Συσχέτιση Παλινδρόμηση

Ανάλυση συνεχών μεταβλητών

Γεωργία Σαλαντή

Λέκτορας

Εργαστήριο υγιεινής και Επιδημιολογίας

Περιεχόμενα

- Συσχέτιση μεταξύ δύο συνεχών μεταβλητών
- Παλινδρόμηση μεταξύ
  - Μίας συνεχούς μεταβλητής
  - Μιας (ή και περισσότερων) μεταβλητών (συνεχών, διχότομων κ.τ.λ)

Εισαγωγικά

- y το αποτέλεσμα (ή δεσμευμένη μεταβλητή) που μας ενδιαφέρει
  - Π.χ. πίεση, τριγλυκερίδια
- x η ανεξάρτητη μεταβλητή
  - Π.χ. ηλικία, φύλο

Ο συντελεστής συσχέτισης κοιτά το πως μεταβάλλεται το y σε σχέση με το x

Τυχαίες μεταβλητές και συσχέτιση

- Όσο πιο ψηλός τόσο πιο βαρύς
- Όσο πιο χαμηλό το βιοτικό επίπεδο, τόσο πιο υψηλή η παιδική θνησιμότητα
- Όσο πιο πολύ το x

  τόσο πιο πολύ/λίγο το y

Συσχέτιση και παλινδρόμηση

- Συσχέτιση: Πόσο έντονα μία αλλαγή στο ένα μέγεθος επηρεάζει το άλλο μέγεθος;
- Παλινδρόμηση: Αν ξέρουμε την τιμή του x μπορούμε να προβλέψουμε το y?

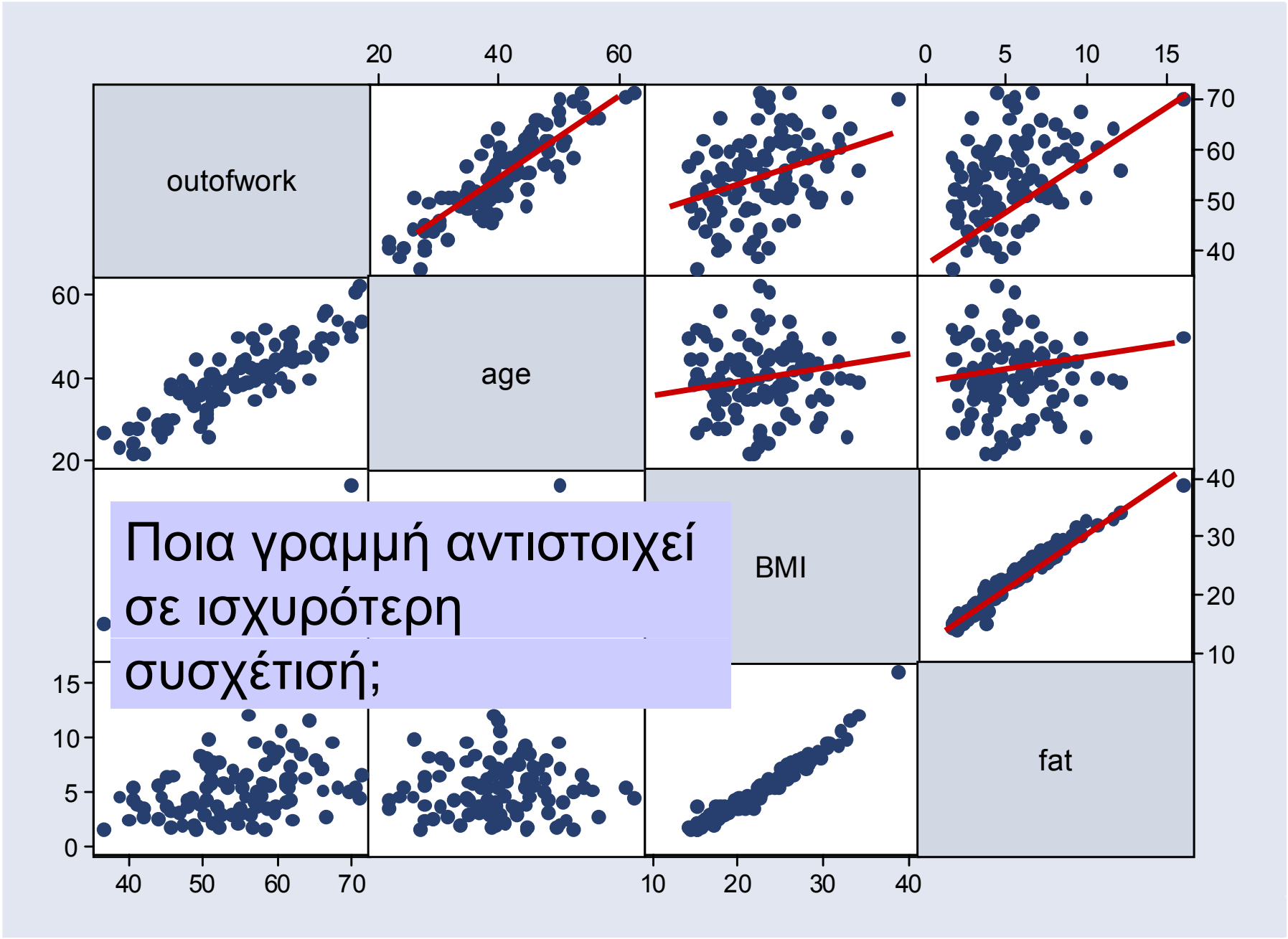
Ξεκινάμε με ένα διάγραμμα διασποράς (scatterplot)

