ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ

ΕΑΡΙΝΟ ΕΞΑΜΗΝΟ

ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ

- Δεδομένα εκλογών στην Αμερική 2000 -

(Εργασία 1)

Διδάσκοντες: Ι. Ντζούφρας – Κ. Πατέρας

Φοιτητής: Νικόλαος Παπαγεωργίου 3200131

Έτος σπουδών: 4ο

8 Ιουνίου, 2024

Περιεχόμενα:

1.Εισαγωγή και περιγραφή της μελέτης του προβλήματος

Αντικείμενο της μελέτης , αποτελούν τα δεδομένα των εκλογών στις Ηνωμένες Πολιτείες, σε επίπεδο πολιτείας, στις εκλογές του 2000. Τα αποτελέσματα αυτών των εκλογών, παρουσιάζουν σημαντικές πληροφορίες σχετικά με τα δημογραφικά στοιχεία και τις συμπεριφορές των πολιτών στις διάφορες πολιτείες της Αμερικής. Η ανάλυση μας, θα βοηθήσει να καταλάβουμε τους παράγοντες που επηρεάζουν την ψήφο των πολιτών στις προεδρικές εκλογές και να αναδείξουμε πιθανές σχέσεις των μεταβλητών.

Κατά την διάρκεια της μελέτης:

- Θα παρουσιάσουμε τα κατάλληλα περιγραφικά μέτρα για τις μεταβλητές.
- Θα εξετάσουμε τις σχέσεις μεταξύ των μεταβλητών (ανά δύο).
- Θα εξετάσουμε πιθανές εξαρτήσεις μεταξύ μεταβλητών.
- Θα επαναλάβουμε τα παραπάνω βήματα χωρίζοντας τις πολιτείες σε αυτές που έχουν μεγάλο αριθμό πληθυσμού και αυτές που έχουν μικρότερο.
- Θα κάνουμε εκτίμηση γραμμικού υποδείγματος.

Πίνακας 1:Πίνακας Δεδομένων

Αριθμός μεταβλητής	Όνομα	Τύπος μεταβλητής	Σημασία
0	Id	Categorical	Id κάθε πολιτείας
1	State	Character	Πολιτείες Αμερικής σε μορφή παράγοντα
2	Bush	Numeric	Ποσοστό που ψήφισε George Bush ανά πολιτεία
3	Male	Numeric	Ποσοστό ανδρών ανά πολιτεία
4	Population	Numeric	Πληθυσμός σε κάθε πολιτεία
5	Rural	Numeric	Ποσοστό πληθυσμού σε μη αστικές περιοχές
6	Below_Poverty ή bpovl	Numeric	Ποσοστό πληθυσμού/πολιτεία με εισόδημα κάτω από τα επίπεδα της φτώχειας
7	CLFU	Numeric	Ποσοστό ανεργίας ανά πολιτεία
8	Management_18 ή mgt18	Numeric	Ποσοστό ανδρών μεγαλύτερων από 18 ετών
9	Percentage_65 ή pgt65	Numeric	Ποσοστό πληθυσμού μεγαλύτεροι από 65 ετών
10	Num_75 ή numgt75	Numeric	Ποσοστό πληθυσμού με εισόδημα μεγαλύτερο από 75K

2. Περιγραφική ανάλυση

Μέρος 2.1 : Αρχικό σετ δεδομένων

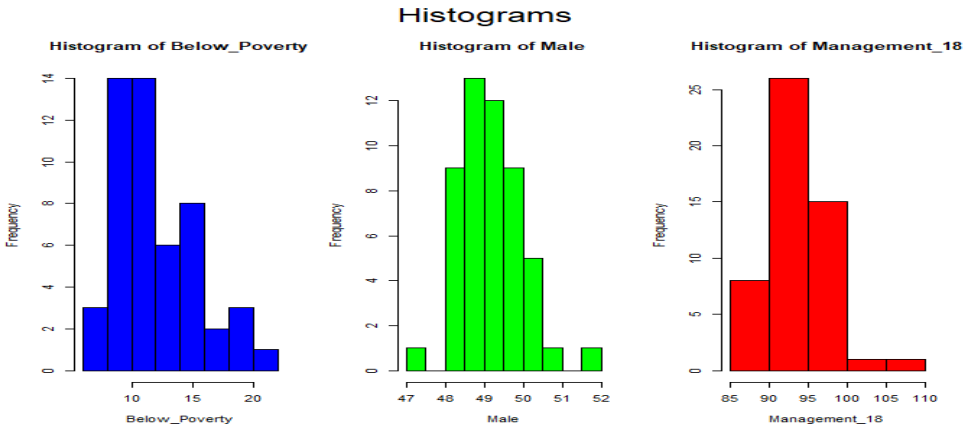
Για την ανάλυσή μας θα χρησιμοποιήσουμε το στατιστικό πακέτο R, που θα μας βοηθήσει στον υπολογισμό των ελέγχων αλλά και στη διαγραμματική απεικόνιση των αποτελεσμάτων μας. Θα εισάγουμε τα δεδομένα μας και τα πακέτα/βιβλιοθήκες που θα χρειαστούμε για την ανάλυση μας. Αρχικά οφείλουμε να εξετάσουμε κάθε μεταβλητή που έχουμε και τις τιμές της. Την μεταβλητή Num_75(Πληθυσμός με εισόδημα μεγαλύτερο από 75K), θα την μετασχηματίσουμε σε ποσοστό, ώστε να είναι ευκολότερα συγκρίσιμη με τις υπόλοιπες μεταβλητές που είναι ήδη εκφρασμένες σε ποσοστά και την Bush αρχικά θα την στρογγυλοποιήσουμε σε δύο δεκαδικά ψηφία και μετά θα την πολλαπλασιάσουμε με το 100 ώστε να είναι σε κοινή κλίμακα με τις υπόλοιπες και να εκφράζει «καθαρά» ποσοστά. Όπως μπορούμε να δούμε και από τον παραπάνω πίνακα έχουμε Numeric, δηλαδή ποσοτικές μεταβλητές και θα μελετήσουμε την μέση τιμή την τυπική απόκλιση ,τη διάμεσο την ασυμμετρία αλλά και την κύρτωση που παρουσιάζουν (Πίνακας 2) . Τιμές κοντά στο μηδέν, τόσο για την ασυμμετρία όσο και για την κύρτωση φανερώνουν ότι μια μεταβλητή προσεγγίζει (ακολουθεί) τη κανονική κατανομή.

Πίνακας 2: Πίνακας περιγραφικών μέτρων

Μεταβλητή	Μέσο	Τυπική Απόκλιση	Διάμεσος	Μικρότερη Τιμή	Μεγαλύτερη τιμή	Ασυμμετρία	Κύρτωση
Bush	49.63	10.4	50	9.0	68	-1.06	2.73
Male	49.14	0.81	49.06	47.09	51.69	0.55	0.84
Population	5518076.6	6164025.78	4012012.00	493782.00	33871648.00	2.45	7.32
Rural	32	21.16	29.60	0.00	76.76	0.39	2.96
Below_Poverty	12.10	3.30	11.40	6.50	20.20	0.79	-0.12
CLFU	5.66	1.31	5.60	3.50	10.80	1.37	3.29
Management_18	93.96	3.95	93.30	86.10	107.60	0.93	1.35
Percentage_65	12.53	1.89	12.70	5.70	17.60	-0.69	2.53
Num_75	2.21	1.19	1.96	0.92	5.99	1.25	0.94

Μπορούμε να παρατηρήσουμε ότι για την μεταβλητή bpovl(Below_Poverty) έχουμε χαμηλές τιμές Ασυμμετρίας και Κύρτωσης, οπότε πιθανόν να προσεγγίζει την κανονική κατανομή, έχοντας όμως μια δεξιά ασυμμετρία και μια διαφορά ανάμεσα στην Μέση τιμή και την Διάμεσο της μεταβλητής. (Βλέπε Πίνακα 2). Στις υπόλοιπες μεταβλητές έχουμε μεγαλύτερες τιμές στα χαρακτηριστικά αυτά και στις περισσότερες (Population,Rural,CLFU,Management_18,Num_75) διακρίνουμε δεξιά ασυμμετρία, ενώ στις Bush και Percentage_65 αριστερή. Σε χαμηλές τιμές στις τιμές της Ασυμμετρίας και Κύρτωσης ακολουθούν την bpovl, οι μεταβλητές Male και Management_18. Παράλληλα, όλες οι μεταβλητές εκτός της Below_Poverty είναι λεπτόκυρτες (Πίνακας 2). Τις παραπάνω υποθέσεις που κάναμε μπορούμε να τις επιβεβαιώσουμε ή να τις απορρίψουμε μέσω διαγραμμάτων και ελέγχων. Στα διαγράμματα θα περιλάβουμε μεταβλητές που είναι πιθανόν να προσεγγίζουν την κανονική κατανομή και παρουσιάζουν όσο το δυνατόν χαμηλότερες τιμές στην Ασυμμετρία και τη Κύρτωση και μικρή

διαφορά ανάμεσα στη Μέση τιμή τους και την Διάμεσο τους. Άρα θα ασχοληθούμε με την Male, bpovl(Below_Poverty), Management_18(Mgt_18) και για αρχή θα παρουσιάσουμε τα ιστογράμματα τους.



Σχήμα 1: Ιστογράμματα μεταβλητών Male,Below_Poverty,Management18

Με βάση τα ιστογράμματα (Σχήμα 1) που βλέπουμε η μεταβλητή Male φαίνεται να προσεγγίζει περισσότερο την κανονική κατανομή έχοντας αρκετές τιμές κοντά στο κέντρο και λιγότερες στα άκρα , παρουσιάζοντας όμως μια ελαφριά δεξιά ασυμμετρία, ενώ οι άλλες δύο μεταβλητές, δηλαδή η Below_Poverty και η Management_18 φαίνεται να έχουν περισσότερες μικρές τιμές (λεπτόκυρτη) και λιγότερες μεγάλες. Για να είμαστε πιο βέβαιοι για τα συμπεράσματα μας θα πραγματοποιήσουμε και ελέγχους (Shapiro–Wilk και Kolmogorov – Smirnov) καθώς και τα αντίστοιχα QQplots. Στους ελέγχους θα χρησιμοποιήσουμε/υποθέσουμε επίπεδο εμπιστοσύνης 95%. Από τους ελέγχους παρατηρούμε ότι μόνο η μεταβλητή Male ακολουθεί την κανονική κατανομή (S−W pvalues = 0.1528>0.05 και K – S(Lillie)pvalue = 0.1952>0.05), ενώ οι μεταβλητές Below_Poverty(S−W pvalues <0.01 και K–S(Lillie)pvalue=0.026<0.05)και Management_18(S−W pvalues = 0.022<0.05) αποκλίνουν. Μόνο στην Male δεν απορρίπτουμε την κανονικότητα.

Στη συνέχεια θα δούμε τα διαγράμματα πλαισίου για κάθε ποσοτική μεταβλητή προκειμένου να εντοπίσουμε τυχόν ακραίες τιμές ,τιμές δηλαδή που αποκλίνουν από τις υπόλοιπες (Βλέπε Σχήμα 4,5,6). Η μεμονωμένη ανάλυση μεταβλητών είναι σίγουρα χρήσιμη και βοηθάει στο να κατανοήσουμε τη συμπεριφορά τους, ακολούθως όμως θα ασχοληθούμε και με την μεταξύ τους σχέση που θα μας βοηθήσει στην αποκόμιση ακόμα περισσότερων συμπερασμάτων.

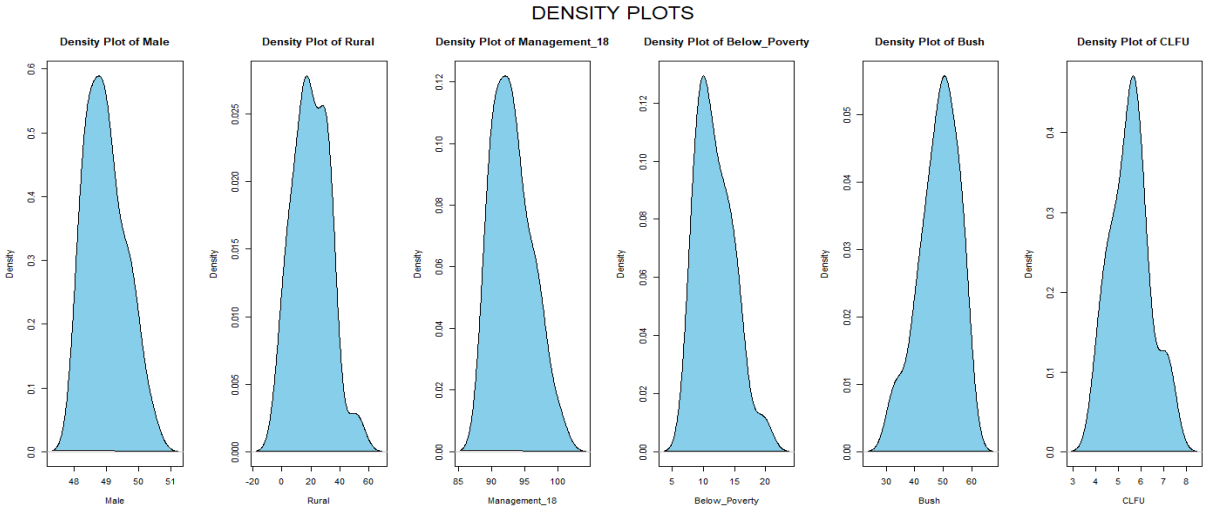
Μέρος 2.2 :Subsets (Πολιτείες με μεγάλο ή μικρό αριθμό πληθυσμού)

Θα χωρίσουμε τις πολιτείες σε δύο υποομάδες αυτές με μεγάλο πληθυσμό και αυτές με μικρότερο πληθυσμό χρησιμοποιώντας την διάμεσο της Population. Θα συγκρίνουμε τις δύο υποομάδες στην περιγραφική ανάλυση που θα κάνουμε για την κάθε μια και στις σχέσεις των μεταβλητών σε αυτές.

Από τους πίνακες περιγραφικής ανάλυσης που πραγματοποιήσαμε για κάθε υποομάδα (Βλέπε Πίνακα 10-11),παρατηρούμε ξανά τις τιμές της Ασυμμετρίας και της Κύρτωσης για κάθε μεταβλητή ξεχωριστά (Υπενθύμιση: τιμές κοντά στο 0 και στις δύο αυτές μετρικές δείχνουν ότι πιθανόν η

εκάστωτε μεταβλητή προσεγγίζει την κανονική κατανομή). Παράλληλα,παρατηρώντας τα διαγράμματα πυκνότητας πιθανότητας (Σχήμα 2) που δείχνουν για τις μεταβλητές που απεικονίζονται να μην έχουν ουρά στα άκρα και να έχουν συμμετρικές κατανομές και κάνοντας τους κατάλληλους ελέγχους κανονικότητας για κάθε μεταβλητή (Βλέπε Πίνακα 10-11) βλέπουμε (κάνουμε Shapiro.test εξαιτίας μεγέθους δείγματος <50):

- <u>Για πολιτείες με υψηλό πληθυσμό φαίνεται να προσεγγίζουν την κανονική κατανομή</u> :
    - Η μεταβλητή Male(Shapiro.test: P-value: 0.18>0.05)
    - Η μεταβλητή Rural (Shapiro.test: P-value: 0.29>0.05)
    - Η μεταβλητή Management_18 (Shapiro.test: P-value: 0.14>0.05,)
    - Η μεταβλητή Below_Poverty (Shapiro.test: P-value: 0.14>0.05,)
    - Η μεταβλητή Bush (Shapiro.test: P-value: 0.36>0.05)
    - Η μεταβλητή CLFU (Shapiro.test: P-value: 0.48>0.05)



Σχήμα 2 : Διαγράμματα πυκνότητας πιθανότητας μεταβλητών σε υψηλό αριθμό πληθυσμού

- <u>Για πολιτείες με μικρό πληθυσμό φαίνεται να προσεγγίζουν την κανονική κατανομή:</u>
    - Η μεταβλητή Rural (Shapiro.test: P-value: 0.21>0.05)
    - Η μεταβλητή Management_18(Shapiro.test: P-value: 0.49>0.05)
    - Η μεταβλητή Male(Shapiro.test: P-value: 0.9>0.05)

Σε αυτά τα συμπεράσματα μας οδήγησαν τα διαγράμματα πυκνότητας πιθανότητας (Σχήμα 2, Σχήμα 19) και τα QQplots που δημιουργήσαμε (Βλέπε Σχήμα 16, Σχήμα 20). Παρατηρούμε λοιπόν, διαφορές ανάμεσα στις δύο υποομάδες και οι μόνες μεταβλητές που ακολουθούν κανονική κατανομή και στις δύο, είναι η Rural (Ποσοστό πληθυσμού σε μη αστικές περιοχές) και η Management_18(Ποσοστό ανδρών μεγαλύτερων από 18 ετών) και η Male. Είναι σημαντικό να αναφέρουμε ότι ανάμεσα στις υποομάδες σημαντική διαφορά στο μέσο και την διάμεσο παρουσιάζει το ποσοστό πληθυσμού σε μη αστικές περιοχές, κάτι που περιμέναμε εξαιτίας του διαχωρισμού που κάναμε σε πολιτείες με μεγάλο αριθμό πληθυσμού και σε αυτές με μικρότερο. Στις υπόλοιπες μεταβλητές δεν παρατηρούμε μεγάλες αποκλίσεις ανάμεσα στα περιγραφικά μέτρα τους.

