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.

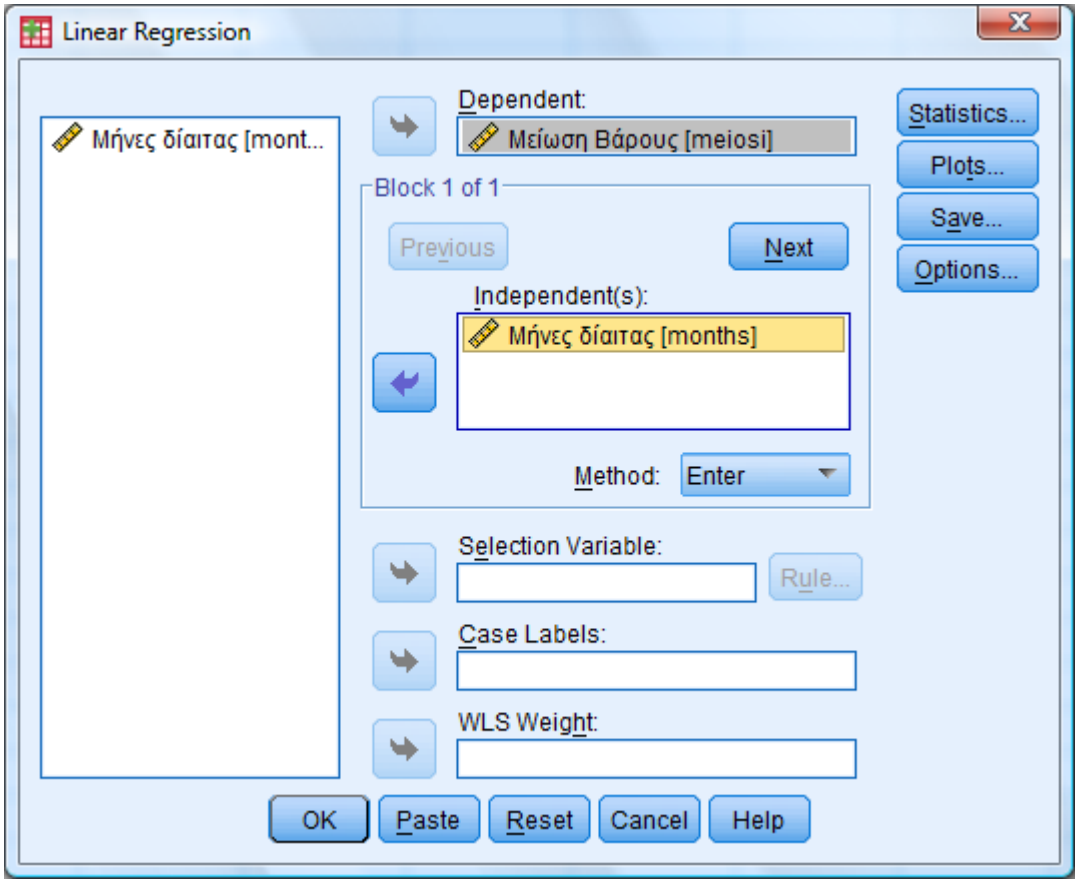
Υλοποίηση στο S.P.S.S. (βλέπε Καρακώστας, 2002, σελ. 22)

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.

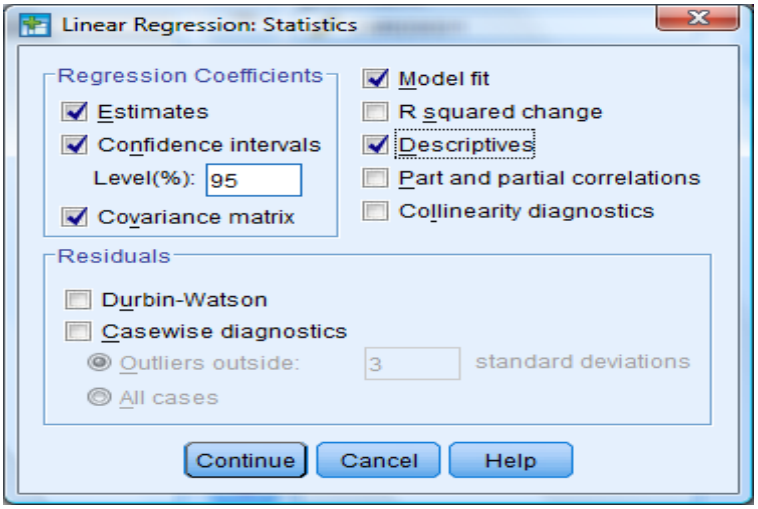
Μήνες Δίαιτας	Μείωση Βάρους
4	17
17	64
14	53
1	1
10	45
22	71
9	38
12	40
4	11
7	24

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.

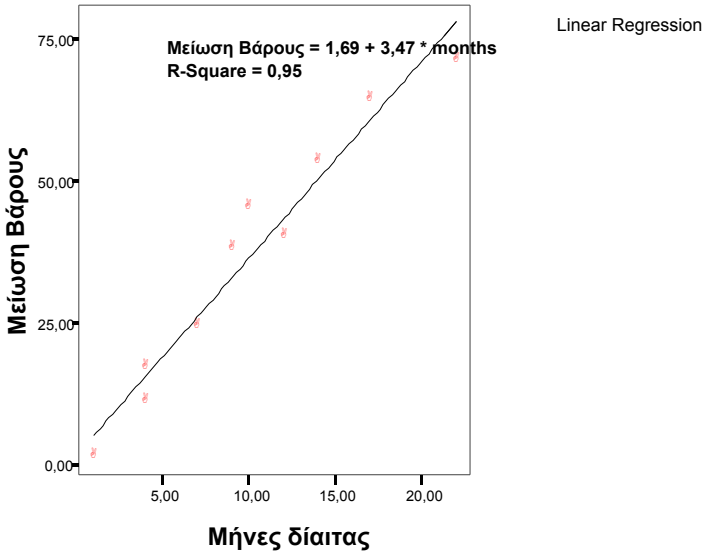


4. Από την επιλογή Statistics επιλέγουμε, προς το παρόν, τα ακόλουθα, τα αποτελέσματα των οποίων θα δούμε μέσω της ερμηνεία των αποτελεσμάτων, πατάμε Continue και OK:



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

πεδίου τιμών αυτής με το οποίο δημιουργήθηκε το μαθηματικό μας μοντέλο. Δηλαδή για το συγκεκριμένο παράδειγμα δε μπορεί να προβλεφθεί η απώλεια βάρους για π.χ. 35 μήνες δίαιτας.

Επιπρόσθετα προκύπτει ότι οι μήνες δίαιτας εξηγούν το 95.3% της μεταβλητότητας της μείωσης βάρους (βλέπε πίνακα Model Summary, $R^2 = SS_{reg} / SS_{tot} = 0.953$). Το αποτέλεσμα αυτό είναι αρκετά ικανοποιητικό.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,976(a)	,953	,947	5,28060

a Predictors: (Constant), Μήνες δίαιτας

<u>Σχόλιο</u>: Για να είναι εφικτή η σύγκριση μοντέλων που έχουν την ίδια εξαρτημένη μεταβλητή και διαφορετικό αριθμό ανεξάρτητων μεταβλητών έχει υιοθετηθεί ο προσαρμοσμένος συντελεστής R^2. Αυτός υπολογίζεται από τη σχέση Adjusted $R^2 = R^2 - \frac{(1-R^2)(k-1)}{(n-k)}$, όπου k το πλήθος των ανεξάρτητων μεταβλητών συμπεριλαμβανομένου του σταθερού όρου και n το μέγεθος του δείγματος.

Επιπλέον, υπό την προϋπόθεση ότι τα σφάλματα ε_i, $i = 1,...,10,$, ακολουθούν κανονική κατανομή με μέση τιμή 0, σταθερή διακύμανση σ^2 και είναι ασυσχέτιστα μεταξύ τους ανά δύο, δηλαδή $Cov(\varepsilon_i, \varepsilon_j) = 0$ για $i, j = 1,...,10,$ με $i \neq j$, υποθέσεις τρόπους ελέγχους των οποίων θα δούμε στην επόμενη ενότητα, ισχύουν τα ακόλουθα:

α) το 95% Διάστημα Εμπιστοσύνης για τους συντελεστές του μοντέλου της παλινδρόμησης είναι (-5.674, 9.0591) και (2.843,4.099), αντίστοιχα (βλέπε στήλη 95% Confidence Interval for B).