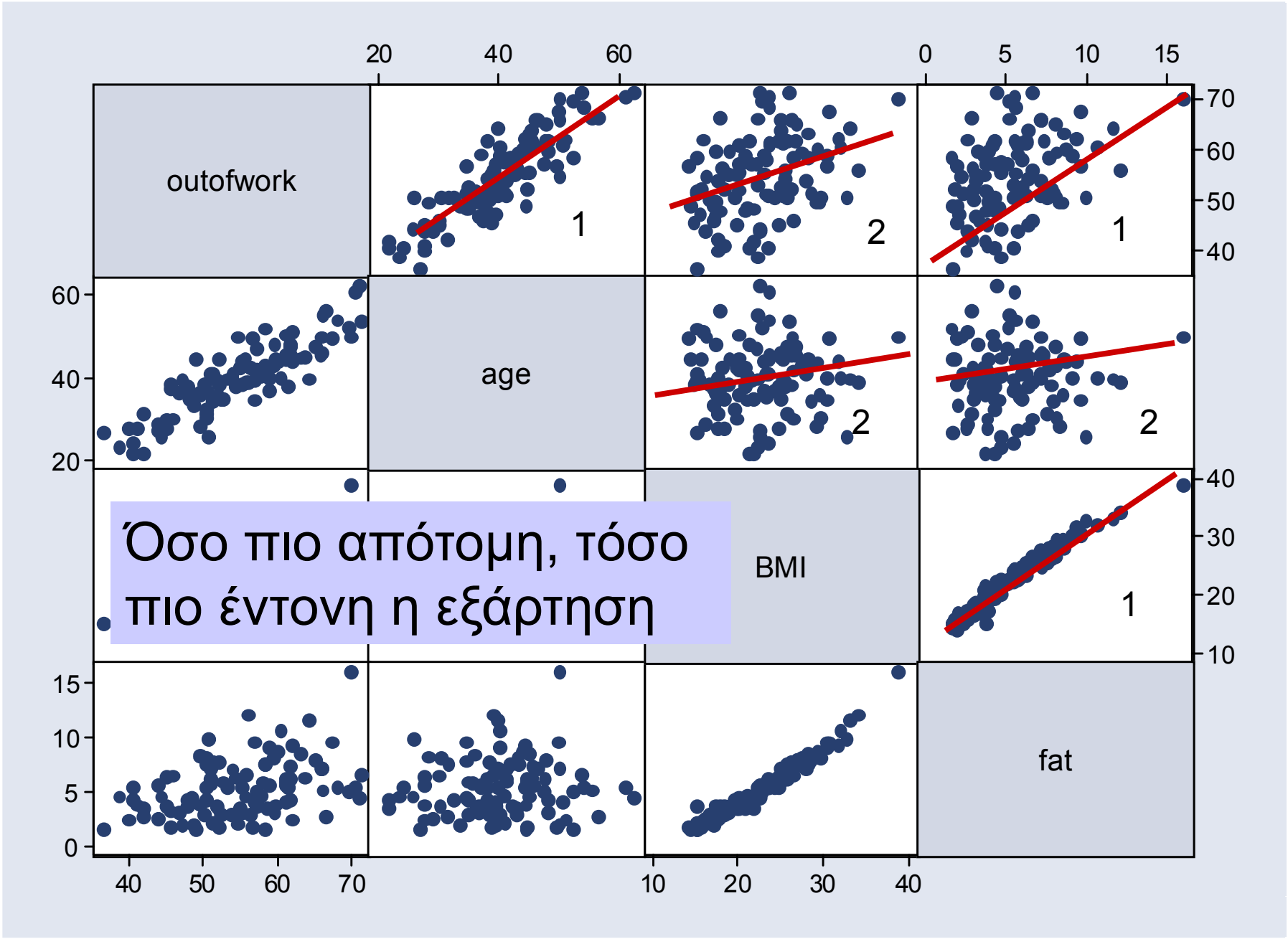
n παρατηρήσεις από το x : $x_1, x_2, \ldots, x_n$

