ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ

ΠΡΟΓΡΑΜΜΑ ΠΡΟΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

ΜΑΙΟΣ 2023

‘ΕΡΕΥΝΑ ΓΙΑ ΤΟΥΣ ΠΟΛΙΤΕΣ ΤΟΥ ILINOIS’

Γουέρα Ευαγγελία-Ζωή

Ανάλυση Δεδομένων

Διδάσκοντες: Ι.Ντζούφρας, Ξ.Πεντελή

Περιεχόμενα

ΠΕΡΙΛΗΨΗ

Ο στόχος της παρούσας έρευνας είναι να αναλύσει το πως επηρεάζεται ο δείκτης πλούτου των ατόμων με βάση τα χαρακτηριστικά τους όπως το φύλο, η ηλικία, το ζώδιο, η οικογενειακή κατάσταση, η κατάσταση υγείας αλλά και ο χρόνος ολοκλήρωσης των σπουδών τους και να μπορέσει να το προβλέψει. Για αυτόν τον σκοπό χρησιμοποιήθηκε ένα σετ δεδομένων για πολίτες του Illinois το οποίο αναλύθηκε κατάλληλα. Παρατηρήσαμε, λοιπόν, ότι ο δείκτης πλούτου επηρεάζεται από την οικογενειακή κατάσταση, την ηλικία και την ηλικία ολοκλήρωσης σπουδών λαμβάνοντας υπόψιν και την σχέση αλληλεπίδρασης μεταξύ της οικογενειακής κατάστασης και της ηλικίας. Επίσης, κατά την διάρκεια της έρευνας ,αξιολογήθηκε η ηλικία ολοκλήρωσης σπουδών για άτομα διαφορετικού φύλου και ζωδίου καταλήγοντας στο ότι η ηλικία ολοκλήρωσης σπουδών κατά μέσο όρο διαφέρει μεταξύ γυναικών και ανδρών ενώ για άτομα διαφορετικού ζωδίου δεν διαφέρει.

1. ΕΙΣΑΓΩΓΗ

Ο δείκτης πλούτου αποτελεί ένα σημαντικό μέτρο για την ανίχνευση και την ανάλυση της οικονομικής ανισότητας μεταξύ των πολιτών μιας περιοχής. Στο πλαίσιο αυτής της εργασίας, θέλουμε να εξετάσουμε πώς διαφορετικοί παράγοντες όπως το φύλο, η ηλικία, το ζώδιο, η οικογενειακή κατάσταση, η κατάσταση υγείας και ο χρόνος ολοκλήρωσης των σπουδών επηρεάζουν την οικονομική κατάσταση των πολιτών ώστε να αναδείξουμε τυχόν πρότυπα ή τάσεις και να εκτιμήσουμε ένα μοντέλο για το πλούτο των ατόμων σε σχέση με τα παραπάνω χαρακτηριστικά. Για το σκοπό αυτό θα χρησιμοποιήσουμε ένα σετ δεδομένων που προέρχεται από έρευνα την εργατικού δυναμικού που πραγματοποιείται κάθε έτος στη πολιτεία Ilinois των ΗΠΑ περιέχοντας πληροφορίες σχετικά με τα παραπάνω χαρακτηριστικά των ατόμων. Το σετ δεδομένων μετά από τροποποιήσεις , που εξηγούνται στο επόμενο κεφάλαιο, περιέχει 774 παρατηρήσεις και 7 μεταβλητές (Πίνακα 1).

Αριθμός μεταβλητής	**Όνομα**	**Τύπος Μεταβλητής**	**Σημασία**	**Τιμές**
1	marital	κατηγορική	Οικογενειακή κατάσταση	παντρεμένος, χήρος, χωρισμένος, σε διάσταση, όχι παντρεμένος
2	age	αριθμητική	Ηλικία	
3	educ	αριθμητική	Ηλικία ολοκλήρωσης σπουδών	
4	sex	κατηγορική	Φύλο	άνδρας, γυναίκα
5	health	κατηγορική	Κατάσταση υγείας	εξαιρετική, καλή, μέτρια, κακή
6	wealth	αριθμητική	Δείκτης πλούτου	
7	zodiac	κατηγορική	Ζώδιο ερωτώμενου	κριός ,ταύρος , δίδυμος, καρκίνος, λέων,…, ιχθύς

Πίνακας 1: Πίνακας Δεδομένων

2. ΑΠΟΤΕΛΕΣΜΑΤΑ

2.1. Περιγραφική Ανάλυση

Για να υλοποιήσουμε την ανάλυσή μας θα χρησιμοποιήσουμε το στατιστικό πακέτο R το οποίο θα μας βοηθήσει στον υπολογισμό των ελέγχων αλλά και στη διαγραμματική απεικόνιση των αποτελεσμάτων μας. Αρχικά θα εισάγουμε τα δεδομένα μας και όποιες βιβλιοθήκες χρειάζονται για την ανάλυση μας. Παρατηρούμε ότι οι μεταβλητές wealth, education και age είναι ποσοτικές μεταβλητές ενώ στα δεδομένα μας έχουν καταγραφεί ως κατηγορικές. Για αυτόν τον λόγο τις μετατρέπουμε σε ποσοτικές. Επίσης παρατηρούμε την μεταβλητή id η οποία δεν είναι χρήσιμη και για αυτό αφαιρείται αλλά και την ύπαρξη μη διαθέσιμων τιμών (NA), οι οποίες θα χρειαστεί να αφαιρεθούν χάνοντας έτσι και ένα μέρος της πληροφορίας του σετ δεδομένων. Αρά από το αρχικό δείγμα 1000 παρατηρήσεων καταλήγω σε ένα δείγμα 773 παρατηρήσεων. Τα δεδομένα διαμορφώνονται όπως βλέπουμε παραπάνω στον Πίνακα 1.

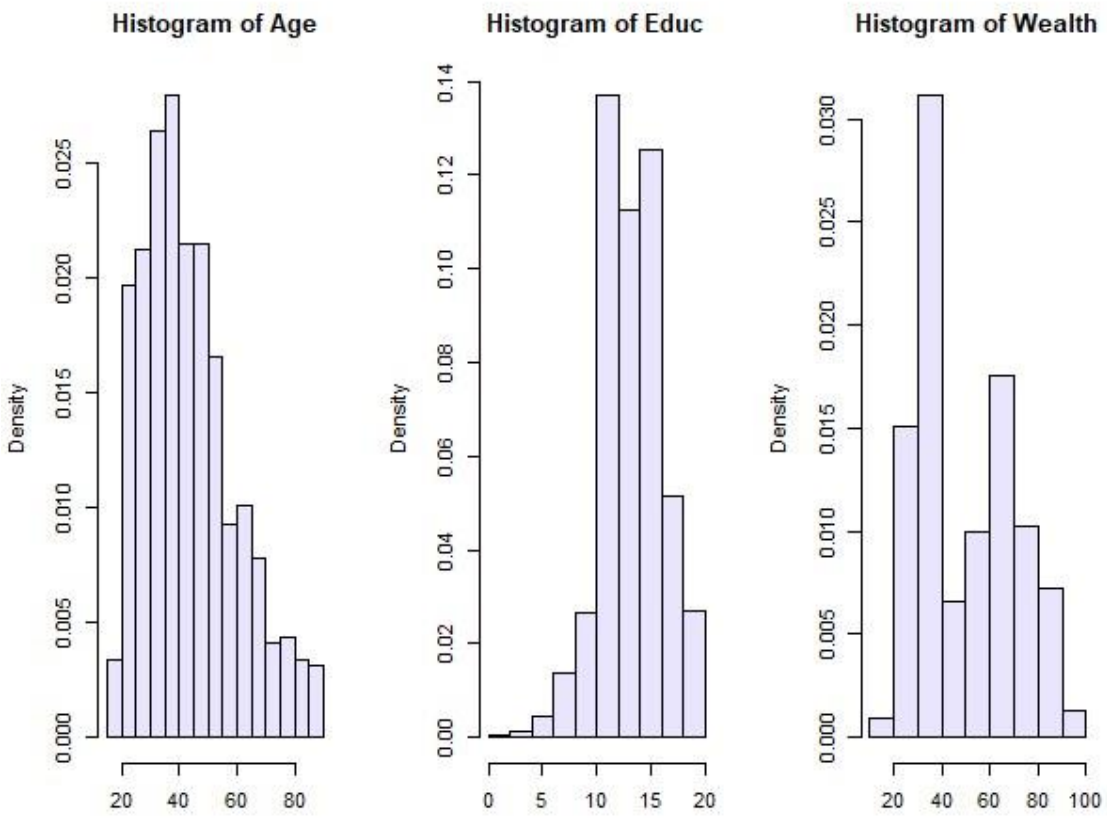
Συνεχίζοντας, θέλουμε να εξετάσουμε κάθε μεταβλητή ξεχωριστά και να δούμε τις τιμές που περιέχει. Για τις κατηγορικές μεταβλητές μπορούμε να δούμε τις συχνότητες με τις οποίες έχει εμφανιστεί η κατηγορία της κάθε μεταβλητής. Ενώ για τις μεταβλητές που είναι ποσοτικές μπορούμε να δούμε την μέση τιμή ,την τυπική απόκλιση ,τη διάμεσο ,την ασυμμετρία αλλά και την κύρτωση (Πίνακας 2). Ιδιαίτερο ενδιαφέρον έχουν οι δύο τελευταίες καθώς για τιμές κοντά στο μηδέν οι κατανομές των μεταβλητών προσεγγίζουν την κανονική κατανομή. Από τις τιμές των μεταβλητών παρατηρούμε ότι καμία δεν προσεγγίζει την κανονική κατανομή.

Στήλη1	**Μέσος**	**Τυπική απόκλιση**	**Διάμεσος**	**Μικρότερη τιμή**	**Μεγαλύτερη τιμή**	**Ασυμμετρία**	**Κύρτωση**
age	43,8	16,09	41	19	89	0,76	-0,02
educ	13,9	2,92	14	0	20	-0,28	0,71
wealth	49,97	19,41	45,25	17,1	97,2	0,45	-0,98

Πίνακας 2: Πίνακας περιγραφικών μέτρων ποιοτικών μεταβλητών

Εξετάζουμε την υπόθεση αυτή κάνοντας τον έλεγχο Shapiro − Wilk και τα ανάλογα διαγράμματα (Ιστόγραμμα ή QQplot) δίνοντας μεγαλύτερη βαρύτητα στα διαγράμματα καθώς όταν το μέγεθος του δείγματος είναι μεγάλο, τα τεστ τείνουν να είναι πιο ευαίσθητα. Διαγραμματικά, αλλά και με την χρήση

του παραπάνω ελέγχου, επιβεβαιώνουμε ότι ούτε η age, ούτε η education ,ούτε και η wealth προσεγγίζουν την κανονική κατανομή (p-value = 3.223e-16, p-value = 9.123e-13, p-value < 2.2e-16)(Σχήμα 1). Παρατηρούμε συγκεκριμένα ότι η age είναι θετικά ασύμμετρη και πλατύκυρτη , η educ είναι αρνητικά ασύμμετρη και λεπτόκυρτη και η wealth είναι θετικά ασύμμετρη και πλατύκυρτη.



