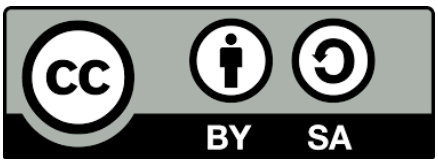
ΠΑΝΕΠΙΣΤΗΜΙΟ ΙΩΑΝΝΙΝΩΝ
ΑΝΟΙΚΤΑ ΑΚΑΔΗΜΑΪΚΑ ΜΑΘΗΜΑΤΑ

Τίτλος Μαθήματος: Στατιστική Ανάλυση Δεδομένων

Ενότητα: □□□/□

Διδάσκων: Επίκ. Καθ. Απόστολος Μπατσίδης

Τμήμα: Μαθηματικών



Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

ΚΕΦΑΛΑΙΟ ΟΓΔΟΟ

Γραμμική παλινδρόμηση

Σε προηγούμενο κεφάλαιο είδαμε ότι η γραφική παράσταση δύο μεταβλητών είναι ένα πρώτο βήμα για τη διαπίστωση της ύπαρξης μίας σχέσης μεταξύ δύο μεταβλητών. Στην παλινδρόμηση το ενδιαφέρον επικεντρώνεται στην εύρεση του καλύτερου γραμμικού μοντέλου που μας δείχνει τον τρόπο με τον οποίο p το πλήθος ανεξάρτητες μεταβλητές επιδρούν σε μία ποσοτική μεταβλητή. Αναζητούμε, επομένως, το μαθηματικό μοντέλο που περιγράφει με τον καλύτερο δυνατό τρόπο τις τιμές της εξαρτημένης μεταβλητής συναρτήσει των τιμών των ανεξάρτητων μεταβλητών. Η εύρεση ενός τέτοιου μοντέλου μας δίνει τη δυνατότητα τόσο να μοντελοποιήσουμε ένα φυσικό-τυχαίο φαινόμενο όσο και να κάνουμε προβλέψεις για τις τιμές της εξαρτημένης μεταβλητής όταν οι ανεξάρτητες θεωρούνται δεδομένες.

Όταν έχουμε μόνο μία ανεξάρτητη μεταβλητή λέμε ότι έχουμε το μοντέλο της απλής γραμμικής παλινδρόμησης. Το μοντέλο αυτό χρησιμοποιείται για την πρόβλεψη των τιμών μίας εξαρτημένης μεταβλητής από τις τιμές μίας ανεξάρτητης μεταβλητής, όταν αυτές είναι συσχετισμένες. Η ανεξάρτητη μεταβλητή μπορεί να είναι είτε κατηγορική είτε συνεχής, ενώ η εξαρτημένη είναι συνεχής. Γενίκευση του μοντέλου της απλής γραμμικής παλινδρόμησης για p το πλήθος ανεξάρτητες μεταβλητές αποτελεί η πολλαπλή παλινδρόμηση.

Σχόλιο: Μία ανεξάρτητη κατηγορική μεταβλητή με k κατηγορίες- τιμές υπεισέρχεται στο μοντέλο της γραμμικής παλινδρόμησης με τη χρήση k-1 δείκτριων μεταβλητών, ενώ όταν η εξαρτημένη μεταβλητή είναι κατηγορική τότε χρησιμοποιούνται μεθοδολογίες της Λογιστικής Παλινδρόμησης. Οι μεθοδολογίες αυτές ξεφεύγουν από το σκοπό αυτών των σημειώσεων.

8.1 Προσαρμογή του μοντέλου της απλής γραμμικής παλινδρόμησης

Στην ενότητα αυτή θα περιγράψουμε τη μεθοδολογία που ακολουθείται για την προσαρμογή ενός μοντέλου απλής γραμμικής παλινδρόμησης.

