Τ.Ε.Ι. ΑΘΗΝΑΣ

ΓΕΝΙΚΟ ΤΜΗΜΑ ΜΑΘΗΜΑΤΙΚΩΝ

ΣΗΜΕΙΩΣΕΙΣ ΓΙΑ ΤΟ ΕΡΓΑΣΤΗΡΙΟ

ΓΡΑΜΜΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ

ΔΙΔΑΣΚΟΥΣΑ: ***Ε. ΠΑΠΑΓΕΩΡΓΙΟΥ***

1. ΑΠΛΗ ΠΑΛΙΝΔΡΟΜΗΣΗ (Simple Regression)

1.1. Με χρήση στατιστικού λογισμικού S.P.S.S.

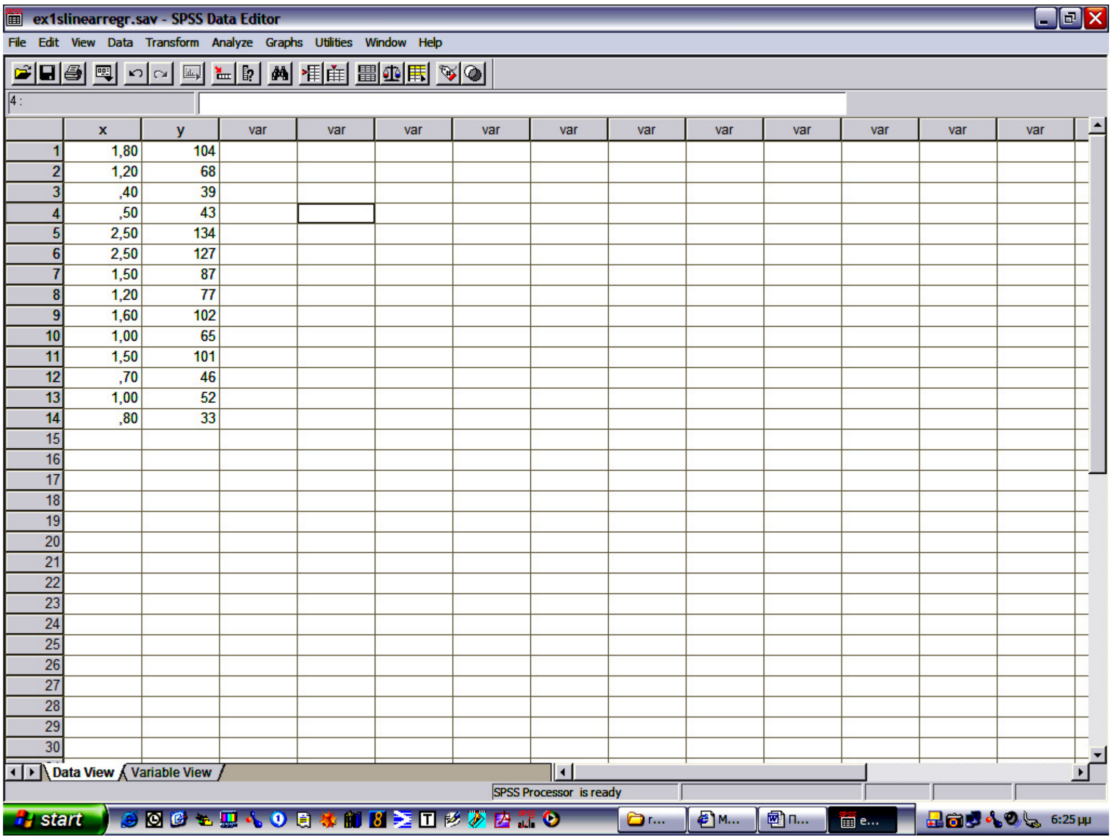
Όπως γνωρίζουμε από τη θεωρία στην *απλή παλινδρόμηση* απαιτούνται δύο ***ποσοτικές μεταβλητές*** εκ των οποίων η μία θεωρείται ***ανεξάρτητη-independent (X)*** και η άλλη ***εξαρτημένη -dependent (Y).*** Η διερεύνηση της μορφής της παλινδρόμησης είναι το βασικό πρόβλημα το οποίο κατ' αρχάς θα πρέπει να επιλυθεί. Είναι δηλαδή απαραίτητο να προσδιορίσουμε αν τα ζεύγη τιμών (X, Y) προσαρμόζονται καλύτερα σε μια ευθεία ή παραβολή ή έλλειψη ή υπερβολή κ.λ.π.

Αν υποθέσουμε ότι η κατάλληλη μορφή παλινδρόμησης, για κάποια συγκεκριμένα ζεύγη τιμών, είναι η γραμμική, τότε για να υπολογίσουμε τους συντελεστές της παλινδρόμησης και τα διάφορα στατιστικά μέτρα τα οποία είναι απαραίτητα, η διαδικασία την οποία πρέπει να ακολουθήσουμε είναι η επόμενη:

Αρχικά ανοίγουμε ένα αρχείο δεδομένων SPSS, δηλ. ένα αρχείο με extension .sav.

Έστω ότι η ***ανεξάρτητη-independent (X)*** εκφράζει τα έξοδα διαφήμισης που έκανε μια εταιρεία για την προώθηση ενός προϊόντος της (σε εκατομμύρια $), ενώ η ***εξαρτημένη - dependent (Y)*** εκφράζει τις πωλήσεις που πραγματοποιήθηκαν για τον πρώτο χρόνο για το ίδιο προϊόν (σε εκατομμύρια $).

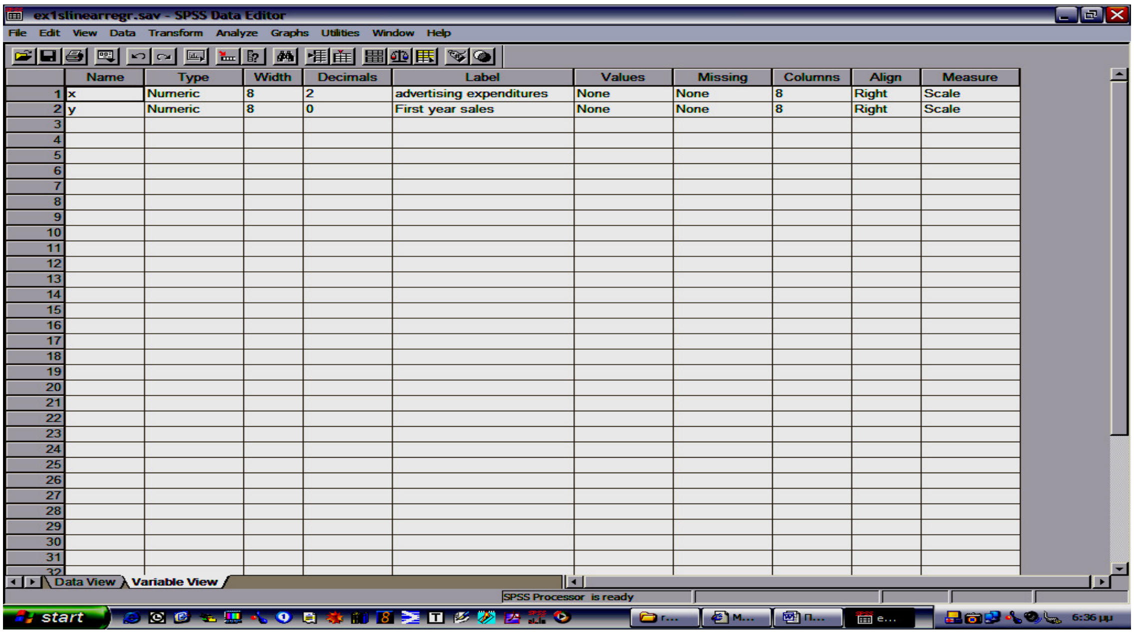
Πληκτρολογούμε τα δεδομένα του προβλήματος όπως στο σχήμα 1.1.



	x	y
1	1,80	104
2	1,20	68
3	,40	39
4	,50	43
5	2,50	134
6	2,50	127
7	1,50	87
8	1,20	77
9	1,60	102
10	1,00	65
11	1,50	101
12	,70	46
13	1,00	52
14	,80	33

Σχήμα 1.1.

Στη συνέχεια καθορίζουμε το όνομα και τον τύπο των μεταβλητών όπως στο σχήμα 1.2.



	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	x	Numeric	8	2	advertising expenditures	None	None	8	Right	Scale
2	y	Numeric	8	0	First year sales	None	None	8	Right	Scale

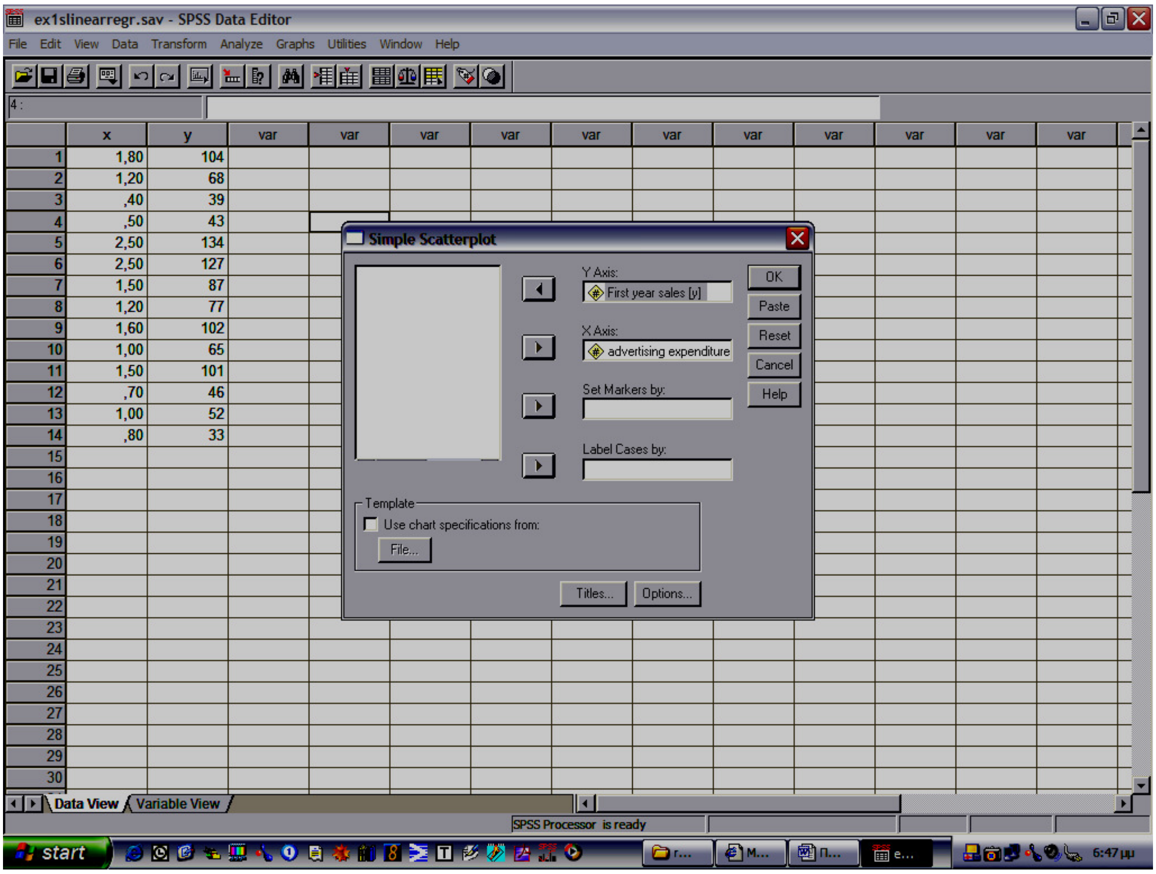
Σχήμα 1.2.

Στη συνέχεια κατασκευάζουμε ένα ***scatter plot*** το οποίο κρίνεται απαραίτητο προκειμένου να αναζητήσουμε αν υπάρχει κάποιου είδους σχέση μεταξύ των δύο μεταβλητών ή αν αυτές εμφανίζονται τυχαία κατανεμημένες.

Αυτό κατασκευάζεται ως εξής:

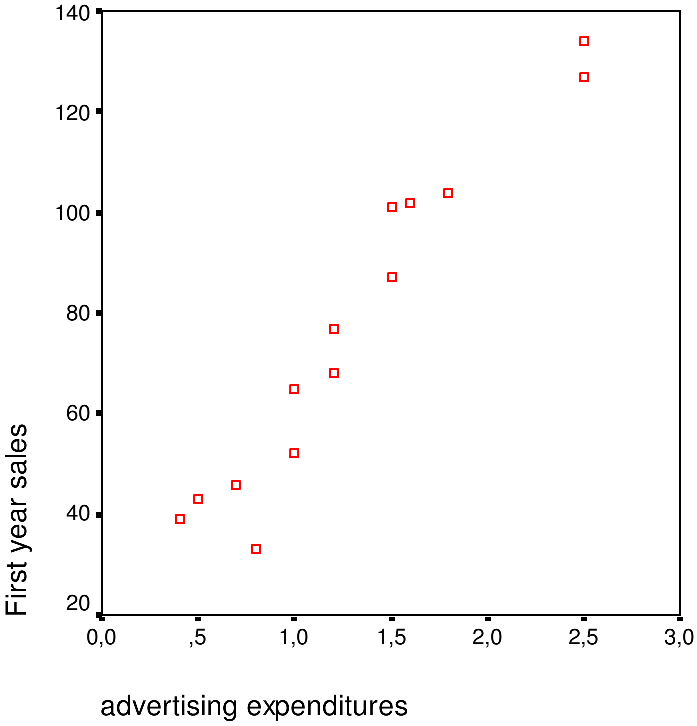
Από το **menu *Graphs*** επιλέγουμε ***Scatter*.**

Εμφανίζεται το επόμενο παράθυρο στο οποίο εισάγουμε όπως φαίνεται στο σχήμα 1.3. τις μεταβλητές μας και πατάμε **O.K.**



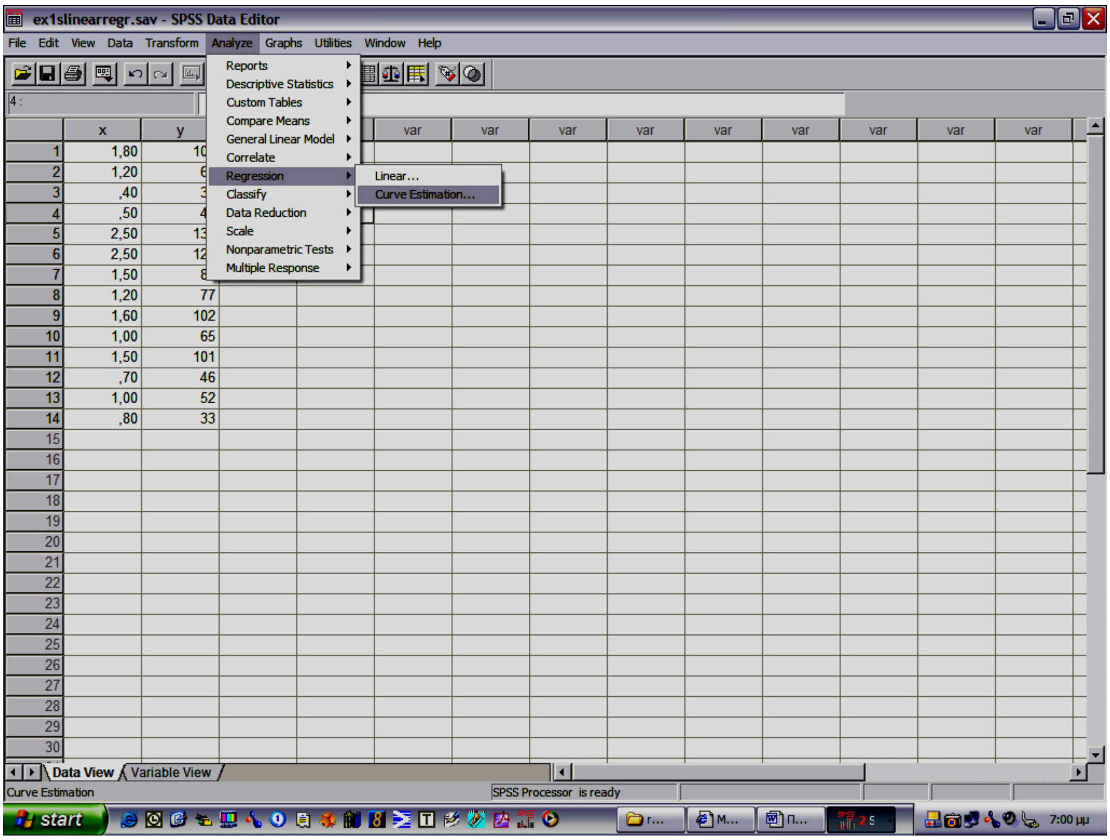
Σχήμα 1.3.

