Simple linear regression

Απλή γραμμική παλινδρόμηση

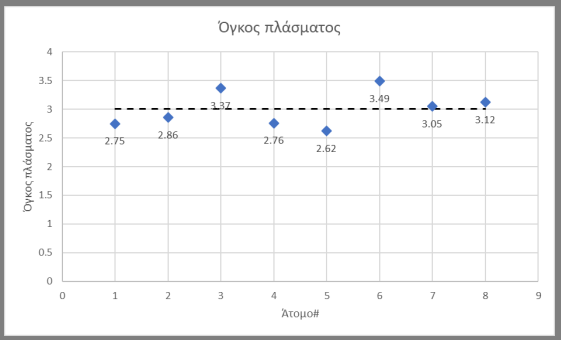
Ζιντζαράς Ηλίας, M.Sc., Ph.D.

Καθηγητής Βιομαθηματικών-Βιομετρίας
Εργαστήριο Βιομαθηματικών
Τμήμα Ιατρικής
Πανεπιστήμιο Θεσσαλίας

Institute for Clinical Research and Health Policy Studies
Tufts University School of Medicine
Boston, MA, USA

Θεόδωρος Μπρότσης, MSc, PhD
Εντεταλμένος Διδάσκων
(http://biomath.med.uth.gr)
Πανεπιστήμιο Θεσσαλίας
Email: tmprotsis@uth.gr

Τα βασικά


Όγκος πλάσματος
Όγκος πλάσματος
Άτομο#
0
0.5
1
1.5
2
2.5
3
3.5
4
0
1
2
3
4
5
6
7
8
9
2.75
2.86
3.37
2.76
2.62
3.49
3.05
3.12

Παράδειγμα

Έστω ότι θέλουμε να δημιουργήσουμε ένα μοντέλο το οποίο θα μας επιτρέπει να προβλέπουμε την τιμή του όγκου του πλάσματος δοθέντος του βάρους του σώματος ενός υγιή άντρα

Παράδειγμα

- Τα δεδομένα που συλλέχτηκαν φαίνονται στον διπλανό πίνακα.
- Αυτό που διαπιστώθηκε εκ των υστέρων είναι πως συλλέχτηκαν μόνο τα δεδομένα για τον όγκο του πλάσματος

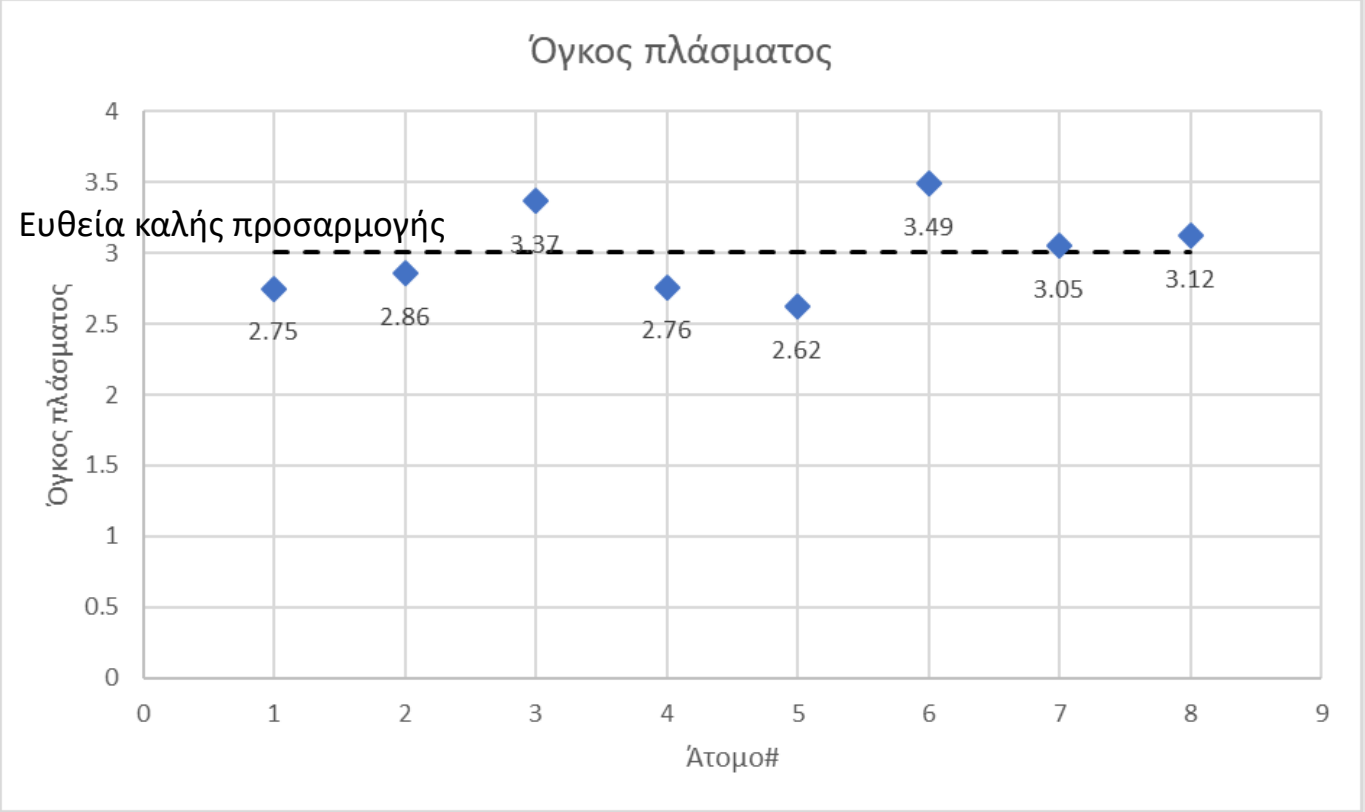
Πως μπορούμε να προβλέψουμε τον όγκο του πλάσματος για τον επόμενο υγιή άντρα, μόνο με βάση τα δεδομένα που συλλέχτηκαν;

Άτομο#	Όγκος πλάσματος σε lt (y)
1	2.75
2	2.86
3	3.37
4	2.76
5	2.62
6	3.49
7	3.05
8	3.12

Παράδειγμα

Άτομο#	Όγκος πλάσματος σε lt (y)
1	2.75
2	2.86
3	3.37
4	2.76
5	2.62
6	3.49
7	3.05
8	3.12

$$\bar{y} = 3.0025$$

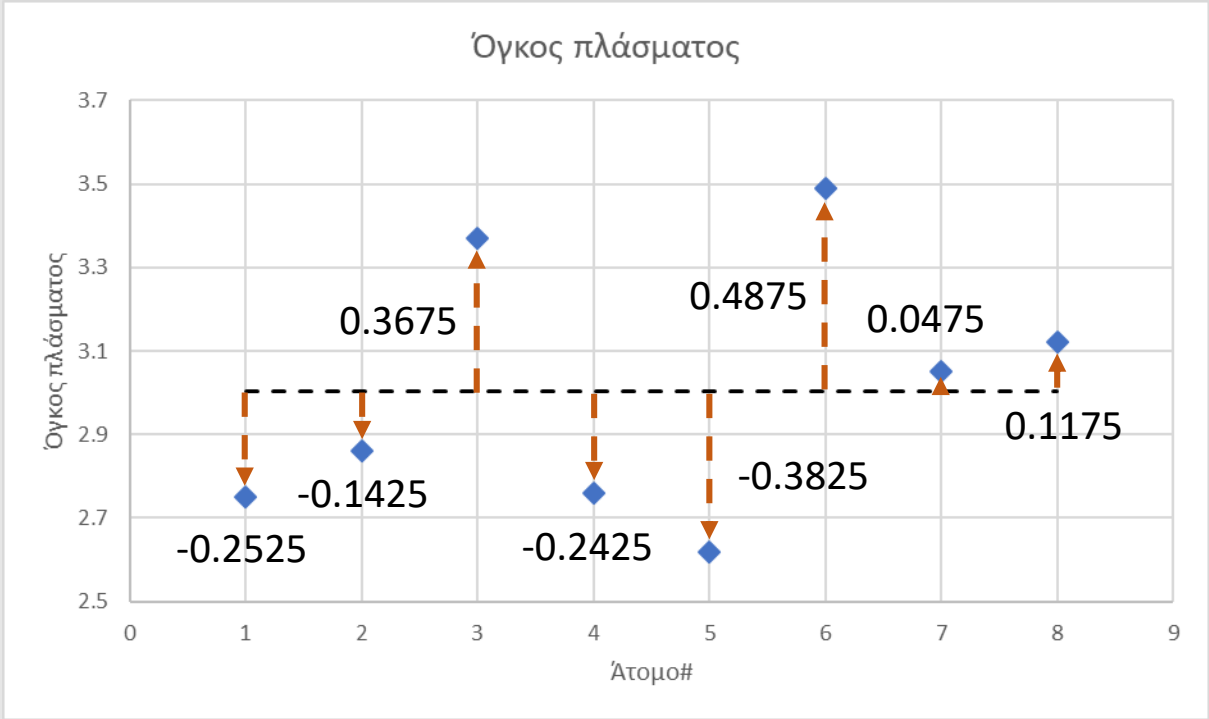


Με μόνο μία μεταβλητή, η μέση τιμή είναι ο καλύτερος προγνωστικός παράγοντας για την επόμενη τιμή. Η μεταβλητότητα των όγκων πλάσματος μπορεί να εξηγηθεί μόνο από τους ίδιους όγκους πλάσματος.

Αξιολόγηση της προσαρμογής (goodness of fit)

Άτομο#	Όγκος πλάσματος σε lt (y)
1	2.75
2	2.86
3	3.37
4	2.76
5	2.62
6	3.49
7	3.05
8	3.12

$$\bar{y} = 3.0025$$

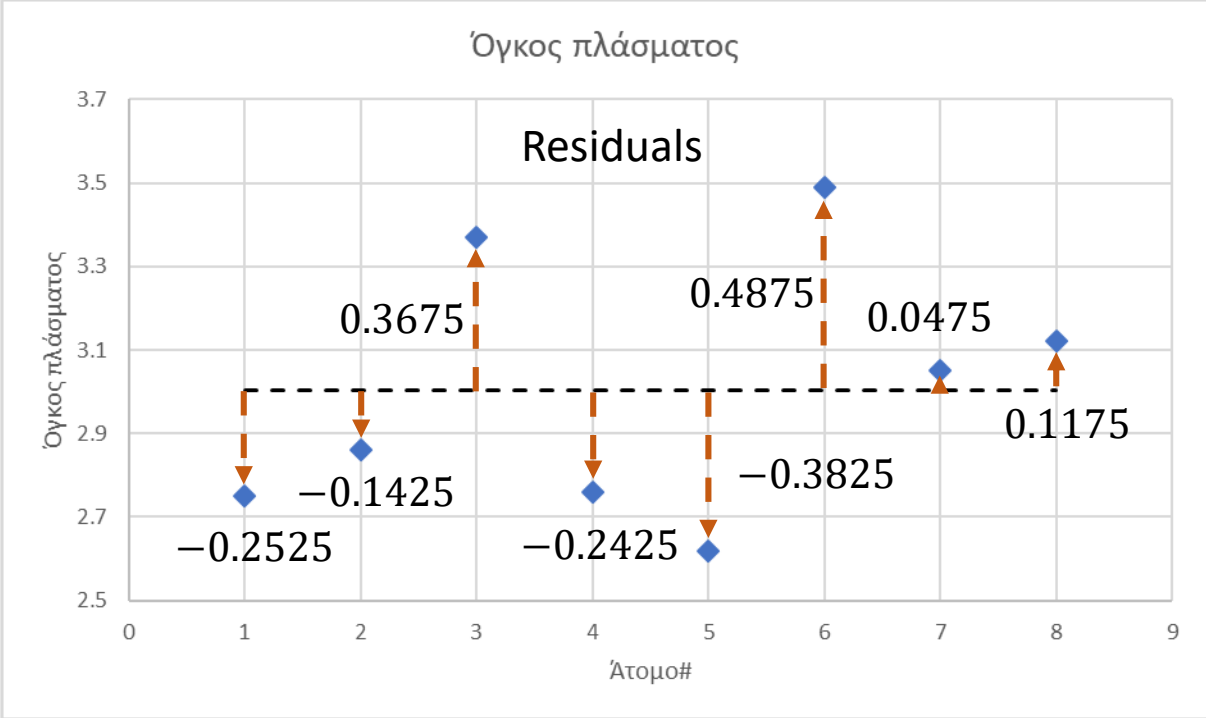


1. Οι παρατηρηθείσες τιμές δεν βρίσκονται πάνω στη ευθεία
2. Κάποιες βρίσκονται πάνω από τη γραμμή και κάποιες από κάτω
3. Αυτό μας δείχνει πόσο καλά προσαρμόζεται η ευθεία στις παρατηρηθείσες τιμές
4. Ένας τρόπος να το κάνουμε αυτό είναι να μετρήσουμε τις αποστάσεις των παρατηρήσεων από την ευθεία καλής προσαρμογής (τυπική απόκλιση)

Αξιολόγηση της προσαρμογής (goodness of fit)

Άτομο#	Όγκος πλάσματος σε lt (y)
1	2.75
2	2.86
3	3.37
4	2.76
5	2.62
6	3.49
7	3.05
8	3.12

$$\bar{y} = 3.0025$$

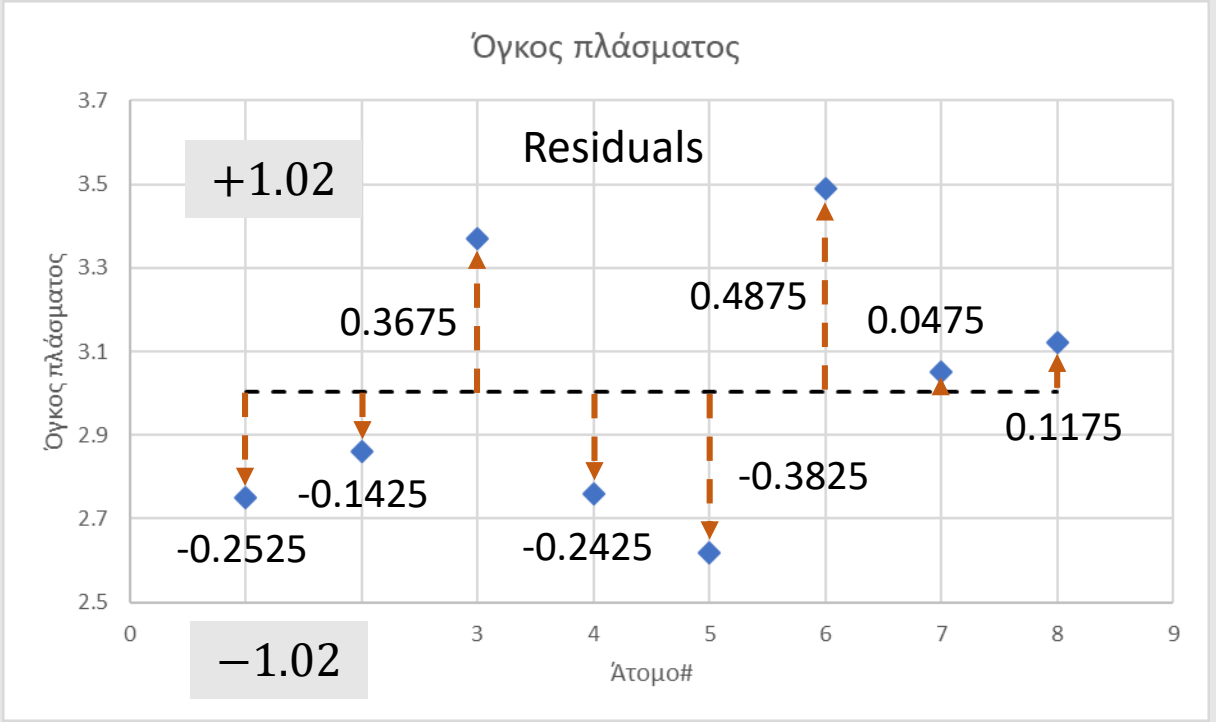


- Αυτές οι αποστάσεις μεταξύ της ευθείας και των παρατηρήσεων ονομάζονται υπόλοιπα (Residuals)
- Επίσης ονομάζονται και σφάλματα (error) καθώς μας δείχνουν πόσο μακριά βρίσκονται οι παρατηρήσεις από την ευθεία καλής προσαρμογής

