

ΑΝΑΛΥΣΗ ΠΑΛΙΝΔΡΟΜΗΣΗΣ, 11-4-02

Άσκηση 3. Δίνονται τα παρακάτω δεδομένα

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
X	13	18	15	4	20	2	15	11	5	1	9	2	12	11	17
Y	12.0	5.0	2.4	17.1	2.7	21.7	6.9	2.4	14.3	52.6	12.3	26.6	10.2	8.4	2.2

α) Να γίνει το διάγραμμα διασποράς

β) εξετάστε τα μοντέλα

$$Y = \beta_0 + \beta_1 X + \varepsilon, \text{ (linear), } Y = \beta_0 + \beta_1 \ln X + \varepsilon, \text{ (logarithmic),}$$

$$Y = \beta_0 + \beta_1/X + \varepsilon, \text{ (inverse), } \ln Y = \beta_0 + \beta_1 X + \varepsilon.$$

με $\varepsilon \sim N(0, \sigma^2)$.

γ. Ποιες οι εκτιμήσεις για τις άγνωστες παραμέτρους β_0, β_1 σε κάθε μοντέλο;

δ. Ποιο μοντέλο έχει τον μεγαλύτερο συντελεστή προσδιορισμού;

Άσκηση 4. Πριν αντικατασταθούν μεταλλικοί σωλήνες από πλαστικούς ή από άλλα συνθετικά υλικά, μελετήθηκε η αντοχή 20 πειραματικών τεμαχίων από πολυεστέρα. Έγινε έλξη του υλικού με ταχύτητα X (cm/min) και μετρήθηκε η αντοχή Y σε κιλά ανά τετραγωνικό εκατοστό (kgf/cm²). Οι μετρήσεις έδωσαν:

X	Y	X	Y
0.25	5520	10	7100
0.25	5390	10	7150
0.25	5730	10	7260
0.25	4940	10	7650
0.25	5810	10	8210
1	6840	10	7960
1	5720	20	7950
1	6120	20	7470
1	6400	20	7720
1	6240	20	8460

Δίνονται τα δύο μοντέλα:

i) $Y = \beta_0 + \beta_1 X + \varepsilon$, ii) $Y = \beta_0 + \beta_1 \ln X + \varepsilon$

Πραγματοποιήστε έλεγχο καλής προσαρμογής του μοντέλου (lack-of-fit) σε στάθμη $\alpha=0.05$ και στις δύο περιπτώσεις.

α) Στο μοντέλο που θα γίνει δεκτό, βρείτε 95% δ.ε. για τη μέση και ατομική πρόβλεψη στο σημείο $X=5$.

β) Ελέγξτε την υπόθεση ($\alpha=0.05$) $H_0: \beta_1=0$, $H_1: \beta_1 \neq 0$, ή $H_1: \beta_1 > 0$.

γ) Ελέγξτε την υπόθεση ($\alpha=0.05$) $H_0: \beta_1=2.5$, $H_1: \beta_1 \neq 2.5$, ή $H_1: \beta_1 < 2.5$.

δ) Υπάρχουν έκτροπες παρατηρήσεις στα κατάλοιπα;

Άσκηση 5. Δημιουργήστε μία μεταβλητή e από 100 τυχαίους αριθμούς από $N(0,3)$.

Δημιουργήστε μία μεταβλητή X από 100 τυχαίους αριθμούς από $U(-5,5)$.

Δημιουργήστε τη μεταβλητή $Y = 2.4 - 4.1X + e$

α) Εξετάστε το γραμμικό μοντέλο $Y = \beta_0 + \beta_1 X + \varepsilon$ και ελέγξτε τις υποθέσεις $H_0: \beta_1 = -4.1$, $H_1: \beta_1 \neq -4.1$, $H_0: \beta_0 = 2.4$, $H_1: \beta_0 > 2.4$ σε ε.σ. $\alpha = 0.05$.

β) Εξετάστε αν τα κατάλοιπα ακολουθούν κανονική κατανομή (μέσω ενός Q-Q, P-P plot, K-S τεστ)

γ) Αν κατά λάθος γράψουμε $10y_5$ αντί y_5 , πώς αλλάζουν οι εκτιμήσεις των παραμέτρων; Από τον έλεγχο των καταλοίπων μπορούμε να εντοπίσουμε το λάθος μας;

δ) Δημιουργήστε ένα αρχείο με τις παρατηρήσεις $Z = 2.4 - 4.1X + X e$. Αν θεωρήσουμε το μοντέλο $Z = \beta_0 + \beta_1 X + \varepsilon$, εξετάστε τα κατάλοιπα. Είναι το μοντέλο αποδεκτό; Πως θα βελτιωθεί;

Άσκηση 6. Για τη μελέτη μιας ασθένειας της καρδιάς στα μικρά παιδιά, μετρήθηκε η ηλικία X σε μήνες και ένας δείκτης Y για 21 παιδιά. Τα δεδομένα ήταν:

α/α	X	Y
1	15	95
2	26	71
3	10	83
4	9	91
5	15	102
6	20	87
7	18	93
8	11	100
9	8	104
10	20	94
11	7	113

α/α	X	Y
12	9	96
13	10	83
14	11	84
15	11	102
16	10	100
17	12	105
18	42	57
19	17	121
20	11	86
21	10	100

- α) Να γίνει η γραφική παράσταση (X, Y) και να εντοπιστούν οι έκτροπες παρατηρήσεις.
 β) Εφαρμόστε το απλό γραμμικό μοντέλο και υπολογίστε το leverage p_{ii} για όλες τις παρατηρήσεις. Ποιες παρατηρήσεις επηρεάζουν περισσότερο το μοντέλο ($p_{ii} > 3 \cdot 2/n$) και ποιες θεωρούνται ασυνήθιστες ($|\hat{e}_i^*| > 2$);
 γ) Να γίνει έλεγχος κανονικότητας (P-P, K-S) και έλεγχος ανεξαρτησίας (runs-test) των παρατηρήσεων.

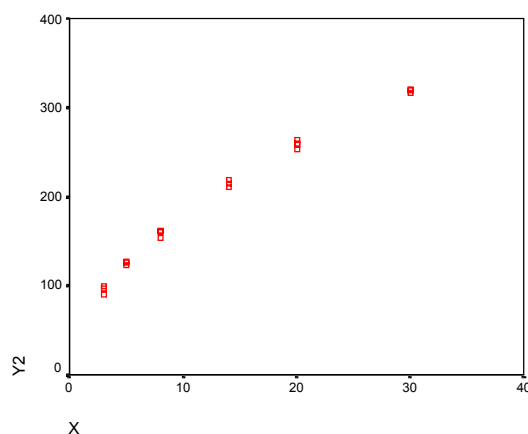
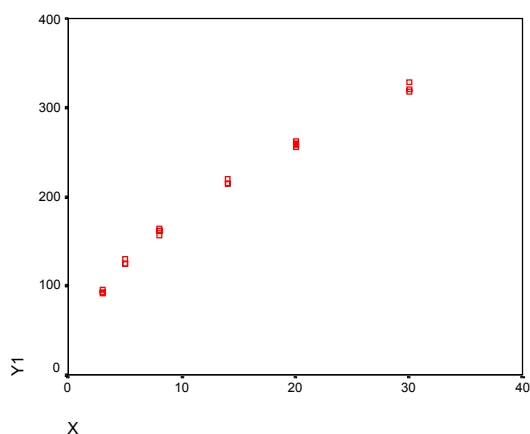
Άσκηση 7. Μία εταιρία έστειλε τμηματικά το προσωπικό του λογιστηρίου να εξασκηθεί σε δύο νέα λογιστικά πακέτα. Μετά το τέλος των σεμιναρίων, έγιναν εξετάσεις για να διαπιστωθεί ο βαθμός εκμάθησης των πακέτων. Ο παρακάτω πίνακας δίνει την βαθμολογία Y_1, Y_2 και τη διάρκεια X των σεμιναρίων σε ώρες.

