



ΠΑΝΕΠΙΣΤΗΜΙΟ ΙΩΑΝΝΙΝΩΝ



ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
**ΜΗΧΑΝΙΚΩΝ Η/Υ
& ΔΙΚΤΥΩΝ**

**ΠΑΝΕΠΙΣΤΗΜΙΟ ΙΩΑΝΝΙΝΩΝ
ΣΧΟΛΗ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ
ΠΜΣ ΜΗΧΑΝΙΚΩΝ Η/Υ ΚΑΙ ΔΙΚΤΥΩΝ**

ΜΕΤΑΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

**ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ PANEL ΜΕ ΧΡΗΣΗ ΤΗΣ
ΓΛΩΣΣΑΣ ΠΡΟΓΡΑΜΜΑΤΙΣΜΟΥ R**

Αναστασία Αντωνίου

Επιβλέπουσα: Φουτσιτζή Γεωργία

Καθηγήτρια

Άρτα, Μάρτιος, 2021



ΠΑΝΕΠΙΣΤΗΜΙΟ ΙΩΑΝΝΙΝΩΝ

**ΠΑΝΕΠΙΣΤΗΜΙΟ ΙΩΑΝΝΙΝΩΝ
ΣΧΟΛΗ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ
ΠΜΣ ΜΗΧΑΝΙΚΩΝ Η/Υ ΚΑΙ ΔΙΚΤΥΩΝ**

ΜΕΤΑΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

**ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ PANEL ΜΕ ΧΡΗΣΗ ΤΗΣ
ΓΛΩΣΣΑΣ ΠΡΟΓΡΑΜΜΑΤΙΣΜΟΥ R**

Αναστασία Αντωνίου

Επιβλέπουσα: Φουτσιτζή Γεωργία

Καθηγήτρια

Άρτα, Μάρτιος, 2021

PANEL DATA ANALYSIS USING R

Εγκρίθηκε από τριμελή εξεταστική επιτροπή

Άρτα, 2021

ΕΠΙΤΡΟΠΗ ΑΞΙΟΛΟΓΗΣΗΣ

1. Επιβλέπουσα καθηγήτρια

Φουτσιτζή Γεωργία,

Καθηγήτρια

2. Μέλος επιτροπής

Αντωνιάδης Νικόλαος,

Καθηγητής

3. Μέλος επιτροπής

Γκόγκος Χρήστος,

Αναπληρωτής Καθηγητής

Ο Διευθυντής του ΠΜΣ

Στύλιος Χρυσόστομος,

Καθηγητής

Υπογραφή

© Αντωνίου, Αναστασία, 2021.

Με επιφύλαξη παντός δικαιώματος. All rights reserved.

Δήλωση μη λογοκλοπής

Δηλώνω υπεύθυνα και γνωρίζοντας τις κυρώσεις του Ν. 2121/1993 περί Πνευματικής Ιδιοκτησίας, ότι η παρούσα μεταπτυχιακή εργασία είναι εξ ολοκλήρου αποτέλεσμα δικής μου ερευνητικής εργασίας, δεν αποτελεί προϊόν αντιγραφής ούτε προέρχεται από ανάθεση σε τρίτους. Όλες οι πηγές που χρησιμοποιήθηκαν (κάθε είδους, μορφής και προέλευσης) για τη συγγραφή της περιλαμβάνονται στη βιβλιογραφία.

Αντωνίου, Αναστασία

Υπογραφή

ΕΥΧΑΡΙΣΤΙΕΣ

Στο σημείο αυτό, θα ήθελα να ευχαριστήσω όλους όσους συνέβαλαν στην ολοκλήρωση της μεταπτυχιακής διπλωματικής μου εργασίας.

Ευχαριστώ θερμά την επιβλέπουσα της διπλωματικής, Καθηγήτρια κ. Φουτσιτζή Γεωργία, για την αδιάλειπτη επίβλεψη της εργασίας, την επιστημονική της καθοδήγηση, τις υποδείξεις της, την επιμονή της, την αμέριστη υποστήριξη και την υπομονή που έδειξε από την αρχή μέχρι το τέλος.

Επίσης, ευχαριστώ τους καθηγητές της εξεταστικής επιτροπής, τον κ. Αντωνιάδη Νικόλαο καθώς επίσης και τον κ. Γκόγκο Χρήστο που δέχτηκαν να είναι μέλη της επιτροπής.

Επιπλέον, θα ήθελα να ευχαριστήσω την κ. Μαργαρίτη Σπυριδούλα που με παρότρυνε και με ενθάρρυνε να ακολουθήσω το συγκεκριμένο πρόγραμμα μεταπτυχιακών σπουδών ώστε να αποκτήσω επιπλέον γνώσεις σχετικά με το αντικείμενο των σπουδών μου.

Τέλος θα ήθελα να ευχαριστήσω τους δικούς μου ανθρώπους, τους γονείς μου, τον αδερφό μου καθώς και τους φίλους μου που με υποστήριζαν να συνεχίσω και ήταν δίπλα μου όλο αυτόν τον καιρό και μου παρείχαν κυρίως ηθική υποστήριξη ώστε να φτάσω στο τέλος αυτής της σημαντικής προσπάθειας.

ΠΕΡΙΛΗΨΗ

Τα τελευταία χρόνια ο όγκος και η πολυπλοκότητα των δεδομένων που συλλέγουμε και καλούμαστε να αναλύσουμε αυξάνονται με εκρηκτικούς ρυθμούς. Η ιδέα της άντλησης χρήσιμων γνώσεων από μεγάλους όγκους δεδομένων είναι κοινή σε πολλούς κλάδους της επιστήμης, από τη Στατιστική έως τη Φυσική, από την Οικονομετρία έως την Αναγνώριση Προτύπων και τον Προσαρμοστικό Έλεγχο. Η ανάλυση δεδομένων είναι η διαδικασία συλλογής και οργάνωσης δεδομένων προκειμένου να εξαχθούν χρήσιμα συμπεράσματα από αυτήν. Η επιστήμη της ανάλυσης δεδομένων χρησιμοποιεί αναλυτικό και λογικό συλλογισμό για να αποκτήσει πληροφορίες από τα δεδομένα. Ο κύριος σκοπός της ανάλυσης δεδομένων είναι να μετατρέψει την πληροφορία σε χρήσιμη γνώση, έτσι ώστε η λαμβανόμενη γνώση να μπορεί να χρησιμοποιηθεί για τη λήψη τεκμηριωμένων αποφάσεων.

Από την άλλη πλευρά, η στατιστική γλώσσα R διαθέτει μια σειρά από τις πιο σύγχρονες τεχνικές για τη στατιστική ανάλυση δεδομένων. Η R μπορεί να χειριστεί μεγάλο όγκο και διαφορετικούς τύπους δεδομένων και δίνει τη δυνατότητα δημιουργίας υψηλής ποιότητας γραφικών. Είναι μια γλώσσα ανοικτού κώδικα που ενσωματώνει συνεχώς νέες βιβλιοθήκες (και κατά συνέπεια νέες μεθοδολογίες ανάλυσης) με αποτέλεσμα να εξελίσσεται σε δυνατότητες και εύρος εφαρμογών. Τέλος, υπάρχει ένα τεράστιο δίκτυο έμπειρων χρηστών που μπορεί να παρέχει υψηλού επιπέδου υποστήριξη.

Σε αυτήν την εργασία ασχολούμαστε με την ανάλυση δεδομένων πάνελ (panel data) με χρήση της R. Τα δεδομένα πάνελ παρέχουν πληροφορίες για μεγέθη που αφορούν τόσο σε επιμέρους οντότητες (επιχειρήσεις, άτομα, γεωγραφικές περιοχές κλπ) όσο και στην χρονική εξέλιξή τους. Δηλαδή, τα δεδομένα panel έχουν δύο διαστάσεις: τη διαστρωματική (cross-sectional) διάσταση και τη διάσταση της χρονολογικής σειράς (time-series). Για το λόγο αυτό, η εφαρμογή μεθόδων παλινδρόμησης σε δεδομένα panel είναι πιο πολύπλοκη από αυτή των απλών διαστρωματικών δεδομένων. Μια ποικιλία μεθόδων εκτίμησης για τα μοντέλα δεδομένων panel έχουν αναπτυχθεί στη βιβλιογραφία που επιτρέπουν την χαλάρωση πολλών περιοριστικών παραδοχών του μοντέλου διαστρωματικών δεδομένων, και οδηγούν σε εναλλακτικές προϋποθέσεις για την αποτελεσματικότητα (efficiency) τους. Οι μέθοδοι αυτές περιλαμβάνουν το μοντέλο ενιαίων δεδομένων (pooled model), το μοντέλο σταθερών επιδράσεων (Fixed Effect model) και μοντέλο τυχαίων επιδράσεων (Random Effect Model).

Αναλυτικότερα, η τρέχουσα μελέτη αποτελείται από πέντε κεφάλαια. Κατ' αρχάς, το πρώτο κεφάλαιο εξετάζει ορισμένες βασικές έννοιες της γλώσσας προγραμματισμού R αλλά και ορισμένων πλεονεκτημάτων ως προς τη χρήση της γλώσσας αυτής. Εν συνεχεία, στο δεύτερο κεφάλαιο, αναλύονται τα υποδείγματα εκείνα που αφορούν σε δεδομένα πάνελ. Ειδικότερα, επεξηγούνται τα εξής ουσιώδη στοιχεία: α) η έννοια των πάνελ δεδομένων, β) τα πλεονεκτήματα και μειονεκτήματα αυτών, γ) η εκτίμηση των μοντέλων παλινδρόμησης δεδομένων πάνελ. Έπειτα, στο τρίτο κεφάλαιο δίνονται οι κυριότερες εντολές που διαθέτει η γλώσσα προγραμματισμού R και ειδικότερα το πακέτο `plm`, για τη διαχείριση και την ανάλυση δεδομένων πάνελ. Στο τέταρτο κεφάλαιο, πραγματοποιείται μια εμπειρική ανάλυση για τη μελέτη των παραγόντων που επηρεάζουν τις Εξαγωγές Υψηλής Τεχνολογίας ως μελέτη περίπτωσης. Οι Εξαγωγές Υψηλής Τεχνολογίας (High Technology Exports HTX) θεωρούνται σημαντικός παράγοντας για τη βιώσιμη οικονομική ανάπτυξη μιας χώρας. Για το λόγο αυτό, η ανάγκη διερεύνησης των πιθανών προσδιοριστικών παραγόντων που επηρεάζουν τις εξαγωγές υψηλής τεχνολογίας σε διάφορες χώρες του πλανήτη έχει τραβήξει το ενδιαφέρον των ερευνητών κατά τα τελευταία χρόνια. Για τη διερεύνηση των παραγόντων αυτών προβήκαμε στην ανάλυση δεδομένων πάνελ από 15 χώρες της Ευρωπαϊκής Ένωσης για το χρονικό διάστημα 2007-2018. Τέλος, στο πέμπτο κεφάλαιο παρουσιάζονται τα συμπεράσματα που προκύπτουν από τη συγκεκριμένη μελέτη.

Λέξεις-κλειδιά: δεδομένα panel, γλώσσα R, στατιστική ανάλυση

ABSTRACT

In recent years, the volume and complexity of the data we collect and are called to analyze are increasing at an explosive rate. The idea of extracting useful knowledge from large volumes of data is common in many disciplines of science, from Statistics to Physics, Econometrics to Pattern Recognition and Adaptive Control. Data analysis is the process of collecting and organizing data in order to draw useful conclusions from it. The science of data analysis uses analytical and logical reasoning to obtain information from data. The main purpose of data analysis is to turn information into useful knowledge so that the knowledge gained can be used to make informed decisions.

Statistical language R, on the other hand, has several state-of-the-art techniques for statistical data analysis. R can handle large volumes and different types of data and enables the creation of high-quality graphics. It is an open source language that constantly integrates new libraries (and consequently new analysis methodologies) thus evolving in capabilities and range of applications. Finally, there is a huge network of experienced users who can provide a high level of support for R.

In this work we deal with the analysis of panel data using R. Panel data provides information on sizes that relate to both entities (businesses, individuals, geographical areas, etc.) and their evolution over time. That is, the panel data has two dimensions: the cross-sectional dimension and the time-series dimension. For this reason, the application of regression methods to panel data is more complex than that of simple layered data. A variety of estimation methods for panel data models have been developed in the literature that allow for the relaxation of many restrictive assumptions of the single stratified data model, and lead to alternative efficiency measures. These methods include the pooled model, the Fixed Effect model, and the Random Effect Model.

More specifically, the current study consists of five chapters. First, the first chapter examines some of the key concepts of the R programming language but also some of the advantages of using this language. Then, in the second chapter, those models based on data panels are analyzed. In particular, the following essential elements are explained: a) the concept of data panels, b) their advantages and disadvantages, c) the evaluation of data panel regression models. Then, in the third chapter the main command sand in particular the plm package, available in the R for the management and analysis of panel data are given. In addition, in the fourth chapter, an empirical analysis is carried out to analyze the determinants of High Technology Exports. High Technology Exports are considered an

important factor for the sustainable economic development of a country. For this reason, the need to investigate the possible determinants that affect high-tech exports to various countries around the world has attract researcher's interest in recent years. To investigate these determinants, we analyzed panel data from 15 countries of the European Union for the period 2007-2018. Finally, the fifth chapter presents the conclusions that emerge from this study.

Keywords: panel data, R, statistics analysis

Πίνακας περιεχομένων

ΕΥΧΑΡΙΣΤΙΕΣ.....	vii
ΠΕΡΙΛΗΨΗ.....	viii
ABSTRACT	x
ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ	xiv
ΚΑΤΑΛΟΓΟΣ ΕΙΚΟΝΩΝ.....	1
ΚΑΤΑΛΟΓΟΣ ΔΙΑΓΡΑΜΜΑΤΩΝ.....	2
1 Η ΓΛΩΣΣΑ ΠΡΟΓΡΑΜΜΑΤΙΣΜΟΥ R	2
1.1 ΕΙΣΑΓΩΓΗ.....	2
1.2 ΤΙ ΕΙΝΑΙ Η R;	3
1.3 ΙΣΤΟΡΙΚΗ ΑΝΑΔΡΟΜΗ ΤΗΣ R	3
1.4 ΤΟ ΠΕΡΙΒΑΛΛΟΝ ΤΗΣ R.....	5
1.5 ΒΑΣΙΚΕΣ ΕΝΤΟΛΕΣ – ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ ΤΗΣ R.....	7
1.6 ΤΕΛΕΣΤΕΣ ΣΤΗΝ R	9
1.7 ΤΕΛΕΣΤΕΣ ΠΡΑΞΕΩΝ ΣΤΗΝ R.....	9
1.7.1 Αριθμητικοί Τελεστές.....	9
1.7.2 Τελεστές σύγκρισης.....	10
1.7.3 Λογικοί Τελεστές.....	12
1.7.4 Σειρά προτεραιότητας τελεστών.....	13
1.8 ΤΥΠΟΙ ΚΑΙ ΔΟΜΕΣ ΔΕΔΟΜΕΝΩΝ	13
1.8.1 ΔΙΑΝΥΣΜΑΤΑ.....	13
1.8.2 ΠΙΝΑΚΕΣ.....	15
1.8.3 ΠΛΑΙΣΙΑ ΔΕΔΟΜΕΝΩΝ.....	24
1.9 ΣΥΝΑΡΤΗΣΕΙΣ ΣΤΗΝ R	25
1.10 ΑΠΟΘΗΚΕΥΣΗ ΑΝΤΙΚΕΙΜΕΝΩΝ ΣΤΗΝ R.....	27
2 ΔΕΔΟΜΕΝΑ PANEL	29
2.1 ΕΙΣΑΓΩΓΗ.....	29
2.2 ΟΡΙΣΜΟΣ ΚΑΙ ΠΕΡΙΓΡΑΦΗ ΔΕΔΟΜΕΝΩΝ PANEL.....	29
2.3 ΠΛΕΟΝΕΚΤΗΜΑΤΑ ΚΑΙ ΜΕΙΟΝΕΚΤΗΜΑΤΑ	32
2.4 ΜΟΝΤΕΛΑ ΔΕΔΟΜΕΝΩΝ PANEL	34
2.4.1 Το Ενιαίο Μοντέλο (Pooled Model).....	35
2.4.2 Το Μοντέλο Σταθερών Επιδράσεων (Fixed Effect Model).....	36
2.5 Το Μοντέλο Τυχαίων Επιδράσεων (Random Effect Model).....	41
2.6 ΈΛΕΓΧΟΙ ΚΑΤΑΛΛΗΛΟΤΗΤΑΣ ΤΩΝ ΜΟΝΤΕΛΩΝ	45
2.6.1 Μοντέλο Σταθερών ή Τυχαίων Επιδράσεων;	45

2.6.2	Μοντέλο Τυχαίων Επιδράσεων ή Ενιαίο Μοντέλο;	46
3	ΑΝΑΛΥΣΗΣ ΔΕΔΟΜΕΝΩΝ PANEL ME THN R.....	47
3.1	ΕΙΣΑΓΩΓΗ.....	47
3.2	ΦΟΡΤΩΣΗ ΤΩΝ ΚΑΤΑΛΛΗΛΩΝ ΠΑΚΕΤΩΝ.....	47
3.3	ΟΡΓΑΝΩΝΟΝΤΑΣ ΤΑ ΔΕΔΟΜΕΝΑ ΩΣ ΠΑΝΕΛ	47
3.4	ΔΙΕΡΕΥΝΗΤΙΚΗ ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ	49
3.4.1	Ετερογένεια μεταξύ των χωρών	50
3.4.2	Ετερογένεια με την πάροδο των ετών.....	51
3.5	ΕΚΤΙΜΗΣΗ ΜΟΝΤΕΛΩΝ ΔΕΔΟΜΕΝΩΝ ΠΑΝΕΛ ME ΧΡΗΣΗ ΤΗΣ R	52
3.5.1	Το Ενιαίο Μοντέλο.....	52
3.5.2	Το Μοντέλο Σταθερών Επιδράσεων.....	54
3.5.3	Το Μοντέλο Τυχαίων Επιδράσεων.....	55
3.6	ΈΛΕΓΧΟΣ ΚΑΤΑΛΛΗΛΟΤΗΤΑΣ ME ΧΡΗΣΗ ΤΗΣ R	57
3.6.1	Μοντέλο Σταθερών ή Τυχαίων Επιδράσεων;	57
3.6.2	Μοντέλο Τυχαίων Επιδράσεων ή Ενιαίο Μοντέλο;	58
3.7	ΈΛΕΓΧΟΣ ΕΤΕΡΟΣΚΕΔΑΣΤΙΚΟΤΗΤΑΣ.....	58
4	ΜΕΛΕΤΗ ΠΕΡΙΠΤΩΣΗΣ ΑΝΑΛΥΣΗΣ ΔΕΔΟΜΕΝΩΝ ΠΑΝΕΛ: Παράγοντες Επίδρασης Εξαγωγών Υψηλής Τεχνολογίας.....	61
4.1	ΕΙΣΑΓΩΓΗ.....	61
4.2	ΤΑ ΔΕΔΟΜΕΝΑ.....	62
4.3	ΠΕΡΙΓΡΑΦΙΚΗ ΑΝΑΛΥΣΗ ΕΞΑΓΩΓΩΝ ΥΨΗΛΗΣ ΤΕΧΝΟΛΟΓΙΑΣ	63
4.4	ΑΝΑΛΥΣΗ ΠΑΛΙΝΔΡΟΜΗΣΗΣ ΤΩΝ ΕΞΑΓΩΓΩΝ ΥΨΗΛΗΣ ΤΕΧΝΟΛΟΓΙΑΣ (ΕΚΤΙΜΗΣΗ-ΕΡΜΗΝΕΙΑ-ΣΥΜΠΕΡΑΣΜΑΤΑ)	75
5	ΕΠΙΛΟΓΟΣ – ΣΥΜΠΕΡΑΣΜΑΤΑ	86
6	ΠΑΡΑΡΤΗΜΑ ΑΠΟΤΕΛΕΣΜΑΤΩΝ.....	90
	ΒΙΒΛΙΟΓΡΑΦΙΑ	95

ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ

Πίνακας 1.5.1: Αριθμητικοί τελεστές στην R	10
Πίνακας 1.5.2: Τελεστές σύγκρισης στην R	11
Πίνακας 1.5.3: Πίνακας αληθείας για τα X και Y	13
Πίνακας 1.5.4: Σειρά προτεραιότητας διαφόρων τελεστών.....	13
Πίνακας 1.6.1: Ειδικές συναρτήσεις και τελεστές για πράξεις με πίνακες	21
Πίνακας 2.5.1: Παράδειγμα τιμών όρων διαταραχής σε μοντέλο τυχαίων επιδράσεων.....	42
Πίνακας 3.3.1: Δείγμα των δεδομένων	49
Πίνακας 4.3.1: Συνδιακύμανση μεταβλητών του υποδείγματος εξαγωγών υψηλής τεχνολογίας ...	72
Πίνακας 4.3.2: Γραφική μήτρα συσχέτισης μεταβλητών του υποδείγματος εξαγωγών υψηλής τεχνολογίας	72
Πίνακας 4.3.3: Βασικές στατιστικές των μεταβλητών του υποδείγματος εξαγωγών υψηλής τεχνολογίας	73
Πίνακας 4.4.1: Αποτελέσματα παλινδρόμησης Ενιαίου Μοντέλου.....	76
Πίνακας 4.4.2: Αποτελέσματα παλινδρόμησης με τη μέθοδο των Σταθερών Επιδράσεων.....	77
Πίνακας 4.4.3: Αποτελέσματα παλινδρόμησης με τη μέθοδο των Τυχαίων Επιδράσεων.....	79
Πίνακας 4.4.4: Έλεγχος Hausman καταλληλότητας μοντέλου	81
Πίνακας 4.4.5: Έλεγχος Breusch-Pagan Lagrangian Multiplier καταλληλότητας του μοντέλου	81
Πίνακας 4.4.6: Έλεγχος ετεροσκεδαστικότητας	82

ΚΑΤΑΛΟΓΟΣ ΕΙΚΟΝΩΝ

Εικόνα 1.1: Το RStudio περιβάλλον ανάπτυξης της R	6
Εικόνα 1.2: Παράδειγμα δημιουργίας διανύσματος με την εντολή <code>c</code>	14
Εικόνα 1.3: Παράδειγμα δημιουργίας διανύσματος με τη συνάρτηση <code>vector</code>	14
Εικόνα 1.4: Παράδειγμα δημιουργίας πίνακα	15
Εικόνα 1.5: Παράδειγμα ανακύκλωσης στην R	17
Εικόνα 1.6: Παράδειγμα δημιουργίας διαγώνιου πίνακα	17
Εικόνα 1.7: Πράξεις σε πίνακες	18
Εικόνα 1.8: Παράδειγμα επιμέρους στοιχείων πίνακα	19
Εικόνα 1.9: Παράδειγμα εντολής <code>diag()</code>	20
Εικόνα 1.10: Παράδειγμα με πράξεις μεταξύ πινάκων	21
Εικόνα 1.11: Παραδείγματα χρήσης συναρτήσεων στην R	22
Εικόνα 1.12: Παράδειγμα της εντολής <code>is.matrix</code>	23
Εικόνα 1.13: Παράδειγμα δημιουργίας πολυδιάστατου πίνακα	23
Εικόνα 1.14: Δημιουργία πολυδιάστατου πίνακα με χρήση της εντολής <code>dim()</code>	24
Εικόνα 1.15: Παράδειγμα δημιουργίας πλαισίου δεδομένων (data frame)	25
Εικόνα 1.16: Ορισμός συνάρτησης από τον χρήστη	26
Εικόνα 1.17: Πέρασμα ορισμάτων με διαφορετική σειρά	26
Εικόνα 1.18: Χρήση του ορίσματος ... σε συνάρτηση	27
Εικόνα 2.1: Παράδειγμα δεδομένων Panel	31

ΚΑΤΑΛΟΓΟΣ ΔΙΑΓΡΑΜΜΑΤΩΝ

Διάγραμμα 4.1: Διαχρονική μεταβολή των εξαγωγών υψηλής τεχνολογίας (HTX) για τις ΕΕ-15 χώρες	63
Διάγραμμα 4.2: Ιστόγραμμα εξαγωγών υψηλής τεχνολογίας	64
Διάγραμμα 4.3: Ιστόγραμμα Δαπανών Έρευνας και Ανάπτυξης.....	65
Διάγραμμα 4.4: Ιστόγραμμα εξαγωγής αγαθών Τεχνολογίας Πληροφορικής και Τηλεπικοινωνιών	65
Διάγραμμα 4.5: Ιστόγραμμα Δείκτη Πραγματικής Συναλλαγματικής Ισοτιμίας.....	65
Διάγραμμα 4.6: Ιστόγραμμα Εργατικού δυναμικού με προχωρημένη εκπαίδευση, γυναίκα	66
Διάγραμμα 4.7: Ιστόγραμμα Δείκτη Πρόσβασης Χρηματοπιστωτικών Ιδρυμάτων.....	66
Διάγραμμα 4.8: Ιστόγραμμα Δείκτη Ανθρώπινης Ανάπτυξης.....	66
Διάγραμμα 4.9: Ιστόγραμμα φυσικού λογαρίθμου εξαγωγών υψηλής τεχνολογίας.....	67
Διάγραμμα 4.10: Ιστόγραμμα φυσικού λογαρίθμου Δαπανών για Έρευνα και Ανάπτυξη.....	68
Διάγραμμα 4.11: Ιστόγραμμα φυσικού λογαρίθμου εξαγωγής αγαθών Τεχνολογίας Πληροφορικής και Τηλεπικοινωνιών.....	68
Διάγραμμα 4.12: Ιστόγραμμα φυσικού λογαρίθμου Δείκτη Πραγματικής Συναλλαγματικής Ισοτιμίας.....	69
Διάγραμμα 4.13: Ιστόγραμμα φυσικού λογαρίθμου Εργατικού δυναμικού με προχωρημένη εκπαίδευση, γυναίκα	69
Διάγραμμα 4.14: Ιστόγραμμα φυσικού λογαρίθμου Δείκτη Πρόσβασης Χρηματοπιστωτικών Ιδρυμάτων	69
Διάγραμμα 4.15: Ιστόγραμμα φυσικού λογαρίθμου Δείκτη Ανθρώπινης Ανάπτυξης.....	70
Διάγραμμα 4.16: Έλεγχος ύπαρξης ετερογένειας μεταξύ διαστρωματικών στοιχείων (εδώ χωρών) στο υπόδειγμα εξαγωγές υψηλής τεχνολογίας	74
Διάγραμμα 4.17: Έλεγχος ύπαρξης ετερογένειας διαχρονικά στο υπόδειγμα εξαγωγές υψηλής τεχνολογίας	74
Διάγραμμα 6.1:Θηκόγραμμα εξαγωγών υψηλής τεχνολογίας.....	90
Διάγραμμα 6.2:Θηκόγραμμα φυσικού λογαρίθμου εξαγωγών υψηλής τεχνολογίας.....	90
Διάγραμμα 6.3:Θηκόγραμμα Δαπανών για Έρευνα και Ανάπτυξη.....	90
Διάγραμμα 6.4:Θηκόγραμμα φυσικού λογαρίθμου Δαπανών για Έρευνα και Ανάπτυξη.....	91
Διάγραμμα 6.5: Θηκόγραμμα εξαγωγής αγαθών Τεχνολογίας Πληροφορικής και Τηλεπικοινωνιών	91
Διάγραμμα 6.6: Θηκόγραμμα φυσικού λογαρίθμου εξαγωγής αγαθών Τεχνολογίας Πληροφορικής και Τηλεπικοινωνιών.....	91
Διάγραμμα 6.7: Θηκόγραμμα Δείκτη Πραγματικής Συναλλαγματικής.....	92
Διάγραμμα 6.8: Θηκόγραμμα λογαρίθμου Δείκτη Πραγματικής Συναλλαγματικής.....	92
Διάγραμμα 6.9: Θηκόγραμμα Εργατικού δυναμικού με προχωρημένη εκπαίδευση, γυναίκα	92
Διάγραμμα 6.10: Θηκόγραμμα φυσικού λογαρίθμου Εργατικού δυναμικού με προχωρημένη εκπαίδευση, γυναίκα	93
Διάγραμμα 6.11: Θηκόγραμμα Δείκτη Πρόσβασης Χρηματοπιστωτικών Ιδρυμάτων.....	93
Διάγραμμα 6.12: Θηκόγραμμα φυσικού λογαρίθμου Δείκτη Πρόσβασης Χρηματοπιστωτικών Ιδρυμάτων	93
Διάγραμμα 6.13: Θηκόγραμμα Δείκτη Ανθρώπινης Ανάπτυξης.....	94
Διάγραμμα 6.14: Θηκόγραμμα φυσικού λογαρίθμου Δείκτη Ανθρώπινης Ανάπτυξης.....	94

ΚΕΦΑΛΑΙΟ 1

1 Η ΓΛΩΣΣΑ ΠΡΟΓΡΑΜΜΑΤΙΣΜΟΥ R

1.1 ΕΙΣΑΓΩΓΗ

Η στατιστική είναι μια επιστήμη με ραγδαία ανάπτυξη τα τελευταία χρόνια. Η ανάπτυξη αυτή οφείλεται σε μεγάλο βαθμό στη ραγδαία εξέλιξη των ηλεκτρονικών υπολογιστών που με τη σειρά της έφερε και την ανάπτυξη στατιστικών πακέτων, ικανών να αναλύσουν δεδομένα μεγάλης κλίμακας.

Η στατιστική άνθισε κυρίως το πρώτο μισό του 20^{ου} αιώνα, μέσα από μια σειρά ορισμένων θεωρητικών αποτελεσμάτων, τα οποία όμως βρήκαν πολύ λίγες εφαρμογές, διότι πολλά από αυτά, για να γίνουν πράξη, απαιτούσαν αρκετό χρόνο και επίπονες πράξεις. Ακόμα και το να αντιστραφεί ένας πίνακας μέτρων ή μικρών διαστάσεων ήταν μια δύσκολη διαδικασία, που απαιτούσε αρκετή οξυδέρκεια και εξειδίκευση.[8]

Οπότε μπορούμε να πούμε ότι η στατιστική και η ανάπτυξη των υπολογιστών εξελίχθηκαν ταυτόχρονα. Κάποιες σημαντικές ανακαλύψεις και ορισμένοι από τους πρώτους υπολογιστές αναπτύχθηκαν, για να επιλύσουν και να βοηθήσουν προβλήματα στατιστικής φύσης, όπως για παράδειγμα η πινακοποίηση και η ανάλυση των αποτελεσμάτων της απογραφής.

Η ανάπτυξη των γλωσσών προγραμματισμού οδήγησε στη σταδιακή ανάπτυξη εκείνων των εργαλείων που ήταν απαραίτητα για τη συγγραφή πιο σύνθετων προγραμμάτων και αυτή με τη σειρά της οδήγησε την ανάπτυξη πιο εξειδικευμένων προγραμμάτων που ειδικεύονταν σε συγκεκριμένα προβλήματα και κατά συνέπεια εμφανίστηκαν και τα πρώτα στατιστικά πακέτα.

Τα πρώτα στατιστικά πακέτα απευθύνονταν σε εξειδικευμένο προσωπικό και ως εκ τούτου γράφτηκαν σε μια γλώσσα προγραμματισμού με συγκεκριμένες δυνατότητες συγγραφής και προγραμματισμού. Έπειτα, άρχισαν να γίνονται πιο φιλικά στον απλό χρήστη και μετά την ευρεία είσοδο των παραθυρικών εφαρμογών, οι προγραμματιστικές τους ικανότητες σχεδόν παραμελήθηκαν. Στις μέρες μας, τα περισσότερα από τα στατιστικά πακέτα δίνουν τη δυνατότητα προγραμματισμού με τη χρήση απλών εντολών.

Η R είναι μια γλώσσα προγραμματισμού που παρέχει σημαντικές δυνατότητες προγραμματισμού και συγγραφής προγραμμάτων για την επίλυση πολύπλοκων στατιστικών προβλημάτων. Η R μπορεί να χειριστεί μεγάλο όγκο και διαφορετικούς τύπους δεδομένων και δίνει τη δυνατότητα δημιουργίας υψηλής ποιότητας γραφικών.

Είναι μια γλώσσα ανοικτού κώδικα που ενσωματώνει συνεχώς νέες βιβλιοθήκες (και κατά συνέπεια νέες μεθοδολογίες ανάλυσης) με αποτέλεσμα να εξελίσσεται σε δυνατότητες και εύρος εφαρμογών. Τέλος, υπάρχει ένα τεράστιο δίκτυο έμπειρων χρηστών σε διάφορες διαδικτυακές ιστοσελίδες που μπορεί να παρέχει υψηλού επιπέδου υποστήριξη και να απαντήσει στις όποιες απορίες [8].

Το στατιστικό πακέτο R περιλαμβάνει έτοιμες προς υλοποίηση τις περισσότερες μεθόδους που χρειάζονται στη στατιστική ανάλυση υπό μορφή εντολών. Επιπλέον, μπορούμε να φέρουμε εις πέρας οποιονδήποτε στατιστικό υπολογισμό επιθυμούμε γράφοντας απλό κώδικα. Τέλος το στατιστικό πακέτο της R διατίθεται ελεύθερα από το Internet για χρήση από τον οποιονδήποτε[3].

1.2 ΤΙ ΕΙΝΑΙ Η R;

Η R είναι μια γλώσσα προγραμματισμού που χρησιμοποιείται κυρίως για ανάλυση δεδομένων και προσφέρει μια σειρά από πλεονεκτήματα.

- Διαθέτει ποικιλία στις σύγχρονες στατιστικές μεθόδους
- Περιλαμβάνει σύγχρονες τεχνικές σχετικές με θέματα στατιστικής μοντελοποίησης
- Διαθέτει υψηλής ποιότητας γραφικά
- Είναι εύκολη στον προγραμματισμό των νέων συναρτήσεων και μεθοδολογιών

Η R μπορεί να προσφέρει στους χρήστες ένα ολοκληρωμένο σύνολο από υπηρεσίες λογισμικού για ανάλυση δεδομένων, υπολογισμών καθώς και γραφημάτων. Αν και χρησιμοποιείται κυρίως στην στατιστική, οι δημιουργοί του προτιμούν να το αποκαλούν εργαλείο για ανάλυση δεδομένων τονίζοντας ότι περιλαμβάνει τόσο μοντέρνες όσο και κλασικές στατιστικές μεθοδολογίες[1][8].

Είναι ένα GNU πρόγραμμα παρόμοιο με τη γλώσσα S που αναπτύχθηκε στα εργαστήρια της Bell από τον Rick Becker, John Chambers και τον Allan Wilks. Η R μπορεί να θεωρηθεί ως μια διαφορετική υλοποίηση της S, αλλά ουσιαστικά η R και η S δεν είναι 100% συμβατές. Μοιραία η R συγκρίνεται με την S+ ή το εμπορικό πακέτο S-plus που είναι και αυτό βασισμένο στην γλώσσα S[11].

1.3 ΙΣΤΟΡΙΚΗ ΑΝΑΔΡΟΜΗ ΤΗΣ R

Η γλώσσα R αναπτύχθηκε στις αρχές της δεκαετίας του 1990 από τους Ross Ihaka και Robert Gentleman στο Πανεπιστήμιο Όκλαντ στη Νέα Ζηλανδία. Η R βασίστηκε σε

παλαιότερη γλώσσα προγραμματισμού γνωστή ως γλώσσα S. Η S αποτελούσε μια γλώσσα υψηλού επιπέδου που αναπτύχθηκε στο τέλος της δεκαετίας του 1980 και αρχές του 1990 και η οποία πρόσφερε εξαιρετικές των δυνατότητες ανάλυσης δεδομένων και γραφικών. Στην πραγματικότητα η R βασίστηκε στην S, παρόλα αυτά όμως έχουν σημαντικές διαφορές μεταξύ τους, κυρίως σε ότι έχει να κάνει με τον τρόπο που διαχειρίζονται τη διαθέσιμη μνήμη του υπολογιστή.

Το 1994 εμφανίστηκε η πρώτη έκδοση της S, σαν πακέτο λογισμικού, και την ονόμασαν “R” με σκοπό να χρησιμοποιηθεί σε ένα εισαγωγικό μάθημα στατιστικής. Στην αρχή, οι δημιουργοί της διέθεταν την R δωρεάν σε όποιον ενδιαφερόταν. Σύντομα όμως πολλοί ερευνητές συμμετείχαν στη βελτίωση της με τους δικούς τους κώδικες. Έτσι το 1996 οι χρήστες της R δημιούργησαν στη Βιέννη ένα αποθετήριο προγραμμάτων της R.

Ουσιαστικά είχε ξεκινήσει μια νέα εποχή. Οι δυο βασικοί προγραμματιστές άφησαν πια ελεύθερο το πεδίο σε όποιον άλλο χρήστη ήθελε να δημιουργεί τους δικούς του κώδικες. Το Φεβρουάριο του 2000 θεωρήθηκε ότι είχε ωριμάσει η πρόοδος και ότι υπήρχε μια αρκετά σταθερή έκδοση της R. Έτσι, η πρώτη έκδοση 1.0 ήταν πλέον στη διάθεση οποιουδήποτε ήθελε να την κατεβάσει. Την εποχή αυτή 20 προγραμματιστές ασχολούνται με την ανάπτυξη της R, ενώ συγχρόνως, πολλοί ερευνητές από την περιοχή της Υπολογιστικής Στατιστικής συνεισφέρουν σε αυτή [1][8][11].

Στα μετέπειτα χρόνια, η R θα αποτελέσει πολύτιμο εργαλείο διδασκαλίας σε όλο τον κόσμο, με αποτέλεσμα οι χρήστες να αυξάνονται. Συγχρόνως, η δυνατότητα ενσωμάτωσης καινούριων βιβλιοθηκών με εξαιρετική ευκολία, αλλά κυρίως, η δυνατότητα που δίνει στους χρήστες να φτιάχνουν και να αποθέτουν βιβλιοθήκες, οδήγησε στην τεράστια ανάπτυξη της R.

Το γεγονός ότι η R διατίθεται ελεύθερα δημιουργεί κάποια προβλήματα και περιορισμούς στη χρήση της για εμπορικούς σκοπούς, καθώς η νομική ευθύνη δεν είναι ξεκάθαρο σε ποιον ανήκει. Παρόλα αυτά, τα τελευταία χρόνια υπάρχουν κάποιες εταιρείες που προσφέρουν πιστοποίηση και άρα αναλαμβάνουν αυτές τις ευθύνες, οδηγώντας σε ολοένα μεγαλύτερη χρήση. Επίσης, μεγάλοι οργανισμοί χρησιμοποιούν την R για στατιστική ανάλυση.

Θα πρέπει επίσης να μη ξεχάσουμε ότι, αν και η όλη ανάπτυξη της R έχει αναφερθεί ως μια γλώσσα εντολών, όμως τα τελευταία χρόνια υπάρχουν γραφικά περιβάλλοντα της R που υποστηρίζουν τη γλώσσα και προσφέρουν επιλογές μέσω παραθύρων για μεγαλύτερη ευκολία στο χρήστη.

Το Ιανουάριο του 2017 υπήρχαν διαθέσιμα πάνω από 10000 πακέτα στην R. Πλέον θεωρείται εξαιρετικά συνηθισμένο ένας ερευνητής να προτείνει μια νέα μεθοδολογία, για την οποία δημιουργεί και ανεβάζει το πακέτο που την υποστηρίζει στην R. Έτσι νέες μεθοδολογίες είναι πλέον διαθέσιμες σχεδόν ταυτόχρονα με τη δημοσίευσή τους κάτι που στο παρελθόν ήταν αδύνατον.

Τέλος, είναι σημαντικό να αναφέρουμε ότι η R αποτελεί σημαντική επιλογή για τη μετάβαση σε αυτό που είναι πια γνωστό ως η «Πρόκληση των Μεγάλων Δεδομένων» (Big Data Challenge). Στις μέρες μας ο όγκος των διαθέσιμων δεδομένων έχει αυξηθεί με εντυπωσιακό ρυθμό, ενώ καινούριες δομές δεδομένων είναι διαθέσιμες για στατιστική επεξεργασία και ανάλυση σε τεράστιους όγκους (π.χ. κείμενα). Η R αποτελεί σημαντική επιλογή των χρηστών προς αυτή την κατεύθυνση, καθώς διαθέτει ήδη βιβλιοθήκες που επιτρέπουν παράλληλη επεξεργασία των δεδομένων, ενώ το γεγονός ότι καθένας μπορεί να επέμβει στους κώδικες και να τους τροποποιήσει, καθιστά εξαιρετικά εύκολο το να δουλέψει με τεράστιους όγκους δεδομένων[8][11].

