



ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

**ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ
ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ**

ΤΟΜΕΑΣ ΜΑΘΗΜΑΤΙΚΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΣΤΑΤΙΣΤΙΚΗ ΑΝΑΛΥΣΗ ΤΩΝ ΔΕΔΟΜΕΝΩΝ ΤΗΣ ΕΡΕΥΝΑΣ ΟΙΚΟΓΕΝΕΙΑΚΩΝ ΠΡΟΫΠΟΛΟΓΙΣΜΩΝ

ΠΗΝΕΛΟΠΗ ΚΑΣΙΔΟΚΩΣΤΑ

**ΕΠΙΒΛΕΠΩΝ: Δ.ΦΟΥΣΚΑΚΗΣ, ΕΠΙΚΟΥΡΟΣ ΚΑΘΗΓΗΤΗΣ
ΕΠΙΤΡΟΠΗ: Ι. ΣΠΗΛΙΩΤΗΣ, ΑΝΑΠΛΗΡΩΤΗΣ ΚΑΘΗΓΗΤΗΣ
Ι. ΚΟΛΕΤΣΟΣ, ΕΠΙΚΟΥΡΟΣ ΚΑΘΗΓΗΤΗΣ**

Απρίλιος, 2014

Στη Γαρυφαλλιά

ΠΕΡΙΕΧΟΜΕΝΑ

Πρόλογος.....	6
1 ΕΙΣΑΓΩΓΗ.....	8
1.1. Τι είναι η Έρευνα Οικογενειακών Προϋπολογισμών.....	8
1.2. Σκοπός της έρευνας.....	9
1.3. Σχεδιασμός της έρευνας.....	9
1.4. Μέγεθος του δείγματος.....	10
1.5. Χρόνος διενέργειας της έρευνας και περίοδοι αναφοράς των δαπανών και των εισοδημάτων.....	12
2 ΑΝΑΛΥΣΗ ΜΕΤΑΒΛΗΤΩΝ.....	13
2.1. Περιγραφικά στοιχεία μεταβλητών.....	13
2.1.1. Περιοχή.....	14
2.1.2. Βαθμός πυκνότητας της περιοχής.....	15
2.1.3. Κύρια πηγή εισοδήματος του νοικοκυριού.....	17
2.1.4. Αριθμός οικονομικά ενεργών μελών του νοικοκυριού.....	19
2.1.5. Κοινωνικο – οικονομική κατάσταση του προσώπου αναφοράς.....	21
2.1.6. Σύνολο εισοδήματος του νοικοκυριού.....	24
2.1.7. Σύνολο καταναλωτικής δαπάνης του νοικοκυριού.....	26
2.1.8. Κατηγορίες δαπανών των νοικοκυριών.....	29
2.2. Συσχετίσεις Μεταβλητών.....	41
2.2.1. Περιοχή – Κύρια πηγή εισοδήματος.....	43
2.2.2. Περιοχή – Βαθμός πυκνότητας.....	45
2.2.3. Περιοχή – Αριθμός οικονομικά ενεργών μελών του νοικοκυριού.....	47
2.2.4. Περιοχή – Κοινωνικο – οικονομική κατάσταση του προσώπου αναφοράς του νοικοκυριού.....	49
2.2.5. Κύρια πηγή εισοδήματος – Βαθμός πυκνότητας.....	51
2.2.6. Κύρια πηγή εισοδήματος – Αριθμός οικονομικά ενεργών μελών του νοικοκυριού.....	54
2.2.7. Κύρια πηγή εισοδήματος – Κοινωνικο – οικονομική κατάσταση του προσώπου αναφοράς του νοικοκυριού.....	57
2.2.8. Βαθμός πυκνότητας – Αριθμός οικονομικά ενεργών μελών του νοικοκυριού.....	60
2.2.9. Βαθμός πυκνότητας – Κοινωνικο – οικονομική κατάσταση του προσώπου αναφοράς του νοικοκυριού.....	62
2.2.10. Αριθμός οικονομικά ενεργών μελών του νοικοκυριού – Κοινωνικο – οικονομική κατάσταση του προσώπου αναφοράς του νοικοκυριού.....	64
2.2.11. Κατηγορίες δαπανών του νοικοκυριού.....	67
3 ΣΤΑΤΙΣΤΙΚΑ ΜΟΝΤΕΛΑ ΠΑΛΙΝΔΡΟΜΗΣΗΣ.....	70
3.1. Εισαγωγή.....	70
3.2. Γραμμική Παλινδρόμηση.....	70
3.2.1. Θεωρητικό υπόβαθρο της Γραμμικής Παλινδρόμησης.....	72
3.2.2. Εκτίμηση των παραμέτρων $b_0, b_1, \dots, b_{p-1}$ και σ^2	73
3.2.3. Έλεγχοι υποθέσεων και διαστήματα εμπιστοσύνης για τις παραμέτρους του μοντέλου.....	75
3.2.4. Ερμηνεία της συνολικής μεταβλητότητας του μοντέλου.....	76
3.2.5. Σημειακή και Αναμενόμενη Μέση πρόβλεψη της Y	77

3.2.6. Εξέταση της ορθότητας του μοντέλου	78
3.2.7. Πολυσυγγραμμικότητα.....	79
3.2.8. Χρήση ψευδομεταβλητών ή εικονικών μεταβλητών.....	79
3.3. Ανάλυση Διασποράς.....	80
3.3.1. Ανάλυση Διασποράς κατά έναν Παράγοντα.....	80
3.3.1.1. Προσδιορισμός του μοντέλου.....	81
3.3.1.2. Υποθέσεις του μοντέλου.....	81
3.3.1.3. Έλεγχος υποθέσεων.....	82
3.3.1.4. Υπολογισμός των αθροισμάτων των τετραγώνων.....	82
3.3.1.5. Τεστ F.....	83
3.3.1.6. Χρήση ψευδομεταβλητών.....	84
3.3.2. Ανάλυση Διασποράς κατά δύο Παράγοντες.....	84
3.3.2.1. Προσδιορισμός του μοντέλου.....	85
3.3.2.2. Υποθέσεις του μοντέλου.....	85
3.3.2.3. Έλεγχος υποθέσεων.....	86
3.3.2.4. Υπολογισμός των αθροισμάτων των τετραγώνων.....	86
3.4. Πολυμεταβλητή Ανάλυση Διασποράς (MANOVA).....	87
3.4.1. Σύγκριση MANOVA και ANOVA.....	88
3.4.2. Έλεγχοι υποθέσεων MANOVA.....	88
3.4.3. Άλλα πολυμεταβλητά κριτήρια.....	90
4 ΕΦΑΡΜΟΓΗ ΤΩΝ ΣΤΑΤΙΣΤΙΚΩΝ ΜΟΝΤΕΛΩΝ ΠΑΛΙΝΔΡΟΜΗΣΗΣ ΜΕ ΤΑ ΔΕΔΟΜΕΝΑ ΤΟΥ ΔΕΙΓΜΑΤΟΣ.....	91
4.1. Εφαρμογή της Ανάλυση Διασποράς (ANOVA).....	91
4.1.1. Επίδραση της περιοχής στη συνολική καταναλωτική δαπάνη των νοικοκυριών.....	91
4.1.2. Επίδραση της κύριας πηγής εισοδήματος στη συνολική καταναλωτική δαπάνη των νοικοκυριών.....	98
4.1.3. Επίδραση του αριθμού των οικονομικά ενεργών μελών στη συνολική καταναλωτική δαπάνη των νοικοκυριών.....	107
4.1.4. Επίδραση της κοινωνικο – οικονομικής κατάστασης του προσώπου αναφοράς στη συνολική καταναλωτική δαπάνη των νοικοκυριών.....	115
4.2. Εφαρμογή της Πολυμεταβλητής Ανάλυσης Διασποράς.....	123
4.2.1. Επίδραση της περιοχής στις επιμέρους κατηγορίες δαπανών των νοικοκυριών.....	123
4.2.2. Επίδραση της κύριας πηγής εισοδήματος στις επιμέρους κατηγορίες των δαπανών των νοικοκυριών.....	128
4.2.3. Επίδραση του αριθμού των οικονομικά ενεργών μελών στις επιμέρους κατηγορίες των δαπανών των νοικοκυριών.....	134
4.2.4. Επίδραση της κοινωνικο – οικονομικής κατάστασης του προσώπου αναφοράς στις επιμέρους κατηγορίες των δαπανών των νοικοκυριών.....	139
4.3. Εφαρμογή της Γραμμικής Παλινδρόμησης.....	145
4.3.1. Απλή Γραμμική Παλινδρόμηση.....	145
4.3.2. Πολλαπλή Γραμμική Παλινδρόμηση.....	155
5 ΑΝΑΛΥΣΗ ΚΥΡΙΩΝ ΣΥΝΙΣΤΩΣΩΝ.....	169
5.1 Εισαγωγή.....	169
5.2 Η βασική ιδέα της μεθόδου.....	170
5.3 Κύριες Συνιστώσες από τον πίνακα συνδιασποράς Σ.....	170

5.4 Κύριες Συνιστώσες από τον πίνακα συσχέτισης ρ	174
5.5 Βήματα της Ανάλυσης σε Κύριες Συνιστώσες.....	176
5.5.1 Έλεγχος συσχετίσεων.....	176
5.5.2 Επιλογή κατάλληλου πίνακα.....	177
5.5.3 Υπολογισμός ιδιοτιμών και ιδιοδιανυσμάτων.....	177
5.5.4 Απόφαση για το πλήθος των Κύριων Συνιστωσών.....	177
5.5.5 Εύρεση των κυρίων συνιστωσών.....	180
5.5.6 Ερμηνεία των κυρίων συνιστωσών.....	180
5.5.7 Δημιουργία νέων μεταβλητών.....	181
5.6 Χρήσιμα αποτελέσματα.....	181
5.7 Εφαρμογές της Ανάλυσης των Κυρίων Συνιστωσών.....	182
6 ΕΦΑΡΜΟΓΗ ΤΗΣ ΑΝΑΛΥΣΗ ΚΥΡΙΩΝ ΣΥΝΙΣΤΩΣΩΝ ΜΕ ΤΑ ΔΕΔΟΜΕΝΑ ΤΟΥ ΔΕΙΓΜΑΤΟΣ.....	183
6.1 Εύρεση Κυρίων Συνιστωσών.....	183
6.2 Αριθμός Κυρίων Συνιστωσών.....	187
6.3 Ερμηνεία κυρίων Συνιστωσών.....	188
6.4 Περαιτέρω χρήση των αποτελεσμάτων της Ανάλυσης των Κυρίων Συνιστωσών.....	188
ΒΙΒΛΙΟΓΡΑΦΙΑ.....	195

ΠΡΟΛΟΓΟΣ

Η πολυμεταβλητή στατιστική ανάλυση αποτελεί ένα πολύτιμο εργαλείο σε πολλές επιστήμες. Οι μέθοδοι της πολυμεταβλητής ανάλυσης αναφέρονται σε διαδικασίες και μεθοδολογίες που χρησιμοποιούνται προκειμένου να πραγματοποιηθεί στατιστική συμπερασματολογία με τη χρήση πολλών μεταβλητών. Οι μέθοδοι αυτοί, είναι ιδιαίτερα χρήσιμες, γιατί βοηθούν στην εξαγωγή συμπερασμάτων από ένα πλήθος δεδομένων και στον περιορισμό της αβεβαιότητας, καθώς και στην εύρεση και την ερμηνεία συσχετίσεων μεταξύ των μεταβλητών.

Η παρούσα διπλωματική εργασία αφορά στην εφαρμογή ποσοτικών μεθόδων της πολυμεταβλητής στατιστικής ανάλυσης σε μεγάλο αριθμό δεδομένων από μια έρευνα που διεξάγεται ετησίως από την Ελληνική Στατιστική Υπηρεσία, την Έρευνα Οικογενειακών Προϋπολογισμών (ΕΟΠ). Σκοπός της εν λόγω εργασίας είναι η διερεύνηση του τρόπου που δαπανάται το οικογενειακό εισόδημα των νοικοκυριών στις 12 κύριες κατηγορίες δαπανών στην Ελλάδα, οι οποίες είναι είδη διατροφής, οισοπνευματώδη ποτά – καπνός, είδη ένδυσης και υπόδησης, στέγαση, διαρκή αγαθά, υγεία, μεταφορές, επικοινωνίες, αναψυχή – πολιτισμός, εκπαίδευση, ξενοδοχεία – καφενεία – εστιατόρια και διάφορα άλλα αγαθά και υπηρεσίες. Επιπρόσθετα, η μελέτη τόσο της συνολικής καταναλωτικής δαπάνης όσο και των επιμέρους δαπανών των νοικοκυριών σε σχέση με το εισόδημά τους καθώς και με κάποια κοινωνικά και δημογραφικά χαρακτηριστικά τους.

Στο πρώτο κεφάλαιο παρουσιάζεται η Έρευνα Οικογενειακών Προϋπολογισμών.

Στο δεύτερο κεφάλαιο γίνεται η περιγραφική ανάλυση των μεταβλητών με κατάλληλες αριθμητικές και γραφικές μεθόδους και ελέγχονται πιθανές συσχετίσεις μεταξύ των μεταβλητών αυτών.

Στο τρίτο κεφάλαιο αναλύονται θεωρητικά στατιστικά μοντέλα παλινδρόμησης και συγκεκριμένα, η γραμμική παλινδρόμηση και η ανάλυση διασποράς.

Στο τέταρτο κεφάλαιο πραγματοποιείται ανάλυση διασποράς για τη συνολική καταναλωτική δαπάνη των νοικοκυριών αλλά και τις 12 επιμέρους κατηγορίες δαπανών ως προς την περιοχή διαμονής, την κύρια πηγή εισοδήματος, τον αριθμό των οικονομικά ενεργών μελών και της κοινωνικο – οικονομικής κατάστασης του προσώπου αναφοράς του νοικοκυριού. Επίσης, εφαρμόζεται η απλή γραμμική παλινδρόμηση για τη συνολική καταναλωτική δαπάνη και το συνολικό εισόδημα των νοικοκυριών και η πολλαπλή γραμμική παλινδρόμηση για τη συνολική καταναλωτική δαπάνη σε σχέση με το εισόδημα, την περιοχή, την κύρια πηγή εισοδήματος, τον αριθμό οικονομικά ενεργών μελών και την κοινωνικο – οικονομική κατάσταση του προσώπου αναφοράς του νοικοκυριού.

Στο πέμπτο κεφάλαιο παρουσιάζεται θεωρητικά η ανάλυση σε κύριες συνιστώσες.

Και στο έκτο κεφάλαιο εφαρμόζεται στα δεδομένα του δείγματος, έτσι ώστε να εξεταστεί πιθανή μείωση των 12 μεταβλητών που αφορούν στις επιμέρους κατηγορίες δαπανών του νοικοκυριού.

Η ανάλυση των δεδομένων έγινε με χρήση του στατιστικού πακέτου R.

Θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή μου, κ. Δ. Φουσκάκη, Επίκουρο Καθηγητή της Σχολής Εφαρμοσμένων Μαθηματικών και Φυσικών Επιστημών του Εθνικού Μετσοβίου Πολυτεχνείου, για την ανάθεση της εργασίας, τις λεπτομερείς διορθώσεις και την καθοδήγηση κατά τη διάρκεια της παρούσας διπλωματικής εργασίας, που ήταν πάντοτε πρόθυμος να απαντήσει στις ερωτήσεις μου.

Τέλος, θα ήθελα να ευχαριστήσω την οικογένειά μου για την κατανόηση και την υποστήριξη σε όλες τις φάσεις της υλοποίησης αυτής της διπλωματικής εργασίας.

ΚΕΦΑΛΑΙΟ 1

ΕΙΣΑΓΩΓΗ

Σε αυτό το εισαγωγικό κεφάλαιο θα εξηγηθεί αναλυτικά η προέλευση των δεδομένων για την εν λόγω διπλωματική εργασία, η ταυτότητα, ο σκοπός και ο σχεδιασμός της έρευνας που χρησιμοποιήθηκε.

1.1. Τι είναι η Έρευνα Οικογενειακών Προϋπολογισμών

Η «Έρευνα Οικογενειακών Προϋπολογισμών» (Household Budget Survey) είναι μία στατιστική έρευνα με την οποία συγκεντρώνονται πληροφορίες από ένα αντιπροσωπευτικό δείγμα νοικοκυριών, για τη σύνθεσή τους, την απασχόληση των μελών τους, τις συνθήκες στέγασης και, κυρίως, για τις δαπάνες διαβίωσής τους, καθώς και για τα εισοδήματά τους. Οι πληροφορίες για τις δαπάνες που συγκεντρώνονται από τα νοικοκυριά είναι πολύ αναλυτικές. Δεν συγκεντρώνονται, δηλαδή, πληροφορίες για τις κατηγορίες δαπανών συνολικά, όπως «δαπάνες διατροφής», «είδη ένδυσης – υπόδησης», «δαπάνες για υγεία» κλπ., αλλά για καθεμιά δαπάνη ξεχωριστά, π.χ. ψωμί άσπρο, γάλα νωπό πλήρες, μοσχάρι νωπό κλπ., υποδήματα ανδρικά, γυναικεία κλπ., ή μικροβιολογικές εξετάσεις, φάρμακα κλπ.

Η Έρευνα Οικογενειακών Προϋπολογισμών (ΕΟΠ) είναι η πρώτη δειγματοληπτική έρευνα νοικοκυριών που διενεργήθηκε από την Εθνική Στατιστική Υπηρεσία της Ελλάδας. Η πρώτη ΕΟΠ στην Ελλάδα διενεργήθηκε κατά τα έτη 1957/58, είχε διάρκεια ένα χρόνο και το δείγμα ανήλθε σε 2.500 περίπου νοικοκυριά των αστικών περιοχών της χώρας. Η έρευνα αυτή συνεχίστηκε κατά τα επόμενα έτη και μέχρι το έτος 1972 μόνο σε δήμους άνω των 30.000 κατοίκων, σε μικρότερο όμως δείγμα νοικοκυριών.

Τον Απρίλιο του έτους 1963, παράλληλα με την έρευνα στις αστικές περιοχές, άρχισε ευρείας έκτασης έρευνα στις ημιαστικές και στις αγροτικές περιοχές της χώρας, δηλαδή σε δήμους και κοινότητες με πληθυσμό κάτω των 10.000 κατοίκων, διήρκεσε ένα χρόνο, περιελήφθησαν 3.755 νοικοκυριά των περιοχών αυτών και συνεχίστηκε μέχρι το έτος 1972, σε μικρότερο όμως δείγμα νοικοκυριών. Κατά τα έτη 1974, 1981/82, 1987/88, 1993/94, 1998/99, 2004/2005 πραγματοποιήθηκαν έρευνες Οικογενειακών Προϋπολογισμών, οι οποίες κάλυψαν όλες τις περιοχές της χώρας, σε δείγμα περίπου 7.500 νοικοκυριών για την πρώτη και σε δείγμα 6.000 έως 6.800 νοικοκυριών για τις επόμενες πέντε και είχαν διάρκεια ένα έτος.

Από το έτος 2008 αποφασίστηκε, για εθνικές ανάγκες (κατάρτιση Δείκτη Τιμών Καταναλωτή με μεγαλύτερη αξιοπιστία, παραγωγή συγκρίσιμων στατιστικών για τις ανάγκες των Εθνικών Λογαριασμών), η διενέργεια της έρευνας να είναι ετήσια και συνεχής, δηλαδή να έχει διάρκεια ένα χρόνο και να πραγματοποιείται κάθε χρόνο. Η συγκρισιμότητα των στοιχείων επιβάλλει σε όλες τις χώρες μέλη, κατά τον σχεδιασμό και την κατάρτιση των ερωτηματολογίων, να λαμβάνουν υπ' όψιν τις συστάσεις της Eurostat για την εναρμόνιση των ερευνών της, καθώς και την

ταξινόμηση των αγαθών και υπηρεσιών που έχει προταθεί από τη Eurostat, αφού προσαρμοστούν στις ιδιαιτερότητες της κάθε χώρας.

1.2. Σκοπός της έρευνας

Με την έρευνα αυτή συγκεντρώνονται αναλυτικές πληροφορίες για την αξία των αγορών και των σε είδος απολαβών των νοικοκυριών, καθώς και για διάφορα χαρακτηριστικά των νοικοκυριών και των κατοικιών τους. Τα στοιχεία που προκύπτουν από την έρευνα χρησιμοποιούνται για τους παρακάτω αντικειμενικούς σκοπούς:

- την αναθεώρηση του Δείκτη Τιμών Καταναλωτή που καταρτίζεται από την Εθνική Στατιστική Υπηρεσία.
- τη συμπλήρωση των διαθέσιμων στατιστικών στοιχείων για την εκτίμηση της συνολικής ιδιωτικής κατανάλωσης.
- την ανάλυση των μεταβολών του επιπέδου διαβίωσης των νοικοκυριών σε σχέση με τις προηγούμενες έρευνες.
- τη μελέτη της σχέσης μεταξύ των αγορών και των σε είδος απολαβών των νοικοκυριών.
- τη μελέτη των ορίων χαμηλού εισοδήματος κατά διάφορες κοινωνικο – οικονομικές κατηγορίες και ομάδες του πληθυσμού.
- τη μελέτη των αλλαγών στις διατροφικές συνήθειες των νοικοκυριών της χώρας.

Τα αποτελέσματα της έρευνας σε μορφή πινάκων δημοσιεύονται στη χώρα μας και, παράλληλα, υπάρχουν και σε δημοσίευμα της Eurostat, στο οποίο περιλαμβάνονται συγκριτικά στοιχεία από τις αντίστοιχες έρευνες όλων των κρατών μελών της Ευρωπαϊκής Ένωσης.

1.3. Σχεδιασμός της έρευνας

Η έρευνα είναι δειγματοληπτική, με «περιστροφικό ενιαίο σχεδιασμό» (rotational integrated design), ο οποίος επιλέχθηκε ως ο πλέον κατάλληλος για ενιαία συγχρονική και διαχρονική έρευνα. Η τελική δειγματοληπτική μονάδα είναι το νοικοκυριό και οι μονάδες ανάλυσης τα νοικοκυριά και τα μέλη τους. Η ΕΟΠ βασίζεται σε δισταδιακή στρωματοποιημένη δειγματοληψία νοικοκυριών από πλαίσιο δειγματοληψίας, το οποίο έχει δημιουργηθεί βάσει των στοιχείων του μόνιμου πληθυσμού της Απογραφής 2001 και καλύπτει πλήρως τον πληθυσμό αναφοράς, έτσι ώστε να εξασφαλίζεται η αντιπροσωπευτικότητα του δείγματος.

Ο σχεδιασμός της δειγματοληψίας περιλαμβάνει δύο κριτήρια στρωμάτωσης:

- Το πρώτο κριτήριο στρωμάτωσης του ερευνώμενου πληθυσμού είναι η γεωγραφική διαίρεση της χώρας. Ως μείζονα στρώματα χρησιμοποιήθηκαν οι 13 περιφέρειες (σύμφωνα με την κατηγοριοποίηση της Ευρωπαϊκής Ένωσης NUTS II), η πρώην Περιφέρεια της Πρωτεύουσας και το Πολεοδομικό Συγκρότημα της Θεσσαλονίκης.

- Το δεύτερο κριτήριο στρωμάτωσης είναι ο βαθμός αστικότητας. Σε κάθε περιφέρεια (NUTS II), τα νοικοκυριά κατανεμήθηκαν ανάλογα με το βαθμό αστικότητας σε 3 κατηγορίες: τοπικά διαμερίσματα με πληθυσμό πάνω από 10.000 κατοίκους, από 2.000 έως 9.999 κατοίκους και μέχρι 1.999 κατοίκους. Η στρωμάτωση κατά βαθμό αστικότητας έγινε σύμφωνα με τον πληθυσμό των τοπικών διαμερισμάτων, εκτός των πολεοδομικών συγκροτημάτων των δύο μεγάλων πόλεων (πρώην Περιφέρεια της Πρωτεύουσας και Πολεοδομικό Συγκρότημα της Θεσσαλονίκης).

Η πρώην Περιφέρεια της Πρωτεύουσας χωρίστηκε σε 31 στρώματα ίσου, περίπου, μεγέθους (ίσους αριθμούς νοικοκυριών), με βάση τους καταλόγους – πλαίσια με τα οικοδομικά τετράγωνα των Δήμων και σύμφωνα με κοινωνικοοικονομικά κριτήρια. Ομοίως, το πρώην Πολεοδομικό Συγκρότημα της Θεσσαλονίκης χωρίστηκε σε 9 ίσου μεγέθους στρώματα. Τα πρώην Πολεοδομικά Συγκροτήματα αυτών των δύο μεγάλων πόλεων αποτελούν, περίπου, το 40% του συνολικού πληθυσμού και κατέχουν ακόμα υψηλότερο ποσοστό σε κάποιες κοινωνικοοικονομικές μεταβλητές. Ο αριθμός των στρωμάτων, ο οποίος προέκυψε από την εφαρμογή των δύο κριτηρίων στρωμάτωσης του πληθυσμού, ανέρχεται σε 79.

- 1^ο στάδιο δειγματοληψίας:
Στο πρώτο στάδιο δειγματοληψίας, επιλέγονται από κάθε στρώμα οι πρωτογενείς μονάδες (επιφάνειες) με πιθανότητα επιλογής ανάλογη του μεγέθους τους (πλήθος νοικοκυριών, σύμφωνα με την Απογραφή Πληθυσμού έτους 2001).
- 2^ο στάδιο δειγματοληψίας:
Στο δεύτερο στάδιο δειγματοληψίας, σε κάθε επιλεγείσα πρωτογενή δειγματοληπτική μονάδα (μονάδα επιφάνειας), από τον ενημερωμένο κατάλογο – πλαίσιο των νοικοκυριών επιλέγεται το δείγμα νοικοκυριών με ίσες πιθανότητες και με την εφαρμογή της συστηματικής δειγματοληψίας. Στην πραγματικότητα, στο δεύτερο στάδιο επιλέγεται δείγμα κατοικιών. Εντούτοις, στις περισσότερες περιπτώσεις, υπάρχει μια αντιστοίχιση μεταξύ νοικοκυριού και κατοικίας. Εάν η επιλεγείσα κατοικία αποτελείται από περισσότερα του ενός νοικοκυριά, τότε ερευνώνται όλα τα νοικοκυριά.

1.4. Μέγεθος του δείγματος

Η Έρευνα Οικογενειακών Προϋπολογισμών (ΕΟΠ) του έτους 2010, από την οποία προέρχονται και τα δεδομένα που έχουν χρησιμοποιηθεί στην παρούσα διπλωματική εργασία, είναι η ενδέκατη κατά σειρά έρευνα που έχει διενεργηθεί στην Ελλάδα. Το μέγεθος του δείγματος για την εν λόγω έρευνα τελικά ανήλθε στα 3.512 νοικοκυριά (κλάσμα δειγματοληψίας 0,08%), το οποίο ισοκατανεμήθηκε μέσα στο έτος, έτσι ώστε να επιλεγούν 4 ισοδύναμα ανεξάρτητα δείγματα, τα οποία αντιστοιχούν στα 4 τρίμηνα του έτους. Ο συνολικός αριθμός των δήμων και των κοινοτήτων που επιλέχθηκαν στην έρευνα ανήλθε σε 337, των οικισμών σε 360 ενώ ο αριθμός των μονάδων επιφάνειας ανήλθε σε 612.

Η έρευνα καλύπτει όλα τα ιδιωτικά νοικοκυριά της χώρας με τα μέλη τους, ανεξάρτητα από το μέγεθος ή οποιαδήποτε οικονομικά και κοινωνικά χαρακτηριστικά τους.

Ως νοικοκυριό λαμβάνεται ένα άτομο που ζει μόνο του σε μία κατοικία ή μια ομάδα ατόμων συγγενικών ή μη τα οποία διαμένουν στην ίδια κατοικία και εξασφαλίζουν από κοινό προϋπολογισμό τη διατροφή τους και άλλα είδη που είναι απαραίτητα για τη διαβίωσή τους.

Άτομα που διαμένουν στην ίδια κατοικία, τα οποία όμως δεν προμηθεύονται από κοινού τα απαραίτητα για τη συντήρησή τους, ούτε τρώνε μαζί, θεωρούνται ξεχωριστά νοικοκυριά, με την προϋπόθεση της αποκλειστικής χρήσης ενός, τουλάχιστον, δωματίου. Σε αντίθετη περίπτωση αποτελούν ένα νοικοκυριό. Πιο συγκεκριμένα:

Μονοπρόσωπα νοικοκυριά θεωρούνται:

- Ένα άτομο που ζει σε μία κατοικία ή σε ένα δωμάτιο κατοικίας και προμηθεύεται μόνο του τα απαραίτητα είδη για τη συντήρησή του.
- Δύο ή περισσότερα άτομα συγγενικά ή μη τα οποία διαμένουν σε μία κατοικία, αλλά δεν προμηθεύονται από κοινού τα απαραίτητα για τη συντήρησή τους, το καθένα από αυτά έχει την αποκλειστική χρήση ενός, τουλάχιστον, δωματίου και δεν χρησιμοποιούν κάποιον κύριο κοινό χώρο (καθιστικό, γραφείο – όχι όμως χωλ – κλπ.).

Πολυπρόσωπα νοικοκυριά θεωρούνται:

- Ένα ζευγάρι μόνο του ή γονείς με τα παιδιά τους ή ένας από τους γονείς με τα παιδιά του.
- Ένα ζευγάρι με ή χωρίς παιδιά, οι γονείς του ζευγαριού και η εσωτερική οικιακή βοηθός, αν υπάρχει και προτίθεται να παραμείνει για χρονικό διάστημα μεγαλύτερο των έξι μηνών.
- Δύο ή περισσότερα άτομα, όχι απαραίτητα συγγενικά μεταξύ τους, τα οποία διαμένουν σε μία κατοικία και προμηθεύονται από κοινού τα είδη που είναι απαραίτητα για τη διαβίωσή τους.
- Μία οικογένεια με έναν έως πέντε – το πολύ – οικότροφους.

Εξαιρούνται από την έρευνα:

- Οι συλλογικές κατοικίες, όπως π.χ. ξενοδοχεία, πανσιόν, νοσοκομεία, γηροκομεία, στρατόπεδα, αναμορφωτήρια κλπ. Συλλογικές κατοικίες θα θεωρηθούν και τα νοικοκυριά που παρέχουν στέγη με διατροφή σε πέντε τρόφιμους και άνω.
- Τα νοικοκυριά με μέλη ξένους υπηκόους που υπηρετούν σε ξένες διπλωματικές αποστολές.

1.5. Χρόνος διενέργειας της έρευνας και περίοδοι αναφοράς των δαπανών και των εισοδημάτων

Η συλλογή των στοιχείων της έρευνας πραγματοποιήθηκε κατά το χρονικό διάστημα Ιανουάριος – Δεκέμβριος 2010.

Περίοδοι αναφοράς των δαπανών και των εισοδημάτων είναι τα χρονικά διαστήματα, με συγκεκριμένη ημέρα έναρξης και λήξης, στα οποία αναφέρονται οι δαπάνες και τα εισοδήματα του νοικοκυριού. Στην εν λόγω έρευνα έχουν χρησιμοποιηθεί διάφορες περίοδοι αναφοράς που έχουν καθοριστεί, ανάλογα με τη συχνότητα που πραγματοποιούνται από τα νοικοκυριά τα διάφορα είδη δαπανών ή λαμβάνονται τα εισοδήματα, με στόχο τη μείωση των δειγματοληπτικών σφαλμάτων μνήμης.

Περίοδοι αναφοράς θεωρούνται:

- Οι δεκατέσσερις ημέρες της έρευνας για τις καθημερινές δαπάνες του νοικοκυριού, όπως είδη διατροφής και οινοπνευματώδη ποτά, είδη καθαριότητας, φαρμακευτικά είδη, προσωπικά έξοδα των μελών του νοικοκυριού καθώς και οι δαπάνες σε εστιατόρια, ταβέρνες, καφεενεία κλπ.
- Ο μήνας, το δίμηνο, το τρίμηνο, το τετράμηνο, το εξάμηνο και το έτος για τις δαπάνες που πληρώνονται σε τακτά χρονικά διαστήματα για υπηρεσίες που παρασχέθηκαν ή θα παρασχεθούν στο μέλλον, όπως λογαριασμοί ηλεκτρικού, νερού, τηλεφώνου, κοινοχρήστων, τέλη κυκλοφορίας και ασφάλιστρα αυτοκινήτων, ενοίκιο κύριας ή δευτερεύουσας κατοικίας κλπ.
- Οι τελευταίες τριάντα ημέρες, οι τελευταίοι τρεις, έξι ή δώδεκα μήνες πριν τη λήξη της έρευνας στο νοικοκυριό για τις δαπάνες επίπλων, ειδών ένδυσης και υπόδησης, υγείας και εκπαίδευσης, διακοπών και παραθερισμού καθώς και τα εισοδήματα από μισθωτές υπηρεσίες, ελεύθερα επαγγέλματα, γεωργικές εκμεταλλεύσεις, συντάξεις, κοινωνικά επιδόματα κλπ.

ΚΕΦΑΛΑΙΟ 2

ΑΝΑΛΥΣΗ ΜΕΤΑΒΛΗΤΩΝ

2.1. Περιγραφικά στοιχεία μεταβλητών

Σκοπός της περιγραφικής ανάλυσης είναι να δώσει μια συνοπτική παρουσίαση των τιμών του δείγματος, καθώς επίσης και να ελέγξει την ορθότητα αυτών. Αυτό γίνεται με τη βοήθεια περιγραφικών πινάκων και διαγραμμάτων. Η επιλογή των κατάλληλων αριθμητικών και γραφικών μεθόδων εξαρτάται από τον τύπο της μεταβλητής που εξετάζεται.

Αρχικά, τα δεδομένα ήταν στο αρχείο SPSS HBS_HH_BASIC_2010_GR.sav. Με το πακέτο foreign στην R γίνεται η ανάγνωση του αρχείου με την εντολή read.spss() και η αποθήκευσή του σε έναν πίνακα με το όνομα file, η μετατροπή του πίνακα αυτού σε πλαίσιο δεδομένων με την εντολή as.data.frame() και η αποθήκευσή του σε μορφή αρχείου .txt με την εντολή write.table() με το όνομα HH_basic.txt.

```
> library (foreign)
> file <- read.spss ('HBS_HH_BASIC_2010_GR.sav')
> file <- as.data.frame (file)
> write.table (file, file='HH_basic.txt')
```

Στη συνέχεια, «διαβάζεται» το αρχείο στην R με την εντολή read.table() και αποθηκεύεται ως πλαίσιο δεδομένων με το όνομα HH_basic. Από το αρχείο αυτό, το οποίο αφορά στη δαπάνη των νοικοκυριών συνολικά ανά κατηγορία, επιλέγονται οι μεταβλητές που θα χρησιμοποιηθούν για περαιτέρω ανάλυση, δημιουργείται ένα καινούριο πλαίσιο δεδομένων με το όνομα dataHH με την εντολή data.frame() και δίνονται χαρακτηριστικά ονόματα στις μεταβλητές αυτές με την εντολή names().

```
> HH_basic <- read.table ('HH_basic.txt')
> dataHH <- data.frame (HH_basic$HA09, HH_basic$HI11, HH_basic$HD20,
HH_basic$HC23, HH_basic$HH099, HH_basic$HA04, HH_basic$HA08,
HH_basic$HE00, HH_basic$HE01, HH_basic$HE02, HH_basic$HE03, HH_basic$HE04,
HH_basic$HE05, HH_basic$HE06, HH_basic$HE07, HH_basic$HE08, HH_basic$HE09,
HH_basic$HE10, HH_basic$HE11, HH_basic$HE12)
> names (dataHH) <- c ('density', 'source', 'members', 'situation', 'income',
'household', 'region', 'total', 'food', 'alcohol', 'clothing', 'housing', 'equipment',
'health', 'transport', 'communication', 'culture', 'education', 'hotels', 'goods')
> write.table (dataHH, file = 'dataHH.txt')
```

Στη συνέχεια, παρουσιάζονται σε πίνακες τα περιγραφικά στοιχεία για καθεμιά μεταβλητή ξεχωριστά σε συνδυασμό με τα αντίστοιχα διαγράμματα, δηλαδή πραγματοποιείται η περιγραφική και μονομεταβλητή ανάλυση των δεδομένων.

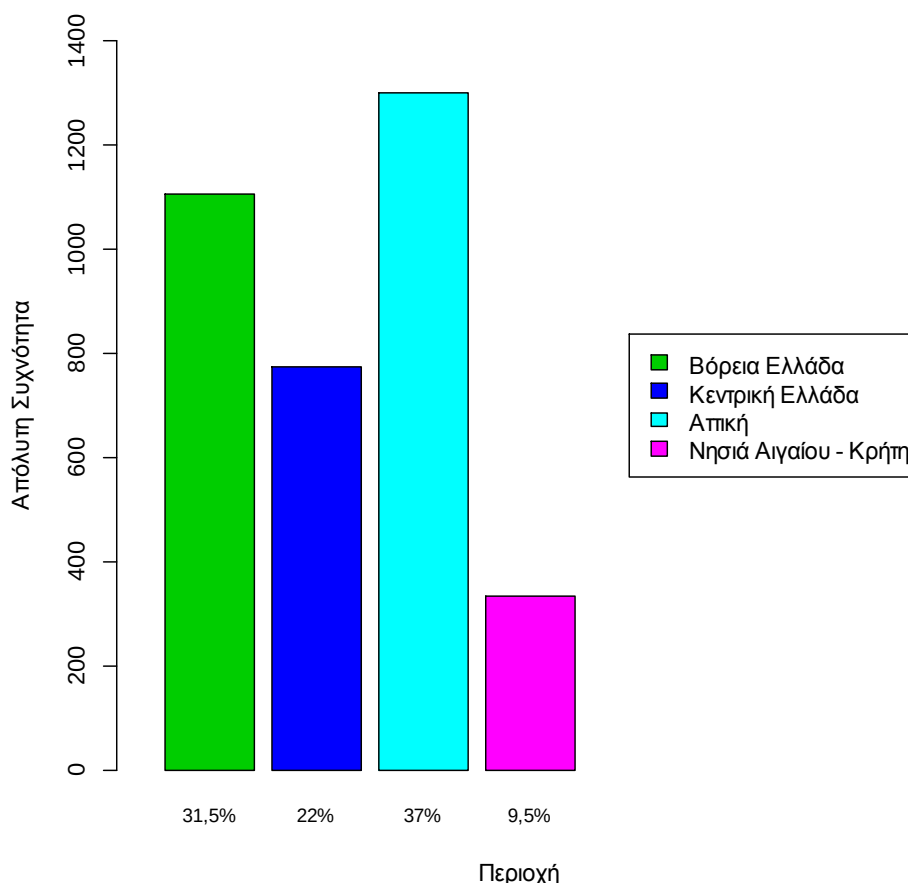
2.1.1. Περιοχή

Το δείγμα, όσον αφορά στην περιοχή, είναι χωρισμένο στις 4 περιφέρειες της Ελλάδας σύμφωνα με την κατηγοριοποίηση της Ευρωπαϊκής Ένωσης (NUTS1): Βόρεια Ελλάδα, Κεντρική Ελλάδα, Αττική και Νησιά Αιγαίου – Κρήτη.

Πίνακας 2.1: Πίνακας συχνοτήτων για τη μεταβλητή περιοχή

Περιοχή	Απόλυτη Συχνότητα	Ποσοστιαία Συχνότητα (%)
Βόρεια Ελλάδα	1105	31.5
Κεντρική Ελλάδα	774	22
Αττική	1300	37
Νησιά Αιγαίου – Κρήτη	333	9.5
Σύνολο	3512	100.0

Διάγραμμα 2.1: Ραβδόγραμμα για τη μεταβλητή περιοχή



Από τον Πίνακα 2.1 αλλά και το Διάγραμμα 2.1 προκύπτει ότι το μεγαλύτερο ποσοστό των νοικοκυριών του δείγματος προέρχεται από την Αττική (37%). Ένα εξίσου σημαντικό ποσοστό προέρχεται από τη Βόρεια Ελλάδα (31.5%), εφ'όσον θεωρείται μείζον στρώμα το Πολεοδομικό Συγκρότημα της Θεσσαλονίκης. Από την Κεντρική Ελλάδα προέρχεται ένα αρκετά μικρότερο ποσοστό (22%) και τέλος, από τα Νησιά Αιγαίου και την Κρήτη μόνο το 1/10 περίπου του συνολικού δείγματος (9.5%).

Εντολές στην R

Η μεταβλητή “region” είναι μια κατηγορική μεταβλητή, η οποία λαμβάνει τις τιμές 1 για την Βόρεια Ελλάδα, 2 για την Κεντρική Ελλάδα, 3 για την Αττική και 4 για τα Νησιά Αιγαίου – Κρήτη.

Η μεταβλητή “region” μετατρέπεται σε κατηγορική με την εντολή factor(), οι απόλυτες συχνότητες κάθε κατηγορίας υπολογίζονται με την εντολή table() καθώς και τα αντίστοιχα ποσοστά με την εντολή round(). Τέλος, η γραφική μέθοδος αναπαράστασης των τιμών για την εν λόγω μεταβλητή είναι το ραβδόγραμμα το οποίο κατασκευάζεται με την εντολή barplot() και η προσθήκη λεζάντας γίνεται με την εντολή legend().

```
> region <- dataHH$region
> region <- factor(region)
> table (region)
region
 1  2  3  4
1105 774 1300 333
> region_perc <- round (prop.table (table (region))*100, 1)
> region_perc
region
 1  2  3  4
31.5 22.0 37.0 9.5
> barplot (table (region), names.arg = c ('31,5%', '22%', '37%', '9,5%'), col = c (3, 4, 5, 6), cex.names = 0.8, ylim = c (0, 1400), xlim = c (0, 5), width = 0.6, xlab = 'Περιοχή', ylab = 'Απόλυτη Συχνότητα')
> legend ('right', fill = c (3, 4, 5, 6), c ('Βόρεια Ελλάδα', 'Κεντρική Ελλάδα', 'Αττική', 'Νησιά Αιγαίου - Κρήτη'), cex = 1)
```

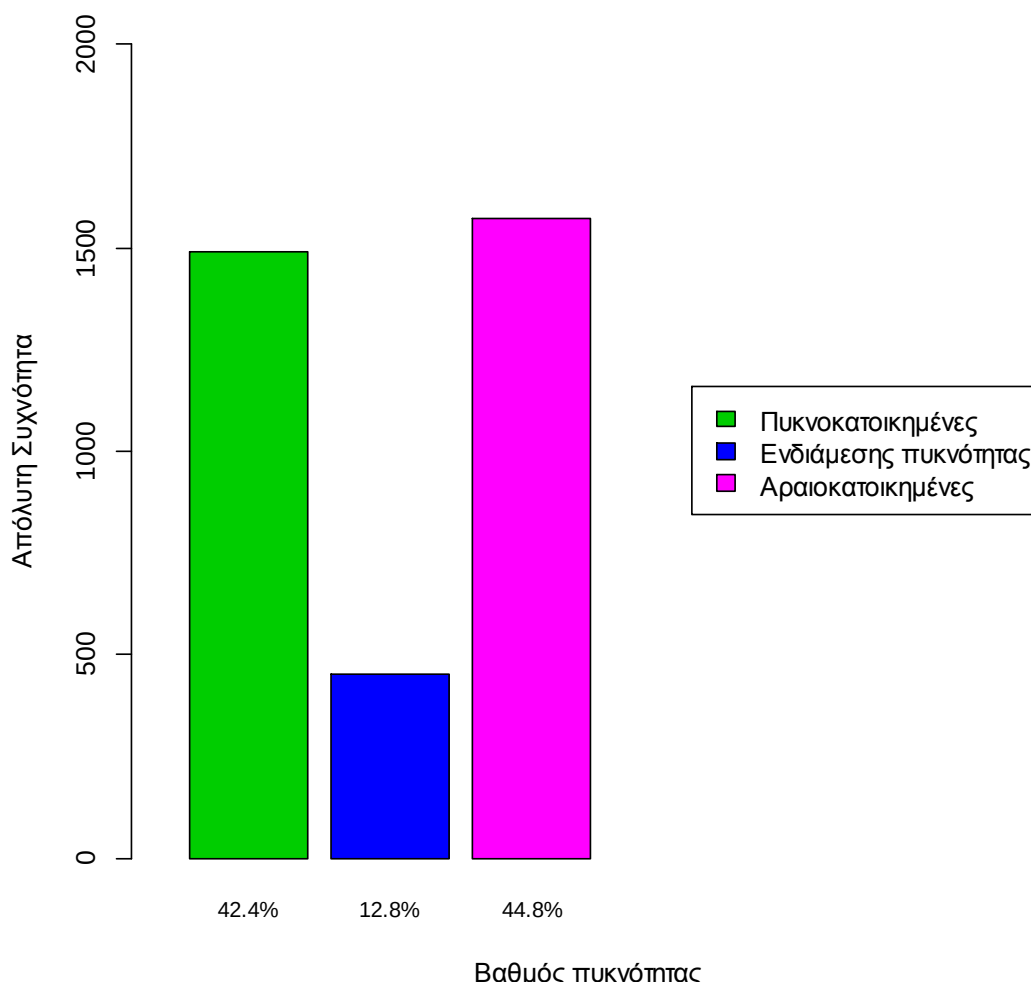
2.1.2. Βαθμός πυκνότητας της περιοχής

Οι περιοχές του δείγματος, όσον αφορά στον βαθμό πυκνότητας, κατατάσσονται σε πυκνοκατοικημένες περιοχές (≥ 500 κάτοικοι/km²), σε περιοχές ενδιάμεσης πυκνότητας (100–499 κάτοικοι/km²) και σε αραιοκατοικημένες περιοχές (≤ 100 /km²).

Πίνακας 2.2: Πίνακας συχνοτήτων για τη μεταβλητή βαθμός πυκνότητας

Βαθμός Πυκνότητας	Απόλυτη Συχνότητα	Ποσοστιαία Συχνότητα (%)
Πυκνοκατοικημένες	1489	42.4
Ενδιάμεσης πυκνότητας	451	12.8
Αραιοκατοικημένες	1572	44.8
Σύνολο	3512	100.0

Διάγραμμα 2.2: Ραβδόγραμμα για τη μεταβλητή βαθμός πυκνότητας



Από τον Πίνακα 2.2 αλλά και το Διάγραμμα 2.2 προκύπτει ότι το μεγαλύτερο ποσοστό των περιοχών είναι αραιοκατοικημένες (44.8%) λόγω της μεγάλης κάλυψης της έρευνας στην επαρχία και, με μικρή διαφορά, πυκνοκατοικημένες (42.4%) που οφείλεται κυρίως στα δύο μεγάλα αστικά κέντρα της περιφέρειας της Αττικής και της Θεσσαλονίκης. Ένα σημαντικά μικρότερο ποσοστό είναι αυτό των περιοχών με ενδιάμεση πυκνότητα (12.8%).

Εντολές στην R

Η μεταβλητή “density” είναι μια κατηγορική μεταβλητή, η οποία λαμβάνει τις τιμές 1 για πυκνοκατοικημένες περιοχές (≥ 500 κάτοικοι/km²), 2 για περιοχές ενδιάμεσης πυκνότητας (100-499 κάτοικοι/km²) και 3 για αραιοκατοικημένες περιοχές (≤ 100 /km²).

Η μεταβλητή “density” μετατρέπεται σε κατηγορική με την εντολή factor(), οι συχνότητες κάθε κατηγορίας υπολογίζονται με την εντολή table() καθώς και τα αντίστοιχα ποσοστά με την εντολή round(). Τέλος, η γραφική μέθοδος αναπαράστασης των τιμών για την εν λόγω μεταβλητή είναι το ραβδόγραμμα το οποίο κατασκευάζεται με την εντολή barplot() και η προσθήκη της λεζάντας γίνεται με την εντολή legend().


```

> density <- dataHH$density
> density <- factor (density)
> table (density)
density
 1  2  3
1489 451 1572
> density_perc <- round (prop.table (table (density))*100, 1)
> density_perc
density
 1  2  3
42.4 12.8 44.8
> barplot (table (density), names.arg = c ('42.4%', '12.8%', '44.8%'), col = c (3, 4, 6),
cex.names = 0.8, ylim = c (0, 2000), xlim = c (0, 5), width = 0.7, xlab = 'Βαθμός
πυκνότητας', ylab = 'Απόλυτη Συχνότητα')
> legend ('right', fill = c (3, 4, 6), c ('Πυκνοκατοικημένες', 'Ενδιάμεσης πυκνότητας',
'Αραιοκατοικημένες'), cex = 1)

```

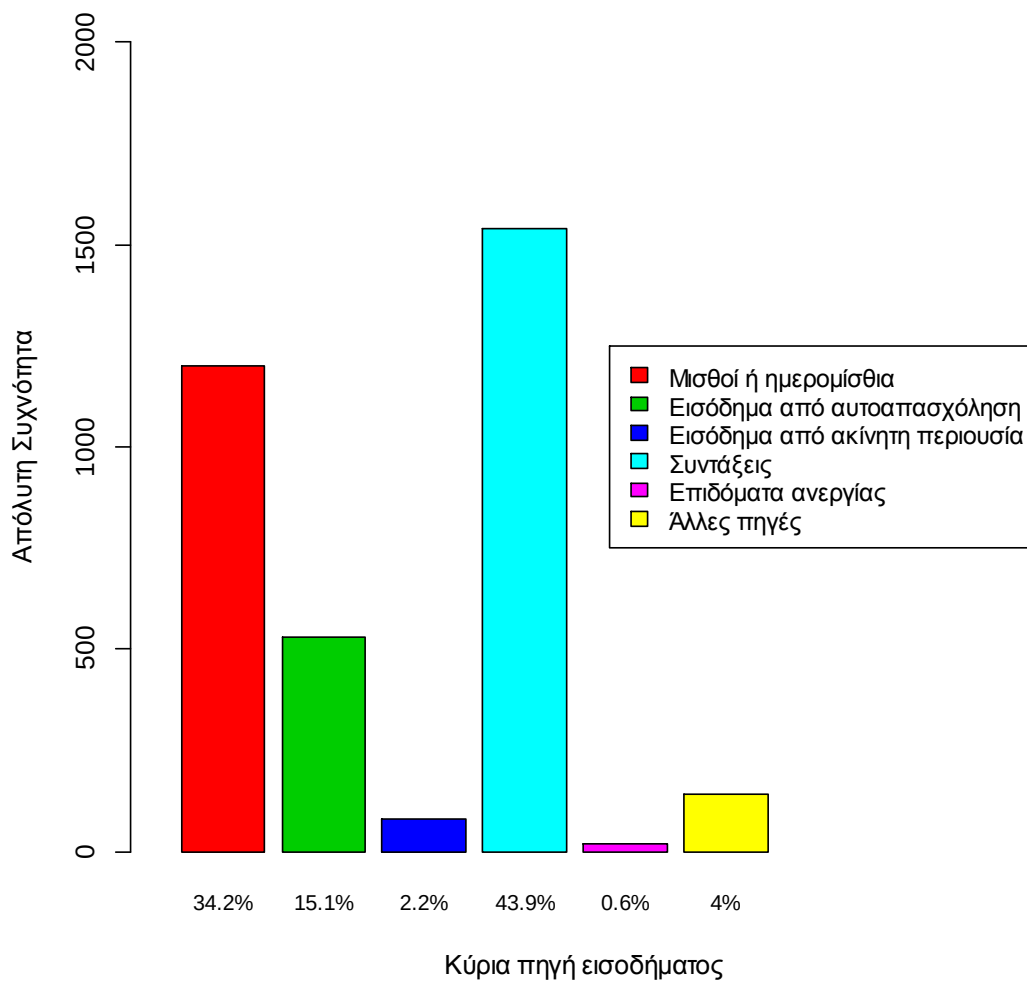
2.1.3. Κύρια πηγή εισοδήματος του νοικοκυριού

Ως κύρια πηγή εισοδήματος θεωρείται αυτή η οποία αποδίδει το μεγαλύτερο εισόδημα στο νοικοκυριό συγκριτικά με τις υπόλοιπες. Η κύρια πηγή εισοδήματος για το κάθε νοικοκυριό μπορεί να είναι οι μισθοί ή τα ημερομίσθια, το εισόδημα από αυτοαπασχόληση, το εισόδημα από ακίνητη περιουσία, οι συντάξεις, τα επιδόματα ανεργίας ή άλλα επιδόματα/εισοδήματα από άλλες πηγές.

Πίνακας 2.3: Πίνακας συχνοτήτων για τη μεταβλητή κύρια πηγή εισοδήματος

Κύρια Πηγή Εισοδήματος	Απόλυτη Συχνότητα	Ποσοστιαία Συχνότητα (%)
Μισθοί ή ημερομίσθια	1201	34.2
Εισόδημα από αυτοαπασχόληση	531	15.1
Εισόδημα από ακίνητη περιουσία	79	2.2
Συντάξεις	1541	43.9
Επιδόματα ανεργίας	20	0.6
Άλλα επιδόματα ή εισοδήματα από άλλες πηγές	140	4.0
Σύνολο	3512	100.0

Διάγραμμα 2.3: Ραβδόγραμμα για τη μεταβλητή κύρια πηγή εισοδήματος



Από τον Πίνακα 2.3 αλλά και το Διάγραμμα 2.3 προκύπτει ότι το μεγαλύτερο ποσοστό των νοικοκυριών έχει ως κύρια πηγή εισοδήματος τις συντάξεις (43.9%) και ακολουθεί ένα ελαφρώς μικρότερο ποσοστό με τους μισθούς ή ημερομίσθια (34.2%). Στη συνέχεια, ένα σημαντικό ποσοστό των νοικοκυριών δηλώνει ως κύρια πηγή τα εισοδήματα από αυτοαπασχόληση (15.1%). Τέλος, πολύ μικρότερα ποσοστά έχουν ως κύρια πηγή εισοδήματος επιδόματα από άλλες πηγές (4%), τα εισοδήματα από ακίνητη περιουσία (2.2%) και τα επιδόματα ανεργίας (0.6%).

Εντολές στην R

Η μεταβλητή “source” είναι μια κατηγορική μεταβλητή, η οποία λαμβάνει τις τιμές 1 για μισθούς ή ημερομίσθια, 2 για εισοδήματα από αυτοαπασχόληση, 3 για εισοδήματα από ακίνητη περιουσία, 4 για συντάξεις, 5 για επιδόματα ανεργίας και 6 για άλλα επιδόματα ή εισοδήματα από άλλες πηγές.

Η μεταβλητή “source” μετατρέπεται σε κατηγορική με την εντολή `factor()`, οι συχνότητες κάθε κατηγορίας υπολογίζονται με την εντολή `table()` καθώς και τα αντίστοιχα ποσοστά με την εντολή `round()`. Τέλος, η γραφική μέθοδος αναπαράστασης των τιμών για την εν λόγω μεταβλητή είναι το ραβδόγραμμα το

οποίο κατασκευάζεται με την εντολή `barplot()` και η προσθήκη λεζάντας γίνεται με την εντολή `legend()`.

```
> source <- dataHH$source
> source <- factor (source)
> table (source)
source
 1  2  3  4  5  6
1201 531 791541 20 140
> source_perc <- round (prop.table (table (source))*100, 1)
> source_perc
source
 1  2  3  4  5  6
34.2 15.1 2.2 43.9 0.6 4.0
> barplot (table (source), names.arg = c ('34.2%', '15.1%', '2.2%', '43.9%', '0.6%',
'4%'), col = c (2, 3, 4, 5, 6, 7), cex.names = 0.8, ylim = c (0, 2000), xlim = c (0, 8), width
= 0.8, ylab = 'Απόλυτη Συχνότητα', xlab = 'Κύρια πηγή εισοδήματος')
> legend ('right', fill = c (2, 3, 4, 5, 6, 7), c ('Μισθοί ή ημερομίσθια', 'Εισόδημα από
αυτοαπασχόληση', 'Εισόδημα από ακίνητη περιουσία', 'Συντάξεις', 'Επιδόματα
ανεργίας', 'Άλλες πηγές'), cex = 0.9)
```

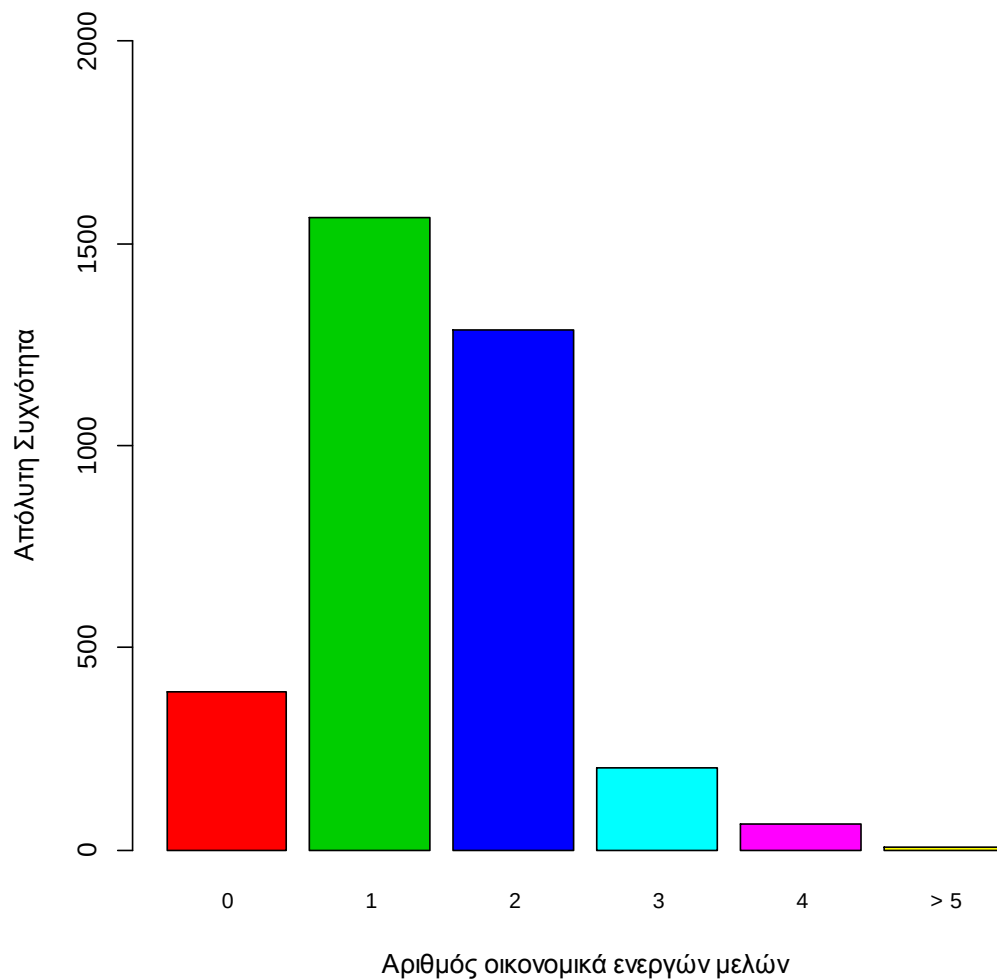
2.1.4. Αριθμός οικονομικά ενεργών μελών του νοικοκυριού

Ως οικονομικά ενεργά μέλη θεωρούνται τα μέλη ηλικίας 14 ετών και άνω τα οποία, κατά την προηγούμενη εβδομάδα από την έναρξη της έρευνας, είχαν εργασία ή, αν δεν είχαν, αναζητούσαν εργασία, δηλαδή οι εργαζόμενοι και οι άνεργοι.

Πίνακας 2.4: Πίνακας συχνοτήτων για τη μεταβλητή αριθμός οικονομικά ενεργών μελών του νοικοκυριού

Αριθμός οικονομικά ενεργών μελών	Απόλυτη Συχνότητα	Ποσοστιαία Συχνότητα (%)
0	391	11.1
1	1562	44.5
2	1287	36.7
3	203	5.8
4	64	1.8
5 ή περισσότερα	5	0.1
Σύνολο	3512	100.0

Διάγραμμα 2.4: Ραβδόγραμμα για τη μεταβλητή αριθμός των οικονομικά ενεργών μελών του νοικοκυριού



Όπως προκύπτει από τον Πίνακα 2.4 αλλά και το Διάγραμμα 2.4, το μεγαλύτερο ποσοστό νοικοκυριών έχει 1 (44.5%) ή 2 (36.7%) οικονομικά ενεργά μέλη. Στη συνέχεια, σχεδόν 1 στα 10 νοικοκυριά (11.1%) δεν έχουν κανένα οικονομικά ενεργό μέλος. Σε μικρότερα ποσοστά εμφανίζονται νοικοκυριά τα οποία δηλώνουν ότι έχουν 3 (5.8%), 4 (1.8%) και 5 ή περισσότερα (0.1%) οικονομικά ενεργά μέλη.

Εντολές στην R

Η μεταβλητή “members” είναι μια κατηγορική μεταβλητή, η οποία λαμβάνει τις τιμές 0, 1, 2, 3, 4 και 5 ανάλογα με τον αριθμό των οικονομικά ενεργών μελών του νοικοκυριού.

Η μεταβλητή “members” μετατρέπεται σε κατηγορική με την εντολή `factor()`, οι συχνότητες κάθε κατηγορίας υπολογίζονται με την εντολή `table()` καθώς και τα αντίστοιχα ποσοστά με την εντολή `round()`. Τέλος, η γραφική μέθοδος αναπαράστασης των τιμών για την εν λόγω μεταβλητή είναι το ραβδόγραμμα το οποίο κατασκευάζεται με την εντολή `barplot()`.

```
> members <- dataHH$members
```

```

> members <- factor (members)
> table (members)
members
 0  1  2  3  4  5
391 1562 1287 203  64  5
> members_perc <- round (prop.table (table (members))*100, 1)
> members_perc
members
 0  1  2  3  4  5
11.1 44.5 36.7 5.8 1.8 0.1
> barplot (table (members), names.arg = c ('0', '1', '2', '3', '4', '> 5'), col = c (2, 3, 4, 5, 6, 7), cex.names = 0.8, ylim = c (0, 2000), ylab = 'Απόλυτη Συχνότητα', xlab = 'Αριθμός οικονομικά ενεργών μελών')

```

2.1.5. Κοινωνικο – οικονομική κατάσταση του προσώπου αναφοράς του νοικοκυριού

Ως πρόσωπο αναφοράς του νοικοκυριού θεωρείται:

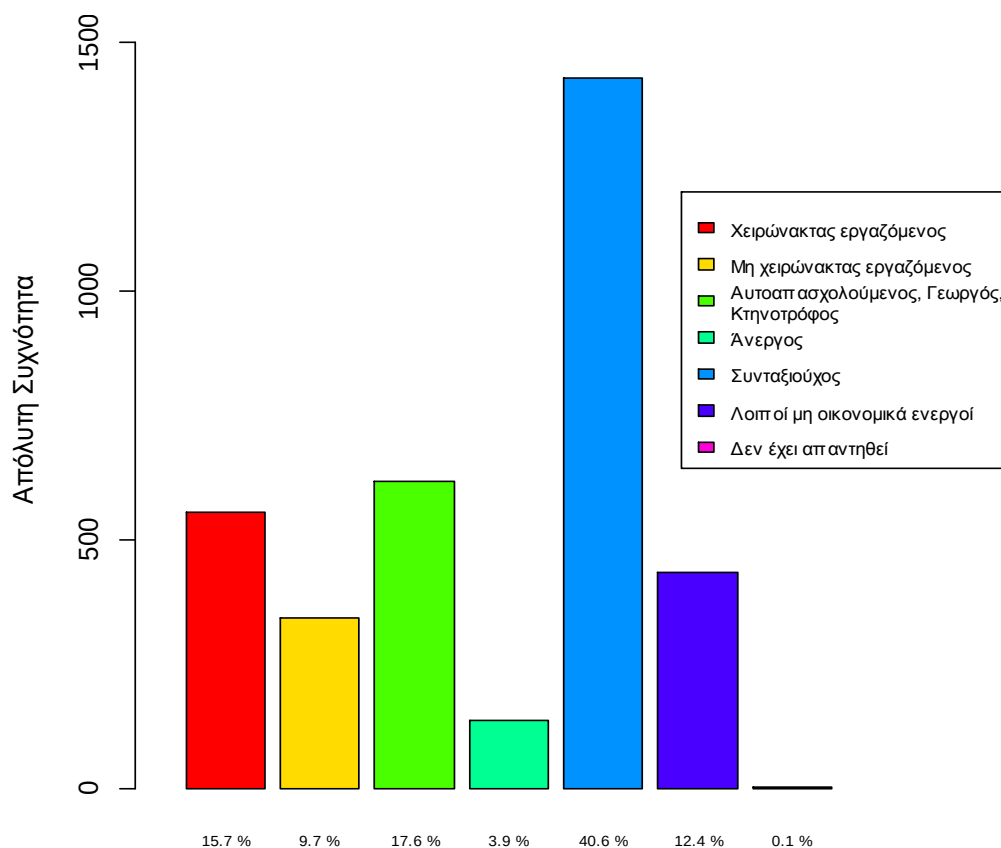
- Ο υπεύθυνος του νοικοκυριού, εφ' όσον είναι εργαζόμενος ή ψάχνει να βρει εργασία. Αν δεν εργάζεται αλλά ούτε αναζητεί εργασία, για να θεωρηθεί πρόσωπο αναφοράς θα πρέπει να μην υπάρχει άλλο άτομο οικονομικά ενεργό στο νοικοκυριό αυτό. Στις περισσότερες περιπτώσεις νοικοκυριών, τα οποία αποτελούνται από γονείς με παιδιά (άγαμα ή έγγαμα), ο πατέρας λαμβάνεται ως υπεύθυνος του νοικοκυριού.
- Ο/η σύζυγος ή σύντροφος του υπεύθυνου, εφ' όσον εργάζεται ή αναζητεί εργασία, αν το πρόσωπο που δηλώνει υπεύθυνος του νοικοκυριού δεν είναι οικονομικά ενεργό.
- Το μεγαλύτερο σε ηλικία ενήλικο μέλος του νοικοκυριού το οποίο είναι οικονομικά ενεργό, αν ούτε ο/η σύζυγος του υπεύθυνου ανήκει στον οικονομικά ενεργό πληθυσμό. Επίσης, αν το νοικοκυριό αποτελείται από συγγενικά ή μη άτομα ή αν κανένας δεν εργάζεται τότε το μεγαλύτερο σε ηλικία ενήλικο μέλος θεωρείται πρόσωπο αναφοράς.
- Ο υπεύθυνος του νοικοκυριού, αν κανένα από τα μέλη του νοικοκυριού δεν ανήκουν στον οικονομικά ενεργό πληθυσμό.

Το πρόσωπο αναφοράς του νοικοκυριού, ως προς την κοινωνικο – οικονομική του κατάσταση, μπορεί να είναι χειρώνακτας εργαζόμενος, μη χειρώνακτας εργαζόμενος, αυτοαπασχολούμενος – γεωργός – κτηνοτρόφος, άνεργος, συνταξιούχος και σπουδαστής – στρατιώτης – νοικοκυρά.

Πίνακας 2.5: Πίνακας συχνοτήτων για τη μεταβλητή κοινωνικο – οικονομική κατάσταση του προσώπου αναφοράς του νοικοκυριού

Κοινωνικο – οικονομική κατάσταση του προσώπου αναφοράς	Απόλυτη Συχνότητα	Ποσοστιαία Συχνότητα (%)
Χειρώνακτας εργαζόμενος	553	15.7
Μη χειρώνακτας εργαζόμενος	341	9.7
Αυτοαπασχολούμενος, γεωργός, κτηνοτρόφος	617	17.6
Άνεργος	136	3.9
Συνταξιούχος	1427	40.6
Σπουδαστής, στρατιώτης, νοικοκυρά	435	12.4
Δεν έχει απαντηθεί	3	0.1
Σύνολο	3512	100.0

Διάγραμμα 2.5: Ραβδόγραμμα για τη μεταβλητή κοινωνικο – οικονομική κατάσταση του προσώπου αναφοράς του νοικοκυριού



Κοινωνικο - οικονομική κατάσταση του προσώπου αναφοράς

Από τον Πίνακα 2.5 αλλά και το Διάγραμμα 2.5 φαίνεται ότι στα περισσότερα νοικοκυριά το πρόσωπο αναφοράς είναι συνταξιούχος (40.6%). Ακολουθούν τα νοικοκυριά στα οποία το πρόσωπο αναφοράς είναι αυτοαπασχολούμενος – γεωργός – κτηνοτρόφος (17.6%), χειρώνακτας εργαζόμενος (15.7%) και σπουδαστής – στρατιώτης – νοικοκυρά (12.4%). Στη συνέχεια, με μικρότερα ποσοστά είναι τα

νοικοκυριά τα οποία δηλώνουν ότι το πρόσωπο αναφοράς είναι μη χειρώνακτας εργαζόμενος (9.7%) και άνεργος (3.9%). Τέλος, υπάρχει ένα πολύ μικρό ποσοστό νοικοκυριών που δεν έχει απαντήσει στο ερώτημα (0.1%).

Εντολές στην R

Η μεταβλητή “situation” είναι μια κατηγορική μεταβλητή, η οποία λαμβάνει τις τιμές 1 για χειρώνακτα εργαζόμενο στον ιδιωτικό τομέα, 2 για μη χειρώνακτα εργαζόμενο στον ιδιωτικό τομέα, 3 για χειρώνακτα εργαζόμενο στον δημόσιο τομέα, 4 για μη χειρώνακτα εργαζόμενο στον δημόσιο τομέα, 5 για αυτοαπασχολούμενο, 6 για γεωργό ή κτηνοτρόφο, 7 για άνεργο, 8 για συνταξιούχο, 9 για σπουδαστή ή στρατιώτη, 10 για νοικοκυρά, 11 για ανίκανο για εργασία και 99 αν δεν έχει απαντηθεί.

Ωστόσο, στην εν λόγω εργασία έχει δημιουργηθεί μια νέα κατηγορική μεταβλητή “sitnew” με λιγότερες κατηγορίες, η οποία λαμβάνει τις τιμές 1 για χειρώνακτα εργαζόμενο (στον ιδιωτικό και τον δημόσιο τομέα), 2 για μη χειρώνακτα εργαζόμενο (στον ιδιωτικό και τον δημόσιο τομέα), 3 για αυτοαπασχολούμενο, γεωργό ή κτηνοτρόφο, 4 για άνεργο, 5 για συνταξιούχο, 6 για άλλους μη οικονομικά ενεργούς (σπουδαστή ή στρατιώτη, νοικοκυρά και ανίκανο για εργασία) και 7 αν δεν έχει απαντηθεί.

Η μεταβλητή “sitnew” μετατρέπεται σε κατηγορική με την εντολή factor(), οι συχνότητες κάθε κατηγορίας υπολογίζονται με την εντολή table() καθώς και τα αντίστοιχα ποσοστά με την εντολή round(). Τέλος, η γραφική μέθοδος αναπαράστασης των τιμών για την εν λόγω μεταβλητή είναι το ραβδόγραμμα το οποίο κατασκευάζεται με την εντολή barplot() και η προσθήκη λεζάντας γίνεται με την εντολή legend().

```
> situation <- dataHH$situation
> situation <- factor(situation)
> table(situation)
situation
 1  2  3  4  5  6  7  8  9 10 11 99
460 131 93 210 465 152 136 1427 76 316 43 3
> sitnew<-rep(0,3512)
> sitnew[situation==1 | situation==3]<-1
> sitnew[situation==2 | situation==4]<-2
> sitnew[situation==5 | situation==6]<-3
> sitnew[situation==7]<-4
> sitnew[situation==8]<-5
> sitnew[situation==9 | situation==10 | situation==11]<-6
> sitnew[situation==99]<-7
> length(situation)
[1] 3512
> length(sitnew)
[1] 3512
> table(sitnew)
```

```

sitnew
  1  2  3  4  5  6  7
553 341 617 136 1427 435  3
> sitnew_perc <- round (prop.table (table (sitnew))*100, 1)
> sitnew_perc
sitnew
  1  2  3  4  5  6  7
15.7 9.7 17.6 3.9 40.6 12.4 0.1
> barplot (table (sitnew), names.arg = paste (sitnew_perc, '%'), cex.names = 0.6, col =
rainbow (7), ylim = c (0, 1600), xlim = c (0, 11), ylab = 'Απόλυτη Συχνότητα', xlab =
'Κοινωνικο - οικονομική κατάσταση του προσώπου αναφοράς')
> legend (6.5, 1200, fill = rainbow (7), cex = 0.7, c ('Χειρώνακτας εργαζόμενος', 'Μη
χειρώνακτας εργαζόμενος', 'Αυτοαπασχολούμενος, Γεωργός, \ηΚτηνοτρόφος',
'Άνεργος', 'Συνταξιούχος', 'Λοιποί μη οικονομικά ενεργοί' , 'Δεν έχει απαντηθεί'))

```

2.1.6. Σύνολο εισοδήματος του νοικοκυριού

Ως εισόδημα θεωρείται το σύνολο των απολαβών σε χρήμα ή και σε είδος των μελών του νοικοκυριού που προέρχονται από εργασία ή και άλλες πηγές. Εισοδήματα θεωρούνται οι μισθοί, οι συντάξεις, τα κέρδη από γεωργικές, κτηνοτροφικές, εμπορικές κλπ. επιχειρήσεις, τα έσοδα από την ενοικίαση της ακίνητης περιουσίας, τους τόκους και τα μερίσματα, τα κοινωνικά επιδόματα, οι τακτικές μεταβιβάσεις μεταξύ των νοικοκυριών κλπ. Για τη συλλογή αξιόπιστων πληροφοριών έχει ζητηθεί το καθαρό εισόδημα (δηλαδή μετά την αφαίρεση του φόρου και των κοινωνικών ασφαλιστικών εισφορών) κατά πηγή, για κάθε μέλος 14 ετών και άνω, έτσι ώστε να υπολογιστεί το συνολικό εισόδημα του νοικοκυριού. Συνεπώς, το συνολικό εισόδημα του νοικοκυριού έχει υπολογιστεί αθροίζοντας για όλα τα μέλη του τα εξής: τα εισοδήματα σε είδος από εργασία (μισθοί και ημερομίσθια σε είδος), το εισόδημα σε είδος που δεν προέρχεται από εργασία, το τεκμαρτό ενοίκιο και το καθαρό εισόδημα από τις πηγές μετά την αφαίρεση των φόρων.

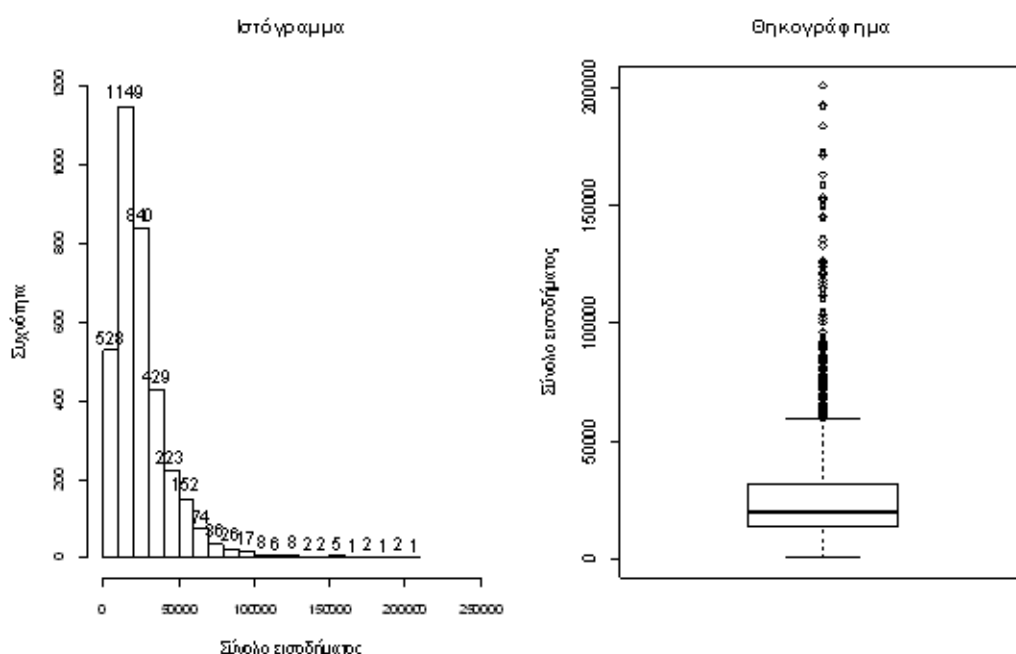
Πίνακας 2.6: Αριθμητικά μέτρα για τη μεταβλητή σύνολο εισοδήματος (σε €) του νοικοκυριού

Μέτρα Θέσης	Δειγματική Διάμεσος	20590
	Δειγματικός Μέσος	25770
Μέτρα Μεταβλητότητας	Μικρότερη Παρατήρηση	450
	Μεγαλύτερη Παρατήρηση	200900
	Εύρος Δείγματος	200450
	1ο τεταρτημόριο	13730
	3ο τεταρτημόριο	32220
	Ενδοτεταρτημοριακό Εύρος	18490
	Δειγματική Διασπορά	414331282
	Δειγματική Τυπική Απόκλιση	20355.13

Ο δειγματικός μέσος είναι 25770€ ενώ η δειγματική διάμεσος 20590€ και άρα, αφού το 50% των παρατηρήσεων και πλέον είναι τιμές μικρότερες του δειγματικού

μέσου, υπάρχει μεγαλύτερη συγκέντρωση των παρατηρήσεων στο διάστημα εκείνο που είναι αριστερά από τον δειγματικό μέσο. Επίσης, το ελάχιστο συνολικό ετήσιο εισόδημα είναι 450€ σε αντίθεση με το μέγιστο το οποίο είναι 200900€ και επομένως, υπάρχει μεγάλο εύρος δείγματος ίσο με 200450€, το οποίο επηρεάζεται προφανώς από τις ακραίες τιμές. Επειδή το 3ο τεταρτημόριο είναι 32220€, δηλαδή η παρατήρηση που είναι ίση ή μεγαλύτερη από το 75% ακριβώς των παρατηρήσεων, επαληθεύεται ότι, εν γένει, το δείγμα είναι μετατοπισμένο προς τα αριστερά με κάποιες ακραίες μεγαλύτερες τιμές. Το ενδοτεταρτημοριακό εύρος, το οποίο δεν επηρεάζεται από ακραίες τιμές, είναι ίσο με 18490€. Επιπρόσθετα, η διασπορά των τιμών είναι αρκετά μεγάλη, όπως φαίνεται από τη δειγματική διασπορά, η οποία είναι 414331282 και από την τυπική απόκλιση, η οποία είναι 20355.13€ και ισούται με τη θετική τετραγωνική ρίζα της διασποράς.

Διάγραμμα 2.6: Ιστόγραμμα και Θηκογράφημα για τη μεταβλητή σύνολο του εισοδήματος που έχουν λάβει τα νοικοκυριά κατά το έτος 2010



Από τα παραπάνω διαγράμματα, το ιστογράμμο συχνοτήτων για το σύνολο του εισοδήματος των νοικοκυριών και το αντίστοιχο θηκογράφημα, προκύπτει ότι τα περισσότερα νοικοκυριά (1989) συγκεντρώνουν ετήσιο εισόδημα κατά το έτος 2010 από 10000€ έως 30000€. Επιπρόσθετα, επιβεβαιώνονται τα συμπεράσματα από τον Πίνακα 2.6.