Σχήμα 1: Διαγράμματα πυκνότητας πιθανότητας για τις μεταβλητές age, educ, wealth

2.2. Ανάλυση Σχέσεων ανά δύο

Η διερεύνηση της κάθε μεταβλητής ξεχωριστά δεν μας βοηθάει να διερευνήσουμε τις σχέσεις μεταξύ των μεταβλητών ,γι' αυτό θέλουμε να εξετάσουμε αυτές τις σχέσεις πιο διεξοδικά. Αρχικά κάνοντας τον πίνακα συσχετίσεων του Spearman (μη παραμετρικός έλεγχος συσχέτισης), αφού για τις μεταβλητές δεν ισχύει η κανονικότητα, βλέπουμε την ύπαρξη γραμμικής συσχέτισης μεταξύ wealth και education (ρ=0,59) ενώ για τις υπόλοιπες δεν παρατηρούμε γραμμική συσχέτιση. Άρα στην ανάλυση μας δεν θα υπάρχει

πρόβλημα πολυσυγραμμικότητας. Για το πρόβλημα μας, οι σχέσεις που θα είχε ενδιαφέρον να ερευνήσουμε είναι οι εξής:

- Δείκτης πλούτου κάθε ερωτώμενου με βάση την οικογενειακή κατάσταση (wealth-marital)
- Δείκτης πλούτου κάθε ερωτώμενου με βάση το φύλο (wealth-sex)
- Δείκτης πλούτου κάθε ερωτώμενου με βάση την κατάσταση υγείας (wealth-health)
- Δείκτης πλούτου κάθε ερωτώμενου με βάση το ζώδιο (wealth-zodiac)

Για τη μελέτη των παραπάνω σχέσεων, εξετάζοντας σχέση κατηγορικής και ποσοτικής μεταβλητής, θα χρησιμοποιήσουμε ελέγχους t-test και anova , διαγράμματα qqnorm και boxplot, ελέγχους Shapiro-wilk για την κανονικότητα ,μη παραμετρικούς ελέγχους Kruskal-wallis και Wilcoxon test στην περίπτωση που απορρίπτεται αλλά και pairwise-wilcox test για σύγκριση μέσων/διαμέσων. Αρχικά για την σχέση πλούτου και οικογενειακής κατάστασης εφαρμόζουμε anova λόγω πολλών κατηγοριών και παρατηρούμε από τον έλεγχο αλλά και διαγραμματικά ότι η κανονικότητα απορρίπτεται (S-W p-value = 4.197e-16). Ενώ τα δείγματα για τις ομάδες είναι μεγάλα (n>50) συμπεραίνουμε ότι ο μέσος δεν είναι κατάλληλο μέτρο περιγραφής της κεντρικής θέσης λόγω ύπαρξης ακραίων τιμών. Γι' αυτό το λόγο πραγματοποιούμε τον μη παραμετρικό έλεγχο Kruskal-Wallis για να ελέγξουμε την ισότητα των διαμέσων από όπου και συμπεραίνουμε ότι διαφέρουν. Για αυτό εφαρμόζουμε pairwise-wilcox test για να συγκρίνουμε τις διαμέσους και να δούμε ποιοι συγκεκριμένα διαφέρουν καταλήγοντας στο ότι ο μέσος δείκτης πλούτου διαφέρει ανάλογα την οικογενειακή κατάσταση και συγκεκριμένα ότι οι παντρεμένοι έχουν διαφορετικό μέσο δείκτη από τους υπολοίπους.

Για την σχέση πλούτου και φύλου παρατηρούμε από τον έλεγχο αλλά και διαγραμματικά ότι η κανονικότητα απορρίπτεται (S-W p-value = 1.121e-12, S-W p-value= 1.009e-12). Έχοντας δύο κατηγορίες για το φύλο, εφαρμόζουμε μη παραμετρικό έλεγχο Wilcox test για την ισότητα των διαμέσων, καθώς παρατηρούμε ακραίες τιμές ενώ έχουμε μεγάλο δείγμα στις ομάδες . Από τον έλεγχο συμπεραίνουμε ότι οι διάμεσοι είναι ίσοι άρα ότι ο μέσος δείκτης πλούτου για τους άνδρες δεν διαφέρει από το μέσο δείκτη πλούτου για τις γυναίκες.

Για την σχέση πλούτου και κατάσταση υγείας εφαρμόζουμε anova λόγω πολλών κατηγοριών και παρατηρούμε από τον έλεγχο αλλά και διαγραμματικά ότι η κανονικότητα απορρίπτεται (S-W p-value < 2.2e-16). Ενώ τα δείγματα για τις ομάδες είναι μεγάλα (n>50) συμπεραίνουμε ότι ο μέσος δεν είναι κατάλληλο μέτρο περιγραφής της κεντρικής θέσης λόγω ύπαρξης ακραίων τιμών. Γι' αυτό το λόγο πραγματοποιούμε τον μη παραμετρικό έλεγχο Kruskal-Wallis για να ελέγξουμε την ισότητα των διαμέσων από όπου και συμπεραίνουμε ότι διαφέρουν. Εφαρμόζουμε pairwise-wilcox test για να συγκρίνουμε τις διαμέσους συμπεραίνοντας ότι ο μέσος δείκτης πλούτου διαφέρει με βάση την κατάσταση υγείας και

συγκεκριμένα ότι ο μέσος δείκτης πλούτου διαφέρει για αυτούς που έχουν εξαιρετική κατάσταση (Excellent) και διαφέρει μεταξύ αυτών που έχουν καλή και μερική κατάσταση (Good-Fair).

Τέλος, για την σχέση πλούτου και ζωδίου εφαρμόζουμε anova λόγω πολλών κατηγοριών και παρατηρούμε από τον έλεγχο αλλά και διαγραμματικά ότι η κανονικότητα απορρίπτεται (S-W p-value <2.2e-16). Ενώ τα δείγματα για τις ομάδες είναι μεγάλα (n>50) συμπεραίνουμε ότι ο μέσος δεν είναι κατάλληλο μέτρο περιγραφής της κεντρικής θέσης λόγω ύπαρξης ακραίων τιμών. Γι' αυτό το λόγο πραγματοποιούμε τον μη παραμετρικό έλεγχο Kruskal-Wallis για να ελέγξουμε την ισότητα των διαμέσων από όπου και συμπεραίνουμε ότι δεν διαφέρουν άρα ότι ο μέσος δείκτης πλούτου δεν διαφέρει μεταξύ των ζωδίων.

Πέρα από τις παραπάνω σχέσεις, ζητήθηκε να αναλυθούν και οι σχέσεις:

- Ηλικία ολοκλήρωσης σπουδών για άτομα διαφορετικού φύλου (educ-sex)
- Ηλικία ολοκλήρωσης σπουδών για άτομα διαφορετικού ζωδίου (educ-zodiac)

Για την σχέση ηλικίας ολοκλήρωσης σπουδών και φύλου παρατηρούμε από τον έλεγχο αλλά και διαγραμματικά ότι η κανονικότητα απορρίπτεται (S-W p-value = 8.113e-08, S-W p-value = 1.064e-08). Έχοντας δύο κατηγορίες για το φύλο, εφαρμόζουμε μη παραμετρικό έλεγχο Wilcox test για την ισότητα των διαμέσων, καθώς παρατηρούμε ακραίες τιμές ενώ έχουμε μεγάλο δείγμα στις ομάδες . Από τον έλεγχο συμπεραίνουμε ότι οι διάμεσοι διαφέρουν άρα ότι η μέση ηλικία ολοκλήρωσης σπουδών για τα δύο φύλα. Ακόμη ,για την σχέση ηλικίας ολοκλήρωσης σπουδών και ζωδίου πραγματοποιούμε anova λόγω πολλών κατηγοριών και παρατηρούμε από τον έλεγχο αλλά και διαγραμματικά ότι η κανονικότητα απορρίπτεται (S-W p-value = 4.544e-09). Ενώ τα δείγματα για τις ομάδες είναι μεγάλα (n>50) συμπεραίνουμε ότι ο μέσος δεν είναι κατάλληλο μέτρο περιγραφής της κεντρικής θέσης λόγω ύπαρξης ακραίων τιμών. Γι' αυτό το λόγο πραγματοποιούμε τον μη παραμετρικό έλεγχο Kruskal-Wallis για να ελέγξουμε την ισότητα των διαμέσων από όπου και συμπεραίνουμε ότι δεν διαφέρουν άρα ότι η μέση ηλικία ολοκλήρωσης σπουδών δεν διαφέρει μεταξύ των ζωδίων.

2.3. Στατιστικό Μοντέλο

Αφού έχουμε ελέγξει τα δεδομένα μας εκτενώς και έχουμε κατανοήσει τις σχέσεις μεταξύ των μεταβλητών και κυρίως του δείκτη πλούτου με τις υπόλοιπες μεταβλητές μπορούμε να προχωρήσουμε στη

προσαρμογή ενός μοντέλου για να εκτιμήσουμε το πλούτο των ατόμων σε σχέση με τα χαρακτηριστικά τους. Αρχικά παίρνουμε το πλήρες μοντέλο το οποίο θα είναι της μορφής:

Δείκτης πλούτου=βο + β1*(Οικογενειακή κατάσταση) + β2*(Ηλικία) +β3*(Ηλικία ολοκλήρωσης σπουδών) +β4*(Φύλο) +β5*(Κατάσταση υγείας) +β6*(Ζώδιο)

Πριν ελέγξουμε τις προϋποθέσεις που πρέπει να ικανοποιεί το μοντέλο, θα ελέγξουμε ποιες μεταβλητές είναι στατιστικά σημαντικές για την εκτίμηση του δείκτη. Εφαρμόζοντας anova στο πλήρες μοντέλο παρατηρούμε από τα αποτελέσματα του p-value για την κάθε μεταβλητή ότι το φύλο, η κατάσταση υγείας και τα ζώδια δεν είναι στατιστικά σημαντικά. Άρα δεν αποτελούν χαρακτηριστικά σημαντικά για την εκτίμηση του δείκτη πλούτου. Επίσης με την χρήση μεθόδου επιλογής μοντέλων stepwise regression επιβεβαιώνουμε τα παραπάνω αποτελέσματα καταλήγοντας στο ότι το καλύτερο μοντέλο θα είναι το:

Δείκτης πλούτου=βο + β1*(Οικογενειακή κατάσταση) + β2*(Ηλικία) +β3*(Ηλικία ολοκλήρωσης σπουδών)

Στο παραπάνω μοντέλο ,αρχικά, παρατηρούμε τις ψευδομεταβλητές που έχουν δημιουργηθεί για τη κατηγορική μεταβλητή ‘οικογενειακή κατάσταση’ η οποία αποτελείται από πέντε κατηγορίες (Married, Widowed, Divorced, Separated, Never Married). Συγκεκριμένα, για αυτά τα επίπεδα διακρίνουμε τέσσερις ψευδομεταβλητές για τις οποίες θα ισχύει ότι:

1. 1 αν είναι windowed, 0 αλλιώς
2. 1 αν είναι divorced, 0 αλλιώς
3. 1 αν είναι separated, 0 αλλιώς
4. 1 αν είναι never married, 0 αλλιώς ψευδομεταβλητές

Το πρώτο επίπεδο που είναι το married θα ισχύει για windowed=divorced=separated=never married = 0. Άρα το τελικό μοντέλο με την χρήση των ψευδομεταβλητών παίρνει την μορφή:

Δείκτης πλούτου=βο + β1*(Ψευδομεταβλητη1) + β2*(Ψευδομεταβλητη2) + β3*(Ψευδομεταβλητη3) + β4*(Ψευδομεταβλητη4) + β5*(Ηλικία) +β6*(Ηλικία ολοκλήρωσης σπουδών)

Στην συνέχεια, παρατηρούμε ότι το επίπεδο Widowed δεν είναι στατιστικά σημαντικό στο τελικό μοντέλο λόγω του p-value όμως δεν θα το αφαιρέσουμε καθώς δεν γίνεται να αφαιρέσούμε ένα επίπεδο.

Γενικά από το μοντέλο μπορούμε να βγάλουμε ολόκληρη μεταβλητή και όχι επίπεδα. Έτσι αφού η μεταβλητή 'οικογενειακή κατάσταση' είναι στατιστικά σημαντική θα την κρατήσουμε μαζί με όλα τα επίπεδα της.

Αναλύοντας τα χαρακτηριστικά του τελικού μοντέλου βλέπουμε ότι υπερτερεί του σταθερού μοντέλου (Δείκτης πλούτου=βο) από την F-statistic (p-value< 2.2e-16). Παρόλα αυτά δεν προσαρμόζει και τόσο καλά τα δεδομένα (Adjusted R-squared= 0.3882) και επιπλέον βλέπουμε ότι η σταθερά δεν βγάζει νόημα αν και είναι στατιστικά σημαντική. Αν πάμε να ερμηνεύσουμε τη σταθερά καταλήγουμε στο συμπέρασμα ότι για μηδενική οικογενειακή κατάσταση ,για ηλικία μηδέν και ηλικία ολοκλήρωσης σπουδών μηδέν η αναμενόμενη τιμή του δείκτη πλούτου θα είναι περίπου -8 ποσοστιαίες μονάδες. Γι' αυτό θα κεντροποιήσουμε τις ποιοτικές μας μεταβλητές, δηλαδή θα αφαιρέσουμε από αυτές τους μέσους τους για να μας βοηθήσει στην ερμηνεία. Με την κεντροποιηση παρατηρούμε ότι δεν αλλάζει τίποτα στο μοντέλο πέρα από την σταθερά και την ερμηνεία του. Για το ποσοστό της μεταβλητότας που εξηγεί το μοντέλο (Adjusted R-squared= 0.3882) θα μπορούσαμε να εφαρμόσουμε κάποιο μετασχηματισμό αλλά δεν θα μας έδινε πολύ επιπλέον πληροφορία.

