)\right) = \left[\widehat{Y}_i - t_{n-p-1;\frac{\alpha}{2}} \cdot S_{\widehat{Y}_i}, \widehat{Y}_i + t_{n-p-1;\frac{\alpha}{2}} \cdot S_{\widehat{Y}_i}\right]$.

Έχοντας μία νέα παρατήρηση $\mathbf{X}_{n+1} = (1, X_{1,n+1}, X_{2,n+1}, \ldots, X_{p,n+1})^{\mathrm{T}} \in \mathbb{R}^{p+1}$, θέλουμε να κατασκευάσουμε διάστημα πρόβλεψης για το $Y_{n+1} = \mathbf{X}_{n+1}^{\mathrm{T}}\boldsymbol{\beta} + \varepsilon_{n+1}$

την οποία δεν έχουμε παρατηρήσει, όπου $\varepsilon_{n+1} \sim N\left(0, \sigma^2\right)$ ανεξάρτητο από τα $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$. Αυτό το διάστημα πρόβλεψης καλείται και **διάστημα ατομικής πρόβλεψης** για το Y_i. Με βάση τις προηγούμενες n παρατηρήσεις $Y_1, Y_2, \ldots, Y_n$, έχουμε υπολογίσει την εκτιμήτρια ελαχίστων τετραγώνων $\widehat{\boldsymbol{\beta}}$. Με βάση αυτές τις εκτιμήτριες, έχουμε ορίσει πρόβλεψη $\widetilde{Y}_{n+1} = \mathbf{X}_{n+1}^{\mathrm{T}} \widehat{\boldsymbol{\beta}}$ και σφάλμα πρόβλεψης $\widetilde{\varepsilon}_{n+1} = Y_{n+1} - \widetilde{Y}_{n+1} = Y_{n+1} - \mathbf{X}_{n+1}^{\mathrm{T}} \widehat{\boldsymbol{\beta}}$.

Το σφάλμα πρόβλεψης $\widetilde{\varepsilon}_{n+1}$ είναι γραμμικός συνδυασμός των Y_{n+1} και $\widehat{\boldsymbol{\beta}}$, το οποίο με τη σειρά του είναι γραμμικός συνδυασμός του $\mathbf{Y}$. Εφόσον οι παρατηρήσεις $Y_1, Y_2, \ldots, Y_n, Y_{n+1}$ είναι ανεξάρτητες και κανονικά κατανεμημένες, συμπεραίνουμε ότι και το $\widetilde{\varepsilon}_{n+1}$ θα είναι κανονικά κατανεμημένο με μέση τιμή και διασπορά που έχουμε υπολογίσει στην πρόταση 2.9, δηλαδή:

$$\widetilde{\varepsilon}_{n+1} \sim N\left(0, \sigma^2_{\widehat{\widetilde{\varepsilon}}_{n+1}}\right), \text{ όπου } \sigma^2_{\widehat{\widetilde{\varepsilon}}_{n+1}} = \sigma^2 \left[1 + \mathbf{X}_{n+1}^{\mathrm{T}} \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1} \mathbf{X}_{n+1}\right] \Rightarrow$$

$$Z = \frac{Y_{n+1} - \mathbf{X}_{n+1}^{\mathrm{T}} \widehat{\boldsymbol{\beta}}}{\sigma_{\widehat{\widetilde{\varepsilon}}_{n+1}}} \sim N(0,1).$$

Εφόσον το $\left(\widehat{\boldsymbol{\beta}}, S^2\right)$ είναι συνάρτηση των $Y_1, Y_2, \ldots, Y_n$, θα είναι ανεξάρτητο από το Y_{n+1}. Όμως, το $\widehat{\boldsymbol{\beta}}$ είναι ανεξάρτητο από το S^2, οπότε τα $\widehat{\boldsymbol{\beta}}$, S^2 και Y_{n+1} είναι αμοιβαία ανεξάρτητα μεταξύ τους. Επομένως, η τυχαία μεταβλητή S^2 είναι ανεξάρτητη από το $\left(\widehat{\boldsymbol{\beta}}, Y_{n+1}\right)$. Συμπεραίνουμε ότι η τυχαία μεταβλητή Z είναι ανεξάρτητη από την τυχαία μεταβλητή $Q = \frac{(n-p-1)S^2}{\sigma^2} \sim \chi^2_{n-p-1}$, οπότε:

$$T = \frac{Z}{\sqrt{\frac{Q}{n-p-1}}} = \frac{Y_{n+1} - \mathbf{X}_{n+1}^{\mathrm{T}} \widehat{\boldsymbol{\beta}}}{\sigma_{\widehat{\widetilde{\varepsilon}}_{n+1}}} \sim t_{n-p-1}, \text{ όπου}$$

$$S^2_{\widetilde{\varepsilon}_{n+1}} = S^2 \left[1 + \mathbf{X}_{n+1}^{\mathrm{T}} \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1} \mathbf{X}_{n+1}\right].$$

Ζητάμε σταθερές $c_1, c_2 \in \mathbb{R}$ τέτοιες, ώστε $P(c_1 \leqslant T \leqslant c_2) = 1 - \alpha$. Όπως και πριν, παίρνουμε ότι $c_1 = -t_{n-p-1;\frac{\alpha}{2}}$ και $c_2 = t_{n-p-1;\frac{\alpha}{2}}$. Επομένως, το διάστημα εμπιστοσύνης ίσων ουρών δίνεται από τη σχέση:

$$c_1 \leqslant \frac{Y_{n+1} - \widetilde{Y}_{n+1}}{S_{\widetilde{\varepsilon}_{n+1}}} \leqslant c_2 \Leftrightarrow \widetilde{Y}_{n+1} + c_1 \cdot S_{\widetilde{\varepsilon}_{n+1}} \leqslant Y_{n+1} \leqslant \widetilde{Y}_{n+1} + c_2 \cdot S_{\widetilde{\varepsilon}_{n+1}}.$$

Τελικά, παίρνουμε $I_{1-\alpha}\left(Y_{n+1}\right) = \left[\widetilde{Y}_{n+1} - t_{n-p-1;\frac{\alpha}{2}} \cdot S_{\widetilde{\varepsilon}_{n+1}}, \widetilde{Y}_{n+1} + t_{n-p-1;\frac{\alpha}{2}} \cdot S_{\widetilde{\varepsilon}_{n+1}}\right]$.

Παρατήρηση 2.4. Θέλουμε να συγκρίνουμε τα διαστήματα μέσης και ατομικής πρόβλεψης για το Y_{n+1}. Παρατηρούμε ότι και τα δύο διαστήματα πρόβλεψης είναι εστιασμένα γύρω από τον κοινό μέσο $\widehat{Y}_{n+1} = \widetilde{Y}_{n+1} = \mathbf{X}_{n+1}^{\mathrm{T}} \widehat{\boldsymbol{\beta}}$. Όμως, βλέπουμε ότι

$S^2_{\widehat{\varepsilon}_{n+1}} = S^2 + S^2_{\widehat{Y}_{n+1}}$. Το διάστημα μέσης πρόβλεψης για το Y_{n+1} έχει μήκος:

$$\lambda\left(I_{1-\alpha}\left(E(Y_{n+1})\right)\right) = 2t_{n-p-1;\frac{\alpha}{2}} S_{\widehat{Y}_{n+1}}.$$

Το αντίστοιχο διάστημα ατομικής πρόβλεψης για το Y_{n+1} έχει μήκος:

$$\begin{aligned}\lambda\left(I_{1-\alpha}\left(Y_{n+1}\right)\right) &= 2t_{n-p-1;\frac{\alpha}{2}} S_{\widetilde{\varepsilon}_{n+1}} \\ &= 2t_{n-p-1;\frac{\alpha}{2}} \sqrt{S^2 + S^2_{\widehat{Y}_{n+1}}} > 2t_{n-p-1;\frac{\alpha}{2}} S_{\widehat{Y}_{n+1}} = \lambda\left(I_{1-\alpha}\left(E(Y_{n+1})\right)\right).\end{aligned}$$

Επομένως, το μήκος του διαστήματος ατομικής πρόβλεψης για το Y_{n+1} είναι πάντα μεγαλύτερο από αυτό του αντίστοιχου διαστήματος μέσης πρόβλεψης. Αυτή η διαφορά στα μήκη των διαστημάτων οφείλεται στις διαφορετικές πηγές αβεβαιότητας που λαμβάνουν υπόψη. Το διάστημα μέσης πρόβλεψης για το Y_{n+1} λαμβάνει υπόψη του μόνο την αβεβαιότητα για τη μέση τιμή $E(Y_{n+1})$. Από την άλλη μεριά, το διάστημα ατομικής πρόβλεψης για το Y_{n+1} συνυπολογίζει και την αβεβαιότητα για την ίδια την παρατήρηση Y_{n+1}, η οποία είναι άγνωστη.

2.10 Πολυσυγγραμμικότητα

Πολυσυγγραμμικότητα είναι το φαινόμενο που εμφανίζεται στη γραμμική παλινδρόμηση όταν υπάρχει ισχυρή γραμμική εξάρτηση ανάμεσα σε ένα σύνολο επεξηγηματικών μεταβλητών.

Έστω ότι έχουμε το πολλαπλό γραμμικό μοντέλο $Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \varepsilon_i$ με κανονικά κατανεμημένα τυχαία σφάλματα. Αν οι επεξηγηματικές μεταβλητές X_1 και X_2 είναι τέλεια γραμμικά εξαρτημένες, τότε ο δειγματικός συντελεστής συσχέτισης του Pearson μεταξύ τους είναι κατά απόλυτη τιμή ίσος με τη μονάδα και ισχύει ότι $X_{2,i} = \lambda_0 + \lambda_1 X_{1,i}$ για κάποιες σταθερές $\lambda_0, \lambda_1 \in \mathbb{R}$. Σε αυτήν την περίπτωση, οι στήλες του πίνακα σχεδιασμού $\mathbf{X} \in \mathbb{R}^{n\times 3}$ είναι γραμμικά εξαρτημένος, οπότε ο πίνακας δεν είναι πλήρους τάξης. Κατά συνέπεια, ο πίνακας $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ είναι μη-αντιστρέψιμος και η εκτιμήτρια ελαχίστων τετραγώνων $\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y} \in \mathbb{R}^3$ δεν υπάρχει. Αντικαθιστώντας στο γραμμικό μοντέλο, βλέπουμε ότι:

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2\left(\lambda_0 + \lambda_1 X_{1,i}\right) + \varepsilon_i = \underbrace{\beta_0 + \lambda_0\beta_2}_{\gamma_0} + \underbrace{\left(\beta_1 + \lambda_1\beta_2\right)}_{\gamma_1} X_{1,i} + \varepsilon_i.$$

Επομένως, θέτοντας $\gamma_0 = \beta_0 + \lambda_0\beta_2$, $\gamma_1 = \beta_1 + \lambda_1\beta_2$ και παραλείποντας την επεξηγηματική μεταβλητή X_2, παίρνουμε ένα καινούργιο γραμμικό μοντέλο χωρίς το φαινόμενο της πολυσυγγραμμικότητας.

Στην πιο ρεαλιστική περίπτωση, όπου υπάρχει ισχυρή, αλλά όχι τέλεια, γραμ-

μική εξάρτηση ανάμεσα σε δύο ή περισσότερες επεξηγηματικές μεταβλητές, ο πίνακας σχεδιασμού είναι πλήρους τάξης, αλλά λόγω αριθμητικής υπολογιστή, ο υπολογισμός του αντιστρόφου του πίνακα $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ είναι αριθμητικά ασταθής. Με άλλα λόγια, μικρές αλλαγές στα δεδομένα, μπορεί να οδηγήσουν σε τεράστιες αλλαγές στις εκτιμήσεις $\widehat{\boldsymbol{\beta}}$.

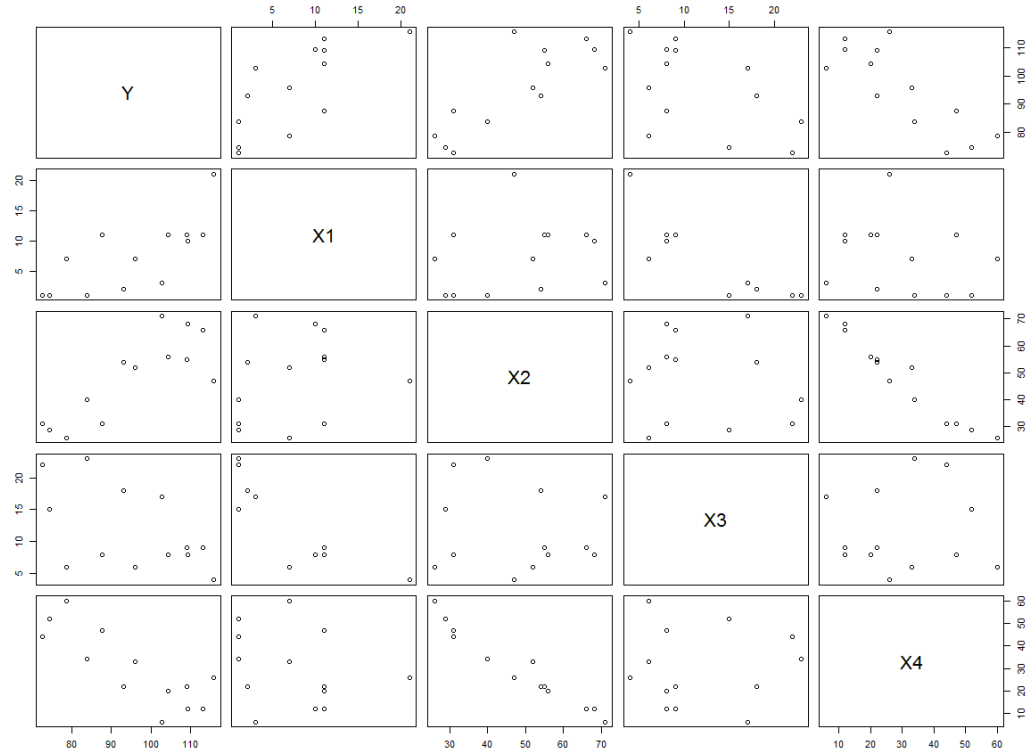
Επιπλέον, αυτή η αριθμητική αστάθεια στον υπολογισμό του αντιστρόφου, μπορεί να οδηγήσει σε μεγάλα τυπικά σφάλματα $S_{\widehat{\beta}_i}$ για τις εκτιμήτριες $\widehat{\beta}_i$ και λανθασμένα συμπεράσματα στους ελέγχους στατιστικής σημαντικότητας, εφόσον γνωρίζουμε ότι $S^2_{\widehat{\beta}_i} = S^2 \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}_{i,i}$. Όπως είδαμε παραπάνω, μία άλλη συνέπεια της πολυσυγγραμμικότητας είναι η εισαγωγή περιττών επεξηγηματικών μεταβλητών στο γραμμικό μοντέλο, οι οποίες θα μπορούσαν κάλλιστα να παραλειφθούν, χωρίς να μειωθεί σημαντικά η προβλεπτική ικανότητα του γραμμικού μοντέλου.

Τέλος, επηρεάζονται σημαντικά οι ερμηνείες των συντελεστών β_i, οι οποίοι, όπως είδαμε, αντιπροσωπεύουν τη μεταβολή στην αναμενόμενη τιμή της αποκριτικής μεταβλητής για αύξηση της αντίστοιχης επεξηγηματικής μεταβλητής X_i κατά μία μονάδα, κρατώντας όλες τις υπόλοιπες επεξηγηματικές μεταβλητές σταθερές. Όταν, όμως, υπάρχει τέτοια ισχυρή γραμμική εξάρτηση μεταξύ επεξηγηματικών μεταβλητών, είναι αδύνατον να επηρεάσεις τη μία, κρατώντας τις υπόλοιπες σταθερές, αφού θα μεταβάλλονται πάντα όλες μαζί.

Προκειμένου να ανιχνεύσουμε το φαινόμενο της πολυσυγγραμμικότητας μεταξύ των επεξηγηματικών μεταβλητών μας, ένα πρώτο βήμα που πρέπει πάντα να ακολουθούμε είναι ο υπολογισμός του δειγματικού πίνακα συσχέτισης των δεδομένων μας και η γραφική αναπαράσταση των δεδομένων μέσω διαγραμμάτων διασποράς όλων των μεταβλητών ανά δύο.

Έστω ότι έχουμε 4 επεξηγηματικές μεταβλητές X_1, X_2, X_3, X_4. Σχεδιάζουμε αρχικά τον πίνακα διαγραμμάτων διασποράς ο οποίος φαίνεται στο σχήμα 2.1. Σε κάθε κελί του πίνακα είναι σχεδιασμένο το διάγραμμα διασποράς με τη μεταβλητή που αναγράφεται στη στήλη του κελιού να βρίσκεται στον άξονα των x και τη μεταβλητή που αναγράφεται στη γραμμή του κελιού να βρίσκεται στον άξονα των y. Τα διαγράμματα που είναι σχεδιασμένα στα κελιά κάτω από τη διαγώνιο ταυτίζονται, προφανώς, με τα αντίστοιχα διαγράμματα που είναι σχεδιασμένα στα κελιά πάνω από τη διαγώνιο, οπότε αρκούμαστε να ελέγξουμε τα τελευταία.

Στην πρώτη γραμμή του πίνακα, βλέπουμε τα διαγράμματα διασποράς της αποκριτικής μεταβλητής Y με καθεμία από τις επεξηγηματικές μεταβλητές ξεχωριστά. Βλέπουμε ότι η αποκριτική μεταβλητή έχει ισχυρή θετική συσχέτιση με τη X_2, παρόμοια ισχυρή αρνητική συσχέτιση με τη X_4, λιγότερο ισχυρή θετική συσχέτιση με τη X_1 και, τέλος, ασθενή αρνητική συσχέτιση με τη X_3.



Σχημα 2.1: Πίνακας Διαγραμμάτων Διασποράς (Scatter Plot Matrix)

Στα υπόλοιπα 6 κελιά που βρίσκονται πάνω από τη διαγώνιο του πίνακα, βλέπουμε σχεδιασμένα τα διαγράμματα διασποράς όλων των επεξηγηματικών μεταβλητών ανά δύο. Ανάμεσα σε αυτά τα 6, μας ανησυχεί ιδιαιτέρως το διάγραμμα διασποράς της X_2 με τη X_4, από το οποίο φαίνεται ότι υπάρχει σχεδόν τέλεια αρνητική συσχέτιση μεταξύ τους. Επομένως, αν δεν είμαστε προσεκτικοί, θα οδηγηθούμε σίγουρα στο φαινόμενο της πολυσυγγραμμικότητας. Μας ανησυχεί σε μικρότερο βαθμό και το διάγραμμα διασποράς της X_1 με τη X_3, το οποίο δείχνει μία αρκετά ισχυρή αρνητική συσχέτιση.

$$\mathbf{R} = \begin{array}{c} \\ Y \\ X_1 \\ X_2 \\ X_3 \\ X_4 \end{array} \begin{array}{c} \begin{array}{ccccc} Y & X_1 & X_2 & X_3 & X_4 \end{array} \\ \begin{bmatrix} 1.00 & 0.73 & 0.82 & -0.53 & -0.82 \\ 0.73 & 1.00 & 0.23 & -0.82 & -0.25 \\ 0.82 & 0.23 & 1.00 & -0.14 & -0.97 \\ -0.53 & -0.82 & -0.14 & 1.00 & 0.03 \\ -0.82 & -0.25 & -0.97 & 0.03 & 1.00 \end{bmatrix} \end{array}.$$

Στη συνέχεια, υπολογίζουμε τον δειγματικό πίνακα συσχέτισης των δεδομένων, προκειμένου να επαληθεύσουμε και ποσοτικά τις συσχετίσεις που είδαμε να απεικονίζονται στα διαγράμματα διασποράς. Βλέπουμε ότι και ο δειγματι-

κός πίνακας συσχέτισης είναι, προφανώς, συμμετρικός και οι συσχετίσεις που εμφανίζονται στη διαγώνιο είναι σαφώς ίσες με τη μονάδα.

Επαληθεύουμε ότι τα μεγέθη των συσχετίσεων της αποκριτικής μεταβλητής Y με καθεμία από τις επεξηγηματικές μεταβλητές, τα οποία εμφανίζονται στην πρώτη γραμμή του δειγματικού πίνακα συσχέτισης, συμφωνούν απόλυτα με όσα παρατηρήσαμε μέσω των διαγραμμάτων διασποράς. Επαληθεύουμε, επίσης, ότι η δειγματική συσχέτιση $r_{X_2X_4} = -0.97$ είναι ανησυχητικά κοντά στο -1, ενώ η δειγματική συσχέτιση $r_{X_1X_3} = -0.82$ είναι και αυτή πολύ ισχυρή.

Γενικότερα, δειγματικές συσχετίσεις κατά απόλυτη τιμή μεγαλύτερες του **0.4** μεταξύ επεξηγηματικών μεταβλητών μπορούν να ερμηνευθούν ως ενδείξεις πολυσυγγραμμικότητας. Όπως συζητήσαμε, όμως, το φαινόμενο της πολυσυγγραμμικότητας εμφανίζεται και όταν υπάρχει ισχυρή γραμμική εξάρτηση μεταξύ μίας ομάδας περισσότερων από δύο επεξηγηματικών μεταβλητών. Στην περίπτωση αυτή, η εξάρτηση μπορεί να μη φαίνεται καθόλου μέσω των διαγραμμάτων διασποράς και του δειγματικού πίνακα συσχετίσεων, αφού ελέγχουν τη συσχέτιση των μεταβλητών μόνο κατά ζεύγη.

Για τον λόγο αυτό, θυμόμαστε τα variance inflation factors που ορίσαμε στην παρατήρηση 2.3. Προφανώς, όσο πιο κοντά στη μονάδα είναι το R_j^2 για κάποια επεξηγηματική μεταβλητή X_j, τόσο πιο μεγαλύτερο της μονάδας είναι το αντίστοιχο VIF_j και τόσο πιο ισχυρά εξαρτημένη είναι η X_j από τις υπόλοιπες επεξηγηματικές μεταβλητές. Επομένως, αν $\text{VIF}_j > 5$ για κάποια επεξηγηματική μεταβλητή X_j, τότε αυτό θεωρούμε πως είναι **ένδειξη πολυσυγγραμμικότητας**.

Όσον αφορά την αντιμετώπιση του φαινομένου, στην περίπτωση όπου υπάρχει μικρό πλήθος εξαρτημένων επεξηγηματικών μεταβλητών, θα μπορούσαμε απλά να αφαιρέσουμε κάποιες από αυτές από το γραμμικό μοντέλο. Αν, όμως, το πλήθος των εξαρτημένων επεξηγηματικών μεταβλητών είναι μεγάλο, τότε θα μπορούσαμε να εφαρμόσουμε κάποια μέθοδο μείωσης διάστασης, όπως η ανάλυση κυρίων συνιστωσών, προκειμένου να συνοψίσουμε την πληροφορία που περιέχουν οι επεξηγηματικές μεταβλητές σε ένα μικρό πλήθος νέων μεταβλητών.

2.11 Κριτήρια Επιλογής Μοντέλου

Έστω ότι έχουμε ένα σύνολο από k υποψήφιες επεξηγηματικές μεταβλητές για μία μεταβλητή ενδιαφέροντος Y και θέλουμε να επιλέξουμε μόνο p από αυτές για να κατασκευάσουμε ένα γραμμικό μοντέλο. Το πλήθος όλων των πιθανών γραμμικών μοντέλων που μπορούμε να κατασκευάσουμε με βάση αυτές τις k υποψήφιες επεξηγηματικές μεταβλητές είναι $2^k - 1$, οπότε καταλαβαίνουμε ότι αυτή η σύγκριση θα ήταν αδύνατο να γίνει στο χαρτί ακόμα και για μικρό k.

Έχουμε ήδη δει τον προσαρμοσμένο συντελεστή προσδιορισμού, ο οποίος χρησιμοποιείται για να επιλέξουμε το μοντέλο που εξηγεί ένα μεγάλο ποσοστό από τη συνολική μεταβλητότητα των δεδομένων με όσο το δυνατόν μικρότερο πλήθος επεξηγηματικών μεταβλητών. Εκτιμώντας όλα τα πιθανά $2^k - 1$ γραμμικά μοντέλα και υπολογίζοντας τον προσαρμοσμένο συντελεστή προσδιορισμού για καθένα από αυτά, θα επιλέγαμε εκείνο που επιτυγχάνει τη μέγιστη δυνατή τιμή.

Κάποια γενικότερα κριτήρια επιλογής μοντέλου, τα οποία βρίσκουν εφαρμογή και εκτός τους πλαισίου των γραμμικών μοντέλων, είναι τα λεγόμενα **κριτήρια πληροφορίας**, τα οποία βασίζονται στη μέγιστη πιθανοφάνεια ενός μοντέλου. Με τον όρο **μέγιστη πιθανοφάνεια** εννοούμε την τιμή που επιτυγχάνει η συνάρτηση πιθανοφάνειας $L(\boldsymbol{\beta}, \sigma^2|\mathbf{y})$ για $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}$ και $\sigma^2 = \widehat{\sigma}^2$, δηλαδή η πιθανοφάνεια υπολογισμένη στο σημείο της εκτίμησης μέγιστης πιθανοφάνειας.

Έστω $\mathbf{X}_p \in \mathbb{R}^{n\times(p+1)}$ ο πίνακας σχεδιασμού που αντιστοιχεί στις επεξηγηματικές μεταβλητές $X_1, X_2, \ldots, X_p$. Έχουμε υπολογίσει τη συνάρτηση πιθανοφάνειας ενός κανονικού γραμμικού μοντέλου με p επεξηγηματικές μεταβλητές ως:

$$L\left(\boldsymbol{\beta}_p, \sigma_p^2|\mathbf{y}\right) = \left(2\pi\sigma_p^2\right)^{-\frac{n}{2}} \exp\left\{-\frac{\|\mathbf{y} - \mathbf{X}_p\boldsymbol{\beta}_p\|^2}{2\sigma_p^2}\right\} \Rightarrow$$

$$\ell\left(\boldsymbol{\beta}_p, \sigma_p^2|\mathbf{y}\right) = \log L\left(\boldsymbol{\beta}_p, \sigma_p^2|\mathbf{y}\right) = -\frac{n}{2}\log\left(2\pi\sigma_p^2\right) - \frac{\|\mathbf{y} - \mathbf{X}_p\boldsymbol{\beta}_p\|^2}{2\sigma_p^2}.$$

Επιπλέον, έχουμε υπολογίσει τις εκτιμήσεις μέγιστης πιθανοφάνειας:

$$\widehat{\boldsymbol{\beta}}_p = \left(\mathbf{X}_p^{\mathrm{T}}\mathbf{X}_p\right)^{-1}\mathbf{X}_p^{\mathrm{T}}\mathbf{y}, \quad \widehat{\sigma}_p^2 = \frac{\|\mathbf{y} - \mathbf{X}_p\widehat{\boldsymbol{\beta}}_p\|^2}{n} = \frac{\mathrm{SSE}_p}{n}.$$

Έστω $\widehat{\ell}_p = \ell\left(\widehat{\boldsymbol{\beta}}_p, \widehat{\sigma}_p^2|\mathbf{y}\right)$ η μέγιστη λογαριθμοπιθανοφάνεια του γραμμικού μοντέλου με p επεξηγηματικές μεταβλητές. Τότε,

$$\widehat{\ell}_p = -\frac{n}{2}\log\left(2\pi\widehat{\sigma}_p^2\right) - \frac{\|\mathbf{y} - \mathbf{X}_p\widehat{\boldsymbol{\beta}}_p\|^2}{2\widehat{\sigma}_p^2} = -\frac{n}{2}\log\frac{2\pi\mathrm{SSE}_p}{n} - \frac{n}{2}.$$

Προφανώς, όσο μεγαλύτερη πιθανοφάνεια έχει ένα μοντέλο, τόσο πιο πιθανό είναι να έχουμε παρατηρήσει τα δεδομένα που παρατηρήσαμε από αυτό το μοντέλο, οπότε αυτό είναι και το μοντέλο που θέλουμε να επιλέξουμε. Όμως, όπως ο συντελεστής προσδιορισμού, έτσι και η μέγιστη πιθανοφάνεια του γραμμικού μοντέλου αυξάνεται συνεχώς όσο προστίθενται καινούργιες επεξηγηματικές μεταβλητές μέσα στο μοντέλο. Για τον λόγο αυτό, τα κριτήρια πληροφορίας συνυπολογίζουν και το πλήθος των επεξηγηματικών του γραμμικού μοντέλου, ώστε να μας βοηθήσουν να επιλέξουμε ένα μοντέλο που έχει μεγάλη πιθανοφάνεια με όσο το δυνατόν λιγότερο πλήθος επεξηγηματικών μεταβλητών.

Το **κριτήριο πληροφορίας του Akaike** (AIC - Akaike Information Criterion) ορίζεται ως εξής:

$$\mathrm{AIC}_p = -2\widehat{\ell}_p + 2(p+2) = n\log\frac{2\pi\mathrm{SSE}_p}{n} + n + 2(p+2).$$

Ο όρος $p+2$ αντιπροσωπεύει το πλήθος των αγνώστων παραμέτρων που εκτιμήσαμε στο μοντέλο με p επεξηγηματικές μεταβλητές, δηλαδή $p+1$ συντελεστές παλινδρόμησης και τη διασπορά σ^2. Σκοπός μας είναι να μεγιστοποιήσουμε την τιμή $\widehat{\ell}_p$, προσπαθώντας παράλληλα να ελαχιστοποιήσουμε τον όρο $p+2$, οπότε καλύτερο μοντέλο είναι αυτό που επιτυγχάνει τη μικρότερη δυνατή τιμή AIC_p.

Το **Μπεϋζιανό κριτήριο πληροφορίας** (BIC - Bayesian Information Criterion), το οποίο είναι γνωστό και ως κριτήριο πληροφορίας του Schwarz, ορίζεται ως:

$$\mathrm{BIC}_p = -2\widehat{\ell}_p + (p+2)\log n = n\log\frac{2\pi\mathrm{SSE}_p}{n} + n + (p+2)\log n.$$

Βλέπουμε ότι η μόνη διαφορά μεταξύ αυτών των δύο κριτηρίων πληροφορίας είναι ο συντελεστής του $p+2$. Με άλλα λόγια, τα δύο κριτήρια θέτουν διαφορετική ποινή στο πλήθος των παραμέτρων προς εκτίμηση. Μάλιστα, παρατηρούμε ότι $\log n > 2 \Leftrightarrow n > e^2 \approx 7.39$, δηλαδή η ποινή που θέτει το BIC γίνεται πιο αυστηρή από αυτή που θέτει το AIC, καθώς αυξάνεται το μέγεθος του δείγματος. Για τον λόγο αυτό, το BIC τείνει να επιλέγει μοντέλα με μικρότερο πλήθος παραμέτρων σε σύγκριση με το AIC. Προφανώς, καλύτερο μοντέλο είναι πάλι αυτό που επιτυγχάνει τη μικρότερη δυνατή τιμή BIC_p.

Ειδικότερα, όταν το μέγεθος του δείγματος είναι μικρό, το AIC τείνει να επιλέγει μοντέλα με υπερβολικά μεγάλο πλήθος αγνώστων παραμέτρων. Για τον λόγο αυτό έχει προταθεί το διορθωμένο κριτήριο πληροφορίας του Akaike (AICc - Corrected Akaike Information Criterion), το οποίο στην περίπτωση του πολλαπλού γραμμικού μοντέλου παίρνει τη μορφή:

$$\mathrm{AICc}_p = \mathrm{AIC}_p + \frac{2(p+2)(p+3)}{n-p-3}.$$

Ένα άλλο κριτήριο που χρησιμοποιείται στην περίπτωση της γραμμικής παλινδρόμησης είναι το **$\mathrm{C_p}$ του Mallows**, το οποίο ορίζεται ως:

$$C_p = \frac{\mathrm{SSE}_p}{S^2} - n + 2(p+1), \text{ όπου } S^2 = \frac{\mathrm{SSE}_k}{n-k-1},$$

δηλαδή S^2 είναι η αμερόληπτη εκτιμήτρια της διασποράς που προκύπτει από το πολλαπλό γραμμικό μοντέλο με όλες τις k δυνατές επεξηγηματικές μεταβλητές. Όταν ένα γραμμικό μοντέλο περιέχει όλες εκείνες τις μεταβλητές που είναι απα-

ραίτητες για την πρόβλεψη της αποκριτικής μεταβλητής Y, τότε μπορεί να δειχθεί ότι $E(C_p) \approx p+1$. Για το μοντέλο που περιέχει όλες τις k δυνατές επεξηγηματικές μεταβλητές ισχύει πάντα ότι:

$$C_k = \frac{\text{SSE}_k}{S^2} - n + 2(k+1) = n - k - 1 - n + 2(k+1) = k + 1.$$

Επομένως, εξετάζουμε γραμμικά μοντέλα με μικρότερο πλήθος επεξηγηματικών μεταβλητών και επιλέγουμε ως καλύτερο γραμμικό μοντέλο εκείνο που έχει τη μικρότερη τιμή του p για την οποία $C_p \approx p + 1$.

Όλα αυτά τα κριτήρια που περιγράψαμε αξιολογούν με διαφορετικούς τρόπους την ποιότητα ενός οποιουδήποτε γραμμικού μοντέλου, οπότε οδηγούν γενικά στην επιλογή διαφορετικών μοντέλων. Η απόφαση για το ποιο μοντέλο είναι τελικά βέλτιστο ανάμεσα στα διαφορετικά μοντέλα που επέλεξαν τα παραπάνω κριτήρια εξαρτάται από πολλούς παράγοντες.

