Γεωργικός Πειραματισμός (Κωδ. 3515)
Βασικές Στατιστικές Μέθοδοι και Εργαλεία Ανάλυσης Δεδομένων
3. Γραμμική Συσχέτιση και Γραμμική Παλινδρόμηση

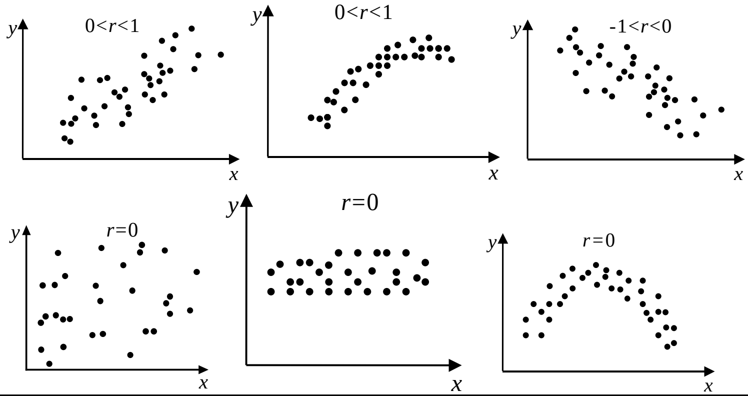
Σύντομη ανασκόπηση βασικών εννοιών, προτάσεων και τύπων

<table>
<tr>
<td>Συντελεστής γραμμικής συσχέτισης του Pearson (Pearson’s correlation coefficient)</td>
<td>
$(x_i, y_i),\ i = 1,2,3,\ldots,\nu$

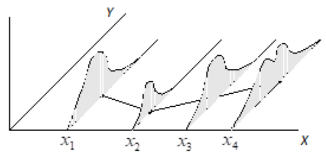
$$r = \frac{s_{xy}}{s_x s_y} = \frac{\sum_{i=1}^{\nu} x_i y_i - \nu\,\bar{x}\,\bar{y}}{\sqrt{\sum_{i=1}^{\nu} x_i^2 - \nu\,\bar{x}^2}\sqrt{\sum_{i=1}^{\nu} y_i^2 - \nu\,\bar{y}^2}}\ ,\ -1 \le r \le +1$$

$$s_{xy} = \frac{1}{\nu-1}\left(\sum_{i=1}^{\nu}(x_i-\bar{x})(y_i-\bar{y})\right) = \frac{1}{\nu-1}\left(\sum_{i=1}^{\nu} x_i y_i - \nu\,\bar{x}\,\bar{y}\right)$$

$$s_x^2 = \frac{1}{\nu-1}\sum_{i=1}^{\nu}(x_i-\bar{x})^2 = \frac{1}{\nu-1}\left(\sum_{i=1}^{\nu} x_i^2 - \nu\,\bar{x}^2\right)$$

$$s_y^2 = \frac{1}{\nu-1}\sum_{i=1}^{\nu}(y_i-\bar{y})^2 = \frac{1}{\nu-1}\left(\sum_{i=1}^{\nu} y_i^2 - \nu\,\bar{y}^2\right)$$
</td>
</tr>
<tr>
<td>Διάγραμμα διασποράς (scatterplot)</td>
<td>


</td>
</tr>
<tr>
<td>Η γενική υπόθεση για ένα μοντέλο παλινδρόμησης</td>
<td>
Η X μετριέται χωρίς σφάλμα ενώ η Y για κάθε επίπεδο x_i της X, είναι τυχαία μεταβλητή με πεπερασμένη μέση τιμή και διακύμανση.



</td>
</tr>
<tr>
<td>Το στατιστικό γραμμικό μοντέλο (stochastic or probabilistic model)</td>
<td>
Στο στατιστικό γραμμικό μοντέλο θεωρούμε ότι οι παρατηρήσεις $y_1, y_2, \ldots, y_\nu$ είναι τυχαίες παρατηρήσεις από πληθυσμούς με μέσες τιμές $\alpha + \beta\, x_i$ και σχετίζονται με τα x_i μέσω των σχέσεων

$$y_i = \alpha + \beta\, x_i + \varepsilon_i,\ i = 1, 2, \ldots, \nu$$

Y = ντετερμινιστικό μοντέλο + τυχαίο σφάλμα

Τα $\varepsilon_i = y_i - (\alpha + \beta\, x_i)$ είναι τυχαίες μεταβλητές με

$$E(\varepsilon_i) = 0\,,\ V(\varepsilon_i) = \sigma^2,\ Cov(\varepsilon_i, \varepsilon_j) = 0,\ i \ne j$$

ονομάζονται τυχαία σφάλματα (random errors) και αντίστοιχα εκφράζουν την η απόκλιση της παρατήρησης y_i από τη μέση τιμή $\alpha + \beta\, x_i$ του πληθυσμού από τον οποίο αυτή προέρχεται (του πληθυσμού όλων των δυνατών y_i όταν το X έχει τιμή x_i).

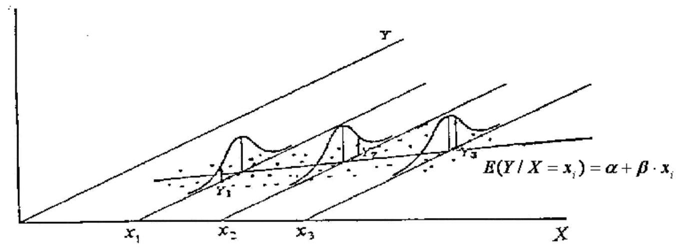
Προφανώς έχουμε

$$E(Y_i) = \alpha + \beta\, x_i \ \text{ και } \ V(Y_i) = \sigma^2 = V(\varepsilon_i)$$
</td>
</tr>
<tr>
<td>Η ευθεία παλινδρόμησης της Y πάνω στην X</td>
<td>

$$E(Y_i) = \alpha + \beta\, x_i$$
</td>
</tr>
</table>

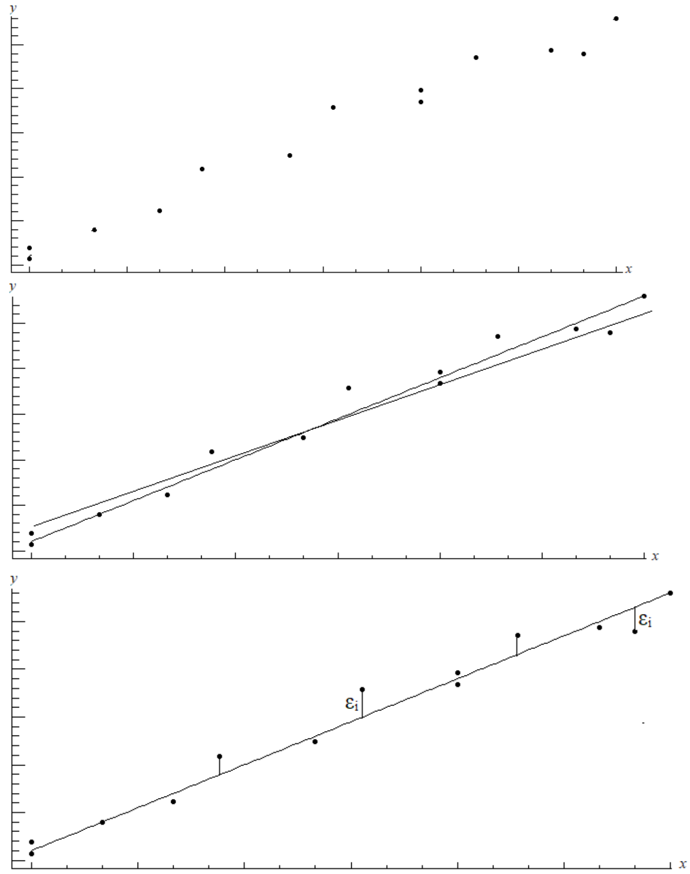
<table>
<tr>
<td>Το κανονικό γραμμικό μοντέλο</td>
<td>

$$Y_i \sim N(\alpha + \beta\, x_i, \sigma^2),\ i = 1, 2, \ldots, \nu$$



</td>
</tr>
<tr>
<td>Η ευθεία ελαχίστων τετραγώνων</td>
<td>

$$(x_i, y_i),\ i = 1,2,3,\ldots,\nu$$



$$g(\alpha, \beta) = \sum_{i=1}^{\nu} \varepsilon_i^2$$

$$\hat{y} = \hat{\alpha} + \hat{\beta}\, x$$

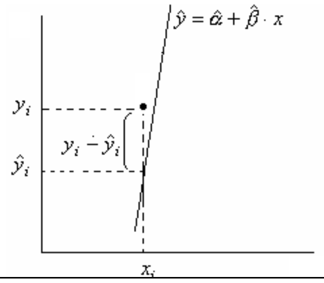
$$\hat{\beta} = \frac{s_{xy}}{s_x^2} = \frac{\sum_{i=1}^{\nu} x_i y_i - \nu\, \bar{x}\, \bar{y}}{\sum_{i=1}^{\nu} x_i^2 - \nu\, \bar{x}^2} \quad \text{και} \quad \hat{\alpha} = \bar{y} - \hat{\beta}\, \bar{x}$$

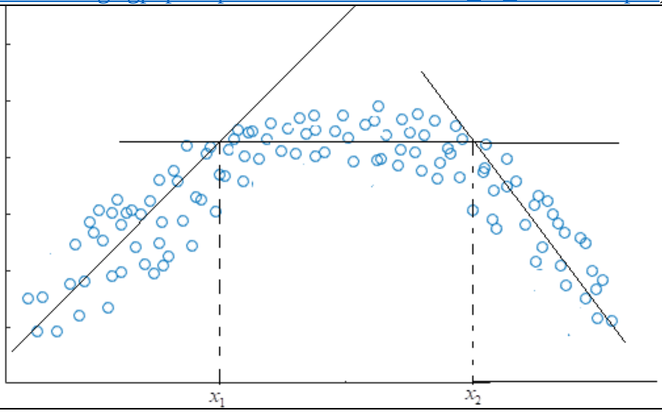
$$\hat{y} = \bar{y} + \hat{\beta}(x - \bar{x})$$

<u>Ερμηνεία των παραμέτρων:</u> Το $\hat{\beta}$ εκφράζει την αναμενόμενη (μέση) μεταβολή της εξαρτημένης μεταβλητής *Y* όταν η τιμή της ανεξάρτητης μεταβλητής *X* αυξηθεί κατά μια μονάδα. Έτσι, όταν η τιμή της *X* αυξηθεί κατά μια μονάδα το $\hat{y}$ αυξάνεται κατά $\hat{\beta}$ μονάδες αν $\hat{\beta} > 0$ ή ελαττώνεται κατά $\hat{\beta}$ μονάδες αν $\hat{\beta} < 0$.

