

ΣΧΟΛΗ ΝΑΥΤΙΚΩΝ ΔΟΚΙΜΩΝ

Γραμμική Παλινδρόμηση
Δήμητρα Κουλουμπού, Γιώργος Κατσουλέας

Συσχέτιση Μεταξύ μεταβλητών

- Στα διάφορα προβλήματα που εξετάσαμε έως τώρα στη Στατιστική ασχοληθήκαμε κάθε φορά με μία μόνο μεταβλητή (ξεχωριστά), π.χ. με το βάρος, με το ύψος, με την επίδοση μαθητών κτλ.

Συσχέτιση Μεταξύ μεταβλητών

- Για καθεμιά μεμονωμένη μεταβλητή αρκεστήκαμε στη μελέτη της κατανομής συχνοτήτων, στον υπολογισμό διάφορων μέτρων όπως μέση τιμή, διάμεσος, διακύμανση, συντελεστής ασυμμετρίας κτλ.

Συσχέτιση Μεταξύ μεταβλητών

- Σε αρκετές όμως περιπτώσεις εξίσου ενδιαφέρουσα είναι και η ταυτόχρονη μελέτη δύο ή περισσότερων μεταβλητών, για να προσδιορίσουμε με ποιο τρόπο οι μεταβλητές αυτές σχετίζονται μεταξύ τους.

Συσχέτιση Μεταξύ μεταβλητών

Για παράδειγμα:

- Η ηλικία και το βάρος ενός παιδιού έχουν κάποια θετική εξάρτηση (συσχέτιση) μεταξύ τους με την έννοια ότι όσο πιο μεγάλο είναι το παιδί τόσο μεγαλύτερο βάρος θα έχει.

Συσχέτιση Μεταξύ μεταβλητών

■ Η διάρκεια ζωής των ζώωντων οργανισμών σε μια περιοχή

και το ποσοστό μόλυνσης της περιοχής έχουν αρνητική εξάρτηση μεταξύ τους, με την έννοια ότι όσο πιο μεγάλο είναι το ποσοστό μόλυνσης της περιοχής τόσο μικρότερη είναι η διάρκεια ζωής των οργανισμών που ζουν στην περιοχή.

■ Όσο μεγαλύτερη (μέχρι ένα ανώτερο όριο) είναι η περιεκτικότητα σε φθόριο στο πόσιμο νερό, τόσο μικρότερες είναι οι περιπτώσεις στη φθορά των δοντιών των μικρών παιδιών.

Συσχέτιση Μεταξύ μεταβλητών

- Η συνολική παραγωγή καλαμποκιού εξαρτάται από τη θέση του χωραφιού, από την ποσότητα λιπάσματος, από την επίδραση της θερμοκρασίας, της υγρασίας κτλ.
- Το ύψος των αποδοχών των υπαλλήλων μιας εταιρείας δεν εξαρτάται από το βάρος τους.

Έτσι λοιπόν είναι ενδιαφέρον να εξεταστούν οι επιδράσεις που κάποιες μεταβλητές ασκούν σε κάποιες άλλες μεταβλητές.

Ανάλυση Παλινδρόμησης

- Η ύπαρξη μιας συναρτησιακής σχέσης (εξίσωσης) μεταξύ των μεταβλητών μπορεί να είναι εξαιρετικά πολύτιμη για την πρόβλεψη των τιμών μιας μεταβλητής από τις γνώσεις που διαθέτουμε για τις άλλες μεταβλητές, όταν ισχύουν κάποιες συγκεκριμένες συνθήκες.

Ανάλυση Παλινδρόμησης

- Η κατασκευή – εκτίμηση ενός μαθηματικού στατιστικού μοντέλου που περιγράφει την εξέλιξη των τιμών μιας μεταβλητής σε σχέση με τις τιμές άλλων μεταβλητών ονομάζεται **ανάλυση παλινδρόμησης (regression analysis)**.

Ανάλυση Παλινδρόμησης

Ιστορικά, ο όρος “regression” χρησιμοποιήθηκε για πρώτη φορά από τον Άγγλο ανθρωπολόγο Galton (1822-1911) το 1885.

Με τη μελέτη του ύψους των παιδιών σε σχέση με το ύψος των γονέων. Διαπιστώθηκε ότι παιδιά ψηλών γονέων τείνουν, κατά μέσο όρο, να είναι κοντύτερα των γονιών τους, ενώ παιδιά κοντών γονέων τείνουν, κατά μέσο όρο, να γίνονται ψηλότερα των γονιών τους.

Ανάλυση Παλινδρόμησης

- Η ανάλυση παλινδρόμησης εφαρμόζεται για την εκτίμηση των τιμών μίας μεταβλητής Y (**εξαρτημένη μεταβλητή**), με βάση τις τιμές μίας ή περισσότερων μεταβλητών X (**ανεξάρτητη μεταβλητή - Predictor**).

Ανάλυση Παλινδρόμησης

- Η ανάλυση παλινδρόμησης εφαρμόζεται για την εκτίμηση των τιμών μίας μεταβλητής Y (**εξαρτημένη μεταβλητή**), με βάση τις τιμές μίας ή περισσότερων μεταβλητών X (**ανεξάρτητη μεταβλητή - Predictor**).

Ανάλυση Παλινδρόμησης

- Στόχος είναι να προσδιορισθεί ένα μαθηματικό μοντέλο - σχέση, που συνδέει τις δύο μεταβλητές το οποίο με βάση τιμές της μεταβλητής X (ή των μεταβλητών X) θα δίνει εκτιμήσεις για τις τιμές της μεταβλητής Y .
- Η κατασκευή του μοντέλου γίνεται με βάση τα δεδομένα ενός δείγματος τιμών (x_i, y_i) των μεταβλητών.

Ανάλυση Παλινδρόμησης

- Στόχος είναι να προσδιορισθεί ένα μαθηματικό μοντέλο - σχέση, που συνδέει τις δύο μεταβλητές το οποίο με βάση τιμές της μεταβλητής X (ή των μεταβλητών X) θα δίνει εκτιμήσεις για τις τιμές της μεταβλητής Y .
- Η κατασκευή του μοντέλου γίνεται με βάση τα δεδομένα ενός δείγματος τιμών (x_i, y_i) των μεταβλητών.

Διάγραμμα Διασποράς (Scatter Plot)

Πριν προχωρήσουμε στην κατασκευή – εκτίμηση οποιoδήποτε μοντέλου καλό είναι να έχουμε μία διαίσθηση για την σχέση μεταξύ των μεταβλητών X και Y . Μια τέτοια διαίσθηση μπορούμε να την αποκτήσουμε απο το γράφημα διασποράς τους (XY – scatter plot).

Διάγραμμα Διασποράς (Scatter Plot)

Παράδειγμα 1

Ας υποθέσουμε ότι μια επιχείρηση θέλει να γνωρίζει πώς το ύψος X των δαπανών που κάνει για διαφήμιση στη διάρκεια ενός μήνα επηρεάζει την αξία Y των πωλήσεών της στη διάρκεια του ίδιου μήνα.

Θεωρούμε ότι στο χρονικό αυτό διάστημα οι τιμές των προϊόντων της δεν μεταβάλλονται, οπότε οι μεταβολές της Y οφείλονται αποκλειστικά σε μεταβολές του όγκου των πωλήσεων

Διάγραμμα Διασποράς (Scatter Plot)

Παράδειγμα 1

Σε 18 διαδοχικούς μήνες παρατηρήσαμε την αξία των πωλήσεων Y μιας επιχείρησης σε επιλεγμένα επίπεδα διαφημιστικής δαπάνης X , απ' όπου προέκυψαν τα ακόλουθα 18 ζεύγη (δεδομένα ενός δείγματος τιμών) (X_i, Y_i) .

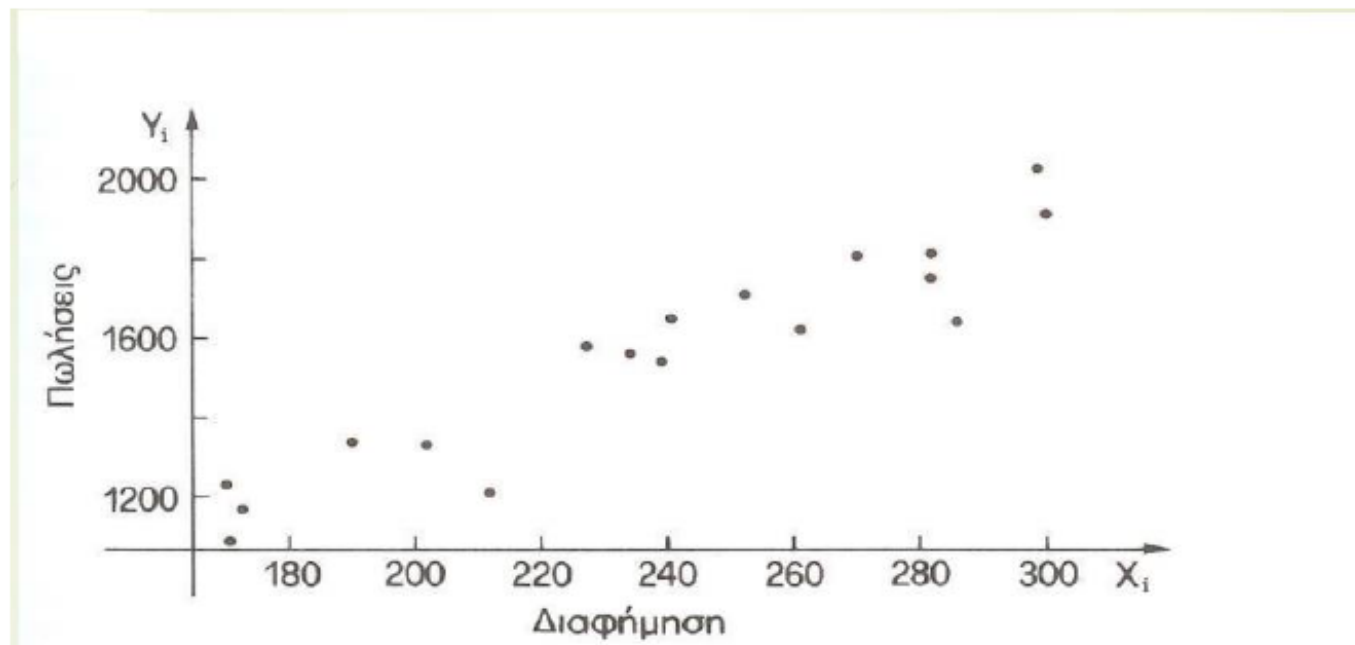
Διάγραμμα Διασποράς (Scatter Plot)

Μήνας	Διαφημιστική Δαπάνη X_i	Αξία Πολήσεων Y_i
1	170	1100
2	172	1180
3	169.5	1240
4	190	1350
5	212	1220
6	228	1590
7	202	1340
8	235	1570
9	241.5	1660

Μήνας	Διαφημιστική Δαπάνη X_i	Αξία Πολήσεων Y_i
10	240.2	1550
11	283.3	1820
12	300.4	2040
13	283.4	1760
14	262.7	1630
15	301.7	1920
16	253.2	1720
17	287.4	1650
18	271.8	1820

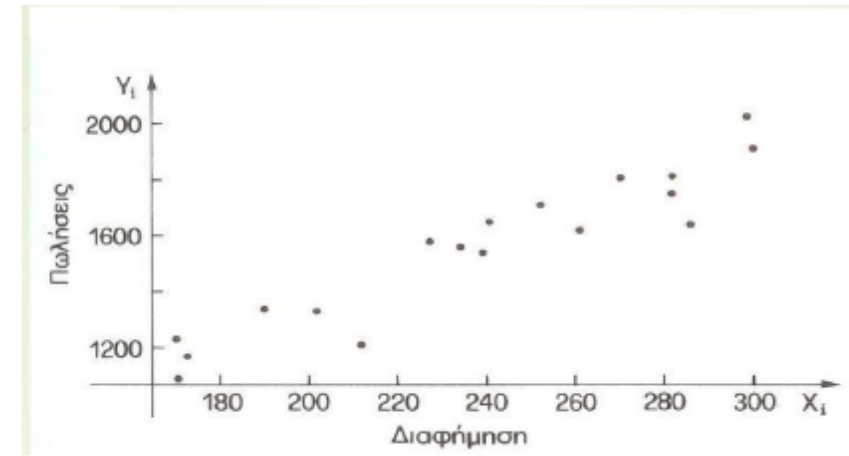
Διάγραμμα Διασποράς (Scatter Plot)

Η γραφική παράσταση των σημείων $(X_i, Y_i), i = 1, 2, \dots, 18$ στο ορθογώνιο σύστημα καρτεσιανών συντεταγμένων ονομάζεται διάγραμμα διασποράς (XY – scatter plot) .



Διάγραμμα Διασποράς (Scatter Plot)

- Παρατηρούμε ότι τα 18 σημεία τείνουν να συγκεντρώνονται γύρω από μια ευθεία γραμμή.
- Μπορούμε επομένως να χρησιμοποιήσουμε μια εξίσωση ευθείας για να περιγράψουμε την παρατηρούμενη σχέση.



Διάγραμμα Διασποράς (Scatter Plot)

- Θα πρέπει να βρεθεί μια μέθοδος εκτίμησης που να υπολογίζει την ίδια ευθεία για ένα σύνολο δεδομένων ανεξάρτητα από τον άνθρωπο που την εφαρμόζει
- Η μέθοδος που χρησιμοποιείται είναι η **μέθοδος των Ελαχίστων Τετραγώνων.**

Συντελεστής Γραμμικής Συσχέτισης του Pearson

Ο συντελεστής γραμμικής συσχέτισης του **Pearson** δύο μεταβλητών X και Y ορίζεται με βάση ένα δείγμα N ζευγών παρατηρήσεων (x_i, y_i) , $i = 1, 2, \dots, N$ συμβολίζεται με r και δίνεται από τον τύπο:

$$r = \frac{s_{xy}}{s_x s_y} = \frac{\sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^N (x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^N (y_i - \bar{y})^2}}$$

όπου $\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i$, $\bar{y} = \frac{1}{N} \sum_{i=1}^N y_i$, οι μέσες τιμές των συγκρινόμενων παραμέτρων.

Συντελεστής Γραμμικής Συσχέτισης του Pearson

Ο συντελεστής συσχέτισης είναι καθαρός αριθμός, δηλαδή δεν εκφράζεται σε συγκεκριμένες μονάδες μέτρησης, επομένως είναι ανεξάρτητος των χρησιμοποιούμενων μονάδων μέτρησης των μεταβλητών X και Y . Επί πλέον ισχύει πάντοτε ότι:

$$-1 \leq r \leq 1$$

Πιο συγκεκριμένα όταν:

Συντελεστής Γραμμικής Συσχέτισης του Pearson

- Όταν, $0 < r < 1$ τότε οι X, Y είναι **θετικά** γραμμικά συσχετισμένες.
- Όταν, $-1 < r < 0$ τότε οι X, Y είναι **αρνητικά** γραμμικά συσχετισμένες.
- Όταν $r = 1$ τότε έχουμε **τέλεια θετική γραμμική συσχέτιση** και όλα τα σημεία βρίσκονται πάνω σε μια ευθεία με θετική κλίση.
- Όταν $r = -1$ τότε έχουμε **τέλεια αρνητική γραμμική συσχέτιση** και όλα τα σημεία βρίσκονται πάνω σε μια ευθεία με αρνητική κλίση.
- Όταν $r = 0$ τότε **δεν υπάρχει γραμμική συσχέτιση** μεταξύ των μεταβλητών. Οι μεταβλητές δηλαδή X, Y είναι **γραμμικά ασυσχέτιστες**.

Συντελεστής Γραμμικής Συσχέτισης του Pearson

Ειδικά

- Αν $r = \pm 1$ τότε υπάρχει τέλεια γραμμική συσχέτιση.
- Αν $|r| \leq 0.3$ τότε λέμε ότι δεν υπάρχει γραμμική συσχέτιση.
- Αν $0.3 < |r| \leq 0.5$ τότε υπάρχει ασθενής γραμμική συσχέτιση.
- Αν $0.5 < |r| \leq 0.7$ τότε υπάρχει μέση γραμμική συσχέτιση.
- Αν $0.7 < |r| \leq 0.8$ τότε υπάρχει ισχυρή γραμμική συσχέτιση.
- Αν $0.8 < |r| < 1$, τότε υπάρχει πολύ ισχυρή γραμμική συσχέτιση.

Εμπειρική κατηγοριοποίηση του συντελεστή συσχέτισης.



Ισχυρή αρνητική συσχέτιση	Μέτρια αρνητικά συσχετισμένες	Ελαφρά αρνητικά συσχετισμένες	Ασυσχέτιστες ή πολύ ελαφρά συσχετισμένες	Ελαφρά θετικά συσχετισμένες	Μέτρια θετικά συσχετισμένες	Ισχυρή θετική συσχέτιση
$< -0,80$	$-0,80 \dots -0,50$	$-0,50 \dots -0,20$	$-0,20 \dots +0,20$	$+0,20 \dots +0,50$	$+0,50 \dots +0,80$	$> +0,80$
Τιμές του συντελεστή συσχέτισης						

Συντελεστής Γραμμικής Συσχέτισης του Pearson

Παρατηρήσεις

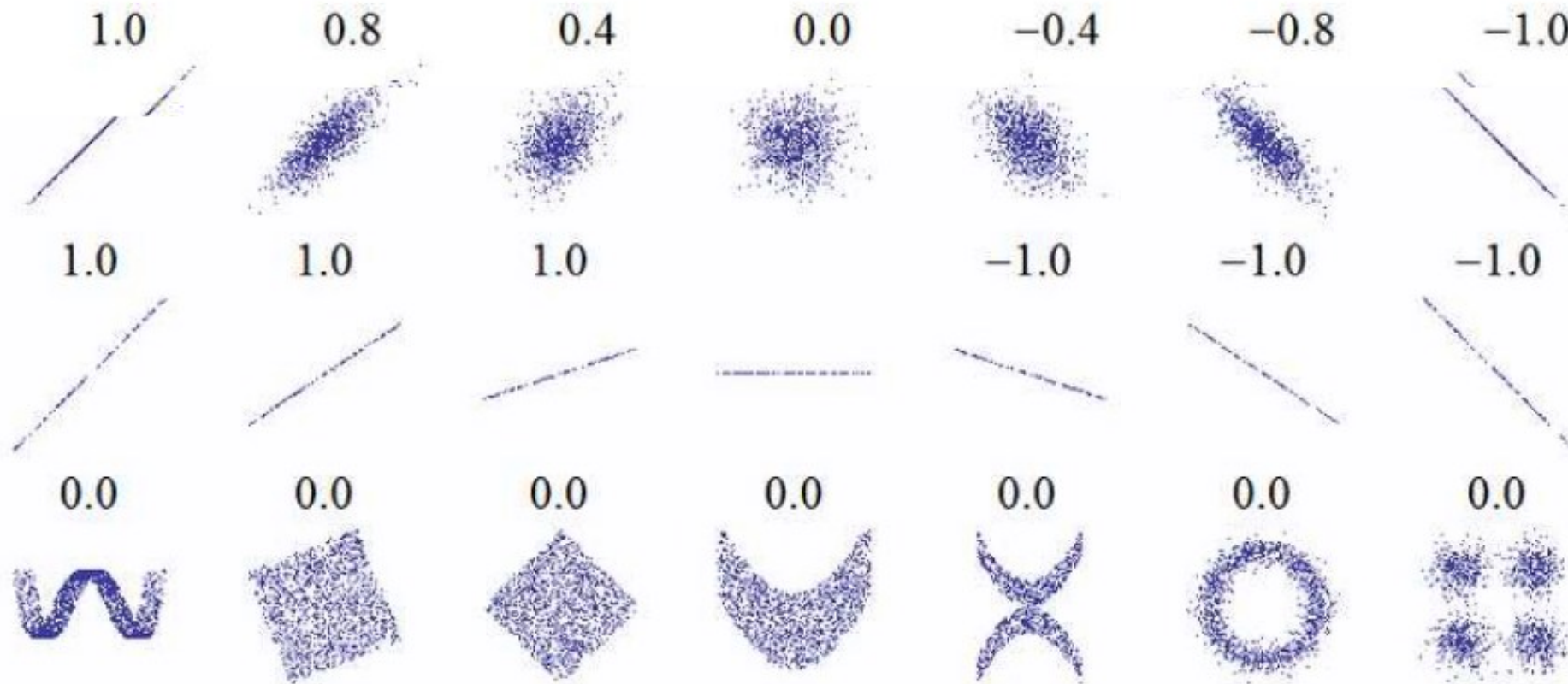
- Θετικές τιμές του r δεν υποδηλώνουν, κατ' ανάγκην μεγαλύτερο βαθμό γραμμικής συσχέτισης.
- Ο βαθμός γραμμικής συσχέτισης καθορίζεται από την απόλυτη τιμή του r και όχι από το πρόσημό του.

