Κεφάλαιο 13

Απλή Γραμμική Παλινδρόμηση

,

Στόχοι

- **Σε αυτό το κεφάλαιο μαθαίνετε:**
- Πώς να χρησιμοποιείτε την ανάλυση παλινδρόμησης για να προβλέψετε την τιμή μιας εξαρτημένης μεταβλητής με βάση την τιμή μιας ανεξάρτητης μεταβλητής
- Να κατανοείτε την σημασία των συντελεστών παλινδρόμησης b_0 και b_1
- Να αξιολογείτε τις υποθέσεις της ανάλυσης παλινδρόμησης και να γνωρίζετε τι να κάνετε όταν οι υποθέσεις παραβιάζονται
- Να βγάζετε συμπεράσματα σχετικά με την κλίση και το συντελεστή συσχέτισης
- Να εκτιμάτε τις μέσες τιμές και να προβλέπετε τις μεμονωμένες τιμές

,

Μέθοδος Ελαχίστων Τετραγώνων

Ο στόχος του διαγράμματος διασποράς είναι η μέτρηση της ισχύος και της κατεύθυνσης της γραμμικής συσχέτισης.

Και τα δύο μπορούν πιο εύκολα να εξαχθούν σχεδιάζοντας μια ευθεία γραμμή μέσα στα δεδομένα.

Χρειαζόμαστε μια αντικειμενική μέθοδο για τη δημιουργία αυτής της ευθείας.

Μία τέτοια μέθοδος είναι η **μέθοδος ελαχίστων τετραγώνων**.

,

Μέθοδος Ελαχίστων Τετραγώνων

Θυμίζουμε ότι, η εξίσωση μιας ευθείας με γνωστή κλίση δίνεται από τον τύπο:

$$y = mx + b$$

όπου:

m είναι η κλίση της ευθείας

b είναι το σημείο τομής με τον άξονα y.

Εάν γνωρίζουμε ότι υπάρχει γραμμική σχέση μεταξύ δύο μεταβλητών με γνωστή συνδιασπορά και γνωστό συντελεστή συσχέτισης, μπορούμε να καθορίσουμε τη γραμμική συνάρτηση της σχέσης;

,

Η Μέθοδος Ελαχίστων Τετραγώνων ...

…δημιουργεί μια ευθεία γραμμή ανάμεσα στα σημεία ώστε το άθροισμα των τετραγώνων των αποκλίσεων μεταξύ των σημείων και της ευθείας να ελαχιστοποιείται. Η ευθεία ορίζεται από την εξίσωση:

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

b_0 ("b" μηδέν) είναι το σημείο τομής με τον άξονα y,

b_1 είναι η κλίση, και

$\hat{y}$ ("y" καπέλο) είναι η τιμή του y που καθορίζει η ευθεία.

,

Μέθοδος Ελαχίστων Τετραγώνων

Οι συντελεστές b_0 και b_1 δίνονται από:

$$b_1 = \frac{s_{xy}}{s_x^2}$$

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

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

,

Σταθερό και Μεταβλητό Κόστος

Το σταθερό κόστος είναι το κόστος που πρέπει να πληρωθεί ανεξάρτητα από τον κύκλο δραστηριοτήτων.

Το κόστος αυτή είναι “σταθερό” για μια συγκεκριμένη χρονική περίοδο ή για συγκεκριμένο εύρος παραγωγής.

Το μεταβλητό κόστος είναι αυτό που μεταβάλλεται σε σχέση με το πλήθος των παραγόμενων προϊόντων.

,

Σταθερό και Μεταβλητό Κόστος

Υπάρχουν ωστόσο και μεικτά έξοδα.

Υπάρχουν αρκετοί τρόποι να διαχωρίσουμε το μεικτό κόστος στη σταθερή και στη μεταβλητή συνιστώσα του. Μια τέτοια μέθοδος είναι η ευθεία ελαχίστων τετραγώνων. Δηλαδή, εκφράζουμε το συνολικό κόστος ως

$$y = b_0 + b_1 x$$

όπου y = συνολικό μεικτό κόστος, b_0 = σταθερό κόστος και b_1 = μεταβλητό κόστος, και x το πλήθος των παραγόμενων μονάδων

,

Παράδειγμα 4.17

Ένα μικρό μηχανουργείο κατασκευάζει εργαλεία.

Σκέφτεται να επεκτείνει τις δραστηριότητές του και χρειάζεται ανάλυση του κόστους παραγωγής.

Μια πηγή κόστους είναι το ηλεκτρικό, το οποίο απαιτείται για τη λειτουργία των μηχανών και το φωτισμό. (Μερικές εργασίες απαιτούν ιδιαίτερα ισχυρό φωτισμό.)

Έχει καταγράψει το καθημερινό κόστος ηλεκτρισμού και τον αριθμό των εργαλείων που παράγει. Να καθορίσετε το σταθερό και το μεταβλητό κόστος ηλεκτρισμού. [Xm04-17]

,

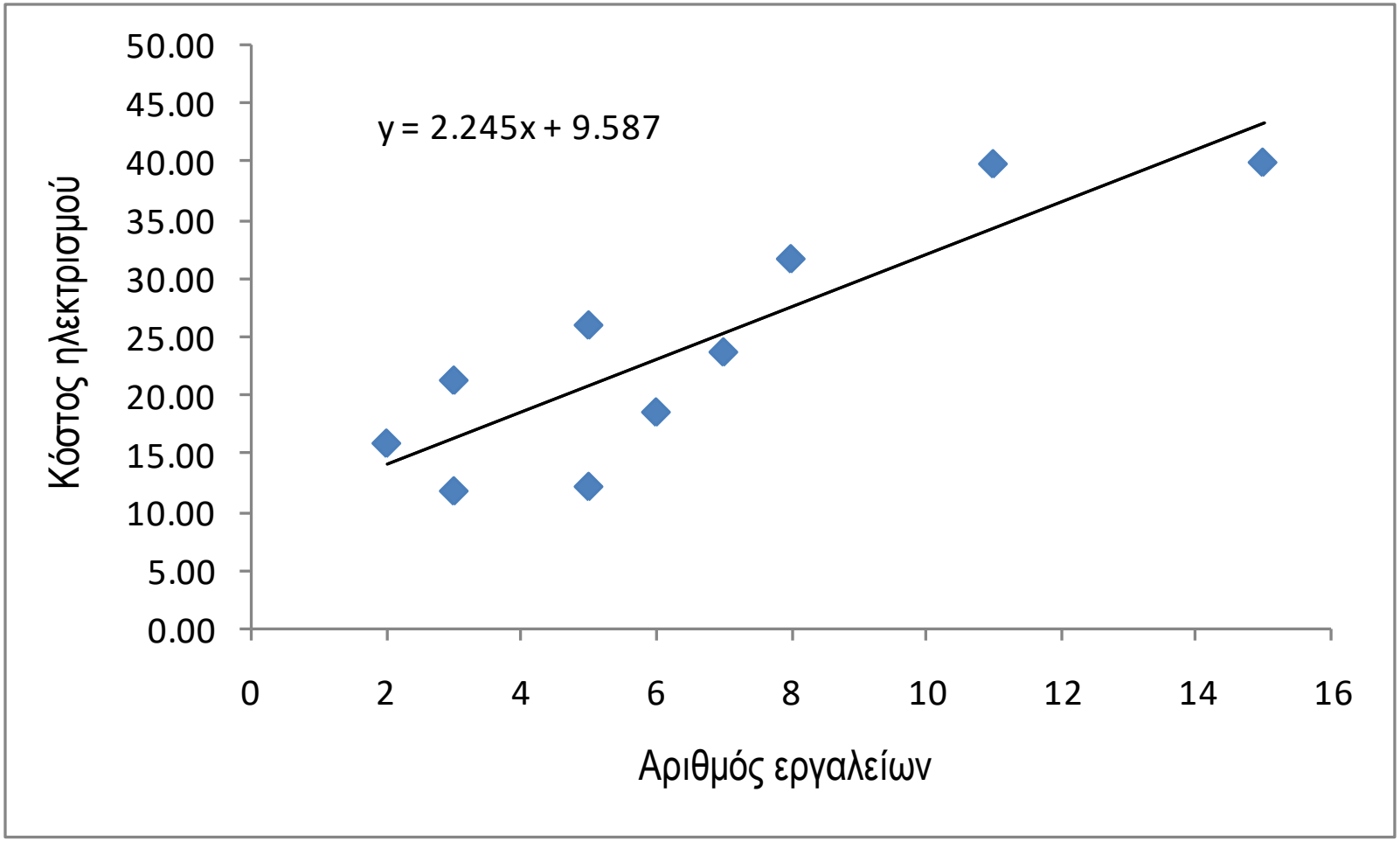
Παράδειγμα 4.17

Λαμβάνει ένα δείγμα 10 ημερών και καταγράφει για κάθε μέρα τον αριθμό παραγόμενων εργαλείων και το σχετικό κόστος παραγωγής.

Ημέρα	Αριθμός Εργαλείων	Κόστος Ηλεκτρισμού
1	7	23,80
2	3	11,89
3	2	15,98
4	5	26,11
5	8	31,79
6	11	39,93
7	5	12,27
8	15	40,06
9	3	21,38
10	6	18,65

,

Παράδειγμα 4.17



,

Παράδειγμα 4.17

$$\hat{y} = 9.587 + 2.245x$$

Η κλίση δείχνει μεταβολή/μονάδα, που σημαίνει ότι είναι η μεταβολή του y (αύξηση) για κάθε 1-μονάδα αύξησης του x .

Η κλίση μετράει την *οριακή* μεταβολή της εξαρτημένης μεταβλητής. Η οριακή μεταβολή αναφέρεται στο αποτέλεσμα της αύξησης της ανεξάρτητης μεταβλητής κατά μία επιπλέον μονάδα.

Στο παράδειγμα αυτό η κλίση είναι 2.25, που σημαίνει ότι για κάθε 1-μονάδα αύξησης στον αριθμό των εργαλείων, η οριακή αύξηση στο κόστος ηλεκτρισμού είναι 2.25. Άρα, το εκτιμώμενο μεταβλητό κόστος είναι $2.25 ανά εργαλείο.

,

Παράδειγμα 4.17

$$\hat{y} = 9.59 + 2.25x$$

Το σημείο τομής με τον άξονα y είναι 9.59.
Αυτή είναι απλά η τιμή όταν x = 0.

Ωστόσο, όταν x = 0 δεν έχουμε παραγωγή εργαλείων και συνεπώς το εκτιμώμενο σταθερό κόστος ηλεκτρισμού είναι $9.59 ανά ημέρα.

,

Συντελεστής Προσδιορισμού

Όταν είδαμε τον συντελεστή συσχέτισης σημειώσαμε ότι εκτός από τις τιμές −1, 0, και +1, δεν μπορούμε να ερμηνεύσουμε τη σημασία του.

Μπορούμε να κρίνουμε τον συντελεστή συσχέτισης μόνο σε σχέση με το πόσο πλησιάζει στο −1, 0, και +1 .

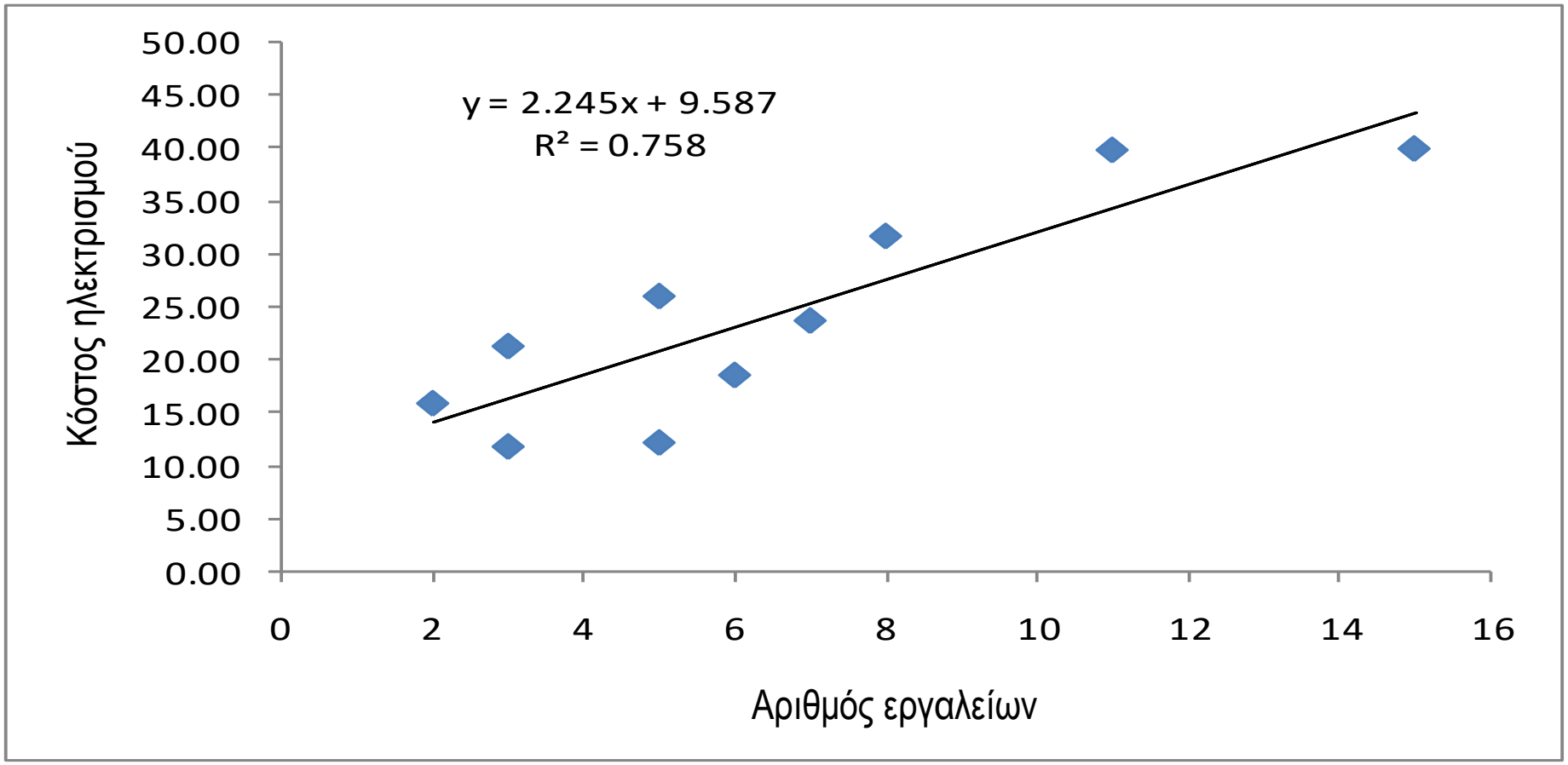
Ευτυχώς, έχουμε άλλο ένα δείκτη που μπορεί να ερμηνευθεί καλύτερα. Είναι ο ***συντελεστής προσδιορισμού***, ο οποίος υπολογίζεται υψώνοντας στο τετράγωνο τον συντελεστή συσχέτισης. Γι' αυτό συμβολίζεται R^2 .

