

Στατιστικά Προγράμματα, 19-12-07, Homework #2

ΜΕΡΟΣ Α. Με βάση τα δεδομένα (αρχείο q2007data.sav) από τις απαντήσεις του ερωτηματολογίου (questionnaire2007.doc) που μοιράστηκε στους 4ετείς φοιτητές του τμήματος στις 31-10-07, να κάνετε τα ακόλουθα:

- **Αριθμήστε τα cases από 1-152** (χρησιμ. \$CASENUM) **και αφαιρέστε τα cases με αύξοντα αριθμό 10+x, 20+x, 30+x, 40+x, όπου x το τελευταίο ψηφίο του ΑΜ σας**

(A0) Σε όσες από τις ερωτήσεις 21 έως και 31 οι απαντήσεις (q21-q31) είναι:

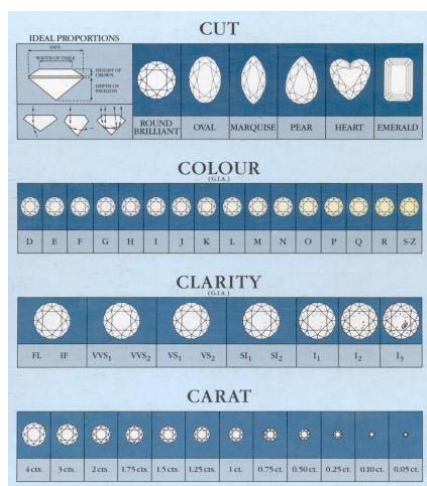
- (i) στην κλίμακα 1-6, να επανακωδικοποιηθούν στη νέα κλίμακα 0-1 με 1,2,3→0 και 4,5,6→1.
- (ii) στην κλίμακα 1-5, να επανακωδικοποιηθούν στη νέα κλίμακα 0-1 με 1,2→0 και 3,4,5→1.
- (iii) στην κλίμακα 1-4, να επανακωδικοποιηθούν στη νέα κλίμακα 0-1 με 1,2→0 και 3,4→1.

Να ονομάσετε q21c, q22c, ..., q31c τις νέες μεταβλητές.

(A1) Να εξετάσετε τις απαντήσεις της (επανακωδικοποιημένης) ερώτησης 23 (q23c) *ανά φύλο*. Συγκεκριμένα, (α) Να κατασκευάσετε τον αντίστοιχο πίνακα συνάφειας με τις παρατηρούμενες συχνότητες, τις αναμενόμενες (υπό την υπόθεση της ανεξαρτησίας) συχνότητες, και τα ποσοστά των απαντήσεων ανά φύλο (% within q1). (β) Αν οι απαντήσεις ήταν ανεξάρτητες του φύλου, πόσοι φοιτητές αναμένονταν να δώσουν μία από τις τρεις πρώτες απαντήσεις ("Εξαντλητικά", "Πάρα πολύ", "Πολύ") και πόσοι τελικά έδωσαν; (γ) Τι ποσοστό των απαντήσεων των φοιτητών ήταν "Εξαντλητικά" ή "Πάρα πολύ" ή "Πολύ" και ποιο το αντίστοιχο ποσοστό των απαντήσεων των φοιτητριών; (δ) Να εξετάσετε αν οι φοιτητές και οι φοιτητρίες απαντούν με διαφορετικό τρόπο την (επανακωδικοποιημένη) ερώτηση 23. (ε) Να απαντήσετε και πάλι τα παραπάνω (α) και (δ), ξεχωριστά για τα άτομα που προήλθαν από την Θετική, Θεωρητική και Τεχνολογική κατεύθυνση (χρησιμοποιήστε κατάλληλο layer).

(A2) (α) Να εξετάσετε σε ποιες από τις (επανακωδικοποιημένες) ερωτήσεις 21-31, οι απαντήσεις εξαρτώνται από το φύλο (q1). Συγκεκριμένα, να παρουσιάσετε τους πίνακες συνάφειας (γραμμές: q1, στήλες: q21-q31) μαζί με τα αντίστοιχα ποσοστά (ανά γραμμή) **μόνο** στις περιπτώσεις που βρεθεί στατιστικά σημαντική διαφορά στις απαντήσεις ανά φύλο. Εάν είναι εύκολο, προσπαθήστε να δώσετε μία διαισθητική ερμηνεία στις περιπτώσεις που βρεθεί στατιστικά σημαντική διαφορά. (β) Να εξετάσετε σε ποιες από τις (επανακωδικοποιημένες) ερωτήσεις 21-31, οι απαντήσεις εξαρτώνται από την καταγωγή (q2), τη σειρά δήλωσης του τμήματος (σε στάθμες «υψηλή», «μεσαία», «χαμηλή», όπως η q3b που κατασκευάστηκε στην 1^η εργασία), την κατεύθυνση εισαγωγής (q4) και την πρόθεση για μεταπτυχιακές σπουδές (q5). Παρουσιάστε όμοια με το (α) **μόνο** τις περιπτώσεις που βρεθεί στατιστικά σημαντική διαφορά και δώστε διαισθητική ερμηνεία.

