

# Λογαριθμιστική παλινδρόμηση

Γεωργία Σαλαντή

# Διχότομα δεδομένα (ναι/όχι)

- Borchgrevink 1966

Εγκεφαλικό	Ναι	Όχι	Σύνολο
Omega 3	10	90	100
Τίποτα	14	86	100
	24	186	200

# Κίνδυνος

- 20 φοιτητές πίνουν ρούμι πριν τις εξετάσεις, 5 κόβονται
- **Κίνδυνος (risk)**
  - = 5 κόβονται/20 όλοι
  - =  $5/20 = 1/4 = 0.25 = 25\%$

$$\text{Κίνδυνος} = \frac{\# \text{ γεγονότων}}{\# \text{ παρατηρήσεων}}$$

- Η πιθανότητα να κοπείς είναι 25% αν πιεις ρούμι την προηγούμενη νύχτα

# Αναλογία

- 20 φοιτητές πίνουν ρούμι πριν τις εξετάσεις, 5 κόβονται
- **Αναλογία (odds)**

= 5 κόβονται/15 περνάνε

=  $5/15 = 1/3 = 0.33$

$$\text{Αναλογία} = \frac{\# \text{ γεγονότων}}{\# \text{ παρατηρήσεων} - \# \text{ γεγονότων}}$$

- Η πιθανότητα να κοπείς όταν πίνεις ρούμι είναι ένα τρίτο της πιθανότητας να περάσεις
- Ένας θα κοπεί για κάθε τρεις που θα περνάνε
- Οι πιθανότητες είναι τρία προς ένα κατά

# Λόγος κινδύνων και λόγος αναλογιών

$$\Lambda\text{Κ} = \frac{\# \text{ Κίνδυνος Ομάδα 1}}{\# \text{ Κίνδυνος Ομάδα 2}} \quad \textbf{Risk Ratio (RR)}$$

$$\Lambda\text{Α} = \frac{\# \text{ Αναλογία Ομάδα 1}}{\# \text{ Αναλογία Ομάδα 2}} \quad \textbf{Odds Ratio (OR)}$$

	Γεγονός	Όχι γεγονός	Συνολο
Νέα θεραπεία	<i>a</i>	<i>b</i>	<i>a+b</i>
Παλιά θεραπεία	<i>c</i>	<i>d</i>	<i>c+d</i>
	<i>a+c</i>	<i>b+d</i>	<i>n</i>

$$\Lambda\text{Κ} = \frac{a (c+d)}{c (a+b)}$$

$$\Lambda\text{Α} = \frac{a d}{c b}$$

# Ερμηνεία λόγων

- $OR = 1$  σημαίνει ότι η έκθεση δεν σχετίζεται με τη νόσο
- Ο κίνδυνος είναι ο ίδιος στις δύο ομάδες
- $OR=2$  : Η αναλογία στην ομάδα A είναι διπλάσια σε σχέση με την ομάδα B

# Λίγο ακόμα για λόγους

- $\pi_1$  και  $\pi_2$  οι πιθανότητες στις ομάδες 1 και 2

$$OR = \frac{\frac{\pi_1}{1-\pi_1}}{\frac{\pi_2}{1-\pi_2}} \quad \log(OR) = \log\left(\frac{\pi_1}{1-\pi_1}\right) - \log\left(\frac{\pi_2}{1-\pi_2}\right)$$

