Πολλαπλή γραμμική εξάρτηση

Γιώτα Τουλούμη
Καθηγήτρια Βιοστατιστικής και Επιδημιολογίας
Εργ. Υγιεινής, Επιδημιολογίας και Ιατρικής Στατιστικής
Ιατρική Σχολή, ΕΚΠΑ
gtouloum@med.uoa.gr

Βάνα Σύψα
Καθηγήτρια Επιδημιολογίας και Ιατρικής Στατιστικής
Εργ. Υγιεινής, Επιδημιολογίας και Ιατρικής Στατιστικής
Ιατρική Σχολή, ΕΚΠΑ
vsipsa@med.uoa.gr

Στατιστικές δοκιμασίες για τη διερεύνηση σχέσης μεταξύ 2 παραγόντων

Παράγοντας 1	**Παράγοντας 2**	
	Ποσοτική (π.χ. χοληστερόλη)	Ποιοτική (π.χ. καρκίνος ναι/όχι)
Ποιοτική (π.χ. κάπνισμα ναι/όχι)	**t-test** (αν η ποιοτική έχει 2 επίπεδα)	**χ^2-test**
Ποσοτική (π.χ. ηλικία)	› **Συντελεστής συσχέτισης με στατιστική αξιολόγηση** › **Εξάρτηση**	

<u>Σημείωση</u>: Οι περισσότερες δοκιμασίες για **ποσοτικά** χαρακτηριστικά υποθέτουν την **κανονική** κατανομή των χαρακτηριστικών

Εισαγωγή στην πολλαπλή γραμμική εξάρτηση

- Χαρακτηριστικά των ατόμων όπως:

  ύψος, βάρος, επίπεδα χοληστερόλης, αρτηριακής πίεσης κλπ.

  παρουσιάζουν μεγάλη μεταβλητότητα

- Ιδανικά, θα θέλαμε να μπορούμε να εξηγήσουμε αυτή τη μεταβλητότητα,
  - π.χ. γιατί τα άτομα διαφέρουν πολύ ως προς το ύψος τους: ηλικία, φύλο, φυλή, ύψος γονιών κλπ

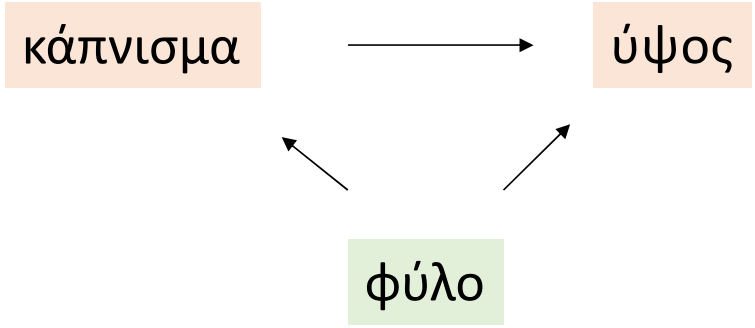
- Μέχρι τώρα έχουμε ασχοληθεί με το αν **μία μεταβλητή** σχετίζεται **με μία άλλη**, π.χ.
  - Το ύψος παιδιών με το φύλο τους
    - Μέσο ύψος σε αγόρια και κορίτσια – στατιστική αξιολόγηση με t-test
  - Το ύψος παιδιών με το ύψος του πατέρα
    - Συσχέτιση ή απλή γραμμική εξάρτηση

Ποια είναι τα μειονεκτήματα αυτών των μεθόδων;

1. Η μεταβλητότητα που παρουσιάζει ένα χαρακτηριστικό, π.χ. ύψος δεν οφείλεται σε ένα μόνο παράγοντα αλλά στην ταυτόχρονη επίδραση πολλών παραγόντων

2. Κάποιες σχέσεις μπορεί να μην είναι πραγματικές αλλά να οφείλονται σε ένα τρίτο παράγοντα (συγχυτικός παράγοντας),

π.χ. καπνιστές ψηλότεροι από τους μη-καπνιστές:



Πολλαπλή γραμμική εξάρτηση/παλινδρόμηση (Multiple linear regression)

Ιδανικά, θα θέλαμε

- να αξιολογήσουμε ταυτόχρονα την επίδραση πολλών παραγόντων στο ύψος και το βαθμό στον οποίο εξηγούν τη μεταβλητότητα από άτομο σε άτομο

- να αξιολογήσουμε την επίδραση κάθε παράγοντα λαμβάνοντας υπόψη και τους άλλους ώστε να **αποκλείσουμε** σχέσεις που οφείλονται σε **συγχυτικούς** παράγοντες

→ **Πολλαπλή γραμμική εξάρτηση**

Πολλαπλή γραμμική εξάρτηση

- **Αντικείμενο**

  Διερεύνηση του τρόπου μεταβολής μια μεταβλητής (εξαρτημένης) συναρτήσει των μεταβολών άλλων μεταβλητών (ανεξάρτητων)

  Διερευνάται η γραμμική σχέση μιας εξαρτημένης μεταβλητής με περισσότερες από μία ανεξάρτητες μεταβλητές

  Εξαρτημένη μεταβλητή→ υποχρεωτικά ποσοτική

  Οι ανεξάρτητες μεταβλητές μπορούν να είναι ποσοτικές ή/και ποιοτικές

- **Προϋπόθεση**

  Η **εξαρτημένη** μεταβλητή (Y) να κατανέμεται κανονικά για κάθε συνδυασμό τιμών των ανεξάρτητων μεταβλητών (X_i)

Μοντέλο πολλαπλής γραμμικής εξάρτησης

- Έστω $X_1, X_2, \ldots, X_p$ αντιπροσωπεύουν p ανεξάρτητες μεταβλητές, τότε

$$\hat{Y}_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + \quad \ldots + b_p X_{pi}$$

i=1,2,3,…,n άτομα

$\hat{Y}_i$: **Προβλεπόμενη τιμή** για το άτομο i

$b_1, b_2, \ldots, b_p$: **Συντελεστές <u>μερικής</u> εξάρτησης για τις ανεξάρτητες μεταβλητές $X_1, X_2, .., X_p$ αντίστοιχα**

(δε θα συζητήσουμε πώς εκτιμώνται από τα δεδομένα)

Ερμηνεία συντελεστών μερικής εξάρτησης

$$\hat{Y}_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + \quad \ldots + b_p X_{pi}$$

Ο συντελεστής b_j (j=1,..,p) εκφράζει την **αναμενόμενη** μεταβολή της εξαρτημένης μεταβλητής (Y) όταν η αντίστοιχη ανεξάρτητη X_j μεταβληθεί κατά **μία** μονάδα

→ όπως ερμηνεύαμε και στην απλή γραμμική εξάρτηση

και

δεδομένου ότι οι **υπόλοιπες ανεξάρτητες μεταβλητές παραμένουν σταθερές**

→ Επιπλέον για πολλαπλή γραμμική εξάρτηση

Παράδειγμα

Y = ύψος παιδιού

X_1=ύψος πατέρα

X_2=ηλικία παιδιού

$$\hat{Y} = b_0 + b_1 \cdot \text{ύψος πατέρα} + b_2 \cdot \text{ηλικία παιδιού}$$

b_1: πόσο διαφέρει κατά μέσο όρο το ύψος ενός παιδιού από ένα άλλο όταν ο πατέρας του είναι **1 cm ψηλότερος**, δεδομένου ότι τα παιδιά είναι **ίδιας** ηλικίας

b_2: πόσο διαφέρει κατά μέσο όρο το ύψος ενός παιδιού **ενός έτους μεγαλύτερου** από ένα άλλο παιδί όταν οι πατέρες έχουν το **ίδιο** ύψος

Αν έχουμε 2 παιδιά ίδιας ηλικίας (π.χ. 10 ετών) και ο πατέρας του ενός είναι 1.75 m και του άλλου 1.76 m, πόσο αναμένεται να διαφέρουν κατά μέσο όρο τα ύψη τους;

$$\hat{Y}_1 = b_0 + b_1 * 175 + b_2 * 10$$

$$\hat{Y}_2 = b_0 + b_1 * 176 + b_2 * 10$$

$$\hat{Y}_2 - \hat{Y}_1 = b_0 + b_1 * 176 + b_2 * 10 - (b_0 + b_1 * 175 + b_2 * 10)$$

$$\hat{Y}_2 - \hat{Y}_1 = \cancel{b_0} + b_1 * 176 + \cancel{b_2 * 10} - \cancel{b_0} - b_1 * 175 - \cancel{b_2 * 10})$$

$$\hat{Y}_2 - \hat{Y}_1 = b_1(176 - 175)$$

$$\hat{Y}_2 - \hat{Y}_1 = b_1$$

→ **Θα διαφέρουν κατά μέσο όρο b_1 cm**

Αν έχουμε 2 παιδιά ίδιας ηλικίας (π.χ. 10 ετών) και ο πατέρας του ενός είναι 1.75 m και του άλλου 1.85 m, πόσο αναμένεται να διαφέρουν τα ύψη τους κατά μέσο όρο;

