Ασκήσεις Κεφαλαίου 5

1. Δημιουργείστε $M = 1000$ δείγματα μεγέθους $n = 20$ ζευγαρωτών παρατηρήσεων των (X, Y) από διμεταβλητή κανονική κατανομή με μέσες τιμές $\mu_X = 0$, $\mu_Y = 0$, τυπικές αποκλίσεις $\sigma_X = 1$, $\sigma_Y = 1$ και για δύο τιμές του συντελεστή συσχέτισης, $\rho = 0$ και $\rho = 0.5$.

   (α΄) Υπολογίστε το παραμετρικό 95% διάστημα εμπιστοσύνης κάνοντας χρήση του μετασχηματισμού Fisher για κάθε ένα από τα M δείγματα. Κάνετε το ιστόγραμμα για κάθε ένα από τα δύο άκρα του διαστήματος εμπιστοσύνης. Σε τι ποσοστό το διάστημα εμπιστοσύνης περιέχει τη πραγματική τιμή του ρ; Κάνετε τους υπολογισμούς ξεχωριστά για $\rho = 0$ και $\rho = 0.5$.

   (β΄) Κάνετε έλεγχο της υπόθεσης για μηδενική συσχέτιση των X και Y χρησιμοποιώντας το στατιστικό της κατανομής Student t της σχέσης (5.5) για κάθε ένα από τα M δείγματα. Σε τι ποσοστό απορρίπτεται η μηδενική υπόθεση; Κάνετε τους υπολογισμούς ξεχωριστά για $\rho = 0$ και $\rho = 0.5$.

   (γ΄) Επαναλάβεται τους παραπάνω υπολογισμούς για δείγματα μεγέθους $n = 200$. Υπάρχει διαφορά στα αποτελέσματα του διαστήματος εμπιστοσύνης και στατιστικής υπόθεσης;

   (δ΄) Επαναλάβεται τους παραπάνω υπολογισμούς για δείγματα μεγέθους $n = 20$ και $n = 200$ αλλά παίρνοντας τα τετράγωνα των παρατηρήσεων, δηλαδή θεωρείστε τις τ.μ. X^2 και Y^2. Υπάρχει διαφορά στα αποτελέσματα του διαστήματος εμπιστοσύνης και στατιστικής υπόθεσης από τα αντίστοιχα για τις τ.μ. X και Y;

2. Μελετήσαμε τον παραμετρικό έλεγχο για ανεξαρτησία (ή καλύτερα μηδενική συσχέτιση) δύο τ.μ. X και Y κάνοντας χρήση του στατιστικού t της σχέσης (5.5) και θεωρώντας ότι ακολουθεί κατανομή Student. Μπορούμε να κάνουμε τον έλεγχο χωρίς να θεωρήσουμε γνωστή κατανομή του στατιστικού κάτω από τη μηδενική υπόθεση και αυτός λέγεται μη-παραμετρικός έλεγχος. Θα χρησιμοποιήσουμε έναν τέτοιο έλεγχο που λέγεται έλεγχος τυχαιοποίησης και θα δημιουργήσουμε L τυχαιοποιημένα δείγματα από το αρχικό μας διμεταβλητό δείγμα των X και Y σύμφωνα με την μηδενική υπόθεση. Για αυτό θα αλλάξουμε τυχαία τη σειρά όλων των παρατηρήσεων της μιας από τις δύο τ.μ. στο δείγμα, και θα το κάνουμε αυτό L φορές. Στη συνέχεια θα υπολογίσουμε το στατιστικό t της σχέσης (5.5) στο αρχικό δείγμα, έστω t_0, αλλά και στα L τυχαιοποιημένα δείγματα, έστω $t_1, \ldots, t_L$. Η μηδενική υπόθεση απορρίπτεται αν η τιμή t_0 δεν περιέχεται στην κατανομή των $t_1, \ldots, t_L$,

δηλαδή στην εμπειρική (μη-παραμετρική) κατανομή του t κάτω από τη μηδενική υπόθεση της ανεξαρτησίας των X και Y. Συγκεκριμένα για επίπεδο σημαντικότητας a θα εξετάσουμε αν το t_0 είναι μεταξύ των $a/2\%$ και $(1-a/2)\%$ ποσοστιαίων σημείων, δηλαδή μεταξύ των σημείων με σειρά $La/2$ και $L(1-a/2)$ (προσεγγιστικά στον πλησιέστερο ακέραιο), όταν βάζουμε τα $t_1, \ldots, t_L$ σε αύξουσα σειρά.