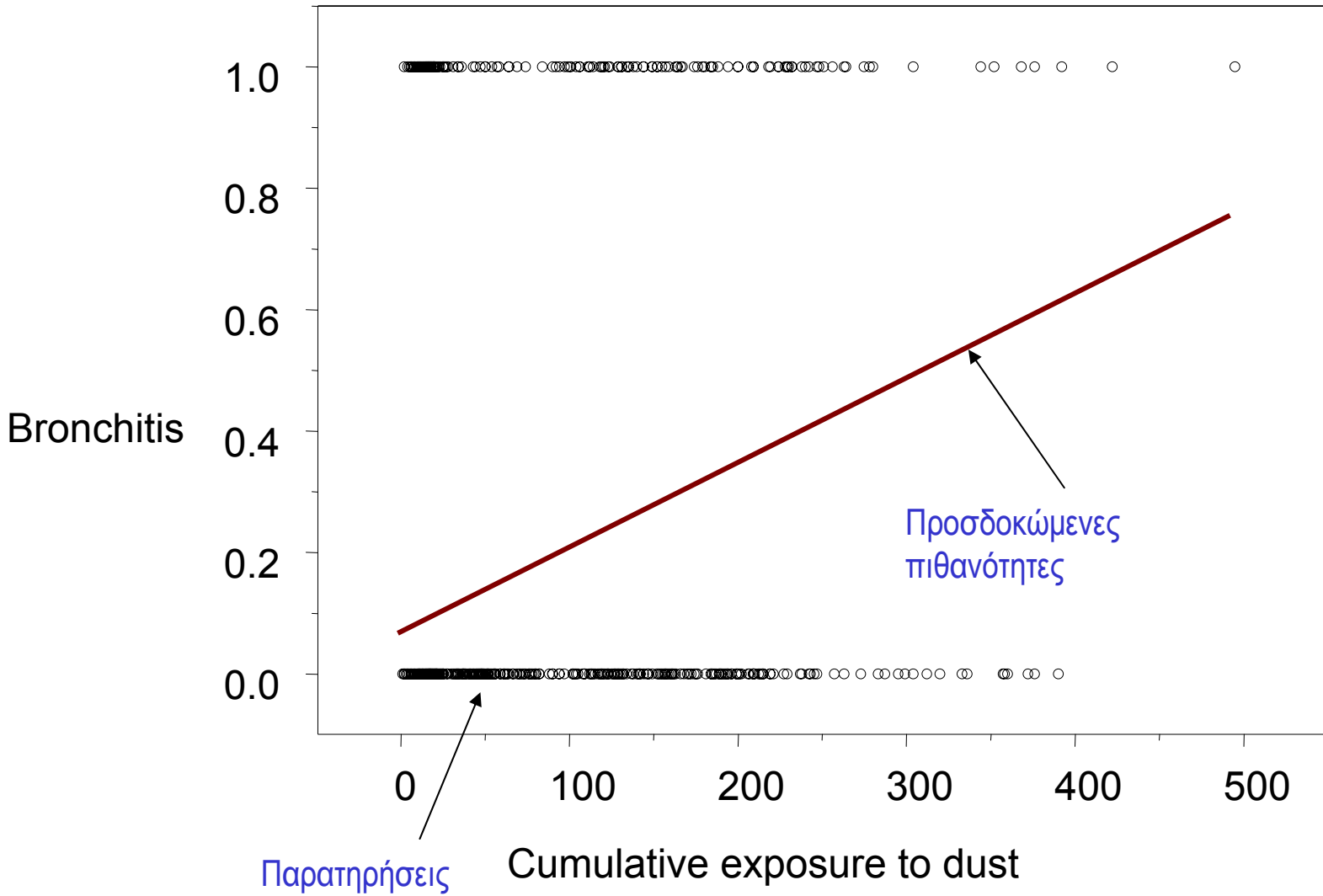
# Συμβολισμοί

- $y$  το **διχότομο** αποτέλεσμα (ή δεσμευμένη μεταβλητή) που μας ενδιαφέρει
  - Π.χ. χρόνια βρογχίτιδα 0/1 = ναι/όχι
- $x$  η ανεξάρτητη μεταβλητή (συνεχής η διχότομη)
  - Π.χ. ηλικία, φύλο



# Παρατηρήσεις

- Για κάθε άτομο παρατηρούμε 0 ή 1
- $Y_1, Y_2, \dots, Y_n$   
 $0, 1, \dots, 1$
- Από τους 100 εργαζόμενους στην μεταλλουργία, οι 20 παρουσίασαν χρόνια βρογχίτιδα
- Πιθανότητα βρογχίτιδας:  $P(Y=1) = \pi = 20\%$
- Μας ενδιαφέρει η πιθανότητα για κάθε άτομο:  $P(Y_i=1) = \pi_i$
- **Λογαριθμιστική παλινδρόμηση:** το  $\pi_i$  σαν συνάρτηση μιας ανεξάρτητης μεταβλητής  $x$  (π.χ. αθροιστική έκθεση στη σκόνη)