Το $\hat{\alpha}$ εκφράζει την αναμενόμενη (μέση) τιμή της εξαρτημένης μεταβλητής *Y* όταν η τιμή της ανεξάρτητης μεταβλητής *X* πάρει την τιμή 0

</td>
</tr>
</table>

Υπόλοιπα (residuals)	$$\hat{\varepsilon}_i = y_i - \hat{y}_i = y_i - (\hat{\alpha} + \hat{\beta}\, x_i)$$ 
Συντελεστής προσδιορισμού (Coefficient of determination)	$$\sum_{i=1}^{\nu}(y_i - \bar{y})^2 = \sum_{i=1}^{\nu}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{\nu}(y_i - \hat{y}_i)^2$$ $$SSTO = SSR + SSE$$ $$r^2 = \frac{SSR}{SSTO}$$ <u>Ερμηνεία του συντελεστή προσδιορισμού r^2</u>**:** οι παρατηρήσεις y_i δεν είναι όλες μεταξύ τους ίσες, παρουσιάζουν μεταβλητότητα *(SSTO*). Ένα μέρος αυτής της μεταβλητότητας εξηγείται από τα x_i μέσω του γραμμικού μοντέλου *(SSR).* Το υπόλοιπο μέρος της συνολικής μεταβλητότητας δεν εξηγείται από το γραμμικό μοντέλο *(SSE)* αποδίδεται στην τύχη και εκφράζει την τυχαιότητα των παρατηρήσεων y_i, δηλαδή τις τυχαίες αποκλίσεις τους από τον μέσο τους (τα ε_i). Ο συντελεστής προσδιορισμού r^2 εκφράζει το ποσοστό της συνολικής μεταβλητότητας της εξαρτημένης μεταβλητής *Y* που εξηγείται από την ανεξάρτητη μεταβλητή *X* μέσω του γραμμικού μοντέλου. <u>Σημείωση</u>: Στο *SSE* ενσωματώνεται (α) η επίδραση παραγόντων που συνεισφέρουν στη μεταβλητότητα των y_i αλλά δεν συμπεριλαμβάνονται στο μοντέλο (δηλαδή, ενσωματώνεται η επίδραση μεταβλητών πέραν της *X*) και (β) η επίδραση από πιθανή αστοχία στην επιλογή μοντέλου (μπορεί η σχέση να περιγράφεται «καλύτερα» από ένα μη γραμμικό μοντέλο).
Μέσο τετραγωνικό υπόλοιπο (residual mean square ή mean square error)	$$s^2 = MSE = \frac{SSE}{\nu - 2} = \frac{\nu - 1}{\nu - 2}\left(s_y^2 - \frac{s_{xy}^2}{s_x^2}\right)$$ Αποτελεί μια αμερόληπτη εκτίμηση της διακύμανσης $V(Y_i) = \sigma^2 = V(\varepsilon_i)$
Τυπικό σφάλμα της εκτίμησης (standard error of the estimate)	$s = \sqrt{MSE}$
Διαστήματα Εμπιστοσύνης για τις παραμέτρους α, β	$(1-\alpha)100\%$ Διάστημα εμπιστοσύνης για την παράμετρο α $$\hat{\alpha} \pm SE(\hat{\alpha})t_{\nu-2;\alpha/2}$$ $(1-\alpha)100\%$ Διάστημα εμπιστοσύνης για την παράμετρο β $$\hat{\beta} \pm SE(\hat{\beta})t_{\nu-2;\alpha/2}$$ Όπου, $SE(\hat{\alpha}) = s\sqrt{\frac{1}{\nu} + \frac{\bar{x}^2}{(\nu-1)s_x^2}}$ και $SE(\hat{\beta}) = \frac{s}{s_x\sqrt{\nu-1}}$
Διάστημα Εμπιστοσύνης για την πρόβλεψη της μέσης τιμής E(Y₀) της Y για X = x₀	$(1-\alpha)100\%$ **Διάστημα Εμπιστοσύνης** για την πρόβλεψη της μέσης τιμής *$E(Y_0)$* για $X = x_0$ $$\hat{y}_0 \pm s(\hat{y}_0)t_{\nu-2;\alpha/2}$$ Όπου, $\hat{y}_0 = \hat{\alpha} + \hat{\beta}\, x_0$ και $s^2(\hat{y}_0) = s^2\left(\frac{1}{\nu} + \frac{(x_0 - \bar{x})^2}{(\nu-1)s_x^2}\right)$

Διάστημα Πρόβλεψης για την τιμή της Y όταν η X πάρει τιμή x₀.	$(1-\alpha)100\%$ **Διάστημα Πρόβλεψης** για την τιμή της *Y* όταν $X = x_0$ $$\hat{y}_0 \pm s\sqrt{\left(1+\frac{1}{\nu}+\frac{(x_0-\bar{x})^2}{(\nu-1)s_x^2}\right)}\, t_{\nu-2;\alpha/2}$$ Όπου, $\hat{y}_0 = \hat{\alpha}+\hat{\beta}x_0$		
Περιοχή απόρριψης της $H_0: \beta=\beta_0$	***έναντι της*** $H_1: \beta \neq \beta_0$	***έναντι της*** $H_1: \beta > \beta_0$	***έναντι της*** $H_1: \beta < \beta_0$
	$\dfrac{\lvert\hat{\beta}-\beta_0\rvert\sqrt{\nu-1}}{s/s_x} \geq t_{\nu-2,\alpha/2}$	$\dfrac{(\hat{\beta}-\beta_0)\sqrt{\nu-1}}{s/s_x} \geq t_{\nu-2,\alpha}$	$\dfrac{(\hat{\beta}-\beta_0)\sqrt{\nu-1}}{s/s_x} \leq -t_{\nu-2,\alpha}$
Περιοχή απόρριψης της $H_0: \rho = 0$	***έναντι της*** $H_1: \rho \neq 0$	***έναντι της*** $H_1: \rho > 0$	***έναντι της*** $H_1: \rho < 0$
	$\left\lvert r\sqrt{\dfrac{\nu-2}{1-r^2}}\right\rvert \geq t_{\nu-2,\alpha/2}$	$r\sqrt{\dfrac{\nu-2}{1-r^2}} \geq t_{\nu-2,\alpha}$	$r\sqrt{\dfrac{\nu-2}{1-r^2}} \leq -t_{\nu-2,\alpha}$
Σύγκριση των κλίσεων δύο ευθειών παλινδρόμησης $H_0: \beta_1 = \beta_2$ $H_1: \beta_1 \neq \beta_2$	Περιοχή απόρριψης της $H_0: \beta_1 = \beta_2$ $$\frac{\lvert\hat{\beta}_1-\hat{\beta}_2\rvert}{s_0\sqrt{\dfrac{1}{(\nu_1-1)s_{x_1}^2}+\dfrac{1}{(\nu_2-1)s_{x_2}^2}}} \geq t_{\nu_1+\nu_2-4;\alpha/2}$$ *Όπου,* $s_0^2 = \dfrac{(\nu_1-2)s_1^2+(\nu_2-2)s_2^2}{\nu_1+\nu_2-4}$ s_1^2, s_2^2, *τα μέσα τετραγωνικά υπόλοιπα των δύο μοντέλων και* $s_{x_1}^2$, $s_{x_2}^2$, *οι δειγματικές διακυμάνσεις των δύο ανεξάρτητων μεταβλητών*		
Συντελεστής γραμμικής συσχέτισης και παλινδρόμηση	Όταν έχουμε πειραματικά δεδομένα όπου ο ερευνητής ελέγχει-καθορίζει τις τιμές της μιας μεταβλητής, τότε η μεταβλητή αυτή είναι ανεξάρτητη (*X*) και η τυχαία μεταβλητή απόκρισης *(Y)* εξαρτημένη. Σε αυτή την περίπτωση εκτιμάμε την ευθεία παλινδρόμησης **της *Y* πάνω στη *X***. Όταν και οι δύο μεταβλητές *(X* και *Y)* είναι τυχαίες, τότε μπορούμε να θεωρήσουμε ως ανεξάρτητη οποιαδήποτε από τις δύο και να μελετήσουμε είτε την παλινδρόμηση **της *Y* πάνω στη *X*** είτε την παλινδρόμηση **της *X* πάνω στη *Y***. Έτσι, αν $\hat{y}=\hat{\alpha}+\hat{\beta}x$ και $\hat{x}=\hat{a}+\hat{b}y$ τότε $\hat{\beta}=\dfrac{s_{xy}}{s_x^2}$ και $\hat{b}=\dfrac{s_{xy}}{s_y^2}$ και επομένως: $\hat{\beta}\hat{b}=r^2$		
Αιτιότητα	Η συσχέτιση δεν συνεπάγεται κατ' ανάγκη αιτιότητα. (http://www.aua.gr/gpapadopoulos/files/association_or_causation.pdf)		
Προεκβολή (extrapolation)			

Προβλήματα και Ασκήσεις

1. Στον πίνακα που ακολουθεί φαίνεται η ποσότητα αζωτούχου λιπάσματος (x) που χρησιμοποιήθηκε σε καθένα από 8 πειραματικά αγροτεμάχια καλλιέργειας βρόμης καθώς και η απόδοση (y) κάθε αγροτεμαχίου.

x *(σε 100άδες pounds/στρέμμα)*	1	1	2	2	3	3	4	4
y *(σε bushels/στρέμμα)*	22	19	38	41	57	54	68	65