Αξιολόγηση της προσαρμογής (goodness of fit)

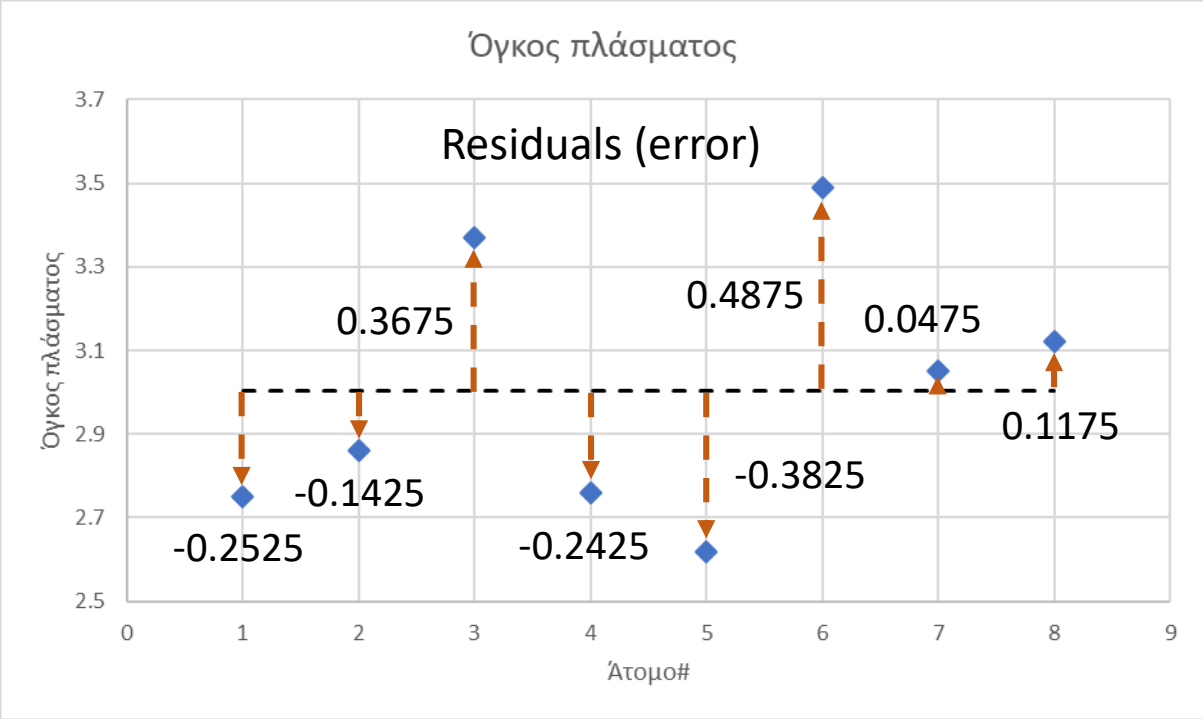
Άτομο#	Όγκος πλάσματος σε lt (y)
1	2.75
2	2.86
3	3.37
4	2.76
5	2.62
6	3.49
7	3.05
8	3.12

$$\bar{y} = 3.0025$$



- Αν προσθέσουμε τα υπόλοιπα που βρίσκονται πάνω από τη γραμμή έχουμε $+1.02$
- Αν προσθέσουμε τα υπόλοιπα που βρίσκονται κάτω από τη γραμμή έχουμε -1.02
- Το άθροισμα των υπολοίπων είναι πάντα μηδέν

Τετραγωνίζοντας τα υπόλοιπα (error)



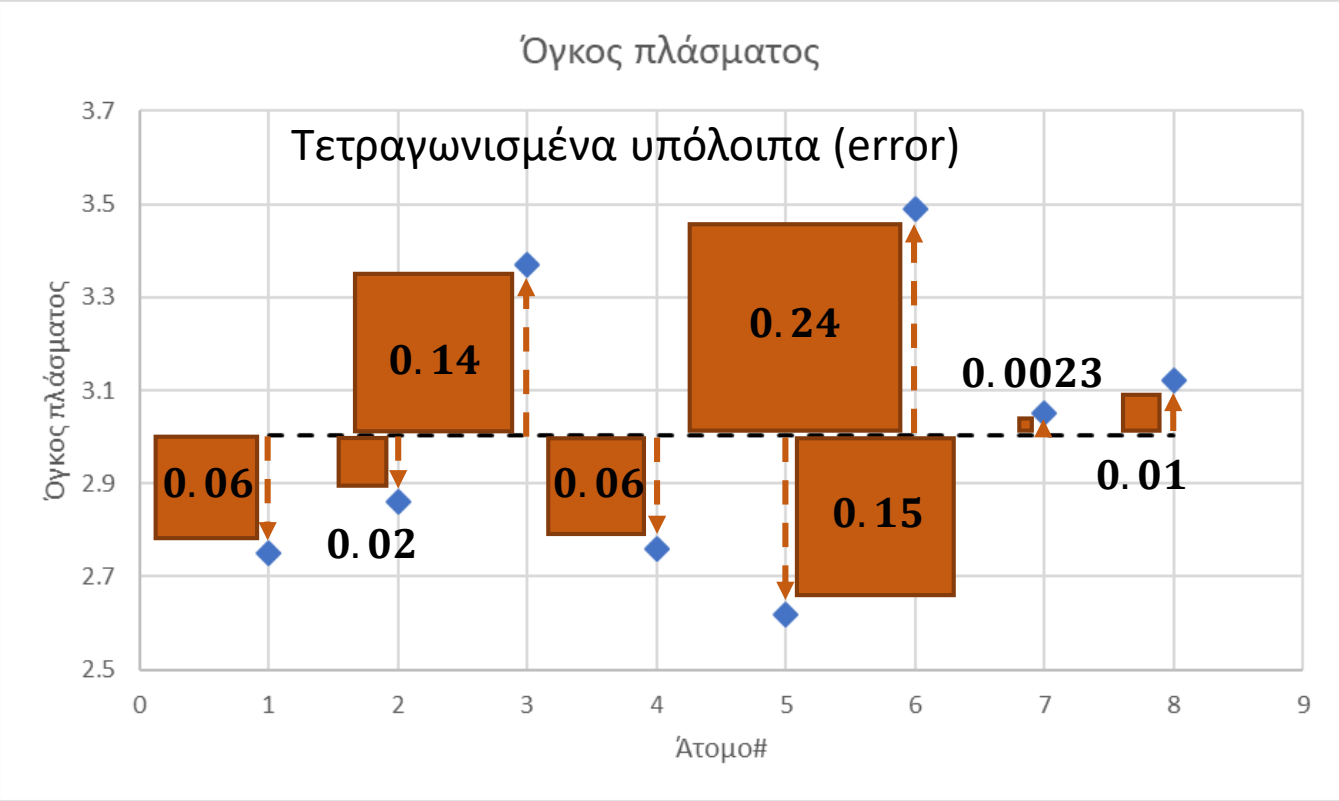
Άτομο#	Residual	Residual²
1	−0.2525	0.06
2	−0.1425	0.02
3	0.3675	0.14
4	−0.2425	0.06
5	−0.3825	0.15
6	0.4875	0.24
7	0.0475	0
8	0.1175	0.01

Γιατί τετραγωνίζουμε τα υπόλοιπα;

1. Για να τα κάνουμε θετικά
2. Για να δώσουμε έμφαση στις μεγαλύτερες αποκλίσεις

$$Sum\ of\ squared\ errors\ (SSE) = 0.68$$

Τετραγωνίζοντας τα υπόλοιπα (error)

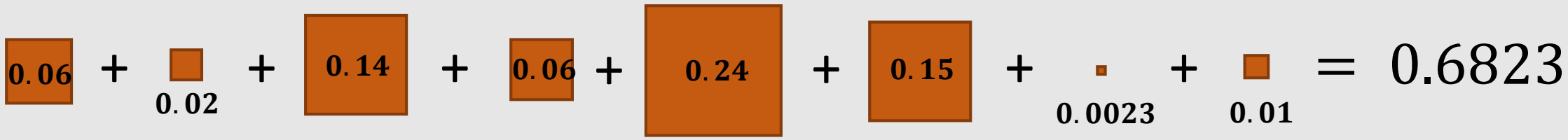


Άτομο#	Residual	Residual^2
1	−0.2525	0.06
2	−0.1425	0.02
3	0.3675	0.14
4	−0.2425	0.06
5	−0.3825	0.15
6	0.4875	0.24
7	0.0475	0.0023
8	0.1175	0.01

Όταν τετραγωνίζουμε τα υπόλοιπα το εννοούμε κυριολεκτικά

$$Sum\ of\ squared\ errors\ (SSE) = 0.68$$

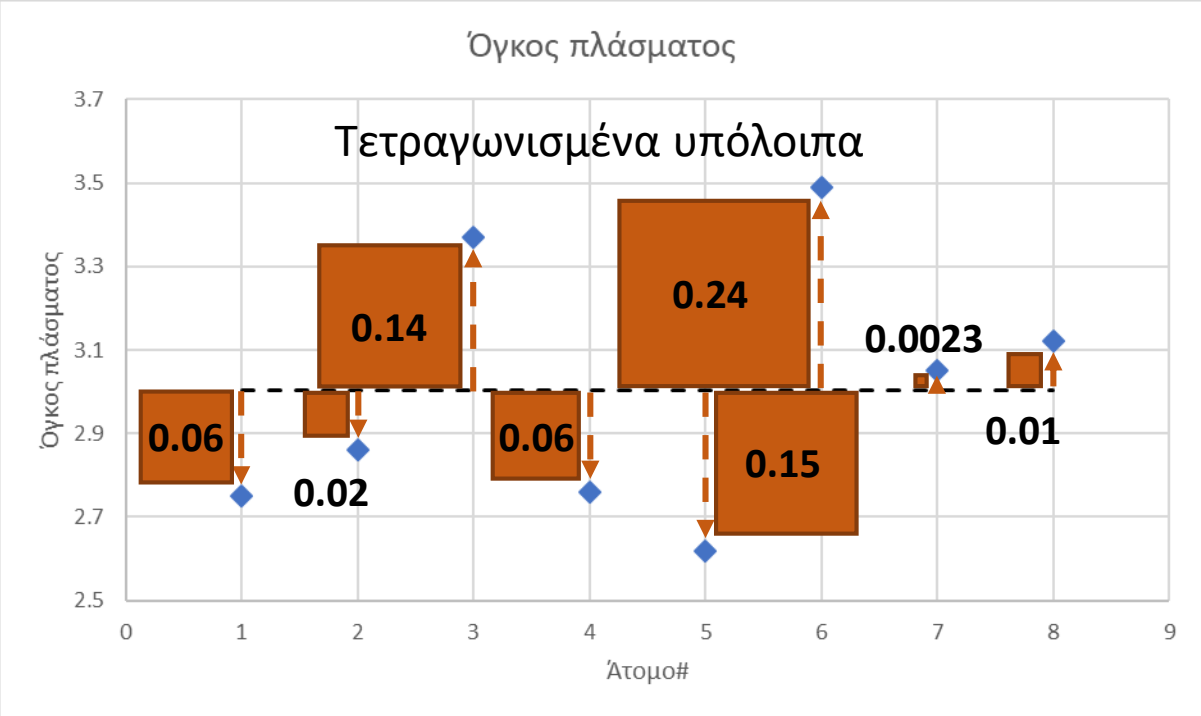
Τετράγωνα των υπολοίπων



Σκοπός της απλής γραμμικής παλινδρόμησης είναι να δημιουργηθεί ένα γραμμικό μοντέλο το οποίο θα ελαχιστοποιεί το άθροισμα των τετραγώνων των υπολοίπων $(sum\ of\ squares\ of\ the\ residuals\ (errors)\ /\ error\ (\boldsymbol{SSE}))$.

Εάν το μοντέλο της παλινδρόμησης είναι στατιστικά σημαντικό, θα "καταναλώσει" ένα μεγάλος μέρος από το SSE που είχαμε όταν υποθέσαμε ότι η ανεξάρτητη μεταβλητή δεν υπάρχει. Η γραμμή παλινδρόμησης **θα/πρέπει να** προσαρμόζει κυριολεκτικά τα δεδομένα καλύτερα. Θα ελαχιστοποιεί τα υπόλοιπα (residuals).

Πολύ σημαντικό



$$Sum\ of\ squared\ errors\ (residuals) = 0.6823$$

Όταν πραγματοποιείται η απλή γραμμική παλινδρόμηση με **ΔΥΟ** μεταβλητές, προσδιορίζουμε πόσο καλά ταιριάζει αυτή η ευθεία στα δεδομένα **συγκρίνοντάς τη με ΑΥΤΟΝ ΤΟΝ ΤΥΠΟ, όπου υποθέτουμε πως η ανεξάρτητη μεταβλητή δεν υπάρχει.**

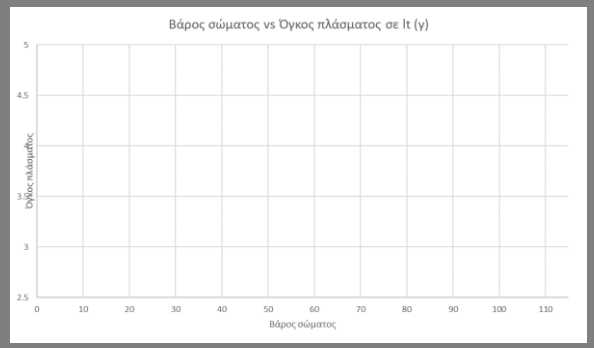
Αν ένα μοντέλο απλής γραμμικής παλινδρόμησης με δύο μεταβλητές φαίνεται σαν το διπλανό παράδειγμα, τι κάνει η άλλη μεταβλητή για να εξηγήσει την εξαρτημένη μεταβλητή;