Λόγω του υπερβολικά μεγάλου αριθμού γραμμικών μοντέλων που πρέπει να συγκριθούν, ώστε να καταλήξουμε στο καλύτερο μοντέλο με βάση κάποιο επιλεγμένο κριτήριο επιλογής, αυτή η διαδικασία είναι πολλές φορές χρονοβόρα ακόμα και για έναν υπολογιστή. Για τον λόγο αυτό, κάνουμε συχνά χρήση κάποιων απλούστερων μεθόδων **βηματικής παλινδρόμησης**. Ξεκινώντας από κάποιο αρχικό γραμμικό μοντέλο, αυτές οι μέθοδοι προσθέτουν ή αφαιρούν σε κάθε βήμα κάποια επεξηγηματική μεταβλητή με βάση κάποιο προεπιλεγμένο κριτήριο, μέχρι να καταλήξουν σε κάποιο γραμμικό μοντέλο το οποίο δε βελτιώνεται από καμία προσθήκη ή διαγραφή επεξηγηματικής μεταβλητής.

Η **μέθοδος backward** χρησιμοποιεί ως σημείο εκκίνησης το γραμμικό μοντέλο το οποίο περιέχει όλες τις k δυνατές επεξηγηματικές μεταβλητές. Ας υποθέσουμε ότι χρησιμοποιούμε ως κριτήριο επιλογής το AIC. Σε κάθε βήμα της μεθόδου, δοκιμάζουμε να αφαιρέσουμε καθεμία από τις επεξηγηματικές μεταβλητές του γραμμικού μοντέλου ξεχωριστά και υπολογίζουμε το AIC των γραμμικών μοντέλων που δημιουργούνται. Επιλέγουμε να αφαιρέσουμε εκείνη την επεξηγηματική μεταβλητή που θα οδηγήσει στη μεγαλύτερη δυνατή μείωση της τιμής του AIC. Αν, σε κάποιο βήμα της μεθόδου, η διαγραφή οποιασδήποτε επεξηγηματικής μεταβλητής οδηγεί σε αύξηση του AIC σε σύγκριση με το τρέχον μοντέλο, τότε η μέθοδος τερματίζεται και επιλέγουμε το τρέχον μοντέλο ως βέλτιστο.

Ως εναλλακτικό κριτήριο επιλογής θα μπορούσαμε να χρησιμοποιήσουμε τα p-value των ελέγχων στατιστικής σημαντικότητας των συντελεστών παλινδρόμησης εκτός του σταθερού όρου και κάποιο προεπιλεγμένο ε.σ.σ. α. Γνωρίζουμε ότι όσο μεγαλύτερο p-value έχει μία παράμετρος στον έλεγχο στατιστικής σημαντικότητας τόσο λιγότερο στατιστικά σημαντική είναι. Σε κάθε βήμα της μεθόδου,

πραγματοποιούμε τους επιμέρους ελέγχους στατιστικής σημαντικότητας όλων των συντελεστών της παλινδρόμησης εκτός του σταθερού όρου. Επιλέγουμε να αφαιρέσουμε την επεξηγηματική μεταβλητή της οποίας ο συντελεστής έχει το μεγαλύτερο p-value. Αν, σε κάποιο βήμα της μεθόδου, τα p-value όλων των ελέγχων στατιστικής σημαντικότητας είναι μικρότερα από α, τότε δεν αφαιρούμε καμία επεξηγηματική μεταβλητή, η μέθοδος τερματίζεται και επιλέγουμε το τρέχον μοντέλο ως βέλτιστο.

Η **μέθοδος forward** χρησιμοποιεί ως σημείο εκκίνησης το μοντέλο το οποίο δεν περιέχει καμία επεξηγηματική μεταβλητή. Ας υποθέσουμε ότι χρησιμοποιούμε ως κριτήριο επιλογής το AIC. Σε κάθε βήμα της μεθόδου, δοκιμάζουμε να προσθέσουμε καθεμία από τις υποψήφιες επεξηγηματικές μεταβλητές ξεχωριστά και υπολογίζουμε το AIC των γραμμικών μοντέλων που δημιουργούνται. Επιλέγουμε να προσθέσουμε εκείνη την επεξηγηματική μεταβλητή που θα οδηγήσει στη μεγαλύτερη δυνατή μείωση της τιμής του AIC. Αν, σε κάποιο βήμα της μεθόδου, η προσθήκη οποιασδήποτε επεξηγηματικής μεταβλητής οδηγεί σε αύξηση του AIC σε σύγκριση με το τρέχον μοντέλο, τότε η μέθοδος τερματίζεται και επιλέγουμε το τρέχον μοντέλο ως βέλτιστο.

Ως εναλλακτικό κριτήριο επιλογής θα μπορούσαμε πάλι να χρησιμοποιήσουμε τα p-value των ελέγχων στατιστικής σημαντικότητας των συντελεστών παλινδρόμησης. Γνωρίζουμε ότι όσο μικρότερο p-value έχει μία παράμετρος στον έλεγχο στατιστικής σημαντικότητας τόσο περισσότερο στατιστικά σημαντική είναι. Σε κάθε βήμα της μεθόδου, δοκιμάζουμε να προσθέσουμε καθεμία από τις υποψήφιες επεξηγηματικές μεταβλητές ξεχωριστά και πραγματοποιούμε τον έλεγχο στατιστικής σημαντικότητας του συντελεστή παλινδρόμησης που προστέθηκε σε καθένα από τα γραμμικά μοντέλα που δημιουργούνται. Επιλέγουμε να προσθέσουμε την επεξηγηματική μεταβλητή της οποίας ο συντελεστής έχει το μικρότερο p-value. Αν, σε κάποιο βήμα της μεθόδου, τα p-value όλων των ελέγχων στατιστικής σημαντικότητας είναι μεγαλύτερα από α, τότε δεν προσθέτουμε καμία επεξηγηματική μεταβλητή, η μέθοδος τερματίζεται και επιλέγουμε το τρέχον μοντέλο ως βέλτιστο.

Η **μέθοδος stepwise** είναι βελτίωση της μεθόδου forward. Σε κάθε βήμα, εκτός από την προσθήκη κάποιας υποψήφιας επεξηγηματικής μεταβλητής, ελέγχουμε και την πιθανή διαγραφή κάποιας από τις επεξηγηματικές μεταβλητές που έχουμε ήδη προσθέσει στο γραμμικό μοντέλο. Η επιλογή ανάμεσα στην προσθήκη ή την αφαίρεση επεξηγηματικής μεταβλητής γίνεται πάλι με στόχο τη βελτιστοποίηση ενός προεπιλεγμένου κριτηρίου επιλογής μοντέλου.

2.12 Διαγνωστικοί Έλεγχοι Γραμμικής Παλινδρόμησης

Έχοντας επιλέξει και εκτιμήσει ένα κατάλληλο κανονικό πολλαπλό γραμμικό μοντέλο για την Y με κάποια από τις μεθόδους που περιγράψαμε στην προηγούμενη παράγραφο, η δουλειά μας δεν έχει τελειώσει ακόμα. Ο βασικός λόγος είναι οι 4 υποθέσεις για τα τυχαία σφάλματα ε_i πάνω στις οποίες έχουμε στηριχτεί για να κατασκευάσουμε το εν λόγω γραμμικό μοντέλο, δηλαδή ότι είναι κανονικά κατανεμημένα, με μέση τιμή 0, κοινή διασπορά σ^2 και ανεξάρτητα. Αν, έχοντας εκτιμήσει το γραμμικό μοντέλο μας, διαπιστώσουμε ότι κάποια από αυτές τις υποθέσεις δεν ευσταθεί, τότε αυτό σημαίνει, προφανώς, ότι όλο το γραμμικό μοντέλο το οποίο έχουμε κατασκευάσει δεν είναι έγκυρο και πρέπει να επεξεργαστεί με διαφορετικό τρόπο.

Εφόσον τα τυχαία σφάλματα ε_i είναι μη-παρατηρήσιμα και μη-υπολογίσιμα, τις 4 αυτές υποθέσεις τις επαληθεύουμε μέσω των εκτιμημένων σφαλμάτων $\widehat{\varepsilon}_i$. Επειδή τα κατάλοιπα $\widehat{\varepsilon}_i$ δεν έχουν κοινή διασπορά, όπως έχουμε δείξει, δηλαδή δεν είναι ισόνομα, χρησιμοποιούμε κάποιες τυποποιημένες εκδοχές τους για την επικύρωση των υποθέσεων της γραμμικής παλινδρόμησης. Συγκεκριμένα, γνωρίζουμε ότι:

$$\mathrm{Var}\left(\widehat{Y}_i\right) = \mathrm{Cov}\left(Y_i, \widehat{Y}_i\right) = \sigma^2 \mathbf{P}_{i,i} \Rightarrow$$

$$\mathrm{Var}\left(\widehat{\varepsilon}_i\right) = \mathrm{Var}\left(Y_i\right) + \mathrm{Var}\left(\widehat{Y}_i\right) - 2\mathrm{Cov}\left(Y_i, \widehat{Y}_i\right) = \mathrm{Var}\left(Y_i\right) - \mathrm{Var}\left(\widehat{Y}_i\right) = \sigma^2\left(1 - \mathbf{P}_{i,i}\right),$$

οπότε $\widehat{\varepsilon}_i = Y_i - \widehat{Y}_i \sim N\left(0, \sigma^2\left(1 - \mathbf{P}_{i,i}\right)\right)$, όπου $\mathbf{P}_{i,i} = \mathbf{X}_i^{\mathrm{T}}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}_i$ το i-οστό διαγώνιο στοιχείο του πίνακα $\mathbf{P} = \mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}$ και $\mathbf{X}_i = (1, X_{1,i}, X_{2,i}, \ldots, X_{p,i})^{\mathrm{T}}$. Επομένως,

$$\frac{\widehat{\varepsilon}_i}{\sigma\sqrt{1 - \mathbf{P}_{i,i}}} \sim N(0,1).$$

Αντικαθιστώντας την άγνωστη διασπορά σ^2 από την αμερόληπτη εκτιμήτρια S^2, παίρνουμε τα **εσωτερικά τυποποιημένα κατάλοιπα** (internally studentised residuals):

$$t_i = \frac{\widehat{\varepsilon}_i}{S\sqrt{1 - \mathbf{P}_{i,i}}} = \frac{Y_i - \widehat{Y}_i}{S\sqrt{1 - \mathbf{P}_{i,i}}}.$$

Δυστυχώς, τα εσωτερικά τυποποιημένα κατάλοιπα δεν ακολουθούν κάποια γνωστή κατανομή, καθώς το S^2 δεν είναι ανεξάρτητο από την παρατήρηση Y_i. Τα κατάλοιπα δείξαμε ότι έχουν πάντα μέση τιμή 0, οπότε αυτή είναι η μόνη υπόθεση που δε χρειάζεται να ελέγξουμε.

Την υπόθεση ότι τα τυχαία σφάλματα είναι κανονικά κατανεμημένα μπορούμε να την ελέγξουμε κάνοντας χρήση διαφόρων ελέγχων κανονικότητας, όπως ο **έλεγχος Shapiro - Wilk**, ο οποίος έχει δειχθεί ότι έχει τη μεγαλύτερη ισχύ ανάμεσα στους γνωστούς ελέγχους κανονικότητας. Με βάση ένα δείγμα $V_1, V_2, \ldots, V_n$, οι

έλεγχοι κανονικότητας λαμβάνουν απόφαση για τις υποθέσεις:

$$\begin{cases} H_0 : V_1, V_2, \ldots, V_n \text{ τυχαίο δείγμα από την κανονική κατανομή vs.} \\ H_1 : V_1, V_2, \ldots, V_n \text{ όχι τυχαίο δείγμα από την κανονική κατανομή.} \end{cases}$$

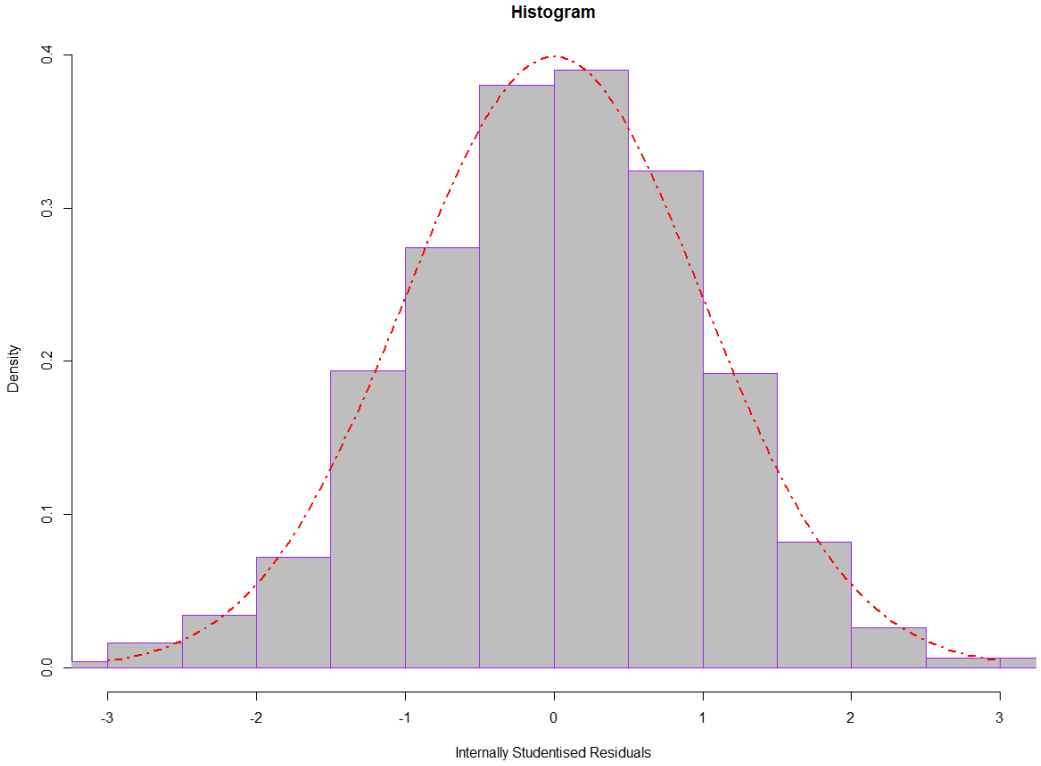
Όλα τα στατιστικά λογισμικά, όπως η R, έχουν ενσωματωμένη τη δυνατότητα πραγματοποίησης του έλεγχο Shapiro - Wilk για δεδομένο δείγμα παρατηρήσεων και δίνουν ως αποτέλεσμα το p-value του ελέγχου. Εφαρμόζουμε τον έλεγχου Shapiro - Wilk στο δείγμα $t_1, t_2, \ldots, t_n$ των εσωτερικά τυποποιημένων καταλοίπων και λαμβάνουμε το p-value. Για δεδομένο ε.σ.σ. α, έχουμε τα εξής ενδεχόμενα:

- Αν p-value $< \alpha$, τότε απορρίπτουμε την H_0, οπότε τα εσωτερικά τυποποιημένα κατάλοιπα δεν είναι κανονικά κατανεμημένα. Σε αυτήν την περίπτωση, η υπόθεση ότι τα τυχαία σφάλματα προέρχονται από την κανονική κατανομή δεν ευσταθεί, οπότε θα μπορούσαμε να καταφύγουμε στην κατασκευή ενός γενικευμένου γραμμικού μοντέλου.
- Αν p-value $> \alpha$, τότε δεν μπορούμε να απορρίψουμε την H_0, δηλαδή την υπόθεση ότι τα εσωτερικά τυποποιημένα κατάλοιπα είναι κανονικά κατανεμημένα. Επομένως, μπορούμε να δεχτούμε ότι το γραμμικό μοντέλο που έχουμε κατασκευάσει έχει τυχαία σφάλματα που προέρχονται από την κανονική κατανομή.

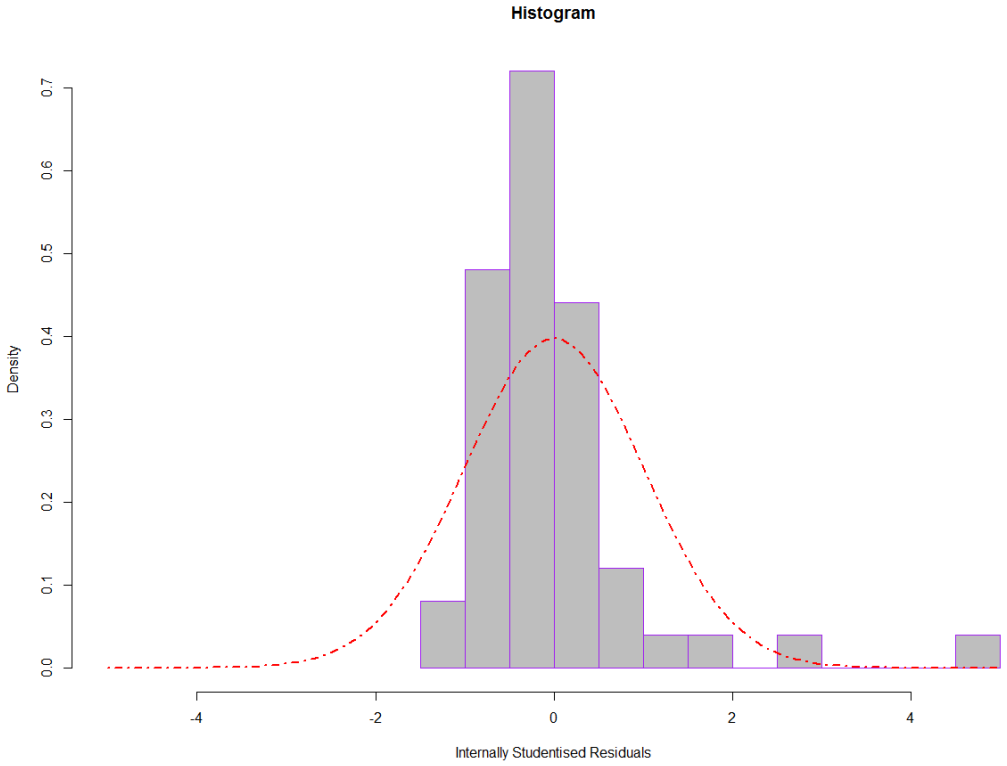
Οι έλεγχοι κανονικότητας καλό είναι πάντα να συνοδεύονται και από διάφορους γραφικούς ελέγχους, όπως ιστογράμματα (histograms), θηκογράμματα (boxplots) και Q-Q plots (quantile - quantile plots).

Πάνω από το **ιστόγραμμα** σχεδιάζουμε την καμπύλη της συνάρτησης πυκνότητας πιθανότητας της κανονικής κατανομής. Αν το ιστόγραμμα συμφωνεί με την καμπύλη της τυποποιημένης κανονικής κατανομής $N(0, 1)$, όπως φαίνεται στο σχήμα 2.2, τότε αυτό είναι ένδειξη ότι τα κατάλοιπα είναι κανονικά κατανεμημένα. Αντιθέτως, στο σχήμα 2.3, βλέπουμε ότι τα κατάλοιπα δεν είναι συμμετρικά γύρω από το μηδέν, όπως θα έπρεπε, ενώ υπάρχουν κιόλας κάποια κατάλοιπα τα οποία είναι υπερβολικά απομακρυσμένα από το μηδέν προς τα θετικά, πράγμα το οποίο δε συμφωνεί με την κανονική κατανομή.

Σε ένα **θηκόγραμμα**, όπως αυτά που φαίνονται στα σχήματα 2.4 και 2.5, το ορθογώνιο πλαίσιο αντιπροσωπεύει το ενδοτεταρτημοριακό εύρος (IR - interquartile range), δηλαδή το κεντρικό διάστημα στο οποίο ανήκει το 50% των παρατηρήσεων. Με άλλα λόγια, αν διατάξουμε τις παρατηρήσεις μας σε αύξουσα σειρά, τότε το 25% των παρατηρήσεων θα βρίσκεται αριστερά του διαστήματος που ορίζει το ορθογώνιο πλαίσιο και το υπόλοιπο 25% των παρατηρήσεων θα βρίσκεται δε-



Σχημα 2.2: Ιστόγραμμα Κανονικά Κατανεμημένων Καταλοίπων



Σχημα 2.3: Ιστόγραμμα Μη-Κανονικά Κατανεμημένων Καταλοίπων

ξιά αυτού του διαστήματος. Το αριστερό άκρο αυτού του διαστήματος καλείται πρώτο τεταρτημόριο (1Q - first quartile) και το δεξί άκρο καλείται τρίτο τεταρτημόριο (3Q - third quartile) του διανύσματος των παρατηρήσεων.

Μέσα στο ορθογώνιο πλαίσιο είναι σχεδιασμένη με μαύρη γραμμή η διάμεσος (median) του διανύσματος των παρατηρήσεων, δηλαδή η μεσαία διατεταγμένη παρατήρηση. Η διάμεσος αποτελεί το δεύτερο τεταρτημόριο (2Q - second quartile) του διανύσματος των παρατηρήσεων.

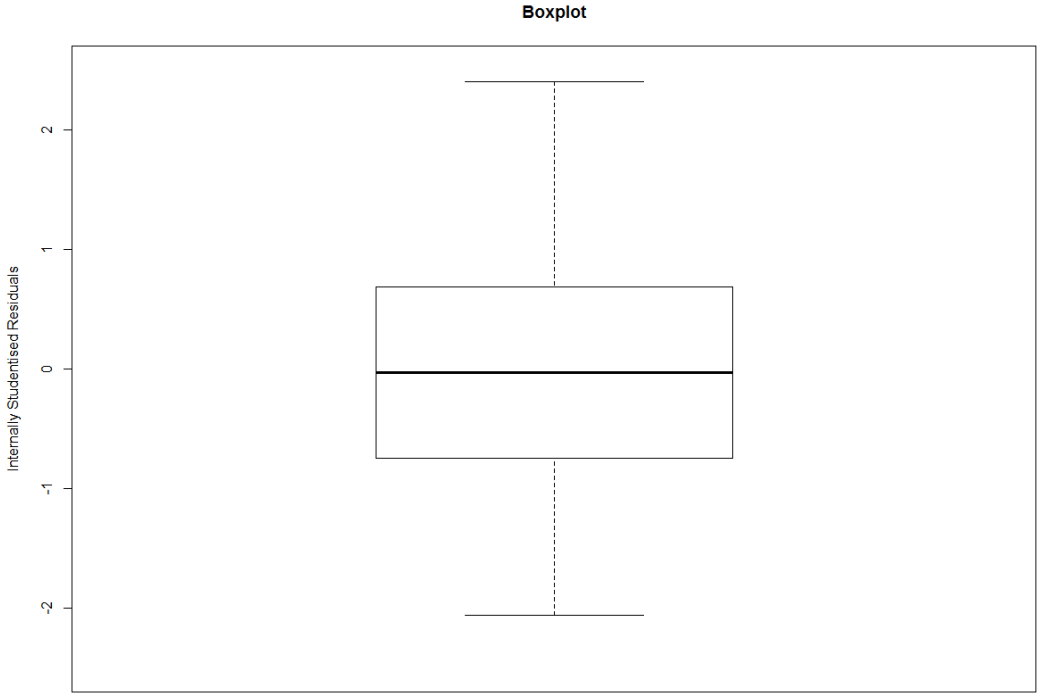
Οι απολήξεις που βρίσκονται σχεδιασμένες εκτός του ορθογωνίου πλαισίου έχουν μήκη $\min\left\{1Q - x_{(1)}, 1.5 \cdot \text{IR}\right\}$ και $\min\left\{x_{(n)} - 3Q, 1.5 \cdot \text{IR}\right\}$ αντίστοιχα, όπου $x_{(1)}$ είναι η ελάχιστη παρατήρηση και $x_{(n)}$ η μέγιστη παρατήρηση του δείγματος $\mathbf{x} = (x_1, x_2, \ldots, x_n)$. Επομένως, υποδεικνύουν την ελάχιστη και τη μέγιστη παρατήρηση αν όλες οι παρατηρήσεις εκτός του ορθογωνίου έχουν απόσταση το πολύ $1.5 \cdot \text{IR}$ από το πρώτο ή το τρίτο τεταρτημόριο αντίστοιχα. Διαφορετικά, αν υπάρχουν παρατηρήσεις με απόσταση μεγαλύτερη από $1.5 \cdot \text{IR}$, τότε σχεδιάζονται η καθεμία ξεχωριστά εκτός των ορίων των απολήξεων.

Αν η διάμεσος βρίσκεται περίπου στη μέση του ορθογώνιου πλαισίου, οι απολήξεις έχουν περίπου ίσα μήκη και δεν υπάρχουν παρατηρήσεις εκτός των ορίων των απολήξεων, όπως φαίνεται στο σχήμα 2.4, τότε τα κατάλοιπα φαίνεται να προέρχονται από την κανονική κατανομή. Αντιθέτως, στο σχήμα 2.5, η διάμεσος είναι πιο κοντά στην κάτω πλευρά του ορθογωνίου, οι απολήξεις έχουν άνισα μήκη και υπάρχουν 3 πολύ απομακρυσμένες παρατηρήσεις στο άνω μέρος του διαγράμματος, οπότε τα κατάλοιπα σίγουρα δεν ακολουθούν κανονική κατανομή.

Σε ένα Normal Q-Q plot, έχουμε στον άξονα των y τα δειγματικά ποσοστιαία σημεία, δηλαδή τις διατεταγμένες παρατηρήσεις $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$, ενώ στον άξονα των x τα θεωρητικά ποσοστιαία σημεία $\Phi^{-1}\left(\frac{k-0.5}{n}\right)$ της τυποποιημένης κανονικής κατανομής $N(0,1)$ για $k = 1, 2, \ldots, n$.

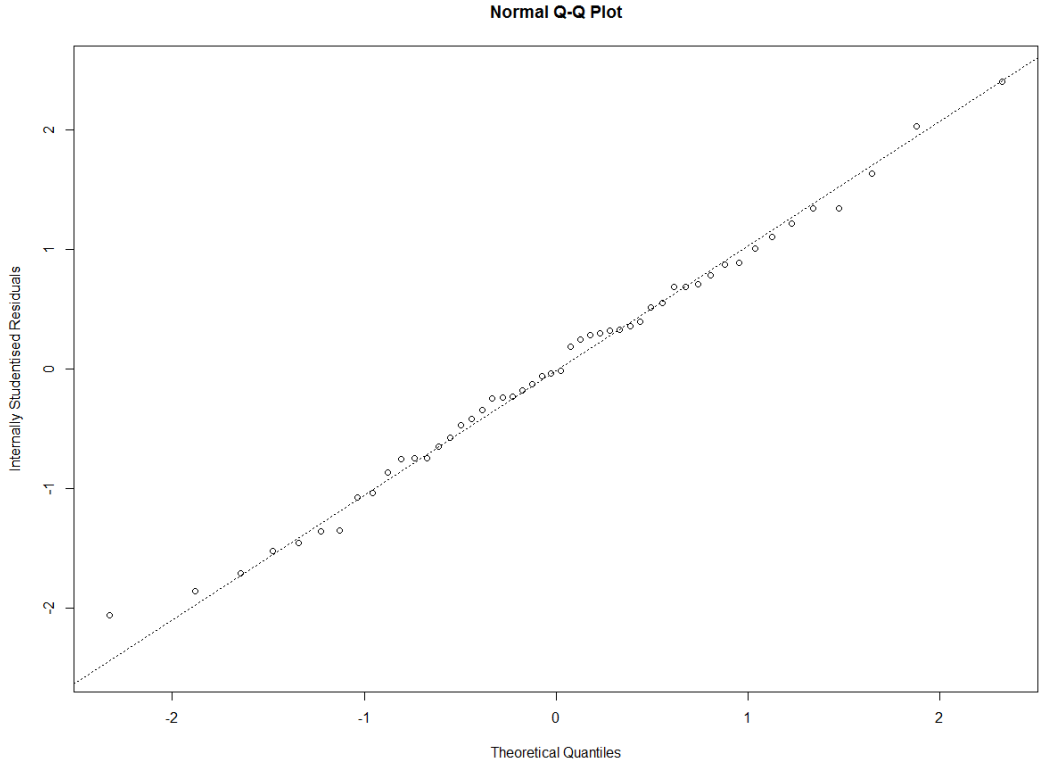
Αν η δειγματική κατανομή και η θεωρητική κατανομή, δηλαδή η κανονική κατανομή, συμφωνούν μεταξύ τους, τότε όλα τα σημεία του γραφήματος θα βρίσκονται συγκεντρωμένα πολύ κοντά σε μία ευθεία, όπως φαίνεται στο σχήμα 2.6. Αυτό είναι ένδειξη ότι τα εσωτερικά τυποποιημένα κατάλοιπα είναι, όντως, κανονικά κατανεμημένα.

Σε αντίθετη περίπτωση, μπορεί να υπάρχουν σημεία που απέχουν πολύ από την ευθεία που απεικονίζεται στο γράφημα. Ειδικότερα, αν τα σημεία απέχουν πολύ από την ευθεία στις ουρές τις κατανομής, δηλαδή στο άνω δεξί και το κάτω αριστερό μέρος του γραφήματος, όπως φαίνεται στο σχήμα 2.7, τότε αυτό είναι ένδειξη ότι τα κατάλοιπα προέρχονται από κάποια κατανομή με πιο "παχιές" ουρές από την κανονική κατανομή, όπως η κατανομή t του Student.

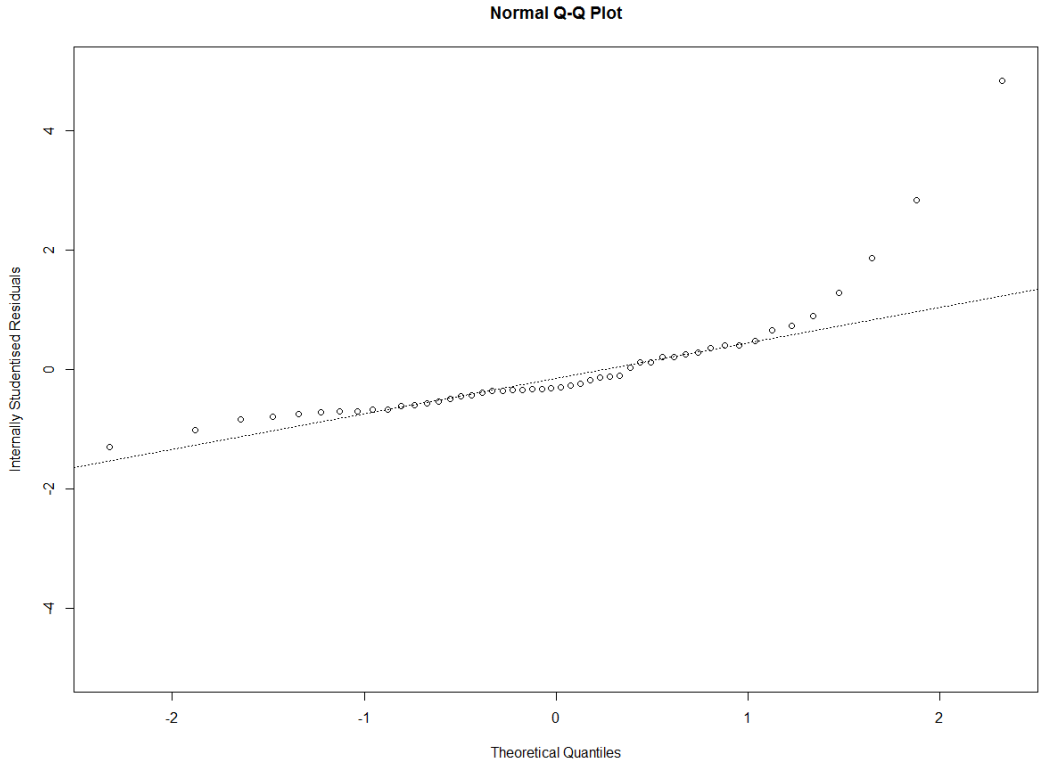


Σχημα 2.4: Θηκόγραμμα Κανονικά Κατανεμημένων Καταλοίπων

Σχημα 2.5: Θηκόγραμμα Μη-Κανονικά Κατανεμημένων Καταλοίπων



Σχημα 2.6: Normal Q-Q Plot Κανονικά Κατανεμημένων Καταλοίπων



Σχημα 2.7: Normal Q-Q Plot Μη-Κανονικά Κατανεμημένων Καταλοίπων

Έστω ότι έχουμε δεδομένα χρονοσειράς, δηλαδή μεγέθη που έχουμε συλλέξει από την ίδια μονάδα σε διαδοχικές χρονικές περιόδους. Τότε, είναι εύλογο να περιμένουμε ότι κάθε τυχαίο σφάλμα ε_i θα εξαρτάται από το αμέσως προηγούμενο τυχαίο σφάλμα ε_{i-1}. Ενδεχομένως, μάλιστα, να εξαρτάται και από άλλα παρελθοντικά τυχαία σφάλματα ε_{i-2}, ε_{i-3} και ούτω καθεξής. Σε αυτήν την περίπτωση, η υπόθεση ότι τα τυχαία σφάλματα είναι ανεξάρτητα προφανώς δεν ευσταθεί.

Αντιθέτως, όταν έχουμε διαστρωματικά δεδομένα, δηλαδή μεγέθη που έχουμε συλλέξει από διαφορετικές μονάδες την ίδια χρονική στιγμή, τότε δεν υπάρχει καμία εύλογη σειρά διάταξης των παρατηρήσεων, στην πλειοψηφία των περιπτώσεων. Σε αυτήν την περίπτωση, δεν μπορούμε σαφώς να θεωρήσουμε μία τέτοια μορφή εξάρτησης μεταξύ των τυχαίων σφαλμάτων και είναι εύλογο να θεωρήσουμε ότι είναι όλα ανεξάρτητα μεταξύ τους. Παρόλα αυτά, ακόμα και στην περίπτωση των διαστρωματικών δεδομένων, θα μπορούσε ενδεχομένως να υπάρχει κάποια σαφής χωρική διάταξη μεταξύ των παρατηρήσεων, η οποία να επιφέρει αυτή τη μορφή εξάρτησης.

