ΚΕΦΑΛΑΙΟ 6

ΧΡΗΣΗ ΓΡΑΜΜΙΚΩΝ ΜΟΝΤΕΛΩΝ ΚΑΙ ΓΡΑΜΜΙΚΗΣ ΠΑΛΙΝΔΡΟΜΗΣΗΣ

6.1 Εισαγωγή

Σε πολλές στατιστικές εφαρμογές συναντάται το πρόβλημα της μελέτης της σχέσης δυο ή περισσότερων τυχαίων μεταβλητών. Η σχέση αυτή μπορεί να προσδιοριστεί με βάση ορισμένες παρατηρήσεις. Σε κάθε σύστημα, στο οποίο οι μεταβλητές ποσότητες αλλάζουν ή μεταβάλλονται, έχει ενδιαφέρον να εξεταστούν οι επιδράσεις που κάποιες μεταβλητές ασκούν (ή φαίνεται ότι ασκούν) σε κάποιες άλλες μεταβλητές. Στις φυσικές διαδικασίες η ύπαρξη μιας τέτοιας απλής σχέσης αποτελεί μάλλον την εξαίρεση παρά τον κανόνα. Συνήθως υπάρχει μια συναρτησιακή σχέση η οποία είναι τόσο πολύπλοκη ώστε δεν μπορεί να γίνει κατανοητή ή να περιγραφεί με απλούς όρους. Σε μια τέτοια περίπτωση, το καλύτερο είναι να γίνει προσέγγιση αυτής της πολύπλοκης συναρτησιακής σχέσης με μια απλή μαθηματική συνάρτηση, όπως πολυώνυμο, που να περιλαμβάνει τις κατάλληλες μεταβλητές και να προσεγγίζει την κανονική συνάρτηση για κάποιο περιορισμένο εύρος μεταβλητών που εμπλέκονται στη διαδικασία. Η εξέταση αυτής της απλής μαθηματικής συνάρτησης μπορεί να δώσει περισσότερες πληροφορίες για την υπάρχουσα αληθινή σχέση και οδηγήσει σε εκτίμηση των ξεχωριστών αλλά και των κοινών επιδράσεων που παράγονται από τις αλλαγές σε συγκεκριμένες σημαντικές μεταβλητές.

Ακόμα και όταν λογικά δεν υπάρχει φυσική σχέση μεταξύ των μεταβλητών, μπορεί να θέλουμε να τις συσχετίσουμε με κάποια μαθηματική εξίσωση. Ενώ η εξίσωση μπορεί να στερείται φυσικής έννοιας, μπορεί να είναι εξαιρετικά πολύτιμη για την *πρόβλεψη* των τιμών κάποιων μεταβλητών από τις γνώσεις που διαθέτουμε για άλλες μεταβλητές, όταν ισχύουν κάποιες συγκεκριμένες συνθήκες.

Προσδιορίζοντας τη σχέση των μεταβλητών, λέμε ότι έχουμε προσδιορίσει ένα μοντέλο. Όταν μας ενδιαφέρει να προσδιορίσουμε ένα μοντέλο για να ερμηνεύσουμε τη συμπεριφορά της μεταβλητής Y με βάση τη συμπεριφορά των μεταβλητών $X_1,\ldots,X_\kappa$, τότε ονομάζουμε την Y *εξαρτημένη* μεταβλητή και τις $X_1,\ldots,X_\kappa$ *ανεξάρτητες ή ερμηνευτικές* μεταβλητές. Η εκτίμηση ενός στατιστικού μοντέλου γίνεται με ανάλυση στατιστικών δεδομένων, δηλαδή παρατηρήσεων της εξαρτημένης μεταβλητής Y σε επιλεγμένα επίπεδά της ή των ερμηνευτικών μεταβλητών. Οι στατιστικές τεχνικές που χρησιμοποιούνται για το σκοπό αυτό αναφέρονται ως ανάλυση παλινδρόμησης. (Ι. Πανάρετος, 1997)

Η γενική μορφή ενός μοντέλου παλινδρόμησης για την τυχαία μεταβλητή Y με ερμηνευτικές μεταβλητές τις $X_1,X_2,\ldots,X_\kappa$ είναι :

$$Y = f(X_1,X_2,\ldots,X_\kappa) + \varepsilon \quad , \quad \text{με } E(\varepsilon) = 0 \quad \text{και} \quad E(Y) = f(X_1,X_2,\ldots,X_\kappa)$$

Η τυχαία μεταβλητή εκφράζεται ως άθροισμα μιας σχέσης που εκφράζει τη μέση τιμή της εξαρτημένης ως συνάρτηση των ερμηνευτικών μεταβλητών $X_1,X_2,\ldots,X_\kappa$ και ενός τυχαίου όρου. Η $f(X_1,X_2,\ldots,X_\kappa)$ ονομάζεται συνάρτηση παλινδρόμησης της Y επί των $X_1,X_2,\ldots,X_\kappa$. Αν το μοντέλο που εξετάζουμε είναι τέτοιας μορφής που η τυχαία μεταβλητή Y είναι γραμμική συνάρτηση των παραμέτρων του μοντέλου, τότε μιλάμε για ένα *γραμμικό μοντέλο* . Η απλούστερη μορφή τέτοιας σχέσης είναι :

$$Y = \alpha + \beta X + \varepsilon \quad , \text{με } E(\varepsilon) = 0 \quad , \text{όπου } \alpha , \beta \text{ σταθερές } (a, b \text{ εκτιμήσεις})$$

Η μέση τιμή της Y για ορισμένη τιμή της X βρίσκεται πάνω σε μια ευθεία με σταθερό όρο α και κλίση β. Όταν το μοντέλο περιέχει παραπάνω από μια ερμηνευτικές μεταβλητές ονομάζεται *πολυμεταβλητό γραμμικό μοντέλο* :

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_\kappa X_\kappa + \varepsilon \text{ και } E(Y) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_\kappa X_\kappa$$

Το α είναι ο σταθερός όρος της συνάρτησης παλινδρόμησης, δηλαδή η τιμή της Y όταν $X_1 = X_2 = \ldots = X_\kappa = 0$. Ο συντελεστής β_j (j=1,2,…κ) είναι η μεταβολή της Y όταν η ερμηνευτική μεταβλητή X_j αυξηθεί κατά μια μονάδα και οι υπόλοιπες μεταβλητές παραμείνουν σταθερές.

Η διαδικασία εκτίμησης του μοντέλου περιλαμβάνει τα εξής στάδια :

- Επιλογή των ερμηνευτικών μεταβλητών με βάση τη σχετική θεωρία ή τη γνώση μας για τη διαδικασία παραγωγής των τιμών της Y. Κάποιες απ' αυτές μπορούν να κριθούν περιττές ή να προστεθούν επιπλέον ερμηνευτικές μεταβλητές.
- Επιλέγεται ο κατάλληλος μαθηματικός τύπος για τον προσδιορισμό μέρους της σχέσης της Y με την ερμηνευτική μεταβλητή (ή μεταβλητές).
- Επιλογή τιμών για τις παραμέτρους του μοντέλου, έτσι ώστε αυτό να εκτιμά τις παρατηρήσεις $Y_1,...Y_\kappa$ καλύτερα, σύμφωνα με κάποιο κριτήριο. Στο γραμμικό μοντέλο όταν η συμπεριφορά του τυχαίου μέρους ικανοποιεί ορισμένες συνθήκες, επιλέγεται ως κριτήριο καλής προσαρμογής η ελαχιστοποίηση του αθροίσματος για όλες τις τιμές Y_i των τετραγωνικών σφαλμάτων εκτίμησης. Ισοδύναμα επιλέγεται ως μέθοδος εκτίμησης η μέθοδος των ελάχιστων τετραγώνων.
- Τα αποτελέσματα του προηγούμενου σταδίου γενικεύονται σ' όλες τις παρατηρήσεις Y_i της Y, οι οποίες θα μπορούσαν να είχαν συλλεχθεί με τον ίδιο τρόπο που έχουν συλλεχθεί οι παρατηρήσεις των δεδομένων μας.
- Ελέγχεται η καταλληλότητα του εκτιμηθέντος μοντέλου. Εξετάζεται αν οι υποθέσεις που έγιναν για τη συνάρτηση παλινδρόμησης και για τον όρο σφάλματος, στηρίζονται από τα δεδομένα ή αν θα πρέπει να γίνουν τροποποιήσεις στην εξειδίκευση του μοντέλου . Οι τροποποιήσεις μπορεί να υπαγορεύουν άλλη μέθοδο εκτίμησης από αυτή των ελάχιστων τετραγώνων.

6.2 Δημιουργία μοντέλων

Τα Μαθηματικά είναι μαζί με την Ελληνική Γλώσσα τα βασικότερα μαθήματα στο Δημοτικό Σχολείο. Πολλοί εκπαιδευτικοί λαμβάνουν υπόψη

τους την επίδοση των μαθητών στα μαθήματα αυτά και καθορίζουν αποφασιστικά τη στάση τους για την αξιολόγηση των μαθητών στα υπόλοιπα μαθήματα. Ο τελικός βαθμός δηλαδή είναι ο ίδιος ή περίπου ο ίδιος με αυτόν των Μαθηματικών ή της Ελληνικής Γλώσσας. Έχει διαπιστωθεί επίσης ότι τα Μαθηματικά δυσκολεύουν περισσότερο τους μαθητές σε σχέση με τα υπόλοιπα μαθήματα. Ενδιαφέρον παρουσιάζει η εύρεση και περιγραφή της σχέσης μεταξύ του βαθμού του Απολυτηρίου (εξαρτημένη μεταβλητή) και του βαθμού στα Μαθηματικά (ανεξάρτητη μεταβλητή).

