Οικονομικό Πανεπιστήμιο Αθηνών

Ανάλυση Δεδομένων 2023 - Εργασία 4

Προεδρικες Εκλογές ΗΠΑ 2000: Μία Στατιστική Μελέτη και Ανάλυση

Ονοματεπώνυμο:
Ιωάννης Γκιώνης
(p3190044)

Διδάσκοντες:
Ι. Ντζούφρας, Ξ. Πεντελή

4 Ιουνίου 2023

Περιεχόμενα

1 Εισαγωγή – περιγραφή μελέτης και προβλήματος

Με την ραγδαία ανάπτυξη της επιστήμης της στατιστικής τις τελευταίες δεκαετίες, ένας τομέας της που έχει τραβήξει αρκετά το ενδιαφέρον του κόσμου είναι αυτός των δημοσκοπήσεων και των πολιτικών προβλέψεων. Με την ανάπτυξη καινούριων μεθόδων ανάλυσης δεδομένων, στατιστικών μοντέλων αλλά και αλγορίθμων μηχανικής μάθησης (Machine Learning) και βαθιάς μηχανικής μάθησης (Deep Learning), ο κλάδος αυτός έχει αποκτήσει πολλά καινούρια "εργαλεία" τα οποία βοηθούν τους data scientists να κάνουν την δουλειά τους πιο γρήγορα αλλά και πιο αποτελεσματικά. Η πρόβλεψη αποτελεσμάτων εκλογών είναι πολύ σημαντική, καθώς η κοινή γνώμη επηρεάζεται αρκετά από τις δημοσκοπήσεις και η τελική απόφαση ψήφου αρκετών ανθρώπων μπορεί να εξαρτάται σε κάποιο βαθμό από το τί αποτέλεσμα περιμένουν, άρα η σωστή συλλογή, επεξεργασία και απεικόνιση των δεδομένων είναι πολύ σημαντική.

Η συγκεκριμένη μελέτη έχει ως αντικείμενο τις εκλογές των ΗΠΑ το 2000 και έχει ως σκοπό την έυρεση συσχέτισης ανάμεσα στα έτη διαμονής σε συγκεκριμένη πολιτεία και τις πολιτικές απόψεις/πρόθεση ψήφου των ερωτηθέντων. Επιπλέον, θα μελετηθεί το ενδεχόμενο κατασκευής γραμμικού μοντέλου πρόβλεψης του εισοδήματος ενός ερωτώμενου με δεδομένα ένα υποσύνολο των υπολοίπων μεταβλητών. Το σετ δεδομένων αποτελείται απο 1000 παρατηρήσεις 8 μεταβλητών και περιγράφεται στον πίνακα που βρίσκεται στην επόμενη σελίδα (σημείωση: η μεταβλητή ID παραλείπεται απο την υπόλοιπη μελέτη καθώς δεν μας προσφέρει καμία πληροφορία):

Table 1: Πίνακας Δεδομένων - Μεταβλητών

Όνομα	Τύπος	Σημασία	Τιμές - Εύρος Τιμών
ID	αριθμητική	Κωδικός ερωτώμενου	1 - 1000
Age	αριθμητική	Ηλικία	18 - 90
Years of residence	αριθμητική	Πόσο καιρό μένει κάποιος στην πολιτεία	0 - 72
Income	αριθμητική	Ετήσιο Εισόδημα (σε δολάρια)	2000 - 125000
Gender	κατηγορική	Φύλο	Αρσενικό, Θηλυκό
Vote	κατηγορική	Ψήφος	Al Gore George W. Bush Pat Buchanan Ralph Nader Other / Did Not Answer
Political orientation	κατηγορική	Πολιτικές Πεποιθήσεις	Extremely Liberal Liberal, Slightly Liberal Moderate Slightly Conservative Conservative Extremely Conservative Other / Did Not Answer
Marital Status	κατηγορική	Οικογενειακή Κατάσταση	Married, Widowed Divorced, Separated Never Married Partnered Other / Did Not Answer

2 Περιγραφική Ανάλυση

Για την περιγραφική ανάλυση των μεταβλητών χρησιμοποιούμε το στατιστικό πακέτο R, το οποίο μας παρέχει εργαλεία για ανάλυση (Data Analysis) και απεικόνιση (Data Presentation) του σετ δεδομένων μας. Αρχικά, εισάγουμε τα δεδομένα μας μέσω του πακέτου Haven. Ύστερα, αφού κοιτάξουμε τη δομή του DataFrame, και εφόσον δεν θα χρειαστεί να αναλύσουμε κάτι σε κάποια συγκεκριμένη εγγραφή, διαγράφουμε την στήλη (Μεταβλητή) ID.

Κοιτώντας τον πίνακα 1 εντοπίζουμε τις κατηγορικές μεταβλητές: Gender, Vote, Political Orientation, Marital Status και ορίζουμε τα κατάλληλα περιγραφικά μέτρα γι αυτές τα οποία επίσης απεικονίζονται στον παραπάνω πίνακα.[1] Η κατηγορική μεταβλητή Gender είναι δίτιμη (0 = αρσενικό, 1 = θηλυκό), ενώ οι άλλες 3 έχουν πολλές τιμές, με την Political Orientation να είναι η μόνη η οποία μπορεί να θεωρηθεί διατάξιμη, με τις τιμές της(1-7) να βρίσκονται σε σειρά στο πολιτικό φάσμα από τον φιλελευθερισμό μέχρι τον συντηρητισμό.

Συνεχίζοντας την ανάλυση αυτή τη φορα για ποσοτικές μεταβλητές, ελέγχουμε το εύρος τιμών κάθε μεταβλητής ξεχωριστά, κάνοντας ελέγχους Shapiro-Wilk [6] και Lillie [4] για κανονικότητα. Παρακάτω ακολουθούν τα διαγράμματα πυκνότητας πιθανότητας για τις 3 ποσοτικές μεταβλητές (Age, Years of residence, Income) στα οποία περιέχονται τιμές των Shapiro tests καθώς και ένας πίνακας με τα περιγραφικά τους μέτρα.

Table 2: Πίνακας Περιγραφικών Μέτρων Ποσοτικών Μεταβλητών

Μεταβλητή	Ελλιπείς Τιμές	Μέσος	Τυπική Απόκλιση	Ελάχιστη Τιμή	Μέγιστη Τιμή	Ασσυμετρία	Κύρτωση
Age	0	47.572	17.213	18	90	0.2	-0.84
Years of residence	1	11.773	12.722	0	72	1.41	1.6
Income	126	55604	37649	2000	125000	0.54	-0.87

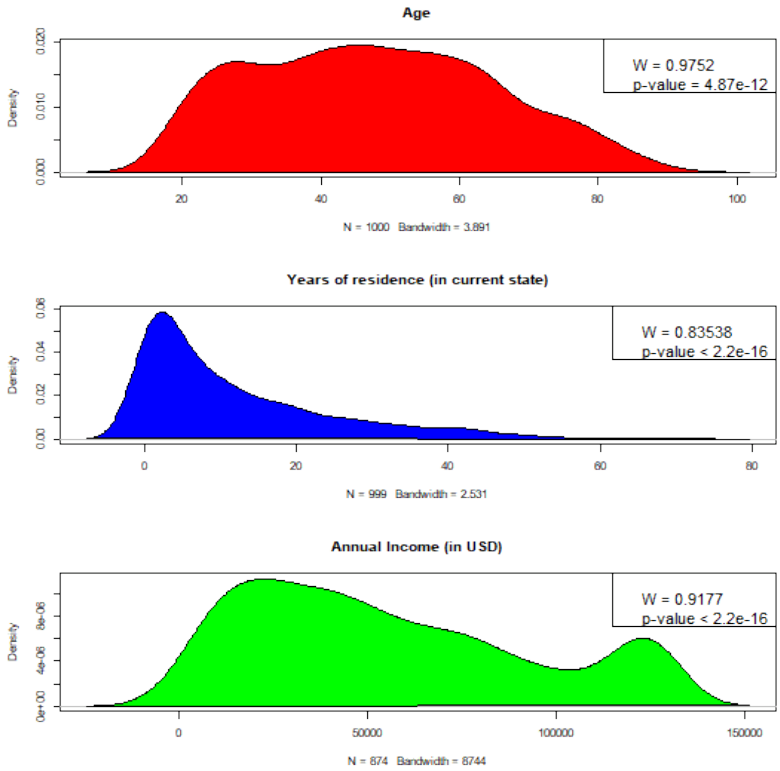
3 Σχέσεις Μεταβλητών ανα **2**

3.1 Γενικά

Αρχικά, μελετούμε τις σχέσεις ανάμεσα στις τρεις ποσοτικές μεταβλητές. Η μόνη που παρουσιάζει κάποια συσχέτιση είναι η $Age \sim Yearsofresidence$, καθώς στον έλεγχο Pearson παρουσιάζεται $pvalue \approx 0$ και $r = 0.57$, το οποίο δείχνει μια θετική συσχέτιση, δηλαδή κατά κανόνα, όσο μεγαλύτερο γίνεται το Age, μεγαλώνει και το Years of residence. Στις άλλες 2 σχέσεις ($Age \sim Income$ και $Yearsofresidence \sim Income$) δεν παρατηρείται κάποια σημαντική συσχέτιση ανάμεσα στις μεταβλητές, ενώ τα pvalues των ελέγχων Pearson είναι 0.5 και 0.35 αντίστοιχα, άρα δεν μπορούμε να καταλήξουμε σε κάποιο συμπέρασμα.

