Αναπλ. Καθηγήτρια, Ελένη Κανδηλώρου

Αθήνα 10-3-2017

Σημειώσεις

Εκτίμηση των Παραμέτρων β_0 & β_1

Απλό γραμμικό υπόδειγμα:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad (1)$$

Η αναμενόμενη τιμή του Y_i, δηλαδή, μέση τιμή του Y_i, δίνεται παρακάτω:

$$E(Y_i) = \beta_0 + \beta_1 X_i = E(Y_i \mid X_i) = \mu_{Y_i|X_i} \qquad (2)$$

Η εκτίμηση το $E(Y_i)$ είναι:

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i \quad \acute{\eta} \quad \widehat{Y}_i = b_0 + b_1 X_i \qquad (3)$$

Το σφάλμα της εκτίμησης είναι e_i και ισούται με:

$$e_i = Y_i - \widehat{Y}_i \qquad (4)$$

Αναζητούμε τις εκτιμήσεις των β_0 και β_1 που ελαχιστοποιούν τα e_i και συγκεκριμένα ελαχιστοποιούν το άθροισμα των τετραγώνων της e_i, δηλαδή,

$$\min \sum_{i=1}^{n} (Y_i - \widehat{Y})^2 = Q \qquad (5)$$

$$\partial Q \Big/ \partial \widehat{\beta}_0 = -2 \sum_{i=1}^{n} (Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i) = 0 \qquad (6)$$

και

$$\partial Q \Big/ \partial \widehat{\beta}_1 = -2 \sum_{i=1}^{n} (Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i) = 0 \qquad (7)$$

Από την εξίσωση *(6)* έχουμε:

$$\sum_{i=1}^{n} Y_i = n\widehat{\beta}_0 + \widehat{\beta}_1 \sum_{i=1}^{n} X_i$$

$$\sum_{i=1}^{n} Y_i - \widehat{\beta}_1 \sum_{i=1}^{n} X_i = n\widehat{\beta}_0 \Rightarrow \widehat{\beta}_0 = \frac{\sum_{i=1}^{n} Y_i}{n} - \widehat{\beta}_1 \frac{\sum_{i=1}^{n} X_i}{n} = \overline{Y} - \widehat{\beta}_1 \overline{X}$$

Από την εξίσωση *(7)* έχουμε:

$$\sum_{i=1}^{n} Y_i X_i - \hat{\beta}_0 \sum_{i=1}^{n} X_i - \hat{\beta}_1 \sum_{i=1}^{n} X_i^2 = 0$$

$$\sum_{i=1}^{n} Y_i X_i = \hat{\beta}_0 \sum_{i=1}^{n} X_i + \hat{\beta}_1 \sum_{i=1}^{n} X_i^2 = (\bar{Y} - \hat{\beta}_1 \bar{X}) \sum_{i=1}^{n} X_i + \hat{\beta}_1 \sum_{i=1}^{n} X_i^2 \Rightarrow$$

$$\hat{\beta}_1 (\sum_{i=1}^{n} X_i^2 - \bar{X} \sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} Y_i X_i - \bar{Y} \sum_{i=1}^{n} X_i \Rightarrow$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} Y_i X_i - \bar{Y} \sum_{i=1}^{n} X_i}{\sum_{i=1}^{n} X_i^2 - \bar{X} \sum_{i=1}^{n} X_i} = \frac{\sum_{i=1}^{n} Y_i X_i - \sum_{i=1}^{n} Y_i \frac{\sum_{i=1}^{n} X_i}{n}}{\sum_{i=1}^{n} X_i^2 - \bar{X} \frac{\sum_{i=1}^{n} X_i}{n} n} =$$

$$\frac{\sum_{i=1}^{n} Y_i X - \sum_{i=1}^{n} Y \frac{\sum_{i=1}^{n} X_i}{n}}{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2} = \frac{n \sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} Y_i \sum_{i=1}^{n} X_i}{n \sum_{i=1}^{n} X_i^2 - (\sum_{i=1}^{n} (X_i)^2)}$$

Τα $\hat{\beta}_0$ και $\hat{\beta}_1$ που εκτιμήσαμε για τις παραμέτρους β_0 και β_1 λέγονται ***σημειακές εκτιμήσεις.*** Για μεγαλύτερη σιγουριά στην εκτίμησή μας θα πρέπει να υπολογίσουμε τα αντίστοιχα Δ.Ε. και να διεξάγουμε στατιστικούς ελέγχους υποθέσεων για τις παραμέτρους τού υπό εκτίμηση υποδείγματος.

