Οικονομικό
Πανεπιστήμιο
Αθηνών

ΕΡΓΑΣΙΑ

Εκτίμηση αξίας μεταπώλησης σπιτιών με ανάλυση δεδομένων

Παληάς Ευστράτιος

ΑΘΗΝΑ 2014

ΠΕΡΙΕΧΟΜΕΝΑ

1) Εισαγωγή

2) Περιγραφική Ανάλυση

3) Σχέσεις Μεταβλητών ανά 2

4) Προβλεπτικά / Ερμηνευτικά Μοντέλα

5) Συμπεράσματα και Συζήτηση

Κεφάλαιο 1ο:Εισαγωγή

Στην εργασία αυτή μας έχουν δοθεί τα δεδομένα 8 μεταβλητών για 117 σπίτια. Σκοπός της εργασίας είναι η εκτίμηση της αξίας μεταπώλησης των σπιτιών βάσει των άλλων 7 χαρακτηριστικών τους όπως είναι η ηλικία, η θέση και το μέγεθός τους προκειμένου να εκτιμήσουμε αν και κατά πόσο η καθεμία από αυτές επηρεάζει την αξία τους. Επίσης ενδιαφερόμαστε να εξετάσουμε και τυχόν ακραίες τιμές (outliers) εφόσον η εξαίρεσή τους από την ανάλυση μπορεί να μας δώσει πιο αξιόπιστα αποτελέσματα. Για την ανάλυσή μας θα χρησιμοποιήσουμε το στατιστικό πακέτο SPSS.

Κεφάλαιο 2ο:Περιγραφική Ανάλυση

Ξεκινώντας την ανάλυσή μας θα υπολογίσουμε τα βασικά περιγραφικά μέτρα για κάθε μεταβλητή. Θα πρέπει όμως να σημειωθεί ότι για δύο μεταβλητές έχουμε και άγνωστες τιμές (missing values) τις οποίες δεν είμαστε σε θέση να γνωρίζουμε και δεν θα μπορέσουμε να τις συμπεριλάβουμε στην ανάλυσή μας. Προκύπτει ο παρακάτω πίνακας.

Descriptive Statistics

	N	Range	Minimum	Maximum	Mean	Variance
price	117	1610,00	540,00	2150,00	1062,7350	144732,300
sqft	117	2913,00	837,00	3750,00	1653,8547	274285,574
age	68	52	1	53	14,97	160,507
feats	117	8	0	8	3,53	1,975
tax	107	1542,00	223,00	1765,00	793,4860	94975,082
Valid N (listwise)	66					

1 Μέση τιμή, διακύμανση και εύρος για τις μεταβλητές.

Χρησιμοποιήσαμε μόνο τις 5 μεταβλητές αφού οι υπόλοιπες όντας κατηγορικές ψευδομεταβλητές (dummy variables) έπαιρναν μόνο τις τιμές 0 και 1. Από τον πίνακα παρατηρούμε ότι η τιμή (price), τα τετραγωνικά (sqft) και ο φόρος (tax) παρουσιάζουν αρκετά μεγάλη διακύμανση. Θα ήταν εξίσου χρήσιμο να βρούμε τις ακραίες τιμές τους. Για το σκοπό αυτό θα χρησιμοποιήσουμε το ιστόγραμμα (histogram)[1].

Κεφάλαιο 3^{ο}:Σχέσεις Μεταβλητών ανά 2

Στη συνέχεια της ανάλυσής μας θα εξετάσουμε τις ανά δύο σχέσεις μεταξύ των μεταβλητών μας[2]. Έχουμε οχτώ επομένως όλοι οι πιθανοί συνδυασμοί είναι 28. Από αυτούς έχει νόημα να εξεταστούν οι παρακάτω:

- Τιμή και καθεμία από τις υπόλοιπες.
- Φόρος και καθεμία από τις υπόλοιπες.

Ο πίνακας συσχετίσεων (πίνακας 4 του παραρτήματος) μας δίνει τα παρακάτω:

- Υψηλή συσχέτιση της τιμής με την έκταση, τα χαρακτηριστικά (feats), το αν το σπίτι είναι γωνιακό και το φόρο.
- Χαμηλή συσχέτιση της τιμής με την ηλικία, το αν έχει ξαναπουληθεί και την περιοχή που βρίσκεται.
- Υψηλή συσχέτιση του φόρου με την τιμή, την έκταση, τα χαρακτηριστικά και το αν το σπίτι είναι γωνιακό.
- Χαμηλή συσχέτιση του φόρου με την ηλικία, το αν έχει ξαναπουληθεί και την περιοχή που βρίσκεται.

Τις σχέσεις αυτές μπορούμε να τις δούμε και διαγραμματικά χρησιμοποιώντας το διάγραμμα διασπορών (scatter plot).[3]

Όπως αναμέναμε τα ζευγάρια μεταβλητών που σύμφωνα με τον πίνακα 4 έχουν υψηλό δείκτη συσχέτισης του Pearson (εδώ μεγαλύτερο του 0,5) προσεγγίζουν την εικόνα μίας ευθείας στο αντίστοιχο γράφημα ενώ για τις υπόλοιπες παρατηρείται ένα <<νέφος>> σημείων (όπως για τις μεταβλητές feats και sqft). Πρέπει επίσης να σημειωθεί ότι σε κανένα διάγραμμα δεν παρατηρούνται ακραίες παρατηρήσεις.

Όσο για τις τρεις ψευδομεταβλητές επειδή παίρνουν μόνο δύο τιμές θα χρησιμοποιήσουμε το διάγραμμα πλαισίου απολήξεων για να δούμε αν παρουσιάζουν διαφοροποίηση[4].

Από τα διαγράμματά μας παρατηρήθηκε σημαντική διαφοροποίηση της τιμής και του φόρου μόνο όσον αφορά τα γωνιακά σπίτια. Οι άλλες δύο μεταβλητές (resale και area) δεν δείχνουν να τις επηρεάζουν πολύ.

Από τις παραπάνω σχέσεις μας έχει ζητηθεί να εξετάσουμε λεπτομερέστερα αυτή μεταξύ τιμής και περιοχής. Για το σκοπό αυτό θα κάνουμε έναν έλεγχο για να το διαπιστώσουμε. Θα πρέπει όμως να προηγηθεί ένας δεύτερος για να δούμε αν η διαφορά των μέσων για τις δύο περιοχές ακολουθεί την κανονική κατανομή. Χρησιμοποιώντας τους ελέγχους Shapiro-Wilk και Kolmogorov- Smirnov και ορίζοντας επίπεδο σημαντικότητας 5% παίρνουμε τα παρακάτω:

Tests of Normality

		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	area	Statistic	df	Sig.	Statistic	df	**Sig.**
price	0	,176	39	,004	,855	39	**,000**
	1	,138	78	,001	,863	78	**,000**

a. Lilliefors Significance Correction

2 Έλεγχος Kolmogorov-Smirnov και Shapiro-Wilk για τον έλεγχο κανονικότητας της τιμής των σπιτιών.

Το **Sig.** σύμφωνα και με τους δύο ελέγχους είναι σχεδόν 0 και για τα δύο δείγματα κάτι το οποίο μας κάνει να απορρίψουμε την υπόθεση περί κανονικότητας των δεδομένων μας. Θα χρησιμοποιήσουμε επομένως τον μη παραμετρικό έλεγχο του Wilcoxon ο οποίος δεν προϋποθέτει κανονικότητα των δεδομένων. Το επίπεδο σημαντικότητας θα το ορίσουμε πάλι στο 5%. Ο έλεγχος μας δίνει τα παρακάτω:

Ranks

	area	N	Mean Rank	Sum of Ranks
price	0	39	51,81	2020,50
	1	78	62,60	4882,50
	Total	117		

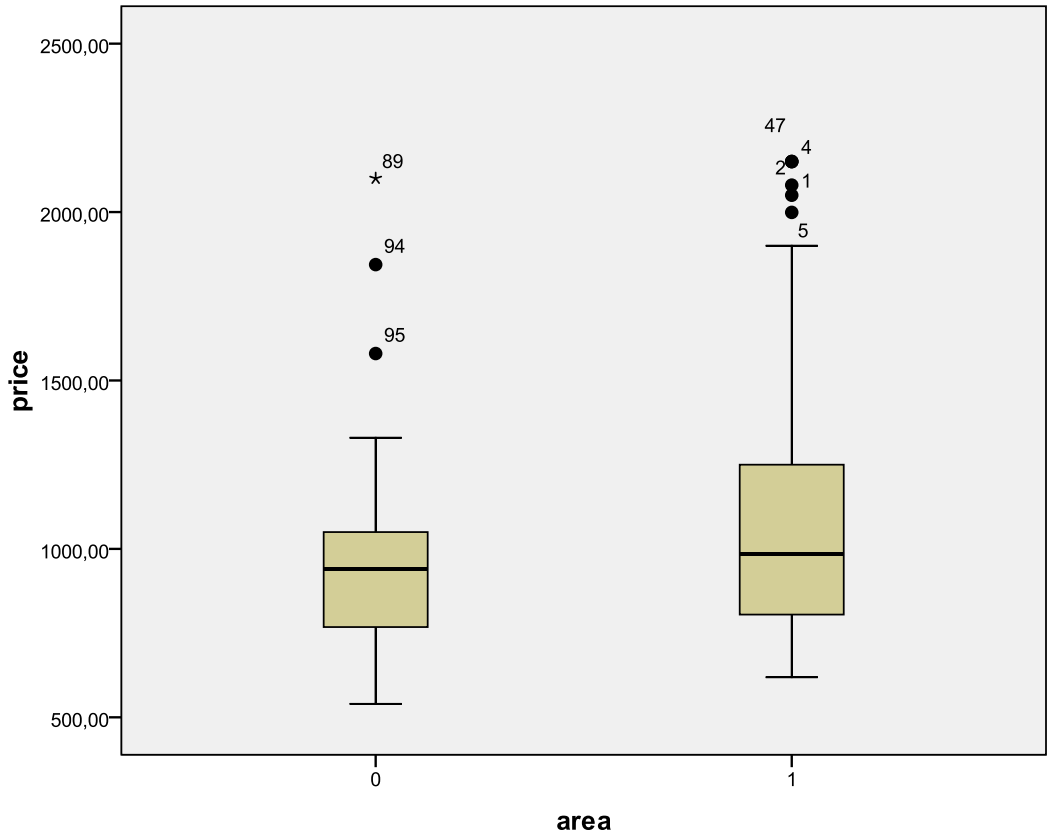
Test Statistics[a]

	price
Mann-Whitney U	1240,500
Wilcoxon W	2020,500
Z	-1,622
Asymp. Sig. (2-tailed)	**,105**

a. Grouping Variable: area

3 Έλεγχος ισότητας μέσων για τις τιμές των σπιτιών μεταξύ των δύο περιοχών.

Αφού το **Asymp. Sig. (2-tailed)** είναι 0,105 (οπότε μεγαλύτερο του 0.05) δεν απορρίπτουμε την υπόθεση περί ισότητας των μέσων. Εξάλλου και από το διάγραμμα box plot το οποίο παραθέτουμε στη συνέχεια (πίνακας 6 του παραρτήματος) δεν παρατηρείται σημαντική διαφορά.



