Βασική ανάλυση δεδομένων και OLS στην R

1. Χρήση ενσωματωμένων δεδομένων στην R

Η R περιέχει έτοιμα datasets για εξάσκηση. Ένα από τα πιο γνωστά είναι το **mtcars**, το οποίο περιέχει **χαρακτηριστικά αυτοκινήτων**, όπως *κατανάλωση*, *ιπποδύναμη*, *βάρος* κ.ά.

```
mtcars
names(mtcars)
dim(mtcars)
```

Τι κάνει κάθε εντολή

- mtcars : εμφανίζει όλο το dataset.
- names(mtcars) : εμφανίζει τα ονόματα των μεταβλητών.
- dim(mtcars) : εμφανίζει τις διαστάσεις του πίνακα δεδομένων, δηλαδή:
    - αριθμό γραμμών = παρατηρήσεις
    - αριθμό στηλών = μεταβλητές

2. Επιλογή δεδομένων από πίνακα

Στην R μπορούμε να επιλέξουμε μεταβλητές ή συγκεκριμένες τιμές με δύο βασικούς τρόπους.

object$column_name

column_name[.,θέση]

Επιλογή ολόκληρης μεταβλητής hp

```
mtcars$hp
# or
mtcars[,4]
```

- mtcars$hp : παίρνει τη στήλη hp με βάση το όνομά της.
- mtcars[,4] : παίρνει την 4η στήλη του πίνακα.

Επιλογή μιας συγκεκριμένης τιμής

```
mtcars$hp[3]
# or
mtcars[3,4]
```

- mtcars$hp[3] : παίρνει την 3η τιμή της μεταβλητής hp.
- mtcars[3,4] : παίρνει το στοιχείο που βρίσκεται στην 3η γραμμή και 4η στήλη.

Άρα:

- dataset$variable → επιλογή μεταβλητής με όνομα
- dataset[i,j] → επιλογή στοιχείου/γραμμής/στήλης με δείκτες

3. Βασικά γραφήματα στην R

Ιστόγραμμα

hist(mtcars$hp)

Το ιστόγραμμα δείχνει **πώς κατανέμονται οι τιμές** της μεταβλητής hp.

Πυκνότητα kernel

plot(density(mtcars$hp))

Η εντολή density() υπολογίζει μια **ομαλή εκτίμηση της πυκνότητας** της κατανομής.
Είναι μια πιο "ομαλή" εναλλακτική του ιστογράμματος.

4. Δημιουργία νέας μεταβλητής

Η εντολή mtcars$lt_100km = 235.215/mtcars$mpg δημιουργεί νέα στήλη στο mtcars, μετατρέποντας την κατανάλωση από miles per gallon (mpg) σε λίτρα ανά 100 χιλιόμετρα. Η πράξη γίνεται για όλες τις παρατηρήσεις ταυτόχρονα.

mtcars$lt_100km = 235.215/mtcars$mpg

mtcars

Η μεταβλητή mpg σημαίνει **miles per gallon**.

Με τον τύπο:

$$lt\_100km = \frac{235.215}{mpg}$$

τη μετατρέπουμε σε **λίτρα ανά 100 χιλιόμετρα**, που είναι πιο οικεία μονάδα κατανάλωσης.

Άρα δημιουργούμε μια νέα μεταβλητή:

- lt_100km = κατανάλωση σε λίτρα / 100 χλμ.

5. Συνδυασμός ιστογράμματος και πυκνότητας

Δύο γραφήματα δίπλα-δίπλα

```
par(mfrow=c(1,2))
hist(mtcars$lt_100km, freq=FALSE)
plot(density(mtcars$lt_100km), col="red")
```

- par(mfrow=c(1,2)) : χωρίζει το παράθυρο γραφημάτων σε 1 γραμμή και 2 στήλες.
- freq=FALSE : το ιστόγραμμα εμφανίζει πυκνότητες αντί για απόλυτες συχνότητες.