$$\hat{Y}_1 = b_0 + b_1 * 175 + b_2 * 10$$

$$\hat{Y}_2 = b_0 + b_1 * 185 + b_2 * 10$$

$$\hat{Y}_2 - \hat{Y}_1 = b_1(185 - 175)$$

$$\hat{Y}_2 - \hat{Y}_1 = 10 * b_1$$

→ Θα διαφέρουν κατά μέσο όρο **$10*b_1$** cm

Αν έχουμε 2 παιδιά των οποίων οι πατέρες έχουν το <u>ίδιο ύψος</u> (π.χ. 1.75) και η <u>ηλικία τους διαφέρει κατά 2 έτη</u> (π.χ. 10 & 12 ετών), πόσο αναμένεται κατά μέσο όρο να διαφέρουν τα ύψη τους;

$$\hat{Y}_1 = b_0 + b_1 * 175 + b_2 * 10$$

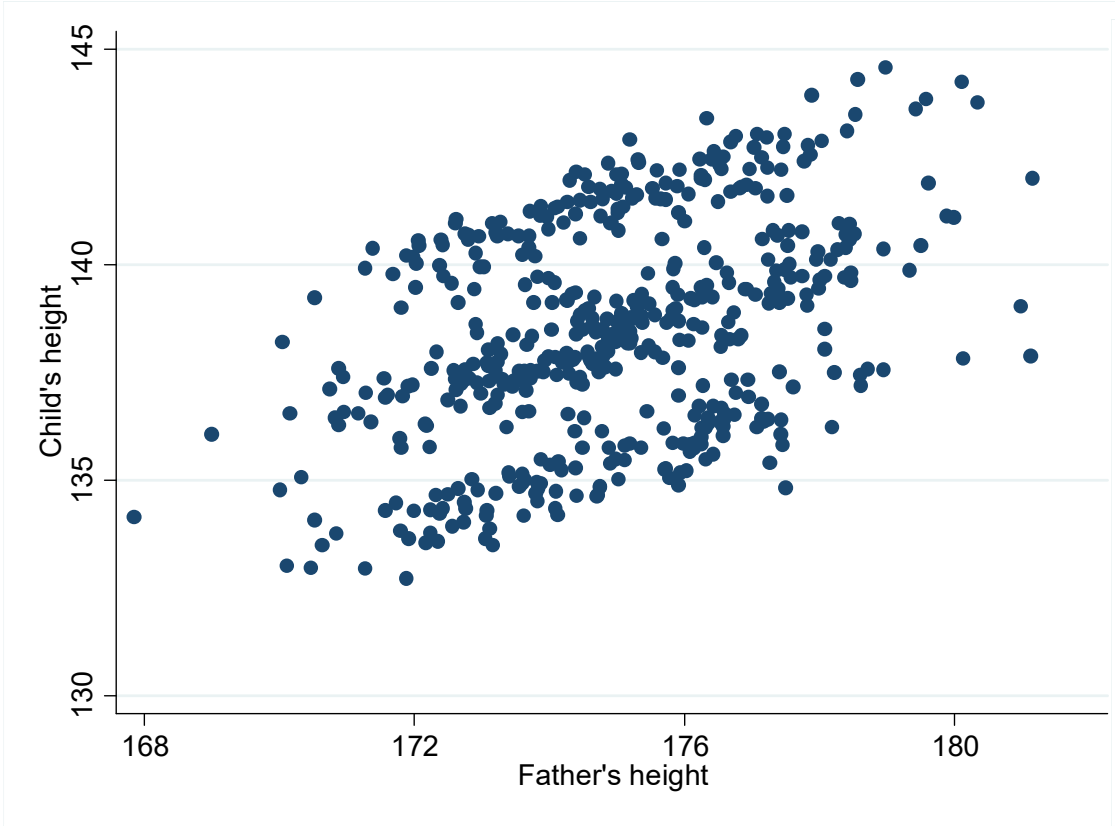
$$\hat{Y}_2 = b_0 + b_1 * 175 + b_2 * 12$$

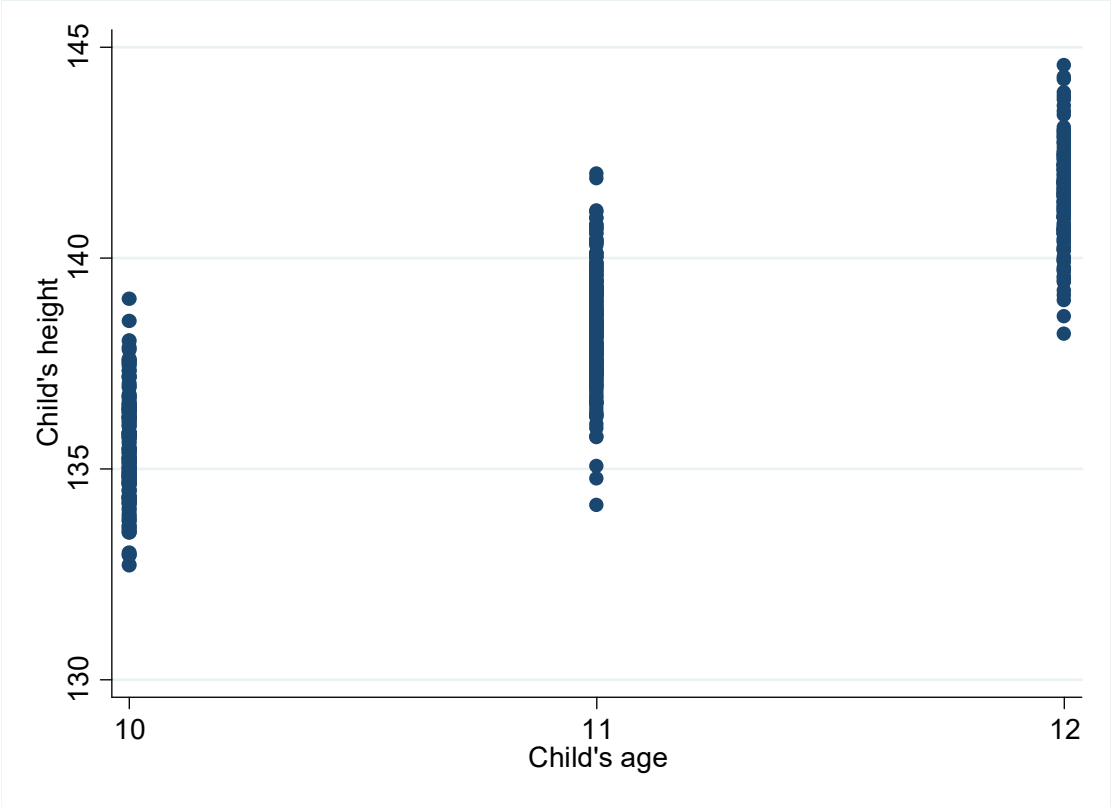
$$\hat{Y}_2 - \hat{Y}_1 = b_2(12 - 10)$$