ΤΙΠΟΤΑ

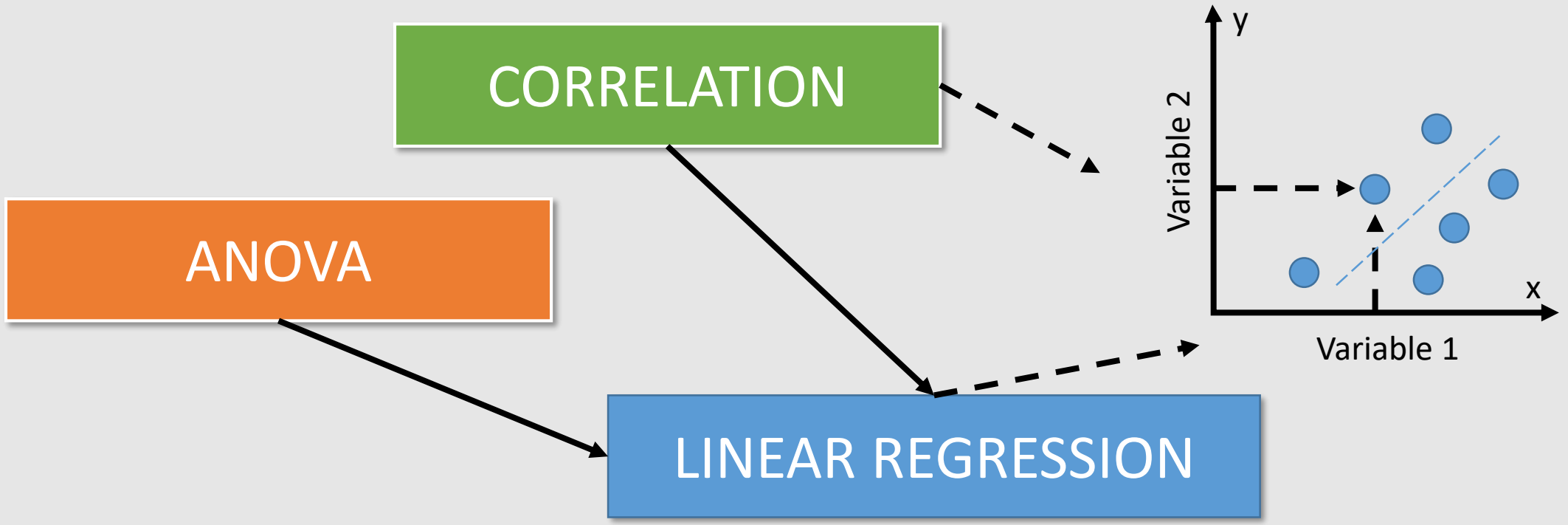
Γρήγορη ανασκόπηση

- Η **απλή γραμμική παλινδρόμηση** αποτελεί στην πραγματικότητα μία σύγκριση **δύο μοντέλων**
  - Το ένα μοντέλο είναι αυτό στο οποίο δεν υπάρχει η ανεξάρτητη μεταβλητή
  - Το άλλο μοντέλο κάνει χρήση της ευθείας καλής προσαρμογής στο οποίο γίνεται χρήση της δεύτερης μεταβλητής (ανεξάρτητης)
- Όταν υπάρχει μόνο μία μεταβλητή, η καλύτερη πρόβλεψη για νέες τιμές είναι η **μέση τιμή** της εξαρτημένης μεταβλητής
- Η διαφορά μεταξύ της της ευθείας καλής προσαρμογής και της παρατηρηθείσας τιμής ονομάζεται **υπόλοιπο** (residual) (ή error)
- Τα υπόλοιπα (residuals) τετραγωνίζονται και στη συνέχεια αθροίζονται για να δημιουργηθεί το άθροισμα των τετραγώνων (κυριολεκτικά) (sum of squares residuals / errors, SSE)
- Η **απλή γραμμική παλινδρόμηση** έχει σχεδιαστεί για να βρίσκει την καλύτερη ευθεία καλής προσαρμογής δια μέσω των δεδομένων που ελαχιστοποιεί το **SSE**

Άλγεβρα, εξισώσεις και γραφήματα



Bivariate Στατιστική



Η τιμή της **μίας μεταβλητής** είναι συνάρτηση **της άλλης μεταβλητής**
Η τιμή της $\boldsymbol{y}$ είναι συνάρτηση της $\boldsymbol{x}$; $y = f(x)$
Η τιμή της **εξαρτημένης μεταβλητής** είναι συνάρτηση της **ανεξάρτητης μεταβλητής**; $y = f(x)$

Επισκόπηση άλγεβρας: ευθείες

Η Κλίση της ευθείας (slope) και το σημείο τομής με τον κατακόρυφο άξονα (intercept) σχηματίζουν μία ευθεία

$y = mx + b$

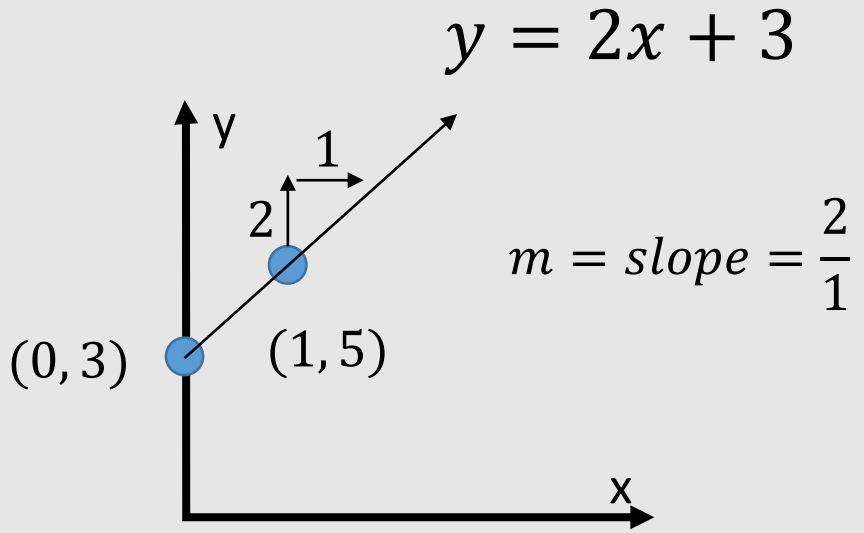
$x =$ μια τυχαία μεταβλητή

$m =$ κλίση της ευθείας (slope)

$b = y -$ σημείο τομής με κατακόρυφο άξονα (intercept)

$y -$ σημείο τομής με κατακόρυφο άξονα είναι όπου $x = 0$

Συντεταγμένη του $(0, y)$



$y = 2(0) + 3$ $\quad$ $y = 2(1) + 3$

$y = 3$ $\quad$ $y = 5$

$(0, 3)$ $\quad$ $(1, 5)$

Μοντέλο απλής γραμμικής παλινδρόμησης

$y = mx + b$ → $y = \beta_0 + \beta_1 x + \varepsilon$

$\beta_0 = y - intercept$ της παραμέτρου του **πληθυσμού**

$\beta_1 = slope$ της παραμέτρου του **πληθυσμού**

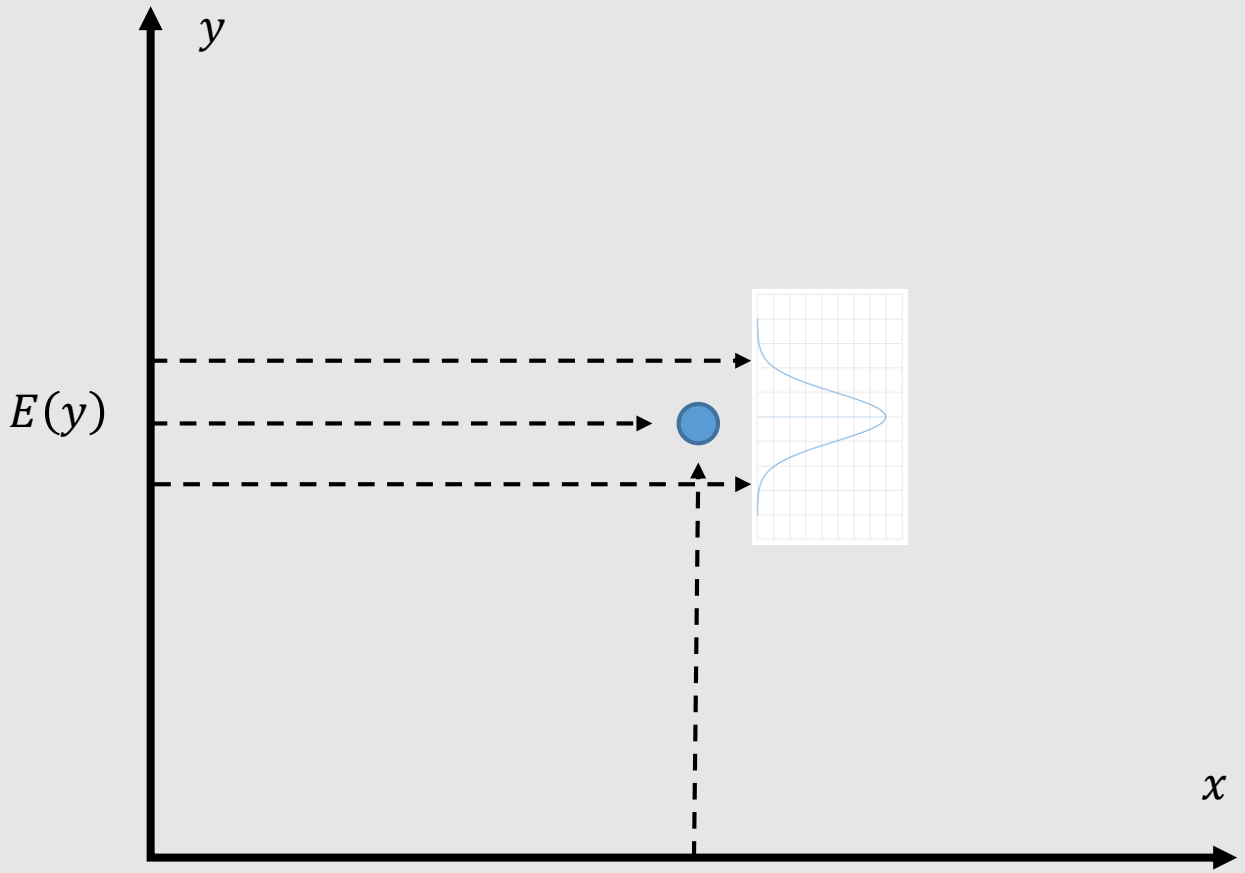
$\varepsilon =$ σφάλμα, διακύμανση που δεν μπορεί να εξηγηθεί στο y

Εξίσωση απλής γραμμικής παλινδρόμησης

$$E(y) = \beta_0 + \beta_1 x$$

$E(y) =$ είναι η μέση τιμή ή η αναμενόμενη τιμή της y, για δοθείσα τιμή της x

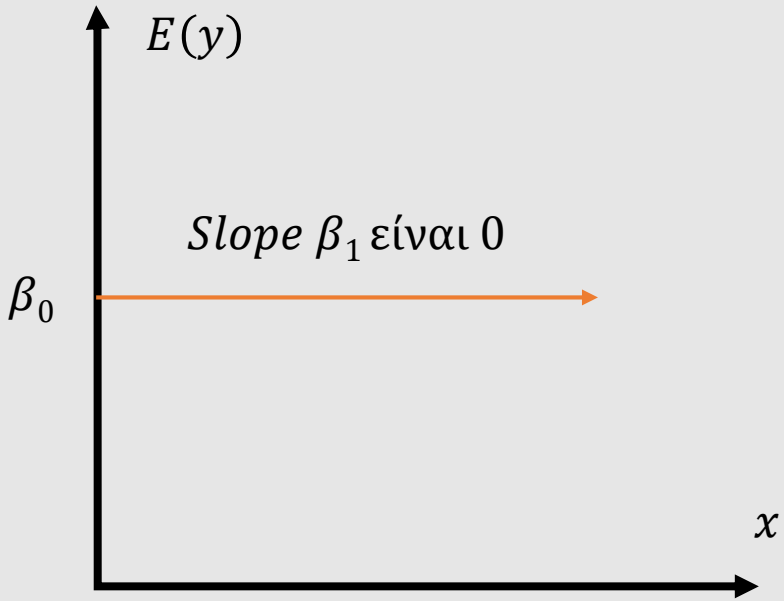
Κατανομή των $y-$ τιμών



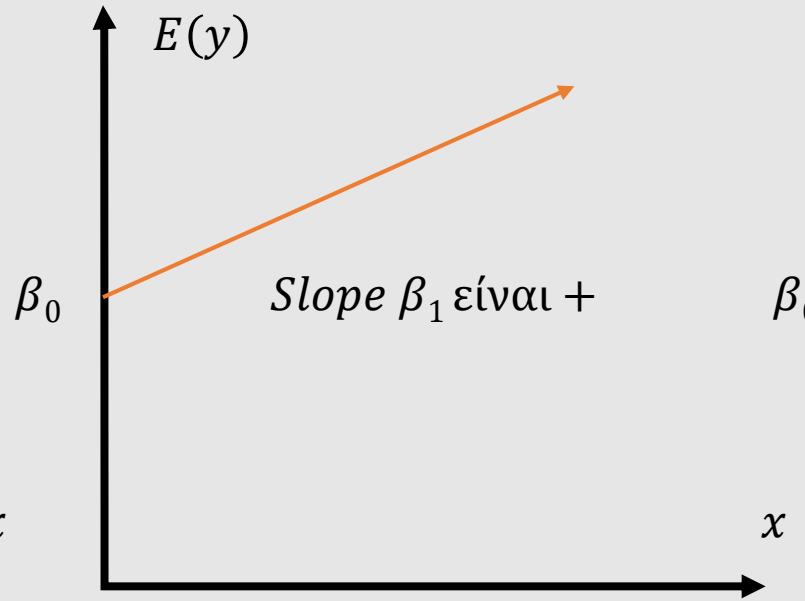
- Η αναμενόμενη τιμή είναι η **μέση τιμή**. Τι ακριβώς σημαίνει αυτό;
- Αν επιλέξουμε μία τιμή για x αυτή θα αντιστοιχεί σε μία τιμή y. Αυτή όμως είναι η αναμενόμενη τιμή του y. Και αυτό δεν είναι τόσο απλό καθώς υπάρχει μία κατανομή από y για αυτή τη x.
- Θυμηθείτε πως το μοντέλο μας δεν πρόκειται να είναι τέλειο. Κάθε αναμενόμενη τιμή για την y είναι και μία εκτίμηση της y. Οπότε όταν αναφερόμαστε σε μία αναμενόμενη τιμή εννοούμε τη μέση τιμή μίας μικρής κατανομής για την y.

