ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
**Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών**
ΙΔΡΥΘΕΝ ΤΟ 1837

**Γραμμικά Μοντέλα**

**Αθήνα 2024**

# **Δείκτες Παγκόσμιας Ανάπτυξης** - Μια Στατιστική Ανάλυση για CorControl

**Ιωάννης Τσαγκαρόπουλος**

**Α.Μ.: 1112 2021 00180**

**Επιβλέπουσα: Λ. Μελιγκοτσίδου**

## Abstract

Η εργασία αυτή μελετάει τους δείκτες παγκόσμιας ανάπτυξης, και συγκεκριμένα Έλεγχος Διαφθοράς (Corruption Control – CorControl). Χρησιμοποιήθηκε, ανάλυση πολλαπλής παλινδρόμησης για να διευκρινιστεί και να προβλευθεί ο έλεγχος της διαφθοράς με βάση διάφορους παγκόσμιους δείκτες ανάπτυξης. Η έρευνα επιλέγει μεθοδολογικά το βέλτιστο μοντέλο χρησιμοποιώντας το Κριτήριο Πληροφοριών Akaike (AIC) και χρησιμοποιεί την αντίστροφη βηματική επιλογή. Το μοντέλο παλινδρόμησης που προκύπτει ενσωματώνει διάφορους προγνωστικούς παράγοντες που αντιπροσωπεύουν διαφορετικούς δείκτες παγκόσμιας ανάπτυξης, με συντελεστές που εκτιμώνται μέσω των συνήθων ελαχίστων τετραγώνων. Οι παραδοχές της γραμμικότητας, της ανεξαρτησίας, της ομοσκεδαστικότητας και της κανονικότητας των καταλοίπων ελέγχθηκαν αυστηρά για να επικυρωθεί η αξιοπιστία του μοντέλου. Η συνολική σημαντικότητα του μοντέλου αξιολογήθηκε με τη χρήση ενός F-test, ενώ η σημαντικότητα των επιμέρους προβλεπτικών παραγόντων αξιολογήθηκε με τη χρήση t-tests για τους συντελεστές παλινδρόμησης.

Επιπλέον, η μελέτη αξιολογεί τις διαφορές στην Πολιτική Σταθερότητας μεταξύ των κατηγορικών προβλεπτών "Κατηγορία γονιμότητας" (FertCat) και "Κατηγορία πληθωρισμού" (InflCat), που ορίζονται στην Ενότητα 1, χρησιμοποιώντας ανάλυση διακύμανσης (ANOVA). Η ANOVA διακρίνει εάν υπάρχουν σημαντικές διαφορές στους μέσους όρους της Πολιτικής Σταθερότητας μεταξύ των ομάδων που ορίζονται από κάθε κατηγορικό προγνωστικό παράγοντα. Η μεταβλητότητα της Πολιτικής Σταθερότητας αναλύεται με την κατάτμησή της σε συνιστώσες εντός και μεταξύ των ομάδων. Η συνάρτηση ελέγχου F, που υπολογίζεται ως ο λόγος του μέσου τετραγώνου μεταξύ των ομάδων προς το μέσο τετράγωνο εντός των ομάδων, ελέγχει τη μηδενική υπόθεση των ίσων μέσων όρων των ομάδων. Ένας σημαντικός έλεγχος F υποδηλώνει διαφορές μεταξύ των μέσων όρων των ομάδων, γεγονός που σημαίνει σημαντική επίδραση του κατηγορικού προβλεπτικού παράγοντα στην πολιτική σταθερότητα. Για τον εντοπισμό συγκεκριμένων ομαδικών διαφορών διενεργήθηκαν post-hoc δοκιμές Honestly Significant Difference (HSD) του Tukey.



## Λέξεις κλειδιά

Λέξεις-Κλειδιά (Keywords): Δείκτες Παγκόσμιας Ανάπτυξης, Γραμμική παλινδρόμηση, Επιλογή κατάλληλου Μοντέλου, Έλεγχος Διαφθοράς, Κανονικό Πολλαπλό Γραμμικό Μοντέλο.

# Περιεχόμενα

# 1 Εισαγωγή

Οι Δείκτες Παγκόσμιας Ανάπτυξης (World Development Indicators - WDIs) αποτελούν μία πηγή πληροφοριών ζωτικής σημασίας για την παρακολούθηση της προόδου μίας χώρας, τον εντοπισμό τάσεων, και τη διαμόρφωση πολιτικών. Οι WDIs περιλαμβάνουν ένα ευρύ φάσμα κοινωνικοοικονομικών δεικτών που εκτείνονται σε διάφορους τομείς όπως η εκπαίδευση, η υγεία, το περιβάλλον, η οικονομία και η διακυβέρνηση, παρέχοντας μια συνολική εικόνα της παγκόσμιας αναπτυξιακής δυναμικής.

Σε αυτή τη μελέτη θα χρησιμοποιήσουμε τη βάση δεδομένων WDI της Παγκόσμιας Τράπεζας . Συγκεκριμένα, δίνεται ένα υποσύνολο των δεδομένων (data_tidy.csv), το οποίο αποτελείται από 27 μεταβλητές για 120 χώρες, το έτος 2018.

### Δομή Εργασίας

Το project αξιοποιεί:

- data_tidy.csv: Το κύριο αρχείο δεδομένων.
- data_new.csv: Ένα αρχείο με 8 ακόμα χώρες που θα χρειαστείτε στην Ενότητα 2, Ερώτημα 8.
- metadata.pdf: Ένα αρχείο με πληροφορίες για τα δεδομένα, που θα χρειαστείτε στην Ενότητα 1.

και τα παραδοτέα της άσκησης είναι:

- analysis.R: Ένα αρχείο κώδικα στο οποίο θα υλοποιηθεί η ανάλυση σύμφωνα με τα ερωτήματα της εργασίας.
- Η παρούσα αναφορά (report) με την παρουσίαση των αποτελεσμάτων σε μορφή pdf.

# 2 Δεδομένα-Περιγραφική στατιστική

Σε αυτή τη μελέτη θα χρησιμοποιήσουμε τη βάση δεδομένων WDI της Παγκόσμιας Τράπεζας . Συγκεκριμένα, δίνεται ένα υποσύνολο των δεδομένων (data_tidy.csv), το οποίο αποτελείται από 27 μεταβλητές για 120 χώρες, το έτος 2018. Οι μεταβλητές που περιλαμβάνει το αρχείο αυτό είναι οι εξής:

## 2.1 Metadata

1. **ElectrAccess** (Πρόσβαση στην ηλεκτρική ενέργεια): Ποσοστό του πληθυσμού που έχει πρόσβαση στην ηλεκτρική ενέργεια. Τα στοιχεία για τον εξηλεκτρισμό συλλέγονται από τη βιομηχανία, εθνικές έρευνες και διεθνείς πηγές.

2. **CookAccess** (Πρόσβαση σε καθαρά καύσιμα και τεχνολογίες για το μαγείρεμα, % του πληθυσμού): Μετρά το ποσοστό του πληθυσμού που χρησιμοποιεί κυρίως καθαρά καύσιμα και τεχνολογίες μαγειρέματος, εξαιρουμένης της κηροζίνης σύμφωνα με τις κατευθυντήριες γραμμές του WHO.

3. **AgriLand** (Γεωργική γη, % της έκτασης της γης): Αναφέρεται στο μερίδιο της έκτασης της γης που είναι καλλιεργήσιμη, υπό μόνιμες καλλιέργειες και υπό μόνιμους βοσκότοπους. Η καλλιεργήσιμη γη περιλαμβάνει τη γη που ορίζεται από τον FAO ως γη υπό προσωρινές καλλιέργειες (οι εκτάσεις με διπλή καλλιέργεια υπολογίζονται μία φορά), τα προσωρινά λιβάδια για κούρεμα ή για βοσκότοπο, τη γη υπό λαχανόκηπους ή λαχανόκηπους και τη γη που βρίσκεται προσωρινά σε αγρανάπαυση. Εξαιρούνται οι εκτάσεις που έχουν εγκαταλειφθεί λόγω της μετακινούμενης καλλιέργειας. Οι εκτάσεις υπό μόνιμες καλλιέργειες είναι εκτάσεις που καλλιεργούνται με καλλιέργειες που καταλαμβάνουν τη γη για μεγάλα χρονικά διαστήματα και δεν χρειάζεται να ξαναφυτεύονται μετά από κάθε συγκομιδή, όπως το κακάο, ο καφές και το καουτσούκ. Η κατηγορία αυτή περιλαμβάνει τη γη κάτω από ανθοφόρους θάμνους, οπωροφόρα δέντρα, καρυδιές και αμπέλια, αλλά δεν περιλαμβάνει τη γη κάτω από δέντρα που καλλιεργούνται για ξύλο ή ξυλεία.Μόνιμος βοσκότοπος είναι η γη που χρησιμοποιείται για πέντε ή περισσότερα χρόνια για ζωοτροφές, συμπεριλαμβανομένων των φυσικών και των καλλιεργούμενων καλλιεργειών.

4. **BirthRate** (Ρυθμός γεννήσεων, ακαθάριστος, ανά 1.000 άτομα): Ο ακαθάριστος δείκτης γεννήσεων αναφέρεται στον αριθμό των γεννήσεων ζώντων που συμβαίνουν κατά τη διάρκεια του έτους, ανά 1.000 κατοίκους που υπολογίζονται στα μέσα του έτους. Η αφαίρεση του ακατέργαστου ποσοστού θανάτων από το ακατέργαστο ποσοστό γεννήσεων παρέχει το ποσοστό φυσικής αύξησης, το οποίο ισούται με το ποσοστό μεταβολής του πληθυσμού ελλείψει μετανάστευσης.

5. **CO2** (Εκπομπές CO2, μετρικοί τόνοι ανά κάτοικο): Περιλαμβάνει τις εκπομπές διοξειδίου του άνθρακα από την καύση ορυκτών καυσίμων και την κατασκευή τσιμέντου, συμπεριλαμβανομένης κάθε κατανάλωσης στερεών, υγρών και αερίων καυσίμων και της καύσης αερίων.

6. **CompEdu** (Υποχρεωτική εκπαίδευση, διάρκεια, έτη): Υποχρεωτική σχολική φοίτηση (υποχρεωτική σχολική εκπαίδευση): Υποδεικνύει τον αριθμό των ετών που τα παιδιά είναι υποχρεωμένα από το νόμο να φοιτούν στο σχολείο.

7. **DeathRate** (Ποσοστό θνησιμότητας, ακατέργαστο, ανά 1.000 άτομα): Ο ακαθάριστος δείκτης θνησιμότητας αναφέρεται τον αριθμό των θανάτων κατά τη διάρκεια του έτους, ανά 1.000 κατοίκους που εκτιμάται στα μέσα του έτους. Η αφαίρεση του ακατέργαστου ποσοστού θανάτων από το ακατέργαστο ποσοστό γεννήσεων παρέχει το ποσοστό φυσικής αύξησης, το οποίο ισούται με το ποσοστό μεταβολής του πληθυσμού ελλείψει μετανάστευσης.

8. **FoodExports** (Εξαγωγές τροφίμων, % των εξαγωγών εμπορευμάτων): Περιλαμβάνει τα εμπορεύματα των τμημάτων SITC 0 (τρόφιμα και ζώντα ζώα), 1 (ποτά και καπνός) και 4 (ζωικά και φυτικά έλαια και λίπη) και του τμήματος SITC 22 (ελαιούχοι σπόροι, ελαιούχοι καρποί και ελαιοπυρήνες).

9. **Telephone** (συνδρομές σταθερής τηλεφωνίας, ανά 100 άτομα): Αναφέρονται στο άθροισμα του ενεργού αριθμού των αναλογικών σταθερών τηλεφωνικών γραμμών, των συνδρομών voice-over-IP (VoIP), των συνδρομών σταθερού ασύρματου τοπικού βρόχου (WLL), των ισοδύναμων φωνητικών καναλιών ISDN και των σταθερών δημόσιων καρτοτηλεφώνων.

10. **Internet** (άτομα που χρησιμοποιούν το Διαδίκτυο, % του πληθυσμού): Οι χρήστες του Διαδικτύου είναι άτομα που χρησιμοποίησαν το Διαδίκτυο (από οποιαδήποτε τοποθεσία) τους τελευταίους 3 μήνες. Το Διαδίκτυο μπορεί να χρησιμοποιηθεί μέσω υπολογιστή, κινητού τηλεφώνου, προσωπικού ψηφιακού βοηθού, παιχνιδομηχανής, ψηφιακής τηλεόρασης κ.λπ.

11. **PopGrowth** (Αύξηση του πληθυσμού, ετήσιο %): Ο ετήσιος ρυθμός αύξησης του πληθυσμού για το έτος $t$ είναι ο εκθετικός ρυθμός αύξησης του πληθυσμού στα μέσα του έτους από το έτος $t-1$ έως το $t$, εκφρασμένος ως ποσοστό . Ο πληθυσμός βασίζεται στον de facto ορισμό του πληθυσμού, ο οποίος καταμετρά όλους τους κατοίκους ανεξάρτητα από το νομικό καθεστώς ή την ιθαγένεια.

12. **BusinessTime** (Χρόνος που απαιτείται για την έναρξη μιας επιχείρησης, ημέρες): Πρόκειται για τον αριθμό των ημερολογιακών ημερών που απαιτούνται για την ολοκλήρωση των διαδικασιών για τη νόμιμη λειτουργία μιας επιχείρησης.

13. **AdolFertRate** (Ποσοστό γονιμότητας εφήβων, γεννήσεις ανά 1.000 γυναίκες ηλικίας 15-19 ετών): Αυτό μετρά τον αριθμό των γεννήσεων ανά 1.000 γυναίκες ηλικίας 15-19 ετών.

14. **Βροχόπτωση** (Μέση βροχόπτωση σε βάθος, mm ανά έτος): Η μέση βροχόπτωση είναι ο μακροχρόνιος μέσος όρος σε βάθος (στο χώρο και στο χρόνο) της ετήσιας βροχόπτωσης στη χώρα. Ως βροχόπτωση ορίζεται κάθε είδος νερού που πέφτει από τα σύννεφα ως υγρό ή στερεό.

15. **CorControl** (Έλεγχος της διαφθοράς: εκτίμηση): Ο έλεγχος της διαφθοράς καταγράφει τις αντιλήψεις σχετικά με τον βαθμό στον οποίο η δημόσια εξουσία ασκείται για ιδιωτικό όφελος, συμπεριλαμβανομένων τόσο των μικροδιαφθορών όσο και των μεγάλων μορφών διαφθοράς, καθώς και της "άλωσης" του κράτους από τις ελίτ και τα ιδιωτικά συμφέροντα. Η εκτίμηση δίνει τη βαθμολογία της χώρας στον συνολικό δείκτη.

16. **EmployerPerc** (Εργοδότες, σύνολο, % της συνολικής απασχόλησης, μοντελοποιημένη εκτίμηση της ILO): Εργοδότες είναι οι εργαζόμενοι που, εργαζόμενοι για δικό τους λογαριασμό ή με έναν ή λίγους συνεργάτες, κατέχουν το είδος των θέσεων εργασίας που ορίζονται ως αυτοαπασχόληση", δηλαδή θέσεις εργασίας στις οποίες η αμοιβή εξαρτάται άμεσα από τα κέρδη που προέρχονται από τα παραγόμενα αγαθά και υπηρεσίες)

17. **FertRate** (Συνολικό ποσοστό γονιμότητας (γεννήσεις ανά γυναίκα)): Αντιπροσωπεύει τον αριθμό των παιδιών που θα γεννιόταν από μια γυναίκα εάν ζούσε μέχρι το τέλος της αναπαραγωγικής της ηλικίας και γεννούσε παιδιά σύμφωνα με τους ειδικούς κατά ηλικία δείκτες γονιμότητας του συγκεκριμένου έτους.

18. **GDPperc** (αύξηση του κατά κεφαλήν ΑΕΠ, ετήσια %): Ετήσιος ποσοστιαίος ρυθμός αύξησης του κατά κεφαλήν ΑΕΠ με βάση το σταθερό τοπικό νόμισμα. Το κατά κεφαλήν ΑΕΠ είναι το ακαθάριστο εγχώριο προϊόν διαιρούμενο με τον πληθυσμό στο μέσο του έτους. Το ΑΕΠ σε τιμές αγοραστή είναι το άθροισμα της ακαθάριστης προστιθέμενης αξίας όλων των κατοίκων παραγωγών της οικονομίας συν τους φόρους επί των προϊόντων και μείον τις επιδοτήσεις που δεν περιλαμβάνονται στην αξία των προϊόντων. Υπολογίζεται χωρίς αφαιρέσεις για την απόσβεση των κατασκευασμένων περιουσιακών στοιχείων ή για την εξάντληση και υποβάθμιση των φυσικών πόρων.

19. **GDPdollars** (κατά κεφαλήν ΑΕΠ, τρέχοντα δολάρια ΗΠΑ): Το κατά κεφαλήν ΑΕΠ είναι το ακαθάριστο εγχώριο προϊόν διαιρούμενο με τον πληθυσμό στο μέσο του έτους. Το ΑΕΠ είναι το άθροισμα της ακαθάριστης προστιθέμενης αξίας όλων των κατοίκων παραγωγών της οικονομίας συν τους φόρους επί των προϊόντων και μείον τις επιδοτήσεις που δεν περιλαμβάνονται στην αξία

των προϊόντων. Υπολογίζεται χωρίς να γίνονται αφαιρέσεις για την απόσβεση των κατασκευασμένων περιουσιακών στοιχείων ή για την εξάντληση και την υποβάθμιση των φυσικών πόρων. Τα στοιχεία είναι σε τρέχοντα δολάρια ΗΠΑ.

20. **Inflation** (Πληθωρισμός, τιμές καταναλωτή, ετήσιος %):Ο πληθωρισμός, όπως μετράται από τον δείκτη τιμών καταναλωτή, αντικατοπτρίζει την ετήσια ποσοστιαία μεταβολή του κόστους για τον μέσο καταναλωτή της απόκτησης ενός καλαθιού αγαθών και υπηρεσιών που μπορεί να είναι σταθερό ή να μεταβάλλεται σε συγκεκριμένα χρονικά διαστήματα, όπως το έτος. Γενικά χρησιμοποιείται ο τύπος Laspeyres.