$$\hat{Y}_2 - \hat{Y}_1 = 2 * b_2$$

→ Θα διαφέρουν κατά μέσο όρο **2*b_2** cm

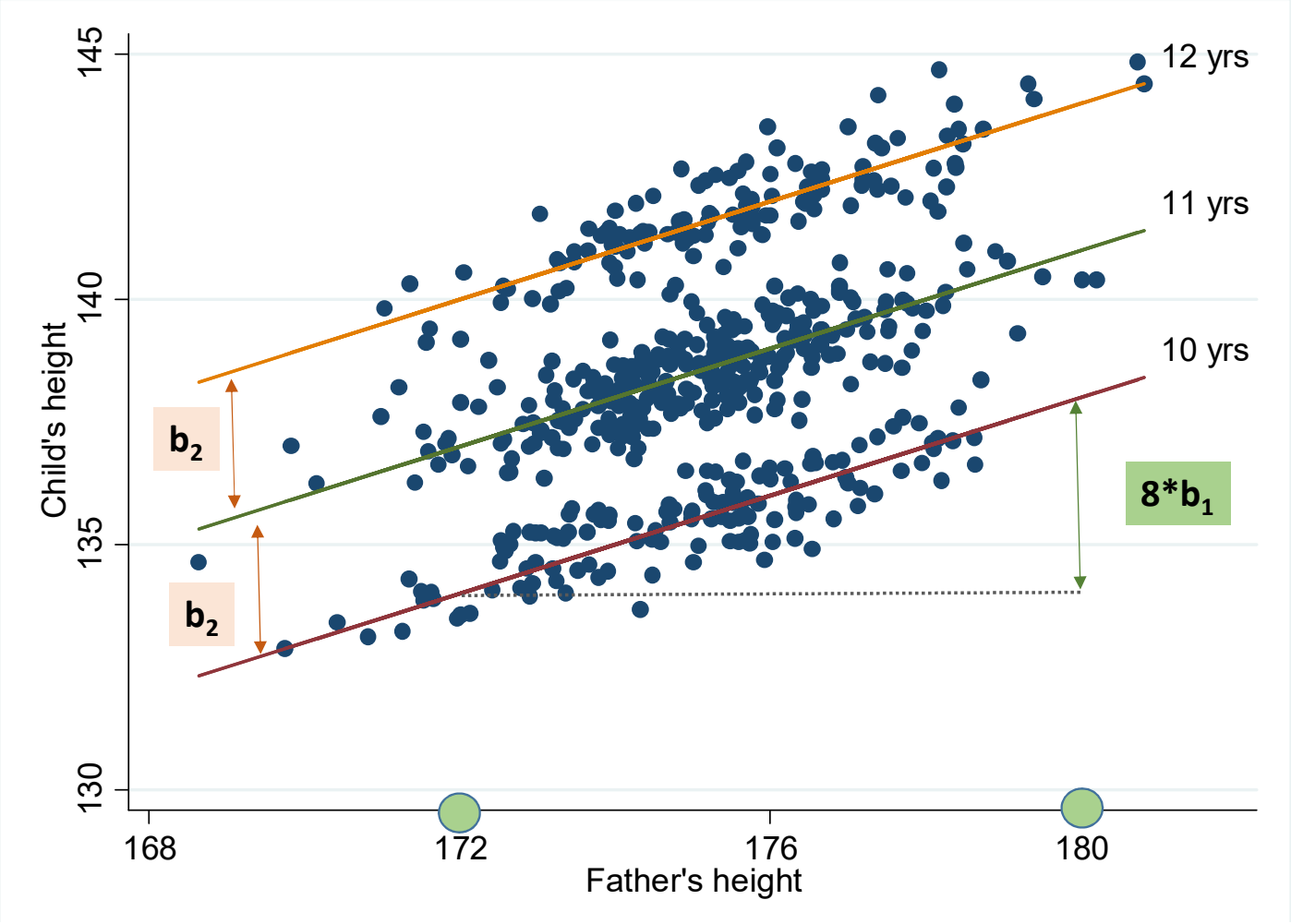
Π.χ. Μελετάμε σε 500 παιδιά το ανάστημά τους σε σχέση με το ανάστημα πατέρα και την ηλικία τους





Από εφαρμογή πολλαπλής γραμμικής παλινδρόμησης:

$$\hat{Y} = b_0 + b_1 \cdot \text{ύψος πατέρα} + \text{b}_2 \cdot \text{ηλικία παιδιού}$$



Υπόλοιπα

Τα παρατηρηθέντα υπόλοιπα e_i (**διαφορά πραγματικής τιμής από εκτιμώμενη τιμή**) υπολογίζονται αντίστοιχα με την απλή γραμμική εξάρτηση:

$$e_i = (Y_i - \hat{Y}_i) = (Y_i - (b_o + b_1 X_{1i} + b_2 X_{2i} + \ldots + b_p X_{pi}))$$

Στην πράξη, γιατί είναι χρήσιμη μία τέτοια μεθοδολογία;

Π.χ. Εφαρμογή πολλαπλής γραμμικής εξάρτησης για την εύρεση παραγόντων που σχετίζονται με τα επίπεδα χοληστερόλης στα άτομα

→ ποιες παράμετροι εξηγούν τη μεταβλητότητα των επιπέδων χοληστερόλης μεταξύ των ατόμων;

Γιατί μας ενδιαφέρει να απαντήσουμε σε τέτοιου είδους ερωτήματα;

- Βαθύτερη κατανόηση των μηχανισμών που συντελούν σε αυξημένα επίπεδα χοληστερόλης
- Αν βρεθεί ένα μοντέλο που εξηγεί ικανοποιητικά τη μεταβλητότητα, μπορεί να χρησιμοποιηθεί για πρόβλεψη
- Αν εντοπιστούν παράγοντες που σχετίζονται με τα επίπεδα χοληστερόλης οι οποίοι μπορούν να τροποποιηθούν (π.χ. διατροφή, βάρος κλπ), τα αποτελέσματα μπορεί να αποτελέσουν βάση για συστάσεις-παρεμβάσεις

Παράδειγμα

- Σε μελέτη για τη διερεύνηση της επίδρασης του μολύβδου στην σωματομετρική ανάπτυξη των παιδιών, μελετήθηκαν παιδιά σχολικής ηλικίας από τρεις περιοχές:
  - Λαύριο, Ελευσίνα και Λουτράκι

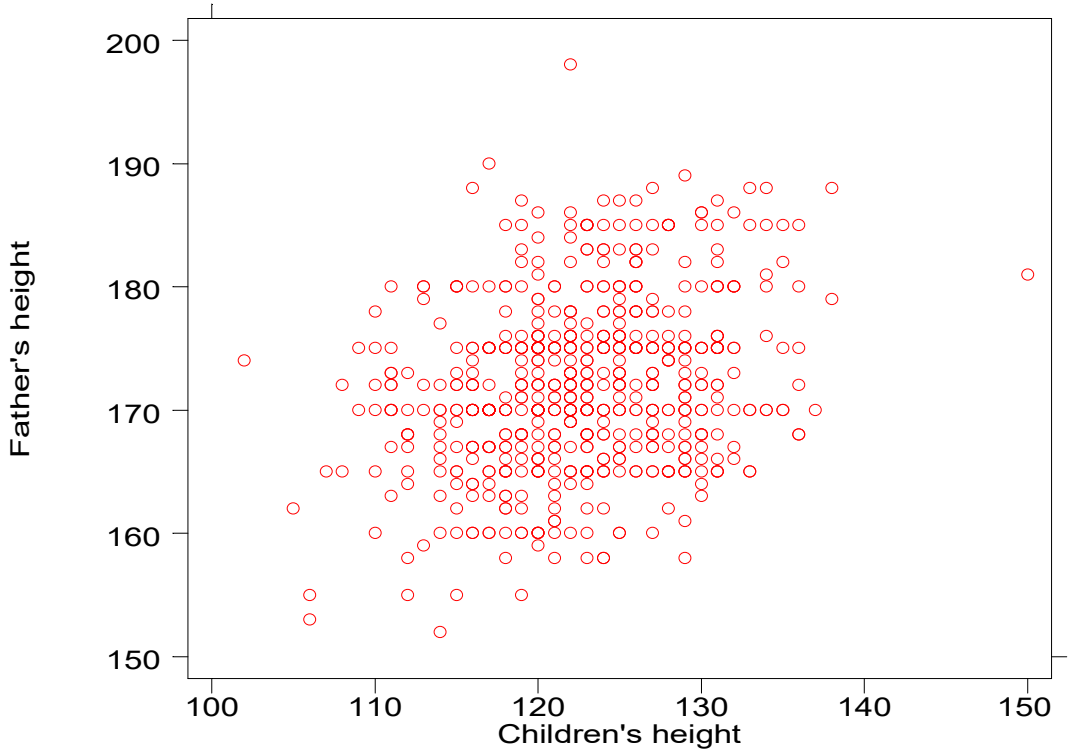
- Το συνολικό δείγμα αποτελείται από 522 παιδιά, 274 αγόρια και 248 κορίτσια ηλικίας 6-9 χρονών. Μέρος των δεδομένων παρουσιάζεται στον πίνακα που ακολουθεί

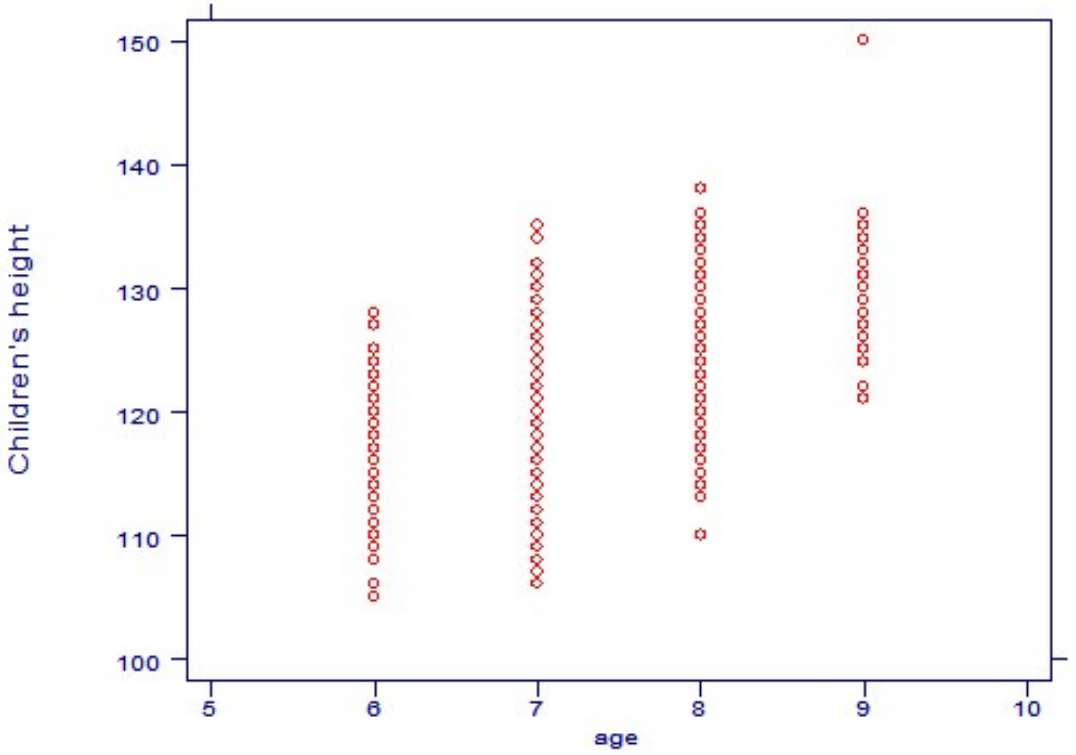
(Kafourou et al, Archives of Environmental health, 1997; 52: 377- 383).