Γενικές ευθείες παλινδρόμησης

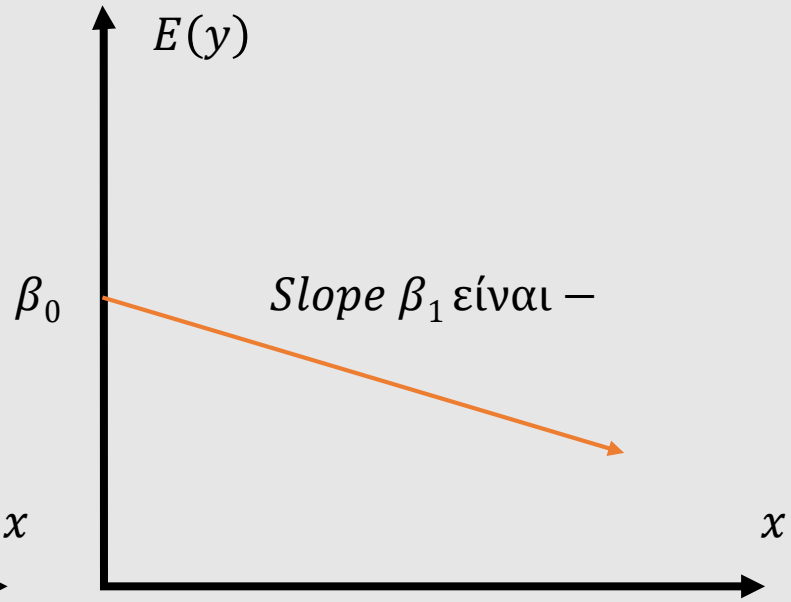
$$E(y) = \beta_0 + \beta_1 x$$



$$E(y) = \beta_0 + (0)x$$



$$E(y) = \beta_0 + \beta_1 x$$



$$E(y) = \beta_0 - \beta_1 x$$

Εξίσωση παλινδρόμησης με εκτιμήσεις

Αν γνωρίζαμε τις παραμέτρους του πληθυσμού, β_0 και β_1, θα κάναμε χρήση της εξίσωσης της απλής γραμμικής παλινδρόμησης

$$E(y) = \beta_0 + \beta_1 x$$

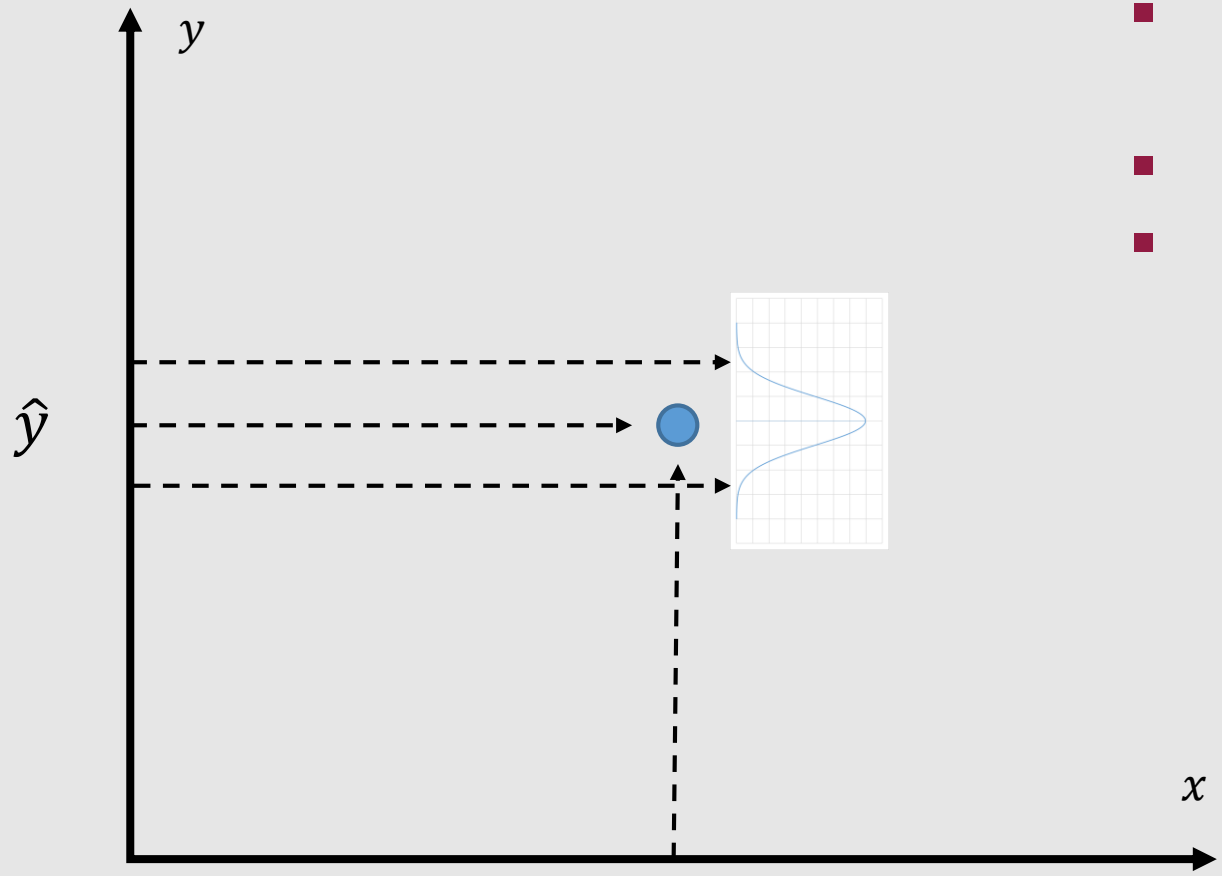
Επειδή στην πραγματικότητα δεν γνωρίζουμε τις παραμέτρους του πληθυσμού θα πρέπει να τις εκτιμήσουμε χρησιμοποιώντας δείγματα. Στην περίπτωση αυτή η εξίσωση αλλάζει λίγο και γίνεται

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

Το $\hat{y}$ (y καπέλο) είναι η εκτίμηση της αναμενόμενης τιμής της $E(y)$.

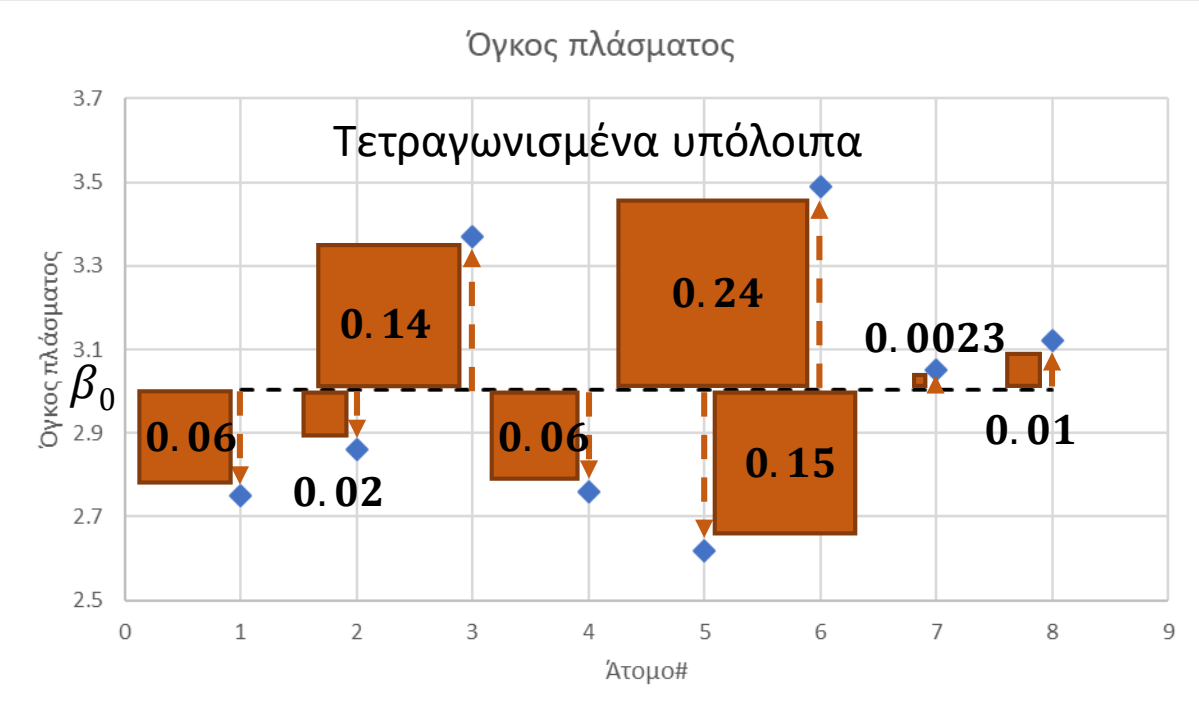
Η $\hat{y}$ είναι η μέση τιμή των y για μία δοθείσα τιμή x

Κατανομή των y −τιμών



- Κάνουμε χρήση του $\hat{y}$ καθώς χρησιμοποιούμε δείγματα
- Η βασική ιδέα παραμένει όμως η ίδια
- Η $\hat{y}$ είναι η μέση τιμή των αναμενόμενων τιμών y για οποιαδήποτε τιμή x

Όταν το slope, $\boldsymbol{\beta_1 = 0}$



$$Sum\ of\ squared\ errors\ (residuals) = 0.6823$$

Όταν πραγματοποιείται απλή γραμμική παλινδρόμηση με **ΔΥΟ** μεταβλητές, προσδιορίζουμε πόσο καλά ταιριάζει αυτή η ευθεία στα δεδομένα **συγκρίνοντάς τη με αυτόν τον τύπο, όπου υποθέτουμε πως η ανεξάρτητη μεταβλητή δεν υπάρχει**˙

$$\boldsymbol{slope\ =\ 0, \beta_1 = 0}$$

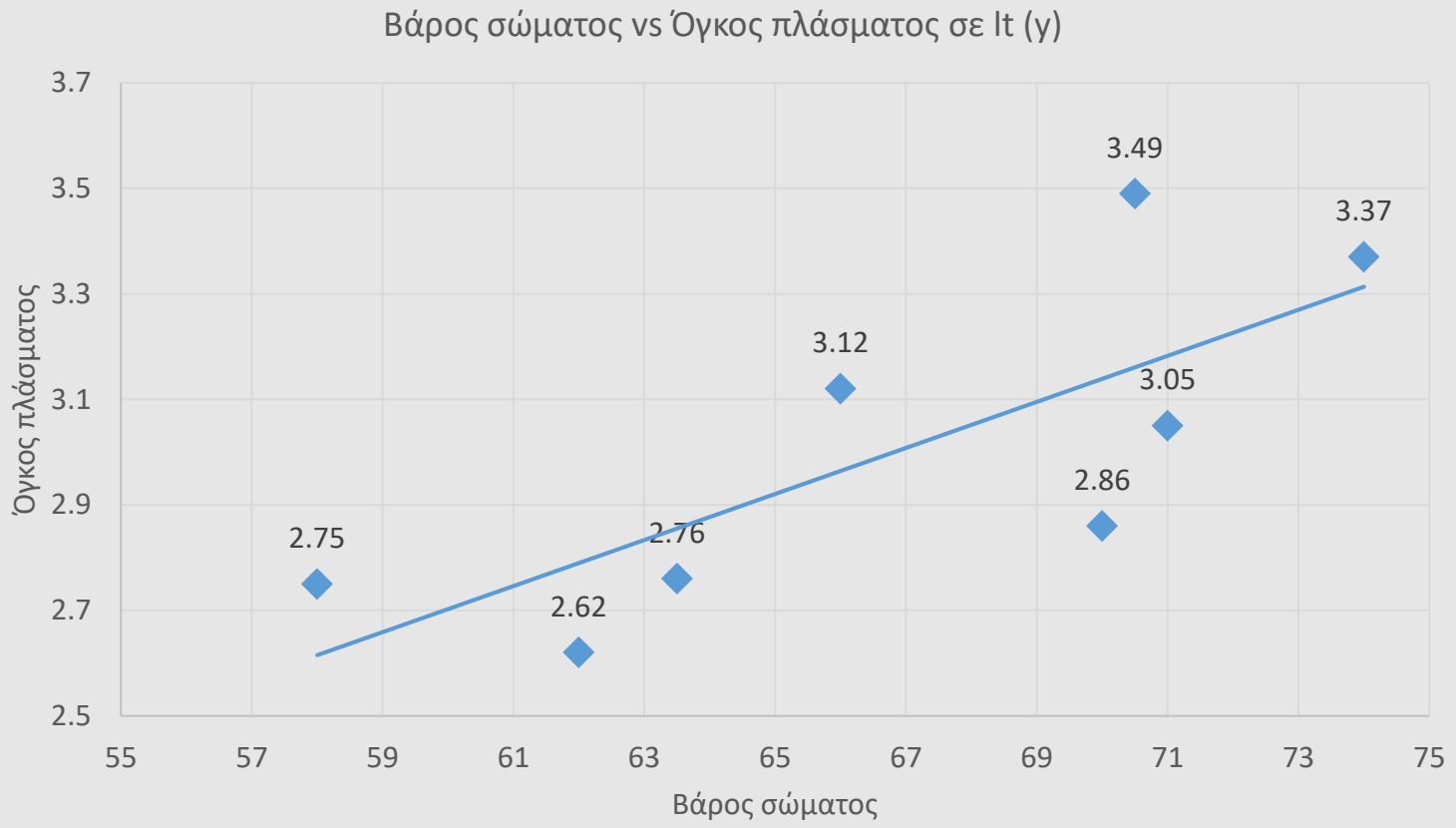
Στην περίπτωση αυτή, η τιμή της $\hat{y}$ είναι 3.0025 **για κάθε τιμή** x.

$$\hat{y} = b_0 + b_1 x \qquad b_0 = 3.0025$$

$$\hat{y} = b_0 + (0)x \qquad \hat{y} = 3.0025$$

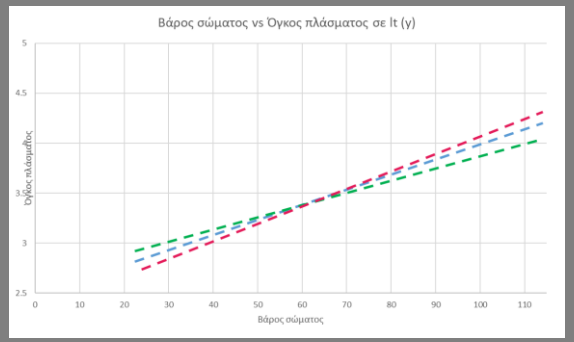
$$\hat{y} = b_0$$

Προετοιμασία για τη μέθοδο ελάχιστων τετραγώνων



Βάρος σε Kg (x)	Όγκος πλάσματος σε lt (y)
58.0	2.75
70.0	2.86
74.0	3.37
63.5	2.76
62.0	2.62
70.5	3.49
71.0	3.05
66.0	3.12

Μέθοδος ελάχιστων τετραγώνων



Παράδειγμα

Μέχρι τώρα είχαμε μόνο τον όγκο του πλάσματος. Καταφέραμε όμως να βρουμε και το βάρος του σώματος. Οπότε τώρα δουλεύουμε με δύο μεταβλητές που είναι σε ζεύγη

Θέλουμε να γνωρίζουμε σε ποιο βαθμό μπορεί να προβλεφθεί ο όγκος του πλάσματος από το βάρος του σώματος

Οπότε η **ΕΞΑΡΤΗΜΕΝΗ ΜΕΤΑΒΛΗΤΗ** είναι ο όγκος τους πλάσματος και η **ΑΝΕΞΑΡΤΗΤΗ ΜΕΤΑΒΛΗΤΗ** το βάρος του σώματος

Βάρος σε Kg (x)	Όγκος πλάσματος σε lt (y)
58.0	2.75
70.0	2.86
74.0	3.37
63.5	2.76
62.0	2.62
70.5	3.49
71.0	3.05
66.0	3.12

Μέθοδος ελάχιστων τετραγώνων

$$min \sum (y_i - \hat{y}_i)^2$$

y_i = παρατηρηθείσα τιμή για την εξαρτημένη μεταβλητή (όγκος πλάσματος)

$\hat{y}_i$ = εκτιμώμενη (προβλεπόμενη) τιμή για την εξαρτημένη μεταβλητή (όγκος πλάσματος)

- Αυτό που παρατηρούμε είναι πως έχουμε δύο τιμές για κάθε x τιμή του γραφήματος
- Έχουμε την παρατηρηθείσα τιμή y_i και την τιμή $\hat{y}_i$ που εκτιμά το μοντέλο
- Οι τιμές αυτές δεν είναι ίδιες. Θα υπάρχουν διαφορές μεταξύ τους
- Τετραγωνίζουμε τις διαφορές και στην συνέχεια τις αθροίζουμε
- Θέλουμε το άθροισμα να είναι όσο πιο ελάχιστο γίνεται

