ANOVA/Regression I

Review

Τα αρχεία δεδομένων του STATA έχουν κατάληξη “.dta” . Τα αρχεία αυτά ανοίγουν είτε μέσα από το STATA επιλέγοντας file/open κλπ ή με την εντολή use, είτε με διπλό κλικ στο ίδιο το αρχείο. Μπορούμε όμως να εισάγουμε στο STATA και αρχεία άλλων τύπων. Τα πιο συνηθισμένα είναι αρχεία EXCEL και ASCII. Τα αρχεία EXCEL μπορούμε να τα περάσουμε στο STATA κάνοντας copy/paste στον editor του STATA. Άλλη επιλογή είναι να τα σώσουμε ως αρχείο ASCII (comma delimited - .csv ή tab delimited - .txt) και να τα εισάγουμε στο STATA με την εντολή insheet.

Πριν διαβάσουμε το αρχείο με τα δεδομένα μας ας ορίσουμε ένα φάκελο στον οποίο θα έχουμε τα αρχεία που θα χρειαστούμε ή θα δημιουργήσουμε.

**cd “c:\...\stata\_review”**

Στο φάκελο αυτό υπάρχει το αρχείο με τα δεδομένα με τα οποία θα δουλέψουμε σήμερα. Είναι σε μορφή ASCII και είναι το αρχείο data.txt .

**insheet using data.txt, clear**

Το αρχείο περιλαμβάνει τις μεταβλητές

id byte %8.0g

age byte %8.0g

sex byte %8.0g

calories float %9.0g

chol float %9.0g

Η μεταβλητή sex παίρνει τις τιμές 1-άνδρας και 2-γυναίκα. Η μεταβλητή calories έχει τις τιμές της ημερήσιας πρόσληψης θερμίδων. Θέλουμε να εξετάσουμε αν μέση ημερήσια πρόσληψη θερμίδων διαφοροποιείται ανά φύλο.

Βασικά μέτρα θέσης και διασποράς για την μεταβλητή calories παίρνουμε με την εντολή **sum**

**by sex : sum calories**

----------------------------------------------------------------------------

-> sex = 1

 Variable | Obs Mean Std. Dev. Min Max

-------------+--------------------------------------------------------

 calories | 41 2045.873 583.9667 827.9 3457.2

----------------------------------------------------------------------------

-> sex = 2

 Variable | Obs Mean Std. Dev. Min Max

-------------+--------------------------------------------------------

 calories | 39 1576.497 524.1219 740.6 3144.8

Τι παρατηρείτε;

Κατόπιν θα ελέγξουμε αν οι τιμές της μεταβλητής calories κατανέμονται κανονικά ανά φύλο. Το ιστόγραμμα μας δίνει μια εικόνα της κατανομής

**hist calories, by(sex) bin(7) normal**

Δεν παρατηρούμε μεγάλη ασυμμετρία ή πολυκόρυφες κατανομές και με δεδομένο το ικανοποιητικό μέγεθος δείγματος μπορούμε να πούμε ότι δεν έχουμε σημαντική απόκλιση από την κανονικότητα. Μπορούμε να κάνουμε και το test Shapiro-Wilk για ένα πιο τυπικό έλεγχο. Η μηδενική υπόθεση του ελέγχου είναι

Η0: “το δείγμα προέρχεται από πληθυσμό που κατανέμεται κανονικά”

**by sex: swilk calories**

----------------------------------------------------------------------------

-> sex = 1

 Shapiro-Wilk W test for normal data

 Variable | Obs W V z Prob>z

-------------+-------------------------------------------------

 calories | 41 0.98270 0.697 -0.761 0.77680

----------------------------------------------------------------------------

-> sex = 2

 Shapiro-Wilk W test for normal data

 Variable | Obs W V z Prob>z

-------------+-------------------------------------------------

 calories | 39 0.96407 1.393 0.697 0.24305

Η υπόθεση της κανονικότητας δεν απορρίπτεται.

Μια εικόνα της κατανομής μπορούμε να πάρουμε και με ένα θηκόγραμμα

**graph box calories, over(sex)**

όπως και με το ιστόγραμμα έτσι και τώρα δεν παρατηρούμε σημαντικές διαφορές στη διακύμανση των κατανομών. Μπορούμε να κάνουμε και ένα F test για τον έλεγχο των διασπορών ( F= s21/s22 , με n1-1, n2-1 βαθμούς ελευθερίας )

**sdtest calories, by(sex)**

Variance ratio test

------------------------------------------------------------------------------

 Group | Obs Mean Std. Err. Std. Dev. [95% Conf. Interval]

---------+--------------------------------------------------------------------

 1 | 41 2045.873 91.20028 583.9667 1861.551 2230.196

 2 | 39 1576.497 83.92668 524.1219 1406.597 1746.398

---------+--------------------------------------------------------------------

combined | 80 1817.052 67.13267 600.4529 1683.428 1950.677

------------------------------------------------------------------------------

Degrees of freedom: 40, 38

 Ho: sd(1) = sd(2)

 Ha: sd(1) < sd(2) Ha: sd(1) != sd(2) Ha: sd(1) > sd(2)

 F = 1.2414 F = 1.2414 F = 1.2414

 P < F = 0.7475 **2\*(P > F) = 0.5050** P > F = 0.2525

Τι παρατηρούμε;

Με βάση τα παραπάνω στοιχεία, για να ελέγξουμε την πρόσληψη θερμίδων ανά φύλο μπορούμε να κάνουμε t-test υποθέτοντας ίσες διασπορές.

