ΠΜΣ ΕΠΑΓΓΕΛΜΑΤΙΚΗ ΚΑΙ ΠΕΡΙΒΑΛΛΟΝΤΙΚΗ ΥΓΕΙΑ, ΔΙΑΧΕΙΡΙΣΗ ΚΑΙ ΟΙΚΟΝΟΜΙΚΗ ΑΠΟΤΙΜΗΣΗ

ΑΚ. ΕΤΟΣ 2006-2007, 3ο εξάμηνο

ΒΙΟΣΤΑΤΙΣΤΙΚΗ ΙΙ

ΜΑΘΗΜΑ 12

ΕΡΓΑΣΤΗΡΙΟ 4

ΑΠΛΗ ΓΡΑΜΜΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ ΜΕ ΤΗΝ ΧΡΗΣΗ SPSS

1

Στόχος μαθήματος:

Πως μπορούμε να συσχετίσουμε μία συνεχή μεταβλητή απόκρισης (Y) που ακολουθεί την κανονική κατανομή, με μία (ή περισσότερες) επεξηγηματική μεταβλητή (X) (είτε συνεχή είτε κατηγορική)

Στο συγκεκριμένο παράδειγμα που ακολουθεί θα ασχοληθούμε με την απλή παλινδρόμηση (δηλαδή μία μόνο X), όταν η επεξηγηματική μεταβλητή είναι συνεχής.

2

1. Απλή γραμμική παλινδρόμηση

Παράδειγμα 6: Χρόνος παράδοσης φορτίου

Ο υπεύθυνος των logistics μιας εταιρείας, ενδιαφέρεται να διερευνήσει τη σχέση του χρόνου παράδοσης (άρα και το αντίστοιχο κόστος) φορτίων με την απόσταση μεταξύ της αποθήκης και του τόπου προορισμού .

Για το λόγο αυτό πήρε ένα τυχαίο δείγμα 10 φορτωτικών και κατέγραψε την απόσταση σε μίλια και τον αριθμό των ημερών που χρειάστηκε για να παραδοθούν.

Ερώτηση: Μπορεί να υπολογιστεί ένα μοντέλο που θα βοηθήσει τον υπεύθυνο της εταιρείας να ποσοτικοποιήσει τη σχέση αυτή;

Ακολουθούν οι μετρήσεις...

3

1.2 Παράδειγμα 6 (συνέχεια)

Φορτωτική	1	2	3	4	5	6	7	8	9	10
Απόσταση σε μίλια	825	214	1070	550	480	92	1350	325	670	1215
Χρόνος παράδοσης σε ημέρες	3.5	1.0	4.0	2.0	1.0	3.0	4.5	1.5	3.0	5.0

4

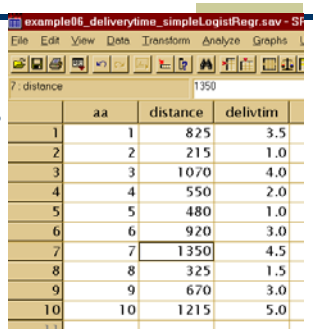
1.3 Δεδομένα

- Μονάδα μελέτης: φορτωτική
- Μέγεθος δείγματος: n=10 φορτωτικές
- Χαρακτηριστικά (μεταβλητές):
  - ✓ Κωδικός (ή αύξων) αριθμός φορτωτικής
  - ✓ Απόσταση (μίλια)
  - ✓ Χρόνος παράδοσης (μέρες)
- Ποια είναι X & ποια Y;

5

1.4 Εισαγωγή δεδομένων στο SPSS

- Αριθμός μεταβλητών: 3 ή 2?
- Κωδικοποίηση έχουμε?
- Φτιάξιμο μεταβλητών (όνομα, «ετικέτα», επεξήγηση κατηγοριών)
- Εισαγωγή στοιχείων



	aa	distance	delivtim
1	1	825	3.5
2	2	215	1.0
3	3	1070	4.0
4	4	550	2.0
5	5	480	1.0
6	6	920	3.0
7	7	1350	4.5
8	8	325	1.5
9	9	670	3.0
10	10	1215	5.0

