

Στατιστικά Προγράμματα, 19-12-08, Homework #2

ΜΕΡΟΣ Α. Θα βασιστούμε στα δεδομένα της εργασίας 1 (αρχείο **Share1.sav**) τα οποία προέρχονται από το πρόγραμμα SHARE (*Survey of Health, Ageing and Retirement in Europe*, www.share-project.org).

Επιλέξτε τα δεδομένα που αφορούν όλες τις χώρες εκτός αυτής με κωδικό 11+x όπου x το τελευταίο ψηφίο του AM σας (αν π.χ. 11 + x = 11 τότε διαγράφετε τα δεδομένα από Αυστρία κ.ο.κ.). **Επίσης επιλέξτε μόνο τα άτομα ηλικίας 50 έως και 60 ετών και απαντήστε ανά φύλο, δηλαδή ξεχωριστά για γυναίκες - άντρες (π.χ. χρησιμ. layer).** **Να σχολιάσετε σύντομα όποια αποτελέσματα νομίζετε ότι είναι άξια σχολιασμού.**

(A1) Να εξετάσετε αν το κάπνισμα (csmoke) σχετίζεται με την κατανάλωση αλκοόλ (> 2 ποτήρια/ημέρα, drinkin2). Συγκεκριμένα, **(α)** Να κατασκευάσετε τον αντίστοιχο πίνακα συνάφειας με τις παρατηρούμενες και τις αναμενόμενες (υπό την υπόθεση της ανεξαρτησίας) συχνότητες, καθώς και τα ποσοστά των απαντήσεων ανά κάπνισμα και ανά κατανάλωση αλκοόλ (ανά φύλο). **(β)** Τι ποσοστό των καπνιστών πίνουν περισσότερα από 2 ποτήρια αλκοόλ/ημέρα και ποιο το αντίστοιχο ποσοστό σε αυτούς που δεν κάπνισαν ποτέ και σε αυτούς που έκοψαν το κάπνισμα; Η διαφορά που παρατηρείται στα ποσοστά είναι στατιστικά σημαντική (δηλ. το κάπνισμα σχετίζεται με την κατανάλωση αλκοόλ); (ανά φύλο)

(A2) Οι μεταβλητές που αφορούν την συναισθηματική υγεία του ατόμου είναι :

euro1 (depression)	δηλώνει λυπημένο ή καταθλιπτικό (0:Όχι, 1:Ναι)
euro2 (pessimism)	δεν αναφέρει κάποια ελπίδα για το μέλλον (0:Όχι, 1:Ναι)
euro3 (suicidality)	δηλώνει τάσεις αυτοκτονικού ιδεασμού ή επιθυμίας θανάτου (0:Όχι, 1:Ναι)
euro4 (guilt)	αισθάνεται ένοχο για κάτι (0:Όχι, 1:Ναι)
euro5 (sleep)	πρόσφατα έχει προβλήματα ύπνου (0:Όχι, 1:Ναι)
euro6 (interest)	τελευταία δεν έχει μεγάλο ενδιαφέρον για ο,τι συμβαίνει γύρω του (0:Όχι, 1:Ναι)
euro7 (irritability)	υπήρξε περισσότερο ευέξαπτο πρόσφατα (0:Όχι, 1:Ναι)
euro8 (appetite)	μειώθηκε η επιθυμία του για φαγητό (0:Όχι, 1:Ναι)
euro9 (fatigue)	έχει τελευταία ελαττωμένη διάθεση για να κάνει τα πράγματα που θέλει (0:Όχι, 1:Ναι)
euro10(concentration)	αναφέρει δυσκολία συγκέντρωσης (0:Όχι, 1:Ναι)
euro11 (enjoyment)	δεν έχει κάποια ευχάριστη δραστηριότητα (0:Όχι, 1:Ναι)
euro12 (tearfulness)	αναφέρει ότι τον τελευταίο μήνα έχει κλάψει (0:Όχι, 1:Ναι)

Να κατασκευάσετε μια νέα μεταβλητή MentalHealth που να δείχνει αν το άτομο παρουσιάζει καλή συναισθηματική υγεία ή όχι. Συγκεκριμένα, η MentalHealth να παίρνει την τιμή 1 αν $(euro1 + euro2 + \dots + euro12) \leq 3$, και 0 διαφορετικά.