Ο συντελεστής προσδιορισμού εκφράζει σε ποιο βαθμό η μεταβλητότητα της εξαρτημένης μεταβλητής εξηγείται από τη μεταβλητότητα της ανεξάρτητης.

,

Παράδειγμα 4.17

Υπολογίστε τον συντελεστή προσδιορισμού για το Παράδειγμα 4.17



,

Παράδειγμα 4.17

Ο συντελεστής προσδιορισμού είναι

$$R^2 = .758$$

Αυτό μας λέει ότι το 75.8% του κόστους ηλεκτρισμού εξηγείται από τον αριθμό των εργαλείων που κατασκευάζονται.

Το υπόλοιπο 24.2% οφείλεται σε άλλους λόγους.

,

Ερμηνεία της Συσχέτισης ...

Λόγω της σημασίας της, θυμίζουμε τη σωστή ερμηνεία την ανάλυσης της σχέσης μεταξύ δύο συνεχών μεταβλητών.

Δηλαδή, εάν δύο μεταβλητές είναι γραμμικά συσχετισμένες αυτό δεν σημαίνει ότι η Χ προκαλεί τη μεταβολή της Υ. Μπορεί να σημαίνει ότι μια άλλη μεταβλητή επηρεάζει και την Χ και την Υ ή ότι η Υ προκαλεί τη μεταβολή της Χ.

Θυμηθείτε

"Συσχέτιση δεν είναι αιτιολόγηση"

,

SPSS output, Παράδειγμα 4.17

Regression

Descriptive Statistics

	Mean	Std. Deviation	N
Ecost	24,1860	10,33103	10
Ntools	6,5000	4,00694	10

Correlations

		Ecost	Ntools
Pearson Correlation	Ecost	1,000	**,871**
	Ntools	,871	1,000
Sig. (1-tailed)	Ecost	.	**,001**
	Ntools	,001	.
N	Ecost	10	10
	Ntools	10	10

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	**,871[a]**	**,759**	,729	5,38185

a. Predictors: (Constant), Ntools

b. Dependent Variable: Ecost

,

SPSS output, Παράδειγμα 4.17 (2)

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	728,856	1	728,856	25,164	,001[a]
	Residual	231,715	8	28,964		
	Total	960,571	9			

a. Predictors: (Constant), Ntools

b. Dependent Variable: Ecost

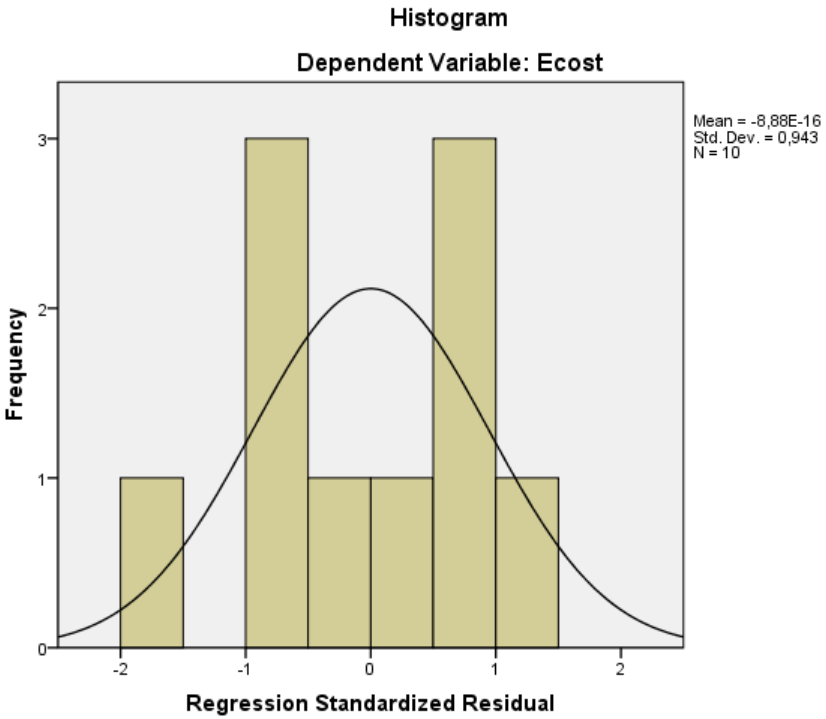
Coefficients[a]

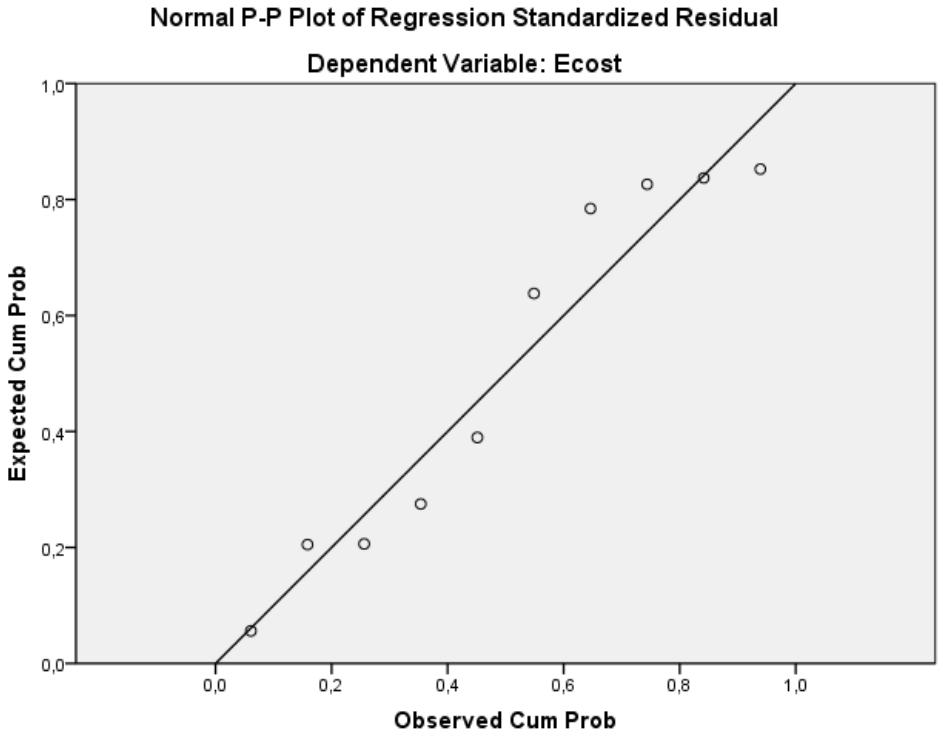
Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B	
	B	Std. Error	Beta			Lower Bound	Upper Bound
1 (Constant)	**9,588**	3,371		2,844	**,022**	1,814	17,362
Ntools	**2,246**	,448	,871	5,016	**,001**	1,213	3,278

a. Dependent Variable: Ecost

,

SPSS output, Παράδειγμα 4.17 (3)





,

Ανάλυση Παλινδρόμησης ...

Ο στόχος μας είναι να ***αναλύσουμε τη σχέση*** μεταξύ συνεχών μεταβλητών. Η ***ανάλυση παλινδρόμησης*** είναι το πρώτο εργαλείο που θα μελετήσουμε.

Η ανάλυση παλινδρόμησης χρησιμοποιείται για την πρόβλεψη της τιμής μιας μεταβλητής (***εξαρτημένη μεταβλητή***) με βάση την τιμή άλλων μεταβλητών (***ανεξάρτητες μεταβλητές***).

Εξαρτημένη μεταβλητή: συμβολίζεται με **Y**

Ανεξάρτητες μεταβλητές: συμβολίζονται με $\mathbf{X_1, X_2, \ldots, X_k}$

,

Ανάλυση Παλινδρόμησης ...

Εάν μας ενδιαφέρει να καθορίσουμε ***μόνο*** το εάν ***υπάρχει*** σχέση, χρησιμοποιούμε ***ανάλυση συσχέτισης***, μια τεχνική που έχουμε ήδη δει.

Στο κεφάλαιο αυτό θα εξετάσουμε τη σχέση μεταξύ ***δύο μεταβλητών***, με ***απλή γραμμική παλινδρόμηση***.

Οι μαθηματικές εξισώσεις που περιγράφουν αυτές τις σχέσεις καλούνται και ***μοντέλα***, και ταξινομούνται σε δύο κατηγορίες: ντετερμινιστικά ή πιθανοθεωρητικά.

,

Μοντέλα ...

Ντετερμινιστικό Μοντέλο: μια εξίσωση ή σύνολο εξισώσεων που μας επιτρέπει να ***καθορίσουμε πλήρως*** την τιμή της εξαρτημένης μεταβλητής από τις τιμές των ανεξάρτητων μεταβλητών.

Σε αντίθεση με …

Πιθανοθεωρητικό Μοντέλο: μια μέθοδος για να υπολογίσουμε την ***τυχαιότητα*** που υπάρχει στις πραγματικές διαδικασίες.

Π.χ. έχουν πωληθεί όλα τα σπίτια ίσου εμβαδού στην ίδια ακριβώς τιμή;

,

Ένα Μοντέλο ...

Για να δημιουργήσουμε ένα πιθανοθεωρητικό μοντέλο, ξεκινάμε με ένα ντετερμινιστικό μοντέλο το οποίο ***προσεγγίζει τη σχέση*** που θέλουμε να προσδιορίσουμε και προσθέτουμε έναν ***τυχαίο όρο*** που μετράει το ***σφάλμα*** της ντετερμινιστικής συνιστώσας;.

