Γραμμικά Μοντέλα ~ R
Scatter-Correlation

Γραμμικά Μοντέλα ~ R

Γραμμικό Μοντέλο

Γενική μορφή γραμμικού μοντέλου

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

X: ανεξάρτητη μεταβλητή - predictor
Y: εξαρτημένη μεταβλητή - response

Σκοπός: Εκτίμηση των **παραμέτρων** α και β,που αναφέρονται στον πληθυσμό χρησιμοποιώντας τις παρατηρήσεις του δείγματος με τη μέθοδο των ελάχιστων τετράγωνων.
Υποθέσεις για τα σφάλματα ε_i :
ακολουθούν, ανεξάρτητα και ισόνομα, την κανονική κατανομή - $\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$ και είναι ανεξάρτητα της μεταβλητής X.

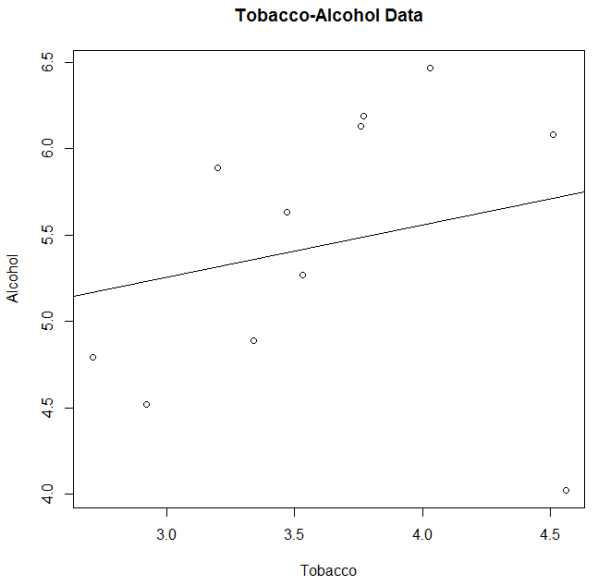
Γραμμικά Μοντέλα ~ R

Διάγραμμα Διασποράς- Scatter

scatter *vi*	(go in all directions)	σκορπίζομαι, διαλύομαι *ρ αμ*

Ένα διάγραμμα διασποράς αποτυπώνει τη σχέση μεταξύ δύο μεταβλητών, τη διασπορά τους και πιθανά προβλήματα, όπως έκτροπες τιμές των δεδομένων

Η χρησιμότητα ενός διαγράμματος διασποράς είναι να ανιχνεύσουμε μία πιθανή Μη γραμμική σχέση μεταξύ δύο μεταβλητών




```
> plot(Tobacco,Alcohol,main="Tobacco-Alcohol Data")
```

Γραμμικά Μοντέλα ~ R

Συντελεστής **Γραμμικής** Συσχέτισης ~ Pearson

Ο συντελεστής **γραμμικής** συσχέτισης ρ ορίζεται ως

$$Corr(X,Y) = \rho_{XY} = \frac{Cov(X,Y)}{\sqrt{\sigma_X^2 \sigma_Y^2}} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

Ο δειγματικός συντελεστής **γραμμικής** συσχέτισης r ορίζεται ως

$$r_{XY} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} = \frac{S_{XY}}{\sqrt{S_{XX}S_{YY}}} = \frac{S_{XY}}{(n-1)S_X S_Y}$$

Γραμμικά Μοντέλα ~ R

Συντελεστής **Γραμμικής** Συσχέτισης ~ Pearson

Εφόσον **Δεν** ανιχνεύσουμε απόκλιση από τη γραμμικότητα μπορούμε:

✓να υπολογίσουμε το **μέτρο** της γραμμικής συσχέτισης δύο μεταβλητών με χρήση του Pearson 's *r* και

✓να κάνουμε έλεγχο για την υπόθεση $H_0 : \rho = 0$ $H_0 : \rho \neq 0$

•Εντολή **cor**(x, y, .., method = c("**pearson(default)** ", "kendall", "spearman"))

```
> cor(Alcohol,Tobacco )
```

Εντολή **cor.test**(x, y, ..., method = c("**pearson**", "kendall", "spearman"), ...)

```
> cor.test(Alcohol,Tobacco )
```