4 Box plot για την τιμή και την περιοχή.

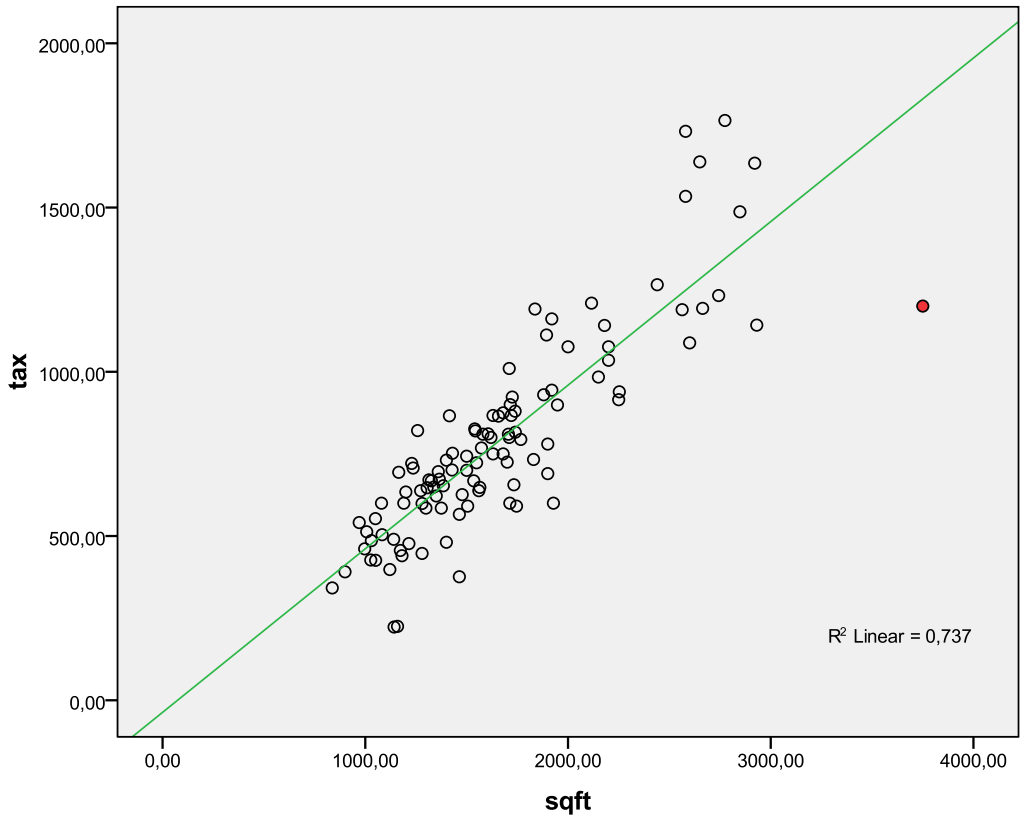
Συνεπώς είναι εύλογο να υποθέσουμε ότι η περιοχή στην οποία βρίσκεται το σπίτι δεν επηρεάζει, σημαντικά τουλάχιστον, την αξία του.

Κεφάλαιο 4ο:Προβλεπτικά/Ερμηνευτικά Μοντέλα

Στο κεφάλαιο αυτό θα εφαρμόσουμε μοντέλα για να μετρήσουμε το βαθμό που οι μεταβλητές επηρεάζονται η μία από την άλλη. Θα σταθούμε σε δύο από αυτά. Αυτό της τιμής σε σχέση με τις υπόλοιπες μεταβλητές και αυτό του φόρου σε σχέση με τις τιμές.

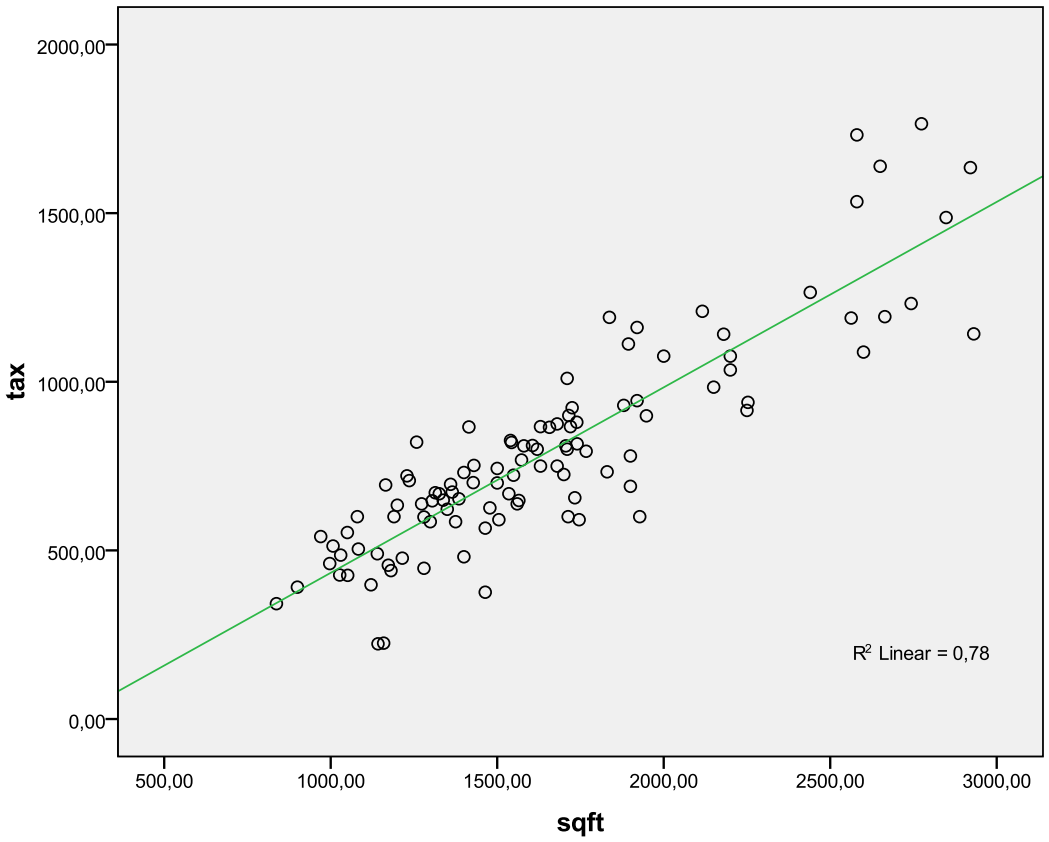
Για το πρώτο μοντέλο θα χρησιμοποιήσουμε τη διαδικασία της πολλαπλής παλινδρόμησης δηλαδή τη χρήση πολλών μεταβλητών για την πρόβλεψη/υπολογισμό μίας (της τιμής των σπιτιών). Το πρώτο πράγμα που θα πρέπει να ελέγξουμε όμως είναι η εξάρτηση μεταξύ των μεταβλητών πέραν της τιμής αφού αν συμπεριλάβουμε δύο ισχυρά εξαρτημένες στο ίδιο μοντέλο θα αλλοιώσουν την όλη εικόνα που θα μας δίνει αυτό για την πρόβλεψή μας.

Από τον πίνακα 4 του παραρτήματος παρατηρούμε ότι η μόνη ισχυρή σχέση που υπάρχει για τις μεταβλητές πέραν της τιμής (price) έγκειται στο φόρο με την έκταση όπου ο συντελεστής του Pearson είναι 0,859. Επίσης το διάγραμμα διασπορών τους παρακάτω προσεγγίζει αρκετά την ευθεία με εξαίρεση τη δεξιά <<ουρά>> στην οποία οι παρατηρήσεις μας φαίνεται να έχουν μεγαλύτερη διακύμανση. Με το πράσινο χρώμα φαίνεται η ευθεία ελαχίστων τετραγώνων η οποία μας επιτρέπει να κάνουμε τις προβλέψεις της μίας μεταβλητής μέσω της άλλης.



5 Διάγραμμα διασπορών του φόρου των σπιτιών με την έκτασή τους.

Κάτι αξιοσημείωτο που βλέπουμε στο διάγραμμα είναι η ύπαρξη μίας μόνο ακραίας παρατήρησης την οποία έχουμε σημειώσει με κόκκινο. Πρόκειται για ένα σπίτι με πολύ μεγαλύτερη έκταση από τα υπόλοιπα. Αν το εξαιρέσουμε θα πάρουμε το παρακάτω διάγραμμα:



6 Διάγραμμα διασποράς του φόρου και της έκτασης χωρίς ακραίες παρατηρήσεις.

Με την αφαίρεση που κάναμε παρατηρούμε ότι το R^2, δηλαδή ο συντελεστής συσχέτισης των δύο μεταβλητών, αυξήθηκε σημαντικά από 0,737 σε 0,78. Η ευθεία μας όμως δεν άλλαξε σημαντικά κλίση.

Θα ξεκινήσουμε με ένα μοντέλο γραμμικής παλινδρόμησης χωρίς σταθερό όρο για την αξία. Από το SPSS παίρνουμε τα παρακάτω:

Model Summary

Model	R	R Square[b]	Adjusted R Square	Std. Error of the Estimate
1	,993[a]	,985	,984	158,65627

a. Predictors: tax, resale, cor, age, area, feats, sqft

b. For regression through the origin (the no-intercept model), R Square measures the proportion of the variability in the dependent variable about the origin explained by regression. This CANNOT be compared to R Square for models which include an intercept.

Coefficients[a,b]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	sqft	,363	,095	,535	3,820	,000
	age	,219	1,806	,004	,121	,904
	feats	14,685	14,712	,050	,998	,322
	area	-28,410	45,641	-,018	-,622	,536
	cor	156,246	49,596	,062	3,150	,003
	resale	-62,003	47,397	-,024	-1,308	,196
	tax	,530	,155	,398	3,427	,001

a. Dependent Variable: price

b. Linear Regression through the Origin

7 Μοντέλο γραμμικής παλινδρόμησης της αξίας με τις υπόλοιπες 7 μεταβλητές.

Από τα αποτελέσματα παίρνουμε το παρακάτω μοντέλο:

price=0,363*sqft+0,219*age+14,685*feats-28,410*area+156,246*cor-62,003*resale+0,530*tax

Ο συντελεστής που βρίσκεται μπροστά από κάθε μεταβλητή μας δείχνει κατά πόσο αυτή επηρεάζει την αξία. Για παράδειγμα το αν το σπίτι είναι γωνιακό (μεταβλητή cor) δείχνει να ασκεί μεγαλύτερη επίδραση από το αν έχει ξαναπουληθεί (resale) αφού οι συντελεστές τους είναι 156,246 και -28,410 αντίστοιχα. Αυτό πάλι σημαίνει ότι τα σπίτια που έχουν ξαναπουληθεί (και επομένως έχουν resale=1) αναμένεται να έχουν μικρότερη αξία από τα υπόλοιπα ενώ αυτά που είναι γωνιακά (οπότε cor=1)

θα έχουν μεγαλύτερη. Στο Model Summary το R^2 μας δείχνει το ποσοστό της μεταβλητής price που εξηγείται από το μοντέλο δηλαδή πόσο καλά προσαρμόζεται αυτό στα δεδομένα μας. Με 0,985 σημαίνει ότι η προσαρμογή είναι αρκετά καλή.

Παρ' όλα αυτά όπως προαναφέραμε ο φόρος και η έκταση έχουν μεγάλη συσχέτιση μεταξύ τους. Αυτό σημαίνει ότι αν μία εξ' αυτών εξαιρεθεί από το μοντέλο δεν θα επηρεάσει σημαντικά τα αποτελέσματά μας. Για παράδειγμα αν τρέξουμε ξανά την παραπάνω παλινδρόμηση στο SPSS αλλά αυτή τη φορά δεν συμπεριλάβουμε την έκταση θα πάρουμε:

Model Summary

Model	R	R Square[b]	Adjusted R Square	Std. Error of the Estimate
1	,991[a]	**,982**	,980	175,70813

a. Predictors: tax, resale, cor, age, area, feats

b. For regression through the origin (the no-intercept model), R Square measures the proportion of the variability in the dependent variable about the origin explained by regression. This CANNOT be compared to R Square for models which include an intercept.

Coefficients[a,b]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	age	4,076	1,659	,065	2,457	,017
	feats	38,171	14,802	,129	2,579	,012
	area	-35,377	50,506	-,023	-,700	,486
	cor	176,192	54,621	,070	3,226	,002
	resale	-32,319	51,781	-,012	-,624	,535
	tax	1,073	,067	,805	15,960	,000

a. Dependent Variable: price

b. Linear Regression through the Origin

8 Γραμμική παλινδρόμηση της αξία με τις υπόλοιπες μεταβλητές εκτός της έκτασης.

Παρατηρούμε ότι με την αφαίρεση της μεταβλητής έκταση (sqft) το R^2 μειώθηκε ελάχιστα από 0,985 σε 0,982. Αυτό ήταν αναμενόμενο αφού η ισχυρή σχέση μεταξύ φόρου και έκτασης τις έκανε να δίνουν σχεδόν την ίδια πληροφορία για την αξία. Επειδή και άλλες μεταβλητές μπορεί να παρουσιάζουν το ίδιο πρόβλημα ίσως θα ήταν καλύτερο να

χρησιμοποιήσουμε μία διαδικασία πρόσθεσης ή αφαίρεσης μεταβλητών στο μοντέλο μας προκειμένου να καταλήξουμε με όσο λιγότερες μεταβλητές γίνεται αποφεύγοντας έτσι τις αλληλεπιδράσεις τους. Μία από αυτές τις μεθόδους που θα εφαρμόσουμε είναι αυτή της forward βάσει της οποίας στο μοντέλο εισάγονται διαδοχικά οι μεταβλητές που είναι στατιστικά σημαντικές[5].

