Πανεπιστήμιο Δυτικής Μακεδονίας

Τμήμα Μηχανολόγων Μηχανικών

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΜΕΛΕΤΗ ΤΗΣ ΑΚΑΔΗΜΑΪΚΗΣ ΕΠΙΔΟΣΗΣ ΤΩΝ ΦΟΙΤΗΤΩΝ ΤΟΥ ΤΜΗΜΑΤΟΣ ΜΗΧΑΝΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ

ΚΟΝΤΙΖΑΣ ΣΤΥΛΙΑΝΟΣ (ΑΕΜ: 1650)

ΕΠΙΒΛΕΠΟΥΣΑ ΚΑΘΗΓΗΤΡΙΑ: ΠΑΝΑΓΙΩΤΙΔΟΥ ΣΟΦΙΑ

ΚΟΖΑΝΗ, ΝΟΕΜΒΡΙΟΣ 2021

Περίληψη

Στην παρούσα διπλωματική εργασία γίνεται μελέτη της ακαδημαϊκής επίδοσης των φοιτητών του Τμήματος Μηχανολόγων Μηχανικών του Πανεπιστημίου Δυτικής Μακεδονίας, έπειτα από χρήση δεδομένων προερχόμενων από 135 αποφοίτους. Σκοπός της εν λόγω διπλωματικής, αξιοποιώντας ποικίλα στατιστικά εργαλεία όπως έλεγχους υποθέσεων, έλεγχους προσαρμογής δεδομένων, ανάλυση μεταβλητότητας ANOVA και μοντέλων παλινδρόμησης, είναι να δημιουργηθούν κατάλληλα μοντέλα πρόβλεψης του αναμενόμενου βαθμού αποφοίτησης ή/και της αναμενόμενης διάρκειας σπουδών. Αξιολογώντας τις μεταβλητές οι οποίες επηρεάζουν περισσότερο τη συνολική επίδοση των φοιτητών, ερευνάται το πώς μπορούν να αξιοποιηθούν αυτές οι πληροφορίες για την πρόβλεψη της επίδοσής τους. Όσον αφορά τις μεταβλητές οι οποίες ενδέχεται να αξιοποιηθούν στο πλαίσιο της διπλωματικής, διερευνάται η συσχέτιση της διάρκειας σπουδών και του βαθμού αποφοίτησης, αρχικά μεταξύ τους, με το κάθε μάθημα, με το κατά πόσο υπήρξε καθυστέρηση στην επιτυχή ολοκλήρωση των μαθημάτων, καθώς και με τις βαθμολογίες των διπλωματικών εργασιών. Στη συνέχεια, αφότου γίνουν κατάλληλες διερευνήσεις σχέσεων εξάρτησης με τον βαθμό πτυχίου και τη διάρκεια σπουδών, γίνεται προσδιορισμός χρήσιμων ανεξάρτητων μεταβλητών για την πρόβλεψή τους. Με την αξιοποίηση και τον έλεγχο καταλληλότητας των ορισμένων ανεξάρτητων μεταβλητών επιτυγχάνεται η δημιουργία μοντέλων πρόβλεψης ανά έτος φοίτησης. Στο τέλος, αφότου γίνει εκτίμηση του σφάλματος μελλοντικών δεδομένων, γίνεται αξιολόγηση της επάρκειας των μοντέλων πρόβλεψης.

Λέξεις Κλειδιά: Επιβλεπόμενη μάθηση, Πολλαπλή παλινδρόμηση, Βηματική παλινδρόμηση, Διασταυρούμενη επικύρωση τμημάτων, Training and test error

Abstract

The current thesis studies the academic performance of the undergraduates of the department of Mechanical Engineering of the University of Western Macedonia by using data from 135 graduates. The purpose of this thesis is to build appropriate predictive models of the expected degree grades and/or expected dates of graduation by utilizing various statistical tools like hypothesis and goodness-of-fit testing, analysis of variance ANOVA and regression analysis. First off, by evaluating the most influential variables for the overall academic performance, we examine how we can make use of this information so we can properly predict the undergraduates performance. As for the independent variables that we may use for the purposes of this thesis, we examine the correlations of the degree grades and dates of graduation with each other, with each course, with the delay of the successful competion of courses and with the thesis grades. Consequently, by properly examining the statistical relationships of dependence with the degree grades and dates of graduation we can specify suitable independent variables for their prediction. With the use of the aforementioned independent variables and after we select the proper subsets of them, we can build predictive models by year of studies. Finally, once we estimate the error of future sets of data, we can determine the adequacy of our predictive models.

Keywords: Supervised learning, Multiple regression, Stepwise regression, k-fold Cross-Validation, Training and test error

Ευχαριστίες

Θα ήθελα να ευχαριστήσω την επιβλέπουσα καθηγήτρια της διπλωματικής μου εργασίας, Παναγιωτίδου Σοφία, για τη συνεχή καθοδήγηση και υποστήριξη καθ' όλη τη διάρκεια της εργασίας αλλά και των σπουδών μου. Επίσης, θα ήθελα να την ευχαριστήσω θερμά για την ανάθεση του θέματος της εργασίας, καθώς μου δόθηκε η δυνατότητα να διευρύνω τις γνώσεις μου σε ένα διαρκώς αναπτυσσόμενο επιστημονικό πεδίο.

Επιπλέον, θα ήθελα να ευχαριστήσω όλους όσοι ήταν δίπλα μου και προπάντων την οικογένειά μου, η οποία με ενθάρρυνε και μου έδωσε τη δύναμη και την ωριμότητα να γίνω αυτό που επιθυμώ.

Περιεχόμενα

Κεφάλαιο 1: Εισαγωγή

Στον επιχειρηματικό κόσμο η ικανότητα χρήσης δεδομένων για την ορθή πρόβλεψη μελλοντικών τιμών έχει γίνει πλέον αναπόσπαστο κομμάτι του. Οι προβλέψεις παίρνουν συνήθως τη μορφή της πρόβλεψης του ενδεχόμενου κέρδους ή οικονομικής ζημίας με απώτερο σκοπό τη δημιουργία κατάλληλων στρατηγικών και λήψης αποφάσεων. Η διαδικασία η οποία ακολουθείται είναι αρχικά η συγκέντρωση χρήσιμων δεδομένων και στη συνέχεια η ανάλυση και η αξιοποίησή τους για τις επιθυμητές προβλέψεις.

Η χρησιμότητα των προβλέψεων όμως δεν περιορίζεται μόνο στις επιχειρήσεις. Για την καλύτερη παρακολούθηση και αξιολόγηση της πορείας των φοιτητών του τμήματος των Μηχανολόγων Μηχανικών του Πανεπιστημίου Δυτικής Μακεδονίας, επιλέχθηκε η ανάλυση δεδομένων 135 αποφοίτων και η χρήση αυτών για τη δημιουργία μοντέλων πρόβλεψης του βαθμού αποφοίτησης και της διάρκειας σπουδών.

Για την καλύτερη ερμηνεία των αποτελεσμάτων της παρούσας διπλωματικής εργασίας, αρχικά επιλέχθηκε η ανάλυση των μαθημάτων του τμήματος στο κεφάλαιο 2. Η εν λόγω ανάλυση επιτεύχθηκε με τη σύγκριση της διακύμανσης των μαθημάτων και στη συνέχεια με τη σύγκριση των μέσων όρων τους.

Στο κεφάλαιο 3 πραγματοποιήθηκε διερεύνηση συσχετίσεων με τη χρήση των συντελεστών Pearson και Spearman. Αναλυτικότερα, οι συσχετίσεις οι οποίες διερευνήθηκαν είναι αυτές των μαθημάτων με τον βαθμό αποφοίτησης και αυτές των μαθημάτων, του βαθμού αποφοίτησης και της διάρκειας σπουδών με το κατά πόσο υπήρξε καθυστέρηση στην επιτυχή εξέταση των μαθημάτων. Ταυτόχρονα διερευνήθηκε το κατά πόσο συσχετίζεται ο βαθμός αποφοίτησης και η διάρκεια σπουδών μεταξύ τους και με τους βαθμούς των διπλωματικών εργασιών.

Εν συνεχεία, στο κεφάλαιο 4 αρχικά έγινε εισαγωγή στην έννοια της παλινδρόμησης και έγινε ανάλυση της μεθοδολογίας της. Με την προσαρμογή μοντέλων παλινδρόμησης, επιτεύχθηκε η ανάλυση σχέσεων εξάρτησης ανεξάρτητων μεταβλητών προερχόμενων από τη συνολική διάρκεια σπουδών των αποφοίτων, αλλά και μεταβλητών της υποχρεωτικής φοίτησης των οποίων εξετάζεται η χρησιμότητα για τη δημιουργία ικανοποιητικών μοντέλων πρόβλεψης.

Στο 5° κεφάλαιο ορίστηκαν οι εξεταζόμενες ανεξάρτητες μεταβλητές για την πρόβλεψη του βαθμού αποφοίτησης και της διάρκειας σπουδών ανά τα έτη για τα πρώτα 5 ακαδημαϊκά έτη φοίτησης. Για κάθε σετ δεδομένων επιλέχθηκε υποσύνολο μεταβλητών εισόδου με τη βοήθεια της μεθόδου της προς τα πίσω απαλοιφής. Έπειτα, πραγματοποιήθηκε επαναλαμβανόμενη διασταυρούμενη επικύρωση τμημάτων με σκοπό την εκτίμηση του σφάλματος μελλοντικών δεδομένων εκτός των υπαρχόντων και έγινε αξιολόγηση των μοντέλων πρόβλεψης.

Τελικό κομμάτι της εργασίας αποτελεί η διερεύνηση αποτελεσματικότερων μοντέλων πρόβλεψης με την προσθήκη αλληλεπιδράσεων των ανεξάρτητων μεταβλητών στα μοντέλα. Τα αποτελέσματα της παραπάνω διερεύνησης δίνονται στο κεφάλαιο 6, ακολουθώντας την ίδια διαδικασία με αυτή του κεφαλαίου 5.

Τέλος, το κεφάλαιο 7 ανακεφαλαιώνει τα αποτελέσματα της εργασίας καταλήγοντας σε συμπεράσματα, ενώ ταυτόχρονα γίνονται προτάσεις για μελλοντική έρευνα.

Κεφάλαιο 2: Ανάλυση ισότητας μεταβλητότητας και μέσου όρου μαθημάτων

Πριν προχωρήσουμε σε περαιτέρω αναλύσεις, στο κεφάλαιο αυτό θα γίνει σύγκριση των μαθημάτων του τμήματος με τη χρήση δεδομένων 135 αποφοίτων, διευρύνοντας έτσι την ικανότητα της αξιολόγησης αποτελεσμάτων στο σύνολο της εργασίας. Η εν λόγω σύγκριση θα γίνει με τον έλεγχο του κατά πόσο τα μαθήματα παρουσιάζουν σταθερή μεταβλητότητα συγκρίνοντας τις διακυμάνσεις των μαθημάτων. Εφόσον διαπιστωθεί αν οι μεταβλητότητες είναι ίσες ή διαφορετικές για το σύνολο των μαθημάτων, μπορούμε να προχωρήσουμε στη σύγκριση των μέσων όρων τους. Για τη σύγκριση των διακυμάνσεων επιλέχθηκε το τεστ του Levene, ενώ για τη σύγκριση των μέσων όρων η μέθοδος του one-sample τεστ Z.

2.1 Τεστ ισότητας μεταβλητότητας βαθμών μαθημάτων

Για τον έλεγχο της ισότητας των διακυμάνσεων μιας μεταβλητής διαφορετικών παραγόντων θα χρησιμοποιηθεί το τεστ του Levene. Μηδενική υπόθεση αποτελεί πως οι μεταβλητότητες των διαφορετικών ομάδων δειγμάτων πληθυσμού κ είναι ίσες $H_0 : \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_\kappa^2$ και εναλλακτική υπόθεση αποτελεί η περίπτωση τουλάχιστον μία μεταβλητότητα να διαφέρει σημαντικά από τις άλλες. Για επίπεδο σημαντικότητας 0.05, αν το p-value του τεστ προκύψει μικρότερο του 0.05, οι διαφορές των διακυμάνσεων είναι σχεδόν απίθανο να προέκυψαν από δειγματοληψία πληθυσμού με ίσες διακυμάνσεις. Επομένως οδηγούμαστε στην απόρριψη της μηδενικής υπόθεσης και στο συμπέρασμα της διαφορετικότητας των διακυμάνσεων. Καθώς στο υποκεφάλαιο 2.2 θα ακολουθήσει σύγκριση των μέσων τιμών των βαθμών των μαθημάτων του τμήματος, χρήσιμη θα ήταν και η σύγκριση των διακυμάνσεών τους για την περίπτωση κοινής πληθυσμιακής διακύμανσης.

Για τον έλεγχο της ύπαρξης ομοιογένειας διακυμάνσεων το τεστ Levene λειτουργεί όπως και η ANOVA (Analysis of Variance), η οποία θα αναλυθεί στο κεφάλαιο 4, με τη χρήση του όρου W που είναι ισοδύναμος του F της ANOVA. Για τον προσδιορισμό του W πρέπει πρώτα να γίνει μετασχηματισμός των τιμών x_{ij} για κάθε ομάδα με i=1,2,…,k και k αριθμό διαφορετικών ομάδων και με παρατηρήσεις της κάθε ομάδας j=1,2,…,N_i, ως εξής:

$Z_{ij} = \left| x_{ij} - \overline{x}_i \right|$ (2.1) με :

$\overline{x}_i$: μέσος όρος παρατηρήσεων ομάδας i.

Έτσι λοιπόν για:

N: Συνολικός αριθμός παρατηρήσεων των ομάδων,

N_i: Αριθμός παρατηρήσεων ομάδας i,

k: Αριθμός διαφορετικών ομάδων δειγμάτων,

$Z_i. = \dfrac{\sum_{j=1}^{N_i} Z_{ij}}{N_i}$ (2.2) : Μέσος όρος των Z_{ij} για την ομάδα i,

$$Z.. = \frac{\sum_{i=1}^{\kappa}\sum_{j=1}^{N_i} Z_{ij}}{N}$$ (2.3) : Μέσος όρος των Z_{ij} όλων των ομάδων,

ο όρος W υπολογίζεται από τον παρακάτω τύπο [1]:

$$W = \frac{(N-k)\sum_{i=1}^{k} N_i (Z_i. - Z..)^2}{(k-1)\sum_{i=1}^{k}\sum_{j=1}^{Ni} (Z_{ij} - Z_i.)^2}$$ (2.1.4)

Ο όρος W ακολουθεί κατανομή F, για επίπεδο σημαντικότητας α, με k-1 και N-k βαθμούς ελευθερίας.

Σύγκριση διακυμάνσεων των βαθμών των μαθημάτων

Χρησιμοποιώντας τη μέθοδο του Lavene τεστ, θα προχωρήσουμε σε τεστ ισότητας μεταβλητότητας των μαθημάτων. Τα μαθήματα για τα οποία έγινε τεστ ισότητας μεταβλητότητας είναι όσα μαθήματα είχαν 15 και πάνω επιτυχείς εξετάσεις και τα αποτελέσματά τους δίνονται στον πίνακα 2.1. Για τα μαθήματα επίσης δίνεται ο αριθμός επιτυχών εξετάσεων, η τυπική τους απόκλιση και το 95% διάστημα εμπιστοσύνης της τυπικής απόκλισής τους. Το p-value του τεστ που προέκυψε είναι μηδενικό, υποδεικνύοντας ισχυρή ένδειξη διαφορετικότητας των διακυμάνσεων.

Πίνακας 2.1: Lavene τεστ ισότητας μεταβλητότητας μαθημάτων

95% Bonferroni Confidence Intervals for Standard Deviations

Sample	N	StDev	CI
101	133	1.44251	(1.21843, 1.75288)
103	133	1.66799	(1.44560, 1.97539)
104	134	1.30820	(1.00192, 1.75284)
105	134	1.49692	(1.22860, 1.87163)
113	135	1.25929	(1.00791, 1.61428)
144	125	1.40506	(1.15619, 1.75554)
141	135	1.32267	(1.14024, 1.57417)
102	134	1.48950	(1.22060, 1.86525)
109	135	1.32373	(1.08301, 1.66002)
111	134	1.34705	(1.05989, 1.75687)
146	121	1.28901	(1.02291, 1.67157)
134	63	1.62265	(1.32018, 2.10893)
142	135	1.60618	(1.41247, 1.87394)
145	62	1.38063	(0.95942, 2.10277)
149	72	1.38648	(1.14822, 1.75769)
110	134	1.08564	(0.76022, 1.59096)
107	134	1.39723	(1.09859, 1.82359)
119	134	1.29619	(1.01817, 1.69334)
132	134	1.42548	(1.22272, 1.70539)
135	134	1.64256	(1.45464, 1.90335)
148	27	1.86224	(1.30900, 3.03363)
112	135	0.87645	(0.54114, 1.45643)
108	135	1.13865	(0.81474, 1.63270)
120	134	1.44122	(1.17875, 1.80829)
137	134	1.36912	(1.10983, 1.73322)
114	135	1.14689	(0.86853, 1.55385)
106	134	1.65777	(1.42017, 1.98580)
118	135	1.24214	(0.96112, 1.64706)
140	134	1.34566	(1.07543, 1.72790)
147	134	1.74649	(1.50325, 2.08225)
116	135	1.27067	(0.99387, 1.66680)
138	134	1.37711	(1.10426, 1.76237)
133	135	1.18490	(0.89922, 1.60193)
117	135	1.51770	(1.31132, 1.80224)
123	135	1.64761	(1.43979, 1.93445)
127	135	1.21429	(0.96573, 1.56650)
131	133	1.28789	(1.04419, 1.63040)
250	83	1.26389	(0.90619, 1.83855)
204	135	1.14915	(0.90406, 1.49864)
207	135	1.59027	(1.34786, 1.92506)
219	135	1.49564	(1.28113, 1.79145)
254	87	1.43727	(1.08435, 1.98304)
228	49	1.42723	(1.08478, 2.01871)
206	118	1.20367	(0.91320, 1.63390)
230	54	1.65902	(1.24773, 2.35507)
372	64	1.43985	(1.15315, 1.89935)
377	21	1.27615	(0.66331, 2.93295)
205	135	1.13130	(0.88040, 1.49150)
255	16	1.42850	(0.92479, 2.80659)
224	17	0.56230	(0.37983, 1.04208)
251	119	1.75064	(1.53401, 2.05698)
241	134	1.54724	(1.36086, 1.80524)
210	72	1.41829	(1.11371, 1.89624)
252	87	1.42245	(1.17933, 1.78589)
256	28	0.85758	(0.42880, 1.95380)
240	89	0.89841	(0.64487, 1.30165)
246	27	1.90385	(1.41837, 2.92622)
249	131	1.23674	(0.96946, 1.62000)
253	18	1.75943	(1.22107, 3.12992)
327	65	1.08530	(0.83959, 1.48086)
391	66	1.40802	(1.02605, 2.03781)
318	52	1.37676	(0.98960, 2.05023)
350	51	1.25729	(0.86894, 1.94998)
380	62	1.18212	(0.89670, 1.64939)
371	32	1.26841	(0.96923, 1.85862)
356	99	0.69011	(0.48729, 1.01233)
389	88	1.03889	(0.76075, 1.47608)
376	70	1.03049	(0.74227, 1.50412)
387	25	1.42001	(0.80750, 2.89296)
382	33	0.87039	(0.26773, 3.15690)
379	17	1.75681	(1.22516, 3.15372)
349	95	1.25667	(0.96855, 1.69140)
367	17	0.19648	(0.05979, 0.80824)
316	48	1.52197	(1.17662, 2.11973)
383	27	1.86186	(1.27225, 3.11998)
390	35	1.29121	(0.91359, 2.02257)
392	20	1.72416	(1.23894, 2.89444)
352	45	1.23991	(0.96710, 1.72045)
381	29	0.93903	(0.33950, 2.94455)
386	21	1.25096	(0.79148, 2.36189)

Individual confidence level = 99.9375%

Method

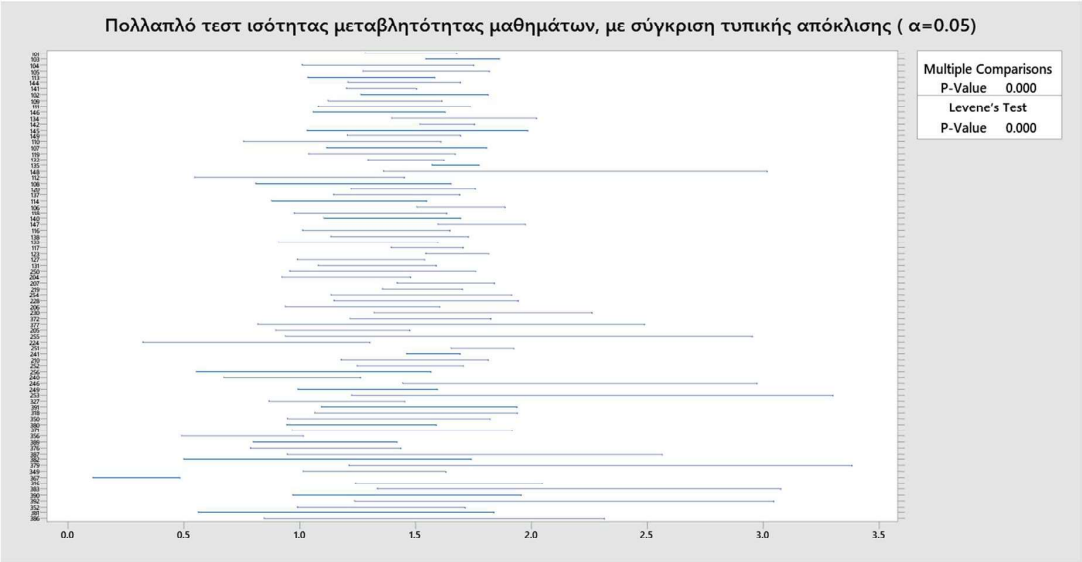
Null hypothesis	All variances are equal
Alternative hypothesis	At least one variance is different
Significance level	α = 0.05

Tests

Method	Test Statistic	P-Value
Multiple comparisons	—	0.000
Levene	6.73	0.000

Στο σχήμα 2.1 φαίνονται τα διαστήματα εμπιστοσύνης των τυπικών αποκλίσεων των μαθημάτων, με τις τυπικές αποκλίσεις των μαθημάτων να διαφέρουν και ως προς το μέγεθος αλλά και ως προς το εύρος των διαστημάτων εμπιστοσύνης τους.

Σχήμα 2.1: Σύγκριση μεταβλητότητας μαθημάτων



2.2 Τεστ ισότητας μέσων τιμών βαθμών μαθημάτων

Σε αυτό το υποκεφάλαιο θα γίνει έλεγχος ισότητας των μέσων τιμών των βαθμών των μαθημάτων, με τη συνολική μέση τιμή όλων των επιτυχών εξετάσεων αυτών και για γνωστή αλλά διαφορετική διακύμανση. Καθώς η πλειοψηφία των μαθημάτων είχε περισσότερες από 30 επιτυχείς εξετάσεις θα προτιμηθεί το τεστ one-sample Z, υποθέτοντας κανονική κατανομή των δειγμάτων και του πληθυσμού ως το σύνολο των δειγμάτων. Για τον πληθυσμό ο μέσος όρος είναι: $\mu_\pi = 6.7624$ και η τυπική απόκλιση είναι : $\sigma_\pi = 1.61278$.

Εφόσον αναλύθηκε η διαφορά των διακυμάνσεων ανά μάθημα στο υποκεφάλαιο 2.1, θα ορισθεί η εκτίμηση της τυπικής απόκλισης για τη διαφορά των μέσων όρων του κάθε μαθήματος με τον πληθυσμό. Έτσι λοιπόν για κάθε μάθημα i, η εκτίμηση της τυπικής απόκλισης της διαφοράς των μέσων όρων είναι:

$$s_i = \sqrt{\frac{\sigma_i^2}{N_i} + \frac{\sigma_\pi^2}{N_\pi}}$$ (2.4) με:

N_i: αριθμός επιτυχών εξετάσεων για κάθε μάθημα i,

N_π: συνολικός αριθμός επιτυχών εξετάσεων πληθυσμού μαθημάτων,

σ_i: τυπική απόκλιση για κάθε μάθημα i,

σ_π: τυπική απόκλιση πληθυσμού.

Η τιμή της στατιστικής δείγματος Z [2] προκύπτει από το πηλίκο της διαφοράς των μέσων όρων των μαθημάτων i με τον μέσο όρο του πληθυσμού π, με την εκτίμηση της τυπικής απόκλισης όπως ορίστηκε στον τύπο 2.5:

$Z_i = \frac{\overline{x}_i - \overline{x}_\pi}{s_i}$ (2.5) με:

$\overline{x}_i$: μέσος όρος επιτυχών εξετάσεων μαθημάτων i,

$\overline{x}_\pi$: μέσος όρος επιτυχών εξετάσεων πληθυσμού.

Στο one-sample τεστ Z τη μηδενική υπόθεση αποτελεί οι μέσοι όροι να είναι ίσοι για επίπεδο σημαντικότητας α και στην περίπτωσή μας το επίπεδο σημαντικότητας ορίστηκε 0.05. Αν το p-value προκύψει μικρότερο του 0.05, συνάγεται το συμπέρασμα πως οι μέσοι όροι διαφέρουν σημαντικά.

Σύγκριση μέσων όρων μαθημάτων

Παρακάτω στον πίνακα 2.3 δίνεται ο μέσος όρος, η τυπική απόκλιση και το τυπικό σφάλμα για κάθε μάθημα, με επιτυχείς εξετάσεις περισσότερες από 15, καθώς και το 95% διάστημα εμπιστοσύνης των μέσων τιμών τους. Στον πίνακα 2.3 επίσης πραγματοποιήθηκε το one-sample τεστ Z για την ισότητα των μέσων τιμών των μαθημάτων με τη μέση τιμή του πληθυσμού. Το αποτέλεσμα της σύγκρισης αυτής είναι 23 από τους 80 εξεταζόμενους μέσους όρους μαθημάτων να μην διαφέρουν από τον μέσο όρο του πληθυσμού, ενώ οι υπόλοιποι 57 να διαφέρουν.

Πίνακας 2.3: Στατιστική απεικόνιση μαθημάτων και One sample test Z ισότητας των μέσων όρων τους

Descriptive Statistics

Sample	N	Mean	StDev	SE Mean	95% CI for μ
101	133	6.820	1.443	0.140	(6.545, 7.094)
103	133	6.996	1.668	0.140	(6.722, 7.270)
104	134	6.022	1.308	0.139	(5.749, 6.295)
105	134	6.459	1.497	0.139	(6.186, 6.732)
113	135	6.167	1.259	0.139	(5.895, 6.439)
144	125	6.440	1.405	0.144	(6.157, 6.723)
141	135	7.241	1.323	0.139	(6.969, 7.513)
102	134	6.571	1.490	0.139	(6.298, 6.844)
109	135	6.319	1.324	0.139	(6.046, 6.591)
111	134	6.187	1.347	0.139	(5.913, 6.460)
146	121	6.132	1.289	0.147	(5.845, 6.420)
134	63	7.008	1.623	0.203	(6.610, 7.406)
142	135	7.380	1.606	0.139	(7.108, 7.652)
145	62	6.290	1.381	0.205	(5.889, 6.692)
149	72	7.236	1.386	0.190	(6.864, 7.609)
110	134	5.720	1.086	0.139	(5.447, 5.993)
107	134	6.209	1.397	0.139	(5.936, 6.482)
119	134	6.048	1.296	0.139	(5.775, 6.321)
132	134	6.925	1.425	0.139	(6.652, 7.198)
135	134	7.313	1.643	0.139	(7.040, 7.587)
148	27	6.944	1.862	0.310	(6.336, 7.553)
112	135	5.478	0.876	0.139	(5.206, 5.750)
108	135	5.878	1.139	0.139	(5.606, 6.150)
120	134	6.220	1.441	0.139	(5.947, 6.493)
137	134	6.642	1.369	0.139	(6.369, 6.915)
114	135	5.926	1.147	0.139	(5.654, 6.198)
106	134	6.809	1.658	0.139	(6.536, 7.082)
118	135	6.199	1.242	0.139	(5.927, 6.471)
140	134	6.313	1.346	0.139	(6.040, 6.587)
147	134	6.892	1.746	0.139	(6.619, 7.165)
116	135	6.218	1.271	0.139	(5.946, 6.490)
138	134	6.205	1.377	0.139	(5.932, 6.478)
133	135	6.054	1.185	0.139	(5.782, 6.326)
117	135	6.826	1.518	0.139	(6.554, 7.098)
123	135	7.074	1.648	0.139	(6.802, 7.346)
127	135	6.170	1.214	0.139	(5.898, 6.442)
131	133	6.331	1.288	0.140	(6.057, 6.605)
250	83	6.012	1.264	0.177	(5.665, 6.359)
204	135	5.836	1.149	0.139	(5.564, 6.108)
207	135	6.770	1.590	0.139	(6.498, 7.042)
219	135	6.904	1.496	0.139	(6.632, 7.176)
254	87	6.155	1.437	0.173	(5.816, 6.494)
228	49	6.673	1.427	0.230	(6.222, 7.125)
206	118	5.915	1.204	0.148	(5.624, 6.206)
230	54	6.750	1.659	0.219	(6.320, 7.180)
372	64	7.828	1.440	0.202	(7.433, 8.223)
377	21	8.643	1.276	0.352	(7.953, 9.333)
205	135	6.000	1.131	0.139	(5.728, 6.272)
255	16	7.406	1.429	0.403	(6.616, 8.196)
224	17	8.765	0.562	0.391	(7.998, 9.531)
251	119	7.832	1.751	0.148	(7.542, 8.122)
241	134	7.690	1.547	0.139	(7.417, 7.963)
210	72	6.403	1.418	0.190	(6.030, 6.775)
252	87	7.611	1.422	0.173	(7.273, 7.950)
256	28	7.929	0.858	0.305	(7.331, 8.526)
240	89	8.270	0.898	0.171	(7.935, 8.605)
246	27	7.019	1.904	0.310	(6.410, 7.627)
249	131	6.141	1.237	0.141	(5.865, 6.417)
253	18	8.417	1.759	0.380	(7.672, 9.162)
327	65	8.615	1.085	0.200	(8.223, 9.007)
391	66	8.220	1.408	0.199	(7.831, 8.609)
318	52	7.106	1.377	0.224	(6.667, 7.544)
350	51	6.137	1.257	0.226	(5.695, 6.580)
380	62	7.935	1.182	0.205	(7.534, 8.337)
371	32	7.188	1.268	0.285	(6.629, 7.746)
356	99	9.431	0.690	0.162	(9.114, 9.749)
389	88	8.966	1.039	0.172	(8.629, 9.303)
376	70	8.557	1.030	0.193	(8.179, 8.935)
387	25	8.568	1.420	0.323	(7.936, 9.200)
382	33	9.652	0.870	0.281	(9.101, 10.202)
379	17	7.647	1.757	0.391	(6.880, 8.414)
349	95	6.395	1.257	0.165	(6.070, 6.719)
367	17	9.912	0.196	0.391	(9.145, 10.678)
316	48	7.198	1.522	0.233	(6.742, 7.654)
383	27	8.204	1.862	0.310	(7.595, 8.812)
390	35	8.543	1.291	0.273	(8.009, 9.077)
392	20	7.330	1.724	0.361	(6.623, 8.037)
352	45	7.189	1.240	0.240	(6.718, 7.660)
381	29	9.603	0.939	0.299	(9.016, 10.190)
386	21	7.310	1.251	0.352	(6.620, 7.999)

μ: mean of 101, 103, 104, 105, 113, 144, 141, 102, 109, 111, 146, 134, 142, 145, 149, 110, 107, 119, 132, 135, 148, 112, 108, 120, 137, 114, 106, 118, 140, 147, 116, 138, 133, 117, 123, 127, 131, 250, 204, 207, 219, 254, 228, 206, 230, 372, 377, 205, 255, 224, 251, 241, 210, 252, 256, 240, 246, 249, 253, 327, 391, 318, 350, 380, 371, 356, 389, 376, 387, 382, 379, 349, 367, 316, 383, 390, 392, 352, 381, 386

Known standard deviation = 1.61278

Test

Null hypothesis H_0: μ = 6.76240

Alternative hypothesis H_1: μ ≠ 6.76240

Sample	Z-Value	P-Value
101	0.41	0.683
103	1.67	0.094
104	-5.31	0.000
105	-2.18	0.029
113	-4.29	0.000
144	-2.23	0.025
141	3.45	0.001
102	-1.37	0.169
109	-3.20	0.001
111	-4.13	0.000
146	-4.30	0.000
134	1.21	0.227
142	4.45	0.000
145	-2.30	0.021
149	2.49	0.013
110	-7.48	0.000
107	-3.97	0.000
119	-5.13	0.000
132	1.17	0.242
135	3.96	0.000
148	0.59	0.558
112	-9.25	0.000
108	-6.37	0.000
120	-3.89	0.000
137	-0.87	0.387
114	-6.03	0.000
106	0.33	0.738
118	-4.06	0.000
140	-3.22	0.001
147	0.93	0.353
116	-3.92	0.000
138	-4.00	0.000
133	-5.10	0.000
117	0.46	0.647
123	2.25	0.025
127	-4.27	0.000
131	-3.09	0.002
250	-4.24	0.000
204	-6.67	0.000
207	0.06	0.954
219	1.02	0.309
254	-3.51	0.000
228	-0.39	0.700
206	-5.71	0.000
230	-0.06	0.955
372	5.29	0.000
377	5.34	0.000
205	5.49	0.000
255	1.60	0.110
224	5.12	0.000
251	7.23	0.000
241	6.66	0.000
210	-1.89	0.058
252	4.91	0.000
256	3.83	0.000
240	8.82	0.000
246	0.83	0.409
249	-4.41	0.000
253	4.35	0.000
327	9.26	0.000
391	7.34	0.000
318	1.54	0.125
350	-2.77	0.006
380	5.73	0.000
371	1.49	0.136
356	16.47	0.000
389	12.82	0.000
376	9.31	0.000
387	5.60	0.000
382	10.29	0.000
379	2.26	0.024
349	-2.22	0.026
367	8.05	0.000
316	1.87	0.061
383	4.64	0.000
390	6.53	0.000
392	1.57	0.116
352	1.77	0.076
381	9.49	0.000
386	1.55	0.120

Κεφάλαιο 3: Διερεύνηση συσχετίσεων

3.1 Συντελεστής συσχέτισης Pearson

Για την εύρεση της γραμμικής συσχέτισης δύο τυχαίων μεταβλητών X και Y, προτιμάται ο συντελεστής συσχέτισης ρ τον οποίο ανέπτυξε ο Karl Pearson. Ο προαναφερόμενος συντελεστής υπολογίζεται ως το πηλίκο της συνδιακύμανσης (με τη συντομογραφία cov: μέτρο βαθμού συσχετίσεως) των μεταβλητών X και Y με το γινόμενο των τυπικών τους αποκλίσεων. Η συνδιακύμανση ορίζεται ως η αναμενόμενη τιμή του γινομένου των διαφορών των μεταβλητών X και Y με τις αναμενόμενες μέσες τους τιμές μ_X και μ_Y αντίστοιχα. Σύμφωνα με τα παραπάνω, ο συντελεστής συσχέτισης Pearson μπορεί να γραφεί ως εξής:

$$\rho = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{\mathrm{E}[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y} \quad (3.1)$$

Ο συντελεστής συσχέτισης Pearson μας δείχνει πόσο ισχυρή είναι η γραμμική σχέση των μεταβλητών και παίρνει τιμές από -1 (για μία τελείως αντίστροφη γραμμική σχέση) μέχρι +1 (για μία τελείως αύξουσα γραμμική σχέση). Στην περίπτωση ανεξαρτησίας των μεταβλητών ο συντελεστής συσχέτισης παίρνει την τιμή 0.

Για ένα δείγμα από n μετρήσεις των X και για κάθε μέτρησή του να αντιστοιχεί μία τιμή Y, ο συντελεστής μετασχηματίζεται για i=1,2, …,n ως εξής:

$$\rho = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} \quad (3.2)$$

Έτσι σύμφωνα με τον παραπάνω τύπο, αν και μόνο αν οι διαφορές των τιμών των μεταβλητών x_i και y_i με τις αντίστοιχες μέσες του τιμές έχουν το ίδιο πρόσημο ταυτοχρόνως, το γινόμενό τους προκύπτει θετικό. Στην αντίθετη περίπτωση το γινόμενό τους προκύπτει αρνητικό και όσο πιο ισχυρές είναι οι τάσεις να έχουν ίδιο ή αντίθετο πρόσημο, τόσο μεγαλύτερη η απόλυτη τιμή του ρ.

Όπως αναφέρθηκε, ο συντελεστής συσχέτισης Pearson προτιμάται για την εύρεση γραμμικών συσχετίσεων μεταβλητών, δηλαδή όταν οι μεταβλητές κινούνται προς την ίδια ή αντίθετη κατεύθυνση με σταθερό ρυθμό. Αν κάτι τέτοιο δεν ισχύει τότε η σχέση τους θεωρείται μονοτονική και ο συντελεστής συσχέτισης Pearson δεν θεωρείται επαρκής.

3.2 Συντελεστής συσχέτισης Spearman

Ένας ακόμα συντελεστής συσχέτισης είναι ο συντελεστής Spearman και πήρε το όνομά του από τον Charles Spearman. Σε αντίθεση με τον Pearson, ο Spearman αξιολογεί το κατά πόσο καλά μπορεί να περιγραφεί η σχέση δύο μεταβλητών με μία μονότονη συνάρτηση. Όπως και στον συντελεστή Pearson, το πρόσημο της συσχέτισης δείχνει το κατά πόσο η σχέση της εξαρτημένης με την ανεξάρτητη μεταβλητή είναι ανάλογη (θετικό) ή αντιστρόφως ανάλογη (αρνητικό). Στην περίπτωση μίας τέλειας μονοτονικής σχέσης, ο συντελεστής Spearman παίρνει σε απόλυτη τιμή την τιμή 1, ενώ

στην περίπτωση ανυπαρξίας τάσης αύξησης ή μείωσης της Y, έπειτα από αύξηση της X, παίρνει την τιμή 0.

Για την εύρεση του συντελεστή Spearman, απαραίτητη προϋπόθεση αποτελεί η ταξινόμηση των τιμών των μεταβλητών X και Y σε αύξουσα σειρά ξεχωριστά. Αν παρουσιαστούν ισοψηφίες, τότε οι τιμές αυτές αρχικά κατατάσσονται σε αύξουσα σειρά και μετά ορίζονται από κοινού ως οι μέσοι όροι των θέσεων που θα είχαν. Έτσι ο συντελεστής Spearman, για ένα δείγμα από n μετρήσεις, έχει τον ίδιο τύπο υπολογισμού με τον συντελεστή Pearson όπως περιγράφηκε στην εξίσωση 3.2 με τη διαφορά πως για κάθε x_i και yi χρησιμοποιείται η τιμή της σειράς κατάταξής τους.

Ο συντελεστής Spearman χρησιμοποιείται όταν οι εξεταζόμενες μεταβλητές έχουν μονοτονική σχέση, δηλαδή όταν οι μεταβλητές κινούνται προς την ίδια ή αντίθετη κατεύθυνση με μη σταθερό ρυθμό. Επιπροσθέτως όταν η σχέση των μεταβλητών δεν είναι ξεκάθαρη είναι πάλι προτιμότερο να χρησιμοποιηθεί ο εν λόγω συντελεστής έναντι του Pearson. Τέλος, μία ακόμη διαφορά του με τον συντελεστή Pearson αποτελεί η ευαισθησία του σε ακραίες τιμές, όντας λιγότερο ευαίσθητος λόγω της ταξινόμησης των μεταβλητών και δίνοντας παρόμοια βαρύτητα σε όλες τις τιμές.

3.3 ΄Ελεγχος υπόθεσης συσχέτισης

Στην περίπτωση γραμμικών συσχετίσεων οι οποίοι υπολογίζονται από τον συντελεστή Pearson, η ύπαρξη ή μη γραμμικής συσχέτισης μεταβλητών εισόδου X και εξόδου Y, εξετάζεται με τον έλεγχο της μηδενικής υπόθεσης ρ=0. Στην περίπτωση απόρριψης της μηδενικής υπόθεσης, οδηγούμαστε στην εναλλακτική υπόθεση ρ≠0 και σημαίνει πως η συσχέτιση των X και Y είναι στατιστικά σημαντική.

Για τον έλεγχο ύπαρξης γραμμικής συσχέτισης χρησιμοποιείται η στατιστική δείγματος για κάθε δείγμα:

$$t_0 = \frac{\rho\sqrt{\nu-2}}{\sqrt{1-\rho^2}}$$ (3.3) με:

ν: αριθμός παρατηρήσεων δείγματος.

Η στατιστική δείγματος του τύπου 3.3 στην περίπτωση που ισχύει η μηδενική υπόθεση, ακολουθεί κατανομή Student, έχοντας ν-2 βαθμούς ελευθερίας. Για επίπεδο σημαντικότητας α, η μηδενική υπόθεση απορρίπτεται αν $|t_0| > t_{1-a/2,\nu-2}$. Εάν η τιμή p-value προκύψει μικρότερη του επιπέδου σημαντικότητας α, τότε θεωρείται σημαντική η ένδειξη γραμμικής συσχέτισης των μεταβλητών.

3.4 Ανάλυση συσχετίσεων

3.4.1 Συσχετίσεις βαθμών μαθημάτων με τον βαθμό αποφοίτησης

Σε αυτήν την υποενότητα γίνεται ανάλυση των συσχετίσεων των μαθημάτων με τον βαθμό αποφοίτησης. Η ανάλυση των συσχετίσεων έγινε για όσα μαθήματα τηρούν τις εξής προϋποθέσεις: i) μετά από έλεγχο σημαντικότητας συσχέτισης διάφορης του μηδενός, για επίπεδο σημαντικότητας α=0.01, προέκυψε p-value ≤ 0.01 ii) τα μαθήματα είχαν περισσότερες από 15 επιτυχείς εξετάσεις. Η συσχέτισή τους υπολογίστηκε με τη μέθοδο Pearson.

Η ελάχιστη συσχέτιση min(ρ) είχε την τιμή 0.2484 και η μέγιστη συσχέτιση max(ρ) την τιμή 0.7199. Στον πίνακα 3.1 χωρίστηκαν τα μαθήματα σε χαμηλές συσχετίσεις για min(ρ) ≤ ρ ≤ 0.4, μέτριες συσχετίσεις για 0.4 < ρ ≤ 0.55 και ισχυρές συσχετίσεις για 0.55 < ρ ≤ max(ρ). Για τις τρεις περιπτώσεις των χαμηλών, μέτριων και υψηλών συσχετίσεων παρατηρήθηκαν μαθήματα τα οποία ανήκουν και στους 3 κύκλους σπουδών. Συγχρόνως, μαθήματα τα οποία κατατάχθηκαν στην κατηγορία των χαμηλών συσχετίσεων είναι μαθήματα τα οποία παρουσίασαν συγκέντρωση της πλειοψηφίας των βαθμολογιών σε περιορισμένα εύρη, ενώ στην κατηγορία των ισχυρών συσχετίσεων τα μαθήματα παρουσίασαν μεγαλύτερη διασπορά των βαθμών.

Πίνακας 3.1: Συσχετίσεις βαθμών μαθημάτων με τον βαθμό αποφοίτησης

Συσχετίσεις μαθημάτων με τον βαθμό αποφοίτησης, με αριθμό επιτυχών εξετάσεων 15 και πάνω και σημαντική ένδειξη συσχέτισης ρ διάφορης του μηδενός (p-value ≤ 0.01)			
Χαμηλή συσχέτιση για min(ρ) ≤ ρ ≤ 0.4			
Κωδικός μαθήματος	Αριθμός επιτυχών εξετάσεων	Συσχέτιση ρ	Έλεγχος σημαντικότητας συσχέτισης ρ ≠0
103	133	0.2484	0.001910715
105	134	0.2873	0.000358574
146	121	0.2958	0.000462102
134	63	0.3122	0.006184841
112	135	0.3283	4.4841E-05
147	134	0.3066	0.000143984
127	135	0.3254	5.22574E-05
252	87	0.3528	0.000364675
391	66	0.3295	0.003298634
376	70	0.3128	0.004032607
Χαμηλότερη συσχέτιση μαθημάτων μετά τον έλεγχο σημαντικότητας min(ρ)=0.2484			

Μέτρια συσχέτιση για 0.4 < ρ ≤0.55			
Κωδικός μαθήματος	Αριθμός επιτυχών εξετάσεων	Συσχέτιση ρ	Έλεγχος σημαντικότητας συσχέτισης ρ ≠0
101	133	0.4951	3.02729E-10
104	134	0.5319	5.81191E-12
144	125	0.4437	6.9625E-08
141	135	0.5182	2.13978E-11
109	135	0.4320	5.43564E-08
142	135	0.5241	1.13957E-11
145	62	0.4036	0.000505388
149	72	0.4750	8.91126E-06
119	134	0.4567	8.30294E-09
114	135	0.5103	4.91627E-11
116	135	0.5324	4.60798E-12
204	135	0.4481	1.50048E-08
219	135	0.4093	2.94962E-07
230	54	0.4596	0.000194165
372	64	0.4413	0.000107631
205	135	0.4672	2.95674E-09
251	119	0.4774	1.09509E-08
241	134	0.5396	2.45004E-12
249	131	0.4986	2.93761E-10
327	65	0.4413	9.5435E-05
318	52	0.5450	9.41763E-06
356	99	0.4806	1.43766E-07
389	88	0.4727	1.09436E-06
316	48	0.5272	4.19478E-05
371	32	0.4450	0.004994659

Ισχυρή συσχέτιση για 0.55 < ρ ≤max(ρ)			
Κωδικός μαθήματος	Αριθμός επιτυχών εξετάσεων	Συσχέτιση ρ	Έλεγχος σημαντικότητας συσχέτισης ρ ≠0 (p-value ≤0.01)
102	134	0.5824	1.23235E-14
110	134	0.5728	4.36318E-14
107	134	0.5532	5.00266E-13
132	134	0.6444	0
120	134	0.5725	4.56302E-14
137	134	0.5930	2.88658E-15
118	135	0.5705	4.70735E-14
140	134	0.5519	5.86331E-13
138	134	0.5635	1.42553E-13
117	135	0.6461	0
123	135	0.6336	0
131	133	0.6744	0
133	135	0.6115	0
106	134	0.5971	0
207	135	0.5684	6.19504E-14
254	87	0.5804	6.09432E-10
228	49	0.6590	4.04159E-08
206	118	0.5908	1.67089E-13
255	16	0.6604	0.002112166
210	72	0.6251	5.5758E-10
250	83	0.6845	3.41949E-14
246	27	0.6074	0.00027714
350	51	0.7199	1.61261E-10
377	21	0.5781	0.002569416
387	25	0.6536	0.000123076
349	95	0.6067	7.43694E-12
383	27	0.6424	9.42695E-05
390	35	0.6382	9.71835E-06
386	21	0.6413	0.000628407
Υψηλότερη συσχέτιση μαθημάτων μετά τον έλεγχο σημαντικότητας max(ρ)=0.7199			

3.4.2 Συσχετίσεις συντελεστή καθυστέρησης με βαθμό μαθημάτων και αποφοίτησης

Στο πλαίσιο της υποενότητας 3.4.2, θα δοθεί ο ορισμός του συντελεστή καθυστέρησης και θα γίνει περαιτέρω ανάλυση συσχετίσεων με αυτόν. Πιο συγκεκριμένα ως συντελεστής καθυστέρησης για κάθε μάθημα ορίστηκε η διαφορά:

$$\Sigma.\Kappa. = (\Epsilon.\Pi. + 1) - (\Epsilon.\Epsilon. + \Epsilon.\Sigma.)$$ (3.4) με:

Ε.Π. : εξεταστική περίοδος, δηλαδή το ημερολογιακό έτος κατά το οποίο εξετάστηκε επιτυχώς το κάθε μάθημα για κάθε απόφοιτο,

Ε.Ε. : ημερολογιακό έτος εισαγωγής του κάθε αποφοίτου,

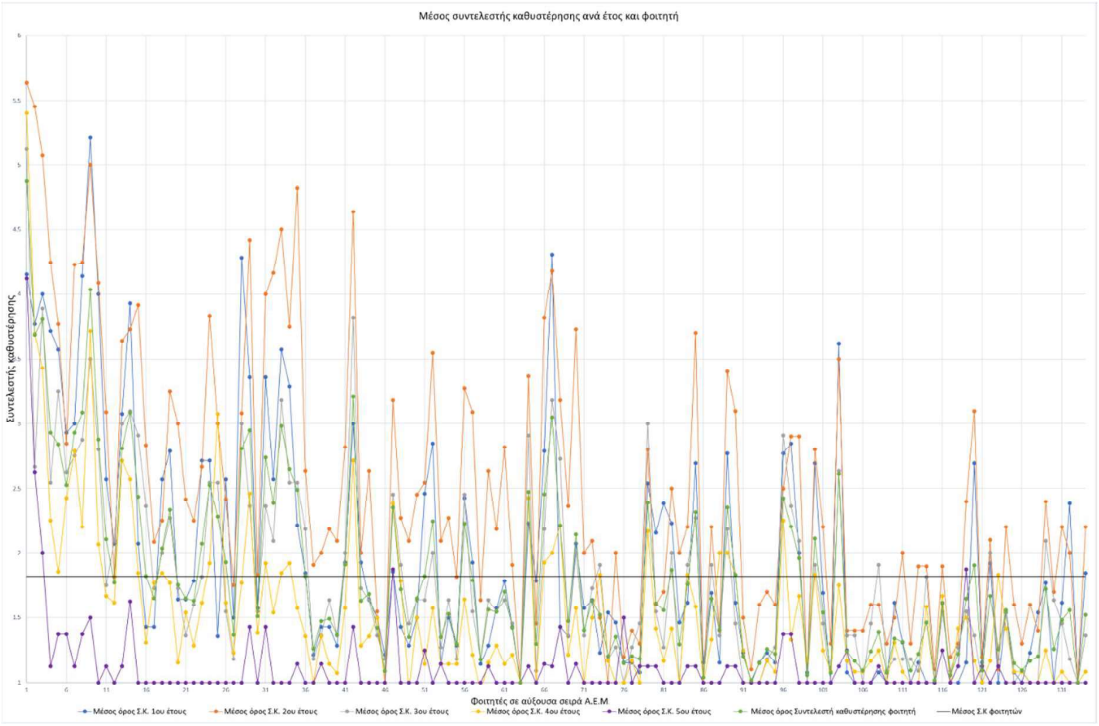
Ε.Σ. : σε πιο έτος σπουδών ανήκει το κάθε μάθημα(1, 2…, 5).