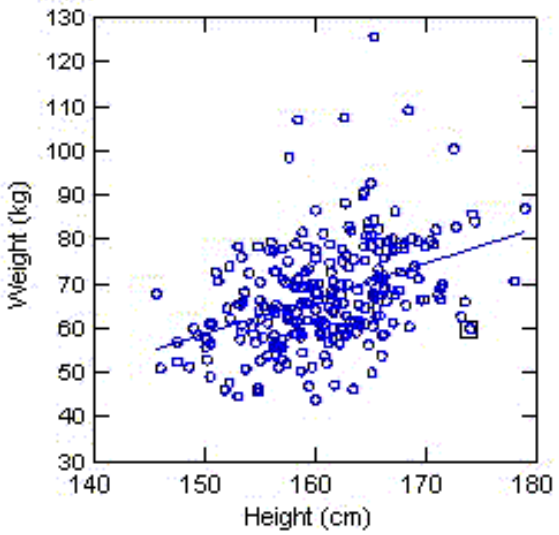
Γραμμικά Μοντέλα ~ R

Ιδιότητες Συντελεστής Γραμμικής Συσχέτισης ~ Pearson

1. Δεν εξαρτάται από τις μονάδες μέτρησης των X, Y
2. Δεν εξαρτάται από το ποια αποκαλούμε X και ποια Y
3. Παίρνει τιμές μεταξύ -1 και $+1$
4. Αν $r < 0$ τότε έχουμε αρνητική γραμμική συσχέτιση (κλίση)
5. Αν $r > 0$ τότε έχουμε θετική γραμμική συσχέτιση (κλίση)
6. Αν $r \cong \pm 1$ τότε έχουμε πολύ ισχυρή γραμμική συσχέτιση
7. Το r μεταβάλλεται όταν μετασχηματίσουμε X ή/και Y μη γραμμικά
8. Το r μετρά το μέγεθος της συσχέτισης μεταξύ των X, Y αλλά η συσχέτιση δεν οδηγεί αυτόματα και σε αιτιολόγηση.

Γραμμικά Μοντέλα ~ R

Διάγραμμα Διασποράς ~ Συντελεστής Συσχέτισης



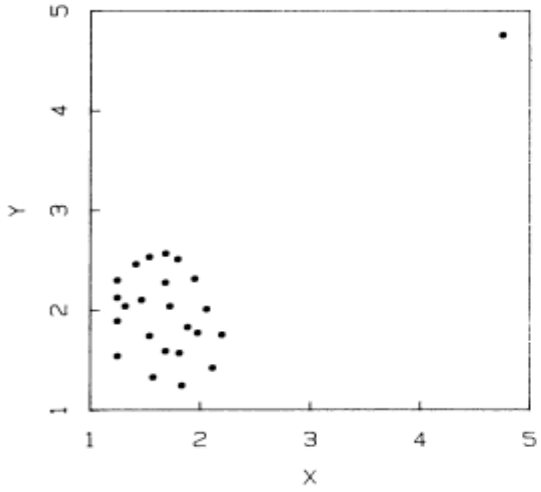
Αν καθώς αυξάνουν οι τιμές της μίας μεταβλητής παρατηρούμε μία «τάση» να αυξάνουν και οι τιμές της άλλης μεταβλητής, τότε έχουμε **θετική** συσχέτιση

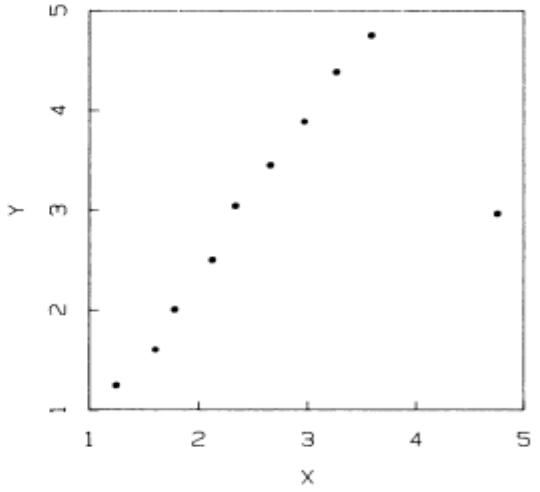
Αν καθώς αυξάνουν οι τιμές της μίας μεταβλητής παρατηρούμε μία «τάση» να μικραίνουν οι τιμές της άλλης μεταβλητής, τότε έχουμε **αρνητική** συσχέτιση

