

Έλεγχος του Chow

Γραμμική Παλινδρόμηση — Wooldridge: sleep75

Η παρακάτω εξίσωση χαρακτηρίζεται ως εξίσωση ύπνου:

$$\text{sleep} = \beta_0 + \beta_1 \cdot \text{totwrk} + \beta_2 \cdot \text{educ} + \beta_3 \cdot \text{age} + \beta_4 \cdot \text{age}^2 + \beta_5 \cdot \text{yngkid} + e$$

Όπου:

Μεταβλητή	Ερμηνεία
sleep	τα λεπτά νυχτερινού ύπνου ανά εβδομάδα
totwrk	τα λεπτά εργασίας ανά εβδομάδα
educ	τα χρόνια σπουδών
age	η ηλικία σε χρόνια
yngkid	δυναμική μεταβλητή που καταγράφει την ύπαρξη μικρού παιδιού στην οικογένεια
male	δυναμική μεταβλητή που καταγράφει το φύλο

Ερωτήματα του αποσπάσματος

1. Εκτιμήστε την παραπάνω εξίσωση ύπνου **ξεχωριστά** για τους άνδρες και τις γυναίκες του δείγματος. Φαίνεται να υπάρχουν σημαντικές διαφορές μεταξύ των συντελεστών των δύο εξισώσεων;
2. Εκτελέστε τον έλεγχο του **Chow** για να ελέγξετε την ισότητα των συντελεστών των δύο εξισώσεων ύπνου. Για να υπολογίσετε το στατιστικό, χρησιμοποιήστε το μοντέλο που περιλαμβάνει τη δυναμική μεταβλητή “male” και όλους τους όρους αλληλεπίδρασης της με τις υπόλοιπες μεταβλητές, χρησιμοποιώντας όλες τις παρατηρήσεις, ανδρών και γυναικών μαζί.
3. Επιτρέψτε την ύπαρξη **ξεχωριστών σταθερών** στις εξισώσεις ύπνου ανδρών και γυναικών και ελέγξτε κατά πόσο όλοι οι όροι αλληλεπίδρασης της “male” με τις υπόλοιπες μεταβλητές είναι από κοινού στατιστικά σημαντικοί.
4. Με βάση τα αποτελέσματα των ερωτημάτων 2 και 3, ποιο μοντέλο θα επιλέγατε;

1) Χωριστές OLS παλινδρομήσεις για άνδρες και γυναίκες

Στο πρώτο βήμα εκτιμούμε δύο παλινδρομήσεις: μία για τους άνδρες και μία για τις γυναίκες. Αυτό γίνεται χρησιμοποιώντας το subset της μεταβλητής male.

```
rm(list=ls())
library(wooldridge)
#help(sleep75)
names(sleep75)

ols1a = lm(sleep ~ totwrk + educ + age + agesq + yngkid,
           data = sleep75, subset = male == 1)
summary(ols1a)

ols1b = lm(sleep ~ totwrk + educ + age + agesq + yngkid,
           data = sleep75, subset = male == 0)
summary(ols1b)
```

Για τους άνδρες εκτιμάται η σχέση:

$$sleep = \beta_0^M + \beta_1^M \cdot totwrk + \beta_2^M \cdot educ + \beta_3^M \cdot age + \beta_4^M \cdot age^2 + \beta_5^M \cdot yngkid + e$$

Για τις γυναίκες εκτιμάται η σχέση:

$$sleep = \beta_0^f + \beta_1^f \cdot totwrk + \beta_2^f \cdot educ + \beta_3^f \cdot age + \beta_4^f \cdot age^2 + \beta_5^f \cdot yngkid + e$$

Μεταβλητή	Άνδρες: εκτίμηση	Γυναίκες: εκτίμηση	Σχόλιο
Intercept	3648.20826	4238.72933	Η σταθερά είναι υψηλότερη στις γυναίκες.
totwrk	-0.18212	-0.13995	Η εργασία μειώνει τον ύπνο και στις δύο ομάδες.
educ	-13.05238	-10.20514	Η επίδραση της εκπαίδευσης είναι αρνητική και παρόμοια.
age	7.15659	-30.35657	Η επίδραση της ηλικίας διαφοροποιείται ως προς τη μορφή της καμπύλης.
agesq	-0.04477	0.36794	Στους άνδρες η παραβολή στρέφει προς τα κάτω, στις γυναίκες προς τα πάνω.
yngkid	60.38021	-118.28256	Η ύπαρξη μικρού παιδιού συνδέεται με διαφορετική κατεύθυνση επίδρασης.

