Statistics for Economics and Business

Γρηγόριος Καρράς

Ανάλυση Πολλαπλής Παλινδρόμησης

Διάλεξη – 4 Μαΐου 2021

Στην αιχμή της επιστήμης (& της τέχνης)

- Μετά την μελέτη της απλής γραμμικής παλινδρόμησης, που είναι η εύρεση μιας ***καλύτερης γραμμής*** μέσω ενός δείγματος δεδομένων για μια εξαρτημένη μεταβλητή Y η οποία εξαρτάται από μια *ανεξάρτητη*, *ερμηνευτική* ή *επεξηγηματική*, μεταβλητή x …

- Αντιλαμβανόμαστε ότι ο κόσμος είναι ένας πολύπλοκος χώρος όπου υπάρχουν πολλές ερμηνευτικές μεταβλητές από επηρεάζουν μια εξαρτημένη μεταβλητή.

- Τι κάνουμε όταν μια **μεταβλητή Y** εξαρτάται σε δύο ή περισσότερες **ερμηνευτικές μεταβλητές (x) ?**

  Πως «*μοντελοποιούμε*» αυτό το φαινόμενο?
  Πως μπορούμε να χρησιμοποιήσουμε γραμμική άλγεβρα να βρούμε την καλύτερη γραμμή (**best fit**?) Πως μπορούμε να ερμηνεύσουμε τα αποτελέσματα ενός τέτοιου εκτιμημένου μοντέλου ?

Περιεχόμενα

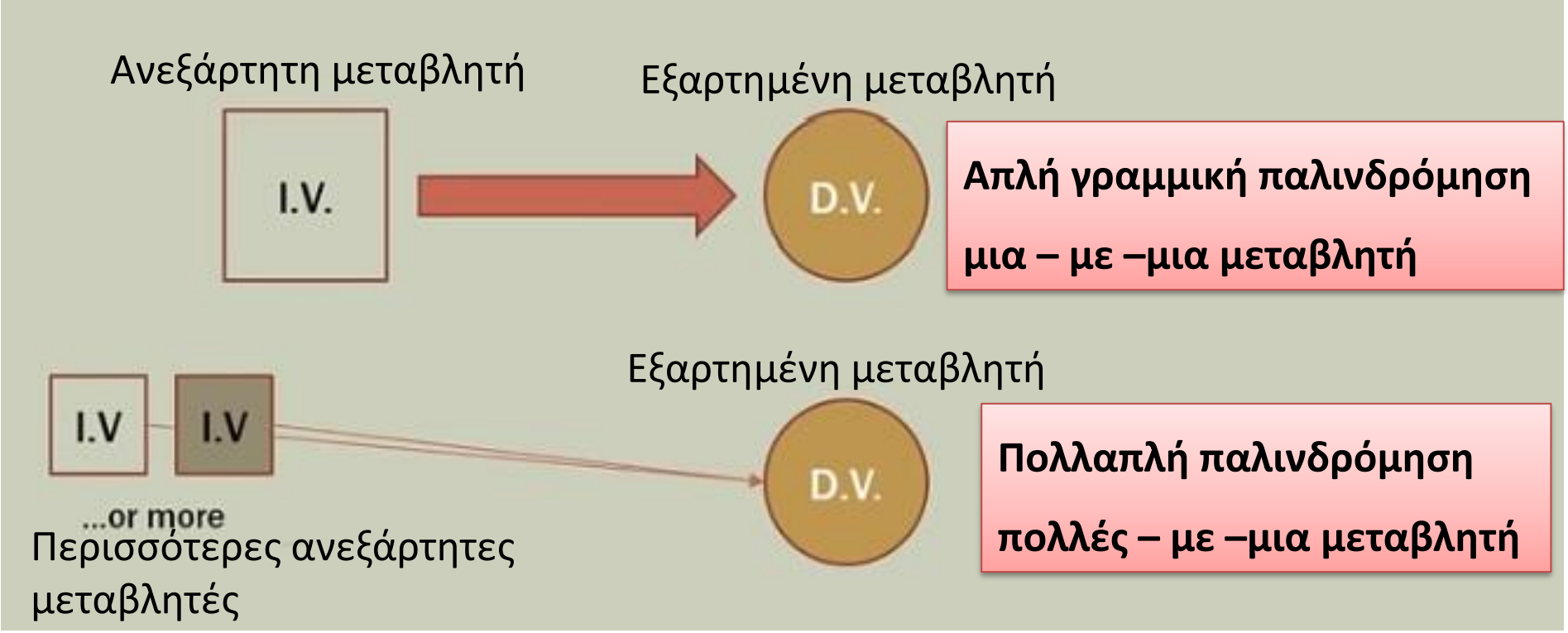
1. Εισαγωγή
2. Το μοντέλο της πολλαπλής παλινδρόμησης
3. Γραφική Αναπαράσταση
4. Μια μελέτη περίπτωσης - ερμηνεία

Παράδειγμα εφαρμογής πολλαπλής παλινδρόμησης:

ΕΧ 1 – Εμπορία – Διανομή κατεψυγμένων γλυκών επιδορπίων

Μία εταιρεία - διανομέας κατεψυγμένων γλυκών επιδορπίων θέλει να αναπτύξει μια νέα μάρκα (brand). Πριν κάνει αυτή την επένδυση ο Γενικός Διευθυντής θέλει να εκτιμήσει τους παράγοντες που επηρεάζουν την ζήτηση για κατεψυγμένα γλυκά επιδόρπια. Η εταιρία συγκεντρώνει δεδομένα πωλήσεων τέτοιων γλυκών (πωλήσεις μονάδων - τεμαχίων ανά εβδομάδα), την τιμή κάθε μονάδας σε € και την διαφημιστική δαπάνη (ή διαφημιστική επένδυση σε 100€). Τα δεδομένα συγκεντρώθηκαν για 15 εβδομάδες και με βάση αυτά η εταιρεία θα ήθελε να εκτιμήσει (υπολογίσει) τον συνολικό αριθμό γλυκών που πωλούνται ανά εβδομάδα βασισμένο σε δύο παράγοντες 1) την τιμή κάθε μονάδας γλυκού και 2) την διαφημιστική επένδυση (που έγινε στην ίδια εβδομάδα)

Η πολλαπλή παλινδρόμηση είναι η επέκταση της απλής παλινδρόμησης



Έχοντας περισσότερες ανεξάρτητες μεταβλητές περιπλέκει κάπως τα πράγματα … Έτσι πρέπει να κάνουμε **νέες σκέψεις - προβληματισμούς**:

Νέες σκέψεις / προβληματισμοί (1/2)

