**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]](#footnote-0) 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

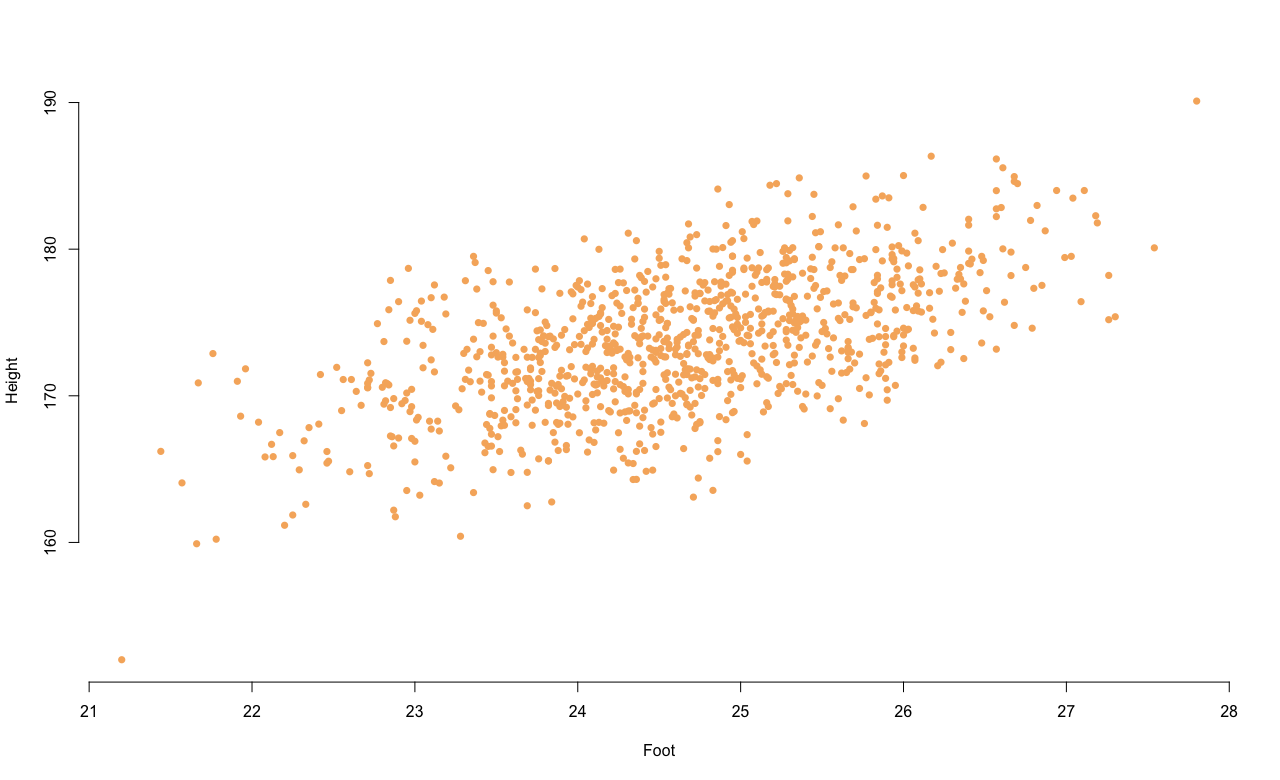
india\_foot\_height <- **read.table**("india\_foot\_height.dat")

**names**(india\_foot\_height)[1:2] <- **c**("Foot", "Height")

**attach**(india\_foot\_height)

#Plot data

**plot**(Foot, Height, pch = 16, bty = "n", col = "#f6b26b")



1. 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).
2. 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**).
3. Η παρατηρούμενη σχέση στο παραπάνω γράφημα καθώς και ο αντίστοιχος **δειγματικός συντελεστής γραμμικής συσχέτισης r**  των δύο μεταβλητών:
   1. **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 and plot all possible combinations