Εντολές στην R

Η μεταβλητή “income” είναι μια ποσοτική μεταβλητή μετρημένη σε €. Η μικρότερη παρατήρηση, το 1ο τεταρτημόριο, η δειγματική διάμεσος, ο δειγματικός μέσος, το 3ο τεταρτημόριο, η μεγαλύτερη παρατήρηση υπολογίζονται με την εντολή `summary()`, η (αμερόληπτη) δειγματική διασπορά με την εντολή `var()` και η τυπική απόκλιση με την εντολή `sd()`. Τέλος, οι γραφικές μέθοδοι αναπαράστασης των τιμών

για την εν λόγω μεταβλητή είναι το ιστόγραμμα συχνοτήτων και το θηκογράφημα τα οποία κατασκευάζονται με τις εντολές `hist()` και `boxplot()` αντίστοιχα.

```
> income <- dataHH$income
> summary (income)
  Min. 1st Qu.  Median   Mean 3rd Qu.   Max.
  450  13730  20590  25770  32220 200900
> var (income)
[1] 414331282
> sd (income)
[1] 20355.13
> par (mfrow = c (1, 2))
> hist (income, nclass = 21, labels = TRUE, xlab = 'Σύνολο εισοδήματος', ylab =
'Sυχνότητα', xlim = c (0, 250000), ylim = c (0, 1200), cex.axis = 0.8, cex.names = 1,
main = 'Ιστόγραμμα')
> boxplot (income, ylab = 'Σύνολο εισοδήματος', main = 'Θηκογράφημα')
```

2.1.7. Σύνολο καταναλωτικής δαπάνης του νοικοκυριού

Δαπάνες του νοικοκυριού είναι η αξία σε χρήμα, των αγαθών και υπηρεσιών που αγόρασε το νοικοκυριό ή έλαβε σε είδος (από τη δική του παραγωγή, από το δικό του κατάστημα, από τον εργοδότη ή από αλλού) για να καλύψει τις οικογενειακές και τις κοινωνικές του ανάγκες. Δεν θεωρούνται δαπάνες:

- Οι πληρωμές που αυξάνουν τα περιουσιακά στοιχεία του νοικοκυριού ή μειώνουν τα χρέη του σε τρίτους, όπως η αγορά και η επέκταση κατοικίας, οι τραπεζικές καταθέσεις, η εξόφληση δανείων, οι άμεσοι φόροι κλπ.
- Οι μεταβιβάσεις χρημάτων σε μέλη και μη μέλη του νοικοκυριού.
- Οι επαγγελματικές δαπάνες, όπως σε εργαλεία απαραίτητα για την άσκηση του επαγγέλματος, σε επαγγελματικά ταξίδια, σε καύσιμα αυτοκινήτων που χρησιμοποιούνται επαγγελματικά, σε σπόρους, σε γεωργικά φάρμακα κλπ. για καλλιέργεια της γης, σε ζωοτροφές για οικόσιτη κτηνοτροφία κλπ.

Στις περιπτώσεις όπου οι επαγγελματικές δαπάνες είναι κοινές με τις οικογενειακές, όπως συμβαίνει στη συστέγαση της οικογενειακής επιχείρησης με το νοικοκυριό, στη χρήση του επαγγελματικού αυτοκινήτου για διακοπές κλπ., γίνεται διαχωρισμός της δαπάνης προκειμένου να υπολογισθεί στο νοικοκυριό το τμήμα της δαπάνης που αναλογεί (π.χ. ενοίκιο, λογαριασμοί ηλεκτρικού, νερού, καύσιμα, τέλη κυκλοφορίας, ασφάλιστρα κλπ.).

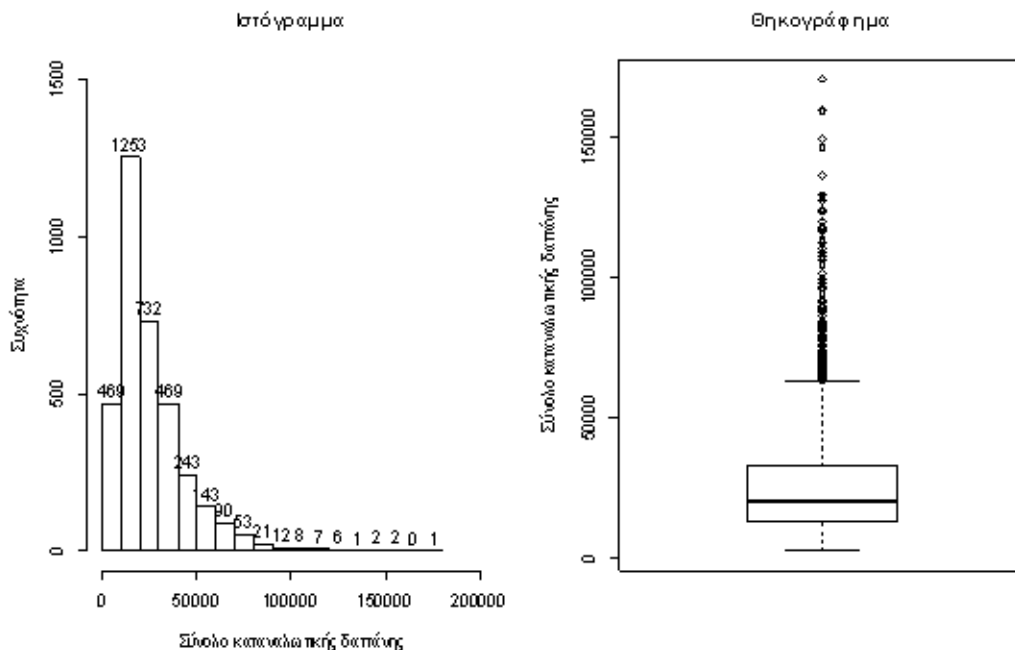
Η συνολική δαπάνη για το κάθε νοικοκυριό ισούται με το άθροισμα των δαπανών για τις 12 βασικές κατηγορίες, οι οποίες είναι είδη διατροφής και μη αλκοολούχα ποτά, οινοπνευματώδη ποτά και καπνός, είδη ένδυσης και υπόδησης, στέγαση, διαρκή αγαθά οικιακής χρήσης, υγεία, μεταφορές, επικοινωνίες, αναψυχή και πολιτισμός, εκπαίδευση, ξενοδοχεία και εστιατόρια και τέλος, διάφορα αγαθά και υπηρεσίες.

Πίνακας 2.7: Αριθμητικά μέτρα για τη μεταβλητή σύνολο καταναλωτικής δαπάνης (σε €) του νοικοκυριού

Μέτρα Θέσης	Δειγματική Διάμεσος	20310
	Δειγματικός Μέσος	25920
Μέτρα Μεταβλητότητας	Μικρότερη Παρατήρηση	2503
	Μεγαλύτερη Παρατήρηση	170400
	Εύρος Δείγματος	167897
	1ο τεταρτημόριο	12940
	3ο τεταρτημόριο	33020
	Ενδοτεταρτημοριακό Εύρος	18490
	Δειγματική Διασπορά	349341494
	Δειγματική Τυπική Απόκλιση	18690.68

Ο δειγματικός μέσος είναι 25920€ ενώ η δειγματική διάμεσος 20310€ και άρα, αφού το 50% των παρατηρήσεων και πλέον είναι τιμές μικρότερες του δειγματικού μέσου, υπάρχει μεγαλύτερη συγκέντρωση των παρατηρήσεων στο διάστημα εκείνο που είναι αριστερά από τον δειγματικό μέσο. Επίσης, το ελάχιστο συνολικό ετήσιο εισόδημα είναι 2503€ σε αντίθεση με το μέγιστο το οποίο είναι 170400€ και επομένως, υπάρχει μεγάλο εύρος δείγματος ίσο με 167897€. Επειδή το 3ο τεταρτημόριο είναι 33020€, δηλαδή η παρατήρηση που είναι ίση ή μεγαλύτερη από το 75% ακριβώς των παρατηρήσεων, επαληθεύεται ότι, εν γένει, το δείγμα είναι μετατοπισμένο προς τα αριστερά με κάποιες ακραίες μεγαλύτερες τιμές. Επιπρόσθετα, η διασπορά των τιμών είναι αρκετά μεγάλη όπως φαίνεται από τη δειγματική διασπορά, η οποία είναι 349341494 και από την τυπική απόκλιση η οποία είναι 18690.68€ και ισούται με τη θετική τετραγωνική ρίζα της διασποράς.

Διάγραμμα 2.7: Ιστόγραμμα και Θηκογράφημα για τη μεταβλητή σύνολο της καταναλωτικής δαπάνης των νοικοκυριών κατά το έτος 2010



Από τα παραπάνω διαγράμματα, το ιστόγραμμα συχνοτήτων για το σύνολο της καταναλωτικής δαπάνης των νοικοκυριών και το αντίστοιχο θηκογράφημα, προκύπτει ότι τα περισσότερα νοικοκυριά (1985) δαπανούν ετησίως το έτος 2010 από 10000€ έως 30000€. Επιπρόσθετα, επιβεβαιώνονται τα συμπεράσματα από τον Πίνακα 2.7. Συνεπώς, υπάρχει αντιστοιχία, όπως ήταν αναμενόμενο, με τα εισοδήματα που έχουν λάβει τα νοικοκυριά κατά το ίδιο έτος.

Εντολές στην R

Η μεταβλητή “total” είναι μια ποσοτική μεταβλητή μετρημένη σε €. Η μικρότερη παρατήρηση, το 1ο τεταρτημόριο, η δειγματική διάμεσος, ο δειγματικός μέσος, το 3ο τεταρτημόριο, η μεγαλύτερη παρατήρηση υπολογίζονται με την εντολή `summary()`, η (αμερόληπτη) δειγματική διασπορά με την εντολή `var()` και η τυπική απόκλιση με την εντολή `sd()`. Τέλος, οι γραφικές μέθοδοι αναπαράστασης των τιμών για την εν λόγω μεταβλητή είναι το ιστόγραμμα συχνοτήτων και το θηκογράφημα τα οποία κατασκευάζονται με τις εντολές `hist()` και `boxplot()` αντίστοιχα.

```
> total <- dataHH$total
> summary (total)
  Min. 1st Qu.  Median   Mean 3rd Qu.   Max.
 2503  12940  20310  25920  33020 170400
> var (total)
[1] 349341494
> sd (total)
[1] 18690.68
> par (mfrow = c (1, 2))
```

```
> hist (total, labels = TRUE, nclass = 18, main = 'Ιστόγραμμα', xlab = 'Σύνολο
καταναλωτικής δαπάνης', ylab = 'Συχνότητα', xlim = c (0, 200000), ylim = c (0, 1500))
> boxplot (total, main = 'Θηκογράφημα', ylab = 'Σύνολο καταναλωτικής δαπάνης')
```

2.1.8. Κατηγορίες δαπανών των νοικοκυριών

- **Είδη διατροφής και μη αλκοολούχα ποτά**

Στην κατηγορία αυτή έχουν καταγραφεί αναλυτικά δαπάνες:

- για είδη διατροφής, όπως αλεύρι, ψωμί, δημητριακά, κρέας, ψάρια, γαλακτοκομικά προϊόντα και αυγά, έλαια και λίπη, φρούτα, λαχανικά, ζάχαρη, μαρμελάδα, μέλι, σιρόπια, σοκολάτα και ζαχαρωτά, λοιπά είδη διατροφής.
- για μη οιοπνευματώδη προϊόντα, όπως καφές, τσάι, κακάο, μεταλλικά νερά, αναψυκτικά, χυμοί φρούτων και λαχανικών.

- **Οιοπνευματώδη ποτά και καπνός**

Στην κατηγορία αυτή έχουν καταγραφεί αναλυτικά δαπάνες:

- για οιοπνευματώδη ποτά, όπως αποστάγματα, κρασί και μύρα.
- για καπνό, όπως τσιγάρα, πούρα και λοιπά προϊόντα καπνού.

- **Είδη ένδυσης και υπόδησης**

Στην κατηγορία αυτή έχουν καταγραφεί αναλυτικά δαπάνες:

- για είδη ένδυσης, όπως υφάσματα, ενδύματα (ανδρικά, γυναικεία, παιδικά και βρεφικά), λοιπά είδη ένδυσης, επιδιορθώσεις, καθάρισμα και ενοικιάσεις ειδών ένδυσης.
- για είδη υπόδησης (ανδρικά, γυναικεία, παιδικά και βρεφικά), επιδιορθώσεις και ενοικιάσεις ειδών υπόδησης.

- **Στέγαση, ύδρευση, καύσιμα και ηλεκτρισμός**

Στην κατηγορία αυτή έχουν καταγραφεί αναλυτικά δαπάνες:

- για το καταβαλλόμενο ενοίκιο κύριας και δευτερεύουσας κατοικίας.
- για το τεκμαρτό ενοίκιο της ιδιοκατοικούμενης κατοικίας και άλλα τεκμαρτά ενοίκια.
- για τη συντήρηση και επισκευές κατοικίας, όπως τα υλικά και οι υπηρεσίες συντήρησης και επισκευής της κατοικίας.
- για την ύδρευση, την αποκομιδή και διάθεση απορριμμάτων και λοιπές υπηρεσίες σχετικά με την κατοικία.
- για ηλεκτρισμό, αέριο και λοιπά καύσιμα, όπως υγραέριο και φυσικό αέριο, υγρά καύσιμα, στερεά καύσιμα και θερμική ενέργεια.

- **Διαρκή αγαθά οικιακής χρήσης – οικιακά είδη άμεσης κατανάλωσης και οικιακές υπηρεσίες**

Στην κατηγορία αυτή έχουν καταγραφεί αναλυτικά δαπάνες:

- για έπιπλα, καλύμματα δαπέδου, φωτιστικά και είδη διακόσμησης, επισκευή επίπλων, διακοσμητικών ειδών και καλυμμάτων δαπέδου.
- για είδη επίπλωσης και κλινοστρωμνής.
- για ηλεκτρικές ή μη οικιακές συσκευές και την επισκευή αυτών, όπως ψυγεία, πλυντήρια ρούχων, συσκευές θέρμανσης και κλιματισμού και λοιπές οικιακές συσκευές.
- για πιατικά, υαλικά, μαχαιροπήρουνα, οικιακά σκεύη και την επισκευή υαλικών επιτραπέζιων σκευών και σκευών οικιακής χρήσης.
- για εργαλεία (μικρά ή μεγάλα), εξοπλισμό για την κατοικία και τον κήπο και διάφορα εξαρτήματα.
- για αγαθά και υπηρεσίες για τη συνήθη οικιακή συντήρηση, όπως προϊόντα καθαρισμού και συντήρησης, οικιακές υπηρεσίες, υπηρεσίες οικιακής μέριμνας και λοιπά οικιακά είδη.

- **Υγεία**

Στην κατηγορία αυτή έχουν καταγραφεί αναλυτικά δαπάνες:

- για φάρμακα, φαρμακευτικά είδη, θεραπευτικές συσκευές και εξοπλισμό και λοιπά ιατρικά προϊόντα.
- για ιατρικές και οδοντιατρικές υπηρεσίες, υπηρεσίες μικροβιολογικών εργαστηρίων και ακτινολογικών κέντρων και λοιπές μη νοσοκομειακές υπηρεσίες.
- για νοσοκομειακή περίθαλψη.

- **Μεταφορές**

Στην κατηγορία αυτή έχουν καταγραφεί αναλυτικά δαπάνες:

- για αγορά μεταφορικών μέσων, όπως αυτοκίνητα, μοτοσικλέτες, ποδήλατα και άμαξες.
- για κίνηση και συντήρηση μεταφορικών μέσων, όπως ανταλλακτικά, εξαρτήματα, καύσιμα, λιπαντικά, επισκευές και άλλες υπηρεσίες (ενοίκιο γκαράζ, διόδια, μαθήματα οδήγησης κλπ.).
- για μετακινήσεις και μεταφορές επιβατών με τρένο, μετρό, τραμ, λεωφορείο, τρόλεϊ, ταξί, αεροπλάνο, πλοίο και λοιπές μετακινήσεις και μεταφορές.

- **Επικοινωνίες**

Στην κατηγορία αυτή έχουν καταγραφεί αναλυτικά δαπάνες:

- για ταχυδρομικές υπηρεσίες.
- για συσκευές τηλεφώνου και τέλεφαξ.
- για υπηρεσίες τηλεφωνίας και τηλεομοιοτυπίας.

- **Αναψυχή και πολιτισμός**

Στην κατηγορία αυτή έχουν καταγραφεί αναλυτικά δαπάνες:

- για συσκευές λήψης, εγγραφής και αναπαραγωγής ήχου και εικόνας, φωτογραφικό, κινηματογραφικό εξοπλισμό και οπτικά όργανα, συσκευές επεξεργασίας δεδομένων, μέσα εγγραφής ήχου και εικόνας και επισκευή αυτών.
- για μεγάλα είδη αναψυχής και πολιτισμού, μουσικά όργανα και λοιπά μεγάλα αγαθά για ψυχαγωγία στο σπίτι και συντήρηση αυτών.
- για παιχνίδια, εξοπλισμό για αθλητισμό, κατασκήνωση και αναψυχή, κήπους, φυτά και άνθη, κατοικίδια ζώα και συναφή προϊόντα και κτηνιατρικές υπηρεσίες.
- για υπηρεσίες αθλητισμού και αναψυχής, υπηρεσίες πολιτισμού (κινηματογράφος, θέατρο, συναυλίες, μουσεία, ζωολογικοί κήποι κλπ.) και τυχερά παιχνίδια.
- για εφημερίδες, περιοδικά, βιβλία, είδη χαρτοπωλείου, γραφική ύλη και είδη σχεδίου.
- για ομαδικά ταξίδια.

- **Εκπαίδευση**

Στην κατηγορία αυτή έχουν καταγραφεί αναλυτικά δαπάνες:

- για προσχολική και πρωτοβάθμια εκπαίδευση.
- για δευτεροβάθμια εκπαίδευση.
- για μεταδευτεροβάθμια εκπαίδευση μη τριτοβάθμιου επιπέδου.
- για τριτοβάθμια εκπαίδευση.
- για εκπαίδευση που δεν μπορεί να ταξινομηθεί με βάση το επίπεδο.

- **Εστιατόρια και ξενοδοχεία**

Στην κατηγορία αυτή έχουν καταγραφεί αναλυτικά δαπάνες:

- για υπηρεσίες τροφοδοσίας, όπως εστιατόρια, καφενεία, καντίνες και παρόμοια καταστήματα.
- για υπηρεσίες ξενοδοχείων, πανσιόν και ενοικιαζόμενων δωματίων.

- **Διάφορα αγαθά και υπηρεσίες**

Στην κατηγορία αυτή έχουν καταγραφεί αναλυτικά δαπάνες:

- για ατομικό ευπρεπισμό, όπως κομμωτήρια και υπηρεσίες καλλωπισμού, ηλεκτρικές ή μη συσκευές και είδη ατομικού ευπρεπισμού.
- για προσωπικά είδη, όπως κοσμήματα, ρολόγια χειρός και τοίχου και λοιπά προσωπικά είδη.
- για υπηρεσίες κοινωνικής προστασίας και βρεφονηπιακών σταθμών.
- για ασφάλεια σχετικά με την κατοικία, ασφάλεια υγείας, αυτοκινήτου και ταξιδίων και λοιπές ασφάλειες.
- για οικονομικές υπηρεσίες.
- για διάφορες άλλες υπηρεσίες.

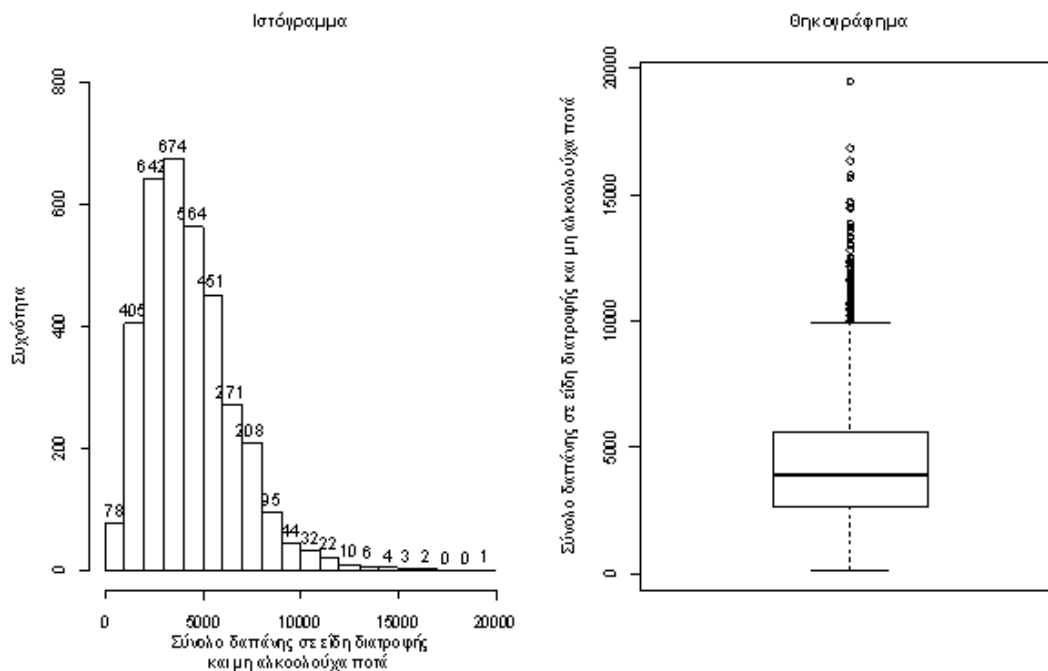
Στη συνέχεια, στον Πίνακα 2.8 δίνονται τα αριθμητικά μέτρα για τις 12 βασικές κατηγορίες δαπανών, οι οποίες είναι κατά σειρά: είδη διατροφής και μη αλκοολούχα ποτά, οινοπνευματώδη ποτά και καπνός, είδη ένδυσης και υπόδησης, στέγαση, ύδρευση, καύσιμα και ηλεκτρισμός, διαρκή αγαθά οικιακής χρήσης, υγεία, μεταφορές, επικοινωνίες, αναψυχή και πολιτισμός, εκπαίδευση, εστιατόρια και ξενοδοχεία και διάφορα αγαθά και υπηρεσίες.

Πίνακας 2.8: Αριθμητικά μέτρα για τις μεταβλητές κατηγορίες δαπανών (σε €) του νοικοκυριού

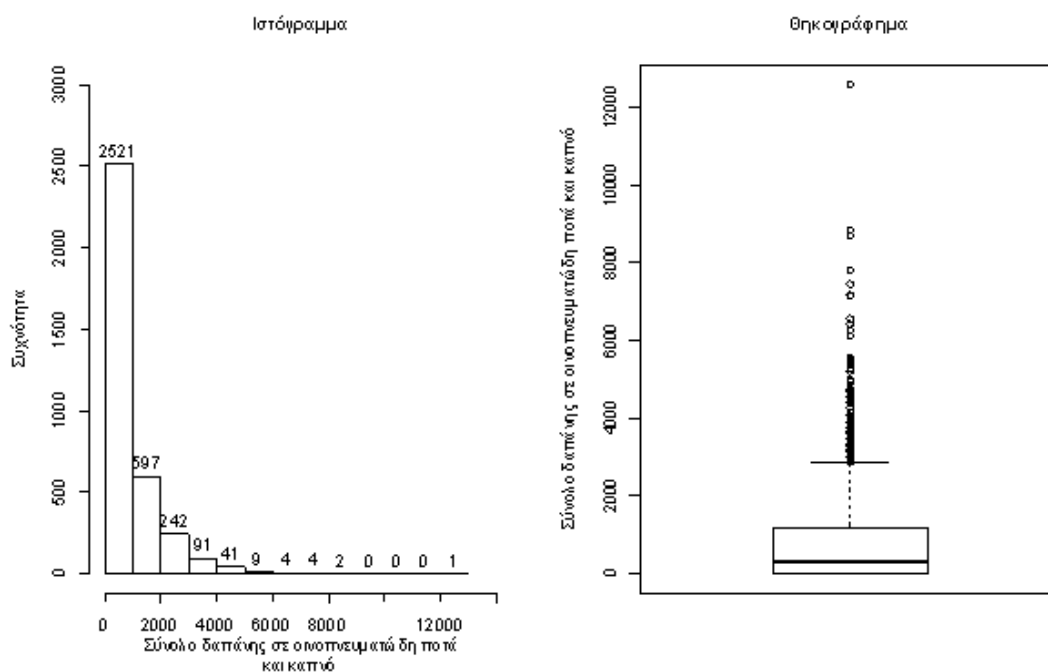
	Μέτρα Θέσης		Μέτρα Μεταβλητότητας							
	Δειγματική Διάμεσος	Δειγματικός Μέσος	Μικρότερη Παρατήρηση	Μεγαλύτερη Παρατήρηση	Εύρος Δείγματος	1ο τεταρτημόριο	3ο τεταρτημόριο	Ενδοτεταρτημοριακό Εύρος	Δειγματική Διασπορά	Δειγματική Τυπική Απόκλιση
1η	3922.0	4317.0	135.6	19490.0	19354.4	2652.0	5581.0	2929	5265780	2294.729
2η	282.7	739.8	0.0	12600.0	12600	0.0	1144.0	1144	1116657	1056.72
3η	756.0	1478.0	0.0	25610.0	25610	160.0	2040.0	1880	4313977	2077.012
4η	5962.0	7421.0	660.0	67540.0	66880	4254.0	8833.0	4579	28075392	5298.622
5η	644.8	1476.0	0.0	35360.0	35360.0	250.4	1564.0	1313.6	6766649	2601.278
6η	730.3	1473.0	0.0	33100.0	33100.0	204.1	1783.0	1578.9	5823885	2413.273
7η	1304.0	2556.0	0.0	67760.0	67760.0	228.2	2759.0	2530.8	23249227	4821.745
8η	642.0	852.8	0.0	7627.0	7627.0	360.0	1163.0	803	489202.2	699.4299
9η	300.1	965.5	0.0	32270.0	32270.0	52.2	1090.0	1037.8	4398901	2097.356
10η	0.0	661.7	0.0	30000.0	30000.0	0.0	0.0	0.0	3892708	1972.995
11η	1693.0	2398.0	0.0	30440.0	30440.0	521.6	3373.0	2851.4	6867253	2620.544
12η	947.8	1578.0	0.0	38280.0	38280.0	395.0	1949.0	1554	4668743	2160.727

Για όλες τις κατηγορίες δαπανών ο δειγματικός μέσος είναι μεγαλύτερος από τη δειγματική διάμεσο και άρα, υπάρχει μεγαλύτερη συγκέντρωση των παρατηρήσεων στο διάστημα εκείνο που είναι αριστερά από τον δειγματικό μέσο, εφ' όσον το 50% των παρατηρήσεων και πλέον είναι τιμές μικρότερες από αυτόν. Επίσης, η ελάχιστη συνολική ετήσια δαπάνη για κάθε μία από αυτές τις κατηγορίες είναι πολύ μικρότερη από τη μέγιστη και άρα, υπάρχει μεγάλο εύρος δείγματος σε όλες τις επιμέρους κατηγορίες. Επειδή το 3ο τεταρτημόριο, δηλαδή η παρατήρηση που είναι ίση ή μεγαλύτερη από το 75% ακριβώς των παρατηρήσεων, είναι για όλες τις κατηγορίες αρκετά μικρότερο από τη μέγιστη παρατήρηση επαληθεύεται ότι, εν γένει, το δείγμα είναι μετατοπισμένο προς τα αριστερά με κάποιες ακραίες μεγαλύτερες τιμές. Επιπρόσθετα, η διασπορά των τιμών είναι σχετικά μεγάλη όπως φαίνεται από τη δειγματική διασπορά για όλες τις επιμέρους κατηγορίες και από τη δειγματική τυπική απόκλιση, η οποία ισούται με τη θετική τετραγωνική ρίζα της διασποράς.

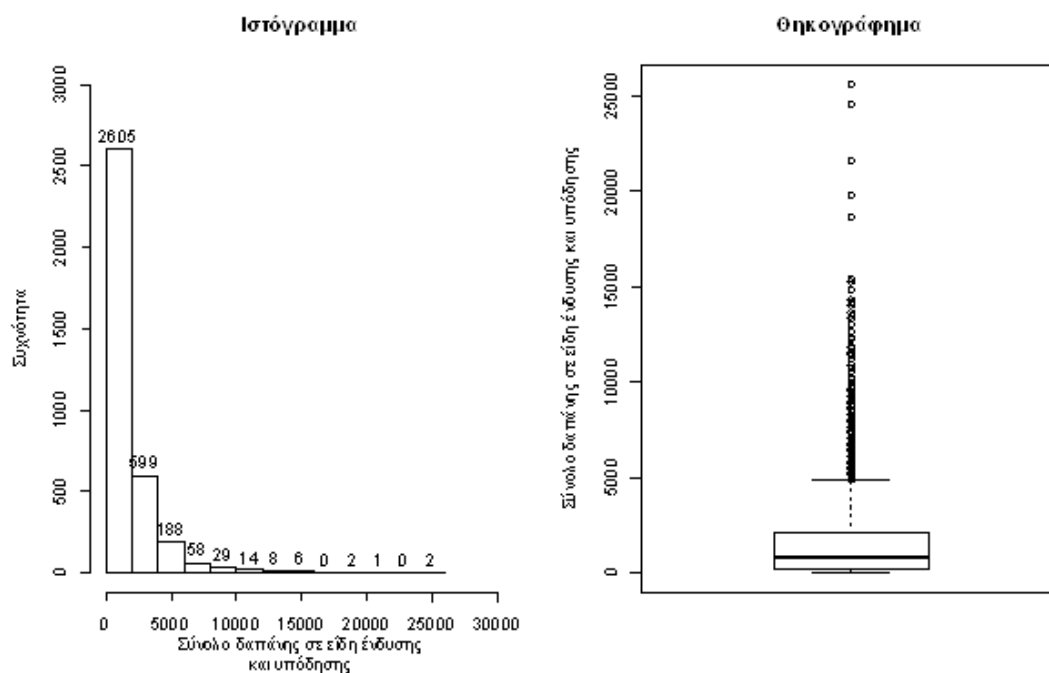
Διάγραμμα 2.8: Ιστογράμμο και Θηκογράφημα για τη μεταβλητή δαπάνη των νοικοκυριών σε είδη διατροφής και μη αλκοολούχα ποτά κατά το έτος 2010



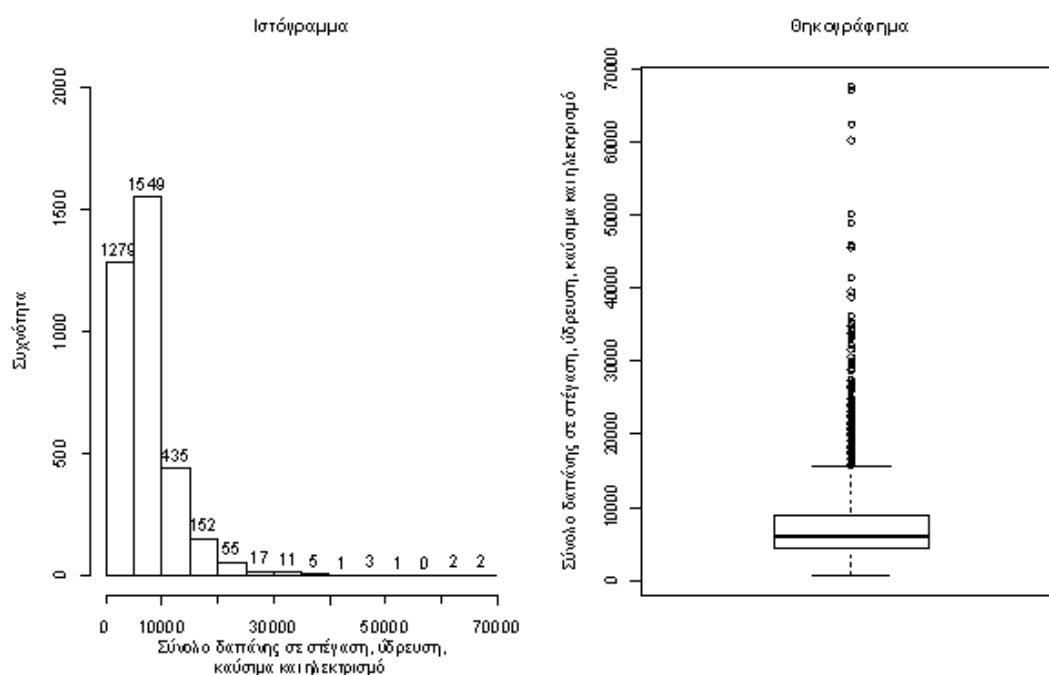
Διάγραμμα 2.9: Ιστογράμμο και Θηκογράφημα για τη μεταβλητή δαπάνη των νοικοκυριών σε οινοπνευματώδη ποτά και καπνό κατά το έτος 2010



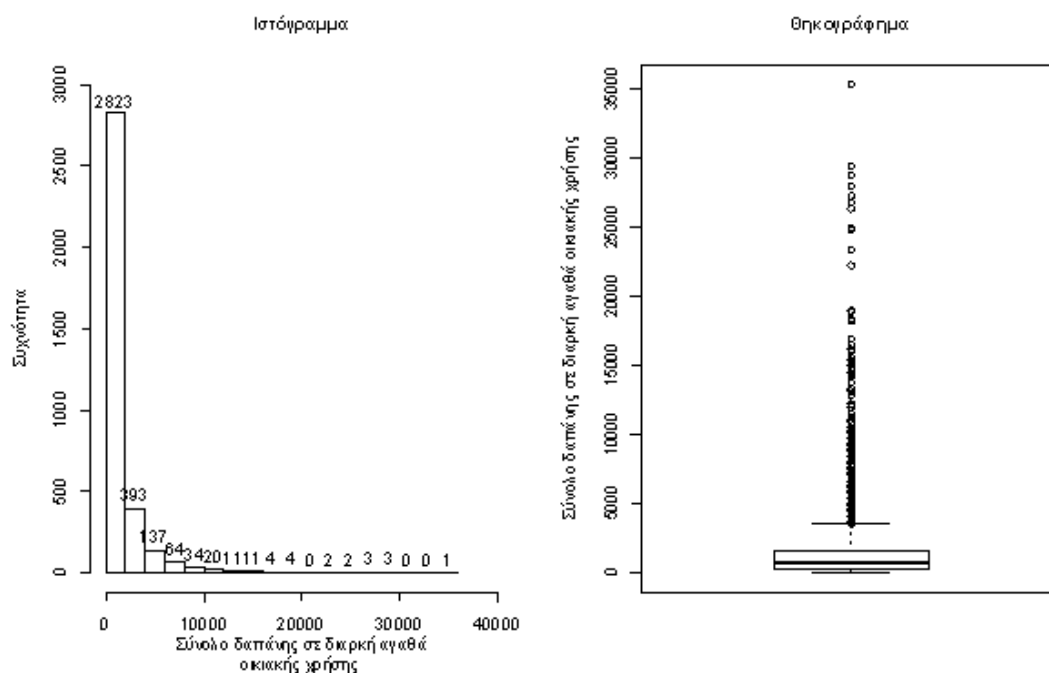
Διάγραμμα 2.10: Ιστόγραμμα και Θηκογράφημα για τη μεταβλητή δαπάνη των νοικοκυριών σε είδη ένδυσης και υπόδησης κατά το έτος 2010



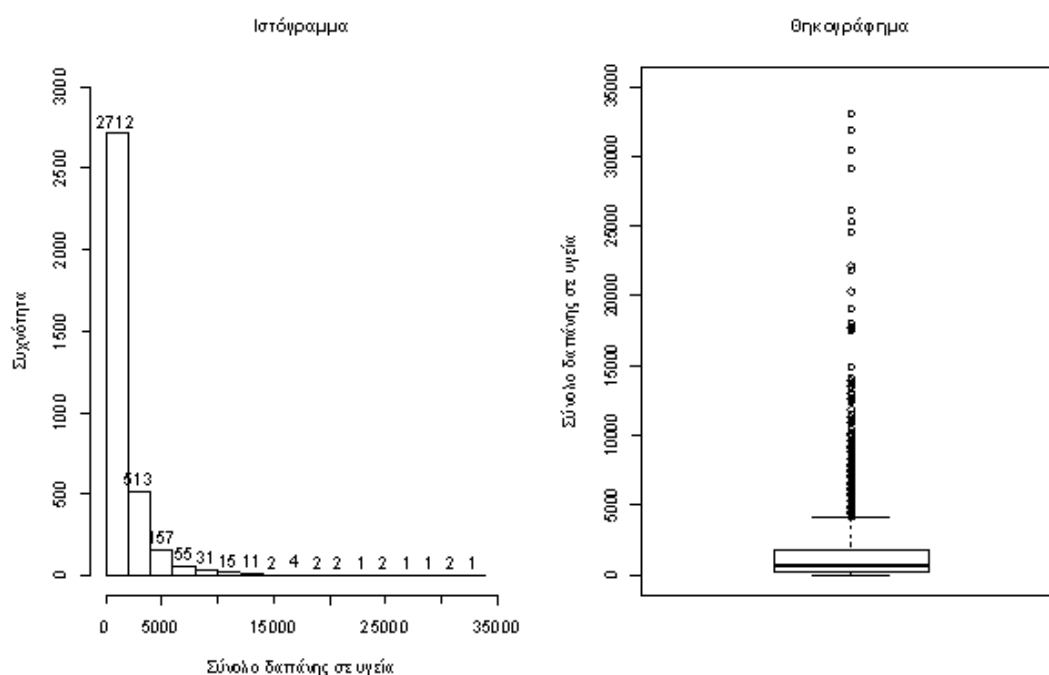
Διάγραμμα 2.11: Ιστόγραμμα και Θηκογράφημα για τη μεταβλητή δαπάνη των νοικοκυριών σε στέγαση, ύδρευση, καύσιμα και ηλεκτρισμό κατά το έτος 2010



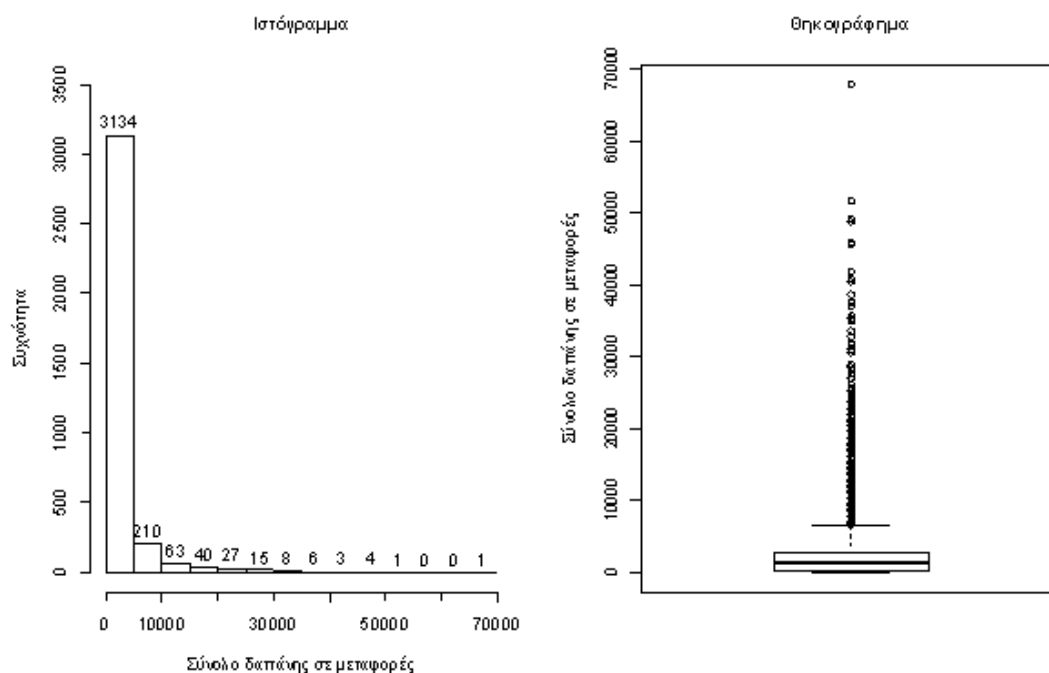
Διάγραμμα 2.12: Ιστόγραμμα και Θηκογράφημα για τη μεταβλητή δαπάνη των νοικοκυριών σε διαρκή αγαθά οικιακής χρήσης κατά το έτος 2010



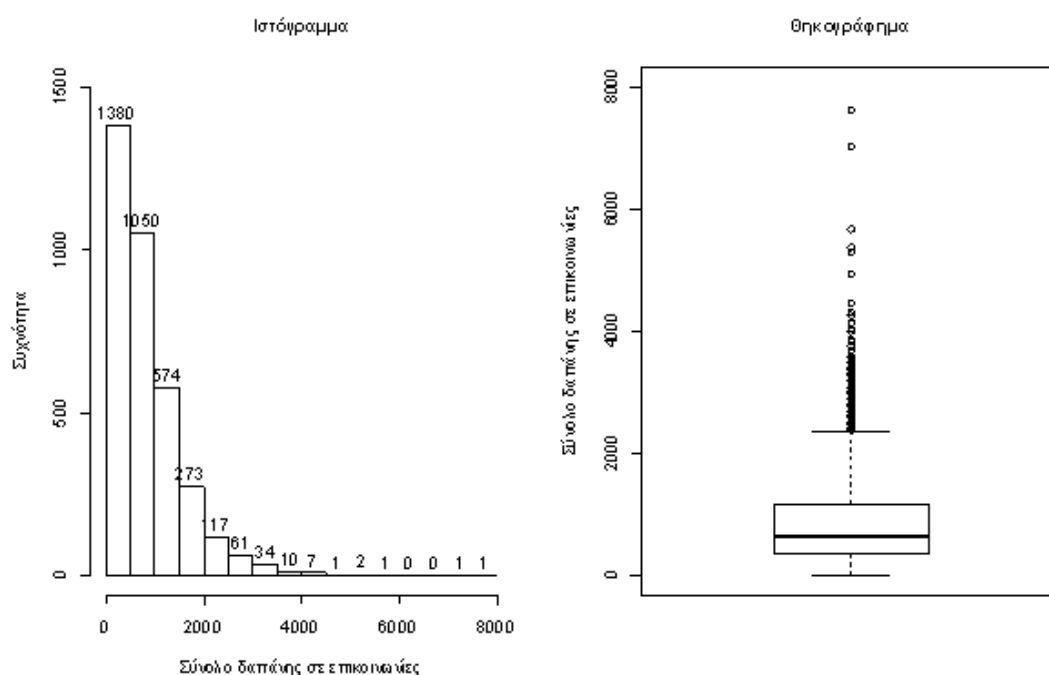
Διάγραμμα 2.13: Ιστόγραμμα και Θηκογράφημα για τη μεταβλητή δαπάνη των νοικοκυριών σε υγεία κατά το έτος 2010



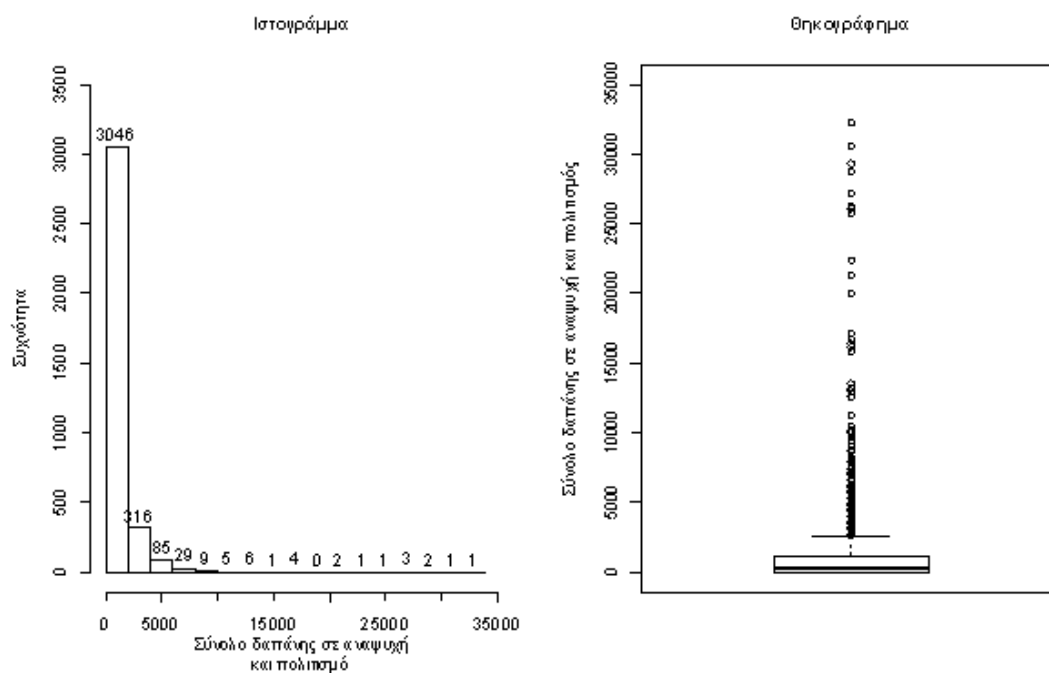
Διάγραμμα 2.14: Ιστόγραμμα και Θηκογράφημα για τη μεταβλητή δαπάνη των νοικοκυριών σε μεταφορές κατά το έτος 2010



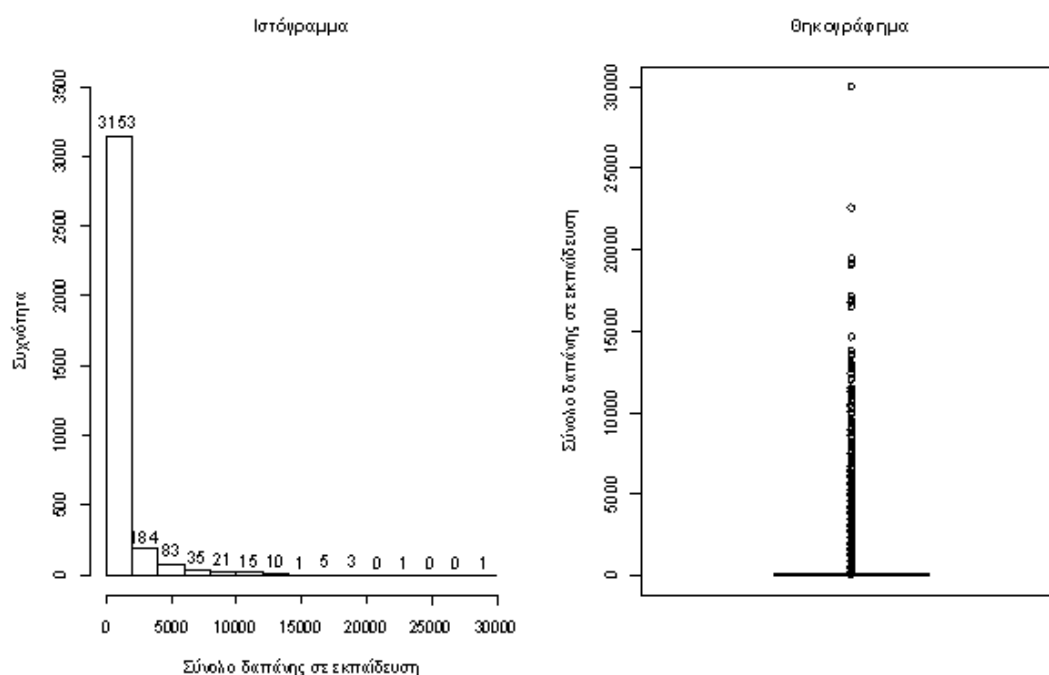
Διάγραμμα 2.15: Ιστόγραμμα και Θηκογράφημα για τη μεταβλητή δαπάνη των νοικοκυριών σε επικοινωνίες κατά το έτος 2010



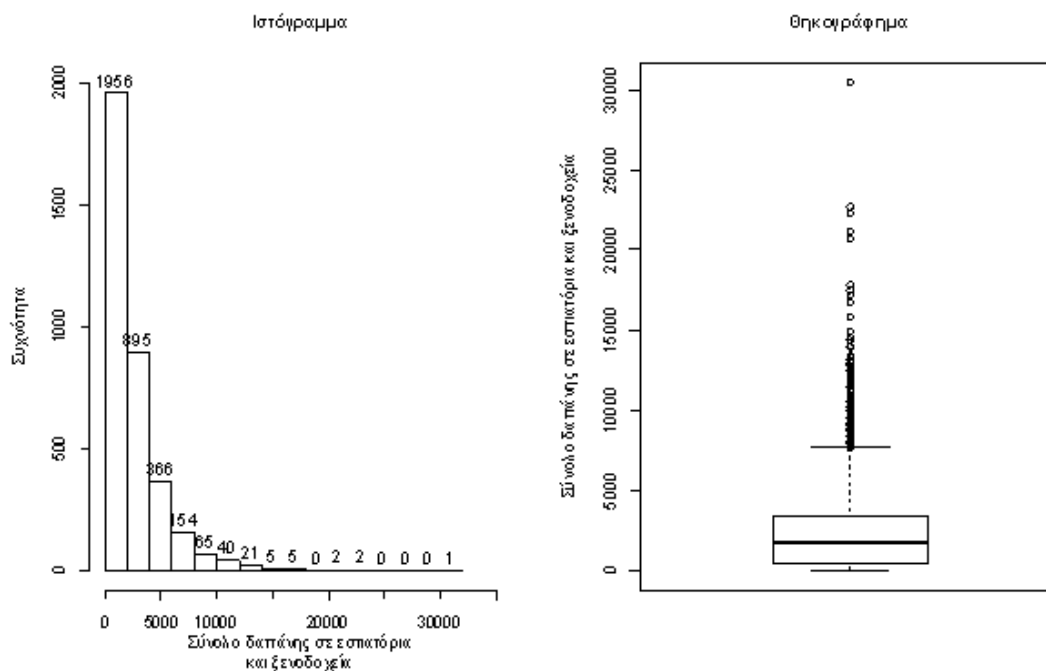
Διάγραμμα 2.16: Ιστόγραμμα και Θηκογράφημα για τη μεταβλητή δαπάνη των νοικοκυριών σε αναψυχή και πολιτισμό κατά το έτος 2010



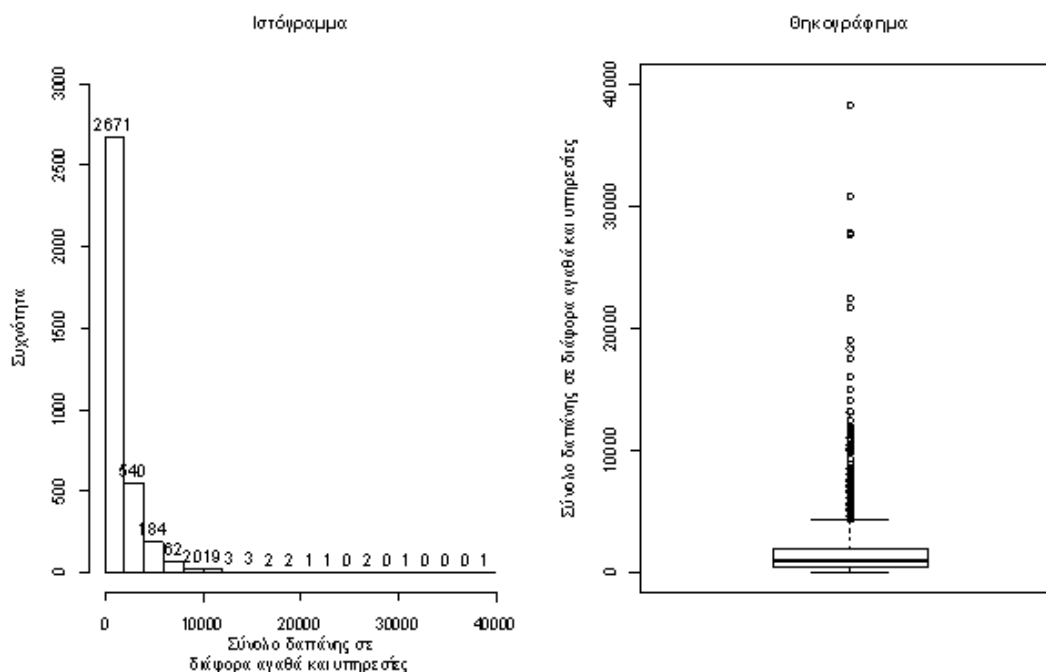
Διάγραμμα 2.17: Ιστόγραμμα και Θηκογράφημα για τη μεταβλητή δαπάνη των νοικοκυριών σε εκπαίδευση κατά το έτος 2010



Διάγραμμα 2.18: Ιστόγραμμα και Θηκογράφημα για τη μεταβλητή δαπάνη των νοικοκυριών σε εστιατόρια και ξενοδοχεία κατά το έτος 2010



Διάγραμμα 2.19: Ιστόγραμμα και Θηκογράφημα για τη μεταβλητή δαπάνη των νοικοκυριών σε διάφορα αγαθά και υπηρεσίες κατά το έτος 2010



Από τα παραπάνω διαγράμματα, το ιστόγραμμα συχνοτήτων και το αντίστοιχο θηκογράφημα για κάθε κατηγορία δαπανών, προκύπτει ότι ετησίως τα περισσότερα νοικοκυριά δαπανούν το έτος 2010 από 2000€ έως 5000€ σε είδη διατροφής και μη αλκοολούχα ποτά (1880), από 0€ έως 1000€ σε οινόπνευματώδη ποτά και καπνό

(2521), από 0€ έως 2000€ σε είδη ένδυσης και υπόδησης (2605), από 0€ έως 10000€ σε στέγαση, ύδρευση, καύσιμα και ηλεκτρισμό (2828), από 0€ έως 2000€ σε διαρκή αγαθά οικιακής χρήσης (2823), από 0€ έως 2000€ σε υγεία (2712), από 0€ έως 5000€ σε μεταφορές (3134), από 0€ έως 1000€ σε επικοινωνίες (2430), από 0€ έως 2000€ σε αναψυχή και πολιτισμό (3046), από 0€ έως 2000€ σε εκπαίδευση (3153), από 0€ έως 2000€ σε εστιατόρια και ξενοδοχεία (1956) και από 0€ έως 2000€ σε διάφορα αγαθά και υπηρεσίες (2671).

Τέλος, είναι σημαντικό να επισημανθεί ότι το έτος 2010 τα νοικοκυριά έχουν σημειώσει μεγαλύτερη δαπάνη για τη στέγαση, την ύδρευση, τα καύσιμα και τον ηλεκτρισμό καθώς επίσης και για τα είδη διατροφής. Ενώ αντίθετα, έχουν σημειώσει μικρότερη δαπάνη κατά το ίδιο έτος για τα οινοπνευματώδη ποτά και την εκπαίδευση.

Εντολές στην R

Όλες οι μεταβλητές είναι ποσοτικές μεταβλητές μετρημένες σε €. Η μικρότερη παρατήρηση, το 1ο τεταρτημόριο, η δειγματική διάμεσος, ο δειγματικός μέσος, το 3ο τεταρτημόριο, η μεγαλύτερη παρατήρηση υπολογίζονται με την εντολή `summary()`, η (αμερόληπτη) δειγματική διασπορά με την εντολή `var()` και η τυπική απόκλιση με την εντολή `sd()`. Τέλος, οι γραφικές μέθοδοι αναπαράστασης των τιμών για τις μεταβλητές αυτές είναι το ιστόγραμμα συχνοτήτων και το θηκογράφημα τα οποία κατασκευάζονται με τις εντολές `hist()` και `boxplot()` αντίστοιχα.

```
> summary (dataHH [, 9:20])
  food    alcohol    clothing    housing
Min. : 135.6 Min. : 0.0 Min. : 0 Min. : 660
1st Qu.: 2651.8 1st Qu.: 0.0 1st Qu.: 160 1st Qu.: 4254
Median : 3922.0 Median : 282.7 Median : 756 Median : 5962
Mean : 4317.1 Mean : 739.8 Mean : 1478 Mean : 7421
3rd Qu.: 5581.2 3rd Qu.: 1143.8 3rd Qu.: 2040 3rd Qu.: 8833
Max. : 19487.0 Max. : 12595.6 Max. : 25606 Max. : 67540
equipment    health    transport    communication
Min. : 0.0 Min. : 0.0 Min. : 0.0 Min. : 0.0
1st Qu.: 250.4 1st Qu.: 204.1 1st Qu.: 228.2 1st Qu.: 360.0
Median : 644.8 Median : 730.3 Median : 1304.2 Median : 642.0
Mean : 1476.1 Mean : 1472.9 Mean : 2555.6 Mean : 852.8
3rd Qu.: 1564.3 3rd Qu.: 1782.8 3rd Qu.: 2758.6 3rd Qu.: 1163.5
Max. : 35359.0 Max. : 33096.7 Max. : 67761.1 Max. : 7627.2
culture    education    hotels    goods
Min. : 0.0 Min. : 0.0 Min. : 0.0 Min. : 0.0
1st Qu.: 52.2 1st Qu.: 0.0 1st Qu.: 521.6 1st Qu.: 394.9
Median : 300.1 Median : 0.0 Median : 1693.4 Median : 947.8
Mean : 965.5 Mean : 661.7 Mean : 2398.5 Mean : 1577.9
3rd Qu.: 1089.8 3rd Qu.: 0.0 3rd Qu.: 3372.8 3rd Qu.: 1949.2
Max. : 32271.6 Max. : 30000.0 Max. : 30438.1 Max. : 38281.4
> par (mfrow = c (1, 2))
```

```

> hist (food, main = 'Ιστόγραμμα', ylab = 'Συχνότητα', xlab = 'Σύνολο δαπάνης σε είδη
διατροφής\ηκαι μη αλκοολούχα ποτά', xlim = c (0, 20000), ylim = c (0, 800), labels =
TRUE, nclass = 20, cex.main = 1, cex.axis = 0.9)
> boxplot (food, main = 'Θηκογράφημα', ylab = 'Σύνολο δαπάνης σε είδη διατροφής
και μη αλκοολούχα ποτά', cex.main = 1, cex.axis = 0.9)
> par (mfrow = c (1, 2))
> hist (alcohol, main = 'Ιστόγραμμα', ylab = 'Συχνότητα', xlab = 'Σύνολο δαπάνης σε
οινοπνευματώδη ποτά\ηκαι καπνό', labels = TRUE, xlim = c (0, 14000), ylim = c (0,
3000), cex.main = 1, nclass = 13)
> boxplot (alcohol, main = 'Θηκογράφημα', ylab = 'Σύνολο δαπάνης σε
οινοπνευματώδη ποτά και καπνό', cex.main = 1, cex.axis = 0.9)
> par (mfrow = c (1, 2))
> hist (clothing, labels = TRUE, main = 'Ιστόγραμμα', xlab = 'Σύνολο δαπάνης σε είδη
ένδυσης\ηκαι υπόδησης', ylab = 'Συχνότητα', xlim = c (0, 30000), ylim = c(0, 3000),
nclass = 13)
> boxplot(clothing, main='Θηκογράφημα', ylab='Σύνολο δαπάνης σε είδη ένδυσης
και υπόδησης')
> par (mfrow = c (1, 2))
> hist (housing, labels = TRUE, main = 'Ιστόγραμμα', ylab = 'Συχνότητα', xlab =
'Σύνολο δαπάνης σε στέγαση, ύδρευση,\ηκαύσιμα και ηλεκτρισμό', ylim = c (0,
2000), xlim = c (0, 70000), cex.main = 1, nclass = 14)
> boxplot (housing, main = 'Θηκογράφημα', ylab = 'Σύνολο δαπάνης σε στέγαση,
ύδρευση, καύσιμα και ηλεκτρισμό', cex.main = 1, cex.axis = 0.9)
> par (mfrow = c (1, 2))
> hist (equipment, main = 'Ιστόγραμμα', ylab = 'Συχνότητα', xlab = 'Σύνολο δαπάνης
σε διαρκή αγαθά\ηοικιακής χρήσης', labels = TRUE, xlim = c (0, 40000), ylim = c (0,
3000), cex.main = 1, nclass = 18)
> boxplot (equipment, main = 'Θηκογράφημα', ylab = 'Σύνολο δαπάνης σε διαρκή
αγαθά οικιακής χρήσης', cex.main = 1, cex.axis = 0.9)
> par (mfrow = c (1, 2))
> hist (health, main = 'Ιστόγραμμα', ylab = 'Συχνότητα', xlab = 'Σύνολο δαπάνης σε
υγεία', labels = TRUE, ylim = c (0, 3000), xlim = c (0, 35000), nclass = 17, cex.main = 1)
> boxplot (health, main = 'Θηκογράφημα', ylab = 'Σύνολο δαπάνης σε υγεία', cex.axis
= 0.9, ylim = c (0, 35000), cex.main = 1)
> par (mfrow = c (1, 2))
> hist (transport, labels = TRUE, main = 'Ιστόγραμμα', xlab = 'Σύνολο δαπάνης σε
μεταφορές', ylab = 'Συχνότητα', xlim = c (0, 70000), ylim = c (0, 3500), nclass = 14,
cex.main = 1)
> boxplot (transport, main = 'Θηκογράφημα', ylab = 'Σύνολο δαπάνης σε
μεταφορές', cex.axis = 0.9, cex.main = 1)
> par (mfrow = c (1, 2))
> hist (communication, main = 'Ιστόγραμμα', xlab = 'Σύνολο δαπάνης σε
επικοινωνίες', ylab = 'Συχνότητα', labels = TRUE, xlim = c (0, 8000), ylim = c (0, 1500),
nclass = 16, cex.main = 1)
> boxplot (communication, main = 'Θηκογράφημα', ylab = 'Σύνολο δαπάνης σε
επικοινωνίες', ylim = c (0, 8000), cex.axis = 0.9, cex.main = 1)
> par (mfrow = c (1, 2))

```



```

> hist(culture, main='Ιστογράμμο', ylab='Συχνότητα', xlab='Σύνολο δαπάνης σε
αναψυχή\ηκαι πολιτισμό', xlim=c(0,35000), ylim=c(0,3500), labels=TRUE, nclass =
17, cex.main = 1)
> boxplot(culture, main='Θηκογράφημα', ylab='Σύνολο δαπάνης σε αναψυχή και
πολιτισμός', ylim=c(0,35000), cex.axis=0.9, cex.main = 1)
> par(mfrow = c(1, 2))
> hist(education, labels = TRUE, main = 'Ιστόγραμμα', ylab = 'Συχνότητα', xlab =
'Σύνολο δαπάνης σε εκπαίδευση', ylim = c(0, 3500), xlim = c(0, 30000), nclass = 15,
cex.main = 1, cex.axis = 0.9)
> boxplot(education, main = 'Θηκογράφημα', ylab = 'Σύνολο δαπάνης σε
εκπαίδευση', cex.main = 1)
> par(mfrow = c(1, 2))
> hist(hotels, main = 'Ιστόγραμμα', ylab = 'Συχνότητα', xlab = 'Σύνολο δαπάνης σε
εστιατόρια\ηκαι ξενοδοχεία', labels = TRUE, xlim = c(0, 35000), ylim = c(0, 2000),
nclass = 16, cex.axis = 0.9, cex.main = 1)
> boxplot(hotels, main = 'Θηκογράφημα', ylab = 'Σύνολο δαπάνης σε εστιατόρια και
ξενοδοχεία', cex.main = 1)
> par(mfrow = c(1, 2))
> hist(goods, labels=TRUE, main='Ιστόγραμμα', ylab='Συχνότητα', xlab='Σύνολο
δαπάνης σε\ηδιάφορα αγαθά και υπηρεσίες', xlim=c(0,40000), ylim=c(0,3000),
nclass = 20, cex.axis = 0.9, cex.main = 1)
> boxplot(goods, main='Θηκογράφημα', ylab='Σύνολο δαπάνης σε διάφορα αγαθά
και υπηρεσίες', ylim=c(0,40000), cex.main = 1)

```

2.2. Συσχετίσεις Μεταβλητών

Στη συνέχεια, θα εξεταστεί πιθανή συσχέτιση ή όχι μεταξύ των μεταβλητών, ποιοτικών και ποσοτικών, της εν λόγω έρευνας.

Εν γένει, για τον έλεγχο ύπαρξης συσχέτισης μεταξύ των κατηγορικών μεταβλητών, δηλαδή την περιοχή, το βαθμό πυκνότητας, την κύρια πηγή εισοδήματος, τον αριθμό των οικονομικά ενεργών μελών και την κοινωνικο – οικονομική κατάσταση του προσώπου αναφοράς του νοικοκυριού, εφαρμόστηκε ο έλεγχος ανεξαρτησίας X^2 (independence test X^2) για όλες τις μεταβλητές ανά δύο. Ο έλεγχος αυτός έχει ως μηδενική υπόθεση H_0 : οι μεταβλητές είναι ανεξάρτητες έναντι της εναλλακτικής H_1 : οι μεταβλητές είναι εξαρτημένες. Απαραίτητη προϋπόθεση, όμως, για να εφαρμοσθεί ο εν λόγω έλεγχος είναι όλες οι αναμενόμενες συχνότητες να είναι ≥ 5 . Το στατιστικό ελέγχου, το οποίο υπολογίζεται είναι:

$$X^2 = \sum \frac{(\text{παρατηρηθείσες συχνότητες} - \text{αναμενόμενες συχνότητες})^2}{\text{αναμενόμενες συχνότητες}}$$

όπου ισχύει ότι οι:

$$\text{αναμενόμενες συχνότητες} = \frac{(\text{άθροισμα γραμμής}) \times (\text{άθροισμα στήλης})}{\text{μέγεθος δείγματος}}$$

και το άθροισμα είναι ως προς όλα τα κελιά. Το στατιστικό αυτό ακολουθεί προσεγγιστικά την χ^2 κατανομή με 1 βαθμό ελευθερίας, κάτω από τη μηδενική υπόθεση.

Ωστόσο, όταν δεν ικανοποιούνται οι προϋποθέσεις για τον έλεγχο ανεξαρτησίας χ^2 , τότε εφαρμόζεται ο ακριβής έλεγχος του Fisher (Fisher's exact test). Στον έλεγχο αυτό, η κάθε συχνότητα, δεδομένων των περιθωρίων συχνοτήτων, ακολουθεί την υπεργεωμετρική κατανομή.

Από την άλλη πλευρά, για τον έλεγχο ύπαρξης συσχέτισης μεταξύ των ποσοτικών μεταβλητών, δηλαδή τη συνολική καταναλωτική δαπάνη και τις 12 επιμέρους κατηγορίες δαπανών, υπολογίσθηκε ο συντελεστής γραμμικής συσχέτισης Pearson ρ κάνοντας χρήση του πακέτου Hmisc, ώστε να γίνει ταυτόχρονα και ο αντίστοιχος έλεγχος υπόθεσης ότι ο συντελεστής συσχέτισης είναι διάφορος του μηδενός. Ο έλεγχος αυτός έχει ως μηδενική υπόθεση H_0 ότι οι μεταβλητές είναι ασυσχέτιστες έναντι της εναλλακτικής H_1 ότι υπάρχει γραμμική συσχέτιση. Ο συντελεστής γραμμικής συσχέτισης του Pearson εκτιμάται από τον αντίστοιχο δειγματικό συντελεστή συσχέτισης, ο οποίος ισούται με:

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}} , \quad i = 1, 2, \dots, n \quad (2.1)$$

όπου τα x_i, y_i αντιπροσωπεύουν τις τιμές του δείγματος που λαμβάνουν οι αντίστοιχες μεταβλητές X, Y και τα $\bar{x}, \bar{y}$ τους δειγματικούς μέσους των εν λόγω μεταβλητών.

Η τιμή του συντελεστή γραμμικής συσχέτισης του Pearson κυμαίνεται μεταξύ -1 και +1 και δεν επηρεάζεται από μεταβολές στην κλίμακα των μεταβλητών. Το πρόσημο του συντελεστή αυτού δείχνει αν υπάρχει αρνητική ή θετική γραμμική συσχέτιση. Αν $r > 0$ τότε οι δύο μεταβλητές X, Y συσχετίζονται θετικά, δηλαδή η αύξηση των τιμών της X συνεπάγεται κατά μέσο όρο αύξηση των τιμών της Y . Όσο πιο κοντά στην τιμή 1 είναι ο εν λόγω συντελεστής, τόσο ισχυρότερη είναι η θετική γραμμική συσχέτιση των μεταβλητών. Αν $r < 0$ τότε οι δύο μεταβλητές X, Y συσχετίζονται αρνητικά, δηλαδή η αύξηση των τιμών της X συνεπάγεται κατά μέσο όρο μείωση των τιμών της Y . Όσο πιο κοντά είναι στο -1 ο εν λόγω συντελεστής, τόσο ισχυρότερη είναι η αρνητική γραμμική συσχέτιση των μεταβλητών. Τέλος, αν $r \approx 0$ τότε οι δύο μεταβλητές θεωρούνται ασυσχέτιστες, δηλαδή δεν υπάρχει γραμμική σχέση μεταξύ τους.

Είναι αξιοσημείωτο να διευκρινιστεί ότι, κατά την ερμηνεία του συντελεστή του Pearson, η ισχυρή συσχέτιση δε σημαίνει κατ' ανάγκη ότι υπάρχει σχέση αιτίου – αιτιατού μεταξύ των δύο συσχετιζόμενων μεταβλητών.

Επιπρόσθετα, είναι σημαντικό να γίνει και ο έλεγχος (έστω σε επίπεδο σημαντικότητας α) ότι οι δύο τυχαίες μεταβλητές είναι ασυσχέτιστες με εναλλακτική ότι δεν είναι ή με άλλη διατύπωση, να γίνει ο έλεγχος για την $H_0: \rho = 0$ με εναλλακτική την $H_1: \rho \neq 0$. Κάτω από τη μηδενική υπόθεση ισχύει ότι:

$$T = \frac{r}{\sqrt{1-r^2}} \sqrt{n-2} \sim St(n-2) \quad (2.2)$$

Βάσει, λοιπόν, των πινάκων της Student κατανομής με $(n-2)$ βαθμούς ελευθερίας υπολογίζεται η P – τιμή του εν λόγω ελέγχου, η οποία είναι ίση με 2 φορές την πιθανότητα δεξιά της παρατηρηθείσας τιμής του $|T|$.

2.2.1. Περιοχή – Κύρια πηγή εισοδήματος

Πίνακας 2.9: Πίνακας συνάφειας περιοχής – κύριας πηγής εισοδήματος

Περιοχή	Κύρια πηγή εισοδήματος							Σύνολο
		Μισθοί	Αυτοαπασχόληση	Ακίνητη περιουσία	Συντάξεις	Επιδόματα ανεργίας	Άλλα επιδόματα	
Περιοχή	Βόρεια Ελλάδα	299	194	24	530	7	51	1105
		0.271	0.176	0.022	0.480	0.006	0.046	0.315
		0.249	0.365	0.304	0.344	0.350	0.364	
		0.085	0.055	0.007	0.151	0.002	0.015	
	Κεντρική Ελλάδα	176	153	10	397	3	35	774
		0.227	0.198	0.013	0.513	0.004	0.045	0.220
		0.147	0.288	0.127	0.258	0.150	0.250	
		0.050	0.044	0.003	0.113	0.001	0.010	
	Αττική	621	136	35	455	6	47	1300
		0.478	0.105	0.027	0.350	0.005	0.036	0.370
		0.517	0.256	0.443	0.295	0.300	0.336	
		0.177	0.039	0.010	0.130	0.002	0.013	
	Νησιά Αιγαίου - Κρήτη	105	48	10	159	4	7	333
		0.315	0.144	0.030	0.477	0.012	0.021	0.095
		0.087	0.090	0.127	0.103	0.200	0.050	
		0.030	0.014	0.003	0.045	0.001	0.002	
	Σύνολο	1201	531	79	1541	20	140	3512
		0.342	0.151	0.022	0.439	0.006	0.040	

Σε κάθε κελί:

Συχνότητα

Σχετική συχνότητα γραμμής

Σχετική συχνότητα στήλης

Σχετική συχνότητα κελιού

Από τον Πίνακα 2.9 προκύπτει ότι:

- Από το 31.5% του συνόλου των νοικοκυριών που κατοικεί στη Βόρεια Ελλάδα το 27.1% έχει ως κύρια πηγή εισοδήματος το μισθό, το 17.6% τα εισοδήματα από αυτοαπασχόληση, το 2.2% τα εισοδήματα από ακίνητη περιουσία, το 48% τη σύνταξη, το 0.6% το επίδομα ανεργίας και το 4.6% άλλα επιδόματα.
- Από το 22% του συνόλου των νοικοκυριών που κατοικεί στην Κεντρική Ελλάδα το 22.7% έχει ως κύρια πηγή εισοδήματος το μισθό, το 19.8% τα εισοδήματα από αυτοαπασχόληση, το 1.3% τα εισοδήματα από ακίνητη περιουσία, το 51.3% τη σύνταξη, το 0.4% το επίδομα ανεργίας και το 4.5% άλλα επιδόματα.
- Από το 37% του συνόλου των νοικοκυριών που κατοικεί στην Αττική το 47.8% έχει ως κύρια πηγή εισοδήματος το μισθό, το 10.5% τα εισοδήματα από αυτοαπασχόληση, το 2.7% τα εισοδήματα από ακίνητη περιουσία, το 35% τη σύνταξη, το 0.5% το επίδομα ανεργίας και το 3.6% άλλα επιδόματα.
- Από το 9.5% του συνόλου των νοικοκυριών που κατοικεί στα Νησιά Αιγαίου – Κρήτη το 31.5% έχει ως κύρια πηγή εισοδήματος το μισθό, το 14.4% τα

εισοδήματα από αυτοαπασχόληση, το 0.3% τα εισοδήματα από ακίνητη περιουσία, το 47.7% τη σύνταξη, το 1.2% το επίδομα ανεργίας και το 2.1% άλλα επιδόματα.

Κατά συνέπεια, σε κάθε περιοχή της Ελλάδας περίπου το 50% των νοικοκυριών δηλώνουν ως κύρια πηγή εισοδήματος τη σύνταξη με εξαίρεση την Αττική όπου το μεγαλύτερο ποσοστό των νοικοκυριών δηλώνει το μισθό.

Όσον αφορά στη συσχέτιση των δύο μεταβλητών, χρησιμοποιώντας τον ακριβή έλεγχο του Fisher, υπολογίζεται η p-τιμή του αμφίπλευρου ελέγχου ότι είναι 0.0004998 σε επίπεδο σημαντικότητας 5%, οπότε υπάρχουν σοβαρές ενδείξεις για να απορριφθεί η μηδενική υπόθεση. Άρα, υπάρχει συσχέτιση μεταξύ της περιοχής και της κύριας πηγής εισοδήματος του νοικοκυριού.

Εντολές στην R

Με την εντολή `CrossTable()` δίνεται ο πίνακας συνάφειας των δύο μεταβλητών όπου για κάθε κελί υπολογίζονται η συχνότητα, η σχετική συχνότητα της γραμμής, η σχετική συχνότητα της στήλης και η σχετική συχνότητά του σε όλο τον πίνακα. Επιπρόσθετα, υπολογίζεται το σύνολο ανά γραμμή και ανά στήλη. Εφαρμόζοντας τον έλεγχο ανεξαρτησίας X^2 στα δεδομένα με την εντολή `chisq.test()` υπήρχε ένα προειδοποιητικό μήνυμα λάθους, γεγονός που σημαίνει ότι η προσέγγιση μπορεί να μην είναι αξιόπιστη, αφού δεν ισχύει η προϋπόθεση ότι όλες οι αναμενόμενες συχνότητες είναι μεγαλύτερες του 5. Γι' αυτό το λόγο, εφαρμόζεται ο μη παραμετρικός ακριβής έλεγχος του Fisher με την εντολή `fisher.test()`. Τέλος, με την εντολή `expected()` διαπιστώνεται ότι όντως υπήρχαν αναμενόμενες συχνότητες μικρότερες του 5.

```
> CrossTable (table (region, source))
> c <- chisq.test (table (region, source))
Warning message:
In chisq.test (table (region, source)) :
  Chi-squared approximation may be incorrect
> c
      Pearson's Chi-squared test
data:  table (region, source)
X-squared = 202.8046, df = 15, p-value < 2.2e-16
> fisher.test (table (region, source), simulate.p.value = TRUE)
      Fisher's Exact Test for Count Data with simulated p-value (based on
      2000 replicates)

data:  table (region, source)
p-value = 0.0004998
alternative hypothesis: two.sided
> attach (c)
> expected
      source
region  1    2    3    4    5    6
```

1 377.8773 167.07147 24.856207 484.8534 6.292711 44.04897
 2 264.6851 117.02563 17.410592 339.6167 4.407745 30.85421
 3 444.5615 196.55467 29.242597 570.4157 7.403189 51.82232
 4 113.8761 50.34823 7.490604 146.1142 1.896355 13.27449

2.2.2. Περιοχή – Βαθμός πυκνότητας

Πίνακας 2.10: Πίνακας συνάφειας περιοχής – βαθμού πυκνότητας

Περιοχή	Βαθμός πυκνότητας				Σύνολο
		> 500 κάτοικοι/km ²	100 - 499 κάτοικοι/km ²	< 100 κάτοικοι/km ²	
Περιοχή	Βόρεια Ελλάδα	345	139	621	1105
		0.312	0.126	0.0562	0.315
		0.232	0.308	0.395	
		0.098	0.040	0.177	
	Κεντρική Ελλάδα	156	15	603	774
		0.202	0.019	0.779	0.220
		0.105	0.033	0.384	
		0.044	0.004	0.172	
	Αττική	900	283	117	1300
		0.692	0.218	0.090	0.370
		0.604	0.627	0.074	
		0.256	0.081	0.033	
	Νησιά Αιγαίου - Κρήτη	88	14	231	333
		0.264	0.042	0.694	0.095
		0.059	0.031	0.147	
		0.025	0.004	0.066	
	Σύνολο	1489	451	1572	3512
		0.424	0.128	0.448	

Σε κάθε κελί:

Συχνότητα

Σχετική συχνότητα γραμμής

Σχετική συχνότητα στήλης

Σχετική συχνότητα κελιού

Από τον Πίνακα 2.10 προκύπτει ότι:

- Από το 31.5% του συνόλου των νοικοκυριών που κατοικεί στη Βόρεια Ελλάδα το 31.2% ζει σε πυκνοκατοικημένη περιοχή, το 12.6% ζει σε ενδιάμεσης πυκνότητας περιοχή και το 56.2% σε αραιοκατοικημένη περιοχή.
- Από το 22% του συνόλου των νοικοκυριών που κατοικεί στην Κεντρική Ελλάδα το 20.2% ζει σε πυκνοκατοικημένη περιοχή, το 1.9% σε ενδιάμεσης πυκνότητας περιοχή και το 77.9% σε αραιοκατοικημένη περιοχή.
- Από το 37% του συνόλου των νοικοκυριών που κατοικεί στην Αττική το 69.2% ζει σε πυκνοκατοικημένη περιοχή, το 21.8% σε ενδιάμεσης πυκνότητας περιοχή και το 9% σε αραιοκατοικημένη περιοχή.
- Από το 9.5% του συνόλου των νοικοκυριών, που κατοικεί στα Νησιά Αιγαίου – Κρήτη, το 26.4% ζει σε πυκνοκατοικημένη περιοχή, το 4.2% σε ενδιάμεσης πυκνότητας περιοχή και το 69.4% σε αραιοκατοικημένη περιοχή.

Κατά συνέπεια, σε κάθε περιοχή το μεγαλύτερο ποσοστό των νοικοκυριών ζει σε αραιοκατοικημένες περιοχές με εξαίρεση την Αττική όπου ζει σε πυκνοκατοικημένες, όπως είναι και προφανές λόγω της μεγάλης συγκέντρωσης του πληθυσμού στην περιοχή αυτή.

Όσον αφορά στη συσχέτιση των δύο μεταβλητών, χρησιμοποιώντας τον έλεγχο ανεξαρτησίας χ^2 , υπολογίζεται η p – τιμή του αμφίπλευρου ελέγχου ότι είναι $< 2.2e-16$ σε επίπεδο σημαντικότητας 5%, οπότε υπάρχουν σοβαρές ενδείξεις για να απορριφθεί η μηδενική υπόθεση. Άρα, υπάρχει συσχέτιση μεταξύ της περιοχής και του βαθμού πυκνότητας.