Θεωρείστε $L = 1000$ τυχαιοποιημένα δείγματα και επαναλάβετε τον έλεγχο, αλλά τώρα τυχαιοποιημένο αντί για παραμετρικό, για την άσκηση 1, για την περίπτωση $n = 20$, και για τα ζευγάρια (X, Y) και (X^2, Y^2). Υπάρχουν διαφορές στα αποτελέσματα;

Βοήθεια (`matlab`*):* Για τη δημιουργία ενός τυχαιοποιημένου δείγματος αρκεί να αλλάξετε τυχαία τη σειρά των παρατηρήσεων της μιας τ.μ.. Μπορείτε να δημιουργείστε μια τυχαιοποιημένη σειρά δεικτών $1, \ldots, n$ με την συνάρτηση `randperm` και όρισμα n. Έτσι αν x είναι το διάνυσμα των n παρατηρήσεων της X και y το αντίστοιχο της Y, τότε αν r = randperm(n) είναι το διάνυσμα τυχαίων δεικτών, θέτοντας xr = x(r) δημιουργούμε το τυχαιοποιημένο δείγμα των xr και y.

3. Δίνονται οι μέσες μηνιαίες τιμές θερμοκρασίας και βροχόπτωσης στη Θεσσαλονίκη για την περίοδο 1959 - 1997. Ελέγξετε αν υπάρχει συσχέτιση μεταξύ της θερμοκρασίας και της βροχόπτωσης για κάθε μήνα ξεχωριστά. Χρησιμοποιείστε το στατιστικό t της σχέσης (5.5) και κάνετε παραμετρικό και έλεγχο τυχαιοποίησης (σύμφωνα με την άσκηση 2). Τα δεδομένα δίνονται στην ιστοσελίδα του μαθήματος σε δύο αρχεία πινάκων, ένας για τη θερμοκρασία και ένας για τη βροχόπτωση, που έχουν 39 γραμμές για τα 39 έτη και 12 στήλες για τους 12 μήνες κάθε έτους, από Ιανουάριο ως Δεκέμβριο.

4. Στο αρχείο `lightair.dat` στην ιστοσελίδα του μαθήματος δίνονται 100 μετρήσεις της πυκνότητας του αέρα (σε $\mathrm{kg/m^3}$) σε διαφορετικές συνθήκες θερμοκρασίας και πίεσης (στην πρώτη στήλη) και οι αντίστοιχες μετρήσεις της ταχύτητας φωτός (-299000 km/sec) στη δεύτερη στήλη.

   (α′) Σχεδιάστε το κατάλληλο διάγραμμα διασποράς και υπολογίστε τον αντίστοιχο συντελεστή συσχέτισης.

   (β′) Εκτιμείστε το μοντέλο γραμμικής παλινδρόμησης με τη μέθοδο ελαχίστων τετραγώνων για τη γραμμική εξάρτηση της ταχύτητας φωτός από την πυκνότητα του αέρα. Υπολογίστε παραμετρικό διάστημα εμπιστοσύνης σε επίπεδο 95% για τους δύο συντελεστές της ευθείας ελαχίστων τετραγώνων (διαφορά ύψους β_0 και κλίση β_1).

(γ′) Σχηματίστε στο διάγραμμα διασποράς την ευθεία ελαχίστων τετραγώνων, τα όρια πρόβλεψης σε επίπεδο 95% για τη μέση ταχύτητα φωτός καθώς και για μια τιμή της ταχύτητας φωτός. Επίσης κάνετε πρόβλεψη για πυκνότητα αέρα $1.29\ \mathrm{kg/m^3}$ δίνοντας και τα όρια μέσης τιμής και μιας παρατήρησης της ταχύτητας φωτός.