β) Επιπλέον, συμπεραίνουμε ότι στο μοντέλο δεν πρέπει να συμπεριληφθεί σταθερός όρος , καθώς δεν απορρίπτεται η υπόθεση $H_0 : \beta_0 = 0 \left(t = \frac{\hat{\beta}_0}{\sqrt{\hat{V}ar(\hat{\beta}_0)}} = 0.530, \text{ p-τιμή=0.611>0.05} \right)$, ενώ δικαιολογείται το μοντέλο της

παλινδρόμησης καθώς απορρίπτεται η μηδενική υπόθεση $H_0 : \beta_1 = 0$ $\left(t = \frac{\hat{\beta}_1}{\sqrt{\hat{V}ar(\hat{\beta}_1)}} = 12.7, \text{ p-τιμή}<0.001 \right)$.

Σχόλιο: Από τον πίνακα Coefficient Correlations υπολογίζονται οι εκτιμητές των $Cov(\hat{\beta}_i, \hat{\beta}_j)$, $i, j = 1,2.$

8.2 Έλεγχος των υποθέσεων της απλής γραμμικής παλινδρόμησης

Το μοντέλο της γραμμικής παλινδρόμησης στηρίζεται στις ακόλουθες υποθέσεις για τα σφάλματα ε_i, $i = 1,...,n$:

α) ακολουθούν κανονική κατανομή (με μέση τιμή 0),

β) έχουν σταθερή διακύμανση σ^2, και

γ) είναι ασυσχέτιστα μεταξύ τους ανά δύο.

Επιπλέον, υποθέτουμε ότι

δ) $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, $i = 1,...,n$, η ανεξάρτητη μεταβλητή συνδέεται με τις εξαρτημένες μέσω της γραμμικής σχέσης

Στις προηγούμενες υποθέσεις θα πρέπει να προσθέσουμε την υπόθεση ότι:
ε) δεν υπάρχουν ακραίες τιμές στα δεδομένα μας και
στ) δεν υπάρχουν επηρεάζουσες παρατηρήσεις.

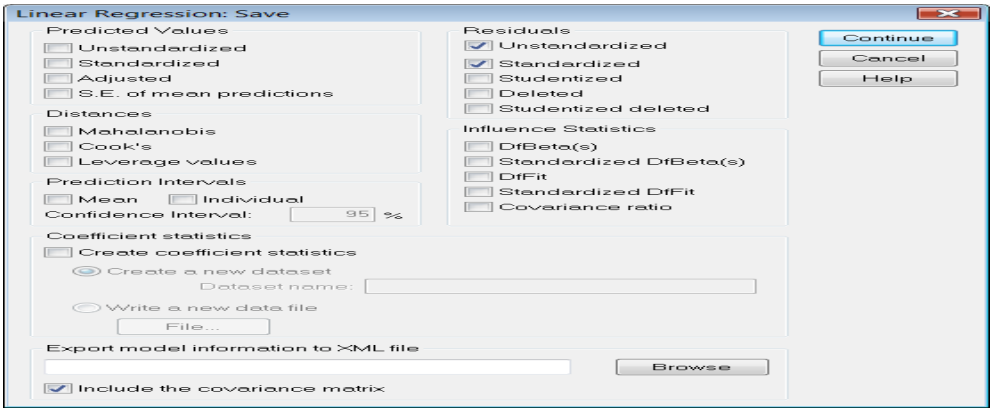
Σε αυτήν την παράγραφο θα υποδείξουμε τρόπους ελέγχου των παραπάνω υποθέσεων. Επιπλέον, θα σχολιάσουμε εν συντομία τα προβλήματα που δημιουργούνται όταν δεν ικανοποιούνται και τέλος θα προτείνουμε ή θα παραπέμψουμε τον αναγνώστη σε λύσεις για την αντιμετώπισή τους.

8.2.1 Έλεγχος κανονικότητας των σφαλμάτων

Ο έλεγχος της κανονικής κατανομής των σφαλμάτων γίνεται με την βοήθεια είτε των υπολοίπων $e_i = Y_i - \hat{Y}_i,\ i = 1,...,n,$ είτε των τυποποιημένων υπολοίπων $e_{si} = \frac{e_i}{\sqrt{MSres}}$. Με τη βοήθεια του S.P.S.S. μπορούμε να προχωρήσουμε τόσο σε γραφικό όσο και σε στατιστικό έλεγχο, χρησιμοποιώντας τις πιο πάνω ποσότητες και τη διαδικασία Explore.

<u>Υλοποίηση στο S.P.S.S.</u>

Κατά τη διαδικασία προσαρμογής του μοντέλου της παλινδρόμησης από το παράθυρο Linear Regression και από την επιλογή Save, ζητούμε την αποθήκευση είτε των Unstandardized Residuals είτε των Standardized Residuals (μη τυποποιημένα και τυποποιημένα υπόλοιπα αντίστοιχα).



Έπειτα, ελέγχουμε με γραφικούς τρόπους (βλέπε Q-Q plot και Detrended Q-Q plot) και με το στατιστικό τεστ των Shapiro-Wilk (βλέπε διαδικασία Explore, 2^{ο} Κεφάλαιο).

Για το παράδειγμα της προηγούμενης παραγράφου, προκύπτει ότι η υπόθεση της κανονικότητας των υπολοίπων δεν απορρίπτεται (τεστ Shapiro-Wilk, p-τιμή=0.784>0,05).

Tests of Normality

	Kolmogorov-Smirnov(a)			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Unstandardized Residual	,155	10	,200(*)	,960	10	,784

* This is a lower bound of the true significance.
a Lilliefors Significance Correction

Τρόποι διόρθωσης του προβλήματος

Εάν η υπόθεση της κανονικότητας των σφαλμάτων του μοντέλου μας δεν μπορεί να γίνει δεκτή τότε καταφεύγουμε σε ένα μετασχηματισμό των τιμών της εξαρτημένης μεταβλητής έτσι ώστε να επιτευχθεί η κανονικότητα. Η μορφή του ιστογράμματος των υπολοίπων ίσως μας υποδεικνύει ποιος μετασχηματισμός είναι κατάλληλος. Ενδεικτικά αν έχουμε ιστόγραμμα με ουρά για μεγάλες τιμές, τότε είναι κατάλληλος ο μετασχηματισμός του λογαρίθμου, ενώ αν η ουρά παρατηρείται για τις μικρές τιμές, θεωρούμε το μετασχηματισμό της ρίζας. Εναλλακτικά είναι διαθέσιμος ο μετασχηματισμός των Box and Cox (1964).

Αν το μέγεθος του δείγματος είναι μεγάλο (λόγω του Κεντρικού Οριακού Θεωρήματος) χρησιμοποιούμε την κανονικότητα των σφαλμάτων προσεγγιστικά, με τη διαφοροποίηση ότι οι κρίσιμες πιθανότητες (οι p-τιμές δηλαδή) είναι προσεγγιστικές και όχι ακριβείς.

Συνέπειες της μη κανονικότητας των σφαλμάτων

Η μη κανονικότητα των σφαλμάτων έχει τις ακόλουθες συνέπειες:

α) Λάθος διαστήματα εμπιστοσύνης και μη σωστοί έλεγχοι υποθέσεων για τις παραμέτρους του μοντέλου.

β) Οι εκτιμητές ελαχίστων τετραγώνων δεν είναι Α.Ο.Ε.Δ.

Παρατήρηση: Διάφορα άλλα προβλήματα παραβίασης των υποθέσεων του μοντέλου μπορούν να έχουν ως συνέπεια τη μη κανονικότητα των υπολοίπων. Μέλημά μας είναι η διόρθωση των υπόλοιπων προβλημάτων και όχι της μη κανονικότητας, ειδικά αν έχουμε μεγάλο σε μέγεθος δείγμα.

8.2.2 Έλεγχος σταθερής διακύμανσης των σφαλμάτων

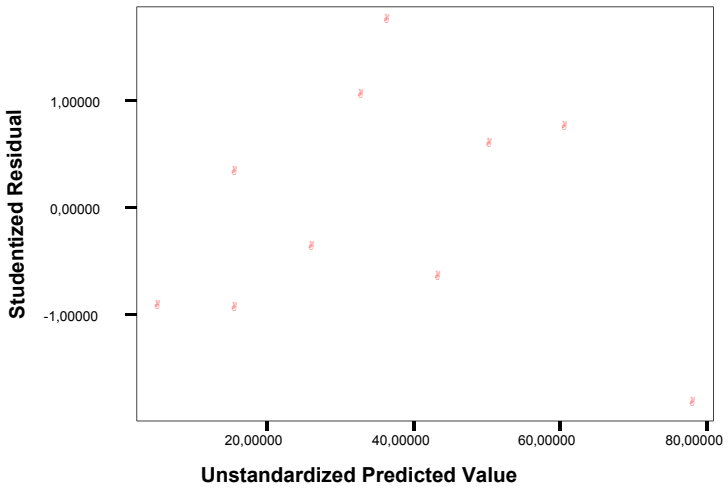
Ο έλεγχος της σταθερής διακύμανσης των σφαλμάτων γίνεται (βλέπε μεταξύ άλλων Seber, 1977, σελ. 165) με τη γραφική παράσταση είτε των τυποποιημένων υπολοίπων (Standardized Residuals) είτε των μαθητικοποιημένων υπολοίπων (Studentized Residuals)

ως προς τις εκτιμώμενες τιμές (Unstandardized Predicted Values). Αν η διακύμανση είναι σταθερή στο γράφημα που προκύπτει παρατηρούμε ότι τα υπόλοιπα κατανέμονται τυχαία γύρω από μία οριζόντια γραμμή που περνά από το 0.