Εντολές στην R

Με την εντολή `CrossTable()` δίνεται ο πίνακας συνάφειας των δύο μεταβλητών όπου για κάθε κελί υπολογίζονται η συχνότητα, η σχετική συχνότητα της γραμμής, η σχετική συχνότητα της στήλης και η σχετική συχνότητά του σε όλο τον πίνακα. Επιπρόσθετα, υπολογίζεται το σύνολο ανά γραμμή και ανά στήλη. Στη συνέχεια, εφαρμόζεται ο έλεγχος ανεξαρτησίας χ^2 στα δεδομένα με την εντολή `chisq.test()`. Τέλος, με την εντολή `expected` επιβεβαιώνεται ότι, πράγματι, όλες οι αναμενόμενες συχνότητες είναι μεγαλύτερες του 5.

```
> CrossTable (table (region, density))
> c <- chisq.test (table (region, density))
> c
      Pearson's Chi-squared test
data:  table (region, density)
X-squared = 1174.173, df = 6, p-value < 2.2e-16
> attach (c)
> expected
      density
region    1    2    3
  1 468.4923 141.90063 494.6071
  2 328.1566  99.39465 346.4487
  3 551.1674 166.94191 581.8907
  4 141.1837  42.76281 149.0535
```

2.2.3. Περιοχή – Αριθμός οικονομικά ενεργών μελών του νοικοκυριού

Πίνακας 2.11: Πίνακας συνάφειας περιοχής – αριθμού οικονομικά ενεργών μελών του νοικοκυριού

		Αριθμός οικονομικά ενεργών μελών του νοικοκυριού						
Περιοχή		0	1	2	3	4	> 5	Σύνολο
	Βόρεια Ελλάδα	132 0.119 0.338 0.038	474 0.429 0.303 0.135	417 0.377 0.324 0.119	55 0.050 0.271 0.016	27 0.024 0.422 0.008	0 0.0 0.0 0.0	1105 0.315
	Κεντρική Ελλάδα	86 0.111 0.220 0.024	307 0.397 0.197 0.087	305 0.394 0.237 0.087	57 0.074 0.281 0.016	17 0.022 0.266 0.005	2 0.003 0.400 0.001	774 0.220
	Αττική	146 0.112 0.373 0.042	625 0.481 0.400 0.178	430 0.331 0.334 0.122	79 0.061 0.389 0.022	18 0.014 0.281 0.005	2 0.002 0.400 0.001	1300 0.370
	Νησιά Αιγαίου - Κρήτη	27 0.081 0.069 0.008	156 0.468 0.100 0.044	135 0.405 0.105 0.038	12 0.036 0.059 0.003	2 0.006 0.031 0.001	1 0.003 0.200 0.000	333 0.095
	Σύνολο	391 0.111	1562 0.445	1287 0.366	203 0.058	64 0.018	5 0.001	3512

Σε κάθε κελί:

Συχνότητα

Σχετική συχνότητα γραμμής

Σχετική συχνότητα στήλης

Σχετική συχνότητα κελιού

Από τον Πίνακα 2.11 προκύπτει ότι:

- Από το 31.5% του συνόλου των νοικοκυριών που κατοικεί στη Βόρεια Ελλάδα το 11.9% δεν έχει κανένα οικονομικά ενεργό μέλος, το 42.9% έχει ένα οικονομικά ενεργό μέλος, το 37.7% έχει δύο οικονομικά ενεργά μέλη, το 5% έχει τρία οικονομικά ενεργά μέλη, το 2.4% έχει τέσσερα οικονομικά ενεργά μέλη και δεν υπάρχουν νοικοκυριά με πέντε ή περισσότερα οικονομικά ενεργά μέλη.
- Από το 22% του συνόλου των νοικοκυριών που κατοικεί στην Κεντρική Ελλάδα το 11.1% δεν έχει κανένα οικονομικά ενεργό μέλος, το 39.7% έχει ένα οικονομικά ενεργό μέλος, το 39.4% έχει δύο οικονομικά ενεργά μέλη, το 7.4% έχει τρία οικονομικά ενεργά μέλη, το 2.2% έχει τέσσερα οικονομικά ενεργά μέλη και το 0.3% έχει πέντε ή περισσότερα οικονομικά ενεργά μέλη.
- Από το 37% του συνόλου των νοικοκυριών που κατοικεί στην Αττική το 11.2% δεν έχει κανένα οικονομικά ενεργό μέλος, το 48.1% έχει ένα οικονομικά ενεργό μέλος, το 33.1% έχει δύο οικονομικά ενεργά μέλη, το 6.1% έχει τρία οικονομικά ενεργά μέλη, το 1.4% έχει τέσσερα οικονομικά ενεργά μέλη και το 0.2% έχει πέντε ή περισσότερα οικονομικά ενεργά μέλη.

- Από το 9.5% του συνόλου των νοικοκυριών που κατοικεί στα Νησιά Αιγαίου – Κρήτη το 8.1% δεν έχει κανένα οικονομικά ενεργό μέλος, το 46.8% έχει ένα οικονομικά ενεργό μέλος, το 40.5% έχει δύο οικονομικά ενεργά μέλη, το 3.6% έχει τρία οικονομικά ενεργά μέλη, το 0.6% έχει τέσσερα οικονομικά ενεργά μέλη και το 0.3% έχει πέντε ή περισσότερα οικονομικά ενεργά μέλη.

Κατά συνέπεια, σε κάθε περιοχή το μεγαλύτερο ποσοστό των νοικοκυριών έχει ένα ή δύο οικονομικά ενεργά μέλη.

Όσον αφορά στη συσχέτιση των δύο μεταβλητών, χρησιμοποιώντας τον ακριβή έλεγχο του Fisher, υπολογίζεται η p – τιμή του αμφίπλευρου ελέγχου ότι είναι 0.001999 σε επίπεδο σημαντικότητας 5%, οπότε υπάρχουν σοβαρές ενδείξεις για να απορριφθεί η μηδενική υπόθεση. Άρα, υπάρχει συσχέτιση μεταξύ της περιοχής και του αριθμού των οικονομικά ενεργών μελών του νοικοκυριού.

Εντολές στην R

Με την εντολή `CrossTable()` δίνεται ο πίνακας συνάφειας των δύο μεταβλητών όπου για κάθε κελί υπολογίζονται η συχνότητα, η σχετική συχνότητα της γραμμής, η σχετική συχνότητα της στήλης και η σχετική συχνότητά του σε όλο τον πίνακα. Επιπρόσθετα, υπολογίζεται το σύνολο ανά γραμμή και ανά στήλη. Εφαρμόζοντας τον έλεγχο ανεξαρτησίας X^2 στα δεδομένα με την εντολή `chisq.test()` υπήρχε ένα προειδοποιητικό μήνυμα λάθους, γεγονός που σημαίνει ότι η προσέγγιση μπορεί να μην είναι αξιόπιστη, αφού δεν ισχύει η προϋπόθεση ότι όλες οι αναμενόμενες συχνότητες είναι μεγαλύτερες του 5. Γι' αυτό το λόγο, εφαρμόζεται ο μη παραμετρικός ακριβής έλεγχος του Fisher με την εντολή `fisher.test()`. Τέλος, με την εντολή `expected()` διαπιστώνεται ότι όντως υπήρχαν αναμενόμενες συχνότητες μικρότερες του 5.

```
> CrossTable (table (region, members))
> c <- chisq.test (table (region, members))
Warning message:
In chisq.test (table (region, members)) :
  Chi-squared approximation may be incorrect
> c
      Pearson's Chi-squared test
data:  table (region, members)
X-squared = 37.5777, df = 15, p-value = 0.001041
> fisher.test (table (region, members), simulate.p.value = TRUE)
      Fisher's Exact Test for Count Data with simulated p-value (based on
      2000 replicates)
data:  table (region, members)
p-value = 0.001999
alternative hypothesis: two.sided
> attach(c)
> expected
      members
region    0    1    2    3    4    5
```


1 123.02249 491.4607 404.9359 63.87101 20.136674 1.5731777
 2 86.17141 344.2449 283.6384 44.73861 14.104784 1.1019362
 3 144.73235 578.1891 476.3952 75.14237 23.690205 1.8507973
 4 37.07375 148.1054 122.0305 19.24801 6.068337 0.4740888

2.2.4. Περιοχή – Κοινωνικο – οικονομική κατάσταση του προσώπου αναφοράς του νοικοκυριού

Πίνακας 2.12: Πίνακας συνάφειας περιοχής – κοινωνικο – οικονομικής κατάστασης του προσώπου αναφοράς του νοικοκυριού

		Κοινωνικο – οικονομική κατάσταση του προσώπου αναφοράς του νοικοκυριού							
Περιοχή		Χειρώνακτας	Μη χειρώνα- κτας	Αυτο- απασχο- λούμενος	Άνεργος	Συντα- ξιούχος	Λοιποί μη οικονομικά ενεργοί	Δεν έχει απαντηθεί	Σύνολο
	Βόρεια Ελλάδα	119	87	214	46	499	138	2	1105
		0.108	0.079	0.194	0.042	0.452	0.125	0.002	0.315
		0.215	0.255	0.347	0.338	0.350	0.317	0.667	
		0.034	0.025	0.061	0.013	0.142	0.039	0.001	
	Κεντρική Ελλάδα	82	51	169	16	360	96	0	774
		0.106	0.066	0.218	0.021	0.465	0.124	0.0	0.220
		0.148	0.150	0.274	0.118	0.252	0.221	0.0	
		0.023	0.015	0.048	0.005	0.103	0.027	0.0	
	Αττική	302	173	176	63	420	166	0	1300
0.232		0.133	0.135	0.048	0.323	0.128	0.0	0.370	
0.546		0.507	0.285	0.463	0.294	0.382	0.0		
0.086		0.049	0.050	0.018	0.120	0.047	0.0		
Νησιά Αιγαίου - Κρήτη	50	30	58	11	148	35	1	333	
	0.150	0.090	0.174	0.033	0.444	0.105	0.003	0.095	
	0.090	0.088	0.094	0.081	0.104	0.080	0.333		
	0.014	0.009	0.017	0.003	0.042	0.010	0.0		
Σύνολο	553	341	617	136	1427	435	3	3512	
	0.157	0.097	0.176	0.039	0.406	0.124	0.001		

Σε κάθε κελί:

Συχνότητα

Σχετική συχνότητα γραμμής

Σχετική συχνότητα στήλης

Σχετική συχνότητα κελιού

Από τον Πίνακα 2.12 προκύπτει ότι:

- Από το 31.5% του συνόλου των νοικοκυριών, που κατοικεί στη Βόρεια Ελλάδα, το 10.8% έχει ως πρόσωπο αναφοράς χειρώνακτα εργαζόμενο, το 7.9% έχει μη χειρώνακτα εργαζόμενο, το 19.4% έχει αυτοαπασχολούμενο, γεωργό ή κτηνοτρόφο, το 4.2% έχει άνεργο, το 45.2% έχει συνταξιούχο, το 12.5% έχει άλλους μη οικονομικά ενεργούς και 0.2% δεν έχει απαντήσει.
- Από το 22% του συνόλου των νοικοκυριών που κατοικεί στην Κεντρική Ελλάδα το 10.6% έχει ως πρόσωπο αναφοράς χειρώνακτα εργαζόμενο, το 6.6% έχει μη χειρώνακτα εργαζόμενο, το 21.8% έχει αυτοαπασχολούμενο, γεωργό ή κτηνοτρόφο, το 2.1% έχει άνεργο, το 46.5% έχει συνταξιούχο, το 12.4% έχει άλλους μη οικονομικά ενεργούς και δεν υπάρχουν νοικοκυριά που δεν έχουν απαντήσει.
- Από το 37% του συνόλου των νοικοκυριών που κατοικεί στην Αττική το 23.2% έχει ως πρόσωπο αναφοράς χειρώνακτα εργαζόμενο, το 13.3% έχει μη χειρώνακτα εργαζόμενο, το 13.5% έχει αυτοαπασχολούμενο, γεωργό ή κτηνοτρόφο, το 4.8% έχει άνεργο, το 32.3% έχει συνταξιούχο, το 12.8% έχει άλλους μη οικονομικά ενεργούς και δεν υπάρχουν νοικοκυριά που δεν έχουν απαντήσει.

- Από το 9.5% του συνόλου των νοικοκυριών που κατοικεί στα Νησιά Αιγαίου – Κρήτη το 15% έχει ως πρόσωπο αναφοράς χειρώνακτα εργαζόμενο, το 9% έχει μη χειρώνακτα εργαζόμενο, το 17.4% έχει αυτοαπασχολούμενο, γεωργό ή κτηνοτρόφο, το 3.3% έχει άνεργο, το 44.4% έχει συνταξιούχο, το 10.5% έχει άλλους μη οικονομικά ενεργούς και 0.3% που δεν έχουν απαντήσει.

Κατά συνέπεια, σε κάθε περιοχή το μεγαλύτερο ποσοστό των νοικοκυριών έχει συνταξιούχο ως πρόσωπο αναφοράς.

Όσον αφορά στη συσχέτιση των δύο μεταβλητών, χρησιμοποιώντας τον ακριβή έλεγχο του Fisher, υπολογίζεται η p – τιμή του αμφίπλευρου ελέγχου ότι είναι 0.0004998 σε επίπεδο σημαντικότητας 5%, οπότε υπάρχουν σοβαρές ενδείξεις για να απορριφθεί η μηδενική υπόθεση. Άρα, υπάρχει συσχέτιση μεταξύ της περιοχής και της κοινωνικο – οικονομικής κατάστασης του προσώπου αναφοράς του νοικοκυριού.

Εντολές στην R

Με την εντολή `CrossTable()` δίνεται ο πίνακας συνάφειας των δύο μεταβλητών όπου για κάθε κελί υπολογίζονται η συχνότητα, η σχετική συχνότητα της γραμμής, η σχετική συχνότητα της στήλης και η σχετική συχνότητά του σε όλο τον πίνακα. Επιπρόσθετα, υπολογίζεται το σύνολο ανά γραμμή και ανά στήλη. Εφαρμόζοντας τον έλεγχο ανεξαρτησίας χ^2 στα δεδομένα με την εντολή `chisq.test()` υπήρχε ένα προειδοποιητικό μήνυμα λάθους, γεγονός που σημαίνει ότι η προσέγγιση μπορεί να μην είναι αξιόπιστη, αφού δεν ισχύει η προϋπόθεση ότι όλες οι αναμενόμενες συχνότητες είναι μεγαλύτερες του 5. Γι' αυτό το λόγο, εφαρμόζεται ο μη παραμετρικός ακριβής έλεγχος του Fisher με την εντολή `fisher.test()`. Τέλος, με την εντολή `expected()` διαπιστώνεται ότι όντως υπήρχαν αναμενόμενες συχνότητες μικρότερες του 5.

```
> CrossTable (table (region, sitnew))
> c <- chisq.test (table (region, sitnew))
Warning message:
In chisq.test (table (region, sitnew)) :
  Chi-squared approximation may be incorrect
> c
  Pearson's Chi-squared test
data: table (region, sitnew)
X-squared = 179.5645, df = 18, p-value < 2.2e-16
> fisher.test (table (region, sitnew), simulate.p.value = TRUE)
  Fisher's Exact Test for Count Data with simulated p-value (based on
  2000 replicates)
data: table (region, sitnew)
p-value = 0.0004998
alternative hypothesis: two.sided
> attach (c)
> expected
  sitnew
```

region 1 2 3 4 5 6 7
 1 173.99345 107.29072 194.13013 42.79043 448.9849 136.86646 0.9439066
 2 121.87415 75.15205 135.97893 29.97267 314.4926 95.86845 0.6611617
 3 204.69818 126.22437 228.38838 50.34169 528.2175 161.01936 1.1104784
 4 52.43423 32.33286 58.50256 12.89522 135.3050 41.24573 0.2844533

2.2.5. Κύρια πηγή εισοδήματος – Βαθμός πυκνότητας

Πίνακας 2.13: Πίνακας συνάφειας κύριας πηγής εισοδήματος – βαθμού πυκνότητας

Κύρια πηγή εισοδήματος	Βαθμός πυκνότητας				Σύνολο
		> 500 κάτοικοι/km ²	100 - 499 κάτοικοι/km ²	< 100 κάτοικοι/km ²	
Μισθοί		651	180	370	1201
		0.542	0.150	0.308	0.342
		0.437	0.399	0.235	
		0.182	0.051	0.105	
	Αυτοαπασχόληση	156	57	318	531
		0.294	0.107	0.599	0.151
		0.105	0.126	0.202	
		0.044	0.016	0.091	
Ακίνητη περιουσία		46	9	24	79
		0.582	0.114	0.304	0.022
		0.031	0.020	0.015	
		0.013	0.003	0.007	
Συντάξεις		533	188	820	1541
		0.346	0.122	0.532	0.439
		0.358	0.417	0.522	
		0.152	0.054	0.233	
Επιδόματα ανεργίας		9	2	9	20
		0.450	0.100	0.450	0.006
		0.006	0.004	0.006	
		0.003	0.001	0.003	
Άλλα επιδόματα		94	15	31	140
		0.671	0.107	0.221	0.040
		0.063	0.033	0.020	
		0.027	0.004	0.009	
Σύνολο		1489	451	1572	3512
		0.424	0.128	0.448	

Σε κάθε κελί:

Συχνότητα

Σχετική συχνότητα γραμμής

Σχετική συχνότητα στήλης

Σχετική συχνότητα κελιού

Από τον Πίνακα 2.13 προκύπτει ότι:

- Από το 34.2% του συνόλου των νοικοκυριών που έχει ως κύρια πηγή εισοδήματος το μισθό το 54.2% ζει σε πυκνοκατοικημένη περιοχή, το 15% ζει σε ενδιάμεσης πυκνότητας περιοχή και το 30.8% ζει σε αραιοκατοικημένη περιοχή.

- Από το 15.1% του συνόλου των νοικοκυριών που έχει ως κύρια πηγή το εισόδημα από αυτοαπασχόληση το 29.4% ζει σε πυκνοκατοικημένη περιοχή, το 10.7% ζει σε ενδιάμεσης πυκνότητας περιοχή και το 59.9% ζει σε αραιοκατοικημένη περιοχή.
- Από το 2.2% του συνόλου των νοικοκυριών που έχει ως κύρια πηγή το εισόδημα από ακίνητη περιουσία το 58.2% ζει σε πυκνοκατοικημένη περιοχή, το 11.4% ζει σε ενδιάμεσης πυκνότητας περιοχή και το 30.4% ζει σε αραιοκατοικημένη περιοχή.
- Από το 43.9% του συνόλου των νοικοκυριών που έχει ως κύρια πηγή εισοδήματος τη σύνταξη το 34.6% ζει σε πυκνοκατοικημένη περιοχή, το 12.2% ζει σε ενδιάμεσης πυκνότητας περιοχή και το 53.2% ζει σε αραιοκατοικημένη περιοχή.
- Από το 0.6% του συνόλου των νοικοκυριών που έχει ως κύρια πηγή εισοδήματος το επίδομα ανεργίας το 45% ζει σε πυκνοκατοικημένη περιοχή, το 10% ζει σε ενδιάμεσης πυκνότητας περιοχή και το 45% ζει σε αραιοκατοικημένη περιοχή.
- Από το 4% του συνόλου των νοικοκυριών που έχει ως κύρια πηγή εισοδήματος άλλα επιδόματα το 67.1% ζει σε πυκνοκατοικημένη περιοχή, το 10.7% ζει σε ενδιάμεσης πυκνότητας περιοχή και το 22.1% ζει σε αραιοκατοικημένη περιοχή.

Κατά συνέπεια, το μεγαλύτερο ποσοστό των νοικοκυριών με κύρια πηγή εισοδήματος τους μισθούς, τα εισοδήματα από ακίνητη περιουσία και άλλα επιδόματα ζει σε πυκνοκατοικημένες περιοχές. Ενώ, το μεγαλύτερο ποσοστό των νοικοκυριών με κύρια πηγή εισοδήματος τη σύνταξη και τα εισοδήματα από αυτοαπασχόληση ζει σε αραιοκατοικημένες περιοχές. Τέλος, τα νοικοκυριά με κύρια πηγή εισοδήματος τα επιδόματα ανεργίας ζουν τόσο σε πυκνοκατοικημένες όσο και σε αραιοκατοικημένες περιοχές.

Όσον αφορά στη συσχέτιση των δύο μεταβλητών, χρησιμοποιώντας τον ακριβή έλεγχο του Fisher, υπολογίζεται η p – τιμή του αμφίπλευρου ελέγχου ότι είναι 0.0004998 σε επίπεδο σημαντικότητας 5%, οπότε υπάρχουν σοβαρές ενδείξεις για να απορριφθεί η μηδενική υπόθεση. Άρα, υπάρχει συσχέτιση μεταξύ της κύριας πηγής εισοδήματος του νοικοκυριού και του βαθμού πυκνότητας.

Εντολές στην R

Με την εντολή `CrossTable()` δίνεται ο πίνακας συνάφειας των δύο μεταβλητών όπου για κάθε κελί υπολογίζονται η συχνότητα, η σχετική συχνότητα της γραμμής, η σχετική συχνότητα της στήλης και η σχετική συχνότητά του σε όλο τον πίνακα. Επιπρόσθετα, υπολογίζεται το σύνολο ανά γραμμή και ανά στήλη. Εφαρμόζοντας τον έλεγχο ανεξαρτησίας X^2 στα δεδομένα με την εντολή `chisq.test()` υπήρχε ένα προειδοποιητικό μήνυμα λάθους, γεγονός που σημαίνει ότι η προσέγγιση μπορεί να μην είναι αξιόπιστη, αφού δεν ισχύει η προϋπόθεση ότι όλες οι αναμενόμενες συχνότητες είναι μεγαλύτερες του 5. Γι' αυτό το λόγο, εφαρμόζεται ο μη παραμετρικός ακριβής έλεγχος του Fisher με την εντολή `fisher.test()`. Τέλος, με την

εντολή `expected()` διαπιστώνεται ότι όντως υπήρχαν αναμενόμενες συχνότητες μικρότερες του 5.

```
> CrossTable (table (source, density))
> c <- chisq.test (table (source, density))
Warning message:
In chisq.test (table (source, density)) :
  Chi-squared approximation may be incorrect
> c
      Pearson's Chi-squared test
data:  table (source, density)
X-squared = 238.8158, df = 10, p-value < 2.2e-16
> fisher.test (table (source, density), simulate.p.value = TRUE)
      Fisher's Exact Test for Count Data with simulated p-value (based on
      2000 replicates)
data:  table(source, density)
p-value = 0.0004998
alternative hypothesis: two.sided
> attach (c)
> expected
      density
source      1      2      3
  1 509.193907 154.228645 537.577449
  2 225.130695  68.189351 237.679954
  3  33.494021  10.144932  35.361048
  4 653.345387 197.890376 689.764237
  5  8.479499  2.568337  8.952164
  6 59.356492 17.978360 62.665148
```

2.2.6. Κύρια πηγή εισοδήματος – Αριθμός οικονομικά ενεργών μελών του νοικοκυριού

Πίνακας 2.14: Πίνακας συνάφειας κύριας πηγής εισοδήματος – αριθμού οικονομικά ενεργών μελών του νοικοκυριού

Κύρια πηγή εισοδήματος	Αριθμός οικονομικά ενεργών μελών του νοικοκυριού							Σύνολο
		0	1	2	3	4	> 5	
Κύρια πηγή εισοδήματος	Μισθοί	32	534	520	86	27	2	1201
		0.027	0.445	0.433	0.072	0.022	0.002	0.342
		0.082	0.342	0.404	0.424	0.422	0.400	
		0.009	0.152	0.148	0.024	0.008	0.001	
	Αυτοαπασχόληση	1	223	232	52	21	2	531
		0.002	0.420	0.437	0.098	0.040	0.004	0.151
		0.003	0.143	0.180	0.256	0.328	0.400	
		0.0	0.063	0.066	0.015	0.006	0.001	
	Ακίνητη περιουσία	18	38	18	4	1	0	79
		0.228	0.481	0.228	0.051	0.013	0.0	0.022
		0.046	0.024	0.014	0.020	0.016	0.0	
		0.005	0.011	0.005	0.001	0.0	0.0	
	Συντάξεις	213	739	513	60	15	1	1541
		0.138	0.480	0.333	0.039	0.010	0.001	0.439
		0.545	0.473	0.399	0.296	0.234	0.200	
		0.061	0.210	0.146	0.017	0.004	0.0	
	Επιδόματα ανεργίας	12	8	0	0	0	0	20
		0.600	0.400	0.0	0.0	0.0	0.0	0.006
		0.031	0.005	0.0	0.0	0.0	0.0	
		0.003	0.002	0.0	0.0	0.0	0.0	
	Άλλα επιδόματα	115	20	4	1	0	0	140
		0.821	0.143	0.029	0.007	0.0	0.0	0.040
		0.294	0.013	0.003	0.005	0.0	0.0	
		0.033	0.006	0.001	0.0	0.0	0.0	
	Σύνολο	391	1562	1287	203	64	5	3512
		0.111	0.445	0.366	0.058	0.018	0.001	

Σε κάθε κελί:

Συχνότητα

Σχετική συχνότητα γραμμής

Σχετική συχνότητα στήλης

Σχετική συχνότητα κελιού

Από τον Πίνακα 2.14 προκύπτει ότι:

- Από το 34.2% του συνόλου των νοικοκυριών που έχει ως κύρια πηγή εισοδήματος το μισθό το 2.7% δεν έχει κανένα οικονομικά ενεργό μέλος, το 44.5% έχει ένα οικονομικά ενεργό μέλος, το 43.3% έχει δύο οικονομικά ενεργά μέλη, το 7.2% έχει τρία οικονομικά ενεργά μέλη, το 2.2% έχει τέσσερα οικονομικά ενεργά μέλη και το 0.2% έχει πέντε ή περισσότερα οικονομικά ενεργά μέλη.
- Από το 15.1% του συνόλου των νοικοκυριών που έχει ως κύρια πηγή το εισόδημα από αυτοαπασχόληση το 0.2% δεν έχει κανένα οικονομικά ενεργό μέλος, το 42% έχει ένα οικονομικά ενεργό μέλος, το 43.7% έχει δύο οικονομικά ενεργά μέλη, το 9.8% έχει τρία οικονομικά ενεργά μέλη, το 4%

έχει τέσσερα οικονομικά ενεργά μέλη και το 0.4% έχει πέντε ή περισσότερα οικονομικά ενεργά μέλη.

- Από το 2.2% του συνόλου των νοικοκυριών που έχει ως κύρια πηγή το εισόδημα από ακίνητη περιουσία το 22.8% δεν έχει κανένα οικονομικά ενεργό μέλος, το 48.1% έχει ένα οικονομικά ενεργό μέλος, το 22.8% έχει δύο οικονομικά ενεργά μέλη, το 5.1% έχει τρία οικονομικά ενεργά μέλη, το 1.3% έχει τέσσερα οικονομικά ενεργά μέλη και δεν υπάρχουν νοικοκυριά με πέντε ή περισσότερα οικονομικά ενεργά μέλη.
- Από το 43.9% του συνόλου των νοικοκυριών που έχει ως κύρια πηγή εισοδήματος τη σύνταξη το 13.8% δεν έχει κανένα οικονομικά ενεργό μέλος, το 48% έχει ένα οικονομικά ενεργό μέλος, το 33.3% έχει δύο οικονομικά ενεργά μέλη, το 3.9% έχει τρία οικονομικά ενεργά μέλη, το 1% έχει τέσσερα οικονομικά ενεργά μέλη και το 0.1% έχει πέντε ή περισσότερα οικονομικά ενεργά μέλη.
- Από το 0.6% του συνόλου των νοικοκυριών που έχει ως κύρια πηγή εισοδήματος το επίδομα ανεργίας το 60% δεν έχει κανένα οικονομικά ενεργό μέλος και το 40% έχει ένα οικονομικά ενεργό μέλος. Ενώ δεν υπάρχουν νοικοκυριά με δύο, τρία, τέσσερα, πέντε ή περισσότερα οικονομικά ενεργά μέλη.
- Από το 4% του συνόλου των νοικοκυριών που έχει ως κύρια πηγή εισοδήματος άλλα επιδόματα το 82.1% δεν έχει κανένα οικονομικά ενεργό μέλος, το 14.3% έχει ένα οικονομικά ενεργό μέλος, το 2.9% έχει δύο οικονομικά ενεργά μέλη και το 0.7% έχει τρία οικονομικά ενεργά μέλη. Ενώ δεν υπάρχουν νοικοκυριά με τέσσερα, πέντε ή περισσότερα οικονομικά ενεργά μέλη.

Κατά συνέπεια, το μεγαλύτερο ποσοστό των νοικοκυριών με κύρια πηγή εισοδήματος τους μισθούς και τα εισοδήματα από αυτοαπασχόληση έχει ένα ή δύο οικονομικά ενεργά μέλη. Επίσης, το μεγαλύτερο ποσοστό των νοικοκυριών με κύρια πηγή εισοδήματος τη σύνταξη και τα εισοδήματα από ακίνητη περιουσία έχει ένα οικονομικά ενεργό μέλος. Τέλος, το μεγαλύτερο ποσοστό των νοικοκυριών με κύρια πηγή εισοδήματος τα επιδόματα ανεργίας και άλλα επιδόματα δεν έχει κανένα οικονομικά ενεργό μέλος.

Όσον αφορά στη συσχέτιση των δύο μεταβλητών, χρησιμοποιώντας τον ακριβή έλεγχο του Fisher, υπολογίζεται η p – τιμή του αμφίπλευρου ελέγχου ότι είναι 0.0004998 σε επίπεδο σημαντικότητας 5%, οπότε υπάρχουν σοβαρές ενδείξεις για να απορριφθεί η μηδενική υπόθεση. Άρα, υπάρχει συσχέτιση μεταξύ της κύριας πηγής εισοδήματος και του αριθμού των οικονομικά ενεργών μελών του νοικοκυριού.

Εντολές στην R

Με την εντολή `CrossTable()` δίνεται ο πίνακας συνάφειας των δύο μεταβλητών όπου για κάθε κελί υπολογίζονται η συχνότητα, η σχετική συχνότητα της γραμμής, η σχετική συχνότητα της στήλης και η σχετική συχνότητά του σε όλο τον πίνακα. Επιπρόσθετα, υπολογίζεται το σύνολο ανά γραμμή και ανά στήλη. Εφαρμόζοντας

τον έλεγχο ανεξαρτησίας X^2 στα δεδομένα με την εντολή `chisq.test()` υπήρχε ένα προειδοποιητικό μήνυμα λάθους, γεγονός που σημαίνει ότι η προσέγγιση μπορεί να μην είναι αξιόπιστη, αφού δεν ισχύει η προϋπόθεση ότι όλες οι αναμενόμενες συχνότητες είναι μεγαλύτερες του 5. Γι' αυτό το λόγο, εφαρμόζεται ο μη παραμετρικός ακριβής έλεγχος του Fisher με την εντολή `fisher.test()`. Τέλος, με την εντολή `expected()` διαπιστώνεται ότι όντως υπήρχαν αναμενόμενες συχνότητες μικρότερες του 5.

```
> CrossTable (table (source, members))
> c <- chisq.test (table (source, members))
Warning message:
In chisq.test (table (source, members)) :
  Chi-squared approximation may be incorrect
> c
      Pearson's Chi-squared test
data:  table (source, members)
X-squared = 1008.933, df = 25, p-value < 2.2e-16
> fisher.test (table (source, members), simulate.p.value = TRUE)
      Fisher's Exact Test for Count Data with simulated p-value (based on
      2000 replicates)
data:  table (source, members)
p-value = 0.0004998
alternative hypothesis: two.sided
> attach (c)
> expected
      members
source  0    1    2    3    4    5
  1 133.710421 534.157745 440.115888 69.419989 21.8861048 1.7098519
  2  59.117597 236.167995 194.589123 30.692768  9.6765376 0.7559795
  3  8.795273 35.136105 28.950171  4.566344  1.4396355 0.1124715
  4 171.563497 685.376424 564.711560 89.072608 28.0820046 2.1939066
  5  2.226651  8.895216  7.329157  1.156036  0.3644647 0.0284738
  6 15.586560 62.266515 51.304100  8.092255  2.5512528 0.1993166
```


2.2.7. Κύρια πηγή εισοδήματος – Κοινωνικο – οικονομική κατάσταση του προσώπου αναφοράς του νοικοκυριού

Πίνακας 2.15: Πίνακας συνάφειας κύριας πηγής εισοδήματος – κοινωνικο – οικονομικής κατάστασης του προσώπου αναφοράς του νοικοκυριού

Κύρια πηγή εισοδήματος		Κοινωνικο – οικονομική κατάσταση του προσώπου αναφοράς του νοικοκυριού							Σύνολο
		Χειρώνα-κτας	Μη χειρώνα-κτας	Αυτο-απασχο-λούμενος	Άνεργος	Συντα-ξιούχος	Λοιποί μη οικονομικά ενεργοί	Δεν έχει απαντηθεί	
	Μισθοί	532 0.443 0.962 0.151	323 0.269 0.947 0.092	98 0.082 0.159 0.028	89 0.074 0.654 0.025	103 0.086 0.072 0.029	55 0.046 0.126 0.016	1 0.001 0.333 0.0	1201 0.342
	Αυτο-απασχόληση	6 0.011 0.011 0.002	6 0.011 0.018 0.002	476 0.896 0.771 0.136	5 0.009 0.037 0.001	30 0.056 0.021 0.009	6 0.011 0.014 0.002	2 0.004 0.667 0.001	531 0.151
	Ακίνητη περιουσία	4 0.051 0.007 0.001	5 0.063 0.015 0.001	12 0.152 0.019 0.003	6 0.076 0.044 0.002	31 0.392 0.022 0.009	21 0.266 0.048 0.006	0 0.0 0.0 0.0	79 0.022
	Συντάξεις	6 0.004 0.011 0.002	4 0.003 0.012 0.001	25 0.016 0.041 0.007	5 0.003 0.037 0.001	1255 0.814 0.879 0.357	246 0.160 0.566 0.070	0 0.0 0.0 0.0	1541 0.439
	Επιδόματα ανεργίας	1 0.050 0.002 0.0	1 0.050 0.003 0.0	0 0.0 0.0 0.0	16 0.800 0.118 0.005	1 0.050 0.001 0.0	1 0.050 0.002 0.0	0 0.0 0.0 0.0	20 0.006
	Άλλα επιδόματα	4 0.029 0.007 0.001	2 0.014 0.006 0.001	6 0.043 0.010 0.002	15 0.107 0.110 0.004	7 0.050 0.005 0.002	106 0.757 0.244 0.030	0 0.0 0.0 0.0	140 0.040
	Σύνολο	553 0.157	341 0.097	617 0.176	136 0.039	1427 0.406	435 0.124	3 0.001	3512

Σε κάθε κελί:

Συχνότητα

Σχετική συχνότητα γραμμής

Σχετική συχνότητα στήλης

Σχετική συχνότητα κελιού

Από τον Πίνακα 2.15 προκύπτει ότι:

- Από το 34.2% του συνόλου των νοικοκυριών, που έχει ως κύρια πηγή εισοδήματος το μισθό, το 44.3% έχει ως πρόσωπο αναφοράς χειρώνακτα εργαζόμενο, το 26.9% έχει μη χειρώνακτα εργαζόμενο, το 8.2% έχει αυτοαπασχολούμενο, γεωργό ή κτηνοτρόφο, το 7.4% έχει άνεργο, το 8.6% έχει συνταξιούχο, το 4.6% έχει άλλους μη οικονομικά ενεργούς και 0.1% δεν έχει απαντήσει.
- Από το 15.1% του συνόλου των νοικοκυριών που έχει ως κύρια πηγή το εισόδημα από αυτοαπασχόληση το 1.1% έχει ως πρόσωπο αναφοράς χειρώνακτα εργαζόμενο, το 1.1% έχει μη χειρώνακτα εργαζόμενο, το 89.6% έχει αυτοαπασχολούμενο, γεωργό ή κτηνοτρόφο, το 0.9% έχει άνεργο, το 5.6% έχει συνταξιούχο, το 1.1% έχει άλλους μη οικονομικά ενεργούς και 0.4% δεν έχει απαντήσει.
- Από το 2.2% του συνόλου των νοικοκυριών που έχει ως κύρια πηγή το εισόδημα από ακίνητη περιουσία το 5.1% έχει ως πρόσωπο αναφοράς χειρώνακτα εργαζόμενο, το 6.3% έχει μη χειρώνακτα εργαζόμενο, το 15.2% έχει αυτοαπασχολούμενο, γεωργό ή κτηνοτρόφο, το 7.6% έχει άνεργο, το

39.2% έχει συνταξιούχο, το 26.6% έχει άλλους μη οικονομικά ενεργούς και δεν υπάρχουν νοικοκυριά που δεν έχουν απαντήσει.

- Από το 43.9% του συνόλου των νοικοκυριών που έχει ως κύρια πηγή εισοδήματος τη σύνταξη το 0.4% έχει ως πρόσωπο αναφοράς χειρώνακτα εργαζόμενο, το 0.3% έχει μη χειρώνακτα εργαζόμενο, το 1.6% έχει αυτοαπασχολούμενο, γεωργό ή κτηνοτρόφο, το 0.3% έχει άνεργο, το 81.4% έχει συνταξιούχο, το 16% έχει άλλους μη οικονομικά ενεργούς και δεν υπάρχουν νοικοκυριά που δεν έχουν απαντήσει.
- Από το 0.6% του συνόλου των νοικοκυριών που έχει ως κύρια πηγή εισοδήματος το επίδομα ανεργίας το 5% έχει ως πρόσωπο αναφοράς χειρώνακτα εργαζόμενο, το 5% έχει μη χειρώνακτα εργαζόμενο, το 80% έχει άνεργο, το 5% έχει συνταξιούχο, το 5% έχει άλλους μη οικονομικά ενεργούς και δεν υπάρχουν νοικοκυριά που το πρόσωπο αναφοράς να είναι αυτοαπασχολούμενος, γεωργός ή κτηνοτρόφος ή που δεν έχουν απαντήσει.
- Από το 4% του συνόλου των νοικοκυριών που έχει ως κύρια πηγή εισοδήματος άλλα επιδόματα το 2.9% έχει ως πρόσωπο αναφοράς χειρώνακτα εργαζόμενο, το 1.4% έχει μη χειρώνακτα εργαζόμενο, το 4.3% έχει αυτοαπασχολούμενο, γεωργό ή κτηνοτρόφο, το 10.7% έχει άνεργο, το 5% έχει συνταξιούχο, το 75.7% έχει άλλους μη οικονομικά ενεργούς και δεν υπάρχουν νοικοκυριά που δεν έχουν απαντήσει.

Κατά συνέπεια, το μεγαλύτερο ποσοστό των νοικοκυριών με κύρια πηγή εισοδήματος το μισθό έχει χειρώνακτα ή μη χειρώνακτα εργαζόμενο ως πρόσωπο αναφοράς. Ακόμα, το μεγαλύτερο ποσοστό των νοικοκυριών με κύρια πηγή εισοδήματος τα εισοδήματα από αυτοαπασχόληση έχει αυτοαπασχολούμενο, γεωργό ή κτηνοτρόφο ως πρόσωπο αναφοράς. Επίσης, το μεγαλύτερο ποσοστό των νοικοκυριών με κύρια πηγή εισοδήματος τη σύνταξη και τα εισοδήματα από ακίνητη περιουσία έχει συνταξιούχο ή άλλους μη οικονομικά ενεργούς ως πρόσωπο αναφοράς. Τέλος, το μεγαλύτερο ποσοστό των νοικοκυριών με κύρια πηγή εισοδήματος τα επιδόματα ανεργίας και άλλα επιδόματα έχουν άνεργο ως πρόσωπο αναφοράς.

Όσον αφορά στη συσχέτιση των δύο μεταβλητών, χρησιμοποιώντας τον ακριβή έλεγχο του Fisher, υπολογίζεται η p – τιμή του αμφίπλευρου ελέγχου ότι είναι 0.0004998 σε επίπεδο σημαντικότητας 5%, οπότε υπάρχουν σοβαρές ενδείξεις για να απορριφθεί η μηδενική υπόθεση. Άρα, υπάρχει συσχέτιση μεταξύ της κύριας πηγής εισοδήματος και της κοινωνικο – οικονομικής κατάστασης του προσώπου αναφοράς του νοικοκυριού.

Εντολές στην R

Με την εντολή `CrossTable()` δίνεται ο πίνακας συνάφειας των δύο μεταβλητών όπου για κάθε κελί υπολογίζονται η συχνότητα, η σχετική συχνότητα της γραμμής, η σχετική συχνότητα της στήλης και η σχετική συχνότητά του σε όλο τον πίνακα. Επιπρόσθετα, υπολογίζεται το σύνολο ανά γραμμή και ανά στήλη. Εφαρμόζοντας τον έλεγχο ανεξαρτησίας χ^2 στα δεδομένα με την εντολή `chisq.test()` υπήρχε ένα προειδοποιητικό μήνυμα λάθους, γεγονός που σημαίνει ότι η προσέγγιση μπορεί να

μην είναι αξιόπιστη, αφού δεν ισχύει η προϋπόθεση ότι όλες οι αναμενόμενες συχνότητες είναι μεγαλύτερες του 5. Γι' αυτό το λόγο, εφαρμόζεται ο μη παραμετρικός ακριβής έλεγχος του Fisher με την εντολή `fisher.test()`. Τέλος, με την εντολή `expected()` διαπιστώνεται ότι όντως υπήρχαν αναμενόμενες συχνότητες μικρότερες του 5.

```
> CrossTable (table (source, sitnew))
> c <- chisq.test (table (source, sitnew))
Warning message:
In chisq.test (table (source, sitnew)) :
  Chi-squared approximation may be incorrect
> c
      Pearson's Chi-squared test
data:  table (source, sitnew)
X-squared = 5544.718, df = 30, p-value < 2.2e-16
> fisher.test (table (source, sitnew), simulate.p.value = TRUE)
      Fisher's Exact Test for Count Data with simulated p-value (based on
      2000 replicates)
data:  table (source, sitnew)
p-value = 0.0004998
alternative hypothesis: two.sided
> attach (c)
> expected
      sitnew
source    1    2    3    4    5    6
  1 189.109624 116.611902 210.995729 46.5079727 487.991743 148.757118
  2  83.611333  51.557802  93.287870 20.5626424 215.756549  65.770216
  3  12.439351   7.670558  13.878986  3.0592255  32.099374   9.785023
  4 242.646071 149.624431 270.728075 59.6742597 626.140945 190.869875
  5  3.149203  1.941913  3.513667 0.7744875  8.126424  2.477221
  6 22.044419 13.593394 24.595672  5.4214123 56.884966 17.340547
      sitnew
source    7
  1 1.02591116
  2 0.45358770
  3 0.06748292
  4 1.31634396
  5 0.01708428
  6 0.11958998
```

2.2.8. Βαθμός πυκνότητας – Αριθμός οικονομικά ενεργών μελών του νοικοκυριού

Πίνακας 2.16: Πίνακας συνάφειας βαθμού πυκνότητας – αριθμού οικονομικά ενεργών μελών του νοικοκυριού

Βαθμός πυκνότητας	Αριθμός οικονομικά ενεργών μελών του νοικοκυριού							Σύνολο
		0	1	2	3	4	> 5	
> 500 κάτοικοι / km ²		216	697	473	81	20	2	1489
		0.145	0.468	0.318	0.054	0.013	0.001	0.424
		0.552	0.446	0.368	0.399	0.312	0.400	
		0.062	0.198	0.135	0.023	0.006	0.001	
	100 - 499 κάτοικοι / km ²	43	213	163	26	6	0	451
		0.095	0.472	0.361	0.058	0.013	0.0	0.128
		0.110	0.136	0.127	0.128	0.094	0.0	
		0.012	0.061	0.046	0.007	0.002	0.0	
	< 100 κάτοικοι / km ²	132	652	651	96	38	3	1572
		0.084	0.415	0.414	0.061	0.024	0.002	0.448
		0.338	0.417	0.506	0.473	0.594	0.600	
		0.038	0.186	0.185	0.027	0.011	0.001	
	Σύνολο	391	1562	1287	203	64	5	3512
		0.111	0.445	0.366	0.058	0.018	0.001	

Σε κάθε κελί:

Συχνότητα

Σχετική συχνότητα γραμμής

Σχετική συχνότητα στήλης

Σχετική συχνότητα κελιού

Από τον Πίνακα 2.16 προκύπτει ότι:

- Από το 42.4% του συνόλου των νοικοκυριών που ζει σε πυκνοκατοικημένη περιοχή το 14.5% δεν έχει κανένα οικονομικά ενεργό μέλος, το 46.8% έχει ένα οικονομικά ενεργό μέλος, το 31.8% έχει δύο οικονομικά ενεργά μέλη, το 5.4% έχει τρία οικονομικά ενεργά μέλη, το 1.3% έχει τέσσερα οικονομικά ενεργά μέλη και το 0.1% έχει πέντε ή περισσότερα οικονομικά ενεργά μέλη.
- Από το 12.8% του συνόλου των νοικοκυριών που ζει σε ενδιάμεσης πυκνότητας περιοχή το 9.5% δεν έχει κανένα οικονομικά ενεργό μέλος, το 47.2% έχει ένα οικονομικά ενεργό μέλος, το 36.1% έχει δύο οικονομικά ενεργά μέλη, το 5.8% έχει τρία οικονομικά ενεργά μέλη, το 1.3% έχει τέσσερα οικονομικά ενεργά μέλη και δεν υπάρχουν νοικοκυριά με πέντε ή περισσότερα οικονομικά ενεργά μέλη.
- Από το 44.8% του συνόλου των νοικοκυριών που ζει σε αραιοκατοικημένη περιοχή, το 8.4% δεν έχει κανένα οικονομικά ενεργό μέλος, το 41.5% έχει ένα οικονομικά ενεργό μέλος, το 41.4% έχει δύο οικονομικά ενεργά μέλη, το 6.1% έχει τρία οικονομικά ενεργά μέλη, το 2.4% έχει τέσσερα οικονομικά ενεργά μέλη και το 0.2% έχει πέντε ή περισσότερα οικονομικά ενεργά μέλη.

Κατά συνέπεια, το μεγαλύτερο ποσοστό των νοικοκυριών που ζει σε πυκνοκατοικημένη, ενδιάμεσης πυκνότητας ή αραιοκατοικημένη περιοχή έχει ένα ή δύο οικονομικά ενεργά μέλη.

Όσον αφορά στη συσχέτιση των δύο μεταβλητών, χρησιμοποιώντας τον ακριβή έλεγχο του Fisher, υπολογίζεται η p – τιμή του αμφίπλευρου ελέγχου ότι είναι 0.0004998 σε επίπεδο σημαντικότητας 5%, οπότε υπάρχουν σοβαρές ενδείξεις για να απορριφθεί η μηδενική υπόθεση. Άρα, υπάρχει συσχέτιση μεταξύ του βαθμού πυκνότητας και του αριθμού των οικονομικά ενεργών μελών του νοικοκυριού.

Εντολές στην R

Με την εντολή `CrossTable()` δίνεται ο πίνακας συνάφειας των δύο μεταβλητών όπου για κάθε κελί υπολογίζονται η συχνότητα, η σχετική συχνότητα της γραμμής, η σχετική συχνότητα της στήλης και η σχετική συχνότητά του σε όλο τον πίνακα. Επιπρόσθετα, υπολογίζεται το σύνολο ανά γραμμή και ανά στήλη. Εφαρμόζοντας τον έλεγχο ανεξαρτησίας X^2 στα δεδομένα με την εντολή `chisq.test()` υπήρχε ένα προειδοποιητικό μήνυμα λάθους, γεγονός που σημαίνει ότι η προσέγγιση μπορεί να μην είναι αξιόπιστη, αφού δεν ισχύει η προϋπόθεση ότι όλες οι αναμενόμενες συχνότητες είναι μεγαλύτερες του 5. Γι' αυτό το λόγο, εφαρμόζεται ο μη παραμετρικός ακριβής έλεγχος του Fisher με την εντολή `fisher.test()`. Τέλος, με την εντολή `expected()` διαπιστώνεται ότι όντως υπήρχαν αναμενόμενες συχνότητες μικρότερες του 5.

```
> CrossTable (table (density, members))
> c <- chisq.test (table (density, members))
Warning message:
In chisq.test (table (density, members)) :
  Chi-squared approximation may be incorrect
> c
      Pearson's Chi-squared test
data:  table (density, members)
X-squared = 59.0751, df = 10, p-value = 5.421e-09
> fisher.test (table (density, members), simulate.p.value = TRUE)
      Fisher's Exact Test for Count Data with simulated p-value (based on
      2000 replicates)
data:  table (density, members)
p-value = 0.0004998
alternative hypothesis: two.sided
> attach (c)
> expected
      members
density  0    1    2    3    4    5
1 165.77420 662.2489 545.6558 86.06691 27.134396 2.1198747
2  50.21099 200.5871 165.2725 26.06862  8.218679 0.6420843
3 175.01481 699.1640 576.0718 90.86446 28.646925 2.2380410
```

2.2.9. Βαθμός πυκνότητας – Κοινωνικο – οικονομική κατάσταση του προσώπου αναφοράς του νοικοκυριού

Πίνακας 2.17: Πίνακας συνάφειας βαθμού πυκνότητας – κοινωνικο – οικονομικής κατάστασης του προσώπου αναφοράς του νοικοκυριού

Βαθμός πυκνότητας	Κοινωνικο – οικονομική κατάσταση του προσώπου αναφοράς του νοικοκυριού								Σύνολο
		Χειρώνακτας	Μη χειρώνακτας	Αυτο-απασχολούμενος	Άνεργος	Συνταξιούχος	Λοιποί μη οικονομικά ενεργοί	Δεν έχει απαντήσει	
	> 500 κάτοικοι /km ²	300 0.201 0.542 0.085	188 0.126 0.551 0.054	206 0.138 0.334 0.059	73 0.049 0.537 0.021	492 0.330 0.345 0.140	230 0.154 0.529 0.065	0 0.0 0.0 0.0	1489 0.424
	100 - 499 κάτοικοι /km ²	69 0.153 0.125 0.020	63 0.140 0.185 0.018	68 0.151 0.110 0.019	20 0.044 0.147 0.006	178 0.395 0.125 0.051	53 0.118 0.122 0.015	0 0.0 0.0 0.0	451 0.128
	< 100 κάτοικοι /km ²	184 0.117 0.333 0.052	90 0.057 0.264 0.026	343 0.218 0.556 0.098	43 0.027 0.316 0.012	757 0.482 0.530 0.216	152 0.097 0.349 0.043	3 0.002 1.0 0.001	1572 0.448
	Σύνολο	553 0.157	341 0.097	617 0.176	136 0.039	1427 0.406	435 0.124	3 0.001	3512

Σε κάθε κελί:

Συχνότητα

Σχετική συχνότητα γραμμής

Σχετική συχνότητα στήλης

Σχετική συχνότητα κελιού

Από τον Πίνακα 2.17 προκύπτει ότι:

- Από το 42.4% του συνόλου των νοικοκυριών που ζει σε πυκνοκατοικημένη περιοχή το 20.1% έχει ως πρόσωπο αναφοράς χειρώνακτα εργαζόμενο, το 12.6% έχει μη χειρώνακτα εργαζόμενο, το 13.8% έχει αυτοαπασχολούμενο, γεωργό ή κτηνοτρόφο, το 4.9% έχει άνεργο, το 33% έχει συνταξιούχο, το 15.4% έχει άλλους μη οικονομικά ενεργούς και δεν υπάρχουν νοικοκυριά που δεν έχουν απαντήσει.
- Από το 12.8% του συνόλου των νοικοκυριών που ζει σε ενδιάμεσης πυκνότητας περιοχή το 15.3% έχει ως πρόσωπο αναφοράς χειρώνακτα εργαζόμενο, το 14% έχει μη χειρώνακτα εργαζόμενο, το 15.1% έχει αυτοαπασχολούμενο, γεωργό ή κτηνοτρόφο, το 4.4% έχει άνεργο, το 39.5% έχει συνταξιούχο, το 11.8% έχει άλλους μη οικονομικά ενεργούς και δεν υπάρχουν νοικοκυριά που δεν έχουν απαντήσει.
- Από το 44.8% του συνόλου των νοικοκυριών που ζει σε αραιοκατοικημένη περιοχή το 11.7% έχει ως πρόσωπο αναφοράς χειρώνακτα εργαζόμενο, το 5.7% έχει μη χειρώνακτα εργαζόμενο, το 21.8% έχει αυτοαπασχολούμενο, γεωργό ή κτηνοτρόφο, το 2.7% έχει άνεργο, το 48.2% έχει συνταξιούχο, το 9.7% έχει άλλους μη οικονομικά ενεργούς και 0.2% δεν έχει απαντήσει.

Κατά συνέπεια, το μεγαλύτερο ποσοστό των νοικοκυριών που ζει σε πυκνοκατοικημένη, ενδιάμεσης πυκνότητας ή αραιοκατοικημένη περιοχή έχει συνταξιούχο ως πρόσωπο αναφοράς.

Όσον αφορά στη συσχέτιση των δύο μεταβλητών, χρησιμοποιώντας τον ακριβή έλεγχο του Fisher, υπολογίζεται η p – τιμή του αμφίπλευρου ελέγχου ότι είναι 0.0004998 σε επίπεδο σημαντικότητας 5%, οπότε υπάρχουν σοβαρές ενδείξεις για να απορριφθεί η μηδενική υπόθεση. Άρα, υπάρχει συσχέτιση μεταξύ του βαθμού πυκνότητας και της κοινωνικο – οικονομικής κατάστασης του προσώπου αναφοράς του νοικοκυριού.

Εντολές στην R

Με την εντολή `CrossTable()` δίνεται ο πίνακας συνάφειας των δύο μεταβλητών όπου για κάθε κελί υπολογίζονται η συχνότητα, η σχετική συχνότητα της γραμμής, η σχετική συχνότητα της στήλης και η σχετική συχνότητά του σε όλο τον πίνακα. Επιπρόσθετα, υπολογίζεται το σύνολο ανά γραμμή και ανά στήλη. Εφαρμόζοντας τον έλεγχο ανεξαρτησίας X^2 στα δεδομένα με την εντολή `chisq.test()` υπήρχε ένα προειδοποιητικό μήνυμα λάθους, γεγονός που σημαίνει ότι η προσέγγιση μπορεί να μην είναι αξιόπιστη, αφού δεν ισχύει η προϋπόθεση ότι όλες οι αναμενόμενες συχνότητες είναι μεγαλύτερες του 5. Γι' αυτό το λόγο, εφαρμόζεται ο μη παραμετρικός ακριβής έλεγχος του Fisher με την εντολή `fisher.test()`. Τέλος, με την εντολή `expected()` διαπιστώνεται ότι όντως υπήρχαν αναμενόμενες συχνότητες μικρότερες του 5.

```
> CrossTable (table (density, sitnew))
> c <- chisq.test (table (density, sitnew))
Warning message:
In chisq.test (table (density, sitnew)) :
  Chi-squared approximation may be incorrect
> c
      Pearson's Chi-squared test
data:  table (density, sitnew)
X-squared = 188.764, df = 12, p-value < 2.2e-16
> fisher.test (table (density, sitnew), simulate.p.value = TRUE)
      Fisher's Exact Test for Count Data with simulated p-value (based on
      2000 replicates)
data:  table (density, sitnew)
p-value = 0.0004998
alternative hypothesis: two.sided
> attach (c)
> expected
      sitnew
density  1    2    3    4    5    6    7
  1 234.45814 144.57546 261.5925 57.66059 605.0122 184.42910 1.2719248
  2  71.01452  43.79015  79.2332 17.46469 183.2509  55.86133 0.3852506
  3 247.52733 152.63440 276.1743 60.87472 638.7369 194.70957 1.3428246
```

2.2.10. Αριθμός οικονομικά ενεργών μελών του νοικοκυριού – Κοινωνικο – οικονομική κατάσταση του προσώπου αναφοράς του νοικοκυριού

Πίνακας 2.18: Πίνακας συνάφειας αριθμού οικονομικά ενεργών μελών– κοινωνικο – οικονομικής κατάστασης του προσώπου αναφοράς του νοικοκυριού

Αριθμός οικονομικά ενεργών μελών	Κοινωνικο – οικονομική κατάσταση του προσώπου αναφοράς του νοικοκυριού								
		Χειρώνα-κτας	Μη χειρώ-νακτας	Αυτο-απασχο-λούμενος	Άνεργος	Συντα-ξιούχος	Λοιποί μη οικονομικά ενεργοί	Δεν έχει απαντηθεί	Σύνολο
	0	0	0	0	62	0	329	0	391
		0.0	0.0	0.0	0.159	0.0	0.841	0.0	
		0.0	0.0	0.0	0.456	0.0	0.756	0.0	
		0.0	0.0	0.0	0.018	0.0	0.094	0.0	
	1	296	154	228	63	730	90	1	1562
		0.190	0.099	0.146	0.040	0.467	0.058	0.001	
		0.535	0.452	0.370	0.463	0.512	0.207	0.333	
		0.084	0.044	0.065	0.018	0.208	0.026	0.0	
	2	219	171	289	8	584	16	0	1287
		0.170	0.133	0.225	0.006	0.454	0.012	0.0	
		0.396	0.501	0.468	0.059	0.409	0.037	0.0	
		0.062	0.049	0.082	0.002	0.166	0.005	0.0	
	3	29	14	71	2	86	0	1	203
		0.143	0.069	0.350	0.010	0.424	0.0	0.005	
		0.052	0.041	0.115	0.015	0.060	0.0	0.333	
		0.008	0.004	0.020	0.001	0.024	0.0	0.0	
	4	8	2	27	1	25	0	1	64
		0.125	0.031	0.422	0.016	0.391	0.0	0.016	
		0.014	0.006	0.044	0.007	0.018	0.0	0.333	
		0.002	0.001	0.008	0.0	0.007	0.0	0.0	
	>5	1	0	2	0	2	0	0	5
		0.200	0.0	0.400	0.0	0.400	0.0	0.0	
		0.002	0.0	0.003	0.0	0.001	0.0	0.0	
		0.0	0.0	0.001	0.0	0.001	0.0	0.0	
	Σύνολο	553	341	617	136	1427	435	3	3512
		0.157	0.097	0.176	0.039	0.406	0.124	0.001	

Σε κάθε κελί:

Συχνότητα

Σχετική συχνότητα γραμμής

Σχετική συχνότητα στήλης

Σχετική συχνότητα κελιού

Από τον Πίνακα 2.18 προκύπτει ότι:

- Από το 11.1% του συνόλου των νοικοκυριών που δεν έχουν κανένα οικονομικά ενεργό μέλος το 15.9% έχει ως πρόσωπο αναφοράς άνεργο, το 84.1% έχει άλλους μη οικονομικά ενεργούς και δεν υπάρχουν νοικοκυριά με χειρώνακτα, μη χειρώνακτα, αυτοαπασχολούμενο, συνταξιούχο ως πρόσωπο αναφοράς ή που δεν έχουν απαντήσει.
- Από το 44.5% του συνόλου των νοικοκυριών που έχουν ένα οικονομικά ενεργό μέλος το 19% έχει ως πρόσωπο αναφοράς χειρώνακτα εργαζόμενο, το 9.9% έχει μη χειρώνακτα εργαζόμενο, το 14.6% έχει αυτοαπασχολούμενο, γεωργό ή κτηνοτρόφο, το 4% έχει άνεργο, το 46.7% έχει συνταξιούχο, το 5.8% έχει άλλους μη οικονομικά ενεργούς και το 0.1% δεν έχουν απαντήσει.
- Από το 36.6% του συνόλου των νοικοκυριών που έχουν δύο οικονομικά ενεργά μέλη το 17% έχει ως πρόσωπο αναφοράς χειρώνακτα εργαζόμενο, το 13.3% έχει μη χειρώνακτα εργαζόμενο, το 22.5% έχει αυτοαπασχολούμενο, γεωργό ή κτηνοτρόφο, το 0.6% έχει άνεργο, το 45.4% έχει συνταξιούχο, το 1.2% έχει άλλους μη οικονομικά ενεργούς και δεν υπάρχουν νοικοκυριά που δεν έχουν απαντήσει.

- Από το 5.8% του συνόλου των νοικοκυριών που έχουν τρία οικονομικά ενεργά μέλη το 14.3% έχει ως πρόσωπο αναφοράς χειρώνακτα εργαζόμενο, το 6.9% έχει μη χειρώνακτα εργαζόμενο, το 35% έχει αυτοαπασχολούμενο, γεωργό ή κτηνοτρόφο, το 1% έχει άνεργο, το 42.4% έχει συνταξιούχο, δεν υπάρχουν νοικοκυριά με άλλο μη οικονομικά ενεργό μέλος ως πρόσωπο αναφοράς και το 0.5% δεν έχουν απαντήσει.
- Από το 1.8% του συνόλου των νοικοκυριών που έχουν τέσσερα οικονομικά ενεργά μέλη το 12.5% έχει ως πρόσωπο αναφοράς χειρώνακτα εργαζόμενο, το 3.1% έχει μη χειρώνακτα εργαζόμενο, το 42.2% έχει αυτοαπασχολούμενο, γεωργό ή κτηνοτρόφο, το 1.6% έχει άνεργο, το 39.1% έχει συνταξιούχο, δεν υπάρχουν νοικοκυριά με άλλο μη οικονομικά ενεργό μέλος ως πρόσωπο αναφοράς και το 1.6% δεν έχουν απαντήσει.
- Από το 44.5% του συνόλου των νοικοκυριών που έχουν πέντε ή περισσότερα οικονομικά ενεργά μέλη το 20% έχει ως πρόσωπο αναφοράς χειρώνακτα εργαζόμενο, το 40% έχει αυτοαπασχολούμενο, γεωργό ή κτηνοτρόφο, το 40% έχει συνταξιούχο και δεν υπάρχουν νοικοκυριά με μη χειρώνακτα, άνεργο και άλλο μη οικονομικά ενεργό μέλος ως πρόσωπο αναφοράς ή που δεν έχουν απαντήσει.

Κατά συνέπεια, τα νοικοκυριά με κανένα οικονομικά ενεργό μέλος έχουν άνεργο ή άλλο οικονομικά μη ενεργό μέλος ως πρόσωπο αναφοράς. Ακόμα, το μεγαλύτερο ποσοστό των νοικοκυριών με ένα ή δύο οικονομικά ενεργά μέλη έχει συνταξιούχο ως πρόσωπο αναφοράς. Επίσης, το μεγαλύτερο ποσοστό των νοικοκυριών με τρία ή τέσσερα οικονομικά ενεργά μέλη έχει αυτοαπασχολούμενο ή συνταξιούχο ως πρόσωπο αναφοράς. Τέλος, τα νοικοκυριά με πέντε ή περισσότερα οικονομικά ενεργά μέλη έχουν χειρώνακτα εργαζόμενο, αυτοαπασχολούμενο ή συνταξιούχο ως πρόσωπο αναφοράς.

Όσον αφορά στη συσχέτιση των δύο μεταβλητών, χρησιμοποιώντας τον ακριβή έλεγχο του Fisher, υπολογίζεται η p – τιμή του αμφίπλευρου ελέγχου ότι είναι 0.0004998 σε επίπεδο σημαντικότητας 5%, οπότε υπάρχουν σοβαρές ενδείξεις για να απορριφθεί η μηδενική υπόθεση. Άρα, υπάρχει συσχέτιση μεταξύ του αριθμού των οικονομικά ενεργών μελών και της κοινωνικο – οικονομικής κατάστασης του προσώπου αναφοράς του νοικοκυριού.

Εντολές στην R

Με την εντολή `CrossTable()` δίνεται ο πίνακας συνάφειας των δύο μεταβλητών όπου για κάθε κελί υπολογίζονται η συχνότητα, η σχετική συχνότητα της γραμμής, η σχετική συχνότητα της στήλης και η σχετική συχνότητά του σε όλο τον πίνακα. Επιπρόσθετα, υπολογίζεται το σύνολο ανά γραμμή και ανά στήλη. Εφαρμόζοντας τον έλεγχο ανεξαρτησίας X^2 στα δεδομένα με την εντολή `chisq.test()` υπήρχε ένα προειδοποιητικό μήνυμα λάθους, γεγονός που σημαίνει ότι η προσέγγιση μπορεί να μην είναι αξιόπιστη, αφού δεν ισχύει η προϋπόθεση ότι όλες οι αναμενόμενες συχνότητες είναι μεγαλύτερες του 5. Γι' αυτό το λόγο, εφαρμόζεται ο μη παραμετρικός ακριβής έλεγχος του Fisher με την εντολή `fisher.test()`. Τέλος, με την

εντολή `expected()` διαπιστώνεται ότι όντως υπήρχαν αναμενόμενες συχνότητες μικρότερες του 5.

```
> CrossTable (table (members, sitnew))
> c <- chisq.test (table (members, sitnew))
Warning message:
In chisq.test (table (members, sitnew)) :
  Chi-squared approximation may be incorrect
> c
      Pearson's Chi-squared test
data:  table (members, sitnew)
X-squared = 2517.846, df = 30, p-value < 2.2e-16
> fisher.test (table (members, sitnew), simulate.p.value = TRUE)
      Fisher's Exact Test for Count Data with simulated p-value (based on
      2000 replicates)
data:  table (members, sitnew)
p-value = 0.0004998
alternative hypothesis: two.sided
> attach (c)
> expected
      sitnew
members    1      2      3      4      5      6
0 61.5669134 37.9644077 68.6921982 15.1412301 158.871583 48.4296697
1 245.9527335 151.6634396 274.4174260 60.4874715 634.673690 193.4709567
2 202.6511959 124.9621298 226.1044989 49.8382688 522.935364 159.4091686
3 31.9644077 19.7104214 35.6637244 7.8610478 82.483200 25.1437927
4 10.0774487 6.2141230 11.2437358 2.4783599 26.004556 7.9271071
5 0.7873007 0.4854784 0.8784169 0.1936219 2.031606 0.6193052
      sitnew
members      7
0 0.333997722
1 1.334282460
2 1.099373576
3 0.173405467
4 0.054669704
5 0.004271071
```

2.2.11. Κατηγορίες δαπανών του νοικοκυριού

Πίνακας 2.19: Πίνακας του συντελεστή συσχετίσεων του Pearson και της p – τιμής των 12 κατηγοριών δαπανών δαπάνης των νοικοκυριών

	Διατροφή	Ποτά	Ένδυση	Στέγαση	Οικιακή χρήση	Υγεία
Διατροφή	1.00	0.32 β0.001	0.42 β0.001	0.30 β0.001	0.25 β0.001	0.20 β0.001
Ποτά		1.00	0.29 β0.001	0.17 β0.001	0.19 β0.001	0.11 β0.001
Ένδυση			1.00	0.36 β0.001	0.42 β0.001	0.27 β0.001
Στέγαση				1.00	0.31 β0.001	0.24 β0.001
Οικιακή χρήση					1.00	0.25 β0.001
Υγεία						1.00
Μεταφορές						
Επικοινωνίες						
Αναψυχή						
Εκπαίδευση						
Εστιατόρια						
Αγαθά						

	Μεταφορές	Επικοινωνίες	Αναψυχή	Εκπαίδευση	Εστιατόρια	Αγαθά
Διατροφή	0.26 β0.001	0.42 β0.001	0.30 β0.001	0.28 β0.001	0.37 β0.001	0.36 β0.001
Ποτά	0.20 β0.001	0.29 β0.001	0.23 β0.001	0.13 β0.001	0.33 β0.001	0.24 β0.001
Ένδυση	0.37 β0.001	0.46 β0.001	0.45 β0.001	0.29 β0.001	0.55 β0.001	0.53 β0.001
Στέγαση	0.30 β0.001	0.47 β0.001	0.34 β0.001	0.31 β0.001	0.38 β0.001	0.39 β0.001
Οικιακή χρήση	0.24 β0.001	0.28 β0.001	0.35 β0.001	0.21 β0.001	0.35 β0.001	0.33 β0.001
Υγεία	0.18 β0.001	0.21 β0.001	0.29 β0.001	0.14 β0.001	0.21 β0.001	0.27 β0.001
Μεταφορές	1.00	0.36 β0.001	0.34 β0.001	0.24 β0.001	0.38 β0.001	0.32 β0.001
Επικοινωνίες		1.00	0.41 β0.001	0.36 β0.001	0.55 β0.001	0.46 β0.001
Αναψυχή			1.00	0.37 β0.001	0.43 β0.001	0.37 β0.001
Εκπαίδευση				1.00	0.31 β0.001	0.24 β0.001
Εστιατόρια					1.00	0.46 β0.001
Αγαθά						1.00

Σε κάθε κελί:

Τιμή του συντελεστή συσχέτισης

P – τιμή

Από τον Πίνακα 2.19 παρατηρούνται στατιστικά σημαντικές γραμμικές συσχετίσεις μεταξύ των ποσοτικών μεταβλητών. Η μηδενική υπόθεση H_0 για το συντελεστή συσχέτισης Pearson απορρίπτεται και κατά συνέπεια, τα δεδομένα έχουν συσχέτιση.

Εντολές στην R

Προκειμένου να γίνει ο έλεγχος της γραμμικής συσχέτισης για τις ποσοτικές μεταβλητές της εν λόγω έρευνας χρησιμοποιήθηκε η εντολή `rcorr()` του πακέτου `Hmisc`, για να κατασκευαστεί ο πίνακας συσχετίσεων.

```
> Y<- dataHH[,9:20]
> Y<-as.matrix (Y)
> rcorr (Y)
      food alcohol clothing housing equipment health transport
food      1.00  0.32   0.42   0.30   0.25  0.20   0.26
alcohol    0.32  1.00   0.29   0.17   0.19  0.11   0.20
clothing    0.42  0.29   1.00   0.36   0.42  0.27   0.37
housing     0.30  0.17   0.36   1.00   0.31  0.24   0.30
equipment    0.25  0.19   0.42   0.31   1.00  0.25   0.24
health      0.20  0.11   0.27   0.24   0.25  1.00   0.18
transport    0.26  0.20   0.37   0.30   0.24  0.18   1.00
communication 0.42  0.29   0.46   0.47   0.28  0.21   0.36
culture      0.30  0.23   0.45   0.34   0.35  0.29   0.34
education    0.28  0.13   0.29   0.31   0.21  0.14   0.24
hotels       0.37  0.33   0.55   0.38   0.35  0.21   0.38
goods        0.36  0.24   0.53   0.39   0.33  0.27   0.32
      communication culture education hotels goods
food            0.42  0.30   0.28  0.37  0.36
alcohol          0.29  0.23   0.13  0.33  0.24
clothing         0.46  0.45   0.29  0.55  0.53
housing          0.47  0.34   0.31  0.38  0.39
equipment        0.28  0.35   0.21  0.35  0.33
health           0.21  0.29   0.14  0.21  0.27
transport        0.36  0.34   0.24  0.38  0.32
communication    1.00  0.41   0.36  0.55  0.46
culture           0.41  1.00   0.37  0.43  0.37
education        0.36  0.37   1.00  0.31  0.24
hotels           0.55  0.43   0.31  1.00  0.46
goods            0.46  0.37   0.24  0.46  1.00
```

n= 3512

P

```
      total food alcohol clothing housing equipment health transport
total      0  0  0  0  0  0  0
food       0  0  0  0  0  0  0
alcohol    0  0  0  0  0  0  0
clothing   0  0  0  0  0  0  0
housing    0  0  0  0  0  0  0
```

equipment	0	0	0	0	0		0	0
health	0	0	0	0	0	0	0	
transport	0	0	0	0	0	0	0	
communication	0	0	0	0	0	0	0	0
culture	0	0	0	0	0	0	0	0
education	0	0	0	0	0	0	0	0
hotels	0	0	0	0	0	0	0	0
goods	0	0	0	0	0	0	0	0
	communication culture education hotels goods							
total	0		0	0	0	0		
food	0		0	0	0	0		
alcohol	0		0	0	0	0		
clothing	0		0	0	0	0		
housing	0		0	0	0	0		
equipment	0		0	0	0	0	0	
health	0		0	0	0	0		
transport	0		0	0	0	0		
communication			0	0	0	0	0	
culture	0		0	0	0	0		
education	0		0		0	0		
hotels	0		0	0		0		
goods	0		0	0	0			

ΚΕΦΑΛΑΙΟ 3

ΣΤΑΤΙΣΤΙΚΑ ΜΟΝΤΕΛΑ ΠΑΛΙΝΔΡΟΜΗΣΗΣ

3.1. Εισαγωγή

Μεγάλο μέρος της επιστήμης της Στατιστικής ασχολείται με την κατασκευή και την ανάλυση στατιστικών μοντέλων (ή προτύπων). Με τον όρο μοντέλο καλείται η μορφή της σχέσης μεταξύ δύο ή περισσότερων μεταβλητών.

Το στατιστικό μοντέλο αποτελεί ένα πολύ σημαντικό εργαλείο της Στατιστικής για πολλαπλούς λόγους:

- Αρκετά συχνά σκοπός της ανάλυσης είναι να βρεθεί ένας τρόπος μελλοντικής πρόβλεψης της τιμής μιας μεταβλητής.
- Μπορεί να μην είναι φανερό ποιες μεταβλητές είναι πράγματι σημαντικές για την επεξήγηση ενός φαινομένου και με τη σύγκριση διαφόρων μοντέλων μπορεί να εξαχθεί ένα συμπέρασμα.
- Εξίσου σημαντική είναι και η διαπίστωση ότι δύο μεταβλητές δεν σχετίζονται μεταξύ τους.
- Τέλος, υπάρχουν περιπτώσεις στις οποίες η μορφή του στατιστικού μοντέλου προτείνεται βάσει της θεωρίας και απαιτείται η επιβεβαίωση του εν λόγω μοντέλου.

3.2. Γραμμική Παλινδρόμηση

Σε πολλές στατιστικές εφαρμογές συναντάται το πρόβλημα της μελέτης της σχέσης δύο ή περισσότερων τυχαίων μεταβλητών. Παραδείγματα τέτοιας σχέσης είναι η μελέτη του ύψους και του βάρους μιας ομάδας ανθρώπων, το εισόδημα και η καταναλωτική δαπάνη των εργαζομένων κλπ. Το πρόβλημα που πρέπει να επιλυθεί είναι αν όντως υπάρχει σχέση σύνδεσης μεταξύ των εξεταζόμενων μεταβλητών και στη συνέχεια, να προσδιοριστεί η σχέση αυτή βάσει ορισμένων παρατηρήσεων που έχουν συλλεχθεί μέσω του δείγματος της έρευνας. Ένας από τους κύριους λόγους που η μελέτη αυτή είναι σημαντική, είναι ότι τα αποτελέσματά της χρησιμοποιούνται συχνά για μελλοντικές προβλέψεις.

Σκοπός της Γραμμικής Παλινδρόμησης (Linear Regression) είναι η κατασκευή ενός μοντέλου, το οποίο να περιγράφει ικανοποιητικά τη σχέση μεταξύ μιας εξαρτημένης συνεχούς μεταβλητής Y και μίας ή περισσότερων συνεχών ανεξάρτητων μεταβλητών $X_1, X_2, \dots, X_p$. Έστω γνωστό $\mathbf{X} = (x_1, x_2, \dots, x_p)^T$. Αν η μέση τιμή της εξαρτημένης μεταβλητής Y (μεταβλητή απόκρισης), δεδομένων των τιμών των ανεξάρτητων μεταβλητών (επεξηγηματικές μεταβλητές), είναι γραμμική συνάρτηση των $X_1, X_2, \dots, X_p$ τότε η περιγραφή της σχέσης αυτής γίνεται βάσει ενός γραμμικού μοντέλου. Στην περίπτωση που το μοντέλο περιλαμβάνει δύο μεταβλητές, την ανεξάρτητη ή επεξηγηματική X και την εξαρτημένη ή μεταβλητή απόκριση Y οι οποίες συνδέονται με τη γραμμική συνάρτηση παλινδρόμησης, καλείται απλό γραμμικό μοντέλο. Όμως, όταν το μοντέλο περιλαμβάνει την

ταυτόχρονη ύπαρξη περισσότερων από δύο ανεξάρτητων ή επεξηγηματικών μεταβλητών καλείται πολλαπλό γραμμικό μοντέλο, το οποίο είναι, εν γένει, της μορφής:

$$E[Y | X_1, X_2, \dots, X_p] = b_0 + b_1x_1 + b_2x_2 + \dots + b_px_p. \quad (3.1)$$

Συνήθως η ανεξάρτητη ή οι ανεξάρτητες μεταβλητές δεν είναι γνωστές εκ των προτέρων, αλλά πρέπει να επιλεγούν οι καταλληλότερες από ένα σύνολο μεταβλητών του δείγματος.

Η διαδικασία δημιουργίας μιας μαθηματικής εξίσωσης για την περιγραφή ενός φαινομένου μπορεί να είναι ιδιαίτερα περίπλοκη. Αυτό οφείλεται στο γεγονός ότι για την κατασκευή του μοντέλου απαιτείται κάποια γνώση της φύσης της σχέσης μεταξύ των μεταβλητών.

Η διαδικασία προσδιορισμού του μοντέλου της Γραμμικής Παλινδρόμησης στοχεύει να περιγραφεί με τον βέλτιστο τρόπο η δοθείσα πληροφορία μέσω των δεδομένων. Στην πραγματικότητα, όμως, κανένα μοντέλο, και κατά συνέπεια κανένα γραμμικό μοντέλο δεν μπορεί να περιγράψει πλήρως το σύνολο των πληροφοριών του εξεταζόμενου σετ δεδομένων. Όσο καλά προσαρμοσμένη και αν είναι η γραμμή της (πολλαπλής) παλινδρόμησης στα δεδομένα, πάντα θα υπάρχει ένα μέρος της πληροφορίας που θα εξακολουθεί να μην ερμηνεύεται μέσω του μοντέλου. Ο παράγοντας που δεν ερμηνεύεται από το γραμμικό μοντέλο ονομάζεται σφάλμα της παλινδρόμησης και βέλτιστα δεν κρύβει μέσα του καμία συστηματική σχέση.

Επομένως, λαμβάνοντας υπ' όψιν και τον παράγοντα του τυχαίου σφάλματος το πολλαπλό μοντέλο της Γραμμικής Παλινδρόμησης λαμβάνει τη μορφή:

$$Y = b_0 + b_1X_1 + b_2X_2 + \dots + b_pX_p + \varepsilon \quad (3.2)$$

όπου ε είναι το τυχαίο σφάλμα.

Πριν την εφαρμογή της τεχνικής της Πολλαπλής Γραμμικής Παλινδρόμησης πρέπει να πραγματοποιηθεί αναλυτική διερεύνηση του σετ των δεδομένων στο οποίο θα εφαρμοστεί η τεχνική, να μελετηθεί το είδος της εξαρτημένης και των ανεξάρτητων μεταβλητών, η ποιότητα των δεδομένων καθώς και η ύπαρξη ελλειψουσών τιμών.

Η Πολλαπλή Γραμμική Παλινδρόμηση δεν μπορεί να εφαρμοστεί σε οποιαδήποτε δεδομένα. Από τα πρώτα βήματα περιγραφής της τεχνικής αναφέρθηκε ότι μελετά τη σχέση μεταξύ μίας εξαρτημένης συνεχούς μεταβλητής και μίας ή περισσότερων ανεξάρτητων μεταβλητών.

Αναλυτικότερα για την εφαρμογή της Πολλαπλής Γραμμικής Παλινδρόμησης πρέπει να ισχύουν οι παρακάτω υποθέσεις:

- Η εξαρτημένη μεταβλητή, με δεδομένες τις ανεξάρτητες μεταβλητές, πρέπει να είναι συνεχής και να ακολουθεί την κανονική κατανομή με σταθερή διακύμανση, δηλαδή $Y | X_1, X_2, \dots, X_p \sim N(b_0 + b_1x_1 + \dots + b_px_p, \sigma^2)$.

- Οι ανεξάρτητες μεταβλητές μπορούν να είναι είτε συνεχείς είτε κατηγορικές. Στην περίπτωση των κατηγορικών μεταβλητών αυτές εισάγονται στο μοντέλο με τη μορφή των ψευδομεταβλητών.
- Η κάθε ανεξάρτητη μεταβλητή πρέπει να συνδέεται γραμμικά με τη μέση τιμή της εξαρτημένης μεταβλητής.
- Οι ανεξάρτητες μεταβλητές πρέπει να είναι ισχυρά συσχετισμένες με την εξαρτημένη μεταβλητή ενώ μεταξύ τους πρέπει να μην εμφανίζουν ισχυρή συσχέτιση.
- Για το τυχαίο σφάλμα πρέπει να ισχύει ότι $\varepsilon \sim N(0, \sigma^2)$.

Η μη ικανοποίηση κάποιων υποθέσεων κάνει την τεχνική λιγότερο αξιόπιστη και τα αποτελέσματα αμφισβητήσιμα.

3.2.1. Θεωρητικό υπόβαθρο της Γραμμικής Παλινδρόμησης

Έστω η εξαρτημένη μεταβλητή Y , η οποία θεωρείται ότι επηρεάζεται από περισσότερες από μία επεξηγηματικές μεταβλητές (έστω p). Μπορεί να χρησιμοποιηθεί ένα γραμμικό μοντέλο για να διερευνηθεί η εξάρτηση της Y από τις $X_1, X_2, \dots, X_p$. Το μοντέλο αυτό, το οποίο αποτελεί γενίκευση του απλού μοντέλου $Y = b_0 + b_1X + \varepsilon$, θα έχει τη μορφή:

$$Y = b_0 + b_1X_1 + b_2X_2 + \dots + b_pX_p + \varepsilon \quad (3.3)$$

για κάποιες παραμέτρους $b_0, b_1, b_2, \dots, b_p$.