Συντελεστής Γραμμικής Συσχέτισης του Pearson

Παρατηρήσεις

- Το πρόσημο του r καθορίζει το είδος, μόνο, της συσχέτισης (θετική ή αρνητική).
Μας πληροφορεί δηλαδή για το αν αύξηση της μιας μεταβλητής αντιστοιχεί σε αύξηση ή σε μείωση της άλλης

Στικτά διαγράμματα για διαφορετικές τιμές του συντελεστή συσχέτισης



Συντελεστής γραμμικής συσχέτισης του Pearson

Παράδειγμα 1

Να υπολογιστεί και να ερμηνευτεί ο συντελεστής συσχέτισης r μεταξύ των μεταβλητών X και Y με βάση τις παρακάτω τιμές:

X	10	13	17	21	25	28	30
Y	21	24	29	25	36	33	40

Συντελεστής γραμμικής συσχέτισης του Pearson

Λύση

Για να διευκολυνθούμε στον υπολογισμό του συντελεστή συσχέτισης μεταξύ των X

και Y κατασκευάζουμε στο **excel** τον παρακάτω πίνακα:

Συντελεστής γραμμικής συσχέτισης του Pearson

Λύση

	X	Y	X-Xbar	Y-bar	(X-Xbar)(Y-bar)	(X-Xbar) ²	(Y-bar) ²
	10	21	-10,57	-8,71	92,0647	111,7249	75,8641
	13	24	-7,57	-5,71	43,2247	57,3049	32,6041
	17	29	-3,57	-0,71	2,5347	12,7449	0,5041
	21	25	0,43	-4,71	-2,0253	0,1849	22,1841
	25	36	4,43	6,29	27,8647	19,6249	39,5641
	28	33	7,43	3,29	24,4447	55,2049	10,8241
	30	40	9,43	10,29	97,0347	88,9249	105,8841
Αθροίσματα	144	208			285,1429	345,7143	287,4287
Μέση Τιμή	20,57143	29,71429					16,95372

Συντελεστής γραμμικής συσχέτισης του Pearson

Λύση

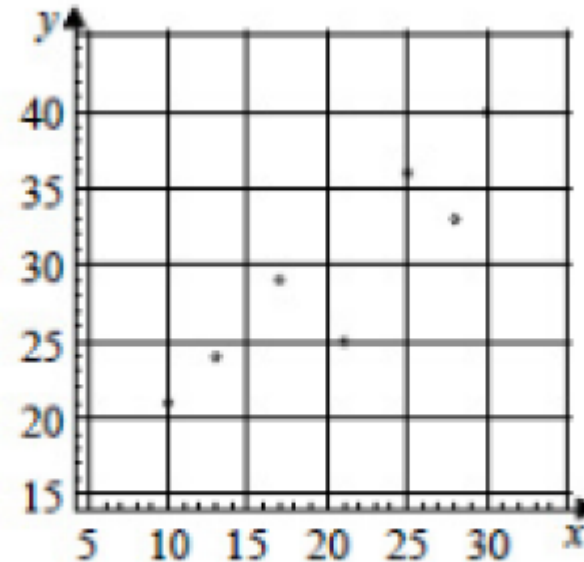
Ο συντελεστής γραμμικής συσχέτισης του Pearson δίνεται από τον τύπο:

$$r = \frac{\sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^N (x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^N (y_i - \bar{y})^2}} =$$
$$\frac{285,1429}{\sqrt{345,7143} \sqrt{287,4287}} = 0,904$$

Συντελεστής γραμμικής συσχέτισης του Pearson

Λύση

Η υψηλή τιμή του r μας δείχνει ότι υπάρχει πολύ έντονη θετική γραμμική συσχέτιση μεταξύ των μεταβλητών X και Y , όπως εξάλλου μπορούμε να το διαπιστώσουμε και από το αντίστοιχο διάγραμμα διασποράς.



Απλή Γραμμική Παλινδρόμηση

Η απλούστερη περίπτωση παλινδρόμησης είναι η **απλή γραμμική παλινδρόμηση (simple linear regression)**, κατά την οποία υπάρχει μόνο μια **ανεξάρτητη μεταβλητή X (independent or input variable)**, και η **εξαρτημένη μεταβλητή Y (dependent or response variable)**, η οποία μπορεί να προσεγγιστεί ικανοποιητικά από μία γραμμική συνάρτηση του X .

Απλή Γραμμική Παλινδρόμηση

Η Απλή Γραμμική Παλινδρόμηση εμφανίζεται τόσο σε πειραματικές όσο και σε μη πειραματικές μελέτες.

Στις πειραματικές μελέτες ο ερευνητής καθορίζει, για παράδειγμα, από πριν τις δόσεις ενός φαρμάκου (ανεξάρτητη μεταβλητή) που δίνει στα πειραματόζωα και μετρά τις αντιδράσεις τους (εξαρτημένη μεταβλητή). Με την παλινδρόμηση ενδιαφέρεται να προσδιορίσει μία σχέση δόσης-αντίδρασης για το συγκεκριμένο φάρμακο.

Απλή Γραμμική Παλινδρόμηση

- Στις μη πειραματικές μελέτες ή δειγματοληψίες, γίνονται μετρήσεις σε δύο χαρακτηριστικά (μεταβλητές) για κάθε άτομο (μονάδα) του δείγματος.

Απλή Γραμμική Παλινδρόμηση

- Για παράδειγμα, σε ένα δείγμα 10 μαθητών μετράμε, για παράδειγμα, το βάρος και το ύψος τους. Αν αυτό που μας ενδιαφέρει είναι το “τι συμβαίνει με το βάρος των παιδιών όταν αλλάζει το ύψος τους”, τότε θεωρούμε ως ανεξάρτητη μεταβλητή X το ύψος και ως εξαρτημένη μεταβλητή Y το βάρος. Οπότε, ενδιαφερόμαστε για την **παλινδρόμηση του βάρους (Y) πάνω στο ύψος (X)**.

Απλή Γραμμική Παλινδρόμηση

- Αντίθετα, αν μας ενδιαφέρει το “τι συμβαίνει με το ύψος των παιδιών όταν αλλάζει το βάρος τους”, τότε θεωρούμε ως ανεξάρτητη μεταβλητή X το βάρος και ως εξαρτημένη μεταβλητή Y το ύψος. Τότε έχουμε παλινδρόμηση του ύψους (Y) πάνω στο βάρος (X).

Το μοντέλο της απλής γραμμικής παλινδρόμησης

- Μία εξίσωση που περιγράφει πώς μία μεταβλητή Y (εξαρτημένη μεταβλητή) σχετίζεται με μία μεταβλητή X (ανεξάρτητη ή ερμηνευτική) χρησιμοποιώντας και κάποιον πρόσθετο όρο σφάλματος καλείται **μοντέλο παλινδρόμησης (regression model)**.
- Το μοντέλο της απλής γραμμικής παλινδρόμησης (simple linear regression model) εκφράζεται θεωρητικά ως εξής:

$$Y = \beta_0 + \beta_1 X + \varepsilon,$$

Όπου:

οι όροι β_0, β_1 καλούνται **παράμετροι του μοντέλου** και

ο όρος ε αποτελεί τυχαία μεταβλητή και καλείται **όρος σφάλματος**.

Σχόλια για τον όρο σφάλματος και τους συντελεστές της γραμμικής παλινδρόμησης

- Η λογική της εισαγωγής του όρου σφάλματος ε στο μοντέλο της απλής γραμμικής παλινδρόμησης $Y = \beta_0 + \beta_1 X + \varepsilon$, υποδηλώνει ότι η σχέση της ανεξάρτητης μεταβλητής με την εξαρτημένη δεν είναι **ντετερμινιστική**.
- Οι συντελεστές β_0, β_1 εξαρτώνται από τις μονάδες μέτρησης των δύο υπό εξέταση μεταβλητών – επομένως, λαμβάνουν διαφορετικές τιμές, όταν το μέγεθος X (λόγου χάριν, ύψος) μετριέται σε εκατοστά ή ίντσες.

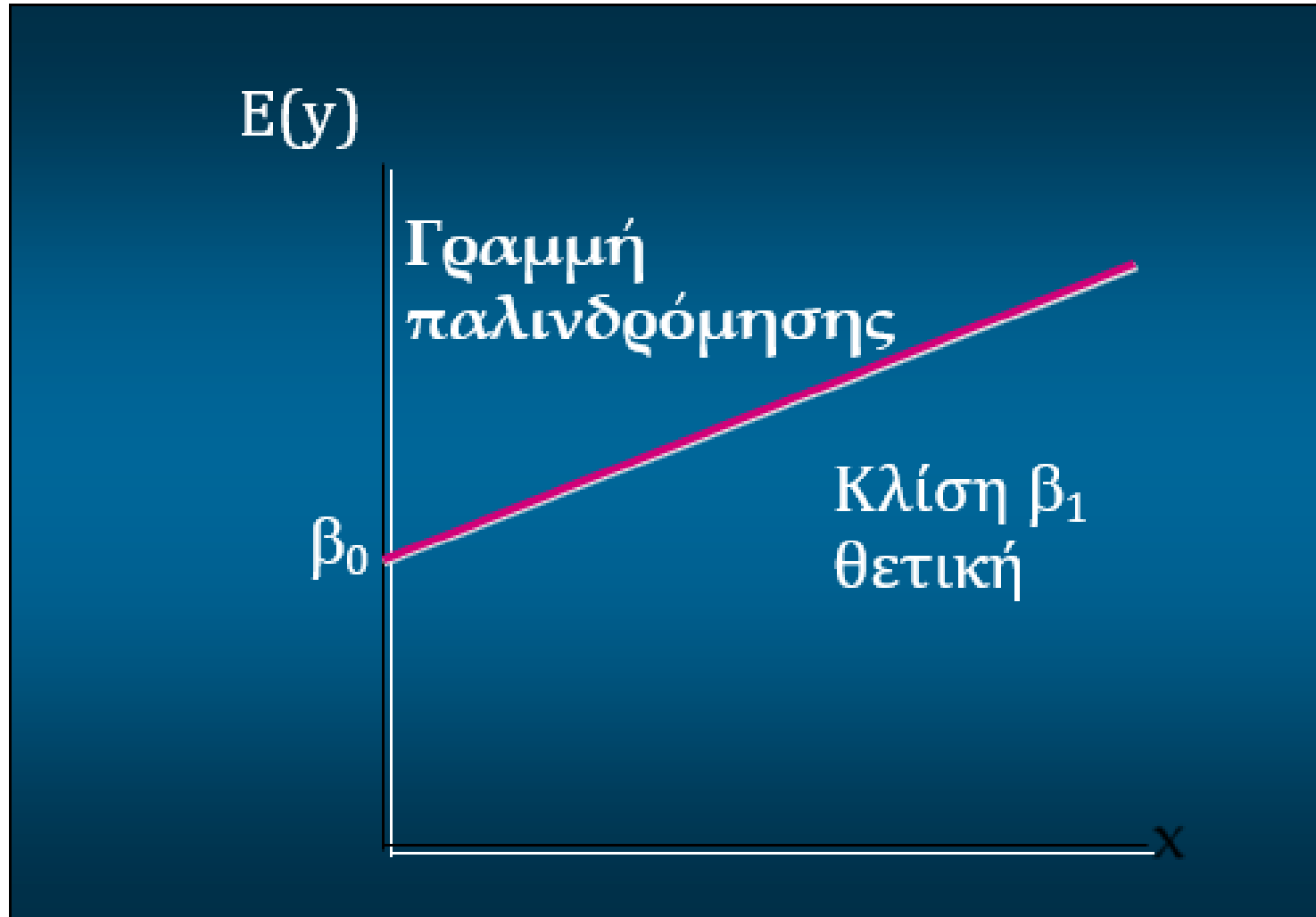
Σχόλια για τον όρο σφάλματος και τους συντελεστές της γραμμικής παλινδρόμησης (2)

Η εξίσωση της απλής γραμμικής παλινδρόμησης λαμβάνει τη μορφή:

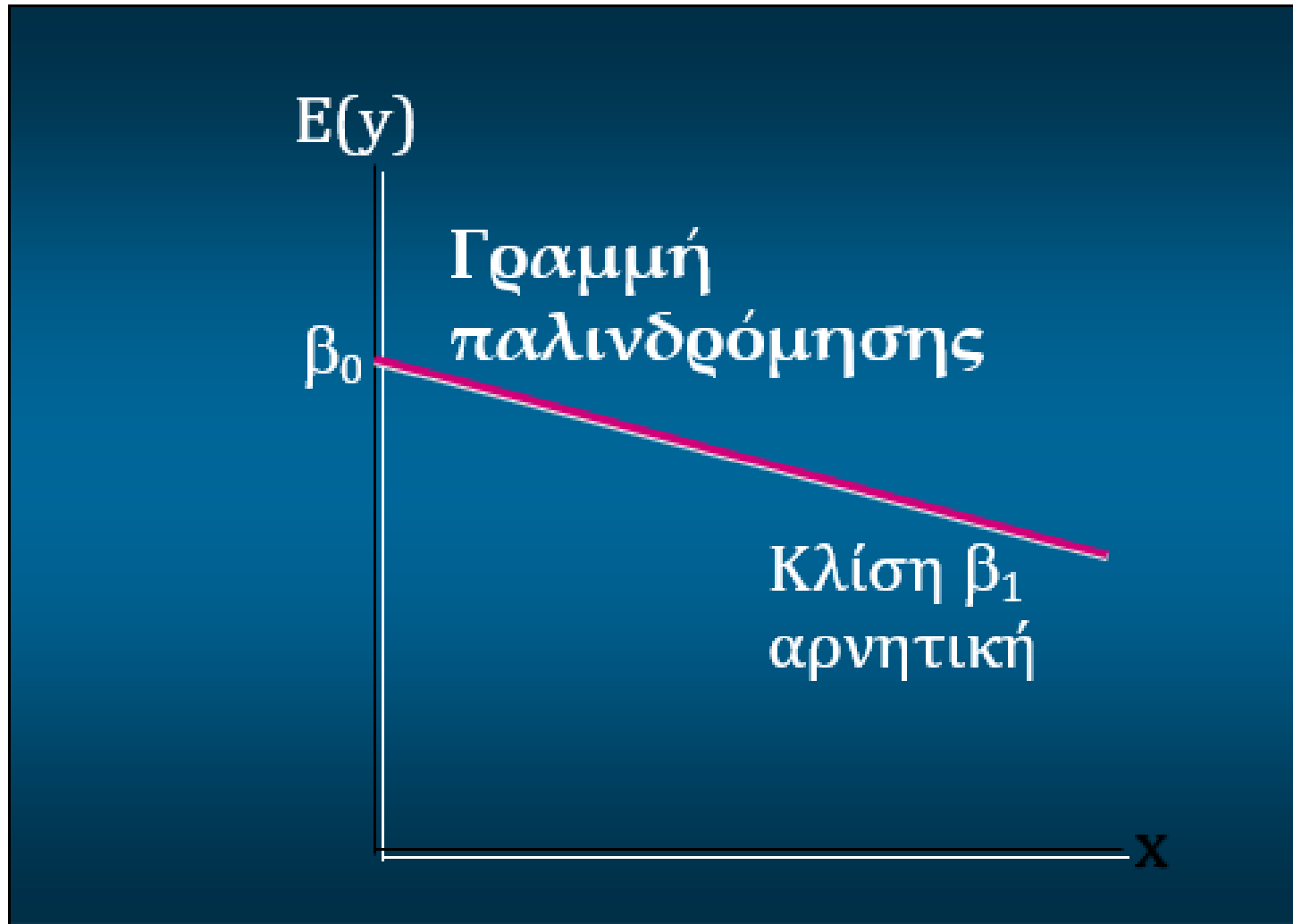
$$E(Y) = \beta_0 + \beta_1 X.$$

- Το γράφημα της εξίσωσης αυτής αποτελεί μία ευθεία γραμμή στο επίπεδο.
- Ο όρος β_0 αντιστοιχεί στο σημείο στο οποίο η ευθεία της παλινδρόμησης τέμνει τον κατακόρυφο άξονα (κατακόρυφη αποτέμνουσα).
- Ο όρος β_1 αποτελεί την κλίση της γραμμής παλινδρόμησης.
- Η ποσότητα $E(Y)$ αποτελεί την αναμενόμενη τιμή της μεταβλητής Y για κάποια δεδομένη τιμή $X = x$ της ανεξάρτητης μεταβλητής.

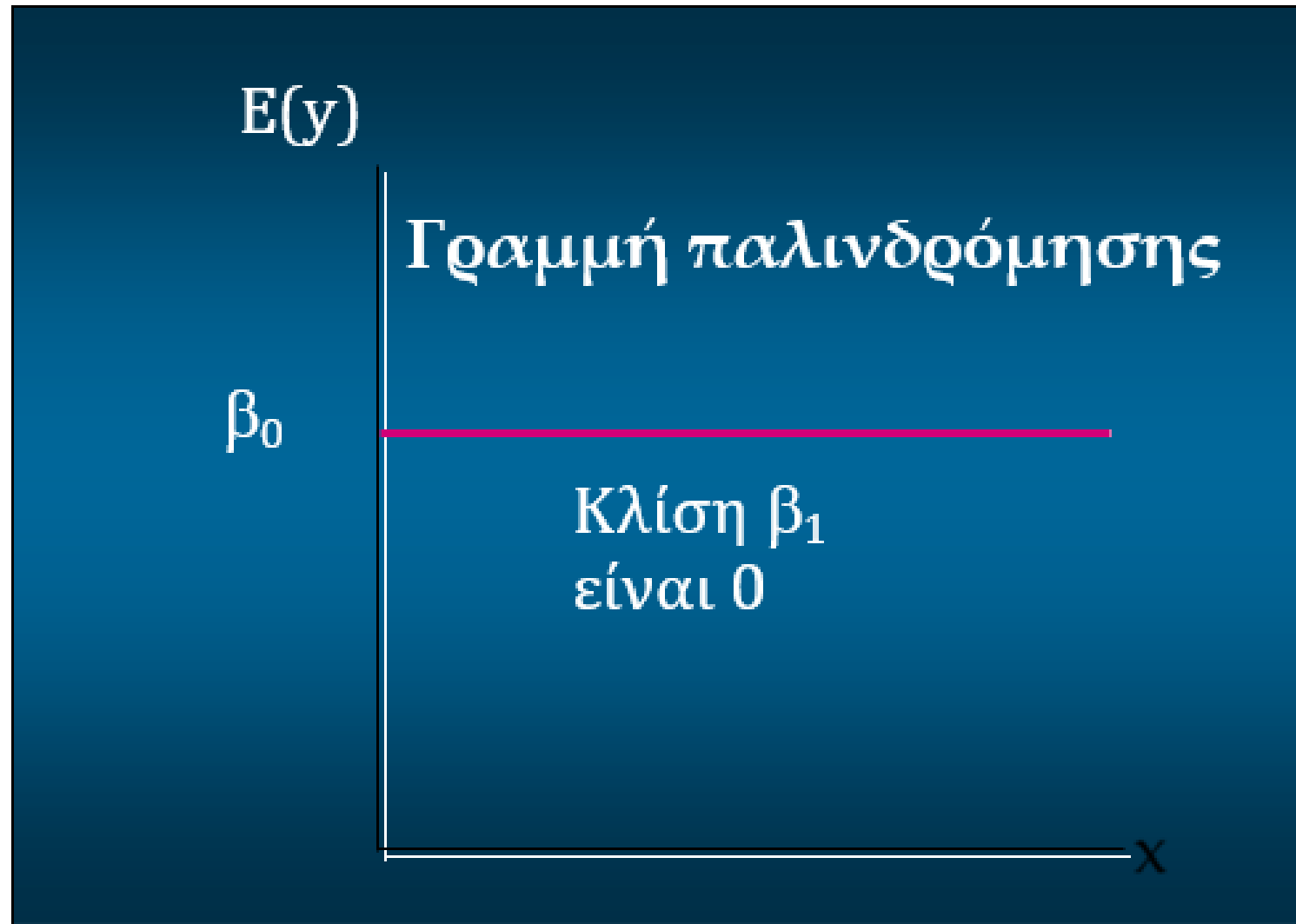
Μοντέλο γραμμικής παλινδρόμησης (θετική γραμμική σχέση)



Μοντέλο γραμμικής παλινδρόμησης (αρνητική γραμμική σχέση)



Μοντέλο γραμμικής παλινδρόμησης (απουσία γραμμικής σχέσης)



Εκτίμηση της εξίσωσης της απλής γραμμικής παλινδρόμησης

Ως εκτίμηση της εξίσωσης της απλής γραμμικής παλινδρόμησης από κάποια δεδομένα λογίζεται η σχέση:

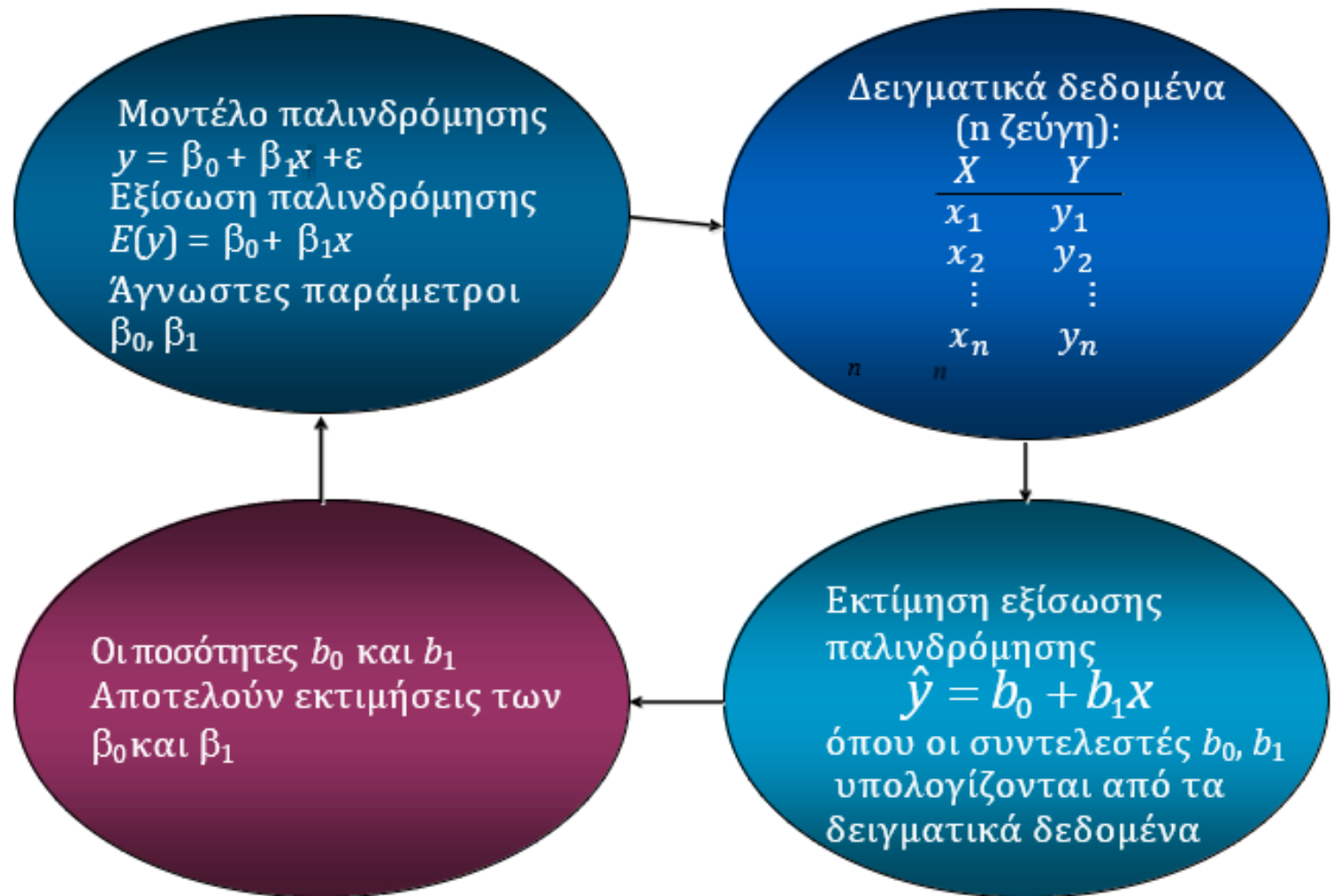
$$\hat{y} = b_0 + b_1 x.$$

- Το γράφημα της εξίσωσης αυτής αποτελεί μία εκτίμηση της γραμμής παλινδρόμησης.
- Οι όροι b_0, b_1 πρόκειται να προσδιοριστούν με χρήση των στατιστικών δεδομένων.
- Ο όρος b_0 αντιστοιχεί στο σημείο στο οποίο η εκτίμηση της γραμμής της παλινδρόμησης τέμνει τον κατακόρυφο άξονα (και αποτελεί εκτίμηση του αντίστοιχου όρου β_0 της γραμμής παλινδρόμησης).
- Ο όρος b_1 αποτελεί εκτίμηση της κλίσης β_1 της γραμμής παλινδρόμησης.
- Η ποσότητα $\hat{y}$ αποτελεί την εκτίμηση της τιμής που λαμβάνει η μεταβλητή Y για κάποια δεδομένη τιμή $X = x$ της ανεξάρτητης μεταβλητής.

Διαδικασία εκτίμησης παραμέτρων απλής γραμμικής παλινδρόμησης

- Θεωρούμε ένα τυχαίο δείγμα από n σε πλήθος ζεύγη παρατηρήσεων $\{(x_i, y_i); i = 1, 2, \dots, n\}$.
- Επιλέγοντας επιπλέον δείγματα με ακριβώς τις ίδιες τιμές x_i ($i = 1, \dots, n$) για την ανεξάρτητη μεταβλητή X , είναι εύλογο ότι οι αντίστοιχες παρατηρήσεις για την y θα λάμβαναν διαφορετικές τιμές. Συνεπώς, η τιμή y_i στο διατεταγμένο ζεύγος (x_i, y_i) είναι η τιμή κάποιας τ.μ. Y_i .
- Η εκτίμηση της γραμμής παλινδρόμησης είναι τόσο καλύτερη, όσο μεγαλύτερο είναι και το μέγεθος δείγματος που έχουμε στα χέρια μας.

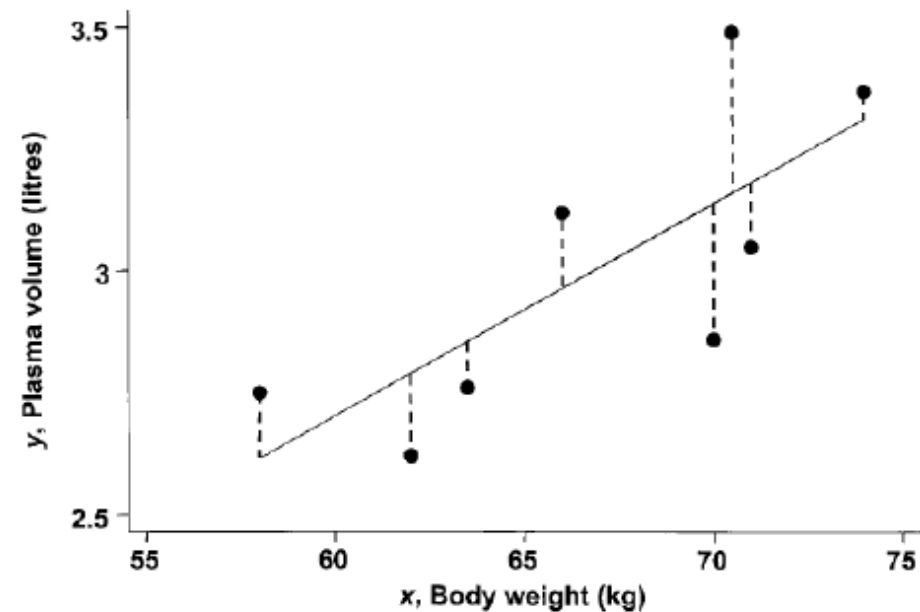
Διαδικασία εκτίμησης παραμέτρων απλής γραμμικής παλινδρόμησης (2)



Παράδειγμα δεδομένων για εφαρμογή απλής γραμμικής παλινδρόμησης

Παράδειγμα: Ο ακόλουθος Πίνακας περιλαμβάνει παρατηρήσεις για το σωματικό βάρος και τον όγκο του πλάσματος σε οκτώ υγιείς ενήλικους.

Subject	Body Weight (kg)	Plasma Volume (litres)
1	58	2,75
2	70	2,86
3	74	3,37
4	63,5	2,76
5	62	2,62
6	70,5	3,49
7	71	3,05
8	66	3,12



Υπολογιστικοί τύποι για τις ποσότητες b_0 και b_1

- Το κριτήριο με την οποία τα κατάλοιπα χρησιμοποιούνται για να προσδιοριστούν οι ποσότητες b_0, b_1 καλείται **κριτήριο ελαχίστων τετραγώνων**.
- Η διαδικασία αυτή θα περιγραφεί αναλυτικότερα στη συνέχεια, αλλά γίνεται αυτόματα με τη χρήση προγραμμάτων (SPSS) στον υπολογιστή.
- Όπως θα δούμε, οι ποσότητες b_0, b_1 υπολογίζονται βάσει των τύπων:

$$\text{Κλίση: } b_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$\text{Αποτέμνουσα: } b_0 = \bar{y} - b_1 \bar{x},$$

όπου $\bar{x}, \bar{y}$ η μέση τιμή της ανεξάρτητης και της εξαρτημένης μεταβλητής, αντίστοιχα.

Μέθοδος Ελαχίστων Τετραγώνων

Η πρώτη αναφορά με ολοκληρωμένη ανάπτυξη της μεθόδου των ελαχίστων τετραγώνων εμφανίζεται το 1805 σε μια εργασία του Γάλλου μαθηματικού Legendre, (1752-1833) και αμέσως μετά από το Γερμανό μαθηματικό Gauss, (1777-1855) στην αστρονομική του πραγματεία “Theoria Motus” για τον προσδιορισμό της τροχιάς του μικρού πλανήτη Δήμητρα.

Μάλιστα ο Gauss αναφέρει ότι χρησιμοποίησε την αρχή των ελαχίστων τετραγώνων πριν από το 1794 (σε ηλικία μόλις 17 ετών), έτσι ώστε να προηγείται του Legendre ως προς την ανακάλυψη αυτής της μεθόδου.

Η μέθοδος των ελαχίστων τετραγώνων

- Χάριν πληρότητας, στην ενότητα αυτή παρουσιάζουμε πώς προκύπτουν οι τύποι

$$\text{Κλίση: } b_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$\text{Αποτέμνουσα: } b_0 = \bar{y} - b_1 \bar{x},$$

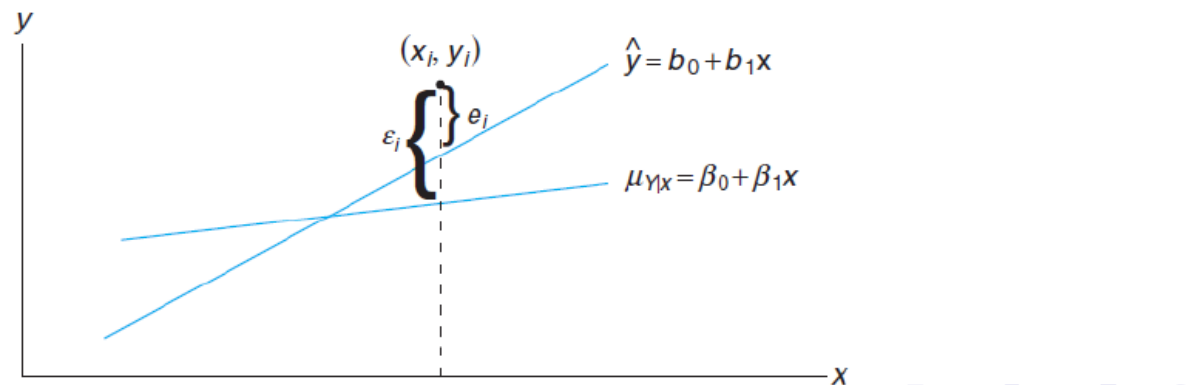
για την εκτίμηση των β_1, β_0 που παρουσιάστηκαν στην προηγούμενη διαφάνεια.

Η έννοια των καταλοίπων

- Δοθέντος ενός συνόλου από ζεύγη παρατηρήσεων $\{(x_i, y_i); i = 1, 2, \dots, n\}$ και μίας εκτίμησης του μοντέλου παλινδρόμησης, $\hat{y} = b_0 + b_1x$, το i -στό κατάλοιπο e_i (κατακόρυφη απόκλιση του σημείου- i από την γραμμή παλινδρόμησης) δίνεται από τη σχέση

$$e_i = y_i - \hat{y}_i, i = 1, 2, \dots, n.$$

- Ισοδύναμα, $y_i = b_0 + b_1x_i + e_i$.
- Όσο μικρότερα είναι τα κατάλοιπα αυτά, τόσο καλύτερη είναι η εφαρμογή του μοντέλου στα δεδομένα.



Υπολογιστικοί τύποι για b_0 και b_1 (Απόδειξη)

- Η παράμετρος β_i εκτιμάται από την ποσότητα b_i ($i = 0, 1$) με τέτοιο τρόπο, ώστε το άθροισμα των τετραγωνισμένων καταλοίπων

$$SSE = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - b_0 - b_1 x_i)^2$$

να ελαχιστοποιείται.

- Παραγωγίζοντας την ποσότητα SSE ως προς b_0 και b_1 , έχουμε

$$\begin{cases} \frac{\partial SSE}{\partial b_0} = -2 \sum_{i=1}^n (y_i - b_0 - b_1 x_i), \\ \frac{\partial SSE}{\partial b_1} = -2 \sum_{i=1}^n (y_i - b_0 - b_1 x_i) x_i, \end{cases}$$

- Θέτοντας τις μερικές αυτές παραγώγους ίσες με το μηδέν και ανδιατάσσοντας τους όρους, καταλήγουμε στο ακόλουθο γραμμικό σύστημα εξισώσεων (ως προς τους αγνώστους b_0 και b_1)

$$\begin{cases} nb_0 + (\sum_{i=1}^n x_i) b_1 = \sum_{i=1}^n y_i, \\ (\sum_{i=1}^n x_i) b_0 + (\sum_{i=1}^n x_i^2) b_1 = \sum_{i=1}^n x_i y_i, \end{cases}$$

Γραμμική παλινδρόμηση

Υπολογιστικοί τύποι για b_0 και b_1 (Απόδειξη-τέλος)

- Η βέλτιστη ευθεία που προσαρμόζεται στα δεδομένα έχει την μορφή

$$\hat{y} = b_0 + b_1 x,$$

όπου

$$\begin{cases} b_1 = \frac{n \sum_{i=1}^n x_i y_i - (\sum_{i=1}^n x_i)(\sum_{i=1}^n y_i)}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \\ b_0 = \frac{\sum_{i=1}^n y_i - b_1 \sum_{i=1}^n x_i}{n} = \bar{y} - b_1 \bar{x}. \end{cases}$$

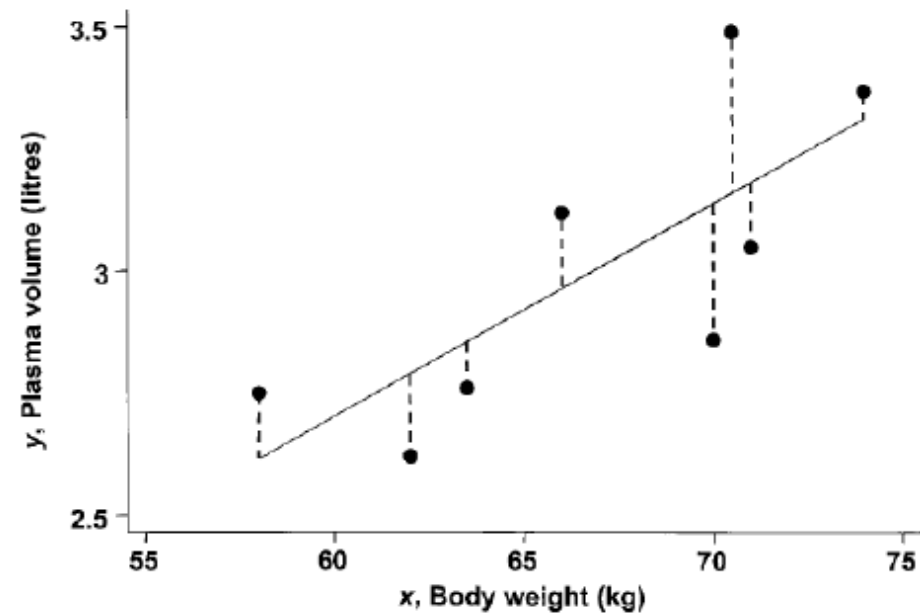
Ερμηνεία κλίσης μοντέλου απλής γραμμικής παλινδρόμησης

- Η κλίση του μοντέλου παλινδρόμησης (b_1) ερμηνεύεται ως η αναμενόμενη μεταβολή της Y (εξαρτημένης μεταβλητής) που επέρχεται από μία αύξηση της ανεξάρτητης μεταβολής (X) κατά μία μονάδα.
- Θετική κλίση υποδεικνύει θετική σχέση μεταξύ των X και Y . Δηλαδή υψηλότερες τιμές της X οδηγούν σε πρόβλεψη για υψηλότερες τιμές της Y (και αντίστροφα, χαμηλές τιμές της X οδηγούν σε πρόβλεψη χαμηλών τιμών της Y).
- Αρνητική κλίση υποδεικνύει αντίστροφη σχέση μεταξύ των X και Y . Δηλαδή υψηλότερες τιμές της X οδηγούν σε πρόβλεψη για χαμηλές τιμές της Y (και χαμηλές τιμές της X οδηγούν σε πρόβλεψη υψηλών τιμών της Y).
- Μηδενική κλίση υποδεικνύει ότι η ανεξάρτητη και η εξαρτημένη μεταβλητή δε συνδέονται από κάποια ερμηνευτική σχέση.
- Η μη κανονικοποιημένη κλίση της ευθείας παλινδρόμησης εξαρτάται, όπως έχουμε πει, από τις μονάδες μέτρησης και μπορεί θεωρητικά να λάβει οποιαδήποτε τιμή.

Παράδειγμα εκτίμησης μοντέλου απλής γραμμικής παλινδρόμησης

Παράδειγμα: Ο ακόλουθος Πίνακας περιλαμβάνει παρατηρήσεις για το σωματικό βάρος και τον όγκο του πλάσματος σε οκτώ υγιείς ενήλικους.

Subject	Body Weight (kg)	Plasma Volume (litres)
1	58	2,75
2	70	2,86
3	74	3,37
4	63,5	2,76
5	62	2,62
6	70,5	3,49
7	71	3,05
8	66	3,12



Παράδειγμα εκτίμησης μοντέλου απλής γραμμικής παλινδρόμησης (συνέχεια)

- Παρατηρώντας ότι $\sum_{i=1}^8 x_i = 535$, $\sum_{i=1}^8 y_i = 24.02$, $\sum_{i=1}^8 x_i^2 = 35983.5$ και $\sum_{i=1}^8 x_i y_i = 1615.295$, έχουμε

$$\begin{cases} b_1 = \frac{8 \cdot 1615.295 - 535 \cdot 24.02}{8 \cdot 35983.5 - (535)^2} = 0.044, \\ b_0 = \frac{24.02 - 0.044 \cdot 535}{8} = 0.086. \end{cases}$$

Subject	Body Weight (kg)		Plasma Volume (litres)		
	x_i	y_i	x_i^2	$x_i y_i$	
1	58	2,75	3364	159,5	
2	70	2,86	4900	200,2	
3	74	3,37	5476	249,38	
4	63,5	2,76	4032,25	175,26	
5	62	2,62	3844	162,44	
6	70,5	3,49	4970,25	246,045	
7	71	3,05	5041	216,55	
8	66	3,12	4356	205,92	
	535	24,02	35983,5	1615,295	

- Η βέλτιστη ευθεία που προσαρμόζεται στα δεδομένα είναι

$$Plasma_vol = 0.86 + 0.044 \cdot Weight$$

- Οι τιμές των συντελεστών b_0 και b_1 υπόκεινται σε δειγματοληπτική διακύμανση (αφού υπολογίζονται βάσει των παρατηρήσεων).
- Η ακρίβεια με την οποία οι παράμετροι β_0 και β_1 εκτιμώνται από τις ποσότητες b_0 και b_1 μετράται από τα τυπικά σφάλματα

Σχόλια για την κλίση μοντέλου απλής γραμμικής παλινδρόμησης

Σε σχέση με το μοντέλο

$$Plasma_vol = 0.86 + 0.044 \cdot Weight$$

που εκτιμήσαμε, έχουμε τις ακόλουθες παρατηρήσεις:

- Η κλίση του μοντέλου παλινδρόμησης ($b_1 = 0.044$) είναι θετική, το οποίο υποδηλώνει θετική σχέση μεταξύ των μεταβλητών X (βάρος) και Y (όγκος πλάσματος). Το γεγονός αυτός συμπίπτει με τη διαίσθησή μας.
- Ειδικότερα, η τιμή $b_1 = 0.044$ ερμηνεύεται ως εξής: μεταξύ δύο ενήλικες με βάρη που διαφέρουν κατά ένα κιλό, ο βαρύτερος από αυτός εμφανίζει όγκο πλάσματος στο σώμα του αυξημένο κατά 0.044 λίτρα σε σχέση με τον ελαφρύτερο (κατά μέσο όρο).
- Ο ίδιος συλλογισμός επεκτείνεται και για ευρύτερες διαφορές κατά τον προφανή τρόπο: Ένα άτομο που είναι βαρύτερο από κάποιο άλλο κατά 10 κιλά, εμφανίζει όγκο πλάσματος στο σώμα του αυξημένο κατά 0.44 λίτρα σε σχέση με τον ελαφρύτερο (κατά μέσο όρο).