Κωδικός	Πόλη	Ηλικία (έτη)	Ανάστημα πατέρα (cm)	Μόλυβδος (μg/mL)	Ανάστημα παιδιού (cm)
353	2	8	172	23.42	116
419	2	.	165	51.17	107
19	1	8	152	.	114
26	1	7	177	5.94	122
506	2	7	155	20.21	119
683	3	8	170	4.16	117
612	3	7	164	9.78	112
97	1	8	164	.	121
504	2	7	172	17.29	113
469	2	9	170	26.98	124
498	2	7	160	13.24	110
565	2	8	168	22.94	123
140	1	8	162	2.86	115
374	2	6	155	26.59	112
673	3	6	172	5.69	119
644	3	8	167	11.87	123
507	2	8	177	10.19	125
711	3	7	165	4.15	124

Οπου υπάρχει . υποδεικνύει ελλείπουσα τιμή. Για την πόλη: 1 σημαίνει Λουτράκι, 2 Λαύριο και 3 Ελευσίνα.

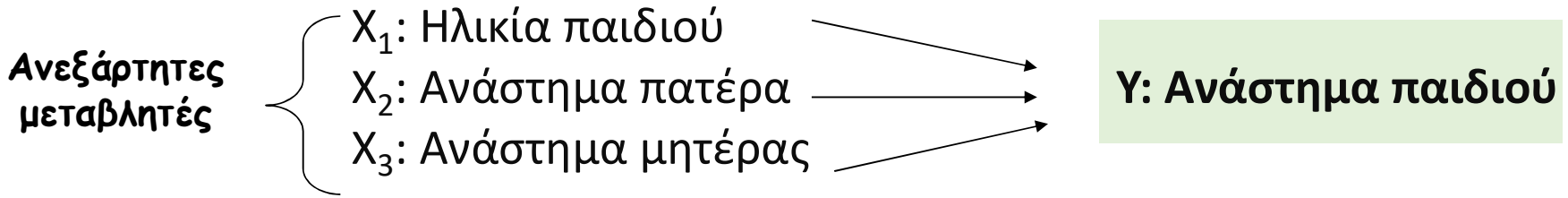
Διάγραμμα συσχέτισης (ή στικτόγραμμα) του αναστήματος του παιδιού με (Α) το ανάστημα του πατέρα και (Β) την ηλικία του παιδιού





Παράδειγμα

- Ας υποθέσουμε ότι θέλουμε να διερευνήσουμε την εξάρτηση του **αναστήματος παιδιού (Υ)** από την ηλικία του (X_1), το ανάστημα του πατέρα του (X_2) και από το ανάστημα της μητέρας του (X_3)



- Με χρήση της κατάλληλης μεθοδολογίας στα δεδομένα, εκτιμούμε την ευθεία πολλαπλής γραμμικής εξάρτησης:

$$\hat{Y} = 20,50 + 4,56 * X_1 + 0,20 * X_2 + 0,21 * X_3$$

Ηλικία παιδιού Ανάστημα πατέρα Ανάστημα μητέρας

Παράδειγμα

$$\hat{Y} = 20,50 + 4,56 * X_1 + 0,20 * X_2 + 0,21 * X_3$$

Ηλικία παιδιού — Ανάστημα πατέρα — Ανάστημα μητέρας

- Για κάθε ένα έτος αύξησης της ηλικίας του παιδιού παρατηρείται **μέση** αύξηση του αναστήματος κατά 4.56 cm, **ανεξάρτητα από το ανάστημα των γονέων**

- Το παιδί ενός πατέρα υψηλότερου κατά 10cm από τον πατέρα ενός άλλου παιδιού, είναι **κατά μέσο όρο** υψηλότερο κατά 2 cm (=0.20*10), **ανεξάρτητα από την ηλικία του και το ανάστημα της μητέρας του**.

Στατιστική αξιολόγηση συντελεστών μερικής εξάρτησης

- Παρόμοια με την απλή γραμμική εξάρτηση, αξιολογούμε το πηλίκο:

$$t = \frac{|b_j|}{SE_{b_j}}$$ όπου j=1,2,…,p μεταβλητές

Στην πολλαπλή γραμμική εξάρτηση όπου έχουμε χρησιμοποιήσει p ανεξάρτητες μεταβλητές $X_1, X_2, .. X_p$, αυτό το πηλίκο ακολουθεί την **t κατανομή στους n-p-1 βαθμούς ελευθερίας** όπου

n: αριθμός παρατηρήσεων

p: αριθμός ανεξάρτητων μεταβλητών → άρα p+1 εκτιμώμενοι συντελεστές (σταθερά **+** p συντελεστές για τις p μεταβλητές)

Στατιστική αξιολόγηση συντελεστών μερικής εξάρτησης

Ελέγχουμε την τιμή $\frac{|b_j|}{SE_{b_j}}$ στον πίνακα με τις οριακές τιμές [5% επίπεδο σημαντικότητας σε n-p-1 ΒΕ]

Αν $\frac{|b_j|}{SE_{b_j}}$ ≥οριακή τιμή → απορρίπτω H_0 και συμπεραίνω ότι η σχέση της X_j με την εξαρτημένη μεταβλητή **είναι** στατιστικά σημαντική

Αν $\frac{|b_j|}{SE_{b_j}}$ <οριακή τιμή → η σχέση της X_j με την εξαρτημένη μεταβλητή **δεν είναι** στατιστικά σημαντική

Αποτελέσματα πολλαπλής εξάρτησης του αναστήματος παιδιών από 3 μεταβλητές

Ανεξάρτητη μεταβλητή	**b**	**SE_b**
Ηλικία παιδιού (έτη)	4,56	0,25
Ανάστημα πατέρα (cm)	0,20	0,03
Ανάστημα μητέρας (cm)	0,21	0,03

- Ερμηνεία π.χ. για ηλικία παιδιού:

  Για κάθε ένα έτος αύξησης της ηλικίας του παιδιού παρατηρείται **μέση** αύξηση του αναστήματος κατά 4.56 cm, **ανεξάρτητα από το ανάστημα των γονέων**

- Εναλλακτικά

  Ένα παιδί που είναι ένα έτος μεγαλύτερο από ένα άλλο και οι γονείς τους έχουν ίδιο ανάστημα, αναμένεται να είναι 4,56 cm ψηλότερο κατά μέσο όρο

ΕΝΑΙ ΑΥΤΗ ΣΧΕΣΗ ΣΤΑΤΙΣΤΙΚΑ ΣΗΜΑΝΤΙΚΗ;

Ανεξάρτητη μεταβλητή	**t-test (b/SE_b)**	**B.E**
Ηλικία παιδιού (έτη)	17,93	518

Επίπεδο σημαντικότητας

BE	0.1	0.05	0.01
518	1.65	1.96	2.576

$b_i/SE_{bi}=17.93>2.576$

Στατιστική σημαντική σχέση στο 1% επίπεδο σημαντικότητας ($p<0.01$)

Επίπεδο σημαντικότητας

	0.1	0.05	0.02	0.01
df =1	6.314	12.706	31.821	63.656
2	2.920	4.303	6.965	9.925
3	2.353	3.182	4.541	5.841
4	2.132	2.776	3.747	4.604
5	2.015	2.571	3.365	4.032
6	1.943	2.447	3.143	3.707
7	1.895	2.365	2.998	3.499
8	1.860	2.306	2.896	3.355
9	1.833	2.262	2.821	3.250
10	1.812	2.228	2.764	3.169
11	1.796	2.201	2.718	3.106
12	1.782	2.179	2.681	3.055
13	1.771	2.160	2.650	3.012
14	1.761	2.145	2.624	2.977
15	1.753	2.131	2.602	2.947
16	1.746	2.120	2.583	2.921
17	1.740	2.110	2.567	2.898
18	1.734	2.101	2.552	2.878
19	1.729	2.093	2.539	2.861
20	1.725	2.086	2.528	2.845
21	1.721	2.080	2.518	2.831
22	1.717	2.074	2.508	2.819
23	1.714	2.069	2.500	2.807
24	1.711	2.064	2.492	2.797
25	1.708	2.060	2.485	2.787
26	1.706	2.056	2.479	2.779
27	1.703	2.052	2.473	2.771
28	1.701	2.048	2.467	2.763
29	1.699	2.045	2.462	2.756
30	1.697	2.042	2.457	2.750
60	1.671	2.000	2.390	2.660
120	1.658	1.980	2.358	2.617
∞	1.645	1.960	2.326	2.576