Ντετερμινιστικό Μοντέλο:

Το κόστος κατασκευής ενός νέου σπιτιού είναι περίπου \$100 ανά τετραγωνικό πόδι (ft^2) ενώ ο περίβολος αυξάνει την αξία περίπου κατά \$100,000. Επομένως, η εκτιμώμενη τιμή πώλησης (**y**) είναι:

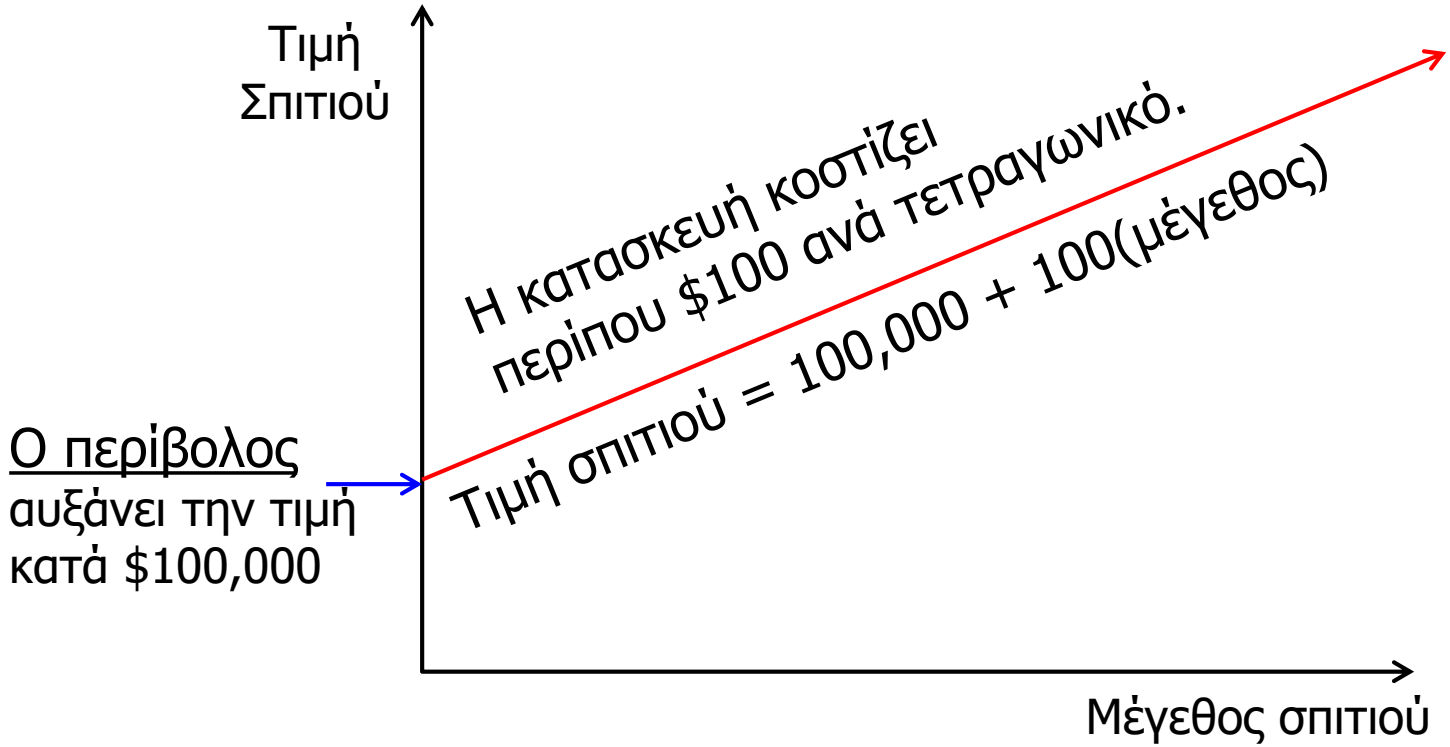
$$\mathbf{y = \$100{,}000 + (100\$/ft^2)(x)}$$

(όπου **x** τα τετραγωνικά πόδια του σπιτιού)

,

Ένα Μοντέλο ...

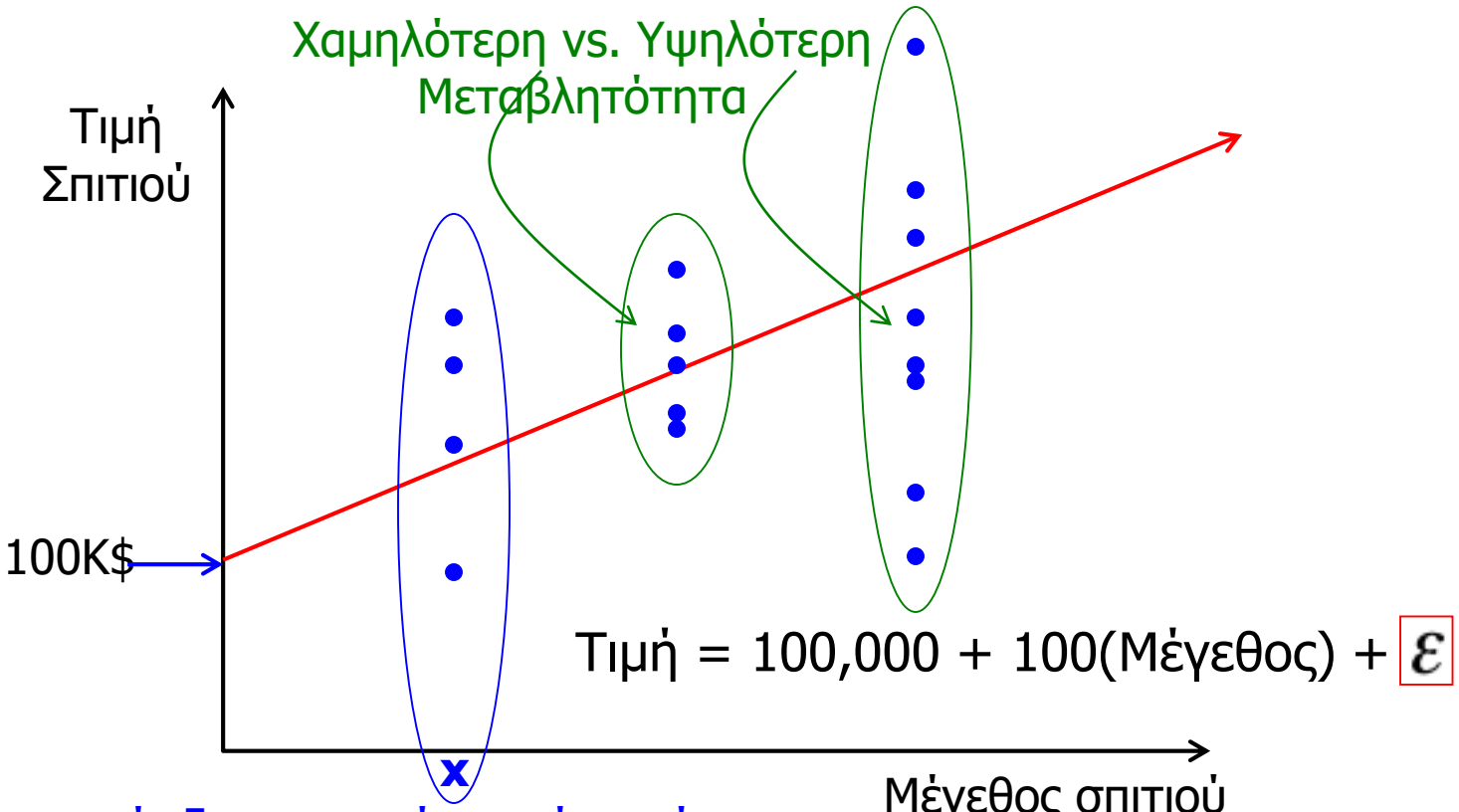
Ένα μοντέλο της σχέσης μεταξύ του μεγέθους του σπιτιού (ανεξάρτητη μεταβλητή) και της τιμής του (εξαρτημένη μεταβλητή) θα ήταν:



Στο μοντέλο αυτό, η τιμή είναι πλήρως **καθορισμένη** από το μέγεθος.

Ένα Μοντέλο ...

Στην πραγματικότητα όμως, η τιμή του σπιτιού διαφοροποιείται ακόμα και μεταξύ σπιτιών ίδιου μεγέθους:



Τα ίδια τετραγωνικά, διαφορετικά σημεία τιμής (π.χ. επιλογές διακόσμησης, τοποθεσία ...)

Τυχαίος Όρος ...

Αναπαριστούμε την τιμή ενός σπιτιού ως συνάρτηση του μεγέθους του στο Πιθανοθεωρητικό Μοντέλο:

$$y = 100{,}000 + 100x + \varepsilon$$

όπου ε είναι ο ***τυχαίος όρος*** (***ή μεταβλητή σφάλματος***).

Είναι η διαφορά μεταξύ της ***πραγματικής*** τιμής πώλησης και της ***εκτιμώμενης*** με βάση το μέγεθος του σπιτιού. Η τιμή του διαφοροποιείται μεταξύ των πωλήσεων, ακόμα κι αν τα τετραγωνικά (δηλ. **x**) παραμένουν ίδια.

,

Απλή Γραμμική Παλινδρόμηση ...

Ένα μοντέλο ευθείας γραμμής με μια ανεξάρτητη μεταβλητή καλείται ***γραμμικό μοντέλο πρώτης τάξης*** ή ***μοντέλο απλής γραμμικής παλινδρόμησης***. Δίνεται από την:

εξαρτημένη μεταβλητή

ανεξάρτητη μεταβλητή

$$y = \beta_0 + \beta_1 x + \varepsilon$$

Τομή με y

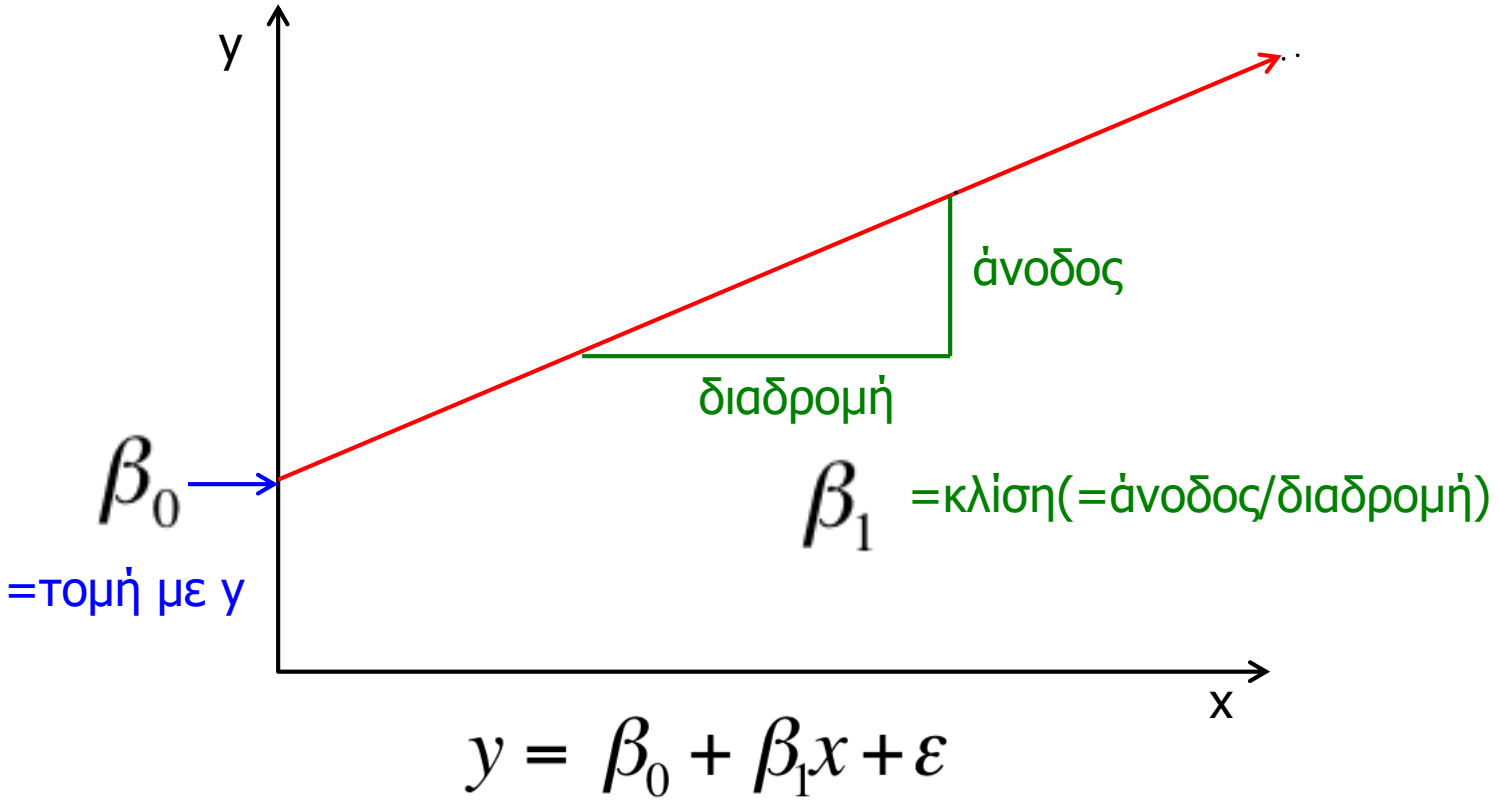
Κλίση ευθείας

Μεταβλητή σφάλματος

,

Απλή Γραμμική Παλινδρόμηση ...

Σημειώστε ότι και το β_0 και το β_1 είναι ***παράμετροι του πληθυσμού*** οι οποίες είναι συνήθως άγνωστες και επομένως ***εκτιμώνται*** από τα δεδομένα.



$$y = \beta_0 + \beta_1 x + \varepsilon$$

,

Εκτίμηση Συντελεστών ...