6.2.1 Δημιουργία μοντέλου για το σύνολο των μαθητών

Θα καθοριστεί αν από τα υπό ανάλυση δεδομένα προκύπτει γραμμική σχέση μεταξύ του βαθμού του Απολυτηρίου και του βαθμού στα μαθηματικά, όπως αυτός αναγράφεται στο απολυτήριο. Δηλαδή διερευνάται αν ο βαθμός στα μαθηματικά είναι ικανός να προβλέψει το βαθμό που έχουν οι μαθητές στο απολυτήριό τους.

Ο παρακάτω πίνακας μας επιτρέπει να εξετάσουμε αν οι δυο μεταβλητές συνδέονται με γραμμική σχέση, επιπλέον δε πόσο στενή είναι η σχέση αυτή.

Πίνακας 43
Συσχέτιση βαθμού μαθηματικών και βαθμού απολυτηρίου

		ΒΑΘΜΟΣ ΜΑΘΗΜΑΤΙΚΩΝ	ΒΑΘΜΟΣ ΑΠΟΛΥΤΗΡΙΟΥ
ΒΑΘΜΟΣ ΜΑΘΗΜΑΤΙΚΩΝ	Pearson (συσχέτιση)	1,000	,935
	Παρατηρούμενο επίπεδο σημαντικότητας	,	,000
	N	343	343
ΒΑΘΜΟΣ ΑΠΟΛΥΤΗΡΙΟΥ	Pearson (συσχέτιση)	,935	1,000
	Παρατηρούμενο επίπεδο σημαντικότητας	,000	,
	N	343	343

Φαίνεται ότι ο συντελεστής συσχέτισης είναι 0,935. Επομένως υπάρχει μεγάλη θετική γραμμική σχέση μεταξύ του βαθμού απολυτηρίου των μαθητών και του βαθμού που έχουν στα μαθηματικά. Ο συντελεστής αυτός δείχνει ότι οι μαθητές με μεγάλους βαθμούς στα μαθηματικά έχουν κατά κανόνα μεγάλους βαθμούς και στο απολυτήριο, οι μαθητές με μέσους βαθμούς στα

μαθηματικά έχουν μέσους βαθμούς και στο απολυτήριο και οι μαθητές με μικρούς βαθμούς στα μαθηματικά έχουν μικρούς βαθμούς και στο απολυτήριο. Από τον παραπάνω πίνακα φαίνεται ότι το παρατηρούμενο επίπεδο σημαντικότητας, που αντιστοιχεί στο στατιστικό έλεγχο για την ύπαρξη ή όχι γραμμικής σχέσης μεταξύ των δυο μεταβλητών, με μηδενική υπόθεση H_0 : ο συντελεστής συσχέτισης είναι 0 έναντι της εναλλακτικής H_1 : ο συντελεστής συσχέτισης διαφέρει από το 0, έχει τιμή 0. Επομένως απορρίπτεται η μηδενική υπόθεση σε επίπεδο σημαντικότητας 5% και γίνεται δεκτή η εναλλακτική.

Θα προσπαθήσουμε να βρούμε μια εκτίμηση της ευθείας παλινδρόμησης, έστω $\hat{y} = a + b \cdot x$. Πρέπει να προσδιορίσουμε τους συντελεστές a και b (οι οποίοι είναι σημειακές εκτιμήσεις των α και β), ώστε η γραμμή $\hat{y} = a + b \cdot x$ να είναι όσο το δυνατό πλησιέστερα προς την πραγματική ευθεία παλινδρόμησης της οποίας η συναρτησιακή μορφή είναι E(Y) = α + βX .

Ο παρακάτω πίνακας δίνει πληροφορίες για τις εκτιμήσεις των παραμέτρων α και β του γραμμικού μοντέλου.

Πίνακας 44
Εκτίμηση παραμέτρων

	Σημειακές εκτιμήσεις παραμέτρων				95% διάστημα εμπιστοσύνης εκτιμήσεων παραμέτρων	
	B	Τυπικά σφάλματα εκτιμήσεων	t	Παρατηρούμενο επίπεδο σημαντικότητας	Κατώτερο όριο	Ανώτερο όριο
(σταθερά)	,770	,162	4,759	,000	,452	1,089
βαθμός μαθηματικών	,904	,019	48,7	,000	,868	,941

Οι εκτιμήσεις των παραμέτρων α , β είναι a = 0,770 και b = 0,904. Τα τυπικά σφάλματα των εκτιμώμενων παραμέτρων είναι 0,162 και 0,019 αντίστοιχα. Στην τέταρτη στήλη υπάρχουν οι τιμές των ελεγχοσυναρτήσεων t για κάθε παράμετρο, οι οποίες ελέγχουν εάν κάθε παράμετρος διαφέρει στατιστικά από το μηδέν. Από τον παραπάνω πίνακα φαίνεται ότι τα παρατηρούμενα επίπεδα σημαντικότητας, που αντιστοιχούν στο στατιστικό έλεγχο για τη στατιστική σημαντικότητα ή όχι των παραμέτρων, με μηδενική υπόθεση H_0 : η παράμετρος είναι μηδέν έναντι της εναλλακτικής H_1 : η παράμετρος

διαφέρει από το μηδέν, έχουν τιμή 0. Επομένως απορρίπτονται οι μηδενικές υποθέσεις σε επίπεδο σημαντικότητας 5% και γίνονται δεκτές οι εναλλακτικές. Στις τελευταίες στήλες υπάρχουν τα διαστήματα εμπιστοσύνης των παραμέτρων του μοντέλου. Τα 95% διαστήματα εμπιστοσύνης είναι : (0,452 , 1,089) για το α και (0,868 , 0,941) για το β.

Ο παρακάτω πίνακας δίνει πληροφορίες για τους συντελεστές προσδιορισμού του γραμμικού μοντέλου.

Πίνακας 45
Συντελεστές προσδιορισμού

	Συντελεστής συσχέτισης	Συντελεστής προσδιορισμού	Προσαρμοσμένος συντελεστής προσδιορισμού	Τυπικό σφάλμα της εκτίμησης
	,935	,874	,874	,4550

Ο συντελεστής προσδιορισμού, όπως και ο προσαρμοσμένος συντελεστής προσδιορισμού, έχουν την τιμή 0,874. Οι υψηλές τιμές των συντελεστών δείχνουν ότι το μοντέλο ερμηνεύει πολύ καλά τις μεταβολές της εξαρτημένης μεταβλητής με τη βοήθεια της ανεξάρτητης. Το 87,4% της μεταβλητότητας που υπάρχει ερμηνεύεται με το μοντέλο. Η βασιμότητα ή όχι της υπόθεσης της γραμμικής παλινδρόμησης μπορεί να ελεγχθεί και από τον πίνακα ανάλυσης διακύμανσης (ANOVA).

Πίνακας 46
Ανάλυση διακύμανσης

Πηγή μεταβλητότητας		Άθροισμα τετραγώνων	Βαθμοί ελευθερίας	Μέσο τετράγωνο	Τιμή F	Παρατηρούμενο επίπεδο σημαντικότητας
	Παλινδρόμηση	491,706	1	491,706	2375,051	,000
	Λάθη	70,597	341	,207		
	Σύνολο	562,304	342			

Η συνολική μεταβλητότητα των δεδομένων χωρίζεται και αποδίδεται εν μέρει στο μοντέλο και εν μέρει στα λάθη. Η μεταβλητότητα της εξαρτημένης μεταβλητής μετράται με βάση το άθροισμα των παρακάτω αποκλίσεων :

- Μεταβλητότητα της εξαρτημένης εξαιτίας της ανεξάρτητης μεταβλητής :

$$SSR = \sum\left(\hat{Y}_i - \overline{Y}\right)^2 = 491{,}706$$

- Μεταβλητότητα της εξαρτημένης εξαιτίας τυχαίων παραγόντων :

$$SSE = \sum\left(Y_i - \hat{Y}_i\right)^2 = 70{,}597$$

- Συνολική μεταβλητότητα της εξαρτημένης μεταβλητής :

$$SST = \sum\left(Y_i - \overline{Y}\right)^2 = 562{,}304$$

Ο πίνακας μας δίνει τους βαθμούς ελευθερίας για τη μεταβλητότητα εξαιτίας της παλινδρόμησης (αριθμός μεταβλητών (k)-1), εξαιτίας των λαθών (αριθμός παρατηρήσεων (n) - αριθμός μεταβλητών (k)) και για το σύνολο (n-1) . Διαιρώντας τα αθροίσματα των τετραγώνων της δεύτερης στήλης με τους αντίστοιχους βαθμούς ελευθερίας παίρνουμε τα αντίστοιχα μέσα τετράγωνα (διακυμάνσεις). Η προτελευταία στήλη μας δίνει το λόγο F ως τη διαίρεση της διακύμανσης από τις μεταβολές της ανεξάρτητης μεταβλητής προς τη διακύμανση από τα λάθη. Η τιμή της ελεγχοσυνάρτησης F και το αντίστοιχο παρατηρούμενο επίπεδο σημαντικότητας αντιστοιχούν στο στατιστικό έλεγχο της μηδενικής υπόθεσης $H_0 : \beta = 0$ έναντι της εναλλακτικής $H_1 : \beta \neq 0$. Απορρίπτεται η μηδενική υπόθεση σε επίπεδο σημαντικότητας 5% και γίνεται δεκτή η εναλλακτική.

Προκειμένου να ελεγχθεί ότι τα αποτελέσματα της παλινδρόμησης είναι αξιόπιστα πρέπει να γίνει έλεγχος προϋποθέσεων σχετικά με τα σφάλματα εκτίμησης (κατάλοιπα). Τα σφάλματα πρέπει να κατανέμονται κανονικά να είναι ομοσκεδαστικά και τυχαία.