- Η προσθήκη περισσότερων ανεξάρτητων μεταβλητών σε ένα μοντέλο πολλαπλής παλινδρόμησης δεν σημαίνει απαραίτητα ότι το μοντέλο θα είναι καλύτερο ή ότι θα κάνει καλύτερες προβλέψεις – στην πράξη μπορεί να κάνει τα πράγματα χειρότερα. Αυτό λέγεται *"υπερβολική τοποθέτηση" / "υπερβολικό ταίριασμα"* ***(overfitting***)
- Η προσθήκη περισσότερων ανεξάρτητων μεταβλητών δημιουργεί περισσότερες σχέσεις αναμεταξύ τους. Έτσι όχι μόνο οι ανεξάρτητες μεταβλητές δυνητικά σχετίζονται με την εξαρτημένη μεταβλητή αλλά επίσης δυνητικά *συσχετίζονται αναμεταξύ τους* (αλληλοσχετίζονται) Αυτό λέγεται ***πολυσυγραμμικότητα***
- Το <u>ιδανικό</u> είναι όλες οι ανεξάρτητες να *σχετίζονται με την εξαρτημένη αλλά ΟΧΙ αναμεταξύ τους*

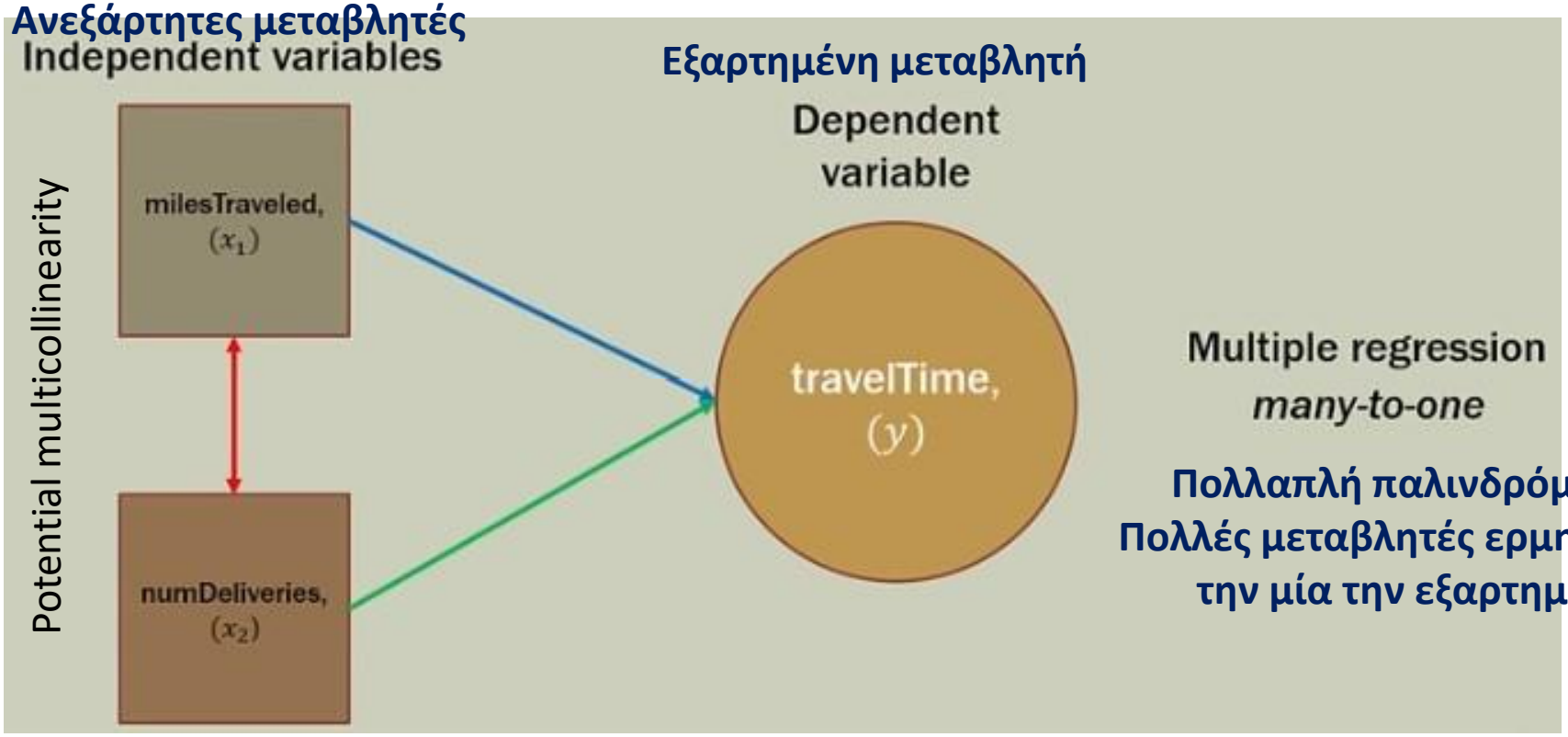
Νέες σκέψεις / προβληματισμοί (1/2)

Εξαιτίας της πολυσυγραμμικότητας και του overfitting υπάρχει μια σημαντική προεργασία που πρέπει να προηγηθεί πριν την εκτέλεση της πολλαπλής παλινδρόμησης