(α) Να προσαρμόσετε κατάλληλο γραμμικό μοντέλο μέσω του οποίου να μπορούμε να εκτιμήσουμε τη μέση απόδοση βρόμης για συγκεκριμένη ποσότητα αζωτούχου λιπάσματος.
(β) Να ερμηνεύσετε τον σταθερό όρο και την κλίση του γραμμικού μοντέλου που εκτιμήσατε στο (α).
(γ) Να δώσετε ένα 95% διάστημα εμπιστοσύνης για την κλίση του μοντέλου. Πώς αντιλαμβάνεστε (ερμηνεύετε) αυτό το διάστημα;
(δ) Ποιο ποσοστό της μεταβλητότητας της απόδοσης της καλλιέργειας εξηγείται από την ποσότητα αζωτούχου λιπάσματος μέσω του μοντέλου που εκτιμήσατε στο (α);
(ε) Πώς αξιολογείτε το μοντέλο που εκτιμήσατε στο (α); Είναι στατιστικά σημαντικό; Χρησιμοποιείστε επίπεδο σημαντικότητας 0,05.
(στ) Για ποσότητα αζωτούχου λιπάσματος ίση με 250 *pounds/στρέμμα* τι απόδοση βρόμης αναμένετε κατά μέσο όρο ανά στρέμμα;
(ζ) Για την εκτίμηση που ζητείται στο (στ) δώστε ένα 95% διάστημα εμπιστοσύνης.
(η) Για ποσότητα αζωτούχου λιπάσματος ίση με 500 *pounds/στρέμμα* τι απόδοση βρόμης αναμένετε κατά μέσο όρο ανά στρέμμα; Τι αξία μπορεί να έχει αυτή η εκτίμηση;

2. Στον πίνακα που ακολουθεί φαίνεται ο υπό ανόργανη μορφή φώσφορος (x) σε 16 δείγματα εδαφών και αντίστοιχα ο φώσφορος (y) που αφομοιώθηκε από φυτά καλαμποκιού που καλλιεργήθηκαν σε αυτά τα εδάφη *(οι τιμές είναι σε ppm).*

Έδαφος	x	y
1	0,40	64
2	0,40	60
3	3,10	71
4	0,60	61
5	4,70	54
6	1,70	77
7	9,40	81
8	10,10	93

Έδαφος	x	y
9	10,90	76
10	23,10	96
11	23,10	77
12	21,60	93
13	23,10	95
14	1,90	54
15	29,90	99
16	11,60	93

(α) Να προσαρμόσετε κατάλληλο γραμμικό μοντέλο μέσω το οποίου να μπορούμε να εκτιμήσουμε την ποσότητα φωσφόρου που κατά μέσο όρο απορροφούν τα φυτά καλαμποκιού από την ποσότητα φωσφόρου που βρίσκεται σε ανόργανη μορφή στο έδαφος.
(β) Να ερμηνεύσετε τον σταθερό όρο και την κλίση του γραμμικού μοντέλου που εκτιμήσατε στο (α).
(γ) Να δώσετε μια εκτίμηση με 95% εμπιστοσύνη της μέσης αύξησης της ποσότητας φωσφόρου που αφομοιώνεται από τα φυτά καλαμποκιού για κάθε επιπλέον *ppm* φωσφόρου στο έδαφος.

(δ) Να υπολογίσετε και να ερμηνεύσετε το συντελεστή προσδιορισμού του μοντέλου που εκτιμήσατε στο (α).
(ε) Πώς αξιολογείτε το μοντέλο που εκτιμήσατε στο (α); Είναι στατιστικά σημαντικό; Χρησιμοποιείστε επίπεδο σημαντικότητας 0,05.
(στ) Για ποσότητα φωσφόρου (ανόργανης μορφής) στο έδαφος ίση με 20 *ppm* τι ποσότητα φωσφόρου εκτιμάτε ότι θα απορροφήσουν κατά μέσο όρο τα φυτά καλαμποκιού; Δώστε ένα 95% διάστημα εμπιστοσύνης.
Να απαντήσετε χρησιμοποιώντας τα παρακάτω outputs από την ανάλυση των δεδομένων με το στατιστικό πακέτο Statgraphics.

Coefficients

	Least Squares	*Standard*	*T*	
Parameter	*Estimate*	*Error*	*Statistic*	*P-Value*
Intercept	63,9423	3,70219	17,2715	0,0000
Slope	1,2581	0,25234	4,98574	0,0002

Analysis of Variance

Source	*Sum of Squares*	*Df*	*Mean Square*	*F-Ratio*	*P-Value*
Model	2400,83	1	2400,83	24,86	0,0002
Residual	1352,17	14	96,5833		
Total (Corr.)	3753,0	15			

Predicted Values

		95,00%		*95,00%*	
	Predicted	*Prediction*	*Limits*	*Confidence*	*Limits*
X	*Y*	*Lower*	*Upper*	*Lower*	*Upper*
20,0	89,1044	66,835	111,374	81,9192	96,2895

3. Στον πίνακα που ακολουθεί φαίνεται η απόδοση μιας ποικιλίας σιταριού *(Yecora)* και μιας ποικιλίας κουκιών *(KY-188)* καθώς και οι τιμές (αντίστοιχα) του *Δείκτη Υδατικού Δυναμικού (Water Potential Index-WPI)*[1].

Σιτάρι (Yecora)	
WPI (MPa)	Απόδοση (Mg/ha)
-1,49	6,12
-1,33	6,27
-1,29	7,1
-1,74	4,73
-1,69	4,08
-1,79	3,47
-1,37	6,88
-1,33	5,67
-1,44	6,47
-2,01	4,47
-1,91	3,47
-2,1	2,17

Κουκιά (KY-188)	
WPI (MPa)	Απόδοση (Mg/ha)
-1,615	0,656
-1,59	0,638
-1,564	0,838
-1,521	1,069
-0,967	2,076
-0,984	2,433
-1,01	3,158
-1,135	1,729
-1,159	1,094
-1,208	2,403
-1,019	1,966
-0,994	1,931
-1,016	2,354

(α) Στα δεδομένα που αφορούν το σιτάρι ποικιλίας *Yecora* (και αντίστοιχα στα δεδομένα που αφορούν στην ποικιλία κουκιών *KY*-188) να προσαρμόσετε κατάλληλο γραμμικό μοντέλο μέσω του οποίου να μπορούμε να εκτιμήσουμε τη μέση απόδοση σιταριού *Yecora* (και αντίστοιχα τη μέση απόδοση κουκιών *KY*-188) για συγκεκριμένη τιμή του *Δείκτη Υδατικού Δυναμικού*.

[1] *Karamanos, A. J. & Papatheohari, A. Y. (1999). Assessment of drought resistance of crop genotypes by means of the water potential index. Crop science, 39(6), 1792-1797.*

(β) Για καθένα από τα μοντέλα που εκτιμήσατε στο (α) να ερμηνεύσετε τον σταθερό όρο και την κλίση του, να υπολογίσετε και να ερμηνεύσετε το συντελεστή προσδιορισμού του και να εξετάσετε αν είναι στατιστικά σημαντικό. Πώς συγκρίνονται οι κλίσεις των δύο γραμμικών μοντέλων; Χρησιμοποιείστε επίπεδο σημαντικότητας 0,05.

Για τα δεδομένα που αφορούν στο σιτάρι δίνονται:

$\sum x_i = -19{,}49$, $\sum y_i = 60{,}9$, $s_{xy} = 0{,}4096$, $s_x^2 = 0{,}082$, $s_y^2 = 2{,}4768$

Για τα δεδομένα που αφορούν στα κουκιά δίνονται:

$\sum x_i = -15{,}78$, $\sum y_i = 22{,}34$, $s_{xy} = 0{,}1731$, $s_x^2 = 0{,}067$, $s_y^2 = 0{,}631$

4. Ένας γεωπόνος-ερευνητής σε φυτώριο πειραματικού σταθμού, έχει επινοήσει μια κλίμακα μέτρησης του «βαθμού φρεσκάδας» που διατηρούν οι τριανταφυλλιές αφότου συσκευαστούν και αποθηκευτούν μέχρι να μεταφυτευθούν. Στον πίνακα που ακολουθεί φαίνεται ο χρόνος αποθήκευσης σε ημέρες (x) καθεμιάς από δέκα τριανταφυλλιές και αντίστοιχα ο «βαθμός φρεσκάδας» (y).

x	5	5	10	10	15	15	20	20	25	25
y	15,3	16,8	13,6	13,8	9,8	8,7	5,5	4,7	1.8	1,0

(α) Να προσαρμόσετε κατάλληλο γραμμικό μοντέλο μέσω του οποίου να μπορούμε να εκτιμήσουμε τον αναμενόμενο «βαθμό φρεσκάδας» των τριανταφυλλιών για συγκεκριμένο χρόνο αποθήκευσης.

(β) Να ερμηνεύσετε τον σταθερό όρο και την κλίση του γραμμικού μοντέλου που εκτιμήσατε στο (α).

(γ) Να υπολογίσετε και να ερμηνεύσετε το συντελεστή προσδιορισμού του μοντέλου που εκτιμήσατε στο (α).

(δ) Πώς αξιολογείτε το μοντέλο που εκτιμήσατε στο (α); Είναι στατιστικά σημαντικό; Χρησιμοποιείστε επίπεδο σημαντικότητας 0,05.

(ε) Τι «βαθμό φρεσκάδας» αναμένετε να έχουν οι τριανταφυλλιές μετά από 18 ημέρες αποθήκευσης;

(στ) Για την εκτίμηση που ζητείται στο (ε) δώστε ένα 95% διάστημα εμπιστοσύνης.

(ζ) Τι «βαθμό φρεσκάδας» αναμένετε να έχουν οι τριανταφυλλιές μετά από 30 ημέρες αποθήκευσης; Τι αξία μπορεί να έχει αυτή η εκτίμηση;

Δίνονται: $\sum x_i = 150$, $\sum y_i = 91$, $\sum x_i^2 = 2750$, $\sum y_i^2 = 1120$, $\sum x_i y_i = 986$.