<u>Έλεγχος κανονικότητας καταλοίπων</u>

Ο έλεγχος της υπόθεσης της κανονικότητας των καταλοίπων θα γίνει με τη χρήση του ελέγχου Kolmogorov - Smirnov.

Οι προς έλεγχο υποθέσεις είναι :

H_0 : Τα κατάλοιπα κατανέμονται κανονικά

H_1 : Τα κατάλοιπα δεν κατανέμονται κανονικά

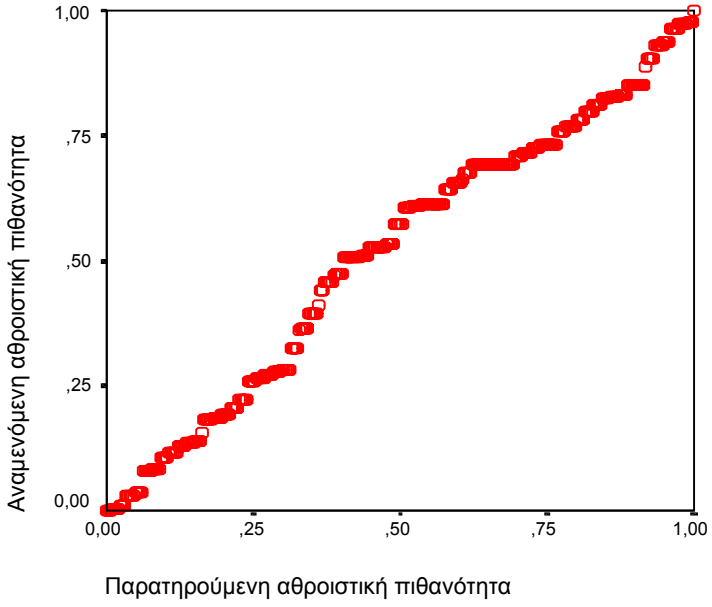
Πίνακας 47
Έλεγχος κανονικότητας καταλοίπων

	Kolmogorov-Smirnov[a]		
	Statistic	Βαθμοί ελευθερίας	Παρατηρούμενο επίπεδο σημαντικότητας
Τυποποιημένα κατάλοιπα	,046	343	,081

a. Lilliefors Significance Correction

Σε επίπεδο σημαντικότητας 5% δεν απορρίπτεται η μηδενική υπόθεση, δηλαδή τα κατάλοιπα κατανέμονται κανονικά.

Διάγραμμα 24
Έλεγχος κανονικότητας καταλοίπων



Από το παραπάνω διάγραμμα φαίνεται ότι τα κατάλοιπα είναι συγκεντρωμένα (σχεδόν) γύρω από μια ευθεία.

<u>Έλεγχος ομοσκεδαστικότητας και τυχαιότητας των καταλοίπων</u>

Η υπόθεση των ίσων διασπορών ονομάζεται πολλές φορές υπόθεση της ομοιογένειας των διασπορών ή ομοσκεδαστικότητα. Όταν υπάρχει ανισότητα

διασπορών το "άπλωμα" του y τείνει, συχνά, να αυξάνεται όσο μεγαλύτερες είναι οι τιμές της μεταβλητής χ. Ένας τρόπος ελέγχου της ισότητας των διασπορών είναι να χωριστούν τα κατάλοιπα σε δυο ομάδες, η μια από τις οποίες θα προέρχεται από τις "μικρές" τιμές του χ και η άλλη από τις "μεγάλες" τιμές του χ. Στη συνέχεια υπολογίζονται οι δειγματικές διασπορές που αντιστοιχούν στις δυο ομάδες των καταλοίπων. Κατόπιν χρησιμοποιείται ο έλεγχος F για την ισότητα δυο διασπορών, ώστε να ελεγχθεί αν τα δυο σύνολα των καταλοίπων έχουν ίσες διασπορές στους αντίστοιχους πληθυσμούς. Η ομοσκεδαστικότητα των καταλοίπων φαίνεται από τον παρακάτω πίνακα, κάνοντας τον έλεγχο :

H_0 : $S_1^2 = S_2^2$, οι δειγματικές διασπορές είναι ίσες

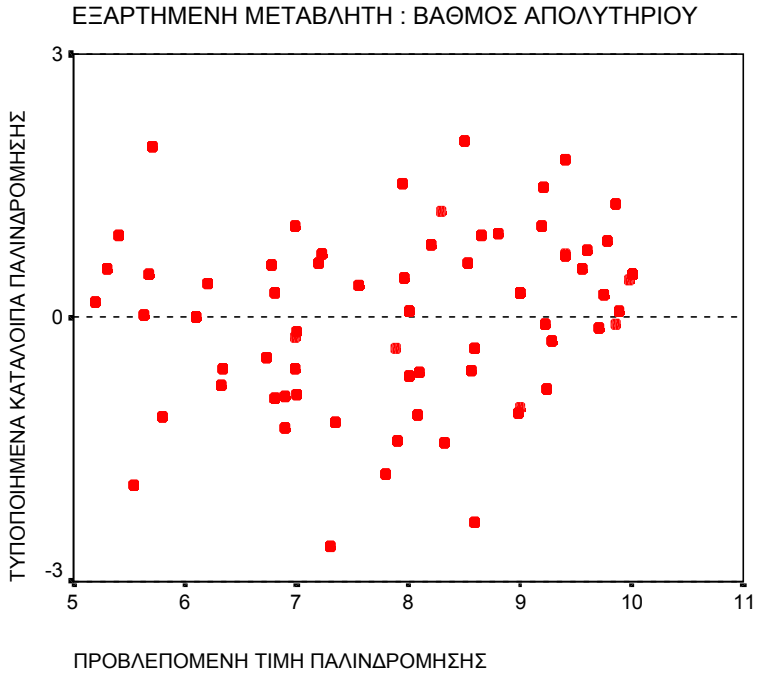
H_1 : $S_1^2 \neq S_2^2$, οι δειγματικές διασπορές είναι άνισες

Πίνακας 48
Έλεγχος ισότητας διασπορών των καταλοίπων

		Levene's Test για ισότητα διασπορών	
		F	Παρατηρούμενο επίπεδο σημαντικότητας
Τυποποιημένα κατάλοιπα	Προϋπόθεση ίσων διασπορών	,754	,386

Σε επίπεδο σημαντικότητας 5% δεν απορρίπτεται η μηδενική υπόθεση και γίνεται δεκτό ότι οι διασπορές των λαθών είναι ίσες. Από το παρακάτω διάγραμμα φαίνεται ότι τα κατάλοιπα είναι ομοιόμορφα κατανεμημένα και δεν παρουσιάζουν ανομοιογένεια των διασπορών. Επίσης η γραφική παράσταση των καταλοίπων δείχνει μια τυχαία τοποθέτησή τους γύρω από τη γραμμή που αντιστοιχεί σε κατάλοιπο "0".

Διάγραμμα 25
Γραφική παράσταση των καταλοίπων



Η τυχαιότητα των καταλοίπων φαίνεται και από τον παρακάτω πίνακα, ελέγχοντας τη μηδενική υπόθεση ότι τα κατάλοιπα είναι τυχαία, έναντι της εναλλακτικής ότι δεν είναι.

Πίνακας 49
Έλεγχος των ροών

	Τυποποιημένα κατάλοιπα
Test Value [a]	,1825895
Cases < Test Value	168
Cases >= Test Value	175
Total Cases	343
Number of Runs	164
Z	-,912
Παρατηρούμενο επίπεδο σημαντικότητας	,362

a. Median

Δεν απορρίπτεται η μηδενική υπόθεση και γίνεται δεκτό ότι τα κατάλοιπα είναι τυχαία.

Η εξίσωση της ευθείας ελάχιστων τετραγώνων είναι η :

$$\hat{Y} = 0,770 + 0,904 \cdot X$$

Σύμφωνα με το μοντέλο αυτό, η αύξηση της βαθμολογίας στα μαθηματικά κατά μια μονάδα, θα επιφέρει αύξηση του βαθμού στο απολυτήριο των μαθητών κατά 0,904 μονάδες. Η τιμή 0,770 είναι το σημείο στο οποίο η ευθεία ελάχιστων τετραγώνων τέμνει τον άξονα Y για X = 0.

6.2.2 Δημιουργία μοντέλου διαφοροποιημένο ως προς το φύλο

Στο πέμπτο κεφάλαιο διαπιστώθηκε ότι η επίδοση των μαθητών, τόσο ως προς το βαθμό του απολυτηρίου όσο και ως προς το βαθμό των μαθηματικών, διαφοροποιείται ως προς το φύλο. Οι μαθήτριες παρουσιάζονται να υπερτερούν έναντι των μαθητών τόσο στα μαθηματικά και την ελληνική γλώσσα όσο και στη γενική επίδοση. Ενδιαφέρον παρουσιάζει η εύρεση και περιγραφή της σχέσης μεταξύ του βαθμού του απολυτηρίου (εξαρτημένη μεταβλητή) και του βαθμού στα μαθηματικά (ανεξάρτητη μεταβλητή), διαφοροποιημένη όμως ως προς το φύλο.

Ο πίνακας συσχέτισης δείχνει αν οι δυο μεταβλητές συνδέονται με γραμμική σχέση.