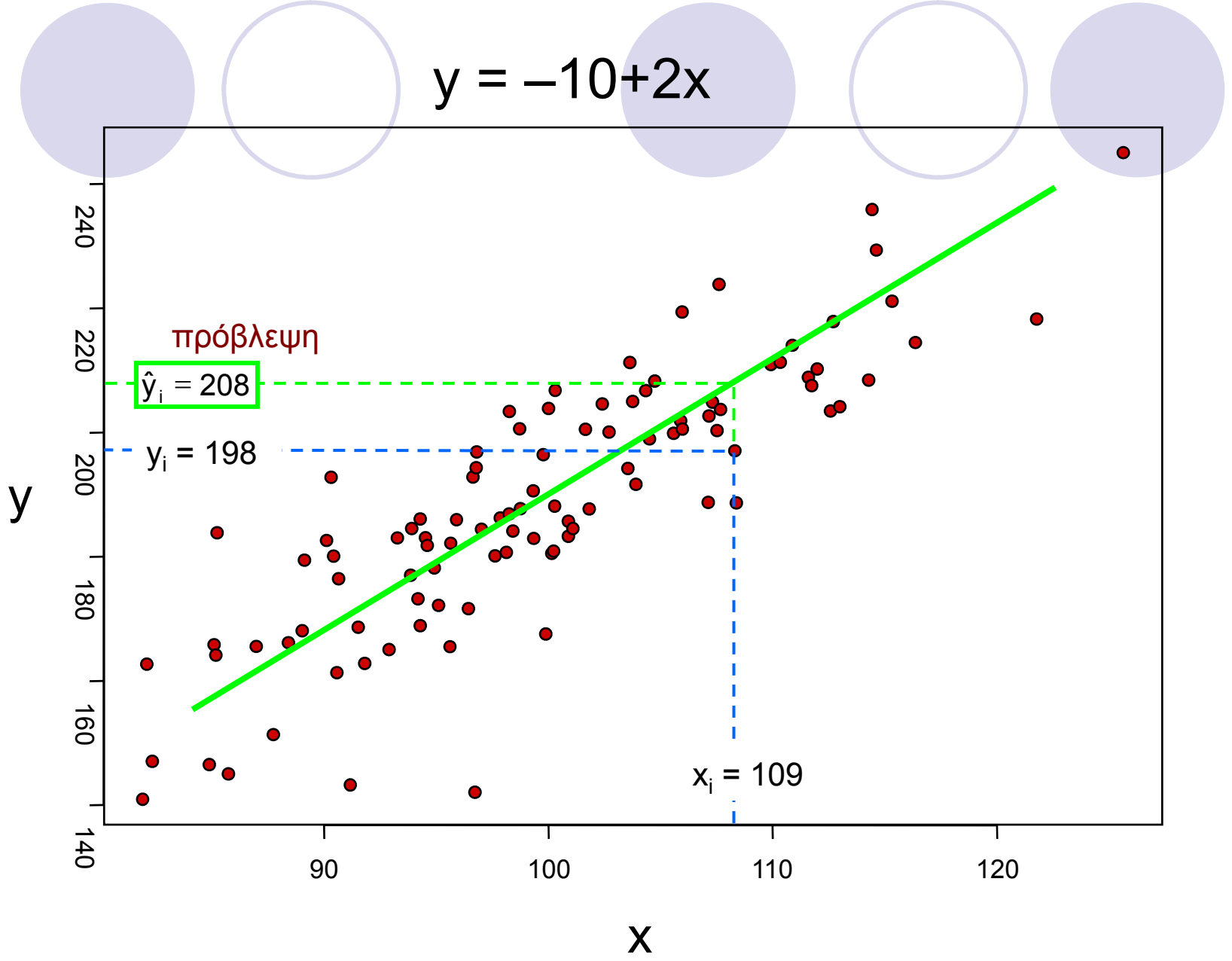
n παρατηρήσεις από το y : $y_1, y_2, \ldots, y_n$

Γραμμική παλινδρόμηση

Γεωργία Σαλαντή


outofwork
age
BMI
fat
Ποια γραμμή αντιστοιχεί σε ισχυρότερη συσχέτισή;


outofwork
age
BMI
fat
Όσο πιο απότομη, τόσο πιο έντονη η εξάρτηση


y = –10+2x
πρόβλεψη
$\hat{y}_i = 208$
$y_i = 198$
$x_i = 109$
y
x
240
220
200
180
160
140
90
100
110
120

Γραμμή παλινδρόμηση

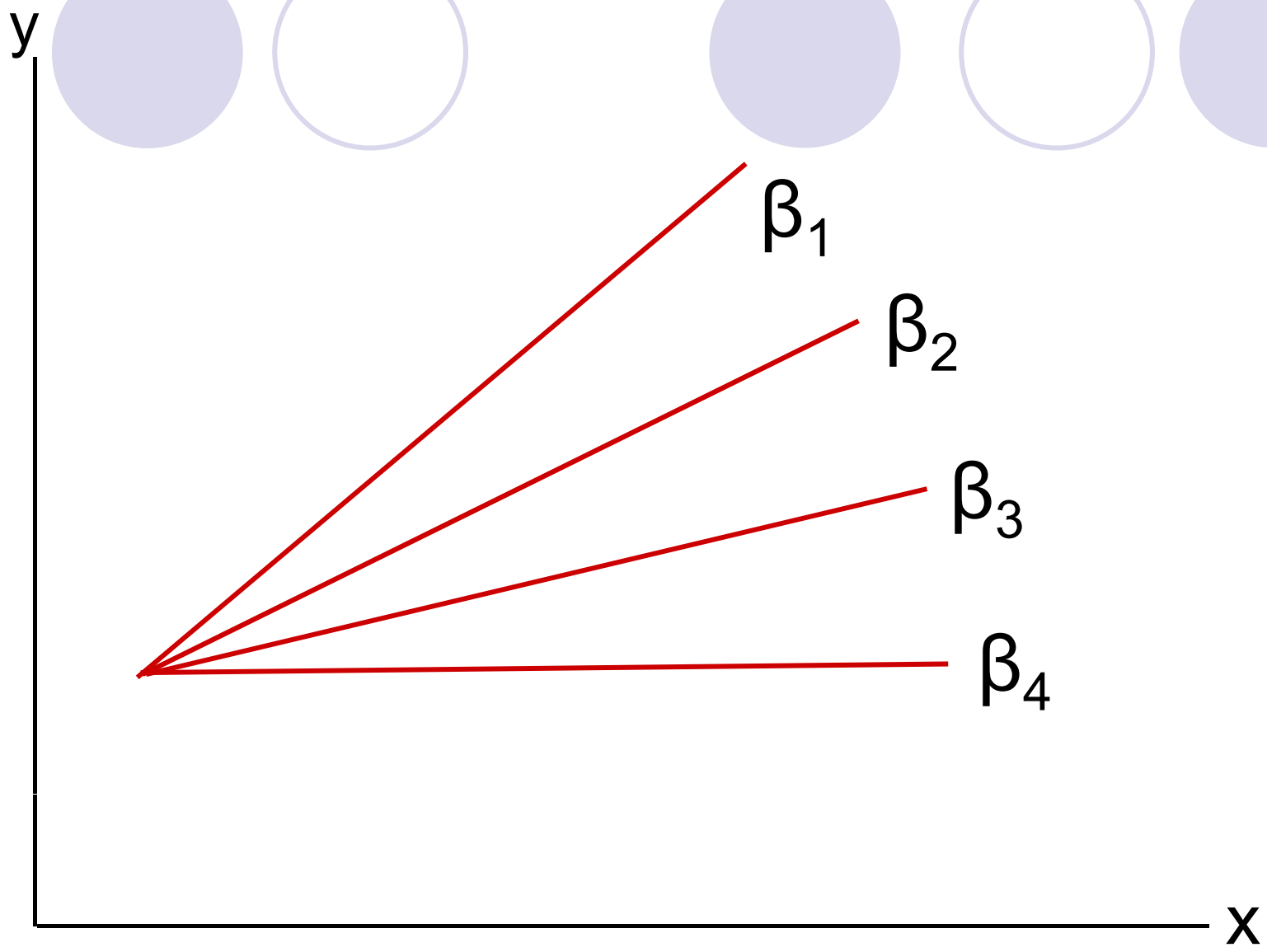
$$\hat{y}_i = \alpha + \beta x_i + \varepsilon_i$$

