Στατιστική Μοντελοποίηση

- **ΔΠΜΣ Επιστήμη Δεδομένων & Μηχανική Μάθηση**
- **Ορφανουδάκης Φίλιππος Σκόβελεφ**
- **ΣΕΙΡΑ 1**

A)

1)

Για το απλο γραμμικό μοντέλο έχω:

- $R^2 = \frac{SSR}{SST}$ και $r^2_{xy} = \frac{S^2_{xy}}{S_{xx} S_{yy}}$.

Ισχύει ότι $SST = S_{yy}$ (1) και

$SSR = \hat{\beta}_1^2 S_{xx}$ (2) και

$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \Rightarrow \hat{\beta}_1^2 = \frac{S^2_{xy}}{S^2_{xx}}$ (3)

Από (2),(3) $\Rightarrow$ $SSR = \frac{S^2_{xy}}{S_{xx}}$ (4)

Από (1),(4) $\Rightarrow$ $\frac{SSR}{SST} = \frac{S^2_{xy}}{S_{xx} S_{yy}} \Rightarrow R^2 = r^2_{xy}$

2)

Για το απλο γραμμικο μοντελο ακολουθουμε την μεθοδο ελαχιστων τετραγωνων η την μεθοδο μεγιστης πιθανοφανειας για την ευρεση των παραμετρων του μοντελου.

Συμφωνα με την μεθοδο ελαχιστων τετραγωνων, εχουμε οτι:

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - E(y_i))^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

Και παραγωγιζουμε ως προς β_0:

$$\frac{\partial}{\partial \beta_0} S(\beta_0, \beta_1)\Big|_{(\beta_0=\hat{\beta_0}, \beta_1=\hat{\beta_1})} = -2 \sum_{i=1}^{n} (y_i - \hat{\beta_0} - \hat{\beta_1} x_i)$$

και θετω αυτη τη παραγωγο ιση με μηδεν:

$$\sum_{i=1}^{n} (y_i - \hat{\beta_0} - \hat{\beta_1} x_i) = 0 \Rightarrow \sum_{i=1}^{n} y_i - \sum \hat{y}_i = 0$$

$$\Rightarrow \sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{y}_i$$

3)

$$\hat{\beta}_1 = \frac{\sum_1^n x_i y_i - n\bar{x}\bar{y}}{\sum_1^n x_i^2 - n\bar{x}^2}$$

$$= c \cdot \left(\sum_1^n x_i y_i - \bar{x} n \bar{y} \right) = c \cdot \left(\sum_1^n x_i y_i - \bar{x} \sum_1^n y_i \right)$$

$$= c \cdot \sum_1^n (x_i - \bar{x}) y_i$$

Και, $\bar{y} = \frac{1}{n} \sum_1^n y_i$

Άρα,

$$\text{cov}(\bar{y}, \hat{\beta}_1) = \frac{1}{n} \cdot c \cdot \text{cov}\left(\sum_1^n y_i, \sum_1^n (x_i - x) y_i \right)$$

$$= \frac{c}{n} \sum_1^n (x_i - \bar{x}) \cdot \text{cov}(y_i, y_j)$$

$$= \frac{c}{n} \sum_1^n (x_i - \bar{x}) \cdot \text{Var}(y_i) = \frac{c \cdot \sigma^2}{n} \sum_1^n (x_i - \bar{x})$$

Όμως $\sum_1^n (x_i - \bar{x}) = n\bar{x} - n\bar{x} = 0$

4)

Έχουμε ότι $\sum_{1}^{n} (\hat{y}_i - \bar{y}_i)(y_i - \hat{y}_i) =$

$$\underbrace{\sum \hat{y}_i (y_i - \hat{y}_i)}_{(1)} - \underbrace{\bar{y}_i \sum (y_i - \hat{y}_i)}_{(2)}$$

Για την 2 έχουμε αποδείξει στο ερώτημα 2 ότι ισούται με μηδέν.