Πυκνότητα πάνω στο ιστόγραμμα

```
par(mfrow=c(1,1))
hist(mtcars$lt_100km, freq=FALSE)
lines(density(mtcars$lt_100km), col="red")
```

Εδώ:

- φτιάχνουμε πρώτα το ιστόγραμμα
- μετά προσθέτουμε πάνω του την καμπύλη πυκνότητας με lines()

Αυτό βοηθά να βλέπουμε καλύτερα το σχήμα της κατανομής.

6. Απλή και πολλαπλή γραμμική παλινδρόμηση (OLS)

6.1 Απλό διάγραμμα διασποράς

```
plot(mtcars$hp , mtcars$lt_100km)
```

Το διάγραμμα διασποράς δείχνει τη σχέση μεταξύ:

- hp = ιπποδύναμη
- lt_100km = κατανάλωση

Μας βοηθά να δούμε αν υπάρχει περίπου γραμμική σχέση.

6.2 Απλή παλινδρόμηση

```
ols = lm(lt_100km ~ hp, data=mtcars)
summary(ols)
```

Η `lm(lt_100km ~ hp, data = mtcars)`

εκτιμά το υπόδειγμα:

$$lt\_100km_i = \beta_0 + \beta_1 hp_i + u_i$$

όπου:

- β_0 = σταθερός όρος
- β_1 = οριακή επίδραση της ιπποδύναμης στην κατανάλωση
- u_i = σφάλμα

Η summary(ols) δίνει:

- εκτιμήσεις συντελεστών
- τυπικά σφάλματα
- t-tests
- p-values
- R^2
- F-test

Από τα αποτελέσματα παίρνουμε:

$$\widehat{lt\_100km} = 6.44909 + 0.04299\, hp$$

Ερμηνεία

Αν ο συντελεστής του hp είναι θετικός, τότε μεγαλύτερη ιπποδύναμη συνδέεται με μεγαλύτερη κατανάλωση.

Ερμηνεία συντελεστών

- **Σταθερός όρος = 6.44909**
  Αν υποθετικά $hp = 0$, η προβλεπόμενη κατανάλωση είναι περίπου **6.45 λίτρα/100 χλμ.**
  Συνήθως η οικονομική/πρακτική σημασία του σταθερού όρου δεν είναι μεγάλη, αλλά χρειάζεται στο μοντέλο.

- **Συντελεστής του hp = 0.04299**
  Για **κάθε επιπλέον 1 μονάδα ιπποδύναμης**, η αναμενόμενη κατανάλωση αυξάνεται κατά περίπου:

$$0.043 \text{ λίτρα/100 χλμ.}$$

Άρα η σχέση είναι **θετική**: πιο ισχυρός κινητήρας → μεγαλύτερη κατανάλωση.

Παράδειγμα: αν η ιπποδύναμη αυξηθεί κατά 10 hp, τότε η κατανάλωση αυξάνεται κατά περίπου:

$$10 \times 0.04299 = 0.4299$$

δηλαδή περίπου **0.43 λίτρα/100 χλμ.**

6.3 Πολλαπλή παλινδρόμηση

```
ols1 = lm(lt_100km ~ hp + disp + wt + qsec, data=mtcars)
summary(ols1)
confint(ols1)
```

Το υπόδειγμα γίνεται:

$$lt\_100km_i = \beta_0 + \beta_1 hp_i + \beta_2 disp_i + \beta_3 wt_i + \beta_4 qsec_i + u_i$$

όπου:

- disp = κυβισμός
- wt = βάρος
- qsec = χρόνος στο 1/4 του μιλίου

Γιατί βάζουμε περισσότερες μεταβλητές;

Για να ελέγξουμε για περισσότερους παράγοντες που επηρεάζουν την κατανάλωση.