[1] Σε κάποιες από τις παραπάνω μεταβλητές υπάρχουν μη-χρησιμοποιούμενες δεσμευμένες τιμές οι οποίες παραλείπονται από αυτή την ανάλυση καθώς δεν παρατηρούνται παραδείγματα που τις χρησιμοποιούν στο segment των δεδομένων που μας έχει δοθεί

Figure 1: Διαγράμματα Πυκνότητας Πιθανότητας



Ύστερα, κοιτώντας πίνακα συσχετίσεων του Pearson 6 βλέπουμε ότι υπάρχουν αρκετά μικρά p-values, άρα αρκετή συσχέτιση μεταβλητών στο σετ δεδομένων μας. Αγνοώντας τα ζεύγη μεταβλητών με υψηλά p-values και έχοντας υπ' όψην ότι οι μεταβλητές vote-political orientation και age-years of residency[2] σχετίζονται λογικά προσπαθούμε να βρούμε σχέσεις που θα βγάζουν νόημα να μελετηθούν. Οι σχέσεις:

- Χρόνια διαμονής σε πολιτεία και ψήφος ($Yearsofresidence \sim Vote$)
- Χρόνια διαμονής σε πολιτεία και πολιτικές πεποιθήσεις ($Yearsofresidence \sim PoliticalOrientation$)

θα μελετηθούν περαιτέρω στην ενότητα 3.2. Τελικά, οι σχέσεις ανάμεσα σε ζεύγη μεταβλητών τις οποίες θα διερευνήσουμε είναι οι εξής:

1. Ετήσιο εισόδημα και φύλο ($Income \sim Gender$)

[2]οι υποθέσεις επιβεβαιώνονται απο ελέγχους X^2 και Pearson αντίστοιχα

2. Ηλικία και φύλο ($Age \sim Gender$)
3. Ηλικία και οικογενειακή κατάσταση ($Age \sim MaritalStatus$)
4. Ηλικία και ψήφος ($Age \sim Vote$)
5. Ετήσιο Εισόδημα και οικογενειακή κατάσταση ($Income \sim MaritalStatus$)
6. Πολιτικές πεποιθήσεις και οικογενειακή κατάσταση ($PoliticalOrientation \sim MaritalStatus$)
7. Ετήσιο εισόδημα και ψήφος ($Income \sim Vote$)
8. Φύλο και ψήφος ($Gender \sim Vote$)
9. Ψήφος και πολιτικές πεποιθήσεις ($Vote \sim PoliticalOrientation$)
10. Οικογενειακή κατάσταση και ψήφος ($MaritalStatus \sim Vote$)
11. Ετήσιο εισόδημα και πολιτικές πεποιθήσεις ($Income \sim PolticalOrientation$)

Τα διαγράμματα (barplots, boxplots) καθώς και τα αποτελέσματα των στατιστικών ελέγχων (Kruskal-Wallis test [2] για ζεύγη ποσοτικών και κατηγορικών μεταβλητών και Chisquared [5] + Fisher tests [1] για ζεύγη κατηγορικών) βρίσκονται στο παράρτημα στην ενότητα 6. Από τις παραπάνω σχέσεις, αυτές που αποδείχθηκαν να έχουν κάποια στατιστικά σημαντική εξάρτηση μεταξύ τους είναι οι: 1,3,5,6,7,8,9,10,11 [3]

3.2 Συγκεκριμένες Σχέσεις Μεταβλητών

Σε αυτή την ενότητα θα αναλύσουμε τις δύο παρακάτω σχέσεις μεταξύ μεταβλητών:

1. Χρόνια διαμονής σε πολιτεία και ψήφος ($Yearsofresidence \sim Vote$)
2. Χρόνια διαμονής σε πολιτεία και πολιτικές πεποιθήσεις ($Yearsofresidence \sim PoliticalOrientation$)

Όπως παρατηρούμε και στον πίνακα 1, και οι 2 σχέσεις πρόκειται για σχέσεις ποσοτικής και κατηγορικής μεταβλητής. Οι ελέγχοι που μπορούμε να εφαρμόσουμε είναι οι Kruskal test και ANOVA. Αρχικά κατασκευάζουμε τα αρχικά μοντέλα ANOVA

[3] $pvalue \leq 0.05$ σε ελέγχους kruskal ή chisq+fisher

και ύστερα τρέχουμε ελέγχους κανονικότητας και ομοσκεδαστικότητας[4]. Παρατηρούμε με ελέγχους Shapiro/Lillie ότι τα κατάλοιπα και των δύο μοντέλων δεν ακολουθούν την κανονική κατανομή, όπως άλλωστε φαίνεται στα διαγραμματα 22 και 23. Επίσης από τα διαγράμματα παρατηρούμε και συμπαιρένουμε ότι ο διάμεσος είναι πιο κατάλληλο μέτρο περιγραφής έναντι του μέσου. Δίνουμε λοιπόν έμφαση στους ελέγχους Kruskal, τα αποτελέσματα των οποίων βρίσκονται παρακάτω.

1. $H = 1.5788$, $df = 4$, $pvalue = 0.8126$, η H_0 δεν μπορεί να απορριφθεί.

2. $H = 18.082$, $df = 6$, $pvalue = 0.006$, η H_0 απορρίπτεται.

Όπως φαίνεται από τα αποτελέσματα των ελέγχων Kruskal-Wallis, υπάρχει μια σημαντική εξάρτηση μεταξύ των δύο μεταβλητών της δεύτερης σχέσης, σε αντίθεση με αυτές της πρώτης. Δηλαδή, φαίνεται οι πολιτικές πεποιθήσεις ενός ανθρώπου να επηρεάζονται στατιστικά από το πόσο καιρό είναι κάτοικος της πολιτείας στην οποία μένει, ενώ το ίδιο δεν ισχύει για το την ψήφο του στις εκλογές του 2000.

4 Προβλεπτικά ή ερμηνευτικά μοντέλα

Για την κατασκευή ενός γραμμικού μοντέλου πρόβλεψης της πραγματικής τιμής της συνεχούς μεταβλητής Income χρειάζεται αρχικά να ελέγξουμε για τυχόν ακραίες τιμές (outliers). Μέσω του διαγράμματος 7 αλλά και μέσω ελέγχων διαπιστώνουμε ότι δεν υπάρχουν τέτοιες τιμές στο σετ δεδομένων μας, οπότε είμαστε έτοιμοι να δούμε ποιές συσχετίσεις της μεταβλητής Income είναι αρκετά σημαντικές για να χρησιμοποιηθούν στο μοντέλο μας. Από την ανάλυση σχέσεων ανά 2 της προηγούμενης ενότητας βλέπουμε ότι ο συντελεστής συσχέτισης Pearson έχει χαμηλές απόλυτες τιμες (cor1= 0.022, cor2=-0.031) για τις συσχετίσεις $Income \sim Age$ και $Income \sim Yearsofresidence$, ενώ τα μεγάλα p-values κάνουν αυτές τις μικρές συσχετίσεις ακόμα πιο ασήμαντες. Άρα, στα γραμμικά υποδείγματα που θα φτιαχθούν, δεν θα ληφθούν υπ' όψιν οι λοιπές συνεχείς αριθμητικές μεταβλητές. Στην επόμενη σελίδα βρίσκονται τα 4 boxplots για την μεταβλητή income και κάθε μια απο τις 4 κατηγορικές μεταβλητές. Γνωρίζουμε ήδη ότι η μεταβλητή Income είναι εξαρτημένη από όλες τις κατηγορικές μεταβλητές, δηλαδή τις: Gender, Vote, Marital Status και Political Orientation, κάτι το οποίο είναι εμφανές στα παρακάτω διαγράμματα.