(1): Θα χρησιμοποιήσουμε την μέθοδο ελαχίστων τετραγώνων.

$$\frac{\partial S(\beta_0, \beta_1)}{\partial \beta_1}\Big|_{\beta_0 = \hat{\beta}_0, \beta_1 = \hat{\beta}_1} = -\sum 2 (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i = 0$$

$$\Rightarrow \sum x_i (y_i - \hat{y}_i) = 0 \,, \quad x_i = \frac{1}{\hat{\beta}_1} \hat{y}_i - \frac{\hat{\beta}_0}{\hat{\beta}_1}$$

Άρα προκύπτει: $\frac{1}{\hat{\beta}_1} \underbrace{\sum \hat{y}_i (y_i - \hat{y}_i)}_{(1)} - \frac{\hat{\beta}_0}{\hat{\beta}_1} \underbrace{\sum (y_i - \hat{y}_i)}_{(2)} = 0$

(2) = 0 Άρα (1) = 0

Άρα έχουμε το ζητούμενο

5)

Στο απλό γραμμικό μοντέλο, μια από τις υποθέσεις μας είναι ότι η y_i ακολουθεί κανονική κατανομή, άρα

το SSR: $\sum (y_i - \hat{y}_i)^2$ ακολουθεί χ^2 κατανομή με $(n-2)$ βαθμούς ελευθερίας

$$\frac{SSR}{\sigma^2} \sim \chi^2(n-2), \quad E(SSR) = (n-2)\sigma^2$$

οπότε μια αμερόληπτη εκτιμήτρια της σ^2 είναι η

$$S^2_{Y|x} = \frac{SSR}{n-2}.$$

Για το ζητούμενο θέλω να αποδείξω αρχικά ότι

$$SSR = S_{yy} - \frac{S^2_{xy}}{S_{xx}}.$$

$$SSR = \sum^{n} (y_i - \hat{y}_i)^2 = \sum^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

$$= \sum^{n} (y_i - \bar{y} + \hat{\beta}_1 \bar{x} - \hat{\beta}_1 x_i)^2$$

$$= \sum^{n} [(y_i - \bar{y}) - \hat{\beta}_1 (x_i - \bar{x})]^2$$

$$= \sum^{n} (y_i - \bar{y})^2 + \hat{\beta}_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 - 2\hat{\beta}_1 \sum^{n} (x_i - \bar{x})(y_i - \bar{y})$$