Με μια πρώτη ματιά βλέπουμε πως οι συντελεστές της μεταβλητής “yngkid” διαφέρουν αρκετά. Με τους άνδρες να κερδίζουν σχεδόν μία ώρα ύπνου την εβδομάδα, ενώ οι γυναίκες να χάνουν σχεδόν δύο ώρες ύπνου όταν υπάρχει μικρό παιδί στην οικογένεια.

Η ηλικία εμφανίζεται να έχει διαφορετική συμπεριφορά. Στους άνδρες το agesq είναι αρνητικό, άρα η παραβολή στρέφει προς τα κάτω. Στις γυναίκες το agesq είναι θετικό, άρα η παραβολή στρέφει προς τα πάνω.

Γράφημα της τετραγωνικής σχέσης της ηλικίας

Για να απεικονιστεί η μη γραμμική σχέση της ηλικίας με τον ύπνο, κατασκευάζεται η μεταβλητή years και σχεδιάζονται οι δύο τετραγωνικές συναρτήσεις.

```
years = min(sleep75$age) : max(sleep75$age)

plot(years, ols1a$coeff[4]*years + ols1a$coeff[5]*years^2,
     type="l", ylab="min of sleep", col="blue", ylim=c(-650,280))

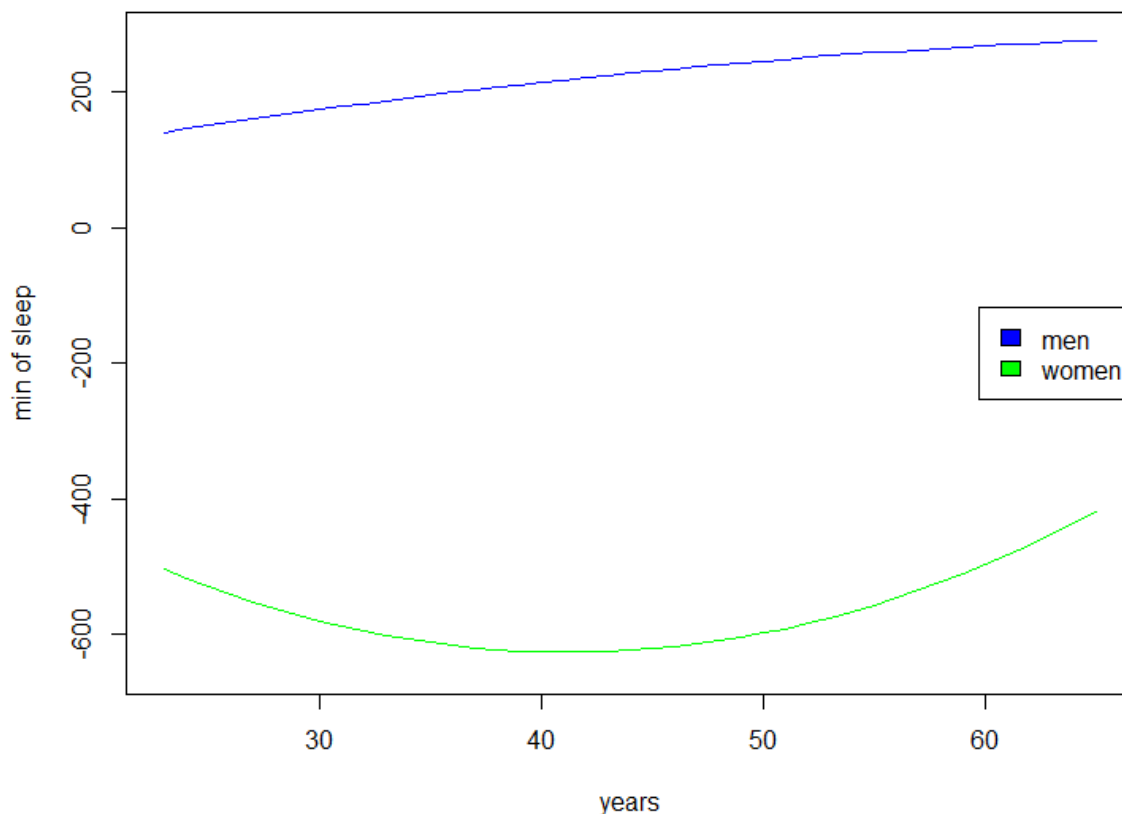
lines(years, ols1b$coeff[4]*years + ols1b$coeff[5]*years^2, col="green")

legend("right", c("men","women"), fill=c("blue","green"))

-ols1b$coeff[4] / (2*ols1b$coeff[5]) # global minimum for women quadratic
relation
```