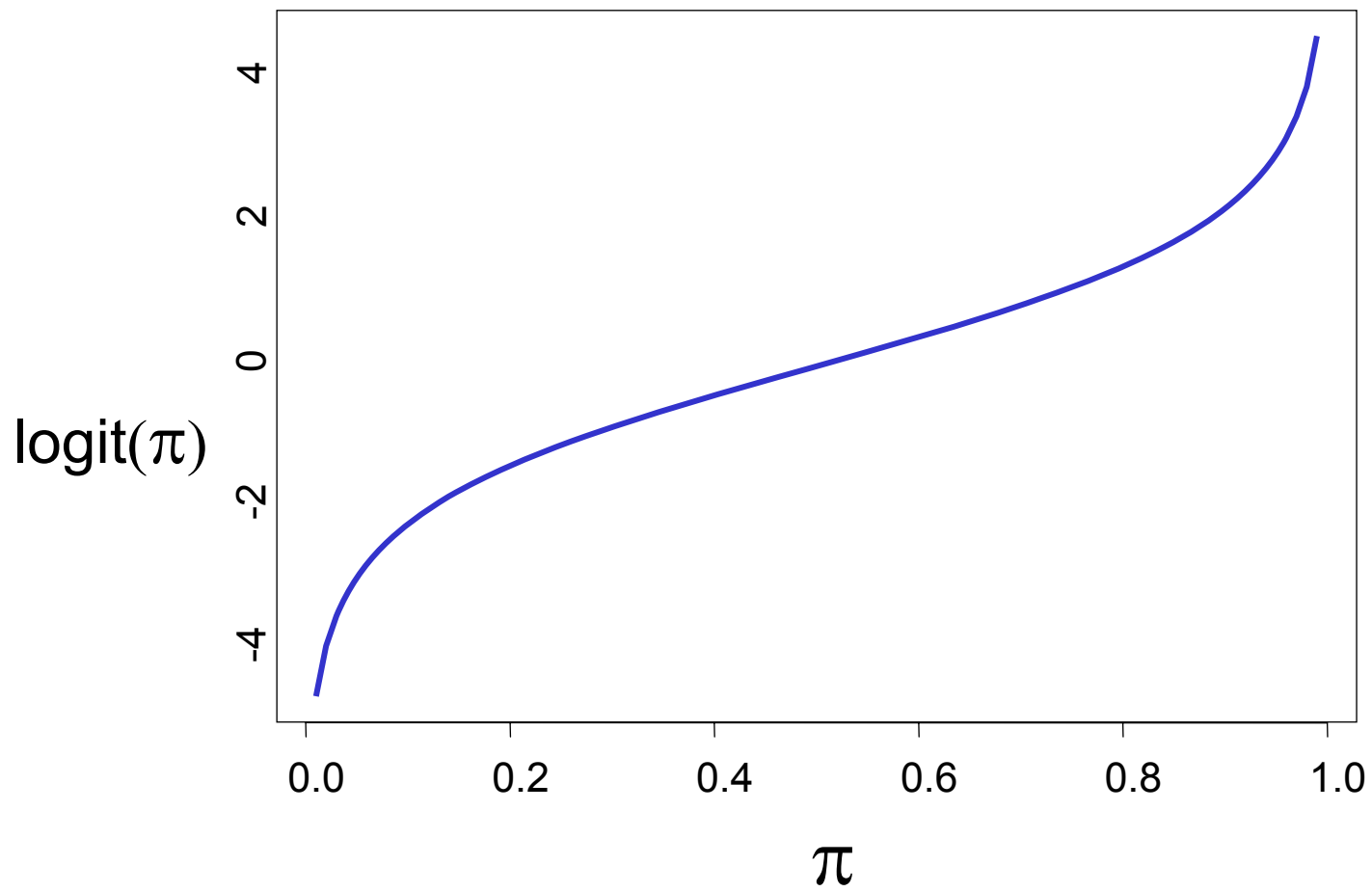


# Λογαριθμιστική Συνάρτηση- Logit

- Εξαρτημένη μετ/τή: αντί για  $\pi$  χρησιμοποιούμε το  $\text{logit}(\pi)$

$$\text{logit}(\pi) = \log\left(\frac{\pi}{1 - \pi}\right)$$

- Γιατι;
  - Έχει καλύτερες ιδιότητες (χωρίς όρια, κανονική κατανομή)
  - Δίνει  $\log\text{OR}$  ( $\log\text{LA}$ ) που