$$= S_{yy} + \hat{\beta}_1^2 S_{xx} - 2\hat{\beta}_1 S_{xy}$$

Έχω ότι $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$

Άρα $SSR = S_{yy} + \frac{S^2_{xy}}{S_{xx}} - 2\frac{S^2_{xy}}{S_{xx}}$

$$= S_{yy} - \frac{S^2_{xy}}{S_{xx}} = S_{yy} - S_{yy} r^2_{xy}$$

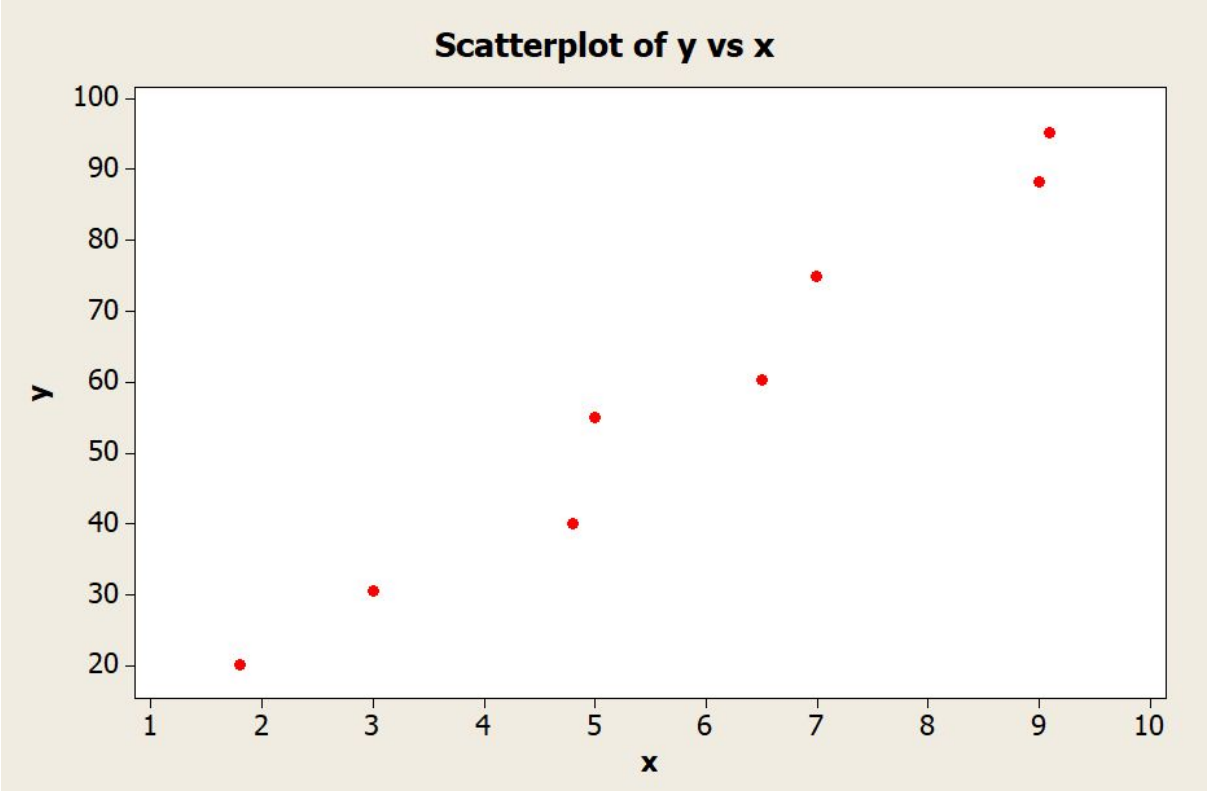
Όριζος

$$r_{xy} = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$$

Β)

Για την εξαγωγή των αποτελεσμάτων κάναμε χρήση του MINITAB.

1)



Από το διάγραμμα διασποράς παρατηρούμε μια έντονη γραμμική σχέση μεταξύ της εξαρτημένης μεταβλητής y και της ανεξάρτητης μεταβλητής x.

2)

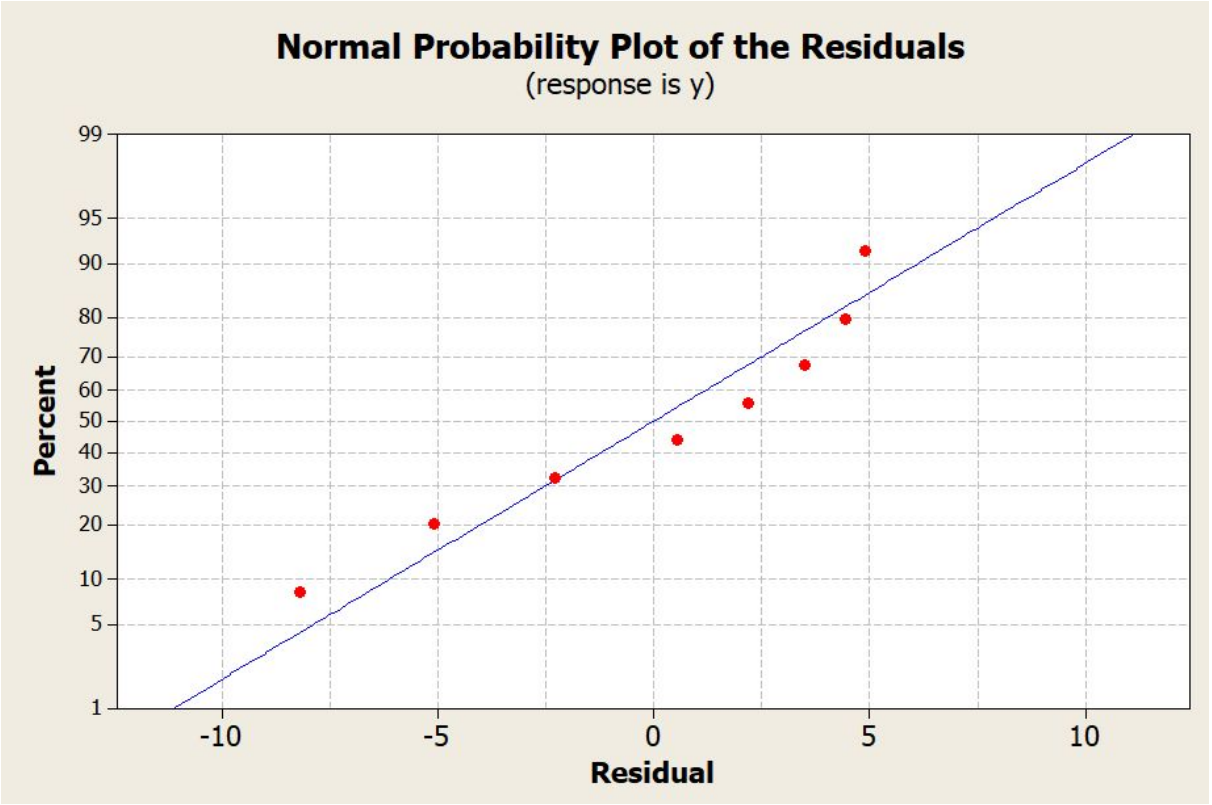
Η εκτίμηση της εξίσωσης παλινδρόμησης :