3. Σχέσεις μεταβλητών ανά δύο

Μέρος 3.1 Αρχικό σετ δεδομένων

Αρχικά υπολογίζουμε τον δείκτη συσχέτισης Pearson (Βλέπε Σχήμα 7) μαθαίνοντας ποίες μεταβλητές παρουσιάζουν μεγάλη συσχέτιση μεταξύ τους. Έπειτα υλοποιούμε Scatterplots ανάμεσα στις ποσοτικές μεταβλητές που παρουσιάζουν σημαντική συσχέτιση και στις μεταβλητές Bush-Rural (Ποσοστό/πολιτεία που ψήφισε τον G.Bush-Ποσοστό πληθυσμού που ζει σε μη αστικές περιοχές) ,Bush-CLFU(Ποσοστό/ πολιτεία που ψήφισε τον G.Bush-Ποσοστό ανεργίας ανά πολιτεία), που θα μας απασχολήσουν.
Από τα αποτελέσματα που λάβαμε και απ'τά διαγράμματα (Βλέπε Σχήμα 7,8,9) παρατηρούμε υψηλή συσχέτιση (θα θεωρήσουμε υψηλή συσχέτιση : >0.6 ή <-0.6) ανάμεσα στις:

- Bush-Numgt75 αρνητική συσχέτιση -0.75 και Bush-Rural θετική συσχέτιση 0.6
- Male-Management_18 θετική συσχέτιση 0.99
- Rural-Numgt75 αρνητική συσχέτιση -0.61
- Below_Poverty-CLFU θετική συσχέτιση 0.67

Είναι ενδιαφέρον να εξετάσουμε περαιτέρω αν το ποσοστό που έλαβε ο George Bush εξαρτάται από το ποσοστό πληθυσμού που ζει σε μη αστικές περιοχές ανά πολιτεία ή αν εξαρτάται από το ποσοστό ανεργίας της κάθε πολιτείας. Σύμφωνα με τους δείκτες Pearson λαμβάνουμε: Bush-Rural:0.6 και Bush-CLFU:-0.33. Βλέπουμε σχετικά δυνατή «σχέση» ανάμεσα στα ποσοστά του G.Bush και του ποσοστού του πληθυσμού που ζει σε μη αστικές περιοχές, ενώ μια μικρότερη, αρνητική σχέση ανάμεσα στα ποσοστά του G.Bush και στα ποσοστά ανεργίας. Μπορούμε επομένως να καταλάβουμε ότι η εξάρτηση της μεταβλητής Bush από την μεταβλητή Rural είναι μεγαλύτερη σε σχέση με αυτήν από τη μεταβλητή CLFU. Επεκτείνοντας την ανάλυση μας δημιουργούμε μια καινούργια κατηγορική μεταβλητή Bush_Over που θα δέχεται τιμές 0 και 1 ανάλογα με τα ποσοστά που έχει σε κάθε πολιτεία η μεταβλητή Bush. Παρατηρήσεις με μεγάλα ποσοστά του George Bush θα έχουν τιμή 1 και αυτές με μικρότερα, τιμή 0. Η τιμή με την οποία θα γίνει διαχωρισμός των μεταβλητών είναι το 50(50%), που εκφράζει τη Διάμεσο και το Μέσο της μεταβλητής Bush (Βλέπε Πίνακα1).
Οπότε θα πραγματοποιήσουμε τους απαραίτητους ελέγχους και στις δύο περιπτώσεις που μας ζητούνται (Βλέπε Πίνακα 6,7,8) , μέσω της Bush_over. Από τα αποτελέσματα, φαίνεται ότι Bush_over δεν εξαρτάται σημαντικά από τη μεταβλητή CLFU, δηλαδή το ποσοστό ανεργίας ανά πολιτεία, (t.test p-value=0.88>0.05) (Βλέπε Πίνακα 6,7). Από την άλλη η Rural ,το ποσοστό πληθυσμού που ζει σε μη αστικές περιοχές, φαίνεται να επηρεάζει τη μεταβλητή Bush_over(Wilcoxn p-value = 2.2e-16<0.05) (Βλέπε Πίνακα 8,Σχήμα 13).Με τους παραπάνω ελέγχους και τους δείκτες συσχέτισης που προαναφέραμε καταλήγουμε ότι η μεταβλητή Bush και κατεπέκταση η Bush_over εξαρτάται (η αλλιώς επηρεάζεται) από την μεταβλητή Rural. Ισχυρές εξαρτήσεις μεταξύ των Bush-Rural, Bush-Num_75 και Rural-Num_75 επιβεβαιώνονται από Πίνακα 9 παραρτήματος όπου μετατρέψαμε σε

δίτιμες κατηγορικές με βάση την διάμεσό τους τις μεταβλητές (μικρά-μεγάλα ποσοστά) ελέγχοντας τις εξαρτήσεις.

Μέρος 3.2 Subsets (Πολιτείες με μεγάλο ή μικρό αριθμό πληθυσμού)

Εκτός των ατομικών αναλύσεων που κάναμε στα δύο subsets, θα ασχοληθούμε και με τις σχέσεις των μεταβλητών πραγματοποιώντας Scatterplots και βρίσκοντας για κάθε ζεύγος μεταβλητών το δείκτη συσχέτισης Pearson(Βλέπε Σχήμα 22,23 24-27).Δείκτες υψηλής συσχέτισης Pearson (θεωρούμε υψηλή συσχέτιση για :>0.6 ή <-0.6) στις πολιτείες με μικρό αριθμό πληθυσμού (Βλέπε Σχήμα 22):

- Bush-Male θετική συσχέτιση 0.61
- Bush-Rural θετική συσχέτιση 0.62
- Bush-Num_75 αρνητική συσχέτιση -0.73
- Male-Management_18 θετική συσχέτιση 0.99
- Below_Poverty-CLFU θετική συσχέτιση 0.65

Δείκτες υψηλής συσχέτισης Pearson (υψηλή συσχέτιση: για >0.6 ή για <-0.6) στις πολιτείες με μεγάλο αριθμό πληθυσμού (Βλέπε Σχήμα 23):

- Bush-Rural θετική συσχέτιση 0.7
- Bush-Num_75 αρνητική συσχέτιση -0.85
- Rural-Num_75 αρνητική συσχέτιση -0.69
- Male-Management_18 θετική συσχέτιση 1
- Below_Poverty-CLFU θετική συσχέτιση 0.76

Παρατηρούμε ότι και στις δύο ομάδες τα ζευγάρια Below_Poverty-CLFU,Bush-Rural,Male-Management_18 και Bush-Num_75 εμφανίζουν υψηλές συσχετίσεις. Αντίθετα, η Bush φαίνεται να έχει υψηλή συσχέτιση με την Male μόνο σε πολιτείες με μικρό αριθμό πληθυσμού. Όσον αφορά τις Rural-Num_75 επίσης φαίνεται στις πολιτείες με υψηλό πληθυσμό μεγαλύτερη συσχέτιση (αρνητική).

Αφού παρατηρήσαμε ποιες μεταβλητές επηρεάζουν σημαντικά την μεταβλητή Bush στις υποομάδες ,θα ελέγξουμε το κατά πόσο η μεταβλητή Bush εξαρτάται από το ποσοστό πληθυσμού που ζει σε μη αστικές περιοχές(Rural) και το ποσοστό ανεργίας ανά πολιτεία(CLFU), εκμεταλλευόμενοι την Bush_over που δημιουργήσαμε. Παρατηρώντας τα Scatterplots για τα δύο ζευγάρια μεταβλητών(Bush-Rural,Bush-CLFU) για κάθε υποομάδα(Βλέπε Σχήμα 26, Σχήμα 27), βλέπουμε μεγαλύτερη επίδραση και βαθμό συσχέτισης ανάμεσα στην Bush και την Rural και στις δύο υποομάδες με εντονότερη σε αυτήν με μεγάλο αριθμό πληθυσμού (Δείκτης συσχέτισης Pearson:0,7 έναντι 0.62). Αντίθετα, στις δύο ομάδες πληθυσμού οι μεταβλητές Bush-CLFU δεν δείχνουν πολύ στενή σχέση, όμως στις πολιτείες με μικρό αριθμό πληθυσμού δεν είναι αμελητέα(δείκτη συσχέτισης Pearson:-0.42) ,όπως σε αυτές με μεγάλο αριθμό πληθυσμού που είναι σχεδόν μηδενική (δείκτη συσχέτισης Pearson:-0.06).

Με τους ελέγχους που πραγματοποιήσαμε ανάμεσα στην Bush_over και τις μεταβλητές Rural και

CLFU στις δύο υποομάδες πληθυσμού συμπεραίνουμε: Η επίδραση της CLFU,του ποσοστού ανεργίας ανά πολιτεία, είναι διαφορετική στις υποομάδες. Ενώ φαίνεται ότι η Bush_over εξαρτάται από τη μεταβλητή CLFU στο subset με χαμηλό αριθμό πληθυσμού (Wilcoxon:p-value=3.427e-10<0.05),(Βλέπε Πίνακα 13) δεν ισχύει το ίδιο για το subset με μεγάλο αριθμό πληθυσμού.(t.test:p-value=0.92>0.05) (Βλέπε Πίνακας 12).Από την άλλη, για την μεταβλητή Rural,του ποσοστού του πληθυσμού σε μη αστικές περιοχές ανά πολιτεία, φαίνεται και στις δύο περιπτώσεις των δύο υποομάδων η μεταβλητή Bush_over να εξαρτάται από αυτήν, επομένως δεν έχουμε σημαντική διαφορά στην εξάρτηση των Bush-Rural στα δύο subsets.(Βλέπε Πίνακα 14, Πίνακα 15).

4. Προβλεπτικά ή ερμηνευτικά μοντέλα

Πραγματοποιώντας περιγραφική ανάλυση του αρχικού dataset και των subsets που δημιουργήσαμε, αναλύοντας παράλληλα τις πιθανές εξαρτήσεις και συσχετίσεις των μεταβλητών μεταξύ τους σε αυτά, καλούμαστε να εκτιμήσουμε ένα γραμμικό υπόδειγμα/μοντέλο. Θα εξετάσουμε τη σχέση της μεταβλητής Bush(Ποσοστό που έλαβε ο G.Bush ανά πολιτεία) με τις υπόλοιπες μεταβλητές.

Παρατηρώντας ότι έχουμε μόνο ποσοτικές μεταβλητές και «κατά πλειοψηφία» εκφράζουν ποσοστά, θα δημιουργήσουμε μια κατηγορική μεταβλητή, με σκοπό να εμπλουτίσουμε της ανάλυση μας. Επιλέγουμε να μετατρέψουμε την Population(pop: Ποσοστό πληθυσμού ανά πολιτεία), καθώς αποτελεί την μοναδική μεταβλητή που δεν εκφράζει ποσοστά και η μετατροπή της θα είναι βοηθητική. Ταυτόχρονα, οι τιμές της είναι αρκετά διαφορετικές ανά πολιτεία και θα μπορέσουμε να δούμε πιθανόν διαφοροποιήσεις στην μεταβλητή Bush ανάλογα με τις κατηγορίες της Population (low,medium,high). Οι κατηγορίες θα προκύψουν με βάση τα τεταρτημόρια του boxplot της Population (Βλέπε Σχήμα 38),έτσι θα είναι στατιστικά αντιπροσωπευτικές. Μέσω ελέγχων Anova ανάμεσα στην μεταβλητή Bush και Population (Βλέπε Πίνακα 16) και διαγραμμάτων παρατηρούμε ότι δεν υπάρχουν σημαντικές διαφορές στις διαμέσους και τους μέσους της μεταβλητής Bush μεταξύ των κατηγοριών της Population (Kruskal p-value=0.21>0.05&Oneway.testp-value=0.13>0.05).

Πλέον έχοντας ελέγξει εκτενώς τα δεδομένα και τις σχέσεις ανάμεσα στις μεταβλητές ,εστιάζοντας στην σχέση της Bush με τις υπόλοιπες, προχωράμε στην προσαρμογή ενός μοντέλου για την επιρροή των μεταβλητών στο ποσοστό που έλαβε ο G.Bush στις εκλογές ανά πολιτεία. Στην ανάλυση δε θα λάβουμε υπόψη τις : ID(Αριθμός πολιτείας),State(Όνομα πολιτειας) καθώς δεν μας προσδίδουν περαιτέρω πληροφορία. Το αρχικό πλήρες μοντέλο για το ποσοστό του G.Bush θα είναι της μορφής :

Bush(Ποσοστό G.Bush) = βο+β1*male(Ποσοστό ανδρών)+β2*Population(Κατηγορία μεγέθους πληθυσμού)+β3*Rural(Ποσοστό πληθυσμού σε μη αστικές περιοχές)+β4*Below_poverty(Ποσοστό με εισόδημα κάτω από όρια της φτώχειας)+β5*Clfu(Ποσοστό ανεργίας)+β6*Management_18(Ποσοστό ανδρών άνω των 18 ετών)+β7*Percentage_65(Ποσοστό πληθυσμού άνω των 65 ετών)+β8* Num_75(Ποσοστό πληθυσμού με εισόδημα μεγαλύτερο από 75K) + ε , ε~N(0,σ^2)

Μέσω των ελέγχων κανονικότητας, ομοσκεδαστικότητας, ανεξαρτησίας και πολυσυγγραμμικότητας αντλούμε πληροφορίες για το πόσο καλό είναι το αρχικό μοντέλο με όλες τις μεταβλητές. Σε συνδυασμό με τα διαγράμματα (Βλέπε Σχήμα 43,44) και τους ελέγχους, φαίνεται να ικανοποιεί τις υποθέσεις της κανονικότητας (Shapiro:p-value=0,67&Lillie:p-value=0.84>0.05), ομοσκεδαστικότητας (Levens :p-value0.35>0.05) και ανεξαρτησίας καταλοίπων (D-W:p-value=0.28>0.05), εμφανίζοντας προβλήματα πολυσυγγραμικότητας και πιθανόν γραμμικότητας (Βλέπε Πίνακα 17). Αυτό, πιθανόν να οφείλεται στο ότι περιλαμβάνονται μεταβλητές που ήδη γνωρίζαμε ότι εμφανίζουν πολύ υψηλά ποσοστά συσχέτισης (Male- Management_18). Για την ερμηνεία του μοντέλου θα κεντροποιήσουμε τις ποσοτικές μεταβλητές, αφαιρώντας από αυτές τους μέσους τους, αλλάζοντας μόνο τους συντελεστές στο μοντέλο και την ερμηνεία των μεταβλητών .

		Bush	
Predictors	*Estimates*	*std. Error*	*p*
(Intercept)	46.96	1.98	**<0.001**
Male	20.16	14.54	0.173
Population [medium]	2.55	2.49	0.312
Population [high]	5.54	2.91	0.065
Rural	0.17	0.06	**0.007**
Below Poverty	0.50	0.48	0.303
CLFU	-2.49	1.25	0.054
Management 18	-3.60	2.94	0.229
Percentage 65	-1.18	0.54	**0.035**
Num 75	-3.10	1.30	**0.022**
Observations	51		
R^2 / R^2 adjusted	0.788 / 0.741		
AIC	325.687		

Πίνακας 3: Μοντέλο πολλαπλής παλινδρόμησης με όλες τις μεταβλητές

Το β0 λαμβάνει τιμή 47. Αυτό υποδηλώνει ότι το ποσοστό του George Bush σε πολιτείες με χαμηλό αριθμό πληθυσμού που έχουν μηδενικό ποσοστό ανδρών, μηδενικό ποσοστό πληθυσμού που ζει σε μη αστικές περιοχές, μηδενικό ποσοστό πληθυσμού/πολιτεία με εισόδημα κάτω από το επίπεδο της φτώχιας, μηδενικό ποσοστό ανεργίας ανά πολιτεία, μηδενικό ποσοστό ανδρών μεγαλύτερων από 18 ετών, μηδενικό ποσοστό του πληθυσμού που είναι μεγαλύτεροι από 65 ετών, μηδενικό ποσοστό πληθυσμού με εισόδημα μεγαλύτερο από 75K αναμένεται να είναι 47%.

Για την παράμετρο β1 του μοντέλου μας ,του ποσοστού ανδρών ανά πολιτεία, συμπεραίνουμε ότι αν αυξηθεί κατά 1% το ποσοστό ανδρών σε μια πολιτεία με χαμηλό αριθμό πληθυσμού κρατώντας σταθερές όλες τις υπόλοιπες μεταβλητές, το ποσοστό πληθυσμού που ζει σε μη αστικές περιοχές, το ποσοστό πληθυσμού/πολιτεία με εισόδημα κάτω από το επίπεδο της φτώχιας, το ποσοστό ανεργίας ανά πολιτεία, το ποσοστό ανδρών μεγαλύτερων από 18 ετών, το ποσοστό του πληθυσμού που είναι μεγαλύτεροι από 65 ετών, το ποσοστό πληθυσμού με εισόδημα μεγαλύτερο από 75K, τότε το αναμενόμενο ποσοστό του G.Bush αναμένεται να αυξηθεί κατά 20%.

Για την β2, δηλαδή τις πολιτείες με μεσαίο μέγεθος πληθυσμού, συμπεραίνουμε ότι μια πολιτεία με μεσαίο αριθμό πληθυσμού, αναμένεται να έχει μια αύξηση 2.55% στο ποσοστό του G.Bush σε σχέση με μια πολιτεία με μικρό αριθμό πληθυσμού άμα βρίσκονται κάτω από τις ίδιες συνθήκες.

Για την β3, δηλαδή τις πολιτείες με μεγάλο αριθμό πληθυσμού, συμπεραίνουμε ότι μια πολιτεία με μεγάλο αριθμό πληθυσμού, αναμένεται να έχει μια αύξηση 5.54% στο ποσοστό του G.Bush σε σχέση με μια πολιτεία με μικρό αριθμό πληθυσμού άμα βρίσκονται κάτω από τις ίδιες συνθήκες.

Για την β4, δηλαδή το ποσοστό πληθυσμού σε μη αστικές περιοχές ανά πολιτεία, συμπεραίνουμε ότι αύξηση 1% στο ποσοστό πληθυσμού σε μη αστικές περιοχές, αναμένεται να έχει μια πολύ μικρή αύξηση 0.17% στο ποσοστό του G.Bush σε σχέση με πριν άμα βρίσκονται υπό ίδιες συνθήκες.