Θα εκτιμήσουμε το β_0 χρησιμοποιώντας το b_0 και το β_1 χρησιμοποιώντας το b_1, την τομή με y και την κλίση (αντίστοιχα) της ***ευθείας ελαχίστων τετραγώνων*** ή ***ευθείας παλινδρόμησης*** που δίνεται από την:

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

(Θυμίζουμε ότι είναι μια εφαρμογή της μεθόδου ελαχίστων τετραγώνων και δημιουργεί μια ευθεία η οποία ***ελαχιστοποιεί*** το άθροισμα των τετραγώνων των αποκλίσεων μεταξύ των σημείων και της ευθείας)

,

Παράδειγμα 16.1

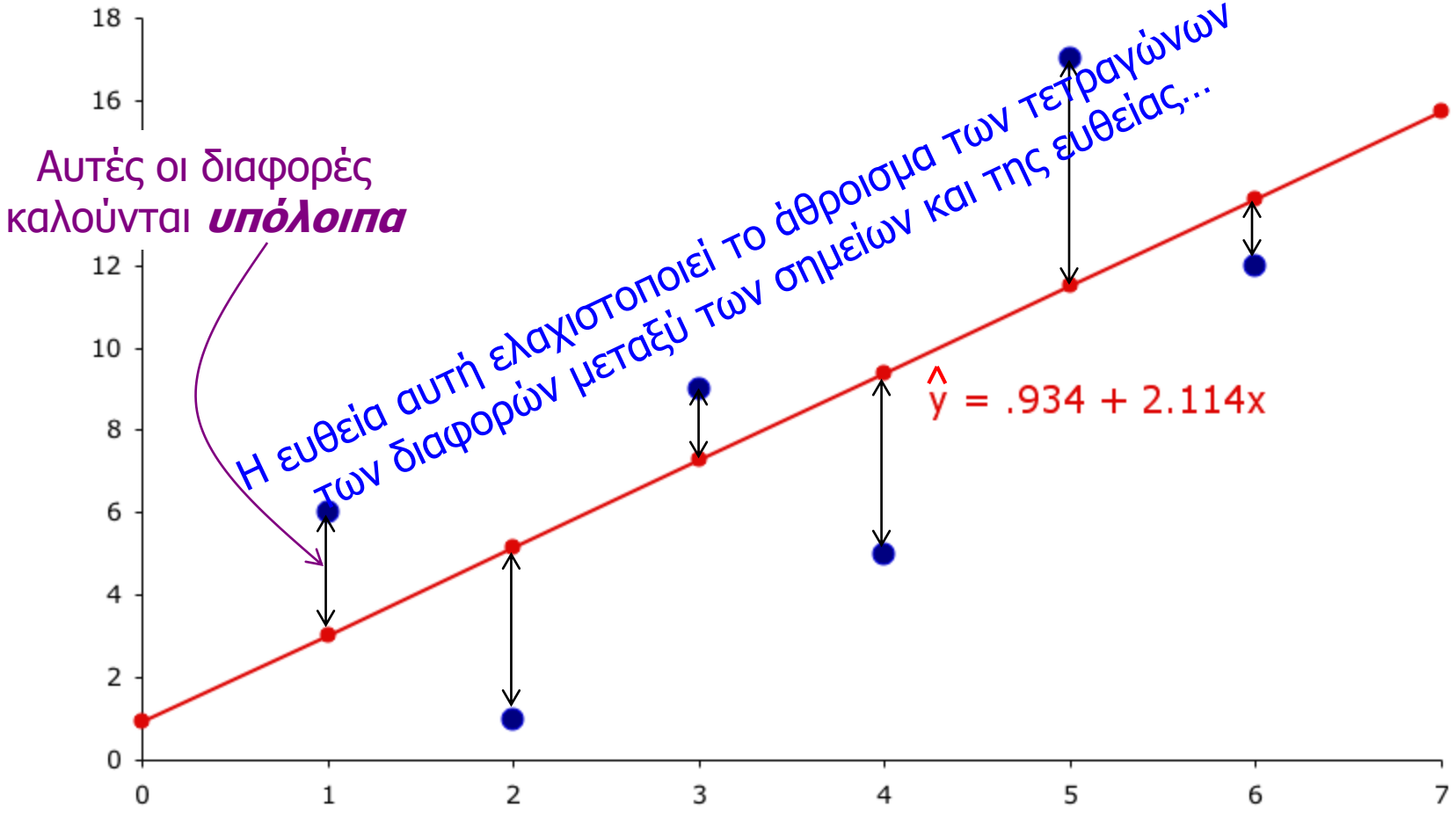
Το πριμ απόδοσης (σε χιλιάδες δολάρια) έξι υπαλλήλων με διαφορετικά χρόνια προϋπηρεσίας δίνεται στον παρακάτω πίνακα. Θέλουμε να καθορίσουμε τη γραμμική σχέση μεταξύ του πριμ και των ετών προϋπηρεσίας.

Έτη προϋπηρεσίας	x	1	2	3	4	5	6
Ετήσιο πριμ	y	6	1	9	5	17	12

,

Ευθεία Ελαχίστων Τετραγώνων...

Παράδειγμα 16.1



,

Παράδειγμα 16.2…

Οι έμποροι αυτοκινήτων στη Βόρεια Αμερική χρησιμοποιούν το «Κόκκινο Βιβλίο» για να καθορίσουν την τιμή των μεταχειρισμένων αυτοκινήτων που παίρνουν με ανταλλαγή όταν οι πελάτες τους αγοράζουν καινούργιο.

Το βιβλίο εκδίδεται κάθε μήνα και καταγράφει τις τιμές για όλα τα βασικά μοντέλα αυτοκινήτων.

Έχει εναλλακτικά τιμές για κάθε μοντέλο ανάλογα με την κατάστασή του και τον εξοπλισμό του.

Οι τιμές καθορίζονται με βάση τη μέση τιμή σε πρόσφατες δημοπρασίες.

,

Παράδειγμα 16.2…

Ωστόσο, το Κόκκινο Βιβλίο δεν δίνει την τιμή με βάση την ένδειξη των χιλιομέτρων, παρά το γεγονός ότι αυτό αποτελεί σημαντικό παράγοντα για τους αγοραστές μεταχειρισμένων αυτοκινήτων.

Για να εξετάσει αυτό το θέμα, ένας πωλητής επέλεξε τυχαία 100 αυτοκίνητα Toyota Camrys τριών ετών, τα οποία πωλήθηκαν σε δημοπρασίες τον τελευταίο μήνα.

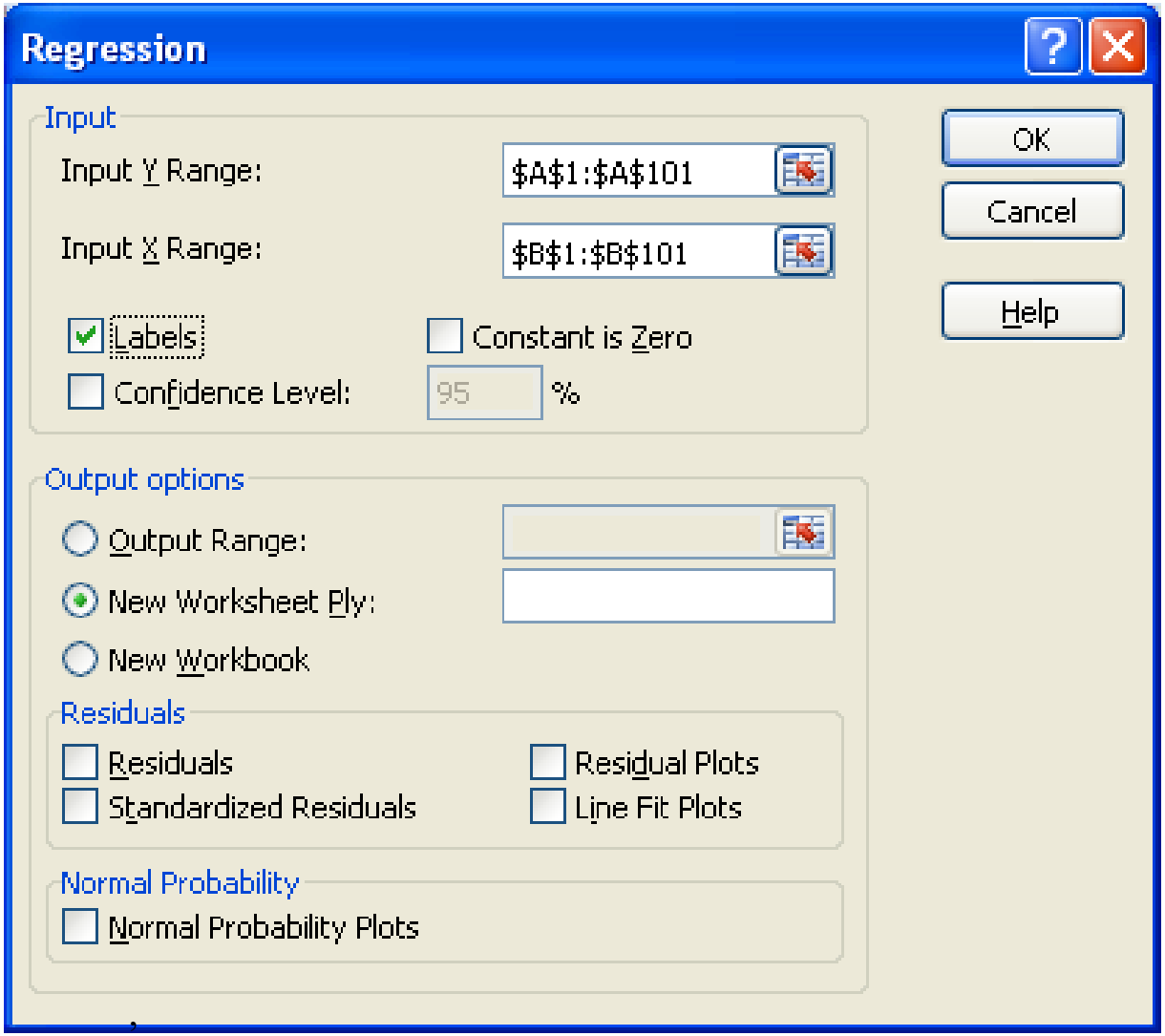
Ο πωλητής κατέγραψε την τιμή πώλησης (σε χιλιάδες δολάρια) και τον αριθμό των μιλίων (σε χιλιάδες) του κοντέρ. (Xm16-02).

Ο πωλητής θέλει να υπολογίσει την ευθεία παλινδρόμησης.

,

Παράδειγμα 16.2…

Χρησιμοποιεί το Excel ….



Παράδειγμα 16.2…

	A	B	C	D	E	F
1	SUMMARY OUTPUT					
2						
3	*Regression Statistics*					
4	Multiple R	0.8052				
5	R Square	0.6483				
6	Adjusted R Square	0.6447				
7	Standard Error	0.3265				
8	Observations	100				
9						
10	ANOVA					
11		*df*	*SS*	*MS*	*F*	*Significance F*
12	Regression	1		19.26	180.64	5.75E-24
13	Residual	98	.45	0.11		
14	Total	99	29.70			
15						
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	
17	Intercept	17.25	0.182	94.73	3.57E-98	
18	Odometer	-0.0669	0.0050	-13.44	5.75E-24	

Από όλα όσα έχει υπολογίσει μας ενδιαφέρουν μόνο αυτά

$$\hat{y} = b_0 + b_1 x = 17.250 - 0.0669x$$

,

Παράδειγμα 16.2…

ΕΡΜΗΝΕΙΑ

Όπως ήταν αναμενόμενο …

Ο συντελεστής κλίσης, b_1, είναι –0.0669, δηλαδή, κάθε επιπλέον μίλι στο κοντέρ μειώνει την τιμή κατά 0.0669 δολάρια ή 6.69 σεντς.

Το ίχνος στον y, b_0, είναι 17,250. Μια ερμηνεία είναι ότι όταν x = 0 (δεν έχει κινηθεί) η τιμή πώλησης είναι $17,250. Ωστόσο, δεν έχουμε δεδομένα για αυτοκίνητα με λιγότερα από 19,100 μίλια, επομένως αυτή η εκτίμηση δεν είναι σωστή.

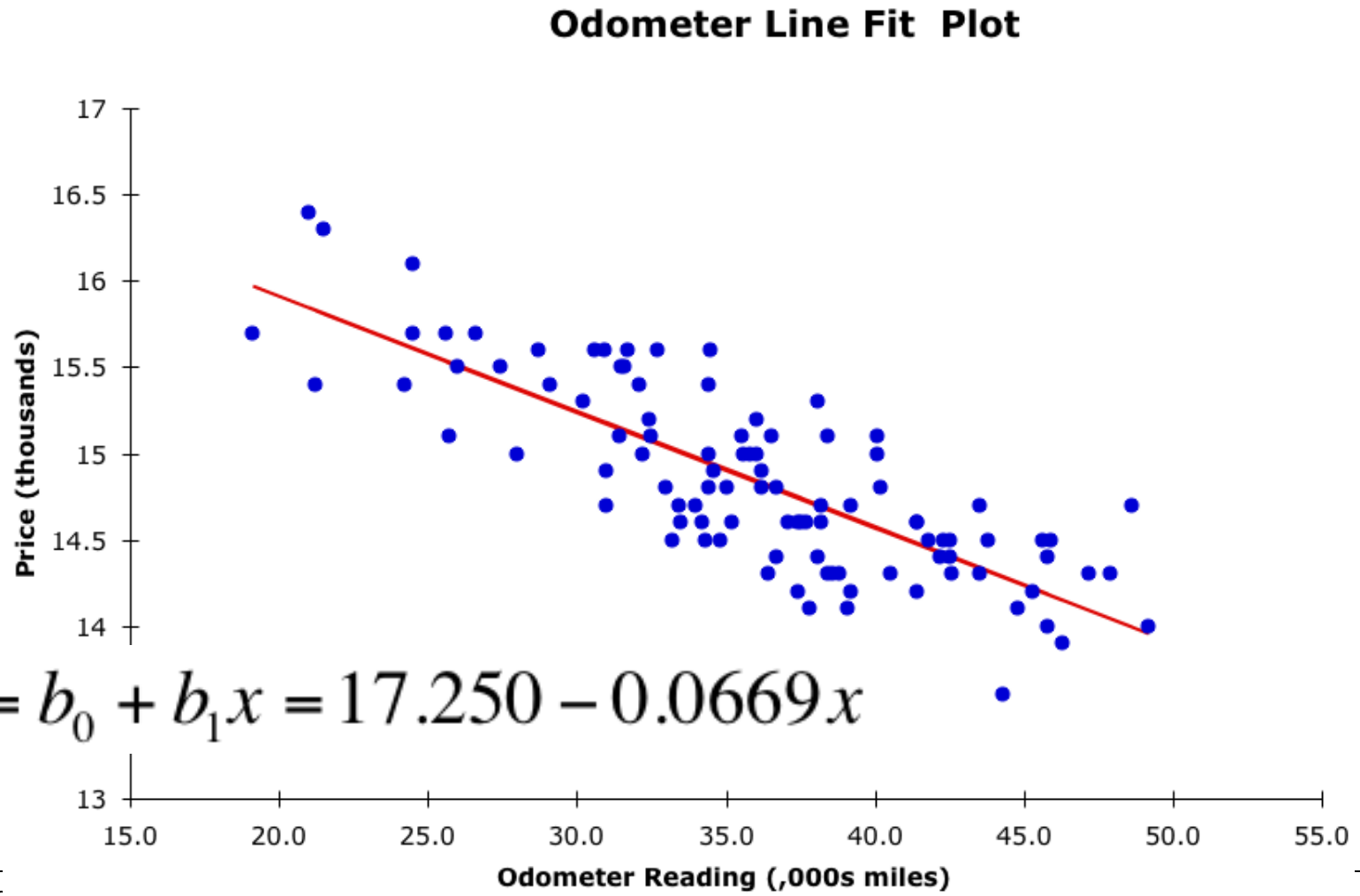
$$\hat{y} = b_0 + b_1 x = 17.250 - 0.0669x$$