21. **WaterStress** (Επίπεδο υδατικού στρες: απόληψη γλυκού νερού ως ποσοστό των διαθέσιμων πόρων γλυκού νερού): Το επίπεδο υδατικής πίεσης: η απόληψη γλυκού νερού ως ποσοστό των διαθέσιμων πόρων γλυκού νερού είναι ο λόγος μεταξύ του συνολικού γλυκού νερού που αποσύρεται από όλους τους κύριους τομείς και των συνολικών ανανεώσιμων πόρων γλυκού νερού, αφού ληφθούν υπόψη οι περιβαλλοντικές απαιτήσεις σε νερό. Οι κύριοι τομείς, όπως ορίζονται από τα πρότυπα ISIC, περιλαμβάνουν τη γεωργία, τη δασοκομία και την αλιεία, τη μεταποίηση, τη βιομηχανία ηλεκτρικής ενέργειας και τις υπηρεσίες. Ο δείκτης αυτός είναι επίσης γνωστός ως ένταση απόληψης νερού.

22. **LifeExp** (Προσδόκιμο ζωής κατά τη γέννηση, συνολικά, έτη): Το προσδόκιμο ζωής κατά τη γέννηση υποδηλώνει τον αριθμό των ετών που θα ζούσε ένα νεογέννητο βρέφος εάν τα επικρατούντα πρότυπα θνησιμότητας κατά τη στιγμή της γέννησής του παρέμεναν τα ίδια καθ' όλη τη διάρκεια της ζωής του.

23. **MortRate** (ποσοστό θνησιμότητας, βρεφική θνησιμότητα, ανά 1.000 γεννήσεις ζώντων):Ο δείκτης βρεφικής θνησιμότητας είναι ο αριθμός των βρεφών που πεθαίνουν πριν από τη συμπλήρωση ενός έτους, ανά 1.000 γεννήσεις ζώντων σε ένα δεδομένο έτος.

24. **PolStab** (Πολιτική σταθερότητα και απουσία βίας/τρομοκρατίας: εκτίμηση): Η πολιτική σταθερότητα και η απουσία βίας/τρομοκρατίας μετρά τις αντιλήψεις σχετικά με την πιθανότητα πολιτικής αστάθειας και/ή βίας με πολιτικά κίνητρα, συμπεριλαμβανομένης της τρομοκρατίας. Η εκτίμηση δίνει τη βαθμολογία της χώρας στο συνολικό δείκτη, σε μονάδες μιας τυπικής κανονικής κατανομής, δηλαδή κυμαίνεται περίπου από -2,5 έως 2,5.

25. **PopPerc14** (Πληθυσμός ηλικίας 0-14 ετών, % του συνολικού πληθυσμού): Πληθυσμός ηλικίας 0 έως 14 ετών ως ποσοστό του συνολικού πληθυσμού. Ο πληθυσμός βασίζεται στον de facto ορισμό του πληθυσμού.

26. **VoiceAcc** (Φωνή και λογοδοσία: εκτίμηση): Η "Φωνή και Λογοδοσία" καταγράφει τις αντιλήψεις σχετικά με τον βαθμό στον οποίο οι πολίτες μιας χώρας μπορούν να συμμετέχουν στην επιλογή της κυβέρνησής τους, καθώς και την ελευθερία της έκφρασης, την ελευθερία του συνεταιρίζεσθαι και τα ελεύθερα μέσα ενημέρωσης. Η εκτίμηση δίνει τη βαθμολογία της χώρας στον συνολικό δείκτη, σε μονάδες μιας τυπικής κανονικής κατανομής, δηλαδή κυμαίνεται περίπου από -2,5 έως 2,5.

27. **WomBusiness** (Women Business and the Law Index Score, κλίμακα 1-100): Ο δείκτης μετρά τον τρόπο με τον οποίο οι νόμοι και οι κανονισμοί επηρεάζουν τις οικονομικές ευκαιρίες των γυναικών. Οι συνολικές βαθμολογίες υπολογίζονται με τη μέση βαθμολογία κάθε δείκτη (Κινητικότητα, Χώρος εργασίας, Αμοιβή, Γάμος, Γονιμότητα, Επιχειρηματικότητα, Περιουσιακά στοιχεία και Σύνταξη), με το 100 να αντιπροσωπεύει την υψηλότερη δυνατή βαθμολογία.

## 2.2 Ποσοτικές Μεταβλητές

Οι ποσοτικές μεταβλητές είναι οι

1. Έλεγχος Διαφθοράς (Corruption Control – CorControl),
2. Πολιτική Σταθερότητα (Political Stability – PolStab),
3. Φωνή και Υπευθυνότητα (Voice and Accountability – VoiceAcc).

Για κάθε μία από αυτές, θα παρουσιάσουμε:

- κάποια συνοπτικά στατιστικά (δειγματικός μέσος, διάμεσος, τυπική απόκλιση, ποσοστημόρια κ.ά.),
- ένα ιστόγραμμα (histogram),
- ένα θηκόγραμμα (boxplot),
- εκτέλεση ελέγχου κανονικότητας Shapiro-Wilk και ερμηνεία των αποτελεσμάτων.

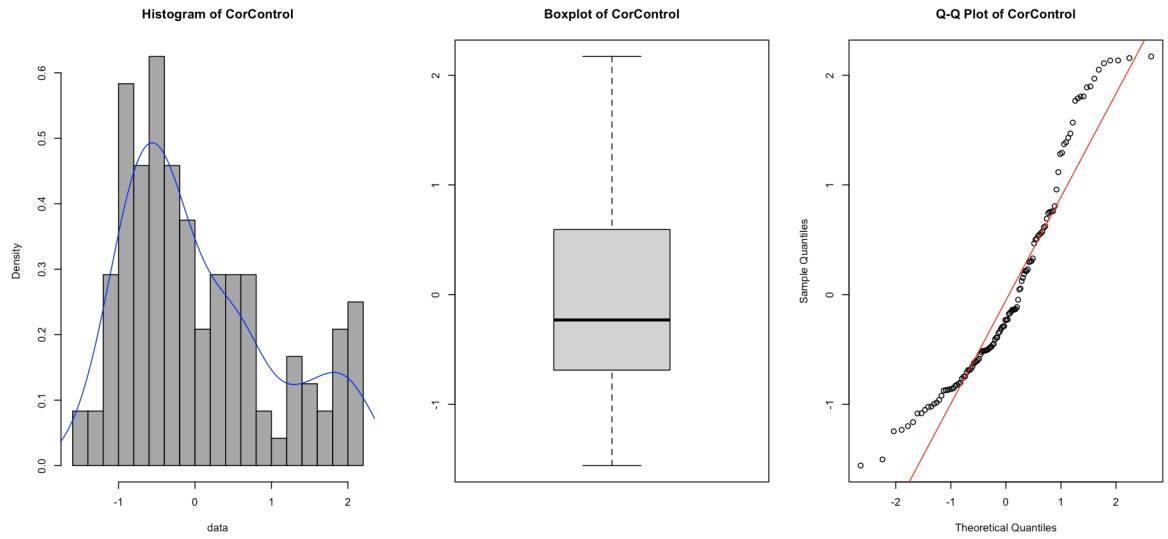
### 2.2.1 Έλεγχος Διαφθοράς

Τα περιγραφικά στατιστικά μέτρα για τη μεταβλητή "Έλεγχος Διαφθοράς (CorControl)" είναι:

| **Στατιστικό Μέτρο** | **CorControl** |
|---|---|
| Μέση τιμή | 0.03603002 |
| Διασπορά | 0.9311955 |
| Τυπική Απόκλιση | 0.9649847 |
| Διάμεσος | -0.2312463 |
| **Ποσοστιμόρια** | |
| 0% | -1.5588410 |
| 25% | -0.6882838 |
| 50% (Διάμεσος) | -0.2312463 |
| 75% | 0.5830534 |
| 100% | 2.1713212 |
| Shapiro-Wilk p-value | $2.180\,701 \times 10^{-6}$ |
| Normality Test Result | Όχι κανονικά κατανεμημένη |

Πίνακας 1: Περιγραφικά στατιστικά στοιχεία και έλεγχος κανονικότητας για τη μεταβλητή CorControl

Τα γραφήματα για τη μεταβλητή αυτή είναι:




Σχήμα 1: Γραφήματα για τη μεταβλητή CorControl

**Ερμηνεία:** Στο ιστόγραμμα, βλέπουμε ότι τα κατάλοιπα δεν είναι συμμετρικά γύρω από το μηδέν, όπως θα έπρεπε, ενώ υπάρχουν κιόλας κάποια κατάλοιπα τα οποία είναι υπερβολικά απομακρυσμένα από το μηδέν προς τα θετικά, πράγμα το οποίο δε συμφωνεί με την κανονική κατανομή. Στο Θηκόγραμμα, είναι μεν θετικό ότι δεν φαίνεται να υπάρχουν παρατηρήσεις εκτός των ορίων των απολήξεω, ωστόσο, η διάμεσος είναι πιο κοντά στην κάτω πλευρά του ορθογωνίου, οι απολήξεις έχουν άνισα μήκη, οπότε αυτά δεν είναι καλά σημάδια για να δηλώσουμε ότι η κατονομή είναι κανονική. Για να έχουμε κανονική κατανομή, θα έπρεπε στο Q-Q Plot τα σημεία του γραφήματος να βρίσκονται συγκεντρωμένα πολύ κοντά στην κόκκινη ευθεία, ενώ εμείς παρατηρούμε ότι υπάρχουν σημεία που απέχουν πολύ από την ευθεία που απεικονίζεται στο γράφημα. Ειδικότερα, τα σημεία απέχουν πολύ από την ευθεία στις ουρές τις κατανομής, δηλαδή στο άνω δεξί και το κάτω αριστερό μέρος του γραφήματος, όπως φαίνεται στο 1, τότε αυτό είναι ένδειξη ότι τα κατάλοιπα προέρχονται από κάποια κατανομή με πιο ”παχιές” ουρές από την κανονική κατανομή, πιθανώς μία t-Student. Συνολικά, φαίνεται ότι η κατανομή δεν ακολουθεί κανονική κατανομή.

Για να επιβεβαιώσουμε τις εικασίες μας, κάνουμε το τεστ κανονικότητας Shapiro-Wilk το οποίο μας βγάζει αποτέλεσμα $p - value = 2.180701e - 06$. Το $p - value$ είναι $2.180701e - 06$ και σε επίπεδο σημαντικότητας 5% απορρίπτω τη μηδενική υπόθεση, ότι δηλαδή τα δεδομένα μου ακολουθούν κανονική κατανομή. Άρα η μεταβλητή CorControl δεν είναι κανονικά κατανεμημένη. Όταν τα δεδομένα δεν ακολουθούν την κανονική κατανομή, χρησιμοποιούμε την διάμεσο και τα ποσοστημόρια, καθώς αυτά προσφέρουν μια πιο αξιόπιστη και ακριβή περιγραφή της κατανομής των δεδομένων.

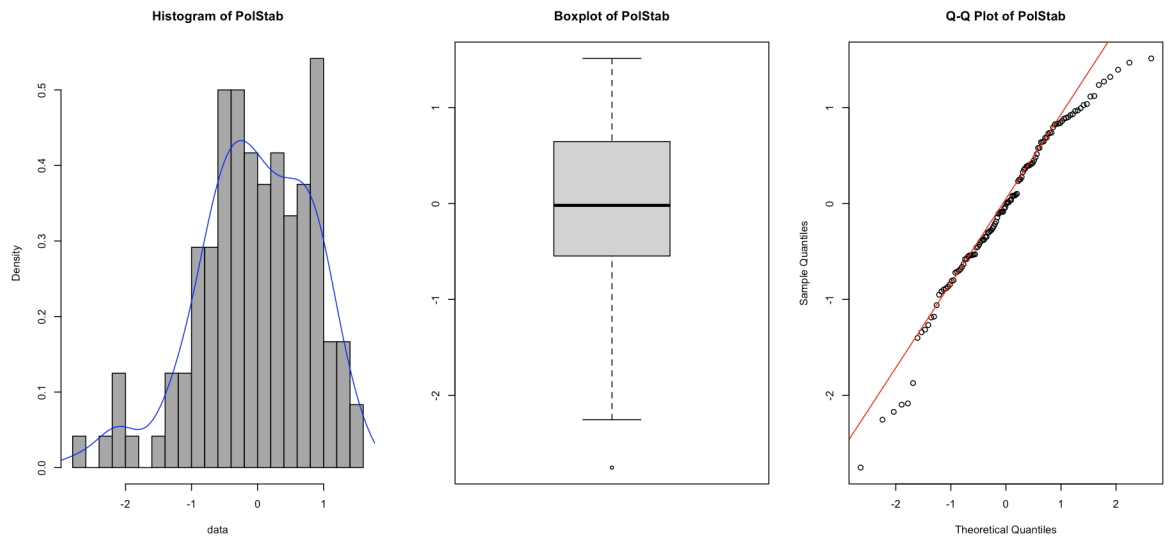
### 2.2.2 Πολιτική Σταθερότητα

Τα περιγραφικά στατιστικά μέτρα για τη μεταβλητή "Πολιτική Σταθερότητα (Political Stability – PolStab)" είναι:

| Στατιστικό Μέτρο | PolStab |
|---|---|
| Μέση τιμή | -0.04803885 |
| Διασπορά | 0.7359813 |
| Τυπική Απόκλιση | 0.8578935 |
| Διάμεσος | -0.01948036 |
| **Ποσοστιμόρια** | |
| 0% | -2.75326204 |
| 25% | -0.54483178 |
| 50% (Διάμεσος) | -0.01948036 |
| 75% | 0.64384830 |
| 100% | 1.51216090 |
| Shapiro-Wilk p-value | 0.009 248 762 |
| Normality Test Result | Όχι κανονικά κατανεμημένη |

Πίνακας 2: Περιγραφικά στατιστικά στοιχεία και έλεγχος κανονικότητας για τη μεταβλητή PolStab

Τα γραφήματα για τη μεταβλητή αυτή είναι:




Σχήμα 2: Γραφήματα για τη μεταβλητή PolStab

**Ερμηνεία:** Στο ιστόγραμμα, βλέπουμε ότι τα κατάλοιπα δεν είναι συμμετρικά γύρω από το μηδέν, όπως θα έπρεπε, ενώ υπάρχουν κιόλας κάποια κατάλοιπα τα οποία είναι υπερβολικά απομακρυσμένα από το μηδέν προς τα αρνητικά, πράγμα το οποίο δε συμφωνεί με την κανονική κατανομή. Στο Θηκόγραμμα, είναι μεν θετικό ότι δεν υπάρχουν παρατηρήσεις εκτός των ορίων των απολήξεων, και ότι η διάμεσος είναι κοντά στο μέση του ορθογωνίου, οι απολήξεις έχουν άνισα μήκη, οπότε αυτά δεν είναι καλό σημάδι για να δηλώσουμε ότι η κατονομή είναι κανονική. Για να έχουμε κανονική κατανομή, θα έπρεπε στο Q-Q Plot τα σημεία του γραφήματος να βρίσκονται συγκεντρωμένα πολύ κοντά στην κόκκινη ευθεία, ενώ εμείς παρατηρούμε ότι υπάρχουν σημεία που απέχουν πολύ από την ευθεία που απεικονίζεται στο γράφημα. Ειδικότερα, τα σημεία απέχουν πολύ από την ευθεία στις ουρές τις κατανομής, δηλαδή στο άνω δεξί και το κάτω αριστερό μέρος του γραφήματος, όπως φαίνεται στο 2, τότε αυτό είναι ένδειξη ότι τα κατάλοιπα προέρχονται από κάποια κατανομή με πιο ”παχιές” ουρές από την κανονική κατανομή, πιθανώς μία t-Student. Συνολικά, φαίνεται ότι η κατανομή δεν ακολουθεί κανονική κατανομή.

Για να επιβεβαιώσουμε τις εικασίες μας, κάνουμε το τεστ κανονικότητας Shapiro-Wilk το οποίο μας βγάζει αποτέλεσμα $p-value = 0.009248762$. Το $p-value$ είναι $0.009248762$ και σε επίπεδο σημαντι-

κότητας 5% απορρίπτω τη μηδενική υπόθεση, ότι δηλαδή τα δεδομένα μου ακολουθούν κανονική κατανομή. Άρα η μεταβλητή PolStab δεν είναι κανονικά κατανεμημένη. Όταν τα δεδομένα δεν ακολουθούν την κανονική κατανομή, χρησιμοποιούμε την διάμεσο και τα ποσοστημόρια, καθώς αυτά προσφέρουν μια πιο αξιόπιστη και ακριβή περιγραφή της κατανομής των δεδομένων.