Για την β5, δηλαδή το ποσοστό πληθυσμού που ζει κάτω από τα όρια της φτώχειας ανά πολιτεία, συμπεραίνουμε ότι αν αυξηθεί το ποσοστό πληθυσμού που ζει κάτω από τα όρια της φτώχειας κατά 1%, αναμένεται να έχει μια μικρή αύξηση 0.5% στο ποσοστό του G.Bush σε σχέση με πριν άμα βρίσκονται κάτω από τις ίδιες συνθήκες. (παραμείνουν όλες οι υπόλοιπες μεταβλητές σταθερές).

Για την β6, δηλαδή το ποσοστό ανεργίας ανά πολιτεία, συμπεραίνουμε ότι αν αυξηθεί το ποσοστό αυτό της ανεργίας κατά 1% , αναμένεται να έχει μια μείωση 2.50% στο ποσοστό του G.Bush σε σχέση με πριν άμα βρίσκονται κάτω από τις ίδιες συνθήκες (σταθερές υπόλοιπες μεταβλητές).

Για την β7, δηλαδή το ποσοστό ανδρών μεγαλύτερων από 18 ετών ανά πολιτεία, συμπεραίνουμε ότι αν αυξηθεί το ποσοστό αυτό των ανδρών μεγαλύτερων από 18 ετών κατά 1%, αναμένεται να έχει μια μείωση 3.6% στο ποσοστό του G.Bush σε σχέση με πριν άμα βρίσκονται κάτω από τις ίδιες συνθήκες (δηλαδή παραμείνουν σταθερές όλες οι υπόλοιπες τιμές των μεταβλητών του μοντέλου).

Για την β8, δηλαδή το ποσοστό πληθυσμού μεγαλύτεροι από 65 ετών ανά πολιτεία, συμπεραίνουμε ότι αν αυξηθεί το ποσοστό αυτό κατά 1%, αναμένεται να έχει μια μικρή μείωση 1.18% στο ποσοστό του G.Bush σε σχέση με πριν άμα βρίσκονται κάτω από τις ίδιες συνθήκες (δηλαδή παραμείνουν σταθερές όλες οι υπόλοιπες τιμές των μεταβλητών του μοντέλου).

Για την β9, του ποσοστό πληθυσμού με εισόδημα μεγαλύτερο από 75K ανά πολιτεία, συμπεραίνουμε ότι αν αυξηθεί το ποσοστό αυτό του πληθυσμού με εισόδημα μεγαλύτερο από 75K κατά 1%, αναμένεται να έχει μια μείωση 3.1% στο ποσοστό του G.Bush σε σχέση με πριν άμα βρίσκονται κάτω από τις ίδιες συνθήκες (δηλαδή παραμείνουν σταθερές οι τιμές των μεταβλητών).

Με βάση το πίνακα του μοντέλου, βλέπουμε ότι κάποιες από τις μεταβλητές είναι στατιστικά μη σημαντικές, ενώ ταυτόχρονα παρουσιάζονται προβλήματα πολυσυγγραμμικότητας και γραμμικότητας. Για να τα ξεπεράσουμε θα χρησιμοποιήσουμε μια μέθοδο επιλογής μεταβλητών η οποία προτείνει μοντέλα με βάση διαφορετικά κριτήρια όπως το AIC(akaike−criterion),BIC. Η μέθοδος που θα επιλέξουμε είναι η Stepwise-regression, η οποία συγκρίνοντας τις διαφορετικές τιμές AIC,BIC των διαφορετικών πιθανόν μοντέλων μας οδηγούν στο βέλτιστο. Πραγματοποιώντας την stepwise regression και για τις δύο τιμές AIC, BIC παρατηρούμε ότι σύμφωνα με το BIC το βέλτιστο μοντέλο δε περιλαμβάνει την μεταβλητή Population ενώ με το AIC την περιλαμβάνει. Επιλέγουμε, την stepwise regression με άξονα το AIC ώστε να διατηρήσουμε την μοναδική κατηγορική μεταβλητή μας και ταυτόχρονα την πληροφορία που μας δίνει. Μέσω του έλεγχου πολυσυγγραμμικότητας αφαιρέσαμε την μεταβλητή Male(εμφάνιζε την μεγαλύτερη τιμή) και απαλλαχθήκαμε από το πρόβλημα, ενώ στην συνέχεια αφαιρέσαμε την στατιστικά μη σημαντική μεταβλητή Management_18(p-value=0.47>0.05). Οπότε το μοντέλο στο οποίο οδηγούμαστε είναι το παρακάτω:

	Bush		
Predictors	*Estimates*	*std. Error*	*p*
(Intercept)	46.85	1.81	**<0.001**
Population [medium]	2.53	2.25	0.266
Population [high]	6.02	2.70	**0.031**
Rural	0.17	0.06	**0.005**
CLFU	-2.55	0.61	**<0.001**
Percentage 65	-1.59	0.42	**<0.001**
Num 75	-4.79	0.93	**<0.001**
Observations	51		
R^2 / R^2 adjusted	0.768 / 0.736		
AIC	324.336		

Πίνακας 4: Μοντέλο Πολλαπλής Παλινδρόμησης μετά από Model – Selection(Stepwise regression βασιζόμενοι στο δείκτη AIC)

Βλέπουμε, ότι το ποσοστό των ψήφων που έλαβε ο G.Bush ανά πολιτεία, εξαρτάται από το μέγεθος του αριθμού του πληθυσμού της, το ποσοστό πληθυσμού που ζει σε μη αστικές περιοχές, το ποσοστό ανεργίας ανά πολιτεία, το ποσοστό του πληθυσμού που είναι μεγαλύτεροι από 65 ετών και από το ποσοστό πληθυσμού με εισόδημα μεγαλύτερο από 75K σε αυτήν.

Το αναμενόμενο ποσοστό του G.Bush σε πολιτείες χαμηλού αριθμού πληθυσμού, με μηδενικό ποσοστό πληθυσμού που ζει σε μη αστικές περιοχές, με μηδενικό ποσοστό ανεργίας ανά πολιτεία,με μηδενικό ποσοστό του πληθυσμού που είναι μεγαλύτεροι από 65 ετών και με μηδενικό ποσοστό πληθυσμού με εισόδημα μεγαλύτερο από 75K αναμένεται να είναι 46.85%. Η αναμενόμενη αύξηση του ποσοστού του G.Bush, αν μια πολιτεία έχει μεσαίο μέγεθος πληθυσμού σε σχέση με μια που έχει μικρό, αναμένεται να είναι 2.53%, ,δεδομένου ότι θα μείνουν σταθερά τα υπόλοιπα χαρακτηριστικά. Η αναμενόμενη αύξηση του ποσοστού του G.Bush, αν μια πολιτεία έχει μεγάλο μέγεθος πληθυσμού σε σχέση με μια που έχει μικρό, αναμένεται να είναι 6%, ,δεδομένου ότι θα μείνουν σταθερά τα υπόλοιπα χαρακτηριστικά. Η αναμενόμενη αύξηση του ποσοστού του G.Bush, αν αυξηθεί κατά 1% το ποσοστό πληθυσμού που ζει σε μη αστικές περιοχές και δεδομένου ότι παραμείνουν σταθερά όλα τα υπόλοιπα χαρακτηριστικά, πρόκειται να είναι 0.17%. Η αναμενόμενη μείωση του ποσοστού του G.Bush, αν αυξηθεί κατά 1% το ποσοστό ανεργίας/πολιτεία, δεδομένου ότι παραμείνουν σταθερά όλα τα υπόλοιπα χαρακτηριστικά, πρόκειται να είναι 2.55%. Η αναμενόμενη μείωση ποσοστού του G.Bush, αν αυξηθεί κατά 1% το ποσοστό πληθυσμού που είναι μεγαλύτεροι από 65 ετών, δεδομένου ότι παραμείνουν σταθερά τα υπόλοιπα χαρακτηριστικά, πρόκειται να είναι 1.6% και ομοίως αύξηση 1% στο ποσοστό πληθυσμού με εισόδημα άνω των 75K, αναμένεται να μειώσει 4.8% το ποσοστό του G.Bush.

Για το μοντέλο που δημιουργήσαμε θα ελέγξουμε κατά πόσο «συμφιλιώνεται» με την κανονικότητα καταλοίπων, ομοσκεδαστικότητα καταλοίπων, ανεξαρτησία καταλοίπων, πολυσυγγραμμικότητα μεταβλητών, καθώς και γραμμικότητα. Με βάση τα διαγράμματα του μοντέλου (Βλέπε Σχήμα 45,46), καθώς και με τους ελέγχους που χρησιμοποιήσαμε λαμβάνουμε: Κανονικότητα Καταλοίπων (S-W: p-value=0.20>0.05&Lillie:p-value=0.54>0.05),Ομοσκεδαστικότητα Καταλοίπων (Levenes:p-value:0.18>0.05), Ανεξαρτησία καταλοίπων (D-W:p-value:0.15) και απορρίπτουμε την πολυσυγγραμμικότητα (τιμές <10 για ποσοτικές και <3.16 για κατηγορικές με πολλά επίπεδα). Η

γραμμή στο plot(Residuals vs Fitted values) δεν είναι απόλυτα επίπεδη, αλλά δεν υπάρχουν μεγάλα μοτίβα ή patterns , που σημαίνει ότι δεν υπάρχουν σημαντικές μη γραμμικότητες. (Βλέπε Πίνακα 18 για όλους τους ελέγχους). Τέλος, το ποσοστό της μεταβλητότητας που εξηγεί το μοντέλο είναι R2 = 0.77, το οποίο είναι ενθαρρυντικό για την ανάλυση που πραγματοποιούμε.

Παρά τα επιθυμητά αποτελέσματα του μοντέλου στις υποθέσεις ελέγχου που πραγματοποιήσαμε, θα ήταν χρήσιμο να αναλογιστούμε σενάρια μετασχηματισμού με σκοπό να το κάνουμε ακόμα καλύτερο βελτιώνοντας και την προβλεπτικότητα του. Μετασχηματισμοί που πραγματοποιήθηκαν πάνω στην μεταβλητή Bush(τετραγωνικής ρίζας, λογαρίθμου κ.α), οδηγούσαν σε απόρριψη υποθέσεων και σε πιο αδύναμα μοντέλα. Μετά από δοκιμές πάνω στις επεξηγηματικές μεταβλητές βρήκαμε ενδιαφέρον την λογαριθμική μετατροπή της μεταβλητής Num_75,ποσοστό πληθυσμού με εισόδημα άνω των 75K, η οποία όπως θα δούμε και από τον Πίνακα 2 εμφάνιζε υψηλή ασυμμετρία(skewness).

	Bush		
Predictors	*Estimates*	*std. Error*	*p*
(Intercept)	46.83	1.74	**<0.001**
Population [medium]	2.29	2.18	0.299
Population [high]	6.58	2.57	**0.014**
Rural c	0.15	0.06	**0.011**
CLFU c	-2.80	0.60	**<0.001**
Percentage 65 c	-1.68	0.40	**<0.001**
log Num 75 c	-12.63	2.27	**<0.001**
Observations	51		
R^2 / R^2 adjusted	0.781 / 0.751		
AIC	321.246		

Πίνακας 5: Μοντέλο πολλαπλής παλινδρόμησης με μετασχηματισμό log(Num_75), λογαριθμικό μετασχηματισμό του ποσοστού πληθυσμού με εισόδημα μεγαλύτερο από 75K

Ο μετασχηματισμός αυτός όχι μόνο εμφανίζει καλύτερο AIC και BIC (πιο μικρά από πριν) (Βλέπε Πίνακα 5), αλλά μας έδινε και καλύτερα αποτελέσματα στους ελέγχους υποθέσεων σε σχέση με κάθε άλλο λογαριθμικό μετασχηματισμό επεξηγηματικής μεταβλητής που δοκιμάστηκε (Βλέπε Πίνακα 19) . Παρέμεινε υψηλά το R2/R2adj = 0.78/0.75 και όλες οι μεταβλητές που είχαμε από το προηγούμενο μοντέλο διατηρήθηκαν στατιστικά σημαντικές. Με σκοπό η ερμηνεία του μοντέλου μας να παραμείνει ρεαλιστική πραγματοποιήσαμε κεντροποίηση των ποσοτικών μεταβλητών. Εάν προσπαθούσαμε να ερμηνεύσουμε το μοντέλο αυτό ως προς τις τιμές β0,β1…β6, θα ακολουθούσαμε ίδια τακτική με τις ερμηνείες μοντέλων που έγιναν παραπάνω, εκτός του δείκτη β6, που αφορά τη μεταβλητή log(Num_75) που θα ακολουθήσει την εξής ερμηνεία: Για κάθε μια μονάδα αύξησης της τιμής της μεταβλητής log(Num_75), το ποσοστό ψήφων του George Bush μειώνεται κατά 12.63%, αν όλα τα υπόλοιπα χαρακτηριστικά (μεταβλητές) παραμείνουν σταθερά στο μοντέλο.

Είναι σημαντικό να τονίσουμε ότι εξαιτίας των λίγων παρατηρήσεων που έχουμε στο dataset ,πιθανόν να υπερεκτιμήσουμε τα μοντέλα και το κατά πόσο αυτά επαληθεύουν τους ελέγχους υποθέσεων που πραγματοποιούμε. Το μοντέλο που προκύπτει είναι:

Bush(Ποσοστό G.Bush) = 46.83+2.29*Population_medium(Κατηγορία μεγέθους πληθυσμού μεσαία)+6.58*Population_high(Κατηγορία μεγέθους πληθυσμού μεγάλη)+0.15*Rural(Ποσοστό σε μη αστικές περιοχές) -2.80*Clfu(Ποσοστό ανεργίας)-1.68 *Percentage_65(Ποσοστό πληθυσμού άνω των 65 ετών)-12.63*log(Num_75)(λογαριθμικό ποσοστό πληθυσμού με εισόδημα μεγαλύτερο από 75K) + ε , ε~N(0, 5.194^2).

5. Συμπεράσματα και συζήτηση

Η ανάλυση που πραγματοποιήσαμε είχε ως σκοπό την εκτίμηση ενός υποδείγματος για τους παράγοντες που επηρέασαν το ποσοστό που έλαβε ο George Bush στις εκλογές του 2000 ανά πολιτεία και τις μεταξύ τους σχέσεις των παραγόντων αυτών. Φάνηκαν λοιπόν μια σειρά από μεταβλητές, οι οποίες επηρεάζουν τα αποτελέσματα μιας εκλογικής διαδικασίας και διαμορφώνουν την τελική έκβαση της. Τόσο οικονομικοί παράγοντες όπως τα ποσοστά ανεργίας ανά πολιτεία, τα ποσοστά του πληθυσμού με υψηλό μισθό, όσο και γεωγραφικοί/δημογραφικοί που αφορούν το πληθυσμό που κατοικεί σε μη αστικές περιοχές, το μέγεθος του αριθμού του πληθυσμού και η ηλικία και το φύλο των ψηφοφόρων, είναι σημαντικές συνιστώσες στην έκβαση των εκλογών. Όλα τα παραπάνω είναι στοιχεία που δείχνουν την κατάσταση στην οποία βρίσκεται μια πολιτεία και πιθανές παθογένειες της, τις οποίες ένας πολιτικός αρχηγός οφείλει να αναλογιστεί για να δυναμώσει την ισχύ του.

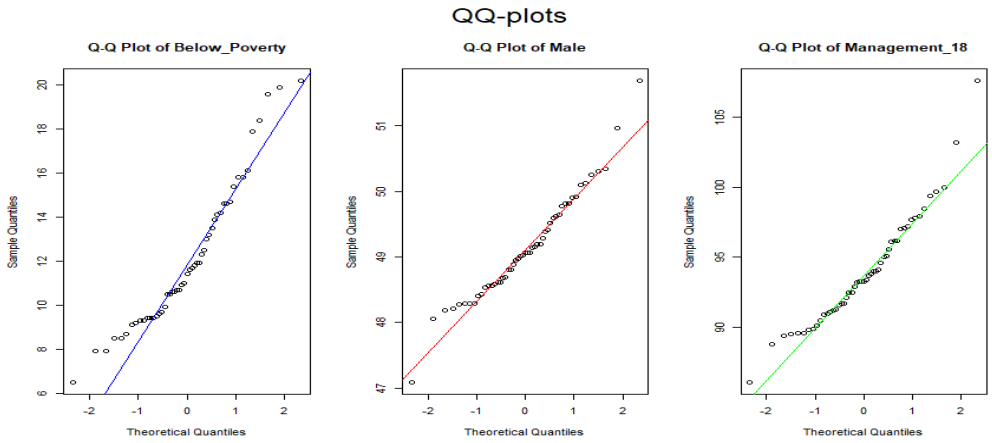
Ελπίζουμε τα αποτελέσματα της μελέτης να βοηθήσουν μελλοντικά, νέα πολιτικά πρόσωπα ως προς την ενημέρωση τους, στους παράγοντες που διαμορφώνουν το ποσοστό ενός κόμματος ή πολιτικού. Επιπλέον, να δώσει το έναυσμα για μελέτες και σε άλλες χώρες/κράτη ανά τον κόσμο επεκτείνοντας την έρευνα στα πολιτικά δρώμενα, διαμορφώνοντας μια πιο ξεκάθαρη εικόνα ως προς το τι κρύβεται πίσω από την επιτυχία ή αποτυχία ενός πολιτικού.

