ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ

ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ

Εργασία **4:** Δεδομένα Νοσοκομειακής Περίθαλψης

ΕΑΡΙΝΟ ΕΞΑΜΗΝΟ *2021-2022*

Μαρίνα Σαμπροβαλάκη

Έτος Σπουδών: 4ο

Α.Μ 3180234

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

Συγγραφή: LaTeX

Contents

1 Εισαγωγή – περιγραφή μελέτης και προβλήματος

Μία άποψη που ενστερνίζεται μεγάλο μέρος του πληθυσμού είναι ότι οι μονάδες υγείας στις αγροτικές περιοχές δεν είναι στο ίδιο επίπεδο με τις αντίστοιχες των αστικών περιοχών μιας και οι ασθενείς είναι λιγότεροι αριθμητικά, συνεπώς και τα έσοδα, με αποτέλεσμα να μην μπορούν να διατηρηθούν σε λειτουργία. Όσο αυτή η άποψη βρίσκει υποστηρικτές, οι περικοπές στις κρατικές δαπάνες και η γενικότερη υποχρηματοδότηση συνεχίζονται ακάθεκτες με αποτέλεσμα νοσοκομεία σε αγροτικές περιοχές να κλείνουν και χιλιάδες άνθρωποι να μην έχουν πρόσβαση στην υγεία. Ο στόχος της παρούσας μελέτης είναι η αξιολόγηση της σχέσης μεταξύ των νοσοκομείων των αστικών κέντρων των αγροτικών περιοχών με κριτήρια τα έσοδα, τα έξοδα και τα κέρδη. Τα αποτελέσματα μίας μελέτης σαν αυτή, θα μπορούσαν να βοηθήσουν στην κατάρριψη της άποψης όσον αναφορά την διαφορά ανάμεσα στο επίπεδο των νοσοκομείων σε διαφορετικές περιοχές και στην επανεκκίνηση της χρηματοδότησης τους. Για τον σκοπό αυτό θα χρησιμοποιήσουμε ένα σετ δεδομένων που περιέχει πληροφορίες σχετικά με τα οικονομικά στοιχεία των νοσοκομείων σε αστικές και αγροτικές περιοχές. Το σετ δεδομένων περιέχει 52 παρατηρήσεις και 7 μεταβλητές και είναι το εξής:

Πίνακας 1: Πίνακας Δεδομένων

Αριθμός Μεταβλητής	Όνομα	Τύπος	Σημασία	Τιμές
1	BED	αριθμητική	αριθμός κρεβατιών	-
2	MCDAYS	αριθμητική	ετήσιος αριθμός ιατρικής περίθαλψης μέσα στο νοσοκομείο (x100)	-
3	TDAYS	αριθμητική	ετήσιος αριθμός ιατρικής περίθαλψης (στο νοσοκομείο ή στο σπίτι) (x100)	-
4	PCREV	αριθμητική	ετήσιο συνολικό εισόδημα περίθαλψης (σε εκατοντάδες $)	-
5	NSAL	αριθμητική	ετήσιο συνολικό ποσό μισθών νοσοκόμων (σε εκατοντάδες $)	-
6	FEXP	αριθμητική	ετήσια έξοδα εγκαταστάσεων (σε εκατοντάδες $)	-
7	RURAL	κατηγορική	τοποθεσία	1:αγροτική 0:μη αγροτική

2 Περιγραφική ανάλυση

Αρχικά θα εισάγουμε τα δεδομένα μας. Σε αυτό το σημείο πρέπει να αναφέρουμε ότι η κατηγορική μεταβλητή RURAL, οι τιμές της οποίας αντιπροσωπεύουν μια απάντηση, στα δεδομένα μας έχει καταγραφεί ως αριθμοί και για αυτό την μετατρέπουμε σε κατάλληλες τιμές. Συνεχίζοντας θέλουμε να εξετάσουμε κάθε μεταβλητή ξεχωριστά και να δούμε τις τιμές που περιέχει. Για να κρατήσουμε πιο απλά τα νούμερα μας, δεν πολλαπλασιάσαμε με το 100 τις τιμές των μεταβλητών
MCDAYS, TDAYS, PCREV, NSAL, FEXP. Για τις αριθμητικές μεταβλητές μπορούμε να δούμε την μέση τιμή, την τυπική απόκλιση, την διάμεσο, την ασυμμετρία, και την κύρτωση (βλ. Πίνακα 2). Όσον αφορά την κατηγορική μεταβλητή, μπορούμε να δούμε τις συχνότητες με τις οποίες έχει εμφανιστεί η κατηγορία της κάθε μεταβλητής. Παρατηρούμε ότι οι αγροτικές περιοχές είναι 34 ενώ οι υπόλοιπες 18 είναι μη αγροτικές.

Αξίζει να αναφέρουμε ότι οι μεταβλητές MCDAYS και TDAYS θα μπορούσαν να συνδυαστούν αφαιρώντας την μία από την άλλη. Με αυτόν τον τρόπο δημιουργείται η μεταβλητή home η οποία περιγράφει τον ετήσιο αριθμό ιατρικής περίθαλψης σε μέρες στο σπίτι. Ακόμα, δημιουργήσαμε την μεταβλητή outgoings η οποία αναπαριστά τα συνολικά έξοδα του νοσοκομείου, δηλαδή είναι το άθροισμα των μισθών των νοσηλευτών και των εξόδων των εγκαταστάσεων. Μία ακόμα μεταβλητή που δημιουργήσαμε είναι η μεταβλητή profit, η οποία προκύπτει από την αφαίρεση των συνολικών εξόδων (outgoings) από το ετήσιο συνολικό εισόδημα (PCREV) και αναπαριστά το κέρδος κάθε νοσοκομειακής μονάδας. Τέλος δημιουργήσαμε μία δίτιμη μεταβλητή ονόματι finance, η οποία παίρνει την τιμή 1 αν τα έσοδα του νοσοκομείου είναι περισσότερα από τα συνολικά έξοδα και την τιμή 0 αλλιώς.

Στην συνέχεια πρέπει να ελέγξουμε αν οι υποθέσεις μας για τις κατανομές ισχύουν. Συνεπώς θα κάνουμε Shapiro-Wilk Test (S), Kolmogorov-Smirnov Test καθώς και τα αντίστοιχα QQ-Plots. Στα Shapiro-Wilk Test και Kolmogorov-Smirnov Test, αν το pvalue είναι μεγαλύτερο από το 0.05, τότε τα δεδομένα ακολουθούν την κανονική κατανομή. Όσο για τα QQ-Plots, τα δεδομένα ακολουθούν την κανονική κατανομή όταν τα σημεία τους είναι πάνω σε μια ευθεία διαγώνια γραμμή. Στα δεδομένα μας, δεν υπάρχει καμία μεταβλητή που να έχει pvalue>0.05 στα tests, συνεπώς καμία αριθμητική μεταβλητή δεν ακολουθεί την κανονική κατανομή.

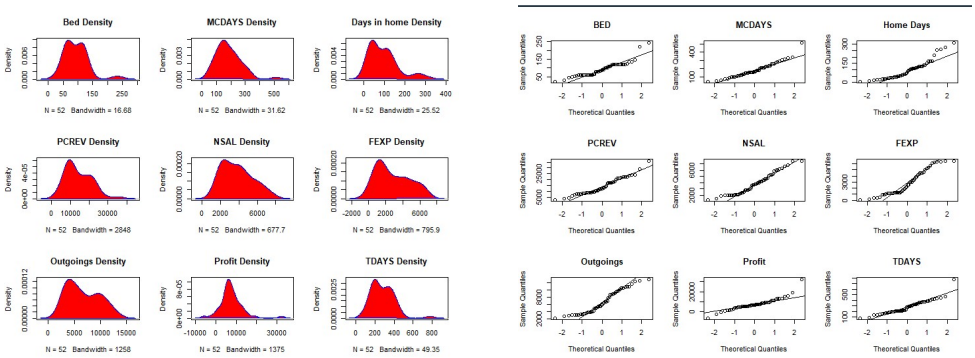


Figure 1: Διαγράμματα πυκνότητας πιθανότητας και QQ-Plots της κάθε μεταβλητής ξεχωριστά όπου βλέπουμε ότι καμια μεταβλητή δεν ακολουθεί την κανονική κατανομή

Πίνακας 2: Πίνακας Περιγραφικών Μέτρων Αριθμητικών Μεταβλητών								
Μεταβλητή	NaN	Μέση Τιμή	Τυπική απόκλιση	Διάμεσος	Min	Max	Ασυμμετρία	Κύρτωση
BED	0	93.27	40.85	88	25	244	1.37	6.23
MCDAYS	0	183.9	87.03	164.5	48	514	1.12	5.32
TDAYS	0	280.2	120.85	279	83	776	1.28	6.55
PCREV	0	14210	6974	12384	2853	36029	0.74	3.28
NSAL	0	3813	1659	3696	1288	7489	0.515	2.285
FEXP	0	2848	1949	2378	137	6442	0.47	1.91
home	0	96.31	71.46	83.50	9	314	1.205	4.11
outgoings	0	6660	3082	6017	2333	12936	0.41	1.87
profit	0	7550	5718	6952	-5995	32198	1.40	8.51

Παρατηρούμε ότι στα διαγράμματα φαίνεται ότι υπάρχουν πολλές μικρές τιμές συγκεντρωμένες (λεπτόκυρτη καμπύλη) και λίγες μεγάλες (δεξιά συμμετρία).
Στη συνέχεια θα δούμε τα διαγράμματα πλαισίου και απολήξεων (boxplots) για την κάθε ποσοτική μεταβλητή προκειμένου να εντοπίσουμε τυχόν ακραίες τιμές (βλ. Παράρτημα). Ωστόσο, η ανάλυση της κάθε μεταβλητής ξεχωριστά μάς δίνει περιορισμένη πληροφορία σε σύγκριση με την πληροφορία που εξάγουμε από την μεταξύ τους σχέση.

3 Σχέσεις μεταβλητών ανά δύο

Η διερεύνηση της κάθε μεταβλητής ξεχωριστά δεν μάς βοηθάει να διερευνήσουμε τις σχέσεις μεταξύ των μεταβλητών και για αυτό θέλουμε να τις εξετάσουμε πιο διεξοδικά. Αρχικά κάνοντας τον πίνακα συσχετίσεων του Pearson παρατηρούμε ότι υπάρχουν κάποιες γραμμικές σχέσεις που είναι ιδιαίτερα αξιοσημείωτες.

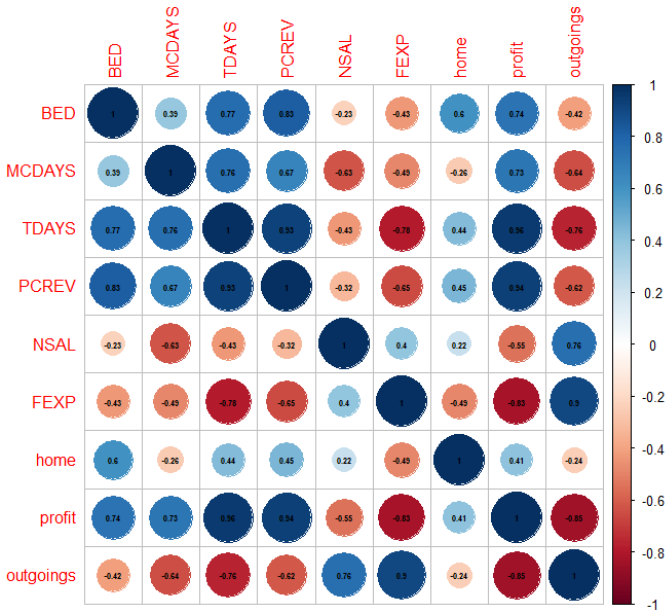


Figure 2: Pearson Matrix στον οποίο φαίνονται οι γραμμικές σχέσεις με πιο έντονο χρωματισμό. Με μπλε φαίνεται η συσχέτιση και με κόκκινο η ανεξαρτησία.

Παρακάτω φαίνονται οι σχέσεις που πιστεύουμε ότι έχει νόημα να μελετήσουμε και να φτιάξουμε boxplots ανά κατηγορία των κατηγορικών μεταβλητών.