είναι *μαθηματικώς* καλύτερο από το  $\text{OR}$  ( $\text{LA}$ )



# Λογαριθμιστική Παλινδρόμηση

$\alpha$  : αρχή

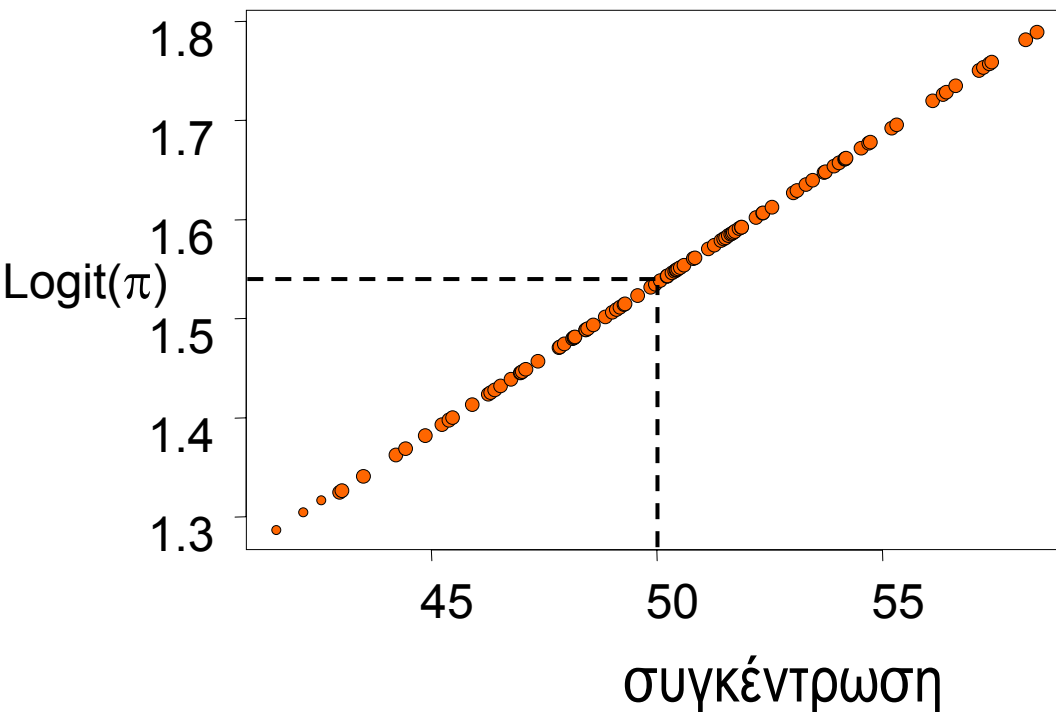
$\beta$  : κλίση

$$\text{logit}(\pi_i) = \alpha + \beta x_i + \varepsilon_i$$

$$\pi_i = \frac{\exp(\alpha + \beta x_i + \varepsilon_i)}{1 + \exp(\alpha + \beta x_i + \varepsilon_i)}$$