$$\text{Άνδρες: } sleep = 7.157 \cdot age - 0.045 \cdot age^2$$

$$\text{Γυναίκες: } sleep = -30.357 \cdot age + 0.368 \cdot age^2$$



Άρα, με δεδομένο ότι όλες οι άλλες μεταβλητές παραμένουν σταθερές, για τους άνδρες μεγαλύτερη ηλικία συνεπάγεται περισσότερο ύπνο. Αντίθετα, για τις γυναίκες μέχρι την ηλικία των 41.25 ετών υπάρχει αρνητική σχέση επιπλέον έτους σπουδών/ηλικίας με τα λεπτά ύπνου, αλλά από αυτή την ηλικία και μετά κάθε επιπλέον έτος συνεπάγεται περισσότερο ύπνο. Οι συντελεστές των μεταβλητών “totwrk” και “educ” δεν φαίνεται να έχουν πολύ μεγάλες διαφορές μεταξύ ανδρών και γυναικών.

2) Πρώτος έλεγχος Chow: ισότητα όλων των συντελεστών

Ο έλεγχος του Chow χρησιμοποιείται για να ελέγξουμε κατά πόσο δύο παλινδρομήσεις σε διαφορετικά σετ δεδομένων έχουν στατιστικά ίσους συντελεστές. Εδώ εξετάζουμε αν οι εξισώσεις ύπνου ανδρών και γυναικών έχουν τους ίδιους συντελεστές.

$$H_0: \beta_i^M = \beta_i^f, \text{ για κάθε } i = 0, 1, 2, 3, 4, 5$$

$$H_1: \beta_i^M \neq \beta_i^f, \text{ για κάποιο } i = 0, 1, 2, 3, 4, 5$$

Ο έλεγχος γίνεται με ένα “ολοκληρωμένο” ή “μεγάλο” μοντέλο, που περιλαμβάνει τη μεταβλητή male και όλους τους όρους αλληλεπίδρασης της male με τις υπόλοιπες μεταβλητές:

$$\begin{aligned} \text{sleep} = & \beta_0 + \beta_1 \cdot \text{totwrk} + \beta_2 \cdot \text{educ} + \beta_3 \cdot \text{age} + \beta_4 \cdot \text{age}^2 + \beta_5 \cdot \text{yngkid} + \beta_6 \cdot \text{male} + \beta_7 \cdot \text{male} \cdot \text{totwrk} + \beta_8 \cdot \text{male} \cdot \text{educ} \\ & + \beta_9 \cdot \text{male} \cdot \text{age} + \beta_{10} \cdot \text{male} \cdot \text{age}^2 + \beta_{11} \cdot \text{male} \cdot \text{yngkid} + e \end{aligned}$$

Σε αυτό το υπόδειγμα, οι μεταβλητές που πολλαπλασιάζονται με τη “male” συγκρίνουν την πληροφορία που διαφοροποιεί τους άνδρες από τις γυναίκες. Για παράδειγμα, στους άνδρες η επίδραση της εκπαίδευσης είναι $\beta_2 + \beta_8$, ενώ στις γυναίκες είναι β_2 .

$$H_0: \beta_i = 0, \text{ για κάθε } i = 6, 7, 8, 9, 10, 11$$

$$H_1: \beta_i \neq 0, \text{ για κάποιο } i = 6, 7, 8, 9, 10, 11$$

```
ols2a = lm(sleep ~ totwrk + educ + age + agesq + yngkid + male +
           male:totwrk + male:educ + male:age + male:agesq + male:yngkid,
           data = sleep75)
summary(ols2a)

ols2b = lm(sleep ~ totwrk + educ + age + agesq + yngkid, data = sleep75)
summary(ols2b)

anova(ols2a, ols2b)
```

Μοντέλο	Εξίσωση	Ρόλος στον έλεγχο
ols2a	sleep ~ totwrk + educ + age + agesq + yngkid + male + όλες οι αλληλεπιδράσεις male	Πλήρες / unrestricted μοντέλο
ols2b	sleep ~ totwrk + educ + age + agesq + yngkid	Περιορισμένο / restricted μοντέλο

Έλεγχος	F	p-value	Συμπέρασμα
anova(ols2a, ols2b)	2.1164	0.04949	p-value = 0.04949 < 0.05, άρα απορρίπτουμε την H_0 . Υπάρχουν ενδείξεις ότι οι εξισώσεις ύπνου ανδρών και γυναικών διαφέρουν στατιστικά, έστω και οριακά.