	B	**Std.error**	**p-value**
Intercept	52,6	0,84	<2e-16
marital			
widowed	-2,71	2,33	0,246
divorced	-5,22	1,56	0,00089
separated	-7,64	2,64	0,0039
never married	-4,28	1,45	0,0032
age.centered	0,12	0,04	0,0029
educ.centered	4,01	0,19	<2e-16
R-squared/Adjusted R-squared	0.393/0.388		
AIC	6.416		

Πίνακας 3: Μοντέλο Πολλαπλής Παλινδρόμησης με τις στατιστικά σημαντικές μεταβλητές και την κεντροποίηση

Από τους συντελεστές του μοντέλου που εκτιμήθηκαν παρατηρούμε ότι:

- Η αναμενομένη τιμή του δείκτη πλούτου όταν η κεντροποιημένη ηλικία του ατόμου αυξηθεί κατά 1 έτος, με τα υπόλοιπα χαρακτηριστικά του να παραμένουν σταθερά, θα αυξηθεί περίπου κατά 0,12 ποσοστιαίες μονάδες . (β5)

- Η αναμενομένη τιμή του δείκτη πλούτου όταν η κεντροποιημένη ηλικία ολοκλήρωσης σπουδών του ατόμου αυξηθεί κατά 1 έτος, με τα υπόλοιπα χαρακτηριστικά να παραμένουν σταθερά, θα αυξηθεί περίπου κατά 4 ποσοστιαίες μονάδες . (β6)

- Η διαφορά της αναμενομένης τιμής του δείκτη πλούτου αν συγκρίνω τα widowed και married, με σταθερά τα υπόλοιπα , θα είναι περίπου 2,7 ποσοστιαίες μονάδες. Άρα για το επίπεδο widowed η αναμενομένη τιμή του δείκτη θα μειωθεί κατά 2,7 μονάδες . Δηλαδή αυτοί που είναι χήροι θα έχουν κατά μέσο όρο 2,7 σε ποσοστιαίες μονάδες μικρότερο δείκτη πλούτου από τους παντρεμένους . (β1)

- Η διαφορά της αναμενομένης τιμής του δείκτη πλούτου αν συγκρίνω τα divorced και married, με σταθερά τα υπόλοιπα , θα είναι περίπου 5,2 ποσοστιαίες μονάδες. Άρα για το επίπεδο divorced η αναμενομένη τιμή του δείκτη θα μειωθεί κατά 5,2 μονάδες . Δηλαδή αυτοί που είναι χωρισμένοι θα έχουν κατά μέσο όρο 5,2 σε ποσοστιαίες μονάδες μικρότερο δείκτη πλούτου από τους παντρεμένους. (β2)

- Η διαφορά της αναμενομένης τιμής του δείκτη πλούτου αν συγκρίνω τα separated και married, με σταθερά τα υπόλοιπα , θα είναι περίπου 7,6 ποσοστιαίες μονάδες. Άρα για το επίπεδο separated η αναμενομένη τιμή του δείκτη θα μειωθεί κατά 7,6 μονάδες . Δηλαδή αυτοί που είναι σε διάσταση θα έχουν κατά μέσο όρο 7,6 σε ποσοστιαίες μονάδες μικρότερο δείκτη πλούτου από τους παντρεμένους. (β3)

- Η διαφορά της αναμενομένης τιμής του δείκτη πλούτου αν συγκρίνω τα never married και married, με σταθερά τα υπόλοιπα , θα είναι περίπου 4,2 ποσοστιαίες μονάδες. Άρα για το επίπεδο never married η αναμενομένη τιμή του δείκτη θα μειωθεί κατά 4,2 μονάδες . Δηλαδή αυτοί που δεν είναι παντρεμένοι θα έχουν κατά μέσο όρο 4,2 σε ποσοστιαίες μονάδες μικρότερο δείκτη πλούτου από τους παντρεμένους. (β4)

- Η αναμενομένη τιμή του δείκτη πλούτου θα είναι περίπου 52,6 ποσοστιαίες μονάδες για κάποιον που ανήκει στο επίπεδο αναφοράς, δηλαδή για κάποιον που είναι παντρεμένος , όταν η κεντροποιημένη ηλικία κα η κεντροποιημένη ηλικία ολοκλήρωσης σπουδών είναι ίσα με μηδέν. (βο)

Όσον αφορά τις προϋποθέσεις που πρέπει να ικανοποιεί το μοντέλο, ελέγχουμε την κανονικότητα, την ομοσκεδαστικότητα, την γραμμικότητα και την τυχαιότητα των τυποποιημένων καταλοίπων. Αρχικά ,για την κανονικότητα, εφαρμόζοντας έλεγχο Shapiro-Wilk παρατηρούμε ότι απορρίπτεται (p-value = 0.0005288). Για την ομοσκεδαστικότητα, εφαρμόζοντας κατάλληλα τον έλεγχο Levene-test παρατηρούμε

επίσης ότι απορρίπτεται (p-value= 4.3e-08). Η υπόθεση της τυχαιότητας/ασυσχέτισης, από την άλλην, εφαρμόζοντας έλεγχο Durbin-Watson παρατηρούμε ότι δεν απορρίπτεται (p-value= 0.748). Κατασκευάζοντας και τα κατάλληλα διαγράμματα για τους παραπάνω ελέγχους ,με την χρήση της εντολής plot, δεν παρατηρούμε έντονα την παραβίαση των παραπάνω υποθέσεων. Αυτό συμβαίνει κυρίως λόγω μεγάλου δείγματος. Θα θεωρήσουμε ,λοιπόν, ότι δεν υπάρχει πρόβλημα με τις υποθέσεις του μοντέλου. Ακόμη, μετά από διαγραμματικό έλεγχο Cook's distance/Leverage βλέπουμε κάποιες ακραίες τιμές οι οποίες όμως δεν αποτελούν σημεία επιρροής άρα δεν μας προβληματίζουν.

Στην συνέχεια της ανάλυσης μας, σκεφτήκαμε να εφαρμόσουμε και το ρεαλιστικό μοντέλο. Δηλαδή το μοντέλο με τις αλληλεπιδράσεις των επίπεδων της κατηγορικής μεταβλητής με τις κεντροποιημένες ποιοτικές. Παρατηρούμε, με χρήση anova, ως στατιστικά σημαντικές τις μεταβλητές marital, educ και την αλληλεπίδραση age:marital. Επιβεβαιώνουμε τη σημαντικότητα εφαρμόζοντας την μέθοδο επιλογής μοντέλων stepwise regression η οποία δείχνει συγκεκριμένα ότι οι στατιστικά σημαντικές μεταβλητές είναι τα επίπεδα Divorced και Separated από την μεταβλητή marital, το educ και η αλληλεπίδραση της age με το επίπεδο never married της marital. Άρα στο τελικό μοντέλο θα κρατήσουμε όλα τα επίπεδα της marital και όλα τα επίπεδα της αλληλεπίδρασης age:marital αφού έστω και ένα από αυτά βγήκε στατιστικά σημαντικό. Επίσης, μπορεί το age να μην είναι στατιστικά σημαντικό αλλά επειδή είναι στατιστικά σημαντική η αλληλεπίδραση του προφανώς και θα περιέχετε στο μοντέλο. Έτσι, το τελικό μοντέλο για τις αλληλεπιδράσεις θα έχει την μορφή:

Δείκτης πλούτου=βο + β1*(Οικογενειακή κατάσταση) + β2*(Ηλικία) +β3*(Ηλικία ολοκλήρωσης σπουδών) + β4*(Ηλικία* Οικογενειακή κατάσταση)

Ή αλλιώς

Δείκτης πλούτου=βο + β1*(Ψευδομεταβλητη1) + β2*(Ψευδομεταβλητη2) + β3*(Ψευδομεταβλητη3) + β4*(Ψευδομεταβλητη4) + β5*(Ηλικία) + β6*(Ηλικία ολοκλήρωσης σπουδών) + β7*(Ηλικία* Ψευδομεταβλητή1) + β8*(Ηλικία*Ψευδομεταβλητή2)+β9*(Ηλικία*Ψευδομεταβλητή3)+ β10*(Ηλικία*Ψευδομεταβλητή4)

		B	Std.error	p-value
Intercept	52,6		0,83	<2e-16
age.centered	0,11		0,05	0,05
marital				
widowed	-3,58		3,77	0,34
divorced	-4,78		1,58	0,0026
separated	-7,81		2,63	0,003
never married	-1,7		1,73	0,32
educ.centered	3,96		0,19	<2e-16
age.centered:marital				
age.centered:widowed	0,03		0,13	0,78
age.centered:divorced	-0,18		0,11	0,12
age.centered:separated	-0,28		0,22	0,2
age.centered:never married	0,23		0,11	0,03
R-squared/Adjusted R-squared	0.4015/0.3936			
AIC	6.414			

Πίνακας 4: Μοντέλο Πολλαπλής Παλινδρόμησης με αλληλεπιδράσεις

Αναλύοντας τα χαρακτηριστικά του παραπάνω μοντέλου βλέπουμε ότι υπερτερεί του σταθερού μοντέλου (Δείκτης πλούτου=βο) από την F-statistic (p-value< 2.2e-16). Παρατηρούμε ότι προσαρμόζει τα δεδομένα λίγο καλύτερα από το προηγούμενο μοντέλο (Adjusted R-squared=0.3936) και επιπλέον από τους συντελεστές του που εκτιμήθηκαν παρατηρούμε ότι:

- Η αναμενόμενη τιμή του δείκτη πλούτου όταν η κεντροποιημένη ηλικία του ατόμου αυξηθεί κατά 1 έτος, θα αυξηθεί περίπου κατά 0,11 ποσοστιαίες μονάδες για το επίπεδο αναφοράς, δηλαδή για τους παντρεμένους. (β5)

- Η αναμενόμενη τιμή του δείκτη πλούτου όταν η κεντροποιημένη ηλικία ολοκλήρωσης σπουδών του ατόμου αυξηθεί κατά 1 έτος, θα αυξηθεί περίπου κατά 3,9 ποσοστιαίες μονάδες για το επίπεδο αναφοράς, δηλαδή για τους παντρεμένους. (β6)

- Η αναμενόμενη μεταβολή του δείκτη πλούτου για το επίπεδο widowed θα είναι ίση με 0,11+0,03=0,14 όταν η κεντροποιημένη ηλικία αυξηθεί 1 έτος. Δηλαδή όταν ηλικία αυξηθεί κατά 1 έτος από την μέση τιμή της , ο δείκτης πλούτου για τους χήρους θα αυξηθεί κατά 0,14. (β7+β5)

- Η αναμενόμενη μεταβολή του δείκτη πλούτου για το επίπεδο divorced θα είναι ίση με 0,11-0,18=-0,07 όταν η κεντροποιημένη ηλικία αυξηθεί 1 έτος. Δηλαδή όταν ηλικία αυξηθεί κατά 1 έτος από την μέση τιμή της , ο δείκτης πλούτου για τους χωρισμένους θα μειωθεί κατά 0,07. (β8+β5)

- Η αναμενόμενη μεταβολή του δείκτη πλούτου για το επίπεδο separated θα είναι ίση με 0,11-0,28=-0,14 όταν η κεντροποιημένη ηλικία αυξηθεί 1 έτος. Δηλαδή όταν ηλικία αυξηθεί κατά 1 έτος από την μέση τιμή της , ο δείκτης πλούτου για αυτούς που είναι σε διάσταση θα μειωθεί κατά 0,14. (β9+β5)

- Η αναμενόμενη μεταβολή του δείκτη πλούτου για το επίπεδο never married θα είναι ίση με 0,11+0,23=0,34 όταν η κεντροποιημένη ηλικία αυξηθεί 1 έτος. Δηλαδή όταν ηλικία αυξηθεί κατά 1 έτος από την μέση τιμή της , ο δείκτης πλούτου για αυτούς που δεν είναι παντρεμένοι θα αυξηθεί κατά 0,34. (β10+β5)