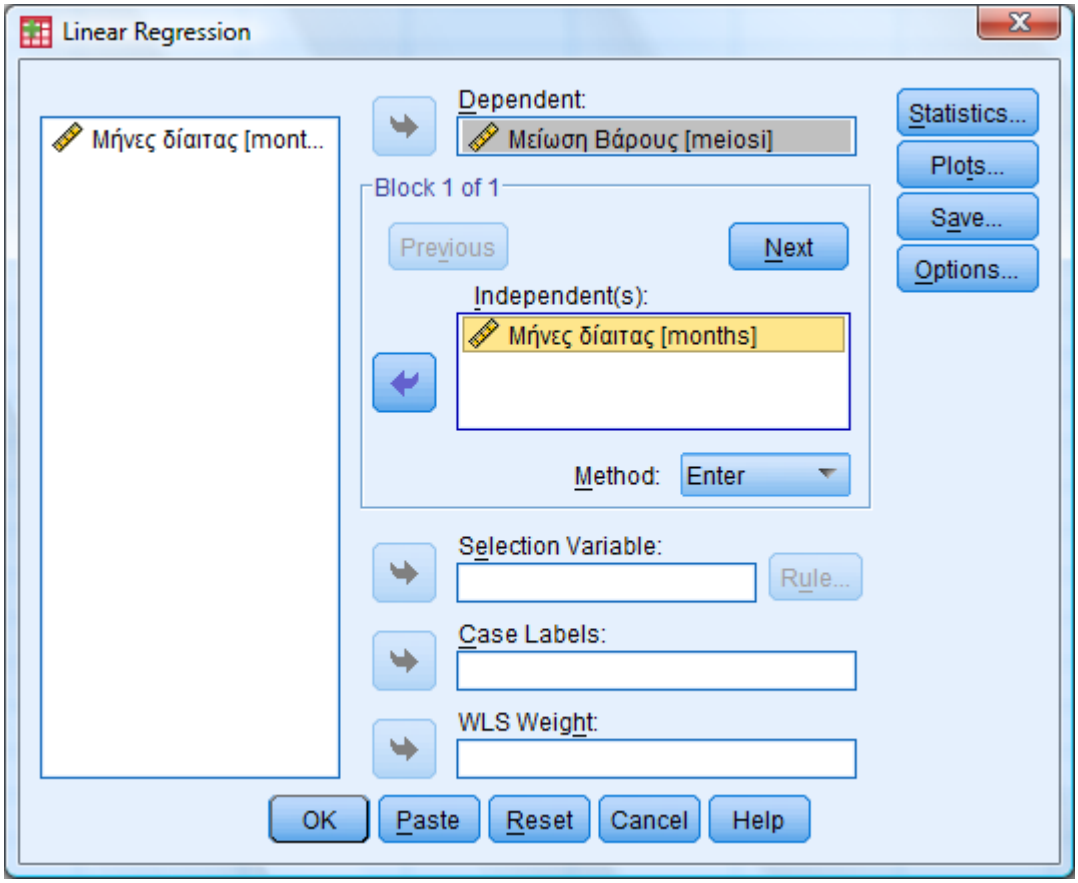
Υλοποίηση στο S.P.S.S. (βλέπε Καρακώστας, 2002, σελ. 22)

Οι παρακάτω τιμές είναι το βάρος (σε λίβρες) που έχασαν 10 άτομα αφού ακολούθησαν κάποια δίαιτα για ορισμένους μήνες. Είναι δυνατή η πρόβλεψη της απώλειας βάρους από τους μήνες δίαιτας.

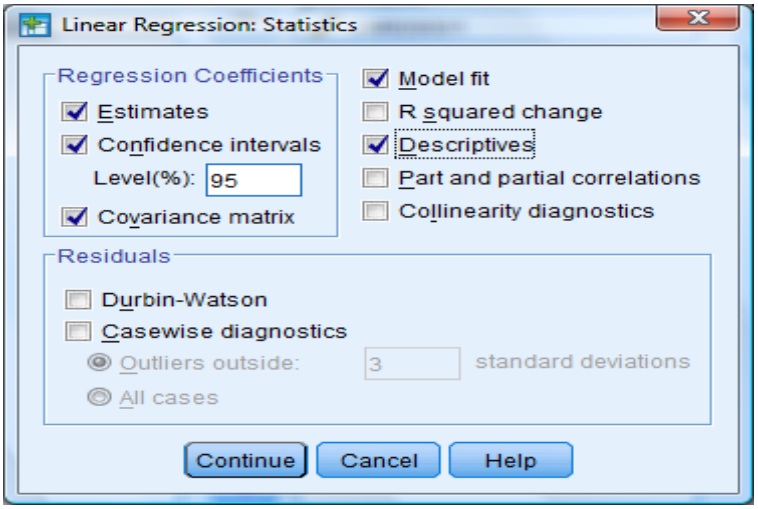
Μήνες Δίαιτας	Μείωση Βάρους
4	17
17	64
14	53
1	1
10	45
22	71
9	38
12	40
4	11
7	24