Ανεξάρτητη μεταβλητή	**b**	**SE_b**	**t-test (b/SE_b)**	**B.E**	P
Ηλικία παιδιού (έτη)	4,56	0,25	17,93	522-4 =518	$<10^{-6}$
Ανάστημα πατέρα (cm)	0,20	0,03	6,76	518	$<10^{-6}$
Ανάστημα μητέρας (cm)	0,21	0,03	6,13	518	$<10^{-6}$

Και οι 3 μεταβλητές είναι στατιστικά σημαντικές

Όρια αξιοπιστίας συντελεστών μερικής εξάρτησης

- Τα **95% CI** υπολογίζονται από τον τύπο:

$$\mathbf{b_j \pm t_{0.05,(n-p-1)} SE(b_j)}$$

Π.χ. στο παράδειγμα, 95% CI για το b_i της ηλικίας παιδιού

$$b_j \pm t_{0.05,(n-p-1)} SE(b_j)$$

$$4{,}56 \pm 1{,}96 * 0{,}25$$

$$4{,}56 \pm 0{,}49$$

4,07 5,05

Με 95% πιθανότητα η μέση αύξηση του αναστήματος για αύξηση της ηλικίας του παιδιού κατά ένα έτος, και ανεξαρτήτως του ύψους των γονέων, βρίσκεται μεταξύ 4,07 και 5,07 cm.

Ποιοτικές ανεξάρτητες μεταβλητές

- Σε ένα μοντέλο πολλαπλής γραμμικής παλινδρόμησης, οι ανεξάρτητες μεταβλητές μπορεί να είναι
  - Ποσοτικές
  - Ποιοτικές

  Π.χ. πως εξαρτάται το ανάστημα παιδιού από την ηλικία του (ποσοτική), το φύλο του (ποιοτική με 2 επίπεδα) και το επάγγελμα του πατέρα (ποιοτική με 3 επίπεδα)

- Πώς εισάγονται οι ποιοτικές μεταβλητές στο μοντέλο και πώς ερμηνεύονται οι συντελεστές μερικής εξάρτησης;

Ποιοτικές μεταβλητές με 2 επίπεδα

Αν η ποιοτική μεταβλητή έχει 2 επίπεδα (π.χ. φύλο):

- Αποφασίζουμε ποια σύγκριση επιθυμούμε να κάνουμε π.χ. άνδρες σε σχέση με γυναίκες ή αντίστροφα
  - Π.χ. αγόρια σε σχέση με κορίτσια →
    κατηγορία αναφοράς: κορίτσια

- Η κατηγορία αναφοράς κωδικοποιείται με 0 και η άλλη κατηγορία με 1 (γενικά η κατηγορία αναφοράς κωδικοποιείται με τη μικρότερη τιμή)

- Ο συντελεστής b για το φύλο εκφράζει πόσο διαφέρουν κατά μέσο όρο τα αγόρια από τα κορίτσια ως προς το ανάστημά τοους (κατηγορία αναφοράς: κορίτσια)

Ποιοτικές μεταβλητές με >2 επίπεδα

Δημιουργία ψευδομεταβλητών (dummy variables- indicator variables)

Παράδειγμα: επάγγελμα πατέρα
1: Ανειδίκευτος / 2: Ειδικευμένος / 3: Πανεπιστημιακής μόρφωσης

Η μεταβλητή έχει 3 επίπεδα → Δημιουργία 3 ψευδομεταβλητών

1. **Ανειδίκευτος (X_1)**
   = 1 αν ο πατέρας είναι ανειδίκευτος
   = 0 άλλο
2. **Ειδικευμένος (X_2)**
   = 1 αν ο πατέρας είναι ειδικευμένος
   = 0 άλλο
3. **Πανεπιστημιακής μόρφωσης (X_3)**
   = 1 αν ο πατέρας έχει πανεπιστημιακή μόρφ.
   = 0 άλλο

Αν γνωρίζουμε τις 2 από αυτές τις μεταβλητές, μπορούμε να συμπεράνουμε την τιμή της τρίτης, π.χ. αν $X_2=0$ και $X_3=0$ → $X_1=1$

Πως εισάγουμε ποιοτικές μεταβλητές στο μοντέλο;

- Στο μοντέλο εισάγονται **Κ-1** ψευδομεταβλητές, όπου Κ ο αριθμός των επιπέδων της ποιοτικής μεταβλητής.
  - **Αυτή που δεν εισέρχεται: κατηγορία αναφοράς (reference category)**

- Εκτιμώνται τα b_j για τη σύγκριση κάθε μίας κατηγορίας προς την κατηγορία αναφοράς, π.χ.

  - ειδικευμένοι σε σχέση με ανειδίκευτους → ο συντελεστής b_2 της X_2

  - πανεπιστ. μόρφωσης σε σχέση με ανειδίκευτους → ο συντελεστής b_3 της X_3

Π.χ. 2 παιδιά, το 1 με πατέρα ανειδίκευτο και το άλλο με πατέρα ειδικευμένο

	X_1 (ref. category)	X_2	X_3
Ανειδικευτος	1	0	0
Ειδικευμενος	0	1	0
Παν. μόρφωση	0	0	1

Παιδί με ανειδίκευτο πατέρα $\hat{Y}_1 = b_0 + b_2 * 0 + b_3 * 0$

Παιδί με ειδικευμένο πατέρα $\hat{Y}_2 = b_0 + b_2 * 1 + b_3 * 0$

Διαφορά στο ανάστημα των 2 παιδιών (ειδικευμένος – ανειδικευτος)

$$\hat{Y}_2 - \hat{Y}_1 = b_2$$

Π.χ. 2 παιδιά, το 1 με πατέρα ανειδίκευτο και το άλλο με παν/κη μόρφωση

	X_1 (ref. category)	X_2	X_3
Ανειδικευτος	1	0	0
Ειδικευμενος	0	1	0
Παν. μόρφωση	0	0	1

Παιδί με ανειδίκευτο πατέρα

$$\hat{Y}_1 = b_0 + b_2 * 0 + b_3 * 0$$

Παιδί με πατέρα με Πανεπιστημιακή μόρφωση

$$\hat{Y}_2 = b_0 + b_2 * 0 + b_3 * 1$$

Διαφορά στο ανάστημα των 2 παιδιών (πανεπιστημιακή – ανειδίκευτος)

$$\hat{Y}_2 - \hat{Y}_1 = b_3$$

Ερμηνεία ψευδομεταβλητών: Παράδειγμα

Ύψος παιδιού (Υ) σε σχέση με επάγγελμα πατέρα

Ανειδίκευτος (X_1) = 1 αν ο πατέρας είναι ανειδίκευτος/ = 0 άλλο

Ειδικευμένος (X_2) = 1 αν ο πατέρας είναι ειδικευμένος/ = 0 άλλο

Πανεπ. μόρφ. (X_3) = 1 αν ο πατέρας έχει πανεπ. μόρφ. / = 0 άλλο

$$\mathbf{Y} = 85 + 2.469\,\mathbf{X}_2 + 2.437^{*}\,\mathbf{X}_3$$

- Κατηγορία αναφοράς: X_1 (ανειδίκευτος)
- Τα παιδιά των ειδικευμένων έχουν **κατά μέσο όρο** 2,469 cm υψηλότερο ανάστημα από τα παιδιά των ανειδίκευτων
- Τα παιδιά με πατέρα πανεπ. μόρφωσης έχουν **κατά μέσο όρο** 2,437 cm υψηλότερο ανάστημα από τα παιδιά των ανειδίκευτων

Ερμηνεία ψευδομεταβλητών

Ύψος παιδιού (Υ) σε σχέση με φύλο παιδιού και επάγγελμα πατέρα

- Επάγγελμα πατέρα:

  X_1 = 1 αν ο πατέρας είναι ανειδίκευτος/ 0 άλλο,

  X_2= 1 αν ο πατέρας είναι ειδικευμένος/ = 0 άλλο

  X_3=1 αν ο πατέρας έχει πανεπ. μόρφ. / = 0 άλλο
- Φύλο παιδιού (X_4): Αγόρι=1, Κορίτσι=2

$$\mathbf{Y} = 90.7 - 2.41^{*}\,\mathbf{X_1} + 0.17\,^{*}\,\mathbf{X_2} - 0.66\,\mathbf{X_4}$$

- Φύλο: κατηγορία αναφοράς→ **αγόρι**
- Επάγγελμα πατέρα: κατηγορία αναφοράς → **X_3 (πανεπ. μόρφωσης)**

Ερμηνεία ψευδομεταβλητών

$$Y = 90.7 - 2.41 * X_1 + 0.17 * X_2 - 0.66 X_4$$

Παράγοντας	b	SE(b)	b/SE(b)
Επάγγελμα πατέρα			
Πανεπ. Μόρφωσης	Κατηγορία αναφοράς		
Ανειδίκευτος	**-2.41**	**0.89**	**-2.71**
Ειδικευμένος	**+0.17**	**0.83**	**0.20**
Φύλο			
Αγόρι	Κατηγορία αναφοράς		
Κορίτσι	**-0.66**	**0.44**	**-1.50**