6

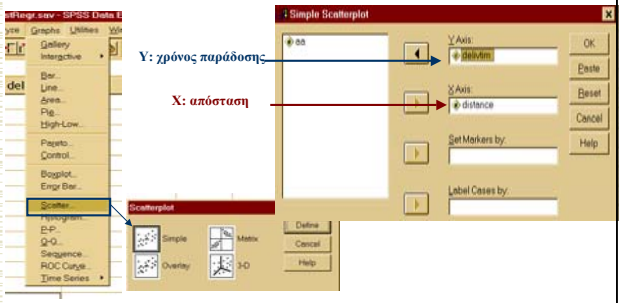
1.5 Ανάλυση

- Διαγραμματική απεικόνιση (Scatter-plot)
- Δείκτες συσχέτισης
- Μοντέλο Παλινδρόμησης
- Έλεγχος Προϋποθέσεων (Ανάλυση καταλοίπων)

7

1.5.1 Διαγραμματική απεικόνιση

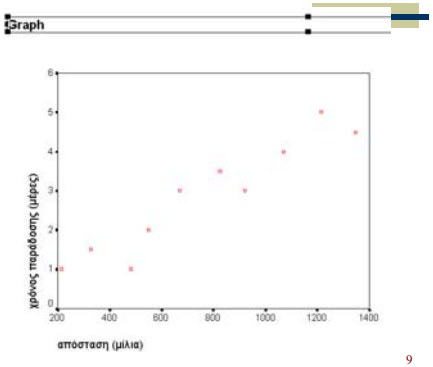
- Graphs>Scatter: simple



8

1.5.1 Διαγραμματική απεικόνιση

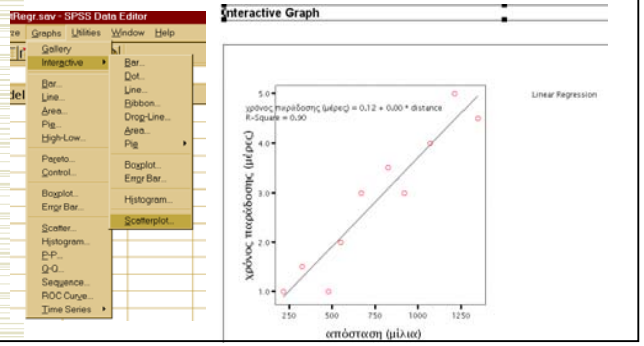
Διαφαίνεται *γραμμική σχέση* μεταξύ της απόστασης και του χρόνου παράδοσης. *Πως μπορεί να ποσοτικοποιηθεί η σχέση αυτή?*



9

1.5.2 Διαγραμματική απεικόνιση & γραμμή παλινδρόμησης

- Graphs>Interactive>Scatterplot

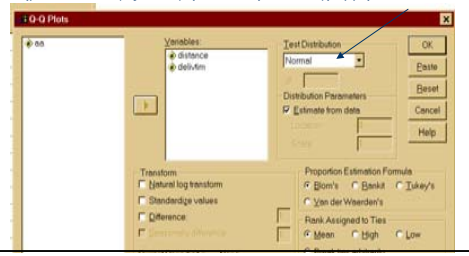


5

1.5.3 Σύντομοι προκαταρκτικοί έλεγχοι: α/Ελεγχος κανονικότητας: QQ Plots

Graphs>QQ

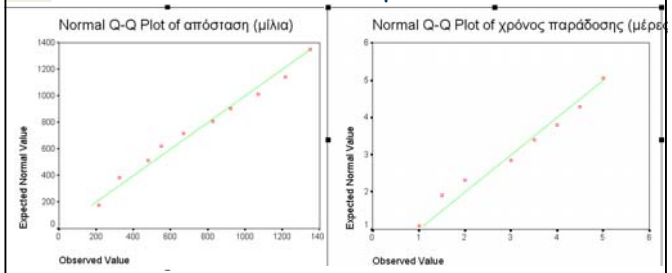
Οι παρατηρούμενες τιμές της μεταβλητής που δίνουμε απεικονίζονται διαγραμματικά σε σχέση με τις αναμενόμενες τιμές αν το δείγμα προερχόταν από την κανονική κατανομή. Αν το δείγμα προέρχεται από κανονική κατανομή τότε τα σημεία θα συνοψίζονται γύρω από την ευθεία γραμμή.