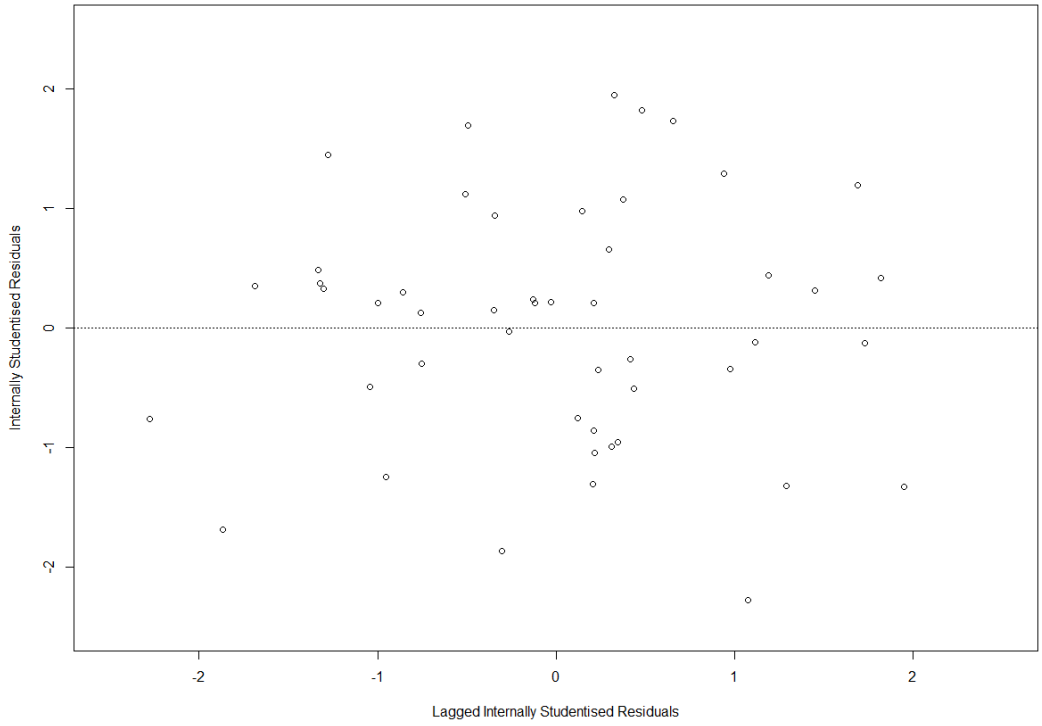
Για να ελέγξουμε αν τα κατάλοιπα είναι ανεξάρτητα, μπορούμε να εφαρμόσουμε διάφορους ελέγχους αυτοσυσχέτισης, όπως ο **έλεγχος Durbin - Watson** και ο **έλεγχος Breusch - Godfrey**. Οι έλεγχοι αυτοί έχουν ως μηδενική υπόθεση ότι οι παρατηρήσεις είναι ασυσχέτιστες, ενώ ως εναλλακτική υπόθεση ότι υπάρχει σειριακή συσχέτιση μεταξύ τους.

Όλα τα στατιστικά λογισμικά, όπως η R, έχουν ενσωματωμένη τη δυνατότητα πραγματοποίησης των ελέγχων αυτοσυσχέτισης για δεδομένο μοντέλο παλινδρόμησης και δίνουν ως αποτέλεσμα τα p-value των ελέγχων. Εφαρμόζουμε τους ελέγχους Durbin - Watson και Breusch - Godfrey στο γραμμικό μοντέλο και λαμβάνουμε τα p-value. Για δεδομένο ε.σ.σ. α, έχουμε τα εξής ενδεχόμενα:

- Αν p-value $< \alpha$, τότε απορρίπτουμε την H_0, δηλαδή την υπόθεση ότι τα κατάλοιπα είναι ασυσχέτιστα. Σε αυτήν την περίπτωση, η υπόθεση ότι τα τυχαία σφάλματα είναι ανεξάρτητα δεν ευσταθεί, οπότε είμαστε αναγκασμένοι να καταφύγουμε στην κατασκευή ενός αυτοπαλίνδρομου μοντέλου ή ενός μοντέλου κινητού μέσου.
- Αν p-value $> \alpha$, τότε δεν μπορούμε να απορρίψουμε την H_0, οπότε τα εσωτερικά τυποποιημένα κατάλοιπα μπορούμε να θεωρήσουμε ότι είναι ασυσχέτιστα. Επομένως, η υπόθεση ότι τα τυχαία σφάλματα του γραμμικού μοντέλου είναι ανεξάρτητα ευσταθεί.

Προκειμένου να ελέγξουμε και γραφικά την ανεξαρτησία των εσωτερικά τυποποιημένων καταλοίπων, μπορούμε να σχεδιάσουμε γραφήματα των t_i με τα t_{i-1} και με τον χρόνο i της παρατήρησης. Αν τα κατάλοιπα είναι εντελώς τυχαία διε-

σπαρμένα γύρω από την οριζόντια ευθεία $y = 0$, όπως φαίνεται στα σχήματα 2.8 και 2.10, τότε συμπεραίνουμε ότι τα κατάλοιπα είναι ανεξάρτητα. Διαφορετικά, αν εμφανίζουν κάποια φθίνουσα ή αύξουσα τάση, όπως φαίνεται στα σχήματα 2.9 και 2.11, τότε συμπεραίνουμε σαφώς ότι υπάρχει σειριακή εξάρτηση ανάμεσα στα κατάλοιπα.

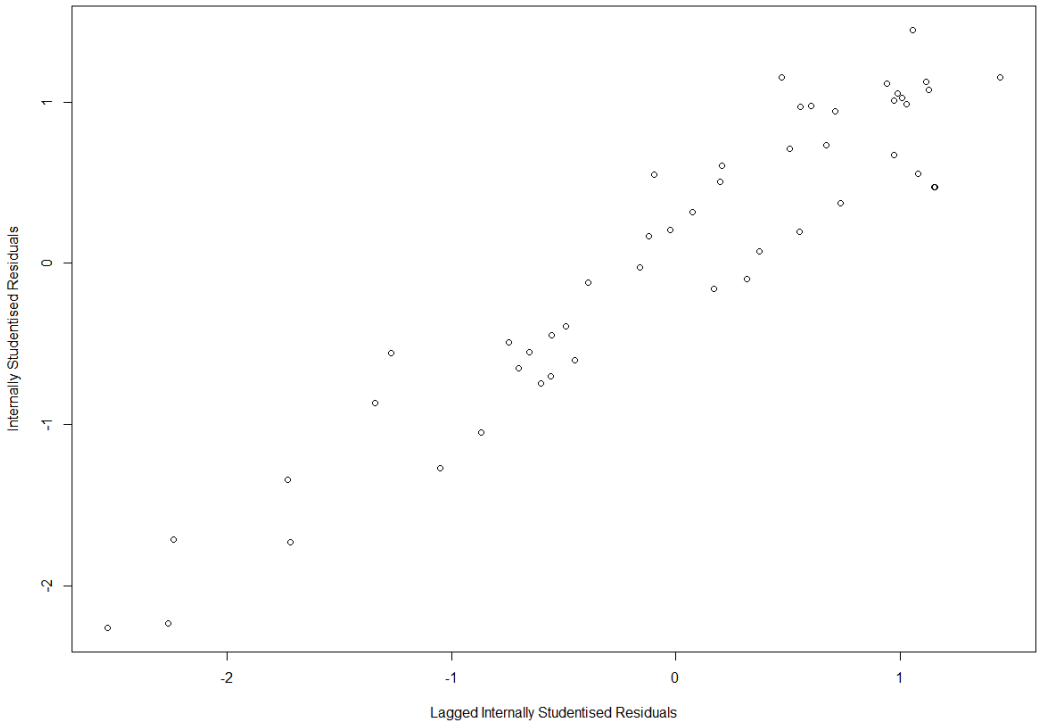


Σχημα 2.8: Γράφημα των t_i με τα t_{i-1}

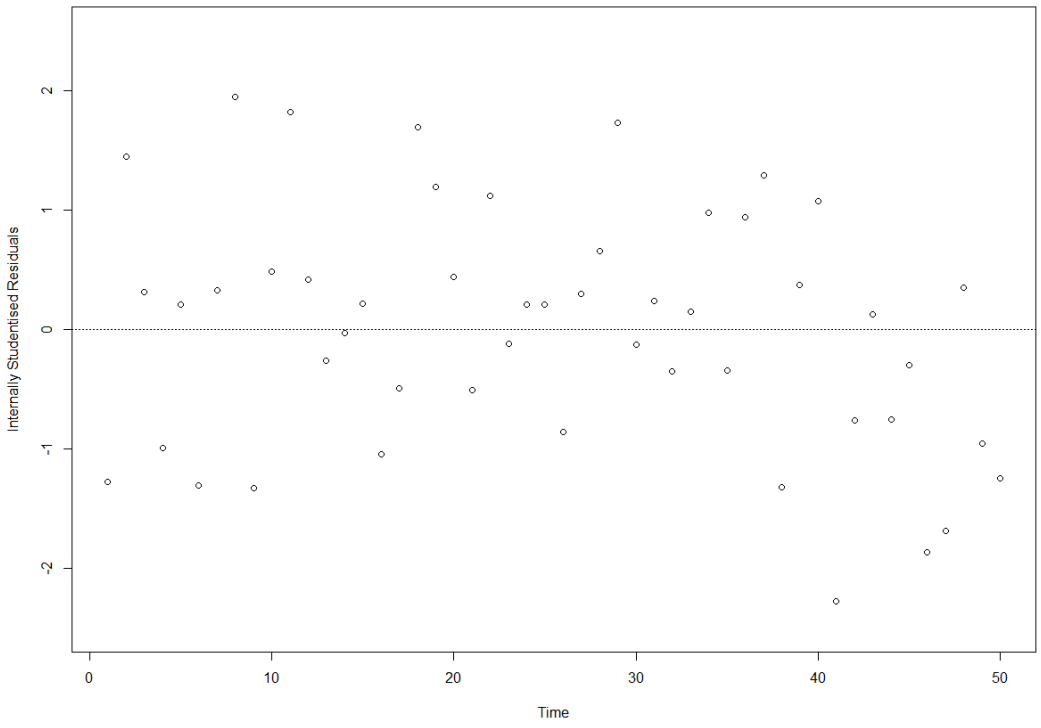
Η διασπορά των τυχαίων σφαλμάτων μπορεί ενδεχομένως να είναι γραμμική ή τετραγωνική συνάρτηση μίας ή περισσότερων από τις διαθέσιμες επεξηγηματικές μεταβλητές. Για να ελέγξουμε αυτή τη μορφή ετεροσκεδαστικότητας, μπορούμε να πραγματοποιήσουμε τον **έλεγχο Breusch - Pagan** και τον **έλεγχο White**. Οι έλεγχοι αυτοί έχουν ως μηδενική υπόθεση ότι τα τυχαία σφάλματα είναι ομοσκεδαστικά, ενώ ως εναλλακτική υπόθεση ότι εμφανίζουν ετεροσκεδαστικότητα.

Όλα τα στατιστικά λογισμικά, όπως η R, έχουν ενσωματωμένη τη δυνατότητα πραγματοποίησης των ελέγχων ετεροσκεδαστικότητας για δεδομένο μοντέλο παλινδρόμησης και δίνουν ως αποτέλεσμα τα p-value των ελέγχων. Εφαρμόζουμε τους ελέγχους Breusch - Pagan και White στο γραμμικό μοντέλο που έχουμε κατασκευάσει και λαμβάνουμε τα p-value. Για δεδομένο ε.σ.σ. α, έχουμε τα εξής ενδεχόμενα:

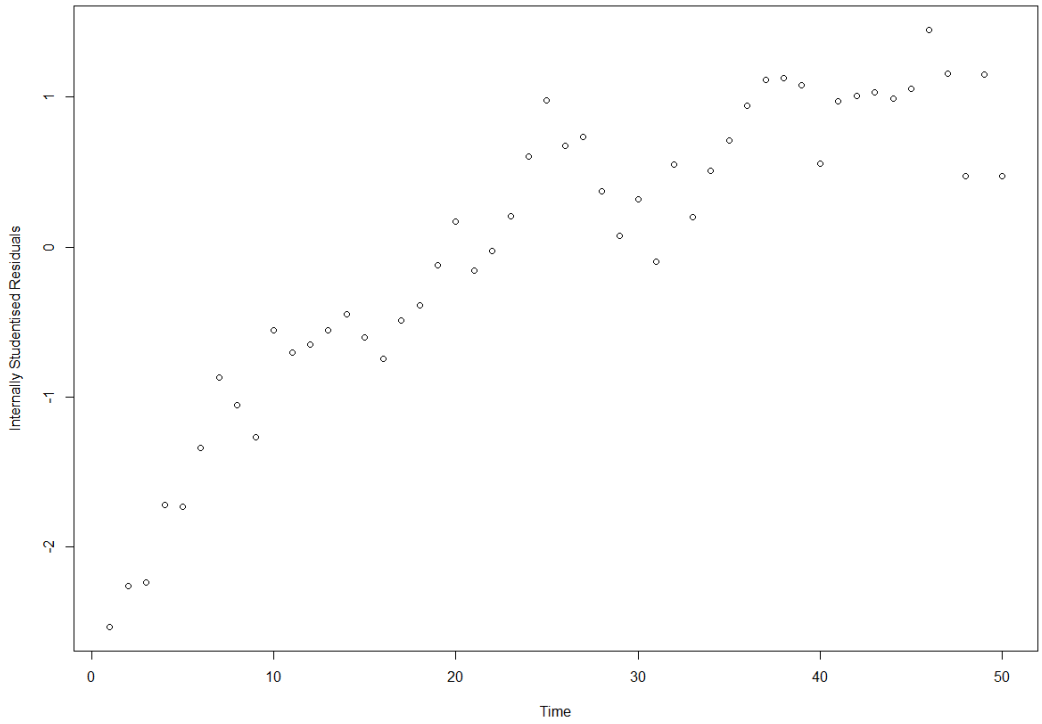
- Αν p-value $< \alpha$, τότε απορρίπτουμε την H_0, δηλαδή την υπόθεση ότι τα τυχαία σφάλματα είναι ομοσκεδαστικά.



Σχημα 2.9: Γράφημα των t_i με τα t_{i-1}



Σχημα 2.10: Γράφημα των Καταλοίπων με τον Χρόνο



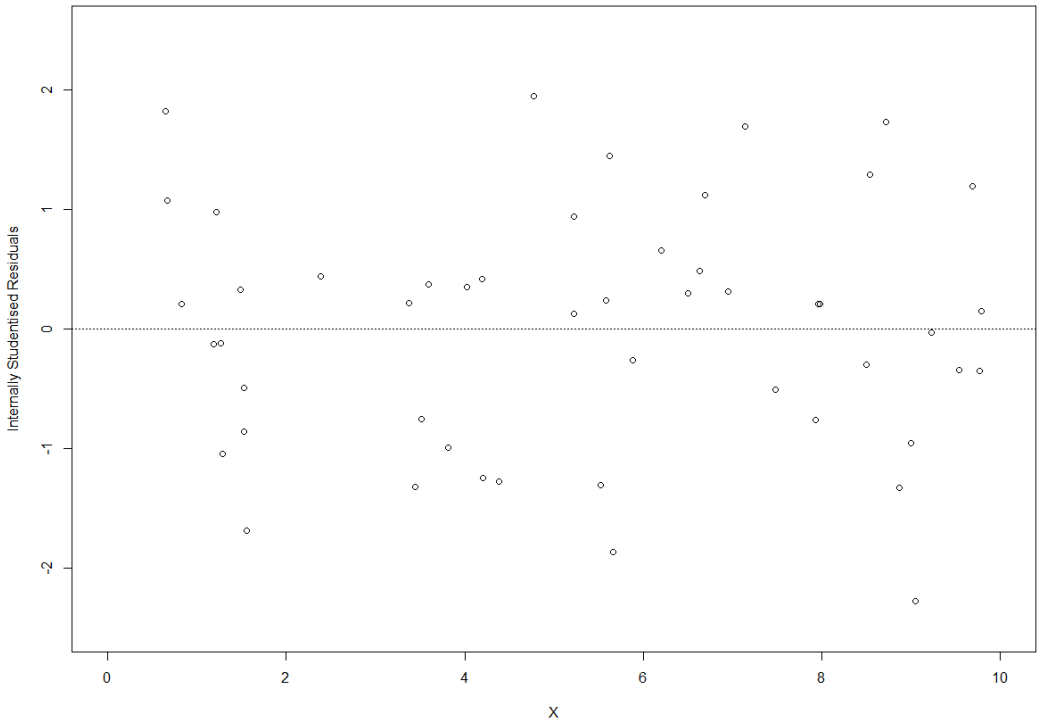
Σχημα 2.11: Γράφημα των Καταλοίπων με τον Χρόνο

- Αν p-value $> \alpha$, τότε δεν μπορούμε να απορρίψουμε την H_0, οπότε τα τυχαία σφάλματα μπορούμε να θεωρήσουμε ότι είναι ομοσκεδαστικά.

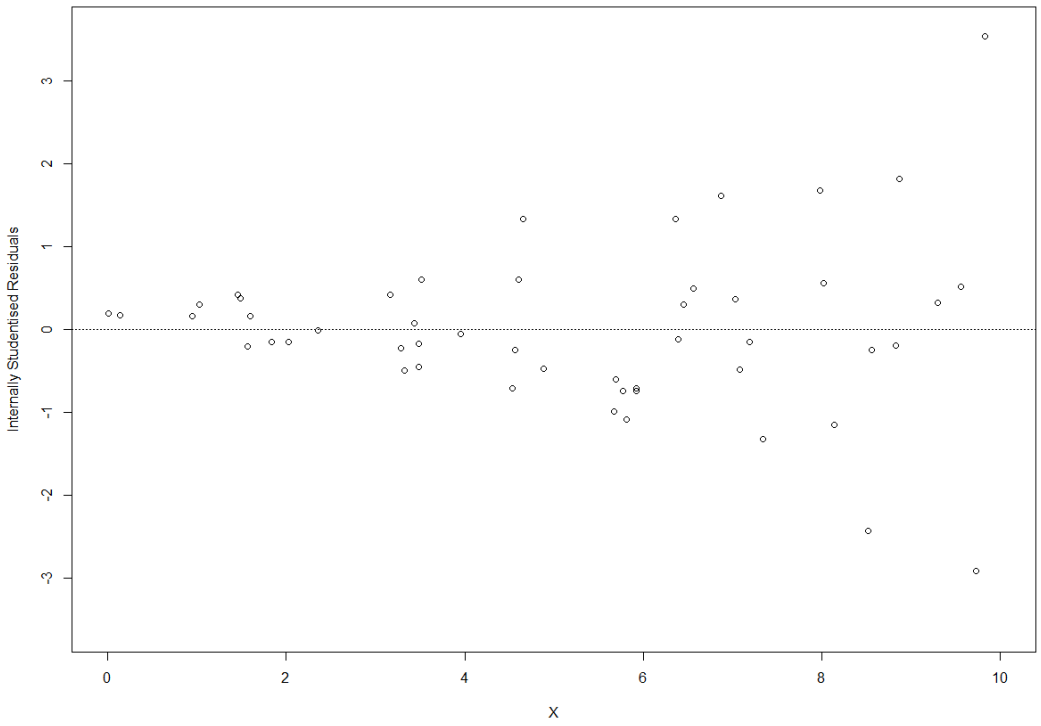
Προκειμένου να ελέγξουμε γραφικά από ποιες επεξηγηματικές μεταβλητές θα μπορούσε να εξαρτάται η διασπορά των τυχαίων σφαλμάτων, σχεδιάζουμε γραφήματα των t_i με τις παρατηρήσεις X_i από κάθε διαθέσιμη επεξηγηματική μεταβλητή X. Αν τα κατάλοιπα έχουν σταθερή διασπορά γύρω από την οριζόντια ευθεία $y = 0$ σε όλα τα γραφήματα, όπως φαίνεται στο σχήμα 2.12, τότε συμπεραίνουμε ότι τα κατάλοιπα είναι ομοσκεδαστικά. Διαφορετικά, αν η διασπορά τους εμφανίζει κάποια φθίνουσα ή αύξουσα τάση, όπως φαίνεται στο σχήμα 2.13, τότε συμπεραίνουμε σαφώς ότι υπάρχει ετεροσκεδαστικότητα.

Στην περίπτωση όπου έχουμε δεδομένα χρονοσειράς, συμπεριλαμβάνουμε και τον χρόνο της παρατήρησης στις διαθέσιμες επεξηγηματικές μεταβλητές, οπότε μπορούμε να τον συμπεριλάβουμε στους προαναφερθέντες ελέγχους ετεροσκεδαστικότητας. Στο σχήμα 2.10, φαίνεται ότι τα κατάλοιπα έχουν σταθερή διασπορά στον χρόνο, ενώ στο σχήμα 2.14 φαίνεται ότι η διασπορά τους έχει αύξουσα τάση, οπότε εκεί συμπεραίνουμε σαφώς ότι υπάρχει ετεροσκεδαστικότητα.

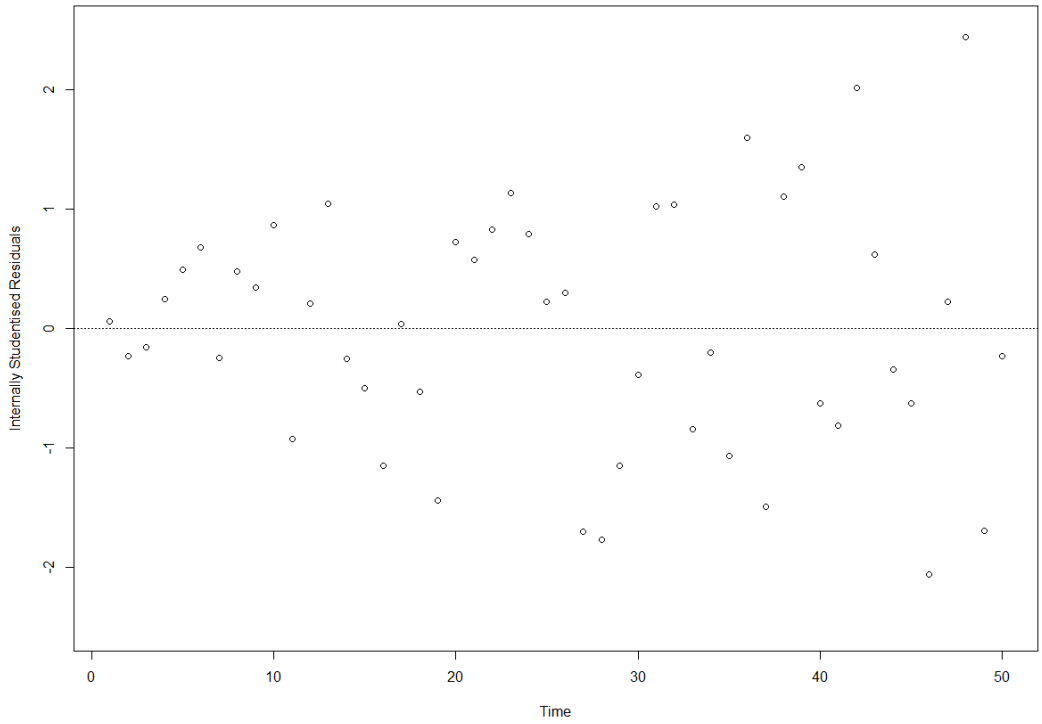
Ένα εναλλακτικό σενάριο ετεροσκεδαστικότητας, στην περίπτωση δεδομένων χρονοσειράς, είναι να υπάρχει κάποια γνωστή χρονική στιγμή n^* τέτοια, ώστε τα τυχαία σφάλματα $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_{n^*}$ να έχουν διασπορά σ_1^2, ενώ τα τυχαία σφάλματα



Σχημα 2.12: Γράφημα των Καταλοίπων με την Επεξηγηματική Μεταβλητή X



Σχημα 2.13: Γράφημα των Καταλοίπων με την Επεξηγηματική Μεταβλητή X



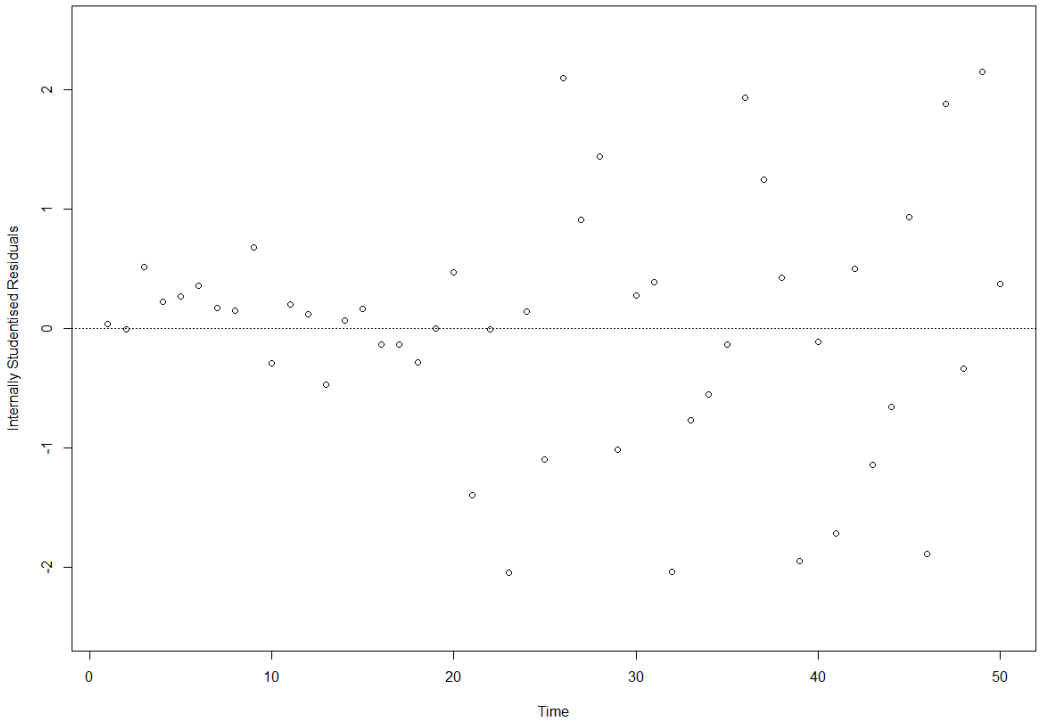
Σχημα 2.14: Γράφημα των Καταλοίπων με τον Χρόνο

$\varepsilon_{n^*+1}, \varepsilon_{n^*+2}, \ldots, \varepsilon_n$ να έχουν διαφορετική διασπορά σ_2^2. Για να ελέγξουμε τη μηδενική υπόθεση $H_0 : \sigma_1^2 = \sigma_2^2$ έναντι της εναλλακτικής υπόθεσης $H_1 : \sigma_1^2 < \sigma_2^2$ ή της $H_1 : \sigma_1^2 > \sigma_2^2$, εφαρμόζουμε τον **έλεγχο Goldfeld - Quandt**.

Όλα τα στατιστικά λογισμικά, όπως η R, έχουν ενσωματωμένη τη δυνατότητα πραγματοποίησης του ελέγχου Goldfeld - Quandt για δεδομένο μοντέλο παλινδρόμησης και δίνουν ως αποτέλεσμα το p-value του ελέγχου. Εφαρμόζουμε τον έλεγχο στο γραμμικό μοντέλο που έχουμε κατασκευάσει και λαμβάνουμε τα p-value. Για δεδομένο ε.σ.σ. α, έχουμε τα εξής ενδεχόμενα:

- Αν p-value $< \alpha$, τότε απορρίπτουμε την H_0, δηλαδή την υπόθεση ότι τα τυχαία σφάλματα είναι ομοσκεδαστικά.
- Αν p-value $> \alpha$, τότε δεν μπορούμε να απορρίψουμε την H_0, οπότε τα τυχαία σφάλματα μπορούμε να θεωρήσουμε ότι είναι ομοσκεδαστικά.

Προκειμένου να εντοπίσουμε τη χρονική στιγμή n^*, σχεδιάζουμε γραφήματα των t_i με τον χρόνο i της παρατήρησης. Στο σχήμα 2.15, φαίνεται να υπάρχει δραστική αύξηση στη διασπορά των καταλοίπων γύρω στη χρονική στιγμή $n^* = 20$, οπότε πραγματοποιούμε τον έλεγχο Goldfeld - Quandt με $n^* = 20$ και εναλλακτική υπόθεση $H_1 : \sigma_1^2 < \sigma_2^2$. Σε αυτήν την περίπτωση, περιμένουμε σαφώς ότι ο έλεγχος θα δώσει p-value $< \alpha$, δηλαδή θα απορρίψουμε την H_0.

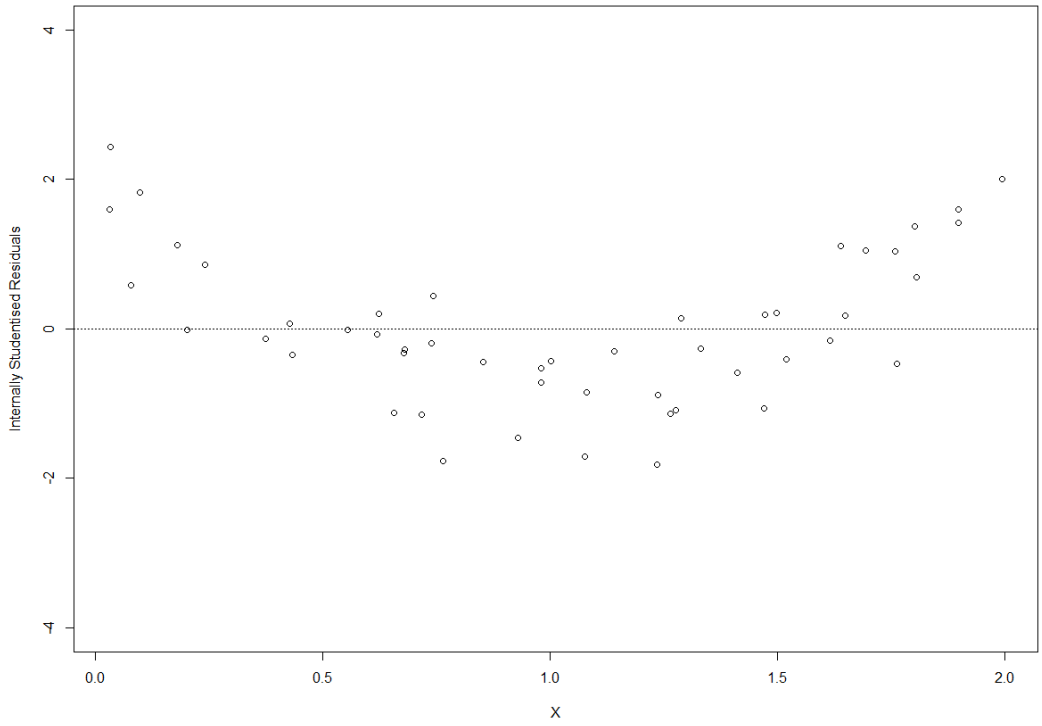


Σχημα 2.15: Γράφημα των Καταλοίπων με τον Χρόνο

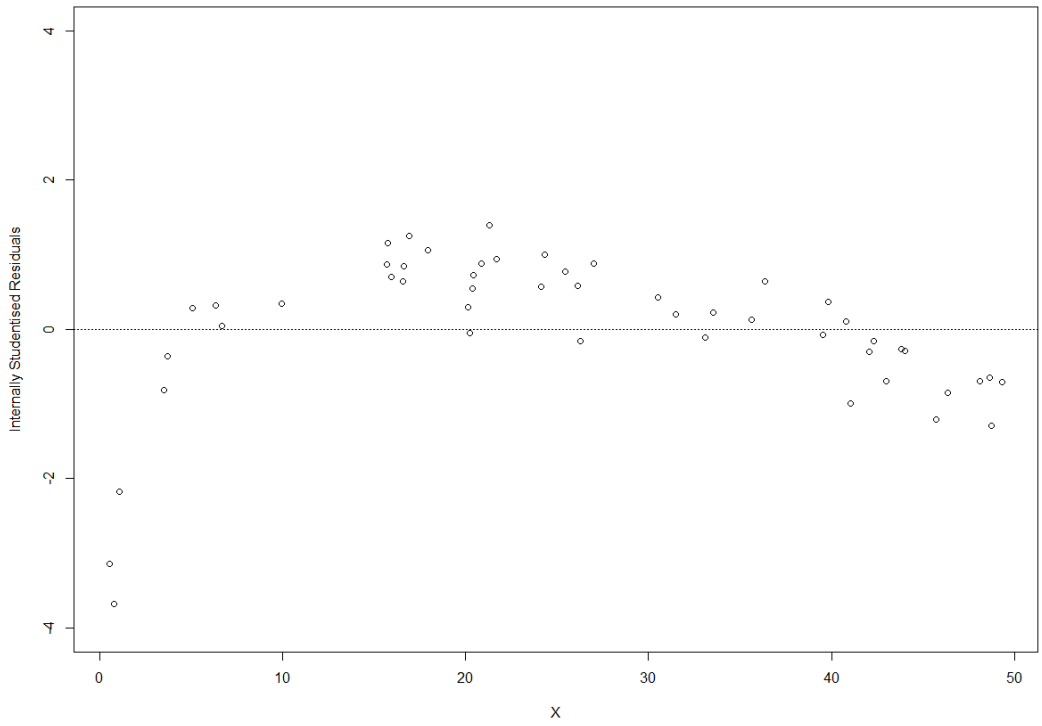
Τα γραφήματα των καταλοίπων με κάποια επεξηγηματική μεταβλητή X του γραμμικού μοντέλου μπορούν να μας υποδείξουν και την ύπαρξη μη-γραμμικής σχέσης μεταξύ αυτής και της αποκριτικής μεταβλητής Y. Σε αυτή την περίπτωση, επιβάλλεται να προβούμε σε μετασχηματισμό της επεξηγηματικής μεταβλητής X. Για παράδειγμα, αν παίρναμε το γράφημα που φαίνεται στο σχήμα 2.16, θα σκεφτόμασταν να θέσουμε $X^* = X^2$ και να χρησιμοποιήσουμε αυτή ως επεξηγηματική μεταβλητή αντί της X, ενώ αν παίρναμε αυτό που φαίνεται στο σχήμα 2.17, θα σκεφτόμασταν αντίστοιχα να θέσουμε $X^* = \log X$.

