ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ

ΓΡΑΜΜΙΚΑ ΜΟΝΤΕΛΑ

ΕΡΓΑΣΤΗΡΙΑΚΗ ΑΣΚΗΣΗ #1

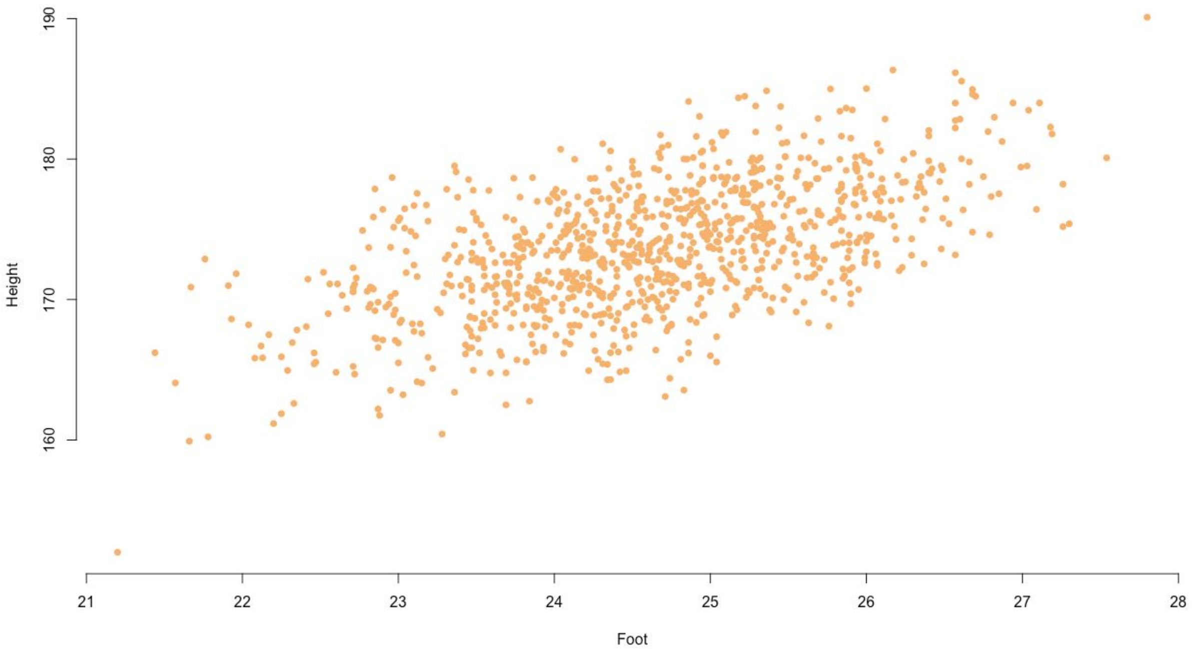
ΧΑΡΑΛΑΜΠΟΣ ΡΟΔΗΣ
6180128

Ο στόχος μας είναι να μελετήσουμε τη σχέση του `συνολικού ύψους` ενός ανθρώπου με το `μήκος των ποδιών` του[1] σε ένα δείγμα κατοίκων της Ινδίας.

1. Αρχικά θα **εισάγουμε τα δεδομένα** στην R και στη συνέχεια θα σχεδιάσουμε ένα **διάγραμμα διασποράς** των δύο μεταβλητών που μας ενδιαφέρουν.

```
#Load data
india_foot_height <- read.table("india_foot_height.dat")
names(india_foot_height)[1:2] <- c("Foot", "Height")
attach(india_foot_height)

#Plot data
plot(Foot, Height, pch = 16, bty = "n", col = "#f6b26b")
```



   i. Τα σημεία στο γράφημα αναπαριστούν όλα τα **ζεύγη τιμών** των μεταβλητών μας (δηλαδή για κάθε τιμή της μεταβλητής **`Foot`**, ποια τιμή της μεταβλητής **`Height`** αντιστοιχεί στα δεδομένα του δείγματός μας).

   ii. Στο γράφημα παρατηρούμε μια **θετική γραμμική σχέση** ανάμεσα στις δύο μεταβλητές (δηλαδή όσο μεγαλύτερη η τιμή της μεταβλητής **`Foot`**, τόσο μεγαλύτερη αναμένουμε να είναι η τιμή της μεταβλητής **`Height`**).