Πίνακας 50

Συσχέτιση βαθμού μαθηματικών και βαθμού απολυτηρίου ως προς το φύλο

ΦΥΛΟ			ΒΑΘΜΟΣ ΜΑΘΗΜΑΤΙΚΩΝ	ΒΑΘΜΟΣ ΑΠΟΛΥΤΗΡΙΟΥ
ΜΑΘΗΤΡΙΕΣ	ΒΑΘΜΟΣ ΜΑΘΗΜΑΤΙΚΩΝ	Pearson (συσχέτιση)	1,000	,937**
		Παρατηρούμενο επίπεδο σημαντικότητας	,	,000
		N	168	168
	ΒΑΘΜΟΣ ΑΠΟΛΥΤΗΡΙΟΥ	Pearson (συσχέτιση)	,937**	1,000
		Παρατηρούμενο επίπεδο σημαντικότητας	,000	,
		N	168	168
ΜΑΘΗΤΕΣ	ΒΑΘΜΟΣ ΜΑΘΗΜΑΤΙΚΩΝ	Pearson (συσχέτιση)	1,000	,933**
		Παρατηρούμενο επίπεδο σημαντικότητας	,	,000
		N	175	175
	ΒΑΘΜΟΣ ΑΠΟΛΥΤΗΡΙΟΥ	Pearson (συσχέτιση)	,933**	1,000
		Παρατηρούμενο επίπεδο σημαντικότητας	,000	,
		N	175	175

**. Η συσχέτιση είναι σημαντική στο 0,01 επίπεδο σημαντικότητας

Ο συντελεστής συσχέτισης είναι 0,937 για τις μαθήτριες και 0,933 για τους μαθητές. Υπάρχει μεγάλη θετική γραμμική σχέση μεταξύ του βαθμού απολυτηρίου των μαθητών και του βαθμού που έχουν στα μαθηματικά. Από τον παραπάνω πίνακα φαίνεται ότι το παρατηρούμενο επίπεδο σημαντικότητας, που αντιστοιχεί στο στατιστικό έλεγχο για την ύπαρξη ή όχι γραμμικής σχέσης μεταξύ των δυο μεταβλητών, με μηδενική υπόθεση H_0 : ο συντελεστής συσχέτισης είναι 0 , έναντι της εναλλακτικής H_1 : ο συντελεστής συσχέτισης διαφέρει από το 0, έχει τιμή 0, τόσο για τους μαθητές όσο και για τις μαθήτριες. Επομένως απορρίπτεται η μηδενική υπόθεση σε επίπεδο σημαντικότητας 5% και γίνεται δεκτή η ύπαρξη γραμμικής σχέσης μεταξύ της βαθμολογίας στα μαθηματικά και του βαθμού του απολυτηρίου.

Θα προσπαθήσουμε να βρούμε μια εκτίμηση της ευθείας παλινδρόμησης, έστω $\hat{y} = a + b \cdot x$. Πρέπει να προσδιορίσουμε τους συντελεστές a και b. Ο παρακάτω πίνακας δίνει πληροφορίες για τις εκτιμήσεις των παραμέτρων α και β του γραμμικού μοντέλου.

Πίνακας 51
Εκτίμηση παραμέτρων

ΦΥΛΟ	Μοντέλο		Σημειακές εκτιμήσεις παραμέτρων		t	Παρατηρούμενο επίπεδο σημαντικότητας	95% διάστημα εμπιστοσύνης εκτιμήσεων παραμέτρων	
			B	Τυπικά σφάλματα εκτιμήσεων			Κατώτερο όριο	Ανώτερο όριο
ΜΑΘΗΤΡΙΕΣ	1	(σταθερά)	1,071	,226	4,734	,000	,624	1,517
		βαθμός μαθηματικών	,879	,025	34,591	,000	,828	,929
ΜΑΘΗΤΕΣ	1	(σταθερά)	,640	,229	2,798	,006	,189	1,092
		βαθμός μαθηματικών	,911	,027	34,034	,000	,858	,964

Οι εκτιμήσεις των παραμέτρων α , β είναι a = 1,071 , b = 0,879 για τις μαθήτριες και a = 0,640 , b = 0,911 για τους μαθητές. Τα τυπικά σφάλματα των εκτιμώμενων παραμέτρων είναι 0,226 και 0,025 για τις μαθήτριες και 0,229 και 0,027 για τους μαθητές. Στην τέταρτη στήλη υπάρχουν οι τιμές των ελεγχοσυναρτήσεων t για κάθε παράμετρο, οι οποίες ελέγχουν εάν κάθε παράμετρος διαφέρει στατιστικά από το μηδέν. Από τον παραπάνω πίνακα

φαίνεται ότι τα παρατηρούμενα επίπεδα σημαντικότητας, που αντιστοιχούν στο στατιστικό έλεγχο για τη στατιστική σημαντικότητα ή όχι των παραμέτρων, με μηδενική υπόθεση H_0 : η παράμετρος είναι μηδέν έναντι της εναλλακτικής H_1 : η παράμετρος διαφέρει από το μηδέν, έχουν τιμή 0. Επομένως απορρίπτονται οι μηδενικές υποθέσεις σε επίπεδο σημαντικότητας 5% και γίνονται δεκτές οι εναλλακτικές, δηλαδή οι παράμετροι διαφέρουν στατιστικά σημαντικά από το μηδέν. Ο πίνακας δίνει διαστήματα εμπιστοσύνης των παραμέτρων του κάθε μοντέλου. Τα 95% διαστήματα εμπιστοσύνης είναι :

(0,624 , 1,517) για το α και (0,828 , 0,929) για το β (μαθήτριες)

(0,189 , 1,092) για το α και (0,858 , 0,964) για το β (μαθητές)

Ο παρακάτω πίνακας δίνει πληροφορίες για τους συντελεστές προσδιορισμού του γραμμικού μοντέλου.

Πίνακας 52
Συντελεστές προσδιορισμού

ΦΥΛΟ	Μοντέλο	Συντελεστής συσχέτισης	Συντελεστής προσδιορισμού	Προσαρμοσμένος συντελεστής προσδιορισμού	Τυπικό σφάλμα της εκτίμησης
ΜΑΘΗΤΡΙΕΣ	1	,937	,878	,877	,4179
ΜΑΘΗΤΕΣ	1	,933	,870	,869	,4780

Οι υψηλές τιμές των συντελεστών δείχνουν ότι τα μοντέλα ερμηνεύουν πολύ καλά τις μεταβολές της εξαρτημένης μεταβλητής με τη βοήθεια της ανεξάρτητης. Το 87,8% και το 87% της μεταβλητότητας που υπάρχει ερμηνεύεται με τα μοντέλα. Η βασιμότητα ή όχι της υπόθεσης της γραμμικής παλινδρόμησης και για τα δυο μοντέλα μπορεί να ελεγχθεί και από τον πίνακα ανάλυσης διακύμανσης (ANOVA).

Πίνακας 53
Ανάλυση διακύμανσης

ΦΥΛΟ	Πηγή μεταβλητότητας	Άθροισμα τετραγώνων	Βαθμοί ελευθερίας	Μέσο τετράγωνο	Τιμή F	Παρατηρούμενο επίπεδο σημαντικότητας
ΜΑΘΗΤΡΙΕΣ	Παλινδρόμηση	208,987	1	208,987	1196,6	,000
	Λάθη	28,993	166	,175		
	Σύνολο	237,980	167			
ΜΑΘΗΤΕΣ	Παλινδρόμηση	264,664	1	264,664	1158,3	,000
	Λάθη	39,529	173	,228		
	Σύνολο	304,193	174			

Ο πίνακας ανάλυσης διακύμανσης δίνει πληροφορίες για τη συνολική μεταβλητότητα των δεδομένων. Τα παρατηρούμενα επίπεδα σημαντικότητας έχουν τιμή μηδέν, οπότε σε επίπεδο σημαντικότητας 5% γίνεται δεκτό ότι υπάρχει γραμμική σχέση.

Τα αποτελέσματα της παλινδρόμησης είναι αξιόπιστα με την προϋπόθεση ότι τα κατάλοιπα κατανέμονται κανονικά, είναι ομοσκεδαστικά και τυχαία.

Έλεγχος κανονικότητας καταλοίπων

Με τον έλεγχο Kolmogorov - Smirnov ελέγχονται οι υποθέσεις :

H_0 : Τα κατάλοιπα κατανέμονται κανονικά

H_1 : Τα κατάλοιπα δεν κατανέμονται κανονικά

Πίνακας 54
Έλεγχος κανονικότητας καταλοίπων

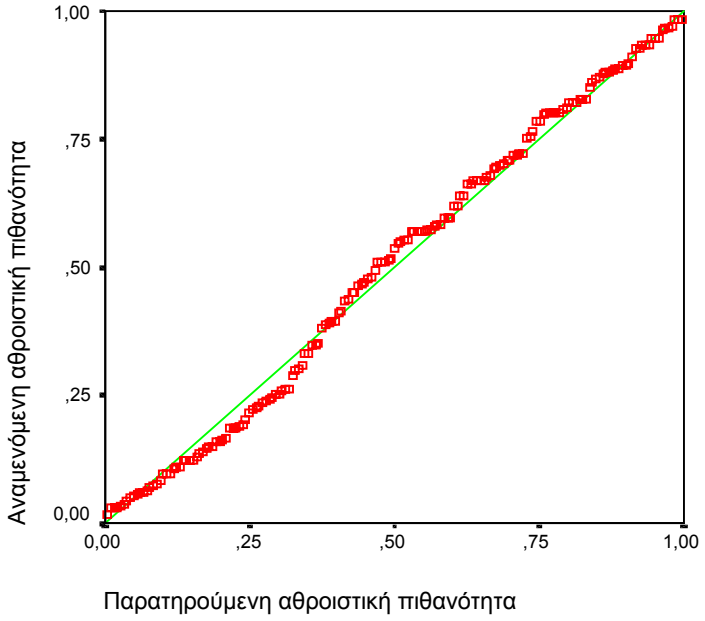
		Kolmogorov-Smirnov[a]		
	ΦΥΛΟ	Statistic	Βαθμοί ελευθερίας	Παρατηρούμενο επίπεδο σημαντικότητας
Τυποποιημένα κατάλοιπα	ΜΑΘΗΤΡΙΕΣ	,058	168	,200*
	ΜΑΘΗΤΕΣ	,055	175	,200*

*. This is a lower bound of the true significance.