Να διερευνήσετε τη σχέση της συναισθηματικής υγείας με τη συμβίωση με σύζυγο ή σύντροφο. Συγκεκριμένα: **(α)** Να ελέγξετε (ε.σ. α) αν η συναισθηματική υγεία σχετίζεται με την συμβίωση με σύζυγο ή σύντροφο. Τι ποσοστό των ατόμων που (i) ζουν με σύζυγο (spouse) παρουσιάζουν καλή συναισθηματική υγεία, και ποιο είναι το αντίστοιχο ποσοστό στα άτομα που (ii) ζουν με σύντροφο (partner) και στα άτομα που (iii) ζουν μόνα (single); (ανά φύλο)

(β) Να βρείτε με ποιες από τις μεταβλητές euro1 – euro12 σχετίζεται η συμβίωση με σύζυγο ή σύντροφο (ε.σ. α). Για το σκοπό αυτό, να κατασκευάσετε έναν πίνακα (για κάθε φύλο) της ακόλουθης μορφής:

	ζουν με σύζυγο	ζουν με σύντροφο	ζουν μόνοι	p-value
Euro1:Depression	...%	...%	...%	...
Euro2: Pessimism	...%	...%	...%	...
...	...	...	...	...

Οι γραμμές του πίνακα αντιστοιχούν στις μεταβλητές euro1 – euro12 και παρουσιάζουν τα ποσοστά αυτών που έχουν απαντήσει «Ναι» ανάμεσα σε αυτούς που (i) ζουν με σύζυγο, (ii)

ζουν με σύντροφο, και (iii) ζουν μόνοι, καθώς και το αντίστοιχο p-value. Ο πίνακας αυτός να περιλαμβάνει στις γραμμές **μόνο** όσες από τις μεταβλητές euro1 – euro12 φαίνεται να σχετίζονται με τη συμβίωση με σύζυγο ή σύντροφο σε ε.σ. α (υποδ: (i) βάλτε όλες μαζί τις euro1 – euro12 στο column του crosstabs, (ii) ο πίνακας αυτός δεν κατασκευάζεται αυτόματα από το SPSS, θα πρέπει να χρησιμοποιήσετε τιμές από κατάλληλους πίνακες).

(A3) Να διερευνήσετε (ανά φύλο) τη σχέση της συναισθηματικής υγείας με την κατανάλωση αλκοόλ. Συγκεκριμένα, να κάνετε και πάλι τα αντίστοιχα των (α), (β) του (A2) αυτή τη φορά ως προς την κατανάλωση αλκοόλ (π.χ. εδώ αντί των απαντήσεων (i) ζουν με σύζυγο, (ii) ζουν με σύντροφο, και (iii) ζουν μόνοι, θα έχουμε τις (i) πίνουν ≤ 2 ποτήρια αλκοόλ την ημέρα, (ii) πίνουν > 2 ποτήρια αλκοόλ την ημέρα)

(A4) Να διερευνήσετε (ανά φύλο) τη σχέση της συναισθηματικής υγείας με το κάπνισμα. Συγκεκριμένα, να κάνετε και πάλι τα αντίστοιχα των (α), (β) του (A2) ως προς το κάπνισμα.

(A5) Να διερευνήσετε (ανά φύλο) τη σχέση της συναισθηματικής υγείας με την κατάσταση της απασχόλησης (current job situation). Συγκεκριμένα, να κάνετε το αντίστοιχο του (α) του (A2) αυτή τη φορά ως προς την κατάσταση της απασχόλησης.