y = - 0.40 + 10.1 x , δηλαδή β_0 = -0.40 και β_1 = 10.1

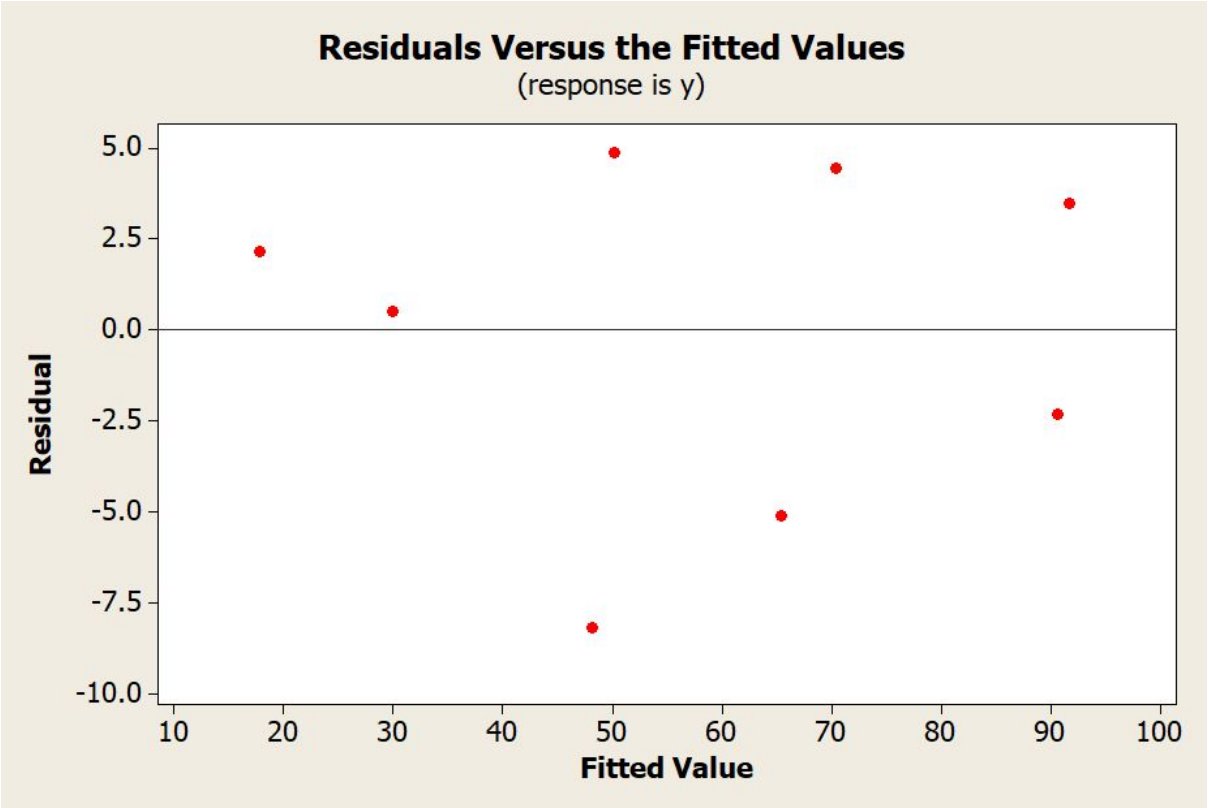
3)