Στους άνδρες η προηγούμενη επίδραση της educ στις γυναίκες προσαυξάνεται με τον συντελεστή της αλληλεπίδρασης male:educ. Δηλαδή: για τις γυναίκες $\text{educ} = \beta_2$, ενώ για τους άνδρες $\text{educ} = \beta_2 + \beta_8$.

```
# "educ" για τους άνδρες
ols1a$coeff[3]
ols2a$coeff[3] + ols2a$coeff[9]

# "educ" για τις γυναίκες
ols1b$coeff[3]
ols2a$coeff[3]
```

3) Δεύτερος έλεγχος Chow: ξεχωριστές σταθερές, κοινές κλίσεις

Στο τρίτο ερώτημα επιτρέπεται η ύπαρξη ξεχωριστών σταθερών στις εξισώσεις ύπνου ανδρών και γυναικών. Δηλαδή, κρατάμε τη μεταβλητή male στο μοντέλο, ώστε οι δύο ομάδες να μπορούν να έχουν διαφορετικό intercept, αλλά ελέγχουμε μόνο τους όρους αλληλεπίδρασης της male με τις υπόλοιπες μεταβλητές.

$$H_0: \beta_i = 0, \text{ για κάθε } i = 7, 8, 9, 10, 11$$

$$H_1: \beta_i \neq 0, \text{ για κάποιο } i = 7, 8, 9, 10, 11$$

Το δεύτερο, **περιορισμένο** μοντέλο του ελέγχου είναι:

$$sleep = \beta_0 + \beta_1 \cdot totwrk + \beta_2 \cdot educ + \beta_3 \cdot age + \beta_4 \cdot age^2 + \beta_5 \cdot yngkid + \beta_6 \cdot male + e$$

```
ols3b = lm(sleep ~ totwrk + educ + age + agesq + yngkid + male,
           data = sleep75)
summary(ols3b)

anova(ols2a, ols3b)
```

Μεταβλητή	Εκτίμηση στο ols3b	Στατιστική σημαντικότητα
Intercept	3840.85207	στατιστικά σημαντικό
totwrk	-0.16342	στατιστικά σημαντικό
educ	-11.71327	στατιστικά σημαντικό στο 5%
age	-8.69740	μη στατιστικά σημαντικό
agesq	0.12844	μη στατιστικά σημαντικό
yngkid	-0.02280	μη στατιστικά σημαντικό
male	87.75455	στατιστικά σημαντικό στο 5%

Έλεγχος	F	p-value	Συμπέρασμα
anova(ols2a, ols3b)	1.2558	0.2814	p-value = 0.2814 > 0.05, άρα δεν απορρίπτουμε την H_0 . Δεν υπάρχουν ενδείξεις ότι οι συντελεστές κλίσης διαφέρουν στατιστικά ανάμεσα σε άνδρες και γυναίκες, εφόσον επιτρέπουμε διαφορετική σταθερά.

4) Επιλογή μοντέλου

Το F-test έδειξε πως, από τη στιγμή που επιτρέπουμε στις σταθερές των δύο εξισώσεων ύπνου να διαφέρουν με την προσθήκη της μεταβλητής “male”, δεν υπάρχουν ενδείξεις πως οι συντελεστές των μεταβλητών, δηλαδή των κλίσεων, διαφέρουν μεταξύ τους.

Άρα, η πιο συνετή επιλογή είναι να κρατήσουμε το υπόδειγμα που περιλαμβάνει τη μεταβλητή “male”, χωρίς όλους τους όρους αλληλεπίδρασης. Το μοντέλο αυτό επιτρέπει διαφορετικό intercept για άνδρες και γυναίκες, αλλά διατηρεί κοινές κλίσεις.

$$sleep = \beta_0 + \beta_1 \cdot totwrk + \beta_2 \cdot educ + \beta_3 \cdot age + \beta_4 \cdot age^2 + \beta_5 \cdot yngkid + \beta_6 \cdot male + e$$

Συνοπτική ερμηνεία του ελέγχου Chow

Εκδοχή ελέγχου	Τι ελέγχει	p-value	Συμπέρασμα
1ος Chow test	Αν όλα τα coefficients, μαζί και η σταθερά, είναι ίδια για άνδρες και γυναίκες.	0.04949	Απορρίπτεται οριακά η H_0 . Υπάρχουν ενδείξεις συνολικής διαφοράς.
2ος Chow test	Αν οι κλίσεις είναι ίδιες, αφού επιτραπεί διαφορετική σταθερά με τη male.	0.2814	Δεν απορρίπτεται η H_0 . Οι κλίσεις μπορούν να θεωρηθούν κοινές.