Αντίθετα αν διαπιστώσουμε για παράδειγμα είτε αύξηση είτε ελάττωση της διακύμανσης με τις εκτιμώμενες τιμές, υπάρχει πρόβλημα σταθερής διακύμανσης. Κάτι τέτοιο δεν πρέπει να θεωρείται ασυνήθιστο. Είναι αναμενόμενο να συμβεί, για παράδειγμα, αν έχουμε ως εξαρτημένη μεταβλητή τις Αποδοχές ενός υπαλλήλου και ως ανεξάρτητη τα χρόνια των σπουδών (βλέπε και πρόβλημα συσχετισμένων σφαλμάτων).

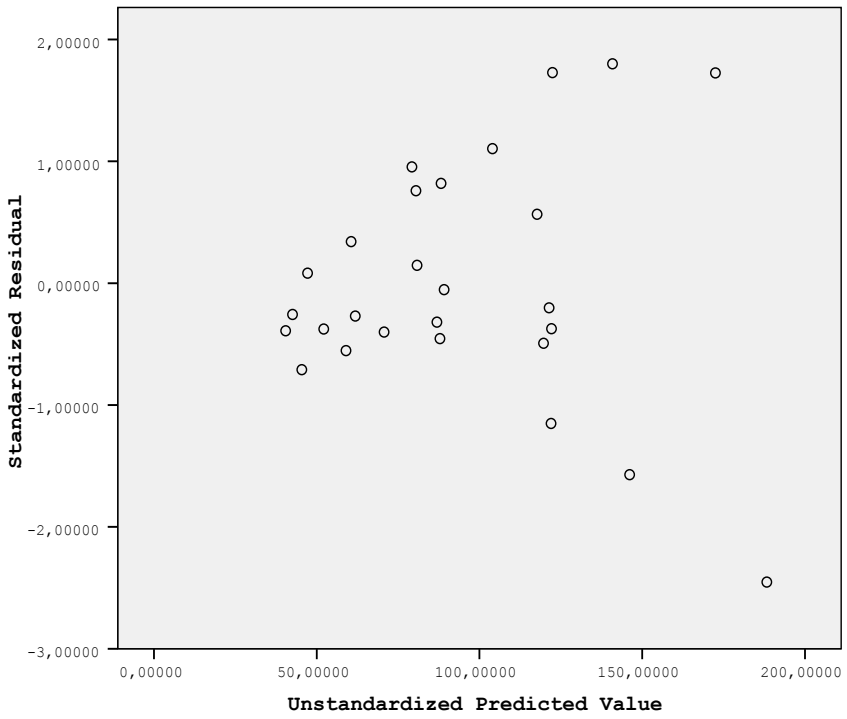
Υλοποίηση στο S.P.S.S.

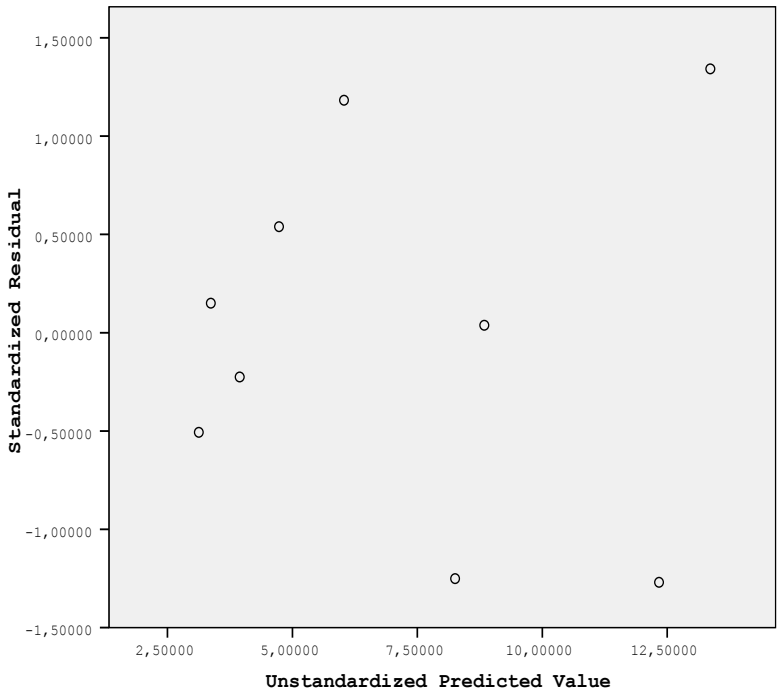
Κατά τη διαδικασία προσαρμογής του μοντέλου της παλινδρόμησης από το παράθυρο Linear Regression και από την επιλογή Save, ζητούμε την αποθήκευση των Unstandardized Predicted Values και των Studentized Residuals (μη τυποποιημένες εκτιμώμενες τιμές και μαθητικοποιημένα υπόλοιπα, αντίστοιχα). Έπειτα, κάνουμε τη γραφική παράσταση αυτών π.χ. μέσω της διαδικασίας Graphs →Interactive →Scatter plot.



Από το γράφημα αυτό παρατηρούμε ότι τα υπόλοιπα κατανέμονται τυχαία γύρω από το μηδέν, αλλά υπάρχει και μία παρατήρηση η οποία είναι κάπως ασυνήθιστη (βλέπε δεξιά κάτω γωνία).

Στη συνέχεια παρατίθενται κάποιες ενδεικτικές γραφικές παραστάσεις βασισμένες σε παραδείγματα των Chatterjee, S. and Price, B. (1977), από τις οποίες συμπεραίνουμε ότι η υπόθεση της σταθερής διακύμανσης απορρίπτεται.





Τρόποι διόρθωσης του προβλήματος

Η χρήση των γενικευμένων εκτιμητών ελαχίστων τετραγώνων ή η προσθήκη κάποιου όρου στο μοντέλο ή ένας κατάλληλος μετασχηματισμός των τιμών της εξαρτημένης μεταβλητής συνιστούν τρόπους διόρθωσης του προβλήματος της μη σταθερής διακύμανσης. Ενδεικτικά αναφέρουμε (βλέπε για περισσότερες λεπτομέρειες Rawlings (1988) και Καρακώστας (2002)) ότι οι συνηθέστεροι μετασχηματισμοί είναι: η τετραγωνική ρίζα (όταν η εξαρτημένη μεταβλητή περιγράφει τον αριθμό των γεγονότων σε κάποιο χρονικό διάστημα δηλ. ακολουθεί Poisson κατανομή), ο λογάριθμος (το εύρος των τιμών της εξαρτημένης είναι μεγάλο και λαμβάνει θετικές τιμές) και ο αντίστροφος μετασχηματισμός (η πλειοψηφία των τιμών κοντά στο μηδέν αλλά υπάρχουν και κάποιες αρκετά μεγάλες τιμές).

Συνέπειες της μη σταθερής διακύμανσης των σφαλμάτων

Η μη σταθερή διακύμανση των σφαλμάτων έχει τις ακόλουθες συνέπειες:

α) Λάθος εκτίμηση της διακύμανσης των εκτιμητών των παραμέτρων του μοντέλου.