Πραγματοποιούμε τον γραφικό έλεγχο της Κανονικής κατανομής των υπολοίπων.

1. Παρατηρούμε από την γραφική παράσταση ότι τα σημεία τείνουν να βρίσκονται πολύ κοντά σε μια ευθεία, το οποίο σημαίνει ότι τα υπόλοιπα ακολουθούν Κανονική Κατανομή. Αν δεν συνέβαινε αυτό τότε το μοντέλο που έχουμε προσαρμόσει πιθανότατα δεν είναι κατάλληλο για να περιγράψει τα δεδομένα μας.



2. Για να είμαστε σίγουροι ότι το μοντέλο μας είναι κατάλληλο θα πραγματοποιήσουμε ακόμα έναν έλεγχο για τα υπόλοιπα μας. Θα πραγματοποιηθεί ο έλεγχος της ομοσκεδαστικότητας. Θέλουμε στο παρακάτω διάγραμμα την τυχαία και ομοσκεδαστική κατανομή των υπολοίπων γύρω από το μηδέν, το οποίο και συμβαίνει. Έτσι δεχόμαστε την υπόθεση ότι τα τυχαία σφάλματα έχουν κοινή διασπορά $V(\varepsilon_i) = \sigma^2$.



4)

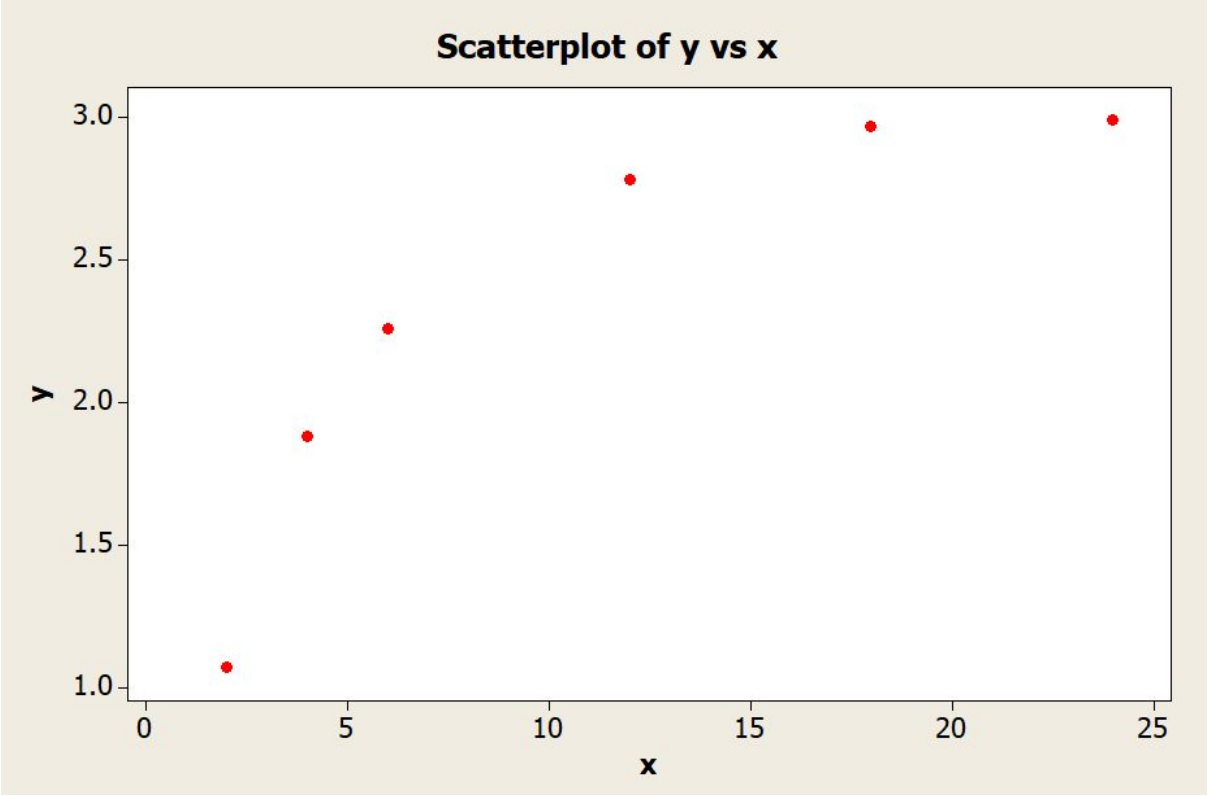
Για x_o = 8 η εκτίμηση για την Y είναι 80.57 και το 95% διάστημα εμπιστοσύνης για την παρατήρηση Y είναι (66.61, 94.53) και για την μέση τιμή E(Y) είναι (74.57, 86.57) και επιβεβαιώνουμε ότι το δ.ε. της μέσης τιμής περιέχεται στο δ.ε. της παρατήρησης.

```
Predicted Values for New Observations

New
Obs    Fit  SE Fit      95% CI          95% PI
  1  80.57    2.45  (74.57, 86.57)  (66.61, 94.53)


Values of Predictors for New Observations

New
Obs     x
  1  8.00
```