X	3	3	3	5	5	5	8	8	8	8
Y_1	92	96	93	124	125	130	162	161	156	164
Y_2	99	97	90	123	127	126	160	161	153	161
X	14	14	14	20	20	20	20	30	30	30
Y_1	214	215	220	262	259	258	256	318	320	328
Y_2	215	211	219	253	263	260	258	319	320	316

- α) Να γίνουν τα διαγράμματα διασποράς $(X, Y_1), (X, Y_2)$
 β) Εξετάστε ποια από τα τρία μοντέλα γίνονται αποδεκτά:
 i) $Y = \beta_0 + \beta_1 X + \varepsilon$, ii) $Y = \beta_0 + \beta_1 \sqrt{X} + \varepsilon$, iii) $Y = \beta_0 + \beta_1 \ln X + \varepsilon$,
 Στα μοντέλα που γίνονται αποδεκτά (lof τεστ):
 γ) Εκτιμήστε τις παραμέτρους και τη διασπορά των σφαλμάτων.
 δ) Ελέγξτε την υπόθεση: $H_0: \beta_1 = 55$, $H_1: \beta_1 > 55$.
 ε) Κάνετε έλεγχο της κανονικότητας των καταλοίπων με P-P, Q-Q plots, K-S τεστ.
 στ) Ελέγξτε αν τα κατάλοιπα ακολουθούν κατανομή Laplace με Q-Q plot.
 ζ) Ελέγξτε αν τα κατάλοιπα ακολουθούν ομοιόμορφη κατανομή με Q-Q plot, K-S, χ^2 test.
 η) Δώστε 95% δ.ε. για τη μέση και ατομική πρόβλεψη της απόδοσης στα δύο σεμινάρια, αν η διάρκεια είναι 12 ώρες.
 θ) Εξετάστε με τα κατάλληλα διαγράμματα αν η διασπορά των σφαλμάτων εξαρτάται από τη διάρκεια ή από τη σειρά παρακολούθησης.
 ι) Εντοπίστε τις έκτροπες παρατηρήσεις που επηρεάζουν ή που είναι ασυνήθιστες.
 ια) Εξετάστε αν τα σφάλματα είναι ασυσχέτιστα, με διάγραμμα και με τεστ.

Απαντήσεις Άσκησης 7.

α) Χρησιμοποιώντας την επιλογή Graphs/scatterplot λαμβάνουμε:



Θα εξετάσουμε ξεχωριστά τη σχέση που υπάρχει μεταξύ της Y_1 και της X και μεταξύ της Y_2 και της X (παρατηρώντας ότι η σχέση μεταξύ των Y_1, X φαίνεται παρόμοια με τη σχέση μεταξύ των Y_2, X θα ήταν ίσως καλύτερο να εξετάσουμε τη σχέση μεταξύ της μέσης βαθμολογίας $Y = (Y_1 + Y_2)/2$ και του X)

β) Παρατηρούμε ότι στα δεδομένα μας υπάρχουν επαναλαμβανόμενες παρατηρήσεις οπότε μπορούμε να χρησιμοποιήσουμε το lack-of-fit τεστ για τον έλεγχο καλής προσαρμογής του μοντέλου.

ια) Για το μοντέλο $Y_1 = \beta_0 + \beta_1 X + \varepsilon$ ($R^2 = 0.977$) έχουμε ότι

$$F^* = \frac{(SSE - SSPE)/(r - 2)}{SSPE/(n - r)}, \quad SSPE = \sum_{j=1}^r SSE_j$$

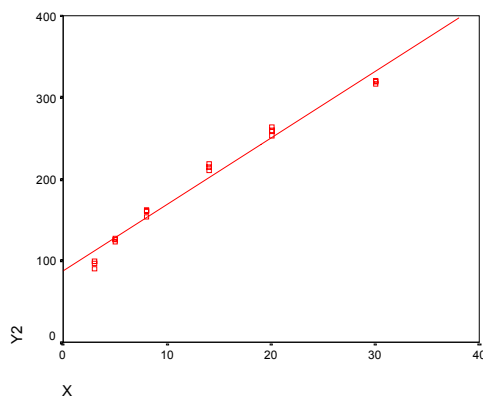
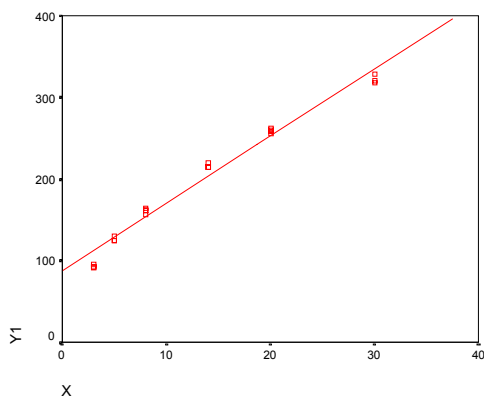
όπου $SSE_j = \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}_{j\cdot})^2$ είναι η δειγματική διασπορά του Y_1 για $X = X_j$ ($X_1=3, X_2=5, X_3=8, X_4=14, X_5=20, X_6=30$). Το F^* δεν δίνεται αυτόματα από το πακέτο οπότε θα πρέπει να το υπολογίσουμε. Ένας σύντομος τρόπος να βρούμε το SSPE είναι μέσω της διαδικασίας Analyze/compare means/one-way ANOVA με Dependent: Y_1 , Factor: X . Στον πίνακα ANOVA που προκύπτει ισχύει ότι $SSPE = \text{Within Groups Sum of Square} = 159.5$. Το SSE ως γνωστό βρίσκεται εκτελώντας παλινδρόμηση (βρίσκεται στον πίνακα ANOVA που προκύπτει από τη διαδικασία Analyze/Regression/Linear, Dependent: Y_1 , Independent: X) και είναι ίσο με 2616,276. Επομένως

$$F^* = \frac{(SSE - SSPE)/(r - 2)}{SSPE/(n - r)} = \frac{(2616.276 - 159.5)/4}{159.5/14} \approx 53.9 > F_{r-2, n-r; \alpha} = F_{4, 14; 0.05} \approx 3.11$$

και άρα **απορρίπτουμε** την H_0 : σωστό μοντέλο.