β) Μη αξιόπιστα διαστήματα εμπιστοσύνης για τις παραμέτρους του μοντέλου,

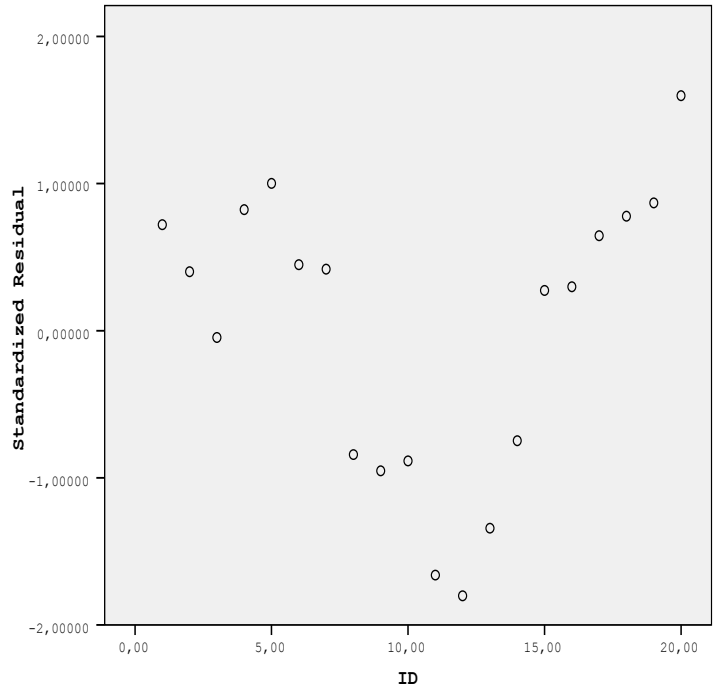
γ) Μη αξιόπιστοι έλεγχοι υποθέσεων για τις παραμέτρους του μοντέλου.

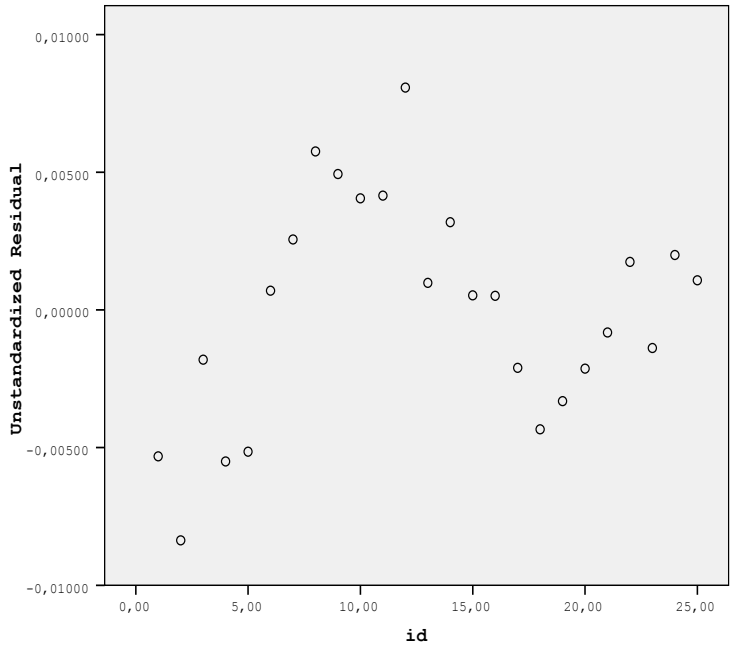
8.2.3 Έλεγχος ασυσχέτιστου των σφαλμάτων

Η ύπαρξη συσχετισμένων σφαλμάτων μπορεί να οφείλεται σε πολλούς λόγους. Είναι σύνηθες φαινόμενο στην περίπτωση που τα δεδομένα έχουν καταγραφεί σε χρονολογική σειρά. Στη συνέχεια παραθέτουμε κάποιους από τους τρόπους ελέγχου της ύπαρξης ή μη συσχετισμένων σφαλμάτων, υλοποιώντας τους στο S.P.S.S. για το παράδειγμα της προηγούμενης παραγράφου.

1. Γραφικά ο έλεγχος της ύπαρξης ή μη αυτοσυσχέτιστων σφαλμάτων γίνεται με την γραφική παράσταση των υπολοίπων (ή των μαθητικοποιημένων υπολοίπων) ως προς την χρονολογική σειρά των παρατηρήσεων. Αν η εικόνα μιας τέτοιας γραφικής παράστασης έχει κυματοειδή μορφή ή οι τιμές των υπολοίπων σχετίζονται με τη χρονολογική σειρά (π.χ. στην αρχή μεγάλες τιμές έπειτα μικρές κ.ο.κ.) τότε οδηγούμαστε στο συμπέρασμα ότι υπάρχει αυτοσυσχέτιση μεταξύ των σφαλμάτων του μοντέλου μας. Ενδεικτικές τέτοιες γραφικές

παραστάσεις είναι οι ακόλουθες που είναι βασισμένες σε παραδείγματα των Chatterjee, S. and Price, B. (1977)

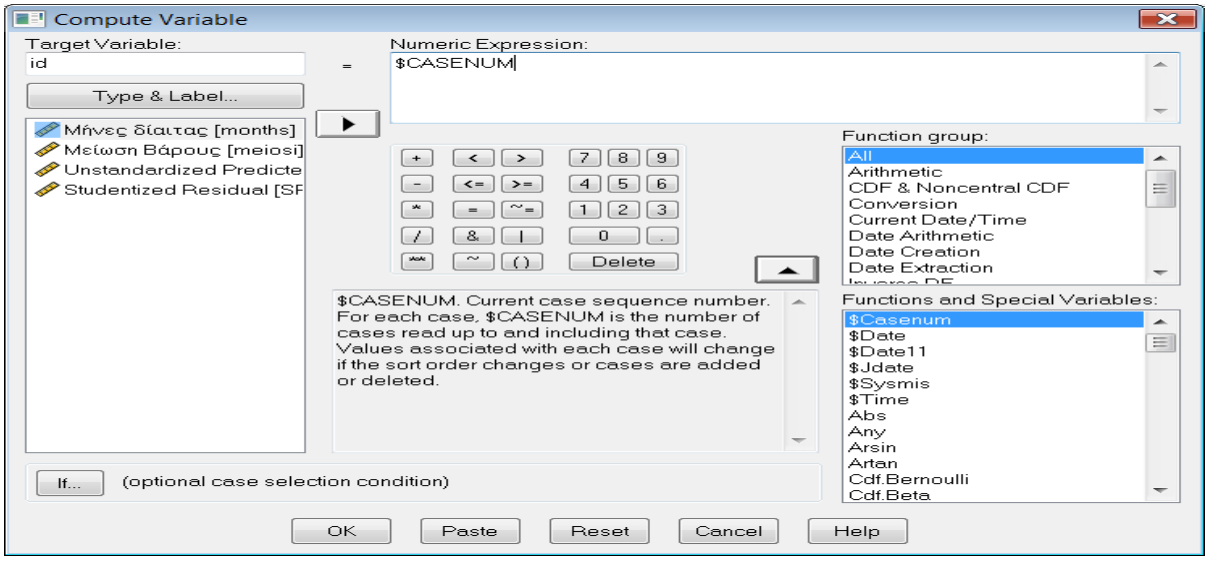




<u>Υλοποίηση στο S.P.S.S.</u>

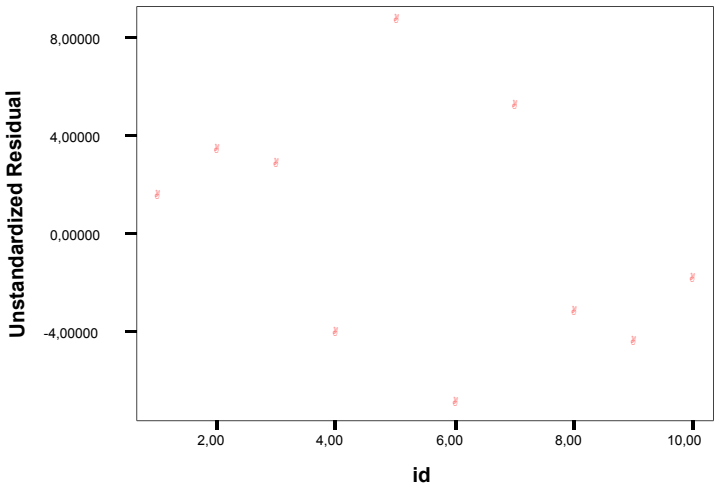
Δημιουργούμε μία νέα στήλη-μεταβλητή με την ονομασία π.χ. ID. Στη στήλη αυτή καταγράφεται ο αύξων αριθμός της παρατήρησης (συνάρτηση $CASENUM) και έπειτα

μέσω π.χ. της διαδικασίας Graphs →Interactive →Scatter plot αποκτούμε το γράφημα που επιθυμούμε.