Σχόλια για την σταθερά μοντέλου απλής γραμμικής παλινδρόμησης

Σε σχέση με το μοντέλο

$$Plasma_vol = 0.86 + 0.044 \cdot Weight$$

που εκτιμήσαμε, έχουμε τις ακόλουθες παρατηρήσεις:

- Η σταθερά του μοντέλου (b_0) εν γένει μπορεί να μην έχει φυσική σημασία. Αντιστοιχεί στην αναμενόμενη τιμή της εξαρτημένης μεταβλητής, όταν η ανεξάρτητη μεταβλητή (X) λαμβάνει μηδενική τιμή.
- Στην συγκεκριμένη περίπτωση ($b_0 = 0.86$), αντιστοιχεί στον αναμενόμενο όγκο πλάσματος (σε λίτρα) που έχει ένας ενήλικας με σωματικό βάρος 0 κιλά. Προφανώς, η φυσική αυτή ερμηνεία δεν έχει νόημα.
- Εν γένει, το μοντέλο που προκύπτει δε θα πρέπει να χρησιμοποιείται για προβλέψεις, εφαρμόζοντας τιμές στην ανεξάρτητη μεταβλητή X που βρίσκονται εκτός του εύρους των τιμών που παρατηρήθηκαν στο δείγμα. (Για το λόγο αυτό, και η περίπτωση $X = 0$ στην προηγούμενη περίπτωση υπήρξε προβληματική.)

Εκτίμηση μοντέλου με χρήση MS Excel

- Ο προσδιορισμός του μοντέλου με το χέρι, όπως παρουσιάστηκε προηγουμένως είναι ιδιαίτερα επίπονος και εφαρμόσιμος μόνο σε περιπτώσεις μικρών δειγμάτων.
- Η ερμηνευτική ικανότητα του μοντέλου είναι τόσο μεγαλύτερη, όσο μεγαλύτερο το δείγμα που χρησιμοποιήθηκε.
- Συνεπώς, στην πράξη χρησιμοποιούνται μεγάλα δείγματα και η διαδικασία διευκολύνεται με χρήση υπολογιστικών προγραμμάτων.
- Ακολούθως παρουσιάζεται η διαδικασία με χρήση του προγράμματος SPSS.

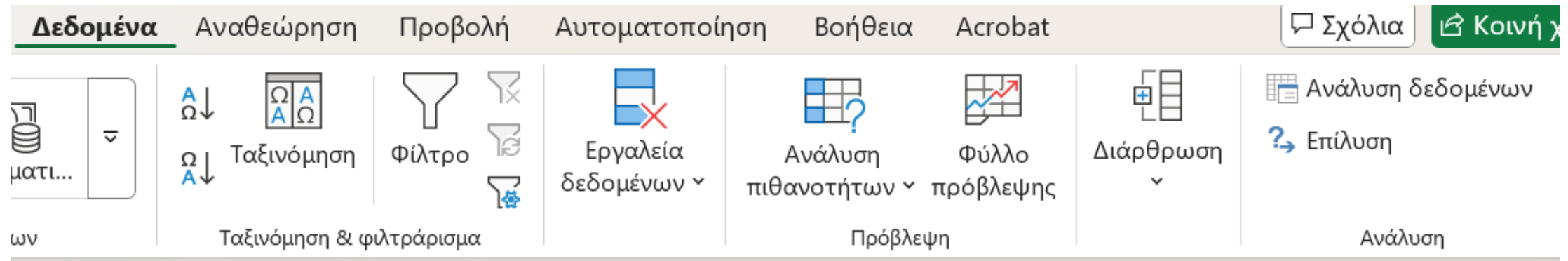
Καταχώριση δεδομένων

- Οι τιμές των δύο ποσοτικών μεταβλητών σε υπολογιστικό φύλλο του Excel:

A	D
weight	volume
58	2,75
70	2,86
74	3,37
64	2,76
62	2,62
71	3,49
71	3,05
66	3,12

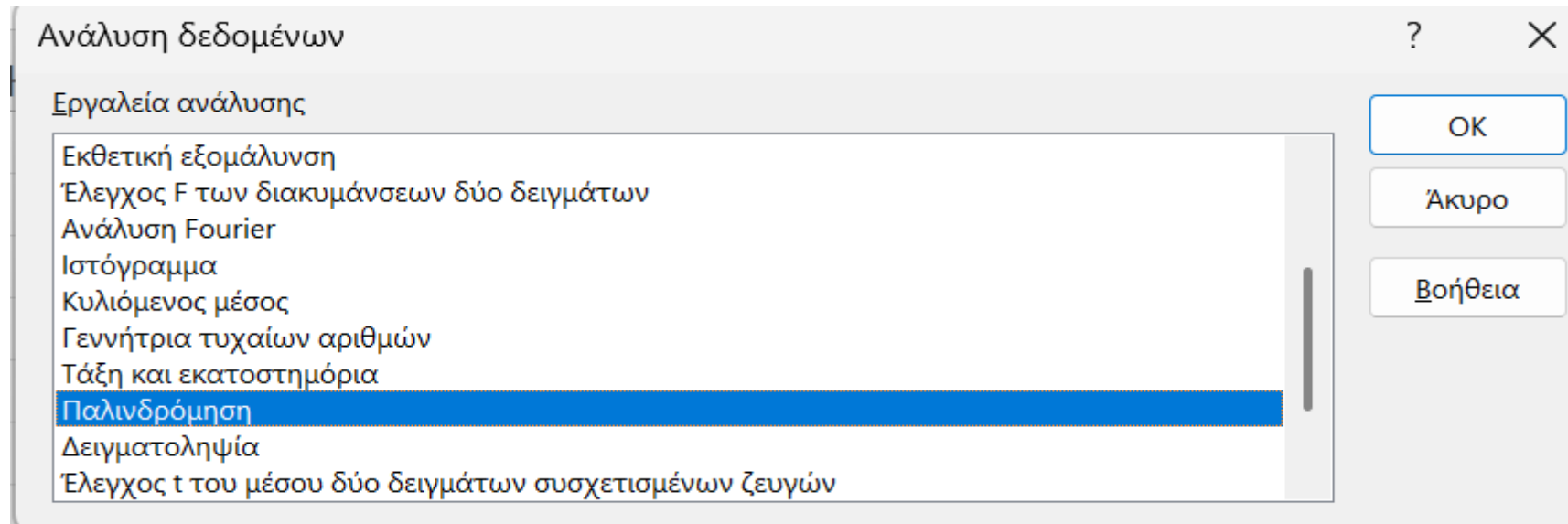
Γραμμική παλινδρόμηση στο Ms Excel: Διαδρομή

- Στην καρτέλα επιλογών «Δεδομένα», επιλέγουμε «Ανάλυση δεδομένων»:



Γραμμική παλινδρόμηση στο Ms Excel: Διαδρομή

- Μεταξύ των παρεχόμενων επιλογών για «Ανάλυση δεδομένων», επιλέγουμε «Παλινδρόμηση»:



Γραμμική παλινδρόμηση στο Ms Excel

- Επιλέγουμε καταλλήλως τις περιοχές όπου αναφέρονται οι τιμές τις ανεξάρτητης (X-weight) και εξαρτημένης (Y-volume) μεταβλητής:

The screenshot shows the 'Παλινδρόμηση' (Regression) dialog box in SPSS. The background is an Excel spreadsheet with two columns: 'weight' and 'volume'. The data points are as follows:

weight	volume
58	2,75
70	2,86
74	3,37
64	2,76
62	2,62
71	3,49
71	3,05
66	3,12

The 'Παλινδρόμηση' dialog box is open, showing the following settings:

- Εισόδος (Input):**
 - Περιοχή εισόδου Y: \$B\$2:\$B\$9
 - Περιοχή εισόδου X: \$A\$2:\$A\$9
 - ☐ Ετικέτες
 - ☐ Βαθμός εμπιστοσύνης: 95 %
 - ☐ Ο σταθερός όρος είναι μηδέν
- Επιλογές εξόδου (Output):**
 - ☐ Περιοχή εξόδου:
 - ☒ Νέο φύλλο:
 - ☐ Νέο βιβλίο εργασίας
- Υπόλοιπα (Residuals):**
 - ☐ Υπόλοιπα
 - ☐ Τυποποιημένα υπόλοιπα
 - ☐ Διάγραμμα υπολοίπων
 - ☐ Διάγραμμα προσαρμογής γραμμής
- Κανονική πιθανότητα (Normality):**
 - ☐ Διάγραμμα κανονικής πιθανότητας

Buttons: OK, Άκυρο, Βοήθεια

Γραμμική παλινδρόμηση στο Ms Excel

- Η επιλογή εξόδου των αποτελεσμάτων σε «Νέο Φύλλο» οδηγεί σε σχηματισμό νέας καρτέλας, όπου θα αναφέρονται τα σχετικά αποτελέσματα της ανάλυσης παλινδρόμησης:

Παλινδρόμηση

Είσοδος

Περιοχή εισόδου Y:

Περιοχή εισόδου X:

☐ Ετικέτες ☐ Ο σταθερός όρος είναι μηδέν

☐ Βαθμός εμπιστοσύνης %

Επιλογές εξόδου

☐ Περιοχή εξόδου:

☒ Νέο φύλλο:

☐ Νέο βιβλίο εργασίας

Υπόλοιπα

☐ Υπόλοιπα ☐ Διάγραμμα υπολοίπων

☐ Τυποποιημένα υπόλοιπα ☐ Διάγραμμα προσαρμογής γραμμής

Κανονική πιθανότητα

☐ Διάγραμμα κανονικής πιθανότητας

OK

Άκυρο

Βοήθεια

Γραμμική παλινδρόμηση στο Excel: Εξαγόμενα

ΕΞΟΔΟΣ ΣΥΜΠΕΡΑΣΜΑΤΟΣ									
Στατιστικά παλινδρόμησης									
Πολλαπλό R	0,759126577								
R Τετράγωνο	0,576273159								
Προσαρμοσμένο R Τετράγωνο	0,505652019								
Τυπικό σφάλμα	0,218809511								
Μέγεθος δείγματος	8								
ΑΝΑΛΥΣΗ ΔΙΑΚΥΜΑΝΣΗΣ									
	βαθμοί ελευθερίας	SS	MS	F	Σημαντικότητα F				
Παλινδρόμηση	1	0,390684388	0,390684388	8,160065927	0,028930913				
Υπόλοιπο	6	0,287265612	0,047877602						
Σύνολο	7	0,67795							
	Συντελεστές	Τυπικό σφάλμα	t	τιμή-P	Κατώτερο 95%	Υψηλότερο 95%	Κατώτερο 95,0%	Υψηλότερο 95,0%	
Τεταγμένη επί την αρχή	0,085724285	1,023998015	0,083715284	0,936006021	-2,419908594	2,591357164	-2,419908594	2,591357164	
Μεταβλητή Χ 1	0,043615338	0,015268361	2,856582911	0,028930913	0,006255005	0,08097567	0,006255005	0,08097567	

Γραμμική παλινδρόμηση στο Excel: Εξαγόμενα (2)

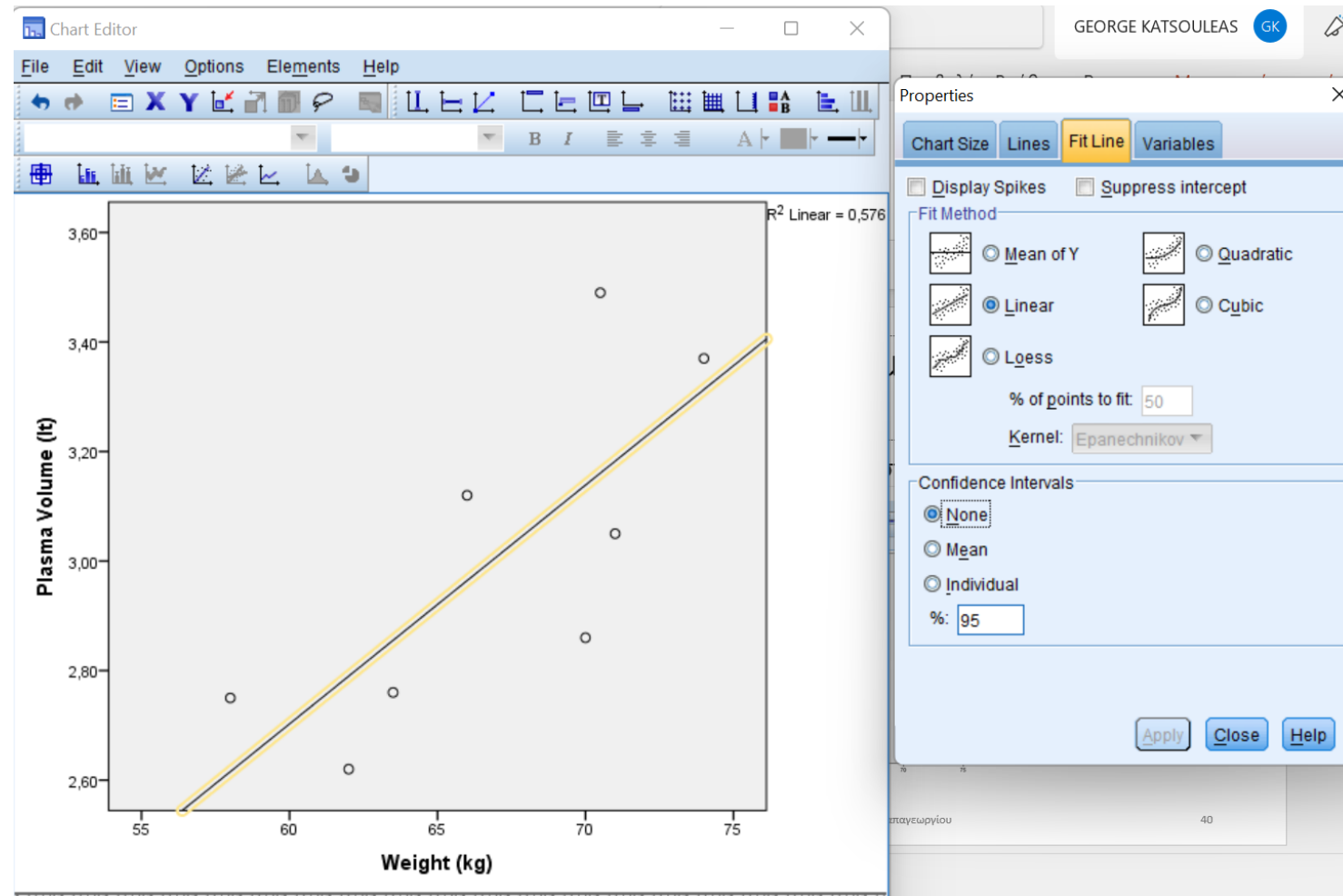
- Οι συντελεστές του μοντέλου

$$Plasma_vol = 0.86 + 0.044 \cdot Weight$$

προκύπτουν άμεσα στην 1η στήλη του τελευταίου από τους εξαγόμενους Πίνακες:

	Συντελεστές	Τυπικό σφάλμα	t	τιμή-P	Κατώτερο 95%	Υψηλότερο 95%	Κατώτερο 95,0%	Υψηλότερο 95,0%
Τεταγμένη επί την αρχή	0,085724285	1,023998015	0,083715284	0,936006021	-2,419908594	2,591357164	-2,419908594	2,591357164
Μεταβλητή X 1	0,043615338	0,015268361	2,856582911	0,028930913	0,006255005	0,08097567	0,006255005	0,08097567

Στικτό γράφημα με προσαρμοσμένη ευθεία

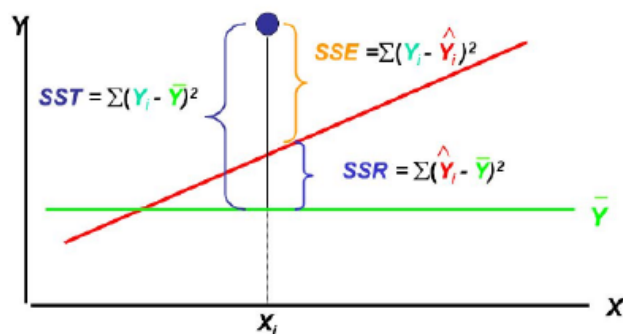


Συντελεστής προσδιορισμού

- Η προσέγγιση της **Ανάλυσης Διακύμανσης (Analysis-of-Variance)**: Στη διαδικασία αυτή, η συνολική μεταβλητότητα της εξαρτημένης μεταβλητής υποδιαιρείται σε δύο συνιστώσες.
- Ειδικότερα, αποδεικνύεται ότι

$$SST = SSR + SSE$$

$$\underbrace{\sum_{i=1}^n (y_i - \bar{y})^2}_{\text{Total Sum of Squares}} = \underbrace{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}_{\text{Sum of Squares due to regression}} + \underbrace{\sum_{i=1}^n (y_i - \hat{y}_i)^2}_{\text{Sum of Squares due to error}}$$



- **Συντελεστής προσδιορισμού (Coefficient of determination)**: $R^2 = 1 - \frac{SSE}{SST} \in [0, 1]$

Συντελεστής προσδιορισμού (συνέχεια)

Συντελεστής προσδιορισμού (Coefficient of determination):

$$R^2 = 1 - \frac{SSE}{SST} = \frac{SSR}{SST} \in [0, 1],$$

όπου

- SSR : Άθροισμα Τετραγώνων λόγω Παλινδρόμησης (Sum of Squares due to regression), $\sum_{i=1}^n (y_i - \bar{y})^2$
- SST : Συνολικό Άθροισμα Τετραγώνων (Total Sum of Squares) $\sum_{i=1}^n (y_i - \bar{y})^2$
- SSE : Συνολικό Άθροισμα Τετραγώνων λόγω Σφάλματος (Sum of Squares due to error) $\sum_{i=1}^n (y_i - \hat{y}_i)^2$

Αξιολόγηση καταλληλότητας μοντέλου

- Προφανώς το μοντέλο έχει ερμηνευτική ικανότητα, όταν ο συντελεστής κλίσης του (β_1) είναι **μη μηδενικός**.
- Στην πράξη, η αντίστοιχη εκτίμηση b_1 του β_1 θα είναι κάποιος (μη μηδενικός) αριθμός. Ακόμα και όταν ο αριθμός αυτός είναι (κατ' απόλυτο τιμή) μεγάλος, πρέπει να ελεγχθεί κατά πόσον επαρκεί, προκειμένου να θεωρήσουμε ότι η αντίστοιχη εκτιμώμενη ποσότητα είναι $\beta_1 \neq 0$.
- Ακολούθως, περιγράφεται ένας τρόπος ελέγχου των υποθέσεων:

$$H_0 : \beta_1 = 0 \text{ έναντι της εναλλακτικής } H_1 : \beta_1 \neq 0$$

- Στατιστικό κριτήριο:

$$F = \frac{SSR/1}{SSE/(n-2)} = \frac{MSR}{MSE} \sim F_{(1,n-2)}$$

- Απορριπτική περιοχή $R = (f_{\alpha;(1,n-2)}, \infty)$ σε στάθμη σημαντικότητας α .

ANOVA Table

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	Computed f
Regression	SSR	1	$MSR = \frac{SSR}{1}$	$f = \frac{MSR}{MSE}$
Error	SSE	$n - 2$	$MSE = \frac{SSE}{n-2}$	
Total	SST	$n - 1$		

Αξιολόγηση καταλληλότητας μοντέλου στο Excel: Εξαγόμενα

ΕΞΟΔΟΣ ΣΥΜΠΕΡΑΣΜΑΤΟΣ								
Στατιστικά παλινδρόμησης								
Πολλαπλό R	0,759126577							
R Τετράγωνο	0,576273159							
Προσαρμοσμένο R Τετράγωνο	0,505652019							
Τυπικό σφάλμα	0,218809511							
Μέγεθος δείγματος	8							
ΑΝΑΛΥΣΗ ΔΙΑΚΥΜΑΝΣΗΣ								
	βαθμοί ελευθερίας	SS	MS	F	Σημαντικότητα F			
Παλινδρόμηση	1	0,390684388	0,390684388	8,160065927	0,028930913			
Υπόλοιπο	6	0,287265612	0,047877602					
Σύνολο	7	0,67795						
	Συντελεστές	Τυπικό σφάλμα	t	τιμή-P	Κατώτερο 95%	Υψηλότερο 95%	Κατώτερο 95,0%	Υψηλότερο 95,0%
Τεταγμένη επί την αρχή	0,085724285	1,023998015	0,083715284	0,936006021	-2,419908594	2,591357164	-2,419908594	2,591357164
Μεταβλητή X 1	0,043615338	0,015268361	2,856582911	0,028930913	0,006255005	0,08097567	0,006255005	0,08097567

Συντελεστής προσδιορισμού στο Excel

- Στον 1ο από τους προκύπτοντες πίνακες, παρατηρούμε ότι $R^2 = 0.576$.