ΜΕΡΟΣ Β. Το αρχείο **cardata.txt** (βλ. ιστοσελίδα μαθήματος) περιλαμβάνει διάφορα χαρακτηριστικά 428 τύπων αυτοκινήτων που ήταν διαθέσιμα στην αγορά των ΗΠΑ το έτος 2003 (Kiplinger's Personal Finance, Dec2003, vol.57, no.12, pp.104-123). Περιλαμβάνονται οι μεταβλητές (στοιχισμένες σε στήλες):

1	Vehicle	Vehicle Name
2	Sports	Sports Car? (1=yes, 0=no)
3	Sport_utility	Sport Utility Vehicle? (1=yes, 0=no)
4	Wagon	Wagon? (1=yes, 0=no)
5	Minivan	Minivan? (1=yes, 0=no)
6	Pickup	Pickup? (1=yes, 0=no)
7	AllWheel	All-Wheel Drive? (1=yes, 0=no)
8	Rear_Wheel	Rear-Wheel Drive? (1=yes, 0=no)
9	Retail_price	Suggested Retail Price (U.S. Dollars)
10	Dealer_Cost	invoice price, what the dealership pays the manufacturer (U.S. Dollars)
11	Engine_size	Engine Size (liters)
12	Cylinders	Number of Cylinders
13	Horsepower	Horsepower
14	CmilesperGallon	City Miles Per Gallon
15	HmilesperGallon	Highway Miles Per Gallon
16	Weight	Weight (Pounds)
17	Wheel	Wheel Base (inches)
18	Length	Length (inches)
19	Width	Width (inches)

Εισάγετε κατάλληλα τα δεδομένα στο SPSS (προσοχή στον ορθό καθορισμό των κάθετων γραμμών – vertical lines – που χωρίζουν τις μεταβλητές στον text import wizard, step 4) και καθορίστε τα ονόματα (από τον παραπάνω πίνακα) και τους τύπους των μεταβλητών. Στη συνέχεια **αριθμήστε τα cases από 1-428** (χρησιμ. \$CASENUM) και **διαγράψτε τα 6 cases με αύξοντα αριθμό 10+x, 20+x, ... , 60+x, όπου x το τελευταίο ψηφίο του AM σας.**

(B1) Αρχικός σκοπός είναι να διερευνηθεί η εξάρτηση της κατανάλωσης (CmilesperGallon) από κατάλληλες ερμηνευτικές μεταβλητές. Ξεκινάμε εξετάζοντας τη σχέση της κατανάλωσης με το βάρος του οχήματος. Εφαρμόστε το γραμμικό μοντέλο

$$(M1) \quad \text{CmilesperGallon} = b_0 + b_1 * \text{Weight} + \varepsilon \quad (\varepsilon \sim N(0, \sigma^2))$$

(α) Κατασκευάστε το Scatterplot των δεδομένων με την (εκτιμημένη) ευθεία γραμμικής παλινδρόμησης. **(β)** Επιδρά το βάρος στην κατανάλωση; Τι ποσοστό της μεταβλητότητας της κατανάλωσης ερμηνεύεται από το μοντέλο; Κατασκευάστε το γράφημα των καταλοίπων σε σχέση με τις προσαρμοσμένες τιμές της κατανάλωσης και ελέγξτε αν τα σφάλματα είναι κανονικά (*Histogram, Q-Q plot, K-S test*). Τι παρατηρείτε;

(B2) (α) Κάνετε και πάλι το (α) του (B1) για το μοντέλο

$$(M2) \quad \text{CmilesperGallon} = b_0 + b_1 * (\text{Weight})^{-1} + \varepsilon$$

Υπάρχουν έκτροπες παρατηρήσεις, και ποιος ειδικός λόγος συντρέχει για αυτές; **Στη συνέχεια και για όσα ακολουθούν στο ΜΕΡΟΣ Β διαγράψτε αυτές τις έκτροπες παρατηρήσεις (με τυποποιημένο κατάλοιπο εκτός του (-3,3)).** **(β)** Κάνετε και πάλι το (β) του (B1) για το μοντέλο (M2). Είναι αυτό το μοντέλο καλύτερο από το (M1) και γιατί;

(B3) Ας θεωρήσουμε το μοντέλο

$$(M3) \quad \text{CmilesperGallon} = c_0 + b_1 * (\text{Weight})^{-1} + b_2 * (\text{Wheel})^{-1} + b_3 * (\text{Horsepower})^{-1} + b_4 * (\text{Engine_size})^{-1} + \varepsilon$$

για τα οχήματα με πίσω κίνηση ($\text{Rear_Wheel} = 1$), και το

$$(M4) \quad \text{CmilesperGallon} = b_0 + b_1 * (\text{Weight})^{-1} + b_2 * (\text{Wheel})^{-1} + b_3 * (\text{Horsepower})^{-1} + b_4 * (\text{Engine_size})^{-1} + \varepsilon$$