Ερμηνεία αποτελεσμάτων

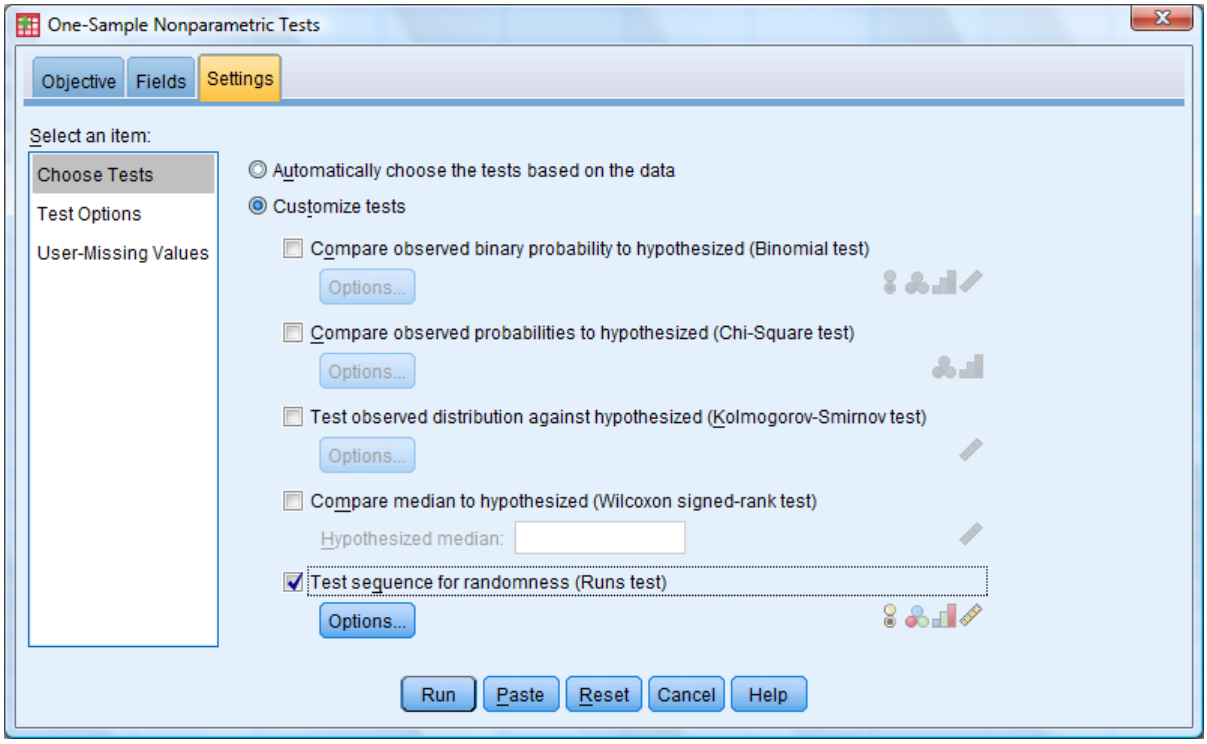
Η εικόνα της γραφικής παράστασης των υπολοίπων ως προς την χρονολογική σειρά δεν έχει κάποια ιδιαίτερη κυματοειδή μορφή, επομένως φαίνεται να μην απορρίπτεται η υπόθεση των ασυσχέτιστων σφαλμάτων.



2. Υπάρχουν και στατιστικά τεστ που ελέγχουν αν τα σφάλματα είναι συσχετισμένα ή όχι. Ένα από τα γνωστά τεστ είναι το τεστ των ροών που στηρίζεται στην ακολουθία-διάταξη των προσήμων των υπολοίπων (είναι διαταγμένα σε χρονολογική σειρά).

Υλοποίηση στο S.P.S.S.

Κατά τη διαδικασία προσαρμογής του μοντέλου της παλινδρόμησης από το παράθυρο Linear Regression και από την επιλογή Save, ζητούμε την αποθήκευση των Standardized Residuals (τυποποιημένα υπόλοιπα). Έπειτα, από το κεντρικό παράθυρο διαλόγου επιλέγουμε: Analyze→NonParametric Tests→One Sample και από το πλαίσιο Settings να επιλέξουμε το Runs test



Από την p-τιμή του ελέγχου αποφασίζουμε αν υπάρχει αυτοσυσχέτιση ή όχι (αν p-τιμή>0.05 δεν υπάρχει αυτοσυσχέτιση).

Ερμηνεία αποτελεσμάτων

Η υπόθεση της τυχαιότητας των σφαλμάτων δεν απορρίπτεται με το τεστ των ροών (p-τιμή>0.05)

3. Ένας άλλος στατιστικός τρόπος εξέτασης της αυτοσυσχέτισης πρώτου βαθμού επιτυγχάνεται με το στατιστικό των Durbin-Watson (Linear Regression Statistics). Το στατιστικό αυτό ελέγχει την μηδενική υπόθεση της μη ύπαρξης αυτοσυσχέτισης έναντι της

εναλλακτικής ότι υπάρχει θετική αυτοσυσχέτιση πρώτου βαθμού (γραμμική). Η τιμή d αυτού του στατιστικού συγκρίνεται με τις τιμές d_l και d_u που δίνονται από κατάλληλους πίνακες. Ισχύει ότι αν $d < d_l$ τότε απορρίπτεται η υπόθεση των ασυσχέτιστων σφαλμάτων. Αν $d > d_u$ η υπόθεση δεν μπορεί να απορριφθεί, ενώ αν $d_l < d < d_u$ δεν μπορούμε να πάρουμε απόφαση.

Παρατήρηση Ο έλεγχος της μηδενικής υπόθεσης της μη ύπαρξης αυτοσυσχέτισης έναντι της εναλλακτικής ότι υπάρχει αρνητική αυτοσυσχέτιση πρώτου βαθμού γίνεται ανάλογα χρησιμοποιώντας την τιμή $d^* = 4 - d$.

Υλοποίηση στο S.P.S.S.

Κατά τη διαδικασία προσαρμογής του μοντέλου της παλινδρόμησης από το παράθυρο Linear Regression και από την επιλογή Statistics, επιλέγουμε το πλαίσιο Durbin Watson.

Ερμηνεία αποτελεσμάτων

Model Summary(b)

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	,976(a)	,953	,947	5,28060	3,070

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

Η τιμή του στατιστικού των Durbin-Watson είναι ίση με 3.070 και καθώς $d_l = 0.87913$ και $d_u = 1.31971$ (τιμές που προκύπτουν από ειδικούς πίνακες, βλέπε Καρακώστας 2002) η υπόθεση των ασυσχέτιστων σφαλμάτων δεν απορρίπτεται.

4. Ένας εναλλακτικός γραφικός τρόπος ελέγχου της ύπαρξης αυτοσυσχέτισης k βαθμού αποτελεί η γραφική παράσταση των υπολοίπων $e_1, e_2, ..., e_n$ ως προς τις τιμές $(-, -, ..., e_k, ..., e_{n-1})$. Αν από αυτό το γράφημα προκύπτει μία γραμμική τάση τότε έχουμε αυτοσυσχέτιση k βαθμού.

Υλοποίηση στο S.P.S.S.

Είναι απαραίτητος ο σχηματισμός, η δημιουργία μίας νέας στήλης όπου θα δίνονται οι τιμές των υπολοίπων $(-,-,...,e_k,...,e_{n-1})$. Επιτυγχάνεται με χρήση της συνάρτησης LAG(Variable,k), όπου στο πλαίσιο Variable εισάγουμε τη μεταβλητή των υπολοίπων και στο πλαίσιο k το βαθμό της αυτοσυσχέτισης που θέλουμε να ελέγξουμε. Έπειτα μέσω π.χ. της διαδικασίας Graphs →Interactive →Scatter plot αποκτούμε το γράφημα που επιθυμούμε.

Τρόποι διόρθωσης του προβλήματος

Η άρση της αυτοσυσχέτισης επιτυγχάνεται μεταξύ άλλων είτε με κατάλληλο μετασχηματισμό των μεταβλητών είτε με εισαγωγή νέων μεταβλητών. Για λεπτομέρειες σχετικά με αυτούς τους τρόπους παραπέμπουμε τον αναγνώστη στο σύγγραμμα των Chaterjee and Price (1977). Ένας άλλος τρόπος είναι με χρήση γενικευμένων εκτιμητών ελαχίστων τετραγώνων (βλέπε Rawlings (1988)). Στο πλαίσιο αυτού του προπτυχιακού μαθήματος απλά θα επισημαίνουμε την ύπαρξη αυτοσυσχέτισης και τις συνέπειες αυτής και θα προβαίνουμε σε διόρθωση του προβλήματος στην ειδική περίπτωση της ύπαρξης αυτοσυσχέτισης πρώτου βαθμού (βλέπε Άσκηση 1 ενότητα 8.3).

Συνέπειες

Η ύπαρξη αυτοσυσχέτισης μεταξύ των σφαλμάτων του μοντέλου έχει τις ακόλουθες συνέπειες (Chaterjee and Price (1977)):

α) Οι εκτιμητές ελαχίστων τετραγώνων είναι αμερόληπτοι, αλλά όχι ΑΟΕΔ.