Στη συνέχεια εμφανίζεται το ακόλουθο διάγραμμα διασποράς (σχήμα 1.4.). Από το διάγραμμα αυτό είναι φανερό ότι η σχέση που συνδέει τις δύο μεταβλητές είναι γραμμική και άρα προχωράμε σε γραμμική παλινδρόμηση.



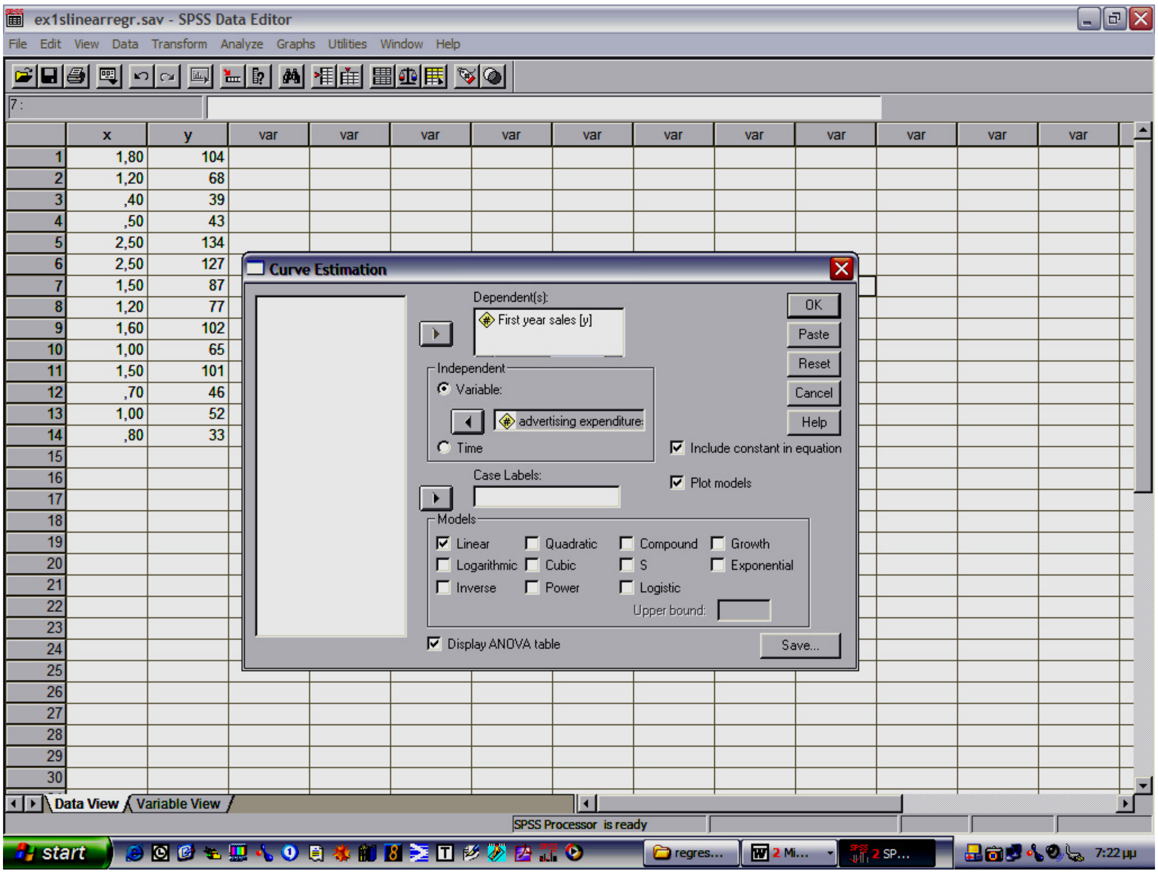
Σχήμα 1.4.

Από το **menu** ***Analyze*** επιλέγουμε ***Regression*** και στη συνέχεια ***Curve Estimation*** όπως στο σχήμα 1.5.



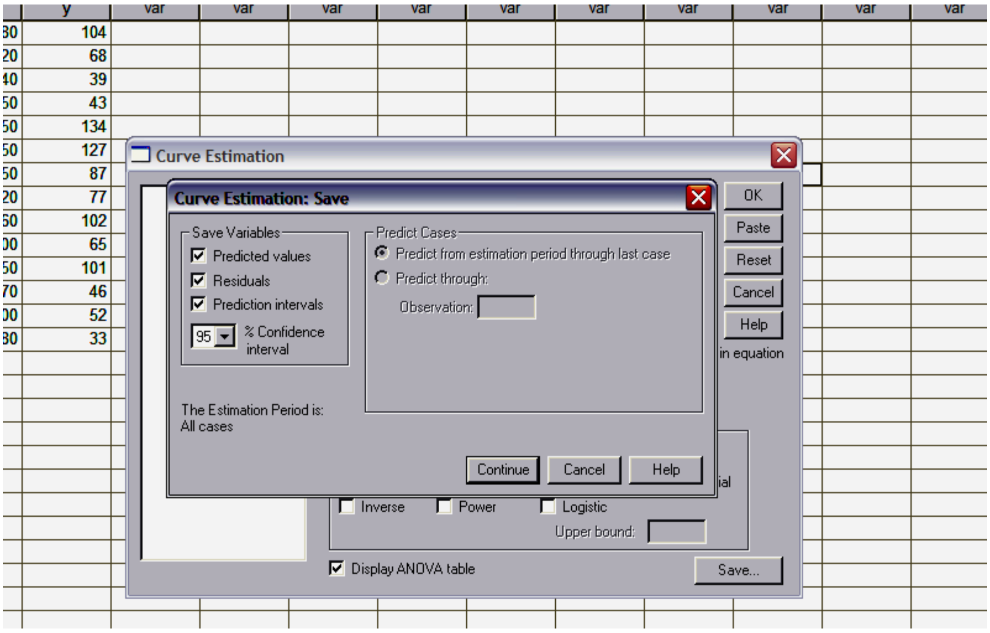
Σχήμα 1.5.

- Στο αριστερό παράθυρο βρίσκονται οι δύο μεταβλητές
- Στο παράθυρο ***Dependent*** μεταφέρουμε την εξαρτημένη μεταβλητή και
- Στο παράθυρο ***Independent*** μεταφέρουμε την ανεξάρτητη μεταβλητή
- Τσεκάρουμε ***Linear*** από τις διαθέσιμες μορφές παλινδρόμησης στη θέση ***Models***.
- Τσεκάρουμε ***include constant in equation*** για να μας δώσει την τιμή του σταθερού όρου.
- Τσεκάρουμε ***Plots Models*** για να μας δώσει το γράφημα
- Τσεκάρουμε ***Display Anova Table*** και η μορφή είναι πλέον η επόμενη:



Σχήμα 1.6.

Αν στο σχήμα 1.6. επιλέξουμε ***Save***, θα έχουμε την επόμενη φόρμα:



Σχήμα 1.7.

- Επιλέγουμε ***Predicted values*** για να πάρουμε τις προβλεπόμενες τιμές της εξαρτημένης μεταβλητής Y.
- Επιλέγουμε ***Residuals*** για να πάρουμε τα σφάλματα
- Τσεκάρουμε ***Prediction intervals*** και επιλέγουμε ***% Confidence interval*** για να πάρουμε το επιθυμητό διάστημα εμπιστοσύνης.
- ***Continue*** και στη συνέχεια με ***O.K*** εμφάνιση των αποτελεσμάτων στον πίνακα που ακολουθεί:

```
MODEL:  MOD_1.
_
Dependent variable.. Y                     Method.. LINEAR

Listwise Deletion of Missing Data

Multiple R           ,96414
R Square             ,92956
Adjusted R Square    ,92369
Standard Error      9,10612
```

```
           Analysis of Variance:

              DF   Sum of Squares      Mean Square

Regression      1        13130,942        13130,942
Residuals      12          995,058           82,921

F =     158,35397       Signif F =  ,0000

-------------------- Variables in the Equation --------------------

Variable                  B        SE B       Beta         T  Sig T

X                 48,597122    3,861856    ,964136    12,584  ,0000
(Constant)        13,823741    5,579203                2,478  ,0291

The following new variables are being created:

  Name        Label

  FIT_1       Fit for Y with X from CURVEFIT, MOD_1 LINEAR
  ERR_1       Error for Y with X from CURVEFIT, MOD_1 LINEAR
  LCL_1       95% LCL for Y with X from CURVEFIT, MOD_1 LINEAR
  UCL_1       95% UCL for Y with X from CURVEFIT, MOD_1 LINEAR
```

Σχήμα 1.8.

Στον προηγούμενο πίνακα βλέπουμε τα αποτελέσματα:

- Στη θέση ***Dependent Variable*** υπάρχει το όνομα της στήλης ***Y***.
- Στη θέση ***Method*** υπάρχει η ένδειξη ***Linear***, ενώ,
- Η τιμή ***Multiple R (Συντελεστής γραμμικής συσχέτισης του Pearson)*** είναι 0.96414,
- Η τιμή ***R square (Δείκτης προσδιορισμού)*** είναι 0.92956,
- Η τιμή ***Adjusted R square (Διορθωμένη τιμή δείκτη προσδιορισμού)*** είναι 0.92369, και
- Η τιμή ***Standard Error (Τυπικό σφάλμα της εκτίμησης)*** είναι 9.10612.

Στη συνέχεια, στον ίδιο πίνακα, βλέπουμε τα αποτελέσματα από την ***Ανάλυση της Διακύμανσης*** -**Analysis of Variance- (ANOVA),** τα αποτελέσματα της οποίας θα ερμηνεύσουμε.

- Στη στήλη ***DF (Degree Of Freedom)*** βλέπουμε τους αριθμούς 1 και 12 οι οποίοι εκφράζουν τους ***βαθμούς ελευθερίας οι οποίοι αντιστοιχούν στο άθροισμα τετραγώνων που ερμηνεύεται από την παλινδρόμηση (Regression,*** **1)** ***και στο άθροισμα τετραγώνων που δεν ερμηνεύεται από την παλινδρόμηση (Residuals, ν-2).***

- Στη στήλη ***Sum. Of Squares*** οι αριθμοί 13130,942 και 995,058 εκφράζουν το άθροισμα τετραγώνων που ερμηνεύεται από την παλινδρόμηση ***(Regression)*** και στο άθροισμα τετραγώνων που δεν ερμηνεύεται από την παλινδρόμηση ***(Residuals)***.
- Στη στήλη ***Mean square*** οι αριθμοί 13130,942 και 82.921 εκφράζουν τα μέσα αθροίσματα τετραγώνων των προηγούμενων τιμών (13130,942/1 και 995,058/12) αντίστοιχα.
- Η τιμή ***F= 158.35397 (F κατανομή)*** είναι το πηλίκο των τιμών της στήλης Mean square.
- Τέλος η τιμή ***Signif. F = 0,000*** είναι η ***κρίσιμη τιμή*** με βάση την οποία ***αποδεχόμαστε*** ή ***απορρίπτουμε την μηδενική υπόθεση***. Εδώ θα πρέπει να αναφέρουμε ότι κατά τη μηδενική υπόθεση ***Ho*** δεν υπάρχει γραμμική σχέση μεταξύ των μεταβλητών Χ και Υ, ενώ κατά την εναλλακτική υπόθεση H_1 υπάρχει γραμμική σχέση μεταξύ των μεταβλητών. Αν το τέστ το κάνουμε σε επίπεδο σημαντικότητας **α = 0,05**, τότε κάθε φορά που ***Signif. F < 0,05*** θα απορρίπτουμε τη μηδενική υπόθεση και δεχόμαστε την εναλλακτική. Στην περίπτωσή μας ***Signif. F = 0,000 < 0,05***, άρα δεχόμαστε ότι υπάρχει γραμμική εξάρτηση μεταξύ των τιμών Χ και Υ.