Το τελικό μας μοντέλο διαμορφώνεται ως εξής:

price= 0,371*sqft + 0,545*tax + 160,954*cor

Σύμφωνα με αυτό βλέπουμε ότι μεγαλύτερη επίδραση στην αξία των σπιτιών έχει το αν είναι γωνιακά όπως μπορούμε επίσης να παρατηρήσουμε και από τον πίνακα 7 του παραρτήματος όπου τα γωνιακά σπίτια έχουν φανερά μεγαλύτερη αξία από τα υπόλοιπα. Οι υπόλοιπες τέσσερις μεταβλητές δεν ήταν στατιστικά σημαντικές και γι' αυτό δεν εισήχθησαν στο μοντέλο μας. Θα πρέπει τώρα να ελέγξουμε και αν ισχύουν οι προϋποθέσεις του μοντέλου[6].

Στη συνέχεια θα κατασκευάσουμε ένα μοντέλο που θα μας υπολογίζει τη σχέση φόρου και αξίας. Αυτή τη φορά θα χρησιμοποιήσουμε την απλή παλινδρόμηση με σταθερά έχοντας να κάνουμε μόνο με δύο μεταβλητές. Όπως και στο προηγούμενο θα πρέπει να ελέγξουμε τις προϋποθέσεις του[7].

Βλέπουμε λοιπόν ότι οι προϋποθέσεις ισχύουν και επομένως το μοντέλο tax = 0,733*price είναι κατάλληλο για την περιγραφή της σχέσης μεταξύ φόρου και αξίας των σπιτιών.

Τέλος θα υπολογίσουμε την αξία που έχει ένα μέσο (συνηθισμένο) σπίτι μέσω του μοντέλου που βγάλαμε θέτοντας κάθε μεταβλητή ίση με τη μέση τιμή της (πίνακας 1). Το αποτέλεσμα είναι **1083,049** χιλιάδες δολάρια.

Κεφάλαιο 5ο:Συμπεράσματα και Συζήτηση

Στο τελευταίο κεφάλαιο θα συνοψίσουμε τα συμπεράσματά μας από τα προηγούμενα προκειμένου να βγάλουμε μία τελική εικόνα για τα δεδομένα μας και την ανάλυση που κάναμε. Πρέπει να υπενθυμίσουμε βέβαια ότι τα missing values δεν λήφθηκαν υπ' όψιν.

Αρχικά μία σύντομη αναφορά στις συσχετίσεις:

- Υψηλή συσχέτιση της τιμής με την έκταση, τα χαρακτηριστικά (feats), το αν το σπίτι είναι γωνιακό και το φόρο.
- Χαμηλή συσχέτιση της τιμής με την ηλικία, το αν έχει ξαναπουληθεί και την περιοχή που βρίσκεται.
- Υψηλή συσχέτιση του φόρου με την τιμή, την έκταση, τα χαρακτηριστικά και το αν το σπίτι είναι γωνιακό.
- Χαμηλή συσχέτιση του φόρου με την ηλικία, το αν έχει ξαναπουληθεί και την περιοχή που βρίσκεται.

Όσο για τα δύο μοντέλα προβλέψεων που μας ζητήθηκαν βρήκαμε:

1) Για το μοντέλο της τιμής με τις υπόλοιπες:

$$price = 0{,}371*sqft + 0{,}545*tax + 160{,}954*cor$$

2)Για το μοντέλο του φόρου με την τιμή:

$$tax = 0{,}733*price$$

Και τα δύο μοντέλα τηρούν τις προϋποθέσεις που προαναφέραμε οπότε μπορούν να χρησιμοποιηθούν για ορθές προβλέψεις.

Όσο για τα outliers βρήκαμε μόνο ένα στον πίνακα 5 και αφαιρώντας το (πίνακας 6) παρατηρήσαμε σημαντική βελτίωση του συντελεστή συσχέτισης. Παρ' όλα αυτά δεν κρίναμε απαραίτητη την αφαίρεσή του από το μοντέλο.

Ως τελευταίο αποτέλεσμα υπολογίσαμε την αξία που έχει ένα μέσο (συνηθισμένο) σπίτι μέσω του μοντέλου που βγάλαμε και τη βρήκαμε ίση με 1083,049 χιλιάδες δολάρια.

Έτσι ολοκληρώσαμε την ανάλυση των δεδομένων μας. Φυσικά σταθήκαμε κυρίως στα σημεία που μας είχαν οριστεί από τα ζητούμενά μας προσέχοντας να βρούμε τις ακριβείς σχέσεις που συνδέουν τις μεταβλητές μας και να κάνουμε αξιόπιστες προβλέψεις με αυτές.

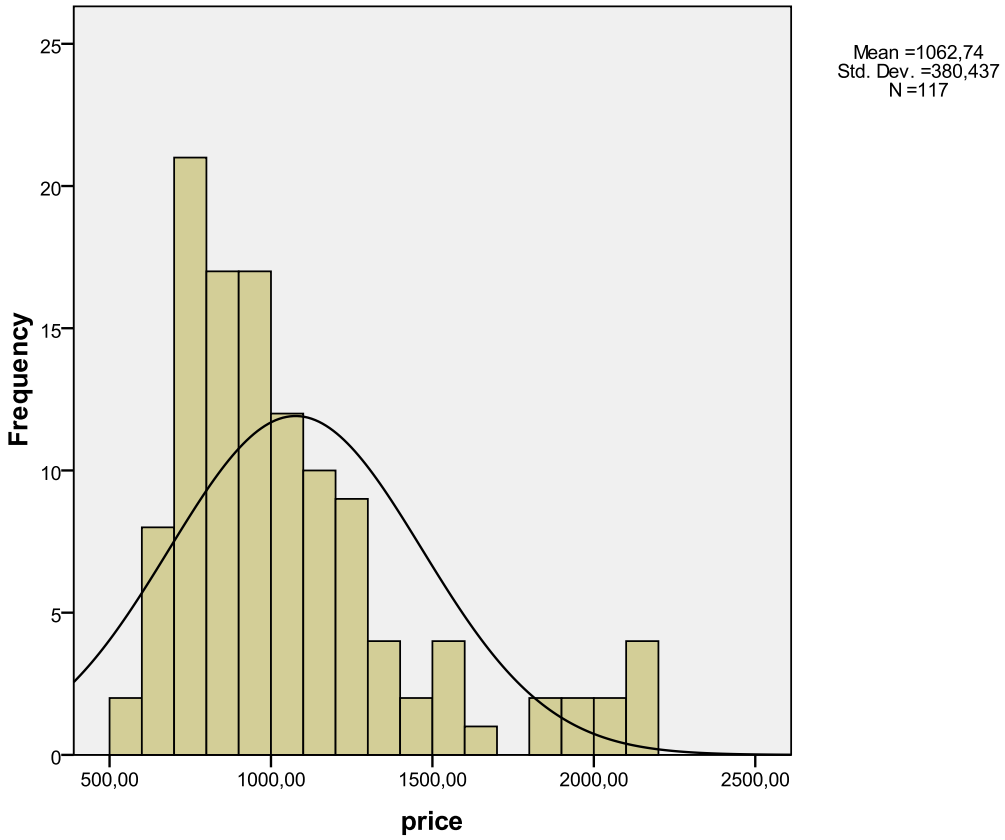
Οικονομικό
Πανεπιστήμιο
Αθηνών

Παράρτημα Εργασίας

Παληάς Ευστράτιος

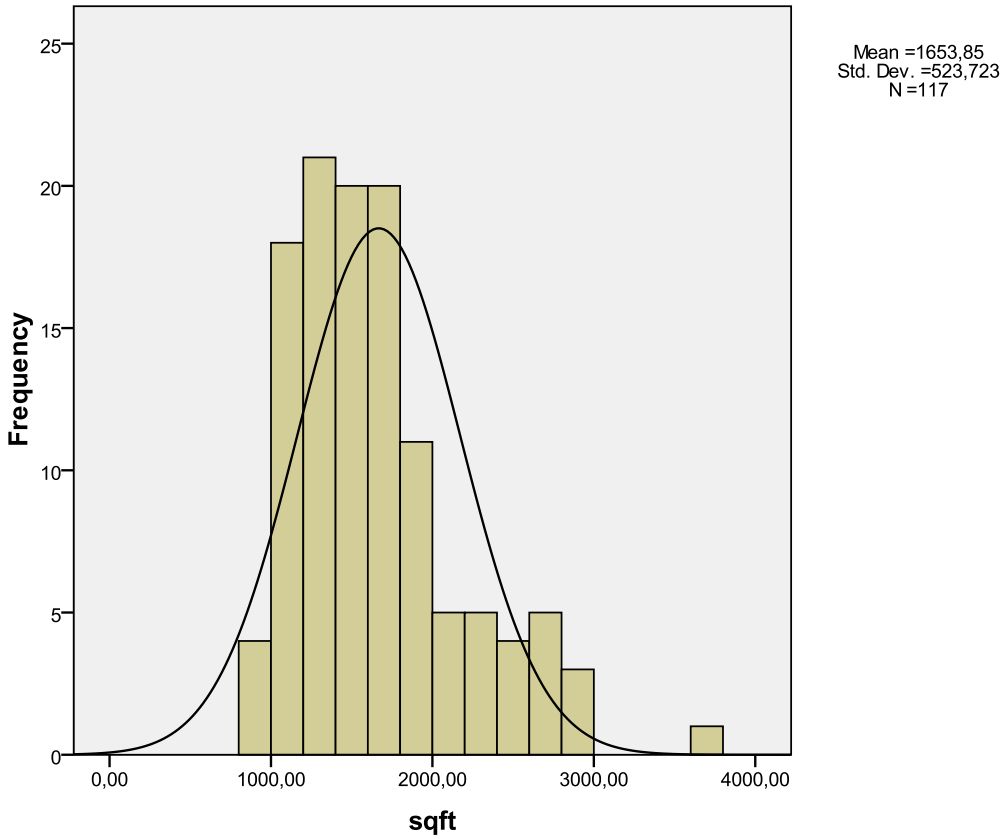
1

Παραθέτουμε εδώ τα ιστογράμματα για την αξία, την έκταση και το φόρο των σπιτιών. Η καμπύλη που συμπεριλάβαμε είναι αυτή της κανονικής κατανομής με μέση τιμή και διακύμανση αυτή των παρατηρήσεών μας. Έτσι παίρνουμε μία εικόνα για το κατά πόσο τα δεδομένα μας είναι κανονικά κατανεμημένα.



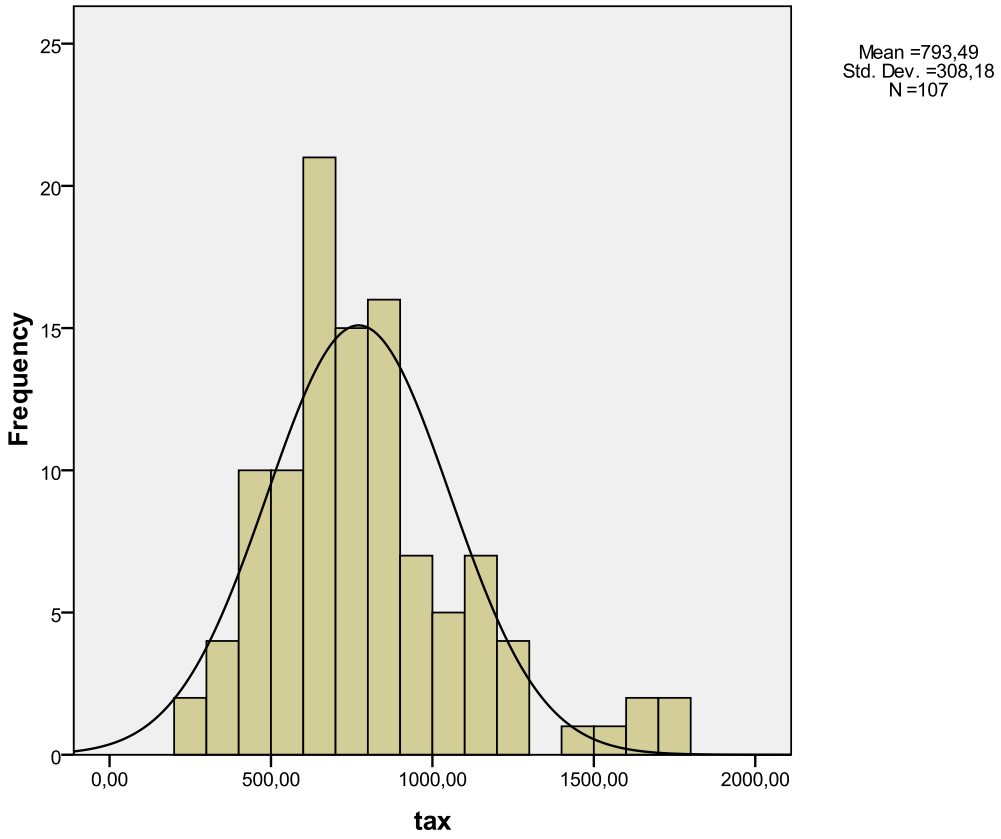
9 Ιστόγραμμα για την αξία.

