

Στατιστικά Προγράμματα, 20-12-06, Homework #2

A. Στο αρχείο **titanic.txt** περιέχονται δεδομένα που αφορούν τους επιβάτες του Τιτανικού που ναυάγησε στις αρχές του προηγούμενου αιώνα (εξαιρείται το πλήρωμα του πλοίου). Από τις μεταβλητές που καταγράφονται μας ενδιαφέρουν οι: SURVIVED: με τιμή 0 αν ο επιβάτης δεν διασώθηκε και τιμή 1 αν το άτομο διασώθηκε, CLASS: κλάση ή θέση επιβίβασης επιβάτη με τιμές 1η, 2η, 3η, και GENDER: φύλο επιβάτη με 0 - female, 1 - male (επίσης καταγράφονται και οι μεταβλητές NAME και AGE που δεν θα χρειαστούμε στην ανάλυση αυτή). Να αριθμήσετε τα cases (μεταβλητή i) και αν x το τελευταίο ψηφίο του AM σας, **να αφαιρέσετε τα cases με $i = 100 + x, 500 + x, 1000 + x$.**

A1. Να κατασκευάσετε έναν πίνακα συχνοτήτων (crosstabulation) που να περιλαμβάνει τις μεταβλητές CLASS (rows), SURVIVED (columns), GENDER (layer 1). Σε κάθε γραμμή του πίνακα να φαίνονται και τα ποσοστά αυτών που επέζησαν και δεν επέζησαν. Σε ποια κατηγορία καταγράφεται το μεγαλύτερο και σε ποιά το μικρότερο ποσοστό επιβίωσης;

A2. (α) Να κάνετε ένα πίνακα συχνοτήτων και ένα ραβδόγραμμα (Barchart/stacked) που να απεικονίζει πόσοι επέζησαν και πόσοι όχι ανά φύλο. Στον πίνακα να φαίνονται και τα αντίστοιχα ποσοστά ανά φύλο, ενώ σε κάθε ράβδο του ραβδογράμματος να φαίνεται και το αντίστοιχο πλήθος επιβατών. Αν η διάσωση ενός επιβάτη ήταν ανεξάρτητη από το φύλο, ποιες θα ήταν οι αναμενόμενες συχνότητες; Ήταν τελικά η διάσωση ανεξάρτητη από το φύλο; **(β)** Να επαναλάβετε τα παραπάνω, αυτή τη φορά για κάθε θέση επιβίβασης ξεχωριστά (split file). Σχολιάστε τα αποτελέσματα.

A3. (α) Να επαναλάβετε το A2(α), τώρα ανά θέση επιβίβασης αντί ανά φύλο. **(β)** Να επαναλάβετε το A3(α), αυτή τη φορά για κάθε φύλο ξεχωριστά (split file). Σχολιάστε τα αποτελέσματα.

B. Κατά την διάρκεια μιας προεκλογικής περιόδου έχει προγραμματισθεί να πραγματοποιηθεί ένα debate μεταξύ των αρχηγών των δύο μεγαλύτερων κομμάτων (ΝΔ και ΠΑΣΟΚ). Μία εταιρία δημοσκοπήσεων διεξαγάγει μία έρευνα σχετικά με το αν η παρακολούθηση του συγκεκριμένου debate επιδρά στις προτιμήσεις των ψηφοφόρων. Για το σκοπό αυτό ερωτήθηκαν 750 ψηφοφόροι σχετικά με την προτίμησή τους για τον έναν ή τον άλλο αρχηγό πριν και μετά την παρακολούθηση του debate. Τα αποτελέσματα φαίνονται στον ακόλουθο πίνακα (υποθετικά δεδομένα, x είναι το τελευταίο ψηφίο του AM σας).

		Προτίμηση μετά το debate	
		Παπανδρέου	Καραμανλής
Προτίμηση πριν το debate	Παπανδρέου	270	$90 - x$
	Καραμανλής	$110 + x$	280

Υπήρξε τελικά στατιστικά σημαντική επίδραση του debate στις προτιμήσεις των ψηφοφόρων και αν ναι, υπέρ ποιού;

G1. Στο University of California at Berkeley έκαναν αίτηση για να γίνουν δεκτοί για σπουδές (στο τμήματα Law και Business) 700 αγόρια και 500 κορίτσια. Από αυτούς έγιναν τελικά δεκτοί 770 όπως φαίνεται στον παρακάτω πίνακα (βλ. άρθρο: Bickel P.J., Hammel P.A., O'Connell J.W. (1975) "Sex bias in Graduate Admissions: Data from Berkeley", Science, ή το άρθρο: Wagner, C.H. (1982) "Simpson's paradox in real life", The American Statistician)