# Λογαριθμιστική Παλινδρόμηση

$$\text{logit}(\pi_i) = 0.04 + 0.03 \times \text{συγκέντρωση}_i$$

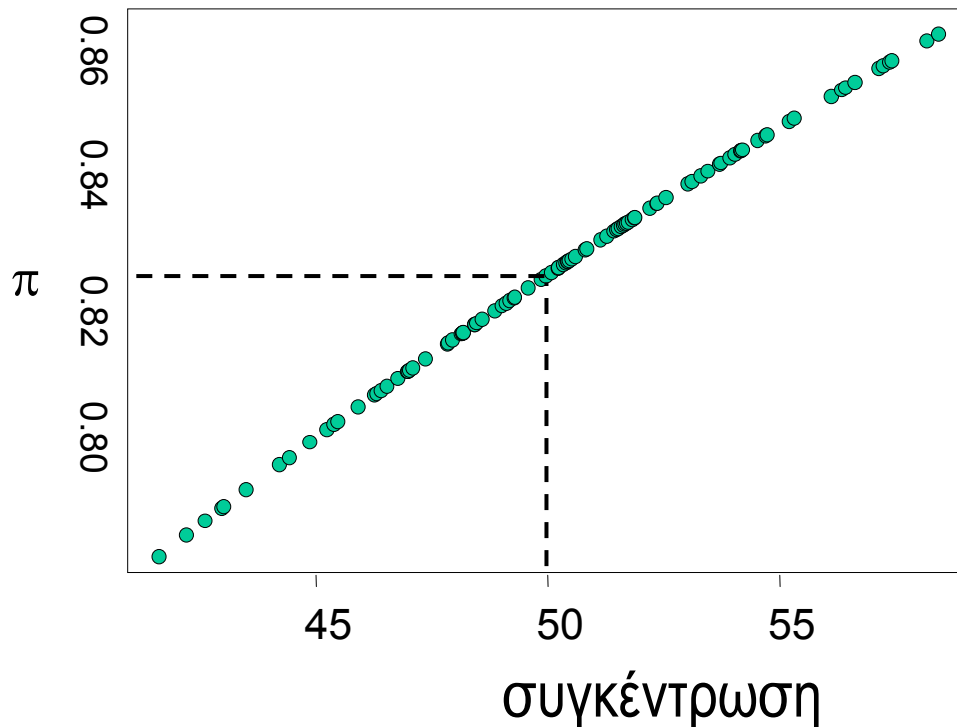


Για συγκέντρωση<sub>i</sub> = 50<sub>mg/cm<sup>3</sup></sub>

$$y_i = \text{logit}(\pi_i) = 1.54$$

# Λογαριθμιστική Παλινδρόμηση

$$\text{logit}(\pi_i) = 0.04 + 0.03 \times \text{συγκέντρωση}_i$$



Για συγκέντρωση<sub>i</sub> = 50 mg/cm<sup>3</sup>

$$\pi_i = 0.82 = 82\%$$

# OR και λογαριθμιστικό μοντέλο

- $\pi_1$  και  $\pi_2$  οι πιθανότητες στις ομάδες 1 και 2

$$\log(\text{OR}) = \log\left(\frac{\pi_1}{1-\pi_1}\right) - \log\left(\frac{\pi_2}{1-\pi_2}\right)$$

$$\log(\text{OR}) = \text{logit}(\pi_1) - \text{logit}(\pi_2)$$

$$\text{OR} = \exp(\text{logit}(\pi_1) - \text{logit}(\pi_2))$$

$$\text{OR} = \exp(\alpha + \beta x_1 - \alpha - \beta x_2)$$

$$\text{OR} = \exp(\beta (x_1 - x_2))$$



# Ερμηνεία των συντελεστών

Διχότομη ανεξάρτητη μεταβλητή: το OR είναι  $\exp(\beta)$

$$\text{logit}(\pi_i) = 0.04 + 0.03 \times \text{καπν}_i \quad \left\{ \begin{array}{l} 0 \text{ μη καπνιστής} \\ 1 \text{ καπνιστής} \end{array} \right.$$

μη καπνιστής

καπνιστής

$$\text{logit}(\pi_{\text{MK}}) = 0.04$$

$$\text{logit}(\pi_{\text{K}}) = 0.07$$

$$\log\text{OR} = \text{logit}(\pi_{\text{K}}) - \text{logit}(\pi_{\text{MK}}) = 0.03$$