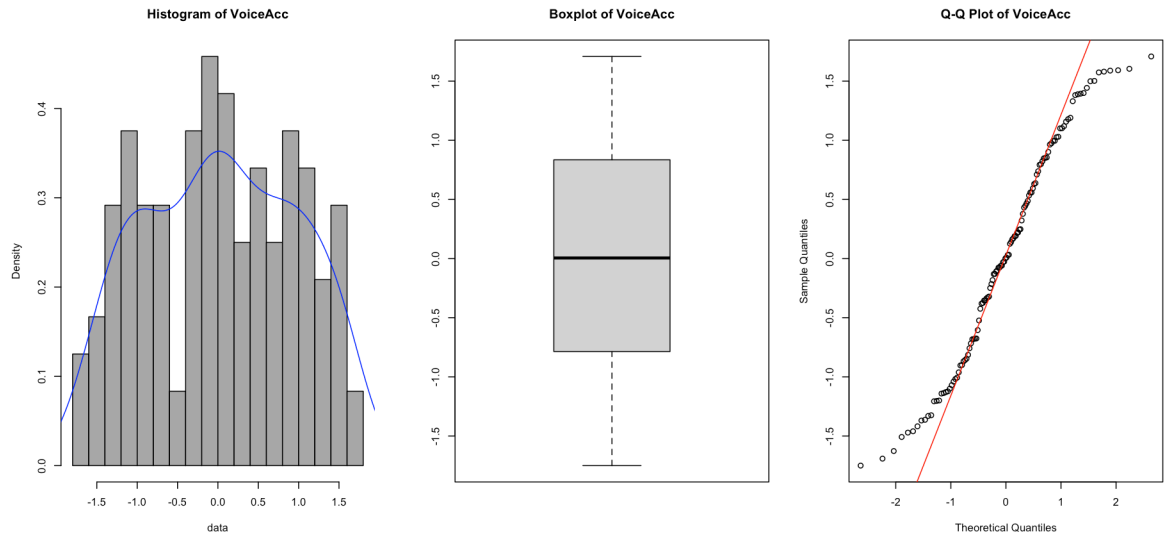
### 2.2.3 Φωνή και Υπευθυνότητα

Τα περιγραφικά στατιστικά μέτρα για τη μεταβλητή "Φωνή και Υπευθυνότητα (Voice and Accountability – VoiceAcc)" είναι:

| **Στατιστικό Μέτρο** | **VoiceAcc** |
|---|---|
| Μέση τιμή | 0.03067167 |
| Διασπορά | 0.8775737 |
| Τυπική Απόκλιση | 0.936789 |
| Διάμεσος | 0.005002012 |
| **Ποσοστιμόρια** | |
| 0% | -1.750232816 |
| 25% | -0.772283003 |
| 50% (Διάμεσος) | 0.005002012 |
| 75% | 0.828630462 |
| 100% | 1.708544374 |
| Shapiro-Wilk p-value | 0.003 363 745 |
| Normality Test Result | Όχι κανονικά κατανεμημένη |

Πίνακας 3: Περιγραφικά στατιστικά στοιχεία και έλεγχος κανονικότητας για τη μεταβλητή VoiceAcc

Τα γραφήματα για τη μεταβλητή αυτή είναι:




Σχήμα 3: Γραφήματα για τη μεταβλητή VoiceAcc

**Ερμηνεία:** Στο ιστόγραμμα, βλέπουμε ότι τα κατάλοιπα είναι συμμετρικά γύρω από το μηδέν, όπως θα έπρεπε, αλλά στα άκρα οι ουρές είναι αρκετά πιο παχιές από το αναμενόμενο. Το θηκόγραμμα φαίνεται αψεγάδιαστο καθώς η διάμεσος βρίσκεται στη μέση του ορθογωνίου, οι ουρές έχουν ίσα σχεδόν μήκη και δεν υπάρχουν παρατηρήσεις που ξεφεύγουν από τα άκρα. Όμως για να έχουμε κανονική κατανομή, θα έπρεπε στο Q-Q Plot τα σημεία του γραφήματος να βρίσκονται συγκεντρωμένα πολύ κοντά

στην κόκκινη ευθεία, ενώ εμείς παρατηρούμε ότι υπάρχουν σημεία που απέχουν πολύ από την ευθεία που απεικονίζεται στο γράφημα. Ειδικότερα, τα σημεία απέχουν πολύ από την ευθεία στις ουρές τις κατανομής, δηλαδή στο άνω δεξί και το κάτω αριστερό μέρος του γραφήματος, όπως φαίνεται στο 2, τότε αυτό είναι ένδειξη ότι τα κατάλοιπα προέρχονται από κάποια κατανομή με πιο "παχιές" ουρές από την κανονική κατανομή, πιθανώς μία t-Student. Συνολικά, φαίνεται ότι η κατανομή δεν ακολουθεί κανονική κατανομή.

Για να επιβεβαιώσουμε τις εικασίες μας, κάνουμε το τεστ κανονικότητας Shapiro-Wilk το οποίο μας βγάζει αποτέλεσμα $p-value = 0.003363745$. Το $p-value$ είναι $0.003363745$ και σε επίπεδο σημαντικότητας 5% απορρίπτω τη μηδενική υπόθεση, ότι δηλαδή τα δεδομένα μου ακολουθούν κανονική κατανομή. Άρα η μεταβλητή VoiceAcc δεν είναι κανονικά κατανεμημένη. Όταν τα δεδομένα δεν ακολουθούν την κανονική κατανομή, χρησιμοποιούμε την διάμεσο και τα ποσοστημόρια, καθώς αυτά προσφέρουν μια πιο αξιόπιστη και ακριβή περιγραφή της κατανομής των δεδομένων.

## 2.3 Ποιοτικές Μεταβλητές

Οι ποιοτικές μεταβλητές είναι οι:

1. Κατηγορία Γονιμότητας (Fertility Category, δώστε το όνομα FertCat)
2. Κατηγορία Πληθωρισμού (Inflation Category – δώστε το όνομα InflCat).

Για κάθε μία από αυτές, θα παρουσιάσουμε:

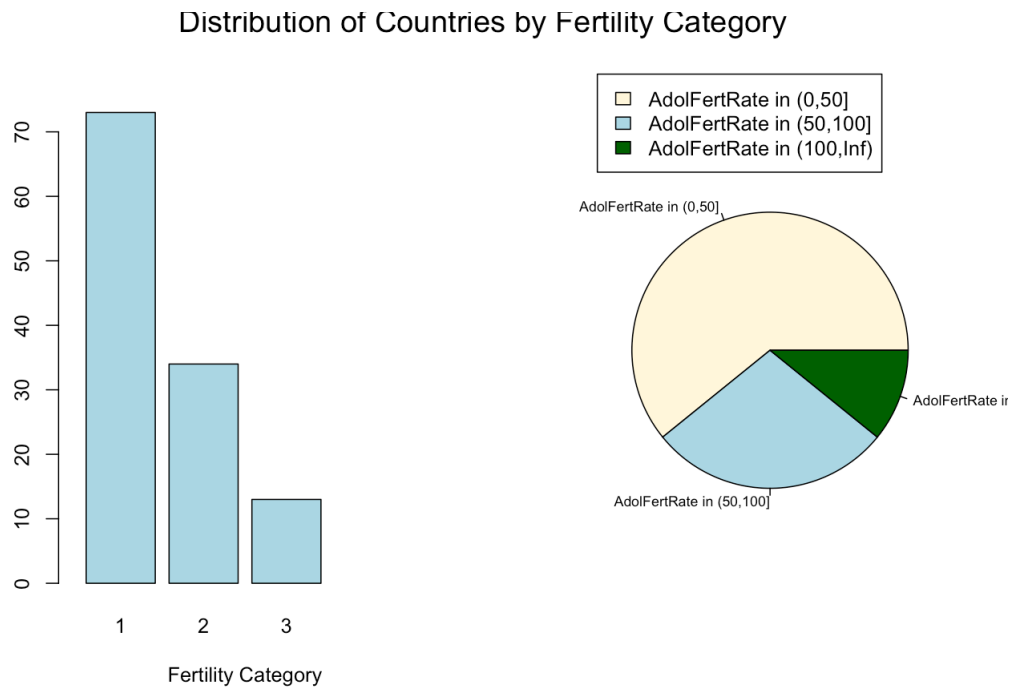
- τον πίνακα συχνοτήτων,
- ένα ραβδόγραμμα (barplot).

### 2.3.1 Κατηγορία Γονιμότητας

Για την κατηγοριοποίηση της "Κατηγορίας Γονιμότητας" (FertCat), θα χρησιμοποιηθούν τα εξής κριτήρια που αφορούν στον Ρυθμό Εφηβικής Γονιμότητας (Adolescent Fertility Rate - AdolFertRate) της εκάστοτε χώρας:

- Τιμή 1, όταν ο Ρυθμός Εφηβικής Γονιμότητας βρίσκεται στο διάστημα $(0, 50]$.
- Τιμή 2, όταν ο Ρυθμός Εφηβικής Γονιμότητας βρίσκεται στο διάστημα $(50, 100]$.
- Τιμή 1, όταν ο Ρυθμός Εφηβικής Γονιμότητας βρίσκεται στο διάστημα $(100, \infty)$.

Με βάση αυτά τα κριτήρια, κάθε χώρα θα κατηγοριοποιείται σε μία από τις τρεις κατηγορίες γονιμότητας, ανάλογα με τον Ρυθμό Εφηβικής Γονιμότητας (AdolFertRate) της. Τα περιγραφικά στατιστικά μέτρα για τη μεταβλητή "Κατηγορία Γονιμότητας" (FertCat) είναι τα εξής:




Σχήμα 4: Περιγραφικά στατιστικά μέτρα για τη μεταβλητή FertCat

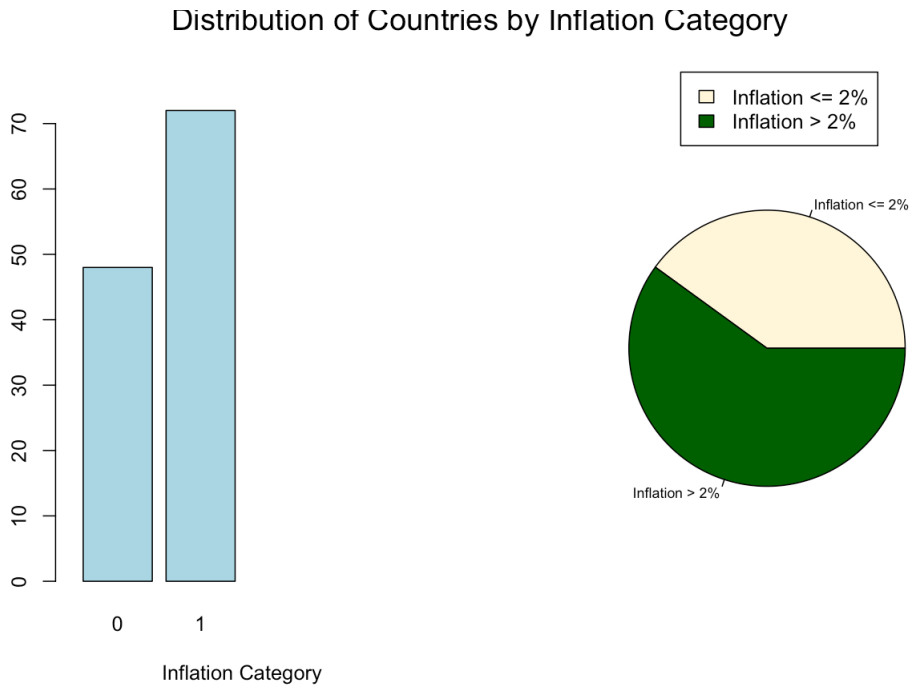
**Ερμηνεία:** Παρατηρούμε ότι κυριαρχεί η τάση οι γυναίκες σε ηλικίες 15-19 ετών να γεννούν λιγότερα παιδιά (0-50 παιδιά ανά 1000 γυναίκες ηλικίας 15-19 ετών).

### 2.3.2 Κατηγορία Πληθωρισμού

Για την κατηγοριοποίηση του "Κατηγορία Πληθωρισμού" (InflCat), θα χρησιμοποιηθούν τα εξής κριτήρια που αφορούν στον Πληθωρισμό (Inflation) της εκάστοτε χώρας :

- Τιμή 0, όταν ο Πληθωρισμός της χώρας είναι μικρότερος ή ίσος από 2%.
- Τιμή 1, όταν ο Πληθωρισμός της χώρας είναι μεγαλύτερος από 2%.

Με βάση αυτά τα κριτήρια, κάθε χώρα θα κατηγοριοποιείται σε μία από τις δύο κατηγορίες πληθωρισμού, ανάλογα με το ποσοστό πληθωρισμού της. Τα περιγραφικά στατιστικά μέτρα για τη μεταβλητή "Κατηγορία Πληθωρισμού" (InfCat) είναι τα εξής:




Σχήμα 5: Περιγραφικά στατιστικά μέτρα για τη μεταβλητή FertCat

**Ερμηνεία:** Στις χώρες που μελετούνται επικρατεί πληθωρισμός μεγαλύτερος του 2%.

## 2.4 Ανάλυση Περιγραφικών Στατιστικών ανά Κατηγορία Γονιμότητας

Χρησιμοποιώντας τη συνάρτηση describeBy του πακέτου psych προκύπτουν περιγραφικά συστατικά (δειγματικό μέσο και τυπική απόκλιση) για τις μεταβλητές Έλεγχος Διαφθοράς (Corruption Control – CorControl), Πολιτική Σταθερότητα (Political Stability – PolStab), και Φωνή και Υπευθυνότητα (Voice and Accountability – VoiceAcc) ανά Κατηγορία Γονιμότητας (Fertility Category – FertCat). Τα αποτελέσματα φαίνονται παρακάτω :

| Κατηγορία | Μέγεθος (Δείγματος) | Μέση τιμή | Τυπική απόκλιση |
|---|---|---|---|
| 1 | 73 | 0.4325478 | 0.9789551 |
| 2 | 34 | -0.4912491 | 0.5542745 |
| 3 | 13 | -0.8115320 | 0.3419991 |

Πίνακας 4: Περιγραφικά συστατικά για τη μεταβλητή Έλεγχος Διαφθοράς (Corruption Control – CorControl

| Κατηγορία | Μέγεθος (Δείγματος) | Μέση τιμή | Τυπική απόκλιση |
|---|---|---|---|
| 1 | 73 | 0.2175080 | 0.8213295 |
| 2 | 34 | -0.2730087 | 0.6618923 |
| 3 | 13 | -0.9508037 | 0.7687288 |

Πίνακας 5: Περιγραφικά συστατικά για τη μεταβλητή Πολιτική Σταθερότητα (Political Stability – PolStab)

| Κατηγορία | Μέγεθος (Δείγματος) | Μέση τιμή | Τυπική απόκλιση |
|---|---|---|---|
| 1 | 73 | 0.1927643 | 1.0577918 |
| 2 | 34 | -0.1210896 | 0.6966411 |
| 3 | 13 | -0.4826266 | 0.3794007 |

Πίνακας 6: Περιγραφικά συστατικά για τη μεταβλητή Φωνή και Υπευθυνότητα (Voice and Accountability – VoiceAcc)

# 3 Μέθοδοι

Για να εξηγηθεί και να προβλεφθεί η Ελέγχου Διαφθοράς από τους άλλους Παγκόσμιους Δείκτες Ανάπτυξης, διενεργήθηκε ανάλυση πολλαπλής παλινδρόμησης. Συγκεκριμένα, η επιλογή του βέλτιστου μοντέλου έγινε με τη χρήση του κριτηρίου AIC και τη διαδικασία βηματικής επιλογής προς τα πίσω (backwards). Το μοντέλο παλινδρόμησης περιλαμβάνει διάφορους προγνωστικούς παράγοντες, καθένας από τους οποίους αντιπροσωπεύει διαφορετικούς Παγκόσμιους Δείκτες Ανάπτυξης, και οι συντελεστές εκτιμήθηκαν με τη μέθοδο των συνήθων ελαχίστων τετραγώνων. Οι υποθέσεις της γραμμικότητας, της ανεξαρτησίας, της ομοσκεδαστικότητας και της κανονικότητας των καταλοίπων ελέγχθηκαν για να διασφαλιστεί η εγκυρότητα του μοντέλου. Η συνολική σημαντικότητα του μοντέλου αξιολογήθηκε με τη χρήση του F-test και η σημασία των επιμέρους προβλεπτικών παραγόντων αξιολογήθηκε με τη χρήση t-test για τους συντελεστές παλινδρόμησης.

Για την αξιολόγηση των διαφορών στην Πολιτική Σταθερότητας σε διαφορετικά επίπεδα των κατηγορικών προγνωστικών παραγόντων "Κατηγορία Γονιμότητας" (Fertility Category - FertCat) και "Κατηγορία Πληθωρισμού" (Inflation Category –InflCat) που δημιουργήθηκαν στην 1η Ενότητα, πραγματοποιήθηκε ανάλυση διακύμανσης (ANOVA). Η ANOVA μας επιτρέπει να προσδιορίσουμε εάν υπάρχουν στατιστικά σημαντικές διαφορές μεταξύ των μέσων όρων της Πολιτικής Σταθερότητας στις ομάδες που ορίζονται από τον εκάστοτε κατηγορικό προγνωστικό παράγοντα. Η συνολική μεταβλητότητα της Πολιτικής Σταθερότητας χωρίζεται σε μεταβλητότητα εντός των ομάδων και σε μεταβλητότητα μεταξύ των ομάδων.

Η ελεγχοσυνάρτηση F, η οποία είναι ο λόγος του μέσου τετραγώνου μεταξύ των ομάδων προς το μέσο τετράγωνο εντός των ομάδων, χρησιμοποιείται για τον έλεγχο της μηδενικής υπόθεσης ότι όλοι οι μέσοι όροι των ομάδων είναι ίσοι. Ένα στατιστικά σημαντικό F-test, υποδεικνύει ότι υπάρχει τουλάχιστον ένας μέσος όρος ομάδας που διαφέρει από τους άλλους, γεγονός που υποδηλώνει ότι ο κατηγορηματικός προγνωστικός παράγοντας που ελέγχεται έχει σημαντική επίδραση στην Πολιτική Σταθερότητα. Πραγματοποιήθηκαν δοκιμές post-hoc (Tukey's HSD test) για τον εντοπισμό των ζευγών των ομάδων που διαφέρουν σημαντικά.

Ακολουθεί θεωρητικό υπόβαθρο που χρησιμοποιήθηκε, το οποίο μπορεί να προσπεραστεί.

## 3.1 Πολλαπλό Γραμμικό Μοντέλο

Όταν δεν έχουμε μόνο μία επεξηγηματική μεταβλητή $X$, αλλά $p$ απαντητικές μεταβλητές $X_1, X_2, \ldots X_p$ και θέλουμε να τις χρησιμοποιήσουμε όλες για να προβλέψουμε τις τιμές της αποκριτικής μεταβλητής , τότε κατασκευάζουμε ένα πολλαπλό γραμμικό μοντέλο ή μοντέλο πολλαπλής γραμμικής παλινδρόμησης. Ορίζουμε το πολλαπλό γραμμικό μοντέλο

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \ldots + \beta_p X_{p,i} + \varepsilon_i, i = 1, 2, \ldots, n.$$

όπου $\beta_0, \beta_1, \ldots, \beta_p$ οι $p + 1$ συντελεστές παλινδρόμησης και $\varepsilon_i$ τα τυχαία σφάλ- ματα, για τα οποία θεωρούμε (πρέπει να επικυρωθεί) ότι είναι κανονικά κατανεμημένα με μέση τιμή 0, ομοσκεδαστικά και ανεξάρτητα, ή ισοδύναμα ασυσχέτιστα. Δηλαδή, $\varepsilon_i$ $N(0, \sigma^2)$ ανεξάρτητα για $i = 1, 2, \ldots, n$.