Μέθοδος ελάχιστων τετραγώνων

$$min \sum (y_i - \hat{y}_i)^2$$

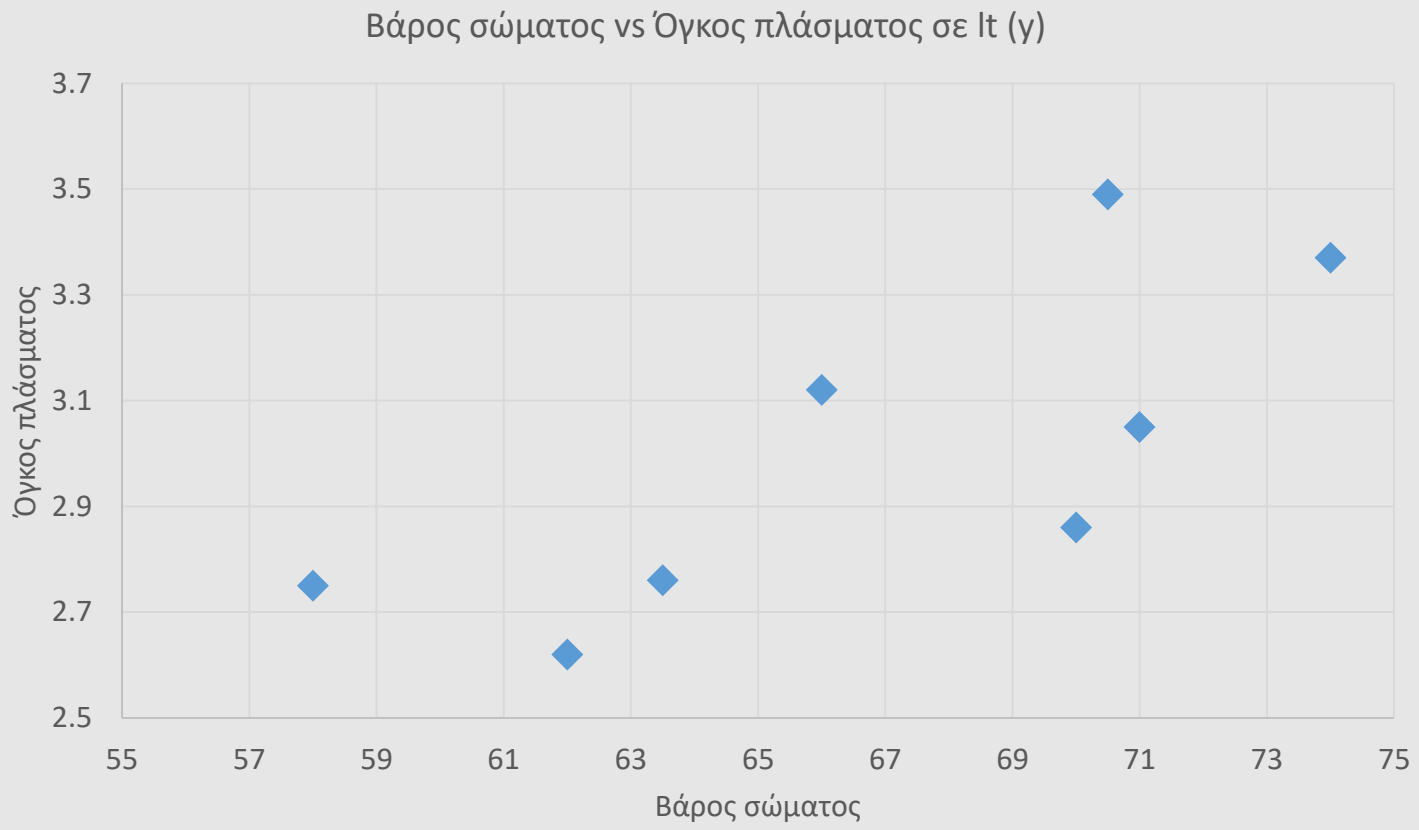
$y_i = \text{παρατηρηθείσα τιμή για την εξαρτημένη μεταβλητή (όγκος πλάσματος)}$

$\hat{y}_i = \text{εκτιμώμενη (προβλεπόμενη) τιμή για την εξαρτημένη μεταβλητή (όγκος πλάσματος)}$

Σκοπός είναι να ελαχιστοποιηθεί το άθροισμα των τετραγωνισμένων διαφορών μεταξύ της παρατηρηθείσας τιμής για την εξαρτημένη μεταβλητή (y_i) και της εκτιμώμενης/προβλεπόμενης τιμής της εξαρτημένης μεταβλητής $(\hat{y}_i)$ που παρέχεται από την γραμμή παλινδρόμησης. Άθροισμα των τετραγώνων των υπολοίπων.

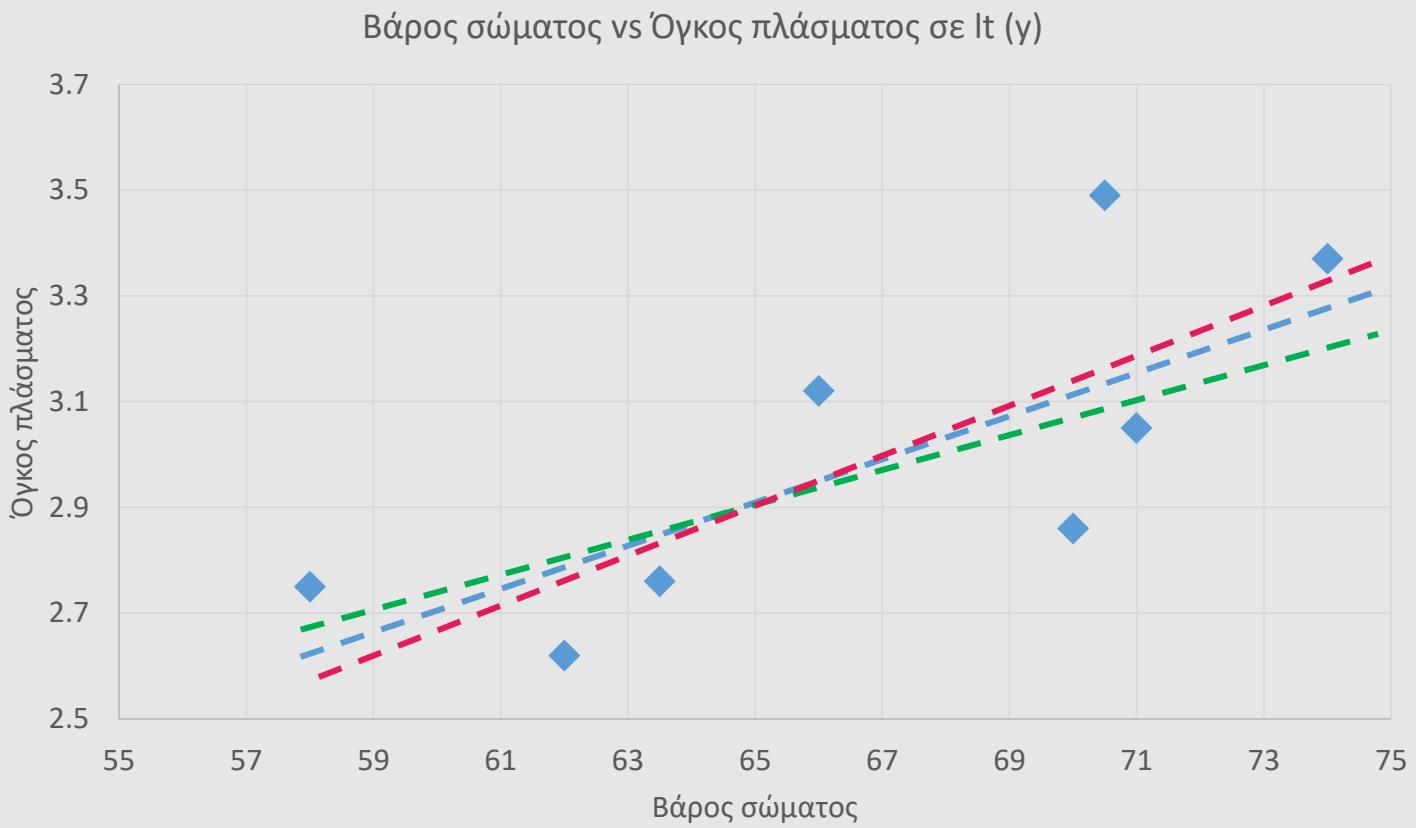
Και όχι μόνο αυτό, αλλά το άθροισμα των τετραγώνων των υπολοίπων θα πρέπει να είναι πολύ μικρότερο από αυτό που κάναμε χρήση όταν είχαμε μόνο την εξαρτημένη μεταβλητή· $\beta_1 = 0, \hat{y} = 3.0025$ για όλες τις τιμές του x. Το άθροισμα των τετραγώνων των υπολοίπων ήταν 0.6823.

Βήμα 1: Διάγραμμα συσχέτισης (Scatter plot)



Βάρος σε Kg (x)	Όγκος πλάσματος σε lt (y)
58.0	2.75
70.0	2.86
74.0	3.37
63.5	2.76
62.0	2.62
70.5	3.49
71.0	3.05
66.0	3.12

Βήμα 2: Αναζήτηση μίας γραμμής



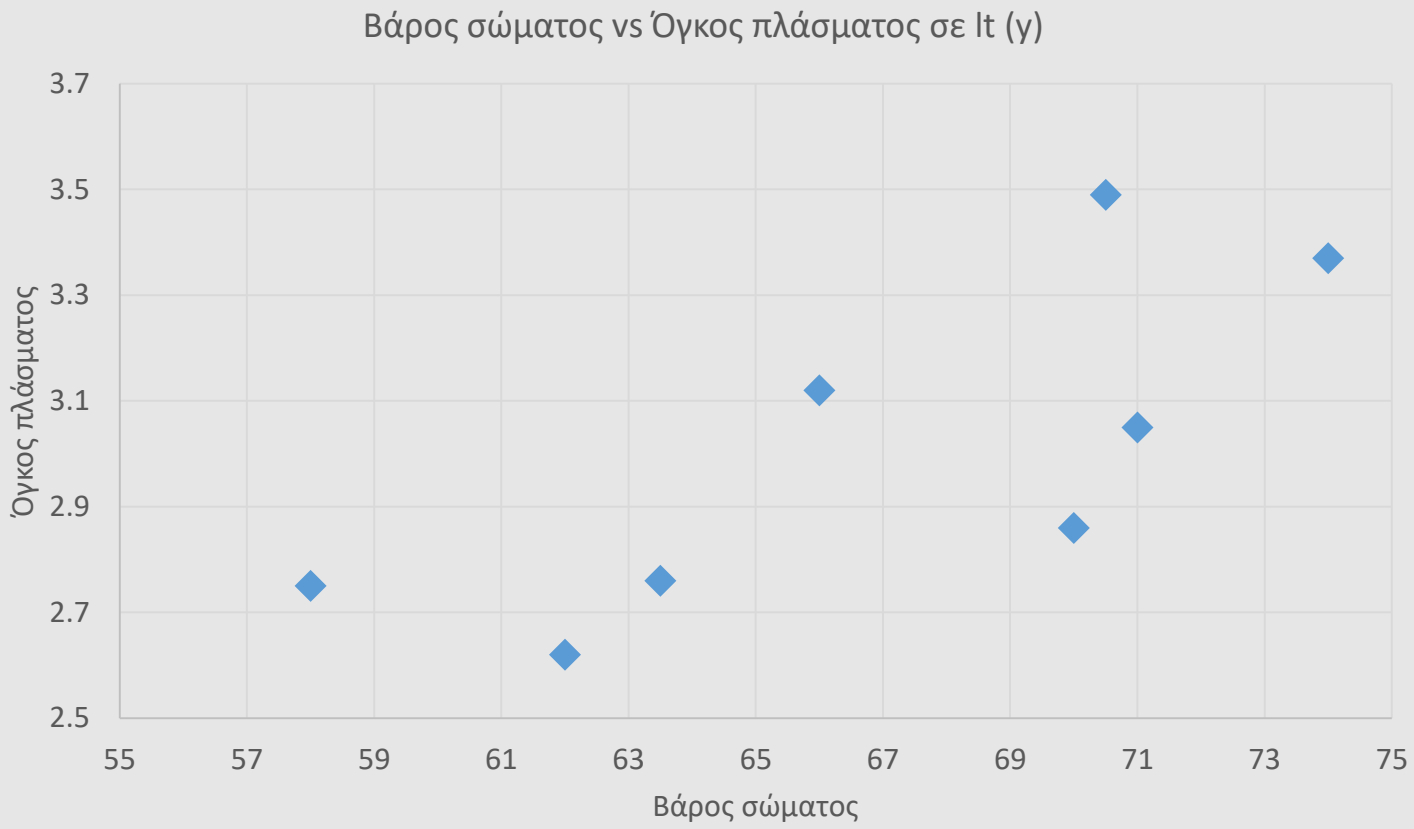
Φαίνεται να υπάρχει κάποια ευθεία που να διαπερνάει τα δεδομένα;

Στην περίπτωση αυτή, **ΝΑΙ**! Οπότε, προχωράμε.

Δεν γνωρίζουμε όμως ποια από αυτές τις ευθείες είναι η γραμμή παλινδρόμησης.

Αν όμως δεν φαίνεται κάτι τέτοιο, τότε σταματάμε.

Βήμα 3: Συσχέτιση (Correlation) (προαιρετικό)



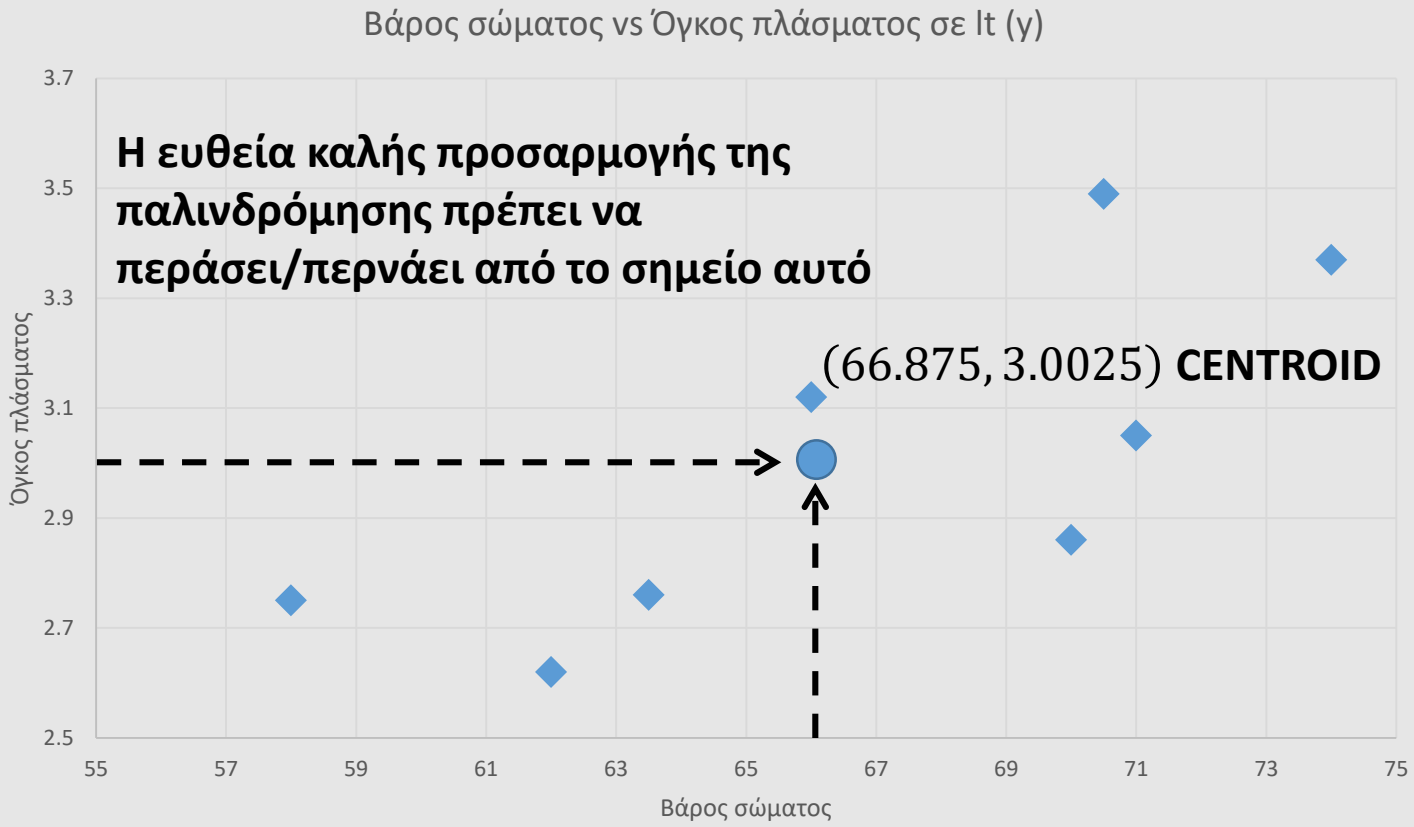
Ποιος είναι ο συντελεστής συσχέτισης r;

Στην περίπτωση αυτή,
$r = 0.743$

Είναι αυτή η σχέση ισχυρή;

Στην περίπτωση αυτή, **ΝΑΙ**!