[4]Τελικά δεν χρειάζεται έλεγχος ομοσκεδαστικότητας καθώς τα κατάλοιπα κανενός μοντέλου δεν ακολουθούν την κανονική κατανομη. Οι ελέγχοι γίνονται στον πηγαίο κώδικα και τα αποτελέσματά τους υπάρχουν σε comment

Figure 2: Boxplot Ετήσιου Εισοδήματος και Ψήφου

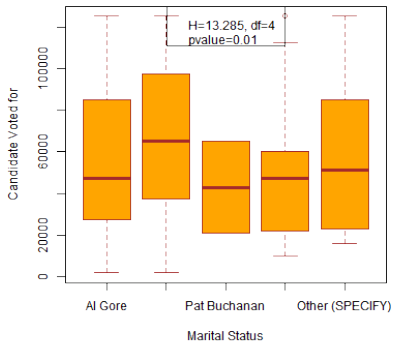


Figure 3: Boxplot Ετήσιου Εισοδήματος και Οικογενειακής Κατάστασης

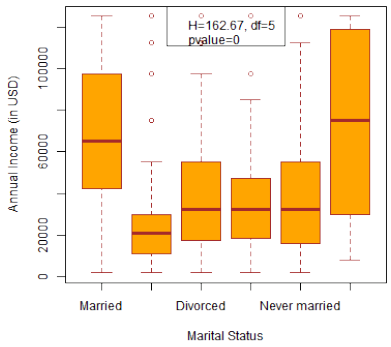


Figure 4: Boxplot Ετήσιου Εισοδήματος και Φύλου

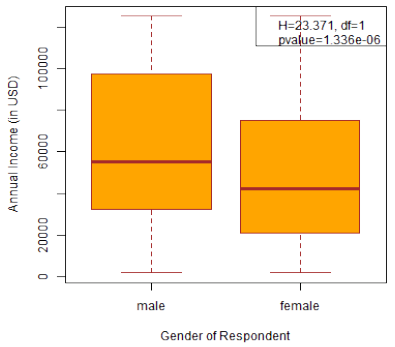
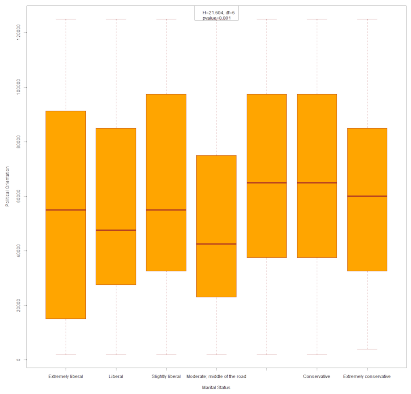


Figure 5: Boxplot Ετήσιου Εισοδήματος και Πολιτικού Προσανατολισμού



Επιπλέον, παρατηρούμε στον έλεγχο X^2 που έγινε στις μεταβλητές Poltical Orientation και Vote ότι ο συντελεστής συσχέτισης X^2 ισούται με 279.7 και το pvalue ισούται με 0, το οποίο δείχνει μια πάρα πολύ σημαντική συσχέτιση των δύο μεταβλητών, κάτι που άλλωστε βγάζει νόημα λογικά. Επειδή οι δύο αυτές μεταβλητές εκφράζουν κάτι πάρα πολύ παρόμοιο και οι τιμές της μίας επηρεάζονται πάρα πολύ από τις τιμές της άλλης, θα χρησιμοποιήσουμε μόνο μία από τις 2 στα μοντέλα μας, έστω την Vote.

Table 3: Πίνακας Πρώτου Μοντέλου Γραμμικής Παλλινδρόμισης

	Dependent variable:
	Income
Gender: female	$-7{,}916.522^{***}$
	(2,920.591)
Vote: George W. Bush	4,728.246
	(2,936.674)
Vote: Pat Buchanan	653.727
	(24,382.570)
Vote: Ralph Nader	$-6{,}980.956$
	(9,164.055)
Vote: Other (SPECIFY)	$-1{,}432.393$
	(14,132.430)
Marital_Status: Widowed	$-33{,}862.190^{***}$
	(5,609.388)
Marital_Status: Divorced	$-25{,}634.470^{***}$
	(4,222.485)
Marital_Status: Separated	$-25{,}586.400^{***}$
	(8,516.669)
Marital_Status: Never married	$-22{,}148.090^{***}$
	(4,006.045)
Marital_Status: Partnered, not married {VOL}	$26{,}435.220^{**}$
	(12,296.670)
Observations	593
R^2	0.170
Adjusted R^2	0.156
Residual Std. Error	34,147.320 (df = 582)
F Statistic	11.922^{***} (df = 10; 582)
Note:	*p<0.1; **p<0.05; ***p<0.01

Κατασκευάζουμε το γραμμικό μοντέλο παλλινδρόμησης με τις 3 μεταβλητές[5] ($Income \sim Gender + MaritalStatus + Vote$. Όπως και θα παρατηρήσουμε στον παραπάνω πίνακα, όλες οι τιμές της μεταβλητής Vote παίρνουν πολύ μεγάλα pvalues(< 0.1), άρα δοκιμάζουμε να κατασκευάσουμε ένα αντίστοιχο μοντέλο χωρίς την μεταβλητή Vote. Ο αντίστοιχος πίνακας γι΄ αυτό το μοντέλο καθώς και τα αντίστοιχα QQPlots των καταλοίπων των δύο μοντέλων βρίσκονται στο παράρτημα. Για την σύγκριση των δύο μοντέλων χρησιμοποιούμε έλεγχο ANOVA, τα αποτελέσματα του οποίου μας δείχνουν ότι τα δύο μοντέλα είναι πολύ παρόμοια, άρα θα κρατήσουμε ως βέλτιστο αυτό με τις λιγότερες μεταβλητές, άρα το δεύτερο. Ελέγχουμε τις προϋποθέσεις παλλινδρόμησης κάνοντας ελέγχους κανονικότητας (Shapiro/Lillie) στα κατάλοιπα αλλά και ελέγχους ομοσκεδαστηκότητας Levene [3], μέσω των οποίων καταλήγουμε στο συμπέρασμα ότι τα κατάλοιπα δεν ειναι κανονικά και υπάρχει ομοσκεδαστηκότητα αλλά όχι σε μεγάλο βαθμό. Για να επιβεβαιώσουμε ότι το μοντέλο που κατασκευάσαμε είναι το βέλτιστο, τρέχουμε μια step-wise function[6] η οποία μας επιστρέφει το βέλτιστο μοντέλο, το οποίο φαίνεται πως είναι ακριβώς το ίδιο με το δέυτερο μοντέλο που κατασκευάσαμε. Είναι χρήσιμο να σημειωθεί ότι ακόμα και το βέλτιστο προβλεπτικό μοντέλο δεν έχει μεγάλη ακρίβεια, καθώς το στατιστικό R^2_{adj} ισούται με 0.156, ενώ το R^2 ισούται με 0.165.

Συμπερασματικά, το βέλτιστο μοντέλο είναι αυτό που χρησιμοποιεί μόνο τις κατηγορικές μεταβλητές Gender και Marital Status. Εάν θεωρήσουμε εξίσωση γραμμικής παλινδρόμησης $Income_{pred} = b0 + b1GenderFemale + b2MaritalWidowed + b3MaritalDivorced + b4Separated + b5NeverMarried + b6Partnered$ με default values τα Gender = Male και Marital Status = Married, τότε οι συντελεστές ισούνται με: $b0 = 76984$, $b1 = -8128$, $b2 = -34629$, $b3 = -26518$, $b4 = 26468$, $b5 = -23403$, $b6 = 25128$, άρα πρακτικά αυτό που καταλαβαίνουμε από το μοντέλο είναι ότι ο μέσος παντρεμένος άντρας βγάζει 76984 δολλάρια το χρόνο, ποσό που μειώνεται εάν μιλάμε για γυναίκα, για χωρισμένο/-η κλπ με μόνη περίπτωση αύξησης να απαντάται στην περίπτωση που η οικογενειακή κατάσταση ισούται με Partnered/Not Married, περίπτωση που απαντάται πολύ λιγες φορές στο dataset μας, κάτι που επιβεβαιώνεται από το pvalue που ισούται με 0.04, τιμή η οποία βρίσκεται ακριβώς πάνω στα όρια του στατιστικά σημαντικού.