(δ′) Η πραγματική σχέση της ταχύτητας φωτός στον αέρα με την πυκνότητα του αέρα είναι:

$$c_{air} = c\left(1 - 0.00029\frac{d}{d_0}\right),$$

όπου οι δύο σταθερές είναι:

- $c = 299792.458\ \mathrm{km/sec}$, η ταχύτητα φωτός στο κενό, και
- $d_0 = 1.29\ \mathrm{kg/m^3}$, η πυκνότητα του αέρα σε θερμοκρασία και πίεση δωματίου.

Από την παραπάνω σχέση υπολογίστε την εξίσωση της πραγματικής ευθείας παλινδρόμησης της ταχύτητας φωτός στον αέρα ως προς την πυκνότητα του αέρα. Στη συνέχεια κάνετε έλεγχο ξεχωριστά για κάθε συντελεστή της πραγματικής ευθείας παλινδρόμησης αν τον δεχόμαστε με βάση το δείγμα των 100 ζευγαρωτών παρατηρήσεων (σύμφωνα με τις εκτιμήσεις τους στο 4β′). Είναι οι πραγματικές μέσες τιμές της ταχύτητας φωτός μέσα στα όρια μέσης πρόβλεψης για κάθε τιμή πυκνότητας αέρα στο δείγμα; (σύμφωνα με το διάστημα μέσης πρόβλεψης που υπολογίσατε και σχηματίσατε στο 4γ′).

5. Θα υπολογίσουμε μη παραμετρικό διάστημα εμπιστοσύνης για τους δύο συντελεστές της ευθείας ελαχίστων τετραγώνων (διαφορά ύψους β_0 και κλίση β_1) και θα το εφαρμόσουμε στα δεδομένα της προηγούμενης άσκησης (έξάρτηση της ταχύτητας φωτός από την πυκνότητα του αέρα). Η μέθοδος που θα χρησιμοποιήσουμε λέγεται bootstrap και για τον υπολογισμό των διαστημάτων εμπιστοσύνης των β_0 και β_1 ορίζεται ως εξής:

   - Πάρε ένα νέο δείγμα 100 τυχαίων ζευγαρωτών παρατηρήσεων από το δείγμα των 100 ζευγαρωτών παρατηρήσεων (πυκνότητας αέρα και ταχύτητας φωτός). Το κάθε ζευγάρι επιλέγεται τυχαία με επανάθεση, δηλαδή το ίδιο ζευγάρι μπορεί να εμφανιστεί πολλές φορές στο νέο δείγμα (και άλλο ζευγάρι να μην εμφανιστεί καθόλου).

- Υπολόγισε τις εκτιμήσεις b_0 και b_1 των συντελεστών της ευθείας ελαχίστων τετραγώνων για αυτό το νέο δείγμα.
- Επανέλαβε τα παραπάνω δύο βήματα $M = 1000$ φορές.
- Από τις M τιμές για το b_0 υπολόγισε τα όρια του $(1 - a)\%$ (εδώ $a = 0.05$) διαστήματος εμπιστοσύνης για το β_0 από τα ποσοστιαία σημεία $a/2\%$ και $(1-a/2)\%$. Για αυτό βάλε τις $M = 1000$ τιμές σε αύξουσα σειρά και βρες τις τιμές για τάξη $Ma/2\%$ και $M(1-a/2)\%$. Κάνε το ίδιο για το b_1.

Σύγκρινε τα bootstrap διαστήματα εμπιστοσύνης των β_0 και β_1 με τα παραμετρικά που βρήκες στην προηγούμενη άσκηση.

Βοήθεια (`matlab`*):* Για τη δημιουργία n τυχαίων αριθμών από 1 ως N με επανάθεση, χρησιμοποίησε τη συνάρτηση `unidrnd` με κατάλληλα ορίσματα ως `unidrnd(`N`,`n`,1)`. Θα δώσει το διάνυσμα δεικτών για το νέο δείγμα από τα n στοιχεία του αρχικού δείγματος (εδώ τα στοιχεία είναι οι ζευγαρωτές παρατηρήσεις).

6. Στον παρακάτω πίνακα δίνονται τα δεδομένα για το ποσοστό υψηλής επίδοσης που ακόμα έχουν ελαστικά (με ακτινωτή ενίσχυση) ενώ έχουν ήδη χρησιμοποιηθεί για τα αντίστοιχα χιλιόμετρα.