1.4 ΤΟ ΠΕΡΙΒΑΛΛΟΝ ΤΗΣ R

Το βασικό εργαλείο που θα χρησιμοποιηθεί σε αυτή την εργασία είναι η γλώσσα προγραμματισμού R. Η απλή κονσόλα της R, μαζί με τα βασικά έτοιμα πακέτα είναι ελεύθερα διαθέσιμη μέσω της επίσημης ιστοσελίδας:

<http://www.r-projects.org>

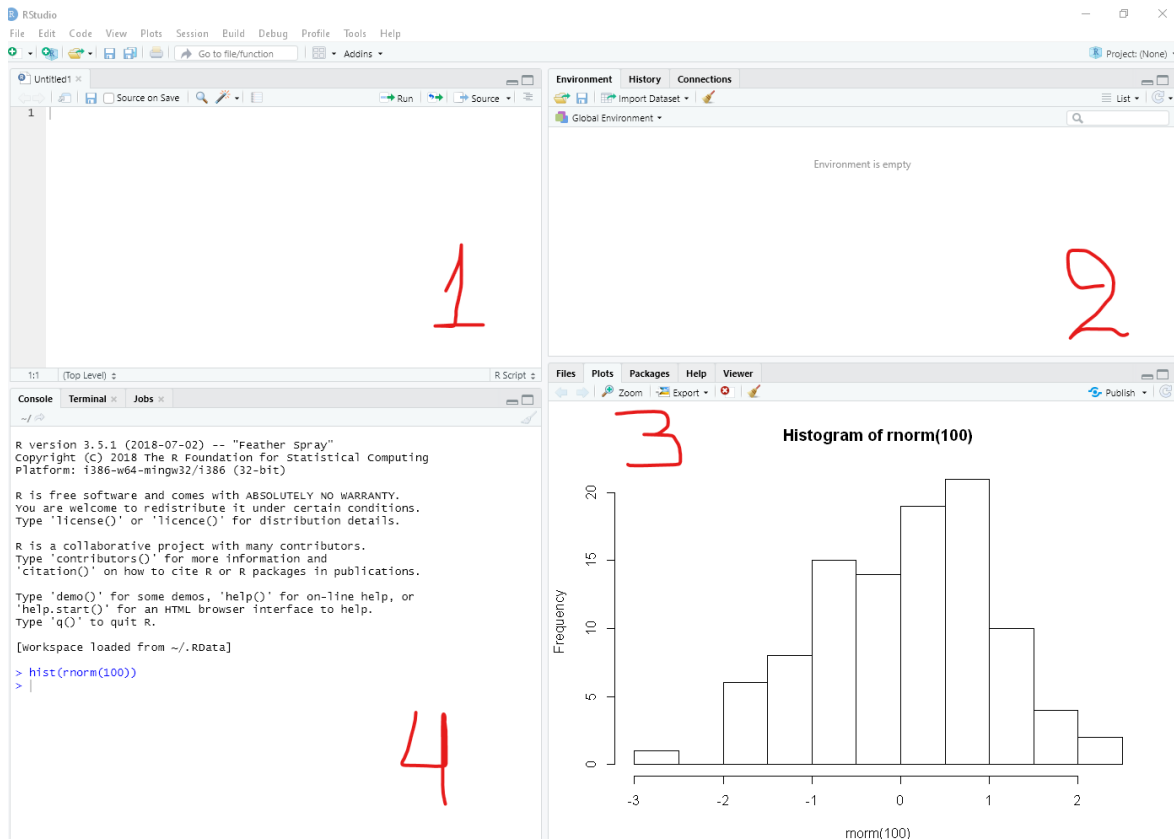
Το R λειτουργεί σχεδόν σε κάθε διαθέσιμη πλατφόρμα, συμπεριλαμβανομένων των ευρέως διαθέσιμων συστημάτων Windows, Mac OS X και Linux. Θα πρέπει να επιλέξουμε τη σωστή έκδοση ανάλογα με το λειτουργικό μας σύστημα.

Υπάρχει επίσης ένα ολοκληρωμένο περιβάλλον ανάπτυξης διαθέσιμο για το R που έχει κατασκευαστεί από το RStudio. Το αρχείο εγκατάστασης του RStudio μπορούμε να το κατεβάσουμε από τον σύνδεσμο:

<https://www.rstudio.com/products/rstudio/download/>. Θα πρέπει και πάλι να επιλέξουμε τη σωστή έκδοση ανάλογα με το λειτουργικό μας σύστημα.[12][13]

Στην Εικόνα 1.1 φαίνεται η αρχική οθόνη του Rstudio, όπου όλοι οι υπολογισμοί γίνονται μέσω εντολών, που πληκτρολογούμε στο παράθυρο R Console. Στο πλαίσιο 2 εμφανίζονται οι μεταβλητές και οι συναρτήσεις. Επίσης μπορούμε να δούμε και το παράθυρο γραφικών, στο οποίο εμφανίζονται τα διαγράμματα, τα οποία κατασκευάζουμε και φαίνεται αντίστοιχο παράδειγμα στο πλαίσιο 3 της Εικόνας 1.1. Παράλληλα μπορούμε

να δούμε και ποια πακέτα έχουμε κατεβάσει ή χρειάζονται ενημέρωση μέσω της καρτέλας “Packages” πάλι από το πλαίσιο 3. Επιπλέον, μέσω της καρτέλας “Help” μπορούμε να βρούμε πληροφορίες και βοήθεια για κάποια συνάρτηση ή πακέτο. Τέλος, στο πλαίσιο 4 βρίσκεται η κλασική κονσόλα της R.[1][11]



Εικόνα 1.1: Το RStudio περιβάλλον ανάπτυξης της R

Ορισμένες χρήσιμες λειτουργίες της R είναι οι ακόλουθες:

- Ο χρήστης έχει τη δυνατότητα να αποθηκεύσει το χώρο εργασίας του με τα δεδομένα αλλά και τις εντολές που έχει χρησιμοποιήσει. Η αποθήκευση γίνεται από το μενού File->Save και το αρχείο παίρνει την ονομασία .Rdata στο φάκελο που αποθηκεύεται.
- Ο χρήστης μπορεί μέσω της επιλογής save history του πλαισίου 2 της Εικόνας 1.1 να αποθηκεύει ότι εντολές έχει χρησιμοποιήσει μέχρι εκείνη τη στιγμή.
- Μέσω του μενού Edit ο χρήστης έχει τη δυνατότητα να καθορίσει την εμφάνιση των παραθύρων, αλλά και να κάνει αντιγραφή και επικόλληση εντολών.
- Από το μενού Tools, πατώντας Install Packages και αφού υπάρχει σύνδεση στο διαδίκτυο ο χρήστης μπορεί να κατεβάσει και να φορτώσει βιβλιοθήκες, δηλαδή προγράμματα άλλων χρηστών για διευκόλυνση της δικής του έρευνας.

- Τέλος, από το μενού Help, μπορεί κανείς να πάρει βοήθεια σε διάφορα formats.[1][8]

1.5 ΒΑΣΙΚΕΣ ΕΝΤΟΛΕΣ – ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ ΤΗΣ R

Στις περισσότερες γλώσσες προγραμματισμού υπάρχουν μεταβλητές και τύποι μεταβλητών. Ωστόσο, η R βλέπει τα πάντα ως αντικείμενα (object), τα οποία ανήκουν σε μια κλάση (class). Δηλαδή, για την R τα αντικείμενα είναι οι μεταβλητές, ενώ η κλάση είναι ο τύπος τους. Στην R δεν απαιτείται η ρητή δήλωση της κλάσης στην οποία ανήκουν τα αντικείμενα. Αυτή καθορίζεται αυτόματα από την τιμή που θα ανατεθεί στο αντικείμενο[1][7][11][13].

Οι κύριες δομές αντικειμένων στην R είναι:

- Διάνυσμα (vector)
- Πίνακας (matrix)
- Πίνακας πολλαπλών διαστάσεων (arrays)
- Λίστα (list)
- Πλαίσιο δεδομένων (data frames)

Επίσης ο διερμηνευτής της R αφού αξιολογήσει μία έκφραση επιστρέφει μία τιμή η οποία μπορεί να είναι:

- ✓ Λογική (logical)
- ✓ Αριθμητική (arithmetic)
- ✓ Μιγαδική (complex)
- ✓ Κείμενο (character)

Βασικά Χαρακτηριστικά της R:[8]

Μερικά από τα βασικά χαρακτηριστικά της R είναι τα ακόλουθα:

- Όταν εργαζόμαστε με τη γραμμή εντολών της R (R Console), στην αρχή βλέπουμε ένα από τα παρακάτω σύμβολα:
 - > το οποίο είναι το σύμβολο αναμονής και σημαίνει ότι το πρόγραμμα αναμένει την εντολή του χρήστη,
 - + το οποίο είναι το σύμβολο που υποδηλώνει ότι η εντολή που κάναμε εισαγωγή είναι ατελής και περιμένει να συνεχιστεί σε αυτή τη γραμμή.

Επιπλέον, η ένδειξη [1] πριν τα αποτελέσματα δηλώνει το πλήθος των εξαγόμενων αποτελεσμάτων (στην περίπτωση μας ένα). Όταν τα αποτελέσματα δίνονται σε μορφή

διανύσματος, τότε το νούμερο μέσα στις αγκύλες υποδηλώνει τον αύξοντα αριθμό του αντικειμένου που τυπώνει. Σε πιο σύνθετα αντικείμενα, η μορφή αλλάζει ανάλογα με την διάταξη του αντικειμένου.

- Ένα από τα σημαντικότερα σύμβολα σε όλες τις γλώσσες προγραμματισμού είναι το σύμβολο της εκχώρησης ή ανάθεσης. Στην R, οι εκχωρήσεις ή αναθέσεις σε ένα αντικείμενο γίνονται με το σύμβολο `<-` του βέλους, σχηματισμένου από 2 χαρακτήρες (« `<` » και « `-` »). Δηλαδή με το σύμβολο `<-` εκχωρούμε την τιμή του δεξιού μέλους στο αριστερό μέλος. Για παράδειγμα η εντολή `x<-3` σημαίνει πως στο αντικείμενο `x` εκχωρούμε την τιμή 3. Επίσης, στις τελευταίες εκδόσεις της R προστέθηκε ως σύμβολο εκχώρησης και το σύμβολο της ισότητας `=` που χρησιμοποιείται σε πολλές άλλες γλώσσες προγραμματισμού.
- Η R είναι ευαίσθητη στα κεφαλαία γράμματα. Δηλαδή το `x` και το `X` είναι διαφορετικά αντικείμενα για την R.
- Αν χρησιμοποιήσουμε τα βελάκια μπορούμε να επανακτήσουμε προηγούμενες εντολές χωρίς να χρειαστεί να τις ξανά πληκτρολογήσουμε. Παρόμοιες δυνατότητες υπάρχουν και μέσα από το αρχείο ιστορικού, όπου ο χρήστης μπορεί να δει όλες τις εντολές που έχει πληκτρολογήσει μέχρι τη στιγμή της τελευταίας αποθήκευσης του ιστορικού.
- Η εντολή `help` είναι ίσως η πιο σημαντική εντολή της R. Η R διαθέτει βοήθεια με λεπτομέρειες για όλες τις εντολές μαζί με παραδείγματα για τη χρήσης τους.
- Στην R για να γράψουμε περισσότερες από μία εντολές σε μια γραμμή πρέπει να τις χωρίσουμε με το χαρακτήρα `«;»`.
- Μέσω του παραθύρου R editor μπορεί να γίνει συγγραφή σειράς εντολών, δηλαδή ενός προγράμματος (script). Αυτό μπορεί να γίνει από το μενού `File>New script`. Σε παράθυρο script μπορούμε να γράψουμε πολλές εντολές και να τις τρέξουμε όλες μαζί, αρκεί να τις μαρκάρουμε και στη συνέχεια να πατήσουμε το πλήκτρο `Run`.
- Η γλώσσα R είναι διαθέσιμη στο διαδίκτυο δωρεάν και η υποστήριξή της γίνεται με συνεισφορά των εθελοντών.
- Επιπλέον η R μπορεί να υποστηριχθεί σε διαφορετικές πλατφόρμες (πχ. Windows, Linux, MacOS).

- Η R το αγνοεί οτιδήποτε ακολουθεί μετά το χαρακτήρα #, καθώς το καταλαβαίνει σαν σχόλιο και όχι σαν εντολή. Με τη χρήση του χαρακτήρα # μπορούμε ακόμα και να προσθέσουμε κάποιο σχόλιο στην ίδια γραμμή με μια εντολή.
- Ένα βασικό χαρακτηριστικό της R είναι τα αντικείμενα. Ουσιαστικά, οι μεταβλητές, οι εντολές αλλά και οι συναρτήσεις είναι αντικείμενα. Τα αντικείμενα είναι αυτόνομα και ουσιαστικά είναι δομές που οργανώνουν τα δεδομένα μας. Αποθηκεύονται στο χώρο εργασίας (workspace) της R και μπορούν να ανακληθούν άμεσα, όποτε χρειαστεί.

1.6 ΤΕΛΕΣΤΕΣ ΣΤΗΝ R

Η R όπως αναφέραμε και πιο πάνω μας δίνει τη δυνατότητα να πραγματοποιήσουμε από τις πιο απλές έως τις πιο σύνθετες μαθηματικές πράξεις. Μια απλή εντολή έχει τη μορφή:

```
> 7+7 #προσθεση στοιχειων
[1] 14
```

Με την παραπάνω εντολή ουσιαστικά ζητήσαμε να προσθέσει 2 αριθμούς. Οποιαδήποτε ακολουθεί μετά τον χαρακτήρα # η R αγνοείται, καθώς θεωρείται ότι είναι σχόλιο του χρήστη που βοηθάει στην κατανόηση ή επεξήγηση των εντολών. Βέβαια η παράσταση μπορεί να γίνει και πιο σύνθετη περιλαμβάνοντας και συναρτήσεις, όπως για παράδειγμα η επόμενη,

```
> abs(-6)-5*log(3)+sin(0.5)
[1] 0.9863641
```

η οποία περιλαμβάνει υπάρχουσες μαθηματικές συναρτήσεις όπως είναι η απόλυτη τιμή (abs), ο νεπέρειος λογάριθμος (log) καθώς και το ημίτονο (sin).

Με τις εντολές αυτές εμφανίζεται το αποτέλεσμα στην οθόνη χωρίς όμως να καταχωρείται κάπου. Αν θέλουμε να καταχωρήσουμε το αποτέλεσμα σε κάποιο αντικείμενο, τότε πρέπει να δώσουμε την εντολή ως εξής:

```
> y<-abs(-6)-5*log(3)+sin(0.5)
> y
[1] 0.9863641
```

1.7 ΤΕΛΕΣΤΕΣ ΠΡΑΞΕΩΝ ΣΤΗΝ R

1.7.1 Αριθμητικοί Τελεστές

Με τις κάτωθι εντολές ορίζουμε αρχικές τιμές στις μεταβλητές x και y και έπειτα κάνουμε πράξεις με αυτά. Ουσιαστικά, τα ορίζουμε να έχουν μία μόνο τιμή, αλλά αυτό για την R

ισοδυναμεί με τον ορισμό ενός διανύσματος με μία μόνο τιμή ή ένα μόνο στοιχείο.

```
> x<-5
> y<-3
> x+y
[1] 8
> x/y
[1] 1.666667
> x%%y
[1] 2
> x%/%y
[1] 1
```

Οι αριθμητικές πράξεις που μπορούμε να χρησιμοποιήσουμε στην R δίνονται στον Πίνακα 1.5.1 που έχουμε παρακάτω:

Σύμβολο	Πράξη
+	Πρόσθεση
-	Αφαίρεση
*	Πολλαπλασιασμός
/	Διαίρεση
^	Ύψωση σε δύναμη
%/%	Ακέραια διαίρεση, επιστρέφει δηλαδή το ακέραιο μέρος της διαίρεσης
%%	Υπόλοιπο διαίρεσης

Πίνακας 1.7.1: Αριθμητικοί τελεστές στην R

Για να αποφύγουμε απροσδόκητα ή ανεπιθύμητα αποτελέσματα και για να είμαστε σίγουροι ότι οι πράξεις θα εκτελεστούν με προτεραιότητα μπορούμε να χρησιμοποιήσουμε παρενθέσεις. Οι εκφράσεις μέσα σε παρενθέσεις αποκτούν απόλυτη προτεραιότητα. Παρακάτω θα δούμε κάποια παραδείγματα για να κατανοήσουμε τη σημασία χρήσης των παρενθέσεων.

```
> 3+5*8 %/%3
[1] 13
> (3+5)*(8 %/%3)
[1] 16
> ((3+5)*8) %/%3
[1] 21
> 6/0
[1] Inf
```

Ενδιαφέρον παρουσιάζει η τελευταία εντολή στην οποία η διαίρεση με το 0 δίνει το άπειρο ως αποτέλεσμα [1][7][8][13].

1.7.2 Τελεστές σύγκρισης

Εκτός από τους αριθμητικούς τελεστές, υπάρχουν και τελεστές με τους οποίους κάνουμε σύγκριση. Για παράδειγμα, ο τελεστής `=` εξετάζει αν δυο στοιχεία είναι ίσα και μας δίνει

την τιμή TRUE, αν ισχύει η ισότητα και την τιμή FALSE, αν δεν ισχύει. Δηλαδή τα αποτελέσματα των τελεστών σύγκρισης είναι ένα αποτέλεσμα που μας δείχνει, αν ισχύει ή όχι, η σχέση που ελέγχουμε και ονομάζεται λογική(logical) σταθερά ή τιμή. Άλλοι τελεστές συγκρίσεων δίνονται στον Πίνακα 1.5.2 που ακολουθεί.

Τελεστής	Ερμηνεία
>	Μεγαλύτερο από
<	Μικρότερο από
>=	Μεγαλύτερο ή ίσο
<=	Μικρότερο ή ίσο
==	Ίσο
!=	Όχι ίσο

Πίνακας 1.7.2: Τελεστές σύγκρισης στην R

Μερικές χρήσιμες παρατηρήσεις είναι οι ακόλουθες:

- Όταν κάνουμε αναθέσεις ή ελέγχους, μπορούμε να χρησιμοποιήσουμε ισοδύναμα μόνο το πρώτο γράμμα των FALSE και TRUE. Δηλαδή μπορούμε να γράψουμε F αντί για FALSE και T αντί για TRUE.
- Η τιμή FALSE είναι μικρότερη της τιμής TRUE.
- Έχουμε τη δυνατότητα να κάνουμε συγκρίσεις στο δεξί μέλος της εντολής εκχώρησης και να κάνουμε αριθμητικές πράξεις. Στις πράξεις αυτές, οι λογικές τιμές μετατρέπονται σε αριθμητικές με το FALSE να γίνεται μηδέν (0) και το TRUE να γίνεται ίσο με ένα (1). Αυτό προσφέρει τη δυνατότητα στο χρήστη να γράψει περίπλοκες κλαδωτές συναρτήσεις με μια απλή εντολή.
- Η χρήση του ίσον σε σύγκριση αριθμητικών τιμών είναι παράτολμη, λόγω στρογγυλοποίησης, αλλά και συσσώρευσης ορισμένων λαθών που μπορεί να πάρουμε την τιμή FALSE σε περιπτώσεις που δεν θα έπρεπε και το ανάποδο. Συνήθως, για να ελέγξουμε την ύπαρξη μιας ισότητας όταν εμπλέκονται πραγματικοί αριθμοί χρησιμοποιούμε μια ανισότητα της μορφής, ότι η απόλυτη διαφορά μεταξύ δυο αριθμών είναι μικρότερη από έναν μικρότερο αριθμό. Δείτε ένα παράδειγμα στις εντολές που ακολουθούν:

```

> 15.0000000001 == 15
[1] FALSE
> (15.0000000001 - 15) < 10^(-4)
[1] TRUE
> 15.0000000000 == 15
[1] TRUE

```

Στα ακόλουθα παραδείγματα γίνεται χρήση των αριθμητικών συγκρίσεων και στη συνέχεια ξανά σύγκριση των λογικών τιμών. Για παράδειγμα, στην πρώτη έκφραση ελέγχουμε αν το 4 είναι μικρότερο του 5 (αποτέλεσμα TRUE) και αν το 3 είναι μικρότερο του 2 (αποτέλεσμα FALSE). Μετά συγκρίνουμε τα δυο αποτελέσματα, δηλαδή αν το TRUE είναι μεγαλύτερο του FALSE που ισχύει (δηλαδή το τελικό αποτέλεσμα είναι TRUE), εφόσον στο πρώτο ορίζεται η τιμή 1 και στο δεύτερο η τιμή 0. [8][12]

```

> (4<5) > (3<2)
[1] TRUE
> (4<5) > (3>2)
[1] FALSE
> (3+2) > 4
[1] TRUE
> 3+(5>6)
[1] 3

```

1.7.3 Λογικοί Τελεστές

Μπορούμε να συνδυάσουμε δύο ή περισσότερες συγκρίσεις με χρήση των λογικών τελεστών & και |. Οπότε, χρησιμοποιώντας το X&Y παίρνουμε την τιμή TRUE, αν ισχύουν και οι δύο συνθήκες που εμπλέκονται στην έκφραση (δηλαδή και η X και η Y πρέπει να είναι αληθείς), ενώ με το X|Y θα πάρουμε TRUE, αν ισχύει τουλάχιστον μία από αυτές (δηλαδή αν η X ή η Y συνθήκη είναι αληθής).

```

> (x<-5) | (x>5)
[1] TRUE

```

Ο τελεστής ! υποδηλώνει άρνηση, δηλαδή το αντίθετο της λογικής έκφρασης ή της σύγκρισης που θα γίνει. Άρα, η παράσταση !(4>5) λαμβάνει την τιμή TRUE, διότι η παράσταση (4>5) έχει την τιμή FALSE. Πρακτικά ο τελεστής ! μετατρέπει τη λογική τιμή TRUE σε FALSE και το αντίθετο.

Στον Πίνακα 1.5.3 φαίνονται διάφοροι συνδυασμοί λογικών τελεστών σε συνδυασμό με τελεστές σύγκρισης. Ένας πίνακας αυτής της μορφής ονομάζεται πίνακας αληθείας. Από τον πίνακα βλέπουμε την τιμή διαφόρων εκφράσεων με βάση την τιμή των αρχικών εκφράσεων.

X	Y	X&Y	X Y	!(X&Y)	!X	!Y	!X&!Y	!(X Y)	(!X) (!Y)
F	F	F	F	T	T	T	T	T	T
F	T	F	T	T	T	F	F	F	T
T	F	F	T	T	F	T	F	F	T
T	T	T	T	F	F	F	F	F	F

Πίνακας 1.7.3: Πίνακας αληθείας για τα X και Y

Ο πίνακας παραπάνω ουσιαστικά θα επιβεβαιώσει τους κανόνες, που είναι γνωστοί και ως κανόνες De Morgan.

- Η άρνηση της τομής δύο ενδεχομένων είναι ίση με την ένωση των αρνήσεων των δυο ενδεχομένων χρησιμοποιώντας τελεστές της R.

$$!(X \& Y) = (!X) | (!Y)$$

και

- Η άρνηση της ένωσης δυο ενδεχομένων είναι ίση με την τομή της άρνησης των δύο ενδεχομένων. [8][11]

$$!(X | Y) = (!X) \& (!Y)$$

1.7.4 Σειρά προτεραιότητας τελεστών

Κλείνουμε την αναφορά μας στους τελεστές (αριθμητικούς, σύγκρισης και λογικούς) δίνοντας στον Πίνακα 1.5.4 τη σειρά προτεραιότητας τους μέσα σε μία έκφραση.[8]

Διάφοροι Τελεστές	Ερμηνεία αυτών
^	Ύψωση σε δύναμη
% / %, %%	Ακέραια διαίρεση, υπόλοιπο
/, *	Διαίρεση, πολλαπλασιασμός
+, -	Πρόσθεση, αφαίρεση
<, >, <=, >=, ==, !=	Συγκρίσεις αριθμών
!	Λογική άρνηση
&,	Λογικοί τελεστές

Πίνακας 1.7.4: Σειρά προτεραιότητας διαφόρων τελεστών

1.8 ΤΥΠΟΙ ΚΑΙ ΔΟΜΕΣ ΔΕΔΟΜΕΝΩΝ

1.8.1 ΔΙΑΝΥΣΜΑΤΑ

Όπως έχουμε ήδη αναφέρει, η R υποστηρίζει βασικές δομές δεδομένων ως κλάσεις αντικειμένων. Η πιο βασική μορφή της R είναι το διάνυσμα (vector). Ουσιαστικά, αν ένα μοναδικό στοιχείο εκχωρείται σε μία μεταβλητή, αυτή στην ουσία είναι ένα διάνυσμα με ένα στοιχείο.

Διάνυσμα είναι οποιοδήποτε αντικείμενο της R, που περιλαμβάνει ένα σύνολο στοιχείων

ίδιου τύπου, καταχωρημένα με συγκεκριμένη σειρά, την οποία και έχει ορίσει ο χρήστης. Δημιουργείται με τη χρήση της εντολής `c`, από την αγγλική λέξη `combine` που σημαίνει συνδυάζω μοναδικές τιμές σε ένα διάνυσμα. Τα στοιχεία μέσα στην παρένθεση της εντολής `c()` διαχωρίζονται με κόμμα.

```
> x<-c("this", "is", "a", "character", "vector")
> x
[1] "this"      "is"
[3] "a"         "character"
[5] "vector"
> y<-c(1, 2, 3, 5, 7)
> y
[1] 1 2 3 5 7
```

Εικόνα 1.2: Παράδειγμα δημιουργίας διανύσματος με την εντολή `c`

Εναλλακτικά, μπορεί να χρησιμοποιηθεί η συνάρτηση `vector`. Γενικά, η δεικτοδότηση των διάφορων δομών στην R ξεκινούν από το 1 και όχι από το 0. **[Error! Unknown switch argument.][8][7][11][13]**

```
> # Arxikopoihsh logikou dianismatos mhkous 6
> y<-vector(mode="logical", length= 6)
> y
[1] FALSE FALSE FALSE FALSE
[5] FALSE FALSE
> y[1]<-TRUE
> y
[1] TRUE FALSE FALSE FALSE
[5] FALSE FALSE
```

Εικόνα 1.3: Παράδειγμα δημιουργίας διανύσματος με τη συνάρτηση `vector`

Ιδιότητες και Χαρακτηριστικά Διανυσμάτων[8]

Στα αντικείμενα της R δίνουμε ονόματα τα οποία αποτελούνται είτε οποιονδήποτε συνδυασμό γραμμάτων είτε αριθμών είτε ακόμα και τελειών και πρέπει να αρχίζουν με κάποιο γράμμα. Ο χρήστης ακόμα μπορεί να χρησιμοποιήσει για όνομα αντικειμένου και μία δεσμευμένη λέξη(π.χ. το όνομα μίας έτοιμης συνάρτησης), γι' αυτό απαιτείται προσοχή στην ονοματολογία διότι η R δεν θα εμποδίσει κάτι τέτοιο.

Επιπλέον τα διανύσματα έχουν τις ακόλουθες ιδιότητες:

- ❖ **Ονόματα στοιχείων:** Στην R έχουμε τη δυνατότητα να δώσουμε ονομασίες σε κάθε στοιχείο του διανύσματος με την εντολή `names(x)`. Συνήθως, όταν αναφερόμαστε σε στατιστικά στοιχεία, οι ονομασίες αυτές μπορεί να είναι τα ονόματα των ατόμων που συμπεριλαμβάνονται στο δείγμα.
- ❖ **Μήκος διανύσματος:** Ως μήκος ενός διανύσματος ορίζουμε το συνολικό αριθμό των

στοιχείων ενός διανύσματος. Αν υπάρξουν τιμές οι οποίες λείπουν (κωδικοποιημένες από την R ως NA) τότε μετρώνται και αυτές. Μπορούμε να ανακτήσουμε το μήκος ενός διανύσματος με την εντολή `length(x)`, καθώς επίσης και να το τροποποιήσουμε θέτοντας για παράδειγμα `length(x)<-8`.

- ❖ **Τύπος διανύσματος:** Όπως ήδη αναφέραμε, μπορούμε να το εξετάσουμε με την εντολή `mode(x)` και πιθανές τιμές να είναι οι «numeric», «logical», «character» και «complex».

Χαρακτηριστικά διανυσμάτων[8]

- ✓ Length: Μήκος (Αριθμός τιμών/δεδομένων στο διάνυσμα)
- ✓ Mode: Τύπος διανύσματος
- ✓ Names: Ονόματα (όνομα για κάθε τιμή)

1.8.2 ΠΙΝΑΚΕΣ

Όπως έχουμε ήδη αναφέρει, η R εκτός από τα απλά διανύσματα έχει και πολλά διαφορετικά αντικείμενα. Σε αυτό το σημείο θα περιγράψουμε τους κύριους τύπους αντικειμένων που διαφέρουν αρκετά τόσο στη μορφή όσο και στον τρόπο αποθήκευσης των δεδομένων τους. Έτσι θα αναφερθούμε διεξοδικά στους πίνακες (ή αλλιώς μήτρες, *matrices*), πολυδιάστατους πίνακες (*arrays*) και πλαίσια δεδομένων (*data frames*).

A. Δισδιάστατοι πίνακες[1][7][8][11]

Οι πίνακες (πίνακας – *matrix*) ορίζονται ως δομές δεδομένων, που έχουν τα στοιχεία τους διατεταγμένα τόσο σε σειρές όσο και σε στήλες. Κάθε πίνακας περιλαμβάνει στοιχεία του ίδιου τύπου (αριθμούς, λογικές τιμές ή *character*).

Δημιουργούνται, αν εφαρμόσουμε την εντολή `matrix( )` σε ένα διάνυσμα και αν δηλώσουμε πόσες γραμμές και στήλες θα επιθυμούσαμε να έχει ο πίνακας αυτός.

```
> x<-c(2:4, 10:12, 24:26, 34:36)
> x
 [1]  2  3  4 10 11 12 24 25
 [9] 26 34 35 36
> y<-matrix(x, ncol = 4)
> y
      [,1] [,2] [,3] [,4]
[1,]    2   10   24   34
[2,]    3   11   25   35
[3,]    4   12   26   36
```

```
> y<-matrix(x, nrow = 3)
> y
      [,1] [,2] [,3] [,4]
[1,]     2    10    24    34
[2,]     3    11    25    35
[3,]     4    12    26    36
```

Εικόνα 1.4: Παράδειγμα δημιουργίας πίνακα

Οι παράμετροι `nrow` και `ncol` δηλώνουν τον αριθμό των σειρών και τον αριθμό των στηλών αντίστοιχα. Μόνο η μία παράμετρος είναι αρκετή. Η άλλη υπολογίζεται αυτόματα μέσω της R.

Τις διαστάσεις ενός πίνακα μπορούμε να τις επιστρέψουμε με την εντολή `dim()`.

```
> dim(y)
[1] 3 4
```

Το αποτέλεσμα είναι ένα διάνυσμα μήκους δυο στοιχείων. Με την ίδια εντολή μπορούμε να δημιουργήσουμε πίνακες από διανύσματα. Έτσι, αν σε ένα διάνυσμα προσθέσουμε τη διάσταση, τότε τα στοιχεία του διανύσματος θα διαμορφώσουν τον πίνακα αντίστοιχης διάστασης, γεμίζοντας διαδοχικά τις στήλες του κάθε πίνακα.

Αν ένας πίνακας που ζητήσουμε να δημιουργηθεί έχει περισσότερα στοιχεία από το διάνυσμα από το οποίο παράγεται, σε αυτή τη περίπτωση η R κάνει ξανά ανακύκλωση και μας δίνει μια σχετική προειδοποίηση.

```
> matrix(c(13, -2, 5, 6, 7, -1, 0, 4), ncol = 3)
      [,1] [,2] [,3]
[1,]    13     6     0
[2,]    -2     7     4
[3,]     5    -1    13
warning message:
In matrix(c(13, -2, 5, 6, 7, -1, 0, 4), ncol = 3) :
  data length [8] is not a sub-multiple or multiple of the number of rows [3]
```

Αρκετοί από τους πίνακες δημιουργούνται και με τη συγχώνευση διανυσμάτων. Τα διανύσματα μπαίνουν το ένα δίπλα στο άλλο σαν στήλες, αν δοθούν ως παράμετροι της εντολής `cbind()`, και το ένα κάτω από το άλλο σαν γραμμές, αν δοθούν ως παράμετροι της εντολής `rbind()`. Αναλόγως την εντολή `cbind()` ή `rbind()` που θα χρησιμοποιήσουμε, τα ονόματα από τα διανύσματα γίνονται ονόματα στηλών ή γραμμών, αντίστοιχα. Σε περίπτωση όμως που τα διανύσματα δεν έχουν το ίδιο μήκος, τότε η R θα εφαρμόσει ανακύκλωση, έτσι ώστε να καταφέρει να τα φέρει όλα στο ίδιο μήκος.

```

> k<-c(4,8,2,9)
> m<-c(-2,-1,0,4)
> cbind(k,m)
      k      m
[1,]  4     -2
[2,]  8     -1
[3,]  2      0
[4,]  9      4
> cbind(m,k)
      m      k
[1,] -2      4
[2,] -1      8
[3,]  0      2
[4,]  4      9
> rbind(k,m,k)
      [,1] [,2] [,3] [,4]
k      4      8      2      9
m     -2     -1      0      4
k      4      8      2      9
> k<-c(4,8,2)
> m<-c(-2,-1,0,4,-2)
> cbind(k,m)
      k      m
[1,]  4     -2
[2,]  8     -1
[3,]  2      0
[4,]  4      4
[5,]  8     -2
warning message:
In cbind(k, m) :
  number of rows of result is not a multiple of vector length (arg 1)
> cbind(m,k)
      m      k
[1,] -2      4
[2,] -1      8
[3,]  0      2
[4,]  4      4
[5,] -2      8
warning message:
In cbind(m, k) :
  number of rows of result is not a multiple of vector length (arg 2)

```

Εικόνα 1.5: Παράδειγμα ανακύκλωσης στην R

Στα δυο τελευταία παραδείγματα βλέπουμε ότι γίνεται ανακύκλωση.

Τέλος, έχουμε τη δυνατότητα να δημιουργήσουμε διαγώνιους πίνακες σύμφωνα με την εντολή `diag( )`. Αν η παράμετρος που ορίσουμε είναι διάνυσμα, τότε τα στοιχεία του διανύσματος τοποθετούνται στη διαγώνιο. Στη περίπτωση όμως που η παράμετρος είναι αριθμός, τότε δημιουργείται τετραγωνικός πίνακας με τον αριθμό γραμμών και στηλών όσο ο αριθμός του διανύσματος και παίρνει την τιμή 1 σε όλα τα διαγώνια στοιχεία.

```

> k
[1] 4 8 2
> diag(k)
      [,1] [,2] [,3]
[1,]    4    0    0
[2,]    0    8    0
[3,]    0    0    2
> diag(4)
      [,1] [,2] [,3] [,4]
[1,]    1    0    0    0
[2,]    0    1    0    0
[3,]    0    0    1    0
[4,]    0    0    0    1

```

Εικόνα 1.6: Παράδειγμα δημιουργίας διαγώνιου πίνακα

Οι μαθηματικές καθώς και οι λογικές πράξεις οι οποίες εφαρμόζονται σε πίνακες, εφαρμόζονται σε κάθε στοιχείο τους και το αποτέλεσμα είναι πίνακας με ίδιες διαστάσεις με τους αρχικούς.

```

> t<-matrix(10:18,ncol = 3)
> t
      [,1] [,2] [,3]
[1,]   10   13   16
[2,]   11   14   17
[3,]   12   15   18
> t^2
      [,1] [,2] [,3]
[1,]  100  169  256
[2,]  121  196  289
[3,]  144  225  324
> sqrt(t)
      [,1] [,2]
[1,] 3.162278 3.605551
[2,] 3.316625 3.741657
[3,] 3.464102 3.872983
      [,3]
[1,] 4.000000
[2,] 4.123106
[3,] 4.242641
> t/0
      [,1] [,2] [,3]
[1,]  Inf  Inf  Inf
[2,]  Inf  Inf  Inf
[3,]  Inf  Inf  Inf
> t==12
      [,1] [,2] [,3]
[1,] FALSE FALSE FALSE
[2,] FALSE FALSE FALSE
[3,]  TRUE FALSE FALSE
> t>10 & t<14
      [,1] [,2] [,3]
[1,] FALSE  TRUE FALSE
[2,]  TRUE FALSE FALSE
[3,]  TRUE FALSE FALSE
> v<-matrix(-2:6,ncol = 3)
> v
      [,1] [,2] [,3]
[1,]   -2    1    4
[2,]   -1    2    5
[3,]    0    3    6
> v>2 & v>0
      [,1] [,2] [,3]
[1,] FALSE FALSE  TRUE
[2,] FALSE FALSE  TRUE
[3,] FALSE  TRUE  TRUE

```

Εικόνα 1.7:Πράξεις σε πίνακες

Ορισμένα από τα στοιχεία του πίνακα εξάγονται με τρόπο αντίστοιχο των διανυσμάτων, μόνο που μέσα στις αγκύλες πρέπει να αναφερόμαστε και σε γραμμές και σε στήλες, χωρίζοντας τις συνθήκες ή τους αριθμούς που δίνουμε με κόμματα.

```

> z<-matrix(-4:7,ncol = 4,row = 3)
> z
      [,1] [,2] [,3] [,4]
[1,]   -4   -1    2    5
[2,]   -3    0    3    6
[3,]   -2    1    4    7
> z[1,1]
[1] -4
> z[1:2,c(2,4)]
      [,1] [,2]
[1,]   -1    5
[2,]    0    6
> z[z<=0]
[1] -4 -3 -2 -1  0
> z[-2,-1]
      [,1] [,2] [,3]
[1,]   -1    2    5
[2,]    1    4    7

```

Εικόνα 1.8: Παράδειγμα επιμέρους στοιχείων πίνακα

Παρατηρούμε πως αν χρησιμοποιήσουμε λογική συνθήκη, τότε τα αποτελέσματα επιστρέφονται ως διάνυσμα. Με την εντολή `z[ , ]` δηλώνουμε ότι δεν επιθυμούμε τη δεύτερη γραμμή, ούτε και την πρώτη στήλη του πίνακα. Αν όμως παραλείψουμε να δώσουμε αριθμό για κάποια διάσταση, τότε είναι σαν να δηλώνουμε το πλήρες εύρος της διάστασης.

```

> z[2,]
[1] -3  0  3  6
> z[,3]
[1] 2 3 4
> z[,]
      [,1] [,2] [,3] [,4]
[1,]   -4   -1    2    5
[2,]   -3    0    3    6
[3,]   -2    1    4    7

```

Ένας ακόμα τρόπος για να εξάγουμε στοιχεία πιο «ακανόνιστα» κατανομημένα σε έναν πίνακα, είναι να μπορούμε να κατασκευάσουμε έναν δεύτερο πίνακα με δύο στήλες. Κάθε γραμμή του πίνακα να δίνει τον αριθμό γραμμής και στήλης ενός στοιχείου. Στο επόμενο παράδειγμα ζητάμε τα στοιχεία `[2,3]` (δηλ πρώτη γραμμή, τρίτη στήλη) και `[2,5]` (δηλ δεύτερη γραμμή, τέταρτη στήλη) του πίνακα `z`.

```

> z<-matrix(c(2,3,2,5),ncol = 2,byrow = T)
> z
      [,1] [,2]
[1,]    2    3
[2,]    2    5

```

Είδαμε πως μπορεί κανείς να χρησιμοποιήσει τους δείκτες για να αποκτήσει πρόσβαση σε συγκεκριμένα στοιχεία ή ακόμα και γραμμές (στήλες) ενός πίνακα. Αυτό που θα πρέπει να τονιστεί είναι πως αν από λάθος χρησιμοποιηθεί ένας μόνο δείκτης (χωρίς όμως κόμμα), τότε η R αντιλαμβάνεται τον πίνακα ως διάνυσμα, το οποίο κατασκευάζεται, βάζοντας σε

σειρά τις στήλες . Έτσι, έχουμε:

```
> f<-matrix(1:12,3,4)
> f
      [,1] [,2] [,3] [,4]
[1,]     1     4     7    10
[2,]     2     5     8    11
[3,]     3     6     9    12
> f[,4]
[1] 10 11 12
> f[3,]
[1]  3  6  9 12
> f[7]
[1]  7
```

Δηλαδή όταν ζητήσουμε το `f [7]`, ο πίνακας `f` γίνεται διάνυσμα και η σειρά των στοιχείων καθορίζεται από τις στήλες, ουσιαστικά είναι ισοδύναμο με το να ορίσουμε `c (f [,1], f [,2], f [,3], f [,4] )`.