Business and Law Schools		
	Admit	Deny
Male	490	210
Female	280	220

Να κατασκευάσετε τον παραπάνω πίνακα μαζί με τα ποσοστά αυτών που έγιναν δεκτοί (ανά γραμμή). Να ελέγξετε αν τα αγόρια και τα κορίτσια γίνονται δεκτά με την ίδια πιθανότητα. Ποιοι φαίνεται να «ευνοούνται»;

G2. Ας εξετάσουμε τώρα ακριβώς τα ίδια δεδομένα αλλά λαμβάνοντας υπόψη και τη σχολή προτίμησης:

		Admit Business + Law	Deny Business + Law
	Male	$480 + 10$	$120 + 90$
	Female	$180 + 100$	$20 + 200$

(α) Κατασκευάστε τώρα δύο πίνακες παρόμοιους με αυτόν στο G1 παραπάνω, ένα για κάθε σχολή (χρησιμοποιήστε κατάλληλο layer). Να ελέγξετε, σε κάθε σχολή ξεχωριστά, αν τα αγόρια και τα κορίτσια γίνονται δεκτά με την ίδια πιθανότητα. Ποιοι φαίνεται να «ευνοούνται» σε κάθε σχολή; **(β)** Υπάρχει κάποια αντίφαση ανάμεσα στα συμπεράσματα των G1, G2(α); Ποιο συμπέρασμα είναι αποδεκτό; Μπορείτε να δώσετε μία ερμηνεία;

Δ. Ο δείκτης FEV (Forced Expiratory Volume) μετράει τον όγκο του αέρα που εκπνέεται από ένα άτομο που φυσά συνεχώς και με πίεση για ένα δευτερόλεπτο. Ο δείκτης FEV έχει μετρηθεί σε ένα σύνολο παιδιών με ηλικίες από 6 έως 19 ετών (Childhood Respiratory Disease Study in 1980 in East Boston, Massachusetts). Συνολικά, σε καθένα άτομο καταγράφονται οι μεταβλητές: Age (years), FEV (litres), Height (cm), Sex (1 = Male, 0 = Female), Smoker (0 = nonsmoker, 1 = current smoker). Χρησιμοποιήστε το αρχείο **fev.sav**. Να αριθμήσετε τα cases (μεταβλητή i) και ο φοιτητής-τρια με x το τελευταίο ψηφίο του AM του να **αφαιρέσει τα cases με $i = 100 + x, 200 + x, \dots, 600 + x$.** Στο ερώτημα Δ θα διερευνήσουμε τη σχέση μεταξύ του ύψους και της ηλικίας. Για όλο το ερώτημα Δ, **κρατήστε μόνο τις παρατηρήσεις με ηλικίες 6 – 13 ετών.**

Δ1. Να εφαρμόσετε τα δύο μοντέλα ($\varepsilon \sim N(0, \sigma^2)$),

$$(M1) \text{ Height} = b_0 + b_1 \text{ Age} + \varepsilon \quad (\text{γραμμικό}), \quad (M2) \text{ Height} = b_0 + b_1 \text{ Age} + b_2 \text{ Age}^2 + \varepsilon \quad (\text{τετραγωνικό})$$

(α) Να κάνετε τα αντίστοιχα διαγράμματα διασποράς των παρατηρήσεων στο επίπεδο μαζί με τις εκτιμημένες καμπύλες παλινδρόμησης (γραμμική και τετραγωνική) και εκτιμήστε τα b_i και το σ^2 σε κάθε περίπτωση. **(β)** Το ύψος μεταβάλλεται με την ηλικία; Τι ποσοστό της μεταβλητότητας των παρατηρήσεων του ύψους ερμηνεύεται από το κάθε μοντέλο και ποιο από τα δύο θεωρείτε καλύτερο; **(γ)** Να κάνετε και πάλι το διάγραμμα διασποράς των παρατηρήσεων, αυτή τη φορά με διαφορετικό χρώμα τα σημεία που αντιστοιχούν σε κορίτσια και αγόρια. Φαίνεται ότι πρέπει να λάβουμε υπόψη και το φύλο;

Δ2. Έστω ότι το τετραγωνικό μοντέλο στα κορίτσια και στα αγόρια είναι αντίστοιχα

$$\text{Height} = a_0 + a_1 \text{Age} + a_2 \text{Age}^2 + \varepsilon \quad \text{και} \quad \text{Height} = (a_0 + c_0) + (a_1 + c_1) \text{Age} + (a_2 + c_2) \text{Age}^2 + \varepsilon$$

(c_i είναι οι διαφορές των παραμέτρων μεταξύ των δύο μοντέλων). Ενώνουμε τα δύο παραπάνω σε ένα ισοδύναμο μοντέλο:

$$(M3) \quad \text{Height} = (a_0 + c_0 \text{Sex}) + (a_1 + c_1 \text{Sex}) \text{Age} + (a_2 + c_2 \text{Sex}) \text{Age}^2 + \varepsilon \\ = a_0 + c_0 \text{Sex} + a_1 \text{Age} + c_1 \text{Sex} * \text{Age} + a_2 \text{Age}^2 + c_2 \text{Sex} * \text{Age}^2 + \varepsilon$$