α : Αρχή (origin)

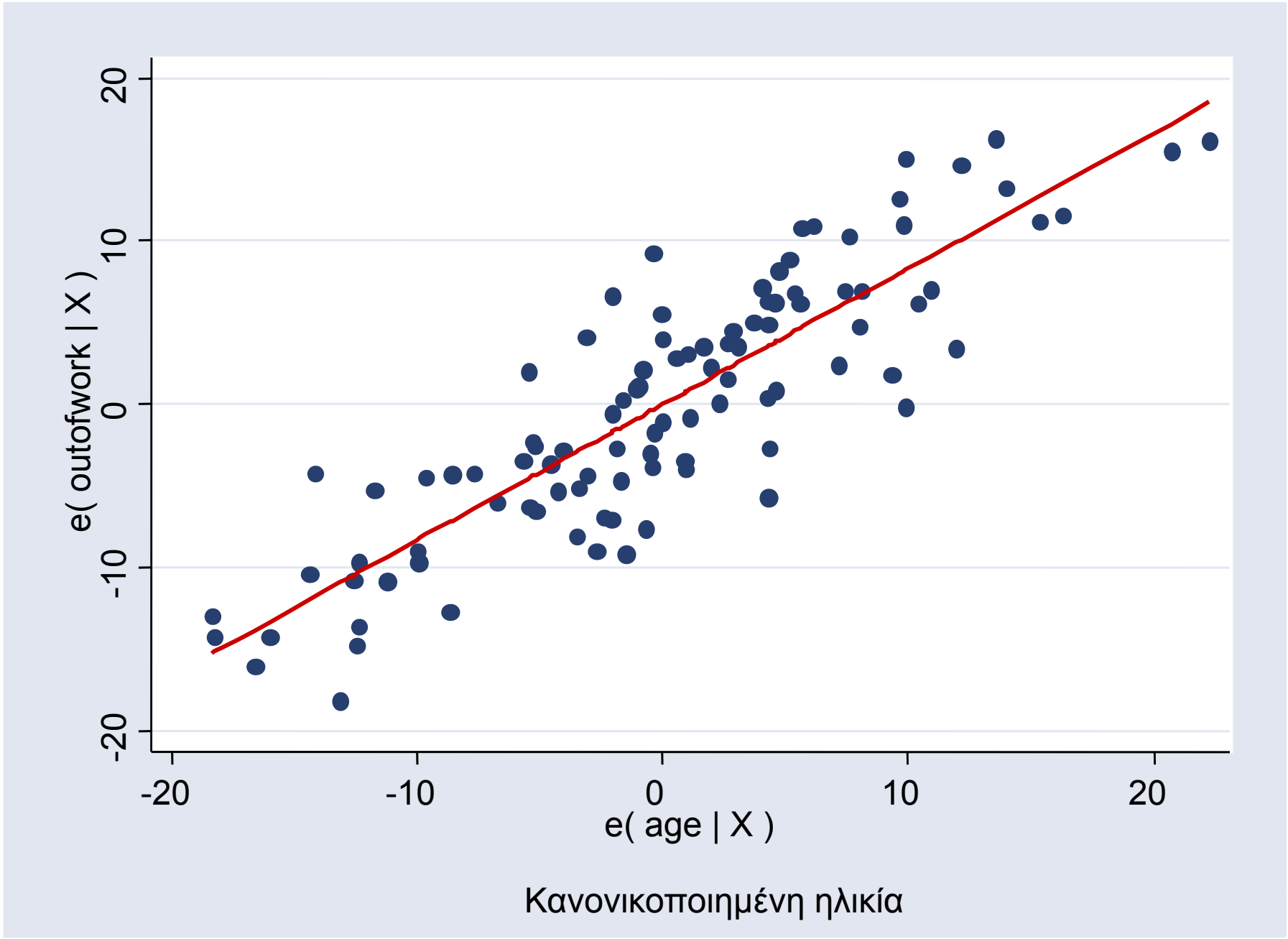
β : Κλίση (slope)

