

ΑΣΚΗΣΕΙΣ

1. Ο διευθυντής προσωπικού μιας μεγάλης εταιρείας πιστεύει ότι ίσως υφίσταται κάποια σχέση μεταξύ των ημερών απουσίας και της ηλικίας των εργαζομένων. Με βάση την υπόθεση αυτή ενδιαφέρεται να κατασκευάσει ένα μοντέλο το οποίο θα χρησιμοποιεί για να προβλέπει τον αριθμό των ημερών απουσίας των εργαζομένων στην εταιρεία κατά τη διάρκεια του έτους. Ένα τυχαίο δείγμα 10 εργαζομένων έδωσε τα εξής στοιχεία που αναφέρονται σε ηλικία και σε αριθμό ημερών απουσίας:

<u>Πρόβλημα απουσιών</u>		
Εργαζόμενος	Ηλικία (σε έτη)	Ημέρες απουσίας Y
1	27	15
2	61	6
3	37	10
4	23	18
5	46	9
6	58	7
7	29	14
8	36	11
9	64	5
10	40	8

- α) Να κατασκευασθεί το διάγραμμα σημείων του προβλήματος αυτού.
- β) Υποθέτοντας ότι υφίσταται μια γραμμική σχέση μεταξύ των δύο αυτών ποσοτήτων, να χρησιμοποιηθεί η μέθοδος των ελαχίστων τετραγώνων για να καθορισθούν οι συντελεστές παλινδρόμησης a και b .
- γ) Να ερμηνευθεί η έννοια της κλίσης b στο πρόβλημα αυτό.
- δ) Για ένα εργαζόμενο ηλικίας 40 ετών, ποιος θα ήταν ο μέσος αριθμός ημερών απουσίας ανά έτος που προβλέπεται με τη χρήση του μοντέλου;
- ε) Να υπολογισθεί ο συντελεστής προσδιορισμού r^2 , όπως επίσης και ο προσαρμοσμένος συντελεστής προσδιορισμού r_{adj}^2 και να ερμηνευθούν.
- στ) Να καθορισθεί αν από τα δεδομένα του προβλήματος προκύπτουν ενδείξεις γραμμικής σχέσης μεταξύ της ηλικίας και των ημερών απουσίας στο $\alpha=0.01$ επίπεδο σημαντικότητας.
- ζ) Με τη χρήση της ανάλυσης των καταλοίπων να εξετασθεί η καταλληλότητα του μοντέλου.

2. Μια εταιρεία που κατασκευάζει ανταλλακτικά ενδιαφέρεται να διαμορφώσει ένα μοντέλο ώστε να εκτιμά τον αριθμό των εργατοωρών που απαιτούνται για την κατασκευή παρτίδων ανταλλακτικών (συσκευασιών) διαφορετικού μεγέθους. (Ως μέγεθος συσκευασίας θεωρείται ο αριθμός των ανταλλακτικών που περιέχει η συγκεκριμένη συσκευασία). Ένα τυχαίο δείγμα από 14 ακολουθίες κατασκευών ανταλλακτικών έδωσε τα αποτελέσματα που φαίνονται στον πίνακα που ακολουθεί: (Για κάθε μια από τις συσκευασίες μεγέθους 20, 30, 40, 50, 60, 70 και 80 ανταλλακτικών επελέγησαν δύο τυχαία δείγματα).

α) Να κατασκευασθεί το διάγραμμα σημείων.

β) Με την υπόθεση ότι υφίσταται μια γραμμική σχέση μεταξύ των δύο αυτών μεγεθών, να προσδιορισθούν οι συντελεστές παλινδρόμησης a και b με τη μέθοδο των ελαχίστων τετραγώνων.

γ) Να δοθεί η ερμηνεία των συντελεστών αυτών.

<u>Πρόβλημα εργατοωρών</u>	
Μέγεθος συσκευασίας	Εργατοώρες
20	50
20	55
30	73
30	67
40	87
40	95
50	108
50	112
60	128
60	135
70	148
70	160
80	170
80	162

δ) Να προβλεφθεί ο μέσος αριθμός εργατοωρών που απαιτείται για να παραχθεί μια συσκευασία με 45 αντικείμενα.

ε) Να δικαιολογηθεί γιατί είναι ακατάλληλο να προβλεφθεί ο μέσος αριθμός εργατοωρών που απαιτούνται για να συμπληρωθεί μια συσκευασία μεγέθους 100 ανταλλακτικών.

στ) Να κατασκευασθεί ένα 95% διάστημα εμπιστοσύνης για τον μέσο αριθμό εργατωρών που απαιτούνται για να συμπληρωθεί μια συσκευασία μεγέθους 45 ανταλλακτικών.