Επιπλέον, με την εντολή `diag( )` επιστρέφονται τα στοιχεία της κύριας διαγωνίου του πίνακα.

```
> z
      [,1] [,2] [,3] [,4]
[1,]    -4    -1     2     5
[2,]    -3     0     3     6
[3,]    -2     1     4     7
> diag(z)
[1] -4  0  4
```

Εικόνα 1.9: Παράδειγμα εντολής `diag( )`

Οι πράξεις μεταξύ πινάκων δεν πραγματοποιούνται με χρήση της άλγεβρας πινάκων. Αντίθετα, όπως και στα διανύσματα, γίνονται μεταξύ αντιστοίχων στοιχείων. Οι πίνακες όμως, θα πρέπει να έχουν τις ίδιες διαστάσεις, γιατί αλλιώς δεν πραγματοποιείται ανακύκλωση.

```
> z
      [,1] [,2] [,3] [,4]
[1,]    -4    -1     2     5
[2,]    -3     0     3     6
[3,]    -2     1     4     7
> diag(z)
[1] -4  0  4
> v<-matrix(-2:6,ncol = 3)
> v
      [,1] [,2] [,3]
[1,]    -2     1     4
[2,]    -1     2     5
[3,]     0     3     6
> diag(v)
[1] -2  2  6
> z<-matrix(1:9,ncol = 3)
> v<-matrix(9:1,ncol = 3)
> z+v
```

```

      [,1] [,2] [,3]
[1,]   10   10   10
[2,]   10   10   10
[3,]   10   10   10
> z^v
      [,1] [,2] [,3]
[1,]     1 4096 343
[2,]   256 3125   64
[3,]  2187 1296    9
> z>v
      [,1] [,2] [,3]
[1,] FALSE FALSE TRUE
[2,] FALSE FALSE TRUE
[3,] FALSE  TRUE TRUE
> z>v & z<6
      [,1] [,2] [,3]
[1,] FALSE FALSE FALSE
[2,] FALSE FALSE FALSE
[3,] FALSE FALSE FALSE

```

Εικόνα 1.10: Παράδειγμα με πράξεις μεταξύ πινάκων

Για την εφαρμογή της άλγεβρας πινάκων υπάρχουν ειδικές συναρτήσεις στην R που δίνουμε στον παρακάτω πίνακα.

Εντολή	Ερμηνεία
t (Q)	Ανάστροφος του πίνακα
solve (Q)	Αντίστροφος του πίνακα (αν υπάρχει)
% * %	Πολλαπλασιασμός πινάκων (αν έχουμε συμβατές διαστάσεις)

Πίνακας 1.8.1: Ειδικές συναρτήσεις και τελεστές για πράξεις με πίνακες

Παραδείγματα χρήσης των συναρτήσεων αυτών δίνονται στον ακόλουθο κώδικα της R.

```

> w<-matrix(1:12,ncol = 2)
> p<-matrix(c(1,3,1,2),nrow = 2)
> w
      [,1] [,2]
[1,]     1     7
[2,]     2     8
[3,]     3     9
[4,]     4    10
[5,]     5    11
[6,]     6    12
> p
      [,1] [,2]
[1,]     1     1
[2,]     3     2

```

```

> w %*% p
      [,1] [,2]
[1,]    22  15
[2,]    26  18
[3,]    30  21
[4,]    34  24
[5,]    38  27
[6,]    42  30
> p %*% w
Error in p %*% w : non-conformable arguments
> w<-matrix(c(4,0,0,4),ncol = 2)
> w
      [,1] [,2]
[1,]     4     0
[2,]     0     4
> solve(w)
      [,1] [,2]
[1,] 0.25 0.00
[2,] 0.00 0.25
> w<-matrix(-4:7,ncol = 4)
> w
      [,1] [,2] [,3] [,4]
[1,]    -4    -1     2     5
[2,]    -3     0     3     6
[3,]    -2     1     4     7
> t(w)
      [,1] [,2] [,3]
[1,]    -4    -3    -2
[2,]    -1     0     1
[3,]     2     3     4
[4,]     5     6     7

```

Εικόνα 1.11: Παραδείγματα χρήσης συναρτήσεων στην R

Ένα διάνυσμα μπορεί να θεωρηθεί είτε ως πίνακας-γραμμή, είτε ως πίνακας-στήλη. Η R θα το χειριστεί ανάλογα με τη θέση του σε μια πράξη.

```

> n<-1:3
> n%*%w
      [,1] [,2] [,3] [,4]
[1,]   -16     2    20    38
> g<-matrix(1:12,nrow = 4)
> g
      [,1] [,2] [,3]
[1,]     1     5     9
[2,]     2     6    10
[3,]     3     7    11
[4,]     4     8    12
> g%*%n
      [,1]
[1,]    38
[2,]    44
[3,]    50
[4,]    56

```

Στο παράδειγμα που δώσαμε παραπάνω, στον πολλαπλασιασμό με το w, το n θεωρήθηκε ως πίνακας-γραμμή, ενώ στον πολλαπλασιασμό με το g, το n θεωρήθηκε πίνακας-στήλη.

Η εντολή `is.matrix`[Error! Reference source not found.]

Η εντολή `is.matrix` εξετάζει αν το αντικείμενο είναι τύπου `matrix` ή όχι και επιστρέφει `TRUE` ή `FALSE` αντίστοιχα. Με όμοιο τρόπο η εντολή `is.vector` εξετάζει αν ένα αντικείμενο είναι τύπου `vector`. Αυτές οι εντολές είναι πολύ χρήσιμες για να μπορέσουμε να ελέγξουμε τη μορφή που έχει το αντικείμενο και πιθανότατα να μπορέσουμε να επιλέξουμε την κατάλληλη μέθοδο.

```
> x<-c(1,4,7,10,12)
> y<-matrix(1:12,3,4)
> is.matrix(x)
[1] FALSE
> is.matrix(y)
[1] TRUE
> is.vector(x)
[1] TRUE
> is.vector(y)
[1] FALSE
```

Εικόνα 1.12: Παράδειγμα της εντολής `is.matrix`

B. Πολυδιάστατοι πίνακες[1][7][8][11]

Πολυδιάστατος πίνακας (`array`) θεωρείται ένας πίνακας που έχει τρεις ή περισσότερες διαστάσεις. Ένας πίνακας με τρεις διαστάσεις για παράδειγμα, θα εμφανίζει τόσους δισδιάστατους πίνακες, όσους και η τιμή της τρίτης διάστασης. Η R έχει ως όριο 8 διαστάσεις. Όλοι αυτοί οι πίνακες δημιουργούνται σύμφωνα με την εντολή `array( )` ή την εντολή `dim( )`. Η εντολή `dim( )` δέχεται ως παράμετρο το μέγεθος της κάθε διάστασης. Η εντολή `array( )` αντίστοιχα δέχεται ως παράμετρο την `dim=` με ένα διάνυσμα με τα μεγέθη των διαστάσεων.

```
> a3<-array(1:24,dim = c(3,4,2))
> a3
, , 1
      [,1] [,2] [,3] [,4]
[1,]    1    4    7   10
[2,]    2    5    8   11
[3,]    3    6    9   12

, , 2
      [,1] [,2] [,3] [,4]
[1,]   13   16   19   22
[2,]   14   17   20   23
[3,]   15   18   21   24
```

Εικόνα 1.13: Παράδειγμα δημιουργίας πολυδιάστατου πίνακα

Παρατηρούμε ότι η πρώτη διάσταση «κινείται» πιο γρήγορα, μετά η δεύτερη και τελευταία η τρίτη. Δηλαδή στην ουσία ο πρώτος πίνακας `[, , 1]` «γεμίζει» πρώτα, μέσα

στον οποίο «γεμίζουν» πρώτα οι στήλες και μετά οι γραμμές, έπειτα «γεμίζει» ο δεύτερος πίνακας, κ.ο.κ.

Με χρήση της εντολής `dim()` φτιάχνουμε τον ίδιο πίνακα ως εξής:

```
> a3<-1:24
> dim(a3)<-c(3,4,2)
> a3
, , 1
      [,1] [,2] [,3] [,4]
[1,]    1    4    7   10
[2,]    2    5    8   11
[3,]    3    6    9   12
, , 2
      [,1] [,2] [,3] [,4]
[1,]   13   16   19   22
[2,]   14   17   20   23
[3,]   15   18   21   24
```

Εικόνα 1.14: Δημιουργία πολυδιάστατου πίνακα με χρήση της εντολής `dim()`

Η R χειρίζεται τόσο τις δομές τύπου `array`, όσο και αυτές τύπου `matrix`. Μπορούν επίσης να δοθούν ονόματα στις διαστάσεις τους, οι πράξεις γίνονται στοιχείο-στοιχείο κάθε φορά, επιπλέον στοιχεία εξάγονται με χρήση της αγκύλης, αριθμητικοί ή λογικοί δείκτες μπορούν να χρησιμοποιηθούν κλπ.

1.8.3 ΠΛΑΙΣΙΑ ΔΕΔΟΜΕΝΩΝ

Τα πλαίσια δεδομένων (`data frames`) χρησιμοποιούνται για την αποθήκευση δεδομένων σε μορφή πίνακα. Πρόκειται για δισδιάστατους πίνακες, των οποίων, όμως, τα στοιχεία μπορούν να έχουν διαφορετικό τύπο δεδομένων σε κάθε στήλη τους. Ένας λογικός περιορισμός που θέτουν είναι ότι κάθε στήλη μπορεί να περιέχει μόνο αντικείμενα της ίδιας κλάσης. Για τη δημιουργία ενός πλαισίου δεδομένων χρησιμοποιούμε τη συνάρτηση ***data.frame***.

Τα στοιχεία σε ένα πλαίσιο δεδομένων μπορούν να προσπελαστούν με τον ίδιο τρόπο που γίνεται η δεικτοδότηση στα μητρώα(εννοούμε πολλά διανύσματα ενωμένα). Μια ακόμα αξιοσημείωτη ιδιότητα τους είναι ότι σε κάθε στήλη μπορεί να ανατεθεί ένα όνομα. Από εκεί και έπειτα η αναφορά μπορεί να γίνει και με το όνομα της κάθε στήλης.[Error! Unknown switch argument.][8][Error! Unknown switch argument.]

```

> x<-c("ntina", "nikos", "vasw")
> y<-c(14, 19, 30)
> z<-c(TRUE, FALSE, FALSE)
> df<-data.frame(username=x, age=y, adult=z)
> df
  username age adult
1   ntina  14  TRUE
2   nikos  19 FALSE
3   vasw   30 FALSE
> #Proti grammi
> df[,1]
[1] ntina nikos vasw
3 Levels: nikos ... vasw
> #stili hlikias
> df$age
[1] 14 19 30

```

Εικόνα 1.15: Παράδειγμα δημιουργίας πλαισίου δεδομένων (data frame)

1.9 ΣΥΝΑΡΤΗΣΕΙΣ ΣΤΗΝ R

Σε όλα τα έτοιμα πακέτα που είναι διαθέσιμα στην R, υπάρχουν διαφορετικές συναρτήσεις, οι οποίες χρησιμοποιούνται για διαφορετική λειτουργία η κάθε μια. Διαθέτει μια μεγάλη ποικιλία συναρτήσεων οι οποίες αφορούν, αφενός σε απλές μαθηματικές συναρτήσεις (π.χ. συνημίτονο, τετραγωνική ρίζα), αφετέρου και σε στατιστικές συναρτήσεις (π.χ. διακύμανση, μέση τιμή), αλλά και σε πιο πολύπλοκες μαθηματικές διαδικασίες (π.χ. επίλυση εξισώσεων). Ένα από τα μεγαλύτερα πλεονεκτήματα της R είναι, ότι επιτρέπει στον χρήστη να ορίσει δικές του συναρτήσεις για να επιλύσει το πρόβλημα με το οποίο ασχολείται, αλλά επιπλέον και να τροποποιήσει ήδη υπάρχουσες συναρτήσεις.

Οι συναρτήσεις ορίζονται, χρησιμοποιώντας την λέξη `function` και αποθηκεύονται και αυτές ως αντικείμενα, όπως γίνεται σχεδόν με τα πάντα στην R. Πιο συγκεκριμένα, οι συναρτήσεις είναι αντικείμενα που ανήκουν στην κλάση “function”. Στη συνέχεια θα δούμε ότι συναρτήσεις μπορούν να χρησιμοποιηθούν ως ορίσματα άλλων συναρτήσεων.

Η βασική ιδέα, για να χρησιμοποιήσουμε συναρτήσεις, είναι να «σπάσουμε» ένα πρόβλημα σε πολλά μικρότερα προβλήματα. Ορισμένα πλεονεκτήματα της χρήσης συναρτήσεων δίνονται παρακάτω:

- Οι συναρτήσεις θεωρούνται ως μικρές και αυτόνομες. Έτσι μπορούν να επαληθευτούν και να διορθωθούν πιο εύκολα.
- Χρησιμοποιώντας συναρτήσεις αποφεύγουμε την επανάληψη της ίδιας ομάδας εντολών πολλές φορές.
- Επίσης με χρήση των συναρτήσεων μπορούμε να βελτιώσουμε τη συνάρτηση μια

φορά αντί να ψάχνουμε μέσα στο πρόγραμμα τις ομάδες εντολών και να κάνουμε πολλές διορθώσεις.

- Τέλος, χρησιμοποιώντας συναρτήσεις βοηθάμε στην καλύτερη διαχείριση της μνήμης. Στο παράδειγμα που ακολουθεί βλέπουμε τη δήλωση μιας απλής συνάρτησης, η οποία δέχεται έναν αριθμό και εκτυπώνει τόσες φορές ένα μήνυμα.

```
> myPrinter<-function(y){  
+  
+   for(i in seq_len(y)){  
+     print("Hello world!!")  
+   }  
+ }  
>  
> myPrinter(4)  
[1] "Hello world!!"  
[1] "Hello world!!"  
[1] "Hello world!!"  
[1] "Hello world!!"
```

Εικόνα 1.16: Ορισμός συνάρτησης από τον χρήστη

Κατά τη δήλωση μιας συνάρτησης μπορούμε να δώσουμε προεπιλεγμένες (default) τιμές, σε περίπτωση που ο χρήστης δεν δώσει τιμή για το αντίστοιχο όρισμα της συνάρτησης. Στο προηγούμενο παράδειγμα, αν ο χρήστης καλούσε τη συνάρτηση χωρίς όρισμα, τότε το μήνυμα θα εκτυπωνόταν 4 φορές, αφού η προεπιλεγμένη τιμή του ορίσματος είναι το 4. Μπορούμε να καλέσουμε μια συνάρτηση, δίνοντας τα ορίσματα με διαφορετική σειρά, εφόσον όμως αναφέρουμε ρητά τα ονόματα τους.

```
> volume<-function(y=4, z=4, k=4){  
+   print(y*z*k)  
+ }  
> volume(z=4, k=5, y=12)  
[1] 240  
> volume()  
[1] 64
```

Εικόνα 1.17: Πέρασμα ορισμάτων με διαφορετική σειρά

Στην περίπτωση που ο αριθμός των ορισμάτων μιας συνάρτησης δεν είναι σταθερός, χρησιμοποιούμε το όρισμα "...", το οποίο μεταφράζεται ως 1 ή περισσότερα ορίσματα ακολουθούν. Το όρισμα "." χρησιμοποιείται ως έχει μέσα στη συνάρτηση, όπως θα δούμε και στο παράδειγμα που ακολουθεί. Ένα θέμα που προκύπτει από τη χρήση του ορίσματος "." είναι ότι όσα ορίσματα ακολουθούν μετά από αυτό θα πρέπει να αναφερθούν ρητά κατά το πέρασμα τιμών στη συνάρτηση.**[Error! Unknown switch argument.][7][8]**

```

> myPrinter<-function(..., mes){
+ print(sum(...))
+ print(mes)
+ }
> myPrinter(4, 9, 15, mes = "Hello")
[1] 28
[1] "Hello"

```

Εικόνα 1.18: Χρήση του ορίσματος ... σε συνάρτηση

1.10 ΑΠΟΘΗΚΕΥΣΗ ΑΝΤΙΚΕΙΜΕΝΩΝ ΣΤΗΝ R

Το R παρέχει την δυνατότητα αποθήκευσης των αντικειμένων τα οποία αποθηκεύονται είτε σε ένα προκαθορισμένο πρότυπο που χρησιμοποιεί το R είτε σε ένα απλό αρχείο κειμένου. Η πλέον εύκολη αποθήκευση αντικειμένων γίνεται με την εντολή:

```
> save (zz, file= «zz.RData»),
```

όπου το zz είναι το όνομα του αντικειμένου και στο «zz.RData» το όνομα του αρχείου που θα αποθηκευτεί. Για να μπορούμε να χρησιμοποιήσουμε το αποθηκευμένο αντικείμενο και σε άλλα στατιστικά πακέτα προσθέτουμε άλλη μια παράμετρο:

```
> save (zz, file= «zz.RData», ascii = TRUE),
```

που σημαίνει ότι θα το αποθηκεύσει ως ένα απλό αρχείο κειμένου.

Για να «φορτώσουμε», επανακτήσουμε, το προηγούμενο αντικείμενο δίνουμε την εντολή:

```
> load («zz.RData»).^1
```

Ένας άλλος τρόπος αποθήκευσης αντικειμένων είναι και με τη χρήση της εντολής **write()** αρκεί να δώσουμε πλήρη διαδρομή αποθήκευσης. Με την εντολή αυτή έχουμε να αποθηκεύσουμε διανύσματα (αριθμητικά, χαρακτήρων ή λογικά), να κάνουμε αποθήκευση σε δισδιάστατους πίνακες με τη διαφορά ότι η εντολή αναστρέφει το πίνακα και άρα είναι σαν κάνουμε αποθήκευση του ανάστροφου πίνακα και τέλος μπορούμε να αποθηκεύουμε και πλαίσια δεδομένων αλλά η εντολή τροποποιείται σε **write.table**.

Όπως και στην αποθήκευση έτσι και στην ανάκτηση πρέπει να δώσουμε την πλήρη διαδρομή του αρχείου από όπου θέλουμε να ανακτήσουμε δεδομένα, εκτός και αν το αρχείο βρίσκεται στο φάκελο που δουλεύουμε την R οπότε το όνομά του αρκεί. Για διανύσματα και δισδιάστατους πίνακες χρησιμοποιούμε την εντολή **scan()**.

Επιπλέον υπάρχουν και μερικές συναρτήσεις για εισαγωγή δεδομένων στην R και μάλιστα θεωρούνται και πιο δημοφιλείς.[12]

- **read.csv**, για αρχεία στα οποία τα δεδομένα διαχωρίζονται με κόμμα.

¹ https://mathbooksgsr.files.wordpress.com/2011/11/r_giagos.pdf

- **read.delim**, για δεδομένα τα οποία βρίσκονται σε αρχεία διαχωρισμένα με στηλοθέτες.
- **readLines**, για εισαγωγή γραμμών ενός αρχείου.

Για πλαίσια δεδομένων χρησιμοποιούμε την εντολή **read.table()**.
[Error! Unknown switch argument.]

- **file**: το όνομα του αρχείου, ή η πλήρης διαδρομή αυτού.
- **header**: δηλώνει αν στην πρώτη γραμμή του αρχείου δίνονται τα ονόματα ή όχι.
- **sep**: δηλώνει πως γίνεται ο διαχωρισμός των δεδομένων.
- **colClasses**: ένα διάνυσμα χαρακτήρων, το οποίο δηλώνει τον τύπο των δεδομένων για κάθε στήλη.
- **nrows**: ο αριθμός των γραμμών του πλαισίου δεδομένων.
- **comment.char**: ένα διάνυσμα χαρακτήρων, το οποίο δηλώνει τον χαρακτήρα, που αποτελεί σχόλιο.
- **skip**: ο αριθμός των γραμμών, που θέλουμε να εξαιρεθούν από την αρχή του πλαισίου δεδομένων.
- **Strings As Factors**: θα έπρεπε τα διανύσματα χαρακτήρων να μετατραπούν σε κατηγορικά διανύσματα?

ΚΕΦΑΛΑΙΟ 2

2 ΔΕΔΟΜΕΝΑ PANEL

2.1 ΕΙΣΑΓΩΓΗ

Στην εργασία αυτή θα ασχοληθούμε με την ανάλυση δεδομένων πάνελ με χρήση της γλώσσας R. Αδιαμφισβήτητα, η επιτυχία κάθε εμπειρικής ανάλυσης εξαρτάται από τη διαθεσιμότητα των δεδομένων. Γενικά, υπάρχουν τρεις τύποι δεδομένων, που χρησιμοποιούνται για εμπειρική ανάλυση: α) τα δεδομένα χρονοσειρών(times series data), β) τα διαστρωματικά δεδομένα(cross- section data) καθώς και γ) τα πάνελ δεδομένα (panel data).[6]

Τα δεδομένα χρονοσειρών και τα διαστρωματικά δεδομένα μπορούν να θεωρηθούν ως ειδικές περιπτώσεις δεδομένων panel που βρίσκονται σε μία μόνο διάσταση (μία οντότητα ή ένα άτομο των δεδομένων panel αποτελούν δεδομένα χρονοσειρών, ενώ ένα χρονικό σημείο αποτελούν διαστρωματικά δεδομένα panel).

Συγκεκριμένα, τα δεδομένα χρονοσειρών εξετάζουν τις τιμές μιας ή περισσότερων μεταβλητών κατά τη διάρκεια μιας χρονικής περιόδου (λ.χ. εξέταση του Α.Ε.Π. για αρκετά τρίμηνα ή έτη). Ενώ στα διαστρωματικά δεδομένα, οι τιμές μιας ή περισσότερων μεταβλητών συλλέγονται για αρκετά δείγματα μονάδων ή οντοτήτων, την ίδια χρονική στιγμή (λ.χ. εξέταση του Α.Ε.Π. των χωρών που βρίσκονται στην ευρωζώνη για ένα έτος).[6]

2.2 ΟΡΙΣΜΟΣ ΚΑΙ ΠΕΡΙΓΡΑΦΗ ΔΕΔΟΜΕΝΩΝ PANEL

Στη στατιστική και την οικονομετρία, τα δεδομένα πάνελ, ή αλλιώς δυναμικά διαστρωματικά στοιχεία, αποτελούν ένα σύνολο δεδομένων στο οποίο η συμπεριφορά των εξεταζόμενων οντοτήτων (ή στρωμάτων), π.χ. κρατών, εταιρειών, ιδιωτών κλπ, παρατηρείται στο πέρασμα του χρόνου. Όταν η χρονική περίοδος αναφοράς των οντοτήτων είναι η ίδια τότε έχουμε χρονικά ισορροπημένα (balanced) δυναμικά διαστρωματικά στοιχεία. Μπορούμε όμως να αναλύσουμε και περιπτώσεις με διαφορετικές χρονικές περιόδους ανά οντότητα. Δεδομένα τέτοιας μορφής έχουν πιο πολύπλοκη εφαρμογή σε παλινδρομήσεις για οικονομετρικά μοντέλα σε σχέση με τα απλά σύνολα δεδομένων, παρ' όλα αυτά χρησιμοποιούνται όλο και περισσότερο στην εφαρμοσμένη έρευνα.

Οι Baltagi (2001) και Hsiao (2003) ισχυρίστηκαν ότι τα δεδομένα πάνελ δίνουν λύση στο πρόβλημα του στατιστικού σφάλματος που προκύπτει από την απαρατήρητη ετερογένεια, ένα πρόβλημα που προκύπτει πολύ συχνά όταν χρησιμοποιούμε διαστρωματικά δεδομένα για τις εκτιμήσεις μας. Επιπλέον, η χρήση των δεδομένων πάνελ στα μοντέλα παλινδρόμησης βοηθά να προσδιοριστεί εύκολα η χρονικά μεταβαλλόμενη σχέση μεταξύ εξαρτημένης και ανεξάρτητης μεταβλητής.[15][20]

Επιπλέον, ένας άλλος λόγος για το μεγάλο επιστημονικό ενδιαφέρον σχετικά με τα δεδομένα πάνελ είναι ότι με τη χρήση αυτών, μπορούν να αναδειχθούν δυναμικές που θα ήταν δύσκολο να εντοπιστούν με χρήση διαστρωματικών δεδομένων. Για παράδειγμα, εάν κάποιος έχει διαστρωματικά δεδομένα για έναν αριθμό ενηλίκων, θα διαπιστώσει ότι μερικοί είναι εργαζόμενοι, κάποιιοι είναι άνεργοι, και οι υπόλοιποι είναι οικονομικά ανενεργοί. Για λόγους πολιτικής, θα θέλαμε να κάνουμε διάκριση μεταξύ της προσωρινής ανεργίας και της μακροχρόνιας ανεργίας. Η προσωρινή ανεργία είναι αναπόφευκτη σε μια μεταβαλλόμενη οικονομία, αλλά η μακροχρόνια ανεργία αναδεικνύει ένα κοινωνικό πρόβλημα που πρέπει να αντιμετωπιστεί. Προκειμένου να υπάρξει ο σχεδιασμός μιας αποτελεσματικής πολιτικής για την αντιμετώπιση της μακροπρόθεσμης ανεργίας θα πρέπει να γνωρίζουμε τα χαρακτηριστικά των ατόμων που πλήττονται ή κινδυνεύουν.

Αρχικά, οι απαραίτητες πληροφορίες θα μπορούσαν να συλλεχθούν με μια έρευνα διαστρωματικών δεδομένων χρησιμοποιώντας ερωτήσεις σχετικά με το καθεστώς απασχόλησης κατά το παρελθόν. Αλλά πρακτικά αυτό είναι περιοριστικό. Όσο πιο πίσω πηγαίνουμε στο παρελθόν, τόσο χειρότερα είναι τα προβλήματα της έλλειψης των απαραίτητων πληροφοριών και τόσο μεγαλύτερο γίνεται το πρόβλημα του σφάλματος μέτρησης. Οι μελέτες δεδομένων πάνελ αποφεύγουν αυτό το πρόβλημα, καθώς η ανάγκη ανάκλησης πληροφοριών περιορίζεται στο χρονικό διάστημα από την προηγούμενη συνέντευξη που συχνά είναι όχι περισσότερο από ένας χρόνος.

Ένας τρίτος λόγος για τη χρήση δεδομένων πάνελ είναι ότι συνήθως έχουν πολύ μεγάλο αριθμό παρατηρήσεων. Εάν υπάρχουν n μονάδες παρατήρησης και εάν η έρευνα διεξάγεται σε T χρονικές περιόδους, υπάρχουν δυνητικά nT παρατηρήσεις που αποτελούνται από χρονοσειρές μήκους T σε n παράλληλες μονάδες.

Για παράδειγμα, έστω ότι έχουμε ένα σύνολο δεδομένων πάνελ, που αφορούν τους μισθούς ατόμων, τις ώρες εργασίας, το επίπεδο εκπαίδευσης, καθώς και άλλους παράγοντες. Αφού συγκεντρώσουμε τα δεδομένα με τυχαία επιλογή των προσώπων από έναν πληθυσμό σε μία δεδομένη χρονική στιγμή, στη συνέχεια από τα ίδια άτομα

λαμβάνουμε επανειλημμένες συνεντεύξεις σε διαδοχικές χρονικές στιγμές. Με αυτό τον τρόπο προκύπτουν τα δεδομένα για τους μισθούς, τις ώρες εργασίας, το επίπεδο εκπαίδευσης και ούτω καθεξής, για την ίδια μονάδα ατόμων σε διαφορετικές χρονικές περιόδους. Επομένως, συμβολίζοντας τις παρατηρήσεις με x_{it} όπου $i=1,2,\dots,n$ η κάθε διαστρωματική μονάδα και $t=1,\dots,T_i$ οι χρονικές περίοδοι που έλαβε χώρα η έρευνα, τα δεδομένα θα έχουν την παρακάτω μορφή:

1^η μονάδα ενδιαφέροντος - $\{x_{11}, x_{12}, \dots, x_{1T_1}\}$

2^η μονάδα ενδιαφέροντος - $\{x_{21}, x_{22}, \dots, x_{2T_2}\}$

.....

n^η μονάδα ενδιαφέροντος - $\{x_{n1}, x_{n2}, \dots, x_{nT_n}\}$.

Ο δείκτης T_i , καθώς παρέχει την πληροφορία για το πλήθος των χρονικών περιόδων που έλαβε χώρα η έρευνα στην ουσία δείχνει το πλήθος των παρατηρήσεων που έχει συλλεχθεί για την κάθε διαστρωματική μονάδα. Στην περίπτωση που όλες οι μονάδες έλαβαν μέρος στην έρευνα σε όλες τις χρονικές περιόδους και άρα $T_1=T_2=\dots=T_n=T$ τότε το πάνελ δεδομένων ονομάζεται *ισορροπημένο πάνελ (balanced panel)* και το μέγεθος του δείγματος είναι ίσο με $N=nT$, ενώ αν το πλήθος των παρατηρήσεων διαφέρει από μονάδα σε μονάδα τότε το πάνελ δεδομένων ονομάζεται *μη-ισορροπημένο (unbalanced panel)* και το μέγεθος του δείγματος είναι ίσο με $T_1+T_2+\dots+T_n$. [9]

Στην εικόνα 2.1 παρουσιάζεται ένα μέρος του συνόλου των δεδομένων της παρούσας μελέτης το οποίο μας βοηθά να κατανοήσουμε την μορφή των panel data.

Εταιρεία	Έτος	Αριθμός υπαλλήλων	Αριθμός μετόχων	Αριθμός δωματίων ξενοδοχείου	Παρατήρηση
1	2012	130	5	292	
1	2011	130	5	292	
1	2010	130	5	292	
1	2009	130	5	292	
1	2008	145	5	292	
1	2007	145	5	292	
1	2006	145	5	292	
1	2005	145	5	292	
1	2004	145	5	292	
2	2012	112	4	98	
2	2011	112	4	98	
2	2010	165	4	98	
2	2009	200	4	98	
2	2008	225	4	98	
2	2007	225	4	98	
2	2006	225	4	98	
2	2005	225	4	98	
2	2004	225	4	98	

Εικόνα 2.1: Παράδειγμα δεδομένων Panel

Ουσιαστικά θα πρέπει να φροντίσουμε ώστε η διάταξη των δεδομένων να γίνει με τέτοιο

τρόπο ώστε να μπορεί να γίνει εύκολα σύνδεση των διαφορετικών χρονικών περιόδων για την ίδια μονάδα διαστρωμάτωσης (άτομο, επιχείρηση, πόλη, κοκ).

2.3 ΠΛΕΟΝΕΚΤΗΜΑΤΑ ΚΑΙ ΜΕΙΟΝΕΚΤΗΜΑΤΑ

Τα Πλεονεκτήματα των δεδομένων panel είναι τα ακόλουθα:

1. Καταγραφή της ατομικής ανομοιογένειας (*individual heterogeneity*). Η ύπαρξη *panel data* δηλώνει από μόνη της, ότι οι οικονομικές μονάδες, οι περιφέρειες ή οι χώρες έχουν διαφορετικά γηγενή χαρακτηριστικά. Η χρήση διαστρωματικών στοιχείων ή απλών χρονολογικών σειρών αγνοεί αυτή τη διαφορετικότητα μεταξύ τους, γεγονός που μπορεί να μας οδηγήσει σε μεροληπτικές εκτιμήσεις (Moulton, 1986; 1987). Παραδείγματος χάρη οι Baltagi και Levin (1992) προσπάθησαν να διερευνήσουν την κατανάλωση τσιγάρων σε 46 αμερικανικές πολιτείες κατά την περίοδο 1963-1988. Η κατανάλωση τσιγάρων στο εμπειρικό τους υπόδειγμα εκφράστηκε σαν συνάρτηση της κατανάλωσης της προηγούμενης περιόδου, της τιμής των τσιγάρων και του ατομικού εισοδήματος. Όλες αυτές οι μεταβλητές μεταβάλλονται τόσο ως προς τον χρόνο όσο και ως προς τις πολιτείες. Παρόλα αυτά υπάρχουν όμως αρκετοί παράγοντες οι οποίοι επηρεάζουν επίσης την κατανάλωση τσιγάρων οι οποίοι όμως δεν μεταβάλλονται ούτε ως προς τον χρόνο ούτε ως προς τις επιμέρους πολιτείες. Παραδείγματα τέτοιων μεταβλητών οι οποίες δεν μεταβάλλονται με το πέρασμα των ετών είναι το θρήσκευμα και το επίπεδο εκπαίδευσης του πληθυσμού σε κάθε πολιτεία. Από την άλλη πλευρά οι διαφημιστικές εκστρατείες των καπνοβιομηχανιών λαμβάνουν χώρα σε ομοσπονδιακό επίπεδο και επομένως δεν διαφοροποιούνται ως προς τις επιμέρους πολιτείες. Όμως και οι τρεις αυτές μεταβλητές αναμένουμε να επηρεάζουν *apriori* το επίπεδο κατανάλωσης τσιγάρων. Επιπλέον είναι πιθανόν να υπάρχουν κάποιες άλλες μεταβλητές οι οποίες να επηρεάζουν την κατανάλωση αλλά να μην είναι εύκολα μετρήσιμες ή προσβάσιμες. Ο αποκλεισμός όμως τέτοιων μεταβλητών από την συνάρτηση κατανάλωσης μπορεί να οδηγήσει σε σφάλματα κατά την οικονομετρική της εκτίμηση. Η χρήση *panel data* δίνει την δυνατότητα να λάβουμε έμμεσα υπόψη μας κατά την οικονομετρική εκτίμηση του υποδείγματος όλες εκείνες τις μεταβλητές οι οποίες δεν μπορούν αντικειμενικά να συμπεριληφθούν στο εμπειρικό υπόδειγμα άμεσα. [16][23]
2. Η χρήση δεδομένων *panel* μας παρέχει περισσότερες πληροφορίες για τις οικονομικές μονάδες, περισσότερη διακύμανση στις μεταβλητές, λιγότερη συγγραμμικότητα

μεταξύ των μεταβλητών, περισσότερους βαθμούς ελευθερίας και περισσότερη αποτελεσματικότητα στην οικονομετρική εκτίμηση. Τα οικονομετρικά υποδείγματα που βασίζονται αποκλειστικά σε στοιχεία χρονολογικών σειρών υποφέρουν σημαντικά από την ύπαρξη έντονης πολυσυγγραμμικότητας (*multicollinearity*). Στο παράδειγμα της κατανάλωσης τσιγάρων υπάρχει έντονη συγγραμμικότητα διαχρονικά μεταξύ της τιμής των τσιγάρων και του ατομικού εισοδήματος. Κάτι τέτοιο είναι όμως λιγότερο πιθανό να συμβεί όταν τα στατιστικά δεδομένα αφορούν και τις επιμέρους πολιτείες για την ίδια χρονική περίοδο, καθώς η διαστρωματική διάσταση προσδίδει περισσότερη διακύμανση στις μεταβλητές καθώς και περισσότερες πληροφορίες αναφορικά με τις τιμές και το εισόδημα. [5]

3. Η χρήση δεδομένων *panel* μας παρέχει την δυνατότητα να αναλύσουμε την διαδικασία και την δυναμική προσαρμογής των οικονομικών δεδομένων διαχρονικά, αφού διατηρήσουμε παράλληλα την ατομική ετερογένεια στα χαρακτηριστικά των οικονομικών μονάδων. Παραδείγματος χάρι για την μέτρηση της ανεργίας η χρήση των απλών διαστρωματικών στοιχείων μπορεί να προσδιορίσει το επίπεδο ανεργίας σε κάθε δεδομένη χρονική στιγμή. Αντιθέτως η χρήση δεδομένων *panel* επιτρέπει την ανάλυση του επιπέδου ανεργίας για μία δεδομένη χρονική περίοδο και πώς αυτή συνδέεται με τα ιδιαίτερα χαρακτηριστικά των οικονομικών μονάδων.[5]
4. Επιπλέον η ύπαρξη δεδομένων *panel* επιτρέπει την εμπειρική εξειδίκευση περισσότερο πολύπλοκων θεωρητικών υποδειγμάτων για την ανάλυση της συμπεριφοράς των οικονομικών μονάδων, η οποία είναι αδύνατη με τη χρήση αποκλειστικά χρονολογικών σειρών ή διαστρωματικών δεδομένων.[5]
5. Τέλος, λαμβάνοντας υπόψη ότι τα στατιστικά δεδομένα συλλέγονται σε επίπεδο οικονομικής μονάδας μειώνεται σημαντικά το ενδεχόμενο σφάλματος κατά τη διαδικασία άθροισης των δεδομένων σε εθνικό ή περιφερειακό επίπεδο.[5]

Μειονεκτήματα [Error! Unknown switch argument.]

Τα δεδομένα πάνελ προσφέρουν πολλά πλεονεκτήματα σε σχέση με τα αμιγώς διαστρωματικά δεδομένα ή τα αμιγώς δεδομένα χρονοσειρών. Παρόλα αυτά όμως παρουσιάζουν και κάποια μειονεκτήματα κυρίως όσον αφορά στη δημιουργία τους και στην ανάλυσή τους. Τα μειονεκτήματα αυτά παρουσιάζονται στη συνέχεια.

1. Ένα από τα κυριότερα προβλήματα είναι αυτό της φθοράς (*attrition*), όπου κάποια ή κάποιες από τις μονάδες που μελετώνται αποχωρεί από το πείραμα πριν από τη λήξη

του. Αιτία της «φθοράς» μπορεί να είναι η χρεοκοπία κάποιας ή κάποιων επιχειρήσεων, η απώλεια κάποιου ή κάποιων ανθρώπων, η συγχώνευση οργανισμών ή οικονομικών μονάδων, η αποστροφή κάποιων μονάδων του πειράματος για την έρευνα κλπ. Αν η φθορά έχει συστηματικό χαρακτήρα όπως για παράδειγμα αν από τον κλάδο τροφίμων αποχωρήσουν για κάποιο λόγο όλες οι εταιρίες που κατασκευάζουν τυροκομικά προϊόντα, τότε υπάρχει εμφάνιση μεροληψίας επιλογής και το σύνολο των δεδομένων πάνελ να μην είναι τόσο αντιπροσωπευτικό αφού δεν μπορεί να χαρακτηριστεί τυχαίο.

2. Η ύπαρξη σφαλμάτων επιλεκτικότητας (*selectivity problems*), που προκύπτουν είτε από τη μη απόκριση των επιλεγμένων διαστρωματικών μονάδων είτε από την υποκειμενικότητα της κάθε διαστρωματικής μονάδας.
3. Η ύπαρξη των σφαλμάτων μέτρησης, που μπορεί να οφείλονται σε ασαφείς ερωτήσεις, σε σκόπιμη διαστρέβλωση των απαντήσεων, σε κακή διαλογή των ερωτώμενων, στη μη σωστή καταγραφή των απαντήσεων από τον ερευνητή, καθώς και σε διαφορετικές μονάδες μέτρησης. Αυτό είναι ένα γενικότερο πρόβλημα στη συλλογή δεδομένων.

