Ανάλυση Δεδομένων

Εισαγωγή σε Κολέγια

Γεώργιος Παυλής

Τμήμα Στατιστικής
Οικονομικό Πανεπιστήμιο Αθηνών
Μάιος 2023

Εισαγωγή

Η τριτοβάθμια εκπαίδευση αποτελεί μία από τις κυριότερες περιόδους της ζωής ενός ανθρώπου η οποία καθορίζει για πολλούς την μετέπειτα καριέρα τους και κατ'επέκταση την καθημερινότητα του. Συνεπώς θα ήταν ενδιαφέρων να αναλύσουμε δεδομένα εισαγωγής φοιτητών σε κολέγια και να εντοπίσουμε διαφορές με βάση το φύλο, την δευτεροβάθμια εκπαίδευση, την φυλή αν παρουσιάζουν ευχέρεια σε συγκεκριμένα μαθήματα είτε την επιλογή τους στο πρόγραμμα σπουδών.

Διακρίνοντας τέτοιου είδους μοτίβα, προτιμήσεις δίνει την ευκαιρία στο εν λόγω κολέγιο να ανανεώσει το πρόγραμμα σπουδών του σύμφωνα με αυτά τα κριτήρια.

Όνομα	Περιγραφή	Τιμές
id	Κωδικός Μαθητή	1:200
genre	Το φύλο του μαθητή	male/female
race	Η φυλή στην οποία ανήκει	hispanic/asian/african-amer/white
schtyp	Ο τύπος σχολείου που έχει αποφοιτήσει	public/private
prog	Το πρόγραμμα σπουδών που ακολούθησε στον προηγ. κύκλο σπουδ. του	general/academic/vocation
write	Βαθμός που πήρε στο τεστ μαθηματικών	0:100
math	Βαθμός που πήρε στο τεστ των κοινωνικών επιστημών	0:100
socst	Βαθμός που πήρε στην έκθεση	0:100

Περιγραφή Μεταβλητών

Όπως είδαμε θα εργαστούμε με 9 μεταβλητές τις οποίες θα χωρίσουμε σε ονομαστικές και διατάξιμες. Ονομαστικές ονομάζονται εκείνες που υπάγονται σε μια κατηγορία και αντιπροσωπεύονται από ένα αριθμό π.χ. για την genre μπορούμε να αντιστοιχίσουμε το 1 για τους άνδρες και το 2 για τις γυναίκες, έτσι μπορούμε να τις χρησιμοποιήσουμε στην ανάλυση. Όμοια για την φυλή έχουμε 1:'hispanic', 2:'asian', 3:'african-americ', 4:'white'. Για το schtyp 1:'public', 2:'private'. Τέλος για το prog 1:'general', 2:'academic', 3:'vocation'.

Οι διατάξιμες είναι κατηγορίες που μπορούν να διατεταχθούν σε μία λογική σειρά όπως εδώ οι βαθμοί διαγωνισμάτων. Γενικότερα οι βαθμοί δεν θεωρούνται διακριτοί/ακέραιοι αριθμοί αλλά σαν κατηγορικές, διότι δεν μοιράζονται τις ίδιες αναλογίες, δηλαδή θα έπρεπε να ισχύει ότι οι διαφορά των μαθητών που έγραψαν 40 και 20 είναι ίδια με αυτούς που έγραψαν 100 και 80 κάτι που δεν ισχύει.

Ξεκινώντας με την περιγραφή των ονομαστικών έχουμε τις συχνότητες κάθε μεταβλητές, πώς οι μαθητές διαμοιράζονται στις κατηγορίες.

Φύλο	Φυλή	Τύπος σχολείου	Πρόγραμμα
Άνδρες: 91	Ισπανοί: 24	Δημόσιο: 168	Γενικό: 45
Γυναίκες: 109	Ασιάτες: 11	Ιδιωτικά: 32	Ακαδημαϊκό: 105
	Αφροαμερικανοί: 20		Επαγγελματικό: 50
	Λευκοί: 145		

Παρατηρούμε οι άνδρες με τις γυναίκες είναι ισορροπημένοι, οι καυκάσιοι αποτελούν το 72% των μαθητών, οι απόφοιτοι των δημόσιων το 84% και αυτοί που ακολούθησαν προγενέστερα των σπουδών τους ακαδημαϊκό πρόγραμμα, μετράνε σ το 52%.

Στη συνέχεια θα κοιτάξουμε τις διατάξιμες μεταβλητές.

Μεταβλητή	Μέσος Όρος	Επικρατούσα Τιμή	Διάμεσος	Τυπική Απόκλιση	Εύρος	Ασυμμετρία	Κύρτωση
write	52.77	59	54	9.48	31-67	-0.48	-0.78
math	52.65	57	52	9.37	33-75	0.29	-0.69
socst	52.41	61	52	10.74	26-71	-0.38	-0.57

Τα περιγραφικά μέτρα αποτελούνται από: Τον μέσο όρο. Την επικρατούσα τιμή, η τιμή με την μεγαλύτερη συχνότητα εμφάνισης. Την τυπική απόκλιση, που μετράει το άπλωμα των τιμών της μεταβλητής γύρω από τον μέσο όρο. Την διάμεσο ως την κεντρική τιμή των διατεταγμένων βαθμολογιών. Το εύρος εκφράζει την ελάχιστη και την μικρότερη τιμή, η ασυμμετρία και η κύρτωση συγκρίνουν την κατανομή των βαθμολογιών, δηλαδή την πιθανότητα εμφάνισης κάθε βαθμολογίας, με την κανονική κατανομή. Η πρώτη μετράει πόσο αποκλίνει από την συμμετρία και η δεύτερη την απόκλιση της κατανομής των πιθανοτήτων ως προς το σχήμα της κανονικής.

Μερικά μέτρα όπως ο μέσος, η κύρτωση, ασυμμετρία χρησιμοποιούνται περισσότερο σε συνεχή δεδομένα (π.χ ο χρόνος) όχι τόσο σε κατηγορικά αλλά επειδή έχουμε μεγάλο εύρος κατηγοριών στις βαθμολογίες (1:100) μπορούμε να τα χρησιμοποιήσουμε.

Παρατηρούμε ότι ο μέσος όρος με την διάμεσο δεν διαφέρουν πολύ σε αντίθεση με την επικρατούσα τιμή που είναι πάντα μεγαλύτερη υποδεικνύοντας μία αριστερή ασυμμετρία. Η τυπική απόκλιση σε όλα τα μαθήματα είναι περίπου ίσες, συνεπώς έχουμε όμοια διασπορά από τον μέσο. Οι ασυμμετρία και κύρτωση ταυτίζονται με την κανονική στην τιμή 0, άρα για την έκθεση και την κοινωνική επιστήμη έχουμε αριστερή ασυμμετρία, η κατανομή γέρνει προς τα δεξιά με τις περισσότερες τιμές. Από την κύρτωση φαίνεται να έχουμε πλατύκυρτες κατανομές με πολλές τιμές με μεγάλη συχνότητα. Όπως φαίνονται από τα διαγράμματα.

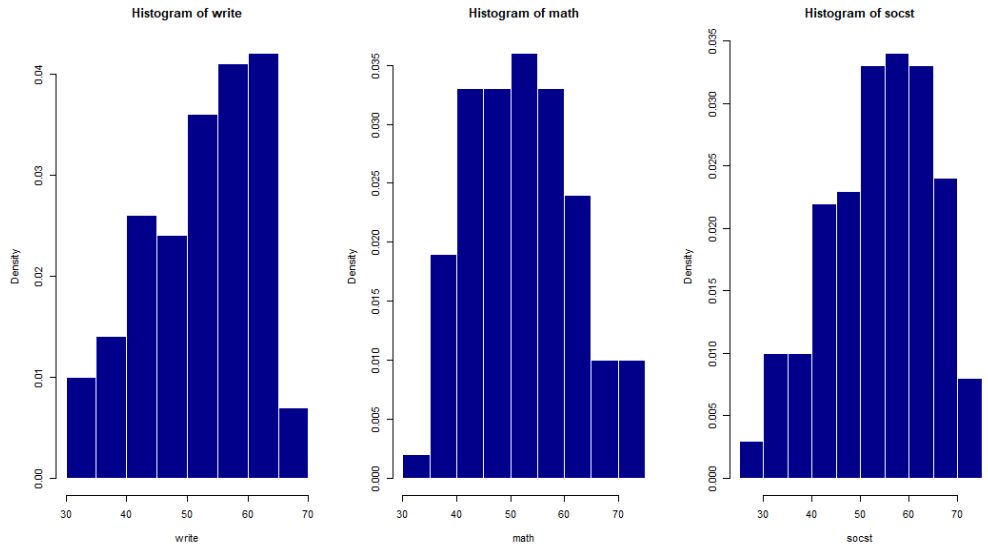

Figure 1: Ιστόγραμμα βαθμολογιών

Σχέσεις ανά δύο

Μερικές χρήσιμες σχέσεις που μπορούμε να εξετάσουμε είναι: Schtyp $\sim$ Math,Socst,Write αν ο τύπος σχολείου παίζει ρόλο στις επιδόσεις των μαθημάτων. Race $\sim$ Math,Socst,Write αν κάποιο φυλή έχει κάποια προτίμηση σε συγκεκριμένο μάθημα επιτυγχάνοντας καλύτερες βαθμολογίες. Schtyp $\sim$ Prog αναλόγως τον τύπο σχολείου που φοίτησαν τι πρόγραμμα ακολούθησαν. Race $\sim$ Prog όμοια με το προηγούμενο σε σχέση τώρα με την φυλή.