5. Στον πίνακα που ακολουθεί φαίνονται τα δεδομένα που προέκυψαν από ένα πείραμα για τη διερεύνηση της σχέσης μεταξύ της φωτεινής διαπερατότητας (y) των φύλλων μιας ποικιλίας ρυζιού και της επιφάνειας του φύλλων όπως αυτή περιγράφεται μέσω ενός δείκτη (x).

x	y
0,50	75,0
0,60	72,0
1,80	42,0
2,50	29,0
2,80	27,0
5,45	10,0

x	y
5,60	9,0
7,20	5,0
8,75	2,0
9,60	2,0
10,40	1,0
12,00	0,9

(α) Να προσαρμόσετε κατάλληλο γραμμικό μοντέλο μέσω του οποίου να μπορούμε να εκτιμήσουμε τη φωτεινή διαπερατότητα των φύλλων ρυζιού για συγκεκριμένη τιμή του δείκτη x.

(β) Να ερμηνεύσετε τον σταθερό όρο και την κλίση του γραμμικού μοντέλου που εκτιμήσατε στο (α).

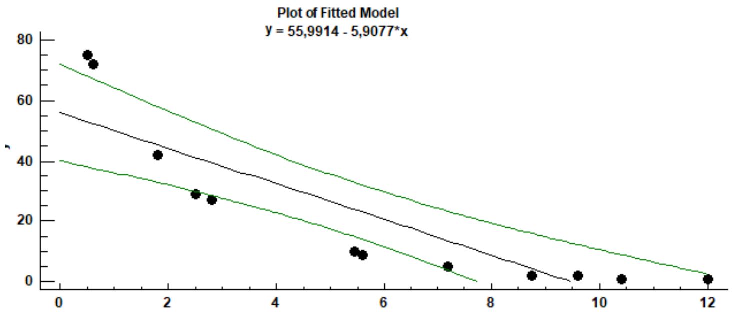
(γ) Να κατασκευάσετε ένα 95% διάστημα εμπιστοσύνης για την κλίση του μοντέλου. Πώς αντιλαμβάνεστε (ερμηνεύετε) αυτό το διάστημα;

(δ) Να υπολογίσετε και να ερμηνεύσετε το συντελεστή προσδιορισμού του μοντέλου που εκτιμήσατε στο (α).

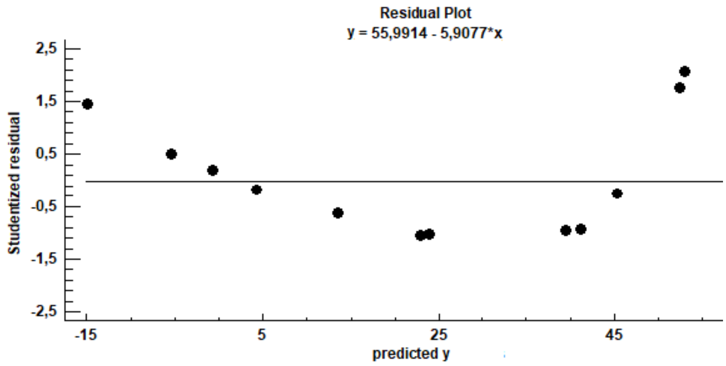
(ε) Πώς αξιολογείτε το μοντέλο που εκτιμήσατε στο (α); Είναι στατιστικά σημαντικό; Χρησιμοποιείστε επίπεδο σημαντικότητας 0,05.

(στ) Για τη διάγνωση πιθανών αποκλίσεων από τις υποθέσεις του κανονικού γραμμικού μοντέλου που προσαρμόσαμε στο (α), κατασκευάσαμε *(με το στατιστικό πακέτο Statgraphics)* το ***διάγραμμα διασποράς* (scatter plot)** των δεδομένων και την ευθεία ελαχίστων τετραγώνων, το ***διάγραμμα υπολοίπων* (residual plot)** και το ***κανονικό διάγραμμα πιθανότητας* (Normal Probability Plot)** των υπολοίπων *(σχήματα 1, 2 και 3 αντίστοιχα).* Υποδεικνύουν αυτά τα διαγράμματα παραβίαση κάποιας (ή κάποιων) από τις υποθέσεις-παραδοχές του κανονικού γραμμικού μοντέλου *(γραμμικότητα, κανονικότητα, ομοσκεδαστικότητα, ανεξαρτησία);*

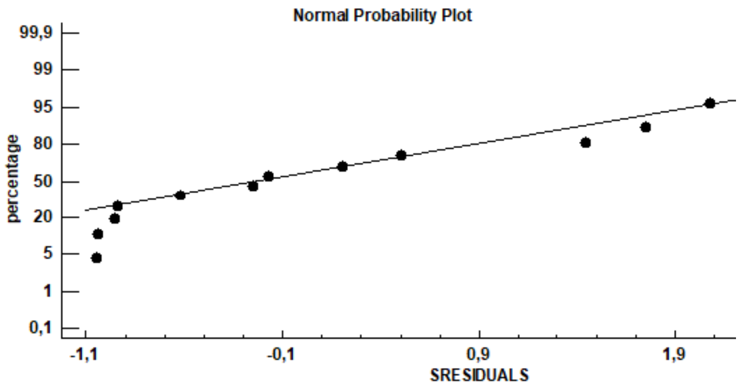
<u>Δίνονται:</u> $\sum x_i = 67$, $\sum y_i = 275$, $\sum x_i^2 = 552$, $\sum y_i^2 = 14359$, $\sum x_i y_i = 503$.



Σχήμα-1



Σχήμα-2



Σχήμα-3

6. Ένας γεωπόνος προκειμένου να μελετήσει τον τρόπο που η ποσότητα λιπάσματος επηρεάζει την απόδοση μιας καλλιέργειας, πειραματίσθηκε με διαφορετικές ποσότητες λιπάσματος σε 10 αγροτεμάχια ίδιου εμβαδού. Μερίμνησε επίσης, στα αγροτεμάχια να επικρατούν ίδιες ή παρόμοιες καλλιεργητικές συνθήκες έτσι ώστε οι όποιες διαφοροποιήσεις στην παραγωγή των αγροτεμαχίων να οφείλονται κατά κύριο λόγο στις διαφορετικές ποσότητες λιπάσματος. Στον πίνακα που ακολουθεί δίνεται η παραγωγή κάθε αγροτεμαχίου (y) και η αντίστοιχη ποσότητα λιπάσματος (x) που χρησιμοποιήθηκε.

x *(σε κιλά)*	20	10	26	8	20	16	20	12	8	24
y *(σε 100άδες κιλά)*	706	550	790	517	694	634	715	571	529	754

(α) Να προσαρμόσετε κατάλληλο γραμμικό μοντέλο μέσω του οποίου να μπορούμε να εκτιμήσουμε τη μέση απόδοση της καλλιέργειας για συγκεκριμένη ποσότητα λιπάσματος.
(β) Να ερμηνεύσετε τον σταθερό όρο και την κλίση του γραμμικού μοντέλου που εκτιμήσατε στο (α).
(γ) Να δώσετε μια εκτίμηση με 95% εμπιστοσύνη της μέσης αύξησης της παραγωγής αν χρησιμοποιηθεί ένα επιπλέον κιλό λιπάσματος.
(δ) Να υπολογίσετε και να ερμηνεύσετε το συντελεστή προσδιορισμού του μοντέλου που εκτιμήσατε στο (α).
(ε) Πώς αξιολογείτε το μοντέλο που εκτιμήσατε στο (α); Είναι στατιστικά σημαντικό; Χρησιμοποιείστε επίπεδο σημαντικότητας 0,05.
(στ) Για ποσότητα λιπάσματος ίση με 9 *κιλά/αγροτεμάχιο* τι απόδοση αναμένετε κατά μέσο όρο ανά αγροτεμάχιο;
(ζ) Για την εκτίμηση που ζητείται στο (στ) δώστε ένα 95% διάστημα εμπιστοσύνης.
(η) Για ποσότητα λιπάσματος ίση με 2 *κιλά/αγροτεμάχιο* τι απόδοση αναμένετε; Τι αξία μπορεί να έχει αυτή η εκτίμηση;
Δίνονται:
$\sum x_i = 164$, $\sum y_i = 6460$, $\sum x_i^2 = 3080$, $\sum y_i^2 = 4261540$, $\sum x_i y_i = 111800$.

7. Στο πλαίσιο μιας μελέτης για τη σχέση του ύψους και της ηλικίας των δένδρων μιας δασικής έκτασης, συγκεντρώθηκαν τα ακόλουθα δεδομένα.

Ηλικία *(σε m)*	3	1	2	5	4
Ύψος *(σε έτη)*	9	5	7	14	10

(α) Να προσαρμόσετε κατάλληλο γραμμικό μοντέλο μέσω του οποίου να μπορούμε (για τη συγκεκριμένη δασική έκταση) να εκτιμήσουμε το μέσο ύψος των δένδρων συγκεκριμένης ηλικίας.
(β) Να ερμηνεύσετε τον σταθερό όρο και την κλίση του γραμμικού μοντέλου που εκτιμήσατε στο (α).
(γ) Να δώσετε ένα 95% διάστημα εμπιστοσύνης για την κλίση του μοντέλου. Πώς αντιλαμβάνεστε (ερμηνεύετε) αυτό το διάστημα;
(δ) Ποιο ποσοστό της μεταβλητότητας του ύψους των δένδρων εξηγείται από την ηλικία τους μέσω του μοντέλου που εκτιμήσατε στο (α);.
(ε) Πώς αξιολογείτε το μοντέλο που εκτιμήσατε στο (α); Είναι στατιστικά σημαντικό; Χρησιμοποιείστε επίπεδο σημαντικότητας 0,05.
Τα παρακάτω outputs προέκυψαν από σχετική ανάλυση των δεδομένων με το στατιστικό πακέτο SPSS. Να απαντήσετε στα ερωτήματα χρησιμοποιώντας αυτά τα outputs (και αφού συμπληρώσετε τα κενά).