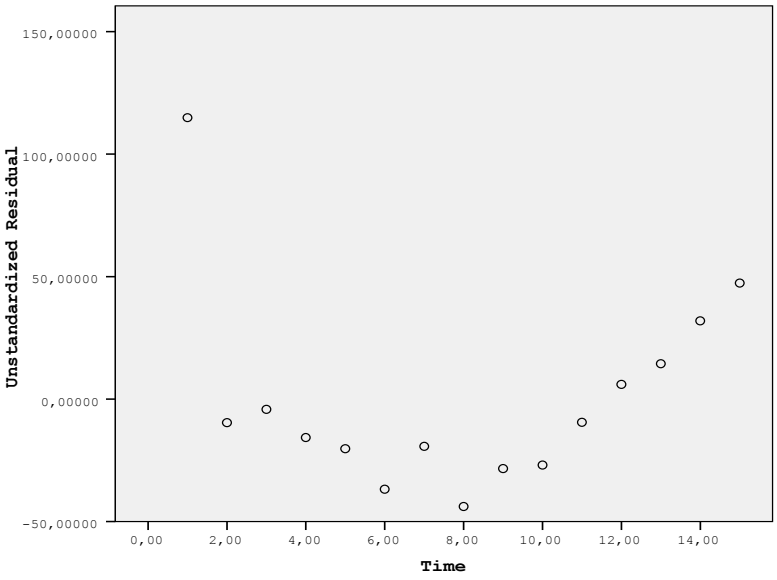
β) Ο εκτιμητής του σ και τα τυπικά σφάλματα των συντελεστών της παλινδρόμησης μπορεί να υποεκτιμούνται. Αυτό οδηγεί σε μη αξιόπιστα αποτελέσματα για τα διαστήματα εμπιστοσύνης και για τους ελέγχους υποθέσεων για τις παραμέτρους του μοντέλου.

8.2.4 Έλεγχος ορθότητας μοντέλου

Γραφικά, ο έλεγχος της ορθότητας του μοντέλου γίνεται (βλέπε μεταξύ άλλων Norusis (2002)) με την γραφική παράσταση των υπολοίπων ως προς την ανεξάρτητη

μεταβλητή. Αν δεν παρατηρηθεί κάποια ιδιαίτερη μορφή και τα σημεία βρίσκονται τυχαία γύρω από το μηδέν το μοντέλο μπορεί να θεωρηθεί ορθό. Αν δούμε κάποια ιδιαίτερη γραφική παράσταση τότε η εξαρτημένη και η ανεξάρτητη μεταβλητή μπορεί να μην συνδέονται με μία γραμμική σχέση.

Μία ενδεικτική γραφική παράσταση που υποδεικνύει πρόβλημα ορθότητας μοντέλου (εισαγωγή δευτεροβάθμιου όρου) βασισμένη σε παράδειγμα των Chatterjee, S. and Price, B. (1977)) είναι η ακόλουθη:



Παρατήρηση: Από τη γραφική παράσταση των υπολοίπων ως προς τις τιμές μίας ανεξάρτητης μεταβλητής που δεν είναι στο μοντέλο μπορούμε να αποφασίσουμε αν πρέπει η συγκεκριμένη μεταβλητή να συμπεριληφθεί στο μοντέλο ή όχι. Έτσι αν παρατηρηθεί κάποια σχέση τότε ίσως πρέπει να συμπεριληφθεί αυτή η ανεξάρτητη μεταβλητή στο μοντέλο.

Υλοποίηση στο S.P.S.S.

Κατά τη διαδικασία προσαρμογής του μοντέλου της παλινδρόμησης από το παράθυρο Linear Regression και από την επιλογή Save, ζητούμε την αποθήκευση των Unstandardized Residuals (μη τυποποιημένα υπόλοιπα). Έπειτα, κάνουμε τη γραφική παράσταση αυτών π.χ. μέσω της διαδικασίας Graphs →Interactive →Scatter plot.

Τρόποι διόρθωσης του προβλήματος

Κάποιες μη γραμμικές σχέσεις είναι δυνατό να αναχθούν σε γραμμικές με κατάλληλους μετασχηματισμούς. Στον παρακάτω πίνακα αναφέρονται κάποιες από αυτές καθώς και ο μετασχηματισμός που οδηγεί σε γραμμικά μοντέλα (βλέπε Rawlings (1988, σελ. 306-308), Chaterjee and Price (1977, σελ. 29)). Φυσικά όλες οι μη γραμμικές σχέσεις δεν είναι δυνατό να μετατραπούν σε γραμμικές. Η μελέτη και προσαρμογή μη γραμμικών μοντέλων ξεφεύγει από τους σκοπούς αυτών των σημειώσεων.

Μη γραμμική σχέση	Κατάλληλος Μετασχηματισμός
$Y_i = aX_i^{\beta}\varepsilon_i$	Λογάριθμος
$Y_i = a\exp(\beta X_i)\varepsilon_i$	Λογάριθμος
$Y_i = \dfrac{X_i}{a+\beta X_i+\varepsilon_i}$	Αντίστροφος
$Y_i = \dfrac{a}{1+\gamma\exp(-\beta X_i)\varepsilon_i}$	$Y^* = \ln\left(\dfrac{a}{Y}-1\right)$

Συνέπειες του μη ορθού μοντέλου

Οι συνέπειες ενός μη ορθού μοντέλου είναι:

α) λάθος ερμηνεία των παραμέτρων του μοντέλου,

β) λάθος προβλέψεις,

γ) λάθος εκτίμηση της κοινής διακύμανσης των σφαλμάτων.

Από το τελευταίο προκύπτει ως επακόλουθη συνέπεια

δ) η μη εγκυρότητα των όποιων διαστημάτων εμπιστοσύνης και ελέγχων υποθέσεων για τις παραμέτρους του μοντέλου.

8.2.5 Έλεγχος ακραίων τιμών

Σε μερικές περιπτώσεις το μοντέλο φαίνεται να είναι ορθό για την πλειοψηφία των δεδομένων, αλλά υπάρχει ένα υπόλοιπο που η απόλυτη τιμή του είναι πολύ μεγαλύτερη από τα άλλα υπόλοιπα. Κάτι τέτοιο μπορεί να οφείλεται σε λάθος καταγραφή των δεδομένων αλλά και όχι. Στη δεύτερη περίπτωση η μελέτη της ακραίας τιμής είναι εξίσου σημαντική όσο και η μελέτη του υπόλοιπου συνόλου δεδομένων καθώς μπορεί να μας δώσει σημαντικές πληροφορίες. Έτσι, η αυτόματη απομάκρυνση των ακραίων τιμών δεν συνίσταται.

Ένας τρόπος ελέγχου της ύπαρξης ή μη ακραίων παρατηρήσεων στα δεδομένα μας γίνεται με τη βοήθεια των τυποποιημένων ή μαθητικοποιημένων υπολοίπων. Τότε (βλέπε μεταξύ άλλων Field, 2005, σελ. 164) παρατηρήσεις των οποίων η απόλυτη τιμή των υπολοίπων αυτών είναι μεγαλύτερη του τρία (για να είμαστε περισσότερο ακριβείς του 3.29) θεωρούνται ακραίες και συνηθέστερα αποκλείονται από την περαιτέρω ανάλυση. Αν περισσότερο από 1% των τυποποιημένων υπολοίπων έχουν απόλυτες τιμές μεγαλύτερες του 2.5 (για την ακρίβεια του 2.58) υποδεικνύεται ότι το μοντέλο έχει κακή προσαρμογή. Στο ίδιο συμπέρασμα καταλήγουμε αν 5% των διαθέσιμων παρατηρήσεων έχουν απόλυτες τιμές των τυποποιημένων υπολοίπων μεγαλύτερες του 2 (του 1.96 για την ακρίβεια όταν το επίπεδο σημαντικότητας είναι 5%). Τέλος, παρατηρήσεις με απόλυτες τιμές των τυποποιημένων υπολοίπων μεταξύ 2 και 3 (1.96 και 3.29 αν θέλουμε να είμαστε πιο ακριβείς) θεωρούνται ως πιθανές ακραίες. Η τελική απόφαση για το αν είναι ακραίες ή όχι γίνεται με τη βοήθεια ενός στατιστικού ελέγχου.

Ο στατιστικός έλεγχος για την ύπαρξη ακραίων παρατηρήσεων γίνεται με την βοήθεια των μαθητικοποιημένων διαγραφόμενων υπολοίπων (studentized deleted residuals). Αν τα σφάλματα του μοντέλου ακολουθούν κανονική κατανομή, τότε η κατανομή των studentized deleted residuals είναι t-κατανομή με n-p-1 βαθμούς ελευθερίας. Απόλυτες τιμές των μαθητικοποιημένων διαγραφόμενων υπολοίπων για μία παρατήρηση μεγαλύτερες του $t_{n-p-1,a/2}$ = IDF.T(1-a/2,n-p-1) υποδεικνύουν τη συγκεκριμένη παρατήρηση ως ακραία.