2.4 ΜΟΝΤΕΛΑ ΔΕΔΟΜΕΝΩΝ ΠΑΝΕΛ

Σε αυτή την ενότητα θα παρουσιάσουμε μοντέλα ανάλυσης δεδομένων πάνελ που απαντώνται πιο συχνά στην βιβλιογραφία. Όπως έχουμε αναφέρει, τα δεδομένα πάνελ περιέχουν πληροφορίες τόσο για τη χρονική όσο και για τη διαστρωματική διάσταση. Η χρονική διάσταση είναι η περίοδος κατά την οποία πραγματοποιούνται οι μετρήσεις, όπως για παράδειγμα μήνες, τρίμηνα, έτη κλπ. Η χωρική διάσταση είναι η διαστρωματική μονάδα των παρατηρήσεων όπως άνθρωποι, επιχειρήσεις, χώρες κλπ.

Το γενικό μοντέλο παλινδρόμησης δεδομένων πάνελ μπορεί να εκφραστεί ως

$$Y_{it} = a + \beta_1 X_{it,1} + \beta_2 X_{it,2} + \dots + \beta_k X_{it,k} + u_{it}$$

ή ισοδύναμα

$$Y_{it} = a + \beta' X_{it} + u_{it}, \quad (2.1)$$

όπου ο δείκτης $i = 1, 2, \dots, n$, αναφέρεται στις διαστρωματικές μονάδες και $t = 1, 2, \dots, T$ αναφέρεται στις χρονικές περιόδους για τις οποίες έχουμε παρατηρήσεις. Η παράμετρος a

είναι η τεταγμένη της αρχής, β είναι το διάνυσμα ($k \times 1$) των αγνώστων συντελεστών, X_{it} είναι η it -παρατήρηση της k ανεξάρτητης μεταβλητής και u_{it} είναι ο διαταρακτικός όρος του μοντέλου.

Στην οικονομετρία, η εκτίμηση μιας συνάρτησης παλινδρόμησης εξαρτάται από τις υποθέσεις που πραγματοποιούνται αναφορικά με τη τεταγμένη της αρχής a (intercept), τις κλίσεις (slope) β της ευθείας καθώς και τον όρο σφάλματος (error term) u_{it} . Επίσης, όπως έχουμε αναφέρει για την ανάλυση των δεδομένων πάνελ δεν μπορούμε να υποθέσουμε ότι οι παρατηρήσεις κατανέμονται ανεξάρτητα στο χρόνο. Για παράδειγμα, οι μη παρατηρούμενοι παράγοντες (όπως η ικανότητα ενός ατόμου) που επηρεάζουν το μισθό κάποιου το 1990 θα επηρεάσουν και το μισθό του ίδιου ατόμου το 1991. Οι μη παρατηρούμενοι παράγοντες που επηρεάζουν το ποσοστό εγκληματικότητας μιας πόλης το 1985 θα επηρεάσουν και το δείκτη εγκληματικότητας της ίδιας πόλης το 1990. Γι' αυτό το λόγο, έχουν αναπτυχθεί ειδικά μοντέλα και μέθοδοι για την ανάλυση των δεδομένων πάνελ.[6][28]

Οι μέθοδοι αυτές περιλαμβάνουν το μοντέλο ενιαίων δεδομένων (pooled model), το μοντέλο σταθερών επιδράσεων (Fixed Effect model) και μοντέλο τυχαίων επιδράσεων(Random Effect Model).

2.4.1 Το Ενιαίο Μοντέλο (Pooled Model)

Η πιο απλή και βασική μέθοδος εκτίμησης της εξίσωσης (2.1) είναι να θεωρήσουμε όλα τα δεδομένα ως μια ενιαία ομάδα και να εφαρμόσουμε τη μέθοδο ελαχίστων τετραγώνων OLS. Για να εκτιμήσουμε την εξίσωση (2.1) χρησιμοποιώντας τη μέθοδο OLS πρέπει να υποθέσουμε ότι μεταξύ του διαταρακτικού όρου u_{it} και των ερμηνευτικών μεταβλητών X_{it} δεν υπάρχει συσχέτιση [26]. Αυτό σημαίνει ότι μόνο όταν δεν υπάρχουν διαστρωματικές και διαχρονικές επιδράσεις, μπορούμε να εφαρμόσουμε το ενιαίο μοντέλο.

Η εξίσωση παλινδρόμησης (2.1) για το ενιαίο μοντέλο (pooled model) παίρνει τη μορφή

$$Y = a + \beta'X + u, \quad (2.2)$$

όπου οι δείκτες i και t εξαλείφθηκαν από την εξίσωση λόγω της παραπάνω υπόθεσης.

Το ενιαίο μοντέλο έχει κάποια μειονεκτήματα. Τα δεδομένα πάνελ περιέχουν τις πληροφορίες τόσο για τη διαστρωματική όσο και για τη διαχρονική διάσταση. Το ενιαίο μοντέλο όμως αγνοεί αυτή την πληροφορία. Επιπλέον, η υπόθεση των ομαδοποιημένων

OLS δεν είναι ρεαλιστική, επειδή δεν είναι δυνατόν να μετρηθούν όλες οι χρονικά-αμετάβλητες ατομικές επιδράσεις ή/και οι διαστρωματικά-αμετάβλητες διαχρονικές επιδράσεις, και να συμπεριληφθούν στο μοντέλο. Έτσι χρησιμοποιώντας το ενιαίο μοντέλο για την ανάλυση δεδομένων πάνελ και αγνοώντας τις αμετάβλητες (χρονικά ή διαστρωματικά) επιδράσεις οδηγούμαστε σε μη αποτελεσματικές εκτιμήσεις και στην περίπτωση που οι διαστρωματικές επιδράσεις συσχετίζονται με κάποιες από τις ερμηνευτικές μεταβλητές σε μεροληπτικές εκτιμήσεις [26]. Παρόλα αυτά, το ενιαίο μοντέλο χρησιμοποιείται συχνά στη βιβλιογραφία επειδή η ομαδοποίηση διαστρωματικών δεδομένων αυξάνει τον αριθμό των παρατηρήσεων σημαντικά και σε πολλές περιπτώσεις, εξασφαλίζει περισσότερες παραλλαγές των βασικών ανεξάρτητων μεταβλητών, εξασφαλίζοντας με αυτό τον τρόπο κατάλληλες εκτιμήσεις[6][28].

2.4.2 Το Μοντέλο Σταθερών Επιδράσεων (Fixed Effect Model)

Μια εναλλακτική μέθοδο για την ανάλυση δεδομένων πάνελ είναι το *μοντέλο σταθερών επιδράσεων (Fixed Effect Model)* το οποίο απαιτεί λιγότερο περιοριστικές υποθέσεις. Το μοντέλο αυτό υποθέτει ότι όλες οι καταχωρήσεις ανάμεσα στις διαστρωματικές μονάδες (οντότητες) μπορούν να διαφέρουν αλλά παραμένουν σταθερές με την πάροδο του χρόνου και έχει σχεδιαστεί για να μελετά τις αιτίες των αλλαγών μέσα σε κάθε διαστρωματική μονάδα. Σύμφωνα με το μοντέλο αυτό, όλες οι διαστρωματικές μονάδες (π.χ. επιχειρήσεις, χώρες κλπ) έχουν ίδιους συντελεστές για κάθε μεταβλητή της εξίσωσης εκτός από τη τεταγμένη της αρχής a . Πιο συγκεκριμένα, αν υποθέσουμε ότι έχουμε $i = 1, 2, \dots, n$ διαστρωματικές μονάδες και $t = 1, 2, \dots, T$ χρονικές περιόδους, η εξίσωση γραμμικής παλινδρόμησης για το μοντέλο σταθερών επιδράσεων έχει την ακόλουθη μορφή:

$$Y = a_i + \beta' X_{it} + v_{it}, \quad (2.3)$$

όπου

Y_{it} = η τιμή της εξαρτημένης μεταβλητής για την επιχείρηση i στο χρόνο t

X_{it} = η τιμή της ανεξάρτητης ερμηνευτικής μεταβλητής για την επιχείρηση i στο χρόνο t

a_i = οι μη παρατηρούμενοι παράγοντες (απαρατήρητες επιδράσεις) που επηρεάζουν την εξαρτημένη μεταβλητή και οι οποίοι δεν μεταβάλλονται χρονικά

v_{it} = ο διαταρακτικός όρος που εκφράζει μη παρατηρούμενους παράγοντες που μεταβάλλονται διαχρονικά και επηρεάζουν την εξαρτημένη μεταβλητή

Το a_i ονομάζεται «ειδικές-διαστρωματικές επιδράσεις» (“the individual-specific effects”) ή

«σταθερή επίδραση στρώματος» (“state fixed effect”) και είναι η σταθερή (fixed) επίδραση του στρώματος i . Η κύρια υπόθεση που αφορά στην σταθερή επίδραση a_i είναι η ακόλουθη

$$Cov(a_i, X_{it}) \neq 0, i = 1, 2, \dots, k, \quad t = 1, 2, \dots, T \quad (2.4)$$

δηλαδή η συνδιακύμανση των ανεξάρτητων μεταβλητών και των παρατηρήσιμων επιδράσεων που επηρεάζουν την εξαρτημένη μεταβλητή είναι διάφορη του μηδενός. Το γεγονός αυτό έχει ως συνέπεια η μέθοδος ελαχίστων τετραγώνων να μη δίνει συνεπείς εκτιμητές των παραμέτρων.

Το μοντέλο σταθερών επιδράσεων είναι ένα κλασικό μοντέλο παλινδρόμησης. Η επιπλοκή για την εφαρμογή της μεθόδου ελαχίστων τετραγώνων προκύπτει από το γεγονός ότι το n μπορεί να είναι πολύ μεγάλο, έτσι ώστε οι συνήθεις τύποι για τον υπολογισμό των συντελεστών ελαχίστων τετραγώνων (OLS) είναι δύσκολο να εφαρμοστούν. Το μοντέλο μπορεί να εκτιμηθεί σε μια απλούστερη μορφή με την αξιοποίηση της άλγεβρας ελαχίστων τετραγώνων. Στη συνέχεια θα περιγραφούν τρεις εκδοχές του μοντέλου σταθερών επιδράσεων. Στις δύο πρώτες από αυτές, το μοντέλο διαμορφώνεται με τέτοιο τρόπο ώστε οι παρατηρήσιμες επιδράσεις a_i να εξαλειφθούν [6][22][28].

Μοντέλο Σταθερών Επιδράσεων “εντός” (Within-groups Fixed Effect Model)

Σε αυτό το μοντέλο οι παράμετροι κλίσης β μπορούν να εκτιμηθούν μετασχηματίζοντας τόσο την εξαρτημένη όσο και την ανεξάρτητη μεταβλητή σε αποκλίσεις από το μέσο όρο κάθε διαστρωματικής μονάδας (within variations) και διεξάγοντας τη μέθοδο ελαχίστων τετραγώνων χωρίς τον σταθερό όρο a_i . Πιο συγκεκριμένα, ορίζουμε τις ακόλουθες μεταβλητές:

$$\begin{aligned} \bar{Y}_i &= \frac{1}{T} \sum_{t=1}^T Y_{it}, & \bar{X}_i &= \frac{1}{T} \sum_{t=1}^T X_{it}, & \bar{v}_i &= \frac{1}{T} \sum_{t=1}^T v_{it} \\ \tilde{Y}_{it} &\equiv Y_{it} - \bar{Y}_i, & \tilde{X}_{it} &\equiv X_{it} - \bar{X}_i, & \tilde{v}_{it} &\equiv v_{it} - \bar{v}_i \end{aligned} \quad (2.5)$$

Στη συνέχεια για κάθε i , λαμβάνουμε τον μέσο όρο της σχέσης (2.3) διαχρονικά. Τελικά προκύπτει

$$\bar{Y}_i = a_i + \beta' \bar{X}_i + \bar{v}_i \quad (2.6)$$

Επειδή το a_i είναι σταθερό, εμφανίζεται και στη σχέση (2.3) και στη (2.6). Αν αφαιρέσουμε από τη σχέση (2.3) τη (2.6) παίρνουμε

$$Y_{it} - \bar{Y}_i = \beta'(X_{it} - \bar{X}_i) + (v_{it} - \bar{v}_i), \quad t = 1, 2, \dots, T$$

ή ισοδύναμα

$$\tilde{Y}_{it} = \beta \tilde{X}_{it} + \tilde{v}_{it}, \quad t = 1, 2, \dots, T \quad (2.7)$$

όπου τα $\tilde{Y}_{it}$ ορίζονται στη (2.5) και είναι τα **διαχρονικά προσαρμοσμένα δεδομένα** για τη μεταβλητή Y , ομοίως τα $\tilde{X}_{it}$ και $\tilde{v}_{it}$. Είναι σημαντικό ότι στην εξίσωση (2.7) οι απαραίτητες επιδράσεις α_i έχουν εξαλειφθεί. Αυτό δείχνει ότι μπορούμε να εκτιμήσουμε τη (2.7) με τη μέθοδο των ομαδοποιημένων ελαχίστων τετραγώνων OLS. Έτσι, ο **εκτιμητής σταθερών επιδράσεων “εντός”** προκύπτει αν εφαρμόσουμε τη μέθοδο ελαχίστων τετραγώνων στο μοντέλο με τους **μετασχηματισμούς “εντός”** (εξίσωση (2.7)). Εάν το μοντέλο περιέχει διαχρονικές επιδράσεις (αντί ή επιπλέον των διαστρωματικών επιδράσεων), οι διαχρονικές επιδράσεις μπορούν να εξαλειφθούν με παρόμοιο τρόπο όπως εξαλείφθηκαν οι διαστρωματικές επιδράσεις.[27]

Κάτω από μια υπόθεση για αυστηρά εξωγενείς ερμηνευτικές μεταβλητές, το μοντέλο των σταθερών επιδράσεων “εντός” δίνει αμερόληπτες εκτιμήσεις αφού σε γενικές γραμμές ο διαταρακτικός όρος v_{it} δεν πρέπει να σχετίζεται με κάθε ερμηνευτική μεταβλητή σε όλες τις χρονικές στιγμές (Για τις ακριβείς υποθέσεις του μοντέλου σταθερών επιδράσεων βλ. [26]). Το μοντέλο των σταθερών επιδράσεων “εντός” επιτρέπει την αυθαίρετη συσχέτιση ανάμεσα στις απαραίτητες επιδράσεις α_i και τις ερμηνευτικές μεταβλητές σε μια οποιαδήποτε χρονική περίοδο (βλ. σχ. (2.4)). Εξαιτίας αυτού του γεγονότος, κάθε ερμηνευτική μεταβλητή που είναι διαχρονικά σταθερή για όλα τα i εξαλείφεται με τον μετασχηματισμό “εντός”. Αυτό όμως είναι ένα μειονέκτημα για το μοντέλο των σταθερών επιδράσεων “εντός” αφού μεταβλητές όπως το “φύλο” ή η “εκπαίδευση” δεν μπορούν να συμπεριληφθούν σε αυτό.

Οι άλλες υποθέσεις που πρέπει να πληρούνται για να είναι αμερόληπτη η εκτίμηση ελαχίστων τετραγώνων για την εξίσωση (2.7) είναι ότι τα σφάλματα v_{it} παρουσιάζουν ομοσκεδαστικότητα και αυτοσυσχέτιση για όλα τα t . Η διακύμανση στη $(X_{it} - \bar{X}_i)$ όμως μπορεί κάλλιστα να είναι πολύ μικρότερη από την διακύμανση στη X_{it} . Σε αυτή την περίπτωση, η επίδραση του όρου διαταραχής v_{it} μπορεί να είναι σχετικά μεγάλη, προκαλώντας ανακριβείς εκτιμήσεις. Η κατάσταση επιδεινώνεται στην περίπτωση σφάλματος μέτρησης, καθώς αυτό θα οδηγήσει σε μεροληψία, και η μεροληψία είναι μεγαλύτερη όσο μικρότερη είναι η διακύμανση της επεξηγηματικής μεταβλητής σε σύγκριση με τη διακύμανση του σφάλματος μέτρησης.

Ένα τρίτο πρόβλημα είναι ότι χάνουμε έναν σημαντικό αριθμό βαθμών ελευθερίας όταν

χρησιμοποιούμε το μοντέλο για να εξαλείψουμε τις απαρατήρητες επιδράσεις: χάνουμε έναν βαθμό ελευθερίας για κάθε άτομο στο δείγμα. Όταν εκτιμάμε την χρονικά προσαρμοσμένη εξίσωση (2.7) με την μέθοδο των ομαδοποιημένων OLS, έχουμε nT συνολικά παρατηρήσεις και k ανεξάρτητες μεταβλητές (υπόψη ότι δεν υπάρχει τεταγμένη επί της αρχής στην (2.7)). Συνεπώς πρέπει να έχουμε $nT - k$ βαθμοί ελευθερίας. Ωστόσο, κατά τη χρήση του μοντέλου, ο αριθμός των βαθμών ελευθερίας μειώνεται κατά n . Πράγματι, για κάθε i , τα προσαρμοσμένα σφάλματα $\tilde{v}_{it}$ έχουν άθροισμα ίσο με το μηδέν όταν αθροιστούν για κάθε t και συνεπώς χάνουν ένα βαθμό ελευθερίας (δεν υπάρχει περιορισμός για τα αρχικά σφάλματα v_{it}). Έτσι, ο πραγματικός αριθμός βαθμών ελευθερίας θα είναι $nT - n - k = n(T - 1) - k$. Εάν το T είναι μικρό, η επίδραση μπορεί να είναι μεγάλη. Ευτυχώς, τα διάφορα λογισμικά παλινδρόμησης δεδομένων πάνελ κάνουν αυτόματα την προσαρμογή στους βαθμούς ελευθερίας κατά την εφαρμογή του μοντέλου σταθερών επιδράσεων “εντός”.

Μοντέλο Σταθερών Επιδράσεων “μεταξύ”(Between Fixed Effect Model)

Ο εκτιμητής σταθερών επιδράσεων “μεταξύ” χρησιμοποιεί τις μεταβολές μεταξύ των στρωμάτων και προκύπτει αν εφαρμόσουμε τη μέθοδο ελαχίστων τετραγώνων (OLS) στην εξίσωση (2.3) όπου συμπεριλαμβάνουμε μια τεταγμένη της αρχής a . Χρησιμοποιούμε τους χρονικούς μέσους για τη μεταβλητή Y_{it} και X και στη συνέχεια πραγματοποιούμε διαστρωματική παλινδρόμηση. Η εξίσωση παλινδρόμησης για το μοντέλο αυτό έχει την ακόλουθη μορφή:

$$\bar{Y}_i = a + \beta' \bar{X}_i + (a_i - a + \bar{v}_i) \quad (2.8)$$

Όπως είναι φανερό από τη (2.8) οι διαχρονικές μεταβολές δεν λαμβάνονται υπόψη και τα δεδομένα συσσωρεύονται σε μια παρατήρηση ανά στρώμα.

Όταν οι α_i σχετίζονται με τους $\bar{X}_i$ ο εκτιμητής σταθερών επιδράσεων “μεταξύ” είναι μεροληπτικός. Αλλά και στην περίπτωση που οι α_i δε σχετίζονται με τους $\bar{X}_i$ είναι προτιμότερο να χρησιμοποιήσουμε το ενιαίο μοντέλο ή τον εκτιμητή τυχαίων επιδράσεων με τον οποίο θα ασχοληθούμε παρακάτω. Τέλος, ο εκτιμητής σταθερών επιδράσεων “μεταξύ” δεν λαμβάνει υπόψη σημαντικές πληροφορίες για τη διαχρονική μεταβολή των μεταβλητών.

Μοντέλο Πρώτων Διαφορών (First differences model)

Στη τρίτη έκδοση της προσέγγισης σταθερών επιδράσεων, το μοντέλο παλινδρόμησης πρώτων διαφορών, η απαρατήρητη επίδραση a_i εξαλείφεται αφαιρώντας την παρατήρηση για την προηγούμενη χρονική περίοδο από την παρατήρηση για την τρέχουσα χρονική περίοδο, για όλες τις χρονικές περιόδους. Για τη διαστρωματική μονάδα i στη χρονική περίοδο t , το μοντέλο μπορεί να γραφτεί:

$$Y_{it} = a_i + \beta'X_{it} + v_{it}, \quad (2.8)$$

Για την προηγούμενη χρονική περίοδο, η σχέση γράφεται:

$$Y_{i,t-1} = a_i + \beta'X_{i,t-1} + v_{i,t-1}. \quad (2.9)$$

Αφαιρώντας τη (2.8) από τη (2.9), προκύπτει η παρακάτω εξίσωση:

$$Y_{it} - Y_{i,t-1} = \beta'(X_{it} - X_{i,t-1}) + (v_{it} - v_{i,t-1}),$$

Η ισοδύναμη

$$\Delta Y_{it} = \beta' \Delta X_{it} + (v_{it} - v_{i,t-1}), \quad (2.10)$$

όπου και πάλι οι απαρατήρητες επιδράσεις a_i έχουν εξαφανιστεί. Ωστόσο, τα υπόλοιπα προβλήματα παραμένουν. Συγκεκριμένα, η τεταγμένη επί της αρχής και οποιαδήποτε μεταβλητή X που παραμένει σταθερή για κάθε διαστρωματική μονάδα θα εξαφανιστούν και θα χαθούν n βαθμοί ελευθερίας από το μοντέλο. Αυτό συμβαίνει επειδή δεν ορίζεται η πρώτη παρατήρηση για κάθε διαστρωματική μονάδα. Επιπλέον, αυτός ο τύπος διαφορών πρώτης τάξης δημιουργεί αυτοσυσχέτιση εάν το v_{it} ικανοποιεί τις υποθέσεις του μοντέλου πολλαπλής παλινδρόμησης OLS. Το σφάλμα για τη διαφορά ΔY_{it} είναι $(v_{it} - v_{i,t-1})$ ενώ για την προηγούμενη παρατήρηση είναι $(v_{i,t-1} - v_{i,t-2})$. Έτσι, οι δύο όροι του σφάλματος έχουν και οι δύο μια κοινή συνιστώσα $v_{i,t-1}$ με αντίθετο πρόσημο. Με αυτόν τον τρόπο παράγεται ένας αρνητικός κινητός μέσος όρος αυτοσυσχέτισης. Ωστόσο, εάν το αρχικό σφάλμα v_{it} υπόκειται σε αυτοσυσχέτιση:

$$v_{it} = \rho v_{i,t-1} + \varepsilon_{it}$$

όπου το ε_{it} είναι καλά συμπεριφερόμενος παράγοντας, ο κινητός μέσος του σφάλματος είναι ίσος με $\varepsilon_{it} - (1 - \rho)v_{i,t-1}$. Επομένως, στην περίπτωση που έχουμε απόλυτη αυτοσυσχέτιση ($\rho \rightarrow 1$), ο όρος $(1 - \rho)v_{i,t-1}$ είναι μικρός και έτσι το μοντέλο πρώτων διαφορών είναι προτιμότερο από το μοντέλο σταθερών επιδράσεων “εντός”.

2.5 Το Μοντέλο Τυχαίων Επιδράσεων (Random Effect Model)

Όπως αναλύσαμε πιο πάνω, όταν οι επεξηγηματικές μεταβλητές είναι σταθερές για κάθε διαστρωματική μονάδα (π.χ. το φύλο), το μοντέλο σταθερών επιδράσεων δεν είναι αποτελεσματικό, επειδή τέτοιες μεταβλητές δεν μπορούν να συμπεριληφθούν σε αυτό. Σε αυτήν την ενότητα θα εξετάσουμε μια εναλλακτική προσέγγιση, γνωστή ως μοντέλο τυχαίων επιδράσεων που μπορεί, υπό δύο προϋποθέσεις, να παρέχει μια λύση σε αυτό το πρόβλημα.

Η πρώτη προϋπόθεση είναι ότι οι απαρατήρητες επιδράσεις a_i , μπορούν να αντιμετωπιστούν ως μια τυχαία μεταβλητή (εξ ου και το όνομα αυτού του μοντέλου) η οποία προέρχονται από μια δεδομένη κατανομή και μπορούμε να ξαναγράψουμε το μοντέλο ως

$$\begin{aligned} Y_{it} &= a + \beta'X_{it} + a_i + v_{it}, \\ &= a + \beta'X_{it} + \varepsilon_{it}, \end{aligned} \quad (2.11)$$

όπου

$$\varepsilon_{it} = a_i + v_{it}. \quad (2.12)$$

Με αυτό τον τρόπο συμπεριλάβαμε τις απαρατήρητες επιδράσεις a_i στον όρο σφάλματος.

Εδώ, τα ε_{it} ονομάζονται «σφάλματα ιδιοσυγκρασίας» (idiosyncratic errors) επειδή αλλάζουν με το χρόνο t αλλά και με τις διαστρωματικές μονάδες i .

Η δεύτερη προϋπόθεση που πρέπει να ισχύει είναι ότι οι απαρατήρητες επιδράσεις a_i είναι ανεξάρτητες από όλες τις ερμηνευτικές μεταβλητές X σε όλες τις χρονικές στιγμές. Εάν αυτό δεν συμβαίνει, τα a_i , και συνεπώς τα ε_{it} , δε θα είναι ασυσχέτιστα με τις ερμηνευτικές μεταβλητές X_{it} και κατά συνέπεια ο εκτιμητής των τυχαίων επιδράσεων θα είναι μεροληπτικός και ασυνεπής. Στην περίπτωση αυτή, θα πρέπει να χρησιμοποιήσουμε το μοντέλο σταθερών επιδράσεων, ακόμα κι αν πληρείται η πρώτη προϋπόθεση.

Εάν πληρούνται οι δύο αυτές προϋποθέσεις, μπορούμε να χρησιμοποιήσουμε τη (2.11) ως εκτιμητή, αλλά όμως οι όροι σφάλματος ε_{it} θα υπόκεινται σε μια ειδικής μορφής αυτοσυσχέτισης και θα πρέπει να χρησιμοποιήσουμε μια τεχνική εκτίμησης που θα λαμβάνει υπόψη της αυτό το γεγονός [6][22][28].

Αρχικά, ας ελέγξουμε τις άλλες συνθήκες του μοντέλου παλινδρόμησης που σχετίζονται με τον όρο διαταραχής. Δεδομένης της υπόθεσής μας ότι ο όρος v_{it} ικανοποιεί τις γνωστές συνθήκες του μοντέλου γραμμικής παλινδρόμησης (OLS), μπορούμε να δούμε ότι ο όρος

σφάλματος ε_{it} ικανοποιεί την συνθήκη ότι η αναμενόμενη τιμή του είναι μηδενική, αφού

$$E(\varepsilon_{it}) = E(a_i + v_{it}) = E(a_i) + E(v_{it}) = 0, \text{ για όλα τα } i \text{ και } t \quad (2.13)$$

Στο σημείο αυτό υποθέσαμε χωρίς απώλεια της γενικότητας ότι $E(a_i) = 0$, και όποτε οποιοδήποτε μη μηδενική συνιστώσα απορροφάται από την τεταγμένη της αρχής, a (για αυτό το λόγο τη συμπεριλάβαμε ρητά στο μοντέλο (2.11)). Επιπλέον, ο όρος ε_{it} θα πρέπει να ικανοποιεί την συνθήκη ότι η διακύμανση του είναι σταθερή, αφού

$$\sigma_{\varepsilon_{it}}^2 = \sigma_{a_i+v_{it}}^2 = \sigma_a^2 + \sigma_v^2 + 2\sigma_{av} = \sigma_a^2 + \sigma_v^2, \text{ για όλα τα } i \text{ και } t \quad (2.14)$$

Ο όρος σ_{av} μηδενίζεται με την προϋπόθεση ότι οια a_i κατανέμονται ανεξάρτητα από τους όρους v_{it} . Το ε_{it} θα ικανοποιεί επίσης την συνθήκη του μοντέλου γραμμικής παλινδρόμησης δηλαδή ότι κατανέμεται ανεξάρτητα από τις τιμές του X_{it} , αφού τόσο οια a_i όσο και οι v_{it} . θεωρείται ότι πληρούν αυτήν την συνθήκη.

Ωστόσο, υπάρχει ένα πρόβλημα με την προϋπόθεση που απαιτεί το μοντέλο παλινδρόμησης (OLS) ότι η τιμή του ε_{it} σε οποιαδήποτε παρατήρηση, είναι ανεξάρτητη από την τιμή του σε όλες τις υπόλοιπες παρατηρήσεις. Για όλες τις παρατηρήσεις που σχετίζονται με μια συγκεκριμένη διαστρωματική μονάδα i , το a_i θα έχει την ίδια τιμή, αντικατοπτρίζοντας τα αμετάβλητα απαρατήρητα χαρακτηριστικά της διαστρωματικής μονάδας. Αυτό απεικονίζεται στον Πίνακα 2.5.1 για την περίπτωση που έχουμε τέσσερις διαστρωματικές μονάδες (άτομα) και τρεις χρονικές περιόδους.

Διαστρωματική Μονάδα	Χρονική Περίοδος	u
1	1	$a_1 + v_{11}$
1	2	$a_1 + v_{12}$
1	3	$a_1 + v_{13}$
2	1	$a_2 + v_{21}$
2	2	$a_2 + v_{22}$
2	3	$a_2 + v_{23}$
3	1	$a_3 + v_{31}$
3	2	$a_3 + v_{32}$
3	3	$a_3 + v_{33}$
4	1	$a_4 + v_{41}$
4	2	$a_4 + v_{42}$
4	3	$a_4 + v_{43}$

Πίνακας 2.5.1: Παράδειγμα τιμών όρων διαταραχής σε μοντέλο τυχαίων επιδράσεων

Δεδομένου ότι οι όροι διαταραχής για τη διαστρωματική μονάδα i έχουν ένα κοινό στοιχείο a_i , οι όροι συσχετίζονται. Για τη διαστρωματική μονάδα i στην περίοδο t , ο όρος

διαταραχής είναι $(a_i + v_{it})$. Για την ίδια τη διαστρωματική μονάδα σε οποιαδήποτε άλλη περίοδο t' είναι $(a_i + v_{it'})$. Η διακύμανση του πληθυσμού μεταξύ τους είναι

$$\sigma_{\varepsilon_{it}, \varepsilon_{it'}} = \sigma_{(a_i + v_{it}), (a_i + v_{it'})} = \sigma_{a_i, a_i} + \sigma_{a_i, v_{it'}} + \sigma_{v_{it}, a_i} + \sigma_{v_{it}, v_{it'}} = \sigma_a^2, \quad (2.15)$$

Για παρατηρήσεις που αναφέρονται σε διαφορετικές διαστρωματικές μονάδες (χώρες), δεν υπάρχει πρόβλημα επειδή τότε οι συνιστώσες a θα είναι διαφορετικές και ανεξάρτητες.

Θα μπορούσαμε για την αντιμετώπιση αυτού του προβλήματος της παραβίασης της συνθήκης του μοντέλου παλινδρόμησης (OLS) να χρησιμοποιήσουμε γενικευμένα μοντέλα ελαχίστων τετραγώνων (GLS). Για να έχει ο εκτιμητής καλές ιδιότητες θα πρέπει να έχει μεγάλο αριθμό διαστρωματικών μονάδων n και σχετικά μικρό αριθμό χρονικών περιόδων T . Σε διαφορετική περίπτωση, το μοντέλο ελαχίστων τετραγώνων (OLS) παραμένει αμερόληπτο και συνεπές, αλλά είναι αναποτελεσματικό και τα τυπικά σφάλματα OLS υπολογίζονται λανθασμένα.

Στην περίπτωση αυτή, η λύση είναι να μετασχηματίσει το μοντέλο έτσι ώστε ο μετασχηματισμένος όρος διαταραχής να ικανοποιεί την συνθήκη του μοντέλου παλινδρόμησης (GLS). Ωστόσο, για το μετασχηματισμό των γενικευμένων ελαχίστων τετραγώνων που εξαλείφει την αυτοσυσχέτιση απαιτείται ανωτέρου επιπέδου άλγεβρα πινάκων. Όμως αυτός καθ'αυτός ο μετασχηματισμός είναι απλός [26].

Ορίζουμε τον συντελεστή

$$\lambda = 1 - \sqrt{\frac{\sigma_\varepsilon^2}{(\sigma_\varepsilon^2 + T\sigma_a^2)}}, \quad (2.16)$$

που παίρνει τιμές μεταξύ του μηδενός και της μονάδας. Τότε το μετασχηματισμένο μοντέλο αποδεικνύεται ότι είναι

$$Y_{it} - \lambda \bar{Y}_i = a(1 - \lambda) + \beta'(X_{it} - \lambda \bar{X}_i) + (u_{it} - \lambda \bar{u}_i), \quad (2.17)$$

όπου η παύλα πάνω από μια μεταβλητή συμβολίζει πάλι τους χρονικούς μέσους όρους (βλ. (2.4)).

Το μοντέλο των σταθερών επιδράσεων αφαιρεί τους χρονικούς μέσους από την αντίστοιχη μεταβλητή ενώ το μοντέλο τυχαίων επιδράσεων αφαιρεί ένα μέρος αυτών των χρονικών μέσων από το μέρος που εξαρτάται από το σ_ε^2 , το σ_a^2 και το πλήθος των χρονικών περιόδων T . Ο εκτιμητής με τη μέθοδο των γενικευμένων τετραγώνων είναι απλώς ο εκτιμητής των ομαδοποιημένων ελαχίστων τετραγώνων (2.17).

Ο μετασχηματισμός (2.17) λαμβάνει υπόψη την περίπτωση που οι ερμηνευτικές

μεταβλητές είναι διαχρονικά σταθερές, και αυτό είναι ένα πλεονέκτημα του μοντέλου τυχαίων επιδράσεων σε σχέση με το μοντέλο σταθερών επιδράσεων. Αυτό συμβαίνει γιατί το μοντέλο τυχαίων επιδράσεων υποθέτει ότι οι παρατηρήσιμες επιδράσεις a_i δε συσχετίζονται με όλες τις ερμηνευτικές μεταβλητές, είτε αυτές είναι είτε δεν είναι σταθερές. Έτσι για παράδειγμα σε μια εκτίμηση των αποδοχών υπαλλήλων μπορούμε να συμπεριλάβουμε ως ερμηνευτική μεταβλητή το επίπεδο μόρφωσης των υπαλλήλων ακόμα και αν αυτό δε μεταβάλλεται διαχρονικά. Όμως υποθέτουμε ότι το επίπεδο μόρφωσης δε σχετίζεται με τις παρατηρήσιμες επιδράσεις a_i , οι οποίες μπορεί να αφορούν πχ. την ικανότητα των υπαλλήλων και το οικογενειακό τους υπόβαθρο. Σε πολλές εφαρμογές η λογική για τη χρήση δεδομένων πάνελ είναι να επιτρέψει στις παρατηρήσιμες επιδράσεις να σχετίζονται με τις ερμηνευτικές μεταβλητές.

Η παράμετρος λ δεν είναι γνωστή στην πράξη αλλά μπορεί να εκτιμηθεί χρησιμοποιώντας για παράδειγμα τη μέθοδο ομαδοποιημένων ελαχίστων τετραγώνων. Πολλά λογισμικά (όπως το πακέτο `plm` της R) υποστηρίζουν την εκτίμηση του μοντέλου τυχαίων επιδράσεων και υπολογίζουν αυτόματα κάποια εκδοχή του $\hat{\lambda}$. Ο εκτιμητής των εφικτών γενικευμένων ελαχίστων τετραγώνων για το λ ονομάζεται **εκτιμητής τυχαίων επιδράσεων**. Αν ισχύουν οι προϋποθέσεις του μοντέλου τυχαίων επιδράσεων, ο εκτιμητής είναι συνεπής (όχι αμερόληπτος) και κατανέμεται ασυμπτωματικά κανονικά καθώς το n μεγαλώνει όταν το T είναι σταθερό. Οι ιδιότητες του εκτιμητή τυχαίων επιδράσεων για μικρό n και μεγάλο T είναι σε μεγάλο βαθμό άγνωστες αν και έχει χρησιμοποιηθεί και σε αυτές τις περιπτώσεις.

Η εξίσωση (2.17) μας επιτρέπει να συσχετίσουμε το μοντέλο τυχαίων επιδράσεων με το μοντέλο ομαδοποιημένων ελαχίστων τετραγώνων και το μοντέλο σταθερών επιδράσεων. Το μοντέλο ομαδοποιημένων ελαχίστων τετραγώνων λαμβάνεται όταν $\lambda=0$ και το μοντέλο σταθερών επιδράσεων “εντός” λαμβάνεται όταν $\lambda=1$. Στην πράξη η εκτίμηση $\hat{\lambda}$ δεν είναι ποτέ μηδέν ή ένα. Όταν όμως το $\hat{\lambda}$ είναι κοντά στο μηδέν, τότε οι εκτιμητές τυχαίων επιδράσεων θα είναι κοντά στους εκτιμητές των ομαδοποιημένων ελαχίστων τετραγώνων. Αυτό συμβαίνει όταν η παρατηρήσιμη επίδραση a_i είναι σχετικά ασήμαντη (αφού έχει μικρή διακύμανση σε σχέση με την σ_ε^2). Αλλά είναι πιο σύνηθες η σ_a^2 να είναι πιο μεγάλη σε σχέση με τη σ_ε^2 , οπότε το $\hat{\lambda}$ θα είναι πιο κοντά στη μονάδα. Καθώς το T μεγαλώνει, το $\hat{\lambda}$ τείνει προς τη μονάδα και αυτό έχει ως συνέπεια οι εκτιμήσεις του μοντέλου τυχαίων επιδράσεων να είναι παρόμοιες με τις εκτιμήσεις του μοντέλου σταθερών επιδράσεων.

2.6 ΈΛΕΓΧΟΙ ΚΑΤΑΛΛΗΛΟΤΗΤΑΣ ΤΩΝ ΜΟΝΤΕΛΩΝ

2.6.1 Μοντέλο Σταθερών ή Τυχαίων Επιδράσεων;

Αρκετές φορές, οι ερευνητές αντιμετωπίζουν το δίλημμα σχετικά με το ποιο από τα δυο μοντέλα που παρουσιάστηκαν παραπάνω (Μοντέλο Σταθερών Επιδράσεων ή Μοντέλο Τυχαίων Επιδράσεων) είναι το καλύτερο. Πότε πρέπει να χρησιμοποιείτε η εκτίμηση σταθερών επιδράσεων και όχι η εκτίμηση τυχαίων αποτελεσμάτων ή το αντίστροφο; Καταρχήν, το μοντέλο τυχαίων επιδράσεων είναι πιο ελκυστικό επειδή τα παρατηρούμενα χαρακτηριστικά που παραμένουν σταθερά για κάθε διαστρωματική μονάδα (οντότητα), διατηρούνται στο μοντέλο παλινδρόμησης. Στο μοντέλο σταθερών επιδράσεων, θα πρέπει να εξαλειφθούν.

Επίσης, με το μοντέλο τυχαίων επιδράσεων δε χάνουμε βαθμούς ελευθερίας, όπως συμβαίνει με τα μοντέλα σταθερών επιδράσεων.

Ωστόσο, εάν παραβιάζεται οποιαδήποτε από τις προϋποθέσεις για τη χρήση του μοντέλου τυχαίων επιδράσεων, θα πρέπει να χρησιμοποιήσουμε το μοντέλο σταθερών επιδράσεων. Μια προϋπόθεση είναι ότι οι παρατηρήσεις επιλέγονται τυχαία από έναν δεδομένο πληθυσμό. Αυτή η προϋπόθεση είναι λογική στην περίπτωση για παράδειγμα που συλλέγουμε δεδομένα για το μισθό, τη μόρφωση και την προϋπηρεσία των εργαζομένων μιας χώρας στον δημόσιο τομέα. Αντίθετα, η προϋπόθεση αυτή δεν θα ήταν λογική εάν οι μονάδες παρατήρησης στο σύνολο των δεδομένων πάνελ ήταν εργαζόμενοι της πρωτοβάθμιας εκπαίδευσης. Αυτοί οι εργαζόμενοι σίγουρα δεν μπορούν να θεωρηθούν ότι αντιπροσωπεύουν ένα τυχαίο δείγμα των κυρίαρχων κλάδων του δημοσίου τομέα.