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	2,700	,835		3,235	,048
	x	2,100	,252	,979	8,345	**?**

a. Dependent Variable: y

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	**?**	1	44,100	69,632	,004[a]
	Residual	1,900	3	**?**		
	Total	46,000	4			

a. Predictors: (Constant), x

b. Dependent Variable: y

8. Στον πίνακα που ακολουθεί φαίνεται η απόδοση σε σπόρο (y) 22 φυτών μιας ποικιλίας σκληρού σιταριού. Επίσης, για κάθε φυτό φαίνονται τέσσερα συστατικά της απόδοσης: ο αριθμός σταχιών, ο αριθμός σταχιδίων ανά στάχυ και ο αριθμός σπόρων ανά σταχίδιο.

y (σε gr)	Στάχια/ φυτό (x)	Σταχίδια/ στάχυ (l)	Σπόροι/ σταχίδιο (w)	y (σε gr)	Στάχια/ φυτό (x)	Σταχίδια/ στάχυ (l)	Σπόροι/ σταχίδιο (w)
10,0	14,4	9,8	1,9	13,6	15,9	13,2	2,0
6,8	9,0	14,5	1,8	16,2	15,3	16,0	2,4
18,8	15,6	12,5	2,9	14,1	15,0	12,5	2,0
10,0	10,6	20,1	1,9	11,2	12,2	14,8	2,4
13,6	13,3	13,2	2,3	15,3	16,3	12,5	2,6
4,1	5,0	22,3	1,4	12,0	12,4	16,4	1,8
17,4	19,4	11,7	2,2	10,1	9,6	16,1	2,1
12,0	12,5	18,4	2,1	10,9	13,8	13,3	2,0
11,0	12,8	14,2	1,9	9,5	9,6	16,4	1,9
12,3	13,4	14,7	1,8	10,7	14,7	11,0	1,8
7,6	8,3	17,8	1,9	10,9	12,3	14,3	2,1

(α) Με βάση τα δεδομένα του πίνακα να γίνει ανάλυση συσχέτισης των χαρακτηριστικών απόδοσης της συγκεκριμένης ποικιλίας σκληρού σιταριού.
(β) Να προσαρμόσετε κατάλληλο γραμμικό μοντέλο μέσω του οποίου να μπορούμε να εκτιμήσουμε τη μέση απόδοση σε σπόρο ανά φυτό αν μας είναι γνωστός ο αριθμός σταχιών ανά φυτό.
(γ) Να ερμηνεύσετε τον σταθερό όρο και την κλίση του γραμμικού μοντέλου που εκτιμήσατε στο (β).
(δ) Να υπολογίσετε και να ερμηνεύσετε το συντελεστή προσδιορισμού του μοντέλου που εκτιμήσατε στο (β).
(ε) Πώς αξιολογείτε το μοντέλο που εκτιμήσατε στο (β); Είναι στατιστικά σημαντικό; Χρησιμοποιείστε επίπεδο σημαντικότητας 0,05.
Δίνονται: $\sum x_i = 281$, $\sum y_i = 258$, $\sum x_i^2 = 3080$, $\sum y_i^2 = 3272$, $\sum x_i y_i = 3502$.

9. Ο Sir Francis Galton σε εργασία του το 1894 διερεύνησε τη σχέση μεταξύ της διαμέτρου του σπόρου αρακά (d_1) και της διαμέτρου του παραγόμενου καρπού

(d_2). Για να συγκεντρώσει πολλά δεδομένα, ανέθεσε σε πολλούς φίλους του να φυτέψουν σπόρους διαμέτρων 0.15-0.21 ιντσών. Στον πίνακα που ακολουθεί φαίνεται ένα μέρος από τα δεδομένα που συγκέντρωσε *(οι μετρήσεις είναι σε ίντσες).*

d1	*d2*	*d1*	*d2*	*d1*	*d2*	*d1*	*d2*
0.21	0.1467	0.19	0.1407	0.17	0.1392	0.15	0.1377
0.21	0.1567	0.19	0.1507	0.17	0.1492	0.15	0.1477
0.21	0.1667	0.19	0.1607	0.17	0.1592	0.15	0.1577
0.21	0.1767	0.19	0.1707	0.17	0.1692	0.15	0.1677
0.21	0.1867	0.19	0.1807	0.17	0.1792	0.15	0.1777
0.21	0.1967	0.19	0.1907	0.17	0.1892	0.15	0.1877
0.21	0.2067	0.19	0.2007	0.17	0.1992	0.15	0.1977
0.21	0.2267	0.19	0.2207	0.16	0.1428	0.20	0.2066
0.20	0.1466	0.18	0.1435	0.16	0.1528	0.20	0.2266
0.20	0.1566	0.18	0.1535	0.16	0.1628	0.18	0.1935
0.20	0.1666	0.18	0.1635	0.16	0.1728	0.18	0.2035
0.20	0.1766	0.18	0.1735	0.16	0.1828	0.20	0.1966
0.20	0.1866	0.18	0.1835	0.16	0.1928	0.16	0.2028

Να διερευνήσετε τη σχέση μεταξύ της διαμέτρου του σπόρου-γονέα και της διαμέτρου του παραγόμενου καρπού.

<u>Δίνονται:</u> $\sum d_{1i} = 9{,}42$, $\sum d_{2i} = 9{,}12$, $\sum d_{1i}^2 = 1{,}73$, $\sum d_{2i}^2 = 1{,}63$, $\sum d_{1i} d_{2i} = 1{,}66$.

10. Στο πλαίσιο ενός πειράματος *(Linnik, 1961)* διερευνήθηκε αν και πώς η διαλυτότητα του νιτρικού νατρίου ($NaNO_3$) εξαρτάται από τη θερμοκρασία. Στον πίνακα που ακολουθεί φαίνονται για εννέα διαφορετικές θερμοκρασίες (σε βαθμούς κελσίου) τα μέρη νιτρικού νατρίου που διαλύθηκαν σε 100 μέρη νερού.

Θερμοκρασία *(x)*	Μέρη $NaNO_3$ *(y)*
0	67
4	71
10	76
15	81
21	86
29	93
36	99
51	114
68	125

(α) Να βρείτε (εκτιμήσετε) την ευθεία ελαχίστων τετραγώνων μέσω της οποίας να μπορούμε να εκτιμήσουμε τη μέση διαλυτότητα νιτρικού νατρίου για συγκεκριμένη θερμοκρασία.
(β) Να ερμηνεύσετε τον σταθερό όρο και την κλίση του γραμμικού μοντέλου που εκτιμήσατε στο (α).
(γ) Να δώσετε ένα 95% διάστημα εμπιστοσύνης για την κλίση του μοντέλου. Πώς αντιλαμβάνεστε (ερμηνεύετε) αυτό το διάστημα;
(δ) Ποιο ποσοστό της μεταβλητότητας της διαλυτότητας του νιτρικού νατρίου εξηγείται από τη θερμοκρασία μέσω του μοντέλου που εκτιμήσατε στο (α);
(ε) Πώς αξιολογείτε το μοντέλο που εκτιμήσατε στο (α); Είναι στατιστικά σημαντικό; Χρησιμοποιείστε επίπεδο σημαντικότητας 0,05.
(στ) Δώστε ένα 95% διάστημα εμπιστοσύνης για την αναμενόμενη διαλυτότητα νιτρικού νατρίου σε θερμοκρασία 30°C.

<u>Δίνονται:</u>

$\sum x_i = 234$, $\sum y_i = 812$, $\sum x_i^2 = 10144$, $\sum y_i^2 = 76334$, $\sum x_i y_i = 24640$.

11. Σε δείγματα μελιού έγιναν επεμβάσεις με *malathion* και *fluvalinate* σε συνθήκες *incubator* και *storage*. Για να μελετηθεί ο ρυθμός αποδόμησης των ουσιών αυτών, έγιναν μετρήσεις της συγκέντρωσης y κάθε ουσίας σε διάφορους χρόνους t μετά την αντίστοιχη επέμβαση. Τα αποτελέσματα των μετρήσεων αυτών φαίνονται στους παρακάτω πίνακες[2].

Malathion

Χρόνος μετά την αγωγή t_i *(σε εβδ.)*	Συγκέντρωση y_i *(σε ppb)*	
	Incubator	Storage
0	98.7	99.3
1	96.0	97.5
3	83.7	81.9
4	77.7	81.2
5	40.3	52.4
6	29.7	37.7
7	17.0	25.4
8	10.3	18.8
9	5.7	15.9
10	3.3	10.2

Fluvalinate

Χρόνος μετά την επέμβαση t_i *(σε εβδ.)*	Συγκέντρωση y_i *(σε ppb)*	
	Incubator	Storage
0	193.5	202.5
4	179.9	203.7
8	160.2	142.6
12	79.3	98.1
16	37.4	95.6
20	18.3	76.1
24	6.9	62.3

(α) Οι ερευνητές προσάρμοσαν στις πειραματικές μετρήσεις και για κάθε περίπτωση ξεχωριστά *(malathion* σε *incubator, malathion* σε *storage, fluvalinate* σε *incubator, fluvalinate* σε *storage),* το απλό γραμμικό μοντέλο παλινδρόμησης. Να εκτιμήσετε αυτά τα μοντέλα γραμμικής παλινδρόμησης και να ερμηνεύσετε τις τιμές των παραμέτρων τους.
(β) Να αξιολογήσετε τα μοντέλα που εκτιμήσατε στο (α).
(γ) Για κάθε περίπτωση, να εκτιμήσετε τη μέση συγκέντρωση της ουσίας δύο εβδομάδες μετά την αντίστοιχη επέμβαση.
(δ) Να ελέγξετε αν υπάρχει στατιστικώς σημαντική διαφορά μεταξύ των ρυθμών αποδόμησης i) της ουσίας *malathion* σε συνθήκες *incubator* και σε συνθήκες *storage* ii) της ουσίας *fluvalinate* σε συνθήκες *incubator* και σε συνθήκες σε *storage* iii) της ουσίας *malathion* σε συνθήκες *incubator* και της ουσίας *fluvalinate* σε συνθήκες *incubator*.

12. Στον πίνακα που ακολουθεί φαίνονται για έξι διαφορετικές θερμοκρασίες περιβάλλοντος οι ρυθμοί κατανάλωσης οξυγόνου που παρατηρήθηκαν σε πουλιά ορισμένου είδους.