Βήμα 4: Περιγραφική Στατιστική / Κεντρική τιμή (Centroid)



$\bar{y} = 3.0025$

$\bar{x} = 66.875$

Βάρος σε Kg (x)	Όγκος πλάσματος σε lt (y)
58.0	2.75
70.0	2.86
74.0	3.37
63.5	2.76
62.0	2.62
70.5	3.49
71.0	3.05
66.0	3.12
$\bar{x} = 66.875$	$\bar{y} = 3.0025$

Βήμα 5: Υπολογισμοί

Intercept (τομή με y)

$$\hat{y}_i = b_0 + b_1 x_i$$

Slope (κλίση)

$$b_0 = \bar{y} - b_1 \bar{x}$$

$$b_1 = \frac{\sum (x_i - \bar{x})\,(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

$\bar{x}$ = μέση τιμή της ανεξάρτητης μεταβλητής

$\bar{y}$ = μέση τιμή της εξαρτημένης μεταβλητής

y_i = τιμή της εξαρτημένης μεταβλητής

x_i = τιμή της ανεξάρτητης μεταβλητής

Βήμα 5: Υπολογισμοί

$$b_1 = \frac{\sum(x_i - \bar{x})\,(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

1. Για κάθε σημείο
2. Παίρνουμε την x τιμή και την αφαιρούμε την από τη μέση τιμή των x
3. Παίρνουμε την y τιμή και την αφαιρούμε από τη μέση τιμή των y
4. Πολλαπλασιάζουμε το βήμα 2 και 3
5. Προσθέτουμε όλους τους πολλαπλασιασμούς

---

1. Για κάθε σημείο
2. Παίρνουμε την x τιμή και την αφαιρούμε από τη μέση τιμή των x
3. Τετραγωνίζουμε το βήμα 2
4. Προσθέτουμε όλα τα τετράγωνα από το βήμα 2

$$b_0 = \bar{y} - b_1\bar{x}$$

Βήμα 5: Υπολογισμοί

Άτομο	Βάρος	Όγκος πλάσματος σε lt (y)	Αποκλίσεις βάρους	Αποκλίσεις όγκου πλάσματος	Γινόμενα αποκλίσεων	Τετραγωνισμένες αποκλίσεις βάρους
	x	y	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
1	58	2.75	-8.875	-0.2525	2.2409375	78.765625
2	70	2.86	3.125	-0.1425	-0.4453125	9.765625
3	74	3.37	7.125	0.3675	2.6184375	50.765625
4	63.5	2.76	-3.375	-0.2425	0.8184375	11.390625
5	62	2.62	-4.875	-0.3825	1.8646875	23.765625
6	70.5	3.49	3.625	0.4875	1.7671875	13.140625
7	71	3.05	4.125	0.0475	0.1959375	17.015625
8	66	3.12	-0.875	0.1175	-0.1028125	0.765625
	$\bar{x} = \mathbf{66.875}$	$\bar{y} = \mathbf{3.0025}$			$\sum = \mathbf{8.9575}$	$\sum = \mathbf{205.375}$

Υπολογισμοί β_1 (slope)

$$b_1 = \frac{\sum(x_i - \bar{x})\,(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

$$b_1 = \frac{8.9575}{205.375}$$

$$b_1 = 0.04362$$

Η κλίση (slope) της ευθείας παλινδρόμησής μας είναι 0.04362

Γινόμενα αποκλίσεων	Τετραγωνισμένες αποκλίσεις βάρους
$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
2.2409375	78.765625
−0.4453125	9.765625
2.6184375	50.765625
0.8184375	11.390625
1.8646875	23.765625
1.7671875	13.140625
0.1959375	17.015625
−0.1028125	0.765625
∑ = 8.9575	**∑ = 205.375**

Υπολογισμοί β_0 (y − intercept)

$$b_0 = \bar{y} - b_1\bar{x}$$

$$b_0 = 3.0025 - 0.04362 \times 66.875$$

$$b_0 = 0.0857$$

Βάρος	Όγκος πλάσματος σε lt (y)
x	y
58	2.75
70	2.86
74	3.37
63.5	2.76
62	2.62
70.5	3.49
71	3.05
66	3.12
$\bar{x} =$ **66.875**	$\bar{y} =$ **3.0025**

Γραμμή παλινδρόμησης

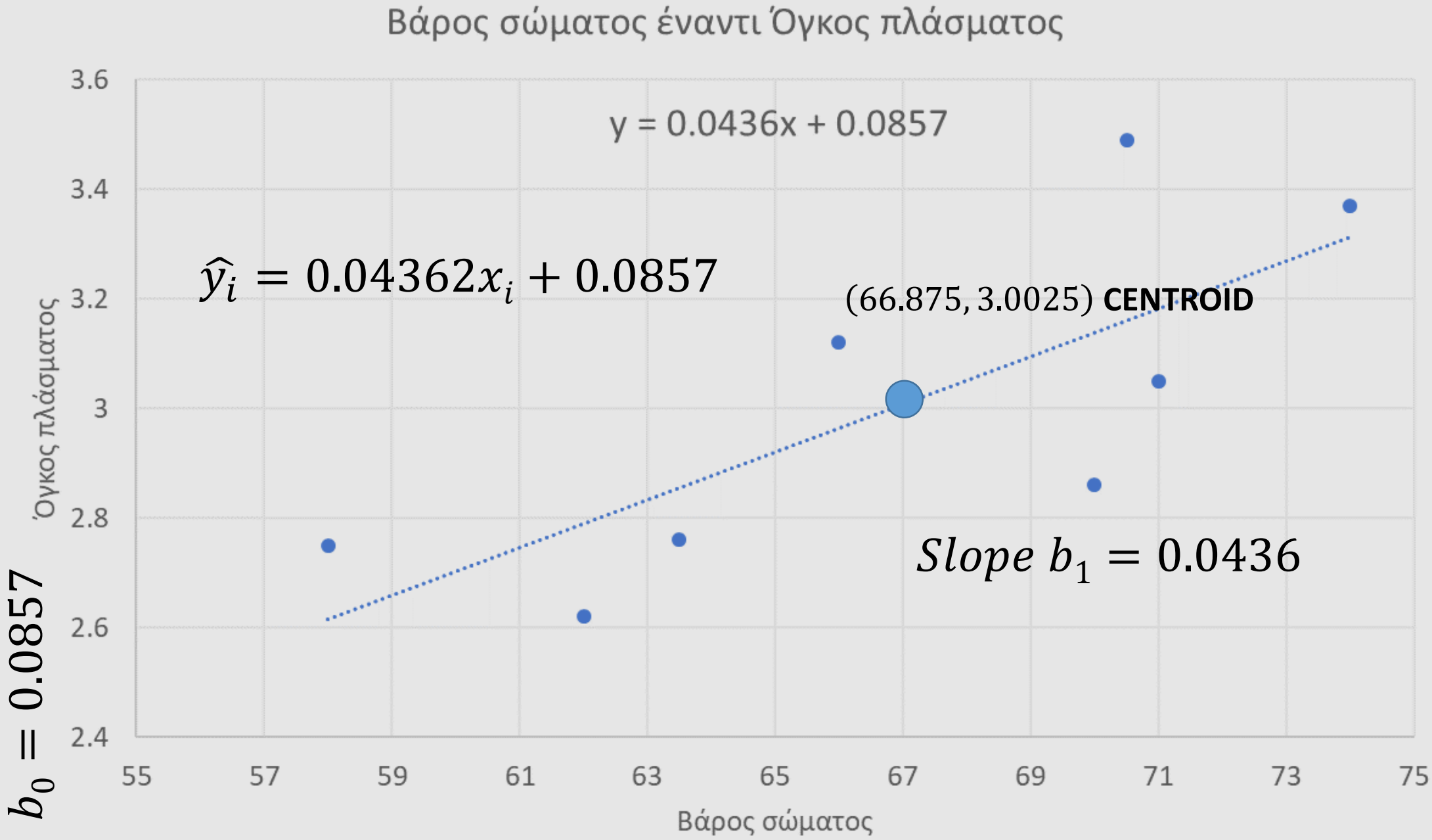
$\hat{y}_i = b_0 + b_1 x_i$ $\quad b_0 = 0.0857$ $\quad b_1 = 0.04362$

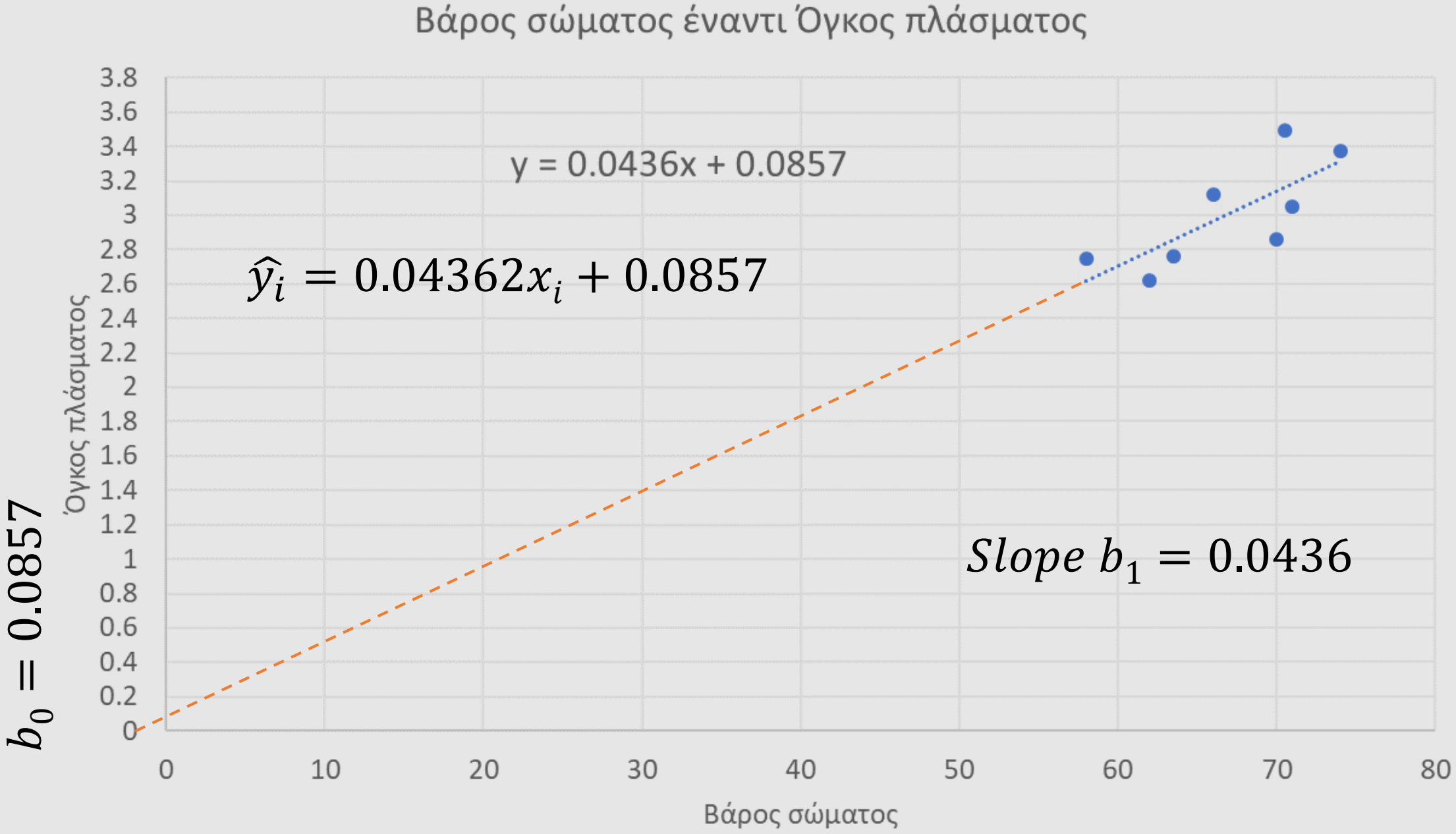
intercept slope

$$\hat{y}_i = 0.0857 + 0.04362 x_i$$

ή

$$\hat{y}_i = 0.04362 x_i + 0.0857$$


Βάρος σώματος έναντι Όγκος πλάσματος
y = 0.0436x + 0.0857
$\hat{y}_i = 0.04362x_i + 0.0857$
(66.875, 3.0025) CENTROID
$Slope\ b_1 = 0.0436$
$b_0 = 0.0857$
Όγκος πλάσματος
Βάρος σώματος
3.6
3.4
3.2
3
2.8
2.6
2.4
55
57
59
61
63
65
67
69
71
73
75


Βάρος σώματος έναντι Όγκος πλάσματος
y = 0.0436x + 0.0857
$\hat{y}_i = 0.04362x_i + 0.0857$
$b_0 = 0.0857$
$Slope\ b_1 = 0.0436$
Όγκος πλάσματος
Βάρος σώματος
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
2.2
2.4
2.6
2.8
3
3.2
3.4
3.6
3.8
0
10
20
30
40
50
60
70
80

Γρήγορη ερμηνεία

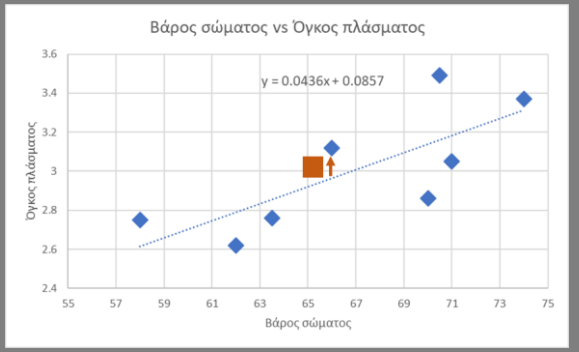
$$\hat{y}_i = 0.04362x_i + 0.0857$$

Για κάθε $1kgr$ αύξησης του βάρους του σώματος (x_i), θα αναμέναμε μία αύξηση του όγκου του πλάσματος κατά 0.04362.