Η άλλη προϋπόθεση είναι ότι οι παρατηρήσιμες επιδράσεις a_i κατανέμονται ανεξάρτητα από τις ερμηνευτικές μεταβλητές X_{it} . Πως μπορούμε να διαπιστώσουμε ότι αυτή η προϋπόθεση ισχύει; Η τυπική διαδικασία είναι μια ακόμη εφαρμογή του τεστ Durbin – Wu – Hausman (DWH) που ως γνωστόν χρησιμοποιείται για να μας βοηθήσει να επιλέξουμε μεταξύ της εκτίμησης ελαχίστων τετραγώνων (OLS) και της εκτίμησης των τεχνητών μεταβλητών IV σε περιπτώσεις όπου υπάρχει υποψία σφάλματος μέτρησης ή ενδογενών ερμηνευτικών μεταβλητών (οι ενδογενείς μεταβλητές έχουν τιμές οι οποίες καθορίζονται από άλλες μεταβλητές του συστήματος). Στην ανάλυση δεδομένων πάνελ, το τεστ DWH, συνήθως αναφέρεται ως έλεγχος Hausman και μας βοηθάει να επιλέξουμε ανάμεσα στο μοντέλο σταθερών και τυχαίων επιδράσεων. Ο έλεγχος Hausman βασίζεται στην παρακάτω σχέση:

$$\varepsilon_{it} = a_i + v_{it}, \quad (2.19)$$

όπου a_i είναι η ατομική επίδραση (Individual effect) και εκφράζει τη συμπεριφορά της καθεμίας διαστρωματικής μονάδας και v_{it} μεταβάλλεται διαχρονικά και ανά διαστρωματική μονάδα. Ο Hausman λοιπόν πρότεινε τον παρακάτω έλεγχο:

H_0 : ο όρος a_i δεν συσχετίζεται με τις ερμηνευτικές μεταβλητές X_{it}

H_1 : ο όρος a_i συσχετίζεται με τις ερμηνευτικές μεταβλητές X_{it}

Τα περισσότερα οικονομετρικά λογισμικά εκτελούν αυτό το τεστ. Σε περίπτωση που ο στατιστικός έλεγχος οδηγήσει στη μη-απόρριψη της H_0 ($p\text{-value} > \alpha$), τότε επιλέγεται η προσέγγιση των τυχαίων επιδράσεων. Αν όμως οδηγήσει στην απόρριψη της H_0 ($p\text{-value} < \alpha$), τότε το υπόδειγμα θα πρέπει να εκτιμηθεί με τη μέθοδο των σταθερών επιδράσεων [5].

2.6.2 Μοντέλο Τυχαίων Επιδράσεων ή Ενιαίο Μοντέλο;

Ας υποθέσουμε ότι ο έλεγχος Hausman δείχνει ότι μπορούμε να χρησιμοποιήσουμε το μοντέλο τυχαίων επιδράσεων. Θα πρέπει λοιπόν να εξετάσουμε εάν δεν υπάρχουν καθόλου απαραίτητες επιδράσεις. Μπορεί απλά το μοντέλο να έχει καθοριστεί τόσο καλά ώστε ο όρος διαταραχής (2.19) αποτελείται μόνο από την καθαρά τυχαία συνιστώσα v_{it} και δεν υπάρχουν όροι ατομικής επίδρασης a_i . Στην περίπτωση αυτή, θα πρέπει να χρησιμοποιήσουμε το ενιαίο μοντέλο το οποίο παρέχει δύο πλεονεκτήματα. Το μοντέλο αυτό είναι αποτελεσματικό, επειδή δεν θεωρούμε ανύπαρκτη αυτοσυσχέτιση εντός των ομάδων και μπορούμε να επωφεληθούμε από τις ιδιότητες του μοντέλου OLS, αντί να πρέπει να βασιστούμε στις ασυμπτωτικές ιδιότητες του μοντέλου τυχαίων επιδράσεων. Για τον έλεγχο της ύπαρξης ή όχι τυχαίων επιδράσεων οι Breusch και Pagan (1980) έχουν προτείνει ένα στατιστικό τεστ βασιζόμενοι στους πολλαπλασιαστές Lagrange (LM)

Ο έλεγχος Breusch-Pagan βασίζεται στις παρακάτω υποθέσεις:

$$H_0: \sigma_v^2 = 0$$

$$H_1: \sigma_v^2 \neq 0$$

Τα περισσότερα οικονομετρικά λογισμικά εκτελούν αυτό τον έλεγχο. Αν το στατιστικό LM είναι στατιστικά σημαντικό ($p\text{-value} < \alpha$), τότε θα πρέπει να χρησιμοποιήσουμε το μοντέλο τυχαίων επιδράσεων αντί για το ενιαίο μοντέλο (pooled model) της γραμμικής παλινδρόμησης [5].

ΚΕΦΑΛΑΙΟ 3

3 ΑΝΑΛΥΣΗΣ ΔΕΔΟΜΕΝΩΝ PANEL ΜΕ ΤΗΝ R

3.1 ΕΙΣΑΓΩΓΗ

Σε αυτό το κεφάλαιο θα δώσουμε τις κυριότερες εντολές που διαθέτει η γλώσσα προγραμματισμού R και ειδικότερα το πακέτο `plm` για τη διαχείριση και την ανάλυση δεδομένων πάνελ. Το `plm` είναι ένα ολοκληρωμένο πακέτο που διαθέτει η R για τις τυπικές λειτουργίες που απαιτούνται για την οικονομετρική ανάλυση των δεδομένων πάνελ. Το `plm` δίνει τη δυνατότητα για την εκτίμηση μιας μεγάλης ποικιλίας μοντέλων δεδομένων πάνελ με απλό τρόπο και για την εξαγωγή (ισχυρών) συμπερασμάτων. Παρακάτω αρχικά γίνεται μια πρώτη ανάλυση των βασικών εντολών που απαιτούνται για την ανάλυση δεδομένων πάνελ. Την ανάλυση κάθε εντολής ακολουθούν παραδείγματα για την πιο απλή κατανόηση του τρόπου που λειτουργεί.

3.2 ΦΟΡΤΩΣΗ ΤΩΝ ΚΑΤΑΛΛΗΛΩΝ ΠΑΚΕΤΩΝ

Πρωταρχικός στόχος της ανάλυσης δεδομένων πάνελ με χρήση της R είναι η φόρτωση των κατάλληλων πακέτων που απαιτούνται για την ορθή ανάλυση αυτών. Ορισμένα από τα σημαντικότερα πακέτα φαίνονται στο παρακάτω πίνακα: [29][31]

```
library(plm)    # Panel data analysis library
library(gplots) # various programing tools for plotting data
library(tidyverse) # Modern data science library
library(broom)  # for `glance()` and `tidy()`
library(car)    # Companion to applied regression |
```

3.3 ΟΡΓΑΝΩΝΟΝΤΑΣ ΤΑ ΔΕΔΟΜΕΝΑ ΩΣ ΠΑΝΕΛ

Όπως έχουμε αναφέρει, τα δεδομένα πάνελ αποτελούν ένα σύνολο δεδομένων στο οποίο η συμπεριφορά των εξεταζόμενων οντοτήτων (N στρωμάτων) παρατηρείται στο πέρασμα του χρόνου (T χρονικών περιόδων). Τα δεδομένα πάνελ δηλαδή μπορούν να θεωρηθούν ως ένα πλαίσιο δεδομένων. Τα δεδομένα πάνελ μπορούν να οργανωθούν κυρίως ως ένα **ευρύ** πλαίσιο δεδομένων όπου η διαστρωματική διάσταση (N) είναι πολύ μεγαλύτερη από τη διαμήκη διάσταση (T), ή ως ένα **μακρύ** πλαίσιο δεδομένων όπου ισχύει το

αντίθετο. Οι περισσότερες μέθοδοι δεδομένων πάνελ απαιτούν τη μακριά μορφή, αλλά πολλές πηγές δεδομένων παρέχουν έναν πίνακα ευρείας μορφής για κάθε μεταβλητή. Η συναρμολόγηση των δεδομένων από διαφορετικές πηγές σε ένα πλαίσιο δεδομένων μεγάλου μήκους είναι πολύ σημαντικό θέμα για την περεταίρω ανάλυσή τους. Το πακέτο `plm` της R δίνει τη δυνατότητα οργάνωσης των δεδομένων υπό τη μορφή δεδομένων πάνελ ώστε να ξεκινήσει η ανάλυση τους.[31]

Με την επόμενη ακολουθία κώδικα αρχικά γίνεται ανάκτηση των δεδομένων μας² και στη συνέχεια δημιουργείται μια δομή πίνακα πάνελ για το σύνολο των δεδομένων μας χρησιμοποιώντας την εντολή `pdata.frame` του `plm` πακέτου:

```
dataPanel101<- read.csv("dataPanel101.csv")
dataPanel101<-data.frame(dataPanel101,index=c("country","year"))
```

Με την εντολή `pdata.frame` ουσιαστικά ορίζεται ένα πλαίσιο δεδομένων(`data.frame`), αλλά χρησιμοποιώντας το επιπλέον όρισμα “`index`” παρέχεται ένα επιπλέον χαρακτηριστικό ευρετηρίου που περιγράφει τις διαστρωματικές και χρονικές διαστάσεις.

Με τις παρακάτω εντολές επιλέγουμε ένα υποσύνολο από τα δεδομένα μας και κατόπιν τα παρουσιάζουμε σε ένα πίνακα:

```
library(xtable)
library(knitr)
smp1<- dataPanel101[country%in%c("A","B"), c(1:6, 7:8)]
tbl<- xtable(smp1)
kable(tbl, digits=4, align="c", caption="Δείγμα Δεδομένων")
```

ζ

²Τα δεδομένα που χρησιμοποιούμε ανακτήθηκαν από την ιστοσελίδα <https://raw.githubusercontent.com/ds777/sample-datasets/master/dataPanel101.csv>

	country	year	y	y_bin	x1	x2	x3	opinion
A-1990	A	1990	1342787840	1	0.27790365	-1.1079559	0.28255358	Str agree
A-1991	A	1991	-1899660544	0	0.32068470	-0.9487200	0.49253848	Disag
A-1992	A	1992	-11234363	0	0.36346573	-0.7894840	0.70252335	Disag
A-1993	A	1993	2645775360	1	0.24614404	-0.8855330	-0.09439092	Disag
A-1994	A	1994	3008334848	1	0.42462304	-0.7297683	0.94613063	Disag
A-1995	A	1995	3229574144	1	0.47721413	-0.7232460	1.02968040	Str agree
A-1996	A	1996	2756754176	1	0.49980500	-0.7815716	1.09228810	Disag
A-1997	A	1997	2771810560	1	0.05162839	-0.7048455	1.41590080	Str agree
A-1998	A	1998	3397338880	1	0.36641079	-0.6983712	1.54872270	Disag
A-1999	A	1999	39770336	1	0.39584252	-0.6431540	1.79419800	Str disag
B-1990	B	1990	-5934699520	0	-0.08184998	1.4251202	0.02342812	Agree
B-1991	B	1991	-711623744	0	0.10616001	1.6496018	0.26036251	Str agree
B-1992	B	1992	-1933116160	0	0.35378519	1.5937191	-0.23439877	Agree
B-1993	B	1993	3072741632	1	0.72677696	1.6917576	0.25622433	Str disag
B-1994	B	1994	3768078848	1	0.71939486	1.7414261	0.41174951	Disag
B-1995	B	1995	2837581312	1	0.67154658	1.7083139	0.53584301	Str disag
B-1996	B	1996	577199360	1	0.81985730	1.5324961	-0.49964902	Str agree
B-1997	B	1997	1786851584	1	0.88016719	1.5021962	-0.57626772	Disag
B-1998	B	1998	-149072048	0	0.70451611	1.4236463	-0.44841924	Agree
B-1999	B	1999	-1174480128	0	0.23696731	1.4545859	-0.04936399	Str disag

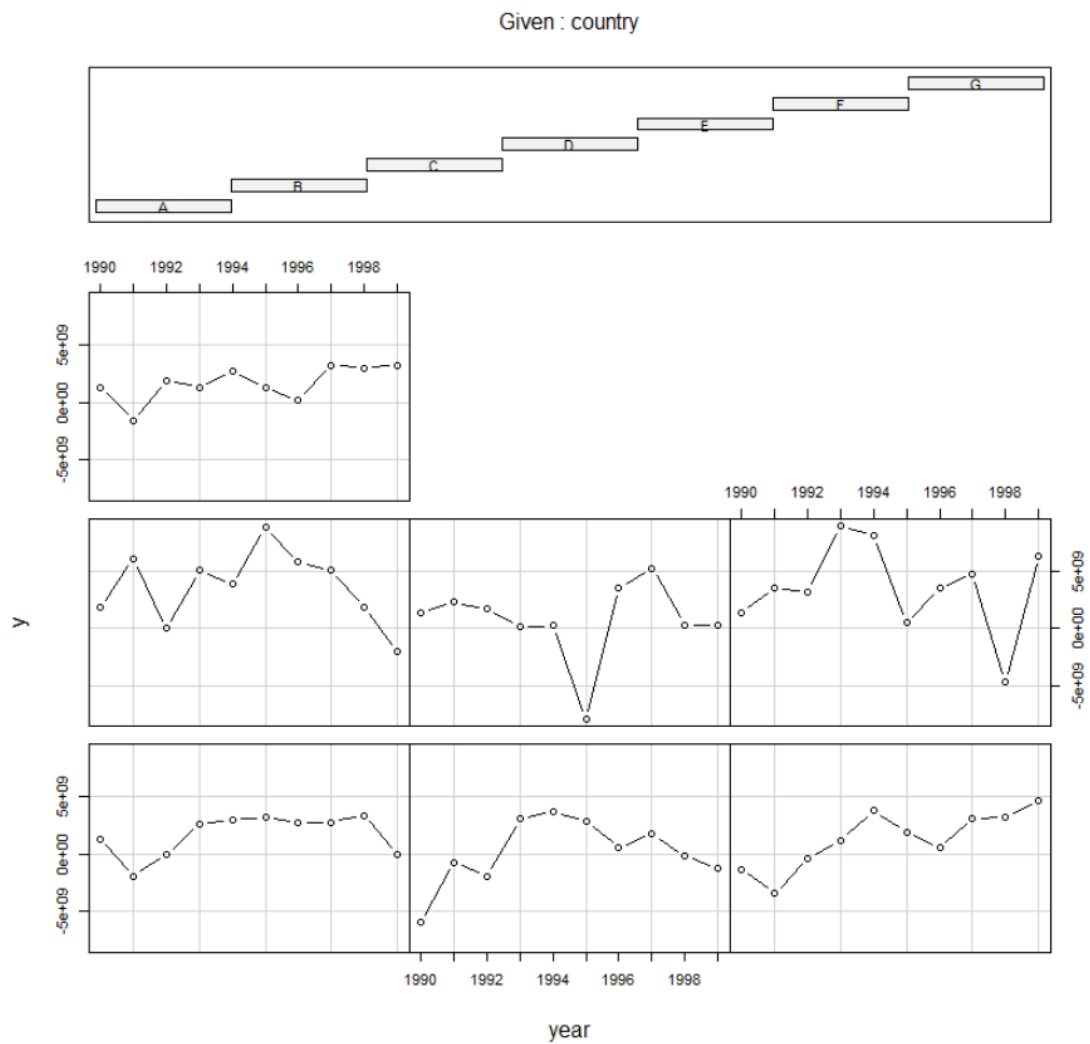
Πίνακας 3.3.1: Δείγμα των δεδομένων

3.4 ΔΙΕΡΕΥΝΗΤΙΚΗ ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ

Σε επόμενο βήμα θα ξεκινήσουμε την ουσιαστικότερη ανάλυση των δεδομένων μας. Στη περίπτωση αυτή η εντολή που θα πληκτρολογήσουμε για την διερευνητική ανάλυση των δεδομένων είναι η εξής:

```
coplot(y ~ year|country, type="b", data=dataPanel101)
```

Το αποτέλεσμα αυτής φαίνεται στη συνέχεια:



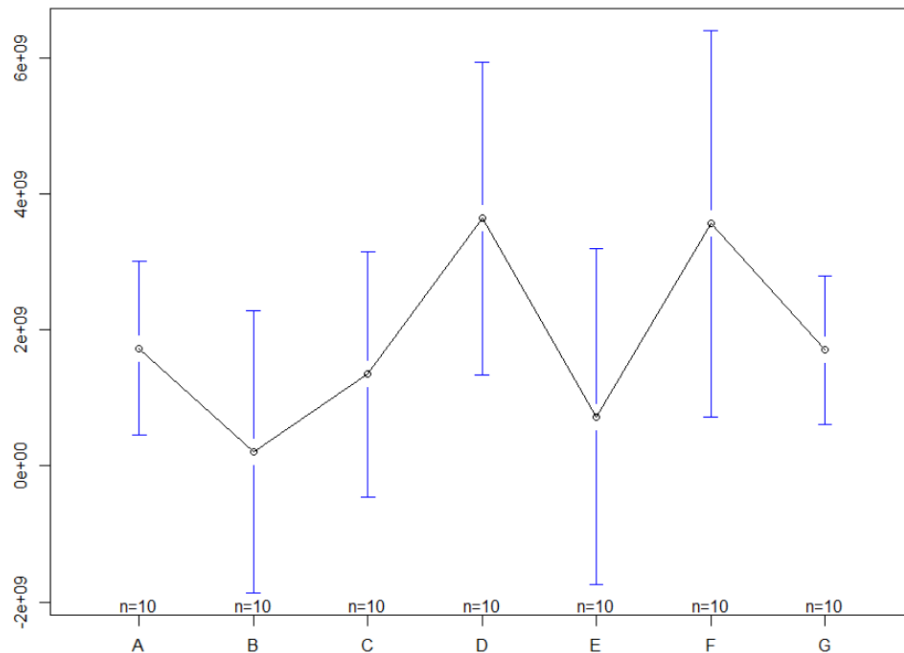
Στο παραπάνω σχήμα, οι μπάρες στην κορυφή υποδεικνύουν τη θέση των χωρών από αριστερά προς τα δεξιά ξεκινώντας από την κάτω σειρά και μας δείχνουν τη διαχρονική μεταβολή των χωρών για την εξαρτημένη μεταβολή.[29]

3.4.1 Ετερογένεια μεταξύ των χωρών

Σύμφωνα με την παρακάτω εντολή κάνουμε έλεγχο για ύπαρξη ετερογένειας μεταξύ διαστρωματικών στοιχείων.

```
plotmeans (y ~ country, data = dataPanel101)
```

Το αποτέλεσμα της εκτέλεσης αυτής εντολής φαίνεται παρακάτω:



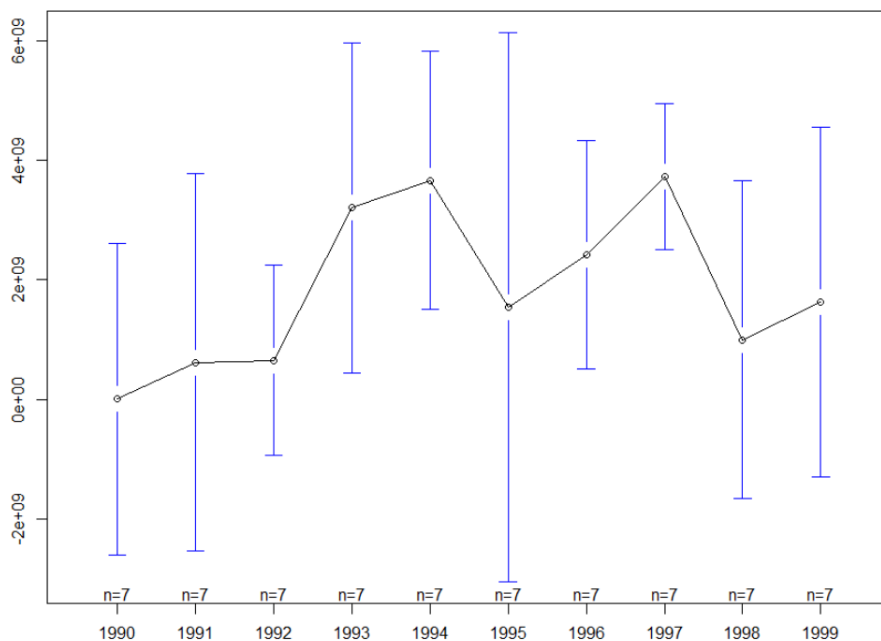
Στο παραπάνω διάγραμμα, με μαύρο κύκλο, συμβολίζεται ο μέσος όρος της εξαρτημένης μεταβλητής για κάθε διαστρωματικό στοιχείο (π.χ. χώρα). Οι μπλε κατακόρυφες γραμμές αποτελούν ένα 95% διάστημα εμπιστοσύνης για κάθε μέσο όρο. Η γραμμή που ενώνει αυτούς τους μέσους μας δίνει πληροφορίες για το αν υπάρχει ή όχι διακύμανση μεταξύ των διαστρωματικών στοιχείων (π.χ. χωρών). Μέσα από αυτό το διάγραμμα μπορούμε να κατανοήσουμε αν το μοντέλο σταθερών επιδράσεων (Fixed-effects model) είναι ή όχι κατάλληλο για την ανάλυσή μας.[29]

3.4.2 Ετερογένεια με την πάροδο των ετών

Ένα ακόμη στοιχείο που εξετάζουμε είναι και η ετερογένεια με την πάροδο των ετών. Με παρόμοιο τρόπο δίνεται η εντολή για την περίπτωση αυτή και η οποία φαίνεται παρακάτω:

```
plotmeans(y ~ year, data = dataPanel101)
```

Το αποτέλεσμα αυτής της ανάλυσης είναι το εξής:



Και πάλι, οι μαύρες γραμμές που ενώνουν τους διαχρονικούς μέσους της εξαρτημένης μεταβλητής και δείχνουν αν υπάρχει απόκλιση ή όχι της εξαρτημένης μεταξύ των ετών.

3.5 ΕΚΤΙΜΗΣΗ ΜΟΝΤΕΛΩΝ ΔΕΔΟΜΕΝΩΝ ΠΑΝΕΛ ΜΕ ΧΡΗΣΗ ΤΗΣ R

Η εκτίμηση μοντέλων δεδομένων πάνελ με χρήση της R γίνεται με τη συνάρτηση **plm** με κατάλληλη επιλογή του ορίσματος **model**.

3.5.1 Το Ενιαίο Μοντέλο

Το ενιαίο μοντέλο μπορεί να εκτιμηθεί στην R χρησιμοποιώντας τη συνάρτηση **plm()**, επιλέγοντας το όρισμα **model = "pooling"** όπως απεικονίζει η παρακάτω ακολουθία κώδικα. Για τον ορισμό του μοντέλου χρησιμοποιούμε την έκφραση **y~x1** που σημαίνει ότι τα δεδομένα για τη μεταβλητή απόκρισης είναι στο διάνυσμα **y** ενώ για την επεξηγηματική μεταβλητή είναι στο διάνυσμα **x1**. [29][30]

```
pooled <- plm(y~x1, data=dataPanel101, model="pooling")
summary(pooled)
```

Με την εντολή **summary()** η R μας παρουσιάζει διάφορα αποτελέσματα στατιστικής συμπερασματολογίας του μοντέλου. Το αποτέλεσμα αυτής της εντολής απεικονίζεται στο παρακάτω πίνακα:

```

Pooling Model

Call:
plm(formula = y ~ x1, data = dataPanel101, model = "pooling")

Balanced Panel: n = 7, T = 10, N = 70

Residuals:
    Min.    1st Qu.    Median        Mean    3rd Qu.     Max.
-9.55e+09 -1.58e+09  1.55e+08  0.00e+00  1.42e+09  7.18e+09

Coefficients:
              Estimate Std. Error t-value Pr(>|t|)
(Intercept) 1524319101  621072623   2.4543  0.01668 *
x1           494988866   778861258   0.6355  0.52722
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares:    6.2729e+20
Residual Sum of Squares: 6.2359e+20
R-Squared:    0.0059046
Adj. R-Squared: -0.0087145
F-statistic: 0.403897 on 1 and 68 DF, p-value: 0.52722

```

Αρχικά, η R μας παρουσιάζει διάφορους περιγραφικούς δείκτες (ελάχιστη τιμή, 1^ο τεταρτημόριο, διάμεσος, 3^ο τεταρτημόριο, μέγιστη τιμή) για τα υπόλοιπα. Από τους εν λόγω δείκτες μπορούμε να συμπεράνουμε ότι η υπόθεση συμμετρικής κατανομής για τα σφάλματα με μηδενική μέση τιμή φαίνεται ρεαλιστική. Ακολουθούν αποτελέσματα για τους συντελεστές του μοντέλου. Στην πρώτη στήλη από την ενότητα Coefficient δίνονται οι σημειακές εκτιμήσεις (Estimate) των συντελεστών. Οι συντελεστές αυτοί αποδίδονται σε καθέναν παλινδρομητή και μας δείχνουν πόσο μεταβάλλεται η εξαρτημένη μεταβλητή όταν ο εκάστοτε παλινδρομητής αυξάνεται κατά μία μονάδα. Κατά συνέπεια, έχουμε τη δυνατότητα να προβλέψουμε την πορεία της εξαρτημένης μεταβλητής βασισμένη στην πορεία μιας ανεξάρτητης, τουλάχιστον μεταξύ των μελών του δείγματός μας. Στη δεύτερη στήλη (Std. Error) δίνονται τα τυπικά σφάλματα παλινδρόμησης για καθένα εκτιμητή, στη τρίτη στήλη (t-value) δίνονται οι τιμές του στατιστικού ελέγχου που χρησιμοποιείται για τον έλεγχο της μηδενικής υπόθεσης (ότι οι συντελεστές ισούνται με μηδέν) και την τελευταία στήλη οι αντίστοιχες p-τιμές του ελέγχου. Ο αριθμός της p-value εκφράζει την

αξιοπιστία που υπάρχει στη σχέση μεταξύ δύο μεταβλητών, δηλαδή την «αντιπροσωπευτικότητα» ενός εκτιμώμενου αποτελέσματος συσχέτισης σε ολόκληρο τον πληθυσμό. Πιο συγκεκριμένα, η p.value, ως δείκτης αξιοπιστίας του αποτελέσματος, αντιπροσωπεύει την πιθανότητα του λάθους που περιλαμβάνεται στην αποδοχή του παρατηρηθέντος αυτού αποτελέσματος. Για παράδειγμα, μια p.value της τιμής 0.05 μας δείχνει ότι υπάρχει πιθανότητα 5% η σχέση μεταξύ των μεταβλητών στο δείγμα μας να είναι «ψευδής».

Στη συνέχεια η R δίνει το ολικό άθροισμα τετραγώνων και το άθροισμα των τετραγώνων των υπολοίπων καθώς και τον συντελεστή προσδιορισμού και τον προσαρμοσμένο συντελεστή προσδιορισμού.

Τέλος η R παρουσιάζει αποτελέσματα και του F-test: την παρατηρηθείσα τιμή του στατιστικού ελέγχου F, τους βαθμούς ελευθερίας της κατανομής Snedecor που ακολουθεί καθώς και την p-τιμή για τον εν λόγω έλεγχο.

3.5.2 Το Μοντέλο Σταθερών Επιδράσεων

Το μοντέλο σταθερών επιδράσεων λαμβάνει υπόψη διαφορές ανάμεσα στα στρώματα, οι οποίες εκφράζονται με διαφορετικές σταθερές (intercepts) της γραμμής παλινδρόμησης για διαφορετικά στρώματα. Ωστόσο, για να εκτιμήσουμε ένα μοντέλο σταθερών επιδράσεων στην R, δεν χρειάζεται να δημιουργήσουμε εικονικές μεταβλητές, αλλά να χρησιμοποιήσουμε την επιλογή **model="within"** στη συνάρτηση `plm()`. Το ακόλουθο τμήμα κώδικα χρησιμοποιεί ολόκληρο το δείγμα.[29][30]

```
fixed <- plm(y ~ x1, data=dataPanel101, model="within")
summary(fixed)
```

Το αποτέλεσμα αυτού του κώδικα απεικονίζεται στο παρακάτω πίνακα:

```

Oneway (individual) effect within Model

Call:
plm(formula = y ~ x1, data = dataPanel101, model = "within")

Balanced Panel: n = 7, T = 10, N = 70

Residuals:
      Min.      1st Qu.      Median      Mean      3rd Qu.      Max.
-8.63e+09 -9.70e+08  5.40e+08  0.00e+00  1.39e+09  5.61e+09

Coefficients:
      Estimate Std. Error t-value Pr(>|t|)
x1  2475617742 1106675596   2.237  0.02889 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares:    5.2364e+20
Residual Sum of Squares: 4.8454e+20
R-Squared:              0.074684
Adj. R-Squared:        -0.029788
F-statistic: 5.00411 on 1 and 62 DF, p-value: 0.028892

```

Αρχικά, όπως παρατηρούμε η R δίνει πληροφορίες για τα δεδομένα (Ισορροπημένα ή μη, πλήθος στρωμάτων, χρονικών περιόδων κλπ). Στη συνέχεια, όπως πριν, η R μας παρουσιάζει περιγραφικούς δείκτες για τα υπόλοιπα, αποτελέσματα για τους συντελεστές του μοντέλου με την ίδια ερμηνεία, το ολικό άθροισμα τετραγώνων και το άθροισμα των τετραγώνων των υπολοίπων καθώς και τον συντελεστή προσδιορισμού και τον προσαρμοσμένο συντελεστή προσδιορισμού. Τέλος, όπως και πριν, η R παρουσιάζει αποτελέσματα του F-test.

3.5.3 Το Μοντέλο Τυχαίων Επιδράσεων

Το μοντέλο **τυχαίων επιδράσεων** προσαρμόζει το μοντέλο σταθερών επιδράσεων αναγνωρίζοντας ότι, δεδομένου ότι τα άτομα στα δεδομένα πάνελ επιλέγονται τυχαία, τα χαρακτηριστικά τους, μετρούνται από την τομή a_i με τον άξονα των τεταγμένων όπου θα πρέπει επίσης να είναι τυχαία. Ένα σημαντικό χαρακτηριστικό του μοντέλου τυχαίων επιδράσεων είναι ότι η χρονική συσχέτιση στα σφάλματα δεν μειώνεται με την πάροδο του χρόνου.

Για να εκτιμήσουμε το μοντέλο τυχαίων επιδράσεων χρησιμοποιούμε την ίδια

συνάρτηση `plm()` με τη διαφορά ότι θέτουμε ως όρισμα `model="random"` και επιλέγοντας `random.method` ως μία από τις τέσσερις δυνατότητες: “swar” (προεπιλογή), “amemiya”, “walhus” ή “nerlove”. Οι επόμενες γραμμές κώδικα υπολογίζουν το μοντέλο τυχαίων επιδράσεων.[29][30][31]

```
random <- plm(y ~ x1, data=dataPanel101, random.method="swar",
model="random")
summary(random)
```

Τα αποτελέσματα που προκύπτουν από την εκτέλεση των παραπάνω εντολών δίνονται παρακάτω:

```
Oneway (individual) effect Random Effect Model
(Swamy-Arora's transformation)

Call:
plm(formula = y ~ x1, data = dataPanel101, model = "random",
random.method = "swar")

Balanced Panel: n = 7, T = 10, N = 70

Effects:
              var   std.dev share
idiosyncratic 7.815e+18 2.796e+09 0.873
individual    1.133e+18 1.065e+09 0.127
theta: 0.3611

Residuals:
      Min.   1st Qu.   Median     Mean   3rd Qu.    Max.
-8.94e+09 -1.51e+09  2.82e+08  0.00e+00  1.56e+09  6.63e+09

Coefficients:
              Estimate Std. Error z-value Pr(>|z|)
(Intercept) 1037014329  790626206  1.3116   0.1896
x1           1247001710  902145599  1.3823   0.1669

Total Sum of Squares:    5.6595e+20
Residual Sum of Squares: 5.5048e+20
R-Squared:      0.02733
Adj. R-Squared: 0.013026
Chisq: 1.91065 on 1 DF, p-value: 0.16689
```

Σε αυτή την περίπτωση, η R δίνει τα ίδια αποτελέσματα για το μοντέλο τυχαίων επιδράσεων με αυτά για το μοντέλο σταθερών επιδράσεων με την επιπλέον ενότητα Effect. Η ενότητα αυτή μας δίνει αποτελέσματα (διακύμανση, τυπική απόκλιση) για τις απαραίτητες επιδράσεις (individual) και τις επιδράσεις ιδιοσυγκρατικού σφάλματος (idiosyncratic) καθώς και το ποσοστό (share) του συνολικού σφάλματος που αντιστοιχεί σε καθεμία από αυτές. Επιπλέον, παρέχεται και ο συντελεστής theta που αποτελεί εκτίμηση της παραμέτρου λ στον εκτιμητή τυχαίων επιδράσεων (2.17). Υπενθυμίζεται ότι τιμές του theta κοντά στο μηδέν, σημαίνει ότι το μοντέλο τυχαίων επιδράσεων είναι κοντά στο ενιαίο μοντέλο, ενώ τιμές του theta κοντά στη μονάδα, σημαίνει ότι το μοντέλο τυχαίων επιδράσεων είναι κοντά στο μοντέλο σταθερών επιδράσεων.

3.6 ΈΛΕΓΧΟΣ ΚΑΤΑΛΛΗΛΟΤΗΤΑΣ ΜΕ ΧΡΗΣΗ ΤΗΣ R

3.6.1 Μοντέλο Σταθερών ή Τυχαίων Επιδράσεων;

Για να αποφασίσουμε μεταξύ σταθερών ή τυχαίων επιδράσεων μπορούμε να εκτελέσουμε τον έλεγχο Hausman όπου η μηδενική υπόθεση είναι ότι το προτιμώμενο μοντέλο είναι αυτό των τυχαίων επιδράσεων έναντι της εναλλακτικής ότι είναι το μοντέλο των σταθερών επιδράσεων. Πιο συγκεκριμένα, ο έλεγχος αυτός ελέγχει εάν τα οι απαραίτητες επιδράσεις α_i συσχετίζονται με τις ερμηνευτικές μεταβλητές, η μηδενική υπόθεση είναι ότι δεν συσχετίζονται. Εάν η τιμή p είναι σημαντική ($p\text{-value} < 0,05$), τότε χρησιμοποιούμε το μοντέλο σταθερών επιδράσεων, διαφορετικά χρησιμοποιούμε το μοντέλο τυχαίων επιδράσεων. Παρακάτω δίνουμε τον κατάλληλο κώδικα καθώς επίσης και το αποτέλεσμα της εκτέλεσής του.[29]

```
phtest(fixed, random)
      Hausman Test

data:  y ~ x1
chisq = 3.674, df = 1, p-value = 0.05527
alternative hypothesis: one model is inconsistent
```

Ελέγχοντας το αποτέλεσμα παρατηρούμε ότι το $p\text{-value} = 0.05527 > 0.05$ άρα το μοντέλο που θα πρέπει να χρησιμοποιήσουμε είναι αυτό των τυχαίων επιδράσεων.

3.6.2 Μοντέλο Τυχαίων Επιδράσεων ή Ενιαίο Μοντέλο;

Ο έλεγχος *Breusch - Pagan* (LM) μας βοηθά να αποφασίσουμε μεταξύ μιας παλινδρόμησης τυχαίων επιδράσεων και μιας απλής παλινδρόμησης ενιαίου μοντέλου.

Η μηδενική υπόθεση στο τεστ LM είναι ότι οι διακυμάνσεις μεταξύ των στρωμάτων είναι μηδέν. Ένα παράδειγμα εντολής για το συγκεκριμένο τεστ φαίνεται στη συνέχεια καθώς και το αποτέλεσμα της εντολής αυτής.[29]

```
# Breusch-Pagan Lagrange Multiplier for random effects. Null is no panel effect (i.e. OLS
better).

plmtest(pooling, type=c("bp"))

      Lagrange Multiplier Test - (Breusch-Pagan) for balanced
panels

data:  y ~ x1
chisq = 2.6692, df = 1, p-value = 0.1023
alternativehypothesis: significanteffects
```

Στο συγκεκριμένο παράδειγμα αποτύχαμε να απορρίψουμε τη μηδενική υπόθεση και να καταλήξουμε στο συμπέρασμα ότι το μοντέλο τυχαίων επιδράσεων δεν είναι κατάλληλο. Αυτό σημαίνει ότι δεν υπάρχει ένδειξη σημαντικών διαφορών μεταξύ των διαστρωματικών οντοτήτων, επομένως μπορούμε να εκτελέσουμε μια απλή παλινδρόμηση ενιαίου μοντέλου.

3.7 ΈΛΕΓΧΟΣ ΕΤΕΡΟΣΚΕΔΑΣΤΙΚΟΤΗΤΑΣ

Ένας τελευταίος έλεγχος που πραγματοποιούμε πριν τα τελικά συμπεράσματα είναι και αυτός για ετεροσκεδαστικότητα στο μοντέλο τυχαίων επιδράσεων.

Ένας σημαντικός έλεγχος που πρέπει να εκτελέσουμε κατά τη διάρκεια μιας ανάλυσης δεδομένων πάνελ είναι και αυτός για τον έλεγχο ετεροσκεδαστικότητας. Αν υποθέσουμε ότι η μηδενική υπόθεση για το τεστ Breusch-Pagan είναι η ομοσκεδαστικότητα, τότε η εναλλακτική θα είναι η ύπαρξη ετεροσκεδαστικότητας. Εκτελώντας την κατάλληλη εντολή που βλέπουμε παρακάτω θα αντλήσουμε και το αντίστοιχο αποτέλεσμα.

```
library(lmtest)
bptest(y~x1 +factor(country), data = dataPanel101,studentize=F)
      Breusch-Pagan test

data:  y ~ x1 + factor(country)
BP = 14.606, df = 7, p-value = 0.04139
```

Επειδή η τιμή $p < 0,05$, εντοπίζουμε ετεροσκεδαστικότητα.

Αφού ανιχνεύθηκε ετεροσκεδαστικότητα, θα πρέπει να χρησιμοποιήσουμε έναν εύρωστο (robust) πίνακα συνδιακύμανσης (Sandwich estimator) για τους συντελεστές της παλινδρόμησης ώστε να την αντιμετωπίσουμε.

Με χρήση του ορίσματος **vcovHC** στην συνάρτηση **coeftest** της R προκύπτουν συνεπείς συντελεστές του μοντέλου Τυχαίων Επιδράσεων με την παρουσία ετεροσκεδαστικότητας.

Η **vcovHC** είναι μια συνάρτηση του πακέτου **plm** της R για την εκτίμηση ενός εύρωστου (robust) πίνακα συνδιακύμανσης των παραμέτρων του μοντέλου τυχαίων επιδράσεων (αλλά και του μοντέλου σταθερών επιδράσεων). Η συνάρτηση **-vcovHC-** έχει επιλογή μεταξύ τριών εκτιμητών συνδιακύμανσης που είναι συνεπείς σε παρουσία ετεροσκεδαστικότητας:

- **white1** "- κατάλληλη για γενική ετεροσκεδαστικότητα αλλά χωρίς σειριακή συσχέτιση. Συνιστάται για το μοντέλο τυχαίων επιδράσεων.
- **white2** "- είναι "white1" εκτιμητής με περιορισμό κοινής διακύμανση εντός των ομάδων. Συνιστάται για το μοντέλο τυχαίων επιδράσεων.
- **arellano** "- κατάλληλη τόσο σε ετεροσκεδαστικότητα όσο και σε σειριακή συσχέτιση. Συνιστάται για το μοντέλο σταθερών επιδράσεων.