,

Παράδειγμα 16.2…

ΕΡΜΗΝΕΙΑ

Κατασκευάζουμε το διάγραμμα των δεδομένων και την ευθεία παλινδρόμησης



$$\hat{y} = b_0 + b_1 x = 17.250 - 0.0669x$$

,

Απαιτούμενες Συνθήκες ...

Για να ισχύουν τα προηγούμενα, πρέπει να πληρούνται τέσσερεις συνθήκες :

- Η κατανομή πιθανοτήτων του ε να είναι κανονική.
- Ο μέσος της κατανομής είναι 0, δηλαδή $E(\varepsilon) = 0$.
- Η τυπική απόκλιση του ε, σ_ε, είναι σταθερή για κάθε τιμή του x.
- Η τιμή του ε που αντιστοιχεί σε κάθε τιμή του y είναι ανεξάρτητη του ε για κάθε άλλη τιμή του y.

,

Μέτρα Μεταβλητότητας

- Η συνολική μεταβλητότητα αποτελείται από δύο μέρη:

$$SST = SSR + SSE$$

Συνολικό Άθροισμα Τετραγώνων	Άθροισμα Τετραγώνων Παλινδρόμησης	Άθροισμα Τετραγώνων των Σφαλμάτων
$SST = \sum(Y_i - \bar{Y})^2$	$SSR = \sum(\hat{Y}_i - \bar{Y})^2$	$SSE = \sum(Y_i - \hat{Y}_i)^2$

όπου:

$\bar{Y}$ = Μέση τιμή της εξαρτημένης μεταβλητής

Y_i = Παρατηρούμενη τιμή της εξαρτημένης μεταβλητής

$\hat{Y}_i$ = Προβλεπόμενη τιμή του Y για τη δεδομένη τιμή X_i

,

Μέτρα Μεταβλητότητας

(συνέχεια)

- SST = συνολ. άθροισμα τετραγώνων (Συνολική Μεταβλητότητα)

  Μετρά την μεταβλητότητα των τιμών Y_i γύρω από τους μέσους $\bar{Y}$

- SSR = άθροισμα τετραγώνων παλινδρόμησης
  (Μεταβλητότητα που εξηγείται από το μοντέλο)

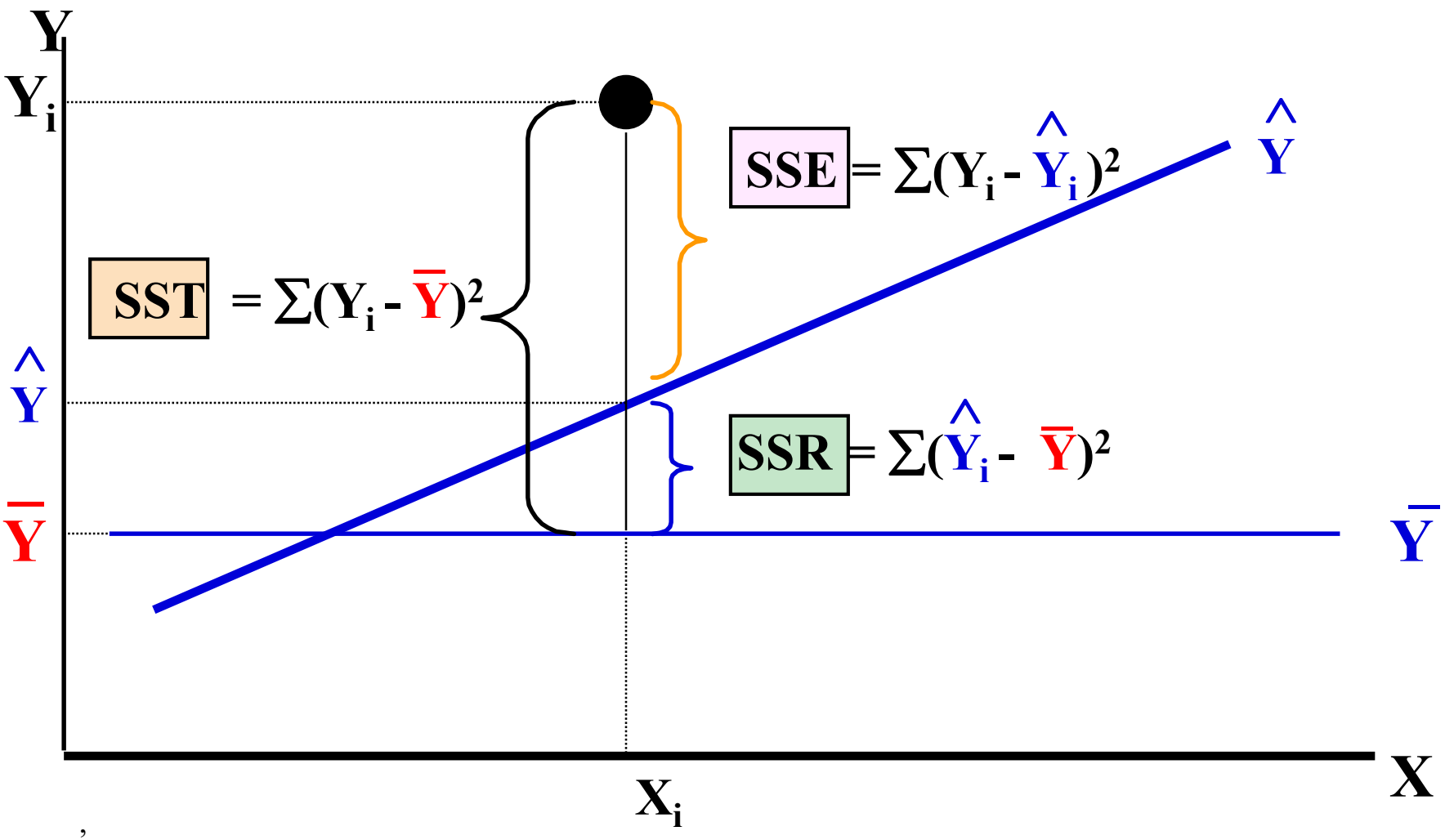
  Μεταβλητότητα που μπορεί να αποδοθεί στην σχέση μεταξύ X και Y

- SSE = άθροισμα τετραγώνων των σφαλμάτων
  (Μεταβλητότητα που δεν εξηγείται από το μοντέλο)

  Μεταβλητότητα του Y που αποδίδεται σε παράγοντες άλλους από το X

Μέτρα Μεταβλητότητας

(συνέχεια)



,

Αξιολόγηση του μοντέλου ...

Η μέθοδος ελαχίστων τετραγώνων δημιουργεί πάντοτε μια ευθεία, ακόμα κι αν δεν υπάρχει σχέση μεταξύ των μεταβλητών, ή κι αν η σχέση είναι μη γραμμική.

Επομένως, πέρα από τον καθορισμό των συντελεστών της ευθείας ελαχίστων τετραγώνων, πρέπει να την αξιολογήσουμε για να δούμε πόσο καλά "ταιριάζει" στα δεδομένα. Θα δούμε στη συνέχεια αυτές τις μεθόδους αξιολόγησης. Βασίζονται στο άθροισμα των τετραγώνων των σφαλμάτων (SSE).

,

Άθροισμα Τετραγώνων Σφάλματος (SSE)…

Το άθροισμα τετραγώνων σφάλματος υπολογίζεται ως:

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

Και χρησιμοποιείται για τον υπολογισμό του ***τυπικού σφάλματος εκτίμησης***:

$$s_\varepsilon = \sqrt{\frac{SSE}{n-2}} \qquad \text{ή}$$

$$S_{YX} = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}{n-2}}$$

Αν s_ε ή S_{YX} είναι μηδέν, τότε όλα τα σημεία είναι πάνω στην ευθεία παλινδρόμησης.

,

Τυπικό Σφάλμα Εκτίμησης …

	A	B	C	D		
1	SUMMARY OUTPUT					
2						
3	*Regression Statistics*					
4	Multiple R	0.8052				
5	R Square	0.6483				
6	Adjusted R Square	0.6447				
7	Standard Error	0.3265				
8	Observations	100				
9						
10	ANOVA					
11		*df*	*SS*	*MS*	*F*	*Significance F*
12	Regression	1	19.26	19.26	180.64	0.00
13	Residual	98	10.45	0.11		
14	Total	99	29.70			
15						
16		***Coefficients***	*Standard Error*	*t Stat*	*P-value*	
17	**Intercept**	**17.25**	0.182	94.73	0.0000	
18	**Odometer**	**-0.0669**	0.0050	-13.44	0.0000	

s_ε

Αν s_ε ή s_{YX} είναι μικρό, η προσαρμογή είναι εξαιρετική και το γραμμικό μοντέλο μπορεί να χρησιμοποιηθεί για πρόβλεψη. Αν s_εήs_{xy} είναι μεγάλο, το μοντέλο μας δεν είναι καλό…

Αλλά πότε είναι ***μικρό*** και πότε είναι ***μεγάλο*** ;

,

Τυπικό Σφάλμα Εκτίμησης ...

Κρίνουμε την τιμή του s_ε ή s_{YX} συγκρίνοντάς το με το μέσο της εξαρτημένης μεταβλητής $\bar{y}$.

Στο παράδειγμά μας,

$s_\varepsilon = .3265$ και

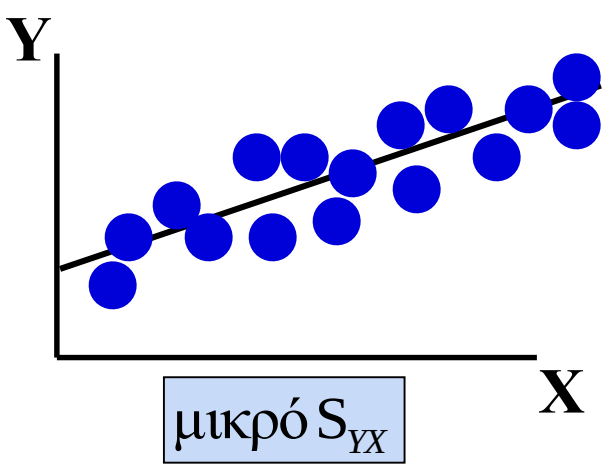
$\bar{y} = 14.841$

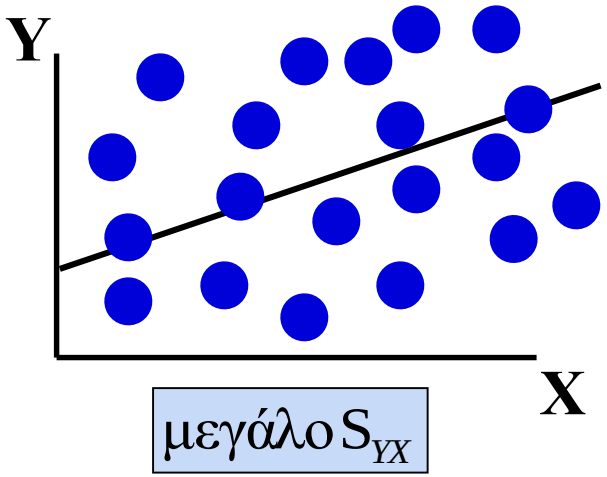
άρα (μιλώντας σχετικά) φαίνεται να είναι “μικρό”, άρα το μοντέλο γραμμικής παλινδρόμησης της τιμής πώλησης των αυτοκινήτων ως συνάρτηση της ένδειξης του κοντέρ είναι “καλό”.