Έχουμε αποδείξει στην πρόταση 2.5, ότι τα κατάλοιπα του γραμμικού μοντέλου είναι ασυσχέτιστα με τις προσαρμοσμένες τιμές $\widehat{Y}_i$. Επομένως, αν όλες οι υποθέσεις που έχουμε κάνει για το γραμμικό μοντέλο ευσταθούν και σχεδιάσουμε ένα γράφημα που έχει στον άξονα των x τις προσαρμοσμένες τιμές και στον άξονα των y τα κατάλοιπα, τότε τα κατάλοιπα θα πρέπει να είναι εντελώς τυχαία διεσπαρμένα γύρω από την οριζόντια ευθεία $y = 0$ με κοινή διασπορά και χωρίς να εμφανίζουν κάποια τάση, όπως ακριβώς φαίνεται στο σχήμα 2.18.

Το γράφημα αυτό είναι γνωστό και ως γράφημα **Residual vs Fitted**. Στην αντίθετη περίπτωση, όπου κάποια από τις υποθέσεις του γραμμικού μοντέλου παραβιάζεται, τότε το γράφημα μπορεί να παρουσιάζει κάποια από τις τάσεις που δείξαμε στα σχήματα 2.13, 2.15, 2.16 ή 2.17. Τότε, συμπεραίνουμε αντίστοιχα

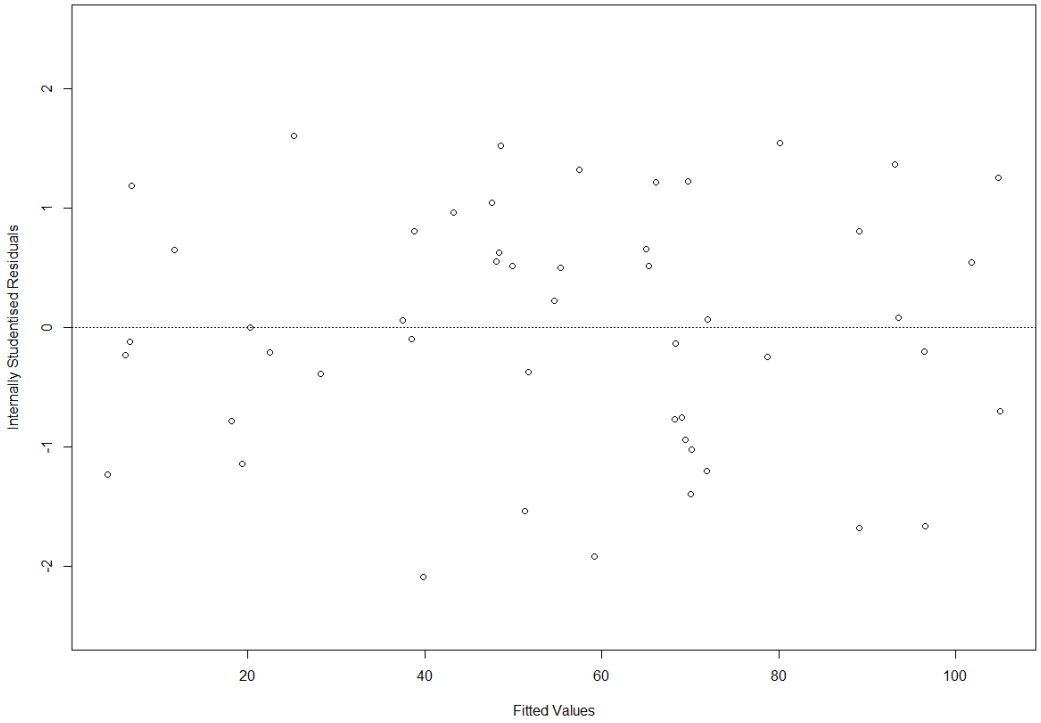


Σχημα 2.16: Γράφημα των Καταλοίπων με την Επεξηγηματική Μεταβλητή X



Σχημα 2.17: Γράφημα των Καταλοίπων με την Επεξηγηματική Μεταβλητή X

ότι υπάρχει πρόβλημα ετεροσκεδαστικότητας ή μη-γραμμικής σχέσης μεταξύ αποκριτικής και επεξηγηματικής μεταβλητής.



Σχημα 2.18: Γράφημα των Καταλοίπων με τις Προσαρμοσμένες Τιμές $\widehat{Y}_i$

Πολλές φορές, όταν επιχειρούμε να κατασκευάσουμε ένα μοντέλο γραμμικής παλινδρόμησης, υπάρχουν στο δείγμα μας ορισμένες παρατηρήσεις που απέχουν δυσανάλογα από το κέντρο βάρους του δείγματος και μπορούν ενδεχομένως να ασκήσουν μεγάλη επιρροή στην εκτιμημένη εξίσωση παλινδρόμησης. Αυτές οι παρατηρήσεις χωρίζονται σε έκτροπες παρατηρήσεις (outliers), σημεία μοχλούς (high leverage points) και σημεία επιρροής (influential observations).

Οι **έκτροπες παρατηρήσεις** είναι παρατηρήσεις που έχουν πολύ μεγάλη απόσταση από την εκτιμημένη εξίσωση παλινδρόμησης σε σύγκριση με τις υπόλοιπες παρατηρήσεις. Με άλλα λόγια, έκτροπες είναι οι παρατηρήσεις που δίνουν πολύ μεγάλα κατά απόλυτη τιμή κατάλοιπα. Για πιο αξιόπιστο εντοπισμό των έκτροπων παρατηρήσεων, κάνουμε χρήση μίας διαφορετικής τεχνικής τυποποίησης των καταλοίπων από αυτή που είδαμε για τα εσωτερικά τυποποιημένα κατάλοιπα.

Έστω $S^2_{(-i)} = \frac{\mathrm{SSE}_{(-i)}}{n-p-2}$ η αμερόληπτη εκτιμήτρια της διασποράς που προκύπτει από το γραμμικό μοντέλο που προσαρμόζουμε στο δείγμα μας, έχοντας αφαιρέσει την παρατήρηση $\left(Y_i, \mathbf{X}_i^{\mathrm{T}}\right)$. Η εκτιμήτρια $S^2_{(-i)}$ είναι προφανώς ανεξάρτητη από την παρατήρηση Y_i. Επιπλέον, μπορούμε να δείξουμε ότι $Q = \frac{(n-p-2)S^2_{(-i)}}{\sigma^2} \sim \chi^2_{n-p-2}$ και ότι το $S^2_{(-i)}$ είναι ανεξάρτητο από την εκτιμήτρια $\widehat{\boldsymbol{\beta}}$ του $\boldsymbol{\beta}$ που προκύπτει από

το γραμμικό μοντέλο που προσαρμόζουμε σε ολόκληρο το δείγμα. Επομένως, ορίζουμε τα **εξωτερικά τυποποιημένα κατάλοιπα** (externally studentised residuals) ως:

$$t_{(-i)} = \frac{\widehat{\varepsilon}_i}{S_{(-i)}\sqrt{1-\mathbf{P}_{i,i}}} \sim t_{n-p-2}.$$

Σε αντίθεση με τα εσωτερικά τυποποιημένα κατάλοιπα, τα εξωτερικά τυποποιημένα κατάλοιπα ακολουθούν την κατανομή t του Student. Μία παρατήρηση $\left(Y_i, \mathbf{X}_i^{\mathrm{T}}\right)$ καλείται **έκτροπη** (outlier) αν το αντίστοιχο εξωτερικά τυποποιημένο κατάλοιπο βρίσκεται στην ουρά της κατανομής t_{n-p-2}, δηλαδή $\left|t_{(-i)}\right| > t_{n-p-2;\frac{\alpha}{2}}$.

Για να αποφύγουμε τον υπολογισμό των εκτιμήσεων $S^2_{(-i)}$, μπορούμε να χρησιμοποιήσουμε τον ακόλουθο τύπο για τον απευθείας υπολογισμό των εξωτερικά τυποποιημένων καταλοίπων μέσω των αντίστοιχων εσωτερικά τυποποιημένων:

$$t_{(-i)} = t_i\sqrt{\frac{n-p-2}{n-p-1-t_i^2}}.$$

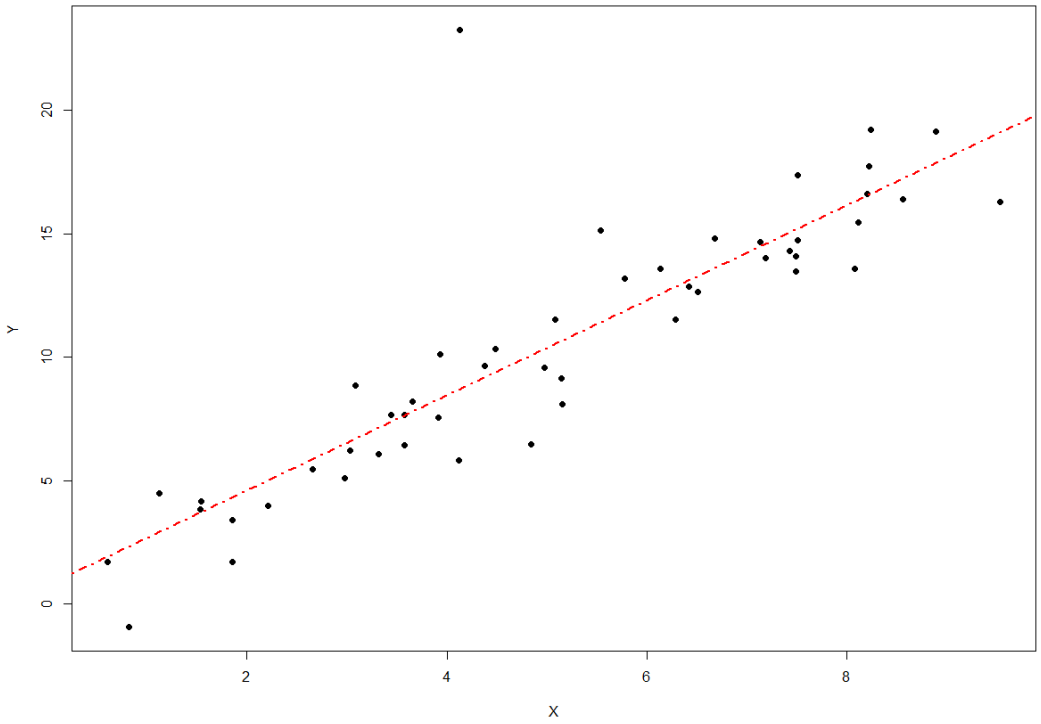
Βλέπουμε ότι το $t_{(-i)}$ είναι αύξουσα συνάρτηση του t_i και $t_i = 1 \Leftrightarrow t_{(-i)} = 1$. Στην περίπτωση όπου $t_i > 1$, τότε βλέπουμε ότι $t_{(-i)} > t_i$, οπότε το $t_{(-i)}$ μεγεθύνει τις αποκλίσεις των εσωτερικά τυποποιημένων καταλοίπων από τη μονάδα και καθιστά πιο εύκολο τον εντοπισμό των έκτροπων παρατηρήσεων.

Στο σχήμα 2.19, βλέπουμε καθαρά την ύπαρξη μίας έκτροπης παρατήρησης στο πάνω μέρος του γραφήματος. Στη γενικότερη περίπτωση όπου δεν έχουμε μόνο μία επεξηγηματική μεταβλητή, οι έκτροπες παρατηρήσεις γίνονται εμφανείς κυρίως μέσω του γραφήματος Residual vs Fitted. Οι έκτροπες παρατηρήσεις δε συμφωνούν με την υπόθεση της κανονικότητας των τυχαίων σφαλμάτων, οπότε εμφανίζονται στα γραφήματα κανονικότητας των καταλοίπων που περιγράψαμε στην αρχή της παραγράφου και οδηγούν συνήθως σε απόρριψη της μηδενικής υπόθεσης στους ελέγχους κανονικότητας.

Η **μόχλευση** (leverage) μίας παρατήρησης ποσοτικοποιεί τη σχετική απομάκρυνση των τιμών των επεξηγηματικών μεταβλητών της συγκεκριμένης παρατήρησης από το κέντρο βάρους του δείγματος των επεξηγηματικών μεταβλητών. Υπολογίζεται από τα διαγώνια στοιχεία του πίνακα ορθογώνιας προβολής (hat matrix) $\mathbf{P} = \mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}$, δηλαδή $\mathbf{P}_{i,i} = \mathbf{X}_i^{\mathrm{T}}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}_i$ είναι η μόχλευση της παρατήρησης i, όπου $\mathbf{X}_i = (1, X_{1,i}, X_{2,i}, \ldots, X_{p,i})^{\mathrm{T}}$. Αν η μόχλευση της παρατήρησης i είναι δυσανάλογα μεγάλη σε σύγκριση με τις μοχλεύσεις των υπόλοιπων παρατηρήσεων, τότε η παρατήρηση $\left(Y_i, \mathbf{X}_i^{\mathrm{T}}\right)$ καλείται σημείο μοχλός.

Πρόταση 2.17. (Μόχλευση)

i. $\sum_{i=1}^{n} P_{i,i} = p+1$, δηλαδή $\frac{1}{n}\sum_{i=1}^{n} P_{i,i} = \frac{p+1}{n}$.



Σχημα 2.19: Έκτροπη Παρατήρηση

ii. $\frac{1}{n} \leqslant P_{i,i} \leqslant 1$. Η ισότητα $P_{i,i} = \frac{1}{n}$ επιτυγχάνεται αν και μόνο αν ισχύει ότι $\mathbf{X}_i = \overline{\mathbf{X}} = \left(1, \overline{X}_1, \overline{X}_2, \ldots, \overline{X}_p\right)^{\mathrm{T}}$.

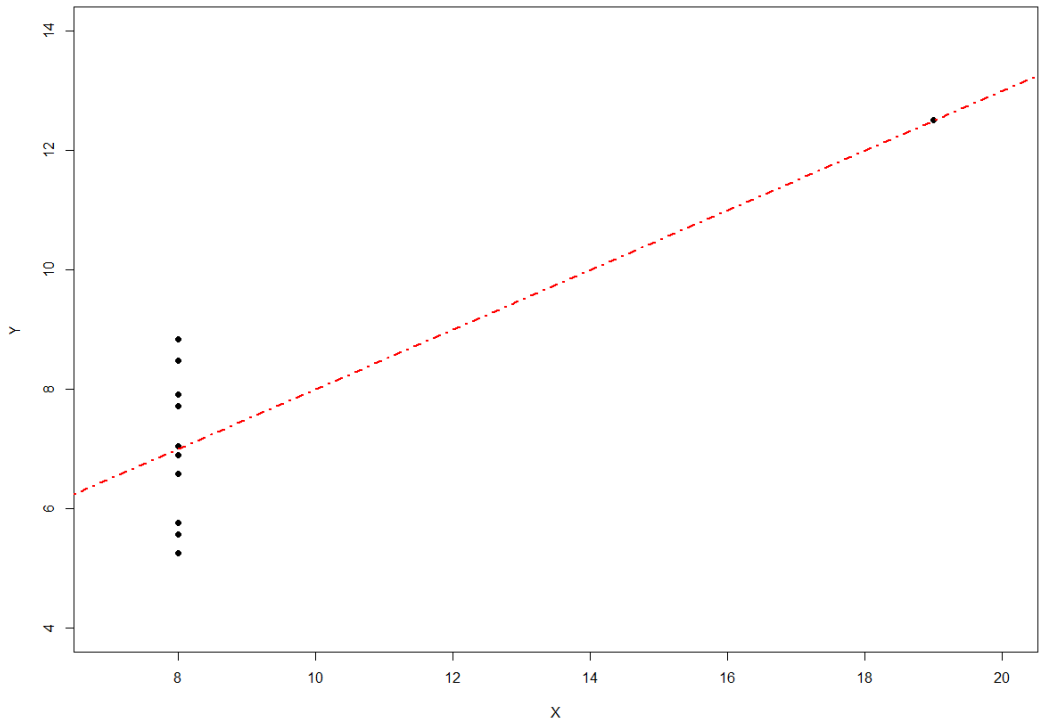
iii. Έστω $\mathbf{X}_{(-i)} \in \mathbb{R}^{(n-1)\times(p+1)}$ ο πίνακας σχεδιασμού που προκύπτει αν διαγράψουμε από τον $\mathbf{X}$ την i-οστή γραμμή, δηλαδή την i-οστή παρατήρηση. Η ισότητα $P_{i,i} = 1$ επιτυγχάνεται αν και μόνο αν $\operatorname{rank}\left[\mathbf{X}_{(-i)}\right] < p+1$, δηλαδή ο $\mathbf{X}_{(-i)}$ δεν είναι πλήρους τάξης. Με άλλα λόγια, αν αφαιρέσουμε μία παρατήρηση με μόχλευση 1 από το γραμμικό μοντέλο, τότε το γραμμικό μοντέλο που προκύπτει δεν είναι καλά ορισμένο.

Σύμφωνα με την παραπάνω πρόταση, η μόχλευση μίας παρατήρησης ισούται κατά μέσο όρο με $\frac{p+1}{n}$. Επομένως, η παρατήρηση $\left(Y_i, \mathbf{X}_i^{\mathrm{T}}\right)$ συνηθίζεται να καλείται **σημείο μοχλός** (high leverage point) αν έχει μόχλευση αρκετά πάνω από τον μέσο όρο, δηλαδή αν ισχύει ότι $\mathbf{P}_{i,i} > 2 \cdot \frac{p+1}{n}$ ή $\mathbf{P}_{i,i} > 3 \cdot \frac{p+1}{n}$. Όπως οι έκτροπες παρατηρήσεις καθορίζονται αποκλειστικά από τις τιμές των καταλοίπων, έτσι και τα σημεία μοχλοί καθορίζονται αποκλειστικά από τις τιμές των επεξηγηματικών μεταβλητών του γραμμικού μοντέλου.

Παρότι ασχολούμαστε με τη μόχλευση των παρατηρήσεων ενός γραμμικού μοντέλου, ένα σημείο μοχλός δεν αποτελεί πάντα πρόβλημα, εκτός αν έχει μόχλευση πολύ κοντά στη μονάδα. Σε αυτήν την περίπτωση, η ύπαρξη του μοχλού ασκεί πολύ μεγάλη επιρροή στο γραμμικό μοντέλο, καθώς, όπως είδαμε στην προηγού-

μενη πρόταση, η αφαίρεση μόνο αυτής της μίας παρατήρησης μπορεί να οδηγήσει στην κατάρρευση ολόκληρου του μοντέλου.

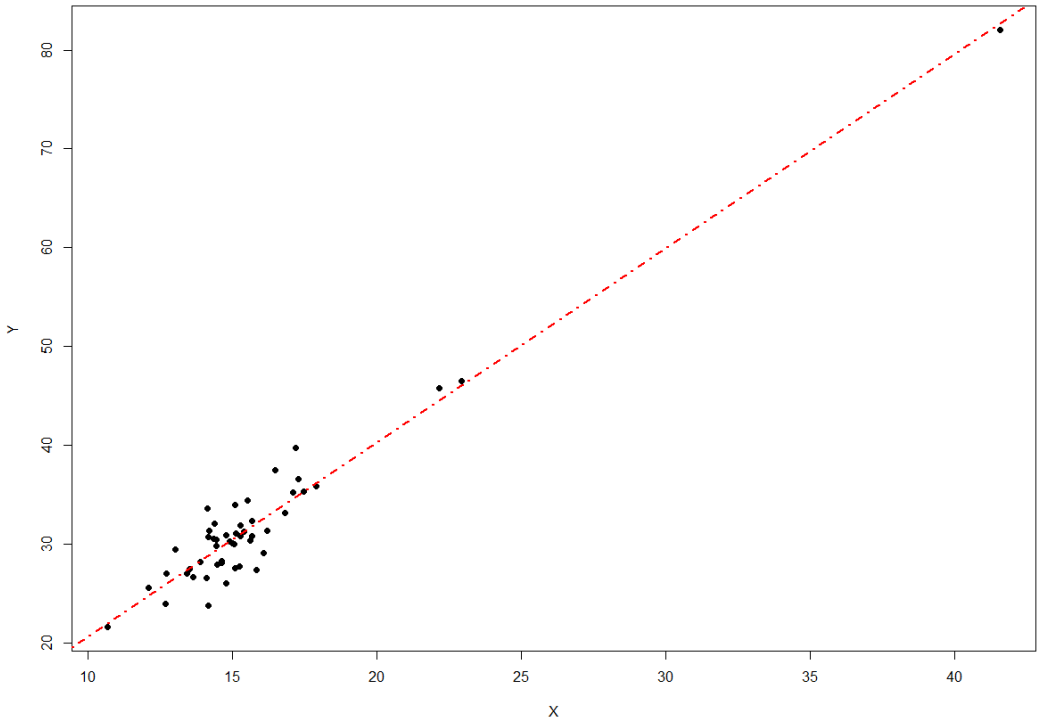
Για παράδειγμα, στο σχήμα 2.20, η παρατήρηση που απεικονίζεται στο δεξί μέρος του γραφήματος υπολογίζουμε ότι έχει μόχλευση ίση με τη μονάδα. Το αποτέλεσμα αυτό είναι και λογικό, καθώς όλες οι υπόλοιπες παρατηρήσεις είναι συγκεντρωμένες πάνω στην ίδια κατακόρυφη ευθεία. Επομένως, η σχετική απομάκρυνση της από το $\overline{X}$ θα πρέπει να είναι η μέγιστη δυνατή, δηλαδή μονάδα. Αν αφαιρούσαμε αυτήν την παρατήρηση από το δείγμα μας, τότε οι υπόλοιπες παρατηρήσεις δε θα μπορούσαν να υποστηρίξουν την κατασκευή ενός γραμμικού μοντέλου, αφού όλα τα εναπομείναντα X_i θα ήταν ίσα μεταξύ τους. Με άλλα λόγια, η παρατήρηση αυτή καθορίζει εξ ολοκλήρου την εκτιμημένη ευθεία παλινδρόμησης που είναι σχεδιασμένη με κόκκινο χρώμα.



Σχημα 2.20: Παρατήρηση με Μόχλευση Μονάδα

Μία παρατήρηση με αρκετά μεγάλη μόχλευση, αλλά όχι κοντά στη μονάδα, θα ήταν ανησυχητική μόνο στην περίπτωση όπου και η αντίστοιχη τιμή Y_i είναι πολύ απομακρυσμένη από αυτό που προβλέπει η εκτιμημένη ευθεία παλινδρόμησης. Για παράδειγμα, στο σχήμα 2.21, η παρατήρηση πάνω δεξιά έχει υπερβολικά μεγάλη απομάκρυνση από το $\overline{X}$ και αντίστοιχα μεγάλη μόχλευση. Παρόλα αυτά, όμως, η παρατήρηση αυτή συμφωνεί με τη γραμμική σχέση που ορίζεται από τις υπόλοιπες παρατηρήσεις, οπότε η απομάκρυνση της από το δείγμα δε θα

επηρέαζε σημαντικά την εκτίμηση της ευθείας παλινδρόμησης.



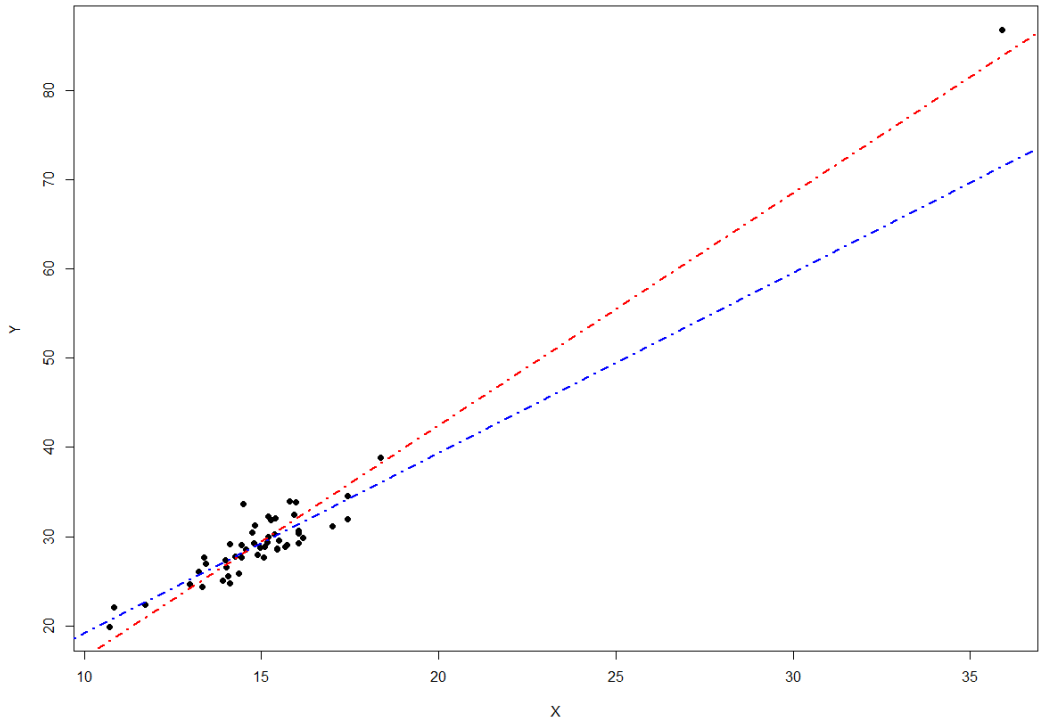
Σχημα 2.21: Σημείο Μοχλός

Αντιθέτως, στο σχήμα 2.22, βλέπουμε ένα σημείο μοχλό που ξεφεύγει πολύ από τη γραμμική τάση που επιδεικνύουν οι υπόλοιπες παρατηρήσεις, οπότε ασκεί μεγάλη επιρροή στην εκτίμηση της ευθείας παλινδρόμησης. Με κόκκινο χρώμα βλέπουμε σχεδιασμένη την εκτιμημένη ευθεία παλινδρόμησης πάνω σε ολόκληρο το δείγμα, ενώ με μπλε χρώμα βλέπουμε την εκτιμημένη ευθεία παλινδρόμησης που προκύπτει αν αφαιρέσουμε το σημείο μοχλό από το δείγμα. Διαπιστώνουμε σαφώς ότι υπάρχει πολύ μεγάλη διαφορά ανάμεσα σε αυτές τις δύο εκτιμήσεις.

Μία παρατήρηση καλείται σημείο επιρροής αν η αφαίρεσή της από το δείγμα μεταβάλλει σημαντικά την εκτίμηση των συντελεστών της παλινδρόμησης. Το σημείο μοχλός που εμφανίζεται στο σχήμα 2.22, αποτελεί προφανώς σημείο επιρροής για το γραμμικό μοντέλο που απεικονίζεται. Η μεταβολή αυτή μετριέται μέσω της **απόστασης του Cook** (Cook's distance):

$$C_i = \frac{\|\widehat{\mathbf{Y}} - \widehat{\mathbf{Y}}_{(-i)}\|^2}{(p+1)S^2},$$

όπου $\widehat{\mathbf{Y}}_{(-i)}$ είναι το διάνυσμα των προσαρμοσμένων τιμών που προκύπτει από το γραμμικό μοντέλο που προσαρμόζουμε στο δείγμα μας, έχοντας αφαιρέσει την παρατήρηση $(Y_i, \mathbf{X}_i^{\mathrm{T}})$. Μία παρατήρηση $(Y_i, \mathbf{X}_i^{\mathrm{T}})$ λέγεται **σημείο επιρροής**



Σχημα 2.22: Σημείο Επιρροής

(influential observation) αν ισχύει ότι $C_i > F_{p+1,n-p-1;0.5}$.

Για να αποφύγουμε τον υπολογισμό των προσαρμοσμένων τιμών $\widehat{\mathbf{Y}}_{(-i)}$, μπορούμε να χρησιμοποιήσουμε τον ακόλουθο τύπο για τον απευθείας υπολογισμό της απόστασης του Cook μέσω των αντίστοιχων μοχλεύσεων και εσωτερικά τυποποιημένων καταλοίπων:

$$C_i = \frac{\mathbf{P}_{i,i}}{1-\mathbf{P}_{i,i}} \cdot \frac{t_i^2}{p+1}.$$

Βλέπουμε, λοιπόν, ότι η απόσταση του Cook συνυπολογίζει τη μόχλευση και το εσωτερικά τυποποιημένο κατάλοιπο μίας παρατήρησης προκειμένου να μας δώσει ένα μέτρο της επίδρασης που έχει η παρατήρηση πάνω στην εκτίμηση της εξίσωσης παλινδρόμησης.

2.13 Χρήση Ποιοτικών Μεταβλητών

Έστω ότι έχουμε δείγμα από μία ποσοτική μεταβλητή ενδιαφέροντος Y, μία ποσοτική μεταβλητή X και μία δίτιμη ποιοτική μεταβλητή W. Για παράδειγμα η μεταβλητή Y θα μπορούσε να είναι το βάρος ενός ατόμου, η μεταβλητή X το ύψος ενός ατόμου και η μεταβλητή W το φύλο του. Αν θέλουμε να κατασκευάσουμε

ένα γραμμικό μοντέλο για την πρόβλεψη της Y μέσω των άλλων δύο μεταβλητών, τότε θα πρέπει να ακολουθήσουμε μία πιο περίπλοκη διαδικασία από αυτή που έχουμε μάθει έως τώρα, λόγω της ύπαρξης της ποιοτικής μεταβλητής W.

Η δίτιμη ποιοτική μεταβλητή W χωρίζει ουσιαστικά το δείγμα μας σε δύο ξένες μεταξύ τους ομάδες. Οι παρατηρήσεις που ανήκουν στην πρώτη ομάδα θα μπορούσαν κάλλιστα να ακολουθούν διαφορετική ευθεία παλινδρόμησης από αυτές που ανήκουν στη δεύτερη ομάδα. Για λόγους απλότητας, θα θεωρήσουμε ότι όλες οι παρατηρήσεις, ανεξαρτήτως ομάδας, έχουν κοινή διασπορά. Διαφορετικά, θα μπορούσαμε να θεωρήσουμε ότι οι παρατηρήσεις στην πρώτη ομάδα έχουν διαφορετική διασπορά από αυτές στην δεύτερη ομάδα και να κατασκευάσουμε ένα ανάλογο γραμμικό μοντέλο.