- Οικονομική κατάσταση του νοσοκομείου ανά περιοχή (finance $\sim$ RURAL)
- Ετήσια έσοδα ανά περιοχή (PCREV $\sim$ RURAL)
- Αριθμός κρεβατιών ανά περιοχή (BED $\sim$ RURAL)
- Ετήσια έσοδα ανά περιοχή (outgoings $\sim$ RURAL)
- Ετήσιο κέρδος ανά περιοχή (profit $\sim$ RURAL)
- Μέρες στο σπίτι ανά περιοχή (home $\sim$ RURAL)

- Μέρες στο νοσκομείο ανά περιοχή (RURAL $\sim$ MCDAYS)
- Έξοδα εγκαταστάσεων ανά περιοχή (RURAL $\sim$ FEXP)
- Συνολικό ποσό μισθών νοσοκόμων ανά περιοχή (RURAL $\sim$ NSAL)

Για τη μελέτη των παραπάνω σχέσεων θα χρησιμοποιήσουμε ραβδογράμματα ανά κατηγορίες της επεξηγηματικής μεταβλητής, Chi-Square tests για την ανεξαρτησία καθώς και Kruskal tests.Αξίζει να σημειώσουμε ότι δεν χρησιμοποιήσαμε ANOVA tests καθώς απαιτεί αρχικά κανονικότητα στα δεδομένα, κάτι που εμείς δεν έχουμε σε απόλυτο βαθμό.

Έχοντας κάνει αυτούς τους ελέγχους και έχοντας ερμηνεύσει και τα αντίστοιχα γραφήματα, καταλήγουμε στο ότι οι πιο σημαντικές σχέσεις είναι οι εξής:

- Ετήσια έσοδα ανά περιοχή (PCREV $\sim$ RURAL)
- Ετήσια έσοδα ανά περιοχή (outgoings $\sim$ RURAL)
- Μέρες στο σπίτι ανά περιοχή (home $\sim$ RURAL)
- Συνολικό ποσό μισθών νοσοκόμων ανά περιοχή (RURAL $\sim$ NSAL)
- Συνολικό κέρδος νοσοκόμων ανά περιοχή (RURAL $\sim$ profit)
- Συνολική οικονομική κατάσταση του νοσοκομείου ανά περιοχή (RURAL $\sim$ profit)

Ελέγχοντας τις παραπάνω σχέσεις, παρατηρούμε κάποιες διαφορές στα χαρακτηριστικά των νοσοκομειακών μονάδων ανά περιοχή.Αρχικά, σύμφωνα με τα αποτελέσματά μας είναι φανερό ότι τα έσοδα του νοσοκομείου είναι περισσότερα σε μη αγροτικές περιοχές, το ίδιο όμως ισχύει και για τα συνολικά έξοδα. Οι μισθοί άλλωστε είναι περισσότερο υψηλοί σε νοσοκομειακές μονάδες σε μη αγροτικές περιοχές, ενώ τα έξοδα για τις εγκαταστάσεις δεν διαφέρουν σημαντικά. Ακόμα, όσον αφορά τις μέρες νοσηλείας στο νοσοκομείο,δεν διαφέρουν σε μεγάλο βαθμό συγκριτικά με το αν το νοσοκομείο βρίσκεται σε αστική περιοχή ή όχι. Ωστόσο αυτό δεν ισχύει για τις μέρες νοσηλείας στο σπίτι, οι οποίες είναι περισσότερες στις μη αγροτικές περιοχές. Όσον αφορά τον αριθμό των κρεβατιών, επίσης υπάρχουν σημαντικές διαφορές ανά περιοχή. Συμπεράινουμε πως η οικονομική κατάσταση, δηλαδή η σχέση εσόδων και εξόδων, ενός νοσοκομείου φαίνεται να σχετίζεται με την τοποθεσία του. Ωστόσο από το Chi-Squared test φάνηκε ότι η ικανότητα ενός νοσοκομείου να υπερκαλύπτει με τα έσοδα τα έξοδα του, δεν φαίνεται να σχετίζεται με την τοποθεσία του. Σε αυτό το σημείο, θα προχωρήσουμε στην δημιουργία μερικών γραμμικών μοντέλων για να επαληθεύσουμε ή όχι αυτή την άποψη.

4 Προβλεπτικά ή ερμηνευτικά μοντέλα

Αφού αναλύσαμε τις ανα δύο σχέσεις μεταξύ των μεταβλητών και καταλήξαμε ποιες είναι οι πιο σημαντικές, θέλουμε να φτιάξουμε ένα προβλεπτικό μοντέλο για να ερμηνεύσουμε την συμπεριφορά του οικονομικού κόστους των νοσοκομείων σε σχέση με τα έσοδα και τα έξοδα του νοσοκομείου ανά περιοχή. Αρχικά πρέπει να κάνουμε κάποιους ελέγχους υποθέσεων για να διερευνήσουμε περαιτέρω τις σχέσεις των εσόδων και των εξόδων και του κόστους με την τοποθεσία του νοσοκομείου. Για να δούμε ποια είναι η σχέση μεταξύ των εσόδων και την τοποθεσίας, πρέπει να ελέγξουμε την κανονικότητα των εσόδων για κάθε κατηγορία της περιοχής. Παρατηρούμε ότι τα έσοδα στις αγροτικές περιοχές δεν ακολουθούν την κανονική κατανομή (S-pvalue = 0.00156), ενώ για τις μη αγροτικές, την ακολουθούν (S-p-value=0.4943). Δεδομένου ότι απορρίφθηκε η κανονικότητα και ότι τα δείγματα για τις δύο ομάδες δεν είναι αρκετά μεγάλα, συμπεραίνουμε ότι ο μέσος όρος δεν είναι κατάλληλο μέτρο περιγραφής της κεντρικής θέσης για τις δύο ομάδες. Για αυτό το λόγο πραγματοποιούμε τον μη παραμετρικό έλεγχο Wilcoxon-test για να ελέγξουμε την ισότητα των διαμέσων. Βλέπουμε ότι είναι στατιστικά σημαντική η διαφορά μεταξύ των διαμέσων(pvalue = 0.0253). Η τοποθεσία, δηλαδή, της νοσοκομειακής μονάδας επηρεάζει στατιστικά σημαντικά τα έσοδα της,όπως φαίνεται και στο παρακάτω boxplot. Στο ίδιο αποτέλεσμα είχαμε φτάσει και στην προηγούμενη ενότητα με το Kruskal Test.

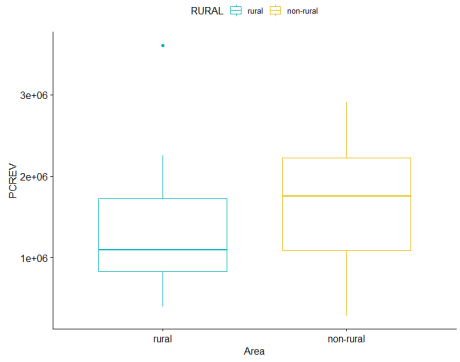


Figure 3: Boxplot των εσόδων σε κάθε περιοχή.

Στην συνέχεια θέλουμε να διερευνήσουμε την σχέση των εξόδων (outgoings) με την τοποθεσία της νοσοκομειακής μονάδας. Αφού ελέγξαμε την κανονικότητα των εξόδων σε κάθε περιοχή, συμπεραίνουμε ότι για τις αγροτικές περιοχές, δεν ακολουθείται η κανονική κατανομη (S-pvalue=0.02138), ενώ για τις μη αγροτικές περιοχές

ακολουθείται (s-pvalue=0.06799). Το Wilcoxon-test επέστρεψε pvalue=0.02279. Καθώς το wilcoxon-test επέστρεψε τόσο, βλέπουμε πως η διαφορά ανάμεσα στις διαμέσους είναι στατιστικά σημαντική. Η τοποθεσία, δηλαδή, της νοσοκομειακής μονάδας επηρεάζει στατιστικά σημαντικά τα έξοδα της, όπως φαίνεται και στο παρακάτω boxplot. Επαληθεύτηκε συνεπώς το αποτέλεσμα του Kruskal test.

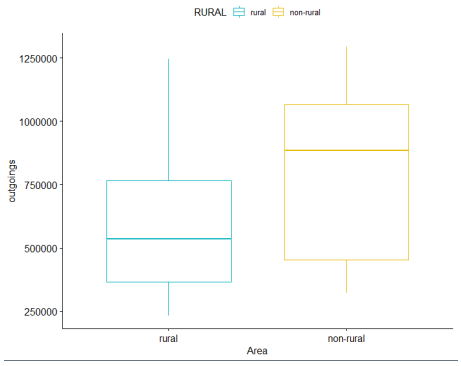


Figure 4: Boxplot των εξόδων σε κάθε περιοχή.

Τέλος, ακολουθήσαμε την ίδια διαδικασία για να διευρευνήσουμε την σχέση του κέρδους (profit) με την τοποθεσία της νοσοκομειακής μονάδας και παρατηρήσαμε ότι το κέρδος ακολουθεί την κανονική κατανομή για τις μη αγροτικές περιοχές (S-pvalue=0.1123), ενώ για τις αγροτικές περιοχές όχι (S-pvalue=2.636e-05). Επί προσθέτως, το Wilcoxon-test μάς επέστρεψε pvalue=0.02279, συνεπώς η διαφορά μεταξύ των διαμέσων είναι στατιστικά σημαντική. Άρα, το αν η νοσοκομειακή μονάδα είναι ή όχι σε αγροτική περιοχή επηρεάζει το κέρδος της. Αυτό φαίνεται και στο παρακάτω boxplot. Στο ίδιο αποτέλεσμα είχαμε φτάσει και στην προηγούμενη ενότητα με το Kruskal Test.

Αφού ελέγξαμε τα δεδομένα μας και κατανοήσαμε τις σχέσεις μεταξύ των μεταβλητών, μπορούμε να προχωρήσουμε στην προσαρμογή ενός μοντέλου για να ερμηνεύσουμε την επιρροή των μεταβλητών στα έσοδα, τα έξοδα και το κέρδος των νοσοκομειακών μονάδων και στη συνέχεια μπορούμε να δημιουργήσουμε ένα μοντέλο για την πρόβλεψή τους. Αρχικά παίρνουμε το πλήρες μοντέλο για τα έσοδα του νοσοκομείου όπου θέλουμε να είναι της μορφής:

PCREV=β0 + β1 * BED + β2 * MCDAYS + β3 * home + β4 * outgoings + β5 * RURAL + β6 * finance

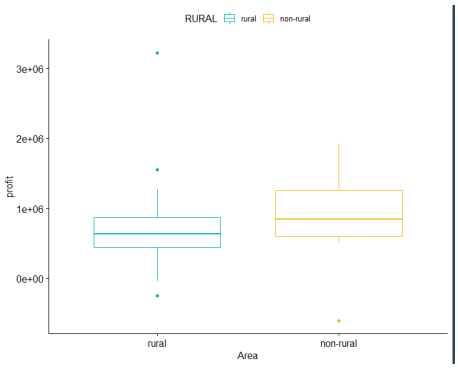


Figure 5: Boxplot του κέρδους σε κάθε περιοχή.

Όπως μπορείτε να δείτε, δεν συμπεριλάβαμε την μεταβλητή profit στο μοντέλο μας καθώς είχε πολύ υψηλό correlation με την μεταβλητή BED και οδηγούσε το μοντέλο σε overfit.

Μπορούμε να παρατηρήσουμε ότι τα residuals του παραπάνω μοντέλου δεν ακολουθούν την κανονική κατανομή (S-pvalue=0.002849), είναι ομοσκεδαστικά (ncvTest pvalue=0.90141) αλλά δεν είναι ανεξάρτητα (DW pvalue=0.038). Αφού δεν έχουμε χρονολογικά δεδομένα αυτό το πρόβλημα μπορεί να προέρχεται είτε από κάποιο λάθος κατά τη διάρκεια της συλλογής των δεδομένων, είτε να λείπει μια μεταβλητή που να μην έχουμε συμπεριλάβει στο μοντέλο είτε τα δεδομένα να έχουν ταξινομηθεί με βάση μία μεταβλητή. Στην συγκεκριμένη περίπτωση τα δεδομένα μας έχουν ταξινομηθεί με βάση τα έσοδα,σε αύξουσα σειρά,το οποίο δημιουργεί πρόβλημα στους ελέγχους αυτοσυσχέτισης(DarwinWatson test). Προκειμένου να λύσουμε το πρόβλημα της αυτοσυσχέτισης των καταλοίπων θα αναταξινομήσουμε τα δεδομένα μας με τυχαία σειρά. Στην συνέχεια, βλέπουμε ότι η σταθερά παρότι είναι στατιστικά σημαντική, δε βγάζει νόημα. Μιας και τα residuals του μοντέλου μας δεν ακολουθούν την κανονική κατανομη, αποφασίσαμε να φτιάξουμε ένα BoxCox. Αν πάμε να την ερμηνεύσουμε, καταλήγουμε στο συμπέρασμα ότι για μηδέν κρεβάτια και μέρες νοσηλείας, για μηδέν έξοδα και για τοποθεσία σε αγροτική περιοχή, η αναμενόμενη τιμή των εσόδων θα μειωθεί. Για αυτόν τον λόγο, θα κεντροποιήσουμε τις αριθμητικές μας μεταβλητές, δηλαδή θα αφαιρέσουμε από αυτές τους μέσους τους για να μας βοηθήσει στην ερμηνεία.Το μόνο που αλλάζει είναι οι συντελεστές στο μοντέλο μας και η ερμηνεία των μεταβλητών. Το μοντέλο φαίνεται παρακάτω.