Στο τελευταίο μέρος του πίνακα -***Variables ιn the equation***- βλέπουμε κατά σειρά:

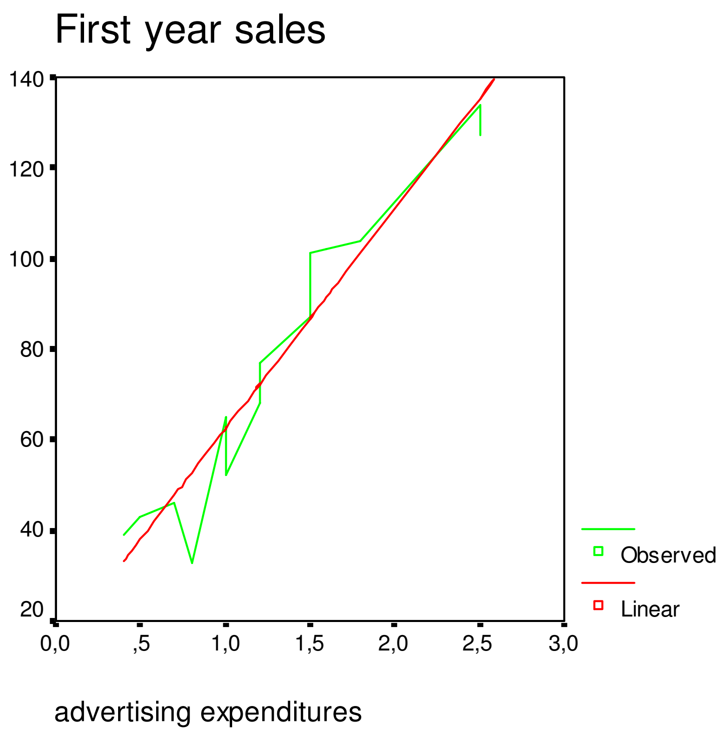
- Την τιμή 48,597122 που βρίσκεται απέναντι από το *Χ* και είναι ο ***συντελεστής της γραμμικής παλινδρόμησης (κλίση)***,
- Την τιμή 13,823741 που είναι *ο* ***σταθερός όρος.***
- Με αυτά τα δεδομένα, η ευθεία της παλινδρόμησης είναι η επόμενη:

$$\boldsymbol{Y = 13{,}823741 + 48{,}597122\, X}$$

- Τις τιμές στη στήλη ***T (T- test)*** που είναι 12,584 και 2,478 για το Χ και τη σταθερά αντίστοιχα και, τέλος,
- Τις τιμές της στήλης ***Sig. T***, οι οποίες αν είναι μικρότερες του επιπέδου σημαντικότητας που ορίσαμε απορρίπτουμε την μηδενική υπόθεση (συντελεστής Χ = 0 και σταθερά = 0). Στο παράδειγμά μας είναι και οι δύο τιμές <0,05, άρα απορρίπτεται η μηδενική υπόθεση και θεωρούμε ότι και ο συντελεστής του Χ και η σταθερά είναι στατιστικά σημαντικά διαφορετικές από το 0. Δηλαδή αν ***Sig. T < α ισχύει η*** H_1.

Στο γράφημα που ακολουθεί, η ευθεία γραμμή είναι η ευθεία της Παλινδρόμησης

(Regression Line) , ***Y* = 13,823741 + 48,597122 *X*** και η τεθλασμένη είναι η γραμμή που προέκυψε από τα πραγματικά ζεύγη τιμών των μεταβλητών X και Y.

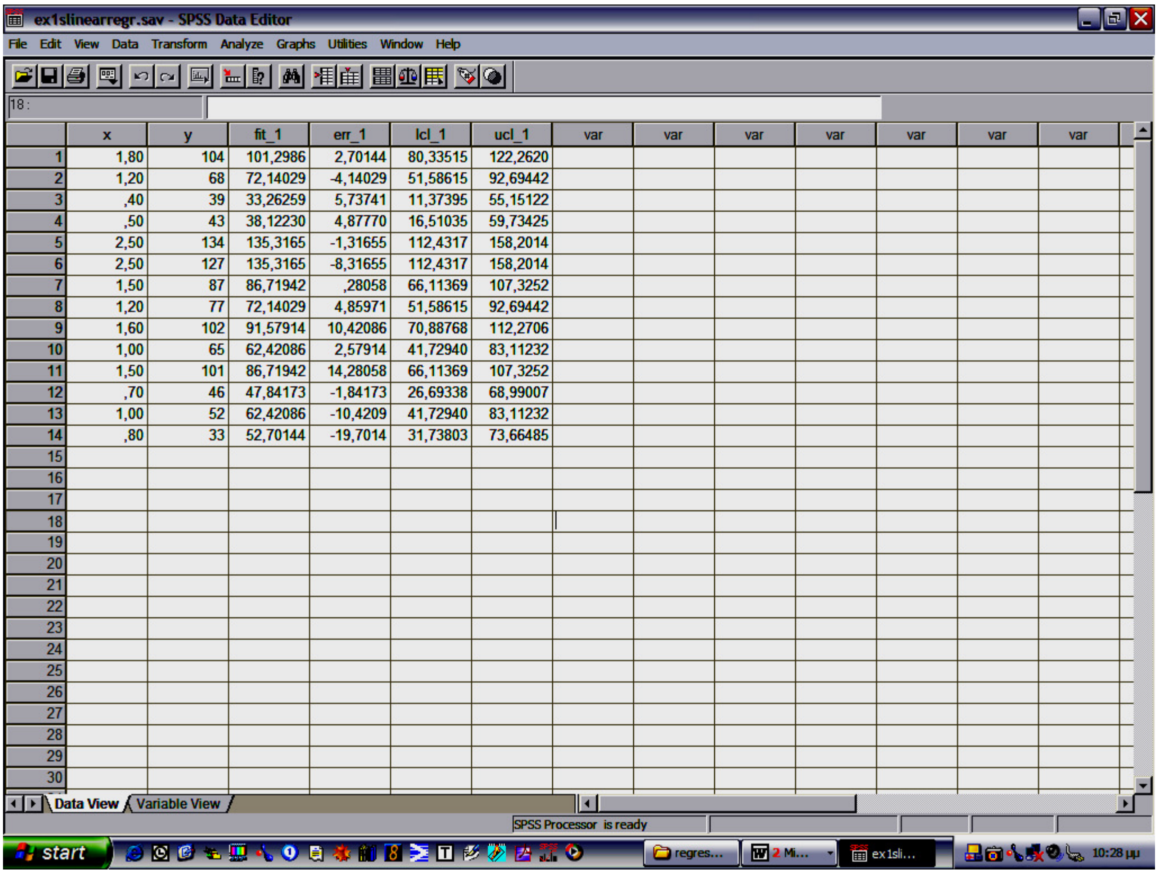


Σχήμα 1.9.

Επιστρέφοντας στο ***Data Editor,*** θα δούμε δίπλα στις αρχικές στήλες των μεταβλητών X και Y και τις στήλες:

- ***fit_1,*** στην οποία αναγράφονται οι ***θεωρητικές τιμές*** οι οποίες προέκυψαν με βάση την ευθεία παλινδρόμησης που ζητήσαμε,
- ***err_1,*** στην οποία αναγράφονται τα σφάλματα (διαφορές εμπειρικών και θεωρητικών τιμών),
- ***lcl_1,*** στην οποία εμφανίζονται τα κάτω άκρα του διαστήματος εμπιστοσύνης που ζητήσαμε (95%) για τις θεωρητικές τιμές του Y και
- ***ucl_1,*** στην οποία εμφανίζονται τα άνω άκρα του διαστήματος εμπιστοσύνης.

Στην εικόνα που ακολουθεί βλέπουμε το **Data Editor** με τις νέες στήλες:



	x	y	fit_1	err_1	lcl_1	ucl_1
1	1,80	104	101,2986	2,70144	80,33515	122,2620
2	1,20	68	72,14029	-4,14029	51,58615	92,69442
3	,40	39	33,26259	5,73741	11,37395	55,15122
4	,50	43	38,12230	4,87770	16,51035	59,73425
5	2,50	134	135,3165	-1,31655	112,4317	158,2014
6	2,50	127	135,3165	-8,31655	112,4317	158,2014
7	1,50	87	86,71942	,28058	66,11369	107,3252
8	1,20	77	72,14029	4,85971	51,58615	92,69442
9	1,60	102	91,57914	10,42086	70,88768	112,2706
10	1,00	65	62,42086	2,57914	41,72940	83,11232
11	1,50	101	86,71942	14,28058	66,11369	107,3252
12	,70	46	47,84173	-1,84173	26,69338	68,99007
13	1,00	52	62,42086	-10,4209	41,72940	83,11232
14	,80	33	52,70144	-19,7014	31,73803	73,66485

Σχήμα 1.10.

Αν αντί Γραμμικής, θέλουμε άλλη μορφή παλινδρόμησης, στο σχήμα 1.6., στη θέση Models θα τσεκάρουμε την επιθυμητή μορφή και θα δουλέψουμε με τον ίδιο τρόπο.

1.2. Με χρήση στατιστικού λογισμικού Minitab

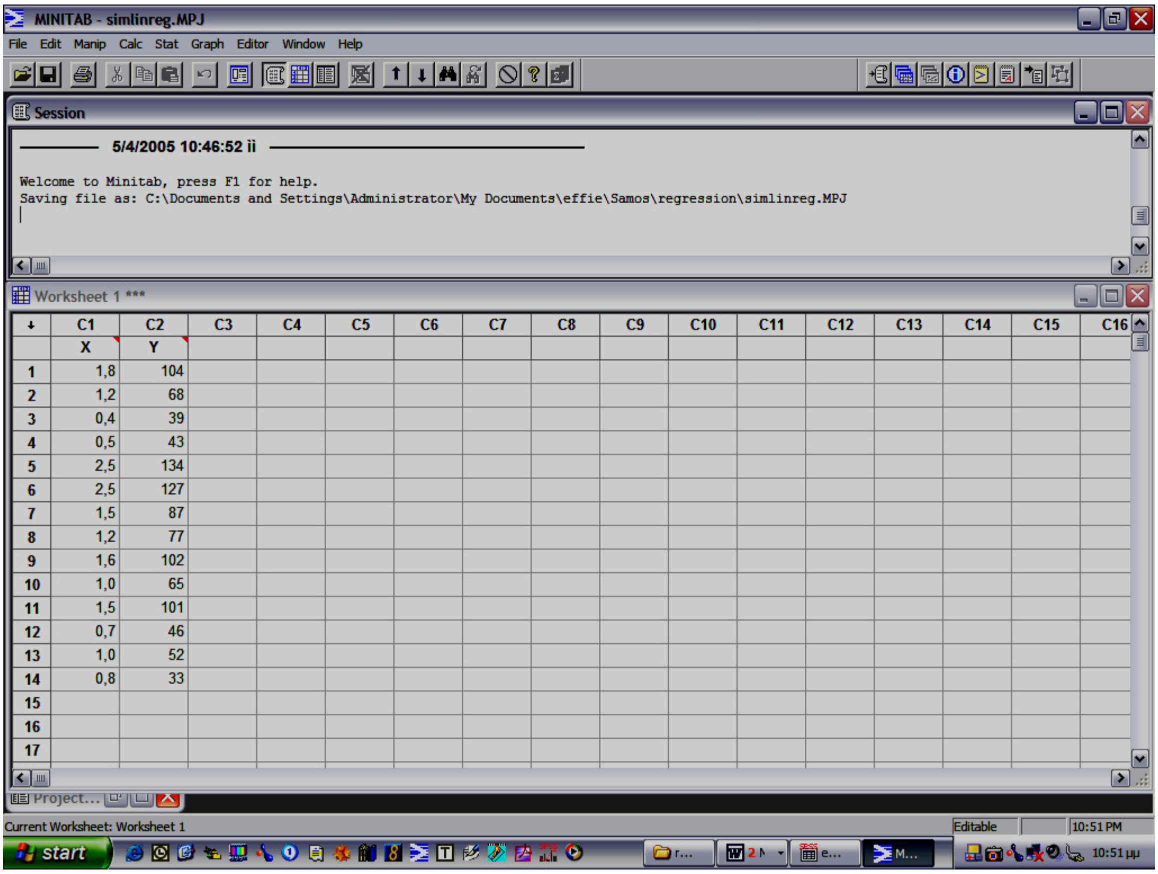
1.2.1 Απλή Γραμμική Παλινδρόμηση

Η ανάλυση αυτή, όπως αναφέρθηκε παραπάνω εξετάζει τη σχέση μεταξύ μιας μεταβλητής απόκρισης ***Y (response variable)*** και μιας ανεξάρτητης μεταβλητής ***X (predictor variable)***. Το **Minitab** παρέχει διαδικασίες για την παλινδρόμηση ελαχίστων τετραγώνων (όταν η μεταβλητή απόκρισης είναι συνεχής). Επίσης, είναι δυνατόν να προσαρμοστεί μια γραμμή παλινδρόμησης καθώς επίσης και γραφικές παραστάσεις των υπολοίπων ***(residuals plots).***

Θα λύσουμε με τη βοήθεια του Minitab το ίδιο πρόβλημα που λύσαμε και στην προηγούμενη παράγραφο με τη βοήθεια του SPSS.

Αρχικά ανοίγουμε ένα αρχείο δεδομένων Minitab, δηλ. ένα αρχείο με extension .mpj.

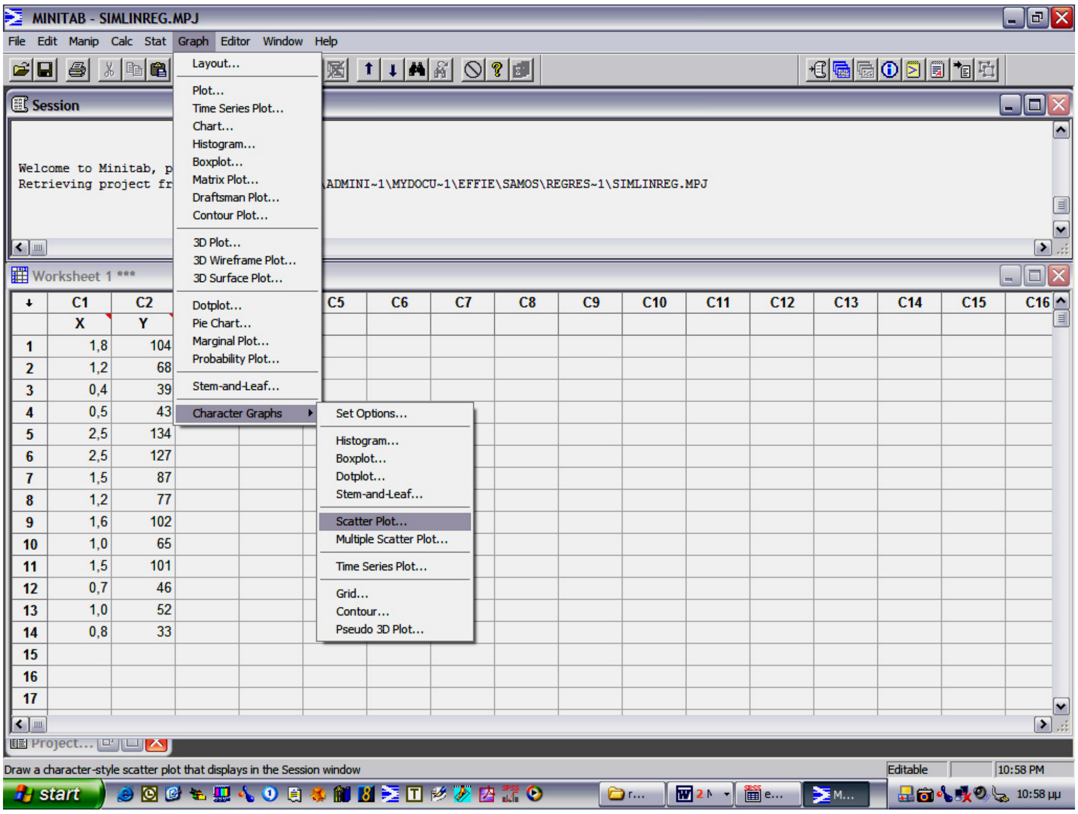
Πληκτρολογούμε τα δεδομένα του προβλήματος όπως στο σχήμα 1.11.