,

Σύγκριση Τυπικών Σφαλμάτων

Το S_{YX} είναι ένα μέτρο μεταβλητότητας των παρατηρούμενων τιμών Y από την γραμμή παλινδρόμησης





Το μέγεθος του S_{YX} θα πρέπει πάντα να κρίνεται σε σχέση με το μέγεθος των τιμών Y στα δεδομένα του δείγματος

,

Έλεγχος της κλίσης ...

Εάν δεν υπάρχει γραμμική σχέση μεταξύ δύο μεταβλητών, θα περιμέναμε η ευθεία παλινδρόμησης να είναι ***οριζόντια***, δηλαδή, να έχουμε ***μηδενική κλίση***.

Θέλουμε να δούμε εάν υπάρχει γραμμική σχέση, δηλαδή να δούμε εάν η κλίση (β_1) είναι διαφορετική από το μηδέν. Η υπόθεσή μας γίνεται:

$$H_1: \beta_1 \neq 0$$

Άρα η μηδενική υπόθεση είναι:

$$H_0: \beta_1 = 0$$

,

Έλεγχος της κλίσης ...

Ο στατιστικός έλεγχος για την υπόθεσή μας:

$$t = \frac{b_1 - \beta_1}{s_{b_1}}$$

όπου s_{b_1} είναι η τυπική απόκλιση του b_1, ορισμένη ως:

$$s_{b_1} = \frac{s_\varepsilon}{\sqrt{(n-1)s_x^2}}$$

Εάν το σφάλμα (ε) ακολουθεί κανονική κατανομή, ο έλεγχος ακολουθεί την **t**-κατανομή με n–2 βαθμούς ελευθερίας. Η περιοχή απόρριψης εξαρτάται από το εάν έχουμε μονόπλευρο ή αμφίπλευρο έλεγχο (συνήθως έχουμε αμφίπλευρο).

,

Παράδειγμα 16.2…

Να ελέγξετε εάν υπάρχει γραμμική σχέση μεταξύ της τιμής και της ένδειξης του κοντέρ. (5% επίπεδο σημαντικότητας)

Έχουμε:

$H_1: \beta_1 \neq 0$

$H_0: \beta_1 = 0$

(εάν η μηδενική υπόθεση ισχύει, δεν υπάρχει γραμμική σχέση)

Η περιοχή απόρριψης είναι:

$$t < -t_{\alpha/2,\nu} = -t_{.025,98} \approx -1.984 \quad or \quad t > t_{\alpha/2,\nu} = t_{.025,98} \approx 1.984$$

,

Παράδειγμα 16.2...

ΥΠΟΛΟΓΙΣΜΟΣ

Μπορούμε να υπολογίσουμε το **t** με το χέρι ή με το Excel …

$$t < -t_{\alpha/2,\nu} = -t_{.025,98} \approx -1.984 \quad or \quad t > t_{\alpha/2,\nu} = t_{.025,98} \approx 1.984$$

		Coefficients	Standard Error	t Stat	P-value
17	Intercept	17.25	0.182	94.73	0.0000
18	Odometer	-0.0669	0.0050	-13.44	0.0000

p-τιμή

Συγκρίνουμε

Βλέπουμε ότι το **t** για το "κοντέρ" (δηλ. την κλίση b_1) είναι –13.44 που είναι μεγαλύτερο από το $\mathbf{t}_{\text{Critical}} = -1.984$. Παρατηρούμε ότι η p-τιμή is 0.000.

Υπάρχουν πολλά στοιχεία που οδηγούν στο ότι υπάρχει γραμμική σχέση μεταξύ της τιμής και της ένδειξης του κοντέρ.

,

Έλεγχος της κλίσης ...

Εάν θέλουμε να ελέγξουμε για ***θετική*** ή ***αρνητική*** γραμμική σχέση κάνουμε μονόπλευρους ελέγχους, δηλαδή οι υποθέσεις μας είναι:

$H_1: \beta_1 < 0$ (έλεγχος για αρνητική κλίση)

ή

$H_1: \beta_1 > 0$ (έλεγχος για θετική κλίση)

Φυσικά, η μηδενική υπόθεση παραμένει: $H_0: \beta_1 = 0$.

,

Συντελεστής Προσδιορισμού ...

Μέχρι τώρα οι έλεγχοι δείχνουν εάν ***υπάρχει*** μια γραμμική σχέση. Είναι χρήσιμο να μετρήσουμε και το ***πόσο ισχυρή είναι αυτή η σχέση***. Αυτό γίνεται υπολογίζοντας τον ***συντελεστή προσδιορισμού*** R^2.

$$R^2 = \frac{s_{xy}^2}{s_x^2 s_y^2} \quad \text{ή} \quad R^2 = 1 - \frac{SSE}{\sum (y_i - \bar{y})^2}$$

Ο συντελεστής προσδιορισμού είναι το τετράγωνο του συντελεστή συσχέτισης (**r**), συνεπώς $R^2 = (r)^2$

,

Συντελεστής Προσδιορισμού ...

Όπως είδαμε στην ανάλυση διασποράς, μπορούμε να χωρίσουμε την μεταβλητότητα του y σε δύο μέρη:

Μεταβλητότητα του y (SST) = SSE + SSR

SSE – Sum of Squares Error – μέτρο της μεταβλητότητας του y που παραμένει ανεξήγητη (Άθροισμα Τετραγώνων Σφάλματος)

SSR – Sum of Squares Regression – μέτρο της μεταβλητότητας του y που εξηγείται από τη μεταβλητότητα της ***ανεξάρτητης μεταβλητής*** x. (Άθροισμα Τετραγώνων Παλινδρόμησης)

Συντελεστής Προσδιορισμού ΥΠΟΛΟΓΙΣΜΟΣ

Το υπολογίζουμε με το χέρι ή με το Excel…

$$s_{xy} = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}\right] = \frac{1}{100-1}\left[53{,}155.9 - \frac{(3{,}601.1)(1{,}484.1}{100}\right] = -2.909$$

$$R^2 = \frac{s_{xy}^2}{s_x^2 \cdot s_y^2} \quad \text{ή} \quad R^2 = 1 - \frac{SSE}{SST}$$

$$R^2 = \frac{s_{xy}^2}{s_x^2 s_y^2} = \frac{(-2.909)^2}{(43.509)(.3000)} = .6483$$

	A	B
1	SUMMARY OUTPUT	
2		
3	*Regression Statistics*	
4	Multiple R	0.8052
5	R Square	0.6483
6	Adjusted R Square	0.6447
7	Standard Error	0.3265
8	Observations	100

$$s_x^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right] = \frac{1}{100-1}\left[133{,}986.59 - \frac{(3{,}601.1)^2}{100}\right] = 43.509$$

$$s_y^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}\right] = \frac{1}{100-1}\left[22{,}055.23 - \frac{(1{,}484.1)^2}{100}\right] = .3000$$

,

Συντελεστής Προσδιορισμού

ΕΡΜΗΝΕΙΑ

R^2 έχει τιμή .6483. Άρα το 64.83% της μεταβλητότητας των τιμών πώλησης (y) ερμηνεύεται από τη μεταβλητότητα των ενδείξεων του κοντέρ (x). Το υπόλοιπο 35.17% είναι ***ανεξήγητο***, δηλαδή οφείλεται σε σφάλμα.

Αντίθετα από τους ελέγχους, ο ***συντελεστής προσδιορισμού*** **δεν** έχει ***κρίσιμη τιμή*** η οποία να μας επιτρέπει να βγάλουμε συμπεράσματα.

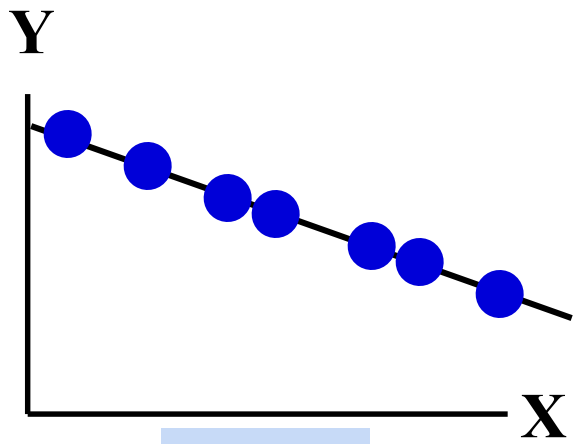
Γενικά, όσο μεγαλύτερη είναι η τιμή του R^2, τόσο ***καλύτερα*** το μοντέλο προσαρμόζεται στα δεδομένα.

$R^2 = 1$: Απόλυτη ταύτιση της ευθείας και των δεδομένων.

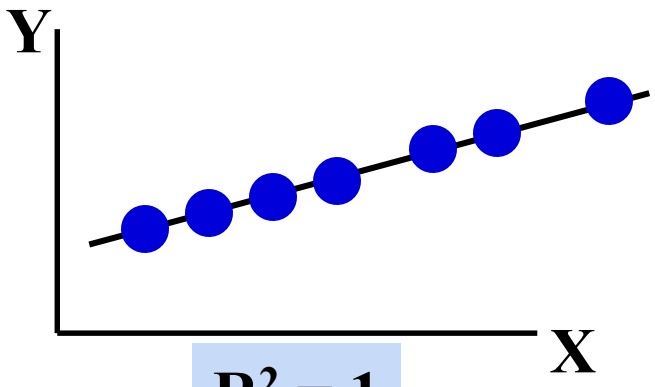
$R^2 = 0$: Δεν υπάρχει γραμμική σχέση μεταξύ x και y.

,

Παραδείγματα Προσεγγιστικών R^2 Τιμών



$R^2 = 1$

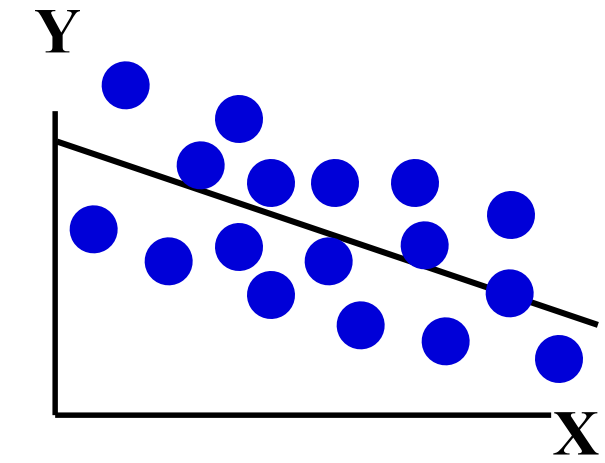


$R^2 = 1$

Ισχυρή γραμμική σχέση μεταξύ X και Y:

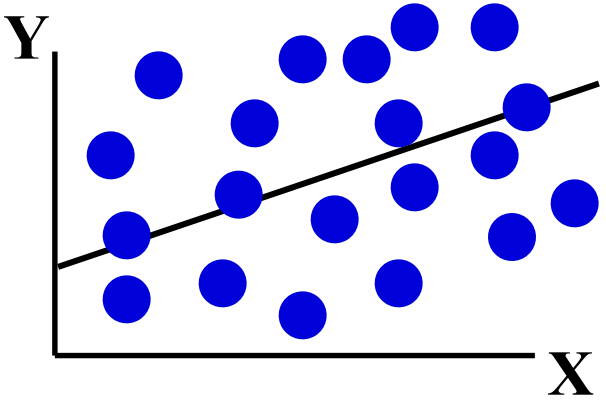
100% της μεταβλητότητας του Y εξηγείται από την μεταβλητότητα του X

Παραδείγματα Προσεγγιστικών R^2 Τιμών



$0 < R^2 < 1$

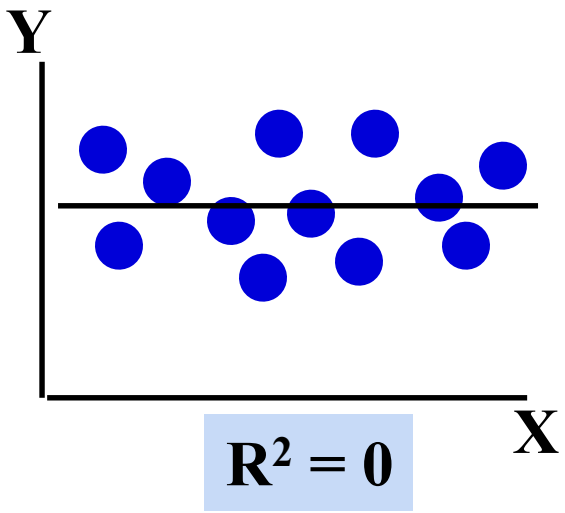
Πιο αδύναμες γραμμικές σχέσεις μεταξύ X και Y:



Ορισμένες αλλά όχι όλες οι μεταβλητότητες στο Y εξηγούνται από την μεταβλητότητα στο X

,

Παραδείγματα Προσεγγιστικών R^2 Τιμών



$R^2 = 0$

$R^2 = 0$

Δεν υπάρχει γραμμική σχέση μεταξύ X και Y:

Η τιμή του Y δεν εξαρτάται από το X. (Καμία από τις μεταβλητότητες του Y δεν εξηγείται από τη μεταβλητότητα στο X)

,

Συντελεστής Συσχέτισης

Μπορούμε να χρησιμοποιήσουμε τον ***συντελεστή συσχέτισης*** για να ελέγξουμε την ύπαρξη γραμμικής σχέσης μεταξύ δύο μεταβλητών.