Βλέπουμε ότι η σταθερά του μοντέλου μας, δηλαδή η παράμετρος β0, έχει τιμή 80.67.Αυτό σημαίνει ότι τα αναμενόμενα έσοδα ενός νοσοκομείου που είναι

	PCREV		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	80.67	49.59 – 111.74	**<0.001**
BED	0.24	-0.07 – 0.56	0.123
MCDAYS	35.54	25.55 – 45.54	**<0.001**
home	39.15	24.90 – 53.40	**<0.001**
outgoings	0.54	0.27 – 0.81	**<0.001**
RURAL [rural]	3.66	-11.82 – 19.13	0.636
finance [1]	62.66	32.80 – 92.52	**<0.001**
Observations	52		
R^2 / R^2 adjusted	0.905 / 0.892		

Figure 6: Μοντέλο πολλαπλής παλινδρόμησης εσόδων με όλες τις μεταβλητές.

σε αγροτική περιοχή, με 0 κρεβάτια και 0 έξοδα, είναι 80.67.

Έχοντας παρατηρήσει τις παραμέτρους του μοντέλου μας συμπεραίνουμε ότι οι περισσότερες μεταβλητές έχουν ισχυρή σχέση με την εξαρτημένη μας μεταβλητή παρ' όλα αυτά υπάρχουν μεταβλητές που δεν προσφέρουν πολλά στο μοντέλο. Για να λύσουμε αυτό το πρόβλημα θα χρησιμοποιήσουμε μεθόδους επιλογής μοντέλων οι οποίες προτείνουν μοντέλα με βάση διαφορετικά κριτήρια. Οι μεθόδοι που θα χρησιμοποιηθούν είναι το subset selection και το stepwise regression.Όλες οι μέθοδοι προτείνουν ένα μοντέλο με 5 μεταβλητές,έχοντας αφαιρέσει την μεταβλητή RURAL,καθώς δεν προσφέρει αρκετή πληροφορία στο μοντέλο. Οπότε μετά τη χρήση μεθόδων επιλογής μεταβλητών καταλήγουμε στο παρακάτω μοντέλο.

	PCREV		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	114.87	86.39 – 143.35	**<0.001**
BED	0.60	0.15 – 1.05	**0.011**
MCDAYS	33.28	20.95 – 45.61	**<0.001**
home	21.72	4.92 – 38.51	**0.013**
outgoings	0.44	0.15 – 0.73	**0.005**
finance [1]	14.24	-15.33 – 43.82	0.332
Observations	34		
R^2 / R^2 adjusted	0.943 / 0.933		

Figure 7: Μοντέλο Πολλαπλής Παλινδρόμησης εσόδων μετά από Model-Selection.

Σε αυτό το σημείο παίρνουμε το πλήρες μοντέλο για τα έξοδα του νοσοκομείου όπου θέλουμε να είναι της μορφής:

outgoings=β0 + β1 * BED + β2 * MCDAYS + β3 * home + β4 * PCREV + β5 * RURAL + β6 * finance

Μπορούμε να παρατηρήσουμε ότι τα residuals του παραπάνω μοντέλου ακολουθούν την κανονική κατανομή (S-pvalue= 0.06738). Ωστόσο δεν είναι ομοσκεδαστικά (studentized Breusch-Pagan pvalue=0.01193), αλλά είναι ανεξάρτητα (DW pvalue=0.47). Σε αυτό το σημείο θα αλλάξουμε το μοντέλο μας στο παρακάτω, λόγω της ετεροσκεδαστικότητας.

log(outgoings)=β0 + β1 * MCDAYS + β2 * home + β3 * PCREV + β4 * RURAL + β5 * finance

	log(outgoings)		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	9.07	8.56 – 9.57	**<0.001**
MCDAYS	-0.00	-0.00 – -0.00	**0.036**
home	-0.00	-0.00 – 0.00	0.090
PCREV	0.00	0.00 – 0.00	**<0.001**
RURAL [rural]	-0.11	-0.34 – 0.12	0.324
finance [1]	-0.91	-1.38 – -0.44	**<0.001**
Observations	52		
R^2 / R^2 adjusted	0.537 / 0.486		

Figure 8: Μοντέλο πολλαπλής παλινδρόμησης εξόδων με όλες τις μεταβλητές.

Βλέπουμε ότι η σταθερά του μοντέλου μας, δηλαδή η παράμετρος β0, έχει τιμή 9.07.Αυτό σημαίνει ότι τα αναμενόμενα έξοδα ενός νοσοκομείου που είναι σε αγροτική περιοχή, με 0 κρεβάτια και 0 έσοδα, είναι 9.07.

Όμοια με παραπάνω, συμπεραίνουμε ότι οι περισσότερες μεταβλητές έχουν ισχυρή σχέση με την εξαρτημένη μας μεταβλητή, όχι όμως όλες.Προκειμένουν να λύσουμε αυτό το πρόβλημα θα χρσιμοποιήσουμε ξανά τις μεθόδους subset selection και stepwise regression.Όλες οι μέθοδοι προτείνουν ένα μοντέλο με 4 μεταβλητές ,έχοντας αφαιρέσει την μεταβλητή RURAL,καθώς δεν προσφέρει αρκετή πληροφορία στο μοντέλο. Οπότε μετά τη χρήση μεθόδων επιλογής μεταβλητών καταλήγουμε στο παρακάτω μοντέλο.

	log(outgoings)		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	8.97	8.51 – 9.44	**<0.001**
PCREV	0.00	0.00 – 0.00	**<0.001**
finance [1]	-0.91	-1.38 – -0.45	**<0.001**
MCDAYS	-0.00	-0.00 – -0.00	**0.027**
home	-0.00	-0.00 – 0.00	0.128
Observations	52		
R^2 / R^2 adjusted	0.527 / 0.486		

Figure 9: Μοντέλο Πολλαπλής Παλινδρόμησης εξόδων μετά από Model-Selection.

Τέλος, παίρνουμε το πλήρες μοντέλο για τα κέρδη του νοσοκομείου όπου θέλουμε να είναι της μορφής:

$$profit=\beta0 + \beta1 * PCREV + \beta2 * RURAL + \beta3 * finance$$

Όπως μπορούμε να δούμε, δεν συμπεριλάβαμε τις μεταβλητές BED,MCDAYS και home καθώς έχουν υψηλό correlation με την μεταβλητή PCREV.

Μπορούμε να παρατηρήσουμε ότι τα residuals του παραπάνω μοντέλου δεν ακολουθούν την κανονική κατανομή (S-pvalue= 0.06738). Ωστόσο δεν είναι ομοσκεδαστικά (studentized Breusch-Pagan pvalue=0.01193) αλλά δεν είναι και ανεξάρτητα (DW pvalue=0.04). Όμοια με τα δύο προηγούμενα μοντέλα, θα αναταξινομήσουμε τα δεδομένα μας με τυχαία σειρά. Ακόμα θα κεντροποιήσουμε τις αριθμητικές μας μεταβλητές και θα για να λύσουμε το πρόβλημα της ετεροσκεδαστικότητας, θα μετατρέψουμε το μοντέλο μας στο ακόλουθο:

$$log(profit)=\beta0 + \beta1 * PCREV + \beta2 * RURAL + \beta3 * finance$$

Σε αυτό το σημείο πρέπει να αναφέρουμε ότι το μοντέλο μας είναι overfitted, συνεπώς θα χρησιμοποιήσουμε ξανά τις μεθόδους subset selection και stepwise regression για να λύσουμε αυτό το πρόβλημα.Όλες οι μέθοδοι προτείνουν ένα μοντέλο με 3 μεταβλητές, τις PCREV και outgoings. Το μοντέλο φαίνεται παρακάτω.

Τα παραπάνω μοντέλα είναι δύσκολο να ερμηνευτούν αλλά ο σκοπός τους δεν είναι να ερμηνεύσουν τα έσοδα, τα έξοδα και το κέρδος αλλά να τα προβλέψουν ώστε να δούμε αν το νοσοκομείο θα είναι σε καλή κατάσταση, αν θα μπορεί δηλαδή να υπερκαλύπτει με τα έσοδα του τα έξοδα του ώστε να είναι επικερδές.

	log(profit)		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	8.76	8.66 – 8.86	**<0.001**
PCREV	0.00	0.00 – 0.00	**<0.001**
outgoings	-0.00	-0.00 – -0.00	**<0.001**
Observations	49		
R^2 / R^2 adjusted	0.754 / 0.744		

Figure 10: Μοντέλο Πολλαπλής Παλινδρόμησης κέρδους μετά από Model-Selection.

5 Συμπεράσματα και Συζήτηση

Η παραπάνω μελέτη είχε ως σκοπό την εκτίμηση ενός υποδείγματος για την λειτουργία των νοσοκομειακών μονάδων, τα έσοδα τους, τα έξοδα και το κέρδος συγκριτικά με το αν η μονάδα βρίσκεται σε αγροτική ή αστική περιοχή. Το συμπέρασμα που βγάζουμε από την παραπάνω ανάλυση είναι ότι τα έσοδα ενός νοσοκομείου επηρεάζονται από την τοποθεσία του νοσοκομείου, κάτι που δεν ισχύει όμως για τα έξοδα, τα οποία επηρεάζονται περισσότερο από τις ημέρες νοσηλείας. Όσον αφορά το κέρδος, βλέπουμε ότι δεν επηρεάζεται από την περιοχή που βρίσκεται το νοσοκομείο, αλλά από τα ίδια τα έσοδα και τα έξοδα της. Συνεπώς όσο το νοσοκομείο καταφέρνει να αυξάνει τα έσοδα του και να μειώνει τα έξοδα του, θα αυξάνονται και τα κέρδη του ανεξάρτητα με το αν είναι χτισμένο σε αγροτική ή αστική περιοχή. Συνεπώς η άποψη ότι μόνο τα νοσοκομεία στα αστικά κέντρα είναι σε καλή κατάσταση οικονομικά, καταρρίπτεται. Συμπερασματικά, στο ερώτημα ποια νοσοκομεία βρίσκονται σε καλή κατάσταση, η απάντηση είναι αυτά που καταφέρνουν να έχουν μεγαλύτερο κέρδος, σε αντίθεση με την κυριάρχουσα γνώμη που θέλει επιτυχημένα οικονομικά μονό εκείνα των αστικών περιοχών. Τα αποτελέσματα αυτής της μελέτης μπορούν να παρακινήσουν τις κυβερνήσεις να χρηματοδοτήσουν τις αγροτικές περιοχές χωρίς την προκατάληψη ότι δεν είναι στο επίπεδο των αστικών κέντρων. Η υγεία είναι δικαίωμα για όλους και όχι αγαθό για τους λίγους.

6 Παράρτημα

Πίνακας 3: Συγκεντρωτικός Πίνακας Δεδομένων				
Αριθμός Μεταβλητής	Όνομα	Τύπος	Σημασία	Τιμές
1	BED	αριθμητική	αριθμός κρεβατιών	-
2	MCDAYS	αριθμητική	ετήσιος αριθμός ιατρικής περίθαλψης μέσα στο νοσοκομείο (x100)	-
3	TDAYS	αριθμητική	ετήσιος αριθμός ιατρικής περίθαλψης (στο νοσοκομείο ή στο σπίτι) (x100)	-
4	PCREV	αριθμητική	ετήσιο συνολικό εισόδημα περίθαλψης (σε εκατοντάδες (dollars)	-
5	NSAL	αριθμητική	ετήσιο συνολικό ποσό μισθών νοσοκόμων (σε εκατοντάδες dollars)	-
6	FEXP	αριθμητική	ετήσια έξοδα εγκαταστάσεων (σε εκατοντάδες dollars)	-
7	RURAL	κατηγορική	τοποθεσία	1:αγροτική 0:μη αγροτική
8	profit	αριθμητική	καθαρό κέρδος νοσοκομειακών μονάδων	-
9	home	αριθμητική	ετήσιος αριθμός ιατρικής περίθαλψης στο σπίτι (x100)	-
10	finance	κατηγορική	οικονομική κατάσταση νοσοκομείου	1: καλή 0:κακή
11	outgoings	αριθμητική	αθροισμα εξόδων του νοσοκομείου (μισθοί και εγκαταστάσεις)	-

Σε αυτό το σημείο θα παρουσιάσουμε τα διαγράμματα των κατηγορικών μεταβλητών.