Παρατηρούμε μία ελαφρά δεξιά ασυμμετρία εφόσον ένα μέρος των τιμών είναι συγκεντρωμένο στη δεξιά «ουρά» του διαγράμματος .



10 Ιστόγραμμα για την έκταση

Κι εδώ υπάρχει μία δεξιά ασυμμετρία παρόμοια με αυτή του πίνακα 1 με αρκετές τιμές να κατανέμονται δεξιά των υπολοίπων.



11 Ιστόγραμμα για το φόρο

Η κατανομή των παρατηρήσεών μας φαίνεται να είναι αρκετά συμμετρική σε αντίθεση με τους δύο προηγούμενους πίνακες. Επίσης δείχνει να ακολουθεί αρκετά πιστά την κανονική καμπύλη.

2

Correlations

		price	age	sqft	resale	feats	area	cor	tax
price	Pearson Correlation	1	-,169	,845**	-,079	,420**	,168	,555**	,876**
	Sig. (2-tailed)		,169	,000	,395	,000	,070	,000	,000
	N	117	68	117	117	117	117	117	107
age	Pearson Correlation	-,169	1	-,040	,136	-,188	,227	-,012	-,292*
	Sig. (2-tailed)	,169		,748	,267	,125	,063	,924	,017
	N	68	68	68	68	68	68	68	66
sqft	Pearson Correlation	,845**	-,040	1	,041	,395**	,145	,520**	,859**
	Sig. (2-tailed)	,000	,748		,664	,000	,119	,000	,000
	N	117	68	117	117	117	117	117	107
resale	Pearson Correlation	-,079	,136	,041	1	-,042	-,077	-,004	-,060
	Sig. (2-tailed)	,395	,267	,664		,656	,407	,966	,539
	N	117	68	117	117	117	117	117	107
feats	Pearson Correlation	,420**	-,188	,395**	-,042	1	,190*	,242**	,442**
	Sig. (2-tailed)	,000	,125	,000	,656		,040	,009	,000
	N	117	68	117	117	117	117	117	107
area	Pearson Correlation	,168	,227	,145	-,077	,190*	1	,043	,197*
	Sig. (2-tailed)	,070	,063	,119	,407	,040		,645	,042
	N	117	68	117	117	117	117	117	107
cor	Pearson Correlation	,555**	-,012	,520**	-,004	,242**	,043	1	,470**
	Sig. (2-tailed)	,000	,924	,000	,966	,009	,645		,000
	N	117	68	117	117	117	117	117	107
tax	Pearson Correlation	,876**	-,292*	,859**	-,060	,442**	,197*	,470**	1
	Sig. (2-tailed)	,000	,017	,000	,539	,000	,042	,000	
	N	107	66	107	107	107	107	107	107

**. Correlation is significant at the 0.01 level (2-tailed).

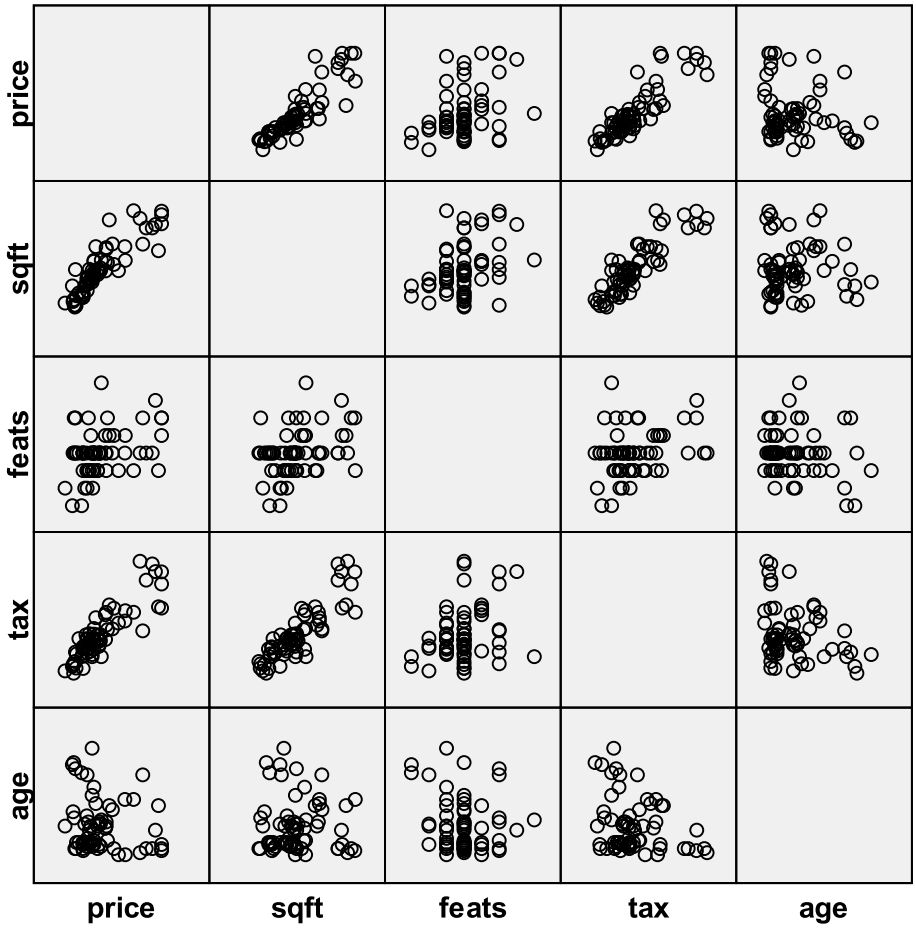
*. Correlation is significant at the 0.05 level (2-tailed).

12 Πίνακας ανά 2 συσχετίσεων για τις 8 μεταβλητές.

Κάθε κελί του πίνακα 4 μας δείχνει για τις δύο μεταβλητές (αυτή της γραμμής και αυτή της στήλης στην οποία βρίσκεται) το συντελεστή συσχέτισης του Pearson. Κελιά που έχουν υψηλό συντελεστή i.e. κοντά στη μονάδα υποδηλώνουν υψηλή θετική συσχέτιση και αυτά με συντελεστή κοντά στο 0 υποδηλώνουν χαμηλή ή και μηδενική συσχέτιση. Με κόκκινο χρώμα έχουμε υπογραμμίσει τις σχέσεις που έχουν νόημα να εξετάσουμε. Ο αριθμός της γραμμής Sig. (2-tailed) μας δείχνει αν η σχέση αυτή είναι στατιστικά σημαντική. Συγκεκριμένα αν ο αριθμός αυτός είναι μικρότερος από το επίπεδο σημαντικότητας (εδώ 0,05), τότε η σχέση είναι στατιστικά σημαντική, διαφορετικά δεν είναι. Είναι προφανές ότι όσο πιο υψηλός είναι ο συντελεστής του Pearson τόσο πιο χαμηλή θα είναι η τιμή του Sig. (2-tailed).

3

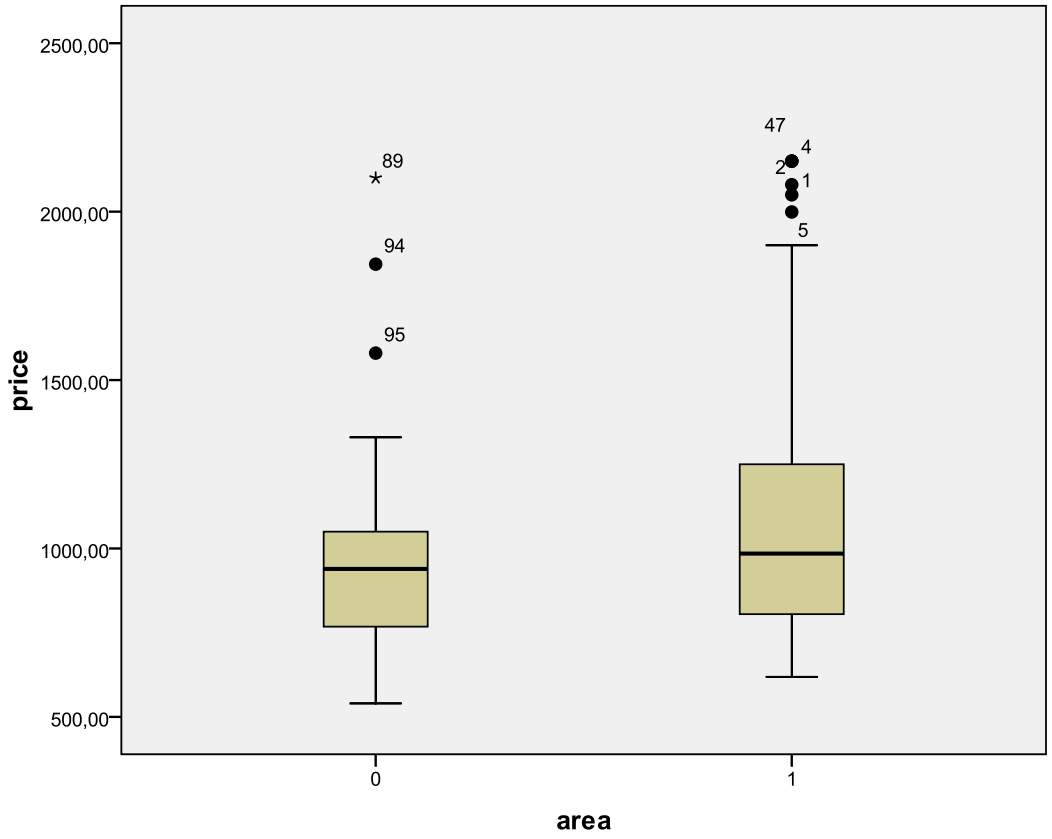
Ένας πίνακας με όλα τα διαγράμματα διασπορών φαίνεται παρακάτω. Οι μεταβλητές που παρουσιάζουν υψηλή συσχέτιση έχουν διάγραμμα που μοιάζει με ευθεία. Ένα παράδειγμα φανερής συσχέτισης είναι αυτό της τιμής (price) και του μεγέθους (sqft). Δεν συμπεριλάβαμε τις τρεις ψευδομεταβλητές μας αφού έχουν μόνο δύο τιμές. Γι' αυτές θα χρησιμοποιήσουμε το διάγραμμα πλαισίου απολήξεων (box plot) στη συνέχεια.



13 Πίνακας διαγραμμάτων διασποράς για 5 μεταβλητές.