Σχήμα 1.11.

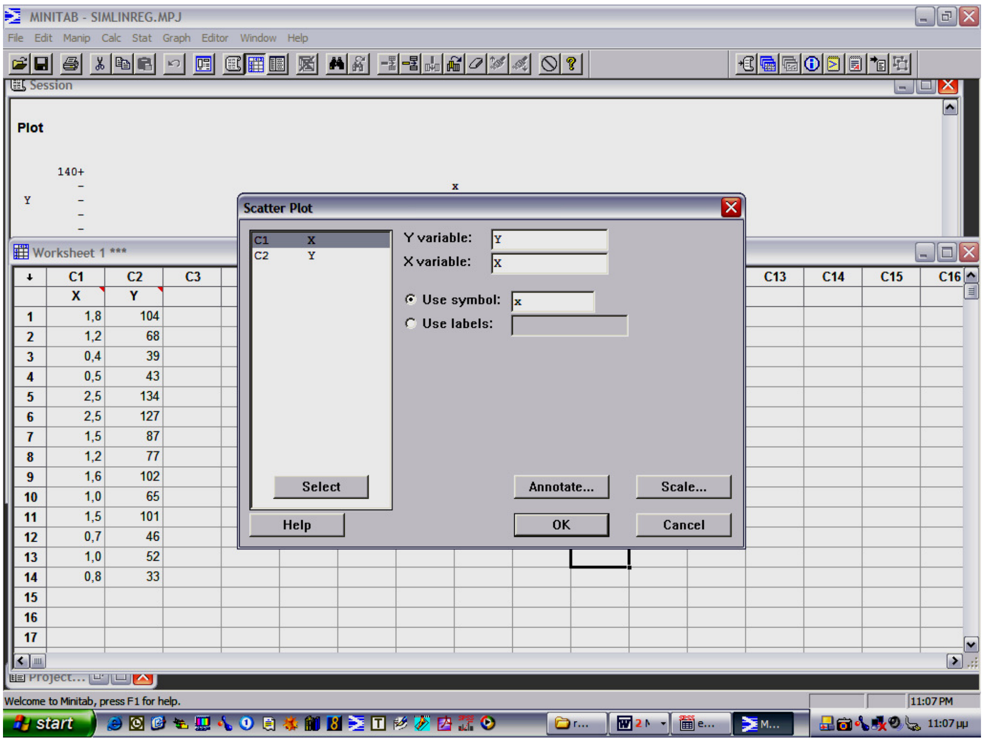
Στη συνέχεια κατασκευάζουμε το ***scatter plot*** για να δούμε τη σχέση μεταξύ των δύο μεταβλητών ως εξής:

Από το **menu *Graph*** επιλέγουμε ***Character Graphs*** και ***Scatter Plot*** όπως δείχνει το επόμενο σχήμα 1.12.



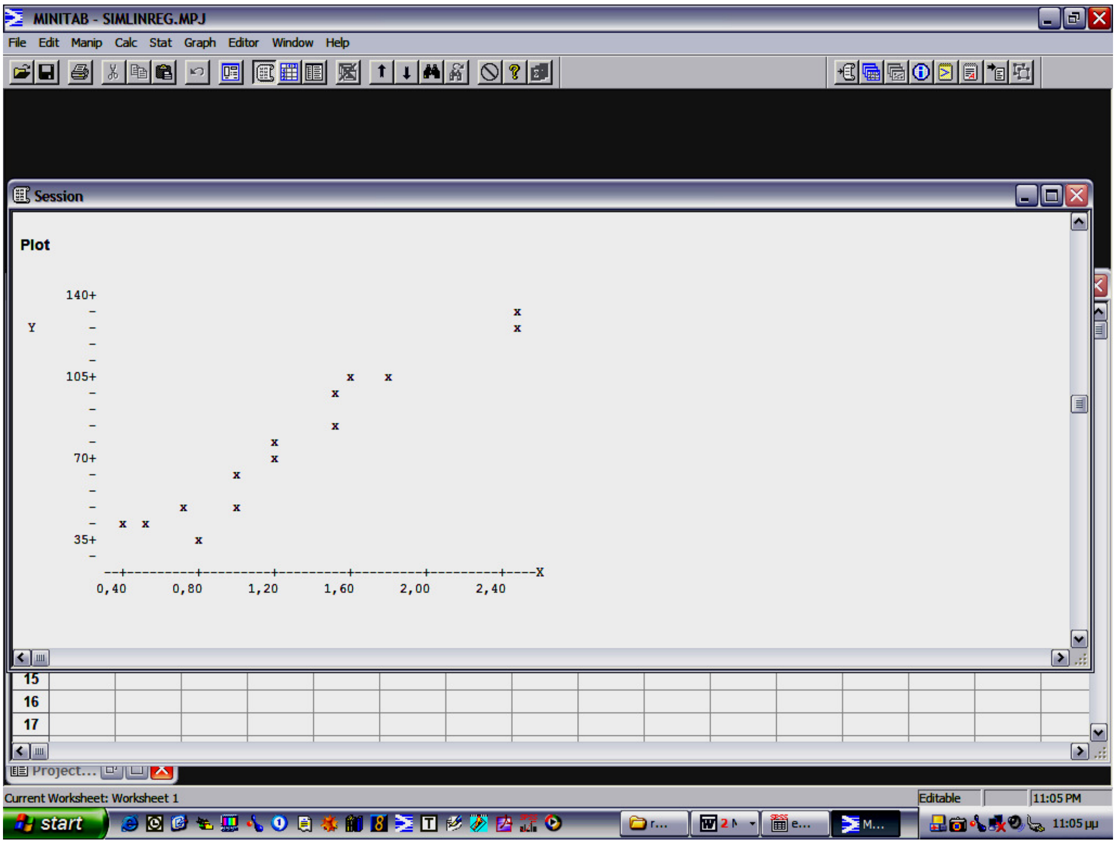
Σχήμα 1.12.

Εμφανίζεται το επόμενο παράθυρο στο οποίο εισάγουμε όπως φαίνεται στο σχήμα 1.13. τις μεταβλητές μας και πατάμε **O.K.**



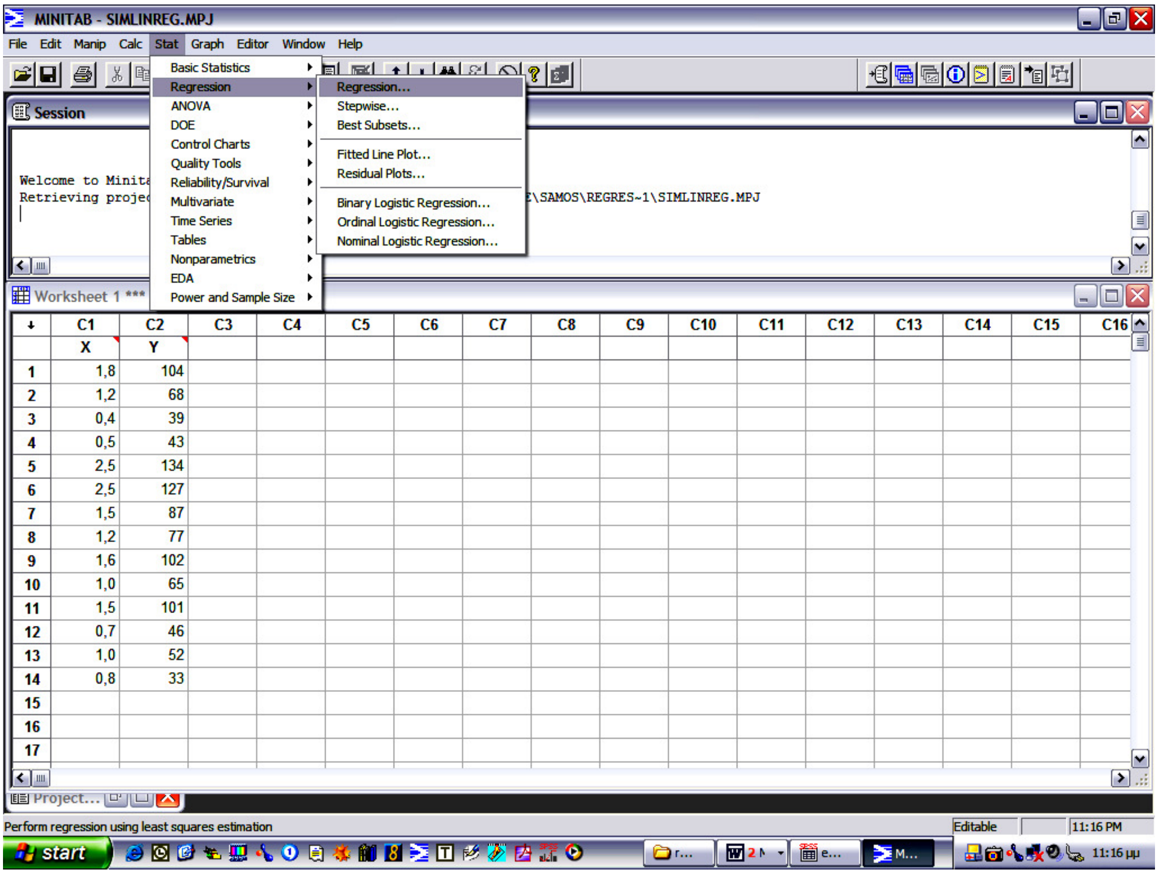
Σχήμα 1.13

Στη συνέχεια εμφανίζεται το ακόλουθο διάγραμμα διασποράς (σχήμα 1.14.). Από το διάγραμμα αυτό είναι φανερό ότι η σχέση που συνδέει τις δύο μεταβλητές είναι γραμμική και άρα προχωράμε σε γραμμική παλινδρόμηση.



Σχήμα 1.14

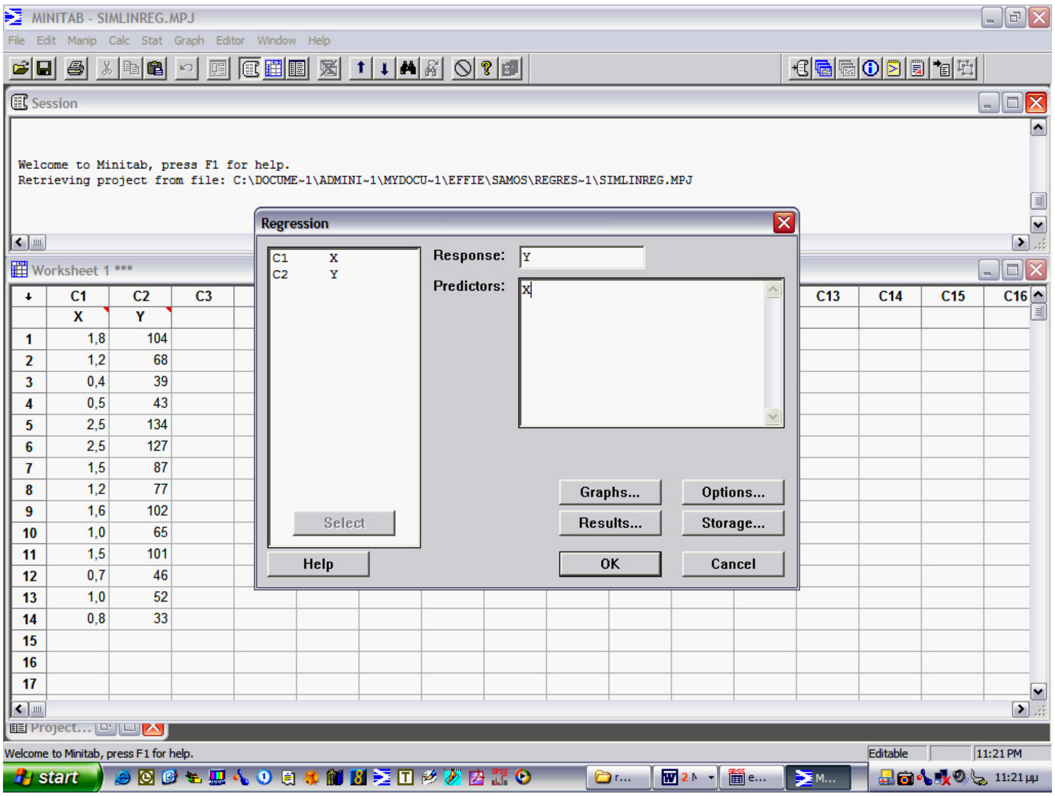
Από το **menu** ***Stat*** επιλέγουμε ***Regression*** και ξανά ***Regression*** όπως δείχνει το επόμενο σχήμα 1.15.



Σχήμα 1.15

Εμφανίζεται η φόρμα του σχήματος 1.16

- Στο ***Response***, εισάγουμε τη στήλη που περιέχει τη μεταβλητή απόκρισης (**Y**).
- Στο ***Predictors*** εισάγουμε τη στήλη που περιέχει τη μεταβλητή πρόβλεψης (**X**).



Σχήμα 1.16

Προαιρετικά μια ή περισσότερες από τις παρακάτω επιλογές, και κατόπιν κλικ **OK**.