Ας μην ξεχνάμε ότι τα κατάλοιπα $\hat{\varepsilon_i}$, είναι οι εκτιμήσεις των τυχαίων σφαλμάτων, $\varepsilon_i$. Κάτω από τις υποθέσεις του γραμμικού μοντέλου $E(\varepsilon_i) = 0, V(\varepsilon_i) = \sigma^2, Cov(\varepsilon_i, \varepsilon_j) = 0, i \neq j$. Τα κατάλοιπα, λοιπόν, πρέπει να είναι τυχαία κατανεμημένα γύρω από το 0, ασυσχέτιστα και ομοσκεδαστικά. Αν αυτά δεν ισχύουν, τότε το μοντέλο που έχουμε προσαρμόσει δεν είναι κατάλληλο για τα δεδομένα μας.

Το αν ισχύουν οι υποθέσεις για τους τυχαίους όρους ελέγχεται με γραφικό έλεγχο καταλοίπων (όπως θα δούμε παρακάτω), δηλαδή με τα γραφήματα των καταλοίπων ως προς τα $_i$, ή τα $Y_i$.

Ένα γράφημα καταλοίπων που επιβεβαιώνει τις κλασικές υποθέσεις παρουσιάζει την εικόνα σύννεφου τυχαίων σημείων γύρω από τη γραμμή του 0 και δεν έχει τίποτα συστηματικό. Ό, τι συστηματικό ανιχνευθεί στο γράφημα καταλοίπων αντιστοιχεί σε απόκλιση από τις κλασικές υποθέσεις και πρέπει να μοντελοποιηθεί.

### 3.2 Εύρεση εκτιμήτριας διανυσματικής παραμέτρου $\beta$

Έχοντας, κάνει υπόθεση για την κατανομή των τυχαίων σφαλμάτων $\varepsilon_i$ μπορούμε να κάνουμε χρήση άλλων γνωστών μεθόδων εκτίμησης εκτός από τη μέθοδο ελαχίστων τετραγώνων, όπως η μέθοδος μέγιστης πιθανοφάνειας που γνωρίζουμε από τη μαθηματική στατιστική.

Σε πινακική μορφή, η συνάρτηση του αθροίσματος των τετραγωνικών αποκλίσεων των παρατηρήσεων από τις μέσες τιμές τους γράφεται ως:

$$Q(B) = ||\varepsilon||^2 = ||Y–E(Y)||^2 = ||Y–X\beta||2 = (Y–X\beta)^T(Y–X\beta).$$

Τη συνάρτηση αυτή θέλουμε να την ελαχιστοποιήσουμε για να προσδιορίσουμε την εκτιμήτρια ελαχίστων τετραγώνων $\hat{\beta}$ της διανυσματικής παραμέτρου $\beta$. Για να το επιτύχουμε αυτό, θα πρέπει να παραγωγίσουμε τη συνάρτηση $Q(\beta)$ ως προς τη διανυσματική παράμετρο $\beta$.

### 3.3 Εκτίμηση της Διασποράς

Η αμερόληπτη εκτιμήτρια της διασποράς $\sigma^2$ προκύπτει, αν αντικαταστήσουμε τους $p+1$ άγνωστους συντελεστές παλινδρόμησης $\beta_0, \beta_1, \ldots, \beta_p$ από τις αμερόληπτες εκτιμήτριες $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ και αφαιρέσουμε $p+1$ βαθμούς ελευθερίας από τον παρονομαστή για τις $p+1$ παραμέτρους που εκτιμήσαμε. Η αμερόληπτη εκτιμήτρια του $\sigma^2$ που προκύπτει είναι το λεγόμενο μέσο τετραγωνικό σφάλμα (mean squared error):

$$MSE = S^2 = \frac{||Y - X\hat{\beta}||^2}{n-p-1} = \frac{||Y - \hat{Y}||^2}{n-p-1} = \frac{||\hat{\varepsilon}||^2}{n-p-1} = \frac{\hat{\varepsilon}^T\hat{\varepsilon}}{n-p-1} = \frac{1}{n-p-1}\sum_{i=1}^{n}\hat{\varepsilon}^2.$$

### 3.4 Συντελεστής Προσδιορισμού

**Ορισμός 1** (Αθροίσματα Τετραγώνων)**.**

- Ορίζουμε $SST = \sum_{i=1}^{2}(Y - \bar{Y})^2$ το συνολικό άθροισμα τετραγώνων
- Ορίζουμε $SSR = \sum_{i=1}^{2}(\hat{Y}_i - \bar{Y})^2$ το άθροισμα τετραγώνων που οφείλεται στην παλινδρόμηση (sum of squares due to regression).
- Ορίζουμε $SSE = \sum_{i=1}^{2}(Y_i - \hat{Y}_i)^2$ το άθροισμα τετραγώνων των καταλοίπων (sum of squared errors).

Ισχύει ότι:

$$\underbrace{\sum_{i=1}^{2}(Y - \bar{Y})^2}_{SST} = \underbrace{\sum_{i=1}^{2}(\hat{Y}_i - \bar{Y})^2}_{SSR} + \underbrace{\sum_{i=1}^{2}(Y_i - \hat{Y}_i)^2}_{SSE}$$

***Παρατήρηση.*** Παρατηρούμε ότι: SSE

$$S^2 = MSE = \frac{SSE}{n-p-1}$$

**Ορισμός 2** (Συντελεστής προσδιορισμού (R-squared))**.**

$$R^2 = \frac{SSR}{SST} = 1 - \frac{}{SSE}SST.$$

Το $R^2$ εκφράζει το ποσοστό συνολικής μεταβλητότης των $Y_i$ που οφείλεται στην παλινδρόμηση. Παίρνει τιμές στο $(0, 1)$. Είναι ένα μέτρο καλής προσαρμογής του μοντέλου (measure of goodness of model fit).

**Ορισμός 3** (Προσαρμοσμένος συντελεστής προσδιορισμού (Adjusted R-squared))**.**

$$R^2_{adj} = 1 - \frac{n-1}{n-p-1}\frac{SSE}{SST} < R^2.$$

Ο $R^2_{adj}$ λαμβάνει υπόψιν του και το πλήθος των αγνώστων παραμέτρων του μοντέλου σε συνδυασμό με το πλήθο των παρατηρήσεων (διόρθωση σε σχέση με το $R^2$), συνυπολογίζει, δηλαδή, και το πλήθος η των επεξηγηματικών μεταβλη τών που χρησιμοποιεί το γραμμικό μοντέλο.

***Παρατήρηση.*** Αν προσθέσουμε ανεξάρτητες μεταβλητές στο μοντέλο, το $R^2$ πάντα αυξάνει ενώ ο $R^2_{adj}$ όχι απαραίτητα.

**Ερμηνεία**: Έστω ότι έχουμε ένα σύνολο από $k$ υποψήφιες επεξηγηματικές μεταβλητές για μία μεταβλητή ενδιαφέροντος και θέλουμε να επιλέξουμε μόνο η από αυτές για να κατασκευάσουμε ένα γραμμικό μοντέλο που να εξηγεί όσο το δυνατόν μεγαλύτερο ποσοστό της μεταβλητότητας των δεδομένων. Ξεκινώντας από ένα απλό γραμμικό μοντέλο που κάνει χρήση μόνο μίας εκ των $k$ υποψήφιων επεξηγηματικών μεταβλητών και προσθέτοντας σταδιακά τις υπόλοιπες επεξηγηματικές μεταβλητές, θα παρατηρούσαμε ότι ο συντελεστής $R^2$ θα αυξανόταν συνεχώς. Συνεπώς, θα έπαιρνε τη μέγιστη δυνατή τιμή του όταν θα είχαμε προσθέσει όλες τις και υποψήφιες επεξηγηματικές μεταβλητές στο γραμμικό μοντέλο.

Αντιθέτως, βλέπουμε ότι ο προσαρμοσμένος συντελεστής προσδιορισμού είναι φθίνουσα συνάρτηση του $p$. Επομένως, ακολουθώντας την ίδια διαδικασία, θα καταλήγαμε σε ένα γραμμικό που εξηγεί ένα μεγάλο ποσοστό από τη μεταβλητότητα των δεδομένων με όσο το δυνατόν μικρότερο πλήθος επεξηγηματικών μεταβλητών.

### 3.5 Συμπερασματολογία εκτιμημένων παραμέτρων

**Πρόταση 1.** Η εκτιμήτρια μέγιστης πιθανοφάνειας του $\beta$ ταυτίζεται με την εκτιμήτρια ελαχίστων τετραγώνων $\hat{\beta}$, ενώ η εκτιμήτρια μέγιστης πιθανόφανειας της διασποράς $\sigma^2$ δίνεται από τον τύπο:

$$\hat{\sigma}^2 = \frac{SSE}{n} = \frac{(n-p-1)S^2}{n}.$$

**Πρόταση 2.** (Κατανομή του $\hat{\beta}$ με χρήση του $S^2$) Ισχύει ότι:

i.
$$\frac{(\hat{\beta}-\beta)^T X^T X(\hat{\beta}-\beta)}{(p+1)S^2} = \frac{||X(\hat{\beta}-\beta)||^2}{(p+1)S^2} = \frac{||\hat{Y}-X\beta||^2}{(p+1)S^2} \sim F_{p+1,n-p-1}.$$

ii.
$$\frac{\hat{\beta}_j - \beta_j}{S_{\hat{\beta}_j}} \sim t_{n-p-1} \text{ για} j = 0,1,\ldots,p, \text{ όπου } S^2_{\hat{\beta}_j} = S^2(X^TX)^{-1}_{j+1,J=1}.$$

Τα παραπάνω μπορούν να χρησιμοποιηθούν σε ελέγχους στατιστικής σημαντικότητας των παραμέτρων του μοντέλου καθώς και σε ελέγχους υποθέσεων.

### 3.6 Διαστήματα Εμπιστοσύνης και Έλεγχοι Υποθέσεων για τις εκτιμηθείσες παραμέτρους

Χρησιμοποιώντας την Πρόταση 2, μπορούμε άμεσα να κατασκευάσουμε διαστήματα εμπιστοσύνηςκαι περιοχές εμπιστοσύνης για τις παραμέτρους του μοντέλου.

**Πρόταση 2.13.** *(Διαστήματα και Περιοχές Εμπιστοσύνης)*

i. Ένα $100(1-\alpha)\%$ διάστημα εμπιστοσύνης ίσων ουρών για το $\beta_j$ δίνεται από τη σχέση:

$$I_{1-\alpha}(\beta_j) = \left[\hat{\beta}_j - t_{n-p-1;\frac{\alpha}{2}} \cdot S_{\hat{\beta}_j}, \hat{\beta}_j + t_{n-p-1;\frac{\alpha}{2}} \cdot S_{\hat{\beta}_j}\right], \quad j = 0,1,\ldots,p.$$

**Πρόταση 2.14.** Υπό τη μηδενική υπόθεση $H_0 : \beta_j = \beta_{j,0}$ για $j = 0, 1, \ldots, p$, ισχύει ότι

$$T = \frac{\hat{\beta}_j - \beta_{j,0}}{S_{\hat{\beta}_j}} \sim t_{n-p-1}.$$

Αντικαθιστώντας τις τυχαίες μεταβλητές $Y_1, Y_2, \ldots, Y_n$ που εμφανίζονται στην ελεγχοσυνάρτηση $T$ από τις παρατηρήσεις $y_1, y_2, \ldots, y_n$, υπολογίζουμε την παρατηρούμενη τιμή $t = \frac{\hat{\beta}_j - \beta_{j,0}}{S_{\hat{\beta}_j}}$.

i. Έστω ότι θέλουμε να πραγματοποιήσουμε τον έλεγχο με την αμφίπλευρη εναλλακτική υπόθεση $H_1 : \beta_j \neq \beta_{j,0}$. Τότε, απορρίπτουμε την $H_0$ σε ε.σ. α αν και μόνο αν $|t| > t_{n-p-1;\frac{\alpha}{2}}$ ή $p$-value$^{(\neq)} = P(|T| \geq |t|) < \alpha$ ή $\beta_{j,0} \notin I_{1-\alpha}(\beta_j)$.

ii. Έστω ότι θέλουμε να πραγματοποιήσουμε τον έλεγχο με την μονόπλευρη εναλλακτική υπόθεση $H_1 : \beta_j > \beta_{j,0}$. Τότε, απορρίπτουμε την $H_0$ σε ε.σ. α αν και μόνο αν $t > t_{n-p-1;\alpha}$ ή $p$-value$^{(>)} = P(T > t) < \alpha$.

iii. Έστω ότι θέλουμε να πραγματοποιήσουμε τον έλεγχο με την μονόπλευρη εναλλακτική υπόθεση $H_1 : \beta_j < \beta_{j,0}$. Τότε, απορρίπτουμε την $H_0$ σε ε.σ. α αν και μόνο αν $t < -t_{n-p-1;\alpha}$ ή $p$-value$^{(<)} = P(T \leq t) < \alpha$.

**Πίνακας ANOVA για την Πολλαπλή Παλινδρόμηση**

| Πηγή Μεταβλητότητας | SS | d.f. | MS | F |
|---|---|---|---|---|
| Παλινδρόμηση | $SSR$ | $p$ | $SSR/p$ | $MSR/MSE$ |
| Σφάλματα | $SSE$ | $n-p-1$ | $SSE/(n-p-1)$ | |
| Σύνολο | $SST$ | $n-1$ | | |

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2)$$

$p + 1$ παράμετροι.

- $SSR$: Άθροισμα τετραγώνων που οφείλεται στην παλινδρόμηση (sum of squares due to regression).
- $SSE$: Άθροισμα τετραγώνων των καταλοίπων (sum of squared errors).
- $SST$: Συνολικό άθροισμα τετραγώνων (total sum of squares).

$p$-value: η πιθανότητα μια τ.μ. $F_{(p,n-p-1)}$ να πάρει τιμή τόσο ακραία ή περισσότερο ακραία από $F^*$.
Έλεγχος: $H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$, Δεν υπάρχει γραμμική σχέση.

$$F = \frac{MSR}{MSE} \sim F(p, n-p-1).$$

Απορρίπτουμε την $H_0$ σε επίπεδο στατιστικής σημαντικότητας $\alpha$ αν $F^* > F_\alpha(p, n-p-1)$ (ή ισοδύναμα αν $p$-value $< \alpha$).

## 3.7 Κριτήρια Επιλογής Μοντέλου

Έστω ότι έχουμε ένα σύνολο από $k$ υποψήφιες επεξηγηματικές μεταβλητές για μία μεταβλητή ενδιαφέροντος και θέλουμε να επιλέξουμε μόνο η από αυτές για να κατασκευάσουμε ένα γραμμικό μοντέλο. Το πλήθος όλων των πιθανών γραμμικών μοντέλων που μπορούμε να κατασκευάσουμε με βάση αυτές τις Κ υποψήφιες επεξηγηματικές μεταβλητές είναι $2^k - 1$, οπότε καταλαβαίνουμε ότι αυτή η σύγκριση θα ήταν αδύνατο να γίνει στο χαρτί ακόμα και για μικρό $k$.

Έχουμε ήδη δει τον προσαρμοσμένο συντελεστή προσδιορισμού, ο οποίος χρησιμοποιείται για να επιλέξουμε το μοντέλο που εξηγεί ένα μεγάλο ποσοστό από τη συνολική μεταβλητότητα των δεδομένων με όσο το δυνατόν μικρότερο πλήθος επεξηγηματικών μεταβλητών. Εκτιμώντας όλα τα πιθανά $2^k - 1$ γραμμικά μοντέλα και υπολογίζοντας τον προσαρμοσμένο συντελεστή προσδιορισμού για καθένα από αυτά, θα επιλέγαμε εκείνο που επιτυγχάνει τη μέγιστη δυνατή τιμή.

Κάποια γενικότερα κριτήρια επιλογής μοντέλου, τα οποία βρίσκουν εφαρμογή και εκτός τους πλαισίου των γραμμικών μοντέλων, είναι τα λεγόμενα κριτήρια πληροφορίας, τα οποία βασίζονται στη μέγιστη πιθανοφάνεια ενός μοντέλου. Με τον όρο μέγιστη πιθανοφάνεια εννοούμε την τιμή που επιτυγχάνει η συνάρτηση πιθανοφάνειας $L(\beta_p, \sigma_p^2|y)$ για $\beta = \hat{\beta}$ και $\sigma^2 = \hat{\sigma}^2$, δηλαδή η πιθανοφάνεια υπολογισμένη στο σημείο της εκτίμησης μέγιστης πιθανοφάνειας.

Έστω $\mathbf{X}_p \in \mathbb{R}^{n\times(p+1)}$ ο πίνακας σχεδιασμού που αντιστοιχεί στις επεξηγηματικές μεταβλητές $X_1, X_2, \ldots, X_p$. Έχουμε υπολογίσει τη συνάρτηση πιθανοφάνειας ενός κανονικού γραμμικού μοντέλου με η επεξηγηματικές μεταβλητές ως:

$$L(\beta_p, \sigma_p^2|y) = (2\pi\sigma_p^2)^{-\frac{n}{2}} exp\{-\frac{||y - X_p\beta_p||^2}{2\sigma_p^2}\} \Rightarrow$$

$$\ell(\beta_p, \sigma_p^2|y) = \log L(\beta_p, \sigma_p^2|y) = -\frac{n}{2}\log\big(2\pi\sigma_p^2\big) - \frac{\|y - X_p\beta_p\|^2}{2\sigma_p^2}.$$

Επιπλέον, έχουμε υπολογίσει τις εκτιμήσεις μέγιστης πιθανοφάνειας:

$$\hat{\beta_p} = (X_p^T X_p)^{-1} X_p^T y, \quad \hat{\sigma_p}^2 = \frac{\|y - X_p\hat{\beta_p}\|^2}{n} = \frac{SSE_p}{n}.$$

Έστω $\hat{\ell}_p = \ell\left(\hat{\beta_p}, \hat{\sigma_p^2}|y\right)$ η μέγιστη λογαριθμοπιθανότητα του γραμμικού μοντέλου με p επεξηγηματικές μεταβλητές. Τότε,

$$\hat{\ell}_p = -\frac{n}{2}\log\left(2\pi\hat{\sigma_p^2}\right) - \frac{\|y - X_p\hat{\beta_p}\|^2}{2\hat{\sigma_p^2}} = -\frac{n}{2}\log\frac{2\pi SSE_p}{n} - \frac{n}{2}.$$

Προφανώς, όσο μεγαλύτερη πιθανότητα έχει ένα μοντέλο, τόσο πιο πιθανό είναι να έχουμε παρατηρήσει τα δεδομένα που παρατηρήσαμε από αυτό το μοντέλο, οπότε αυτό είναι και το μοντέλο που θέλουμε να επιλέξουμε. Όμως, όπως ο συντελεστής προσδιορισμού, έτσι και η μέγιστη πιθανότητα του γραμμικού μοντέλου αυξάνεται συνεχώς όσο προστίθενται καινούργιες επεξηγηματικές μεταβλητές μέσω του μοντέλου. Για τον λόγο αυτό, τα κριτήρια πληροφορίας συνυπολογίζουν και το πλήθος των επεξηγηματικών του γραμμικού μοντέλου, ώστε να μας βοηθήσουν να επιλέξουμε ένα μοντέλο που έχει μεγάλη πιθανότητα με όσο το δυνατόν λιγότερο πλήθος επεξηγηματικών μεταβλητών.