$t_{0.05,518}=1.96$

Παράγοντας	b	SE(b)	b/SE(b)
Επάγγελμα πατέρα			
Πανεπ. Μόρφωσης	Κατηγορία αναφοράς		
Ανειδίκευτος	**-2.41**	**0.89**	**-2.71**
Ειδικευμένος	**+0.17**	**0.83**	**0.20**
Φύλο			
Αγόρι	**Κατηγορία αναφοράς**		
Κορίτσι	**-0.66**	**0.44**	**-1.50**

$t_{0.05,518}=1.96$

- Τα παιδιά των ανειδίκευτων, **ανεξαρτήτως του φύλου τους**, έχουν κατά μέσο όρο **χαμηλότερο** ανάστημα από τα παιδιά αυτών με πανεπιστημιακή μόρφωση κατά 2,41 cm.
- Το ύψος των παιδιών των ειδικευμένων **ΔΕΝ ΔΙΑΦΕΡΕΙ** σε βαθμό στατιστικά σημαντικό από το ύψος των παιδιών που ο πατεράς τους έχει πανεπιστημιακή μόρφωση.
- Ανεξαρτήτως του επαγγέλματος του πατέρα, τα κορίτσια έχουν **χαμηλότερο** ύψος από τα αγόρια κατά μέσο όρο κατά 0,66 cm **αλλά η διαφορά δεν είναι στατιστικά σημαντική**

Επιλογή των ανεξάρτητων μεταβλητών

- Α. Λόγοι εισαγωγής

  - **Να έχει ιδιαίτερο ενδιαφέρον από μόνη της** (όπως τα επίπεδα μολύβδου στη μελέτη διερεύνησης της σχέσης μολύβδου και σωματομετρικών παραμέτρων των παιδιών)

  - Να αποτελεί **σημαντικό προγνωστικό παράγοντα** και κατά συνέπεια το να συμπεριληφθεί στο μοντέλο να έχει ως αποτέλεσμα τη μείωση της διακύμανσης των υπολοίπων και την αύξηση της συνολικής προγνωστικής αξίας του μοντέλου (όπως η ηλικία του παιδιού στο παράδειγμά μας)

  - **Τον έλεγχο των πιθανών συγχυτικών επιδράσεων** (για παράδειγμα το ανάστημα του πατέρα συσχετίζεται σε βαθμό στατιστικά σημαντικό τόσο με το ύψος του παιδιού όσο και με τα επίπεδα του μολύβδου στα παιδιά)

Επιλογή των ανεξάρτητων μεταβλητών

- **Β. Τρόποι επιλογής**
  - Ο αριθμός των παραμέτρων πρέπει να είναι σαφώς μικρότερος του αριθμού των παρατηρήσεων n
    (το πολύ ίσος με $n/10$ ή $\sqrt{n}$)
    - Το τελικό μοντέλο πολλαπλής γραμμικής παλινδρόμησης πρέπει να είναι επεξηγηματικό και ταυτόχρονα **λιτό**

  - Πως θα περιορίσω τον αριθμό των ανεξάρτητων μεταβλητών;
    - Μπορεί η συνεισφορά κάποιων μεταβλητών στο μοντέλο να μην είναι σημαντική
    - Επιλογή με στατιστικά κριτήρια ή στατιστικές μεθόδους αυτόματης επιλογής μέσω στατιστικών προγραμμάτων, π.χ. forward (αθροιστική), backward (αφαιρετική), stepwise regression (συνδυασμός)

Μειονεκτήματα της αυτόματης επιλογής ανεξάρτητων μεταβλητών

ΕΚΤΟΣ ΥΛΗΣ

- Δύο ανεξάρτητες μεταβλητές συσχετίζονται ισχυρά μεταξύ τους (π.χ. ρ>0.7) **Συγγραμικότητα (collinearity)**
- Οι ποιοτικές μεταβλητές δεν αξιολογούνται συνολικά
- Μεταβλητές που αποτελούν γνωστούς συγχυτικούς παράγοντες μπορεί με καθαρά στατιστικά κριτήρια να μην εισαχθούν στο μοντέλο, όμως με επιδημιολογικά κριτήρια θα έπρεπε να εισαχθούν στο τελικό μοντέλο

→**Η επιλογή των ανεξάρτητων μεταβλητών δεν θα πρέπει να στηρίζεται αποκλειστικά σε στατιστικά κριτήρια, αλλά σε ένα συνδυασμό στατιστικών κριτηρίων και προϋπάρχουσας γνώσης**

Είναι καλό το μοντέλο;
→Συντελεστής πολλαπλής συσχέτισης R^2

- R^2: Το ποσοστό της μεταβλητότητας της εξαρτημένης μεταβλητής που ερμηνεύεται από τις ανεξάρτητες μεταβλητές (μπορεί να αποδοθεί στις ανεξάρτητες μεταβλητές)

  - Εννοιολογικά αντίστοιχος του συνηθισμένου συντελεστή συσχέτισης δύο μεταβλητών - Υπολογίζεται όταν έχουμε περισσότερες από δύο μεταβλητές

  - **Παίρνει τιμές από 0 (το μοντέλο δεν εξηγεί καθόλου τη μεταβλητότητα) έως 1 (το μοντέλο εξηγεί 100% τη μεταβλητότητα)**

  - R^2 αυξάνεται κάθε φορά που προστίθεται μία μεταβλητή στο μοντέλο, ανεξάρτητα από το πόσο σημαντική είναι αυτή η μεταβλητή → Προσαρμοσμένο (adjusted) R^2: **ΔΕΝ ΑΥΞΑΝΕΙ** πάντα με τη προσθήκη νέας μεταβλητής στο μοντέλο

Συντελεστής πολλαπλής συσχέτισης

- Παράδειγμα εξάρτησης αναστήματος παιδιών από ύψος πατέρα, μητέρας και ηλικία παιδιού: **R^2=0.43**

  →**Ερμηνεία:** Το 43% της μεταβλητότητας του αναστήματος των παιδιών μπορεί να ερμηνευτεί από τις 3 ανεξάρτητες μεταβλητές

- Προσαρμοσμένο **R^2adjusted=0.42**

- Αν προσθέσω στο μοντέλο και τα επίπεδα διαστολικής αρτηριακής πίεσης: **R^2adjusted=0.42**

  → αυτός ο παράγοντας δε βελτιώνει το ποσοστό μεταβλητότητας που εξηγεί το μοντέλο

Multiple linear regression model on the impact of lead and other variables on children's height in 3 cities in Greece

Table 3.—Final Regression Model for Height

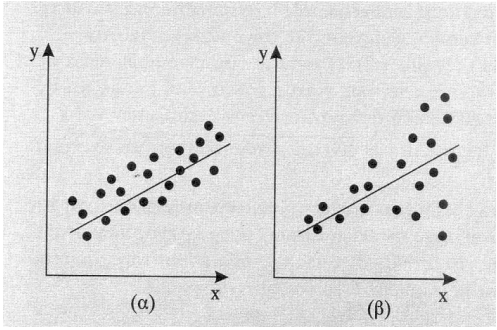
Independent variables	Regression coefficient (*b*)	*SE* (*b*)	*t* ratio	*p*
Lead (μg/dl)	–0.086	0.037	–2.34	.020
Sex				
Females versus males	–0.616	0.426	–1.45	.149
Father's height (cm)	0.233	0.032	7.27	.000
Father's job				
Skilled professional versus unskilled	2.038	0.487	4.18	.000
Age (y)	4.654	0.275	16.93	.000
Hb (g/dl)	0.736	0.235	3.14	.002
City*				
Lavrion	1.179	0.719	1.64	.102
Elefsina	–0.214	0.612	–0.35	.726
Intercept	36.280	6.607	5.49	.000

*Reference category: Loutraki; R^2 = .45.

Kafourou et al, 1997

Έλεγχος των προϋποθέσεων της γραμμικής εξάρτησης

1. Η διασπορά των σημείων πάνω και κάτω από την ευθεία να είναι περίπου ομοιόμορφη σε όλο το μήκος της γραμμής.