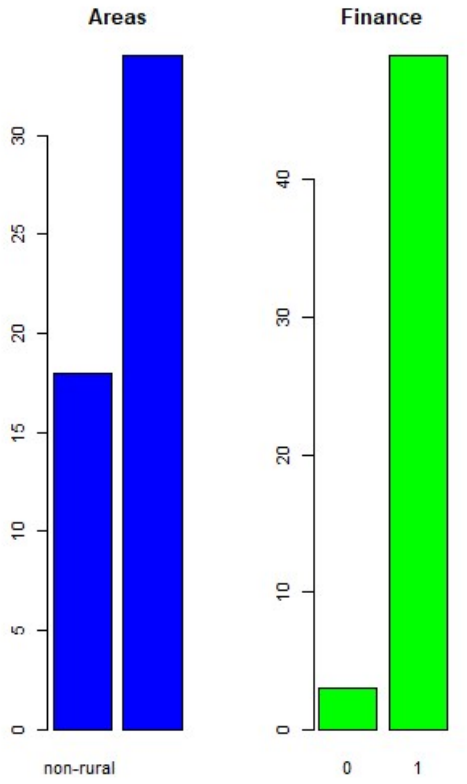


Figure 11: Plots της κατηγορικής μεταβλητής RURAL όπου βλέπουμε ότι τα περισσότερα νοσοκομεία είναι σε αγροτικές περιοχές καθώς και της finance όπου βλέπουμε ότι τα περισσότερα νοσοκομεία υπερκαλύπτουν τα έξοδα τους.

Σε αυτό το σημείο έχουμε το barplot για τις 2 κατηγορικές μεταβήτές RURAL και finance.

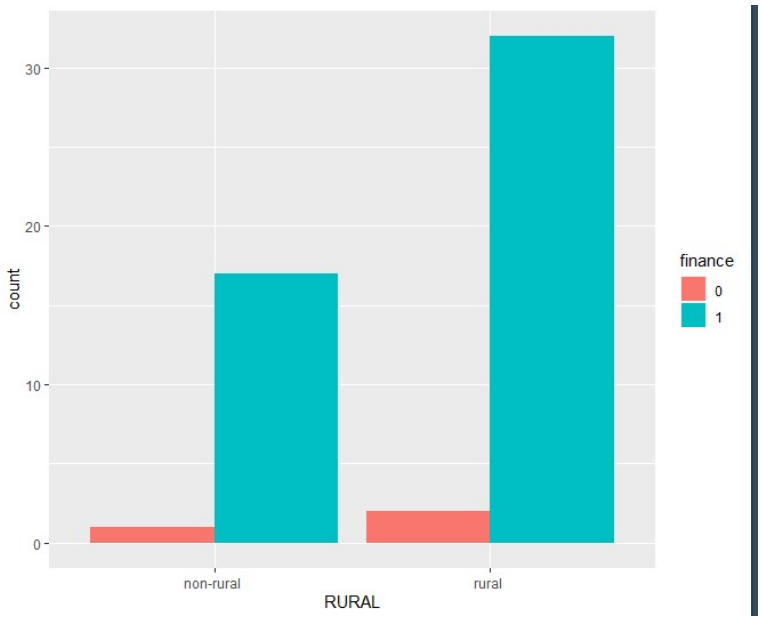
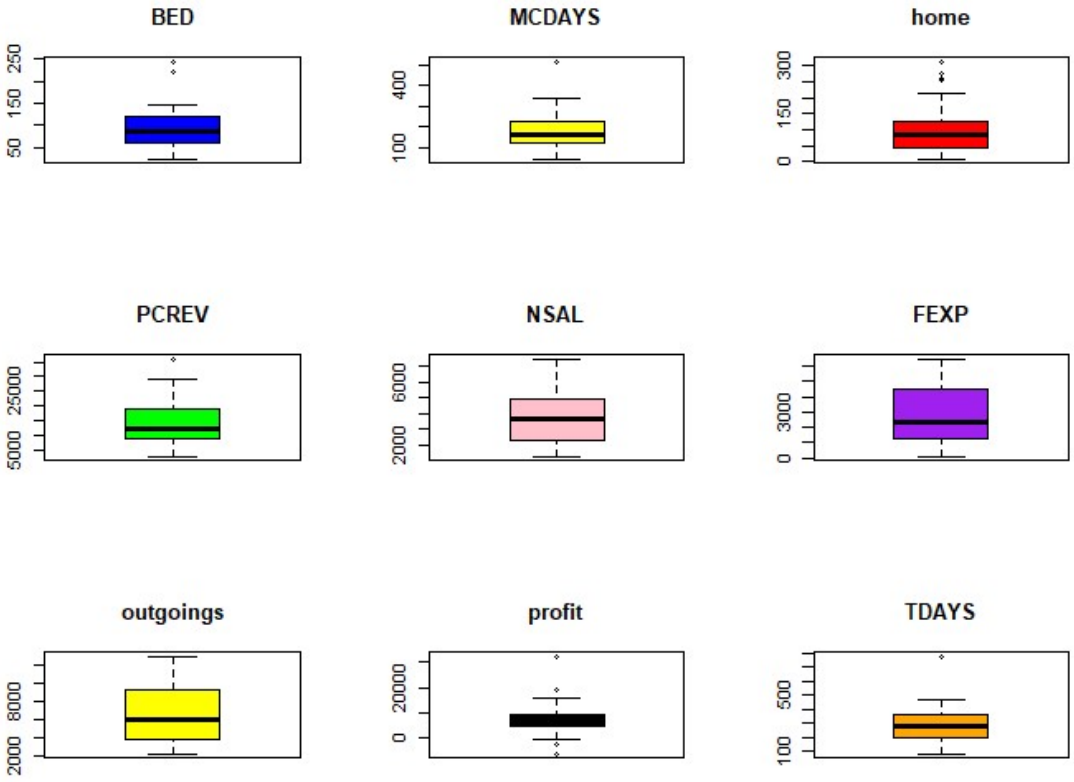


Figure 12: Barplot των 2 κατηγορικών μεταβλητών

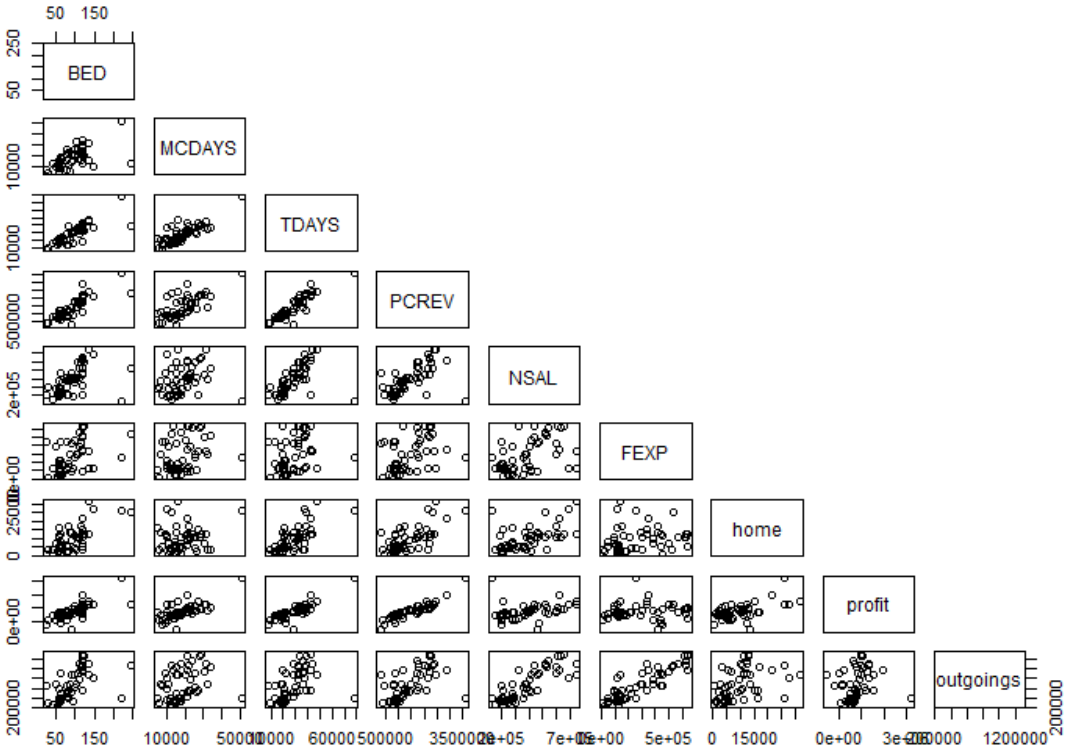
Σε αυτό το σημείο παρουσιάζουμε ta boxplots για τις αριθμητικές μεταβλητές.



(a)

Figure 13: Boxplot και το QQ-Plot της profit.

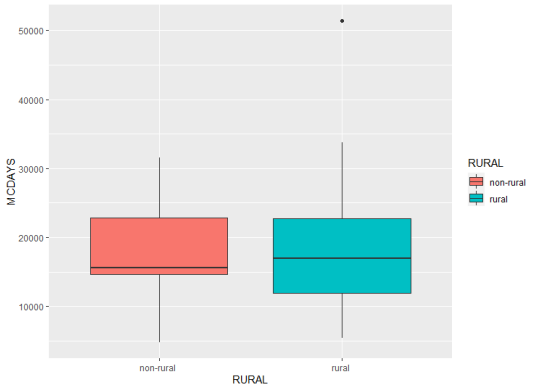
Σε αυτό το σημείο παρουσιάζουμε το ανα δύο Scatterplot για τις αριθμητικές μεταβλητές.



(a)

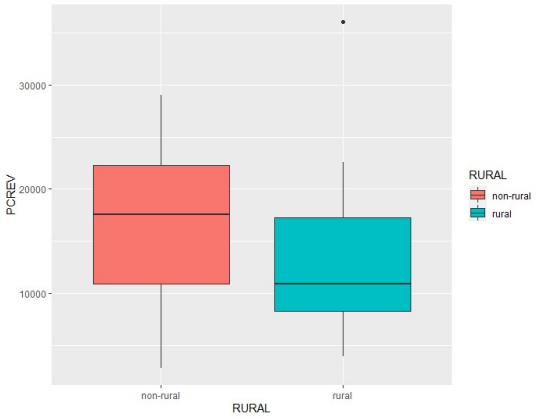
Figure 14: Ανά δύο Scatterplot για τους συνδυασμούς των αριθμητικών μεταβλητών

Σε αυτό το σημείο θα παραθέσουμε τα boxplots κι errorbars των μεταβλητών



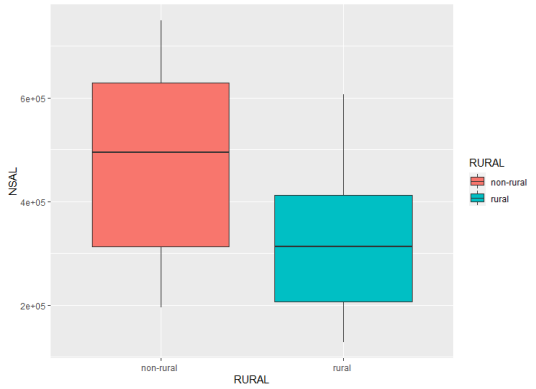
(a)

Figure 15: Boxplot της MCDAYS για κάθε κατηγορία της RURAL



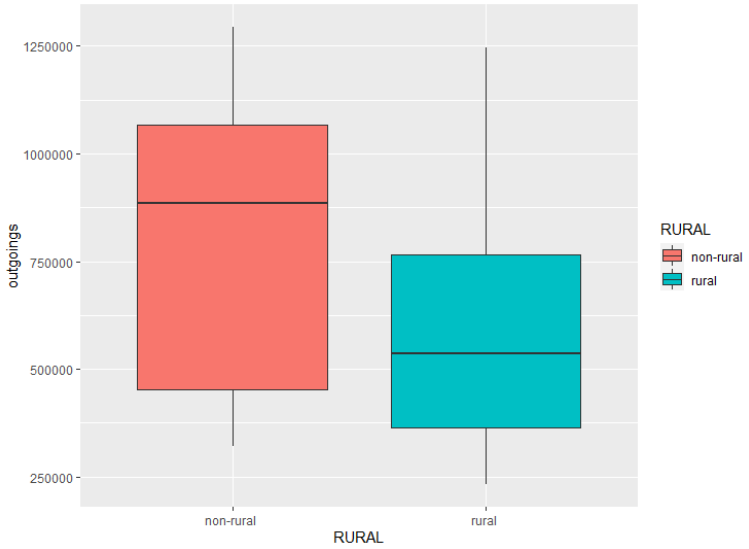
(a)