Το **κριτήριο πληροφορίας του Akaike** (AIC - Akaike Information Criterion) ορίζεται ως εξής:

$$\mathrm{AIC}_p = -2\hat{\ell}_p + 2(p+2) = n\log\frac{2\pi SSE_p}{n} + n + 2(p+2).$$

Ο όρος $p+2$ αντιπροσωπεύει το πλήθος των αγνώστων παραμέτρων που εκτιμήσαμε στο μοντέλο με p επεξηγηματικές μεταβλητές, δηλαδή $p+1$ συντελεστές παλινδρόμησης και τη διασπορά $\sigma^2$. Σκοπός μας είναι να μεγιστοποιήσουμε την τιμή $\hat{\ell}_p$, προσπαθώντας παράλληλα να ελαχιστοποιήσουμε τον όρο $p+2$, οπότε καλύτερο μοντέλο είναι αυτό που επιτυγχάνει την μικρότερη δυνατή τιμή $\mathrm{AIC_p}$.

Η απόφαση για το ποιο μοντέλο είναι τελικά βέλτιστο ανάμεσα στα διαφορετικά μοντέλα που επέλεξαν τα παραπάνω κριτήρια εξαρτάται από πολλούς παράγοντες.

Λόγω του υπερβολικά μεγάλου αριθμού γραμμικών μοντέλων που πρέπει να συγκριθούν, ώστε να καταλήξουμε στο καλύτερο μοντέλο με βάση κάποιο επιλεγμένο κριτήριο επιλογής, αυτή η διαδικασία είναι πολλές φορές χρονοβόρα ακόμα και για έναν υπολογιστή. Για τον λόγο αυτό, κάνουμε συχνά χρήση κάποιων απλούστερων μεθόδων βηματικής παλινδρόμησης. Ξεκινώντας από κάποιο αρχικό γραμμικό μοντέλο, αυτές οι μέθοδοι προσθέτουν ή αφαιρούν σε κάθε βήμα κάποια επεξηγηματική μεταβλητή με

βάση κάποιο προεπιλεγμένο κριτήριο, μέχρι να καταλήξουν σε κάποιο γραμμικό μοντέλο το οποίο δε βελτιώνεται από καμία προσθήκη ή διαγραφή επεξηγηματικής μεταβλητής.

Η μέθοδος backward, που χρησιμοποιούμε σε αυτή την ανάλυση, χρησιμοποιεί ως σημείο εκκίνησης το γραμμικό μοντέλο το οποίο περιέχει όλες τις $k$ δυνατές επεξηγηματικές μεταβλητές. Ας υποθέσουμε ότι χρησιμοποιούμε ως κριτήριο επιλογής το AIC. Σε κάθε βήμα της μεθόδου, δοκιμάζουμε να αφαιρέσουμε καθεμία από τις επεξηγηματικές μεταβλητές του γραμμικού μοντέλου ξεχωριστά και υπολογίζουμε το AIC των γραμμικών μοντέλων που δημιουργούνται. Επιλέγουμε να αφαιρέσουμε εκείνη την επεξηγηματική μεταβλητή που θα οδηγήσει στη μεγαλύτερη δυνατή μείωση της τιμής του AIC. Αν, σε κάποιο βήμα της μεθόδου, η διαγραφή οποιασδήποτε επεξηγηματικής μεταβλητής οδηγεί σε αύξηση του AIC σε σύγκριση με το τρέχον μοντέλο, τότε η μέθοδος τερματίζεται και επιλέγουμε το τρέχον μοντέλο ως βέλτιστο.

## 3.8 Διαγνωστικοί έλεγχοι

Έχοντας επιλέξει και εκτιμήσει ένα κατάλληλο κανονικό πολλαπλό γραμμικό μοντέλο για την με τη μέθοδοπου περιγράψαμε στην προηγούμενη παράγραφο, η δουλειά μας δεν έχει τελειώσει ακόμα. Ο βασικός λόγος είναι οι 4 υποθέσεις για τα τυχαία σφάλματα $\varepsilon_i$ πάνω στις οποίες έχουμε στηριχτεί για να κατασκευάσουμε το εν λόγω γραμμικό μοντέλο, δηλαδή ότι είναι κανονικά κατανεμημένα, με μέση τιμή $0$, κοινή διασπορά $\sigma^2$ και ανεξάρτητα. Αν, έχοντας εκτιμήσει το γραμμικό μοντέλο μας, διαπιστώσουμε ότι κάποια από αυτές τις υπο-θέσεις δεν ευσταθεί, τότε αυτό σημαίνει, προφανώς, ότι όλο το γραμμικό μοντέλο το οποίο έχουμε κατασκευάσει δεν είναι έγκυρο και πρέπει να επεξεργαστεί με διαφορετικό τρόπο.

Εφόσον τα τυχαία σφάλματα $\varepsilon_i$ είναι μη-παρατηρήσιμα και μη-υπολογίσιμα, τις 4 αυτές υποθέσεις τις επαληθεύουμε μέσω των εκτιμημένων σφαλμάτων $\hat{\varepsilon}_i$. Επειδή τα κατάλοιπα $\hat{\varepsilon}_i$, δεν έχουν κοινή διασπορά, δηλαδή δεν είναι ισόνομα, χρησιμοποιούμε κάποιες τυποποιημένες εκδοχές τους για την επικύρωση των υποθέσεων της γραμμικής παλινδρόμησης.

Τα κατάλοιπα δείξαμε έχουν πάντα μέση τιμή $0$, οπότε αυτή είναι η μόνη υπόθεση που δε χρειάζεται να ελέγξουμε.

### 3.8.1 Shapiro-Wilk

Την υπόθεση ότι τα τυχαία σφάλματα είναι κανονικά κατανεμημένα μπορούμε να την ελέγξουμε κάνοντας χρήση διαφόρων ελέγχων κανονικότητας, όπως ο έλεγ χος Shapiro - Wilk, ο οποίος έχει δειχθεί ότι έχει τη μεγαλύτερη ισχύ ανάμεσα στους γνωστούς ελέγχους κανονικότητας. Με βάση ένα δείγμα $V_1, V_2, \ldots, V_n$, οι έλεγχοι κανονικότητας λαμβάνουν απόφαση για τις υποθέσεις: $_0 : V_1, V_2, \ldots, V_n$ τυχαίο δείγμα από την κανονική κατανομή vs. $_1 : V_1, V_2, \ldots, V_n$ όχι τυχαίο δείγμα από την κανονική κατανομή.

Εφαρμόζουμε τον έλεγχου Shapiro - Wilk στο δείγμα $t_1, t_2, \ldots, t_n$ των εσωτερικά τυποποιημένων καταλοίπων και λαμβάνουμε το $p - value$. Για δεδομένο ε.σ.σ. $\alpha$, έχουμε τα εξής ενδεχόμενα:

- Αν $p - value < \alpha$, τότε απορρίπτουμε την Ho, οπότε τα εσωτερικά τυποποιημένα κατάλοιπα δεν είναι κανονικά κατανεμημένα. Σε αυτήν την περίπτωση, η υπόθεση ότι τα τυχαία σφάλματα προέρχονται από την κανονική κατανομή δεν ευσταθεί, οπότε θα μπορούσαμε να καταφύγουμε στην κατασκευή ενός γενικευμένου γραμμικού μοντέλου.

- Αν $p-value > \alpha$, τότε δεν μπορούμε να απορρίψουμε την $_0$, δηλαδή την υπόθεση ότι τα εσωτερικά τυποποιημένα κατάλοιπα είναι κανονικά κατανεμημένα. Επομένως, μπορούμε να δεχτούμε ότι το γραμμικό μοντέλο που έχουμε κατασκευάσει έχει τυχαία σφάλματα που προέρχονται από την και

### 3.8.2 Γραφικοί έλεγχοι κανονικότητας

Οι έλεγχοι κανονικότητας καλό είναι πάντα να συνοδεύονται και από διάφορους γραφικούς ελέγχους, όπως ιστογράμματα (histograms), θηκογράμματα (boxplots) και Q-Q plots (quantile - quantile plots).

**Ιστόγραμμα** Πάνω από το ιστόγραμμα σχεδιάζουμε την καμπύλη της συνάρτησης πυκνότητας πιθανότητας της κανονικής κατανομής. Αν το ιστόγραμμα συμφωνεί με την καμπύλη της τυποποιημένης κανονικής κατανομής $(0, 1)$, τότε αυτό είναι ένδειξη ότι τα κατάλοιπα είναι κανονικά κατανεμημένα. Αντιθέτως, αν τα κατάλοιπα δεν είναι συμμετρικά γύρω από το μηδέν, όπως θα έπρεπε, ή υπάρχουν κάποια κατάλοιπα τα οποία είναι υπερβολικά απομακρυσμένα από το μηδέν προς τα θετικά ή τα αρνητικά, μάλλον τα κατάλοιπα δε θα ακολουθούν κανονική κατανομή.

**Θηκόγραμμα** Σε ένα θηκόγραμμα, το ορθογώνιο πλαίσιο αντιπροσωπεύει το ενδοτεταρτημοριακό εύρος (IR - interquartile range), δηλαδή το κεντρικό διάστημα στο οποίο ανήκει το 50% των παρατηρήσεων.

Μέσα στο ορθογώνιο πλαίσιο είναι σχεδιασμένη με μαύρη γραμμή η διάμεσος (median) του διανύσματος των παρατηρήσεων, δηλαδή η μεσαία διατεταγμένη παρατήρηση.

Αν η διάμεσος βρίσκεται περίπου στη μέση του ορθογώνιου πλαισίου, οι απολήξεις έχουν περίπου ίσα μήκη και δεν υπάρχουν παρατηρήσεις εκτός των ορίων των απολήξεων, τότε τα κατάλοιπα φαίνεται να προέρχονται από την κανονική κατανομή.

**Q-Q Plot** Σε ένα Normal Q-Q plot, έχουμε στον άξονα των y τα δειγματικά ποσοστιαία σημεία, ενώ στον άξονα των α τα θεωρητικά ποσοστιαία σημεία της τυποποιημένης κανονικής κατανομής.

Αν η δειγματική κατανομή και η θεωρητική κατανομή, δηλαδή η κανονική κατανομή, συμφωνούν μεταξύ τους, τότε όλα τα σημεία του γραφήματος θα βρίσκονται συγκεντρωμένα πολύ κοντά σε μία ευθεία. Αυτό είναι ένδειξη ότι τα εσωτερικά τυποποιημένα κατάλοιπα είναι, όντως, κανονικά κατανεμημένα.

Σε αντίθετη περίπτωση, μπορεί να υπάρχουν σημεία που απέχουν πολύ από την ευθεία που απεικονίζεται στο γράφημα. Ειδικότερα, αν τα σημεία απέχουν πολύ από την ευθεία στις ουρές τις κατανομής, δηλαδή στο άνω δεξί και το κάτω αριστερό μέρος του γραφήματος, τότε αυτό είναι ένδειξη ότι τα κατάλοιπα προέρχονται από κάποια κατανομή με πιο παχιές" ουρές από την κανονική κατανομή, όπως η κατανομή t του Student.

### 3.8.3 Έλεγχος ετεροσκεδαστικότητας

Η διασπορά των τυχαίων σφαλμάτων μπορεί ενδεχομένως να είναι γραμμική ή τετραγωνική συνάρτηση μίας ή περισσότερων από τις διαθέσιμες επεξηγηματικές μεταβλητές. Για να ελέγξουμε αυτή τη μορφή ετεροσκεδαστικότητας, μπορούμε να πραγματοποιήσουμε τον έλεγχο Breusch - Pagan. Ο έλεγχος αυτός έχει ως μηδενική υπόθεση ότι τα τυχαία σφάλματα είναι ομοσκεδαστικά, ενώ ως εναλλακτική υπόθεση ότι εμφανίζουν ετεροσκεδαστικότητα. Εφαρμόζουμε τους ελέγχους Breusch - Pagan στο γραμμικό μοντέλο που έχουμε και τασκευάσει και λαμβάνουμε τα $p - value$. Για δεδομένο ε.σ.σ. $\alpha$, έχουμε τα εξής ενδεχόμενα:

- Αν $p - value < \alpha$, τότε απορρίπτουμε την $H_0$, δηλαδή την υπόθεση ότι τα τυχαία σφάλματα είναι ομοσκεδαστικά.
- Αν $p - value < \alpha$, τότε δεν μπορούμε να απορρίψουμε την $H_0$, οπότε τα τυχαία.

Προκειμένου να ελέγξουμε γραφικά από ποιες επεξηγηματικές μεταβλητές θα μπορούσε να εξαρτάται η διασπορά των τυχαίων σφαλμάτων, σχεδιάζουμε γραφήματα των $t_i$ με τις παρατηρήσεις $Z$ από κάθε διαθέσιμη επεξηγηματική μεταβλητή $X$. Αν τα κατάλοιπα έχουν σταθερή διασπορά γύρω από την οριζόντια ευθεία $y = 0$ σε όλα τα γραφήματα, τότε συμπε ραίνουμε ότι τα κατάλοιπα είναι ομοσκεδαστικά. Διαφορετικά, αν η διασπορά τους εμφανίζει κάποια φθίνουσα ή αύξουσα τάσητότε συμπεραίνουμε σαφώς ότι υπάρχει ετεροσκεδαστικότητα.

## 3.9 Ανάλυση διασποράς - ANOVA

Η ανάλυση παλινδρόμησης μελετά τη στατιστική σχέση ανάμεσα σε μία ή περισσότερες ανεξάρτητες μεταβλητές και μια εξαρτημένη μεταβλητή. Συγκεκριμένα, η αναμενόμενη τιμή της εξαρτημένης μεταβλητής εκφράζεται ως γραμμική συνάρτηση των ανεξάρτητων μεταβλητών. Η ανάλυση διακύμανσης

είναι ένα πιο γενικό στατιστικό εργαλείο. Επίσης μελετά τη στατιστική σχέση ανάμεσα σε μία ή περισσότερες ανεξάρτητες μεταβλητές και μια εξαρτημένη μεταβλητή, χωρίς όμως να υποθέτει απαραίτητα κάποιο συγκεκριμένο μοντέλο για την περιγραφή της σχέσης αυτής.

Είδαμε την ανάλυση διακύμανσης στα πλαίσια του γραμμικού μοντέλου. Γενικά η ανάλυση διακύμανσης στα πλαίσια κάποιου παραμετρικού μοντέλου χρησιμοποιείται σαν ένα μετρο καλής προσαρμογής. Δείχνει κατά πόσο η μεταβλητότητα της εξαρτημένης μεταβλητής εξηγείται από το μοντέλο που υποθέσαμε.

Συχνά όμως στην ανάλυση διακύμανσης οι ανεξάρτητες μεταβλητές είναι ποιοτικές (παράγοντες) και το ενδιαφέρον μας εστιάζει στο κατά πόσο ο κάθε παράγοντας και τα επίπεδά του επηρεάζουν κάποια απαντητική μεταβλητή. Η ανάλυση διακύμανσης κατά παράγοντες χρησιμοποιείται πολύ στο σχεδιασμό πειραμάτων.

Τα μοντέλα ανάλυσης διασποράς (ANOVA) αποτελούν ειδική περίπτωση γραμμικών μοντέλων όπου όλες οι διαθέσιμες επεξηγηματικές μεταβλητές για την αποκριτική μεταβλητή είναι ποιοτικές, δηλαδή παίρνουν ένα πεπερασμένο πλήθος διαφορετικών τιμών. Σε αυτήν την περίπτωση, κάθε δυνατός συνδυασμός τιμών των ποιοτικών επεξηγηματικών μεταβλητών ορίζει μία υποομάδα του πληθυσμού από τον οποίο προέρχεται το δείγμα μας. Η μεταβλητή ενδιαφέροντος μπορεί εν γένει να έχει διαφορετική μέση τιμή και διαφορετική διασπορά σε καθεμία από τις υποομάδες που ορίζουν οι επεξηγηματικές μεταβλητές.

Θεωρούμε ότι αποκριτική μεταβλητή έχει κοινή διασπορά σε όλες τις υποομάδες, αλλά διαφορετική μέση τιμή στην καθεμία. Σκοπός μας είναι να διαπιστώσουμε αν όντως τα διαφορετικά επίπεδα του παράγοντα επηρεάζουν τη διαδικασία, δηλαδή αν όντως οι μέσοι με των επιπέδων διαφέρουν, ή αν ο μέσος είναι σταθερός για όλα τα επίπεδα του παράγοντα. Στο ερώτημα αυτό θα απαντήσουμε χρησιμοποιώντας στατιστικά εργαλεία, βασιζόμενοι σε ένα τυχαίο δείγμα τιμών της απαντητικής μεταβλητής από τα διάφορα επίπεδα του παράγοντα.

## 3.10 Έλεγχος ισότητας μέσων

Έλεγχος Ισότητας Μέσων Η υπόθεση που μας ενδιαφέρει να ελέγξουμε είναι $H_0 : \mu_i = \mu, i = 1, \ldots, m$ vs. $H_1 : \mu_i \neq \mu$, για ένα τουλάχιστον i

Για να ελέγξουμε την παραπάνω υπόθεση θα αναλύσουμε τη συνολική διασπορά των δεδομένων σε δύο συνιστώσες: τη διασπορά μέσα στα επίπεδα και τη διασπορά ανάμεσα στα επίπεδα. Διαισθητικά περιμένουμε ότι αν η διασπορά ανάμεσα στα επίπεδα είναι μεγαλύτερη από τη διασπορά μέσα στα επίπεδα, τότε ο μέσος δεν θα είναι σταθερός για όλα τα επίπεδα.