india\_foot\_height$Foot\_inch <- Foot \* 0.393701

india\_foot\_height$Height\_inch <- Height \* 0.393701

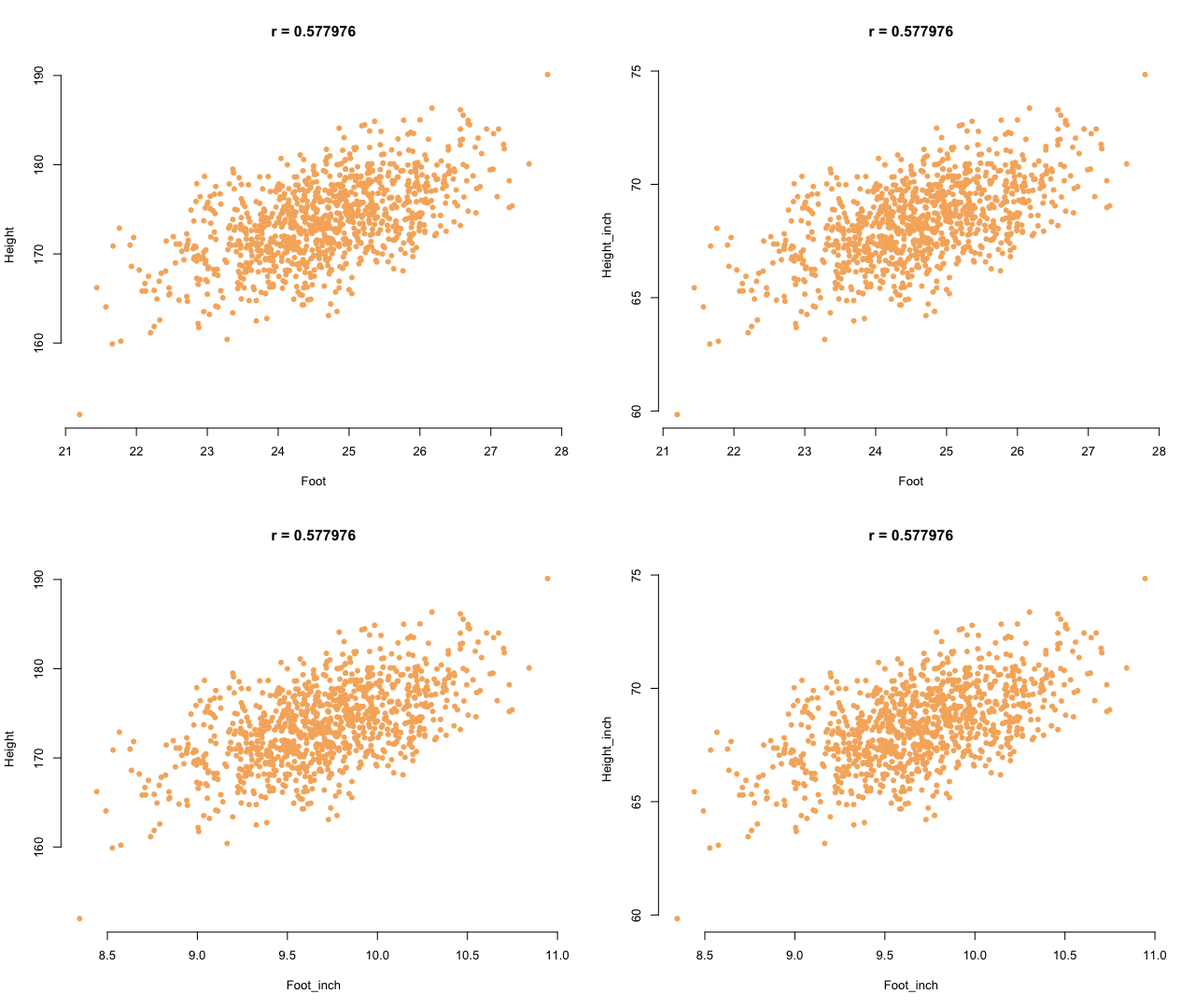
**attach**(india\_foot\_height)