Για το λόγο αυτό πρέπει να προσδιοριστεί η κατανομή των $\hat{\beta}_0$ και $\hat{\beta}_1$ και να γίνουν οι πιο κάτω υποθέσεις:

i) Τα σφάλματα (errors) $\varepsilon_i \sim N(0, \sigma^2)$ για κάθε ***i*** . Η υπόθεση αυτή για τα ε_i (αντίστοιχα και για τα ***Yi***), δηλώνει ότι κάθε ένα από αυτά έχουν την ίδια διακύμανση. Για το λόγο αυτό τα σφάλματα ονομάζονται ***ομοσκεδαστικά*** *(homoscedastic)*. Παραβίαση της υπόθεσης αυτής οδηγεί στο πρόβλημα της ***ετεροσκεδαστικότητας*** *(heteroscedasticity)*.
ii) Τα σφάλματα (ε_i) είναι μεταξύ τους ανεξάρτητα. Δηλαδή, $Cov(\varepsilon_i, \varepsilon_j) = 0$, για κάθε i ≠ j (τα οποία δεν αυτοσυχ/νται).

Από τη σχέση $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ βλέπουμε ότι οι παραπάνω υποθέσεις συνεπάγονται τις παρακάτω υποθέσεις:
iii) Τα **Yi** $\sim N(\beta_0 + {}_1 X_i, \sigma^2)$, i = 1,2,...
iv) Τα ***Yi*** είναι ανεξάρτητα μεταξύ τους. Δηλαδή, τα ***Yi*** είναι, αντίστοιχα, ασυσχέτιστα).

ν*)** Τα ***Xi είναι προκαθορισμένα.

<u>Διάστημα Εμπιστοσύνης για το *β*</u>

***1*) <u>Πώς προσδιορίζεται ένα Δ.Ε για το *β*</u> ;**

(εκτιμητής) ± (κριτική τιμή πινάκων)(τυπ. σφάλ. εκτιμητή)*:

$$S_{\hat{\beta}_1} = \frac{S_e}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}}$$

$$\hat{\beta}_1 - S_{\hat{\beta}_1} * t_{n-2,1-\alpha/2} \leq \beta_1 \leq \hat{\beta}_1 + S_{\hat{\beta}_1} * t_{n-2,1-\alpha/2}$$

2) <u>Πώς προσδιορίζεται ένας στατιστικός έλεγχος;</u>

$\hat{\beta}_1$?

$$H_0 : \beta_1 = c$$
$$H_1 : \beta_1 \neq c$$

$$| t_{n-2} | = \frac{\hat{\beta}_1 - c}{S_{\hat{\beta}_1}}$$

όπου c είναι η τιμή της παραμέτρου $\boldsymbol{\beta_1}$ κάτω από την υπόθεση μηδέν**.**

Ύλη 6ης Διάλεξης

1. Παράδειγμα για:
   1. Εκτιμήσεις παραμέτρων
   2. Υπολογισμός των $\hat{Y}_i$
   3. Υπολογισμός εκτιμημένων σφαλμάτων
   4. Πρόβλεψη τιμών της Y_i
2. Εκτίμηση Διακύμανσης του Σφάλματος

2

Δεδομένα

Yi	*Xi*	Xi^2	Yi^2	XiYi n=16
1050	***32***	1024	1102500	33600
1260	***47***	2209	1587600	59220
1470	***23***	529	2160900	33810
2160	***68***	4624	4665600	146880
1950	***32***	1024	3802500	62400
2400	***17***	289	5760000	40800
2370	***58***	3364	5616900	137460
3150	***75***	5625	9922500	236250
3570	***98***	9604	12744900	349860
4410	***43***	1849	19448100	189630
4500	***76***	5776	20250000	342000
5610	***89***	7921	31472100	499290
5190	***108***	11664	26936100	560520
5670	***76***	5776	32148900	430920
5160	***65***	4225	26625600	335400
6840	***93***	8649	46785600	636120
56760	**1000**	**74152**	**251029800**	**4094160**

Ζητούμενο

1. Εκτιμήσεις των παραμέτρων β_1 και β_2:

1 $\hat{\beta}_1 = ?$

2 $\hat{\beta}_0 = ?$

2. Υπολογισμός του:

1 $\hat{Y}_i$ =? ($\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$)

2 $\hat{e}_i = Y_i - \hat{Y}_i = ?$ (εκτιμημένα σφάλματα)

Συνολική Μεταβλητότητα του Υποδείγματος

Yi	Xi		$(Y_i - \hat{Y}_i)$	$(Yi - \hat{Y}_i)^2$
1050	32	2116,5757	-1066,5757	1137583,724
1260	47	2820,30896	-1560,30896	2434564,051
1470	23	1694,33574	-224,33574	50326,52424
2160	68	3805,53553	-1645,53553	2707787,18
1950	32	2116,5757	-166,5757	27747,46383
2400	17	1412,84243	987,15757	974480,068
2370	58	3336,38002	-966,38002	933890,3431
3150	75	4133,94439	-983,94439	968146,5626
3570	98	5213,00206	-1643,00206	2699455,769
4410	43	2632,64676	1777,35324	3158984,54
4500	76	4180,85994	319,14006	101850,3779
5610	89	4790,7621	819,2379	671150,7368
5190	108	5682,15757	-492,15757	242219,0737
5670	76	4180,85994	1489,14006	2217538,118
5160	65	3664,78888	1495,21112	2235656,293
6840	93	4978,4243	1861,5757	3465464,087
				24026844,91

3

3. Υπολογισμός προβλέψεων: των τιμών ***της εξαρτημένης μεταβλητής***.
Με δεδομένη την τιμή της ***X,*** μπορούμε να προβλέψουμε την τιμή της ***Y.***

Απαντήσεις

1. Εκτιμήσεις:

$$\hat{\beta}_1 = \frac{n\sum Y_i X_i - \sum Y_i \sum X_i}{n\sum X_i^2 - (\sum (X_i))^2} = 46,916$$

$$\hat{\beta}_0 = \bar{Y} - \beta_1 \bar{X} = 615,278$$

Ερμηνεία των παραπάνω αποτελεσμάτων

2. $$\hat{e}_i = Y_i - \hat{Y}_i$$

Τα αποτελέσματα δίνονται στον προηγούμενο Πίνακα, στην 4η στήλη.

Παράδοση 21-3-2017

Σύντομη αναφορά σε διδαχθείσες έννοιες

Μέχρι στιγμής έχουμε μάθει ότι εφαρμόζοντας τη Μέθοδο των Ελαχίστων Τετραγώνων δηλαδή, την ελαχιστοποίηση του αθροίσματος των τετραγώνου των καταλοίπων), μπορούμε να υπολογίσουμε:

1. τους εκτιμητές: $\hat{\beta}_1 = \frac{\sum_i (\mathrm{X}_i - \bar{\mathrm{X}})(Y_i - \bar{Y})}{\sum_i (X_i - \bar{X})^2} = \frac{S_{XY}}{S_{XX}}$ & $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 X_i$
2. τα κατάλοιπα e_i ισούνται με $e_i = Y_i - \hat{Y}_i$. Το άθροισμα των τετραγώνων τους $\sum_i e_i^2 = \sum_i (Y_i - \hat{Y}_i)^2 = \sum_i (Y_i - \hat{\beta}_0 - \hat{\beta}_1)^2 = \sum_i Y_i^2 - \hat{\beta}_0 \sum_i Y_i - \hat{\beta}_1 \sum Y_i X_i$ και
3. τα αθροίσματα τετραγώνων ισούνται με:
   - $SST = \sum_i (Y_i - \bar{Y})^2$
   - $SSR = SST\text{-}SSE$ και
   - $SSE = \sum_i (\hat{Y}_i - \bar{\hat{Y}})^2$
4. για ευκολία γράφουμε:
   - $S_{XX} = \sum_i (X_i - \bar{X})^2 = \sum X_i^2 - \frac{(\sum_i X_i)^{2\partial}}{n}$
   - $S_{YY} = \sum_i (Y_i - \bar{Y})^2 = \sum_i Y_i^2 - \frac{(\sum_i Y_i)^2}{n}$
   - $S_{YX} = \sum_i (Y_i - \bar{Y})\ (X_i - \bar{X}) = \sum_i (X_i - \bar{X}) Y_i$