Στατιστικά παλινδρόμησης	
Πολλαπλό R	0,759126577
R Τετράγωνο	0,576273159
Προσαρμοσμένο R Τετράγωνο	0,505652019
Τυπικό σφάλμα	0,218809511
Μέγεθος δείγματος	8

- Το μέτρο αυτό είναι ενδεικτικό της ερμηνευτικής ικανότητας του μοντέλου. Σημαίνει ότι 57,6% της μεταβλητότητας της εξαρτημένης μεταβλητής (όγκος πλάσματος) ερμηνεύεται από την ανεξάρτητη μεταβλητή (σωματικό βάρος). Η υπόλοιπη μεταβλητότητα μπορεί να αποδοθεί σε άλλους παράγοντες.

Συντελεστής προσδιορισμού στο Excel (2)

- Σημειώνουμε επίσης ότι $R^2 = \frac{SSR}{SST} = \frac{0.391}{0.678} = 0.576$, χρησιμοποιώντας τα δεδομένα του πίνακα ANOVA.

Στατιστικά παλινδρόμησης					
Πολλαπλό R	0,759126577				
R Τετράγωνο	0,576273159				
Προσαρμοσμένο R Τετράγωνο	0,505652019				
Τυπικό σφάλμα	0,218809511				
Μέγεθος δείγματος	8				
ΑΝΑΛΥΣΗ ΔΙΑΚΥΜΑΝΣΗΣ					
	βαθμοί ελευθερίας	SS	MS	F	Σημαντικότητα F
Παλινδρόμηση	1	0,390684388	0,390684388	8,160065927	0,028930913
Υπόλοιπο	6	0,287265612	0,047877602		
Σύνολο	7	0,67795			

Αξιολόγηση καταλληλότητας μοντέλου στο Excel: Ερμηνεία πίνακα ANOVA

- Για τον έλεγχο των υποθέσεων:

$$H_0 : \beta_1 = 0 \text{ έναντι της εναλλακτικής } H_1 : \beta_1 \neq 0,$$

στον πίνακα ANOVA, είναι άμεσα υπολογισμένο το στατιστικό κριτήριο:

$$F = \frac{SSR/1}{SSE/(n-2)} = \frac{MSR}{MSE} = 8.160$$

- Άμεσα παρέχεται και το αντίστοιχο p-value: Sig=0.029
- Αφού $\text{Sig}=0.029 < \alpha = 0.05$, μπορούμε να πούμε ότι η μηδενική υπόθεση απορρίπτεται σε στάθμη σημαντικότητας $\alpha = 0.05$.
- Συνεπώς, το μοντέλο κρίνεται επαρκές.

ΑΝΑΛΥΣΗ ΔΙΑΚΥΜΑΝΣΗΣ					
	βαθμοί ελευθερίας	SS	MS	F	Σημαντικότητα F
Παλινδρόμηση	1	0,390684388	0,390684388	8,160065927	0,028930913
Υπόλοιπο	6	0,287265612	0,047877602		
Σύνολο	7	0,67795			

Σχέση συντελεστή προσδιορισμού r^2 με δειγματικό συντελεστή συσχέτισης X, Y

Ο δειγματικός συντελεστής συσχέτισης των μεταβλητών X, Y βρίσκεται από τα
εξαγόμενα του SPSS ως εξής:

$$r_{XY} = (\text{Πρόσημο του } b_1) \sqrt{\text{Συντελεστής προσδιορισμού}},$$

$$r_{XY} = (\text{Πρόσημο του } b_1) \sqrt{r^2},$$

όπου

b_1 : η κλίση του εκτιμημένου μοντέλου παλινδρόμησης.

Στο Παράδειγμά μας είναι $r = +\sqrt{0.576} = 0.759$ και δίνεται στην πρώτη στήλη του
ακόλουθου πίνακα:

Στατιστικά παλινδρόμησης	
Πολλαπλό R	0,759126577
R Τετράγωνο	0,576273159
Προσαρμοσμένο R Τετράγωνο	0,505652019
Τυπικό σφάλμα	0,218809511
Μέγεθος δείγματος	8

Υποθέσεις σχετικά με τον όρο σφάλματος ϵ

- Ο όρος σφάλματος στο μοντέλο

$$Y = b_0 + b_1X + \epsilon$$

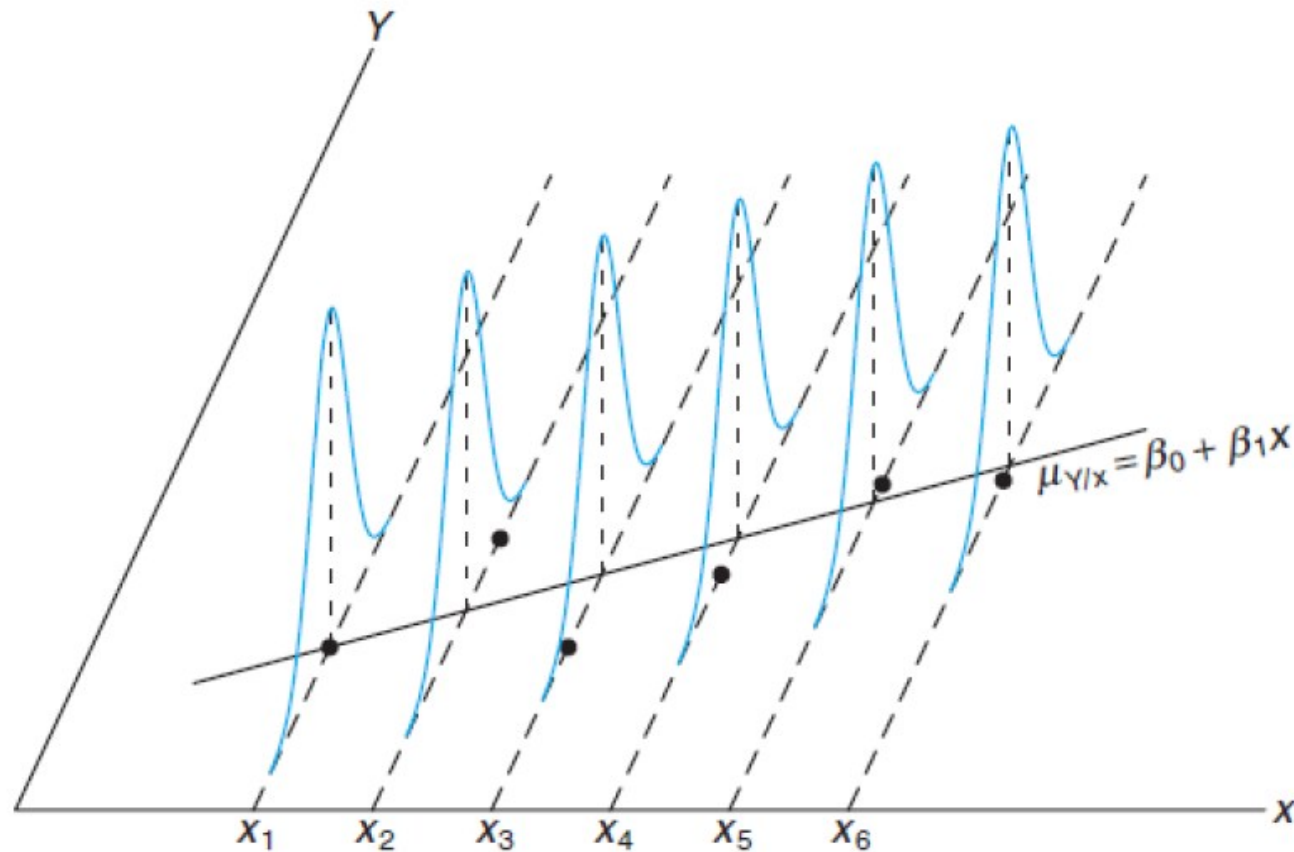
αποτελεί τυχαία μεταβλητή (τ.μ.) με μηδενική μέση τιμή.

- Η διακύμανση του όρου σφάλματος είναι σταθερή (συμβολίζεται με σ^2) για όλες τις τιμές της ανεξάρτητης μεταβλητής.
- Οι τιμές του ϵ είναι ανεξάρτητες.
- Ο όρος σφάλματος ακολουθεί την κανονική κατανομή.

Συνοπτικά, μπορούμε να συμβολίσουμε ότι:

$$\epsilon \sim \mathcal{N}(0, \sigma^2).$$

Κατανομή μεμονωμένων παρατηρήσεων γύρω από τη θεωρητική γραμμή παλινδρόμησης



Μία εκτίμηση s^2 για την διακύμανση σ^2 του ε

- Αποδεικνύεται ότι το μέσο τετραγωνικό σφάλμα (MSE - mean square error)

$$MSE = \frac{SSE}{n - 2},$$

όπου

$$SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2,$$

αποτελεί εκτημήτρια του σ^2 .

- Για το λόγο αυτό, συμβολίζεται επίσης:

$$MSE \equiv s^2.$$

- Υπνεθυμίζουμε ότι η ποσότητα αυτή βρίσκεται στον πίνακα ANOVA (μεταξύ των εξαγομένων του προγράμματος).

ANOVA Table				
Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	Computed f
Regression	SSR	1	$MSR = \frac{SSR}{1}$	$f = \frac{MSR}{MSE}$
Error	SSE	$n - 2$	$MSE = \frac{SSE}{n-2}$	
Total	SST	$n - 1$		

Τυπικό σφάλμα (Standard error of the estimate)

- Προκειμένου να εκτιμήσουμε την τυπική απόκλιση του όρου σφάλματος ϵ , αρκεί να πάρουμε την τετραγωνική ρίζα του MSE .
- Δηλαδή, η ποσότητα

$$s = \sqrt{MSE} = \sqrt{\frac{SSE}{n-2}}$$

αναφέρεται ως **τυπικό σφάλμα εκτίμησης (standard error of the estimate)**.

- Όπως θα δούμε ακολούθως, η ποσότητα s διαδραματίζει κεντρικό ρόλο στο σχηματισμό διαστημάτων εμπιστοσύνης, στον έλεγχο υποθέσεων και γενικότερα στην επαγωγή με χρήση του μοντέλου παλινδρόμησης.

Παράδειγμα 2

Παράδειγμα 2

Ένας οικονομικός αναλυτής θέλει να διερευνήσει τη σχέση μεταξύ των πωλήσεων μιας εταιρίας και των δαπανών διαφήμισής της.

Τα δεδομένα που έχει στην διάθεση του αφορούν το διάστημα των επτά προηγούμενων ετών και δίνονται από τον ακόλουθο πίνακα.

(η μεταβλητή X εκφράζει τις ετήσιες δαπάνες διαφήμισης της εταιρίας σε χρηματικές μονάδες και η μεταβλητή Y τις ετήσιες πωλήσεις της στις ίδιες χρηματικές μονάδες).

Παράδειγμα 2

Τα δεδομένα δίνονται στον διπλανό Πίνακα.

Ετήσιες Πωλήσεις, σε χ.μ. (Y)	Ετήσιες Δαπάνες Διαφήμισης, σε χ.μ. (X)
21	5
23	7
27	9
30	10
35	12
50	13
45	14

Παράδειγμα 2

1. Δεδομένου ότι η σχέση μεταξύ των μεταβλητών X και Y είναι γραμμική να εκτιμήσετε τις παραμέτρους β_0 και β_1 του μοντέλου $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ με την μέθοδο των ελαχίστων τετραγώνων και να ερμηνευθούν οι συντελεστές.
2. Με βάση το γραμμικό μοντέλο $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ να εκτιμήσετε τις ετήσιες πωλήσεις της εταιρίας όταν οι ετήσιες δαπάνες για διαφήμιση είναι 10 χρηματικές μονάδες.
3. Να υπολογισθεί ο συντελεστής συσχέτισης και να ερμηνευθεί.
4. Να υπολογισθεί ο συντελεστής προσδιορισμού και να ερμηνευθεί.

Παράδειγμα 2

Λύση

1. Για διευκόλυνση στους υπολογισμούς μας δημιουργούμε τον παρακάτω πίνακα

X_i	Y_i	$X_i - \bar{X}$	$(X_i - \bar{X})^2$	$Y_i - \bar{Y}$	$(Y_i - \bar{Y})^2$	$(X_i - \bar{X})(Y_i - \bar{Y})$	$(X_i - \bar{X})Y_i$	X_i^2	Y_i^2	$X_i Y_i$
5	21	-5	25	-12	144	60	-105	25	441	105
7	23	-3	9	-10	100	30	-69	49	529	161
9	27	-1	1	-6	36	6	-27	81	729	243
10	30	0	0	-3	9	0	0	100	900	300
12	35	2	4	2	4	4	70	144	1225	420
13	50	3	9	17	289	51	150	169	2500	650
14	45	4	16	12	144	48	180	196	2025	630
70	231	0	64	0	726	199	199	764	8349	2509

Παράδειγμα 2

Χρήσιμες ποσότητες:

$$\bar{X} = \frac{70}{7} = 10$$

$$\bar{Y} = \frac{231}{7} = 33$$

$$S_{XX} = \sum_{i=1}^7 (X_i - \bar{X})^2 = 64$$

$$S_{YY} = \sum_{i=1}^7 (Y_i - \bar{Y})^2 = 726$$

$$S_{XY} = S_{YX} = \sum_{i=1}^7 (X_i - \bar{X})(Y_i - \bar{Y}) = 199$$

Παράδειγμα 2

Ή εναλλακτικά

$$S_{XX} = \sum_{i=1}^7 X_i^2 - \frac{(\sum_{i=1}^7 X_i)^2}{7} = 764 - 700 = 64$$

$$S_{YY} = \sum_{i=1}^7 Y_i^2 - \frac{(\sum_{i=1}^7 Y_i)^2}{7} = 8349 - 7623 = 726$$

$$S_{XY} = S_{YX} = \sum_{i=1}^7 X_i Y_i - \frac{(\sum_{i=1}^7 X_i)(\sum_{i=1}^7 Y_i)}{7} = 2509 - 2310 = 199$$

Παράδειγμα 2

Για την ευθεία ελαχίστων τετραγώνων

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

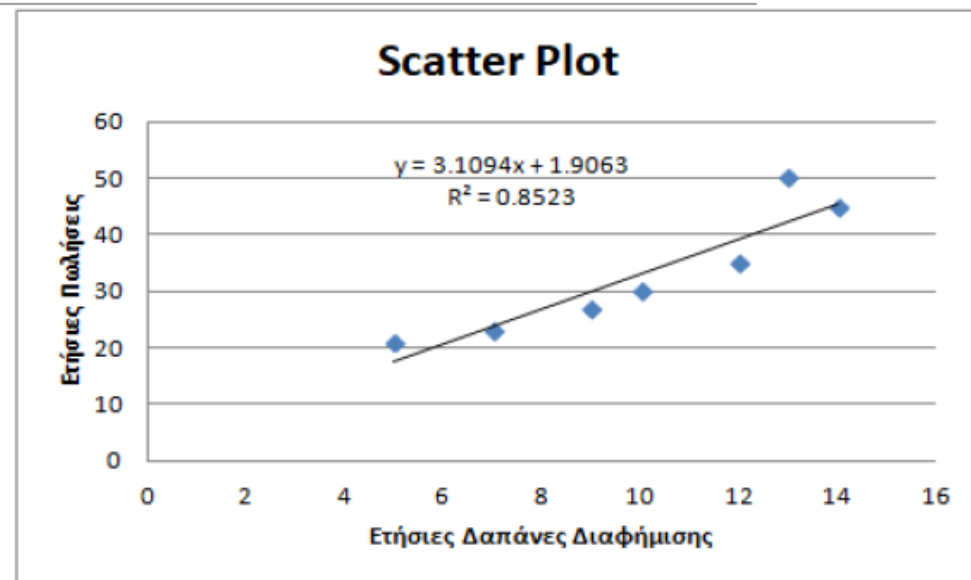
Έχουμε ότι

$$\hat{\beta}_1 = \frac{S_{XY}}{S_{XX}} = \frac{199}{64} = 3,11$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} = 33 - 3,11 \cdot 10 = 1,9$$

Άρα η ευθεία ελάχιστων τετραγώνων είναι η

$$\hat{y} = 1,9 + 3,11x$$



Παράδειγμα 2

Ερμηνεία συντελεστών:

- $\widehat{\beta}_0 = 1,9$: Δεδομένου ότι η εταιρία δεν θα έχει δαπάνες για διαφημίσεις (δηλαδή $X = 0$) το αναμενόμενο ύψος των ετήσιων πωλήσεων της εταιρίας είναι 1,9 χρηματικές μονάδες.

Παράδειγμα 2

Ερμηνεία συντελεστών:

- $\hat{\beta}_1 = 3,11$: Βάσει του μοντέλου το ύψος αναμενόμενων ετήσιων πωλήσεων της εταιρίας αυξάνεται κατά 3,11 χρηματικές μονάδες για κάθε μοναδιαία αύξηση των δαπανών για διαφήμιση.

Παράδειγμα 2

2. Θέλουμε να εκτιμήσουμε την τιμή του Y γνωρίζοντας ότι $X = 10$. Γνωρίζοντας ότι η ευθεία παλινδρόμησης είναι η

$$\hat{y} = 1,9 + 3,11x$$

Παράδειγμα 2

Για $x = 10$ έχουμε:

$$\hat{y} = 1,9 + 3,11 \cdot 10 = 1,9 + 31,1 = 33$$

Άρα όταν οι ετήσιες δαπάνες της διαφήμισης είναι 10 χρηματικές μονάδες, τότε η αξία των αναμενόμενων πωλήσεων είναι 33 χρηματικές μονάδες.

Παράδειγμα 2

3. Ο συντελεστής συσχέτισης του Pearson υπολογίζεται ως εξής:

$$r = \frac{S_{xy}}{S_x S_y} = \frac{199}{\sqrt{64} \cdot \sqrt{726}} = 0,92$$

Άρα συμπεραίνουμε ότι υπάρχει ισχυρή θετική συσχέτιση μεταξύ των ετήσιων πωλήσεων της εταιρίας και των αντίστοιχων δαπανών διαφήμισης.

Παράδειγμα 2

4. Ο συντελεστής προσδιορισμού R^2 υπολογίζεται ως εξής:

$$R^2 = \frac{(S_{XY})^2}{S_{XX}S_{YY}} = \frac{(199)^2}{64 \cdot 726} = 0,8523$$

Παράδειγμα 2

4. Επομένως συμπεραίνουμε ότι οι ετήσιες δαπάνες διαφήμισης της εταιρίας ερμηνεύουν το 85,23% της μεταβλητότητας των δεδομένων από τις ετήσιες πωλήσεις της.

Με άλλα λόγια μέσω των δεδομένων των ετήσιων δαπανών το γραμμικό μοντέλο έχει επεξηγηματική ικανότητα για τα δεδομένα πωλήσεων περίπου 85,%

Παράδειγμα 2

Προσοχή:

Γενικά δεν ισχύει ότι $R^2 = r^2$.

Παρόλα αυτά η ισότητα ισχύει στην περίπτωση της απλής γραμμικής παλινδρόμησης.

Το κουαρτέτο του Anscombe

- Όπως θα δούμε στη συνέχεια, η επιβεβαίωση της ισχύος των υποθέσεων αυτών γίνεται σε μεγάλο βαθμό με την εξέταση των γραφημάτων των καταλοίπων της παλινδρόμησης.
- Προτού εκθέσουμε τις υποθέσεις της γραμμικής παλινδρόμησης, παρουσιάζουμε πρώτα ένα περίφημο σχετικό σετ δεδομένων που κατασκεύασε το 1973 σε μία διάσημη εργασία του ο Anscombe (Graphs in Statistical Analysis. American Statistician, 27, 17-21) που καταδεικνύει τη σημασία της επαλήθευσης των υποθέσεων της γραμμικής παλινδρόμησης.
- Το σετ αυτό αποτελείται από 4 ζεύγη μεταβλητών $(x_1, y_1), (x_2, y_2), (x_3, y_3), (x_4, y_4)$ με 11 παρατηρήσεις η καθεμία.
- Οι μεταβλητές x_1, x_2, x_3, x_4 εμφανίζουν ακριβώς ίδια βασικά περιγραφικά στατιστικά, όπως επίσης και οι y_1, y_2, y_3, y_4 .
- Επιπλέον, καθένα από τα ζεύγη $(x_1, y_1), (x_2, y_2), (x_3, y_3), (x_4, y_4)$ εμφανίζει πανομοιότυπο συντελεστή συσχέτισης.