Ονομαστικές Κατηγορικές

- Schtyp $\sim$ Prog. Μέσω ελέγχων X^2 και Fisher exact test σε πίνακες συνάφειας των κατηγορικών μεταβλητών βρέθηκε ότι οι προτιμήσεις των παιδιών για το πρόγραμμα σπουδών έχει στατιστικά σημαντική διαφορά αναλόγως αν φοίτησαν σε ιδιωτικό ή δημόσιο σχολείο

	Γενικό	Ακαδημαϊκό	Επαγγελματικό
Δημόσια	23.2%	48.2 %	28.6%
Ιδιωτικά	18.8%	75%	6.2%

  Όπως φαίνεται και από τον πίνακα συνάφειας με τα ποσοστά κάθε κατηγορία ανά γραμμή υπάρχουν μεγάλες διαφορές. Η μεγαλύτερη διαφορά εμφανίζεται ανάμεσα στο ποσοστό των παιδιών απο ιδιωτικά που ακολούθησαν επαγγελματική καριέρα είναι της τάξης του 6.2% ,ενώ το ποσοστό των παιδιών από δημόσια είναι στο 28.6%.

Βαθμολογίες

Στις βαθμολογίες μέσω ελέγχων για 2 εξαρτημένα δείγματα (Paired-t.test & Wilcoxon test), δηλαδή όπως στο παρόν ανά δύο βαθμολογίες σε διαφορετικά μαθήματα για τον ίδιο φοιτητή, δεν βρέθηκε καμία στατιστικά σημαντική διαφορά στα μαθήματα. Συνεπώς ο βαθμός ανάμεσα και στα τρία μαθήματα δεν διαφοροποιείται πολύ για τον εκάστοτε φοιτητή.

Βαθμολογίες και κατηγορικές

Τώρα θα εξετάσουμε αν για τις ομάδες των κατηγορικών μεταβλητών μας υπάρχουν διαφορές σε κάθε μάθημα ή όπως ονομάζεται ανάλυση διακύμανσης κατά ένα παράγοντα. Με την βοήθεια των ελέγχων ισότητας μέσων και διαμέσων ANOVA & Kruskal Wallis αντίστοιχα, με την πρώτη να υποθέτουμε κανονικότητα και ομοσκεδαστικότητα.

- Math $\sim$ Race. Υπάρχει αντίθεση στις βαθμολογίες των μαθηματικών ανάμεσα σε διαφορετικές φυλές. Τις χειρότερες βαθμολογίες παρουσιάζουν οι Αφροαμερικανοί και Ισπανοί χωρίς να ξεχωρίζουν κατά μέσο όρο 45-50, επόμενοι είναι οι Λευκοί με μέσο όρο 54 και πρώτοι οι Ασιάτες με μεγάλοι διακύμανση στον μέσο βαθμό να βρίσκεται στο 50-65.

- Socst $\sim$ Race. Επίσης παρατηρείται στο μάθημα των Κοινωνικών Επιστημών να προηγούνται οι λευκοί από τους Ισπανούς .

- Write $\sim$ Race. Ακόμα στην Έκθεση έχουμε κατά μέσο όρο στο 45-50 τους Ισπανούς και Αφροαμερικάνους, τους Λευκούς στο 54 και χώρις να είμαστε βέβαιοι ότι υπάρχει διαφορά οι Ασιάτες στο 50-65 με μεγαλύτερη διακύμανση. Παρόμοια με τα μαθηματικά

- Math $\sim$ Prog. Στα μαθηματικά αλλά τώρα για ομάδες με διαφορετικό πρόγραμμα σπουδών όλες οι ομάδες ξεχωρίζουν με πρώτη να έρχεται η ακαδημαϊκή, ακολουθεί η γενική και τέλος η εργασιακή.

- Socst $\sim$ Prog. Ομοίως και στις Κοινωνικές Επιστήμες με την ίδια σειρά.

- Write $\sim$ Prog. Ομοίως το ίδιο και στην Έκθεση.

- Write $\sim$ Genre + Prog. Υπάρχουν διαφορές στην έκθεση ανάλογα τώρα με το φύλο και το πρόγραμμα σπουδών που έχουν ακολουθήσει;

  Ναι υπάρχουν. Μάλιστα οι άνδρες με επαγγελματικό πρόγραμμα γράφουν λιγότερο από κάθε κατηγορία, προηγούνται επόμενοι οι άνδρες με γενικό πρόγραμμα και τέλος στην κορυφή βρίσκονται οι γυναίκες με ακαδημαϊκή εκπαίδευση.

Γραμμικά Μοντέλα

Δεδομένου των συσχετίσεων που είδαμε στο προηγούμενο ερώτημα έχουμε την δυνατότητα να μοντελοποίησουμε την βαθμολογία ενός μαθήματος γνωρίζοντας τις υπόλοιπες μεταβλητές και να διαπιστώσουμε ποιες πραγματικά συντελούν σε αυτό, κάνοντας ακόμα και πρόβλεψη για τον αναμενόμενο βαθμό.

Αρχικά για να έχουμε ξεκάθαρη εικόνα πρέπει να δούμε τις συσχετίσεις μεταξύ των βαθμολογιών ώστε δεδομένου κάποιων να αποκτήσω πληροφορία για αυτή που θέλω να προβλέψω. Στις σχέσεις ανά δύο δεν καταφέραμε να διακρίνουμε σημαντικές διαφορές, ευνοώντας τον στόχο μας. Μοιράζονται δηλαδή πληροφορία και όπως ειπώθηκε έχουν την ίδιο βαθμό περίπου και στα τρία μαθήματα το οποίο μπορώ να εκμεταλλευτώ. Όπως κάθε μοντέλο πρέπει να θέσουμε κάποιες παραδοχές για να λειτουργήσει στην προκειμένη περίπτωση πρέπει τα κατάλοιπα, οι αποκλίσεις των προβλεπόμενων από τις πραγματικές τιμές, να ακολουθούν κανονική κατανομή με μέση τιμή 0 και κοινή διακύμανση. Επίσης ανεξαρτησία των καταλοίπων δηλαδή οι προηγούμενες να μην επηρεάζουν τις επόμενες να είναι δηλαδή τυχαία. Τέλος η σχέση της μεταβλητής που θέλω να προβλέψω να έχει γραμμική σχέση με το μοντέλο μου, καθώς μιλάμε πάντα για γραμμικό μοντέλο.