ιβ) Όμοια, για το μοντέλο $Y_2 = \beta_0 + \beta_1 X + \varepsilon$ ($R^2 = 0.977$) έχουμε ότι

$$F^* = \frac{(2546.37 - 191.75)/4}{191.75/14} \approx 42.98 > F_{4, 14; 0.05} \approx 3.11$$



και άρα **απορρίπτουμε** την H_0 : σωστό μοντέλο. (Τα αντίστοιχα γραφήματα έγιναν με τη χρήση της διαδικασίας Graphs/scatter/simple όπου με διπλό κλικ στο σχήμα που παράγεται ανοίγουμε τον SPSS Chart Editor και επιλέγουμε Chart/options/Fit Line: Total με Fit Options: Linear Regression)

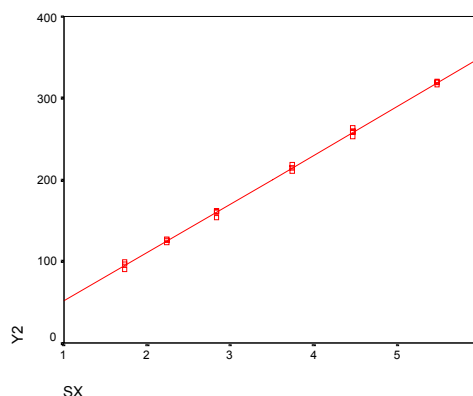
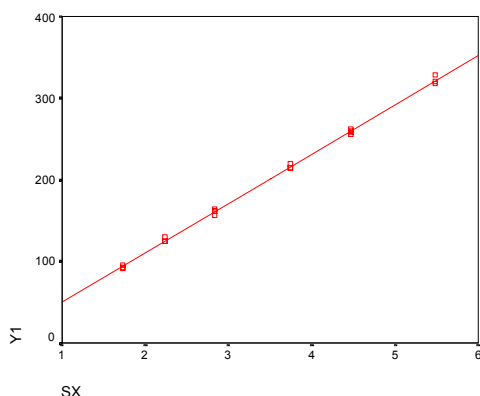
iiα) Για το μοντέλο $Y = \beta_0 + \beta_1 \sqrt{X} + \varepsilon$ ($R^2 = 0.998$) έχουμε ότι (το SSPE είναι ίδιο με αυτό του iα)

$$F^* = \frac{(178.9 - 159.5)/4}{159.5/14} \approx 0.425 < F_{4,14;0.05} \approx 3.11$$

και άρα **δεν μπορούμε να απορρίψουμε** την H_0 : σωστό μοντέλο.

iiβ) Για το μοντέλο $Y = \beta_0 + \beta_1 \sqrt{X} + \varepsilon$ ($R^2 = 0.998$) έχουμε ότι (το SSPE είναι ίδιο με αυτό του iβ)

$$F^* = \frac{(202.23 - 191.75)/4}{191.75/14} \approx 0.1913 < F_{4,14;0.05} \approx 3.11$$



και άρα **δεν μπορούμε να απορρίψουμε** την H_0 : σωστό μοντέλο.

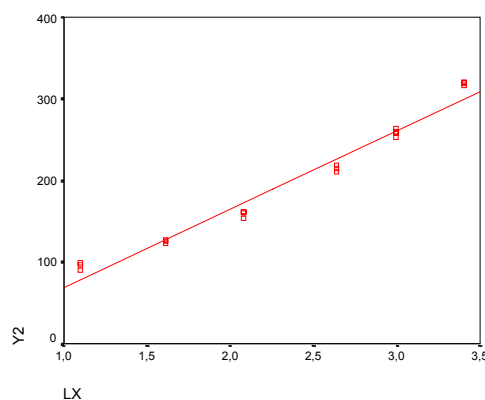
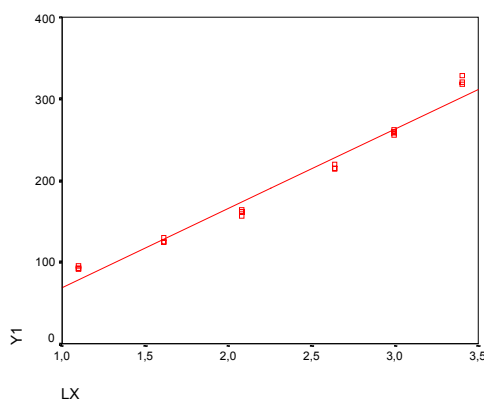
iiiα) Για το μοντέλο $Y = \beta_0 + \beta_1 \ln X + \varepsilon$ ($R^2 = 0.972$) έχουμε ότι (το SSPE είναι ίδιο με αυτό του iα)

$$F^* = \frac{(3199.1 - 159.5)/4}{159.5/14} \approx 66.7 > F_{4,14;0.05} \approx 3.11$$

και άρα **απορρίπτουμε** την H_0 : σωστό μοντέλο.

iiiβ) Για το μοντέλο $Y = \beta_0 + \beta_1 \ln X + \varepsilon$ ($R^2 = 0.971$) έχουμε ότι (το SSPE είναι ίδιο με αυτό του iβ)

$$F^* = \frac{(3295 - 191.75)/4}{191.75/14} \approx 56.64 > F_{4,14;0.05} \approx 3.11$$



και άρα **απορρίπτουμε** την H_0 : σωστό μοντέλο.

Συνεπώς αποδεκτό γίνεται μόνο το μοντέλο $Y = \beta_0 + \beta_1 \sqrt{X} + \varepsilon$, κάτι που είναι φανερό και από τα γραφήματα.

γ) Μοντέλο: $Y_1 = \beta_0 + \beta_1 \sqrt{X} + \varepsilon$. Εκτελούμε τη διαδικασία Analyze/regression/Linear (Y1, SX= $\sqrt{X}$) από όπου λαμβάνουμε:

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	-10,267	2,054		-5,000	,000	-14,582	-5,953
	SX	60,473	,561	,999	107,798	,000	59,295	61,652

a. Dependent Variable: Y1

ANOVA^b

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	115491,7	1	115491,654	11620,427	,000 ^a
	Residual	178,896	18	9,939		
	Total	115670,6	19			

a. Predictors: (Constant), SX

b. Dependent Variable: Y1

και άρα $\hat{\beta}_0 = -10.267$ με $s(\hat{\beta}_0) = 2,054$, $\hat{\beta}_1 = 60.473$ με $s(\hat{\beta}_1) = 0,561$, $\hat{\sigma}^2 = MSE = 9.939$.

Μοντέλο: $Y_2 = \beta_0 + \beta_1 \sqrt{X} + \varepsilon$. Αντίστοιχα λαμβάνουμε:

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	-8,676	2,183		-3,973	,001	-13,263	-4,089
	SX	59,691	,596	,999	100,076	,000	58,437	60,944

a. Dependent Variable: Y2

ANOVA^b

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	112520,7	1	112520,718	10015,107	,000 ^a
	Residual	202,232	18	11,235		
	Total	112722,9	19			

a. Predictors: (Constant), SX

b. Dependent Variable: Y2

και άρα $\hat{\beta}_0 = -8.676$ με $s(\hat{\beta}_0) = 2,183$, $\hat{\beta}_1 = 59.691$ με $s(\hat{\beta}_1) = 0.596$, $\hat{\sigma}^2 = MSE = 11.235$.