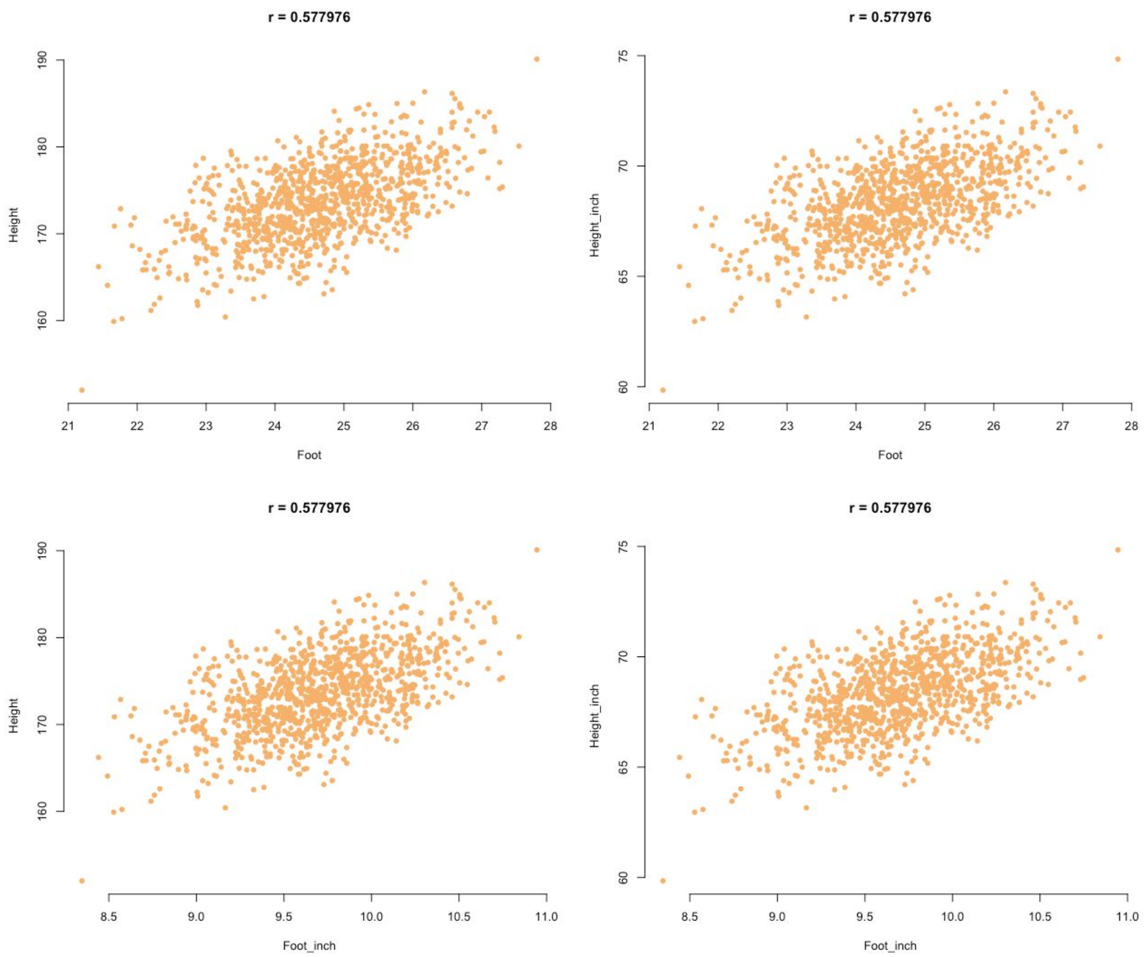
[1] Τουλάχιστον αυτή είναι η δική μου ερμηνεία της μεταβλητής **`foot`** των δεδομένων.

2. Η παρατηρούμενη σχέση στο παραπάνω γράφημα καθώς και ο αντίστοιχος **δειγματικός συντελεστής γραμμικής συσχέτισης** `r` των δύο μεταβλητών:

  i. **Δεν εξαρτάται από τις μονάδες μέτρησης των δύο μεταβλητών**. Αυτό μπορούμε να το επιβεβαιώσουμε μετατρέποντας τα δεδομένα μας σε άλλες μονάδες μέτρησης και σχεδιάζοντας τα αντίστοιχα διαγράμματα διασποράς.