[5] η συνάρτηση lm() του προγραμματιστικού περιβάλλοντος R κατασκευάζει από μόνη της ψευδομεταβλητές για τις κατηγορικές μεταβλητές που του δίνουμε, όπως άλλοστε φαίνεται και στον πίνακα 3

[6] Χρησιμοποιούμε modified dataframe με τις 3 μεταβλητές που έχουν μείνει. Εάν χρησιμοποιήσουμε το αρχικό dataframe το αποτέλεσμα θα είναι παρόμοιο αλλά η περικοπή εγγραφών λόγω null values σε irrelevant στήλες οδηγεί σε χαμηλότερο R^2_{adj}, άρα δεν είναι η καλύτερη επιλογή

5 Συμπεράσματα

Η στατιστική μελέτη αυτή είχε ως σκοπό την πλήρη ανάλυση των μεταβλητών του σετ δεδομένων καθώς και των μεταξύ τους συσχετίσεων αλλά και την κατασκευή ενός γραμμικού μοντέλου με σκοπό την πρόβλεψη του ετησίου εισοδήματος ενός αμερικάνου πολίτη με βάση τα λοιπά στοιχεία που υπάρχουν στο σετ δεδομένων. Ενώ το μοντέλο έχει αρκετά χαμηλές τιμές R^2 και R^2_{adj} (0.165 και 0.156 αντίστοιχα), κάτι που δείχνει ότι η προσαρμογή του μοντέλου δεν είναι και η καλύτερη, καταφέρνει αρκετά καλά να περιγράψει τις συσχετίσεις που υπάρχουν στο σετ δεδομένων.

Από το αρχικό μοντέλο, φαίνεται πόσο δυνατές είναι οι συσχετίσεις με τις 2 κατηγορικές μεταβλητές που τελικά καταλήγουμε να χρησιμοποιούμε, καθώς τα pvalues τους στον πίνακα 3 είναι πολύ κοντά στο 0, σε αντίθεση με τις υπόλοιπες μεταβλητές. Επίσης, από το τελικό μοντέλο φαίνεται το πόσο σημαντικό ρόλο παίζει η οικογένεια στην δυτική της μορφή στο τελικό εισόδημα ενός ανθρώπου, καθώς παρατηρείται ότι οι γυναίκες, παρόλο που θεσμικά πληρώνονται ακριβώς το ίδιο με τους άντρες, καταλήγουν να έχουν αρκετά χαμηλότερο ετήσιο εισόδημα, γεγονός που αποδίδεται συνήθως στις αυξημένες ευθύνες και στον μειωμένο χρόνο λόγω της ιδιότητας των γυναικών ως μητέρες αλλά και την μεγαλύτερη συμβολή τους στα οικιακά. Επιπλέον, παρατηρείται διαφορά στο εισόδημα ανάμεσα σε ανθρώπους διαφορετικών οικογενειακών καταστάσεων, αλλά το μέγεθος του δείγματος, το πλήθος των διαφορετικών τιμών(6) αλλά και το γεγονός ότι το μεγαλύτερο κομμάτι του δείγματος παίρνει μια συγκεκριμένη τιμή (παντρεμένος/-η) δεν μας αφήνει να βγάλουμε κάποιο πραγματικό συμπέρασμα.