Σύμφωνα με τον τύπο 3.4, ο συντελεστής καθυστέρησης μπορεί να πάρει την ελάχιστη τιμή 1 αν το μάθημα εξετάστηκε επιτυχώς στο έτος σπουδών που του αναλογούσε. Για κάθε χρόνο που υπήρξε καθυστέρηση στην επιτυχή εξέταση ενός μαθήματος, ο συντελεστής καθυστέρησης αυξάνεται κατά 1.

Συντελεστής καθυστέρησης ανά φοιτητή

Πριν προχωρήσουμε στην ανάλυση συσχετίσεων με τον συντελεστή καθυστέρησης, χρήσιμο θα ήταν να δοθεί μία γενική εικόνα του Σ.Κ. για να εντοπιστούν τυχών τάσεις. Κάτι τέτοιο μπορεί να επιτευχθεί με τη δημιουργία ενός διαγράμματος, στο οποίο θα αναλύεται ο συντελεστής καθυστέρησης για κάθε φοιτητή σε αύξουσα σειρά Α.Ε.Μ. Αναλυτικότερα όπως φαίνεται στο σχήμα 3.1, για κάθε φοιτητή παρουσιάστηκε: i) ο μέσος όρος του συντελεστή καθυστέρησης όλων των μαθημάτων ii) ο μέσος όρος του συντελεστή καθυστέρησης των μαθημάτων ανά το έτος σπουδών στο οποίο ανήκαν και έγινε σύγκρισή τους με τον συνολικό μέσο συντελεστή καθυστέρησης όλων των φοιτητών. Από την γενική εικόνα του σχήματος 3.1 φαίνεται πως τα μαθήματα τα οποία αντιστοιχούν στο 2° έτος σπουδών παρουσιάζουν τον μεγαλύτερο συντελεστή καθυστέρησης, ενώ τα μαθήματα του 5^{ου} έτους σπουδών τον μικρότερο. Ταυτόχρονα φαίνεται πως τα μαθήματα του 5^{ου} έτους παρουσιάζουν, εκτός ελαχίστων περιπτώσεων, μικρότερο συντελεστή καθυστέρησης από αυτόν του μέσου συντελεστή καθυστέρησης όλων των φοιτητών. Τέλος παρατηρείται ότι αν ένας φοιτητής παρουσιάσει μεγάλο συντελεστής καθυστέρησης σε μαθήματα ενός έτους σπουδών, είναι πιθανό να παρουσιάσει αντιστοίχως μεγαλύτερους Σ.Κ. μαθημάτων και για τα υπόλοιπα έτη σπουδών, συγκριτικά πάντα με τους υπόλοιπους φοιτητές και σύμφωνα με τη γενική εικόνα των μαθημάτων ανάλογα με το έτος σπουδών.

Σχήμα 3.1: Διαγραμματική απεικόνιση των μέσων συντελεστών καθυστέρησης ανά έτος και φοιτητή



Συσχέτιση συντελεστή καθυστέρησης μαθημάτων του 1^{ου} κύκλου σπουδών

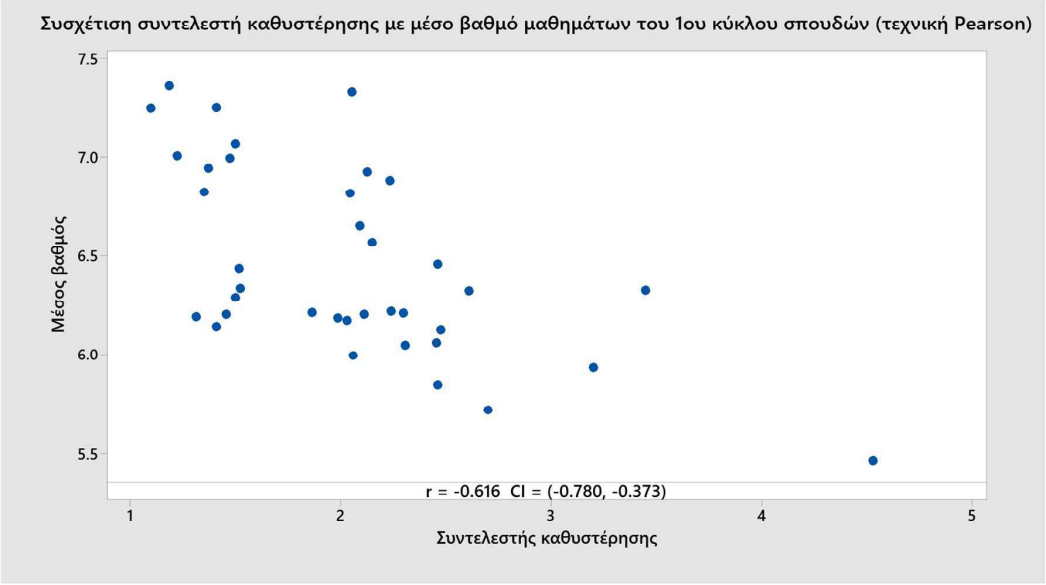
Για τα μαθήματα του 1^{ου} κύκλου σπουδών επιλέχθηκε να γίνει περαιτέρω ανάλυση του συντελεστή καθυστέρησης, καθώς αυτά τα μαθήματα ως επί το πλείστον ήταν κοινά για όλους τους αποφοίτους. Τα μαθήματα αυτά παρουσιάζονται στον πίνακα 3.2 εν συνοδεία του αριθμού των επιτυχών εξετάσεών τους, τον μέσο όρο βαθμολογιών τους και τον μέσο όρο Σ.Κ. που παρουσίασαν. Για αυτά τα μαθήματα δίνεται η συσχέτιση (μαζί με το 95% διάστημα εμπιστοσύνης της) του Σ.Κ. με τον μέσο όρο των βαθμολογιών τους στο σχήμα 3.2, μαζί με το διάγραμμα διασποράς τους. Για τις συσχετίσεις

χρησιμοποιήθηκε η τεχνική Pearson, για τους λόγους που αναλύθηκαν στο υποκεφάλαιο 3.1. Ο συντελεστής συσχέτισης έχει την τιμή ρ=-0.616, η οποία υποδηλώνει υψηλά αρνητική συσχέτιση.

Πίνακας 3.2: Ανάλυση συντελεστών καθυστέρησης των μαθημάτων του 1ου κύκλου σπουδών

	Σύγκριση μέσου συντελεστή καθυστέρησης και μέσου βαθμού μαθημάτων του 1ου κύκλου σπουδών με 15 και πάνω επιτυχής εξετάσεις			
	Κωδικός μαθήματος	Μέσος συντελεστής καθυστέρησης	Μέσος βαθμός μαθήματος	Αριθμός επιτυχών εξετάσεων
1ο Ακαδημαϊκό έτος	101	2.045	6.820	133
	103	1.474	6.996	133
	104	2.060	5.996	133
	105	2.463	6.459	134
	113	2.030	6.172	134
	144	1.516	6.435	124
	141	1.097	7.250	134
	102	2.149	6.571	134
	109	3.448	6.328	134
	111	1.985	6.187	134
	146	2.475	6.125	120
	134	1.222	7.008	63
	142	1.187	7.360	134
	145	1.500	6.290	62
	149	1.408	7.254	71
2ο Ακαδημαϊκό έτος	110	2.701	5.720	134
	107	2.299	6.209	134
	119	2.306	6.048	134
	132	2.127	6.925	134
	135	2.053	7.331	133
	148	1.370	6.944	27
	112	4.530	5.466	134
	108	2.463	5.847	134
	120	2.239	6.220	134
	137	2.090	6.654	133
	114	3.201	5.933	134
3ο Ακαδημαϊκό έτος	118	2.112	6.204	134
	140	2.112	6.204	133
	147	2.609	6.323	133
	116	2.233	6.883	134
	138	1.455	6.204	133
	133	1.865	6.214	134
	117	2.455	6.062	134
	123	1.351	6.825	134
	127	1.500	7.067	134
	131	1.410	6.142	132
	133	1.523	6.337	134
	106	2.455	6.062	125
	250	1.312	6.191	82

Σχήμα 3.2: Διάγραμμα διασποράς των μέσων συντελεστών καθυστέρησης με τις μέσες βαθμολογίες των μαθημάτων του 1ου κύκλου σπουδών

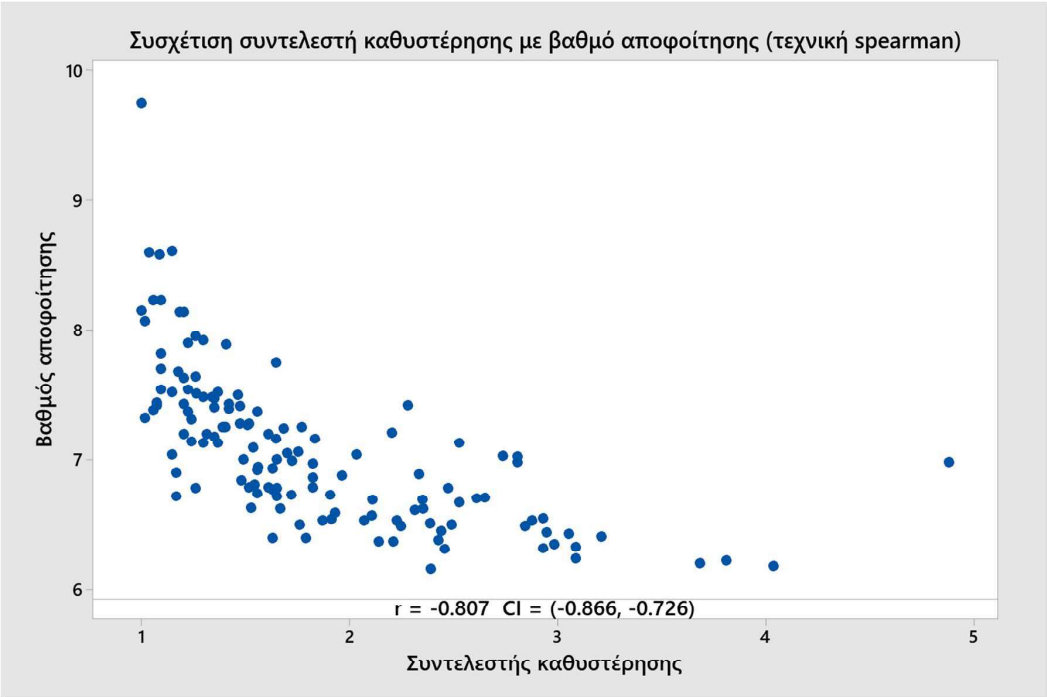


Συσχέτιση συντελεστή καθυστέρησης με βαθμό αποφοίτησης

Μία ακόμη χρήσιμη ανάλυση συσχέτισης του συντελεστή καθυστέρησης, είναι αυτή του μέσου συντελεστή καθυστέρησης ανά φοιτητή με τον βαθμό αποφοίτησης για κάθε φοιτητή. Όπως φαίνεται

στο σχήμα 3.3 από το διάγραμμα διασποράς του, η σχέση τους είναι μη γραμμική, με αποτέλεσμα να προτιμηθεί η τεχνική Spearman. Ο συντελεστής συσχέτισης προκύπτει ρ= -0.807 και υποδηλώνει ισχυρά αρνητική συσχέτιση μεταξύ των δύο. Το παραπάνω μας οδηγεί στο συμπέρασμα πως μαθήματα τα οποία εξετάσσονται επιτυχώς στο έτος φοίτησης που τους αναλογεί θα παρουσιάσουν δυνητικά και αυξημένες βαθμολογίες, ενισχύοντας την αναγκαιότητα διαρκούς παρακολούθησης των μαθημάτων.

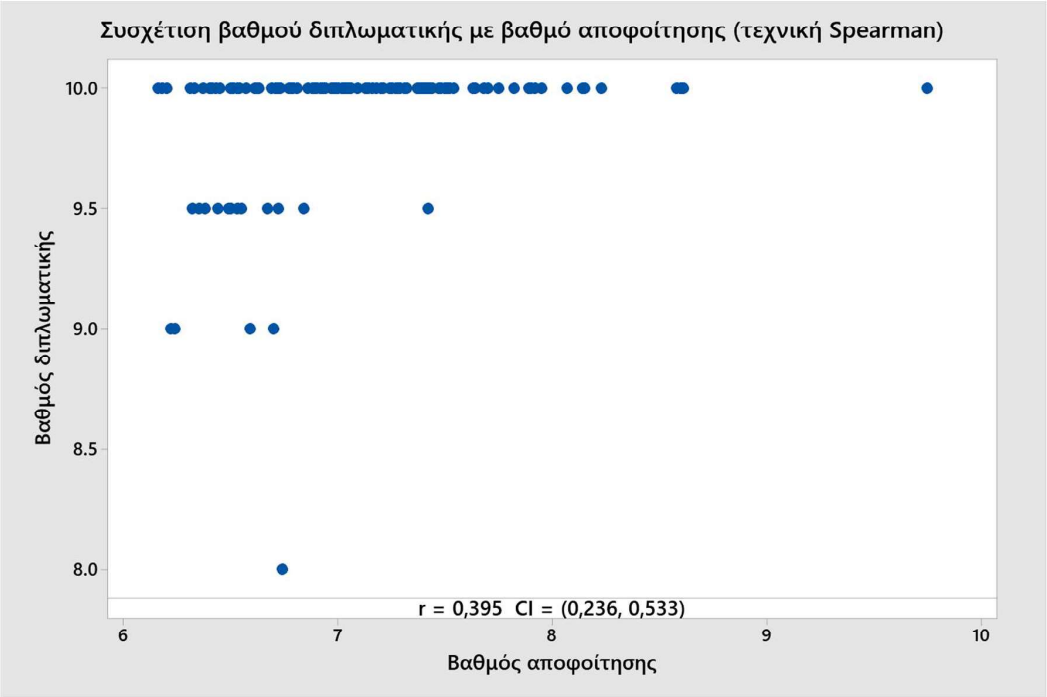
Σχήμα 3.3: Διάγραμμα διασποράς συντελεστή καθυστέρησης με βαθμό αποφοίτησης



3.4.3 Συσχετίσεις βαθμών διπλωματικής με τον βαθμό αποφοίτησης και τη διάρκεια σπουδών

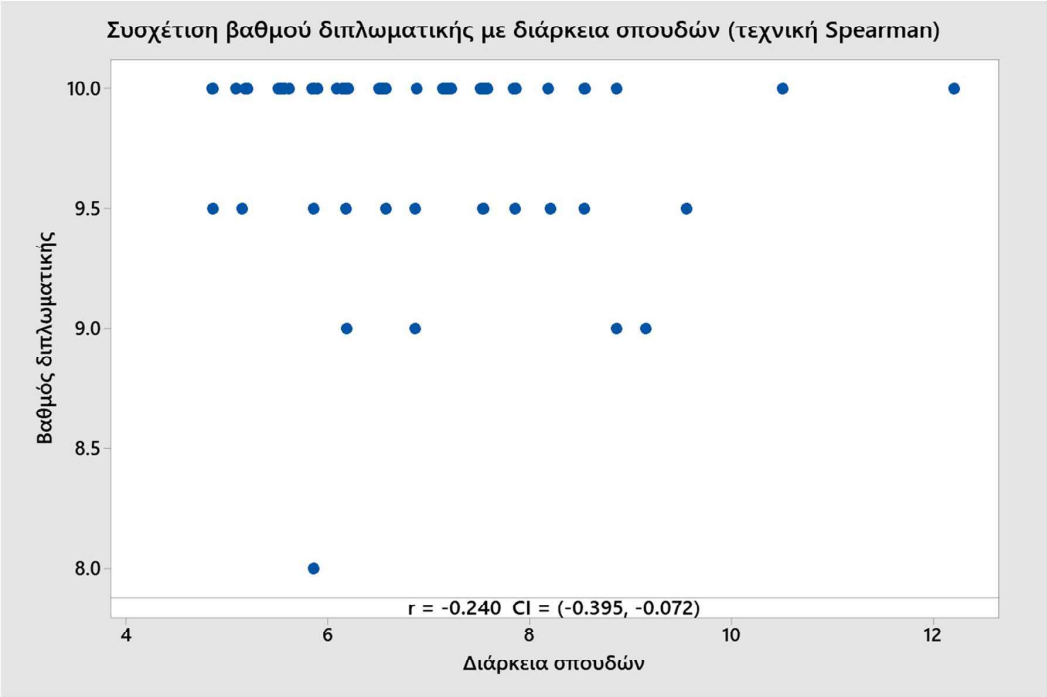
Σε αυτήν την υποενότητα ερευνάται η συσχέτιση των βαθμών διπλωματικής των αποφοίτων με τους αντίστοιχους βαθμούς αποφοίτησης και διάρκεια σπουδών τους. Αρχικά, όσον αφορά τη συσχέτιση με τον βαθμό αποφοίτησης, λόγω της εικόνας του διαγράμματος διασποράς που παρουσιάζεται στο σχήμα 3.4, επιλέχθηκε η μέθοδος προσδιορισμού συντελεστή συσχέτισης Spearman. Ειδικότερα, ο συντελεστής συσχέτισης προέκυψε ρ= 0.395, δηλαδή θετική συσχέτιση μεταξύ των δύο, όχι όμως αρκετά υψηλή. Κάτι τέτοιο σημαίνει πως ο βαθμός διπλωματικής τείνει να έχει υψηλές τιμές κοντά στο 10 ανεξαρτήτως της μέσης βαθμολογίας των μαθημάτων.

Σχήμα 3.4: Διάγραμμα διασποράς βαθμού διπλωματικής με βαθμό αποφοίτησης



Στο σχήμα 3.5 δίνεται η συσχέτιση των βαθμών διπλωματικής των φοιτητών με τις αντίστοιχες διάρκειες σπουδών τους. Για τη διερεύνηση της συσχέτισης προτιμήθηκε η τεχνική Spearman λόγω μη ξεκάθαρης σχέσης των μεταβλητών, αν και φαινομενικά πάλι φαίνεται να παρουσιάζουν μονοτονική σχέση. Ο συντελεστής συσχέτισης ο οποίος παρουσιάζεται και στο σχήμα είναι ρ= -0.240, με αποτέλεσμα να θεωρηθεί η σχέση τους αρνητικά συσχετιζόμενη, όχι όμως σε σημαντικό βαθμό.

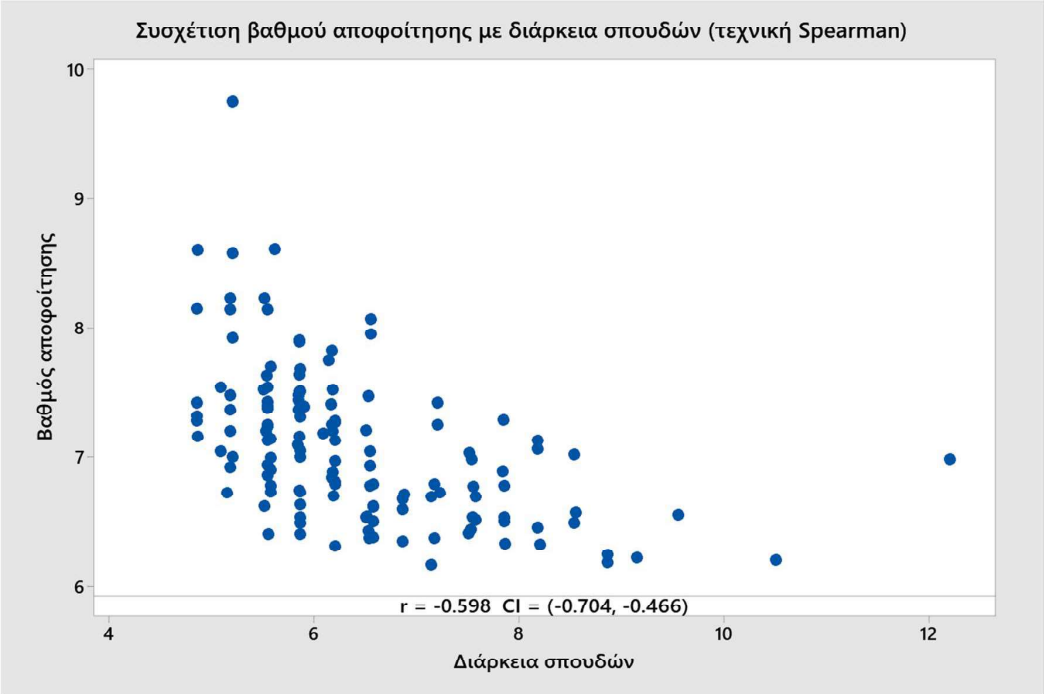
Σχήμα 3.5: Διάγραμμα διασποράς βαθμού διπλωματικής με διάρκεια σπουδών



3.4.4 Συσχέτιση βαθμού αποφοίτησης με διάρκεια σπουδών

Η τελευταία συσχέτιση η οποία διερευνάται είναι αυτή του βαθμού αποφοίτησης με τη διάρκεια σπουδών. Η εικόνα του διαγράμματος διασποράς στο σχήμα 3.6 φαίνεται να υποδηλώνει μονοτονική σχέση μεταξύ των δύο και έχει παρόμοια εικόνα με το διάγραμμα διασποράς στο σχήμα 3.2 της υποενότητας 3.4.2. Ο συντελεστής συσχέτισης των δύο είναι ρ= -0.598, ως αποτέλεσμα οι δύο μεταβλητές θεωρούνται σχετικά ισχυρά αρνητικά συσχετιζόμενες.

Σχήμα 3.6: Διάγραμμα διασποράς βαθμού αποφοίτησης με διάρκεια σπουδών



Κεφάλαιο 4: Διερεύνηση σχέσεων εξάρτησης

4.1 Απλή γραμμική παλινδρόμηση

Η προσέγγιση της απλής γραμμικής παλινδρόμησης αρχικά υποθέτει γραμμική σχέση μεταξύ μίας εξαρτημένης μεταβλητής y με μία ανεξάρτητη μεταβλητή x. Σκοπός της είναι να μοντελοποιήσει τη σχέση των δύο μεταβλητών με τη χρήση ενός απλού μαθηματικού μοντέλου, το οποίο μοντέλο μπορεί να χρησιμοποιηθεί για την πρόβλεψη της εξαρτημένης μεταβλητής y, δεδομένων νέων τιμών x. Το εν λόγω μοντέλο, το οποίο περιγράφει τη γραμμική αυτή σχέση των μεταβλητών είναι το εξής:

$y = \beta_0 + \beta_1 x + \varepsilon$ (4.1) με:

β_0, β_1: σταθερές οι οποίες ονομάζονται συντελεστές παλινδρόμησης. Οι συντελεστές αυτοί είναι άγνωστοι και σκοπός μας είναι ο προσδιορισμός των κατάλληλων τιμών τους. Όπως και στην εξίσωση ευθείας ο σταθερός όρος β_0 καθορίζει το σημείο στο οποίο η ευθεία του μοντέλου τέμνει τον άξονα y και ο όρος β_1 την κλίση της ευθείας,

ε: τυχαίο σφάλμα.

Βασική μηδενική υπόθεση της γραμμικής παλινδρόμησης είναι η: H_0: $\beta_1=0$ με εναλλακτική H_1: $\beta_1 \neq 0$. Έτσι άμα ισχύει η μηδενική υπόθεση, η μεταβλητή x δεν επηρεάζει την y, ενώ εναλλακτικά αλλαγές στη μεταβλητή της x οδηγούν σε αλλαγές της μεταβλητής y. Ταυτόχρονα για να κάνουμε γραμμική παλινδρόμηση θεωρούμε πως για κάθε τιμή της x, η y παρουσιάζει σταθερή διασπορά και ίση με τη μεταβλητότητα σ^2. Το παραπάνω ισχύει καθώς τα τυχαία σφάλματα ε θεωρούνται ανεξάρτητες τυχαίες μεταβλητές με μέση τιμή 0 και μεταβλητότητα σταθερή και ίση με σ^2, για όλα τα x. Τέλος θεωρούμε πως για κάθε x οι τιμές των y ακολουθούν κανονική κατανομή, με όλα τα παραπάνω να πρέπει να επαληθευτούν.

Για τις προβλέψεις η εξίσωση 4.1 θα πρέπει να μετασχηματιστεί στην ευθεία ελαχίστων τετραγώνων ή παλινδρόμησης:

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ (4.2) με:

$\hat{y}$: εκτίμηση τιμής y δεδομένου τιμής x,

$\hat{\beta}_0$, $\hat{\beta}_1$: εκτιμήσεις συντελεστών παλινδρόμησης.

Οι εκτιμήτριες των συντελεστών παλινδρόμησης $\hat{\beta}_0$, $\hat{\beta}_1$ υπολογίζονται με τη μέθοδο των ελαχίστων τετραγώνων, με τη χρήση της οποίας στόχος είναι η εύρεση μίας ευθείας που ελαχιστοποιεί την απόσταση κάθε σημείου (x_i, y_i), με i=1,…,n τις παρατηρήσεις και n το σύνολο των παρατηρήσεων. Για τα υπόλοιπα πρόβλεψης : $\hat{\varepsilon}_i = y_i - \hat{y}_i$, οι εκτιμήτριες $\hat{\beta}_0$, $\hat{\beta}_1$ είναι αυτές που ελαχιστοποιούν το άθροισμα τετραγώνων των υπολοίπων πρόβλεψης RSS (residual sum of squares) :

$$RSS(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x)^2 \quad (4.3)$$

Για την ελαχιστοποίηση του RSS χρησιμοποιείται απειροστικός λογισμός, δημιουργώντας σύστημα εξισώσεων των πρώτων μερικών παραγώγων του RSS ως προς τις εκτιμήτριες συντελεστών παλινδρόμησης και την ισότητά τους με το 0. Τη λύση του συστήματος αποτελούν οι εκτιμήτριες $\hat{\beta}_0$, $\hat{\beta}_1$:

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 x$ (4.4) και

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}, S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i, S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2 \quad (4.5\text{-}4.7)$$

όπου $\bar{x}$ και $\bar{y}$ οι μέσες τιμές των x_i και y_i.

4.1.1 Έλεγχος σημαντικότητας απλής γραμμικής παλινδρόμησης

Πριν προχωρήσουμε στον έλεγχο σημαντικότητας της γραμμικής παλινδρόμησης θα πρέπει να δοθούν οι εκτιμήτριες των τυπικών σφαλμάτων των $\hat{\beta}_0$ και $\hat{\beta}_1$. Αρχικά αμερόληπτη εκτιμήτρια της μεταβλητότητας του σφάλματος αποδεικνύεται πως είναι η:

$$\hat{\sigma}^2 = \frac{RSS}{n-2} \quad (4.8)$$

Οι μεταβλητότητες (Var:Variation) των εκτιμητριών των συντελεστών παλινδρόμησης υπολογίζονται από τις εξισώσεις:

$$Var(\hat{\beta}_0) = \sigma^2(\frac{1}{n} + \frac{\bar{x}}{S_{xx}}) = \sigma^2(\frac{1}{n} + \frac{\bar{x}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}) \quad (4.9) \text{ και}$$

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}} = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad (4.10) \text{ με}$$

σ^2: μεταβλητότητα του σφάλματος.

Χρησιμοποιώντας την αμερόληπτη εκτιμήτρια της μεταβλητότητας του σφάλματος της εξίσωσης 4.8 στις τετραγωνικές ρίζες των εξισώσεων 4.9 και 4.10, προκύπτουν οι εκτιμήτριες των τυπικών σφαλμάτων των συντελεστών $\hat{\beta}_0$ και $\hat{\beta}_1$:

$$s_{\hat{\beta}_0} = \sqrt{\hat{\sigma}^2(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2})} \quad (4.11) \text{ και } s_{\hat{\beta}_1} = \sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \quad (4.12)$$

Όπως προαναφέρθηκε στην αρχή του κεφαλαίου 4.1, μηδενική υπόθεση είναι η: H_0: $\beta_1=0$ με εναλλακτική H_1: $\beta_1 \neq 0$. Στη γενική περίπτωση για τον έλεγχο υπόθεσης H_0: $\beta_1=\beta_{1(0)}$, $\beta_1 \neq \beta_{1(0)}$, με $\beta_{1(0)}$ να παίρνει οποιαδήποτε τιμή, απορρίπτεται η μηδενική υπόθεση αν:

$$\left|\frac{\hat{\beta}_1-\hat{\beta}_{1(0)}}{s_{\hat{\beta}_1}}\right| > t_{1-\frac{a}{2},n-2}$$ (4.13) (κατανομή student, n-2 βαθμοί ελευθερίας για επίπεδο σημαντικότητας α)

Οπότε για : H_0: $\beta_1=0$ με εναλλακτική H_1: $\beta_1\neq 0$, η μηδενική υπόθεση απορρίπτεται αν:

$$\left|\frac{\hat{\beta}_1-0}{s_{\hat{\beta}_1}}\right| > t_{1-\frac{a}{2},n-2}$$ (4.14)

4.1.2 Καταλληλότητα γραμμικής παλινδρόμησης

Λόγω των αναμενόμενων σφαλμάτων ε για κάθε ζεύγος παρατηρήσεων (x_i, y_i), ακόμα και αν γνωρίζαμε την πραγματική ευθεία που περιγράφει τη σχέση των x,y, δεν θα μπορούσαμε να προβλέψουμε με απόλυτη ακρίβεια τις τιμές της y δεδομένου τιμών της x. Για αυτό απαραίτητη προϋπόθεση αποτελεί ο έλεγχος καταλληλότητας του προτύπου της γραμμικής παλινδρόμησης. Κάτι τέτοιο μπορεί να επιτευχθεί με τη χρήση των μέτρων RSE, RMSE, MAE και R^2, καθώς και της ανάλυσης μεταβλητότητας ANOVA, για την οποία θα γίνει ανάλυση της μεθοδολογίας στο κεφάλαιο 4.3.

RSE, RMSE και MAE

Από τον ορισμό του αθροίσματος τετραγώνων σφάλματος $RSS(\hat{\beta}_0,\hat{\beta}_1)$ στο κεφάλαιο 4.1, γίνεται κατανοητό πως το RSS χαρακτηρίζεται από n-2 βαθμούς ελευθερίας, λόγω των εκτιμητριών παλινδρόμησης. Το RSE (residual standard error) αποτελεί μέτρο εκτίμησης της μεταβλητότητας των σφαλμάτων και μας δείχνει το κατά πόσο θα αποκλίνουν οι εκτιμήσεις κατά μέσο όρο από την πραγματική γραμμή της παλινδρόμησης. Δηλαδή το πηλίκο : $\frac{RSE}{\bar{y}}$ θα μας έδειχνε το σφάλμα σε ποσοστό της εκατό. Για τον υπολογισμό του χρησιμοποιείται η εξίσωση :

$$RSE=\sqrt{\frac{RSS}{n-2}}=\sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}$$ (4.15)

Όσο πιο κοντά βρίσκονται οι τιμές των προβλέψεων στις πραγματικές τιμές, τόσο μικρότερη και η τιμή του RSE.

Η διαφορά του RMSE (Root mean squared error) με το RSE είναι πως δεν συμπεριλαμβάνει τους βαθμούς ελευθερίας, άρα:

$$RMSE=\sqrt{\frac{RSS}{n}}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}$$ (4.16)

Το RMSE σαν μέτρο μπορεί να πάρει τιμές από το μηδέν ως το άπειρο, ανεξαρτήτως της πορείας των σφαλμάτων. Να σημειωθεί πως το RMSE είναι πιο ευαίσθητο στην ύπαρξη ακραίων τιμών συγκριτικά με το μέσο απόλυτο σφάλμα:

$$MAE=\frac{1}{n}\sum_{i=1}^{n}|y_i-\hat{y}_i|$$ (4.17)

και η χρήση του ενός έναντι του άλλου εξαρτάται από την περίπτωση. Τέλος από τον ορισμό τους ισχύει : $MAE \leq RMSE$ και μόνο εάν τα σφάλματα πρόβλεψης έχουν όλα την ίδια βαρύτητα, ισχύει η ισότητα.

Συντελεστής προσδιορισμού R^2 και προσαρμοσμένου R^2

Στο υποκεφάλαιο 4.1 είχαμε ορίσει ως RSS το άθροισμα των τετραγώνων των υπολοίπων πρόβλεψης, δηλαδή τις αποκλίσεις των πραγματικών τιμών από την γραμμή της παλινδρόμησης. Ως συνολικό άθροισμα τετραγώνων TSS (total sum of squares) ορίζουμε την ποσότητα :

$$TSS = \sum_{i=1}^{n} (y_i - \overline{y})^2 \quad (4.18)$$

Το συνολικό άθροισμα τετραγώνων μετράει το μέγεθος της μεταβλητότητας της μεταβλητής εξόδου y, πριν καν γίνει η παλινδρόμηση, ενώ το RSS μετράει το μέγεθος ανεξήγητης μεταβλητότητας μετά την παλινδρόμηση.

Ως συντελεστής προσδιορισμού R^2 ορίζεται:

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS} \quad (4.19)$$

Σύμφωνα με τα παραπάνω η διαφορά TSS-RSS μετράει το μέγεθος της εξηγήσιμης μεταβλητότητας μετά την παλινδρόμηση και επομένως ο συντελεστής προσδιορισμού R^2 μας δείχνει τον βαθμό της μεταβλητότητας στις τιμές y που μπορούν να εξηγηθούν από τις τιμές x. Ο R^2 κυμαίνεται μεταξύ του 0 και του 1, με την τιμή του να τείνει στη μονάδα όσο καλύτερη η προσαρμογή της ευθείας της παλινδρόμησης. Πολύ υψηλή τιμή RSS θα σήμαινε πως τα πραγματικά σημεία απέχουν πολύ από τη γραμμή της παλινδρόμησης και η τιμή του R^2 θα ήταν χαμηλή.

Σημαντικό μειονέκτημα του R^2 αποτελεί το γεγονός πως στην περίπτωση της πολλαπλής γραμμικής παλινδρόμησης, που θα αναλυθεί στο υποκεφάλαιο 4.2, η τιμή του δεν μειώνεται ποτέ όσες ανεξάρτητες μεταβλητές και αν προσθέσουμε. Πιο συγκεκριμένα περιττές μεταβλητές, οι οποίες θα περιέπλεκαν το μοντέλο της πρόβλεψης, θα οδηγούσαν σε αύξηση ή απλά στασιμότητα της τιμής του R^2.

Για την αντιμετώπιση της αδυναμίας εντοπισμού περιττών μεταβλητών εισόδου χρησιμοποιείται ο προσαρμοσμένος συντελεστής προσδιορισμού R^2 (adjusted R^2). Για k αριθμό ανεξάρτητων μεταβλητών :

$$R_{adjusted}{}^2 = 1 - \frac{RSS \frac{1}{(n-k-1)}}{TSS \frac{1}{(n-1)}} \quad (4.20)$$

Άμα μία μεταβλητή περιπλέκει άδικα το μοντέλο θα επιφέρει μικρές μειώσεις στο RSS, αλλά σύμφωνα με τον παραπάνω τύπο, τέτοιες μεταβλητές λόγω της αύξησης του αριθμού ανεξάρτητων μεταβλητών k, δυνητικά θα αύξαναν το κλάσμα RSS/(n-k-1) και θα οδηγούσαν σε χαμηλότερο προσαρμοσμένο R^2. Εμπειρικά ο τρόπος αξιολόγησης του προσαρμοσμένου R^2 προκύπτει από τη σύγκρισή του με τον

απλό R^2. Δηλαδή για μικρές αυξήσεις του R^2 ο προσαρμοσμένος μπορεί να παραμείνει σταθερός ή και να μειωθεί, υποδεικνύοντας τη λανθασμένη ένδειξη χρησιμότητας μίας μεταβλητής. Αν υπάρξει σημαντική αύξηση του R^2 με την προσθήκη νέας ανεξάρτητης μεταβλητής και ταυτόχρονα υπάρξει σημαντική αύξηση του προσαρμοσμένου, αποτελεί ισχυρή ένδειξη πως η εν λόγω μεταβλητή είναι χρήσιμη στην πρόβλεψη της μεταβλητής εξόδου.

4.2 Πολλαπλή γραμμική παλινδρόμηση

Η πολλαπλή γραμμική παλινδρόμηση αποτελεί γενίκευση των μεθόδων της απλής γραμμικής παλινδρόμησης και εφαρμόζεται στην περίπτωση ύπαρξης k ανεξάρτητων μεταβλητών. Οπότε αντί να γίνει ξεχωριστά παλινδρόμηση για κάθε μία από τις ανεξάρτητες μεταβλητές, η πολλαπλή γραμμική παλινδρόμηση αξιοποιεί την επίδραση όλων των μεταβλητών ταυτόχρονα. Το εν λόγω μοντέλο, χρησιμοποιώντας διαφορετικούς συντελεστές παλινδρόμησης για κάθε μεταβλητή z=1,2,…,κ , γράφεται ως εξής:

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon$ (4.21) με:

β_0 : συντελεστής που καθορίζει τις αρχικές συνθήκες του επιπέδου,

$\beta_{z=1,2,\ldots,k}$: συντελεστές πολλαπλής γραμμικής παλινδρόμησης των ανεξάρτητων μεταβλητών. Ο κάθε συντελεστής β_z μας δείχνει την αναμενόμενη μέση μεταβολή της μεταβλητής εξόδου για κάθε μοναδιαία αλλαγή της ανεξάρτητης μεταβλητής x_z, εάν οι υπόλοιπες μεταβλητές παραμείνουν σταθερές.

Βασική μηδενική υπόθεση της πολλαπλής γραμμικής παλινδρόμησης είναι η: H_0: $\beta_1=\beta_2=\ldots=\beta_\kappa=0$ με εναλλακτική H_1: $\exists\ \beta_{(z=1,2,\ldots,k)}\neq 0$. Επομένως άμα ισχύει η μηδενική υπόθεση, οι μεταβλητές x_z δεν επηρεάζουν τη y, ενώ εναλλακτικά μεταβολές των μεταβλητών x_z οδηγούν σε αλλαγές της μεταβλητής y. Επιπροσθέτως, για να γίνει πολλαπλή γραμμική παλινδρόμηση θα πρέπει ο αριθμός συνολικών παρατηρήσεων n, για κάθε μεταβλητή, να είναι μεγαλύτερος του αριθμού μεταβλητών εισόδου k. Τα σφάλματα ε που προκύπτουν από τις αντίστοιχες τιμές παρατηρήσεων για κάθε μεταβλητή, θεωρούμε πως ακολουθούν κανονική κατανομή ως ανεξάρτητες τυχαίες μεταβλητές με μέση τιμή 0 και μεταβλητότητα σταθερή και ίση με σ^2.

Οι συντελεστές της πολλαπλής γραμμικής παλινδρόμησης είναι άγνωστοι και θα πρέπει να εκτιμηθούν, οδηγώντας μας στην εξίσωση πρόβλεψης :

$\hat{y}_{i=1,2,\ldots,n} = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \ldots + \hat{\beta}_k x_{ik}$ (4.22) με

$\hat{\beta}_0$, $\hat{\beta}_{z=1,2,\ldots,k}$: εκτιμήσεις συντελεστών πολλαπλής γραμμικής παλινδρόμησης,

$i = 1, 2, \ldots, n$: παρατήρηση.

Οι εκτιμήτριες των συντελεστών πολλαπλής γραμμικής παλινδρόμησης $\hat{\beta}_0$, $\hat{\beta}_z$ υπολογίζονται και πάλι με τη μέθοδο των ελαχίστων τετραγώνων, με στόχο την ελαχιστοποίηση του RSS, δηλαδή του

αθροίσματος των τετραγώνων των σφαλμάτων πρόβλεψης $\widehat{\varepsilon_i} = y_i - \widehat{y}_i$. Οπότε οι εκτιμήσεις των συντελεστών προκύπτει από την ελαχιστοποίηση της ποσότητας:

$$RSS(\widehat{\beta}_0, \widehat{\beta}_Z) = \sum_{i=1}^{n} \widehat{\varepsilon}_i^2 = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \widehat{\beta}_0 - \sum_{z=1}^{k} \widehat{\beta}_z \widehat{x}_{iz})^2 \quad (4.23)$$

Για την ελαχιστοποίηση του RSS θα πρέπει να ικανοποιούνται οι συνθήκες των πρώτων μερικών παραγώγων :

$$\frac{\partial RSS}{\partial \widehat{\beta}_0} = -2\sum_{i=1}^{n} (y_i - \widehat{\beta}_0 - \sum_{z=1}^{k} \widehat{\beta}_j x_{iz}) \text{ και}$$

$$\frac{\partial RSS}{\partial \widehat{\beta}_{z=1,2,\ldots,k}} = -2\sum_{i=1}^{n} (y_i - \widehat{\beta}_0 - \sum_{z=1}^{k} \widehat{\beta}_j x_{iz}) x_{iz} = 0$$

Οι παραπάνω συνθήκες μας οδηγούν στο σύστημα των κανονικών εξισώσεων ελαχίστων τετραγώνων, με τη λύση των οποίων υπολογίζουμε τις εκτιμήσεις των συντελεστών πολλαπλής γραμμικής παλινδρόμησης $\widehat{\beta}_0$, $\widehat{\beta}_{z=1,2,\ldots,k}$, με τις εξισώσεις να είναι οι εξής :

$$n\widehat{\beta}_0 + \widehat{\beta}_1 \sum_{i=1}^{n} x_{i1} + \widehat{\beta}_2 \sum_{i=1}^{n} x_{i2} + \ldots + \widehat{\beta}_k \sum_{i=1}^{n} x_{ik} = \sum_{i=1}^{n} y_i$$

$$\widehat{\beta}_0 \sum_{i=1}^{n} x_{i1} + \widehat{\beta}_1 \sum_{i=1}^{n} x_{i1} + \widehat{\beta}_2 \sum_{i=1}^{n} x_{i1}x_{i2} + \ldots + \widehat{\beta}_k \sum_{i=1}^{n} x_{i1}x_{ik} = \sum_{i=1}^{n} x_{i1}y_i$$

.....

$$\widehat{\beta}_0 \sum_{i=1}^{n} x_{i1} + \widehat{\beta}_1 \sum_{i=1}^{n} x_{i\kappa} + \widehat{\beta}_2 \sum_{i=1}^{n} x_{i\kappa}x_{i2} + \ldots + \widehat{\beta}_k \sum_{i=1}^{n} x_{i\kappa}x_{ik} = \sum_{i=1}^{n} x_{i\kappa}y_i$$

Μετά από χρήση γραμμικής άλγεβρας και για τους πίνακες:

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \ldots \\ y_n \end{pmatrix}, X = \begin{pmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1k} \\ 1 & x_{21} & x_{22} & \ldots & x_{2k} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nk} \end{pmatrix}, B = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \ldots \\ \beta_n \end{pmatrix}, e = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \ldots \\ \varepsilon_n \end{pmatrix}$$

ο πίνακας $\widehat{\mathrm{B}}$ των εκτιμήσεων των συντελεστών πολλαπλής γραμμικής παλινδρόμησης προκύπτει από την εξίσωση πινάκων :

$$\widehat{\mathrm{B}} = (X'X)^{-1}X'Y \quad (4.24)$$

Ακόμα το RSS είναι μετά από πράξεις :

$$RSS = Y'Y - \widehat{\mathrm{B}}'X'Y \quad (4.25)$$

Οι βαθμοί ελευθερίας του RSS είναι n-k-1 και αποδεικνύεται πως:

$\mathrm{E}(RSS) = \sigma^2 (n-k-1)$ (4.26) με:

$\mathrm{E}(RSS)$: μέση τιμή του RSS

και άρα αμερόληπτη εκτιμήτρια της μεταβλητότητας του σφάλματος είναι η :

$$\hat{\sigma}^2 = \frac{\mathrm{E}(RSS)}{(n-k-1)} \quad (4.27)$$

Ταυτόχρονα αμερόληπτες εκτιμήτριες των συντελεστών πολλαπλής γραμμικής παλινδρόμησης είναι οι τιμές του πίνακα $\hat{\mathrm{B}}$, εφόσον η μέση τιμή του προκύπτει να ισούται με τον πίνακα B :

$\mathrm{E}(\hat{\mathrm{B}}) = \mathrm{E}[(X'X)^{-1}X'Y] = \mathrm{E}[(X'X)^{-1}X'(\mathrm{XB}+e] = \mathrm{B}$ (4.28) για:

$\mathrm{E}(e) = 0$ και $(X'X)^{-1}X'X = 1$

4.3 Analysis of Variance (Anova)

Με την ανάλυση διακύμανσης Anova και του F-test μπορούμε να αξιολογήσουμε το κατά πόσο το εξεταζόμενο μοντέλο εξηγεί μεγάλο μέρος της διακύμανσης της εξόδου. Ουσιαστικά με τη μεθοδολογία της Anova πραγματοποιούμε έλεγχο σημαντικότητας της παλινδρόμησης. Η εν λόγω μεθοδολογία βασίζεται στον μετασχηματισμό της εξίσωσης:

$y_i = \hat{y}_i + (y_i - \hat{y}_i)$ (4.29) για κάθε παρατήρηση i, σε : $\sum_{i=1}^{n}(y_i-\bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2 + \sum_{i=1}^{n}(y_i-\hat{y})^2$ (4.30)

Για $ESS = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})$ ως το άθροισμα των τετραγώνων των σφαλμάτων, η εξίσωση δηλαδή μετασχηματίστηκε σε : $TSS = ESS + RSS$. Το συνολικό άθροισμα τετραγώνων δηλαδή χωρίστηκε σε δύο αθροίσματα τετραγώνων, με το ESS να υποδεικνύει το ανεξήγητο σφάλμα και το RSS τις αποκλίσεις που εξηγούνται από την προσαρμογή της παλινδρόμησης.