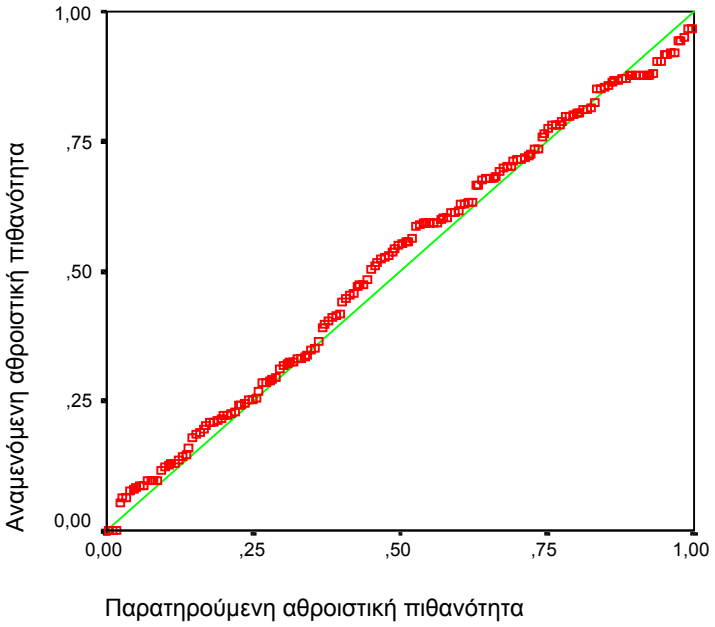
a. Lilliefors Significance Correction

Δεν απορρίπτεται η μηδενική υπόθεση, δηλαδή τα κατάλοιπα ακολουθούν την κανονική κατανομή. Η υπόθεση της κανονικότητας των καταλοίπων μπορεί να ελεγχθεί και διαγραμματικά. Στα παρακάτω διαγράμματα φαίνεται ότι δεν υπάρχει ένδειξη παραβίασης της υπόθεσης της κανονικότητας των καταλοίπων, αφού είναι συγκεντρωμένα γύρω από μια ευθεία.

Διάγραμμα 26
Έλεγχος κανονικότητας καταλοίπων για τις μαθήτριες



Διάγραμμα 27
Έλεγχος κανονικότητας καταλοίπων για τους μαθητές



Έλεγχος ομοσκεδαστικότητας και τυχαιότητας των καταλοίπων

Η ομοσκεδαστικότητα των καταλοίπων φαίνεται από τον παρακάτω πίνακα, κάνοντας τον έλεγχο :

H_0 : $S_1^2 = S_2^2$, οι δειγματικές διασπορές είναι ίσες

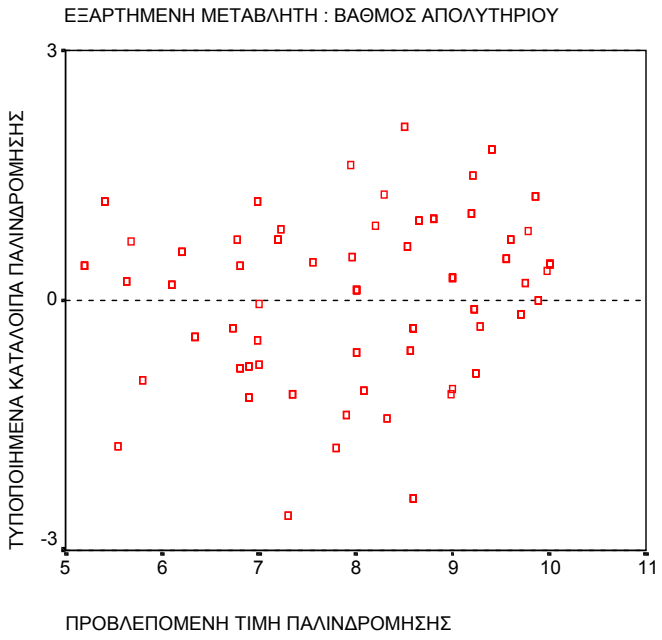
H_1 : $S_1^2 \neq S_2^2$, οι δειγματικές διασπορές είναι άνισες

Πίνακας 55
Έλεγχος ισότητας διασπορών των καταλοίπων

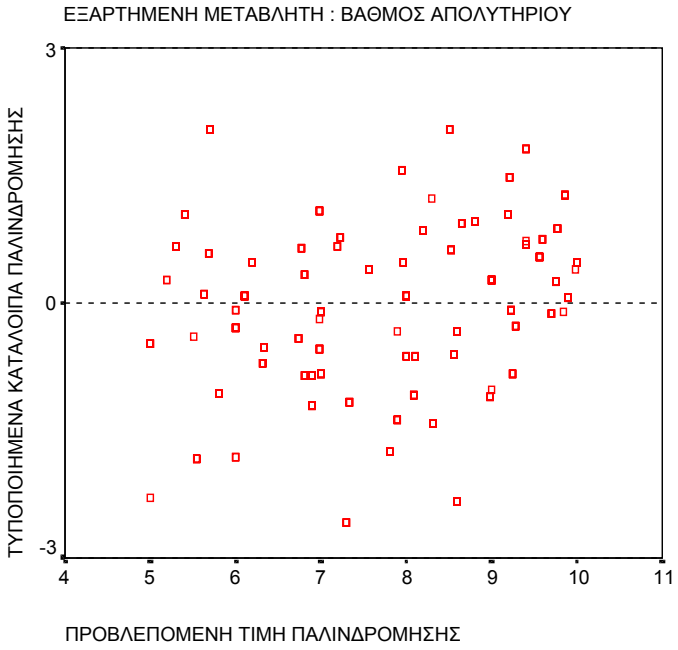
			Levene's Test για ισότητα διασπορών	
ΦΥΛΟ			F	Παρατηρούμενο επίπεδο σημαντικότητας
ΜΑΘΗΤΡΙΕΣ	Τυποποιημένα κατάλοιπα	Προϋπόθεση ίσων διασπορών	,157	,692
ΜΑΘΗΤΕΣ	Τυποποιημένα κατάλοιπα	Προϋπόθεση ίσων διασπορών	,514	,474

Τα παρατηρούμενα επίπεδα σημαντικότητας έχουν τιμή 0,692 και 0,474 για τις μαθήτριες και τους μαθητές αντίστοιχα. Σε οποιοδήποτε επίπεδο σημαντικότητας δεν απορρίπτεται η μηδενική υπόθεση, δηλαδή δεχόμαστε ότι οι διασπορές των καταλοίπων είναι ίσες. Τα κατάλοιπα είναι ομοιόμορφα κατανεμημένα και δεν παρουσιάζουν ανομοιογένεια των διασπορών. Η γραφική παράσταση των καταλοίπων δείχνει μια τυχαία τοποθέτησή τους γύρω από τη γραμμή που αντιστοιχεί σε κατάλοιπο "0".

Διάγραμμα 28
Γραφική παράσταση των καταλοίπων για τις μαθήτριες



Διάγραμμα 29
Γραφική παράσταση των καταλοίπων για τους μαθητές



Η τυχαιότητα των καταλοίπων φαίνεται και από τον παρακάτω πίνακα, ελέγχοντας τη μηδενική υπόθεση ότι τα κατάλοιπα είναι τυχαία, έναντι της εναλλακτικής ότι δεν είναι.

Πίνακας 56
Έλεγχος των ροών

ΦΥΛΟ		Τυποποιημένα κατάλοιπα
ΜΑΘΗΤΡΙΕΣ	Test Value	,2259687
	Cases < Test Value	83
	Cases >= Test Value	85
	Total Cases	168
	Number of Runs	90
	Z	,776
	Παρατηρούμενο επίπεδο σημαντικότητας	,438
ΜΑΘΗΤΕΣ	Test Value	,2778335
	Cases < Test Value	87
	Cases >= Test Value	88
	Total Cases	175
	Number of Runs	87
	Z	-,227
	Παρατηρούμενο επίπεδο σημαντικότητας	,820

Σε οποιοδήποτε επίπεδο σημαντικότητας δεν απορρίπτεται η μηδενική υπόθεση και γίνεται δεκτό ότι τα κατάλοιπα είναι τυχαία.

Η εξίσωση της ευθείας ελάχιστων τετραγώνων είναι η :

$$\hat{Y} = 1{,}071 + 0{,}879 \cdot X \quad \text{για τις μαθήτριες}$$

$$\hat{Y} = 0{,}640 + 0{,}911 \cdot X \quad \text{για τους μαθητές}$$

και εκφράζει τη σχέση μεταξύ των βαθμών στα μαθηματικά και των βαθμών του απολυτηρίου ,τόσο για τις μαθήτριες όσο για τους μαθητές.

Σύμφωνα με τα μοντέλα αυτά, η αύξηση της βαθμολογίας στα μαθηματικά κατά μια μονάδα, θα επιφέρει αύξηση του βαθμού στο απολυτήριο των μαθητριών κατά 0,879 μονάδες, ενώ αύξηση της βαθμολογίας στα μαθηματικά κατά μια μονάδα, θα επιφέρει αύξηση του βαθμού στο απολυτήριο των μαθητών κατά 0,911 μονάδες.