Α/Α	*Απόσταση* σε χιλιάδες km	*ποσοστό δυνατότητας* χρήσης
1	2	98.2
2	3	91.7
3	8	81.3
4	16	64.0
5	32	36.4
6	48	32.6
7	64	17.1
8	80	11.3

(α′) Κάνε το διάγραμμα διασποράς και προσάρμοσε το κατάλληλο μοντέλο για τον προσδιορισμό του ποσοστού δυνατότητας χρήσης (υψηλής επίδοσης) προς τα αντίστοιχα χιλιόμετρα χρήσης του ελαστικού. Για να ελέγξεις την καταλληλότητα του μοντέλου, κάνε διαγνωστικό διάγραμμα διασποράς των τυποποιημένων σφαλμάτων προσαρμογής του επιλεγμένου μοντέλου προς την εξαρτημένη μεταβλητή (ποσοστό δυνατότητας χρήσης).

(β′) Πρόβλεψε το ποσοστό δυνατότητας χρήσης για ελαστικό που χρησιμοποιήθηκε για 25000km.

7. Ο θερμοστάτης είναι αντιστάτης με αντίσταση που εξαρτιέται από τη θερμοκρασία. Είναι φτιαγμένος (συνήθως) από ημι-αγωγό υλικό με ενεργειακό διάκενο E_g. Η αντίσταση R του θερμοστάτη αλλάζει σύμφωνα με τη σχέση

$$R \propto R_0 e^{E_g/2kT},$$

όπου T είναι η θερμοκρασία (σε ^{o}K) και R_0, k είναι σταθερές. Για κατάλληλες παραμέτρους β_0 και β_1 ($\beta_1 = 2k/E_g$) η παραπάνω εξίσωση μπορεί να απλοποιηθεί στη γραμμική μορφή

$$\frac{1}{T} = \beta_0 + \beta_1 \ln(R).$$

Το ενεργειακό διάκενο E_g έχει κάποια μικρή εξάρτηση από τη θερμοκρασία έτσι ώστε η παραπάνω έκφραση να μην είναι ακριβής. Διορθώσεις μπορούν να γίνουν προσθέτοντας πολυωνυμικούς όρους του $\ln(R)$.

Στον παρακάτω πίνακα δίνονται 32 μετρήσεις της αντίστασης R (σε Ωμ) και της θερμοκρασίας σε ^{o}C (θα πρέπει να μετατραπούν σε ^{o}K, δηλαδή να προστεθεί σε κάθε τιμή 273.15).

(α′) Βρείτε το κατάλληλο πολυωνυμικό μοντέλο της παλινδρόμησης του $1/T$ ως προς $\ln(R)$, κάνοντας διαγνωστικό έλεγχο με το διάγραμμα διασποράς των τυποποιημένων υπολοιπών προς $1/T$ για κάθε μοντέλο που δοκιμάζετε (πρώτου βαθμού, δευτέρου βαθμού κτλ).

(β′) Συγκρίνετε την προσαρμογή και κατλληλότητα του μοντέλου που καταλήξατε με το μοντέλο του Steinhart-Hart που δίνεται από την εξίσωση

$$\frac{1}{T} = \beta_0 + \beta_1 \ln(R) + \beta_3 (\ln(R))^3.$$

Α/Α	*Αντίσταση*	*θερμοκρασία (σε ^oC)*
1	0.76	110
2	0.86	105
3	0.97	100
4	1.11	95
5	1.45	85
6	1.67	80
7	1.92	75
8	2.23	70
9	2.59	65
10	3.02	60
11	3.54	55
12	4.16	50
13	4.91	45
14	5.83	40
15	6.94	35
16	8.31	30
17	10.00	25
18	12.09	20
19	14.68	15
20	17.96	10
21	22.05	5
22	27.28	0
23	33.89	-5
24	42.45	-10
25	53.39	-15
26	67.74	-20
27	86.39	-25
28	111.30	-30
29	144.00	-35
30	188.40	-40
31	247.50	-45
32	329.20	-50

8. Μετρήθηκε το βάρος και 10 δείκτες σώματος σε 22 άνδρες νεαρής ηλικίας. Τα δεδομένα δίνονται στην ιστοσελίδα του μαθήματος στο αρχείο `physical.txt`. Η πρώτη γραμμή του πίνακα του αρχείου έχει τα ονόματα των δεικτών σε κάθε στήλη δεδομένων και δίνονται στον παρακάτω πίνακα.