Θυμίζουμε:

Το εύρος του συντελεστή συσχέτισης είναι

μεταξύ –1 and +1.

- Αν r = –1 (αρνητική συσχέτιση) ή r = +1 (θετική συσχέτιση) κάθε σημείο είναι πάνω στην ευθεία παλινδρόμησης.
- Αν r = 0 δεν υπάρχει γραμμικό μοτίβο

,

Συντελεστής Συσχέτισης

Ο συντελεστής συσχέτισης του ***πληθυσμού*** συμβολίζεται με ρ (rho)

Εκτιμάμε την τιμή του από τα δεδομένα με τον ***συντελεστή συσχέτισης του δείγματος***:

$$r = \frac{s_{xy}}{s_x s_y}$$

Ό στατιστικός έλεγχος για $\rho = 0$ είναι:

$$t = r\sqrt{\frac{n-2}{1-r^2}},\ \nu = n-2$$

δηλαδή **t**-κατανομή με n–2 βαθμούς ελευθερίας.

,

Παράδειγμα 16.2...

Μπορούμε να διεξάγουμε **t**-έλεγχο του ***συντελεστή συσχέτισης*** για να καθορίσουμε με διαφορετικό τρόπο εάν η τιμή πώλησης και η ένδειξη του κοντέρ είναι ***γραμμικά εξαρτημένες***.

Η υπόθεσή μας είναι:

$$H_1: \rho \neq 0$$

(δηλ. υπάρχει γραμμική σχέση) και η μηδενική υπόθεση είναι:

$$H_0: \rho = 0$$

(δηλ. δεν υπάρχει γραμμική σχέση)

,

Παράδειγμα 16.2…

ΥΠΟΛΟΓΙΣΜΟΣ

Έχουμε ήδη αποδείξει ότι:

$$s_{xy} = -2.909 \qquad s_x^2 = 43.509 \Rightarrow s_x = \sqrt{43.509} = 6.596$$
$$s_y^2 = .3000 \Rightarrow s_y = \sqrt{.3000} = .5477$$

Επομένως ο συντελεστής συσχέτισης είναι:

$$r = \frac{s_{xy}}{s_x s_y} = \frac{-2.909}{(6.596)(.5477)} = -.8052$$

και η τιμή για τον έλεγχο γίνεται:

$$t = r\sqrt{\frac{n-2}{1-r^2}} = -.8052\sqrt{\frac{100-2}{1-(-.8052)^2}} = -13.44$$

,

Παράδειγμα 16.2…

ΥΠΟΛΟΓΙΣΜΟΣ

Μπορούμε να χρησιμοποιήσουμε και το Excel παίρνοντας το output:

	A	B
1	**Correlation**	
2		
3	*Price and Odometer*	
4	Pearson Coefficient of Correlation	-0.8052
5	t Stat	-13.44
6	df	98
7	P(T<=t) one tail	0
8	t Critical one tail	1.6606
9	P(T<=t) two tail	0
10	t Critical two tail	1.9845

Μπορούμε επίσης να κάνουμε μονόπλευρο έλεγχο για θετική ή αρνητική γραμμική σχέση

p-τιμή

συγκρίνουμε

Πάλι, απορρίπτουμε τη μηδενική υπόθεση (ότι δεν υπάρχει γραμμική συσχέτιση) και αποδεχόμαστε την εναλλακτική (ότι οι μεταβλητές μας συνδέονται με γραμμικό τρόπο).

,

Χρήση της εξίσωσης Παλινδρόμησης ...

Θα μπορούσαμε να χρησιμοποιήσουμε την εξίσωση παλινδρόμησης:

$$\hat{y} = 17.250 - .0669x$$

για να προβλέψουμε την τιμή πώλησης ενός αυτοκινήτου με 40 (,000) μίλια:

$$\hat{y} = 17.250 - .0669x = 17.250 - .0669(40) = 14,574$$

Καλούμε αυτήν την τιμή ($14,574) ***σημειακή πρόβλεψη***. Μάλλον όμως η πραγματική τιμή πώλησης θα είναι διαφορετική …

,

Προϋποθέσεις Παλινδρόμησης...

Υπάρχουν τρεις προϋποθέσεις για την εφαρμογή της ανάλυσης παλινδρόμησης. Αυτές είναι:

- Η μεταβλητή του σφάλματος πρέπει να ακολουθεί την κανονική κατανομή,
- Η μεταβλητή του σφάλματος πρέπει να έχει σταθερή διασπορά, &
- Τα σφάλματα πρέπει να είναι ανεξάρτητα μεταξύ τους.

Πώς μπορούμε να διαγνώσουμε περιπτώσεις όπου αυτές οι προϋποθέσεις δεν ισχύουν;

➔Η **Ανάλυση Υπολοίπων** εξετάζει τις ***διαφορές*** ανάμεσα στα πραγματικά δεδομένα και σε αυτά που προβλέπονται ἀπό την εξίσωση παλινδρόμησης…

Ανάλυση Υπολοίπων...

Θυμίζουμε ότι οι παρεκκλίσεις των σημείων των πραγματικών δεδομένων από τη γραμμή παλινδρόμησης λέγονται ***υπόλοιπα***. Το Excel υπολογίζει υπόλοιπα σαν μέρος της ανάλυσης παλινδρόμησης:

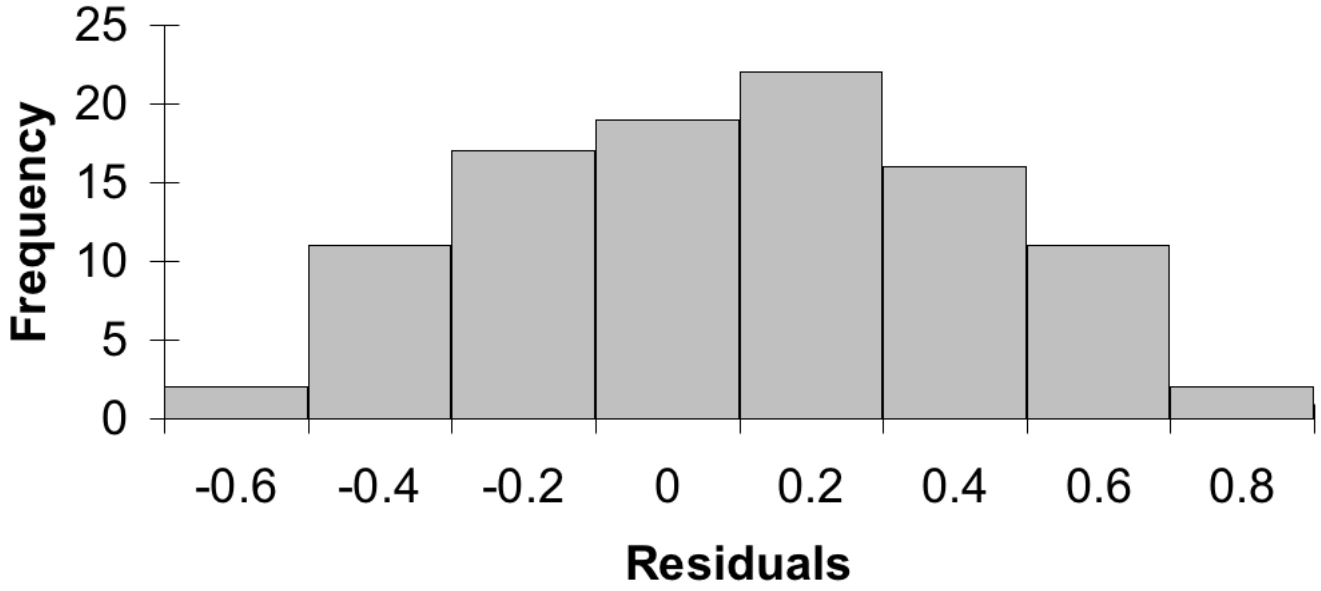
ResidualsExample.xls

	A	B	C	D
1	RESIDUAL OUTPUT			
2				
3	*Observation*	*Predicted Price*	*Residuals*	*Standard Residuals*
4	1	14.739	-0.139	-0.451
5	2	14.282	-0.182	-0.589
6	3	14.220	-0.220	-0.712
7	4	15.141	0.459	1.488
8	5	15.092	0.508	1.648

Μπορούμε να χρησιμοποιήσουμε αυτά τα υπόλοιπα για να εξετάσουμε αν η μεταβλητή σφάλματος είναι μη κανονική, αν η διασπορά του σφάλματος είναι σταθερή και αν τα σφάλματα είναι ανεξάρτητα…

Μη κανονικότητα...

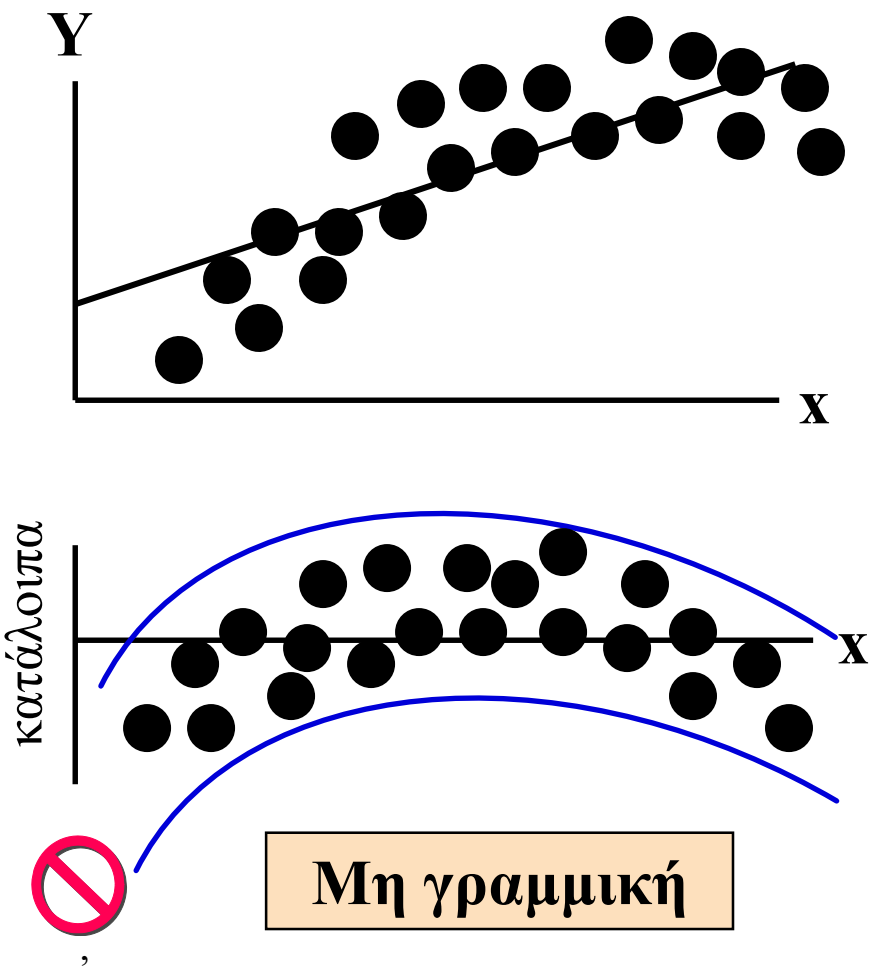
Μπορούμε να παραστήσουμε τα υπόλοιπα σε ένα ιστόγραμμα για να ελέγξουμε γραφικά την κανονικότητα…