- Η αναμενόμενη τιμή του δείκτη πλούτου θα είναι περίπου 52,6 ποσοστιαίες μονάδες για κάποιον που ανήκει στο επίπεδο αναφοράς, δηλαδή για κάποιον που είναι παντρεμένος, όταν η κεντροποιημένη ηλικία κα η κεντροποιημένη ηλικία ολοκλήρωσης σπουδών είναι ίσα με μηδέν. (β0)

- Η διαφορά της αναμενόμενης τιμής του δείκτη πλούτου μεταξύ never married και married όταν η κεντροποιημένη ηλικία είναι ίση με μηδέν , δηλαδή όταν η ηλικία είναι ίση με την μέση τιμή της, θα είναι περίπου 1,7 ποσοστιαίες μονάδες. Άρα για την μέση ηλικία , αυτοί που δεν είναι παντρεμένοι θα έχουν κατά 1,7 ποσοστιαίες μονάδες μικρότερο δείκτη πλούτου από τους παντρεμένους. (β4)

- Η διαφορά της αναμενόμενης τιμής του δείκτη πλούτου μεταξύ separated και married όταν η κεντροποιημένη ηλικία είναι ίση με μηδέν , δηλαδή όταν η ηλικία είναι ίση με την μέση τιμή της, θα είναι περίπου 7,8 ποσοστιαίες μονάδες. Άρα για την μέση ηλικία, αυτοί που είναι σε διάσταση θα έχουν κατά 7,8 ποσοστιαίες μονάδες μικρότερο δείκτη πλούτου από τους παντρεμένους. (β3)

- Η διαφορά της αναμενόμενης τιμής του δείκτη πλούτου μεταξύ divorced και married όταν η κεντροποιημένη ηλικία είναι ίση με μηδέν , δηλαδή όταν η ηλικία είναι ίση με την μέση τιμή της, θα είναι περίπου 4,7 ποσοστιαίες μονάδες. Άρα για την μέση ηλικία, αυτοί που είναι χωρισμένοι θα έχουν κατά 4,7 ποσοστιαίες μονάδες μικρότερο δείκτη πλούτου από τους παντρεμένους. (β2)

- Η διαφορά της αναμενόμενης τιμής του δείκτη πλούτου μεταξύ widowed και married όταν η κεντροποιημένη ηλικία είναι ίση με μηδέν , δηλαδή όταν η ηλικία είναι ίση με την μέση τιμή της, θα είναι περίπου 3,5 ποσοστιαίες μονάδες. Άρα για την μέση ηλικία, αυτοί που είναι χήροι θα έχουν κατά 3,5 ποσοστιαίες μονάδες μικρότερο δείκτη πλούτου από τους παντρεμένους. (β1)

Στο τελικό στάδιο της ανάλυσης, για να μπορέσουμε να καταλήξουμε στο καλύτερο εκτιμητικά μοντέλο για το δείκτη πλούτου, θα συγκρίνουμε τα δύο μοντέλα που έχουμε βρει. Με την χρήση anova για την

σύγκριση των δύο φωλιασμένων μοντέλων αλλά και συγκεκριμένα με την τιμή AIC (6416.436, 6413.563) για το κάθε ένα, παρατηρούμε ότι το καλύτερο μοντέλο είναι το μοντέλο με τις αλληλεπιδράσεις .

Ελέγχοντας τις προϋποθέσεις που πρέπει να ικανοποιεί το μοντέλο παρατηρούμε αρχικά για την κανονικότητα, εφαρμόζοντας έλεγχο Shapiro-Wilk , ότι απορρίπτεται (p-value = 0.002917). Για την ομοσκεδαστικότητα, εφαρμόζοντας κατάλληλα τον έλεγχο Levene-test παρατηρούμε επίσης ότι απορρίπτεται (p-value= 5.443e-09). Η υπόθεση της τυχαιότητας/ασυσχέτισης, από την άλλην, εφαρμόζοντας έλεγχο Durbin-Watson παρατηρούμε ότι δεν απορρίπτεται (p-value= 0.436). Κατασκευάζοντας και τα κατάλληλα διαγράμματα για τους παραπάνω ελέγχους ,με την χρήση της εντολής plot, δεν παρατηρούμε έντονα την παραβίαση των παραπάνω υποθέσεων. Θα θεωρήσουμε ,λοιπόν, λόγω μεγάλου δείγματος, ότι δεν υπάρχει πρόβλημα με τις υποθέσεις του μοντέλου. Ακόμη, μετά από διαγραμματικό έλεγχο Cook’s distance/Leverage βλέπουμε κάποιες ακραίες τιμές οι οποίες όμως δεν αποτελούν σημεία επιρροής άρα δεν μας προβληματίζουν.

3. ΣΥΜΠΕΡΑΣΜΑΤΑ

Η παραπάνω μελέτη είχε ως σκοπό την εκτίμηση ενός υποδείγματος για τους παράγοντες που επηρεάζουν το δείκτη πλούτου αλλά και την σχέση των παραγόντων αυτών μεταξύ τους. Το τελικό μοντέλο δεν έχει και τόσο καλή προσαρμογή (Adjusted R-squared=0.3936) αλλά ερμηνεύει τις σχέσεις μεταξύ των μεταβλητών αρκετά καλά και κυρίως προβλέπει τον αναμενόμενο δείκτη πλούτου. Το συμπέρασμα που βγάζουμε από την παραπάνω ανάλυση είναι ότι ο δείκτης πλούτου επηρεάζεται έντονα από την οικογενειακή κατάσταση και την ηλικία ολοκλήρωσης σπουδών . Επίσης φαίνεται να υπάρχει διαφορά στον αναμενόμενο δείκτη πλούτου ανάλογα την ηλικία των ατόμων για την κάθε οικογενειακή κατάσταση. Παρατηρούμε ότι για την μέση ηλικία, οι παντρεμένοι θα έχουν μεγαλύτερο δείκτη πλούτου από τους υπόλοιπους ανεξαρτήτως της ηλικίας ολοκλήρωσης σπουδών . Ύστερα επίσης από μαθηματική ανάλυση παρατηρούμε ότι αν η ηλικία αυξηθεί κατά 1 έτος από την μέση τιμή της θα ισχύει ότι:

- Οι χήροι θα έχουν μικρότερο δείκτη πλούτου από τους παντρεμένους κατά 3,47 ποσοστιαίες μονάδες. (διαφορά αναμενόμενης τιμής δείκτη πλούτου μεταξύ χήρων και παντρεμένων=0,14-0,11-3,5)

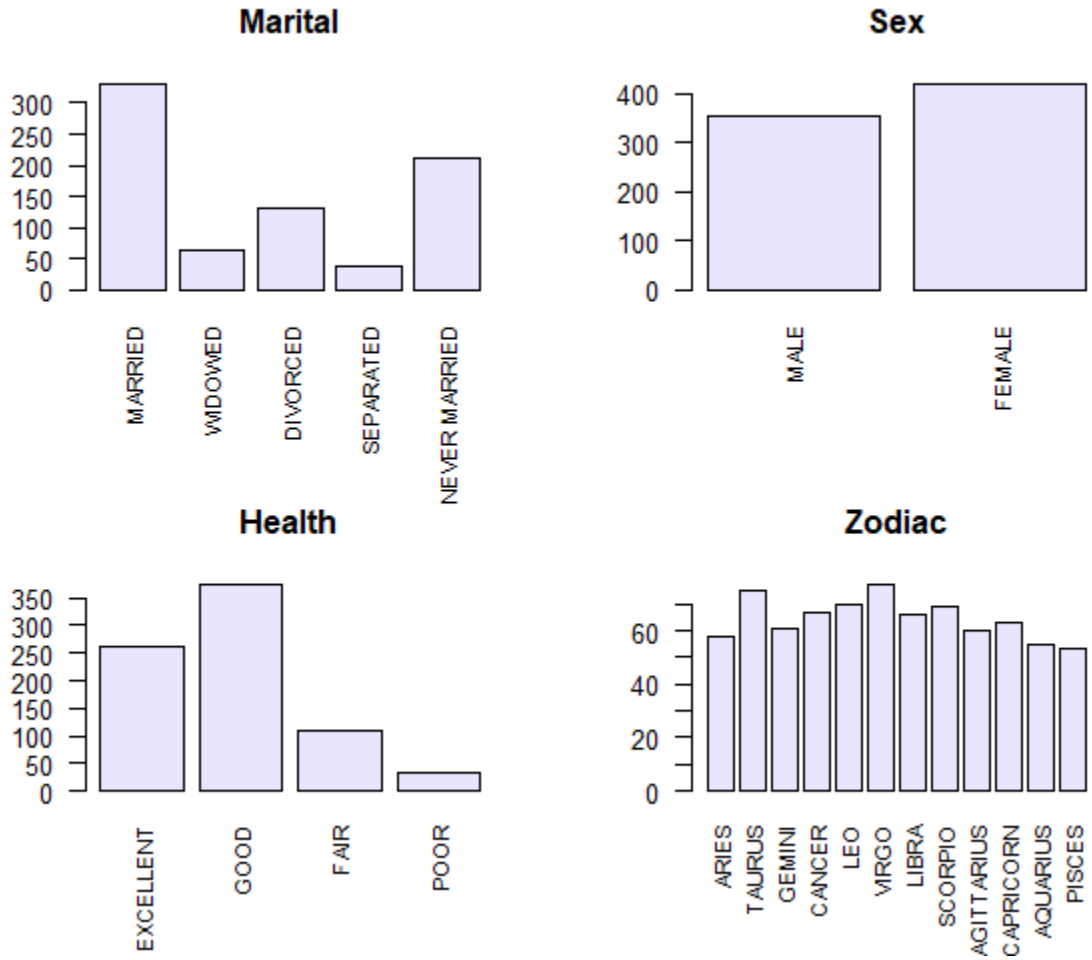
- Οι χωρισμένοι θα έχουν μικρότερο δείκτη πλούτου από τους παντρεμένους κατά 4,88 ποσοστιαίες μονάδες. (διαφορά αναμενόμενης τιμής δείκτη πλούτου μεταξύ χωρισμένων και παντρεμένων=-0,07-0,11-4,7)

- Αυτοί που είναι σε διάσταση θα έχουν μικρότερο δείκτη πλούτου από τους παντρεμένους κατά 8,05 ποσοστιαίες μονάδες. (διαφορά αναμενόμενης τιμής δείκτη πλούτου μεταξύ χωρισμένων και παντρεμένων=-0,14-0,11-7,8)

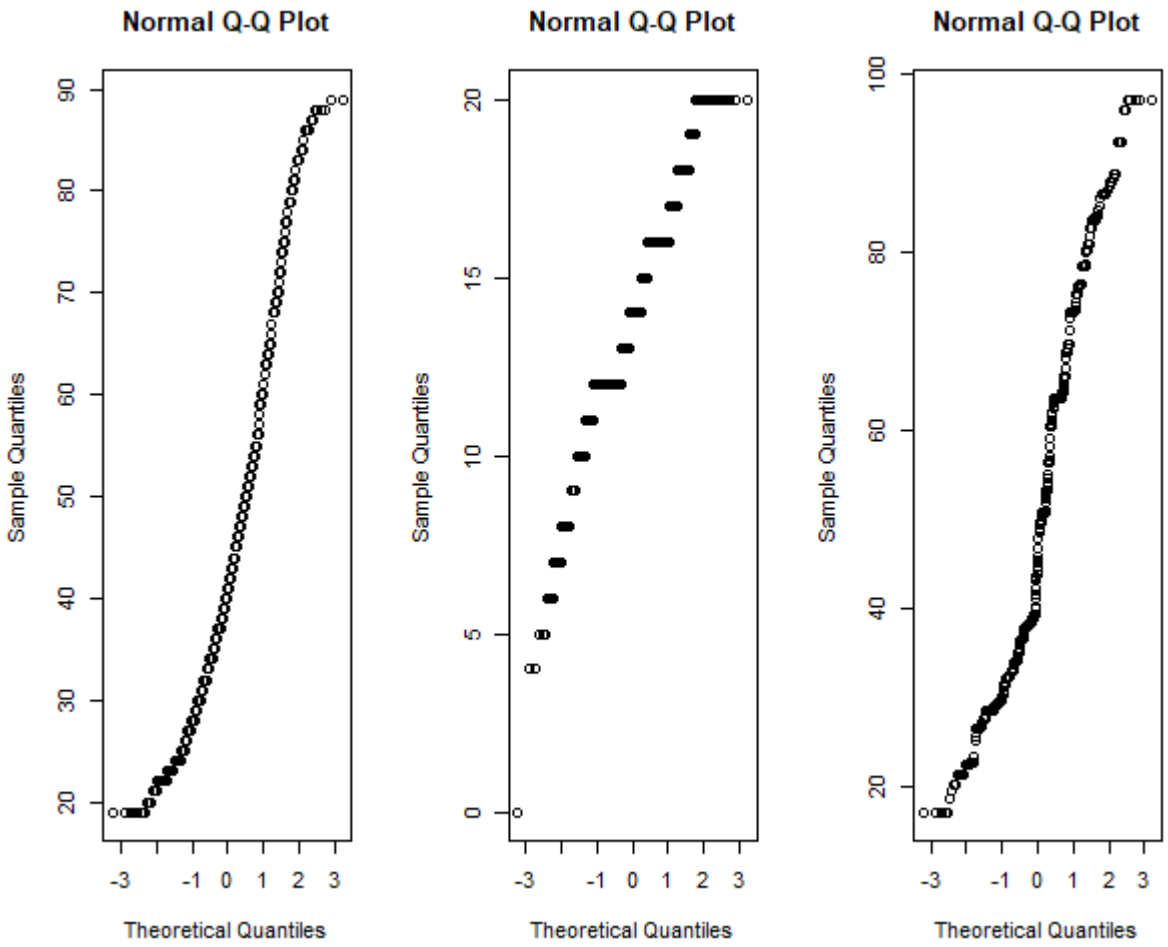
- Αυτοί που δεν είναι παντρεμένοι θα έχουν μικρότερο δείκτη πλούτου από τους παντρεμένους κατά 1,47 ποσοστιαίες μονάδες. (διαφορά αναμενόμενης τιμής δείκτη πλούτου μεταξύ χωρισμένων και παντρεμένων=0,34-0,11-1,7)