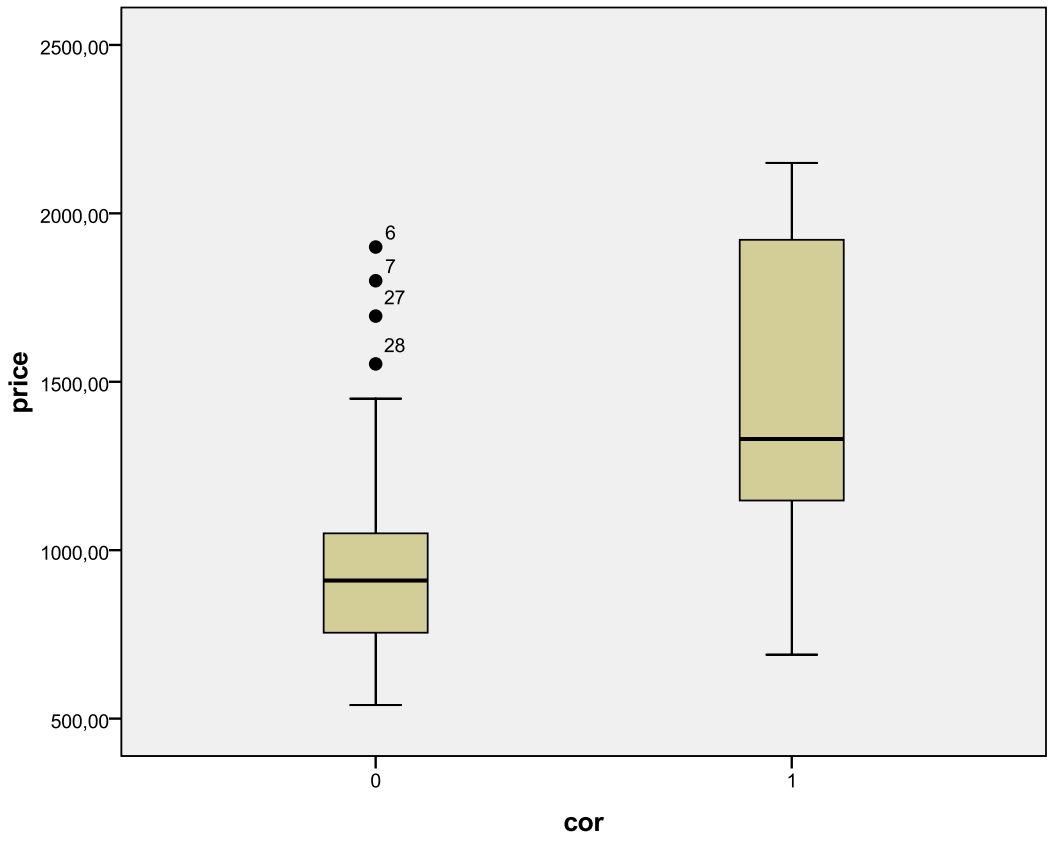
4

Θα πάρουμε δύο διαγράμματα για κάθε μία από τις τρεις ψευδομεταβλητές. Ένα για την τιμή και ένα για το φόρο. Θα ξεκινήσουμε από την τιμή.



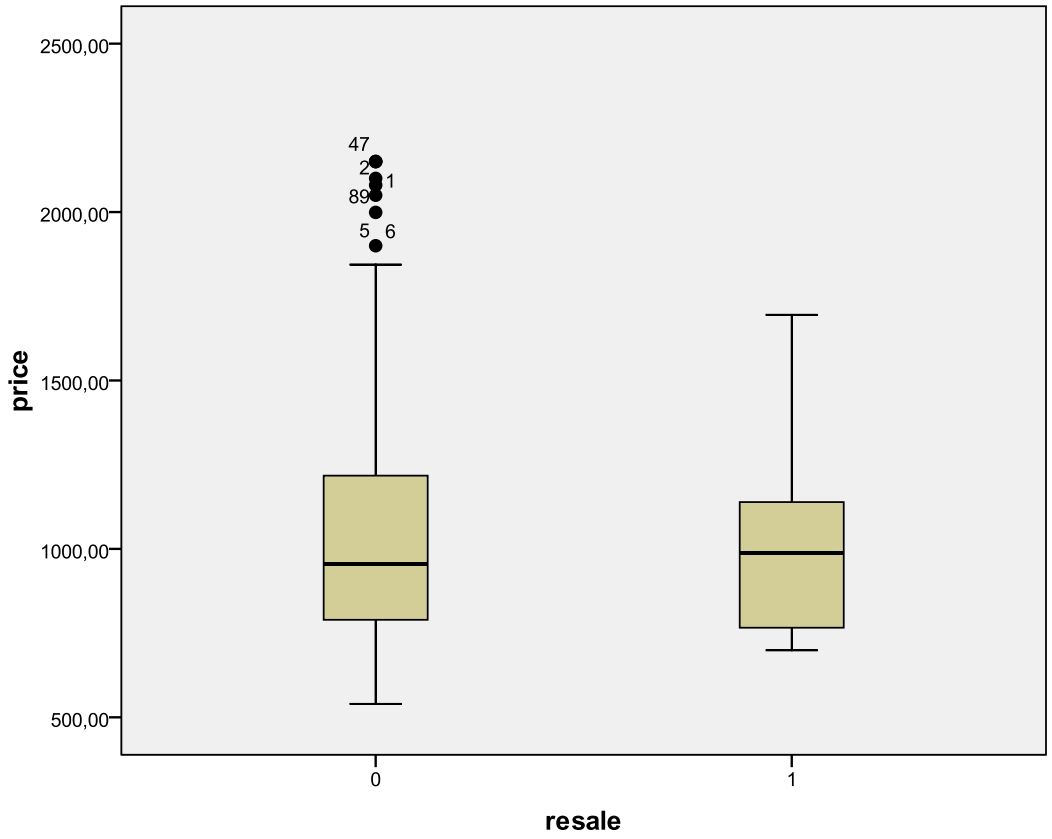
14 Box plot για την τιμή και την περιοχή.

Στο διάγραμμα αυτό για τις δύο περιοχές (0 και 1) έχουμε ένα παραλληλόγραμμο σχήμα με δύο απολήξεις (πάνω και κάτω). Τα πάνω και κάτω άκρα του παραλληλογράμμου μας δείχνουν το τρίτο και πρώτο τεταρτημόριο αντίστοιχα ενώ η οριζόντια γραμμή μέσα σε αυτό μας δείχνει τη διάμεσο. Εν ολίγοις το γράφημα αυτό μας δείχνει το εύρος στο οποίο κατανέμονται οι παρατηρήσεις μας. Βλέπουμε εδώ ότι η περιοχή 1 (καλύτερο σημείο σύμφωνα με τα δεδομένα μας) παρουσιάζει μεγαλύτερες τιμές από την περιοχή 0. Παρόλα αυτά η διάμεσος καθώς και το εύρος δεν φαίνεται να διαφέρουν πολύ. Οι τιμές που εμφανίζονται με τη μορφή κουκίδας είναι τα outliers.



15 Box plot για την τιμή και το αν το σπίτι είναι γωνιακό.

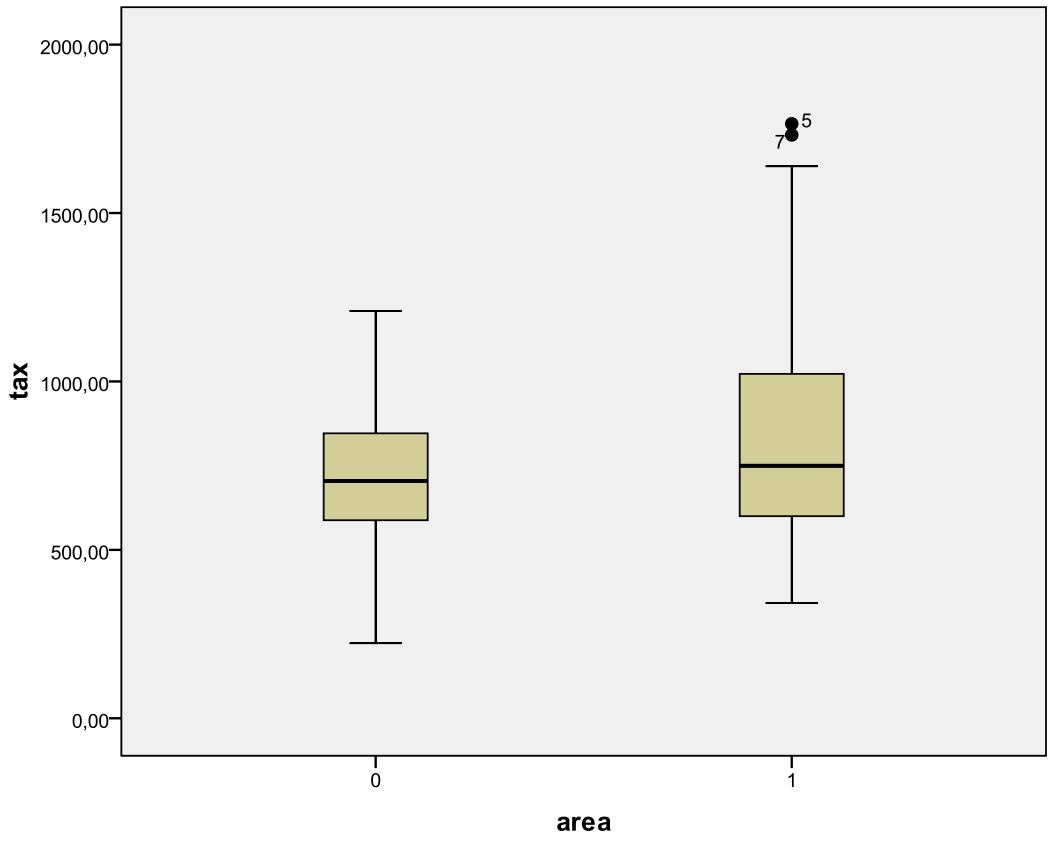
Παρατηρούμε μία φανερή αύξηση της τιμής στα γωνιακά σπίτια από τα υπόλοιπα τόσο στη διάμεσο όσο και στο εύρος. Αξίζει επίσης να σημειωθεί η ύπαρξη ακραίων τιμών μόνο στα μη γωνιακά σπίτια.



16 Box plot για την τιμή και το αν το σπίτι έχει ξαναπουληθεί.

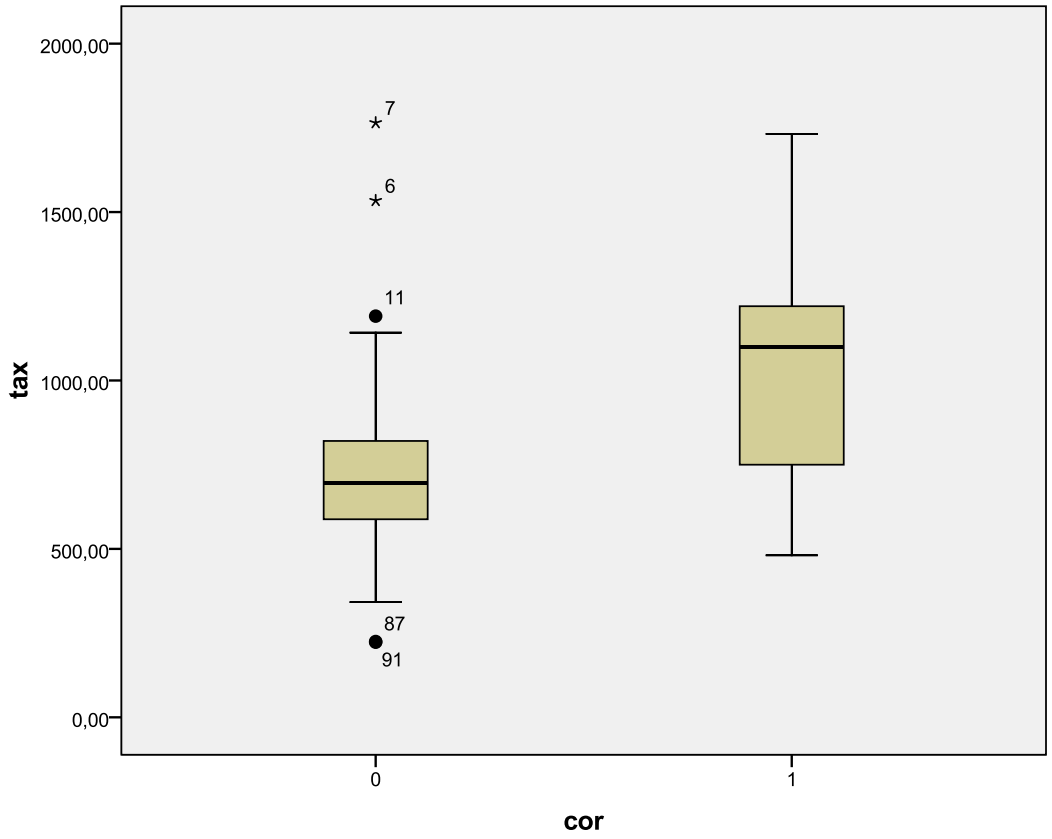
Εδώ δε φαίνεται να υπάρχει σημαντική διαφοροποίηση της τιμής μεταξύ των δύο αυτών κατηγοριών. Η μόνη αλλαγή έγκειται στο εύρος και στις ακραίες παρατηρήσεις της πρώτης χωρίς όμως αυτές να την επηρεάζουν σημαντικά.