Η λέξη «τάση» χρησιμοποιήθηκε για να δηλώσουμε ότι μας ενδιαφέρει η σχέση για τον κύριο όγκο των τιμών και όχι ένα μεμονωμένο ζευγάρι τιμών

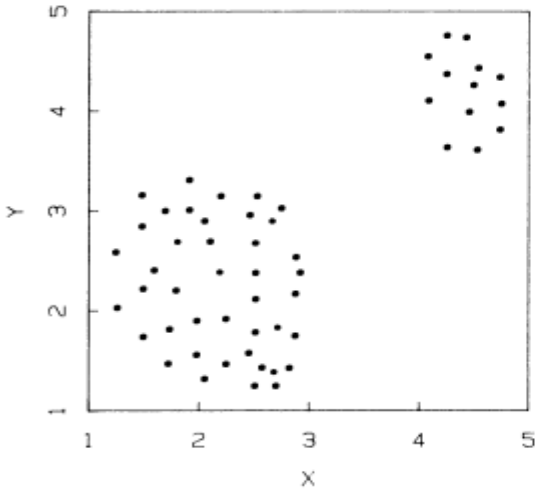
Γραμμικά Μοντέλα ~ R

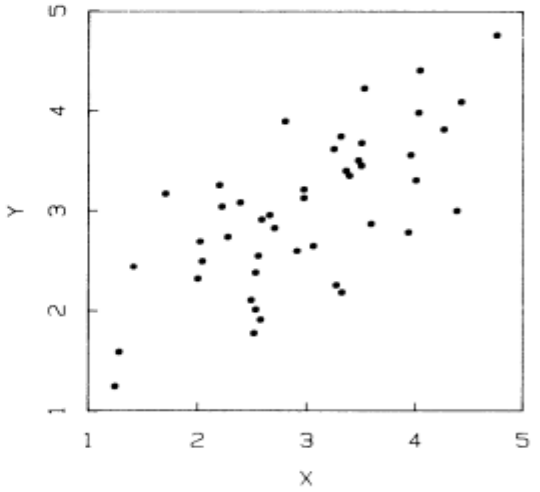
Διάγραμμα Διασποράς ~ Συντελεστής Γραμμικής Συσχέτισης





r=0.7!!!