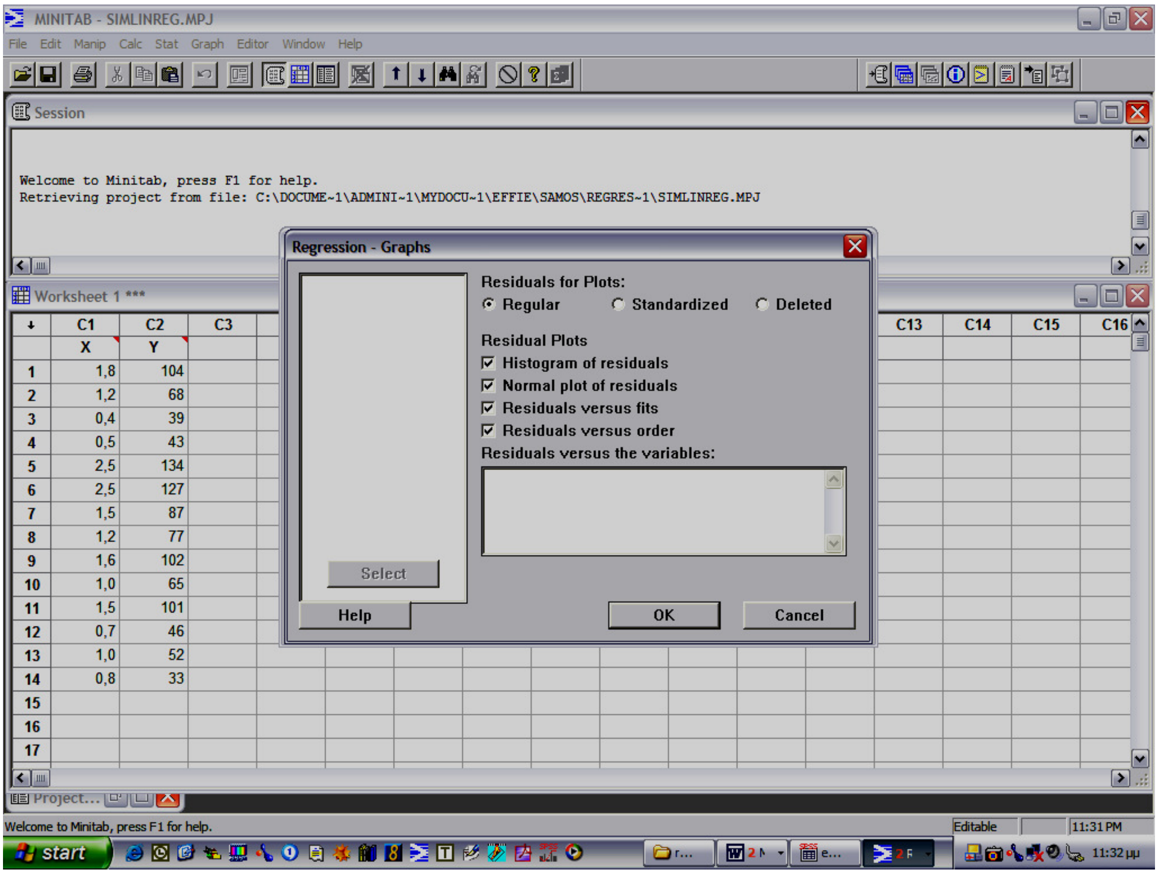
Επιλογές

- ***<u>Graphs</u>***

Υπάρχουν τέσσερις διαφορετικές γραφικές παραστάσεις για τα υπόλοιπα.

- ιστόγραμμα (histogram)
- γραφική παράσταση κανονικής πιθανότητας (normal probability plot)
- γραφική παράσταση των υπολοίπων με τις προσαρμοσμένες τιμές
- γραφική παράσταση των υπολοίπων με τα δεδομένα,

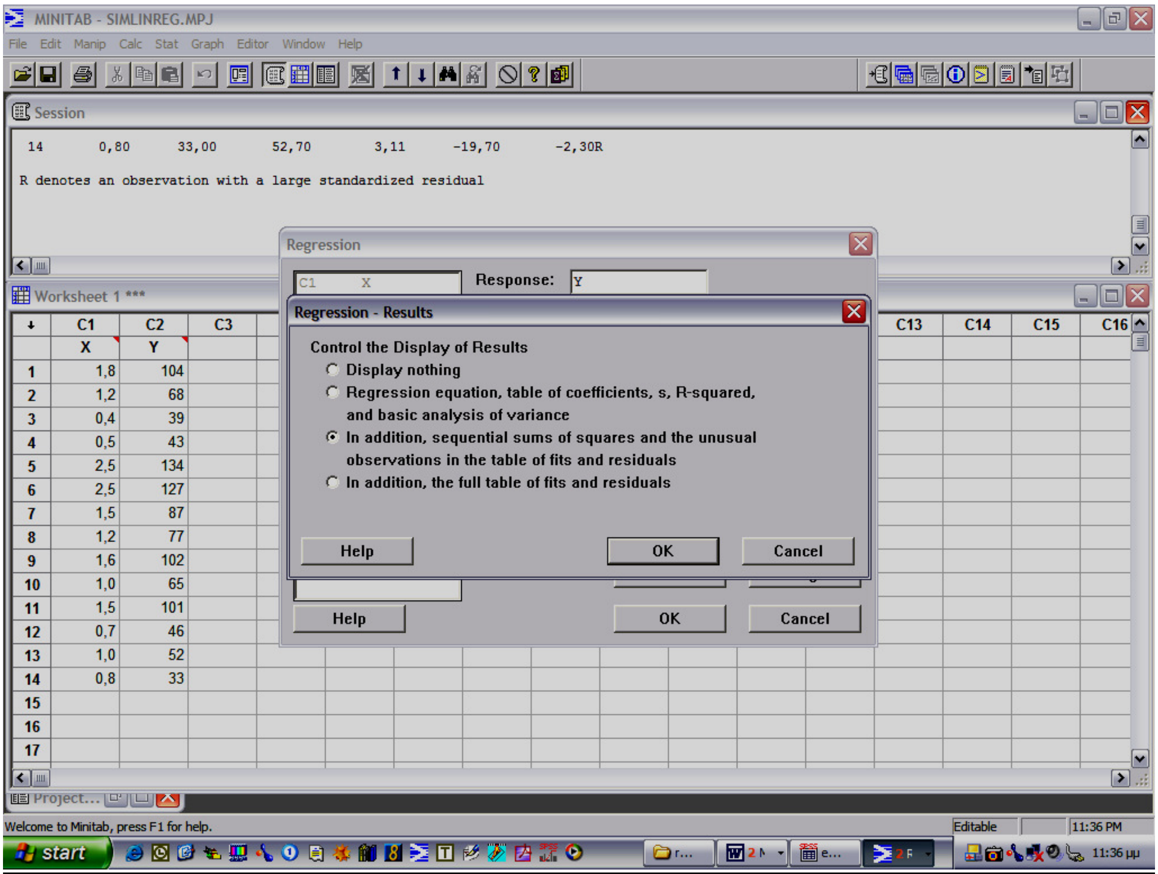
όπως φαίνεται στο παρακάτω σχήμα 1.17.



Σχήμα 1.17

- ***Results***

Επιλογές στο πως θα διαμορφωθεί το ***Session Window Output*** (αν θα περιέχει τους συντελεστές, το s, *το* R^2, τον πίνακα ανάλυσης διασποράς, την εξίσωση παλινδρόμησης, τις προσαρμοσμένες τιμές κλπ.), όπως φαίνεται στο παρακάτω σχήμα 1.18.



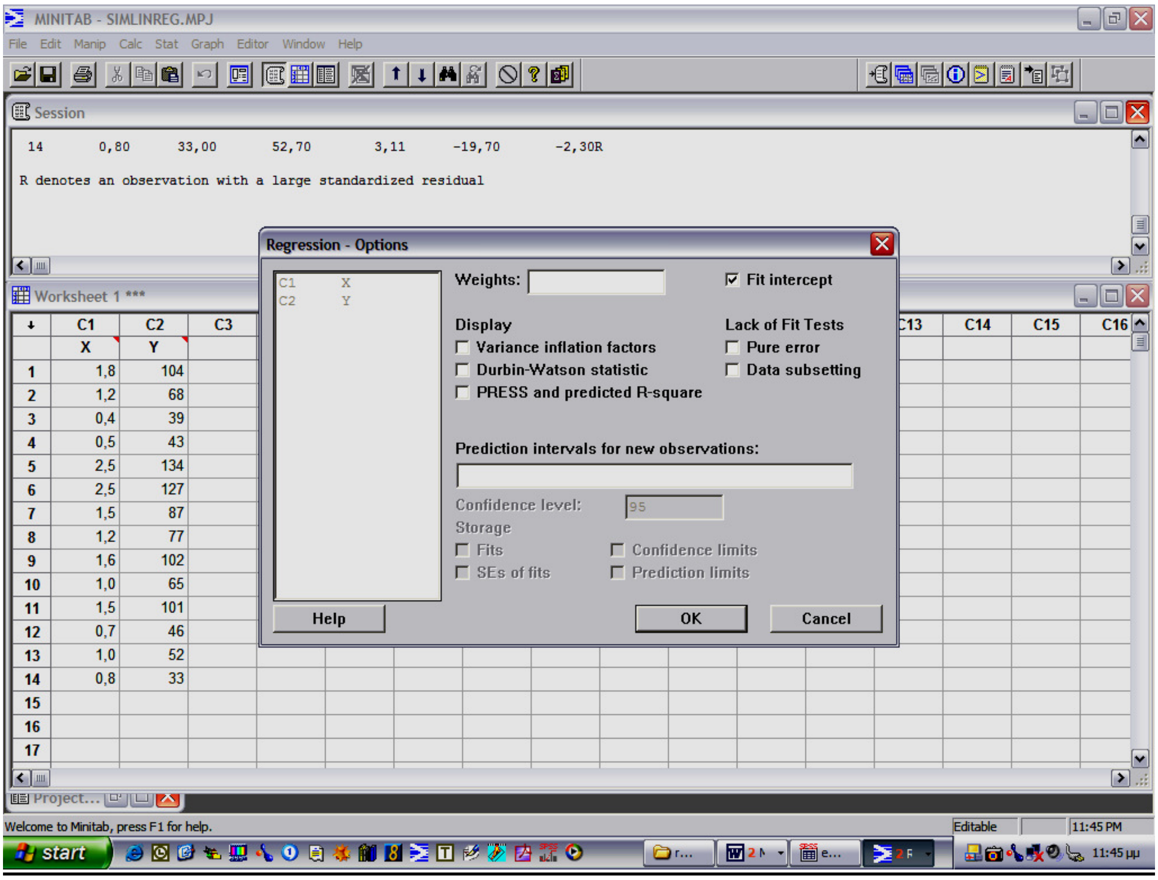
Σχήμα 1.18

- ***Options***

Περιλαμβάνει ορισμένες υποεπιλογές:

- σταθμισμένη παλινδρόμηση
- σταθερά παλινδρόμησης
- Durbin-Watson statistic
- έλεγχος για την ακρίβεια του μοντέλου
- πρόβλεψη για ης νέες παρατηρήσεις

όπως φαίνεται στο παρακάτω σχήμα 1.19.

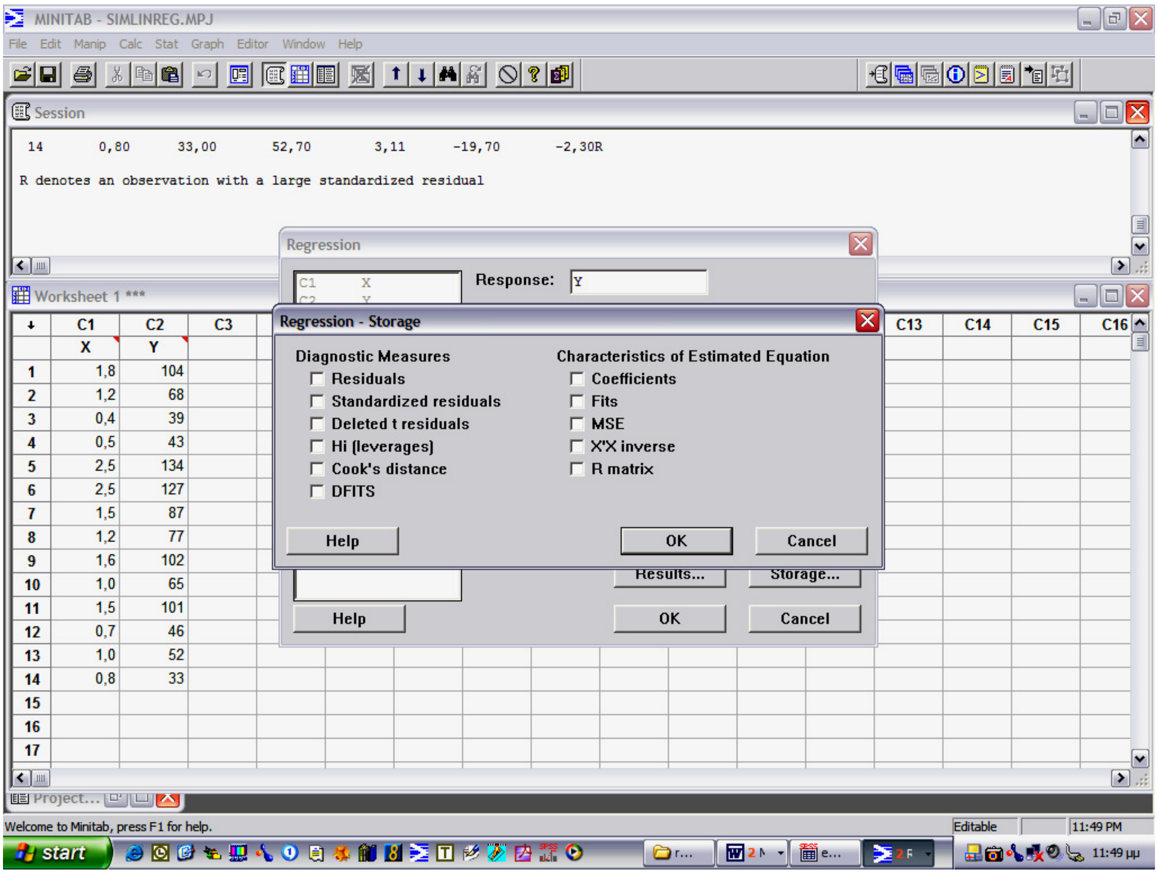


Σχήμα 1.19

- ***Storage***

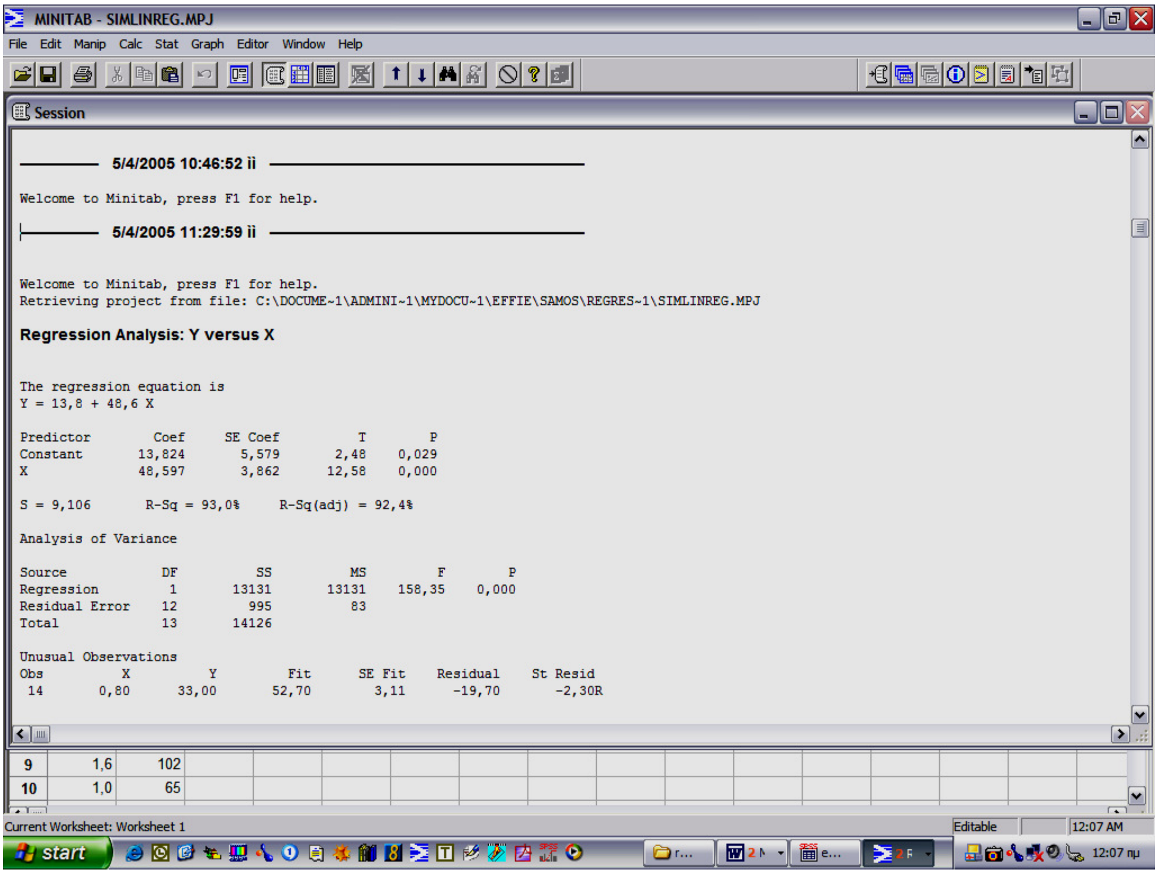
Υπάρχουν 11 υποεπιλογές σχετικά με τους συντελεστές, τις προσαρμοσμένες τιμές, τα υπόλοιπα, την αναγνώριση απομονωμένων σημείων, τον πίνακα R, κλπ. όπως φαίνεται στο παρακάτω σχήμα 1.20.

