ΔΠΜΣ Επιστήμη Δεδομένων και Μηχανική Μάθηση

Στατιστική Μοντελοποίηση

1η Σειρά Ασκήσεων

Ευάγγελος Τσόγκας

03400120

Ευάγγελος Τσόγκας
ΑΜ: 03400120

1η Σειρά Ασκήσεων

Α)

1) $R^2 = r_{xy}^2$

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2} \quad \text{και} \quad r_{xy} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 \sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

Ισχύει ότι $\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2 = SSR = \hat{b}_1^2 \sum_{i=1}^{n}(x_i-\bar{x})^2 \quad (1)$

όπου με τη μέθοδο των ελαχίστων τετραγώνων βρίσκουμε ότι:

$$\hat{b}_1 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \quad (2)$$

Άρα έχουμε ότι:

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2} \overset{(1)}{=} \frac{\hat{b}_1^2 \sum_{i=1}^{n}(x_i-\bar{x})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2} \overset{(2)}{=}$$

$$= \frac{\left[\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})\right]^2 \cancel{\sum_{i=1}^{n}(x_i-\bar{x})^2}}{\left[\sum_{i=1}^{n}(x_i-\bar{x})^2\right]^{\cancel{2}} \sum_{i=1}^{n}(y_i-\bar{y})^2} = \frac{\left[\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})\right]^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2 \sum_{i=1}^{n}(y_i-\bar{y})^2} =$$

$$= r_{xy}^2$$

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

Η εκτιμημένη συνάρτηση παλινδρόμησης είναι η:

$$\hat{y}_i = \hat{b}_0 + \hat{b}_1 x_i \quad (1) \quad , i=1,\dots,n$$

Μία από τις κανονικές εξισώσεις στις οποίες καταλήγουμε με τη μέθοδο των ελαχίστων τετραγώνων είναι η:

$$n\hat{b}_0 + \hat{b}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \quad (2)$$

Άρα, $\sum_{i=1}^{n} \hat{y}_i = \sum_{i=1}^{n} (\hat{b}_0 + \hat{b}_1 x_i) = \sum_{i=1}^{n} \hat{b}_0 + \sum_{i=1}^{n} \hat{b}_1 x_i = n\hat{b}_0 + \hat{b}_1 \sum_{i=1}^{n} x_i =$

$= \sum_{i=1}^{n} y_i$ //

3) $cov(\bar{y}, \hat{b}_1) = 0$

Η εκτιμήτρια $\hat{b}_1$ δίνεται από τη σχέση:

$$\hat{b}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) y_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) y_i}{S_{xx}}$$

Άρα, $cov(\bar{y}, \hat{b}_1) = cov\left(\frac{\sum_{i=1}^{n} y_i}{n}, \frac{\sum_{j=1}^{n} (x_j - \bar{x}) y_j}{S_{xx}} \right) = \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{(x_j - \bar{x})}{n S_{xx}} cov(y_i, y_j)$

αλλά, $cov(y_i, y_j) = 0 \quad \forall i \neq j$, άρα:

$$cov(\bar{y}, \hat{b}_1) = \sum_{i=1}^{n} \frac{(x_i - \bar{x}) V(y_i)}{n S_{xx}} = \sum_{i=1}^{n} \frac{(x_i - \bar{x}) \sigma^2}{n S_{xx}} = \frac{\sigma^2}{n S_{xx}} \underbrace{\sum_{i=1}^{n} (x_i - \bar{x})}_{0} = 0$$

//

4) $\sum_{i=1}^{n} (y_i-\hat{y}_i)(\hat{y}_i-\bar{y})=0$

Από τις κανονικές εξισώσεις προκύπτουν οι περιορισμοί:

$$\sum_{i=1}^{n}(y_i-\hat{y}_i)=0 \quad (1) \qquad \text{και} \qquad \sum_{i=1}^{n} x_i(y_i-\hat{y}_i)=0 \quad (2)$$

Άρα, $\sum_{i=1}^{n}(y_i-\hat{y}_i)(\hat{y}_i-\bar{y}) = \sum_{i=1}^{n}(y_i-\hat{y}_i)\hat{y}_i - \bar{y}\cancelto{0\,(1)}{\sum_{i=1}^{n}(y_i-\hat{y}_i)} =$

$= \sum_{i=1}^{n}(y_i-\hat{y}_i)\hat{y}_i = \sum_{i=1}^{n}(y_i-\hat{y}_i)(\hat{b}_0+\hat{b}_1 x_i) =$