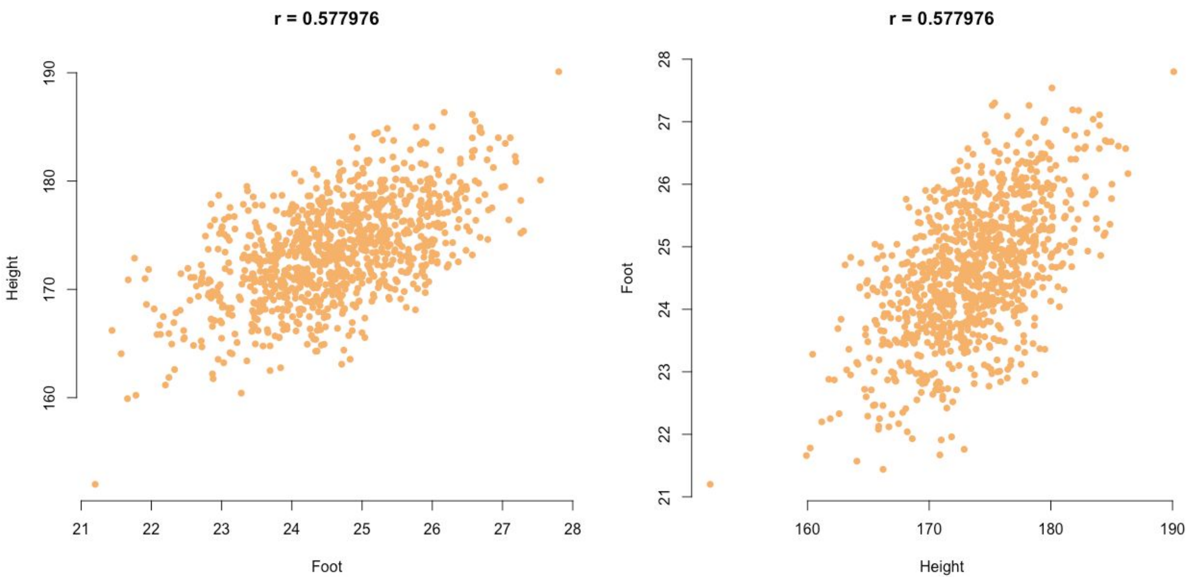
```
#Convert cm to inch
india_foot_height$Foot_inch <- Foot * 0.393701
india_foot_height$Height_inch <- Height * 0.393701

#Plot and compare all possible combinations of units
par(mfrow=c(2,2), pch = 16, bty = "n", col = "#f6b26b")
plot(Foot, Height, main = paste("r =", round(cor(Foot, Height), 6)))
plot(Foot, Height_inch, main = paste("r =", round(cor(Foot, Height_inch),
6)))
plot(Foot_inch, Height, main = paste("r =", round(cor(Foot_inch, Height),
6)))
plot(Foot_inch, Height_inch, main = paste("r =", round(cor(Foot_inch,
Height_inch), 6)))
```



ii. **Δεν αλλάζει αν αντιστρέψουμε το γράφημα**, δηλαδή ποια μεταβλητή έχουμε στον κάθετο (ορίζουμε ως *εξαρτημένη*) και ποια στον οριζόντιο άξονα (ορίζουμε ως *ανεξάρτητη*). Μπορούμε να το επιβεβαιώσουμε με τα παρακάτω γραφήματα.