6 Παράρτημα

Figure 6: Πίνακας Συσχετίσεων του Pearson

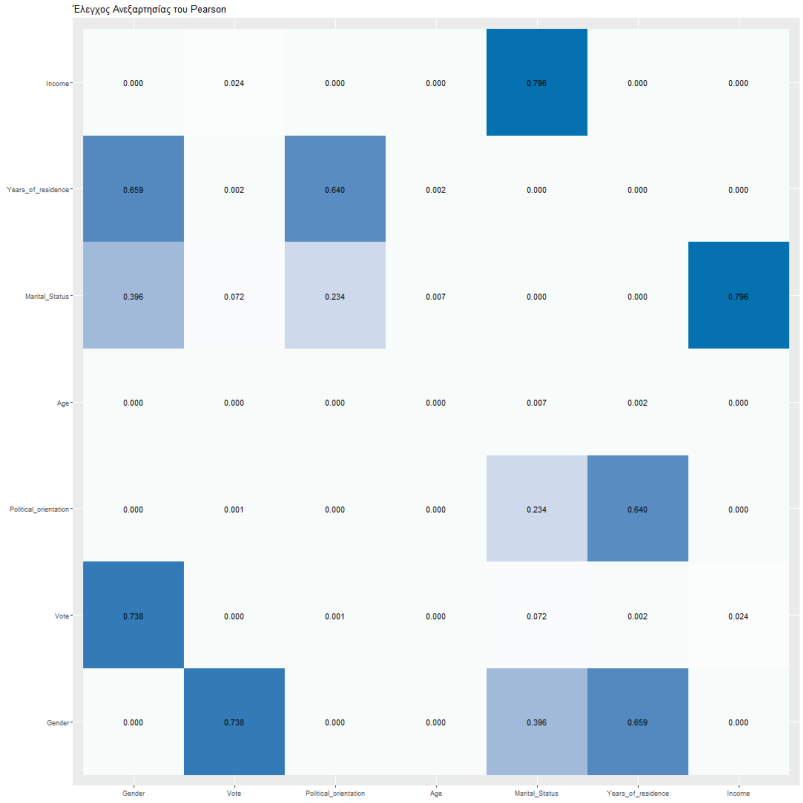


Figure 7: Διάγραμμα (Boxplot) Ετήσιου Εισοδήματος

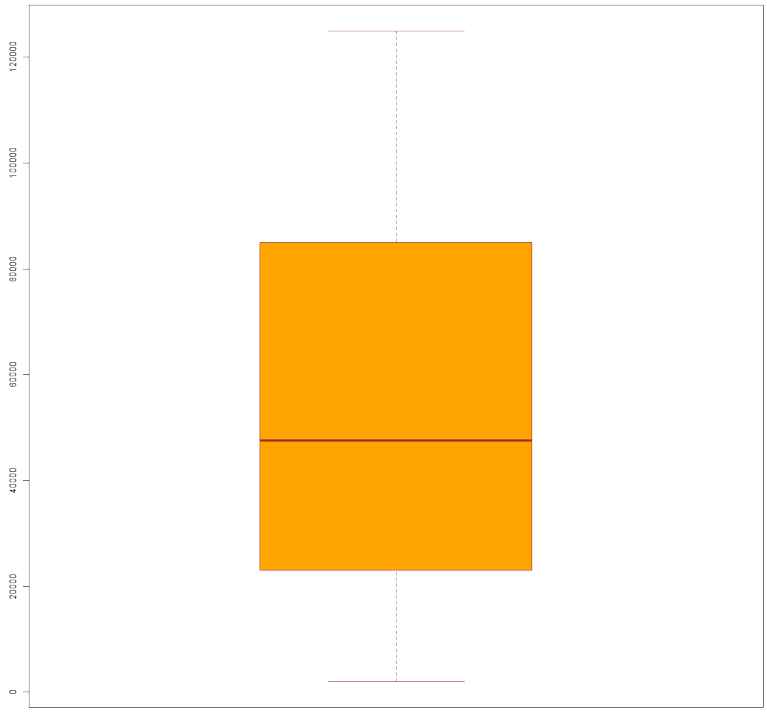


Figure 8: Διάγραμμα Διασποράς Ηλικίας και Χρόνων διαμονής σε πολιτεία

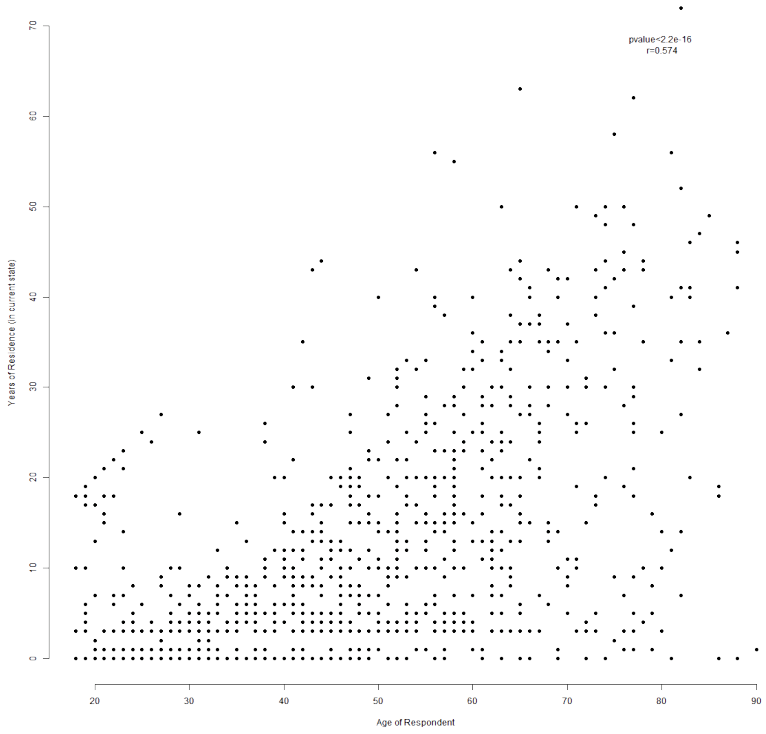


Figure 9: Διάγραμμα Διασποράς Ηλικίας και Ετήσιου Εισοδήματος

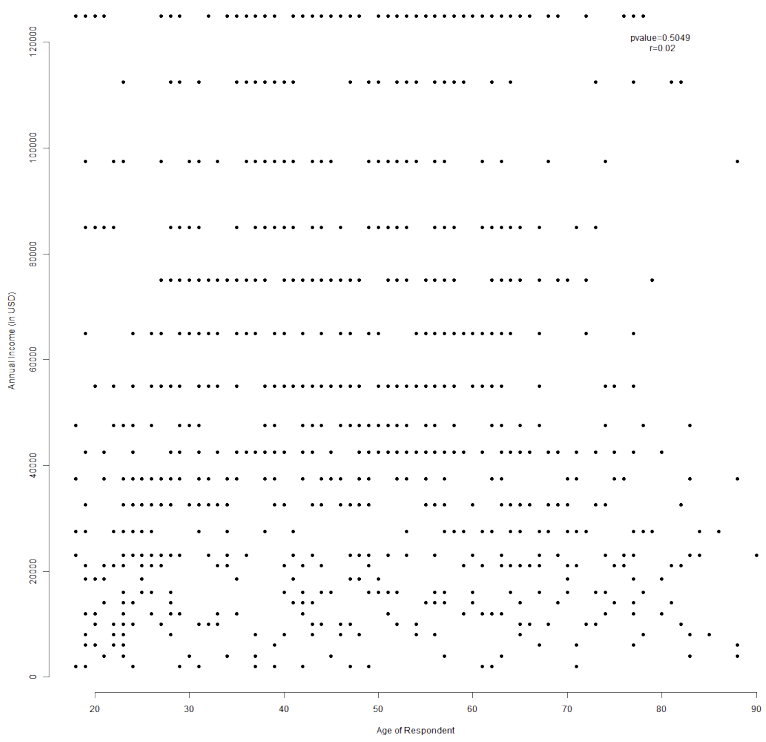


Figure 10: Διάγραμμα Διασποράς Χρόνων διαμονής σε πολιτεία και Ετήσιου Εισοδήματος