$$\text{OR} = \exp(0.03)$$

# Ερμηνεία των συντελεστών

Συνεχής ανεξάρτητη μετ/τη: το OR για μια μονάδα διαφοράς είναι  $\exp(\beta)$

$$\text{logit}(\pi_i) = 0.04 + 0.03 \times \text{συγκέντ}_i$$

$$\text{συγκέντρωση} = 50 \text{ mg/cm}^3$$

$$\text{logit}(\pi_{50}) = 1.54$$

$$\text{συγκέντρωση} = 51 \text{ mg/cm}^3$$

$$\text{logit}(\pi_{51}) = 1.57$$

$$\log\text{OR} = \text{logit}(\pi_{51}) - \text{logit}(\pi_{50}) = 0.03$$

$$\text{OR} = \exp(0.03)$$

# Εκτίμηση των $\alpha$ και $\beta$

- Εκτίμηση μεγιστοποιώντας την συνάρτηση πιθανοφάνειας
- Πιθανοφάνεια (likelihood): Πόσο πιθανοί είναι οι συντελεστές που εκτιμούμε όταν παρατηρούμε τα δεδομένα
- Έστω  $n$  παρατηρήσεις που είναι  $y_i=0$  ή  $y_i=1$

$$L = \prod_{i=1}^n \pi_i^{y_i} (1 - \pi_i)^{1-y_i}$$

$$L = \prod_{i=1}^n (1 + \exp(-\alpha - \beta x_i))^{-y_i} \left[ 1 - (1 + \exp(-\alpha - \beta x_i))^{-1} \right]^{1-y_i}$$

Μεγιστοποιούμε την  $L$  χρησιμοποιώντας την μέθοδο Newton-Raphson και εκτιμούμε  $\alpha$  και  $\beta$

# Πιθανοφάνεια, Deviance και προσαρμογή του μοντέλου

- Πιθανοφάνεια: όσο πιο μεγάλη τόσο καλύτερο το μοντέλο
- Deviance  $D = -2 \log(\text{Likelihood})$ 
  - ο Όσο μικρότερη, τόσο καλύτερο το μοντέλο
- Καλή προσαρμογή = μεγάλη πιθανοφάνεια = μικρή Deviance
- Όσο πιο πολλές μεταβλητές έχει το μοντέλο, τόσο καλύτερη η προσαρμογή του, αλλά γίνεται πιο περίπλοκο

# Λογαριθμιστική Παλινδρόμηση με πολλές μεταβλητές

$$\text{logit}(\pi_i) = \alpha + \beta_1 \times \text{καπν}_i + \beta_2 \times \text{συγκ}_i$$

Έλεγχοι και επιλογή μοντέλου:

- Έλεγχος για κάθε  $\beta_i$ ; **Wald test, Likelihood ratio test (LRT)**
- Έλεγχος ενός μοντέλου έναντι του άλλου: **LRT**
- Έλεγχος για όλους τους συντελεστές σύγχρονος: **LRT**

# Wald τεστ

Για κάθε ανεξάρτητη μεταβλητή ξεχωριστά

$H_0: \beta_i = 0$  ή ότι η ανεξάρτητη μεταβλητή  $x$  δεν έχει προγνωστική αξία ως προς την  $y$

$$\frac{\beta_i}{SE(\beta_i)} \sim \chi^2_{1df}$$

# Επιλογή μεταβλητών – έλεγχοι μοντέλων

- Για να αποφασίσουμε αν μια μεταβλητή  $X$  είναι σημαντική για το μοντέλο θα συγκρίνουμε δυο μοντέλα με βάση την προσαρμοστικότητα τους στα δεδομένα (χρησιμοποιώντας την deviance)
  - Το μοντέλο **χωρίς** την μεταβλητή  $X$
  - Το μοντέλο **με** την μεταβλητή  $X$
- Μοντέλο 1: μοντέλο με το κάπνισμα  $\Rightarrow D_1$
- Μοντέλο 2: μοντέλο με το κάπνισμα και την συγκέντρωση σκόνης  $\Rightarrow D_2$
- Προτιμούμε το μοντέλο που έχει τη μικρότερη Deviance

# Επιλογή μεταβλητών και σύγκριση μοντέλων

- Το μοντέλο με τις πιο πολλές μεταβλητές θα έχει τη μικρότερη Deviance
  - $D_2 < D_1$
- Όμως είναι και το πιο περίπλοκο
  - Πολλές μεταβλητές
- Χρειαζόμαστε έναν κανόνα για να αποφασίσουμε εάν η μείωση στην deviance 'αξίζει' την αυξημένη περιπλοκότητα του μοντέλου