Για τους φόρους έχουμε τα παραάτω:



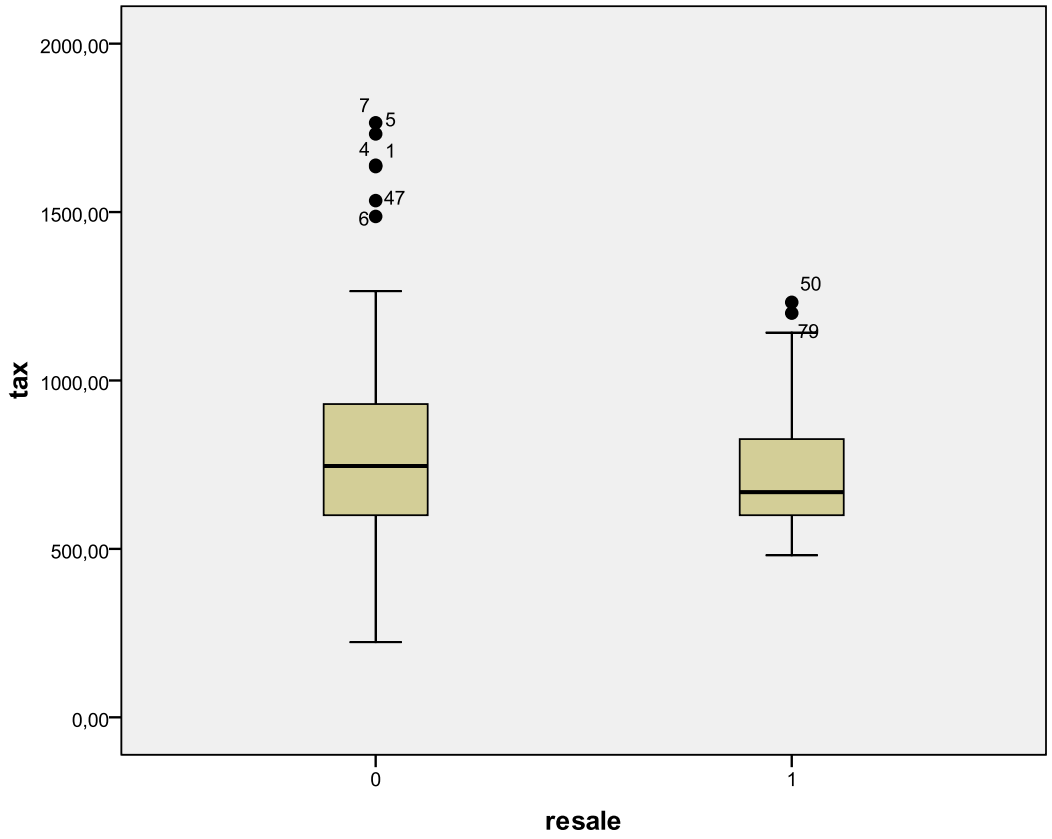
17 Box plot για τον φόρο και την περιοχή.

Στο διάγραμμα δεν φαίνεται σημαντική διαφοροποίηση το φόρου μεταξύ των δύο περιοχών. Στην area 1 οι τιμές δείχνουν ελαφρά μόνο πιο αυξημένες.



18 Box plot για τον φόρο και το αν το σπίτι είναι γωνιακό.

Εδώ παρατηρούμε μία φανερή αύξηση του φόρου όταν το σπίτι είναι γωνιακό.



19 Box plot για τον φόρο και το αν το σπίτι έχει ξαναπουληθεί.

Όπως και στον πίνακα 9 δεν φαίνεται σημαντική διαφοροποίηση το φόρου μεταξύ των δύο κατηγοριών παρά μόνο μία μικρή αύξηση του εύρους και της διαμέσου για τα σπίτια που δεν έχουν ξαναπουληθεί.

5

Τα αποτελέσματα που παίρνουμε είναι τα ακόλουθα:

Variables Entered/Removed[a,b]

Model	Variables Entered	Variables Removed	Method
1	sqft	.	Forward (Criterion: Probability-of-F-to-enter <= ,050)
2	tax	.	Forward (Criterion: Probability-of-F-to-enter <= ,050)
3	cor	.	Forward (Criterion: Probability-of-F-to-enter <= ,050)

a. Dependent Variable: price

b. Linear Regression through the Origin

Model Summary

Model	R	R Square[b]	Adjusted R Square	Std. Error of the Estimate
1	,988[a]	,977	,976	189,66555
2	,991[c]	,982	,981	169,30391
3	,992[d]	**,984**	,984	157,67831

a. Predictors: sqft

b. For regression through the origin (the no-intercept model), R Square measures the proportion of the variability in the dependent variable about the origin explained by regression. This CANNOT be compared to R Square for models which include an intercept.

c. Predictors: sqft, tax

d. Predictors: sqft, tax, cor

Coefficients[a,b]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	sqft	,670	,013	,988	52,307	,000
2	sqft	,384	,069	,566	5,549	,000
	tax	,570	,136	,428	4,192	,000
3	sqft	,371	,065	,547	5,742	,000
	tax	,545	,127	,409	4,296	,000
	cor	160,954	49,010	,064	3,284	,002

a. Dependent Variable: price

b. Linear Regression through the Origin

20 Γραμμική παλινδρόμηση για την αξία με τη μέθοδο stepwise.

Οι μεταβλητές που προστέθηκαν στο μοντέλο είναι διαδοχικά η έκταση (sqft), ο φόρος (tax) και το αν το σπίτι είναι γωνιακό (cor). Το μοντέλο μας διαμορφώνεται ως εξής:

$$price = 0,371*sqft + 0,545*tax + 160,954*cor$$

με $R^2 = 0,984$ το οποίο υποδηλώνει αρκετά καλή προσαρμογή

Ερμηνεία: Η αξία ενός σπιτιού που είναι γωνιακό είναι κατά μέσο όρο προσαυξημένη κατά 160,954 από ένα που δεν είναι υποθέτωντας ότι έχουν την ίδια έκταση και τον ίδιο φόρο.

Αντίστοιχα για τις άλλες δύο μεταβλητές μας έχουμε: Η διαφορά της αξίας μεταξύ δύο σπιτιών που έχουν ίδιο φόρο και ίδιο cor (δηλαδή είτε είναι και τα δύο γωνιακά είτε δεν είναι) αλλά διαφέρουν κατά 1 στην έκταση αναμένεται να είναι 0.371.

Τέλος αν συγκρίνουμε δύο σπίτια με το ίδιο cor και την ίδια έκταση αλλά που διαφέρουν κατά μία μονάδα στο φόρο αναμένουμε διαφορά κατά 0.545 στις αξίες τους.

6

Οι προϋποθέσεις τις οποίες θα πρέπει να ελέγξουμε είναι:

1)Η κανονικότητα των καταλοίπων.

2)Η ανεξαρτησία των καταλοίπων.

3)Η ομοσκεδαστικότητα των καταλοίπων.

4)Η γραμμικότητα μεταξύ των μεταβλητών του μοντέλου.

Τα κατάλοιπα είναι η διαφορά της προβλεπόμενης τιμής της αξίας των σπιτιών σύμφωνα με το μοντέλο από την πραγματική (μεταβλητή price).

Η ισχύς τους είναι πολύ σημαντική γιατί αν παραβιάζονται υπάρχει μεγάλη πιθανότητα να πάρουμε ανακριβή αποτελέσματα.

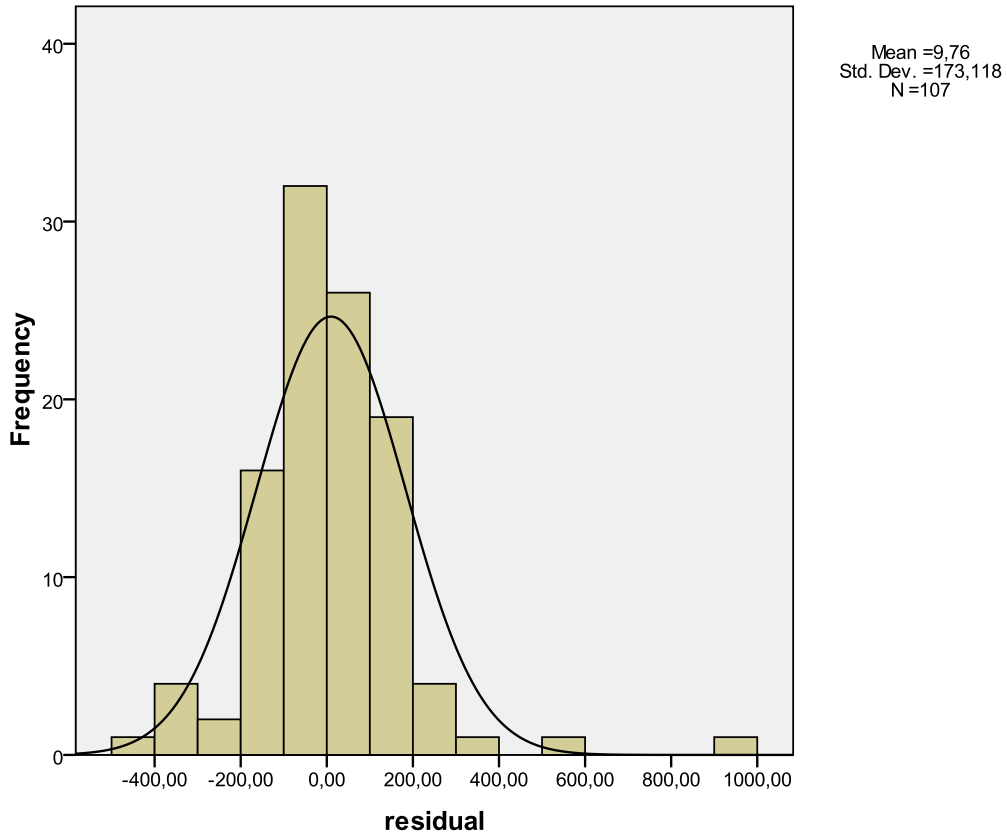
Ξεκινώντας από την κανονικότητα των καταλοίπων θα κοιτάξουμε το ιστόγραμμά τους και θα κάνουμε και τον κατάλληλο έλεγχο κανονικότητας.

Τα κατάλοιπα θα τα παρουσιάσουμε με μία καινούργια μεταβλητή την **residual** η οποία θα είναι η διαφορά μεταξύ της πραγματικής αξίας των σπιτιών (μεταβλητή price) και της προβλεπόμενης από το μοντέλο αξίας τους (μεταβλητή pricehat). Θα την υπολογίσουμε ως εξής:

pricehat = 0,371*sqft + 0,545*tax + 160,954*cor

residual = price – pricehat

Έχουμε λοιπόν το ιστόγραμμα:



21 Ιστόγραμμα των καταλοίπων του μοντέλου της αξίας με το φόρο, την έκταση και το αν το σπίτι είναι γωνιακό.

Βλέπουμε ότι τα κατάλοιπα ακολουθούν αρκετά πιστά την κανονική καμπύλη. Για τη συνέχεια θα κάνουμε και τον έλεγχο κανονικότητας Kolmogorov-Smirnov ο οποίος μας δίνει τα παρακάτω:

One-Sample Kolmogorov-Smirnov Test

		residual
N		107
Normal Parameters[a,,b]	Mean	9,7600
	Std. Deviation	173,11814
Most Extreme Differences	Absolute	,111
	Positive	,082
	Negative	-,111
Kolmogorov-Smirnov Z		1,148
Asymp. Sig. (2-tailed)		**,143**

a. Test distribution is Normal.

b. Calculated from data.