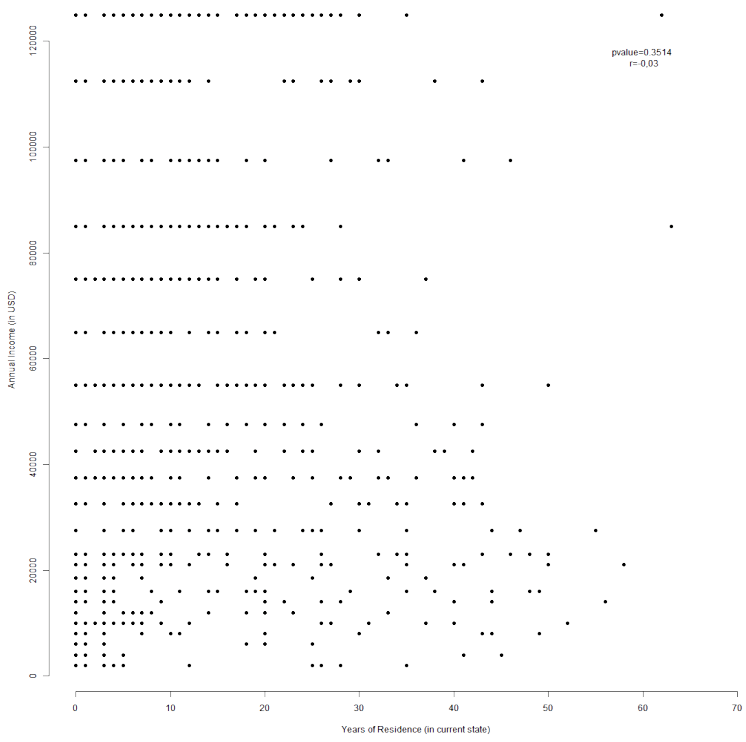


Figure 11: Boxplot Ηλικίας και Φύλου

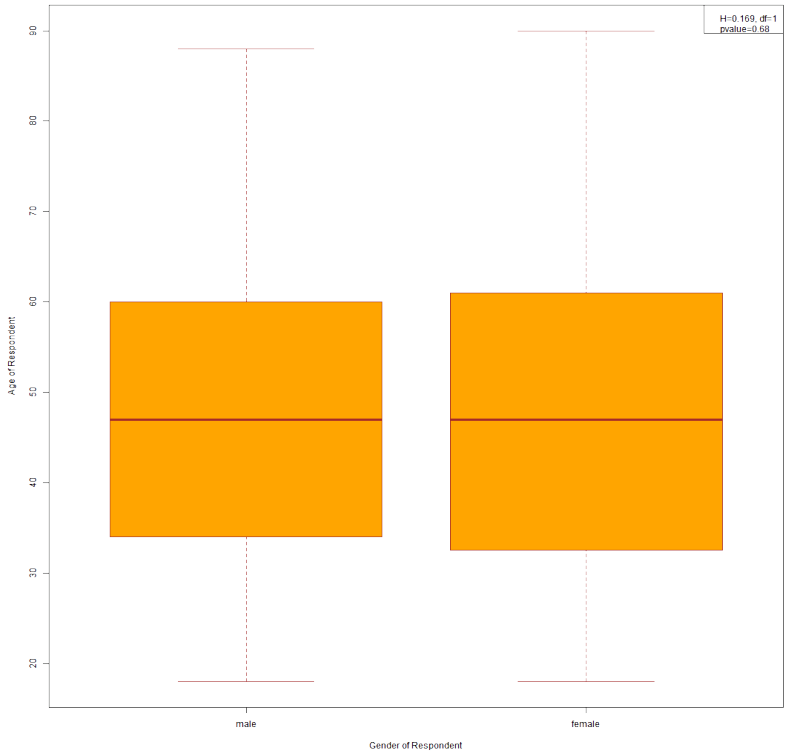


Figure 12: Boxplot Ηλικίας και Οικογενειακής Κατάστασης

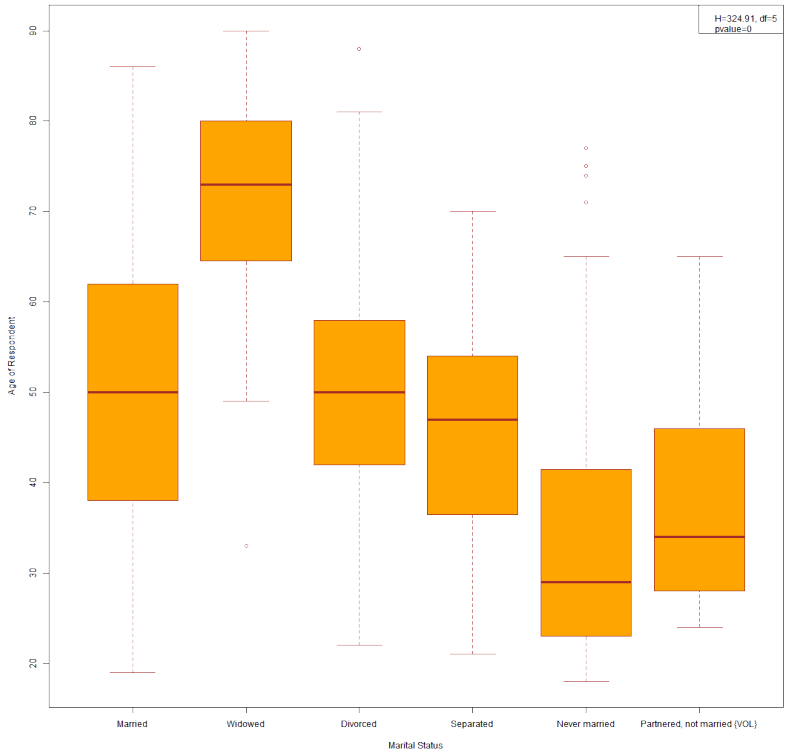


Figure 13: Boxplot Ηλικίας και Ψήφου

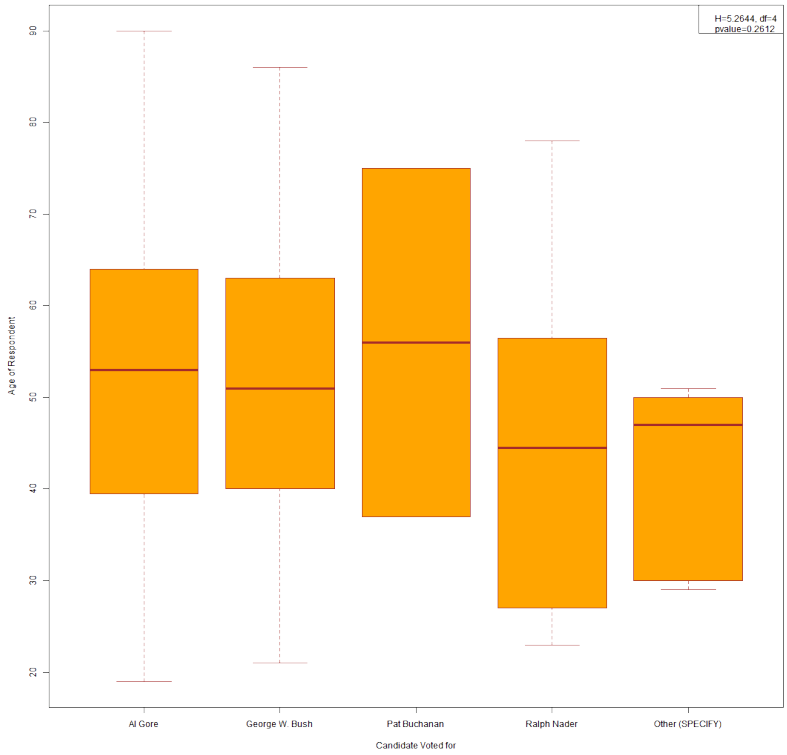


Figure 14: Barplot Φύλου και Ψήφου

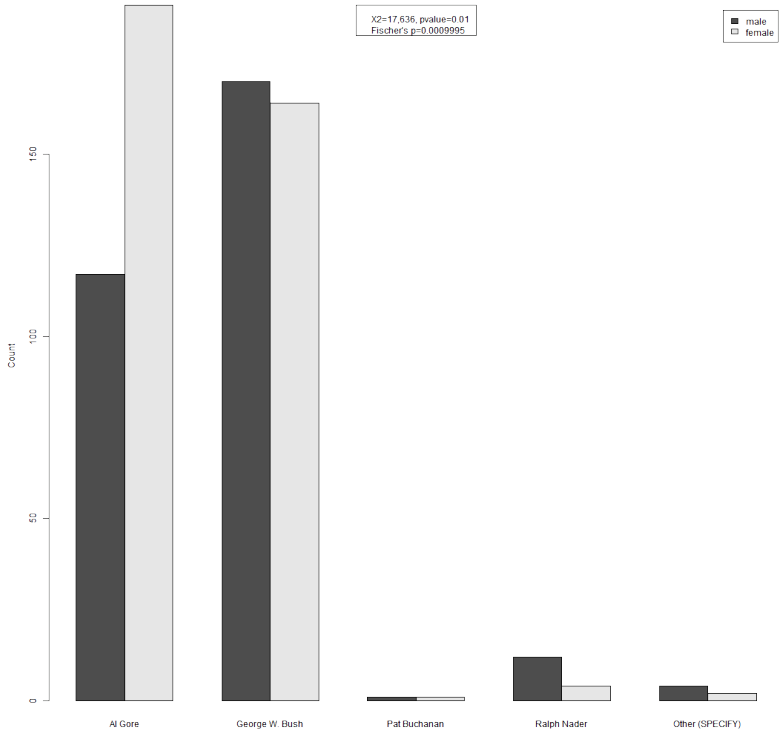


Figure 15: Barplot Πολιτικών Πεποιθήσεων και Οικογενειακής Κατάστασης

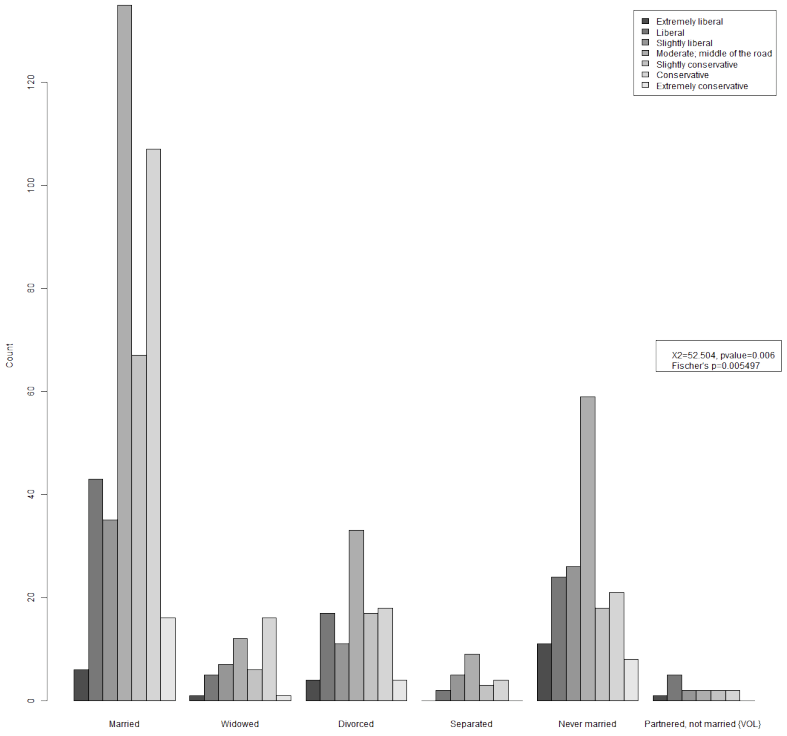


Figure 16: Barplot Ψήφου και Πολιτικών Πεποιθήσεων