$\varepsilon_i \sim N(0, \sigma^2)$ Τα σφάλματα

Το β δείχνει πόσο απότομη είναι η παλινδρόμηση

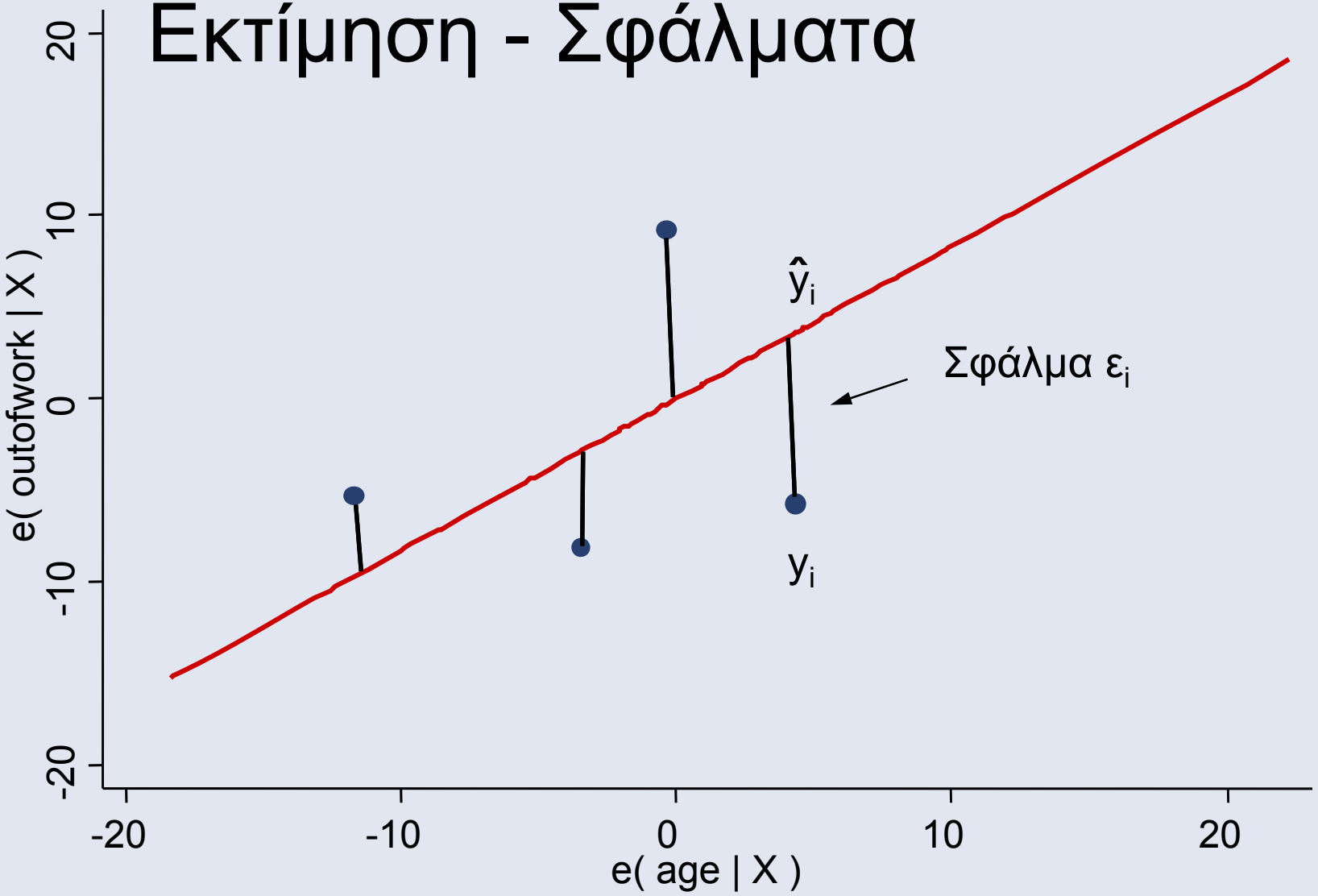


$\beta_1 > \beta_2 > \beta_3 > \beta_4$ και $\beta_4 = 0$



coeff = .83, SE = .05, t = 17.45

Εκτίμηση - Σφάλματα



Θέλουμε να ελαχιστοποιήσουμε τα σφάλματα

Εκτίμηση: πως βρίσκουμε τα α και β

- Με την μέθοδο των ελάχιστων τετραγώνων

Ελαχιστοποιούμε την $$\sum (y_i - \hat{y}_i)^2 = \sum (y_i - \alpha - \beta x_i)^2$$

$$\frac{\partial \sum (y_i - \hat{y}_i)^2}{\partial \beta} = 0 \Rightarrow \quad \beta = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

$$\frac{\partial \sum (y_i - \hat{y}_i)^2}{\partial \alpha} = 0 \Rightarrow \quad \alpha = \bar{y} - \beta \bar{x}$$

Έλεγχος: το F τεστ

H_0 υπόθεση : $\beta = 0$

Μέθοδος: ανάλυση διασποράς (ANOVA)

πηγή	Άθροισμα τετραγ (SoS)	β.ε. df	Μέσοι τετραγώνων	$F_{1,n-2}$
Εξηγείται από την παλ/μηση	$\sum (\hat{y}_i - \bar{y})^2$	1	$MS_{regr} = \frac{SoS_{regr}}{1}$	$\frac{MS_{regr}}{MS_{res}}$
Σφάλματα	$\sum (y_i - \hat{y}_i)^2$	n – 2	$MS_{res} = \frac{SoS_{res}}{n-2}$	
Σύνολο	$\sum (y_i - \bar{y})^2$	n – 1		

(Παρένθεση)

- Το F-τεστ χρησιμοποιείται και για σύγκριση πολλών μέσων (σαν προέκταση του t-τεστ)
  - Για να συγκρίνουμε την μέση επιβίωση στην Ευρώπη, Ασία και Αμερική (E_E, $E_{Aσ}$, $E_{Aμ}$), εξετάζουμε100 άτομα από κάθε περιοχή
  - *F* = Παρατηρηθείσα μετ/τητα των E_E, $E_{Aσ}$, $E_{Aμ}$ / Προσδοκόμενη μεταβλητότητα των E_E, $E_{Aσ}$, $E_{Aμ}$
  - $F_{\text{ομαδες-1, δείγμα-ομάδες}}$
  - $F_{2,297}$