δ) Για να ελέγξουμε την υπόθεση $H_0: \beta_1 = 55$ έναντι της $H_1: \beta_1 > 55$ με τη βοήθεια του πακέτου θέτουμε $\gamma = \beta_1 - 55$ και το μοντέλο γίνεται

$$Y = \beta_0 + \beta_1 \sqrt{X} + \varepsilon \Leftrightarrow Y = \beta_0 + (\gamma + 55)\sqrt{X} + \varepsilon \Leftrightarrow Y - 55\sqrt{X} = \beta_0 + \gamma \sqrt{X} + \varepsilon \Leftrightarrow Y' = \beta_0 + \gamma \sqrt{X} + \varepsilon.$$

Στο παραπάνω γραμμικό μοντέλο αρκεί τώρα να ελέγξουμε αν $H_0: \gamma = 0$ έναντι της $H_1: \gamma > 0$. Από το Transform/Compute κατασκευάζουμε τη νέα μεταβλητή $Y'_1 = Y_1 - 55\sqrt{X}$ και εκτελώντας τη διαδικασία Analyze/Regression/Linear (Y'_1 , SX= $\sqrt{X}$) λαμβάνουμε:

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-10,267	2,054		-5,000	,000
	SX	5,473	,561	,917	9,757	,000

a. Dependent Variable: Y1PRIME

Επειδή το $T^* = 5.473$ είναι θετικό, το p -value του μονόπλευρου ελέγχου $H_0: \gamma = 0 \Leftrightarrow \beta_1 = 55$ με $H_1: \gamma > 0 \Leftrightarrow \beta_1 > 55$ είναι

$$p\text{-value} = P(T > T^* | T \sim t_{n-2}) = \frac{1}{2} P(|T| > T^* | T \sim t_{n-2}) = \frac{1}{2} \text{Sig} \approx \frac{1}{2} 0.000 \approx 0$$

και άρα απορρίπτουμε την H_0 έναντι της H_1 (δηλ. $\beta_1 > 55$)

Όμοια, για το δεύτερο σεμινάριο $Y_2 = \beta_0 + \beta_1 \sqrt{X} + \varepsilon$, λαμβάνουμε τον πίνακα ($Y'_2 = Y_2 - 55\sqrt{X}$)

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-8,676	2,183		-3,973	,001
	SX	4,691	,596	,880	7,864	,000

a. Dependent Variable: Y2PRIME

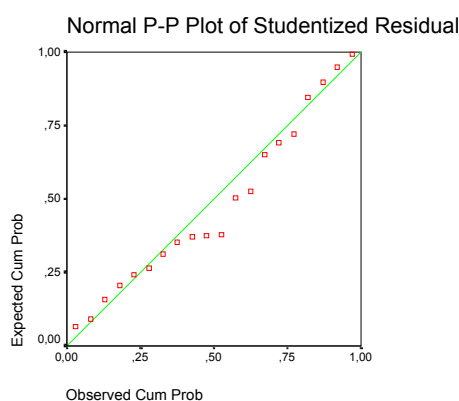
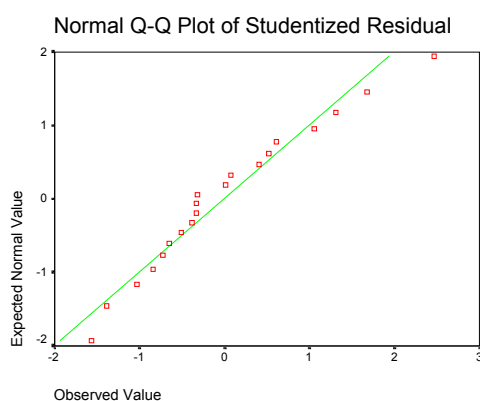
Επειδή και εδώ $T^* = 4.691 > 0$, το p -value του μονόπλευρου ελέγχου $H_0: \beta_1 = 55$ με $H_1: \beta_1 > 55$ είναι

$$p\text{-value} = \frac{1}{2} \text{Sig} \approx \frac{1}{2} 0.000 \approx 0$$

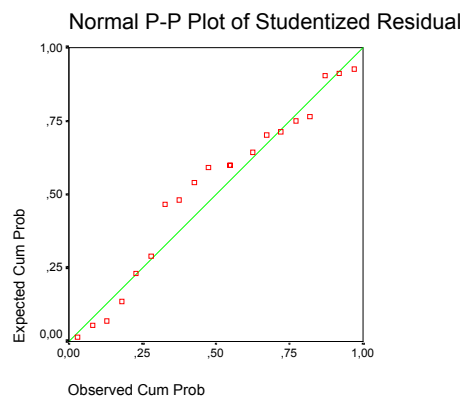
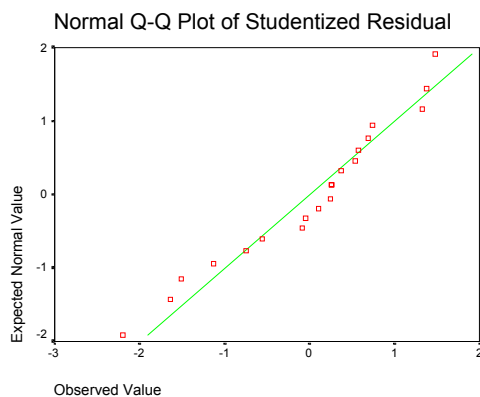
και άρα και πάλι απορρίπτουμε την H_0 έναντι της H_1 (δηλ. $\beta_1 > 55$).

ε) Αρχικά εισάγουμε τη στήλη με τα κατάλοιπα (residuals) ως εξής: Εκτελούμε τη διαδικασία Analyze/Regression/Linear (με Y1,SX και Y2,SX) και στο Save επιλέγουμε τα Studentized residuals. Από το Graphs/Q-Q Plot και το Graphs/P-P Plot (test distribution: Normal) λαμβάνουμε τα γραφήματα Q-Q, P-P για τα δύο σεμινάρια:

Γραφήματα για τα κατάλοιπα του μοντέλου $Y_1 = \beta_0 + \beta_1 \sqrt{X} + \varepsilon$:



Γραφήματα για τα κατάλοιπα του μοντέλου $Y_2 = \beta_0 + \beta_1 \sqrt{X} + \varepsilon$:



Από τα παραπάνω γραφήματα δεν έχουμε κάποια ισχυρή ένδειξη για την απόρριψη της υπόθεσης της κανονικότητας των καταλοίπων. Για να μπορέσουμε να αποφανθούμε με δεδομένη πιθανότητα ε-σφαλμένης απόρριψης της H_0 (π.χ. $\alpha = 5\%$) θα πρέπει να προχωρήσουμε σε έναν έλεγχο με βάση το τεστ Kolmogorov-Smirnov. Συγκεκριμένα, χρησιμοποιώντας τη διαδικασία Analyze/non parametric tests/ One-Sample Kolmogorov-Smirnov test (test distribution Normal) προκύπτει (για τα δύο σεμινάρια) ότι:

One-Sample Kolmogorov-Smirnov Test

		SRE_1 Studentized Residual
N		20
Normal Parameters ^{a,b}	Mean	1,8612E-03
	Std. Deviation	1,0356153
Most Extreme Differences	Absolute	,172
	Positive	,172
	Negative	-,065
Kolmogorov-Smirnov Z		,767
Asymp. Sig. (2-tailed)		,599

a. Test distribution is Normal.

b. Calculated from data.