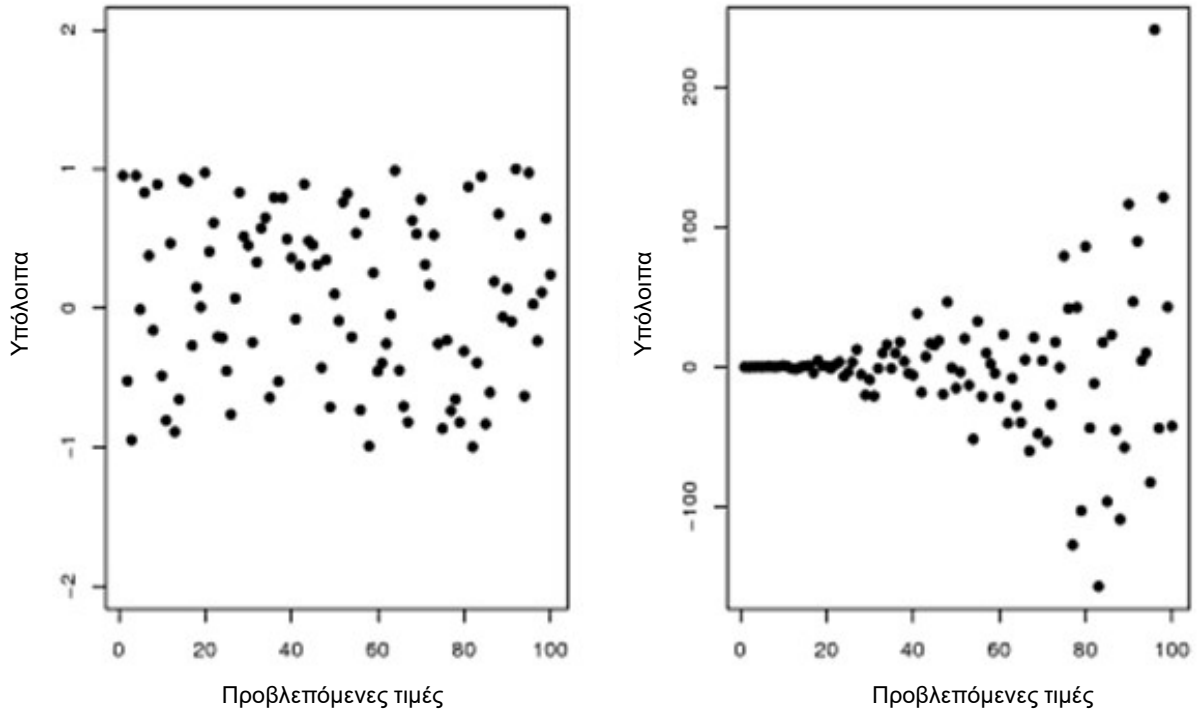
2. Η κατανομή συχνοτήτων της εξαρτημένης Υ, οι οποίες αντιστοιχούν σε ορισμένη τιμή της Χ, πρέπει να είναι κατά προσέγγιση κανονική

3. Γραμμικότητα

→ Ο έλεγχος των προϋποθέσεων της γραμμικής εξάρτησης μπορεί να γίνει με τον έλεγχο των παρατηρηθέντων **υπολοίπων** (αφού δηλαδή την εφαρμόσω)

Έλεγχος των προϋποθέσεων της γραμμικής εξάρτησης: Ομοσκεδαστικότητα (homoscedasticity)

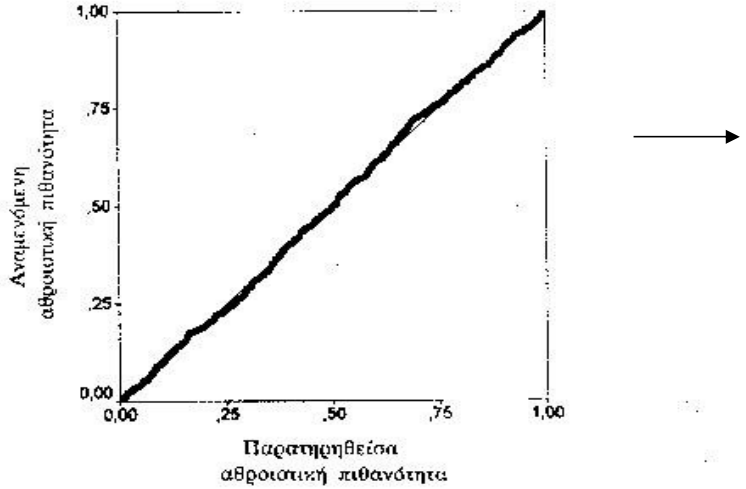
Διάγραμμα υπολοίπων του μοντέλου πολλαπλής γραμμικής εξάρτησης (e_i) προς τις προβλεπόμενες από το μοντέλο τιμές ($\hat{Y}$)



Έλεγχος των προϋποθέσεων της γραμμικής εξάρτησης: Κανονικότητα

- **Διάγραμμα κανονικής κατανομής (normal plot)**

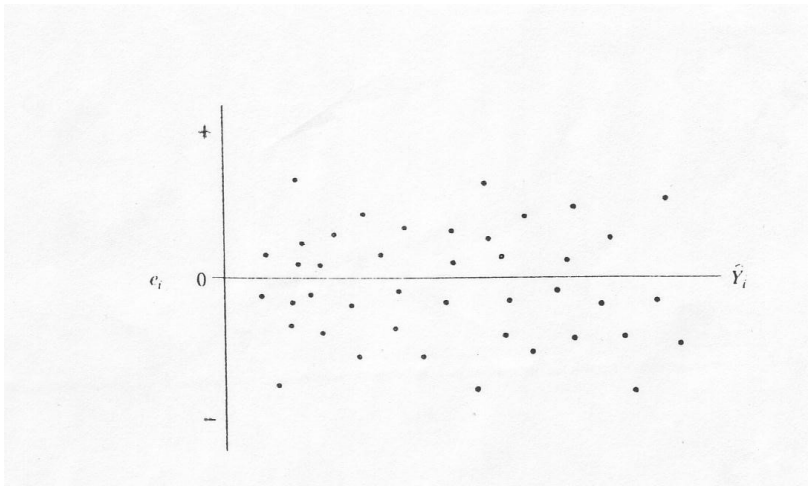
  Γράφημα κανονικής κατανομής των υπολοίπων του μοντέλου πολλαπλής γραμμικής εξάρτησης του αναστήματος των παιδιών από την ηλικία τους και το ύψος των γονέων τους

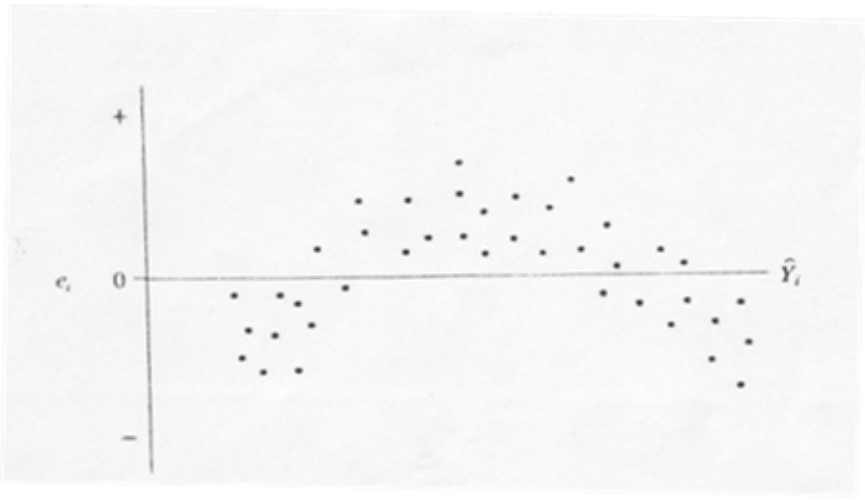


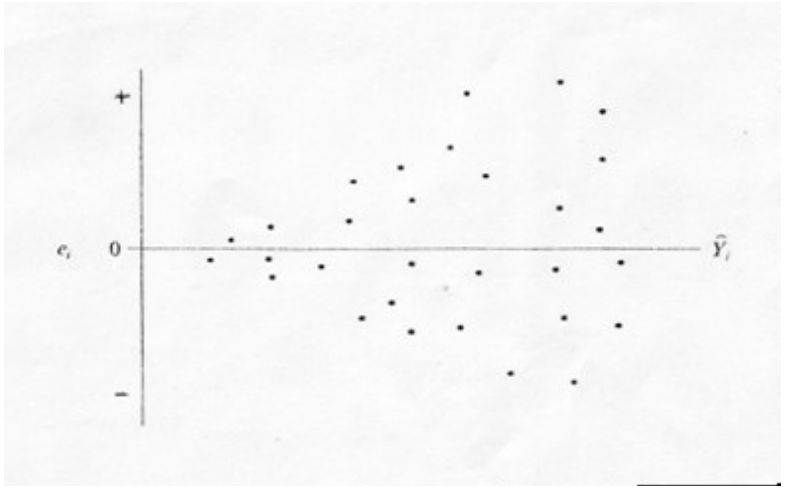
Σχημα 5.1 Γράφημα κανονικής κατανομής των υπολοίπων της πολλαπλής γραμμικής εξάρτησης του πίνακα 5.1

Τα σημεία θα πρέπει να σχηματίζουν μία ευθεία γραμμή. Αν όχι, ένδειξη μη κανονικότητας

Διαγνωστικά διαγράμματα: Διάγραμμα υπολοίπων εξάρτησης προς αναμενόμενες τιμές







Πηγή: Rawlings J.O., Applied regression analysis, 1988.