Η πληθυσμιακή διακύμανση του σφάλματος είναι η παράμετρος που καθορίζει την ένταση της εξάρτησης της Y από την X **(**σ_ε^2**).**

Ωστόσο, επειδή το σ_ε, το τυπικό σφάλμα εκτίμησης της εξίσωσης παλινδρόμησης, δεν είναι γνωστό, χρησιμοποιούμε την εκτίμηση S_e από τα δεδομένα. Η εκτίμηση αυτή θα βασισθεί στο άθροισμα των τετραγώνων των σφαλμάτων γύρω από τη γραμμή παλινδρόμησης, δηλαδή το SSE:

$$\sqrt{S_e^2} = \sqrt{\frac{\sum_i (Y_i - \hat{Y})^2}{n-2}} = \sqrt{\frac{SSE}{n-2}}$$

Προσδιορισμό τυπικού σφάλματος της κατανομής δειγματοληψίας του συντελεστή $\hat{\beta}_1$

Στο σημείο αυτό και βασισμένοι στα παραπάνω, θα ασχοληθούμε με τον προσδιορισμό του τυπικού σφάλματος της κατανομής δειγματοληψίας του συντελεστή $\hat{\beta}_1$, το οποίο συμβολίζεται με $\sigma_{\hat{\beta}_1}$ και δίνεται από τη σχέση:

$$\sigma_{\hat{\beta}_1} = \frac{\sigma_\varepsilon}{\sqrt{\sum_i (\mathrm{X}_i - \bar{\mathrm{X}})^2}},$$

βασισμένοι στο $\sqrt{Var(\hat{\beta}_1)} = \sqrt{\frac{\sum_i (\mathrm{X}_i - \bar{\mathrm{X}})(Y_i - \bar{Y})}{\sum_i (X_i - \bar{X})^2}}$ ή

$$S_{\hat{\beta}_1} = \frac{S_e}{\sqrt{\sum_i (X_i - \bar{X})}} = \frac{\sqrt{\frac{SSE}{n-2}}}{\sum_i X^2 - (\sum_i X)^2 / n}$$

Το $S_{\hat{\beta}_1}$ ονομάζεται και ***τυπικό σφάλμα εκτίμησης*** του συντελεστή παλινδρόμησης. Η κατανομή του $\hat{\beta}_1$ δίνεται παρακάτω

$$\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{S_{XX}})$$

Άρα το ***(1-α)% Δ. Ε.*** του $\hat{\beta}_1$ ακολουθεί

$$\hat{\beta}_1 \pm t_{n-2,1-\alpha/2} S_{\hat{\beta}_1} .$$

Αντίστοιχα η ελεγχοσυνάρτηση για τον στστιστικό έλεγχο του $\hat{\beta}_1$ είναι:

$$T_0 = \frac{\hat{\beta}_1 - \beta_1}{S\sqrt{\frac{1}{S_{XX}}}} \sim t_{n-2}$$

Οι υποθέσεις που συνεπάγονται τη χρήση του παραπάνω τύπου, διατυπώνονται παρακάτω:

α) $H_0 : \beta_1 = c$, $H_1 : \beta_1 \neq c$ Η H_0 απορρίπτεται αν $|T_0| > t_{n-2,1-\alpha/2}$

β) $H_0 : \beta_1 = c$, $H_1 : \beta_1 > c$ Η H_0 απορρίπτεται αν $T_0 > t_{n-2,1-\alpha}$

γ) $H_0 : \beta_1 = c$, $H_1 : \beta_1 < c$ Η H_0 απορρίπτεται αν $T_0 < -t_{n-2,1-\alpha}$

Προσδιορισμό της κατανομής δειγματοληψίας του συντελεστή $\hat{\beta}_0$

Η κατανομή του $\hat{\beta}_0$ δίνεται παρακάτω

$$\hat{\beta}_0 \sim N(\beta_0, \sigma^2\left(\frac{1}{n}+\frac{\bar{X}^2}{S_{XX}}\right))$$