Το κουαρτέτο του Anscombe

	y1	x1	y2	x2	y3	x3	y4	x4	
1	8,04	10	9,14	10	7,46	10	6,58	8	
2	6,95	8	8,14	8	6,77	8	5,76	8	
3	7,58	13	8,74	13	12,74	13	7,71	8	
4	8,81	9	8,77	9	7,11	9	8,84	8	
5	8,33	11	9,26	11	7,81	11	8,47	8	
6	9,96	14	8,10	14	8,84	14	7,04	8	
7	7,24	6	6,13	6	6,08	6	5,25	8	
8	4,26	4	3,10	4	5,39	4	12,50	19	
9	10,84	12	9,13	12	8,15	12	5,56	8	
10	4,82	7	7,26	7	6,42	7	7,91	8	
11	5,68	5	4,74	5	5,73	5	6,89	8	
12									

Το κουαρτέτο του Anscombe (Περιγραφικά στατιστικά)

Descriptive Statistics

	N	Mean	Std. Deviation
X1	11	9,00	3,317
X2	11	9,00	3,317
X3	11	9,00	3,317
X4	11	9,00	3,317
Valid N (listwise)	11		

Descriptive Statistics

	N	Mean	Std. Deviation
Y1	11	7,5009	2,03157
Y2	11	7,5009	2,03166
Y3	11	7,5000	2,03042
Y4	11	7,5009	2,03058
Valid N (listwise)	11		

Το κουαρτέτο του Anscombe (Συντελεστές συσχέτισης)

Correlations

		Y1	X1
Y1	Pearson Correlation	1	,816**
	Sig. (2-tailed)		,002
	N	11	11
X1	Pearson Correlation	,816**	1
	Sig. (2-tailed)	,002	
	N	11	11

** . Correlation is significant at the 0.01 level (2-tailed).

Correlations

		Y3	X3
Y3	Pearson Correlation	1	,816**
	Sig. (2-tailed)		,002
	N	11	11
X3	Pearson Correlation	,816**	1
	Sig. (2-tailed)	,002	
	N	11	11

** . Correlation is significant at the 0.01 level (2-tailed).

Correlations

		Y2	X2
Y2	Pearson Correlation	1	,816**
	Sig. (2-tailed)		,002
	N	11	11
X2	Pearson Correlation	,816**	1
	Sig. (2-tailed)	,002	
	N	11	11

** . Correlation is significant at the 0.01 level (2-tailed).

Correlations

		Y4	X4
Y4	Pearson Correlation	1	,817**
	Sig. (2-tailed)		,002
	N	11	11
X4	Pearson Correlation	,817**	1
	Sig. (2-tailed)	,002	
	N	11	11

** . Correlation is significant at the 0.01 level (2-tailed).

Παλινδρομήσεις στα σεντ του Anscombe

- Το ενδιαφέρον είναι ότι αν τρέξουμε απλή γραμμική παλινδρόμηση σε καθένα από τα ζεύγη $(x_1, y_1), (x_2, y_2), (x_3, y_3), (x_4, y_4)$, προκύπτουν ακριβώς τα ίδια γραμμικά μοντέλα, αλλά πανομοιότυποι πίνακες ANOVA για την αξιολόγηση των μοντέλων:

Παλινδρομήσεις στα σετ του Anscombe (2)

ANOVA^a

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	27,510	1	27,510	17,990	,002 ^b
	Residual	13,763	9	1,529		
	Total	41,273	10			

a. Dependent Variable: Y1

b. Predictors: (Constant), X1

ANOVA^a

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	27,500	1	27,500	17,966	,002 ^b
	Residual	13,776	9	1,531		
	Total	41,276	10			

a. Dependent Variable: Y2

b. Predictors: (Constant), X2

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	3,000	1,125		2,667	,026
	X1	,500	,118	,816	4,241	,002

a. Dependent Variable: Y1

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	3,001	1,125		2,667	,026
	X2	,500	,118	,816	4,239	,002

a. Dependent Variable: Y2

Παλινδρομήσεις στα σετ του Anscombe (3)

ANOVA^a

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	27,470	1	27,470	17,972	,002 ^b
	Residual	13,756	9	1,528		
	Total	41,226	10			

a. Dependent Variable: Y3

b. Predictors: (Constant), X3

ANOVA^a

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	27,490	1	27,490	18,003	,002 ^b
	Residual	13,742	9	1,527		
	Total	41,232	10			

a. Dependent Variable: Y4

b. Predictors: (Constant), X4

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	3,002	1,124		2,670	,026
	X3	,500	,118	,816	4,239	,002

a. Dependent Variable: Y3

Coefficients^a

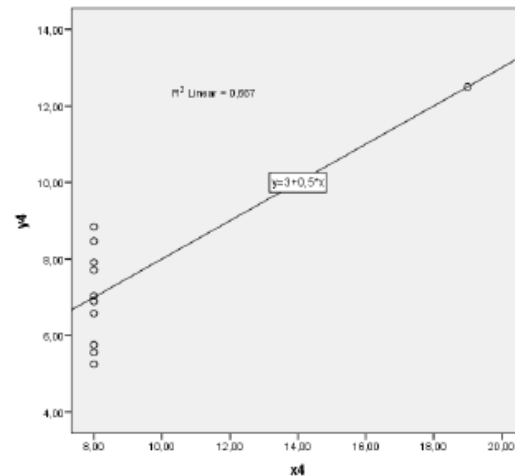
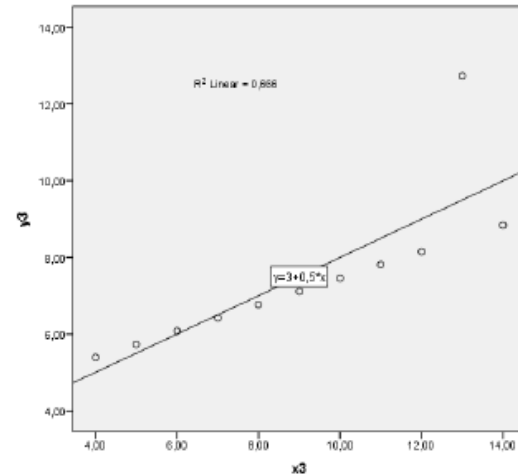
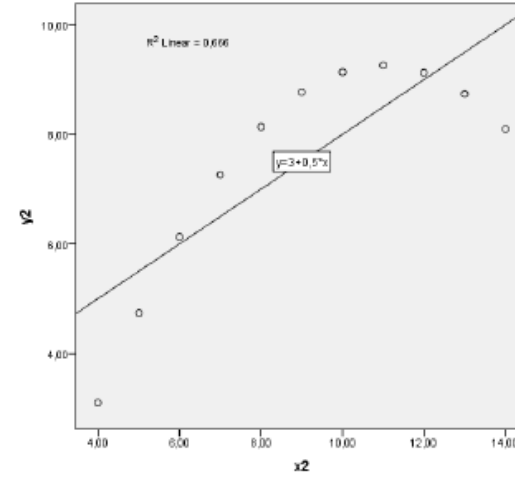
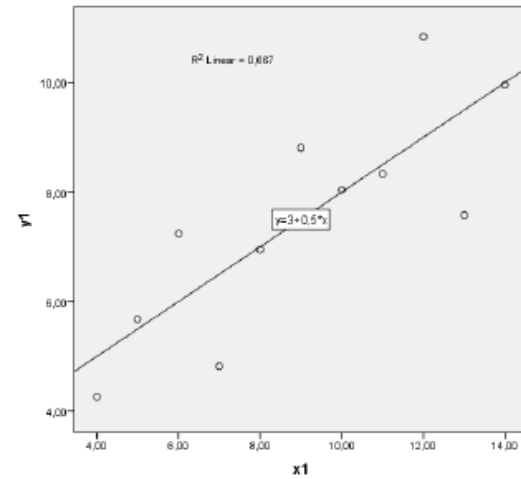
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	3,002	1,124		2,671	,026
	X4	,500	,118	,817	4,243	,002

a. Dependent Variable: Y4

Παλινδρομήσεις στα σεντ του Anscombe (4)

- Και στις τέσσερις περιπτώσεις καταλήγουμε στο ίδιο μοντέλο:
$$Y_i = 3 + 0,5 X_i \quad (i = 1,2,3,4).$$
- Με βάση τα εξαγόμενα που παρουσιάστηκαν στις προηγούμενες διαφάνειες, θα έλεγε κανείς ότι και τα τέσσερα μοντέλα είναι το ίδιο ικανοποιητικά στην περιγραφή της σχέσης των ζευγών $(x_1, y_1), (x_2, y_2), (x_3, y_3), (x_4, y_4)$.
- Είναι όμως έτσι?
- Μία σημαντική παράμετρος που έχει αγνοηθεί στην μέχρι τώρα ανάλυση των τεσσάρων σεντ είναι η οπτική τους εικόνα.
- Ας δούμε τα αντίστοιχα στικτά διαγράμματα με τις αντίστοιχες ευθείες παλινδρόμησης:

Οπτικοποίηση των σετ του Anscombe



Η σημασία της ανάλυσης των καταλοίπων

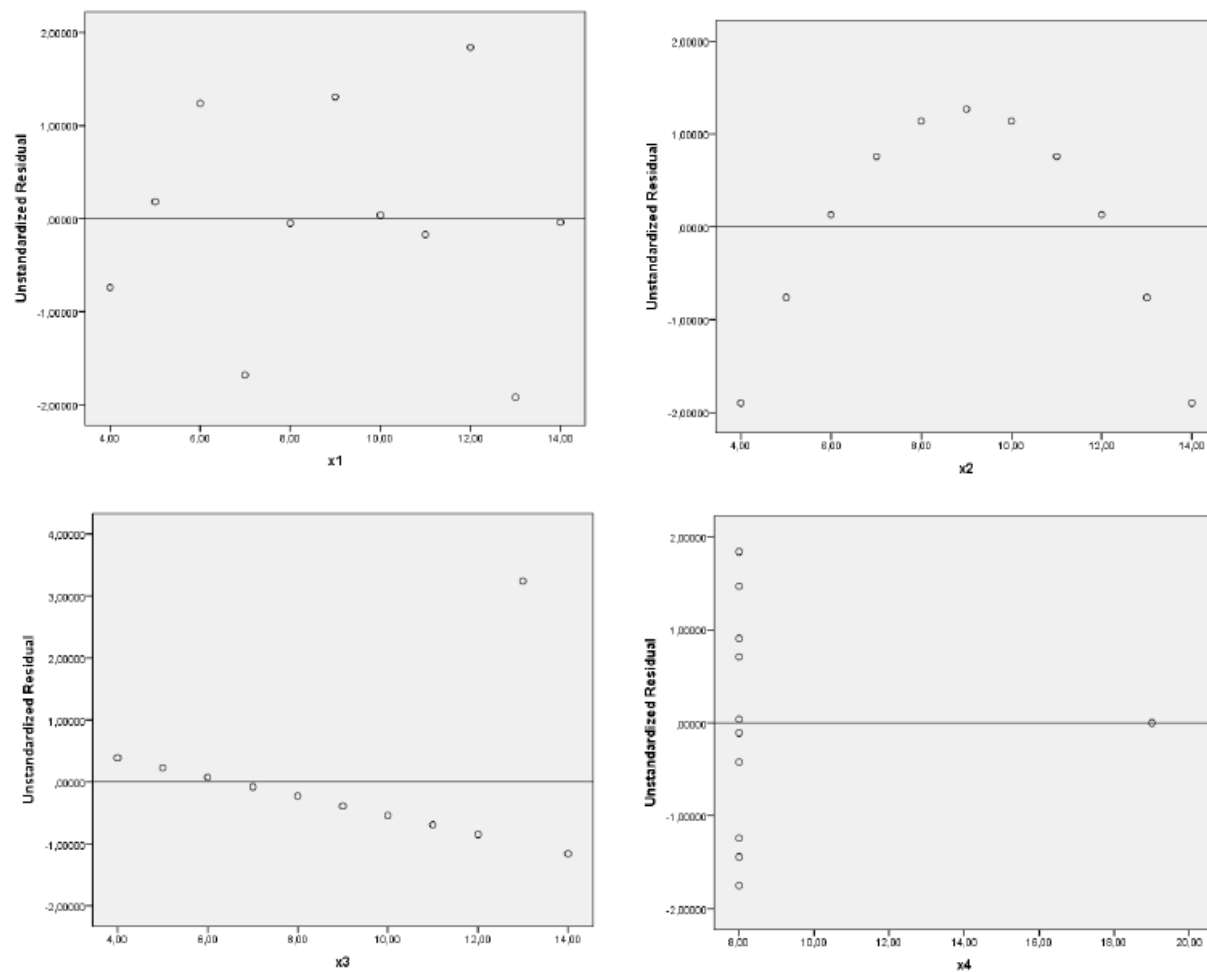
- Πώς αξιολογούμε τώρα το ταίριασμα του μοντέλου

$$Y_i = 3 + 0,5 X_i \quad (i = 1,2,3,4)$$

σε καθεμιά από τις 4 περιπτώσεις?

- Η εικόνα μπορεί να γίνει ακόμα περισσότερο ξεκάθαρη αν, αντί των προηγούμενων στικτών διαγραμμάτων των αρχικών ζευγών (x_i, y_i) , $(i = 1,2,3,4)$, σχηματίσουμε τα αντίστοιχα στικτά διαγράμματα για τα ζεύγη (x_i, e_i) , $(i = 1,2,3,4)$, όπου $e_i = y_i - \hat{y}_i$ τα αντίστοιχα κατάλοιπα.

Διαγράμματα καταλοίπων στα σετ του Anscombe



Υποθέσεις γραμμικής παλινδρόμησης

ΥΠΟΘΕΣΕΙΣ ΓΡΑΜΜΙΚΗΣ ΠΑΛΙΝΔΡΟΜΗΣΗΣ:

- Η σχέση μεταξύ της ανεξάρτητης μεταβλητής (X) και της εξαρτημένης (Y) είναι γραμμική.
 - Τα σφάλματα ε_i αποτελούν τυχαίες μεταβλητές με μέση τιμή 0.
 - Τα σφάλματα ε_i είναι **ομοσκεδαστικά** (δηλαδή, η διασπορά τους παραμένει σταθερή στα διάφορα επίπεδα της ανεξάρτητης μεταβλητής).
 - Τα σφάλματα ε_i είναι ασυσχέτιστα μεταξύ τους, δηλαδή
$$\text{Cov}(\varepsilon_i, \varepsilon_j) = 0, \quad i \neq j.$$
 - Τα σφάλματα ε_i ακολουθούν την κανονική κατανομή, δηλαδή ειδικότερα
$$\varepsilon_i \sim N(0, \sigma^2),$$
- Όπου σ^2 η σταθερή από την υπόθεση της ομοσκεδαστικότητας διακύμανση των σφαλμάτων.
- Απουσία ακραίων σημείων (outliers) και επιδραστικών σημείων (influential points).

Επαλήθευση υποθέσεων γραμμικής παλινδρόμησης

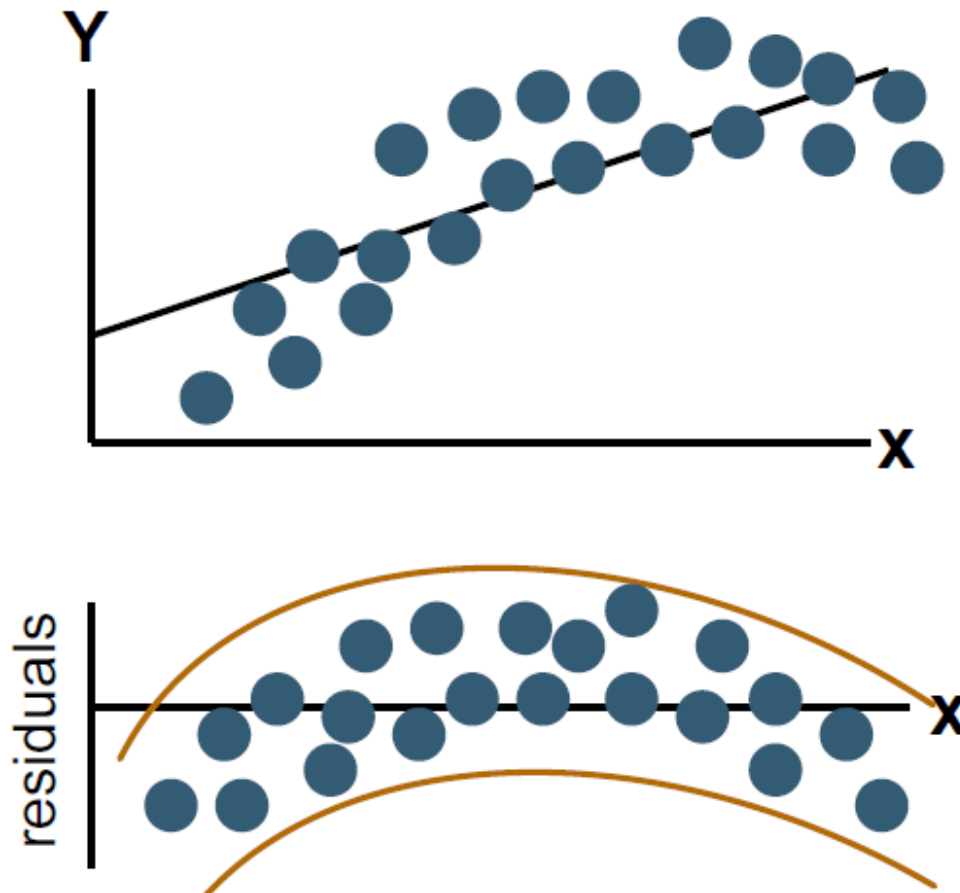
- Οι υποθέσεις της γραμμικής παλινδρόμησης αφορούν κατά κύριο λόγο στα σφάλματα ε_i .
- Τα σφάλματα ε_i αποτελούν θεωρητικές ποσότητες, οπότε ο έλεγχος των σχετικών υποθέσεων επαληθεύεται μέσα από τη διερεύνηση των καταλοίπων

$$e_i = y_i - \hat{y}_i,$$

τα οποία αποτελούν υπολογίσιμες ποσότητες (ως διαφορές των παρατηρήσεων της εξαρτημένης μεταβλητής Y από τις αντίστοιχες ατομικές προβλέψεις για αυτές μέσα από το γραμμικό μοντέλο), όπως επιχειρήθηκε στο προηγούμενο παράδειγμα του Anscombe.

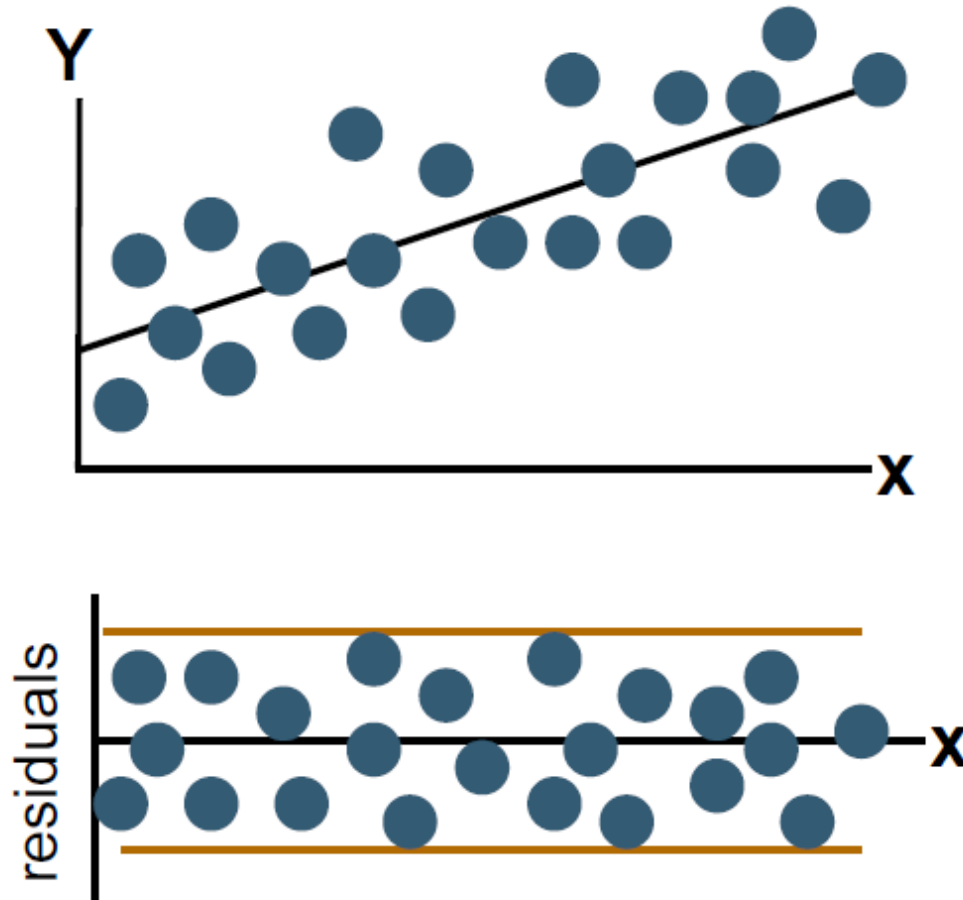
- Ακόμα και οι υποθέσεις που εκ πρώτης όψεως δεν αφορούν στα υπόλοιπα (όπως η υπόθεση της γραμμικότητας της σχέσης των X , Y και εκείνη που αφορά στα ακραία σημεία), μπορούν να επαληθευτούν μέσα από τη μελέτη των καταλοίπων.