Για να διερευνηθεί η σχέση μεταξύ της εξαρτημένης και των ανεξάρτητων μεταβλητών λαμβάνεται δείγμα μεγέθους n και για κάθε μονάδα του δείγματος καταγράφονται οι τιμές των συγκεκριμένων μεταβλητών. Έτσι, εφαρμόζεται το μοντέλο:

$$Y_i = b_0 + b_1X_{i1} + b_2X_{i2} + \dots + b_pX_{ip} + \varepsilon_i, \quad i = 1, 2, \dots, n \quad (3.4)$$

όπου τα «σφάλματα» $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$ θεωρούνται ανεξάρτητες και ισόνομες τυχαίες μεταβλητές από την $N(0, \sigma^2)$.

Το παραπάνω μοντέλο γράφεται με τη βοήθεια πινάκων στην απλούστερη μορφή:

$$\mathbf{Y} = \mathbf{X}\mathbf{b} + \boldsymbol{\varepsilon} \quad (3.5)$$

$$\text{Όπου } \mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} 1 & X_{11} & \dots & X_{1,p} \\ 1 & X_{21} & \dots & X_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & \dots & X_{n,p} \end{bmatrix} \quad \mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_p \end{bmatrix} \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Σύμφωνα με τα παραπάνω πρέπει να ισχύουν οι υποθέσεις των Gauss – Markov :

Υποθέσεις Gauss – Markov για το τυχαίο διάνυσμα $\boldsymbol{\varepsilon}$:

- Το $\boldsymbol{\varepsilon}$ είναι τυχαίο διάνυσμα.
- $E(\boldsymbol{\varepsilon}) = \mathbf{0}$.
- $Var(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I}_n$.

Δηλαδή ε_i είναι ανεξάρτητες τυχαίες μεταβλητές και $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n)$.

Υποθέσεις Gauss – Markov για το τυχαίο διάνυσμα $\mathbf{Y}$:

- Το $\mathbf{Y} | \mathbf{X}$ είναι τυχαίο διάνυσμα.
- $E(\mathbf{Y} | \mathbf{X}) = \mathbf{X}\mathbf{b}$.
- $Var(\mathbf{Y} | \mathbf{X}) = Var(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I}_n$.

Δηλαδή ισχύει ότι $\mathbf{Y} | \mathbf{X} \sim N(\mathbf{X}\mathbf{b}, \sigma^2 \mathbf{I}_n)$.

3.2.2. Εκτίμηση των παραμέτρων $\mathbf{b}_0, \mathbf{b}_1, \dots, \mathbf{b}_p$ και σ^2

Το τυχαίο σφάλμα $\boldsymbol{\varepsilon}$ αποτελείται από n ανεξάρτητες τυχαίες μεταβλητές οι οποίες ακολουθούν την $N(0, \sigma^2)$ και επομένως, θα έχουν από κοινού συνάρτηση πυκνότητας πιθανότητας $N(\mathbf{0}, \sigma^2 \mathbf{I}_n)$, δηλαδή ακολουθεί μία πολυδιάστατη κανονική κατανομή όπου $\mathbf{I}_n$ είναι ο μοναδιαίος πίνακας διάστασης n .

Η πολυδιάστατη κανονική κατανομή δίνεται από τον τύπο:

$$f(X_1, X_2, \dots, X_k) = \frac{1}{(2\pi)^{k/2} |\boldsymbol{\Sigma}|^{1/2}} e^{\frac{1}{2}(\mathbf{X}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{X}-\boldsymbol{\mu})} \quad (3.6)$$

$$\text{όπου } \mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_k \end{bmatrix} \quad \boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_k \end{bmatrix} \quad \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1k} \\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{k1} & \sigma_{k2} & \dots & \sigma_{kk} \end{bmatrix}.$$

Επομένως ως συνέπεια το τυχαίο διάνυσμα $\mathbf{Y} | \mathbf{X}$ θα ακολουθεί την πολυδιάστατη κανονική κατανομή $\mathbf{Y} \sim N(\mathbf{X}\mathbf{b}, \sigma^2 \mathbf{I}_n)$, με συνάρτηση πιθανοφάνειας:

$$L(\mathbf{b}, \sigma^2) = f(y_1, y_2, \dots, y_n; \mathbf{b}, \sigma^2) = \frac{1}{(2\pi)^{n/2} (\sigma^2)^{n/2}} e^{-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\mathbf{b})^T(\mathbf{y}-\mathbf{X}\mathbf{b})} \quad (3.7)$$

η οποία μεγιστοποιείται ως προς $\mathbf{b}$ (Εκτιμητής Μεγίστης Πιθανοφάνειας του $\mathbf{b}$) όταν ελαχιστοποιείται το :

$$(\mathbf{Y} - \mathbf{X}\mathbf{b})^T (\mathbf{Y} - \mathbf{X}\mathbf{b}) = \boldsymbol{\varepsilon}^T \boldsymbol{\varepsilon} = \sum_{i=1}^n \varepsilon_i^2. \quad (3.8)$$

Το παραπάνω γράφεται:

$$(\mathbf{Y}^T - \mathbf{b}^T \mathbf{X}^T) (\mathbf{Y} - \mathbf{X}\mathbf{b}) = \mathbf{Y}^T \mathbf{Y} - \mathbf{Y}^T \mathbf{X}\mathbf{b} - \mathbf{b}^T \mathbf{X}^T \mathbf{Y} + \mathbf{b}^T \mathbf{X}^T \mathbf{X}\mathbf{b} \quad (3.9)$$

και παραγωγίζοντας ως προς $\mathbf{b}$ προκύπτει η παρακάτω παράγωγος:

$$\frac{d}{d\mathbf{b}} (\mathbf{Y} - \mathbf{X}\mathbf{b})^T (\mathbf{Y} - \mathbf{X}\mathbf{b}) = -2\mathbf{X}^T \mathbf{Y} + 2\mathbf{X}^T \mathbf{X}\mathbf{b} \quad (3.10)$$

η οποία είναι ίση με το 0 όταν:

$$\mathbf{X}^T \mathbf{X}\mathbf{b} = \mathbf{X}^T \mathbf{Y}. \quad (3.11)$$

Το σύστημα των $p + 1$ εξισώσεων με p αγνώστους $\mathbf{X}^T \mathbf{X}\mathbf{b} = \mathbf{X}^T \mathbf{Y}$ έχει μοναδική λύση όταν υπάρχει ο αντίστροφος του $\mathbf{X}^T \mathbf{X}$ και σε αυτή την περίπτωση οι εκτιμήτριες μεγίστης πιθανοφάνειας για το $\mathbf{b} = (b_0, b_1, \dots, b_p)^T$ θα είναι:

$$\hat{\mathbf{b}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}. \quad (3.12)$$

Η παραπάνω μέθοδος συμπίπτει με τη μέθοδο των ελαχίστων τετραγώνων.

Ο εκτιμητής μεγίστης πιθανοφάνειας για το σ^2 δίνεται από τον τύπο:

$$\hat{\sigma}^2 = \frac{1}{n} (\mathbf{Y} - \mathbf{X}\hat{\mathbf{b}})^T (\mathbf{Y} - \mathbf{X}\hat{\mathbf{b}}) = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{b}_0 - \hat{b}_1 X_{i1} - \dots - \hat{b}_p X_{ip})^2. \quad (3.13)$$

Οι προβλεπόμενες τιμές για τα Y_i , όπως και στο απλό γραμμικό μοντέλο, δίνονται από τις εκτιμήσεις των:

$$E(Y_i | X_i) = b_0 + b_1 X_{i1} + b_2 X_{i2} + \dots + b_p X_{ip} \quad (3.14)$$

δηλαδή:

$$\hat{y}_i = \hat{b}_0 + \hat{b}_1 x_{i1} + \dots + \hat{b}_p x_{ip}, \quad i = 1, 2, \dots, n \quad (3.15)$$

και συνοπτικά η σχέση μεταξύ πινάκων γίνεται:

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\mathbf{b}} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} = \mathbf{P}\mathbf{Y} \quad \text{όπου } \mathbf{P} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T. \quad (3.16)$$

Οι διαφορές των εκτιμώμενων $\hat{y}_i$ από τις παρατηρούμενες y_i καλούνται υπόλοιπα (residuals) ή εκτιμώμενα σφάλματα και συμβολίζονται με

$$\hat{\varepsilon}_i = y_i - \hat{y}_i \quad \text{με } i = 1, 2, \dots, n. \quad (3.17)$$

Η παραπάνω σχέση με πίνακες γράφεται:

$$\hat{\varepsilon} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\hat{\mathbf{b}} = \mathbf{Y} - \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} = (\mathbf{I}_n - \mathbf{P})\mathbf{Y} \quad (3.18)$$

όπου $\mathbf{P}$ είναι ο γνωστός πίνακας $\mathbf{P} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$.

3.2.3. Έλεγχοι υποθέσεων και διαστήματα εμπιστοσύνης για τις παραμέτρους του μοντέλου

Υποθέτοντας ότι το διάνυσμα των σφαλμάτων $\varepsilon \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n)$ αποδεικνύεται ότι:

$$\hat{\mathbf{b}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} \sim N(\mathbf{b}, \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1}) \quad (3.19)$$

δηλαδή το τυχαίο διάνυσμα $\hat{\mathbf{b}}$ ακολουθεί την πολυδιάστατη κανονική κατανομή με μέση τιμή τον πίνακα $\mathbf{b} = (b_0, b_1, \dots, b_p)^T$ και πίνακα διασποράς – συνδιασποράς τον $\sigma^2 (\mathbf{X}^T \mathbf{X})^{-1}$.

Επίσης αποδεικνύεται ότι:

$$\frac{1}{\sigma^2} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \frac{1}{\sigma^2} (\mathbf{Y} - \mathbf{X}\hat{\mathbf{b}})^T (\mathbf{Y} - \mathbf{X}\hat{\mathbf{b}}) = \frac{1}{\sigma^2} \mathbf{Y}^T (\mathbf{I}_n - \mathbf{P}) \mathbf{Y} \sim X_{n-p-1}^2. \quad (3.20)$$

Επομένως συνήθως χρησιμοποιείται ως εκτιμητής του σ^2 ο αμερόληπτος εκτιμητής

$$\hat{\sigma}^2 = S^2 = \frac{1}{n-p-1} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2. \quad (3.21)$$

Αν τα διαγώνια στοιχεία του πίνακα $(\mathbf{X}^T \mathbf{X})^{-1}$ συμβολίζονται με c_{ii} , $i = 1, 2, \dots, p$ τότε προφανώς $\hat{b}_i \sim N(b_i, \sigma^2 c_{ii})$, $i = 1, 2, \dots, p$. Από το γεγονός ότι $\frac{(n-p)S^2}{\sigma^2} \sim X_{n-p-1}^2$ με S^2 ανεξάρτητη των $\hat{b}_i$ και ότι $\hat{b}_i \sim N(b_i, \sigma^2 c_{ii})$, $i = 1, 2, \dots, p$ προκύπτει ότι:

$$\frac{\hat{b}_i - b_i}{S\sqrt{c_{ii}}} \sim t_{n-p-1}, \quad i = 1, 2, \dots, p. \quad (3.22)$$

Επομένως, τα διαστήματα εμπιστοσύνης (δ.ε.) για τα $b_0, b_1, \dots, b_p$ αντίστοιχα σε επίπεδο σημαντικότητας $1 - \alpha$ δίνονται από τον τύπο:

$$\left(\hat{b}_i - S\sqrt{c_{ii}} t_{n-p-1, (\frac{\alpha}{2})}, \hat{b}_i + S\sqrt{c_{ii}} t_{n-p-1, (\frac{\alpha}{2})} \right), \quad i = 1, 2, \dots, p. \quad (3.23)$$

Για το δίπλευρο έλεγχο της υπόθεσης $H_0: b_i = 0$ σε επίπεδο σημαντικότητας α προκύπτουν οι αντίστοιχες περιοχές απόρριψης:

$$K: |T_i| > t_{n-p-1, (\frac{\alpha}{2})} \quad \text{όπου } T_i = \frac{\hat{b}_i}{S\sqrt{c_{ii}}}, \quad i = 1, 2, \dots, p. \quad (3.24)$$

Αν για κάποιο i απορριφθεί η υπόθεση $H_0: b_i = 0$, δοθέντων των υπολοίπων επεξηγηματικών μεταβλητών, τότε προκύπτει ότι η μεταβλητή Y εξαρτάται γραμμικά από τη X_i .

3.2.4. Ερμηνεία της συνολικής μεταβλητότητας του μοντέλου

Είναι σημαντικό, επίσης, να εξεταστεί κατά πόσο η μεταβλητότητα των τιμών της εξαρτημένης μεταβλητής Y εξηγείται από τη μεταβλητότητα των ανεξάρτητων μεταβλητών $X_i, i = 1, 2, \dots, p$.

Το συνολικό άθροισμα τετραγώνων (SST – total sum of squares) μπορεί να αναλυθεί σε άθροισμα δύο συνιστωσών μεταβλητότητας, του αθροίσματος τετραγώνων λόγω παλινδρόμησης (SSR – sum of squares due to regression) και του αθροίσματος τετραγώνων λόγω υπολοίπων ή σφάλματος (SSE – sum of squares due to error or residuals sum of squares) ως εξής:

$$\sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \quad (3.25)$$

τα οποία συμβολίζονται αντίστοιχα:

$$SST = \sum_{i=1}^n (Y_i - \bar{Y})^2 \quad SSE = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \quad SSR = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2.$$

Το SST εκφράζει τη συνολική παρατηρούμενη μεταβλητότητα των Y_i , το SSR εκφράζει τη μεταβλητότητα των εκτιμώμενων τιμών $\hat{Y}_i$ και το SSE εκφράζει τη μεταβλητότητα των Y_i σε σχέση με τις αντίστοιχες εκτιμώμενες τιμές $\hat{Y}_i$. Η μεταβλητότητα για το SSR ερμηνεύεται από το μοντέλο ενώ για το SSE όχι.

Το η^2 καλείται συντελεστής ή δείκτης προσδιορισμού και εκφράζει το ποσοστό της μεταβλητότητας των τ.μ. Y_i που ερμηνεύεται από το μοντέλο. Οι τιμές, τις οποίες μπορεί να λάβει ο R^2 , είναι μεταξύ του 0 και του 1. Όσο πιο «κοντά» στο 1 είναι τόσο ισχυρότερη η γραμμική σχέση εξάρτησης των τ.μ. Y και X είναι.

Προηγουμένως αναφέρθηκε ότι:

$$\frac{1}{\sigma^2} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \sim X_{n-p-1}^2 \quad (3.26)$$

δηλαδή:

$$\frac{SSE}{\sigma^2} \sim X_{n-p-1}^2. \quad (3.27)$$

Επίσης αν $b_1 = b_2 = \dots = b_p = 0$ αποδεικνύεται ότι:

$$\frac{SSR}{\sigma^2} \sim X_p^2 \quad \text{και} \quad \frac{SST}{\sigma^2} \sim X_{n-1}^2. \quad (3.28)$$

Επομένως αν $b_1 = b_2 = \dots = b_p = 0$ τότε:

$$F = \frac{\frac{SSR}{\sigma^2}/p}{\frac{SSE}{\sigma^2}/(n-p-1)} = \frac{SSR/p}{SSE/(n-p-1)} \sim F_{p,n-p-1}. \quad (3.29)$$

Σημειώνεται ότι SSR και SSE είναι ανεξάρτητες τυχαίες μεταβλητές.

Από τα παραπάνω κατασκευάζεται ένας έλεγχος για την υπόθεση $H_0 : b_1 = b_2 = \dots = b_p = 0$, δηλαδή ότι η Y δεν εξαρτάται γραμμικά από καμία από τις $X_1, X_2, \dots, X_p$. Η μηδενική υπόθεση απορρίπτεται όταν η στατιστική συνάρτηση F λαμβάνει μεγάλες τιμές υπό την H_0 με επίπεδο σημαντικότητας α , δηλαδή όταν:

$$F = \frac{SSR/p}{SSE/(n-p-1)} > F_{p,n-p-1}(\alpha) \quad (3.30)$$

όπου το $F_{p,n-p-1}(\alpha)$ είναι το άνω α - σημείο της κατανομής F με p και $n-p-1$ βαθμούς ελευθερίας.

Όλες οι παραπάνω ποσότητες συνοψίζονται στον Πίνακα 3.1.

Πίνακας 3.1: Πίνακας ανάλυσης διασποράς για ένα μοντέλο γραμμικής παλινδρόμησης με p επεξηγηματικές μεταβλητές, ANOVA

Πηγή μεταβλητότητας	Άθροισμα τετραγώνων	Βαθμοί ελευθερίας	Μέσο άθροισμα τετραγώνων	Έλεγχος F
Παλινδρόμηση	$SSR = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$	p	$MSR = \frac{SSR}{p}$	$\frac{MSR}{MSE}$
Υπόλοιπα	$SSE = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$	$n-p-1$	$MSE = \frac{SSE}{n-p-1}$	
Σύνολο	$SST = \sum_{i=1}^n (Y_i - \bar{Y})^2$	$n-1$		

3.2.5. Σημειακή και Αναμενόμενη Μέση πρόβλεψη της Y

Αφού έχουν εκτιμηθεί οι συντελεστές $b_0, b_1, b_2, \dots, b_p$ μπορεί να γίνει πρόβλεψη της τιμής της μεταβλητής Y για τις τιμές των μεταβλητών $X_1 = x_1, X_2 = x_2, \dots, X_p = x_p$ δηλαδή:

$$\hat{y} = \hat{b}_0 + \hat{b}_1 x_1 + \dots + \hat{b}_p x_p. \quad (3.31)$$

Η τιμή της Y μπορεί να προσδιοριστεί επίσης μέσω ενός διαστήματος εμπιστοσύνης.

1. Το διάστημα αναμενόμενης μέσης πρόβλεψης της τ.μ. Y όταν: $X_1 = x_1, X_2 = x_2, \dots, X_p = x_p$ είναι ένα διάστημα εμπιστοσύνης με επίπεδο

σημαντικότητας $1 - \alpha$ για το $E(Y | X = x) = b_0 + b_1x_1 + b_2x_2 + \dots + b_px_p = \mathbf{x}^T \mathbf{b}$ όπου $\mathbf{x} = [1, x_1, x_2, \dots, x_p]^T$ και δίνεται από τον τύπο:

$$\left(\mathbf{x}^T \hat{\mathbf{b}} - S\sqrt{\mathbf{x}^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}} \cdot t_{n-p-1, (\frac{\alpha}{2})}, \mathbf{x}^T \hat{\mathbf{b}} + S\sqrt{\mathbf{x}^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}} \cdot t_{n-p-1, (\frac{\alpha}{2})} \right). \quad (3.32)$$

2. Το διάστημα σημειακής πρόβλεψης της τ.μ. Y όταν: $X_1 = x_1, X_2 = x_2, \dots, X_p = x_p$ είναι ένα διάστημα μέσα στο οποίο βρίσκεται η τιμή της τ.μ. Y με βάσει το μοντέλο παλινδρόμησης με πιθανότητα $1 - \alpha$:

$$\left(\mathbf{x}^T \hat{\mathbf{b}} - S\sqrt{1 + \mathbf{x}^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}} \cdot t_{n-p, (\frac{\alpha}{2})}, \mathbf{x}^T \hat{\mathbf{b}} + S\sqrt{1 + \mathbf{x}^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}} \cdot t_{n-p, (\frac{\alpha}{2})} \right). \quad (3.33)$$

Ωστόσο, είναι σημαντικό να σημειωθεί ότι οι προβλέψεις είναι σωστές και αξιόπιστες μόνο για τιμές των μεταβλητών $X_i, i = 1, 2, \dots, p$, οι οποίες είναι σχετικά κοντά με αυτές του δείγματος.

3.2.6. Εξέταση της ορθότητας του μοντέλου

Στην πολλαπλή γραμμική παλινδρόμηση θα πρέπει να είναι βέβαιο ότι οι παρατηρήσεις προσαρμόζονται ικανοποιητικά στο πολλαπλό μοντέλο, έτσι ώστε τα συμπεράσματα που προκύπτουν να θεωρούνται αξιόπιστα. Οι προϋποθέσεις αυτές είναι η γραμμικότητα, η κανονικότητα των σφαλμάτων, η ομοσκεδαστικότητα και η ανεξαρτησία των σφαλμάτων. Πιο αναλυτικά, στο πολλαπλό γραμμικό μοντέλο η δεσμευμένη μέση τιμή της μεταβλητής απόκρισης Y πρέπει να συνδέεται γραμμικά με τις επεξηγηματικές μεταβλητές $\mathbf{X}$ (γραμμικότητα). Ωστόσο, για αξιόπιστες μελλοντικές προβλέψεις και στατιστική συμπερασματολογία των παραμέτρων του μοντέλου πρέπει τα τυχαία σφάλματα να ακολουθούν κανονική κατανομή (κανονικότητα σφαλμάτων). Επίσης, θα πρέπει τα τυχαία σφάλματα να είναι ανεξάρτητες τυχαίες μεταβλητές (ανεξαρτησία σφαλμάτων). Επιπλέον, θα πρέπει η διασπορά της δεσμευμένης κατανομής της τ.μ. Y , δεδομένων των τιμών των τ.μ. $\mathbf{X}$, να παραμένει σταθερή ανεξάρτητα των τιμών των τ.μ. $\mathbf{X}$ (ομοσκεδαστικότητα). Διαφορετικά, θα πρέπει να τροποποιηθεί κατάλληλα το μοντέλο. Για να είναι σωστή η σύγκριση απαιτείται όλα τα υπόλοιπα να έχουν την ίδια κατανομή. Ο έλεγχος της ορθότητας του πολλαπλού μοντέλου δε γίνεται με τα συνήθη υπόλοιπα αλλά με τη χρήση των τυποποιημένων υπολοίπων, τα οποία ορίζονται ως εξής:

$$\hat{\varepsilon}_i = \frac{\hat{\varepsilon}_i}{S\sqrt{1 - p_{ii}}} \quad i = 1, 2, \dots, n \quad (3.34)$$

όπου τα p_{ii} είναι τα διαγώνια στοιχεία του πίνακα $\mathbf{P}$.

Αυτή η τυποποίηση αφαιρεί την ετεροσκεδαστικότητα των συνήθων υπολοίπων.

3.2.7. Πολυσυγγραμικότητα

Η ύπαρξη έντονης συσχέτισης μεταξύ δύο ή περισσότερων επεξηγηματικών μεταβλητών ονομάζεται φαινόμενο της πολυσυγγραμικότητας. Η παρουσία του φαινομένου αυτού οδηγεί σε αυξημένα τυπικά σφάλματα των $\hat{\mathbf{b}}$ και κατά συνέπεια, δυσκολεύει την εκτίμηση της επίδρασης της κάθε επεξηγηματικής μεταβλητής στην εξαρτημένη μεταβλητή, αφού τα διαστήματα εμπιστοσύνης των αντίστοιχων συντελεστών θα είναι μεγάλα σε εύρος. Επίσης, σε τέτοιες περιπτώσεις είναι δύσκολος και ο εντοπισμός των στατιστικά σημαντικών μεταβλητών.

Στην περίπτωση της πολυσυγγραμικότητας, όπου κάποιες από τις ανεξάρτητες μεταβλητές $X_1, X_2, \dots, X_p$ είναι έντονα γραμμικά εξαρτημένες, ο πίνακας $\mathbf{X}^T \mathbf{X}$ δεν αντιστρέφεται, ή ισοδύναμα η ορίζουσά του είναι 0. Αυτό αποτελεί πρόβλημα καθώς για τον υπολογισμό των εκτιμητριών:

$$\hat{\mathbf{b}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} \quad (3.35)$$

απαιτείται ο εν λόγω πίνακας να είναι αντιστρέψιμος.

Σε περιπτώσεις στις οποίες η πολυσυγγραμικότητα οφείλεται στη φύση των μεταβλητών, η ανάλυση παλινδρόμησης μπορεί να πραγματοποιηθεί μόνο μετά την αφαίρεση κάποιων ανεξάρτητων μεταβλητών από το μοντέλο. Έτσι, όμως, χάνεται «πληροφορία». Όταν εντοπιστούν ομάδα ή ομάδες από ισχυρά συσχετισμένες μεταβλητές, τότε θα πρέπει στο τελικό μοντέλο να υπολογισθεί μόνο μία από κάθε ομάδα.

3.2.8. Χρήση ψευδομεταβλητών ή εικονικών μεταβλητών

Σε ένα πολλαπλό γραμμικό μοντέλο η εξαρτημένη μεταβλητή ή μεταβλητή απόκρισης είναι μια μεταβλητή με τιμές γύρω από ένα συνεχές πεδίο τιμών. Ωστόσο, πολλές φορές μια ανεξάρτητη ή επεξηγηματική μεταβλητή είναι κατηγορική, όπως για παράδειγμα το φύλο ή το μορφωτικό επίπεδο στην πρόβλεψη μισθών υπαλλήλων.

Προκειμένου να γίνει ο έλεγχος ως προς το αν υπάρχουν διαφορές ανάμεσα σε δύο ή περισσότερες κατηγορίες του πληθυσμού δημιουργούνται οι λεγόμενες ψευδομεταβλητές ή εικονικές μεταβλητές, οι οποίες βοηθούν στο να διακρίνονται ποιες μετρήσεις αφορούν σε ποια κατηγορία. Στην παρούσα ανάλυση θα χρησιμοποιηθεί μια συγκεκριμένη οικογένεια ψευδομεταβλητών, οι λεγόμενες δείκτριες μεταβλητές.

Έστω μια κατηγορική μεταβλητή K με c κατηγορίες $1, 2, \dots, c$. Οι δείκτριες μεταβλητές είναι οι $c - 1$ μεταβλητές $X_1, X_2, \dots, X_{c-1}$, όπου

$$X_i = \begin{cases} 1, & \text{αν } K = i \\ 0, & \text{διαφορετικά} \end{cases} \quad i = 1, 2, \dots, c - 1$$

δηλαδή το $X_i = 1$ δείχνει ότι η παρατήρηση ανήκει στην i – οστή ομάδα του πληθυσμού, συνεπώς $X_j = 0$ για $i \neq j$.

Η χρήση των $c - 1$ αντι των c ψευδομεταβλητών εξηγείται με τον ακόλουθο τρόπο. Λόγω της σχέσης $\sum_{i=1}^c X_i = 1$, αν η παρατήρηση ανήκει σε μία από τις πρώτες $1, 2, \dots, c - 1$ κατηγορίες, τότε ισχύει ότι $X_c = 0$. Αν δεν ανήκει σε καμία από τις πρώτες $c - 1$ κατηγορίες, τότε ισχύει ότι $X_c = 1$. Έτσι, είναι εμφανές ότι η X_c είναι εντελώς προβλέψιμη από τις υπόλοιπες ψευδομεταβλητές και δεν μπορεί να προσφέρει καμία επιπλέον πληροφορία στο μοντέλο. Προφανώς, μπορεί να παραλειφθεί μια οποιαδήποτε από τις c ψευδομεταβλητές και όχι υποχρεωτικά την τελευταία.

Η ερμηνεία του συντελεστή b_i της δείκτριας μεταβλητής X_i είναι ότι η αναμενόμενη τιμή της τ.μ. Y αυξάνεται κατά b_i μονάδες, όταν η παρατήρηση προέρχεται από την κατηγορία i σε σύγκριση με μια παρατήρηση από την κατηγορία αναφοράς, δηλαδή αυτήν που δεν περιέχεται στην ανάλυση. Η επιλογή της κατηγορίας αναφοράς εξαρτάται από τη φύση του προβλήματος, γιατί συγκεκριμένες επιλογές μπορεί να οδηγήσουν σε μια καλύτερη ερμηνεία των συντελεστών παλινδρόμησης. Αυτό συμβαίνει όταν γίνονται συγκρίσεις με κάποια «ελεγχόμενη» ομάδα, η οποία χρησιμοποιείται ως κατηγορία αναφοράς.

Το γενικό κίνητρο για να συμπεριληφθεί μια εικονική μεταβλητή σ' ένα πρόβλημα παλινδρόμησης είναι το ίδιο με εκείνο που οδηγεί στο να συμπεριληφθεί μια ποσοτική ανεξάρτητη μεταβλητή. Με άλλα λόγια, να μελετηθεί καλύτερα η εξαρτημένη μεταβλητή με την ελάττωση της επίδρασης του παράγοντα που οφείλεται στα λάθη και να αποτραπεί μια μεροληπτική αποτίμηση της επίδρασης μιας ανεξάρτητης μεταβλητής που είναι απόρροια του ότι έχει παραληφθεί από το μοντέλο μια άλλη ανεξάρτητη μεταβλητή η οποία σχετίζεται με αυτή.

3.3. Ανάλυση Διασποράς

Η Ανάλυση Διασποράς (Analysis of Variance – ANOVA) είναι μια μέθοδος ισοδύναμη με την ανάλυση παλινδρόμησης, για τη μελέτη της σχέσης μεταξύ μίας εξαρτημένης μεταβλητής και μίας ή και περισσότερων ανεξάρτητων μεταβλητών, οι οποίες είναι κατηγορικές. Στόχος κάθε στατιστικού πειράματος είναι ο προσδιορισμός της επίδρασης μίας ή περισσότερων ανεξάρτητων μεταβλητών πάνω στην υπό μελέτη εξαρτημένη μεταβλητή. Στην Ανάλυση Διασποράς οι ανεξάρτητες μεταβλητές, οι οποίες ονομάζονται παράγοντες, είναι συνήθως ποιοτικές (π.χ. διαφορετικοί τύποι μπαταριών ως προς τη διάρκεια ζωής τους). Κάποιες φορές, όμως, μπορεί να είναι και ποσοτικές οι οποίες έχουν ομαδοποιηθεί (π.χ. ηλικία).

3.3.1. Ανάλυση Διασποράς κατά έναν Παράγοντα

Ο απλούστερος πειραματικός σχεδιασμός είναι εκείνος που χαρακτηρίζεται από την ανάλυση της διασποράς κατά έναν παράγοντα (one – way ANOVA), στην οποία υπάρχει μία ανεξάρτητη μεταβλητή.

3.3.1.1. Προσδιορισμός του μοντέλου

Στη γενική περίπτωση υποτίθεται ότι υπάρχει μια συνεχής τυχαία μεταβλητή Y (εξαρτημένη μεταβλητή ή μεταβλητή απόκρισης) και μια ανεξάρτητη μεταβλητή X (παράγοντας) με k στάθμες με n_j παρατηρήσεις ανά στάθμη (ή επίπεδο) για $j = 1, \dots, k$. Συμβολίζεται με Y_{ij} η i παρατήρηση της j στάθμης. Σκοπός είναι να γίνει έλεγχος αν κατά μέσο όρο η εξαρτημένη μεταβλητή διαφοροποιείται στις k στάθμες της κατηγορικής μεταβλητής. Ορίζονται, επίσης, τα ακόλουθα βασικά μεγέθη:

$$Y_{.j} = \sum_{i=1}^{n_j} Y_{ij}, \quad \text{το άθροισμα των παρατηρήσεων του } j \text{ επιπέδου}$$

$$\bar{Y}_{.j} = \frac{Y_{.j}}{n_j}, \quad \text{ο δειγματικός μέσος του } j \text{ επιπέδου}$$

$$Y_{..} = \sum_{j=1}^k Y_{.j} = \sum_{j=1}^k \sum_{i=1}^{n_j} Y_{ij}, \quad \text{το άθροισμα όλων των παρατηρήσεων}$$

$$\bar{Y}_{..} = \frac{Y_{..}}{N}, \quad \text{ο δειγματικός μέσος όλων των παρατηρήσεων}$$

$$\text{όπου } N = \sum_{j=1}^k n_j, \quad \text{ο συνολικός αριθμός των παρατηρήσεων.}$$

Το κατά έναν παράγοντα ανάλυσης διασποράς μοντέλο εκφράζει την i παρατήρηση της j στάθμης μέσω της σχέσης:

$$Y_{ij} = \mu + \alpha_j + e_{ij}, \quad 1 \leq i \leq n_j, 1 \leq j \leq k \quad (3.36)$$

όπου μ είναι ο «γενικός μέσος» ανεξαρτήτως της στάθμης, α_j είναι αυτό που ονομάζεται κύρια επίδραση της στάθμης j του παράγοντα, η οποία δηλώνει την απόκλιση που ενδέχεται να έχει η μέση τιμή της τ.μ. Y από το γενικό μέσο για την j στάθμη και e_{ij} είναι το σφάλμα της i παρατήρησης της εξαρτημένης μεταβλητής Y του j επιπέδου του παράγοντα.

3.3.1.2. Υποθέσεις του μοντέλου

Οι υποθέσεις που διέπουν το εν λόγω μοντέλο είναι:

- Οι παρατηρήσεις Y_{ij} κάθε στάθμης αποτελούν k ανεξάρτητα δείγματα από ανεξάρτητους πληθυσμούς.

- Καθένας από τους k πληθυσμούς ακολουθεί την κανονική κατανομή με μέση τιμή $\mu + \alpha_j$ και κοινή διασπορά σ^2 για $j = 1, \dots, k$, δηλαδή $Y_j \sim N(\mu + \alpha_j, \sigma^2)$.
- Οι επιδράσεις από τις στάθμες είναι σταθεροί αριθμοί που ικανοποιούν τη σχέση:

$$\sum_{j=1}^k \alpha_j = 0. \quad (3.37)$$

- Τα e_{ij} είναι ανεξάρτητες τυχαίες μεταβλητές από την κανονική κατανομή με μέση τιμή 0 και διασπορά σ^2 , δηλαδή $e_{ij} \sim N(0, \sigma^2)$.

3.3.1.3. Έλεγχος υποθέσεων

Το ζητούμενο σε τέτοια προβλήματα είναι να εξεταστεί η πιθανή διαφοροποίηση ανάμεσα στις στάθμες του παράγοντα, όπως έχει ήδη αναφερθεί. Ισοδύναμα, πρέπει να πραγματοποιηθεί ο έλεγχος με:

$$H_0 : \alpha_1 = \dots = \alpha_k \quad (3.38)$$

H_1 : τουλάχιστον μία από τις ισότητες δεν ισχύει σε επίπεδο σημαντικότητας α .

3.3.1.4. Υπολογισμός των αθροισμάτων των τετραγώνων

Οι αποκλίσεις των παρατηρήσεων Y_{ij} από τον ολικό μέσο $\bar{Y}_{..}$ μπορούν να χωριστούν σε δύο συνιστώσες, τη μεταβλητότητα εντός των ομάδων και τη μεταβλητότητα μεταξύ των ομάδων:

$$(Y_{ij} - \bar{Y}_{..}) = (Y_{ij} - \bar{Y}_{.j}) + (\bar{Y}_{.j} - \bar{Y}_{..}). \quad (3.39)$$

Αν η μεταβλητότητα μεταξύ των ομάδων είναι μεγάλη και ταυτόχρονα η μεταβλητότητα εντός των ομάδων είναι μικρή τότε η μηδενική υπόθεση απορρίπτεται καθώς οι στάθμες φαίνεται να διαφοροποιούνται ως προς τον εν λόγω παράγοντα. Αντίθετα, αν η μεταβλητότητα μεταξύ των ομάδων είναι μικρή και ταυτόχρονα η μεταβλητότητα εντός των ομάδων είναι μεγάλη τότε η μηδενική υπόθεση δεν απορρίπτεται καθώς οι στάθμες φαίνεται να μη διαφοροποιούνται στατιστικά σημαντικά ως προς τον εν λόγω παράγοντα.

Η συνολική διασπορά γύρω από τον γενικό μέσο δίνεται από το συνολικό άθροισμα των τετραγώνων:

$$SST = \sum_{j=1}^k \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}_{..})^2. \quad (3.40)$$

Προσθέτοντας και αφαιρώντας την τιμή $\bar{Y}_{.j}$ στην προηγούμενη έκφραση προκύπτει:

$$SST = SSA + SSE \quad (3.41)$$

$$\sum_{j=1}^k \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}_{..})^2 = \sum_{j=1}^k n_j (\bar{Y}_{.j} - \bar{Y}_{..})^2 + \sum_{j=1}^k \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}_{.j})^2 \quad (3.42)$$

αφού ο όρος $\sum_{j=1}^k \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}_{.j})(\bar{Y}_{.j} - \bar{Y}_{..}) = 0$ γιατί $\sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}_{.j}) = 0$.

Ο όρος SSA περιέχει το μέρος της διασποράς που οφείλεται στις διάφορες στάθμες του παράγοντα, γι' αυτό αναφέρεται ως άθροισμα τετραγώνων «μεταξύ των ομάδων». Ο όρος SSE περιέχει το υπόλοιπο μέρος της διασποράς, το οποίο οφείλεται σε τυχαία σφάλματα ή αποκλίσεις εντός των στάθμεων και γι' αυτό ονομάζεται άθροισμα τετραγώνων «εντός των ομάδων».

Τέλος είναι δυνατόν να υπολογισθεί ένας εκτιμητής της διασποράς σ^2 για τον πληθυσμό:

$$\hat{\sigma}^2 = \frac{\sum_{j=1}^k \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}_{.j})^2}{N - k} \quad (3.43)$$

όπου $\bar{Y}_{.j}$ είναι η αναμενόμενη τιμή της εξαρτημένης μεταβλητής Y της i παρατήρησης της j στάθμης του παράγοντα βάσει του εν λόγω μοντέλου ανάλυσης διασποράς.

Η θετική τετραγωνική ρίζα της προηγούμενης εκτιμήτριας ονομάζεται τυπικό σφάλμα του μοντέλου και όσο μικρότερη τιμή έχει τόσο καλύτερη προσαρμογή γίνεται από το μοντέλο διασποράς.

3.3.1.5. Τεστ F

Όπως προκύπτει από τα αθροίσματα των τετραγώνων, όσο μεγαλύτερο είναι το SSA ως προς το SSE τόσο πιο αξιοσημείωτες είναι οι διαφορές μεταξύ των ομάδων. Άρα ένα μέτρο σπουδαιότητας του παράγοντα είναι ο λόγος SSA/SSE . Όπως είχε αναφερθεί και στο κεφάλαιο της γραμμικής παλινδρόμησης, υπό την υπόθεση της κανονικής κατανομής τα αθροίσματα τετραγώνων SSA και SSE διαιρεμένα με τη διασπορά σ^2 είναι ανεξάρτητες τυχαίες μεταβλητές της κατανομής X^2 με βαθμούς ελευθερίας $k - 1$ και $N - k$ αντίστοιχα. Έτσι, συνάγεται το συμπέρασμα ότι για την ελεγχουσυνάρτηση F σε επίπεδο σημαντικότητας α ισχύει ότι :

$$F = \frac{SSA/[\sigma^2(k-1)]}{SSE/[\sigma^2(N-k)]} = \frac{MSA}{MSE} \sim F_{k-1, N-k} \quad (3.44)$$

υπό την H_0 .

Με βάση αυτή γίνεται ο έλεγχος για το αν υπάρχει διαφορά μεταξύ των επιπέδων (στάθμεων) του παράγοντα. Η μηδενική υπόθεση απορρίπτεται, δηλαδή ισοδύναμα

ο μέσος της τ.μ. Y διαφοροποιείται στις ομάδες του παράγοντα X , σε επίπεδο σημαντικότητας α όταν $F > F_{k-1, N-k}$. Σε αυτήν την περίπτωση η p -τιμή είναι μικρότερη του α .

Όλες οι παραπάνω ποσότητες συνοψίζονται στον Πίνακα 3.2.

Πίνακας 3.2: Πίνακας ανάλυση διασποράς με έναν παράγοντα

Πηγή μεταβλητότητας	Άθροισμα τετραγώνων	Βαθμοί ελευθερίας	Μέσο άθροισμα τετραγώνων	Έλεγχος F
Παράγοντας (μεταξύ των ομάδων)	$SSA = \sum_{j=1}^k n_j (\bar{Y}_j - \bar{Y}_{..})^2$	$k - 1$	$MSA = \frac{SSA}{k - 1}$	$\frac{MSA}{MSE}$
Υπόλοιπα (εντός των ομάδων)	$SSE = \sum_{j=1}^k \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}_j)^2$	$N - k$	$MSE = \frac{SSE}{N - k}$	
Ολική	$SST = \sum_{j=1}^k \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}_{..})^2$	$N - 1$		

3.3.1.6. Χρήση ψευδομεταβλητών

Η Ανάλυση Διασποράς με έναν παράγοντα συνδέεται με το γενικό γραμμικό μοντέλο παλινδρόμησης, στο οποίο οι διάφορες στάθμες εκφράζονται μέσω των εικονικών μεταβλητών (ή ψευδομεταβλητών). Στην Ανάλυση Διασποράς ένας εναλλακτικός τρόπος μοντελοποίησης είναι να χρησιμοποιηθεί ένα σύνολο ψευδομεταβλητών, τέτοιες ώστε να εκτιμηθούν τα $\hat{\mu}, \hat{\alpha}_1, \dots, \hat{\alpha}_{k-1}$ κατευθείαν ως συντελεστές της συνάρτησης παλινδρόμησης. Η εκτίμηση για την τελευταία κύρια επίδραση προκύπτει από τη σχέση $\sum_{j=1}^k \alpha_j = 0$. Οι ψευδομεταβλητές x_m , $m = 1, \dots, k - 1$ ορίζονται με τον ακόλουθο τρόπο:

$$x_m = \begin{cases} 1, & \text{αν η παρατήρηση είναι από τη στάθμη } m \text{ του παράγοντα } (m \neq k) \\ -1, & \text{αν η παρατήρηση είναι από τη στάθμη } k \text{ του παράγοντα} \\ 0, & \text{διαφορετικά} \end{cases}$$

Με την παραπάνω κωδικοποίηση λαμβάνεται η εξής συνάρτηση παλινδρόμησης:

$$Y = b_0 + b_1 x_1 + \dots + b_{k-1} x_{k-1} + \varepsilon \quad (3.45)$$

και ισχύει ότι:

$$b_0 = \mu, \quad b_1 = \alpha_1, \dots, b_{k-1} = \alpha_{k-1} \quad (3.46)$$

3.3.2. Ανάλυση Διασποράς κατά δύο Παράγοντες

Συχνά είναι σημαντικό να εξετασθεί η επίδραση δύο διαφορετικών παραγόντων και σε αυτήν την περίπτωση εφαρμόζεται η μέθοδος της Ανάλυσης Διασποράς κατά δύο παράγοντες (two – way ANOVA), στην οποία υπάρχουν δύο ανεξάρτητες

μεταβλητές. Είναι πιθανό ανάμεσα στους δύο παράγοντες A και B να υπάρχει αλληλεπίδραση. Ο όρος αλληλεπίδραση χρησιμοποιείται όταν η ταυτόχρονη δράση δύο συγκεκριμένων στάθμεων των παραγόντων A και B δεν ισούται με το άθροισμα των ξεχωριστών παραγόντων των δύο αυτών στάθμεων. Η υπόθεση ότι η δράση του ενός παράγοντα δρα προσθετικά στη δράση του δεύτερου παράγοντα, δηλαδή ότι δεν υπάρχει αλληλεπίδραση, δεν πρέπει να υιοθετείται ελεύθερα αλλά να ελέγχεται. Ακολουθείται η ίδια διαδικασία με αυτήν που περιγράφηκε στην Ανάλυση Διασποράς κατά έναν Παράγοντα.

3.3.2.1. Προσδιορισμός του μοντέλου

Στη γενική περίπτωση υποτίθεται ότι υπάρχει μια συνεχής εξαρτημένη μεταβλητή Y και δύο επεξηγηματικές μεταβλητές, παράγοντες, A και B με c και r στάθμες αντίστοιχα με n_{ij} παρατηρήσεις ανά αγωγή (i, j) . Σκοπός είναι να γίνει έλεγχος αν υπάρχει διαφορά στις τιμές της εξαρτημένης μεταβλητής ως προς τις διαφορετικές στάθμες κάθε παράγοντα ξεχωριστά (κύριες επιδράσεις των δύο παραγόντων) αλλά και αν συγκεκριμένοι συνδυασμοί δύο στάθμεων των δύο παραγόντων δημιουργούν διαφοροποιημένα αποτελέσματα, δηλαδή υπάρχει αλληλεπίδραση, που σημαίνει ότι η κύρια επίδραση ενός παράγοντα δεν είναι η ίδια σε κάθε στάθμη του άλλου παράγοντα. Συμβολίζεται με Y_{ijl} την l παρατήρηση της i στάθμης του A παράγοντα και της j στάθμης του B παράγοντα.

Το μοντέλο Ανάλυσης Διασποράς, που ενσωματώνει και τον δεύτερο παράγοντα, εκφράζει την l παρατήρηση της i στάθμης του παράγοντα A και της j στάθμης του παράγοντα B μέσω της σχέσης:

$$Y_{ijl} = \mu + \alpha_i + \beta_j + \delta_{ij} + \varepsilon_{ijl}, \quad 1 \leq i \leq c, 1 \leq j \leq r, 1 \leq l \leq n_{ij} \quad (3.47)$$

όπου μ είναι ο «γενικός μέσος» ανεξαρτήτως της στάθμης, α_i είναι η κύρια επίδραση της στάθμης i του παράγοντα A , β_j είναι η κύρια επίδραση της j στάθμης του παράγοντα B , η δ_{ij} είναι η αλληλεπίδραση που οφείλεται στον συνδυασμό των δύο παραγόντων A και B στην αγωγή (i, j) και ε_{ijl} είναι το σφάλμα της l παρατήρησης της εξαρτημένης μεταβλητής Y του i επιπέδου του παράγοντα A και του j επιπέδου του παράγοντα B .

3.3.2.2. Υποθέσεις του μοντέλου

Οι υποθέσεις που διέπουν το εν λόγω μοντέλο είναι:

- Η κατανομή που ακολουθεί η εξαρτημένη τυχαία μεταβλητή Y σε κάθε αγωγή είναι κανονική και η διασπορά της είναι η ίδια σε κάθε αγωγή.
- Οι επιδράσεις από τις στάθμες είναι σταθεροί αριθμοί που ικανοποιούν τη σχέση:

$$\sum_{i=1}^c \alpha_i = 0, \quad \sum_{j=1}^r \beta_j = 0, \quad \sum_{i=1}^c \delta_{ij} = 0, \quad \sum_{j=1}^r \delta_{ij} = 0. \quad (3.48)$$

- Τα ε_{ijl} είναι ανεξάρτητες τυχαίες μεταβλητές από την κανονική κατανομή με μέση τιμή 0 και διασπορά σ^2 , δηλαδή $\varepsilon_{ijl} \sim N(0, \sigma^2)$.

3.3.2.3. Έλεγχος υποθέσεων

Το ζητούμενο σε τέτοια προβλήματα είναι να εξεταστεί η πιθανή διαφοροποίηση ανάμεσα στις στάθμες των δύο διαφορετικών παραγόντων όπως έχει ήδη αναφερθεί. Ισοδύναμα, πρέπει να πραγματοποιηθούν οι εξής έλεγχοι:

1. Για την κύρια επίδραση του παράγοντα A :

$$H_0 : \alpha_1 = \dots = \alpha_c \quad (3.49)$$

H_1 : τουλάχιστον μία από τις ισότητες δεν ισχύει σε επίπεδο σημαντικότητας α .

2. Για την κύρια επίδραση του παράγοντα B :

$$H_0 : \beta_1 = \dots = \beta_r \quad (3.50)$$

H_1 : τουλάχιστον μία από τις ισότητες δεν ισχύει σε επίπεδο σημαντικότητας α .

3. Για την αλληλεπίδραση:

$$H_0 : \delta_{1j} = \delta_{2j} = \dots = \delta_{cr} \quad (3.51)$$

H_1 : τουλάχιστον μία από τις ισότητες δεν ισχύει σε επίπεδο σημαντικότητας α .

3.3.2.4. Υπολογισμός των αθροισμάτων των τετραγώνων

Οι εν λόγω έλεγχοι μπορούν να γίνουν και πάλι με μια διάσπαση του συνολικού αθροίσματος των τετραγώνων SST γύρω από τον μέσο, αλλά συμπεριλαμβάνοντας αυτή τη φορά τους δύο παράγοντες και την αλληλεπίδρασή τους. Στην περίπτωση που ο αριθμός των παρατηρήσεων n_{ijl} σε κάθε αγωγή (i, j) είναι ίσος με n τότε το συνολικό άθροισμα των τετραγώνων SST αναλύεται ως εξής:

$$\begin{aligned} SST &= \sum_{i=1}^c \sum_{j=1}^r \sum_{l=1}^n (Y_{ijl} - \bar{Y}_{...})^2 \\ &= \sum_{i=1}^c \sum_{j=1}^r \sum_{l=1}^n (\bar{Y}_{i..} - \bar{Y}_{...})^2 + \sum_{i=1}^c \sum_{j=1}^r \sum_{l=1}^n (\bar{Y}_{.j.} - \bar{Y}_{...})^2 \\ &\quad + \sum_{i=1}^c \sum_{j=1}^r \sum_{l=1}^n (\bar{Y}_{ij.} - \bar{Y}_{i..} - \bar{Y}_{.j.} + \bar{Y}_{...})^2 + \sum_{i=1}^c \sum_{j=1}^r \sum_{l=1}^n (Y_{ijl} - \bar{Y}_{ij.})^2 \\ &= SSA + SSB + SSAB + SSE \quad (3.52) \end{aligned}$$

καθώς οι υπόλοιποι όροι της διάσπασης μηδενίζονται.

Ο όρος SSA είναι το άθροισμα τετραγώνων λόγω του παράγοντα A , ο όρος SSB είναι το άθροισμα τετραγώνων λόγω του παράγοντα B , ο όρος $SSAB$ είναι το

άθροισμα τετραγώνων λόγω της αλληλεπίδρασης μεταξύ των παραγόντων και ο όρος SSE είναι το άθροισμα τετραγώνων των υπολοίπων εντός των στάθμεων.

Τέλος είναι δυνατόν να υπολογισθεί ένας εκτιμητής της διασποράς σ^2 για τον πληθυσμό:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^c \sum_{j=1}^r \sum_{l=1}^n (Y_{ijl} - \bar{Y}_{ijl})^2}{N - cr} \quad (3.53)$$

όπου $\bar{Y}_{ijl}$ είναι η αναμενόμενη τιμή της εξαρτημένης μεταβλητής Y της l παρατήρησης της i στάθμης του παράγοντα A και της j στάθμης του παράγοντα B βάσει του εν λόγω μοντέλου ανάλυσης διασποράς.

Η θετική τετραγωνική ρίζα της προηγούμενης εκτιμήτριας ονομάζεται τυπικό σφάλμα του μοντέλου και όσο μικρότερη τιμή έχει τόσο καλύτερη προσαρμογή γίνεται από το μοντέλο διασποράς.

Ο πίνακας της ανάλυσης διασποράς με δύο παράγοντες με αλληλεπίδραση συνοπτικά παρουσιάζεται στον Πίνακα 3.3.

Πίνακας 3.3: Πίνακας ανάλυσης διασποράς με δύο παράγοντες με αλληλεπίδραση

Πηγή μεταβλητότητας	Άθροισμα τετραγώνων	Βαθμοί ελευθερίας	Μέσο άθροισμα τετραγώνων	Έλεγχος F
Παράγοντας A	$SSA = \frac{1}{nr} \sum_{i=1}^c Y_{i..}^2 - \frac{Y_{...}^2}{n}$	$c - 1$	$MSA = \frac{SSA}{c - 1}$	$F = \frac{MSA}{MSE}$
Παράγοντας B	$SSB = \frac{1}{nc} \sum_{j=1}^r Y_{.j.}^2 - \frac{Y_{...}^2}{n}$	$r - 1$	$MSB = \frac{SSB}{r - 1}$	$F = \frac{MSB}{MSE}$
Αλληλεπίδραση	$SSAB = \frac{1}{n} \sum_{i=1}^c \sum_{j=1}^r Y_{ij.}^2 - SSA - SSB - \frac{Y_{...}^2}{n}$	$(c - 1)(r - 1)$	$MSAB = \frac{SSAB}{(c - 1)(r - 1)}$	$F = \frac{MSAB}{MSE}$
Υπόλοιπα	$SSE = \sum_{i=1}^c \sum_{j=1}^r \sum_{l=1}^n Y_{ijl}^2 - \sum_{i=1}^c \sum_{j=1}^r Y_{ij.}^2$	$N - rc$	$MSE = \frac{SSE}{N - rc}$	
Ολική	$SST = \sum_{i=1}^c \sum_{j=1}^r \sum_{l=1}^n Y_{ijl}^2 - \frac{Y_{...}^2}{n}$	$N - 1$		

3.4. Πολυμεταβλητή Ανάλυση Διασποράς (MANOVA)

Η Πολυμεταβλητή Ανάλυση Διασποράς (Multivariate Analysis of Variance – MANOVA) είναι μια γενίκευση της Ανάλυσης Διασποράς που επιτρέπει στον ερευνητή να αναλύει παραπάνω από μία εξαρτημένες μεταβλητές. Η Πολυμεταβλητή Ανάλυση Διασποράς δημιουργεί μια νέα εξαρτημένη μεταβλητή, βάσει του γραμμικού συνδυασμού όλων των εξαρτημένων μεταβλητών του μοντέλου, η οποία μεγιστοποιεί κατά το δυνατόν τις διαφορές των μέσων όρων μεταξύ των ομάδων – επιπέδων της ανεξάρτητης μεταβλητής. Έτσι, η

Πολυμεταβλητή Ανάλυση Διασποράς πραγματοποιεί ελέγχους των διαφορών των μέσων όρων σε δύο επίπεδα:

- το πολυμεταβλητό επίπεδο (multivariate tests), όπου εξετάζεται η επίδραση της ανεξάρτητης μεταβλητής πάνω στον γραμμικό συνδυασμό όλων των εξαρτημένων μεταβλητών συγχρόνως.
- το μονομεταβλητό επίπεδο (univariate tests or between – subjects effects), όπου εξετάζεται η επίδραση της ανεξάρτητης μεταβλητής πάνω σε καθεμία εξαρτημένη μεταβλητή ξεχωριστά.

3.4.1. Σύγκριση MANOVA και ANOVA

Τα πλεονεκτήματα που έχει η Πολυμεταβλητή Ανάλυση Διασποράς είναι ότι εξετάζοντας όλες τις μεταβλητές μαζί, πραγματοποιείται ένας πιο ισχυρός έλεγχος από το να εξετάζονται η καθεμία ξεχωριστά. Επιπλέον, η ερμηνεία των αποτελεσμάτων μπορεί να είναι ενισχυμένη υπολογίζοντας τα κριτήρια των μεταβλητών ταυτόχρονα. Ένας τρόπος για την κατανόηση της Πολυμεταβλητής Ανάλυσης Διασποράς είναι το γεγονός ότι δημιουργεί ένα νέο κριτήριο για κάθε αντικείμενο μελέτης το οποίο είναι αυτός ο γραμμικός συνδυασμός των p μεταβλητών που μέγιστα χωρίζονται οι ομάδες.

Ωστόσο, υπάρχουν περιπτώσεις όπου η Πολυμεταβλητή Ανάλυση Διασποράς δεν είναι απαραίτητη. Πιο συγκεκριμένα, εάν οι εξαρτημένες μεταβλητές είναι ασυσχέτιστες και το δειγματικό μέγεθος είναι μικρό τότε η εν λόγω μέθοδος είναι σε μειονεκτική θέση σε σχέση με τη μονομεταβλητή ανάλυση. Επιπρόσθετα, τα αποτελέσματα από μία Πολυμεταβλητή Ανάλυση Διασποράς μπορεί να είναι σύνθετα και δύσκολα ως προς την ερμηνεία τους αν και αυτή η πολυπλοκότητα, τελικά, μπορεί να αντανakλά την πραγματικότητα. Βέβαια, ισχύει και το αντίστροφο, ότι η Πολυμεταβλητή Ανάλυση Διασποράς μπορεί να απλουστεύσει τα δεδομένα ώστε να γίνουν πιο κατανοητά. Γι' αυτό το λόγο, ο ερευνητής είναι εκείνος που πρέπει να λάβει υπ' όψιν του τα πλεονεκτήματα και τα μειονεκτήματα της κάθε μεθόδου πριν την εφαρμογή τους.

3.4.2. Έλεγχοι υποθέσεων MANOVA

Ο γενικός έλεγχος της Πολυμεταβλητής Ανάλυσης Διασποράς είναι ανάλογος του γενικού ελέγχου F στη Μονομεταβλητή Ανάλυση Διασποράς. Βέβαια, στην πολυμεταβλητή περίπτωση υπάρχουν αρκετοί στατιστικοί έλεγχοι, οι οποίοι μπορούν να χρησιμοποιηθούν για να εκτιμηθεί η γενική μηδενική υπόθεση. Επειδή οι στατιστικοί αυτοί έλεγχοι βασίζονται σε διαφορετικά μαθηματικά κριτήρια, τα αποτελέσματα μιας μελέτης μπορεί να διαφέρουν λόγω του στατιστικού ελέγχου που έχει επιλεγεί.

Θα εξεταστεί η απλούστερη περίπτωση πολυμεταβλητής ανάλυσης με έναν παράγοντα (one – way MANOVA). Έστω X_{ij} , $j = 1, \dots, k$ και $i = 1, \dots, n_j$ όπου j συμβολίζει την ομάδα και i την παρατήρηση μέσα στην ομάδα. Το μέγεθος του δείγματος δεν είναι το ίδιο αναγκαστικά για όλες τις ομάδες. Για τον πληθυσμό κάθε ομάδας πρέπει να ισχύει ότι $N_p \sim (\mu_j, \Sigma)$ όπου p είναι η διάσταση (αφού

πρόκειται για μια πολυμεταβλητή κανονική κατανομή p διαστάσεων) και μ και Σ είναι το διάνυσμα των μέσων και ο πίνακας διακύμανσης – συνδιακύμανσης αντίστοιχα. Επίσης, ο πίνακας διακύμανσης – συνδιακύμανσης πρέπει να είναι ο ίδιος για όλες τις ομάδες. Άρα, οι δύο βασικές υποθέσεις, οι οποίες απαιτούνται για την Πολυμεταβλητή Ανάλυση Διασποράς, είναι η κανονικότητα των δεδομένων και ο ίδιος πίνακας διακύμανσης – συνδιακύμανσης για όλες τις ομάδες.

Ωστόσο, υπάρχουν κι άλλες προϋποθέσεις για την πολυμεταβλητή περίπτωση. Οι συγκρινόμενες ομάδες πρέπει να είναι ισοπληθείς, καθώς εξαιρετικά μεγάλες διαφορές στα μεγέθη των ομάδων μειώνουν τη διαφοροποιητική δύναμη της εν λόγω ανάλυσης. Όσον αφορά στο μέγεθος του δείγματος, είναι απαραίτητο ο αριθμός των παρατηρήσεων να είναι μεγαλύτερος συγκριτικά με τον αριθμό των εξαρτημένων μεταβλητών για κάθε συνδυαστική ομάδα. (Κάποιοι δέχονται ως ικανοποιητικό αριθμό παρατηρήσεων τις 20.) Οι ανεξάρτητες μεταβλητές πρέπει να είναι κατηγορικές ενώ οι εξαρτημένες να είναι συνεχείς και η κλίμακα μέτρησής τους, τουλάχιστον, ισοδιαστημική (interval). Επίσης, όλες οι συσχετίσεις μεταξύ των εξαρτημένων μεταβλητών πρέπει να είναι ευθύγραμμες. Και τέλος, η κατανομή των υπολοίπων, δηλαδή των διαφορών ανάμεσα στις αναμενόμενες και τις πραγματικές τιμές, πρέπει να είναι τυχαία.

Η πολυμεταβλητή μηδενική υπόθεση για την ισότητα των μέσων είναι:

$$H_0: \begin{matrix} \mu_{11} = & \mu_{12} = & \dots = & \mu_{1k} \\ \mu_{21} = & \mu_{22} = & \dots = & \mu_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ \mu_{p1} = & \mu_{p2} = & \dots = & \mu_{pk} \end{matrix}$$

Με άλλα λόγια, η μηδενική υπόθεση ελέγχει κατά πόσο για κάθε μεταβλητή όλες οι k ομάδες έχουν τον ίδιο πληθυσμιακό μέσο. Επιπλέον, μπορεί να γραφεί σε μορφή πινάκων. Ειδικότερα, έστω ότι μ_j είναι η στήλη διάνυσμα των πληθυσμιακών μέσων της j – ομάδας:

$$\mu_j = \begin{bmatrix} \mu_{1j} \\ \mu_{2j} \\ \vdots \\ \mu_{pj} \end{bmatrix}.$$

Συνεπώς, η μηδενική υπόθεση γίνεται ισοδύναμα με χρήση διανυσμάτων:

$$H_0: \mu_1 = \mu_2 = \dots = \mu_k$$

Η εναλλακτική υπόθεση H_1 είναι ότι για μία τουλάχιστον μεταβλητή υπάρχει μία τουλάχιστον ομάδα με πληθυσμιακό μέσο διαφορετικό από τις υπόλοιπες.

Η πολυμεταβλητή μηδενική υπόθεση για την ισότητα των διασπορών είναι:

$$H_0: \Sigma_1 = \Sigma_2 = \dots = \Sigma_k = \Sigma$$

Η εναλλακτική υπόθεση H_1 είναι ότι τουλάχιστον δύο πίνακες διακύμανσης διαφέρουν μεταξύ τους.

3.4.3. Άλλα πολυμεταβλητά κριτήρια

Υπάρχουν ποικίλα κριτήρια τα οποία χρησιμοποιούνται για τη μελέτη των κύριων επιδράσεων και των αλληλεπιδράσεων των ανεξάρτητων μεταβλητών στο πολυμεταβλητό επίπεδο. Για τον έλεγχο της στατιστικής σημαντικότητάς τους χρησιμοποιείται η F – κατανομή. Τα πιο διαδεδομένα είναι τα εξής:

- Hotelling's T^2 : χρησιμοποιείται όταν η ανεξάρτητη μεταβλητή έχει μόνο δύο τιμές – επίπεδα, οπότε και αποτελεί την πολυμεταβλητή προέκταση του απλού t – test.
- Wilks : είναι ο πιο διαδεδομένος δείκτης, δηλαδή αυτός που εφαρμόζεται στην πλειοψηφία των ερευνών και χρησιμοποιείται όταν οι συγκρινόμενες ομάδες που δημιουργούνται βάσει των τιμών – επιπέδων της ανεξάρτητης μεταβλητής είναι περισσότερες από δύο.
- Pillai's Trace (ή Pillai – Bartlett's Trace) : θεωρείται ο πλέον σταθερός πολυμεταβλητός δείκτης σε περίπτωση που οι συγκρινόμενες ομάδες είναι ανισοπληθείς και για το λόγο αυτό προτείνεται μερικές φορές.
- Roy's Largest Root : είναι ο λιγότερο σταθερός από τα υπόλοιπα πολυμεταβλητά κριτήρια όταν παραβιάζεται η προϋπόθεση της πολυμεταβλητής κανονικότητας.

ΚΕΦΑΛΑΙΟ 4

ΕΦΑΡΜΟΓΗ ΤΩΝ ΣΤΑΤΙΣΤΙΚΩΝ ΜΟΝΤΕΛΩΝ ΠΑΛΙΝΔΡΟΜΗΣΗΣ ΣΤΑ ΔΕΔΟΜΕΝΑ ΤΟΥ ΔΕΙΓΜΑΤΟΣ

4.1. Εφαρμογή της Ανάλυση Διασποράς (ANOVA)

4.1.1. Επίδραση της περιοχής στη συνολική καταναλωτική δαπάνη των νοικοκυριών

ΔΙΑΤΥΠΩΣΗ ΥΠΟΘΕΣΗΣ

Θα εφαρμοσθεί η Ανάλυση Διασποράς με έναν παράγοντα (one – way ANOVA) προκειμένου να ελεγχθεί αν διαφοροποιείται η μέση τιμή της συνολικής καταναλωτικής δαπάνης των νοικοκυριών στις 4 περιοχές της Ελλάδας. Η εξαρτημένη μεταβλητή ή μεταβλητή απόκρισης είναι η συνεχής τυχαία μεταβλητή συνολική καταναλωτική δαπάνη (total) και ο παράγοντας είναι η κατηγορική τυχαία μεταβλητή περιοχή (region), η οποία διαφοροποιείται σε 4 στάθμες: Βόρεια Ελλάδα, Κεντρική Ελλάδα, Αττική και Νησιά Αιγαίου – Κρήτη.

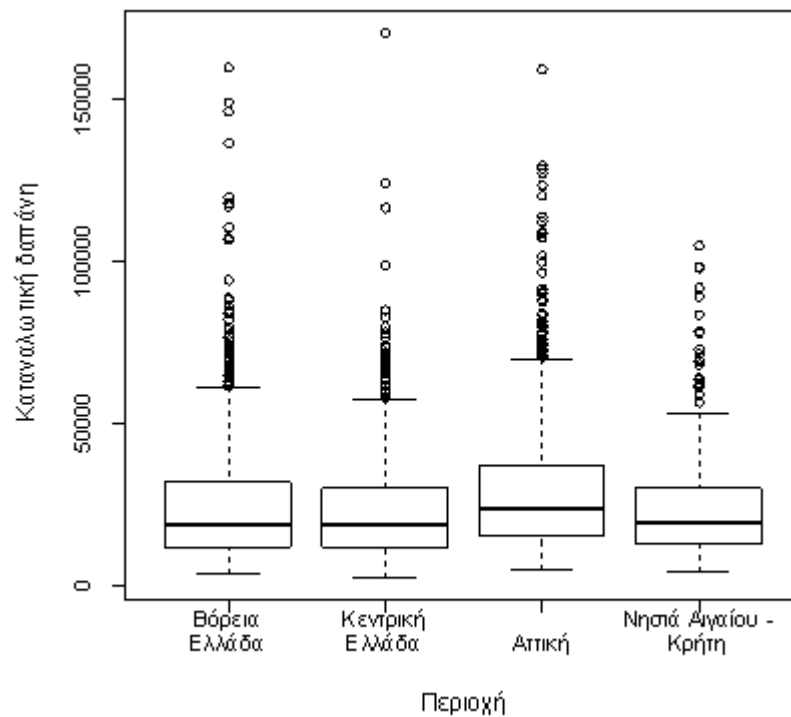
ΑΠΟΤΕΛΕΣΜΑΤΑ – ΕΡΜΗΝΕΙΑ

Αρχικά, παραθέτονται στον Πίνακα 4.1 μέτρα θέσης και μεταβλητότητας της συνολικής καταναλωτικής δαπάνης ως προς τις περιοχές της Ελλάδας, τα οποία υπολογίζονται στην R με τις εντολές `summary()`, `var()` και `sd()`. Επίσης, δίνεται και το αντίστοιχο θηκογράφημα, το οποίο κατασκευάζεται με την εντολή `boxplot()`.

Πίνακας 4.1: Αριθμητικά μέτρα για τη συνολική καταναλωτική δαπάνη (σε €) ανά περιοχή

		Βόρεια Ελλάδα	Κεντρική Ελλάδα	Αττική	Νησιά Αιγαίου - Κρήτη
Μέτρα Θέσης	Δειγματική Διάμεσος	18790	18990	23310	19200
	Δειγματικός Μέσος	24610	23610	28800	24360
Μέτρα Μετα- βλητό- τητας	Μικρότερη Παρατήρηση	3596	2503	5062	3918
	Μεγαλύτερη Παρατήρηση	159400	170400	159100	104600
	Εύρος Δείγματος	155804	167897	154038	100682
	1ο τεταρτημόριο	11850	11800	15250	12810
	3ο τεταρτημόριο	31640	30070	37100	30030
	Ενδοτεταρτημοριακό Εύρος	19790	18270	21850	17220
	Δειγματική Διασπορά	358737955	293735558	372567238	306732767
	Δειγματική Τυπική Απόκλιση	18940.38	17138.72	19302	17513.79

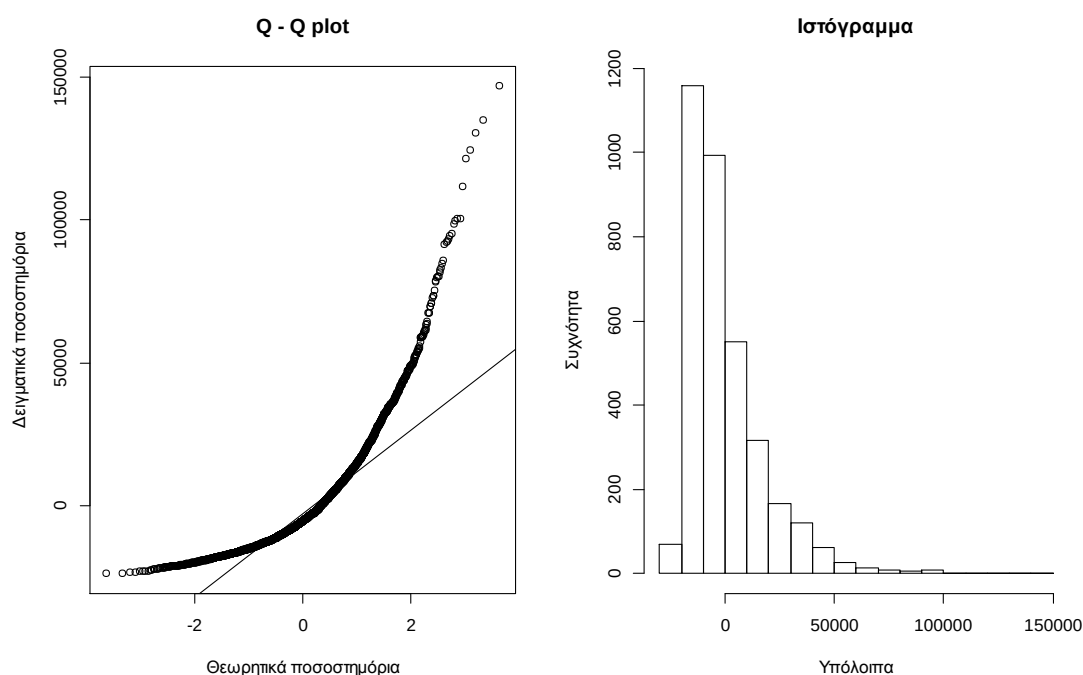
Διάγραμμα 4.1: Θηκογράφημα της συνολικής καταναλωτικής δαπάνης ανά περιοχή



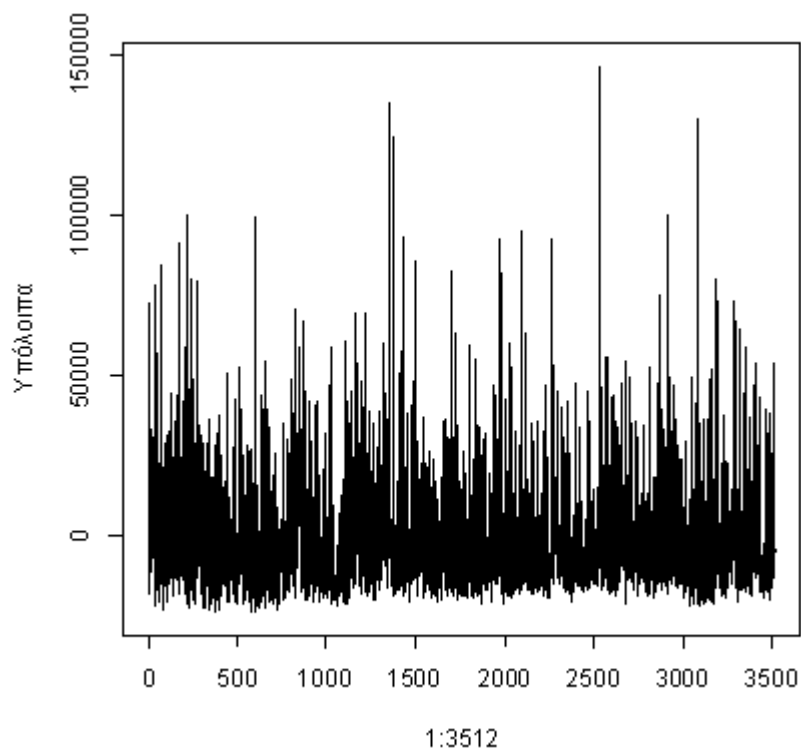
Συμπερασματικά, φαίνεται από το Διάγραμμα 4.1 ότι η συνολική καταναλωτική δαπάνη διαφοροποιείται στην Αττική σε σχέση με τις υπόλοιπες περιοχές της Ελλάδας. Για το λόγο αυτό, θα γίνει έλεγχος για το αν η παρατηρούμενη αυτή διαφορά είναι στατιστικά σημαντική.

Πριν γίνει η ανάλυση διασποράς πρέπει να πραγματοποιηθεί ο έλεγχος αν πληρούνται οι προϋποθέσεις του μοντέλου, δηλαδή τα σφάλματα να ακολουθούν την κανονική κατανομή (κανονικότητα), η διασπορά της εξαρτημένης τυχαίας μεταβλητής δοθείσης της επεξηγηματικής τυχαίας μεταβλητής να είναι ίδια σε όλες τις ομάδες (ομοσκεδαστικότητα) και τα σφάλματα να είναι ανεξάρτητες τυχαίες μεταβλητές (ανεξαρτησία).

Διάγραμμα 4.2: Διάγραμμα Q – Q και Ιστόγραμμα συχνοτήτων των υπολοίπων για τον έλεγχο κανονικότητας των υπολοίπων για το μοντέλο της ανάλυσης διασποράς κατά έναν παράγοντα



Διάγραμμα 4.3: Διάγραμμα για τον έλεγχο ανεξαρτησίας των υπολοίπων για το μοντέλο της ανάλυσης διασποράς κατά έναν παράγοντα



Όσον αφορά στον έλεγχο της κανονικότητας των υπολοίπων κατασκευάζεται Διάγραμμα Q – Q και το ιστόγραμμα συχνοτήτων με τις εντολές `qqnorm()` – `qqline()` και `hist()` αντίστοιχα. Σχετικά με την υπόθεση της ομοσκεδαστικότητας εκτελείται ο

έλεγχος του Levene (Levene's test) μέσω της εντολής `levene.test()` που βρίσκεται στη βιβλιοθήκη `lawstat` καθώς και ο έλεγχος Fligner – Killeen (Fligner – Killeen test), επειδή είναι πιο ανθεκτικός έλεγχος σε περιπτώσεις αποκλίσεων από την κανονικότητα, μέσω της εντολής `fligner.test()`. Και τέλος, για την ανεξαρτησία των υπολοίπων κατασκευάζεται ένα διάγραμμα υπολοίπων σε σχέση με τη σειρά των δεδομένων με την εντολή `plot()`.

Εκτελώντας τους ελέγχους αυτούς απορρίπτεται η υπόθεση τόσο για την κανονικότητα των σφαλμάτων όσο και για την υπόθεση της ομοσκεδαστικότητας. Αυτό φαίνεται από το Διάγραμμα 4.2 καθώς η υπόθεση της κανονικότητας δεν φαίνεται λογική επειδή παρατηρούνται αποκλίσεις. Από τον έλεγχο του Levene προκύπτει η p – τιμή = 0.00354 < 0.05, οπότε υπάρχουν σοβαρές ενδείξεις εναντίον της υπόθεσης της ομοσκεδαστικότητας. Το συμπέρασμα αυτό επιβεβαιώνεται και από τον έλεγχο του Fligner – Killeen, καθώς η p – τιμή = 7.673e-06 < 0.05. Επιπρόσθετα, και από τα προηγούμενα στοιχεία για τις ομάδες φαίνεται η έλλειψη συμμετρίας, η συγκέντρωση προς τα αριστερά των παρατηρήσεων του δείγματος και οι διαφορές στη διασπορά μεταξύ των ομάδων. Από το Διάγραμμα 4.3 μπορεί να εκτιμηθεί η ανεξαρτησία των σφαλμάτων καθώς τα δεδομένα φαίνεται να συμπεριφέρονται τυχαία αλλά και βάσει του τρόπου σχεδιασμού της εν λόγω έρευνας μπορεί να θεωρηθεί ότι ικανοποιείται η υπόθεση αυτή.

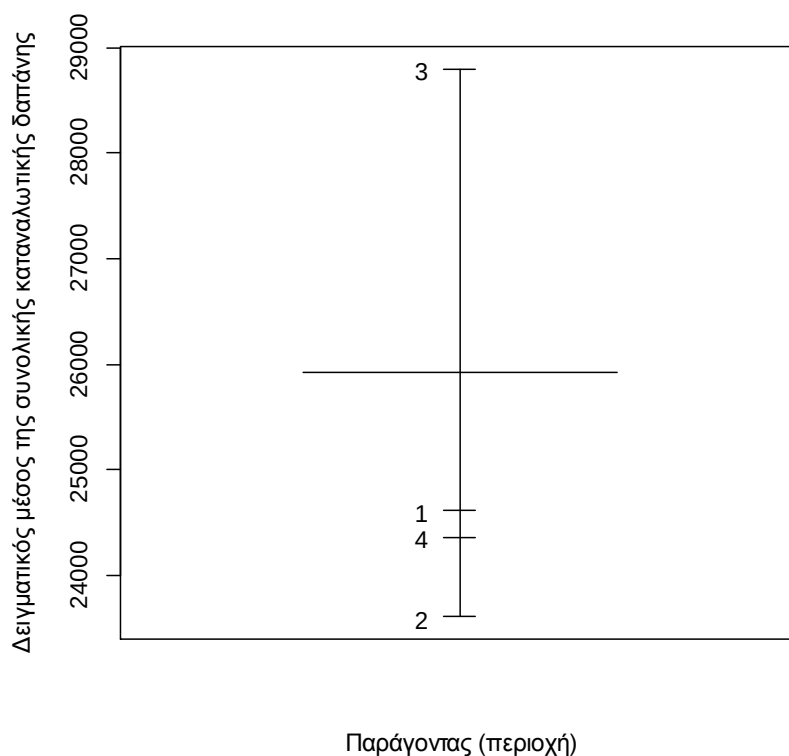
Εφ' όσον, λοιπόν, έχει απορριφθεί η μηδενική υπόθεση της ισότητας των διασπορών καθώς και η κανονικότητα των σφαλμάτων, θα εφαρμοσθεί στη συνέχεια ο μη παραμετρικός έλεγχος ανάλυσης διασποράς κατά έναν παράγοντα, ο οποίος καλείται Kruskal – Wallis test μέσω της εντολής `kruskal.test()`. Βάσει του ελέγχου αυτού, του οποίου η p -τιμή είναι < 2.2e-16, προκύπτει ότι υπάρχουν σοβαρές ενδείξεις για να απορριφθεί η μηδενική υπόθεση και άρα, παρατηρείται διαφοροποίηση στη συνολική καταναλωτική δαπάνη των νοικοκυριών ως προς τις 4 διαφορετικές περιοχές της Ελλάδας. Το ίδιο συμπέρασμα θα προέκυπτε ακόμα και αν είχε γίνει ο παραμετρικός έλεγχος μέσω της εντολής `aov()`, αφού η p -τιμή είναι και πάλι πάρα πολύ μικρή (p – τιμή = 5.38e-11).

Στη συνέχεια, υπολογίζονται οι εκτιμήσεις των επιδράσεων κάθε ομάδας, με τον περιορισμό ότι $\frac{\sum_{j=1}^k n_j a_j}{n} = 0$, μέσω της εντολής `model.tables()`. Με την προσθήκη του ορίσματος `type = 'means'` στην εν λόγω εντολή υπολογίζονται οι δειγματικοί μέσοι σε κάθε ομάδα του παράγοντα και ο αντίστοιχος αριθμός παρατηρήσεων σε κάθε ομάδα. Στον Πίνακα 4.2 και στο Διάγραμμα 4.4 παρουσιάζονται τα αποτελέσματα που προκύπτουν από την εντολή αυτή.