One-Sample Kolmogorov-Smirnov Test

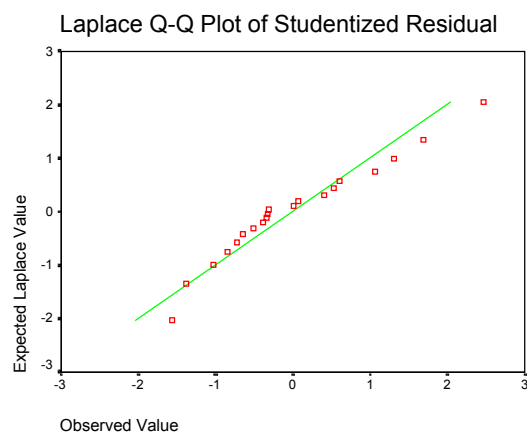
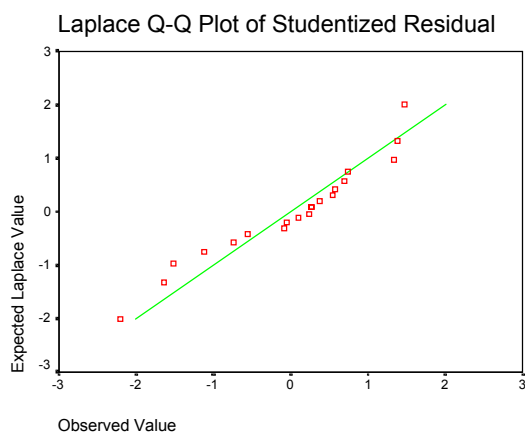
		SRE_2 Studentized Residual
N		20
Normal Parameters ^{a,b}	Mean	2,1079E-03
	Std. Deviation	1,0218689
Most Extreme Differences	Absolute	,167
	Positive	,086
	Negative	-,167
Kolmogorov-Smirnov Z		,745
Asymp. Sig. (2-tailed)		,636

a. Test distribution is Normal.

b. Calculated from data.

Και στα δύο μοντέλα δεν μπορούμε να απορρίψουμε την υπόθεση της κανονικότητας διότι τα αντίστοιχα p-value είναι αρκετά υψηλά (0.599 και 0.636).

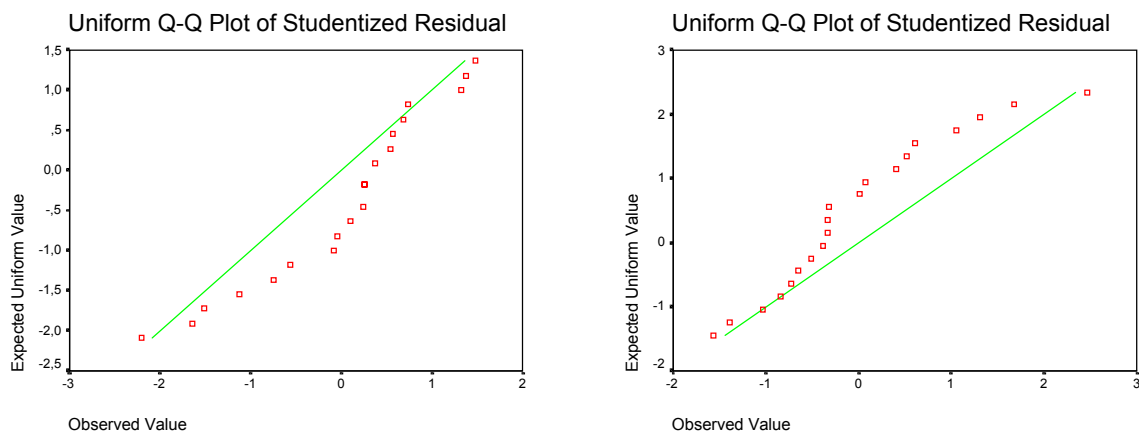
στ) Από το Graphs/Q-Q Plot (test distribution: Laplace) λαμβάνουμε τα γραφήματα Q-Q για τα δύο σεμινάρια:



Από τα παραπάνω γραφήματα δεν διακρίνεται κάποια ισχυρή ένδειξη για την απόρριψη της υπόθεσης ότι τα κατάλοιπα προέρχονται από κατανομή Laplace. Σε αυτό το σημείο θα πρέπει να επισημάνουμε ότι δεν υπάρχει αντίφαση σε σχέση με τον έλεγχο κανονικότητας που έγινε παραπάνω: και στις δύο περιπτώσεις δεν έχουμε αρκετά στοιχεία ώστε να απορρίψουμε ότι τα κατάλοιπα προέρχονται από την

κανονική ή την Laplace κατανομή (για να μπορέσουμε να καταλήξουμε στη μία από τις δύο κατανομές προφανώς χρειαζόμαστε μεγαλύτερο δείγμα).

ζ) Από το Graphs/Q-Q Plot (test distribution: Uniform) λαμβάνουμε τα γραφήματα Q-Q για τα δύο σεμινάρια:



Από τα παραπάνω γραφήματα διακρίνεται κάποια ένδειξη για την απόρριψη της υπόθεσης ότι τα σφάλματα προέρχονται από ομοιόμορφη κατανομή (τα σημεία δεν βρίσκονται «πολύ κοντά» σε μία ευθεία). Για να μπορέσουμε και πάλι να αποφανθούμε με δεδομένη πιθανότητα εσφαλμένης απόρριψης της H_0 (π.χ. $\alpha = 5\%$) θα πρέπει να προχωρήσουμε σε έναν έλεγχο με βάση το τεστ Kolmogorov-Smirnov και το τεστ χ -τετράγωνο.

i) Χρησιμοποιώντας τη διαδικασία Analyze/non parametric tests/ One-Sample Kolmogorov-Smirnov test (test distribution Uniform) προκύπτει (για τα δύο σεμινάρια) ότι:

One-Sample Kolmogorov-Smirnov Test 2

		SRE_1 Studentized Residual	SRE_2 Studentized Residual
N		20	20
Uniform Parameters ^{a,b}	Minimum	-1,56443	-2,20364
	Maximum	2,46888	1,47482
Most Extreme Differences	Absolute	,262	,276
	Positive	,262	,051
	Negative	-,050	-,276
Kolmogorov-Smirnov Z		1,174	1,236
Asymp. Sig. (2-tailed)		,127	,094

a. Test distribution is Uniform.

b. Calculated from data.