.



Σχήμα 1.20

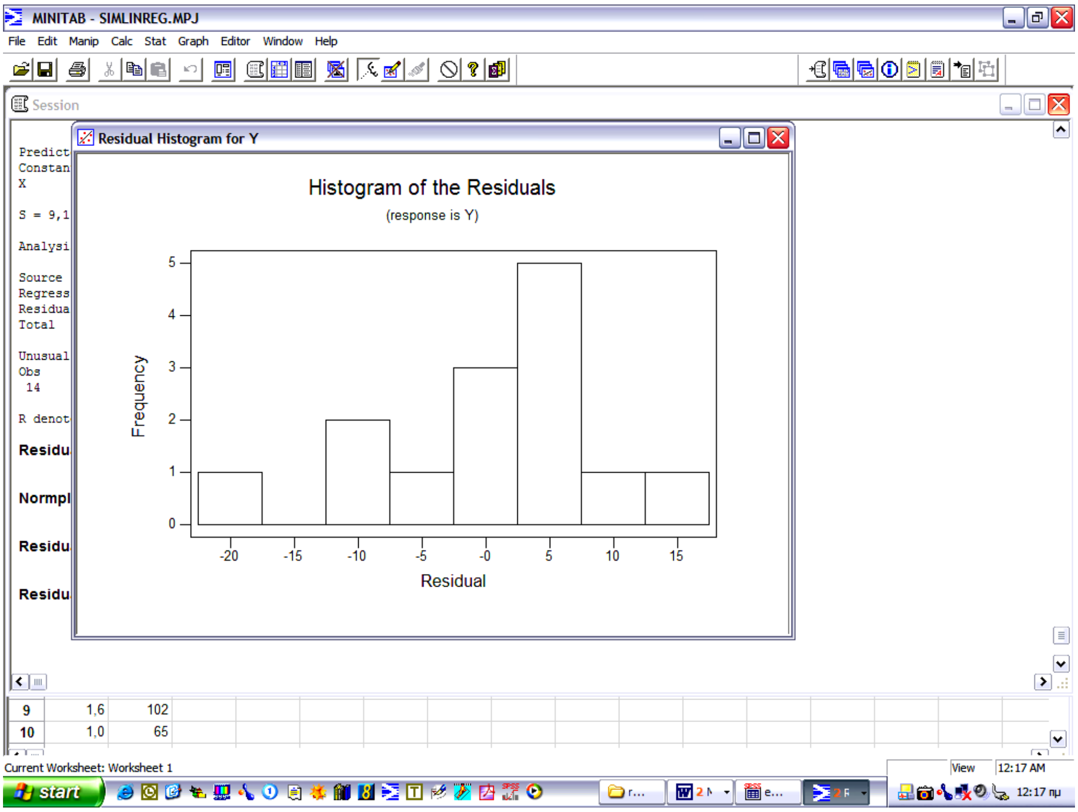
Η εμφάνιση των αποτελεσμάτων που ακολουθεί φαίνεται στα παρακάτω σχήματα 1.21, 1.22, 1.23, 1.24 και 1.25



Σχήμα 1.21

- ***Histogram of residuals.***

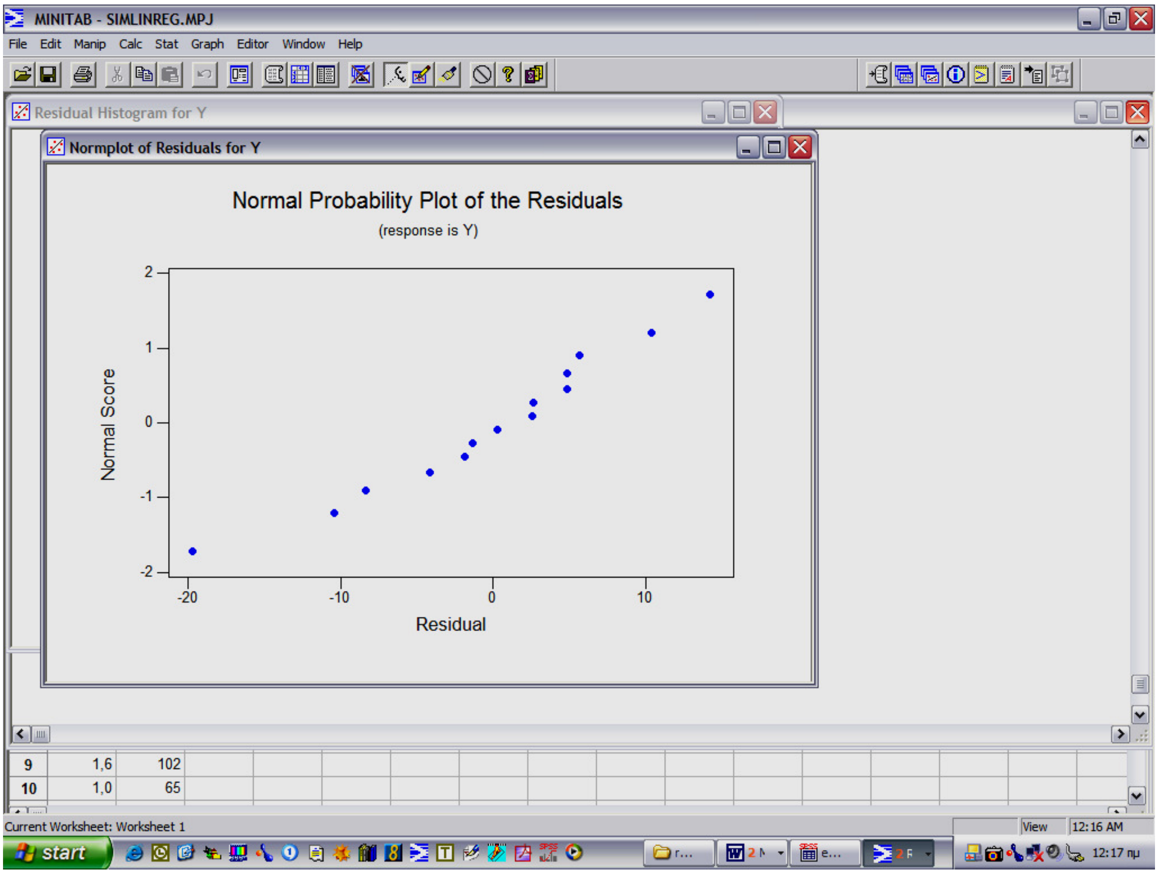
Αν τα υπόλοιπα (residuals) ακολουθούν την κανονική κατανομή, τότε αυτή η γραφική παράσταση παριστάνει μια κανονική κατανομή με μέση τιμή μηδέν.



Σχήμα 1.22

- ***Normal Plot of Residuals.***

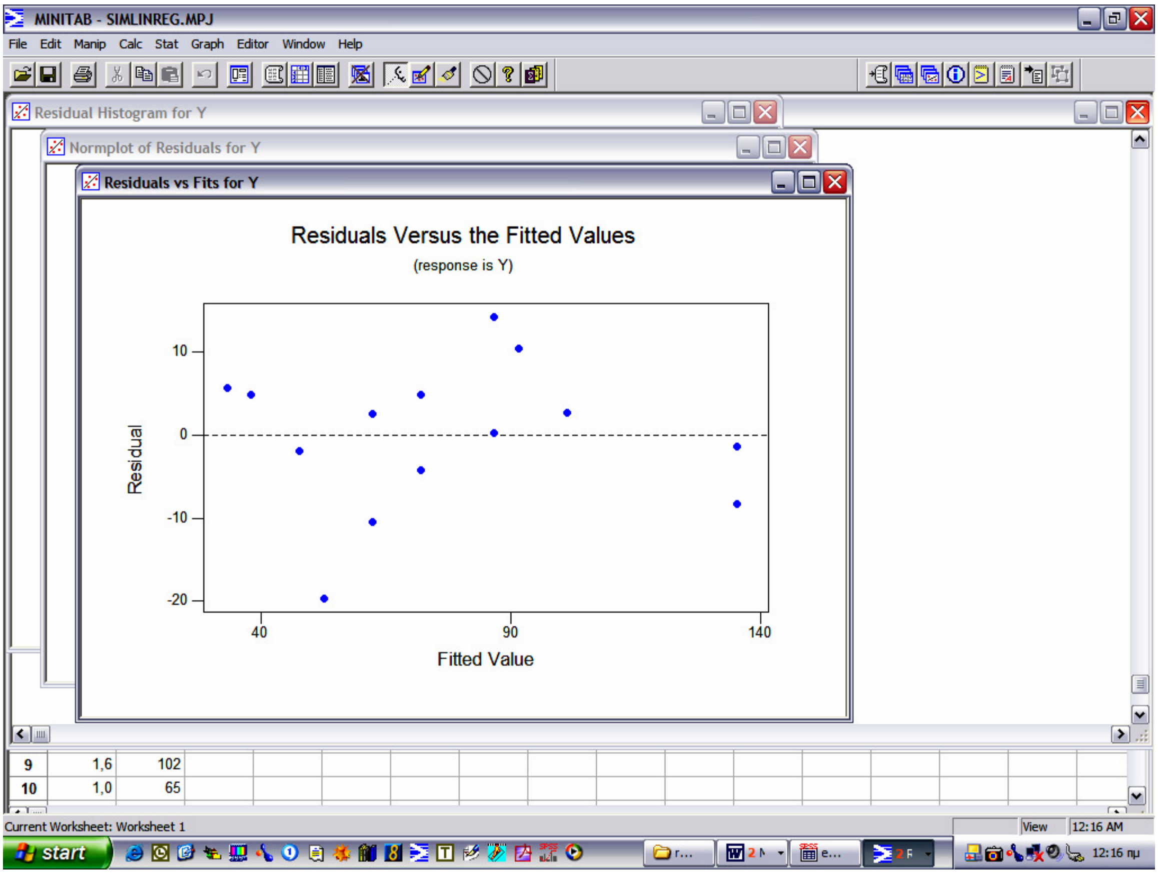
Αν τα υπόλοιπα (residuals) ακολουθούν την κανονική κατανομή, τότε τα σημεία της γραφικής παράστασης πρέπει να κείτονται σε μια ευθεία. Διαφορετικά, η καταλληλότητα του μοντέλου παλινδρόμησης αμφισβητείται.



Σχήμα 1.23

Residuals versus fit.

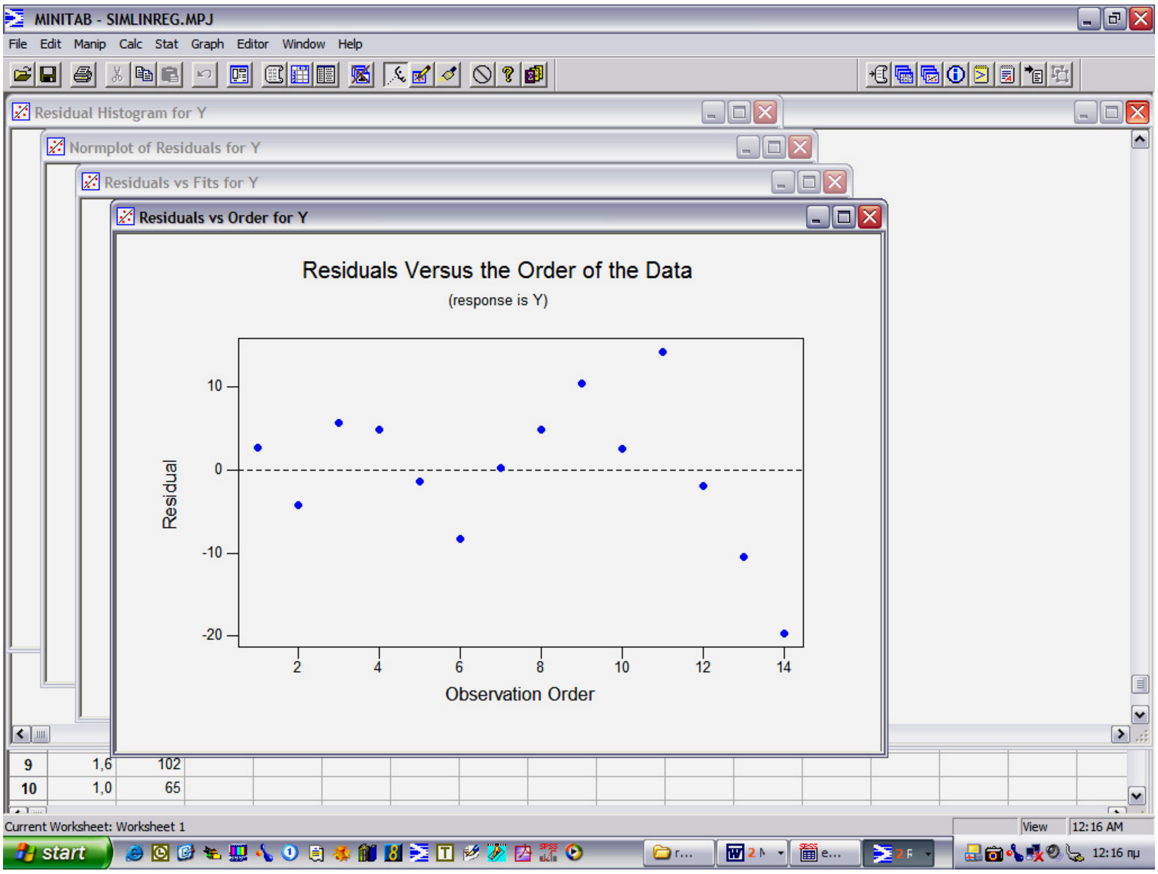
Αναδεικνύει μια τυχαία ομοσκεδαστική μορφή των υπολοίπων.