Αναφορές

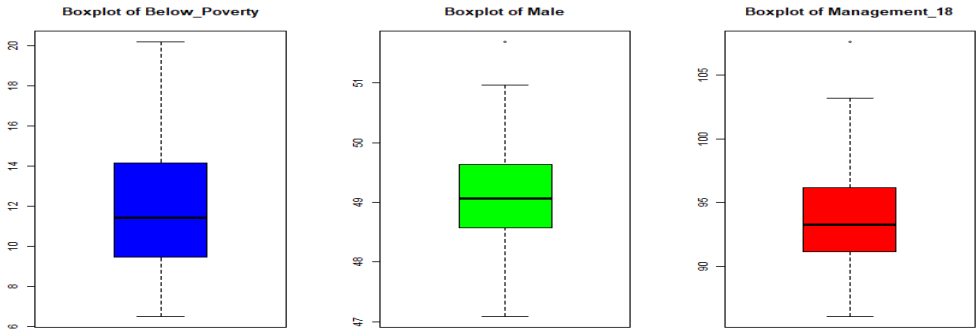
Leip, D (2019) Dave Leip's Atlas of U.S. Presidential Elections.
https://uselectionatlas.org/

Παράρτημα

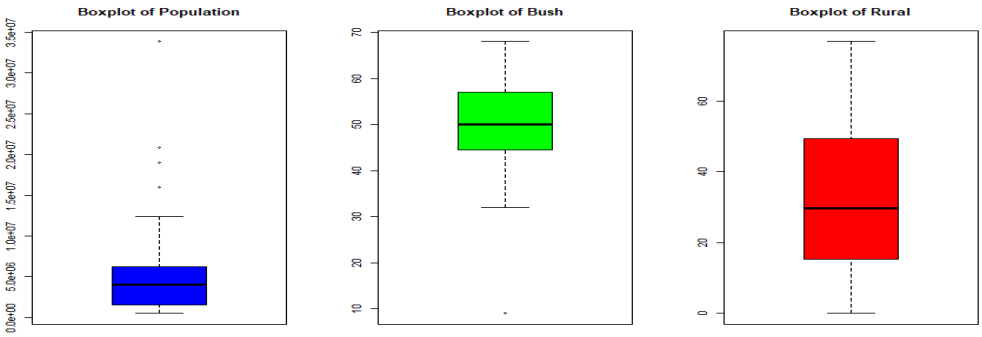
Υπενθύμιση: Να σημειωθεί ότι στο παράρτημα ακολουθήσαμε την εξής τακτική: Πρώτα αναλύσαμε το αρχικό dataset εξολοκλήρου (Περιγραφική ανάλυση+ Σχέσεις ανά δύο) και μετά αναλύσαμε τα δύο subsets που προέκυψαν από την διάμεσο του Population με τον ίδιο τρόπο. Για αυτό μπορεί η σειρά των εικόνων να μην ταυτίζεται πλήρως με την σειρά των ερωτημάτων όπως απαντώνται στην παραπάνω ανάλυση που κάναμε. Χρησιμοποιώντας όμως αναφορές (Βλέπε Σχήμα Χ) κάθε φορά στην εικόνα/πίνακα που αναφερόμαστε προσπαθήσαμε να αποφύγουμε την σύγχυση.



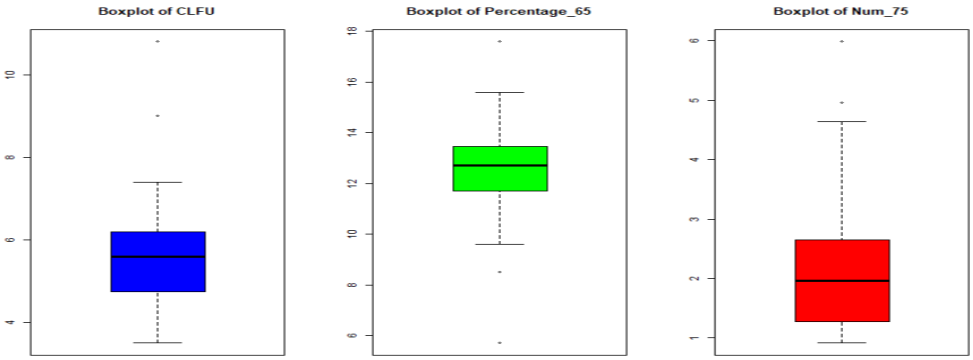
Σχήμα 3: QQ-Plots των μεταβλητών που εξετάζουμε αν προσεγγίζουν ή αποκλίνουν από την κανονική κατανομή. Η μεταβλητή Male φαίνεται να την ακολουθεί, ενώ οι υπόλοιπες να αποκλίνουν.



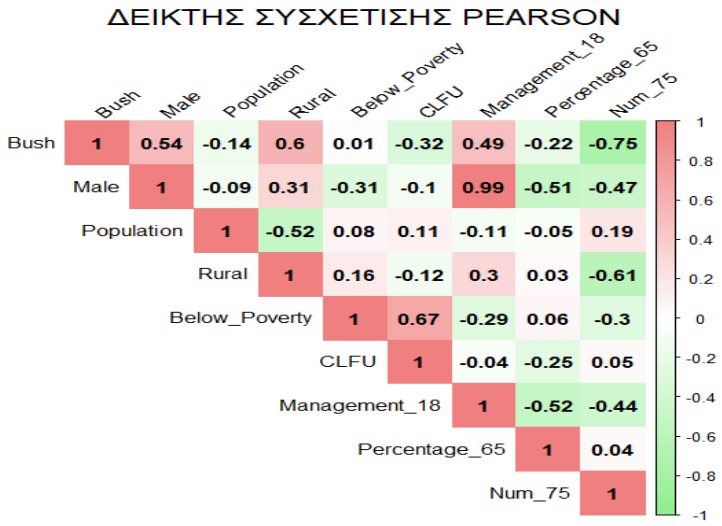
Σχήμα 4: Boxplot μεταβλητών Below_Poverty, Male, Management_18



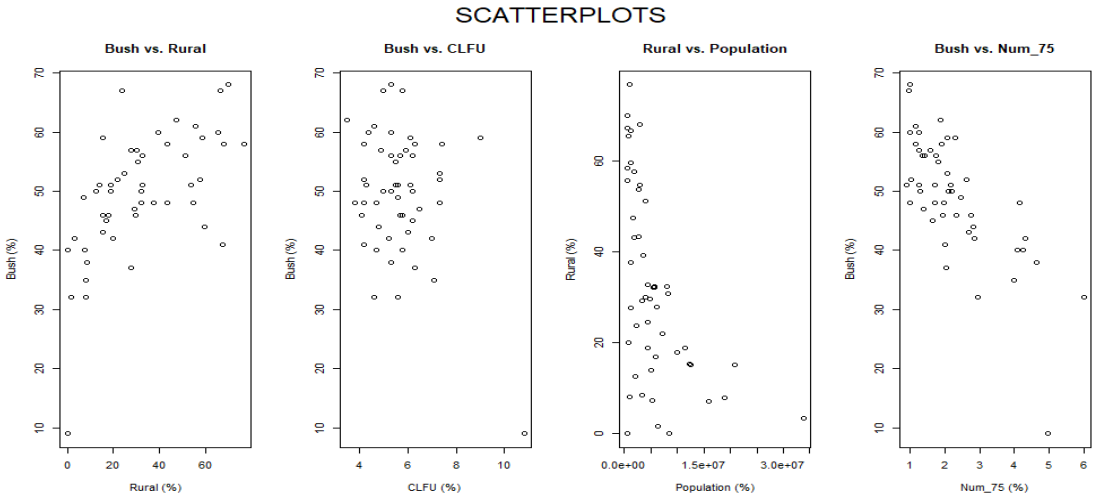
Σχήμα 5: Boxplot μεταβλητών Population, Bush, Rural



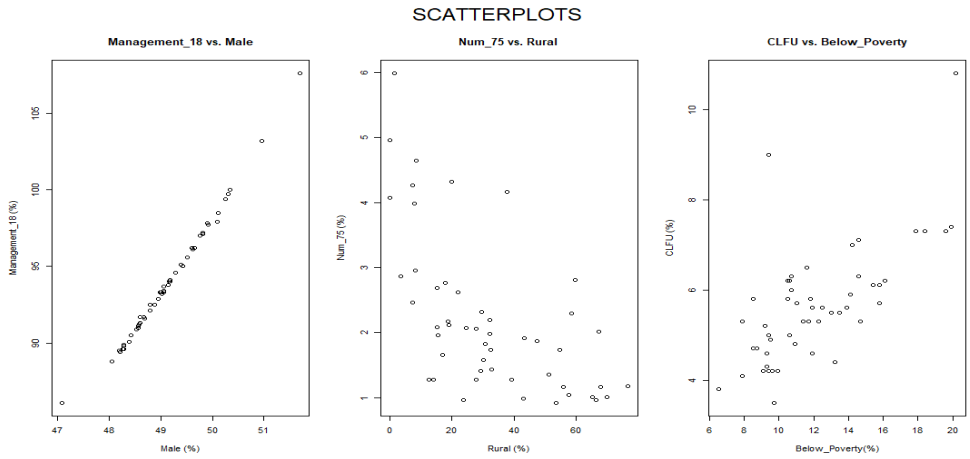
Σχήμα 6: Boxplot μεταβλητών CLFU, Percentage_65, Num_75



Σχήμα 7: Δείκτες συσχέτισης Pearson των μεταβλητών του Dataset



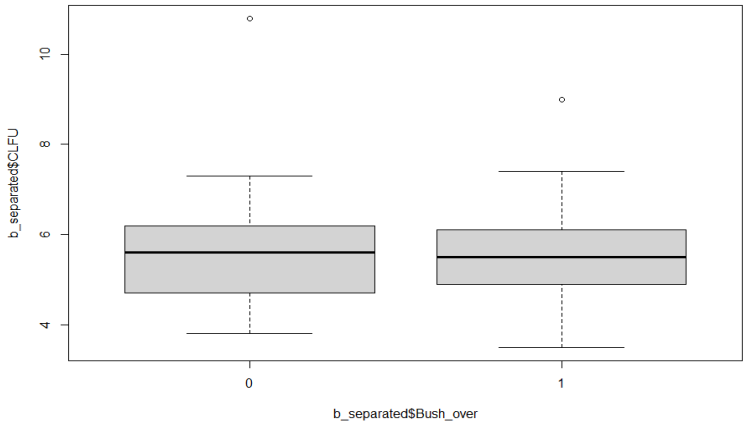
Σχήμα 8: Scatterplots των μεταβλητών Bush-Rural, Bush-CLFU,Rural-Population,Bush-Num_75



Σχήμα 9: Scatterplots των μεταβλητών Male-Management_18,Rural-Num_75,Below_Poverty-CLFU

Μεταβλητές	Έλεγχος κανονικότητας Shapiro.test	Μέγεθος δειγμάτων n1,n2	Μη παραμετρικός έλεγχος Wilcox.test
Bush_over ~ CLFU	**p-value = 0.0015 <0.05 για Bush_over:0 (Απόρριψη μηδενικής υπόθεσης H0 για ύπαρξη κανονικότητας)** **p-value = 0.25>0.05 για Bush_over:1**	**25<50 και 26<50**	**p-value = 2.2e-16<0.05 Απορρίπτουμε την μηδενική υπόθεση για μηδενική διαφορά διαμέσων. Άρα υπάρχουν σημαντικές διαφοροποιήσεις στη διάμεσο των ποσοστών ανεργίας μεταξύ πολιτειών με μεγάλα ποσοστά George Bush και αυτών με μικρομεσαία**

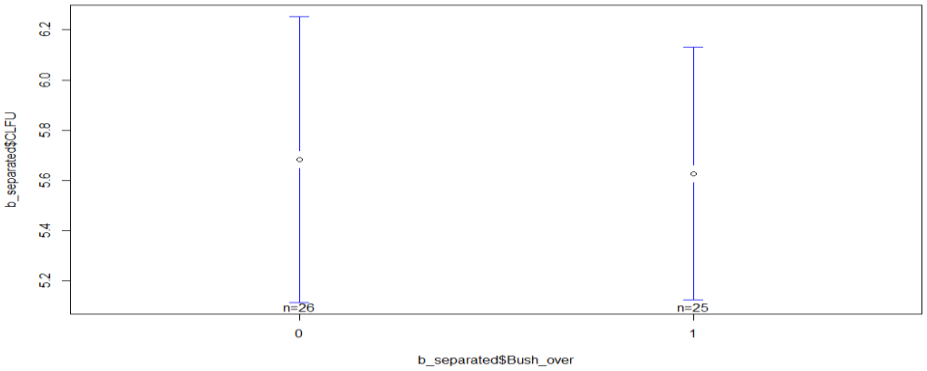
Πίνακας 6: Έλεγχοι για πιθανή εξάρτηση της μεταβλητής Bush_over από την μεταβλητή CLFU(με shapiro.test)



Σχήμα 10: Box-plot CLFU~Bush_over εξαιτίας μη παραμετρικού ελέγχου

Μεταβλητές	Έλεγχος Κανονικότητας Lillie.test	Έλεγχος ίσων διακυμάνσεων	Παραμετρικός έλεγχος t.test
Bush_over ~ CLFU	**p-value = 0.22>0.05 για Bush_over:0 p-value = 0.33 >0.05 για Bush_over:1** **Δεν απορρίπτεται η Κανονικότητα**	**p-value = 0.48>0.05 Δεν απορρίπτουμε την μηδενική υπόθεση H0 για ισότητα διακυμάνσεων**	**p-value = 0.88>0.05 Δεν απορρίπτουμε H0, επομένως δεν υπάρχουν σημαντικές διαφοροποιήσεις στα μέσα των ποσοστών ανεργίας για πολιτεία με μεγάλα ή μη ποσοστά George Bush**

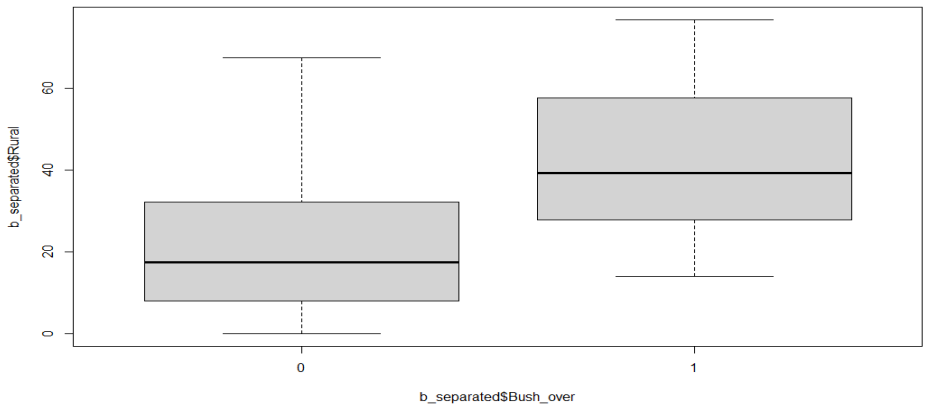
Πίνακας 7: Έλεγχοι για πιθανή εξάρτηση της μεταβλητής Bush_over από την μεταβλητή CLFU(με Lillie.test)



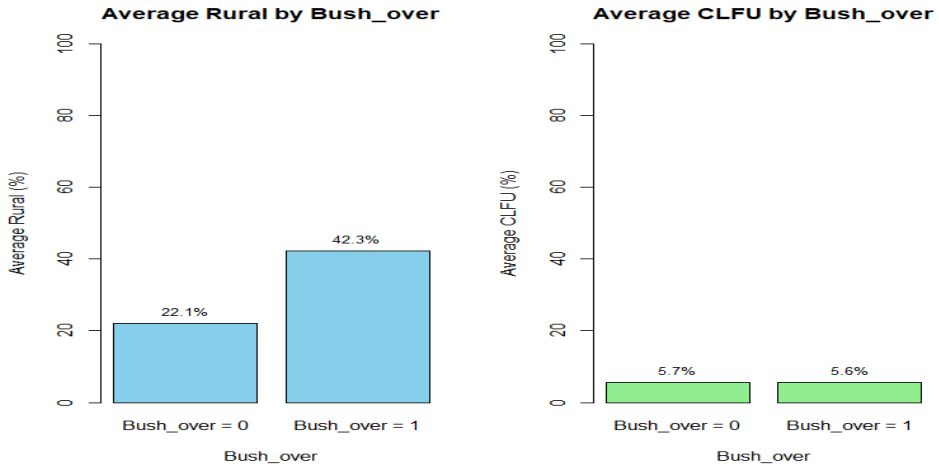
Σχήμα 11: Error-bar CLFU~Bush_over σε περίπτωση παραμετρικού ελέγχου

Μεταβλητές	Έλεγχοι κανονικότητας	Μέγεθος δειγμάτων n1,n2	Μη παραμετρικός έλεγχος Wilcox.test
Bush_over ~ Rural	**Shapiro.test: p-value = 0.022<0.05 (Bush_over:0) p-value = 0.17>0.05 (Bush_over:1)** **Lillie.test: p-value = 0.085>0.05 (Bush_over:0) p-value = 0.047<0.05 (Bush_over:1)** **Απορρίπτουμε και στους δύο ελέγχους την κανονικότητα και την H0(για ύπαρξη κανονικότητας)**	**25<50 και 26<50**	**p-value < 2.2e-16<0.05 Απορρίπτουμε την μηδενική υπόθεση για μηδενική διαφορά διαμέσων. Άρα υπάρχει σημαντική διαφορά στη διάμεσο του ποσοστού πληθυσμού που ζει σε μη αστικές περιοχές μεταξύ των πολιτειών με μεγάλα ποσοστά George Bush και αυτών με μικρομεσαία.**

Πίνακας 8: Έλεγχοι για πιθανή εξάρτηση της μεταβλητής Bush_over από την μεταβλητή Rural