6.3 Γραμμικό μοντέλο με εικονικές μεταβλητές

Ο όρος ανάλυση παλινδρόμησης χρησιμοποιείται όταν τόσο το Y όσο και τα X είναι ποσοτικές μεταβλητές. Σε πολλές περιπτώσεις όμως τα προβλήματα που μελετώνται αναφέρονται και σε ποιοτικές μεταβλητές οι οποίες πρέπει να ποσοτικοποιηθούν προκειμένου να μελετηθούν. Οι μεταβλητές που χρησιμοποιούνται στις εξισώσεις παλινδρόμησης είναι, συνήθως συνεχείς. Πολλές φορές όμως χρειάζεται να χρησιμοποιηθεί κάποιος παράγοντας που εμφανίζεται σε δυο ή περισσότερα διακεκριμένα επίπεδα. Χρειάζεται δηλαδή να ορισθούν μεταβλητές που θα προσδώσουν αριθμητική έκφραση σε ποιοτικά (κατηγορικά) χαρακτηριστικά. Για τις περιπτώσεις αυτές χρησιμοποιούνται οι λεγόμενες *εικονικές μεταβλητές* ή *ψευδομεταβλητές*. Οι μεταβλητές αυτές συνήθως χρησιμοποιούνται για να εκφράσουν δύο κατηγορίες (επίπεδα), οπότε είναι δίτιμες. Η συνήθης επιλογή για τον ορισμό μιας δίτιμης ψευδομεταβλητής είναι η χρησιμοποίηση μιας μεταβλητής δείκτη (0-1), η οποία δείχνει αν μια συγκεκριμένη παρατήρηση ανήκει σ' ένα από τα δύο καθορισμένα επίπεδα, ή κατηγορίες μιας κατηγορικής μεταβλητής. Οι ψευδομεταβλητές χρησιμοποιούνται επίσης για να εκφράσουν μια ποιοτική μεταβλητή που παίρνει τιμές σε περισσότερες από δύο κατηγορίες (επίπεδα). (Ι. Πανάρετος, 2001)

Αν χρειάζεται να περιληφθεί σ' ένα μοντέλο παλινδρόμησης με σταθερό όρο μια κατηγορική (ποιοτική) μεταβλητή με k επίπεδα (κατηγορίες), πρέπει να οριστούν k-1 ψευδομεταβλητές για να εκφράσουν την κατηγορική μεταβλητή. Όταν ορισθούν οι k-1 ψευδομεταβλητές για μια ποιοτική μεταβλητή με k επίπεδα, το επίπεδο (κατηγορία) που μένει ονομάζεται *κατηγορία αναφοράς* ή *κατηγορία βάσης*.

Η επιλογή της κατηγορίας αναφοράς εξαρτάται από τη φύση του προβλήματος, γιατί συγκεκριμένες επιλογές μπορεί να οδηγήσουν σε μια καλύτερη ερμηνεία των συντελεστών παλινδρόμησης. Αυτό συμβαίνει όταν γίνονται συγκρίσεις με κάποια "ελεγχόμενη" ομάδα η οποία χρησιμοποιείται ως κατηγορία αναφοράς.

Το γενικό κίνητρο για να συμπεριληφθεί μια εικονική μεταβλητή σ' ένα πρόβλημα παλινδρόμησης είναι το ίδιο μ' εκείνο που οδηγεί στο να συμπεριληφθεί μια ποσοτική ανεξάρτητη μεταβλητή. Επομένως :

- Να μελετηθεί καλύτερα η εξαρτημένη μεταβλητή με την ελάττωση της επίδρασης του παράγοντα που οφείλεται στα λάθη.
- Να αποτραπεί μια μεροληπτική αποτίμηση της επίδρασης μιας ανεξάρτητης μεταβλητής που είναι απόρροια του ότι έχει παραληφθεί από το μοντέλο μια άλλη ανεξάρτητη μεταβλητή η οποία σχετίζεται με αυτή.

Όπως αναφέρθηκε παραπάνω, χαρακτηριστικότερη περίπτωση χρησιμοποίησης εικονικής μεταβλητής είναι όταν θέλουμε να μελετήσουμε ποιοτικά δεδομένα που είναι ενδεχόμενο να έχουν ένα από δύο διαφορετικά χαρακτηριστικά. Σε τέτοιες περιπτώσεις χρησιμοποιούμε τη μεταβλητή - δείκτη 0-1 ως εικονική μεταβλητή. Όπως είναι προφανές, η ψευδομεταβλητή χρησιμοποιείται ανάλογα με το αν υφίσταται ή όχι κάποιο συγκεκριμένο χαρακτηριστικό.

6.3.1 Δημιουργία μοντέλου για το σύνολο των μαθητών

Μας ενδιαφέρει να εξετάσουμε, αν, μελετώντας τους βαθμούς στην ελληνική γλώσσα και τα μαθηματικά,, υπάρχουν σημαντικές διαφορές στον τελικό βαθμό απόλυσης (βαθμός απολυτηρίου) μεταξύ μαθητών και μαθητριών, πρωτότοκων ή όχι, που φοιτούν σε δημόσια ή ιδιωτικά σχολεία, με διαζευγμένους ή όχι γονείς.

Μας ενδιαφέρει επομένως να εκτιμήσουμε την εξίσωση παλινδρόμησης :

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 D_1 + \beta_4 D_2 + \beta_5 D_3 + \beta_6 D_4 + \varepsilon$$

Όπου Y είναι η τελική επίδοση των μαθητών, δηλαδή ο βαθμός του απολυτηρίου, X_1 είναι ο βαθμός στην ελληνική γλώσσα και X_2 είναι ο βαθμός στα μαθηματικά.

Η δίτιμη ψευδομεταβλητή η οποία εκφράζει το φύλο ορίζεται ως :

$$D_1 = \begin{cases} 1 & \textit{για} \quad \textit{μαθητές} \\ 0 & \textit{για} \quad \textit{μαθήτριες} \end{cases}$$

όπου η κατηγορία των μαθητριών είναι η κατηγορία αναφοράς.

Επίσης για την έκφραση της οικογενειακής κατάστασης, η μεταβλητή ορίζεται ως :

$$D_2 = \begin{cases} 1 & \textit{γονείς διαζευγμένοι} \\ 0 & \textit{μη διαζευγμένοι γονείς} \end{cases}$$

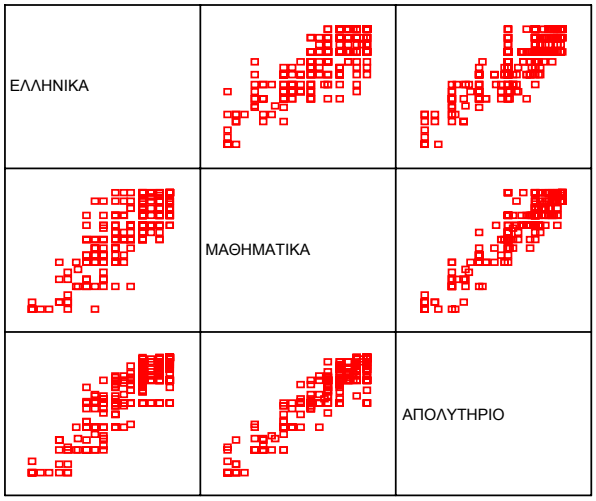
Τέλος για την έκφραση της σειράς γέννησης και του είδους του σχολείου φοίτησης, οι δίτιμες ψευδομεταβλητές μεταβλητή ορίζονται ως :

$$D_3 = \begin{cases} 1 & \textit{πρωτότοκος} \\ 0 & \textit{άλλη σειρά γέννησης} \end{cases}$$

$$D_4 = \begin{cases} 1 & \textit{φοίτηση σε ιδιωτικό σχολείο} \\ 0 & \textit{φοίτηση σε δημόσιο σχολείο} \end{cases}$$

Από την ανάλυση των αποτελεσμάτων της έρευνας θα βρεθεί κατά πόσο οι καταγεγραμμένες μεταβλητές (ανεξάρτητες μεταβλητές) επηρεάζουν τη βαθμολογία του απολυτηρίου (εξαρτημένη μεταβλητή). Από το παρακάτω διάγραμμα διασποράς (της εξαρτημένης μεταβλητής με κάθε μια από τις ανεξάρτητες, όπως και μεταξύ των ανεξαρτήτων), φαίνεται ότι η εξαρτημένη μεταβλητή (απολυτήριο) έχει μεγάλη θετική γραμμική συσχέτιση με τις ανεξάρτητες (βαθμοί στα μαθηματικά και τα ελληνικά). Δε φαίνεται πρόβλημα πολυσυγγραμμικότητας μεταξύ των ανεξάρτητων μεταβλητών.

Διάγραμμα 30
Διασπορά ανεξάρτητης με κάθε μια από τις εξαρτημένες μεταβλητές



Ο παρακάτω πίνακας μας επιτρέπει να εξετάσουμε αν μεταβλητές συνδέονται με γραμμική σχέση, επιπλέον δε πόσο στενή είναι η σχέση αυτή.