διότι

$$E(\hat{\beta}_0) = E(Y - \hat{\beta}_1 X_i) = \beta_0$$

$$Var(\hat{\beta}_0) = S^2(\frac{1}{n}+\frac{\bar{X}^2}{S_{XX}})$$

$$S_{\hat{\beta}_0} = S\sqrt{(\frac{1}{n}+\frac{\bar{X}^2}{S_{XX}})}$$

Άρα το ***(1-α)% Δ. Ε.*** του $\hat{\beta}_0$ ακολουθεί

$$\hat{\beta}_0 \pm t_{n-2,1-\alpha/2} S_{\hat{\beta}_0}.$$

Αντίστοιχα η ελεγχοσυνάρτηση για τον στστιστικό έλεγχο του $\hat{\beta}_0$ είναι:

$$T_0 = \frac{\hat{\beta}_0 - \beta_0}{S\sqrt{\frac{1}{n}+\frac{\bar{X}^2}{S_{XX}}}} \sim t_{n-2}$$

Οι υποθέσεις που συνεπάγονται τη χρήση του παραπάνω τύπου, διατυπώνονται παρακάτω:

α) $H_0 : \beta_0 = \beta^*$, $H_1 : \beta_0 \neq \beta^*$ Η H_0 απορρίπτεται αν $|T_0| > t_{n-2,1-\alpha/2}$

β) $H_0 : \beta_0 = \beta^*$, $H_1 : \beta_0 > \beta^*$ Η H_0 απορρίπτεται αν $T_0 > t_{n-2,1-\alpha}$

γ) $H_0 : \beta_0 = \beta^*$, $H_1 : \beta_0 < \beta^*$ Η H_0 απορρίπτεται αν $T_0 < -t_{n-2,1-\alpha}$

ΟΠΑ
Τμήμα Στατιστικής
Ε. Κανδηλώρου
Αν. Καθηγήτρια

Αθήνα, 24-3-2017

Παράδειγμα
Ο πίνακας που ακολουθεί, δίνει στοιχεία για την ποσότητα σε νερό που χρησιμοποιήθηκε για πότισμα σε ένα χωράφι (σε εκατοστά) και την παραγωγή τριφυλλιού (σε τόνους ανά στρέμμα) στο χωράφι αυτό (που χρησιμοποιήθηκε πειραματικά).

Λύση Άσκησης *(λίπασμα-απόδοση αγροτεμαχίου)*

Νερό:	12	18	24	30	36	42	48
Σοδειά:	5.27	5.68	6.25	7.21	8.05	8.71	8.42

α) Υπάρχει μια γραμμική σχέση $\boldsymbol{Y = \beta_0 + \beta_1 X}$ ανάμεσα στη σοδειά και το νερό;
β) Ερμηνεύονται οι μεταβολές της $\boldsymbol{Y}$ από την $\boldsymbol{X;}$
γ) Ποιό είναι το ποσοστό ερμηνείας του Υ από το Χ;
δ) Ποιό είναι το 90% ΔΕ του συντελεστή κλίσης της εξίσωσης;

Άσκηση
14-3-2017

Ένας αγρότης ενδιαφέρεται να προσδιορίσει τον τρόπο με τον οποίο η ποσότητα του λιπάσματος (σε εκατοντάδες κιλά) που χρησιμοποιείται σε ένα αγροτεμάχιο επηρεάζει την παραγωγή (σε χιλ. κιλά) του αγροκτήματος. Για το σκοπό αυτό πειραματίστηκε με 10 όμοια αγροτεμάχια, έτσι ώστε οι όποιες διαφοροποιήσεις παρατηρούνται στην παραγωγή των αγρών να οφείλονται κατά κύριο λόγο στις διαφορετικές ποσότητες λιπάσματος που χρησιμοποιήθηκαν. τα δεδομένα δίνονται παρακάτω:

Λίπασμα	20	10	26	8	20	16	20	12	8	24
Παραγωγή	706	550	790	517	694	634	715	571	529	754

1. Να σχεδιάσετε το διάγραμμα διασποράς μεταξύ των 2 μεταβλητών. Τι αυτό αποκαλύπτει;
2. Να ελέγξετε τη στατιστική σημαντικότητα του συντελεστή συσχέτισης των 2 μεταβλητών, σε 1-α = 97%;
3. Ποιές είναι οι συνιστώσες της «συνολικής μεταβλητότητα της εξαρτημένης μεταβλητής»;
4. Να εκτιμήσετε το 96% ΔΕ των παραμέτρων β_1 & β_0.
5. Σύμφωνα με τη μέθοδο Ελαχίστων Τετραγώνων, η ευθεία που προσαρμόζεται καλύτερα στα δεδομένα μας είναι αυτή που *min* το *SSE* (άθροισμα των τετραγώνων των καταλοίπων = $\sum_{i=1}^{n} e_i^2$).