Κι άλλο τεστ : t τέστ

$$\frac{\beta}{SE(\beta)} = t_{n-2}$$

$$SE(\beta) = \frac{\sqrt{MS_{res}}}{\sqrt{\sum (x_i - \bar{x})^2}}$$

→ Μπορούμε να υπολογίσουμε και 95% δ.ε. για το β

Το F (για παλινδρόμιση με έναν συντελεστή) και το t τεστ πρέπει να δίνουν τα ίδια αποτελέσματα όσον αφορά την στατιστική σημαντικότητα

Συντελεστής προσαρμογής - Goodness of fit

- Τα συμπεράσματα των F και t τεστ εξαρτώνται από την ισχύ
- Ελέγχουν – δεν δείχνουν το πόσο καλό είναι το μοντέλο (πόσο καλά εφαρμόζει στα δεδομένα)

$$R^2 = \frac{SoS_{regr}}{SoS_{tot}}$$

- $0 \leq R^2 \leq 1$
- Περιγράφει το ποσό της διασποράς που μπορεί να εξηγήσει το μοντέλο (όσο περισσότερο τόσο καλύτερα!)

Παράδειγμα

Ασθενής	Σφυγμοί	Πίεση:
1	83	141
2	86	162
3	88	161
4	92	154
5	94	171
6	98	174
7	101	184
8	114	190
9	117	187
10	121	191

Ερμηνεία

- Παλινδρόμηση της πίεσης με τους σφυγμούς

$$BP = 1.12 \times HR + 60$$

- Για κάθε παλμό παραπάνω, η πίεση αυξάνει κατά 1.12
- Ένα άτομο με σφυγμούς 91 θα έχει πίεση 1.12 mmHg παραπάνω από κάποιον με 90 σφυγμούς
- Παλινδρόμηση της πίεσης με το φύλο (0: άνδρες, 1: γυναίκες

$$BP = 1.5 \times \text{φύλο} + 170$$

- Οι γυναίκες έχουν 1.5 mmHg παραπάνω απο τους άνδρες

Ερμηνεία

- Είναι στατιστικά σημαντική αυτή η αύξηση; (κοιτάμε την p-value=0.0003)
- Είναι κλινικά σημαντική;
- Πόσο καλό είναι το μοντέλο;
- R^2=81% - είναι καλό
- 81% της μεταβλητότητας εξηγείται από την παλινδρόμηση

Πολλαπλή παλινδρόμηση

- Πολλές ανεξάρτητες μεταβλητές, π.χ. $p = 3$ μεταβλητές

$$\hat{y}_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i$$

- Εκτίμηση των κλίσεων και έλεγχοι παρόμοιοι με την απλή παλινδρόμηση

$$\hat{y}_i = \alpha + \underbrace{\beta_1 x_{1i} + \ldots + \beta_k x_{ki} + \ldots + \beta_p x_{pi}}_{\text{μεταβλητές}} + \varepsilon_i$$

Έλεγχος για τις k

Πολλαπλή παλινδρόμηση: Σύγκριση και έλεγχος μοντέλων

- H_0 υπόθεση : $\beta_1=\beta_2=\ldots=\beta_k=0$ από τις p μεταβλητές του μοντέλου
- Φτιάχνουμε δύο μοντέλα: ένα με 0 μεταβλητές και ένα με p μεταβλητές
- Τα συγκρίνουμε

$$F = \frac{(SoS_{regr}^{\text{with the k pred}} - SoS_{regr}^{\text{without the k pred}})/k}{MS_{res}^{\text{with the k pred}}} \sim F_{k,n-p-1}$$

- Έλεγχος των συντελεστών

$$\frac{\beta_j}{SE(\beta_j)} \sim t_{n-p-1}$$

Πολλαπλή παλινδρόμηση: Συντελεστής προσαρμογής

$$R^2 = \frac{SoS_{regr}}{SoS_{total}} \times \frac{n-1}{n-p-1}$$

Πιο γενικά, το R^2 δείχνει την γραμμική συσχέτιση μεταξύ των παρατηρήσεων και των προσδοκόμενων (σύμφωνα με την παλινδρόμηση) τιμών

Ερμηνεία

- Παλινδρόμηση της πίεσης με τους σφυγμούς

$$BP = 1.03 \times HR + 1.1 \times \text{φύλο} + 140$$

- Ένα άτομο με σφυγμούς 91 θα έχει πίεση 1.03 mmHg παραπάνω από κάποιον με 90 σφυγμούς – *αυτή η αύξηση είναι σταθμισμένη για τις διαφορές ανάμεσα στα δύο φύλα*

Υποθέσεις

1. Κανονικότητα: y ακολουθεί κανονική κατανομή ≈ τα σφάλματα ε ακολουθούν κανονική κατανομή
2. Όλες οι παρατηρήσεις προέρχονται από την ίδια κατανομή με διασπορά σ^2
3. Γραμμικότητα: στην συσχέτιση των x και y
4. Ανεξαρτησία των παρατηρήσεων
5. Ανεξαρτησία των ανεξάρτητων μεταβλητών

Οι υποθέσεις 1,2,4 συμπεριλαμβάνονται στην σχέση

$$\varepsilon_i \sim N(0,\sigma^2)$$

Διάφορα άλλα...