Παράδειγμα

multiple linear regression analysis to investigate the effects of height (cm), weight (kg) and sex (0 = boy, 1 = girl) on SBP (mm Hg) in these children.

Variable	Parameter estimate	Standard error	95% CI for parameter	Test statistic	*P*-value
Intercept	79.4395	17.1182	(45.89 to 112.99)	4.6406	0.0001
Height	–0.0310	0.1717	(–0.37 to 0.31)	–0.1807	0.8570
Weight	1.1795	0.2614	(0.67 to 1.69)	4.5123	0.0001
Sex	4.2295	1.6105	(1.07 to 7.39)	2.6261	0.0101

Petrie & Sabin

Διερεύνηση των προϋποθέσεων

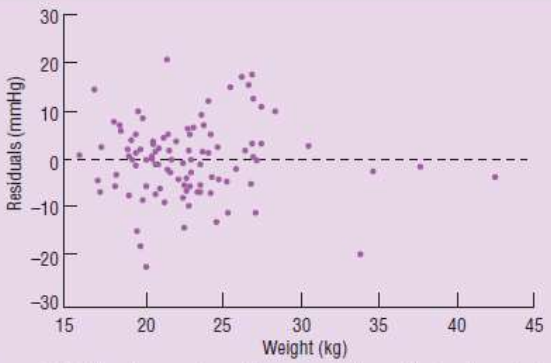


Figure 29.1 There is no systematic pattern to the residuals when plotted against weight. (Note that, similarly to Fig. 28.2, a plot of the residuals from this model against height also shows no systematic pattern.)

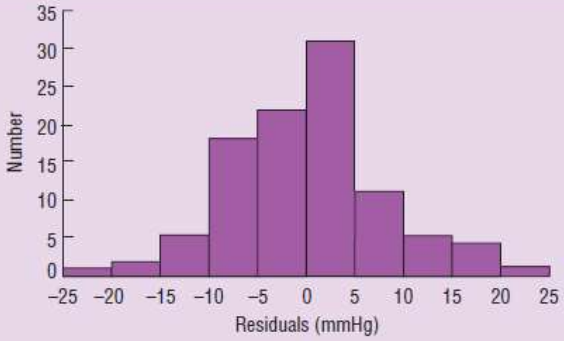


Figure 29.2 The distribution of the residuals is approximately Normal and the variance is slightly less than that from the simple regression model (Chapter 28), reflecting the improved fit of the multiple linear regression model over the simple model.

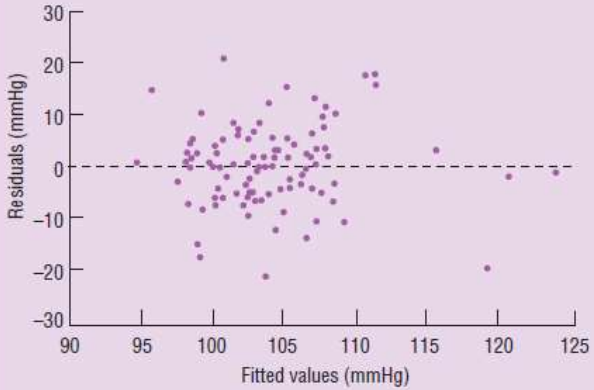


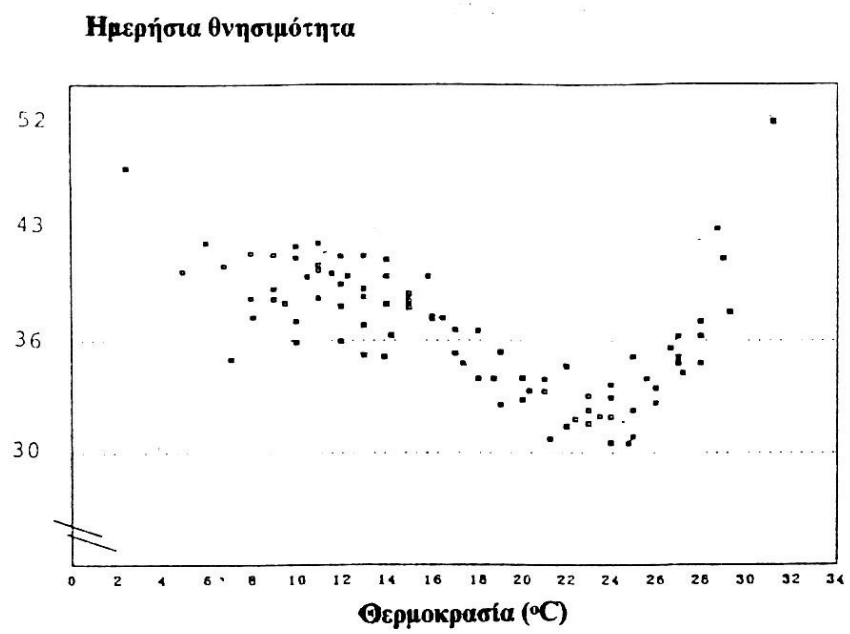
Figure 29.3 As with the univariable model, there is no tendency for the residuals to increase or decrease systematically with fitted values. Hence the constant variance assumption is satisfied.

Figure 29.4 The distribution of the residuals is similar in boys and girls, suggesting that the model fits equally well in the two groups.

Petrie & Sabin

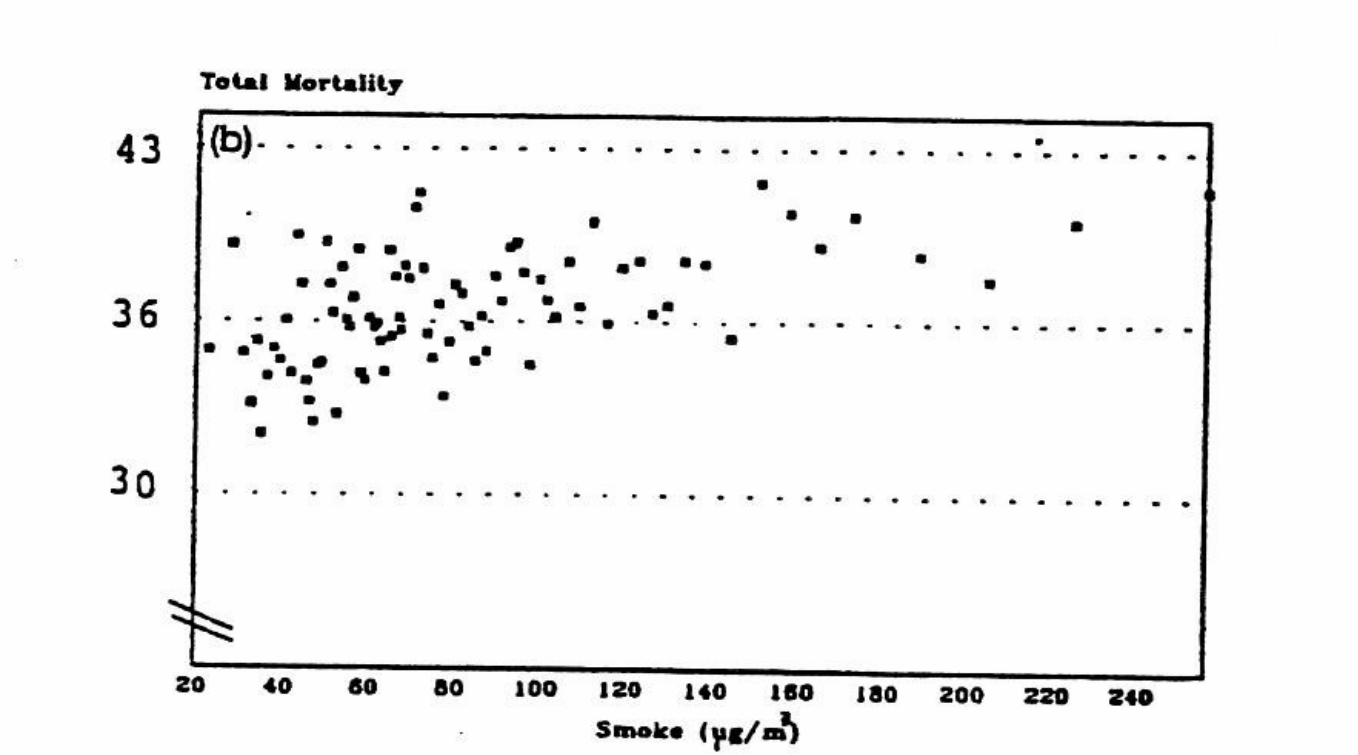
Υπόθεση γραμμικότητας

- Σχέση μεταξύ ημερήσιας θνησιμότητας και ημερήσιας θερμοκρασίας (μετά τον έλεγχο για διαχρονικές και εποχιακές τάσεις και ημέρα της εβδομάδας)



Πρόβλεψη των τιμών του Υ για δοθείσες τιμές του Χ

- Να **αποφεύγονται** οι προβλέψεις για τιμές της Χ **εκτός του εύρους των τιμών της στο δείγμα (extrapolation)**



Πηγή: Touloumi et al, Intern J Epidemiol, Vol.23, No5