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_i))^2$$

Ερωτήσεις-Απαντήσεις

1) Ποιά είναι η ευθεία που προσαρμόζεται καλύτερα στα παραπάνω δεδομένα;

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad (1)$$

Να την εκτιμήσετε.

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i \qquad (2)$$

όπου:

$$\hat{\beta}_1 = \frac{n\sum_{i=1}^{n} X_i Y_i - \sum_{i=2}^{n} Y_i \sum_{i=1}^{n} X_i}{n\sum_{i=1}^{n} X_i^2 - ((\sum_{i=1}^{n} X_i))^2}$$

Υπολογίζοντας τα παραπάνω αθροίσματα:

$$\sum_{i=1}^{n} X_i Y_i = 111800$$

$$\sum_{i=1}^{n} X_i = 164$$

$$\sum_{i=1}^{n} Y_i = 6460$$

$$\sum_{i=1}^{n} X_i^2 = 3080$$

Η ευθεία που προσαρμόζεται καλύτερα στα παραπάνω δεδομένα και δεδομένου ότι:

$$\hat{\beta}_1 = \frac{n\sum_{i=1}^{n} X_i Y_i - \sum_{i=2}^{n} Y_i \sum_{i=1}^{n} X_i}{n\sum_{i=1}^{n} X_i^2 - ((\sum_{i=1}^{n} X_i))^2} = 15$$

&

$$\hat{\beta}_0 = \bar{Y} - 15\bar{X} = \frac{6460}{10} - 15(\frac{164}{10}) = 400$$

Άρα η εκτιμημένη εξίσωση είναι:

$$\hat{Y}_i = 400 + 15X_i$$

2) **Ποιά** είναι η κλίση της γραμμής παλινδρόμησης; $\hat{\beta}_1 = 15$

Τι μετρά; Ο εκτιμητής της παραμέτρου β_1 μετρά την κλίση της γραμμής παλινδρόμησης. Ο εκτιμητής αυτός μετρά το πόσο μεταβάλλεται η εξαρτημένη μεταβλητή (σε μονάδα μέτρησης της), σε μια μεταβολή της ανεξάρτητης κατά μία μονάδα δικής της μέτρησης. Άρα, η παραγωγή θα αυξηθεί κατά 15 χιλ. κιλά (15000 κιλά!), αν το λίπασμα αυξηθεί κατά μία εκατοντάδα κιλά (100 κιλά).

3) Ποιά είναι η προβλεπτική ικανότητα της εξίσωσης; Να ελέγξετε αυτήν την ικανότητα σε α=10%.

Η προβλεπτική ικανότητα της εξίσωσης παλινδρόμησης ή το ποσοστό των μεταβολών της εξαρτημένης μεταβλητής (***Y)*** που οφείλονται στις επιδράσεις της ***(X),*** εκτιμάται από την παρακάτω εξίσωση

$$R^2 = \frac{SSR}{SST} = \frac{87840}{88380} = 0,994$$

Ερμηνεία: το 99,4% των μεταβολών της παραγωγής (*Y*) οφείλεται στη μεταβολή της ποσότητας λιπάσματος (*X*) που χρησιμοποιείται. Το υπόλοιπο 0,6% των μεταβολών της παραγωγής οφείλεται σε άλλες, εκτός της *X*, μεταβλητές.

Να ελέγξετε την προβλεπτική ικανότητα της εξίσωσης, σε α=10%.