<u>Τρόποι διόρθωσης του προβλήματος</u>

Το πρώτο μέλημα μας όταν έχουμε αρκετές ακραίες τιμές είναι η εύρεση ενός μετασχηματισμού που θα διορθώσει το πρόβλημα. Αν η εύρεση ενός τέτοιου

μετασχηματισμού είναι αδύνατη τότε είτε θα απορρίψουμε τις ακραίες τιμές και θα προχωρήσουμε στην ανάλυση των υπόλοιπων δεδομένων (τακτική που οδηγεί πολλές φορές σε απώλεια σημαντικής πληροφορίας) είτε θα χρησιμοποιήσουμε μεθόδους ανθεκτικές στην ύπαρξη ακραίων τιμών (βλέπε Huber (1973)). Υπάρχει βέβαια και η επιλογή της διεξαγωγής της έρευνας τόσο με τις ακραίες όσο και χωρίς τις ακραίες τιμές και την επισήμανση των όποιων διαφορετικών αποτελεσμάτων

Συνέπειες της ύπαρξης ακραίων τιμών

Η παρουσία ακραίων τιμών στο δείγμα μας έχει σαν συνέπεια οι εκτιμητές των παραμέτρων καθώς και οι διακυμάνσεις αυτών να μην έχουν τις γνωστές ιδιότητες των εκτιμητών ελαχίστων τετραγώνων. Άμεση συνέπεια αυτού είναι η μη εγκυρότητα των όποιων διαστημάτων εμπιστοσύνης ή ελέγχου υποθέσεων για τις παραμέτρους του μοντέλου.

8.2.6 Επηρεάζουσες παρατηρήσεις

Είναι πιθανό δύο ή περισσότερες πειραματικές μονάδες να επιδρούν σημαντικά στο μοντέλο παλινδρόμησης. Τέτοιες παρατηρήσεις ονομάζονται επηρεάζουσες.

Έτσι για παράδειγμα οι συντελεστές των παραμέτρων του μοντέλου αλλάζουν αρκετά όταν οι τιμές των συγκεκριμένων πειραματικών μονάδων εξαιρούνται από τον υπολογισμό τους. Μία τέτοια κατάσταση είναι ανεπιθύμητη καθώς θέλουμε ένα μοντέλο παλινδρόμησης που να μην εξαρτάται από τις τιμές ενός μικρού αριθμού πειραματικών μονάδων, αλλά όλες οι πειραματικές μονάδες να συνεισφέρουν όσο γίνεται το ίδιο στον υπολογισμό των συντελεστών αυτών. Θα πρέπει να δοθεί ξεχωριστή σημασία στις συγκεκριμένες πειραματικές μονάδες που είναι επηρεάζουσες παρατηρήσεις και ίσως πρέπει να παρουσιαστούν τα αποτελέσματα των αναλύσεων με και χωρίς αυτές.

Τρόποι ελέγχου

Από το πλαίσιο Influence Statistics του Linear Regression Save μπορούμε να ζητήσουμε την αποθήκευση διάφορων ποσοτήτων για την εξέταση αυτού του προβλήματος:

DfBeta(s): Η διαφορά στις τιμές των συντελεστών της παλινδρόμησης αν δεν ληφθεί υπόψη η συγκεκριμένη πειραματική μονάδα. Υπολογίζεται και για τον σταθερό όρο. Οι τυποποιημένες τιμές παρατίθενται στη στήλη Standardized DfBeta. Απόλυτες τιμές αυτών

μεγαλύτερες από $2/\sqrt{n}$ μας υποδεικνύουν παρατήρηση που επιδρά στην εκτίμηση των συντελεστών της παλινδρόμησης (βλέπε Belsley, Kuh, and Welsch (1980)).

DfFit: Μετρά τη διαφορά στην προσαρμογή, δηλαδή στην εκτιμώμενη τιμή, αν δεν συμπεριληφθεί η συγκεκριμένη παρατήρηση στους υπολογισμούς. Δίνονται και οι αντίστοιχες τυποποιημένες τιμές Standardized DfFit. Απόλυτες τιμές αυτών μεγαλύτερες του $2\sqrt{\frac{p+1}{n}}$ υποδεικνύουν επηρεάζουσες παρατηρήσεις (βλέπε Belsley, Kuh, and Welsch (1980)).

Σχόλιο: Κάποιοι συγγραφείς διαφοροποιούν τα παραπάνω κριτήρια στην περίπτωση που το μέγεθος του δείγματος είναι μικρότερο του 30. Υποστηρίζουν ότι σε μία τέτοια περίπτωση μία παρατήρηση είναι επηρεάζουσα για τιμές των παραπάνω τυποποιημένων δεικτών μεγαλύτερες της μονάδας.

Covariance ratio: Το πηλίκο της ορίζουσας του πίνακα διακυμάνσεων συνδιακυμάνσεων χωρίς η συγκεκριμένη παρατήρηση να λαμβάνεται υπόψη στους υπολογισμούς προς την αντίστοιχη ορίζουσα όταν αυτή η παρατήρηση λαμβάνεται υπόψη. Τιμές μεγαλύτερες (μικρότερες αντίστοιχα) του $1+3\frac{p+1}{n}$ (του $1-3\frac{p+1}{n}$ αντίστοιχα) υποδεικνύουν επηρεάζουσα παρατήρηση (βλέπε Belsley, Kuh, and Welsch (1980)).

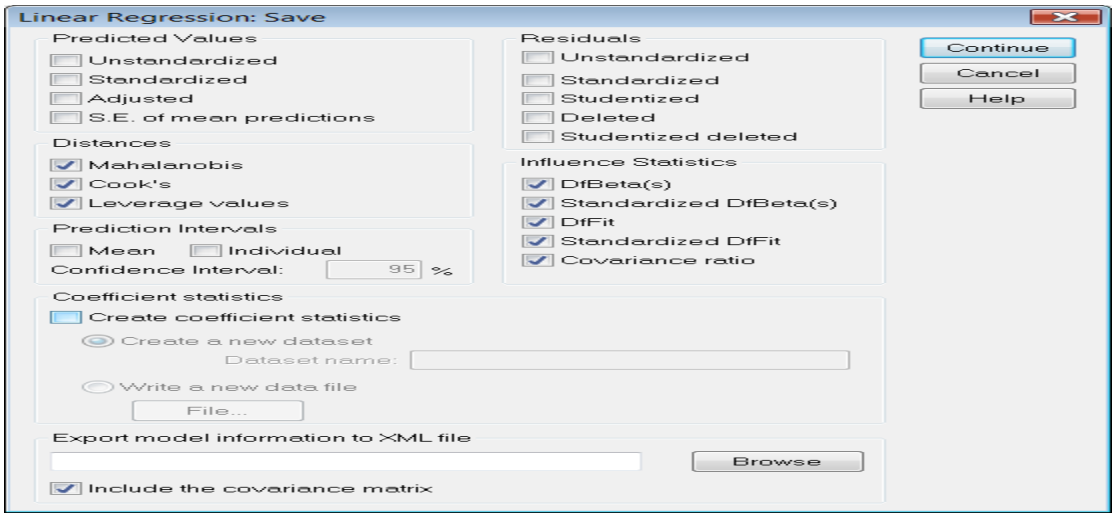
Απόσταση του Cook: καθορίζει πόσο οι τιμές των υπολοίπων όλων των περιπτώσεων θα μεταβληθούν, αν η συγκεκριμένη τιμή δε ληφθεί υπόψη στους υπολογισμούς των συντελεστών του μοντέλου. Αποδεικνύεται ότι το στατιστικό αυτό ακολουθεί μία F κατανομή με 2 και n-2 βαθμούς ελευθερίας. Τιμές της απόστασης του Cook για μία παρατήρηση μεγαλύτερες του $F_{2,n-2,a} = IDF.F(1-a,2,n-2)$ υποδεικνύουν τη συγκεκριμένη παρατήρηση ως επηρεάζουσα.

Απόσταση Mahalanobis: καθορίζει την απόσταση των πειραματικών μονάδων από τη μέση τιμή της προβλεπόμενης τιμής. Το ενδιαφέρον επικεντρώνετε στις πειραματικές μονάδες με μεγάλες τιμές σε αυτή, αλλά δυστυχώς δεν υπάρχει ένα γενικό cut-off point (βλέπε Barnett and Lewis, 1978).

Leverage values: Μετρούν την επίδραση μίας πειραματικής μονάδας στην προσαρμογή του μοντέλου της παλινδρόμησης. Οι κεντρικές Leverage values λαμβάνουν τιμές από 0 (όχι ενδείξεις επίδρασης) έως $(n-1)/n$.