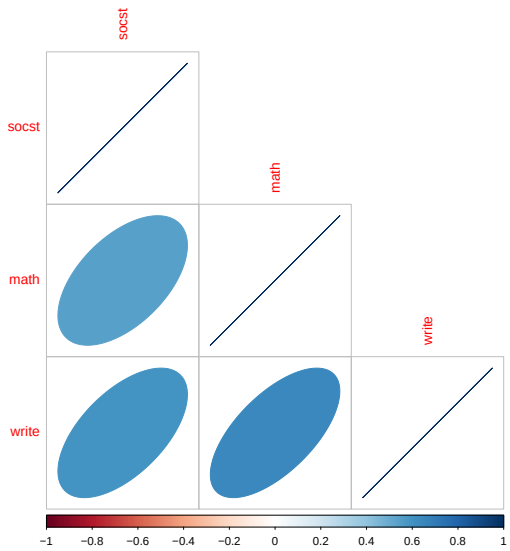

Figure 2: Διάγραμμα Συσχετίσεων

Socst

Για να προβλέψουμε και να ερμηνεύσουμε τον βαθμό στις κοινωνικές επιστήμες καταλήξαμε στο εξής μοντέλο :

socst(Κοινωνικές Επιστήμες)=β_0+β_1*write(Έκθεση)+β_2*Math(Μαθηματικά)+β_3*prog(Πρόγραμμα Σπουδών academic) + $\beta_4$prog(Πρόγραμμα Σπουδών vocation)

Είναι προφανές ότι δεν αξιοποίησα όλες τις υπόλοιπες μεταβλητές, αυτό έγινε διότι δεν προσδίδουν όλες οι μεταβλητές πληροφορία για τον βαθμό στις Κοινωνικές Επιστήμες ή αυτή που παρέχουν να επικαλύπτεται και να ταυτίζεται με μίας άλλης οπότε αυτό την καθιστά πλεονεκτική. Για να αποφασίσω ποιους παράγοντες θα κρατήσω και ποιους θα βγάλω τρέχω μία διαδικασία stepwise, όπου με βάση κάποιο κριτήριο αξιολόγησης μοντέλου (AIC ή BIC) συγκρίνω το πλήρες με τα μοντέλα με μία λιγότερη και επαναλαμβάνω μέχρι να καταλήξω στο φαινομενικά καλύτερο με τις πιο ποιοτικές μεταβλητές.

Έχοντας βρει το μοντέλο για να λειτουργεί πρέπει να ελέγξω τις υποθέσεις που έθεσα. Για την κανονικότητα επέλεξα τον έλεγχο (Lilliefor's p=0.09 $>$ 0.05) λόγω του μεγάλου δείγματος που έχω και δεν απορρίπτω την κανονικότητα σε επίπεδο 5% .Για την κοινή διακύμανση ο έλεγχος δεν την απορρίπτει (Levene's p= 0.44 $>$ 0.05).Τέλος από τα διαγράμματα δεν διακρίνεται κάποια εξάρτηση των καταλοίπων όλα είναι τυχαία κατανεμημένα γύρω από το 0, συνεπώς δεν απορρίπτουμε την ανεξαρτησία των σφαλμάτων.

Εκτιμώντας τις παραμέτρους β με την μέθοδο ελαχίστων τετραγώνων προκύπτουν :

Το intercept αντιστοιχεί στο $\beta_0 = 15.61$ και συμβολίζει την αναμενόμενη βαθμολογία όταν ο φοιτητής έχει ακολουθήσει Γενικό πρόγραμμα σπουδών και έχει γράψει μηδέν στην Έκθεση και στα Μαθηματικά. Η ερμηνεία του β_0 δεν έχει πάντα νόημα ιδιαίτερα όταν δεν έχω δεδομένα για βαθμολογίες στο 0 δεν μπορώ να προβλέψω αποτελεσματικά εκεί που δεν έχω στοιχεία, όμως είναι απαραίτητη στο μοντέλο.

Το $\beta_1 = 0.45$ είναι η αναμενόμενη αύξηση στην βαθμολογία των κοινωνικών Επιστημών αν στην Έκθεση είχα γράψει ένα βαθμό παραπάνω δεδομένου ότι στα μαθηματικά έχω το ίδιο βαθμό και ανήκω στο ίδιο πρόγραμμα σπουδών.

Το $\beta_2 = 0.24$ αντίστοιχα είναι η αναμενόμενη αύξηση στις κοινωνικές επιστήμες αν είχα γράψει στα μαθηματικά ένα βαθμό παραπάνω δεδομένου ότι τα άλλα είναι σταθερά.

Το $\beta_3 = 2.28$ συμβολίζει την αναμενόμενη αύξηση στο βαθμό αν συγκρίνουμε δύο φοιτητές με τους ίδιους βαθμούς σε

	socst		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	15.61	7.43 – 23.79	**<0.001**
write	0.45	0.29 – 0.60	**<0.001**
math	0.24	0.08 – 0.40	**0.004**
prog [2]	2.28	-0.73 – 5.28	0.136
prog [3]	-2.67	-6.02 – 0.68	0.118
Observations	200		
R^2 / R^2 adjusted	0.441 / 0.429		

Figure 3: Μοντέλο για το Socst

μαθηματικά, έκθεση και έχουν ακολουθήσει ακαδημαϊκό πρόγραμμα σπουδών και γενικό, δηλαδή για του ίδιους βαθμούς αυτός με ακαδημαϊκό πρόγραμμα αναμένεται να γράψει 2.28 παραπάνω από αυτόν στο γενικό .

Το $\beta_4 = -2.64$ είναι αντίστοιχα η μείωση στον βαθμό αν συγκρίνουμε φοιτητές από επαγγελματικό και γενικό πρόγραμμα με ίδιους βαθμούς.

Παρατηρώ ότι οι τρεις πρώτοι παράμετροι είναι στατιστικά σημαντικά διάφοροι του μηδενός, έτσι είναι σημαντικοί στο μοντέλο.

Επίσης το $R^2 = 0.441$ ο δείκτης καλής προσαρμογής του μοντέλου απεικονίζει την μεταβλητότητα των βαθμολογιών των Κοινωνικών Επιστημών που εξηγείται από το μοντέλο μου, το οποίο είναι αρκετά καλό.

math

Προχωράμε στο μοντέλο για τον βαθμό στα Μαθηματικά :

Math(Μαθηματικά)=β_0+β_1*genre(Φύλο)+β_2*prog(Πρόγραμμα Σπουδών academic)+β_3*prog(Πρόγραμμα Σπουδών vocation)+β_4*Write(Έκθεση)+β_5*Socst(Κοινωνικές Επιστήμες)

Από τις υποθέσεις του μοντέλου δεν απορρίπτω καμία. Κανονικότητα δεν απορρίπτεται (Lilliefors p=0.061 > 0.05). Ομοσκεδαστικότητα δεν απορρίπτεται (Levene's p=0.13¿0.05). Δεν απορρίπτεται και η ανεξαρτησία των σφαλμάτων.

Εκτιμώντας τις παραμέτρους β με την μέθοδο ελαχίστων τετραγώνων προκύπτουν :

Σε σχέση με το προηγούμενο έχουμε προσθέσει την κατηγορική genre που διαχωρίζει τα φύλα λογικό εφόσον στις ανά δύο στατιστικά σημαντική την σχέση math∼genre.

Το intercept αντιστοιχεί στο $\beta_0 = 20.02$ και συμβολίζει την αναμενόμενη βαθμολογία για έναν άνδρα φοιτητή που έχει ακολουθήσει Γενικό πρόγραμμα σπουδών και έχει γράψει μηδέν στην Έκθεση και στις Κοινωνικές Επιστήμες. Ξανά Η ερμηνεία του β_0 δεν έχει πάντα νόημα, όμως είναι απαραίτητη στο μοντέλο.

Το $\beta_1 = -3.03$ συμβολίζει την αναμενόμενη μείωση στον βαθμό όταν συγκρίνουμε γυναίκες και άνδρες με ίδιο πρόγραμμα σπουδών και ίδιους βαθμούς στα υπόλοιπα.

Το $\beta_1 = 3.54$ είναι η αναμενόμενη αύξηση στον βαθμό όταν συγκρίνουμε δύο φοιτητές ίδιου φύλου με ίδιους βαθμούς στην Έκθεση και Κοινωνικές Επιστήμες αλλά ο ένας είναι από ακαδημαϊκό πρόγραμμα και ο άλλος από γενικό.

	math		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	20.02	13.51 – 26.52	**<0.001**
genre [female]	-3.03	-5.02 – -1.04	**0.003**
prog [academic]	3.54	1.06 – 6.02	**0.005**
prog [vocation]	-0.60	-3.42 – 2.22	0.677
write	0.46	0.33 – 0.60	**<0.001**
socst	0.16	0.04 – 0.27	**0.008**
Observations	200		
R^2 / R^2 adjusted	0.489 / 0.475		

Figure 4: Μοντέλο για το Math

Το $\beta_2 = -0.60$ ομοίως με το β_1 αλλά αντί για ακαδημαϊκό έχουμε επαγγελματικό πρόγραμμα σπουδών.

Το $\beta_3 = 0.46$ είναι η αναμενόμενη αύξηση στον βαθμό των μαθηματικών για δύο φοιτητές με ένα βαθμό διαφορά στην έκθεση και όλα τα υπόλοιπα ταυτόσημα.