22 Έλεγχος κανονικότητας Kolmogorov-Smirnov για τα κατάλοιπα της παλινδρόμησης της αξίας με την έκταση, το φόρο και το αν το σπίτι είναι γωνιακό.

Το **Asymp. Sig. (2-tailed)** είναι 0,143 άρα μεγαλύτερο του 0,05 οπότε δεν απορρίπτουμε την υπόθεση περί κανονικότητας των καταλοίπων.

Για την ανεξαρτησία θα χρησιμοποιήσουμε τον δείκτη Durbin-Watson

Model Summary[e,f]

Model	R	R Square[b]	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	,988[a]	,977	,976	189,66555	
2	,991[c]	,982	,981	169,30391	
3	,992[d]	,984	,984	157,67831	**1,801**

a. Predictors: sqft

b. For regression through the origin (the no-intercept model), R Square measures the proportion of the variability in the dependent variable about the origin explained by regression. This CANNOT be compared to R Square for models which include an intercept.

c. Predictors: sqft, tax

d. Predictors: sqft, tax, cor

e. Dependent Variable: price

f. Linear Regression through the Origin

23 Δείκτης Durbin-Watson της παλινδρόμησης της τιμής με το φόρο, την έκταση και το αν το σπίτι είναι γωνιακό.

Στον παραπάνω πίνακα τα 3 **μοντέλα** είναι αυτά που πήραμε διαδοχικά με την μέθοδο forward και μας ενδιαφέρει το τρίτο (model 3) στο οποίο και καταλήξαμε. Για το μοντέλο αυτό βλέπουμε ότι ο δείκτης του Durbin-Watson είναι 1,801 δηλαδή αρκετά κοντά στο 2 πράγμα που υποστηρίζει σε μεγάλο βαθμό ότι τα κατάλοιπά μας είναι ανεξάρτητα.

Για να ελέγξουμε τώρα την ομοσκεδαστικότητα θα πρέπει να πάρουμε την ανάλυση διακύμανσης χωρίζοντας τις παρατηρήσεις μας σε ομάδες. Διαλέγουμε οι ομάδες αυτές να γίνουν βάσει της μεταβλητής feats δίνοντάς μας 9 ομάδες.

Test of Homogeneity of Variances

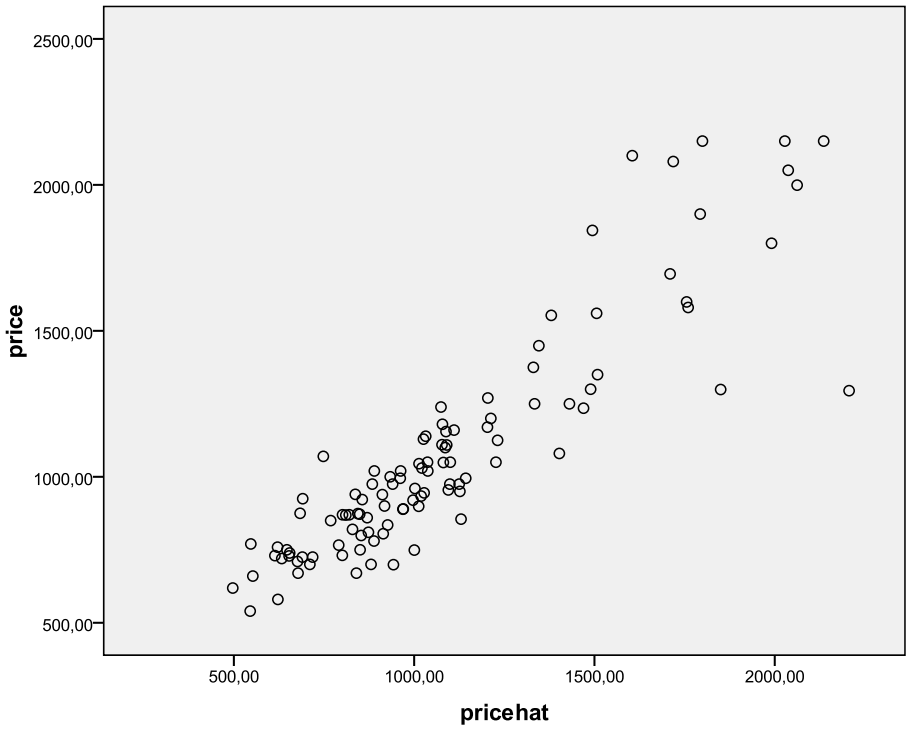
residual

Levene Statistic	df1	df2	Sig.
,607	6	98	,724

24 Έλεγχος του Levene για την ισότητα/ομοσκεδαστικότητα των διακυμάνσεων.

Το **Sig.** είναι 0,724, μεγαλύτερο του 0,05, άρα δεν απορρίπτουμε την υπόθεση περί ομοσκεδαστικότητας των διακυμάνσεων.

Τέλος για τη γραμμικότητα μεταξύ της προβλεπόμενης και πραγματικής αξίας των σπιτιών θα αρκεστούμε στο διάγραμμα διασπορών τους.



25 Διάγραμμα διασπορών μεταξύ προβλεπόμενης και πραγματικής αξίας των σπιτιών.

Η αρκετά πιστή προσέγγιση της ευθείας του διαγράμματος μας δίνει την εικόνα πολύ καλής γραμμικότητας από το μοντέλο μας.

7

Με την απλή παλινδρόμηση παίρνουμε τα παρακάτω:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,876[a]	**,767**	,765	149,53329

a. Predictors: (Constant), price

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	**36,344**	43,237		,841	,402
	price	**,703**	,038	,876	18,581	,000

a. Dependent Variable: tax

26 Γραμμική παλινδρόμηση του φόρου των σπιτιών σε σχέση με την τιμή τους.

Το R^2 είναι 0,767 και το μοντέλο μας διαμορφώνεται ως εξής:

tax = 36,344 + 0,703*price

Έχουμε λοιπόν τα παρακάτω:

1) Αύξηση της τιμής κατά μία μονάδα συνεπάγεται στην αύξηση του φόρου κατά 0,703.

2) Ένα σπίτι με μηδενική τιμή θα έχει φόρο ίσο με 36,344. Αυτό όμως δεν έχει νόημα αφού κανένα από τα σπίτια μας δεν έχει μηδενική αξία. Επομένως θα ήταν ίσως πιο λογικό να ξαναπάρουμε το μοντέλο αλλά χωρίς σταθερά. Έτσι έχουμε:

Model Summary

Model	R	R Square[b]	Adjusted R Square	Std. Error of the Estimate
1	,985[a]	**,969**	,969	149,32617

a. Predictors: price

Model Summary

Model	R	R Square[b]	Adjusted R Square	Std. Error of the Estimate
1	,985[a]	**,969**	,969	149,32617

a. Predictors: price

b. For regression through the origin (the no-intercept model), R Square measures the proportion of the variability in the dependent variable about the origin explained by regression. This CANNOT be compared to R Square for models which include an intercept.

Coefficients[a,b]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	price	**,733**	,013	,985	58,024	,000

a. Dependent Variable: tax

b. Linear Regression through the Origin

27 Γραμμική παλινδρόμηση του φόρου σε σχέση με την τιμή χωρίς σταθερά.

Το καινούργιο μοντέλο (χωρίς πλέον τη σταθερά) είναι το:

tax = 0,733*price

Οπότε τώρα η αύξηση του φόρου κατά μία μονάδα συνεπάγεται στην αύξηση της τιμής κατά 0.733 ενώ το R^2 αυξήθηκε στο 0,969 δίνοντας μας την εικόνα πολύ ισχυρής σχέσης των δύο αυτών μεταβλητών.

Για το γραμμικό μοντέλο που βγάλαμε θα πρέπει τώρα να ελέγξουμε αν τηρούνται οι προϋποθέσεις του. Όπως προαναφέραμε και στο 6 για το προηγούμενο μοντέλο, αυτές είναι οι παρακάτω:

1)Η κανονικότητα των καταλοίπων.

2)Η ανεξαρτησία των καταλοίπων.

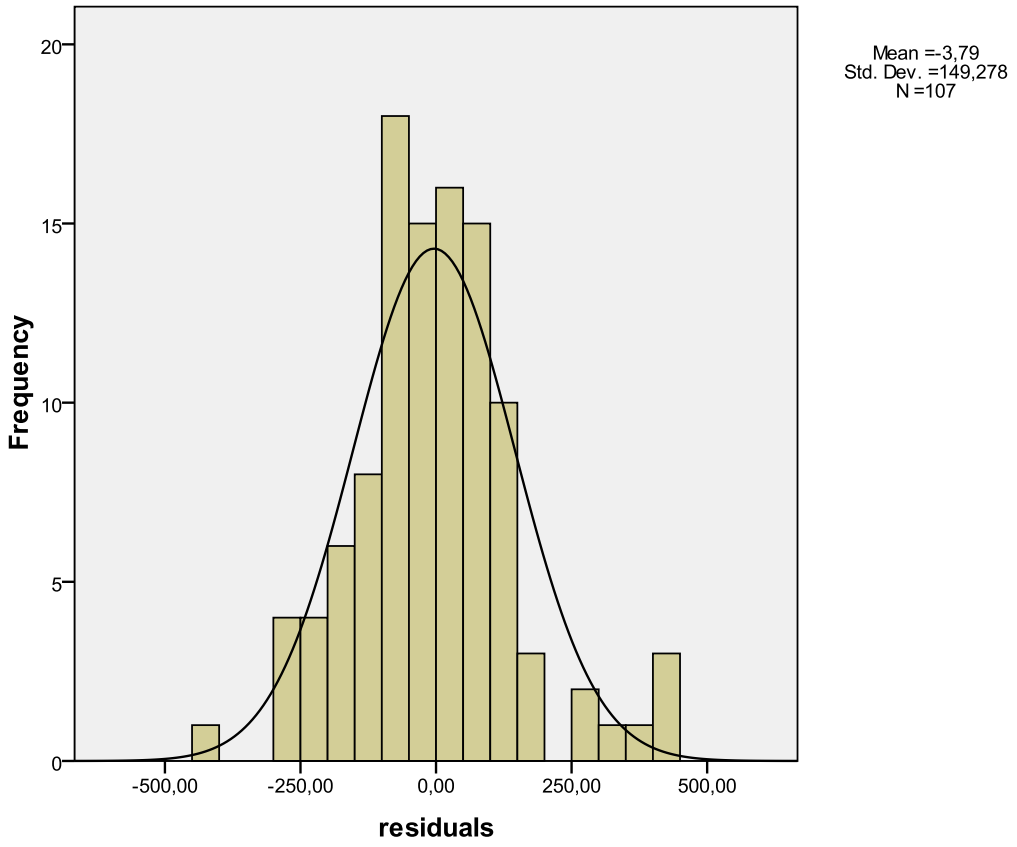
3)Η ομοσκεδαστικότητα των καταλοίπων.

4)Η γραμμικότητα μεταξύ των μεταβλητών του μοντέλου.

Για τον έλεγχο της κανονικότητας θα πάρουμε και πάλι το ιστόγραμμα των καταλοίπων και το Q-Q Plot τα οποία μας δείχνουν πόσο πιστά προσεγγίζουν την κανονική κατανομή. Για να γίνει αυτό θα πρέπει να τα δημιουργήσουμε μία νέα μεταβλητή που θα την ονομάσουμε residuals.