Πίνακας 4.2: Πίνακας των εκτιμήσεων των επιδράσεων της περιοχής στη συνολική καταναλωτική δαπάνη

Περιοχή	Δειγματικός Μέσος	Επίδραση	Αριθμός Παρατηρήσεων	Συνολικός Δειγματικός Μέσος
Βόρεια Ελλάδα	24611	- 1307	1105	25917.19
Κεντρική Ελλάδα	23609	- 2309	774	
Αττική	28801	2884	1300	
Νησιά Αιγαίου - Κρήτη	24359	- 1558	333	

Διάγραμμα 4.4: Γραφική αναπαράσταση των επιδράσεων των ομάδων του παράγοντα περιοχή στην συνολική καταναλωτική δαπάνη



Από τον Πίνακα 4.2 και το Διάγραμμα 4.4 προκύπτει ότι μόνο η επίδραση της ομάδας 3, δηλαδή της Αττικής, αυξάνει τη συνολική καταναλωτική δαπάνη ενώ η επίδραση των υπολοίπων τη μειώνουν.

Τέλος, για να εντοπιστούν οι εκτιμώμενες μέσες διαφορές μεταξύ όλων των ομάδων ανά δύο πραγματοποιείται στην R η εντολή `TukeyHSD()` και επιβεβαιώνεται η αρχική υπόθεση ότι η συνολική καταναλωτική δαπάνη των νοικοκυριών στην Αττική είναι διαφοροποιημένη σε σχέση με τις υπόλοιπες περιοχές της Ελλάδας και συγκεκριμένα, πιο μεγάλη. Πιο αναλυτικά, αυτό προκύπτει από τα 95% διαστήματα

εμπιστοσύνης ως προς τις εκτιμώμενες διαφορές μεταξύ των ομάδων. Τα εν λόγω διαστήματα για τις ομάδες 1, 2 και 4, δηλαδή Βόρεια Ελλάδα, Κεντρική Ελλάδα και Νησιά Αιγαίου – Κρήτη αντίστοιχα, περιέχουν τη μηδενική τιμή, γεγονός που σημαίνει ότι δε διαφοροποιούνται στατιστικά σημαντικά μεταξύ τους. Ενώ αντίθετα, τα διαστήματα αυτών με της Αττικής δεν περιέχουν τη μηδενική τιμή και άρα, διαφοροποιούνται.

Επομένως, στην Αττική σημειώνεται η μεγαλύτερη συνολική καταναλωτική δαπάνη των νοικοκυριών 28800€, η οποία είναι μεγαλύτερη κατά 4190.68€ από ότι η δαπάνη στη Βόρεια Ελλάδα (24609.32€), κατά 5192.75€ από ότι η δαπάνη στην Κεντρική Ελλάδα (23607.25€) και κατά 4442.21€ από ότι η δαπάνη στα Νησιά Αιγαίου – Κρήτη (24357.79€).

Εντολές στην R

```
> boxplot (total ~ region, names = c ('Βόρεια\ηΕλλάδα', 'Κεντρική\ηΕλλάδα', 'Αττική',
'Νησιά Αιγαίου -\ηΚρήτη'), xlab = 'Περιοχή', ylab = 'Καταναλωτική δαπάνη', cex.axis
= 0.9)
```

```
> by (total, region, summary)
```

```
region: 1
```

```
Min. 1st Qu. Median Mean 3rd Qu. Max.
3596 11850 18790 24610 31640 159400
```

```
region: 2
```

```
Min. 1st Qu. Median Mean 3rd Qu. Max.
2503 11800 18990 23610 30070 170400
```

```
region: 3
```

```
Min. 1st Qu. Median Mean 3rd Qu. Max.
5062 15250 23310 28800 37100 159100
```

```
region: 4
```

```
Min. 1st Qu. Median Mean 3rd Qu. Max.
3918 12810 19200 24360 30030 104600
```

```
> by (total, region, var)
```

```
region: 1
```

```
[1] 358737955
```

```
region: 2
```

```
[1] 293735558
```

```
region: 3
```

```
[1] 372567238
```

```
region: 4
```

```
[1] 306732767
```

```
> by (total, region, sd)
```

```
region: 1
```

```
[1] 18940.38
```



```

region: 2
[1] 17138.72
-----
region: 3
[1] 19302
-----
region: 4
[1] 17513.79
> R <- aov (total ~ region)
> par (mfrow = c (1, 2))
> qqnorm (residuals (R), xlab = 'Θεωρητικά ποσοστημόρια', ylab = 'Δειγματικά
ποσοστημόρια', main = 'Q - Q plot')
> qqline (residuals (R))
> hist (residuals (R), main = 'Ιστόγραμμα', xlab = 'Υπόλοιπα', ylab = 'Συχνότητα' )
> lawstat::levene.test (total, region, location = 'mean')
classical Levene's test based on the absolute deviations from the mean
( none not applied because the location is not set to median )
data: total
Test Statistic = 4.5341, p-value = 0.00354
> fligner.test (total ~ region, data = dataHH)
Fligner-Killeen test of homogeneity of variances
data: total by region
Fligner-Killeen:med chi-squared = 26.4511, df = 3, p-value = 7.673e-06
> plot (residuals (R), type = 'l', xlab = '1:3512', ylab = 'Υπόλοιπα')
> kruskal.test (total ~ region)
Kruskal-Wallis rank sum test
data: total by region
Kruskal-Wallis chi-squared = 80.9884, df = 3, p-value < 2.2e-16
> summary (aov (total ~ region))
      Df Sum Sq Mean Sq F value Pr(>F)
region   3 1.763e+10 5.878e+09  17.06 5.38e-11 ***
Residuals 3508 1.209e+12 3.446e+08
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> model.tables (aov (total ~ region), type = 'means')
Tables of means
Grand mean
25917.19
region
  1   2   3   4
24611 23609 28801 24359
rep 1105  774 1300  333
> model.tables (aov (total ~ region))
Tables of effects
region
  1   2   3   4
-1307 -2309 2884 -1558
rep 1105  774 1300  333
> plot.design (total ~ region, xlab = 'Παράγοντας (περιοχή)', ylab = 'Δειγματικός
μέσος της συνολικής καταναλωτικής δαπάνης', xaxt = NA)

```

```

> TukeyHSD (aov (total~region))
Tukey multiple comparisons of means
 95% family-wise confidence level
Fit: aov(formula = total ~ region)
$region
      diff      lwr      upr    p adj
2-1 -1002.0682 -3238.500 1234.363 0.6574336
3-1  4190.6799  2238.371 6142.989 0.0000002
4-1  -251.5272 -3234.299 2731.244 0.9964060
3-2  5192.7480  3026.511 7358.985 0.0000000
4-2   750.5409 -2376.439 3877.520 0.9267178
4-3 -4442.2071 -7372.717 -1511.698 0.0005763

```

4.1.2. Επίδραση της κύριας πηγής εισοδήματος στη συνολική καταναλωτική δαπάνη των νοικοκυριών

ΔΙΑΤΥΠΩΣΗ ΥΠΟΘΕΣΗΣ

Θα εφαρμοσθεί η Ανάλυση Διασποράς με έναν παράγοντα (one – way ANOVA) προκειμένου να ελεγχθεί αν διαφοροποιείται η μέση τιμή της συνολικής καταναλωτικής δαπάνης των νοικοκυριών ως προς την κύρια πηγή του εισοδήματός του. Η εξαρτημένη μεταβλητή ή μεταβλητή απόκρισης είναι η συνεχής τυχαία μεταβλητή συνολική καταναλωτική δαπάνη (total) και ο παράγοντας είναι η κατηγορική τυχαία μεταβλητή κύρια πηγή εισοδήματος (source), η οποία διαφοροποιείται σε 6 στάθμες: μισθοί ή ημερομίσθια, εισόδημα από αυτοαπασχόληση, εισόδημα από ακίνητη περιουσία, συντάξεις, επιδόματα ανεργίας και άλλα επιδόματα ή εισοδήματα από άλλες πηγές.

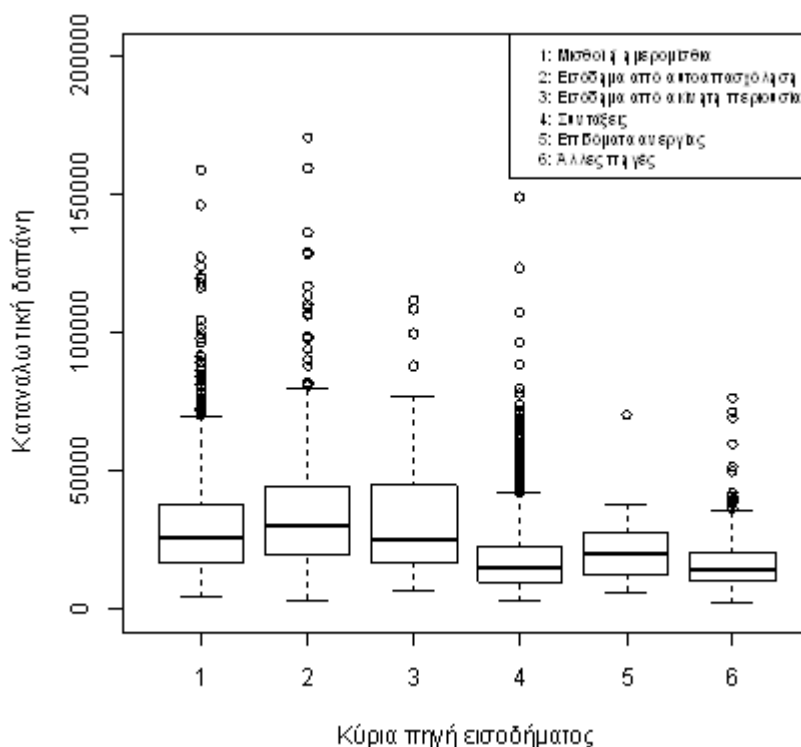
ΑΠΟΤΕΛΕΣΜΑΤΑ – ΕΡΜΗΝΕΙΑ

Αρχικά, παραθέτονται στον Πίνακα 4.3 μέτρα θέσης και μεταβλητότητας της συνολικής καταναλωτικής δαπάνης ως προς τις κύριες πηγές εισοδήματος, τα οποία υπολογίζονται στην R με τις εντολές `summary()`, `var()` και `sd()`. Επίσης, δίνεται και το αντίστοιχο θηκογράφημα, το οποίο κατασκευάζεται με την εντολή `boxplot()`.

Πίνακας 4.3: Αριθμητικά μέτρα για τη συνολική καταναλωτική δαπάνη (σε €) ανά κύρια πηγή εισοδήματος

		Μισθοί ή ημερομίσθια	Εισόδημα από αυτοαπα- σχόληση	Εισόδημα από ακίνητη περιουσία	Συντάξεις	Επιδόματα ανεργίας	Άλλα επιδόματα ή εισοδήματα από άλλες πηγές
Μέτρα Θέσης	Δειγματική Διάμεσος	26090	30180	25600	14940	19780	14270
	Δειγματικός Μέσος	30620	35840	34490	19140	22570	18240
Μέτρα Μετα- βλητό- τητας	Μικρότερη Παρατήρηση	4692	3182	6800	3596	6245	2503
	Μεγαλύτερη Παρατήρηση	159100	170400	112100	149000	70450	76330
	Εύρος Δείγματος	154408	167218	105300	145404	64205	73827
	1ο τεταρτημόριο	17410	20250	16620	10220	13790	10640
	3ο τεταρτημόριο	38450	44430	44800	23030	27080	20690
	Ενδοτεταρτημοριακό Εύρος	21040	24180	28180	12810	13290	10050
	Δειγματική Διασπορά	359490630	499365922	578148778	191619033	201582292	162571977
	Δειγματική Τυπική Απόκλιση	18960.24	22346.5	24044.72	13842.65	14197.97	12750.37

Διάγραμμα 4.5: Θηκογράφημα της συνολικής καταναλωτικής δαπάνης ανά κύρια πηγή εισοδήματος

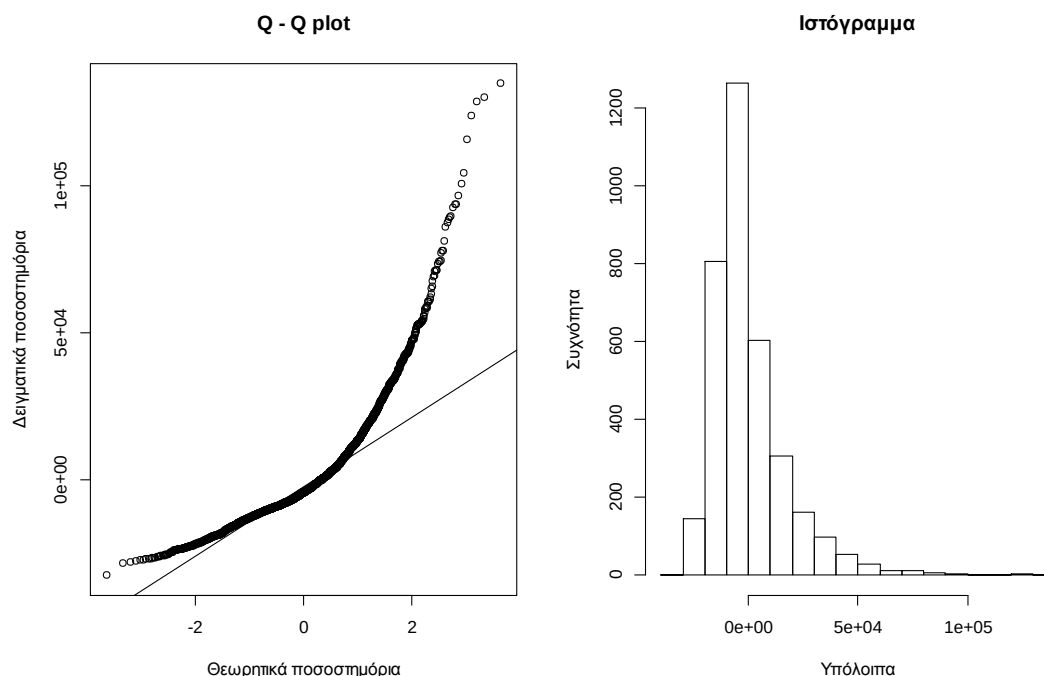


Συμπερασματικά, φαίνεται από το Διάγραμμα 4.5 ότι η συνολική καταναλωτική δαπάνη διαφοροποιείται στα νοικοκυριά με κύρια πηγή εισοδήματος τους μισθούς, τα εισοδήματα από ακίνητη περιουσία και από αυτοαπασχόληση σε σχέση με τα

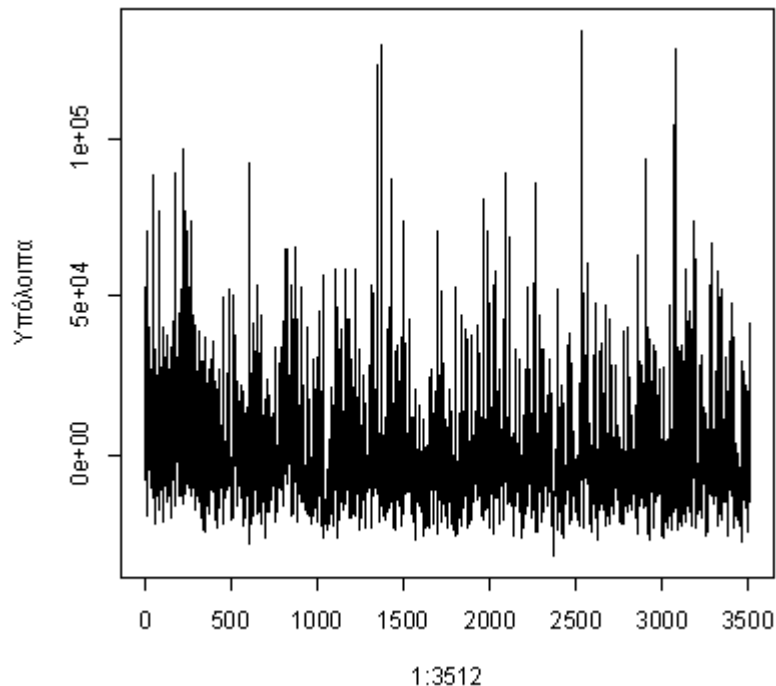
νοικοκυριά με κύρια πηγή εισοδήματος τις συντάξεις, τα επιδόματα ανεργίας και άλλα επιδόματα. Για το λόγο αυτό, θα γίνει έλεγχος για το αν η παρατηρούμενη αυτή διαφορά είναι στατιστικά σημαντική.

Πριν γίνει η ανάλυση διασποράς πρέπει να πραγματοποιηθεί ο έλεγχος αν πληρούνται οι προϋποθέσεις του μοντέλου, δηλαδή τα σφάλματα να ακολουθούν την κανονική κατανομή (κανονικότητα), η διασπορά της εξαρτημένης τυχαίας μεταβλητής δοθείσης της εξηγηματικής τυχαίας μεταβλητής να είναι ίδια σε όλες τις ομάδες (ομοσκεδαστικότητα) και τα σφάλματα να είναι ανεξάρτητες τυχαίες μεταβλητές (ανεξαρτησία).

Διάγραμμα 4.6: Διάγραμμα Q – Q και Ιστόγραμμα συχνοτήτων των υπολοίπων για τον έλεγχο κανονικότητας των υπολοίπων για το μοντέλο της ανάλυσης διασποράς κατά έναν παράγοντα



Διάγραμμα 4.7: Διάγραμμα για τον έλεγχο ανεξαρτησίας των υπολοίπων για το μοντέλο της ανάλυσης διασποράς κατά έναν παράγοντα



Όσον αφορά στον έλεγχο της κανονικότητας των υπολοίπων κατασκευάζεται Διάγραμμα Q – Q και το ιστόγραμμα συχνοτήτων με τις εντολές `qqnorm()` – `qqline()` και `hist()` αντίστοιχα. Σχετικά με την υπόθεση της ομοσκεδαστικότητας εκτελείται ο έλεγχος του Levene (Levene’s test) μέσω της εντολής `levene.test()` που βρίσκεται στη βιβλιοθήκη `lawstat` καθώς και ο έλεγχος Fligner – Killeen (Fligner – Killeen test), επειδή είναι πιο ανθεκτικός έλεγχος σε περιπτώσεις αποκλίσεων από την κανονικότητα, μέσω της εντολής `fligner.test()`. Και τέλος, για την ανεξαρτησία των υπολοίπων κατασκευάζεται ένα διάγραμμα υπολοίπων σε σχέση με τη σειρά των δεδομένων με την εντολή `plot()`.

Εκτελώντας τους ελέγχους αυτούς απορρίπτεται η υπόθεση τόσο για την κανονικότητα των σφαλμάτων όσο και για την υπόθεση της ομοσκεδαστικότητας. Αυτό φαίνεται από το Διάγραμμα 4.6 καθώς η υπόθεση της κανονικότητας δεν φαίνεται λογική επειδή παρατηρούνται αποκλίσεις. Από τον έλεγχο του Levene προκύπτει η p – τιμή $< 2.2e-16 < 0.05$, οπότε υπάρχουν σοβαρές ενδείξεις εναντίον της υπόθεσης της ομοσκεδαστικότητας. Το συμπέρασμα αυτό επιβεβαιώνεται και από τον έλεγχο του Fligner – Killeen, καθώς η p – τιμή είναι η ίδια. Επιπρόσθετα, και από τα προηγούμενα στοιχεία για τις ομάδες φαίνεται η έλλειψη συμμετρίας, η συγκέντρωση προς τα αριστερά των παρατηρήσεων του δείγματος και οι διαφορές στη διασπορά μεταξύ των ομάδων. Από το Διάγραμμα 4.7 μπορεί να εκτιμηθεί η ανεξαρτησία των σφαλμάτων καθώς τα δεδομένα φαίνεται να συμπεριφέρονται τυχαία αλλά και βάσει του τρόπου σχεδιασμού της εν λόγω έρευνας μπορεί να θεωρηθεί ότι ικανοποιείται η υπόθεση αυτή.

Εφ’ όσον, λοιπόν, έχει απορριφθεί η μηδενική υπόθεση της ισότητας των διασπορών καθώς και η κανονικότητα των σφαλμάτων, θα εφαρμοσθεί στη

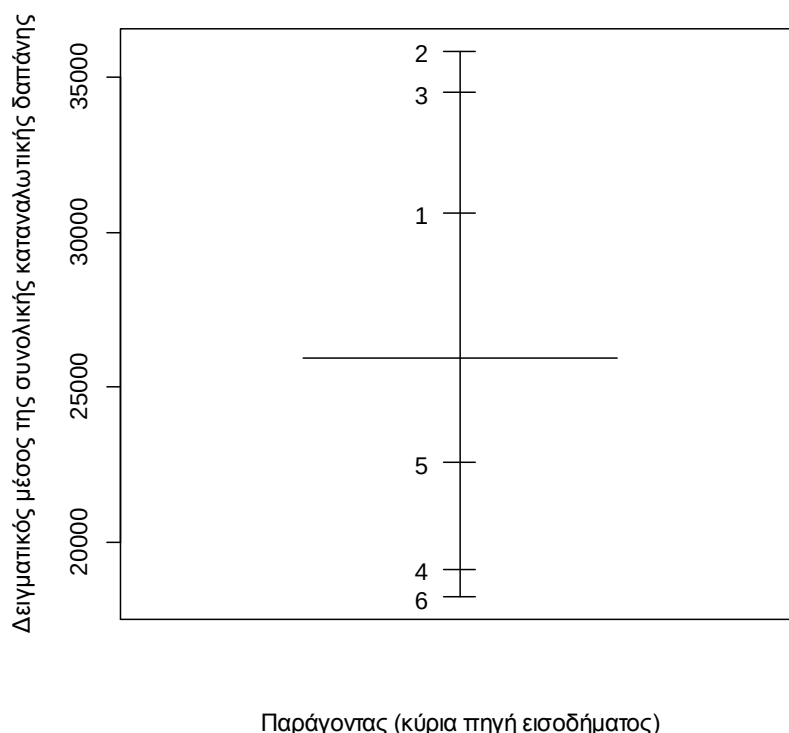
συνέχεια ο μη παραμετρικός έλεγχος ανάλυσης διασποράς κατά έναν παράγοντα, ο οποίος καλείται Kruskal – Wallis test μέσω της εντολής `kruskal.test()`. Βάσει του ελέγχου αυτού, του οποίου η p -τιμή είναι $< 2.2e-16 < 0.05$, προκύπτει ότι υπάρχουν σοβαρές ενδείξεις για να απορριφθεί η μηδενική υπόθεση και άρα, παρατηρείται διαφοροποίηση στη συνολική καταναλωτική δαπάνη των νοικοκυριών ως προς τις 6 κύριες πηγές εισοδήματος των νοικοκυριών. Το ίδιο συμπέρασμα θα προέκυπτε ακόμα και αν είχε γίνει ο παραμετρικός έλεγχος μέσω της εντολής `aov()`, αφού η p -τιμή είναι και πάλι πάρα πολύ μικρή (p -τιμή $< 2e-16$).

Στη συνέχεια, υπολογίζονται οι εκτιμήσεις των επιδράσεων κάθε ομάδας, με τον περιορισμό ότι $\frac{\sum_{j=1}^k n_j a_j}{n} = 0$, μέσω της εντολής `model.tables()`. Με την προσθήκη του ορίσματος `type = 'means'` στην εν λόγω εντολή υπολογίζονται οι δειγματικοί μέσοι σε κάθε ομάδα του παράγοντα και ο αντίστοιχος αριθμός παρατηρήσεων σε κάθε ομάδα. Στον Πίνακα 4.4 και στο Διάγραμμα 4.8 παρουσιάζονται τα αποτελέσματα που προκύπτουν από την εντολή αυτή.

Πίνακας 4.4: Πίνακας των εκτιμήσεων των επιδράσεων της κύριας πηγής εισοδήματος στη συνολική καταναλωτική δαπάνη

Κύρια Πηγή Εισοδήματος	Δειγματικός Μέσος	Επίδραση	Αριθμός Παρατηρήσεων	Συνολικός Δειγματικός Μέσος
Μισθοί ή ημερομίσθια	30617	4700	1201	25917.19
Εισόδημα από αυτοαπασχόληση	35836	9918	531	
Εισόδημα από ακίνητη περιουσία	34490	8573	79	
Συντάξεις	19138	- 6779	1541	
Επιδόματα ανεργίας	22568	- 3349	20	
Άλλα επιδόματα ή εισοδήματα από άλλες πηγές	18240	- 7677	140	

Διάγραμμα 4.8: Γραφική αναπαράσταση των επιδράσεων των ομάδων του παράγοντα κύρια πηγή εισοδήματος στη συνολική καταναλωτική δαπάνη



Από τον Πίνακα 4.4 και το Διάγραμμα 4.8 προκύπτει ότι μόνο η επίδραση των ομάδων 2, 3 και 1, δηλαδή οι ομάδες που έχουν ως κύρια πηγή εισοδήματος τα εισοδήματα από αυτοαπασχόληση, τα εισοδήματα από ακίνητη περιουσία και τους μισθούς ή ημερομίσθια, αυξάνει τη συνολική καταναλωτική δαπάνη ενώ η επίδραση των υπολοίπων τριών τη μειώνουν.

Τέλος, για να εντοπιστούν οι εκτιμώμενες μέσες διαφορές μεταξύ όλων των ομάδων ανά δύο πραγματοποιείται στην R η εντολή TukeyHSD() και επιβεβαιώνεται η αρχική υπόθεση ότι η συνολική καταναλωτική δαπάνη των νοικοκυριών με κύρια πηγή τους μισθούς, τα εισοδήματα από αυτοαπασχόληση και ακίνητη περιουσία είναι διαφοροποιημένη σε σχέση με τις υπόλοιπες κύριες πηγές και συγκεκριμένα, πιο μεγάλη. Πιο αναλυτικά, αυτό προκύπτει από τα 95% διαστήματα εμπιστοσύνης ως προς τις εκτιμώμενες διαφορές μεταξύ των ομάδων. Τα εν λόγω διαστήματα για τις στάθμες 1, 2 και 3, δηλαδή μισθούς, εισοδήματα από αυτοαπασχόληση και από ακίνητη περιουσία αντίστοιχα, περιέχουν τη μηδενική τιμή, γεγονός που σημαίνει ότι δεν διαφοροποιούνται στατιστικά σημαντικά μεταξύ τους. Ενώ αντίθετα, τα διαστήματα αυτών με τις στάθμες 4 και 6, δηλαδή συντάξεις και άλλα επιδόματα, δεν περιέχουν τη μηδενική τιμή και άρα, διαφοροποιούνται. Όσον αφορά στη στάθμη 5, δηλαδή τα επιδόματα ανεργίας, φαίνεται να μην υπάρχει στατιστικά σημαντική διαφοροποίηση και ως προς τους μισθούς και τα εισοδήματα από ακίνητη περιουσία αλλά και ως προς τις συντάξεις και άλλα επιδόματα, καθώς σε όλα αυτά τα διαστήματα περιέχεται η μηδενική τιμή. Θα μπορούσε, ωστόσο, να θεωρηθεί ότι προσεγγίζει καλύτερα τη δεύτερη ομάδα (συντάξεις και άλλα επιδόματα), αφού υπάρχουν πιο ισχυρές ενδείξεις ότι δε διαφοροποιείται ως προς

αυτές τις κατηγορίες και παρουσιάζει μεγαλύτερη απόκλιση από την πρώτη ομάδα (μισθοί, εισοδήματα από αυτοαπασχόληση και από ακίνητη περιουσία).

Επομένως, η συνολική καταναλωτική δαπάνη για τα νοικοκυριά με κύρια πηγή εισοδήματος τους μισθούς είναι 30620€, μικρότερη κατά 5218.32€ από ότι στα νοικοκυριά με κύρια πηγή τα εισοδήματα από αυτοαπασχόληση (35838.32€), μικρότερη κατά 3872.62€ από ότι στα νοικοκυριά με κύρια πηγή τα εισοδήματα από ακίνητη περιουσία (34492.62€), μεγαλύτερη κατά 11479.21€ από ότι στα νοικοκυριά με κύρια πηγή εισοδήματος τις συντάξεις (19140.78€), μεγαλύτερη κατά 8048.89€ από ότι στα νοικοκυριά με κύρια πηγή εισοδήματος τα επιδόματα ανεργίας (22571.11€) και μεγαλύτερη κατά 12377.35€ από ότι στα νοικοκυριά με κύρια πηγή εισοδήματος άλλα επιδόματα (18242.65€).

Εντολές στην R

```
> boxplot (total ~ source, names = c ('1', '2', '3', '4', '5', '6'), ylim = c (0, 200000), xlab = 'Κύρια πηγή εισοδήματος', ylab = 'Καταναλωτική δαπάνη')
> legend ('topright', c ('1: Μισθοί ή ημερομίσθια', '2: Εισόδημα από αυτοαπασχόληση', '3: Εισόδημα από ακίνητη περιουσία', '4: Συντάξεις', '5: Επιδόματα ανεργίας', '6: Άλλες πηγές'), cex = 0.7)
> by (total, source, summary)
source: 1
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 4692  17410  26090  30620  38450 159100
-----
source: 2
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 3182  20250  30180  35840  44430 170400
-----
source: 3
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 6800  16620  25600  34490  44800 112100
-----
source: 4
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 3596  10220  14940  19140  23030 149000
-----
source: 5
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 6245  13790  19780  22570  27080  70450
-----
source: 6
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 2503  10640  14270  18240  20690  76330
> by (total, source, var)
source: 1
[1] 359490630
-----
source: 2
```



```

[1] 499365922
-----
source: 3
[1] 578148778
-----
source: 4
[1] 191619033
-----
source: 5
[1] 201582292
-----
source: 6
[1] 162571977
> by (total, source, sd)
source: 1
[1] 18960.24
-----
source: 2
[1] 22346.5
-----
source: 3
[1] 24044.72
-----
source: 4
[1] 13842.65
-----
source: 5
[1] 14197.97
-----
source: 6
[1] 12750.37
> S <- aov (total ~ source)
> par (mfrow = c (1, 2))
> qqnorm (residuals (S), xlab = 'Θεωρητικά ποσοστημόρια', ylab = 'Δειγματικά
ποσοστημόρια', main = 'Q - Q plot')
> qqline (residuals (S))
> hist (residuals (S), main = 'Ιστόγραμμα', xlab = 'Υπόλοιπα', ylab = 'Συχνότητα')
> lawstat::levene.test (total, source, location = 'mean')
classical Levene's test based on the absolute deviations from the mean
( none not applied because the location is not set to median )
data: total
Test Statistic = 37.6548, p-value < 2.2e-16
> fligner.test (total ~ source, data = dataHH)
Fligner-Killeen test of homogeneity of variances
data: total by source
Fligner-Killeen:med chi-squared = 249.5518, df = 5, p-value < 2.2e-16
> plot (residuals (S), type = 'l', xlab = '1:3512', ylab = 'Υπόλοιπα')
> kruskal.test (total ~ source)
Kruskal-Wallis rank sum test
data: total by source

```

```

Kruskal-Wallis chi-squared = 674.4534, df = 5, p-value < 2.2e-16
> summary (aov (total ~ source))
      Df Sum Sq Mean Sq F value Pr(>F)
source    5 1.639e+11 3.277e+10  108.1 <2e-16 ***
Residuals 3506 1.063e+12 3.031e+08
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> model.tables (aov (total ~ source), type = 'means')
Tables of means
Grand mean
25917.19
source
  1  2  3  4  5  6
30617 35836 34490 19138 22568 18240
rep 1201  531  79 1541  20 140
> model.tables (aov (total ~ source))
Tables of effects
source
  1  2  3  4  5  6
4700 9918 8573 -6779 -3349 -7677
rep 1201  531  79 1541  20 140
> plot.design (total ~ source, xlab = 'Παράγοντας (κύρια πηγή εισοδήματος)', ylab =
'Dειγματικός μέσος της συνολικής καταναλωτικής δαπάνης', xaxt = NA)
> TukeyHSD (aov (total ~ source))
Tukey multiple comparisons of means
95% family-wise confidence level
Fit: aov(formula = total ~ source)
$source
      diff      lwr      upr    p adj
2-1 5218.3204 2631.365 7805.2760 0.0000001
3-1 3872.6197 -1893.103 9638.3420 0.3929129
4-1 -11479.2089 -13389.919 -9568.4993 0.0000000
5-1 -8048.8870 -19240.820  3143.0464 0.3139388
6-1 -12377.3504 -16810.503 -7944.1981 0.0000000
3-2 -1345.7008 -7331.717  4640.3156 0.9878920
4-2 -16697.5294 -19195.458 -14199.6004 0.0000000
5-2 -13267.2074 -24574.206 -1960.2091 0.0107392
6-2 -17595.6708 -22311.779 -12879.5625 0.0000000
4-3 -15351.8286 -21078.159 -9625.4978 0.0000000
5-3 -11921.5066 -24347.263  504.2499 0.0686656
6-3 -16249.9701 -23235.161 -9264.7791 0.0000000
5-4  3430.3220 -7741.369 14602.0131 0.9523897
6-4  -898.1415 -5279.939  3483.6561 0.9920788
6-5 -4328.4634 -16194.748  7537.8212 0.9044010

```

4.1.3. Επίδραση του αριθμού των οικονομικά ενεργών μελών στη συνολική καταναλωτική δαπάνη των νοικοκυριών

ΔΙΑΤΥΠΩΣΗ ΥΠΟΘΕΣΗΣ

Θα εφαρμοσθεί η Ανάλυση Διασποράς με έναν παράγοντα (one – way ANOVA) προκειμένου να ελεγχθεί αν διαφοροποιείται η μέση τιμή της συνολικής καταναλωτικής δαπάνης των νοικοκυριών ως προς τον αριθμό των οικονομικά ενεργών μελών του. Η εξαρτημένη μεταβλητή ή μεταβλητή απόκρισης είναι η συνεχής τυχαία μεταβλητή συνολική καταναλωτική δαπάνη (total) και ο παράγοντας είναι η κατηγορική τυχαία μεταβλητή αριθμός των οικονομικά ενεργών μελών (members), η οποία διαφοροποιείται σε 6 στάθμες: 0, 1, 2, 3, 4 και 5 ή περισσότερα ενεργά οικονομικά μέλη. Επιπρόσθετα, επειδή ο αριθμός των οικονομικά ενεργών μελών είναι μία μεταβλητή διάταξης (ordinal), θα ελεγχθεί μία πιθανή γραμμική τάση (linear trend) μεταξύ των μέσων της εξαρτημένης μεταβλητής total στις διατάξιμες στάθμες του παράγοντα members.

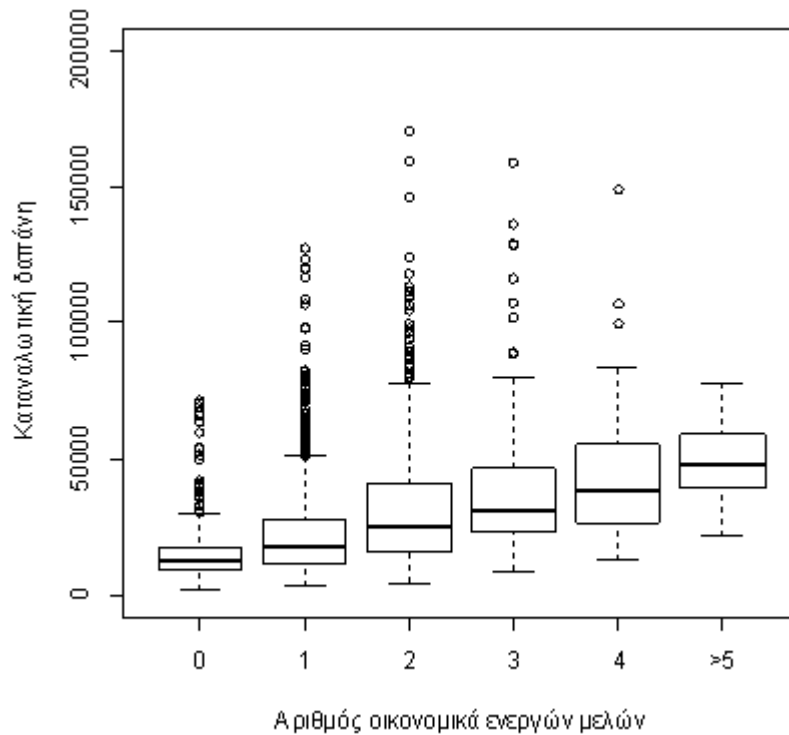
ΑΠΟΤΕΛΕΣΜΑΤΑ – ΕΡΜΗΝΕΙΑ

Αρχικά, παραθέτονται στον Πίνακα 4.5 μέτρα θέσης και μεταβλητότητας της συνολικής καταναλωτικής δαπάνης για τον αριθμό των οικονομικά ενεργών μελών του νοικοκυριού, τα οποία υπολογίζονται στην R με τις εντολές `summary()`, `var()` και `sd()`. Επίσης, δίνεται και το αντίστοιχο θηκογράφημα, το οποίο κατασκευάζεται με την εντολή `boxplot()`.

Πίνακας 4.5: Αριθμητικά μέτρα για τη συνολική καταναλωτική δαπάνη (σε €) ανά αριθμό οικονομικά ενεργών μελών του νοικοκυριού

		0	1	2	3	4	> 5
Μέτρα Θέσης	Δειγματική Διάμεσος	12720	17990	25760	31020	38730	48060
	Δειγματικός Μέσος	15790	21960	30950	37670	44080	49270
Μέτρα Μετα-βλητότητας	Μικρότερη Παρατήρηση	2503	3182	4238	8660	12830	21720
	Μεγαλύτερη Παρατήρηση	71370	127200	170400	159100	149000	77850
	Εύρος Δείγματος	68867	124018	166162	150440	136170	56130
	1ο τεταρτημόριο	9443	11860	16170	23330	26830	39220
	3ο τεταρτημόριο	17910	27490	40930	46350	55140	59510
	Ενδοτεταρτημοριακό Εύρος	8467	15630	24760	23020	28310	20290
	Δειγματική Διασπορά	118637947	231270066	405132559	542042960	604229936	445747065
	Δειγματική Τυπική Απόκλιση	10892.1	15207.57	20127.9	23281.82	24581.09	21112.72

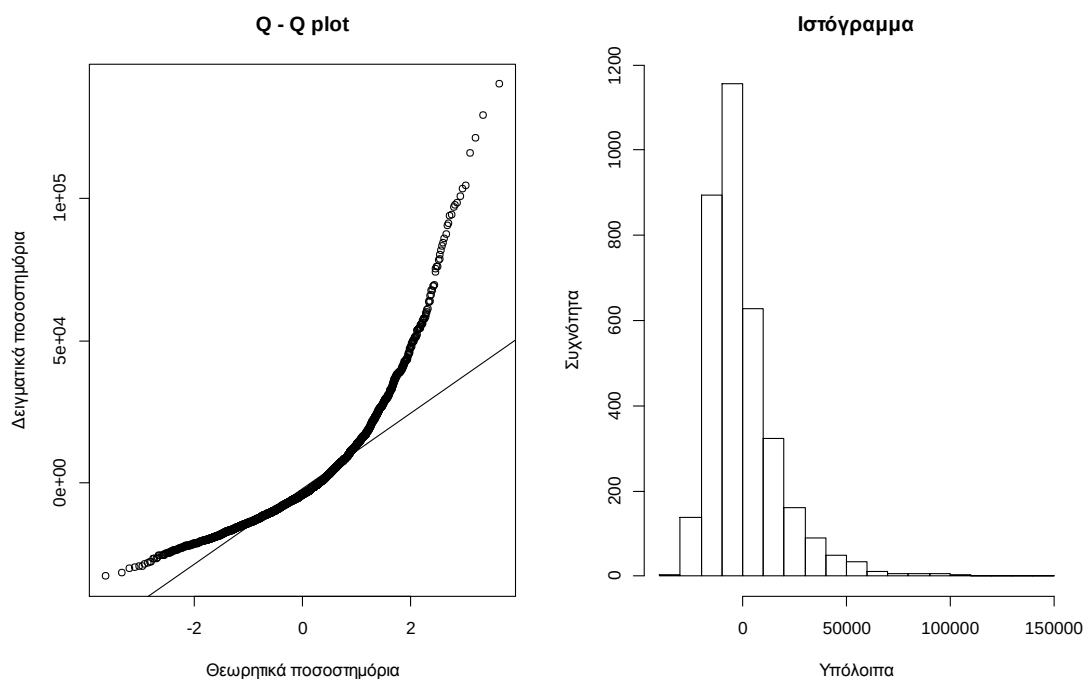
Διάγραμμα 4.9: Θηκογράφημα της συνολικής καταναλωτικής δαπάνης ανά αριθμό των οικονομικά ενεργών μελών



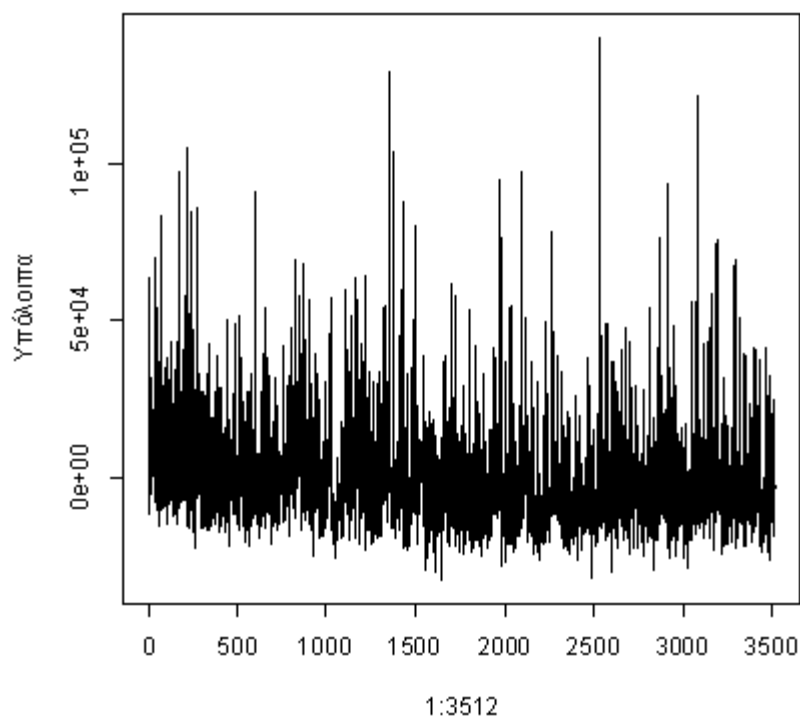
Συμπερασματικά, φαίνεται από το Διάγραμμα 4.9 ότι η συνολική καταναλωτική δαπάνη αυξάνεται όσο αυξάνεται και ο αριθμός των οικονομικά ενεργών μελών των νοικοκυριών. Επομένως, είναι εμφανές ότι υπάρχει γραμμική τάση μεταξύ των στάθμεων του παράγοντα ως προς τη συνολική καταναλωτική δαπάνη. Για το λόγο αυτό, θα γίνει έλεγχος για το αν η παρατηρούμενη αυτή αύξηση είναι στατιστικά σημαντική.

Πριν γίνει η ανάλυση διασποράς πρέπει να πραγματοποιηθεί ο έλεγχος αν πληρούνται οι προϋποθέσεις του μοντέλου, δηλαδή τα σφάλματα να ακολουθούν την κανονική κατανομή (κανονικότητα), η διασπορά της εξαρτημένης τυχαίας μεταβλητής δοθείσης της εξηγηματικής τυχαίας μεταβλητής να είναι ίδια σε όλες τις ομάδες (ομοσκεδαστικότητα) και τα σφάλματα να είναι ανεξάρτητες τυχαίες μεταβλητές (ανεξαρτησία).

Διάγραμμα 4.10: Διάγραμμα Q – Q και Ιστόγραμμα συχνοτήτων των υπολοίπων για τον έλεγχο κανονικότητας των υπολοίπων για το μοντέλο της ανάλυσης διασποράς κατά έναν παράγοντα



Διάγραμμα 4.11: Διάγραμμα για τον έλεγχο ανεξαρτησίας των υπολοίπων για το μοντέλο της ανάλυσης διασποράς κατά έναν παράγοντα



Όσον αφορά στον έλεγχο της κανονικότητας των υπολοίπων κατασκευάζεται Διάγραμμα Q – Q και το ιστόγραμμα συχνοτήτων με τις εντολές `qqnorm()` – `qqline()` και `hist()` αντίστοιχα. Σχετικά με την υπόθεση της ομοσκεδαστικότητας εκτελείται ο

έλεγχος του Levene (Levene's test) μέσω της εντολής `levene.test()` που βρίσκεται στη βιβλιοθήκη `lawstat` καθώς και ο έλεγχος Fligner – Killeen (Fligner – Killeen test), επειδή είναι πιο ανθεκτικός έλεγχος σε περιπτώσεις αποκλίσεων από την κανονικότητα, μέσω της εντολής `fligner.test()`. Και τέλος, για την ανεξαρτησία των υπολοίπων κατασκευάζεται ένα διάγραμμα υπολοίπων σε σχέση με τη σειρά των δεδομένων με την εντολή `plot()`.

Εκτελώντας τους ελέγχους αυτούς απορρίπτεται η υπόθεση τόσο για την κανονικότητα των σφαλμάτων όσο και για την υπόθεση της ομοσκεδαστικότητας. Αυτό φαίνεται από το Διάγραμμα 4.10 καθώς η υπόθεση της κανονικότητας δεν φαίνεται λογική επειδή παρατηρούνται αποκλίσεις. Από τον έλεγχο του Levene προκύπτει η p – τιμή $< 2.2\text{e-}16 < 0.05$, οπότε υπάρχουν σοβαρές ενδείξεις εναντίον της υπόθεσης της ομοσκεδαστικότητας. Το συμπέρασμα αυτό επιβεβαιώνεται και από τον έλεγχο του Fligner – Killeen, καθώς η p – τιμή είναι η ίδια. Επιπρόσθετα, και από τα προηγούμενα στοιχεία για τις ομάδες φαίνεται η έλλειψη συμμετρίας, η συγκέντρωση προς τα αριστερά των παρατηρήσεων του δείγματος και οι διαφορές στη διασπορά μεταξύ των ομάδων. Από το Διάγραμμα 4.11 μπορεί να εκτιμηθεί η ανεξαρτησία των σφαλμάτων καθώς τα δεδομένα φαίνεται να συμπεριφέρονται τυχαία αλλά και βάσει του τρόπου σχεδιασμού της εν λόγω έρευνας μπορεί να θεωρηθεί ότι ικανοποιείται η υπόθεση αυτή.

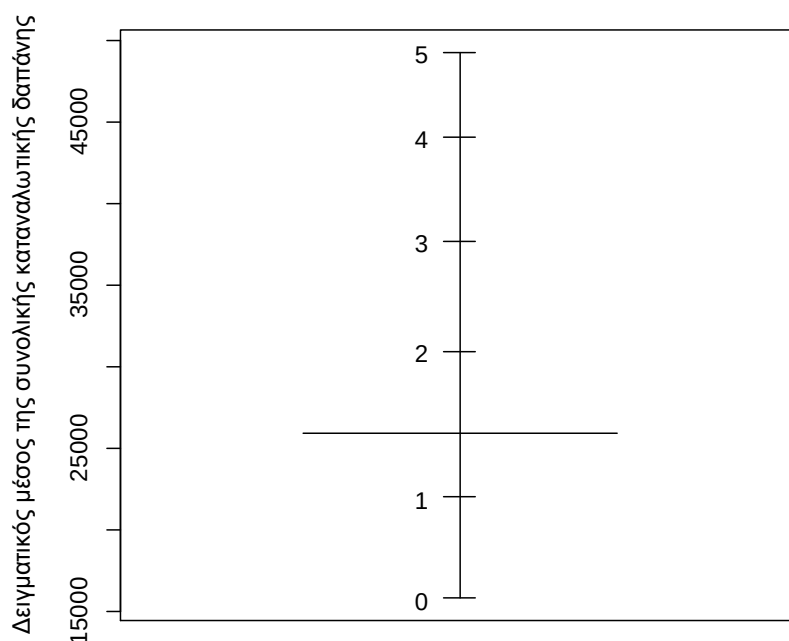
Εφ' όσον, λοιπόν, έχει απορριφθεί η μηδενική υπόθεση της ισότητας των διασπορών καθώς και η κανονικότητα των σφαλμάτων, θα εφαρμοσθεί στη συνέχεια ο μη παραμετρικός έλεγχος ανάλυσης διασποράς κατά έναν παράγοντα, ο οποίος καλείται Kruskal – Wallis test μέσω της εντολής `kruskal.test()`. Βάσει του ελέγχου αυτού, του οποίου η p -τιμή είναι $< 2.2\text{e-}16$, προκύπτει ότι υπάρχουν σοβαρές ενδείξεις για να απορριφθεί η μηδενική υπόθεση και άρα, παρατηρείται διαφοροποίηση στη συνολική καταναλωτική δαπάνη των νοικοκυριών ως προς τον αριθμό των οικονομικά ενεργών μελών των νοικοκυριών. Το ίδιο συμπέρασμα θα προέκυπτε ακόμα και αν είχε γίνει ο παραμετρικός έλεγχος μέσω της εντολής `aov()`, αφού η p -τιμή είναι και πάλι πάρα πολύ μικρή (p – τιμή $< 2\text{e-}16$).

Στη συνέχεια, υπολογίζονται οι εκτιμήσεις των επιδράσεων κάθε ομάδας, με τον περιορισμό ότι $\frac{\sum_{j=1}^k n_j a_j}{n} = 0$, μέσω της εντολής `model.tables()`. Με την προσθήκη του ορίσματος `type = 'means'` στην εν λόγω εντολή υπολογίζονται οι δειγματικοί μέσοι σε κάθε ομάδα του παράγοντα και ο αντίστοιχος αριθμός παρατηρήσεων σε κάθε ομάδα. Στον Πίνακα 4.6 και στο Διάγραμμα 4.12 παρουσιάζονται τα αποτελέσματα που προκύπτουν από την εντολή αυτή.

Πίνακας 4.6: Πίνακας των εκτιμήσεων των επιδράσεων του αριθμού των οικονομικά ενεργών μελών στη συνολική καταναλωτική δαπάνη

Αριθμός οικονομικά ενεργών μελών	Δειγματικός Μέσος	Επίδραση	Αριθμός Παρατηρήσεων	Συνολικός Δειγματικός Μέσος
0	15790	- 10127	391	25917.19
1	21964	- 3954	1562	
2	30945	5028	1287	
3	37665	11748	203	
4	44079	18162	64	
> 5	49273	23356	5	

Διάγραμμα 4.12: Γραφική αναπαράσταση των επιδράσεων των ομάδων του παράγοντα περιοχή στην συνολική καταναλωτική δαπάνη



Παράγοντας (αριθμός οικονομικά ενεργών μελών)

Από τον Πίνακα 4.6 και το Διάγραμμα 4.12 προκύπτει ότι μόνο η επίδραση των ομάδων 0 και 1, δηλαδή τα νοικοκυριά με κανένα ή ένα οικονομικά ενεργό μέλος, μειώνει τη συνολική καταναλωτική δαπάνη ενώ η επίδραση των υπολοίπων την αυξάνει.

Τέλος, για να εντοπιστούν οι εκτιμώμενες μέσες διαφορές μεταξύ όλων των ομάδων ανά δύο πραγματοποιείται στην R η εντολή `TukeyHSD()` και επιβεβαιώνεται η αρχική υπόθεση ότι η συνολική καταναλωτική δαπάνη των νοικοκυριών διαφοροποιείται όσο αυξάνεται ο αριθμός των οικονομικά ενεργών μελών του νοικοκυριού. Πιο αναλυτικά, αυτό προκύπτει από τα 95% διαστήματα εμπιστοσύνης ως προς τις εκτιμώμενες διαφορές μεταξύ των ομάδων. Τα εν λόγω διαστήματα για

τις στάθμες 0, 1 και 2, δηλαδή με κανένα, 1 και 2 οικονομικά ενεργά μέλη αντίστοιχα, δεν περιέχουν τη μηδενική τιμή, γεγονός που σημαίνει ότι διαφοροποιούνται στατιστικά σημαντικά τόσο μεταξύ τους όσο και με τις υπόλοιπες στάθμες. Ενώ αντίθετα, τα διαστήματα για τις στάθμες 3, 4 και 5, δηλαδή με 3, 4 και 5 ή περισσότερα οικονομικά ενεργά μέλη, περιέχουν τη μηδενική τιμή και άρα, δε διαφοροποιούνται μεταξύ τους. Όσον αφορά στη στάθμη 2, φαίνεται να μην υπάρχει στατιστικά σημαντική διαφοροποίηση και ως προς τη στάθμη 5, καθώς στο αντίστοιχο διαστήμα περιέχεται η μηδενική τιμή. Ωστόσο, επειδή παρουσιάζει σχετικά μεγάλη απόκλιση από τη μέση τιμή αλλά και από τις στάθμες 3 και 4, η στάθμη 2 θα θεωρηθεί διαφοροποιημένη από αυτή την ομάδα των τριών κατηγοριών.

Επομένως, η συνολική καταναλωτική δαπάνη για τα νοικοκυριά με κανένα οικονομικά ενεργό μέλος είναι 15790€, μικρότερη κατά 6173.61€ από ότι στα νοικοκυριά με ένα οικονομικά ενεργό μέλος (21963.61€), μικρότερη κατά 15155.56€ από ότι στα νοικοκυριά με δύο οικονομικά ενεργά μέλη (30945.56€), μικρότερη κατά 21875.29€ από ότι στα νοικοκυριά με τρία οικονομικά ενεργά μέλη (37665.29€), μικρότερη κατά 28289.43€ από ότι στα νοικοκυριά με τέσσερα οικονομικά ενεργά μέλη (44079.43€) και μικρότερη κατά 33483.19€ από ότι στα νοικοκυριά με πέντε ή περισσότερα οικονομικά ενεργά μέλη (49273.19€).

Εντολές στην R

```
> boxplot (total ~ members, names = c ('0', '1', '2', '3', '4', '>5'), ylim = c (0, 200000),
xlab = 'Αριθμός οικονομικά ενεργών μελών', ylab = 'Καταναλωτική δαπάνη')
> by (total, members, summary)
```

members: 0

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
2503	9443	12720	15790	17910	71370

members: 1

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
3182	11860	17990	21960	27490	127200

members: 2

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
4238	16170	25760	30950	40930	170400

members: 3

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
8660	23330	31020	37670	46350	159100

members: 4

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
12830	26830	38730	44080	55140	149000

members: 5

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
21720	39220	48060	49270	59510	77850


```

> by (total, members, var)
members: 0
[1] 118637947
-----
members: 1
[1] 231270066
-----
members: 2
[1] 405132559
-----
members: 3
[1] 542042960
-----
members: 4
[1] 604229936
-----
members: 5
[1] 445747065
> by (total, members, sd)
members: 0
[1] 10892.1
-----
members: 1
[1] 15207.57
-----
members: 2
[1] 20127.9
-----
members: 3
[1] 23281.82
-----
members: 4
[1] 24581.09
-----
members: 5
[1] 21112.72
> M <- aov (total ~ members)
> par (mfrow = c (1, 2))
> qqnorm (residuals (M), xlab = 'Θεωρητικά ποσοστημόρια', ylab = 'Δειγματικά
ποσοστημόρια', main = 'Q - Q plot')
> qqline (residuals (M))
> hist (residuals (M), main = 'Ιστόγραμμα', xlab = 'Υπόλοιπα', ylab = 'Συχνότητα' )
> lawstat::levene.test (total, members, location = 'mean')
classical Levene's test based on the absolute deviations from the mean
( none not applied because the location is not set to median )
data: total
Test Statistic = 43.8906, p-value < 2.2e-16
> fligner.test (total ~ members, data = dataHH)
Fligner-Killeen test of homogeneity of variances
data: total by members
Fligner-Killeen:med chi-squared = 299.6602, df = 5, p-value < 2.2e-16

```

```

> plot (residuals (M), type = 'l', xlab = '1:3512', ylab = 'Υπόλοιπα')
> kruskal.test (total ~ source)
    Kruskal-Wallis rank sum test

data: total by source
Kruskal-Wallis chi-squared = 674.4534, df = 5, p-value < 2.2e-16
> summary (aov (total ~ members))
      Df Sum Sq Mean Sq F value Pr(>F)
members    5 1.489e+11 2.978e+10  96.9 <2e-16 ***
Residuals 3506 1.078e+12 3.074e+08
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> model.tables (aov (total ~ members), type = 'means')
Tables of means
Grand mean
25917.19
members
  0  1  2  3  4  5
15790 21964 30945 37665 44079 49273
rep 391 1562 1287 203  64  5
> model.tables (aov (total ~ members))
Tables of effects
members
  0  1  2  3  4  5
-10127 -3954 5028 11748 18162 23356
rep 391 1562 1287 203  64  5
> plot.design (total ~ members, xlab = 'Παράγοντας (αριθμός οικονομικά ενεργών
μελών)', ylab = 'Δειγματικός μέσος της συνολικής καταναλωτικής δαπάνης', xaxt =
NA)
> TukeyHSD (aov (total~members))
Tukey multiple comparisons of means
 95% family-wise confidence level

Fit: aov(formula = total ~ members)
$members
      diff      lwr      upr    p adj
1-0 6173.608 3346.8370 9000.379 0.0000000
2-0 15155.556 12268.9555 18042.157 0.0000000
3-0 21875.285 17550.8913 26199.678 0.0000000
4-0 28289.432 21548.8765 35029.988 0.0000000
5-0 33483.191 10985.2601 55981.122 0.0003250
2-1 8981.948 7100.0991 10863.797 0.0000000
3-1 15701.677 11972.1633 19431.190 0.0000000
4-1 22115.824 15740.5609 28491.088 0.0000000
5-1 27309.583 4918.3843 49700.783 0.0067867
3-2 6719.729 2944.6655 10494.792 0.0000060
4-2 13133.876 6731.8597 19535.893 0.0000001
5-2 18327.635 -4071.1958 40726.466 0.1810545
4-3 6414.148 -752.0041 13580.299 0.1096726
5-3 11607.907 -11021.1797 34236.993 0.6882422
5-4 5193.759 -18018.5278 28406.046 0.9881493

```

4.1.4. Επίδραση της κοινωνικο – οικονομικής κατάστασης του προσώπου αναφοράς στη συνολική καταναλωτική δαπάνη των νοικοκυριών

ΔΙΑΤΥΠΩΣΗ ΥΠΟΘΕΣΗΣ

Θα εφαρμοσθεί η Ανάλυση Διασποράς με έναν παράγοντα (one – way ANOVA) προκειμένου να ελεγχθεί αν διαφοροποιείται η μέση τιμή της συνολικής καταναλωτικής δαπάνης των νοικοκυριών ως προς την κοινωνικο – οικονομική κατάσταση του προσώπου αναφοράς. Η εξαρτημένη μεταβλητή ή μεταβλητή απόκρισης είναι η συνεχής τυχαία μεταβλητή συνολική καταναλωτική δαπάνη (total) και ο παράγοντας είναι η κατηγορική τυχαία μεταβλητή κοινωνικο – οικονομική κατάσταση (situation), η οποία διαφοροποιείται σε 7 στάθμες: χειρώνακτας εργαζόμενος, μη χειρώνακτας εργαζόμενος, αυτοαπασχολούμενος – γεωργός – κτηνοτρόφος, άνεργος, συνταξιούχος, λοιποί μη οικονομικά ενεργοί και δεν έχει απαντηθεί.

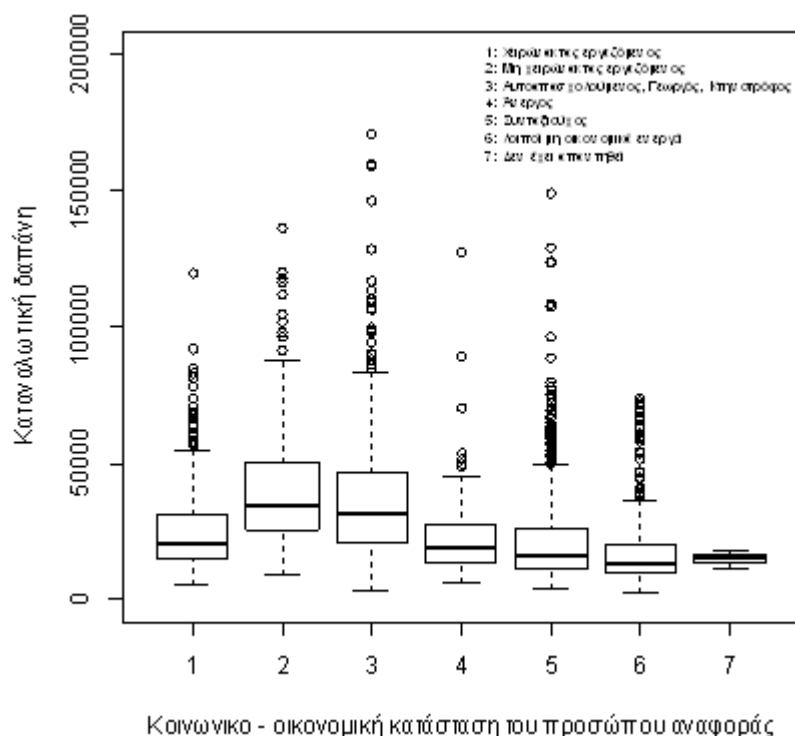
ΑΠΟΤΕΛΕΣΜΑΤΑ – ΕΡΜΗΝΕΙΑ

Αρχικά, παραθέτονται στον Πίνακα 4.7 μέτρα θέσης και μεταβλητότητας της συνολικής καταναλωτικής δαπάνης για την κοινωνικο – οικονομική κατάσταση του προσώπου αναφοράς του νοικοκυριού, τα οποία υπολογίζονται στην R με τις εντολές `summary()`, `var()` και `sd()`. Επίσης, δίνεται και το αντίστοιχο θηκογράφημα, το οποίο κατασκευάζεται με την εντολή `boxplot()`.

Πίνακας 4.7: Αριθμητικά μέτρα για τη συνολική καταναλωτική δαπάνη (σε €) ανά κοινωνικο – οικονομική κατάσταση του προσώπου αναφοράς του νοικοκυριού

		Χειρώνακτας εργαζόμενος	Μη χειρώνακτας εργαζόμενος	Αυτοαπασχολούμενος, γεωργός, κτηνοτρόφος	Άνεργος	Συνταξιούχος	Λοιποί μη οικονομικά ενεργοί	Δεν έχει απαντηθεί
Μέτρα θέσης	Δειγματική Διάμεσος	20770	34400	31770	18900	16350	13470	15570
	Δειγματικός Μέσος	24870	40090	37000	22920	21030	17470	14950
Μέτρα μεταβλη- τότητας	Μικρότερη παρατήρηση	5528	9032	3182	6245	3854	2503	11190
	Μεγαλύτερη παρατήρηση	119700	136400	170400	127200	149000	73800	18090
	Εύρος Δείγματος	114172	127368	167218	120955	145146	71297	6900
	1ο τεταρτημόριο	14920	25350	21200	13810	10890	9800	13380
	3ο τεταρτημόριο	30870	50540	46380	27610	26300	20640	16830
	Ενδοτεταρτημοριακό εύρος	15950	25190	25180	13800	15410	10840	3450
	Δειγματική Διασπορά	219595867	434920588	513653107	238383016	234810827	143766452	12201063
	Δειγματική Τυπική Απόκλιση	14818.77	20854.75	22663.92	15439.66	15323.54	11990.26	3493.002

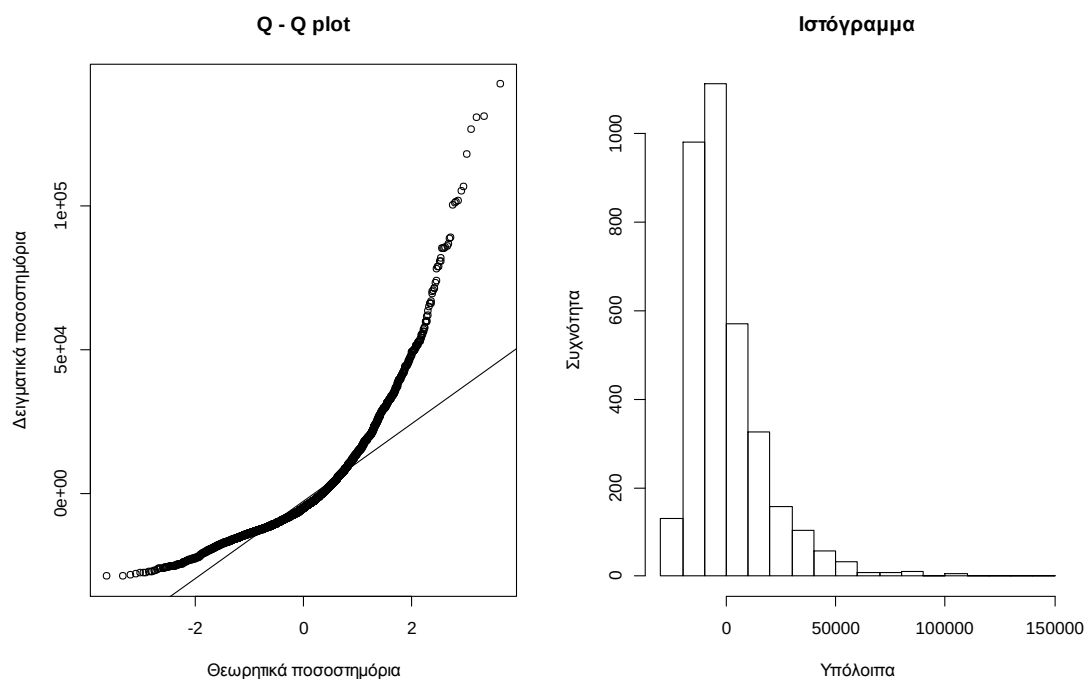
Διάγραμμα 4.13: Θηκογράφημα της συνολικής καταναλωτικής δαπάνης ανά κοινωνικο – οικονομική κατάσταση του προσώπου αναφοράς



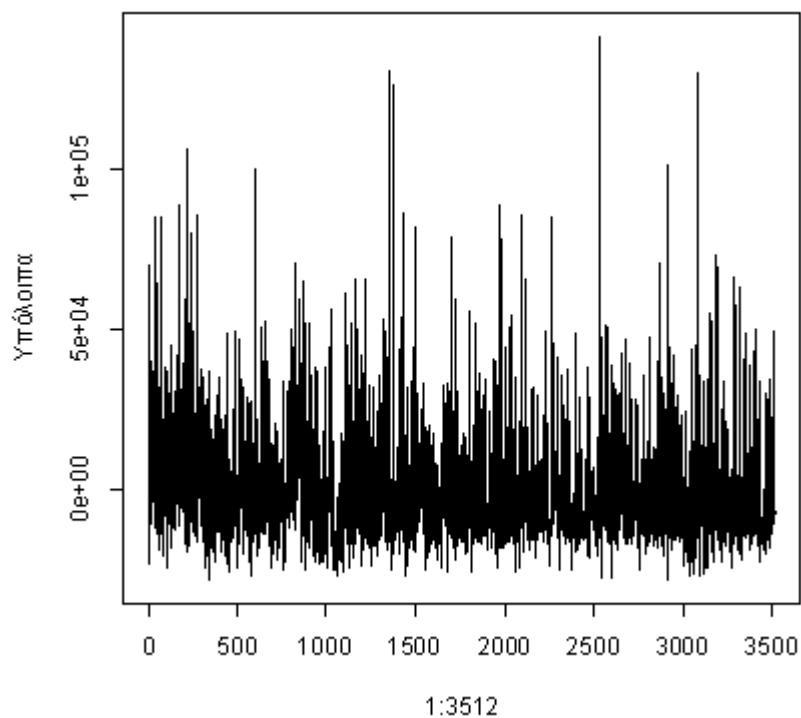
Συμπερασματικά, φαίνεται από το Διάγραμμα 4.13 ότι η συνολική καταναλωτική δαπάνη διαφοροποιείται σε σχέση με την κοινωνικο – οικονομική κατάσταση του προσώπου αναφοράς των νοικοκυριών. Και πιο συγκεκριμένα, για κάποιες στάθμες, όπως οι μη χειρώνακτες εργαζόμενοι και οι αυτοαπασχολούμενοι – γεωργοί – κτηνοτρόφοι, η συνολική καταναλωτική δαπάνη φαίνεται να είναι μεγαλύτερη από ότι για κάποιες άλλες, όπως οι χειρώνακτες εργαζόμενοι, και ακόμη μεγαλύτερη από τις υπόλοιπες. Για το λόγο αυτό, θα γίνει έλεγχος για το αν η παρατηρούμενη αυτή διαφορά είναι στατιστικά σημαντική.

Πριν γίνει η ανάλυση διασποράς πρέπει να πραγματοποιηθεί ο έλεγχος αν πληρούνται οι προϋποθέσεις του μοντέλου, δηλαδή τα σφάλματα να ακολουθούν την κανονική κατανομή (κανονικότητα), η διασπορά της εξαρτημένης τυχαιάς μεταβλητής δοθείσης της επεξηγηματικής τυχαιάς μεταβλητής να είναι ίδια σε όλες τις ομάδες (ομοσκεδαστικότητα) και τα σφάλματα να είναι ανεξάρτητες τυχαιές μεταβλητές (ανεξαρτησία).

Διάγραμμα 4.14: Διάγραμμα Q – Q και Ιστόγραμμα συχνοτήτων των υπολοίπων για τον έλεγχο κανονικότητας των υπολοίπων για το μοντέλο της ανάλυσης διασποράς κατά έναν παράγοντα



Διάγραμμα 4.15: Διάγραμμα για τον έλεγχο ανεξαρτησίας των υπολοίπων για το μοντέλο της ανάλυσης διασποράς κατά έναν παράγοντα



Όσον αφορά στον έλεγχο της κανονικότητας των υπολοίπων κατασκευάζεται Διάγραμμα Q – Q και το ιστόγραμμα συχνοτήτων με τις εντολές `qqnorm()` – `qqline()` και `hist()` αντίστοιχα. Σχετικά με την υπόθεση της ομοσκεδαστικότητας εκτελείται ο έλεγχος του Levene (Levene's test) μέσω της εντολής `levene.test()` που βρίσκεται στη βιβλιοθήκη `lawstat` καθώς και ο έλεγχος Fligner – Killeen (Fligner – Killeen test), επειδή είναι πιο ανθεκτικός έλεγχος σε περιπτώσεις αποκλίσεων από την κανονικότητα, μέσω της εντολής `fligner.test()`. Και τέλος, για την ανεξαρτησία των υπολοίπων κατασκευάζεται ένα διάγραμμα υπολοίπων σε σχέση με τη σειρά των δεδομένων με την εντολή `plot()`.

Εκτελώντας τους ελέγχους αυτούς απορρίπτεται η υπόθεση τόσο για την κανονικότητα των σφαλμάτων όσο και για την υπόθεση της ομοσκεδαστικότητας. Αυτό φαίνεται από το Διάγραμμα 4.14 καθώς η υπόθεση της κανονικότητας δεν φαίνεται λογική επειδή παρατηρούνται αποκλίσεις. Από τον έλεγχο του Levene προκύπτει η p – τιμή $< 2.2e-16 < 0.05$, οπότε υπάρχουν σοβαρές ενδείξεις εναντίον της υπόθεσης της ομοσκεδαστικότητας. Το συμπέρασμα αυτό επιβεβαιώνεται και από τον έλεγχο του Fligner – Killeen, καθώς η p – τιμή είναι η ίδια. Επιπρόσθετα, και από τα προηγούμενα στοιχεία για τις ομάδες φαίνεται η έλλειψη συμμετρίας, η συγκέντρωση προς τα αριστερά των παρατηρήσεων του δείγματος και οι διαφορές στη διασπορά μεταξύ των ομάδων. Από το Διάγραμμα 4.15 μπορεί να εκτιμηθεί η ανεξαρτησία των σφαλμάτων καθώς τα δεδομένα φαίνεται να συμπεριφέρονται τυχαία αλλά και βάσει του τρόπου σχεδιασμού της εν λόγω έρευνας μπορεί να θεωρηθεί ότι ικανοποιείται η υπόθεση αυτή.

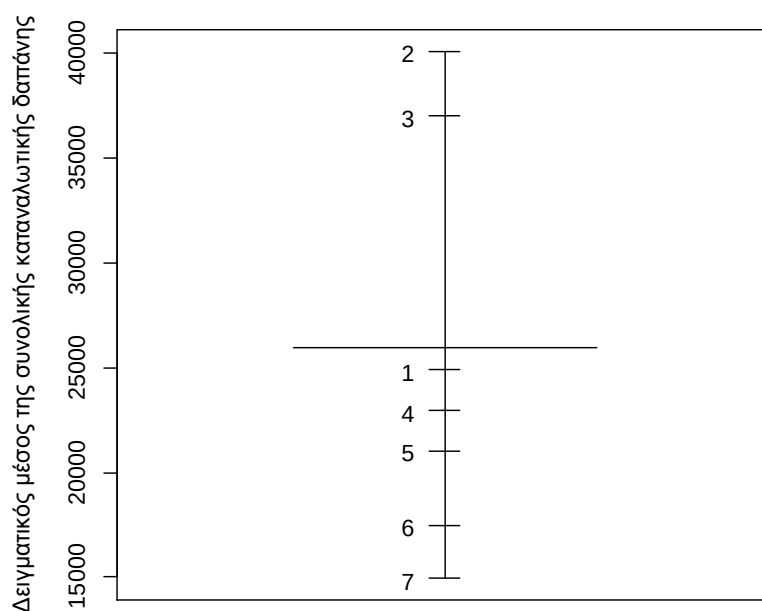
Εφ' όσον, λοιπόν, έχει απορριφθεί η μηδενική υπόθεση της ισότητας των διασπορών καθώς και η κανονικότητα των σφαλμάτων, θα εφαρμοσθεί στη συνέχεια ο μη παραμετρικός έλεγχος ανάλυσης διασποράς κατά έναν παράγοντα, ο οποίος καλείται Kruskal – Wallis test μέσω της εντολής `kruskal.test()`. Βάσει του ελέγχου αυτού, του οποίου η p – τιμή είναι $< 2.2e-16$, προκύπτει ότι υπάρχουν σοβαρές ενδείξεις για να απορριφθεί η μηδενική υπόθεση και άρα, παρατηρείται διαφοροποίηση στη συνολική καταναλωτική δαπάνη των νοικοκυριών ως προς την κοινωνικο – οικονομική κατάσταση του προσώπου αναφοράς του νοικοκυριού. Το ίδιο συμπέρασμα θα προέκυπτε ακόμα και αν είχε γίνει ο παραμετρικός έλεγχος μέσω της εντολής `aov()`, αφού η p – τιμή είναι και πάλι πάρα πολύ μικρή ($< 2e-16$).

Στη συνέχεια, υπολογίζονται οι εκτιμήσεις των επιδράσεων κάθε ομάδας, με τον περιορισμό ότι $\frac{\sum_{j=1}^k n_j a_j}{n} = 0$, μέσω της εντολής `model.tables()`. Με την προσθήκη του ορίσματος `type = 'means'` στην εν λόγω εντολή υπολογίζονται οι δειγματικοί μέσοι σε κάθε ομάδα του παράγοντα και ο αντίστοιχος αριθμός παρατηρήσεων σε κάθε ομάδα. Στον Πίνακα 4.8 και στο Διάγραμμα 4.16 παρουσιάζονται τα αποτελέσματα που προκύπτουν από την εντολή αυτή.

Πίνακας 4.8: Πίνακας των εκτιμήσεων των επιδράσεων της κοινωνικο – οικονομικής κατάστασης του προσώπου αναφοράς στη συνολική καταναλωτική δαπάνη

Κοινωνικο – οικονομική κατάσταση του προσώπου αναφοράς	Δειγματικός Μέσος	Επίδραση	Αριθμός Παρατηρήσεων	Συνολικός Δειγματικός Μέσος
Χειρώννακτας εργαζόμενος	24871	- 1046	553	25917.19
Μη χειρώννακτας εργαζόμενος	40087	14170	341	
Αυτοαπασχολούμενος, γεωργός ή κτηνοτρόφος	37001	11084	617	
Άνεργος	22924	- 2993	136	
Συνταξιούχος	21027	- 4890	1427	
Λοιποί μη οικονομικά ενεργοί	17471	- 8446	435	
Δεν έχει απαντηθεί	14951	- 10966	3	

Διάγραμμα 4.16: Γραφική αναπαράσταση των επιδράσεων των ομάδων του παράγοντα περιοχή στην συνολική καταναλωτική δαπάνη



Παράγοντας (κοινωνικο - οικονομική κατάσταση του προσώπου αναφοράς)

Από τον Πίνακα 4.8 και το Διάγραμμα 4.16 προκύπτει ότι μόνο η επίδραση των ομάδων 2 και 3, δηλαδή των νοικοκυριών με πρόσωπο αναφοράς μη χειρώννακτα εργαζόμενο και αυτοαπασχολούμενο, γεωργό ή κτηνοτρόφο, αυξάνει τη συνολική καταναλωτική δαπάνη ενώ η επίδραση των υπολοίπων τη μειώνουν.

Τέλος, για να εντοπιστούν οι εκτιμώμενες μέσες διαφορές μεταξύ όλων των ομάδων ανά δύο πραγματοποιείται στην R η εντολή TukeyHSD() και επιβεβαιώνεται η αρχική υπόθεση ότι η συνολική καταναλωτική δαπάνη των νοικοκυριών διαφοροποιείται ως προς την κοινωνικο – οικονομική κατάσταση του προσώπου αναφοράς του νοικοκυριού. Πιο αναλυτικά, αυτό προκύπτει από τα 95% διαστήματα εμπιστοσύνης ως προς τις εκτιμώμενες διαφορές μεταξύ των ομάδων. Το εν λόγω διάστημα για τις στάθμες 2 και 3, δηλαδή οι μη χειρώνακτες εργαζόμενοι και οι αυτοαπασχολούμενοι – γεωργοί – κτηνοτρόφοι αντίστοιχα, περιέχει τη μηδενική τιμή, γεγονός που σημαίνει ότι δε διαφοροποιούνται στατιστικά σημαντικά μεταξύ τους. Επίσης, το διάστημα για τις στάθμες 1 και 4, δηλαδή οι χειρώνακτες εργαζόμενοι και οι άνεργοι, περιέχει κι αυτό τη μηδενική τιμή και άρα, δε διαφοροποιούνται στατιστικά σημαντικά μεταξύ τους. Όσον αφορά στη στάθμη 5, δηλαδή οι συνταξιούχοι, φαίνεται να μην υπάρχει στατιστικά σημαντική διαφοροποίηση ως προς τη στάθμη 4 καθώς στο αντίστοιχο διάστημα δεν περιέχεται η μηδενική τιμή. Η στάθμη 6, δηλαδή οι λοιποί μη οικονομικά ενεργοί, φαίνεται να διαφοροποιείται στατιστικά σημαντικά ως προς όλες τις υπόλοιπες κατηγορίες, εφ' όσον κανένα από τα διαστήματα εμπιστοσύνης δεν περιέχει τη μηδενική τιμή. Τέλος, η στάθμη 7, δηλαδή εκείνοι που δεν έχουν απαντήσει το εν λόγω ερώτημα, δε διαφοροποιείται με καμία από τις υπόλοιπες στάθμες, καθώς όλα τα διαστήματα εμπιστοσύνης περιέχουν τη μηδενική τιμή.

Επομένως, η συνολική καταναλωτική δαπάνη για τα νοικοκυριά με χειρώνακτα εργαζόμενο είναι 24870€, μικρότερη κατά 15216.281€ από ότι στα νοικοκυριά με μη χειρώνακτα εργαζόμενο (40086.281€), μικρότερη κατά 12129.791€ από ότι στα νοικοκυριά με αυτοαπασχολούμενο – κτηνοτρόφο – γεωργό (36999.791€), μεγαλύτερη κατά 1947.122€ από ότι στα νοικοκυριά με άνεργο (22922.878€), μεγαλύτερη κατά 3843.505€ από ότι στα νοικοκυριά με συνταξιούχο (21026.495€), μεγαλύτερη κατά 7399.431€ από ότι στα νοικοκυριά με μη οικονομικά ενεργό (17470.569€), μεγαλύτερη κατά 9919.426€ από ότι στα νοικοκυριά που δεν έχουν απαντηθεί στο ερώτημα (14950.574€).