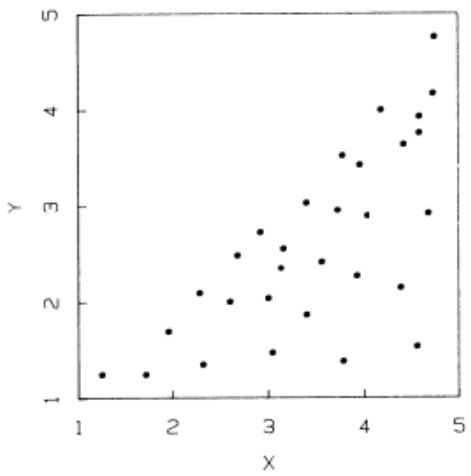


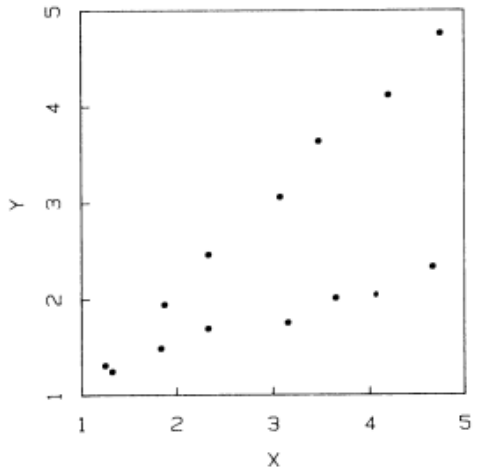


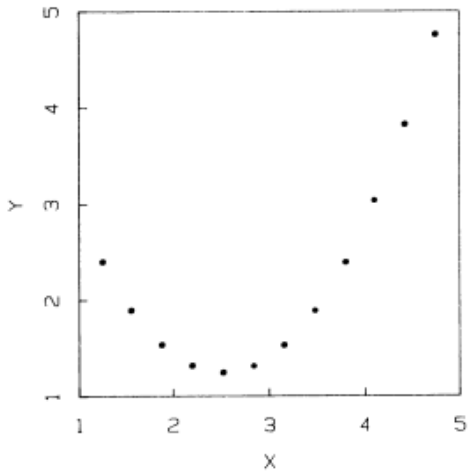
Γραφήματα από Chambers, John, William Cleveland, Beat Kleiner, and Paul Tukey, (1983), *Graphical Methods for Data Analysis*, Wadsworth.

Γραμμικά Μοντέλα ~ R

Διάγραμμα Διασποράς ~ Συντελεστής Γραμμικής Συσχέτισης







r=0.7!!!

Γραφήματα από Chambers, John, William Cleveland, Beat Kleiner, and Paul Tukey, (1983), *Graphical Methods for Data Analysis*, Wadsworth.