Αποδεικνύεται ότι

$$SST = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{..})^2 = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{i.})^2 + \sum_{i=1}^{m} n_i(\bar{Y}_{i.} - \bar{Y}_{..})^2,$$

όπου το άθροισμα των τετραγώνων των αποκλίσεων των παρατηρήσεων από τους αντίστοιχους μέσους των επιπέδων στα οποία ανήκουν, $SSE = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{i.})^2$, εκφράζει τη μεταβλητότητα μέσα στα επίπεδα του παράγοντα, και το άθροισμα των τετραγώνων των αποκλίσεων των μέσων των επιπέδων από το συνολικό μέσο του δείγματος, $SSF = \sum_{i=1}^{m} n_i(\bar{Y}_{i.} - \bar{Y}_{..})^2$, εκφράζει τη μεταβλητότητα ανάμεσα στα επίπεδα. Δηλαδή,

$$SST = SSE + SSF.$$

Η μεταβλητότητα ανάμεσα στα επίπεδα είναι αυτή που οφείλεται στις διαφορές των επιπέδων του παράγοντα και εκφράζεται από το συνολικό άθροισμα τετραγώνων του παράγοντα (factor sum of squares). Η μεταβλητότητα μέσα στα επίπεδα οφείλεται στην τυχαιότητα και εκφράζεται από το συνολικό άθροισμα τετραγώνων των τυχαίων σφαλμάτων (error sum of squares).

Οι συνολικοί βαθμοί ελευθερίας στο πείραμα είναι $n-1$. Μέσα σε κάθε επίπεδο οι βαθμοί ελευθερίας είναι $n_i - 1$, άρα συνολικά μέσα στα επίπεδα έχουμε $n - m$ βαθμούς ελευθερίας. Οι βαθμοί ελευθερίας ανάμεσα στα επίπεδα είναι $m - 1$.

Έχουμε:

$$\frac{SST}{\sigma^2} = \frac{SSE}{\sigma^2} + \frac{SSF}{\sigma^2}.$$

Από το θεώρημα τετραγωνικών μορφών, κάτω από την $H_0$, η συνάρτηση $\frac{SSF}{\sigma^2}$ ακολουθεί [2] κατανομή με $m-1$ βαθμούς ελευθερίας. Άρα

$$E\left[\frac{SSF}{\sigma^2}\right] = m-1 \Rightarrow E\left[\frac{SSF}{m-1}\right] = \sigma^2,$$

δηλαδή η συνάρτηση $\frac{SSF}{m-1}$, κάτω από την $H_0$, είναι μια αμερόληπτη εκτιμήτρια της διασποράς $\sigma^2$.

Επομένως, κάτω από την $H_0$, έχουμε ότι

$$\frac{\frac{SSF}{m-1}}{\frac{SSE}{n-m}} \sim F_{m-1,n-m}.$$

Η ελεγχοσυνάρτηση συνάρτηση $F_0 = \frac{\frac{SSF}{m-1}}{\frac{SSE}{n-m}}$ μπορεί να χρησιμοποιηθεί για τον έλεγχο ισότητας μέσων στην ανάλυση διασποράς. Απορρίπτουμε την αρχική υπόθεση $H_0$ σε επίπεδο στατιστικής σημαντικότητας $\alpha$ αν η παρατηρούμενη τιμή της ελεγχοσυνάρτησης συνάρτησης είναι μεγαλύτερη από το α ποσοστιαίο σημείο της κατανομής $F$ με $m-1$ και $n-m$ βαθμούς ελευθερίας, δηλαδή αν

$$F_0 = \frac{\frac{SSF}{m-1}}{\frac{SSE}{n-m}} > F_{a,m-1,n-m}.$$

(Ισοδύναμο $p$ - value $< \alpha$).

Ακραία (στην ουρά της $F$ κατανομής) παρατηρούμενη τιμή της ελεγχοσυνάρτησης συνάρτησης σημαίνει ότι μεγάλο μέρος της συνολικής μεταβλητότητας των δεδομένων οφείλεται στους παράγοντες (μεταβλητότητα ανάμεσα στα επίπεδα - $SSF$) συγκριτικά με το μέρος που οφείλεται στους τυχαίους όρους (μεταβλητότητα μέσα στα επίπεδα - $SSE$). Αυτό αποτελεί ένδειξη εναντίον της $H_0$.

# 4 Αποτελέσματα

Στην ενότητα αυτή θα παρουσιαστούν τα αποτελέσματα των Ενοτήτων 2 και 3, δηλαδή της ανάλυσης του γραμμικού μοντέλου με εξαρτημένη μεταβλητή τον Έλεγχος Διαφθοράς (Corruption Control – CorControl),

## 4.1 Ανάλυση πλήρους γραμμικού μοντέλου

Αρχικά, εκτιμήθηκε ένα πλήρες γραμμικό μοντέλο με εξαρτημένη μεταβλητή τον Έλεγχος Διαφθοράς (CorControl) και εξαρτημένες με εξαίρεση τις μεταβλητές Πολιτική Σταθερότητα (PolStab) και Φωνή και Υπευθυνότητα (Voice Acc) (και εννοείται εξαιρώντας τη μεταβλητή Country). Παρακάτω βρίσκονται τα αποτελέσματα:

| **Ποσοστιμόριο** | **Τιμή** |
|---|---|
| Min | -0.85666 |
| 1Q | -0.28355 |
| Median | -0.02637 |
| 3Q | 0.24556 |
| Max | 1.13595 |

Πίνακας 7: Ποσοστιμόρια Καταλοίπων πλήρους γραμμικού μοντέλου

| **Μέτρο** | **Τιμή** |
|---|---|
| Residual standard error | 0.4595 on 93 d.f. |
| Multiple R-squared | 0.8228 |
| Adjusted R-squared | 0.7733 |
| F-statistic | 16.61 on 26 and 93 DF |
| p-value | < 2.2e-16 |

Πίνακας 8: Στατιστικά μέτρα ανάλυσης

***Σχολιασμός.*** Το ποσοστό της συνολικής μεταβλητότητας των $Y_i$ που οφείλεται στην παλινδρόμηση, το οποίο εκφράζεται από το $R^2$, είναι 82.28%.

Το $R^2_{adj}$ μετρά ποσοστό της συνολικής μεταβλητότητας των $Y_i$ που οφείλεται στην παλινδρόμηση, αλλά προσαρμόζει το $R^2$ για τον αριθμό των προβλεπτικών παραγόντων στο μοντέλο.

Παρέχει ένα ακριβέστερο μέτρο του πόσο καλά το μοντέλο εξηγεί τη μεταβλητότητα της μεταβλητής αποτελέσματος. Όπως περιμέναμε, είναι χαμηλότερο από το $R^2$, εφόσον το μοντέλο μας έχει πολλές ανεξάρτητες μεταβλητές.

| **Ανεξάρτητη μεταβλητή** | **Εκτίμηση** | **Τυπ. Απόκλιση** | **t value** | **Pr(>\|t\|)** |
|---|---|---|---|---|
| (Intercept) | -2.137e+00 | 3.003e+00 | -0.712 | 0.4784 |
| ElectrAccess | -9.999e-03 | 4.859e-03 | -2.058 | 0.0424 * |
| CookAccess | 1.065e-03 | 4.051e-03 | 0.263 | 0.7932 |
| AgriLand | -2.645e-03 | 2.619e-03 | -1.010 | 0.3153 |
| BirthRate | -5.516e-02 | 6.915e-02 | -0.798 | 0.4271 |
| CO2 | -1.750e-02 | 1.856e-02 | -0.943 | 0.3480 |
| CompEdu | -4.110e-02 | 2.508e-02 | -1.639 | 0.1046 |
| DeathRate | -1.034e-02 | 4.540e-02 | -0.228 | 0.8203 |
| FoodExports | 8.692e-04 | 2.346e-03 | 0.371 | 0.7118 |
| Telephone | 5.740e-03 | 4.978e-03 | 1.153 | 0.2518 |
| Internet | 1.248e-02 | 6.148e-03 | 2.029 | 0.0453 * |
| PopGrowth | -1.412e-01 | 8.569e-02 | -1.648 | 0.1028 |
| BusinessTime | -1.464e-03 | 2.713e-03 | -0.540 | 0.5908 |
| AdolFertRate | -1.233e-02 | 4.881e-03 | -2.526 | 0.0132 * |
| Precipitation | -4.492e-08 | 8.064e-05 | -0.001 | 0.9996 |
| EmployerPerc | 7.618e-03 | 2.323e-02 | 0.328 | 0.7437 |
| FertRate | 1.534e-01 | 3.277e-01 | 0.468 | 0.6408 |
| GDPperc | 2.725e-03 | 2.512e-02 | 0.108 | 0.9138 |
| GDPdollars | 2.670e-05 | 4.284e-06 | 6.232 | 1.33e-08 *** |
| Inflation | -1.533e-02 | 1.511e-02 | -1.015 | 0.3129 |
| WaterStress | -5.844e-05 | 1.505e-04 | -0.388 | 0.6987 |
| LifeExp | 1.342e-02 | 3.230e-02 | 0.416 | 0.6787 |
| MortRate | 4.353e-03 | 7.356e-03 | 0.592 | 0.5555 |
| PopPerc14 | 6.442e-02 | 3.859e-02 | 1.669 | 0.0984 . |
| WomBusiness | 7.241e-03 | 4.503e-03 | 1.608 | 0.1112 |
| FertCat | 2.863e-01 | 2.127e-01 | 1.346 | 0.1815 |
| InflCat | 4.377e-02 | 1.191e-01 | 0.367 | 0.7141 |

Πίνακας 9: Ανάλυση πλήρους γραμμικού μοντέλου

## 4.2 Ανάλυση βέλτιστου γραμμικού μοντέλου

Έπειτα, εφαρμόσαμε τη διαδικασία βηματικής επιλογής προς τα πίσω (backwards) για την επιλογή του βέλτιτσου μοντέλου με τη χρήση του κριτηρίου AIC. Τα αποτελέσματα βρίσκονται παρακάτω:

| Ποσοστιμόριο | Τιμή |
|---|---|
| Min | -0.94538 |
| 1Q | -0.34818 |
| Median | -0.01533 |
| 3Q | 0.28638 |
| Max | 1.26908 |

Πίνακας 10: Ποσοστιμόρια Καταλοίπων πλήρους γραμμικού μοντέλου

| Μέτρο | Τιμή |
|---|---|
| Residual standard error | 0.4426 on 109 d.f. |
| Multiple R-squared | 0.8073 |
| Adjusted R-squared | 0.7896 |
| F-statistic | 45.66 on 10 and 109 DF |
| p-value | < 2.2e-16 |

Πίνακας 11: Στατιστικά μέτρα ανάλυσης

| Ανεξάρτητη μεταβλητή | Εκτίμηση | Τυπ. Απόκλιση | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | -8.472e-01 | 6.708e-01 | -1.263 | 0.20933 |
| ElectrAccess | -1.114e-02 | 3.619e-03 | -3.079 | 0.00263** |
| AgriLand | -3.633e-03 | 2.227e-03 | -1.631 | 0.10569 |
| CO2 | -2.564e-02 | 1.354e-02 | -1.893 | 0.06097 . |
| Internet | 1.305e-02 | 4.690e-03 | 2.783 | 0.00634 ** |
| PopGrowth | -1.236e-01 | 6.079e-02 | -2.034 | 0.04440 * |
| AdolFertRate | -1.346e-02 | 4.175e-03 | -3.224 | 0.00167 ** |
| GDPdollars | 2.933e-05 | 3.586e-06 | 8.179 | 5.74e-13 *** |
| PopPerc14 | 2.624e-02 | 1.349e-02 | 1.945 | 0.05440 |
| WomBusiness | 6.641e-03 | 3.267e-03 | 2.033 | 0.04449 |
| FertCat | 2.630e-01 | 1.896e-01 | 1.387 | 0.16836 |

Πίνακας 12: Ανάλυση πλήρους γραμμικού μοντέλου

Η σύγκριση των τιμών πληροφορίας AIC για τα 2 μοντέλα είναι:

| Μοντέλο | Τιμή AIC |
|---|---|
| Πλήρες μοντέλο | 179.3194 |
| Βέλτιστο κατά AIC | 157.3954 |

Πίνακας 13: Σύγκριση τιμών πληροφορίας AIC για τα 2 μοντέλα

***Σχολιασμός.*** Παρατηρούμε ότι χρησιμοποιώτνας τη μέθοδο βηματικής επιλογής μοντέλου προς τα πίσω, χρησιμοποιώντας το κριτήριο πληροφορίας AIC, επιλέχθηκαν λιγότερες ανεξάρτητες μεταβλητές για την ερμηνεία του μοντέλου, τα σφάλματα απλώθηκαν περισσότερο, κρίνωντας από τα ποσοστημόρια για τα κατάλοιπα (Πίνακας 7 vs. Πίνακας 10).

Αναφορικά με τους συντελεστές προσδιορισμού των μοντέλων, παρατηρούμε ότι μειώθηκε το $R^2$, ενώ αυξήθηκε το $R^2_{adj}$. Αυτό είναι λογικό καθώς όπως ο συντελεστής προσδιορισμού, έτσι και η μέγιστη πιθανότητα του γραμμικού μοντέλου αυξάνεται συνεχώς όσο προστίθενται καινούργιες επεξηγηματικές μεταβλητές μέσω του μοντέλου. Επίσης η αύξηση του $R^2_{adj}$ στο μοντέλο που προέκυψε, πιστοποιεί ότι το καινούριο μοντέλο εξηγεί μεγαλύτερο ποσοστό της μεταβλητότητας των $Y_i$ που οφείλεται στην παλινδρόμηση συμπεριλαμβάνοντας τον αριθμό των ανεξάρτητων μεταβλητών στο μοντέλο.

Γνωρίζουμε ότι καλύτερο μοντέλο είναι αυτό που επιτυγχάνει τη μικρότερη δυνατή τιμή $AIC_p$ και αυτό συμφωνεί με τα ευρήματά μας (Πίνακας 13), αφού το καινούριο μοντέλο, που θεωρείται καλύτερο σύμφωνα με το κριτήριο του $R^2_{adj}$.

### 4.3 Έλεγχος στατιστικής σημαντικότητας συντελεστών βέλτιστου γραμμικού μοντέλου

Για τον έλεγχο στατιστικής σημαντικότητας των συντελεστών των `"AgriLand"` και `"GDPdollars"` φτιάχτηκε η συνάρτηση `t_test`, η οποία υπολογίζει για δοσμένη μεταβλητή την τιμή της ελεγχοσυνάρτησης T ως `t_value <- bhat[var]/bhatse[var]` και συγκρίνει την απόλυτη τιμή της (διπλός έλεγχος) με το αντίστοιχο ποσοστημόριο της $t_{n-p}$. (Εννοείται ότι θα μποορύσαν να χρησιμοποιηθούν κατευθείαν τα $p-values$ των συντελεστών από τον συνολικό πίνακα του μοντέλου). Τα αποτελέσματα είναι τα εξής:

- `> t_test(X, bhat, bhatse, "AgriLand", 0.05)`

  ```
  We fail to reject the null hypothesis H0 for the coefficient of AgriLand at
  significance level alpha = 0.05 so the coefficient is not statistically
  important.
  ```

- `> t_test(X, bhat, bhatse, "GDPdollars", 0.05)`

  ```
  We reject the null hypothesis H0 for the coefficient of GDPdollars in favor
  of the alternative hypothesis H1 at significance level alpha = 0.05 , so the
  coefficient is statistically important.
  ```

***Σχολιασμός.*** Ο έλεγχο στατιστικής σημαντικότητας συντελεστών για τις ανεξάρτητες μεταβλητές `"AgriLand"` και `"GDPdollars"` είχε τα αποτελέσματα ότι:

- Αποδεχόμαστε την $_0$ σε επίπεδο στατιστικής σημαντικότητας $\alpha = 0.05$ του ελέγχου $H_0 : \beta_{AgriLand} = 0$ vs. $H_1 : \beta_{AgriLand} \neq 0$, αποτέλεσμα που συμφωνεί με το $p-value$ του Πίνακα 12.
- Από την άλλη, απορρίπτουμε την $_0$ σε επίπεδο στατιστικής σημαντικότητας $\alpha = 0.05$ του ελέγχου $H_0 : \beta_{GDPdollars} = 0$ vs. $H_1 : \beta_{GDPdollars} \neq 0$, αποτέλεσμα που συμφωνεί με το $p-value$ του Πίνακα 12.