11

1.5.3 Σύντομοι προκαταρκτικοί έλεγχοι: α/κανονικότητα: QQ Plots (συνέχεια)

Αποτελέσματα

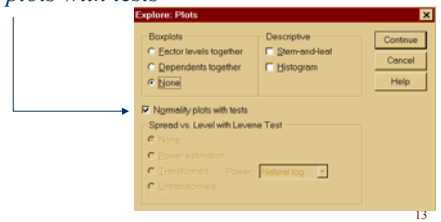


12

1.5.3 Σύντομοι προκαταρκτικοί έλεγχοι: β/κανονικότητα: έλεγχοι υποθέσεων

Εναλλακτικά ακολουθούμε από το μενού:

Analyze> Descriptive Statistics> Explore, και εν συνεχεία από την δυνατότητα «Plots» επιλέγουμε "*Normality plots with tests*"



13

1.5.3 Σύντομοι προκαταρκτικοί έλεγχοι: β/κανονικότητα: έλεγχοι υποθέσεων (συνέχεια)

Εκτός από τα QQ Plots που ξανα-εμφανίζονται και με αυτό τον τρόπο, έχουμε επιπλέον τους ελέγχους υποθέσεων κανονικότητας των Kolmogorov-Smirnov (με correction) & Shapiro - Wilk

Για μεγάλα δείγματα

Για μικρά δείγματα

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
DISTANCE απόσταση (μίλια)	.112	10	.200*	.973	10	.911
DELIVTIM χρόνος παράδοσης (μέρες)	.142	10	.200*	.941	10	.540

*. This is a lower bound of the true significance.

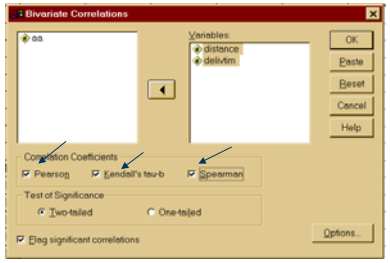
a. Lilliefors Significance Correction

Δεν απορρίπτουμε την υπόθεση της ακολουθίας κανονικής κατανομής για καμία από τις δύο μεταβλητές

14

1.5.3 Σύντομοι προκαταρκτικοί έλεγχοι: γ/συσχέτιση: δείκτης γραμμικής συσχέτισης Pearson

- Analyze> Correlate> Bivariate



15

1.5.3 Σύντομοι προκαταρκτικοί έλεγχοι: γ/συσχέτιση: δείκτης γραμμικής συσχέτισης Pearson (συνέχεια)

Correlations

		DISTANCE απόσταση (μίλια)	DELIVTIM χρόνος παράδοσης (μέρες)
DISTANCE απόσταση (μίλια)	Pearson Correlation	1.000	.949**
	Sig. (2-tailed)	.	.000
	N	10	10
DELIVTIM χρόνος παράδοσης (μέρες)	Pearson Correlation	.949**	1.000
	Sig. (2-tailed)	.000	.
	N	10	10

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

= r

r=0.949, δηλαδή ο συντελεστής συσχέτισης είναι πολύ υψηλός (όσο πιο μεγάλος ο δείκτης αυτός, τόσο ισχυρότερη είναι η συσχέτιση των δύο μεταβλητών (θετική ή αρνητική)

Στατιστικά σημαντικός ο έλεγχος ➔ *απορρίπτουμε την* H_0*:*$\rho=0$, δηλαδή παρατηρείται *ισχυρή (θετική) γραμμική συσχέτιση* μεταξύ των δύο μεταβλητών

16

1.5.3 Σύντομοι προκαταρκτικοί έλεγχοι: γ/συσχέτιση: μη παραμετρικοί δείκτες γραμμικής συσχέτισης