Άμα λάβουμε υπόψη τους βαθμούς ελευθερίας των αθροισμάτων τετραγώνων του σφάλματος και των υπολοίπων πρόβλεψης, για την απλή γραμμική παλινδρόμηση, ορίζουμε ως μέσα τετράγωνα σφάλματος και παλινδρόμησης αντίστοιχα τα εξής :

$MSE = \hat{\sigma}^2 = \frac{ESS}{n-2}$ (4.31) : μέσο τετραγωνικό σφάλμα (αμερόληπτη εκτιμήτρια της μεταβλητότητας του σφάλματος) και

$MSR = \frac{RSS}{1}$ (4.32) : μέσα τετράγωνα λόγω παλινδρόμησης.

Σύμφωνα με τα παραπάνω, προκύπτει ο πίνακας της ανάλυσης μεταβλητότητας ANOVA για την απλή γραμμική παλινδρόμηση:

Πηγή μεταβλητότητας	Βαθμοί ελευθερίας	Άθροισμα τετραγώνων	Μέσο τετράγωνο	F
Παλινδρόμηση	1	RSS	MSR=SSR/1	MSR/MSE
Σφάλμα	n-2	ESS	MSE=SSE/(n-2)	
Σύνολο	n-1	TSS		

Ο όρος F ακολουθεί F κατανομή με 1 και n-2 βαθμούς ελευθερίας και μας χρησιμεύει για τον έλεγχο υπόθεσης $\beta_1 \neq 0$ έναντι της $\beta_1 = 0$. Το p-value το οποίο προκύπτει μετά από σύγκριση του υπολογισμένου F με αυτό της F-κατανομής, μας δείχνει την πιθανότητα να προκύψει τόσο υψηλό F όσο το υπολογισμένο, αν ισχύει η αρχική υπόθεση $\beta_1 = 0$.

Για την πολλαπλή γραμμική παλινδρόμηση και k αριθμό ανεξάρτητων μεταβλητών, ο πίνακας της ανάλυσης μεταβλητότητας είναι :

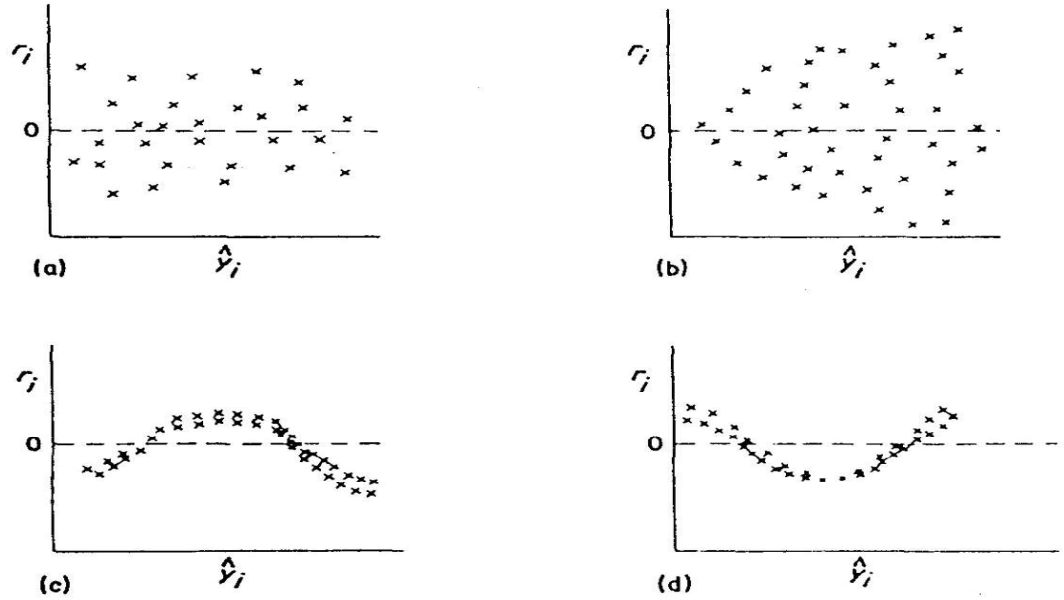
Πηγή μεταβλητότητας	Βαθμοί ελευθερίας	Άθροισμα τετραγώνων	Μέσο τετράγωνο	F
Παλινδρόμηση	k	RSS	MSR=SSR/k	MSR/MSE
Σφάλμα	n-k-1	ESS	MSE=SSE/(n-k-1)	
Σύνολο	n-1	TSS		

4.4 Αξιολόγηση εικόνας υπολοίπων

Τα υπόλοιπα $\hat{\varepsilon}_i = y_i - \hat{y}_i$ (residuals), θεωρούμε στην ANOVA πως ακολουθούν κανονική κατανομή ως ανεξάρτητες τυχαίες μεταβλητές με μηδενική μέση τιμή και σταθερή μεταβλητότητα. Ακόμα τα υπόλοιπα ως αποκλίσεις των παρατηρήσεων από την εκτιμήτρια της μέσης τιμής της y, αποτελούν εκτιμήσεις του σφάλματος ε και μπορούν να χρησιμοποιηθούν για τον έλεγχο καταλληλότητας της παλινδρόμησης. Για να θεωρηθεί η στατιστική εξάρτηση των μεταβλητών σημαντικός γραμμική και για την αξιολόγηση των αποτελεσμάτων της ANOVA, θα πρέπει να γίνει παρουσίαση της εικόνας των υπολοίπων με τη μορφή διαγραμμάτων διασποράς, ιστογραμμάτων και της χρήσης του normal probability plot.

Όσον αφορά τα διαγράμματα διασποράς των υπολοίπων, αυτά μπορούν να γίνουν με τη χρήση των υπολοίπων ή μετά την τυποποίησή τους (standardized residuals). Τα τυποποιημένα υπόλοιπα υπολογίζονται από τη διαίρεση των υπολοίπων προς την εκτιμώμενη τυπική απόκλιση. Η τυποποίηση των υπολοίπων είναι ιδιαίτερα χρήσιμη για τον εντοπισμό ακραίων τιμών. Στα διαγράμματα διασποράς θα πρέπει να γίνει αξιολόγηση της επάρκειας του γραμμικού προτύπου. Η αξιολόγηση αυτή μπορεί να γίνει με την παρατήρηση τυχών τάσεων των υπολοίπων, όπως το να παρουσιάζουν ελλειπτική μορφή. Άλλη μία εικόνα των υπολοίπων που δυνητικά μας ενδιαφέρει είναι να μην παρουσιάζουν σταθερή μορφή. Στην περίπτωση των τυποποιημένων υπολοίπων, κοιτάμε αν η κύρια ζώνη των σημείων βρίσκεται μεταξύ +/- 1 τυπικής απόκλισης και όλα τα σημεία να βρίσκονται στο πεδίο που σχηματίζουν οι +/- 3 τυπικές αποκλίσεις, ούτως ώστε να θεωρηθεί πως ακολουθούν κανονική κατανομή. Στο σχήμα 4.1 [3] φαίνονται διαφορετικές περιπτώσεις μορφής των υπολοίπων. Στην (a) περίπτωση η μορφή των υπολοίπων είναι η επιθυμητή, στην (b) περίπτωση τα υπόλοιπα δεν παρουσιάζουν σταθερή μορφή, ενώ στις περιπτώσεις (c) και (δ) αποδεικνύεται η μη γραμμικότητα των μεταβλητών και η αδυναμία του γραμμικού προτύπου να περιγράψει τη σχέση τους.

Σχήμα 4.1: Διάφορες περιπτώσεις μορφής υπολοίπων



Ειδικότερα, για την περίπτωση (b) του σχήματος 4.1, παρατηρούμε πως η διακύμανση των υπολοίπων δεν είναι σταθερή γύρω από το 0. Έτσι, οποιαδήποτε τάση αύξησης ή ελάττωσης ή γενικότερα μεταβολής της διακύμανσης οδηγούν σε μη ικανοποιητικό μοντέλο.

Μία ακόμη τακτική αξιολόγησης της εικόνας των υπολοίπων αποτελεί η δημιουργία ιστογράμματος αυτών ή των τυποποιημένων τιμών τους. Η επιθυμητή εικόνα του ιστογράμματος των υπολοίπων είναι η ύπαρξη συμμετρίας με τη μορφή καμπάνας όπως και στην κανονική κατανομή. Στην περίπτωση των ιστογραμμάτων των τυποποιημένων υπολοίπων, κοιτάμε επίσης αν οι μεγαλύτερες κλάσεις των υπολοίπων βρίσκονται μεταξύ +/- 1 τυπικής απόκλισης και αν όλες οι κλάσεις βρίσκονται μεταξύ +/- 3 τυπικών αποκλίσεων.

Για τη δημιουργία των Normal probability plot, αρχικά πρέπει τα υπόλοιπα ή οι τυποποιημένες τιμές τους να ταξινομηθούν σε αύξουσα σειρά. Έπειτα πρέπει να υπολογιστεί η αθροιστική πιθανότητα των υπολοίπων ως εξής:

$P_i = \frac{i}{n+1}$ (4.33) με:

i: σειρά κατάταξης τιμής υπολοίπου,

n: αριθμός συνολικών τιμών,

P: αθροιστική πιθανότητα i-οστού υπολοίπου.

Η διαγώνιος ευθεία του Normal probability plot είναι ενδεικτική της τέλειας προσαρμογής των δεδομένων στην κανονική κατανομή. Έτσι, άμα τα υπόλοιπα ακολουθούν κανονική κατανομή θα πρέπει να παρουσιάσουν παρόμοια μορφή με την ευθεία, έχοντας μικρές αποκλίσεις από αυτήν.

Τυχόν απορρίψεις των αρχικών υποθέσεων, μετά από ανάλυση της εικόνας των υπολοίπων, μας οδηγούν στο συμπέρασμα πως η ανάλυση ANOVA δεν ισχύει, επομένως και τα αποτελέσματα στα οποία καταλήξαμε. Έχει αποδειχθεί όμως πως με μικρές αποκλίσεις των υπολοίπων από την κανονική κατανομή ή από τη σταθερή μεταβλητότητά τους, η ANOVA δίνει ικανοποιητικά αποτελέσματα και μόνο εάν οι αποκλίσεις είναι μεγάλες δεν ισχύει η ANOVA.

4.5 Διερεύνηση σχέσεων εξάρτησης συνολικής διάρκειας φοίτησης

Στο υποκεφάλαιο αυτό θα ασχοληθούμε με την προσαρμογή μοντέλων πολλαπλής γραμμικής παλινδρόμησης χρησιμοποιώντας δεδομένα της συνολικής διάρκεια φοίτησης των αποφοίτων. Τα δεδομένα αυτά δεν μπορούν να χρησιμοποιηθούν για προβλέψεις, παρ' όλα αυτά παρουσιάζει ενδιαφέρον η ανάλυσή τους για την εύρεση αξιοποιήσιμων μεταβλητών για τις επιθυμητές προβλέψεις του βαθμού αποφοίτησης και της διάρκειας σπουδών. Μέσω της προσαρμογής μοντέλων πολλαπλής γραμμικής παλινδρόμησης μπορούμε να αξιολογήσουμε τα αποτελέσματα της ανάλυσης μεταβλητότητας (ANOVA), τους συντελεστές παλινδρόμησης και την εικόνα των υπολοίπων.

4.5.1 Πολλαπλές γραμμικές παλινδρομήσεις μέσων συντελεστών καθυστέρησης

Στο 3ο κεφάλαιο παρατηρήσαμε ισχυρές συσχετίσεις οι οποίες σχετίζονται με το κατά πόσο υπήρξε καθυστέρηση στην επιτυχή εξέταση των μαθημάτων. Στην υποενότητα αυτή θα εξετάσουμε αν υπάρχουν σχέσεις εξάρτησης του βαθμού αποφοίτησης και της διάρκειας σπουδών με τους μέσους όρους των συντελεστών καθυστέρησης των μαθημάτων των 5 ακαδημαϊκών ετών.

Ξεκινώντας με τη προσαρμογή μοντέλου πολλαπλής γραμμικής παλινδρόμησης, ορίζουμε ως μεταβλητές εισόδου τους μέσους όρους των συντελεστών καθυστέρησης ανά ακαδημαϊκό έτος και μεταβλητή εξόδου τα δεδομένα των βαθμών αποφοίτησης. Από την ανάλυση μεταβλητότητας στον πίνακα αποτελεσμάτων 4.1 μπορούμε να συμπεράνουμε πως το μέσο τετραγωνικό σφάλμα είναι αρκετά υψηλό. Κάτι τέτοιο όμως ήταν αναμενόμενο καθώς γνωρίζαμε εξαρχής πως ο βαθμός αποφοίτησης εξαρτάται από παραπάνω μεταβλητές, ειδικά από μεταβλητές σχετιζόμενες με βαθμολογίες. Ταυτόχρονα μπορούμε να παρατηρήσουμε πως για τους μέσους όρους συντελεστών καθυστέρησης του 1ου έτους και του 5ου οι συντελεστές παλινδρόμησης προκύπτουν θετικοί ενώ για τα υπόλοιπα έτη αρνητικοί. Το παραπάνω, δηλαδή η ύπαρξη θετικών αλλά και αρνητικών συντελεστών παλινδρόμησης, συμβαίνει λόγω της έντονης συσχέτισης μεταξύ των Σ.Κ. και την ύπαρξη αλληλοεπικάλυψης στην πληροφορία που παρέχουν. Η ισχυρή συσχέτισή τους παρουσιάζεται στον πίνακα 4.1. Στη συνέχεια συγκρίνοντας τα p-value των μεταβλητών, φαίνεται πως το 5ο και το 2ο έτος έχουν μεγάλη επίδραση και για p-value=0.05 φαίνονται τα τυποποιημένα μεγέθη επίδρασης των μεταβλητών στο σχήμα 4.1. Ο R^2 και το τυπικό σφάλμα παρουσιάζουν αποδεκτές τιμές, εφόσον ως μεταβλητές εισόδου χρησιμοποιούνται μόνο οι συντελεστές καθυστέρησης.

Πίνακας 4.1: Συσχετίσεις συντελεστών καθυστέρησης

Correlations

	Μέσος όρος Σ.Κ. 1ου έτους	Μέσος όρος Σ.Κ. 2ου έτους	Μέσος όρος Σ.Κ. 3ου έτους	Μέσος όρος Σ.Κ. 4ου έτους
Μέσος όρος Σ.Κ. 2ου έτους	0.844			
Μέσος όρος Σ.Κ. 3ου έτους	0.813	0.861		
Μέσος όρος Σ.Κ. 4ου έτους	0.701	0.757	0.849	
Μέσος όρος Σ.Κ. 5ου έτους	0.421	0.497	0.569	0.737

Μελετώντας εκτενέστερα τα διαγράμματα του σχήματος 4.1 και ξεκινώντας με την περίπτωση του normal probability plot, μπορούμε να παρατηρήσουμε μικρή απόκλιση των υπολοίπων από την ευθεία. Εν συνεχεία, αξιολογώντας το διάγραμμα διασποράς των υπολοίπων διακρίνουμε την αδυναμία του γραμμικού προτύπου να περιγράψει τη σχέση τους. Τέλος, από τη μελέτη του ιστογράμματος των υπολοίπων δεν παρατηρούμε την επιθυμητή εικόνα της καμπάνας.

Πίνακας αποτελεσμάτων 4.1: Ανάλυση παλινδρόμησης βαθμού αποφοίτησης με μέσους όρους συντελεστών καθυστέρησης μαθημάτων των 5 ακαδημαϊκών ετών

Regression Equation

Βαθμός αποφοίτησης = 7.752 + 0.0073 Μέσος όρος Σ.Κ. 1ου έτους
- 0.3666 Μέσος όρος Σ.Κ. 2ου έτους - 0.191 Μέσος όρος Σ.Κ. 3ου έτους
- 0.007 Μέσος όρος Σ.Κ. 4ου έτους + 0.514 Μέσος όρος Σ.Κ. 5ου έτους

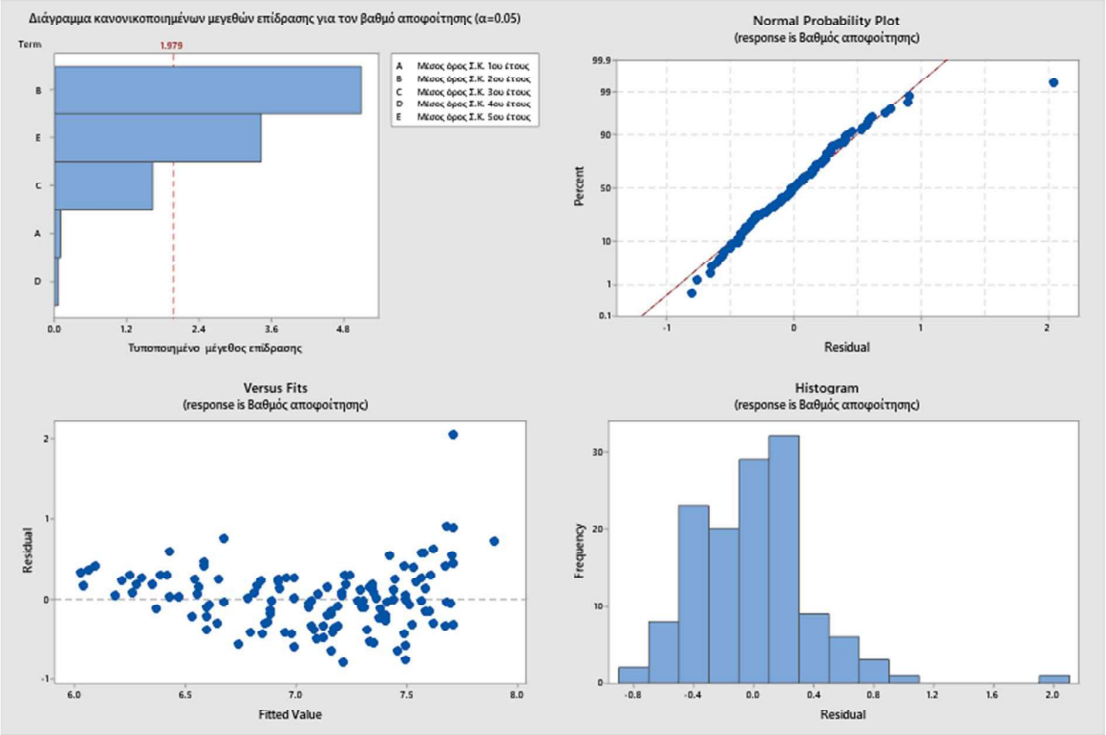
Analysis of Variance

Source	DF	Seq SS	Contribution	Adj SS	Adj MS	F-Value	P-Value
Regression	5	26.8560	57.29%	26.8560	5.37121	34.34	0.000
Μέσος όρος Σ.Κ. 1ου έτους	1	18.0999	38.61%	0.0016	0.00156	0.01	0.921
Μέσος όρος Σ.Κ. 2ου έτους	1	5.9758	12.75%	4.0192	4.01917	25.70	0.000
Μέσος όρος Σ.Κ. 3ου έτους	1	0.0454	0.10%	0.4129	0.41292	2.64	0.107
Μέσος όρος Σ.Κ. 4ου έτους	1	0.9115	1.94%	0.0006	0.00055	0.00	0.953
Μέσος όρος Σ.Κ. 5ου έτους	1	1.8235	3.89%	1.8235	1.82348	11.66	0.001
Error	128	20.0184	42.71%	20.0184	0.15639		
Lack-of-Fit	126	18.4572	39.38%	18.4572	0.14649	0.19	0.994
Pure Error	2	1.5612	3.33%	1.5612	0.78062		
Total	133	46.8745	100.00%				

Model Summary

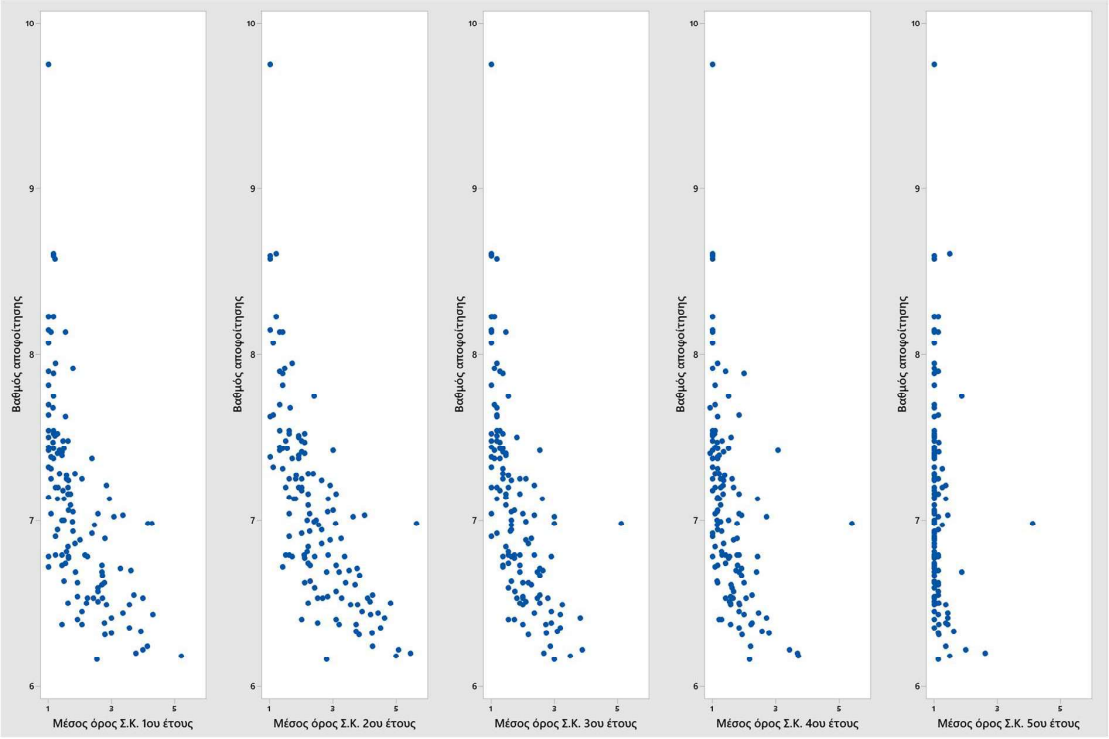
S	R-sq	R-sq(adj)	PRESS	R-sq(pred)	AICc	BIC
0.395467	57.29%	55.63%	22.1155	52.82%	140.41	159.80

Σχήμα 4.1: Διαγράμματα υπολοίπων της παλινδρόμησης του βαθμού αποφοίτησης με τους μέσους Σ.Κ.



Στο σχήμα 4.2 παρουσιάζονται τα διαγράμματα διασποράς των πέντε μεταβλητών εισόδου της υποενότητας με τον βαθμό αποφοίτησης. Για όλα τα έτη μαθημάτων οι απόφοιτοι με τους υψηλότερους βαθμούς αποφοίτησης παρουσίασαν χαμηλούς μέσους όρους στον συντελεστή καθυστέρησης, ενώ αυτοί με τους χαμηλότερους βαθμούς αποφοίτησης υψηλότερους. Για το 2° έτος υπήρξε η μεγαλύτερη διασπορά και οι υψηλότερες τιμές των συντελεστών καθυστέρησης, ενώ αντιθέτως η μικρότερη διασπορά και οι χαμηλότερες τιμές εντοπίστηκαν για το 5° έτος μαθημάτων. Η εικόνα αυτή των δύο παραπάνω διαγραμμάτων διασποράς και καθώς παρουσιάζουν σαφή διαχωρισμό μεταξύ των υψηλών και χαμηλών βαθμολογιών αποφοίτησης, φαίνεται να αποσαφηνίζει τα αποτελέσματα της παλινδρόμησης με τις μεταβλητές του 2ου και 5ου έτους να παρουσιάζουν τη μεγαλύτερη σημαντικότητα στην πρόβλεψη του βαθμού αποφοίτησης. Η διαφορά των δύο είναι πως για το 2° έτος στο διάγραμμα διασποράς υπάρχει σταδιακός διαχωρισμός μεταξύ των υψηλότερων και χαμηλότερων βαθμολογιών αποφοίτησης, ενώ για το 5° ο διαχωρισμός είναι πιο απότομος με τους περισσότερους συντελεστές καθυστέρησης να έχουν τιμή 1 ή κοντά στο 1 με εξαίρεση τους φοιτητές οι οποίοι είχαν τους χαμηλότερους βαθμούς αποφοίτησης.

Σχήμα 4.2: Διαγράμματα διασποράς των μέσων όρων συντελεστών καθυστέρησης των μαθημάτων ανά ακαδημαϊκό έτος με τον βαθμό αποφοίτησης



Διατηρώντας τις μεταβλητές εισόδου ως έχει και χρησιμοποιώντας ως μεταβλητή εξόδου τη διάρκεια σπουδών, μπορούμε να προχωρήσουμε σε ανάλυση μεταβλητότητας. Από τα αποτελέσματα προσαρμογής της παλινδρόμησης στον πίνακα αποτελεσμάτων 4.2 παρατηρούμε μία ικανοποιητική τιμή του R^2=74.76%. Όλες οι μεταβλητές επηρεάζουν θετικά την έξοδο έχοντας θετικούς συντελεστές παλινδρόμησης, με αυτήν του 3^{ου} έτους να αποτελεί εξαίρεση αλλά ταυτόχρονα να μην θεωρείται σημαντική η επίδρασή της έχοντας p-value πολύ υψηλό.

Στο σχήμα 4.3 η εικόνα των υπολοίπων είναι αρκετά ικανοποιητική. Αρχικά φαίνεται να υπάρχουν μικρές αποκλίσεις των υπολοίπων από την ευθεία του normal probability plot και ταυτόχρονα από την μελέτη του διαγράμματος διασποράς τους φαίνεται να είναι τυχαία διασκορπισμένα, παρουσιάζοντας μία σταθερή μορφή. Το ιστόγραμμά τους παρουσιάζει κι αυτό την επιθυμητή μορφή έχοντας την εικόνα της καμπάνας και υποδηλώνοντας κανονικότητα των υπολοίπων.

Πίνακας αποτελεσμάτων 4.2: Ανάλυση παλινδρόμησης διάρκειας σπουδών με μέσους όρους συντελεστών καθυστέρησης μαθημάτων των 5 ακαδημαϊκών ετών

Regression Equation

Διάρκεια σπουδών = 3.562 + 0.192 Μέσος όρος Σ.Κ. 1ου έτους + 0.314 Μέσος όρος Σ.Κ. 2ου έτους
- 0.055 Μέσος όρος Σ.Κ. 3ου έτους + 0.792 Μέσος όρος Σ.Κ. 4ου έτους
+ 0.491 Μέσος όρος Σ.Κ. 5ου έτους

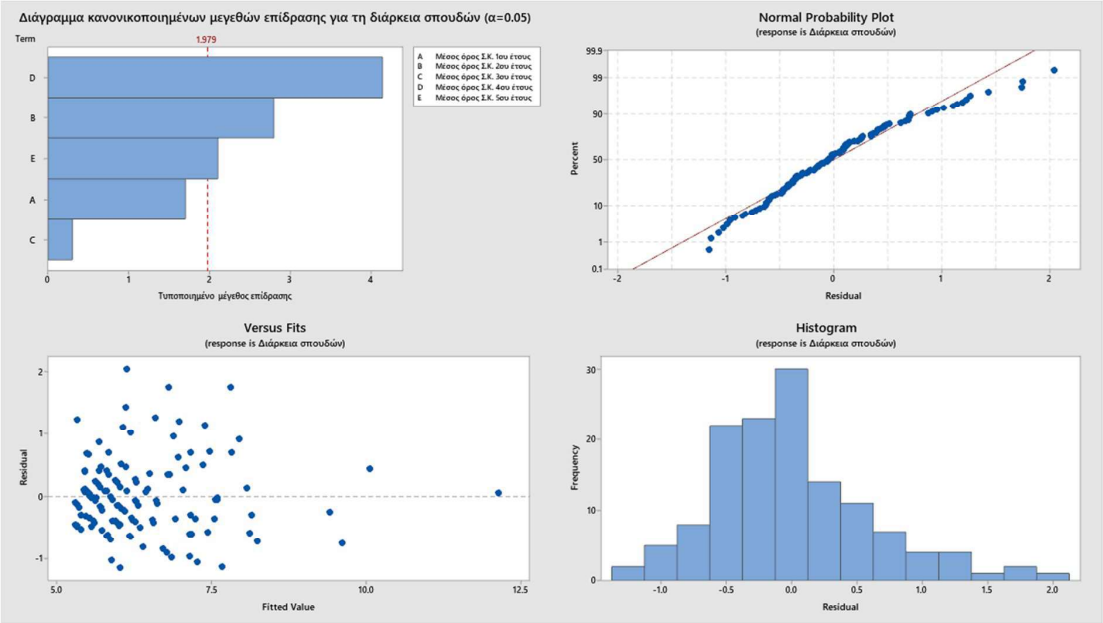
Analysis of Variance

Source	DF	Seq SS	Contribution	Adj SS	Adj MS	F-Value	P-Value
Regression	5	142.674	74.76%	142.674	28.5348	75.83	0.000
Μέσος όρος Σ.Κ. 1ου έτους	1	99.440	52.11%	1.089	1.0893	2.89	0.091
Μέσος όρος Σ.Κ. 2ου έτους	1	18.391	9.64%	2.943	2.9432	7.82	0.006
Μέσος όρος Σ.Κ. 3ου έτους	1	6.498	3.40%	0.035	0.0348	0.09	0.762
Μέσος όρος Σ.Κ. 4ου έτους	1	16.679	8.74%	6.455	6.4548	17.15	0.000
Μέσος όρος Σ.Κ. 5ου έτους	1	1.667	0.87%	1.667	1.6670	4.43	0.037
Error	128	48.168	25.24%	48.168	0.3763		
Lack-of-Fit	126	46.663	24.45%	46.663	0.3703	0.49	0.865
Pure Error	2	1.505	0.79%	1.505	0.7527		
Total	133	190.842	100.00%				

Model Summary

S	R-sq	R-sq(adj)	PRESS	R-sq(pred)	AICc	BIC
0.613445	74.76%	73.77%	52.9260	72.27%	258.06	277.46

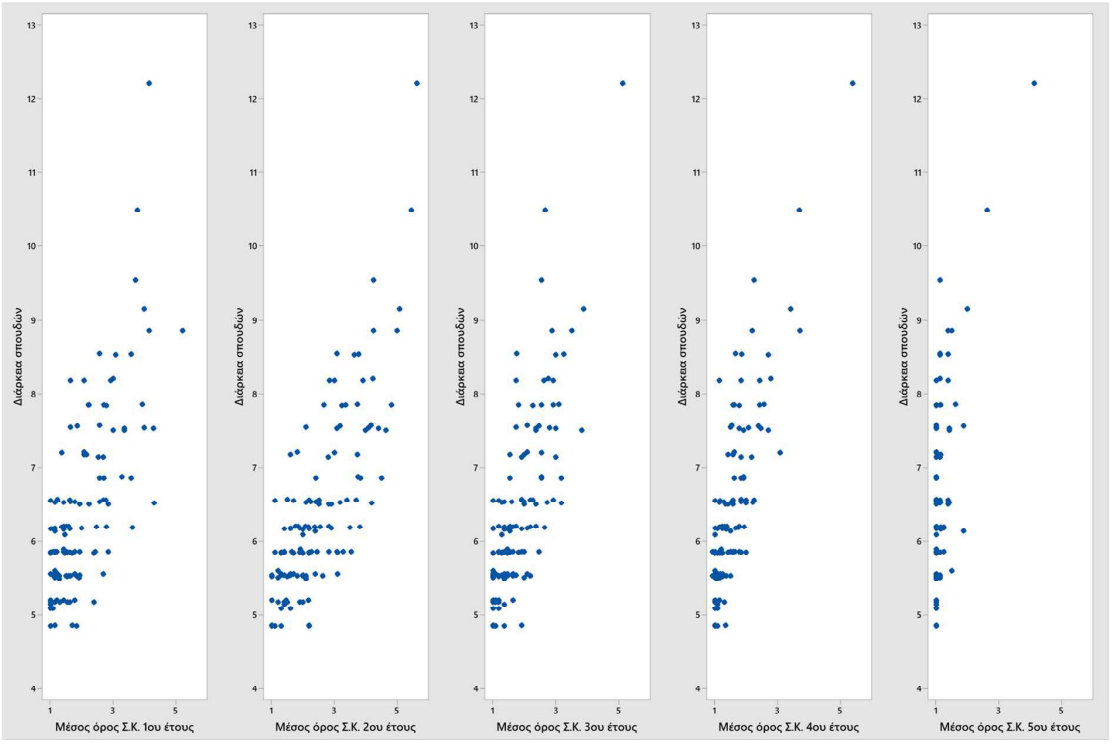
Σχήμα 4.3: Διαγράμματα υπολοίπων της παλινδρόμησης της διάρκειας σπουδών με τους μέσους Σ.Κ.



Μελετώντας τα διαγράμματα διασποράς των μεταβλητών του συντελεστή καθυστέρησης με τη διάρκεια σπουδών στο σχήμα 4.4, μπορούμε να διακρίνουμε ισχυρό διαχωρισμό των αποφοίτων. Πιο συγκεκριμένα απόφοιτοι με μικρότερη διάρκεια σπουδών παρουσίασαν και κατά γενικό κανόνα μικρότερες τιμές συντελεστών καθυστέρησης ανά τα έτη, ενώ αυτοί με τις μεγαλύτερες διάρκειες σπουδών τις μεγαλύτερες τιμές. Ο πιο σαφής διαχωρισμός για την πλειοψηφία των τιμών της

διάρκειας σπουδών παρατηρήθηκε για το 4° έτος έχοντας ταυτόχρονα σχετικά μικρή διασπορά των τιμών. Παρόμοια εικόνα παρατηρήθηκε και για το 2° έτος αλλά με μεγαλύτερη διασπορά των τιμών, ενώ μικρή διασπορά των τιμών υπήρξε και για το 5° έτος αλλά με τον διαχωρισμό τον αποφοίτων να γίνεται απότομα. Για το τρίτο έτος, για το οποίο εντοπίστηκε και το μικρότερο p-value στην παλινδρόμηση, μπορούμε να διακρίνουμε μία στασιμότητα στη διακύμανση των τιμών των συντελεστών καθυστέρησης για αποφοίτους με διάρκειες σπουδών μεγαλύτερες από 7 έτη.

Σχήμα 4.4: Διαγράμματα διασποράς των μέσων όρων συντελεστών καθυστέρησης των μαθημάτων ανά ακαδημαϊκό έτος με τη διάρκεια σπουδών



4.5.2 Πολλαπλές γραμμικές παλινδρομήσεις ποσοστών περασμένων μαθημάτων

Εκτός του συντελεστή καθυστέρησης, την πληροφορία του κατά πόσο υπήρξε καθυστέρηση στην επιτυχή ολοκλήρωση των μαθημάτων μας δίνει και το ποσοστό περασμένων μαθημάτων ανά έτος. Οι εξεταζόμενοι απόφοιτοι είχαν εξεταστεί επιτυχώς στο σύνολο των μαθημάτων μέχρι και τα 11 χρόνια φοίτησης, ως εκ τούτου μεταβλητές εισόδου σχετικές με τα ποσοστά περασμένων μαθημάτων δίνονται μέχρι και το 10ο έτος φοίτησης για να μπορέσει να γίνει προσαρμογή παλινδρόμησης. Ξεκινώντας με έξοδο τον βαθμό αποφοίτησης, ως μεταβλητές εισόδου ορίζονται τα αθροιστικά ποσοστά περασμένων μαθημάτων ανά έτος φοίτησης. Ως συντομογραφία για το αθροιστικό ποσοστό περασμένων μαθημάτων μέχρι το y έτος δίνεται η : Π.Π 1-y έτη, για έτη μεγαλύτερων του πρώτου. Τα μεγαλύτερα p-value προκύπτουν για τα πρώτα 2 χρόνια φοίτησης με R^2=61.98% και τυπικό σφάλμα 0.38 βαθμολογικές μονάδες, όπως και φαίνεται στον πίνακα αποτελεσμάτων 4.3. Τέλος, θετικούς συντελεστές παλινδρόμησης παρουσιάζουν οι μεταβλητές του 2ου, του 3ου, του 4ου ,του 6ου και του 10ου έτους, ενώ οι υπόλοιπες μεταβλητές αρνητικούς. Η ύπαρξη θετικών αλλά και αρνητικών συντελεστών παλινδρόμησης οφείλεται στις υψηλές συσχετίσεις μεταξύ των μεταβλητών όπως και φαίνεται στον πίνακα 4.2.

Πίνακας 4.2: Συσχετίσεις ποσοστών περασμένων μαθημάτων

Correlations

	Π.Π. 1ου έτους	Π.Π. 1-2 έτη	Π.Π. 1-3 έτη	Π.Π. 1-4 έτη	Π.Π. 1-5 έτη	Π.Π. 1-6 έτη	Π.Π. 1-7 έτη	Π.Π. 1-8 έτη
Π.Π. 1-2 έτη	0.896							
Π.Π. 1-3 έτη	0.831	0.931						
Π.Π. 1-4 έτη	0.761	0.839	0.935					
Π.Π. 1-5 έτη	0.623	0.677	0.801	0.897				
Π.Π. 1-6 έτη	0.491	0.504	0.585	0.674	0.814			
Π.Π. 1-7 έτη	0.327	0.318	0.386	0.464	0.611	0.885		
Π.Π. 1-8 έτη	0.233	0.208	0.269	0.340	0.477	0.749	0.947	
Π.Π. 1-9 έτη	0.180	0.136	0.190	0.246	0.370	0.626	0.873	0.964
Π.Π. 1-10 έτη	0.184	0.141	0.193	0.248	0.370	0.625	0.871	0.968

	Π.Π. 1-9 έτη
Π.Π. 1-2 έτη	
Π.Π. 1-3 έτη	
Π.Π. 1-4 έτη	
Π.Π. 1-5 έτη	
Π.Π. 1-6 έτη	
Π.Π. 1-7 έτη	
Π.Π. 1-8 έτη	
Π.Π. 1-9 έτη	
Π.Π. 1-10 έτη	0.998

Πίνακας αποτελεσμάτων 4.3: Ανάλυση παλινδρόμησης βαθμού αποφοίτησης με αθροιστικά ποσοστά περασμένων μαθημάτων ανά έτος φοίτησης

Regression Equation

Βαθμός αποφοίτησης = -34.8 - 0.807 Π.Π. 1ου έτους + 1.609 Π.Π. 1-2 έτη + 1.035 Π.Π. 1-3 έτη
+ 0.809 Π.Π. 1-4 έτη - 0.971 Π.Π. 1-5 έτη + 0.31 Π.Π. 1-6 έτη
- 0.00 Π.Π. 1-7 έτη - 0.23 Π.Π. 1-8 έτη - 33.7 Π.Π. 1-9 έτη
+ 75 Π.Π. 1-10 έτη

Analysis of Variance

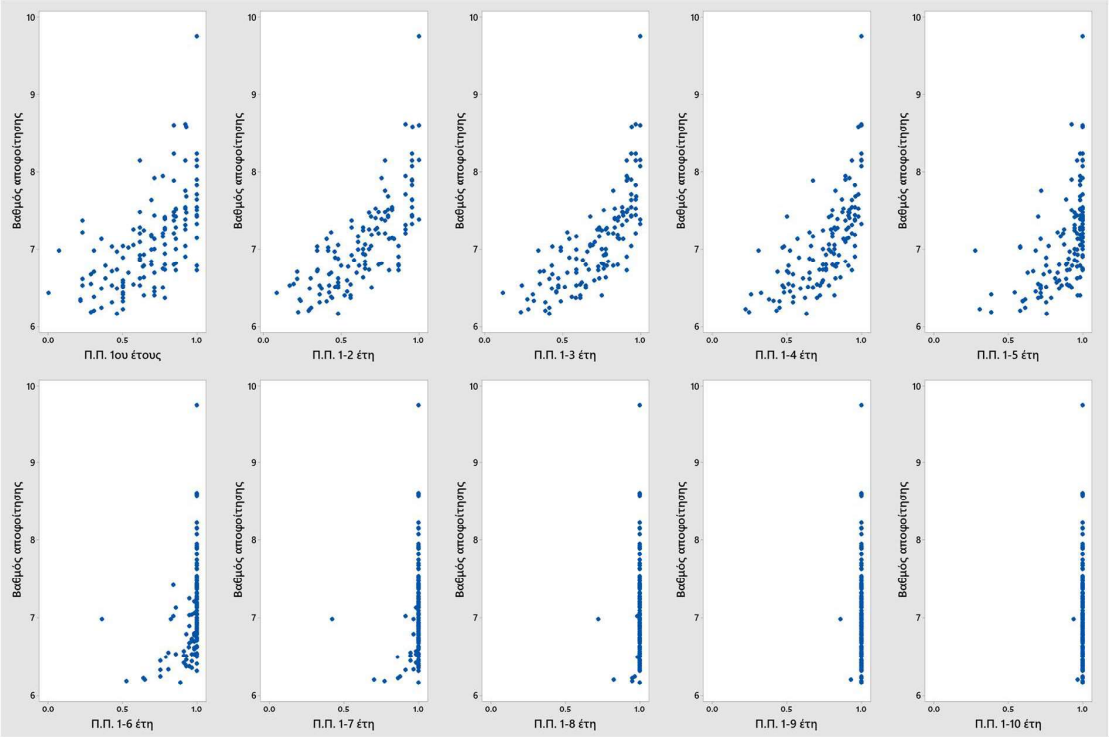
Source	DF	Seq SS	Contribution	Adj SS	Adj MS	F-Value	P-Value
Regression	10	29.0793	61.98%	29.0793	2.90793	20.05	0.000
Π.Π. 1ου έτους	1	17.3235	36.92%	0.9183	0.91834	6.33	0.013
Π.Π. 1-2 έτη	1	10.3337	22.02%	1.3426	1.34260	9.26	0.003
Π.Π. 1-3 έτη	1	0.8511	1.81%	0.3388	0.33882	2.34	0.129
Π.Π. 1-4 έτη	1	0.0013	0.00%	0.2058	0.20583	1.42	0.236
Π.Π. 1-5 έτη	1	0.4396	0.94%	0.3086	0.30859	2.13	0.147
Π.Π. 1-6 έτη	1	0.0093	0.02%	0.0076	0.00759	0.05	0.819
Π.Π. 1-7 έτη	1	0.0132	0.03%	0.0000	0.00000	0.00	1.000
Π.Π. 1-8 έτη	1	0.0048	0.01%	0.0001	0.00009	0.00	0.980
Π.Π. 1-9 έτη	1	0.0319	0.07%	0.0926	0.09259	0.64	0.426
Π.Π. 1-10 έτη	1	0.0710	0.15%	0.0710	0.07096	0.49	0.486
Error	123	17.8408	38.02%	17.8408	0.14505		
Lack-of-Fit	121	16.2795	34.70%	16.2795	0.13454	0.17	0.996
Pure Error	2	1.5612	3.33%	1.5612	0.78062		
Total	133	46.9201	100.00%				

Model Summary

S	R-sq	R-sq(adj)	PRESS	R-sq(pred)	AICc	BIC
0.380850	61.98%	58.88%	20.6696	55.95%	136.66	168.86

Η διαγραμματική απεικόνιση των ποσοστών περασμένων μαθημάτων ανά τα έτη με τον βαθμό αποφοίτησης δίνεται στο σχήμα 4.5 με τη μορφή διαγραμμάτων διασποράς. Για όλα τα έτη υπήρξε σαφής διαχωρισμός των αποφοίτων, με τα χαμηλότερα ποσοστά περασμένων μαθημάτων, για τα πρώτα 5 έτη φοίτησης, να συνδέονται με χαμηλότερους βαθμούς αποφοίτησης. Ενώ για τα πρώτα 4 έτη φοίτησης δεν υπήρξαν σημαντικά πολλοί φοιτητές οι οποίοι είχαν ολοκληρώσει τις υποχρεώσεις των έως τότε διαθέσιμων μαθημάτων, στο τέλος της υποχρεωτικής φοίτησης, δηλαδή στο 5° έτος, υπήρξε απότομη μεταβολή της εικόνας αυτής με αρκετούς φοιτητές να έχουν εξεταστεί επιτυχώς στο σύνολο των μαθημάτων.

Σχήμα 4.5: Διαγράμματα διασποράς αθροιστικών ποσοστών περασμένων μαθημάτων ανά έτος φοίτησης με βαθμό αποφοίτησης



Χρησιμοποιώντας και πάλι ως μεταβλητές εισόδου τα ποσοστά περασμένων μαθημάτων και ως έξοδο τη διάρκεια σπουδών, οδηγούμαστε στο συμπέρασμα πως η διάρκεια σπουδών εξαρτάται από τις εν λόγω μεταβλητές. Αναλυτικότερα, από την εξίσωση της πολλαπλής γραμμικής παλινδρόμησης στον πίνακα αποτελεσμάτων 4.4 παρατηρούμε ισχυρή αύξηση των συντελεστών παλινδρόμησης από το 5° έτος και μετά, με θετικούς συντελεστές να μπορούν να παρατηρηθούν για τις μεταβλητές του 1^{ου}, του 3^{ου}, του 8^{ου} και του 9^{ου} έτους. Εξετάζοντας τα p-value που προκύπτουν μετά από την ανάλυση μεταβλητότητας, τις μεγαλύτερες επιδράσεις φαίνεται να τις έχουν οι μεταβλητές Π.Π. 1-6, Π.Π. 1-8, Π.Π. 1-9 και Π.Π. 1-10 έτη. Ο R^2=78.41% είναι σχετικά υψηλός και το τυπικό σφάλμα ικανοποιητικό με τιμή 0.58 έτη φοίτησης. Να σημειωθεί πως τα αποτελέσματα είναι παρόμοια με αυτά του μοντέλου στο υποκεφάλαιο 4.5.1. όπως και ήταν λογικό.