**ttest calories, by(sex)**

Two-sample t test with equal variances

------------------------------------------------------------------------------

 Group | Obs Mean Std. Err. Std. Dev. [95% Conf. Interval]

---------+--------------------------------------------------------------------

 1 | 41 2045.873 91.20028 583.9667 1861.551 2230.196

 2 | 39 1576.497 83.92668 524.1219 1406.597 1746.398

---------+--------------------------------------------------------------------

combined | 80 1817.052 67.13267 600.4529 1683.428 1950.677

---------+--------------------------------------------------------------------

 diff | 469.3757 124.2787 221.9559 716.7955

------------------------------------------------------------------------------

Degrees of freedom: 78

 Ho: mean(1) - mean(2) = diff = 0

 Ha: diff < 0 Ha: diff != 0 Ha: diff > 0

 t = 3.7768 t = 3.7768 t = 3.7768

 P < t = 0.9998 P > |t| = 0.0003 P > t = 0.0002

Παρατηρούμε ότι η τιμή της Τ= …

Το p-value του ελέγχου **Ha: diff != 0** είναι .… 0.05

Τι συμπεραίνουμε;

Το p-value μπορούμε να το βρούμε αν γνωρίζουμε την τιμή της Τ, με τη συνάρτηση **ttail** που μας δίνει την πιθανότητα Ρ(Τ>t), πολλαπλασιασμένη επί 2

**display 2\*ttail(78, 3.7768)**

**.00030861**

Συγκρίνουμε την Τ με την tn-2, 0.025. Την τιμή της tn-2, 0.025 μπορούμε να τη βρούμε με την συνάρτηση **invttail** (αντίστροφη της **ttail**)

**display invttail(78, 0.025)**

**1.9908471**

Παρατηρούμε ότι Τ> tn-2, 0.025 και άρα απορρίπτουμε τη μηδενική υπόθεση της ισότητας των μέσων.

Για να εξετάσουμε πιθανή σχέση μεταξύ των ποσοτικών μεταβλητών που περιλαμβάνονται στο αρχείο δεδομένων (age-ηλικία, calories και chol-χοληστερόλη) μπορούμε αρχικά να δημιουργήσουμε τα στικτογράμματα των μεταβλητών:

**graph matrix calories chol age**

Τι παρατηρείτε;

Οι συντελεστές γραμμικής συσχέτισης του Pearson δίνονται με την εντολή:

**pwcorr chol cal age, sig**

 | chol calories age

-------------+---------------------------

 chol | 1.0000

 |

 |

 calories | 0.5804 1.0000

 | 0.0000

 |

 age | -0.0916 -0.1360 1.0000

 | 0.4192 0.2290

 |

Βλέπουμε ότι υπάρχει στατιστικά σημαντική θετική συσχέτιση μεταξύ των μεταβλητών calories και chol:

r=0.58, p-value<0.001

Για να βρούμε τους συντελεστές ελαχίστων τετραγώνων της εξίσωσης της ευθείας

Y = a + b\*X για την σχέση chol – calories, εκτελούμε την εντολή regress:

**regress chol calories**

 Source | SS df MS Number of obs = 80

-------------+------------------------------ F( 1, 78) = 39.62

 Model | 479809.732 1 479809.732 Prob > F = 0.0000

 Residual | 944524.043 78 12109.2826 R-squared = 0.3369

-------------+------------------------------ Adj R-squared = 0.3284

 Total | 1424333.78 79 18029.5415 Root MSE = 110.04

------------------------------------------------------------------------------

 chol | **Coef**. Std. Err. t P>|t| [95% Conf. Interval]

-------------+----------------------------------------------------------------

 calories | .1297902 .020619 6.29 0.000 .0887411 .1708394

 \_cons | 31.76056 39.43408 0.81 0.423 -46.74667 110.2678

------------------------------------------------------------------------------

Οι εκτιμήσεις των συντελεστών είναι στην πρώτη στήλη:

a= 31.76 και b = 0.13

Μπορούμε να φτιάξουμε το σχετικό γράφημα της ευθείας ελαχίστων τετραγώνων με την εντολή:

**twoway (scatter chol calories, ytitle(chol)) (lfit chol calories)**