Correlations

			DISTANCE απόσταση (μίλια)	DELIVTIM χρόνος παράδοσης (μέρες)
Kendall's tau_b	DISTANCE απόσταση (μίλια)	Correlation Coefficient	1.000	.841**
		Sig. (2-tailed)	.	.001
		N	10	10
	DELIVTIM χρόνος παράδοσης (μέρες)	Correlation Coefficient	.841**	1.000
		Sig. (2-tailed)	.001	.
		N	10	10
Spearman's rho	DISTANCE απόσταση (μίλια)	Correlation Coefficient	1.000	.945**
		Sig. (2-tailed)	.	.000
		N	10	10
	DELIVTIM χρόνος παράδοσης (μέρες)	Correlation Coefficient	.945**	1.000
		Sig. (2-tailed)	.000	.
		N	10	10

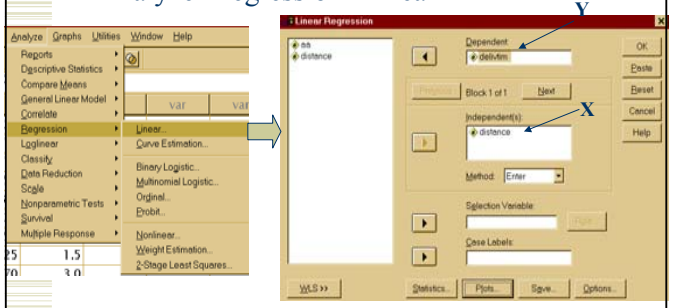
** Correlation is significant at the .01 level (2-tailed).

- Υψηλές τιμές δεικτών
- Απορρίπτουμε την υπόθεση της μη-συσχέτισης→ συνάγουμε υψηλή γραμμική (θετική) συσχέτιση

17

1.5.4 Απλή γραμμική παλινδρόμηση

- Analyze>Regression>Linear



18

1.5.4 Απλή γραμμική παλινδρόμηση: συνοπτικός πίνακας μοντέλου

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.949[a]	.900	.888	.4800

a. Predictors: (Constant), distance

R=Multiple Correlation Coefficient

R^2 (coefficient of determination)= % διακύμανσης της Y που εξηγείται από το μοντέλο

R_{adj}^2= % διακύμανσης της Y που εξηγείται από το μοντέλο διορθωμένο για τον αριθμό των μεταβλητών

Ο τελευταίος αυτός δείκτης:

✓Λαμβάνει υπόψη του τις μεταβλητές

✓Χρησιμοποιείται ως μέτρο καλής προσαρμογής ή πρόβλεψης

✓ΜΠΟΡΕΙ να χρησιμοποιηθεί ως κριτήριο επιλογής μοντέλου (ΓΕΝΙΚΑ)

✓Στην απλής γραμμικής παλινδρόμησης δε διαφέρει πολύ από το R^2.

19

1.5.4 Απλή γραμμική παλινδρόμηση: συνοπτικός πίνακας μοντέλου

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.949[a]	.900	.888	.4800

a. Predictors: (Constant), distance

R^2 (coefficient of determination)= 0.90, δηλαδή η 'απόσταση' εξηγεί το 90% της συνολικής διακύμανσης των 'ημερών παράδοσης'. Το υπόλοιπο 10% της διακύμανσης είναι ανεξήγητο και πρέπει να οφείλεται σε άλλους παράγοντες που δεν λαμβάνονται υπ' όψη στην παρούσα μελέτη

20

1.5.4 Απλή γραμμική παλινδρόμηση: συνοπτικός πίνακας μοντέλου (συνέχεια)

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.949[a]	.900	.888	.4800

a. Predictors: (Constant), distance

Στο συγκεκριμένο παράδειγμα παρατηρούμε πολύ καλή προσαρμογή του μοντέλου (σταθερά + απόσταση(X)), λόγω του υψηλού R_{adj}^2