- Πόσες ανεξάρτητες μεταβλητές; $p \approx n/20$

Όταν βλέπουμε αποτελέσματα παλινδρόμησης

- Έχουμε αρκετά δεδομένα;
- Ικανοποιούνται οι προϋποθέσεις (κανονική κατανομή, γραμμική συσχέτιση;)
- Κοιτάμε
  - τον συντελεστή β (coefficient)
  - το τυπικό σφάλμα του β (SE)
  - την p-value
  - τον συντελεστή προσαρμογής R^2
- Δεν μπορούμε να γενικεύσουμε πέραν των δεδομένων
  - Ερμηνεία των συντελεστών έχει νόημα μόνο μέσα στο πλαίσιο τιμών που εξετάσαμε
- Προσοχή στις ακραίες παρατηρήσεις!

The Communication and Symbolic Behaviour Scales (CSBS)

Table 3 Regression analysis of CSBS total score at 8 and 12 months

Αβεβαιότητα σε αυτή την εκτίμηση

Variable	Coefficient	95% CI	*P*-value
Female	2.97	1.66–4.27	<0.001
Twin birth	−10.20	−16.47 to −3.93	0.001
Premature birth (<36 weeks)	−1.17	−7.77–5.43	0.73
Non-English speaking background	3.58	0.42–6.74	0.03
Mother's education level (reference category: ≤11 years)			0.05
12 years	0.51	−2.49–3.51	
13 years	−0.10	−2.70–2.51	
University degree	−2.16	−4.98–0.65	
Postgraduate degree	−2.42	−5.51–0.68	
Family history of speech/language difficulties	−2.60	−4.20 to −1.01	0.001
Maternal mental health problems	−1.05	−2.52–0.42	0.16

Growth of infant communication between 8 and 12 months: A population study.
J Paediatr Child Health. 2006 Dec;42(12):764-70.

Ερμηνεία

«In multiple regression, neonatal aortic pulse wave velocity remained significantly inversely associated with maternal systolic BP (adjusted beta coefficient: -0.032; 95% CI: -0.040 to -0.024; P<0.001), after adjustment for maternal age, birth weight, length, and neonatal BP (all independently and positively related to aPWV) and for gestational age, maternal weight, and height (unrelated)»

Maternal and Neonatal Influences on, and Reproducibility of, Neonatal Aortic Pulse Wave Velocity. Hypertension. 2006 Nov 6;

$aPWV = \beta_1 \times BP$	$\beta_1 = -0.032$
$+ \beta_2 \times$ maternal age	$\beta_2 > 0$
$+ \beta_3 \times$ birth weight	$\beta_3 > 0$
$+ \beta_4 \times$ length	$\beta_4 > 0$
$+ \beta_5 \times$ neonatal BP	$\beta_5 > 0$
$+ \beta_6 \times$ gestational age	β_6 ?
$+ \beta_7 \times$ maternal weight	β_7 ?
$+ \beta_8 \times$ height	β_8 ?

$aPWV = \beta_1 \times BP$	$p < 0.001$
$+ \beta_2 \times$ maternal age	$p < 0.05$
$+ \beta_3 \times$ birth weight	$p < 0.05$
$+ \beta_4 \times$ length	$p < 0.05$
$+ \beta_5 \times$ neonatal BP	$p < 0.05$
$+ \beta_6 \times$ gestational age	β_6 ?
$+ \beta_7 \times$ maternal weight	β_7 ?
$+ \beta_8 \times$ height	β_8 ?

- Πόσες παρατηρήσεις πρέπει να έχουμε για να κάνουμε μια τέτοια παλινδρόμηση (με τόσες πολλές ανεξάρτητες μεταβλητές);
- $20 \times 8 = 160$ παρατηρήσεις

Table 2: Pakistani Medical Students' Knowledge and Attitude towards Research According to Gender, Type of High School, Year and Mode of Learning at Medical School

		No.	Knowledge		Attitude	
			Mean+SD	p-value	Mean+SD	p-value
Gender	Male	122	49.4+19.8	0.705	57.7+21.7	0.001
	Female	73	48.4+18.6		47.2+19.1	
High school type	HSSC	73	47.5+20.9	0.349	54.2+18.9	0.615
	A-Levels	110	50.0+18.5		53.4+22.4	
	Others	12	55.8+16.2		59.7+20.7	
Mode of learning	PBL	131	45.7+20.9	<0.001	49.0+18.6	<0.001
	LBL	66	55.5+15.3		63.5+23.1	
Year at medical school	1st	47	43.2+19.0		39.2+16.0	
	2nd	46	47.0+21.5		55.4+18.1	
	3rd	38	47.4+22.7		53.7+17.3	
	4th	32	58.4+17.3		60.7+27.2	
	5th	34	52.6+12.9		66.2+18.6	
Total		197	49.0+19.7		53.7+21.4	

HSSC, Higher secondary school certificate; A-levels, Advanced level; PBL, Problem based learning; LBL, Lecture based learning; SD, Standard deviating

Knowledge and attitudes about health research amongst a group of Pakistani medical students - *BMC Medical Education* 2006, **6:**54

Table 3: Predictors of Score on the Knowledge and Attitude Scales among Pakistani Medical Students

	Regression Coefficient (b)	Correlation coefficient (r)	p-value
		Knowledge	
Year at medical school	4.1	0.30	0.019
Age	-0.98	-0.90	0.476
		Attitude	
Year at medical school	6.67	0.45	<0.001
Age	-0.63	-0.05	0.661

$$\text{Knowledge score} = 4.1 \times \text{Years} - 0.98 \times \text{Age} + \alpha$$

$$\text{Attitude score} = 6.7 \times \text{Years} - 0.63 \times \text{Age} + \alpha$$

Knowledge and attitudes about health research amongst a group of Pakistani medical students
BMC Medical Education 2006, **6:**54

Ερώτηση

- Πόσο είναι το σκορ γνώσης για ένα άτομο ηλικίας 19 ετών στο 3ο έτος σπουδών;
- Σκορ = $4.1 \times 3 - 0.98 \times 19 + \alpha$
- Έστω ότι ξέρω α = 55, Σκορ = 48.68
- Σε ποιο έτος θα είναι κάποιος με σκορ 52 ηλικίας 20 ετών;
- Για να το βρούμε αυτό χρειαζόμαστε την παλινδρόμηση του έτους σε σχέση με το σκορ και την ηλικία!