Έστω I_1 το σύνολο των δεικτών των παρατηρήσεων που ανήκουν στην πρώτη ομάδα και I_2 το σύνολο των δεικτών των παρατηρήσεων που ανήκουν στη δεύτερη ομάδα. Τότε, ορίζουμε:

$$Y_i = \begin{cases} \alpha_0 + \alpha_1 X_i + \varepsilon_i, & i \in I_1 \\ \gamma_0 + \gamma_1 X_i + \varepsilon_i, & i \in I_2 \end{cases},$$

όπου τα τυχαία σφάλματα ε_i ακολουθούν τις γνωστές υποθέσεις του κανονικού γραμμικού μοντέλου. Θέλουμε να γράψουμε αυτή τη δίκλαδη συνάρτηση στη γνωστή μορφή $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ για $i = 1, 2, \ldots, n$, ώστε να μπορούμε να δουλέψουμε πάνω στο γραμμικό μοντέλο ακριβώς όπως έχουμε μάθει έως τώρα. Προφανώς, θα πρέπει να ισχύουν τα εξής:

$$\beta_0 = \begin{cases} \alpha_0, & i \in I_1 \\ \gamma_0, & i \in I_2 \end{cases}, \quad \beta_1 = \begin{cases} \alpha_1, & i \in I_1 \\ \gamma_1, & i \in I_2 \end{cases}.$$

Προκειμένου να βρούμε μία ενιαία έκφραση για τους συντελεστές της παλινδρόμησης, ορίζουμε την ψευδομεταβλητή (dummy variable) Z ως εξής:

$$Z_i = \begin{cases} 0, & i \in I_1 \\ 1, & i \in I_2 \end{cases}.$$

Τότε, παρατηρούμε ότι $\beta_0 = \alpha_0 + (\gamma_0 - \alpha_0) Z_i$ και $\beta_1 = \alpha_1 + (\gamma_1 - \alpha_1) Z_i$. Αντικαθι-

στούμε στη γνωστή μορφή του απλού γραμμικού μοντέλου:

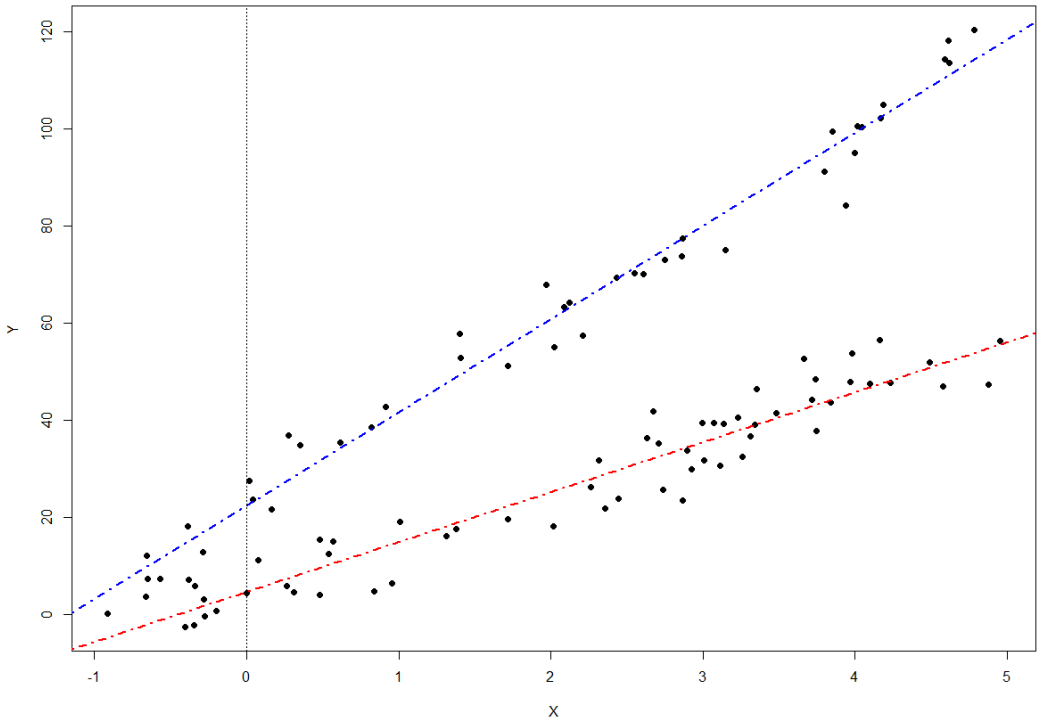
$$\begin{aligned} Y_i &= \alpha_0 + (\gamma_0 - \alpha_0)Z_i + [\alpha_1 + (\gamma_1 - \alpha_1)Z_i]\, X_i + \varepsilon_i \\ &= \underbrace{\alpha_0}_{\delta_0} + \underbrace{(\gamma_0 - \alpha_0)}_{\delta_1} Z_i + \underbrace{\alpha_1}_{\delta_2} X_i + \underbrace{(\gamma_1 - \alpha_1)}_{\delta_3} \underbrace{Z_i X_i}_{V_i} + \varepsilon_i \\ &= \delta_0 + \delta_1 Z_i + \delta_2 X_i + \delta_3 V_i + \varepsilon_i. \end{aligned}$$

Επομένως, καταλήξαμε σε ένα πολλαπλό γραμμικό μοντέλο με επεξηγηματικές μεταβλητές την ψευδομεταβλητή Z, την ποσοτική μεταβλητή και το γινόμενο των προηγούμενων δύο. Καθένας από τους 4 συντελεστές παλινδρόμησης αυτού του γραμμικού μοντέλου έχει τη δική του ερμηνεία, η οποία προκύπτει από τον τρόπο με τον οποίο κατασκευάστηκε:

- Ισχύει ότι $\delta_0 = \alpha_0$, δηλαδή η παράμετρος δ_0 εκφράζει την αναμενόμενη τιμή της Y για $X = 0$ στην πρώτη ομάδα παρατηρήσεων.
- Ισχύει ότι $\delta_1 = \gamma_0 - \alpha_0$, δηλαδή η παράμετρος δ_1 εκφράζει τη διαφορά μεταξύ της αναμενόμενης τιμής της Y για $X = 0$ στη δεύτερη ομάδα παρατηρήσεων και της αναμενόμενης τιμής της Y για $X = 0$ στην πρώτη ομάδα παρατηρήσεων.
- Ισχύει ότι $\delta_2 = \alpha_1$, δηλαδή η παράμετρος δ_2 εκφράζει τη μεταβολή στην αναμενόμενη τιμή της Y για αύξηση της X κατά μία μονάδα στην πρώτη ομάδα παρατηρήσεων.
- Ισχύει ότι $\delta_3 = \gamma_1 - \alpha_1$, δηλαδή η παράμετρος δ_3 εκφράζει τη διαφορά μεταξύ της μεταβολής στην αναμενόμενη τιμή της Y για αύξηση της X κατά μία μονάδα στη δεύτερη ομάδα παρατηρήσεων και της μεταβολής στην αναμενόμενη τιμή της Y για αύξηση της X κατά μία μονάδα στην πρώτη ομάδα παρατηρήσεων.

Αν και οι 4 συντελεστές του γραμμικού μοντέλου που ορίσαμε ήταν στατιστικά σημαντικοί και σχεδιάζαμε το γράφημα της αποκριτικής μεταβλητής Y με την επεξηγηματική μεταβλητή X, τότε θα βλέπαμε τα δεδομένα να συγκεντρώνονται ξεκάθαρα γύρω από δύο διαφορετικές ευθείες, όπως φαίνεται στο σχήμα 2.23. Αυτές οι δύο ευθείες έχουν αρκετά διαφορετικές θετικές κλίσεις και τέμνουν την κατακόρυφη ευθεία $x = 0$ σε διαφορετικά σημεία.

Αν η παράμετρος δ_3 δεν ήταν στατιστικά σημαντική, τότε θα βλέπαμε τις δύο ευθείες να είναι σχεδόν παράλληλες στο γράφημα, όπως φαίνεται στο σχήμα 2.24. Αν η παράμετρος δ_2 δεν ήταν στατιστικά σημαντική, τότε θα βλέπαμε μία από τις δύο ευθείες να είναι σχεδόν οριζόντια στο γράφημα, όπως φαίνεται στο σχήμα 2.25. Αν η παράμετρος δ_1 δεν ήταν στατιστικά σημαντική, θα βλέπαμε τις δύο



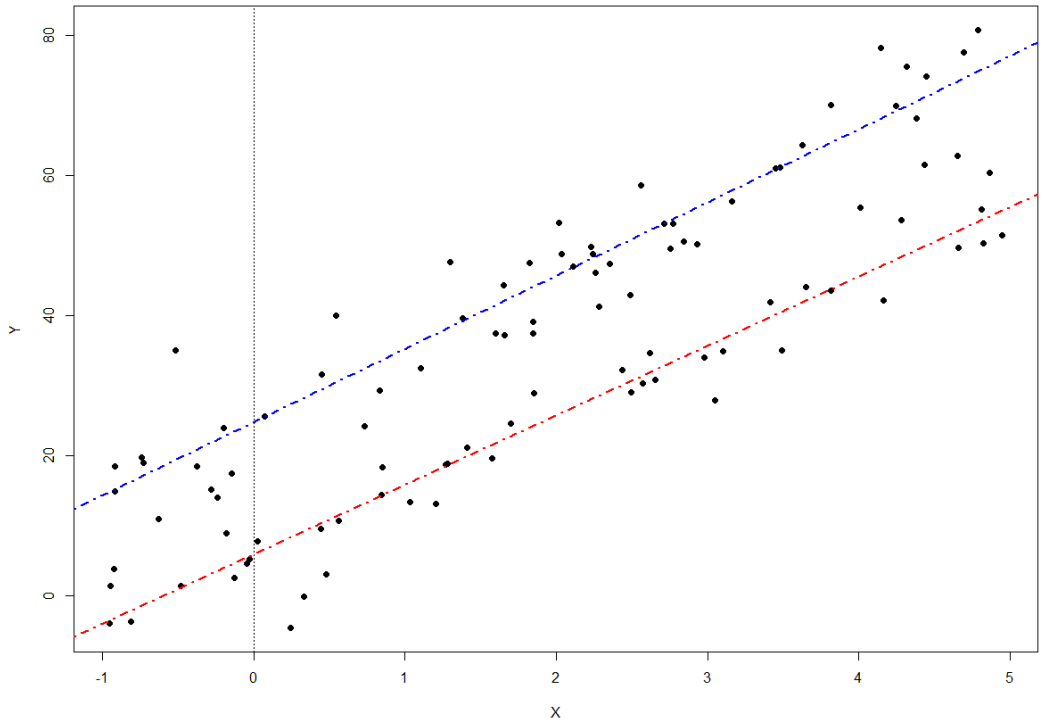
Σχημα 2.23: Δίτιμη Επεξηγηματική Μεταβλητή

ευθείες να τέμνονται σχεδόν στο ίδιο σημείο στον άξονα των x, όπως φαίνεται στο σχήμα 2.26. Επειδή θα μπορούσαμε να συναντήσουμε οποιονδήποτε συνδυασμό των παραπάνω χαρακτηριστικών στα γραφήματα όπου εμπλέκονται ποιοτικές μεταβλητές, πρέπει να γνωρίζουμε πολύ καλά την ερμηνεία καθενός από τους συντελεστές παλινδρόμησης που ορίσαμε.

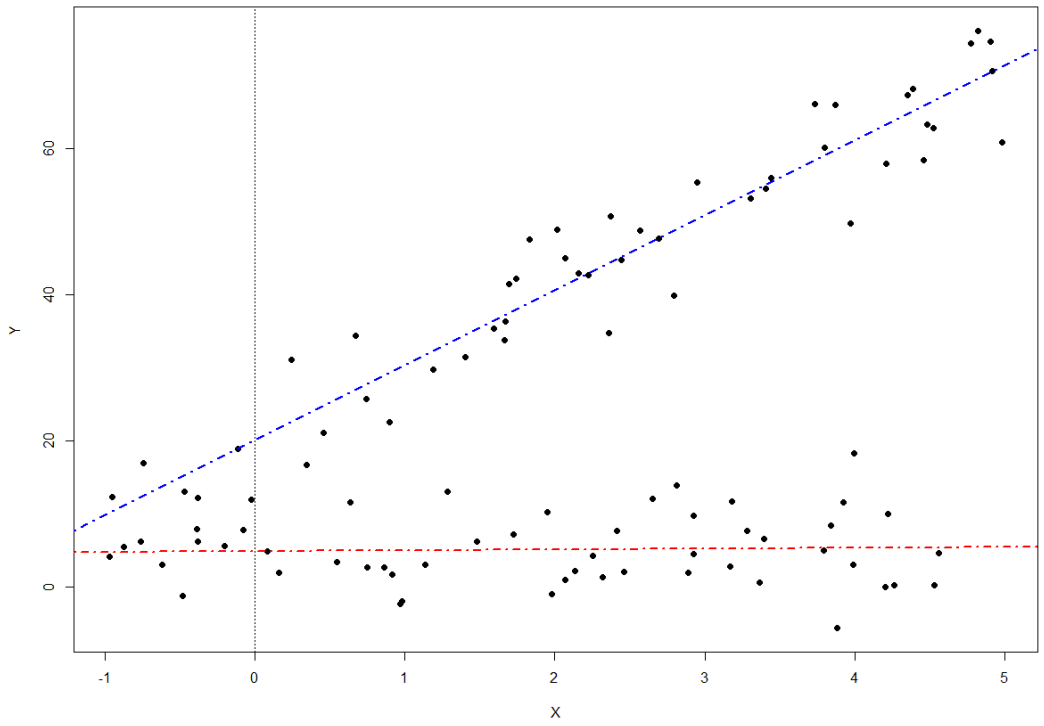
Καταλαβαίνουμε ότι αν είχαμε παραπάνω ποσοτικές επεξηγηματικές μεταβλητές, τότε η διαδικασία θα παρέμενε η ίδια και στην τελική εξίσωση παλινδρόμησης θα εμφανίζονταν τα γινόμενα της ψευδομεταβλητής Z με καθεμία από τις επεξηγηματικές μεταβλητές. Αν η ποιοτική μεταβλητή W είχε παραπάνω από δύο επίπεδα, δηλαδή έπαιρνε παραπάνω από δύο τιμές, θα ακολουθούσαμε την ίδια διαδικασία με μία διαφορά στον ορισμό της ψευδομεταβλητής Z. Ας θεωρήσουμε ότι έχει τρία επίπεδα, τα οποία ορίζουν τα σύνολα δεικτών I_1, I_2 και I_3. Τότε, θέτουμε:

$$Y_i = \begin{cases} \alpha_0 + \alpha_1 X_i + \varepsilon_i, & i \in I_1 \\ \gamma_0 + \gamma_1 X_i + \varepsilon_i, & i \in I_2 \\ \zeta_0 + \zeta_1 X_i + \varepsilon_i, & i \in I_3 \end{cases},$$

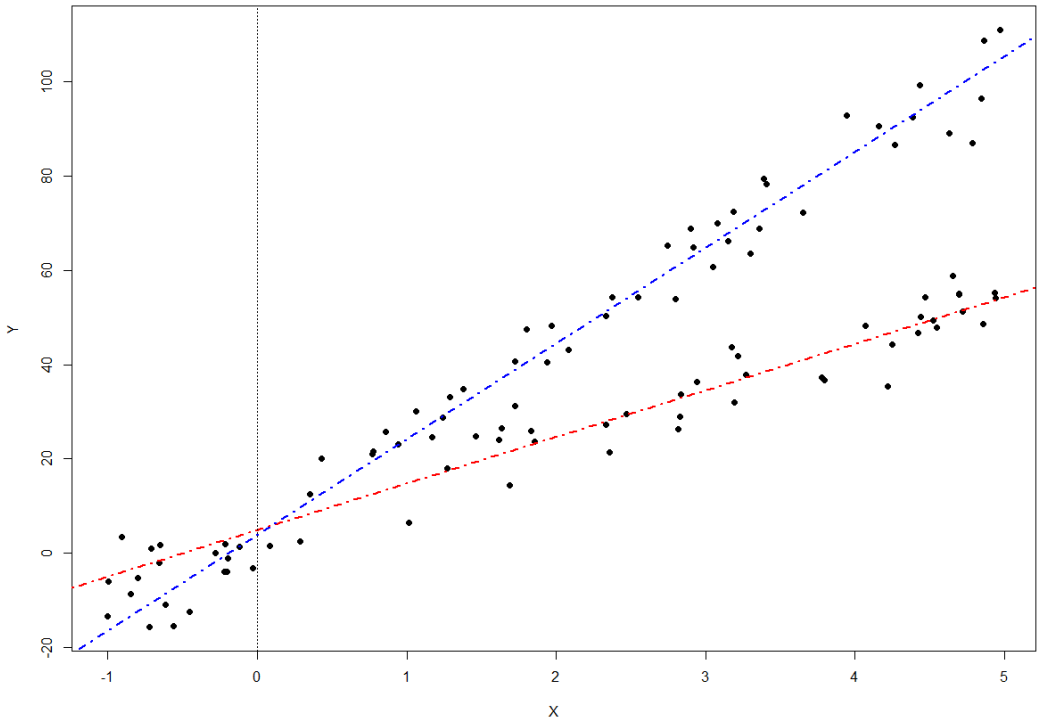
όπου τα τυχαία σφάλματα ε_i ακολουθούν τις γνωστές υποθέσεις του κανονικού



Σχημα 2.24: Παράλληλες Ευθείες Παλινδρόμησης



Σχημα 2.25: Οριζόντια Ευθεία Παλινδρόμησης



Σχημα 2.26: Τέμνουσες Ευθείες Παλινδρόμησης

γραμμικού μοντέλου. Προφανώς, θα πρέπει να ισχύουν τα εξής:

$$\beta_0 = \begin{cases} \alpha_0, & i \in I_1 \\ \gamma_0, & i \in I_2 \\ \zeta_0, & i \in I_3 \end{cases}, \quad \beta_1 = \begin{cases} \alpha_1, & i \in I_1 \\ \gamma_1, & i \in I_2 \\ \zeta_1, & i \in I_3 \end{cases}.$$

Έχοντας μία ποιοτική επεξηγηματική μεταβλητή που έχει m επίπεδα, δηλαδή παίρνει m διαφορετικές τιμές, ορίζουμε $m-1$ ψευδομεταβλητές $Z_1, Z_2, \ldots, Z_{m-1}$, προκειμένου να βρούμε μία ενιαία έκφραση για τους συντελεστές της παλινδρόμησης. Στη συγκεκριμένη περίπτωση, ορίζουμε:

$$Z_{1,i} = \begin{cases} 0, & i \notin I_1 \\ 1, & i \in I_1 \end{cases}, \quad Z_{2,i} = \begin{cases} 0, & i \notin I_2 \\ 1, & i \in I_2 \end{cases}.$$

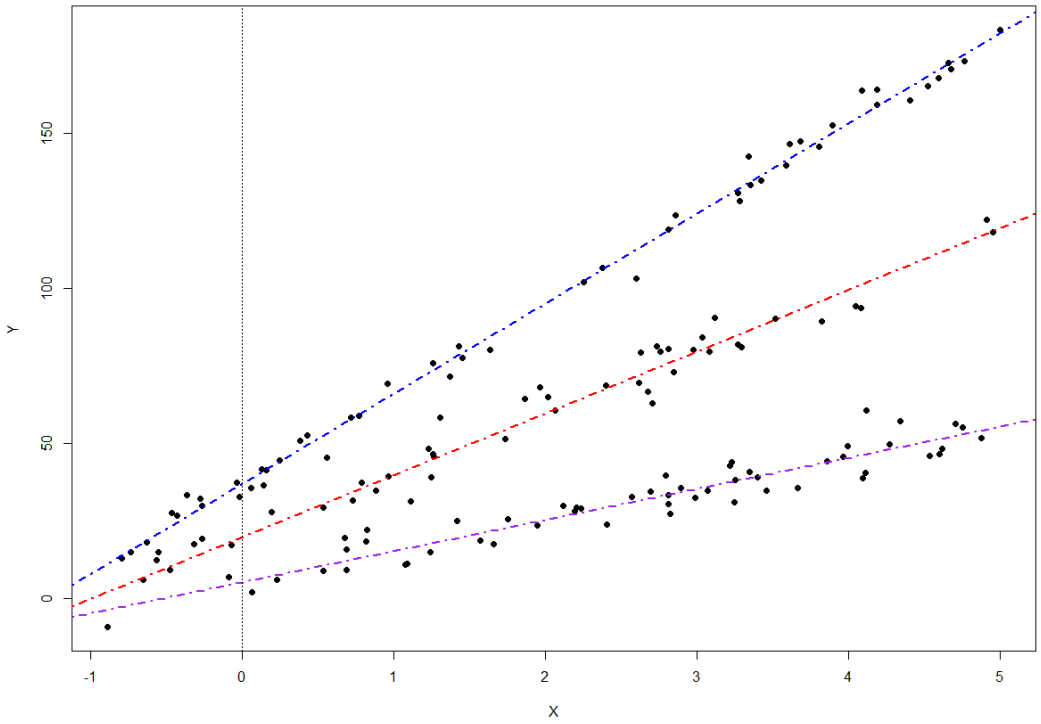
Τότε, παρατηρούμε ότι ισχύουν οι σχέσεις $\beta_0 = \zeta_0 + (\alpha_0 - \zeta_0)Z_{1,i} + (\gamma_0 - \zeta_0)Z_{2,i}$ και $\beta_1 = \zeta_1 + (\alpha_1 - \zeta_1)Z_{1,i} + (\gamma_1 - \zeta_1)Z_{2,i}$. Βλέπουμε ότι η τρίτη ομάδα παρατηρήσεων λειτουργεί ως σημείο αναφοράς, δηλαδή μία παρατήρηση (Y_i, X_i) ανήκει στην τρίτη ομάδα αν και μόνο αν $Z_{1,i} = Z_{2,i} = 0$. Αντικαθιστούμε στη γνωστή μορφή

του απλού γραμμικού μοντέλου:

$$\begin{aligned}
Y_i &= \zeta_0 + (\alpha_0 - \zeta_0)Z_{1,i} + (\gamma_0 - \zeta_0)Z_{2,i} + [\zeta_1 + (\alpha_1 - \zeta_1)Z_{1,i} + (\gamma_1 - \zeta_1)Z_{2,i}]\, X_i + \varepsilon_i \\
&= \underbrace{\zeta_0}_{\delta_0} + \underbrace{(\alpha_0 - \zeta_0)}_{\delta_1} Z_{1,i} + \underbrace{(\gamma_0 - \zeta_0)}_{\delta_2} Z_{2,i} + \underbrace{\zeta_1}_{\delta_3} X_i \\
&\qquad + \underbrace{(\alpha_1 - \zeta_1)}_{\delta_4} \underbrace{Z_{1,i}X_i}_{V_{1,i}} + \underbrace{(\gamma_1 - \zeta_1)}_{\delta_5} \underbrace{Z_{2,i}X_i}_{V_{2,i}} + \varepsilon_i \\
&= \delta_0 + \delta_1 Z_{1,i} + \delta_2 Z_{2,i} + \delta_3 X_i + \delta_4 V_{1,i} + \delta_5 V_{2,i} + \varepsilon_i.
\end{aligned}$$

Επομένως, καταλήξαμε σε ένα πολλαπλό γραμμικό μοντέλο με επεξηγηματικές μεταβλητές τις ψευδομεταβλητές Z_1, Z_2, την ποσοτική μεταβλητή και τα γινόμενα των ψευδομεταβλητών με την X. Καθένας από τους 6 συντελεστές παλινδρόμησης αυτού του γραμμικού μοντέλου έχει τη δική του ερμηνεία, η οποία προκύπτει από τον τρόπο με τον οποίο κατασκευάστηκε:

- Ισχύει ότι $\delta_0 = \zeta_0$, δηλαδή η παράμετρος δ_0 εκφράζει την αναμενόμενη τιμή της Y για $X = 0$ στην τρίτη ομάδα παρατηρήσεων.
- Ισχύει ότι $\delta_1 = \alpha_0 - \zeta_0$, δηλαδή η παράμετρος δ_1 εκφράζει τη διαφορά μεταξύ της αναμενόμενης τιμής της Y για $X = 0$ στην πρώτη ομάδα παρατηρήσεων και της αναμενόμενης τιμής της Y για $X = 0$ στην τρίτη ομάδα παρατηρήσεων.
- Ισχύει ότι $\delta_2 = \gamma_0 - \zeta_0$, δηλαδή η παράμετρος δ_1 εκφράζει τη διαφορά μεταξύ της αναμενόμενης τιμής της Y για $X = 0$ στη δεύτερη ομάδα παρατηρήσεων και της αναμενόμενης τιμής της Y για $X = 0$ στην τρίτη ομάδα παρατηρήσεων.
- Ισχύει ότι $\delta_3 = \zeta_1$, δηλαδή η παράμετρος δ_2 εκφράζει τη μεταβολή στην αναμενόμενη τιμή της Y για αύξηση της X κατά μία μονάδα στην τρίτη ομάδα παρατηρήσεων.
- Ισχύει ότι $\delta_4 = \alpha_1 - \zeta_1$, δηλαδή η παράμετρος δ_3 εκφράζει τη διαφορά μεταξύ της μεταβολής στην αναμενόμενη τιμή της Y για αύξηση της X κατά μία μονάδα στην πρώτη ομάδα παρατηρήσεων και της μεταβολής στην αναμενόμενη τιμή της Y για αύξηση της X κατά μία μονάδα στην τρίτη ομάδα παρατηρήσεων.
- Ισχύει ότι $\delta_5 = \gamma_1 - \zeta_1$, δηλαδή η παράμετρος δ_3 εκφράζει τη διαφορά μεταξύ της μεταβολής στην αναμενόμενη τιμή της Y για αύξηση της X κατά μία μονάδα στη δεύτερη ομάδα παρατηρήσεων και της μεταβολής στην αναμενόμενη τιμή της Y για αύξηση της X κατά μία μονάδα στην τρίτη ομάδα παρατηρήσεων.



Σχημα 2.27: Ποιοτική Επεξηγηματική Μεταβλητή με 3 Επίπεδα

Στο σχήμα 2.27, βλέπουμε ότι οι παρατηρήσεις συγκεντρώνονται γύρω από τρεις πολύ ευδιάκριτες ευθείες, οι οποίες έχουν όλες διαφορετικές θετικές κλίσεις και τέμνουν την οριζόντια ευθεία $x = 0$ σε διαφορετικά σημεία.

Κεφάλαιο 3

Ανάλυση Διασποράς - ANOVA

3.1 Εισαγωγή

Τα μοντέλα ανάλυσης διασποράς (ANOVA) αποτελούν ειδική περίπτωση γραμμικών μοντέλων όπου όλες οι διαθέσιμες επεξηγηματικές μεταβλητές για την αποκριτική μεταβλητή Y είναι ποιοτικές, δηλαδή παίρνουν ένα πεπερασμένο πλήθος διαφορετικών τιμών. Σε αυτήν την περίπτωση, κάθε δυνατός συνδυασμός τιμών των ποιοτικών επεξηγηματικών μεταβλητών ορίζει μία υποομάδα του πληθυσμού από τον οποίο προέρχεται το δείγμα μας. Η μεταβλητή ενδιαφέροντος Y μπορεί εν γένει να έχει διαφορετική μέση τιμή και διαφορετική διασπορά σε καθεμία από τις υποομάδες που ορίζουν οι επεξηγηματικές μεταβλητές.

Στα πλαίσια του συγγράμματος θα θεωρήσουμε ότι η αποκριτική μεταβλητή Y έχει κοινή διασπορά σε όλες τις υποομάδες, αλλά διαφορετική μέση τιμή στην καθεμία. Στόχος μας είναι να ελέγξουμε αν οι διαφορές στις μέσες τιμές των διαφορετικών υποομάδων είναι στατιστικά σημαντικές. Με κατάλληλο μετασχηματισμό των ποιοτικών μεταβλητών σε ψευδομεταβλητές και κατασκευή ενός γραμμικού μοντέλου με επεξηγηματικές μεταβλητές αυτές τις ψευδομεταβλητές, όπως περιγράφηκε στην παράγραφο 2.13, αυτό θα δούμε ότι μπορεί απλά να γίνει μέσω των ελέγχων στατιστικής σημαντικότητας των συντελεστών παλινδρόμησης του γραμμικού μοντέλου, οπότε δεν απαιτεί κάποια καινούργια θεωρία.

Ωστόσο, τα μοντέλα ανάλυσης διασποράς προτιμάται να επεξεργάζονται με έναν πιο ιδιαίτερο τρόπο, ώστε να μπορεί να ερμηνευτεί καλύτερα η συνολική συνεισφορά κάθε ποιοτικής επεξηγηματικής μεταβλητής στο γραμμικό μοντέλο και οι πιθανές αλληλεπιδράσεις που μπορεί να έχουν μεταξύ τους. Αυτό επιτυγχάνεται με την ανάλυση της συνολικής μεταβλητότητας της αποκριτικής μεταβλητής Y σε συνιστώσες και την κατασκευή πινάκων ανάλυσης διασποράς, όπως αυτούς που συναντήσαμε στη γενικότερη περίπτωση των γραμμικών μοντέλων.

3.2 ANOVA κατά Έναν Παράγοντα

Έστω ότι έχουμε μόνο μία ποιοτική επεξηγηματική μεταβλητή X στη διάθεσή μας, η οποία παίρνει m διαφορετικές τιμές. Η ποιοτική μεταβλητή X καλείται **παράγοντας** και οι διαφορετικές τιμές που παίρνει καλούνται **επίπεδα** του παράγοντα. Για λόγους απλότητας θα θεωρήσουμε ότι η μεταβλητή X παίρνει τιμές στο σύνολο $\{1, 2, \ldots, m\}$. Ορίζουμε $m-1$ ψευδομεταβλητές $Z_1, Z_2, \ldots, Z_{m-1}$:

$$Z_{k,i} = \begin{cases} 0, & X_i \neq k \\ 1, & X_i = k \end{cases}, \quad k = 1, 2, \ldots, m-1, \quad i = 1, 2, \ldots, n.$$

Βλέπουμε ότι το m-οστό επίπεδο της X λειτουργεί ως σημείο αναφοράς, δηλαδή μία παρατήρηση Y_i ανήκει στο m-οστό επίπεδο της X αν και μόνο αν $Z_{k,i} = 0$ για $k = 1, 2, \ldots, m-1$. Ορίζουμε το γραμμικό μοντέλο:

$$Y_i = \beta_0 + \beta_1 Z_{1,i} + \beta_2 Z_{2,i} + \cdots + \beta_{m-1} Z_{m-1,i} + \varepsilon_i, \quad i = 1, 2, \ldots, n.$$

Οι συντελεστές παλινδρόμησης $\beta_0, \beta_1, \ldots, \beta_{m-1}$ έχουν τις εξής ερμηνείες:

- Για $Z_1 = Z_2 = \cdots = Z_{m-1} = 0$, έχουμε $E(Y) = \beta_0$, δηλαδή ο συντελεστής β_0 εκφράζει την αναμενόμενη τιμή της εξαρτημένης μεταβλητής στο m-οστό επίπεδο του παράγοντα X.
- Για $Z_k = 1$, έχουμε $E(Y) = \beta_0 + \beta_k \Rightarrow \beta_k = E(Y) - \beta_0$, δηλαδή ο συντελεστής β_k για $k = 1, 2, \ldots, m-1$ εκφράζει τη διαφορά μεταξύ της αναμενόμενης τιμής της Y στο k-οστό επίπεδο και της αναμενόμενης τιμής της Y στο m-οστό επίπεδο.

Θα μπορούσαμε με βάση όσα έχουμε μάθει για τα γραμμικά μοντέλα να πραγματοποιήσουμε τους ελέγχους στατιστικής σημαντικότητας των συντελεστών παλινδρόμησης β_k για $k = 1, 2, \ldots, m-1$. Αν προέκυπτε ότι ο συντελεστής β_k είναι στατιστικά σημαντικός, τότε θα συμπεραίναμε ότι υπάρχει στατιστικά σημαντική διαφορά ανάμεσα στις μέσες τιμές της μεταβλητής Y στα επίπεδα k και m. Αν θέλαμε να πραγματοποιήσουμε αυτές τις συγκρίσεις με διαφορετικό επίπεδο αναφοράς, τότε θα ορίζαμε πάλι με κατάλληλο τρόπο τις ψευδομεταβλητές μας, ώστε να μην αντιστοιχεί καμία ψευδομεταβλητή στο επιθυμητό επίπεδο αναφοράς.

Ιδιαίτερο ενδιαφέρον θα είχε να κατασκευάσουμε τον πίνακα ανάλυσης διασποράς για αυτό το γραμμικό μοντέλο και να πραγματοποιήσουμε τον έλεγχο:

$$\begin{cases} H_0 : \beta_1 = \beta_2 = \cdots = \beta_{m-1} = 0 \text{ vs.} \\ H_1 : \beta_j \neq 0 \text{ για κάποιο } j \in \{1, 2, \ldots, m-1\}. \end{cases}$$

Με αυτόν τον τρόπο ελέγχουμε αν υπάρχει τουλάχιστον μία στατιστικά σημαντική διαφορά ανάμεσα στις μέσες τιμές της Y στα επίπεδα $1, 2, \ldots, m-1$ και στη μέση τιμή της Y στο επίπεδο m. Αν δεν μπορούμε να απορρίψουμε την H_0, τότε συμπεραίνουμε ότι οι μέσες τιμές της Y σε όλα τα επίπεδα δε διαφέρουν σημαντικά από τη μέση τιμή της Y στο επίπεδο m. Μπορούμε έτσι να υποθέσουμε ότι οι μέσες τιμές της Y είναι ίσες για όλα τα επίπεδα, οπότε ο παράγοντας X δε συνεισφέρει καθόλου στο μοντέλο.