Figure 16: Boxplot της PCREV για κάθε κατηγορία της RURAL



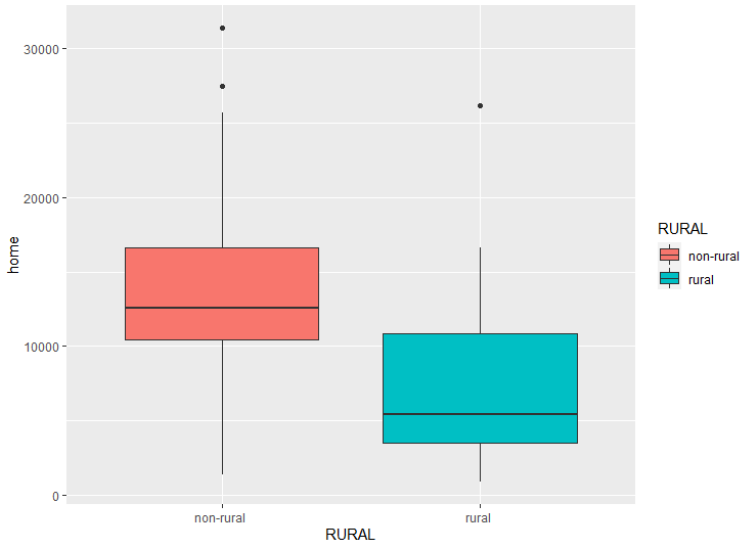
(a)

Figure 17: Boxplot της NSAL για κάθε κατηγορία της RURAL



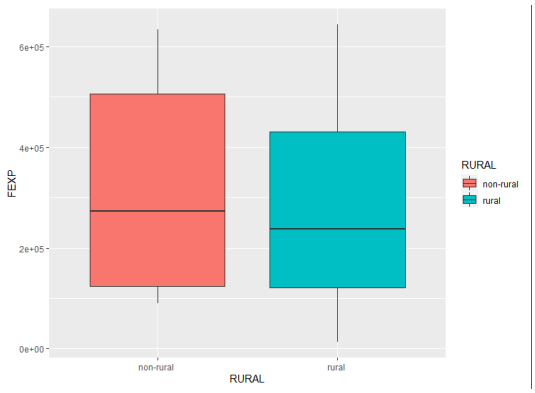
(a)

Figure 18: Boxplot της outgoings για κάθε κατηγορία της RURAL



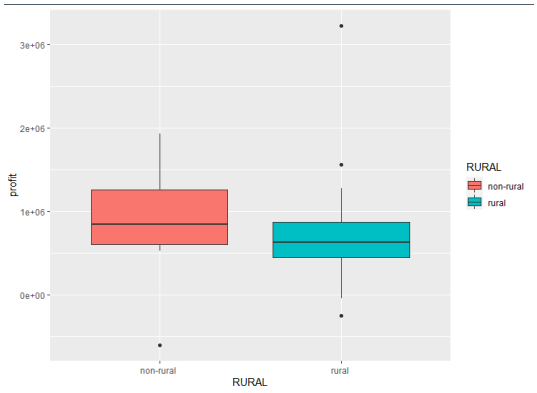
(a)

Figure 19: Boxplot της home για κάθε κατηγορία της RURAL



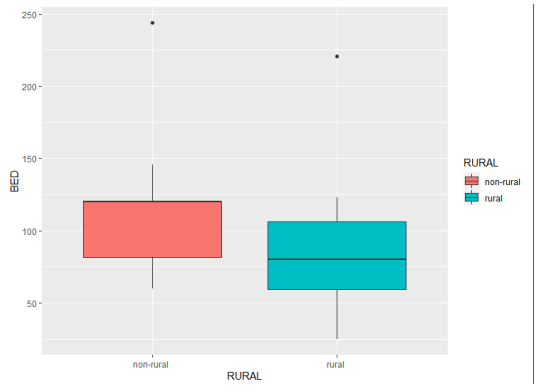
(a)

Figure 20: Boxplots της FEXP για κάθε κατηγορία της RURAL



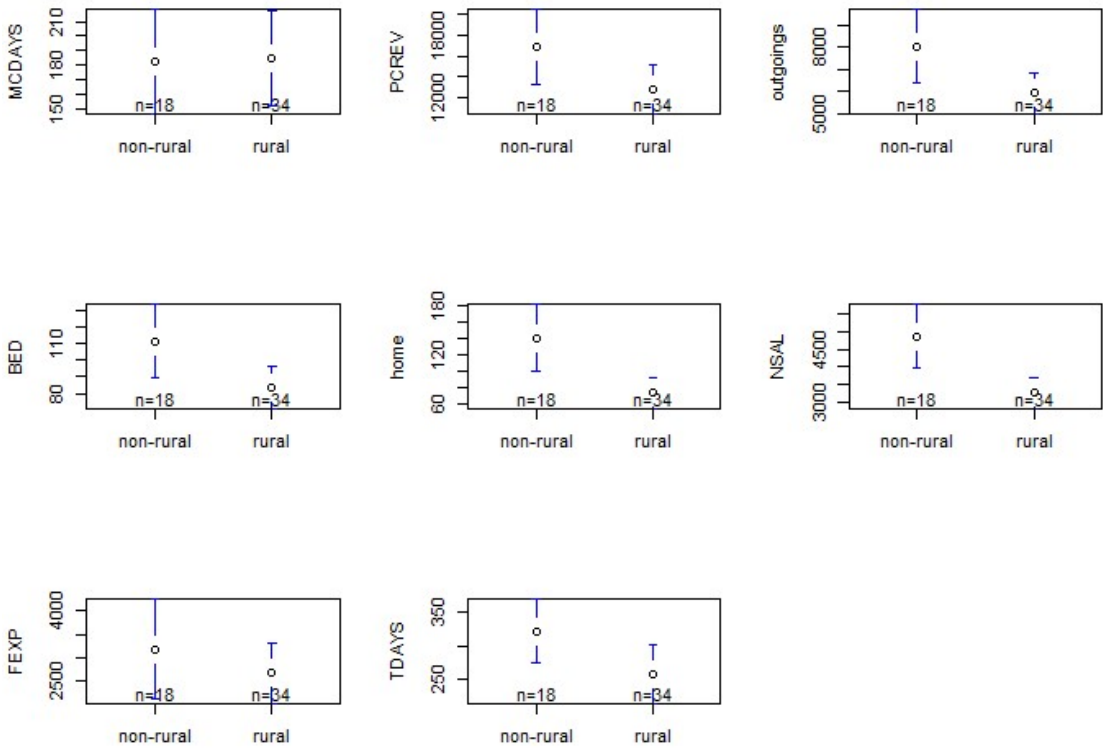
(a)

Figure 21: Boxplots της profit για κάθε κατηγορία της RURAL



(a)

Figure 22: Boxplts της BED για κάθε κατηγορία της RURAL



(a)

Figure 23: Errorbars όλων των μεταβλητών για κάθε κατηγορία της RURAL

Σύγκριση των εσόδων με την κατηγορική μεταβλητή RURAL

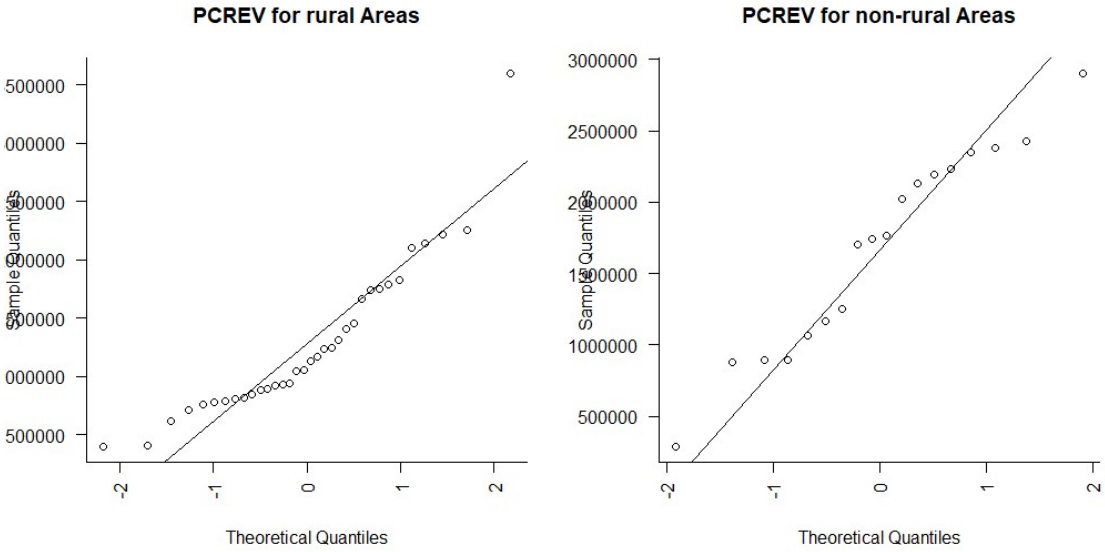


Figure 24: QQplots για να δούμε την κατανομή των εσόδων για κάθε νοσοκομειακή μονάδα σε αγροτικές και μη περιοχές.

Σύγκριση των εξόδων με την κατηγορική μεταβλητή RURAL

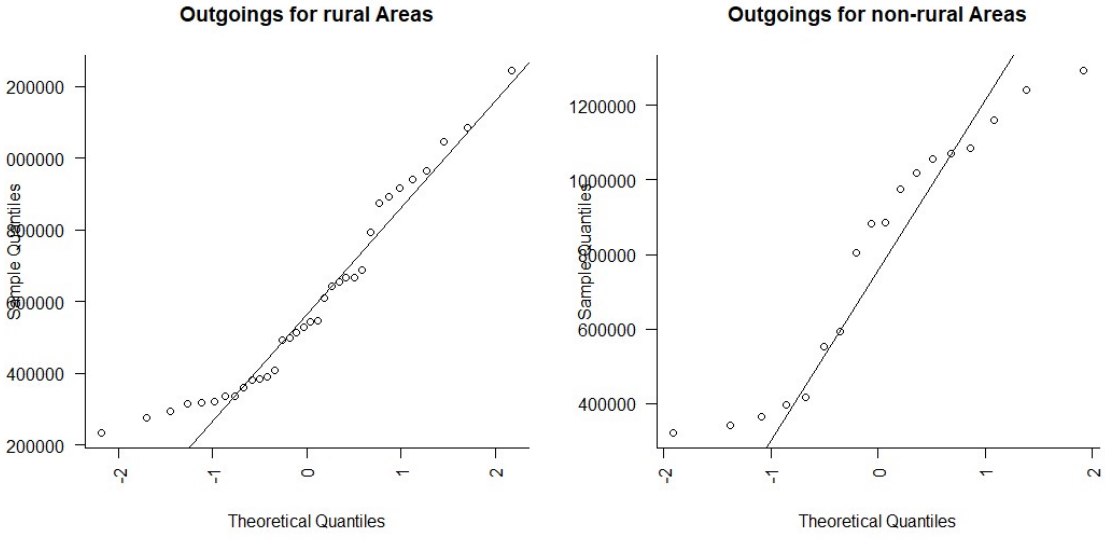


Figure 25: QQplots για να δούμε την κατανομή των εξόδων για κάθε νοσοκομειακή μονάδα σε αγροτικές και μη περιοχές.