ζ) Να κατασκευασθεί ένα 95% διάστημα πρόβλεψης για τις ώρες που χρειάζεται ένας συγκεκριμένος εργαζόμενος να ολοκληρώσει μια συσκευασία 45 ανταλλακτικών.

η) Να υπολογισθεί ο συντελεστής προσδιορισμού r^2 και να ερμηνευθεί.

θ) Να καθορισθεί αν, με βάση τα συγκεκριμένα δεδομένα, υπάρχουν ενδείξεις γραμμικής σχέσης μεταξύ του μεγέθους της συσκευασίας και των εργατωρών που απαιτούνται για να γίνει η συσκευασία αυτή σε επίπεδο σημαντικότητας $\alpha=0.10$.

3. Ένας παραγωγός ενδιαφέρεται να καθορίσει την επίδραση ενός φυσικού οργανικού λιπάσματος στην παραγωγή τομάτας. Για το σκοπό αυτό χρησιμοποιεί πέντε διαφορετικές ποσότητες του λιπάσματος αυτού 0, 10, 20, 30 και 40 κιλών σε ισοδύναμα κομμάτια γης έκτασης 100 τ.μ. το καθένα. Οι ποσότητες του λιπάσματος τοποθετούνται τυχαία στα κομμάτια γης. Τα αποτελέσματα δίνονται στον πίνακα που ακολουθεί.

<u>Πρόβλημα παραγωγής τομάτας</u>		
Πλαίσιο γης	Ποσότητα λιπάσματος X	Παραγωγή Y
1	0	6
2	0	8
3	10	11
4	10	14
5	20	18
6	20	23
7	30	25
8	30	28
9	40	30
10	40	34

α) Να κατασκευαστεί το διάγραμμα σημείων.

β) Να προσδιορισθεί η ευθεία παλινδρόμησης με τη μέθοδο των ελαχίστων τετραγώνων για τα δεδομένα αυτά.

γ) Να κατασκευασθούν 95% διαστήματα εμπιστοσύνης για τις παραμέτρους α και β .

- δ) Να ερμηνευθεί η τιμή b της εκτιμήτριας της παραμέτρου β που προέκυψε με τη μέθοδο των ελαχίστων τετραγώνων.
- ε) Να προβλεφθεί η μέση παραγωγή τομάτος για ένα πλαίσιο αγρού 100τ.μ. στο οποίο χρησιμοποιήθηκαν 15 κιλά του οργανικού αυτού λιπάσματος.
- στ) Να κατασκευαστεί το 95% διάστημα εμπιστοσύνης για την μέση παραγωγή τομάτας που προβλέπεται να προκύψει από τη χρησιμοποίηση 15 κιλών λιπάσματος σε ένα πλαίσιο γης 15τ.μ.
- ζ) Να κατασκευαστεί το 95% διάστημα πρόβλεψης για την παραγωγή τομάτας που θα προκύψει από ένα συγκεκριμένο πλαίσιο αγρού έκτασης 100τ.μ. στο οποίο χρησιμοποιήθηκαν 20 κιλά οργανικού λιπάσματος.
- η) Να προσδιορισθούν οι συντελεστές r^2 και r_{adj}^2 για το πρόβλημα αυτό και να ερμηνευθούν.
- θ) Να προσδιορισθεί αν, με βάση τα στοιχεία αυτά, υπάρχουν επαρκείς ενδείξεις γραμμικής συσχέτισης μεταξύ της ποσότητας του λιπάσματος και την παραγωγή τομάτας σε επίπεδο σημαντικότητας $\alpha=0.10$.
- ι) Με τη βοήθεια της ανάλυσης των καταλοίπων, να εξετασθεί η καταλληλότητα εφαρμογής του μοντέλου αυτού.

4. Ο διευθυντής μιας εταιρείας υπολογιστών ενδιαφέρεται να κατασκευάσει ένα μοντέλο για να προβλέπει τον αριθμό των φορών που καλείται ανά έτος για επισκευές ενός συγκεκριμένου τύπου υπολογιστή που η εταιρεία του προωθεί. Την πρόβλεψη αυτή θέλει να την στηρίξει στην παλαιότητα των υπολογιστών (σε έτη). Ένα τυχαίο δείγμα δέκα τέτοιων υπολογιστών έδωσε τα αποτελέσματα που ακολουθούν στη συνέχεια.