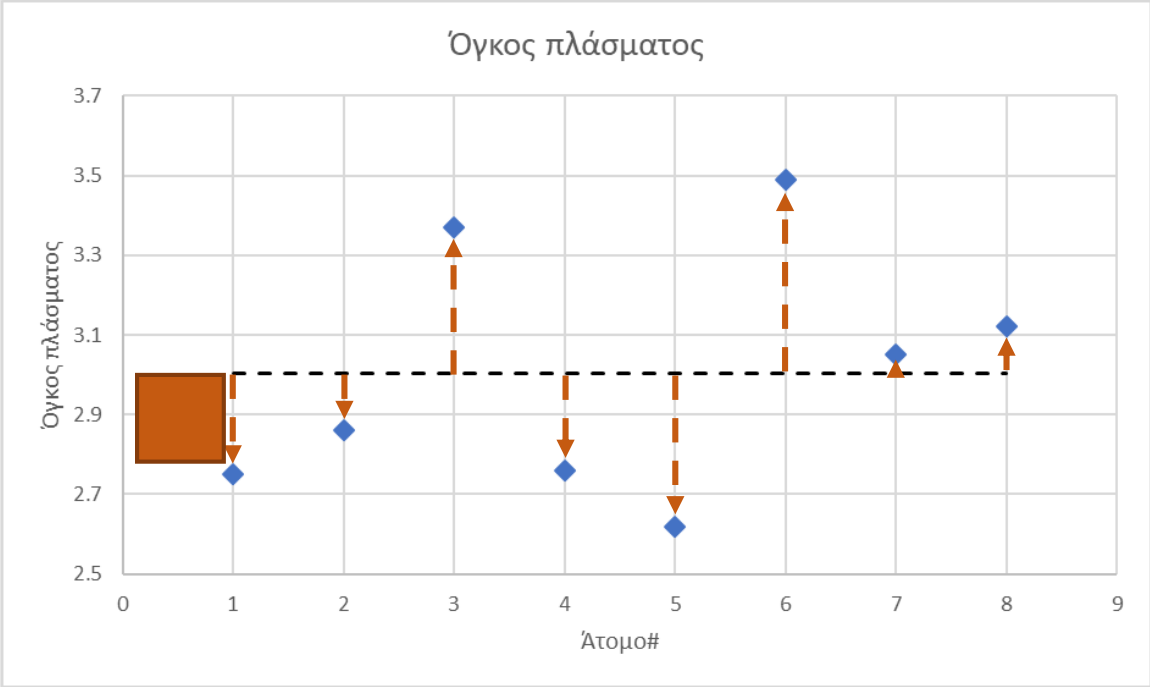
Αν το βάρος είναι μηδέν, τότε η αναμενόμενη/προβλεπόμενη τιμή του όγκου πλάσματος θα είναι 0.0857.
Έχει αυτό νόημα; ΟΧΙ. Η τομή (intercept) δεν έχει νόημα στην ιατρική

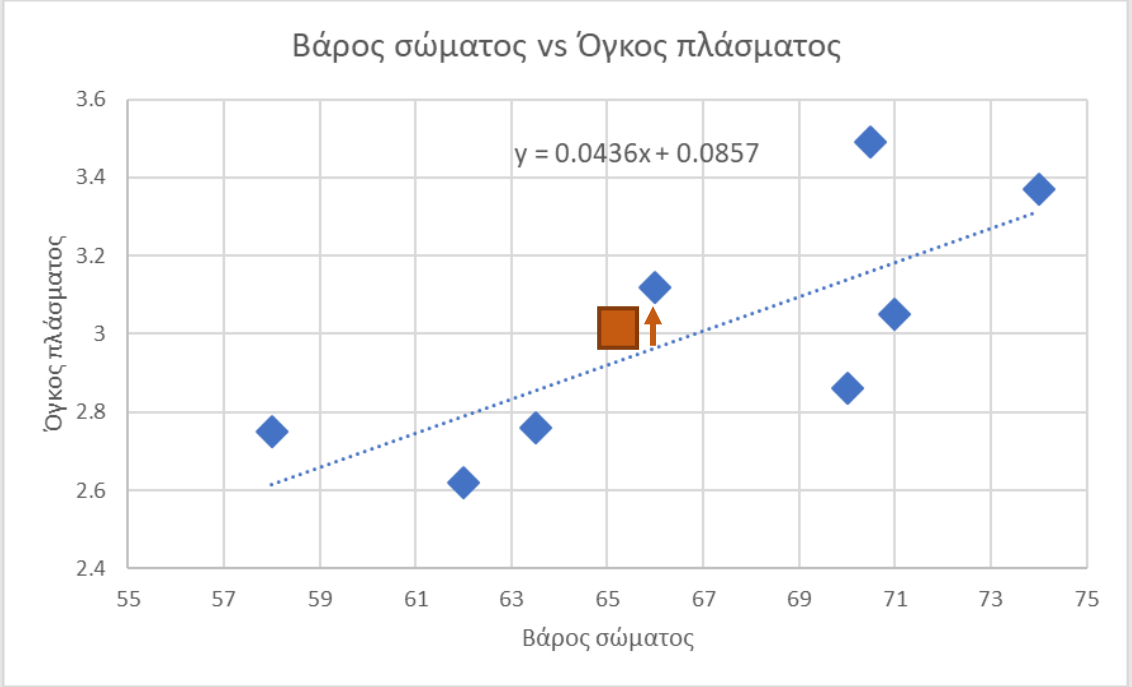
Είναι όμως το μοντέλο αυτό της γραμμικής παλινδρόμησης καλό;

Συντελεστής Προσδιορισμού



Ευθείες παλινδρόμησης





$SSE = 0.6823$

$SSE = SST$

$SST = 0.6823$

Με μόνο μία μεταβλητή, το μόνο άθροισμα των τετραγώνων οφείλεται στο σφάλμα. Επομένως είναι και το συνολικό και ΜΕΓΙΣΤΟ άθροισμα των τετραγώνων για τα δεδομένα υπό ανάλυση.

$SST = 0.6823$

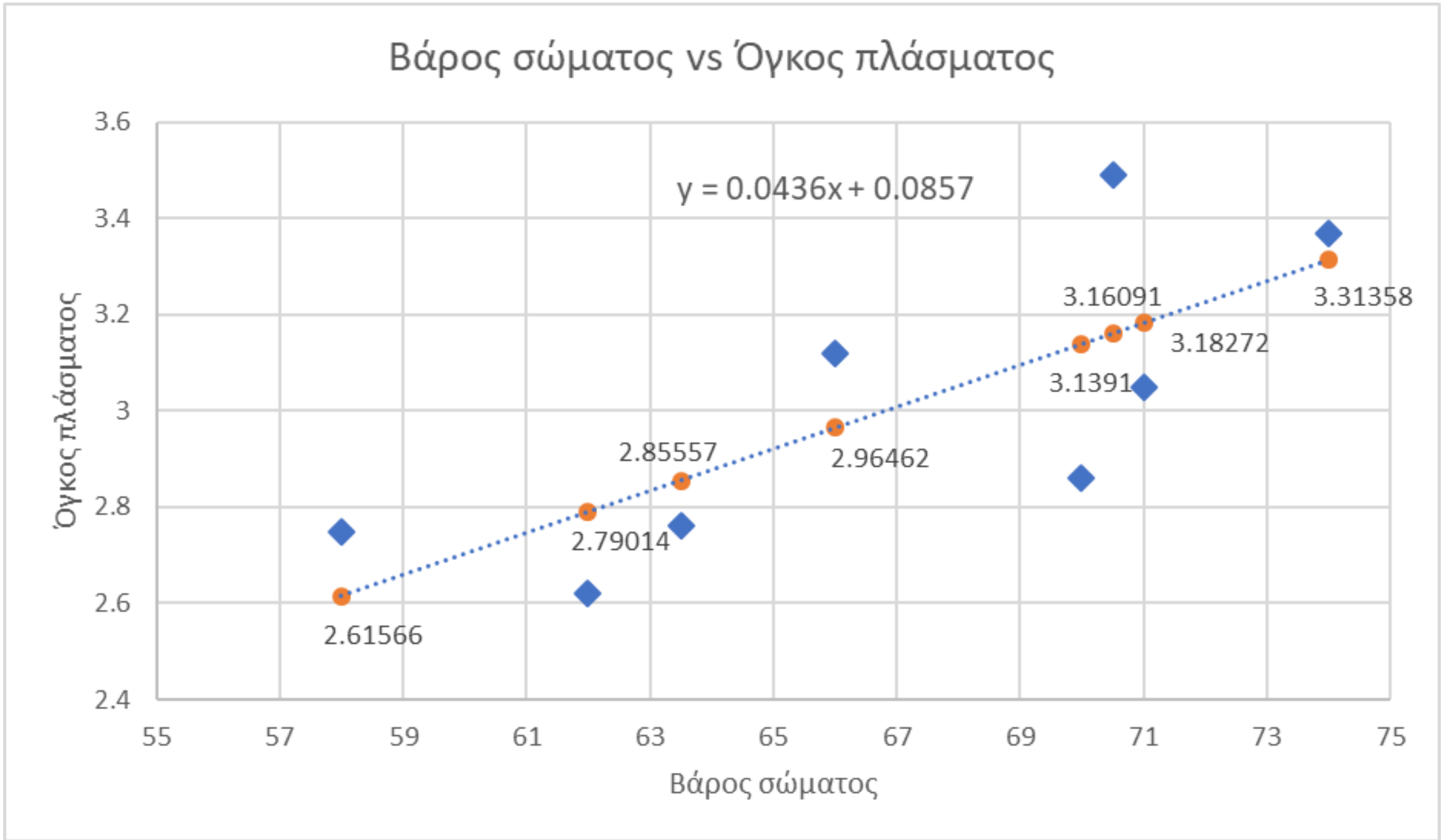
$\mathrm{SSE} = ?$

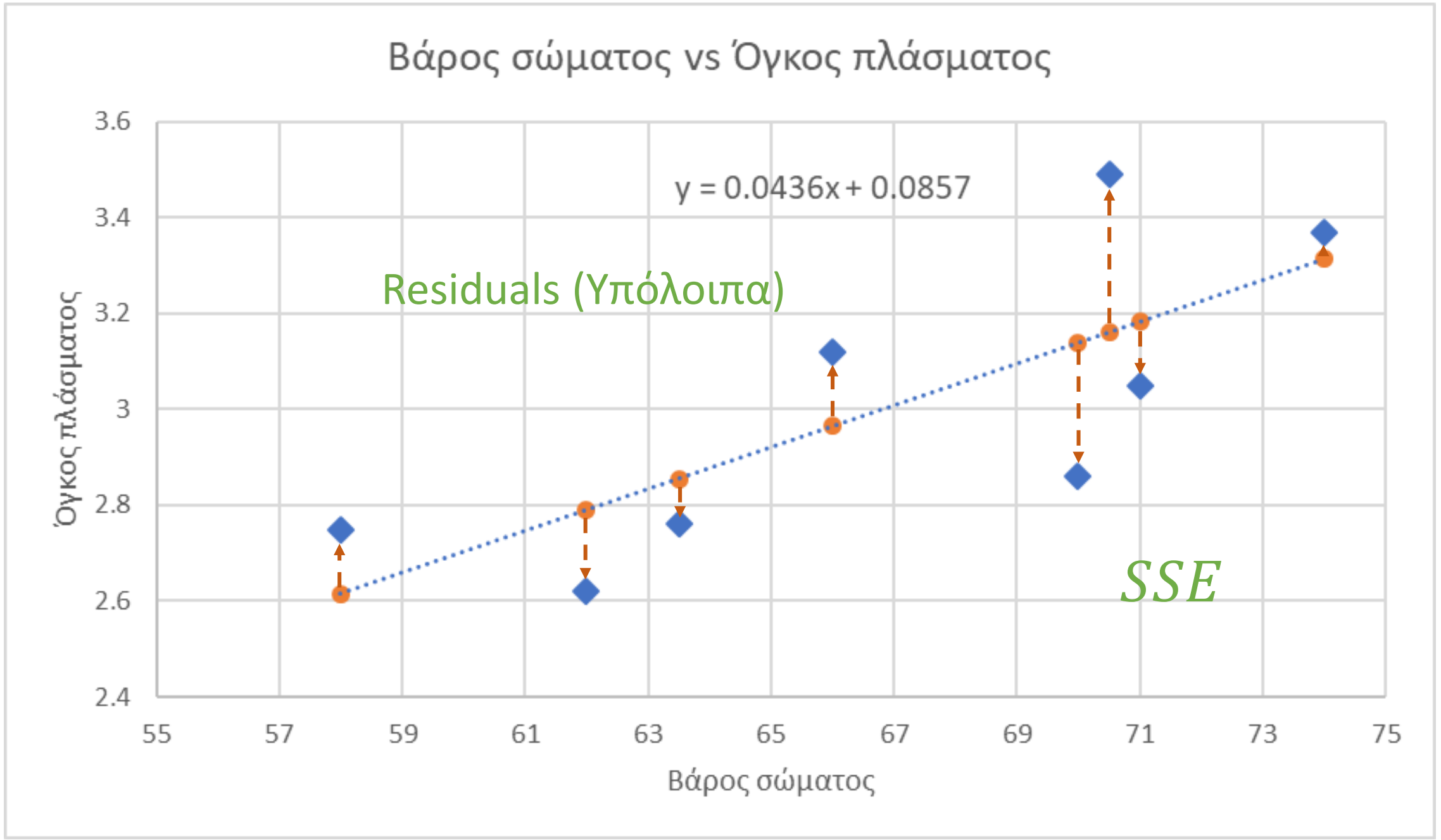
$SST - \mathrm{SSE} = SSR$

Με και τις δύο μεταβλητές το συνολικό άθροισμα τετραγώνων παραμένει το ίδιο. Αλλά (ιδανικά) το άθροισμα των τετραγώνων του σφάλματος μειώνεται σημαντικά, καθώς η ευθεία προσαρμόζεται καλύτερα στα δεδομένα. Η διαφορά μεταξύ του SST και του SSE οφείλεται στην παλινδρόμηση, SSR.