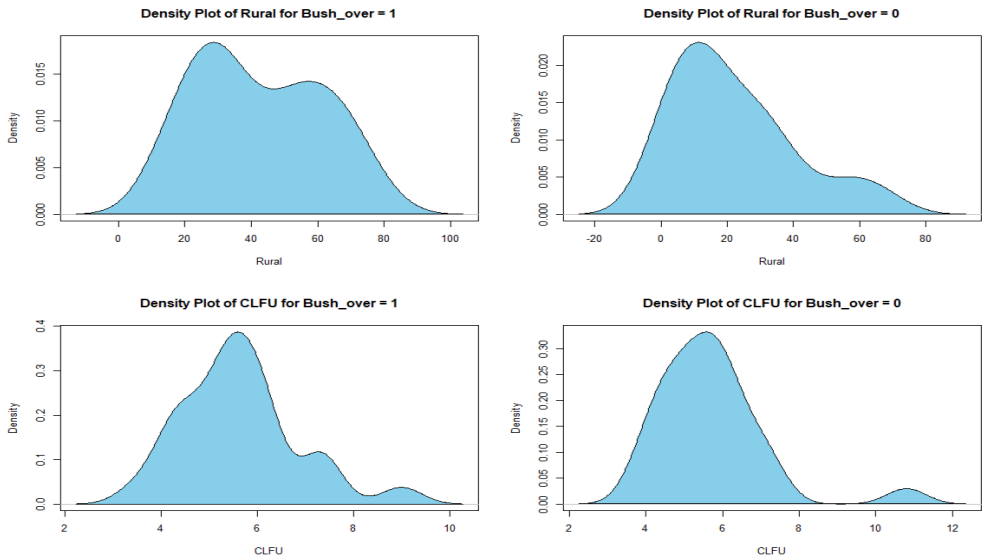


Σχήμα 12: Box-plot Rural~Bush_over εξαιτίας μη παραμετρικού ελέγχου

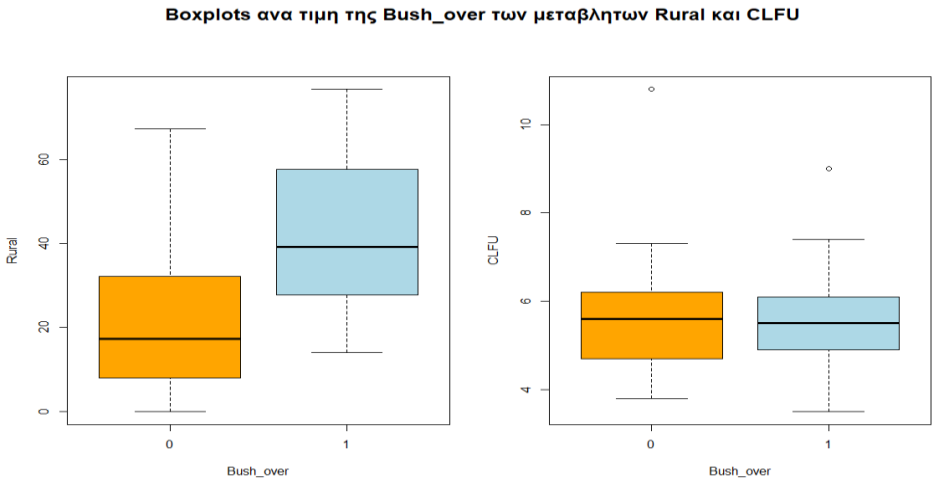


Σχήμα 13:Barplots των Average τιμών της Rural και CLFU ανά τιμή της Bush_over

(σε ποσοστά)



Σχήμα 14: Density plots των μεταβλητών Rural και CLFU για κάθε τιμή της Bush_over



Σχήμα 15: Boxplots ανά τιμή της Bush_over των μεταβλητών Rural και CLFU

FISHERS EXACT TEST	
Bush-Rural	p-value: 0.002<0.05, υπάρχει εξάρτηση
Bush-Num_75	p-value: 0.0002<0.05, υπάρχει εξάρτηση
Rural-Num_75	p-value: 0.0002<0.05, υπάρχει εξάρτηση

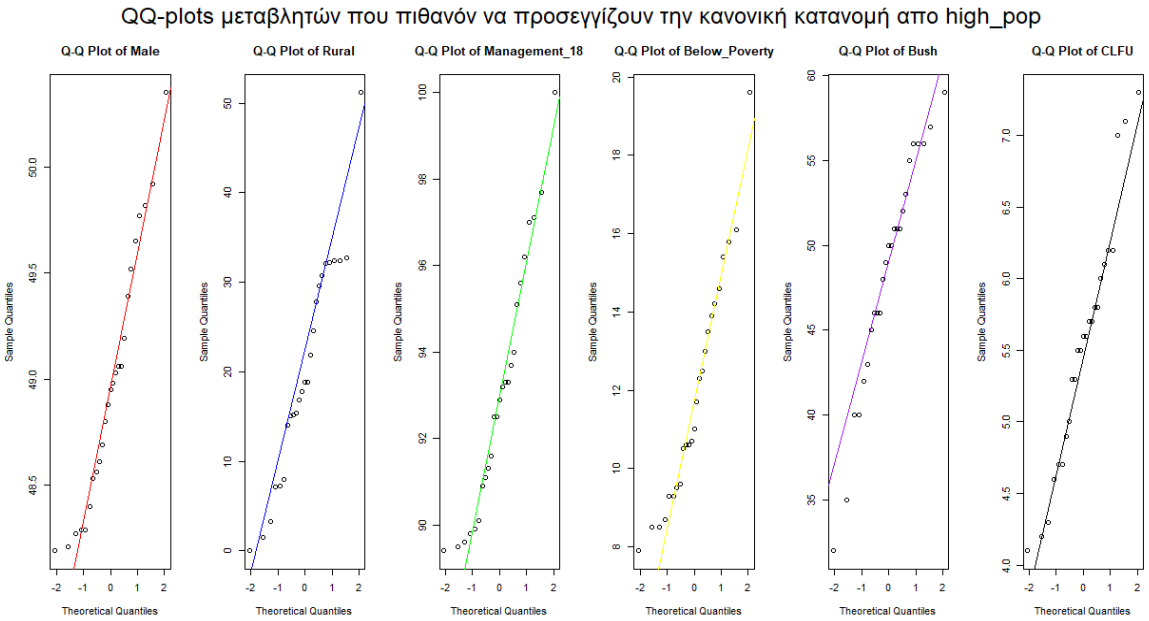
Πίνακας 9: Fischer έλεγχος αν μετατρέψουμε τις μεταβλητές σε δίτιμες κατηγορικές μέσω της διαμέσου τους.

Μεταβλητή	Μέσο	Τυπική Απόκλιση	Διάμεσος	Μικρότερη Τιμή	Μεγαλύτερη τιμή	Ασυμμετρία	Κύρτωση
Bush	48.36	6.93	50	32	59	-0.57	-0.44
Male	49	0.6	49	48.2	50.35	0.50	-0.81
Population	9399217.6	6868019.5	6349097	4041769	33871648	2.02	4.03
Rural	20.28	12.32	18.83	0.00	51.18	0.30	-0.39
Below_Poverty	11.9	2.94	11	7.9	19.60	0.69	-0.24
CLFU	5.53	0.86	5.60	4.10	7.30	0.26	-0.61
Management_18	93.1	2.92	92.90	89.40	100	0.55	-0.69
Percentage_65	12.38	1.74	12.40	9.60	17.60	0.87	1.49
Num_75	2.43	1.10	2.12	1.27	5.99	1.55	2.15

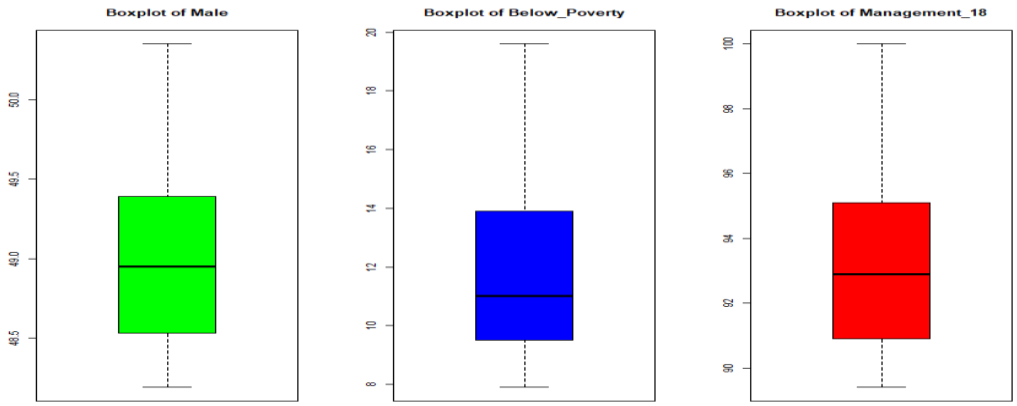
Πίνακας 10: Πίνακας περιγραφικών μέτρων για μεγάλο αριθμό Πληθυσμού (high_pop)

Μεταβλητή	Μέσο	Τυπική Απόκλιση	Διάμεσος	Μικρότερη Τιμή	Μεγαλύτερη τιμή	Ασυμμετρία	Κύρτωση
Bush	51	13	51.5	9	68	-1.19	1.85
Male	49.30	0.96	49.17	47.09	51.69	0.25	0.24
Population	1786210.23	1077203.86	1502608	493782	4012012	0.49	-1.16
Rural	43.26	21.93	45.42	0.00	76.76	-0.36	-1.14
Below_Poverty	12.31	3.67	11.50	6.50	20.20	0.73	-0.49
CLFU	5.78	1.63	5.45	3.50	10.80	1.18	1.48
Management_18	94.79	4.64	94.05	86.10	107.60	0.68	0.48
Percentage_65	12.67	2.04	13.10	5.70	15.30	-1.63	3.17
Num_75	1.99	1.24	1.50	0.92	4.96	1.21	0.10

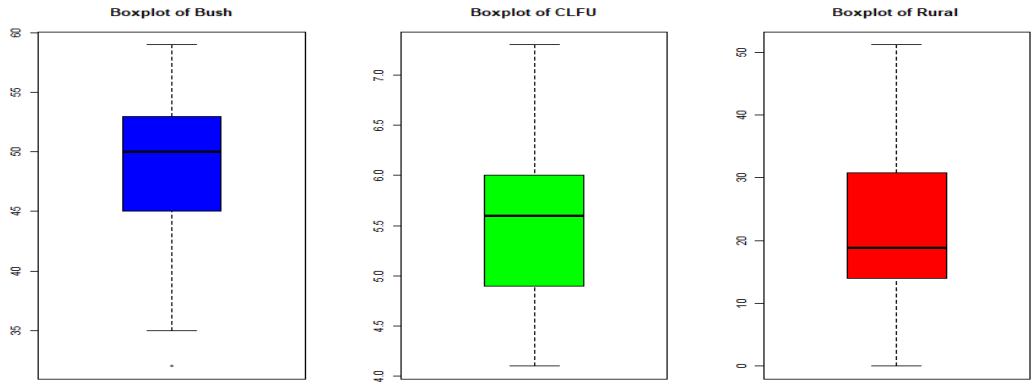
Πίνακας 11: Πίνακας περιγραφικών μέτρων για μικρό αριθμό Πληθυσμού (low_pop)



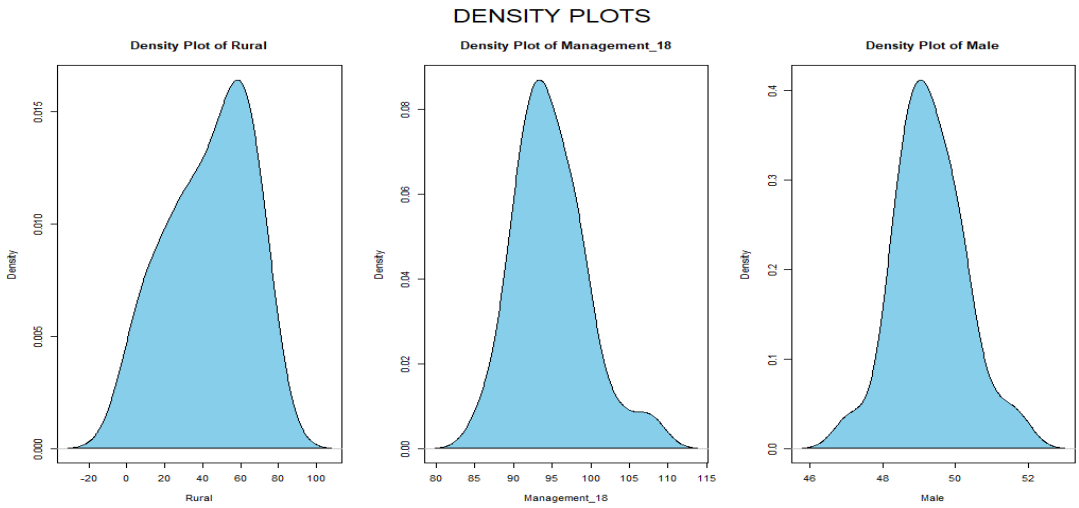
Σχήμα 16: QQ-plots μεταβλητών που πιθανόν να προσεγγίζουν την κανονική κατανομή απο high_pop



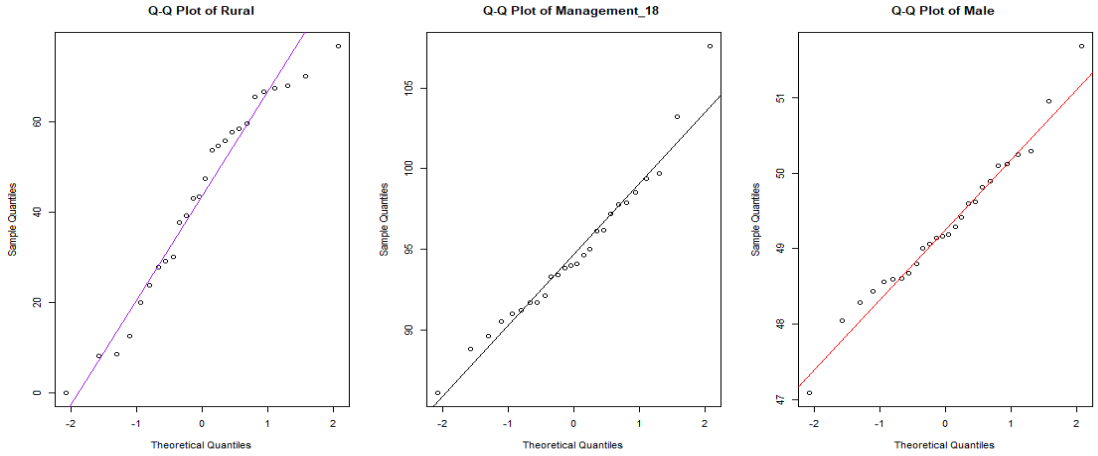
Σχήμα 17: Boxplot μεταβλητών Male, Below_Poverty, Management_18 από το subset High_pop(Μεγάλος αριθμός πληθυσμού)



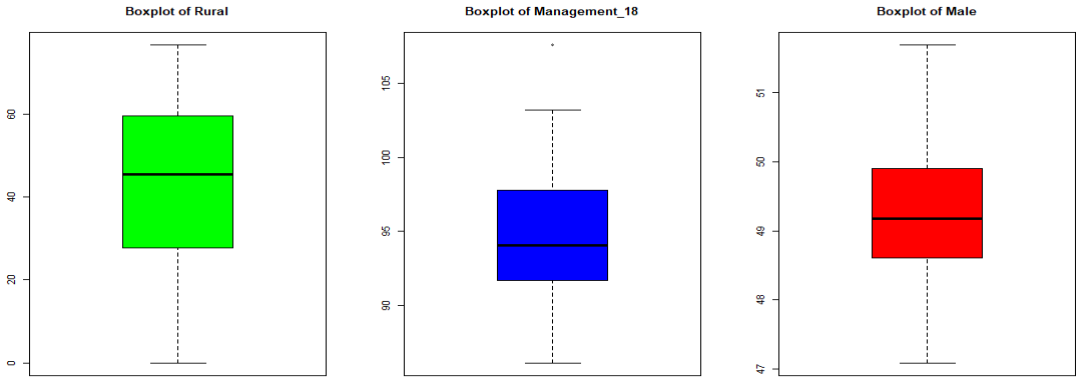
Σχήμα 18: Boxplot μεταβλητών Bush, CLFU, Rural από το subset High_pop(Μεγάλος αριθμός πληθυσμού)



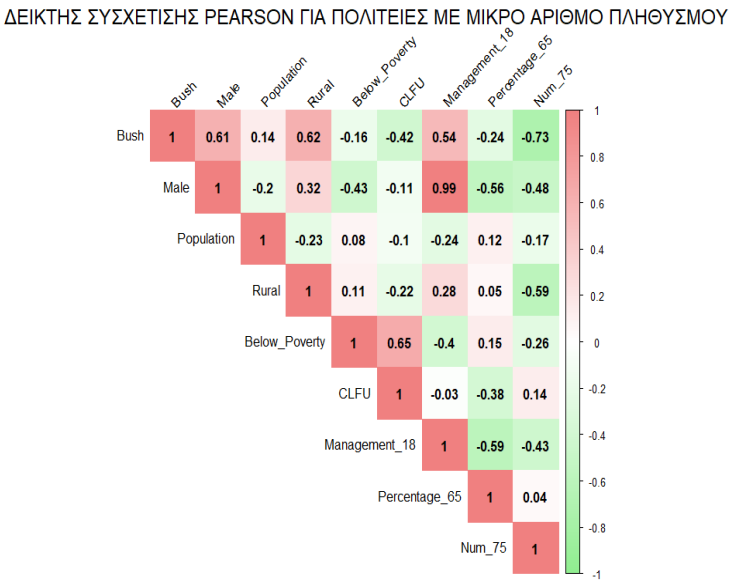
Σχήμα 19: Διαγράμματα πυκνότητας πιθανότητας για μεταβλητές Rural,Management_18,Male για μικρό μέγεθος πληθυσμών



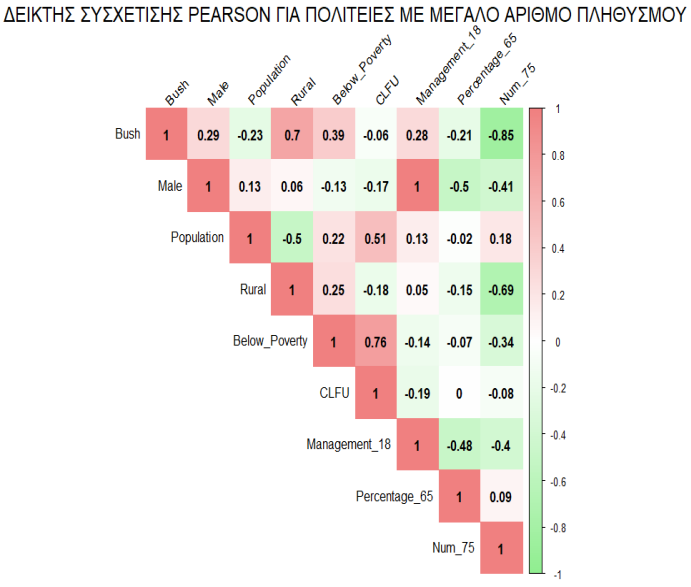
Σχήμα 20: QQ-Plots Rural,Management_18 και Male για low_pop (κανονική κατανομή)