Σύγκριση του κέρδους με την κατηγορική μεταβλητή RURAL

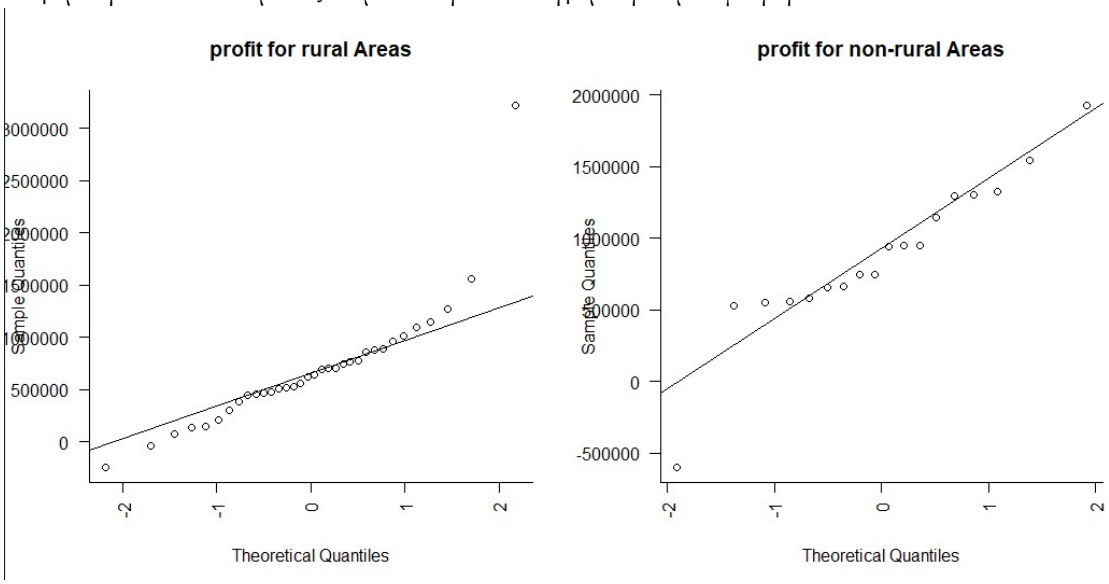


Figure 26: QQplots για να δούμε την κατανομή του κέρδους για κάθε νοσοκομειακή μονάδα σε αγροτικές και μη περιοχές.

Μοντέλο Παλινδρόμησης εσόδων με όλες τις μεταβλητές

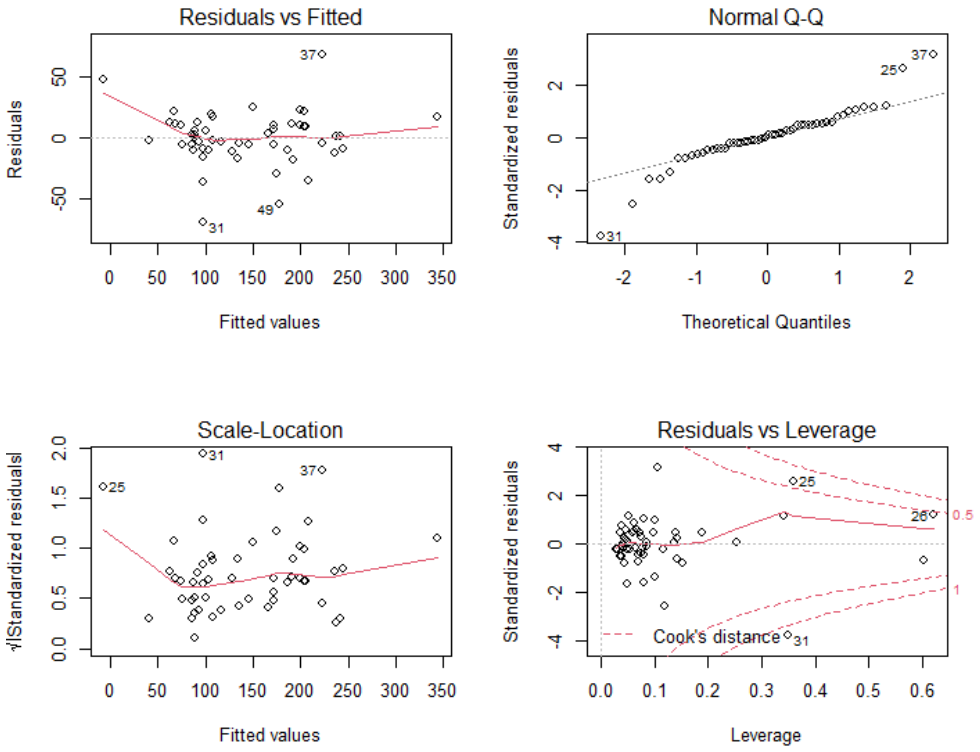


Figure 27: Το plot για τους ελέγχους υποθέσεων του μοντέλου για τα έσοδα με όλες τις μεταβλητές

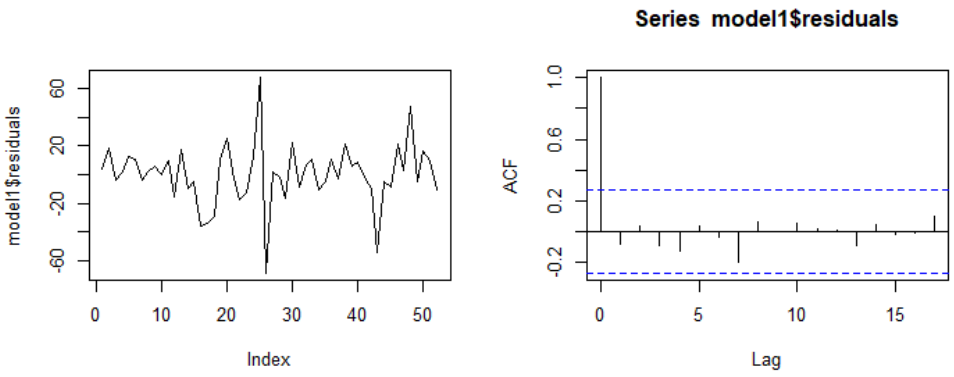


Figure 28: Διάγραμματα ανεξαρτησίας των καταλοίπων και αυτοσυσχέτισης για το πλήρες μοντέλο των εσόδων

Μοντέλο Παλινδρόμησης εσόδων μετά την αναταξινόμηση και την κεντροποίηση των αριθμητικών μεταβλητών

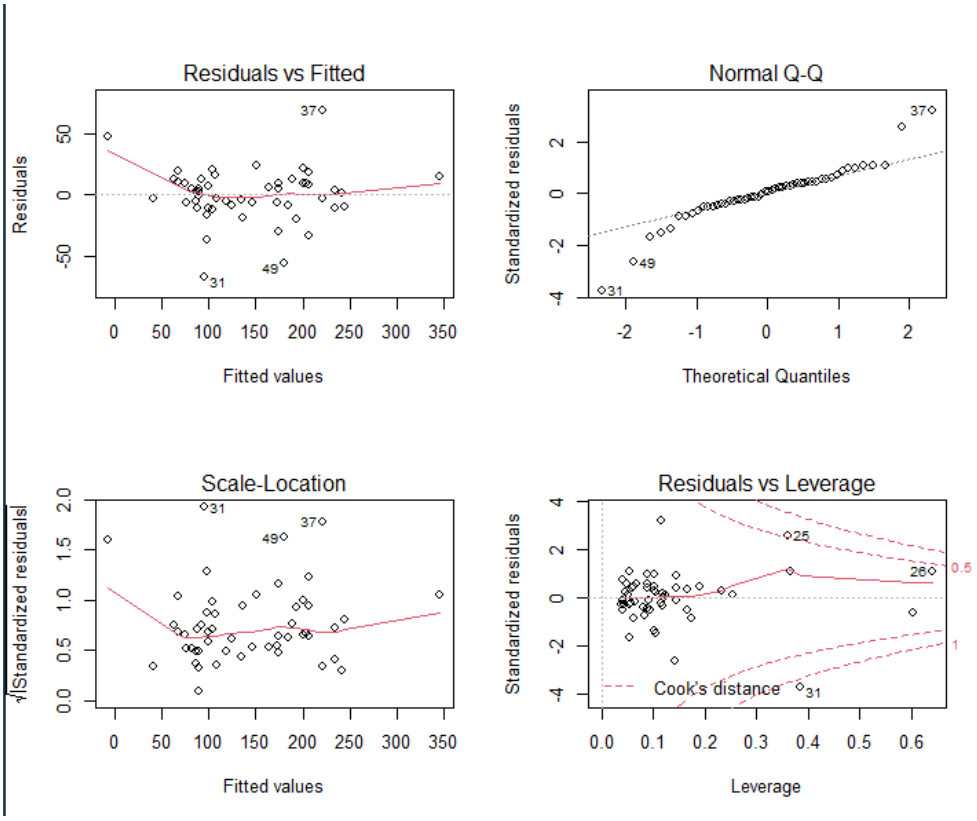


Figure 29: Το plot για τους ελέγχους υποθέσεων μετά την αναταξινόμηση και την κεντροποίηση των αριθμητικών μεταβλητών

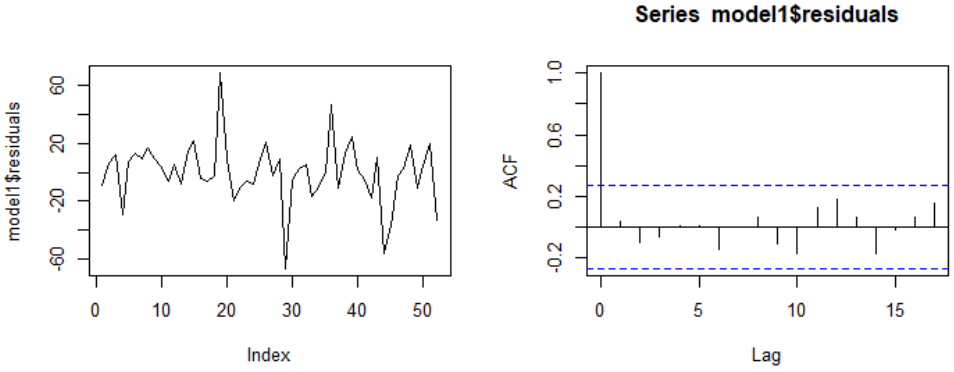


Figure 30: Διάγραμματα ανεξαρτησίας των καταλοίπων και αυτοσυσχέτισης μετά την αναταξινόμηση και την κεντροποίηση των αριθμητικών μεταβλητών

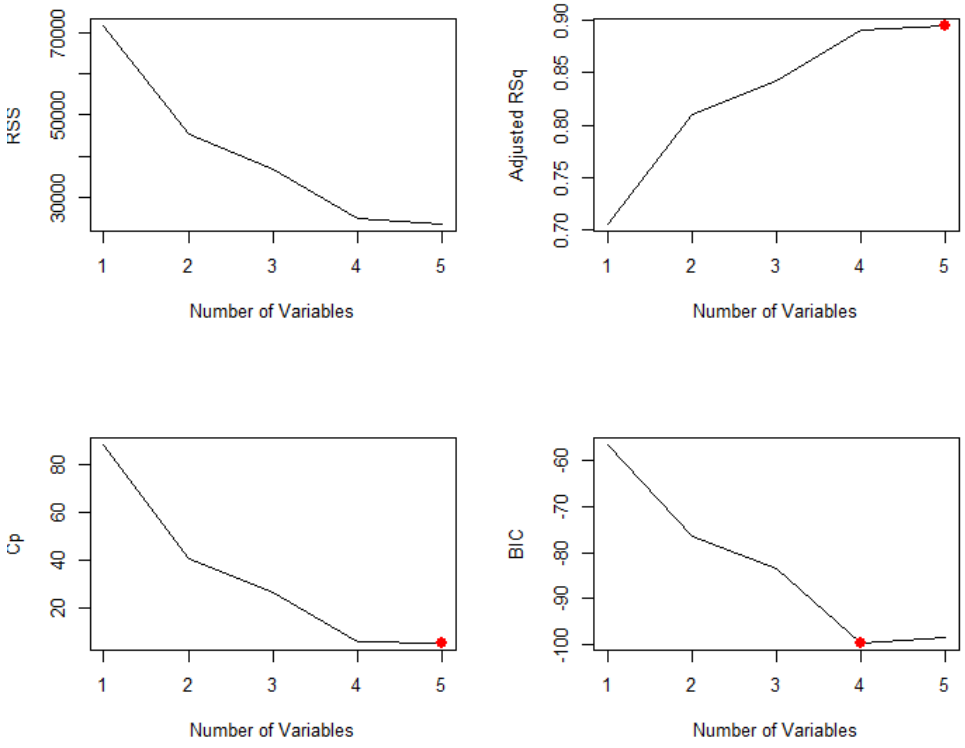


Figure 31: Μέθοδος subsetselection : Διαγράμματα επιλογής αριθμού μεταβλητών ανάλογα με τα αντίστοιχα κριτήρια

Μοντέλο Παλινδρόμησης μετά τις μεθόδους επιλογής μεταβλητών subsetselection,stepwise regression