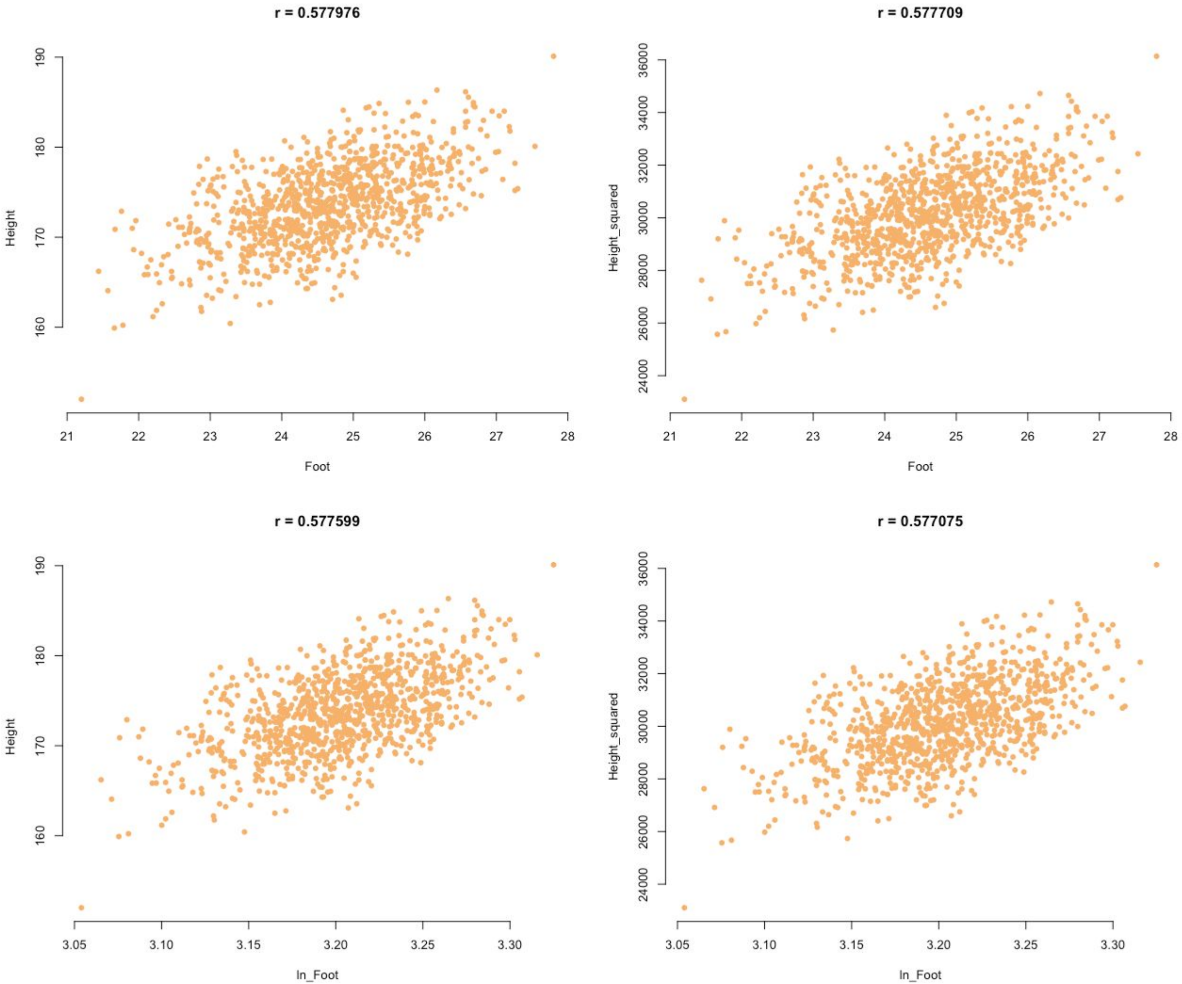
```
#Invert axes of scatterplot and compare r
par(mfrow=c(1,2), pch = 16, bty = "n", col = "#f6b26b")
plot(Foot, Height, main = paste("r =", round(cor(Foot, Height), 6)))
plot(Height, Foot, main = paste("r =", round(cor(Height, Foot), 6)))
```