Θερμοκρασία (^0C) *(x)*	-5	0	5	10	12	20
Κατανάλωση Οξυγόνου *(ml/g/hr)* *(y)*	3,6	3,3	2,6	2,3	2,2	1,9

(α) Να βρείτε (εκτιμήσετε) την ευθεία ελαχίστων τετραγώνων μέσω της οποίας να μπορούμε για το συγκεκριμένο είδος πουλιών να εκτιμήσουμε το μέσο ρυθμό κατανάλωσης οξυγόνου για συγκεκριμένη θερμοκρασία.
(β) Να ερμηνεύσετε τον σταθερό όρο και την κλίση του γραμμικού μοντέλου που εκτιμήσατε στο (α).
(γ) Πώς αξιολογείτε το μοντέλο που βρήκατε στο (α); Χρησιμοποιείστε επίπεδο σημαντικότητας 0,05.

[2] *Balayannis, P. G., & Santas, L. A. (1992). Dissipation of malathion and fluvalinate residues from honey. Journal of Apicultural Research, 31(2), 70-76.*

(δ) Δώστε ένα 95% διάστημα εμπιστοσύνης για το μέσο ρυθμό κατανάλωσης οξυγόνου των πουλιών σε θερμοκρασία περιβάλλοντος 8ºC.
Δίνονται: $\sum x_i = 42$, $\sum y_i = 15{,}9$, $\sum x_i^2 = 694$, $\sum y_i^2 = 44{,}35$, $\sum x_i y_i = 82{,}4$.

13. Σε ένα γνωστό πείραμα (Forbes 1857) μετρήθηκε η θερμοκρασία βρασμού του νερού σε διάφορες τιμές ατμοσφαιρικής πίεσης. Στον πίνακα που ακολουθεί φαίνεται ένα μέρος από τα δεδομένα που προέκυψαν.

Πίεση *(σε ίντσες)* *(x)*	Θερμοκρασία βρασμού *σε (ºF)* *(y)*
20,79	194,5
22,40	197,9
22,67	198.4
23,35	199,9
23,89	200,9
24,02	201,4
25,14	203,6
26,57	204,6
27,76	208,6
28,49	209,5

(α) Να βρείτε (εκτιμήσετε) την ευθεία ελαχίστων τετραγώνων μέσω της οποίας να μπορούμε να εκτιμήσουμε τη μέση θερμοκρασία βρασμού του νερού υπό συγκεκριμένη ατμοσφαιρική πίεση.
(β) Να ερμηνεύσετε τον σταθερό όρο και την κλίση του γραμμικού μοντέλου που εκτιμήσατε στο (α).
(γ) Πώς αξιολογείτε το μοντέλο που βρήκατε στο (α); Χρησιμοποιείστε επίπεδο σημαντικότητας 0,05.
(δ) Δώστε ένα 95% διάστημα εμπιστοσύνης για τη μέση θερμοκρασία βρασμού του νερού υπό ατμοσφαιρική πίεση 23 ίντσες.
Δίνονται:
$\sum x_i = 245$, $\sum y_i = 2019$, $\sum x_i^2 = 6061$, $\sum y_i^2 = 407958$, $\sum x_i y_i = 49593$.

14. Γεωπόνος ενδιαφέρεται να εκτιμήσει την απόδοση (*y*) καλλιέργειας βαμβακιού με βάση τον αριθμό των καρπών/καρυδιών (*x*) στο μέσον της καλλιεργητικής περιόδου. Στον πίνακα που ακολουθεί φαίνεται η απόδοση καθενός από επτά αγροτεμάχια και αντίστοιχα ο αριθμός καρπών κατά το μέσον της καλλιεργητικής περιόδου.

Απόδοση *(σε μπάλες)*	21	17	20	19	15	23	20
Αριθμός καρπών *(σε εκατοντάδες)*	5,5	2,8	4,7	4,3	3,7	6,1	4,5

(α) Να προσαρμόσετε κατάλληλο γραμμικό μοντέλο μέσω του οποίου να μπορεί ο γεωπόνος να εκτιμήσει τη μέση απόδοση βαμβακιού ανά αγροτεμάχιο με βάση τον αριθμό καρπών κατά το μέσον της καλλιεργητικής περιόδου.
(β) Να ερμηνεύσετε τον σταθερό όρο και την κλίση του γραμμικού μοντέλου που εκτιμήσατε στο (α).
(γ) Να δώσετε ένα 95% διάστημα εμπιστοσύνης για την κλίση του μοντέλου. Πώς αντιλαμβάνεστε (ερμηνεύετε) αυτό το διάστημα;
(δ) Να υπολογίσετε και να ερμηνεύσετε το συντελεστή προσδιορισμού του μοντέλου που εκτιμήσατε στο (α).
(ε) Πώς αξιολογείτε το μοντέλο που εκτιμήσατε στο (α); Είναι στατιστικά σημαντικό; Χρησιμοποιείστε επίπεδο σημαντικότητας 0,05.

(στ) Δώστε ένα 95% διάστημα εμπιστοσύνης για την αναμενόμενη απόδοση βαμβακιού (ανά αγροτεμάχιο) αν κατά το μέσον της καλλιεργητικής περιόδου καταμετρηθούν 430 καρποί.
(ζ) Για αριθμό καρπών στο μέσον της καλλιεργητικής περιόδου ίσο με 700 τι απόδοση αναμένετε; Τι αξία μπορεί να έχει αυτή η εκτίμηση;
Δίνονται:
$\sum x_i = 31{,}6$, $\sum y_i = 135$, $\sum x_i^2 = 149{,}82$, $\sum y_i^2 = 2645$, $\sum x_i y_i = 624{,}6$.

15. *(Συνέχεια της άσκησης 14)* Σε ένα κρίσιμο για την ανάπτυξη των φυτών στάδιο, ο γεωπόνος μέτρησε τον αριθμό των επιβλαβών εντόμων ανά αγροτεμάχιο. Στον πίνακα που ακολουθεί φαίνονται τα δεδομένα που προέκυψαν.

Απόδοση *(σε μπάλες)*	21	17	20	19	15	23	20
Αριθμός εντόμων	11	20	13	12	18	10	12

(α) Να προσαρμόσετε κατάλληλο γραμμικό μοντέλο μέσω του οποίου να μπορεί ο γεωπόνος να εκτιμήσει τη μέση απόδοση βαμβακιού ανά αγροτεμάχιο με βάση τον αριθμό των επιβλαβών εντόμων ανά αγροτεμάχιο.
(β) Να ερμηνεύσετε τον σταθερό όρο και την κλίση του γραμμικού μοντέλου που εκτιμήσατε στο (α).
(γ) Να δώσετε ένα 95% διάστημα εμπιστοσύνης για την κλίση του μοντέλου. Πώς αντιλαμβάνεστε (ερμηνεύετε) αυτό το διάστημα;
(δ) Να υπολογίσετε και να ερμηνεύσετε το συντελεστή προσδιορισμού του μοντέλου που εκτιμήσατε στο (α).
(ε) Πώς αξιολογείτε το μοντέλο που εκτιμήσατε στο (α); Είναι στατιστικά σημαντικό; Χρησιμοποιείστε επίπεδο σημαντικότητας 0,05.
Δίνονται: $\sum x_i = 96$, $\sum y_i = 135$, $\sum x_i^2 = 1402$, $\sum y_i^2 = 2645$, $\sum x_i y_i = 1799$.

16. Στον πίνακα που ακολουθεί φαίνεται η ηλικία και το βάρος δώδεκα μοσχαριών

Ηλικία (*σε μήνες)*	0	2	3	3	4	4	6	6	8	12	12	12
Βάρος *(σε Kg)*	20	35	70	55	50	70	100	95	140	180	195	190

(α) Να προσαρμόσετε κατάλληλο γραμμικό μοντέλο μέσω του οποίου να μπορούμε να εκτιμήσουμε το μέσο βάρος των μοσχαριών συγκεκριμένης ηλικίας.
(β) Να ερμηνεύσετε τον σταθερό όρο και την κλίση του γραμμικού μοντέλου που εκτιμήσατε στο (α).
(γ) Να δώσετε ένα 95% διάστημα εμπιστοσύνης για την κλίση του μοντέλου. Πώς αντιλαμβάνεστε (ερμηνεύετε) αυτό το διάστημα;
(δ) Να υπολογίσετε και να ερμηνεύσετε το συντελεστή προσδιορισμού του μοντέλου που εκτιμήσατε στο (α).
(ε) Πώς αξιολογείτε το μοντέλο που εκτιμήσατε στο (α); Είναι στατιστικά σημαντικό; Χρησιμοποιείστε επίπεδο σημαντικότητας 0,05.
(στ) Δώστε ένα 95% διάστημα εμπιστοσύνης για την εκτίμηση του μέσου βάρους των μοσχαριών ηλικίας 7 μηνών.
Δίνονται:
$\sum x_i = 72$, $\sum y_i = 1200$, $\sum x_i^2 = 622$, $\sum y_i^2 = 162100$, $\sum x_i y_i = 9995$.

17. Στον πίνακα που ακολουθεί φαίνονται τα δεδομένα που συγκεντρώθηκαν στο πλαίσιο μελέτης που έγινε για την ποσότητα σωματιδίων ρύπανσης που

απομακρύνονται από τον ατμοσφαιρικό αέρα σε σχέση με την ημερήσια βροχόπτωση.

Ημερήσια βροχόπτωση *(x)* *(σε 0.01cm)*	Σωματίδια *(y)* *(σε $\mu g/m^3$)*
4,3	126
4,5	125
5,9	116
5,6	118
6,1	114
5,2	118
3,8	132
2,1	141
7,5	108

(α) Να βρείτε (εκτιμήσετε) την ευθεία ελαχίστων τετραγώνων μέσω της οποίας να μπορούμε να εκτιμήσουμε τη μέση ποσότητα σωματιδίων που απομακρύνονται από τον ατμοσφαιρικό αέρα για συγκεκριμένη ημερήσια βροχόπτωση.
(β) Να ερμηνεύσετε τον σταθερό όρο και την κλίση του γραμμικού μοντέλου που εκτιμήσατε στο (α).
(γ) Να υπολογίσετε και να ερμηνεύσετε το συντελεστή προσδιορισμού του μοντέλου που εκτιμήσατε στο (α).
(δ) Πώς αξιολογείτε το μοντέλο που βρήκατε στο (α); Χρησιμοποιείστε επίπεδο σημαντικότητας 0,05.
(ε) Δώστε ένα 95% διάστημα εμπιστοσύνης για τη μέση ποσότητα σωματιδίων που απομακρύνονται από τον ατμοσφαιρικό αέρα όταν η ημερήσια βροχόπτωση είναι 4.8 (0.01 cm).
Δίνονται:
$\sum x_i = 45$, $\sum y_i = 1098$, $\sum x_i^2 = 244$, $\sum y_i^2 = 134770$, $\sum x_i y_i = 5366$.