	Sum of Squares	d.f.	Mean Square	$F_{m-1,n-m}$	p-value
R	$\text{SSR} = \|\widehat{\mathbf{Y}} - \overline{Y}\mathbf{1}_n\|^2$	$m-1$	$\text{MSR} = \frac{\text{SSR}}{m-1}$	$f = \frac{\text{MSR}}{\text{MSE}}$	$P(F \geqslant f)$
E	$\text{SSE} = \|\mathbf{Y} - \widehat{\mathbf{Y}}\|^2$	$n-m$	$\text{MSE} = \frac{\text{SSE}}{n-m}$		
T	$\text{SST} = \|\mathbf{Y} - \overline{Y}\mathbf{1}_n\|^2$	$n-1$			

Πινακας 3.1: Πίνακας ANOVA για το γραμμικό μοντέλο με έναν παράγοντα

Για να κατασκευάσουμε το μοντέλο ανάλυσης διασποράς που αντιστοιχεί στο παραπάνω γραμμικό μοντέλο, χωρίζουμε τις παρατηρήσεις $Y_1, Y_2, \ldots, Y_n$ στα επίπεδα $1, 2, \ldots, m$ και παίρνουμε την αναδιάταξη Y_{ij} για $i = 1, 2, \ldots, m$ και $j = 1, 2, \ldots, n_i$, όπου n_i το πλήθος των παρατηρήσεων που ανήκουν στο επίπεδο i. Προφανώς ισχύει ότι $\sum_{i=1}^{m} n_i = n$. Επιπλέον, ορίζουμε:

- Το άθροισμα των παρατηρήσεων στο επίπεδο i:

$$Y_{i\cdot} = \sum_{j=1}^{n_i} Y_{ij}, \quad i = 1, 2, \ldots, m.$$

- Τον δειγματικό μέσο των παρατηρήσεων στο επίπεδο i:

$$\overline{Y}_{i\cdot} = \frac{Y_{i\cdot}}{n_i}, \quad i = 1, 2, \ldots, m.$$

- Το συνολικό άθροισμα των παρατηρήσεων:

$$Y_{\cdot\cdot} = \sum_{i=1}^{m} \sum_{j=1}^{n_i} Y_{ij}.$$

- Τον συνολικό δειγματικό μέσο των παρατηρήσεων:

$$\overline{Y}_{\cdot\cdot} = \frac{Y_{\cdot\cdot}}{n}.$$

Τα παραπάνω συνοψίζονται στον πίνακα 3.2. Τότε, το ισοδύναμο **μοντέλο**

Επίπεδα	Παρατηρήσεις	Αθροίσματα	Δειγματικοί Μέσοι
1	$Y_{11}, Y_{12}, \ldots, Y_{1n_1}$	$Y_{1\cdot}$	$\overline{Y}_{1\cdot}$
2	$Y_{21}, Y_{22}, \ldots, Y_{2n_2}$	$Y_{2\cdot}$	$\overline{Y}_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
m	$Y_{m1}, Y_{m2}, \ldots, Y_{mn_m}$	$Y_{m\cdot}$	$\overline{Y}_{m\cdot}$
		$Y_{\cdot\cdot}$	$\overline{Y}_{\cdot\cdot}$

Πινακας 3.2: Παρατηρήσεις κατά έναν παράγοντα

ανάλυσης διασποράς κατά έναν παράγοντα γράφεται ως:

$$Y_{ij} = \mu_i + \varepsilon_{ij} = \mu + \tau_i + \varepsilon_{ij}, \quad i = 1, 2, \ldots, m, \quad j = 1, 2, \ldots, n_i, \quad \text{όπου:}$$

- μ_i η πληθυσμιακή μέση τιμή της Y στο i-οστό επίπεδο. Χρησιμοποιούμε τον αντίστοιχο δειγματικό μέσο της Y στο i-οστό επίπεδο για να εκτιμήσουμε την άγνωστη πραγματική μέση τιμή μ_i, οπότε $\widehat{Y}_{ij} = \widehat{\mu}_i = \overline{Y}_{i\cdot}$. Επιπλέον, ισχύει ότι $\mu_i = \beta_0 + \beta_i$ για $i = 1, 2, \ldots, m-1$ και $\mu_m = \beta_0$, όπου $\beta_0, \beta_1, \ldots, \beta_{m-1}$ οι συντελεστές του αντίστοιχου γραμμικού μοντέλου με επίπεδο αναφοράς το m. Συμπεραίνουμε ότι οι προσαρμοσμένες τιμές $\widehat{Y}_i$ των Y_i ταυτίζονται με τις προσαρμοσμένες τιμές $\widehat{Y}_{ij}$ των Y_{ij} μετά την αναδιάταξη.
- μ η πληθυσμιακή μέση τιμή της Y. Χρησιμοποιούμε τον συνολικό δειγματικό μέσο της Y για να εκτιμήσουμε την άγνωστη πραγματική μέση τιμή μ, οπότε $\widehat{\mu} = \overline{Y}_{\cdot\cdot}$. Προφανώς, ισχύει ότι:

$$\mu = \frac{1}{n} \sum_{i=1}^{m} n_i \mu_i.$$

- $\tau_i = \mu_i - \mu$ η επίδραση του i-οστού επιπέδου του παράγοντα στην πληθυσμιακή μέση τιμή της Y. Αντικαθιστώντας στην παραπάνω σχέση που συνδέει το μ με τα μ_i, παίρνουμε ότι $\sum_{i=1}^{m} n_i \tau_i = 0$. Επιπλέον, συμπεραίνουμε ότι $\widehat{\tau}_i = \overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot}$.
- $\varepsilon_{ij} \sim N(0, \sigma^2)$ ανεξάρτητα για $i = 1, 2, \ldots, m$ και $j = 1, 2, \ldots, n_i$. Συμπεραίνουμε ότι $Y_{ij} \sim N(\mu_i, \sigma^2)$ ανεξάρτητα για $i = 1, 2, \ldots, m$ και $j = 1, 2, \ldots, n_i$. Τα μη-παρατηρήσιμα τυχαία σφάλματα ε_{ij} εκτιμούνται από τα κατάλοιπα $\widehat{\varepsilon}_{ij} = Y_{ij} - \widehat{Y}_{ij} = Y_{ij} - \overline{Y}_{i\cdot}$.
- $\overline{Y}_{i\cdot} \sim N\left(\mu_i, \frac{\sigma^2}{n_i}\right)$ για $i = 1, 2, \ldots, m$ και $\overline{Y}_{\cdot\cdot} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$.

Όπως κάναμε και στο απλό γραμμικό μοντέλο, θα αναλύσουμε τη συνολική μεταβλητότητα των δεδομένων σε δύο κομμάτια. Το ένα κομμάτι εξηγείται από τον παράγοντα μέσω της απόκλισης των δειγματικών μέσων των επιπέδων από

τον συνολικό δειγματικό μέσο $\overline{Y}_{..}$ και το άλλο κομμάτι παραμένει ανεξήγητο και ποσοτικοποιείται μέσω των καταλοίπων $\widehat{\varepsilon}_{ij}$.

Ορισμός 3.1. (Αθροίσματα Τετραγώνων)

- Ορίζουμε $\text{SST} = \sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(Y_{ij}-\overline{Y}_{..}\right)^2$ το συνολικό άθροισμα τετραγώνων (total sum of squares) των δεδομένων.
- Ορίζουμε $\text{SSF} = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(\widehat{Y}_{ij}-\overline{Y}_{..})^2 = \sum_{i=1}^{m} n_i\left(\overline{Y}_{i.}-\overline{Y}_{..}\right)^2$ το άθροισμα τετραγώνων που οφείλεται στον παράγοντα (sum of squares due to the factor).
- Ορίζουμε $\text{SSE} = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(Y_{ij}-\widehat{Y}_{ij})^2 = \sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(Y_{ij}-\overline{Y}_{i.}\right)^2 = \sum_{i=1}^{m}\sum_{j=1}^{n_i}\widehat{\varepsilon}_{ij}^2$ το άθροισμα τετραγώνων των καταλοίπων (sum of squared errors).

Πρόταση 3.1. (Ανάλυση Διασποράς)

$$\sum_{i=1}^{m}\sum_{j=1}^{n_i}(Y_{ij}-\overline{Y}_{..})^2 = \sum_{i=1}^{m} n_i(\overline{Y}_{i.}-\overline{Y}_{..})^2+\sum_{i=1}^{m}\sum_{j=1}^{n_i}(Y_{ij}-\overline{Y}_{i.})^2, \text{ δηλαδή } \text{SST} = \text{SSF}+\text{SSE}.$$

Απόδειξη. Έχοντας στο μυαλό μας μία τεχνική που εμφανίζεται και στον υπολογισμό του μέσου τετραγωνικού σφάλματος μίας εκτιμήτριας, σκεφτόμαστε να προσθαφαιρέσουμε το $\overline{Y}_{i.}$ στην ποσότητα $Y_{ij}-\overline{Y}_{..}$ που εμφανίζεται στο SST:

$$\begin{aligned}
\text{SST} &= \sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(Y_{ij}-\overline{Y}_{..}\right)^2 = \sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(Y_{ij}-\overline{Y}_{i.}+\overline{Y}_{i.}-\overline{Y}_{..}\right)^2 \\
&= \sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(Y_{ij}-\overline{Y}_{i.}\right)^2 + \sum_{i=1}^{m} n_i\left(\overline{Y}_{i.}-\overline{Y}_{..}\right)^2 + 2\sum_{i=1}^{m}\left[\left(\overline{Y}_{i.}-\overline{Y}_{..}\right)\sum_{j=1}^{n_i}\left(Y_{ij}-\overline{Y}_{i.}\right)\right] \\
&= \text{SSE}+\text{SSF}+2\sum_{i=1}^{m}\left[\left(\overline{Y}_{i.}-\overline{Y}_{..}\right)\left(\cancel{\sum_{j=1}^{n_i}Y_{ij}}-\cancel{n_i\overline{Y}_{i.}}\right)\right] \\
&= \text{SSE}+\text{SSF}. \quad \square
\end{aligned}$$

Παρατήρηση 3.1. Σύμφωνα με την αναδιάταξη που έχουμε κάνει στα δεδομένα για να κατασκευάσουμε το μοντέλο ανάλυσης διασποράς, παρατηρούμε τα εξής:

- $\text{SST} = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(Y_{ij}-\overline{Y}_{..})^2 = \sum_{i=1}^{n}(Y_i-\overline{Y})^2 = \|\mathbf{Y}-\overline{Y}\mathbf{1}_n\|^2$.
- $\text{SSF} = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(\widehat{Y}_{ij}-\overline{Y}_{..})^2 = \sum_{i=1}^{n}(\widehat{Y}_i-\overline{Y})^2 = \|\widehat{\mathbf{Y}}-\overline{Y}\mathbf{1}_n\|^2 = \text{SSR}$.
- $\text{SSE} = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(Y_{ij}-\widehat{Y}_{ij})^2 = \sum_{i=1}^{n}(Y_i-\widehat{Y}_i)^2 = \|\mathbf{Y}-\widehat{\mathbf{Y}}\|^2$.

Με άλλα λόγια, η ανάλυση διασποράς του μοντέλου ANOVA ταυτίζεται με την ανάλυση διασποράς του αντίστοιχου γραμμικού μοντέλου. Επομένως, ο πίνακας

ANOVA που θα κατασκευάσουμε για την πραγματοποίηση του ελέγχου:

$$\begin{cases} H_0 : \mu_1 = \mu_2 = \cdots = \mu_m = \mu \text{ vs.} \\ H_1 : \mu_i \neq \mu \text{ για κάποιο } i \in \{1, 2, \ldots, m\} \end{cases}$$

ή ισοδύναμα του ελέγχου:

$$\begin{cases} H_0 : \tau_1 = \tau_2 = \cdots = \tau_m = 0 \text{ vs.} \\ H_1 : \tau_i \neq 0 \text{ για κάποιο } i \in \{1, 2, \ldots, m\} \end{cases}$$

θα ταυτίζεται με τον πίνακα ANOVA του αντίστοιχου γραμμικού μοντέλου. Προφανώς, ο έλεγχος αυτός θα είναι και ισοδύναμος με τον έλεγχο της ANOVA στο αντίστοιχο γραμμικό μοντέλο.

	Sum of Squares	d.f.	Mean Square	$F_{m-1,n-m}$	p-value
F	$\text{SSF} = \sum_{i=1}^{m} n_i \left(\overline{Y}_{i.} - \overline{Y}_{..}\right)^2$	$m-1$	$\text{MSF} = \frac{\text{SSF}}{m-1}$	$f = \frac{\text{MSF}}{\text{MSE}}$	$P(F \geqslant f)$
E	$\text{SSE} = \sum_{i=1}^{m} \sum_{j=1}^{n_i} \left(Y_{ij} - \overline{Y}_{i.}\right)^2$	$n-m$	$\text{MSE} = \frac{\text{SSE}}{n-m}$		
T	$\text{SST} = \sum_{i=1}^{m} \sum_{j=1}^{n_i} \left(Y_{ij} - \overline{Y}_{..}\right)^2$	$n-1$			

Πινακας 3.3: Πίνακας ANOVA κατά έναν παράγοντα

Λήμμα 3.1. Έστω $Q = Q_1 + Q_2$, όπου $Q \sim \chi^2_r$, $Q_1 \sim \chi^2_{r_1}$ και Q_2 ανεξάρτητη από την Q_1. Τότε, ισχύει ότι $Q_2 \sim \chi^2_{r_2}$, όπου $r_2 = r - r_1$.

Πρόταση 3.2. (Αθροίσματα Τετραγώνων)

i. Ισχύει ότι $\text{SST} = (n-1)S_Y^2$. Συμπεραίνουμε ότι $\frac{\text{SST}}{\sigma^2} = \frac{(n-1)S_Y^2}{\sigma^2} \sim \chi^2_{n-1}$ υπό την $H_0 : \mu_1 = \mu_2 = \cdots = \mu_m = \mu$.

ii. Ισχύει ότι $\text{SSE} = \sum_{i=1}^{m} (n_i - 1)S_i^2$, όπου $S_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_{i.})^2$. Συμπεραίνουμε ότι $\frac{\text{SSE}}{\sigma^2} = \sum_{i=1}^{m} \frac{(n_i-1)S_i^2}{\sigma^2} \sim \chi^2_{n-m}$.

iii. Τα SSF και SSE είναι ανεξάρτητα. Συμπεραίνουμε ότι $\frac{\text{SSF}}{\sigma^2} = \frac{\text{SSE}}{\sigma^2} - \frac{\text{SST}}{\sigma^2} \sim \chi^2_{m-1}$ και $F = \frac{n-m}{m-1} \cdot \frac{\text{SSF}}{\text{SSE}} = \frac{\text{MSF}}{\text{MSE}} \sim F_{m-1,n-m}$ υπό την $H_0 : \mu_1 = \mu_2 = \cdots = \mu_m = \mu$.

Απόδειξη.

i. Σύμφωνα με την προηγούμενη παρατήρηση, $\text{SST} = \sum_{i=1}^{n} (Y_i - \overline{Y})^2$. Υπό την $H_0 : \mu_1 = \mu_2 = \cdots = \mu_m = \mu$, ισχύει ότι $Y_{ij} \sim N(\mu, \sigma^2)$ για $i = 1, 2, \ldots, m$ και $j = 1, 2, \ldots, n_i$, δηλαδή τα $Y_1, Y_2, \ldots, Y_n$ είναι ανεξάρτητες και ισόνομες παρατηρήσεις από την κανονική κατανομή, οπότε $\frac{\text{SST}}{\sigma^2} = \frac{(n-1)S_Y^2}{\sigma^2} \sim \chi^2_{n-1}$.

ii. Το $S_i^2 = \frac{1}{n_i-1}\sum_{j=1}^{n_i}(Y_{ij}-\overline{Y}_{i.})^2$ είναι η δειγματική διασπορά της Y στο i-οστό επίπεδο. Γνωρίζουμε ότι $\frac{(n_i-1)S_i^2}{\sigma^2} \sim \chi^2_{n_i-1}$ για $i = 1, 2, \ldots, m$. Εφόσον οι παρατηρήσεις μεταξύ διαφορετικών επιπέδων είναι ανεξάρτητες, συμπεραίνουμε ότι οι δειγματικές διασπορές $S_1^2, S_2^2, \ldots, S_m^2$ είναι ανεξάρτητες μεταξύ τους. Επομένως, το $\frac{\text{SSE}}{\sigma^2} = \sum_{i=1}^{m}\frac{(n_i-1)S_i^2}{\sigma^2}$ ακολουθεί την κατανομή χ^2 με $\sum_{i=1}^{m}(n_i-1) = n-m$ βαθμούς ελευθερίας.

iii. Η ανεξαρτησία μεταξύ SSF και SSE προκύπτει με κατάλληλη εφαρμογή του θεωρήματος Cochran, όπως και στην πολλαπλή γραμμική παλινδρόμηση. Σύμφωνα με το προηγούμενο λήμμα, συμπεραίνουμε ότι $\frac{\text{SSF}}{\sigma^2} = \frac{\text{SSE}}{\sigma^2} - \frac{\text{SST}}{\sigma^2} \sim \chi^2_{m-1}$ υπό την $H_0 : \mu_1 = \mu_2 = \cdots = \mu_m = \mu$. Η κατανομή της ελεγχοσυνάρτησης $F = \frac{\text{MSF}}{\text{MSE}}$ προκύπτει άμεσα από τα παραπάνω και τον ορισμό της κατανομής F του Snedecor. □

Πρόταση 3.3. (Εκτίμηση της Διασποράς)

i. Ισχύει ότι $E(\text{MSE}) = \sigma^2$, δηλαδή το MSE είναι μία αμερόληπτη εκτιμήτρια της διασποράς σ^2 στο μοντέλο ανάλυσης διασποράς κατά έναν παράγοντα.

ii. Ισχύει ότι $E(\text{MSF}) = \sigma^2 + \frac{1}{m-1}\sum_{i=1}^{m} n_i\tau_i^2 \geqslant \sigma^2$, δηλαδή το MSF είναι μία αμερόληπτη εκτιμήτρια του σ^2 αν και μόνο αν ισχύει η μηδενική υπόθεση $H_0 : \tau_1 = \tau_2 = \cdots = \tau_m = 0$. Στη γενικότερη περίπτωση το MSF υπερεκτιμά το σ^2.

iii. Ισχύει ότι $E\left(\frac{\text{SST}}{n-1}\right) = E(S_Y^2) = \sigma^2 + \frac{1}{n-1}\sum_{i=1}^{m} n_i\tau_i^2 \geqslant \sigma^2$, δηλαδή το S_Y^2 είναι αμερόληπτη εκτιμήτρια του σ^2 αν και μόνο αν ισχύει η μηδενική υπόθεση $H_0 : \tau_1 = \tau_2 = \cdots = \tau_m = 0$. Στη γενικότερη περίπτωση το S_Y^2 υπερεκτιμά το σ^2.

Απόδειξη.

i. Ισχύει ότι $\frac{\text{SSE}}{\sigma^2} = \frac{(n-m)\text{MSE}}{\sigma^2} \sim \chi^2_{n-m}$. Επομένως, $E\left[\frac{(n-m)\text{MSE}}{\sigma^2}\right] = n-m$, δηλαδή παίρνουμε ότι $E(\text{MSE}) = \sigma^2$.

ii. Υπολογίζουμε ότι:

$$\begin{aligned}
\text{SSF} &= \sum_{i=1}^{m} n_i(\overline{Y}_{i.} - \overline{Y}_{..})^2 = \sum_{i=1}^{m} n_i\overline{Y}_{i.}^2 - 2\overline{Y}_{..}\sum_{i=1}^{m} n_i\overline{Y}_{i.} + \overline{Y}_{..}^2\sum_{i=1}^{m} n_i \\
&= \sum_{i=1}^{m} n_i\overline{Y}_{i.}^2 - 2n\overline{Y}_{..}\cdot\frac{1}{n}\sum_{i=1}^{m} Y_{i.} + n\overline{Y}_{..}^2 \\
&= \sum_{i=1}^{m} n_i\overline{Y}_{i.}^2 - 2n\overline{Y}_{..}^2 + n\overline{Y}_{..}^2 = \sum_{i=1}^{m} n_i\overline{Y}_{i.}^2 - n\overline{Y}_{..}^2.
\end{aligned}$$

$$
\begin{aligned}
E(\text{SSF}) &= \sum_{i=1}^{m} n_i E\left(\overline{Y}_{i.}^2\right) - nE\left(\overline{Y}_{..}^2\right) \\
&= \sum_{i=1}^{m} n_i \left[\text{Var}\left(\overline{Y}_{i.}\right) + \left(E\left(\overline{Y}_{i.}\right)\right)^2\right] - n\left[\text{Var}\left(\overline{Y}_{..}\right) + \left(E\left(\overline{Y}_{..}\right)\right)^2\right] \\
&= \sum_{i=1}^{m} n_i \left[\frac{\sigma^2}{n_i} + (\mu + \tau_i)^2\right] - n\left(\frac{\sigma^2}{n} + \mu^2\right) \\
&= m\sigma^2 + \cancel{n\mu^2} + \sum_{i=1}^{m} n_i\tau_i^2 + 2\mu\cancel{\sum_{i=1}^{m} n_i\tau_i} - \sigma^2 - \cancel{n\mu^2} \\
&= (m-1)\sigma^2 + \sum_{i=1}^{m} n_i\tau_i^2 \Rightarrow \\
E(\text{MSF}) &= \frac{1}{m-1}E(\text{SSF}) = \sigma^2 + \frac{1}{m-1}\sum_{i=1}^{m} n_i\tau_i^2.
\end{aligned}
$$

iii. Σχετικά με το SST, υπολογίζουμε ότι:

$$
E\left(\text{SST}\right) = E(\text{SSF}) + E(\text{SSE}) = (m-1)\sigma^2 + \sum_{i=1}^{m} n_i\tau_i^2 + (n-m)\sigma^2 \Rightarrow
$$

$$
E\left(\frac{\text{SST}}{n-1}\right) = E\left(S_Y^2\right) = \sigma^2 + \frac{1}{n-1}\sum_{i=1}^{m} n_i\tau_i^2. \quad \square
$$

Όπως και στη γραμμική παλινδρόμηση, πρέπει πάντα να πραγματοποιούμε ορισμένους διαγνωστικούς ελέγχους μετά την εκτίμηση ενός μοντέλου ανάλυσης διασποράς. Ιδιαίτερο ενδιαφέρον έχει να ελέγξουμε για πιθανή ετεροσκεδαστικότητα των τυχαίων σφαλμάτων μεταξύ διαφορετικών επιπέδων του παράγοντα. Αυτό μπορεί να γίνει μέσω του **ελέγχου Bartlett** ή του **ελέγχου Levene** για τις υποθέσεις:

$$
\begin{cases}
H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_m^2 = \sigma^2 \text{ vs.} \\
H_1 : \sigma_j^2 \neq \sigma^2 \text{ για κάποιο } j \in \{1, 2, \ldots, m\}.
\end{cases}
$$

Όλα τα στατιστικά λογισμικά, όπως η R, έχουν ενσωματωμένη τη δυνατότητα πραγματοποίησης ελέγχων ισότητας διασπορών για κατηγοριοποιημένα δεδομένα και δίνουν ως αποτέλεσμα τα p-value των ελέγχων. Εφαρμόζουμε τους ελέγχους Bartlett και Levene στα δεδομένα μας και λαμβάνουμε τα p-value. Για δεδομένο ε.σ.σ. α, έχουμε τα εξής ενδεχόμενα:

- Αν p-value $< \alpha$, τότε απορρίπτουμε την H_0, δηλαδή την υπόθεση ότι οι διασπορές της Y μεταξύ διαφορετικών επιπέδων του παράγοντα είναι ίσες.
- Αν p-value $> \alpha$, τότε δεν μπορούμε να απορρίψουμε την H_0, οπότε μπορούμε να θεωρήσουμε ότι δεν παραβαίνουμε την υπόθεση ομοσκεδαστικότητας του μοντέλου ανάλυσης διασποράς.

3.3 Συγκρίσεις Μέσων Τιμών

Ορισμός 3.2. Ορίζουμε contrast τη σύγκριση μέσων τιμών που εμπλέκει δύο ή περισσότερα επίπεδα του παράγοντα, δηλαδή $L = \sum_{i=1}^{m} c_i \mu_i$, όπου $\sum_{i=1}^{m} c_i = 0$.

Πρόταση 3.4. Η στατιστική συνάρτηση $\widehat{L} = \sum_{i=1}^{m} c_i \overline{Y}_{i.}$ είναι μία κανονικά κατανεμημένη αμερόληπτη εκτιμήτρια του L με $\mathrm{Var}(\widehat{L}) = \sigma^2 \sum_{i=1}^{m} \frac{c_i^2}{n_i}$.

Απόδειξη. Η $\widehat{L}$ είναι κανονικά κατανεμημένη ως γραμμικός συνδυασμός των $\overline{Y}_{i.}$. Υπολογίζουμε ότι:

$$E\left(\widehat{L}\right) = \sum_{i=1}^{m} c_i E\left(\overline{Y}_{i.}\right) = \sum_{i=1}^{m} c_i \mu_i = L,$$

$$\mathrm{Var}\left(\widehat{L}\right) \overset{\text{ανεξ.}}{=} \sum_{i=1}^{m} \mathrm{Var}\left(c_i \overline{Y}_{i.}\right) = \sum_{i=1}^{m} c_i^2 \mathrm{Var}\left(\overline{Y}_{i.}\right) = \sum_{i=1}^{m} c_i^2 \cdot \frac{\sigma^2}{n_i} = \sigma^2 \sum_{i=1}^{m} \frac{c_i^2}{n_i}. \quad \square$$

Πρόταση 3.5. Ένα $100(1-\alpha)\%$ διάστημα εμπιστοσύνης ίσων ουρών για το L δίνεται από τη σχέση:

$$I_{1-\alpha}(L) = \left[\widehat{L} - t_{n-m;\frac{\alpha}{2}} \cdot S_{\widehat{L}}, \widehat{L} + t_{n-m;\frac{\alpha}{2}} \cdot S_{\widehat{L}}\right], \quad \text{όπου} \quad S_{\widehat{L}}^2 = \mathrm{MSE} \cdot \sum_{i=1}^{m} \frac{c_i^2}{n_i}.$$

Απόδειξη. Γνωρίζουμε ότι:

$$Z = \frac{\widehat{L} - L}{\sigma_{\widehat{L}}} \sim N(0,1), \quad \text{όπου} \quad \sigma_{\widehat{L}}^2 = \sigma^2 \sum_{i=1}^{m} \frac{c_i^2}{n_i},$$

$$Q = \frac{\mathrm{SSE}}{\sigma^2} = \frac{(n-m)S^2}{\sigma^2} = \frac{(n-m)S_{\widehat{L}}^2}{\sigma_{\widehat{L}}^2} \sim \chi_{n-m}^2, \quad \text{όπου} \quad S^2 = \mathrm{MSE}.$$

Η Z εξαρτάται μόνο από τους δειγματικούς μέσους $\overline{Y}_{1.}, \overline{Y}_{2.}, \ldots, \overline{Y}_{m.}$, ενώ η Q εξαρτάται μόνο από τις δειγματικές διασπορές $S_1^2, S_2^2, \ldots, S_m^2$, οπότε η Z είναι ανεξάρτητη από την Q. Επομένως,

$$T = \frac{Z}{\sqrt{\frac{Q}{n-m}}} = \frac{\widehat{L} - L}{\sigma_{\widehat{L}}} \cdot \frac{\sigma_{\widehat{L}}}{S_{\widehat{L}}} = \frac{\widehat{L} - L}{S_{\widehat{L}}} \sim t_{n-m}.$$

Ζητάμε σταθερές $c_1, c_2 \in \mathbb{R}$ τέτοιες, ώστε $P(c_1 \leqslant T \leqslant c_2) = 1 - \alpha$. Συγκεκριμένα, για το διάστημα εμπιστοσύνης ίσων ουρών χρησιμοποιούμε τις σχέσεις:

$$P(T < c_1) = \frac{\alpha}{2} \Rightarrow P(T > c_1) = 1 - \frac{\alpha}{2} \Rightarrow c_1 = t_{n-m;1-\frac{\alpha}{2}} = -t_{n-m;\frac{\alpha}{2}},$$

$$P(T > c_2) = \frac{\alpha}{2} \Rightarrow c_2 = t_{n-m;\frac{\alpha}{2}}.$$

Επομένως, το διάστημα εμπιστοσύνης ίσων ουρών δίνεται από τη σχέση:

$$c_1 \leqslant \frac{\widehat{L} - L}{S_{\widehat{L}}} \leqslant c_2 \Leftrightarrow \widehat{L} - t_{n-m;\frac{\alpha}{2}} \cdot S_{\widehat{L}} \leqslant L \leqslant \widehat{L} + t_{n-m;\frac{\alpha}{2}} \cdot S_{\widehat{L}}.$$

Τελικά, παίρνουμε ότι $I_{1-\alpha}(L) = \left[\widehat{L} - t_{n-m;\frac{\alpha}{2}} \cdot S_{\widehat{L}}, \widehat{L} + t_{n-m;\frac{\alpha}{2}} \cdot S_{\widehat{L}}\right]$. □

Πρόταση 3.6. Υπό τη μηδενική υπόθεση $H_0 : L = 0$ για $i = 0, 1$, γνωρίζουμε ότι $T = \frac{\widehat{L}}{S_{\widehat{L}}} \sim t_{n-m}$. Αντικαθιστώντας τις τυχαίες μεταβλητές $Y_1, Y_2, \ldots, Y_n$ που εμφανίζονται στην ελεγχοσυνάρτηση T από τις παρατηρήσεις $y_1, y_2, \ldots, y_n$, υπολογίζουμε την παρατηρούμενη τιμή $t = \frac{\widehat{L}}{s_{\widehat{L}}}$ της ελεγχοσυνάρτησης.

i. Έστω ότι θέλουμε να πραγματοποιήσουμε τον έλεγχο με την αμφίπλευρη εναλλακτική υπόθεση $H_1 : L \neq 0$. Τότε, απορρίπτουμε την H_0 σε ε.σ.σ. α αν και μόνο αν $|t| > t_{n-m;\frac{\alpha}{2}}$ ή p-value$^{(\neq)} = P(|T| \geqslant |t|) < \alpha$ ή $0 \notin I_{1-\alpha}(L)$.

ii. Έστω ότι θέλουμε να πραγματοποιήσουμε τον έλεγχο με τη μονόπλευρη εναλλακτική υπόθεση $H_1 : L > 0$. Τότε, απορρίπτουμε την H_0 σε ε.σ.σ. α αν και μόνο αν $t > t_{n-m;\alpha}$ ή p-value$^{(>)} = P(T \geqslant t) < \alpha$.

iii. Έστω ότι θέλουμε να πραγματοποιήσουμε τον έλεγχο με τη μονόπλευρη εναλλακτική υπόθεση $H_1 : L < 0$. Τότε, απορρίπτουμε την H_0 σε ε.σ.σ. α αν και μόνο αν $t < -t_{n-m;\alpha}$ ή p-value$^{(<)} = P(T \leqslant t) < \alpha$.

Παρατήρηση 3.2. Θέτοντας $c_i = 1$ για κάποιο $i \in \{1, 2, \ldots, m\}$ και $c_j = 0$ για κάθε $j \neq i$, θα μπορούσαμε πολύ εύκολα να πάρουμε διαστήματα εμπιστοσύνης και να πραγματοποιήσουμε ελέγχους υποθέσεων που εμπλέκουν τη μέση τιμή της Y σε ένα μόνο επίπεδο του παράγοντα. Συγκεκριμένα, θα παίρναμε ότι:

$$T = \frac{\overline{Y}_{i.} - \mu_i}{S}\sqrt{n_i} \sim t_{n-m},$$

οπότε το $I_{1-\alpha}(\mu_i) = \left[\overline{Y}_{i.} - t_{n-m;\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{n_i}}, \overline{Y}_{i.} + t_{n-m;\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{n_i}}\right]$ θα ήταν το $100(1-\alpha)\%$ διάστημα εμπιστοσύνης ίσων ουρών για τη μέση τιμή της Y στο i-οστό επίπεδο. Αν θέλαμε να πραγματοποιήσουμε ελέγχους υποθέσεων με μηδενική υπόθεση την $H_0 : \mu_i = \mu_{i,0}$, τότε θα χρησιμοποιούσαμε την ελεγχοσυνάρτηση:

$$T = \frac{\overline{Y}_{i.} - \mu_{i,0}}{S}\sqrt{n_i} \sim t_{n-m}$$

και θα εφαρμόζαμε την παραπάνω πρόταση.

Παρατήρηση 3.3. (Μέθοδος Ελάχιστα Σημαντικής Διαφοράς - Least Significant Difference) Θέτοντας $c_i = 1$, $c_j = -1$ για κάποια $i, j \in \{1, 2, \ldots, m\}$ και $c_k = 0$ για

κάθε $k \neq i, j$, θα μπορούσαμε πολύ εύκολα να πάρουμε διαστήματα εμπιστοσύνης και να πραγματοποιήσουμε ελέγχους υποθέσεων που εμπλέκουν τη διαφορά των μέσω τιμών της Y σε δύο επίπεδα του παράγοντα. Συγκεκριμένα, θα παίρναμε ότι:

$$T = \frac{(\overline{Y}_{i.} - \overline{Y}_{j.}) - (\mu_i - \mu_j)}{S\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}} \sim t_{n-m}.$$

Επομένως, το $100(1-\alpha)\%$ διάστημα εμπιστοσύνης ίσων ουρών για τη διαφορά των μέσων τιμών της Y στο i-οστό και στο j-οστό επίπεδο θα ήταν:

$$I_{1-\alpha}(\mu_i - \mu_j) = \left[\overline{Y}_{i.} - \overline{Y}_{j.} - t_{n-m;\frac{\alpha}{2}} S\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}, \overline{Y}_{i.} - \overline{Y}_{j.} + t_{n-m;\frac{\alpha}{2}} S\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}\right].$$

Αν θέλαμε να πραγματοποιήσουμε ελέγχους υποθέσεων με μηδενική υπόθεση την $H_0 : \mu_i = \mu_j$, τότε θα χρησιμοποιούσαμε την ελεγχοσυνάρτηση:

$$T = \frac{\overline{Y}_{i.} - \overline{Y}_{j.}}{S\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}} \sim t_{n-m}$$

και θα εφαρμόζαμε την παραπάνω πρόταση.

Το πρόβλημα με αυτήν τη μέθοδο είναι ότι αν θέλαμε να κάνουμε πολλαπλές ταυτόχρονες συγκρίσεις μεταξύ όλων των πιθανών ζευγών μέσων τιμών της Y σε κοινό ε.σ.σ α, τότε η πιθανότητα σφάλματος τύπου I θα αυξανόταν πολύ. Βλέπουμε εύκολα ότι το συνολικό πλήθος πιθανών ζευγών μέσων τιμών είναι $k = \binom{m}{2}$. Έστω X το πλήθος των σφαλμάτων τύπου I που διαπράττουμε σε αυτούς τους k ελέγχους. Κάθε σφάλμα τύπου I διαπράττεται με πιθανότητα α. Αν επιπλέον υποθέταμε ότι κάθε σφάλμα τύπου I διαπράττεται ανεξάρτητα από τα άλλα, θα συμπεραίναμε ότι $X \sim \text{Bin}(k, \alpha)$. Επομένως,

$$P(X = 0) = (1-\alpha)^k \ll 1 - \alpha \Rightarrow P(X \geqslant 1) = 1 - P(X = 0) = 1 - (1-\alpha)^k \gg \alpha.$$

Δηλαδή, η πιθανότητα να πραγματοποιήσουμε τουλάχιστον ένα σφάλμα τύπου I ανάμεσα σε αυτούς τους k ταυτόχρονους ελέγχους θα γινόταν πολύ μεγαλύτερη από το επιθυμητό ε.σ.σ. α. Στην πραγματικότητα τα σφάλματα τύπου I δε διαπράττονται ανεξάρτητα μεταξύ τους, αλλά η πιθανότητα να πραγματοποιήσουμε τουλάχιστον ένα σφάλμα τύπου I θα είναι πάντα μεγαλύτερη από α. Το ίδιο ακριβώς πρόβλημα υπάρχει και στα ισοδύναμα διαστήματα εμπιστοσύνης.

Μέθοδος Scheffé: Είναι μία μέθοδος για την ταυτόχρονη κατασκευή διαστημάτων εμπιστοσύνης ή την ταυτόχρονη πραγματοποίηση ελέγχων υποθέσεων για όλα τα δυνατά contrasts που εμπλέκουν δύο ή περισσότερα επίπεδα του παράγοντα. Σε αντίθεση με τη μέθοδο ελάχιστα σημαντικής διαφοράς, η μέθοδος Scheffé προ-

σαρμόζει το κοινό επίπεδο σημαντικότητας των ελέγχων, ώστε το σφάλμα τύπου I για την ταυτόχρονη πραγματοποίηση όλων των ελέγχων υποθέσεων να είναι ακριβώς ίσο με α.