Αρχικά θα ορίσουμε τη μεταβλητή **taxhat = price*0.733** που είναι ο φόρος που μας δίνει το μοντέλο που βρήκαμε με την παλινδρόμηση. Τα

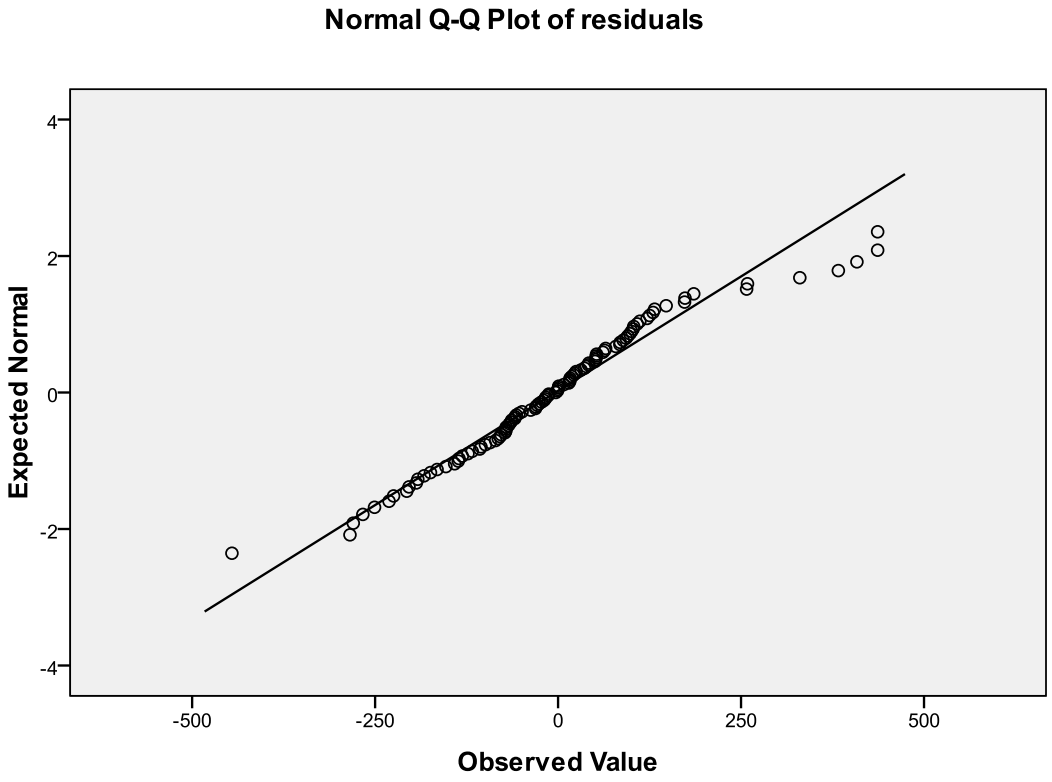
κατάλοιπα επομένως θα οριστούν με τον τύπο **residuals= taxhat - tax**. Έχουμε τα παρακάτω:



28 Ιστόγραμμα των καταλοίπων της παλινδρόμησης του φόρου με την αξία των σπιτιών.

Το ιστόγραμμα μας δείχνει αρκετά καλή προσαρμογή των καταλοίπων στην κανονική κατανομή η οποία συμβολίζεται με την καμπύλη του διαγράμματος.

Το επόμενο διάγραμμα είναι το Q-Q Plot:



29 Q-Q Plot για τα κατάλοιπα της παλινδρόμησης του φόρου με την αξία.

Και σε αυτό το διάγραμμα παρατηρούμε μία αρκετά καλή προσαρμογή των καταλοίπων στην κανονική κατανομή (δηλαδή στην ευθεία που έχει).

Τέλος θα κάνουμε τον έλεγχο του Kolmogorov-Smirnov για τον έλεγχο της κανονικότητας. Παίρνουμε:

	Test of Normality		
	Kolmogorov-Smirnov[a]		
	Statistic	df	Sig.
residuals	,079	107	,094

30 Έλεγχος κανονικότητας για το μοντέλο του φόρου με την αξία.

Το **Sig.** είναι 0,094 οπότε μεγαλύτερο του επιπέδου σημαντικότητας (που έχουμε ορίσει στο 0,05) άρα δεν απορρίπτουμε την υπόθεση περί της κανονικότητας των καταλοίπων.

Για την ανεξαρτησία των καταλοίπων θα χρησιμοποιήσουμε τον έλεγχο Durbin-Watson ο οποίος μας δίνει:

Model Summary[c,d]

Model	R	R Square[b]	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	,985[a]	,969	,969	149,32617	1,566

a. Predictors: price

b. For regression through the origin (the no-intercept model), R Square measures the proportion of the variability in the dependent variable about the origin explained by regression. This CANNOT be compared to R Square for models which include an intercept.

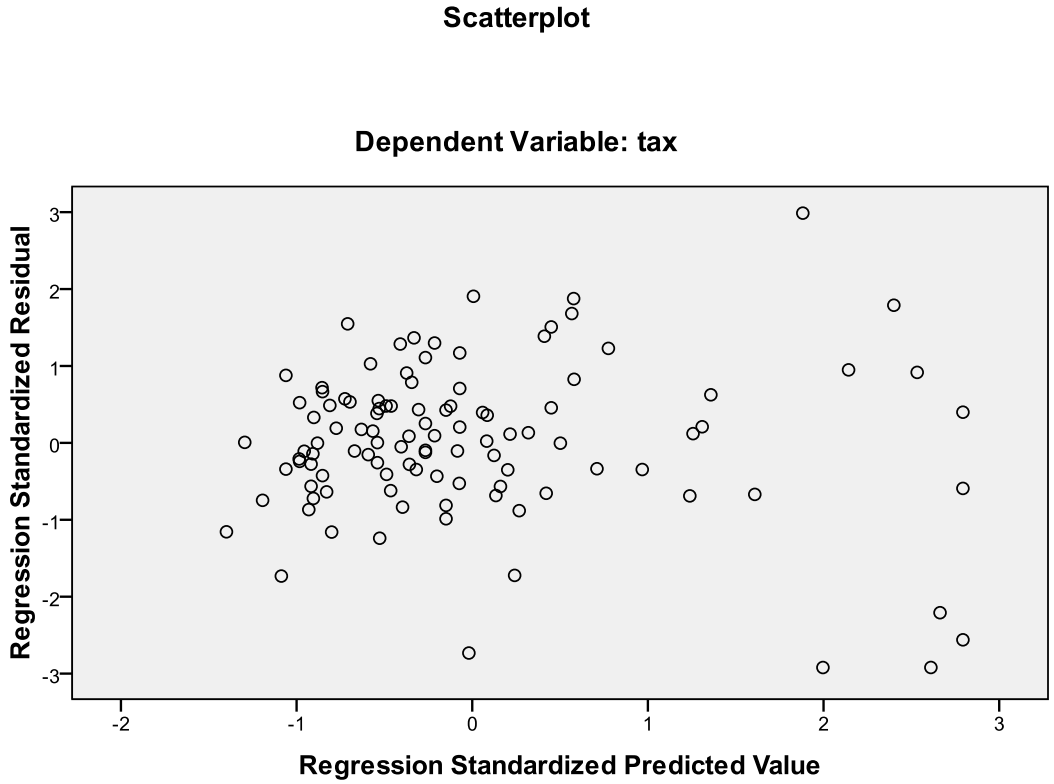
c. Dependent Variable: tax

d. Linear Regression through the Origin

31 Δείκτης Durbin-Watson για το μοντέλο του φόρου και της αξίας.

Η τιμή του δείκτη Durbin-Watson είναι κοντά στο 2 πράγμα που μας κάνει να αποφανθούμε την μη αυτοσυσχέτιση των καταλοίπων δηλαδή την ανεξαρτησία τους.

Μπορούμε επίσης να πάρουμε και το διάγραμμα διασπορών των καταλοίπων της παλινδρόμησης:



32 Διάγραμμα διασπορών των καταλοίπων της παλινδρόμησης του φόρου με την αξία.

Βλέπουμε και από εδώ ένα νέφος σημείων που υποστηρίζει ακόμα περισσότερο την τυχαιότητα και κατ' επέκταση την ανεξαρτησία μεταξύ των καταλοίπων της παλινδρόμησης.

Για την ομοσκεδαστικότητα των καταλοίπων θα κάνουμε τον έλεγχο της ισότητας διακυμάνσεων χωρίζοντας τις τιμές σε ομάδες. Και πάλι θα φτιάξουμε μία ομάδα για κάθε τιμή της μεταβλητής feats επομένως 9. Τα αποτελέσματα:

Test of Homogeneity of Variances

residuals

Levene Statistic	df1	df2	Sig.
1,245	6	98	,290

33 Έλεγχος ισότητας/ομοσκεδαστικότητας των διακυμάνσεων των καταλοίπων του μοντέλου του φόρου με την αξία.

Το **Sig.** είναι υψηλότερο του επιπέδου σημαντικότητας (που είναι 0,05) οπότε δεν απορρίπτουμε την υπόθεση περί ισότητας/ομοσκεδαστικότητας των διακυμάνσεων μεταξύ των ομάδων των καταλοίπων.

ANOVA

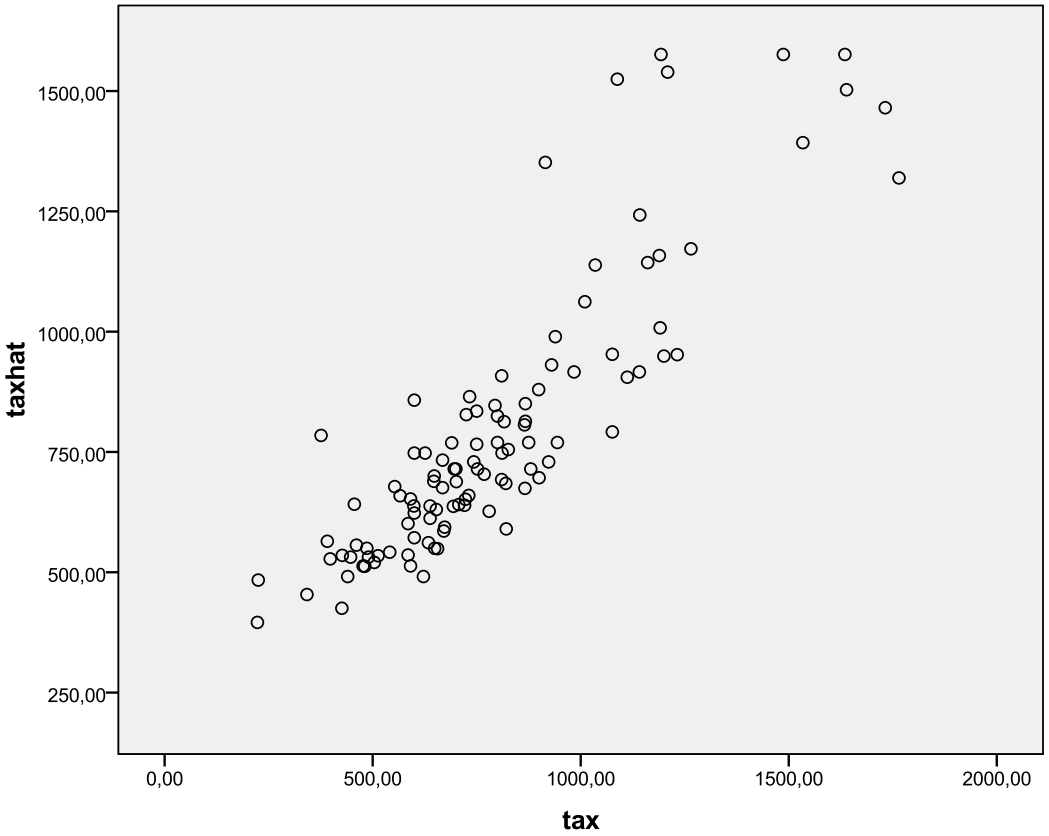
residuals

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	245719,372	8	30714,921	1,422	,197
Within Groups	2116371,789	98	21595,630		
Total	2362091,160	106			

34 Πίνακας ανάλυσης διακύμανσης των καταλοίπων του μοντέλου του φόρου με την αξία.

Ο παραπάνω πίνακας μας δίνει το αποτέλεσμα του ελέγχου της ισότητας των μέσων μεταξύ των ομάδων των καταλοίπων. Με **Sig.** ίσο με 0,197 δεν απορρίπτουμε την υπόθεση της ισότητας των μέσων.

Η τελευταία προϋπόθεση που θα μας απασχολήσει είναι αυτή της γραμμικότητας των μεταβλητών του μοντέλου δηλαδή του φόρου και της αξίας των σπιτιών. Θα ξεκινήσουμε με ένα διάγραμμα διασπορών (scatterplot) το οποίο φαίνεται παρακάτω:



35 Διάγραμμα διασπορών του φόρου και της τιμής του φόρου που προβλέπει το μοντέλο.

Παρατηρούμε αρκετά καλή συσχέτιση των δύο μεταβλητών μας που προσεγγίζει τη γραμμικότητα.