Πίνακας αποτελεσμάτων 4.4: Ανάλυση παλινδρόμησης διάρκειας σπουδών με αθροιστικά ποσοστά περασμένων μαθημάτων ανά έτος φοίτησης

Regression Equation

Διάρκεια σπουδών = 273.8 + 0.071 Π.Π. 1ου έτους - 0.909 Π.Π. 1-2 έτη + 0.34 Π.Π. 1-3 έτη
- 0.13 Π.Π. 1-4 έτη - 3.24 Π.Π. 1-5 έτη - 4.03 Π.Π. 1-6 έτη
- 7.28 Π.Π. 1-7 έτη + 32.7 Π.Π. 1-8 έτη + 135.0 Π.Π. 1-9 έτη
- 421 Π.Π. 1-10 έτη

Analysis of Variance

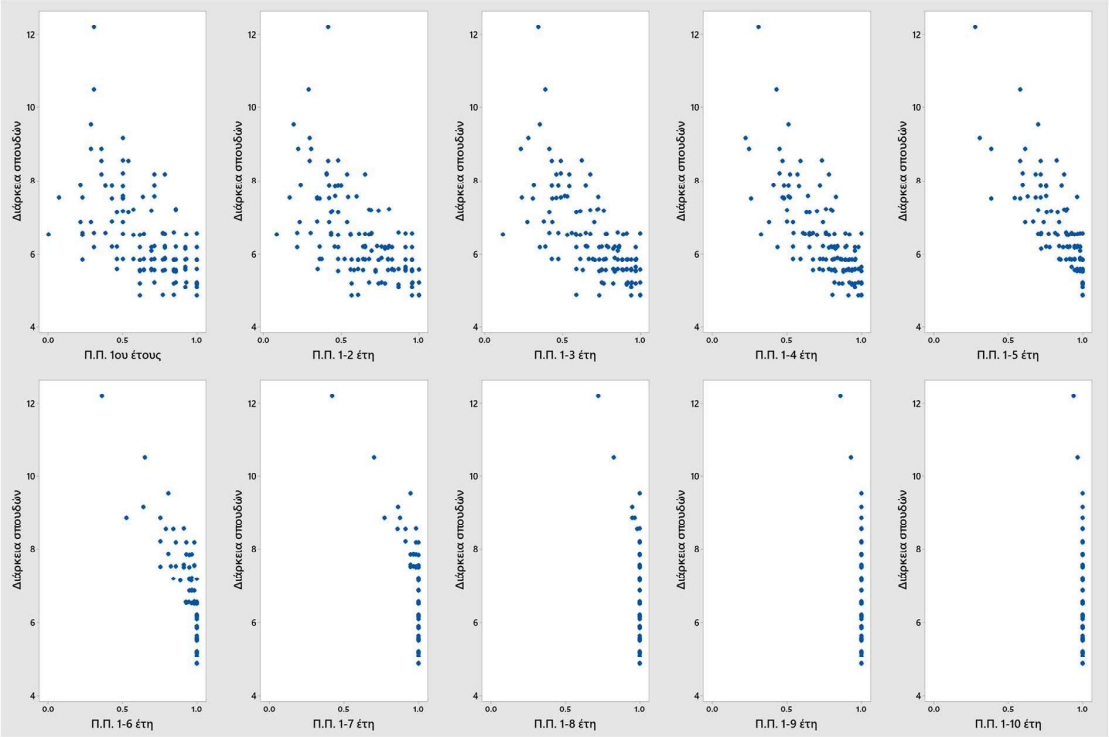
Source	DF	Seq SS	Contribution	Adj SS	Adj MS	F-Value	P-Value
Regression	10	150.145	78.41%	150.145	15.0145	44.66	0.000
Π.Π. 1ου έτους	1	63.902	33.37%	0.007	0.0072	0.02	0.884
Π.Π. 1-2 έτη	1	8.877	4.64%	0.429	0.4286	1.27	0.261
Π.Π. 1-3 έτη	1	18.162	9.48%	0.037	0.0368	0.11	0.741
Π.Π. 1-4 έτη	1	16.439	8.58%	0.005	0.0055	0.02	0.899
Π.Π. 1-5 έτη	1	25.546	13.34%	3.435	3.4354	10.22	0.002
Π.Π. 1-6 έτη	1	12.913	6.74%	1.267	1.2665	3.77	0.055
Π.Π. 1-7 έτη	1	1.246	0.65%	0.726	0.7256	2.16	0.144
Π.Π. 1-8 έτη	1	0.046	0.02%	1.825	1.8245	5.43	0.021
Π.Π. 1-9 έτη	1	0.754	0.39%	1.483	1.4830	4.41	0.038
Π.Π. 1-10 έτη	1	2.258	1.18%	2.258	2.2583	6.72	0.011
Error	123	41.353	21.59%	41.353	0.3362		
Lack-of-Fit	121	39.848	20.81%	39.848	0.3293	0.44	0.894
Pure Error	2	1.505	0.79%	1.505	0.7527		
Total	133	191.498	100.00%				

Model Summary

S	R-sq	R-sq(adj)	PRESS	R-sq(pred)	AICc	BIC
0.579832	78.41%	76.65%	60.6349	68.34%	249.31	281.51

Στα διαγράμματα διασποράς του σχήματος 4.6 των ποσοστών περασμένων μαθημάτων ανά τα έτη φοίτησης με τη διάρκεια σπουδών, ως γενικότερη εικόνα παρατηρείται αντιστρόφως ανάλογη σχέση μεταξύ των δύο. Στην παλινδρόμηση το χαμηλότερο p-value το είχε παρουσιάσει η μεταβλητή Π.Π. 1-5 έτη και αυτό μπορεί να εξηγηθεί από το γεγονός πως όσοι φοιτητές είχαν εξεταστεί επιτυχώς σε όλα τα μαθήματα μέχρι και το 5° έτος, είχαν πλέον μεγάλη πιθανότητα να παρουσιάσουν χαμηλή διάρκεια σπουδών, έχοντας μόνο να ολοκληρώσουν τη διπλωματική τους εργασία.

Σχήμα 4.6: Διαγράμματα διασποράς αθροιστικών ποσοστών περασμένων μαθημάτων ανά έτος φοίτησης με διάρκεια σπουδών



4.5.3 Πολλαπλές γραμμικές παλινδρομήσεις μέσων όρων ετών μαθημάτων

Ολοκληρώνοντας τη διερεύνηση σχέσεων εξάρτησης για τη συνολική διάρκεια φοίτησης, σημαντικό είναι να εξετάσουμε την επίδραση της βαθμολογίας των μαθημάτων στην πρόβλεψη του βαθμού αποφοίτησης. Για την επίτευξη του παραπάνω ορίσαμε ως μεταβλητές εισόδου τους μέσους όρους των μαθημάτων των 5 ακαδημαϊκών ετών για 133 εξεταζόμενους αποφοίτους. Στον πίνακα αποτελεσμάτων 4.5 μπορούμε να διαπιστώσουμε από τη μελέτη της εξίσωσης της παλινδρόμησης πως όλες οι μεταβλητές επηρεάζουν θετικά την έξοδο έχοντας θετικούς συντελεστές παλινδρόμησης. Όλοι οι συντελεστές προσδιορισμού πλησιάζουν το 100% με το τυπικό σφάλμα της παλινδρόμησης να είναι μόλις 0.035 βαθμολογικές μονάδες.

Πίνακας αποτελεσμάτων 4.5: Ανάλυση παλινδρόμησης βαθμού αποφοίτησης με τους μέσους όρους των μαθημάτων ανά ακαδημαϊκό έτος

Regression Equation

Βαθμός αποφοίτησης = 0.8554 + 0.23033 Μέσοι όροι 1ου έτους + 0.14661 Μέσοι όροι 2ου έτους
+ 0.19593 Μέσοι όροι 3ου έτους + 0.21440 Μέσοι όροι 4ου έτους
+ 0.12968 Μέσοι όροι 5ου έτους

Analysis of Variance

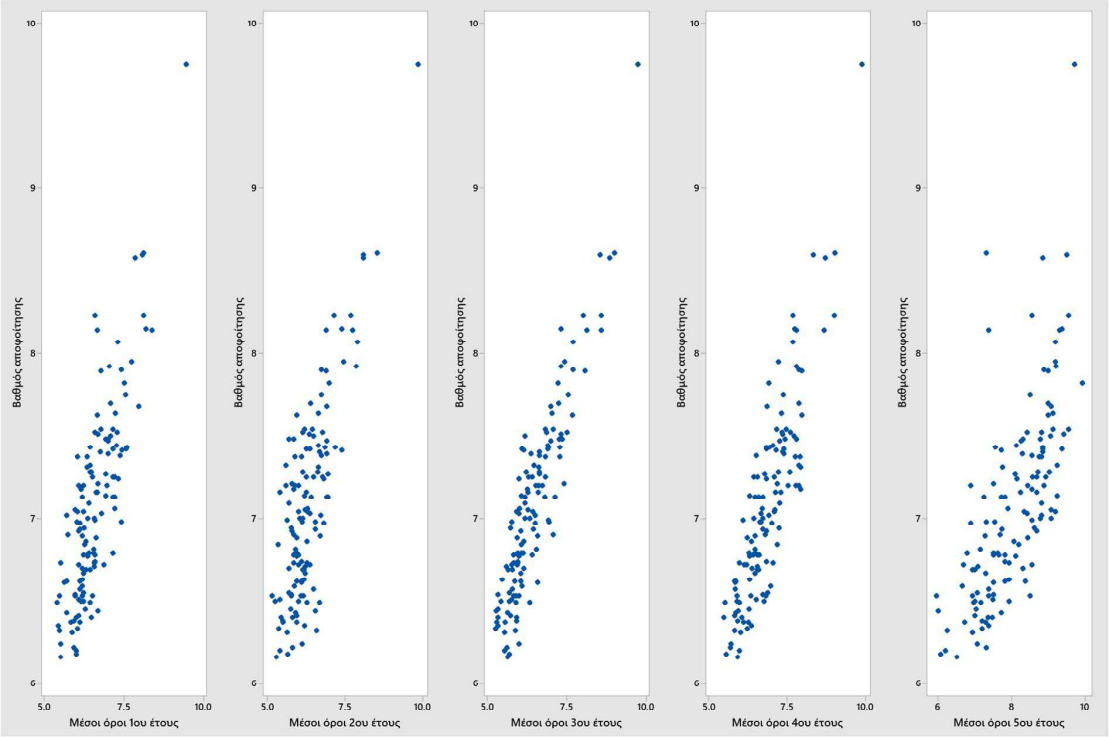
Source	DF	Seq SS	Contribution	Adj SS	Adj MS	F-Value	P-Value
Regression	5	46.7651	99.67%	46.7651	9.35302	7726.28	0.000
Μέσοι όροι 1ου έτους	1	31.1951	66.49%	1.5492	1.54917	1279.72	0.000
Μέσοι όροι 2ου έτους	1	4.7947	10.22%	0.5365	0.53654	443.22	0.000
Μέσοι όροι 3ου έτους	1	7.5675	16.13%	0.9021	0.90209	745.19	0.000
Μέσοι όροι 4ου έτους	1	2.0714	4.41%	1.2699	1.26994	1049.06	0.000
Μέσοι όροι 5ου έτους	1	1.1364	2.42%	1.1364	1.13639	938.74	0.000
Error	128	0.1549	0.33%	0.1549	0.00121		
Total	133	46.9201	100.00%				

Model Summary

S	R-sq	R-sq(adj)	PRESS	R-sq(pred)	AICc	BIC
0.0347929	99.67%	99.66%	0.167035	99.64%	-511.01	-491.61

Όπως και είναι λογικό, καθώς οι βαθμοί αποφοίτησης εξαρτώνται άμεσα από τις βαθμολογίες των μαθημάτων, υπήρξε σαφής διαχωρισμός των αποφοίτων ανά τα έτη όπως και φαίνεται στα διαγράμματα διασποράς των μέσων όρων των μαθημάτων των 5 ακαδημαϊκών ετών με τον βαθμό αποφοίτησης στο σχήμα 4.7. Από την σύγκριση των μεταβλητών εισόδου της παλινδρόμησης, οι μεταβλητές οι οποίες παρουσίασαν το μεγαλύτερο ποσοστό συνεισφοράς στην επεξήγηση των αποκλίσεων από την παλινδρόμηση, παρουσίασαν και τη μικρότερη διασπορά τιμών με τον βαθμό αποφοίτησης.

Σχήμα 4.7: Διαγράμματα διασποράς βαθμού αποφοίτησης με μέσους όρους μαθημάτων ανά ακαδημαϊκό έτος



Όσον αφορά τη διάρκεια σπουδών, η εξίσωση της πολλαπλής γραμμικής παλινδρόμησης φαίνεται στον πίνακα αποτελεσμάτων 4.6. Από τη μελέτη της εν λόγω εξίσωσης δεν προκύπτουν όλες οι μεταβλητές εισόδου να επηρεάζουν αρνητικά την έξοδο, καθώς ο συντελεστής παλινδρόμησης των μέσων όρων του 2ου έτους προέκυψε θετικός. Ταυτόχρονα σημαντική επίδραση σύμφωνα με τον πίνακα ανάλυσης μεταβλητότητας έχουν μόνο οι μεταβλητές του 1ου, του 2ου και του 5ου έτους. Ο R^2=39.85% της παλινδρόμησης προκύπτει αρκετά χαμηλός και το τυπικό σφάλμα έχοντας τιμή 0.95 έτη φοίτησης, αρκετά υψηλό. Συνεπώς, δεν προκύπτει να επηρεάζουν σημαντικά οι βαθμολογίες των μαθημάτων τη διάρκεια σπουδών.

Πίνακας αποτελεσμάτων 4.6: Ανάλυση παλινδρόμησης διάρκειας σπουδών με τους μέσους όρους των μαθημάτων ανά ακαδημαϊκό έτος

Regression Equation

Διάρκεια σπουδών = 13.369 - 0.435 Μέσοι όροι 1ου έτους + 0.453 Μέσοι όροι 2ου έτους
- 0.042 Μέσοι όροι 3ου έτους - 0.323 Μέσοι όροι 4ου έτους
- 0.557 Μέσοι όροι 5ου έτους

Analysis of Variance

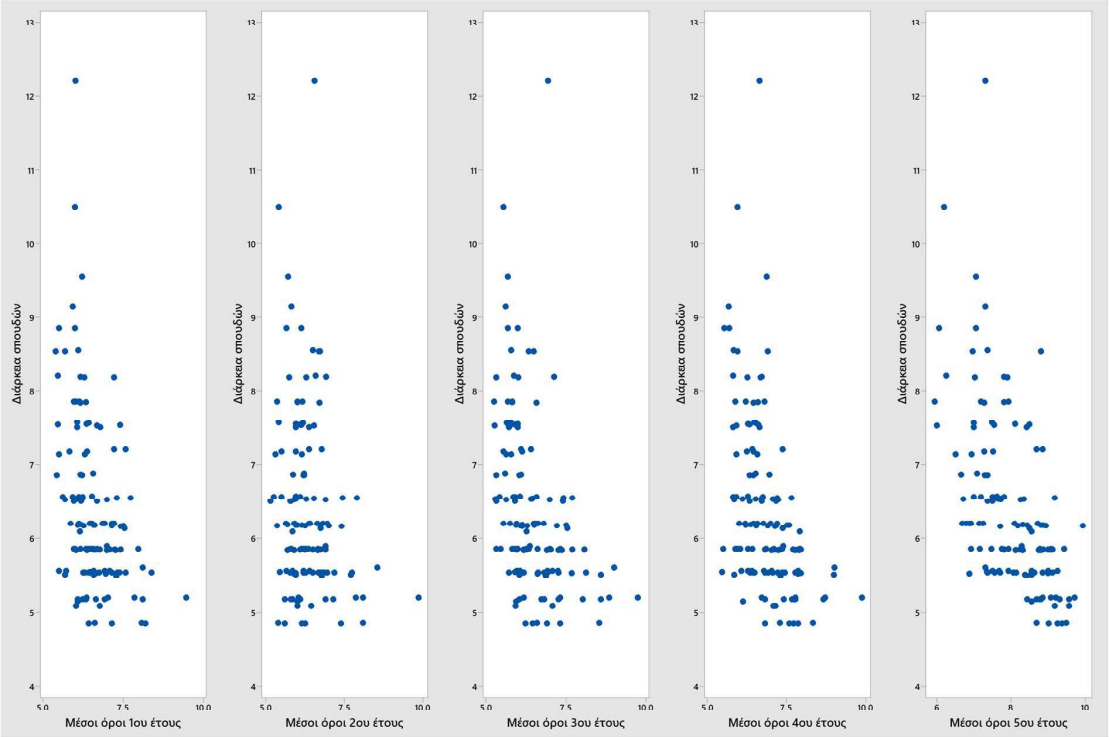
Source	DF	Seq SS	Contribution	Adj SS	Adj MS	F-Value	P-Value
Regression	5	76.304	39.85%	76.304	15.2608	16.96	0.000
Μέσοι όροι 1ου έτους	1	29.863	15.59%	5.538	5.5381	6.15	0.014
Μέσοι όροι 2ου έτους	1	1.638	0.86%	5.130	5.1299	5.70	0.018
Μέσοι όροι 3ου έτους	1	15.233	7.95%	0.041	0.0409	0.05	0.832
Μέσοι όροι 4ου έτους	1	8.624	4.50%	2.886	2.8860	3.21	0.076
Μέσοι όροι 5ου έτους	1	20.945	10.94%	20.945	20.9455	23.27	0.000
Error	128	115.194	60.15%	115.194	0.9000		
Total	133	191.498	100.00%				

Model Summary

S	R-sq	R-sq(adj)	PRESS	R-sq(pred)	AICc	BIC
0.948660	39.85%	37.50%	126.005	34.20%	374.90	394.30

Προχωρώντας στην ανάλυση των διαγραμμάτων διασποράς του σχήματος 4.8 των μέσων όρων μαθημάτων των 5 ακαδημαϊκών ετών με τη διάρκεια σπουδών, φαίνεται να αιτιολογούνται να αποτελέσματα της παλινδρόμησης. Αν και η σχέση της διάρκειας σπουδών με τις βαθμολογίες των μαθημάτων φαίνεται να είναι αντιστρόφως ανάλογη για όλα τα έτη μαθημάτων, ο διαχωρισμός των αποφοίτων δεν είναι σαφής ειδικά για τις περιπτώσεις των ενδιάμεσων τιμών.

Σχήμα 4.8: Διαγράμματα διασποράς διάρκειας σπουδών με μέσους όρους μαθημάτων ανά ακαδημαϊκό έτος



4.6 Διερεύνηση σχέσεων εξάρτησης υποχρεωτικής διάρκειας φοίτησης

Στο υποκεφάλαιο 4.5 διαπιστώσαμε πως τα ποσοστά περασμένων μαθημάτων και οι μέσοι όροι των μαθημάτων αποτελούν χρήσιμες μεταβλητές εισόδου για τις προβλέψεις του βαθμού αποφοίτησης και λιγότερο για τη διάρκεια σπουδών. Για τις προβλέψεις όμως θα πρέπει να χρησιμοποιηθούν αξιοποιήσιμες μεταβλητές εισόδου έχοντας μόνο δεδομένα τα οποία θα μπορούσαμε να έχουμε μελλοντική πρόσβαση σε αυτά. Κατά συνέπεια, αντί να γίνει εξέταση της πορείας των φοιτητών στη συνολική διάρκεια φοίτησης θα γίνει εξέταση της υποχρεωτικής τους φοίτησης ανά έτος φοίτησης. Οι ορισμένες ανεξάρτητες μεταβλητές δίνονται στον πίνακα 4.3 με :

Π.Π.χ.ν, Μ.Μ.χ.ν : αθροιστικά ποσοστά περασμένων μαθημάτων και μέσοι όροι αντίστοιχα, του χ έτους μαθημάτων στο ν έτος φοίτησης,

Π.Π.1-ν, Μ.Μ.1-ν : συνολικά ποσοστά περασμένων μαθημάτων και μέσοι όροι αντίστοιχα, μέχρι το ν έτος φοίτησης.

Για τις ανεξάρτητες μεταβλητές του πίνακα 4.3 θα γίνουν αναλύσεις συσχετίσεων μεταξύ τους και με τον βαθμό αποφοίτησης καθώς και με τη διάρκεια σπουδών, ενώ ταυτόχρονα θα γίνει προσαρμογή μοντέλων παλινδρόμησης για τον έλεγχο χρησιμότητάς τους στις προβλέψεις.

Πίνακας 4.3: Κωδική ονομασία ορισμένων ανεξάρτητων μεταβλητών υποχρεωτικής φοίτησης

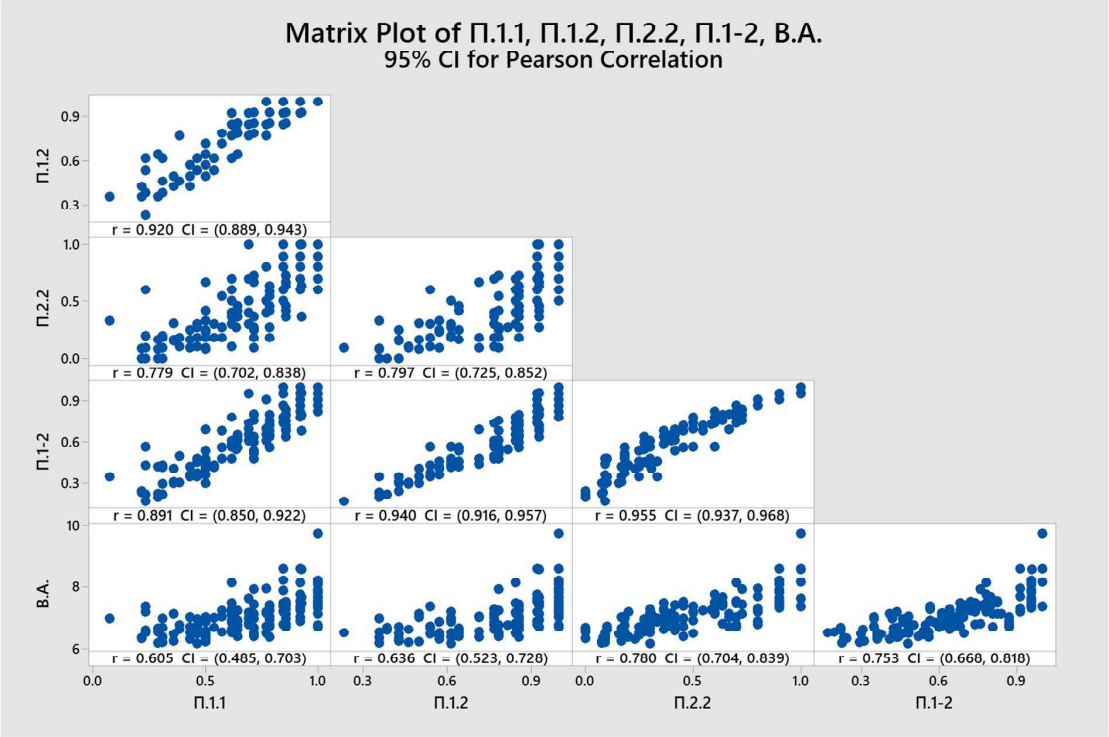
Κωδική ονομασία	Μεταβλητές υποχρεωτικής φοίτησης
Π.1.1	Π.Π. 1ου έτους στο 1ο έτος
Π.1.2	Π.Π. 1ου έτους στο 2ο έτος
Π.1.3	Π.Π. 1ου έτους στο 3ο έτος
Π.1.4	Π.Π. 1ου έτους στο 4ο έτος
Π.1.5	Π.Π. 1ου έτους στο 5ο έτος
Π.2.2	Π.Π. 2ου έτους στο 2ο έτος
Π.2.3	Π.Π. 2ου έτους στο 3ο έτος
Π.2.4	Π.Π. 2ου έτους στο 4ο έτος
Π.2.5	Π.Π. 2ου έτους στο 5ο έτος
Π.3.3	Π.Π. 3ου έτους στο 3ο έτος
Π.3.4	Π.Π. 3ου έτους στο 4ο έτος
Π.3.5	Π.Π. 3ου έτους στο 5ο έτος
Π.4.4	Π.Π. 4ου έτους στο 4ο έτος
Π.4.5	Π.Π. 4ου έτους στο 5ο έτος
Π.5.5	Π.Π. 5ου έτους στο 5ο έτος
Π.1-2	Π.Π. 1-2 Ακαδημαϊκά έτη
Π.1-3	Π.Π. 1-3 Ακαδημαϊκά έτη
Π.1-4	Π.Π. 1-4 Ακαδημαϊκά έτη
Π.1-5	Π.Π. 1-5 Ακαδημαϊκά έτη
Μ.1.1	Μ.Ο. 1ου έτους στο 1ο έτος
Μ.1.2	Μ.Ο. 1ου έτους στο 2ο έτος
Μ.1.3	Μ.Ο. 1ου έτους στο 3ο έτος
Μ.1.4	Μ.Ο. 1ου έτους στο 4ο έτος
Μ.1.5	Μ.Ο. 1ου έτους στο 5ο έτος
Μ.2.2	Μ.Ο. 2ου έτους στο 2ο έτος
Μ.2.3	Μ.Ο. 2ου έτους στο 3ο έτος
Μ.2.4	Μ.Ο. 2ου έτους στο 4ο έτος
Μ.2.5	Μ.Ο. 2ου έτους στο 5ο έτος
Μ.3.3	Μ.Ο. 3ου έτους στο 3ο έτος
Μ.3.4	Μ.Ο. 3ου έτους στο 4ο έτος
Μ.3.5	Μ.Ο. 3ου έτους στο 5ο έτος
Μ.4.4	Μ.Ο. 4ου έτους στο 4ο έτος
Μ.4.5	Μ.Ο. 4ου έτους στο 5ο έτος
Μ.5.5	Μ.Ο. 5ου έτους στο 5ο έτος
Μ.1-2	Μ.Ο. 1-2 Ακαδημαϊκά έτη
Μ.1-3	Μ.Ο. 1-3 Ακαδημαϊκά έτη
Μ.1-4	Μ.Ο. 1-4 Ακαδημαϊκά έτη
Μ.1-5	Μ.Ο. 1-5 Ακαδημαϊκά έτη

4.6.1 Έλεγχοι χρησιμότητας ανεξάρτητων μεταβλητών για πρόβλεψη του βαθμού αποφοίτησης

Συσχετίσεις

Πριν προχωρήσουμε σε προσαρμογή παλινδρόμησης των ορισμένων ανεξάρτητων μεταβλητών με τον βαθμό αποφοίτησης, χρήσιμος είναι ο έλεγχος της εικόνας των συσχετίσεων. Για την αναπαράσταση των συσχετίσεων, για τα πρώτα 2 ακαδημαϊκά έτη δίνονται οι συσχετίσεις των ποσοστών περασμένων μαθημάτων με την μέθοδο Pearson στο σχήμα 4.9. Από την μελέτη των διαγραμμάτων διασποράς μπορούμε να συμπεράνουμε πως οι όλες οι συσχετίσεις είναι σε ικανοποιητικό βαθμό γραμμικές, αιτιολογώντας τη μέθοδο Pearson. Ταυτόχρονα όλες οι συσχετίσεις προέκυψαν ισχυρά θετικές για όλους τους εξεταζόμενους όρους.

Σχήμα 4.9: Διαγράμματα διασποράς και συσχετίσεις μεταξύ των ορισμένων ανεξάρτητων μεταβλητών των ποσοστών περασμένων μαθημάτων των πρώτων 2 ακαδημαϊκών ετών φοίτησης και του βαθμού αποφοίτησης



Επεκτείνοντας τη διερεύνηση συσχετίσεων για όλες τις ορισμένες ανεξάρτητες μεταβλητές των ποσοστών περασμένων μαθημάτων, στον πίνακα 4.4 δίνονται οι συντελεστές συσχέτισης Pearson με τον βαθμό αποφοίτησης. Ως επί το πλείστον, οι συσχετίσεις είναι ισχυρά θετικές με τον βαθμό αποφοίτησης να παρουσιάζει υψηλές συσχετίσεις με όλες τις μεταβλητές με την εξαίρεση αυτής της μεταβλητής Π.5.5. Πιο συγκεκριμένα, η μεταβλητή Π.5.5 παρουσιάζει τις χαμηλότερες συσχετίσεις με όλες τις μεταβλητές και σχετικά χαμηλή συσχέτιση με τον βαθμό αποφοίτησης με τιμή 0.183. Στο τμήμα οι υψηλότερες βαθμολογίες με τη μεγαλύτερη συσσώρευση βαθμολογιών, εντοπίζονται στο 5° ακαδημαϊκό έτος, ενώ ταυτόχρονα σε αυτό το έτος παρουσιάζονται και τα υψηλότερα ποσοστά επιτυχίας. Έτσι σύμφωνα με τα παραπάνω αιτιολογούνται οι χαμηλές συσχετίσεις της μεταβλητής Π.5.5.

Τις υψηλότερες τιμές συσχετίσεων (τιμές μεγαλύτερες του 0.7) με τον βαθμό αποφοίτησης παρουσιάζουν οι μεταβλητές Π.2.2, Π.2.3, Π.3.3, Π.1-2 και Π.1-3, το οποίο υποδεικνύει τις υψηλές συσχετίσεις των μαθημάτων του 2^{ου} και 3^{ου} ακαδημαϊκού έτους, καθώς και της συνολικής πορείας των φοιτητών μέχρι το 2° και μέχρι το 3° έτος φοίτησης, με τον βαθμό αποφοίτησης. Για τη μεταβλητή Π.1-5 μπορεί να παρατηρηθεί αύξηση των συσχετίσεων με τις υπόλοιπες μεταβλητές εισόδου όσο αυξάνονται τα έτη φοίτησης και αυτό μπορεί να αιτιολογηθεί από τον ορισμό των μεταβλητών, με τα ποσοστά περασμένων μαθημάτων να αποτελούν αθροιστικά ποσοστά. Δηλαδή αυτό το οποίο παρατηρείται είναι πως για τη μεταβλητή Π.1-5 οι συσχετίσεις με τις υπόλοιπες μεταβλητές έχουν την εν λόγω εικόνα: $\rho_{\Pi.\chi.(\nu-1)} < \rho_{\Pi.\chi.\nu}$ και $\rho_{\Pi.\chi-(\nu-1)} < \rho_{\Pi.\chi-\nu}$. Το παραπάνω ισχύει και για τις ανεξάρτητες μεταβλητές των μέσων όρων των μαθημάτων, των οποίων η ανάλυση θα γίνει στη συνέχεια, έχοντας δηλαδή την εικόνα για τη μεταβλητή Μ.1-5: $\rho_{M.\chi.(\nu-1)} < \rho_{M.\chi.\nu}$ και $\rho_{M.\chi-(\nu-1)} < \rho_{M.\chi-\nu}$, εφόσον και οι μέσοι όροι δίνονται και πάλι αθροιστικά.

Πίνακας 4.4: Συσχετίσεις μεταξύ των ορισμένων ανεξάρτητων μεταβλητών των ποσοστών περασμένων μαθημάτων της υποχρεωτικής φοίτησης και του βαθμού αποφοίτησης

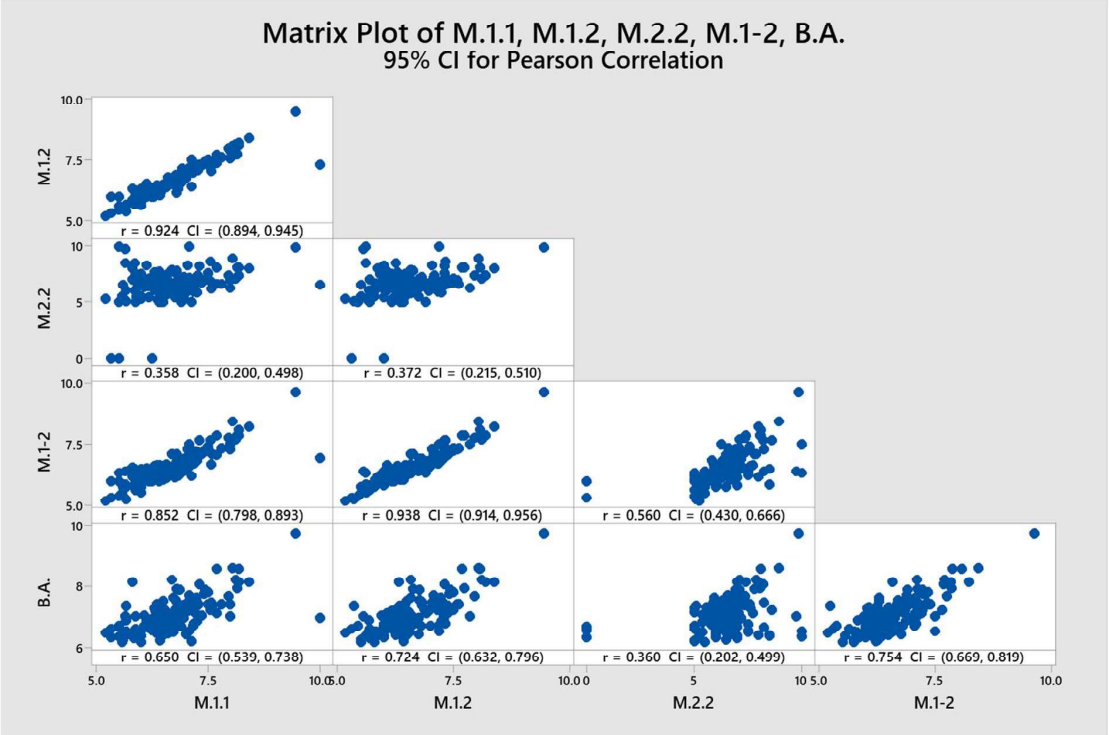
Correlations

	Π.1.1	Π.1.2	Π.1.3	Π.1.4	Π.1.5	Π.2.2	Π.2.3	Π.2.4	Π.2.5	Π.3.3
Π.1.2	0.920									
Π.1.3	0.846	0.922								
Π.1.4	0.774	0.859	0.945							
Π.1.5	0.671	0.746	0.833	0.892						
Π.2.2	0.779	0.797	0.736	0.649	0.563					
Π.2.3	0.719	0.774	0.750	0.673	0.642	0.895				
Π.2.4	0.666	0.730	0.775	0.734	0.694	0.805	0.914			
Π.2.5	0.585	0.631	0.715	0.716	0.759	0.627	0.749	0.856		
Π.3.3	0.693	0.706	0.688	0.643	0.608	0.791	0.778	0.739	0.634	
Π.3.4	0.679	0.679	0.699	0.689	0.677	0.748	0.787	0.772	0.738	0.869
Π.3.5	0.590	0.616	0.703	0.734	0.741	0.566	0.646	0.699	0.795	0.700
Π.4.4	0.559	0.542	0.582	0.547	0.579	0.602	0.656	0.656	0.658	0.731
Π.4.5	0.536	0.553	0.636	0.651	0.697	0.567	0.667	0.710	0.772	0.678
Π.5.5	0.277	0.267	0.332	0.363	0.459	0.305	0.380	0.375	0.472	0.391
Π.1-2	0.891	0.940	0.870	0.789	0.687	0.955	0.886	0.815	0.669	0.793
Π.1-3	0.820	0.873	0.883	0.820	0.759	0.883	0.924	0.887	0.770	0.914
Π.1-4	0.750	0.784	0.837	0.823	0.796	0.788	0.855	0.892	0.843	0.850
Π.1-5	0.623	0.660	0.755	0.785	0.846	0.616	0.725	0.790	0.900	0.711
Βαθμός αποφοίτησης	0.605	0.636	0.608	0.551	0.498	0.780	0.721	0.665	0.534	0.715

	Π.3.4	Π.3.5	Π.4.4	Π.4.5	Π.5.5	Π.1-2	Π.1-3	Π.1-4	Π.1-5
Π.1.2									
Π.1.3									
Π.1.4									
Π.1.5									
Π.2.2									
Π.2.3									
Π.2.4									
Π.2.5									
Π.3.3									
Π.3.4									
Π.3.5	0.841								
Π.4.4	0.763	0.668							
Π.4.5	0.766	0.778	0.805						
Π.5.5	0.486	0.514	0.505	0.658					
Π.1-2	0.758	0.627	0.609	0.596	0.305				
Π.1-3	0.870	0.759	0.723	0.725	0.405	0.928			
Π.1-4	0.914	0.836	0.872	0.843	0.497	0.832	0.933		
Π.1-5	0.826	0.904	0.758	0.924	0.691	0.677	0.807	0.900	
Βαθμός αποφοίτησης	0.634	0.480	0.555	0.501	0.183	0.753	0.754	0.685	0.525

Όσον αφορά τις ανεξάρτητες μεταβλητές των μέσων όρων των μαθημάτων, από τη μελέτη του σχήματος 4.10 με τα επιμέρους διαγράμματα διασποράς για τα πρώτα δύο ακαδημαϊκά έτη, φαίνεται να δικαιολογείται η χρήση του συντελεστή Pearson. Στην περίπτωση αυτή η ύπαρξη ή μη μηδενικών τιμών φαίνεται να επηρεάζει αρνητικά τις τιμές των συσχετίσεων λόγω της ευαισθησίας του συντελεστή στην ύπαρξη ακραίων τιμών, καθώς οι μέσοι όροι μαθημάτων παίρνουν την τιμή 5 και πάνω στην περίπτωση επιτυχούς εξέτασης ενός ή παραπάνω μαθημάτων.

Σχήμα 4.10: Διαγράμματα διασποράς και συσχετίσεις μεταξύ των ορισμένων ανεξάρτητων μεταβλητών των μέσων όρων μαθημάτων των πρώτων 2 ακαδημαϊκών ετών φοίτησης και του βαθμού αποφοίτησης



Στον πίνακα 4.5 παρουσιάζονται οι συσχετίσεις όλων των ανεξάρτητων μεταβλητών των μέσων όρων μαθημάτων με τη διάρκεια σπουδών. Οι χαμηλότερες συσχετίσεις μπορούν να παρατηρηθούν για τις μεταβλητές M.4.4, M.4.5 και M.5.5 και κάτι τέτοιο μπορεί να συνδέεται με την επιλογή κατευθύνσεων στο 4ο ακαδημαϊκό έτος του 2ου κύκλου σπουδών και στη συνέχεια στο 5ο ακαδημαϊκό έτος του 3ου κύκλου σπουδών. Οι μεγαλύτερες συσχετίσεις με τον βαθμό αποφοίτησης (τιμές μεγαλύτερες του 0.8) μπορούν να παρατηρηθούν για τις μεταβλητές M.3.4, M3.5, M.1-3, M.1-4 και M.1-5, παρουσιάζοντας πολύ ισχυρές συσχετίσεις, μεγαλύτερες από αυτές οι οποίες παρατηρήθηκαν από την ανάλυση συσχετίσεων του βαθμού αποφοίτησης με τις μεταβλητές των ποσοστών περασμένων μαθημάτων.

Πίνακας 4.5: Συσχετίσεις μεταξύ των ορισμένων ανεξάρτητων μεταβλητών των μέσων όρων μαθημάτων της υποχρεωτικής φοίτησης και του βαθμού αποφοίτησης

Correlations

	M.1.1	M.1.2	M.1.3	M.1.4	M.1.5	M.2.2	M.2.3	M.2.4	M.2.5	M.3.3
M.1.2	0.924									
M.1.3	0.923	0.984								
M.1.4	0.919	0.975	0.993							
M.1.5	0.901	0.963	0.980	0.986						
M.2.2	0.358	0.372	0.338	0.340	0.341					
M.2.3	0.381	0.438	0.426	0.424	0.438	0.780				
M.2.4	0.400	0.454	0.446	0.466	0.475	0.686	0.857			
M.2.5	0.474	0.538	0.554	0.571	0.585	0.555	0.713	0.858		
M.3.3	0.398	0.412	0.446	0.476	0.481	0.262	0.376	0.501	0.592	
M.3.4	0.444	0.478	0.523	0.549	0.550	0.279	0.390	0.509	0.634	0.822
M.3.5	0.444	0.500	0.540	0.563	0.559	0.299	0.411	0.519	0.649	0.799
M.4.4	0.134	0.155	0.161	0.175	0.204	0.205	0.281	0.257	0.234	0.282
M.4.5	0.301	0.328	0.350	0.358	0.382	0.218	0.328	0.295	0.355	0.459
M.5.5	0.232	0.279	0.304	0.313	0.322	0.090	0.173	0.193	0.228	0.399
M.1-2	0.852	0.938	0.920	0.918	0.914	0.560	0.641	0.667	0.718	0.475
M.1-3	0.760	0.832	0.855	0.868	0.865	0.494	0.637	0.717	0.814	0.729
M.1-4	0.714	0.773	0.798	0.817	0.817	0.480	0.622	0.702	0.783	0.738
M.1-5	0.656	0.726	0.759	0.776	0.784	0.401	0.554	0.637	0.754	0.749
Βαθμός αποφοίτησης	0.650	0.724	0.757	0.776	0.775	0.360	0.481	0.564	0.694	0.728

	M.3.4	M.3.5	M.4.4	M.4.5	M.5.5	M.1-2	M.1-3	M.1-4	M.1-5
M.1.2									
M.1.3									
M.1.4									
M.1.5									
M.2.2									
M.2.3									
M.2.4									
M.2.5									
M.3.3									
M.3.4									
M.3.5	0.979								
M.4.4	0.284	0.271							
M.4.5	0.560	0.568	0.483						
M.5.5	0.432	0.455	0.394	0.625					
M.1-2	0.524	0.549	0.211	0.374	0.269				
M.1-3	0.808	0.820	0.274	0.496	0.378	0.900			
M.1-4	0.845	0.855	0.371	0.609	0.409	0.839	0.964		
M.1-5	0.862	0.886	0.337	0.642	0.555	0.774	0.931	0.966	
Βαθμός αποφοίτησης	0.848	0.872	0.346	0.582	0.557	0.754	0.903	0.938	0.976

Παλινδρόμηση

Χρησιμοποιώντας ως ανεξάρτητες μεταβλητές τα ποσοστά περασμένων και τους μέσους όρους μαθημάτων για τα πρώτα 5 έτη φοίτησης για την πρόβλεψη του βαθμού αποφοίτησης, η εξίσωση της πολλαπλής γραμμικής παλινδρόμησης δίνεται στον πίνακα αποτελεσμάτων 4.7. Εξετάζοντας τα αποτελέσματα του πίνακα της ανάλυσης μεταβλητότητας μπορούμε να παρατηρήσουμε ισχυρή επιρροή αρκετών μεταβλητών εισόδου, έχοντας p-value μικρότερα του 0.05, καθώς και χαμηλό μέσο τετραγωνικό σφάλμα. Το τυπικό σφάλμα προκύπτει μόλις 0.066 βαθμολογικές μονάδες και ο R^2=99.13% σημαντικά υψηλός.

Όσον αφορά την εικόνα των υπολοίπων, μπορεί να γίνει αξιολόγησή τους μέσω των διαγραμμάτων του σχήματος 4.11. Από τη μελέτη του normal probability plot παρατηρούμε μικρές αποκλίσεις από την ευθεία εκτός ελαχίστων εξαιρέσεων, ενώ στο διάγραμμα διασποράς η πλειοψηφία των κανονικοποιημένων υπολοίπων βρίσκεται μεταξύ +/- 1 τυπικής απόκλισης και σχεδόν όλα μεταξύ +/- 3 τυπικών αποκλίσεων. Στο ιστόγραμμα των κανονικοποιημένων υπολοίπων παρατηρείται η επιθυμητή εικόνα της κανονικής κατανομής με σχετικά ασήμαντες αποκλίσεις. Σύμφωνα με τα

παραπάνω, οι ανεξάρτητες μεταβλητές οι οποίες επιλέχθηκαν φαίνεται να είναι χρήσιμες για την πρόβλεψη του βαθμού αποφοίτησης.

Πίνακας αποτελεσμάτων 4.7: Ανάλυση παλινδρόμησης βαθμού αποφοίτησης με τις ορισμένες ανεξάρτητες μεταβλητές για την υποχρεωτική διάρκεια φοίτησης

Regression Equation

Βαθμός αποφοίτησης = 0.472 - 0.0772 Π.1.1 + 0.280 Π.1.2 + 0.332 Π.1.3 - 1.921 Π.1.4
+ 1.849 Π.1.5 + 0.406 Π.2.2 + 0.357 Π.2.3 - 1.569 Π.2.4 + 1.230 Π.2.5
+ 0.383 Π.3.3 - 1.803 Π.3.4 + 1.404 Π.3.5 - 1.945 Π.4.4 + 1.331 Π.4.5
+ 0.657 Π.5.5 - 0.68 Π.1-2 - 1.23 Π.1-3 + 7.47 Π.1-4 - 5.68 Π.1-5
- 0.0074 M.1.1 - 0.0050 M.1.2 - 0.108 M.1.3 + 0.166 M.1.4 - 0.1306 M.1.5
- 0.01899 M.2.2 + 0.0081 M.2.3 - 0.0233 M.2.4 - 0.0083 M.2.5 - 0.0075 M.
3.3 + 0.0848 M.3.4 - 0.1234 M.3.5 - 0.00248 M.4.4 - 0.0542 M.4.5
- 0.0122 M.5.5 + 0.0828 M.1-2 - 0.0530 M.1-3 + 0.103 M.1-4 + 0.9732 M.
1-5

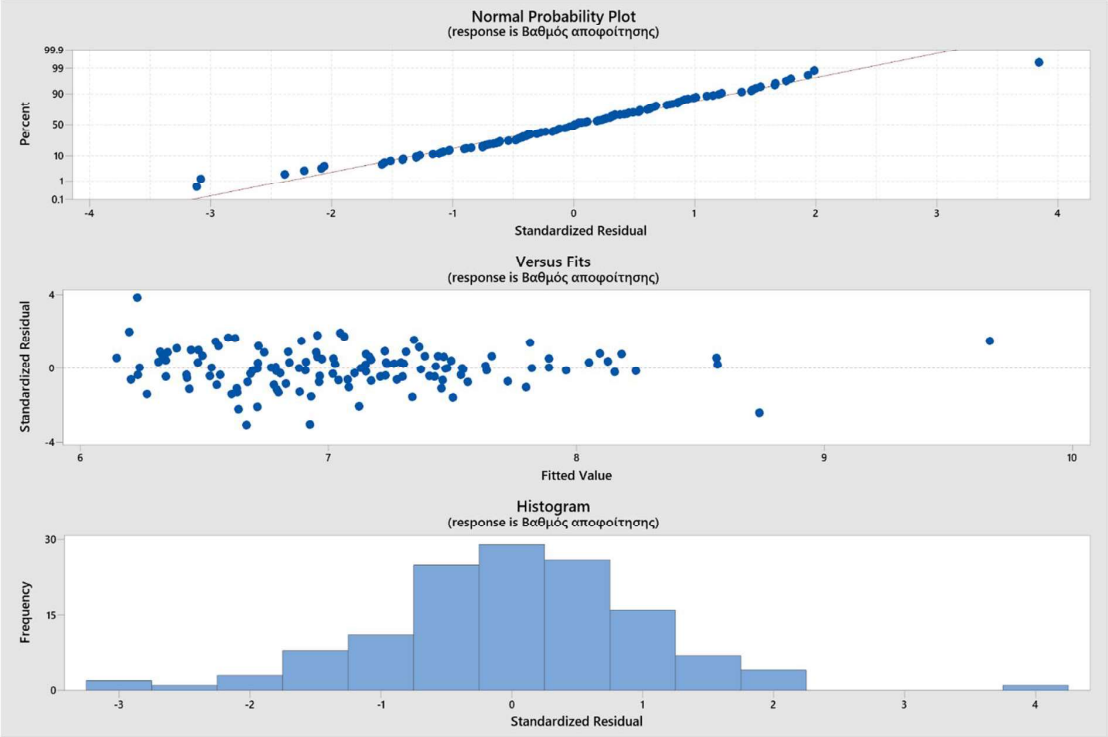
Analysis of Variance

Source	DF	Seq SS	Contribution	Adj SS	Adj MS	F-Value	P-Value
Regression	38	46.0876	99.13%	46.0876	1.21283	280.76	0.000
Π.1.1	1	17.0451	36.66%	0.0038	0.00382	0.88	0.349
Π.1.2	1	1.8979	4.08%	0.0004	0.00036	0.08	0.775
Π.1.3	1	0.1522	0.33%	0.0009	0.00087	0.20	0.655
Π.1.4	1	0.1085	0.23%	0.0242	0.02418	5.60	0.020
Π.1.5	1	0.0856	0.18%	0.0365	0.03645	8.44	0.005
Π.2.2	1	9.3546	20.12%	0.0012	0.00118	0.27	0.603
Π.2.3	1	0.0219	0.05%	0.0017	0.00170	0.39	0.532
Π.2.4	1	0.0015	0.00%	0.0265	0.02651	6.14	0.015
Π.2.5	1	0.0006	0.00%	0.0223	0.02227	5.16	0.025
Π.3.3	1	1.0085	2.17%	0.0018	0.00182	0.42	0.518
Π.3.4	1	0.1693	0.36%	0.0339	0.03389	7.85	0.006
Π.3.5	1	0.0955	0.21%	0.0300	0.02999	6.94	0.010
Π.4.4	1	0.0896	0.19%	0.0286	0.02862	6.63	0.012
Π.4.5	1	0.0252	0.05%	0.0198	0.01976	4.57	0.035
Π.5.5	1	0.7082	1.52%	0.0132	0.01318	3.05	0.084
Π.1-2	1	0.0185	0.04%	0.0006	0.00065	0.15	0.700
Π.1-3	1	0.4399	0.95%	0.0018	0.00182	0.42	0.518
Π.1-4	1	0.0396	0.09%	0.0298	0.02982	6.90	0.010
Π.1-5	1	0.1264	0.27%	0.0195	0.01946	4.51	0.036
M.1.1	1	3.3481	7.20%	0.0003	0.00035	0.08	0.778
M.1.2	1	1.7312	3.72%	0.0000	0.00002	0.00	0.948
M.1.3	1	0.6447	1.39%	0.0044	0.00441	1.02	0.315
M.1.4	1	0.8834	1.90%	0.0091	0.00906	2.10	0.151
M.1.5	1	0.1210	0.26%	0.0173	0.01729	4.00	0.048
M.2.2	1	0.5495	1.18%	0.0214	0.02139	4.95	0.028
M.2.3	1	1.5223	3.27%	0.0010	0.00095	0.22	0.640
M.2.4	1	0.8408	1.81%	0.0036	0.00363	0.84	0.362
M.2.5	1	0.6489	1.40%	0.0004	0.00045	0.10	0.748
M.3.3	1	1.2586	2.71%	0.0013	0.00133	0.31	0.581
M.3.4	1	0.8395	1.81%	0.0131	0.01313	3.04	0.085
M.3.5	1	0.3249	0.70%	0.0249	0.02489	5.76	0.018
M.4.4	1	0.0154	0.03%	0.0006	0.00057	0.13	0.718
M.4.5	1	0.3588	0.77%	0.0705	0.07049	16.32	0.000
M.5.5	1	0.1915	0.41%	0.0031	0.00306	0.71	0.402
M.1-2	1	0.0668	0.14%	0.0069	0.00686	1.59	0.211
M.1-3	1	0.0554	0.12%	0.0016	0.00164	0.38	0.540
M.1-4	1	0.8641	1.86%	0.0042	0.00417	0.97	0.328
M.1-5	1	0.4343	0.93%	0.4343	0.43428	100.53	0.000
Error	94	0.4061	0.87%	0.4061	0.00432		
Total	132	46.4937	100.00%				

Model Summary

S	R-sq	R-sq(adj)	PRESS	R-sq(pred)	AICc	BIC
0.0657259	99.13%	98.77%	1.07035	97.70%	-277.19	-197.23

Σχήμα 4.11: Διαγράμματα κανονικοποιημένων υπολοίπων της παλινδρόμησης του βαθμού αποφοίτησης με τις ορισμένες ανεξάρτητες μεταβλητές.