Πρόταση 3.7. Έστω $C = \left\{\mathbf{c} = (c_1, c_2, \ldots, c_m) \in \mathbb{R}^m : \sum_{i=1}^{m} c_i = 0\right\}$, $L_{\mathbf{c}} = \sum_{i=1}^{m} c_i \mu_i$ για $\mathbf{c} \in C$ και $c_\alpha = \sqrt{(m-1)F_{m-1,n-m;a}}$. Ένα $100(1-\alpha)\%$ ταυτόχρονο διάστημα εμπιστοσύνης ίσων ουρών για το L για κάθε $\mathbf{c} \in C$ δίνεται από τη σχέση:

$$I_{1-\alpha}^{(S)}(L_{\mathbf{c}}) = \left[\widehat{L}_{\mathbf{c}} - c_\alpha \cdot S_{\widehat{L}_{\mathbf{c}}}, \widehat{L}_{\mathbf{c}} + c_\alpha \cdot S_{\widehat{L}_{\mathbf{c}}}\right].$$

Απόδειξη. Ορίζουμε $\overline{Z}_{\cdot\cdot} = \overline{Y}_{\cdot\cdot} - \mu$ και $\overline{Z}_{i\cdot} = \overline{Y}_{i\cdot} - \mu_i$ για $i = 1, 2, \ldots, m$. Χρησιμοποιώντας την ανισότητα Cauchy-Schwarz, υπολογίζουμε ότι:

$$\begin{aligned}
P\left(L_{\mathbf{c}} \in I_{1-\alpha}^{(S)}(L_{\mathbf{c}}),\ \forall \mathbf{c} \in C\right) &= P\left(\widehat{L}_{\mathbf{c}} - c_\alpha \cdot S_{\widehat{L}_{\mathbf{c}}} \leqslant L_{\mathbf{c}} \leqslant \widehat{L}_{\mathbf{c}} + c_\alpha \cdot S_{\widehat{L}_{\mathbf{c}}},\ \forall \mathbf{c} \in C\right) \\
&= P\left(\left|\frac{\widehat{L}_{\mathbf{c}} - L_{\mathbf{c}}}{S_{\widehat{L}_{\mathbf{c}}}}\right| \leqslant c_\alpha,\ \forall \mathbf{c} \in C\right) \\
&= P\left(\max_{\mathbf{c}\in C}\left|\frac{\widehat{L}_{\mathbf{c}} - L_{\mathbf{c}}}{S_{\widehat{L}_{\mathbf{c}}}}\right| \leqslant c_\alpha\right) \\
&= P\left(\max_{\mathbf{c}\in C}\left|\frac{\sum_{i=1}^{m} c_i\left(\overline{Y}_{i\cdot} - \mu_i\right)}{S\sqrt{\sum_{i=1}^{m}\frac{c_i^2}{n_i}}}\right| \leqslant c_\alpha\right) \\
&= P\left(\max_{\mathbf{c}\in C}\left|\frac{\sum_{i=1}^{m} c_i\left[\overline{Y}_{i\cdot} - \mu_i - \left(\overline{Y}_{\cdot\cdot} - \mu\right)\right]}{S\sqrt{\sum_{i=1}^{m}\frac{c_i^2}{n_i}}}\right| \leqslant c_\alpha\right) \\
&= P\left(\frac{1}{S^2}\max_{\mathbf{c}\in C}\frac{\left[\sum_{i=1}^{m}\frac{c_i}{\sqrt{n_i}}\sqrt{n_i}\left(\overline{Z}_{i\cdot} - \overline{Z}_{\cdot\cdot}\right)\right]^2}{\sum_{i=1}^{m}\frac{c_i^2}{n_i}} \leqslant c_\alpha^2\right) \\
&= P\left(\frac{1}{S^2}\sum_{i=1}^{m} n_i\left(\overline{Z}_{i\cdot} - \overline{Z}_{\cdot\cdot}\right)^2 \leqslant (m-1)F_{m-1,n-m;a}\right) \\
&= P\left(\frac{1}{\mathrm{MSE}}\frac{\mathrm{SSF}}{m-1} \leqslant F_{m-1,n-m;a}\right) \\
&= P\left(\frac{\mathrm{MSF}}{\mathrm{MSE}} \leqslant F_{m-1,n-m;a}\right) = 1-\alpha. \quad \square
\end{aligned}$$

Σύμφωνα με τη μέθοδο Scheffé, απορρίπτουμε την $H_0 : L_{\mathbf{c}} = 0$ έναντι της εναλλακτικής $H_1 : L_{\mathbf{c}} \neq 0$ αν και μόνο αν $0 \notin I_{1-\alpha}^{(S)}(L_{\mathbf{c}}) = \left[\widehat{L}_{\mathbf{c}} - c_\alpha \cdot S_{\widehat{L}_{\mathbf{c}}}, \widehat{L}_{\mathbf{c}} + c_\alpha \cdot S_{\widehat{L}_{\mathbf{c}}}\right]$ ή $|t_{\mathbf{c}}| = \left|\frac{\widehat{L}_{\mathbf{c}}}{s_{\widehat{L}_{\mathbf{c}}}}\right| > c_\alpha$, όπου $c_\alpha = \sqrt{(m-1)F_{m-1,n-m;a}}$. Τα διαστήματα εμπιστοσύνης $I_{1-\alpha}^{(S)}(L_{\mathbf{c}})$ που προκύπτουν μέσω της μεθόδου Scheffé είναι πάντα μεγαλύτερα (πιο συντηρητικά) από τα αντίστοιχα μεμονωμένα διαστήματα εμπιστοσύνης $I_{1-\alpha}(L_{\mathbf{c}})$, οπότε αφήνουν μεγαλύτερη αβεβαιότητα για την εκτιμώμενη ποσότητα L. Συνήθως μας ενδιαφέρει να πραγματοποιήσουμε ταυτόχρονους ελέγχους υποθέσεων

μόνο για ένα μικρό πλήθος από contrasts. Σε αυτήν την περίπτωση, η μέθοδος Scheffé δίνει πολύ συντηρητικά διαστήματα εμπιστοσύνης και το σφάλμα τύπου I για την ταυτόχρονη πραγματοποίηση όλων των ελέγχων υποθέσεων είναι πολύ μικρότερο από το ζητούμενο α.

Ειλικρινά Σημαντική Διαφορά - Tukey's Honest Significant Difference: Είναι μία μέθοδος για την ταυτόχρονη κατασκευή διαστημάτων εμπιστοσύνης ή την ταυτόχρονη πραγματοποίηση ελέγχων υποθέσεων για όλα τα δυνατά contrasts που εμπλέκουν δύο επίπεδα του παράγοντα. Σε αντίθεση με τη μέθοδο ελάχιστα σημαντικής διαφοράς, η μέθοδος ειλικρινά σημαντικής διαφοράς προσαρμόζει το κοινό επίπεδο σημαντικότητας των ελέγχων, ώστε το σφάλμα τύπου I για την ταυτόχρονη πραγματοποίηση όλων των ελέγχων υποθέσεων να είναι ακριβώς ίσο με α. Θεωρούμε ότι $n_1 = n_2 = \cdots = n_m$. Για $i = 1, 2, \ldots, m$, ορίζουμε:

$$T_i = \frac{\overline{Y}_{i.} - \mu_i}{S}\sqrt{n_1} \sim t_{n-m}.$$

Τότε, παρατηρούμε ότι:

$$T_{ij} = \frac{(\overline{Y}_{i.} - \overline{Y}_{j.}) - (\mu_i - \mu_j)}{S\sqrt{\frac{2}{n_1}}} = \frac{T_i - T_j}{\sqrt{2}} \Rightarrow |T_{ij}| \leqslant \frac{\max_i T_i - \min_i T_i}{\sqrt{2}}.$$

Η τυχαία μεταβλητή $R = \max_i T_i - \min_i T_i$ ακολουθεί την κατανομή studentized range με παραμέτρους m και $n - m$. Αν $R_{m,n-m;\alpha}$ είναι το άνω α-ποσοστιαίο σημείο αυτής της κατανομής, τότε ένα $100(1-\alpha)\%$ ταυτόχρονο διάστημα εμπιστοσύνης ίσων ουρών για τη διαφορά μέσων τιμών $\mu_i - \mu_j$ με $i \neq j$ δίνεται από τη σχέση:

$$I_{1-\alpha}^{(HSD)}(\mu_i - \mu_j) = \left[\overline{Y}_{i.} - \overline{Y}_{j.} - R_{m,n-m;\alpha}\frac{S}{\sqrt{n_1}}, \overline{Y}_{i.} - \overline{Y}_{j.} + R_{m,n-m;\alpha}\frac{S}{\sqrt{n_1}}\right].$$

3.4 ANOVA κατά Δύο Παράγοντες χωρίς Αλληλεπίδραση

Έστω τώρα ότι έχουμε έναν παράγοντα A με m επίπεδα και έναν παράγοντα B με ℓ επίπεδα. Θεωρούμε ότι δεν υπάρχει αλληλεπίδραση μεταξύ των δύο παραγόντων, δηλαδή η επίδραση του παράγοντα A στη μέση τιμή της αποκριτικής μεταβλητής Y δεν εξαρτάται από το επίπεδο του παράγοντα B και η επίδραση του παράγοντα B δεν εξαρτάται από το επίπεδο του παράγοντα A. Για λόγους απλότητας θεωρούμε ότι έχουμε μόνο μία παρατήρηση για κάθε συνδυασμό επιπέδων των δύο παραγόντων, δηλαδή $n = m\ell$. Γενικότερα, θα μπορούσαμε να θεωρήσουμε ότι έχουμε n_{ij} παρατηρήσεις από τον συνδυασμό του i-οστού επιπέδου του παράγοντα A και του j-οστού επιπέδου του παράγοντα B. Έστω Y_{ij} η

μία παρατήρηση από τον συνδυασμό του i-οστού επιπέδου του παράγοντα A και του j-οστού επιπέδου του παράγοντα B. Ορίζουμε:

- Το άθροισμα των παρατηρήσεων στο επίπεδο i του παράγοντα A:

$$Y_{i.} = \sum_{j=1}^{\ell} Y_{ij}, \quad i = 1, 2, \ldots, m.$$

- Τον δειγματικό μέσο των παρατηρήσεων στο επίπεδο i του παράγοντα A:

$$\overline{Y}_{i.} = \frac{Y_{i.}}{\ell}, \quad i = 1, 2, \ldots, m.$$

- Το άθροισμα των παρατηρήσεων στο επίπεδο j του παράγοντα B:

$$Y_{.j} = \sum_{i=1}^{m} Y_{ij}, \quad j = 1, 2, \ldots, \ell.$$

- Τον δειγματικό μέσο των παρατηρήσεων στο επίπεδο j του παράγοντα B:

$$\overline{Y}_{.j} = \frac{Y_{j.}}{m}, \quad j = 1, 2, \ldots, \ell.$$

- Το συνολικό άθροισμα των παρατηρήσεων:

$$Y_{..} = \sum_{i=1}^{m} \sum_{j=1}^{\ell} Y_{ij}.$$

- Τον συνολικό δειγματικό μέσο των παρατηρήσεων:

$$\overline{Y}_{..} = \frac{Y_{..}}{m\ell}.$$

Τα παραπάνω συνοψίζονται στον πίνακα 3.4. Τότε, το **μοντέλο ανάλυσης διασποράς κατά δύο παράγοντες** γράφεται ως:

$$Y_{ij} = \mu_{ij} + \varepsilon_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij} = \mu_{i.} + \mu_{.j} - \mu + \varepsilon_{ij}$$

για $i = 1, 2, \ldots, m$ και $j = 1, 2, \ldots, \ell$, όπου:

- $\mu_{i.}$ η πληθυσμιακή μέση τιμή της Y στο i-οστό επίπεδο του παράγοντα A. Χρησιμοποιούμε τον αντίστοιχο δειγματικό μέσο της Y στο i-οστό επίπεδο του A για να εκτιμήσουμε την άγνωστη πραγματική μέση τιμή $\mu_{i.}$, οπότε $\widehat{\mu}_{i.} = \overline{Y}_{i.}$.

		Παράγοντας B					
	Επίπεδα	1	2	$\cdots$	ℓ	Αθροίσματα	Μέσοι
Παράγοντας A	1	Y_{11}	Y_{12}	$\cdots$	$Y_{1\ell}$	$Y_{1.}$	$\overline{Y}_{1.}$
	2	Y_{21}	Y_{22}	$\cdots$	$Y_{2\ell}$	$Y_{2.}$	$\overline{Y}_{2.}$
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$
	m	Y_{m1}	Y_{m2}	$\cdots$	$Y_{m\ell}$	$Y_{m.}$	$\overline{Y}_{m.}$
	Αθροίσματα	$Y_{.1}$	$Y_{.2}$	$\cdots$	$Y_{.\ell}$	$Y_{..}$	
	Μέσοι	$\overline{Y}_{.1}$	$\overline{Y}_{.2}$	$\cdots$	$\overline{Y}_{.\ell}$		$\overline{Y}_{..}$

Πινακας 3.4: Παρατηρήσεις κατά δύο παράγοντες χωρίς αλληλεπίδραση

- $\mu_{.j}$ η πληθυσμιακή μέση τιμή της Y στο j-οστό επίπεδο του παράγοντα B. Χρησιμοποιούμε τον αντίστοιχο δειγματικό μέσο της Y στο j-οστό επίπεδο του B για να εκτιμήσουμε την άγνωστη πραγματική μέση τιμή $\mu_{.j}$, οπότε $\widehat{\mu}_{.j} = \overline{Y}_{.j}$.
- μ η πληθυσμιακή μέση τιμή της Y. Χρησιμοποιούμε τον συνολικό δειγματικό μέσο της Y για να εκτιμήσουμε την άγνωστη πραγματική μέση τιμή μ, οπότε $\widehat{\mu} = \overline{Y}_{..}$. Προφανώς, ισχύει ότι:

$$\mu = \frac{1}{m\ell}\sum_{i=1}^{m}\sum_{j=1}^{\ell}\mu_{ij} = \frac{1}{m}\sum_{i=1}^{m}\mu_{i.} = \frac{1}{\ell}\sum_{j=1}^{\ell}\mu_{.j}.$$

- $\alpha_i = \mu_{i.} - \mu$ η επίδραση του i-οστού επιπέδου του παράγοντα A στην πληθυσμιακή μέση τιμή της Y. Αντικαθιστώντας στην παραπάνω σχέση, παίρνουμε ότι $\sum_{i=1}^{m}\alpha_i = 0$. Επιπλέον, συμπεραίνουμε ότι $\widehat{\alpha}_i = \overline{Y}_{i.} - \overline{Y}_{..}$.
- $\beta_j = \mu_{.j} - \mu$ η επίδραση του j-οστού επιπέδου του παράγοντα B στην πληθυσμιακή μέση τιμή της Y. Αντικαθιστώντας στην παραπάνω σχέση, παίρνουμε ότι $\sum_{j=1}^{\ell}\beta_j = 0$. Επιπλέον, συμπεραίνουμε ότι $\widehat{\beta}_j = \overline{Y}_{.j} - \overline{Y}_{..}$.
- $\varepsilon_{ij} \sim N(0, \sigma^2)$ ανεξάρτητα για $i = 1, 2, \ldots, m$ και $j = 1, 2, \ldots, \ell$. Συμπεραίνουμε ότι $Y_{ij} \sim N(\mu_{ij}, \sigma^2)$ ανεξάρτητα για $i = 1, 2, \ldots, m$ και $j = 1, 2, \ldots, \ell$. Οι προσαρμοσμένες τιμές δίνονται ως $\widehat{Y}_{ij} = \widehat{\mu}_{i.} + \widehat{\mu}_{.j} - \widehat{\mu} = \overline{Y}_{i.} + \overline{Y}_{.j} - \overline{Y}_{..}$, ενώ τα μη-παρατηρήσιμα τυχαία σφάλματα ε_{ij} εκτιμούνται από τα κατάλοιπα $\widehat{\varepsilon}_{ij} = Y_{ij} - \widehat{Y}_{ij} = Y_{ij} - \overline{Y}_{i.} - \overline{Y}_{.j} + \overline{Y}_{..}$.
- $\overline{Y}_{i.} \sim N\left(\mu_{i.}, \frac{\sigma^2}{\ell}\right)$ για $i = 1, 2, \ldots, m$, $\overline{Y}_{.j} \sim N\left(\mu_{.j}, \frac{\sigma^2}{m}\right)$ για $j = 1, 2, \ldots, \ell$ και $\overline{Y}_{..} \sim N\left(\mu, \frac{\sigma^2}{m\ell}\right)$.

Δουλεύοντας ομοίως με το μοντέλο ανάλυσης διασποράς κατά έναν παράγο-

ντα, θα αναλύσουμε τη συνολική μεταβλητότητα των δεδομένων σε τρία κομμάτια. Το πρώτο κομμάτι εξηγείται από τον παράγοντα A μέσω της απόκλισης των δειγματικών μέσων $\overline{Y}_{i.}$ από τον συνολικό δειγματικό μέσο $\overline{Y}_{..}$, το δεύτερο κομμάτι εξηγείται από τον παράγοντα B μέσω της απόκλισης των δειγματικών μέσων $\overline{Y}_{.j}$ από τον συνολικό δειγματικό μέσο $\overline{Y}_{..}$ και το τρίτο κομμάτι παραμένει ανεξήγητο και ποσοτικοποιείται μέσω των καταλοίπων $\widehat{\varepsilon}_{ij}$.

Ορισμός 3.3. (Αθροίσματα Τετραγώνων)

- Ορίζουμε $\text{SST} = \sum_{i=1}^{m}\sum_{j=1}^{\ell}\left(Y_{ij}-\overline{Y}_{..}\right)^2$ το συνολικό άθροισμα τετραγώνων (total sum of squares) των δεδομένων.
- Ορίζουμε $\text{SSA} = \ell\sum_{i=1}^{m}\left(\overline{Y}_{i.}-\overline{Y}_{..}\right)^2$ το άθροισμα τετραγώνων που οφείλεται στον παράγοντα A (sum of squares due to factor A).
- Ορίζουμε $\text{SSB} = m\sum_{j=1}^{\ell}\left(\overline{Y}_{.j}-\overline{Y}_{..}\right)^2$ το άθροισμα τετραγώνων που οφείλεται στον παράγοντα B (sum of squares due to factor B).
- Ορίζουμε $\text{SSE} = \sum_{i=1}^{m}\sum_{j=1}^{\ell}(Y_{ij}-\widehat{Y}_{ij})^2 = \sum_{i=1}^{m}\sum_{j=1}^{\ell}\left(Y_{ij}-\overline{Y}_{i.}-\overline{Y}_{.j}+\overline{Y}_{..}\right)^2$ το άθροισμα τετραγώνων των καταλοίπων (sum of squared errors).

Πρόταση 3.8. (Ανάλυση Διασποράς)

$$\sum_{i=1}^{m}\sum_{j=1}^{\ell}\left(Y_{ij}-\overline{Y}_{..}\right)^2 = \ell\sum_{i=1}^{m}\left(\overline{Y}_{i.}-\overline{Y}_{..}\right)^2 + m\sum_{j=1}^{\ell}\left(\overline{Y}_{.j}-\overline{Y}_{..}\right)^2 + \sum_{i=1}^{m}\sum_{j=1}^{\ell}\left(Y_{ij}-\overline{Y}_{i.}-\overline{Y}_{.j}+\overline{Y}_{..}\right)^2, \text{ δηλαδή:}$$

$$\text{SST} = \text{SSA} + \text{SSB} + \text{SSE}.$$

Απόδειξη. Αρκεί να δείξουμε τα εξής:

$$\text{CPAB} = \sum_{i=1}^{m}\sum_{j=1}^{\ell}\left(\overline{Y}_{i.}-\overline{Y}_{..}\right)\left(\overline{Y}_{.j}-\overline{Y}_{..}\right) = \sum_{i=1}^{m}\left[\left(\overline{Y}_{i.}-\overline{Y}_{..}\right)\left(\sum_{j=1}^{\ell}\frac{Y_{.j}}{m}-\frac{Y_{..}}{m}\right)\right] = 0,$$

$$\begin{aligned}\text{CPAE} &= \sum_{i=1}^{m}\sum_{j=1}^{\ell}\left(\overline{Y}_{i.}-\overline{Y}_{..}\right)\left(Y_{ij}-\overline{Y}_{i.}-\overline{Y}_{.j}+\overline{Y}_{..}\right)\\ &= \sum_{i=1}^{m}\left[\left(\overline{Y}_{i.}-\overline{Y}_{..}\right)\left(\sum_{j=1}^{\ell}Y_{ij}-\ell\overline{Y}_{i.}-\sum_{j=1}^{\ell}\overline{Y}_{.j}+\ell\overline{Y}_{..}\right)\right]\\ &= \sum_{i=1}^{m}\left[\left(\overline{Y}_{i.}-\overline{Y}_{..}\right)\left(\frac{Y_{..}}{m}-\sum_{j=1}^{\ell}\frac{Y_{.j}}{m}\right)\right] = 0 \quad \text{και}\end{aligned}$$

$$\begin{aligned}\text{CPBE} &= \sum_{i=1}^{m}\sum_{j=1}^{\ell}\left(\overline{Y}_{\cdot j}-\overline{Y}_{\cdot\cdot}\right)\left(Y_{ij}-\overline{Y}_{i\cdot}-\overline{Y}_{\cdot j}+\overline{Y}_{\cdot\cdot}\right)\\ &= \sum_{j=1}^{\ell}\left[\left(\overline{Y}_{\cdot j}-\overline{Y}_{\cdot\cdot}\right)\left(\cancel{\sum_{i=1}^{m}Y_{ij}}-\sum_{i=1}^{m}\overline{Y}_{i\cdot}-\cancel{m\overline{Y}_{\cdot j}}+m\overline{Y}_{\cdot\cdot}\right)\right]\\ &= \sum_{j=1}^{\ell}\left[\left(\overline{Y}_{\cdot j}-\overline{Y}_{\cdot\cdot}\right)\left(\cancel{\frac{Y_{\cdot\cdot}}{\ell}}-\cancel{\sum_{i=1}^{m}\frac{Y_{i\cdot}}{\ell}}\right)\right]=0. \quad \square\end{aligned}$$

Ενδιαφερόμαστε να πραγματοποιήσουμε τους δύο παρακάτω ελέγχους υποθέσεων:

$$\begin{cases}H_0 : \alpha_1=\alpha_2=\cdots=\alpha_m=0 \text{ vs.}\\ H_1 : \alpha_i\neq 0 \text{ για κάποιο } i\in\{1,2,\ldots,m\},\end{cases}$$

$$\begin{cases}H_0 : \beta_1=\beta_2=\cdots=\beta_\ell=0 \text{ vs.}\\ H_1 : \beta_j\neq 0 \text{ για κάποιο } j\in\{1,2,\ldots,\ell\}.\end{cases}$$

Ο πρώτος έλεγχος μας βοηθάει να αποφασίσουμε αν υπάρχει στατιστικά σημαντική συνεισφορά του παράγοντα A στην πρόβλεψη της μέσης τιμής της αποκριτικής μεταβλητής Y, ενώ ο δεύτερος έλεγχος να αποφασίσουμε αν υπάρχει στατιστικά σημαντική συνεισφορά του παράγοντα B.

	Sum of Squares	d.f.	Mean Square	F
A	$\ell\sum_{i=1}^{m}\left(\overline{Y}_{i\cdot}-\overline{Y}_{\cdot\cdot}\right)^2$	$m-1$	$\frac{\text{SSA}}{m-1}$	$\frac{\text{MSA}}{\text{MSE}}$
B	$m\sum_{j=1}^{\ell}\left(\overline{Y}_{\cdot j}-\overline{Y}_{\cdot\cdot}\right)^2$	$\ell-1$	$\frac{\text{SSB}}{\ell-1}$	$\frac{\text{MSB}}{\text{MSE}}$
E	$\sum_{i=1}^{m}\sum_{j=1}^{\ell}\left(Y_{ij}-\overline{Y}_{i\cdot}-\overline{Y}_{\cdot j}+\overline{Y}_{\cdot\cdot}\right)^2$	$(m-1)(\ell-1)$	$\frac{\text{SSE}}{(m-1)(\ell-1)}$	
T	$\sum_{i=1}^{m}\sum_{j=1}^{\ell}\left(Y_{ij}-\overline{Y}_{\cdot\cdot}\right)^2$	$m\ell-1$		

Πινακας 3.5: Πίνακας ANOVA κατά δύο παράγοντες χωρίς αλληλεπίδραση

Πρόταση 3.9. (Αθροίσματα Τετραγώνων)

i. Ισχύει ότι $\frac{\text{SSA}}{\sigma^2}\sim\chi^2_{m-1}$ υπό την $H_0 : \alpha_1=\alpha_2=\cdots=\alpha_m=0$ και $\frac{\text{SSB}}{\sigma^2}\sim\chi^2_{\ell-1}$ υπό την $H_0 : \beta_1=\beta_2=\cdots=\beta_\ell=0$.

ii. Ισχύει ότι $\frac{\text{SSE}}{\sigma^2}\sim\chi^2_{(m-1)(\ell-1)}$. Επιπλέον, τα SSA, SSB και SSE είναι ανεξάρτητα.

iii. Συμπεραίνουμε ότι $F_A=\frac{(m-1)(\ell-1)}{m-1}\cdot\frac{\text{SSA}}{\text{SSE}}=\frac{\text{MSA}}{\text{MSE}}\sim F_{m-1,(m-1)(\ell-1)}$ υπό την $H_0 : \alpha_1=\alpha_2=\cdots=\alpha_m=0$ και $F_B=\frac{(m-1)(\ell-1)}{\ell-1}\cdot\frac{\text{SSB}}{\text{SSE}}=\frac{\text{MSB}}{\text{MSE}}\sim F_{\ell-1,(m-1)(\ell-1)}$ υπό την $H_0 : \beta_1=\beta_2=\cdots=\beta_\ell=0$.

Απόδειξη.

i. Αγνοώντας την ύπαρξη του παράγοντα B και εργαζόμενοι όπως στην πρόταση 3.2, προκύπτει ότι $\frac{\text{SSA}}{\sigma^2} \sim \chi^2_{m-1}$ υπό την $H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_m = 0$. Ομοίως, αγνοώντας την ύπαρξη του παράγοντα A και εργαζόμενοι ακριβώς όπως στην πρόταση 3.2, προκύπτει ότι $\frac{\text{SSB}}{\sigma^2} \sim \chi^2_{\ell-1}$ υπό τη μηδενική υπόθεση $H_0 : \beta_1 = \beta_2 = \cdots = \beta_\ell = 0$.

ii. Η κατανομή του $\frac{\text{SSE}}{\sigma^2}$ και η ανεξαρτησία μεταξύ SSA, SSB και SSE προκύπτουν με κατάλληλη εφαρμογή του θεωρήματος Cochran, όπως και στην πολλαπλή γραμμική παλινδρόμηση.

iii. Οι κατανομές των τυχαίων μεταβλητών F_A και F_B προκύπτουν άμεσα από τα παραπάνω και τον ορισμό της κατανομής F του Snedecor. □

Πρόταση 3.10. (Εκτίμηση της Διασποράς)

i. Ισχύει ότι $E(\text{MSE}) = \sigma^2$, δηλαδή το MSE είναι μία αμερόληπτη εκτιμήτρια της διασποράς σ^2 στο μοντέλο ανάλυσης διασποράς κατά έναν παράγοντα.

ii. $E(\text{MSA}) = \sigma^2 + \frac{\ell}{m-1}\sum_{i=1}^{m}\alpha_i^2 \geqslant \sigma^2$ και $E(\text{MSB}) = \sigma^2 + \frac{m}{\ell-1}\sum_{j=1}^{\ell}\beta_j^2 \geqslant \sigma^2$, δηλαδή το MSA είναι αμερόληπτη εκτιμήτρια του σ^2 αν και μόνο αν ισχύει η μηδενική υπόθεση $H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_m = 0$ και το MSB είναι αμερόληπτη εκτιμήτρια του σ^2 αν και μόνο αν ισχύει η $H_0 : \beta_1 = \beta_2 = \cdots = \beta_\ell = 0$. Στη γενικότερη περίπτωση τα MSA και MSB υπερεκτιμούν το σ^2.

iii. $E\left(\frac{\text{SST}}{m\ell-1}\right) = E(S_Y^2) = \sigma^2 + \frac{\ell}{m\ell-1}\sum_{i=1}^{m}\alpha_i^2 + \frac{m}{m\ell-1}\sum_{j=1}^{\ell}\beta_j^2 \geqslant \sigma^2$, δηλαδή το S_Y^2 είναι αμερόληπτη εκτιμήτρια του σ^2 αν και μόνο αν ισχύουν οι μηδενικές υποθέσεις $H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_m = 0$ και $H_0 : \beta_1 = \beta_2 = \cdots = \beta_\ell = 0$ ταυτόχρονα. Στη γενικότερη περίπτωση το S_Y^2 υπερεκτιμά το σ^2.

Απόδειξη.

i. Ισχύει ότι $\frac{\text{SSE}}{\sigma^2} = \frac{(m-1)(\ell-1)\text{MSE}}{\sigma^2} \sim \chi^2_{(m-1)(\ell-1)}$. Επομένως, συμπεραίνουμε ότι $E\left[\frac{(m-1)(\ell-1)\text{MSE}}{\sigma^2}\right] = (m-1)(\ell-1)$, δηλαδή $E(\text{MSE}) = \sigma^2$.

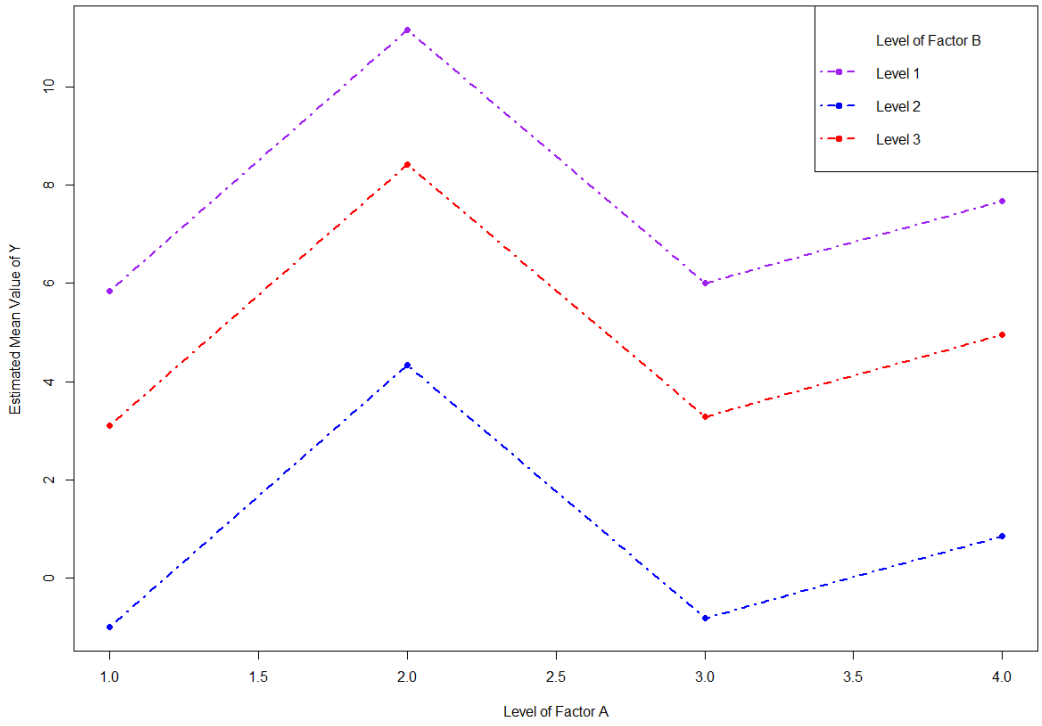
ii. Όμοια με την πρόταση 3.3 (σελίδα 107).

iii. Σχετικά με το SST, υπολογίζουμε ότι:

$$\begin{aligned} E\left(\text{SST}\right) &= E(\text{SSA}) + E(\text{SSB}) + E(\text{SSE}) \\ &= (m-1)\sigma^2 + \ell\sum_{i=1}^{m}\alpha_i^2 + (\ell-1)\sigma^2 + m\sum_{j=1}^{\ell}\beta_j^2 + (m-1)(\ell-1)\sigma^2 \\ &= (m\ell-1)\sigma^2 + \ell\sum_{i=1}^{m}\alpha_i^2 + m\sum_{j=1}^{\ell}\beta_j^2 \Rightarrow \end{aligned}$$

$$E\left(\frac{\text{SST}}{n-1}\right) = E\left(S_Y^2\right) = \sigma^2 + \frac{\ell}{m\ell - 1}\sum_{i=1}^{m}\alpha_i^2 + \frac{m}{m\ell - 1}\sum_{j=1}^{\ell}\beta_j^2. \quad \square$$

Σημείωση 3.1. Αν σχεδιάζαμε ένα γράφημα με τις εκτιμημένες μέσες τιμές της αποκριτικής μεταβλητής Y για όλους τους δυνατούς συνδυασμούς επιπέδων των δύο παραγόντων, τότε θα βλέπαμε ένα σχήμα όπως το 3.1. Σε αυτό το σχήμα, καθεμία από τις τρεις καμπύλες που σχηματίζουν τα τρία επίπεδα του παράγοντα B είναι μία παράλληλη μετατόπιση των άλλων δύο καμπυλών. Αυτό συμβαίνει διότι το μοντέλο ανάλυσης διασποράς χωρίς αλληλεπίδραση δε λαμβάνει καθόλου υπόψη του τις πιθανές διαφορές που μπορεί να έχει η επίδραση του ενός παράγοντα ανάμεσα στα διαφορετικά επίπεδα του άλλου παράγοντα.