1.5.4 Απλή γραμμική παλινδρόμηση: πίνακας ανάλυσης διακύμανσης

Στο συγκεκριμένο πίνακα (απλή παλινδρόμηση) ελέγχουμε την υπόθεση: **H_0: $\beta_1=0$ έναντι της εναλλακτικής H_1: $\beta_1 \neq 0$**, δηλαδή ελέγχουμε *αν το τρέχων μοντέλο διαφέρει από το σταθερό* (δηλαδή το μοντέλο $y=\beta_0+\varepsilon$)

Στο παράδειγμά μας → **απορρίπτουμε** την **H_0: $\beta_1=0$,** γεγονός που σημαίνει ότι η *επίδραση της ανεξάρτητης μεταβλητής είναι σημαντική* και επηρεάζει / καθορίζει τις τιμές της εξαρτημένης

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	16.682	1	16.682	72.396	.000[a]
	Residual	1.843	8	.230		
	Total	18.525	9			

a. Predictors: (Constant), DISTANCE απόσταση (μίλια)

b. Dependent Variable: DELIVTIM χρόνος παράδοσης (μέρες)

1.5.4 Απλή γραμμική παλινδρόμηση: πίνακας εκτίμησης παραμέτρων μοντέλου

Από τον παρακάτω πίνακα προκύπτει η εκτιμώμενη γραμμή παλινδρόμησης:

ΗΜΕΡΕΣ ΠΑΡΑΔΟΣΗΣ = 0.118 + 0.00359 ΜΙΛΙΑ + ε, ε~NORMAL(0, 0.48^2)

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	.118	.355		.333	.748
	DISTANCE απόσταση (μίλια)	3.59E-03	.000	.949	8.509	.000

a. Dependent Variable: DELIVTIM χρόνος παράδοσης (μέρες)

1.5.4 Απλή γραμμική παλινδρόμηση: πίνακας εκτίμησης παραμέτρων μοντέλου (συνέχεια)

Από την τιμή του p-value που αναφέρεται στον έλεγχο της υπόθεσης H_0: $\beta_1=0$, συνάγουμε ότι απορρίπτουμε την H_0, επομένως η απόσταση είναι *στατιστικά σημαντικός παράγοντας για την εκτίμηση των ημερών παράδοσης* (έχουν σημαντική, γραμμικά θετική, σχέση - αφού $\beta_1>0$)

Για το αλλο p-value της σταθεράς σχόλιο?????

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	.118	.355		.333	.748
	DISTANCE απόσταση (μίλια)	3.59E-03	.000	.949	8.509	.000

a. Dependent Variable: DELIVTIM χρόνος παράδοσης (μέρες)

1.5.4 Απλή γραμμική παλινδρόμηση: πίνακας εκτίμησης παραμέτρων μοντέλου (συνέχεια)

Αν θελήσουμε να ποσοτικοποιήσουμε την σχέση των ημερών παράδοσης και μιλίων μπορούμε να πούμε το εξής:

Οι αναμενόμενες ημέρες παράδοσης αυξάνονται κατά 0.0036 μέρες για κάθε αύξηση της απόστασης κατά 1 μίλι, ή

Με κάθε επιπλέον 100 μίλια, ο αναμενόμενος χρόνος παράδοσης αυξάνεται κατά 0.36 μέρες (περίπου 8.6 ώρες)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	.118	.355		.333	.748
	DISTANCE απόσταση (μίλια)	3.59E-03	.000	.949	8.509	.000

a. Dependent Variable: DELIVTIM χρόνος παράδοσης (μέρες)

25

1.5.4 Απλή γραμμική παλινδρόμηση: πίνακας εκτίμησης παραμέτρων μοντέλου (συνέχεια)