- α) Να κατασκευαστεί το διάγραμμα σημείων του προβλήματος.
- β) Να κατασκευαστεί η ευθεία παλινδρόμησης με τη μέθοδο των ελαχίστων τετραγώνων.
- γ) Να ερμηνευθεί η τιμή των παραμέτρων a και b του προβλήματος.
- δ) Χρησιμοποιώντας το μοντέλο παλινδρόμησης της ερώτησης (β), να προβλεφθεί ο μέσος αριθμός κλήσεων για επισκευή για ένα υπολογιστή της κατηγορίας αυτής παλαιότητας 4 ετών.

Πρόβλημα επισκευής υπολογιστών

Υπολογιστής	Αριθμός κλήσεων	Ηλικία (έτη)
1	3	1
2	4	1
3	3	2
4	5	2
5	5	3
6	7	3
7	8	4
8	10	4
9	10	5
10	12	5

ε) Να υπολογισθούν οι τιμές του r^2 και r_{adj}^2 και να ερμηνευθούν.

στ) Να κατασκευαστούν το 95% διαστήματα εμπιστοσύνης για το $\mu_{Y|X=4}$ και το 95% διάστημα πρόβλεψης για τον αριθμό κλήσεων για ένα έτος υπολογιστή ηλικίας 4 ετών. Να συγκριθούν τα διαστήματα αυτά.

ζ) Να εξετασθεί, στο 0.05 επίπεδο σημαντικότητας αν υπάρχουν ενδείξεις, από τα δεδομένα, γραμμικής συσχέτισης μεταξύ της ηλικίας των υπολογιστών αυτών και του αριθμού των κλήσεων για επισκευή.

η) Να χρησιμοποιηθεί η ανάλυση καταλοίπων για να εξετασθεί η καταλληλότητα του μοντέλου που χρησιμοποιήθηκε.

5. Ο διευθυντής μιας εταιρείας έρευνας αγοράς ενδιαφέρεται να κατασκευάσει ένα μοντέλο για να προβλέπει τον αριθμό των ερωτηματολογίων που συμπληρώνουν οι ερευνητές του ανά ημέρα. Ο διευθυντής θεωρεί ότι η εμπειρία των συνεντευκτών (μετρούμενη σε εβδομάδες εμπειρίας) είναι το κύριο στοιχείο που καθορίζει τον αριθμό των ερωτηματολογίων που ο κάθε ερευνητής είναι δυνατόν να συμπληρώσει. Χρησιμοποιώντας ένα τυχαίο δείγμα 10 συνεντευκτών, ο διευθυντής κατέγραψε τον αριθμό των ερωτηματολογίων που συμπλήρωσαν οι συνεντευκτές αυτοί, όπως επίσης και την εμπειρία τους (σε εβδομάδες) με τα αποτελέσματα που δίνονται στη συνέχεια.

Για το πρόβλημα αυτό να απαντηθούν οι ερωτήσεις α-ζ της προηγούμενης άσκησης.

Πρόβλημα απόδοσης συνεντευκτών

Εβδομάδες Εμπειρίας	Αριθμός ερωτηματολογίων
15	4
41	9
58	12
18	6
37	8
52	10
28	6
24	5
45	10
33	7

6. Μια εταιρεία ραδιοταξί ενδιαφέρεται να καθορίσει το χρόνο που απαιτείται για να μεταφερθούν επιβάτες από συγκεκριμένα σημεία της πόλης στο αεροδρόμιο. Ένα τυχαίο δείγμα 12 τέτοιων διαδρομών σε μια συγκεκριμένη μέρα έδωσε τα εξής αποτελέσματα:

Να απαντηθούν τα ερωτήματα α-ζ της άσκησης 5 για τα δεδομένα αυτά.

Πρόβλημα με ταξί

Απόσταση (χιλιόμετρα)	Χρόνος (λεπτά)
10.3	19.71
11.6	18.15
12.1	21.88
14.3	24.21
15.7	27.08
16.1	22.96
18.4	29.38
20.2	37.24
21.8	36.84
24.3	40.59
25.4	41.21
26.7	38.19

7. Για τα δεδομένα των προβλημάτων 1-6, να εξετάσετε αν υπάρχουν ακραίες τιμές. Στις περιπτώσεις που υπάρχουν τέτοιες ακραίες τιμές θεωρείτε ότι αυτές θα πρέπει να απομακρυνθούν και γιατί; Στις περιπτώσεις εκείνες στις οποίες θα απομακρύνετε κάποια ακραία τιμή να προχωρήσετε σε νέα ανάλυση παλινδρόμησης με τα υπόλοιπα δεδομένα και να συγκρίνετε τα αποτελέσματα που θα προκύψουν με αυτά που προέκυψαν με την χρήση όλων των δεδομένων.