1. Το πρώτο βήμα για την ανάλυση του παραπάνω προβλήματος είναι να ορίσουμε ποια είναι η εξαρτημένη και ποια η ανεξάρτητη μεταβλητή. Προφανώς το ρόλο της ανεξάρτητης παίζει η μεταβλητή Μήνες δίαιτας, ενώ το ρόλο της εξαρτημένης η Μείωση Βάρους.
2. Θέλοντας να διαπιστώσουμε αν η προσαρμογή του μοντέλου της απλής γραμμικής παλινδρόμησης αιτιολογείται προβαίνουμε στη γραφική παράσταση των δεδομένων της εξαρτημένης ως προς την ανεξάρτητη (βλέπε παράγραφο για Διάγραμμα Διασποράς). Αν η γραφική αυτή παράσταση μας υποδεικνύει ότι η σχέση των δύο μεταβλητών δεν είναι γραμμική, τότε η υιοθέτηση του μοντέλου της απλής γραμμικής παλινδρόμησης είναι λανθασμένη. Τρόποι αντιμετώπισης αυτού του προβλήματος αναφέρονται στην επόμενη παράγραφο και στην ενότητα «Ορθότητα μοντέλου».
3. Προχωρούμε έπειτα στην προσαρμογή του μοντέλου της απλής γραμμικής παλινδρόμησης επιλέγοντας από το αρχικό παράθυρο του στατιστικού πακέτου S.P.S.S.: **Analyze→Regression→Linear**. Στο νέο παράθυρο διαλόγου που προκύπτει τοποθετείται η Μείωση Βάρους ως εξαρτημένη μεταβλητή (Dependent) και οι Μήνες δίαιτας ως

ανεξάρτητη μεταβλητή (Independent), αντίστοιχα, ενώ στο πεδίο Method επιβεβαιώνουμε ότι η επιλογή Enter έχει καθοριστεί.

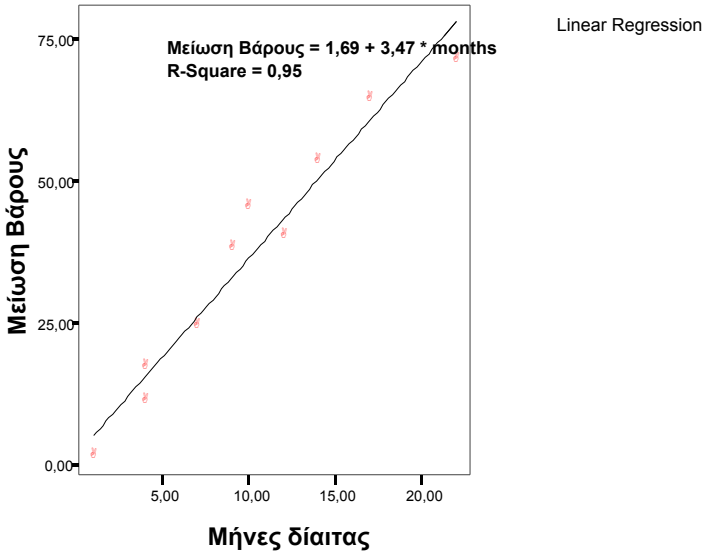


4. Από την επιλογή Statistics επιλέγουμε, προς το παρόν, τα ακόλουθα, τα αποτελέσματα των οποίων θα δούμε μέσω της ερμηνεία των αποτελεσμάτων, πατάμε Continue και OK:



Ερμηνεία αποτελεσμάτων

Η γραφική παράσταση που προκύπτει, αν ζητήσουμε να προσαρμοστεί και η ευθεία της γραμμικής παλινδρόμησης (Elements Fit Line at Total) είναι η ακόλουθη:



Παρατηρούμε ότι η γραφική αυτή παράσταση μας δείχνει ότι η σχέση των δύο μεταβλητών είναι γραμμική σε αρκετά ικανοποιητικό βαθμό και επομένως είναι λογικό να προσαρμόσουμε το μοντέλο της απλής γραμμικής παλινδρόμησης. Στο ίδιο συμπέρασμα καταλήγουμε ερμηνεύοντας και το αποτέλεσμα για το συντελεστή συσχέτισης του Pearson (βλέπε πίνακα Correlations, r=0.976, p-τιμή<0,001), παρότι θα πρέπει να είμαστε επιφυλακτικοί καθώς (όπως έχει ήδη αναφερθεί στο 3^{ο} Κεφάλαιο) αυτός επηρεάζεται από την ύπαρξη ακραίων τιμών, ενώ ο στατιστικός έλεγχος αν υπάρχει στατιστικά σημαντική γραμμική συσχέτιση μεταξύ της μείωσης βάρους και του αριθμού των μηνών που διεξήχθη η δίαιτα υποθέτει την ύπαρξη διδιάστατης κανονικότητας.