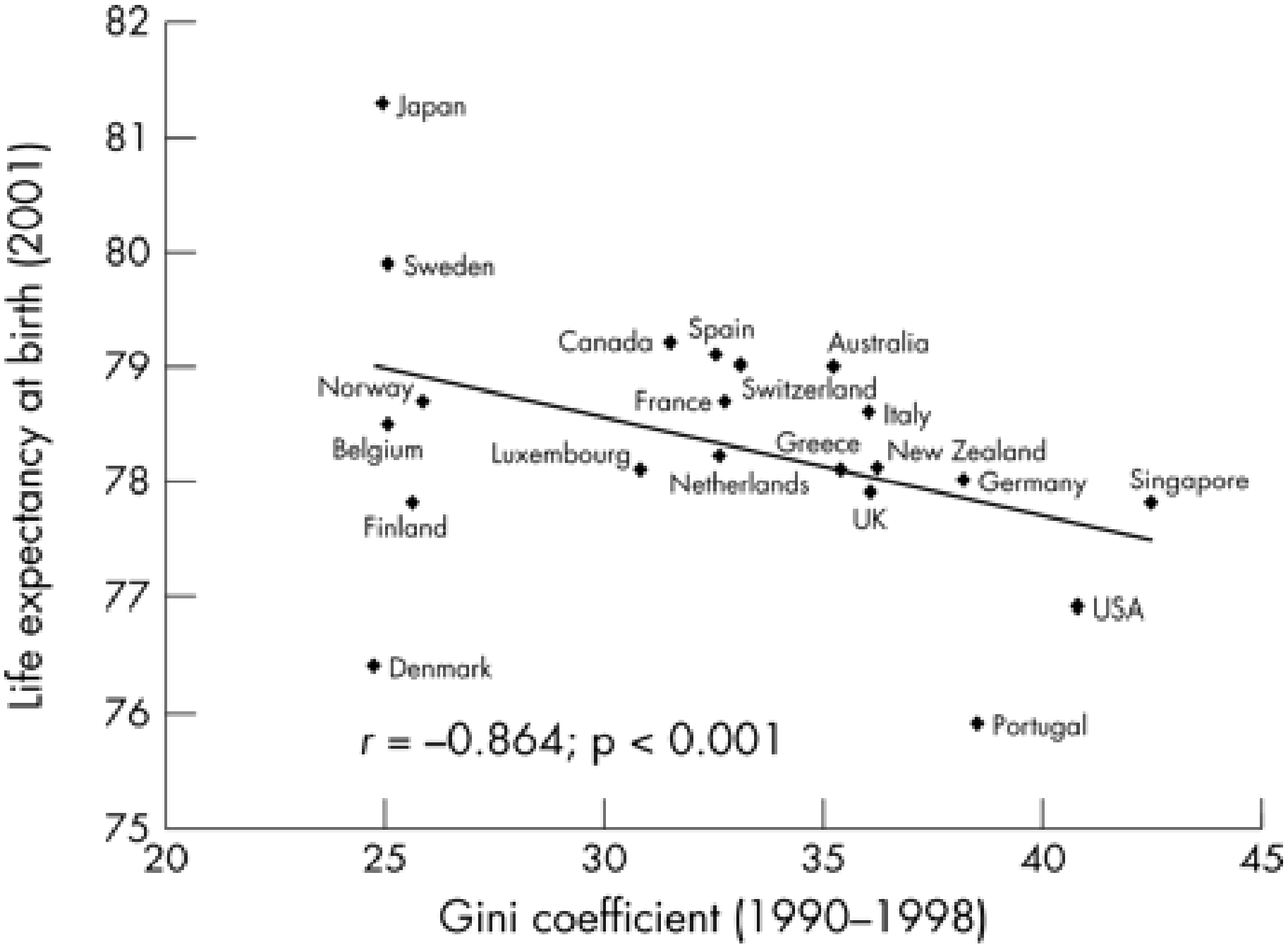

Life expectancy at birth (2001)
82
81
80
79
78
77
76
75
Japan
Sweden
Canada
Spain
Australia
Norway
France
Switzerland
Italy
Belgium
Luxembourg
Greece
New Zealand
Netherlands
Germany
Singapore
Finland
UK
USA
Denmark
Portugal
r = –0.864; p < 0.001
20
25
30
35
40
45
Gini coefficient (1990–1998)

Table 1 Linear regression on life expectancy at birth across Italian regions (n = 20)

	Standardised β coefficients				
	Model 1	Model 2	Model 3	Model 4	Model 5
Gini index	−0.658***	–	–	−0.785***	−0.559***
Per capita income	–	0.538***	–	−0.146*	–
Education	–	–	0.476***	–	0.276***
Constant	82.652	78.287	76.315	83.616	80.187
r^2	0.433	0.290	0.226	0.438	0.500
Model statistics:					
F (df, N)	363.9*** (1, 20)	194.3*** (1, 20)	139.4*** (1, 20)	185.4*** (2, 20)	237.5*** (2, 20

*p<0.05; **p<0.01; ***p<0.001.

Marco Maggiorini, Peter Bartsch, Oswald Oelz: Association between raised body temperature and acute mountain sickness: cross sectional
study. *British Medical Journal*, 315, 403-4.

- Τι μελέτη είναι;
- Εξηγήστε τα αποτελέσματα