Παρατηρούμε ότι τα p-value των ελέγχων στα δύο μοντέλα είναι μεν σχετικά μικρά, αλλά δεν βρίσκονται κάτω του ορίου του 5% (το ένα βρίσκεται κάτω του ορίου του 10%). Συνεπώς και πάλι δεν μπορούμε να απορρίψουμε την υπόθεση H_0 : τα κατάλοιπα προέρχονται από ομοιόμορφη κατανομή σε ε.σ. 5%. Παρόλα αυτά, η πιθανότητα να εμφανιστούν τέτοια δείγματα υπό την H_0 (δηλ. το p-value) είναι μόλις 12,7% και 9,4% αντίστοιχα. Η αδυναμία απόρριψης της H_0 ίσως να οφείλεται στο ότι το δείγμα είναι σχετικά μικρό (αν είχαμε στη διάθεσή μας λίγο μεγαλύτερο δείγμα μάλλον θα απορρίπταμε την H_0).

ii) Υπενθυμίζεται ότι το χ -τετράγωνο τεστ βασίζεται στο γεγονός ότι αν ένα τυχαίο διάνυσμα $(Z_1, Z_2, \dots, Z_k)$ ακολουθεί πολυωνυμική κατανομή με παραμέτρους $(n, p_1, p_2, \dots, p_k)$ με $\sum_{i=1}^k p_i = 1$ (δηλαδή Z_i είναι το πλήθος των επιτυχιών i -είδους σε μία ακολουθία n δοκιμών, η κάθε μία από τις οποίες είναι επιτυχία i -είδους με πιθανότητα $p_i, i=1,2,\dots,k$) τότε η στατιστική συνάρτηση

$$\chi^2 = \sum_{i=1}^k \frac{(Z_i - np_i)^2}{np_i} \sim \chi_{k-1}^2 \quad (\text{ασυμπτωτικά})$$

(απαιτείται επίσης να ισχύει ότι $np_i \geq 5$). Στην περίπτωση μας επιθυμούμε να ελέγξουμε την H_0 : τα studentized residuals $\hat{e}_1^*, \dots, \hat{e}_n^*$ προέρχονται από την ομοιόμορφη κατανομή. Επειδή τα $\hat{e}_i^*$ είναι μεταξύ του $(-2.5, 2.5)$ θα ελέγξουμε αν προέρχονται από την ομοιόμορφη στο $(-2.5, 2.5)$ κατανομή (ή π.χ. θα μπορούσαμε να θεωρήσουμε το $(-3, 3)$). Διαμερίζουμε το $(-2.5, 2.5)$ στα διαστήματα $A_1 = (-2.5, -1.25]$, $A_2 = (-1.25, 0]$, $A_3 = (0, 1.25]$, $A_4 = (1.25, 2.5)$ και θέτουμε $Z_i = \text{πλήθος από τα } \hat{e}_1^*, \dots, \hat{e}_n^* \text{ που ανήκουν στο } A_i, i = 1, 2, 3, 4$. Αν θεωρήσουμε ότι τα $\hat{e}_1^*, \dots, \hat{e}_n^*$ είναι ανεξάρτητα, το τυχαίο διάνυσμα (Z_1, Z_2, Z_3, Z_4) ακολουθεί (κάτω από την H_0) πολυωνυμική κατανομή με παραμέτρους ($n = 20, p_1 = p_2 = p_3 = p_4 = 1/4$). Επομένως, αν ισχύει η H_0 θα ισχύει (προσεγγιστικά) ότι

$$X^2 = \sum_{i=1}^k \frac{(Z_i - np_i)^2}{np_i} = \sum_{i=1}^4 \frac{(Z_i - 5)^2}{5} \sim \chi_3^2$$

(εδώ έχουμε διαμερίσει το $(-2.5, 2.5)$ έτσι ώστε να ισχύει ότι $np_i \geq 5$) και άρα απορρίπτουμε την H_0 όταν

$$X^2 = \sum_{i=1}^4 \frac{(Z_i - 5)^2}{5} > \chi_{3;\alpha}^2$$

(Εναλλακτικά θα μπορούσαμε να θεωρήσουμε ότι τα δύο άκρα του διαστήματος της ομοιόμορφης κατανομής εκτιμώνται από τα δεδομένα και συνεπώς η στατιστική συνάρτηση X^2 ακολουθεί κατανομή $\chi_{3-2}^2 = \chi_1^2$). Παρατηρούμε ότι τα $\hat{e}_i^*$ για τα δύο μοντέλα είναι

$$X^2 = \frac{(2-5)^2}{5} + \frac{(9-5)^2}{5} + \frac{(6-5)^2}{5} + \frac{(3-5)^2}{5} = 6 < \chi_{3;0.05}^2 = 7.81, \quad \text{p-value} = 1 - F_{\chi_3^2}(6) = 0.112$$

$$X^2 = \frac{(3-5)^2}{5} + \frac{(5-5)^2}{5} + \frac{(9-5)^2}{5} + \frac{(3-5)^2}{5} = 4.8 < \chi_{3;0.05}^2 = 7.81, \quad \text{p-value} = 1 - F_{\chi_3^2}(4.8) = 0.187$$

και επομένως δεν μπορούμε να απορρίψουμε την H_0 σε ε.σ. 5% σε κανένα από τα δύο μοντέλα.

Αν θέλαμε να κάνουμε το συγκεκριμένο τεστ με τη βοήθεια του SPSS εργαζόμαστε ως εξής: Αρχικά μετασχηματίζουμε τα δεδομένα χρησιμοποιώντας τη διαδικασία Transform / Recode into Different Variables με $\text{sre_1} \rightarrow \text{e1}$, $\text{sre_2} \rightarrow \text{e2}$, με επιλογή old and new values: Range: -2.5 through $-1.25 \rightarrow 1$, -1.25 through $0 \rightarrow 2$, 0 through $1.25 \rightarrow 3$, 1.25 through $2.5 \rightarrow 4$. Στη συνέχεια εκτελούμε τη διαδικασία Analyze / Non-parametric tests / chi-square (Test variable list: e1, e2, Range: Get from Data, Expected values: All categories equal (διότι έχουμε ομοιόμορφη κατανομή και τα διαστήματα στα οποία έχουμε διαμερίσει το $(-2.5, 2.5)$ έχουν ίσο πλάτος) από όπου προκύπτουν οι πίνακες:

E1				E2			
	Observed N	Expected N	Residual		Observed N	Expected N	Residual
1,00	2	5,0	-3,0	1,00	3	5,0	-2,0
2,00	9	5,0	4,0	2,00	5	5,0	,0
3,00	6	5,0	1,0	3,00	9	5,0	4,0
4,00	3	5,0	-2,0	4,00	3	5,0	-2,0
Total	20			Total	20		