Επιπλέον, αν θέσουμε την τιμή 0 για την απόσταση στην εξίσωση που εκτιμήσαμε, τότε έχουμε αναμενόμενη τιμή για την παράδοση: 0.118 ημέρες (περ 3 ώρες). Αυτό δεν μπορεί να ισχύει με βάση τη λογική. Η εξίσωση παλινδρόμησης προκύπτει από τις παρατηρούμενες τιμές των δύο μεταβλητών και ισχύει μόνο μεταξύ του εύρους των τιμών που παρατηρήθηκαν. Το να συνάγουμε συμπεράσματα για τιμές πέραν αυτού του εύρους, μπορεί να αποβεί παραπλανητικό.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	.118	.355		.333	.748
	DISTANCE απόσταση (μίλια)	3.59E-03	.000	.949	8.509	.000

a. Dependent Variable: DELIVTIM χρόνος παράδοσης (μέρες)

26

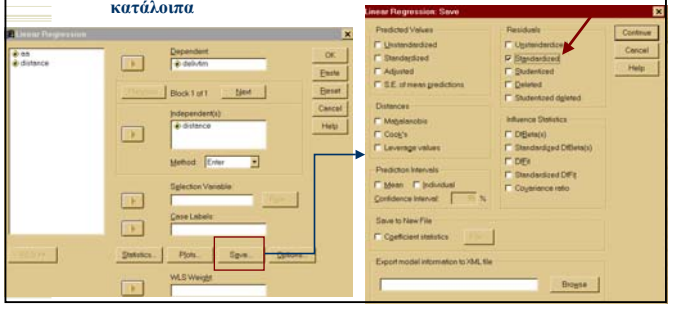
1.5.5 Προϋποθέσεις μοντέλου

- Κανονικότητα σφαλμάτων
- Ανεξαρτησία σφαλμάτων
- Ομοσκεδαστικότητα σφαλμάτων
- Γραμμικότητα X και Y: για οποιαδήποτε τιμή του x, η αντίστοιχη τιμή του y έχει αναμενόμενη τιμή $\alpha+\beta x$, που είναι γραμμική συνάρτηση του x

27

1.5.6 Έλεγχοι προϋποθέσεων μοντέλου: α) κανονικότητα

Προκειμένου να ελέγξουμε την κανονικότητα των καταλοίπων ακολουθούμε τα εξής: **1. Αποθηκεύουμε τα τυποποιημένα κατάλοιπα**



1.5.6 Έλεγχοι προϋποθέσεων μοντέλου: α) κανονικότητα: 1. αποθήκευση καταλοίπων

Το πρόγραμμα υπολογίζει και αποθηκεύει με ένα δικό του όνομα (και label) την νέα μεταβλητή

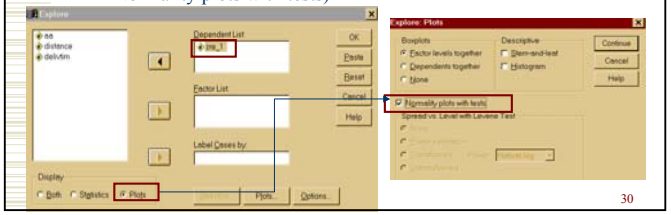
aa	distance	delivtim	zre_1
1	825	3.5	.88358
2	215	1.0	.23138
3	1070	4.0	.09537
4	550	2.0	-.18739
5	480	1.0	-1.74782
6	920	3.0	-.86756
7	1350	4.5	-.95424
8	325	1.5	.45144
9	670	3.0	.99960
10	1215	5.0	1.09564

29

1.5.6 Έλεγχοι προϋποθέσεων μοντέλου: α) κανονικότητα: 2.Έλεγχος Shapiro-Wilk

Προκειμένου να ελέγξουμε την κανονικότητα των καταλοίπων ακολουθούμε τα εξής:

2. **Πραγματοποιούμε έλεγχο της κανονικότητάς τους με Shapiro-Wilk** (Analyze> Descriptive Statistics> Explore|Plots: Normality plots with tests)



30

1.5.6 Έλεγχοι προϋποθέσεων μοντέλου: α) κανονικότητα: 2.Έλεγχος Shapiro-Wilk (συνέχεια)

Tests of Normality

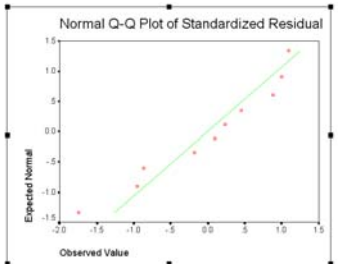
	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
ZRE_1 Standardized Residual	.140	10	.200*	.937	10	.491

*. This is a lower bound of the true significance.
a. Lilliefors Significance Correction

Δεν απορρίπτουμε υπόθεση κανονικότητας

31

- ΝΑ ΤΟ ΒΑΛΩ? ΠΑΡΑΤΗΡΗΣΕΙΣ?