για τα υπόλοιπα οχήματα (τα (M3), (M4) απλώς έχουν διαφορετικές σταθερές c_0, b_0). Τα μοντέλα (M3), (M4) μπορούν να συνδυαστούν σε ένα μοντέλο με τη χρήση της δείκτριας μεταβλητής Rear_Wheel (μεταβλητές τέτοιου τύπου, με τιμές 0-1, ονομάζονται βωβές - *dummy*):

$$(M5) \quad \text{CmilesperGallon} = b_0 + b_1 * (\text{Weight})^{-1} + b_2 * (\text{Wheel})^{-1} + b_3 * (\text{Horsepower})^{-1} + b_4 * (\text{Engine_size})^{-1} + b_5 * \text{Rear_Wheel} + \varepsilon$$

(με $b_5 = c_0 - b_0$) **(α)** Εφαρμόστε το (M5) και εκτιμήστε τα b_i και το σ . Επηρεάζουν όλες οι παραπάνω ερμηνευτικές μεταβλητές την κατανάλωση (ε.σ. α) και με ποιο τρόπο; (π.χ. αύξηση του βάρους, αυξάνει ή μειώνει την CmilesperGallon , κ.ο.κ.). Είναι αυτό το μοντέλο καλύτερο από το (M2) και γιατί; **(β)** Ελέγξτε την ορθότητα του μοντέλου (M5). Συγκεκριμένα, ελέγξτε αν τα σφάλματα είναι κανονικά (*ιστόγραμμα, Q-Q plot και K-S test*) και εξετάστε μέσω κατάλληλου γραφήματος αν υπάρχει σχέση μεταξύ των καταλοίπων και των προσαρμοσμένων τιμών της κατανάλωσης.

(B4) Με βάση το (M5): Αν ένα τυχαία επιλεγμένο αυτοκίνητο έχει βάρος 4000 Pounds, μήκος μεταξύ μπρος και πίσω τροχών (Wheel Base) 120 ίντσες, Ιπποδύναμη 300 ίππους, μηχανή 2000cc (2 λίτρα) και κίνηση εμπρός, εκτιμήστε την κατανάλωσή του (*Μίλια μέσα στην πόλη ανά γαλόνι*). Μεταξύ ποιών ορίων θα βρίσκεται αυτή με πιθανότητα 95%;

(B5) Επιθυμούμε τώρα να διερευνήσουμε την εξάρτηση της *λιανικής τιμής πώλησης* (Retail_price) από τις υπόλοιπες ερμηνευτικές μεταβλητές. Δημιουργήστε μια νέα (*dummy*) μεταβλητή *europaean* με τιμή 1 στα Ευρωπαϊκά αυτοκίνητα (Audi, BMW, Jaguar, Mercedes-Benz, Porsche, Saab, Volkswagen, Volvo) και τιμή 0 στα υπόλοιπα.

Εφαρμόστε το μοντέλο γραμμικής παλινδρόμησης με μεταβλητή απόκρισης την Retail_price και ερμηνευτικές μεταβλητές τις (οι 7 πρώτες είναι *dummy*):

European, Sports, Sport_utility, Wagon, Minivan, AllWheel, Rear_Wheel, Engine_size, Cylinders, Horsepower, CmilesperGallon, HmilesperGallon, Weight, Wheel, Length, Width

Είναι το μοντέλο αυτό αποδεκτό; (κανονικότητα σφαλμάτων)

(B6) Εφαρμόστε τώρα το γραμμικό μοντέλο με μεταβλητή απόκρισης την $\ln(\text{Retail_price})$ και ερμηνευτικές μεταβλητές τις ίδιες με το (B5) παραπάνω. Είναι τώρα το μοντέλο αυτό αποδεκτό; (κανονικότητα σφαλμάτων). Ποιες από τις ερμηνευτικές μεταβλητές φαίνεται να επηρεάζουν την τιμή πώλησης; (ε.σ. α). Στα «Ευρωπαϊκά» αυτοκίνητα διαμορφώνεται διαφορετική λιανική τιμή από τα υπόλοιπα, θεωρώντας ίδια τα υπόλοιπα χαρακτηριστικά.