$= \hat{b}_0\cancelto{0\,(1)}{\sum_{i=1}^{n}(y_i-\hat{y}_i)} + \hat{b}_1\cancelto{0\,(2)}{\sum_{i=1}^{n}x_i(y_i-\hat{y}_i)} = 0$

5) $S^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 = \frac{1}{n-2}\left\{S_{yy}-\frac{S^2_{xy}}{S_{xx}}\right\} = \frac{S_{yy}}{n-2}\left\{1-r^2_{xy}\right\}$

Αρκεί να δείξω ότι: α) $S_{yy}(1-r^2_{xy}) = S_{yy}-\frac{S^2_{xy}}{S_{xx}}$

β) $S_{yy}(1-r^2_{xy}) = \sum_{i=1}^{n}(y_i-\hat{y}_i)^2$

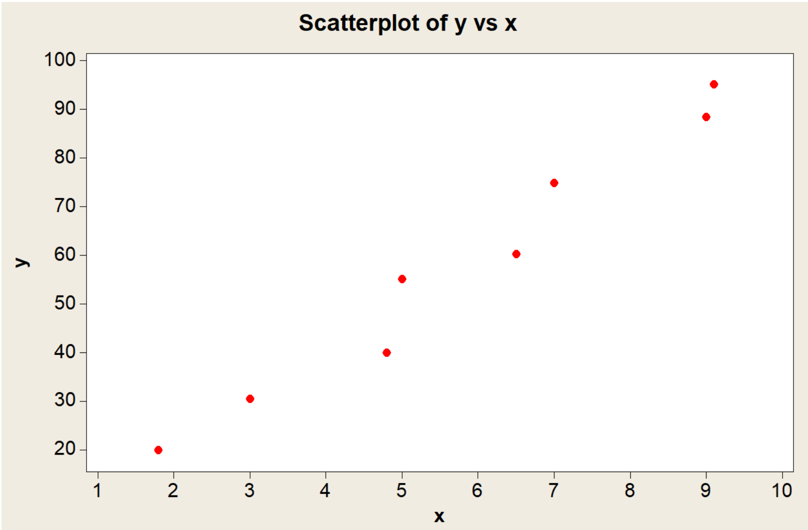
α) $S_{yy}(1-r^2_{xy}) = S_{yy}-S_{yy}r^2_{xy} = S_{yy}-S_{yy}\left(\frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}\right)^2 = S_{yy}-\cancel{S_{yy}}\frac{S^2_{xy}}{S_{xx}\cancel{S_{yy}}} = S_{yy}-\frac{S^2_{xy}}{S_{xx}}$

β) $S_{yy}(1-r^2_{xy}) = S_{yy}-S_{yy}r^2_{xy} = S_{yy}-S_{yy}R^2 = \sum_{i=1}^{n}(y_i-\bar{y})^2 - \cancel{\sum_{i=1}^{n}(y_i-\bar{y})^2}\cdot\frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{\cancel{\sum_{i=1}^{n}(y_i-\bar{y})^2}} =$

$= \sum_{i=1}^{n}(y_i-\bar{y})^2 - \sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2 = SST-SSR = SSE = \sum_{i=1}^{n}(y_i-\hat{y}_i)^2$

B

1) Διάγραμμα διασποράς των μεταβλητών y και x.



Από τη μορφή του διαγράμματος διασποράς παρατηρούμε ότι υπάρχει μια ισχυρή γραμμική σχέση ανάμεσα στις μεταβλητές y και x.

2) Εκτίμηση της εξίσωσης παλινδρόμησης $E(y|x) = \beta_0 + \beta_1 x$.