Σχήμα 1.24

- ***Residuals versus order.***

Χρησιμοποιείται για τον έλεγχο της αυτοσυσχέτισης. (Εδώ δεν παρατηρείται αυτοσυσχέτιση).

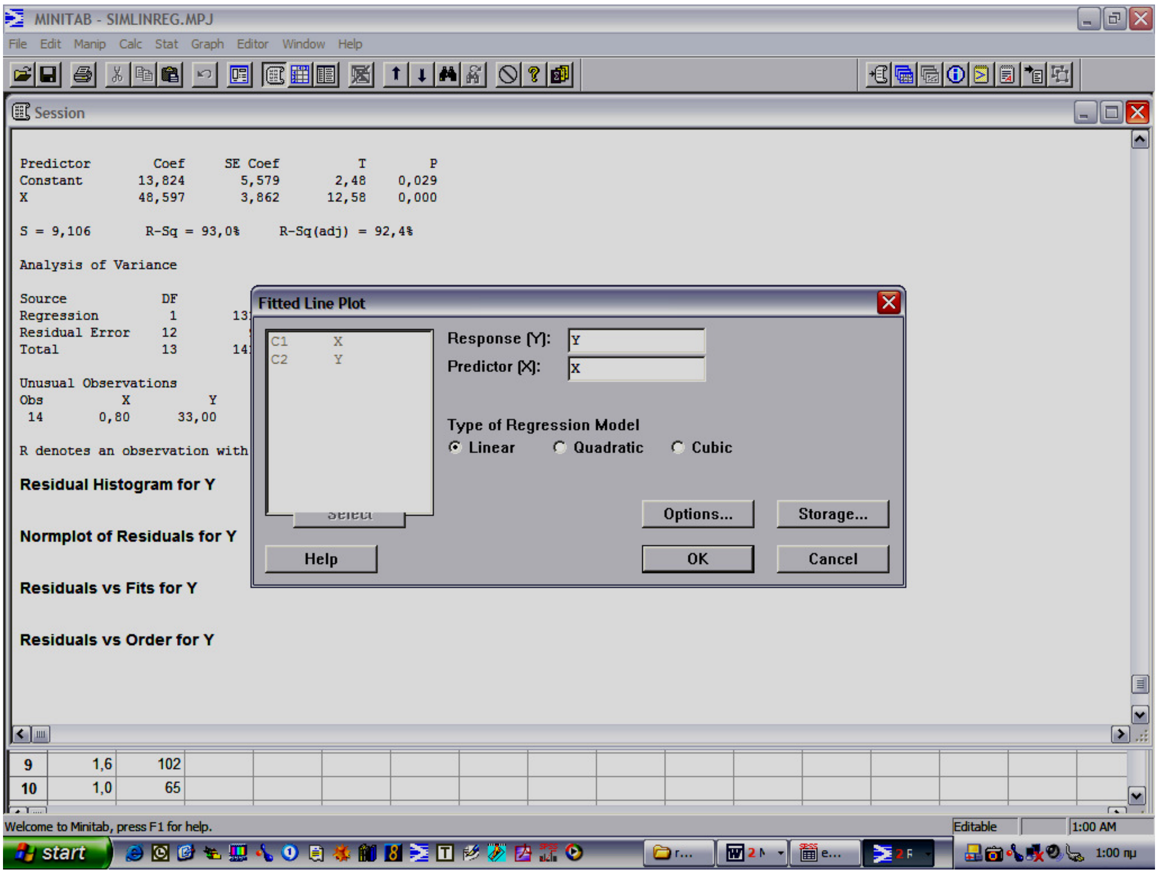


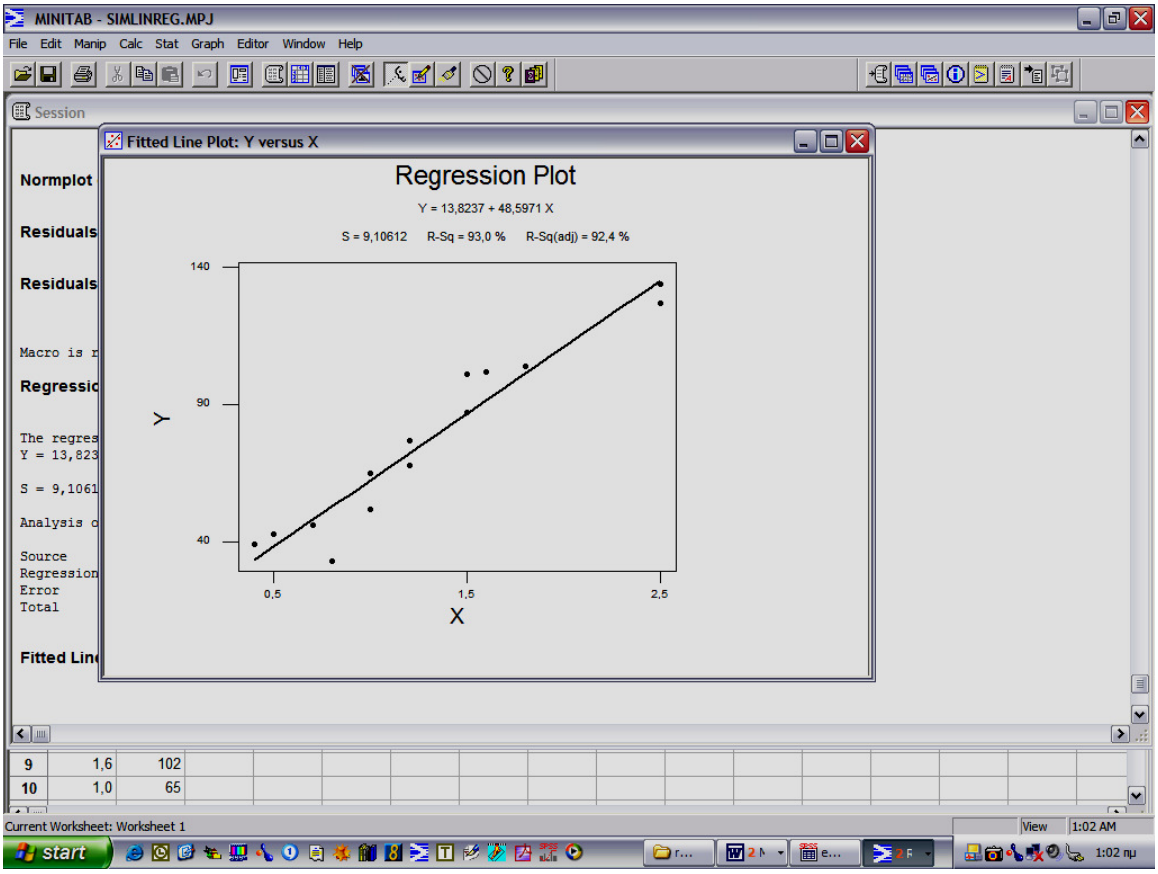
Σχήμα 1.25

Οι παραπάνω είναι οι πιο βασικές γραφικές παραστάσεις, οι οποίες μας παρέχουν πληροφορίες για το προσαρμοσμένο μοντέλο.

Παρατήρηση

Επίσης από την επιλογή ***Stat->Regression->Fitted Line Plot*** είναι δυνατόν η προσαρμογή ενός γραμμικού μοντέλου ή ενός πολυωνυμικού μοντέλου που περιέχει τους όρους X^2 και X^3. Επιπλέον, είναι εφικτή η μετατροπή των δεδομένων σε άλλη μορφή π.χ. λογαριθμική και η ανάλυση της παλινδρόμησης να πραγματοποιηθεί σε αυτή τη μορφή.

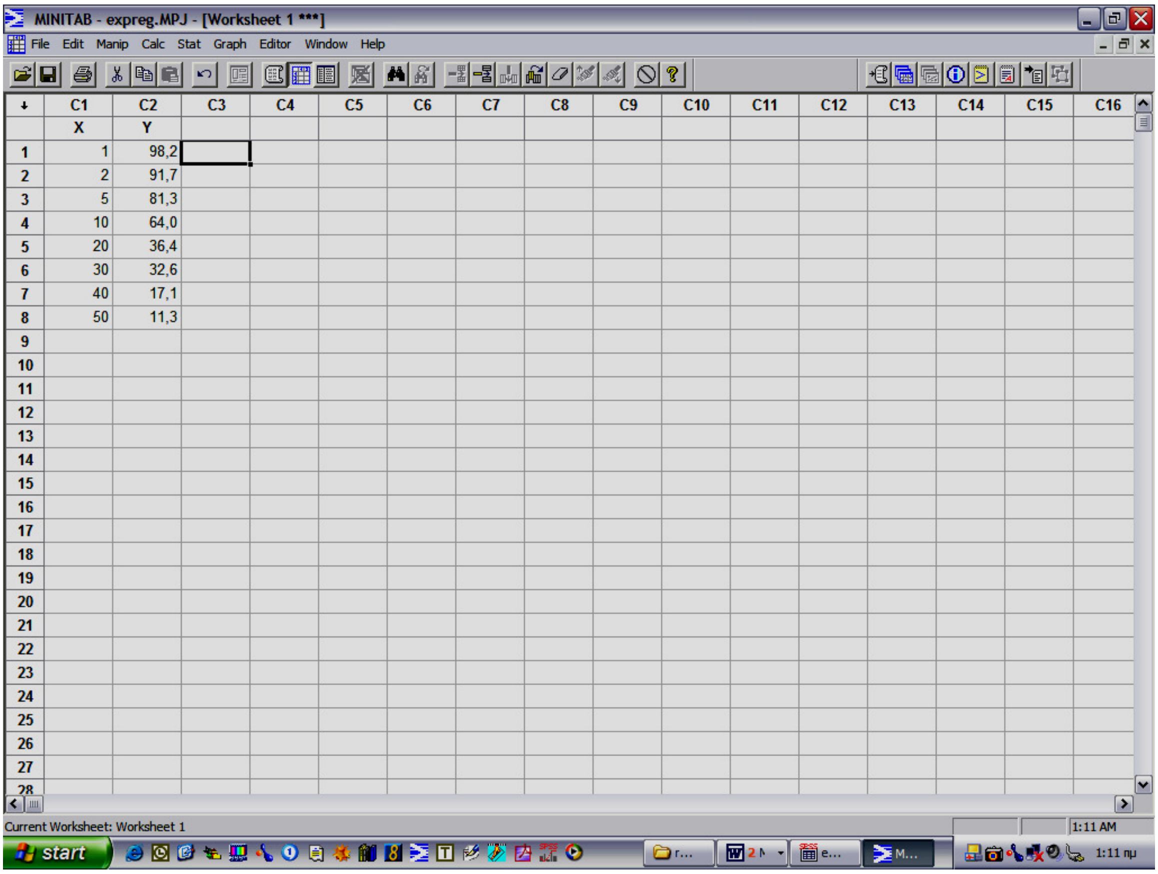

MINITAB - SIMLINREG.MPJ
Fitted Line Plot
Response [Y]: Y
Predictor [X]: X
Type of Regression Model
Linear
Quadratic
Cubic
Options...
Storage...
Help
OK
Cancel
Predictor Coef SE Coef T P
Constant 13,824 5,579 2,48 0,029
X 48,597 3,862 12,58 0,000
S = 9,106 R-Sq = 93,0% R-Sq(adj) = 92,4%
Analysis of Variance
Residual Histogram for Y
Normplot of Residuals for Y
Residuals vs Fits for Y
Residuals vs Order for Y


MINITAB - SIMLINREG.MPJ
Fitted Line Plot: Y versus X
Regression Plot
Y = 13,8237 + 48,5971 X
S = 9,10612 R-Sq = 93,0 % R-Sq(adj) = 92,4 %
140
90
40
Y
0,5
1,5
2,5
X

Σχήμα 1.26

1.2.2 Απλή Εκθετική Παλινδρόμηση

Έστω ότι έχουμε τα δεδομένα του παρακάτω πίνακα:

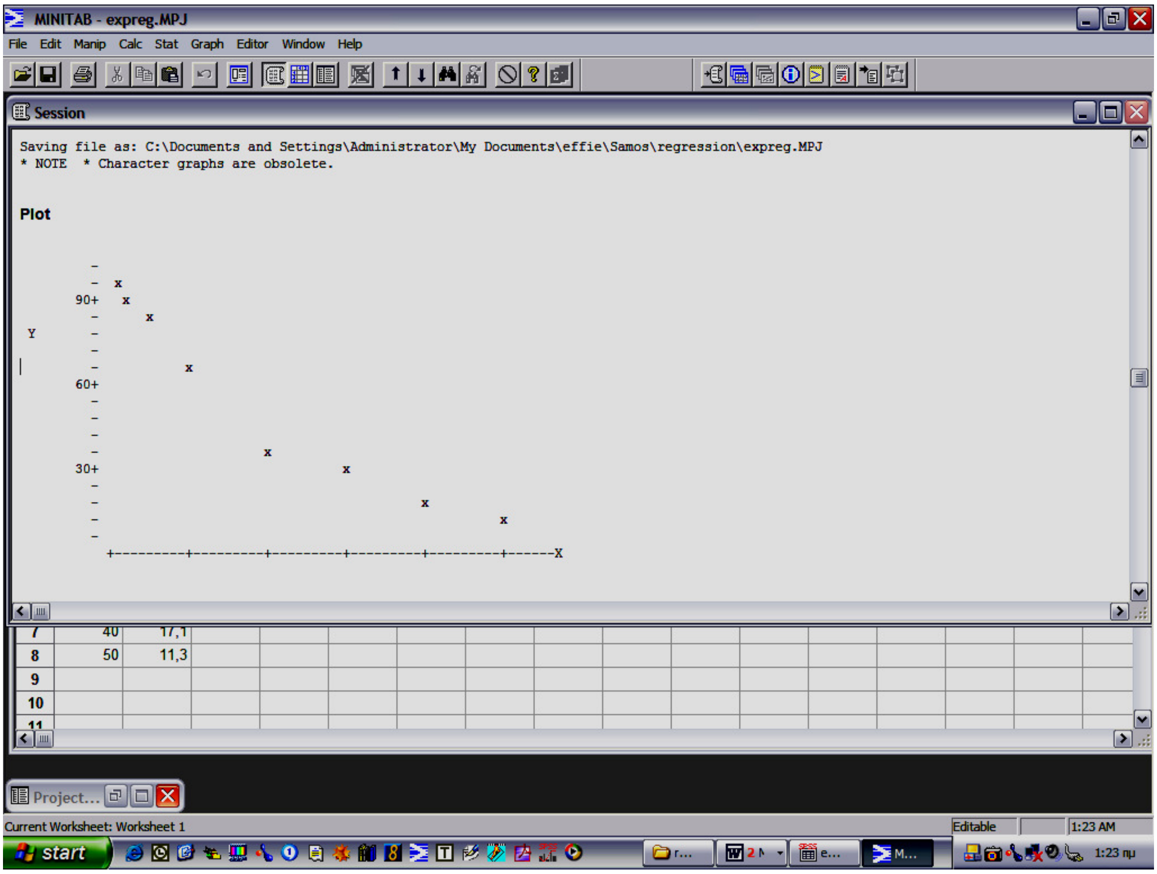


Ζητείται το διάγραμμα διασποράς των δύο μεταβλητών και να προσαρμοστεί το κατάλληλο μοντέλο.