6.4 Εναλλακτικά υποδείγματα

```
ols2 = lm(lt_100km ~ hp + wt + qsec, data=mtcars)
summary(ols2)
confint(ols2)
```

```
ols3 = lm(lt_100km ~ hp + wt, data=mtcars)
summary(ols3)
confint(ols3)
```

7. Διάστημα εμπιστοσύνης συντελεστών

confint(ols3)

Η confint() δίνει **διαστήματα εμπιστοσύνης** για τους συντελεστές του μοντέλου.

8. Κατάλοιπα και διαγνωστικοί έλεγχοι

```
myseries = ols3$residuals
myseries2 = ols3$fitted.values
```

ή ισοδύναμα:

```
myseries = residuals(ols3)
myseries2 = fitted(ols3)
```

8.1 Γραφήματα καταλοίπων

```
par(mfrow=c(2,2))

plot(myseries, type='l', main='Residuals')
abline(h=0)
```

Το γράφημα αυτό δείχνει τα κατάλοιπα με τη σειρά τους.
Ιδανικά:

- να κινούνται τυχαία γύρω από το 0
- να μην εμφανίζουν μοτίβο

Η abline(h=0) βάζει οριζόντια γραμμή στο 0.

8.3 Ιστόγραμμα και πυκνότητα καταλοίπων

```
hist(myseries, freq=FALSE)
lines(density(myseries), col="red")
```

Ελέγχουμε αν τα κατάλοιπα μοιάζουν περίπου κανονικά.

8.4 Έλεγχοι κανονικότητας

8.4.1. Kolmogorov-Smirnov test

```
ks.test(myseries, "pnorm", mean = mean(myseries), sd = sd(myseries))
```

Ελέγχει αν η κατανομή της myseries είναι συμβατή με **κανονική κατανομή** με μέση τιμή mean(myseries) και τυπική απόκλιση sd(myseries).

- H_0: τα κατάλοιπα ακολουθούν κανονική κατανομή
- H_1: τα κατάλοιπα δεν ακολουθούν κανονική κατανομή

2. Shapiro-Wilk test

```
shapiro.test(myseries)
```

Είναι από τους πιο συνηθισμένους και πιο αξιόπιστους ελέγχους κανονικότητας, ειδικά σε μικρά και μεσαία δείγματα.

- H_0: τα δεδομένα είναι κανονικά
- H_1: τα δεδομένα δεν είναι κανονικά

Αν το p-value < 0.05, απορρίπτουμε την κανονικότητα.

3. Jarque-Bera test

```
jarque.bera.test(myseries)
```

Ελέγχει κανονικότητα με βάση:

- την **ασυμμετρία** (skewness)
- την **κύρτωση** (kurtosis)

Σε κανονική κατανομή θέλουμε:

- ασυμμετρία κοντά στο 0
- κύρτωση κοντά στην κανονική
- H_0: τα δεδομένα είναι κανονικά
- H_1: τα δεδομένα δεν είναι κανονικά

Αν το p-value < 0.05, απορρίπτουμε την κανονικότητα.

8.5 Κατάλοιπα ως προς προσαρμοσμένες τιμές

plot(myseries2, myseries)

Δηλαδή:

- στον οριζόντιο άξονα: fitted values
- στον κάθετο άξονα: residuals

Χρήσιμο για έλεγχο:

- ετεροσκεδαστικότητας
- μη γραμμικότητας
- ύπαρξης μοτίβων

Ιδανικά, τα σημεία πρέπει να είναι διάσπαρτα τυχαία γύρω από το 0.

8.6 Αυτοσυσχέτιση καταλοίπων

acf(myseries)

Η acf() δείχνει τη συνάρτηση αυτοσυσχέτισης των καταλοίπων.

Σε διαστρωματικά δεδομένα, συνήθως δεν περιμένουμε έντονη αυτοσυσχέτιση. Σε χρονολογικές σειρές ο έλεγχος είναι χρήσιμος.