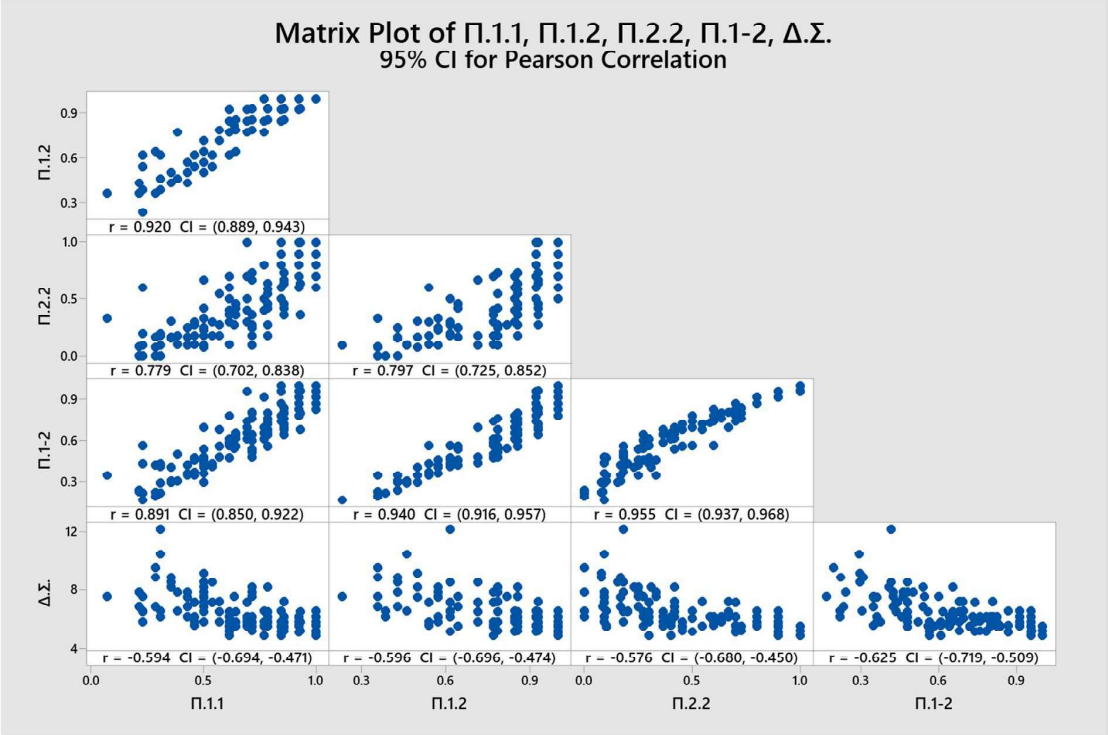


4.6.2 Έλεγχοι χρησιμότητας ανεξάρτητων μεταβλητών για πρόβλεψη της διάρκειας σπουδών

Συσχετίσεις

Όπως και στην υποενότητα 4.6.1 θα προχωρήσουμε στην αναπαράσταση των συσχετίσεων των ποσοστών περασμένων μαθημάτων για τα πρώτα 2 έτη φοίτησης, αλλά αντί για τον βαθμό αποφοίτησης αυτή τη φορά εξετάζεται η διάρκεια σπουδών. Οι συσχετίσεις με τη διάρκεια σπουδών προκύπτουν σημαντικά αρνητικές αλλά με μικρές αποκλίσεις μεταξύ τους, όπως και φαίνεται στο σχήμα 4.12.

Σχήμα 4.12: Διαγράμματα διασποράς και συσχετίσεις μεταξύ των ορισμένων ανεξάρτητων μεταβλητών των ποσοστών περασμένων μαθημάτων των πρώτων 2 ακαδημαϊκών ετών φοίτησης και της διάρκειας σπουδών



Επεκτείνοντας και πάλι τη διερεύνηση συσχετίσεων για όλες τις ορισμένες ανεξάρτητες μεταβλητές των ποσοστών περασμένων μαθημάτων, στον πίνακα 4.6 δίνονται οι συντελεστές συσχέτισης Pearson με τη διάρκεια σπουδών. Εκτός της συσχέτισης της διάρκειας σπουδών με τη μεταβλητή Π.1.4, όσο αυξάνονται τα έτη φοίτησης ν τόσο πιο αρνητικές είναι οι συσχετίσεις της διάρκειας σπουδών με τις μεταβλητές Π.χ.ν για το ίδιο έτος μαθημάτων χ. Ταυτόχρονα όσο αυξάνονται τα έτη φοίτησης τόσο πιο αρνητικές είναι οι συσχετίσεις με τις μεταβλητές Π.1-ν, φτάνοντας σε ισχυρά αρνητική συσχέτιση με τη μεταβλητή Π.1-5 έχοντας $\rho_{\Pi.1\text{-}5}$=-0.829.

Πίνακας 4.6: Συσχετίσεις μεταξύ των ορισμένων ανεξάρτητων μεταβλητών των ποσοστών περασμένων μαθημάτων της υποχρεωτικής φοίτησης και της διάρκειας σπουδών

Correlations

	Π.1.1	Π.1.2	Π.1.3	Π.1.4	Π.1.5	Π.2.2	Π.2.3	Π.2.4	Π.2.5
Π.1.2	0.920								
Π.1.3	0.846	0.922							
Π.1.4	0.774	0.859	0.945						
Π.1.5	0.671	0.746	0.833	0.892					
Π.2.2	0.779	0.797	0.736	0.649	0.563				
Π.2.3	0.719	0.774	0.750	0.673	0.642	0.895			
Π.2.4	0.666	0.730	0.775	0.734	0.694	0.805	0.914		
Π.2.5	0.585	0.631	0.715	0.716	0.759	0.627	0.749	0.856	
Π.3.3	0.693	0.706	0.688	0.643	0.608	0.791	0.778	0.739	0.634
Π.3.4	0.679	0.679	0.699	0.689	0.677	0.748	0.787	0.772	0.738
Π.3.5	0.590	0.616	0.703	0.734	0.741	0.566	0.646	0.699	0.795
Π.4.4	0.559	0.542	0.582	0.547	0.579	0.602	0.656	0.656	0.658
Π.4.5	0.536	0.553	0.636	0.651	0.697	0.567	0.667	0.710	0.772
Π.5.5	0.277	0.267	0.332	0.363	0.459	0.305	0.380	0.375	0.472
Π.1-2	0.891	0.940	0.870	0.789	0.687	0.955	0.886	0.815	0.669
Π.1-3	0.820	0.873	0.883	0.820	0.759	0.883	0.924	0.887	0.770
Π.1-4	0.750	0.784	0.837	0.823	0.796	0.788	0.855	0.892	0.843
Π.1-5	0.623	0.660	0.755	0.785	0.846	0.616	0.725	0.790	0.900
Διάρκεια σπουδών	-0.594	-0.596	-0.672	-0.652	-0.719	-0.576	-0.629	-0.658	-0.776

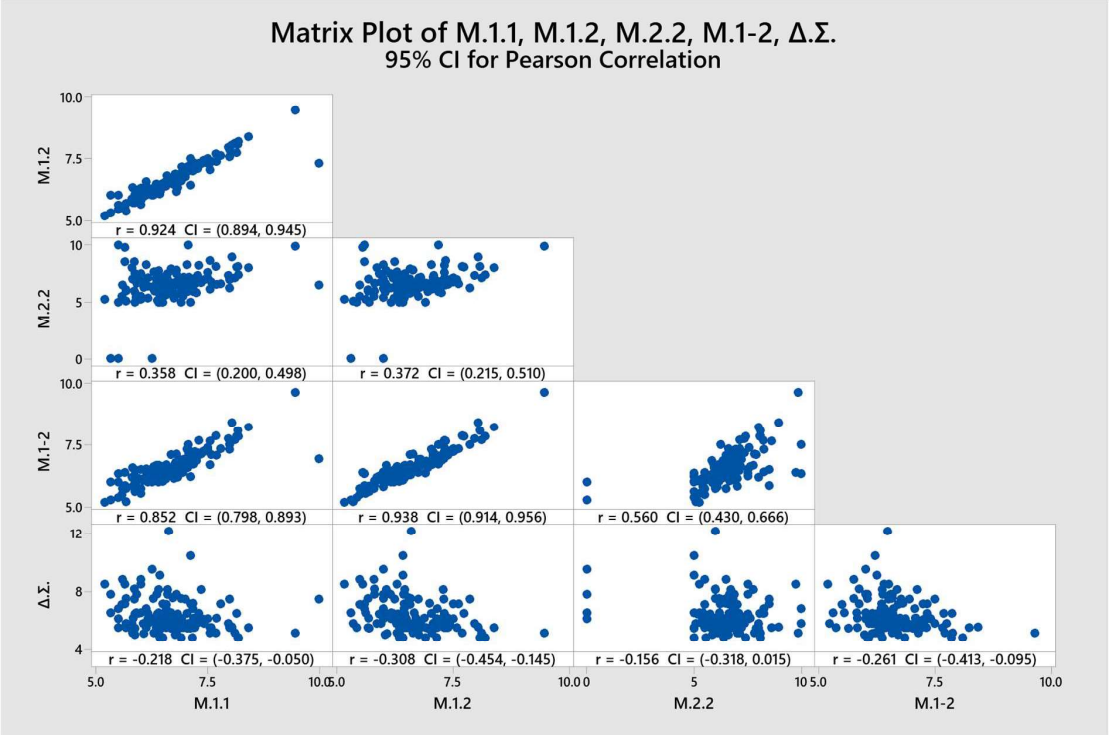
	Π.3.3	Π.3.4	Π.3.5	Π.4.4	Π.4.5	Π.5.5	Π.1-2	Π.1-3	Π.1-4
Π.1.2									
Π.1.3									
Π.1.4									
Π.1.5									
Π.2.2									
Π.2.3									
Π.2.4									
Π.2.5									
Π.3.3									
Π.3.4	0.869								
Π.3.5	0.700	0.841							
Π.4.4	0.731	0.763	0.668						
Π.4.5	0.678	0.766	0.778	0.805					
Π.5.5	0.391	0.486	0.514	0.505	0.658				
Π.1-2	0.793	0.758	0.627	0.609	0.596	0.305			
Π.1-3	0.914	0.870	0.759	0.723	0.725	0.405	0.928		
Π.1-4	0.850	0.914	0.836	0.872	0.843	0.497	0.832	0.933	
Π.1-5	0.711	0.826	0.904	0.758	0.924	0.691	0.677	0.807	0.900
Διάρκεια σπουδών	-0.607	-0.672	-0.752	-0.670	-0.757	-0.539	-0.625	-0.701	-0.759

	Π.1-5
Π.1.2	
Π.1.3	
Π.1.4	
Π.1.5	
Π.2.2	
Π.2.3	
Π.2.4	
Π.2.5	
Π.3.3	
Π.3.4	
Π.3.5	
Π.4.4	
Π.4.5	
Π.5.5	
Π.1-2	
Π.1-3	
Π.1-4	
Π.1-5	
Διάρκεια σπουδών	-0.829

Συνεχίζοντας τη διερεύνηση των συσχετίσεων της διάρκειας σπουδών, αυτή τη φορά με τις μεταβλητές των μέσων όρων των μαθημάτων, δίνονται τα διαγράμματα διασποράς των μεταβλητών με τη διάρκεια σπουδών για τα πρώτα 2 έτη φοίτησης στο σχήμα 4.13. Από τη μελέτη των διαγραμμάτων διασποράς παρατηρούνται οι εικόνες δημιουργίας «συννέφων» μεταξύ των σημείων, με την πιο χαρακτηριστική περίπτωση αυτής του διαγράμματος διασποράς με τη μεταβλητή Μ.2.2.

στο οποίο μπορούμε να παρατηρήσουμε και μηδενικές τιμές. Λόγω του παραπάνω κοντινότερη στο μηδέν τιμή συσχέτισης εντοπίζεται για τη συσχέτιση με τη μεταβλητή Μ.2.2, ενώ και οι συσχετίσεις με τις υπόλοιπες μεταβλητές δεν παρουσιάζουν σημαντικά αρνητικές τιμές.

Σχήμα 4.13: Διαγράμματα διασποράς και συσχετίσεις μεταξύ των ορισμένων ανεξάρτητων μεταβλητών των μέσων όρων μαθημάτων των πρώτων 2 ακαδημαϊκών ετών φοίτησης και της διάρκειας σπουδών



Χρησιμοποιώντας όλες τις ορισμένες ανεξάρτητες μεταβλητές των μέσων όρων των μαθημάτων, οι συσχετίσεις τους με τη διάρκεια σπουδών δίνονται στον πίνακα 4.7. Όλες οι συσχετίσεις προκύπτουν αρνητικές, χωρίς όμως να παρουσιάζουν σημαντικά αρνητικές τιμές ειδικά για την περίπτωση των συσχετίσεων με τις μεταβλητές Μ.2.ν του 2^{ου} έτους μαθημάτων. Οι πιο αρνητικές συσχετίσεις (ρ<-0.4) εντοπίζονται με τις μεταβλητές Μ.4.4, Μ.4.5, Μ.5.5 και Μ.1-5.

Πίνακας 4.7: Συσχετίσεις μεταξύ των ορισμένων ανεξάρτητων μεταβλητών των μέσων όρων μαθημάτων της υποχρεωτικής φοίτησης και της διάρκειας σπουδών

Correlations

	M.1.1	M.1.2	M.1.3	M.1.4	M.1.5	M.2.2	M.2.3	M.2.4	M.2.5
M.1.2	0.924								
M.1.3	0.923	0.984							
M.1.4	0.919	0.975	0.993						
M.1.5	0.901	0.963	0.980	0.986					
M.2.2	0.358	0.372	0.338	0.340	0.341				
M.2.3	0.381	0.438	0.426	0.424	0.438	0.700			
M.2.4	0.400	0.454	0.446	0.466	0.475	0.686	0.857		
M.2.5	0.474	0.538	0.554	0.571	0.585	0.555	0.713	0.858	
M.3.3	0.398	0.412	0.446	0.476	0.481	0.262	0.376	0.501	0.592
M.3.4	0.444	0.478	0.523	0.549	0.550	0.279	0.390	0.509	0.634
M.3.5	0.444	0.500	0.540	0.563	0.559	0.299	0.411	0.519	0.649
M.4.4	0.134	0.155	0.161	0.175	0.204	0.205	0.281	0.257	0.234
M.4.5	0.301	0.328	0.350	0.358	0.382	0.218	0.328	0.295	0.355
M.5.5	0.232	0.279	0.304	0.313	0.322	0.090	0.173	0.193	0.228
M.1-2	0.852	0.938	0.920	0.918	0.914	0.560	0.641	0.667	0.718
M.1-3	0.760	0.832	0.855	0.868	0.865	0.494	0.637	0.717	0.814
M.1-4	0.714	0.773	0.798	0.817	0.817	0.480	0.622	0.702	0.783
M.1-5	0.656	0.726	0.759	0.776	0.784	0.401	0.554	0.637	0.754
Διάρκεια σπουδών	-0.218	-0.308	-0.303	-0.309	-0.320	-0.156	-0.091	-0.057	-0.081

	M.3.3	M.3.4	M.3.5	M.4.4	M.4.5	M.5.5	M.1-2	M.1-3	M.1-4
M.1.2									
M.1.3									
M.1.4									
M.1.5									
M.2.2									
M.2.3									
M.2.4									
M.2.5									
M.3.3									
M.3.4	0.822								
M.3.5	0.799	0.979							
M.4.4	0.282	0.284	0.271						
M.4.5	0.459	0.560	0.568	0.483					
M.5.5	0.399	0.432	0.455	0.394	0.625				
M.1-2	0.475	0.524	0.549	0.211	0.374	0.269			
M.1-3	0.729	0.808	0.820	0.274	0.496	0.378	0.900		
M.1-4	0.738	0.845	0.855	0.371	0.609	0.409	0.839	0.964	
M.1-5	0.749	0.862	0.886	0.337	0.642	0.555	0.774	0.931	0.966
Διάρκεια σπουδών	-0.257	-0.345	-0.381	-0.458	-0.470	-0.668	-0.261	-0.298	-0.339

	M.1-5
M.1.2	
M.1.3	
M.1.4	
M.1.5	
M.2.2	
M.2.3	
M.2.4	
M.2.5	
M.3.3	
M.3.4	
M.3.5	
M.4.4	
M.4.5	
M.5.5	
M.1-2	
M.1-3	
M.1-4	
M.1-5	
Διάρκεια σπουδών	-0.427

Παλινδρόμηση

Το μοντέλο της πολλαπλής γραμμικής παλινδρόμησης της διάρκειας σπουδών με το σύνολο των ορισμένων ανεξάρτητων μεταβλητών για τα πρώτα 5 έτη φοίτησης δίνεται στον πίνακα αποτελεσμάτων 4.8. Στο συνολικό αυτό μοντέλο παλινδρόμησης, εξετάζοντας τον πίνακα ανάλυσης

μεταβλητότητας μπορούμε να διακρίνουμε πως οι μεταβλητές των περασμένων μαθημάτων φαίνεται να έχουν μεγαλύτερη χρησιμότητα για την πρόβλεψη της διάρκειας σπουδών από αυτές των μέσων όρων, με την μεταβλητή Μ.4.4 να είναι η μόνη που παρουσιάζει p-value χαμηλότερο του 0.05 από τις ανεξάρτητες μεταβλητές των μέσων όρων μαθημάτων. Ταυτόχρονα αν και ο R^2=85.44% είναι ικανοποιητικός, παρουσιάζεται ασυνέπεια στη σύγκριση των τιμών μεταξύ των υπολοίπων συντελεστών προσδιορισμού, με τον R-sq(pred)=62.05%, για τον οποίο θα αναφερθούμε περαιτέρω στο κεφάλαιο 5, να υποδηλώνει χαμηλή προβλεπτική ικανότητα του μοντέλου. Τέλος το τυπικό σφάλμα φαίνεται να είναι σχετικά υψηλό αν αναλογιστεί κανείς πως χρησιμοποιήθηκαν δεδομένα από όλα τα υποχρεωτικά χρόνια φοίτησης.

Η εικόνα των υπολοίπων μπορεί να αξιολογηθεί μέσω των διαγραμμάτων του σχήματος 4.14. Αρχικά μπορούμε να διακρίνουμε αρκετά μικρές αποκλίσεις των κανονικοποιημένων υπολοίπων από τη διαγώνιο ευθεία του normal probability plot. Συγχρόνως, εξετάζοντας το διάγραμμα διασποράς των κανονικοποιημένων υπολοίπων, η πλειοψηφία των σημείων παρατηρείται μεταξύ +/- 1 τυπικής απόκλισης με σχεδόν όλα τα σημεία να βρίσκονται μεταξύ +/- 3 τυπικές αποκλίσεις και με το ιστόγραμμα τους να παρουσιάζει την επιθυμητή εικόνα.

Πίνακας αποτελεσμάτων 4.8: Ανάλυση παλινδρόμησης διάρκειας σπουδών με τις ορισμένες ανεξάρτητες μεταβλητές για την υποχρεωτική διάρκεια φοίτησης

Regression Equation

Διάρκεια σπουδών = 12.88 - 0.279 Π.1.1 + 18.91 Π.1.2 + 13.46 Π.1.3 - 6.47 Π.1.4 - 8.68 Π.1.5
+ 13.76 Π.2.2 + 11.46 Π.2.3 - 4.53 Π.2.4 - 7.47 Π.2.5 + 11.55 Π.3.3
- 4.24 Π.3.4 - 5.58 Π.3.5 - 7.18 Π.4.4 - 6.96 Π.4.5 - 3.65 Π.5.5 - 32.3 Π.
1-2 - 38.7 Π.1-3 + 26.4 Π.1-4 + 25.4 Π.1-5 + 0.248 M.1.1 - 1.244 M.1.2
- 0.174 M.1.3 + 0.950 M.1.4 - 0.853 M.1.5 - 0.1295 M.2.2 + 0.072 M.2.3
- 0.349 M.2.4 + 0.231 M.2.5 + 0.075 M.3.3 - 0.306 M.3.4 + 0.044 M.3.5
- 0.0803 M.4.4 - 0.292 M.4.5 + 0.058 M.5.5 + 0.831 M.1-2 + 0.175 M.1-3
+ 1.585 M.1-4 - 1.111 M.1-5

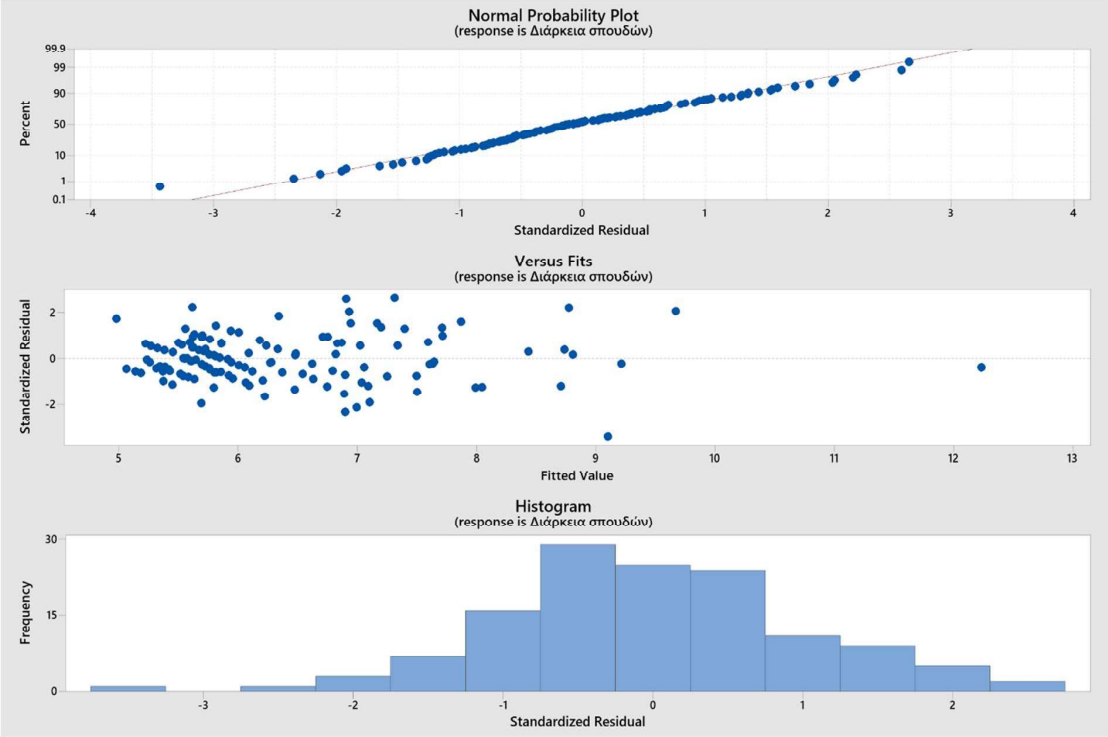
Analysis of Variance

Source	DF	Seq SS	Contribution	Adj SS	Adj MS	F-Value	P-Value
Regression	38	163.220	85.24%	163.220	4.29527	14.29	0.000
Π.1.1	1	67.484	35.24%	0.050	0.04986	0.17	0.685
Π.1.2	1	3.086	1.61%	1.619	1.61858	5.38	0.022
Π.1.3	1	19.362	10.11%	1.423	1.42276	4.73	0.032
Π.1.4	1	0.695	0.36%	0.274	0.27444	0.91	0.342
Π.1.5	1	18.764	9.80%	0.804	0.80448	2.68	0.105
Π.2.2	1	4.573	2.39%	1.353	1.35335	4.50	0.036
Π.2.3	1	1.153	0.60%	1.756	1.75625	5.84	0.018
Π.2.4	1	0.727	0.38%	0.221	0.22103	0.74	0.393
Π.2.5	1	15.372	8.03%	0.820	0.82046	2.73	0.102
Π.3.3	1	0.926	0.48%	1.650	1.65012	5.49	0.021
Π.3.4	1	0.103	0.05%	0.187	0.18729	0.62	0.432
Π.3.5	1	3.951	2.06%	0.473	0.47322	1.57	0.213
Π.4.4	1	2.921	1.53%	0.390	0.38968	1.30	0.258
Π.4.5	1	2.230	1.16%	0.539	0.53946	1.79	0.184
Π.5.5	1	0.628	0.33%	0.407	0.40707	1.35	0.248
Π.1-2	1	4.911	2.56%	1.464	1.46401	4.87	0.030
Π.1-3	1	0.011	0.01%	1.806	1.80603	6.01	0.016
Π.1-4	1	2.604	1.36%	0.371	0.37144	1.24	0.269
Π.1-5	1	1.338	0.70%	0.389	0.38881	1.29	0.258
M.1.1	1	0.022	0.01%	0.389	0.38855	1.29	0.259
M.1.2	1	0.213	0.11%	1.133	1.13328	3.77	0.055
M.1.3	1	0.001	0.00%	0.011	0.01145	0.04	0.846
M.1.4	1	0.047	0.02%	0.296	0.29648	0.99	0.323
M.1.5	1	2.152	1.12%	0.737	0.73712	2.45	0.121
M.2.2	1	0.365	0.19%	0.994	0.99441	3.31	0.072
M.2.3	1	0.112	0.06%	0.074	0.07438	0.25	0.620
M.2.4	1	0.150	0.08%	0.813	0.81298	2.70	0.103
M.2.5	1	0.717	0.37%	0.344	0.34373	1.14	0.288
M.3.3	1	0.012	0.01%	0.132	0.13196	0.44	0.509
M.3.4	1	0.949	0.50%	0.171	0.17067	0.57	0.453
M.3.5	1	0.021	0.01%	0.003	0.00318	0.01	0.918
M.4.4	1	1.209	0.63%	0.596	0.59560	1.98	0.163
M.4.5	1	2.725	1.42%	2.050	2.05041	6.82	0.010
M.5.5	1	0.788	0.41%	0.069	0.06945	0.23	0.632
M.1-2	1	1.743	0.91%	0.692	0.69178	2.30	0.133
M.1-3	1	0.160	0.08%	0.018	0.01788	0.06	0.808
M.1-4	1	0.426	0.22%	0.991	0.99064	3.30	0.073
M.1-5	1	0.566	0.30%	0.566	0.56614	1.88	0.173
Error	94	28.261	14.76%	28.261	0.30064		
Total	132	191.481	100.00%				

Model Summary

S	R-sq	R-sq(adj)	PRESS	R-sq(pred)	AICc	BIC
0.548311	85.24%	79.27%	72.6714	62.05%	287.09	367.05

Σχήμα 4.14: Διαγράμματα κανονικοποιημένων υπολοίπων της παλινδρόμησης της διάρκειας σπουδών με τις ορισμένες ανεξάρτητες μεταβλητές.



Κεφάλαιο 5: Μοντέλα πρόβλεψης

Στο πλαίσιο του κεφαλαίου αυτού θα ασχοληθούμε με τον προσδιορισμό των κατάλληλων μεταβλητών εισόδου για τις προβλέψεις του βαθμού αποφοίτησης και της διάρκειας σπουδών, χρησιμοποιώντας δεδομένα από την κοινή διάρκεια φοίτησης, δηλαδή τα πρώτα πέντε ακαδημαϊκά έτη. Υπενθυμίζεται πως στο κεφάλαιο 4 έγινε ανάλυση σχέσεων εξάρτησης με απώτερο σκοπό τον ορισμό χρήσιμων ανεξάρτητων μεταβλητών για τις προβλέψεις του βαθμού αποφοίτησης και της διάρκειας σπουδών. Οδηγηθήκαμε στο συμπέρασμα πως τα ποσοστά περασμένων μαθημάτων των αποφοίτων και οι μέσοι όροι τους φαίνεται να χρησιμεύουν για τις εν λόγω προβλέψεις με τη χρήση της πολλαπλής γραμμικής παλινδρόμησης.

Εφόσον θα χρησιμοποιηθούν δεδομένα των αποφοίτων από τα πρώτα πέντε ακαδημαϊκά έτη, θα χωρίσουμε τα δεδομένα σε 5 σετ δεδομένων, ένα για κάθε έτος φοίτησης. Ειδικά για το 5° έτος φοίτησης εξετάστηκε και το αν οι απόφοιτοι είχαν ολοκληρώσει και παραδώσει τη διπλωματική τους εργασία, με 69 από τους 133 αποφοίτους να είχαν ολοκληρώσει τις διπλωματικές τους. Η εν λόγω μεταβλητή παραλήφθηκε επειδή δεν βελτίωσε κανένα μοντέλο πρόβλεψης. Αν για κάθε έτος εξεταστούν οι μεταβλητές εισόδου:

Π.Π.χ.ν, Μ.Μ.χ.ν: αθροιστικά ποσοστά περασμένων μαθημάτων και μέσοι όροι αντίστοιχα, του χ έτους μαθημάτων στο ν έτος φοίτησης, με ν ≥ χ και ν_{max}=εξεταζόμενο ακαδημαϊκό έτος,

Π.Π.1-ν, Μ.Μ.1-ν: συνολικά ποσοστά περασμένων μαθημάτων και μέσοι όροι αντίστοιχα, μέχρι το ν έτος φοίτησης,

τότε οι εξεταζόμενες ανεξάρτητες ματαβλητές των 5 σετ δεδομένων θα είναι οι εξής:

Πίνακας 5.1: Κωδικές ονομασίες των ορισμένων ανεξάρτητων μεταβλητών για κάθε έτος φοίτησης

1ο έτος		2ο έτος		3ο έτος		4ο έτος		5ο έτος	
Π.1.1	Π.Π. 1ου έτους στο 1ο έτος	Π.1.1	Π.Π. 1ου έτους στο 1ο έτος	Π.1.1	Π.Π. 1ου έτους στο 1ο έτος	Π.1.1	Π.Π. 1ου έτους στο 1ο έτος	Π.1.1	Π.Π. 1ου έτους στο 1ο έτος
Μ.1.1	Μ.Ο. 1ου έτους στο 1ο έτος	Π.1.2	Π.Π. 1ου έτους στο 2ο έτος	Π.1.2	Π.Π. 1ου έτους στο 2ο έτος	Π.1.2	Π.Π. 1ου έτους στο 2ο έτος	Π.1.2	Π.Π. 1ου έτους στο 2ο έτος
		Π.2.2	Π.Π. 2ου έτους στο 2ο έτος	Π.1.3	Π.Π. 1ου έτους στο 3ο έτος	Π.1.3	Π.Π. 1ου έτους στο 3ο έτος	Π.1.3	Π.Π. 1ου έτους στο 3ο έτος
		Π.1-2	Π.Π. 1-2 Ακαδημαϊκά έτη	Π.2.2	Π.Π. 2ου έτους στο 2ο έτος	Π.1.4	Π.Π. 1ου έτους στο 4ο έτος	Π.1.4	Π.Π. 1ου έτους στο 4ο έτος
		Μ.1.1	Μ.Ο. 1ου έτους στο 1ο έτος	Π.2.3	Π.Π. 2ου έτους στο 3ο έτος	Π.2.2	Π.Π. 2ου έτους στο 2ο έτος	Π.1.5	Π.Π. 1ου έτους στο 5ο έτος
		Μ.1.2	Μ.Ο. 1ου έτους στο 2ο έτος	Π.3.3	Π.Π. 3ου έτους στο 3ο έτος	Π.2.3	Π.Π. 2ου έτους στο 3ο έτος	Π.2.2	Π.Π. 2ου έτους στο 2ο έτος
		Μ.2.2	Μ.Ο. 2ου έτους στο 2ο έτος	Π.1-2	Π.Π. 1-2 Ακαδημαϊκά έτη	Π.2.4	Π.Π. 2ου έτους στο 4ο έτος	Π.2.3	Π.Π. 2ου έτους στο 3ο έτος
		Μ.1-2	Μ.Ο. 1-2 Ακαδημαϊκά έτη	Π.1-3	Π.Π. 1-3 Ακαδημαϊκά έτη	Π.3.3	Π.Π. 3ου έτους στο 3ο έτος	Π.2.4	Π.Π. 2ου έτους στο 4ο έτος
				Μ.1.1	Μ.Ο. 1ου έτους στο 1ο έτος	Π.3.4	Π.Π. 3ου έτους στο 4ο έτος	Π.2.5	Π.Π. 2ου έτους στο 5ο έτος
				Μ.1.2	Μ.Ο. 1ου έτους στο 2ο έτος	Π.4.4	Π.Π. 4ου έτους στο 4ο έτος	Π.3.3	Π.Π. 3ου έτους στο 3ο έτος
				Μ.1.3	Μ.Ο. 1ου έτους στο 3ο έτος	Π.1-2	Π.Π. 1-2 Ακαδημαϊκά έτη	Π.3.4	Π.Π. 3ου έτους στο 4ο έτος
				Μ.2.2	Μ.Ο. 2ου έτους στο 2ο έτος	Π.1-3	Π.Π. 1-3 Ακαδημαϊκά έτη	Π.3.5	Π.Π. 3ου έτους στο 5ο έτος
				Μ.2.3	Μ.Ο. 2ου έτους στο 3ο έτος	Π.1-4	Π.Π. 1-4 Ακαδημαϊκά έτη	Π.4.4	Π.Π. 4ου έτους στο 4ο έτος
				Μ.3.3	Μ.Ο. 3ου έτους στο 3ο έτος	Μ.1.1	Μ.Ο. 1ου έτους στο 1ο έτος	Π.4.5	Π.Π. 4ου έτους στο 5ο έτος
				Μ.1-2	Μ.Ο. 1-2 Ακαδημαϊκά έτη	Μ.1.2	Μ.Ο. 1ου έτους στο 2ο έτος	Π.5.5	Π.Π. 5ου έτους στο 5ο έτος
				Μ.1-3	Μ.Ο. 1-3 Ακαδημαϊκά έτη	Μ.1.3	Μ.Ο. 1ου έτους στο 3ο έτος	Π.1-2	Π.Π. 1-2 Ακαδημαϊκά έτη
						Μ.1.4	Μ.Ο. 1ου έτους στο 4ο έτος	Π.1-3	Π.Π. 1-3 Ακαδημαϊκά έτη
						Μ.2.2	Μ.Ο. 2ου έτους στο 2ο έτος	Π.1-4	Π.Π. 1-4 Ακαδημαϊκά έτη
						Μ.2.3	Μ.Ο. 2ου έτους στο 3ο έτος	Π.1-5	Π.Π. 1-5 Ακαδημαϊκά έτη
						Μ.2.4	Μ.Ο. 2ου έτους στο 4ο έτος	Μ.1.1	Μ.Ο. 1ου έτους στο 1ο έτος
						Μ.3.3	Μ.Ο. 3ου έτους στο 3ο έτος	Μ.1.2	Μ.Ο. 1ου έτους στο 2ο έτος
						Μ.3.4	Μ.Ο. 3ου έτους στο 4ο έτος	Μ.1.3	Μ.Ο. 1ου έτους στο 3ο έτος
						Μ.4.4	Μ.Ο. 4ου έτους στο 4ο έτος	Μ.1.4	Μ.Ο. 1ου έτους στο 4ο έτος
						Μ.1-2	Μ.Ο. 1-2 Ακαδημαϊκά έτη	Μ.1.5	Μ.Ο. 1ου έτους στο 5ο έτος
						Μ.1-3	Μ.Ο. 1-3 Ακαδημαϊκά έτη	Μ.2.2	Μ.Ο. 2ου έτους στο 2ο έτος
						Μ.1-4	Μ.Ο. 1-4 Ακαδημαϊκά έτη	Μ.2.3	Μ.Ο. 2ου έτους στο 3ο έτος
								Μ.2.4	Μ.Ο. 2ου έτους στο 4ο έτος
								Μ.2.5	Μ.Ο. 2ου έτους στο 5ο έτος
								Μ.3.3	Μ.Ο. 3ου έτους στο 3ο έτος
								Μ.3.4	Μ.Ο. 3ου έτους στο 4ο έτος
								Μ.3.5	Μ.Ο. 3ου έτους στο 5ο έτος
								Μ.4.4	Μ.Ο. 4ου έτους στο 4ο έτος
								Μ.4.5	Μ.Ο. 4ου έτους στο 5ο έτος
								Μ.5.5	Μ.Ο. 5ου έτους στο 5ο έτος
								Μ.1-2	Μ.Ο. 1-2 Ακαδημαϊκά έτη
								Μ.1-3	Μ.Ο. 1-3 Ακαδημαϊκά έτη
								Μ.1-4	Μ.Ο. 1-4 Ακαδημαϊκά έτη
								Μ.1-5	Μ.Ο. 1-5 Ακαδημαϊκά έτη

Να σημειωθεί πως τις ενδεχόμενες μεταβλητές εισόδου τις ορίσαμε ως εξεταζόμενες και κάτι τέτοιο ισχύει καθώς δεν είναι εύκολο όλες οι μεταβλητές να επηρεάζουν την εκάστοτε έξοδο, ειδικά στην περίπτωση των μεγαλύτερων ακαδημαϊκών ετών. Το γεγονός αυτό μας οδηγεί στον προσδιορισμό κατάλληλων υποσυνόλων μεταβλητών εισόδου. Ταυτόχρονα καθώς οι εξεταζόμενες μεταβλητές αποτελούν όλο και μεγαλύτερο μέρος των μεταβλητών εξόδου, όσο αυξάνονται τα έτη φοίτησης και ειδικά στην περίπτωση της πρόβλεψης του βαθμού αποφοίτησης, το πρόβλημα της πρόβλεψης γίνεται σταδιακά σχεδόν ντετερμινιστικό και η γραμμική παλινδρόμηση ως βέλτιστη μέθοδος πρόβλεψης. Στο πλαίσιο της εργασίας, εξετάζονται ξεχωριστά όλες οι μεταβλητές και αυτές που σχετίζονται με τα συνολικά δεδομένα μέχρι το ν έτος φοίτησης, δηλαδή μόνο οι μεταβλητές Π.Π.1-ν, Μ.Μ.1-ν. Για τις περιπτώσεις αυτές θα γίνει μεταξύ τους σύγκριση για κάθε σετ δεδομένων, ώστε να εξεταστεί αν μεγαλύτερο πλήθος μεταβλητών εισόδου οδηγεί σε σημαντικά καλύτερες προβλέψεις ώστε να δικαιολογεί τον αριθμό τους.

Αφότου γίνει ο κατάλληλος προσδιορισμός υποσυνόλων των ανεξάρτητων μεταβλητών, με τη χρήση των υποσυνόλων αυτών θα γίνει πρόβλεψη του βαθμού αποφοίτησης και της διάρκειας σπουδών για κάθε σετ δεδομένων με τη χρήση της πολλαπλής γραμμικής παλινδρόμησης. Υπάρχει όμως ο κίνδυνος τα μοντέλα πρόβλεψης τα οποία θα προκύψουν να έχουν προσαρμοστεί υπερβολικά σύμφωνα με τα δεδομένα εξόδου (overfitting). Στη συγκεκριμένη περίπτωση τα αποτελέσματα των προβλέψεων δεν θα ήταν τα επιθυμητά, καθώς σκοπός δεν είναι η επακριβής πρόβλεψη των δεδομένων των αποφοίτων, τα οποία είναι ήδη γνωστά, αλλά η εύρεση κατάλληλων μοντέλων πρόβλεψης για εν ενεργεία φοιτητές. Σύμφωνα με το παραπάνω θα πρέπει να γίνουν ανάλογοι έλεγχοι υπερβολικής προσαρμογής, αλλά και το να γίνει ακριβής προσέγγιση του σφάλματος πρόβλεψης για δεδομένα εκτός των υπαρχόντων, γιατί δεν χρησιμοποιήθηκε το σύνολο των δεδομένων των έως τώρα αποφοίτων του τμήματος ώστε να θεωρείται αυτό γνωστό.

Άσχετα αν χρησιμοποιήθηκαν δεδομένα για τις προβλέψεις από τα πρώτα 5 ακαδημαϊκά έτη, μεγαλύτερο ενδιαφέρον παρουσιάζουν οι προβλέψεις με τη χρήση δεδομένων των χαμηλότερων ετών. Έτσι ισχυρότερο κριτήριο επάρκειας των μοντέλων πρόβλεψης αποτελεί το κατά πόσο γίνονται ικανοποιητικές προβλέψεις στα χαμηλότερα έτη φοίτησης.

5.1 Μέθοδοι βηματικής παλινδρόμησης

Όπως αναφέρθηκε στην αρχή του κεφαλαίου, δεν είναι εύκολο όλες οι εξεταζόμενες μεταβλητές εισόδου να είναι χρήσιμες για τις επιθυμητές προβλέψεις, έτσι με την προσθήκη περιττών μεταβλητών το μοντέλο περιπλέκεται άσκοπα, οδηγώντας σε χειρότερες προβλέψεις. Οι μέθοδοι της βηματικής παλινδρόμησης μεταξύ άλλων, χρησιμοποιούνται για τον εντοπισμό χρήσιμων μεταβλητών εισόδου. Από ανάλυση εκτός των αποτελεσμάτων της εργασίας και σύγκρισης των μεθόδων της βηματικής παλινδρόμησης, δηλαδή της προς τα πίσω απαλοιφής, της προς τα εμπρός επιλογής και της αμφίδρομης βηματικής παλινδρόμησης, αλλά και σύγκριση με τον αλγόριθμο παλινδρόμησης Lasso, προέκυψε πως για τα δεδομένα της εργασίας καλύτερες προβλέψεις επιτυγχάνονται ως επί το πλείστον με τη μέθοδο της προς τα πίσω απαλοιφής. Σκοπός της εργασίας αυτής είναι η πρόταση μίας κοινής μεθοδολογίας προβλέψεων, ανεξαρτήτως έτους φοίτησης, με αποτέλεσμα τη χρήση μόνο της μεθόδου της προς τα πίσω απαλοιφής.

Αναλυτικά, όσον αφορά τη μέθοδο της προς τα πίσω απαλοιφής, αυτή ξεκινάει τον έλεγχο χρησιμότητας των μεταβλητών χρησιμοποιώντας το συνολικό πλήθος k ανεξάρτητων μεταβλητών,

άρα και με το συνολικό μοντέλο ελαχίστων τετραγώνων. Στη συνέχεια αφαιρεί τη λιγότερο χρήσιμη μεταβλητή, μία κάθε φορά σύμφωνα με τον αλγόριθμο [4]:

Αλγόριθμος προς τα πίσω απαλοιφής

1) Συνολικό μοντέλο M, το οποίο συμπεριλαμβάνει το συνολικό πλήθος παραγόντων k.

2) Για p=k,k-1,…,1:

2α) Έλεγχος των p μοντέλων τα οποία συμπεριλαμβάνουν όλες εκτός μίας ανεξάρτητης μεταβλητής του συνολικού μοντέλου M, για το σύνολο p-1 ανεξάρτητων μεταβλητών.

2β) Επιλογή βέλτιστου p μοντέλου, το οποίο παρουσιάζει το χαμηλότερο RSS ή τον μεγαλύτερο R^2, ονομάζοντάς το M_{p-1}.

3) Επιλογή ενός βέλτιστου μοντέλου ανάμεσα στα $M_0,…,M_k$.

Για την επιλογή του βέλτιστου μοντέλου χρησιμοποιείται η σύγκριση των συντελεστών AIC, BIC είτε μέθοδοι διασταυρούμενης επικύρωσης για τους/τις οποίους/ες θα αναφερθούμε στην υποενότητα 5.1.1 και υποκεφάλαιο 5.2 αντίστοιχα και τέλος μπορεί να χρησιμοποιηθεί και η σύγκριση του προσαρμοσμένου R^2.

5.1.1 Καταλληλότητα μοντέλων βηματικής παλινδρόμησης

Για κάθε μία από τις μεθόδους βηματικής παλινδρόμησης πρέπει να υπάρξει ένα κριτήριο επιλογής μοντέλου με τον βέλτιστο αριθμό ανεξάρτητων μεταβλητών. Σκοπός είναι η ελαχιστοποίηση του test error, δηλαδή του σφάλματος άγνωστων δεδομένων κατά τη διάρκεια εκπαίδευσης του εκάστοτε μοντέλου και όχι απαραίτητα η καλύτερη προσαρμογή στα δεδομένα της παλινδρόμησης. Αυτό μπορεί να επιτευχθεί με τη χρήση των κριτηρίων AIC (Akaike information criterion) και BIC (Bayesian information criterion), του προσαρμοσμένου συντελεστή προσδιορισμού R^2, του οποίου ανάλυση έγινε στην υποενότητα 4.1.2 και των R^2(LOOCV), PRESS.

Κριτήριο AIC

Το κριτήριο AIC, το οποίο αναπτύχθηκε από τον Hirotugu Akaike και δημοσιεύτηκε το 1973, χρησιμοποιείται ως εκτίμηση του σφάλματος πρόβλεψης, αξιολογώντας διαφορετικά μοντέλα με τη μεταξύ τους σύγκριση, λαμβάνοντας υπόψη τα ρίσκα του overfitting και underfitting ταυτοχρόνως. Η αξιολόγηση αυτή επιτυγχάνεται με την επιλογή του μοντέλου που ελαχιστοποιεί το AIC. Η τιμή του AIC, για την πολλαπλή γραμμική παλινδρόμηση και με τη χρήση της αρχής της μέγιστης πιθανοφάνειας, υπολογίζεται ως εξής:

$$AIC = \frac{1}{n\hat{\sigma}^2}(RSS + 2d\hat{\sigma}^2) \quad (5.1)$$

Για μικρές τιμές των παρατηρήσεων n, ο AIC τείνει να επιλέγει μοντέλα με μεγάλο αριθμό ανεξάρτητων μεταβλητών και σε αυτές τις περιπτώσεις ενδείκνυται ο μετασχηματισμός του [5] σε:

$AIC_c = AIC + \frac{2k^2 + 2k}{n - k - 1}$ (5.2) με:

$\frac{2k^2 + 2k}{n - k - 1}$: τον όρο να αποτελεί την ποινή του αριθμού παραμέτρων. Για n→∞ ο όρος πρακτικά μηδενίζεται και ο $AIC_c = AIC$.

Κριτήριο BIC

Το κριτήριο BIC, το οποίο αναπτύχθηκε από τον Gideon E. Schwarz και δημοσιεύτηκε σε άρθρο το 1978, αποτελεί κι αυτό μέθοδο επιλογής μοντέλου με τον βέλτιστο αριθμό ανεξάρτητων μεταβλητών. Όπως και στο κριτήριο AIC, έτσι και εδώ το σκοπός αποτελεί η ελαχιστοποίηση του BIC, ούτως ώστε να επιλέξουμε το μοντέλο με το μικρότερο test error. Η τιμή του BIC, για την πολλαπλή γραμμική παλινδρόμηση και με τη χρήση της αρχής της μέγιστης πιθανοφάνειας, υπολογίζεται ως εξής:

$$BIC = \frac{1}{n\hat{\sigma}^2}(RSS + \log(n) d \hat{\sigma}^2) \quad (5.3)$$

R^2_{LOOCV} και PRESS

Μία ακόμα αναγκαία σύγκριση είναι αυτή του συντελεστή προσδιορισμού R^2, με τον συντελεστή R^2 ο οποίος προκύπτει μετά από συστηματική αφαίρεση μίας παρατήρησης. Η μέθοδος η οποία χρησιμοποιείται αποτελεί ειδική περίπτωση της μεθόδου της διασταυρούμενης επικύρωσης τμημάτων, για την οποία θα αναφερθούμε στο υποκεφάλαιο 5.2. Αναλυτικότερα η μέθοδος αυτή ονομάζεται LOOCV (Leave One Out Cross Validation), και το k της διασταυρούμενης επικύρωσης τμημάτων ισούται με τον αριθμό των παρατηρήσεων. Ο λόγος για τον οποίο αποτελεί σημαντική η σύγκριση του R^2 με τον R^2_{LOOCV}, είναι για την αποφυγή του overfitting. Ως overfitting ορίζουμε την υπερβολική προσαρμογή του μοντέλου στα δεδομένα εξόδου, έχοντας ως αποτέλεσμα την εύρεση προτύπων η οποία δεν συμπίπτει με το σύνολο δεδομένων. Όσον αφορά τη σύγκριση των συντελεστών, αν ο R^2_{LOOCV} προκύψει σημαντικά μικρότερος του R^2 σημαίνει πως υπήρξε overfitting.