Εντολές στην R

```
> boxplot (total ~ sitnew, names = c ('1', '2', '3', '4', '5', '6', '7'), ylim = c (0, 200000),
xlab = 'Κοινωνικο - οικονομική κατάσταση του προσώπου αναφοράς', ylab =
'Καταναλωτική δαπάνη')
> legend ('topright', c ('1: Χειρώνακτας εργαζόμενος', '2: Μη χειρώνακτας
εργαζόμενος', '3: Αυτοαπασχολούμενος, Γεωργός, Κτηνοτρόφος', '4: Άνεργος', '5:
Συνταξιούχος', '6: Λοιποί μη οικονομικά ενεργοί', '7: Δεν έχει απαντηθεί'), cex = 0.6,
bty = 'n')
> by (total, sitnew, summary)
sitnew: 1
  Min. 1st Qu.  Median   Mean 3rd Qu.   Max.
 5528 14920 20770 24870 30870 119700
-----
sitnew: 2
  Min. 1st Qu.  Median   Mean 3rd Qu.   Max.
 9032 25350 34400 40090 50540 136400
```



```

-----
sitnew: 3
  Min. 1st Qu. Median   Mean 3rd Qu.  Max.
  3182  21200  31770  37000  46380 170400
-----

sitnew: 4
  Min. 1st Qu. Median   Mean 3rd Qu.  Max.
  6245  13810  18900  22920  27610 127200
-----

sitnew: 5
  Min. 1st Qu. Median   Mean 3rd Qu.  Max.
  3854  10890  16350  21030  26300 149000
-----

sitnew: 6
  Min. 1st Qu. Median   Mean 3rd Qu.  Max.
  2503   9800  13470  17470  20640  73800
-----

sitnew: 7
  Min. 1st Qu. Median   Mean 3rd Qu.  Max.
  11190 13380 15570 14950 16830 18090
> by (total, sitnew, var)
sitnew: 1
[1] 219595867
-----

sitnew: 2
[1] 434920588
-----

sitnew: 3
[1] 513653107
-----

sitnew: 4
[1] 238383016
-----

sitnew: 5
[1] 234810827
-----

sitnew: 6
[1] 143766452
-----

sitnew: 7
[1] 12201063
> by (total, sitnew, sd)
sitnew: 1
[1] 14818.77
-----

sitnew: 2
[1] 20854.75
-----

sitnew: 3
[1] 22663.92
-----

```

```

sitnew: 4
[1] 15439.66
-----
sitnew: 5
[1] 15323.54
-----
sitnew: 6
[1] 11990.26
-----
sitnew: 7
[1] 3493.002
> s <- aov (total ~ sitnew)
> par (mfrow = c (1, 2))
> qqnorm (residuals (s), xlab = 'Θεωρητικά ποσοστημόρια', ylab = 'Δειγματικά
ποσοστημόρια', main = 'Q - Q plot')
> qqline (residuals (s))
> hist (residuals (s), main = 'Ιστόγραμμα', xlab = 'Υπόλοιπα', ylab = 'Συχνότητα' )
> lawstat::levene.test (total, sitnew, location = 'mean')
classical Levene's test based on the absolute deviations from the mean
( none not applied because the location is not set to median )
data: total
Test Statistic = 32.2951, p-value < 2.2e-16
> fligner.test (total ~ sitnew, data = dataHH)
Fligner-Killeen test of homogeneity of variances
data: total by sitnew
Fligner-Killeen:med chi-squared = 222.5668, df = 6, p-value < 2.2e-16
> plot (residuals (s), type = 'l', xlab = '1:3512', ylab = 'Υπόλοιπα')
> kruskal.test (total ~ sitnew)
Kruskal-Wallis rank sum test
data: total by sitnew
Kruskal-Wallis chi-squared = 754.251, df = 6, p-value < 2.2e-16
> summary (aov (total ~ sitnew))
      Df Sum Sq Mean Sq F value Pr(>F)
sitnew  1 8.495e+10 8.495e+10 261.2 <2e-16 ***
Residuals 3510 1.142e+12 3.252e+08
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> model.tables (aov (total ~ sitnew), type = 'means')
Tables of means
Grand mean
25917.19
sitnew
  1   2   3   4   5   6   7
24871 40087 37001 22924 21027 17471 14951
rep 553 341 617 136 1427 435 3
> model.tables (aov (total ~ sitnew))
Tables of effects
sitnew
  1   2   3   4   5   6   7
-1046 14170 11083 -2993 -4890 -8446 -10966

```

```

rep 553 341 617 136 1427 435 3
> plot.design (total ~ sitnew, xlab = 'Παράγοντας (κοινωνικο - οικονομική κατάσταση
του προσώπου αναφοράς)', ylab = 'Δειγματικός μέσος της συνολικής καταναλωτικής
δαπάνης', xaxt = NA)
> TukeyHSD (aov (total~sitnew))
Tukey multiple comparisons of means
95% family-wise confidence level
Fit: aov(formula = total ~ sitnew)
$sitnew
      diff      lwr      upr    p adj
2-1 15216.281 11759.808 18672.7536 0.0000000
3-1 12129.791 9190.166 15069.4162 0.0000000
4-1 -1947.122 -6751.993 2857.7496 0.8961016
5-1 -3843.505 -6358.066 -1328.9437 0.0001366
6-1 -7399.431 -10616.611 -4182.2501 0.0000000
7-1 -9919.426 -38980.948 19142.0965 0.9525804
3-2 -3086.490 -6473.895 300.9155 0.1015284
4-2 -17163.402 -22254.562 -12072.2428 0.0000000
5-2 -19059.786 -22085.696 -16033.8749 0.0000000
6-2 -22615.711 -26246.602 -18984.8205 0.0000000
7-2 -25135.706 -54245.931 3974.5185 0.1429918
4-3 -14076.913 -18832.341 -9321.4847 0.0000000
5-3 -15973.296 -18392.041 -13554.5513 0.0000000
6-3 -19529.222 -22672.080 -16386.3634 0.0000000
7-3 -22049.217 -51102.605 7004.1717 0.2749072
5-4 -1896.383 -6401.460 2608.6935 0.8777499
6-4 -5452.309 -10384.139 -520.4788 0.0192340
7-4 -7972.304 -37273.239 21328.6305 0.9847677
6-5 -3555.926 -6305.322 -806.5293 0.0026409
7-5 -6075.921 -35089.383 22937.5416 0.9962716
7-6 -2519.995 -31602.778 26562.7874 0.9999771

```

4.2. Εφαρμογή της Πολυμεταβλητής Ανάλυσης Διασποράς

4.2.1. Επίδραση της περιοχής στις επιμέρους κατηγορίες δαπανών των νοικοκυριών

ΔΙΑΤΥΠΩΣΗ ΥΠΟΘΕΣΗΣ

Θα εφαρμοσθεί η Πολυμεταβλητή Ανάλυση Διασποράς με έναν παράγοντα (one – way MANOVA) προκειμένου να ελεγχθεί αν διαφοροποιείται η μέση τιμή των 12 κατηγοριών των δαπανών των νοικοκυριών στις 4 περιοχές της Ελλάδας. Οι εξαρτημένες μεταβλητές είναι οι συνεχείς τυχαίες μεταβλητές για τις 12 κατηγορίες δαπανών (food, alcohol, clothing, housing, equipment, health, transport, communication, culture, education, hotels και goods) και η ανεξάρτητη μεταβλητή είναι η κατηγορική τυχαία μεταβλητή περιοχή (region), η οποία διαφοροποιείται σε 4 στάθμες: Βόρεια Ελλάδα, Κεντρική Ελλάδα, Αττική και Νησιά Αιγαίου – Κρήτη.

ΑΠΟΤΕΛΕΣΜΑΤΑ – ΕΡΜΗΝΕΙΑ

Αρχικά, στον Πίνακα 4.9 δίνονται τα μεγέθη των 4 συγκρινόμενων ομάδων της ανεξάρτητης μεταβλητής, τα οποία υπολογίσθηκαν στην R με την εντολή `aggregate()`.

Πίνακας 4.9: Πίνακας για το μέγεθος του δείγματος ανά περιοχή

Περιοχή	Μέγεθος κάθε κατηγορίας δαπανών
Βόρεια Ελλάδα	1105
Κεντρική Ελλάδα	774
Αττική	1300
Νησιά Αιγαίου – Κρήτη	333
Σύνολο	3512

Προφανώς για όλες τις επιμέρους κατηγορίες δαπανών των νοικοκυριών το μέγεθος του δείγματος είναι το ίδιο ανά περιοχή, εφ' όσον έχουν καταγραφεί οι δαπάνες για όλες τις κατηγορίες σε κάθε νοικοκυριό.

Στη συνέχεια, δίνονται στον πίνακα 4.10 περιγραφικά στοιχεία για τις κατηγορίες δαπανών ως προς τις διαφορετικές ομάδες της ανεξάρτητης μεταβλητής, μέσω της εντολής `aggregate()` και συγκεκριμένα η μέση τιμή αυτών (`mean`).

Πίνακας 4.10: Πίνακας για τη μέση τιμή των επιμέρους κατηγοριών των δαπανών (σε €) ανά περιοχή

Περιοχή	Κατηγορίες Δαπανών					
	Είδη Διατροφής	Οινοπνευματώδη Ποτά - Καπνός	Είδη Ένδυσης και Υπόδησης	Στέγαση	Διαρκή Αγαθά	Υγεία
Βόρεια Ελλάδα	4127.163	763.7229	1439.589	6836.509	1394.991	1423.945
Κεντρική Ελλάδα	4424.784	696.5377	1331.490	6607.512	1347.076	1444.429
Αττική	4424.200	778.9729	1606.014	8594.162	1669.904	1530.497
Νησιά Αιγαίου - Κρήτη	4279.289	608.2505	1450.659	6670.609	1288.165	1476.818

Περιοχή	Κατηγορίες Δαπανών					
	Μεταφορές	Επικοινωνίες	Αναψυχή - Πολιτισμός	Εκπαίδευση	Ξενοδοχεία, Καφεενεία, Εστιατόρια	Διάφορα άλλα αγαθά
Βόρεια Ελλάδα	2316.283	778.1719	1034.3704	730.2947	2273.319	1492.304
Κεντρική Ελλάδα	2314.958	793.1169	754.9211	471.3341	2145.931	1276.504
Αττική	2804.028	951.6075	1092.3244	737.1696	2732.308	1880.156
Νησιά Αιγαίου - Κρήτη	2938.665	853.3528	731.1431	581.6047	2097.649	1382.931

Εν γένει, παρατηρούνται διαφοροποιήσεις στη μέση τιμή της κάθε κατηγορίας δαπανών ως προς τις 4 υποομάδες της περιοχής. Για το λόγο αυτό, θα γίνει έλεγχος για το αν οι παρατηρούμενες αυτές διαφορές είναι στατιστικά σημαντικές.

Πίνακας 4.11: Πίνακας για τη στατιστική σημαντικότητα της επίδρασης της ανεξάρτητης μεταβλητής περιοχή

	B.E.	Τιμή	F	p - τιμή
Pillai	3	0.056335	5.5802	< 2.2e – 16
Wilks	3	0.94443	5.6085	< 2.2e – 16
Hotelling – Lawley	3	0.058037	5.6355	< 2.2e – 16
Roy	3	0.038954	11.358	< 2.2e – 16

Αν και, ενδεχομένως, το «καλύτερο» κριτήριο για τη συγκεκριμένη περίπτωση είναι το Wilks, επειδή πρόκειται για ισοπληθή δείγματα με περισσότερες από δύο ομάδες, έχουν εφαρμοσθεί με την εντολή `summary()` και τα 4 κριτήρια (Wilks, Pillai, Hotelling – Lawley, Roy) με τα οποία μπορεί να μελετηθεί η κύρια επίδραση της ανεξάρτητης μεταβλητής στις εξαρτημένες. Προκύπτει, λοιπόν, ότι η εν λόγω επίδραση είναι στατιστικώς σημαντική, αφού η p – τιμή < 0.05 και κατά συνέπεια, απορρίπτεται σε επίπεδο σημαντικότητας 5% η υπόθεση της ισότητας των μέσων τιμών στις ομάδες της ανεξάρτητης μεταβλητής.

Επίσης, είναι σημαντικό να μελετηθεί περαιτέρω ποιες εξαρτημένες μεταβλητές διαφοροποιούνται περισσότερο ως προς την ανεξάρτητη μεταβλητή. Για το λόγο αυτό, θα πραγματοποιηθεί ο F έλεγχος σε μονομεταβλητό επίπεδο μέσω της εντολής `summary.aov()`. Τα αποτελέσματα συνοψίζονται στον Πίνακα 4.12.

Πίνακας 4.12: Πίνακας μονομεταβλητής ανάλυσης της ανεξάρτητης μεταβλητής περιοχή

Εξαρτημένη Μεταβλητή	B.E.	F	p - τιμή
food	3	4.0764	0.006699
alcohol	3	2.9418	0.03185
clothing	3	3.0805	0.02638
housing	3	35.022	< 2.2e – 16
equipment	3	3.9877	0.007576
health	3	0.4344	0.7284
transport	3	3.4079	0.01686
communication	3	14.895	1.236e – 09
culture	3	5.9938	0.000453
education	3	3.6724	0.01172
hotels	3	11.84	1.029e – 07
goods	3	15.162	8.383e – 10

Όπως προκύπτει από Πίνακα 4.12, η επίδραση για κάθε επιμέρους κατηγορία δαπανών είναι στατιστικά σημαντική, αφού η p – τιμή < 0.05 και κατά συνέπεια, απορρίπτεται σε επίπεδο σημαντικότητας 5% η υπόθεση της ισότητας των μέσων τιμών στις ομάδες της ανεξάρτητης μεταβλητής. Μοναδική εξαίρεση αποτελεί η κατηγορία δαπανών όσον αφορά στην Υγεία, αφού φαίνεται να υπάρχουν σοβαρές ενδείξεις προκειμένου να μην απορριφθεί η μηδενική υπόθεση.

Συμπερασματικά, γίνεται αποδεκτό ότι ως προς την περιοχή οι δαπάνες των νοικοκυριών που δε διαφοροποιούνται είναι για τη μεταβλητή της υγείας (health) ενώ αντίθετα εκείνες που διαφοροποιούνται περισσότερο είναι για τις μεταβλητές στέγαση (housing), επικοινωνίες (communication), ξενοδοχεία, καφεενεία και εστιατόρια (hotels) και διάφορα αγαθά και υπηρεσίες (goods).

Εντολές στην R

```
> Y <- cbind (food, alcohol, clothing, housing, equipment, health, transport,
communication, culture, education, hotels, goods)
> Y <- as.matrix (Y)
> length <- aggregate (Y, by = list (region), length)
> length
  Group.1 food alcohol clothing housing equipment health transport
1      1 1105   1105   1105   1105   1105   1105   1105
2      2  774    774    774    774    774    774    774
3      3 1300   1300   1300   1300   1300   1300   1300
4      4  333    333    333    333    333    333    333
communication culture education hotels goods
1      1105   1105   1105   1105   1105
2      774    774    774    774    774
3     1300   1300   1300   1300   1300
4      333    333    333    333    333
> mean <- aggregate (Y, by = list (region), mean)
> mean
  Group.1 food alcohol clothing housing equipment health transport
1      1 4127.163 763.7229 1439.589 6836.509 1394.991 1423.945 2316.283
2      2 4424.784 696.5377 1331.490 6607.512 1347.076 1444.429 2314.958
3      3 4424.200 778.9729 1606.014 8594.162 1669.904 1530.497 2804.028
4      4 4279.289 608.2505 1450.659 6670.609 1288.165 1476.818 2938.665
communication culture education hotels goods
1    778.1719 1034.3704 730.2947 2273.319 1492.304
2    793.1169 754.9211 471.3341 2145.931 1276.504
3    951.6075 1092.3244 737.1696 2732.308 1880.156
4    853.3528 731.1431 581.6047 2097.649 1382.931
> res <- manova (Y~region)
> summary (res, test = 'Pillai')
      Df Pillai approx F num Df den Df Pr(>F)
region  3 0.056335  5.5802  36 10497 < 2.2e-16 ***
Residuals 3508
---
```

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

> summary(res, test = 'Wilks')
      Df Wilks approx F num Df den Df Pr(>F)
region  3 0.94443  5.6085  36 10333 < 2.2e-16 ***
Residuals 3508
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(res, test = 'Hotelling-Lawley')
      Df Hotelling-Lawley approx F num Df den Df Pr(>F)
region  3  0.058037  5.6355  36 10487 < 2.2e-16 ***
Residuals 3508
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(res, test = 'Roy')
      Df Roy approx F num Df den Df Pr(>F)
region  3 0.038954 11.358  12 3499 < 2.2e-16 ***
Residuals 3508
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary.aov(res)
Response food :
      Df Sum Sq Mean Sq F value Pr(>F)
region  3 6.4227e+07 21408892 4.0764 0.006699 **
Residuals 3508 1.8424e+10 5251975
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Response alcohol :
      Df Sum Sq Mean Sq F value Pr(>F)
region  3 9838566 3279522 2.9418 0.03185 *
Residuals 3508 3910743489 1114807
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Response dothing :
      Df Sum Sq Mean Sq F value Pr(>F)
region  3 3.9797e+07 13265531 3.0805 0.02638 *
Residuals 3508 1.5107e+10 4306322
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Response housing :
      Df Sum Sq Mean Sq F value Pr(>F)
region  3 2.8664e+09 955471678 35.022 < 2.2e-16 ***
Residuals 3508 9.5706e+10 27282294
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Response equipment :
      Df Sum Sq Mean Sq F value Pr(>F)
region  3 8.0744e+07 26914624 3.9877 0.007576 **
Residuals 3508 2.3677e+10 6749419
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Response health :
      Df Sum Sq Mean Sq F value Pr(>F)

```

```

region      3 7.5933e+06 2531106 0.4344 0.7284
Residuals 3508 2.0440e+10 5826701
Response transport :
      Df Sum Sq Mean Sq F value Pr(>F)
region  3 2.3720e+08 79067623 3.4079 0.01686 *
Residuals 3508 8.1391e+10 23201492
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Response communication :
      Df Sum Sq Mean Sq F value Pr(>F)
region  3 21603078 7201026 14.895 1.236e-09 ***
Residuals 3508 1695985722 483462
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Response culture :
      Df Sum Sq Mean Sq F value Pr(>F)
region  3 7.8762e+07 26254057 5.9938 0.000453 ***
Residuals 3508 1.5366e+10 4380211
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Response education :
      Df Sum Sq Mean Sq F value Pr(>F)
region  3 4.2789e+07 14263013 3.6724 0.01172 *
Residuals 3508 1.3625e+10 3883839
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Response hotels :
      Df Sum Sq Mean Sq F value Pr(>F)
region  3 2.4169e+08 80561916 11.84 1.029e-07 ***
Residuals 3508 2.3869e+10 6804230
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Response goods :
      Df Sum Sq Mean Sq F value Pr(>F)
region  3 2.0983e+08 69943532 15.162 8.383e-10 ***
Residuals 3508 1.6182e+10 4612920
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

4.2.2. Επίδραση της κύριας πηγής εισοδήματος στις επιμέρους κατηγορίες των δαπανών των νοικοκυριών

ΔΙΑΤΥΠΩΣΗ ΥΠΟΘΕΣΗΣ

Θα εφαρμοσθεί η Πολυμεταβλητή Ανάλυση Διασποράς με έναν παράγοντα (one – way MANOVA) προκειμένου να ελεγχθεί αν διαφοροποιείται η μέση τιμή των 12 κατηγοριών των δαπανών των νοικοκυριών στις 6 ομάδες της κύριας πηγής εισοδήματος. Οι εξαρτημένες μεταβλητές είναι οι συνεχείς τυχαίες μεταβλητές για τις 12 κατηγορίες δαπανών (food, alcohol, clothing, housing, equipment, health, transport, communication, culture, education, hotels και goods) και η ανεξάρτητη

μεταβλητή είναι η κατηγορική τυχαία μεταβλητή κύρια πηγή εισοδήματος (source), η οποία διαφοροποιείται σε 6 στάθμες: μισθοί ή ημερομίσθια, εισόδημα από αυτοαπασχόληση, εισόδημα από ακίνητη περιουσία, συντάξεις, επιδόματα ανεργίας και άλλα επιδόματα ή εισοδήματα από άλλες πηγές.

ΑΠΟΤΕΛΕΣΜΑΤΑ – ΕΡΜΗΝΕΙΑ

Αρχικά, στον Πίνακα 4.13 δίνονται τα μεγέθη των 6 συγκρινόμενων ομάδων της ανεξάρτητης μεταβλητής, τα οποία υπολογίσθηκαν στην R με την εντολή `aggregate()`.

Πίνακας 4.13: Πίνακας για το μέγεθος του δείγματος ανά κύρια πηγή εισοδήματος

Κύρια πηγή εισοδήματος	Μέγεθος κάθε κατηγορίας δαπανών
Μισθοί ή ημερομίσθια	1201
Εισόδημα από αυτοαπασχόληση	531
Εισόδημα από ακίνητη περιουσία	79
Συντάξεις	1541
Επιδόματα ανεργίας	20
Άλλα επιδόματα ή εισοδήματα από άλλες πηγές	140
Σύνολο	3512

Προφανώς για όλες τις επιμέρους κατηγορίες δαπανών των νοικοκυριών το μέγεθος του δείγματος είναι το ίδιο ανά κύρια πηγή εισοδήματος, εφ' όσον έχουν καταγραφεί οι δαπάνες για όλες τις κατηγορίες σε κάθε νοικοκυριό.

Στη συνέχεια, δίνονται στον Πίνακα 4.14 περιγραφικά στοιχεία για τις κατηγορίες δαπανών ως προς τις διαφορετικές ομάδες της ανεξάρτητης μεταβλητής, μέσω της εντολής `aggregate()` και συγκεκριμένα η μέση τιμή αυτών (mean).

Πίνακας 4.14: Πίνακας για τη μέση τιμή των επιμέρους κατηγοριών των δαπανών ανά κύρια πηγή εισοδήματος

Κύρια Πηγή Εισοδήματος	Κατηγορίες Δαπανών					
	Είδη Διατροφής	Οινοπνευματώδη Ποτα, Καπνός	Είδη Ένδυσης και Υπόδησης	Στέγαση	Διαρκή Αγαθά	Υγεία
Μισθοί	4775.058	965.6383	1911.8748	8126.550	1599.6170	1364.0441
Αυτοαπασχόληση	5567.100	1174.3546	2379.3858	9050.952	1998.2581	1808.7627
Ακίνητη περιουσία	4248.620	873.9570	1898.2830	10094.367	1912.6754	2577.6091
Συντάξεις	3692.122	424.9229	864.0417	6368.114	1247.1774	1471.8286
Επιδόματα Ανεργίας	4705.518	980.4240	1149.7380	4996.692	1105.8780	1113.8100
Άλλα επιδόματα	2510.546	510.5357	915.3189	5611.330	761.4403	572.9143

Κύρια Πηγή Εισοδήματος	Κατηγορίες Δαπανών					
	Μεταφορές	Επικοινωνίες	Αναψυχή - Πολιτισμός	Εκπαίδευση	Ξενοδοχεία, Καφεενεία, Εστιατόρια	Διάφορα Άλλα Αγαθά και Υπηρεσίες
Μισθοί	3561.089	1034.6672	1230.5936	1005.8902	3019.655	2022.514
Αυτοαπασχόληση	3702.657	1286.7015	1688.6707	1341.9788	3545.060	2291.630
Ακίνητη περιουσία	3815.452	1086.7641	1606.9610	674.3681	3326.806	2373.949
Συντάξεις	1381.929	573.6488	498.7846	142.9449	1457.528	1014.941
Επιδόματα Ανεργίας	3696.420	615.3480	702.7740	326.0940	1764.126	1411.482
Άλλα επιδόματα	1622.863	621.2906	760.8643	878.8800	2645.018	828.840

Εν γένει, παρατηρούνται διαφοροποιήσεις στη μέση τιμή της κάθε κατηγορίας δαπανών ως προς τις 6 υποομάδες της κύριας πηγής εισοδήματος. Για το λόγο αυτό, θα γίνει έλεγχος για το αν οι παρατηρούμενες αυτές διαφορές είναι στατιστικά σημαντικές.

Πίνακας 4.15: Πίνακας για τη στατιστική σημαντικότητα της επίδρασης της ανεξάρτητης μεταβλητής κύρια πηγή εισοδήματος

	B.E.	Τιμή	F	p - τιμή
Pillai	5	0.29993	18.607	< 2.2e – 16
Wilks	5	0.717	20.088	< 2.2e – 16
Hotelling – Lawley	5	0.37148	21.629	< 2.2e – 16
Roy	5	0.29956	87.346	< 2.2e – 16

Αν και, ενδεχομένως, το «καλύτερο» κριτήριο για τη συγκεκριμένη περίπτωση είναι το Wilks, επειδή πρόκειται για ισοπληθή δείγματα με περισσότερες από δύο ομάδες, έχουν εφαρμοσθεί (summary) και τα 4 κριτήρια (Wilks, Pillai, Hotelling – Lawley, Roy) με τα οποία μπορεί να μελετηθεί η κύρια επίδραση της ανεξάρτητης μεταβλητής στις εξαρτημένες. Προκύπτει, λοιπόν, ότι η εν λόγω επίδραση είναι στατιστικώς σημαντική, αφού η $p - τιμή < 0.05$ και κατά συνέπεια, απορρίπτεται σε

επίπεδο σημαντικότητας 5% η υπόθεση της ισότητας των μέσων τιμών στις ομάδες της ανεξάρτητης μεταβλητής.

Επίσης, είναι σημαντικό να μελετηθεί περαιτέρω ποιες εξαρτημένες μεταβλητές διαφοροποιούνται περισσότερο ως προς την ανεξάρτητη μεταβλητή. Για το λόγο αυτό, θα πραγματοποιηθεί ο F έλεγχος σε μονομεταβλητό επίπεδο μέσω της εντολής `summary.aov()`. Τα αποτελέσματα συνοψίζονται στον Πίνακα 4.16.

Πίνακας 4.16: Πίνακας μονομεταβλητής ανάλυσης της ανεξάρτητης μεταβλητής κύρια πηγή εισοδήματος

Εξαρτημένη Μεταβλητή	B.E.	F	p - τιμή
food	5	91.971	< 2.2e – 16
alcohol	5	63.222	< 2.2e – 16
clothing	5	65.769	< 2.2e – 16
housing	5	36.345	< 2.2e – 16
equipment	5	9.9729	1.726e – 09
health	5	9.965	1.758e – 09
transport	5	39.076	< 2.2e – 16
communication	5	132.36	< 2.2e – 16
culture	5	35.163	< 2.2e – 16
education	5	44.267	< 2.2e – 16
hotels	5	85.138	< 2.2e – 16
goods	5	51.693	< 2.2e – 16

Όπως προκύπτει από τον Πίνακα 4.16, η επίδραση για κάθε επιμέρους κατηγορία δαπανών είναι στατιστικά σημαντική, αφού η p – τιμή < 0.05 και κατά συνέπεια, απορρίπτεται σε επίπεδο σημαντικότητας 5% η υπόθεση της ισότητας των μέσων τιμών στις ομάδες της ανεξάρτητης μεταβλητής.

Συμπερασματικά, γίνεται αποδεκτό ότι ως προς την κύρια πηγή εισοδήματος οι δαπάνες των νοικοκυριών που διαφοροποιούνται περισσότερο σε σχέση με τις υπόλοιπες είναι για τις μεταβλητές διαρκή αγαθά (equipment) και υγεία (health).

Εντολές στην R

```
> length <- aggregate (Y, by = list (source), length)
> length
Group.1 food alcohol clothing housing equipment health transport
1 1 1201 1201 1201 1201 1201 1201 1201
2 2 531 531 531 531 531 531 531
3 3 79 79 79 79 79 79 79
4 4 1541 1541 1541 1541 1541 1541 1541
5 5 20 20 20 20 20 20 20
6 6 140 140 140 140 140 140 140
communication culture education hotels goods
1 1201 1201 1201 1201 1201
```

```

2      531  531   531  531  531
3      79   79    79   79   79
4     1541 1541   1541 1541 1541
5      20   20    20   20   20
6     140  140   140  140  140
> mean <- aggregate (Y, by = list (source), mean)
> mean
  Group.1 food alcohol clothing housing equipment health transport
1  1 4775.058 965.6383 1911.8748 8126.550 1599.6170 1364.0441 3561.089
2  2 5567.100 1174.3546 2379.3858 9050.952 1998.2581 1808.7627 3702.657
3  3 4248.620 873.9570 1898.2830 10094.367 1912.6754 2577.6091 3815.452
4  4 3692.122 424.9229 864.0417 6368.114 1247.1774 1471.8286 1381.929
5  5 4705.518 980.4240 1149.7380 4996.692 1105.8780 1113.8100 3696.420
6  6 2510.546 510.5357 915.3189 5611.330 761.4403 572.9143 1622.863
  communication culture education hotels goods
1 1034.6672 1230.5936 1005.8902 3019.655 2022.514
2 1286.7015 1688.6707 1341.9788 3545.060 2291.630
3 1086.7641 1606.9610 674.3681 3326.806 2373.949
4  573.6488 498.7846 142.9449 1457.528 1014.941
5  615.3480 702.7740 326.0940 1764.126 1411.482
6  621.2906 760.8643 878.8800 2645.018 828.840
> res <- manova (Y~source)
> summary (res, test = 'Pillai')
      Df Pillai approx F num Df den Df  Pr(>F)
source    5 0.29993  18.607   60 17495 < 2.2e-16 ***
Residuals 3506
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary (res, test = 'Wilks')
      Df Wilks approx F num Df den Df  Pr(>F)
source    5 0.717  20.088   60 16369 < 2.2e-16 ***
Residuals 3506
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary (res, test = 'Hotelling-Lawley')
      Df Hotelling-Lawley approx F num Df den Df  Pr(>F)
source    5    0.37148  21.629   60 17467 < 2.2e-16 ***
Residuals 3506
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary (res, test = 'Roy')
      Df Roy approx F num Df den Df  Pr(>F)
source    5 0.29956  87.346   12 3499 < 2.2e-16 ***
Residuals 3506
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary.aov(res)
Response food :
      Df Sum Sq Mean Sq F value  Pr(>F)
source    5 2.1438e+09 428755118 91.971 < 2.2e-16 ***
Residuals 3506 1.6344e+10 4661831

```

```

---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Response alcohol :
      Df Sum Sq Mean Sq F value Pr(>F)
source  5 324252330 64850466 63.222 < 2.2e-16 ***
Residuals 3506 3596329725 1025764
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Response clothing :
      Df Sum Sq Mean Sq F value Pr(>F)
source  5 1.2988e+09 259765343 65.769 < 2.2e-16 ***
Residuals 3506 1.3848e+10 3949671
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Response housing :
      Df Sum Sq Mean Sq F value Pr(>F)
source  5 4.8575e+09 971506419 36.345 < 2.2e-16 ***
Residuals 3506 9.3715e+10 26729940
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Response equipment :
      Df Sum Sq Mean Sq F value Pr(>F)
source  5 3.3316e+08 66631711 9.9729 1.726e-09 ***
Residuals 3506 2.3425e+10 6681274
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Response health :
      Df Sum Sq Mean Sq F value Pr(>F)
source  5 2.8652e+08 57303491 9.965 1.758e-09 ***
Residuals 3506 2.0161e+10 5750469
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Response transport :
      Df Sum Sq Mean Sq F value Pr(>F)
source  5 4.3088e+09 861766221 39.076 < 2.2e-16 ***
Residuals 3506 7.7319e+10 22053396
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Response communication :
      Df Sum Sq Mean Sq F value Pr(>F)
source  5 272734105 54546821 132.36 < 2.2e-16 ***
Residuals 3506 1444854695 412109
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Response culture :
      Df Sum Sq Mean Sq F value Pr(>F)
source  5 7.3752e+08 147503031 35.163 < 2.2e-16 ***
Residuals 3506 1.4707e+10 4194816
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Response education :
      Df Sum Sq Mean Sq F value Pr(>F)
source  5 8.1158e+08 162315358 44.267 < 2.2e-16 ***
Residuals 3506 1.2856e+10 3666777
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Response hotels :

```

```

      Df Sum Sq Mean Sq F value Pr(>F)
source    5 2.6105e+09 522103771 85.138 < 2.2e-16 ***
Residuals 3506 2.1500e+10 6132460

```

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Response goods :

```

      Df Sum Sq Mean Sq F value Pr(>F)
source    5 1.1255e+09 225092011 51.693 < 2.2e-16 ***
Residuals 3506 1.5266e+10 4354391

```

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

4.2.3. Επίδραση του αριθμού των οικονομικά ενεργών μελών στις επιμέρους κατηγορίες των δαπανών των νοικοκυριών

ΔΙΑΤΥΠΩΣΗ ΥΠΟΘΕΣΗΣ

Θα εφαρμοσθεί η Πολυμεταβλητή Ανάλυση Διασποράς με έναν παράγοντα (one – way MANOVA) προκειμένου να ελεγχθεί αν διαφοροποιείται η μέση τιμή των 12 κατηγοριών των δαπανών των νοικοκυριών στις 6 ομάδες του αριθμού των οικονομικά ενεργών μελών του νοικοκυριού. Οι εξαρτημένες μεταβλητές είναι οι συνεχείς τυχαίες μεταβλητές για τις 12 κατηγορίες δαπανών (food, alcohol, clothing, housing, equipment, health, transport, communication, culture, education, hotels και goods) και η ανεξάρτητη μεταβλητή είναι η κατηγορική τυχαία μεταβλητή αριθμός των οικονομικά ενεργών μελών (members), η οποία διαφοροποιείται σε 6 στάθμες: 0, 1, 2, 3, 4 και 5 ή περισσότερα ενεργά οικονομικά μέλη αντίστοιχα.

ΑΠΟΤΕΛΕΣΜΑΤΑ – ΕΡΜΗΝΕΙΑ

Αρχικά, στον Πίνακα 4.17 δίνονται τα μεγέθη των 6 συγκρινόμενων ομάδων της ανεξάρτητης μεταβλητής, τα οποία υπολογίσθηκαν στην R με την εντολή aggregate()).

Πίνακας 4.17: Πίνακας για το μέγεθος του δείγματος ανά αριθμό οικονομικά ενεργών μελών του νοικοκυριού

Αριθμός οικονομικά ενεργών μελών του νοικοκυριού	Μέγεθος κάθε κατηγορίας δαπανών
0	391
1	1562
2	1287
3	203
4	64
> 5	5
Σύνολο	3512

Προφανώς για όλες τις επιμέρους κατηγορίες δαπανών των νοικοκυριών το μέγεθος του δείγματος είναι το ίδιο ανά αριθμό οικονομικά ενεργών μελών του νοικοκυριού, εφ' όσον έχουν καταγραφεί οι δαπάνες για όλες τις κατηγορίες σε κάθε νοικοκυριό.

Στη συνέχεια, δίνονται στον Πίνακα 4.18 περιγραφικά στοιχεία για τις κατηγορίες δαπανών ως προς τις διαφορετικές ομάδες της ανεξάρτητης μεταβλητής, μέσω της εντολής aggregate() και συγκεκριμένα η μέση τιμή αυτών (mean).

Πίνακας 4.18: Πίνακας για τη μέση τιμή των επιμέρους κατηγοριών των δαπανών ανά αριθμό οικονομικά ενεργών μελών του νοικοκυριού

Αριθμός οικονομικά ενεργών μελών	Κατηγορίες Δαπανών					
	Είδη Διατροφής	Οινοπνευματώδη Ποτά, Καπνός	Είδη Ένδυσης και Υπόδησης	Στέγαση	Διαρκή Αγαθά	Υγεία
0	2746.975	300.9922	686.8413	5541.176	883.0318	1039.814
1	3813.199	644.1648	1198.4853	6714.679	1166.4409	1198.235
2	5003.097	859.7611	1825.6865	8559.573	1883.6602	1744.757
3	6087.757	1206.6242	2438.4591	8671.820	2193.2991	2235.505
4	6542.706	1762.3406	3120.2381	8716.886	2217.5869	2970.199
> 5	7588.968	2025.1440	1451.4960	14584.296	1050.7680	1051.944

Αριθμός οικονομικά ενεργών μελών	Κατηγορίες Δαπανών					
	Μεταφορές	Επικοινωνίες	Αναψυχή - Πολιτισμός	Εκπαίδευση	Ξενοδοχεία, Καφεενεία, Εστιατόρια	Διάφορα Άλλα Αγαθά
0	962.6599	500.555	489.9087	360.4536	1487.814	789.6918
1	2084.6364	712.129	739.4286	496.3169	1939.732	1256.0740
2	3105.4447	1002.873	1245.4036	946.8761	2788.584	1979.7533
3	4593.6225	1410.372	1462.8018	680.7606	4084.941	2599.2343
4	6084.7294	1513.515	2124.3000	638.2012	5575.552	2813.0906
> 5	4768.7760	2615.544	1700.0640	1986.0000	7380.792	3069.3120

Εν γένει, παρατηρούνται διαφοροποιήσεις στη μέση τιμή της κάθε κατηγορίας δαπανών ως προς τις 6 υποομάδες του αριθμού των οικονομικά ενεργών μελών του νοικοκυριού. Για το λόγο αυτό, θα γίνει έλεγχος για το αν οι παρατηρούμενες αυτές διαφορές είναι στατιστικά σημαντικές.

Πίνακας 4.19: Πίνακας για τη στατιστική σημαντικότητα της επίδρασης της ανεξάρτητης μεταβλητής αριθμός οικονομικά ενεργών μελών του νοικοκυριού

	B.E.	Τιμή	F	p - τιμή
Pillai	5	0.27164	16.751	< 2.2e – 16
Wilks	5	0.73944	18.166	< 2.2e – 16
Hotelling – Lawley	5	0.33757	19.655	< 2.2e – 16
Roy	5	0.28941	84.387	< 2.2e – 16

Αν και, ενδεχομένως, το «καλύτερο» κριτήριο για τη συγκεκριμένη περίπτωση είναι το Wilks, επειδή πρόκειται για ισοπληθή δείγματα με περισσότερες από δύο ομάδες, έχουν εφαρμοσθεί (summary) και τα 4 κριτήρια (Wilks, Pillai, Hotelling – Lawley, Roy) με τα οποία μπορεί να μελετηθεί η κύρια επίδραση της ανεξάρτητης μεταβλητής στις εξαρτημένες. Προκύπτει, λοιπόν, ότι η εν λόγω επίδραση είναι στατιστικώς σημαντική, αφού η p – τιμή < 0.05 και κατά συνέπεια, απορρίπτεται σε επίπεδο σημαντικότητας 5% η υπόθεση της ισότητας των μέσων τιμών στις ομάδες της ανεξάρτητης μεταβλητής.

Επίσης, είναι σημαντικό να μελετηθεί περαιτέρω ποιες εξαρτημένες μεταβλητές διαφοροποιούνται περισσότερο ως προς την ανεξάρτητη μεταβλητή. Για το λόγο αυτό, θα πραγματοποιηθεί ο F έλεγχος σε μονομεταβλητό επίπεδο (summary.aov). Τα αποτελέσματα συνοψίζονται στον Πίνακα 4.20.

Πίνακας 4.20: Πίνακας μονομεταβλητής ανάλυσης της ανεξάρτητης μεταβλητής αριθμός οικονομικά ενεργών μελών του νοικοκυριού

Εξαρτημένη Μεταβλητή	B.E.	F	p - τιμή
food	5	134.37	$< 2.2e - 16$
alcohol	5	43.197	$< 2.2e - 16$
clothing	5	43.369	$< 2.2e - 16$
housing	5	33.628	$< 2.2e - 16$
equipment	5	19.462	$< 2.2e - 16$
health	5	19.336	$< 2.2e - 16$
transport	5	30.406	$< 2.2e - 16$
communication	5	100.33	$< 2.2e - 16$
culture	5	19.024	$< 2.2e - 16$
education	5	9.9777	$1.706e - 09$
hotels	5	70.276	$< 2.2e - 16$
goods	5	42.322	$< 2.2e - 16$

Όπως προκύπτει από τον Πίνακα 4.20, η επιδράση για κάθε επιμέρους κατηγορία δαπανών είναι στατιστικά σημαντική, αφού η p – τιμή < 0.05 και κατά συνέπεια, απορρίπτεται σε επίπεδο σημαντικότητας 5% η υπόθεση της ισότητας των μέσων τιμών στις ομάδες της ανεξάρτητης μεταβλητής.

Συμπερασματικά, γίνεται αποδεκτό ότι ως προς τον αριθμό των οικονομικά ενεργών μελών του νοικοκυριού οι δαπάνες των νοικοκυριών που διαφοροποιούνται λιγότερο σε σχέση με τις υπόλοιπες είναι για τη μεταβλητή εκπαίδευση (education).

Εντολές στην R

```
> length <- aggregate (Y, by = list (members), length)
> length
Group.1 food alcohol clothing housing equipment health transport
1      0 391    391    391    391    391 391    391
```



```

2  1 1562 1562 1562 1562 1562 1562 1562
3  2 1287 1287 1287 1287 1287 1287 1287
4  3 203 203 203 203 203 203 203
5  4 64 64 64 64 64 64 64
6  5 5 5 5 5 5 5 5
communication culture education hotels goods
1  391 391 391 391 391
2  1562 1562 1562 1562 1562
3  1287 1287 1287 1287 1287
4  203 203 203 203 203
5  64 64 64 64 64
6  5 5 5 5 5
> mean <- aggregate (Y, by = list (members), mean)
> mean
Group.1 food alcohol clothing housing equipment health transport
1  0 2746.975 300.9922 686.8413 5541.176 883.0318 1039.814 962.6599
2  1 3813.199 644.1648 1198.4853 6714.679 1166.4409 1198.235 2084.6364
3  2 5003.097 859.7611 1825.6865 8559.573 1883.6602 1744.757 3105.4447
4  3 6087.757 1206.6242 2438.4591 8671.820 2193.2991 2235.505 4593.6225
5  4 6542.706 1762.3406 3120.2381 8716.886 2217.5869 2970.199 6084.7294
6  5 7588.968 2025.1440 1451.4960 14584.296 1050.7680 1051.944 4768.7760
communication culture education hotels goods
1  500.555 489.9087 360.4536 1487.814 789.6918
2  712.129 739.4286 496.3169 1939.732 1256.0740
3  1002.873 1245.4036 946.8761 2788.584 1979.7533
4  1410.372 1462.8018 680.7606 4084.941 2599.2343
5  1513.515 2124.3000 638.2012 5575.552 2813.0906
6  2615.544 1700.0640 1986.0000 7380.792 3069.3120
> res <- manova (Y~members)
> summary (res, test = 'Pillai')
Df Pillai approx F num Df den Df Pr(>F)
members 5 0.27164 16.751 60 17495 < 2.2e-16 ***
Residuals 3506
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary (res, test = 'Wilks')
Df Wilks approx F num Df den Df Pr(>F)
members 5 0.73944 18.166 60 16369 < 2.2e-16 ***
Residuals 3506
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary (res, test = 'Hotelling-Lawley')
Df Hotelling-Lawley approx F num Df den Df Pr(>F)
members 5 0.33757 19.655 60 17467 < 2.2e-16 ***
Residuals 3506
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary (res, test = 'Roy')
Df Roy approx F num Df den Df Pr(>F)
members 5 0.28941 84.387 12 3499 < 2.2e-16 ***
Residuals 3506

```

```

---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary.aov(res)
Response food :
      Df Sum Sq Mean Sq F value Pr(>F)
members  5 2.9732e+09 594637388 134.37 < 2.2e-16 ***
Residuals 3506 1.5515e+10 4425262
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Response alcohol :
      Df Sum Sq Mean Sq F value Pr(>F)
members  5 227511943 45502389 43.197 < 2.2e-16 ***
Residuals 3506 3693070112 1053357
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Response dothing :
      Df Sum Sq Mean Sq F value Pr(>F)
members  5 8.8223e+08 176445496 43.369 < 2.2e-16 ***
Residuals 3506 1.4264e+10 4068496
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Response housing :
      Df Sum Sq Mean Sq F value Pr(>F)
members  5 4.5110e+09 902200322 33.628 < 2.2e-16 ***
Residuals 3506 9.4062e+10 26828780
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Response equipment :
      Df Sum Sq Mean Sq F value Pr(>F)
members  5 6.4159e+08 128318435 19.462 < 2.2e-16 ***
Residuals 3506 2.3116e+10 6593301
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Response health :
      Df Sum Sq Mean Sq F value Pr(>F)
members  5 5.4872e+08 109743782 19.336 < 2.2e-16 ***
Residuals 3506 1.9899e+10 5675682
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Response transport :
      Df Sum Sq Mean Sq F value Pr(>F)
members  5 3.3925e+09 678493703 30.406 < 2.2e-16 ***
Residuals 3506 7.8236e+10 22314765
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Response communication :
      Df Sum Sq Mean Sq F value Pr(>F)
members  5 214994291 42998858 100.33 < 2.2e-16 ***
Residuals 3506 1502594509 428578
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Response culture :

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
members	5	4.0794e+08	81588726	19.024	< 2.2e-16 ***
Residuals	3506	1.5037e+10	4288818		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Response education :

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
members	5	1.9175e+08	38349984	9.9777	1.706e-09 ***
Residuals	3506	1.3476e+10	3843567		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Response hotels :

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
members	5	2.1963e+09	439264684	70.276	< 2.2e-16 ***
Residuals	3506	2.1915e+10	6250599		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Response goods :

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
members	5	9.3305e+08	186609918	42.322	< 2.2e-16 ***
Residuals	3506	1.5459e+10	4409271		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

4.2.4. Επίδραση της κοινωνικο – οικονομικής κατάστασης του προσώπου αναφοράς στις επιμέρους κατηγορίες των δαπανών των νοικοκυριών

ΔΙΑΤΥΠΩΣΗ ΥΠΟΘΕΣΗΣ

Θα εφαρμοσθεί η Πολυμεταβλητή Ανάλυση Διασποράς με έναν παράγοντα (one – way MANOVA) προκειμένου να ελεγχθεί αν διαφοροποιείται η μέση τιμή των 12 κατηγοριών των δαπανών των νοικοκυριών στις 7 ομάδες της κοινωνικο – οικονομικής κατάστασης του προσώπου αναφοράς του νοικοκυριού. Οι εξαρτημένες μεταβλητές είναι οι συνεχείς τυχαίες μεταβλητές για τις 12 κατηγορίες δαπανών (food, alcohol, clothing, housing, equipment, health, transport, communication, culture, education, hotels και goods) και η ανεξάρτητη μεταβλητή είναι η κατηγορική τυχαία μεταβλητή κοινωνικο – οικονομική κατάσταση (sitnew), η οποία διαφοροποιείται σε 7 στάθμες: χειρώνακτας εργαζόμενος, μη χειρώνακτας εργαζόμενος, αυτοαπασχολούμενος – γεωργός – κτηνοτρόφος, άνεργος, συνταξιούχος, λοιποί μη οικονομικά ενεργοί και δεν έχει απαντηθεί.

ΑΠΟΤΕΛΕΣΜΑΤΑ – ΕΡΜΗΝΕΙΑ

Αρχικά, στον Πίνακα 4.21 δίνονται τα μεγέθη των 7 συγκρινόμενων ομάδων της ανεξάρτητης μεταβλητής, τα οποία υπολογίσθηκαν στην R με την εντολή aggregate()).

Πίνακας 4.21: Πίνακας για το μέγεθος του δείγματος ανά κοινωνικο – οικονομική κατάσταση του προσώπου αναφοράς του νοικοκυριού

Κοινωνικο – οικονομική κατάσταση του προσώπου αναφοράς	Μέγεθος κάθε κατηγορίας δαπανών
Χειρώνακτας εργαζόμενος	553
Μη χειρώνακτας εργαζόμενος	341
Αυτοαπασχολούμενος, γεωργός, κτηνοτρόφος	617
Άνεργος	136
Συνταξιούχος	1427
Λοιποί μη οικονομικά ενεργοί	435
Δεν έχει απαντηθεί	3
Σύνολο	3512

Προφανώς για όλες τις επιμέρους κατηγορίες δαπανών των νοικοκυριών το μέγεθος του δείγματος είναι το ίδιο ανά αριθμό οικονομικά ενεργών μελών του νοικοκυριού, εφ' όσον έχουν καταγραφεί οι δαπάνες για όλες τις κατηγορίες σε κάθε νοικοκυριό.

Στη συνέχεια, δίνονται στον Πίνακα 4.22 περιγραφικά στοιχεία για τις κατηγορίες δαπανών ως προς τις διαφορετικές ομάδες της ανεξάρτητης μεταβλητής, μέσω της εντολής `aggregate()` και συγκεκριμένα η μέση τιμή αυτών (`mean`).

Πίνακας 4.22: Πίνακας για τη μέση τιμή των επιμέρους κατηγοριών των δαπανών ανά κοινωνικο – οικονομική κατάσταση του προσώπου αναφοράς του νοικοκυριού

Κοινωνικο – οικονομική κατάσταση του προσώπου αναφοράς	Κατηγορίες Δαπανών					
	Είδη Διατροφής	Οινοπνευματώδη Ποτά, Καπνός	Είδη Ένδυσης και Υπόδησης	Στέγαση	Διαρκή Αγαθά	Υγεία
Χειρώνακτας εργαζόμενος	4488.264	911.4315	1510.7868	6829.769	1108.8914	1004.354
Μη χειρώνακτας εργαζόμενος	5100.327	919.6888	2500.6093	10836.399	2544.8034	1949.400
Αυτοαπασχολούμενος	5527.429	1218.7002	2443.3645	9073.410	2067.1992	1873.421
Άνεργος	4139.943	942.4579	1319.2032	6238.988	845.6047	1100.351
Συνταξιούχος	3959.261	489.2796	1016.2497	6718.790	1305.2651	1509.208
Λοιποί μη οικονομικά ενεργοί	2994.732	459.2712	838.7007	5851.980	1031.0836	1129.236
Δεν έχει απαντηθεί	4836.880	837.8000	681.3200	3361.520	447.4000	771.000

Κοινωνικο – οικονομική κατάσταση του προσώπου αναφοράς	Κατηγορίες Δαπανών					
	Μεταφορές	Επικοινωνίες	Αναψυχή - Πολιτισμός	Εκπαίδευση	Ξενοδοχεία, Καφεενεία, Εστιατόρια	Διάφορα Άλλα Αγαθά και Υπηρεσίες
Χειρώνακτας εργαζόμενος	2795.876	852.8767	750.3708	685.0939	2438.684	1494.4277
Μη χειρώνακτας εργαζόμενος	4513.215	1294.1349	2055.5159	1818.2133	3793.188	2761.6107
Αυτοαπασχολούμενος	4179.574	1333.9173	1784.8862	1362.2643	3745.037	2391.4141
Άνεργος	2335.331	721.3147	901.9932	685.8715	2170.860	1521.7862
Συνταξιούχος	1725.860	634.0276	577.8605	174.1967	1736.308	1181.0137
Λοιποί μη οικονομικά ενεργοί	1211.541	583.4698	519.8819	327.6808	1594.108	929.7098
Δεν έχει απαντηθεί	1243.520	791.7600	64.3200	0.0000	1447.520	468.3600

Εν γένει, παρατηρούνται διαφοροποιήσεις στη μέση τιμή της κάθε κατηγορίας δαπανών ως προς τις 7 υποομάδες της κοινωνικο – οικονομικής κατάστασης του προσώπου αναφοράς του νοικοκυριού. Για το λόγο αυτό, θα γίνει έλεγχος για το αν οι παρατηρούμενες αυτές διαφορές είναι στατιστικά σημαντικές.

Πίνακας 4.23: Πίνακας για τη στατιστική σημαντικότητα της επίδρασης της ανεξάρτητης μεταβλητής της κοινωνικο – οικονομικής κατάστασης του προσώπου αναφοράς του νοικοκυριού

	B.E.	Τιμή	F	p - τιμή
Pillai	6	0.32007	16.431	< 2.2e – 16
Wilks	6	0.70098	17.821	< 2.2e – 16
Hotelling – Lawley	6	0.39722	19.267	< 2.2e – 16
Roy	6	0.31231	91.063	< 2.2e – 16

Αν και, ενδεχομένως, το «καλύτερο» κριτήριο για τη συγκεκριμένη περίπτωση είναι το Wilks, επειδή πρόκειται για ισοπληθή δείγματα με περισσότερες από δύο ομάδες, έχουν εφαρμοσθεί (summary) και τα 4 κριτήρια (Wilks, Pillai, Hotelling –

Lawley, Roy) με τα οποία μπορεί να μελετηθεί η κύρια επίδραση της ανεξάρτητης μεταβλητής στις εξαρτημένες. Προκύπτει, λοιπόν, ότι η εν λόγω επίδραση είναι στατιστικώς σημαντική, αφού η p – τιμή < 0.05 και κατά συνέπεια, απορρίπτεται σε επίπεδο σημαντικότητας 5% η υπόθεση της ισότητας των μέσων τιμών στις ομάδες της ανεξάρτητης μεταβλητής.

Επίσης, είναι σημαντικό να μελετηθεί περαιτέρω ποιες εξαρτημένες μεταβλητές διαφοροποιούνται περισσότερο ως προς την ανεξάρτητη μεταβλητή. Για το λόγο αυτό, θα πραγματοποιηθεί ο F έλεγχος σε μονομεταβλητό επίπεδο (summary.aov). Τα αποτελέσματα συνοψίζονται στον Πίνακα 4.24.

Πίνακας 4.24: Πίνακας μονομεταβλητής ανάλυσης της ανεξάρτητης μεταβλητής κοινωνικο – οικονομική κατάσταση του προσώπου αναφοράς του νοικοκυριού

Εξαρτημένη Μεταβλητή	B.E.	F	p - τιμή
food	6	73.96	$< 2.2e - 16$
alcohol	6	48.096	$< 2.2e - 16$
clothing	6	60.412	$< 2.2e - 16$
housing	6	50.685	$< 2.2e - 16$
equipment	6	22.063	$< 2.2e - 16$
health	6	10.807	$6.03e - 12$
transport	6	36.061	$< 2.2e - 16$
communication	6	129.39	$< 2.2e - 16$
culture	6	46.943	$< 2.2e - 16$
education	6	53.581	$< 2.2e - 16$
hotels	6	73.662	$< 2.2e - 16$
goods	6	50.389	$< 2.2e - 16$

Όπως προκύπτει από τον Πίνακα 4.24, η επίδραση για κάθε επιμέρους κατηγορία δαπανών είναι στατιστικά σημαντική, αφού η p – τιμή < 0.05 και κατά συνέπεια, απορρίπτεται σε επίπεδο σημαντικότητας 5% η υπόθεση της ισότητας των μέσων τιμών στις ομάδες της ανεξάρτητης μεταβλητής.

Συμπερασματικά, γίνεται αποδεκτό ότι ως προς τον αριθμό των οικονομικά ενεργών μελών του νοικοκυριού οι δαπάνες των νοικοκυριών που διαφοροποιούνται λιγότερο σε σχέση με τις υπόλοιπες είναι για τη μεταβλητή υγεία (health).

Εντολές στην R

```
> length <- aggregate (Y, by = list (sitnew), length)
```

```
> length
```

```
Group.1 food alcohol clothing housing equipment health transport
```

```
1 1 553 553 553 553 553 553 553
```

```
2 2 341 341 341 341 341 341 341
```

```
3 3 617 617 617 617 617 617 617
```

```
4 4 136 136 136 136 136 136 136
```

```

5  5 1427 1427 1427 1427 1427 1427 1427
6  6 435 435 435 435 435 435 435
7  7 3 3 3 3 3 3 3
communication culture education hotels goods
1  553 553 553 553 553
2  341 341 341 341 341
3  617 617 617 617 617
4  136 136 136 136 136
5  1427 1427 1427 1427 1427
6  435 435 435 435 435
7  3 3 3 3 3
> mean <- aggregate (Y, by = list (sitnew), mean)
> mean
Group.1 food alcohol clothing housing equipment health transport
1  1 4488.264 911.4315 1510.7868 6829.769 1108.8914 1004.354 2795.876
2  2 5100.327 919.6888 2500.6093 10836.399 2544.8034 1949.400 4513.215
3  3 5527.429 1218.7002 2443.3645 9073.410 2067.1992 1873.421 4179.574
4  4 4139.943 942.4579 1319.2032 6238.988 845.6047 1100.351 2335.331
5  5 3959.261 489.2796 1016.2497 6718.790 1305.2651 1509.208 1725.860
6  6 2994.732 459.2712 838.7007 5851.980 1031.0836 1129.236 1211.541
7  7 4836.880 837.8000 681.3200 3361.520 447.4000 771.000 1243.520
communication culture education hotels goods
1  852.8767 750.3708 685.0939 2438.684 1494.4277
2  1294.1349 2055.5159 1818.2133 3793.188 2761.6107
3  1333.9173 1784.8862 1362.2643 3745.037 2391.4141
4  721.3147 901.9932 685.8715 2170.860 1521.7862
5  634.0276 577.8605 174.1967 1736.308 1181.0137
6  583.4698 519.8819 327.6808 1594.108 929.7098
7  791.7600 64.3200 0.0000 1447.520 468.3600
> res <- manova (Y~sitnew)
> summary (res, test = 'Pillai')
      Df Pillai approx F num Df den Df Pr(>F)
sitnew  6 0.32007 16.431 72 20994 < 2.2e-16 ***
Residuals 3505
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary (res, test = 'Wilks')
      Df Wilks approx F num Df den Df Pr(>F)
sitnew  6 0.70098 17.821 72 19015 < 2.2e-16 ***
Residuals 3505
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary (res, test = 'Hotelling-Lawley')
      Df Hotelling-Lawley approx F num Df den Df Pr(>F)
sitnew  6 0.39722 19.267 72 20954 < 2.2e-16 ***
Residuals 3505
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary (res, test = 'Roy')
      Df Roy approx F num Df den Df Pr(>F)
sitnew  6 0.31231 91.063 12 3499 < 2.2e-16 ***

```

Residuals 3505

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> summary.aov(res)

Response food :

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
sitnew	6	2.0777e+09	346282755	73.96	< 2.2e-16 ***
Residuals	3505	1.6410e+10	4682014		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Response alcohol :

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
sitnew	6	298237257	49706210	48.096	< 2.2e-16 ***
Residuals	3505	3622344798	1033479		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Response dothing :

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
sitnew	6	1.4196e+09	236593779	60.412	< 2.2e-16 ***
Residuals	3505	1.3727e+10	3916351		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Response housing :

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
sitnew	6	7.8697e+09	1311623433	50.685	< 2.2e-16 ***
Residuals	3505	9.0703e+10	25878163		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Response equipment :

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
sitnew	6	8.6464e+08	144107212	22.063	< 2.2e-16 ***
Residuals	3505	2.2893e+10	6531544		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Response health :

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
sitnew	6	3.7142e+08	61902652	10.807	6.03e-12 ***
Residuals	3505	2.0076e+10	5727887		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Response transport :

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
sitnew	6	4.7460e+09	790994740	36.061	< 2.2e-16 ***
Residuals	3505	7.6882e+10	21934969		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Response communication :

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
sitnew	6	311453059	51908843	129.39	< 2.2e-16 ***
Residuals	3505	1406135741	401180		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Response culture :

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
sitnew	6	1.1488e+09	191464909	46.943	< 2.2e-16 ***
Residuals	3505	1.4296e+10	4078674		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Response education :

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
sitnew	6	1.1483e+09	191380184	53.581	< 2.2e-16 ***
Residuals	3505	1.2519e+10	3571759		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Response hotels :

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
sitnew	6	2.6999e+09	449978998	73.662	< 2.2e-16 ***
Residuals	3505	2.1411e+10	6108716		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Response goods :

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
sitnew	6	1.3017e+09	216942062	50.389	< 2.2e-16 ***
Residuals	3505	1.5090e+10	4305365		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

4.3. Εφαρμογή της Γραμμικής Παλινδρόμησης

4.3.1. Απλή Γραμμική Παλινδρόμηση

ΔΙΑΤΥΠΩΣΗ ΥΠΟΘΕΣΗΣ

Στη συνέχεια, θα εφαρμοσθεί η Απλή Γραμμική Παλινδρόμηση προκειμένου να εκτιμηθεί η συνολική ετήσια καταναλωτική δαπάνη των νοικοκυριών συναρτήσει του συνολικού εισοδήματος του νοικοκυριού. Η εξαρτημένη μεταβλητή ή μεταβλητή απόκρισης (ποσοτική) είναι η συνολική ετήσια καταναλωτική δαπάνη (total) και η εξηγηματική μεταβλητή (ποσοτική) το συνολικό εισόδημα (income). Ορίζεται, λοιπόν, η υπόθεση ότι:

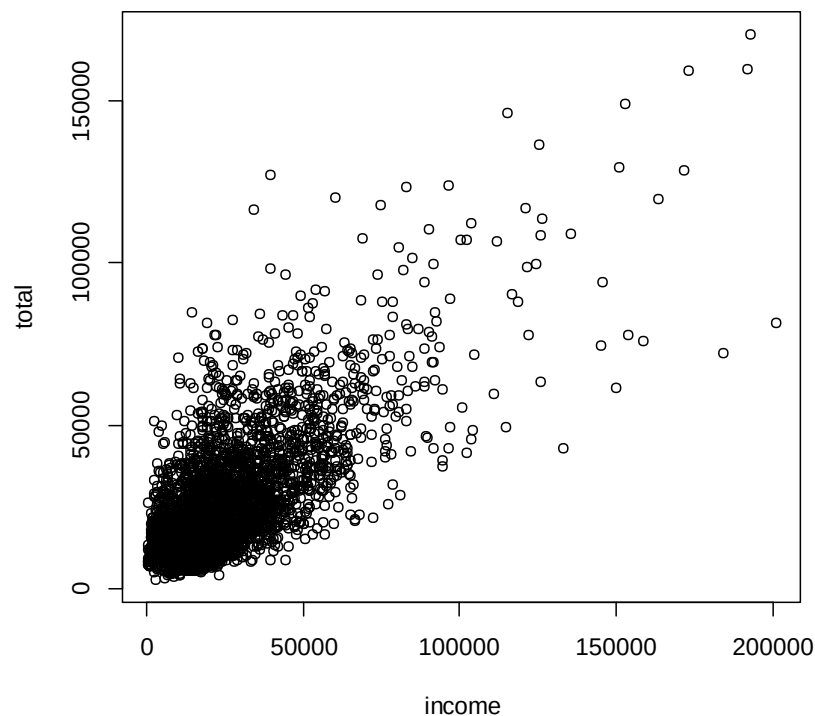
$$total_i = a + b_1 income_i + \varepsilon_i$$

Εντολές στην R

> plot (income, total)

- Με την εντολή plot() κατασκευάζεται το διάγραμμα διασποράς, δηλαδή μια απεικόνιση των σημείων (income, total), για να γίνει ο έλεγχος αν η γραμμική συνάρτηση φαίνεται να είναι κατάλληλη.

Διάγραμμα 4.17: Διάγραμμα διασποράς



> results <- lm (total ~ income)

> results

Call:

lm(formula = total ~ income)

Coefficients:

(Intercept) income

9378.6035 0.6418

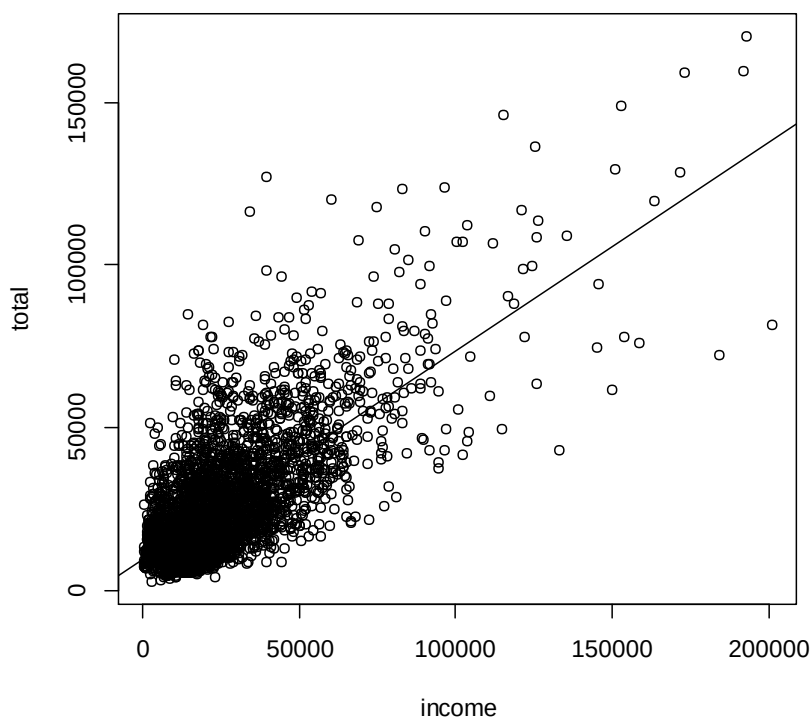
- Με την εντολή lm() προσαρμόζεται το γραμμικό μοντέλο παλινδρόμησης με total τα δεδομένα για τη μεταβλητή απόκρισης και income τα δεδομένα για την επεξηγηματική μεταβλητή.

> plot(income, total)

> abline(results)

- Με την εντολή abline() δίνεται η ευθεία των ελαχίστων τετραγώνων. Από το Διάγραμμα 4.18 γίνεται και ο έλεγχος για τη γραμμικότητα, η οποία φαίνεται ότι είναι λογική.

Διάγραμμα 4.18: Εκτιμώμενη ευθεία της απλής γραμμικής παλινδρόμησης



```
> summary(results)
```

```
Call:
```

```
lm(formula = total ~ income)
```

```
Residuals:
```

```
    Min     1Q  Median     3Q      Max 
-56887 -9175 -2616  6125  92609 
```

```
Coefficients:
```

```
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 9.379e+03  3.640e+02  25.77  <2e-16 ***
income      6.418e-01  1.108e-02   57.90  <2e-16 ***
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 13370 on 3510 degrees of freedom
```

```
Multiple R-squared:  0.4885,    Adjusted R-squared:  0.4883 
```

```
F-statistic: 3352 on 1 and 3510 DF, p-value: < 2.2e-16
```

- Με την εντολή `summary()` δίνονται εκτιμήσεις για τις παραμέτρους του γραμμικού μοντέλου.

```
> confint(results)
```

```
      2.5 %    97.5 %
```

```
(Intercept) 8664.9128205 1.009229e+04
```

```
income      0.6200234 6.634905e-01
```

- Με την εντολή `confint()` υπολογίζονται συμμετρικά 95% διαστήματα εμπιστοσύνης για τις παραμέτρους του μοντέλου.

```
> model.matrix(results)
```

```
(Intercept) income
1           1 23112.60
2           1 20319.96
3           1 15484.56
```

...

3510 1 34800.00

3511 1 20360.04

3512 1 29789.76

Με την εντολή `model.matrix()` εμφανίζονται οι εικονικές μεταβλητές που έχουν δημιουργηθεί για το εν λόγω γραμμικό μοντέλο.

```
> par (mfrow = c (1, 2))
```

```
> qqnorm (results$res)
```

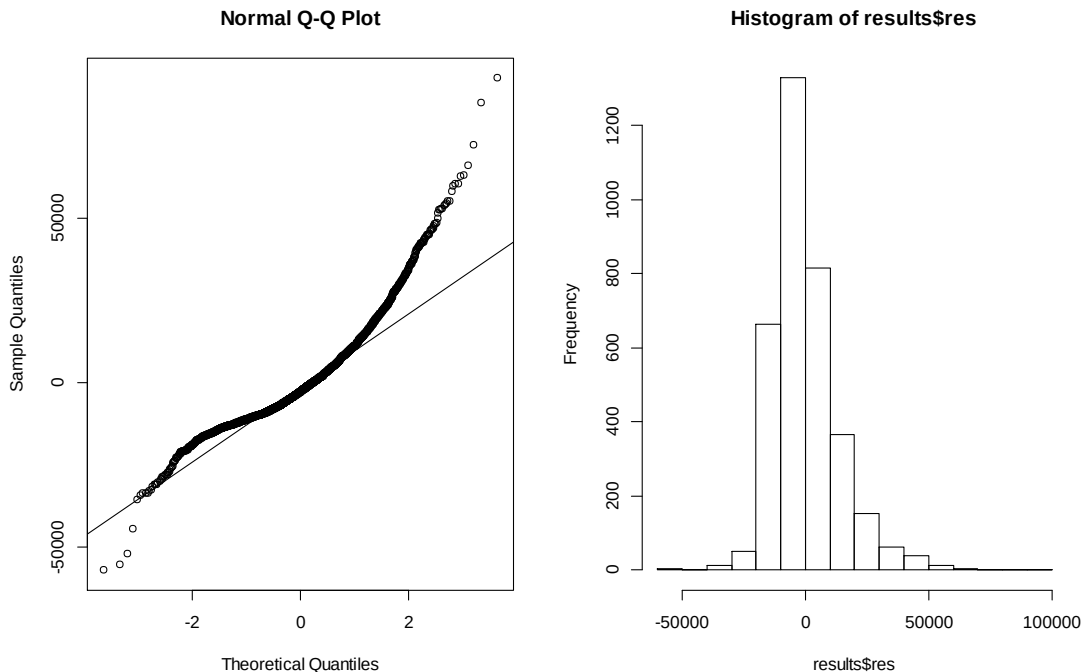
```
> qqline (results$res)
```

- Με την εντολή `qqnorm()` κατασκευάζεται ένα γράφημα του οποίου ο οριζόντιος άξονας αναπαριστά τα θεωρητικά ποσοστημόρια της κανονικής κατανομής και ο κατακόρυφος άξονας τα δειγματικά ποσοστημόρια, όπως αυτά προκύπτουν από το εν λόγω δείγμα. Ενώ με την εντολή `qqline()` προστίθεται η θεωρητική ευθεία στο ήδη υπάρχον γράφημα για τα ποσοστημόρια της κανονικής κατανομής.

```
> hist (results$res)
```

- Με την εντολή `hist()` κατασκευάζεται το ιστόγραμμα των υπολοίπων.

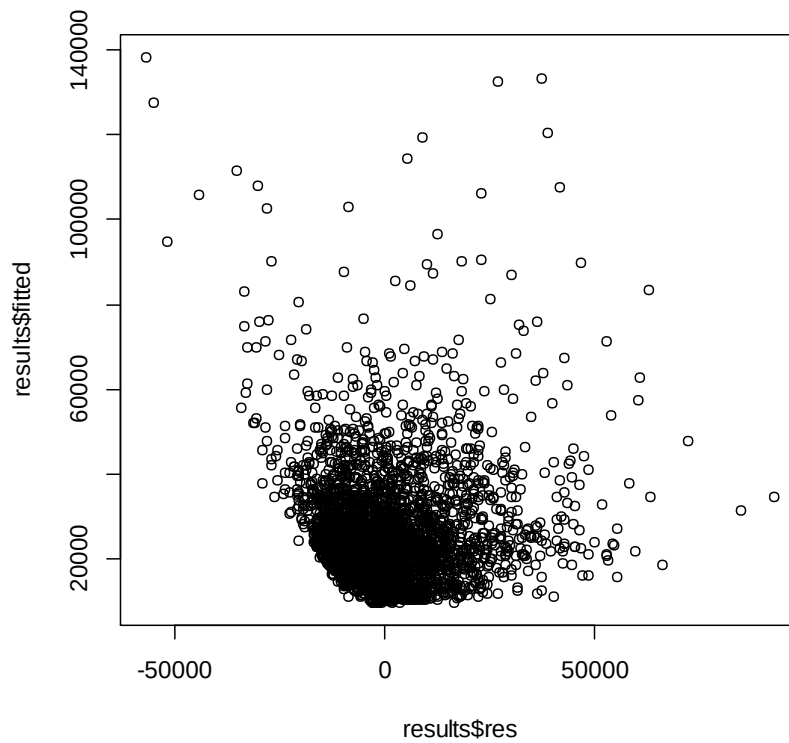
Διάγραμμα 4.19: QQ – plot και Ιστόγραμμα υπολοίπων



```
> plot (results$res, results$fitted)
```

- Με την εντολή `plot()` κατασκευάζεται το γράφημα των υπολοίπων `results$res` συναρτήσει των προβλεπόμενων τιμών `results$fitted`.

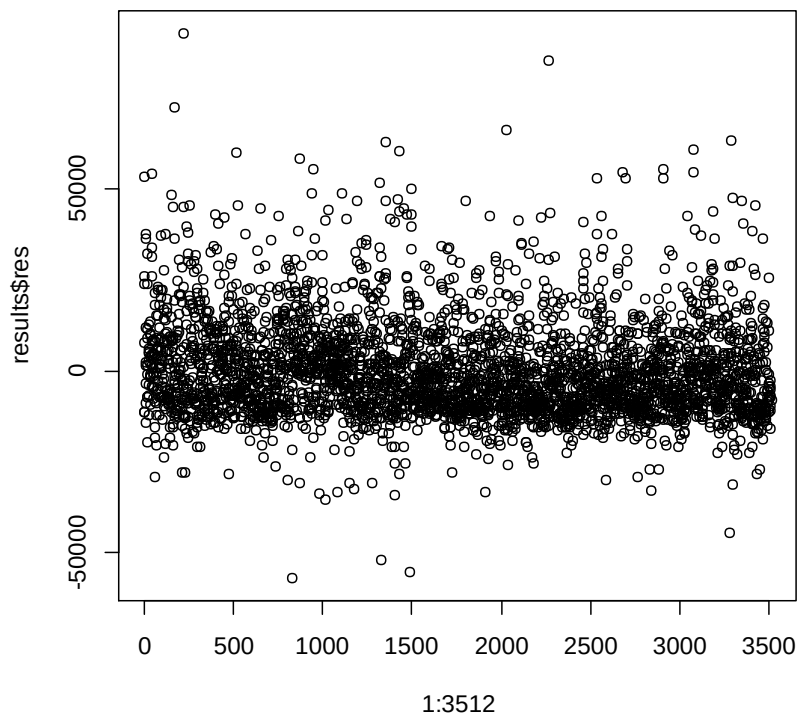
Διάγραμμα 4.20: Διάγραμμα για τον γραφικό έλεγχο της ομοσκεδαστικότητας



```
> plot(1:3512, results$res)
```

- Με την εντολή `plot()` κατασκευάζεται το γράφημα των υπολοίπων `results$res` σε σχέση με τη σειρά των δεδομένων.

Διάγραμμα 4.21: Διάγραμμα για τον γραφικό έλεγχο της ανεξαρτησίας

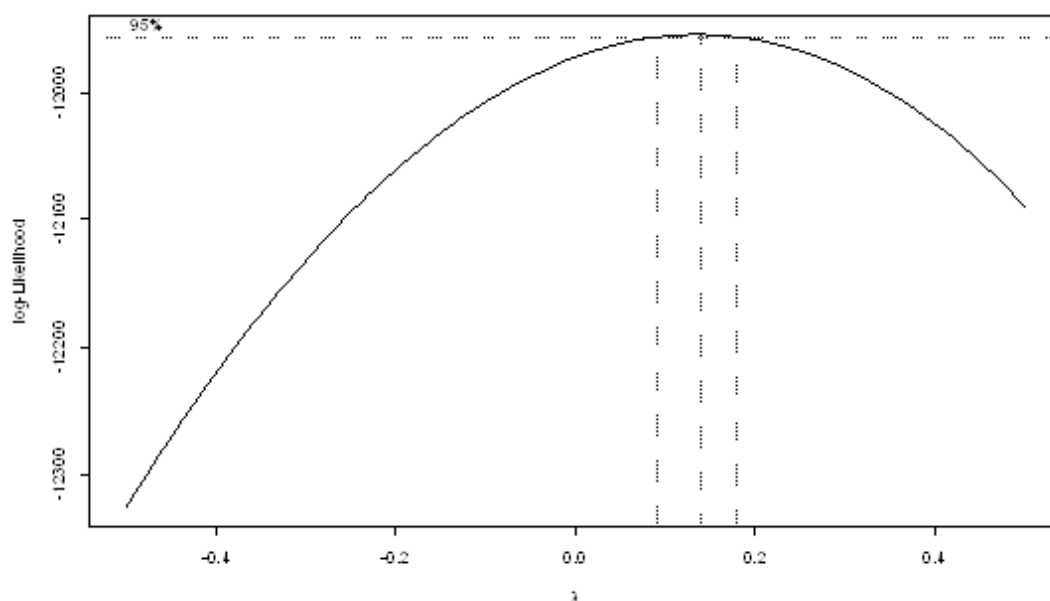


Η υπόθεση της κανονικότητας ελέγχεται με το Διάγραμμα 4.19 από το οποίο και προκύπτει να μην είναι λογική η εν λόγω υπόθεση για τα δεδομένα του δείγματος καθώς παρατηρούνται μεγάλες αποκλίσεις, δηλαδή τα υπόλοιπα δεν ακολουθούν την Κανονική κατανομή. Επίσης, από το ιστόγραμμα συχνοτήτων των υπολοίπων φαίνεται ότι δεν υπάρχει συμμετρία στη διασπορά τους γύρω από το 0. Η υπόθεση της ομοσκεδαστικότητας ελέγχεται με το Διάγραμμα 4.20 από το οποίο φαίνονται τα υπόλοιπα να κατανέμονται με μη τυχαίο ή μη ομοσκεδαστικό τρόπο. Η υπόθεση της ανεξαρτησίας των σφαλμάτων ελέγχεται με το Διάγραμμα 4.21 στο οποίο φαίνεται ότι η σειρά των δεδομένων δεν εμφανίζει κάποιου είδους εξάρτηση και κατανέμονται τυχαία και άρα, η υπόθεση της ανεξαρτησίας επιβεβαιώνεται.