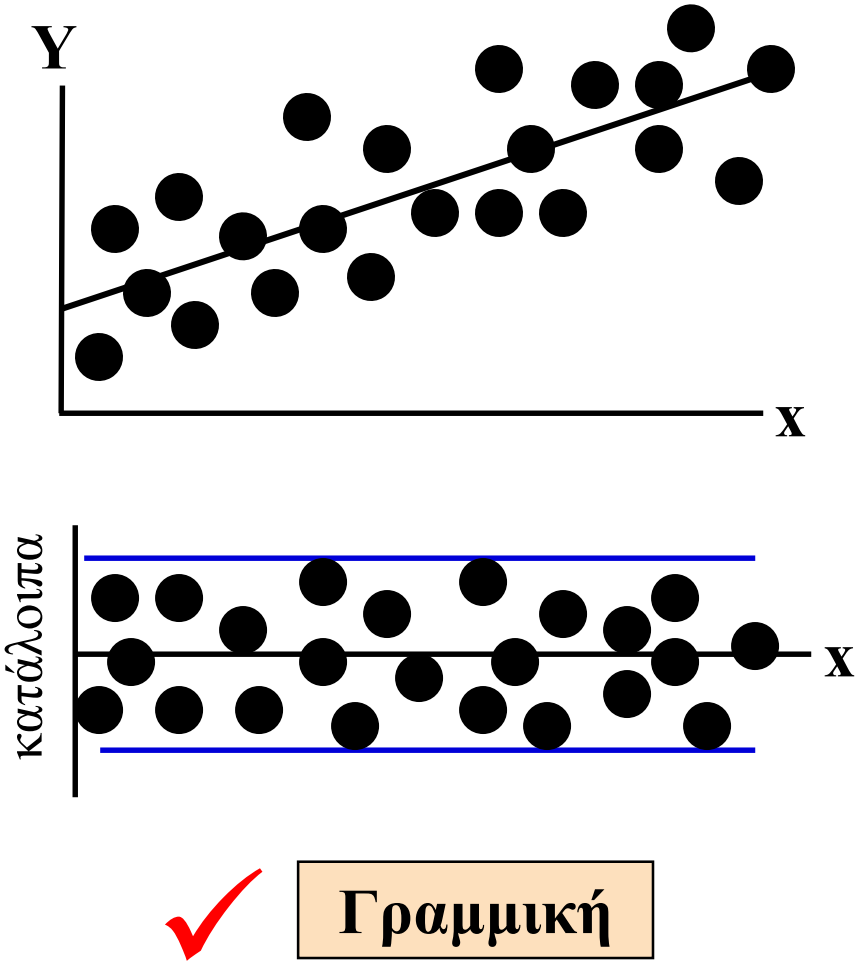


…ψάχνουμε για ένα ιστόγραμμα σε σχήμα καμπάνας με τον μέσο κοντά στο 0. ✓

,

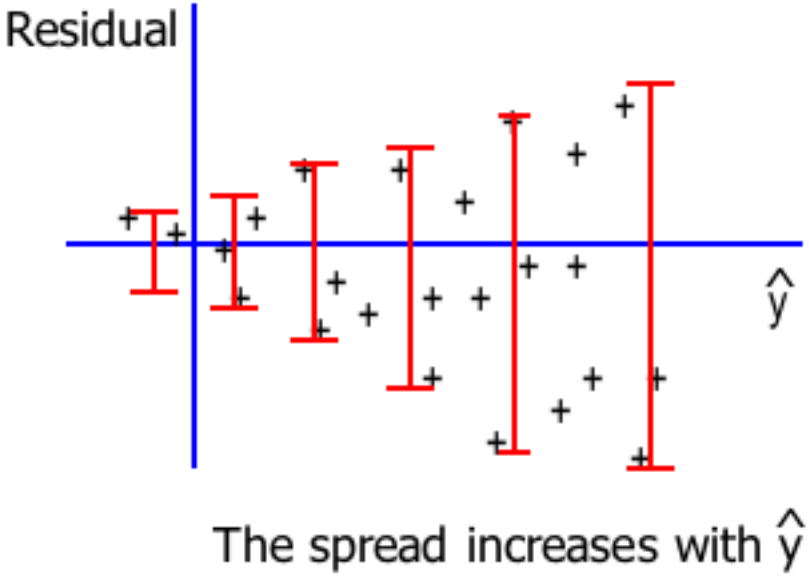
Ανάλυση Καταλοίπων για Γραμμικότητα





Ετεροσκεδαστικότητα...

Όταν η προϋπόθεση της σταθερής διασποράς δεν τηρείται, έχουμε μια κατάσταση ***ετεροσκεδαστικότητας***.

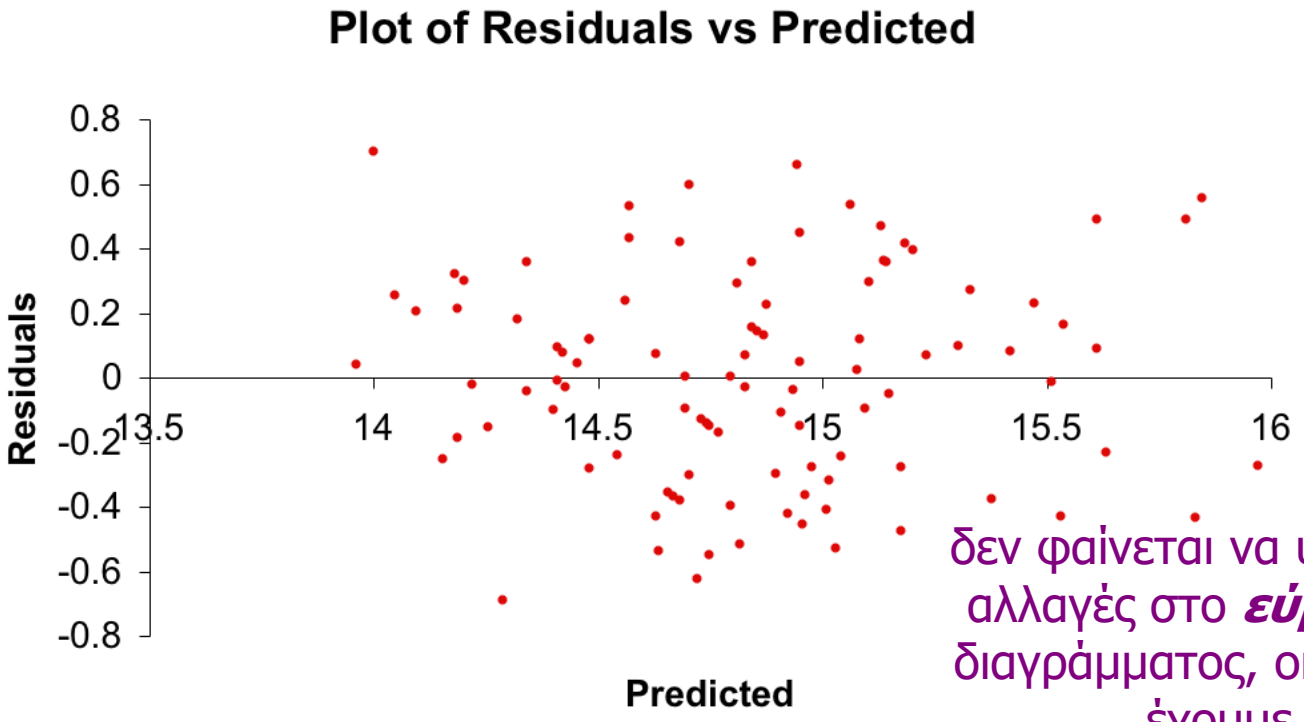


Μπορούμε να διαγνώσουμε ετεροσκεδαστικότητα κατασκευάζοντας διάγραμμα διασποράς με τα υπόλοιπα και τις προβλεπόμενες τιμές της y.

,

Ετεροσκεδαστικότητα...

Αν η διασπορά της μεταβλητής του σ_ε^2 σφάλματος δεν είναι σταθερή, τότε έχουμε "***ετεροσκεδαστικότητα***". Εδώ βλέπουμε το διάγραμμα διασποράς των υπολοίπων με τις προβλεπόμενες τιμές της **y**:



δεν φαίνεται να υπάρχουν αλλαγές στο ***εύρος*** του διαγράμματος, οπότε δεν έχουμε ***ετεροσκεδαστικότητα*** ✓

,

Μη ανεξαρτησία της μεταβλητής σφάλματος

Αν είχαμε να καταγράψουμε τη τιμή πλειστηριασμού των αυτοκινήτων κάθε εβδομάδα για έναν ολόκληρο χρόνο, αυτό θα συνιστούσε ***μία χρονολογική σειρά***.

Όταν τα δεδομένα είναι χρονολογικές σειρές, τα σφάλματα συχνά ***συσχετίζονται***.

Τιμές σφάλματος που συνδέονται μεταξύ τους σε μια χρονολογική σειρά λέμε ότι εμφανίζουν ***αυτοσυσχέτιση***

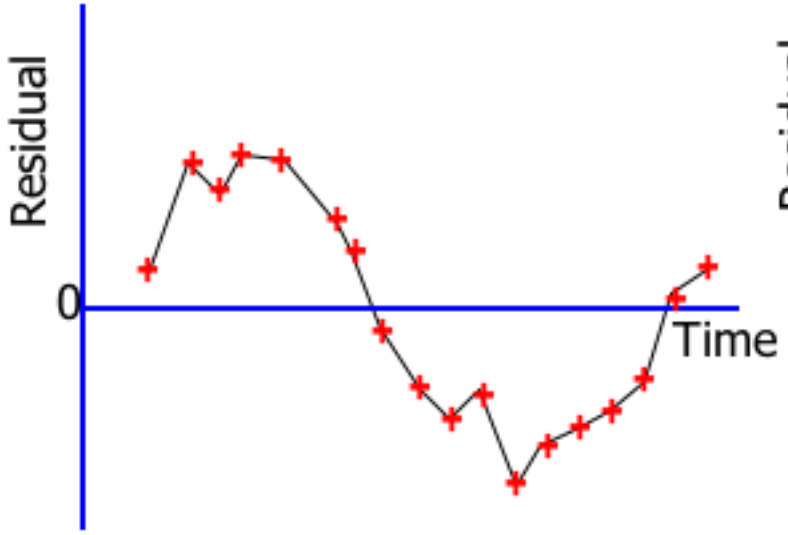
(autocorrelated) ή ***σειριακή συσχέτιση*** (***serially correlated)***.

Μπορούμε συχνά να διαγνώσουμε αυτοσυσχέτιση ***κατασκευάζοντας διάγραμμα των υπολοίπων ενάντια στις χρονικές περιόδους***. Αν στο διάγραμμα έχουμε κάποιο επαναλαμβανόμενο σχήμα, είναι πιθανό να μην υπάρχει ανεξαρτησία.

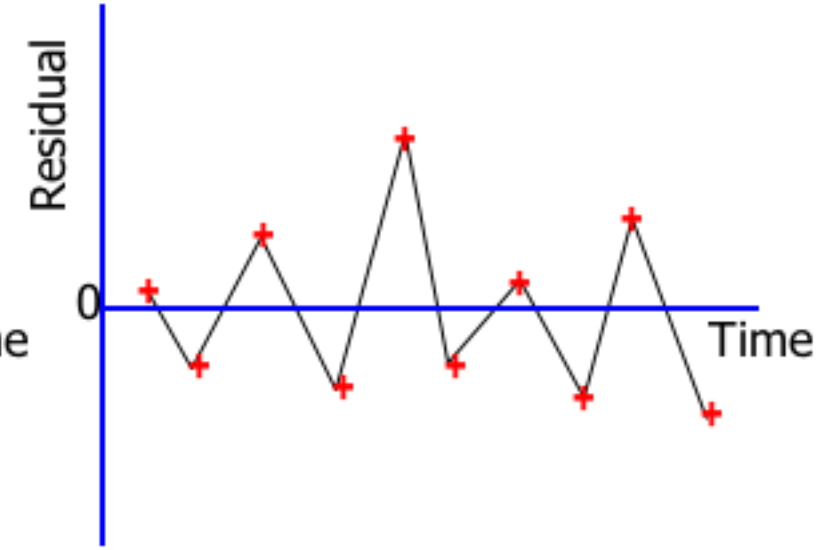
,

Μη Ανεξαρτησία της Μεταβλητής Σφάλματος

Τα σχήματα στην παράσταση των υπολοίπων σε σχέση με τον χρόνο φανερώνουν ότι υπάρχει αυτοσυσχέτιση:



Σημειώστε τη ροή των θετικών υπολοίπων, που αντικαθίσταται από ροή αρνητικών υπολοίπων



Σημειώστε την ταλάντωση των υπολοίπ γύρω από το μηδέν

,

Ακραίες τιμές...

Μία ***ακραία τιμή*** είναι μια παρατήρηση που είναι ***ασυνήθιστα μικρή*** ή ***ασυνήθιστα μεγάλη***.

Π.χ. στο παράδειγμα με τα αυτοκίνητα οι ενδείξεις των οδομέτρων ήταν από 19.1 έως 49.2 χιλιάδες μίλια. Αν είχαμε μια τιμή μόλις 5,000 μιλίων (π.χ. ένα αυτοκίνητο που οδηγείται από ηλικιωμένο μόνο τις Κυριακές ☺) — αυτό το σημείο είναι μια ***ακραία τιμή***.

,

Ακραίες τιμές...

Πιθανές αιτίες της ύπαρξης ακραίων τιμών είναι:

Σφάλμα στην καταγραφή της τιμής

Η παρατήρηση δεν έπρεπε να συμπεριληφθεί στο δείγμα

Ίσως η παρατήρηση είναι πράγματι έγκυρη.

Οι ακραίες τιμές είναι εύκολο να ανιχνευτούν σε ένα διάγραμμα διασποράς.

Αν η απόλυτη τιμή του τυποποιημένου υπολοίπου είναι > 2, υποπτευόμαστε ότι το σημείο μπορεί να είναι ακραία τιμή οπότε να την διερευνήσουμε περαιτέρω.

Πρέπει να διερευνηθούν γιατί μπορεί εύκολα να επηρεάσουν την ευθεία ελαχίστων τετραγώνων…

,

Διαδικασία Ανάλυσης Παλινδρόμησης...

1. Ανάπτυξη ενός μοντέλου με θεωρητική βάση.
2. Συλλογή δεδομένων για τις δύο μεταβλητές του μοντέλου.
3. Σχεδίαση του διαγράμματος διασποράς για να δούμε αν υπάρχει κατάλληλο γραμμικό μοντέλο και να αναγνωρίσουμε πιθανές ακραίες τιμές.
4. Καθορισμός της εξίσωσης παλινδρόμησης.
5. Υπολογισμός των υπολοίπων και έλεγχος των προϋποθέσεων
6. Αξιολόγηση του μοντέλου.
7. ***Αν το μοντέλο ταιριάζει στα δεδομένα***, ***χρήση της εξίσωσης παλινδρόμησης*** για την πρόβλεψη των τιμών της εξαρτημένης μεταβλητής και/ή εκτίμηση του μέσου τους.

,