18. Τα δεδομένα που ακολουθούν προέκυψαν στο πλαίσιο πειράματος για τη μελέτη και διερεύνηση της επίδρασης της θερμοκρασίας συντήρησης (*x*) στην υφή της επιφάνειας (*y*) της φράουλας *(οι τιμές της θερμοκρασίας έχουν κωδικοποιηθεί).*

x	-2	-2	0	2	2
y	4,0	3,5	2,0	0,5	0,0

(α) Να βρείτε (εκτιμήσετε) την ευθεία ελαχίστων τετραγώνων μέσω της οποίας να μπορούμε να εκτιμήσουμε την υφή της επιφάνειας των φραουλών από την θερμοκρασία συντήρησής τους.
(β) Να ερμηνεύσετε τον σταθερό όρο και την κλίση του γραμμικού μοντέλου που εκτιμήσατε στο (α).
(γ) Να υπολογίσετε και να ερμηνεύσετε το συντελεστή προσδιορισμού του μοντέλου που εκτιμήσατε στο (α).
(δ) Πώς αξιολογείτε το μοντέλο που εκτιμήσατε στο (α); Είναι στατιστικά σημαντικό; Χρησιμοποιείστε επίπεδο σημαντικότητας 0,05.
Δίνονται: $\sum x_i = 0$, $\sum y_i = 10$, $\sum x_i^2 = 16$, $\sum y_i^2 = 32{,}5$, $\sum x_i y_i = 14$.

19. Στον πίνακα που ακολουθεί φαίνεται η απόδοση σε σπόρο 20 φυτών ενός πληθυσμού λούπινου. Επίσης φαίνεται το ύψος κάθε φυτού αμέσως μετά τη συγκομιδή, ο αριθμός σπόρων ανά φυτό και ο αριθμός λοβών ανά φυτό[3].

[3] *Χαρακτηρισμός του Λούπινου των Άνδεων και καταγραφή των επικονιαστών του σε δύο περιοχές στην Ελλάδα. Μεταπτυχιακή Μελέτη, Μπάρδα Μυρτώ, 2018. (Επιβλ. Καθηγ. Μπεμπέλη Π.)*

y *Απόδοση* *(gr)*	*x* Ύψος φυτού *(cm)*	*l* Σπόροι/φυτό	*w* Λοβοί/φυτό
1,45	43,5	17	7
4,8	47	39	19
0,65	30	9	3
0,21	23	2	1
1,38	30	11	7
0,66	27	10	3
1,22	33,5	12	6
1,58	44	14	4
2,44	43	19	9
1,52	34,5	29	11
1,43	35	9	6
1,14	31	8	3
0,63	35,5	8	3
1,39	38,5	16	8
2,05	33	19	6
0,92	20	6	3
0,38	30	16	9
0	28	0	2
1,18	32	12	5
0,9	30	10	5

(α) Με βάση τα δεδομένα του πίνακα να γίνει ανάλυση συσχέτισης των τεσσάρων χαρακτηριστικών του συγκεκριμένου πληθυσμού λούπινου.
(β) Να προσαρμόσετε κατάλληλο γραμμικό μοντέλο μέσω του οποίου να μπορούμε να εκτιμήσουμε τη μέση απόδοση σε σπόρο ανά φυτό αν μας είναι γνωστό το ύψος του φυτού.
(γ) Να ερμηνεύσετε τον σταθερό όρο και την κλίση του γραμμικού μοντέλου που εκτιμήσατε στο (β).
(δ) Να υπολογίσετε και να ερμηνεύσετε το συντελεστή προσδιορισμού του μοντέλου που εκτιμήσατε στο (β).
(ε) Πώς αξιολογείτε το μοντέλο που εκτιμήσατε στο (β); Είναι στατιστικά σημαντικό; Χρησιμοποιείστε επίπεδο σημαντικότητας 0,05.
Δίνονται:
$\sum x_i = 668{,}5$, $\sum y_i = 25{,}9$, $\sum x_i^2 = 23282{,}3$, $\sum y_i^2 = 53{,}3$, $\sum x_i y_i = 963{,}5$.

20. Στον πίνακα που ακολουθεί δίνονται τα αποτελέσματα μετρήσεων πέντε μορφολογικών χαρακτηριστικών άγριου κριθαριού (*Hordeum spodaneum).*

Μήκος σπόρου	Πλάτος σπόρου	Μήκος λέπυρου	Μήκος τρίχας λέπυρου	Συμπάγια
9,11	2,8	26,31	19,25	21,29
8,98	2,84	22,63	15,61	21,97
9,28	2,9	27,52	21,41	23
9,65	2,84	25,15	17,81	21,3
8,23	2,97	20,37	15,45	23,76
9,45	2,98	23,15	16,26	25,2
8,85	2,96	23,19	16,11	24,16
9,13	2,99	24,67	17,2	21,45
8,86	2,86	21,75	15,09	25,11
9,18	2,98	23,3	16,07	25,17

(α) Με βάση τα δεδομένα του πίνακα να γίνει ανάλυση συσχέτισης των πέντε συγκεκριμένων χαρακτηριστικών του άγριου κριθαριού.

(β) Να προσαρμόσετε κατάλληλο γραμμικό μοντέλο μέσω του οποίου να μπορούμε να εκτιμήσουμε το μήκος του σπόρου από το μήκος του λέπυρου.

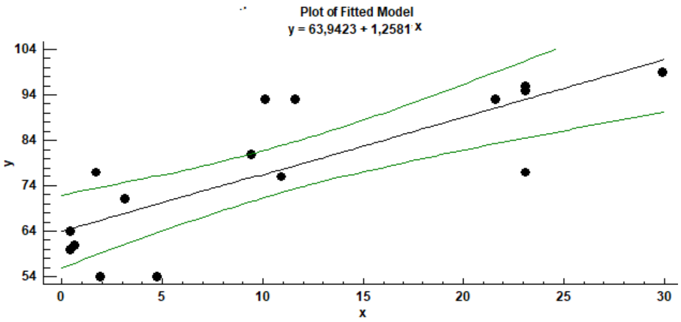
(γ) Να ερμηνεύσετε τον σταθερό όρο και την κλίση του γραμμικού μοντέλου που εκτιμήσατε στο (β).

(δ) Να υπολογίσετε και να ερμηνεύσετε το συντελεστή προσδιορισμού του μοντέλου που εκτιμήσατε στο (β).

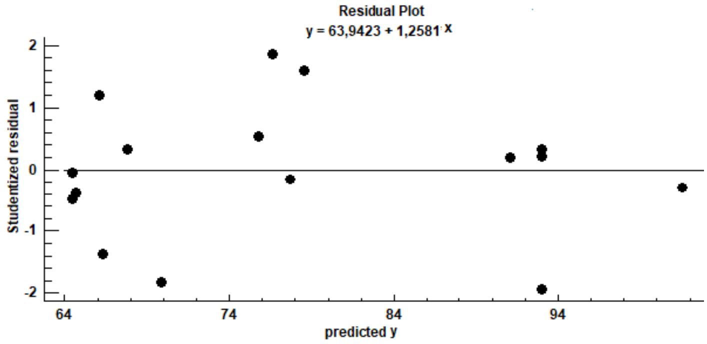
(ε) Πώς αξιολογείτε το μοντέλο που εκτιμήσατε στο (β); Είναι στατιστικά σημαντικό; Χρησιμοποιείστε επίπεδο σημαντικότητας 0,05.

Δίνονται: $\sum x_i = 249$, $\sum y_i = 98{,}5$, $\sum x_i^2 = 5833$, $\sum y_i^2 = 885$, $\sum x_i y_i = 5366$.

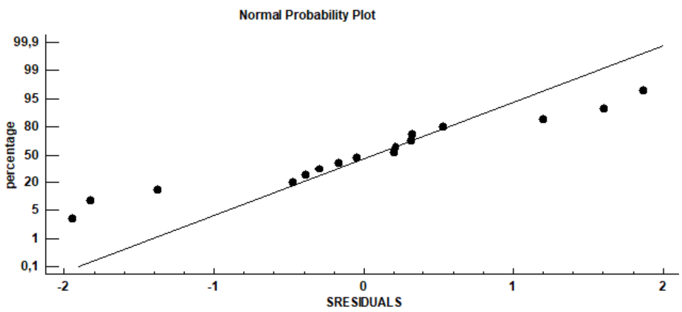
21. *(Συνέχεια της Άσκησης 2)* Για τη διάγνωση πιθανών αποκλίσεων από τις υποθέσεις του κανονικού γραμμικού μοντέλου που προσαρμόσαμε στο ερώτημα (α), κατασκευάσαμε *(με το στατιστικό πακέτο Statgraphics)* το ***διάγραμμα διασποράς*** των δεδομένων και την ευθεία ελαχίστων τετραγώνων, το ***διάγραμμα υπολοίπων*** και το ***κανονικό διάγραμμα πιθανότητας*** των υπολοίπων *(σχήματα 1, 2 και 3 αντίστοιχα)*. Υποδεικνύουν αυτά τα διαγράμματα παραβίαση κάποιας (ή κάποιων) από τις υποθέσεις-παραδοχές του κανονικού γραμμικού μοντέλου *(γραμμικότητα, κανονικότητα, ομοσκεδαστικότητα, ανεξαρτησία);*



Σχήμα-1



Σχήμα-2



Σχήμα-3

<u>Άσκηση 14 – (ενδεικτική απάντηση)</u>

Για τους απαιτούμενους υπολογισμούς διευκολύνει να συμπληρώσουμε τον πίνακα δεδομένων ως εξής:

x_i	y_i	x_i^2	y_i^2	$x_i y_i$
5,5	21	30,25	441	115,5
2,8	17	7,84	289	47,6
4,7	20	22,09	400	94
4,3	19	18,49	361	81,7
3,7	15	13,69	225	55,5
6,1	23	37,21	529	140,3
4,5	20	20,25	400	90
$\sum_{i=1}^{\nu} x_i = 31{,}6$	$\sum_{i=1}^{\nu} y_i = 135$	$\sum_{i=1}^{\nu} x_i^2 = 149{,}82$	$\sum_{i=1}^{\nu} y_i^2 = 2645$	$\sum_{i=1}^{\nu} x_i y_i = 624{,}6$