Υλοποίηση στο S.P.S.S.

Κατά τη διαδικασία προσαρμογής του μοντέλου της παλινδρόμησης από το παράθυρο Linear Regression και από την επιλογή Save, ζητούμε την αποθήκευση των



Για τον εντοπισμό πιθανής επηρεάζουσας παρατήρησης υπολογίζουμε, αρχικά, το λεγόμενο cut-off point (δηλαδή τις ποσότητες $2/\sqrt{n}$, $2\sqrt{\frac{p+1}{n}}$, $1\pm 3\frac{p+1}{n}$ και $F_{2,n-2,a}$). Έπειτα, δημιουργούμε μία νέα στήλη-μεταβλητή με την ονομασία π.χ. ID. Στη στήλη αυτή καταγράφεται ο αύξων αριθμός της παρατήρησης (συνάρτηση $CASENUM). Έπειτα επιλέγουμε Graphs→ Legacy Dialog→ Scatter/Dot και Simple Scatter. Στο νέο παράθυρο διαλόγου που προκύπτει τοποθετούμε στον άξονα των Y π.χ. την απόσταση Cook, ενώ στον άξονα των X την νέα μεταβλητή ID. Κάνοντας διπλό κλικ στο γράφημα που προκύπτει και έπειτα δεξί κλικ ζητούμε την προσθήκη γραμμών αναφοράς στον άξονα Y, Add→ Y Axis Reference Line και στο πλαίσιο Position δηλώνουμε την κατάλληλη τιμή του cut-off point. Αν υπάρχουν σημεία που παραβιάζουν την προς έλεγχο σχέση με το cut-off point με δεξί κλικ και επιλογή του Show Data Labels μας υποδεικνύεται ο αύξων αριθμός της πειραματικής μονάδας (εναλλακτικά αφού επιλέξουμε το σημείο, δεξί κλικ και επιλογή του Go to Case)

8.3 Ασκήσεις

1. Στο αρχείο autocorrelation1.sav καταγράφονται τα τετραμηνιαία δεδομένα από το 1952 έως το 1956 που αφορούν τις δαπάνες και τις αποταμιεύσεις μετρούμενες σε δισ. δολάρια. Οι οικονομολόγοι ενδιαφέρονται για την μεταβολή στις δαπάνες που προκαλούνται από τη μεταβολή στις αποταμιεύσεις (Chatterjee and Price (1980, σελ. 124)).

2. Μία εταιρεία θέλει να κατανοήσει τη σχέση μεταξύ οικοδομικών αδειών (housing starts) και της ανάπτυξης του πληθυσμού. Στο αρχείο autocorrelation2.sav δίνονται τα δεδομένα αυτά για 25 χρόνια. Επιπλέον, σε μία τρίτη στήλη δίνεται η τιμή ενός δείκτη που μετρά την οικονομική δυνατότητα (mortgage money) (Chatterjee and Price (1980, σελ. 133)).

3. Μία εταιρεία της Αμερικής παράγει και πουλάει εξαρτήματα σκι. Θέλει να προβλέψει τις πωλήσεις της με βάση ένα δείκτη (PDI) που μετρά το εισόδημα. Δίνονται στο αρχείο autocorrelation3.sav τα δεδομένα που αφορούν 40 τρίμηνα από το 1964-1973. (Chatterjee and Price (1980, σελ. 138))

4. Στο αρχείο chatterjeep.44.sav καταγράφονται ο αριθμός των προϊστάμενων και υφιστάμενων 27 εταιρειών. Μπορεί να δημιουργηθεί ένα μοντέλο πρόβλεψης του αριθμού των προϊστάμενων από τον αριθμό των υφιστάμενων; (Chatterjee and Price (1980, σελ. 44))

5. Στο αρχείο chatterjeep.40.sav καταγράφονται το ποσοστό των πτήσεων και ο αριθμός των ατυχημάτων 9 αεροπορικών εταιρειών. Μπορεί να δημιουργηθεί ένα μοντέλο πρόβλεψης του αριθμού των ατυχημάτων από το ποσοστό των πτήσεων; (Chatterjee and Price (1980, σελ. 40))

6. Στο αρχείο δεδομένων chatterjee21.sav καταγράφονται 30 παρατηρήσεις και 2 μεταβλητές που αφορούν την ακροαματικότητα πριν το δελτίο ειδήσεων (lead in) και την ακροαματικότητα του δελτίου ειδήσεων (newsrate). Θέλουμε να εξετάσουμε αν το πρόγραμμα πριν τις ειδήσεις επηρεάζει την ακροαματικότητα των ειδήσεων. (Chatterjee and Price (1980, σελ. 21))

7*. Με σκοπό να μελετηθεί αν τα γενικά έξοδα, σε ετήσια βάση, μιας οικογένειας μπορούν να προβλέψουν τα έξοδα που γίνονται για την εκπαίδευση των παιδιών της οικογένειας συγκεντρώθηκαν τα δεδομένα του αρχείου EducSpend.sav. Η μελέτη αυτή έγινε πιλοτικά και οι οικογένειες επιλέχθηκαν τυχαία από μια γεωγραφική περιοχή μιας πόλης. Η

μεταβλητή Pay εκφράζει το ποσό των γενικών εξόδων, ενώ η μεταβλητή Spend εκφράζει τα ειδικά έξοδα για την εκπαίδευση, σε χιλιάδες δολάρια. Με βάση τα δεδομένα του συγκεκριμένου αρχείου να δοθεί μια απάντηση στο αρχικό ερώτημα.

8*. Μια ομάδα γιατρών θέλησε να εξετάσει αν ο ρυθμός θνησιμότητας των γυναικών που πάσχουν από καρκίνο του στήθους επηρεάζεται από την θερμοκρασία. Αν ναι να βρουν ένα μοντέλο με το οποίο θα μπορούν να κάνουν αξιόπιστες προβλέψεις για τον μέσο ρυθμό θνησιμότητας (σε ετήσια βάση) από την μέση ετήσια θερμοκρασία. Για τον σκοπό αυτό κατέγραψαν τα δεδομένα στο αρχείο BreastCancer.sav, όπου mortality είναι ο ρυθμός θνησιμότητας και Temperature η μέση θερμοκρασία (σε βαθμούς Fahrenheit), για την συγκεκριμένη χρονιά. Ζητείται με βάση τα δεδομένα αυτά να διατυπώσουμε τα συμπεράσματά μας σχετικά με το αρχικό ερώτημα των γιατρών.

9*. Το Οικονομικό Επιμελητήριο μιας χώρας θέλησε να εξετάσει αν οι μεταβολές του πληθωρισμού επηρεάζουν και πως τα τραπεζικά επιτόκια, σε μηνιαία βάση. Για τον σκοπό αυτό συγκέντρωσε στοιχεία, για τον πληθωρισμό και τα επιτόκια, για τους τελευταίους 191 μήνες. Τα στοιχεία αυτά περιέχονται στο αρχείο BankInflatRate.sav. Στο αρχείο αυτό BankRate είναι το μέσο τραπεζικό επιτόκιο, (σε μηνιαία βάση) και InflatRate ο μέσος μηνιαίος πληθωρισμός.

10*. Τα δεδομένα τα οποία παρουσιάζονται στο αρχείο HeihgtWeight12.sav είναι ένα τυχαίο δείγμα 63 παιδιών ηλικίας 12 ετών από ένα σχολικό συγκρότημα. Σκοπός μας είναι να ελέγξουμε αν και σε ποιο βαθμό το ύψος (σε ίντσες) ενός παιδιού μπορεί να προσδιορίσει το βάρος του (σε λίβρες). Να αναλυθούν τα δεδομένα και να διατυπωθούν τα όποια συμπεράσματα.

11*. Στο αρχείο BirthRatio.sav έχουν συγκεντρωθεί τα αποτελέσματα μιας έρευνας με σκοπό να εξετασθεί κατά πόσον είναι δυνατόν να προβλέψουμε το λόγο του βάρους προς το ύψος (μεταβλητή Ratio στο αρχείο) σε νεογέννητα παιδιά ηλικίας μερικών μηνών (μεταβλητή Age στο αρχείο).(Ο πληθυσμός στον οποίο αναφέρεται το συγκεκριμένο δείγμα είναι αυτός των γυναικών που γέννησαν σε ένα συγκεκριμένο Νοσοκομείο).

Οι ασκήσεις που επισημαίνονται με * καθώς και τα αντίστοιχα σύνολα δεδομένα προέρχονται από το υλικό διδασκαλίας του κ. Κ. Καρακώστα (βλέπε Καρακώστας (2004)).