iii. **Μεταβάλλεται αν μετασχηματίσουμε μη-γραμμικά μία ή και τις δύο μεταβλητές**. Αν για παράδειγμα υψώσουμε στο τετράγωνο τις τιμές του `Height` ή πάρουμε τον φυσικό λογάριθμο των τιμών του `Foot`, τότε ο αντίστοιχος συντελεστής γραμμικής συσχέτισης θα αλλάξει, όπως μπορούμε να επιβεβαιώσουμε παρακάτω[2].

```
#Convert foot to ln(foot) and height to height^2
india_foot_height$ln_Foot <- log(Foot)
india_foot_height$Height_squared <- Height^2

#Plot and compare all combinations of non-linear transformations
par(mfrow=c(2,2), pch = 16, bty = "n", col = "#f6b26b")
plot(Foot, Height, main = paste("r =", round(cor(Foot, Height), 6)))
plot(Foot, Height_squared, main = paste("r =", round(cor(Foot,
Height_squared), 6)))
plot(ln_Foot, Height, main = paste("r =", round(cor(ln_Foot, Height), 6)))
plot(ln_Foot, Height_squared, main = paste("r =", round(cor(ln_Foot,
Height_squared), 6)))
```



[2] Παρόλο που στα διαγράμματα, λόγω της ελάχιστης μεταβολής του r σε αυτούς τους μη-γραμμικούς μετασχηματισμούς, δεν παρατηρούμε κάποια οφθαλμοφανή αλλαγή στη συσχέτιση των δύο μεταβλητών.

3. Υπολογίζουμε τον **δειγματικό συντελεστή γραμμικής συσχέτισης** του Pearson (`r`) για τις δύο μεταβλητές μας, και **ελέγχουμε τη στατιστική του σημαντικότητα**.

```
#Calculate r and evaluate its statistical significance
cor.test(Foot, Height)
```

```
                    Pearson's product-moment correlation
data:  Foot and Height
t = 22.598, df = 1018, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5355953 0.6174523
sample estimates:
      cor
0.5779759
```

i. Η τιμή του συντελεστή γραμμικής συσχέτισης του Pearson στο δείγμα μας είναι `0.5779759`.

ii. Η τιμή αυτή υποδεικνύει πως υπάρχει μια θετική γραμμική συσχέτιση ανάμεσα στις μεταβλητές `Foot` και `Height`.

iii. Πραγματοποιήθηκε **δίπλευρος έλεγχος t** για τις παρακάτω εναλλακτικές υποθέσεις:

- $\mathbf{H_0}$: $\rho = 0$
- $\mathbf{H_1}$: $\rho \neq 0$

Όπου **ρ** είναι η πραγματική (*πληθυσμιακή*) τιμή του συντελεστή γραμμικής συσχέτισης του Pearson.

iv. Τα αποτελέσματα του ελέγχου, με επίπεδο στατιστικής σημαντικότητας `5%`, μας οδηγούν στην **απόρριψη της μηδενικής υπόθεσης**, δηλαδή έχουμε ισχυρές ενδείξεις πως υπάρχει γραμμική συσχέτιση ανάμεσα στις δύο μεταβλητές στον πληθυσμό.

a. Η πιθανότητα (**p-value**) λήψης ενός αποτελέσματος ίσου ή περισσότερο ακραίου από το παρατηρούμενο στο συγκεκριμένο δείγμα, κάτω από την παραδοχή της $\mathbf{H_0}$, υπολογίζεται στο 2.2×10^{-16}, σημαντικά μικρότερη από το `0.05` του επιπέδου στατιστικής σημαντικότητας του ελέγχου. Άρα μπορούμε να απορρίψουμε την μηδενική υπόθεση.

b. Εκτιμούμε επίσης, με βεβαιότητα `95%`, πως το διάστημα εμπιστοσύνης `[0.5355953, 0.6174523]` περιέχει την **πραγματική τιμή της παραμέτρου ρ**. Άρα, αφού η τιμή ελέγχου βρίσκεται εκτός αυτού του διαστήματος, μπορούμε να απορρίψουμε την μηδενική υπόθεση.

ΔΕΙΤΕ ΤΟΝ ΠΛΗΡΗ ΚΩΔΙΚΑ ΣΕ R ΓΙΑ ΤΗΝ ΕΡΓΑΣΙΑ:

`https://github.com/harrisrodis/linear-models-lab/blob/master/AS01/AS01.R`

ΚΑΤΕΒΑΣΤΕ ΤΗΝ ΕΡΓΑΣΙΑ ΣΕ ΑΡΧΕΙΟ WORD:

`https://www.dropbox.com/s/nu19tv7slik2ez4/Rodis_6180128_LMlab_AS01.docx?dl=1`