Υπολογίζουμε τα στατιστικά $\bar{x}$, $\bar{y}$, s_{xy}, s_x^2, s_y^2 και s^2 που θα μας χρειαστούν στη συνέχεια.

$$\bar{x} = \frac{\sum_{i=1}^{\nu} x_i}{\nu} = \frac{31{,}6}{7} = 4{,}51$$

$$\bar{y} = \frac{135}{7} = 19{,}29$$

$$s_{xy} = \frac{1}{\nu - 1}\left(\sum_{i=1}^{\nu} x_i y_i - \nu\,\bar{x}\,\bar{y}\right) = \frac{1}{6}\left(624{,}6 - 7 \cdot 4{,}51 \cdot 19{,}29\right) = 2{,}6$$

$$s_x^2 = \frac{1}{\nu - 1}\left(\sum_{i=1}^{\nu} x_i^2 - \nu\bar{x}^2\right) = \frac{1}{6}\left(149{,}8 - 7 \cdot 4{,}51^2\right) = 1{,}24$$

$$s_y^2 = \frac{1}{6}\left(2645 - 7 \cdot 19{,}29^2\right) = 6{,}71$$

$$s^2 = MSE = \frac{\nu - 1}{\nu - 2}\left(s_y^2 - \frac{s_{xy}^2}{s_x^2}\right) = \frac{6}{5}\left(6{,}71 - \frac{2{,}6^2}{1{,}24}\right) = 1{,}51$$

$$s = \sqrt{1{,}51} = 1{,}23$$

(α) Για την ευθεία ελαχίστων τετραγώνων

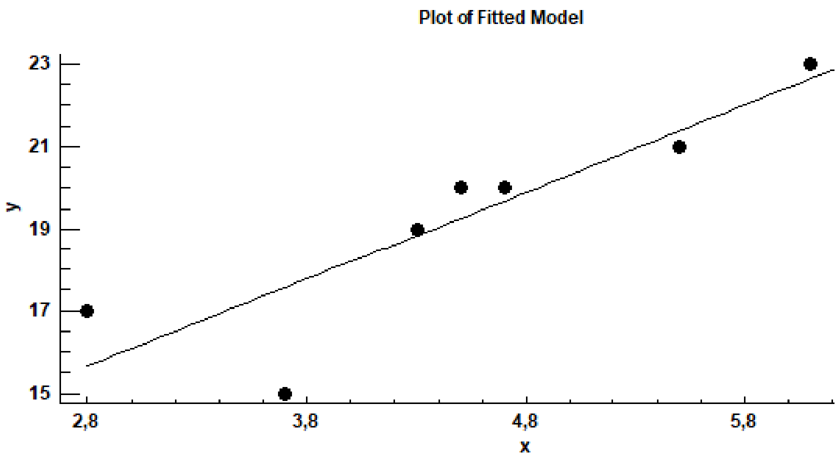
$$\hat{y} = \hat{\alpha} + \hat{\beta}\, x$$

έχουμε

$$\hat{\beta} = \frac{s_{xy}}{s_x^2} = \frac{2{,}6}{1{,}24} = 2{,}1 \text{ και } \quad \hat{\alpha} = \bar{y} - \hat{\beta}\,\bar{x} = 19{,}29 - 2{,}1 \cdot 4{,}51 = 9{,}8$$

Άρα, το ζητούμενο γραμμικό μοντέλο περιγράφεται από την εξίσωση

$$\hat{y} = 9{,}8 + 2{,}1\,x$$



(β) *Ερμηνεία του* $\hat{\beta}$*:* Επειδή $\hat{\beta} = 2{,}1 > 0$, αύξηση του αριθμού των καρπών συνεπάγεται αύξηση της απόδοσης βαμβακιού. Για κάθε επιπλέον 100άδα καρπών η μέση απόδοση εκτιμάται ότι θα αυξάνεται κατά 2,1 μπάλες ανά αγροτεμάχιο.

Ερμηνεία του $\hat{\alpha}$*:* Δίνει την εκτίμηση της μέσης απόδοσης για $x = 0$ (δηλαδή, για μηδέν καρπούς). Επειδή όμως η τιμή 0 είναι εκτός του εύρους των πειραματικών δεδομένων αλλά και επειδή με όρους του φυσικού προβλήματος δεν έχει νόημα (να αναμένουμε απόδοση από μηδέν καρπούς), η εκτίμηση αυτή δεν είναι μια αξιοποιήσιμη πρόβλεψη και ασφαλώς δεν έχει νόημα.

(γ) Το ζητούμενο $(1-\alpha)100\%$ διάστημα εμπιστοσύνης δίνεται από τον τύπο

$$\hat{\beta} \pm SE(\hat{\beta}) t_{\nu-2;\alpha/2}$$

Όπου, $\alpha = 0{,}05$ και $SE(\hat{\beta}) = \dfrac{s}{s_x\sqrt{\nu-1}} = \dfrac{1{,}23}{\sqrt{1{,}24}\sqrt{6}} = 0{,}45$

Άρα ένα 95% διάστημα εμπιστοσύνης για την κλίση του μοντέλου είναι

$2{,}1 \pm 0{,}45 t_{5;0.025}$ ή $2{,}1 \pm 0{,}45 \cdot 2{,}57$ ή $2{,}1 \pm 1{,}16$ ή $[0{,}94 \;\; 3{,}26]$

Αυτό σημαίνει ότι για κάθε επιπλέον 100άδα καρπών είμαστε 95% βέβαιοι ότι η μέση απόδοση θα αυξάνεται από 0,94 έως 3,26 μπάλες ανά αγροτεμάχιο.

(δ) Ο συντελεστής προσδιορισμού είναι

$$r^2 = \frac{s_{xy}^2}{s_x^2 s_y^2} = \frac{2{,}6^2}{1{,}24 \cdot 6{,}71} = 0{,}81.$$

Δηλαδή, ο αριθμός των καρπών (κατά το μέσον της καλλιεργητικής περιόδου) εξηγεί μέσω του γραμμικού μοντέλου το 81% της συνολικής μεταβλητότητας της απόδοσης του βαμβακιού.

(ε) Θα κάνουμε τον έλεγχο της υπόθεσης

$H_0 : \beta = 0$ έναντι της $H_1 : \beta \neq 0$

όπου β η κλίση της ευθείας παλινδρόμησης.

Η απορριπτική περιοχή του ελέγχου ορίζεται από την ανισότητα

$$\frac{\left|\hat{\beta} - \beta_0\right| \sqrt{\nu-1}}{s/s_x} \geq t_{\nu-2,\alpha/2}$$

και επειδή για επίπεδο σημαντικότητας $\alpha = 5\%$ η τιμή $\frac{|2{,}1-0|\sqrt{7-1}}{1{,}23/1{,}11} = 4{,}64$ που παίρνει η στατιστική συνάρτηση ελέγχου ανήκει στην απορριπτική περιοχή αφού $\frac{|2{,}1-0|\sqrt{7-1}}{1{,}23/1{,}11} = 4{,}64 \geq t_{5;0.025} = 2{,}57$, η μηδενική υπόθεση απορρίπτεται, δηλαδή σε επίπεδο σημαντικότητας 5%, το μοντέλο που προσαρμόσαμε είναι στατιστικά σημαντικό. Αυτό σημαίνει ότι ο αριθμός καρπών (κατά το μέσον της καλλιεργητικής περιόδου) μέσω του γραμμικού μοντέλου συνεισφέρει στατιστικά σημαντικά στην πρόβλεψη της μέσης απόδοσης βαμβακιού.

(στ) Για $x_0 = 4{,}2$ το μοντέλο που προσαρμόσαμε στα πειραματικά δεδομένα προβλέπει $\hat{y}_0 = 9{,}8 + 2{,}1 \cdot 4{,}2 = 18{,}62$ μπάλες ανά αγροτεμάχιο.

Αυτό σημαίνει ότι αν στο μέσον της καλλιεργητικής περιόδου καταμετρηθούν 420 καρποί, η αναμενόμενη απόδοση βαμβακιού εκτιμάται (προβλέπεται) ίση με 18,62 μπάλες ανά αγροτεμάχιο. Ένα 95% διάστημα εμπιστοσύνης για αυτή την εκτίμηση δίνεται από τον τύπο

$$\hat{y}_0 \pm s(\hat{y}_0) t_{\nu-2;\alpha/2}$$

όπου, $\hat{y}_0 = 18{,}62$ και

$$s^2(\hat{y}_0) = s^2\left(\frac{1}{\nu} + \frac{(x_0-\bar{x})^2}{(\nu-1)s_x^2}\right) = 1.51\left(\frac{1}{7} + \frac{(4{,}2-4{,}51)^2}{6 \cdot 1{,}24}\right) = 0{,}23$$

δηλαδή, $18{,}62 \pm \sqrt{0{,}23}\,t_{5;0{,}025}$ ή $18{,}62 \pm 1{,}23$.

Αυτό σημαίνει ότι για αριθμό καρπών 420 (στο μέσον της καλλιεργητικής περιόδου) το διάστημα [17,39 19,85] περιέχει τη μέση απόδοση βαμβακιού με πιθανότητα 95%.

(ζ) Για $x_0 = 7$ το μοντέλο προβλέπει $\hat{y}_0 = 9{,}8 + 2{,}1 \cdot 7 = 24{,}5$ μπάλες ανά αγροτεμάχιο όμως αυτή η εκτίμηση δεν είναι μια καλή και αξιοποιήσιμη πρόβλεψη γιατί η τιμή $x_0 = 7$ βρίσκεται εκτός του εύρους των πειραματικών δεδομένων.

Στο σχήμα που ακολουθεί φαίνεται το διάγραμμα διασποράς των δεδομένων, το γραμμικό μοντέλο που προσαρμόσαμε και η 95% ζώνη εμπιστοσύνης.