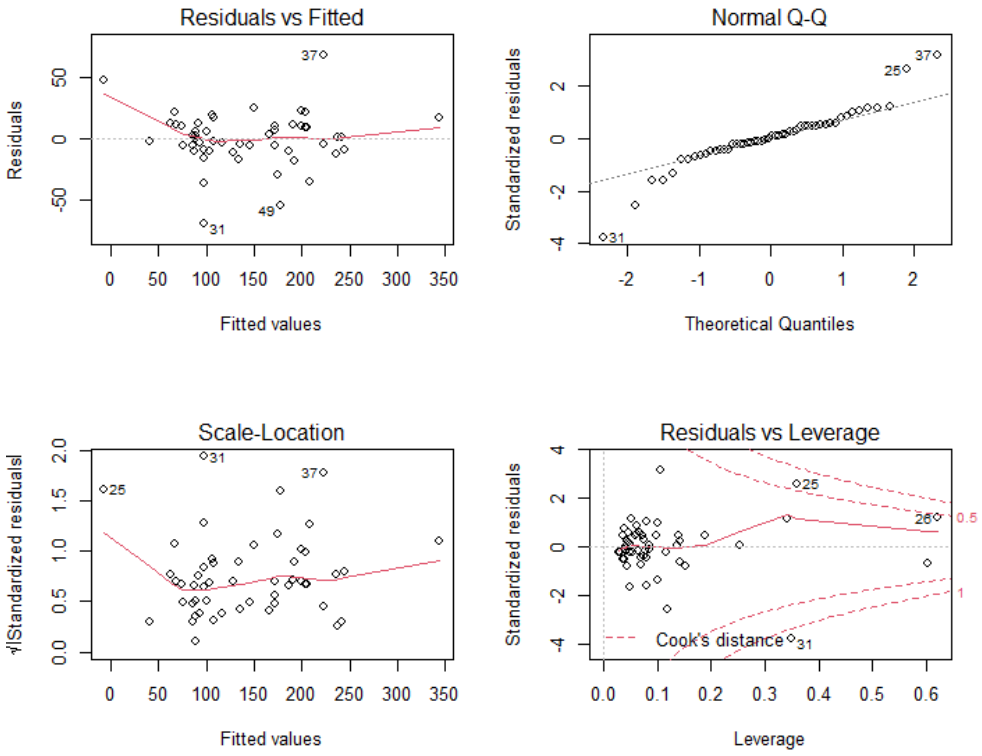


Figure 32: Το plot για τους ελέγχους υποθέσεων μετά τις μεθόδους επιλογής μεταβλητών subsetselection,stepwise regression

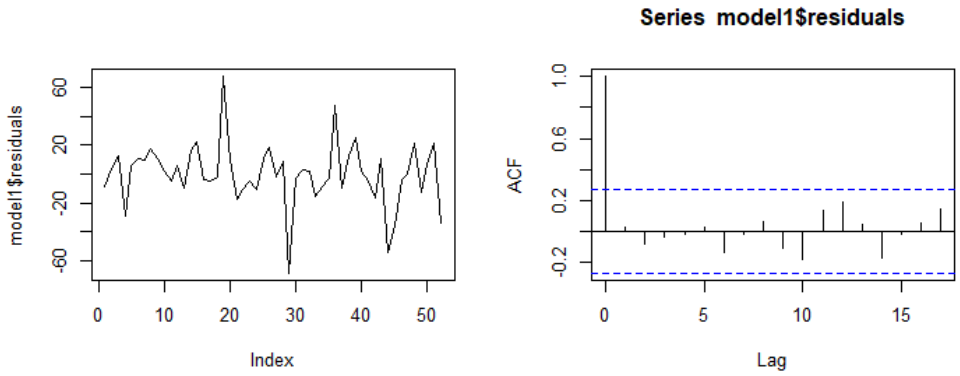


Figure 33: Διάγραμματα ανεξαρτησίας των καταλοίπων και αυτοσυσχέτισης μετά τις μεθόδους επιλογής μεταβλητών subsetselection,stepwise regression

Μοντέλο Παλινδρόμησης εξόδων με όλες τις μεταβλητές

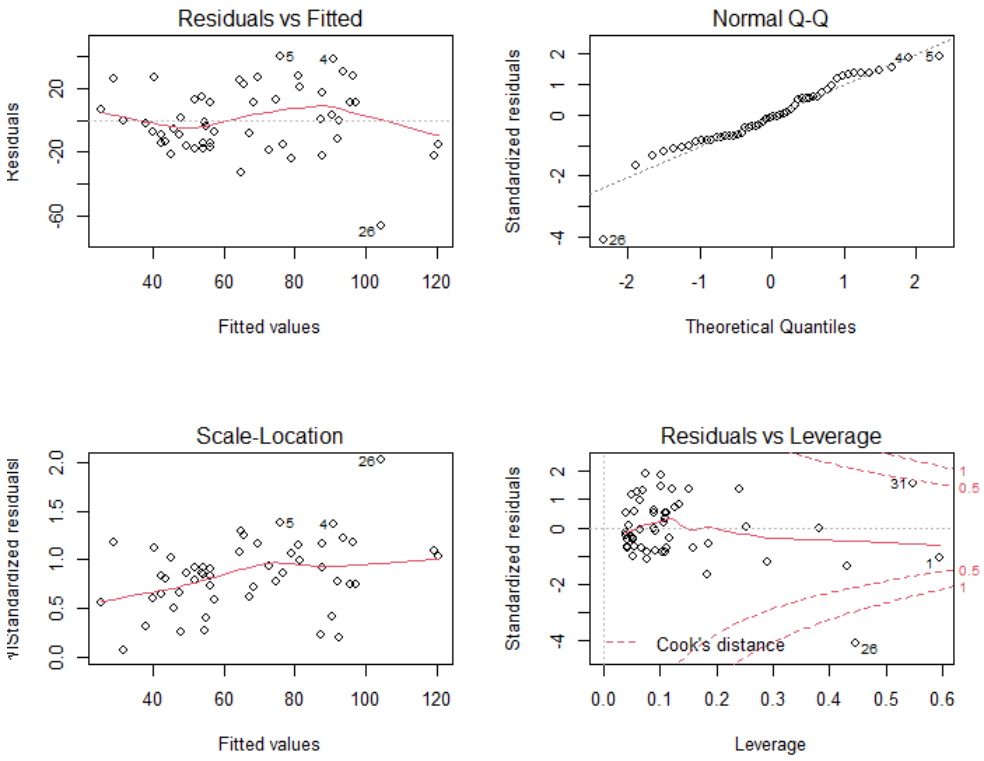


Figure 34: Το plot για τους ελέγχους υποθέσεων του μοντέλου για τα έξοδα με όλες τις μεταβλητές

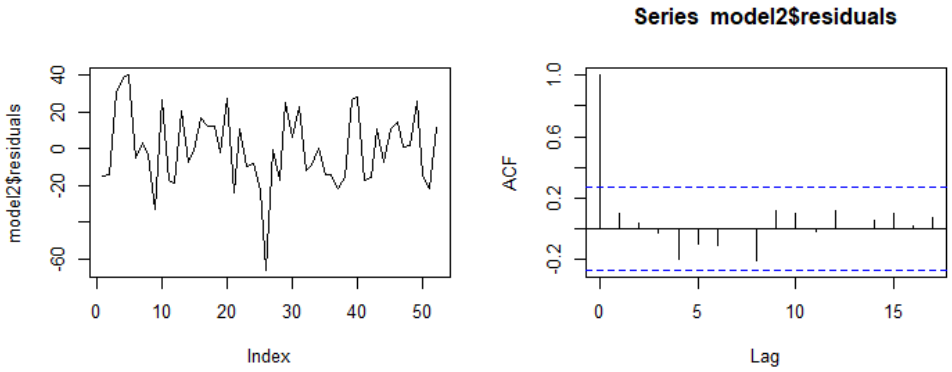


Figure 35: Διάγραμματα ανεξαρτησίας των καταλοίπων και αυτοσυσχέτισης για το πλήρες μοντέλο των εξόδων

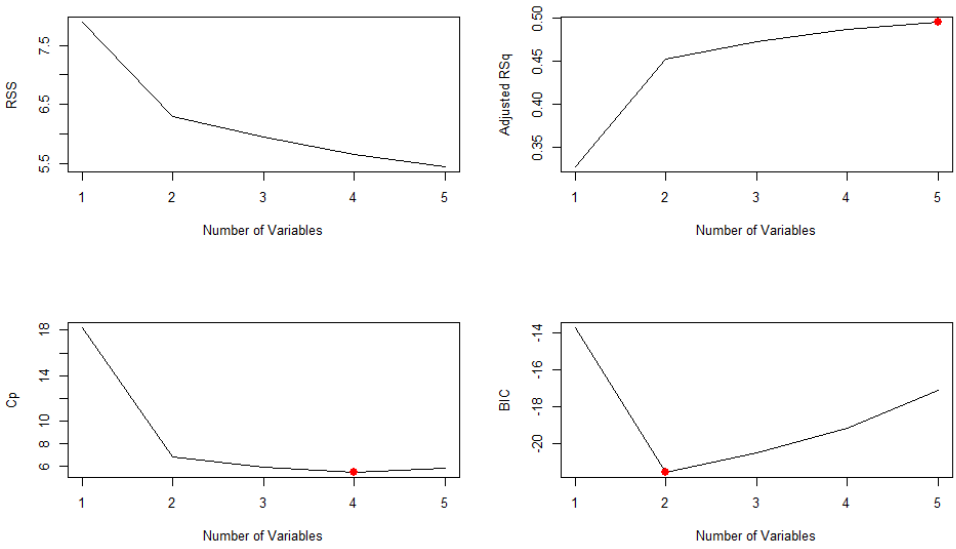


Figure 36: Μέθοδος subsetselection : Διαγράμματα επιλογής αριθμού μεταβλητών ανάλογα με τα αντίστοιχα κριτήρια

Μοντέλο Παλινδρόμησης μετά τις μεθόδους επιλογής μεταβλητών subsetselection,stepwise regression

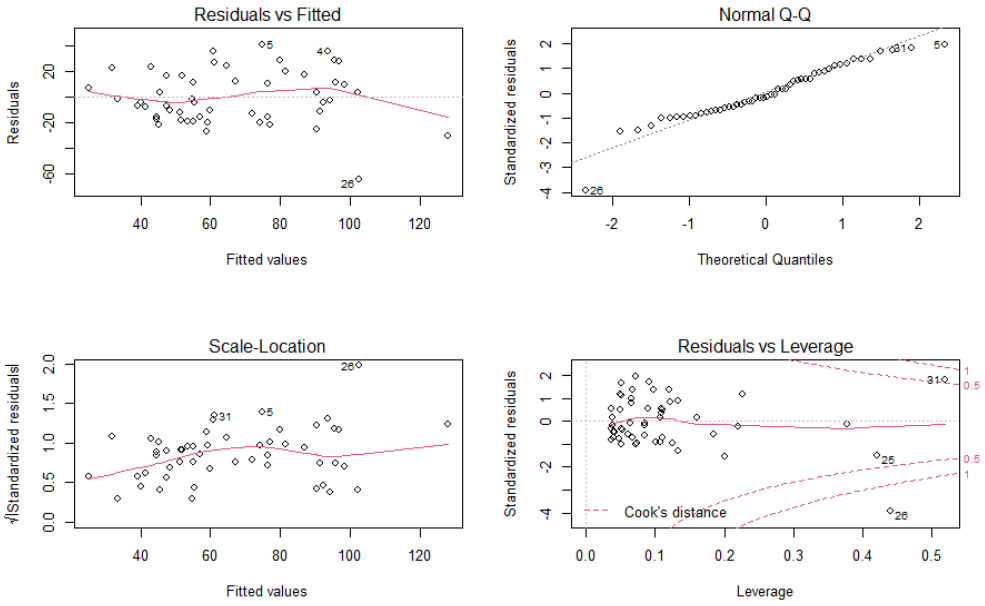


Figure 37: Το plot για τους ελέγχους υποθέσεων μετά τις μεθόδους επιλογής μεταβλητών subsetselection,stepwise regression

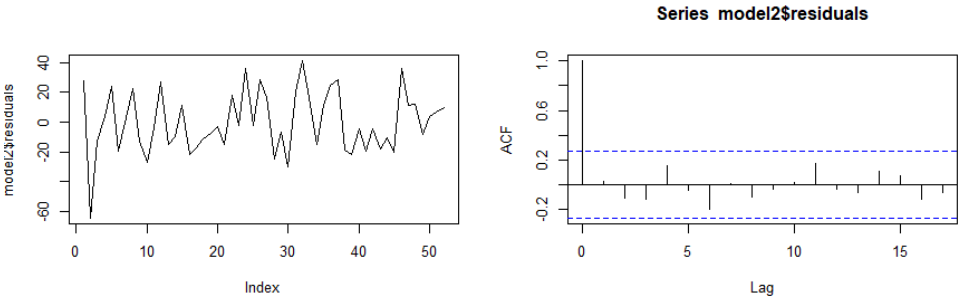


Figure 38: Διάγραμματα ανεξαρτησίας των καταλοίπων και αυτοσυσχέτισης μετά τις μεθόδους επιλογής μεταβλητών subsetselection,stepwise regression