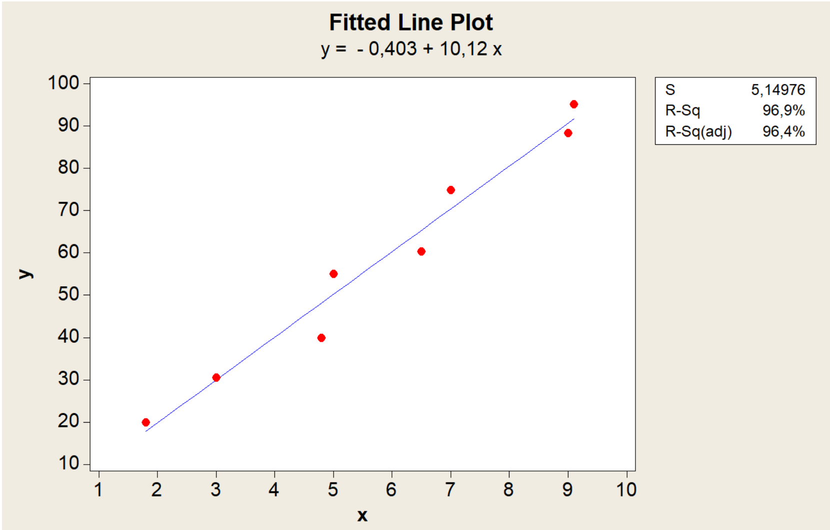
```
The regression equation is
y = - 0,40 + 10,1 x


Predictor      Coef  SE Coef      T      P
Constant     -0,403    4,641  -0,09  0,934
x           10,1218   0,7392  13,69  0,000


S = 5,14976   R-Sq = 96,9%   R-Sq(adj) = 96,4%


Analysis of Variance

Source          DF      SS      MS       F      P
Regression       1  4972,4  4972,4  187,50  0,000
Residual Error   6   159,1    26,5
Total            7  5131,5
```

Η προσαρμοσμένη συνάρτηση παλινδρόμησης στα δεδομένα μας είναι η:

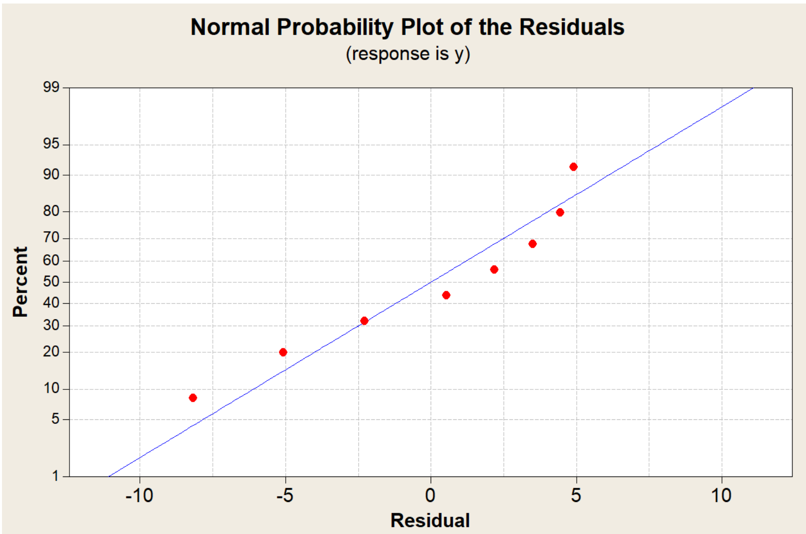
$$\hat{y} = -0.40 + 10.1x$$

άρα $\hat{\beta}_0$ = - 0.40 και $\hat{\beta}_1$ = 10.1.

Επιπλέον, η προσαρμοσμένη συνάρτηση φαίνεται και στο παρακάτω διάγραμμα:



3) Γραφικός έλεγχος Κανονικής κατανομής για τα υπόλοιπα.



Τα σημεία της γραφικής παράστασης κείτονται σε μια ευθεία, άρα συμπεραίνουμε ότι τα υπόλοιπα ακολουθούν την Κανονική κατανομή και επομένως ικανοποιείται η υπόθεση της κανονικότητάς τους.

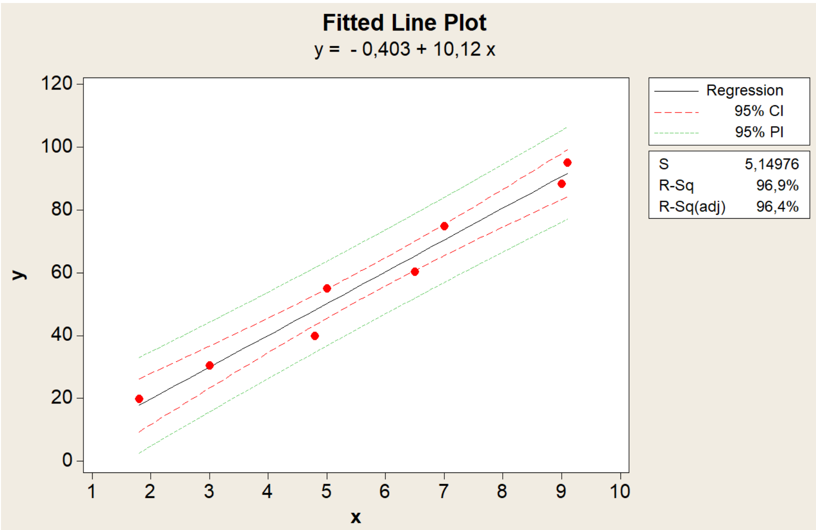
4)

```
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
```