Ανάλυση καταλοίπων για γραμμικότητα (απουσία γραμμικής σχέσης)



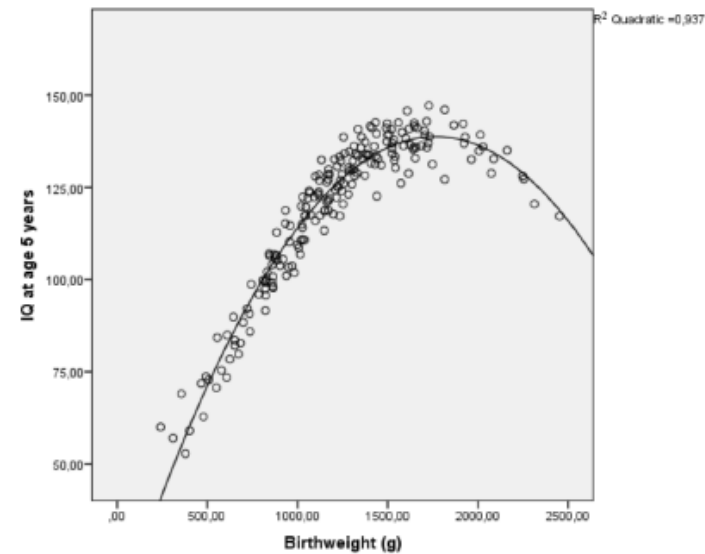
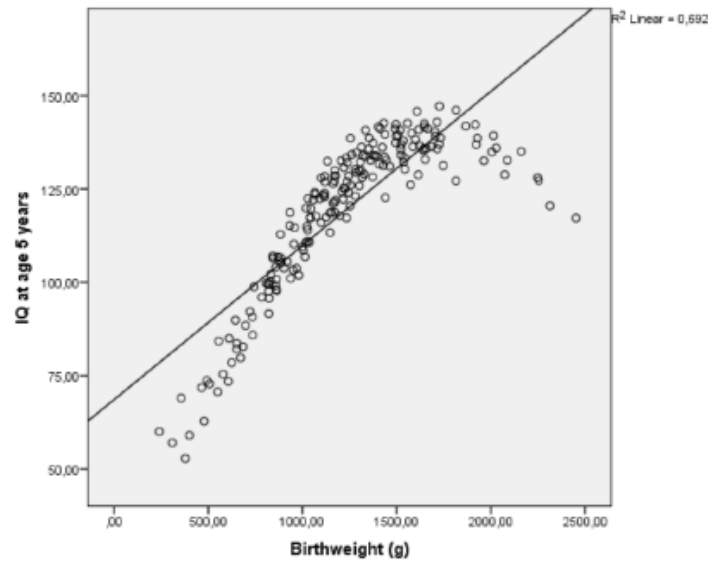
Γραμμική παλινδρόμηση

Ανάλυση καταλοίπων για γραμμικότητα (γραμμική σχέση)



Διερεύνηση ύπαρξης μη γραμμικής σχέσης

- Διερεύνηση του στικτού γραφήματος και εντοπισμός της τάσης που ταιριάζει καλύτερα στα δεδομένα.

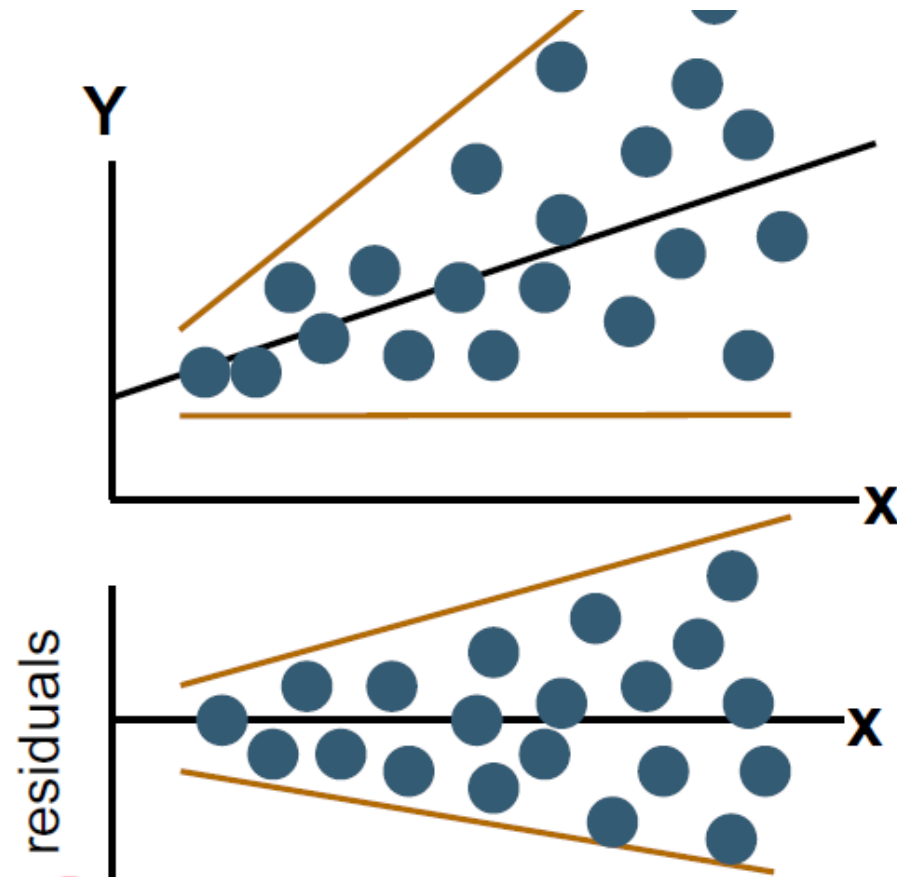


Διερεύνηση ύπαρξης μη γραμμικής σχέσης

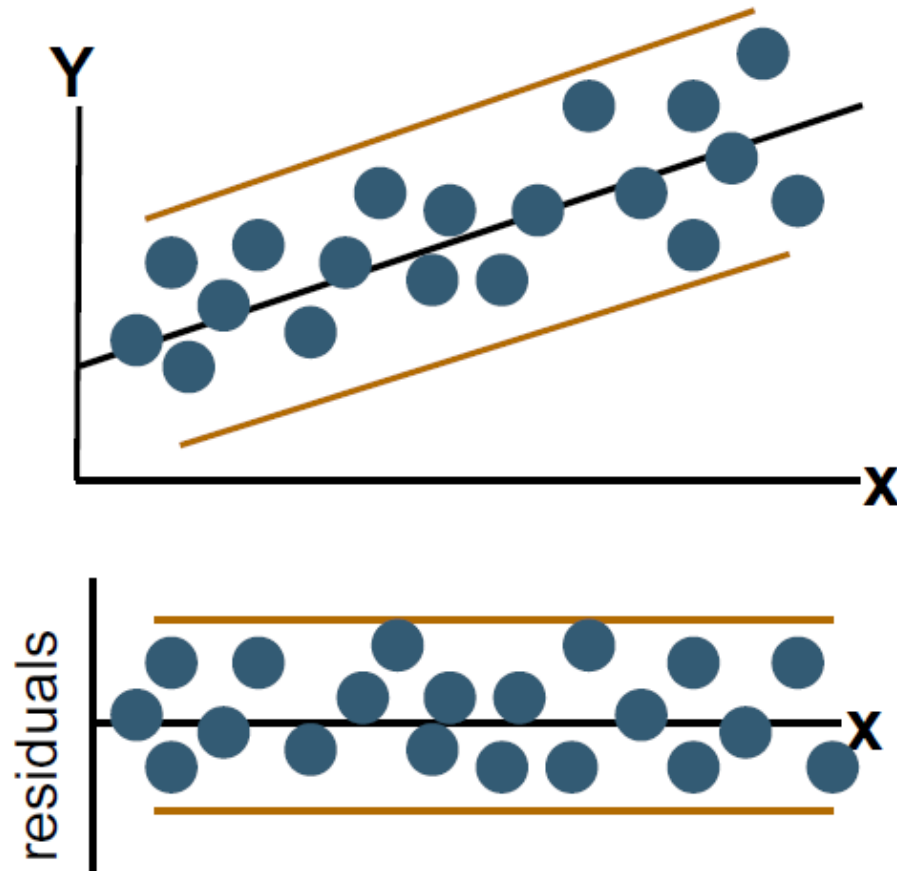
- **Ιεραρχική Πολλαπλή Παλινδρόμηση** (προσθέτοντας πολυωνυμικούς όρους ανώτερης τάξης) → απαιτείται ο υπολογισμός των όρων αυτών.
- Καταλήγουμε σε μοντέλο της μορφής:

$$IQ_i = b_0 + b_1 \cdot Weight_i + b_2 \cdot (Weight_i)^2 + \dots$$

Ανάλυση καταλοίπων για ετεροσκεδαστικότητα (μη σταθερή διακύμανση)

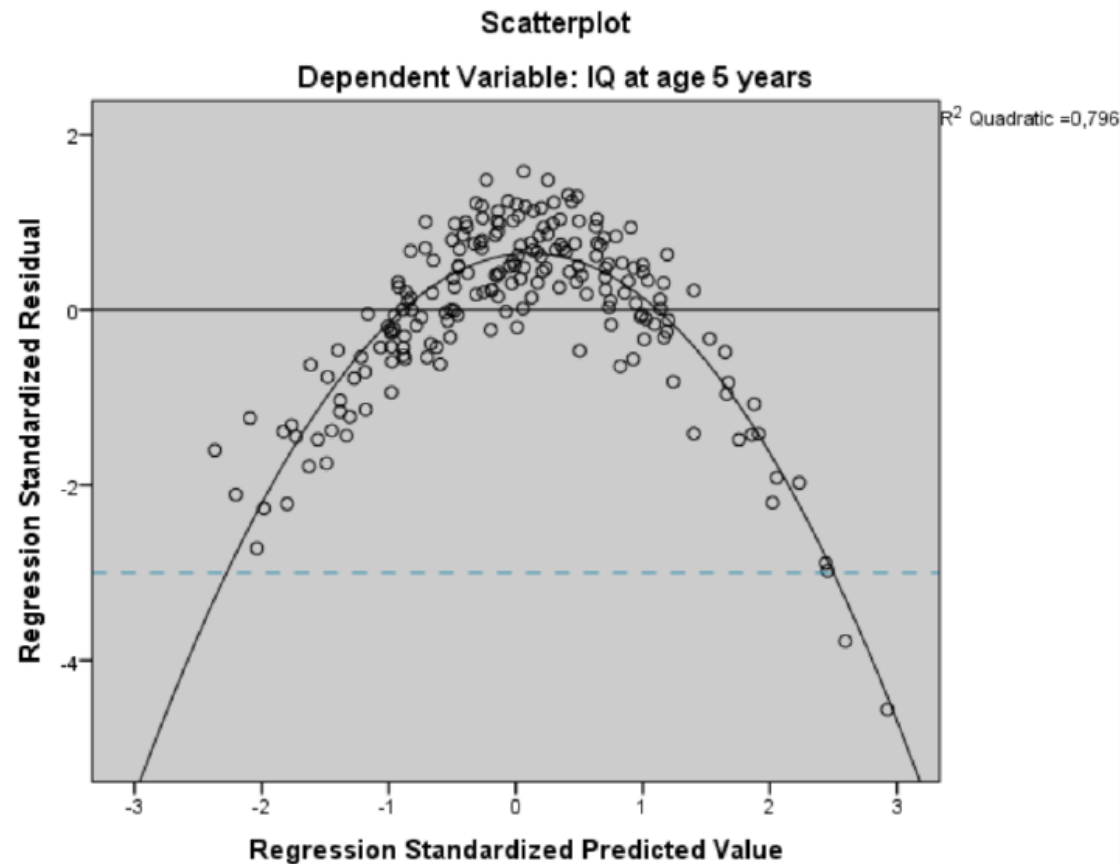


Ανάλυση καταλοίπων για ετεροσκεδαστικότητα (σταθερή διακύμανση)

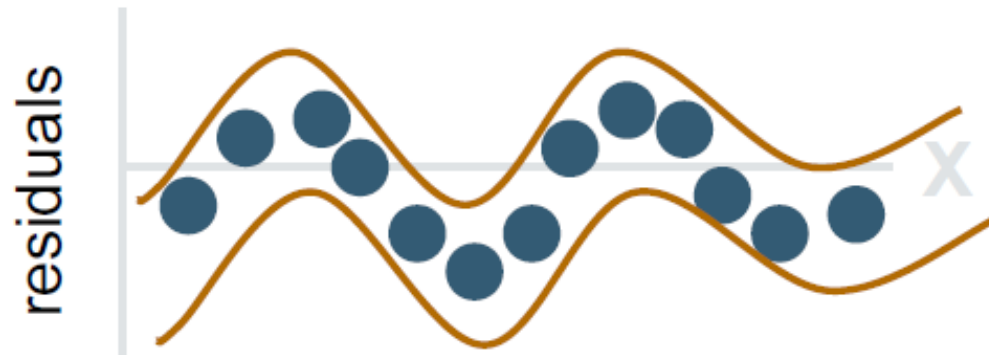
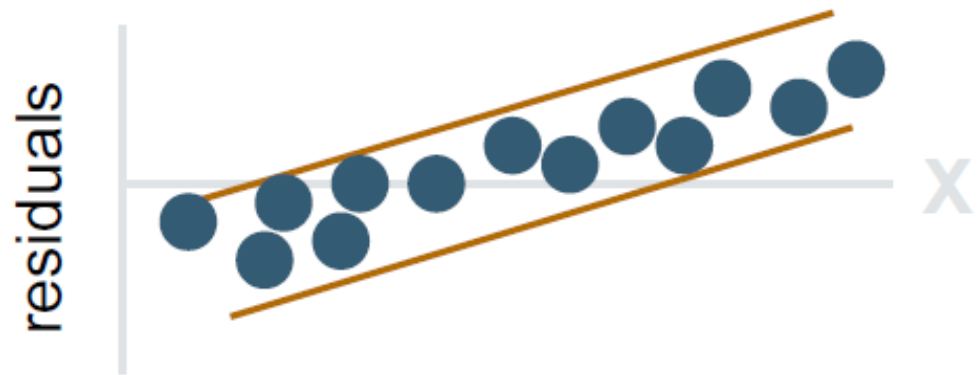


Ανάλυση καταλοίπων για ετεροσκεδαστικότητα

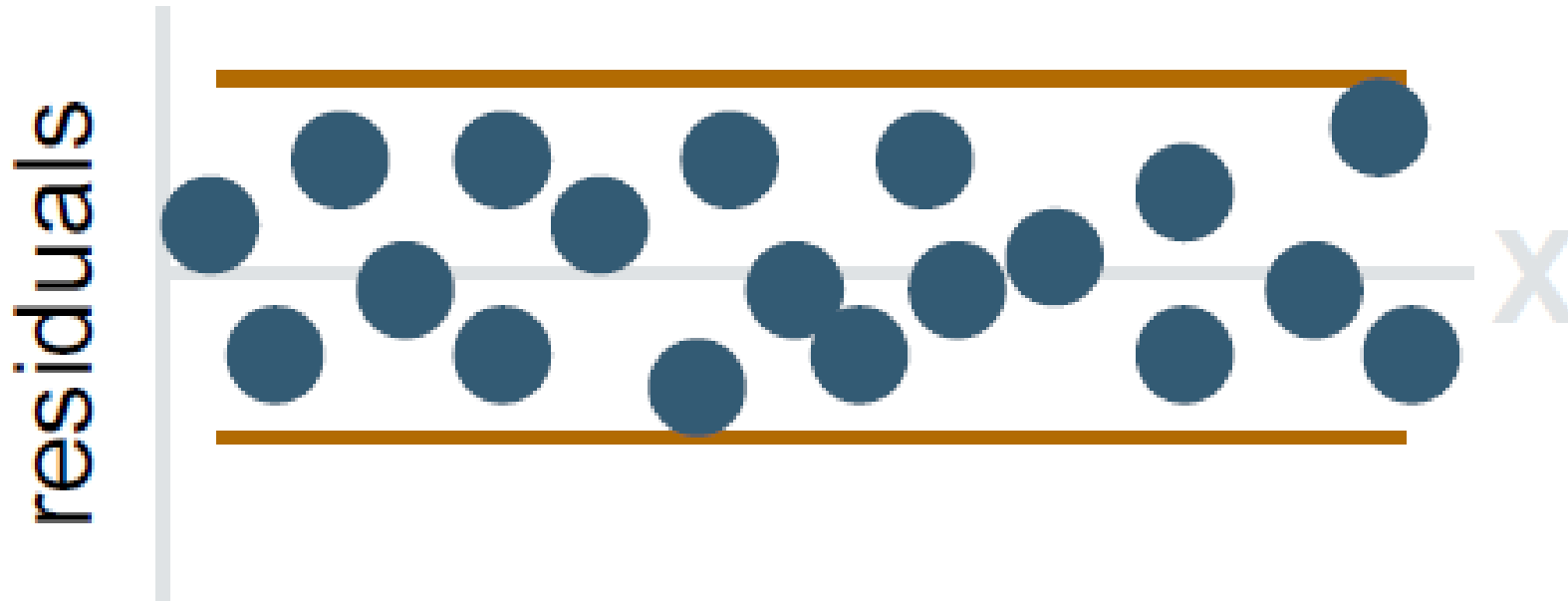
- Οπτικός έλεγχος ομοσκεδαστικότητας



Ανάλυση καταλοίπων για ανεξαρτησία παρατηρήσεων (μη ανεξάρτητες)

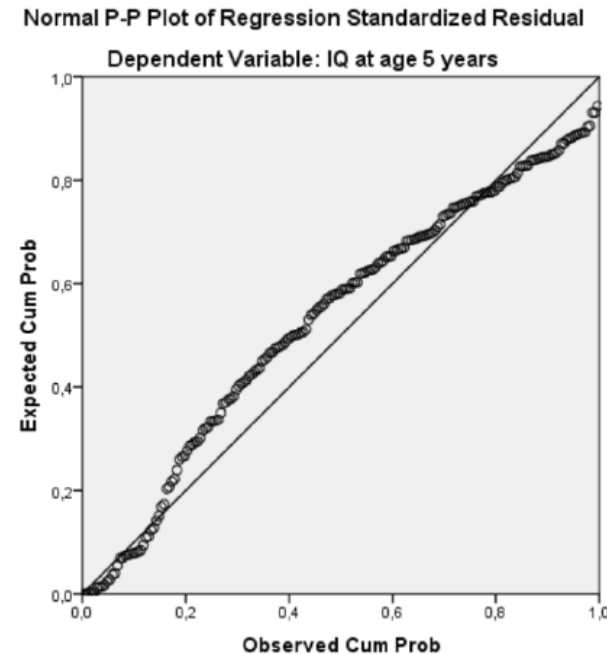
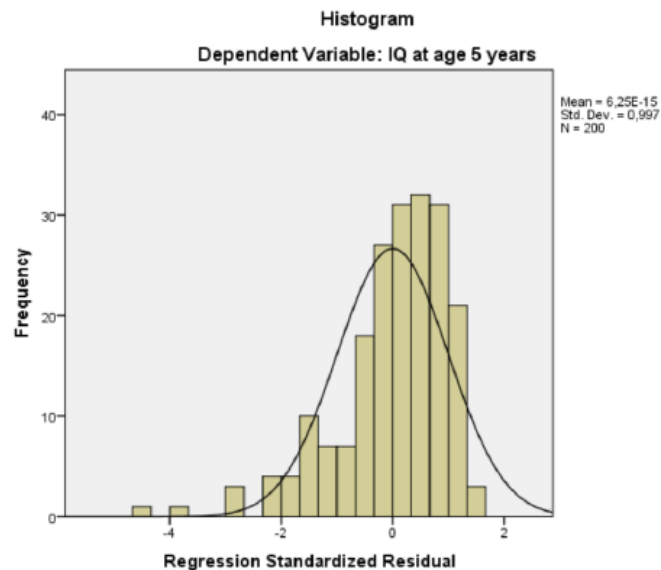


Ανάλυση καταλοίπων για ανεξαρτησία παρατηρήσεων (ανεξάρτητες παρατηρήσεις)



Οπτικοί και στατιστικοί έλεγχοι κανονικότητας καταλοίπων

- Διερεύνηση κανονικότητας καταλοίπων (Ιστόγραμμα, QQ-plot, PP-plot)



- + Έλεγχοι κανονικότητας για τα κατάλοιπα (K-S test, Shapiro Wilk, κλπ..)