Εφαρμόστε το μοντέλο (M3) και εκτιμήστε τα $a_0, c_0, a_1, c_1, a_2, c_2, \sigma^2$. Ποιες από τις ερμηνευτικές μεταβλητές φαίνεται να επηρεάζουν το ύψος; Τι ποσοστό της μεταβλητότητας των παρατηρήσεων του ύψους ερμηνεύεται από το μοντέλο; Είναι το μοντέλο αυτό καλύτερο από το (M2) και γιατί;

Δ3. Εφαρμόστε το (M4) που είναι το μοντέλο (M3) θεωρώντας ότι $c_0 = 0$ και εκτιμήστε τις παραμέτρους του. Είναι το μοντέλο αυτό καλύτερο από το (M3) και γιατί; Στο (M4), το ύψος επηρεάζεται διαφορετικά από την ηλικία στα αγόρια και στα κορίτσια; Ερμηνεύστε τα αποτελέσματα.

Δ4. Στο (M4): Υπάρχουν έκτροπες ή ασυνήθιστες παρατηρήσεις και αν ναι, ποιες είναι; Ελέγξτε αν ισχύει η κανονικότητα των υπολοίπων (ιστόγραμμα, Q-Q plot και K-S τεστ). Εξετάστε (μέσω κατάλληλου γραφήματος) αν υπάρχει σχέση μεταξύ των προσαρμοσμένων τιμών των Y_i και των τυποποιημένων καταλοίπων.

Δ5. Στο (M4): **(α)** Να κατασκευάσετε έναν πίνακα που για κάθε ηλικία (6, 7, ..., 13 ετών) να δίνει την πρόβλεψη (σημειική, και με διάστημα εμπιστοσύνης 95%) για το μέσο ύψος ενός αγοριού και ενός κοριτσιού. Αν επιλεγεί τυχαία (από τον πληθυσμό από τον οποίο προήλθε το δείγμα) ένα κορίτσι 13 ετών, μεταξύ ποιών ορίων θα βρίσκεται το ύψος του; (σ.ε. 95%). **(β)** Μπορούμε να κάνουμε πρόβλεψη με το ίδιο μοντέλο για τα ύψη που αντιστοιχούν σε ηλικίες μεγαλύτερες από 13 έτη; Να κάνετε ένα γράφημα με τις δύο καμπύλες παλινδρόμησης (μία για τα αγόρια και μία για τα κορίτσια) και τις παρατηρήσεις με διαφορετικά χρώματα για αγόρια – κορίτσια (υπόδειξη: στο τέλος των δεδομένων προσθέστε 8 γραμμές με ηλικίες 6, ..., 13 και φύλο male και 8 γραμμές με ηλικίες 6, ..., 13 και φύλο female. Στις αντίστοιχες θέσεις της height προσθέστε τις αντίστοιχες σημειακές προβλέψεις που βρήκατε στο Δ5(α). Φτιάξτε μία νέα μεταβλητή group που να είναι ίση με 2 στα νέα αυτά 16 cases και 0 αλλού. Τέλος κάνετε Scatterplot της height με την Age θέτοντας Markers την sex+group και μετά ενώστε τα σημεία των group 2 και 3).

Ε. Χρησιμοποιώντας τα **πλήρη** δεδομένα του Δ (δηλαδή **όλες τις ηλικίες**) επιθυμούμε τώρα να διερευνήσουμε αν επηρεάζεται ο δείκτης FEV από το κάπνισμα (και **πάλι αφαιρέστε τα cases με $i = 100+x, 200+x, \dots, 600+x$**).

Ε1. Να κάνετε έναν πίνακα με όλα τα διαγράμματα διασποράς των μεταβλητών {FEV, Age, Height} ανά δύο (Scatterplot/matrix) στα οποία οι καπνιστές – μη καπνιστές να είναι με διαφορετικό χρώμα. Με ποιες μεταβλητές φαίνεται να σχετίζεται η μεταβλητή FEV από τα διαγράμματα;

Ε2. Είναι σωστό να κάνουμε t -τεστ για τον έλεγχο του μέσου FEV μεταξύ καπνιστών και μη; Κάνετε t -τεστ και σχολιάστε το αποτέλεσμα. Πού μπορεί να οφείλεται αυτό που παρατηρείτε;