```
> bc <- boxcox (total ~ income, lambda = seq (-0.5, 0.5, 0.01))
```

- Επειδή η υπόθεση της κανονικότητας δε φαίνεται λογική και παρατηρείται, επίσης, πρόβλημα στην ομοσκεδαστικότητα, θα πραγματοποιηθεί μετασχηματισμός της μεταβλητής απόκρισης με τη χρήση του μετασχηματισμού Box – Cox, ο οποίος υλοποιείται στην R με την εντολή `boxcox()`. Τα δεδομένα πρέπει να είναι θετικά, τα οποία στην εν λόγω περίπτωση είναι. Η ιδέα είναι να βρεθεί ο μετασχηματισμός δύναμης, λ , ο οποίος μεγιστοποιεί την πιθανότητα όταν ένα καθορισμένο σύνολο των επεξηγηματικών μεταβλητών είναι προσαρμοσμένο σε $(y^\lambda - 1)/\lambda$ ως απόκριση. Η τιμή του λ μπορεί να είναι θετική ή αρνητική, αλλά δεν μπορεί να είναι μηδέν. Για την περίπτωση $\lambda = 0$ ο μετασχηματισμός Box – Cox ορίζεται ως $\log(y)$. Η ποσότητα που προκύπτει εισάγεται στο μοντέλο της παλινδρόμησης έναντι των επεξηγηματικών μεταβλητών και του υπολογισμού της λογαριθμικής συνάρτησης πιθανότητας. Από το Διάγραμμα 4.22 προκύπτει ότι η βέλτιστη τιμή του λ είναι ίση με 0.14, οπότε και στη συνέχεια πραγματοποιείται ο μετασχηματισμός της μεταβλητής απόκρισης.

Διάγραμμα 4.22: Μετασχηματισμός Box – Cox



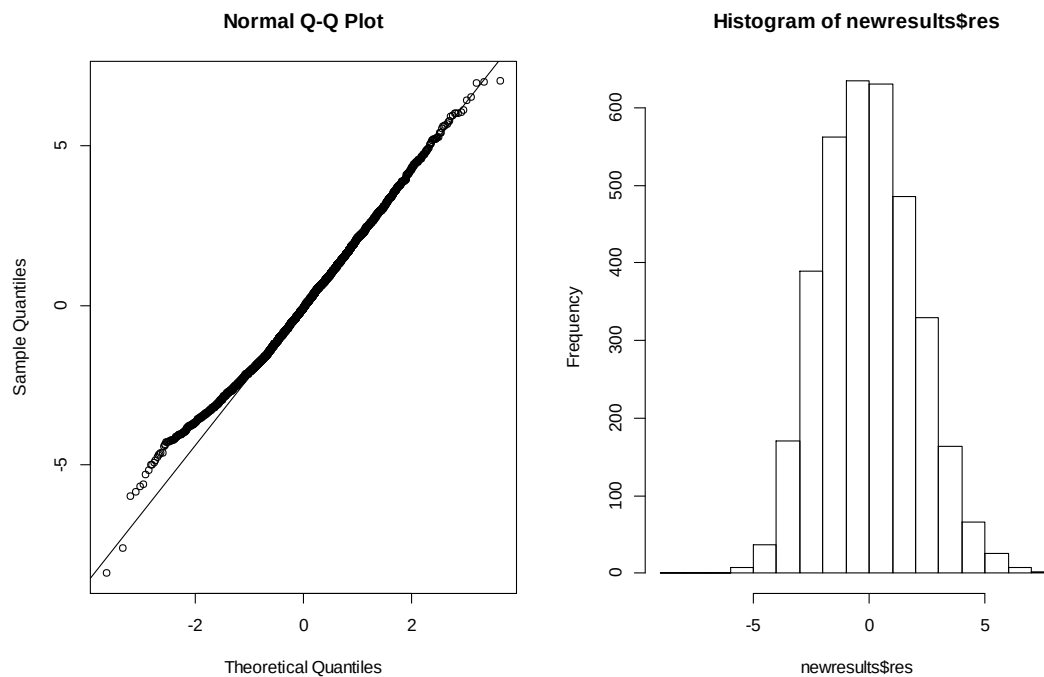
```

> newtotal <- (total ^ 0.14 - 1) / 0.14
> newresults <- lm (newtotal ~ income)
> newresults
Call:
lm(formula = newtotal ~ income)
Coefficients:
(Intercept)  income
 1.964e+01  8.163e-05
  ➤ Με την εντολή lm() προσαρμόζεται εκ νέου το γραμμικό μοντέλο
    παλινδρόμησης με newtotal τα δεδομένα για τη μεταβλητή απόκρισης.
> summary(newresults)
Call:
lm(formula = newtotal ~ income)
Residuals:
  Min   1Q Median   3Q   Max
-8.394 -1.543 -0.080  1.365  7.037
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.964e+01  5.567e-02  352.70  <2e-16 ***
income      8.163e-05  1.695e-06  48.15  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 2.045 on 3510 degrees of freedom
Multiple R-squared:  0.3978,    Adjusted R-squared:  0.3976
F-statistic: 2319 on 1 and 3510 DF,  p-value: < 2.2e-16
  ➤ Με την εντολή summary() δίνονται εκτιμήσεις για τις παραμέτρους του
    γραμμικού μοντέλου.
> confint(newresults)
            2.5 %    97.5 %
(Intercept) 1.952588e+01 1.974418e+01
income      7.830939e-05 8.495704e-05
  ➤ Με την εντολή confint() υπολογίζονται συμμετρικά 95% διαστήματα
    εμπιστοσύνης για τις παραμέτρους του μοντέλου.
> model.matrix(newresults)
(Intercept) income
1          1 23112.60
2          1 20319.96
3          1 15484.56
...
3510         1 134800.00
3511         1 120360.04
3512         1 129789.76
  ➤ Με την εντολή model.matrix() εμφανίζονται οι μεταβλητές για το εν λόγω
    γραμμικό μοντέλο.
> par (mfrow = c (1, 2))
> qqnorm (newresults$res)
> qqline (newresults$res)
> hist(newresults$res)

```

- Με τις εντολές `qqnorm()`, `qqline()` και `hist()` κατασκευάζεται το Διάγραμμα 4.23 για τον έλεγχο της κανονικότητας των δεδομένων, η οποία φαίνεται να ικανοποιείται καλύτερα στα νέα δεδομένα.

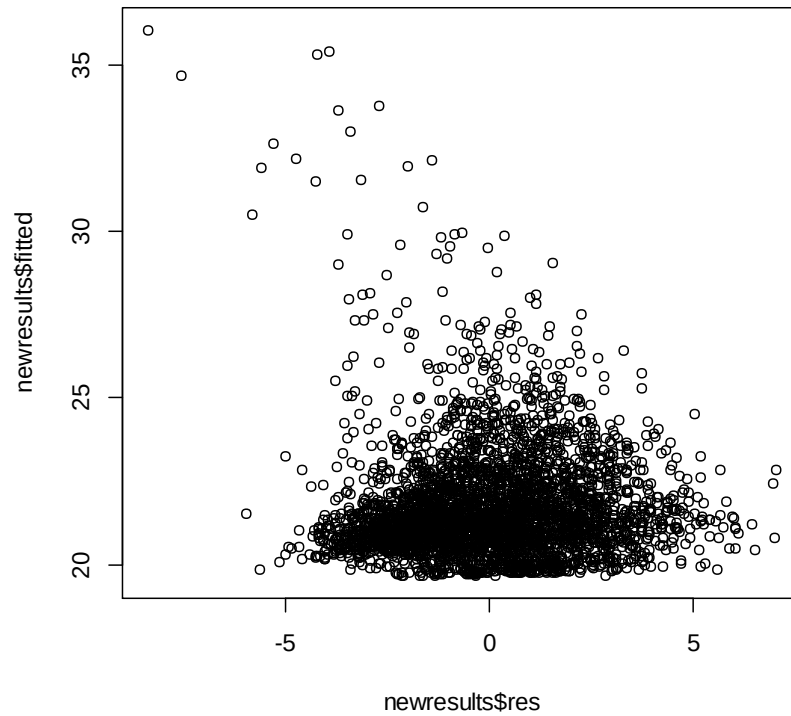
Διάγραμμα 4.23: QQ – plot και Ιστόγραμμα υπολοίπων



`> plot(newresults$res, newresults$fitted)`

- Με την εντολή `plot()` κατασκευάζεται το Διάγραμμα 4.24 για τον έλεγχο της ομοσκεδαστικότητας, η οποία φαίνεται να ικανοποιείται καλύτερα.

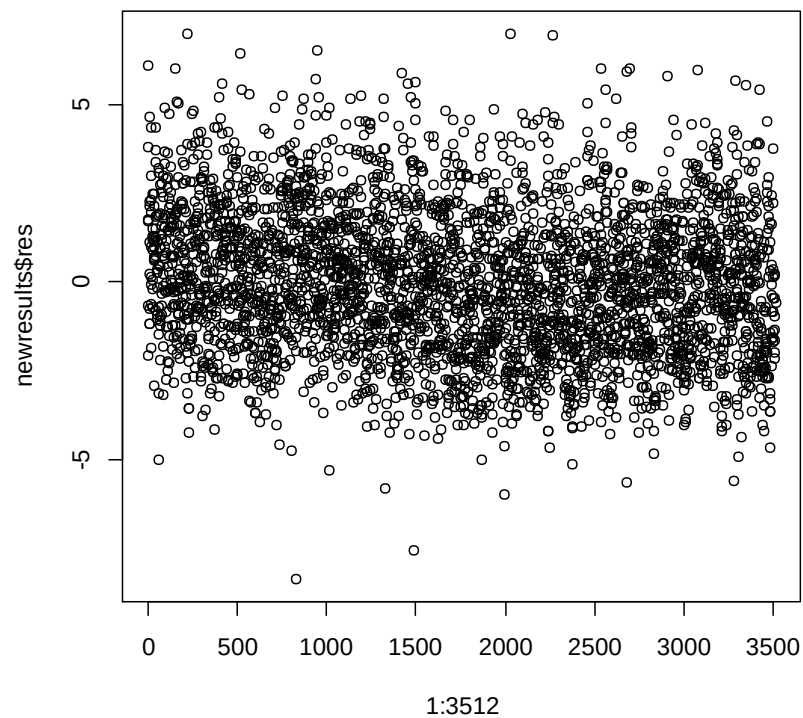
Διάγραμμα 4.24: Διάγραμμα για τον γραφικό έλεγχο της ομοσκεδαστικότητας



> plot (1:3512, newresults\$res)

- Με την εντολή plot() κατασκευάζεται το Διάγραμμα 4.25 για τον έλεγχο της ανεξαρτησίας, η οποία φαίνεται επίσης βελτιωμένη.

Διάγραμμα 4.25: Διάγραμμα για τον γραφικό έλεγχο της ανεξαρτησίας



ΑΠΟΤΕΛΕΣΜΑΤΑ – ΕΡΜΗΝΕΙΑ

Όπως προκύπτει από τα αποτελέσματα της παλινδρόμησης, το μοντέλο για τη μέση συνολική ετήσια καταναλωτική δαπάνη των νοικοκυριών ως προς το εισόδημα παίρνει την εξής μορφή:

$$E(\text{total}|\text{income}) = 1.964 \cdot 10 + 8.163 \cdot 10^{-5} \text{ income}.$$

Συνάγονται, λοιπόν, τα παρακάτω συμπεράσματα:

- Η αναμενόμενη τιμή της μεταβλητής απόκρισης είναι 19.64, όταν $\text{income} = 0$.
- Η αναμενόμενη τιμή της μεταβλητής απόκρισης αυξάνεται κατά 0.00008163 αν το συνολικό ετήσιο εισόδημα αυξηθεί κατά 1€.

Ωστόσο, οι εκτιμήσεις του νέου αυτού μοντέλου δεν είναι ερμηνεύσιμες, αφού έχει πραγματοποιηθεί ο μετασχηματισμός της μεταβλητής απόκρισης.

Επιπρόσθετα, δίνονται τα εκτιμώμενα στοιχεία για κάθε συντελεστή του εν λόγω γραμμικού μοντέλου:

Πίνακας 4.25: Εκτιμήσεις των συντελεστών του γραμμικού μοντέλου

	Εκτίμηση	Τυπικό σφάλμα	Στατιστικό ελέγχου T_1	p – τιμή	95% διάστημα εμπιστοσύνης	
					κάτω όριο	άνω όριο
σταθερά	1.964e+01	5.567e-02	352.70	<2e-16	1.952588e+01	1.974418e+01
income	8.163e-05	1.695e-06	48.15	<2e-16	7.830939e-05	8.495704e-05

- Επειδή οι εκτιμήσεις των παραμέτρων βασίζονται στα συγκεκριμένα δεδομένα, πραγματοποιείται για κάθε συντελεστή ο έλεγχος H_0 , δηλαδή ο συντελεστής να παίρνει τη μηδενική τιμή, με εναλλακτική H_1 να είναι στατιστικά διάφορος του μηδενός. Σε επίπεδο σημαντικότητας 5% όλες οι p – τιμές είναι πολύ μικρές οπότε, πράγματι, αύξηση κατά μία μονάδα στη μεταβλητή income αντίστοιχα σημαίνει και μεταβολή της αναμενόμενης τιμής της μεταβλητής απόκρισης total και η αναμενόμενη τιμή της μεταβλητής απόκρισης total είναι διάφορη του μηδενός όταν $\text{income} = 0$.
- Γεγονός που επαληθεύεται και από τα αντίστοιχα συμμετρικά 95% διαστήματα εμπιστοσύνης, τα οποία δεν περιέχουν την τιμή 0. Επίσης, πραγματοποιείται και το F – test, το οποίο ελέγχει κατά πόσο διαφέρει το εν λόγω γραμμικό μοντέλο από το σταθερό. Στην απλή γραμμική περίπτωση ταυτίζεται με τον έλεγχο για τη μηδενική υπόθεση $H_0 : b_1 = 0$ με εναλλακτική ότι δεν είναι.
- Το τυπικό σφάλμα της παλινδρόμησης $S_{y/x} = 2.045$ είναι σχετικά μεγάλο, γεγονός που σημαίνει ότι υπάρχει αρκετά μεγάλη διασπορά σφαλμάτων στο εν λόγω δείγμα και η προσαρμογή του δεν είναι πολύ καλή από το γραμμικό μοντέλο.
- Ο πολλαπλός συντελεστής προσδιορισμού $R^2 = 0.3978$, ο οποίος εκφράζει το ποσοστό της διασποράς της τ.μ. total που εξηγείται βάσει του μοντέλου

παλινδρόμησης δεν είναι τόσο μεγάλος και άρα, δεν υπάρχει τόσο καλή προσαρμογή της ευθείας. Αντίστοιχα είναι τα συμπεράσματα και παρατηρώντας το διορθωμένο συντελεστή προσδιορισμού $\bar{R} = 0.3976$, ο οποίος χρησιμοποιείται για να αγνοηθεί η εξάρτηση από τον αριθμό των επεξηγηματικών μεταβλητών που εμφανίζεται στον πολλαπλό συντελεστή προσδιορισμού όταν προστίθεται μια νέα επεξηγηματική μεταβλητή στο μοντέλο, όπου και στο πολλαπλό μοντέλο εξηγείται η ερμηνεία του.

4.3.2. Πολλαπλή Γραμμική Παλινδρόμηση

ΔΙΑΤΥΠΩΣΗ ΥΠΟΘΕΣΗΣ

Στη συνέχεια, θα εφαρμοσθεί η Πολλαπλή Γραμμική Παλινδρόμηση προκειμένου να εκτιμηθεί η συνολική ετήσια καταναλωτική δαπάνη των νοικοκυριών συναρτήσει του εισοδήματος, της περιοχής, της κύριας πηγής εισοδήματος, του αριθμού των οικονομικά ενεργών μελών και της κοινωνικο – οικονομικής κατάστασης του προσώπου αναφοράς του νοικοκυριού. Η εξαρτημένη μεταβλητή ή μεταβλητή απόκρισης (ποσοτική) είναι η συνολική ετήσια καταναλωτική δαπάνη (total) και οι επεξηγηματικές μεταβλητές είναι το συνολικό εισόδημα (income), η περιοχή (region), η κύρια πηγή εισοδήματος (source), ο αριθμός των οικονομικά ενεργών μελών (members) και η κοινωνικο – οικονομική κατάσταση του προσώπου αναφοράς (sitnew). Μόνο το εισόδημα είναι ποσοτική μεταβλητή ενώ οι υπόλοιπες είναι κατηγορικές.

Θα δημιουργηθούν, λοιπόν, οι λεγόμενες εικονικές μεταβλητές (ή ψευδομεταβλητές) για τις κατηγορίες των κατηγορικών μεταβλητών, οι οποίες λαμβάνουν τις τιμές 0 και 1. Οπότε ορίζονται οι εξής εικονικές μεταβλητές:

για τη μεταβλητή region:

$$\begin{aligned} \text{region1} &= \begin{cases} 1, \text{Βόρεια Ελλάδα} \\ 0, \text{διαφορετικά} \end{cases} \\ \text{region2} &= \begin{cases} 1, \text{Κεντρική Ελλάδα} \\ 0, \text{διαφορετικά} \end{cases} \\ \text{region4} &= \begin{cases} 1, \text{Νησιά Αιγαίου – Κρήτη} \\ 0, \text{διαφορετικά} \end{cases} \end{aligned}$$

για τη μεταβλητή source:

$$\begin{aligned} \text{source2} &= \begin{cases} 1, \text{εισόδημα από αυτοαπασχόληση} \\ 0, \text{διαφορετικά} \end{cases} \\ \text{source3} &= \begin{cases} 1, \text{εισόδημα από ακίνητη περιουσία} \\ 0, \text{διαφορετικά} \end{cases} \\ \text{source4} &= \begin{cases} 1, \text{συντάξεις} \\ 0, \text{διαφορετικά} \end{cases} \\ \text{source5} &= \begin{cases} 1, \text{επιδόματα ανεργίας} \\ 0, \text{διαφορετικά} \end{cases} \\ \text{source6} &= \begin{cases} 1, \text{άλλα επιδόματα} \\ 0, \text{διαφορετικά} \end{cases} \end{aligned}$$

για τη μεταβλητή members:

$$\begin{aligned} \text{members1} &= \begin{cases} 1, 1 \text{ ενεργό μέλος} \\ 0, \text{διαφορετικά} \end{cases} \\ \text{members2} &= \begin{cases} 1, 2 \text{ ενεργά μέλη} \\ 0, \text{διαφορετικά} \end{cases} \\ \text{members3} &= \begin{cases} 1, 3 \text{ ενεργά μέλη} \\ 0, \text{διαφορετικά} \end{cases} \\ \text{members4} &= \begin{cases} 1, 4 \text{ ενεργά μέλη} \\ 0, \text{διαφορετικά} \end{cases} \\ \text{members5} &= \begin{cases} 1, > 5 \text{ ενεργά μέλη} \\ 0, \text{διαφορετικά} \end{cases} \end{aligned}$$

για τη μεταβλητή sitnew:

$$\begin{aligned} \text{sitnew2} &= \begin{cases} 1, \text{μη χειρώνεκτας εργαζόμενος} \\ 0, \text{διαφορετικά} \end{cases} \\ \text{sitnew3} &= \begin{cases} 1, \text{αυτοαπασχολούμενος} \\ 0, \text{διαφορετικά} \end{cases} \\ \text{sitnew4} &= \begin{cases} 1, \text{άνεργος} \\ 0, \text{διαφορετικά} \end{cases} \\ \text{sitnew5} &= \begin{cases} 1, \text{συνταξιούχος} \\ 0, \text{διαφορετικά} \end{cases} \\ \text{sitnew6} &= \begin{cases} 1, \text{λοιποί μη οικονομικά ενεργοί} \\ 0, \text{διαφορετικά} \end{cases} \end{aligned}$$

$$sitnew7 = \begin{cases} 1, & \text{δεν έχει απαντηθεί} \\ 0, & \text{διαφορετικά} \end{cases}$$

Επομένως, πρέπει να εκτιμηθούν οι συντελεστές του μοντέλου:

$$E[total|income, region, source, members, sitnew] = a + b_1 income + b_2 region1 + \dots + b_{20} sitnew7.$$

Εντολές στην R

```
> region <- factor(region, levels = c(3, 1, 2, 4))
```

- Με την εντολή αυτή αλλάζει η σειρά των κατηγοριών για την εν λόγω μεταβλητή, ώστε να θεωρεί στην παλινδρόμηση ως κατηγορία αναφοράς η περιοχή της Αττικής, για να συγκριθεί με τις υπόλοιπες.

```
> mlm <- lm(total ~ income + region + source + members + sitnew)
```

```
> mlm
```

Call:

```
lm(formula = total ~ income + region + source + members + sitnew)
```

Coefficients:

```
(Intercept)  income  region1  region2  region4  source2
13784.1584    0.6671 -818.7968 -1652.7527 -1712.6415 -6654.4885
source3  source4  source5  source6  members1  members2
-9947.6384 -9633.6978 -1342.5345 -4518.2851 1753.3252 4481.9543
members3  members4  members5  sitnew2  sitnew3  sitnew4
5733.8102 7363.6097 3157.1137 7532.0049 -913.2683 -531.8736
sitnew5  sitnew6  sitnew7
-4278.4867 -2637.4740 -8178.9447
```

- Με την εντολή lm() προσαρμόζεται το πολλαπλό γραμμικό μοντέλο παλινδρόμησης με total τα δεδομένα για τη μεταβλητή απόκρισης και income, region, source, members και sitnew τα δεδομένα για τις επεξηγηματικές μεταβλητές.

```
> summary(mlm)
```

Call:

```
lm(formula = total ~ income + region + source + members + sitnew)
```

Residuals:

```
Min 1Q Median 3Q Max
-63297 -5951 -1291 4130 85984
```

Coefficients:

```
Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.378e+04 1.246e+03 11.062 < 2e-16 ***
income      6.671e-01 1.072e-02 62.247 < 2e-16 ***
region1     -8.188e+02 4.637e+02 -1.766 0.077540 .
region2     -1.653e+03 5.169e+02 -3.198 0.001398 **
region4     -1.713e+03 6.834e+02 -2.506 0.012259 *
source2     -6.654e+03 9.400e+02 -7.079 1.75e-12 ***
source3     -9.948e+03 1.403e+03 -7.090 1.61e-12 ***
source4     -9.634e+03 8.184e+02 -11.771 < 2e-16 ***
source5     -1.343e+03 2.625e+03 -0.511 0.609068
source6     -4.518e+03 1.248e+03 -3.620 0.000299 ***
members1    1.753e+03 1.119e+03 1.567 0.117245
members2    4.482e+03 1.202e+03 3.729 0.000195 ***
```

```

members3  5.734e+03 1.433e+03 4.002 6.41e-05 ***
members4  7.364e+03 1.854e+03 3.972 7.27e-05 ***
members5  3.157e+03 5.098e+03 0.619 0.535806
sitnew2   7.532e+03 7.694e+02 9.790 < 2e-16 ***
sitnew3   -9.133e+02 9.935e+02 -0.919 0.358013
sitnew4   -5.319e+02 1.195e+03 -0.445 0.656412
sitnew5   -4.278e+03 9.083e+02 -4.710 2.57e-06 ***
sitnew6   -2.637e+03 1.129e+03 -2.337 0.019490 *
sitnew7   -8.179e+03 6.439e+03 -1.270 0.204074

```

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11020 on 3491 degrees of freedom

Multiple R-squared: 0.6543, Adjusted R-squared: 0.6524

F-statistic: 330.4 on 20 and 3491 DF, p-value: < 2.2e-16

- Με την `summary()` δίνονται οι εκτιμήσεις για τις παραμέτρους του πολλαπλού γραμμικού μοντέλου.

```

> confint(mlm)
      2.5 %      97.5 %
(Intercept) 11340.97155 16227.3453000
income      0.64608 0.6881038
region1     -1728.01374 90.4202186
region2     -2666.17214 -639.3332572
region4     -3052.62440 -372.6586786
source2     -8497.52142 -4811.4556531
source3     -12698.46539 -7196.8113805
source4     -11238.37124 -8029.0244108
source5     -6489.13871 3804.0696561
source6     -6965.39926 -2071.1710367
members1     -440.68490 3947.3352938
members2     2125.54024 6838.3682701
members3     2924.88940 8542.7309173
members4     3728.91743 10998.3019652
members5     -6839.14553 13153.3730191
sitnew2      6023.52732 9040.4824916
sitnew3     -2861.09076 1034.5541580
sitnew4     -2875.74995 1812.0027191
sitnew5     -6059.32942 -2497.6440203
sitnew6     -4850.11182 -424.8361559
sitnew7     -20803.06017 4445.1708288

```

- Με την εντολή `confint()` υπολογίζονται συμμετρικά 95% διαστήματα εμπιστοσύνης για τις παραμέτρους του εν λόγω μοντέλου.

```

> model.matrix(mlm)
(Intercept) income region1 region2 region4 source2 source3 source4 source5
1      1 23112.60    0    0    0    0    0    0    0
2      1 20319.96    0    0    0    0    0    1    0
3      1 15484.56    0    0    0    0    0    0    0
      source6 members1 members2 members3 members4 members5 sitnew2 sitnew3 sitnew4
1      0      0      1      0      0      0      1      0      0
2      0      0      0      0      0      0      0      0      0
3      0      1      0      0      0      0      1      0      0
      sitnew5 sitnew6 sitnew7

```

```

1  0  0  0
2  0  1  0
3  0  0  0
...
(Intercept) income region1 region2 region4 source2 source3 source4
3509      140241.64    0    0    1    0    0    1
3510      134800.00    0    0    1    1    0    0
3511      120360.04    0    0    1    0    0    1
3512      129789.76    0    0    1    1    0    0
      source5 source6 members1 members2 members3 members4 members5 sitnew2
3509    0    0    1    0    0    0    0    0
3510    0    0    0    1    0    0    0    0
3511    0    0    1    0    0    0    0    0
3512    0    0    1    0    0    0    0    0
      sitnew3 sitnew4 sitnew5 sitnew6 sitnew7
3509    0    0    1    0    0
3510    1    0    0    0    0
3511    0    0    1    0    0
3512    1    0    0    0    0

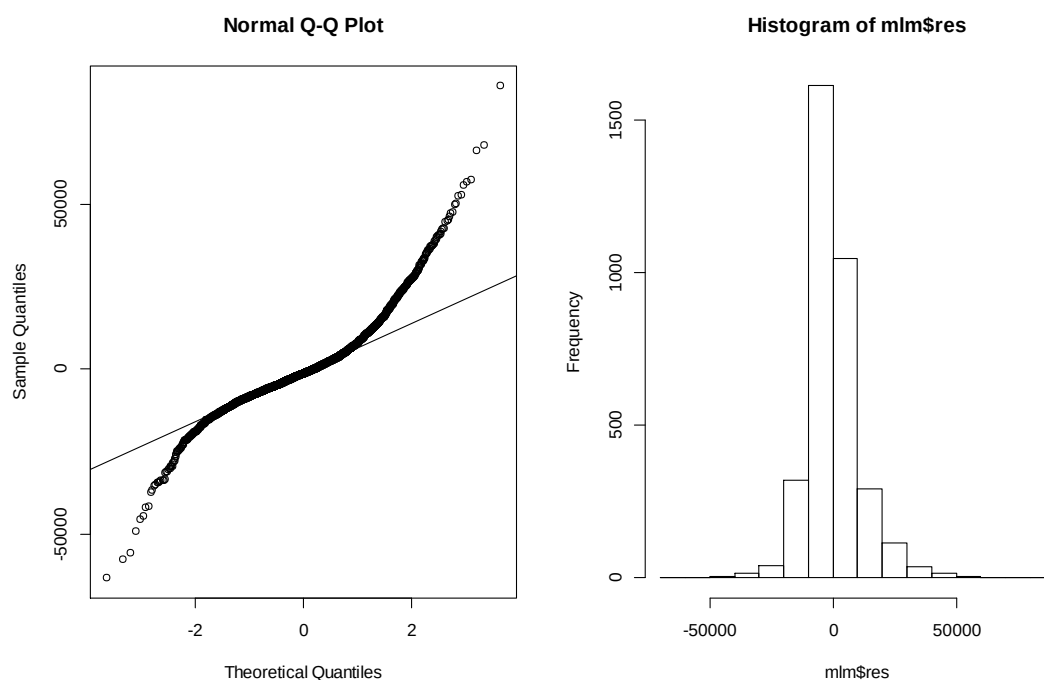
```

- Με την εντολή `model.matrix()` εμφανίζονται οι εικονικές μεταβλητές που έχουν δημιουργηθεί για το εν λόγω πολλαπλό γραμμικό μοντέλο.
- ```

> par (mfrow = c (1, 2))
> qqnorm (mlm$res)
> qqline (mlm$res)
> hist(mlm$res)

```
- Με τις εντολές `qqnorm()`, `qqline()` και `hist()` κατασκευάζεται το Διάγραμμα 4.26 για τον έλεγχο της κανονικότητας των σφαλμάτων.

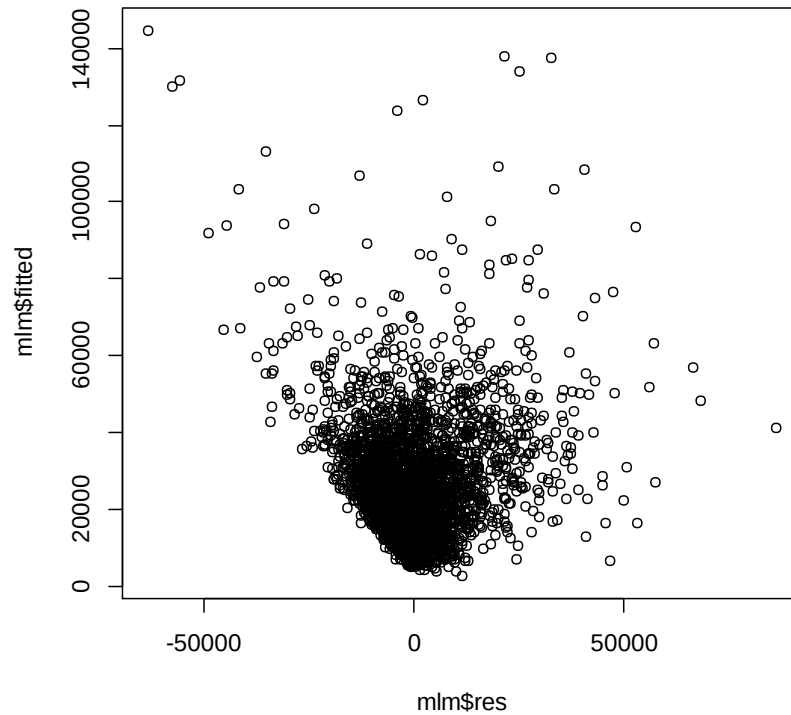
**Διάγραμμα 4.26:** QQ – plot και Ιστόγραμμα υπολοίπων



> plot (mlm\$res, mlm\$fitted)

- Με την εντολή plot() κατασκευάζεται το Διάγραμμα 4.27 για τον έλεγχο της ομοσκεδαστικότητας.

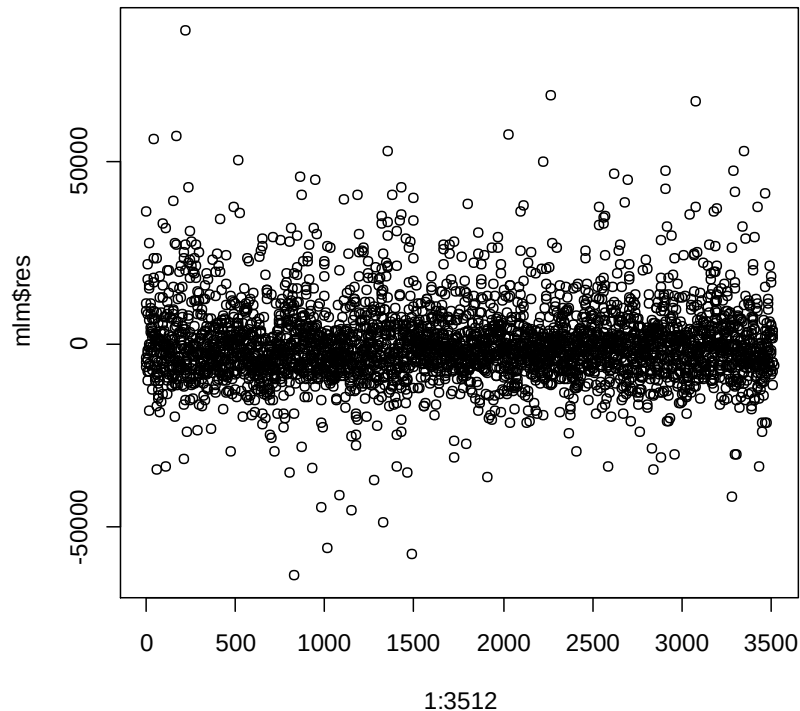
**Διάγραμμα 4.27:** Διάγραμμα για τον γραφικό έλεγχο της ομοσκεδαστικότητας



> plot (1:3512, mlm\$res)

- Με την εντολή plot() κατασκευάζεται το Διάγραμμα 4.28 για τον έλεγχο της ανεξαρτησίας.

**Διάγραμμα 4.28:** Διάγραμμα για τον γραφικό έλεγχο της ανεξαρτησίας



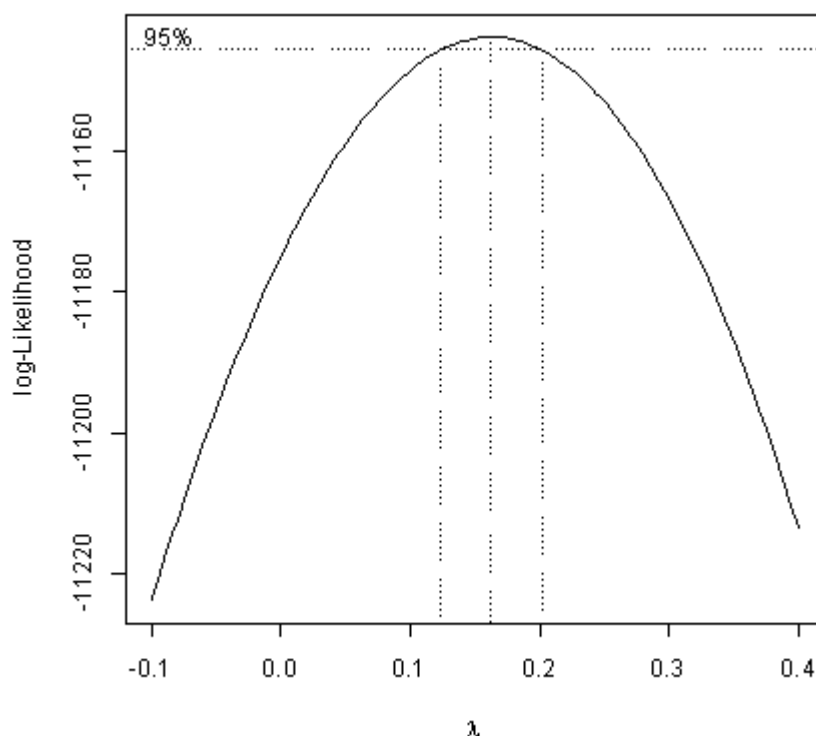
Η υπόθεση της κανονικότητας ελέγχεται με το Διάγραμμα 4.26 από το οποίο και προκύπτει να μην είναι λογική η εν λόγω υπόθεση για τα δεδομένα του δείγματος καθώς παρατηρούνται μεγάλες αποκλίσεις, δηλαδή τα υπόλοιπα δεν ακολουθούν την Κανονική κατανομή. Επίσης, από το ιστόγραμμα συχνοτήτων των υπολοίπων φαίνεται ότι δεν υπάρχει συμμετρία στη διασπορά τους γύρω από το 0. Η υπόθεση της ομοσκεδαστικότητας ελέγχεται με το Διάγραμμα 4.27 από το οποίο φαίνονται τα υπόλοιπα να κατανέμονται με σχετικά τυχαίο ή ομοσκεδαστικό τρόπο με κάποιες μικρές αποκλίσεις. Η υπόθεση της ανεξαρτησίας των σφαλμάτων ελέγχεται με το Διάγραμμα 4.28 στο οποίο φαίνεται ότι η σειρά των δεδομένων δεν εμφανίζει κάποιου είδους εξάρτηση και κατανέμονται τυχαία και άρα, η υπόθεση της ανεξαρτησίας επιβεβαιώνεται.

```
> bc <- boxcox (total ~ income + region + source + members + sitnew, lambda = seq (-0.1, 0.4, 0.01))
```

- Επειδή η υπόθεση της κανονικότητας δε φαίνεται λογική και παρατηρείται, επίσης, ένα μικρό πρόβλημα στην ομοσκεδαστικότητα, θα πραγματοποιηθεί μετασχηματισμός της μεταβλητής απόκρισης με τη χρήση του μετασχηματισμού Box – Cox, ο οποίος υλοποιείται στην R με την εντολή `boxcox()`. Τα δεδομένα πρέπει να είναι θετικά, τα οποία στην εν λόγω περίπτωση είναι. Η ποσότητα που προκύπτει εισάγεται στο μοντέλο της παλινδρόμησης έναντι των επεξηγηματικών μεταβλητών και του υπολογισμού της λογαριθμικής συνάρτησης πιθανότητας. Από το Διάγραμμα 4.29 προκύπτει ότι η βέλτιστη τιμή του  $\lambda$  είναι ίση με 0.14, οπότε και στη συνέχεια πραγματοποιείται ο μετασχηματισμός της μεταβλητής απόκρισης.



**Διάγραμμα 4.29:** Μετασχηματισμός Box – Cox



```
> newtotal2 <- (total^0.16-1)/0.16
> newmlm <- lm (newtotal2 ~ income + region + source + members + sitnew)
> newmlm
```

Call:

```
lm(formula = newtotal2 ~ income + region + source + members +
 sitnew)
```

Coefficients:

```
(Intercept) income region1 region2 region4 source2
 2.231e+01 9.965e-05 -3.918e-01 -4.765e-01 -4.633e-01 -9.373e-01
 source3 source4 source5 source6 members1 members2
-1.484e+00 -1.889e+00 -3.690e-02 -7.863e-01 9.299e-01 1.670e+00
 members3 members4 members5 sitnew2 sitnew3 sitnew4
 1.986e+00 2.194e+00 1.573e+00 1.411e+00 9.838e-04 2.541e-01
 sitnew5 sitnew6 sitnew7
-6.482e-01 -1.211e-01 -1.694e+00
```

- Με την εντολή `lm()` προσαρμόζεται εκ νέου το γραμμικό μοντέλο παλινδρόμησης με `newtotal` τα δεδομένα για τη μεταβλητή απόκρισης. Ωστόσο, οι εκτιμήσεις του νέου αυτού μοντέλου δεν είναι ερμηνεύσιμες, αφού έχει πραγματοποιηθεί ο μετασχηματισμός της μεταβλητής απόκρισης.

```
> summary(newmlm)
```

Call:

```
lm(formula = newtotal2 ~ income + region + source + members +
 sitnew)
```

Residuals:

```
 Min 1Q Median 3Q Max
-11.1586 -1.3323 -0.0752 1.2726 8.1655
```

Coefficients:

|             | Estimate   | Std. Error | t value | Pr(> t )     |
|-------------|------------|------------|---------|--------------|
| (Intercept) | 2.231e+01  | 2.246e-01  | 99.370  | < 2e-16 ***  |
| income      | 9.965e-05  | 1.931e-06  | 51.598  | < 2e-16 ***  |
| region1     | -3.918e-01 | 8.357e-02  | -4.688  | 2.86e-06 *** |
| region2     | -4.765e-01 | 9.315e-02  | -5.116  | 3.29e-07 *** |
| region4     | -4.633e-01 | 1.232e-01  | -3.762  | 0.000171 *** |
| source2     | -9.373e-01 | 1.694e-01  | -5.533  | 3.37e-08 *** |
| source3     | -1.484e+00 | 2.528e-01  | -5.869  | 4.81e-09 *** |
| source4     | -1.889e+00 | 1.475e-01  | -12.808 | < 2e-16 ***  |
| source5     | -3.690e-02 | 4.730e-01  | -0.078  | 0.937828     |
| source6     | -7.863e-01 | 2.249e-01  | -3.496  | 0.000478 *** |
| members1    | 9.299e-01  | 2.017e-01  | 4.611   | 4.15e-06 *** |
| members2    | 1.670e+00  | 2.166e-01  | 7.710   | 1.64e-14 *** |
| members3    | 1.986e+00  | 2.582e-01  | 7.693   | 1.85e-14 *** |
| members4    | 2.194e+00  | 3.341e-01  | 6.568   | 5.87e-11 *** |
| members5    | 1.573e+00  | 9.188e-01  | 1.712   | 0.087005 .   |
| sitnew2     | 1.411e+00  | 1.386e-01  | 10.176  | < 2e-16 ***  |
| sitnew3     | 9.838e-04  | 1.790e-01  | 0.005   | 0.995616     |
| sitnew4     | 2.541e-01  | 2.154e-01  | 1.179   | 0.238322     |
| sitnew5     | -6.482e-01 | 1.637e-01  | -3.960  | 7.63e-05 *** |
| sitnew6     | -1.211e-01 | 2.034e-01  | -0.595  | 0.551697     |
| sitnew7     | -1.694e+00 | 1.160e+00  | -1.460  | 0.144490     |

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.986 on 3491 degrees of freedom

Multiple R-squared: 0.6225, Adjusted R-squared: 0.6203

F-statistic: 287.8 on 20 and 3491 DF, p-value: < 2.2e-16

- Με την εντολή `summary()` δίνονται οι εκτιμήσεις για τις παραμέτρους του πολλαπλού γραμμικού μοντέλου.

> `confint(newmlm)`

|             | 2.5 %         | 97.5 %        |
|-------------|---------------|---------------|
| (Intercept) | 2.187435e+01  | 22.7549187483 |
| income      | 9.586318e-05  | 0.0001034362  |
| region1     | -5.556434e-01 | -0.2279467856 |
| region2     | -6.591467e-01 | -0.2938938039 |
| region4     | -7.048250e-01 | -0.2218733832 |
| source2     | -1.269463e+00 | -0.6052040872 |
| source3     | -1.979502e+00 | -0.9880592017 |
| source4     | -2.178233e+00 | -1.5998821904 |
| source5     | -9.643596e-01 | 0.8905604594  |
| source6     | -1.227338e+00 | -0.3453581817 |
| members1    | 5.345020e-01  | 1.3252589949  |
| members2    | 1.245120e+00  | 2.0944100852  |
| members3    | 1.480024e+00  | 2.4924043535  |
| members4    | 1.539067e+00  | 2.8490692292  |
| members5    | -2.285453e-01 | 3.3742694921  |
| sitnew2     | 1.139084e+00  | 1.6827637482  |
| sitnew3     | -3.500297e-01 | 0.3519972523  |
| sitnew4     | -1.683063e-01 | 0.6764649729  |
| sitnew5     | -9.691626e-01 | -0.3273178923 |
| sitnew6     | -5.197966e-01 | 0.2776741597  |

```
sitnew7 -3.968568e+00 0.5813689094
```

- Με την εντολή `confint()` υπολογίζονται συμμετρικά 95% διαστήματα εμπιστοσύνης για τις παραμέτρους του εν λόγω μοντέλου.

```
> model.matrix(newmlm)
```

```
(Intercept) income region1 region2 region4 source2 source3 source4 source5
1 1 23112.60 0 0 0 0 0 0 0
2 1 20319.96 0 0 0 0 0 1 0
3 1 15484.56 0 0 0 0 0 0 0
source6 members1 members2 members3 members4 members5 sitnew2 sitnew3 sitnew4
1 0 0 1 0 0 0 1 0 0
2 0 0 0 0 0 0 0 0 0
3 0 1 0 0 0 0 1 0 0
sitnew5 sitnew6 sitnew7
1 0 0 0
2 0 1 0
3 0 0 0
...
(Intercept) income region1 region2 region4 source2 source3 source4
3510 1 34800.00 0 0 1 1 0 0
3511 1 20360.04 0 0 1 0 0 1
3512 1 29789.76 0 0 1 1 0 0
source5 source6 members1 members2 members3 members4 members5 sitnew2
3510 0 0 0 1 0 0 0 0
3511 0 0 1 0 0 0 0 0
3512 0 0 1 0 0 0 0 0
sitnew3 sitnew4 sitnew5 sitnew6 sitnew7
3510 1 0 0 0 0
3511 0 0 1 0 0
3512 1 0 0 0 0
```

- Με την εντολή `model.matrix()` εμφανίζονται οι εικονικές μεταβλητές που έχουν δημιουργηθεί για το εν λόγω πολλαπλό γραμμικό μοντέλο.

```
> par (mfrow = c (1, 2))
```

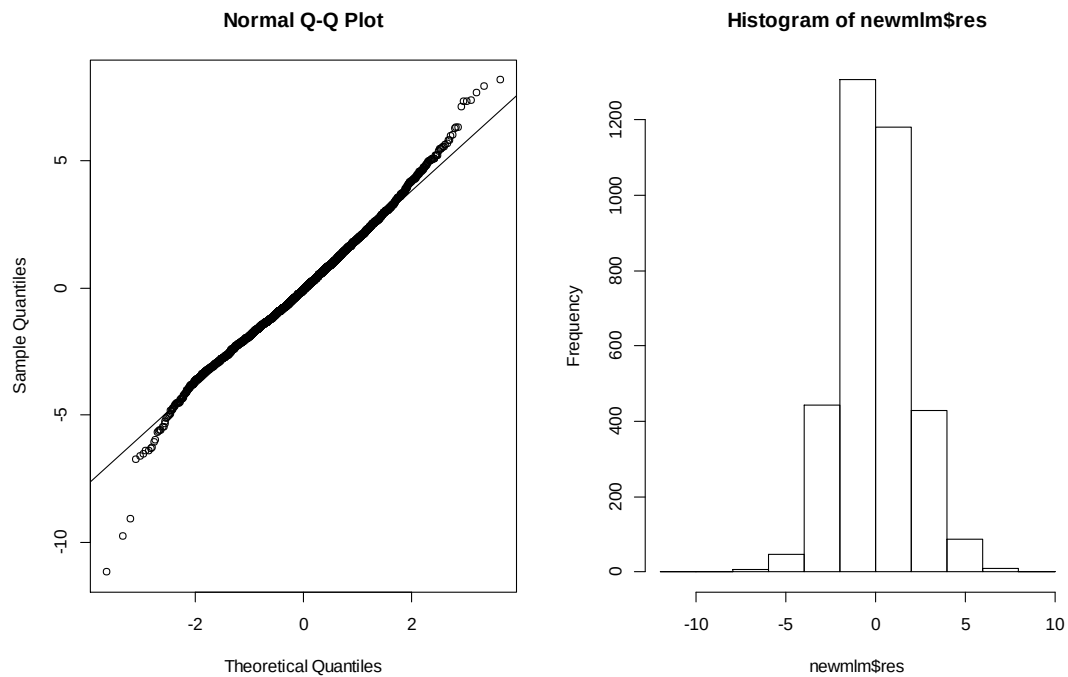
```
> qqnorm (newmlm$res)
```