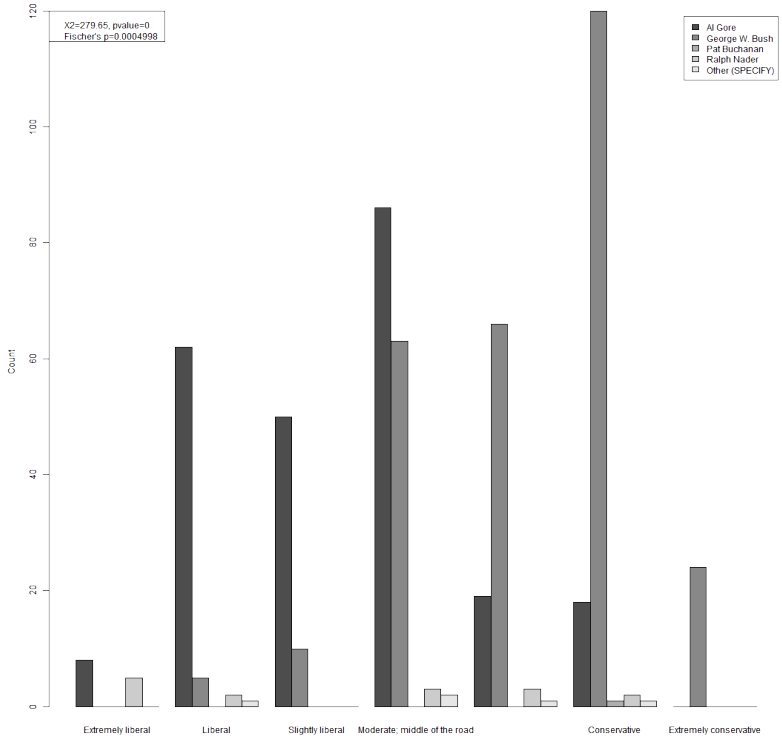


Figure 17: Barplot Οικογενειακής Κατάστασης και Ψήφου

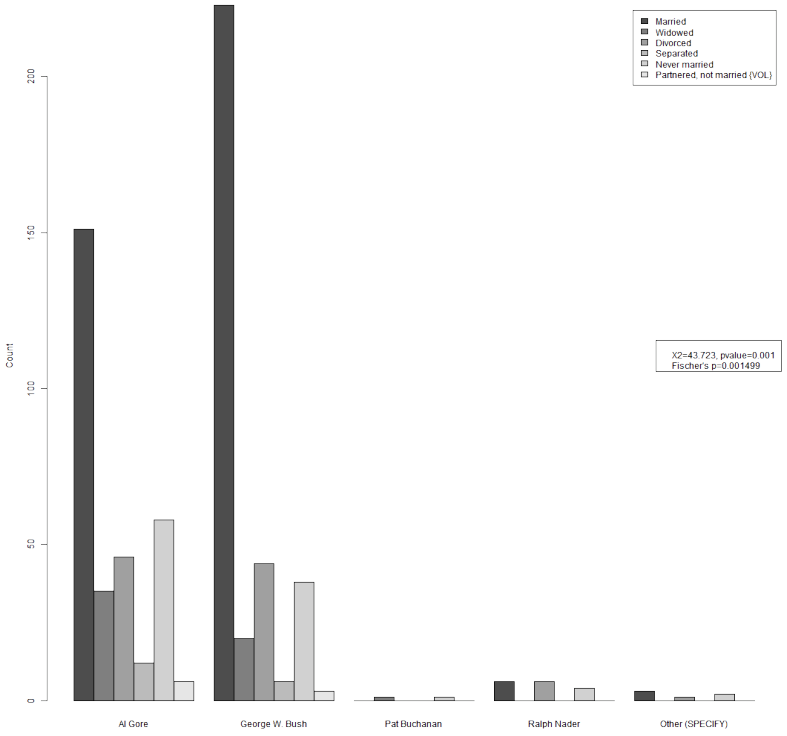


Figure 18: Barplot Ετών Διαμονής σε πολιτεία και Ψήφου

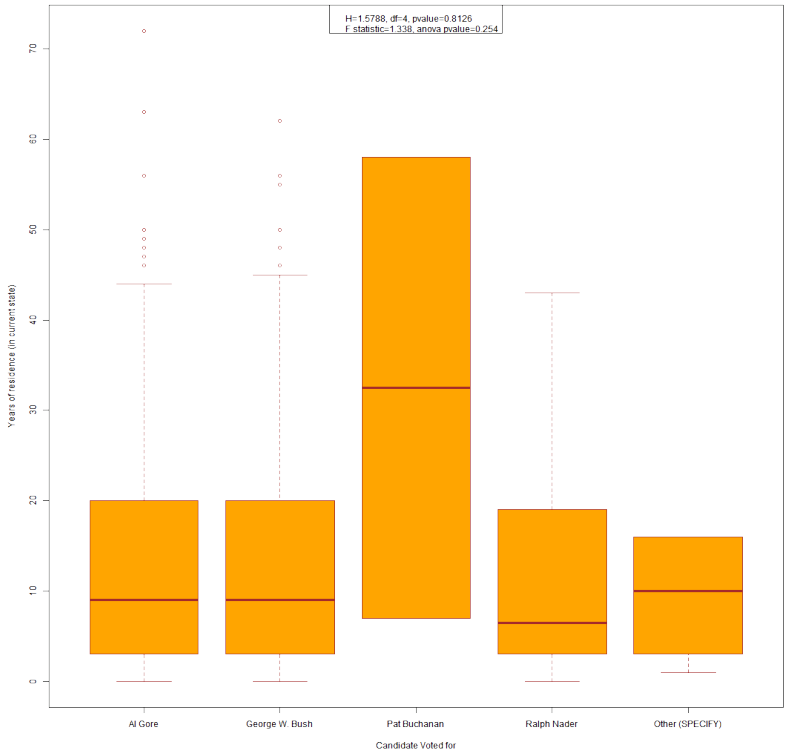


Figure 19: Barplot Ετών Διαμονής σε πολιτεία και Πολιτικών Πεποιθήσεων

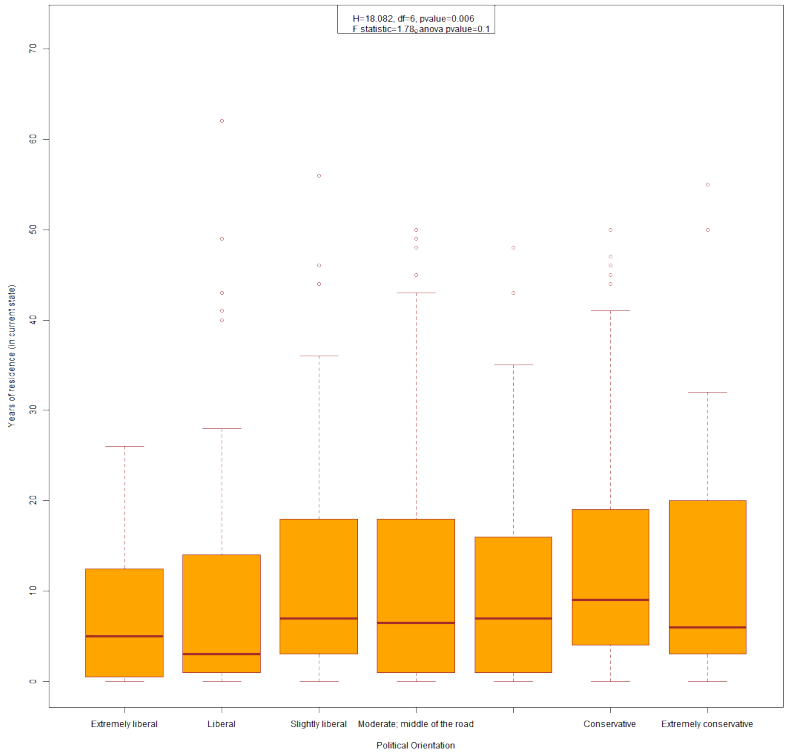


Figure 20: QQPlot Μοντέλου ANOVA1 ($Yearsofresidence \sim Vote$)

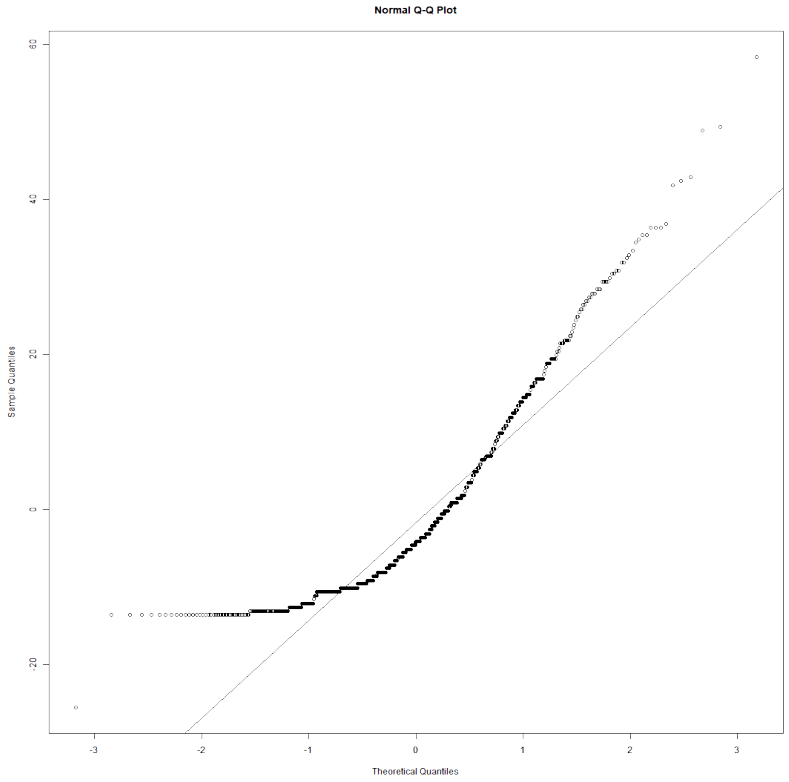


Figure 21: QQPlot Μοντέλου ANOVA2 ($Yearsofresidence \sim PoliticalOrientation$)