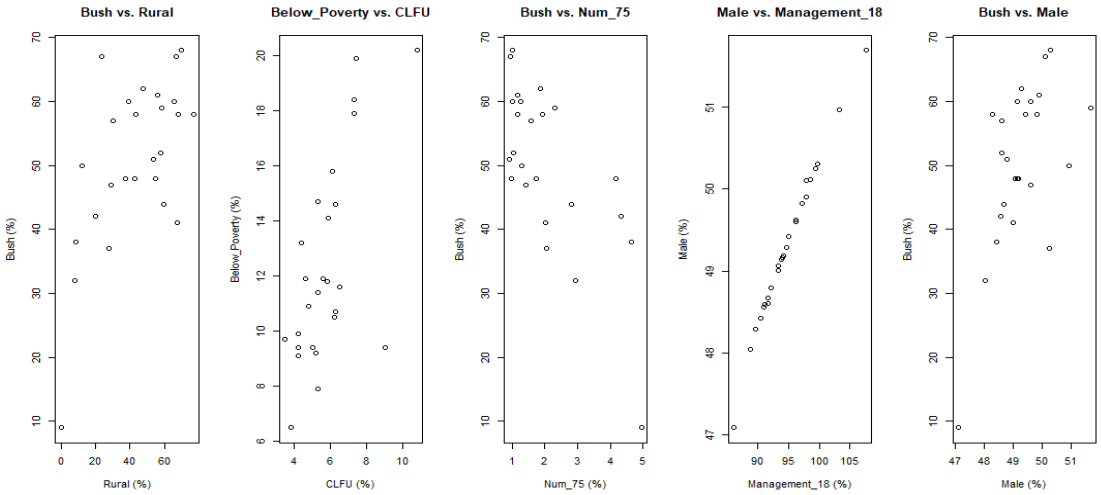
Σχήμα 21: Boxplot μεταβλητών Rural, Management_18,Male για το subset Low_pop(Μικρός αριθμός πληθυσμού)



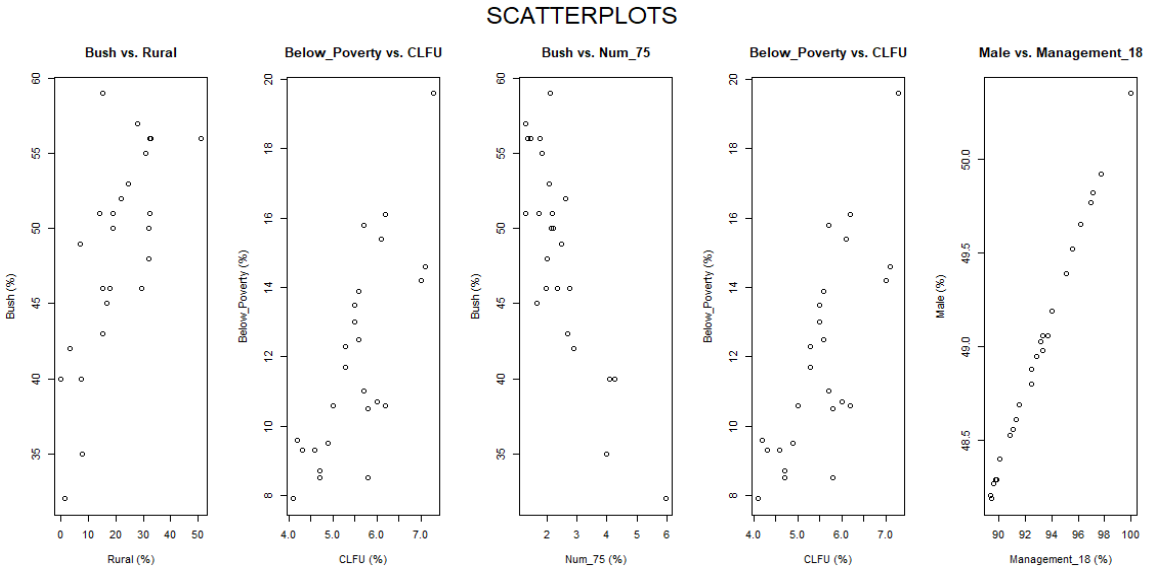
Σχήμα 22: Δείκτης συσχέτισης Pearson για πολιτείες με μικρό αριθμό πληθυσμού



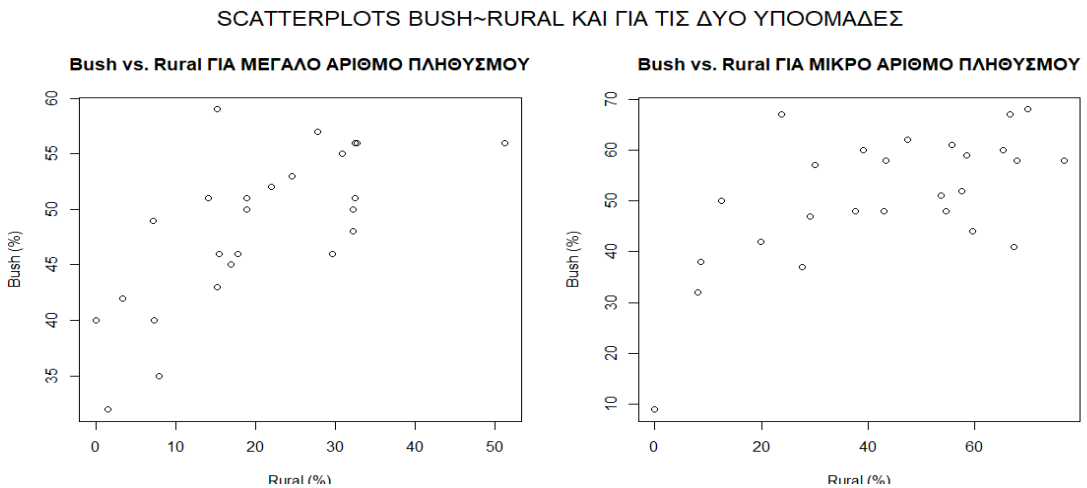
Σχήμα 23: Δείκτης συσχέτισης Pearson για πολιτείες με μεγάλο αριθμό πληθυσμού



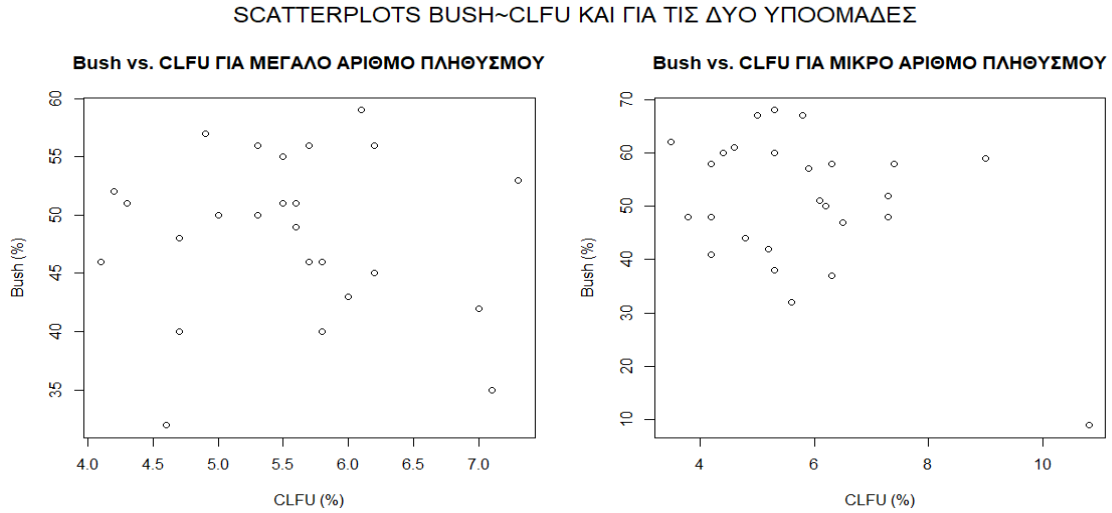
Σχήμα 24: Scatterplots των μεταβλητών με υψηλή συσχέτιση στις πολιτείες με μικρό αριθμό πληθυσμού



Σχήμα 25: Scatterplots των μεταβλητών με υψηλή συσχέτιση στις πολιτείες με μεγάλο αριθμό πληθυσμού



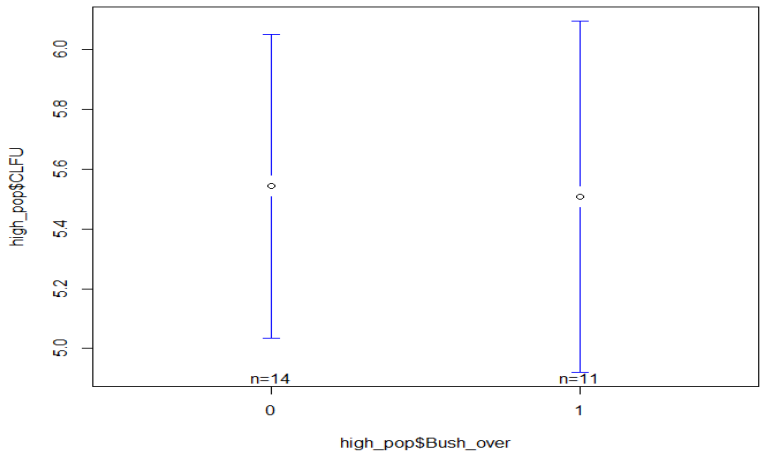
Σχήμα 26: Scatterplots Bush-Rural και στις δύο ομάδες πληθυσμού



Σχήμα 27: Scatterplots Bush-CLFU και στις δύο ομάδες πληθυσμού

Μεταβλητές	Έλεγχος Κανονικότητας Shapiro.test	Έλεγχος ίσων διακυμάνσεων	Παραμετρικός έλεγχος t.test
Bush_over ~ CLFU	**p-value = 0.72>0.05 για Bush_over:0 p-value = 0.66>0.05 για Bush_over:1 Δεν απορρίπτεται η Κανονικότητα**	**p-value = 0.99>0.05 Δεν απορρίπτουμε την μηδενική υπόθεση H0 για ισότητα διακυμάνσεων**	**p-value = 0.92>0.05 Δεν απορρίπτουμε H0, επομένως δεν υπάρχουν σημαντικές διαφοροποιήσεις στο μέσο ποσοστό ανεργίας ανάμεσα σε πολιτείες με μεγάλα ή χαμηλά ποσοστά George Bush**

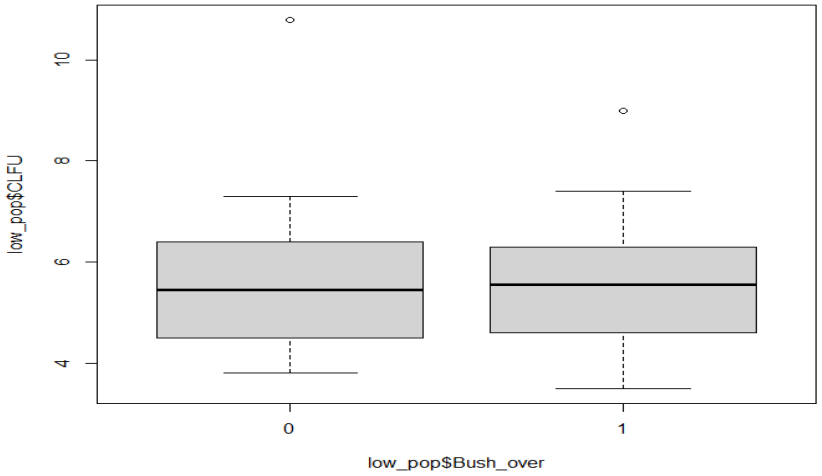
Πίνακας 12: Έλεγχοι για πιθανή εξάρτηση της μεταβλητής Bush_over από την μεταβλητή CLFU για την υποομάδα με τις πολιτείες με Υψηλό Πληθυσμό



Σχήμα 28: Error-bar CLFU~Bush_over σε περίπτωση παραμετρικού ελέγχου για υψηλό αριθμό πληθυσμού

Μεταβλητές	Έλεγχος Κανονικότητας shapiro.test	Έλεγχος μεγέθους δειγμάτων	Μη Παραμετρικός έλεγχος wilcoxon.test
Bush_over ~ CLFU	**p-value = 0.03<0.05 για Bush_over:0 p-value = 0.74>0.05 για Bush_over:1** **Απορρίπτεται η Κανονικότητα**	**n1<50 , n2<50**	**p-value = 3.427e-10<0.05 Απορρίπτουμε Η0, επομένως υπάρχουν σημαντικές διαφοροποιήσεις στις διαμέσους του ποσοστό ανεργίας ανάμεσα σε πολιτείες με μεγάλα ή χαμηλά ποσοστά George Bush**

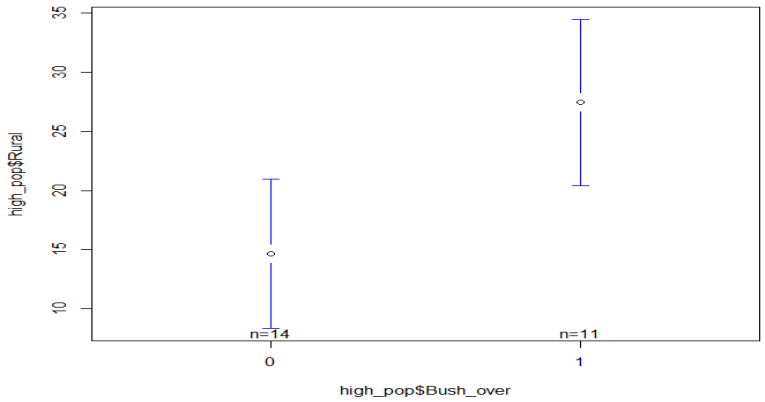
Πίνακας 13: Έλεγχοι για πιθανή εξάρτηση της μεταβλητής Bush_over από την μεταβλητή CLFU για την υποομάδα με τις πολιτείες με Χαμηλό Πληθυσμό



Σχήμα 29: Box-plot CLFU~Bush_over εξαιτίας μη παραμετρικού ελέγχου, σε χαμηλό αριθμό πληθυσμού

Μεταβλητές	Έλεγχος Κανονικότητας shapiro.test	Έλεγχος ίσων διακυμάνσεων	Παραμετρικός έλεγχος t.test
Bush_over ~ Rural	**p-value = 0.20>0.05 για Bush_over:0 p-value = 0.26>0.05 για Bush_over:1 Δεν απορρίπτεται η Κανονικότητα**	**p-value = 0.91>0.05 Δεν απορρίπτουμε την μηδενική υπόθεση H0 για ισότητα διακυμάνσεων**	**p-value = 0.007<0.05 Απορρίπτουμε H0, επομένως υπάρχουν σημαντικές διαφοροποιήσεις στο μέσο ποσοστό πληθυσμού που ζει σε μη αστικές περιοχές ανάμεσα σε πολιτείες με μεγάλα ή χαμηλά ποσοστά George Bush**

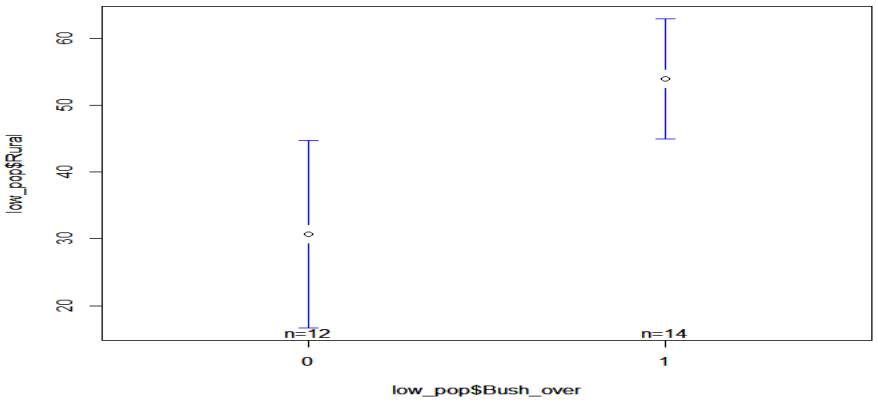
Πίνακας 14: Έλεγχοι για πιθανή εξάρτηση της μεταβλητής Bush_over από την μεταβλητή Rural για την υποομάδα με τις πολιτείες με Υψηλό Πληθυσμό