- Συσχετίσεις μεταβλητών
- Διαγράμματα διασποράς
- Απλές παλινδρομήσεις

Ο υπολογισμός του μοντέλου πολλαπλής παλινδρόμησης είναι το τελευταίο βήμα

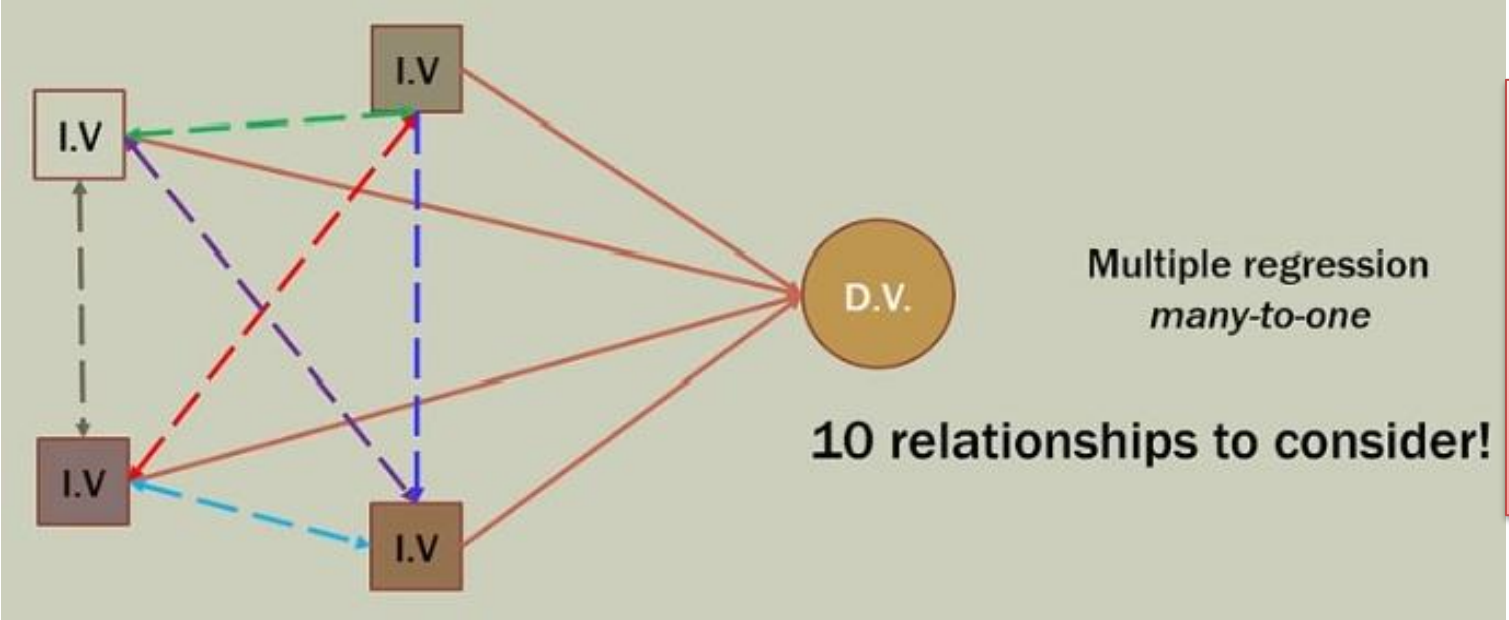
Ανάλυση πολλαπλής Παλινδρόμησης



Δεν έχουμε μόνο 2 σχέσεις ανάμεσα στις μεταβλητές (ανεξάρτητες – εξαρτημένες)

Ίσως «*ανακαλύψουμε*» συσχέτιση ανάμεσα στις ανεξάρτητες: πρέπει να το ελέγξουμε αυτό για να αποφύγουμε την πολυσυγγραμικότητα – πρέπει οι ανεξάρτητες (ερμηνευτικές) μεταβλητές να είναι ***ανεξάρτητες*** αναμεταξύ τους

Ανάλυση πολλαπλής παλινδρόμησης:
περιλαμβάνει πολλές σχέσεις για να μελετηθούν



Με 4 ανεξάρτητες μεταβλητές οι σχέσεις προς μελέτη ανέρχονται σε 10 !

- Με την προσθήκη κάθε επιπλέον ανεξάρτητης μεταβλητής, οι σχέσεις μπορεί να γίνουν πάρα πολλές και πολύπλοκες.
- Η τεχνική της πολλαπλής παλινδρόμησης αποφασίζει ποιες ανεξάρτητες μεταβλητές είναι σημαντικές και ποιες όχι.
- Ορισμένες ανεξάρτητες μεταβλητές, είναι καλύτερες στο να προβλέπουν την εξαρτημένη μεταβλητή από άλλες.
- Μερικές ανεξάρτητες μεταβλητές δεν προσφέρουν τίποτα το σημαντικό.

Ανάλυση Πολλαπλής Παλινδρόμησης:

περιλαμβάνει το χειρισμό πολλών σχέσεων ανάμεσα στις μεταβλητές

Το ιδανικό είναι όλες οι ΑΝΕΞΑΡΤΗΤΕΣ ΜΕΤΑΒΛΗΤΕΣ να συσχετίζονται με την εξαρτημένη

αλλά ΝΑ ΜΗΝ (ΑΛΛΗΛΟ)ΣΥΣΧΕΤΙΖΟΝΤΑΙ ΑΝΑΜΕΣΑ ΤΟΥΣ

Το μοντέλο

Ιδέα: Εξέταση της γραμμικής σχέσης ανάμεσα σε 1 εξαρτημένη μεταβλητή (Y) & 2 ή περισσότερες ανεξάρτητες ή ερμηνευτικές μεταβλητές (X_i)

Το μοντέλο της πολλαπλής παλινδρόμησης με k ανεξάρτητες μεταβλητές:

Ο σταθερός όρος του πληθυσμού Y (intercept)

Οι «κλίσεις» (τάσεις) του πληθυσμού λέγονται συντελεστές (coefficients)

Τυχαίο σφάλμα

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik} + \varepsilon_i$$

Οι συντελεστές του μοντέλου πολλαπλής παλινδρόμησης εκτιμώνται – υπολογίζονται χρησιμοποιώντας ***δείγματα δεδομένων***

Το μοντέλο της πολλαπλής παλινδρόμησης με k ανεξάρτητες μεταβλητές:

Εκτίμηση του σταθερού όρου

Εκτιμήσεις των συντελεστών (coefficients) του πληθυσμού

$$\hat{Y}_{i=} b_0 + b_1 X_{i1} + b_2 X_{i2} + \ldots + b_k X_{ik}$$

Εκτίμηση ή πρόβλεψη της εξαρτημένης μεταβλητής Y

Μοντέλο πολλαπλής παλινδρόμησης

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_p x_p + \epsilon$$

linear parameters — error

Γραμμικές παράμετροι

σφάλμα / κατάλοιπο

Μοντέλο πολλαπλής παλινδρόμησης

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_p x_p$$

error term assumed to be zero

<u>Παραδοχή:</u> Το σφάλμα / κατάλοιπο ***είναι μηδενικό***

Εκτιμημένη εξίσωση πολλαπλής παλινδρόμησης

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots b_p x_p$$

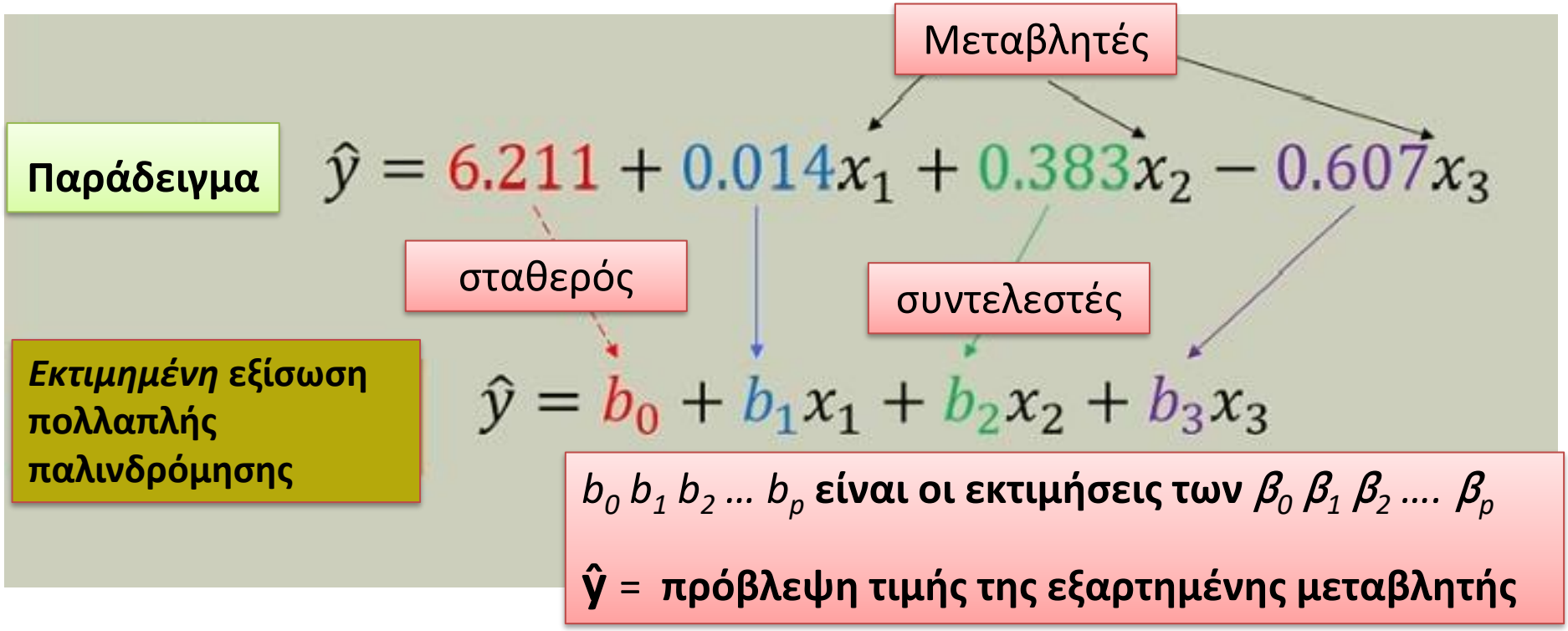
$b_0, b_1, b_2, \ldots b_p$ are the estimates of $\beta_0, \beta_1, \beta_2, \ldots \beta_p$

$\hat{y}$ = predicted value of the dependent variable

$b_0\ b_1\ b_2 \ldots b_p$ **είναι οι εκτιμήσεις των** $\beta_0\ \beta_1\ \beta_2 \ldots \beta_p$

$\hat{y}$ = **πρόβλεψη τιμής της εξαρτημένης μεταβλητής**

Παράδειγμα με αριθμούς



Ερμηνεία των συντελεστών στο μοντέλο της πολλαπλής παλινδρόμησης

Ένα αριθμητικό παράδειγμα

$$\hat{y} = 27 + 9x_1 + 12x_2$$

x_1 = επενδυμένο κεφάλαιο (€000s)

x_2 = δαπάνες marketing (προώθησης) (€000s)

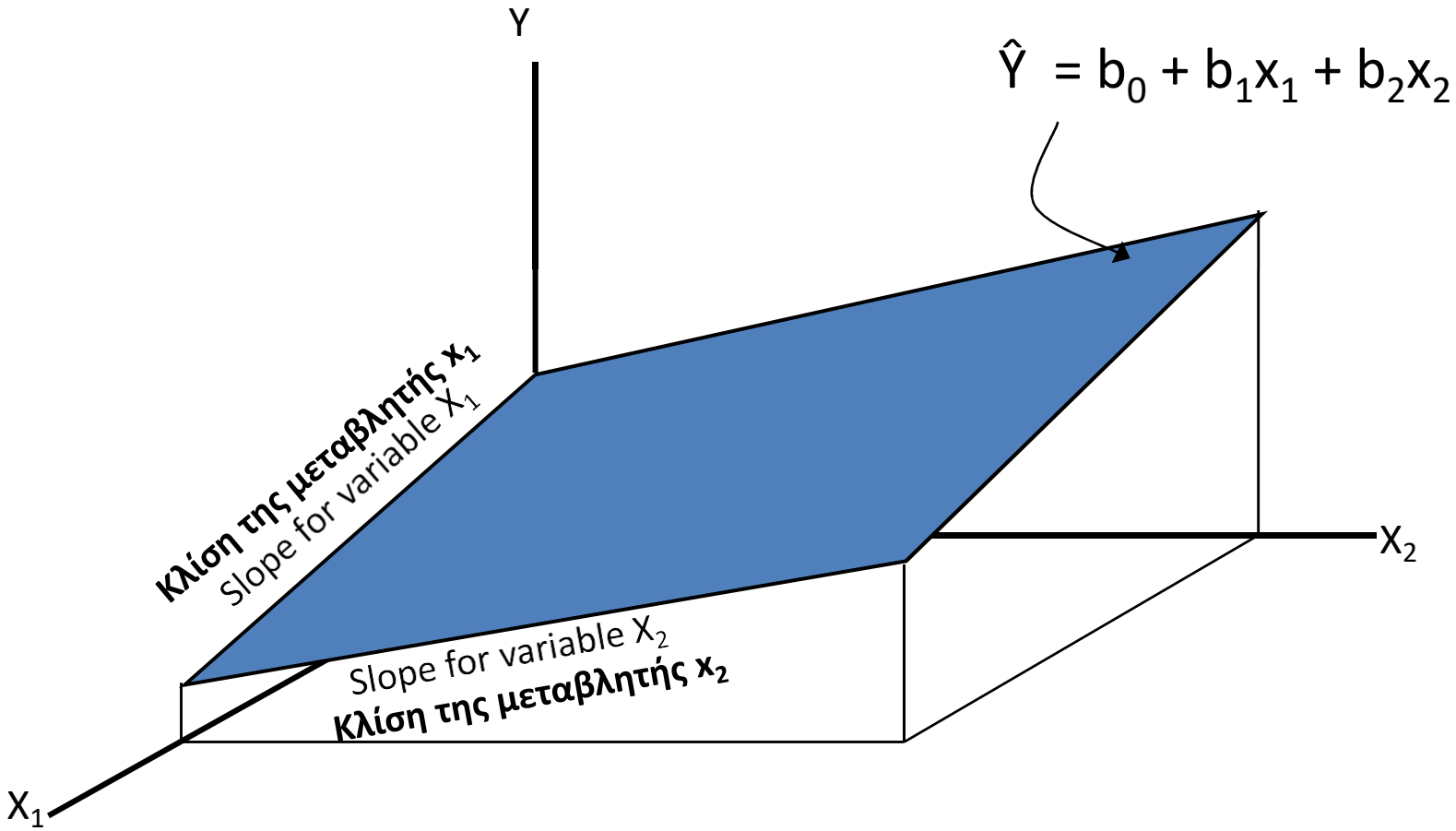
$\hat{y}$ = πρόβλεψη πωλήσεων (€000s)