Το $\beta_4 = 0.46$ ομοίως είναι η αναμενόμενη αύξηση στον βαθμό των μαθηματικών για δύο φοιτητές με ένα βαθμό διαφορά στις Κοινωνικές Επιστήμες και όλα τα υπόλοιπα ταυτόσημα.

Οι πέντε από τους έξι παραμέτρους είναι στατιστικά σημαντικά διάφοροι του μηδενός, έτσι είναι σημαντικοί στο μοντέλο.

Το $R^2 = 0.489$ ο δείκτης καλής προσαρμογής του μοντέλου είναι αρκετά καλό και απεικονίζει ότι το μοντέλο μου περνάει καλά γύρω από τα δεδομένα.

Στη συνέχεια πρέπει να δημιουργήσουμε πάλι μοντέλα για τους βαθμούς στις Κοινωνικές Επιστήμες, Μαθηματικά αλλά τώρα χωρίς την επίδραση της έκθεσης.

Socst

Το μοντέλο που κατέληξα είναι το ακόλουθο :

socst(Κοινωνικές Επιστήμες)=β_0+β_1*Math(Μαθηματικά)+β_2*prog(Πρόγραμμα Σπουδών academic) + β_3*prog(Πρόγραμμα Σπουδών vocation)

Αυτό που είχα και προηγουμένως χωρίς την επίδραση της Έκθεσης.

Από τις υποθέσεις παραβιάζεται μονάχα αυτή της κανονικότητας η κατανομή των καταλοίπων έχει πιο παχιές ουρές από'τι θα έπρεπε δηλαδή έχω μεγαλύτερη πιθανότητα να παρατηρήσω ακραίες τιμές οπότε απορρίπτεται (Lilliefors p=0.02<0.05). Ομοσκεδαστικότητα δεν απορρίπτεται (Levene's p=0.14>0.05).Τέλος τα κατάλοιπα είναι ασυσχέτιστα.

Το intercept αντιστοιχεί στο $\beta_0 = 26.29$ και συμβολίζει την αναμενόμενη βαθμολογία όταν ο φοιτητής έχει ακολουθήσει Γενικό πρόγραμμα σπουδών και έχει γράψει μηδέν στα Μαθηματικά.

	socst		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	26.29	18.43 – 34.14	**<0.001**
math	0.49	0.34 – 0.63	**<0.001**
prog [academic]	2.83	-0.39 – 6.06	0.085
prog [vocation]	-3.83	-7.41 – -0.25	**0.036**
Observations	200		
R^2 / R^2 adjusted	0.348 / 0.338		

Figure 5: Δεύτερο μοντέλο για το Socst

Το $\beta_1 = 0.49$ είναι η αναμενόμενη αύξηση στον βαθμό των Κοινωνικών Επιστημών για δύο φοιτητές με ένα βαθμό διαφορά στα μαθηματικά και όλα τα υπόλοιπα ταυτόσημα.

Το $\beta_2 = 2.83$ συμβολίζει την αναμενόμενη αύξηση στο βαθμό αν συγκρίνουμε δύο φοιτητές με τους ίδιους βαθμούς σε μαθηματικά και έχουν ακολουθήσει ακαδημαϊκό πρόγραμμα σπουδών και γενικό, δηλαδή για του ίδιους βαθμούς αυτός με ακαδημαϊκό πρόγραμμα αναμένεται να γράψει 2.83 παραπάνω από αυτόν στο γενικό .

Το $\beta_3 = -2.83$ είναι αντίστοιχα η μείωση στον βαθμό αν συγκρίνουμε φοιτητές από επαγγελματικό και γενικό πρόγραμμα με ίδιους βαθμούς.

Επίσης το $R^2 = 0.338$ ο δείκτης καλής προσαρμογής του μοντέλου είναι χειρότερος από αυτό του προηγούμενου μοντέλου καθώς αφαιρώντας την επίδραση του write χάσαμε σημαντική πληροφορία.

Απορρίπτοντας την υπόθεση της κανονικότητας των σφαλμάτων παραβιάζουμε μία βασική παραδοχή του μοντέλου οπότε δεν ξέρουμε κατά πόσο μπορούμε να το εμπιστευτούμε.

Math

Το μοντέλο που κατέληξα είναι το ακόλουθο :

Math(Μαθηματικά)=β_0+β_1*race(Φυλή Asian)+β_2*race(Φυλή african-amer)+β_3*race(Φυλή white)+β_4*prog(Πρόγραμμα Σπουδών academic)+β_5*prog(Πρόγραμμα Σπουδών vocation)+β_6*Socst(Κοινωνικές Επιστήμες)

Δεν απορρίπτει καμία υπόθεση.Κανονικότητα (Lilliefor's p = 0.8848>0.05).Ομοσκεδαστικότητα (Levene's p=0.4313>0.05).Δεν απορρίπτω την ανεξαρτησία των καταλοίπων.

Το intercept αντιστοιχεί στο $\beta_0 = 29.62$ και συμβολίζει την αναμενόμενη βαθμολογία όταν ένας Ισπανός φοιτητής έχει ακολουθήσει Γενικό πρόγραμμα σπουδών και έχει γράψει μηδέν στις Κοινωνικές Επιστήμες.

	math		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	29.62	23.41 – 35.82	**<0.001**
race [asian]	8.07	2.82 – 13.33	**0.003**
race [african-amer]	-1.26	-5.60 – 3.07	0.567
race [white]	4.03	0.82 – 7.23	**0.014**
prog [academic]	4.73	2.09 – 7.36	**<0.001**
prog [vocation]	-1.04	-4.06 – 1.98	0.498
socst	0.34	0.23 – 0.44	**<0.001**
Observations	200		
R^2 / R^2 adjusted	0.420 / 0.402		

Figure 6: Δεύτερο μοντέλο για το Math

Το $\beta_1 = 8.07$ είναι η αναμενόμενη αύξηση στον βαθμό των Μαθηματικών για δύο φοιτητές ο ένας Ασιάτης και ο άλλος Ισπανός με ίδιο βαθμό στις Κοινωνικές Επιστήμες.

Το $\beta_2 = -1.26$ είναι η αναμενόμενη μείωση στον βαθμό των Μαθηματικών για δύο φοιτητές ο ένας Αφροαμερικανός και ο άλλος Ισπανός με ίδιο βαθμό στις Κοινωνικές Επιστήμες.

Το $\beta_3 = 4.03$ είναι η αναμενόμενη μείωση στον βαθμό των Μαθηματικών για δύο φοιτητές ο ένας Λευκός και ο άλλος Ισπανός με ίδιο βαθμό στις Κοινωνικές Επιστήμες.

Το $\beta_4 = 4.73$ είναι η αναμενόμενη αύξηση στον βαθμό των Μαθηματικών για δύο φοιτητές ο ένας έχει ακολουθήσει ακαδημαϊκό πρόγραμμα σπουδών και ο άλλος γενικό με ίδιο βαθμό στις Κοινωνικές Επιστήμες.

Το $\beta_5 = -1.04$ είναι η αναμενόμενη μείωση στον βαθμό των Μαθηματικών για δύο φοιτητές ο ένας έχει ακολουθήσει ακαδημαϊκό πρόγραμμα σπουδών και ο άλλος γενικό με ίδιο βαθμό στις Κοινωνικές Επιστήμες.

Το $\beta_6 = 0.34$ είναι η αναμενόμενη αύξηση στον βαθμό για δύο φοιτητές με ένα βαθμό διαφορά στις Κοινωνικές Επιστήμες και όλα τα υπόλοιπα ταυτόσημα.

Αλλαγή Παραμέτρων

Όλα τα παραπάνω μοντέλα χρησιμοποιούσαν την παραμετροποίηση corner constraints για την εισαγωγή των κατηγορικών μεταβλητών δηλαδή είχαν ένα επίπεδο αναφοράς και σύγκριναν οι υπόλοιπες με αυτό.Αν χρησιμοποιήσουμε μία διαφορετική παραμετροποίηση το sum to zero που συγκρίνουμε με τον μέσο της μεταβλητής απόκρισης για όλα τα επίπεδα.