Αρά οι παντρεμένοι έχουν μεγαλύτερο δείκτη πλούτου. Η ηλικία βλέπουμε ότι επηρεάζει μόνο το πόσο θα διαφέρουν οι υπόλοιπες ομάδες από τους παντρεμένους με βάση την μέση τιμή της, δηλαδή το 43. Σύμφωνα με τα παραπάνω παρατηρούμε ακόμη ότι η διαφορά μειώθηκε για τους χήρους και τους μη παντρεμένους. Άρα αν η ηλικία αυξηθεί και άλλο ίσως η διαφορά αυτή να μηδενιστή. Και όντως, αν η ηλικία αυξηθεί κατά 8 έτη από την μέση τιμή της, δηλαδή φτάσει στα 51 έτη, βλέπουμε ότι ο δείκτης πλούτου των μη παντρεμένων θα είναι μεγαλύτερος κατά 0,14 ποσοστιαίες μονάδες από τον δείκτη των παντρεμένων. Για τους χήρους, από την άλλην, παρατηρούμε ότι η διαφορά τους θα μηδενιστή όταν η μέση ηλικία αυξηθεί κατά 117 έτη κάτι το οποίο είναι ακατόρθωτο .

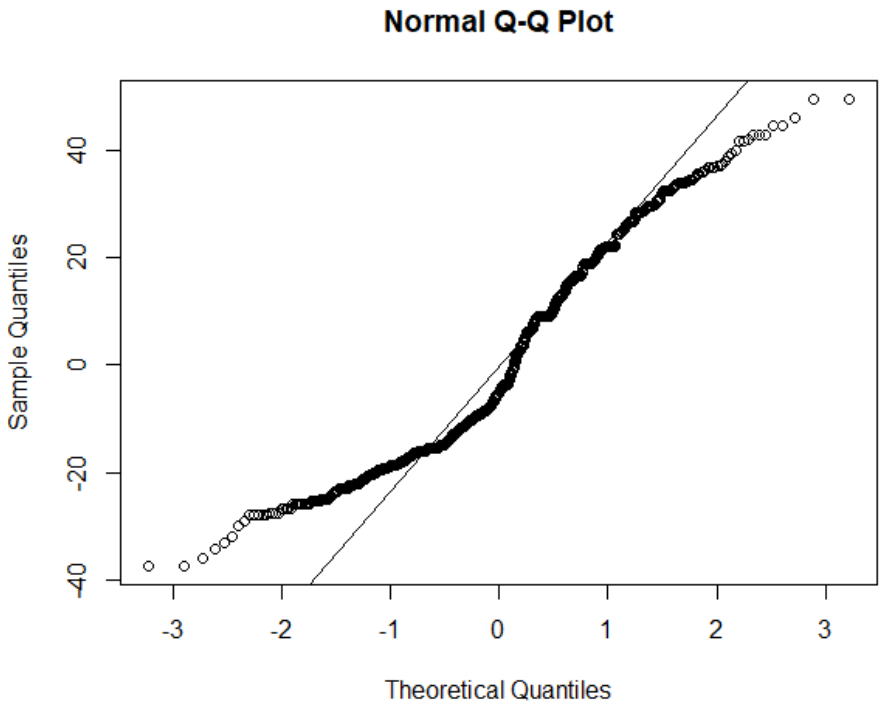
4. ΠΑΡΑΡΤΗΜΑ



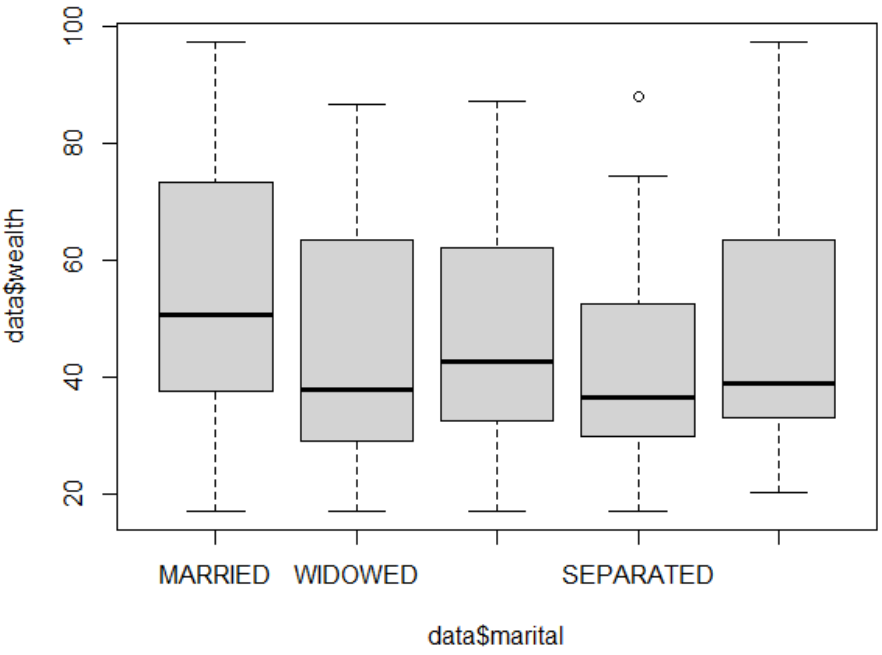
Σχήμα 2: Barplot για τις κατηγορικές μεταβλητές



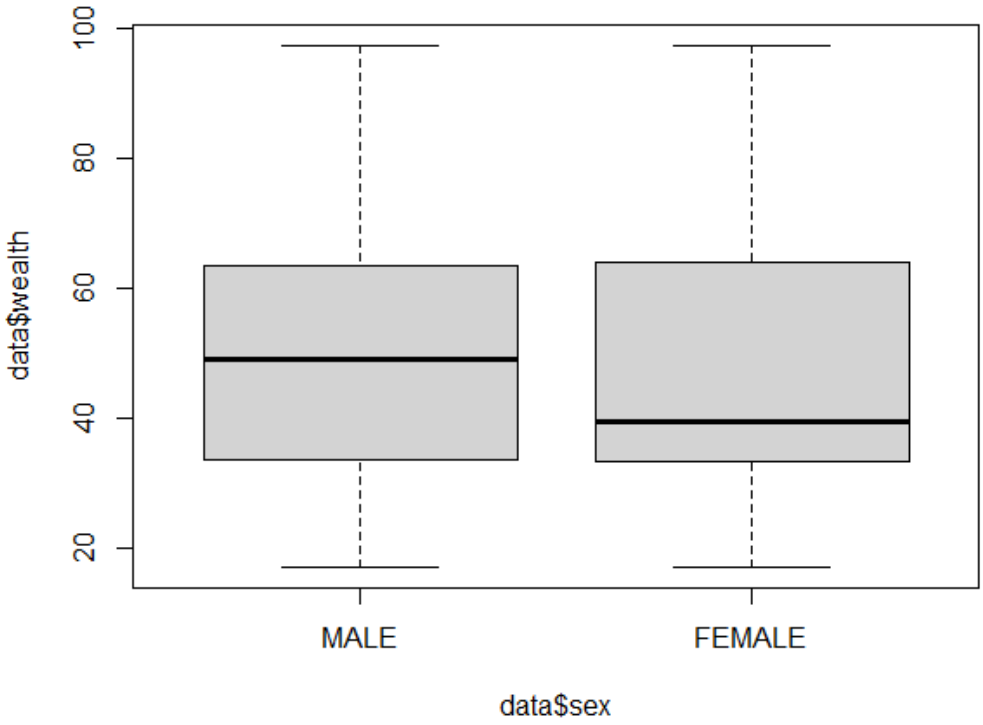
Σχήμα 3: QQplots διαγράμματα για τον έλεγχο της κανονικότητας των ποιοτικών μεταβλητών age, educ και wealth



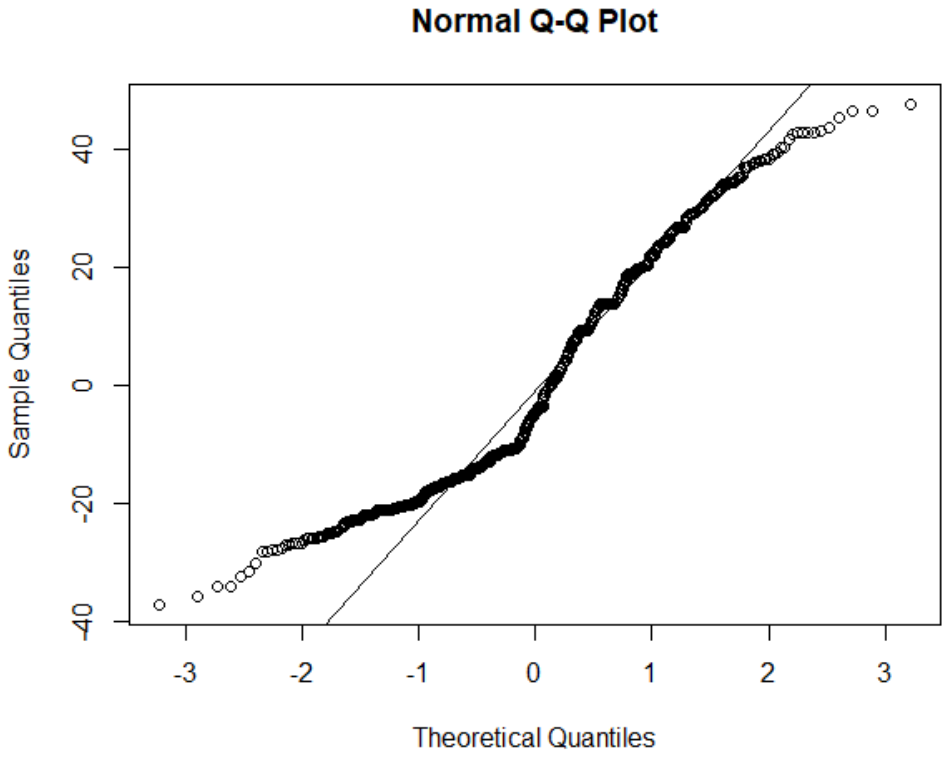
Σχήμα 4: Έλεγχος κανονικότητας των καταλοίπων για την σχέση ανά 2 wealth-marital