(B7) (α) Θεωρείστε και πάλι ως μεταβλητή απόκρισης την $\ln(\text{Retail_price})$ και κρατήστε μόνο ως ερμηνευτικές τις

European, Sports, Sport_utility, AllWheel, Rear_Wheel,
Horsepower, CmilesperGallon, Weight, Length, Width

Είναι το μοντέλο αυτό καλύτερο από τα προηγούμενα και γιατί;

(β) Στο μοντέλο του (α): Για κάθε τύπο αυτοκινήτου, το κατάλοιπο μπορεί να θεωρηθεί ότι αποτελεί έναν δείκτη απόκλισης της λιανικής τιμής του από την προβλεπόμενη από τα παραπάνω χαρακτηριστικά του. Με βάση αυτή την θεώρηση, ποια είναι τα 5 «ακριβότερα» και τα 5 «φθηνότερα» αυτοκίνητα με βάση τα χαρακτηριστικά τους;

ΜΕΡΟΣ Γ. Χρησιμοποιούμε και πάλι τα δεδομένα του προγράμματος SHARE: **Κρατήστε τα δεδομένα που αφορούν μόνο τις 4 χώρες που αντιστοιχούν στον ΑΜ σας (όπως και στην εργασία 1).**

Επιθυμούμε να διερευνήσουμε αν, και με ποιο τρόπο, εξαρτάται το *income* - εισόδημα ενός ατόμου ($\text{income} = \text{Annual gross income from employment}$) από τα έτη σπουδών του (*yeduc*). Χρησιμοποιήστε το αρχείο **Share2.sav** (βλ. ιστοσελίδα μαθήματος) το οποίο είναι το ίδιο με το *Share1.sav* αλλά περιέχει επιπλέον την μεταβλητή *yeduc*.

(Γ1) Σε κάθε μία από τις 4 χώρες ξεχωριστά: Βρείτε τη διάμεσο *D* του *income* στα άτομα που εργάζονται (μεταβλητή *current job situation*) και έχουν μη μηδενικό *income* (με κατάλληλο *select cases*). Στη συνέχεια επιλέξτε μόνο τα άτομα που εργάζονται, έχουν *income* μεγαλύτερο από το 50% του *D* (για να εξαιρέσουμε τους περισσότερους που είναι μερικής απασχόλησης) και έχουν *income* μικρότερο από το 4πλάσιο του *D* (για να εξαιρέσουμε τις ακραίες περιπτώσεις). Δημιουργήστε την (dummy) μεταβλητή *sex* που παίρνει την τιμή 1 ή 0 ανάλογα με το αν το άτομο είναι άνδρας ή όχι. Στα δεδομένα που έχετε επιλέξει:

Εφαρμόστε τα μοντέλα (στις 4 χώρες ξεχωριστά, με *split file*)

$$(M1) \quad \text{Income} = b_0 + b_1 * \text{sex} + b_2 * \text{yeduc} + b_3 * \text{age} + \varepsilon$$

και το

$$(M2) \quad \ln(\text{Income}) = b_0 + b_1 * \text{sex} + b_2 * \text{yeduc} + b_3 * \text{age} + \varepsilon$$

και εξετάστε ποιο παρουσιάζει καλύτερη προσαρμογή στα δεδομένα (ελέγξτε την κανονικότητα των σφαλμάτων).

(Γ2) Στο καλύτερο μοντέλο του (Γ1) και στις 4 χώρες ξεχωριστά: **(α)** Τι ποσοστό της μεταβλητότητας της μεταβλητής απόκρισης ερμηνεύεται από το μοντέλο; Ποιες από τις ερμηνευτικές μεταβλητές φαίνεται να επηρεάζουν το εισόδημα; **(β)** Κάθε επιπλέον έτος σπουδών, πόση αύξηση του εισοδήματος συνεπάγεται κατά μέσο όρο; Αν ένας άνδρας ηλικίας 55 ετών έχει 16 έτη σπουδών, ποιο θα είναι κατά μέσο όρο το εισόδημά του; Δώστε και το αντίστοιχο δ.ε. 95%.

- όλοι οι έλεγχοι να γίνουν σε ε.σ. α = 5%