Γραμμικά Μοντέλα ~ R

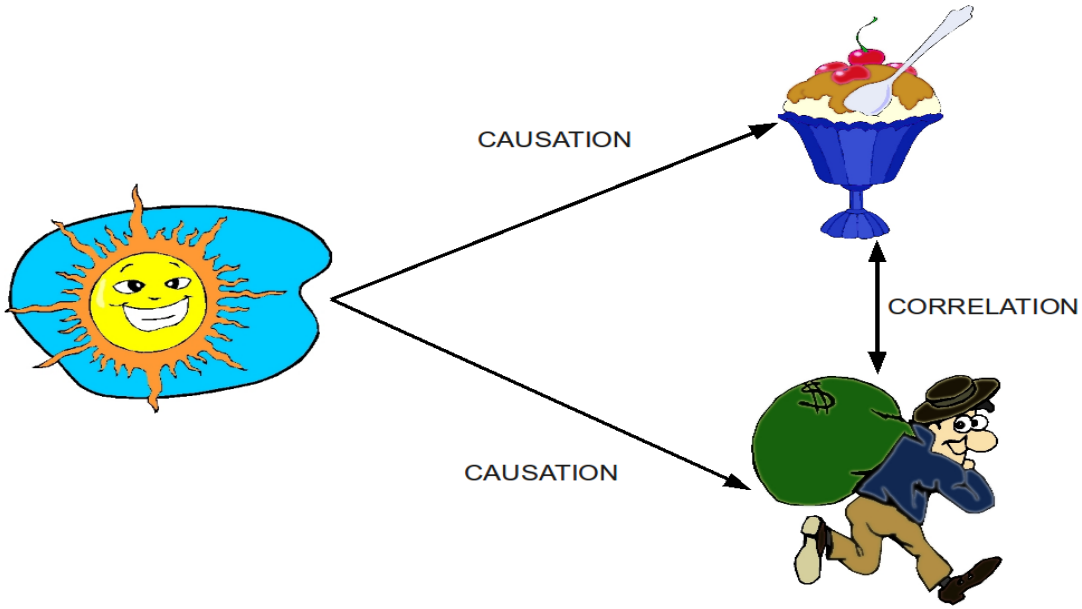
Spurious (Ψευδείς) Correlations

http://tylervigen.com/discover

Γραμμικά Μοντέλα ~ R

Correlation ≠ Causation

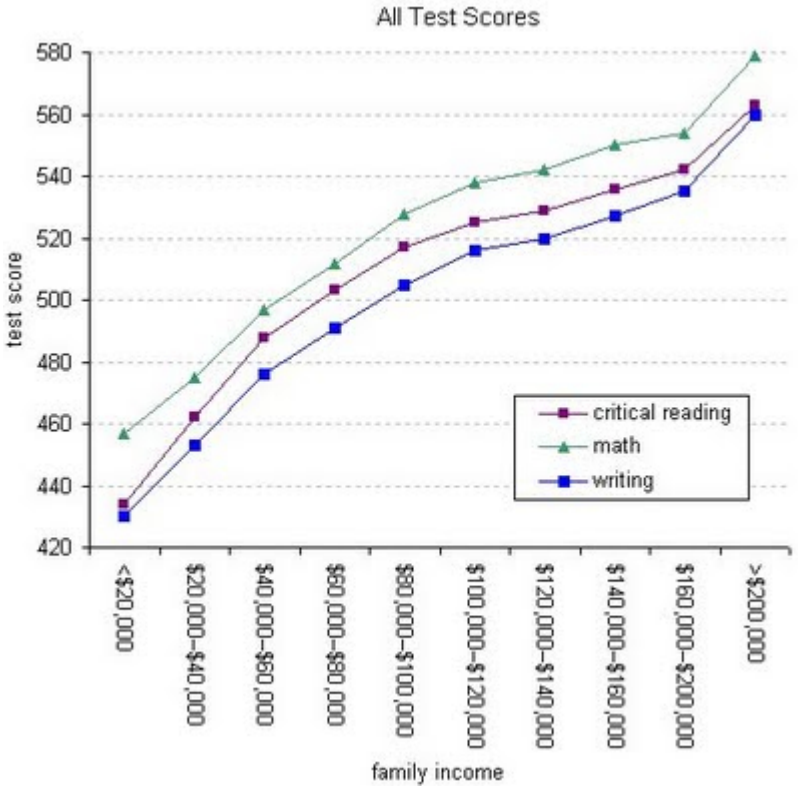
Συσχέτιση ≠ Αιτιότητα

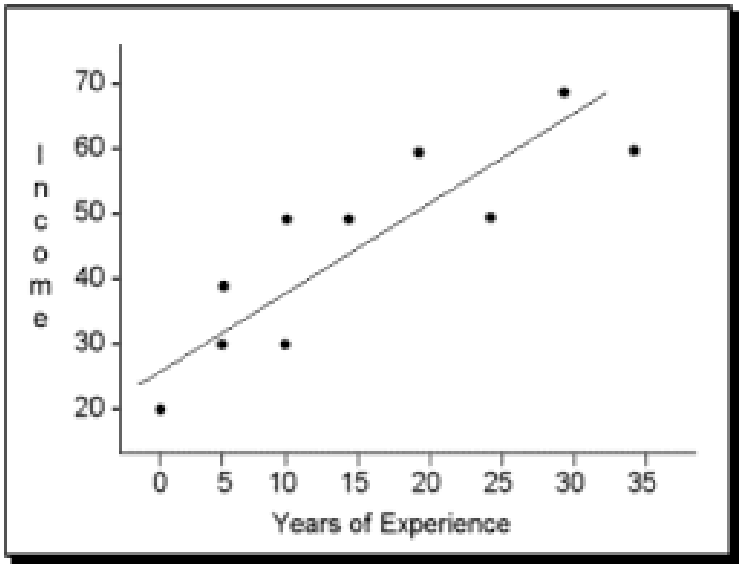


Γραμμικά Μοντέλα ~ R

Correlation ~ Causation

Συσχέτιση ~ Αιτιότητα





Εργασία

Χρησιμοποιώντας τα δεδομένα του αρχείου “india_foot_height.dat”

1. Δώστε το κατάλληλο γράφημα για την παρουσίαση της κατανομής από κοινού των μεταβλητών foot (πόδια) και height(σε cm)
   i. Εξηγήστε το γράφημα
   ii. Τι παρατηρείτε ως προς τη σχέση των δύο μεταβλητών;

2. Επιβεβαιώστε ότι ισχύουν οι ιδιότητες 1, 2 και 7 της διαφάνειας 6 της παρουσίασης

3. Υπολογίστε τον συντελεστή γραμμικής συσχέτισης του Pearson κάνοντας τον κατάλληλο έλεγχο υπόθεσης.
   i. Ποια είναι η τιμή του συντελεστή γραμμικής συσχέτισης του Pearson;
   ii. Ερμηνεύστε την τιμή που υπολογίστηκε.
   iii. Ποιος έλεγχος πραγματοποιήθηκε; Δώστε τη μηδενική και εναλλακτική του Υπόθεση.
   iv. Σε ποιό συμπέρασμα καταλήγουμε ως προς τον έλεγχο
      a. Α. Με χρήση του p-value
      b. Β. Με χρήση του 95% Δ.Ε.