Ασκήσεις (Απλή Γραμμική παλινδρόμηση)

Ασκ. 1 : Δείξτε ότι ο συντελεστής προσδιορισμού R^2, δίνεται επίσης από τις σχέσεις :

α) $$R^2 = \left(\frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i-\bar{x})^2}\sqrt{\sum_i (y_i - \bar{y})^2}} \right)^2 ,$$

δηλ. το τετράγωνο του δειγματικού (εμπειρικού) συντελεστή γραμμικής συσχέτισης μεταξύ των x και y.

β) $$R^2 = \left(\frac{\sum_i (\hat{y}_i - \bar{y})(y_i - \bar{y})}{\sqrt{\sum_i (\hat{y}_i-\bar{y})^2}\sqrt{\sum_i (y_i - \bar{y})^2}} \right)^2$$

Ασκ. 2 : Η παρακάτω στατιστική μελέτη δείχνει τη σχέση μεταξύ των βαρών 10 πατεράδων και των πρωτότοκων γιών τους :

(Π_i)	πατέρας :	65	63	67	64	68	62	70	66	68	67	69	71
(γ_i)	γιος :	68	66	68	65	69	66	68	65	71	67	68	70

Δίνονται τα παρακάτω αποτελέσματα :

$$\sum_{i=1}^{12} \Pi_i = 800 , \quad \sum_{i=1}^{12} \Pi_i^2 = 53418 , \quad \sum_{i=1}^{12} \Pi_i \gamma_i = 54107 , \quad \sum_{i=1}^{12} \gamma_i = 811 , \quad \sum_{i=1}^{12} \gamma_i^2 = 54849$$

Υπολογίστε : α) Την ευθεία ελαχίστων τετραγώνων $\gamma = \hat{b}_1 + \hat{b}_2 \Pi$.

β) Κάντε το ίδιο για $\Pi = \hat{c}_1 + \hat{c}_2 \gamma$.

γ) Βρείτε το γινόμενο $\hat{b}_2 \hat{c}_2$.

Ασκ. 3 : Δώδεκα άτομα έχουν εγγραφεί σε μία σχολή εκπαίδευσης. Στην αρχή της εκπαίδευσης, οι εκπαιδευόμενοι περνάνε μία πρώτη δοκιμασία A που βαθμολογείται με άριστα το 20. Στο τέλος της εκπαίδευσης μία δεύτερη δοκιμασία B του ίδιου επιπέδου δυσκολίας. Τα αποτελέσματα ήταν τα παρακάτω :

δοκιμή A :	3	4	6	7	9	10	9	11	12	13	15	4
δοκιμή B :	8	9	10	13	15	14	13	16	13	19	6	19

α) Κάντε το διάγραμμα διασποράς των σημείων. Καθορίστε την ευθεία ελαχ. τετραγώνων (εετ). Υπολογίστε το R^2. Σχολιάστε το αποτέλεσμα.

β) Δύο εκπαιδευόμενοι φαίνεται να ξεχωρίζουν από τους υπόλοιπους. Διαγράψτε τους, βρείτε την ε.ε.τ. Υπολογίστε πάλι το R^2 και σχολιάστε.

Ασκ. 4 (Το ύψος των ευκαλύπτων).

Θέλουμε να εξηγήσουμε το ύψος (y) σε m, ενός δέντρου ως συνάρτηση της περιφέρειάς του σε cm, σε ύψος 1m 30 cm από το έδαφος.
Έγιναν $n = 1429$ μετρήσεις (x_i, y_i) και βρέθηκε $(\bar{x}, \bar{y}) = (47.3, 21.2)$.
Επίσης $\sum_i (x_i - \bar{x})^2 = 102924$, $\sum_i (y_i - \bar{y})^2 = 8857$, $\sum_i (x_i - \bar{x})(y_i - \bar{y}) = 26466$

α) Βρείτε την ΕΕΤ για το μοντέλο $y_i = b_1 + b_2 x_i + \varepsilon_i$.

β) Υπολογίστε το R^2 και το $\bar{R}^2$. Σχολιάστε την ποιότητα προσαρμογής των δεδομένων στο μοντέλο.

γ) Δίνεται ότι $\sum_i (y_i - \hat{y}_i)^2 = 2052$. Αν υποθέσουμε ότι έχουμε ένα απλό ΚΓΜ, βρείτε έναν αμερόληπτο εκτιμητή $\hat{\sigma}^2$ του σ^2.

δ) Για $i = 1, 2$, υπολογίστε $\hat{\sigma}^2_{\hat{b}_i}$ (μια εκτιμήτρια της διασποράς του $\hat{b}_i$).