H_0: *θεωρητική τιμή του* $R^2=0$

H_1: *θεωρητική τιμή του* $R^2\neq 0$

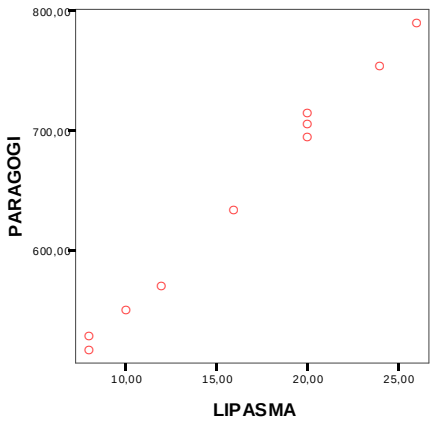
$$F_0 = \frac{\frac{SSR}{k-1}}{\frac{SSE}{n-2}} = \frac{MSR}{MSE} = \frac{87840}{67,5} = 1301,333$$

$$F_0 > F_{\nu_1,\nu_2;\alpha} = F_{1,8;0,01} = 11,2587$$

$\acute{\alpha}\rho\alpha \quad \not{H}_0$

Δεδομένου ότι η μηδενική υπόθεση απορρίπτεται και σε α=1%, έπεται ότι υπόθεση αυτή απορρίπτεται και σε α=10%

4) ***Να*** σχεδιάσετε το διάγραμμα διασποράς μεταξύ των δύο μεταβλητών.



Τι *αποκαλύπτει* αυτό το διάγραμμα;

Αποκαλύπτει με τον πιο εύκολο τρόπο, ότι υπάρχει θετική συσχέτιση μεταξύ των δύο μεταβλητών μας.

5) ***Τι*** *μετρά* ο συντελεστής συσχέτισης; Να τον υπολογίσετε.

Μετρά το βαθμό της γραμμικής συσχέτισης 2 τ.μ. (*X & Y)* με διασπορά $\boldsymbol{\sigma^2_X}$ & σ^2_Y αντίστοιχα & συνδιακύμανση , *Cov.(X,Y) = E(X, Y)-E(X)E(Y).*

ή

Η ποσοτική μέτρηση της γραμμικής σχέσης μεταξύ 2 μεταβλητών ονομάζεται συντλεστής συσχέτισης.

Να ελέγξετε τη στατιστική σημαντικότητα του συντελεστή συσχέτισης των 2 μεταβλητών, σε α =0,03.

$$r = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{\sum \left(X - \overline{X}\right)\left(Y - \overline{Y}\right)}{\sqrt{\sum \left(\left(X - \overline{X}\right)^2\right)\left(\sum \left(Y - \overline{Y}\right)^2\right)}}$$

$$= \frac{n\sum XY - \sum X \sum Y}{\sqrt{\left(n\left(\sum X^2\right) - \left(\sum X\right)^2\right)\left(n\left(\sum Y^2\right) - \left(\sum Y\right)^2\right)}}$$

=0,994

H_0: ρ=0
H_1: $\rho \neq 0$

$$t_{n-2} = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

= 25,704

sig.=0,000 $\Rightarrow H_0$ απορ. διότι sig.< α=0,03.

6) ***Ποιές*** είναι οι συνιστώσες της «συνολικής μεταβλητότητα της εξαρτημένης μεταβλητής»;

Είναι τα: ***SSR, SSE, SST***

7) ***Να*** εκτιμήσετε το 95% ΔΕ των παραμέτρων β_1. Τι σχόλιο έχετε να κάνετε για τους αντίστοιχους στατιστικούς έλεγχους, χωρίς να κάνετε πράξεις;

$$\widehat{\beta}_1 \pm t_{n-2,1-0,05/2} S_{\widehat{\beta}_1} \Rightarrow [14{,}041 \leq \beta_1 \leq 15{,}959)]$$

όπου $S_{\widehat{\beta}_1} = 0{,}416$ διότι $S_{\widehat{\beta}_1} = \frac{S_e}{\sqrt{\sum_i (X_i - \overline{X})}} = \frac{\sqrt{\frac{SSE}{n-2}}}{\sum_i X^2 - (\sum_i X)^2 / n}$

$t_{8,\ 0,975}$=*2,306*

H_0: $\beta_1 = 0$
H_1: $\beta_1 > 0$

H_0 απορρίπτεται, δεδομένου ότι το Δ.Ε. δεν περιλαμβάνει το μηδέν. Άρα, επιβεβαιώνεται η στατιστικά σημαντική θετική σχέση μεταξύ *X* & *Y*.