Παρόμοια λογική ακολουθεί και ο όρος PRESS, για τον οποίο υπολογίζεται με LOOCV, δηλαδή παραλείποντας συστηματικά μία παρατήρηση, υπολογίζοντας το άθροισμα των τετραγώνων των σφαλμάτων πρόβλεψης ως εξής:

$$PRESS = \sum_{i=1}^{n} (y_i - \hat{y}_{i-1})^2 \quad (5.4)$$

Σε σύγκριση μοντέλων, καταλληλότερο μοντέλο είναι αυτό με το μικρότερο PRESS, όπως και στην περίπτωση του RSS.

5.1.2 Αποτελέσματα της προς τα πίσω απαλοιφής

Τα αποτελέσματα της προς τα πίσω απαλοιφής προήλθαν από λογισμικό του οποίου ορίστηκε μέγιστο p-value=0.1, για το οποίο κάθε ανεξάρτητη μεταβλητή θεωρείται χρήσιμη. Η ανεξάρτητη μεταβλητή με το υψηλότερο p-value για κάθε μοντέλο ελαχίστων τετραγώνων, αφαιρείται και ξαναγίνεται έλεγχος μέχρι όλα τα p-value των ανεξάρτητων μεταβλητών να είναι μικρότερα του ορίου

που ορίσαμε. Με τον τρόπο αυτό επιτυγχάνεται σε ικανοποιητικό βαθμό η μεγιστοποίηση του προσαρμοσμένου R^2.

Οι μεταβλητές οι οποίες ορίστηκαν σημαντικές για κάθε έτος δίνονται στους πίνακες 5.2 και 5.3 για την πρόβλεψη του βαθμού αποφοίτησης και της διάρκειας σπουδών αντίστοιχα, ενώ στους πίνακες 5.4 και 5.5 δίνονται χρήσιμοι συντελεστές για την αξιολόγηση των αποτελεσμάτων της μεθόδου. Όπως φαίνεται στους πίνακες έγινε έλεγχος όλων των μεταβλητών και αυτών που σχετίζονται με τα συνολικά δεδομένα μέχρι το ν έτος φοίτησης ξεχωριστά και θα προχωρήσουμε στη σύγκριση των αποτελεσμάτων τους.

Πίνακας 5.2: Επιλογή υποσυνόλου μεταβλητών εισόδου για την πρόβλεψη του βαθμού αποφοίτησης με τη μέθοδο της προς τα πίσω απαλοιφής

Επιλογή υποσυνόλου μεταβλητών εισόδου για πρόβλεψη του βαθμού αποφοίτησης (1 αν επιλέχθηκε, 0 αν όχι)													
1ο έτος		2ο έτος			3ο έτος			4ο έτος			5ο έτος		
Μεταβλητές													
	Όλες		Όλες	Συνολικών επιδόσεων		Όλες	Συνολικών επιδόσεων		Όλες	Συνολικών επιδόσεων		Όλες	Συνολικών επιδόσεων
Π.1.1	1	Π.1.1	1	1	Π.1.1	0	0	Π.1.1	0	0	Π.1.1	1	1
Μ.1.1	1	Π.1.2	1	-	Π.1.2	1	-	Π.1.2	1	-	Π.1.2	0	-
Σύνολο	2	Π.2.2	1	-	Π.1.3	0	-	Π.1.3	0	-	Π.1.3	0	-
		Π.1-2	1	1	Π.2.2	1	-	Π.1.4	0	-	Π.1.4	1	-
		Μ.1.1	0	0	Π.2.3	0	-	Π.2.2	1	-	Π.1.5	1	-
		Μ.1.2	1	-	Π.3.3	0	-	Π.2.3	0	-	Π.2.2	0	-
		Μ.2.2	1	-	Π.1-2	1	0	Π.2.4	0	-	Π.2.3	0	-
		Μ.1-2	1	1	Π.1-3	1	1	Π.3.3	0	-	Π.2.4	1	-
		Σύνολο	7	3	Μ.1.1	0	0	Π.3.4	1	-	Π.2.5	1	-
					Μ.1.2	0	-	Π.4.4	1	-	Π.3.3	0	-
					Μ.1.3	0	-	Π.1-2	1	0	Π.3.4	1	-
					Μ.2.2	1	-	Π.1-3	0	0	Π.3.5	1	-
					Μ.2.3	0	-	Π.1-4	1	1	Π.4.4	1	-
					Μ.3.3	0	-	Μ.1.1	1	0	Π.4.5	1	-
					Μ.1-2	1	0	Μ.1.2	0	-	Π.5.5	1	-
					Μ.1-3	1	1	Μ.1.3	0	-	Π.1-2	0	1
					Σύνολο	7	2	Μ.1.4	0	-	Π.1-3	0	0
								Μ.2.2	1	-	Π.1-4	1	1
								Μ.2.3	0	-	Π.1-5	1	1
								Μ.2.4	0	-	Μ.1.1	0	0
								Μ.3.3	0	-	Μ.1.2	0	-
								Μ.3.4	0	-	Μ.1.3	0	-
								Μ.4.4	1	-	Μ.1.4	1	-
								Μ.1-2	0	1	Μ.1.5	1	-
								Μ.1-3	0	1	Μ.2.2	1	-
								Μ.1-4	1	1	Μ.2.3	0	-
								Σύνολο	10	4	Μ.2.4	0	-
											Μ.2.5	0	-
											Μ.3.3	0	-
											Μ.3.4	1	-
											Μ.3.5	1	-
											Μ.4.4	0	-
											Μ.4.5	1	-
											Μ.5.5	1	-
											Μ.1-2	0	1
											Μ.1-3	0	0
											Μ.1-4	0	0
											Μ.1-5	1	1
											Σύνολο	20	6

Πίνακας 5.3: Επιλογή υποσυνόλου μεταβλητών εισόδου για την πρόβλεψη της διάρκειας σπουδών με τη μέθοδο της προς τα πίσω απαλοιφής

Επιλογή υποσυνόλου μεταβλητών εισόδου για πρόβλεψη της διάρκειας σπουδών (1 αν επιλέχθηκε, 0 αν όχι)													
1ο έτος		2ο έτος			3ο έτος			4ο έτος			5ο έτος		
Μεταβλητές													
	Όλες		Όλες	Συνολικών επιδόσεων		Όλες	Συνολικών επιδόσεων		Όλες	Συνολικών επιδόσεων		Όλες	Συνολικών επιδόσεων
Π.1.1	1	Π.1.1	0	0	Π.1.1	0	0	Π.1.1	0	0	Π.1.1	0	0
Μ.1.1	0	Π.1.2	1	-	Π.1.2	1	-	Π.1.2	1	-	Π.1.2	1	-
Σύνολο	1	Π.2.2	1	-	Π.1.3	1	-	Π.1.3	1	-	Π.1.3	1	-
		Π.1-2	1	1	Π.2.2	1	-	Π.1.4	1	-	Π.1.4	0	-
		Μ.1.1	0	0	Π.2.3	0	-	Π.2.2	1	-	Π.1.5	1	-
		Μ.1.2	0	-	Π.3.3	0	-	Π.2.3	1	-	Π.2.2	1	-
		Μ.2.2	0	-	Π.1-2	1	0	Π.2.4	1	-	Π.2.3	1	-
		Μ.1-2	0	0	Π.1-3	1	1	Π.3.3	1	-	Π.2.4	0	-
		Σύνολο	3	1	Μ.1.1	0	0	Π.3.4	1	-	Π.2.5	1	-
					Μ.1.2	1	-	Π.4.4	1	-	Π.3.3	1	-
					Μ.1.3	1	-	Π.1-2	1	0	Π.3.4	0	-
					Μ.2.2	0	-	Π.1-3	1	0	Π.3.5	1	-
					Μ.2.3	0	-	Π.1-4	1	1	Π.4.4	1	-
					Μ.3.3	0	-	Μ.1.1	0	0	Π.4.5	1	-
					Μ.1-2	0	0	Μ.1.2	0	-	Π.5.5	1	-
					Μ.1-3	0	1	Μ.1.3	0	-	Π.1-2	1	0
					Σύνολο	7	2	Μ.1.4	0	-	Π.1-3	1	0
								Μ.2.2	0	-	Π.1-4	1	0
								Μ.2.3	0	-	Π.1-5	1	1
								Μ.2.4	0	-	Μ.1.1	1	0
								Μ.3.3	0	-	Μ.1.2	1	-
								Μ.3.4	0	-	Μ.1.3	0	-
								Μ.4.4	1	-	Μ.1.4	0	-
								Μ.1-2	0	0	Μ.1.5	0	-
								Μ.1-3	0	0	Μ.2.2	1	-
								Μ.1-4	0	0	Μ.2.3	0	-
								Σύνολο	13	1	Μ.2.4	0	-
											Μ.2.5	0	-
											Μ.3.3	0	-
											Μ.3.4	0	-
											Μ.3.5	0	-
											Μ.4.4	0	-
											Μ.4.5	1	-
											Μ.5.5	0	-
											Μ.1-2	1	1
											Μ.1-3	0	1
											Μ.1-4	1	0
											Μ.1-5	1	1
											Σύνολο	22	4

Πίνακας 5.4: Χρήσιμοι συντελεστές για την αξιολόγηση των αποτελεσμάτων της παλινδρόμησης του βαθμού αποφοίτησης

	Αξιολόγηση αποτελεσμάτων της προς τα πίσω απαλοιφής, για την πρόβλεψη του βαθμού αποφοίτησης και σύγκριση επιλογής μεταβλητών						
	S	R-sq	R-sq(adj)	PRESS	R-sq(pred)	AICc	BIC
	1ο ακαδημαϊκό έτος						
Όλες οι μεταβλητές	0.3978	55.75%	55.07%	22.2868	52.06%	137.52	148.76
	2ο ακαδημαϊκό έτος						
Όλες οι μεταβλητές	0.280402	78.86%	77.68%	11.6587	74.92%	50.42	74.97
Μεταβλητές συνολικών επιδόσεων	0.297945	75.37%	74.80%	12.5789	72.94%	61.76	75.74
	3ο ακαδημαϊκό έτος						
Όλες οι μεταβλητές	0.176303	91.64%	91.18%	4.51471	90.29%	-73.01	-48.46
Μεταβλητές συνολικών επιδόσεων	0.186541	90.35%	90.12%	4.88616	89.49%	-62.79	-48.81
	4ο ακαδημαϊκό έτος						
Όλες οι μεταβλητές	0.123249	96.01%	95.69%	2.34738	94.95%	-164.33	-132.24
Μεταβλητές συνολικών επιδόσεων	0.150238	93.79%	93.59%	3.18475	93.15%	-119.2	-102.53
	5ο ακαδημαϊκό έτος						
Όλες οι μεταβλητές	0.063431	99.03%	98.86%	0.693835	98.51%	-325.8	-271.41
Μεταβλητές συνολικών επιδόσεων	0.09643	97.48%	97.36%	1.43715	96.91%	-234.75	-212.79

Πίνακας 5.5: Χρήσιμοι συντελεστές για την αξιολόγηση των αποτελεσμάτων της παλινδρόμησης της διάρκειας σπουδών

	Αξιολόγηση αποτελεσμάτων της προς τα πίσω απαλοιφής, για την πρόβλεψη της διάρκειας σπουδών και σύγκριση επιλογής μεταβλητών						
	S	R-sq	R-sq(adj)	PRESS	R-sq(pred)	AICc	BIC
	1ο ακαδημαϊκό έτος						
Όλες οι μεταβλητές	0.972905	35.24%	34.75%	128.709	32.78%	374.3	382.79
	2ο ακαδημαϊκό έτος						
Όλες οι μεταβλητές	0.830024	53.59%	52.51%	99.1796	48.20%	334.29	348.27
Μεταβλητές συνολικών επιδόσεων	0.943386	39.11%	38.65%	120.568	37.03%	366.11	374.59
	3ο ακαδημαϊκό έτος						
Όλες οι μεταβλητές	0.762075	62.09%	59.96%	84.2967	55.98%	316.38	340.93
Μεταβλητές συνολικών επιδόσεων	0.849557	51.00%	50.25%	99.0297	48.28%	339.35	350.6
	4ο ακαδημαϊκό έτος						
Όλες οι μεταβλητές	0.715978	68.14%	64.66%	94.0641	50.88%	307.87	347.13
Μεταβλητές συνολικών επιδόσεων	0.787692	57.55%	57.23%	85.067	55.57%	318.13	326.61
	5ο ακαδημαϊκό έτος						
Όλες οι μεταβλητές	0.535316	83.54%	80.25%	50.2864	73.74%	245.07	303.33
Μεταβλητές συνολικών επιδόσεων	0.6563	71.21%	70.31%	60.5722	68.37%	272.99	289.66

5.1.2.1 Αποτελέσματα 1ου ακαδημαϊκού έτους

Βαθμός αποφοίτησης

Για το πρώτο ακαδημαϊκό έτος οι μεταβλητές Π.1.1 και Μ.1.1 συμπίπτουν με τις μεταβλητές Μ.1-1 και Π.1-1, οπότε γίνεται έλεγχος μόνο αυτών των δύο. Για τον βαθμό αποφοίτησης ορίστηκαν σημαντικές και οι 2 μεταβλητές όπως φαίνεται στον πίνακα 5.2. Από την μελέτη του πίνακα 5.4 το μόνο που έχουμε να εξετάσουμε προς το παρόν είναι τις τιμές των διαφορετικών R^2 αλλά και να συμπεράνουμε πως το τυπικό σφάλμα, με τιμή 0.3978 βαθμολογικές μονάδες, φαίνεται να είναι ικανοποιητικό για αυτό το αρχικό σετ δεδομένων. Από την σύγκριση του συντελεστή παλινδρόμησης με τον προσαρμοσμένο, παρατηρούμε πως οι τιμές των 2 είναι πολύ κοντινές, ενώ μεγαλύτερο

ενδιαφέρον εμφανίζει η σύγκριση του συντελεστή παλινδρόμησης με τον R^2_{LOOCV} ο οποίος δίνεται ως R-sq(pred). Από την παραπάνω σύγκριση μπορούμε να συμπεράνουμε το κατά πόσο το μοντέλο θα δώσει ικανοποιητικές προβλέψεις για μελλοντικά δεδομένα με R^2=55.75% και R^2_{LOOCV}=52.06%. Η διαφορά στα ποσοστά των δύο συντελεστών θεωρείται αμελητέα και πιθανό είναι οι προβλέψεις μελλοντικών τιμών να μην απέχουν πολύ από τις προβλέψεις των τιμών του σετ δεδομένων. Ταυτόχρονα ο R^2 αν και εφόσον παρουσιάσει αύξηση στα επόμενα ακαδημαϊκά έτη θεωρείται και αυτό ικανοποιητικό.

Διάρκεια σπουδών

Όσον αφορά τη διάρκεια σπουδών ορίστηκε σημαντική μόνο η μεταβλητή Π.1.1 με αποτέλεσμα να οδηγηθούμε σε απλή γραμμική παλινδρόμηση στην εν λόγω περίπτωση. Αρχικά το πρώτο πράγμα που παρατηρούμε από την μελέτη του πίνακα 5.5 είναι πως το τυπικό σφάλμα παρουσιάζει μεγάλη τιμή κοντά στο 1 έτος φοίτησης. Ταυτόχρονα ο R^2=35.24% είναι αρκετά χαμηλός έχοντας ως αποτέλεσμα να θεωρηθεί πως οι προβλέψεις με το εν λόγω μοντέλο για την διάρκεια σπουδών δεν θα είναι ικανοποιητικές. Ακόμα μικρότερος είναι ο R^2_{LOOCV}=32.78%, όχι όμως μικρότερος σημαντικά από τον R^2 ώστε να υπάρξει ένδειξη σημαντικής διαφοράς στις προβλέψεις μελλοντικών τιμών.

5.1.2.2 Αποτελέσματα 2^{ου} ακαδημαϊκού έτους

Βαθμός αποφοίτησης

Από αυτό το έτος μπορούμε πλέον να προχωρήσουμε σε σύγκριση των αποτελεσμάτων με το πρώτο και σύγκριση όλων των ορισμένων μεταβλητών για αυτό το έτος με τις μεταβλητές οι οποίες σχετίζονται με τις συνολικές επιδόσεις. Στην περίπτωση όλων των μεταβλητών ορίστηκαν σημαντικές 7 μεταβλητές από τις 8 και 3 από τις 4 για την περίπτωση των μεταβλητών των συνολικών επιδόσεων, με όλες τις μεταβλητές των ποσοστών περασμένων μαθημάτων να ορίστηκαν σημαντικές και στις δύο περιπτώσεις. Το παραπάνω μπορεί να εξηγηθεί από τη σημαντικότητα της διαφοράς των επιδόσεων ανά έτος στα ποσοστά περασμένων μαθημάτων. Από την σύγκριση των αποτελεσμάτων του 1^{ου} ακαδημαϊκού έτους με το 2^{ο} παρατηρούμε σημαντικές μειώσεις των όρων AIC_C, BIC και PRESS αλλά και της τιμής του τυπικού σφάλματος και για τις δύο περιπτώσεις, κάτι το οποίο υποδεικνύει την ανωτερότητα των εν λόγων μοντέλων σε σύγκριση με το μοντέλο του 1^{ου} ακαδημαϊκού έτους. Ταυτόχρονα υπήρξε ισχυρή αύξηση σε όλους τους συντελεστές προσδιορισμού, ενώ για το 2^{ο} ακαδημαϊκό έτος οι R^2_{LOOCV} παρουσιάζουν και πάλι αμελητέα διαφορά με τους R^2 και στις δύο περιπτώσεις. Αναφορικά με την σύγκριση όλων των μεταβλητών και των μεταβλητών των συνολικών επιδόσεων, παρατηρούμε πως η προσθήκη τεσσάρων επιπλέον μεταβλητών παρουσίασε σταθερά καλύτερα αποτελέσματα σε όλους του συντελεστές, χωρίς όμως να φάνηκε καμία ισχυρή ανωτερότητα έναντι του μοντέλου το οποίο περιείχε υποσύνολο όλων των εξεταζόμενων μεταβλητών εισόδου.

Διάρκεια σπουδών

Η μέθοδος της προς τα πίσω απαλοιφής όρισε ως υποσύνολο μεταβλητών εισόδου 3 μεταβλητές στην περίπτωση όλων των εξεταζόμενων μεταβλητών για αυτό το έτος και 1 μεταβλητή συνολικών επιδόσεων την Π.1-2. Ξεκινώντας πάλι με τη σύγκριση των αποτελεσμάτων του μοντέλου του 1^{ου} ακαδημαϊκού έτους με αυτά του 2^{ου}, φαίνεται να υπερτερεί καθαρά το μοντέλο με τις 3 μεταβλητές εισόδου, ενώ κάτι τέτοιο δεν ισχύει για αυτό με την μία μεταβλητή εισόδου. Στην 1^{η} περίπτωση

παρατηρούμε αύξηση στους συντελεστές προσδιορισμού ενώ καλύτερη εικόνα παρατηρείται και στα υπόλοιπα αποτελέσματα παρουσιάζοντας ταυτόχρονα χαμηλότερο τυπικό σφάλμα. Στην 2η περίπτωση, αν και υπήρξαν αυξήσεις στους R^2, η διαφορά δεν είναι σημαντικά μεγάλη με το μοντέλο του 1ου έτους με τον συντελεστή προσδιορισμού R^2 =39.11% να παραμένει σε πολύ χαμηλά επίπεδα. Ως εκ τούτου, το μοντέλο με τις 3 μεταβλητές θεωρείται ικανοποιητικά ανώτερο από αυτό του μοντέλου με τη 1, παρουσιάζοντας μείωση στο τυπικό σφάλμα 0.14 χρόνια φοίτησης και αύξηση των R^2 και R^2_{LOOCV} σε 53.59% και 48.20% αντίστοιχα.

5.1.2.3 Αποτελέσματα 3ου ακαδημαϊκού έτους

Βαθμός αποφοίτησης

Στην περίπτωση του 3ου ακαδημαϊκού έτος δεν παρατηρήθηκε αύξηση στις μεταβλητές εισόδου, με 7 και 2 μεταβλητές να ορίστηκαν σημαντικές στις περιπτώσεις εξέτασης όλων των μεταβλητών και αυτών των συνολικών επιδόσεων αντίστοιχα. Αν και ο αριθμός των μεταβλητών παρέμεινε σε παρόμοιο επίπεδο, τα αποτελέσματα συνέχισαν να είναι σημαντικά καλύτερα συγκριτικά με αυτά των μοντέλων του προηγούμενου έτους. Στην περίπτωση του μοντέλου το οποίο αξιοποίησε υποσύνολο όλων των μεταβλητών εισόδου παρατηρούμε μείωση του τυπικού σφάλματος σε 0.175 βαθμολογικές μονάδες, R^2 αύξηση σε 91.64% αλλά και παρόμοιο R^2_{LOOCV}=90.29%. Παρόμοια εικόνα παρουσίασε και το μοντέλο με τις 2 μεταβλητές εισόδου για το 3ο ακαδημαϊκό έτος, με κατά τι χειρότερα αποτελέσματα, όχι όμως σημαντικά χειρότερα ώστε να αιτιολογούν την διαφορά στον αριθμό μεταβλητών εισόδου.

Διάρκεια σπουδών

Συγκριτικά με το 2ο ακαδημαϊκό έτος παρατηρήθηκε αύξηση στις ορισμένες μεταβλητές εισόδου σε 7 και 2 σημαντικές για τις περιπτώσεις εξέτασης όλων των μεταβλητών και αυτών των συνολικών επιδόσεων αντίστοιχα. Παρ' όλα αυτά, τα αποτελέσματα παρέμειναν σε μη ικανοποιητικά επίπεδα, ειδικά άμα ληφθεί υπόψιν πως μεγαλύτερο ενδιαφέρον παρουσιάζουν οι προβλέψεις με τη χρήση δεδομένων στα μικρότερα έτη φοίτησης και εφόσον πλέον βρισκόμαστε κοντά στα 5 υποχρεωτικά έτη φοίτησης. Στην περίπτωση της εξέτασης των μεταβλητών που σχετίζονται με τις συνολικές επιδόσεις, παρατηρούμε χειρότερα αποτελέσματα από αυτών του μοντέλου με τις 3 μεταβλητές εισόδου στο 2ο ακαδημαϊκό έτος. Φαίνεται λοιπόν η χρησιμότητα, στην εν λόγω περίπτωση, της προσθήκης επιπλέον μεταβλητών. Αντιθέτως το μοντέλο με τις 7 ανεξάρτητες μεταβλητές παρουσίασε σημαντικές αυξήσεις στους συντελεστές προσδιορισμού με τους R^2=62.09% και R^2_{LOOCV}=55.98% να μην παρουσιάζουν ανησυχητική διαφορά. Το τυπικό σφάλμα παρέμεινε υψηλό με την τιμή του να είναι 0.76 χρόνια φοίτησης.

5.1.2.4 Αποτελέσματα 4ου ακαδημαϊκού έτους

Βαθμός αποφοίτησης

Οι σημαντικές ανεξάρτητες μεταβλητές αυξήθηκαν στο 4ο έτος σε 10 και 4 στις περιπτώσεις εξέτασης όλων των μεταβλητών και των συνολικών επιδόσεων αντίστοιχα. Για πρώτη φορά παρουσιάζεται σημαντική διαφορά στα αποτελέσματα των δύο περιπτώσεων, ενώ και τα δύο μοντέλα αποδίδουν καλύτερα από αυτά του 3ου ακαδημαϊκού έτους. Αναλυτικότερα στην πρώτη περίπτωση οι συντελεστές προσδιορισμού είναι: R^2=96.01%, $R^2_{adjusted}$= 95.69% και R^2_{LOOCV}=94.95%, ενώ στην 2η είναι : R^2=93.79%%, $R^2_{adjusted}$= 93.59% και R^2_{LOOCV}=93.15% και με τις δύο περιπτώσεις να φαίνεται να υπάρχει

καλή προσαρμογή στα δεδομένα αλλά και σε μελλοντικά. Από τη σύγκριση των AIC_C, BIC και PRESS φαίνεται η καθαρή υπεροχή του 1ου μοντέλου με τις 10 ανεξάρτητες μεταβλητές έναντι του 2ου με τις 4, με αποτέλεσμα να αιτιολογείται η διαφορά στον αριθμό μεταβλητών εισόδου.

Διάρκεια σπουδών

Όσον αφορά τη διάρκεια σπουδών, ορίστηκαν ως σημαντικές 13 ανεξάρτητες μεταβλητές στην περίπτωση εξέτασης όλων των μεταβλητών και μόνο 1 στην περίπτωση των συνολικών επιδόσεων. Ξεκινώντας με την αξιολόγηση των αποτελεσμάτων της πρώτης περίπτωσης, μπορούμε να παρατηρήσουμε μείωση στο τυπικό σφάλμα και αύξηση του R^2 σε σύγκριση με το 3ο έτος, το οποίο υποδεικνύει καλύτερη προσαρμογή της παλινδρόμησης στα δεδομένα. Κάτι τέτοιο όμως δεν ισχύει για την περίπτωση μελλοντικών δεδομένων με τον BIC να παρουσιάζει μικρή αύξηση και τον R^2_{LOOCV}=50.88% μείωση, αλλά και μεγάλη διαφορά με τον R^2=68.14%. Οι παραπάνω συγκρίσεις μας οδηγούν στο συμπέρασμα πως υπήρξε υπερβολική προσαρμογή της παλινδρόμησης στα δεδομένα. Ένας από τους λόγους για τους οποίους θα μπορούσε να εξηγηθεί το εν λόγω overfitting, είναι πως στο 4ο έτος ξεκίνησε ο 2ος κύκλος σπουδών και η επιλογή κατευθύνσεων. Πιο συγκεκριμένα, πλέον η φοίτηση δεν είναι κοινή για όλους με τις κατευθύνσεις ενδεχομένως να παρουσιάζουν διαφορετικά επίπεδα δυσκολίας και ανάλογα με το κατά πόσο ένας φοιτητής επέλεξε την κατάλληλη κατεύθυνση για αυτόν. Επομένως η πορεία των φοιτητών σε αυτό το έτος μπορεί να παρουσίασε μεγαλύτερη τυχαιότητα συγκριτικά με τις επιδόσεις των προηγούμενων ετών. Τέλος, αναφορικά με την περίπτωση της εξέτασης των συνολικών επιδόσεων, αν και δεν υπήρξε υπερβολική προσαρμογή, δεν φάνηκε να παρουσίασε καλύτερα αποτελέσματα από αυτήν του μοντέλου του 3ου ακαδημαϊκού έτους με όλες τις μεταβλητές. Σύμφωνα με τα παραπάνω, φαίνεται πως το σετ δεδομένων του 4ου ακαδημαϊκού έτους δεν βοηθάει στην βελτίωση της πρόβλεψης της διάρκειας σπουδών.

5.1.2.5 Αποτελέσματα 5ου ακαδημαϊκού έτους

Βαθμός αποφοίτησης

Στο έτος αυτό παρουσιάστηκε μεγάλη αύξηση των επιλεγμένων μεταβλητών εισόδου με 20 και 6 μεταβλητές να ορίστηκαν σημαντικές για τις περιπτώσεις εξέτασης όλων των μεταβλητών και των συνολικών επιδόσεων αντίστοιχα. Καθώς πλέον έχουμε χρησιμοποιήσει δεδομένα από τη συνολική υποχρεωτική φοίτηση, ο βαθμός αποφοίτησης μπορεί να υπολογιστεί με σχετική ακρίβεια και για την περίπτωση άγνωστων δεδομένων. Συγκρίνοντας τα 2 μοντέλα του 5ου έτους παρατηρούμε σημαντική υπεροχή στο μοντέλο με τις 20 ανεξάρτητες μεταβλητές στους AIC_C, BIC και PRESS αλλά και στον R^2_{LOOCV}, διαφορές που φαίνεται να αιτιολογούν την προσθήκη των επιπλέον ανεξάρτητων μεταβλητών. Αναφορικά το R^2 για το μοντέλο με τις 20 ανεξάρτητες μεταβλητές ανήλθε σε 99.03%, ενώ το τυπικό σφάλμα μόλις 0.06 βαθμολογικές μονάδες, ενώ το R^2_{LOOCV}=98.51% είναι ελάχιστα μικρότερο του R^2.

Διάρκεια σπουδών

Σε αυτό το έτος καθώς έχουμε φτάσει στο τέλος της υποχρεωτικής φοίτησης είναι λογικό να παρουσιαστούν καλύτερα αποτελέσματα συγκριτικά με τα προηγούμενα έτη, καθώς αρκετοί από τους εξεταζόμενους φοιτητές έχουν πλέον αποφοιτήσει. Αυτό έχει και ως αποτέλεσμα να παρουσιάζει μικρότερο ενδιαφέρον η μελέτη της πρόβλεψης της διάρκειας σπουδών για φοιτητές που έχουν διανύσει τα πρώτα 5 έτη φοίτησης.

Η μέθοδος της προς τα πίσω απαλοιφής όρισε σημαντικές 22 και 4 ανεξάρτητες μεταβλητές για τις περιπτώσεις εξέτασης όλων των μεταβλητών και των συνολικών επιδόσεων αντίστοιχα. Τα αποτελέσματα των μοντέλων παρουσιάζουν σημαντικές βελτιώσεις συγκριτικά με τα προηγούμενα έτη. Από την σύγκριση των μοντέλων του 5ου έτους οδηγούμαστε στο συμπέρασμα πως το μοντέλο με τις 22 ανεξάρτητες μεταβλητές είναι ανώτερο αυτού με τις 4, έχοντας καλύτερο R^2=83.54% έναντι R^2=71.21% και καλύτερο R^2_{LOOCV}=73.74% έναντι R^2_{LOOCV}=68.37%, με τη διαφορά R^2- R^2_{LOOCV} να μην είναι αρκετά μεγάλη ώστε να θεωρηθεί πως έγινε υπερβολική προσαρμογή στα δεδομένα.

5.2.1 Διασταυρούμενη επικύρωση τμημάτων (k-fold Cross-Validation)

Όπως αναφέρθηκε στην αρχή του κεφαλαίου θα πρέπει να γίνει ακριβής προσέγγιση των σφαλμάτων πρόβλεψης για δεδομένα εκτός των υπαρχόντων. Για την επίτευξη του παραπάνω, τα μοντέλα πρόβλεψης θα πρέπει να χαρακτηρίζονται από χαμηλή μεροληψία και χαμηλή διακύμανση. Πιο συγκεκριμένα δηλαδή, δεν επιθυμούμε μοντέλα τα οποία καταλήγουν σε υπεραπλούστευση των δεδομένων ή/και μοντέλα τα οποία ακολούθησαν πολύ προσεκτικά την εικόνα των δεδομένων έχοντας ως αποτέλεσμα τη δυσκολία γενίκευσης της πρόβλεψης. Με τη μέθοδο της διασταυρούμενης επικύρωσης τμημάτων επιτυγχάνουμε μία ικανοποιητική προσέγγιση του test-error, αν και συνήθως υποδεικνύει μικρότερο σφάλμα από το αληθινό, αλλά και ικανοποιητική ισορροπία μεταξύ μεροληψίας και διακύμανσης.

Η διαδικασία της διασταυρούμενης επικύρωσης τμημάτων ξεκινάει με τον τυχαίο διαχωρισμό του σετ δεδομένων σε k αριθμό ομάδων. Για κάθε ξεχωριστή ομάδα χρησιμοποιείται μία ομάδα ως σύνολο επικύρωσης (test data set) και οι υπόλοιπες δημιουργούν το σύνολο εκπαίδευσης (training data set). Το κάθε μοντέλο στη συνέχεια εκπαιδεύεται με τη χρήση των δεδομένων του συνόλου εκπαίδευσης και αξιολογείται από το σύνολο επικύρωσης. Μετέπειτα το αποτέλεσμα της αξιολόγησης συγκρατείται για κάθε μοντέλο, με τελικό αποτέλεσμα τον μέσο όρο των αποτελεσμάτων των ξεχωριστών μοντέλων, δηλαδή έχουμε k εκτιμήσεις του test-error και κρατάμε τον μέσο όρο τους. Στο πλαίσιο της εργασίας, καθώς δίνεται έμφαση στην ύπαρξη ακραίων τιμών, θα χρησιμοποιηθεί το RMSE ως μέτρο του test-error. Σύμφωνα με τα παραπάνω, η εκτίμηση του RMSE μετά από διασταυρούμενη επικύρωση k τμημάτων θα είναι:

$$RMSE_{(k)} = \frac{1}{k}\sum_{i=1}^{k} RMSE_i \quad (5.5)$$

Ειδική περίπτωση της διασταυρούμενης επικύρωσης αποτελεί η LOOCV, για την οποία το k ισούται με τον αριθμό n των παρατηρήσεων και χαρακτηρίζεται από χαμηλή μεροληψία. Όπως αναφέρθηκε όμως, προσδοκούμε χαμηλή μεροληψία ταυτόχρονα με χαμηλή διακύμανση. Έχει αποδειχθεί εμπειρικά, έπειτα από πειραματικές διαδικασίες, πως για την επίτευξη της ισορροπίας μεταξύ των δύο προτιμάται η τιμή k=10.

Ακόμα και με τη χρήση διασταυρούμενης επικύρωσης δέκα τμημάτων, η εκτίμηση της αποτελεσματικότητας ενός μοντέλου μπορεί να εμπεριέχει θόρυβο. Αυτό οφείλεται στον τυχαίο διαχωρισμό του σετ δεδομένων, με διαφορετικούς διαχωρισμούς να οδηγούν σε διαφορετικά αποτελέσματα. Για την αντιμετώπιση του παραπάνω προβλήματος χρησιμοποιείται η επαναλαμβανόμενη διασταυρούμενη επικύρωση δέκα τμημάτων και αξιολογείται η τιμή του μέσου όρου των αποτελεσμάτων τους. Συνεπώς, στην περίπτωση επανάληψης διασταυρούμενης επικύρωσης δέκα τμημάτων δέκα φορές, έχει ως αποτέλεσμα 100 διαφορετικούς συνδυασμούς

συνόλων εκπαίδευσης με επικύρωσης [6]. Οπότε το εκτιμώμενο RMSE για x αριθμό επαναλήψεων διασταυρούμενης επικύρωσης k τμημάτων θα είναι:

$$RMSE_{(x)} = \frac{1}{x}\sum_{j=1}^{x} RMSE_{(k)} \quad (5.6)$$

5.2.2 Σύγκριση μοντέλων μέσω επαναλαμβανόμενης επικύρωσης

Για τη σύγκριση των μοντέλων παλινδρόμησης της υποενότητας 5.1.2, αλλά και την προσέγγιση του test-error, θα χρησιμοποιήσουμε επαναλαμβανόμενη επικύρωση δέκα τμημάτων, δέκα φορές. Η προσέγγιση του σφάλματος πρόβλεψης για δεδομένα εκτός των υπαρχόντων θα γίνει με την εκτίμηση των τιμών RMSE. Για κάθε έτος θα γίνει αξιολόγηση της ποσοστιαίας διαφοράς των RMSE των μοντέλων, αν και εφόσον χρησιμοποιήθηκαν τουλάχιστον δύο. Οι τιμές των RMSE όλων των διασταυρούμενων επικυρώσεων 10 τμημάτων, καθώς και οι μέσοι όροι των δέκα επαναλαμβανόμενων διασταυρούμενων επικυρώσεων για κάθε μοντέλο δίνονται στους πίνακες 5.6 και 5.7 για την πρόβλεψη του βαθμού αποφοίτησης και διάρκειας σπουδών αντίστοιχα.

Πίνακας 5.6: Επαναλαμβανόμενη επικύρωση 10 τμημάτων των σφαλμάτων των μοντέλων για την πρόβλεψη του βαθμού αποφοίτησης

Ακαδημαϊκό έτος/Μοντέλα πρόβλεψης βαθμού αποφοίτησης	RMSE 1ης επικύρωσης 10 τμημάτων	RMSE 2ης επικύρωσης 10 τμημάτων	RMSE 3ης επικύρωσης 10 τμημάτων	RMSE 4ης επικύρωσης 10 τμημάτων	RMSE 5ης επικύρωσης 10 τμημάτων	RMSE 6ης επικύρωσης 10 τμημάτων	RMSE 7ης επικύρωσης 10 τμημάτων	RMSE 8ης επικύρωσης 10 τμημάτων	RMSE 9ης επικύρωσης 10 τμημάτων	RMSE 10ης επικύρωσης 10 τμημάτων	Μέσος όρος RMSE	Ποσοστιαία διαφορά RMSE μοντέλων ανά έτος (1=100%)
1ο Ακαδημαϊκό έτος												
Γραμμική παλινδρόμηση	0.4075	0.4076	0.4157	0.4104	0.4043	0.4115	0.4188	0.4104	0.4115	0.4089	0.4106	
2ο Ακαδημαϊκό έτος												
Γραμμική παλινδρόμηση (όλες οι μεταβλητές εισόδου)	0.2950	0.3034	0.2990	0.3011	0.2945	0.2944	0.2952	0.3005	0.2920	0.2967	0.2972	
Γραμμική παλινδρόμηση (μεταβλητές συνολικών επιδόσεων)	0.3111	0.3129	0.3092	0.3071	0.3098	0.3134	0.3074	0.3146	0.3032	0.3121	0.3101	0.0434
3ο Ακαδημαϊκό έτος												
Γραμμική παλινδρόμηση (όλες οι μεταβλητές εισόδου)	0.1801	0.1846	0.1844	0.1889	0.1878	0.1846	0.1868	0.1891	0.1803	0.1820	0.1849	
Γραμμική παλινδρόμηση (μεταβλητές συνολικών επιδόσεων)	0.1934	0.1957	0.1901	0.1916	0.1948	0.1927	0.1915	0.1941	0.1909	0.1907	0.1926	0.0417
4ο Ακαδημαϊκό έτος												
Γραμμική παλινδρόμηση (όλες οι μεταβλητές εισόδου)	0.1350	0.1324	0.1340	0.1351	0.1380	0.1378	0.1372	0.1349	0.1337	0.1338	0.1352	
Γραμμική παλινδρόμηση (μεταβλητές συνολικών επιδόσεων)	0.1535	0.1575	0.1546	0.1547	0.1610	0.1556	0.1543	0.1551	0.1547	0.1551	0.1556	0.1510
5ο Ακαδημαϊκό έτος												
Γραμμική παλινδρόμηση (όλες οι μεταβλητές εισόδου)	0.0727	0.0736	0.0707	0.0743	0.0740	0.0730	0.0738	0.0774	0.0741	0.0716	0.0735	
Γραμμική παλινδρόμηση (μεταβλητές συνολικών επιδόσεων)	0.1049	0.1049	0.1033	0.1012	0.1081	0.1064	0.1036	0.1046	0.1037	0.1019	0.1043	0.4179

Πίνακας 5.7: Επαναλαμβανόμενη επικύρωση 10 τμημάτων των σφαλμάτων των μοντέλων για την πρόβλεψη της διάρκειας σπουδών

Ακαδημαϊκό έτος/Μοντέλα πρόβλεψης διάρκεια σπουδών	RMSE 1ης επικύρωσης 10 τμημάτων	RMSE 2ης επικύρωσης 10 τμημάτων	RMSE 3ης επικύρωσης 10 τμημάτων	RMSE 4ης επικύρωσης 10 τμημάτων	RMSE 5ης επικύρωσης 10 τμημάτων	RMSE 6ης επικύρωσης 10 τμημάτων	RMSE 7ης επικύρωσης 10 τμημάτων	RMSE 8ης επικύρωσης 10 τμημάτων	RMSE 9ης επικύρωσης 10 τμημάτων	RMSE 10ης επικύρωσης 10 τμημάτων	Μέσος όρος RMSE	Ποσοστιαία διαφορά RMSE μοντέλων ανά έτος (1=100%)
1ο Ακαδημαϊκό έτος												
Γραμμική παλινδρόμηση	0.9845	0.9987	1.0167	0.9732	0.9936	0.9919	0.9911	0.9881	0.9853	0.9833	0.9906	
2ο Ακαδημαϊκό έτος												
Γραμμική παλινδρόμηση (όλες οι μεταβλητές εισόδου)	0.8644	0.8459	0.8584	0.8938	0.8607	0.8672	0.8743	0.8772	0.8725	0.8711	0.8686	
Γραμμική παλινδρόμηση (μεταβλητές συνολικών επιδόσεων)	0.9451	0.9561	0.9505	0.9474	0.9586	0.9543	0.9492	0.9574	0.9648	0.9515	0.9535	0.0978
3ο Ακαδημαϊκό έτος												
Γραμμική παλινδρόμηση (όλες οι μεταβλητές εισόδου)	0.7826	0.7962	0.7983	0.7967	0.7917	0.7996	0.7963	0.7890	0.7788	0.8095	0.7939	
Γραμμική παλινδρόμηση (μεταβλητές συνολικών επιδόσεων)	0.8734	0.8803	0.8575	0.8666	0.8984	0.8555	0.8667	0.8683	0.8578	0.8736	0.8698	0.0956
4ο Ακαδημαϊκό έτος												
Γραμμική παλινδρόμηση (όλες οι μεταβλητές εισόδου)	0.7874	0.8608	0.8257	0.8359	0.8433	0.7791	0.8220	0.8820	0.8500	0.8661	0.8352	
Γραμμική παλινδρόμηση (μεταβλητές συνολικών επιδόσεων)	0.8060	0.8046	0.7976	0.7982	0.8010	0.7983	0.7971	0.7919	0.7939	0.7992	0.7988	
5ο Ακαδημαϊκό έτος												
Γραμμική παλινδρόμηση (όλες οι μεταβλητές εισόδου)	0.6111	0.6327	0.6103	0.6128	0.6577	0.6258	0.6297	0.6363	0.6383	0.6347	0.6289	
Γραμμική παλινδρόμηση (μεταβλητές συνολικών επιδόσεων)	0.6766	0.6644	0.6724	0.6776	0.6781	0.6797	0.6814	0.6766	0.6765	0.6796	0.6763	0.0753

5.2.2.1 Αποτελέσματα επαναλαμβανόμενης επικύρωσης τμημάτων 1ου ακαδημαϊκού έτους

Βαθμός αποφοίτησης

Ξεκινώντας με την πρόβλεψη του βαθμού αποφοίτησης το εκτιμώμενο RMSE, δηλαδή ο μέσος όρος των επιμέρους RMSE, προκύπτει 0.4106, τιμή ικανοποιητική για το αρχικό αυτό σετ δεδομένων. Τα επιμέρους RMSE των δέκα επαναλαμβανόμενων επικυρώσεων δέκα τμημάτων φαίνεται να μην παρουσιάζουν μεγάλη διακύμανση και ο μέσος όρος των RMSE συνάδει με τα συμπεράσματα της υποενότητας 5.1.2.1

Διάρκεια σπουδών

Στην υποενότητα 5.1.2.1 είχαμε καταλήξει στο συμπέρασμα πως οι προβλέψεις της διάρκειας σπουδών δεν θα είναι ικανοποιητικές για αυτό το αρχικό σετ δεδομένων. Το εκτιμώμενο RMSE προκύπτει 0.9906 με τη μέθοδο της απλής γραμμικής παλινδρόμησης, τιμή η οποία επαληθεύει τα συμπεράσματά μας. Να σημειωθεί πως τα επιμέρους RMSE δεν παρουσιάζουν μεγάλη διακύμανση και να γίνει υπενθύμιση πως ο R^2 είχε παρουσιάσει αμελητέα διαφορά με τον R^2_{LOOCV}.

5.2.2.2 Αποτελέσματα επαναλαμβανόμενης επικύρωσης τμημάτων 2ου ακαδημαϊκού έτους

Βαθμός αποφοίτησης

Στο 2ο ακαδημαϊκό έτος, για την πρόβλεψη του βαθμού αποφοίτησης, υπήρξε σημαντική μείωση του εκτιμώμενου RMSE και για τα δύο εξεταζόμενα μοντέλα, αυτό με το υποσύνολο μεταβλητών εισόδου εξετάζοντας όλες τις μεταβλητές έχοντας εκτιμώμενο RMSE=0.2972 και εκείνου με τις μεταβλητές που σχετίζονται με τις συνολικές επιδόσεις έχοντας RMSE=0.3101 . Πιο συγκεκριμένα από τη σύγκριση των δύο προκύπτει πως το δεύτερο, απλούστερο μοντέλο παρουσιάζει μόλις 4.34% μεγαλύτερο RMSE.

Διάρκεια σπουδών

Όσον αφορά την πρόβλεψη της διάρκειας σπουδών, τα εκτιμώμενα RMSE, αν και υπήρξε καθαρή βελτίωση σε σύγκριση με αυτό του 1ου, παραμένουν σε υψηλά επίπεδα με $RMSE_{(υποσυνόλου\ όλων\ των\ μεταβλητών)}$=0.8686 και $RMSE_{(υποσυνόλου\ μεταβλητών\ συνολικών\ επιδόσεων)}$=0.9535. Στη 2η περίπτωση παρατηρούμε 9.78% μεγαλύτερη τιμή RMSE, γεγονός που αιτιολογεί το μεγαλύτερο πλήθος ανεξάρτητων μεταβλητών του 1ου μοντέλου.

5.2.2.3 Αποτελέσματα επαναλαμβανόμενης επικύρωσης τμημάτων 3ου ακαδημαϊκού έτους

Βαθμός αποφοίτησης

Όπως και στο 2ο έτος, έτσι και εδώ υπήρξε καθαρή ελάττωση των εκτιμώμενων RMSE, με τα RMSE των δύο μοντέλων να παρουσιάζουν μικρή διαφορά έχοντας $RMSE_{(υποσυνόλου\ όλων\ των\ μεταβλητών)}$=0.1849 και $RMSE_{(υποσυνόλου\ μεταβλητών\ συνολικών\ επιδόσεων)}$=0.1926 και ποσοστιαία διαφορά 4.17% του δεύτερου έναντι του πρώτου.

Διάρκεια σπουδών

Για τη διάρκεια σπουδών παρατηρούμε $RMSE_{(υποσυνόλου\ όλων\ των\ μεταβλητών)}$=0.7939 και $RMSE_{(υποσυνόλου\ μεταβλητών\ συνολικών\ επιδόσεων)}$=0.8698, με τη δεύτερη περίπτωση να έχει 9.56% μεγαλύτερο RMSE. Ταυτόχρονα μπορούμε να συμπεράνουμε πως στη δεύτερη περίπτωση το εκτιμώμενο RMSE είναι παρόμοιο με αυτό του 2ου ακαδημαϊκού έτους, χρησιμοποιώντας υποσύνολο όλων των μεταβλητών. Αυτό μας οδηγεί στο συμπέρασμα πως η ποσοστιαία διαφορά 10% των RMSE είναι αναμφισβήτητα σημαντική στη συγκεκριμένη περίπτωση της πρόβλεψης της διάρκειας σπουδών.

5.2.2.4 Αποτελέσματα επαναλαμβανόμενης επικύρωσης τμημάτων 4ου ακαδημαϊκού έτους

Βαθμός αποφοίτησης

Η απότομη μείωση των εκτιμώμενων RMSE με την αύξηση των ετών συνεχίζεται και σε αυτό το έτος με $RMSE_{(υποσυνόλου\ όλων\ των\ μεταβλητών)}$=0.1352 και $RMSE_{(υποσυνόλου\ μεταβλητών\ συνολικών\ επιδόσεων)}$=0.1556. Σε αυτό το έτος όμως υπήρξε καθαρή υπεροχή του RMSE της 1ης περίπτωσης έχοντας 10 ανεξάρτητες μεταβλητές, με τη 2η να παρουσιάζει 15.1% μεγαλύτερο RMSE με 6 μεταβλητές λιγότερες. Η διαφορά αυτή επιβεβαιώνει τα αποτελέσματα της υποενότητας 5.1.2.4 και πλέον φαίνεται να αιτιολογείται η χρήση περίπλοκων μοντέλων έναντι των απλούστερων.