Στην πολλαπλή παλινδρόμηση κάθε συντελεστής μιας x ερμηνεύεται ως η *εκτιμώμενη* μεταβολή της y που αντιστοιχεί (*προκαλείται*) από μεταβολή μιας μονάδας αυτής της x - όλες οι άλλες παραμένουν σταθερές

- Έτσι σε αυτό το παράδειγμα μια αύξηση € 1000 στο επενδυμένο κεφάλαιο x_1 (προβλέπεται) να προκαλέσει μια αύξηση € 9.000 στην y δηλαδή στις πωλήσεις με τις δαπάνες marketing x_2 να παραμένουν σταθερές
- Και μια αύξηση € 1000 τις δαπάνες marketing x_2 (προβλέπεται) να προκαλέσει μια αύξηση € 12.000 στην y δηλαδή στις πωλήσεις με το επενδυμένο κεφάλαιο x_1 να παραμένει σταθερό

Γραφική Αναπαράσταση

Μοντέλο με δύο μεταβλητές



Μια μελέτη περίπτωσης

- Ένας διανομέας κατεψυγμένων γλυκών επιδορπίων θέλει να εκτιμήσει παράγοντες που επηρεάζουν την ζήτηση

– **Εξαρτημένη μεταβλητή: Πωλήσεις επιδορπίων (μονάδες ανά εβδομάδα)**

– Ανεξάρτητες μεταβλητές:
  - Τιμή (€)
  - Διαφήμιση (€100's)

Δεδομένα συλλέχθηκαν σε διάστημα 15 εβδομάδων

Ένα μοντέλο πολλαπλής παλινδρόμησης:

$$\text{Πωλήσεις} = b_0 + b_1(\text{Τιμή}) + b_2(\text{Διαφημιστική δαπάνη})$$

Εβδομάδα	Πωλήσεις επιδορπίων	Τιμή (€)	Διαφημιστική δαπάνη ($100s)
1	350	5.50	3.3
2	460	7.50	3.3
3	350	8.00	3.0
4	430	8.00	4.5
5	350	6.80	3.0
6	380	7.50	4.0
7	430	4.50	3.0
8	470	6.40	3.7
9	450	7.00	3.5
10	490	5.00	4.0
11	340	7.20	3.5
12	300	7.90	3.2
13	440	5.90	4.0
14	450	5.00	3.5
15	300	7.00	2.7

ΈΞΟΔΟΣ ΣΥΜΠΕΡΑΣΜΑΤΟΣ

Στατιστικά παλινδρόμησης	
Πολλαπλό R	0,722134292
R Τετράγωνο R^2	0,521477936
Προσαρμοσμένο R Τετράγ	0,441724259
Τυπικό σφάλμα	47,46341263
Μέγεθος δείγματος	15

Η ΕΝΤΟΛΗ ΣΤΟ EXCEL ΕΙΙΝΑΙ:
> DATA
> DATA ANALYSIS
> REGRESSION
ΚΑΙ ΚΑΘΟΡΙΖΕΙΣ:
ΠΟΙΑ ΠΕΡΙΟΧΗ ΕΧΕΙ ΤΙΣ ΤΙΜΕΣ ΤΗΣ ΕΞΑΡΤΗΜΕΝΗΣ Υ
ΚΑΙ ΠΟΙΑ ΠΕΡΙΟΧΗ ΕΧΕΙ ΤΗΝ ΑΝΕΞΑΡΤΗΤΗ΄Η ΤΙΣ ΑΝΕΞΑΡΤΗΤΕΣ Χ

ΑΝΑΛΥΣΗ ΔΙΑΚΥΜΑΝΣΗΣ

	βαθμοί ελευθερίας	*SS*	*MS*	*F*	*Σημαντικότητα F*
Παλινδρόμηση	2	29460,02687	14730,01343	6,53860679	0,012006372
Υπόλοιπο	12	27033,30647	2252,775539		
Σύνολο	14	56493,33333			

	Συντελεστές	*Τυπικό σφάλμα*	*t*	*τιμή-P*	*Κατώτερο 95%*	*Υψηλότερο 95%*	*Κατώτερο 95,0%*	*Υψηλότερο 95,0%*
Τεταγμένη επί την αρχή	306,5261933	114,2538935	2,682851182	0,01993159	57,58834426	555,4640423	57,58834426	555,4640423
Μεταβλητή X 1	-24,97508952	10,83212512	-2,305650022	0,03978846	-48,5762627	-1,373916335	-48,5762627	-1,373916335
Μεταβλητή X 2	74,13095749	25,96731792	2,854779139	0,01449363	17,55303206	130,7088829	17,55303206	130,7088829

ΈΞΟΔΟΣ ΣΥΜΠΕΡΑΣΜΑΤΟΣ

Στατιστικά παλινδρόμησης	
Πολλαπλό R	0,722134292
R Τετράγωνο	0,521477936
Προσαρμοσμένο R Τετράγωνο	0,441724259
Τυπικό σφάλμα	47,46341263
Μέγεθος δείγματος	15

Το εκτιμημένο μοντέλο είναι

Πωλήσεις y = 306,52 -24,975 Τιμή + 74,13 Διαφ Δαπάνη

> Με αυξηση της τιμής κατά 1€ οι πωλήσεις μειώνονται κατά 25 μονάδες (περίπου)

> Με αυξηση της διαφημιστικής δαπάνης κατά 1 μονάδα (των 100€) οι πωλήσεις αυξάνονται κατά 74 μονάδες (περίπου)

ΑΝΑΛΥΣΗ ΔΙΑΚΥΜΑΝΣΗΣ

	βαθμοί ελευθερίας	*SS*	*MS*	*F*	*Σημαντικότητα F*
Παλινδρόμηση	2	29460,0269	14730,01343	6,53860679	0,012006372
Υπόλοιπο	12	27033,3065	2252,775539		
Σύνολο	14	56493,3333			

	Συντελεστές	*υπικό σφάλμ*	*t*	*τιμή-P*	*Κατώτερο 95%*	*Υψηλότερο 95%*	*Κατώτερο 95,0%*	*Υψηλότερο 95,0%*
Τεταγμένη επί την αρχή	306,5261933	114,253894	2,682851182	0,01993159	57,58834426	555,4640423	57,58834426	555,4640423
Μεταβλητή X 1	-24,97508952	10,8321251	-2,305650022	0,03978846	-48,5762627	-1,373916335	-48,5762627	-1,373916335
Μεταβλητή X 2	74,13095749	25,9673179	2,854779139	0,01449363	17,55303206	130,7088829	17,55303206	130,7088829