**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)))

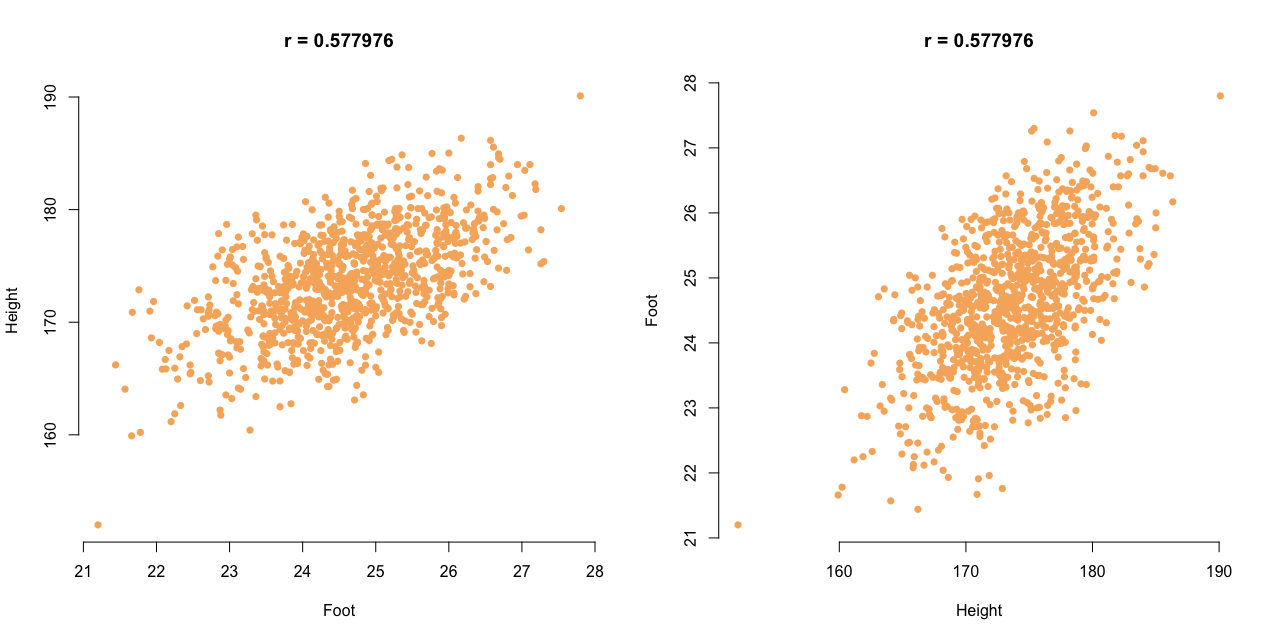
* 1. **Δεν αλλάζει αν αντιστρέψουμε το γράφημα**, δηλαδή ποια μεταβλητή έχουμε στον κάθετο (ορίζουμε ως *εξαρτημένη*) και ποια στον οριζόντιο άξονα (ορίζουμε ως *ανεξάρτητη*). Μπορούμε να το επιβεβαιώσουμε με τα παρακάτω γραφήματα.

#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)))



* 1. **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]](#footnote-1).

#Convert foot to ln(foot) and height to height^2 and plot again all combinations

india\_foot\_height$ln\_Foot <- **log**(Foot)