Γ)

Για την εξαγωγή των αποτελεσμάτων κάναμε χρήση του MINITAB.

1)



Από το διάγραμμα διασποράς παρατηρούμε ότι δεν προκύπτει κάποια γραμμική σχέση μεταξύ των μεταβλητών x,y και θα πρέπει να πραγματοποιηθεί ένας μετασχηματισμός για να μελετηθεί ένα γραμμικό μοντέλο που να μην παραβιάζονται οι υποθέσεις κανονικότητας.

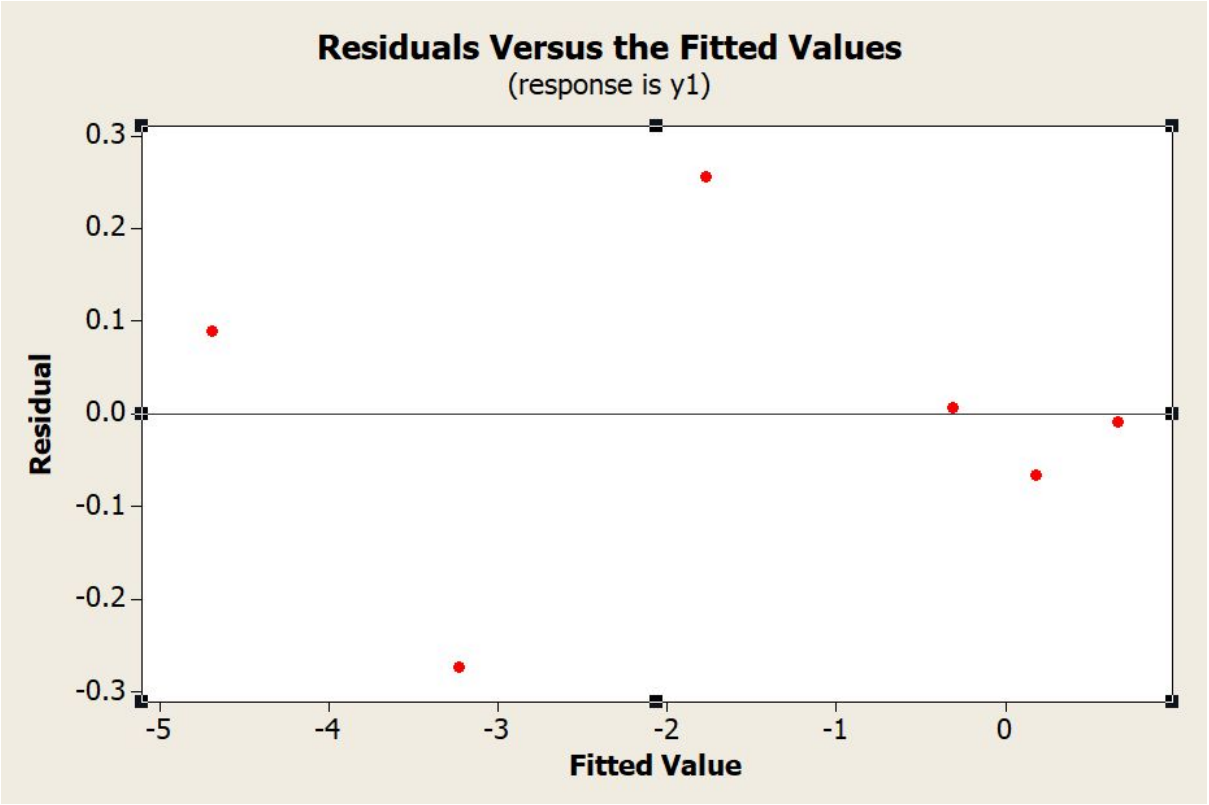
2)