Όσον αφορά στους τύπους στάθμισης, η συνάρτηση περιλαμβάνει τις ακόλουθες επιλογές:

- **HC0** - συνεπής σε ετεροσκεδαστικότητα: Είναι η προεπιλεγμένη.
- **HC1, HC2, HC3** - Συνιστάται για μικρά δείγματα. Το **HC3** δίνει μικρότερο βάρος στις παρατηρήσεις με μεγάλη επιρροή
- **HC4**- Συνιστάται για μικρά δείγματα με παρατηρήσεις μεγάλης επιρροής **HAC-** είναι συνεπής σε ετεροσκεδαστικότητα και αυτοσυσχέτιση

Η παρακάτω εντολή δίνει τους αρχικούς συντελεστές του μοντέλου τυχαίων επιδράσεων:[29][30]

```
# Αρχικοί Συντελεστές
coeftest(random)
t test of coefficients:
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1037014329	790626206	1.3116	0.1941
x1	1247001710	902145599	1.3823	0.1714

Εισάγοντας το όρισμα `vcovHC`, προκύπτουν οι παρακάτω συνεπείς εκτιμήσεις των συντελεστών του μοντέλου χωρίς παρουσία ετεροσκεδαστικότητας

```
coeftest(random, vcovHC)t test of coefficients:
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1037014329	907983024	1.1421	0.2574
x1	1247001710	828970258	1.5043	0.1371

```
# Heteroskedasticity consistent coefficients, type 3
coeftest(random, vcovHC(random, type = "HC3"))
t test of coefficients:
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1037014329	943438278	1.0992	0.2756
x1	1247001710	867137595	1.4381	0.1550

ΚΕΦΑΛΑΙΟ 4

4 ΜΕΛΕΤΗ ΠΕΡΙΠΤΩΣΗΣ ΑΝΑΛΥΣΗΣ ΔΕΔΟΜΕΝΩΝ ΠΑΝΕΛ: Παράγοντες Επίδρασης Εξαγωγών Υψηλής Τεχνολογίας

4.1 ΕΙΣΑΓΩΓΗ

Στο κεφάλαιο αυτό θα προσπαθήσουμε, με τη βοήθεια της R, να εφαρμόσουμε τις μεθόδους ανάλυσης δεδομένων πάνελ, που αναπτύξαμε στο προηγούμενο κεφάλαιο, για τον καθορισμό των παραγόντων που επηρεάζουν τις Εξαγωγές Υψηλής Τεχνολογίας (HTX) σε 15 χώρες της Ευρωπαϊκής Ένωσης για το χρονικό διάστημα 2007-2018.

Οι Εξαγωγές Υψηλής Τεχνολογίας (High Technology Exports HTX) θεωρούνται σημαντικός παράγοντας για τη βιώσιμη οικονομική ανάπτυξη μιας χώρας. Για το λόγο αυτό, η ανάγκη διερεύνησης των πιθανών προσδιοριστικών παραγόντων που επηρεάζουν τις εξαγωγές υψηλής τεχνολογίας σε διάφορες χώρες του πλανήτη έχει οξυνθεί, κατά τα τελευταία χρόνια.

Η ανάλυση που ακολουθεί επιχειρεί να διαπιστώσει κατά πόσο α) οι Δαπάνες για Έρευνα και Ανάπτυξη, β) οι Εξαγωγές αγαθών Τεχνολογίας Πληροφορικής και Τηλεπικοινωνιών, γ) η Συναλλαγματική Ισοτιμία, δ) το γυναικείο Εργατικό Δυναμικό με εξειδικευμένη εκπαίδευση, στ) η Χρηματοοικονομική Ανάπτυξη και ε) το Επίπεδο Ανθρώπινης Ανάπτυξης επηρεάζουν τις Εξαγωγές Υψηλής Τεχνολογίας σε 15 χώρες της Ευρωπαϊκής Ένωσης αναφορικά με τα έτη 2007-2018.

Αρχικά διεξάγουμε μια περιγραφική ανάλυση των μεταβλητών προκειμένου να διαπιστωθεί η σχέση με την εξαρτημένη μεταβλητή, η ύπαρξη ή όχι ετεροσκεδαστικότητας στα δεδομένα και η πιθανή αντιμετώπισή της με κατάλληλο μετασχηματισμό. Στη συνέχεια διεξάγουμε μια διερευνητική ανάλυση των δεδομένων μας για να διαπιστώσουμε την ύπαρξη ή μη σταθερών επιδράσεων. Ακολουθεί η εκτίμηση παλινδρόμησης με χρήση του πακέτου plm, το οποίο πραγματοποιεί ανάλυση σε panel data, για την εκτίμηση των παραμέτρων της γραμμικής παλινδρόμησης στο Ενιαίο Μοντέλο (Pooled Model), στο Μοντέλο Σταθερών Επιδράσεων (Fixed Effects Model) καθώς επίσης και στο Μοντέλο Τυχαίων Επιδράσεων (Random Effects Model). Κατόπιν διεξάγουμε κατάλληλους ελέγχους για να διαπιστώσουμε ποιο από τα μοντέλα παλινδρόμησης είναι κατάλληλο για να ερμηνεύσει τα δεδομένα μας. Τέλος ελέγχουμε την ύπαρξη ετεροσκεδαστικότητας και προβαίνουμε στην αντιμετώπισή της.

4.2 ΤΑ ΔΕΔΟΜΕΝΑ

Τα δεδομένα που χρησιμοποιήθηκαν στην παρούσα εργασία (εκτός από τον Δείκτη Ανθρώπινης Ανάπτυξης) αντλήθηκαν από τη βάση δεδομένων της Παγκόσμιας Τράπεζας (World Bank) διότι είναι ένας αξιόπιστος ερευνητικός φορέας με δειγματικά δεδομένα αντιπροσωπευτικά ως προς τον πληθυσμό. Επιπλέον για τη μέτρηση του επιπέδου ανθρώπινης ανάπτυξης (HDI) χρησιμοποιούμε τον Δείκτη Ανθρώπινης Ανάπτυξης (HDI) ο οποίος αντλήθηκε από τον επίσημο ιστότοπο του United Nations Development Program (UNDP) (2011). Τελικά, για την ανάλυση των εξαγωγών υψηλής τεχνολογίας (HTX) επιλέχθηκαν οι παρακάτω ερμηνευτικές μεταβλητές:

- RD: Δαπάνες για Έρευνα και Ανάπτυξη ((ως ποσοστό του ΑΕΠ) (R&D expenditure (% of GDP))
- ICT: Εξαγωγές αγαθών Τεχνολογίας Πληροφορικής και Τηλεπικοινωνιών ως ποσοστό των συνολικών εξαγωγών (ICT goods exports (% of total goods exports)
- REER: Δείκτη Πραγματικής Συναλλαγματικής Ισοτιμίας (Real effective exchange rate index (2010 = 100))
- FLabor: Εργατικό δυναμικό με προχωρημένη εκπαίδευση, γυναίκα (ως ποσοστό του πληθυσμού γυναικών σε ηλικία εργασίας με προχωρημένη εκπαίδευση) (Labor force with advanced education, female (% of female working-age population with advanced education)
- FIAI: Δείκτης Πρόσβασης Χρηματοπιστωτικών Ιδρυμάτων (Financial Institutions Access Index (FIAI))
- HDI: Δείκτη Ανθρώπινης Ανάπτυξης (Human Development Index (HDI))

Κριτήριο επιλογής των παραπάνω μεταβλητών, ως παράγοντες οι οποίοι επηρεάζουν τις εξαγωγές μιας χώρας, ήταν η χρήση τους στην υπάρχουσα βιβλιογραφία και λόγω της πληρότητας των δεδομένων τους για όλες τις υπό μελέτη χώρες.

Σε ό,τι αφορά στην εξαρτημένη μεταβλητή HTX, χρησιμοποιήσαμε τις Εξαγωγές Υψηλής Τεχνολογίας (ως ποσοστό των εξαγωγών κατασκευαστικού τομέα) (High-technology exports (% of manufactured exports)).

Στο σημείο αυτό αξίζει να σημειωθεί ότι οι χώρες που διερευνούμε είναι η Αυστρία (Austria), το Βέλγιο (Belgium), η Δανία (Denmark), η Φινλανδία (Finland), η Γαλλία (France), η Γερμανία (Germany), η Ελλάδα (Greece), η Ιρλανδία (Ireland), η Ιταλία (Italy),

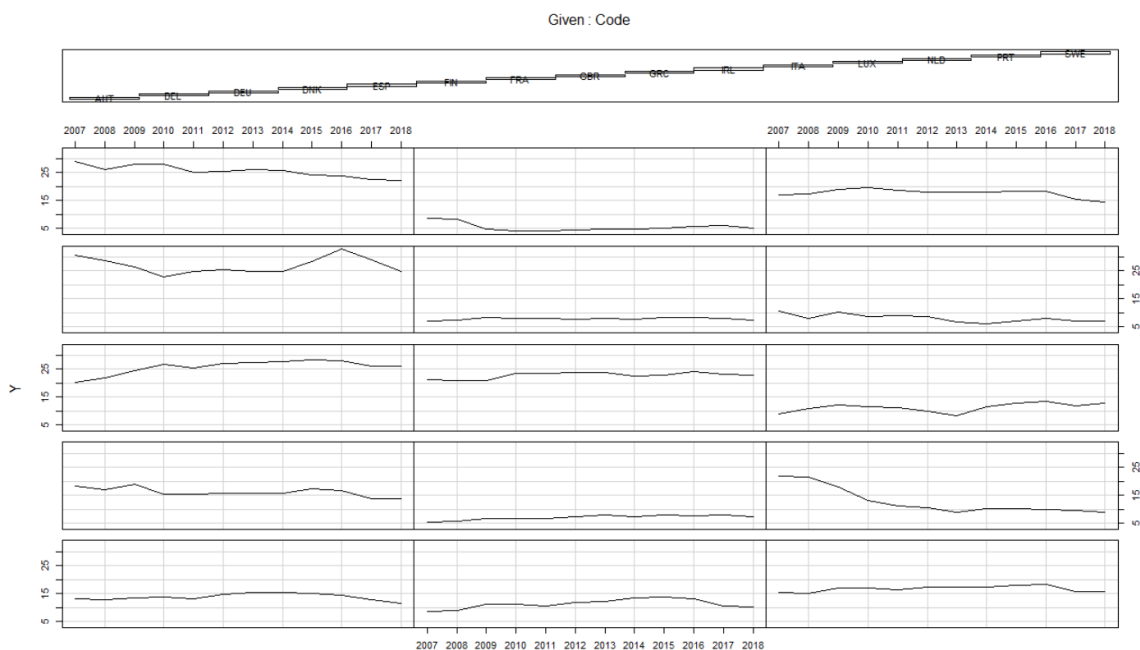
το Λουξεμβούργο (Luxembourg), η Ολλανδία (Netherlands), η Πορτογαλία (Portugal), η Ισπανία (Spain), η Σουηδία (Sweden) και το Ηνωμένο Βασίλειο (United Kingdom).[14][18]

4.3 ΠΕΡΙΓΡΑΦΙΚΗ ΑΝΑΛΥΣΗ ΕΞΑΓΩΓΩΝ ΥΨΗΛΗΣ ΤΕΧΝΟΛΟΓΙΑΣ

Αρχικά δίνουμε την πληροφορία στην R ότι πρόκειται να ασχοληθούμε με δεδομένα πάνελ. Με χρήση της εντολής `pdata.frame` η R δημιουργεί ένα πλαίσιο δεδομένων με έναν χαρακτηριστικό δείκτη που περιγράφει τις διαστρωματικές μονάδες (*id*) και τις διαστάσεις του χρόνου (*t*)

```
> pdata <- pdata.frame(mydata, index=c("id","t"))
```

Στη συνέχεια, παρουσιάζουμε τη διαχρονική μεταβολή της εξαρτημένης μεταβλητής για κάθε χώρα της ΕΕ-15 που εξετάζουμε.

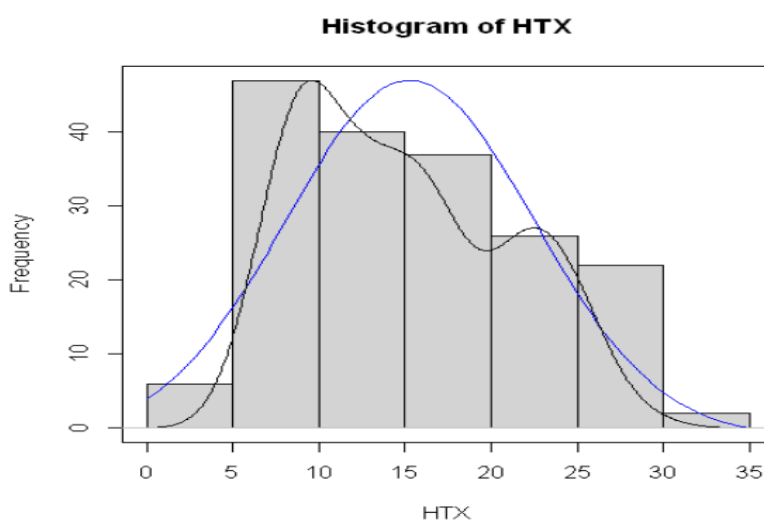


Διάγραμμα 4.1: Διαχρονική μεταβολή των εξαγωγών υψηλής τεχνολογίας (HTX) για τις ΕΕ-15 χώρες

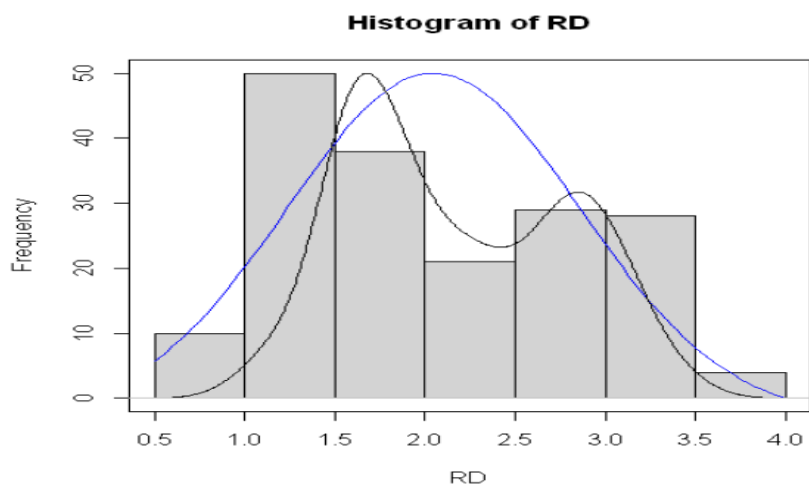
Στο Διάγραμμα 4.1 διαφαίνεται η ανάλυση της διαχρονικής μεταβολής των εξαγωγών υψηλής τεχνολογίας (HTX) για κάθε χώρα. Οι μπάρες (Given:Code) στην κορυφή του Διαγράμματος 4.1 υποδεικνύουν το αντίστοιχο γράφημα της κάθε χώρας (Code=κωδικός χώρας) από αριστερά προς τα δεξιά ξεκινώντας από την κάτω σειρά. Από τα παραπάνω εξάγουμε το συμπέρασμα ότι οι εξαγωγές υψηλής τεχνολογίας (HTX) για κάθε χώρα είναι σχετικά σταθερές με μικρές αυξομειώσεις, με την πιο σημαντική να παρατηρείτε στη

Φιλανδία όπου μέχρι το 2013 ακολουθεί μια πτωτική τάση και στη συνέχεια φαίνεται να βελτιώνεται και να σταθεροποιείται.

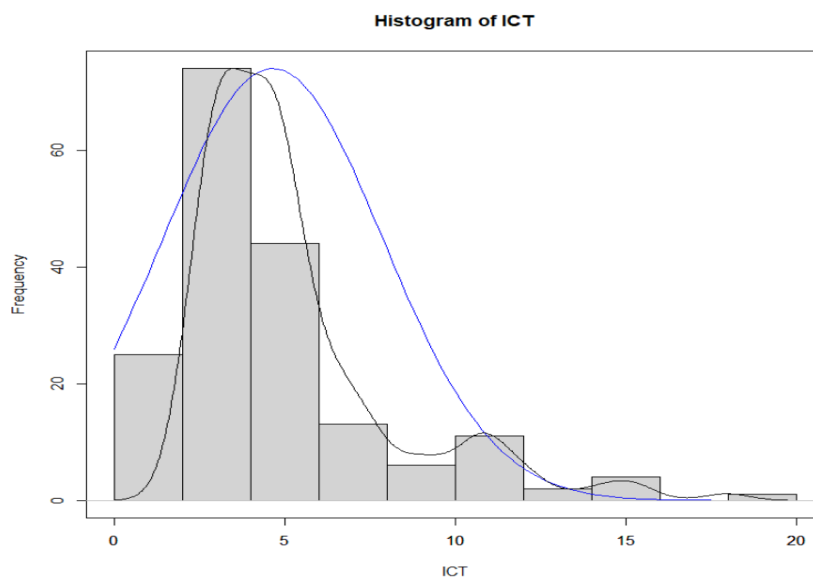
Κατ' αρχάς, εξετάζουμε αν υπάρχει ετεροσκεδαστικότητα (μη σταθερή διακύμανση) στις μεταβλητές του μοντέλου. Αυτό το πρόβλημα γίνεται ορατό μέσω των ιστογραμμάτων τους. Τα ιστογράμματα αποτελούν γραφική απεικόνιση στατιστικών συχνοτήτων περιοχών τιμών ενός μεγέθους. Η επιφάνεια κάθε ορθογωνίου είναι το μέτρο της συχνότητας εμφάνισης της συγκεκριμένης περιοχής τιμών, ενώ το ύψος του ισούται με το λόγο της συχνότητας προς το εύρος των τιμών που αντιπροσωπεύει το ορθογώνιο. Αποτελεί την πιο συνηθισμένη επιλογή γραφικής παράστασης συνεχών μεταβλητών, όπως αυτές των χρονοσειρών. Βασικό χαρακτηριστικό των χρονολογικών σειρών είναι ότι συνήθως η κατανομή τους αποκλίνει από την κανονικότητα. Αυτό σημαίνει ότι μπορεί να υπάρχει θετική ή αρνητική ασυμμετρία ή λεπτοκύρτωση, δηλαδή παρουσία ακραίων τιμών. Ελέγχουμε λοιπόν τις χρονολογικές σειρές για κανονικότητα. Η κανονική κατανομή έχει σταθερό μέσο μ και διακύμανση σ^2 . Επομένως, όσο πιο συμμετρική είναι η καμπύλη μιας μεταβλητής τόσο πιο κοντά στην κανονική κατανομή βρίσκεται και τόσο μικρότερο είναι το πρόβλημα της ετεροσκεδαστικότητας που αντιμετωπίζει. Ένας πρακτικός τρόπος να ελέγξουμε αν η κατανομή κάθε μεταβλητής είναι κανονική ή ασυμμετρική είναι με τα ιστογράμματα. Για καλύτερα ορατά αποτελέσματα μπορούμε να σχεδιάσουμε τα ιστογράμματα σε συνδυασμό με την κανονική κατανομή, για να δούμε σε τι βαθμό διαφέρουν από αυτή οι κατανομές των μεταβλητών μας. Ακολουθούν τα ιστογράμματα:



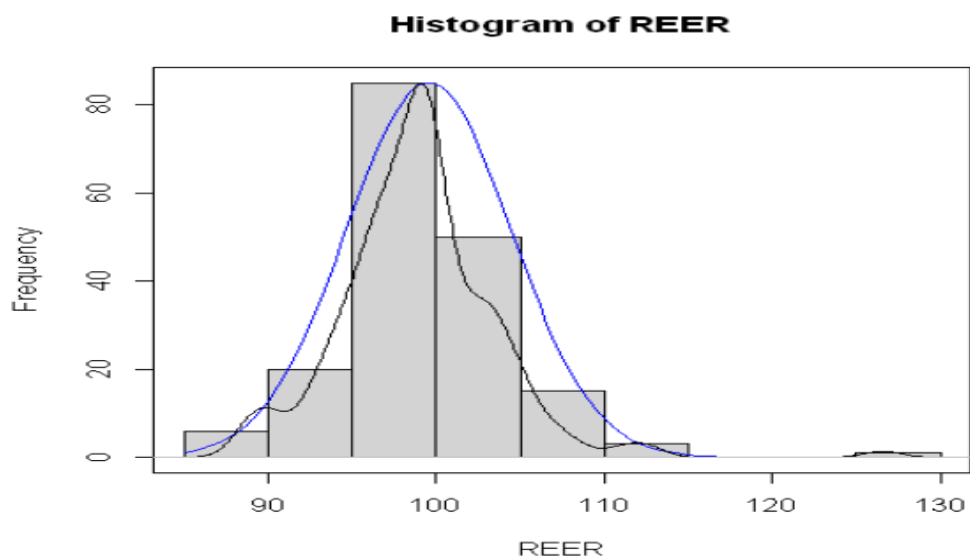
Διάγραμμα 4.2: Ιστόγραμμα εξαγωγών υψηλής τεχνολογίας



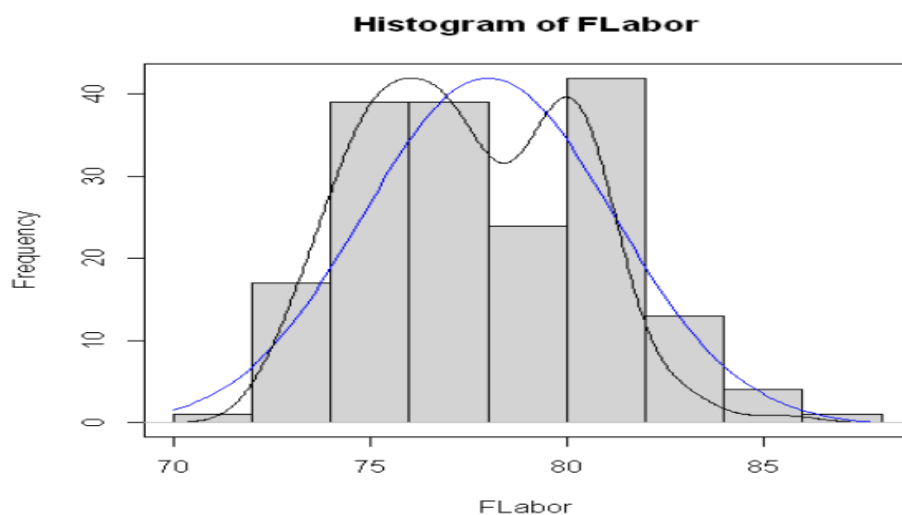
Διάγραμμα 4.3: Ιστόγραμμα Δαπανών Έρευνας και Ανάπτυξης



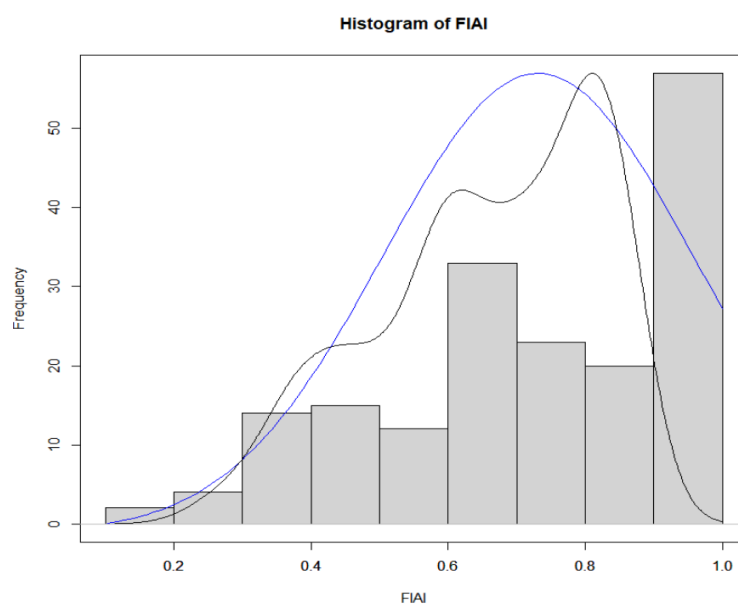
Διάγραμμα 4.4: Ιστόγραμμα εξαγωγής αγαθών Τεχνολογίας Πληροφορικής και Τηλεπικοινωνιών



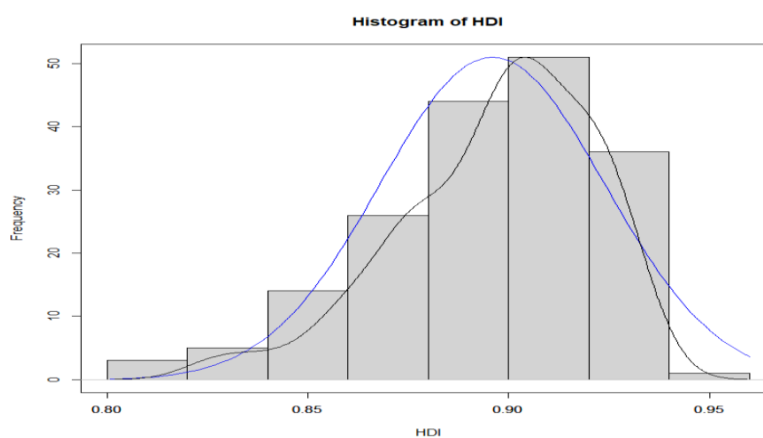
Διάγραμμα 4.5: Ιστόγραμμα Δείκτη Πραγματικής Συναλλαγματικής Ισοτιμίας



Διάγραμμα 4.6: Ιστόγραμμα Εργατικού δυναμικού με προχωρημένη εκπαίδευση, γυναίκα



Διάγραμμα 4.7: Ιστόγραμμα Δείκτη Πρόσβασης Χρηματοπιστωτικών Ιδρυμάτων



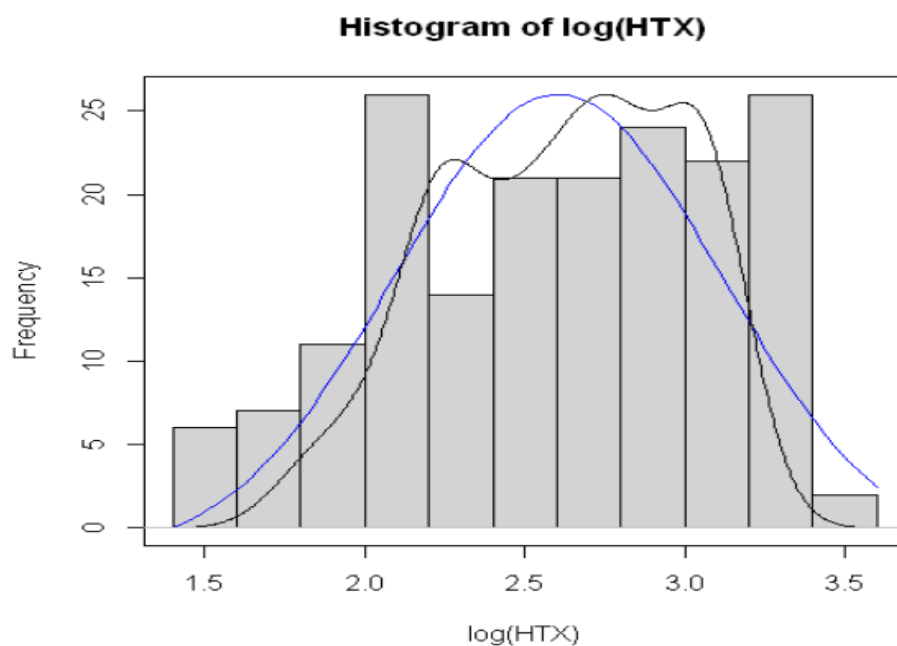
Διάγραμμα 4.8: Ιστόγραμμα Δείκτη Ανθρώπινης Ανάπτυξης

Από τα παραπάνω παρατηρούμε ότι οι κατανομές των μεταβλητών «δείκτης ανθρώπινης

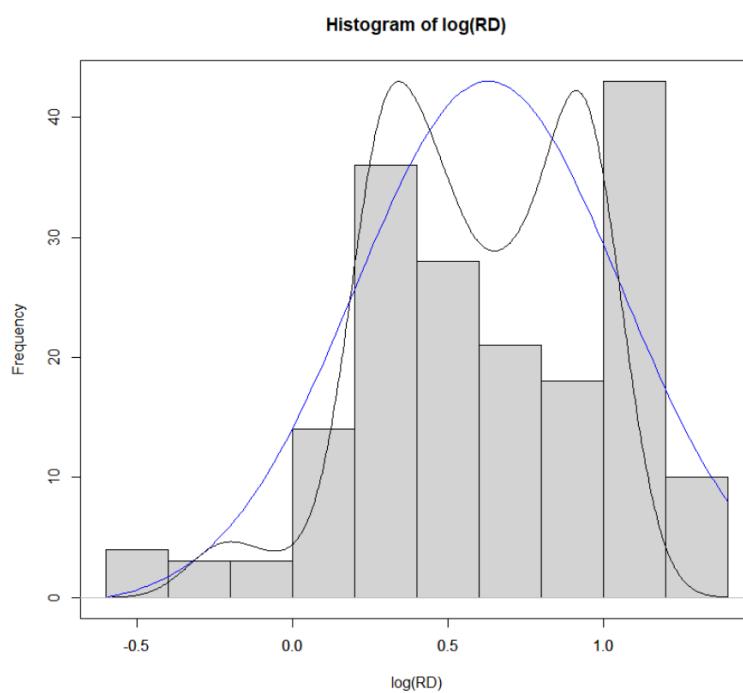
ανάπτυξης» (HDI), «Δείκτη Πραγματικής Συναλλαγματικής Ισοτιμίας» (REER) καθώς και «οι εξαγωγές υψηλής τεχνολογίας» (HTX) πλησιάζουν την κανονική κατανομή (συμμετρική καμπύλη). Θετική ασυμμετρία παρουσιάζουν οι υπόλοιπες μεταβλητές, εκτός από τη μεταβλητή «Δείκτη Πρόσβασης Χρηματοπιστωτικών Ιδρυμάτων» (FIAI) που παρουσιάζει αρνητική ασυμμετρία.

Από τα διαγράμματα συχνοτήτων διαφαίνεται να υπάρχει κάποιος βαθμός ετεροσκεδαστικότητας για όλες ανεξαιρέτως τις μεταβλητές. Κατ' επέκταση, κρίνεται σκόπιμη η χρησιμοποίηση φυσικών λογαρίθμων για την διόρθωση αυτού του προβλήματος. Υπάρχουν πολλοί λόγοι για να χρησιμοποιήσει κανείς φυσικούς λογαρίθμους σε μια ανάλυση παλινδρόμησης. Για παράδειγμα, οι συντελεστές κλίσης μεταβλητών που εμφανίζονται σε λογαριθμική μορφή δεν επηρεάζονται από τις μονάδες μέτρησης. Όταν μια μεταβλητή είναι ένα θετικό χρηματικό ποσό εκφρασμένο σε δολάρια, συνήθως χρησιμοποιείται ο λογάριθμος. Οι αυστηρά θετικές μεταβλητές συχνά ακολουθούν υπό δέσμευση κατανομές που παρουσιάζουν ετεροσκεδαστικότητα και ασυμμετρία, όπως παρατηρήσαμε παραπάνω στις μεταβλητές μας. Η χρήση λογαρίθμων μετριάξει ή και εξαλείφει αυτά τα δυο προβλήματα. Τέλος, η χρήση λογαρίθμων περιορίζει το εύρος της μεταβλητής, κάτι που κάνει τις εκτιμήσεις λιγότερο ευαίσθητες στις ακραίες παρατηρήσεις για την εξαρτημένη ή τις ανεξάρτητες μεταβλητές.[14]

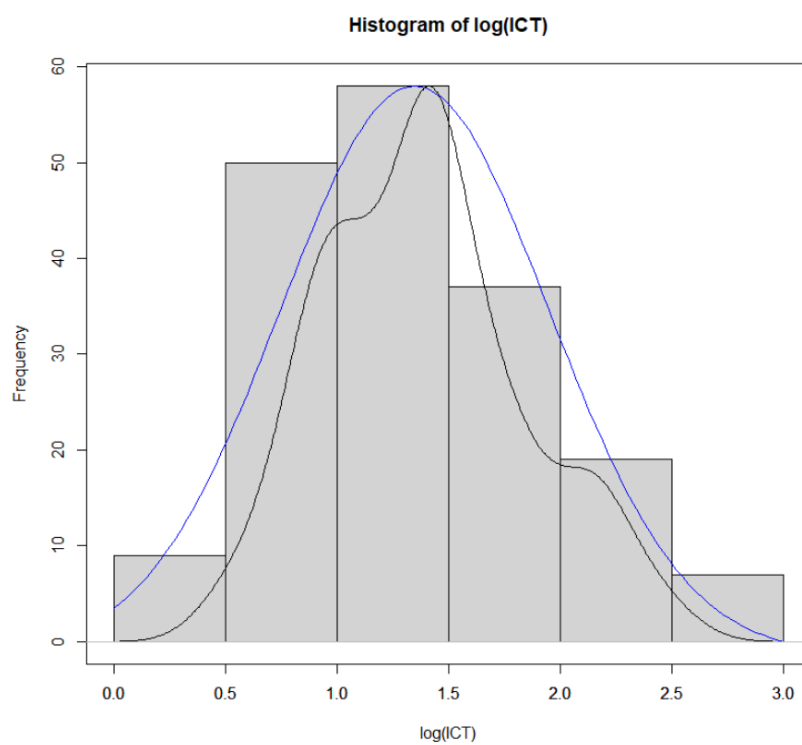
Παρακάτω, παρουσιάζονται τα ιστογράμματα των νέων-λογαριθμισμένων-μεταβλητών.



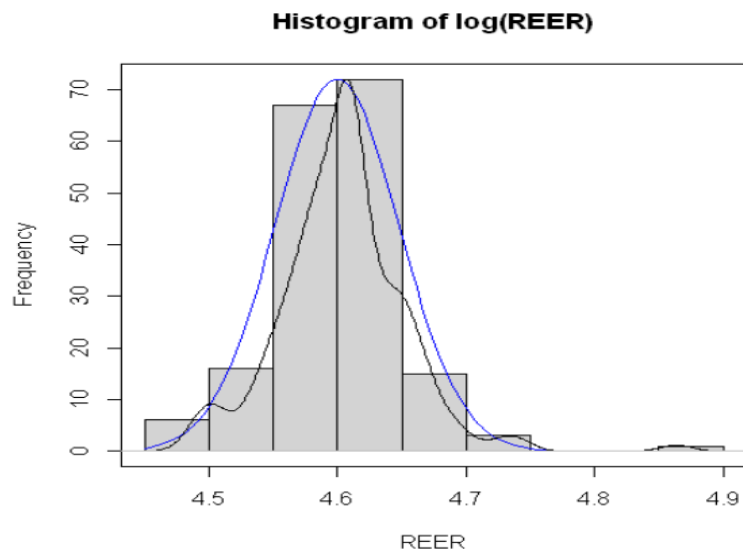
Διάγραμμα 4.9: Ιστόγραμμα φυσικού λογαρίθμου εξαγωγών υψηλής τεχνολογίας



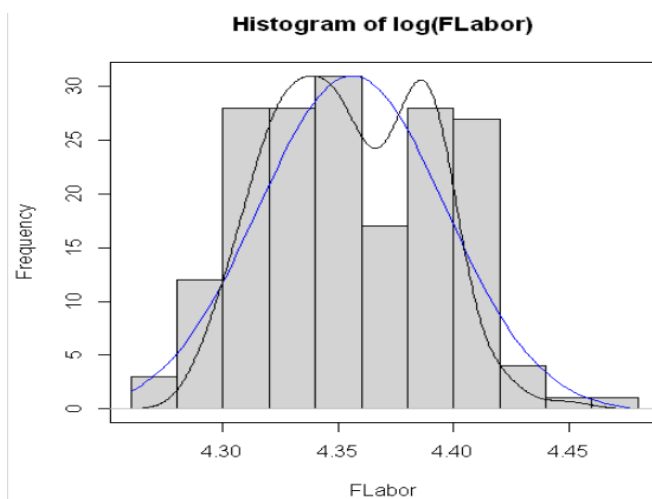
Διάγραμμα 4.10: Ιστόγραμμα φυσικού λογαρίθμου Δαπανών για Έρευνα και Ανάπτυξη



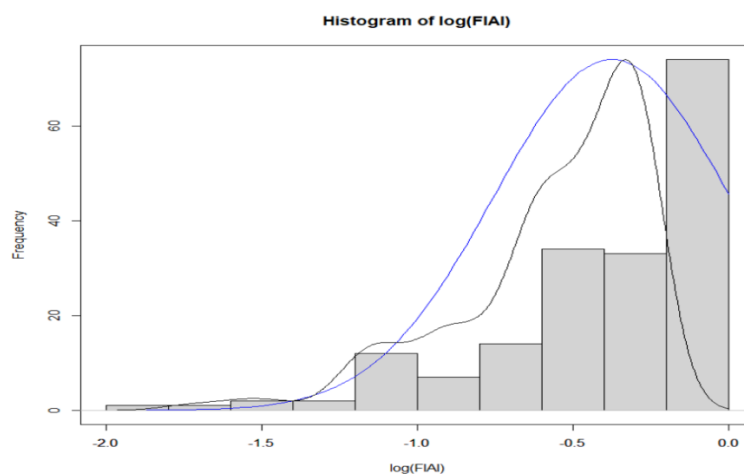
Διάγραμμα 4.11: Ιστόγραμμα φυσικού λογαρίθμου εξαγωγής αγαθών Τεχνολογίας Πληροφορικής και Τηλεπικοινωνιών



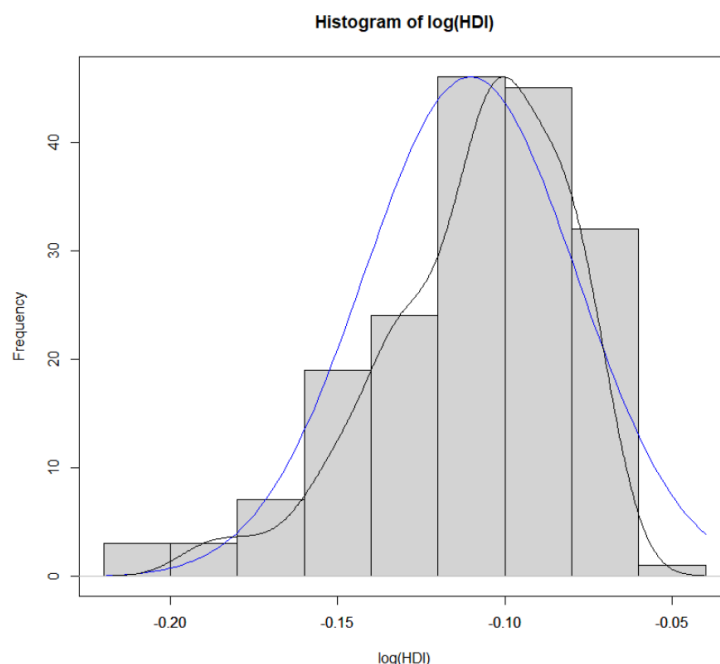
Διάγραμμα 4.12: Ιστόγραμμα φυσικού λογαρίθμου Δείκτη Πραγματικής Συναλλαγματικής Ισοτιμίας



Διάγραμμα 4.13: Ιστόγραμμα φυσικού λογαρίθμου Εργατικού δυναμικού με προχωρημένη εκπαίδευση, γυναίκα



Διάγραμμα 4.14: Ιστόγραμμα φυσικού λογαρίθμου Δείκτη Πρόσβασης Χρηματοπιστωτικών Ιδρυμάτων



Διάγραμμα 4.15: Ιστογράμμο φυσικού λογαρίθμου Δείκτη Ανθρώπινης Ανάπτυξης

Συγκρίνοντας τα ιστογράμματα της κάθε μεταβλητής παρατηρούμε ότι μετά τη χρήση λογαρίθμων κάποιες κατανομές κανονικοποιούνται/ βελτιώνονται («εξαγωγής αγαθών τεχνολογίας πληροφορικής και τηλεπικοινωνιών» (ICT), «δείκτη πραγματικής συναλλαγματικής ισοτιμίας» (REER) και «εργατικού δυναμικού με προχωρημένη εκπαίδευση, γυναίκα» (FLabor)), άλλες απομακρύνονται περισσότερο από την κανονική κατανομή («Δαπάνες Έρευνας και Ανάπτυξης» (RD)) και άλλες διατηρούν τη μορφή της κατανομής τους («εξαγωγών υψηλής τεχνολογίας» (HTX), «δείκτη πρόσβασης χρηματοπιστωτικών ιδρυμάτων» (FIAI) και «δείκτης ανθρώπινης ανάπτυξης» (HDI)).

Ένας ακόμα σύντομος τρόπος να διαπιστώσουμε αν παίρνουμε καλύτερες κατανομές με χρήση φυσικών λογαρίθμων είναι τα θηκογράμματα (boxplot). Αυτά τα διαγράμματα παραθέτουν το 25°, 50° και 75° εκατοστημόριο, σε αντίθεση προς τα διαγράμματα συχνοτήτων που δίνουν μόνο το μέσο. Τα θηκογράμματα είναι γραφήματα τα οποία συνοψίζουν βασικά περιγραφικά μέτρα, όπως η διάμεσος, τα τεταρτημόρια, το ενδοτεταρτημοριακό εύρος, καθώς και οι ακραίες τιμές. Επίσης, μπορούν να προϋποθέσουν για τη σχηματική μορφή της κατανομής ως προς την ασυμμετρία που πιθανώς αυτή εμφανίζει. Ένα θηκόγραμμα αποτελείται από ένα ορθογώνιο παραλληλόγραμμο, όπου η κάτω οριζόντια πλευρά του αντιπροσωπεύει το 25° εκατοστημόριο (Q1), η πάνω οριζόντια πλευρά που αντιπροσωπεύει το 75° εκατοστημόριο (Q3) ενώ η οριζόντια γραμμή στο

εσωτερικό του είναι η διάμεσος.[14]

Επίσης, τα θηκογράμματα διαθέτουν οριζόντιες γραμμές, τους λεγόμενους φράκτες, σε αποστάσεις, ίσες το πολύ με $1,5(Q3 - Q1)$. Αν η μικρότερη ή η μεγαλύτερη τιμή βρίσκονται εντός των περιοχών, τότε οι φράκτες φέρονται ακριβώς στο ύψος των τιμών αυτών. Τιμές που βρίσκονται εκτός των φρακτών ονομάζονται ακραία σημεία ή ακραίες παρατηρήσεις (outliers). Με αυτά τα γραφήματα ελέγχουμε και συγκρίνουμε τις μεταβλητές (αρχικές και λογαριθμικές).

Από τα διαγράμματα (6.1-6.14) στο Α. ΠΑΡΑΡΤΗΜΑ ΑΠΟΤΕΛΕΣΜΑΤΩΝ, εξάγουμε τα παρακάτω συμπεράσματα. Κατ' αρχάς, παρατηρούμε ότι δεν υπάρχει μικρή εξομάλυνση στην κατανομή των μεταβλητών «εξαγωγές υψηλής τεχνολογίας» (HTX) και «Δαπανών για Έρευνα και Ανάπτυξης» (RD) μετά τη χρήση λογαρίθμων. Οι μεταβλητές «εξαγωγής αγαθών τεχνολογίας πληροφορικής και τηλεπικοινωνιών» (ICT) και «Δείκτης πραγματικής συναλλαγματικής» (REER) μετά τη λογαρίθμηση παρουσιάζει λιγότερες ακραίες τιμές και η κατανομή της πλησιάζει την κατανομή. Η λογαριθμική μορφή των μεταβλητών «δείκτη πρόσβασης χρηματοπιστωτικών ιδρυμάτων» (FIAI), και «δείκτη ανθρώπινης ανάπτυξης» (HDI) φαίνεται πως χειροτερεύει την κατανομή τους, δημιουργώντας ακραίες τιμές με αρνητική ασυμμετρία. Ο φυσικός λογάριθμος της «εργατικού δυναμικού με προχωρημένη εκπαίδευση, γυναίκα» (FLabor) εξομαλύνει τις ακραίες παρατηρήσεις και η κατανομή της προσεγγίζει την κανονική κατανομή.

Συμπερασματικά, από όσα προαναφέραμε 3 μεταβλητές βελτιώνουν τις κατανομές τους με τους φυσικούς λογαρίθμους προσεγγίζοντας την κανονική κατανομή, 1 παρουσιάζει περισσότερη ασυμμετρία ενώ 3 φαίνεται να διατηρούν την ίδια. Επομένως, θα κρατήσουμε για τις μεταβλητές που δε βελτιώνεται η κατανομή τους μετά τους λογαρίθμους, την αρχική μορφή τους ενώ για τις μεταβλητές που η κατανομή βελτιώνεται μετά τη λογαρίθμηση, θα κρατήσουμε τις μεταβλητές φυσικού λογαρίθμου.

Ο επόμενος έλεγχος αφορά την ύπαρξη συσχέτισης. Η απλούστερη ένδειξη για την ύπαρξη συσχέτισης μεταξύ των μεταβλητών είναι η συνδιακύμανση, που ορίζεται ως ο μέσος όρος του γινομένου των αποκλίσεων των δεδομένων από τις αντίστοιχες μέσες τιμές τους. Από τον έλεγχο λαμβάνουμε τον ακόλουθο πίνακα:

	HTX	RD	lnICT	lnREER	lnFLabor	FIAI	HDI
HTX	1.00	0.21	0.73	0.04	0.11	-0.34	0.49
RD	0.21	1.00	0.35	0.05	-0.19	-0.63	0.59
lnICT	0.73	0.35	1.00	0.22	0.32	-0.53	0.36
lnREER	0.04	0.05	0.22	1.00	0.26	0.14	-0.14
lnFLabor	0.11	-0.19	0.32	0.26	1.00	0.06	-0.26
FIAI	-0.34	-0.63	-0.53	0.14	0.06	1.00	-0.61
HDI	0.49	0.59	0.36	-0.14	-0.26	-0.61	1.00

Πίνακας 4.3.1: Συνδιακύμανση μεταβλητών του υποδείγματος εξαγωγών υψηλής τεχνολογίας

Η συνδιακύμανση λαμβάνει τιμές μεταξύ -1 και 1. Για τιμές κοντά στο 1, υπάρχει θετική συσχέτιση, αρνητική κοντά στο -1. Στον παραπάνω πίνακα, παρουσιάζεται η συνδιακύμανση μεταξύ όλων των μεταβλητών του υποδείγματος που εξετάζουμε. Η εξαρτημένη μεταβλητή «Εξαγωγές Υψηλής Τεχνολογίας» (HTX) σχετίζεται θετικά με τις ανεξάρτητες μεταβλητές «Δείκτης Έρευνας και Ανάπτυξης» (RD), «φυσικού λογαρίθμου εξαγωγής αγαθών Τεχνολογίας Πληροφορικής και Τηλεπικοινωνιών» (lnICT), «φυσικού λογαρίθμου Δείκτη Πραγματικής Συναλλαγματικής Ισοτιμίας» (lnREER), «φυσικού λογαρίθμου Εργατικού Δυναμικού με Προχωρημένη Εκπαίδευση, γυναίκα» (lnFLabor) και «Δείκτη Ανθρώπινης Ανάπτυξης» (HDI). Τέλος, η μεταβλητή «Εξαγωγές Υψηλής Τεχνολογίας» (HTX) εμφανίζει αρνητική συσχέτιση με τη μεταβλητή «Δείκτης Πρόσβασης Χρηματοπιστωτικών Ιδρυμάτων» (FIAI).

Τελευταία αφήσαμε τα διαγράμματα διασποράς. Το διάγραμμα διασποράς αναδεικνύει τη σχέση μεταξύ δυο μεταβλητών, δηλαδή το βαθμό συσχέτισής τους. Δημιουργούμε μια γραφική αναπαράσταση της μήτρας συσχετίσεων με την από κοινού παρουσίαση των διαγραμμάτων διασποράς για το σύνολο των μεταβλητών.[14]



Πίνακας 4.3.2: Γραφική μήτρα συσχέτισης μεταβλητών του υποδείγματος εξαγωγών υψηλής τεχνολογίας

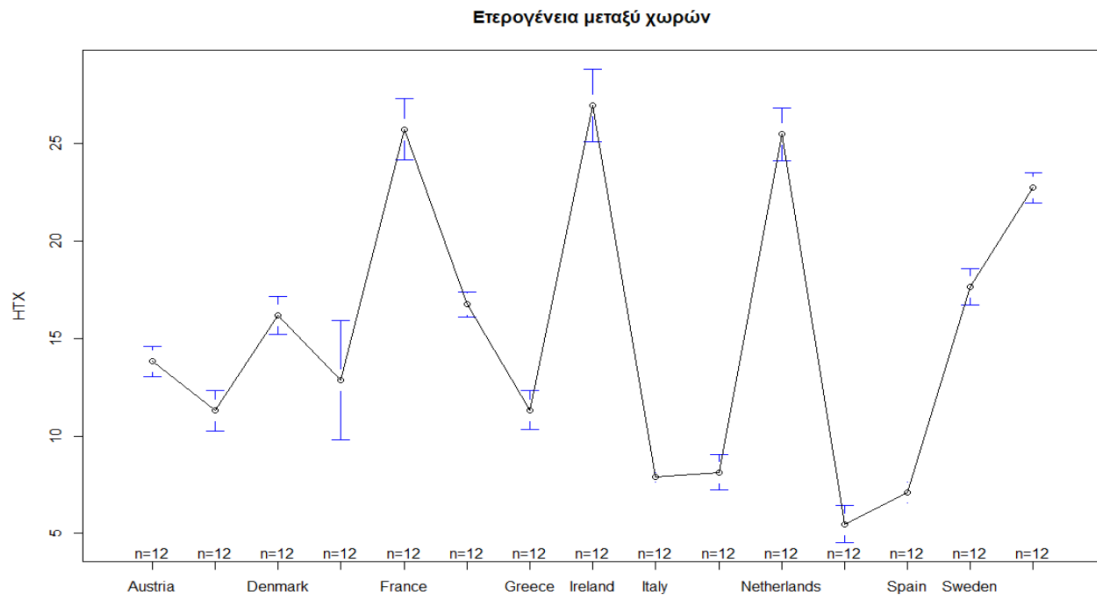
Ακολουθούν κάποιες συνοπτικές στατιστικές των δεδομένων πάνελ όπως μέσος, τυπική απόκλιση κλπ. Όπως μπορούμε να διακρίνουμε από τον Πίνακα 4.3.3, υπολογίζεται η τυπική απόκλιση κι επομένως και διακύμανση για κάθε μια από τις παρατηρήσεις.

Variable	N	Mean	Std. dev.	Min.	Median	Max.
HTX (High-tech exports (% of manufactured exports))	180	15.2804	7.2113	4.056	14.1240	32.7165
RD (R&D exp. % GDP)	180	2.0474	0.8145	0.5765	1.8988	3.7488
lnICT goodsexport	180	1.3489	0.6009	0.0690	1.3528	2.9251
lnRealEff Real effective ER	180	4.5994	0.0482	4.4843	4.6023	4.8541
lnLabor Labor force with advanced education(female)	180	4.3560	0.0407	4.2756	4.3533	4.4743
FIAI (Access index)	180	0.7310	0.2248	0.1600	0.7850	1.0000
HDI (Human Development Index)	180	0.8959	0.0280	0.8100	0.8990	0.9420

Πίνακας 4.3.3: Βασικές στατιστικές των μεταβλητών του υποδείγματος εξαγωγών υψηλής τεχνολογίας

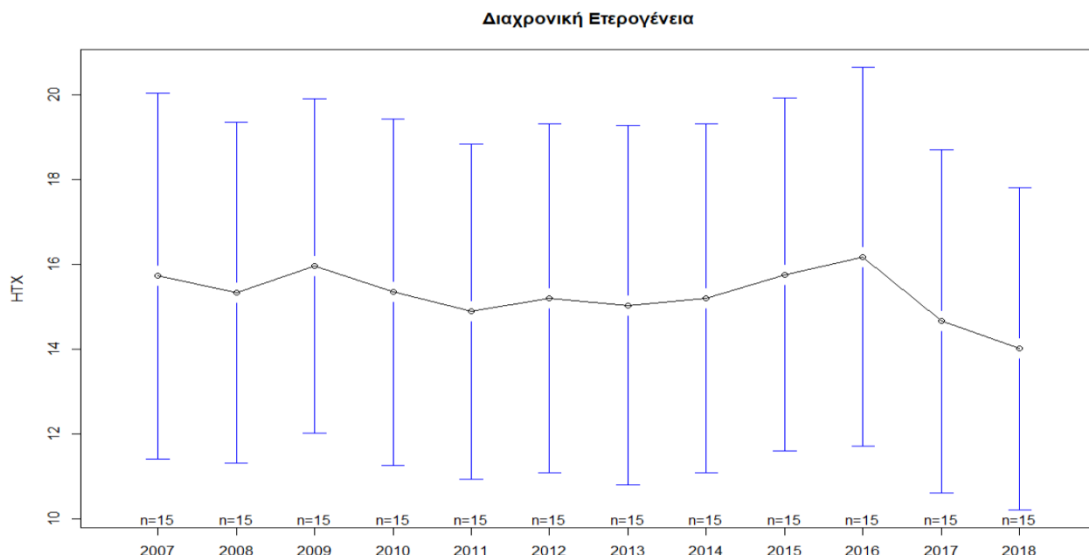
Όπως προκύπτει από τον Πίνακα 4.3.3, η μέση τιμή των εξαγωγών υψηλής τεχνολογίας (HTX) είναι 15.28% των κατασκευασμένων εξαγωγών με σχετικά μεγάλες αποκλίσεις που κυμαίνονται στο 7.21.

Τελευταίο βήμα πριν προχωρήσουμε στην ανάλυση παλινδρόμησης, το οποίο θα μας προϋδεάσει και για την κατάλληλη επιλογή μοντέλου για την ανάλυσή μας, αποτελεί ο έλεγχος ετερογένειας μεταξύ διαστρωματικών στοιχείων (χωρών) και διαχρονικά.



Διάγραμμα 4.16: Έλεγχος ύπαρξης ετερογένειας μεταξύ διαστρωματικών στοιχείων (εδώ χωρών) στο υπόδειγμα εξαγωγές υψηλής τεχνολογίας

Στο παραπάνω διάγραμμα, με μαύρο κύκλο, συμβολίζεται ο μέσος όρος της εξαρτημένης μεταβλητής «εξαγωγές υψηλής τεχνολογίας» (HTX) για κάθε διαστρωματικό στοιχείο (χώρα). Οι μπλε κατακόρυφες γραμμές αποτελούν ένα 95% διάστημα εμπιστοσύνης για κάθε μέσο όρο. Η γραμμή που ενώνει αυτούς τους μέσους καταδεικνύει το γεγονός ότι υπάρχει σημαντική διακύμανση μεταξύ των διαστρωματικών στοιχείων (χωρών). Επομένως, έχουμε μια ένδειξη ότι το μοντέλο σταθερών επιδράσεων (Fixed-effects model) φαίνεται να είναι κατάλληλο για την ανάλυσή μας, αφού προϋποθέτει πως ο σταθερός όρος, α_i διαφέρει μεταξύ διαστρωματικών στοιχείων. Άρα, αυτός ο όρος για την εξαρτημένη μεταβλητή όπως φαίνεται διαφέρει σχεδόν για όλες τις υπό εξέταση χώρες.



Διάγραμμα 4.17: Έλεγχος ύπαρξης ετερογένειας διαχρονικά στο υπόδειγμα εξαγωγές υψηλής τεχνολογίας

Στο παραπάνω διάγραμμα, οι μαύρες γραμμές που ενώνουν τους μέσους της εξαρτημένης μεταβλητής «εξαγωγές υψηλής τεχνολογίας» (HTX) διαχρονικά δείχνουν ότι δεν υπάρχει ουσιαστική απόκλιση της εξαρτημένης μεταξύ των ετών εντός της ίδιας διαστρωματικής μονάδας (χώρας) εκτός ίσως μια μικρή απόκλιση μετά το έτος 2016. Για το λόγο αυτό και για να δούμε κατά πόσο επηρεάζει η χρονική περίοδος 2016-2018 ορίζουμε μια ψευδομεταβλητή After2016. Η μεταβλητή αυτή θα παίρνει την τιμή 1 για το διάστημα 2016-2018 και 0 για το υπόλοιπο χρονικό διάστημα και θα συμπεριληφθεί στις ανεξάρτητες μεταβλητές του μοντέλου μας.

Στη συνέχεια ακολουθεί η ανάλυση παλινδρόμησης με τις προαναφερθείσες μεθόδους και η διεξαγωγή ελέγχων για την επιλογή της κατάλληλης μεθόδου ανάλυσης.[14]

4.4 ΑΝΑΛΥΣΗ ΠΑΛΙΝΔΡΟΜΗΣΗΣ ΤΩΝ ΕΞΑΓΩΓΩΝ ΥΨΗΛΗΣ ΤΕΧΝΟΛΟΓΙΑΣ (ΕΚΤΙΜΗΣΗ-ΕΡΜΗΝΕΙΑ-ΣΥΜΠΕΡΑΣΜΑΤΑ)

Για την ανάλυση της παλινδρόμησης θα επιλέξουμε μεταξύ του Ενιαίου μοντέλου Ελαχίστων Τετραγώνων – Pooled OLS Model του μοντέλου Σταθερών Επιδράσεων (Fixed-Effects Model) και του μοντέλου Τυχαίων Επιδράσεων (Random-Effects Model). Το μοντέλο Τυχαίων Επιδράσεων θεωρείται ως ελκυστικό διότι τα παρατηρούμενα χαρακτηριστικά που είναι σταθερά διατηρούνται στο μοντέλο ενώ στην εκτίμηση σταθερών επιδράσεων πρέπει να εξαλειφθούν. Επιπλέον με το μοντέλο Τυχαίων Επιδράσεων δεν χάνουμε βαθμούς ελευθερίας όπως με των Σταθερών Επιδράσεων. Τέλος μέσα από τους ελέγχους αν διαπιστωθεί ότι το στατιστικό LM είναι σημαντικό, θα πρέπει να χρησιμοποιήσουμε το μοντέλο Τυχαίων Επιδράσεων αντί του Ενιαίου μοντέλου. Ακόμα στον έλεγχο Hausman αν διαπιστωθεί ότι η $p\text{-value} < 0.05$, τότε είναι ασφαλές να χρησιμοποιήσουμε το μοντέλο Τυχαίων Επιδράσεων σε σχέση με το μοντέλο Σταθερών Επιδράσεων.

Στον Πίνακα 4.4.1 δίνονται τα αποτελέσματα που λάβαμε μέσω του πακέτου plm του λογισμικού R με χρήση του εκτιμητή των Ενιαίων Δεδομένων (Pooled OLS Estimator).

Pooling Model

Call:

```
plm(formula = Y ~ X + After2016, data = pdata, model = "pooling")
```

Balanced Panel: n = 15, T = 12, N = 180

Residuals:

Min.	1st Qu.	Median	3rd Qu.	Max.
-8.2421	-2.6128	-0.3909	1.7628	11.3264

Coefficients:

	Estimate	Std. Error	t-value	Pr(> t)	
(Intercept)	14.37434	53.32274	0.2696	0.7878124	
XRD	-1.67448	0.54514	-3.0717	0.0024752	**
XlnICT	9.57099	0.71937	13.3047	< 2.2e-16	***
XlnREER	-17.14400	7.45006	-2.3012	0.0225805	*
XlnFLabor	-9.53252	9.13588	-1.0434	0.2982203	
XFAI	8.07008	2.20313	3.6630	0.0003316	***
XHDI	118.56347	16.03930	7.3921	6.039e-12	***
After2016	-1.96950	0.90345	-2.1800	0.0306175	*

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

Total Sum of Squares: 9308.6

Residual Sum of Squares: 3009.5

R-Squared: 0.6767

Adj. R-Squared: 0.66354

F-statistic: 51.4297 on 7 and 172 DF, p-value: < 2.22e-16

Πίνακας 4.4.1: Αποτελέσματα παλινδρόμησης Ενιαίου Μοντέλου

***συμβολίζει επίπεδο στατιστικής σημαντικότητας 1%(p-value<0.001)

**συμβολίζει επίπεδο στατιστικής σημαντικότητας 1% (p-value<0.01)

*συμβολίζει επίπεδο στατιστικής σημαντικότητας 5%(p-value<0.05)

. συμβολίζει επίπεδο στατιστικής σημαντικότητας 1%(p-value<0.1)

Όπως προκύπτει από τα αποτελέσματα του Πίνακα 4.4.1 για το Ενιαίο Μοντέλο, οι περισσότερες μεταβλητές είναι στατιστικά σημαντικές για το συγκεκριμένο μοντέλο, εκτός από τη μεταβλητή «εργατικού δυναμικού με προχωρημένη εκπαίδευση, γυναίκα» (lnFLabor) και τη σταθερά (Intercept). Ένα ακόμα στοιχείο που εξετάζουμε είναι και ο δείκτης R-squared (συντελεστής προσδιορισμού) ο οποίος εκφράζει το ποσοστό της διακύμανσης της εξαρτημένης μεταβλητής το οποίο ερμηνεύεται από τη διακύμανση των τιμών των ανεξάρτητων μεταβλητών. Δηλαδή στην περίπτωση μας, ο εκτιμητής Ενιαίου Μοντέλου μπορεί να ερμηνεύσει το 67,6% της μεταβλητότητας των εξαγωγών υψηλής τεχνολογίας (HTX). Επίσης για το Ενιαίο Μοντέλο, η R δίνει και τον προσαρμοσμένο συντελεστή προσδιορισμού Adj. R-Squared= 0.66354, ο οποίος δεν διαφοροποιείται σχεδόν από τον R-squared και δείχνει ότι το μοντέλο μπορεί να ερμηνεύσει αρκετά καλά την μεταβλητότητα των εξαγωγών υψηλής τεχνολογίας (HTX).

Στον Πίνακα 4.4.2 δίνονται τα αποτελέσματα που λάβαμε μέσω του πακέτου plm του λογισμικού R με χρήση του εκτιμητή των Σταθερών Επιδράσεων (Fixed Effects Estimator).

```

Oneway (individual) effect within Model

Call:
plm(formula = Y ~ X + After2016, data = pdata, model = "within")

Balanced Panel: n = 15, T = 12, N = 180

Residuals:
    Min.   1st Qu.   Median   3rd Qu.    Max.
-4.89754 -0.75754  0.15448  0.84918  3.15765

Coefficients:
              Estimate Std. Error t-value Pr(>|t|)
XRD             1.05000    0.56665   1.8530  0.065747 .
XlnICT           5.73508    0.44144  12.9918 < 2.2e-16 ***
XlnREER        -10.31682    3.21083  -3.2131  0.001592 **
XlnFLabor      -21.10854    7.96669  -2.6496  0.008878 **
XFIAI           13.55254    2.21112   6.1293 6.774e-09 ***
XHDI            91.28578   15.36605   5.9407 1.749e-08 ***
After2016      -1.56215    0.33331  -4.6867 5.955e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total sum of Squares:    670.67
Residual Sum of Squares: 268.27
R-Squared:              0.6
Adj. R-Squared: 0.54683
F-statistic: 33.8569 on 7 and 158 DF, p-value: < 2.22e-16

```

Πίνακας 4.4.2: Αποτελέσματα παλινδρόμησης με τη μέθοδο των Σταθερών Επιδράσεων

***συμβολίζει επίπεδο στατιστικής σημαντικότητας 1%(p-value<0.001)

**συμβολίζει επίπεδο στατιστικής σημαντικότητας 1% (p-value<0.01)

*συμβολίζει επίπεδο στατιστικής σημαντικότητας 5% (p-value<0.05)

. συμβολίζει επίπεδο στατιστικής σημαντικότητας 10% (p-value<0.1)

Όπως προκύπτει από τα αποτελέσματα για το Μοντέλο Σταθερών Επιδράσεων όλες οι μεταβλητές είναι στατιστικά σημαντικές για το συγκεκριμένο μοντέλο, δηλαδή εξηγούν επαρκώς το μοντέλο και την ανεξάρτητη μεταβλητή, επομένως το μοντέλο μας είναι καλό. Σύμφωνα με τον F έλεγχο και την αντίστοιχη μηδενική p-value διαπιστώνουμε ότι οι επεξηγηματικές μεταβλητές στο σύνολό τους είναι στατιστικά σημαντικές δηλαδή εξηγούν πλήρως το μοντέλο και την εξαρτημένη μεταβλητή. Συνήθως μια p-τιμή η οποία είναι μικρότερη από 0.05 καθιστά το μοντέλο ακριβές και αξιόπιστο.

Από τις p-value του ελέγχου για κάθε μεταβλητή, συμπεραίνουμε με μια πρώτη ματιά ότι οι μεταβλητές «εξαγωγής αγαθών τεχνολογίας πληροφορικής και τηλεπικοινωνιών σε

φυσικούς λογαρίθμους» (lnICT), «δείκτη πρόσβασης χρηματοπιστωτικών ιδρυμάτων» (FIAI), «δείκτη ανθρώπινης ανάπτυξης» (HDI) και «ψευδομεταβλητή After2016» είναι στατιστικά σημαντικές σε επίπεδο σημαντικότητας 1% (p-value<0.001) ενώ οι μεταβλητές «δείκτη πραγματικής συναλλαγματικής ισοτιμίας σε φυσικούς λογαρίθμους» (lnREER) και «εργατικού δυναμικού με προχωρημένη εκπαίδευση, γυναίκα σε φυσικούς λογαρίθμους» (lnFLabor) είναι στατιστικά σημαντικές σε επίπεδο σημαντικότητας 1% (p-value<0.01). Τέλος η μεταβλητή «Δείκτη Έρευνας και Ανάπτυξης» (RD) είναι σημαντική σε επίπεδο σημαντικότητας 10% (p-value<0.1).

Ένα ακόμα στοιχείο που εξετάζουμε είναι και ο δείκτης R-squared ο οποίος εκφράζει το ποσοστό της διακύμανσης της εξαρτημένης μεταβλητής το οποίο ερμηνεύεται από τη διακύμανση των τιμών των ανεξάρτητων μεταβλητών. Δηλαδή στην περίπτωση μας, ο εκτιμητής σταθερών επιδράσεων μπορεί να ερμηνεύσει το 60% της διακύμανσης της εξαρτημένης μεταβλητής. Επίσης για το Υπόδειγμα Σταθερών Επιδράσεων, εξετάσαμε και τον προσαρμοσμένο συντελεστή προσδιορισμού Adj. R-Squared= 0.54683, ο οποίος διαφοροποιείται ελαφρώς από τον R-squared και δείχνει ότι ακριβώς και ο συντελεστής R-squared, αλλά προσαρμόζεται από τον αριθμό των περιπτώσεων και των μεταβλητών. Όταν ο αριθμός των περιπτώσεων είναι αρκετά μεγάλος και ο αριθμός των μεταβλητών είναι μικρός τότε ο προσαρμοσμένος συντελεστή προσδιορισμού Adj.R-Squared «πλησιάζει» τον συντελεστή προσδιορισμού R-squared.

Στον Πίνακα 4.4.3 δίνονται τα αποτελέσματα που λάβαμε μέσω του πακέτου plm του λογισμικού R με χρήση του εκτιμητή Τυχαίων Επιδράσεων (Random Effects Estimator).

```
Oneway (individual) effect Random Effect Model
(Swamy-Arora's transformation)
```

```

call:
plm(formula = Y ~ X + After2016, data = pdata, model = "random")

Balanced Panel: n = 15, T = 12, N = 180

Effects:
              var std.dev share
idiosyncratic 1.698  1.303 0.061
individual    26.126  5.111 0.939
theta: 0.9266

Residuals:
      Min.      1st Qu.        Median      3rd Qu.        Max.
-4.293402 -0.833320   0.035951   0.905967   2.969870

Coefficients:
              Estimate Std. Error z-value Pr(>|z|)
(Intercept)  40.89236   36.97608   1.1059  0.268764
XRD           0.90332    0.54386   1.6610  0.096722 .
XlnICT        5.88268    0.43410  13.5514 < 2.2e-16 ***
XlnREER       -10.13901   3.14692  -3.2219  0.001274 **
XlnFLabor     -18.91134   7.73749  -2.4441  0.014521 *
XFIAI         12.67088    2.11908   5.9794  2.239e-09 ***
XHDI          94.44854   14.74524   6.4054  1.500e-10 ***
After2016     -1.62939    0.32667  -4.9879  6.104e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares:    717.2
Residual Sum of Squares: 291.67
R-Squared:                0.59332
Adj. R-Squared: 0.57677
Chisq: 250.934 on 7 DF, p-value: < 2.22e-16

```

Πίνακας 4.4.3: Αποτελέσματα παλινδρόμησης με τη μέθοδο των Τυχαίων Επιδράσεων

***συμβολίζει επίπεδο στατιστικής σημαντικότητας 1%(p-value<0.001)
 **συμβολίζει επίπεδο στατιστικής σημαντικότητας 1% (p-value<0.01)
 *συμβολίζει επίπεδο στατιστικής σημαντικότητας 5% (p-value<0.05)
 . συμβολίζει επίπεδο στατιστικής σημαντικότητας 1% (p-value<0.1)

Ο F έλεγχος, καθώς και η πιθανότητα για αυτόν ($p < 0,05$), μας οδηγούν στο συμπέρασμα ότι οι μεταβλητές στο σύνολό τους είναι στατιστικά σημαντικές για την ερμηνεία του μοντέλου μας.

Όπως προκύπτει από τα αποτελέσματα για το Μοντέλο Τυχαίων Επιδράσεων όλες οι μεταβλητές εκτός από τη σταθερά είναι στατιστικά σημαντικές. Αυτό σημαίνει ότι εξηγούν επαρκώς το μοντέλο και την εξαρτημένη μεταβλητή, επομένως το Μοντέλο Τυχαίων Επιδράσεων είναι κατάλληλο. Επιπλέον, σύμφωνα με το $\lambda = 92,6\%$ (συμβολίζεται με theta στον Πίνακα 4.4.3), οι εκτιμήσεις του μοντέλου τυχαίων επιδράσεων είναι πολύ πιο κοντά στις εκτιμήσεις του μοντέλου σταθερών επιδράσεων από ό,τι στις εκτιμήσεις του ενιαίου μοντέλου.

Από τις p-value προκύπτει ότι οι μεταβλητές «εξαγωγής αγαθών τεχνολογίας πληροφορικής και τηλεπικοινωνιών σε φυσικούς λογαρίθμους» (lnICT), «δείκτη πρόσβασης χρηματοπιστωτικών ιδρυμάτων» (FIAI), «δείκτη ανθρώπινης ανάπτυξης» (HDI), και «ψευδομεταβλητή After2016» είναι στατιστικά σημαντικές σε επίπεδο σημαντικότητας 1% (p-value<0.001) ενώ η μεταβλητή «δείκτη πραγματικής συναλλαγματικής ισοτιμίας σε φυσικούς λογαρίθμους» (lnREER) είναι στατιστικά σημαντική σε επίπεδο σημαντικότητας 1% (p-value<0.01). Ακόμα η μεταβλητή «και «εργατικού δυναμικού με προχωρημένη εκπαίδευση, γυναίκα σε φυσικούς λογαρίθμους» (lnFLabor) είναι σημαντική σε επίπεδο σημαντικότητας 5% (p-value<0.05) και τέλος η μεταβλητή «Δείκτη Έρευνας και Ανάπτυξης» (RD) είναι σημαντική σε επίπεδο σημαντικότητας 10% (p-value<0.1). Οι παραπάνω συντελεστές αναπαριστούν τη μέση επίδραση κάθε ερμηνευτικής μεταβλητής επάνω στην εξαρτημένη μεταβλητή, όταν αυτή μεταβάλλεται διαχρονικά και μεταξύ διαστρωματικών στοιχείων (εδώ χωρών) κατά μία μονάδα.

Σύμφωνα με τον δείκτη R-squared, ο εκτιμητής τυχαίων επιδράσεων μπορεί να ερμηνεύσει το 59,3% της μεταβολής “εντός”. Επίσης για το μοντέλο Τυχαίων Επιδράσεων, εξετάσαμε και τον προσαρμοσμένο συντελεστή προσδιορισμού Adj. R-Squared= 0.57677, ο οποίος δεν διαφοροποιείται ιδιαίτερα από το R-squared και δείχνει ότι το μοντέλο έχει καλά σημάδια για την εξήγηση της μεταβλητότητας “εντός” της εξαρτημένης μεταβλητής. Τέλος, όπως προκύπτει από τον δείκτη ρ (βλ. αποτελέσματα Effect: share (individual)) στον Πίνακα 4.4.3) μεγάλο ποσοστό 93,9% ερμηνεύεται από τις απαραίτητες επιδράσεις και το υπόλοιπο από τον ιδιοσυγκρατικό όρο του σφάλματος.

Σε κάθε περίπτωση και παρά τα όποια συμπεράσματα βγάζουμε εκ πρώτης από τα αποτελέσματα των παλινδρομήσεων, θα εφαρμόσουμε το τεστ του Hausman για να διαπιστώσουμε έγκυρα ποιο μοντέλο (Fixed or Random-Effects Model) είναι κατάλληλο να χρησιμοποιήσουμε στην ανάλυσή μας. Στον έλεγχο αυτόν η μηδενική υπόθεση είναι H_0 : Το μοντέλο Random-Effects είναι κατάλληλο (δεν υπάρχει συσχέτιση μεταξύ των σφαλμάτων ιδιοσυγκρασίας με τις ανεξάρτητες μεταβλητές) με εναλλακτική υπόθεση H_1 : Το μοντέλο Fixed-Effects είναι κατάλληλο (υπάρχει συσχέτιση).

Hausman Test

```
data: Y ~ X + After2016  
chisq = 5.5896, df = 7, p-value = 0.5884  
alternative hypothesis: one model is inconsistent
```

Πίνακας 4.4.4: Έλεγχος Hausman καταλληλότητας μοντέλου

Παρατηρούμε ότι η τιμή της $p\text{-value}=0.5884$, οπότε δεν μπορούμε να απορρίψουμε τη μηδενική υπόθεση, η οποία δηλώνει πως το Μοντέλο Τυχαίων Επιδράσεων είναι περισσότερο αντιπροσωπευτικό για την τρέχουσα ανάλυση σε σχέση με το Μοντέλο Σταθερών Επιδράσεων.

Εφόσον το μοντέλο των τυχαίων επιδράσεων κρίθηκε καταλληλότερο από εκείνο των σταθερών, θα προχωρήσουμε σε έναν έλεγχο για ύπαρξη ετεροσκεδαστικότητας, τον έλεγχο Lagrangian Multiplier ή LM. Ο προαναφερθείς έλεγχος, μας βοηθά να αποφασίσουμε ποιο μοντέλο παλινδρόμησης είναι καταλληλότερο, αυτό των τυχαίων επιδράσεων ή το ενιαίο μοντέλο. Το ενιαίο μοντέλο (Pooled OLS Regression Model) προκύπτει από τη χρονική ομαδοποίηση των διαστρωματικών δεδομένων. Η μηδενική υπόθεση ενός LM test είναι ότι η διακύμανση μεταξύ διαστρωματικών στοιχείων (εδώ χωρών) είναι μηδενική, δηλ. δεν υπάρχει σημαντική διαφορά ανάμεσα στα διαστρωματικά στοιχεία (panel effect).

Lagrange Multiplier Test - (Honda) for balanced panels

```
data: Y ~ X + After2016  
normal = 26.458, p-value < 2.2e-16  
alternative hypothesis: significant effects
```

Πίνακας 4.4.5: Έλεγχος Breusch-Pagan Lagrangian Multiplier καταλληλότητας του μοντέλου

Η $p\text{-value}<0.05$, άρα πρέπει να απορρίψουμε τη μηδενική υπόθεση, που σημαίνει ότι υπάρχει ένδειξη σημαντικών διαφορών μεταξύ των χωρών, κι επομένως, το υπόδειγμα τυχαίων επιδράσεων θεωρείται κατάλληλο για την ανάλυση μας. Αξίζει να σημειωθεί ότι σύμφωνα με το παραπάνω συμπέρασμα, υπάρχουν σημαντικές διαφορές μεταξύ των διαστρωματικών μονάδων (χωρών).

Συμπερασματικά, το μοντέλο τυχαίων επιδράσεων είναι εκείνο που εξηγεί καλύτερα τα δεδομένα της προκείμενης ανάλυσης σε σύγκριση με το ενιαίο μοντέλο και το μοντέλο σταθερών επιδράσεων.

Έλεγχος για την ύπαρξη ετεροσκεδαστικότητας για δεδομένα πάνελ

Στη συνέχεια και σαν ένα από τα τελευταία βήματα της ανάλυσης είναι ο έλεγχος για ύπαρξη ετεροσκεδαστικότητας στο μοντέλο μας. Ο έλεγχος θεωρεί ως μηδενική υπόθεση H_0 : Τα κατάλοιπα είναι ομοσκεδαστικά και ως εναλλακτική ότι H_1 : τα κατάλοιπα είναι ετεροσκεδαστικά. Ο έλεγχος έγινε με την βοήθεια της συνάρτησης `bptest` και τα αποτελέσματα δίνονται στον Πίνακα 4.4.6.

Breusch-Pagan test

```
data: Y ~ X + factor(id)
BP = 42.762, df = 20, p-value = 0.002197
```

Πίνακας 4.4.6: Έλεγχος ετεροσκεδαστικότητας

Από τον παραπάνω πίνακα, αφού $p\text{-value} < 0.05$, απορρίπτουμε τη μηδενική υπόθεση. Επομένως, σε επίπεδο σημαντικότητας 5% μπορεί εύκολα να εξαχθεί το συμπέρασμα ότι στο μοντέλο υπάρχει ετεροσκεδαστικότητα.

Αφού ανιχνεύθηκε ετεροσκεδαστικότητα, θα πρέπει να χρησιμοποιήσουμε έναν εύρωστο (robust) πίνακα συνδιακύμανσης (Sandwich estimator) για τους συντελεστές της παλινδρόμησης ώστε να την αντιμετωπίσουμε.

Με χρήση του ορίσματος `vcovHC` στην συνάρτηση `coeftest` της R προκύπτουν συνεπείς συντελεστές του μοντέλου Τυχαίων Επιδράσεων με την παρουσία ετεροσκεδαστικότητας.

Με την παρακάτω εντολή παράγουμε τους αρχικούς συντελεστές του μοντέλου τυχαίων επιδράσεων.[25][29]

```
> coeftest(random)
```

```
t test of coefficients:
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	40.89236	36.97608	1.1059	0.270308
XRD	0.90332	0.54386	1.6610	0.098544 .
XlnICT	5.88268	0.43410	13.5514	< 2.2e-16 ***
XlnREER	-10.13901	3.14692	-3.2219	0.001523 **
XlnFLabor	-18.91134	7.73749	-2.4441	0.015530 *
XFIAI	12.67088	2.11908	5.9794	1.261e-08 ***
XHDI	94.44854	14.74524	6.4054	1.384e-09 ***
After2016	-1.62939	0.32667	-4.9879	1.483e-06 ***

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

Εισάγοντας το όρισμα `vcovHC`, προκύπτουν οι παρακάτω συνεπείς εκτιμήσεις των συντελεστών του μοντέλου χωρίς παρουσία ετεροσκεδαστικότητας

```
> coeftest(random, vcovHC)
```

t test of coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	40.89236	51.21386	0.7985	0.425703
XRD	0.90332	1.03533	0.8725	0.384154
XlnICT	5.88268	0.83029	7.0851	3.421e-11 ***
XlnREER	-10.13901	3.47667	-2.9163	0.004015 **
XlnFLabor	-18.91134	10.53986	-1.7943	0.074527 .
XFIAT	12.67088	2.64062	4.7984	3.453e-06 ***
XHDI	94.44854	13.32947	7.0857	3.409e-11 ***
After2016	-1.62939	0.22094	-7.3748	6.666e-12 ***

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

Όπως προκύπτει από τα παραπάνω αποτελέσματα, οι εκτιμήσεις των συντελεστών παλινδρόμησης είναι ακριβώς ίδιοι αλλά υπάρχουν μερικές διαφοροποιήσεις στις τιμές των τυπικών σφαλμάτων και κατά συνέπεια του ελέγχου t και των αντίστοιχων p-τιμών. Ως συνέπεια αυτών των διαφοροποιήσεων η μεταβλητή «Δείκτη Έρευνας και Ανάπτυξης» (RD) δεν είναι στατιστικά σημαντική μεταβλητή στην εξήγηση των Εξαγωγών Υψηλής Τεχνολογίας (HTX), ενώ η μεταβλητή «εργατικού δυναμικού με προχωρημένη εκπαίδευση, γυναίκα σε φυσικούς λογαρίθμους» (lnFLabor) είναι στατιστικά σημαντική σε μικρότερο επίπεδο σημαντικότητας (10% από 1%). Οι υπόλοιπες μεταβλητές παραμένουν στατιστικά σημαντικές στην εξήγηση των Εξαγωγών Υψηλής Τεχνολογίας (HTX) στο ίδιο επίπεδο σημαντικότητας 1%.