```
> qqline(newmlm$res)
```

```
> hist(newmlm$res)
```

- Με τις εντολές `qqnorm()`, `qqline()` και `hist()` κατασκευάζεται το Διάγραμμα 4.30 για τον έλεγχο της κανονικότητας των δεδομένων, η οποία φαίνεται να ικανοποιείται καλύτερα στα νέα δεδομένα με αποκλίσεις στις ουρές.

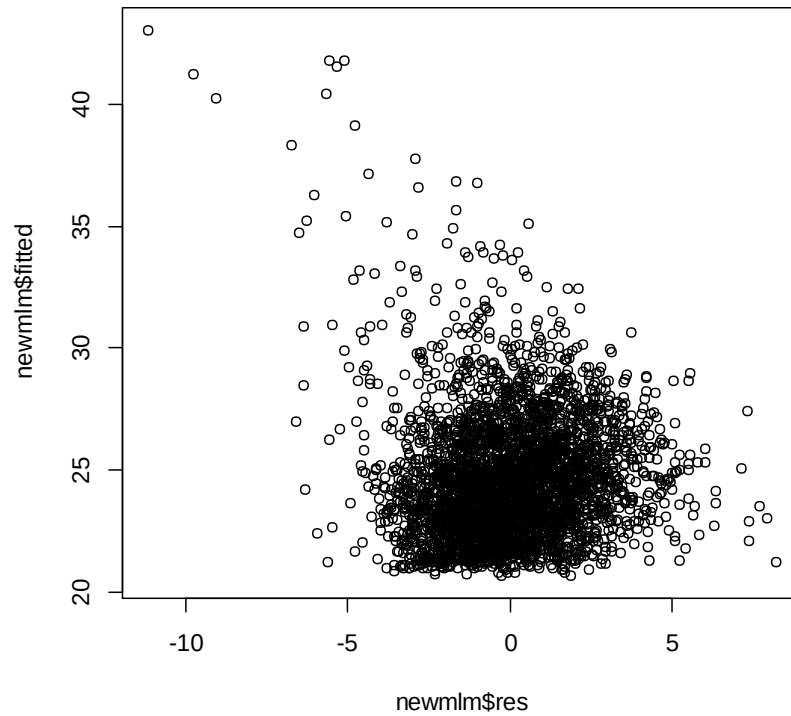
**Διάγραμμα 4.30:** QQ – plot και Ιστόγραμμα υπολοίπων



> plot (newmlm\$res, newmlm\$fitted)

- Με την εντολή plot() κατασκευάζεται το Διάγραμμα 4.31 για τον έλεγχο της ομοσκεδαστικότητας, η οποία φαίνεται να ικανοποιείται καλύτερα σε σχέση με το προηγούμενο γραμμικό μοντέλο παρ' ότι και πάλι παρατηρούνται κάποιες αποκλίσεις γεγονός που οφείλεται στην ύπαρξη κάποιων ακραίων παρατηρήσεων του δείγματος.

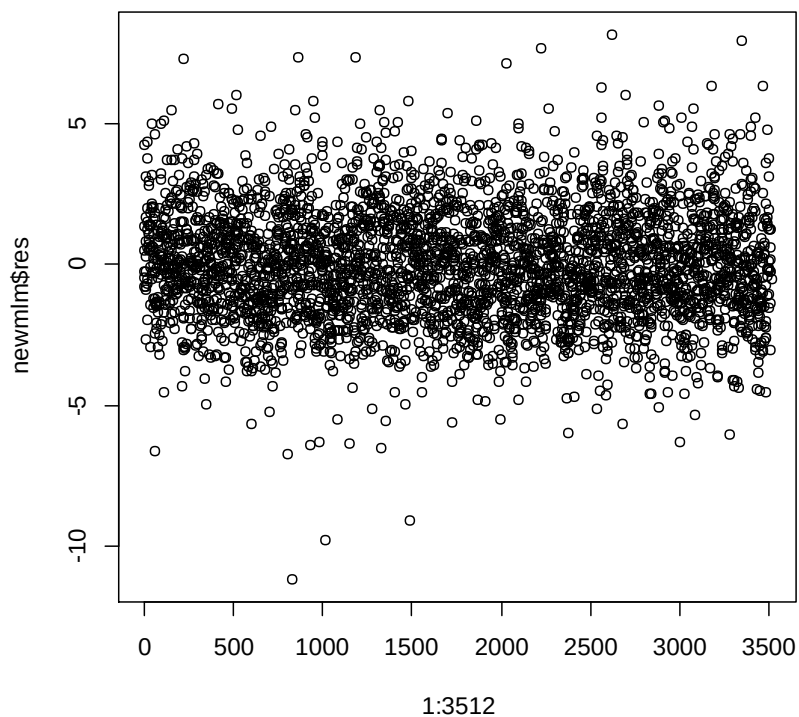
**Διάγραμμα 4.31:** Διάγραμμα για τον γραφικό έλεγχο της ομοσκεδαστικότητας



> plot (1:3512, newmlm\$res)

- Με την εντολή plot() κατασκευάζεται το Διάγραμμα 4.32 για τον έλεγχο της ανεξαρτησίας, η οποία φαίνεται επίσης βελτιωμένη μετά τον μετασχηματισμό της μεταβλητής απόκρισης..

**Διάγραμμα 4.32:** Διάγραμμα για τον γραφικό έλεγχο της ανεξαρτησίας



## ΑΠΟΤΕΛΕΣΜΑΤΑ – ΕΡΜΗΝΕΙΑ

Όπως προκύπτει από τα αποτελέσματα της παλινδρόμησης, το μοντέλο για τη μέση συνολική ετήσια καταναλωτική δαπάνη των νοικοκυριών παίρνει την εξής μορφή:

$$\begin{aligned} E[total|income, region, source, members, sitnes] &= 2.231 \cdot 10 + 9.965 \cdot 10^{-5}income - 3.918 \cdot 10^{-1}region1 - 4.765 \\ &\cdot 10^{-1}region2 - 4.633 \cdot 10^{-1}region4 - 9.373 \cdot 10^{-1}source2 \\ &- 1.484source3 - 1.889source4 - 3.690 \cdot 10^{-2}source5 - 7.863 \\ &\cdot 10^{-1}source6 + 9.299 \cdot 10^{-1}members1 + 1.670members2 \\ &+ 1.986members3 + 2.194members4 + 1.573members5 \\ &+ 1.411sitnew2 + 9.838 \cdot 10^{-4}sitnew3 + 2.541 \cdot 10^{-1}sitnew4 - 6.482 \\ &\cdot 10^{-1}sitnew5 - 1.211 \cdot 10^{-1}sitnew6 - 1.694sitnew7 \end{aligned}$$

Συνάγονται, λοιπόν, τα παρακάτω συμπεράσματα:

- Η αναμενόμενη τιμή της μεταβλητής απόκρισης είναι 22.31 με όλες τις επεξηγηματικές μεταβλητές να είναι μηδενικές.
- Η αναμενόμενη τιμή της μεταβλητής απόκρισης αυξάνεται κατά 0.00009965 αν το συνολικό ετήσιο εισόδημα αυξηθεί κατά 1€.
- Η αναμενόμενη τιμή της μεταβλητής απόκρισης στην Αττική μειώνεται κατά 0.3918 για ένα νοικοκυριό με τα ίδια ακριβώς χαρακτηριστικά αν βρίσκεται στη Βόρεια Ελλάδα, μειώνεται κατά 0.4765 αν βρίσκεται στην Κεντρική Ελλάδα και μειώνεται κατά 0.4633 αν βρίσκεται στα Νησιά Αιγαίου – Κρήτη.
- Η αναμενόμενη τιμή της μεταβλητής απόκρισης με κύρια πηγή εισοδήματος το μισθό μειώνεται κατά 0.9373 αν έχει ως κύρια πηγή εισοδήματος την αυτοαπασχόληση, μειώνεται κατά 1.484 αν έχει ως κύρια πηγή εισοδήματος την ακίνητη περιουσία, μειώνεται κατά 1.889 αν έχει ως κύρια πηγή εισοδήματος τη σύνταξη, μειώνεται κατά 0.0369 αν έχει ως κύρια πηγή εισοδήματος τα επιδόματα ανεργίας και μειώνεται κατά 0.7863 αν έχει ως κύρια πηγή εισοδήματος άλλα επιδόματα δεδομένου ότι όλα τα υπόλοιπα χαρακτηριστικά του δεν μεταβάλλονται.
- Η αναμενόμενη τιμή της μεταβλητής απόκρισης με κανένα οικονομικά ενεργό μέλος είναι μικρότερη κατά 0.9299 από ότι για ένα νοικοκυριό με 1 οικονομικά ενεργό μέλος, είναι μικρότερη κατά 1.670 από ότι για ένα νοικοκυριό με 2 οικονομικά ενεργά μέλη, είναι μικρότερη κατά 1.986 από ότι για ένα νοικοκυριό με 3 οικονομικά ενεργά μέλη, είναι μικρότερη κατά 2.194 από ότι για ένα νοικοκυριό με 4 οικονομικά ενεργά μέλη και είναι μικρότερη κατά 1.573 από ότι για ένα νοικοκυριό με 5 ή περισσότερα οικονομικά ενεργά μέλη με σταθερά όλα τα υπόλοιπα χαρακτηριστικά.
- Η αναμενόμενη τιμή της μεταβλητής απόκρισης με το πρόσωπο αναφοράς να είναι χειρώνακτας εργαζόμενος είναι μικρότερη κατά 1.411 σε σχέση με το νοικοκυριό με πρόσωπο αναφοράς μη χειρώνακτα εργαζόμενο, είναι μικρότερη κατά 0.0009838 σε σχέση με το νοικοκυριό με πρόσωπο αναφοράς αυτοαπασχολούμενο, γεωργό ή κτηνοτρόφο, είναι μικρότερη κατά 0.2541 σε σχέση με το νοικοκυριό με πρόσωπο αναφοράς άνεργο, είναι μεγαλύτερη κατά 0.6482 σε σχέση με το νοικοκυριό με πρόσωπο αναφοράς συνταξιούχο, είναι μεγαλύτερη κατά 0.1211 σε σχέση με το νοικοκυριό με πρόσωπο αναφοράς που ανήκει σε άλλη κατηγορία μη οικονομικά ενεργών

ατόμων και είναι μεγαλύτερη κατά 1.694 σε σχέση με τα νοικοκυριά που δεν έχουν απαντήσει στο εν λόγω ερώτημα.

Ωστόσο, οι εκτιμήσεις του νέου αυτού μοντέλου δεν είναι ερμηνεύσιμες, αφού έχει πραγματοποιηθεί ο μετασχηματισμός της μεταβλητής απόκρισης.

**Πίνακας 4.26:** Εκτιμήσεις των συντελεστών του πολλαπλού γραμμικού μοντέλου

|          | Εκτίμηση   | Τυπικό σφάλμα | Στατιστικό ελέγχου $T_1$ | $p$ – τιμή | 95% διάστημα εμπιστοσύνης |               |
|----------|------------|---------------|--------------------------|------------|---------------------------|---------------|
|          |            |               |                          |            | κάτω όριο                 | άνω όριο      |
| σταθερά  | 2.231e+01  | 2.246e-01     | 99.370                   | < 2e-16    | 2.187435e+01              | 22.7549187483 |
| income   | 9.965e-05  | 1.931e-06     | 51.598                   | < 2e-16    | 9.586318e-05              | 0.0001034362  |
| region1  | -3.918e-01 | 8.357e-02     | -4.688                   | 2.86e-06   | -5.556434e-01             | -0.2279467856 |
| region2  | -4.765e-01 | 9.315e-02     | -5.116                   | 3.29e-07   | -6.591467e-01             | -0.2938938039 |
| region4  | -4.633e-01 | 1.232e-01     | -3.762                   | 0.000171   | -7.048250e-01             | -0.2218733832 |
| source2  | -9.373e-01 | 1.694e-01     | -5.533                   | 3.37e-08   | -1.269463e+00             | -0.6052040872 |
| source3  | -1.484e+00 | 2.528e-01     | -5.869                   | 4.81e-09   | -1.979502e+00             | -0.9880592017 |
| source4  | -1.889e+00 | 1.475e-01     | -12.808                  | < 2e-16    | -2.178233e+00             | -1.5998821904 |
| source5  | -3.690e-02 | 4.730e-01     | -0.078                   | 0.937828   | -9.643596e-01             | 0.8905604594  |
| source6  | -7.863e-01 | 2.249e-01     | -3.496                   | 0.000478   | -1.227338e+00             | -0.3453581817 |
| members1 | 9.299e-01  | 2.017e-01     | 4.611                    | 4.15e-06   | 5.345020e-01              | 1.3252589949  |
| members2 | 1.670e+00  | 2.166e-01     | 7.710                    | 1.64e-14   | 1.245120e+00              | 2.0944100852  |
| members3 | 1.986e+00  | 2.582e-01     | 7.693                    | 1.85e-14   | 1.480024e+00              | 2.4924043535  |
| members4 | 2.194e+00  | 3.341e-01     | 6.568                    | 5.87e-11   | 1.539067e+00              | 2.8490692292  |
| members5 | 1.573e+00  | 9.188e-01     | 1.712                    | 0.087005   | -2.285453e-01             | 3.3742694921  |
| sitnew2  | 1.411e+00  | 1.386e-01     | 10.176                   | < 2e-16    | 1.139084e+00              | 1.6827637482  |
| sitnew3  | 9.838e-04  | 1.790e-01     | 0.005                    | 0.995616   | -3.500297e-01             | 0.3519972523  |
| sitnew4  | 2.541e-01  | 2.154e-01     | 1.179                    | 0.238322   | -1.683063e-01             | 0.6764649729  |
| sitnew5  | -6.482e-01 | 1.637e-01     | -3.960                   | 7.63e-05   | -9.691626e-01             | -0.3273178923 |
| sitnew6  | -1.211e-01 | 2.034e-01     | -0.595                   | 0.551697   | -5.197966e-01             | 0.2776741597  |
| sitnew7  | -1.694e+00 | 1.160e+00     | -1.460                   | 0.144490   | -3.968568e+00             | 0.5813689094  |

- Επειδή οι εκτιμήσεις των παραμέτρων βασίζονται στα συγκεκριμένα δεδομένα, πραγματοποιείται για κάθε συντελεστή ο έλεγχος  $H_0$ , δηλαδή ο συντελεστής να παίρνει τη μηδενική τιμή, με εναλλακτική  $H_1$  να είναι στατιστικά διάφορος του μηδενός. Σε επίπεδο σημαντικότητας 5% για τη σταθερά και τις μεταβλητές income, region1, region2, region4, source2, source3, source4, source6, members1, members2, members3, members4, sitnew2 και sitnew5 οι  $p$  – τιμές είναι < 0.05 οπότε, πράγματι, μεταβολή στην τιμή των μεταβλητών αυτών σημαίνει και μεταβολή της αναμενόμενης τιμής της μεταβλητής απόκρισης. Για τις υπόλοιπες μεταβλητές σε επίπεδο σημαντικότητας 5% δεν υπάρχουν σοβαρές ενδείξεις για να απορριφθεί η μηδενική υπόθεση.
- Γεγονός που επαληθεύεται και από τα αντίστοιχα συμμετρικά 95% διαστήματα εμπιστοσύνης, τα οποία δεν περιέχουν την τιμή 0. Επίσης, πραγματοποιείται και το F – test, το οποίο ελέγχει κατά πόσο διαφέρει το εν λόγω γραμμικό μοντέλο από το σταθερό. Η  $p$  – τιμή < 2.2e-16 οπότε υπάρχουν σοβαρές ενδείξεις για να απορριφθεί η μηδενική υπόθεση, ότι δηλαδή όλοι οι συντελεστές είναι ίσοι με μηδέν.
- Το τυπικό σφάλμα της παλινδρόμησης  $S_{y/x} = 1.986$  είναι σχετικά μεγάλο, γεγονός που σημαίνει ότι υπάρχει αρκετά μεγάλη διασπορά σφαλμάτων στο

εν λόγω δείγμα και η προσαρμογή δεν είναι πολύ καλή από το πολλαπλό γραμμικό μοντέλο.

- Ο πολλαπλός συντελεστής προσδιορισμού  $R^2 = 0.6225$ , ο οποίος εκφράζει το ποσοστό της διασποράς της τ.μ. total που εξηγείται βάσει του μοντέλου παλινδρόμησης δεν είναι αρκετά μεγάλος και άρα, δεν υπάρχει καλή προσαρμογή της ευθείας. Αντίστοιχα είναι τα συμπεράσματα και παρατηρώντας το διορθωμένο συντελεστή προσδιορισμού  $\bar{R} = 0.6203$ , ο οποίος χρησιμοποιείται για να αγνοηθεί η εξάρτηση από τον αριθμό των επεξηγηματικών μεταβλητών που εμφανίζεται στον πολλαπλό συντελεστή προσδιορισμού όταν προστίθεται μια νέα επεξηγηματική μεταβλητή στο μοντέλο, όπου και στο πολλαπλό μοντέλο εξηγείται η ερμηνεία του.



## **ΚΕΦΑΛΑΙΟ 5**

### **ΑΝΑΛΥΣΗ ΚΥΡΙΩΝ ΣΥΝΙΣΤΩΣΩΝ**

#### **5.1 Εισαγωγή**

Η ανάλυση πολυμεταβλητών δεδομένων καθίσταται εξαιρετικά δύσκολη όταν το πλήθος των μεταβλητών είναι μεγάλο. Ένα ακόμη πρόβλημα παρατηρείται όταν υπάρχει υψηλή συσχέτιση μεταξύ τους.

Η μέθοδος των Κυρίων Συνιστωσών (Principal Components Analysis) είναι μία πολυμεταβλητή στατιστική τεχνική ανάλυσης δεδομένων με σκοπό τη δημιουργία καινούριων μεταβλητών, οι οποίες είναι γραμμικοί συνδυασμοί των αρχικών μεταβλητών (έστω  $p$ ), έτσι ώστε να είναι ασυσχέτιστες μεταξύ τους και να περιέχουν όσο το δυνατόν μεγαλύτερο μέρος της διακύμανσης των αρχικών μεταβλητών. Οι νέες αυτές μεταβλητές που κατασκευάζονται ονομάζονται Κύριες Συνιστώσες. Παρ' όλο που απατούνται  $p$  μεταβλητές για να ερμηνευτεί η συνολική μεταβλητότητα, η περισσότερη από αυτή τη μεταβλητότητα μπορεί να ερμηνευτεί από ένα μικρότερο αριθμό,  $k < p$ , κυρίων συνιστωσών.

Πιο αναλυτικά, το βασικό πλεονέκτημα της εν λόγω μεθόδου είναι ότι από ένα σύνολο συσχετισμένων μεταβλητών προκύπτει ένα σύνολο ασυσχέτιστων μεταβλητών, το οποίο και είναι χρήσιμο για αρκετές στατιστικές μεθόδους. Για παράδειγμα, στο πρόβλημα της πολυσυγγραμμικότητας στη γραμμική παλινδρόμηση, αν χρησιμοποιηθούν οι συσχετισμένες μεταβλητές οι εκτιμήσεις που θα προκύψουν δεν θα είναι συνεπείς ενώ αντίθετα, με χρήση των ασυσχέτιστων μεταβλητών αντιμετωπίζεται το πρόβλημα αυτό. Επιπρόσθετα, οι κύριες συνιστώσες που δημιουργούνται μπορούν να ερμηνεύσουν το μεγαλύτερο ποσοστό της διακύμανσης, γεγονός που σημαίνει ότι έχει μειωθεί ο αριθμός των αρχικών μεταβλητών με κόστος, βέβαια, την απώλεια ενός σχετικά μικρού ποσοστού της συνολικής μεταβλητότητας. Αυτό είναι ιδιαίτερα σημαντικό στην περίπτωση δεδομένων με λίγες παρατηρήσεις και πολλές μεταβλητές. Προφανώς αν σε ένα τέτοιο σετ δεδομένων εφαρμοσθεί ένα γενικευμένο γραμμικό μοντέλο, η υπερπαραμετροποίησή του μπορεί να ξεπεραστεί με τη χρήση της μεθόδου των κυρίων συνιστωσών. Ένα σημείο που παρουσιάζει εξίσου μεγάλο ενδιαφέρον στη μέθοδο αυτή είναι το ότι μπορεί να διαπιστωθεί πόσο μοιάζουν ή όχι οι μεταβλητές και δίνοντας ονόματα στις κύριες συνιστώσες, με βάση ποιες από τις αρχικές μεταβλητές έχουν μεγαλύτερη επίδραση, μπορούν να ποσοτικοποιηθούν μη μετρήσιμες ποσότητες, όπως η ευφυΐα, η εμπορευσιμότητα ενός προϊόντος ή άλλες αφηρημένες έννοιες. Ωστόσο, επειδή τέτοιες ερμηνείες εμπεριέχουν σε μεγάλο βαθμό υποκειμενικά κριτήρια πολλοί δεν εμπιστεύονται τα αποτελέσματα της μεθόδου αυτής.

Η μέθοδος των Κυρίων Συνιστωσών, αρχικά, περιγράφηκε το 1901 από τον Karl Pearson και αναπτύχθηκε περισσότερο το 1933 από τον Hotelling. Όμως, η πρακτική χρήση της μεθόδου ακολούθησε μετά την ευρεία διάδοση των ηλεκτρονικών

υπολογιστών, επειδή οι μαθηματικές πράξεις ήταν πολύ δύσκολο να πραγματοποιηθούν διαφορετικά για πολλές μεταβλητές.

Τέλος, αξίζει να σημειωθεί ότι η τεχνική των κυρίων συνιστωσών δεν επιτυγχάνει πάντοτε τη μείωση της διάστασης, για παράδειγμα αυτό συμβαίνει όταν οι αρχικές μεταβλητές είναι ασυσχέτιστες μεταξύ τους. Τότε άλλες τεχνικές πρέπει να αναζητηθούν.

## 5.2 Η βασική ιδέα της μεθόδου

Έστω ένας τετραγωνικός συμμετρικός πίνακας  $A$  διαστάσεων  $p \times p$ . Ο πίνακας αυτός μπορεί να αναπαρασταθεί ως

$$A = P\Lambda P' \quad (5.1)$$

όπου  $\Lambda$  είναι ένας  $p \times p$  διαγώνιος πίνακας και στοιχεία της διαγωνίου του είναι οι ιδιοτιμές του πίνακα  $A$ , δηλαδή

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_p \end{bmatrix}$$

και  $P$  είναι ένας ορθογώνιος  $p \times p$  πίνακας (δηλαδή ισχύει ότι  $P'P = I_p$ ), ο οποίος αποτελείται από τα κανονικοποιημένα ιδιοδιανύσματα των αντίστοιχων ιδιοτιμών. Η παραπάνω αναπαράσταση του πίνακα  $A$  ονομάζεται φασματική ανάλυση του πίνακα  $A$  και επομένως, αφού ο πίνακας  $P$  είναι ορθογώνιος ισχύει  $P^{-1} = P'$ .

Κατά συνέπεια, μπορεί να αποδειχτεί ότι:

$$A = P\Lambda P' \Leftrightarrow P^{-1}A = P^{-1}P\Lambda P' \Leftrightarrow P^{-1}AP = \Lambda P'P = \Lambda \Leftrightarrow \Lambda = P'AP. \quad (5.2)$$

Στη συγκεκριμένη μέθοδο, αυτή η ιδέα προσαρμόζεται με την εξής χρησιμότητα:

- Η φασματική ανάλυση ενός πίνακα διακύμανσης επιτρέπει να πολλαπλασιαστεί το αρχικό διάνυσμα με έναν κατάλληλο πίνακα τέτοιον, ώστε να δημιουργηθεί ένα νέο διάνυσμα το οποίο να είναι ασυσχέτιστο, δηλαδή να έχει διαγώνιο πίνακα διακύμανσης.

## 5.3 Κύριες Συνιστώσες από τον πίνακα συνδιασποράς $\Sigma$

Αλγεβρικά, οι κύριες συνιστώσες είναι συγκεκριμένοι γραμμικοί συνδυασμοί των  $p$  τυχαίων μεταβλητών  $X_1, X_2, \dots, X_p$ . Γεωμετρικά, αυτοί οι γραμμικοί συνδυασμοί παριστάνουν την επιλογή ενός νέου συστήματος συντεταγμένων που δίνεται με την περιστροφή του αρχικού συστήματος το οποίο έχει τις μεταβλητές  $X_1, X_2, \dots, X_p$  ως άξονες συντεταγμένων. Οι νέοι αυτοί άξονες παριστάνουν τις κατευθύνσεις με τη

μέγιστη διασπορά και δίνουν μια πιο απλή περιγραφή της δομής διασποράς – συνδιασποράς των δεδομένων.

Οι κύριες συνιστώσες εξαρτώνται από τον πίνακα συνδιασποράς  $\Sigma$  (ή τον πίνακα συσχετίσεων  $\rho$ ) των μεταβλητών  $X_1, X_2, \dots, X_p$ , γιατί η ανάλυση των κυρίων συνιστωσών βασίζεται στη φασματική ανάλυση ενός τετραγωνικού πίνακα. Σημαντικό είναι το γεγονός πως δεν απαιτείται πολυμεταβλητή κανονική κατανομή των δεδομένων ή κάποια άλλου είδους προϋπόθεση για την εφαρμογή της μεθόδου, αφού πρόκειται για έναν καθαρά μαθηματικό μετασχηματισμό των δεδομένων. Ωστόσο, οι κύριες συνιστώσες που εξάγονται από πληθυσμούς με πολυδιάστατη κανονική κατανομή παρουσιάζουν χρήσιμες γεωμετρικές ερμηνείες.

Έστω, λοιπόν, ένα σύνολο από  $p$  μεταβλητές  $X_1, X_2, \dots, X_p$  το οποίο έχει τον πίνακα συνδιασποράς  $\Sigma$  με ιδιοτιμές  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$  και  $Y_1, Y_2, \dots, Y_p$  οι κύριες συνιστώσες οι οποίες προκύπτουν από τους παρακάτω γραμμικούς συνδυασμούς:

$$\begin{aligned} Y_1 &= \mathbf{l}'_1 \mathbf{X} = l_{11}X_1 + l_{21}X_2 + \dots + l_{p1}X_p \\ Y_2 &= \mathbf{l}'_2 \mathbf{X} = l_{12}X_1 + l_{22}X_2 + \dots + l_{p2}X_p \\ &\dots \\ Y_p &= \mathbf{l}'_p \mathbf{X} = l_{1p}X_1 + l_{2p}X_2 + \dots + l_{pp}X_p. \end{aligned} \quad (5.3)$$

Τότε:

$$Var(Y_i) = \mathbf{l}'_i \Sigma \mathbf{l}_i \quad i = 1, 2, \dots, p \quad (5.4)$$

$$Cov(Y_i, Y_k) = \mathbf{l}'_i \Sigma \mathbf{l}_k \quad i = 1, 2, \dots, p \quad (5.5)$$

όπου  $k$  είναι ο αριθμός των κυρίων συνιστωσών.

Οι κύριες συνιστώσες είναι εκείνοι οι γραμμικοί συνδυασμοί των ασυσχέτιστων τυχαίων μεταβλητών  $Y_1, Y_2, \dots, Y_p$  των οποίων η διασπορά στη σχέση 5.4 είναι όσο το δυνατόν μεγαλύτερη. Πιο συγκεκριμένα, η πρώτη κύρια συνιστώσα είναι εκείνος ο γραμμικός συνδυασμός με τη μεγαλύτερη διασπορά. Έτσι, προκύπτει η μέγιστη διασπορά  $Var(Y_1) = \mathbf{l}'_1 \Sigma \mathbf{l}_1$ . Είναι προφανές ότι η τιμή αυτή μπορεί να μεγαλώσει αν πολλαπλασιαστεί με κάποια σταθερά. Προκειμένου να αποφευχθεί αυτό ορίζονται τα παρακάτω:

Πρώτη Κύρια Συνιστώσα = γραμμικός συνδυασμός  $\mathbf{l}'_1 \mathbf{X}$  που μεγιστοποιεί τη  $Var(\mathbf{l}'_1 \mathbf{X})$  υπό τον περιορισμό  $\mathbf{l}'_1 \mathbf{l}_1 = 1$ .

Δεύτερη Κύρια Συνιστώσα = γραμμικός συνδυασμός  $\mathbf{l}'_2 \mathbf{X}$  που μεγιστοποιεί τη  $Var(\mathbf{l}'_2 \mathbf{X})$  υπό τον περιορισμό  $\mathbf{l}'_2 \mathbf{l}_2 = 1$  και  $Cov(\mathbf{l}'_1 \mathbf{X}, \mathbf{l}'_2 \mathbf{X}) = 0$ .

Και εν γένει για την  $i$  κύρια συνιστώσα ισχύει ότι:

$i$  – οστή κύρια συνιστώσα = γραμμικός συνδυασμός  $\mathbf{l}'_i \mathbf{X}$  που μεγιστοποιεί τη  $Var(\mathbf{l}'_i \mathbf{X})$  υπό τον περιορισμό  $\mathbf{l}'_i \mathbf{l}_i = 1$  και  $Cov(\mathbf{l}'_i \mathbf{X}, \mathbf{l}'_k \mathbf{X}) = 0$  για  $k < i$ .

Επομένως, το πρόβλημα εύρεσης των κυρίων συνιστωσών ουσιαστικά επιλύεται με την εύρεση των στοιχείων του πίνακα:

$$\mathbf{L} = \begin{bmatrix} l_{11} & l_{21} & \cdots & l_{p1} \\ l_{12} & l_{22} & \cdots & l_{p2} \\ \vdots & \vdots & \ddots & \vdots \\ l_{1p} & l_{2p} & \cdots & l_{pp} \end{bmatrix}$$

#### Χρήσιμη πρόταση 1

Έστω ότι ο πίνακας  $\mathbf{\Sigma}$  έχει ζεύγη ιδιοτιμών – ιδιοδιανυσμάτων  $(\lambda_1, \mathbf{e}_1), (\lambda_2, \mathbf{e}_2), \dots, (\lambda_p, \mathbf{e}_p)$ , με  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$ . Τότε υπάρχει ένας  $p \times p$  ορθογώνιος πίνακας  $\mathbf{Q} = (\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_p)$  τέτοιος ώστε:

$$\mathbf{Q}' \mathbf{\Sigma} \mathbf{Q} = \mathbf{\Lambda} = \begin{pmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_p \end{pmatrix} \quad (5.6)$$

Από τη σχέση διαγωνοποίησης προκύπτει ότι τα διανύσματα  $\mathbf{e}_i \in R^p$  είναι τα ορθογώνια ιδιοδιανύσματα του πίνακα  $\mathbf{\Sigma}$  με μοναδιαίο μήκος που αντιστοιχούν στις ιδιοτιμές  $\lambda_i$ .

Άρα, έτσι όπως ορίστηκαν οι κύριες συνιστώσες και βάσει των αποτελεσμάτων της πρότασης 1 παρατηρείται ότι οι κύριες συνιστώσες έχουν την εξής βέλτιστη ιδιότητα. Η πρώτη κύρια συνιστώσα  $Y_1 = \mathbf{e}'_1 \mathbf{X}$  έχει τη μεγαλύτερη διασπορά  $\lambda_1$ , μεταξύ όλων των γραμμικών συνδυασμών της μορφής  $\mathbf{l}' \mathbf{X}$  με  $\mathbf{l}' \mathbf{l} = 1$ . Στη συνέχεια, μεταξύ όλων των γραμμικών συνδυασμών  $\mathbf{l}' \mathbf{X}$  με  $\mathbf{l}' \mathbf{l} = 1$  που είναι ασυσχέτιστοι με την  $Y_1$ , η κύρια συνιστώσα  $Y_2 = \mathbf{e}'_2 \mathbf{X}$  έχει τη μεγαλύτερη διασπορά  $\lambda_2$ . Γενικά, μεταξύ όλων των γραμμικών συνδυασμών  $\mathbf{l}' \mathbf{X}$  με  $\mathbf{l}' \mathbf{l} = 1$  που είναι ασυσχέτιστοι με τις  $Y_1, Y_2, \dots, Y_{m-1}$ , η κύρια συνιστώσα  $Y_m = \mathbf{e}'_m \mathbf{X}$  έχει τη μεγαλύτερη διασπορά  $\lambda_m$ .

Επίσης, είναι αξιοσημείωτο ότι η γραμμική συνάρτηση  $\mathbf{l}' \mathbf{X}$  έχει διασπορά  $Var(\mathbf{l}' \mathbf{X}) = \mathbf{l}' \mathbf{\Sigma} \mathbf{l}$  και η συνθήκη να είναι η  $\mathbf{l}' \mathbf{X}$  ασυσχέτιστη με την  $i$  – κύρια συνιστώσα  $Y_i = \mathbf{e}'_i \mathbf{X}$  σημαίνει ότι:

$$0 = Cov(\mathbf{l}' \mathbf{X}, \mathbf{e}'_i \mathbf{X}) = \mathbf{l}' Cov(\mathbf{X}) \mathbf{e}_i = \mathbf{l}' \mathbf{\Sigma} \mathbf{e}_i = \lambda_i \mathbf{l}' \mathbf{e}_i \quad (5.7)$$

επειδή  $\mathbf{\Sigma} \mathbf{e}_i = \lambda_i \mathbf{e}_i$ , Επομένως πρέπει  $\mathbf{l}' \mathbf{e}_i = 0$ .

Η προαναφερθείσα ιδιότητα των κυρίων συνιστωσών είναι άμεση συνέπεια της επόμενης πρότασης. Η ιδιότητα αυτή μπορεί να χρησιμοποιηθεί και σαν άλλος τρόπος ορισμού των κυρίων συνιστωσών και έχει ευνότητα, από στατιστικής πλευράς, ερμηνεία αναφορικά με τη μέγιστη δυνατή μείωση της διάστασης των δεδομένων.

## Χρήσιμη πρόταση 2

Έστω ο  $p \times p$  ορθογώνιος πίνακας  $\mathbf{Q} = (\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_p)$  τέτοιος ώστε:

$$\mathbf{Q}'\Sigma\mathbf{Q} = \Lambda = \begin{pmatrix} \lambda_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \lambda_p \end{pmatrix} \quad (5.8)$$

όπου  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$ . Τότε ισχύει ότι:

$$\max\{\mathbf{l}'\Sigma\mathbf{l} : \mathbf{l}'\mathbf{l} = 1 \text{ και } \mathbf{l}'\mathbf{e}_i = 0, i = 1, \dots, m-1\} = \lambda_m, \forall m = 1, 2, \dots, p$$

Το μέγιστο επιτυγχάνεται για  $\mathbf{l} = \mathbf{e}_m$ .

Προκύπτει ένα πολύ σημαντικό συμπέρασμα για τη συνολική διασπορά του πληθυσμού από τις προηγούμενες παρατηρήσεις:

$$\sum_{i=1}^p \text{Var}(X_i) = \sum_{i=1}^p \text{Var}(Y_i) \Leftrightarrow \sigma_{11}^2 + \sigma_{22}^2 + \dots + \sigma_{pp}^2 = \lambda_1 + \lambda_2 + \dots + \lambda_p. \quad (5.9)$$

Και κατά συνέπεια, το ποσοστό της συνολικής διασποράς που ερμηνεύει η  $k$  κύρια συνιστώσα θα είναι:

$$\frac{\lambda_k}{\lambda_1 + \lambda_2 + \dots + \lambda_p}, \quad k = 1, 2, \dots, p. \quad (5.10)$$

Εάν το περισσότερο ποσοστό (π.χ. 80 – 90 %) της συνολικής διασποράς του πληθυσμού, για μεγάλο  $p$ , μπορεί να ερμηνευθεί από τις πρώτες μία, δύο ή και τρεις κύριες συνιστώσες τότε αυτές οι συνιστώσες μπορούν να «αντικαταστήσουν» τις αρχικές  $p$  μεταβλητές χωρίς να χαθεί αρκετό μέρος πληροφορίας, το οποίο είναι και το ζητούμενο της εν λόγω μεθόδου.

Τέλος, αξίζει να σημειωθεί η σχέση η οποία δίνει τη συσχέτιση ανάμεσα στην κύρια συνιστώσα  $Y_i$  και στη μεταβλητή  $X_k$  και είναι:

$$\rho_{Y_i, X_k} = \frac{e_{ki}\sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}}, \quad i, k = 1, 2, \dots, p \quad (5.11)$$

όπου  $Q = (e_{ij})$  και  $\sigma_{ii} = \text{Var}(X_i)$ .

Άρα συνάγονται τα εξής συμπεράσματα:

- ✓ Για να κατασκευαστούν οι κύριες συνιστώσες χρειάζεται να υπολογιστούν οι ιδιοτιμές και τα ιδιοδιανύσματα του πίνακα  $\Sigma$ .
- ✓ Η μεγαλύτερη ιδιοτιμή και το ιδιοδιάνυσμά της αντιστοιχούν στην πρώτη κύρια συνιστώσα, η δεύτερη μεγαλύτερη ιδιοτιμή στη δεύτερη κύρια συνιστώσα κ.ο.κ.
- ✓ Η διακύμανση της κάθε κύριας συνιστώσας είναι ίση με την ιδιοτιμή που της αντιστοιχεί, δηλαδή  $Var(Y_i) = \lambda_i$ .
- ✓ Οι κύριες συνιστώσες είναι ασυσχέτιστες μεταξύ τους και συνεπώς, ο πίνακας διακύμανσής τους είναι ο διαγώνιος με διαγώνια στοιχεία τις ιδιοτιμές  $\lambda_i$ .
- ✓ Η συνολική διακύμανση των κυρίων συνιστωσών είναι ίδια με τη συνολική διακύμανση των αρχικών μεταβλητών εξ' αιτίας των ιδιοτήτων του ίχνους συμμετρικού και τετραγωνικού πίνακα και έτσι, η συνολική διακύμανση διατηρείται.
- ✓ Η ποσότητα  $\frac{\lambda_i}{\sum_{i=1}^p \lambda_i}$  δείχνει το ποσοστό της συνολικής διακύμανσης που εξηγεί η  $i$  κύρια συνιστώσα.

#### 5.4 Κύριες Συνιστώσες από τον πίνακα συσχέτισης $\rho$

Η μέθοδος της Ανάλυσης των Κυρίων Συνιστωσών παρουσιάζει ένα σημαντικό μειονέκτημα καθώς οι κύριες συνιστώσες επηρεάζονται αν αλλάξει η κλίμακα μέτρησης των δεδομένων με τη χρήση του πίνακα διακύμανσης, δηλαδή αν αλλάξουν οι μονάδες μέτρησης των δεδομένων. Κατά συνέπεια, αλλάζει και η ερμηνεία τους.

Επίσης, οι κύριες συνιστώσες είναι δύσκολο να ερμηνευτούν όταν οι μεταβλητές μετρούνται σε διαφορετικές μονάδες μέτρησης (π.χ. βάρος σε γραμμάρια, ύψος σε εκατοστά). Ακόμη και όταν όλες οι μεταβλητές έχουν την ίδια μονάδα μέτρησης, μπορεί τα πεδία τιμών τους να είναι πολύ διαφορετικά και το εύρος τους πολύ μεγάλο. Τότε οι κύριες συνιστώσες επηρεάζονται πολύ από τις μεταβλητές που παρουσιάζουν μεγάλη διασπορά και τείνουν να ταυτίζονται με αυτές τις μεταβλητές.

Σε τέτοιες περιπτώσεις είναι προτιμότερο να τυποποιούνται οι μεταβλητές και στη συνέχεια, να εφαρμόζεται η τεχνική των κυρίων συνιστωσών στον πίνακα συνδιασποράς των τυποποιημένων μεταβλητών που δεν είναι άλλος από τον πίνακα συσχέτισης των αρχικών μεταβλητών. Αναλυτικότερα οι τυποποιημένες μεταβλητές είναι:

$$Z_1 = \frac{(X_1 - \mu_1)}{\sqrt{\sigma_{11}}}$$

$$Z_2 = \frac{(X_2 - \mu_2)}{\sqrt{\sigma_{22}}}$$

$$Z_p = \frac{(X_p - \mu_p)}{\sqrt{\sigma_{pp}}}$$

ή σε μορφή πινάκων:

$$\mathbf{Z} = (\mathbf{V}^{1/2})^{-1}(\mathbf{X} - \boldsymbol{\mu}) \quad (5.12)$$

όπου ο  $\mathbf{V}^{1/2}$  είναι ο διαγώνιος πίνακας των τυπικών αποκλίσεων και ισχύουν τα εξής:

$$E(\mathbf{Z}) = \mathbf{0} \quad (5.13)$$

$$Cov(\mathbf{Z}) = (\mathbf{V}^{1/2})^{-1} \boldsymbol{\Sigma} (\mathbf{V}^{1/2})^{-1} = \boldsymbol{\rho} \quad (5.14)$$

όπου  $\boldsymbol{\rho}$  ο πίνακας συσχέτισης των  $X_1, \dots, X_p$ .

Σημειώνεται ότι οι διασπορές των  $Z_i$ , δηλαδή τα διαγώνια στοιχεία του  $\boldsymbol{\rho}$  είναι 1 και συνεπώς, η συνολική μεταβλητότητα των  $Z_i$  είναι ίση με  $p$ , το οποίο είναι και το πλήθος των μεταβλητών. Οι κύριες συνιστώσες του  $\mathbf{Z}$  μπορούν να υπολογισθούν από τα ιδιοδιανύσματα του πίνακα συσχετίσεων  $\boldsymbol{\rho}$  του διανύσματος  $\mathbf{X}$ . Όλα τα προηγούμενα αποτελέσματα μπορούν να εφαρμοσθούν και σε αυτή την περίπτωση με κάποιες απλοποιήσεις.

Συμπερασματικά, λοιπόν, η  $i$  - οστή κύρια συνιστώσα των κανονικοποιημένων μεταβλητών  $Z_1, Z_2, \dots, Z_p$  με  $Cov(\mathbf{Z}) = \boldsymbol{\rho}$  δίνεται από τη σχέση:

$$Y_i = e_i' \mathbf{Z} = e_i' (\mathbf{V}^{1/2})^{-1} (\mathbf{X} - \boldsymbol{\mu}), \quad i = 1, 2, \dots, p. \quad (5.15)$$

Επιπλέον,

$$\sum_{i=1}^p Var(Y_i) = \sum_{i=1}^p Var(Z_i) = p \quad (5.16)$$

και

$$\rho_{Y_i, Z_k} = e_{ki} \sqrt{\lambda_i}, \quad i, k = 1, 2, \dots, p. \quad (5.17)$$

Στην εν λόγω περίπτωση  $(\lambda_1, e_1), (\lambda_2, e_2), \dots, (\lambda_p, e_p)$  είναι τα ζεύγη ιδιοτιμών – ιδιοδιανυσμάτων του πίνακα συσχετίσεων  $\boldsymbol{\rho}$  με  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$ .

Από τη σχέση 5.16 συνάγεται το συμπέρασμα ότι η συνολική πληθυσμιακή διασπορά (για τις τυποποιημένες μεταβλητές) είναι ίση με  $p$ , το άθροισμα δηλαδή των διαγωνίων στοιχείων του πίνακα συσχετίσεων  $\boldsymbol{\rho}$ . Με χρήση της σχέσης 5.10 και αντικαθιστώντας με  $\mathbf{Z}$  στη θέση του  $\mathbf{X}$ , προκύπτει ότι το ποσοστό της συνολικής διασποράς που ερμηνεύεται από την  $k$  κύρια συνιστώσα του  $\mathbf{Z}$  είναι:

$$\frac{\lambda_k}{p}, \quad k = 1, 2, \dots, p \quad (5.18)$$

όπου  $\lambda_k, k = 1, 2, \dots, p$  οι ιδιοτιμές του πίνακα συσχετίσεων  $\rho$ .

Επομένως, για να ξεπεραστούν τα προβλήματα που δημιουργούνται από τη χρήση του πίνακα διακύμανσης μπορεί να χρησιμοποιηθεί ο πίνακας συσχετίσεων καθώς οι συσχετίσεις δεν αλλάζουν αν αλλάξουν οι μονάδες μέτρησης ή ή κλίμακα. Επιπρόσθετα, δίνεται το ίδιο βάρος σε όλες τις μεταβλητές αφού όλα τα στοιχεία της διαγωνίου είναι ίσα με 1.

Ωστόσο, η γενικευμένη χρήση του πίνακα συσχετίσεων δεν ενδείκνυται επειδή η διαφορά στις διακυμάνσεις ενδέχεται να περιέχει πολύτιμη πληροφορία για την έρευνα. Άρα, θέτοντας όλες τις μεταβλητές με το ίδιο βάρος χάνεται χρήσιμη πληροφορία. Στην πράξη, λοιπόν, δεν είναι ξεκάθαρο ποιος από τους δύο πίνακες πρέπει να χρησιμοποιείται στην εν λόγω μέθοδο. Μια καλή στρατηγική θα ήταν να αποφεύγεται ο πίνακας διακύμανσης όταν υπάρχουν μεταβλητές με πολύ μεγαλύτερη διακύμανση από ότι οι υπόλοιπες. Ενώ αν οι διακυμάνσεις διαφέρουν αλλά είναι συγκρίσιμες τότε καλό είναι να αξιοποιείται αυτή η πληροφορία με τη χρήση του πίνακα διακύμανσης.

## 5.5 Βήματα της Ανάλυσης σε Κύριες Συνιστώσες

### 5.5.1 Έλεγχος συσχετίσεων

Είναι σημαντικό να γίνει ο υπολογισμός του πίνακα συσχετίσεων καθώς αν δεν υπάρχουν ικανοποιητικές συσχετίσεις δεν έχει νόημα να συνεχισθεί περαιτέρω η ανάλυση. Μεταβλητές που δεν εμφανίζουν συσχετίσεις με τις υπόλοιπες πρέπει να αφαιρεθούν από την αρχή.

Η έννοια της ικανοποιητικής συσχέτισης μεταξύ δύο μεταβλητών δεν ταυτίζεται απαραίτητα με την ένδειξη της στατιστικής σημαντικότητας βάσει ενός ελέγχου υποθέσεων. Ακόμα και συσχετίσεις πολύ μικρές (π.χ. 0.10) μπορεί να είναι στατιστικά σημαντικές για ένα μέτριο μεγέθους δείγμα. Όσον αφορά στην εφαρμογή της Ανάλυσης σε Κύριες Συνιστώσες συσχετίσεις της τάξης τουλάχιστον του 0.40 θεωρούνται ικανοποιητικές πρακτικά για να πραγματοποιηθεί η εν λόγω μέθοδος. Επομένως, είναι χρήσιμο ο ερευνητής να ελέγξει όχι μόνο αν οι συσχετίσεις είναι στατιστικά σημαντικά διάφορες του 0 αλλά αν είναι και επαρκώς μεγάλες σε απόλυτη τιμή εξ' αρχής.

Ένα μέτρο για να γίνει η αξιολόγηση των συσχετίσεων για ένα σετ δεδομένων είναι το

$$\varphi = \sqrt{\frac{\sum_{i=1}^p \sum_{j=1}^p r_{ij}^2 - p}{p(p-1)}} \quad (5.19)$$



όπου  $r_{ij}$  είναι το  $ij$  στοιχείο του πίνακα συσχετίσεων, δηλαδή η συσχέτιση των μεταβλητών  $X_i$  και  $X_j$  και  $p$  είναι ο αριθμός των αρχικών μεταβλητών. Το στατιστικό  $\varphi$  λαμβάνει τιμές «κοντά» στο 1 αν υπάρχουν μεγάλες συσχετίσεις. Αν δεν υπάρχουν συσχετίσεις, η τιμή είναι «κοντά» στο 0.

Από την άλλη πλευρά, το αντίστοιχο μέτρο, στην περίπτωση της χρήσης του πίνακα διακύμανσης, είναι το

$$\varphi = \sqrt{\frac{\sum_{i=1}^p \sum_{j=1}^p s_{ij}^2 - \sum_{j=1}^p s_{jj}^2}{\sum_{i=1}^p \sum_{j \neq i}^p s_{ii} s_{jj}}} \quad (5.20)$$

όπου  $s_{ij}$  είναι το  $ij$  στοιχείο του πίνακα διακύμανσης, δηλαδή η διακύμανση της  $X_i$  με τη  $X_j$  μεταβλητή και για το οποίο ισχύει, επίσης, ότι λαμβάνει τιμές «κοντά» στο 1 αν υπάρχουν μεγάλες συσχετίσεις ενώ αν δεν υπάρχουν συσχετίσεις, η τιμή είναι «κοντά» στο 0.

### 5.5.2 Επιλογή κατάλληλου πίνακα

Όσον αφορά στην επιλογή μεταξύ του πίνακα διακύμανσης και του πίνακα συσχετίσεων τα κριτήρια έχουν αναφερθεί ήδη. Είναι σαφές, πλέον, ότι ανάλογα με την επιλογή του πίνακα διαφοροποιούνται και τα αποτελέσματα που προκύπτουν. Άρα, είναι σημαντικό να αξιοποιηθεί σωστά οποιαδήποτε πληροφορία περιέχουν τα αρχικά δεδομένα.

### 5.5.3 Υπολογισμός ιδιοτιμών και ιδιοδιανυσμάτων

Ο υπολογισμός των ιδιοτιμών και των ιδιοδιανυσμάτων γίνεται, προφανώς, με βάση τον πίνακα που έχει επιλεγεί για την εφαρμογή της μεθόδου. Τα ιδιοδιανύσματα που υπολογίζουν τα στατιστικά πακέτα είναι κανονικοποιημένα, δηλαδή το άθροισμα τετραγώνων τους είναι ίσο με 1. Επειδή δεν είναι μοναδικά τα ιδιοδιανύσματα που υπολογίζονται, μπορεί να διαφέρει η λύση ανάλογα με το στατιστικό πακέτο.

### 5.5.4 Απόφαση για το πλήθος των Κύριων Συνιστωσών

Το πλήθος των κυρίων συνιστωσών που θα επιλεγεί αποτελεί κρίσιμη απόφαση για τον ερευνητή και δεν έχει, δυστυχώς, μια κοινώς αποδεκτή απάντηση. Εν γένει, το ζητούμενο είναι να επιλεγούν λιγότερες κύριες συνιστώσες από όσες ήταν οι αρχικές μεταβλητές έτσι, ώστε να μειωθούν οι διαστάσεις του προβλήματος αλλά ταυτόχρονα να χαθεί όσο το δυνατόν λιγότερη πληροφορία για τα δεδομένα ώστε να είναι αξιόπιστα τα αποτελέσματα της έρευνας. Στη βιβλιογραφία υπάρχουν κάποια κριτήρια και τα πιο σημαντικά αναφέρονται στη συνέχεια.

### Ποσοστό συνολικής διακύμανσης που εξηγούν οι συνιστώσες

Με βάση αυτό το κριτήριο επιλέγονται τόσες κύριες συνιστώσες όσες απαιτούνται για να εξηγούν αθροιστικά μεγαλύτερο ποσοστό από ένα κάτω όριο που θεωρείται επιθυμητό για την ανάλυση από τον ερευνητή (π.χ. 80 – 85%). Τα αποτελέσματα, ωστόσο, δεν θεωρούνται ικανοποιητικά από το κριτήριο αυτό, αν και είναι πολύ εύκολο στην εφαρμογή, καθώς δεν είναι ξεκάθαρο ποιο ποσοστό θα πρέπει να θεωρηθεί ως κάτω όριο.

### Κριτήριο Kaiser

Με βάση αυτό το κριτήριο ο αριθμός των κυρίων συνιστωσών προσδιορίζεται ως ο αριθμός των ιδιοτιμών που είναι μεγαλύτερες από  $\bar{\lambda} = \sum_{i=1}^p \lambda_i / p$ , δηλαδή μεγαλύτερες από τη μέση τιμή των ιδιοτιμών. Στην περίπτωση που γίνεται χρήση του πίνακα συσχετίσεων ισχύει ότι  $\bar{\lambda} = 1$  και άρα, επιλέγονται τόσες κύριες συνιστώσες όσες και οι ιδιοτιμές μεγαλύτερες της μονάδας. Στην πράξη η υπόθεση αυτή είναι κάπως απλοϊκή καθώς, ακόμα κι αν δεν υπάρχει δομή και όλες οι ιδιοτιμές είναι ίσες με 1, όταν πρόκειται για ένα δείγμα σίγουρα κάποιες από τις ιδιοτιμές θα είναι μεγαλύτερες του 1 αφού το άθροισμά τους πρέπει να είναι ίσο με  $p$ . Συνήθως το εν λόγω κριτήριο υπερεκτιμά το πλήθος των απαιτούμενων συνιστωσών.

### Ποσοστό της διακύμανσης των αρχικών μεταβλητών που ερμηνεύεται

Με τη χρήση αυτού του κριτηρίου επιλέγονται τόσες κύριες συνιστώσες ώστε να ερμηνεύεται για κάθε μεταβλητή, τουλάχιστον, ένα υψηλό ποσοστό, αφού αν διατηρηθούν λιγότερες συνιστώσες από τις αρχικές μεταβλητές ένα μέρος της πληροφορίας χάνεται. Και σε αυτήν την περίπτωση, όμως, το υψηλό αυτό ποσοστό αποτελεί υποκειμενική απόφαση και αν κάποια από τις μεταβλητές δεν ερμηνεύεται σωστά οδηγεί σε μεγάλο πλήθος συνιστωσών γεγονός που δεν είναι αποτελεσματικό για την εφαρμογή της μεθόδου.

### Scree plot (Γραφική μέθοδος)

Το scree plot αποτελεί ένα γράφημα, το οποίο έχει στον κατακόρυφο άξονα των  $y$  την τιμή κάθε ιδιοτιμής  $\lambda_i$  και στον οριζόντιο άξονα των  $x$  τη σειρά, δηλαδή τον αύξοντα αριθμό τους, ο οποίος αυξάνει καθώς φθίνουν οι ιδιοτιμές. Με το κριτήριο αυτό, επιλέγονται τόσες κύριες συνιστώσες μέχρι το γράφημα να αρχίσει να γίνεται περίπου παράλληλο στον οριζόντιο άξονα, ουσιαστικά δηλαδή μέχρι να αλλάξει η κλίση του. Πρέπει να γίνει σαφές ότι ούτε αυτή η μέθοδος είναι ξεκάθαρη ως προς τον αριθμό των κυρίων συνιστωσών που πρέπει να επιλεγεί, εφ' όσον μερικές φορές παίζει ιδιαίτερο ρόλο και πάλι το υποκειμενικό κριτήριο για το πού αλλάζει η κλίση του γραφήματος.

### Μέθοδος του σπασμένου ραβδιού (Broken stick)

Η μέθοδος αυτή στηρίζεται στην παρατήρηση ότι αν ένα ραβδί μεγέθους 1 μονάδας σπάσει σε τυχαία  $p$  κομμάτια τότε το  $k$  μεγαλύτερο από αυτά αναμένεται να έχει μήκος

$$g_k = \frac{1}{p} \sum_{i=k}^p \left( \frac{1}{i} \right). \quad (5.21)$$

Με βάση το εν λόγω κριτήριο επιλέγονται τόσες συνιστώσες για τις οποίες ισχύει

$$\frac{\lambda_k}{\sum_{i=1}^p \lambda_i} > g_k. \quad (5.22)$$

### Κανονική προσέγγιση

Για το κριτήριο αυτό απαιτείται η προϋπόθεση ότι ο πληθυσμός ακολουθεί πολυμεταβλητή κανονική κατανομή και έτσι, μπορεί να κατασκευαστεί ένα διάστημα εμπιστοσύνης για τις ιδιοτιμές με βάση τις δειγματικές ιδιοτιμές. Τα διαστήματα εμπιστοσύνης για τις ιδιοτιμές βασίζονται στα ασυμπτωτικά αποτελέσματα της κανονικής κατανομής. Στην πράξη, στην περίπτωση που το 95% διάστημα εμπιστοσύνης για την  $i$  ιδιοτιμή δεν περιέχει το 1 και συγκεκριμένα, είναι μεγαλύτερο από το 1, τότε η αντίστοιχη κύρια συνιστώσα θεωρείται σημαντική. Προκειμένου να εφαρμοσθεί το εν λόγω κριτήριο πρέπει να γίνεται χρήση του πίνακα συσχέτισεων.

### Bootstrap

Η μέθοδος bootstrap εφαρμόζεται όλο και περισσότερο στην επιστήμη της στατιστικής, αφού επιτρέπει με επαναληπτική δειγματοληψία με επανάθεση να εκτιμηθούν οι ποσότητες του πληθυσμού καθώς και τα τυπικά σφάλματα των εκτιμητριών τους. Το πλεονέκτημα της μεθόδου αυτής είναι ότι μπορεί να κατασκευάσει διαστήματα εμπιστοσύνης για τις ιδιοτιμές χωρίς να απαιτείται η υπόθεση της κανονικής κατανομής. Αρχικά, γίνεται δειγματοληψία με επανάθεση στις παρατηρήσεις έτσι, ώστε να δημιουργηθεί ένα δείγμα ίσου μεγέθους με το αρχικό και στο οποίο, προφανώς, μια παρατήρηση μπορεί να υπάρχει περισσότερες από μία φορές. Στη συνέχεια, κατασκευάζεται ο πίνακας διακύμανσης ή συσχέτισης και υπολογίζονται οι ιδιοτιμές. Με την επαναληπτικότητα αυτής της διαδικασίας δημιουργείται μια σειρά από τιμές της κατανομής των ιδιοτιμών και έτσι, μπορεί να εκτιμηθεί η τυπική απόκλιση των ιδιοτιμών. Τελικά, κατασκευάζονται διαστήματα εμπιστοσύνης και προκειμένου να αποφασισθεί αν μια συνιστώσα είναι σημαντική ή όχι, πραγματοποιείται ο έλεγχος για το αν η τιμή που το κριτήριο του Kaiser θέτει σαν όριο ανήκει στο διάστημα. Αποδεκτές είναι μόνο οι ιδιοτιμές για τις οποίες το διάστημα εμπιστοσύνης βρίσκεται εξ' ολοκλήρου «πάνω» από το όριο του κριτηρίου του Kaiser. Το μειονέκτημα για την εν λόγω μέθοδο είναι ότι έχει υπολογιστικό κόστος.

### Έλεγχοι υποθέσεων του Bartlett

Με την προϋπόθεση ότι ο πληθυσμός ακολουθεί την κανονική κατανομή μπορεί να πραγματοποιηθεί ένας ακόμη έλεγχος ως προς το αν οι  $p - k$  ιδιοτιμές, που δεν θεωρούνται σημαντικές για να χρησιμοποιηθούν ως κύριες συνιστώσες, είναι ίσες ή όχι. Ο εν λόγω έλεγχος έχει ως μηδενική την υπόθεση ότι:

$$H_0: \text{οι τελευταίες } p - k \text{ ιδιοτιμές είναι ίσες}$$

έναντι της εναλλακτικής ότι:

$$H_1: \text{δεν είναι ίσες.}$$

Η ελεγχοσυνάρτηση που υπολογίζεται είναι:

$$\chi = -(n-1) \sum_{j=k+1}^p \ln(\lambda_j) + (n-1)(p-k) \ln\left(\frac{\sum_{j=k+1}^p \lambda_j}{p-k}\right) \quad (5.23)$$

η οποία, κάτω από τη μηδενική υπόθεση, ακολουθεί την κατανομή  $\chi^2$  με  $\frac{1}{2}(p-k-1)(p-k+2)$  βαθμούς ελευθερίας.

### Cross Validation

Πρόκειται για μια επαναληπτική διαδικασία, όπου κάθε φορά αγνοούνται κάποιες από τις παρατηρήσεις των δεδομένων και εξετάζεται η συμπεριφορά των συνιστωσών προκειμένου να προβλεφθούν οι παρατηρήσεις που αγνοήθηκαν. Με την επανάληψη της διαδικασίας αυτής πολλές φορές αποδεικνύεται αν το μοντέλο με  $k$  συνιστώσες δίνει ικανοποιητικά αποτελέσματα. Συνεπώς, συγκρίνοντας τα αποτελέσματα για διάφορες τιμές του  $k$  εντοπίζεται η τιμή εκείνη για την οποία είναι πιο ικανοποιητικά.

#### **5.5.5 Εύρεση των κυρίων συνιστωσών**

Προκειμένου να υπολογισθούν οι κύριες συνιστώσες απαιτείται προηγουμένως η εύρεση των ιδιοτιμών και των ιδιοδιανυσμάτων του πίνακα που χρησιμοποιήθηκε για την ανάλυση βάσει της φασματικής ανάλυσης. Αν και έχει μεγάλο υπολογιστικό φόρτο με τη χρήση του ηλεκτρονικού υπολογιστή, το στάδιο αυτό έχει απλοποιηθεί αρκετά.

#### **5.5.6 Ερμηνεία των κυρίων συνιστωσών**

Είναι σημαντικό, από τη στιγμή που κατασκευάζονται οι κύριες συνιστώσες, να μπορεί ο ερευνητής να αποδώσει κάποια ερμηνεία σε αυτές με βάση τις μεταβλητές. Αποτελεί ζωτικής σημασίας σημείο για επιστήμες, όπως η ψυχολογία και το marketing, όπου μπορούν να ποσοτικοποιηθούν με αυτόν τον τρόπο μη μετρήσιμες μεταβλητές.

### 5.5.7 Δημιουργία νέων μεταβλητών

Όπως έχει ήδη αναφερθεί, οι κύριες συνιστώσες είναι νέες μεταβλητές, οι οποίες μπορούν να χρησιμοποιηθούν για περαιτέρω στατιστική ανάλυση. Για να επιτευχθεί αυτό αρκεί σε κάθε συνιστώσα να αντικατασταθούν οι τιμές με εκείνες που είχε η παρατήρηση για κάθε μεταβλητή.

### 5.6 Χρήσιμα αποτελέσματα

- ✓ Αν μια μεταβλητή παρατηρηθεί να είναι ασυσχέτιστη με τις υπόλοιπες, πρέπει να αφαιρεθεί από την ανάλυση εξ' αρχής καθώς στη συνέχεια, θα ταυτιστεί με μία από τις κύριες συνιστώσες. Αυτό σημαίνει ότι η μεταβλητή αυτή δεν είχε στατιστικά σημαντικές συσχετίσεις με τις υπόλοιπες μεταβλητές της έρευνας και συμβάλλει με το να χρησιμοποιηθεί στη μελέτη.
- ✓ Αν δύο ιδιοτιμές προκύψουν ίσες τότε αυτές αντιστοιχούν σε δύο όμοιες κύριες συνιστώσες γεγονός που αποτελεί πρόβλημα για την ανάλυση. Σε τέτοια περίπτωση πρέπει να ελεγχθεί το δείγμα καθώς μπορεί να υπάρχουν στήλες που να επαναλαμβάνονται περιέχοντας την ίδια πληροφορία.
- ✓ Αν υπάρχουν μηδενικές ιδιοτιμές, σημαίνει ότι κάποιες από τις αρχικές μεταβλητές ήταν γραμμικά εξαρτημένες. Στην πράξη, είναι δύσκολο να προκύψουν μηδενικές ιδιοτιμές αλλά μπορεί να προκύψουν πολύ μικρές, δηλαδή οι αρχικές μεταβλητές να ήταν σχεδόν γραμμικές. Έτσι, λοιπόν, μπορούν να αγνοηθούν και να παραμείνει στην ανάλυση μόνο μία μεταβλητή από αυτές καθώς περιέχουν περίπου την ίδια πληροφορία και τελικά το μεγαλύτερο μέρος από αυτή θα πάει σε κάποια από τις πρώτες κύριες συνιστώσες και η υπόλοιπη θα πάει σε μια συνιστώσα με αμελητέα διακύμανση.
- ✓ Δύο διαφορετικά σετ δεδομένων μπορεί να έχουν τα ίδια ιδιοδιανύσματα και διαφορετικές ιδιοτιμές. Αυτό σημαίνει ότι ενώ οι κύριες συνιστώσες είναι οι ίδιες, η κάθε μία από αυτές εξηγεί διαφορετικό ποσοστό της συνολικής διακύμανσης.
- ✓ Εν γένει, ο πίνακας συσχετίσεων μπορεί να έχει μόνο θετικά στοιχεία, δηλαδή όλες οι συσχετίσεις να είναι θετικές, τότε η πρώτη κύρια συνιστώσα μπορεί να θεωρηθεί ως ένας σταθμικός μέσος όρος των μεταβλητών με σταθμίσεις τους αντίστοιχους συντελεστές.
- ✓ Στην ανάλυση σε κύριες συνιστώσες ο στόχος είναι γραφούν οι κύριες συνιστώσες ως γραμμικοί συνδυασμοί των αρχικών μεταβλητών μέσω ενός πίνακα  $L$ , ο οποίος περιέχει τους συντελεστές. Επειδή, όμως, ο πίνακας αυτός είναι ορθογώνιος με  $L' = L^{-1}$ , μπορούν πολύ εύκολα και οι αρχικές μεταβλητές να γραφούν ως γραμμικοί συνδυασμοί των κύριων συνιστωσών και έτσι, να βρεθούν τα αρχικά δεδομένα.

## 5.7 Εφαρμογές της Ανάλυσης των Κυρίων Συνιστωσών

- Data mining (εξόρυξη γνώσης). Πρόκειται για έναν επιστημονικό χώρο στον οποίο μπορεί να συνδυαστεί η πληροφορική με τη στατιστική με πολύ σημαντικά αποτελέσματα. Η τεράστια πληροφορία των βάσεων δεδομένων μεγάλων εταιρειών και οργανισμών μπορεί να συμπυκνωθεί με τη βοήθεια της στατιστικής σε όσο το δυνατόν λιγότερες διαστάσεις μέσω της ανάλυσης σε κύριες συνιστώσες.
- Γραμμική παλινδρόμηση. Έχει ήδη αναφερθεί το πρόβλημα της πολυσυγγραμμικότητας το οποίο συναντάται στην περίπτωση που οι ανεξάρτητες μεταβλητές σε μια γραμμική παλινδρόμηση είναι συσχετισμένες. Αυτό το πρόβλημα μπορεί να ξεπεραστεί αν αντί για τις αρχικές συσχετισμένες μεταβλητές χρησιμοποιηθούν οι κύριες συνιστώσες, οι οποίες είναι ασυσχέτιστες.
- Γραφική απεικόνιση πολυδιάστατων δεδομένων. Η οπτική παρουσίαση των δεδομένων είναι ένα σημαντικό εργαλείο, το οποίο καθίσταται εξαιρετικά δύσκολο να υλοποιηθεί στην περίπτωση πολλών αρχικών μεταβλητών. Με την ανάλυση σε κύριες συνιστώσες μπορούν τα δεδομένα να οπτικοποιηθούν αποτελεσματικά, εφ' όσον οι πρώτες κύριες συνιστώσες ερμηνεύουν ένα μεγάλο ποσοστό της μεταβλητότητας του συνόλου των δεδομένων.
- Έλεγχος ποιότητας. Πολλές φορές οι εταιρείες ενδιαφέρονται για την πορεία πολλών χαρακτηριστικών ενός προϊόντος κάνοντας χρήση διαγραμμάτων ποιοτικού ελέγχου. Με την ανάλυση σε κύριες συνιστώσες η πληροφορία μπορεί να συμπυκνωθεί σε πολύ μικρότερο όγκο και έτσι, να μπορεί να παρακολουθηθεί το εν λόγω προϊόν πολύ πιο αποτελεσματικά.

## **ΚΕΦΑΛΑΙΟ 6**

### **ΕΦΑΡΜΟΓΗ ΤΗΣ ΑΝΑΛΥΣΗΣ ΚΥΡΙΩΝ ΣΥΝΙΣΤΩΣΩΝ ΣΤΑ ΔΕΔΟΜΕΝΑ ΤΟΥ ΔΕΙΓΜΑΤΟΣ**

#### **6.1 Εύρεση Κυρίων Συνιστωσών**

Η Ανάλυση σε Κύριες Συνιστώσες θα εφαρμοσθεί στα δεδομένα του δείγματος και συγκεκριμένα στις 12 μεταβλητές που αφορούν στις επιμέρους κατηγορίες δαπανών για τις οποίες έχουν καταγραφεί αναλυτικά οι δαπάνες για τα 3512 νοικοκυριά του δείγματος κατά το έτος 2010. Στο κεφάλαιο 2 είχε υπολογισθεί ο πίνακας με τις συσχετίσεις των μεταβλητών αυτών οπότε η εν λόγω ανάλυση έχει νόημα να εφαρμοσθεί. Έτσι, λοιπόν, από τις αρχικές συσχετισμένες μεταβλητές θα δημιουργηθούν νέες ασυσχέτιστες κύριες συνιστώσες με στόχο να περιοριστεί όσο το δυνατόν περισσότερο ο όγκος των δεδομένων και παράλληλα να ερμηνευτεί το μεγαλύτερο δυνατό ποσοστό της συνολικής διακύμανσης.

Στη συνέχεια, έχουν υπολογισθεί τόσο οι διακυμάνσεις – συνδιακυμάνσεις όσο και οι συσχετίσεις για τα δεδομένα του δείγματος όσον αφορά στις 12 μεταβλητές, με τις εντολές `con()` και `cor()` αντίστοιχα στην R. Τα αποτελέσματα παρουσιάζονται στον Πίνακα 6.1 και τον Πίνακα 6.2.

**Πίνακας 6.1:** Πίνακας διακύμανσης – συνδιακύμανσης

|               | food      | alcohol   | clothing  | housing    | equipment | health    |
|---------------|-----------|-----------|-----------|------------|-----------|-----------|
| food          | 5265780.2 |           |           |            |           |           |
| alcohol       | 781840.1  | 1116656.8 |           |            |           |           |
| clothing      | 2023955.1 | 626031.1  | 4313977.4 |            |           |           |
| housing       | 3689883.0 | 951851.3  | 3916734.3 | 28075392.4 |           |           |
| equipment     | 1509316.7 | 535913.6  | 2285856.1 | 4339483.3  | 6766649.1 |           |
| health        | 1103202.1 | 272569.8  | 1364254.5 | 3047718.0  | 1579124.4 | 5823885.1 |
| transport     | 2829777.2 | 1028052.4 | 3752973.7 | 7549059.0  | 2957402.2 | 2052932.9 |
| communication | 679055.9  | 217378.7  | 661668.5  | 1736439.3  | 514177.5  | 346028.1  |
| culture       | 1429847.8 | 499858.0  | 1940960.6 | 3751449.0  | 1918510.3 | 1476502.3 |
| education     | 1287391.7 | 279137.8  | 1189924.8 | 3207261.7  | 1101753.9 | 687027.6  |
| hotels        | 2198884.5 | 907269.3  | 2966951.0 | 5343572.9  | 2379132.2 | 1328959.9 |
| goods         | 1789004.1 | 537278.9  | 2370006.7 | 4501058.8  | 1852925.2 | 1384343.1 |

|               | transport | communication | culture   | education | hotels    | goods     |
|---------------|-----------|---------------|-----------|-----------|-----------|-----------|
| food          |           |               |           |           |           |           |
| alcohol       |           |               |           |           |           |           |
| clothing      |           |               |           |           |           |           |
| housing       |           |               |           |           |           |           |
| equipment     |           |               |           |           |           |           |
| health        |           |               |           |           |           |           |
| transport     | 23249227  |               |           |           |           |           |
| communication | 1197855   | 489202.2      |           |           |           |           |
| culture       | 3395061   | 604098.5      | 4398900.9 |           |           |           |
| education     | 2269520   | 498853.4      | 1521890.2 | 3892707.7 |           |           |
| hotels        | 4779874   | 1001917.0     | 2347148.7 | 1595815.0 | 6867252.5 |           |
| goods         | 3309824   | 694809.4      | 1698784.6 | 1027055.7 | 2582363.5 | 4668742.5 |

**Πίνακας 6.2:** Πίνακας συσχετίσεων

|               | food      | alcohol   | clothing  | housing   | equipment | health    |
|---------------|-----------|-----------|-----------|-----------|-----------|-----------|
| food          | 1         |           |           |           |           |           |
| alcohol       | 0.3224235 | 1         |           |           |           |           |
| clothing      | 0.4246495 | 0.2852312 | 1         |           |           |           |
| housing       | 0.3034718 | 0.1699990 | 0.3558953 | 1         |           |           |
| equipment     | 0.2528495 | 0.1949612 | 0.4230806 | 0.3148388 | 1         |           |
| health        | 0.1992128 | 0.1068837 | 0.2721761 | 0.2383447 | 0.2515493 | 1         |
| transport     | 0.2557506 | 0.2017675 | 0.3747420 | 0.2954783 | 0.2357867 | 0.1764266 |
| communication | 0.4230873 | 0.2941121 | 0.4554675 | 0.4685463 | 0.2826064 | 0.2050032 |
| culture       | 0.2970889 | 0.2255354 | 0.4455595 | 0.3375701 | 0.3516456 | 0.2917129 |
| education     | 0.2843501 | 0.1338853 | 0.2903720 | 0.3067931 | 0.2146703 | 0.1442919 |
| hotels        | 0.3656618 | 0.3276309 | 0.5451047 | 0.3848374 | 0.3490119 | 0.2101425 |
| goods         | 0.3608112 | 0.2353098 | 0.5280933 | 0.3931441 | 0.3296637 | 0.2654834 |

|               | transport | communication | culture   | education | hotels    | goods |
|---------------|-----------|---------------|-----------|-----------|-----------|-------|
| food          |           |               |           |           |           |       |
| alcohol       |           |               |           |           |           |       |
| clothing      |           |               |           |           |           |       |
| housing       |           |               |           |           |           |       |
| equipment     |           |               |           |           |           |       |
| health        |           |               |           |           |           |       |
| transport     | 1         |               |           |           |           |       |
| communication | 0.3551859 | 1             |           |           |           |       |
| culture       | 0.3357154 | 0.4118049     | 1         |           |           |       |
| education     | 0.2385634 | 0.3614955     | 0.3677776 | 1         |           |       |
| hotels        | 0.3782863 | 0.5466333     | 0.4270483 | 0.3056492 | 1         |       |
| goods         | 0.3176879 | 0.4597498     | 0.378576  | 0.2409174 | 0.4560641 | 1     |

Όπως προκύπτει από τον πίνακα 6.1 παρατηρούνται κάποιες διαφορές ανάμεσα στις συνδιακυμάνσεις των μεταβλητών (π.χ. housing – communication) αλλά όλες τις μεταβλητές έχουν την ίδια μονάδα μέτρησης. Από τον πίνακα 6.2 προκύπτει ότι οι συσχετίσεις μπορεί να μην είναι πάρα πολύ μεγάλες αλλά, ωστόσο, αρκούν ώστε να γίνει η ανάλυση λόγω του μεγάλου όγκου του δείγματος (3512 νοικοκυριά).

Στη συνέχεια, έχει πραγματοποιηθεί η ανάλυση σε κύριες συνιστώσες με τη χρήση του στατιστικού πακέτου R (με την εντολή `prcomp`) και παρουσιάζονται τα αποτελέσματα στον Πίνακα 6.3 όπου φαίνονται οι ιδιοτιμές και τα αντίστοιχα ποσοστά της συνολικής διακύμανσης που εξηγούνται από αυτές πρώτα με χρήση του πίνακα διακυμάνσεων και μετά, με χρήση του πίνακα συσχετίσεων.



**Πίνακας 6.3:** Πίνακας ιδιοτιμών και των αντίστοιχων ποσοστών διακύμανσης με χρήση του πίνακα διακυμάνσεων

| A/A | Ιδιοτιμή  | Ποσοστό της διακύμανσης | Αθροιστικό ποσοστό της διακύμανσης |
|-----|-----------|-------------------------|------------------------------------|
| 1   | 6527.2261 | 0.4488                  | 0.4488                             |
| 2   | 4223.5821 | 0.1879                  | 0.6367                             |
| 3   | 3035.9250 | 0.09709                 | 0.7338                             |
| 4   | 2263.9186 | 0.05399                 | 0.7878                             |
| 5   | 2147.8081 | 0.0486                  | 0.8364                             |
| 6   | 1929.0889 | 0.0392                  | 0.8756                             |
| 7   | 1859.0220 | 0.03641                 | 0.9120                             |
| 8   | 1694.8887 | 0.03026                 | 0.9423                             |
| 9   | 1550.5280 | 0.02533                 | 0.9676                             |
| 10  | 1378.0342 | 0.0200                  | 0.9876                             |
| 11  | 956.3357  | 0.00963                 | 0.99724                            |
| 12  | 511.9691  | 0.00276                 | 1.00000                            |

Επίσης, στον Πίνακα 6.4 υπάρχουν τα αντίστοιχα αποτελέσματα της ανάλυσης σε κύριες συνιστώσες και με τη χρήση του πίνακα συσχετίσεων, μέσω της εντολής `prcomp()` στην οποία έχει προστεθεί το όρισμα `scale = T`.

**Πίνακας 6.4:** Πίνακας ιδιοτιμών και των αντίστοιχων ποσοστών διακύμανσης με χρήση του πίνακα συσχετίσεων

| A/A | Ιδιοτιμή | Ποσοστό της διακύμανσης | Αθροιστικό ποσοστό της διακύμανσης |
|-----|----------|-------------------------|------------------------------------|
| 1   | 2.1478   | 0.3844                  | 0.3844                             |
| 2   | 0.99361  | 0.08227                 | 0.46671                            |
| 3   | 0.95476  | 0.07596                 | 0.54267                            |
| 4   | 0.89025  | 0.06605                 | 0.60871                            |
| 5   | 0.86778  | 0.06275                 | 0.67147                            |
| 6   | 0.86191  | 0.06191                 | 0.73338                            |
| 7   | 0.81562  | 0.05544                 | 0.78881                            |
| 8   | 0.79571  | 0.05276                 | 0.84157                            |
| 9   | 0.73310  | 0.04479                 | 0.88636                            |
| 10  | 0.72652  | 0.04399                 | 0.93035                            |
| 11  | 0.66941  | 0.03734                 | 0.96769                            |
| 12  | 0.62269  | 0.03231                 | 1.00000                            |

Από τη σύγκριση του Πίνακα 6.3 και του Πίνακα 6.4 φαίνεται να εξηγείται από τον ίδιο αριθμό κυρίων συνιστωσών μικρότερο ποσοστό της συνολικής διακύμανσης με χρήση του πίνακα συσχετίσεων.

Στον Πίνακα 6.5 φαίνονται οι συντελεστές για τις κύριες συνιστώσες όπως προέκυψαν από την ανάλυση με τη χρήση του πίνακα διακυμάνσεων.

**Πίνακας 6.5:** Πίνακας με τις κύριες συνιστώσες όπως προκύπτουν από την ανάλυση με τον πίνακα διακυμάνσεων

|               | 1            | 2            | 3            | 4            | 5            | 6            | 7           | 8            | 9            | 10           | 11          | 12          |
|---------------|--------------|--------------|--------------|--------------|--------------|--------------|-------------|--------------|--------------|--------------|-------------|-------------|
| food          | -0.17        | 0.01         | -0.29        | 0.28         | -0.30        | <u>0.76</u>  | -0.20       | -0.24        | -0.14        | 0.13         | -0.12       | -0.04       |
| alcohol       | -0.05        | 0.02         | -0.08        | 0.07         | -0.02        | 0.04         | -0.05       | -0.06        | -0.05        | 0.03         | <u>0.99</u> | -0.05       |
| clothing      | -0.19        | 0.05         | -0.32        | 0.12         | -0.01        | -0.04        | -0.14       | 0.22         | -0.03        | <u>-0.88</u> | -0.02       | -0.01       |
| housing       | <u>-0.68</u> | <u>-0.63</u> | 0.37         | -0.02        | 0.02         | 0.00         | -0.03       | -0.02        | -0.03        | -0.03        | 0.01        | -0.03       |
| equipment     | -0.19        | -0.01        | <u>-0.42</u> | -0.34        | <u>0.78</u>  | 0.18         | -0.03       | -0.11        | 0.07         | 0.11         | -0.02       | 0.01        |
| health        | -0.13        | -0.01        | -0.27        | <u>-0.76</u> | <u>-0.52</u> | -0.10        | -0.05       | -0.18        | 0.11         | 0.00         | 0.01        | 0.00        |
| transport     | <u>-0.52</u> | <u>0.77</u>  | <u>0.36</u>  | -0.07        | 0.03         | 0.04         | -0.02       | -0.01        | 0.01         | 0.01         | -0.01       | -0.01       |
| communication | -0.07        | 0.00         | -0.05        | 0.06         | -0.06        | -0.01        | 0.01        | 0.00         | 0.01         | 0.03         | 0.04        | <u>0.99</u> |
| culture       | -0.17        | 0.04         | -0.26        | -0.02        | -0.09        | -0.1         | <u>0.51</u> | 0.27         | <u>-0.73</u> | 0.15         | -0.03       | -0.03       |
| education     | -0.13        | 0.00         | -0.13        | 0.14         | -0.05        | 0.16         | <u>0.75</u> | 0.03         | <u>0.58</u>  | -0.04        | 0.03        | -0.04       |
| hotels        | -0.25        | 0.05         | <u>-0.37</u> | <u>0.41</u>  | -0.09        | <u>-0.58</u> | -0.08       | <u>-0.49</u> | 0.06         | 0.17         | -0.09       | -0.07       |
| goods         | -0.19        | 0.00         | -0.26        | 0.10         | -0.05        | -0.09        | -0.32       | <u>0.73</u>  | 0.28         | 0.38         | -0.01       | -0.04       |

Δηλαδή για την πρώτη κύρια συνιστώσα ισχύει:

$$Y_1 = -0.17(\text{food}) - 0.05(\text{alcohol}) - 0.19(\text{clothing}) - 0.68(\text{housing}) - 0.19(\text{equipment}) - 0.13(\text{health}) - 0.52(\text{transport}) - 0.07(\text{communication}) - 0.17(\text{culture}) - 0.13(\text{education}) - 0.25(\text{hotels}) - 0.19(\text{goods}).$$

Παρατηρώντας τον Πίνακα 6.5 φαίνεται ότι η ενδέκατη κύρια συνιστώσα σχεδόν ταυτίζεται με την 2η μεταβλητή “alcohol” και η δωδέκατη κύρια συνιστώσα σχεδόν ταυτίζεται με την 8η μεταβλητή “communication”. Επίσης, η δέκατη κύρια συνιστώσα έχει πολύ μεγάλο συντελεστή για την 3η μεταβλητή “clothing” συγκριτικά με τις υπόλοιπες. Άρα, λοιπόν, αν στηριχθεί η ανάλυση στον πίνακα διακυμάνσεων τότε τα αποτελέσματα φαίνεται να μην έχουν κάποιο ενδιαφέρον και αυτό οφείλεται στις σχετικά μεγάλες διαφορές των αρχικών μεταβλητών ως προς τις διακυμάνσεις τους.

Στον Πίνακα 6.6 φαίνονται οι συντελεστές για τις κύριες συνιστώσες όπως προέκυψαν από την ανάλυση με τη χρήση του πίνακα συσχετίσεων.

**Πίνακας 6.6:** Πίνακας με τις κύριες συνιστώσες όπως προκύπτουν από την ανάλυση με τον πίνακα συσχετίσεων

|               | 1            | 2            | 3            | 4            | 5            | 6            | 7            | 8            | 9            | 10           | 11           | 12           |
|---------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
| food          | -0.28        | 0.30         | -0.08        | <u>0.37</u>  | -0.26        | -0.05        | <u>0.39</u>  | <u>-0.56</u> | <u>0.34</u>  | 0.02         | -0.01        | 0.20         |
| alcohol       | -0.21        | <u>0.60</u>  | <u>-0.41</u> | 0.30         | 0.25         | -0.07        | <u>-0.42</u> | 0.15         | -0.20        | 0.18         | -0.03        | -0.06        |
| clothing      | <u>-0.35</u> | 0.00         | -0.17        | -0.18        | 0.02         | 0.14         | <u>0.37</u>  | 0.04         | -0.11        | -0.04        | <u>-0.54</u> | <u>-0.60</u> |
| housing       | -0.29        | -0.15        | 0.24         | -0.06        | <u>-0.47</u> | 0.02         | <u>-0.62</u> | -0.14        | 0.11         | 0.21         | <u>-0.38</u> | 0.03         |
| equipment     | -0.26        | -0.29        | -0.28        | -0.09        | 0.24         | <u>0.67</u>  | -0.20        | <u>-0.36</u> | -0.04        | -0.11        | 0.28         | 0.04         |
| health        | -0.20        | <u>-0.60</u> | <u>-0.41</u> | <u>0.39</u>  | -0.06        | <u>-0.47</u> | -0.05        | 0.04         | -0.13        | -0.21        | 0.02         | -0.01        |
| transport     | -0.26        | 0.03         | 0.13         | <u>-0.50</u> | <u>0.42</u>  | <u>-0.53</u> | -0.07        | <u>-0.43</u> | -0.07        | 0.05         | 0.08         | 0.02         |
| communication | <u>-0.34</u> | 0.16         | 0.21         | -0.03        | -0.27        | -0.08        | -0.11        | 0.18         | 0.16         | <u>-0.33</u> | <u>0.58</u>  | <u>-0.47</u> |
| culture       | -0.31        | -0.20        | 0.10         | 0.11         | 0.39         | 0.02         | 0.07         | <u>0.38</u>  | <u>0.62</u>  | <u>0.38</u>  | 0.02         | 0.06         |
| education     | -0.24        | -0.06        | <u>0.64</u>  | <u>0.47</u>  | 0.25         | 0.11         | 0.09         | -0.04        | <u>-0.47</u> | -0.01        | -0.03        | 0.02         |
| hotels        | <u>-0.35</u> | 0.15         | 0.01         | -0.20        | 0.01         | 0.04         | 0.03         | <u>0.33</u>  | 0.01         | <u>-0.59</u> | -0.23        | <u>0.56</u>  |
| goods         | <u>-0.32</u> | -0.04        | -0.13        | -0.24        | <u>-0.34</u> | 0.02         | 0.28         | 0.21         | -0.41        | 0.51         | 0.29         | 0.25         |

Δηλαδή για την πρώτη κύρια συνιστώσα ισχύει:

$$Y_1 = -0.28(\text{food}) - 0.21(\text{alcohol}) - 0.35(\text{clothing}) - 0.29(\text{housing}) - 0.26(\text{equipment}) \\ - 0.20(\text{health}) - 0.26(\text{transport}) - 0.34(\text{communication}) - 0.31(\text{culture}) - \\ 0.24(\text{education}) - 0.35(\text{hotels}) - 0.32(\text{goods}).$$

Παρατηρώντας τον Πίνακα 6.6 οι μεταβλητές δείχνουν να μπερδεύονται μεταξύ τους γεγονός που σημαίνει πως κάθε κύρια συνιστώσα εξαρτάται από πολλές μεταβλητές. Συνεπώς, ο πίνακας συσχετίσεων μοιάζει να είναι μια καλύτερη επιλογή για την εν λόγω έρευνα.

## 6.2 Αριθμός Κυρίων Συνιστωσών

Ένα άλλο σημαντικό ζήτημα είναι ο αριθμός των κυρίων συνιστωσών που θα επιλεγεί. Με τη βοήθεια του στατιστικού πακέτου R θα εφαρμοσθούν τα παρακάτω 3 κριτήρια προκειμένου να προσδιορισθεί ο αριθμός αυτός.

- Ποσοστό συνολικής διακύμανσης που εξηγούν οι συνιστώσες

Σύμφωνα με το κριτήριο αυτό απαιτούνται 5 κύριες συνιστώσες με βάση τον πίνακα διακυμάνσεων και 8 με βάση τον πίνακα συσχετίσεων ώστε να εξηγείται πάνω από το 80% της συνολικής διακύμανσης. Και στις δύο περιπτώσεις, ο αριθμός των κυρίων συνιστωσών είναι σχετικά μεγάλος. Αν το όριο αυτό, ωστόσο, ελαττωθεί σε 70% τότε θα χρειάζονται 3 με βάση τον πίνακα διακυμάνσεων και 6 με βάση τον πίνακα συσχετίσεων.

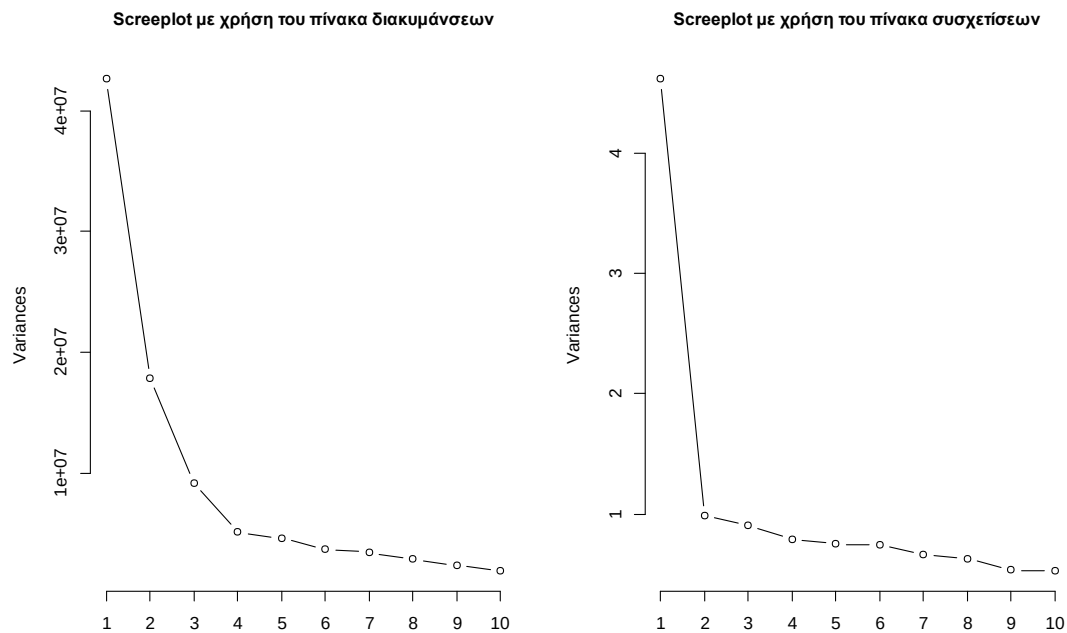
- Κριτήριο Kaiser

Σύμφωνα με το κριτήριο αυτό, με βάση τον πίνακα διακυμάνσεων, ο αριθμός των ιδιοτιμών οι οποίες είναι μεγαλύτερες από  $\bar{\lambda} = \sum_{i=1}^p \lambda_i / p$  είναι 3 και συνεπώς προκύπτουν 3 κύριες συνιστώσες. Ενώ με βάση τον πίνακα συσχετίσεων μόνο μία ιδιοτιμή είναι μεγαλύτερη της 1 και συνεπώς, επιλέγεται 1 κύρια συνιστώσα. Ωστόσο, ο αριθμός αυτός δεν είναι απόλυτα ξεκάθαρος αφού και οι επόμενες δύο ιδιοτιμές είναι πολύ κοντά στη 1.

- Scree plot (Γραφική Μέθοδος)

Στη συνέχεια δίνονται scree plots, μέσω της εντολής `screeplot()`, για τις ιδιοτιμές του πίνακα διακυμάνσεων (αριστερά) και του πίνακα συσχετίσεων (δεξιά). Από το πρώτο γράφημα προκύπτει ότι η κλίση της γραφικής παράστασης αλλάζει και γίνεται περίπου παράλληλη στον οριζόντιο άξονα στην τέταρτη συνιστώσα. Ενώ από το δεύτερο γράφημα προκύπτει ότι η κλίση της γραφικής παράστασης αλλάζει απότομα στη δεύτερη συνιστώσα αλλά δεν μπορεί να θεωρηθεί ακριβώς παράλληλη με τον οριζόντιο άξονα. Φαίνεται μετά την τρίτη συνιστώσα να προσεγγίζει το επίπεδο. Εδώ εμπεριέχεται το υποκειμενικό κριτήριο για τον προσδιορισμό του αριθμού των συνιστωσών από την αλλαγή στην κλίση.

**Διάγραμμα 6.1:** Scree plot για τις ιδιοτιμές του πίνακα διακυμάνσεων και του πίνακα συσχετίσεων



Συμπερασματικά, λοιπόν, δεν υπάρχει συγκεκριμένη απάντηση για τον αριθμό των κυρίων συνιστωσών που πρέπει να χρησιμοποιηθούν. Αυτό ποικίλει ανάλογα με τον πίνακα που χρησιμοποιηθεί στην ανάλυση σε κύριες συνιστώσες αλλά και με το κριτήριο για τον προσδιορισμό του εν λόγω αριθμού.

### 6.3 Ερμηνεία κυρίων Συνιστωσών

Όπως έχει ήδη αναφερθεί ο πίνακας συσχετίσεων φαίνεται να είναι καλύτερη επιλογή για την εν λόγω έρευνα. Οπότε θα δοθεί ερμηνεία για τις κύριες συνιστώσες με βάση τον Πίνακα 6.6. Η πρώτη κύρια συνιστώσα μπορεί να ερμηνευθεί ως ένας σταθμισμένος μέσος όρος των μεταβλητών, αφού όλοι οι συντελεστές τους έχουν το ίδιο πρόσημο και δε φαίνεται να επηρεάζεται ιδιαίτερα από κάποιες συγκεκριμένες μεταβλητές. Η δεύτερη κύρια συνιστώσα έχει μια έντονη επιρροή από τις δύο μεταβλητές “alcohol” και “health”, δηλαδή τα οινόπνευματώδη ποτά – καπνό και την υγεία. Η τρίτη κύρια συνιστώσα επηρεάζεται κατά κύριο λόγο από τις τρεις μεταβλητές “alcohol”, “health” και “education”, δηλαδή τα οινόπνευματώδη ποτά – καπνό, την υγεία και την εκπαίδευση. Με την ίδια λογική και βάσει του Πίνακα 6.6 μπορούν να ερμηνευτούν και οι υπόλοιπες κύριες συνιστώσες.

### 6.4 Περαιτέρω χρήση των αποτελεσμάτων της Ανάλυσης των Κυρίων Συνιστωσών

Αφού γίνει η επιλογή των κυρίων συνιστωσών με τις οποίες χάνεται κάποια πληροφορία, αυτές μπορούν να χρησιμοποιηθούν σε ένα γραμμικό μοντέλο, μέσω της εντολής predict(). Σε αυτό το μοντέλο θα χρησιμοποιηθούν οι τιμές των κυρίων συνιστωσών, οι οποίες υπολογίζονται για κάθε γραμμή των δεδομένων του

δείγματος από το γραμμικό συνδυασμό των τυποποιημένων τιμών των αρχικών μεταβλητών με συντελεστές αυτούς από τις κύριες συνιστώσες με χρήση του πίνακα συσχετίσεων. Στη συνέχεια, δίνονται στον Πίνακα 6.7 τα αποτελέσματα με τη χρήση των 3 πρώτων κυρίων συνιστωσών και με βάση του πίνακα συσχετίσεων, με τις οποίες εξηγείται το 54% της συνολικής διακύμανσης.

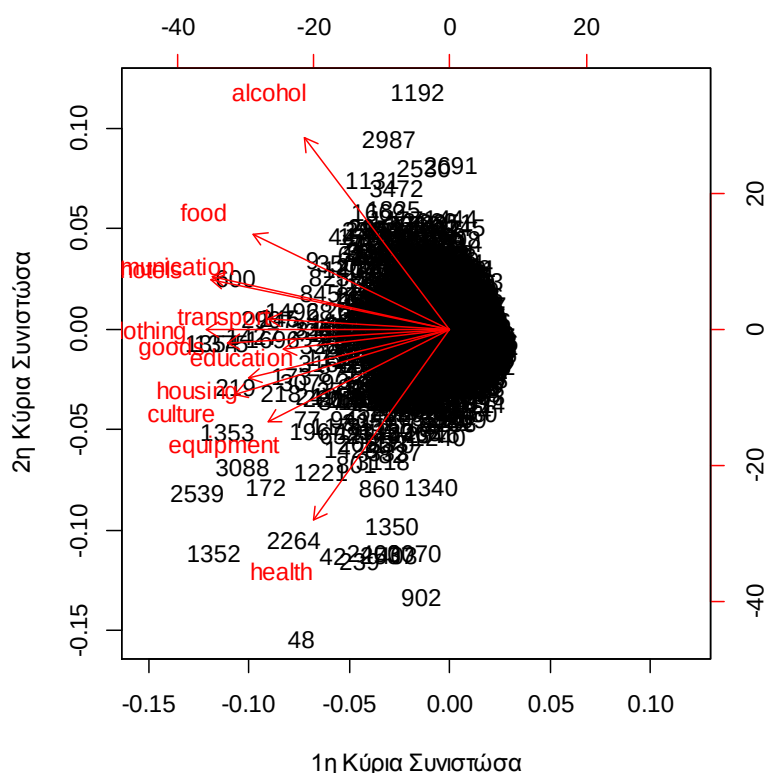
**Πίνακας 6.7:** Πίνακας αποτελεσμάτων του γραμμικού μοντέλου

|      | 1η           | 2η            | 3η            |
|------|--------------|---------------|---------------|
| 1    | -3.145461736 | -1.1751734854 | -9.447423e-01 |
| 2    | 1.954902396  | -0.6909131842 | 1.702631e-01  |
| 3    | -0.473507488 | 0.4522709654  | 3.083551e-02  |
| 4    | -5.305154812 | 1.1278461158  | -2.787583e-02 |
| ...  |              |               |               |
| 3509 | 0.654861336  | 0.4118166342  | 8.996261e-02  |
| 3510 | -0.888005968 | 0.8116567062  | 1.226158e+00  |
| 3511 | 0.797841729  | 0.6684807013  | -2.577721e-01 |
| 3512 | -0.261566768 | 2.0018873139  | -2.361649e-01 |

Τέλος, ένα χρήσιμο γράφημα στην ανάλυση σε κύριες συνιστώσες είναι το biplot, το οποίο κατασκευάζεται με την εντολή biplot(). Πρόκειται για τη γραφική παράσταση των τιμών των κυρίων συνιστωσών και των συντελεστών ταυτόχρονα. Οι μεταβλητές παριστάνονται με βέλη που ξεκινούν από την αρχή των αξόνων. Αν τα βέλη είναι κάθετα μεταξύ τους τότε οι αντίστοιχες μεταβλητές δεν συσχετίζονται ενώ, αντίθετα, αν τα βέλη δείχνουν προς την ίδια (ή αντίθετη) κατεύθυνση τότε οι μεταβλητές συσχετίζονται ισχυρά θετικά (ή αρνητικά).

Επομένως, από το Διάγραμμα 6.2 προκύπτει, για παράδειγμα, ότι η στέγαση και τα οινόπνευματώδη ποτά είναι μεταβλητές που συσχετίζονται λίγο καθώς φαίνονται σχεδόν κάθετα μεταξύ τους. Επίσης, η υγεία και τα είδη διατροφής συσχετίζονται σε μικρό ποσοστό. Αντίθετα, δύο μεταβλητές που φαίνεται να συσχετίζονται αρκετά ισχυρά θετικά, αφού σχεδόν ταυτίζονται, είναι οι επικοινωνίες με τα ξενοδοχεία, καφενεία και εστιατόρια. Δεν υπάρχουν μεταβλητές οι οποίες να έχουν ισχυρή αρνητική συσχέτιση.

Διάγραμμα 6.2: Biplot



Εντολές στην R

```
> data <- dataHH [, 9:20]
> cov (data)
 food alcohol clothing housingequipment health
food 5265780.2 781840.1 2023955.1 3689883.0 1509316.7 1103202.1
alcohol 781840.1 1116656.8 626031.1 951851.3 535913.6 272569.8
clothing 2023955.1 626031.1 4313977.4 3916734.3 2285856.1 1364254.5
housing 3689883.0 951851.3 3916734.3 28075392.4 4339483.3 3047718.0
equipment 1509316.7 535913.6 2285856.1 4339483.3 6766649.1 1579124.4
health 1103202.1 272569.8 1364254.5 3047718.0 1579124.4 5823885.1
transport 2829777.2 1028052.4 3752973.7 7549059.0 2957402.2 2052932.9
communication 679055.9 217378.7 661668.5 1736439.3 514177.5 346028.1
culture 1429847.8 499858.0 1940960.6 3751449.0 1918510.3 1476502.3
education 1287391.7 279137.8 1189924.8 3207261.7 1101753.9 687027.6
hotels 2198884.5 907269.3 2966951.0 5343572.9 2379132.2 1328959.9
goods 1789004.1 537278.9 2370006.7 4501058.8 1852925.2 1384343.1
 transport communication culture education hotels goods
food 2829777 679055.9 1429847.8 1287391.7 2198884.5 1789004.1
alcohol 1028052 217378.7 499858.0 279137.8 907269.3 537278.9
clothing 3752974 661668.5 1940960.6 1189924.8 2966951.0 2370006.7
housing 7549059 1736439.3 3751449.0 3207261.7 5343572.9 4501058.8
equipment 2957402 514177.5 1918510.3 1101753.9 2379132.2 1852925.2
health 2052933 346028.1 1476502.3 687027.6 1328959.9 1384343.1
transport 23249227 1197854.7 3395061.1 2269519.6 4779873.9 3309824.0
communication 1197855 489202.2 604098.5 498853.4 1001917.0 694809.4
culture 3395061 604098.5 4398900.9 1521890.2 2347148.7 1698784.6
education 2269520 498853.4 1521890.2 3892707.7 1595815.0 1027055.7
```

```

hotels 4779874 1001917.0 2347148.7 1595815.0 6867252.5 2582363.5
goods 3309824 694809.4 1698784.6 1027055.7 2582363.5 4668742.5
> cor (data)
 food alcohol clothing housing equipment health
food 1.0000000 0.3224235 0.4246495 0.3034718 0.2528495 0.1992128
alcohol 0.3224235 1.0000000 0.2852312 0.1699990 0.1949612 0.1068837
clothing 0.4246495 0.2852312 1.0000000 0.3558953 0.4230806 0.2721761
housing 0.3034718 0.1699990 0.3558953 1.0000000 0.3148388 0.2383447
equipment 0.2528495 0.1949612 0.4230806 0.3148388 1.0000000 0.2515493
health 0.1992128 0.1068837 0.2721761 0.2383447 0.2515493 1.0000000
transport 0.2557506 0.2017675 0.3747420 0.2954783 0.2357867 0.1764266
communication 0.4230873 0.2941121 0.4554675 0.4685463 0.2826064 0.2050032
culture 0.2970889 0.2255354 0.4455595 0.3375701 0.3516456 0.2917129
education 0.2843501 0.1338853 0.2903720 0.3067931 0.2146703 0.1442919
hotels 0.3656618 0.3276309 0.5451047 0.3848374 0.3490119 0.2101425
goods 0.3608112 0.2353098 0.5280933 0.3931441 0.3296637 0.2654834
 transport communication culture education hotels goods
food 0.2557506 0.4230873 0.2970889 0.2843501 0.3656618 0.3608112
alcohol 0.2017675 0.2941121 0.2255354 0.1338853 0.3276309 0.2353098
clothing 0.3747420 0.4554675 0.4455595 0.2903720 0.5451047 0.5280933
housing 0.2954783 0.4685463 0.3375701 0.3067931 0.3848374 0.3931441
equipment 0.2357867 0.2826064 0.3516456 0.2146703 0.3490119 0.3296637
health 0.1764266 0.2050032 0.2917129 0.1442919 0.2101425 0.2654834
transport 1.0000000 0.3551859 0.3357154 0.2385634 0.3782863 0.3176879
communication 0.3551859 1.0000000 0.4118049 0.3614955 0.5466333 0.4597498
culture 0.3357154 0.4118049 1.0000000 0.3677776 0.4270483 0.3748576
education 0.2385634 0.3614955 0.3677776 1.0000000 0.3086492 0.2409174
hotels 0.3782863 0.5466333 0.4270483 0.3086492 1.0000000 0.4560641
goods 0.3176879 0.4597498 0.3748576 0.2409174 0.4560641 1.0000000
> pc <- prcomp (data)
> summary (pc)
Importance of components:
 PC1 PC2 PC3 PC4 PC5
Standard deviation 6527.2261 4223.5821 3.036e+03 2.264e+03 2147.8081
Proportion of Variance 0.4488 0.1879 9.709e-02 5.399e-02 0.0486
Cumulative Proportion 0.4488 0.6367 7.338e-01 7.878e-01 0.8364
 PC6 PC7 PC8 PC9 PC10
Standard deviation 1929.0889 1.859e+03 1.695e+03 1.551e+03 1378.0342
Proportion of Variance 0.0392 3.641e-02 3.026e-02 2.533e-02 0.0200
Cumulative Proportion 0.8756 9.120e-01 9.423e-01 9.676e-01 0.9876
 PC11 PC12
Standard deviation 956.33571 5.12e+02
Proportion of Variance 0.00963 2.76e-03
Cumulative Proportion 0.99724 1.00e+00
> pc1 <- prcomp (data, scale = T)
> summary (pc1)
Importance of components:
 PC1 PC2 PC3 PC4 PC5 PC6 PC7
Standard deviation 2.1478 0.99361 0.95476 0.89025 0.86778 0.86191 0.81562
Proportion of Variance 0.3844 0.08227 0.07596 0.06605 0.06275 0.06191 0.05544
Cumulative Proportion 0.3844 0.46671 0.54267 0.60871 0.67147 0.73338 0.78881

```

```

 PC8 PC9 PC10 PC11 PC12
Standard deviation 0.79571 0.73310 0.72652 0.66941 0.62269
Proportion of Variance 0.05276 0.04479 0.04399 0.03734 0.03231
Cumulative Proportion 0.84157 0.88636 0.93035 0.96769 1.00000
> pc
Standard deviations:
[1] 6527.2261 4223.5821 3035.9250 2263.9186 2147.8081 1929.0889 1859.0220
[8] 1694.8887 1550.5280 1378.0342 956.3357 511.9691
Rotation:
 PC1 PC2 PC3 PC4 PC5
food -0.16549748 0.010026232 -0.28694878 0.27783223 -0.29731722
alcohol -0.04896361 0.017586983 -0.08167286 0.07095879 -0.02261836
clothing -0.19093474 0.049235411 -0.32354502 0.11752479 -0.01031772
housing -0.68126323 -0.627108217 0.37200321 -0.01564955 0.01934160
equipment -0.19133139 -0.010791650 -0.41683270 -0.33557091 0.78331326
health -0.13202735 -0.011093678 -0.27209216 -0.76433799 -0.51802960
transport -0.51630652 0.774347222 0.35482634 -0.06712086 0.03014213
communication -0.06518606 -0.003088886 -0.05061372 0.05879373 -0.03156843
culture -0.17390896 0.035554410 -0.25637156 -0.02456394 -0.05768117
education -0.12918679 -0.004143120 -0.13259231 0.13827708 -0.09041337
hotels -0.24858298 0.052429487 -0.37326792 0.41158657 -0.05172696
goods -0.18986804 0.002916705 -0.26275275 0.09903216 -0.11222082
 PC6 PC7 PC8 PC9 PC10
food 0.759372927 -0.19759674 -0.236795680 -0.140605996 0.128050619
alcohol 0.035788479 -0.04494321 -0.057313513 -0.053035233 0.026749719
clothing -0.040177206 -0.14212791 0.220016701 -0.027871725 -0.878154986
housing -0.002415660 -0.02692712 -0.023212035 -0.031224079 -0.026577935
equipment 0.180875883 -0.03375121 -0.110065159 0.073624863 0.107896701
health -0.099907163 -0.04626684 -0.179140281 0.107080337 -0.004351411
transport 0.035846208 -0.02257130 -0.014255830 0.013495548 0.013855824
communication -0.008641246 0.01035682 -0.003129288 0.009002555 0.026495093
culture -0.100624992 0.50613295 0.267799152 -0.733407065 0.145612086
education 0.160638742 0.75435630 0.033702330 0.581461557 -0.043328567
hotels -0.576169855 -0.08437640 -0.489561336 0.064124860 0.167995413
goods -0.092156683 -0.31901468 0.731269912 0.280035685 0.383624316
 PC11 PC12
food -0.122407118 -0.040838214
alcohol 0.986133345 -0.050032607
clothing -0.016266859 -0.010018319
housing 0.005012805 -0.025182182
equipment -0.014700261 0.009531626
health 0.011408701 0.003968175
transport -0.007759265 -0.008154893
communication 0.038427877 0.993113667
culture -0.032488567 -0.026038978
education 0.031681372 -0.038014016
hotels -0.090962391 -0.068407178
goods -0.007158724 -0.042941570
> pc1
Standard deviations:
[1] 2.1478370 0.9936143 0.9547563 0.8902502 0.8677761 0.8619116 0.8156216

```



[8] 0.7957102 0.7330978 0.7265153 0.6694084 0.6226856

Rotation:

|               | PC1        | PC2           | PC3          | PC4         | PC5         |
|---------------|------------|---------------|--------------|-------------|-------------|
| food          | -0.2833913 | 2.976813e-01  | -0.084077293 | 0.37027169  | -0.25754502 |
| alcohol       | -0.2094670 | 5.949482e-01  | -0.411594552 | 0.29837499  | 0.25327045  |
| clothing      | -0.3506675 | -8.418986e-05 | -0.167935700 | -0.17914842 | 0.02213840  |
| housing       | -0.2907300 | -1.525653e-01 | 0.242167468  | -0.05959313 | -0.46785933 |
| equipment     | -0.2620494 | -2.867373e-01 | -0.282519624 | -0.09001967 | 0.24388211  |
| health        | -0.1961036 | -5.952188e-01 | -0.405539147 | 0.38621843  | -0.05673659 |
| transport     | -0.2612195 | 3.251742e-02  | 0.133197486  | -0.49593134 | 0.42186631  |
| communication | -0.3418850 | 1.640628e-01  | 0.209353380  | -0.02634445 | -0.26815665 |
| culture       | -0.3099232 | -2.009642e-01 | 0.103107778  | 0.10645783  | 0.39399046  |
| education     | -0.2402294 | -6.214769e-02 | 0.635714326  | 0.47304049  | 0.25373459  |
| hotels        | -0.3450418 | 1.540142e-01  | 0.005034004  | -0.20404477 | 0.01194646  |
| goods         | -0.3213371 | -4.122296e-02 | -0.126627194 | -0.23995993 | -0.34323904 |

|               | PC6         | PC7         | PC8         | PC9         | PC10        |
|---------------|-------------|-------------|-------------|-------------|-------------|
| food          | -0.04581399 | 0.38816736  | -0.55960824 | 0.33595965  | 0.01732010  |
| alcohol       | -0.06653617 | -0.42024965 | 0.14455295  | -0.19936607 | 0.18379206  |
| clothing      | 0.14160248  | 0.36836495  | 0.04263608  | -0.10615410 | -0.03977502 |
| housing       | 0.02351659  | -0.62102477 | -0.14118821 | 0.10672578  | 0.21213278  |
| equipment     | 0.66544099  | -0.20038348 | -0.35518241 | -0.03501165 | -0.11345454 |
| health        | -0.47402390 | -0.04795134 | 0.04321633  | -0.13062402 | -0.21001925 |
| transport     | -0.53340766 | -0.06824912 | -0.43114228 | -0.07144268 | 0.04982661  |
| communication | -0.07641920 | -0.10497529 | 0.18040202  | 0.16338277  | -0.32862740 |
| culture       | 0.01820535  | 0.07336329  | 0.38045688  | 0.62051423  | 0.38288131  |
| education     | 0.11062600  | 0.08631916  | -0.03672732 | -0.47290389 | -0.01185682 |
| hotels        | 0.04414596  | 0.03197671  | 0.32643970  | 0.00731588  | -0.58445389 |
| goods         | 0.02184755  | 0.28171889  | 0.21232323  | -0.40731959 | 0.51337552  |

|               | PC11         | PC12        |
|---------------|--------------|-------------|
| food          | -0.006403825 | 0.20354753  |
| alcohol       | -0.028170552 | -0.05765675 |
| clothing      | -0.539816433 | -0.59534132 |
| housing       | -0.384271692 | 0.02994577  |
| equipment     | 0.277047073  | 0.04171134  |
| health        | 0.018155488  | -0.00680200 |
| transport     | 0.077271727  | 0.02057553  |
| communication | 0.582429277  | -0.46524066 |
| culture       | 0.014478969  | 0.06372871  |
| education     | -0.030843922 | 0.01944747  |
| hotels        | -0.225733342 | 0.55972656  |
| goods         | 0.292750546  | 0.25204039  |

> par (mfrow = c (1, 2))

> screeplot (pc, type = 'line', main = 'Screeplot με χρήση του πίνακα διακυμάνσεων',  
cex.main = 1)

> screeplot (pc1, type = 'line', main = 'Screeplot με χρήση του πίνακα συσχετίσεων',  
cex.main = 1)

> predict (pc1)[, 1:3]

|   | PC1          | PC2           | PC3           |
|---|--------------|---------------|---------------|
| 1 | -3.145461736 | -1.1751734854 | -9.447423e-01 |
| 2 | 1.954902396  | -0.6909131842 | 1.702631e-01  |
| 3 | -0.473507488 | 0.4522709654  | 3.083551e-02  |

```

4 -5.305154812 1.1278461158 -2.787583e-02
5 -5.206586349 -1.3534893421 -7.136567e-02
6 0.410537721 -0.4309143313 2.953375e-01
7 -2.992467664 0.4997338359 8.031528e-01
8 0.575089995 -0.7237295594 -1.937671e-01
9 -8.756021775 2.0772703221 5.603252e-01
10 0.633469070 -0.8937191567 -7.478294e-01
...
3500 1.112890071 0.0907583185 -4.434115e-03
3501 1.199007818 0.0364952594 -1.240014e-02
3502 1.403216575 -0.8736236884 -3.579503e-01
3503 -2.149541428 0.5984957400 1.347674e+00
3504 1.329201014 -0.7264199333 -5.444461e-01
3505 -6.747207194 2.0015700680 3.728164e+00
3506 1.478606096 -0.1074833493 -2.038972e-01
3507 0.348644319 0.3344219905 -1.060730e-01
3508 1.780989329 -1.8826362170 -7.914678e-01
3509 0.654861336 0.4118166342 8.996261e-02
3510 -0.888005968 0.8116567062 1.226158e+00
3511 0.797841729 0.6684807013 -2.577721e-01
3512 -0.261566768 2.0018873139 -2.361649e-01
> biplot (pc1, xlab = '1η Κύρια Συνιστώσα', ylab = '2η Κύρια Συνιστώσα')

```

## **ΒΙΒΛΙΟΓΡΑΦΙΑ**

### Ελληνική

Καρλής Δ. (2004). *Πολυμεταβλητές Τεχνικές*. Εκδόσεις Οικονομικού Πανεπιστημίου Αθηνών.

Κοκολάκης Γ. και Σπηλιώτης Ι. (2002). *Εισαγωγή στις Πιθανότητες*. Εκδόσεις Συμεών.

Κοκολάκης Γ.Ε. και Φουσκάκης Δ.Σ. (2005). *Σημειώσεις Στατιστικής*. Έκδοση Εθνικού Μετσοβίου Πολυτεχνείου.

Οικονόμου Πολυχρόνης και Καρώνη Χρυσής (2009). *Ανάλυση Παλινδρόμησης*. Έκδοση Εθνικού Μετσοβίου Πολυτεχνείου.

Οικονόμου Π. και Καρώνη Χ. (2010). *Στατιστικά Μοντέλα Παλινδρόμησης*. Εκδόσεις Συμεών.

Τσάντας Νίκος, Μωυσιάδης Χρόνης, Μπαγιάτης Ντίνος και Χατζηπαντελής Θόδωρος (1999). *Ανάλυση Δεδομένων με τη βοήθεια στατιστικών πακέτων*. Εκδόσεις Ζήτη.

Φουσκάκης Δ. (2013). *Ανάλυση Δεδομένων με χρήση της R*. Εκδόσεις Τσότρας.

Norman Draper and Harry Smith (1997). *Εφαρμοσμένη Ανάλυση Παλινδρόμησης* (μετάφραση). Εκδόσεις Παπαζήση.

### Ξένα

Andrew Gelman and Jennifer Hill (2006). *Data Analysis Using Regression and Multilevel / Hierarchical Models*. Cambridge University Press.

Barnett V. (1981). *Interpreting Multivariate Data*. Wiley.

Everitt. B.S. (2005). *An R and S – plus companion for multivariate analysis*. Springer.

Johnson D.E. (1998). *Applied Multivariate Methods for Data Analysis*. Duxbury Press.

Johnson R.A. and Wichern D.W. (2002). *Applied Multivariate Statistical Analysis*. Prentice Hall.

Rencher C.A. (2002). *Methods of Multivariate Analysis*. John Wiley and Sons, Inc.

Διαδικτυακές Πηγές

<http://cran.r-project.org/doc/contrib/Faraway-PRA.pdf>

[http://cran.r-project.org/web/packages/HSAUR/vignettes/Ch\\_analysis\\_of\\_variance.pdf](http://cran.r-project.org/web/packages/HSAUR/vignettes/Ch_analysis_of_variance.pdf)

<http://www.statmethods.net/index.html>