### 4.4 Διαστήματα εμπιστοσύνης συντελεστώς βέλτιστου γραμμικού μοντέλου

Τα διαστήματα εμπιστοσύνης που προέκυψαν είναι:

| Ανεξάρτητη μεταβλητή | Κάτω άκρο | Άνω άκρο |
|---|---|---|
| (Intercept) | -2.605895e+00 | 9.115475e-01 |
| ElectrAccess | -2.062889e-02 | -1.654203e-03 |
| AgriLand | -9.470325e-03 | 2.204987e-03 |
| CO2 | -6.115005e-02 | 9.865715e-03 |
| Internet | 7.588021e-04 | 2.534786e-02 |
| PopGrowth | -2.830260e-01 | 3.573709e-02 |
| AdolFertRate | -2.440776e-02 | -2.514836e-03 |
| GDPdollars | 1.992698e-05 | 3.872789e-05 |
| PopPerc14 | -9.136330e-03 | 6.161234e-02 |
| WomBusiness | -1.923583e-03 | 1.520646e-02 |
| FertCat | -2.342096e-01 | 7.601732e-01 |

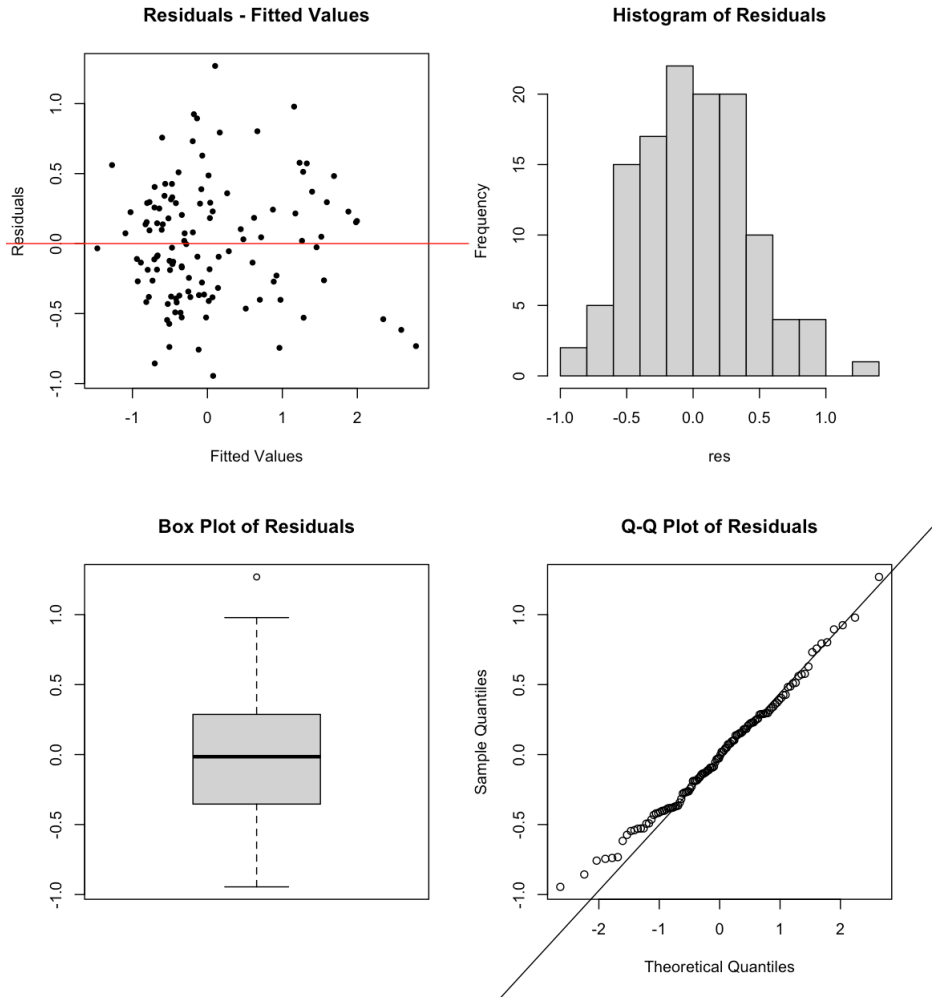
Πίνακας 14: Διαστήματα εμπιστοσύνης βέλτιστου μοντέλου

***Σχολιασμός.*** Επιστάμε την προσοχή μας στο διάστημα εμπιστοσύνης 99% των ανεξάρτητων μεταβλητών `"AgriLand"` και `"GDPdollars"`. Παρατηρούμε ότι όταν γίνουν τόσο στενά τα πλαίσια θεώρησης στατιστικά σημαντικού συντελεστή, πλέον και ο συντελεστής της μεταβλητής `"GDPdollars"` καθίσταται στατιστικά ασήμαντος (το 0 δεν περιέχεται στο διάστημα εμπιστοσύνης του) και φυσικά το ίδιο συμβαίνει και για τον συντελεστή της ανεξάρτητης μεταβλητής `"AgriLand"` (λογικό αφού δεν είναι καν στατιστικά σημαντικός σε ε.σ.σ. $\alpha = 0.05$).

## 4.5 Έλεγχος υποθέσεων μοντέλου μέσω ελέγχου καταλοίπων

### 4.5.1 Γραφικός έλεγχος καταλοίπων

Τα γράφημα καταλοίπων-προσαρμοσμένων τιμών, ιστόγραμμα, θηκόγραμμα και Q-Q Plot για τα κατάλοιπα του βέλτιστου μοντέλου, βρίσκεται παρακάτω:




Σχήμα 6: Γραφήματα καταλοίπων για το βέλτιστο μοντέλο

***Σχολιασμός.*** Από το πρώτο γράφημα φαίνεται ότι τα κατάλοιπα έχουν σταθερή διασπορά γύρω από την οριζόντια ευθεία $y = 0$ και δεν εμφανίζουν κάποια αύξουσα ή φθίνουσα τάξη, δεν εμφανίζεται δηλαδή κάτι συστηματικό σε όλα τα γραφήματα. Φαίνεται λοιπόν πως είναι ομοσκεδαστικά

Το ιστόγραμμα και το θηκόγραμμα συμφωνεί με την καμπύλη της τυποποιημένης κανονικής κατανομής $N(0, 1)$ και αυτό είναι ένδειξη ότι τα κατάλοιπα είναι κανονικά κατανεμημένα. Επίσης, στο Q-Q Plot όλα τα σημεία του γραφήματος θα βρίσκονται συγκεντρωμένα πολύ κοντά σε μία ευθεία. Αυτό είναι ένδειξη ότι τα εσωτερικά τυποποιημένα κατάλοιπα είναι, όντως, κανονικά κατανεμημένα.

### 4.5.2 Στατιστικός έλεγχος καταλοίπων

Οι στατιστικοί έλεγχοι για τα κατάλοιπα δίνουν:

| **Breusch-Pagan test** | **Τιμή** |
|---|---|
| BP | 15.321 |
| df | 10 |
| $p-value$ | 0.1208 |

Πίνακας 15: Breusch-Pagan test για ομοσκεδαστικότητα

| **Shapiro-Wilk test** | **Τιμή** |
|---|---|
| W | 0.99029 |
| $p-value$ | 0.5603 |

Πίνακας 16: Shapiro-Wilk test για κανονικότητα

***Σχολιασμός.*** Τις εικασίες μας για τα κατάλοιπα έρχονται να επαληθεύσουν οι στατιστικοί έλεγχοι που αφορούν τα κατάλοιπα. Πράγματι, το $p-value$ Breusch-Pagan test για ομοσκεδαστικότητα είναι μεγαλύτερο από $0.1$ και άρα δεν μπορούμε να απορρίψουμε την $H_0$, οπότε τα τυχαία σφάλματα μπορούμε να θεωρήσουμε ότι είναι ομοσκεδαστικά.

Επιπρόσθετα,

Αν το $p-value$ του Shapiro-Wilk test για κανονικότητα είναι μεγαλύτερο από $0.1$ και δεν μπορούμε να απορρίψουμε την $H_0$, δηλαδή την υπόθεση ότι τα εσωτερικά τυποποιημένα κατάλοιπα είναι κανονικά κατανεμημένα. Μπορούμε, λοιπόν να δεχτούμε ότι το γραμμικό μοντέλο που έχουμε κατασκευάσει έχει τυχαία σφάλματα που προέρχονται από την κανονική κατανομή.

Επίσης, εφόσον το δείγμα μας είναι τυχαίο, μπορούμε να θεωρήσουμε ότι είναι ασυσχέτιστα τα κατάλοιπα και εφόσον αυτά είναι κανονικά κατανεμημένα, ισοδύναμα, είναι και ανεξάρτητα. Άρα οι υποθέσεις του μοντέλου ισχύουν και δε χρειάζεται να δοκιμάσουμε κάποιο μετασχηματισμό στις ανεξάρτητες μεταβλητές.

## 4.6 Πρόβλεψη για νέα δεδομένα

Αξιοποιώντας τις εκτιμήσεις των συντελεστών του βέλτιστου μοντέλου, μπορούμε να κάνουμε προβλέψεις για νέα δεδομένα.

Παρακάτω βρίσκεται η σημειακή πρόβλεψη για την εξαρτημένη μεταβλητή Έλεγχος Διαφθοράς (Corruption Control – CorControl) καθεμίας από τις 8 χώρες:

| Cyprus | Georgia | Greece | Luxemburg | Mexico | Philippines | Slovenia | Uruguay |
|---|---|---|---|---|---|---|---|
| 0.8863817 | -0.3291076 | 0.4393572 | 3.2028723 | -0.3314785 | -0.4205245 | 0.7223933 | 0.1084105 |

Πίνακας 17: Σημειακή πρόβλεψη για τη μεταβλητή Έλεγχος Διαφθοράς (Corruption Control – CorControl) κάθε χώρας.

Παρακάτω βρίσκονται τα μέσα διαστήματα πρόβλεψης για την εξαρτημένη μεταβλητή Έλεγχος Διαφθοράς (Corruption Control – CorControl) καθεμίας από τις 8 χώρες:

| Χώρα | Σημειακή πρόβλεψη | **95% Μέσο Διάστημα Πρόβλεψης** |
|---|---|---|
| Cyprus | 0.8863817 | (0.7068971, 1.0658664) |
| Georgia | -0.3291076 | (-0.5043345, -0.1538807) |
| Greece | 0.4393572 | (0.2593801, 0.6193342) |
| Luxemburg | 3.2028723 | (2.6563054, 3.7494393) |
| Mexico | -0.3314785 | (-0.5282404, -0.1347165) |
| Philippines | -0.4205245 | (-0.6668829, -0.1741661) |
| Slovenia | 0.7223933 | (0.5652714, 0.8795153) |
| Uruguay | 0.1084105 | (-0.2202040, 0.4370249) |

Πίνακας 18: Μέσα διαστήματα πρόβλεψης για την εξαρτημένη μεταβλητή Έλεγχος Διαφθοράς (Corruption Control – CorControl) καθεμίας από τις 8 χώρες.

Παρακάτω βρίσκονται τα ατομικά διαστήματα πρόβλεψης για την εξαρτημένη μεταβλητή Έλεγχος Διαφθοράς (Corruption Control – CorControl) καθεμίας από τις 8 χώρες:

| Χώρα | Σημειακή πρόβλεψη | 95% Ατομικό Διάστημα Πρόβλεψης |
|---|---|---|
| Cyprus | 0.8863817 | (-0.009041873, 1.7818053) |
| Georgia | -0.3291076 | (-1.223687467, 0.5654723) |
| Greece | 0.4393572 | (-0.456165284, 1.3348796) |
| Luxemburg | 3.2028723 | (2.169284339, 4.2364603) |
| Mexico | -0.3314785 | (-1.230524590, 0.5675677) |
| Philippines | -0.4205245 | (-1.331711176, 0.4906622) |
| Slovenia | 0.7223933 | (-0.168817063, 1.6136037) |
| Uruguay | 0.1084105 | (-0.828369186, 1.0451901) |

Πίνακας 19: Ατομικά διαστήματα πρόβλεψης για την εξαρτημένη μεταβλητή Έλεγχος Διαφθοράς (Corruption Control – CorControl) καθεμίας από τις 8 χώρες.

## 4.7 Ανάλυση διασποράς ANOVA της CorControl με εξαρτημένες μεταβλητές-παράγοντες FertCat και InflCat

### 4.7.1 Υποθέσεις του μοντέλου ANOVA

Έστω ότι το αποτέλεσμα ενός πειράματος εξαρτάται από 2 παράγοντες, FertCat και InflCat. Αρχικά υποθέτουμε ότι δεν υπάρχει αλληλεπίδραση (interaction) μεταξύ των παραγόντων. Θα μελετήσουμε τον παράγοντα FertCat σε $m = 3$ επίπεδα και τον παράγοντα InflCat σε $l = 2$ επίπεδα. Επομένως υπάρχουν $ml$ συνδυασμοί επιπέδων (treatments) για τους οποίους λαμβάνουμε παρατηρήσεις. Με χρήση κώδικα συμπεραίνουμε ότι έχουμε παραπάνω από μία παρατήρηση για κάθε treatment και θα προσαρμόσουμε μοντέλο κατά 2 παράγοντες με αλληλεπίδραση.

Το μοντέλο Anova που χρησιμοποιούμε είναι :

$$Y_{ijr} = \mu_{ij} + \varepsilon_{ijr} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijr},$$

όπου:

- $\sum_{i=1}^{m} \alpha_i = 0$,
- $\sum_{j=1}^{l} \beta_j = 0$,
- $\sum_{i=1}^{m} \alpha_i = 0$ και $\sum_{j=1}^{l} \gamma_j = 0$.

Οι υποθέσεις μας είναι για τα κατάλοιπα είναι:

- κανονικότητα
- ομοσκεδαστικότητα
- τυχαιότητα/ανεξαρτησία

### 4.7.2 ANOVA Table

| Πηγή Μεταβλητότητας | B.E | SS | Mean Sq | F-value ($f^*$) | $Pr(F > f^*)$ |
|---|---|---|---|---|---|
| factor(FertCat) | 2 | 30.27 | 15.135 | 24.081 | 1.89e-09 *** |
| factor(InflCat) | 1 | 5.22 | 5.217 | 8.301 | 0.00473 ** |
| factor(FertCat):factor(InflCat) | 2 | 3.68 | 1.840 | 2.927 | 0.05758 . |
| Residuals | 114 | 71.65 | 0.628 | | |

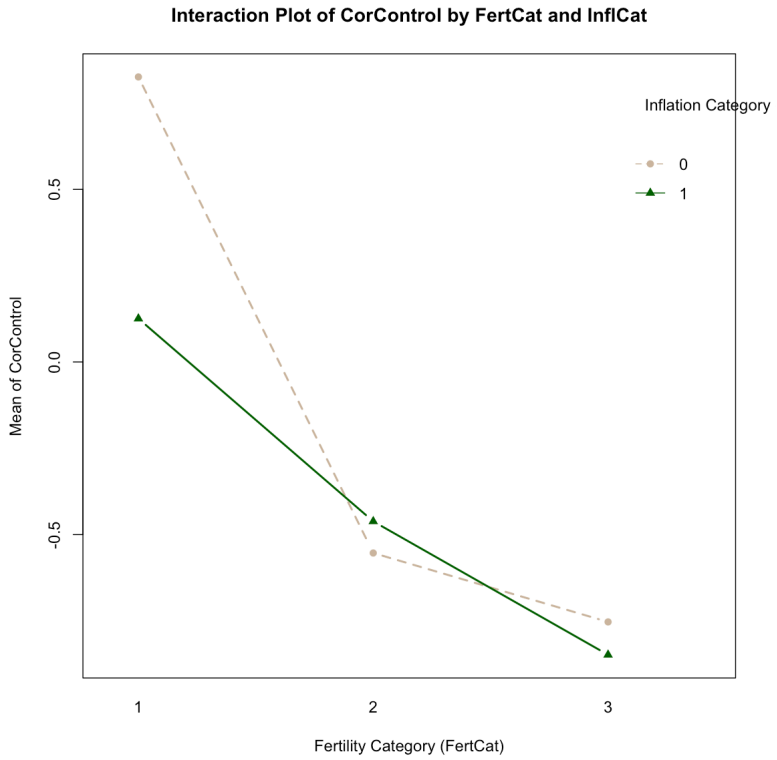
Πίνακας 20: ANOVA Table για μοντέλο με αλληλεπίδραση των παραγόντων FertCat και InflCat στην εξαρτημένη μεταβλητή Έλεγχος Διαφθοράς (Corruption Control – CorControl)

***Σχολιασμός.*** Από τα $p - values$ του ANOVA Table για το μοντέλο με αλληλεπίδραση των παραγόντων FertCat και InflCat στην εξαρτημένη μεταβλητή Έλεγχος Διαφθοράς (Corruption Control – CorControl), συμπεραίνουμε ότι οι επιδράσεις των μεταβλητών `"FertCat"` και `"InflCat"` είναι στατιστικώς σημαντικές σε ε.σ.σ. $\alpha = 0.01$ και ότι η αλληλεπίδραση δεν είναι στατιστικά σημαντική σε ε.σ.σ $\alpha = 0.05$.

## 4.8 Έλεγχος στατιστικής σημαντικότητας αλληλεπίδρασης και κύριων επιδράσεων των παραγόντων

Αρχικά, θα εξετάσουμε την ύπαρξη στατιστικά σημαντικής αλληλεπίδρασης των παραγόντων και έπειτα των επιμέρους παραγόντων.

Για τον έλεγχο της αλληλεπίδρασης των 2 παραγόντων, χρησιμοποιήσαμε το $p-value$ που προκύπτει, το οποίο είναι $p-value = 0.05758 > 0.05$, άρα δεν έχω επαρκή δεδομένα για να απορρίψω την αρχική, ότι δηλαδή η αλληλεπίδραση των παραγόντων είναι στατιστικά ασήμαντη.




Σχήμα 7: Γράφημα αλληλεπίδρασης παραγόντων FertCat και InflCat στην εξαρτημένη μεταβλητή Έλεγχος Διαφθοράς (Corruption Control – CorControl).

Παρόλο που βλέπουμε σημεία τομής μεταξύ των καμπυλών, οι σχεδιασμένες καμπύλες είναι σχεδόν παράλληλες μεταξύ τους, άρα συμπεραίνουμε σωστά ότι δεν υπάρχει σημαντική αλληλεπίδραση.

Για τον έλεγχο στατιστικής σημαντικότητας των συντελεστών των παραγόντων `FertCat` και `InflCat`, ελέγχοντας την τιμή της αντίστοιχης ελεγχοσυνάρτησης $F$ με το αντίστοιχο ποσοστιμόριο, λάβαμε τα εξής αποτελέσματα:

```
We reject the null hypothesis H0 at the level of statistical significance a=0.05,
so the means of different levels of FertCat are identical, so the value of CorControl
is dependent on the level of factor FertCat
```

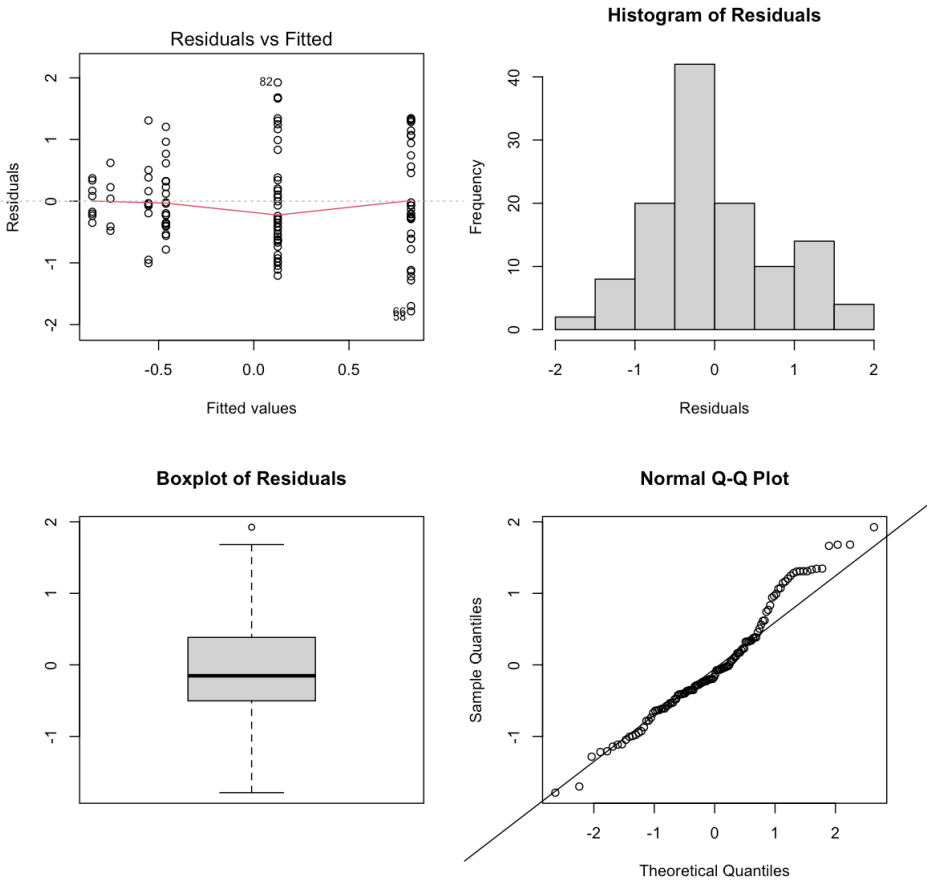
```
We reject the null hypothesis H0 at a level of statistical significance a=0.05,
so the means of different levels of InflCat are identical, so the value of CorControl
is dependent on the level of factor InflCat
```

***Σχολιασμός.*** Από τα $p - values$ του ANOVA Table για το μοντέλο με αλληλεπίδραση των παραγόντων FertCat και InflCat στην εξαρτημένη μεταβλητή Έλεγχος Διαφθοράς (Corruption Control – CorControl), συμπεραίνουμε ότι οι επιδράσεις των μεταβλητών `"FertCat"` και `"InflCat"` είναι στατιστικώς σημαντικές σε ε.σ.σ. $\alpha = 0.01$ και ότι η αλληλεπίδραση δεν είναι στατιστικά σημαντική σε ε.σ.σ $\alpha = 0.05$.