Ε3. Για να ληφθεί υπόψη και η επίδραση των άλλων ερμηνευτικών μεταβλητών, εφαρμόστε το πλήρες μοντέλο (κατασκευάζεται όπως και το ανάλογο στο Δ2):

$$(M5) \quad \text{FEV} = (a_0 + c_0 \text{Sex} + \kappa_0 \text{Smoke}) + (a_1 + c_1 \text{Sex} + \kappa_1 \text{Smoke}) \text{Age} \\ + (a_2 + c_2 \text{Sex} + \kappa_2 \text{Smoke}) \text{Height} + (a_3 + c_3 \text{Sex} + \kappa_3 \text{Smoke}) \text{Height}^2 + \varepsilon$$

Δώστε τις εκτιμήσεις των παραμέτρων του και το R^2 . Ποιες από τις 11 ερμηνευτικές μεταβλητές επηρεάζουν το FEV;

Ε4. Εφαρμόστε το μικρότερο μοντέλο

$$(M6) \quad \text{FEV} = a_0 + c_0 \text{Sex} + a_1 \text{Age} + \kappa_1 \text{Smoke} * \text{Age} + a_2 \text{Height} + a_3 \text{Height}^2 + c_3 \text{Sex} * \text{Height}^2 + \varepsilon$$

Δώστε τις εκτιμήσεις των παραμέτρων του και το R^2 . Μπορεί το (M6) να θεωρηθεί καλύτερο από το (M5) και γιατί; Με βάση το (M6), ο δείκτης FEV τελικά επηρεάζεται από το κάπνισμα;

Ε5. Με βάση το (M6), **(α)** Αν ένα κορίτσι ηλικίας 18 ετών και ύψους 165cm επιλεγεί τυχαία από τον πληθυσμό, κάντε πρόβλεψη (σημειακή και μέσω διαστήματος) για το FEV αν (i) είναι καπνίστρια, (ii) δεν είναι καπνίστρια. **(β)** Να κατασκευάσετε έναν πίνακα ο οποίος να δείχνει για κάθε ηλικία την (εκτιμώμενη μέση) διαφορά του FEV μεταξύ ενός ατόμου που καπνίζει και ενός που δεν καπνίζει.

ΣΤ. Είναι γνωστό ότι το πτηνό «κούκος» γεννά τα αυγά του σε φωλιές άλλων πουλιών τα οποία και «υιοθετούν» τα αυγά. Μέσα από την διαδικασία της φυσικής επιλογής ο κούκος χρησιμοποιεί τις φωλιές συγκεκριμένων ειδών πουλιών. Επιθυμούμε να διερευνήσουμε αν τα αυγά του κούκου αναπτύσσονται το ίδιο σε διαφορετικές φωλιές πουλιών. Για το λόγο αυτό μετρήθηκε το μήκος 115 τυχαία επιλεγμένων αυγών (μεταβλητή Length σε mm) που βρέθηκαν σε φωλιές 6 ειδών πτηνών. Σε κάθε περίπτωση καταγράφεται και το είδος του πτηνού που «υιοθέτησε» το αυγό (μεταβλητή species με τιμές 1: Meadow Pipit, 2: Tree Pipit, 3: Hedge Sparrow, 4: Robin, 5: Pied Wagtail, 6: Wren). Χρησιμοποιήστε τα αρχείο **Cuckoo.sav**. Ο φοιτητής-τρια με x το τελευταίο ψηφίο του ΑΜ του να **αφαιρέσει την $x+1$ παρατήρηση** που αντιστοιχεί σε κάθε ένα από τα 6 είδη πτηνών. Να εφαρμόσετε ένα μοντέλο ανάλυσης διασποράς του μήκους με παράγοντα το είδος του πτηνού. Αν $\mu_1, \mu_2, \dots, \mu_6$ το μέσο μήκος του αυγού του κούκου σε κάθε ένα είδος:

ΣΤ1. Κατασκευάστε ένα κοινό scatterplot και ένα κοινό Boxplot του μήκους ανά είδος πτηνού. Εκτιμήστε τα μ_i , σημειακά και με δ.ε. 95%. Να κατασκευάσετε και ένα γράφημα με τα δ.ε. των μ_i (error bar).

ΣΤ2. Ελέγξτε την υπόθεση $H_0: \mu_1 = \mu_2 = \dots = \mu_6$ και εκτιμήστε τη διασπορά των σφαλμάτων. Να ελέγξετε αν το μήκος των αυγών έχει την ίδια διασπορά σ^2 σε όλες τις φωλιές των διαφόρων ειδών

ΣΤ3. Να δοθούν ταυτόχρονα δ.ε. (με συνολικό σ.ε. 95%) για όλες τις διαφορές $\mu_i - \mu_j$, με την μεθόδους Tukey και LSD. Ποια είναι η διαφορά των δύο μεθόδων; Να γίνει ομαδοποίηση των μέσων ώστε η συνολική πηθ. σφάλματος να είναι 5%.

- όλοι οι έλεγχοι να γίνουν σε ε.σ. $\alpha = 5\%$.