8. Οι ηλικίες (σε χρόνια) ανδρών και γυναικών έξι παντρεμένων ζευγαριών δίνονται στον πίνακα που ακολουθεί:

Ζευγάρι i	1	2	3	4	5	6
Ηλικίας της συζύγου X_i	35	25	51	25	53	42
Ηλικίας του συζύγου Y_i	38	25	49	31	55	44

Να εξετάσετε τη σχέση που υπάρχει μεταξύ των ηλικιών του συζύγου και την συζύγου σε παντρεμένα ζευγάρια, όπως αυτή προκύπτει από τα δεδομένα του προβλήματος αυτού.

9. Ένας αναλυτής, στην προσπάθειά του να κατασκευάσει ένα μοντέλο παλινδρόμησης της ετήσιας συμμετοχής Y σε ένα συνταξιοδοτικό πρόγραμμα των υπαλλήλων μιας εταιρείας, με χρήση ως ανεξάρτητης μεταβλητής X του αριθμού των χρόνων υπηρεσίας στην εταιρεία, κατέληξε σε μια εκτύπωση του υπολογιστή που έδινε το 95% διάστημα εμπιστοσύνης του $E(Y_i)$ για κάθε υπάλληλο. Ο αναλυτής περίμενε να βρει ότι στις 95% των περιπτώσεων, η συμμετοχή του υπαλλήλου στο συνταξιοδοτικό πρόγραμμα θα βρισκόταν στο σχετικό διάστημα εμπιστοσύνης $E(Y_i)$. Αντί όμως για αυτό, ο αναλυτής παρατήρησε ότι κάτι τέτοιο συνέβαινε σε πολύ μικρότερο ποσοστό των περιπτώσεων. Να εξηγηθεί το λάθος στον τρόπο σκέψης του αναλυτή αυτού.

10. Για κάθε μια από τις ερωτήσεις που ακολουθούν, να εξηγηθεί αν ένα διάστημα εμπιστοσύνης για το $E(Y|x)$ ή ένα διάστημα πρόβλεψης για το Y είναι περισσότερο κατάλληλα:

α) Ποιο θα είναι το ποσοστό ανεργίας το επόμενο τετράμηνο, δοθέντος ότι ο δείκτης επιχειρηματικής δραστηριότητας θα είναι 178.6;

β) Πόσες ώρες ανακούφισης από τον πόνο θα έχει ένας ασθενής με τη χρήση κάποιου συγκεκριμένου φαρμάκου όταν η δόση που θα πάρει είναι 10% παραπάνω από το συνηθισμένο επίπεδο;

11. Να αποδειχθεί ότι οι τυχαίες μεταβλητές $\bar{Y}$ και $\hat{\beta}$ του γραμμικού μοντέλου είναι ασυσχέτιστες.

$$\begin{aligned}\text{Λύση: } \text{cov}(\bar{Y}, \hat{\beta}) &= \text{cov}\left(\frac{1}{n} \sum_i Y_i, \hat{\beta}\right) = \\ &= \frac{1}{n} \sum_i \text{cov}(Y_i, \hat{\beta}) = \\ &= \frac{1}{n} \sum_i c_i \sum_j \text{cov}(Y_i, Y_j) = \\ &= \frac{1}{n} \sum_i c_i \text{cov}(Y_i, Y_i) = \\ &= \frac{1}{n} \sigma^2 \sum_i c_i = 0.\end{aligned}$$

12. Να βρεθεί η συνδιακύμανση των εκτιμητριών $\hat{\alpha}$ και $\hat{\beta}$ των συντελεστών του γραμμικού μοντέλου

$$\mu_{Y|x} = \alpha + \beta x.$$

Είναι οι εκτιμήτριες αυτές ασυσχέτιστες;

$$\begin{aligned}\text{Λύση: } \text{Cov}(\hat{\alpha}, \hat{\beta}) &= \text{Cov}(\bar{Y} - \hat{\beta}\bar{x}, \hat{\beta}) = \\ &= \text{Cov}(\bar{Y}, \hat{\beta}) - \bar{x} \text{cov}(\hat{\beta}, \hat{\beta}) = \\ &= -\bar{x}V(\hat{\beta}) = (\text{λόγω του αποτελέσματος της άσκησης 11}) \\ &= -\frac{\sigma_x^2}{S_{xx}}.\end{aligned}$$

13. Να αποδειχθεί ότι οι εκτιμήτριες $\hat{\alpha}^*$ και $\hat{\beta}$ των συντελεστών του γραμμικού μοντέλου παλινδρόμησης αποκλίσεων από το δειγματικό μέσο είναι ασυσχέτιστες.

Λύση: $\text{cov}(\hat{\alpha}^*, \hat{\beta}) = \text{cov}(\bar{Y}, \hat{\beta}) = 0$.