Α/Α	*Όνομα*	*Περιγραφή*
1	Mass	Βάρος σε κιλά
2	Fore	μέγιστη περιφέρεια του πήχυ χεριού
3	Bicep	μέγιστη περιφέρεια του δικέφαλου μυ
4	Chest	περιμετρική απόσταση στήθους (στο ύψος κάτω από τις μασχάλες)
5	Neck	περιμετρική απόσταση λαιμού (στο μέσο ύψος λαιμού)
6	Shoulder	περιμετρική απόσταση ώμου
7	Waist	περιμετρική απόσταση μέσης (οσφίου)
8	Height	ύψος από την κορυφή στα δάχτυλα ποδιού
9	Calf	μέγιστη περιφέρεια κνήμης
10	Thigh	περιμετρική απόσταση γοφού
11	Head	περιμετρική απόσταση κεφαλιού

Διερευνείστε το κατάλληλο μοντέλο γραμμικής παλινδρόμησης για το βάρος. Δοκιμάστε το μοντέλο με τις 10 ανεξάρτητες μεταβλητές και συγκρίνετε το με το μοντέλο που δίνει κάποια μέθοδος βηματικής παλινδρόμησης. Υπολογίστε για το κάθε μοντέλο τις εκτιμήσεις των παραμέτρων, τη διασπορά των σφαλμάτων και το συντελεστή προσδιορισμού (καθώς και τον προσαρμοσμένο συντελεστή προσδιορισμού).

Βοήθεια (`matlab`*):* Για να εφαρμόσετε βηματική παλινδρόμηση, το `matlab` παρέχει γραφικό περιβάλλον με την εντολή `stepwise` και κάνει τους ίδιους υπολογισμούς στη συνάρτηση `stepwisefit`. Για να φορτώσετε τα δεδομένα του αρχείου με την εντολή `load` θα πρέπει πρώτα να διαγράψετε την πρώτη σειρά με τα ονόματα των μεταβλητών.

9. Μετρήθηκαν σε 12 νοσοκομεία των ΗΠΑ οι μηνιαίες ανθρωπο-ώρες που σχετίζονται με την υπηρεσία αναισθησιολογίας, καθώς και άλλοι δείκτες που ενδεχομένως επηρεάζουν την απασχόληση προσωπικού στην υπηρεσία αναισθησιολογίας. Τα δεδομένα δίνονται στην ιστοσελίδα του μαθήματος στο αρχείο `hospital.txt`. Η πρώτη γραμμή του πίνακα του αρχείου έχει τα ονόματα των δεικτών σε κάθε στήλη δεδομένων και δίνονται στον παρακάτω πίνακα.

Α/Α	*Όνομα*	*Περιγραφή*
1	ManHours	οι ανθρωπο-ώρες στην υπηρεσία αναισθησιολογίας
2	Cases	τα περιστατικά χειρουργίου μηνιαία
3	Eligible	ο πληθυσμός που εξυπηρετείται ανά χιλιάδες
4	OpRooms	οι αίθουσες χειρουργείου

Διερευνείστε το κατάλληλο μοντέλο γραμμικής παλινδρόμησης για τις ανθρωπο-ώρες. Δοκιμάστε το μοντέλο με τις 3 ανεξάρτητες μεταβλητές και συγκρίνετε το με το μοντέλο που δίνει κάποια μέθοδος βηματικής παλινδρόμησης. Υπολογίστε για το κάθε μοντέλο τις εκτιμήσεις των παραμέτρων, τη διασπορά των σφαλμάτων και το συντελεστή προσδιορισμού (καθώς και τον προσαρμοσμένο συντελεστή προσδιορισμού). Επίσης διερευνείστε το φαινόμενο πολλαπλής συγγραμικότητας για τους 3 δείκτες.