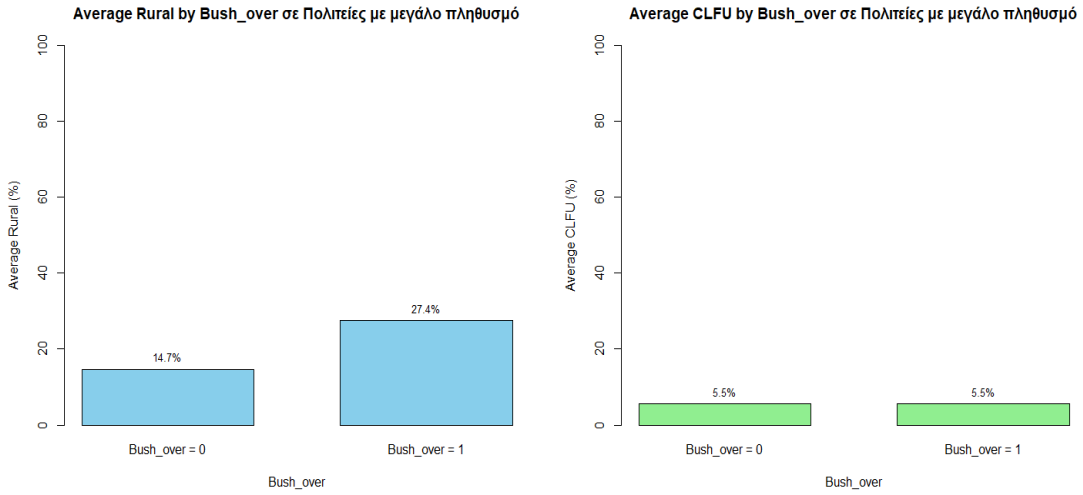
Σχήμα 30: Error-bar Rural~Bush_over σε περίπτωση παραμετρικού ελέγχου για υψηλό αριθμό πληθυσμού

Μεταβλητές	Έλεγχος Κανονικότητας shapiro.test	Έλεγχος ίσων διακυμάνσεων	Παραμετρικός έλεγχος t.test
Bush_over ~ Rural	**p-value = 0.64 >0.05 για Bush_over:0 p-value = 0.67 >0.05 για Bush_over:1** **Δεν απορρίπτεται η Κανονικότητα**	**p-value = 0.23 >0.05 Δεν απορρίπτουμε την μηδενική υπόθεση H0 για ισότητα διακυμάνσεων**	**p-value = 0.004 <0.05 Απορρίπτουμε H0, επομένως υπάρχουν σημαντικές διαφοροποιήσεις στο μέσο ποσοστό πληθυσμού που ζει σε μη αστικές περιοχές ανάμεσα σε πολιτείες με μεγάλα ή χαμηλά ποσοστά George Bush**

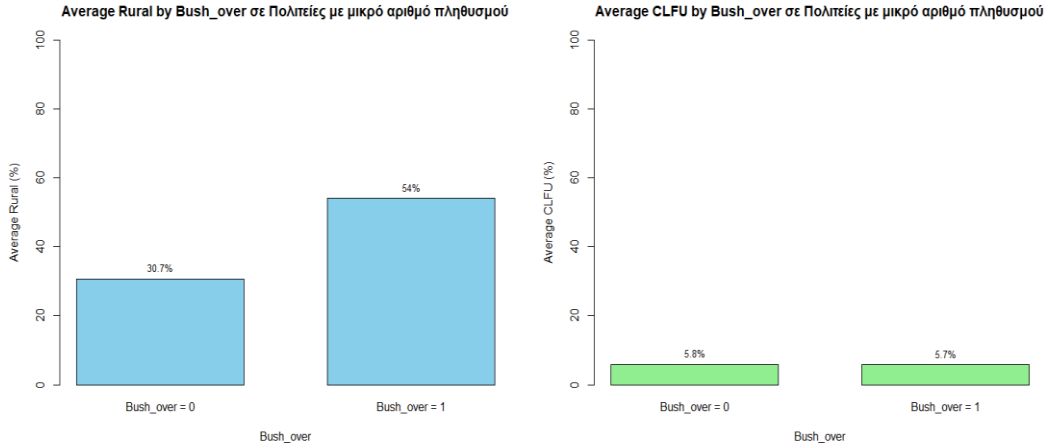
Πίνακας 15: Έλεγχοι για πιθανή εξάρτηση της μεταβλητής Bush_over από την μεταβλητή Rural για την υποομάδα με τις πολιτείες με Χαμηλό Πληθυσμό



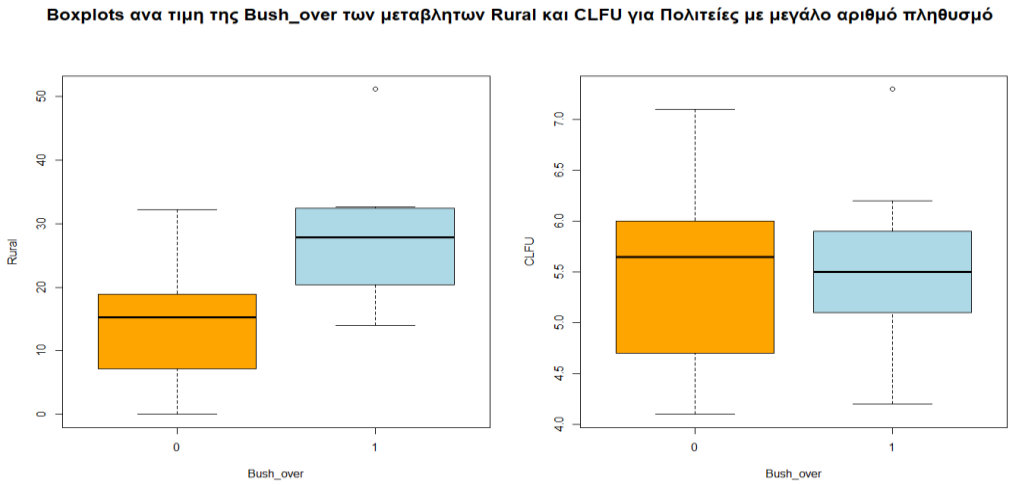
Σχήμα 31: Error-bar Rural~Bush_over σε περίπτωση παραμετρικού ελέγχου για χαμηλό αριθμό πληθυσμού



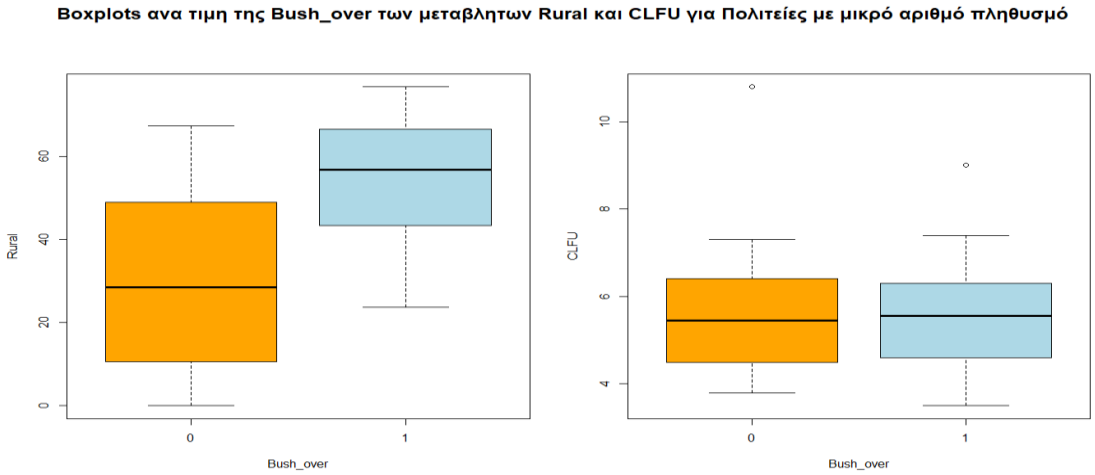
Σχήμα 32: Barplots των μεταβλητών Rural και CLFU ανά τιμή της Bush_over για τις πολιτείες με υψηλό πληθυσμό



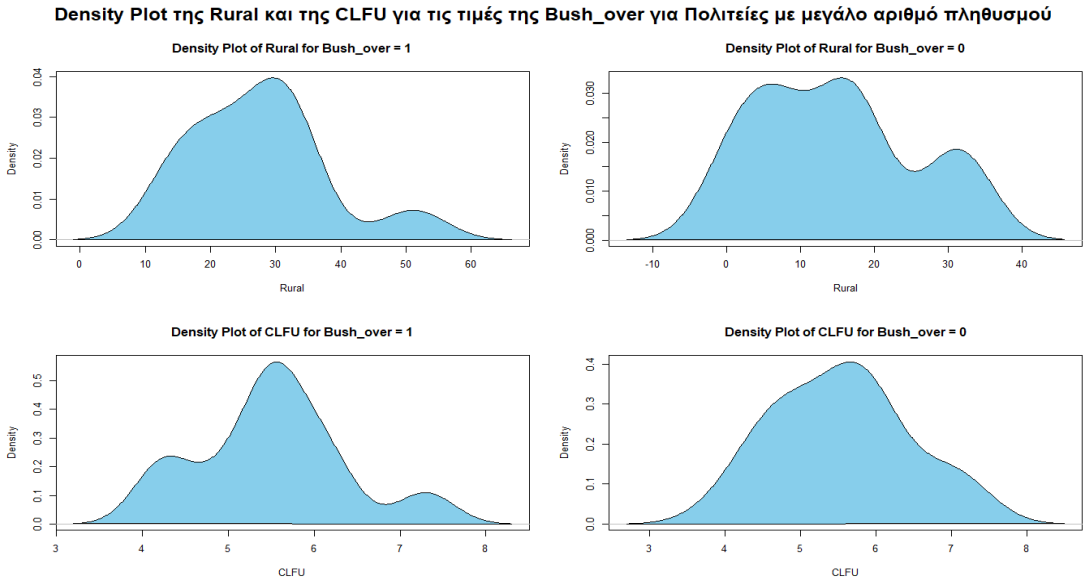
Σχήμα 33: Barplots των μεταβλητών Rural και CLFU ανά τιμή της Bush_over για τις πολιτείες με χαμηλό αριθμό πληθυσμό



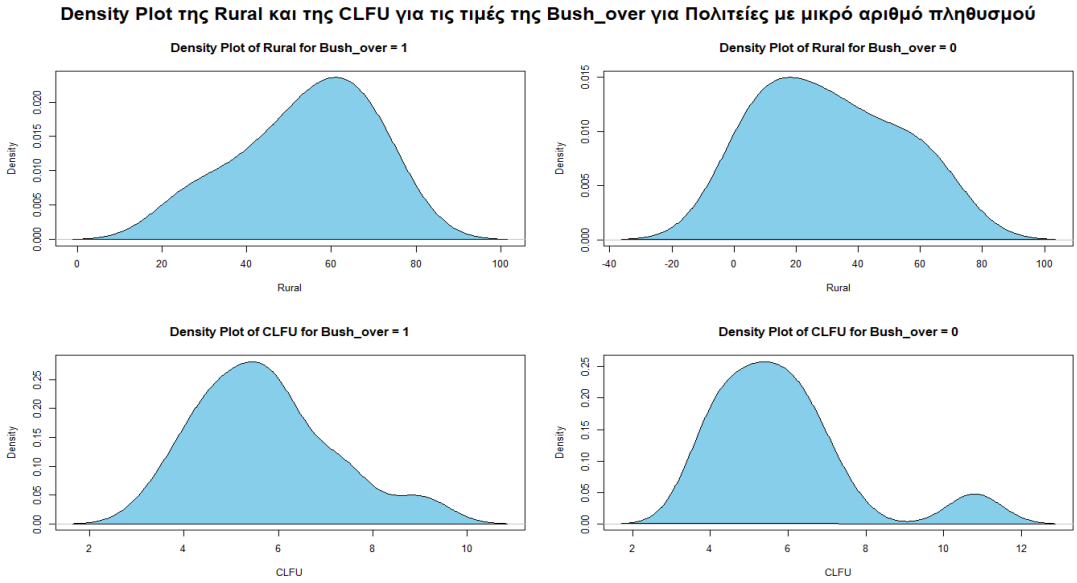
Σχήμα 34: Boxplots ανά τιμή της Bush_over των μεταβλητών Rural και CLFU για Πολιτείες με <u>μεγάλο αριθμό πληθυσμό</u>



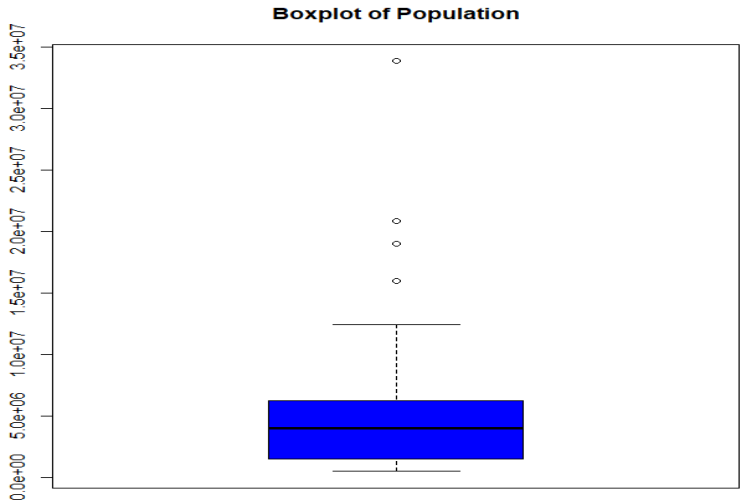
Σχήμα 35: Boxplots ανά τιμή της Bush_over των μεταβλητών Rural και CLFU για Πολιτείες με μικρό αριθμό πληθυσμό



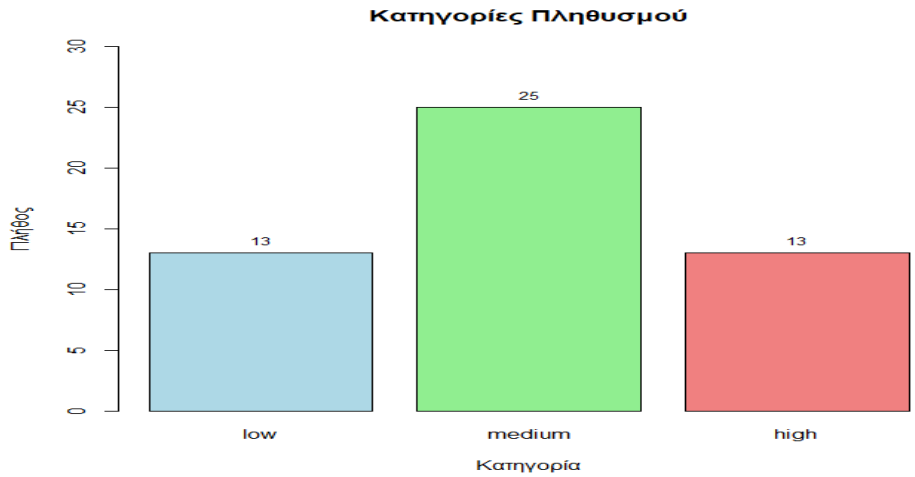
Σχήμα 36: Density Plot της Rural και της CLFU για τις τιμές της Bush_over για Πολιτείες με μεγάλο αριθμό πληθυσμού



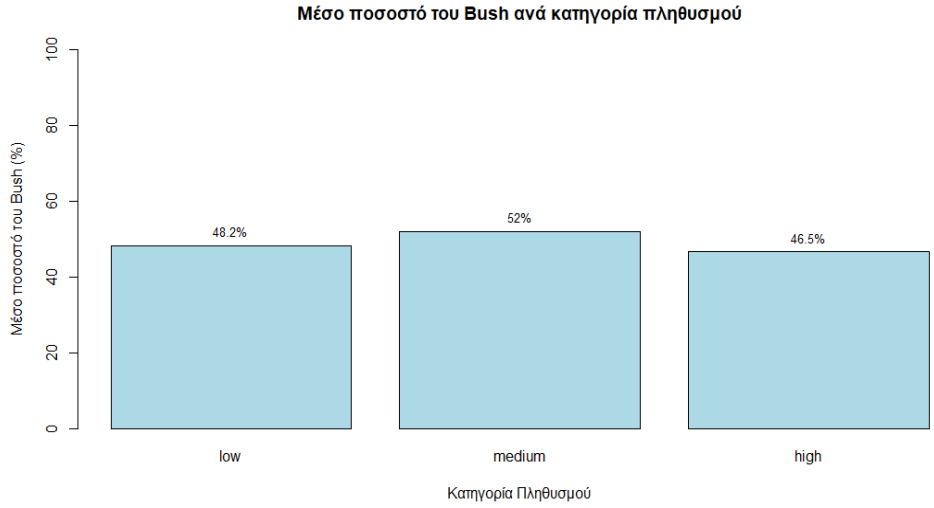
Σχήμα 37: Density Plot της Rural και της CLFU για τις τιμές της Bush_over για Πολιτείες με μικρό αριθμό πληθυσμού



Σχήμα 38: Boxplot της μεταβλητής Population με σκοπό τη μετατροπή της σε κατηγορική



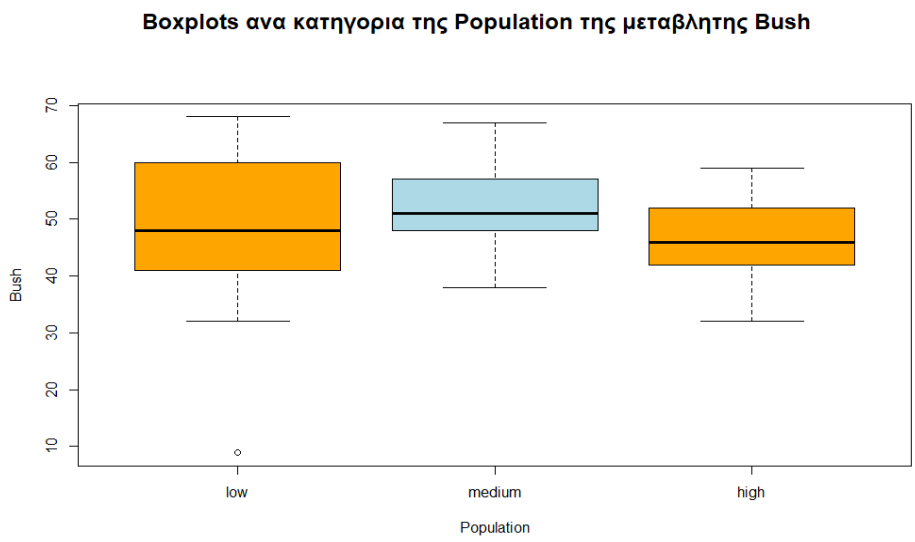
Σχήμα 39: Barplot της μεταβλητής Population μετά τη μετατροπή σε κατηγορική μεταβλητή