Εκτίμηση των τιμών παλινδρόμησης

Άτομο	Βάρος	Όγκος πλάσματος σε lt (y)	$\hat{y}_i = 0.044x_i + 0.0857$	$\hat{y}_i$ (προβλεπόμενος όγκος πλάσματος)
	x	y		
1	58	2.75	$\hat{y}_i = 0.044(58) + 0.0857$	2.61566
2	70	2.86	$\hat{y}_i = 0.044(70) + 0.0857$	3.1391
3	74	3.37	$\hat{y}_i = 0.044(74) + 0.0857$	3.31358
4	63.5	2.76	$\hat{y}_i = 0.044(63.5) + 0.0857$	2.85557
5	62	2.62	$\hat{y}_i = 0.044(62) + 0.0857$	2.79014
6	70.5	3.49	$\hat{y}_i = 0.044(70.5) + 0.0857$	3.16091
7	71	3.05	$\hat{y}_i = 0.044(71) + 0.0857$	3.18272
8	66	3.12	$\hat{y}_i = 0.044(66) + 0.0857$	2.96462
	$\bar{x} = 66.875$	$\bar{y} = 3.0025$		


Βάρος σώματος vs Όγκος πλάσματος
y = 0.0436x + 0.0857
Όγκος πλάσματος
Βάρος σώματος
3.6
3.4
3.2
3
2.8
2.6
2.4
55
57
59
61
63
65
67
69
71
73
75
2.61566
2.79014
2.85557
2.96462
3.1391
3.16091
3.18272
3.31358

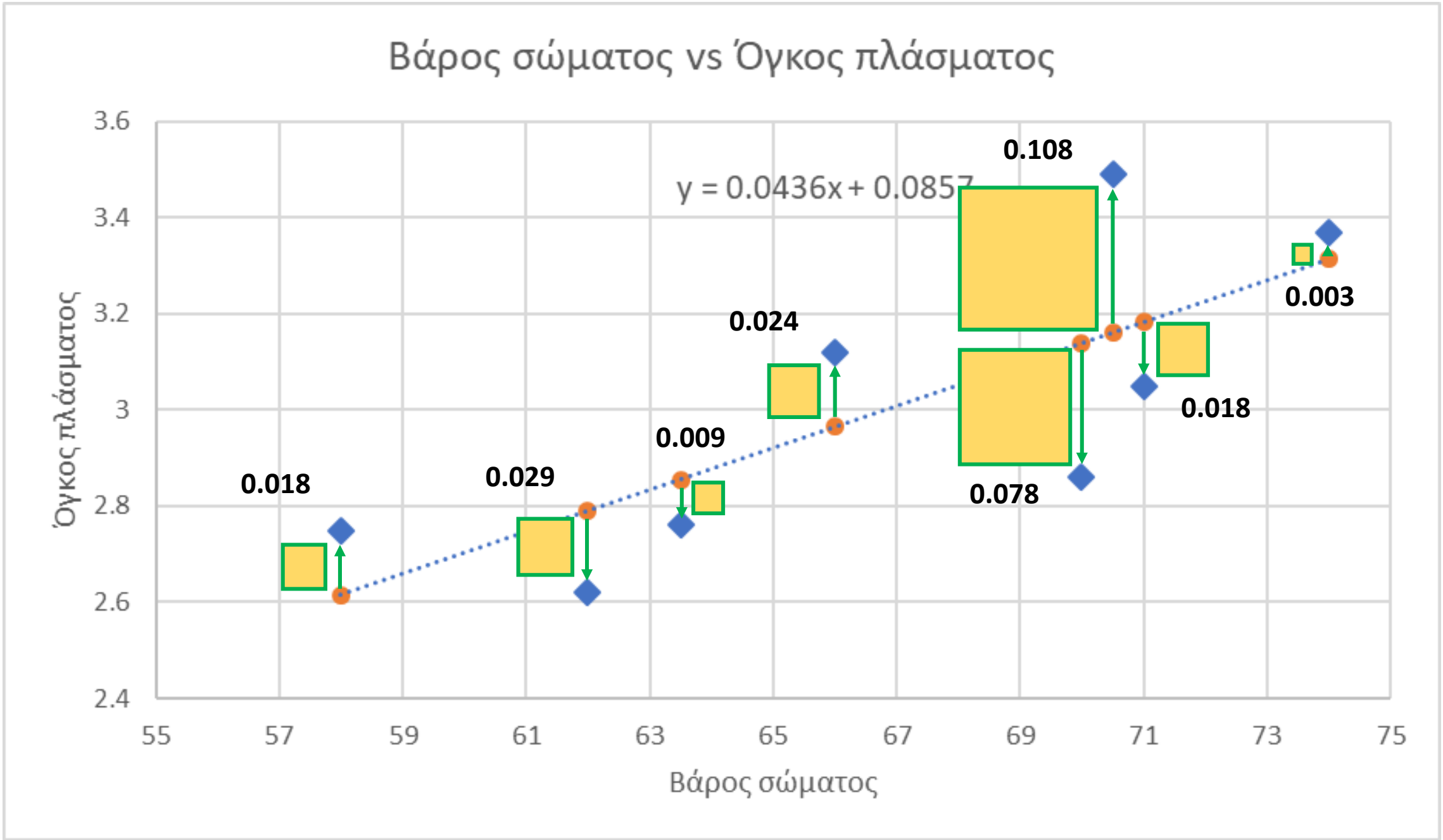

Βάρος σώματος vs Όγκος πλάσματος
y = 0.0436x + 0.0857
Residuals (Υπόλοιπα)
SSE
Όγκος πλάσματος
Βάρος σώματος
3.6
3.4
3.2
3
2.8
2.6
2.4
55
57
59
61
63
65
67
69
71
73
75

Σφάλμα παλινδρόμησης (Residuals / Υπόλοιπα)

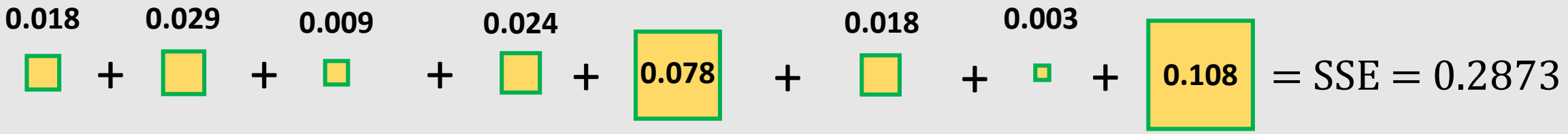
Άτομο	Βάρος	Όγκος πλάσματος σε lt (y)	$\widehat{y_i}$ (προβλεπόμενος όγκος πλάσματος)	Σφάλμα: (παρατηρηθείσα – προβλεπόμενη) $(y_i - \widehat{y_i})$
	x	y		$(y_i - \widehat{y_i})$
1	58	2.75	2.61566	$2.75 - 2.61566 = 0.13434$
2	70	2.86	3.1391	$2.86 - 3.1391 = -0.2791$
3	74	3.37	3.31358	$3.37 - 3.31358 = 0.05642$
4	63.5	2.76	2.85557	$2.76 - 2.85557 = -0.09557$
5	62	2.62	2.79014	$2.62 - 2.79014 = -0.17014$
6	70.5	3.49	3.16091	$3.49 - 3.16091 = 0.32909$
7	71	3.05	3.18272	$3.05 - 3.18272 = -0.13272$
8	66	3.12	2.96462	$3.12 - 2.96462 = 0.15538$
	$\bar{x} = 66.875$	$\bar{y} = 3.0025$		

Τετραγωνισμένο σφάλμα παλινδρόμησης (Residuals / Υπόλοιπα)

Άτομο	Βάρος	Όγκος πλάσματος σε lt (y)	$\hat{y}_i$ (προβλεπόμενος όγκος πλάσματος)	Σφάλμα	Τετραγωνισμένο σφάλμα
	x	y		$(y_i - \hat{y}_i)$	$(y_i - \hat{y}_i)^2$
1	58	2.75	2.61566	0.13434	0.01804724
2	70	2.86	3.1391	−0.2791	0.07789681
3	74	3.37	3.31358	0.05642	0.00318322
4	63.5	2.76	2.85557	−0.09557	0.00913362
5	62	2.62	2.79014	−0.17014	0.02894762
6	70.5	3.49	3.16091	0.32909	0.10830023
7	71	3.05	3.18272	−0.13272	0.0176146
8	66	3.12	2.96462	0.15538	0.02414294
	$\bar{x} = 66.875$	$\bar{y} = 3.0025$		**SSE =**	$\sum = 0.2873$


Βάρος σώματος vs Όγκος πλάσματος
y = 0.0436x + 0.0857
Όγκος πλάσματος
Βάρος σώματος
3.6
3.4
3.2
3
2.8
2.6
2.4
55
57
59
61
63
65
67
69
71
73
75
0.018
0.029
0.009
0.024
0.078
0.108
0.018
0.003

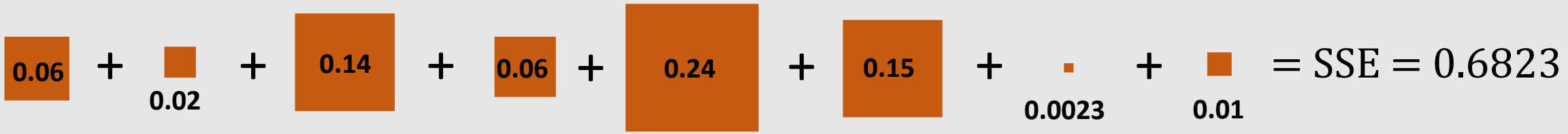
Εξαρτημένη και ανεξάρτητη μεταβλητή (βάρος σώματος και όγκος πλάσματος)

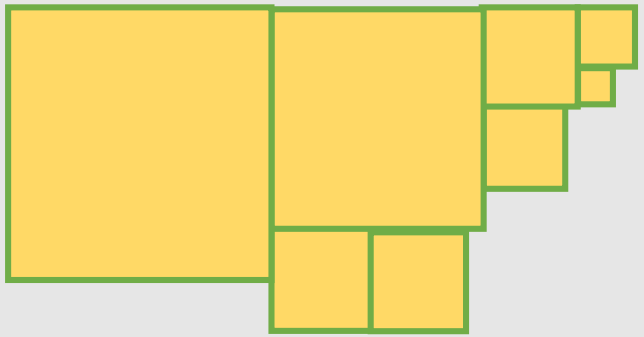


Σύγκριση των SSE των δύο μοντέλων

Μόνο η εξαρτημένη μεταβλητή (όγκος του πλάσματος)

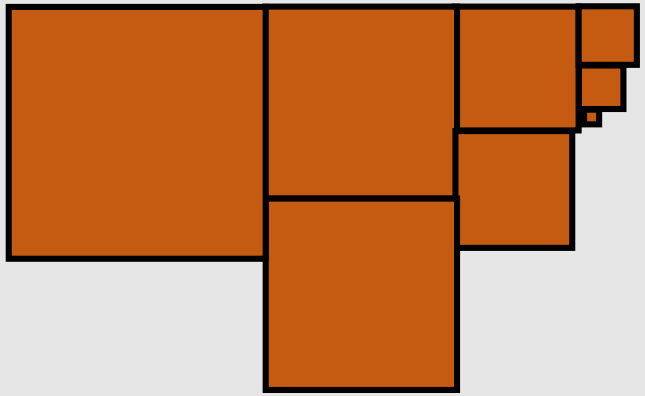
$SSE = SST$





$= 0.2873$

Σύγκριση των SSE των δύο μοντέλων



$= 0.68$

Όταν εκτελέσαμε την παλινδρόμηση, το SSE μειώθηκε από 0.68 σε 0.2873. Αυτό σημαίνει πως 0.2873 από το άθροισμα των τετραγώνων ερμηνεύεται ή κατανέμεται στο ΣΦΑΛΜΑ.

Τα υπόλοιπα 0.3927 που πήγαν;

Τα 0.3927 είναι το άθροισμα των τετραγώνων που οφείλεται στην παλινδρόμηση (SSR).

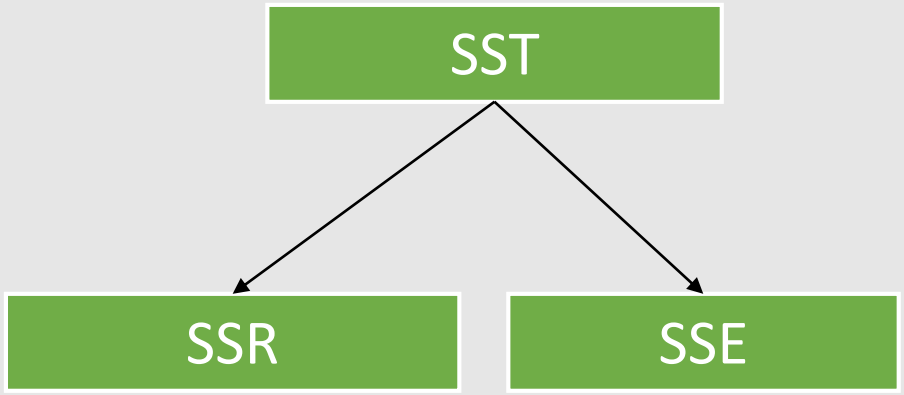
$$SST = SSR + SSE$$

$$0.68 = 0.3927 + 0.2873$$

Συντελεστής προσδιορισμού

Πόσο καλά προσαρμόζει η εκτιμώμενη εξίσωση παλινδρόμησης τα δεδομένα μας;

Σε αυτό το σημείο η παλινδρόμηση αρχίζει να μοιάζει με την ANOVA. Το συνολικό άθροισμα τετραγώνων (SST) διαχωρίζεται σε SSE και SSR



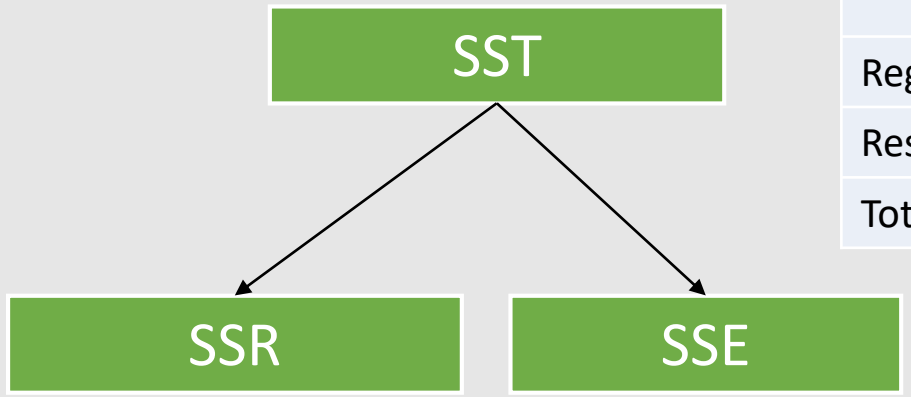
Αν το SSR είναι μεγάλο, τότε χρησιμοποιεί ένα μεγάλο μέρος από το SST, οπότε το SSE είναι μικρότερο σε σχέση με το SST. Ο συντελεστής προσδιορισμού ποσοτικοποιεί αυτή την αναλογία ως ποσοστό.

$$\text{Συντελεστής προσδιορισμού} = r^2 = \frac{SSR}{SST}$$

Συντελεστής προσδιορισμού

Πόσο καλά προσαρμόζει η εκτιμώμενη εξίσωση παλινδρόμησης τα δεδομένα μας;

Σε αυτό το σημείο η παλινδρόμηση αρχίζει να μοιάζει με την ANOVA. Το συνολικό άθροισμα τετραγώνων (SST) διαχωρίζεται σε SSE και SSR



ANOVA					
	df	SS	MS	F	Significance F
Regression	1	0.390684388	0.39068439	8.160065927	0.028930913
Residual	6	0.287265612	0.0478776		
Total	7	0.67795			

Ερμηνεία του Συντελεστής προσδιορισμού r^2

$$\text{Συντελεστής προσδιορισμού} = r^2 = \frac{SSR}{SST}$$

ΚΑΛΗ ΠΡΟΣΑΡΜΟΓΗ!

$$\text{Συντελεστής προσδιορισμού} = r^2 = \frac{0.3927}{0.68}$$

$$\text{Συντελεστής προσδιορισμού} = r^2 = 0.5775 \text{ ή } 57.75\%$$

Συμπεραίνουμε πως 57.75% από το συνολικό άθροισμα τετραγώνων μπορεί να εξηγηθεί χρησιμοποιώντας την εκτιμώμενη εξίσωση παλινδρόμησης για να προβλέψει τον όγκο πλάσματος. Το υπόλοιπο, 42.25%, είναι το σφάλμα.

3 τετραγωνισμένες διαφορές