Test Statistics

	E1	E2
Chi-Square ^a	6,000	4,800
df	3	3
Asymp. Sig.	,112	,187

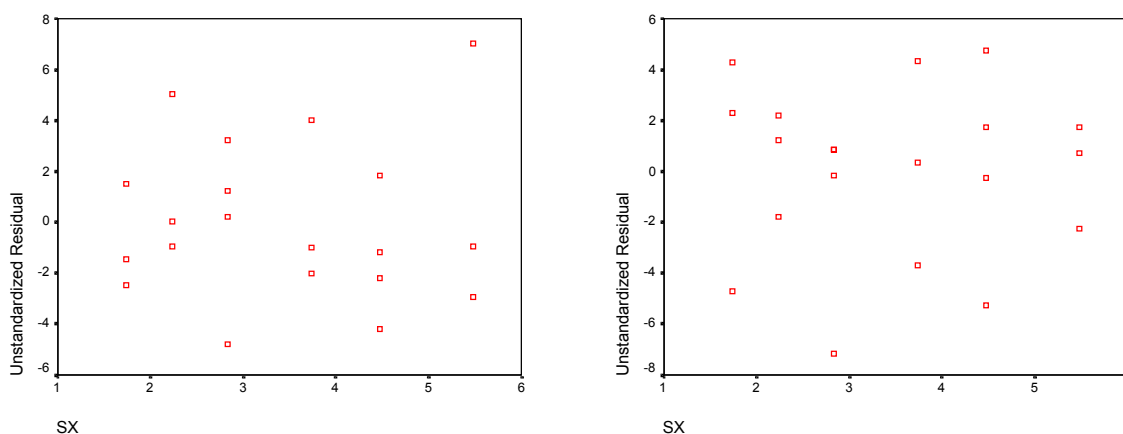
a. 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 5,0.

(Αν χρησιμοποιούσαμε τη διαδικασία Analyze / Non-parametric tests / chi-square με Test variable list: τα αρχικά sre_1, sre_2 και Range: -3, 3 τότε το SPSS προχωρά σε μία αυθαίρετη διαμέριση του (-3, 3) η οποία δεν είναι πάντα η καλύτερη).

η) Στην 21^η γραμμή της στήλης $sx = \sqrt{X}$ εισάγουμε το $\sqrt{12} \approx 3.464$ (οι υπόλοιπες θέσεις στην 21^η γραμμή αφήνονται κενές). Στη συνέχεια εκτελούμε τη διαδικασία Analyze/Regression/Linear επιλέγοντας στο save τα Prediction Intervals με 95%. Στην 21^η γραμμή λαμβάνονται τα αποτελέσματα:

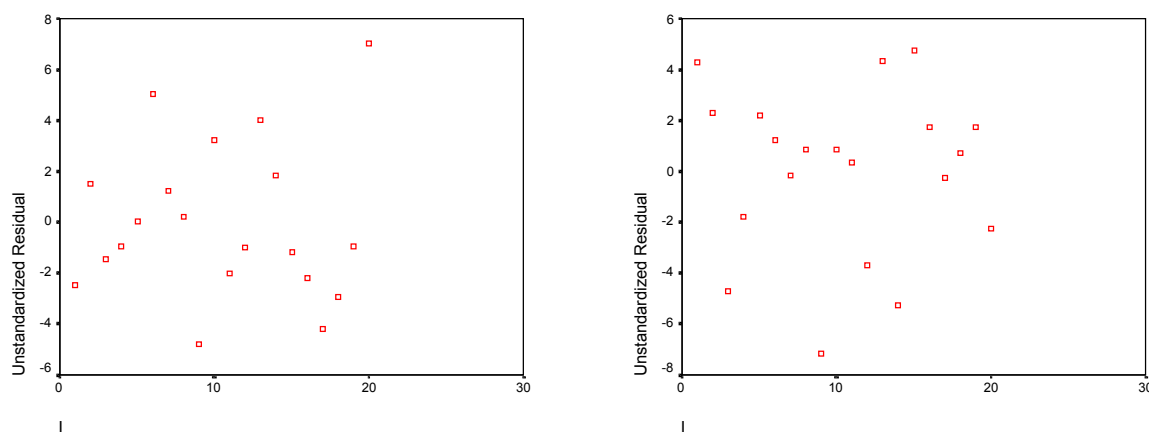
απόδοση 1 ^{ου} σεμιναρίου:	Δ.ε. 95% για την μέση πρόβλεψη: (197,7311, 200,6938)
	Δ.ε. 95% για την ατομική πρόβλεψη: (192,4255, 205,9994)
απόδοση 2 ^{ου} σεμιναρίου:	Δ.ε. 95% για την μέση πρόβλεψη: (196,5173, 199,6672)
	Δ.ε. 95% για την ατομική πρόβλεψη: (190,8762, 205,3082)

θ) Τα scatterplots μεταξύ του $sx = \sqrt{X}$ και του res_1 (unstandardized residuals από 1^ο σεμινάριο) και μεταξύ του $sx = \sqrt{X}$ και του res_2 (unstandardized residuals από 2^ο σεμινάριο) είναι



(θα μπορούσαμε να είχαμε χρησιμοποιήσει και τα studentized residuals). Δεν υπάρχει κάποια εμφανής τάση (π.χ. αυξητική ή μειωτική) στην απόσταση των καταλοίπων από τον οριζόντιο άξονα (τα σημεία φαίνεται να κατανέμονται τυχαία) και συνεπώς δεν έχουμε κάποια ένδειξη ότι η διασπορά των καταλοίπων εξαρτάται από τη διάρκεια.

Τα scatterplots μεταξύ του i (σειρά παρακολούθησης) και του res_1 και μεταξύ του i και του res_2 είναι



Και σε αυτή την περίπτωση δεν υπάρχει κάποια εμφανής τάση στην απόσταση των καταλοίπων από τον οριζόντιο άξονα και συνεπώς δεν έχουμε κάποια ένδειξη ότι η διασπορά των καταλοίπων εξαρτάται από τη σειρά παρακολούθησης.

ι) Ως παρατηρήσεις που επηρεάζουν σημαντικά το μοντέλο (έκτροπες παρατηρήσεις ως προς την ανεξάρτητη μεταβλητή sx) θεωρούμε αυτές με μόχλευση (leverage) $p_{ii} > 3 \cdot 2/n = 0.3$. Επίσης θεωρούμε ως «ασυνήθιστες» τις παρατηρήσεις με studentized residual $|\hat{e}_i^*| > 2$. Οι μοχλεύσεις, καθώς και τα $\hat{e}_i^*$ των $n = 20$ παρατηρήσεων δίνονται στον παρακάτω πίνακα (χρησιμοποιούμε Analyze / Regression / Linear / Dependent: Y1, Independent: SX, Save Leverage Values, Studentized residuals):

i	x	$sx = \sqrt{X}$	p_{ii}	$\hat{e}_i^*$ (Y1,SX)	$\hat{e}_i^*$ (Y2,SX)
1	3	1,7321	0,09217	-0,84785	1,38145
2	3	1,7321	0,09217	0,52207	0,73722
3	3	1,7321	0,09217	-0,50537	-1,51759
4	5	2,2361	0,04576	-0,31865	-0,56359
5	5	2,2361	0,04576	0,01493	0,69137
6	5	2,2361	0,04576	1,68280	0,37763
7	8	2,8284	0,01177	0,40043	-0,04761
8	8	2,8284	0,01177	0,07296	0,26040
9	8	2,8284	0,01177	-1,56443	-2,20364
10	8	2,8284	0,01177	1,05539	0,26040
11	14	3,7417	0,00292	-0,65297	0,10247
12	14	3,7417	0,00292	-0,32703	-1,12377
13	14	3,7417	0,00292	1,30268	1,32872
14	20	4,4721	0,03385	0,60386	-1,64212
15	20	4,4721	0,03385	-0,39034	1,47482
16	20	4,4721	0,03385	-0,72174	0,53974
17	20	4,4721	0,03385	-1,38454	-0,08365
18	30	5,4772	0,13166	-1,03757	0,24317
19	30	5,4772	0,13166	-0,33628	0,57296
20	30	5,4772	0,13166	2,46888	-0,74622