Διάρκεια σπουδών

Στην υποενότητα 5.1.2.4 είχαμε συμπεράνει πως για το έτος αυτό υπήρξε υπερβολική προσαρμογή της παλινδρόμησης στα δεδομένα στην περίπτωση εξέτασης όλων των μεταβλητών και θεωρήσαμε πως κάτι τέτοιο μπορεί να οφείλεται στην έναρξη του 2ου κύκλου σπουδών. Το overfitting φαίνεται να επιβεβαιώνεται και από την τιμή του εκτιμώμενου $RMSE_{(υποσυνόλου\ όλων\ των\ μεταβλητών)}$=0.8352, το οποίο είναι χειρότερο από αυτό του 3ου έτους αν και παρουσίασε καλύτερο R^2 και τυπικό σφάλμα στα αποτελέσματα του 5.1.2. Ακόμα προέκυψε $RMSE_{(υποσυνόλου\ μεταβλητών\ συνολικών\ επιδόσεων)}$=0.7988, τιμή η οποία δεν παρουσιάζει βελτίωση στην ικανότητα πρόβλεψης από το βέλτιστο μοντέλο του 3ου έτους. Για τους παραπάνω λόγους δεν ενδείκνυται η χρήση του σετ δεδομένων του 4ου ακαδημαϊκού έτους για την πρόβλεψη της διάρκειας σπουδών.

5.2.2.5 Αποτελέσματα επαναλαμβανόμενης επικύρωσης τμημάτων 5ου ακαδημαϊκού έτους

Βαθμός αποφοίτησης

Έχοντας φτάσει πλέον στο τέλος της υποχρεωτικής φοίτησης η πρόβλεψη του βαθμού αποφοίτησης φαίνεται να μπορεί να γίνει με σχετική ακρίβεια με $RMSE_{(υποσυνόλου\ όλων\ των\ μεταβλητών)}$=0.0735 και $RMSE_{(υποσυνόλου\ μεταβλητών\ συνολικών\ επιδόσεων)}$=0.1043. Ταυτόχρονα στη 2η περίπτωση το RMSE είναι 41.79% μεγαλύτερο της πρώτης, με τη διαφορά αυτή να είναι πολύ μεγάλη και να δείχνει τη καθαρή υπεροχή του μοντέλου με το υποσύνολο όλων των μεταβλητών, άσχετα αν το μοντέλο είναι αρκετά πιο περίπλοκο έχοντας 20 μεταβλητές εισόδου.

Διάρκεια σπουδών

Ακόμα και με τη χρήση όλων των δεδομένων της υποχρεωτικής φοίτησης δεν φαίνεται να μπορεί να γίνει πρόβλεψη της διάρκειας σπουδών με ακρίβεια με $RMSE_{(υποσυνόλου\ όλων\ των\ μεταβλητών)}$=0.6289 και $RMSE_{(υποσυνόλου\ μεταβλητών\ συνολικών\ επιδόσεων)}$=0.6763, με τη δεύτερη περίπτωση να παρουσιάζει 7.53% μεγαλύτερο RMSE. Τα αποτελέσματα του έτους αυτού ακόμα και αν είναι σημαντικά καλύτερα του βέλτιστου μοντέλου του 3ου έτους, είναι πλέον αργά ώστε να θεωρηθούν χρήσιμες οι προβλέψεις της διάρκειας σπουδών με αρκετούς από τους φοιτητές να έχουν ολοκληρώσει τις φοιτητικές τους υποχρεώσεις και να έχουν αποφοιτήσει.

5.3 Αποτελέσματα βέλτιστων μοντέλων πρόβλεψης

Στα αποτελέσματα του υποκεφαλαίου 5.2 παρατηρήσαμε σημαντικές βελτιώσεις στα εκτιμώμενα RMSE με τη χρήση περιπλοκότερων μοντέλων, ειδικά για τις περιπτώσεις της πρόβλεψης του βαθμού αποφοίτησης με χρήση των συνολικών δεδομένων του 4ου και 5ου έτους φοίτησης και για τα πρώτα 3 έτη της πρόβλεψης της διάρκειας σπουδών. Τα μοντέλα αυτά, για τα οποία χρησιμοποιήθηκαν ως μεταβλητές εισόδου υποσύνολα όλων των εξεταζομένων μεταβλητών για κάθε έτος, θεωρούνται βέλτιστα και πως η περιπλοκότητά τους αιτιολογείται. Για αυτά τα μοντέλα θα γίνει παρουσίαση των διαγραμμάτων διασποράς των τιμών πρόβλεψης σε σχέση με τις πραγματικές τιμές για τυχαίες διασταυρούμενες επικυρώσεις τμημάτων. Συγχρόνως, θα δωθούν οι τιμές R^2 και MAE για περαιτέρω σύγκριση, ενώ η ανάλυση των RMSE θεωρείται περιττή καθώς έγινε εκτενής ανάλυσή τους στο υποκεφάλαιο 5.2.1.

5.3.1.1 Διαγράμματα διασποράς πρόβλεψης βαθμού αποφοίτησης

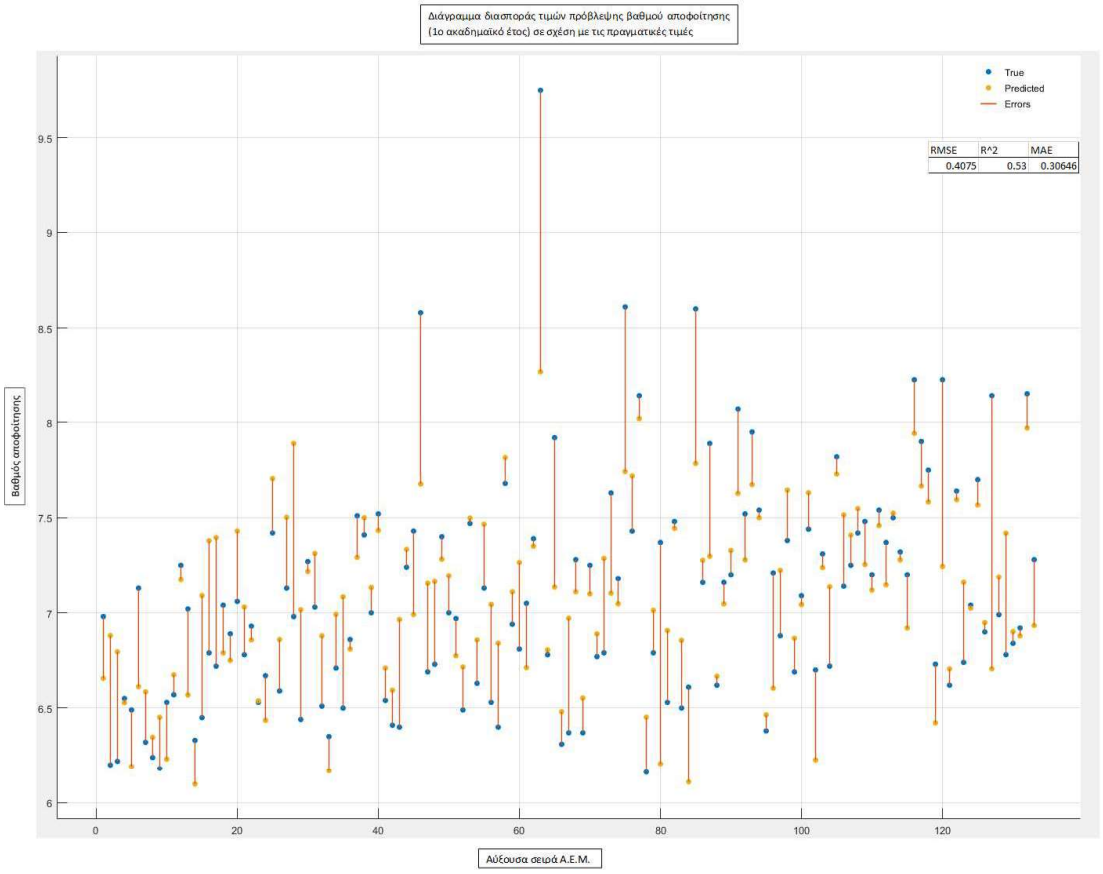
Στο σχήμα 5.1 φαίνεται το διάγραμμα διασποράς τιμών πρόβλεψης του βαθμού αποφοίτησης με τη χρήση δεδομένων του 1ου έτους φοίτησης, σε σχέση με τις πραγματικές τιμές και παρουσιάζοντας διαγραμματικά τα σφάλματα πρόβλεψης για τον κάθε φοιτητή. Η γενική τάση πρόβλεψης φαίνεται να είναι ικανοποιητική και το πλήθος υψηλών σφαλμάτων πρόβλεψης παρόμοιο με αυτό των χαμηλών. Εξαίρεση αποτελεί η τάση της αδυναμίας πρόβλεψης μεγαλύτερων βαθμών αποφοίτησης, προβλέποντας σταθερά μικρότερες βαθμολογίες για πραγματικούς βαθμούς αποφοίτησης μεγαλύτερους του 8. Επιπροσθέτως, ομοίως πολλές φαίνεται να είναι οι προβλέψεις με μεγαλύτερη τιμή από την πραγματική σε σχέση με αυτές που παρουσίασαν μικρότερη. Ο εκτιμώμενος R^2 είναι 0.53 και το MAE είναι 0.30646, τιμές ικανοποιητικές αν ληφθεί υπόψιν πως οι προβλέψεις γίνονται με τη χρήση δεδομένων του 1ου έτους φοίτησης μόνο.

Στο δεύτερο έτος παρουσιάζεται σημαντική βελτίωση στην ικανότητα πρόβλεψης του βαθμού αποφοίτησης, όπως και φαίνεται στο σχήμα 5.2 με εκτιμώμενο R^2=0.75 και MAE=0.23627. Τα μεγέθη των σφαλμάτων πρόβλεψης συρρικνώθηκαν σημαντικά και η τάση υποεκτίμησης βαθμών αποφοίτησης μεγαλύτερων του 8 δεν παραμένει στον ίδιο βαθμό που υπήρξε στο προηγούμενο μοντέλο. Παράλληλα, αντίστοιχα πολλές είναι οι υπερεκτιμήσεις έναντι των υποεκτιμήσεων, χωρίς να φαίνεται κάποια ισχυρή τάση.

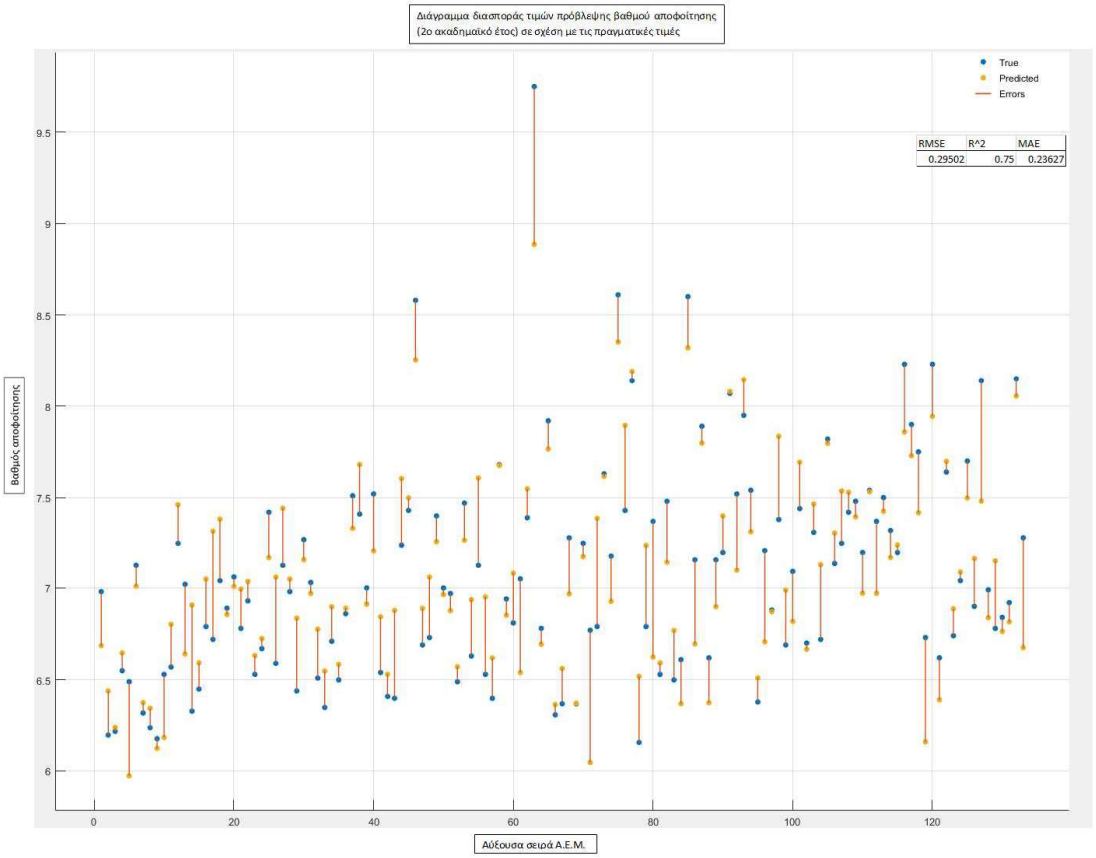
Από τη μελέτη του σχήματος 5.3 με τη χρήση δεδομένων από τα πρώτα 3 έτη φοίτησης, αναμφίβολα πλέον ο βαθμός αποφοίτησης μπορεί να εκτιμηθεί σε πολύ καλό βαθμό. Το παραπάνω συμπέρασμα φαίνεται από τον υψηλό εκτιμώμενο R^2=0.91 αλλά και από το χαμηλό MAE=0.13714. Με τη χρήση του εν λόγω μοντέλου τα σφάλματα πρόβλεψης είναι αρκετά χαμηλά και δεν φαίνεται να υπάρχει κάποια ισχυρή τάση υποεκτιμήσεων ή υπερεκτιμήσεων.

Για το 4° και 5° έτος τα διαγράμματα διασποράς των αντίστοιχων μοντέλων πρόβλεψης φαίνονται στα σχήματα 5.4 και 5.5. Οι τιμές των R^2 παρουσίασαν σταθερή βελτίωση με $R^2_{(4ου\ έτους)}$=0.95 και $R^2_{(5ου\ έτους)}$=0.99, με το ίδιο να ισχύει για τα MAE τα οποία μειώθηκαν διαρκώς με $MAE_{(4ου\ έτους)}$=0.10649 και $MAE_{(5ου\ έτους)}$=0.054712. Σταθερή αύξηση υπήρξε σε αυτά τα έτη και στον αριθμό των προβλέψεων τα οποία παρουσίασαν πολύ μικρά σφάλματα. Συγχρόνως οι προβλέψεις φαίνεται να γίνονται σε πολύ καλό βαθμό και για φοιτητές οι οποίοι δεν ήταν κοντά στην ολοκλήρωση των σπουδών τους, δηλαδή κατά γενικό κανόνα για τους φοιτητές με χαμηλότερο αριθμό Α.Ε.Μ.. Συγκριτικά όμως με τους υπολοίπους, οι φοιτητές με μεγαλύτερη διάρκεια σπουδών παρουσίασαν μικρότερη βελτίωση των σφαλμάτων πρόβλεψης ανά τα έτη.

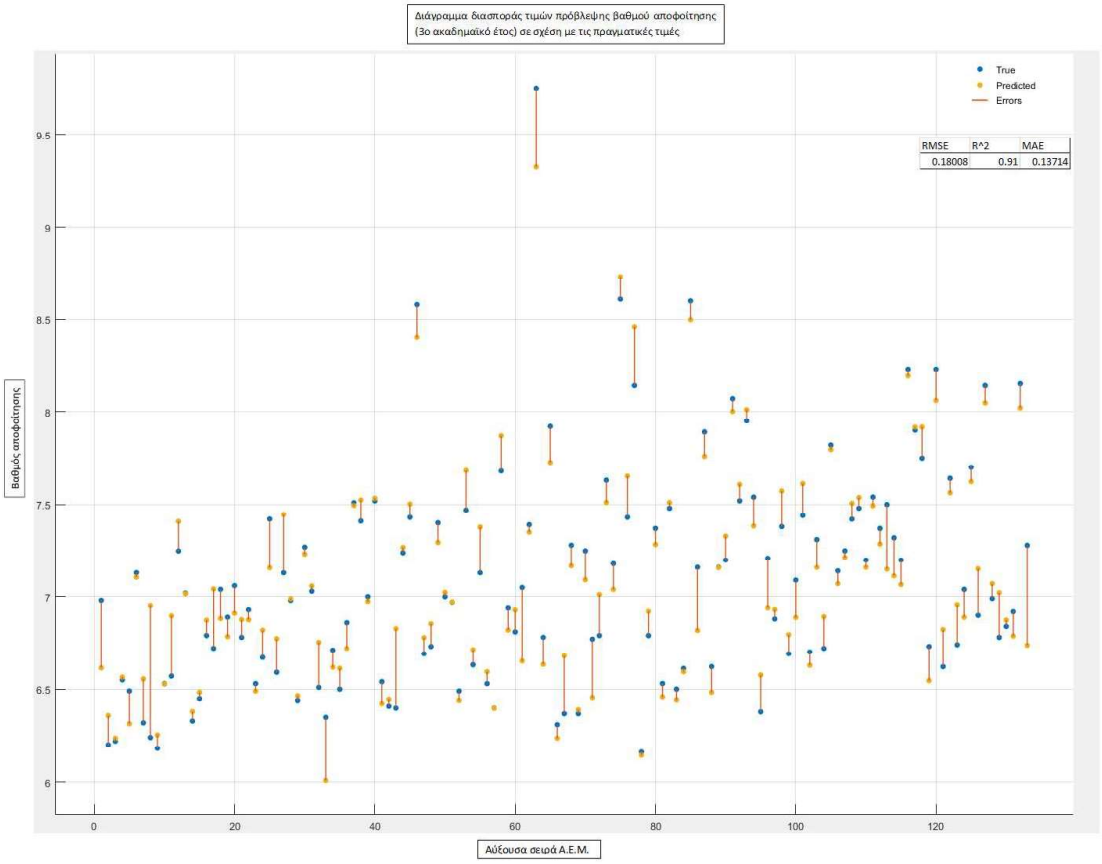
Σχήμα 5.1: Διάγραμμα διασποράς τιμών πρόβλεψης του βαθμού αποφοίτησης σε σχέση με τις πραγματικές τιμές με χρήση δεδομένων του 1^{ου} έτους φοίτησης



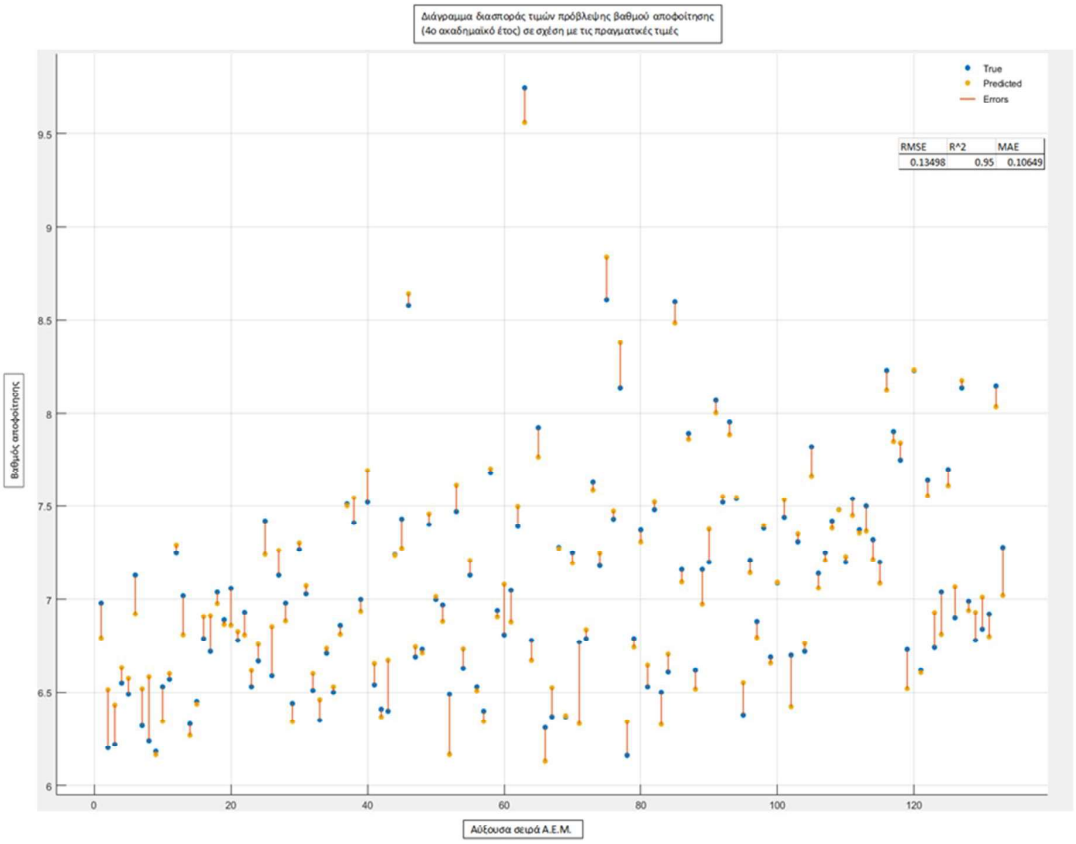
Σχήμα 5.2: Διάγραμμα διασποράς τιμών πρόβλεψης του βαθμού αποφοίτησης σε σχέση με τις πραγματικές τιμές με χρήση δεδομένων του 2^{ου} έτους φοίτησης



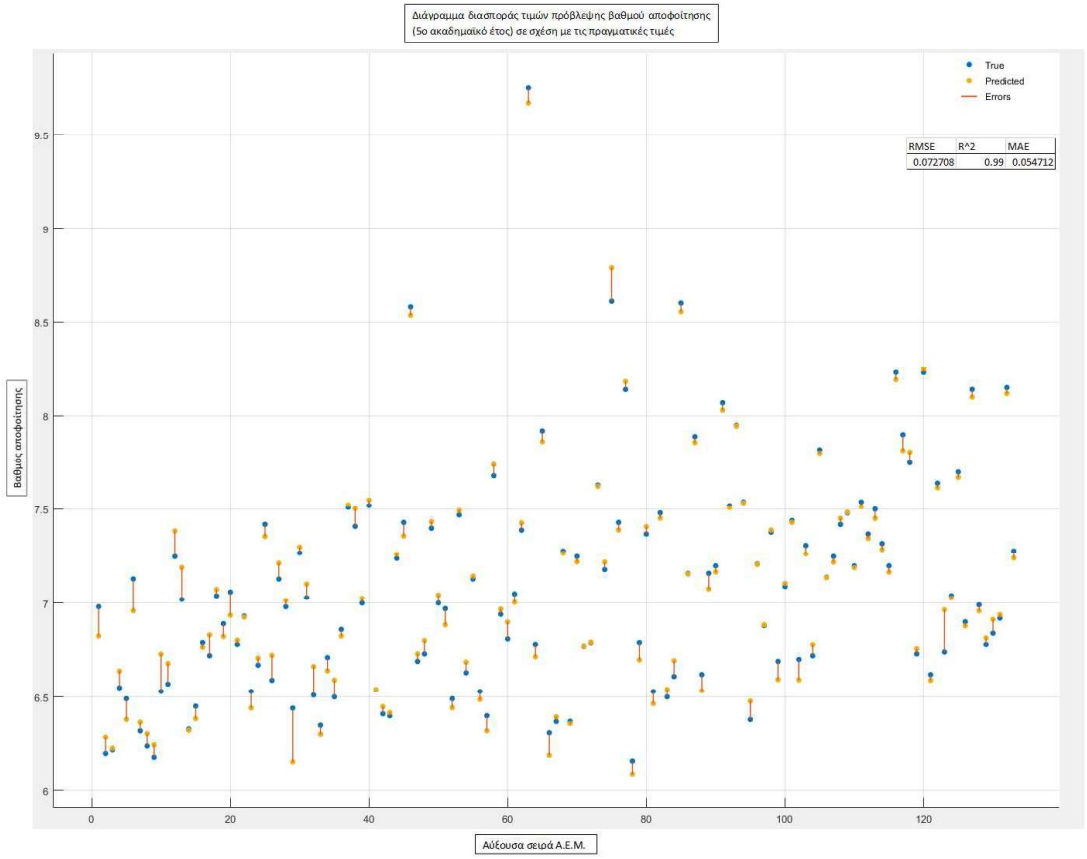
Σχήμα 5.3: Διάγραμμα διασποράς τιμών πρόβλεψης του βαθμού αποφοίτησης σε σχέση με τις πραγματικές τιμές με χρήση δεδομένων του 3^{ου} έτους φοίτησης



Σχήμα 5.4: Διάγραμμα διασποράς τιμών πρόβλεψης του βαθμού αποφοίτησης σε σχέση με τις πραγματικές τιμές με χρήση δεδομένων του 4^{ου} έτους φοίτησης



Σχήμα 5.5: Διάγραμμα διασποράς τιμών πρόβλεψης του βαθμού αποφοίτησης σε σχέση με τις πραγματικές τιμές με χρήση δεδομένων του 5ου έτους φοίτησης



5.3.1.2 Αξιολόγηση αποτελεσμάτων πρόβλεψης βαθμού αποφοίτησης ενός τυχαίου αποφοίτου

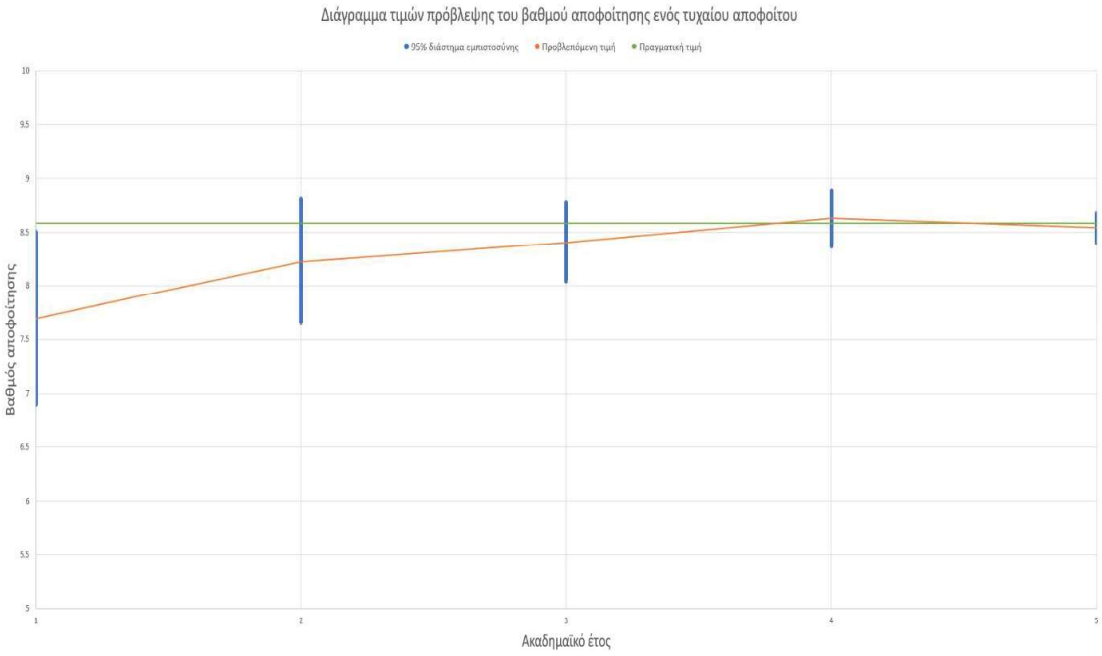
Επιλέγοντας τυχαία έναν απόφοιτο εκ των 133 μπορούμε να προχωρήσουμε σε επιπλέον ανάλυση των τιμών πρόβλεψης του κάθε μοντέλου για αυτόν τον απόφοιτο. Ο απόφοιτος ο οποίος επιλέχθηκε είναι ο υπ' αριθμόν 46 σε αύξουσα σειρά Α.Ε.Μ με πραγματική τιμή βαθμού αποφοίτησης 8.58. Στον πίνακα 5.8 παρουσιάζονται οι προβλεπόμενοι βαθμοί αποφοίτησης ανά έτος φοίτησης εν συνοδεία των 95% διαστημάτων εμπιστοσύνης τους. Τα αποτελέσματα του πίνακα 5.8 παρουσιάζονται διαγραμματικά στο σχήμα 5.6. Ως σύνολο εκπαίδευσης για τα εν λόγω αποτελέσματα χρησιμοποιήθηκαν τα δεδομένα και των 133 αποφοίτων. Γίνεται υπενθύμιση πως για τα πρώτα δύο έτη φοίτησης τα μοντέλα της πρόβλεψης έχουν την τάση να υποεκτιμούν τις πραγματικές τιμές των βαθμών αποφοίτησης μεγαλύτερων του 8. Αυτό γίνεται αντιληπτό και από τις προβλεπόμενες τιμές του βαθμού αποφοίτησης για τα πρώτα 2 έτη αλλά και από τα διαστήματα εμπιστοσύνης των 2 προβλέψεων, με το πρώτο διάστημα εμπιστοσύνης να έχει πάνω όριο 8.5, μικρότερο δηλαδή της πραγματικής τιμής 8.58. Τα διαστήματα εμπιστοσύνης των προβλέψεων παρουσιάζουν συνεχή βελτίωση, με τις προβλεπόμενες τιμές να βρίσκονται πολύ κοντά με την πραγματική τιμή για τις περιπτώσεις του 4ου και 5ου έτους φοίτησης. Το μικρότερο κάτω όριο των διαστημάτων εμπιστοσύνης ανήκει στην προβλεπόμενη τιμή του 1ου έτους φοίτησης με αυτό να αυξάνεται συνεχώς μέχρι και το 5ο έτος για το οποίο εμφανίζεται το μεγαλύτερο κάτω όριο συγκριτικά με τα υπόλοιπα έτη. Όσον αφορά τα πάνω όρια εμπιστοσύνης, αυτά παρουσιάζουν την μικρότερη τιμή για το 1ο έτος και

οφείλεται στην τάση υποεκτίμησης των βαθμών αποφοίτησης μεγαλύτερων του 8 για εκείνο το έτος, ενώ τη 2η μικρότερη τιμή πάνω ορίου, αλλά ταυτόχρονα μεγαλύτερη της πραγματικής τιμής (με αποτέλεσμα η πραγματική τιμή να βρίσκεται μεταξύ των ορίων σε αντίθεση της περίπτωσης του 1ου έτους), την παρατηρούμε στο 5ο έτος.

Πίνακας 5.8: Προβλεπόμενες τιμές του βαθμού αποφοίτησης ενός τυχαίου αποφοίτου

Τιμές πρόβλεψης βαθμού αποφοίτησης, ενός τυχαίου αποφοίτου, εν συνοδεία των 95% διαστημάτων εμπιστοσύνης τους	
Fit	95% PI
1ο ακαδημαϊκό έτος	
7.69996	(6.90429, 8.49563)
2ο ακαδημαϊκό έτος	
8.22611	(7.65298, 8.79924)
3ο ακαδημαϊκό έτος	
8.4062	(8.04543, 8.76696)
4ο ακαδημαϊκό έτος	
8.62879	(8.37415, 8.88342)
5 ακαδημαϊκό έτος	
8.54022	(8.40707, 8.67337)

Σχήμα 5.6: Διαγραμματική απεικόνιση τιμών πρόβλεψης του βαθμού αποφοίτησης για έναν τυχαίο απόφοιτο



5.3.2.1 Διαγράμματα διασποράς πρόβλεψης διάρκειας σπουδών

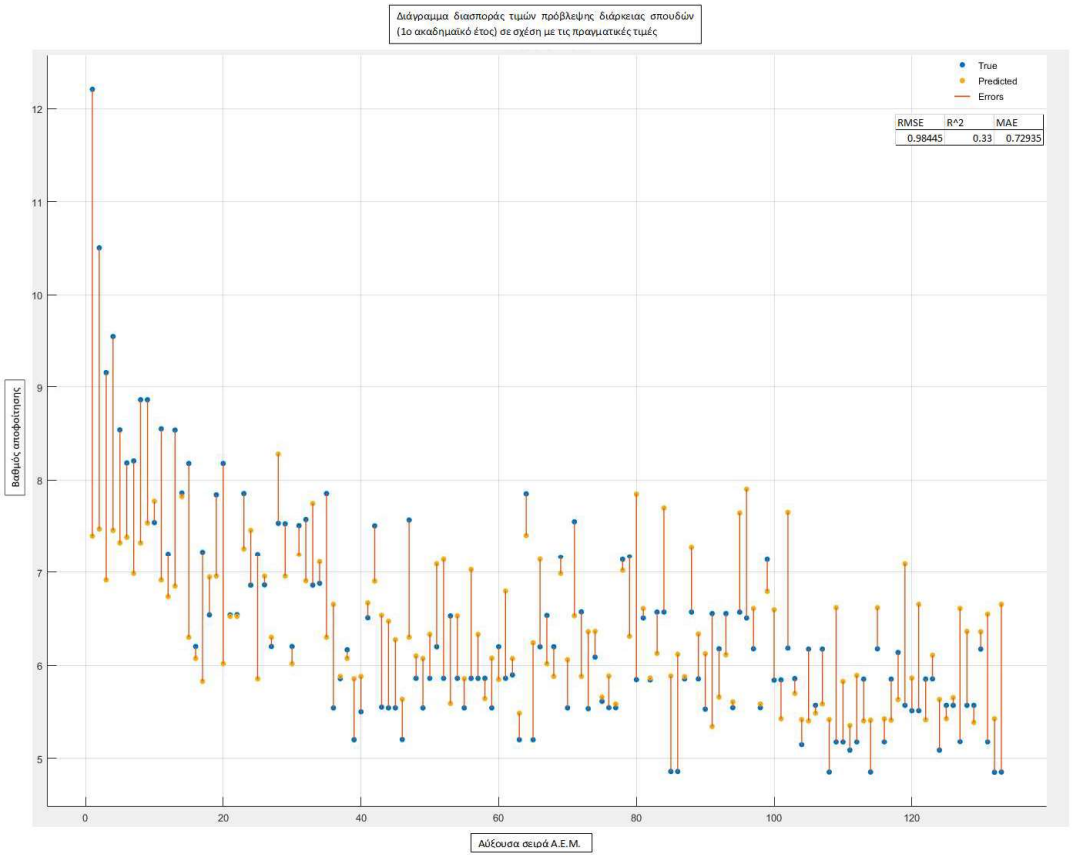
Σε αυτό το σημείο, να γίνει υπενθύμιση πως ως διάρκεια σπουδών δίνεται η χρονολογική διάρκεια σπουδών για τον καλύτερο διαχωρισμό αποφοίτων οι οποίοι δεν εντασσόταν στις περιπτώσεις αποφοίτησης στις καθιερωμένες περιόδους ορκωμοσίας. Στο σχήμα 5.7 παρουσιάζεται το διάγραμμα διασποράς των τιμών πρόβλεψης της διάρκειας σπουδών έναντι των πραγματικών τιμών αξιοποιώντας δεδομένα του 1ου έτους φοίτησης. Βασική παρατήρηση αποτελεί ο πολύ χαμηλός R^2=0.33 και το υψηλό MAE=0.72935. Το μόνο θετικό από την εν λόγω εικόνα της σύγκρισης της διασποράς των τιμών είναι πως υπάρχει σαφής διαχωρισμός των υψηλών τιμών της διάρκειας σπουδών έναντι των χαμηλών, έχοντας υψηλότερες και χαμηλότερες τιμές πρόβλεψης αντίστοιχα εκτός ορισμένων εξαιρέσεων. Όπως ήταν αναμενόμενο σύμφωνα από τις έως τώρα αναλύσεις, τα σφάλματα πρόβλεψης της διάρκειας σπουδών με τη χρήση των δεδομένων του 1ου έτους ήταν σημαντικά υψηλά.

Μελετώντας την εικόνα του σχήματος 5.8 μπορούμε να διακρίνουμε σημαντική βελτίωση στην ικανότητα της πρόβλεψης της διάρκειας σπουδών με την αξιοποίηση δεδομένων του 2ου έτους φοίτησης. Ο διαχωρισμός των υψηλών τιμών πρόβλεψης έναντι των χαμηλών είναι πλέον ακόμα πιο σαφής, αλλά τα υψηλά σφάλματα πρόβλεψης παραμένουν, άλλωστε ο εκτιμώμενος R^2=0.49 είναι πολύ χαμηλός και το MAE=0.60013 σημαντικά υψηλό.

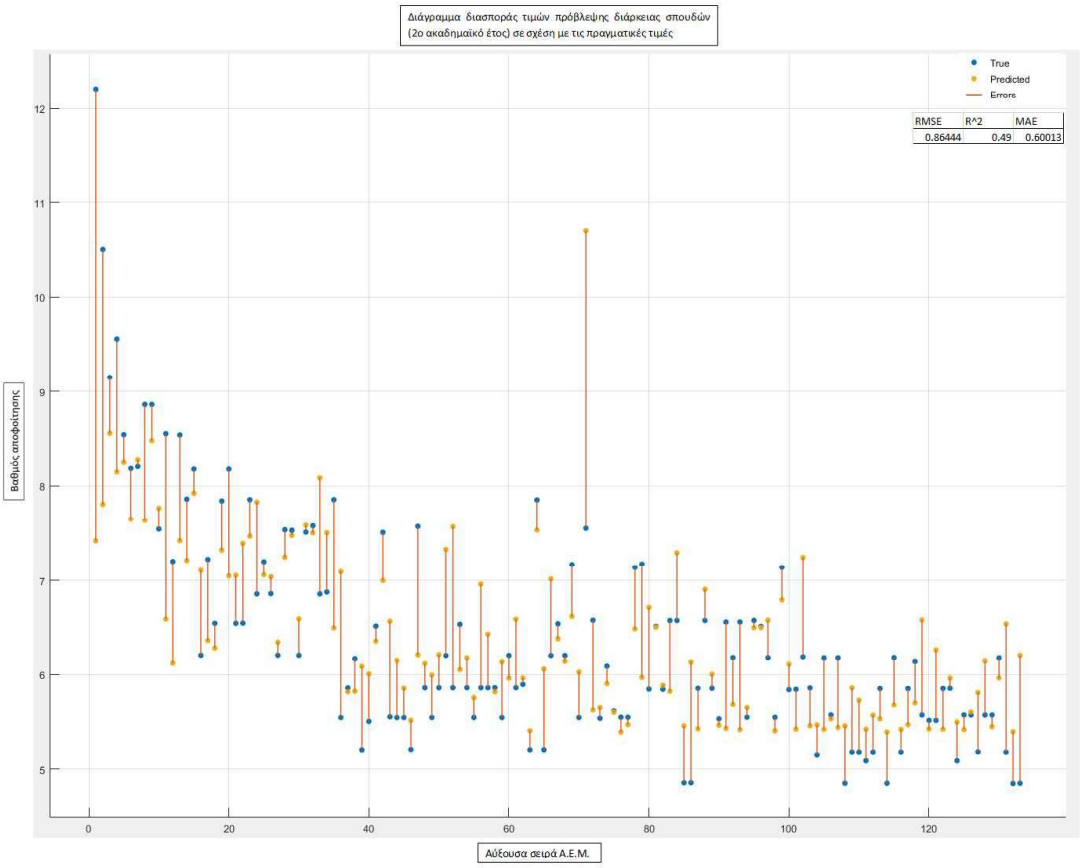
Λόγω της υπερβολικής προσαρμογής για το 4ο έτος φοίτησης, το μοντέλο πρόβλεψης του 4ου έτους θα είναι το ίδιο με αυτό του 3ου. Για αυτά τα έτη φοίτησης φαίνεται το διάγραμμα διασποράς τιμών πρόβλεψης έναντι πραγματικών τιμών στο σχήμα 5.9. Η ικανότητα πρόβλεψης της διάρκειας σπουδών συνεχίζει να μην είναι επαρκής, ειδικά αν αναλογιστεί κανείς πως για τα 4 πρώτα χρόνια φοίτησης υπάρχει το μεγαλύτερο ενδιαφέρον πρόβλεψης με την υποχρεωτική φοίτηση να είναι 5 χρόνια. Ο εκτιμώμενος R^2 είναι 0.59 και το MAE είναι 0.53957 με τα σφάλματα πρόβλεψης να είναι αρκετά μεγάλα για την πλειοψηφία των τιμών.

Αν και πλέον αργά, καταλήγοντας στο τέλος της υποχρεωτικής φοίτησης, η ικανότητα πρόβλεψης της διάρκειας σπουδών είναι για πρώτη φορά ικανοποιητική όπως και φαίνεται στο σχήμα 5.10. Για το μοντέλο πρόβλεψης του έτους αυτού ο εκτιμώμενος R^2 είναι 0.75 και το MAE=0.47738. Για πρώτη φορά υπήρξε σημαντική βελτίωση στις προβλέψεις υψηλών τιμών διάρκειας σπουδών, ενώ αν και αρκετοί ήταν οι φοιτητές οι οποίοι είχαν ολοκληρώσει τις φοιτητικές τους υποχρεώσεις, τα σφάλματα πρόβλεψης για μικρές τιμές διάρκειας σπουδών παραμένουν σε μη ικανοποιητικό επίπεδο. Ανά τα έτη, για τις υψηλές διάρκειες σπουδών υπήρξε τάση υποεκτίμησης και για τις χαμηλές τάση υπερεκτίμησης, με τις τάσεις αυτές να παραμένουν μέχρι και το 5ο έτος αν και σε μικρότερο βαθμό.

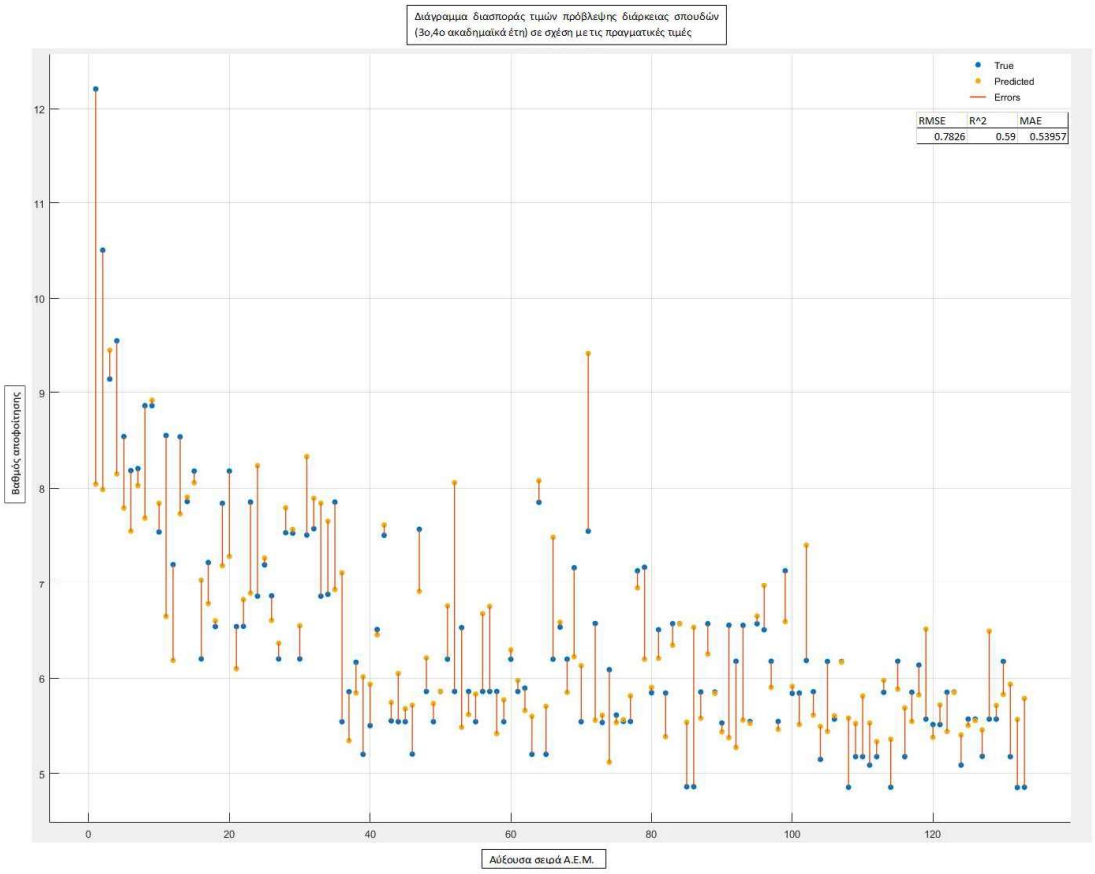
Σχήμα 5.7: Διάγραμμα διασποράς τιμών πρόβλεψης της διάρκειας σπουδών σε σχέση με τις πραγματικές τιμές με χρήση δεδομένων του 1^{ου} έτους φοίτησης



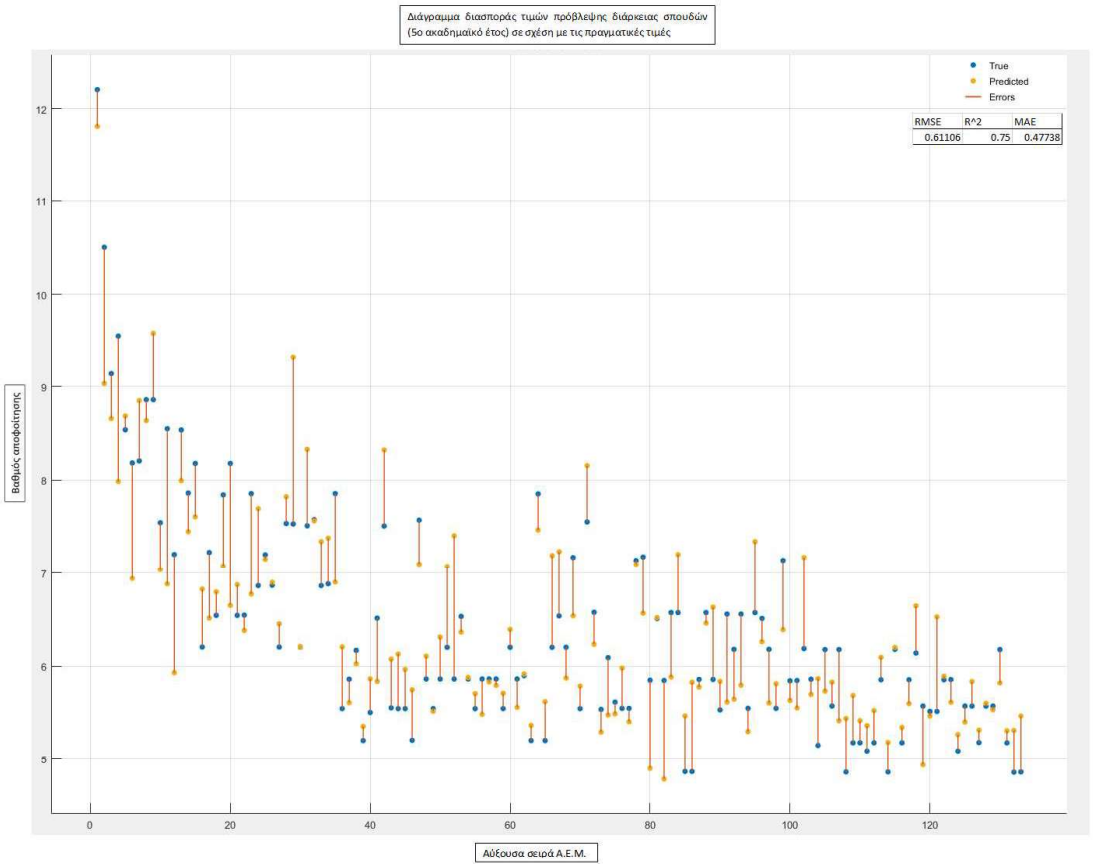
Σχήμα 5.8: Διάγραμμα διασποράς τιμών πρόβλεψης της διάρκειας σπουδών σε σχέση με τις πραγματικές τιμές με χρήση δεδομένων του 2^{ου} έτους φοίτησης



Σχήμα 5.9: Διάγραμμα διασποράς τιμών πρόβλεψης της διάρκειας σπουδών σε σχέση με τις πραγματικές τιμές με χρήση δεδομένων του 3^{ου} έτους φοίτησης



Σχήμα 5.10: Διάγραμμα διασποράς τιμών πρόβλεψης της διάρκειας σπουδών σε σχέση με τις πραγματικές τιμές με χρήση δεδομένων του 5ου έτους φοίτησης



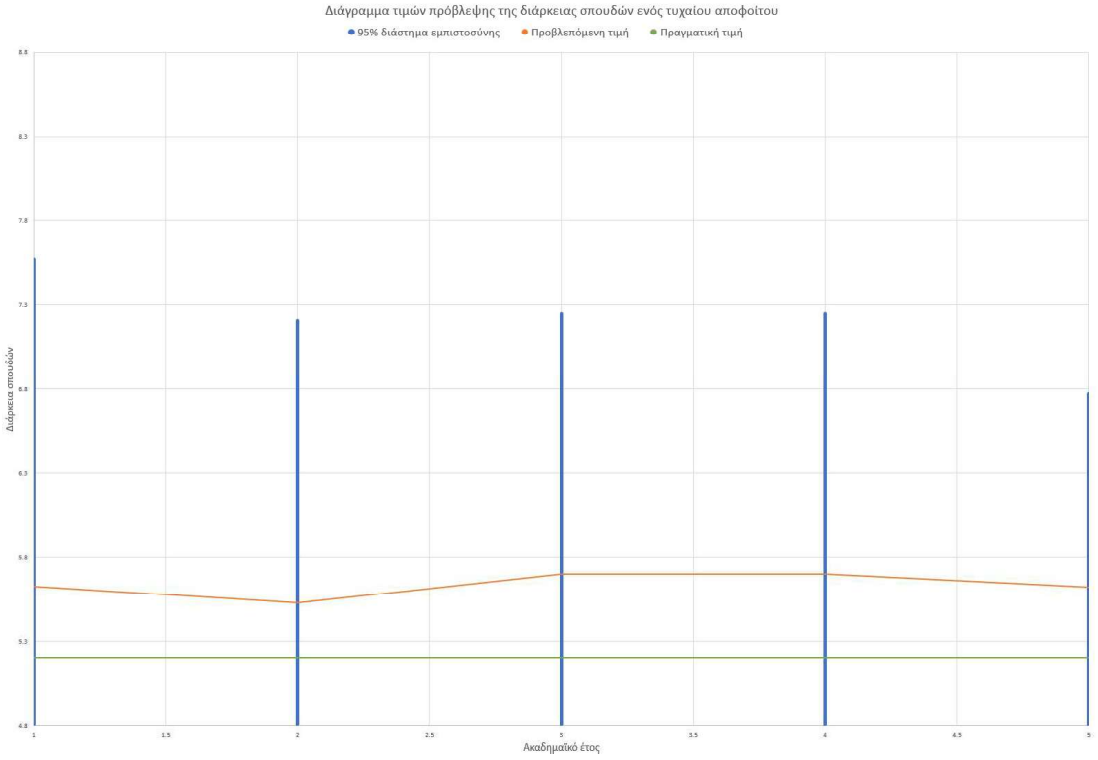
5.3.2.2 Αξιολόγηση αποτελεσμάτων πρόβλεψης διάρκειας σπουδών ενός τυχαίου αποφοίτου

Στην υποενότητα αυτή θα προχωρήσουμε στην αξιολόγηση των αποτελεσμάτων πρόβλεψης της διάρκειας σπουδών για τον απόφοιτο τον οποίο επιλέχθηκε και στην υποενότητα 5.3.1.1. Ο απόφοιτος αυτός είχε πραγματική τιμή διάρκειας σπουδών 5.205 και στον πίνακα 5.9 φαίνονται οι τιμές πρόβλεψης της διάρκειας σπουδών του ανά τα έτη μαζί με τα 95% διαστήματα εμπιστοσύνης τους. Τα αποτελέσματα του πίνακα 5.9 παρουσιάζονται διαγραμματικά στο σχήμα 5.11. Η πρώτη παρατήρηση η οποία προκύπτει από την μελέτη του πίνακα 5.9 είναι πως οι προβλεπόμενες τιμές της διάρκειας σπουδών δεν παρουσιάζουν μεγάλη απόκλιση μεταξύ τους, με την πιο κοντινή στην πραγματική τιμή να είναι αυτή του 2ου έτους. Παρ' όλα αυτά, σημαντική βελτίωση παρατηρήθηκε στα διαστήματα εμπιστοσύνης, με τα πάνω όρια να ελαττώνονται με την αύξηση των ετών και την πραγματική τιμή να εμπεριέχεται σε όλα τα διαστήματα εμπιστοσύνης. Υπενθυμίζεται πως ως διάρκεια σπουδών ορίστηκε η χρονολογική διάρκεια σπουδών, οπότε επιλέχθηκε τα κάτω όρια εμπιστοσύνης να έχουν την τιμή 4.8.