Πρόβλεψη πωλήσεων: για μια εβδομάδα στην οποία η τιμή πώλησης του κάθε επιδορπίου θα είναι € 5,50 και η διαφημιστική δαπάνη θα είναι € 350:

Έχουμε

$\hat{Y} = b_0 + b_1x_1 + b_2x_2 = 306,52 - 24,975$ τιμή (5,50) + 74,13 διαφήμιση (3,5) =

$= 306,52 - 24,975 * 5,50 + 74,13 * 3,5$

$= 306,52 - 137,363 + 259,455 =$ **428,612**

Οι προβλεπόμενες πωλήσεις επιδορπίων είναι 428,612 μονάδες

Σημείωση η διαφημιστική δαπάνη είναι σε €100's, άρα €350 σημαίνει $X_2 = 350/100 = 3,5$

Ο συντελεστής προσδιορισμού R^2 :

Δείχνει το ποσοστό της συνολικής διακύμανσης της εξαρτημένης Y (πωλήσεις επιδορπίων) που ερμηνεύεται από ***όλες*** τις ερμηνευτικές μεταβλητές X συνολικά και συλλογικά

$$r^2 = \frac{SSR}{SST} = \frac{\text{regression sum of squares}}{\text{total sum of squares}}$$

R^2 = SSR Regression sum of squares / SST Total sum of squares

Εδώ είναι $R^2 = 0,524$ άρα το 52,4 % της διακύμανσης της Y (πωλήσεις) ερμηνεύεται από την παλινδρόμηση αυτή

ΈΞΟΔΟΣ ΣΥΜΠΕΡΑΣΜΑΤΟΣ

Στατιστικά παλινδρόμησης	
Πολλαπλό R	0,722134292
R Τετράγωνο	0,521477936
Προσαρμοσμένο R Τετράγ	0,441724259
Τυπικό σφάλμα	47,46341263
Μέγεθος δείγματος	15

R^2 = SSR Regression sum of squares / SST Total sum of squares

Εδώ είναι R^2 = 0,524 άρα το 52,4 % της διακύμανσης της Y (πωλήσεις) ερμηνεύεται από την παλινδρόμηση αυτή

ΑΝΑΛΥΣΗ ΔΙΑΚΥΜΑΝΣΗΣ

	βαθμοί ελευθερίας	*SS*	*MS*	*F*	*Σημαντικότητα F*
Παλινδρόμηση	2	29460,02687	14730,01343	6,53860679	0,012006372
Υπόλοιπο	12	27033,30647	2252,775539		
Σύνολο	14	56493,33333			

	Συντελεστές	*Τυπικό σφάλμα*	*t*	*τιμή-P*	*Κατώτερο 95%*	*Υψηλότερο 95%*	*Κατώτερο 95,0%*	*Υψηλότερο 95,0%*
Τεταγμένη επί την αρχή	306,5261933	114,2538935	2,682851182	0,01993159	57,58834426	555,4640423	57,58834426	555,4640423
Μεταβλητή X 1	-24,97508952	10,83212512	-2,305650022	0,03978846	-48,5762627	-1,373916335	-48,5762627	-1,373916335
Μεταβλητή X 2	74,13095749	25,96731792	2,854779139	0,01449363	17,55303206	130,7088829	17,55303206	130,7088829

Προσαρμοσμένο R^2_{adj}

Δείχνει την αναλογία της διακύμανσης της Υ *που ερμηνεύεται από όλες τις μεταβλητές* Χ «διορθωμένος» για το πλήθος των μεταβλητών Χ που χρησιμοποιήθηκαν και για το μέγεθος του δείγματος

$$r^2_{adj} = 1 - \left[(1 - r^2) \left(\frac{n-1}{n-k-1} \right) \right]$$

(όπου n = μέγεθος δείγματος, k = πλήθος ανεξάρτητων μεταβλητών)

- «Τιμωρεί» υπερβολική χρήση μη σημαντικών ανεξάρτητων μεταβλητών
- Πάντα μικρότερο από το απλό r^2
- Κατάλληλο για σύγκριση ανάμεσα σε διαφορετικά μοντέλα

Στο μοντέλο μας με τα επιδόρπια το R^2_{adj} = 0,44177 δηλαδή μόνο ***το 44,17 % της διακύμανσης της y ερμηνεύεται από τις x***

- F Test για την συνολική σημαντικότητα του μοντέλου

Δείχνει εάν υπάρχει γραμμική σχέση μεταξύ όλων των X μεταβλητών συλλογικά και της Y

Χρήση της στατιστικής F-test

Έλεγχος υποθέσεων:

Έλεγχος υπόθεσης F - test

$\mathbf{H_0}$: $\beta_1 = \beta_2 = \cdots = \beta_k = 0$ (Δεν υπάρχει γραμμική σχέση)

$\mathbf{H_1}$: at least one $\beta_i \neq 0$ (τουλάχιστον μια ανεξάρτητη μεταβλητή x επηρεάζει την εξαρτημένη Y)

$$F_{STAT} = \frac{MSR}{MSE} = \frac{\frac{SSR}{k}}{\frac{SSE}{n-k-1}}$$

$$F_{STAT} = \frac{MSR}{MSE} = \frac{14730.0}{2252.8} = 6.5386$$

ΈΞΟΔΟΣ ΣΥΜΠΕΡΑΣΜΑΤΟΣ

Στατιστικά παλινδρόμησης	
Πολλαπλό R	0,722134292
R Τετράγωνο	0,521477936
Προσαρμοσμένο R Τετράγ	0,441724259
Τυπικό σφάλμα	47,46341263
Μέγεθος δείγματος	15

ΑΝΑΛΥΣΗ ΔΙΑΚΥΜΑΝΣΗΣ

	βαθμοί ελευθερίας	*SS*	*MS*	*F*	*Σημαντικότητα F*
Παλινδρόμηση	2	29460,02687	14730,01343	6,53860679	0,012006372
Υπόλοιπο	12	27033,30647	2252,775539		
Σύνολο	14	56493,33333			

P-value for the F Test

	Συντελεστές	*Τυπικό σφάλμα*	*t*	*τιμή-P*	*Κατώτερο 95%*	*Υψηλότερο 95%*	*Κατώτερο 95,0%*	*Υψηλότερο 95,0%*
Τεταγμένη επί την αρχή	306,5261933	114,2538935	2,682851182	0,01993159	57,58834426	555,4640423	57,58834426	555,4640423
Μεταβλητή X 1	-24,97508952	10,83212512	-2,305650022	0,03978846	-48,5762627	-1,373916335	-48,5762627	-1,373916335
Μεταβλητή X 2	74,13095749	25,96731792	2,854779139	0,01449363	17,55303206	130,7088829	17,55303206	130,7088829