Τι κάνουμε σε περιπτώσεις που οι υποθέσεις της κανονικότητας δεν ικανοποιούνται

- Έλεγχος για λάθη στην κωδικοποίηση ή την καταχώριση των δεδομένων που έχουν οδηγήσει σε ακραίες ή επιδραστικές παρατηρήσεις.
- Διερεύνηση άλλων, τυχόν μη γραμμικών σχέσεων μεταξύ ανεξάρτητης και εξαρτημένης μεταβλητής.
- Ανάλυση ευαισθησίας που εξετάζει κατά πόσον τα συμπεράσματα της ανάλυσης παραμένουν παρόμοια, όταν παραλείπονται όσες παρατηρήσεις έχουν ταυτοποιηθεί ως επιδραστικές.
- Χρήση μετασχηματισμών. Στη βιβλιογραφία υπάρχουν σχετικοί πίνακες μετασχηματισμών που οδηγούν σε γραμμικά μοντέλα (Rawlings (1988, σελ. 306-308), Chatterjee and Price (1977, σελ. 29)). Φυσικά όλες οι μη γραμμικές σχέσεις δεν είναι δυνατό να μετατραπούν σε γραμμικές.

5.10 Πολλαπλή παλινδρόμηση

Το μοντέλο της πολλαπλής παλινδρόμησης

- **Ιδέα:** Διερεύνηση της σχέσης μεταξύ μίας εξαρτημένης (Y) και 2 ή περισσότερων ανεξάρτητων μεταβλητών (X_i).

Μοντέλο πολλαπλής παλινδρόμησης με k ανεξάρτητες μεταβλητές:

The diagram shows the multiple regression equation $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + \varepsilon$. Above the equation, four blue boxes with white text are connected to the equation by blue arrows. The first box, 'Εξαρτημένη μεταβλητή', has an arrow pointing to Y . The second box, 'Σταθερά', has an arrow pointing to β_0 . The third box, 'Παράμετροι κλίσεων', has two arrows pointing to β_1 and β_2 . The fourth box, 'Τυχαίο σφάλμα', has an arrow pointing to ε .

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + \varepsilon$$

Εξίσωση πολλαπλής παλινδρόμησης

- Οι παράμετροι του μοντέλου πολλαπλής παλινδρόμησης εκτιμώνται με χρήση των δειγματικών δεδομένων.

Εκτίμηση εξίσωσης πολλαπλής παλινδρόμησης με k ανεξάρτητες μεταβλητές:

$$\hat{y}_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \dots + b_k x_{ki}$$

- Οι υπολογισμοί των εκτιμήσεων $b_0, b_1, \dots, b_k$ είναι περίπλοκοι και γίνονται πάντοτε με χρήση σχετικού λογισμικού, όπως το SPSS ή το MS Excel.

Ερμηνεία παραμέτρων

- Η σταθερά β_0 ερμηνεύεται ως η αναμενόμενη μέση τιμή της εξαρτημένης μεταβλητής Y αν όλες οι ανεξάρτητες μεταβλητές είναι 0 (δεν έχουν δηλαδή κάποια επίδραση).
- Ο συντελεστής παλινδρόμησης β_j δίνει την μεταβολή της εξαρτημένης μεταβλητής Y για μοναδιαία μεταβολή της ανεξάρτητης μεταβλητής X_j αν όλες οι υπόλοιπες ανεξάρτητες μεταβλητές παραμείνουν σταθερές.

Χρήση ποιοτικών μεταβλητών ως ερμηνευτικών

Σημείωση: Στην περίπτωση όπου υπάρχουν και ποιοτικές μεταβλητές, απαιτείται η χρήση ψευδομεταβλητών (dummy variables) στη θέση κάθε μιας ποιοτικής μεταβλητής. Γενικός κανόνας: Για ποιοτική μεταβλητή k κατηγοριών, χρειάζονται $k-1$ ψευδομεταβλητές.

Έστω ότι έχουμε μια κατηγορική μεταβλητή X με δύο επίπεδα, π.χ. το φύλο (Άνδρας-Γυναίκα). Αποφασίζουμε μια κατηγορία θεωρείται ως κατηγορία αναφοράς (ως προς την οποία θα γίνεται η σύγκριση των αποτελεσμάτων) και κωδικοποιείται με 0 ενώ η άλλη κατηγορία με 1. Υποθέτοντας ως κατηγορία αναφοράς τις γυναίκες, ο εκτιμώμενος συντελεστής παλινδρόμησης εκφράζει την αναμενόμενη μεταβολή της εξαρτημένης μεταβλητής για τους άνδρες ($X=1$) σε σχέση με τις γυναίκες ($X=0$).

Χρήση ποιοτικών μεταβλητών ως ερμηνευτικών

Αντίστοιχα συμπεράσματα θα είχαμε αν υπήρχε μια μεταβλητή με παραπάνω από δύο επίπεδα, π.χ. το επάγγελμα, με κατηγορίες: εργάτης/δημόσιος υπάλληλος/ιδιωτικός υπάλληλος/αγρότης. Εισάγονται $k-1 = 3$ ψευδομεταβλητές στο μοντέλο, στη θέση της αρχικής μεταβλητής που αφορά το επάγγελμα και για κατηγορία αναφοράς τους εργάτες, έχουμε:

$X_1 = 1$, αν είναι δημόσιος υπάλληλος, 0 για όλες τις άλλες κατηγορίες.

$X_2 = 1$, αν είναι ιδιωτικός υπάλληλος, 0 για όλες τις άλλες κατηγορίες.

$X_3 = 1$, αν είναι αγρότης, 0 για όλες τις άλλες κατηγορίες.

Οπότε και θα προέκυπταν τρεις συντελεστές παλινδρόμησης, η ερμηνεία των οποίων θα είναι ως προς την κατηγορία αναφοράς (τους εργάτες). Π.χ. Ο β_2 θα εκφράζει την αναμενόμενη μεταβολή της εξαρτημένης μεταβλητής για τους ιδιωτικούς υπαλλήλους σε σχέση με τους εργάτες.

Υποθέσεις πολλαπλής παλινδρόμησης

- Οι τιμές x_i και οι όροι σφάλματος ε_i είναι ανεξάρτητες.
- Οι όροι σφάλματος αποτελούν τυχαίες μεταβλητές με μέσο 0 με σταθερή διακύμανση σ^2 :

$$E(\varepsilon_i)=0 \text{ και } \text{Var}(\varepsilon_i) = \sigma^2, \text{ για } i=1, \dots, n.$$

Όπως και στην περίπτωση της απλής γραμμικής παλινδρόμησης, η υπόθεση της σταθερής διακύμανσης καλείται **ομοσκεδαστικότητα**.

- Τα τυχαία σφάλματα είναι **ασυσχέτιστα** μεταξύ τους, δηλαδή

$$\text{Cov}(\varepsilon_i, \varepsilon_j)=0, \text{ για } i \neq j.$$

- Δεν υπάρχουν συντελεστές $\lambda_0, \lambda_1, \dots, \lambda_k$, τέτοιοι ώστε

$$\lambda_0 + \lambda_1 x_{1i} + \lambda_2 x_{2i} + \dots + \lambda_k x_{ki}=0.$$

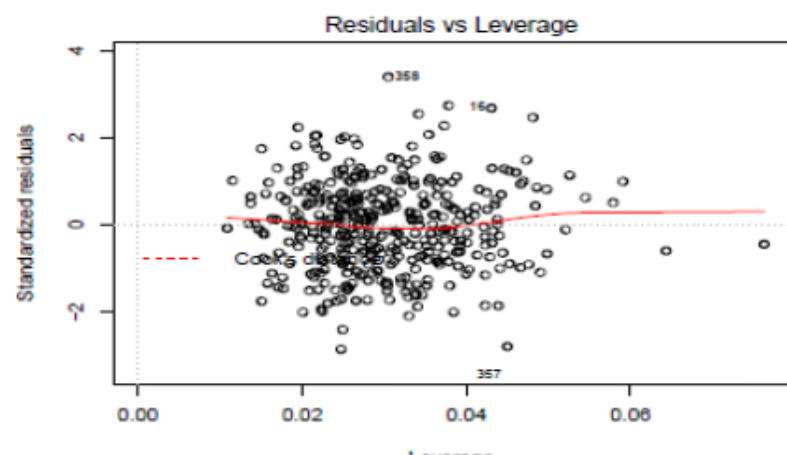
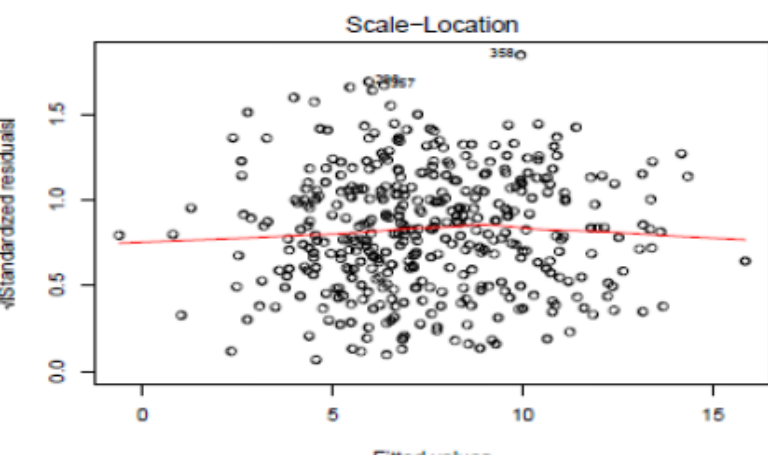
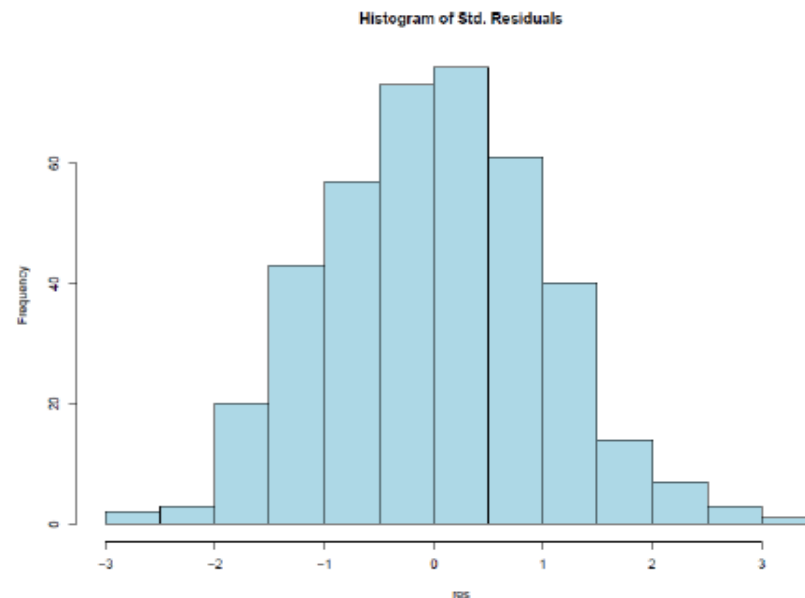
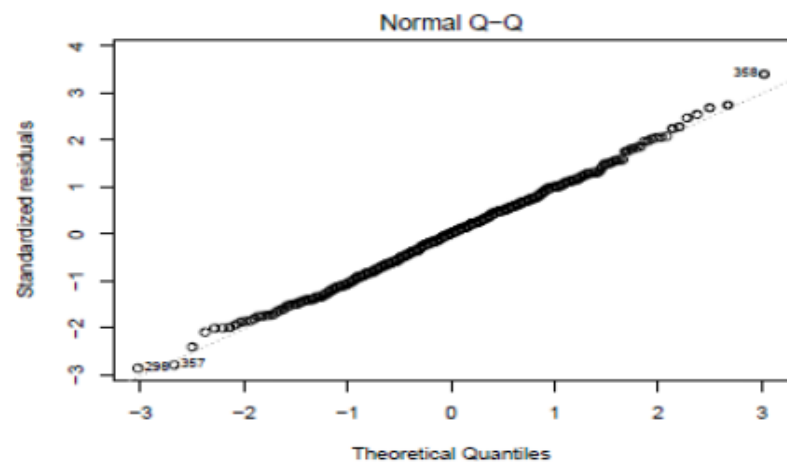
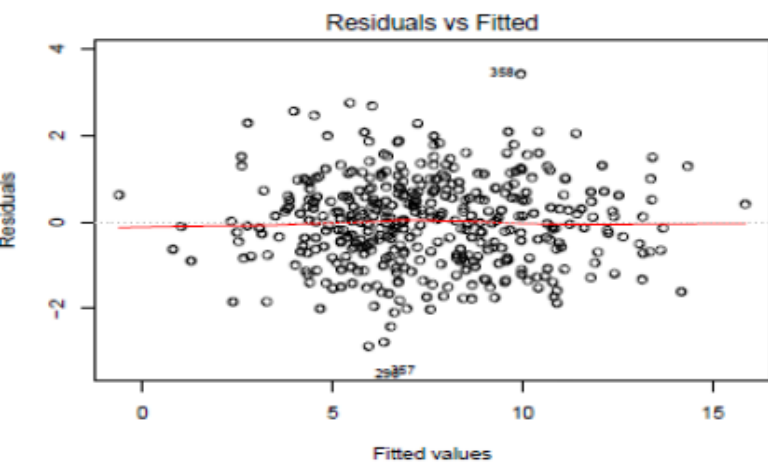
- Με άλλα λόγια, δεν υφίσταται γραμμική σχέση μεταξύ των ερμηνευτικών μεταβλητών $X_1, X_2, \dots, X_k$. Η υπόθεση αυτή εξασφαλίζει ότι η εκτίμηση των παραμέτρων είναι ευσταθής.

Εξέταση ισχύος υποθέσεων πολλαπλής παλινδρόμησης

Ο έλεγχος των υποθέσεων του μοντέλου γίνεται με χρήση κατάλληλων γραφικών και στατιστικών ελέγχων:

- Οι **συσχετίσεις** μεταξύ των επεξηγηματικών μεταβλητών ελέγχονται κυρίως διαγραμματικά ή μέσω του πίνακα συσχετίσεων (correlation matrix).
- Η υπόθεση της **Κανονικότητας** των Y_i , ελέγχεται μέσω του γραφήματος Q-Q plot καθώς και μέσω κάποιου στατιστικού ελέγχου για την κανονικότητα (π.χ. K-S, Shapiro, κτλ.).
- Η υπόθεση της **ομοσκεδαστικότητας** των σφαλμάτων (ότι δηλαδή η διακύμανση των σφαλμάτων παραμένει σταθερή και δεν μεταβάλλεται) ελέγχεται κυρίως μέσω της διαγραμματικής τους απεικόνισης λαμβάνοντας υπόψη και το συνολικό μέγεθος του δείγματος.
- Η υπόθεση της **ανεξαρτησίας** των σφαλμάτων ελέγχεται μέσω του στατιστικού ελέγχου σειριακής συσχέτισης των Durbin-Watson

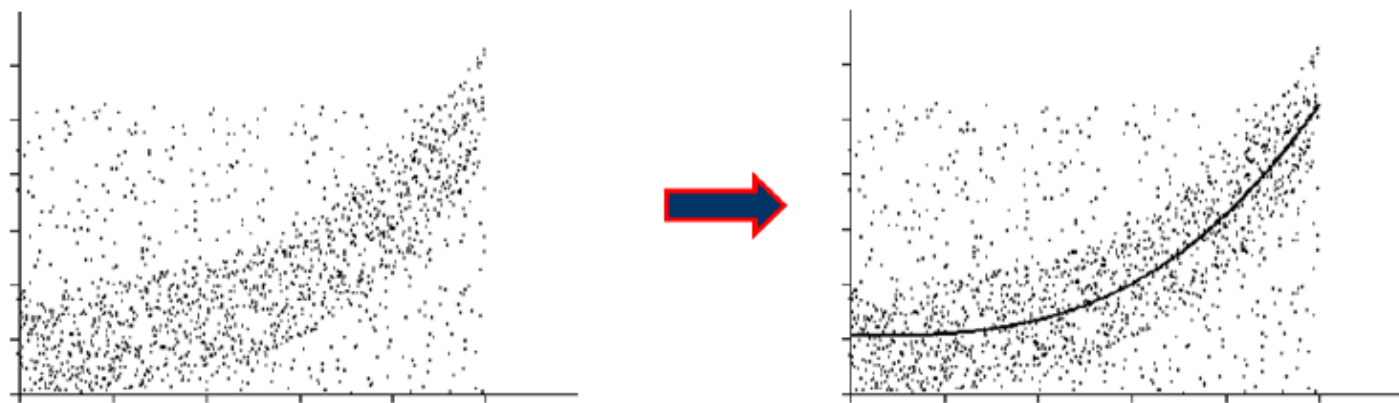
Ενδεικτικά παραδείγματα γραφικών ελέγχων



Μη γραμμικά μοντέλα

➤ Εκτός των γραμμικών μοντέλων που συνδέουν την εξαρτημένη μεταβλητή Y με την X , βάσει μιας γραμμικής σχέσης, υπάρχουν τα μη γραμμικά μοντέλα παλινδρόμησης. Αυτά μπορεί να είναι:

- Πολυωνυμικά μοντέλα
- Εκθετικά μοντέλα
- Λογαριθμικά μοντέλα
- κ.α.



Μη γραμμικά μοντέλα

➤ Η γενική μορφή ενός μοντέλου παλινδρόμησης, μπορεί να γραφεί ως

$$Y_i = f(X_i) + \epsilon_i$$

➤ Οπότε για διαφορετικές μορφές της f μπορούμε να λάβουμε όποιας μορφής μοντέλο επιθυμούμε. Για παράδειγμα:

- $f(x) = \beta_0 + \beta_1 x \rightarrow$ Μοντέλο γραμμικής παλινδρόμησης
- $f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 \rightarrow$ Τετραγωνικό μοντέλο (πολυωνυμικό 2ας τάξης)

➤ ϵ_i είναι η απόκλιση της i -οστής παρατήρησης από την καμπύλη παλινδρόμησης. Συνήθως χρησιμοποιείται και εδώ το κριτήριο ελαχίστων τετραγώνων κατά την εκτίμηση των παραμέτρων του μοντέλου, οπότε ακολουθείται η κλασική υπόθεση για τα σφάλματα $\epsilon_i \sim N(0, \sigma^2)$.

$$\epsilon_i = Y_i - \hat{Y}_i = Y_i - f(X_i)$$

Μη γραμμικά μοντέλα

➤ Σημείωση: Ο όρος «μη γραμμική» παλινδρόμηση, δεν αφορά απαραίτητως τις παραμέτρους του μοντέλου, αλλά τη σχέση της εξαρτημένης μεταβλητής Y με τις αντίστοιχες ανεξάρτητες μεταβλητές.

- ▶ $Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \epsilon_i$ το τετραγωνικό μοντέλο παλινδρόμησης είναι γραμμικό ως προς τις παραμέτρους $\beta_0, \beta_1, \beta_2$.
- ▶ $Y_i = \beta_0 + \beta_1 \log(X_i) + \epsilon_i$ το λογαριθμικό μοντέλο παλινδρόμησης είναι γραμμικό ως προς τις παραμέτρους β_0, β_1 .
- ▶ Αν $\varphi(\cdot), \psi(\cdot)$ είναι γνωστές συναρτήσεις, τότε το μοντέλο

$$\varphi(Y_i) = \beta_0 + \beta_1 \psi(X_i) + \epsilon_i$$

είναι επίσης γραμμικό ως προς τις παραμέτρους β_0, β_1 .

Μη γραμμικά μοντέλα

➤ Παραδείγματα μη γραμμικών μοντέλων:

- ▶ Πολυωνυμικό μοντέλο. Για $p > 1$ και ακέραιο:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \dots + \beta_p X_i^p + \epsilon_i$$

- ▶ Εκθετικό μοντέλο.

$$Y_i = \beta_0 + \beta_1 e^{\beta_2 + \beta_3 X_i} + \epsilon_i$$

- ▶ Λογαριθμικό μοντέλο.

$$Y_i = \beta_0 + \beta_1 \log(\beta_2 + \beta_3 X_i) + \epsilon_i$$

- ▶ Δυναμικό μοντέλο.

$$Y_i = \beta_0 + \beta_1 X_i^{\beta_2} + \epsilon_i$$

- ▶ Τριγωνομετρικό μοντέλο.

$$Y_i = \beta_0 + \beta_1 \sin(\beta_2 X_i) + \beta_3 \cos(\beta_4 X_i) + \epsilon_i$$

Δεν πρέπει να ξεχνάμε ότι ένα πολύπλοκο μοντέλο δεν είναι απαραίτητα και το βέλτιστο, καθώς και ότι πάντα υπάρχει το πρόβλημα της υπερπροσαρμογής, το οποίο σαφώς και πρέπει να αποφεύγεται.