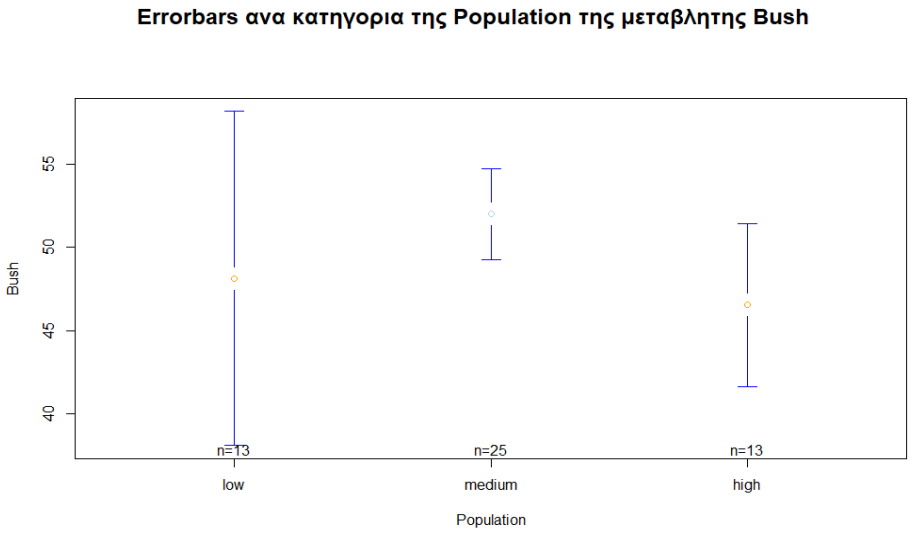
Σχήμα 40: Barplot για το Μέσο ποσοστό του Bush ανά κατηγορία Population

Μεταβλητές	Έλεγχος Κανονικότητας(Lillie&Shapiro)	Μέγεθος δείγματος	Έλεγχος για ισότητα διαμέσων και μέσων
Bush~Population	**Shapiro.test:p-value:0.014<0.05** **Lillie.test:p-value:0.20>0.05** **Αφού απορρίψαμε την μηδενική υπόθεση H0 δηλαδή για ύπαρξη κανονικότητας καταλοίπων σε ένα από τα δύο test θα απορρίψουμε την υπόθεση**	**n>50**	**Kruskal.test:** **p-value: 0.21>0.05** **Άρα συμπεραίνουμε ότι δεν απορρίπτουμε την μηδενική υπόθεση ότι δεν υπάρχουν σημαντικές διαφορές στις διαμέσους της μεταβλητής Bush μεταξύ των κατηγοριών της μεταβλητής Population (Ισότητα διαμέσων)** **Oneway.test(με άνισες διακυμάνσεις):** **p-value = 0.13>0.05** **Άρα δεν απορρίπτουμε μηδενική υπόθεση H0, ότι δεν υπάρχουν σημαντικές διαφορές στους μέσους της μεταβλητής Bush μεταξύ των κατηγοριών της μεταβλητής Population**

Πίνακας 16: Έλεγχος Anova για ύπαρξη σημαντικών διαφορών στην μεταβλητή Bush μεταξύ των κατηγοριών της Population.

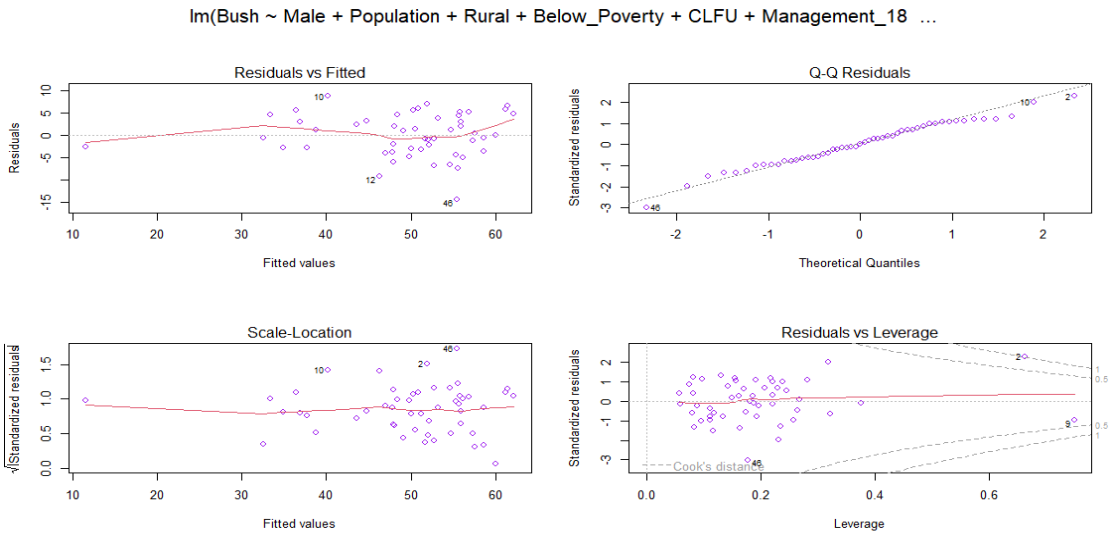


Σχήμα 41: Boxplots ανά κατηγορία της Population της μεταβλητής Bush

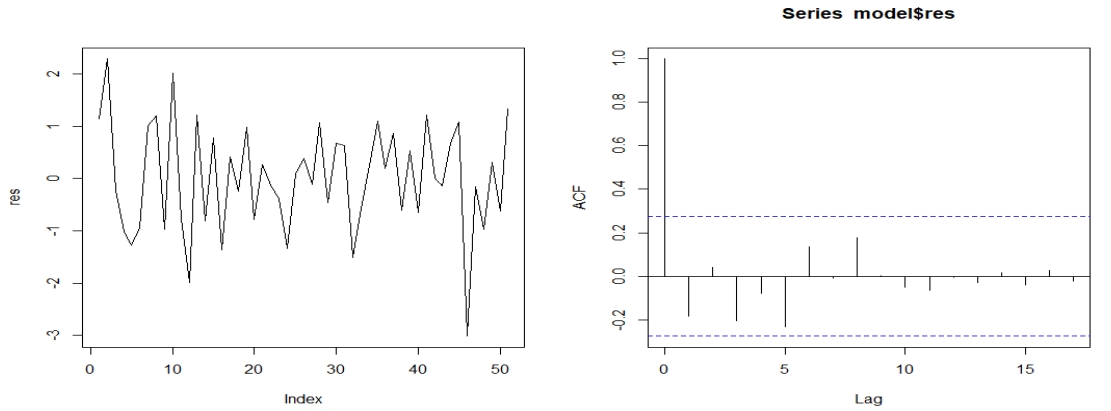


Σχήμα 42: Errorbars ανά κατηγορία της Population της μεταβλητής Bush

Γραμμικά Μοντέλα Παλινδρόμησης:



Σχήμα 43: Το plot για τους ελέγχους υποθέσεων του μοντέλου με όλες τις μεταβλητές

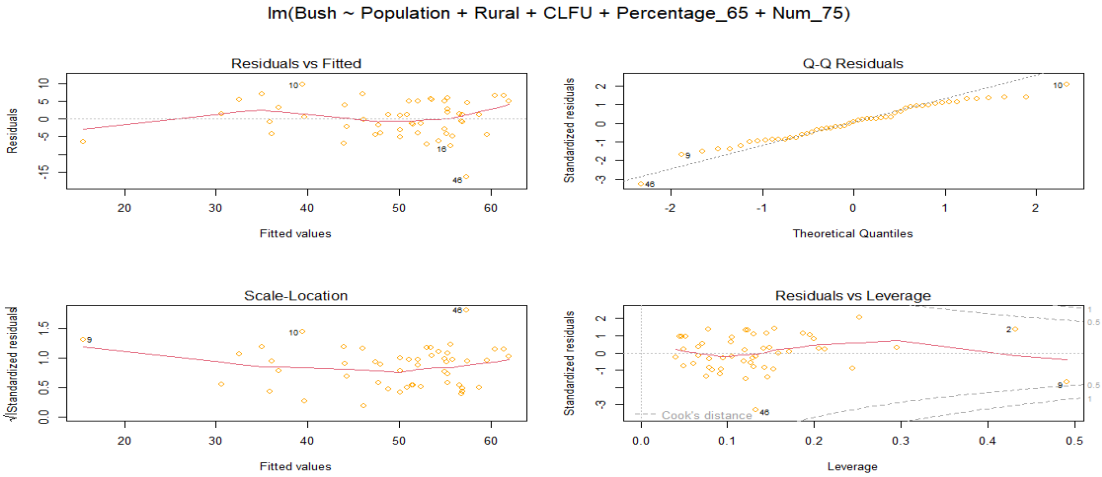


Σχήμα 44: Διαγράμματα ανεξαρτησίας των καταλοίπων και αυτοσυσχέτισης για το πλήρες μοντέλο

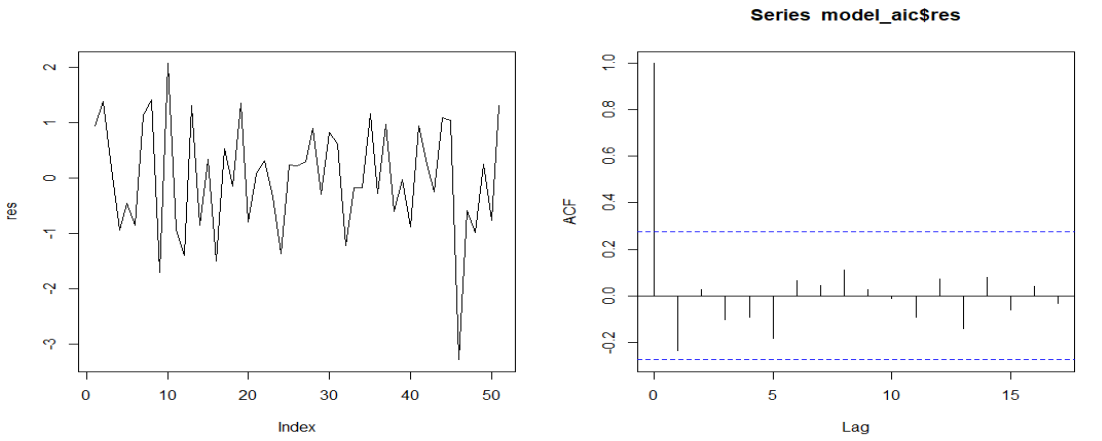
Μοντέλο	Έλεγχος Κανονικότητας καταλοίπων	Έλεγχος Ομοσκεδαστικότητας καταλοίπων	Έλεγχος Ανεξαρτησίας	Έλεγχος Πολυσυγγραμμικότητας
Πλήρες Μοντέλο	Shapiro.test= p-value:0.67>0.05 lillie.test= p-value: 0.84>0.05 Από τους ελέγχους δεν απορρίπτουμε την κανονικότητα	LeveneTest= p-value:0.35>0.05 Από τους ελέγχους δεν απορρίπτουμε την ομοσκεδαστικότητα	D-W statistic p- value: 0.28>0.05 Από τους ελέγχους δεν απορρίπτουμε την ανεξαρτησία ούτε και στα πρώτα 7 lag.	Vif(model): Παρατηρούμε υψηλή πολυσυγγραμμικότητα στις μεταβλητές Male και Management_18,δηλαδή μεγαλύτερη του 10

Πίνακας 17: Έλεγχος υποθέσεων του πλήρες μοντέλου

Μοντέλο Παλινδρόμησης μετά τη μέθοδο επιλογής μεταβλητών AIC stepwise-regression:



Σχήμα 45: Το plot για τους ελέγχους υποθέσεων μετά την μέθοδο επιλογής μεταβλητών stepwise regression

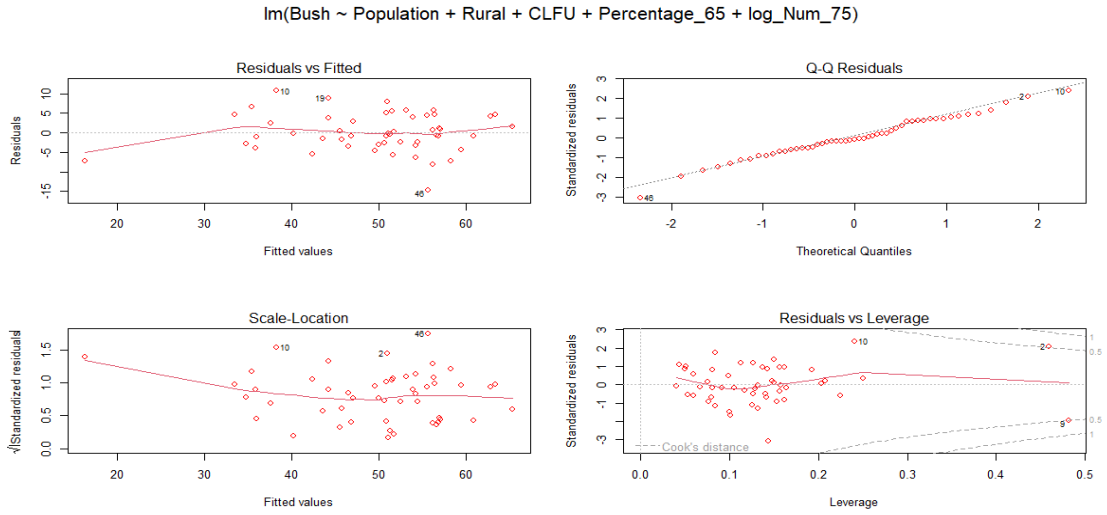


Σχήμα 46: Διάγραμμα ανεξαρτησίας των καταλοίπων και αυτοσυσχέτισης για το μοντέλο που προέκυψε από την stepwise regression.

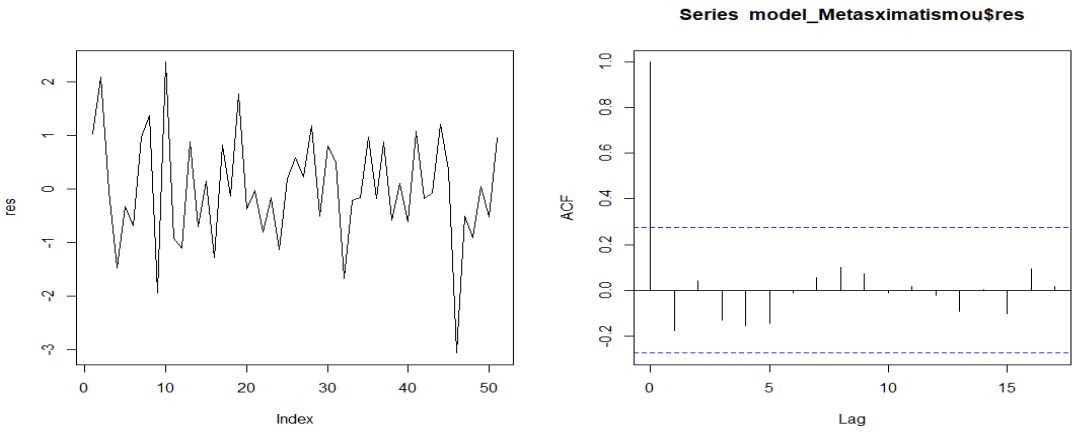
Μοντέλο	Έλεγχος Κανονικότητας καταλοίπων	Έλεγχος Ομοσκεδαστικότητας καταλοίπων	Έλεγχος Ανεξαρτησίας	Έλεγχος Πολυσυγγραμμικότητας
Πλήρες Μοντέλο	Shapiro.test=0.20>0.05 lillie.test= 0.54>0.05 Από τους ελέγχους δεν απορρίπτουμε την κανονικότητα	LeveneTest=0.18>0.5 Από τους ελέγχους δεν απορρίπτουμε την ομοσκεδαστικότητα	D-W statistic p- value: 0.14>0.05 Από τους ελέγχους δεν απορρίπτουμε την ανεξαρτησία ούτε και στα πρώτα 7 lag.	Vif(model): Παρατηρούμε ότι δεν έχουμε πολυσυγγραμμικότητα σε καμία μεταβλητή,δηλαδή τιμές μικρότερες του 10

Πίνακας 18: Έλεγχος υποθέσεων του μοντέλου που προέκυψε από την stepwise regression

Μοντέλο Παλινδρόμησης μετά τον μετασχηματισμό:



Σχήμα 47: Το plot για τους ελέγχους υποθέσεων μετά τον μετασχηματισμό λογαρίθμου της Num_75(Ποσοστό πληθυσμού με εισόδημα μεγαλύτερο από 75K)



Σχήμα 48: Διάγραμματα ανεξαρτησίας των καταλοίπων και αυτοσυσχέτισης μετά τον μετασχηματισμό λογαρίθμου της Num_75(Ποσοστό πληθυσμού με εισόδημα μεγαλύτερο από 75K)

Μοντέλο	Έλεγχος Κανονικότητας καταλοίπων	Έλεγχος Ομοσκεδαστικότητας καταλοίπων	Έλεγχος Ανεξαρτησίας	Έλεγχος Πολυσυγγραμμικότητας
Πλήρες Μοντέλο	Shapiro.test=0.79>0.05 lillie.test= 0.67>0.05 Από τους ελέγχους δεν απορρίπτουμε την κανονικότητα	LeveneTest=0.15>0.05 Από τους ελέγχους δεν απορρίπτουμε την ομοσκεδαστικότητα	D-W statistic p- value: 0.27>0.05 Από τους ελέγχους δεν απορρίπτουμε την ανεξαρτησία ούτε και στα πρώτα 7 lag.	Vif(model): Παρατηρούμε ότι δεν έχουμε πολυσυγγραμμικότητα σε καμία μεταβλητή,δηλαδή τιμές μικρότερες του 10

Πίνακας 19: Έλεγχος υποθέσεων του μοντέλου που προέκυψε μετά τον μετασχηματισμό