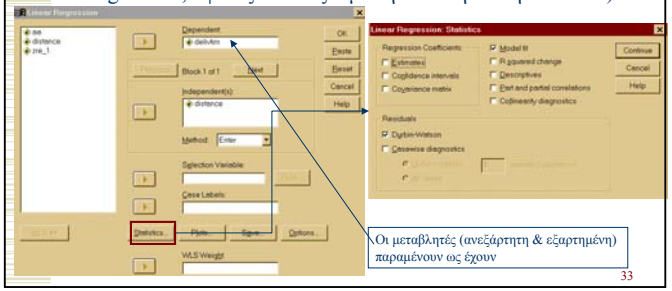


32

1.5.6 Έλεγχοι προϋποθέσεων μοντέλου: β) Ανεξαρτησία καταλοίπων

Δεν είναι εύκολα ελέγξιμη

Υπολογίζουμε τον δείκτη **Durbin-Watson** (από επιλογή linear regression, αφού ξε-επιλέξουμε την αποθήκευση residuals)



33

1.5.6 Έλεγχοι προϋποθέσεων μοντέλου: β) Ανεξαρτησία καταλοίπων (συνέχεια)

Αν D-W =2 ⇔ δεν υπάρχει «αυτοσυσχέτιση» μεταξύ των καταλοίπων

Στην προκειμένη περίπτωση η τιμή είναι πολύ χαμηλή και πιθανόν να υπάρχει συσχέτιση

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.949[a]	.900	.888	.480	.753

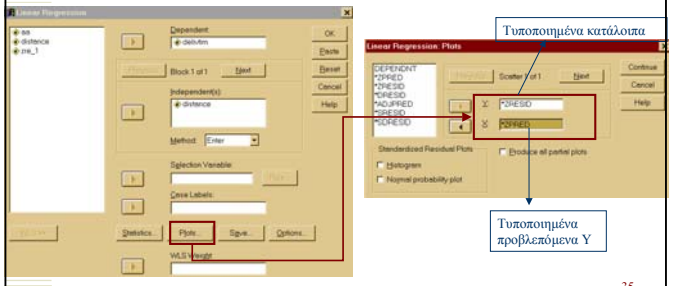
a. Predictors: (Constant), DISTANCE απόσταση (μίλια)

b. Dependent Variable: DELIVTIM χρόνος παράδοσης (μέρες)

34

1.5.6 Έλεγχοι προϋποθέσεων μοντέλου: γ) Ομοσκεδαστικότητα καταλοίπων

- ΔΙΑΓΡΑΜΜΑ RESIDUALS ΑΝΑ $\hat{Y}$



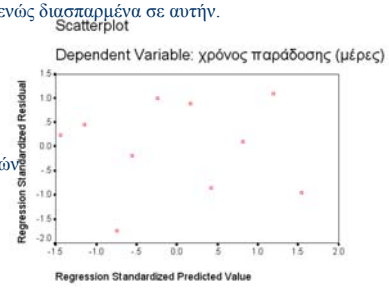
35

1.5.6 Έλεγχοι προϋποθέσεων μοντέλου: γ) Ομοσκεδαστικότητα καταλοίπων (συνέχεια)

Επιθυμητό: τα σημεία να βρίσκονται σε μία οριζόντια ζώνη μεταξύ (-2, 2), και να είναι ομοιογενώς διασπαρμένα σε αυτήν.

Έτσι ελέγχονται:

1. Γραμμικότητα
2. Ομοσκεδαστικότητα
3. Ύπαρξη ακραίων τιμών



36