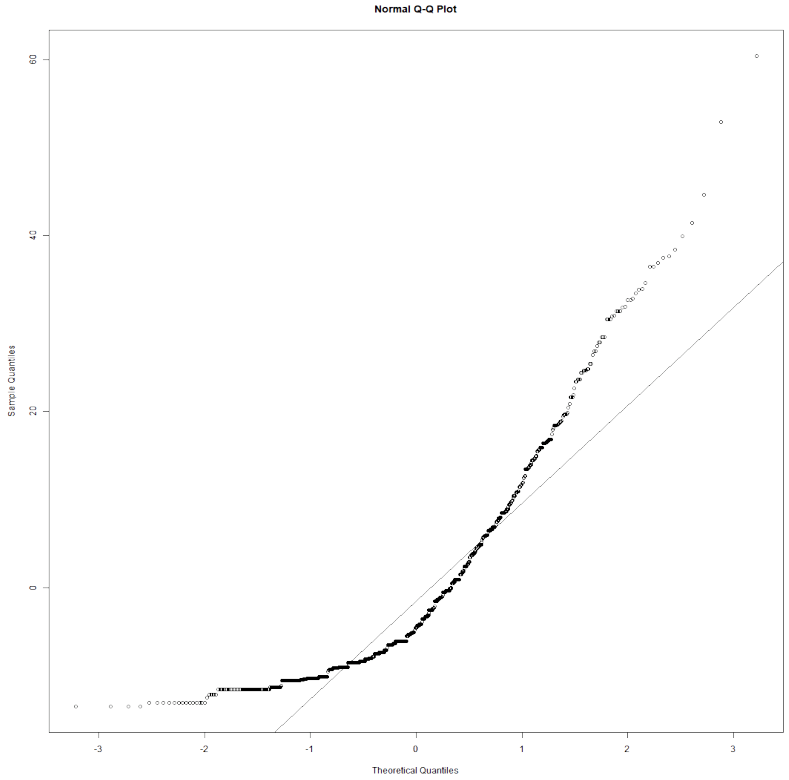


Figure 22: QQPlot Πρώτου Μοντέλου Γραμμικής Παλλινδρόμησης

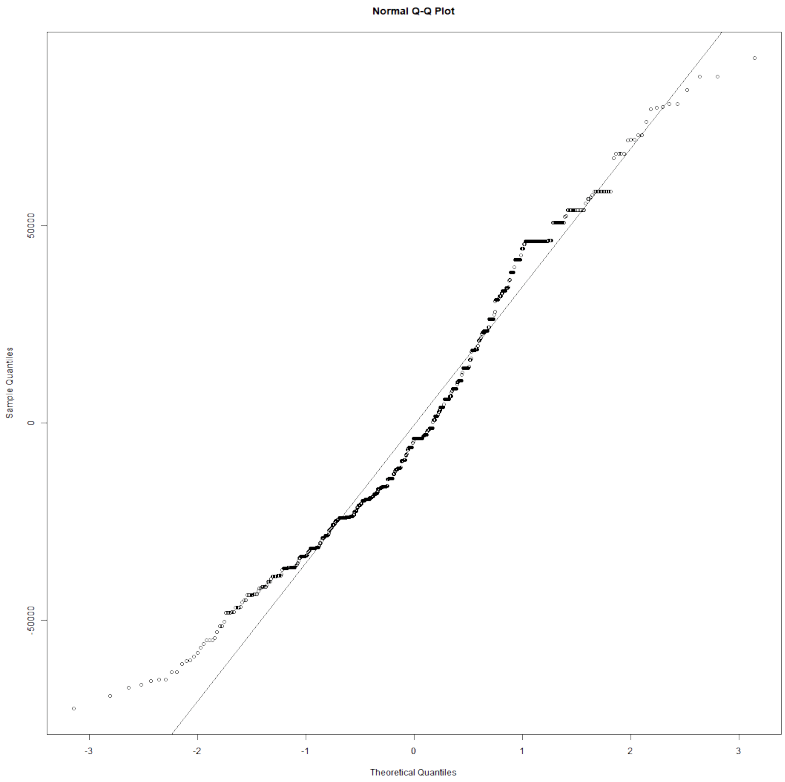
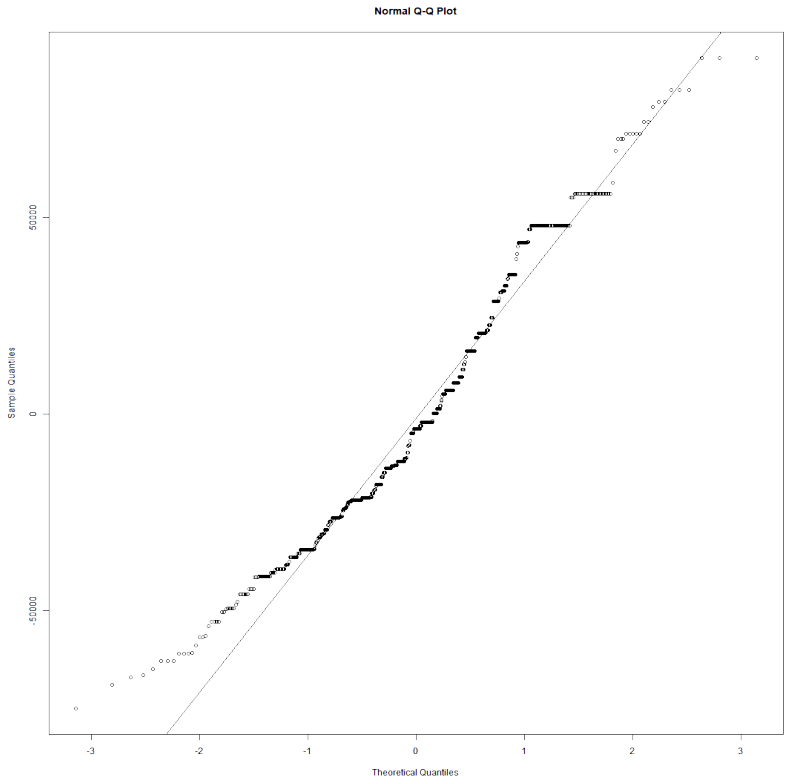


Table 4: Πίνακας Δευτέρου Μοντέλου Γραμμικής Παλλινδρόμισης

	Dependent variable:
	Income
Gender: female	−9,314.418***
	(2,380.520)
Marital_Status: Widowed	−36,510.540***
	(4,692.898)
Marital_Status: Divorced	−25,056.760***
	(3,611.607)
Marital_Status: Separated	−30,188.590***
	(6,202.189)
Marital_Status: Never married	−26,983.850***
	(2,918.340)
Marital_Status: Partnered, not married {VOL}	7,270.042
	(9,038.889)
Constant	73,793.870***
	(2,000.498)
Observations	874
R^2	0.173
Adjusted R^2	0.167
Residual Std. Error	34,363.650 (df = 867)
F Statistic	30.149*** (df = 6; 867)
Note:	*p<0.1; **p<0.05; ***p<0.01

Figure 23: QQPlot Δευτέρου Μοντέλου Γραμμικής Παλλινδρόμησης



Αναφορές

[1] R. A. Fisher. On the interpretation of $\chi 2$ from contingency tables, and the calculation of p. *Journal of the Royal Statistical Society*, 85(1):87–94, 1922.

[2] W. H. Kruskal and W. A. Wallis. Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, 47(260):583–621, 1952.

[3] H. Levene. *Robust Tests for Equality of Variance*, volume 2. 1960.

[4] H. W. Lilliefors. On the kolmogorov-smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*, 62:399–402, 1967.

[5] K. Pearson. X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302):157–175, 1900.

[6] S. S. Shapiro and M. B. Wilk. An analysis of variance test for normality (complete samples)†. *Biometrika*, 52(3-4):591–611, 12 1965.