ε) Κάντε τους ελέγχους υποθέσεων H_0: "$b_i = 0$" vs H_1: "$b_i \neq 0$" σε ε.σ. $a = 0.05$

Ασκ. 5

Λέμε "συχνότητα κατώφλι" ενός ερασιτέχνη αθλητή, τη συχνότητα των παλμών της καρδιάς του μετά από 3/4 της ώρας έντονου τρεξίματος. Θέλουμε να δούμε αν η ηλικία του αθλητή έχει επιρροή στη συχνότητα κατώφλι. Πήραμε 20 μετρήσεις (x_i, y_i), όπου x_i είναι η ηλικία και y_i η συχνότητα κατώφλι του i-αθλητή. Πήραμε $(\bar{x}, \bar{y}) = (35.6, 170.2)$ και

$$\sum_i (x_i - \bar{x})^2 = 1991, \quad \sum_i (y_i - \bar{y})^2 = 189.2, \quad \sum_i (x_i - \bar{x})(y_i - \bar{y}) = -195.4.$$

α) Βρείτε την ΕΕΤ για το μοντέλο $y_i = b_1 + b_2 x_i + \varepsilon_i$.

β) Υπολογίστε το συντελεστή προσδιορισμού R^2. Σχολιάστε την ποιότητα προσαρμογής των δεδομένων στο μοντέλο.

γ) Βρέθηκε ότι $\sum_i (y_i - \hat{y}_i)^2 = 170$. Αν το μοντέλο είναι απλό ΚΓΜ, βρείτε έναν αμερόληπτο εκτιμητή $\hat{\sigma}^2$ του σ^2.

δ) Υπολογίστε τη διασπορά $\hat{\sigma}^2_{\hat{b}_2}$.

ε) Ελέγξτε την υπόθεση H_0: "$b_2 = 0$" vs H_1: "$b_2 \neq 0$" σε ε.σ. $a = 0.05$.
Συμπεράνετε αν η ηλικία επηρεάζει τη συχνότητα κατώφλι.

Ασκ. 6 : Δίνεται το στατιστικό μοντέλο :

$$y_i = b x_i + \varepsilon_i \quad , i=1,\dots,n,$$ όπου υποθέτουμε

ότι ε_i είναι τέτοια ώστε $E(\varepsilon_i)=0$ και $Cov(\varepsilon_i,\varepsilon_j)=\sigma^2\delta_{ij}$,

όπου δ_{ij} το δέλτα του Kronecker.

α) Επιστρέφοντας στον ορισμό των ελαχίστων τετραγώνων, δείξτε ότι η εκτιμ. ελαχ. τετραγ. $\hat{b}$ δίνεται από $\hat{b} = \dfrac{\sum_i x_i y_i}{\sum_i x_i^2}$.

β) Δείξτε ότι η ευθεία που περνά από την αρχή των αξόνων και το κέντρο βάρους $(\bar{x},\bar{y})$ είναι η $y=b^* x$, όπου $b^* = \dfrac{\sum_i y_i}{\sum_i x_i}$

γ) Δείξτε ότι $\hat{b}$ και b^* είναι και οι δύο αμερόληπτες εκτιμήτριες του b.

δ) Δείξτε ότι $Var(\hat{b}) < Var(b^*)$, εκτός από την περίπτωση που όλα τα x_i είναι ίσα μεταξύ τους. Αυτό το αποτέλεσμα ήταν αναμενόμενο ?

Ασκ. 7 : Διαθέτουμε n-σημεία $(x_i,y_i)_{1\le i\le n}$ και υποθέτουμε ένα μοντέλο της μορφής $Y_i = a x_i + b + \varepsilon_i$, όπου $\varepsilon=(\varepsilon_1,\dots\varepsilon_n)^t$ και, είναι τέτοιο ώστε

$$E(\varepsilon)=0 \quad \text{και} \quad Cov(\varepsilon)=\sigma^2 I_n$$, όπου I_n ο ταυτοτικός τάξης-n.

α) Δώστε τη μορφή του $\hat{a}$ και του $\hat{b}$ (εκτιμήτριες ελαχίστων τετραγώνων) και τις διασπορές τους.

β) Σε αυτό το ερώτημα, υποθέτουμε ότι γνωρίζουμε το b αλλά όχι το a.

(i) Επιστρέφοντας στον ορισμό των ελαχίστων τετραγώνων, υπολογίστε την εκτιμήτρια ελαχ. τετραγώνων $\tilde{a}$ του a.

(ii) Υπολογίστε τη διασπορά της $\tilde{a}$ και δείξτε ότι $Var(\tilde{a}) < Var(\hat{a})$. Αντιβαίνει το αποτέλεσμα αυτό την ιδιότητα του $\hat{a}$ από το Θεωρ. Gauss-Markov?

γ) Σε αυτό το ερώτημα, υποθέτουμε ότι γνωρίζουμε το a αλλά όχι το b.

(i) Υπολογίστε όπως στο β-(i), την εκτιμ. ελαχ. τετρ. $\tilde{b}$ του b.

(ii) Υπολογίστε τη διασπορά της $\tilde{b}$ και δείξτε ότι $Var(\tilde{b}) < Var(\hat{b})$.