### 4.8.1 Έλεγχος υποθέσεων μοντέλου μέσω ελέγχου καταλοίπων

### 4.8.2 Γραφικός έλεγχος καταλοίπων

Τα γράφημα καταλοίπων-προσαρμοσμένων τιμών, ιστόγραμμα, θηκόγραμμα και Q-Q Plot για τα κατάλοιπα του βέλτιστου μοντέλου, βρίσκεται παρακάτω:




Σχήμα 8: Γραφήματα καταλοίπων για το βέλτιστο μοντέλο

***Σχολιασμός.*** Στο Residual vs Fitted values, παρατηρούμε ότι τα κατάλοιπα δεν είναι ομοσκεδαστικά. Στο

ιστόγραμμα, βλέπουμε ότι τα κατάλοιπα δεν είναι συμμετρικά γύρω από το μηδέν, όπως θα έπρεπε, ενώ υπάρχουν κιόλας κάποια κατάλοιπα τα οποία είναι υπερβολικά απομακρυσμένα από το μηδέν προς τα θετικά, πράγμα το οποίο δε συμφωνεί με την κανονική κατανομή. Στο Θηκόγραμμα, υπάρχουν κάποιες παρατηρήσεις εκτός ορίων των ορίων των απολήξεων, ένδειξη ετεροσκεδαστικότητας των καταλοίπων.

Για να έχουμε κανονική κατανομή, θα έπρεπε στο Q-Q Plot τα σημεία του γραφήματος να βρίσκονται συγκεντρωμένα πολύ κοντά στην κόκκινη ευθεία, ενώ εμείς παρατηρούμε ότι υπάρχουν σημεία που απέχουν πολύ από την ευθεία που απεικονίζεται στο γράφημα. Ειδικότερα, τα σημεία απέχουν πολύ από την ευθεία στις ουρές τις κατανομής, δηλαδή στο άνω δεξί, τότε αυτό είναι ένδειξη ότι τα κατάλοιπα προέρχονται από κάποια κατανομή με πιο "παχιές" ουρές από την κανονική κατανομή, πιθανώς μία t-Student. Συνολικά, φαίνεται ότι η κατανομή δεν ακολουθεί κανονική κατανομή.

### 4.8.3 Στατιστικός έλεγχος καταλοίπων

Οι στατιστικοί έλεγχοι για τα κατάλοιπα δίνουν:

| | **Df** | **F value** | **Pr(>F)** |
|---|---|---|---|
| **Group** | 5 | 2.7847 | 0.02067 * |
| **Residuals** | 114 | | |

Πίνακας 21: Levene's Test για ομοσκεδαστικότητα (center = median)

| **Shapiro-Wilk test** | **Τιμή** |
|---|---|
| W | 0.96924 |
| $p-value$ | 0.007526 |

Πίνακας 22: Shapiro-Wilk test για κανονικότητα

***Σχολιασμός.*** Τις εικασίες μας για τα κατάλοιπα έρχονται να επαληθεύσουν οι στατιστικοί έλεγχοι που αφορούν τα κατάλοιπα. Πράγματι, το $p-value$ Levene's test για ομοσκεδαστικότητα είναι μικρότερο από $0.05$ και απορρίπτουμε την $H_0$ σε ε.σ.σ. $\alpha = 0.05$ οπότε τα τυχαία σφάλματα θεωρούμε ότι είναι ετεροσκεδαστικά.

Επιπρόσθετα,

Αν το $p-value$ του Shapiro-Wilk test για κανονικότητα είναι μικρότερο από $0.1$ και απορρίπτουμε την $H_0$ σε ε.σ.σ. $\alpha = 0.01$ οπότε τα τυχαία σφάλματα θεωρούμε ότι δεν προέρχονται από την κανονική κατανομή.

Συνολικά οι υποθέσεις του μοντέλου μας δεν ισχύουν και άρα χρειάζεται κάποιο διαφορετικό μοντέλο για να περιγράψει την εξάρτηση της μεταβλητής Έλεγχος Διαφθοράς (Corruption Control – CorControl) από τις ανεξάρτητες κατηγορικές μεταβλητές `"FertCat"` και `"Infl"`.

# 5 Συζήτηση

## 5.1 Ερμηνεία των συντελεστών του γραμμικού μοντέλου

Το βέλτσιο μοντέλο που προέκυψε από τη μέθοδο backward stepwise χρησιμοποιώτνας το κριτήριο AIC αυτό θεωρεί ότι υπάρχει γραμμική εξάρτηση της μεταβλητής Έλεγχος Διαφθοράς (Corruption Control – CorControl) και των μεταβλητών που βρίσκονται στην 1η στήλη του Πίνακα 12.

Σύμφωνα με το μοντέλο όσες μεταβλητές έχουν αρνητική εκτίμηση, σημαίνει ότι όσο αυτές αυξάνονται, τότε ο Έλεγχος Διαφθοράς (Corruption Control – CorControl) μειώνεται γραμμικά (και αντιστρόφως).

Σύμφωνα με το μοντέλο όσες μεταβλητές έχουν θετική εκτίμηση, σημαίνει ότι όσο αυτές αυξάνονται, τότε ο Έλεγχος Διαφθοράς (Corruption Control – CorControl) αυξάνεται γραμμικά (και αντιστρόφως).

Η εκτίμη

Έστω $j \in \{1, 2, \ldots, p\}$. Τια $X_j = X_j + 1$, έχουμε ότι $E(Y) = E(Y_0) + \beta_j \Rightarrow \beta_j = E(Y) - E(Y_0)$. Με άλλα λόγια δηλαδή ο συντελεστής $\beta_j$ εκφράζει τη μεταβολή στην αναμενόμενη τιμή της εξαρτημένης μεταβλητής για αύξηση της ανεξάρτη της μεταβλητής $X_j$ κατά μία μονάδα, κρατώντας όλες τις υπόλοιπες

ανεξάρτητες μεταβλητές σταθερές. Η παράμετρος $\beta_j$ μετριέται σε μονάδα της εξαρτημένης μεταβλητής ανά μονάδα της ανεξάρτητης μεταβλητής $X_j$.

Για παράδειγμα, η εκτίμηση $1.354e-02$ (εκτίμηση συντελεστή ανεξάρτητης μεταβλητής Internet), εκφράζει τη μεταβολή (συγκεκριμένα αύξηση καθώς ο η εκτίμηση είναι θετική) στην αναμενόμενη τιμή της εξαρτημένης μεταβλητής `"CorControl"` για αύξηση της ανεξάρτη της μεταβλητής `"Internet"` κατά μία μονάδα, κρατώντας όλες τις υπόλοιπες ανεξάρτητες μεταβλητές σταθερές. Η παράμετρος $\beta_{Internet}$ μετριέται σε μονάδα της εξαρτημένης μεταβλητής `"CorControl"` ανά μονάδα της ανεξάρτητης μεταβλητής `"Internet"`.

## 5.2 Ερμηνεία των αποτελεσμάτων

Ο έλεγχος της διαφθοράς καταγράφει τις αντιλήψεις σχετικά με τον βαθμό στον οποίο η δημόσια εξουσία ασκείται για ιδιωτικό όφελος, συμπεριλαμβανομένων τόσο των μικροδιαφθορών όσο και των μεγάλων μορφών διαφθοράς, καθώς και της "άλωσης" του κράτους από τις ελίτ και τα ιδιωτικά συμφέροντα. Η εκτίμηση δίνει τη βαθμολογία της χώρας στον συνολικό δείκτη.

- `"Intercept"`: Υπάρχει μια πάγια τιμή Ελέγχου διαφθοράς, οπότε σε όλες τις χώρες υπάχρει η αντίληψη ότι η δημόσια εξουσία ασκείται για ιδιωτικό όφελος, πράγμα που αληθεύει.

- `"ElectrAccess"` : Η αύξηση της πρόσβασης στην ηλεκτρική ενέργεια συνδέεται με μείωση της εξαρτημένης μεταβλητής, `"CorControl"`. Αυτό υποδηλώνει ότι όσο περισσότεροι άνθρωποι αποκτούν πρόσβαση στην ηλεκτρική ενέργεια, ο ελέγχος της διαφθοράς μπορεί να επηρεαστεί αρνητικά. Με την ευρύτερη πρόσβαση σε ηλεκτρικό ρεύμα υπάρχουν περισσότερες ευκαιρίες για διαφθορά κατά τη διάρκεια έργων ηλεκτροδότησης.

  Η αρνητική σχέση μεταξύ `"ElectrAccess"` και `"CorControl"` μπορεί είναι απροσδόκητη, καθώς η υψηλότερη πρόσβαση σε υποδομές θεωρείται συνήθως ότι συσχετίζεται με καλύτερη διακυβέρνηση. Αυτό θα μπορούσε να υποδηλώνει υποκείμενες πολυπλοκότητες ή συγκεκριμένα περιφερειακά ζητήματα.

- `"Agriland"` : Η υψηλή εξάρτηση από τη γεωργία μπορεί να συνδέεται με αγροτικές οικονομίες, όπου η εποπτεία και ο θεσμικός έλεγχος μπορεί να είναι ασθενέστεροι, οδηγώντας ενδεχομένως σε υψηλότερα επίπεδα διαφθοράς στην κατανομή των πόρων και στις γεωργικές επιδοτήσεις. Ωστόσο, ο συνετελεστής είναι αρκετά μικρός και από τα $p-values$ φαίνεται ότι είναι στατιστικά ασήμαντος, οπότε μπορούμε να τον αγνοήσουμε.

- `"CO2"` : Οι υψηλότερες εκπομπές CO2 μπορεί να υποδηλώνουν βιομηχανική δραστηριότητα η οποία, σε ορισμένες χώρες, μπορεί να συνδέεται με αδύναμους περιβαλλοντικούς κανονισμούς και διεφθαρμένες πρακτικές σε βιομηχανικούς τομείς ή περιβαλλοντική συμμόρφωση, αντίθετη άποψη σύμφωνα με τα ευρήματά μας. Ωστόσο και αυτός ο συντελεστής από τα $p-values$ φαίνεται ότι είναι στατιστικά ασήμαντος σε ε.σ.σ. $\alpha = 0.05$, οπότε μπορούμε να τον αγνοήσουμε.

- `"Internet"` : Με την ευρύτερη χρήση του ίντερνετ, μεταφέρεται ευκολότερα η πληφορία και γίνονται ευρέως γνωστά φαινόμενα διαφθοράς την δημόσιας εξουσίας, οπότε με την αύξηση της χρήσης του από κατοίκους μιας χώρας λογικό είναι να αυξάνεται και ο έλεγχος της διαφθοράς.

- `"PopGrowth"` : Η μεγάλη αύξηση του πληθυσμού μπορεί να επιβαρύνει τους πόρους και τις υπηρεσίες, αυξάνοντας δυνητικά τις ευκαιρίες για διεφθαρμένες πρακτικές, καθώς η ζήτηση υπερβαίνει την προσφορά στη διακυβέρνηση και την παροχή υπηρεσιών.

- `"AdolFertRate"` : Τα υψηλά ποσοστά γονιμότητας των εφήβων συχνά υποδηλώνουν χαμηλότερη κοινωνικοοικονομική ανάπτυξη και περιορισμένη πρόσβαση στην εκπαίδευση και τις υπηρεσίες υγείας, που μπορεί να συσχετίζονται με ασθενέστερα θεσμικά πλαίσια και υψηλότερη διαφθορά.

- "GDPdollars" : Το υψηλότερο κατά κεφαλήν ΑΕΠ συνήθως συσχετίζεται με καλύτερη οικονομική ανάπτυξη και ισχυρότερους θεσμούς, οι οποίοι μπορούν να συμβάλουν στη μείωση της διαφθοράς.
- "PopPerc14" : Ένας νεότερος πληθυσμός θα μπορούσε να υποδηλώνει μελλοντικές δυνατότητες για εκπαίδευση και μεταρρυθμίσεις, γεγονός που πιθανόν να υποδηλώνει μελλοντικές βελτιώσεις στον έλεγχο της διαφθοράς.
- "WomBusiness" : Η μεγαλύτερη συμμετοχή των γυναικών στις επιχειρήσεις μπορεί να οδηγήσει σε πιο δίκαιες και διαφανείς επιχειρηματικές πρακτικές, συμβάλλοντας στη μείωση της διαφθοράς.
- "FertCat" : Δεν έχει νόημα να αναλύσουμε με αυτόν τον τρόπο την κατηφορική αυτή μεταβλητή. Θα έπρεπε να εισάγουμε ψευδομεταβλητές που δίνουν πληροφορίες για κάθε επίπεδο της μεταβλητής

## 5.3 Περιορισμοί των μεθόδων

- Ενδέχεται τα μοντέλα να μην είναι επαρκή αν έχουν παραληφθεί άλλες σημαντικές μεταβλητές που επηρεάζουν τον δείκτη "CorControl" που δεν υπήρχαν στη βάση δεδομένων.
- Μπορεί να υπάρχει πολυσυγγραμμικότητα μεταξύ των ανεξάρτητων μεταβλητών, η οποία μπορεί να επηρεάσει τη σταθερότητα των εκτιμήσεων των συντελεστών. Για παράδειγμα, ένα υψηλότερο ποσοστό γεωργικής γης υποδηλώνει συνήθως μεγαλύτερο αγροτικό πληθυσμό, υψηλότερο ΑΕΠ συνδέεται με ευρύτερη χρήση ίντερνετ και ηλεκτρικής ενέργειας και το ποσοστά γονιμότητας των εφήβων έχουν άμεση συσχέτιση με την κατηγορική μεταβλητή "FertCat".

## 5.4 Αξιολόγηση αποτελεσμάτων

- "GDPdollars": υποστηρίζοντας την άποψη ότι η οικονομική ανάπτυξη ενισχύει τον έλεγχο της διαφθοράς.
- "Internet": Σημαντικό και στα δύο μοντέλα, ευθυγραμμίζεται με τις θεωρίες για τη διαφάνεια και τη διάδοση πληροφοριών που μειώνουν τη διαφθορά.
- "AdolFertRate": Σημαντικό και στα δύο μοντέλα, αντανακλώντας ευρύτερες κοινωνικοοικονομικές επιπτώσεις.
- "ElectrAccess": (Σημαντικό αλλά λιγότερο αναμενόμενο) Η αρνητική επίπτωση είναι σημαντική, αλλά απαιτεί περαιτέρω διερεύνηση για την κατανόηση των υποκείμενων λόγων.
- "WomBusiness": Οριακά σημαντική στο μοντέλο καλύτερης προσαρμογής, γεγονός που υποδηλώνει ότι θα μπορούσε να εξαρτάται από το πλαίσιο ή να επηρεάζεται από άλλους μη μετρήσιμους παράγοντες.
- "PopGrowth" : Σημαντικό μόνο στο μοντέλο καλύτερης προσαρμογής, γεγονός που υποδηλώνει κάποια ανθεκτικότητα, αλλά ενδεχομένως επηρεάζεται από το συγκεκριμένο υποσύνολο δεδομένων που χρησιμοποιήθηκε.

## 5.5 Σύγκριση με την πραγματικότητα

Αντιστοίχιση με την πραγματικότητα: Η σημασία των δολαρίων ΑΕΠ και της πρόσβασης στο Διαδίκτυο που συνεπάγονται και αυξημένη χρήση ηλεκτρικής ενέργειας, αυξημένες εκπομπές CO2 και του πληθυσμού και ευθυγραμμίζεται με τις καθιερωμένες θεωρίες για την οικονομική ανάπτυξη και τη διαφάνεια που βελτιώνει τον έλεγχο της διαφθοράς.

Χρειάζεται περαιτέρω διερεύνηση: Ο αρνητικός αντίκτυπος του ElectrAccess και ο ειδικός ρόλος των ποσοστών γονιμότητας των εφήβων ενδέχεται να απαιτούν βαθύτερη ανάλυση του πλαισίου για την κατανόηση των περιφερειακών παραλλαγών και των υποκείμενων αιτιωδών μηχανισμών.

## Αναφορές


[1] D.C. Montgomery, E.A. Peck, and G.G. Vining (2001) *Introduction to Linear Regression Analysis*, John Wiley and Sons, Hardcover.

[2] G.A.F. Seber and A.J. Lee (2003) *Linear Regression Analysis*, Wiley Series in Probability and Statistics, John Wiley and Sons, Hardcover.

[3] D.C. Montgomery (2017) *Design and Analysis of Experiments*, John Wiley and Sons.

[4] T.P. Ryan (2009) *Modern Regression Methods*, Wiley Series in Probability and Statistics, John Wiley and Sons.

[5] C.R. Rao and H. Toutenburg (2002) *Linear Models: Least Squares and Alternatives*, Springer.