H_0: $\beta_1 = \beta_2 = \cdots = \beta_k = 0$ (Δεν υπάρχει γραμμική σχέση)

H_1: at least one $\beta_i \neq 0$ (τουλάχιστον μια ανεξάρτητη μεταβλητή x επηρεάζει την εξαρτημένη Y)

P-value = 0,012 < 0,05 άρα απορρίπτουμε την **H_0** και αποδεχόμαστε την **H_1** δηλ. υπάρχει γραμμική σχέση της **y** με *τουλάχιστον μια ανεξάρτητη* **x**

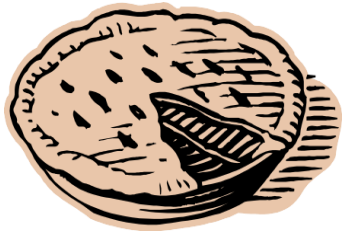
- Χρήση των t tests των κλίσεων (συντελεστών) *για κάθε μια μεταβλητή ξεχωριστά*

Δείχνουν εάν υπάρχει γραμμική σχέση ανάμεσα στην κάθε ανεξάρτητη μεταβλητή X_j και Y με σταθερές τις άλλες μεταβλητές X

Έλεγχος υποθέσεων για κάθε μια μεταβλητή ξεχωριστά

– H_0: $\beta_j = 0$ (δεν υπάρχει γραμμική σχέση)

– H_1: $\beta_j \neq 0$ (υπάρχει γραμμική σχέση μεταξύ X_j και Y)



4. Μια μελέτη περίπτωσης

ΈΞΟΔΟΣ ΣΥΜΠΕΡΑΣΜΑΤΟΣ

Στατιστικά παλινδρόμησης	
Πολλαπλό R	0,722134292
R Τετράγωνο	0,521477936
Προσαρμοσμένο R Τετράγωνο	0,441724259
Τυπικό σφάλμα	47,46341263
Μέγεθος δείγματος	15

Το εκτιμημένο μοντέλο είναι

Πωλήσεις y = 306,52 -24,975 Τιμή + 74,13 Διαφ Δαπάνη

> Με αυξηση της τιμής κατά 1€ οι πωλήσεις μειώνονται κατά 25 μονάδες (περίπου)

> Με αυξηση της διαφημιστικής δαπάνης κατά 1 μονάδα (των 100€) οι πωλήσεις αυξάνονται κατά 74 μονάδες (περίπου)

ΑΝΑΛΥΣΗ ΔΙΑΚΥΜΑΝΣΗΣ

	βαθμοί ελευθερίας	*SS*	*MS*	*F*	*Σημαντικότητα F*
Παλινδρόμηση	2	29460,0269	14730,01343	6,53860679	0,012006372
Υπόλοιπο	12	27033,3065	2252,775539		
Σύνολο	14	56493,3333			

$$t_{STAT} = \frac{b_j - 0}{S_{b_j}}$$

	Συντελεστές	*υπικό σφάλμ*	*t*	*τιμή-P*	*Κατώτερο 95%*	*Υψηλότερο 95%*	*Κατώτερο 95,0%*	*Υψηλότερο 95,0%*
Τεταγμένη επί την αρχή	306,5261933	114,253894	2,682851182	0,01993159	57,58834426	555,4640423	57,58834426	555,4640423
Μεταβλητή X 1	-24,97508952	10,8321251	-2,305650022	0,03978846	-48,5762627	-1,373916335	-48,5762627	-1,373916335
Μεταβλητή X 2	74,13095749	25,9673179	2,854779139	0,01449363	17,55303206	130,7088829	17,55303206	130,7088829

Έλεγχος υποθέσεων για κάθε μια μεταβλητή ξεχωριστά

- **H_0**: $\beta_j = 0$ (δεν υπάρχει γραμμική σχέση)
- **H_1**: $\beta_j \neq 0$ (υπάρχει γραμμική σχέση μεταξύ X_j και Y)

P-value for the t test
Όταν είναι < 0,05 τότε απορρίπτουμε **H_0** την και αποδεχόμαστε την **H_1**
δηλαδή οι συντελεστές μας είναι στατιστικά σημαντικοί

Σφάλματα (κατάλοιπα) από το μοντέλο παλινδρόμησης:

$$e_i = (Y_i - \hat{Y}_i)$$

Παραδοχές:

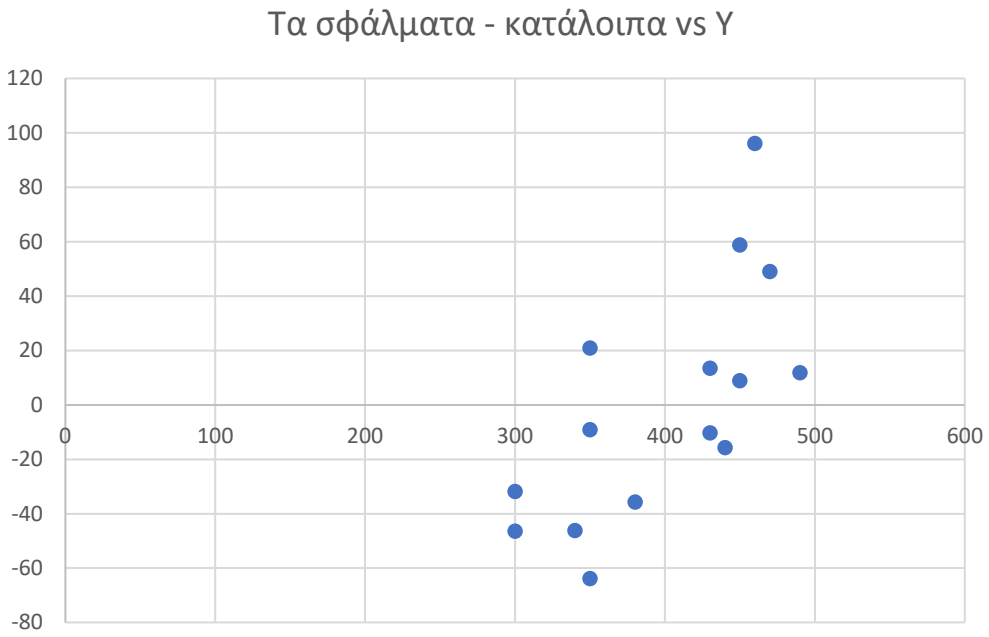
- Ανεξαρτησία των σφαλμάτων
  - Οι τιμές των σφαλμάτων είναι στατιστικά ανεξάρτητες
- Κανονικότητα των σφαλμάτων
  - Οι τιμές των σφαλμάτων ακολουθούν την κανονική κατανομή για οποιαδήποτε δεδομένα - τιμές των X
- Ίση (σταθερή) διακύμανση (*Ομοσκεδαστικότητα* (Homoscedasticity)
  - Τα σφάλματα έχουν σταθερή διακύμανση δηλ. η κατανομή πιθανότητας τους έχει σταθερή διακύμανση