Πίνακας 57
Συντελεστές γραμμικής συσχέτισης

		ΒΑΘΜΟΣ ΕΛΛΗΝΙΚΩΝ	ΒΑΘΜΟΣ ΜΑΘΗΜΑΤΙΚΩΝ	ΒΑΘΜΟΣ ΑΠΟΛΥΤΗΡΙΟΥ
ΒΑΘΜΟΣ ΕΛΛΗΝΙΚΩΝ	Pearson (συσχέτιση)	1,000	,852	,875
	Παρατηρούμενο επίπεδο σημαντικότητας	,	,000	,000
	Μέγεθος	343	343	343
ΒΑΘΜΟΣ ΜΑΘΗΜΑΤΙΚΩΝ	Pearson (συσχέτιση)	,852	1,000	,935
	Παρατηρούμενο επίπεδο σημαντικότητας	,000	,	,000
	Μέγεθος	343	343	343
ΒΑΘΜΟΣ ΑΠΟΛΥΤΗΡΙΟΥ	Pearson (συσχέτιση)	,875	,935	1,000
	Παρατηρούμενο επίπεδο σημαντικότητας	,000	,000	,
	Μέγεθος	343	343	343

Υπάρχει μεγάλη θετική γραμμική συσχέτιση τόσο μεταξύ των ανεξάρτητων μεταβλητών όσο και μεταξύ της εξαρτημένης με κάθε μια από τις ανεξάρτητες. Ο συντελεστής συσχέτισης μεταξύ των βαθμών στα μαθηματικά και των βαθμών που έχουν οι μαθητές στα ελληνικά είναι 0,852 , μεταξύ του βαθμού του απολυτηρίου και των βαθμών στα μαθηματικά είναι 0,935 και τέλος, μεταξύ του βαθμού απόλυσης και των βαθμών στην ελληνική γλώσσα 0,875. Από τον παραπάνω πίνακα φαίνεται ότι τα παρατηρούμενα επίπεδα σημαντικότητας, που αντιστοιχούν στο στατιστικό έλεγχο για την ύπαρξη ή όχι γραμμικής σχέσης, με μηδενική υπόθεση H_0 : ο συντελεστής συσχέτισης είναι μηδέν έναντι της εναλλακτικής H_1: ο συντελεστής συσχέτισης διαφέρει από το μηδέν, έχουν τιμή 0. Επομένως σε επίπεδο σημαντικότητας 5% απορρίπτονται οι μηδενικές υποθέσεις.

Προκειμένου να βρεθεί το βέλτιστο μοντέλο περιγραφής των δεδομένων, εφαρμόζουμε τη μέθοδο βηματικής παλινδρόμησης (Stepwise). Σύμφωνα με τη μέθοδο αυτή, ξεκινάμε με μοντέλο χωρίς καμία μεταβλητή και σταδιακά σε κάθε βήμα μπορούμε είτε να προσθέσουμε είτε να αφαιρέσουμε μεταβλητές, ανάλογα με τη σημαντικότητα αυτών.

Ο παρακάτω πίνακας μας δίνει πληροφορίες για τις εκτιμήσεις των παραμέτρων του μοντέλου.

Πίνακας 58
Εκτίμηση παραμέτρων

Μοντέλο		Σημειακές εκτιμήσεις παραμέτρων		t	Παρατηρούμενο επίπεδο σημαντικότητας	Διαγνωστικά	
		B	Τυπικά σφάλματα εκτιμήσεων			Ανοχή	Παράγοντες μεταβολής (VIF)
1	(σταθερά)	,770	,162	4,759	,000		
	βαθμός μαθηματικών	,904	,019	48,734	,000	1,000	1,000
2	(σταθερά)	,170	,163	1,046	,296		
	βαθμός μαθηματικών	,668	,032	20,780	,000	,274	3,649
	βαθμός ελληνικών	,302	,035	8,611	,000	,274	3,649
3	(σταθερά)	,248	,164	1,515	,131		
	βαθμός μαθηματικών	,668	,032	20,966	,000	,274	3,649
	βαθμός ελληνικών	,291	,035	8,308	,000	,270	3,702
	είδος σχολείου	,233	,089	2,628	,009	,951	1,051
4	(σταθερά)	,385	,170	2,262	,024		
	βαθμός μαθηματικών	,669	,032	21,173	,000	,274	3,649
	βαθμός ελληνικών	,282	,035	8,088	,000	,268	3,736
	είδος σχολείου	,241	,088	2,740	,006	,950	1,053
	φύλο	-,120	,044	-2,708	,007	,971	1,030

Τελικά κρίνονται σημαντικές και επιλέγονται για το μοντέλο οι μεταβλητές X_1 (βαθμός στα ελληνικά) , X_2 (βαθμός στα μαθηματικά) , D_1 (φύλο) , D_4 (είδος σχολείου). Οι εκτιμήσεις των παραμέτρων α, β_1, β_2, β_3, β_6 είναι : a = 0,385 , b_1= 0,282 , b_2= 0,669 , b_3= 0,241 , b_6= -0,120 . Τα τυπικά σφάλματα των εκτιμώμενων παραμέτρων είναι 0,170 , 0,035 , 0,032 , 0,088 και 0,044 αντίστοιχα. Στην τέταρτη στήλη υπάρχουν οι τιμές των ελεγχοσυναρτήσεων t για κάθε παράμετρο, οι οποίες ελέγχουν εάν κάθε παράμετρος διαφέρει στατιστικά από το μηδέν. Από τον παραπάνω πίνακα φαίνεται ότι τα παρατηρούμενα επίπεδα σημαντικότητας, που αντιστοιχούν στο στατιστικό έλεγχο για τη στατιστική σημαντικότητα ή όχι των παραμέτρων, με μηδενική υπόθεση H_0 : η παράμετρος είναι μηδέν έναντι της

εναλλακτικής H_1 : η παράμετρος διαφέρει από το μηδέν, έχουν τιμή πολύ μικρότερη από την τιμή 0,05. Επομένως απορρίπτονται οι μηδενικές υποθέσεις σε επίπεδο σημαντικότητας 5% και γίνονται δεκτές οι εναλλακτικές. Τα 95% διαστήματα εμπιστοσύνης είναι : (0,05 , 0,719) για το α , (0,607 , 0,731) για το β_1 , (0,068 , 0,415) για το β_2 , (0,213 , 0,350) για το β_3 και (-0,208 , -0,033) για το β_6 . Στις δύο τελευταίες στήλες του πίνακα υπάρχουν οι τιμές ανοχής (tolerance = $\frac{1}{VIF}$) και οι τιμές VIF (Variance inflation factor) , δηλαδή αυξητικοί παράγοντες μεταβολής. Είναι VIF = $\frac{1}{1-R_K^2}$, όπου R_k^2 είναι ο συντελεστής προσδιορισμού του γραμμικού μοντέλου $X_k = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_{k-1} X_{k-1} + \beta_k X_k + \ldots \beta_\rho X_\rho$. Υψηλές τιμές VIF υποδηλώνουν πρόβλημα πολυσυγγραμμικότητας. Εδώ δε φαίνεται να υφίσταται τέτοιο πρόβλημα.

Ο παρακάτω πίνακας δίνει πληροφορίες για τους συντελεστές προσδιορισμού του γραμμικού μοντέλου.

Πίνακας 59
Συντελεστές προσδιορισμού

Μοντέλο	Συντελεστής συσχέτισης	Συντελεστής προσδιορισμού	Προσαρμοσμένος συντελεστής προσδιορισμού	Τυπικό σφάλμα της εκτίμησης
1	,935[a]	,874	,874	,4550
2	,947[b]	,897	,896	,4129
3	,948[c]	,899	,898	,4093
4	,949[d]	,901	,900	,4056

a. Μεταβλητές : (σταθερά), Μαθηματικά
b. Μεταβλητές : (σταθερά), Μαθηματικά, Ελληνικά
c. Μεταβλητές : (σταθερά), Μαθηματικά, Ελληνικά, Είδος σχολείου
d. Μεταβλητές: (σταθερά), Μαθηματικά, Ελληνικά, Είδος σχολείου, φύλο

Ο συντελεστής προσδιορισμού έχει τιμή 0,901. Το 90,1% του αθροίσματος των τετραγώνων των αποκλίσεων των τιμών του Y από το $\overline{Y}$ εξηγείται από τους παράγοντες που έχουν χρησιμοποιηθεί στο μοντέλο. Η τιμή του προσαρμοσμένου πολλαπλού συντελεστή είναι 0,9. Είναι καλύτερο μέτρο αφού λαμβάνει υπόψη του και τον αριθμό των μεταβλητών που έχουν

χρησιμοποιηθεί. Η βασιμότητα ή όχι της υπόθεσης της γραμμικής παλινδρόμησης μπορεί να ελεγχθεί και από τον πίνακα ανάλυσης διακύμανσης.

Πίνακας 60
Ανάλυση διακύμανσης

Πηγή μεταβλητότητας		Άθροισμα τετραγώνων	Βαθμοί ελευθερίας	Μέσο τετράγωνο	Τιμή F	Παρατηρούμενο επίπεδο σημαντικότητας
	Παλινδρόμιση	506,709	4	126,667	770,2	,000[a]
	Λάθη	55,594	338	,164		
	Σύνολο	562,304	342			

a. Ανεξάρτητες Μεταβλητές: (σταθερά), Μαθηματικά, Ελληνικά, Είδος σχολείου, Φύλο
Εξαρτημένη μεταβλητή : Απολυτήριο

Η συνολική μεταβλητότητα των δεδομένων χωρίζεται και αποδίδεται εν μέρει στο μοντέλο και εν μέρει στα λάθη. Η μεταβλητότητα της εξαρτημένης μεταβλητής μετράται με βάση το άθροισμα των παρακάτω αποκλίσεων :

- Μεταβλητότητα της εξαρτημένης εξαιτίας των ανεξάρτητων :

$$SSR = \sum\left(\hat{Y}_i - \overline{Y}\right)^2 = 506{,}709$$

- Μεταβλητότητα της εξαρτημένης εξαιτίας τυχαίων παραγόντων :

$$SSE = \sum\left(Y_i - \hat{Y}_i\right)^2 = 55{,}594$$

- Συνολική μεταβλητότητα της εξαρτημένης μεταβλητής :

$$SST = \sum\left(Y_i - \overline{Y}\right)^2 = 562{,}304$$

Ο πίνακας μας δίνει τους βαθμούς ελευθερίας για τη μεταβλητότητα εξαιτίας της παλινδρόμησης (αριθμός μεταβλητών (k)-1), εξαιτίας των λαθών (αριθμός παρατηρήσεων (n) - αριθμός μεταβλητών (k)) και για το σύνολο (n-1) . Για να εξετάσουμε αν κάθε μια από τις ανεξάρτητες μεταβλητές του μοντέλου συνεισφέρει πληροφορίες για την πρόβλεψη του Y πρέπει να ελέγξουμε τη μηδενική υπόθεση H_0 : $\beta_i = 0$ έναντι της εναλλακτικής H_1 : τουλάχιστον μία από τις παραμέτρους διαφέρει από το μηδέν. Η τιμή της ελεγχοσυνάρτησης F είναι 770,2 και είναι μεγαλύτερη από την τιμή της κατανομής F με 4 και 338 βαθμούς ελευθερίας (F=2,4). Επίσης το παρατηρούμενο επίπεδο σημαντικότητας έχει τιμή μηδέν. Απορρίπτεται η μηδενική υπόθεση σε επίπεδο σημαντικότητας 5% και γίνεται δεκτή η

εναλλακτική, δηλαδή τουλάχιστον μια από τις παραμέτρους διαφέρει σημαντικά από το μηδέν.