Ερμηνεία των συντελεστών

Στο σημείο αυτό, θα προσπαθήσουμε να ερμηνεύσουμε τα αποτελέσματα της παλινδρόμησης με βάση το μοντέλο Τυχαίων Επιδράσεων (Πίνακα 4.4.3). Σύμφωνα με την εκτιμηθείσα παλινδρόμηση, παρατηρείται ότι οι «Δαπάνες για Έρευνα και Ανάπτυξη (ως ποσοστό του ΑΕΠ)» (RD) δεν είναι στατιστικά σημαντική μεταβλητή στην εξήγηση των «εξαγωγών υψηλής τεχνολογίας» (HTX). Το γεγονός αυτό έρχεται σε αντίθεση με τα αποτελέσματα της διεθνούς βιβλιογραφία [18]. Επισημαίνουμε όμως ότι στο αρχικό μοντέλο η μεταβλητή RD ήταν στατιστικά σημαντική σε επίπεδο σημαντικότητας 10%.

Επίσης ο λογάριθμος των «Εξαγωγές αγαθών Τεχνολογίας Πληροφορικής και Τηλεπικοινωνιών ως ποσοστό των συνολικών εξαγωγών» (lnICT) καθώς και ο «Δείκτης

Πρόσβασης Χρηματοπιστωτικών Ιδρυμάτων» (FIAI) έχουν θετική επίδραση στις «εξαγωγές υψηλής τεχνολογίας» (HTX). Ειδικότερα, η παραπάνω μελέτη καταδεικνύει ότι η χρηματοοικονομική ανάπτυξη βελτιώνει σημαντικά (σε επίπεδο σημαντικότητας 1%) τις «εξαγωγές υψηλής τεχνολογίας» (HTX) αφού κάθε επιπλέον μονάδα του «Δείκτη Πρόσβασης Χρηματοπιστωτικών Ιδρυμάτων» (FIAI) πάνω από τον μέσο όρο, επιφέρει αύξηση κατά 12.67 μονάδες στις εξαγωγές υψηλής τεχνολογίας» (HTX). Το συμπέρασμα αυτό συμβαδίζει με εκείνο στο οποίο καταλήγουν οι Gaur M. κα. [18]. Επιπροσθέτως, ένας άλλος παράγοντας που φαίνεται να επιδρά θετικά (σε επίπεδο σημαντικότητας 1%) στις «εξαγωγές υψηλής τεχνολογίας» (HTX) είναι ο «Δείκτης Ανθρώπινης Ανάπτυξης» (HDI). Μάλιστα, ο «Δείκτης Ανθρώπινης Ανάπτυξης» (HDI) έχει τη μεγαλύτερη θετική επίδραση, αφού κάθε επιπλέον μονάδα στον δείκτη HDI πάνω από τον μέσο όρο επιφέρει αύξηση κατά 94.45 στις «εξαγωγές υψηλής τεχνολογίας» (HTX). Το συμπέρασμα αυτό είναι πάλι σε συμφωνία με τη διεθνή βιβλιογραφία [17].

Επιπλέον, από τα παραπάνω αποτελέσματα προκύπτει ότι ο λογάριθμος του «Δείκτη Πραγματικής Συναλλαγματικής Ισοτιμίας» (lnREER), έχει αρνητική επίδραση (σε επίπεδο σημαντικότητας 1%) στις «εξαγωγές υψηλής τεχνολογίας» (HTX). Το αποτέλεσμα αυτό συμβαδίζει με τη γενική παραδοχή ότι η αύξηση της συναλλαγματικής ισοτιμίας μειώνει τις καθαρές εξαγωγές μακροχρόνια.

Επίσης, ενδιαφέρον παρουσιάζει η αρνητική επίδραση (σε επίπεδο σημαντικότητας 5%) του λογάριθμου του «Εργατικού δυναμικού με προχωρημένη εκπαίδευση, γυναίκα (ως ποσοστό του πληθυσμού γυναικών σε ηλικία εργασίας με προχωρημένη εκπαίδευση)» (lnFLabor) στις «εξαγωγές υψηλής τεχνολογίας» (HTX). Το συμπέρασμα αυτό συμβαδίζει με εκείνο στο οποίο καταλήγουν οι Gaur M. κα.[18] και καταδεικνύει την ανάγκη ενίσχυσης της πρόσβασης στην ανώτερη εκπαίδευση στις γυναίκες.

Τέλος, από τα παραπάνω αποτελέσματα προκύπτει ότι η ψευδομεταβλητή After2016 έχει αρνητική επίδραση (σε επίπεδο σημαντικότητας 1%) στις «εξαγωγές υψηλής τεχνολογίας» (HTX). Το συμπέρασμα αυτό επιβεβαιώνει την αίσθηση που είχαμε αρχικά για την αρνητική επιρροή της χρονικής περιόδου 2016-2018, όπου την περίοδο εκείνη φάνηκαν οι συνέπειες της αρνητικής επίδρασης στις «εξαγωγές υψηλής τεχνολογίας» (HTX).

Στο σημείο αυτό θέλουμε να επισημάνουμε ότι οι εκτιμήσεις των συντελεστών του μοντέλου τυχαίων επιδράσεων είναι πολύ κοντά στις εκτιμήσεις του μοντέλου σταθερών

επιδράσεων. Το αποτέλεσμα αυτό είναι εύλογο και λογικό αφού ο συντελεστής $\lambda=0.9266$ είναι πολύ κοντά στην μονάδα.

ΚΕΦΑΛΑΙΟ 5

5 ΕΠΙΛΟΓΟΣ – ΣΥΜΠΕΡΑΣΜΑΤΑ

Αντικείμενο της παρούσας διπλωματικής εργασίας ήταν η ανάλυση δεδομένων πάνελ με χρήση της γλώσσας προγραμματισμού R. Όπως είδαμε, τα δεδομένα πάνελ παρέχουν πληροφορίες για μεγέθη που αφορούν τόσο σε επιμέρους οντότητες (χώρες, επιχειρήσεις, άτομα, γεωγραφικές περιοχές κλπ) όσο και στην διαχρονική εξέλιξή τους.

Ένα από τα σημαντικότερα πλεονεκτήματα που προσφέρουν τα δεδομένα πάνελ είναι ότι επιτρέπουν τον έλεγχο της ετερογένειας στο δείγμα. Ένας επιπλέον λόγος για το μεγάλο επιστημονικό ενδιαφέρον σχετικά με τα δεδομένα πάνελ είναι ότι με τη χρήση αυτών, μπορούν να αναδειχθούν δυναμικές που θα ήταν δύσκολο να εντοπιστούν με χρήση μόνο διαστρωματικών δεδομένων ή μόνο δεδομένων χρονολογικών σειρών. Ως εκ τούτου, έχουν γίνει πολλές προσπάθειες για τη συλλογή δεδομένων πάνελ μέσω ερευνών μεγάλης κλίμακας και η διαθεσιμότητα τέτοιων δεδομένων πάνελ υψηλής ποιότητας επιτρέπει στους επιστήμονες από διάφορες ειδικότητες να εκτελέσουν σε βάθος εμπειρική ανάλυση. Από την άλλη πλευρά, τα όλο και περισσότερο διαθέσιμα μεγάλα σύνολα δεδομένων για το οικονομικό και το ανθρώπινο κεφάλαιο και την ανάπτυξη, το κοινωνικό κεφάλαιο, το διεθνές εμπόριο, το χρηματιστήριο κ.α., απαιτούν τη γνώση και χρήση εξειδικευμένων υπολογιστικών πακέτων για την ανάλυσή τους.

Η στατιστική γλώσσα R διαθέτει μια σειρά από τις πιο σύγχρονες τεχνικές για τη στατιστική ανάλυση δεδομένων. Η R μπορεί να χειριστεί μεγάλο όγκο και διαφορετικούς τύπους δεδομένων και δίνει τη δυνατότητα δημιουργίας υψηλής ποιότητας γραφικών. Είναι μια γλώσσα ανοικτού κώδικα που ενσωματώνει συνεχώς νέες βιβλιοθήκες (και κατά συνέπεια νέες μεθοδολογίες ανάλυσης) με αποτέλεσμα να εξελίσσεται σε δυνατότητες και εύρος εφαρμογών. Τέλος, υπάρχει ένα τεράστιο δίκτυο εμπειρών χρηστών που μπορεί να παρέχει υψηλού επιπέδου υποστήριξη.

Προκειμένου να υπάρξει μια ολοκληρωμένη εικόνα στην ανάλυσή μας ώστε να εξαχθούν με αξιοπιστία τα βασικά συμπεράσματα η εργασία ξεκίνησε με μια ανασκόπηση της γλώσσας προγραμματισμού R. Στο δεύτερο κεφάλαιο, μελετήσαμε τις βασικές μεθόδους εκτίμησης μοντέλων με δεδομένα πάνελ: το ενιαίο μοντέλο, το μοντέλο σταθερών επιδράσεων και το μοντέλο τυχαίων επιδράσεων καθώς και ελέγχους καταλληλότητας για τα μοντέλα αυτά.

Όπως είδαμε, το μοντέλο σταθερών επιδράσεων είναι αποτελεσματικό όταν τα σφάλματα παρουσιάζουν αυτοσυσχέτιση (αλλά και ομοσκεδαστικότητα) και δεν διατυπώνουμε υποθέσεις για τη συσχέτιση μεταξύ των απαρατήρητων επιδράσεων a_i και των ερμηνευτικών μεταβλητών. Το μοντέλο τυχαίων επιδράσεων είναι κατάλληλο όταν οι απαρατήρητες επιδράσεις δε σχετίζονται με όλες τις ερμηνευτικές μεταβλητές. Στην περίπτωση αυτή, οι απαρατήρητες επιδράσεις a_i μπορούν να συγχωνευθούν στον όρο του σφάλματος και η προκύπτουσα αυτοσυσχέτιση διαχρονικά μπορεί να αντιμετωπισθεί με μια ομαδοποιημένη παλινδρόμηση επί οιονεί προσαρμοσμένων δεδομένων. Η τιμή της εκτιμηθείσας παραμέτρου μετασχηματισμού, $\hat{\lambda}$, δείχνει αν οι εκτιμήσεις είναι κοντά σε αυτές του ενιαίου μοντέλου ή σε αυτές του μοντέλου σταθερών επιδράσεων. Αν ισχύουν οι κατάλληλες προϋποθέσεις, ο εκτιμητής των τυχαίων επιδράσεων είναι ασυμπτωτικά - καθώς το N μεγαλώνει και το T είναι σταθερό- πιο αποτελεσματικός από τις εκτιμήσεις του ενιαίου μοντέλου ή του μοντέλου σταθερών επιδράσεων (όπου οι δύο είναι αμερόληπτες, συνεπείς και αποτελεσματικές).

Στο τρίτο κεφάλαιο, παρουσιάζεται τα βασικά εργαλεία που διαθέτει η R για την ανάλυση δεδομένων πάνελ. Αρχικά, αναλύονται οι εντολές που διαθέτει για την οργάνωση, τη διερευνητική ανάλυση και τον έλεγχο της ετερογένειας των δεδομένων πάνελ. Τέλος παρουσιάζεται αναλυτικά και μέσω παραδειγμάτων το οικονομετρικό πακέτο plm για την εκτίμηση μοντέλων δεδομένων πάνελ.

Τέλος, το τέταρτο κεφάλαιο, αφορά στο εμπειρικό μέρος της εργασίας. Συγκεκριμένα, αφορά στην εμπειρική ανάλυση των προσδιοριστικών παραγόντων που επηρεάζουν τις Εξαγωγές Υψηλής Τεχνολογίας με χρήση δεδομένων πάνελ από 15 χώρες της Ευρωπαϊκής Ένωσης για τα έτη 2007-2018. Οι Εξαγωγές Υψηλής Τεχνολογίας θεωρούνται σημαντικός παράγοντας για τη βιώσιμη οικονομική ανάπτυξη μιας χώρας. Στη μελέτη αυτή θεωρήσαμε ως ερμηνευτικές μεταβλητές τις ακόλουθες α) τις Δαπάνες για Έρευνα και Ανάπτυξη, β) τις Εξαγωγές αγαθών Τεχνολογίας Πληροφορικής και Τηλεπικοινωνιών, γ) τη Συναλλαγματικής Ισοτιμία, δ) το γυναικείο Εργατικό Δυναμικό με εξειδικευμένη εκπαίδευση, στ) τη Χρηματοοικονομική Ανάπτυξη και ε) το Επίπεδο Ανθρώπινης Ανάπτυξης.

Μέσω της πολύτιμης συμβολής του στατιστικού πακέτου, R, η εμπειρική μελέτη έδειξε ότι το Μοντέλο Τυχαίων Επιδράσεων αποδεικνύεται το πλέον κατάλληλο και αξιόπιστο για την εκτίμηση των παραγόντων που επηρεάζουν τις Εξαγωγές Υψηλής Τεχνολογίας σε σύγκριση με το Μοντέλο Σταθερών Επιδράσεων καθώς επίσης και το

Ενιαίο Μοντέλο. Τα αποτελέσματα εκτίμησης του μοντέλου τυχαίων επιδράσεων έδειξαν ότι ο λογάριθμος των «Εξαγωγές αγαθών Τεχνολογίας Πληροφορικής και Τηλεπικοινωνιών ως ποσοστό των συνολικών εξαγωγών» (lnICT) καθώς και ο «Δείκτης Πρόσβασης Χρηματοπιστωτικών Ιδρυμάτων» (FIAI) έχουν θετική επίδραση στις «εξαγωγές υψηλής τεχνολογίας» (HTX) στις 15 χώρες της Ευρωπαϊκής Ένωσης, αναφορικά με τα έτη 2007-2018. Επιπροσθέτως, ένας άλλος παράγοντας που φαίνεται να επιδρά θετικά (σε επίπεδο σημαντικότητας 1%) στις «εξαγωγές υψηλής τεχνολογίας» (HTX) είναι ο «Δείκτης Ανθρώπινης Ανάπτυξης» (HDI). Μάλιστα, ο «Δείκτης Ανθρώπινης Ανάπτυξης» (HDI) έχει τη μεγαλύτερη θετική επίδραση.

Σε αυτό το σημείο, αξίζει να επισημανθεί ότι το αρνητικό πρόσημο του στατιστικά σημαντικού παράγοντα «Δείκτη Πραγματικής Συναλλαγματικής Ισοτιμίας» (lnREER) συμβαδίζει με τη γενική παραδοχή ότι η αύξηση της συναλλαγματικής ισοτιμίας μειώνει τις καθαρές εξαγωγές μακροχρόνια. Επίσης, ενδιαφέρον παρουσιάζει η αρνητική επίδραση (σε επίπεδο σημαντικότητας 5%) του λογάριθμου του «Εργατικού δυναμικού με προχωρημένη εκπαίδευση, γυναίκα (ως ποσοστό του πληθυσμού γυναικών σε ηλικία εργασίας με προχωρημένη εκπαίδευση)» (lnFLabor) στις εξαγωγές υψηλής τεχνολογίας» (HTX). Το συμπέρασμα αυτό συμβαδίζει με εκείνο στο οποίο καταλήγουν οι (Gaur M. κα. 2020) και καταδεικνύει την ανάγκη ενίσχυσης της πρόσβασης στην ανώτερη εκπαίδευση στις γυναίκες.

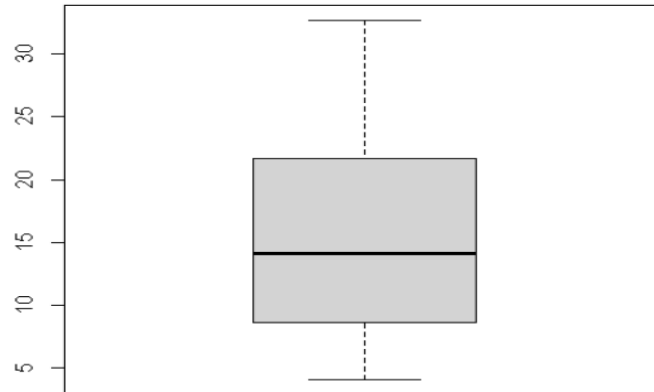
Τελειώνοντας, θα θέλαμε να δώσουμε κάποιες προτάσεις για μελλοντική έρευνα. Η παρούσα εργασία επιχείρησε να παρουσιάσει τις σημαντικότερες μεθόδους ανάλυσης δεδομένων πάνελ. Περαιτέρω θα μπορούσαμε να επεκτείνουμε την ανάλυση σε πιο προχωρημένες μεθόδους ανάλυσης δεδομένων πάνελ όπως π.χ. σε προχωρημένη ανάλυση διακύμανσης, σε μεθόδους συνολοκλήρωσης (cointegration analysis), σε μη γραμμική παλινδρόμηση κλπ. Επιπλέον, θα μπορούσαμε να διερευνήσουμε τις δυνατότητες που δίνει η R για τις αντίστοιχες αναλύσεις αλλά και κάποια άλλη γλώσσα προγραμματισμού, όπως η Python.

Με βάση τα αποτελέσματα της εμπειρικής ανάλυσης, θα μπορούσαμε να δώσουμε προτάσεις για περαιτέρω έρευνα σχετικά με τους παράγοντες που επηρεάζουν τις εξαγωγές υψηλής τεχνολογίας. Κατ' αρχάς, θα μπορούσε να διεξαχθεί η ίδια ανάλυση, με τις ίδιες μεταβλητές (εξαρτημένη και ανεξάρτητη μεταβλητή) και με τις ίδιες χώρες για περισσότερα έτη. Ο έλεγχος αυτός θα προσδώσει πιο ξεκάθαρα αποτελέσματα σχετικά με τη σχέση των εξαγωγών υψηλής τεχνολογίας και των αντίστοιχων προσδιοριστικών

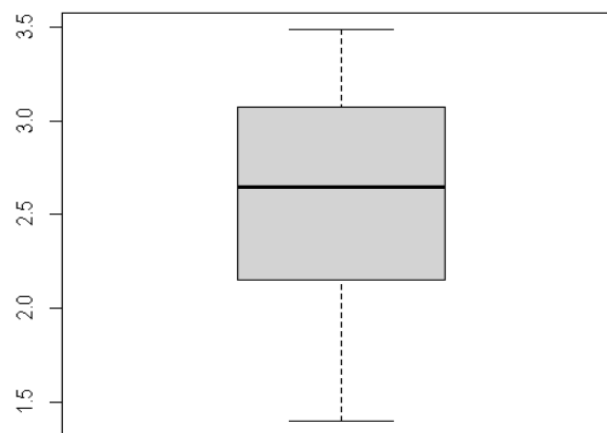
παραγόντων. Μάλιστα με αυτή τη διευρυμένη ανάλυση θα διαλευκανθούν και το ακριβές πρόσημο της μεταβλητής «Δαπάνες για Έρευνα και Ανάπτυξη», η οποία ερχόταν σε αντίθεση με τα αποτελέσματα της διεθνούς βιβλιογραφία (Gaur M. κα. 2020). Επιπλέον, η παρούσα έρευνα θα μπορούσε να επεκταθεί και για το σύνολο των κρατών μελών της ΕΕ ή θα μπορούσε να γίνει διάκριση και σε χώρες εντός και εκτός Ευρωζώνης για να διαπιστωθεί αν οι Εξαγωγές Υψηλής Τεχνολογίας επηρεάζονται από αυτόν τον παράγοντα. Ακόμη, το υπόδειγμά μας θα μπορούσε να εμπλουτισθεί και με άλλες εξηγηματικές μεταβλητές (εφόσον δύναται να βρεθούν δεδομένα) όπως π.χ. με τον δείκτη οικονομικής ελευθερίας, τις εφαρμογές πατέντων κ.α., ώστε το υπόδειγμα να εξηγεί μεγαλύτερο ποσοστό της μεταβλητότητας της εξαρτημένης μεταβλητής.

ΠΑΡΑΡΤΗΜΑ

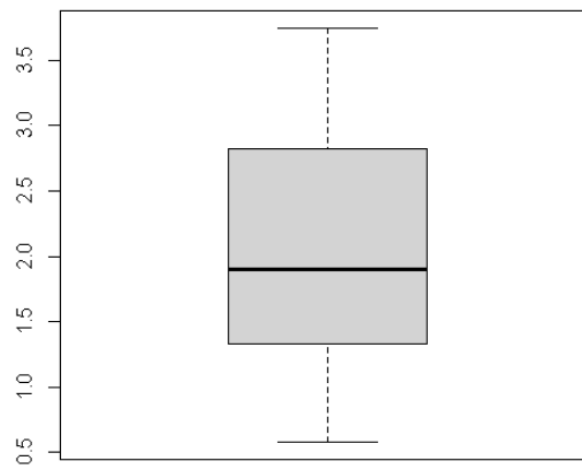
6 ΠΑΡΑΡΤΗΜΑ ΑΠΟΤΕΛΕΣΜΑΤΩΝ



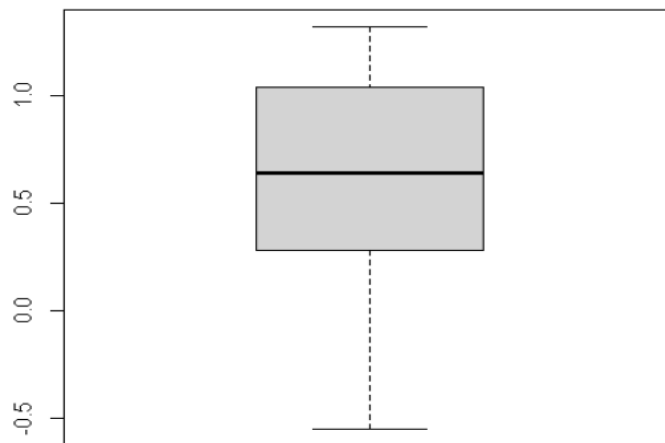
Διάγραμμα 6.1:Θηκόγραμμα εξαγωγών υψηλής τεχνολογίας



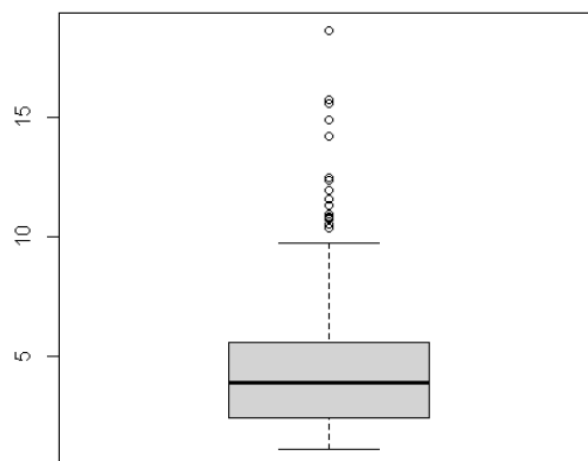
Διάγραμμα 6.2:Θηκόγραμμα φυσικού λογαρίθμου εξαγωγών υψηλής τεχνολογίας



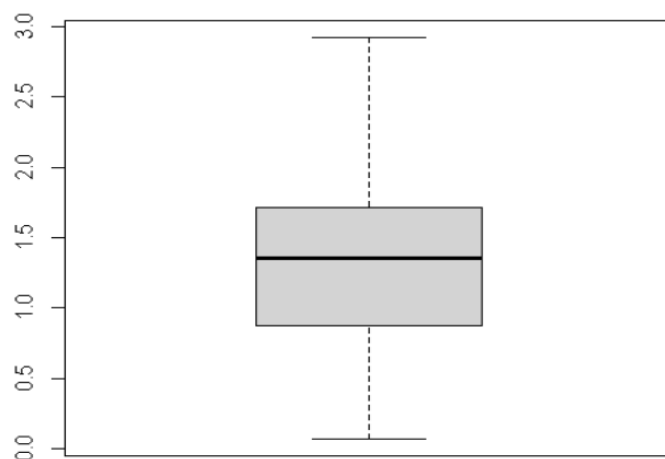
Διάγραμμα 6.3:Θηκόγραμμα Δαπανών για Έρευνα και Ανάπτυξη



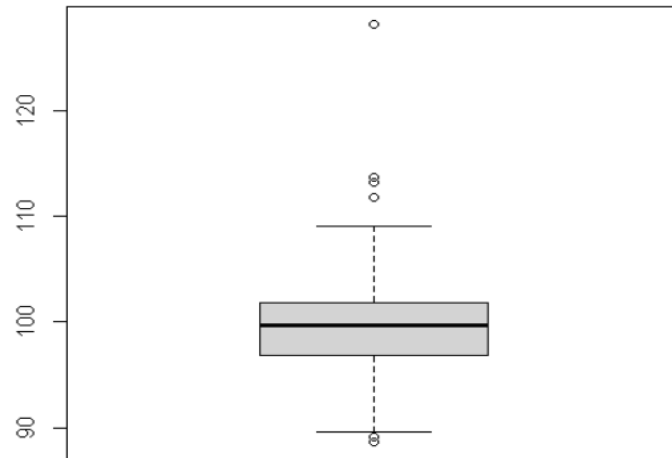
Διάγραμμα 6.4: Θηκόγραμμα φυσικού λογαρίθμου Δαπανών για Έρευνα και Ανάπτυξη



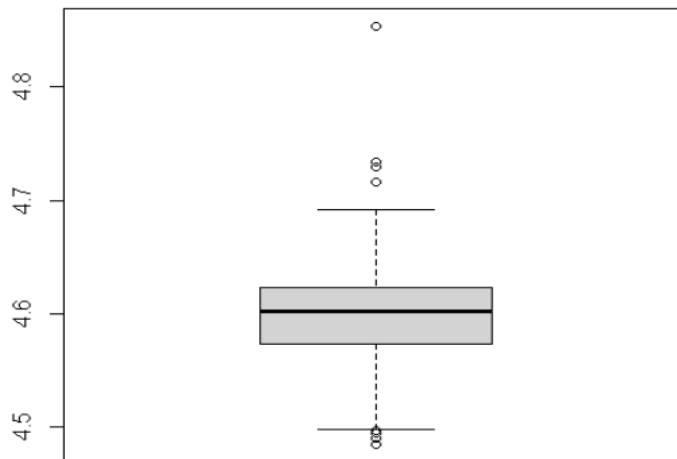
Διάγραμμα 6.5: Θηκόγραμμα εξαγωγής αγαθών Τεχνολογίας Πληροφορικής και Τηλεπικοινωνιών



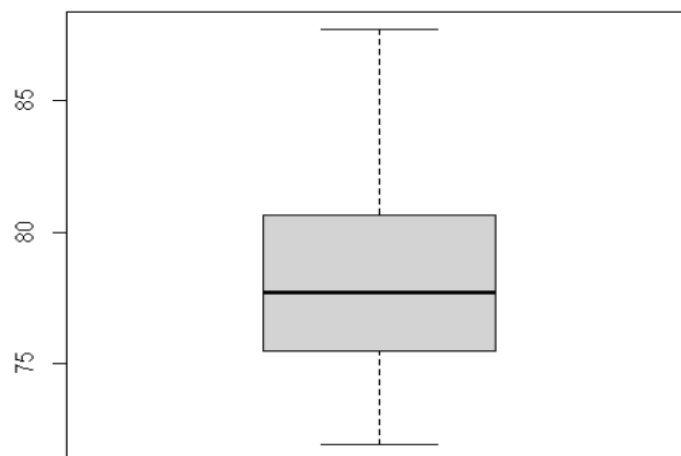
Διάγραμμα 6.6: Θηκόγραμμα φυσικού λογαρίθμου εξαγωγής αγαθών Τεχνολογίας Πληροφορικής και Τηλεπικοινωνιών



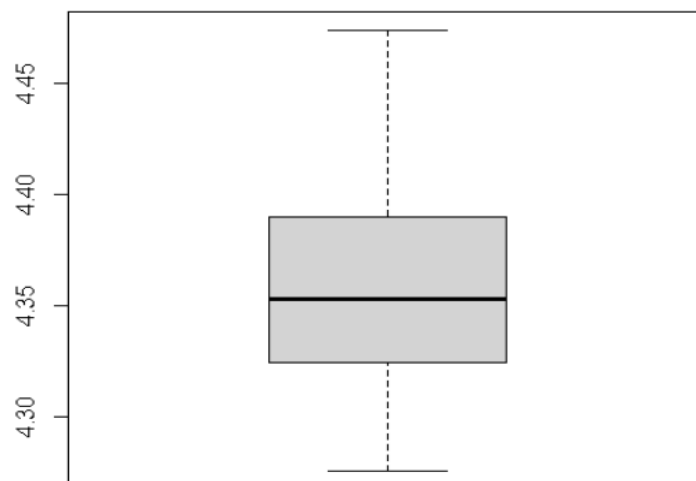
Διάγραμμα 6.7: Θηκόγραμμα Δείκτη Πραγματικής Συναλλαγματικής



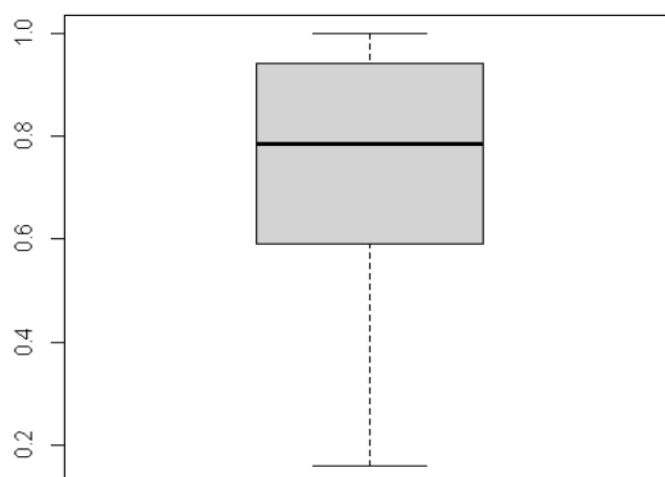
Διάγραμμα 6.8: Θηκόγραμμα λογαρίθμου Δείκτη Πραγματικής Συναλλαγματικής



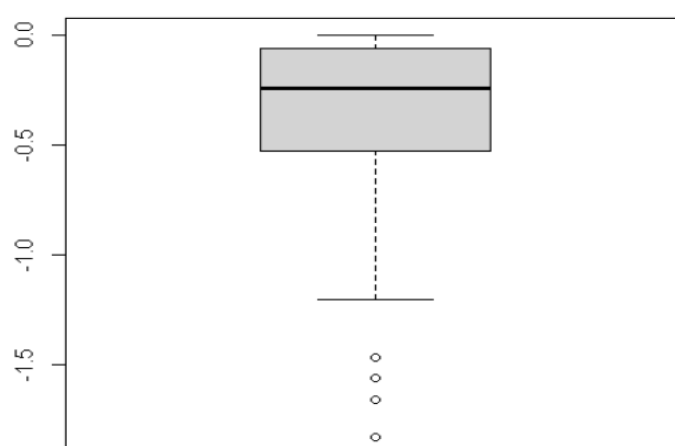
Διάγραμμα 6.9: Θηκόγραμμα Εργατικού δυναμικού με προχωρημένη εκπαίδευση, γυναίκα



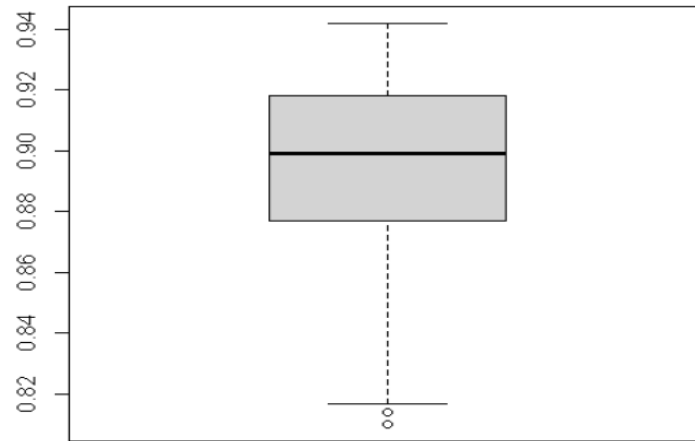
Διάγραμμα 6.10: Θηκόγραμμα φυσικού λογαρίθμου Εργατικού δυναμικού με προχωρημένη εκπαίδευση, γυναίκα



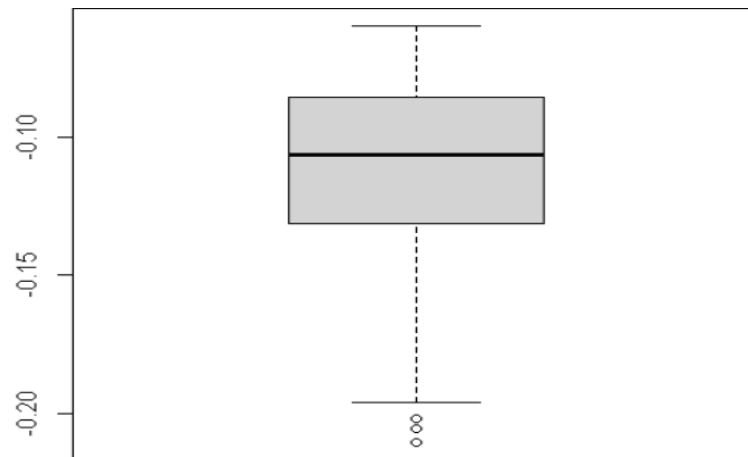
Διάγραμμα 6.11: Θηκόγραμμα Δείκτη Πρόσβασης Χρηματοπιστωτικών Ιδρυμάτων



Διάγραμμα 6.12: Θηκόγραμμα φυσικού λογαρίθμου Δείκτη Πρόσβασης Χρηματοπιστωτικών Ιδρυμάτων



Διάγραμμα 6.13: Θηκόγραμμα Δείκτη Ανθρώπινης Ανάπτυξης



Διάγραμμα 6.14: Θηκόγραμμα φυσικού λογαρίθμου Δείκτη Ανθρώπινης Ανάπτυξης

ΒΙΒΛΙΟΓΡΑΦΙΑ

1. Αντζουλάκος, Δ. (2013). *Ανάλυση Δεδομένων με τη Χρήση Στατιστικών Πακέτων Εισαγωγή στο R*.Σημειώσεις παραδόσεων του Τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς.
2. Βασιλόπουλος, Γ. (2016).*Μία εμπειρική διερεύνηση της επίδρασης των φόρων στην οικονομική μεγέθυνση για τις χώρες του ΟΟΣΑ*(Master' sthesis, Πανεπιστήμιο Πειραιώς).
3. Βερύκιος, Β., Καγκλής, Β., & Σταυρόπουλος, Η. (2015). *Η επιστήμη των δεδομένων μέσα από τη γλώσσα R*.
4. Βουβαλά, Κ. (2012). *Labor force participation*.(Μεταπτυχιακή εργασία). Ανακτήθηκε από Θεσσαλονίκη: Πανεπιστήμιο Μακεδονίας Οικονομικών και Κοινωνικών Επιστημών Τμήμα Οικονομικών Επιστημών: Διατμηματικό Πρόγραμμα Μεταπτυχιακών Σπουδών.
<https://dspace.lib.uom.gr/bitstream/2159/15716/3/VouvalaKonstantiniaMsc2012.pdf>
5. Μάττα, Ε. Κ. (2015).*Stress Testing Τραπεζικής Κερδοφορίας: Μια Ευρωπαϊκή Προσέγγιση*(No. GRI-2015-15052). Aristotle University of Thessaloniki.
6. Νομικού, Ε. (2012).*Παράγοντες επίδρασης των άμεσων ξένων επενδύσεων*(Doctoral dissertation).
7. Ντάνης, Δ. Α. (2016). *Σύγχρονα στατιστικά λογισμικά και οι εφαρμογές τους*.(Πτυχιακή εργασία). Ανακτήθηκε από Αντίρριο: Τεχνολογικό Εκπαιδευτικό Ίδρυμα Δυτικής Ελλάδας Τμήμα Μηχανικών Πληροφορικής Τ.Ε.<http://repository.library.teimes.gr/xmlui/handle/123456789/3150>
8. Ντζούφρας, Ι., Καρλής, Δ., 2015. *Εισαγωγή στον προγραμματισμό και στη στατιστική ανάλυση με R*. [ηλεκτρ. βιβλ.] Αθήνα: Σύνδεσμος Ελληνικών Ακαδημαϊκών Βιβλιοθηκών. Διαθέσιμο στο: <http://hdl.handle.net/11419/2601>
9. Ξανθόπουλος, Π. (2014).*Μελέτη και εκτίμηση υποδειγμάτων με τη μέθοδο διαφορών σε διαφορές*(Master' sthesis, Πανεπιστήμιο Πειραιώς).

10. ΣΗΜΕΙΩΣΕΙΣ, Φ. Χ. Ρ. Π. ΑΕΙ: ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ΤΜΗΜΑ: ΟΙΚΟΝΟΜΙΚΗΣ ΕΠΙΣΤΗΜΗΣ ΗΜΕΡΟΜΗΝΙΑ ΓΣ ΤΜΗΜΑΤΟΣ:
17/06/2015.
11. Σοφρά, Ρ. (2014). Γλώσσα προγραμματισμού R. Εργαλείο εκπαιδευτικής έρευνας.
Συγκριτική μελέτη SPSS με την R.
12. Φουσκάκης, Δ. (2014). Ανάλυση Δεδομένων με χρήση του Στατιστικού Πακέτου
R. Σχολή Εφαρμοσμένων Μαθηματικών και Φυσικών Εφαρμογών, Εθνικό Μετσόβιο
Πολυτεχνείο.
13. Φωκιανός, Κ., Χαράλαμπος, Χ., Μαθηματικών, Τ., & Κύπρου, Σ. Π. (2010).
Εισαγωγή στην R Πρόχειρες Σημειώσεις. Τμήμα μαθηματικών & Στατιστικής
Πανεπιστήμιο Κύπρου.
14. Ψαρομμάτη, Β. (2016). Βιοτικό επίπεδο, εκβιομηχάνιση και απελευθέρωση εμπορίου
στην Νότιο Αφρική: μια οικονομετρική ανάλυση.
15. Baltagi, B.H. (2001), *Econometric Analysis of Panel Data*, Second edition, New York:
Wiley.
16. Baltagi, B. H., & Levin, D. (1992). Cigarette taxation: Raising revenues and reducing
consumption. *Structural Change and Economic Dynamics*, 3(2), 321-335.
17. Gökmen, Y., & Turen, U. (2013). The determinants of high technology exports volume:
A panel data analysis of EU-15 countries, *International Journal of Management,
Economics and Social Sciences*, 2(3), 217-232.
18. Gaur, M., Kant, P. and Verma, N. M. P (2020). Macro-Economic Determinants of High
Technology Exports, *EJBMR, European Journal of Business and Management
Research*, 5 (5), DOI: <http://dx.doi.org/10.24018/ejbmr.2020.5.5.506>
19. Croissant, Y., & Millo, G. (2007). Introduction to plm. cran. r-- project. org.
20. Hsiao C (2003) *Analysis of panel data*, 2nd edn. *Econometric society monographs*, vol
34. Cambridge University Press, Cambridge
21. Ani Katchova (2013). *Panel Data Models Example*.
22. Kunst, R. M. (2010). *Econometric methods for panel data*. Institute for Advanced
Studies Vienna.

23. Moulton, B. R. (1987). *Diagnostics for group effects in regression analysis*. *Journal of Business & Economic Statistics*, 5(2), 275-282.
24. Torres-Reyna, O. (2007). *Panel data analysis fixed and random effects using Stata (v. 4.2)*. Data & Statistical Services, Princeton University, 1-40.
25. Torres-Reyna, O. (2010). *Getting Started in Fixed/Random Effects Models using R (ver. 0.1-Draft)*.
26. Wooldridge, J. M. (2002). *Econometric analysis of cross section and panel data* MIT press. Cambridge, MA, 108.
27. Wooldridge, J. M. (2009). *Advanced panel data methods. Introductory econometrics: a modern approach*. Mason: South-Western Cengage Learning, 489.
28. Wooldridge, J. M. (2016). *Introductory econometrics: A modern approach*. Nelson Education
29. https://rstudio-pubsstatic.s3.amazonaws.com/372492_3e05f38dd3f248e89cdedd317d603b9a.html
30. <https://cran.r-project.org/web/packages/plm/vignettes/plmPackage.html>
31. <https://bookdown.org/ccolonescu/RPoE4/panel-data-models.html>
32. <https://databank.worldbank.org/home.aspx>