Πίνακας 5.9: Προβλεπόμενες τιμές της διάρκειας σπουδών ενός τυχαίου αποφοίτου

Τιμές πρόβλεψης διάρκειας σπουδών, ενός τυχαίου αποφοίτου, εν συνοδεία των 95% διαστημάτων εμπιστοσύνης τους	
Fit	95% PI
1ο ακαδημαϊκό έτος	
5.62811	(4.8, 7.56831)
2ο ακαδημαϊκό έτος	
5.52957	(4.8, 7.20743)
3ο ακαδημαϊκό έτος	
5.70083	(4.8, 7.24450)
4ο ακαδημαϊκό έτος	
5.70083	(4.8, 7.24450)
5ο ακαδημαϊκό έτος	
5.6231	(4.8, 6.77272)

Σχήμα 5.11: Διαγραμματική απεικόνιση τιμών πρόβλεψης της διάρκειας σπουδών για έναν τυχαίο απόφοιτο



5.4 Σχόλια

Συνοψίζοντας, στο 5° κεφάλαιο έγινε διερεύνηση κατάλληλων μοντέλων πρόβλεψης του βαθμού αποφοίτησης και της διάρκειας σπουδών. Αρχικά, επιλέγοντας υποσύνολα ανεξάρτητων μεταβλητών με τη μέθοδο της προς τα πίσω απαλοιφής και συγκρίνοντας τα μοντέλα μέσω επαναλαμβανόμενης διασταυρούμενης επικύρωσης τμημάτων, διαπιστώσαμε πως περιπλοκότερα μοντέλα οδηγούν σε καλύτερες προβλέψεις. Στην περίπτωση της πρόβλεψης του βαθμού αποφοίτησης, τα μοντέλα πρόβλεψης είχαν ικανοποιητικές προβλέψεις σε αντίθεση με αυτά της διάρκειας σπουδών. Πιο συγκεκριμένα για την πρόβλεψη της διάρκειας σπουδών, παρατηρήθηκε στο 4° έτος φοίτησης υπερβολική προσαρμογή δεδομένων με αποτέλεσμα να χρησιμοποιηθεί το ίδιο μοντέλο πρόβλεψης με αυτό του 3^{ου} έτους. Εξαιτίας του παραπάνω και σε συνδυασμό με τις χαμηλές προβλεπτικές ικανότητες της διάρκειας σπουδών ανά τα έτη, δεν ενδείκνυται η πρόβλεψή της, τουλάχιστον όχι με τη συγκεκριμένη μέθοδο. Συνεπώς, θα ήταν σημαντικό να διερευνηθούν εναλλακτικές μέθοδοι πρόβλεψης της διάρκειας σπουδών, ειδικά για τα χαμηλότερα έτη φοίτησης (1^{ος} κύκλος σπουδών) για τα οποία οι προβλέψεις παρουσιάζουν μεγαλύτερο ενδιαφέρον.

Κεφάλαιο 6: Μοντέλα παλινδρόμησης με όρους αλληλεπιδράσεων

Στο 5° κεφάλαιο διαπιστώσαμε πως ο βαθμός αποφοίτησης μπορεί να προβλεφθεί ικανοποιητικά χρησιμοποιώντας μοντέλα πολλαπλής γραμμικής παλινδρόμησης, ενώ τα μοντέλα πρόβλεψης της διάρκειας σπουδών παρουσίασαν χαμηλές προβλεπτικές ικανότητες. Παρ' όλα αυτά παρατηρήθηκαν αδυναμίες στα μοντέλα πρόβλεψης του βαθμού αποφοίτησης, ειδικά για τα πρώτα 2 ακαδημαϊκά έτη για τα οποία υπήρξαν σημαντικές τάσεις υποεκτίμησης και υπερεκτίμησης. Στο κεφάλαιο αυτό διερευνάται το κατά πόσο οι αλληλεπιδράσεις μεταξύ των μεταβλητών Π.Π.χ.ν με Μ.Μ.χ.ν (για ίδιο χ και ν) και Π.Π.1-ν με Μ.Μ.1-ν (για ίδιο ν) μπορούν να βελτιώσουν τυχόν αδυναμίες των μοντέλων.

Για τη διάρκεια σπουδών, οι προβλέψεις έπειτα από την προσθήκη των αλληλεπιδράσεων και ακολουθώντας τη μεθοδολογία που παρουσιάστηκε στο κεφάλαιο 5, παρέμειναν σε μη ικανοποιητικά επίπεδα και δεν θα γίνει ανάλυση των αποτελεσμάτων. Όσον αφορά τον βαθμό αποφοίτησης, οι προβλεπτικές ικανότητες των μοντέλων του 3^{ου}, του 4^{ου} και του 5^{ου} έτους παρέμειναν σε παρόμοια επίπεδα έπειτα από την προσθήκη των αλληλεπιδράσεων, ενώ κάτι τέτοιο δεν ισχύει για τα πρώτα δύο έτη φοίτησης. Για αυτά τα έτη και για την πρόβλεψη του βαθμού αποφοίτησης θα γίνει συνοπτική ανάλυση των αποτελεσμάτων με τη μεθοδολογία που ακολουθήθηκε στο 5° κεφάλαιο, λαμβάνοντας πλέον υπόψιν και τις αλληλεπιδράσεις.

6.1 Αποτελέσματα προς τα πίσω απαλοιφής

Σε αυτό το σημείο είναι σημαντικό να σημειωθεί πως αν και οι εξεταζόμενες μεταβλητές εισόδου αυξήθηκαν κατά 50% (με την προσθήκη των ορισμένων αλληλεπιδράσεων) σε σύγκριση με αυτές του 5^{ου} κεφαλαίου, η μέθοδος της προς τα πίσω απαλοιφής όρισε ως σημαντικές μεταβλητές παρόμοιο αριθμό μεταβλητών και στα 10 μοντέλα παλινδρόμησης με αυτά του 5^{ου} κεφαλαίου. Αναφορικά για την πρόβλεψη του βαθμού αποφοίτησης, ορίστηκαν ως σημαντικές μεταβλητές εισόδου για το 1° έτος φοίτησης η μεταβλητή Π.1.1 και η αλληλεπίδραση Π.1.1*Μ.1.1, ενώ για το 2° έτος φοίτησης οι μεταβλητές Π.1.1, Π.1.2, Π.1-2 και Μ.2.2 και οι αλληλεπιδράσεις Π.1.2*Μ.1.2, Π.2.2*Μ.2.2 και Π.1-2*Μ.1-2. Έτσι λοιπόν ορίστηκαν ως σημαντικές 2 και 7 μεταβλητές εισόδου αντίστοιχα, ακριβώς ο ίδιος αριθμός μεταβλητών με τα μοντέλα που παρουσιάστηκαν στο 5° κεφάλαιο.

Τα αποτέσματα του μοντέλου της προς τα πίσω απαλοιφής, λαμβάνοντας υπόψιν τις ορισμένες εξεταζόμενες αλληλεπιδράσεις, δίνονται στον πίνακα 6.1 σε σύγκριση με τα μοντέλα του 5^{ου} κεφαλαίου για τα πρώτα 2 έτη φοίτησης. Έπειτα από σύγκριση των μοντέλων, μπορούμε να διακρίνουμε την υπεροχή αυτών με τις αλληλεπιδράσεις σε όλα τα αποτελέσματα του πίνακα. Πιο συγκεκριμένα τα τυπικά σφάλματα είναι μικρότερα όπως και οι όροι PRESS, AIC_C και BIC, ενώ ταυτόχρονα παρουσιάζουν υψηλότερους συντελεστές προσδιορισμού.

Πίνακας 6.1: Χρήσιμοι συντελεστές για την αξιολόγηση των αποτελεσμάτων της παλινδρόμησης του βαθμού αποφοίτησης με ή χωρίς αλληλεπιδράσεις

	Αξιολόγηση αποτελεσμάτων της προς τα πίσω απαλοιφής, για την πρόβλεψη του βαθμού αποφοίτησης με την προσθήκη αλληλεπιδράσεων						
	S	R-sq	R-sq(adj)	PRESS	R-sq(pred)	AICc	BIC
	1ο ακαδημαϊκό έτος						
Όλες οι ορισμένες μεταβλητές και αλληλεπιδράσεις	0.362746	63.21%	62.64%	18.0755	61.12%	112.98	124.23
Όλες οι μεταβλητές χωρίς τις αλληλεπιδράσεις	0.3978	55.75%	55.07%	22.2868	52.06%	137.52	148.76
	2ο ακαδημαϊκό έτος						
Όλες οι ορισμένες μεταβλητές και αλληλεπιδράσεις	0.259779	81.86%	80.84%	9.68607	79.17%	30.1	54.65
Όλες οι μεταβλητές χωρίς τις αλληλεπιδράσεις	0.280402	78.86%	77.68%	11.6587	74.92%	50.42	74.97

6.2 Επαναλαμβανόμενη διασταυρούμενη επικύρωση 10 τμημάτων

Συνεχίζοντας την ανάλυση των μοντέλων με την προσθήκη των ορισμένων αλληλεπιδράσεων, επόμενο βήμα αποτελεί η επαναλαμβανόμενη διασταυρούμενη επικύρωση 10 τμημάτων των RMSE. Από τη μελέτη του πίνακα 6.2 παρατηρούμε σημαντική μείωση των RMSE, με τα μοντέλα τα οποία δεν συμπεριέλαβαν αλληλεπιδράσεις μεταβλητών να παρουσιάζουν 11% και 10% μεγαλύτερα μέσα RMSE για το 1° και 2° έτος αντίστοιχα.

Πίνακας 6.2: Επαναλαμβανόμενη επικύρωση 10 τμημάτων των σφαλμάτων για την πρόβλεψη του βαθμού αποφοίτησης με ή χωρίς προσθήκη αλληλεπιδράσεων

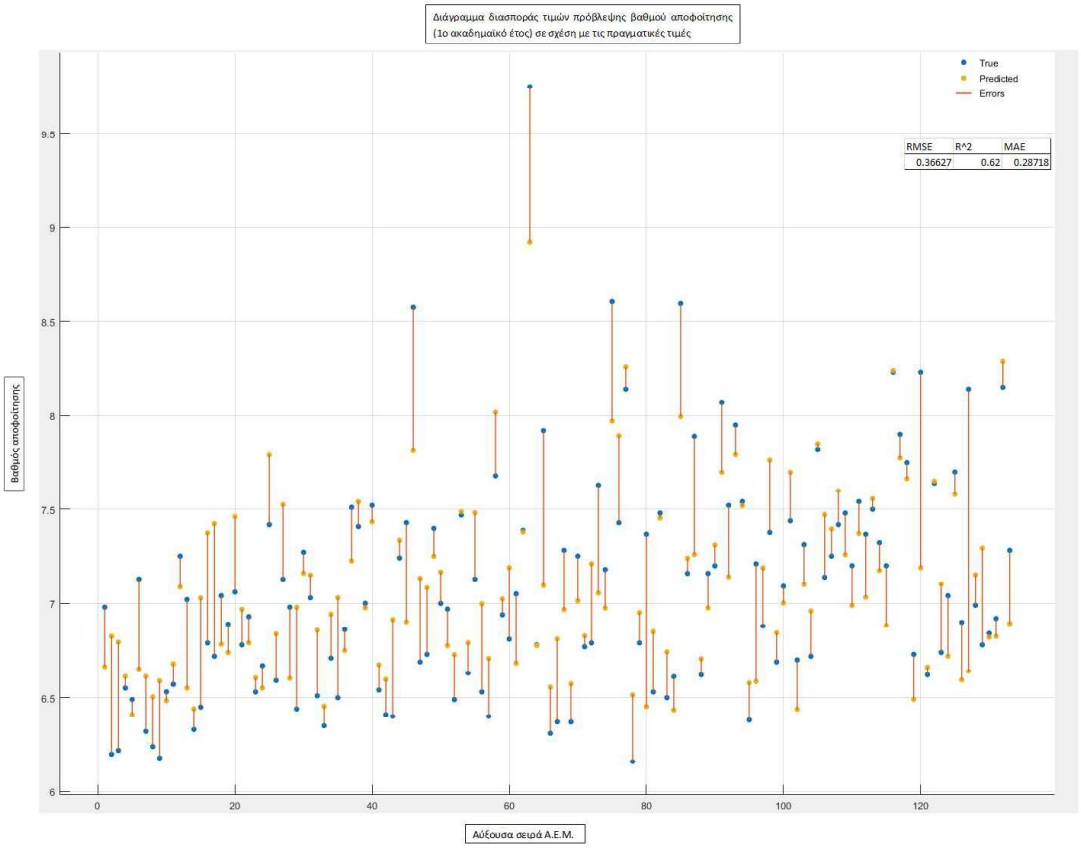
Ακαδημαϊκό έτος/Μοντέλα πρόβλεψης βαθμού αποφοίτησης	RMSE 1ης επικύρωσης 10 τμημάτων	RMSE 2ης επικύρωσης 10 τμημάτων	RMSE 3ης επικύρωσης 10 τμημάτων	RMSE 4ης επικύρωσης 10 τμημάτων	RMSE 5ης επικύρωσης 10 τμημάτων	RMSE 6ης επικύρωσης 10 τμημάτων	RMSE 7ης επικύρωσης 10 τμημάτων	RMSE 8ης επικύρωσης 10 τμημάτων	RMSE 9ης επικύρωσης 10 τμημάτων	RMSE 10ης επικύρωσης 10 τμημάτων	Μέσος όρος RMSE	Ποσοστιαία διαφορά RMSE μοντέλων ανά έτος (1=100%)
1ο Ακαδημαϊκό έτος												
Γραμμική παλινδρόμηση (με την προσθήκη της αλληλεπίδρασης των ορισμένων μεταβλητών εισόδου)	0.3663	0.3687	0.3685	0.3691	0.3678	0.3716	0.3712	0.3670	0.3675	0.3712	0.3689	
Γραμμική παλινδρόμηση	0.4075	0.4076	0.4157	0.4104	0.4043	0.4115	0.4188	0.4104	0.4115	0.4089	0.4106	0.1132
2ο Ακαδημαϊκό έτος												
Γραμμική παλινδρόμηση (εξετάζοντας όλες τις ορισμένες μεταβλητές εισόδου και τις ορισμένες αλληλεπιδράσεις τους)	0.2662	0.2714	0.2698	0.2725	0.2714	0.2699	0.2718	0.2687	0.2674	0.2695	0.2699	
Γραμμική παλινδρόμηση (εξετάζοντας όλες τις ορισμένες μεταβλητές εισόδου)	0.2950	0.3034	0.2990	0.3011	0.2945	0.2944	0.2952	0.3005	0.2920	0.2967	0.2972	0.1013

Για τα πρώτα 2 έτη φοίτησης τα μοντέλα τα οποία συμπεριέλαβαν τις ορισμένες αλληλεπιδράσεις των ανεξάρτητων μεταβλητών μας οδήγησαν σε χαμηλότερα εκτιμώμενα σφάλματα. Επόμενο βήμα αποτελεί η αξιολόγηση της εικόνας των διαγραμμάτων διασποράς των προβλεπόμενων τιμών έναντι των πραγματικών. Στο σχήμα 6.1 παρουσιάζεται το διάγραμμα διασποράς για το μοντέλο του 1ου έτους φοίτησης. Από τη μελέτη του εν λόγω διαγράμματος και από τη σύγκρισή του με αυτό του μοντέλου του 5ου κεφαλαίου, μπορούμε να διακρίνουμε πως ενώ για τους χαμηλότερους βαθμούς αποφοίτησης δεν υπήρξε κάποια βελτίωση όσον αφορά την τάση υπερεκτίμησης, υπήρξε καθαρή βελτίωση σχετικά με την τάση υποεκτίμησης των μεγαλύτερων βαθμών. Συγχρόνως, υπήρξε σημαντική αύξηση του R^2 από 0.53 σε 0.62 και μείωση του MAE από 0.30646 σε 0.28718. Στο σχήμα 6.2 δίνονται διαγραμματικά τα σφάλματα πρόβλεψης των δύο μοντέλων για την ευκολότερη σύγκρισή τους.

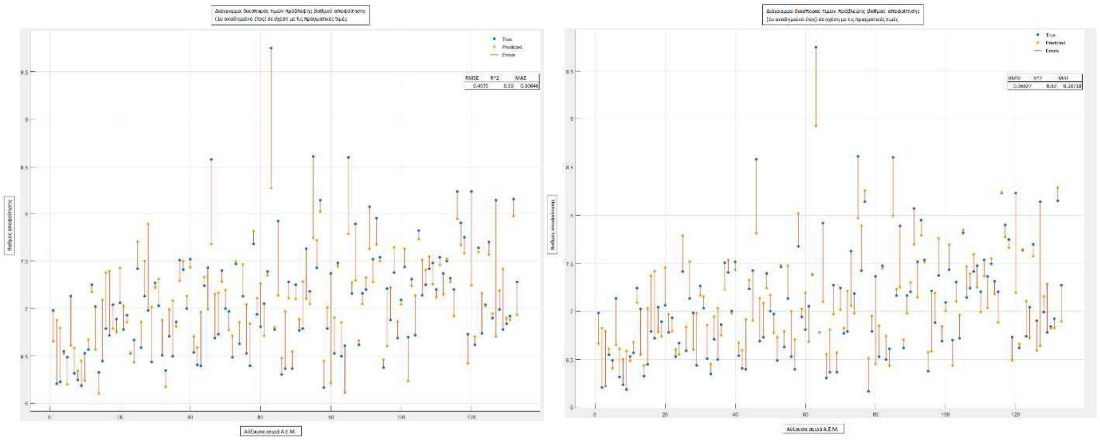
Αντίστοιχα καλύτερη εικόνα παρουσιάζει και το μοντέλο του 2ου έτους έναντι αυτό του 5ου κεφαλαίου, όπως και φαίνεται στο σχήμα 6.3. Ο συντελεστής προσδιορισμού παρουσίασε αύξηση σε 0.80 από 0.75 και το MAE μείωση από 0.23627 σε 0.22098. Τέλος φαίνεται να αντιμετωπίστηκε η τάση υποεκτίμησης των βαθμολογιών μεγαλύτερων του 8, αν και υπήρξε σχετικά μεγαλύτερη συχνότητα υπερεκτιμήσεων των μικρότερων βαθμολογιών. Η διαγραμματική σύγκριση των μοντέλων γίνεται μέσω του σχήματος 6.4.

Συνεπώς, αν και η προσθήκη των αλληλεπιδράσεων δεν βελτίωσε τις προβλεπτικές ικανότητες των μοντέλων στην πρόβλεψη του βαθμού αποφοίτησης για το 3ο έτος φοίτησης και έπειτα, βελτιώθηκαν σημαντικά τα συστηματικά σφάλματα των πρώτων δύο ετών. Ταυτόχρονα, όπως και αναφέρθηκε, τα μοντέλα πρόβλεψης εμπεριείχαν παρόμοιο αριθμό ανεξάρτητων μεταβλητών έπειτα από την εφαρμογή του αλγορίθμου της προς τα πίσω απαλοιφής, έχοντας ως αποτέλεσμα να μην οδηγηθούμε σε περιπλοκότερα μοντέλα με την προσθήκη των ορισμένων αλληλεπιδράσεων.

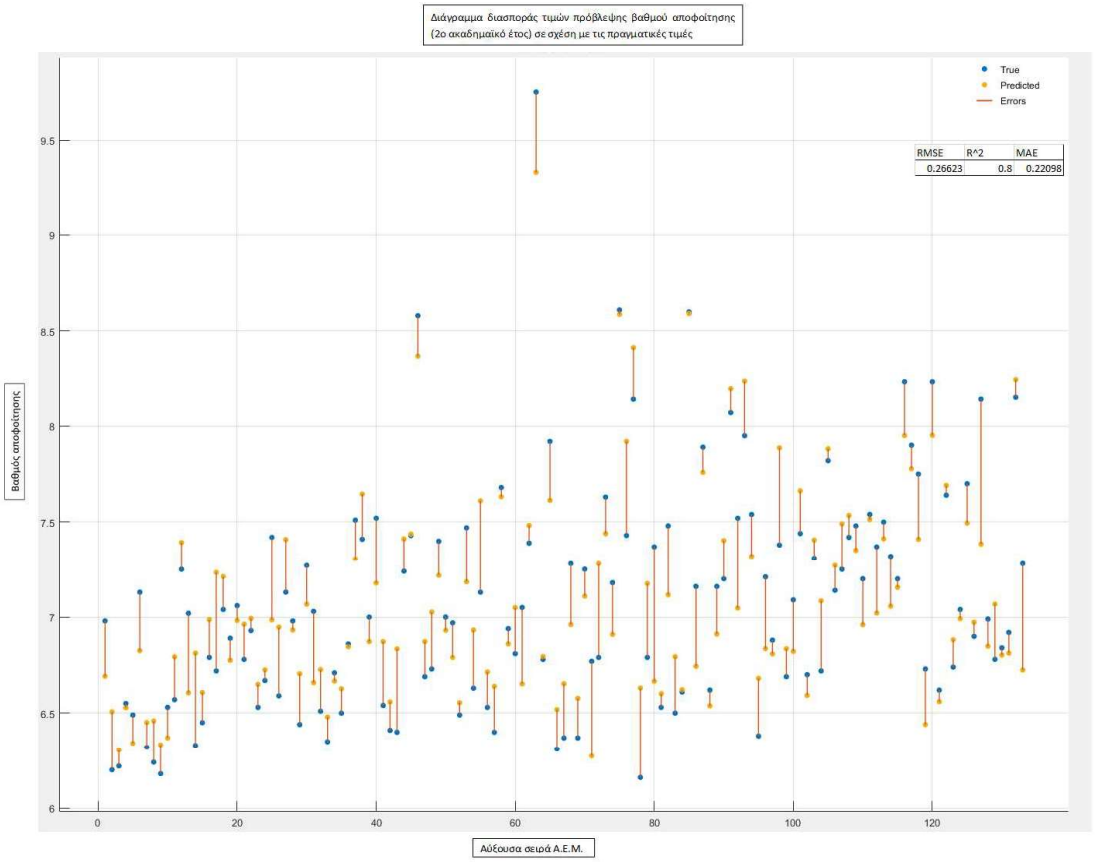
Σχήμα 6.1: Σφάλματα πρόβλεψης του βαθμού αποφοίτησης με χρήση δεδομένων του 1ου έτους φοίτησης και την προσθήκη αλληλεπίδρασης των μεταβλητών εισόδου



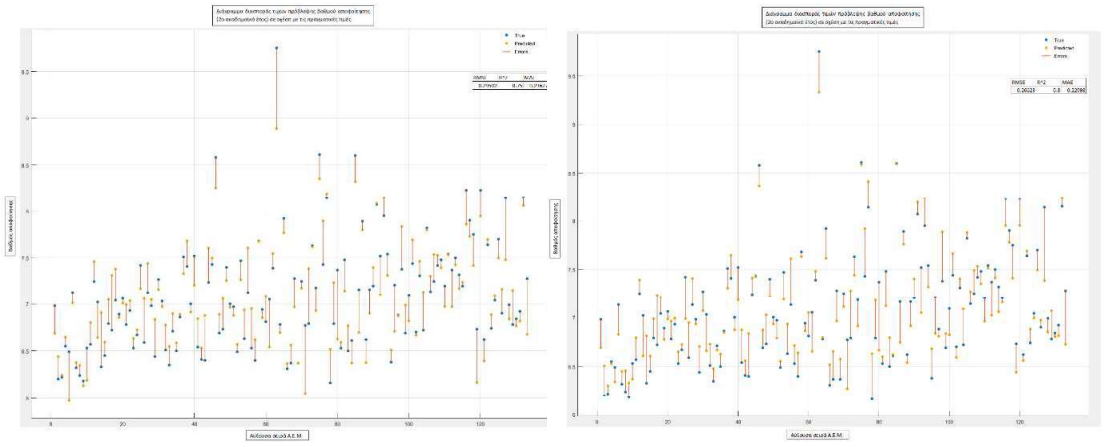
Σχήμα 6.2: Σύγκριση μοντέλων 1ου έτους του 5ου κεφαλαίου αριστερά και του 6ου δεξιά



Σχήμα 6.3: Σφάλματα πρόβλεψης του βαθμού αποφοίτησης με χρήση δεδομένων του 2ου έτους φοίτησης και την προσθήκη αλληλεπιδράσεων των μεταβλητών εισόδου



Σχήμα 6.4: Σύγκριση μοντέλων 2ου έτους του 5ου κεφαλαίου αριστερά και του 6ου δεξιά



Κεφάλαιο 7: Συμπεράσματα και προτάσεις για μελλοντική έρευνα

Ανακεφαλαιώνοντας, στην παρούσα εργασία έγινε μελέτη της ακαδημαϊκής επίδοσης των φοιτητών του Τμήματος Μηχανολόγων Μηχανικών του Πανεπιστημίου Δυτικής Μακεδονίας. Αρχικά έγινε ανάλυση του κατά πόσο διαφέρουν οι βαθμοί των μαθημάτων του τμήματος μεταξύ τους και στη συνέχεια έγινε διερεύνηση συσχετίσεων. Καταλήξαμε στο συμπέρασμα πως υπήρξαν ισχυρές συσχετίσεις μεταξύ του βαθμού αποφοίτησης και της διάρκειας σπουδών και με το κατά πόσο υπήρξε καθυστέρηση στην επιτυχή εξέταση των μαθημάτων. Οι βαθμοί διπλωματικής δεν παρουσίασαν σημαντική συσχέτιση με τον βαθμό αποφοίτησης αλλά και ούτε με τη διάρκεια σπουδών.

Συνεχίζοντας με την ανάλυση σχέσεων εξάρτησης, οδηγηθήκαμε στο συμπέρασμα πως τα ποσοστά περασμένων μαθημάτων και οι βαθμολογίες τους φαίνεται να αποτελούν χρήσιμες ανεξάρτητες μεταβλητές για την πρόβλεψη του βαθμού αποφοίτησης και λιγότερο για τη διάρκεια σπουδών. Έπειτα από σύγκριση και αξιολόγηση μοντέλων πρόβλεψης φάνηκε πως η διάρκεια σπουδών δεν μπορεί να προβλεφθεί ικανοποιητικά, ενώ τα μοντέλα πρόβλεψης του βαθμού αποφοίτησης παρουσίασαν ικανοποιητικές προβλεπτικές ικανότητες.

Αναλυτικά για την πρόβλεψη της διάρκεια σπουδών, διακρίναμε υψηλά εκτιμώμενα σφάλματα μελλοντικών προβλέψεων, ενώ αντιμετωπίστηκαν θέματα υπερπροσαρμογής δεδομένων. Όσον αφορά τον βαθμό αποφοίτησης, τα μοντέλα πρόβλεψης παρουσίασαν ικανοποιητικά εκτιμώμενα σφάλματα ανά τα έτη φοίτησης και ταυτόχρονα μετριάστηκε το πρόβλημα υποεκτίμησης υψηλών βαθμολογιών στα μοντέλα των πρώτων 2 ετών φοίτησης με την προσθήκη αλληλεπιδράσεων των ανεξάρτητων μεταβλητών.

Τέλος, σημαντικό είναι να αναφερθεί πως σκοπό της εργασίας αποτελούσε η πρόταση μίας κοινής μεθοδολογίας πρόβλεψης της ακαδημαϊκής επίδοσης των φοιτητών ανά τα έτη φοίτησης. Συνεπώς, τα μοντέλα τα οποία χρησιμοποιήθηκαν δεν ήταν πάντα βέλτιστα για κάθε έτος φοίτησης. Με τη χρήση αλλά και σύγκριση περαιτέρω μεθόδων μηχανικής μάθησης θα μπορούσε να επιτευχθεί ο εντοπισμός των καταλληλότερων μοντέλων για κάθε περίπτωση. Χαρακτηριστικές περιπτώσεις μεθόδων πρόβλεψης της μηχανικής μάθησης αποτελούν οι αλγόριθμοι των Support Vector Machines και της Gaussian Process Regression, οι οποίες κατά περίπτωση μπορεί να οδηγούν σε καλύτερα αποτελέσματα για κάποια έτη φοίτησης. Στο παράρτημα της εργασίας παρουσιάζονται ενδεικτικά κάποια επιμέρους αποτελέσματα εφαρμογής των εν λόγω μεθόδων, οι οποίες όμως δεν επιλέχθηκαν τελικά ως η προτεινόμενη μεθοδολογία πρόβλεψης για τους λόγους που προαναφέρθηκαν. Ταυτόχρονα, λόγω της σχέσης των ανεξάρτητων μεταβλητών με τις εξαρτημένες, δεν θεωρείται απαραίτητο να εξεταστούν ακόμα πιο ευέλικτες μη γραμμικές τεχνικές παλινδρόμησης όπως η πολυωνυμική παλινδρόμηση και οι Regression Splines.

Συμπερασματικά, τα αποτελέσματα της εργασίας θα μπορούσαν να χρησιμοποιηθούν ως ένα εργαλείο αξιολόγησης της επίδοσης των φοιτητών κατά τη διάρκεια σπουδών τους. Χρησιμοποιώντας δεδομένα από περισσότερους αποφοίτους αλλά και με τη διερεύνηση περισσότερων ανεξάρτητων μεταβλητών όπως το φύλο των φοιτητών, την καταγωγή τους, την ποσοστιαία χρονολογική διάρκεια του συνόλου σπουδών κατά την οποία βρίσκονται στην ίδια πόλη με αυτήν την οποία φοιτούν, την επίδοση τους στις πανελλήνιες εξετάσεις και άλλες, ενδεχομένως να οδηγηθούμε σε ικανοποιητικότερα μοντέλα πρόβλεψης.

Βιβλιογραφία

[1] Levene, Howard. (1960). Robust tests for equality of variances.

[2] Douglas C.Montgomery, George C.Runger. (2014). *Applied Statistics And Probability For Engineers*, 6th edition.

[3] Cook, R. Dennis and Sanford Weisberg. 1981. Residuals and Influence in Regression. New York: Chapman and Hall.

[4] Gareth James, Daniela Witten, Robert Tibshirami, Trevor Hastie. 2017. An Introduction to Statistical Learning: with Applications in R, 7th edition.

[5] Burnham, Kenneth P., Anderson, David R. 2002. Model Selection and Multimodel Inference: A practical information-theoretic approach, 2nd edition.

[6] Max Kuhn, Kjell Johnson. 2013. Applied Predictive Modeling, 1st edition.

Παράρτημα

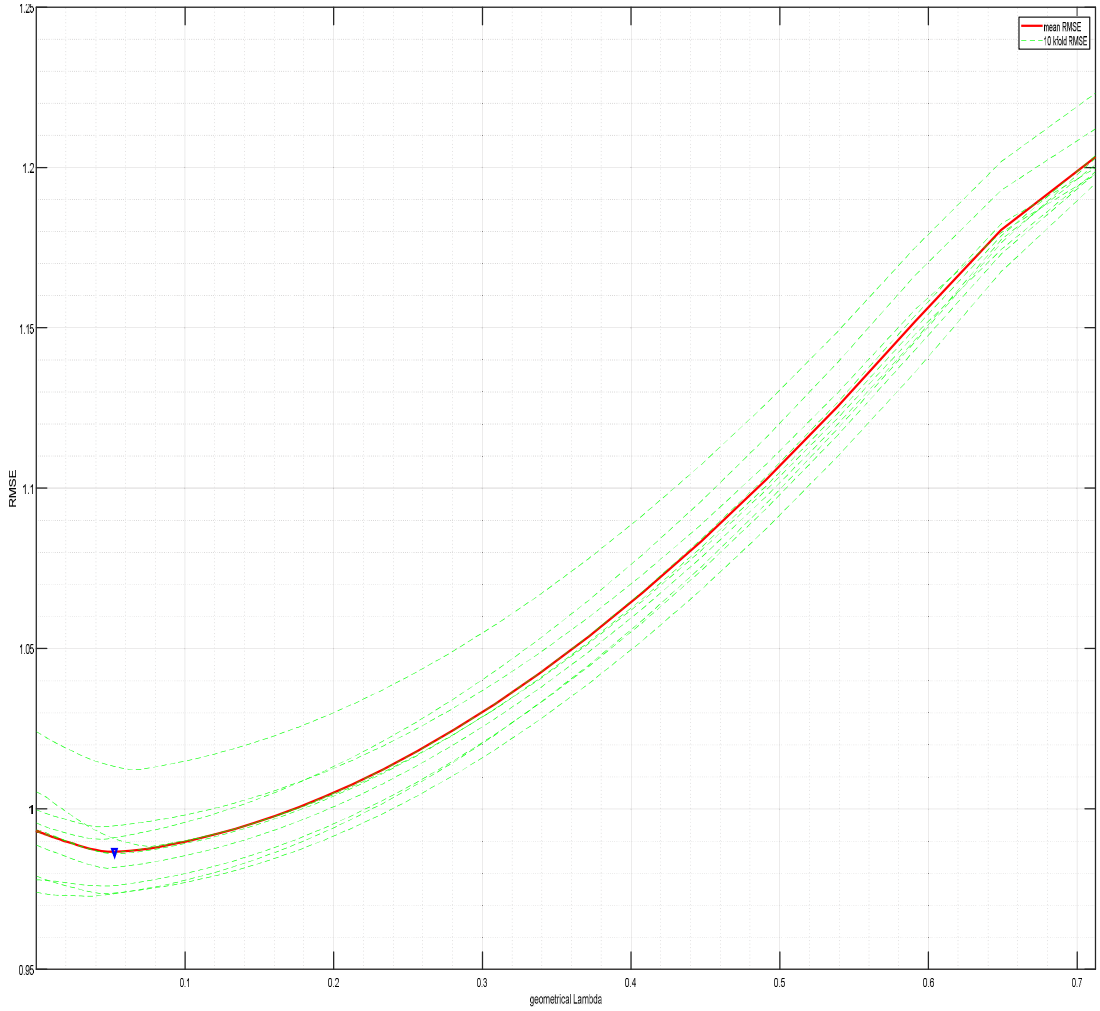
Στο παράρτημα επιλέχθηκε η ανάλυση μεθοδολογιών και αποτελεσμάτων που δεν συμπεριλήφθηκαν στην εργασία. Πιο συγκεκριμένα παρουσιάζονται συνοπτικά ο αλγόριθμος παλινδρόμησης Lasso και η χρήση των αλγορίθμων των Support Vector Machines και της Gaussian Process Regression.

Μία εναλλακτική μέθοδος επιλογής υποσυνόλου μεταβλητών εισόδου από αυτών της βηματικής παλινδρόμησης αποτελεί ο αλγόριθμος παλινδρόμησης Lasso. Το πλεονέκτημα της εν λόγω μεθόδου είναι πως οι συντελεστές παλινδρόμησης συρρικνώνονται αλλά και μηδενίζονται σε αντίθεση με έναν άλλον αλγόριθμο μηχανικής μάθησης τον αλγόριθμο παλινδρόμησης Ridge. Η σταθερά παλινδρόμησης παραμένει ίδια με αυτήν του μοντέλου της γραμμικής παλινδρόμησης και η ποινή εφαρμόζεται στους συντελεστές της παλινδρόμησης. Έτσι ορίζοντας ως ρυθμιστική παράμετρο τη λ και για k αριθμό ανεξάρτητων μεταβλητών η συνάρτηση κόστους του αλγορίθμου Lasso δίνεται παρακάτω:

$$\sum_{i=1}^{n}(y_i-\hat{y}_i)^2=\sum_{i=1}^{n}(y_i-\beta_0-\sum_{j=1}^{k}\beta_j x_{ij})^2+\lambda\sum_{j=1}^{k}\left|\beta_j\right|$$

Όσον αφορά τη ρυθμιστική παράμετρο λ, αυτή μπορεί να πάρει τιμές από το 0 (δεν εφαρμόζεται ποινή) έως το άπειρο και θα πρέπει να γίνει επαναληπτική επικύρωση 10 τμημάτων ώστε να επιλεχθεί η τιμή με το χαμηλότερο σφάλμα. Εξετάζοντας 100 τιμές γεωμετρικής κατανομής της ρυθμιστικής παραμέτρου επιλέγεται αυτή με το χαμηλότερο RMSE όπως και φαίνεται στο Σχήμα 1 για την πρόβλεψη της διάρκειας σπουδών με χρήση δεδομένων του 1ου έτους φοίτησης.

Σχήμα 1: Επιλογή ρυθμιστικής παραμέτρου λ με το χαμηλότερο μέσο RMSE (μπλε τρίγωνο) μετά από επαναληπτική επικύρωση 10 τμημάτων, αξιοποιώντας τις εξεταζόμενες μεταβλητές εισόδου του 1ου ακαδημαϊκού έτους για την πρόβλεψη της διάρκειας σπουδών.



Από την επαναληπτική επικύρωση 10 τμημάτων επιλέχθηκε η τιμή της ρυθμιστικής παραμέτρου λ=0.0526. Στον πίνακα 1 γίνεται ενδεικτικά σύγκριση των συντελεστών παλινδρόμησης με ή χωρίς την εφαρμογή του αλγορίθμου Lasso για την επιλογή υποσυνόλου ανεξάρτητων μεταβλητών και την πρόβλεψη της διάρκειας σπουδών με χρήση δεδομένων του 1ου έτους φοίτησης. Από τη σύγκριση των συντελεστών παρατηρούμε συρρίκνωση του συντελεστή της μεταβλητής Π.1.1 και μηδενισμό του συντελεστή της Μ.1.1.

Πίνακας 1: Σύγκριση συντελεστών παλινδρόμησης με ή χωρίς την εφαρμογή αλγορίθμου Lasso για την πρόβλεψη της διάρκειας σπουδών με χρήση δεδομένων του 1ου έτους φοίτησης

	Πολλαπλή γραμμική παλινδρόμηση	Πολλαπλή γραμμική παλινδρόμηση με εφαρμογή αλγορίθμου Lasso (λ=0.0526)
	Συντελεστές παλινδρόμησης	
Σταθερά παλινδρόμησης	8.180	8.180
Π.1.1	-3.161	-2.856
Μ.1.1	0.054	0.000

Εκτός της απλής και πολλαπλής γραμμικής παλινδρόμησης χρήσιμο θα ήταν να εξεταστούν εναλλακτικές μέθοδοι πρόβλεψης του βαθμού αποφοίτησης και της διάρκειας σπουδών. Στον πίνακα 2 δίνονται ενδεικτικά τα αποτελέσματα επαναλαμβανόμενων επικυρώσεων 10 τμημάτων εναλλακτικών μοντέλων για την πρόβλεψη του βαθμού αποφοίτησης και για τα πρώτα 2 έτη φοίτησης, έπειτα από επιλογή υποσυνόλου μεταβλητών εισόδου με τη μέθοδο της προς τα πίσω απαλοιφής. Να σημειωθεί πως στην περίπτωση των μεταβλητών Μ.χ.ν δεν επιλέχθηκαν οι αθροιστικοί μέσοι όροι μαθημάτων αλλά οι μέσοι όροι μεμονωμένα για κάθε έτος. Στο μοντέλο του 1^{ου} έτους δίνεται το μέσο RMSE με χρήση αλγορίθμου των Support Vector Machines. Από τη σύγκριση με το μέσο RMSE του αντίστοιχου μοντέλου του 5^{ου} κεφαλαίου παρατηρούμε μείωση από 0.4106 σε 0.375. Τέλος, με τη χρήση αλγορίθμου της Gaussian Process Regression στο 2^{ο} έτος παρατηρείται μείωση του μέσου RMSE σε σχέση με το 5^{ο} κεφάλαιο από 0.2972 σε 0.272.

Πίνακας 2: Επαναλαμβανόμενη επικύρωση 10 τμημάτων σφαλμάτων εναλλακτικών μεθόδων πρόβλεψης βαθμού αποφοίτησης για τα πρώτα 2 έτη φοίτησης

	RMSE 1ης επικύρωσης 10 τμημάτων	RMSE 2ης επικύρωσης 10 τμημάτων	RMSE 3ης επικύρωσης 10 τμημάτων	RMSE 4ης επικύρωσης 10 τμημάτων	RMSE 5ης επικύρωσης 10 τμημάτων	RMSE 6ης επικύρωσης 10 τμημάτων	RMSE 7ης επικύρωσης 10 τμημάτων	RMSE 8ης επικύρωσης 10 τμημάτων	RMSE 9ης επικύρωσης 10 τμημάτων	RMSE 10ης επικύρωσης 10 τμημάτων	Μέσος όρος RMSE
1ο Ακαδημαϊκό έτος											
Quadratic SVM	0.37287	0.37237	0.37637	0.36656	0.37402	0.37927	0.38485	0.37825	0.36956	0.37173	0.375
2ο Ακαδημαϊκό έτος											
Rational quadratic GPR	0.2749	0.26797	0.27326	0.27256	0.2748	0.2719	0.27264	0.27004	0.27293	0.27235	0.272

ΔΗΛΩΣΗ ΠΕΡΙ ΜΗ ΠΡΟΣΒΟΛΗΣ ΔΙΚΑΙΩΜΑΤΩΝ

ΠΝΕΥΜΑΤΙΚΗΣ ΙΔΙΟΚΤΗΣΙΑΣ

Δηλώνω ρητά ότι η παρούσα Διπλωματική Εργασία με τίτλο:

" Μελέτη της Ακαδημαϊκής Επίδοσης των Φοιτητών του Τμήματος Μηχανολόγων Μηχανικών *"*

καθώς και τα ηλεκτρονικά αρχεία και πηγαίοι κώδικες που αναπτύχθηκαν ή τροποποιήθηκαν στο πλαίσιο αυτής της εργασίας και αναφέρονται ρητώς μέσα στο κείμενο που συνοδεύουν και η οποία έχει εκπονηθεί στο Τμήμα Μηχανολόγων Μηχανικών του Πανεπιστημίου Δυτικής Μακεδονίας, υπό την επίβλεψη του/της κ./κα. Παναγιωτίδου Σοφίας ***, αποτελεί αποκλειστικά προϊόν προσωπικής εργασίας και δεν προσβάλλει κάθε μορφής πνευματικά δικαιώματα τρίτων και δεν είναι προϊόν μερικής ή ολικής αντιγραφής, οι πηγές δε που χρησιμοποιήθηκαν περιορίζονται στις βιβλιογραφικές αναφορές και μόνον. Τα σημεία όπου έχω χρησιμοποιήσει ιδέες, κείμενο, αρχεία ή / και πηγές άλλων συγγραφέων, αναφέρονται ευδιάκριτα στο κείμενο με την κατάλληλη παραπομπή και η σχετική αναφορά περιλαμβάνεται στο τμήμα των βιβλιογραφικών αναφορών με πλήρη περιγραφή.***

Απαγορεύεται η αντιγραφή, αποθήκευση και διανομή της παρούσας εργασίας, εξ ολοκλήρου ή τμήματος αυτής, για εμπορικό σκοπό. Επιτρέπεται η ανατύπωση, αποθήκευση και διανομή για σκοπό μη κερδοσκοπικό, εκπαιδευτικής ή ερευνητικής φύσης, υπό την προϋπόθεση να αναφέρεται η πηγή προέλευσης. Ερωτήματα που αφορούν τη χρήση της εργασίας για κερδοσκοπικό σκοπό πρέπει να απευθύνονται προς τον συγγραφέα. Οι απόψεις και τα συμπεράσματα που περιέχονται σε αυτό το έγγραφο εκφράζουν τον συγγραφέα και μόνο.

Copyright (C) Ονοματεπώνυμο Φοιτητή & Επιβλέποντα, Ημερομηνία, Πόλη

Υπογραφή Φοιτητή