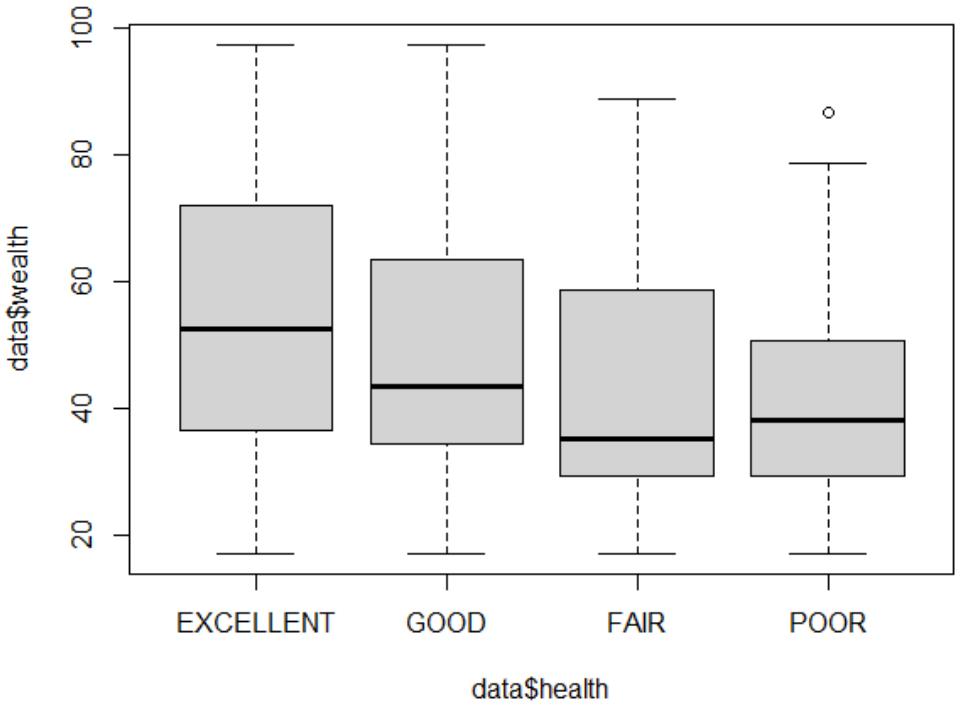
Σχήμα 5: Boxplots για τον έλεγχο διαφοράς διαμέσων για την σχέση ανά 2 wealth-marital



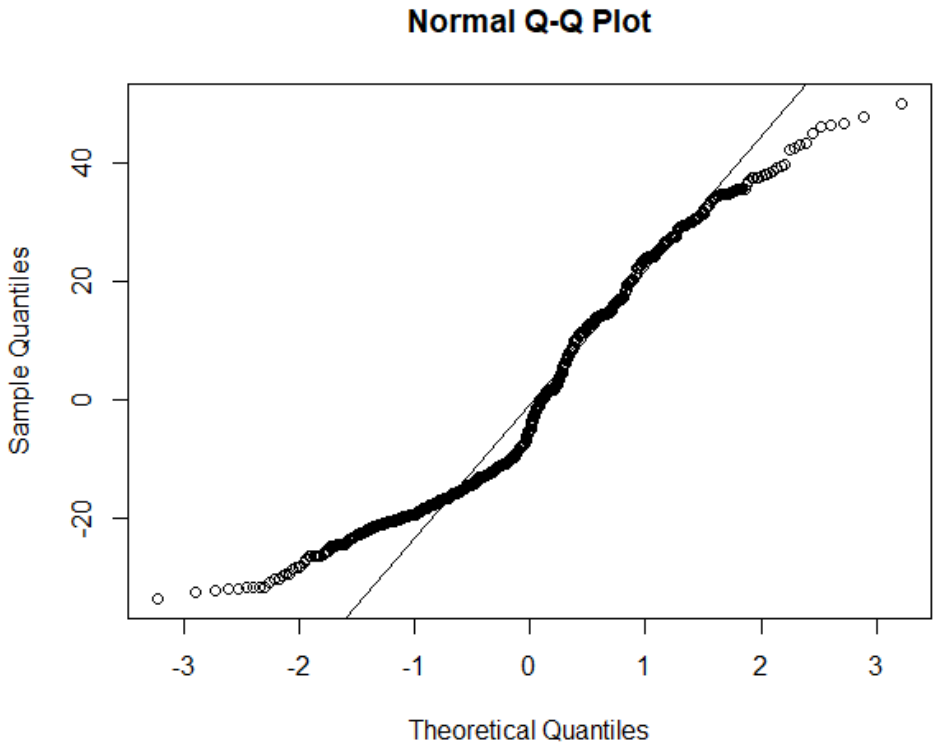
Σχήμα 6: Boxplots για τον έλεγχο διαφοράς διαμέσων για την σχέση ανά 2 wealth-sex



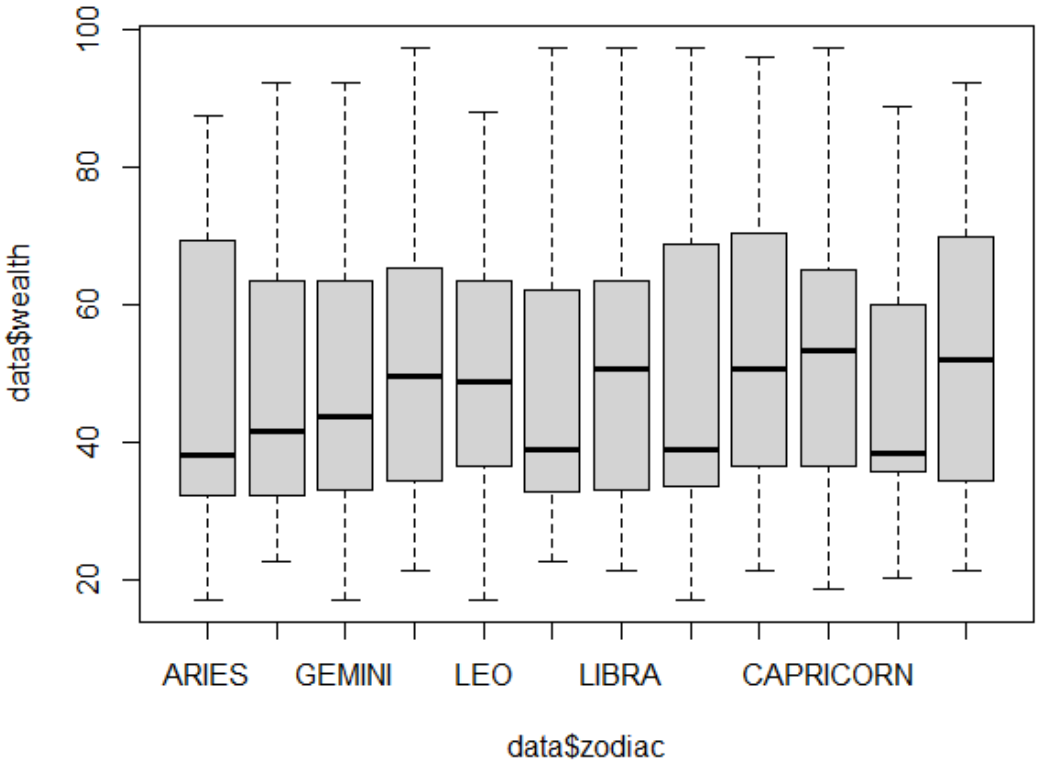
Σχήμα 7: Έλεγχος κανονικότητας των καταλοίπων για την σχέση ανά 2 wealth-health



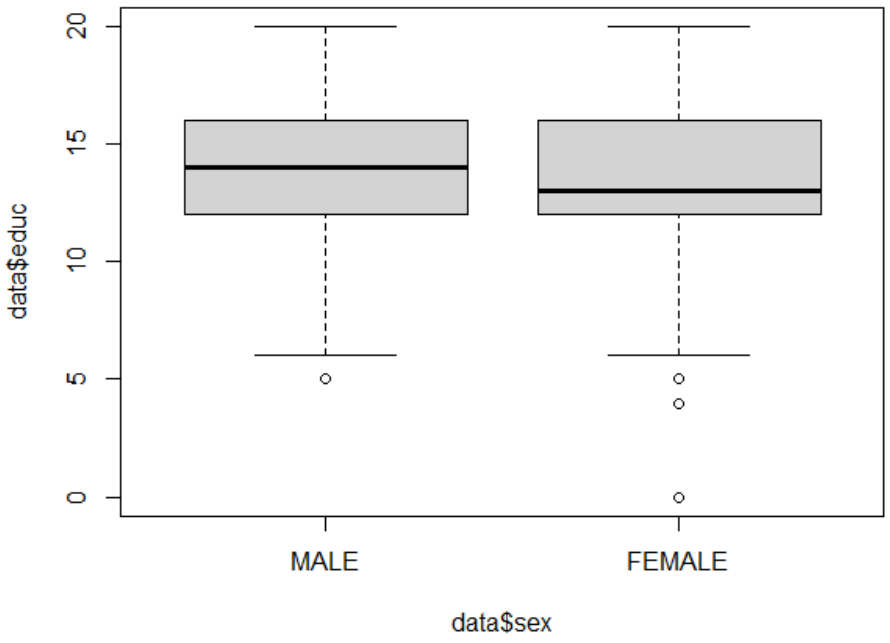
Σχήμα 8: Boxplots για τον έλεγχο διαφοράς διαμέσων για την σχέση ανά 2 wealth-health



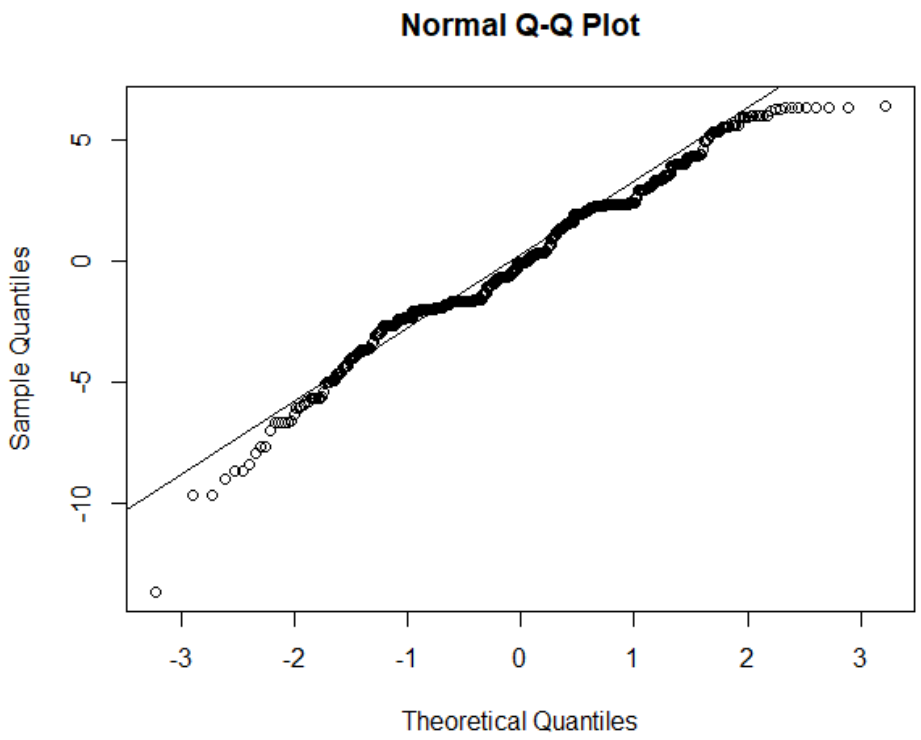
Σχήμα 9: Έλεγχος κανονικότητας των καταλοίπων για την σχέση ανά 2 wealth-zodiac