Για παράδειγμα θα συγκρίνουμε το ίδιο μοντέλο με sum to zero και corner constraints

Αλλάζοντας αυτή την παραμετροποίηση δεν αλλάζει το μοντέλο και η αποδοτικότητα του αλλά η συντελεστές και η ερμηνεία τους, γι'αυτο έχουμε και το ίδιο R^2 . Τώρα το intercept συμβολίζει τον συνολικό μέσο για κάθε ομάδα.

	socst			socst		
Predictors	*Estimates*	*CI*	*p*	*Estimates*	*CI*	*p*
(Intercept)	25.95	18.26 – 33.64	**<0.001**	26.29	18.43 – 34.14	**<0.001**
prog [1]	3.17	1.33 – 5.00	**0.001**			
prog [2]	-3.50	-5.54 – -1.46	**0.001**			
math	0.49	0.34 – 0.63	**<0.001**	0.49	0.34 – 0.63	**<0.001**
prog [academic]				2.83	-0.39 – 6.06	0.085
prog [vocation]				-3.83	-7.41 – -0.25	**0.036**
Observations	200			200		
R^2 / R^2 adjusted	0.348 / 0.338			0.348 / 0.338		

Figure 7: Σύγκριση μοντέλων με διαφορετική παραμετροποίηση

Ο συντελεστής της συνεχής μεταβλητής παρέμεινε ίδιος. Οι συντελεστές των κατηγορικών αναφέρονται στην μέση μεταβολή αυτής της ομάδας από τον μέσο.

Το μοντέλο είναι :

socst(Κοινωνικές Επιστήμες)=β_0+β_1*Math(Μαθηματικά)+β_2*prog[1](Πρόγραμμα Σπουδών academic) + β_3*prog[2](Πρόγραμ Σπουδών vocation)

1. Το $\beta_0 = 25.95$ είναι η μέση βαθμολογία για κάθε πρόγραμμα σπουδών όταν ο βαθμός στα μαθηματικά είναι 0.
2. Το $\beta_1 = 0.49$ είναι το ίδιο όπως πριν.
3. Το $\beta_2 = 3.17$ είναι η διαφορά ενός φοιτητή του ακαδημαϊκού προγράμματος από τον μέσο β_0.
4. Το $\beta_3 = -3.5$ είναι η διαφορά ενός φοιτητή του επαγγελματικού προγράμματος από τον μέσο β_0.

Παράρτημα

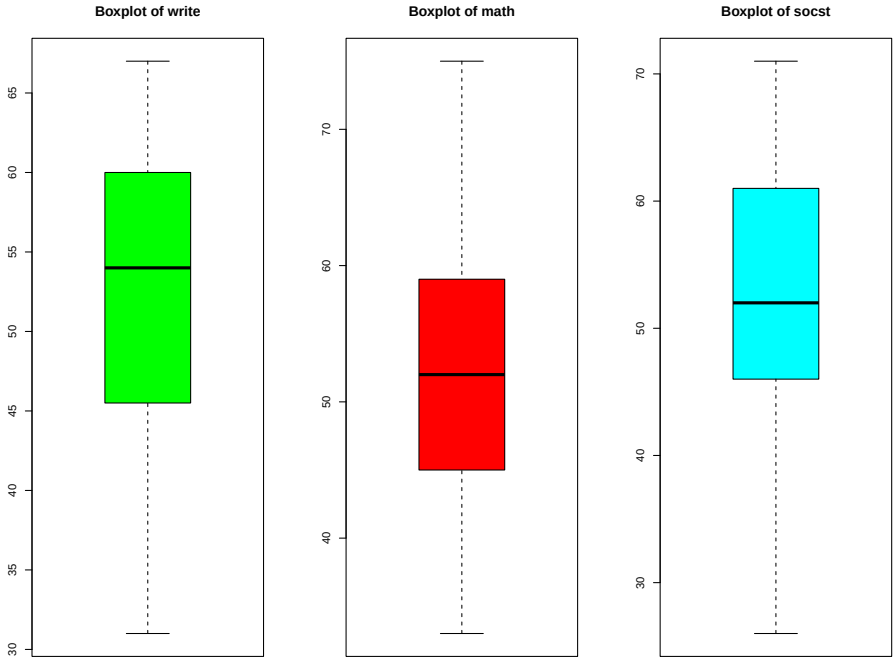

Figure 8: Boxplots των βαθμολογιών

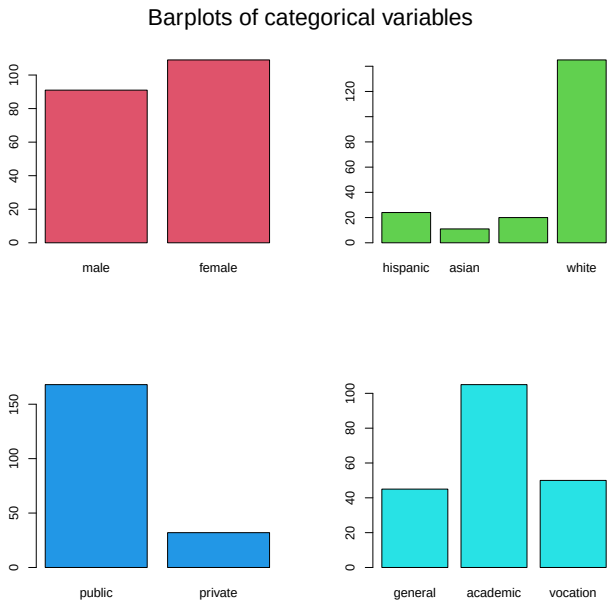

Figure 9: Barplots των κατηγορικών

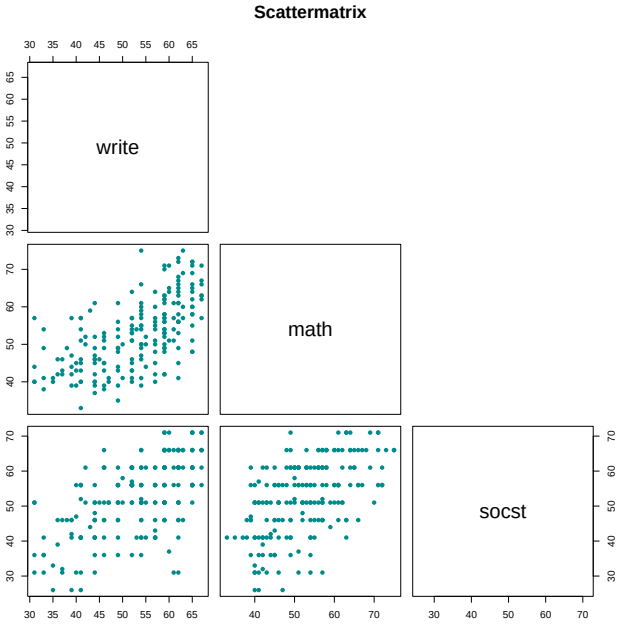

Figure 10: scattermatrix των βαθμολογιών

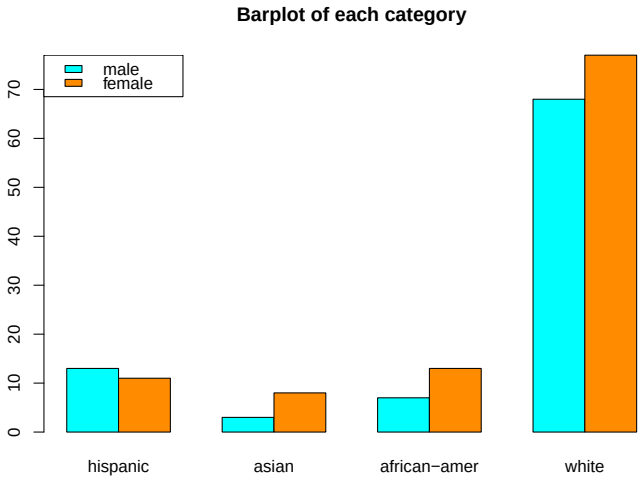

Figure 11:

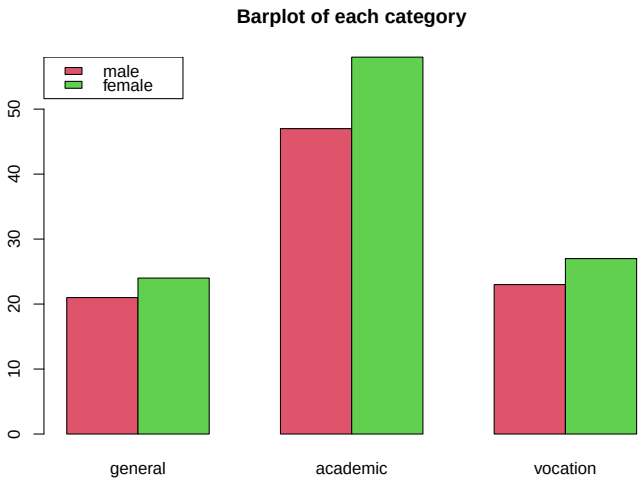

Figure 12:

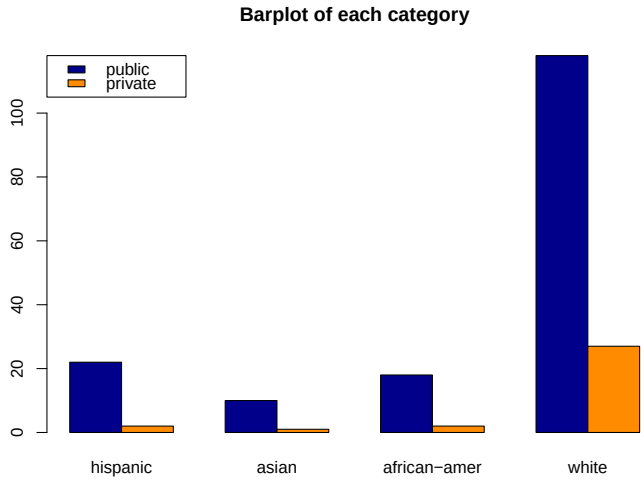

Figure 13:

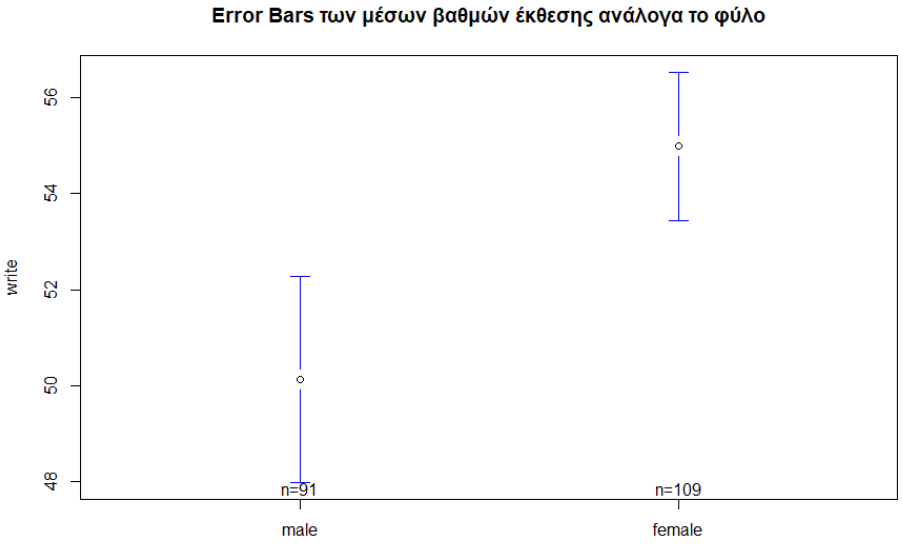

Figure 14:

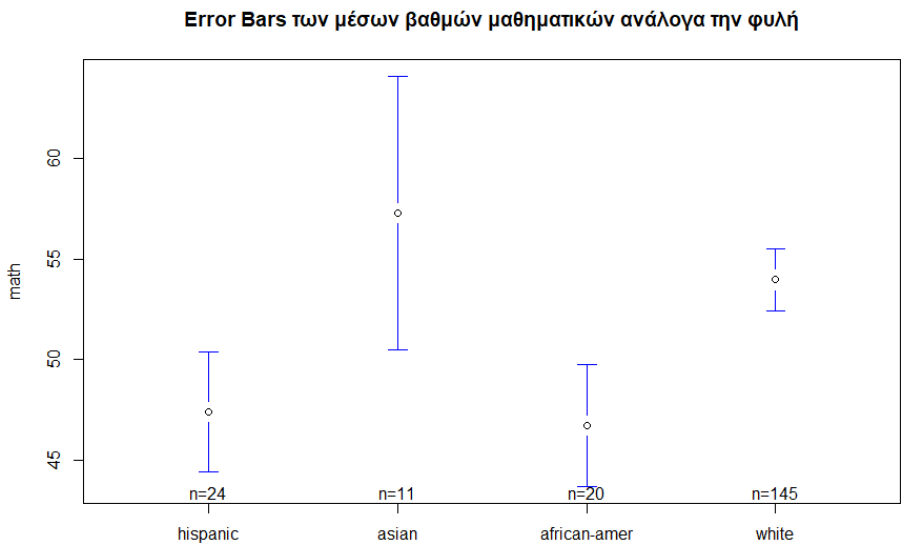

Figure 15:

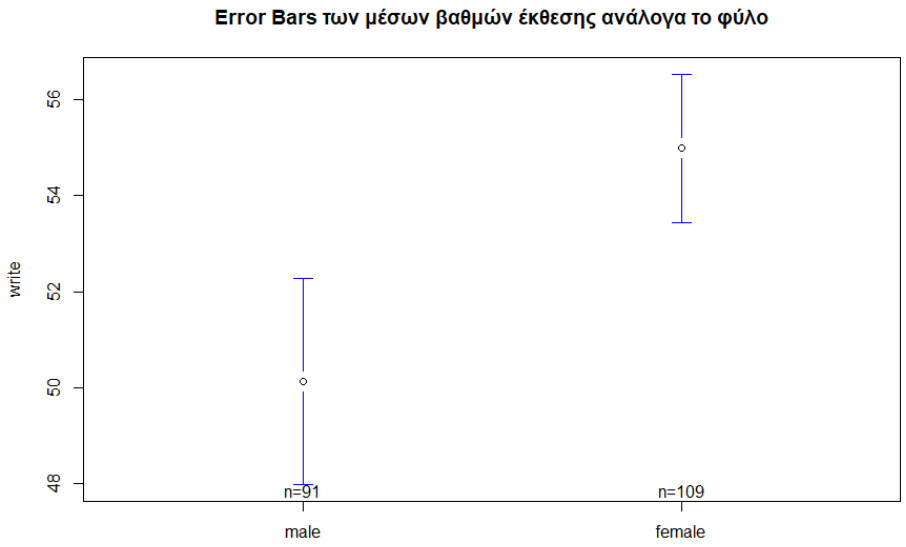

Figure 16:

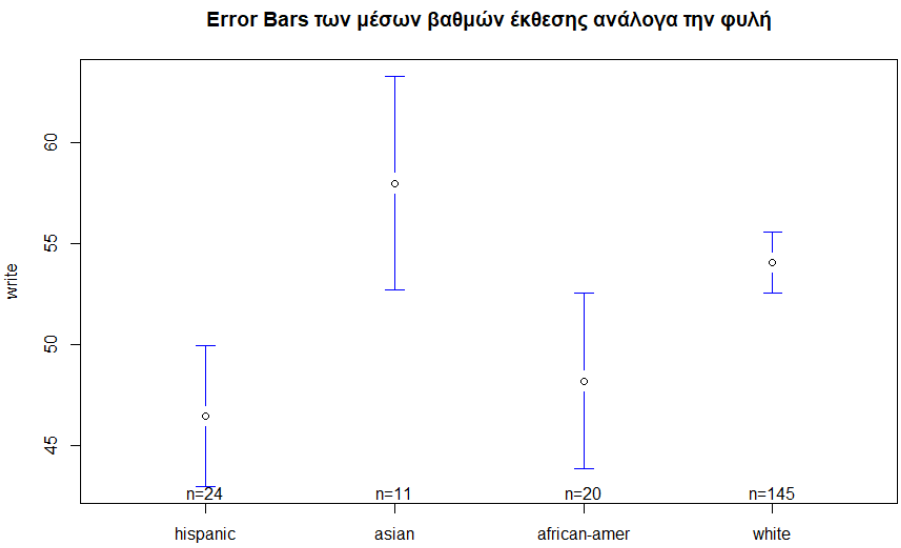

Figure 17:

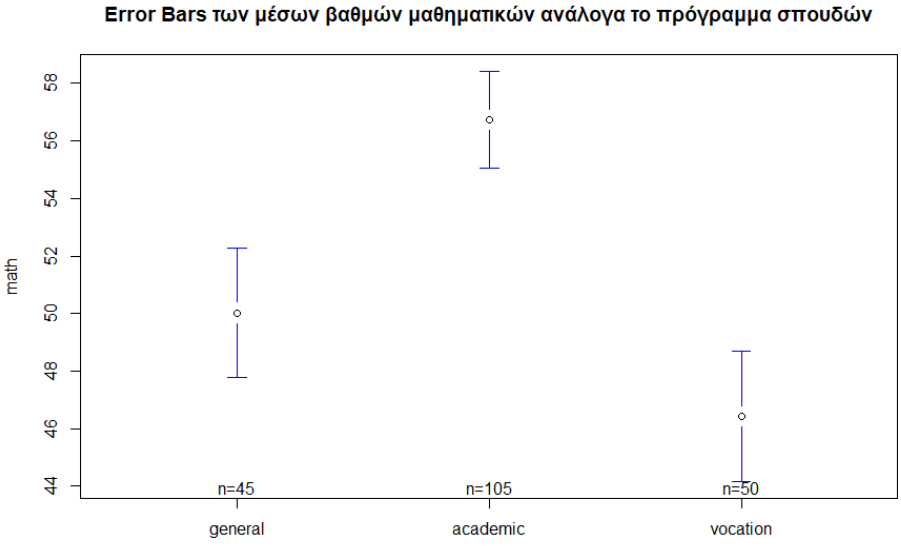

Figure 18:

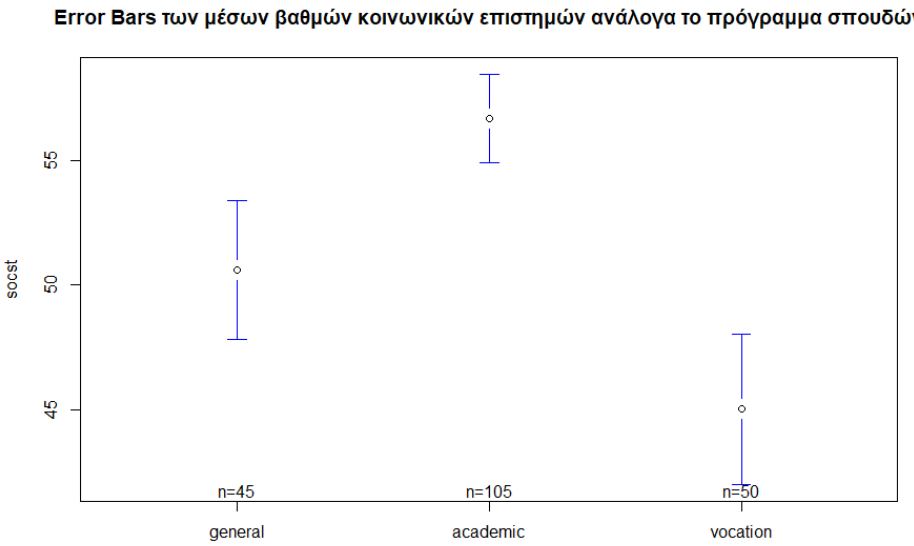

Figure 19:

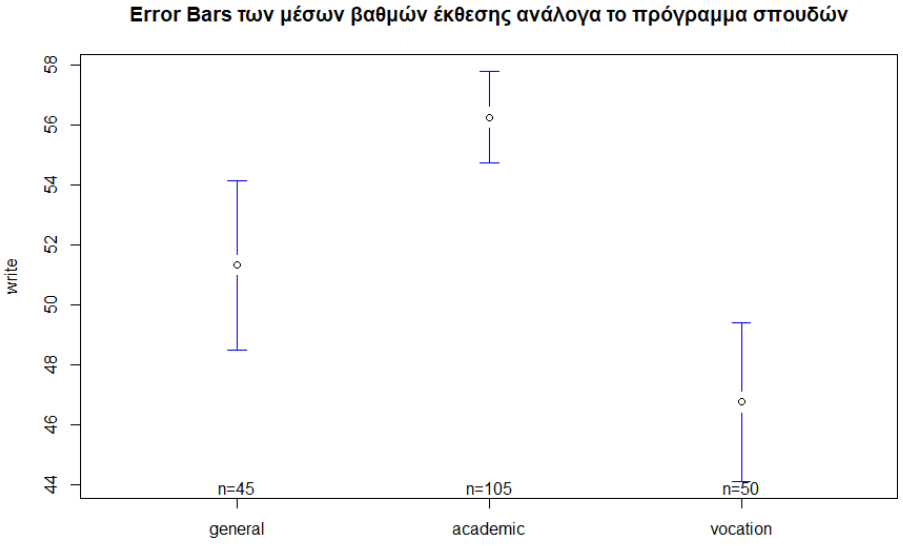

Figure 20:

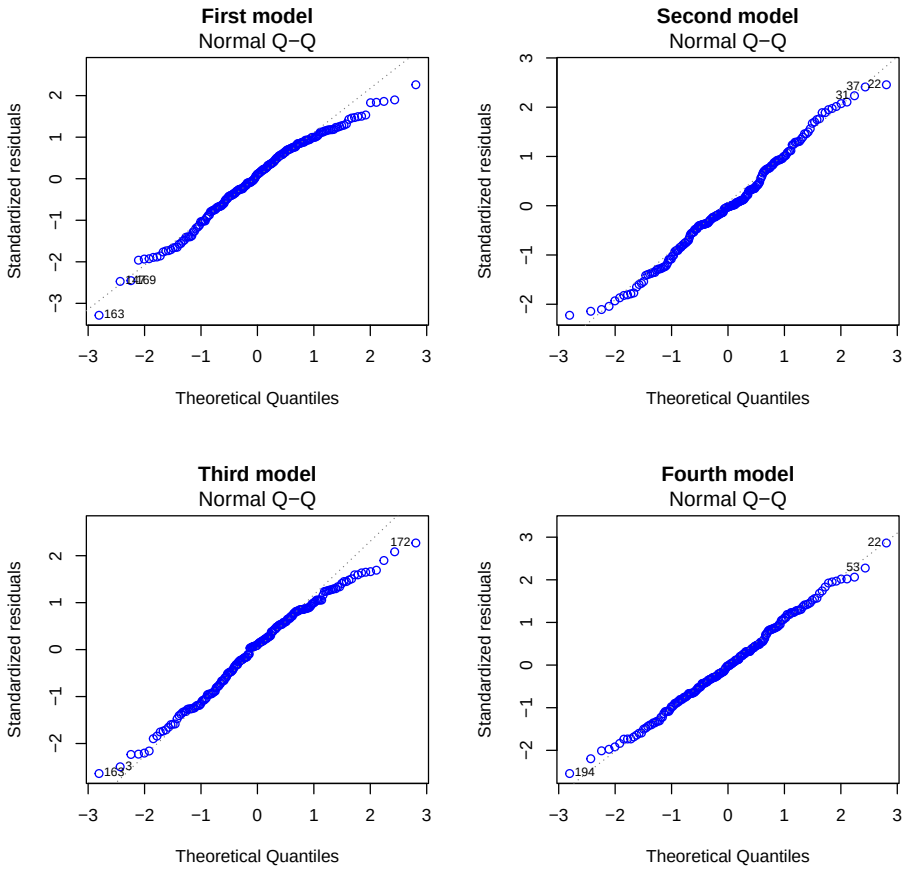

Figure 21: qqplots των μοντέλων

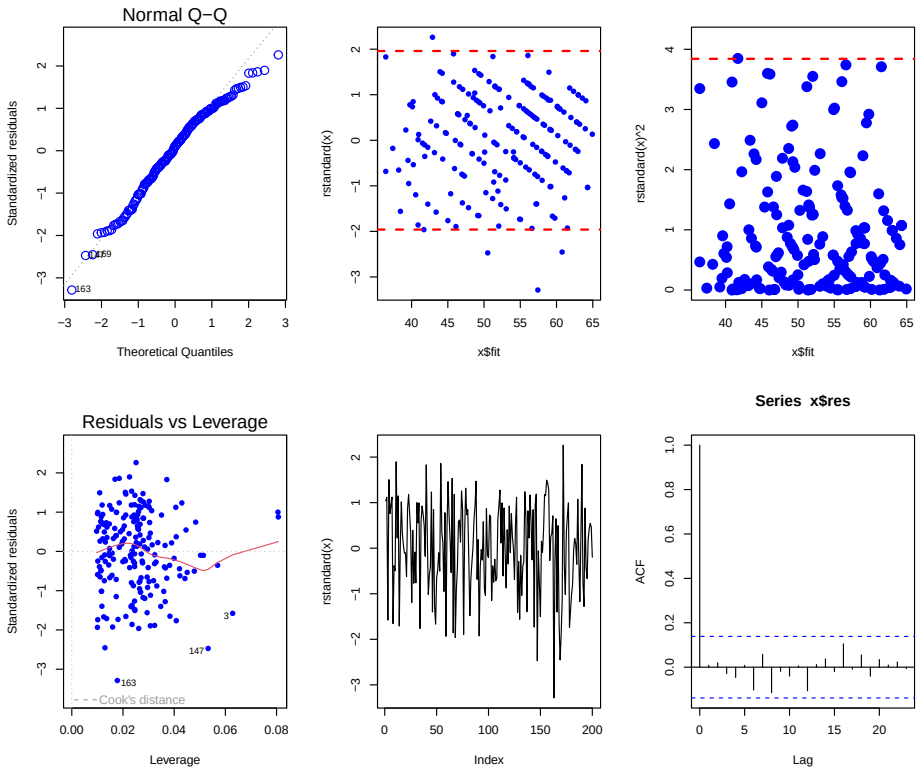

Figure 22: Διαγράμματα ελέγχων υποθέσεων πρώτου μοντέλου

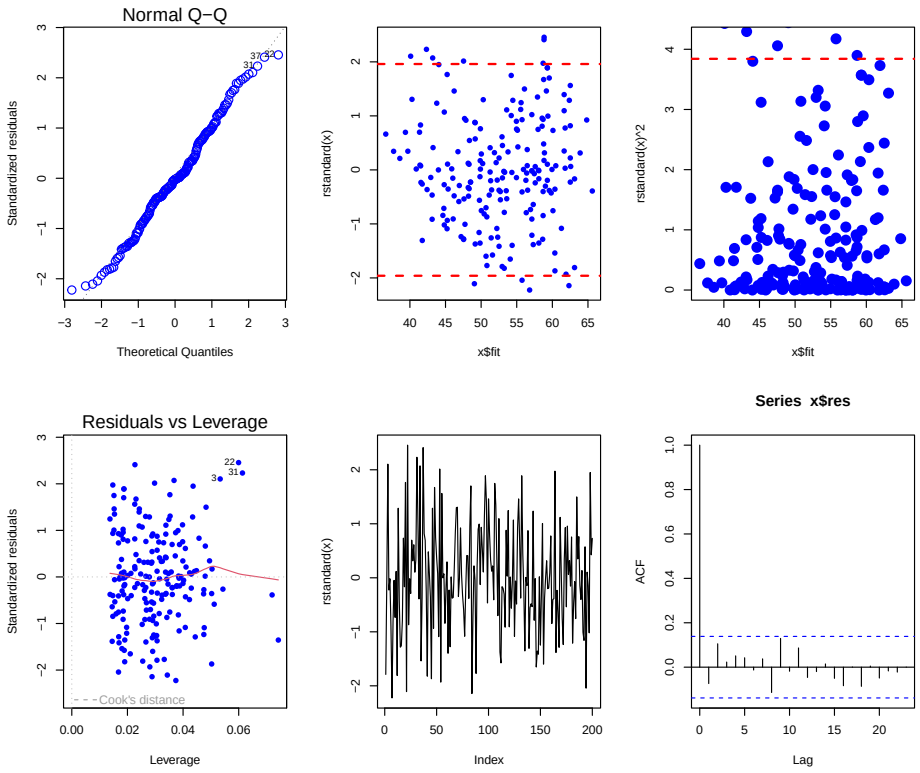

Figure 23: Διαγράμματα ελέγχων υποθέσεων δεύτερου μοντέλου

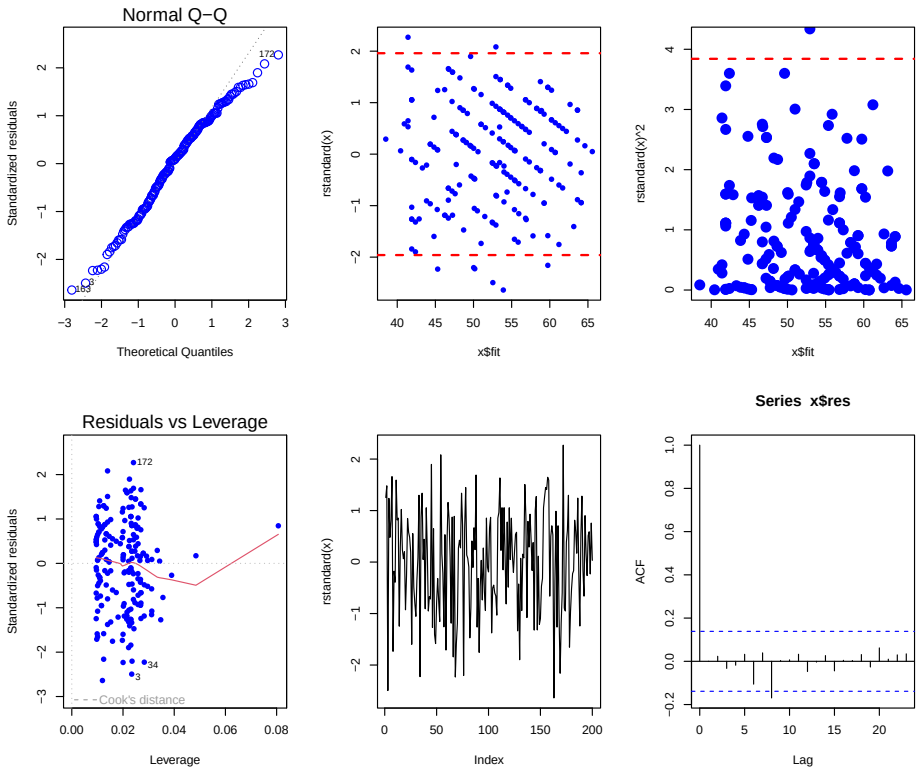

Figure 24: Διαγράμματα ελέγχων υποθέσεων τρίτου μοντέλου

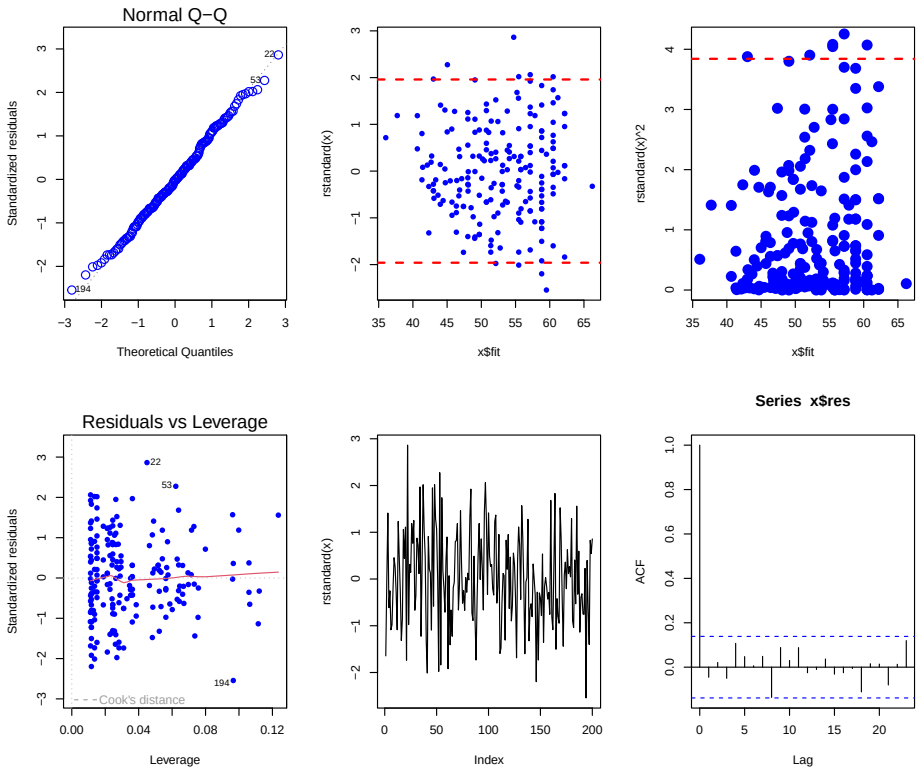

Figure 25: Διαγράμματα ελέγχων υποθέσεων τέταρτου μοντέλου