Σύμφωνα με τον παραπάνω πίνακα, δεν υπάρχουν παρατηρήσεις που επηρεάζουν «σημαντικά» το μοντέλο διότι όλα τα p_{ii} είναι μικρότερα του 0.3 (δηλαδή δεν υπάρχουν κάποιες τιμές του sx που να απέχει σημαντικά από τις υπόλοιπες). Παρατηρούμε απλά ότι οι τρεις τελευταίες παρατηρήσεις επηρεάζουν το μοντέλο περισσότερο από τις υπόλοιπες (αλλά δεν μπορούν να θεωρηθούν ως έκτροπες).

Τέλος, ως «ασυνήθιστη» μπορεί να θεωρηθεί η παρατήρηση 20 στο μοντέλο $Y_1 = \beta_0 + \beta_1 \sqrt{X} + \varepsilon$ διότι $\hat{e}_{20}^* = 2,46888$, και η παρατήρηση 9 στο μοντέλο $Y_2 = \beta_0 + \beta_1 \sqrt{X} + \varepsilon$ διότι $\hat{e}_9^* = -2,20364$ (και στις δύο όμως αυτές περιπτώσεις είναι $|\hat{e}_i^*| < 3$, οπότε οι συγκεκριμένες παρατηρήσεις δεν μπορούν να χαρακτηριστούν ως «έκτροπες» (outliers)).

ια) Στα διαγράμματα του ερωτήματος (θ) δεν υπάρχει κάποιος εμφανής σχηματισμός των σημείων στο επίπεδο οπότε από τα διαγράμματα δεν φαίνεται να υπάρχει κάποια σχέση μεταξύ των καταλοίπων. Για να πάρουμε μία ασφαλέστερη απόφαση θα προχωρήσουμε σε ένα τεστ ροών (runs test) των καταλοίπων. Εδώ, οι τιμές των studentized residuals που είναι μεγαλύτερες του 0 έχουν θεωρηθεί ως + και οι τιμές που είναι μικρότερες του 0 ως -.

$\hat{e}_i^*$ (Y1,SX)	
-0,84785	-
0,52207	+
-0,50537	-
-0,31865	-
0,01493	+
1,68280	+
0,40043	+
0,07296	+
-1,56443	-
1,05539	+
-0,65297	-
-0,32703	-
1,30268	+
0,60386	+
-0,39034	-
-0,72174	-
-1,38454	-
-1,03757	-
-0,33628	-
2,46888	+
Ροές:	10

$\hat{e}_i^*$ (Y2,SX)	
1,38145	+
0,73722	+
-1,51759	-
-0,56359	-
0,69137	+
0,37763	+
-0,04761	-
0,26040	+
-2,20364	-
0,26040	+
0,10247	+
-1,12377	-
1,32872	+
-1,64212	-
1,47482	+
0,53974	+
-0,08365	-
0,24317	+
0,57296	+
-0,74622	-
Ροές:	14

Το πλήθος R των ροών από + και από - στο πρώτο μοντέλο είναι 10 ενώ στο δεύτερο 14. Γνωρίζουμε ότι κάτω από την μηδενική υπόθεση H_0 : «τα σφάλματα e_i είναι ανεξάρτητα» τα κατάλοιπα μπορούν πρακτικά να θεωρηθούν ανεξάρτητα, οπότε μπορούμε (σχεδόν) ισοδύναμα να εξετάσουμε την υπόθεση H'_0 : «τα κατάλοιπα είναι ανεξάρτητα» (τα κατάλοιπα είναι γνωστά, αντίθετα από τα σφάλματα που είναι άγνωστα). Σύμφωνα με τη συναφή θεωρία, το πλήθος R των ροών (υπό την H'_0) ακολουθεί ασυμπτωτικά κανονική κατανομή με

$$\mu = \frac{2n_1n_2}{n_1 + n_2} + 1, \quad \sigma^2 = \frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1 + n_2)^2(n_1 + n_2 - 1)}$$

(n_1 :πλήθος από +, n_2 : πλήθος από -) οπότε απορρίπτουμε σε ε.σ. α την H_0 όταν (θεωρούμε αμφίπλευρη εναλλακτική υπόθεση διότι όταν το R είναι «μικρό» υποδηλώνεται θετική συσχέτιση μεταξύ των σφαλμάτων, ενώ όταν το R είναι «μεγάλο» υποδηλώνεται αρνητική συσχέτιση μεταξύ των σφαλμάτων)

$$|z| = \left| \frac{R - \mu + 0.5}{\sigma} \right| > Z_{\alpha/2}$$

(για καλύτερη προσέγγιση χρησιμοποιούμε και διόρθωση συνέχειας +0.5). Ισοδύναμα, απορρίπτουμε σε ε.σ. α την H_0 όταν

$$p - value = P(|Z| > |z| | Z \sim N(0,1)) = 2(1 - \Phi(|z|)) < \alpha.$$

Στο 1^ο μοντέλο ισχύει ότι $n_1=9$, $n_2=11$. Συνεπώς $\mu = 10.9$, $\sigma^2 \approx 4.63$,

$$z = \frac{R - 10.9 + 0.5}{\sqrt{4.63}} = \frac{10 - 10.9 + 0.5}{\sqrt{4.63}} \approx -0.186,$$

$$p - value = 2(1 - \Phi(|z|)) = 2(1 - \Phi(0.186)) \approx 2(1 - 0.573) = 0.853$$

και άρα δεν μπορούμε να απορρίψουμε την H_0 π.χ. σε ε.σ. 5%. Αντίστοιχα, στο 2^ο μοντέλο είναι $n_1=12$, $n_2=8$ από όπου προκύπτει ότι

$$z \approx 1.391 \quad \text{και} \quad p - value \approx 0.164$$

με συνέπεια να μην μπορούμε και πάλι να απορρίψουμε την H_0 σε ε.σ. 5%.

Όλα τα παραπάνω δίνονται αυτόματα από το SPSS μέσω της διαδικασίας Analyze / non-parametric tests / Runs, test variables: studentized residuals[sre_1], studentized residuals[sre_2], Cut Point: Custom (0) από όπου προκύπτει ο πίνακας:

Runs Test

	SRE_1 Studentized Residual	SRE_2 Studentized Residual
Test Value ^a	0	0
Total Cases	20	20
Number of Runs	10	14
Z	-,186	1,391
Asymp. Sig. (2-tailed)	,853	,164

a. User-specified.

- Εναλλακτικά, θα μπορούσαμε να είχαμε χρησιμοποιήσει τη διάμεσο (median) ή το δειγματικό μέσο (mean) ως cut point (αντί του 0 που χρησιμοποιήσαμε παραπάνω).