X	Y	XY	X*X	Y*Y	(Y-Ymean)
20	706	14120	400	498436	60
10	550	5500	100	302500	-96
26	790	20540	676	624100	144
8	517	4136	64	267289	-129
20	694	13880	400	481636	48
16	634	10144	256	401956	-12
20	715	14300	400	511225	69
12	571	6852	144	326041	-75
8	529	4232	64	279841	-117
24	754	18096	576	568516	108
164	**6460**	**111800**	**3080**	**4261540**	**0**

PRED	$(Y-Ymean)^2$	RES	PRED-646=W	W*W
700	3600	6	54	2916
550	9216	0	-96	9216
790	20736	0	144	20736
520	16641	-3	-126	15876
700	2304	-6	54	2916
640	144	-6	-6	36
700	4761	15	54	2916
580	5625	-9	-66	4356
520	13689	9	-126	15876
760	11664	-6	114	12996
6460	**88380**	**0**		**87840**
	SST			**SSR**

Γραμμικά Μοντέλα (31-3-2017)
Ελένη Κανδηλώρου

Ασκήσεις

1. Δίνονται δύο μεταβλητές με τα παρακάτω αθροίσματα.

$$\sum_{i=1}^{19} X_i = 124,21$$

$$\sum_{i=1}^{19} X_i^2 = 890,16$$

$$\sum_{i=1}^{19} Y_i = 417,21$$

$$\sum_{i=1}^{19} Y_i^2 = 11715,38$$

$$\sum_{i=1}^{19} X_i Y_i = 3163,32$$

α) Να γίνει η εκτίμηση των παραμέτρων β_0, β_1 της παλινδρόμησης $\boldsymbol{Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i}$
β) Να προσδιοριστεί το 95% διάστημα εμπιστοσύνης των $\boldsymbol{\beta_0, \beta_1}$
γ) Να ελεγχθεί η υπόθεση $\boldsymbol{\beta_1 = 0}$ σε επίπεδο σημαντικότητας 1%.

2. Δίνονται οι παρακάτω παρατηρήσεις:

Ηλικία (*X*)	36	38	42	42	47	49	55	56	60	63	68	72
Πίεση αίματος (*Y*)	118	115	125	140	128	145	150	147	155	149	152	160

α) Να σχεδιαστεί το διάγραμμα διασποράς μεταξύ των *X*, *Y*. Δικαιολογείται από το διάγραμμα η εφαρμογή γραμμικού υποδείγματος;
β) Να κατασκευάσετε:
(ι) το διάγραμμα διασποράς των δεδομένων (*X*,*Y*) μαζί με την εκτιμημένη ευθεία γραμμικής παλινδρόμησης και
(ιι) τα διαστήματα εμπιστοσύνης για την ατομική και μέση πρόβλεψη, με πιθανότητα 0,95.
γ) Να μελετήσετε το υπόδειγμα $\boldsymbol{Y = \beta_0 + \beta_1 X + \varepsilon}$. Συγκεκριμένα:
(ι) Να εκτιμήσετε τα $\boldsymbol{\beta_0, \beta_1}$.
(ιι) Ποιό είναι τα όρια του συντελεστή παλινδρόμισης, με πιθανότηατ 0.95,
(ιιι) Να ερμηνεύσετε τις εκτιμήσεις των συντελεστών παλινδρόμισης.
(ιν) Μπορείτε να ισχυριστείτε ότι η $\boldsymbol{X}$ δεν επηρεάζει τις μεταβολές της $\boldsymbol{Y}$; Η μεταβλητή *Y* εξαρτάται από την *X*;
(ν) Να κατασκευάσετε τον πίνακα ανάλυσης διασποράς (ANOVA) και να διαπιστώσετε αν έχει καλή προσαρμογή το υπόδειγμα.
(νι) Τι ποσοστό της μεταβλητότητας των $\boldsymbol{Yi}$ ερμηνεύεται από το υπόδειγμα;
δ) Να υπολογιστούν οι θεωρητικές τιμές των $\boldsymbol{Yi}$ και τα κατάλοιπα.
1) Ποιά είναι η πρόβλεψη της πίεσης του αίματος για γυναίκα ηλικίας $\boldsymbol{X_0 = 52}$ ετών**.**
2) Να γίνει σημειακή πρόβλεψη και να δοθούν τα διαστήματα ατομικής και μέσης πρόβλεψης (95%).
3) Εάν επιλεγεί τυχαία μια γυναίκα 52 ετών από τον πληθυσμό, μεταξύ ποιών ορίων θα βρίσκεται η πίεση του αίματός της (σ.ε. 95%).