Παρατηρούμε ότι:

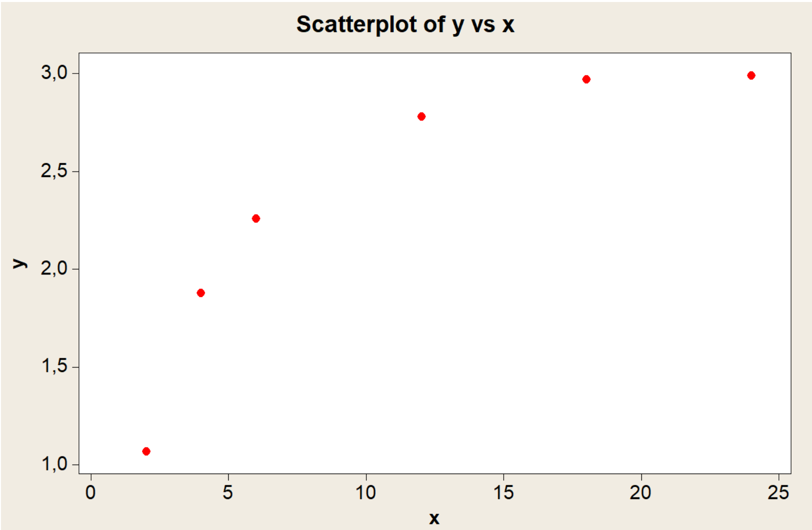
- Για $x_o = 8$ η ενεργειακή κατανάλωση είναι $\mathbf{y_{x0}} = \mathbf{80.57}$.
- Το 95% δ.ε. της πρόβλεψης για την παρατήρηση y_{x0} είναι το **(66.61, 94.53)**.
- Το 95% δ.ε. για τη μέση τιμή της y_{x0} είναι το **(74.57, 86.57)**.

Στο παρακάτω διάγραμμα φαίνονται και γραφικά τα διαστήματα εμπιστοσύνης της y (PI) και της E(y) (CI) :



Γ

1) Διάγραμμα διασποράς μεταξύ y και x.



Από τη μορφή του διαγράμματος διασποράς παρατηρούμε ότι δεν προκύπτει γραμμική σχέση μεταξύ των y και x.

2) Όπως καταλαβαίνουμε από το διάγραμμα διασποράς θα πρέπει να μετασχηματίσουμε τη συνάρτηση κατάλληλα ώστε να αποκτήσουμε ένα γραμμικό μοντέλο ως εξής:

$$y = 3 - ae^{\beta x} \Rightarrow 3 - y = ae^{\beta x} \Rightarrow \ln(3-y) = \ln\left(ae^{\beta x}\right) \Rightarrow \mathbf{ln(3-y)} = \boldsymbol{lna + \beta x}$$

άρα, $\boldsymbol{y^* = \ln(3-y)}$ $\boldsymbol{\kappa\alpha\iota\ \beta_0 = \ln(a)}$

Προσαρμόζοντας το μοντέλο προκύπτουν τα εξής:

```
The regression equation is
y* = 1,15 - 0,244 x
```