india\_foot\_height$Height\_squared <- Height^2

**attach**(india\_foot\_height)

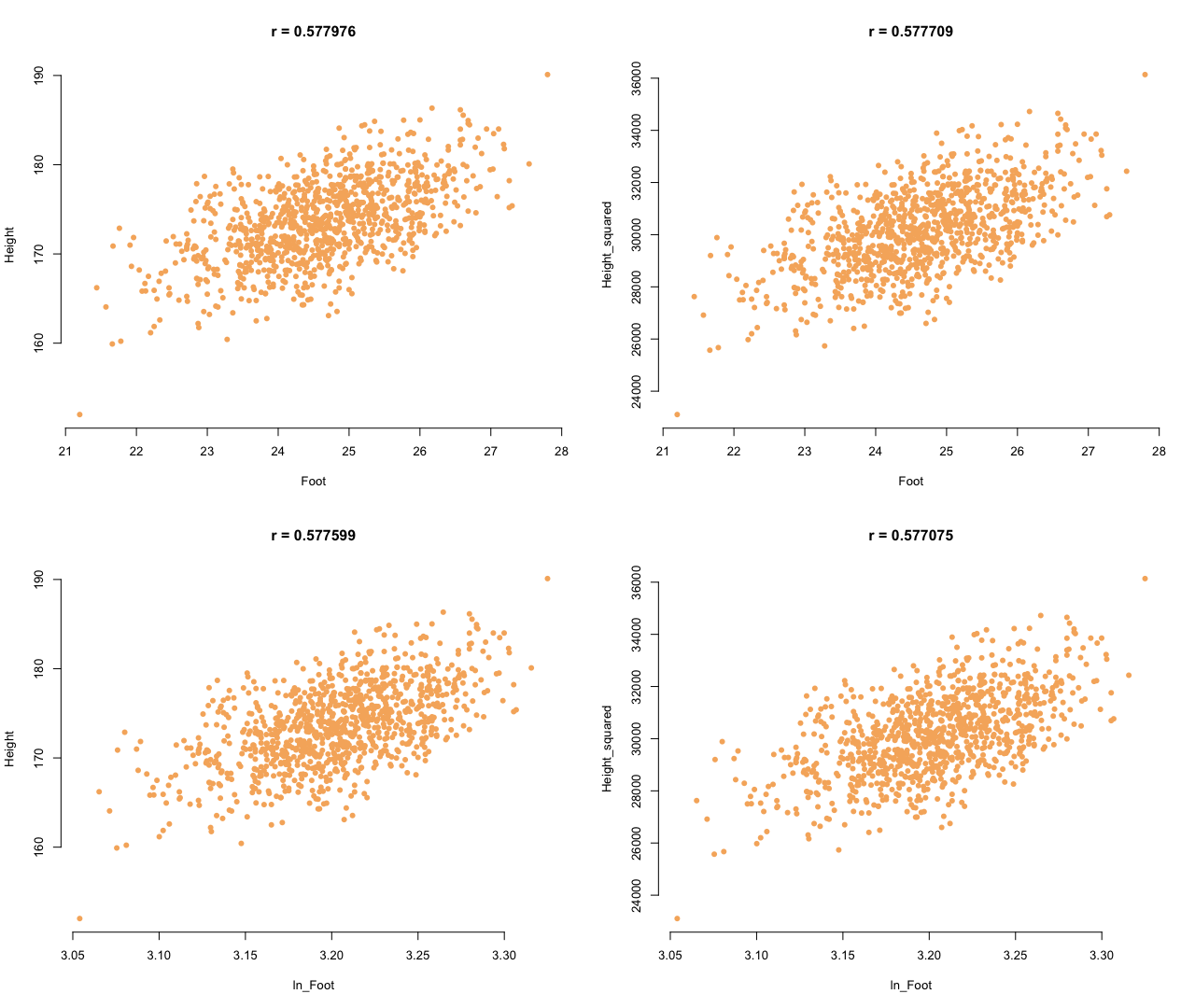
**par**(mfrow=c(2,2), pch = 16, bty = "n", col = "#f6b26b")

**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)))



1. 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

1. Η τιμή του συντελεστή γραμμικής συσχέτισης του Pearson στο δείγμα μας είναι **0.5779759**.
2. Η τιμή αυτή υποδεικνύει πως υπάρχει μια θετική γραμμική συσχέτιση ανάμεσα στις μεταβλητές **Foot** και **Height**.
3. Πραγματοποιήθηκε **δίπλευρος** **έλεγχος t** για τις παρακάτω εναλλακτικές υποθέσεις:
   * + **Η0**: ρ = 0
     + **Η1**: ρ ≠ 0

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

1. 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.
   1. H πιθανότητα (**p-value**) λήψης ενός αποτελέσματος ίσου ή περισσότερο ακραίου από το παρατηρούμενο στο συγκεκριμένο δείγμα, κάτω από την παραδοχή της **Η0**, 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.
   2. 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.

1. Τουλάχιστον αυτή είναι η δική μου ερμηνεία της μεταβλητής **foot** των δεδομένων. [↑](#footnote-ref-0)
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. [↑](#footnote-ref-1)