Θεωρούμε ότι η σχέση του y με το x είναι της μορφής :

$Y = 3 - \alpha e^{\beta x}$ συνεπώς πραγματοποιούμε τις εξής μετατροπές :

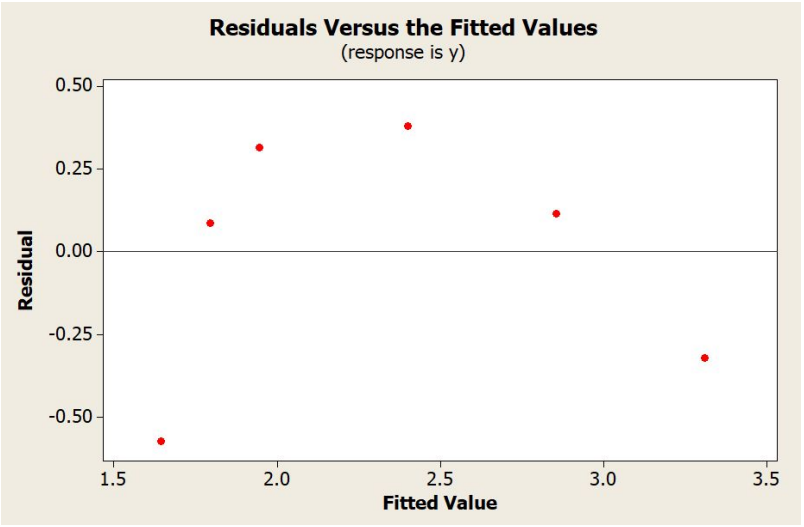
- **$3 - Y = \alpha e^{\beta x}$**
- **$\ln(3 - Y) = \ln\alpha + \beta x$**
- **$Y' = \ln(3-Y)$, $\beta_o = \ln\alpha$, $\beta_1 = \beta$**
- **$Y' = \beta_o + \beta_1 x$**

Και το ζητούμενο διάγραμμα είναι το εξής :



Παρατηρούμε ότι ικανοποιείται ο έλεγχος της ομοσκεδαστικότητας από τον τρόπο που είναι κατανεμημένα τα σημεία στο χώρο.

Αν σχεδιάσουμε το αντίστοιχο διάγραμμα στο αρχικό μοντέλο δεν θα βλέπαμε την ίδια κατανομή, γεγονός που μας κάνει να αμφισβητούμε την καταλληλότητα του.



3)

Για τα ζητούμενα θα πραγματοποιήσουμε την εκτίμηση στο μοντέλο του 2ου ερωτήματος που έχει υποστεί μετασχηματισμό και στη συνέχεια θα πραγματοποιήσουμε αντίστροφο μετασχηματισμό. Τα τελικά αποτελέσματα είναι τα εξής :

- Εκτίμηση άγνωστης μεταβλητης Y : 2.548418765077
- 99% δ.ε. για την για την πρόβλεψη της παρατήρησης Y_{X0} : (1.792804380283, 2.83107492406)