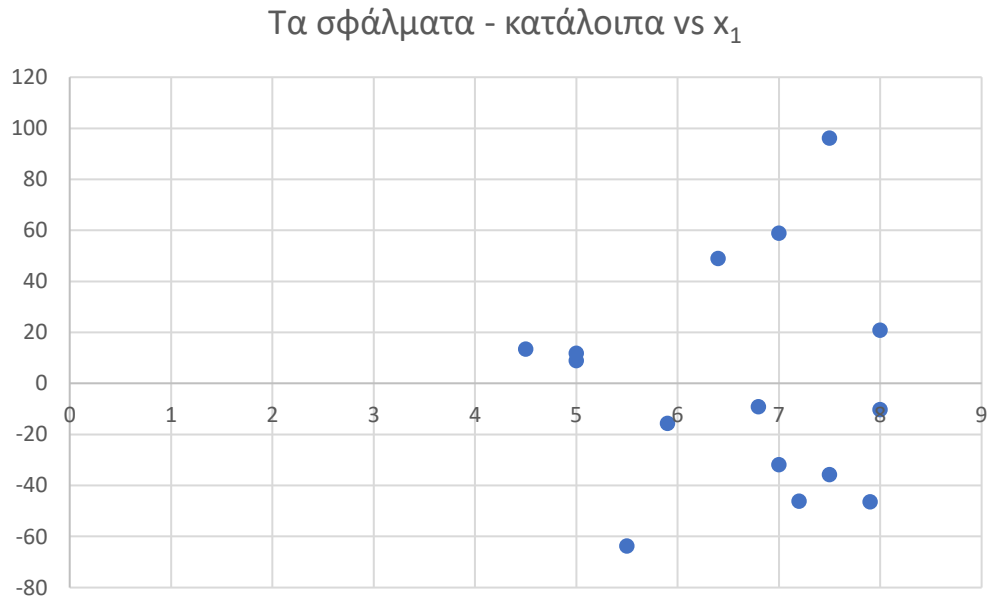
- Τα ακόλουθα διαγράμματα διασποράς χρησιμοποιούνται στην πολλαπλή παλινδρόμηση:
  - κατάλοιπα vs. $\hat{y}_j$
  - κατάλοιπα vs. X_{1i}
  - κατάλοιπα vs. X_{2i}
  - κατάλοιπα *vs. χρόνος (εάν είναι χρονοσειρές)*

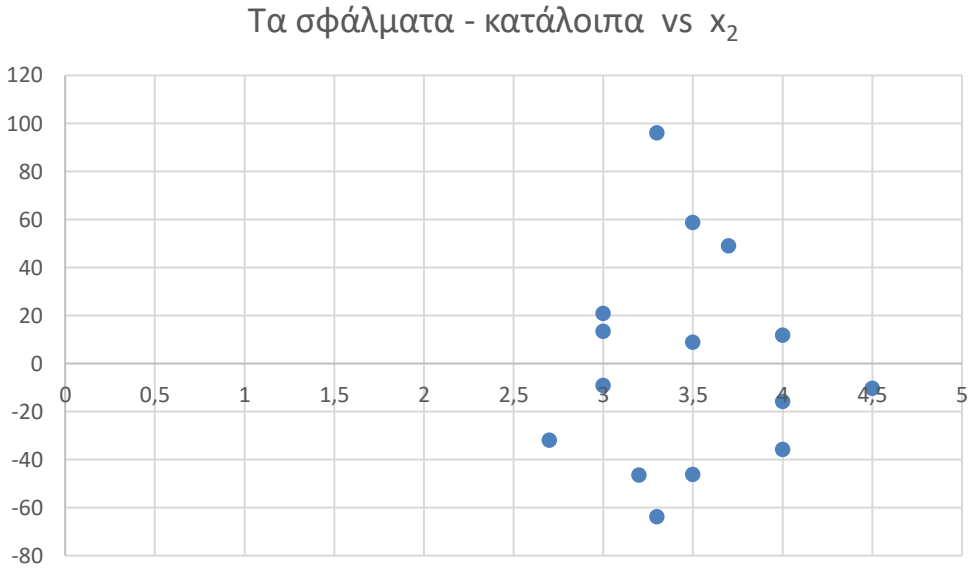
Χρήση των καταλοίπων για έλεγχο των παραδοχών του μοντέλου πολλαπλής παλινδρόμησης

4. Μια μελέτη περίπτωσης

Τα διαγράμματα των σφαλμάτων σε σχέση με τις μεταβλητές παρουσιάζουν «*νέφος*» και δεν δείχνουν καμία «*κανονικότητα*»







Multicollinearity (**πολυσυγραμμικότητα***)* υπάρχει όταν δυο ή περισσότερες ανεξάρτητες ερμηνευτικές μεταβλητές ***έχουν μεταξύ τους υψηλή συσχέτιση***
Όταν υπάρχει τέτοια πολυσυγραμμικότητα οι συντελεστές (coefficients) του μοντέλου πολλαπλής παλινδρόμησης είναι πολύ ασταθείς / μεταβάλλονται σημαντικά όταν υπάρχουν μικρές αλλαγές στα δεδομένα – άρα χαμηλής απόδοσης μοντέλο παλινδρόμησης

Στην εφαρμογή με τα επιδόρπια Ο συντελεστής συσχέτισης των ανεξάρτητων ερμηνευτικών μεταβλητών που είναι τιμή μονάδας x_1 και διαφημιστική δαπάνη x_2 είναι:
=CORREL(E5:E19;F5:F19) = 0,030437581 ιδιαίτερα μικρός που δείχνει ότι ***δεν υπάρχει συσχέτιση ανάμεσα σε αυτές τις μεταβλητές άρα δεν υπάρχει κίνδυνος πολυσυγραμμικότητας***

Γενικά
η Διαδικασία πολλαπλής παλινδρόμησης

- Προσδιορισμός του μοντέλου πολλαπλής παλινδρόμησης
- Έλεγχος - Test της σημαντικότητας του μοντέλου που εκτιμήθηκε
- Έλεγχος - Test της σημαντικότητας των συντελεστών της παλινδρόμησης
- Αξιολόγηση του προσαρμοσμένου r^2
- Μελέτη των κατάλοιπων