ΜΕΡΟΣ Β. Επιθυμούμε να βρούμε ένα ικανοποιητικό μοντέλο τιμολόγησης διαφόρων τύπων διαμαντιών με βάση συγκεκριμένα χαρακτηριστικά τους (Κοπή, Χρώμα, Καθαρότητα, Καράτια). Για το σκοπό αυτό έχουν καταγραφεί σε 308 διαμάντια υψηλής ποιότητας που διατίθενται προς πώληση οι ακόλουθες μεταβλητές (από πραγματική διαφήμιση στην εφημερίδα Singapore's *Business Times* edition of February 18, 2000, βλ. S.Chu (2001) *Journal of Statistics Education*, Volume 9, Number 2):



- (i) **Βάρος σε καράτια (Carats)** (1 carat = 0.2gr),
- (ii) **Χρώμα (Colour)** σε 6 στάθμες: D, E, F, G, H, I, από εντελώς άχρωμο μέχρι σχεδόν άχρωμο ή ελαφρώς κίτρινο.
- (iii) **Καθαρότητα (Clarity)** σε 5 στάθμες: IF (internally flawless - αψεγάδιαστο), VVS1, VVS2 (very very slightly imperfect 1 και 2), VS1, VS2 (very slightly imperfect 1 και 2).
- (iv) Το **πιστοποιητικό (CB, Certification Body)** που εγγυάται τα παραπάνω τρία χαρακτηριστικά με 3 δυνατές τιμές: πιστοποιητικό από GIA (New York based Gemmological Institute of America), από IGI (Antwerp based International Gemmological Institute) και από HRD (Hoge Raad Voor Diamant).
- (v) **Τιμή (Price)**, σε δολάρια Σιγκαπούρης (1 Singapore dollar ≈ 0.692 U.S. dollars) (Όλα τα διαμάντια του δείγματος είναι κοπής μπριγιάν (round diamond cut) και επομένως η τιμολόγηση θα γίνει μόνο για αυτό το είδος κοπής. Άλλα λιγότερο δημοφιλή είδη είναι «πρινσέττα», «καρδιά», «μαρκίζα», «σμαράγδιν», «δάκρυ», «οβάλ»).

Τα δεδομένα βρίσκονται στο αρχείο Diamonds.txt που βρίσκεται στην ιστοσελίδα του μαθήματος.

- **Αριθμήστε τα cases από 1-308 και αφαιρέστε τα cases με αύξοντα αριθμό 10+x, 20+x, 30+x, 40+x, όπου x το τελευταίο ψηφίο του ΑΜ σας.**

(B1) (α) Να εφαρμόσετε τα ακόλουθα τρία μοντέλα ($\varepsilon \sim N(0, \sigma^2)$),

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

$$(M3) \ln(\text{Price}) = b_0 + b_1 \text{ Carats} + b_2 \text{ Carats}^2 + \varepsilon \quad (\text{τετραγωνικό επί του λογαρίθμου της τιμής})$$

Να κάνετε τα αντίστοιχα διαγράμματα διασποράς των παρατηρήσεων στο επίπεδο μαζί με τις εκτιμημένες καμπύλες παλινδρόμησης (γραμμική - linear και τετραγωνική - quadratic) και εκτιμήστε τα b_i και το σ^2 σε κάθε περίπτωση. Τι ποσοστό της μεταβλητότητας των τιμών ερμηνεύεται από το κάθε μοντέλο;

(B2) Ποιο από τα τρία θεωρείτε ορθότερο; Συγκεκριμένα ελέγξτε αν ισχύει η κανονικότητα των υπολοίπων (ιστόγραμμα, Q-Q plot και K-S τεστ) και εξετάστε (μέσω κατάλληλου γραφήματος) αν υπάρχει σχέση μεταξύ των προσαρμοσμένων τιμών των Y_i και των τυποποιημένων καταλοίπων σε κάθε ένα από αυτά.