Σχημα 3.1: Εκτιμημένες Μέσες Τιμές χωρίς Αλληλεπίδραση

3.5 ANOVA κατά Δύο Παράγοντες με Αλληλεπίδραση

Θεωρούμε τώρα ότι υπάρχει αλληλεπίδραση μεταξύ των δύο παραγόντων, δηλαδή η επίδραση του παράγοντα A στη μέση τιμή της αποκριτικής μεταβλητής Y εξαρτάται από το επίπεδο του παράγοντα B και η επίδραση του παράγοντα B στη μέση τιμή της αποκριτικής μεταβλητής Y εξαρτάται από το επίπεδο του παράγοντα A. Προκειμένου να ελέγξουμε για πιθανή ύπαρξη αλληλεπίδρασης μεταξύ των παραγόντων, θα πρέπει αναγκαστικά να έχουμε περισσότερες από μία

παρατηρήσεις για κάθε συνδυασμό επιπέδων των δύο παραγόντων. Για λόγους απλότητας θεωρούμε ότι έχουμε c παρατηρήσεις για κάθε συνδυασμό επιπέδων των δύο παραγόντων, δηλαδή $n = cm\ell$. Γενικότερα, θα μπορούσαμε να θεωρήσουμε ότι έχουμε n_{ij} παρατηρήσεις από τον συνδυασμό του i-οστού επιπέδου του παράγοντα A και του j-οστού επιπέδου του παράγοντα B. Έστω Y_{ijr} η r-οστή παρατήρηση από τον συνδυασμό του i-οστού επιπέδου του παράγοντα A και του j-οστού επιπέδου του παράγοντα B. Ορίζουμε:

- Το άθροισμα των παρατηρήσεων από τον συνδυασμό του i-οστού επιπέδου του παράγοντα A και του j-οστού επιπέδου του παράγοντα B:

$$Y_{ij.} = \sum_{r=1}^{c} Y_{ijr}, \quad i = 1, 2, \ldots, m, \quad j = 1, 2, \ldots, \ell.$$

- Τον δειγματικό μέσο των παρατηρήσεων από τον συνδυασμό του i-οστού επιπέδου του παράγοντα A και του j-οστού επιπέδου του παράγοντα B:

$$\overline{Y}_{ij.} = \frac{Y_{ij.}}{c}, \quad i = 1, 2, \ldots, m, \quad j = 1, 2, \ldots, \ell.$$

- Το άθροισμα των παρατηρήσεων στο επίπεδο i του παράγοντα A:

$$Y_{i..} = \sum_{j=1}^{\ell} \sum_{r=1}^{c} Y_{ijr}, \quad i = 1, 2, \ldots, m.$$

- Τον δειγματικό μέσο των παρατηρήσεων στο επίπεδο i του παράγοντα A:

$$\overline{Y}_{i..} = \frac{Y_{i..}}{c\ell}, \quad i = 1, 2, \ldots, m.$$

- Το άθροισμα των παρατηρήσεων στο επίπεδο j του παράγοντα B:

$$Y_{.j.} = \sum_{i=1}^{m} \sum_{r=1}^{c} Y_{ijr}, \quad j = 1, 2, \ldots, \ell.$$

- Τον δειγματικό μέσο των παρατηρήσεων στο επίπεδο j του παράγοντα B:

$$\overline{Y}_{.j.} = \frac{Y_{.j.}}{cm}, \quad j = 1, 2, \ldots, \ell.$$

- Το συνολικό άθροισμα των παρατηρήσεων:

$$Y_{...} = \sum_{i=1}^{m} \sum_{j=1}^{\ell} \sum_{r=1}^{c} Y_{ij}.$$

- Τον συνολικό δειγματικό μέσο των παρατηρήσεων:

$$\overline{Y}_{\cdots} = \frac{Y_{\cdots}}{cm\ell}.$$

Το **μοντέλο ανάλυσης διασποράς κατά δύο παράγοντες με αλληλεπίδραση** γράφεται ως:

$$Y_{ijr} = \mu_{ij} + \varepsilon_{ijr} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijr}$$

για $i = 1, 2, \ldots, m$, $j = 1, 2, \ldots, \ell$ και $r = 1, 2, \ldots, c$, όπου:

- μ_{ij} η πληθυσμιακή μέση τιμή της Y στον συνδυασμό του i-οστού επιπέδου του παράγοντα A και του j-οστού επιπέδου του παράγοντα B. Χρησιμοποιούμε τον αντίστοιχο δειγματικό μέσο της Y για να εκτιμήσουμε την άγνωστη πραγματική μέση τιμή μ_{ij}, οπότε $\widehat{\mu}_{ij} = \overline{Y}_{ij\cdot}$.

- $\mu_{i\cdot}$ η πληθυσμιακή μέση τιμή της Y στο i-οστό επίπεδο του παράγοντα A. Χρησιμοποιούμε τον αντίστοιχο δειγματικό μέσο της Y για να εκτιμήσουμε την άγνωστη πραγματική μέση τιμή $\mu_{i\cdot}$, οπότε $\widehat{\mu}_{i\cdot} = \overline{Y}_{i\cdot\cdot}$.

- $\mu_{\cdot j}$ η πληθυσμιακή μέση τιμή της Y στο j-οστό επίπεδο του παράγοντα B. Χρησιμοποιούμε τον αντίστοιχο δειγματικό μέσο της Y για να εκτιμήσουμε την άγνωστη πραγματική μέση τιμή $\mu_{\cdot j}$, οπότε $\widehat{\mu}_{\cdot j} = \overline{Y}_{\cdot j\cdot}$.

- μ η πληθυσμιακή μέση τιμή της Y. Χρησιμοποιούμε τον συνολικό δειγματικό μέσο της Y για να εκτιμήσουμε τη μέση τιμή μ, οπότε $\widehat{\mu} = \overline{Y}_{\cdots}$. Προφανώς, ισχύει ότι:

$$\mu = \frac{1}{m\ell}\sum_{i=1}^{m}\sum_{j=1}^{\ell}\mu_{ij} = \frac{1}{m}\sum_{i=1}^{m}\mu_{i\cdot} = \frac{1}{\ell}\sum_{j=1}^{\ell}\mu_{\cdot j}.$$

- $\alpha_i = \mu_{i\cdot} - \mu$ η επίδραση του i-οστού επιπέδου του παράγοντα A στην πληθυσμιακή μέση τιμή της Y. Αντικαθιστώντας στην παραπάνω σχέση, παίρνουμε ότι $\sum_{i=1}^{m}\alpha_i = 0$. Επιπλέον, συμπεραίνουμε ότι $\widehat{\alpha}_i = \overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdots}$.

- $\beta_j = \mu_{\cdot j} - \mu$ η επίδραση του j-οστού επιπέδου του παράγοντα B στην πληθυσμιακή μέση τιμή της Y. Αντικαθιστώντας στην παραπάνω σχέση, παίρνουμε ότι $\sum_{j=1}^{\ell}\beta_j = 0$. Επιπλέον, συμπεραίνουμε ότι $\widehat{\beta}_j = \overline{Y}_{\cdot j\cdot} - \overline{Y}_{\cdots}$.

- $\gamma_{ij} = \mu_{ij} - \mu_{i\cdot} - \mu_{\cdot j} + \mu$ η αλληλεπίδραση μεταξύ του i-οστού επιπέδου του παράγοντα A και του j-οστού επιπέδου του παράγοντα B στην πληθυσμιακή μέση τιμή της Y. Αντικαθιστώντας στην παραπάνω σχέση, παίρνουμε ότι $\sum_{i=1}^{m}\gamma_{ij} = \sum_{j=1}^{\ell}\gamma_{ij} = 0$. Επιπλέον, συμπεραίνουμε ότι $\widehat{\gamma}_{ij} = \overline{Y}_{ij\cdot} - \overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdot j\cdot} + \overline{Y}_{\cdots}$.

- $\varepsilon_{ijr} \sim N(0, \sigma^2)$ ανεξάρτητα σφάλματα για $i = 1, 2, \ldots, m$, $j = 1, 2, \ldots, \ell$

και $r = 1, 2, \ldots, c$. Συμπεραίνουμε ότι $Y_{ijr} \sim N(\mu_{ij}, \sigma^2)$ ανεξάρτητα για $i = 1, 2, \ldots, m$, $j = 1, 2, \ldots, \ell$ και $r = 1, 2, \ldots, c$. Οι προσαρμοσμένες τιμές δίνονται ως $\widehat{Y}_{ijr} = \widehat{\mu}_{ij} = \overline{Y}_{ij.}$, ενώ τα μη-παρατηρήσιμα τυχαία σφάλματα ε_{ijr} εκτιμούνται από τα κατάλοιπα $\widehat{\varepsilon}_{ijr} = Y_{ijr} - \widehat{Y}_{ijr} = Y_{ijr} - \overline{Y}_{ij.}$.

- $\overline{Y}_{ij.} \sim N\left(\mu_{ij}, \frac{\sigma^2}{c}\right)$ για $i = 1, 2, \ldots, m$ και $j = 1, 2, \ldots, \ell$, $\overline{Y}_{i..} \sim N\left(\mu_{i.}, \frac{\sigma^2}{c\ell}\right)$ για $i = 1, 2, \ldots, m$, $\overline{Y}_{.j} \sim N\left(\mu_{.j.}, \frac{\sigma^2}{cm}\right)$ για $j = 1, 2, \ldots, \ell$ και $\overline{Y}_{...} \sim N\left(\mu, \frac{\sigma^2}{cm\ell}\right)$.

Θα αναλύσουμε τη συνολική μεταβλητότητα των δεδομένων σε τέσσερα κομμάτια. Το πρώτο κομμάτι εξηγείται από τον παράγοντα A μέσω της απόκλισης των δειγματικών μέσων $\overline{Y}_{i..}$ από τον συνολικό δειγματικό μέσο $\overline{Y}_{...}$, το δεύτερο κομμάτι εξηγείται από τον παράγοντα B μέσω της απόκλισης των δειγματικών μέσων $\overline{Y}_{.j.}$ από τον συνολικό δειγματικό μέσο $\overline{Y}_{...}$, το τρίτο κομμάτι εξηγείται από την αλληλεπίδραση μεταξύ των δύο παραγόντων και το τέταρτο κομμάτι παραμένει ανεξήγητο και ποσοτικοποιείται μέσω των καταλοίπων $\widehat{\varepsilon}_{ijr}$.

Ορισμός 3.4. (Αθροίσματα Τετραγώνων)

- Ορίζουμε $\text{SST} = \sum_{i=1}^{m} \sum_{j=1}^{\ell} \sum_{r=1}^{c} \left(Y_{ijr} - \overline{Y}_{...}\right)^2$ το συνολικό άθροισμα τετραγώνων (total sum of squares) των δεδομένων.
- Ορίζουμε $\text{SSA} = c\ell \sum_{i=1}^{m} \left(\overline{Y}_{i..} - \overline{Y}_{...}\right)^2$ το άθροισμα τετραγώνων που οφείλεται στον παράγοντα A (sum of squares due to factor A).
- Ορίζουμε $\text{SSB} = cm \sum_{j=1}^{\ell} \left(\overline{Y}_{.j.} - \overline{Y}_{...}\right)^2$ το άθροισμα τετραγώνων που οφείλεται στον παράγοντα B (sum of squares due to factor B).
- Ορίζουμε $\text{SSAB} = c \sum_{i=1}^{m} \sum_{j=1}^{\ell} \left(\overline{Y}_{ij.} - \overline{Y}_{i..} - \overline{Y}_{.j.} + \overline{Y}_{...}\right)^2$ το άθροισμα τετραγώνων που οφείλεται στην αλληλεπίδραση του A με τον B (sum of squares due to the interaction between A and B).
- Ορίζουμε $\text{SSE} = \sum_{i=1}^{m} \sum_{j=1}^{\ell} \sum_{r=1}^{c} (Y_{ijr} - \widehat{Y}_{ijr})^2 = \sum_{i=1}^{m} \sum_{j=1}^{\ell} \sum_{r=1}^{c} \left(Y_{ijr} - \overline{Y}_{ij.}\right)^2$ το άθροισμα τετραγώνων των καταλοίπων (sum of squared errors).

Πρόταση 3.11. (Ανάλυση Διασποράς)

$$\sum_{i=1}^{m} \sum_{j=1}^{\ell} \sum_{r=1}^{c} \left(Y_{ijr} - \overline{Y}_{...}\right)^2 = c\ell \sum_{i=1}^{m} \left(\overline{Y}_{i..} - \overline{Y}_{...}\right)^2 + cm \sum_{j=1}^{\ell} \left(\overline{Y}_{.j.} - \overline{Y}_{...}\right)^2 + c \sum_{i=1}^{m} \sum_{j=1}^{\ell} \left(\overline{Y}_{ij.} - \overline{Y}_{i..} - \overline{Y}_{.j.} + \overline{Y}_{...}\right)^2 + \sum_{i=1}^{m} \sum_{j=1}^{\ell} \sum_{r=1}^{c} \left(Y_{ijr} - \overline{Y}_{ij.}\right)^2, \text{ δηλαδή:}$$

$$\mathrm{SST} = \mathrm{SSA} + \mathrm{SSB} + \mathrm{SSAB} + \mathrm{SSE}.$$

Απόδειξη. Αρκεί να δείξουμε τα εξής:

$$\begin{aligned}
\mathrm{CPAB} &= \sum_{i=1}^{m}\sum_{j=1}^{\ell} \left(\overline{Y}_{i..} - \overline{Y}_{...}\right)\left(\overline{Y}_{.j.} - \overline{Y}_{...}\right) \\
&= \sum_{i=1}^{m} \left[\left(\overline{Y}_{i..} - \overline{Y}_{...}\right) \left(\sum_{j=1}^{\ell} \frac{\cancel{Y_{.j.}}}{\cancel{cm}} - \frac{\cancel{Y_{...}}}{\cancel{cm}} \right) \right] = 0,
\end{aligned}$$

$$\begin{aligned}
\mathrm{CPAAB} &= \sum_{i=1}^{m}\sum_{j=1}^{\ell} \left(\overline{Y}_{i..} - \overline{Y}_{...}\right)\left(\overline{Y}_{ij.} - \overline{Y}_{i..} - \overline{Y}_{.j.} + \overline{Y}_{...}\right) \\
&= \sum_{i=1}^{m} \left[\left(\overline{Y}_{i..} - \overline{Y}_{...}\right) \left(\sum_{j=1}^{\ell} \frac{\cancel{Y_{ij.}}}{c} - \frac{\cancel{Y_{i..}}}{\cancel{c}} - \sum_{j=1}^{\ell} \frac{\cancel{Y_{.j.}}}{\cancel{cm}} + \frac{\cancel{Y_{...}}}{\cancel{cm}} \right) \right] = 0,
\end{aligned}$$

$$\begin{aligned}
\mathrm{CPAE} &= \sum_{i=1}^{m}\sum_{j=1}^{\ell}\sum_{r=1}^{c} \left(\overline{Y}_{i..} - \overline{Y}_{...}\right)\left(Y_{ijr} - \overline{Y}_{ij.}\right) \\
&= \sum_{i=1}^{m}\sum_{j=1}^{\ell} \left(\overline{Y}_{i..} - \overline{Y}_{...}\right)\left(\cancel{Y_{ij.}} - \cancel{c\overline{Y}_{ij.}}\right) = 0,
\end{aligned}$$

$$\begin{aligned}
\mathrm{CPBAB} &= \sum_{i=1}^{m}\sum_{j=1}^{\ell} \left(\overline{Y}_{.j.} - \overline{Y}_{...}\right)\left(\overline{Y}_{ij.} - \overline{Y}_{i..} - \overline{Y}_{.j.} + \overline{Y}_{...}\right) \\
&= \sum_{j=1}^{\ell} \left[\left(\overline{Y}_{.j.} - \overline{Y}_{...}\right) \left(\sum_{i=1}^{m} \frac{\cancel{Y_{ij.}}}{c} - \sum_{i=1}^{m} \frac{\cancel{Y_{i..}}}{c\ell} - \frac{\cancel{Y_{.j.}}}{\cancel{c}} + \frac{\cancel{Y_{...}}}{\cancel{c\ell}} \right) \right] = 0,
\end{aligned}$$

$$\begin{aligned}
\mathrm{CPBE} &= \sum_{i=1}^{m}\sum_{j=1}^{\ell}\sum_{r=1}^{c} \left(\overline{Y}_{.j.} - \overline{Y}_{...}\right)\left(Y_{ijr} - \overline{Y}_{ij.}\right) \\
&= \sum_{i=1}^{m}\sum_{j=1}^{\ell} \left(\overline{Y}_{.j.} - \overline{Y}_{...}\right)\left(\cancel{Y_{ij.}} - \cancel{c\overline{Y}_{ij.}}\right) = 0 \quad \text{και}
\end{aligned}$$

$$\begin{aligned}
\mathrm{CPABE} &= \sum_{i=1}^{m}\sum_{j=1}^{\ell}\sum_{r=1}^{c} \left(\overline{Y}_{ij.} - \overline{Y}_{i..} - \overline{Y}_{.j.} + \overline{Y}_{...}\right)\left(Y_{ijr} - \overline{Y}_{ij.}\right) \\
&= \sum_{i=1}^{m}\sum_{j=1}^{\ell} \left(\overline{Y}_{ij.} - \overline{Y}_{i..} - \overline{Y}_{.j.} + \overline{Y}_{...}\right)\left(\cancel{Y_{ij.}} - \cancel{c\overline{Y}_{ij.}}\right) = 0. \quad \square
\end{aligned}$$

Ενδιαφερόμαστε να πραγματοποιήσουμε τους τρεις παρακάτω ελέγχους υποθέσεων:

$$\begin{cases}
H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_m = 0 \text{ vs.} \\
H_1 : \alpha_i \neq 0 \text{ για κάποιο } i \in \{1, 2, \ldots, m\},
\end{cases}$$

$$\begin{cases} H_0 : \beta_1 = \beta_2 = \cdots = \beta_\ell = 0 \text{ vs.} \\ H_1 : \beta_j \neq 0 \text{ για κάποιο } j \in \{1, 2, \ldots, \ell\}, \end{cases}$$

$$\begin{cases} H_0 : \gamma_{11} = \gamma_{12} = \cdots = \gamma_{m\ell} = 0 \text{ vs.} \\ H_1 : \gamma_{ij} \neq 0 \text{ για κάποια } i \in \{1, 2, \ldots, m\} \text{ και } j \in \{1, 2, \ldots, \ell\}. \end{cases}$$

Ο πρώτος έλεγχος μας βοηθάει να αποφασίσουμε αν υπάρχει στατιστικά σημαντική συνεισφορά του παράγοντα A στην πρόβλεψη της μέσης τιμής της αποκριτικής μεταβλητής Y, ο δεύτερος να αποφασίσουμε αν υπάρχει στατιστικά σημαντική συνεισφορά του παράγοντα B και ο τρίτος έλεγχος αν υπάρχει στατιστικά σημαντική αλληλεπίδραση μεταξύ των δύο παραγόντων στην πρόβλεψη της μέσης τιμής της αποκριτικής μεταβλητής Y.

	Sum of Squares	d.f.	Mean Square	F
A	$c\ell \sum_{i=1}^{m} \left(\overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdot\cdot\cdot}\right)^2$	$m-1$	$\frac{\text{SSA}}{m-1}$	$\frac{\text{MSA}}{\text{MSE}}$
B	$cm \sum_{j=1}^{\ell} \left(\overline{Y}_{\cdot j\cdot} - \overline{Y}_{\cdot\cdot\cdot}\right)^2$	$\ell-1$	$\frac{\text{SSB}}{\ell-1}$	$\frac{\text{MSB}}{\text{MSE}}$
AB	$c \sum_{i=1}^{m} \sum_{j=1}^{\ell} \left(\overline{Y}_{ij\cdot} - \overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdot j\cdot} + \overline{Y}_{\cdot\cdot\cdot}\right)^2$	$(m-1)(\ell-1)$	$\frac{\text{SSAB}}{(m-1)(\ell-1)}$	$\frac{\text{MSAB}}{\text{MSE}}$
E	$\sum_{i=1}^{m} \sum_{j=1}^{\ell} \sum_{r=1}^{c} \left(Y_{ijr} - \overline{Y}_{ij\cdot}\right)^2$	$m\ell(c-1)$	$\frac{\text{SSE}}{m\ell(c-1)}$	
T	$\sum_{i=1}^{m} \sum_{j=1}^{\ell} \sum_{r=1}^{c} \left(Y_{ijr} - \overline{Y}_{\cdot\cdot\cdot}\right)^2$	$cm\ell-1$		

Πινακας 3.6: Πίνακας ANOVA κατά δύο παράγοντες με αλληλεπίδραση

Πρόταση 3.12. (Αθροίσματα Τετραγώνων)

i. Ισχύει ότι $\frac{\text{SSA}}{\sigma^2} \sim \chi^2_{m-1}$ υπό την $H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_m = 0$ και $\frac{\text{SSB}}{\sigma^2} \sim \chi^2_{\ell-1}$ υπό την $H_0 : \beta_1 = \beta_2 = \cdots = \beta_\ell = 0$.

ii. Ισχύει ότι $\text{SSE} = \sum_{i=1}^{m} \sum_{j=1}^{\ell} (c-1) S_{ij}^2$, όπου $S_{ij}^2 = \frac{1}{c-1} \sum_{r=1}^{c} (Y_{ijr} - \overline{Y}_{ij\cdot})^2$. Συμπεραίνουμε ότι $\frac{\text{SSE}}{\sigma^2} \sim \chi^2_{m\ell(c-1)}$.

iii. Ισχύει ότι $\frac{\text{SSAB}}{\sigma^2} \sim \chi^2_{(m-1)(\ell-1)}$. Επιπλέον, τα SSA, SSB, SSAB και SSE είναι ανεξάρτητα.

iv. Ισχύει ότι $F_A = \frac{\text{MSA}}{\text{MSE}} \sim F_{m-1, m\ell(c-1)}$ υπό την $H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_m = 0$, $F_B = \frac{\text{MSB}}{\text{MSE}} \sim F_{\ell-1, m\ell(c-1)}$ υπό τη μηδενική υπόθεση $H_0 : \beta_1 = \beta_2 = \cdots = \beta_\ell = 0$ και $F_{AB} = \frac{\text{MSAB}}{\text{MSE}} \sim F_{(m-1)(\ell-1), m\ell(c-1)}$ υπό την $H_0 : \gamma_{11} = \gamma_{12} = \cdots = \gamma_{m\ell} = 0$.

Απόδειξη.

i. Αγνοώντας την ύπαρξη του παράγοντα B και εργαζόμενοι όπως στην πρό-

ταση 3.2, προκύπτει ότι $\frac{\text{SSA}}{\sigma^2} \sim \chi^2_{m-1}$ υπό την $H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_m = 0$. Ομοίως, αγνοώντας την ύπαρξη του παράγοντα A και εργαζόμενοι ακριβώς όπως στην πρόταση 3.2, προκύπτει ότι $\frac{\text{SSB}}{\sigma^2} \sim \chi^2_{\ell-1}$ υπό τη μηδενική υπόθεση $H_0 : \beta_1 = \beta_2 = \cdots = \beta_\ell = 0$.

ii. Το $S^2_{ij} = \frac{1}{c-1} \sum_{r=1}^{c} (Y_{ijr} - \overline{Y}_{ij.})^2$ είναι η δειγματική διασπορά της Y στον συνδυασμό του i-οστού επιπέδου του παράγοντα A και του j-οστού επιπέδου του παράγοντα B. Έχουμε ότι $\frac{(c-1)S^2_{ij}}{\sigma^2} \sim \chi^2_{c-1}$ για $i = 1, 2, \ldots, m$ και $j = 1, 2, \ldots, \ell$. Εφόσον οι παρατηρήσεις μεταξύ διαφορετικών συνδυασμών επιπέδων είναι ανεξάρτητες, συμπεραίνουμε ότι οι δειγματικές διασπορές $S^2_{11}, S^2_{12}, \ldots, S^2_{m\ell}$ είναι ανεξάρτητες μεταξύ τους. Επομένως, το $\frac{\text{SSE}}{\sigma^2} = \sum_{i=1}^{m} \sum_{j=1}^{\ell} \frac{(c-1)S^2_{ij}}{\sigma^2}$ ακολουθεί την κατανομή χ^2 με $\sum_{i=1}^{m} \sum_{j=1}^{\ell} (c-1) = m\ell(c-1)$ βαθμούς ελευθερίας.

iii. Η κατανομή του $\frac{\text{SSAB}}{\sigma^2}$ και η ανεξαρτησία μεταξύ SSA, SSB, SSAB και SSE προκύπτουν με κατάλληλη εφαρμογή του θεωρήματος Cochran, όπως και στην πολλαπλή γραμμική παλινδρόμηση.

iv. Οι κατανομές των τυχαίων μεταβλητών F_A, F_B και F_{AB} προκύπτουν άμεσα από τα παραπάνω και τον ορισμό της κατανομής F του Snedecor. □

Πρόταση 3.13. (Εκτίμηση της Διασποράς)

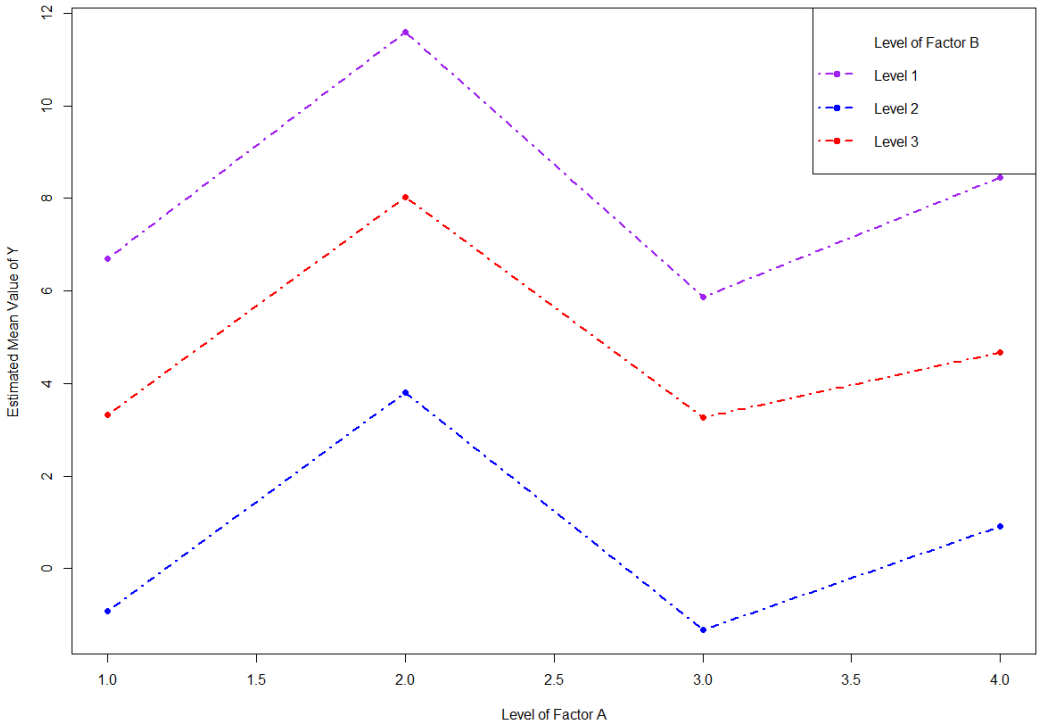
i. Ισχύει ότι $E(\text{MSE}) = \sigma^2$, δηλαδή το MSE είναι μία αμερόληπτη εκτιμήτρια της διασποράς σ^2 στο μοντέλο ανάλυσης διασποράς κατά έναν παράγοντα.

ii. $E(\text{MSA}) = \sigma^2 + \frac{c\ell}{m-1} \sum_{i=1}^{m} \alpha_i^2 \geqslant \sigma^2$ και $E(\text{MSB}) = \sigma^2 + \frac{cm}{\ell-1} \sum_{j=1}^{\ell} \beta_j^2 \geqslant \sigma^2$, δηλαδή το MSA είναι αμερόληπτη εκτιμήτρια του σ^2 αν και μόνο αν ισχύει η μηδενική υπόθεση $H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_m = 0$ και το MSB είναι αμερόληπτη εκτιμήτρια του σ^2 αν και μόνο αν ισχύει η $H_0 : \beta_1 = \beta_2 = \cdots = \beta_\ell = 0$. Στη γενικότερη περίπτωση τα MSA και MSB υπερεκτιμούν το σ^2.

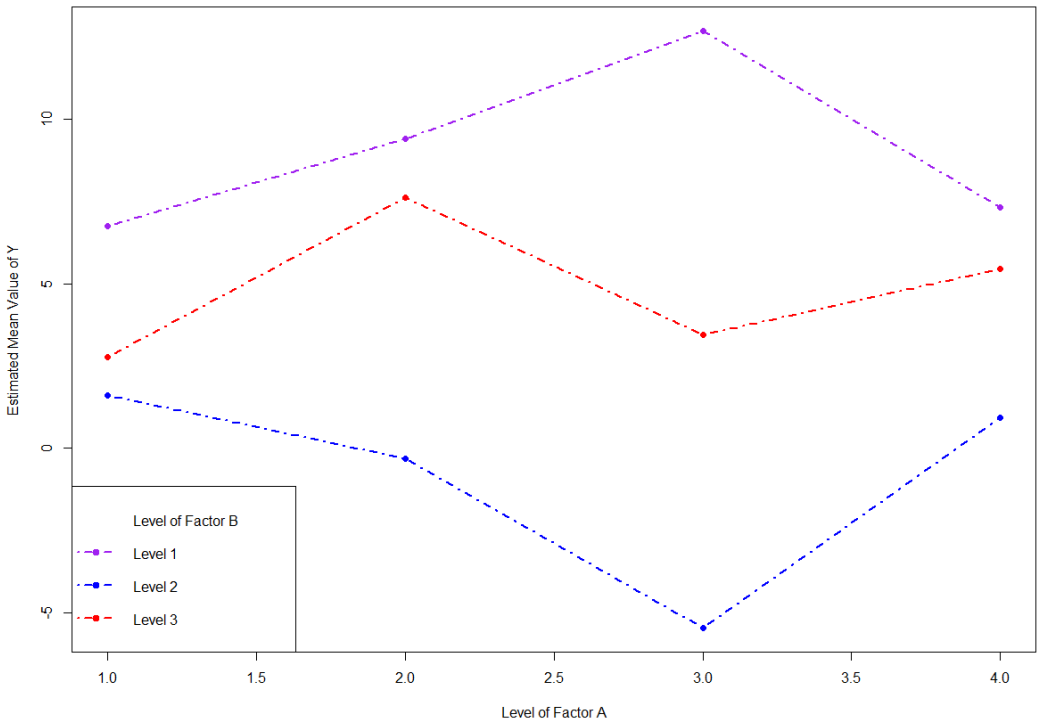
iii. $E(\text{MSAB}) = \sigma^2 + \frac{c}{(m-1)(\ell-1)} \sum_{i=1}^{m} \sum_{j=1}^{\ell} \gamma_{ij}^2 \geqslant \sigma^2$, δηλαδή το MSAB είναι αμερόληπτη εκτιμήτρια του σ^2 αν και μόνο αν ισχύει η $H_0 : \gamma_{11} = \gamma_{12} = \cdots = \gamma_{m\ell} = 0$. Στη γενικότερη περίπτωση το MSAB υπερεκτιμά το σ^2.

iv. $E\left(\frac{\text{SST}}{m\ell-1}\right) = E(S_Y^2) = \sigma^2 + \frac{c\ell}{cm\ell-1} \sum_{i=1}^{m} \alpha_i^2 + \frac{cm}{cm\ell-1} \sum_{j=1}^{\ell} \beta_j^2 + \frac{c}{cm\ell-1} \sum_{i=1}^{m} \sum_{j=1}^{\ell} \gamma_{ij}^2 \geqslant \sigma^2$, δηλαδή το S_Y^2 είναι αμερόληπτη εκτιμήτρια του σ^2 αν και μόνο αν ισχύουν οι μηδενικές υποθέσεις $H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_m = 0$, $H_0 : \beta_1 = \beta_2 = \cdots = \beta_\ell = 0$ και $H_0 : \gamma_{11} = \gamma_{12} = \cdots = \gamma_{m\ell} = 0$ ταυτόχρονα. Στη γενικότερη περίπτωση το S_Y^2 υπερεκτιμά το σ^2.

Απόδειξη. Ομοίως με την πρόταση 3.3 (σελίδα 107).



Σχημα 3.2: Εκτιμημένες Μέσες Τιμές με Ασθενή Αλληλεπίδραση



Σχημα 3.3: Εκτιμημένες Μέσες Τιμές με Ισχυρή Αλληλεπίδραση

Σημείωση 3.2. Όταν έχουμε δύο διαθέσιμους παράγοντες για να κατασκευάσουμε ένα μοντέλο ανάλυσης διασποράς, μία χρήσιμη ένδειξη για την ύπαρξη ή απουσία αλληλεπίδρασης ανάμεσα στους παράγοντες είναι να εκτιμήσουμε το μοντέλο ανάλυσης διασποράς με αλληλεπίδραση και να σχεδιάσουμε ένα γράφημα με τις εκτιμημένες μέσες τιμές της Y για κάθε δυνατό συνδυασμό επιπέδων των δύο παραγόντων. Αν οι καμπύλες που σχεδιάσουμε είναι σχεδόν παράλληλες μεταξύ τους, όπως στο σχήμα 3.2, τότε συμπεραίνουμε ότι δεν υπάρχει σημαντική αλληλεπίδραση. Αντιθέτως, στο σχήμα 3.3, βλέπουμε πολύ μεγάλες αποκλίσεις μεταξύ των καμπυλών, το οποίο συνεπάγεται ισχυρή αλληλεπίδραση.