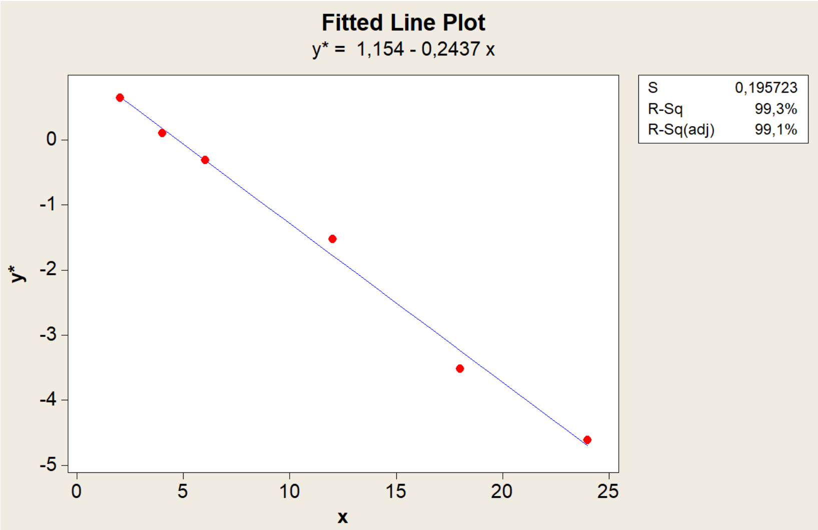
Predictor	Coef	SE Coef	T	P
Constant	1,1544	0,1370	8,42	0,001
x	-0,24367	0,01012	-24,08	0,000

```
S = 0,195723   R-Sq = 99,3%   R-Sq(adj) = 99,1%
```

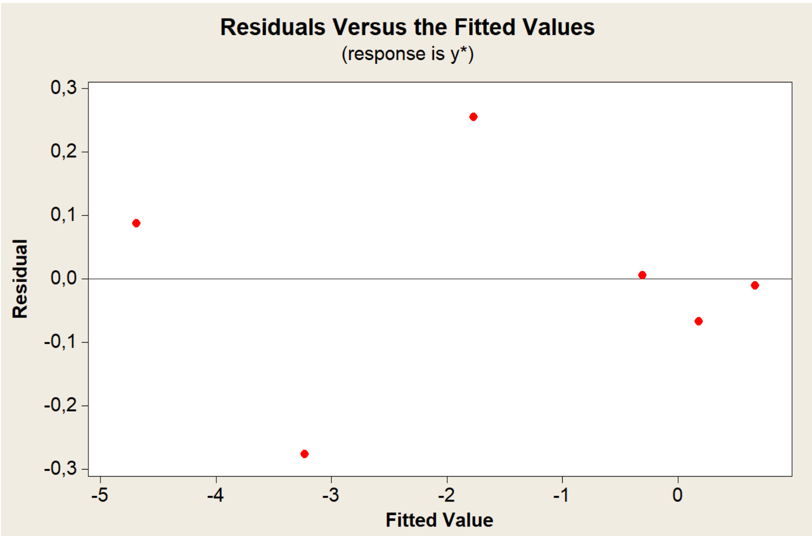
Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	22,206	22,206	579,68	0,000
Residual Error	4	0,153	0,038		
Total	5	22,360			

Άρα, $\hat{y}^* = \mathbf{1.15 - 0.244x}$ η οποία φαίνεται και στο παρακάτω διάγραμμα από το οποίο παρατηρούμε ότι πλέον η σχέση των y* και x είναι γραμμική:



Παρακάτω φαίνεται και η γραφική παράσταση των υπολοίπων e επί των εκτιμηθέντων $\hat{y}^*$ από την οποία παρατηρούμε ότι η διασπορά των υπολοίπων είναι πολύ μεγαλύτερη για μικρές τιμές.



3)

```
Predicted Values for New Observations

New
Obs      Fit   SE Fit         99% CI                 99% PI
  1  -0,7950   0,0855   (-1,1886; -0,4015)   (-1,7783; 0,1883)


Values of Predictors for New Observations

New
Obs     x
  1  8,00
```

Παρατηρούμε πως για $x_0 = 8$:

- Η άγνωστη παρατήρηση είναι $\boldsymbol{y}^*$ **= -0.7950.**
- Το 99% δ.ε. της πρόβλεψης για την παρατήρηση $\boldsymbol{y}^*_{x0}$ είναι το **(-1.7783, 0.1883)**.
- Το 99% δ.ε. για τη μέση τιμή της $\boldsymbol{y}^*_{x0}$ είναι το **(-1.1886, -0.4015)**.

Αντίστοιχα, για το αρχικό μοντέλο με τον αντίστροφο μετασχηματισμό $-e^{y*} + 3$:

- Η άγνωστη παρατήρηση είναι $\boldsymbol{y}$ **= 2.54842.**
- Το 99% δ.ε. της πρόβλεψης για την παρατήρηση $\boldsymbol{y}_{x0}$ είναι το **(1.79280, 2.83107)**.
- Το 99% δ.ε. για τη μέση τιμή της $\boldsymbol{y}_{x0}$ (προσεγγιστικά) είναι το **(2.33068, 2.69535)**.