(B3) (α) Θεωρούμε ότι το παραπάνω μοντέλο M3 έχει διαφορετική σταθερά στις διάφορες στάθμες του χρώματος και συγκεκριμένα,

$$\text{για I colour: } \ln(\text{Price}) = b_0 + b_1 \text{ Carats} + b_2 \text{ Carats}^2 + \varepsilon$$

$$\text{για H colour: } \ln(\text{Price}) = (b_0 + c_1) + b_1 \text{ Carats} + b_2 \text{ Carats}^2 + \varepsilon$$

$$\text{για G colour: } \ln(\text{Price}) = (b_0 + c_2) + b_1 \text{ Carats} + b_2 \text{ Carats}^2 + \varepsilon$$

$$\text{για F colour: } \ln(\text{Price}) = (b_0 + c_3) + b_1 \text{ Carats} + b_2 \text{ Carats}^2 + \varepsilon$$

$$\text{για E colour: } \ln(\text{Price}) = (b_0 + c_4) + b_1 \text{ Carats} + b_2 \text{ Carats}^2 + \varepsilon$$

$$\text{για D colour: } \ln(\text{Price}) = (b_0 + c_5) + b_1 \text{ Carats} + b_2 \text{ Carats}^2 + \varepsilon$$

Μπορούμε να ενώσουμε τα 6 παραπάνω σε ένα ισοδύναμο μοντέλο:

$$(M4) \quad \ln(\text{Price}) = b_0 + c_1 H + c_2 G + c_3 F + c_4 E + c_5 D + b_1 \text{ Carats} + b_2 \text{ Carats}^2 + \varepsilon$$

όπου η μεταβλητή $D = 1$ ή 0 ανάλογα με το αν Colour = D ή όχι, και όμοια κατασκευάζονται και οι μεταβλητές E, F, G, H (πρόκειται για δείκτριες ή βωβές μεταβλητές - dummy variables). Εφαρμόστε το μοντέλο (M4). Ποιες από τις ερμηνευτικές μεταβλητές φαίνεται να επηρεάζουν την τιμή; Είναι το μοντέλο αυτό καλύτερο από το (M3) και γιατί;

(B4) (α) Αν ανάλογα με τις στάθμες του χρώματος στο (B3), θεωρήσουμε το ίδιο για τις στάθμες της καθαρότητας και του πιστοποιητικού λαμβάνουμε το πλήρες μοντέλο (χωρίς αλληλεπιδράσεις):

$$(M5) \quad \ln(\text{Price}) = b_0 + c_1 H + c_2 G + c_3 F + c_4 E + c_5 D + d_1 \text{ VS1} + d_2 \text{ VVS2} + d_3 \text{ VVS1} + d_4 \text{ IF} \\ + e_1 \text{ GIA} + e_2 \text{ IGI} + b_1 \text{ Carats} + b_2 \text{ Carats}^2 + \varepsilon$$

όπου οι μεταβλητές VS1, VVS2, VVS1, IF, GIA, IGI είναι οι δείκτριες (ή βωβές) μεταβλητές των διαμαντιών που είναι VS1, ..., IGI αντίστοιχα. Εφαρμόστε το μοντέλο (M5) και εκτιμήστε τα $b_0, c_1, c_2, c_3, c_4, c_5, d_1, d_2, d_3, d_4, e_1, e_2, b_1, b_2, \sigma^2$. Ποιες από τις ερμηνευτικές μεταβλητές φαίνεται να επηρεάζουν την τιμή; Τι ποσοστό της μεταβλητότητας των παρατηρήσεων της τιμής ερμηνεύεται από το μοντέλο; Είναι το μοντέλο αυτό καλύτερο από το (M4) και γιατί;

(β) Στο (M5): Υπάρχουν έκτροπες ή ασυνήθιστες παρατηρήσεις και αν ναι, ποιες είναι; Ελέγξτε την ορθότητα του μοντέλου (κανονικότητα υπολοίπων, σχέση των προσαρμοσμένων Y_i και των τυποποιημένων καταλοίπων). Ποια είναι η παρατήρηση με την μεγαλύτερη επιρροή; είναι αυτή η επιρροή σημαντική ($>3(\text{πλήθος ερμηνευτικών μεταβλ.})/n$);

(B5) Με βάση το (M6), που ορίζεται ως το (M5) χωρίς τις μεταβλητές GIA, IGI που αφορούν το πιστοποιητικό:

(α) Αν ένα τυχαία επιλεγμένο μπριγιάν διαμάντι 1 καρατίου είναι εντελώς άχρωμο και αψεγάδιαστο, εκτιμήστε την τιμή του (σε \$ Σιγκαπούρης). Μεταξύ ποιών ορίων θα βρίσκεται η τιμή του με πιθανότητα 95%; Μεταξύ ποιών ορίων θα βρίσκεται η μέση τιμή των διαμαντιών με τα παραπάνω χαρακτηριστικά με πιθανότητα 95%; Αν αγοράσαμε ένα τέτοιο διαμάντι προς 17500\$ Σιγκαπούρης, έχουμε, κατά πάσα πιθανότητα, κάνει συμφέρουσα αγορά; (β) Κατασκευάστε έναν πίνακα 6×5 που θα περιέχει τις 30 εκτιμήσεις των τιμών μπριγιάν διαμαντιών, $0.3 + x/10$ καρατίων (x το τελευταίο ψηφίο του AM σας), ως προς το χρώμα (6 γραμμές, για D, E, F, G, H, I) και ως προς την καθαρότητα (5 στήλες, για IF, VVS1, VVS2, VS1, VS2). Το χρώμα ή η καθαρότητα φαίνεται να επηρεάζει περισσότερο την τιμή;

(B6) Τέλος, θεωρήστε το εναλλακτικό μοντέλο

$$\ln(\text{Price}) = b_0 + c \text{ Colour} + d \text{ Clarity} + b_1 \text{ Carats} + b_2 \text{ Carats}^2 + \varepsilon$$

όπου η μεταβλητή Color παίρνει τις τιμές 0, 1, 2, 3, 4, 5 ανάλογα με το αν το χρώμα είναι I, H, G, F, E, D και η μεταβλητή Clarity παίρνει τις τιμές 0, 1, 2, 3, 4 ανάλογα με το αν η καθαρότητα είναι VS2, VS1, VVS2, VVS1, IF. Επιχειρηματολογήστε αν είναι καλύτερο από το (M6) ή όχι.

ΜΕΡΟΣ Γ. Με βάση και πάλι τα δεδομένα από το ΜΕΡΟΣ Α (εκτός των 4 cases που αντιστοιχούν στον AM σας):

(Γ0) Να δημιουργήσετε μία νέα μεταβλητή q18q19 η οποία θα είναι ίση με το πλήθος των μαθημάτων που έχει περάσει κάποιος φοιτητής-τρια (q18) επί το μέσο βαθμό (q19) που δηλώνει. Η νέα αυτή μεταβλητή μπορεί να θεωρηθεί ότι εκφράζει το *σύνολο των βαθμών που έχει λάβει κάποιος φοιτητής-τρια σε όλα τα μαθήματα που έχει περάσει*. Επιθυμούμε να διερευνήσουμε αν υπάρχουν διαφορές στο μέσο σύνολο βαθμών ανάμεσα σε αυτούς που έχουν απαντήσει ότι διαβάζουν (*λίγο πριν και κατά την διάρκεια της εξεταστικής*): (1) Εξαντλητικά, (2) Πάρα πολύ, (3) Πολύ, (4) Μέτρια, (5) Λίγο, (6) Καθόλου (ερώτ. 23). Για το σκοπό αυτό θεωρούμε το μοντέλο $Y_{ij} = \mu_i + \varepsilon_{ij}$, $i = 1, 2, \dots, 6$, όπου μ_i είναι το μέσο σύνολο βαθμών των ατόμων που έχουν δώσει την απάντηση i στην ερώτηση 23 (με $\varepsilon_{ij} \sim N(0, \sigma^2)$).

(Γ1) (α) Κατασκευάστε ένα κοινό ιστόγραμμα (χρησιμ. "Panel by"), ένα κοινό Boxplot και ένα κοινό Error Bar (δ.ε. για το μέσο) του συνόλου των βαθμών ανά επίπεδο διαβάσματος (q23). (β) Εκτιμήστε τα μ_i σημειακά και με δ.ε. 95% και εξετάστε την υπόθεση $H_0: \mu_1 = \mu_2 = \dots = \mu_6$ με H_1 : *διαφορετικά*. (γ) Να δοθούν ταυτόχρονα δ.ε. για τις διαφορές των μέσων συνολικών βαθμών ανά δύο με την μέθοδο Tukey. Να γίνει ομαδοποίηση των μέσων $\mu_1, \mu_2, \dots, \mu_6$ με συνολική πιθανότητα σφάλματος 5%. Σχολιάστε το αποτέλεσμα. (δ) Εκτιμήστε το σ^2 . Να ελέγξετε αν οι διασπορές των συνολικών βαθμών είναι ίδιες ανάμεσα στις 6 κλάσεις των διαφορετικών απαντήσεων της ερώτησης 23.

(Γ2) Να γίνει ομαδοποίηση των μέσων συνολικών βαθμών (με συνολική πιθανότητα σφάλματος 5%) αυτή τη φορά σε σχέση με τις απαντήσεις τις ερώτησης 24, της ερώτησης 30 και της ερώτησης 5. Σχολιάστε τα αποτελέσματα.