<u>Λύση</u>

Άνοιγμα του αρχείου *Expreg*.

Από το **menu *Graph*** επιλέγουμε ***Character Graphs*** και ***Scatter Plot*** σχεδιάζουμε το διάγραμμα διασποράς για το Y και X.



Σχήμα 1.27

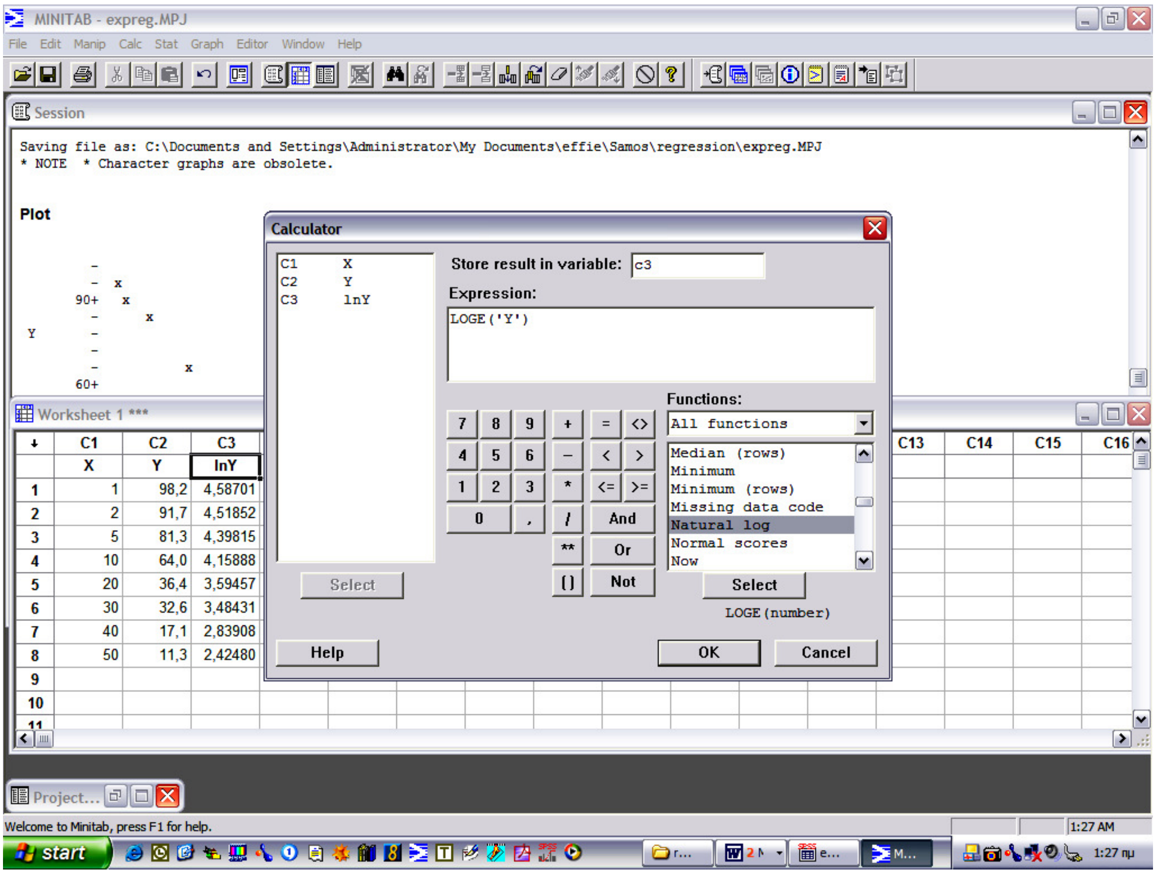
Παρατηρούμε ότι η σχέση μεταξύ τους δεν είναι γραμμική αλλά εκθετική. Άρα το μοντέλο που θα προσαρμοστεί θα είναι της μορφής:

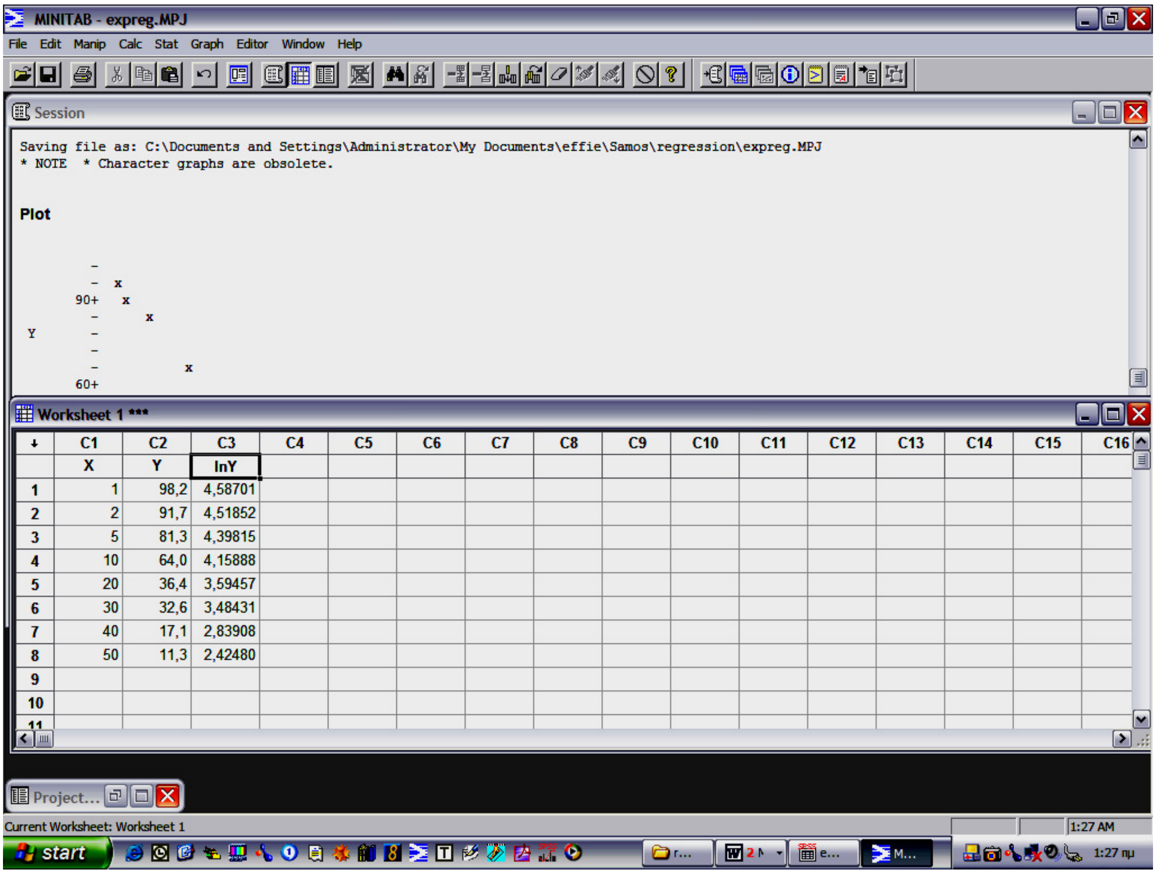
$$y=\alpha\beta^{x}$$

Λογαριθμίζοντας αυτή έχουμε:

$$\ln y=\ln\alpha + x\ln\beta$$

Άρα από την επιλογή ***Calc->Calculator*** γίνεται η λογαριθμική μετατροπή.

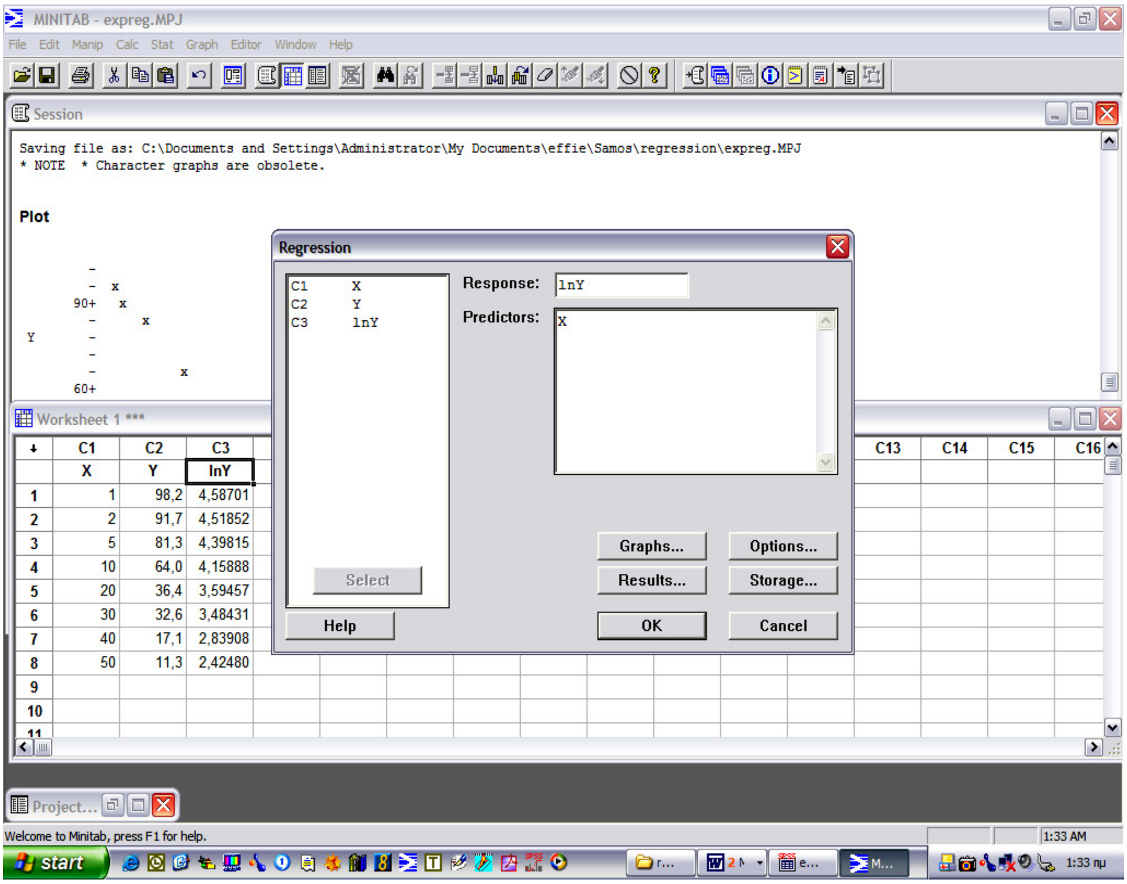


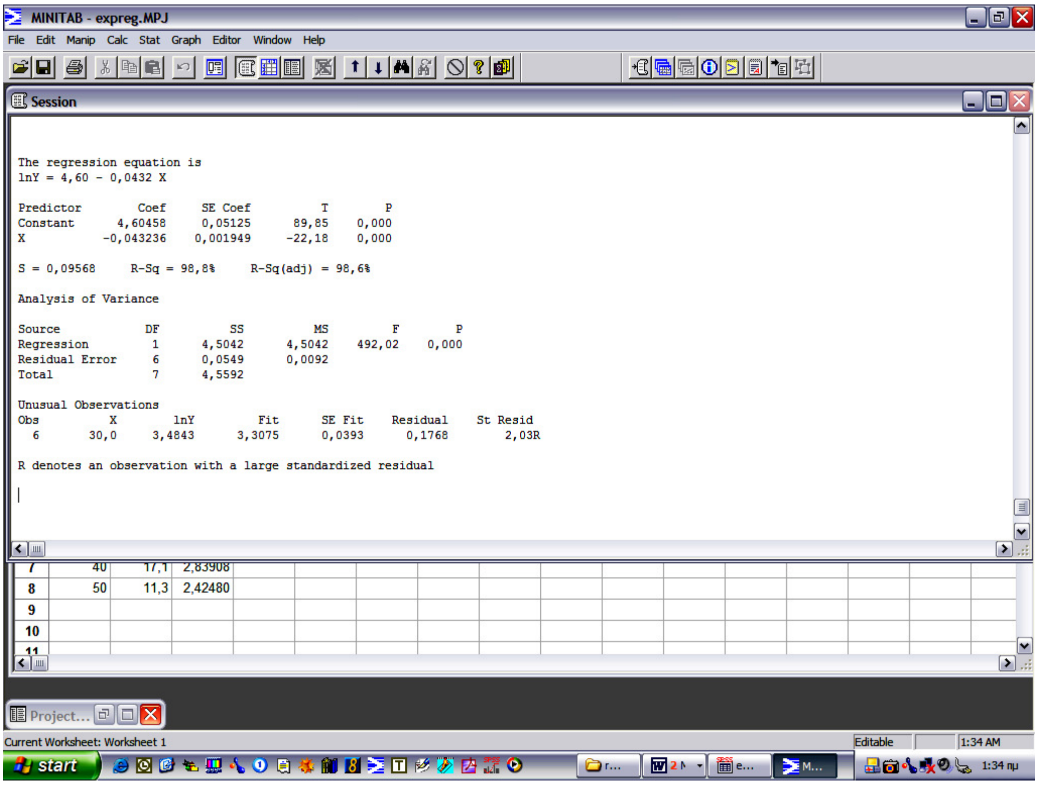


Σχήμα 1.28

Στη συνέχεια από ***Stat->Regression->Regression*** επιλέγουμε ***Response lny*** και ***Predictors***

x. Παίρνουμε τα κάτωθι αποτελέσματα:





Συνεπώς έχουμε:

$\ln\hat{\alpha} = 4.60 \Rightarrow \hat{\alpha} = 99.484$

και

$\ln\hat{\beta} = -0.0432 \Rightarrow \hat{\beta} = 0.9577$