Μοντέλο Παλινδρόμησης κέρδους με όλες τις μεταβλητές

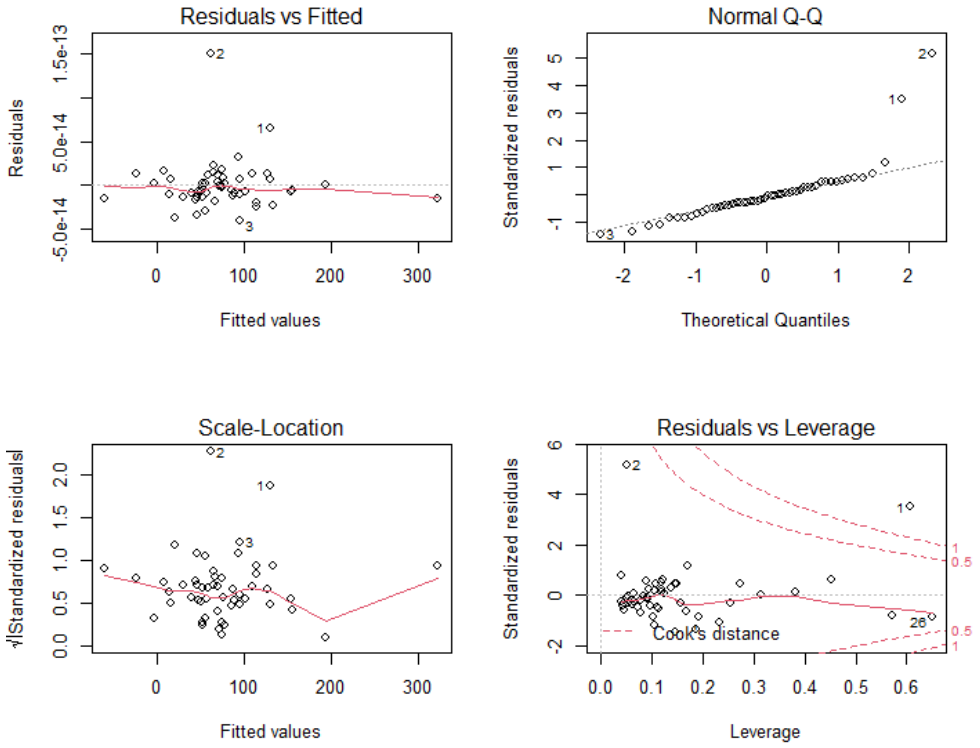


Figure 39: Το plot για τους ελέγχους υποθέσεων του μοντέλου για τα κέρδη με όλες τις μεταβλητές

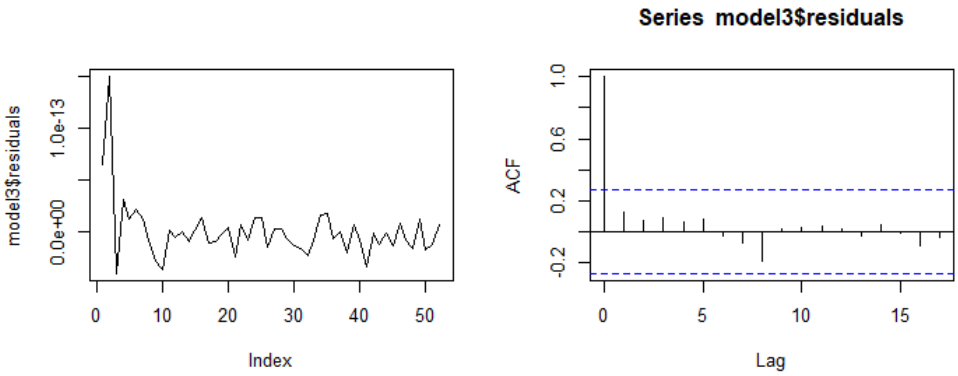


Figure 40: Διάγραμματα ανεξαρτησίας των καταλοίπων και αυτοσυσχέτισης για το πλήρες μοντέλο των κερδών

Μοντέλο Παλινδρόμησης κέρδους μετά την αναταξινόμηση και την κεντροποίηση των αριθμητικών μεταβλητών

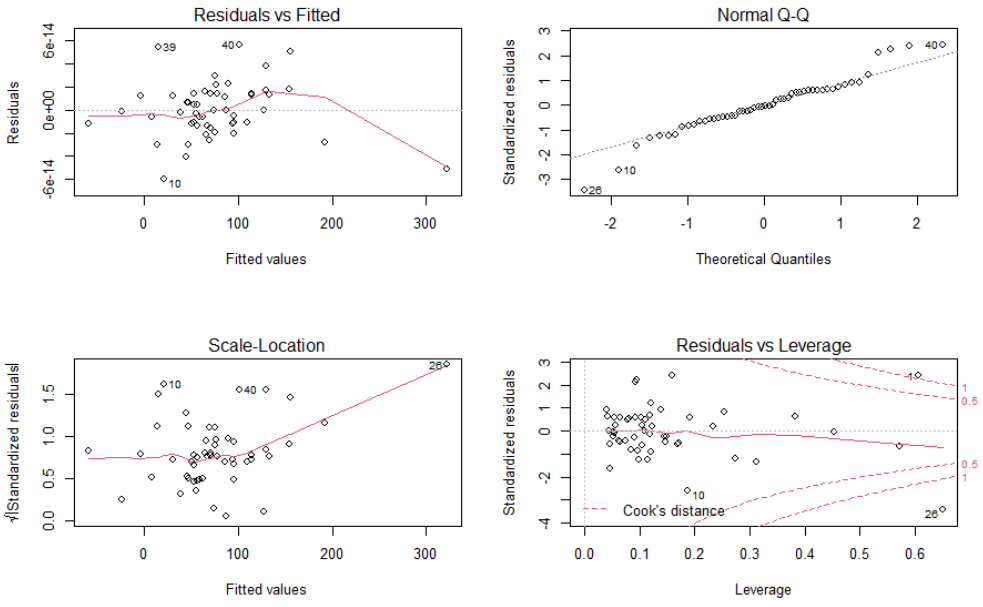


Figure 41: To plot για τους ελέγχους υποθέσεων μετά την αναταξινόμηση και την κεντροποίηση των αριθμητικών μεταβλητών

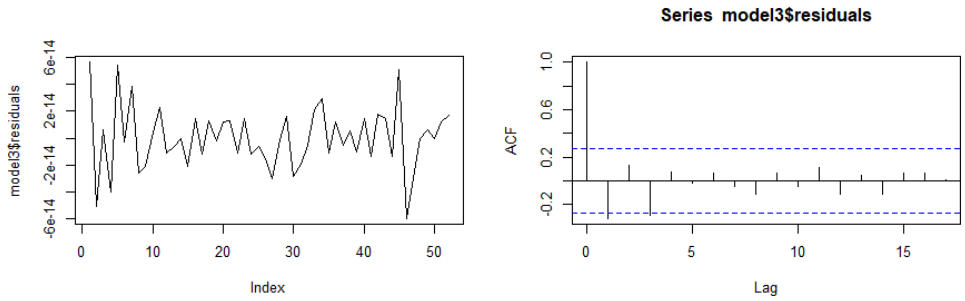


Figure 42: Διάγραμματα ανεξαρτησίας των καταλοίπων και αυτοσυσχέτισης μετά την αναταξινόμηση και την κεντροποίηση των αριθμητικών μεταβλητών

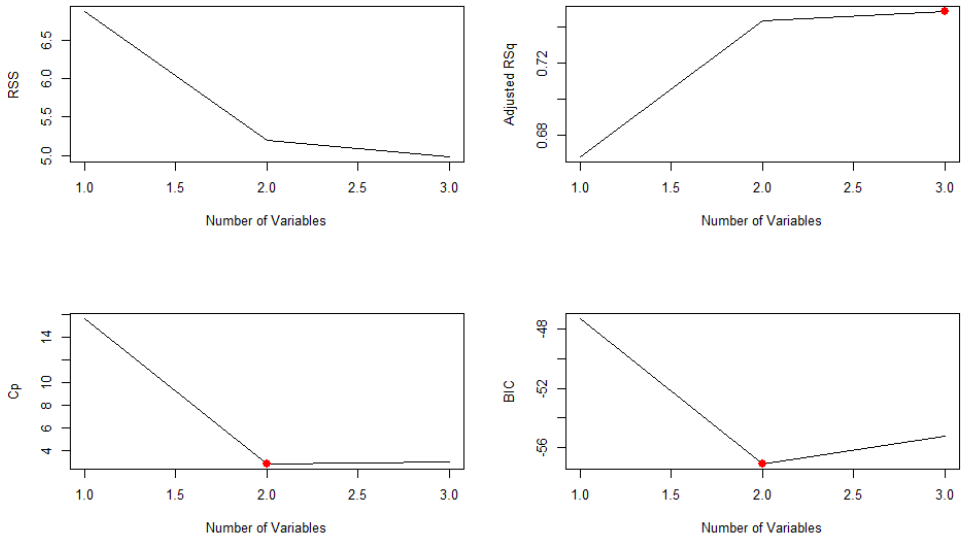


Figure 43: Μέθοδος subsetselection : Διαγράμματα επιλογής αριθμού μεταβλητών ανάλογα με τα αντίστοιχα κριτήρια

Μοντέλο Παλινδρόμησης μετά τις μεθόδους επιλογής μεταβλητών subsetselection,stepwise regression

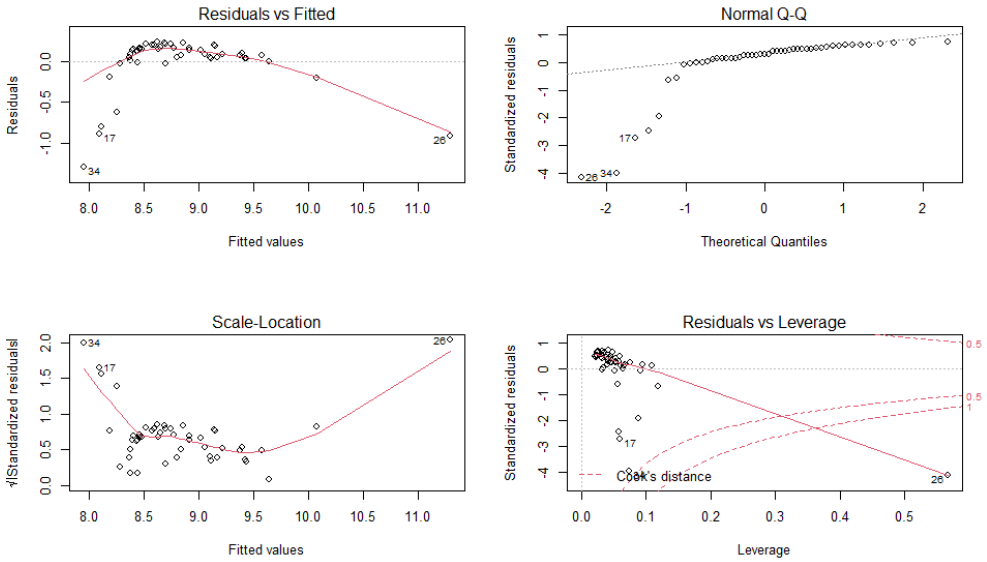


Figure 44: Το plot για τους ελέγχους υποθέσεων μετά τις μεθόδους επιλογής μεταβλητών subsetselection,stepwise regression

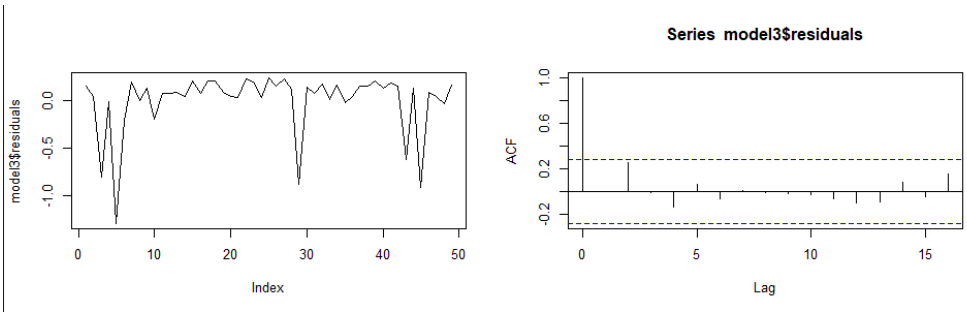


Figure 45: Διάγραμματα ανεξαρτησίας των καταλοίπων και αυτοσυσχέτισης μετά τις μεθόδους επιλογής μεταβλητών subsetselection,stepwise regression