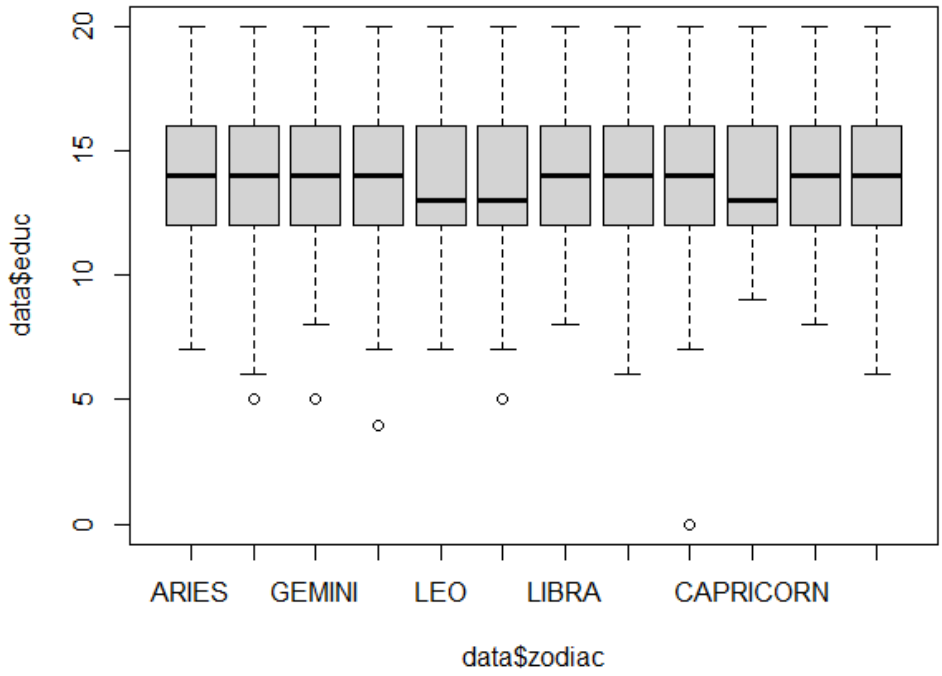
Σχήμα 10: Boxplots για τον έλεγχο διαφοράς διαμέσων για την σχέση ανά 2 wealth-zodiac



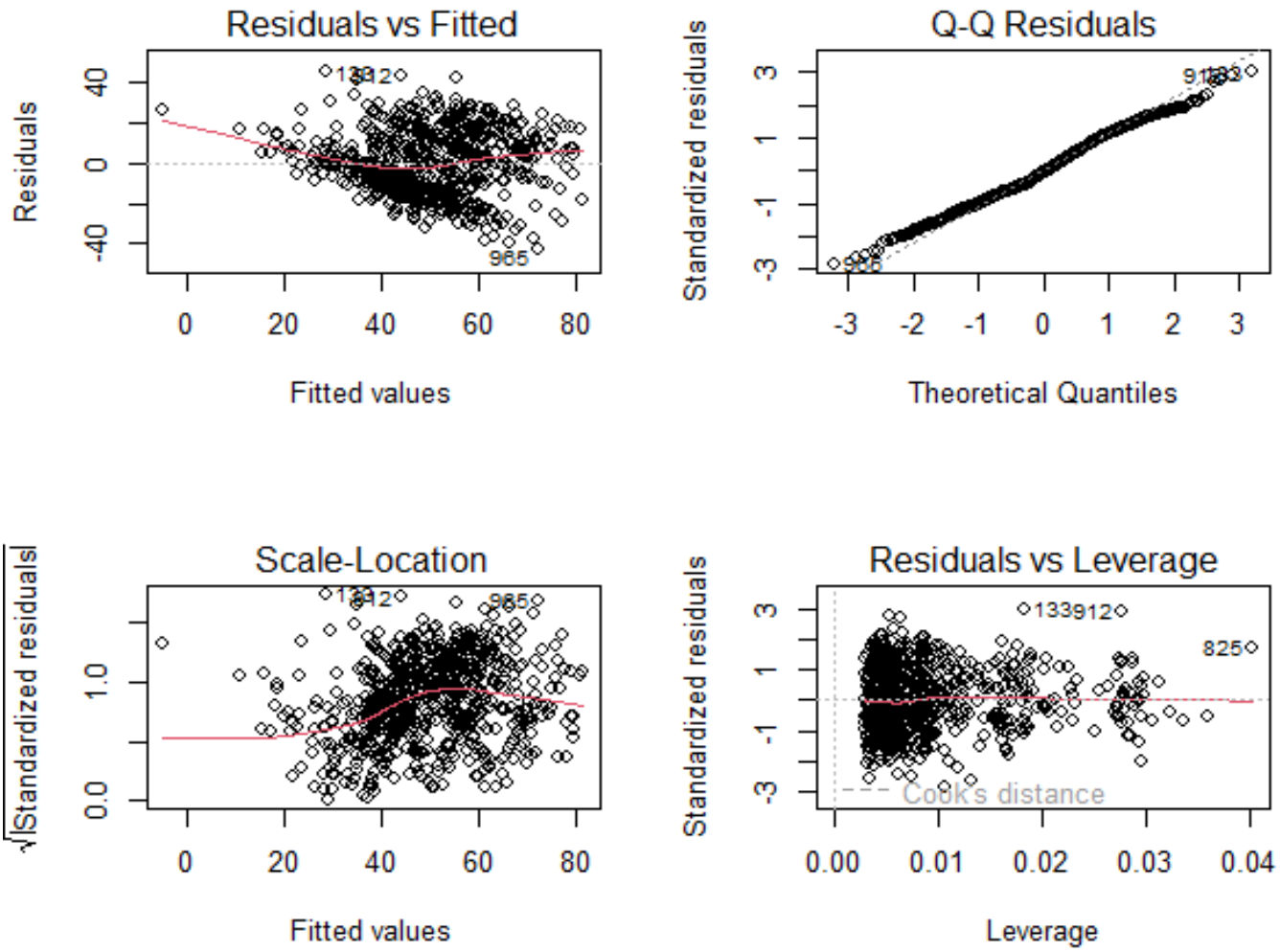
Σχήμα 11: Boxplots για τον έλεγχο διαφοράς διαμέσων για την σχέση ανά 2 educ-sex



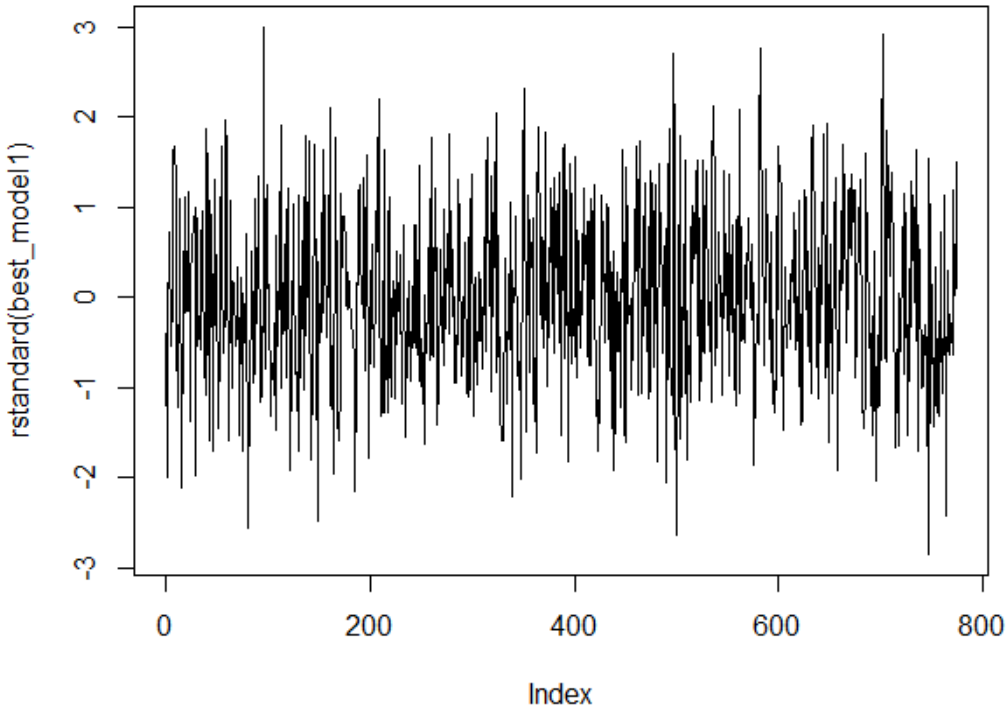
Σχήμα 12: Έλεγχος κανονικότητας των καταλοίπων για την σχέση ανά 2 educ-zodiac



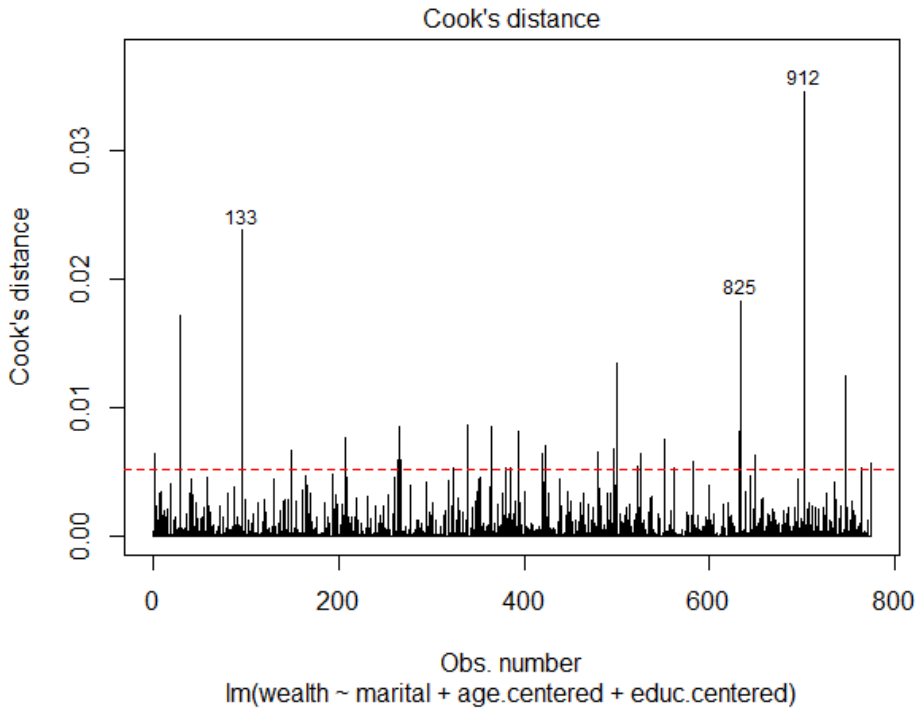
Σχήμα 13: Boxplots για τον έλεγχο διαφοράς διαμέσων για την σχέση ανά 2 educ-zodiac



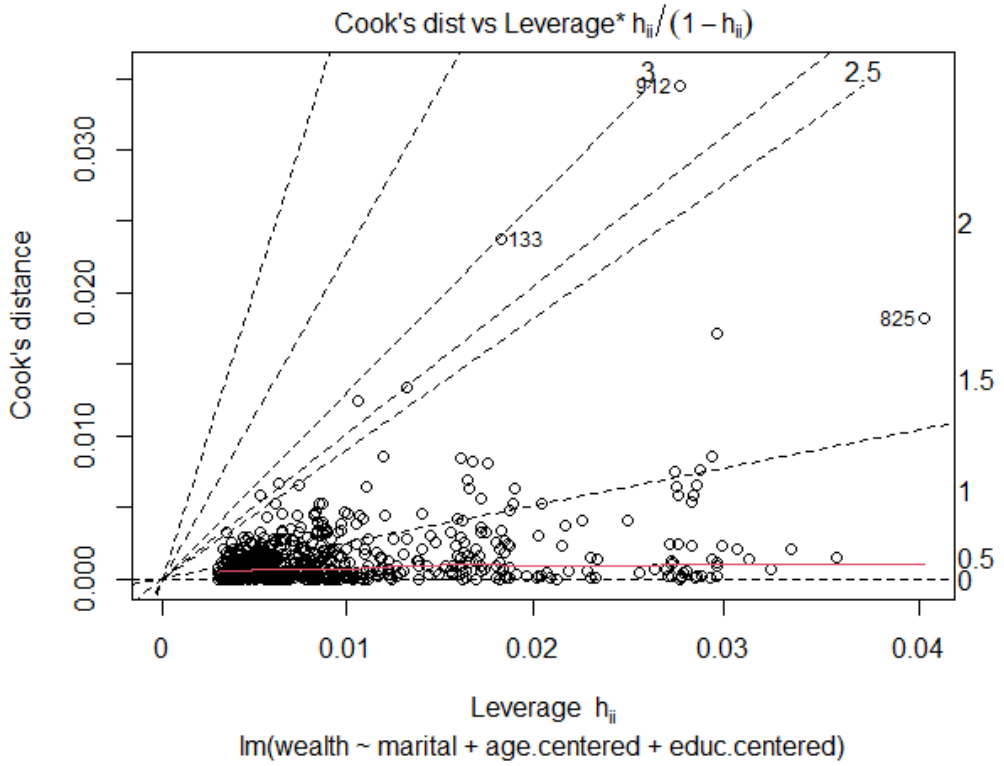
Σχήμα 14: Διαγνωστικοί έλεγχοι για τις υποθέσεις του πολλαπλού γραμμικού μοντέλου για τις στατιστικά σημαντικές μεταβλητές μετά την κεντροποίηση