Correlations

		Μείωση Βάρους	Μήνες δίαιτας
Pearson Correlation	Μείωση Βάρους	1,000	,976
	Μήνες δίαιτας	,976	1,000
Sig. (1-tailed)	Μείωση Βάρους	.	,000
	Μήνες δίαιτας	,000	.
N	Μείωση Βάρους	10	10
	Μήνες δίαιτας	10	10

Θέλοντας να κατασκευάσουμε ένα μοντέλο πρόβλεψης της μείωσης του βάρους από τους μήνες δίαιτας προσαρμόζουμε το μοντέλο:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, i = 1,...,10,$$

όπου Y_i η Μείωση Βάρους του i-οστού ατόμου (η μέση απώλεια βάρους είναι 36,4 κιλά και η τυπική απόκλιση 22,97922 κιλά) και X_i οι Μήνες Δίαιτας του i-οστού ατόμου, αντίστοιχα (η μέση διάρκεια δίαιτας είναι 10 μήνες και η τυπική απόκλιση 6,46357 μήνες, βλέπε πίνακα Descriptive Statistics).

Descriptive Statistics

	Mean	Std. Deviation	N
Μείωση Βάρους	36,4000	22,97922	10
Μήνες δίαιτας	10,0000	6,46357	10

Ο έλεγχος της υπόθεσης ότι δεν υπάρχει παλινδρόμηση έδειξε ότι η υπόθεση αυτή απορρίπτεται (βλέπε Πίνακα ΑΝΑΔΙΑ, $F = \frac{MS_{reg}}{MS_{res}} = 162,430, p - \text{τιμή} < 0.001$).

ANOVA(b)

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4529,322	1	4529,322	162,430	,000(a)
	Residual	223,078	8	27,885		
	Total	4752,400	9			

a Predictors: (Constant), Μήνες δίαιτας
b Dependent Variable: Μείωση Βάρους

<u>Σχόλιο</u>: Από τον πίνακα ANOVA έχουμε όλες τις πληροφορίες που περιέχονται σε ένα ΠΙΝΑΚΑ ΑΝΑΔΙΑ: Άθροισμα Τετραγώνων (Sum of Squares) της Παλινδρόμησης (Regression), των Υπολοίπων (Residual), καθώς και συνολικό άθροισμα τετραγώνων (Total), βαθμοί ελευθερίας (df), μέσα τετράγωνα (Mean Square) της παλινδρόμησης και των υπολοίπων, τιμή του F-στατιστικού τεστ για τον έλεγχο της υπόθεσης $\beta_1 = 0$ και αντίστοιχη p-τιμή).

Με τη μέθοδο των ελαχίστων τετραγώνων προκύπτουν, οι ακόλουθοι εκτιμητές (οι λεγόμενοι εκτιμητές ελαχίστων τετραγώνων των παραμέτρων του μοντέλου, στήλη Unstandardized Coefficients B)

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X} = 1.693$$

και

$$\hat{\beta}_1 = \frac{n\sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2} = 3.471 .$$

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	1,693	3,194		,530	,611	-5,674	9,059
	Μήνες δίαιτας	3,471	,272	,976	12,7	,000	2,843	4,099

a. Dependent Variable: Μείωση Βάρους

Το γεγονός αυτό σημαίνει ότι υπό την προϋπόθεση ότι το εκτιμώμενο μοντέλο είναι σωστό ισχύει ότι:

$$\hat{Y} = 1.693 + 3.471X,$$

δηλαδή μπορούμε να πούμε ότι $\hat{\beta}_0 = 1.693$ κιλά είναι η απώλεια βάρους όταν κάποιος δεν κάνει δίαιτα (άρα γίνεται αντιληπτό ότι το μοντέλο με σταθερό όρο δεν είναι λογικό) και $\hat{\beta}_1 = 3.471$ κιλά είναι η απώλεια βάρους που θα έχει κάποιος αν κάνει <u>ένα</u> μήνα περισσότερο δίαιτα (γενικά ισχύει ότι αν $\hat{\beta}_1 > 0$ αύξηση της τιμής της ανεξάρτητης μεταβλητής κατά μία μονάδα επιφέρει αύξηση των τιμών της εξαρτημένης κατά $\hat{\beta}_1$ μονάδες, ενώ όταν $\hat{\beta}_1 < 0$ αύξηση της τιμής της ανεξάρτητης κατά μία μονάδα επιφέρει ελάττωση των τιμών της εξαρτημένης κατά $\hat{\beta}_1$ μονάδες, και θα πρέπει να ελέγχουμε αν τα αποτελέσματα αυτά συμφωνούν με τη φύση του προβλήματος).

<u>Παρατήρηση</u>: Η εκτιμώμενη εξίσωση $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$ δεν θα πρέπει να χρησιμοποιείται για την πρόβλεψη των τιμών της εξαρτημένης μεταβλητής για τιμές της ανεξάρτητης πέρα του