Προκειμένου να ελεγχθεί ότι τα αποτελέσματα της παλινδρόμησης είναι αξιόπιστα πρέπει να γίνει έλεγχος προϋποθέσεων σχετικά με τα σφάλματα εκτίμησης (κατάλοιπα). Τα σφάλματα πρέπει να κατανέμονται κανονικά να είναι ομοσκεδαστικά και τυχαία.

Έλεγχος κανονικότητας καταλοίπων

Ο έλεγχος της υπόθεσης της κανονικότητας των καταλοίπων θα γίνει με τη χρήση του ελέγχου Kolmogorov - Smirnov.

Οι προς έλεγχο υποθέσεις είναι :

H_0 : Τα κατάλοιπα κατανέμονται κανονικά

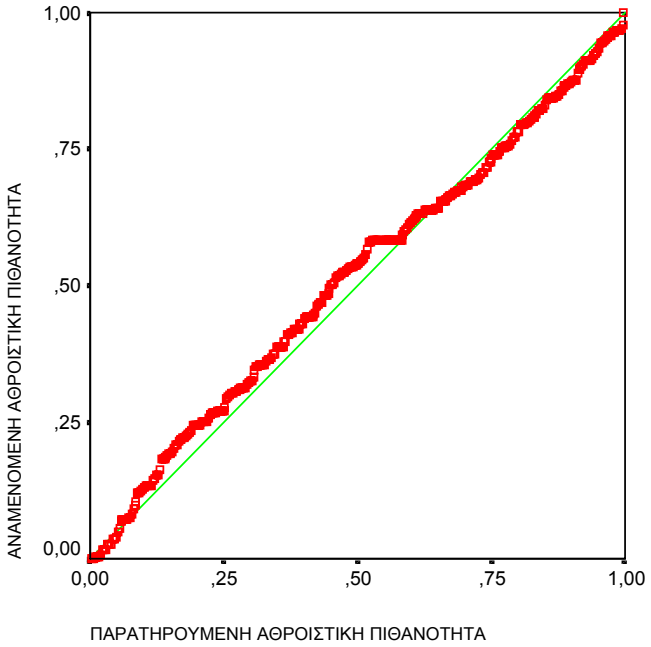
H_1 : Τα κατάλοιπα δεν κατανέμονται κανονικά

Πίνακας 61
Έλεγχος κανονικότητας καταλοίπων

	Kolmogorov-Smirnov		
	Statistic	Βαθμοί ελευθερίας	Παρατηρούμενο επίπεδο σημαντικότητας
Τυποποιημένα κατάλοιπα	,048	343	,059

Σε επίπεδο σημαντικότητας 5% δεν απορρίπτεται η μηδενική υπόθεση, δηλαδή τα κατάλοιπα κατανέμονται κανονικά.

Διάγραμμα 31
Έλεγχος κανονικότητας καταλοίπων



Από το παραπάνω διάγραμμα φαίνεται ότι τα κατάλοιπα είναι συγκεντρωμένα (σχεδόν) γύρω από μια ευθεία.

<u>Έλεγχος ομοσκεδαστικότητας και τυχαιότητας των καταλοίπων</u>

Όπως έχουμε ήδη αναφέρει, κάνοντας τους ελέγχους και στα προηγούμενα μοντέλα, ένας τρόπος ελέγχου της ισότητας των διασπορών είναι να χωριστούν τα κατάλοιπα σε δυο ομάδες, η μια από τις οποίες θα προέρχεται από τις "μικρές" τιμές του χ και η άλλη από τις "μεγάλες" τιμές του χ. Στη συνέχεια υπολογίζονται οι δειγματικές διασπορές που αντιστοιχούν στις δυο ομάδες των καταλοίπων. Κατόπιν χρησιμοποιείται ο έλεγχος F για την ισότητα δυο διασπορών, ώστε να ελεγχθεί αν τα δυο σύνολα των καταλοίπων έχουν ίσες διασπορές στους αντίστοιχους πληθυσμούς. Η ομοσκεδαστικότητα των καταλοίπων φαίνεται από τον παρακάτω πίνακα, κάνοντας τον έλεγχο :

H_0 : $S_1^2 = S_2^2$, οι δειγματικές διασπορές είναι ίσες

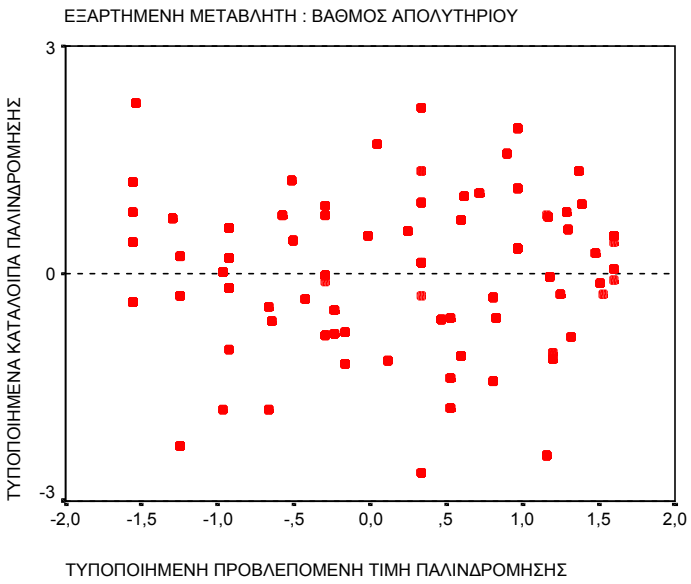
H_1 : $S_1^2 \neq S_2^2$, οι δειγματικές διασπορές είναι άνισες

Πίνακας 62
Έλεγχος ισότητας διασπορών των καταλοίπων

		Levene's Test για ισότητα διασπορών	
		F	Παρατηρούμενο επίπεδο σημαντικότητας
Τυποποιημένα κατάλοιπα	Προϋπόθεση ίσων διασπορών	,873	,351

Σε επίπεδο σημαντικότητας 5% δεν απορρίπτεται η μηδενική υπόθεση και γίνεται δεκτό ότι οι διασπορές των λαθών είναι ίσες. Από το παρακάτω διάγραμμα φαίνεται ότι τα κατάλοιπα είναι ομοιόμορφα κατανεμημένα και δεν παρουσιάζουν ανομοιογένεια των διασπορών. Επίσης η γραφική παράσταση των καταλοίπων δείχνει μια τυχαία τοποθέτησή τους γύρω από τη γραμμή που αντιστοιχεί σε κατάλοιπο "0".

Διάγραμμα 32
Γραφική παράσταση των καταλοίπων



Τέλος πρέπει ν' αναφέρουμε ότι έγινε έλεγχος ύπαρξης ακραίων παρατηρήσεων και παρατηρήσεων μεγάλης επιρροής.

Η εκτιμώμενη ευθεία παλινδρόμησης προσδιορίζεται ως :

$$\hat{Y} = 0,385 + 0,282\, X_1 + 0,669\, X_2 - 0,120\, D_1 + 0,241\, D_4$$

Όπου Y είναι η τελική επίδοση των μαθητών, δηλαδή ο βαθμός του απολυτηρίου, X_1 είναι ο βαθμός στην ελληνική γλώσσα , X_2 είναι ο βαθμός στα μαθηματικά, D_1 είναι η εικονική μεταβλητή για το φύλο και D_4 είναι η εικονική μεταβλητή για το είδος σχολείου φοίτησης.

Οι θετικοί και στατιστικά σημαντικοί συντελεστές των μεταβλητών X_1 και X_2 αποτελούν ένδειξη ότι ο βαθμός του απολυτηρίου αυξομειώνεται ανάλογα με τις αντίστοιχες αυξομειώσεις των βαθμών των μαθητών στα μαθηματικά και την ελληνική γλώσσα. Ο αρνητικός συντελεστής για την εικονική μεταβλητή D_1 αποτελεί ένδειξη ότι οι μαθητές χάνουν μονάδες από το βαθμό στο απολυτήριο, σε σχέση με τις μαθήτριες. Ο θετικός συντελεστής για την εικονική μεταβλητή D_4 αποτελεί ένδειξη ότι η φοίτηση των μαθητών σε ιδιωτικό σχολείο αποτελεί παράγοντα αύξησης της βαθμολογίας του απολυτηρίου σε σχέση με αυτούς που φοιτούν σε δημόσιο σχολείο.