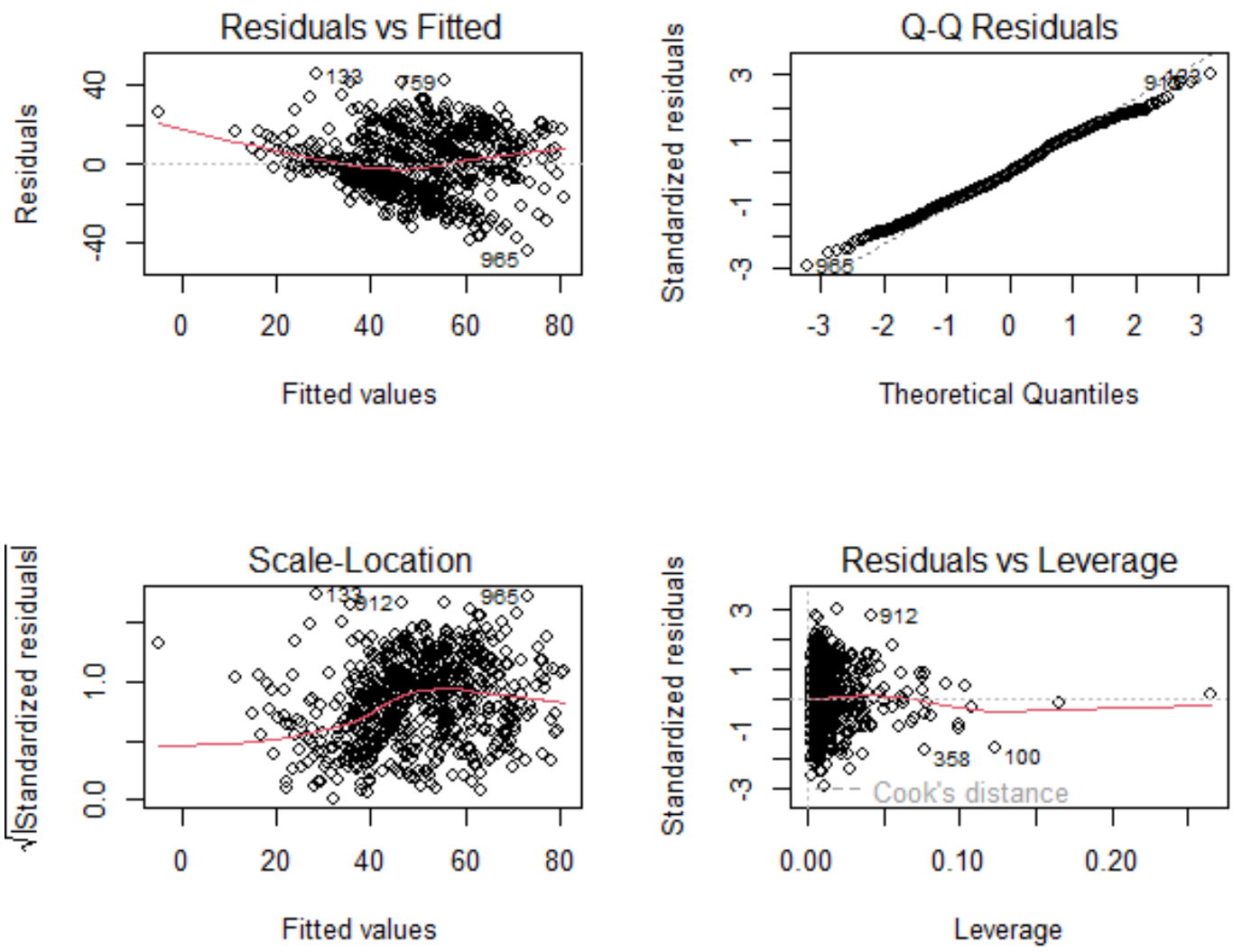
Σχήμα 15: Έλεγχος τυχαιότητας των καταλοίπων του πολλαπλού γραμμικού μοντέλου



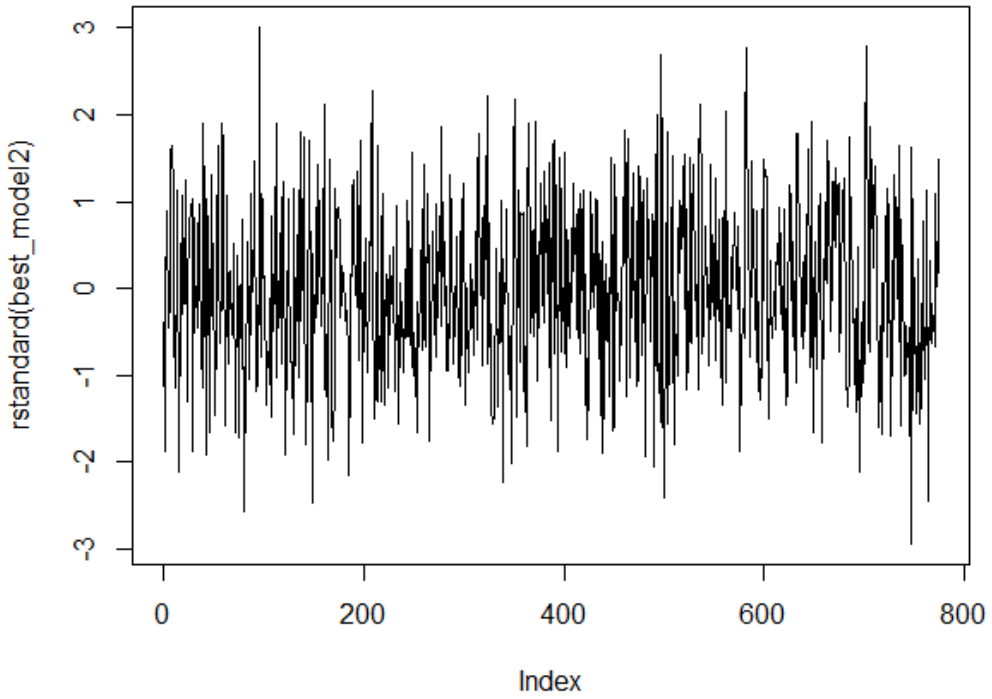
Σχήμα 16: Έλεγχος για ακραίες τιμές και σημεία επιρροής για το πολλαπλό γραμμικό μοντέλο



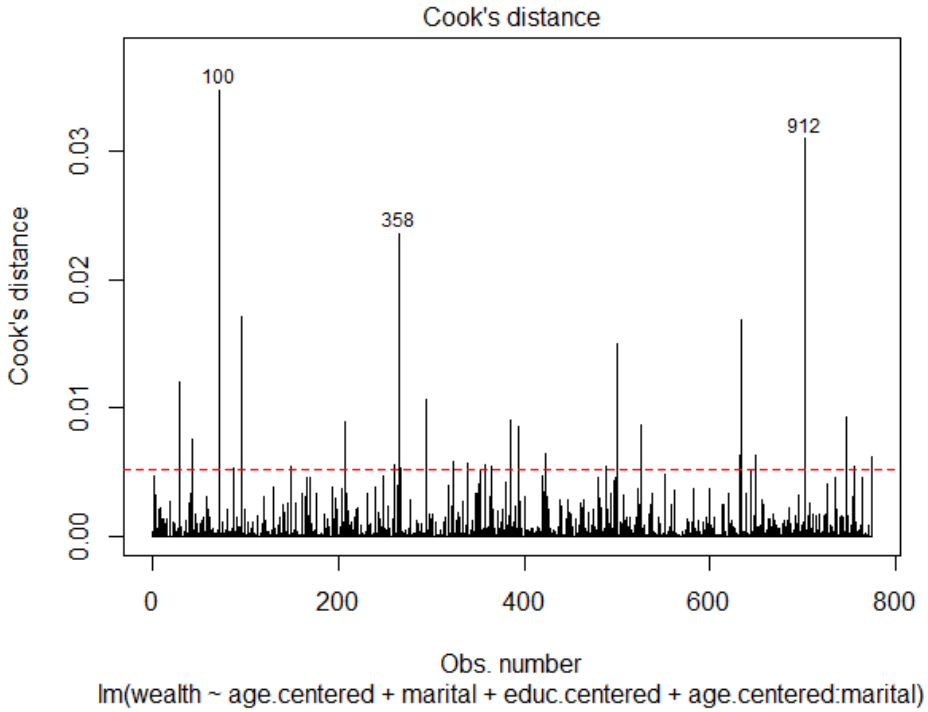
Σχήμα 17: Έλεγχος για ακραίες τιμές και σημεία επιρροής για το πολλαπλό γραμμικό μοντέλο



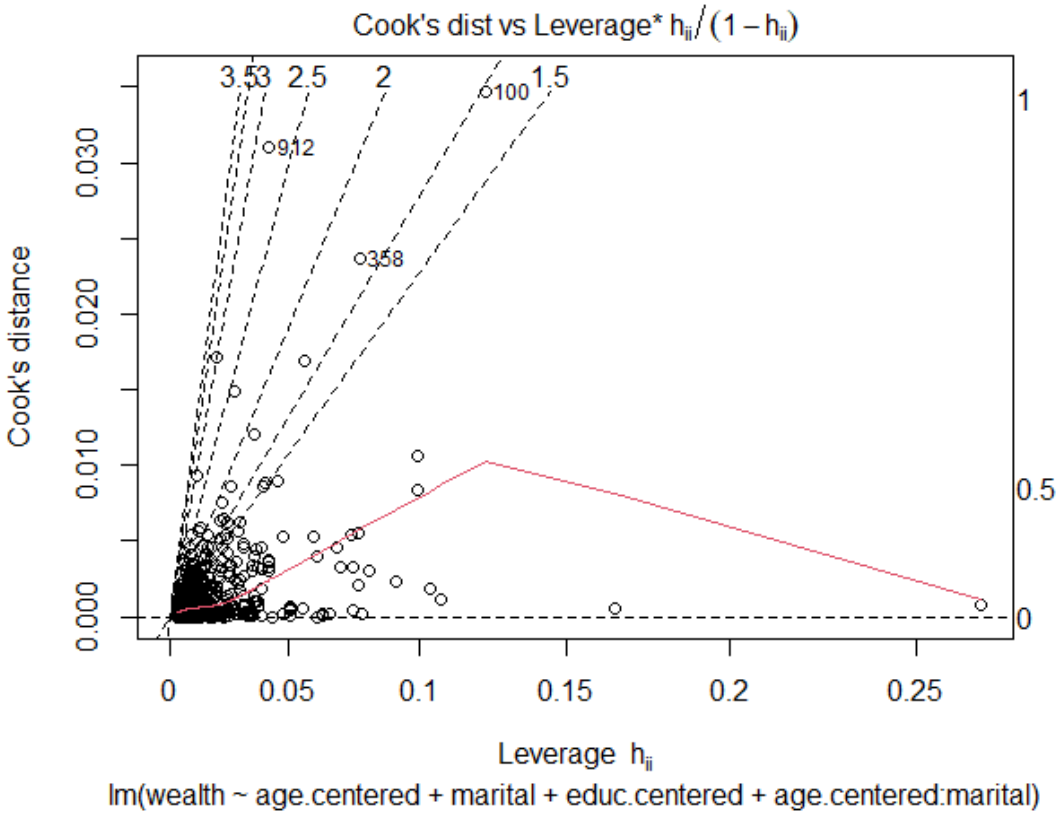
Σχήμα 18: Διαγνωστικοί έλεγχοι για τις υποθέσεις του πολλαπλού γραμμικού μοντέλου με τις αλληλεπιδράσεις



Σχήμα 19: Έλεγχος τυχαιότητας των καταλοίπων του πολλαπλού γραμμικού μοντέλου με τις αλληλεπιδράσεις



Σχήμα 20: Έλεγχος για ακραίες τιμές και σημεία επιρροής για το πολλαπλό γραμμικό μοντέλο με τις αλληλεπιδράσεις



Σχήμα 21: Έλεγχος για ακραίες τιμές και σημεία επιρροής για το πολλαπλό γραμμικό μοντέλο με τις αλληλεπιδράσεις