Likelihood Ratio Test (LRT) [ Λόγος Πιθανοφανειών]



# LRT: γενίκευση

- $LRT = D(\text{μοντέλο χωρίς τις } p \text{ μεταβλητές}) - D(\text{μοντέλο με τις } p \text{ μεταβλητές})$

$LRT \sim \chi^2$  με  $p$  βαθμούς ελευθερίας

- $P\text{-value} < 0.05$  : Το μοντέλο με τις επιπλέον μεταβλητές **είναι** στατιστικά καλύτερο από το μοντέλο χωρίς τις μεταβλητές  $\Rightarrow$  κρατάμε τις μεταβλητές ως σημαντικές
  - $P\text{-value} > 0.05$  : Το μοντέλο με τις επιπλέον μεταβλητές **δεν είναι** στατιστικά καλύτερο από το μοντέλο χωρίς τις μεταβλητές  $\Rightarrow$  πετάμε τις μεταβλητές ως ασήμαντες
- **Προσοχή!** Τέτοιοι ‘απλοϊκοί’ κανόνες δεν είναι πάντα χρήσιμοι – να κοιτάμε και την κλινική πλευρά του θέματος

# Επιλογή μοντέλου: Έλεγχος για την $X_3$

Variables in the model	Deviance	Compare	$\chi^2$	P-value
$X_1 X_2 X_3$	$D_1$			
$X_1 X_2$	$D_2$			
$X_3$	$D_3$			
none	$D_4$			

$$D_1 < D_2 < D_3 < D_4$$

# Επιλογή μοντέλου: Έλεγχος για την $X_3$

Variables in the model	Deviance	Compare	$\chi^2$	P-value
$X_1 X_2 X_3$	$D_1$			Ελέγχει για την $X_3$
$X_1 X_2$	$D_2$			
$X_3$	$D_3$	$D_3 - D_4$	1df	
none	$D_4$			

$$D_1 < D_2 < D_3 < D_4$$

# Επιλογή μοντέλου: Έλεγχος για την $X_3$

Variables in the model	Deviance	Compare	$\chi^2$	P-value
$X_1 X_2 X_3$	$D_1$	$D_1 - D_2$	1df	Ελέγχει για την $X_3$
$X_1 X_2$	$D_2$			
$X_3$	$D_3$			
none	$D_4$			

Αυτός ο έλεγχος είναι 'σταθμισμένος' για τις μεταβλητές  $X_1, X_2$

$$D_1 < D_2 < D_3 < D_4$$

# Επιλογή μοντέλου: παράδειγμα

Variables in the model	Deviance	Compare	$\chi^2$	P-value
$X_1 X_2 X_3$	$D_1$			
$X_1 X_2$	$D_2$			
$X_3$	$D_3$			
none	$D_4$			

$$D_1 < D_2 < D_3 < D_4$$

# Επιλογή μοντέλου: παράδειγμα

Variables in the model	Deviance	Compare	$\chi^2$	P-value
$X_1 X_2 X_3$	$D_1$	$D_1 - D_4$	3 df	Ελέγχει για τις $X_1, X_2, X_3$
$X_1 X_2$	$D_2$			
$X_3$	$D_3$			
none	$D_4$			

$$D_1 < D_2 < D_3 < D_4$$

# Προσαρμογή του μοντέλου στα δεδομένα

- Ψευδο  $R^2 = \frac{\text{Πιθανοφάνεια}_{\text{μοντέλου}} - \text{Πιθανοφάνεια}_{\text{0 μεταβλητες)}}{\text{Πιθανοφάνεια}_{\text{μοντέλου}}}$

Είναι το ποσοστό αύξησης της πιθανοφάνειας με το μοντέλο

- Akaike's Information Criterion

$$\text{AIC} = \text{Deviance} + 2 \times p$$

$p$  ο αριθμός των μεταβλητών στο μοντέλο

- Όσο μικρότερο τόσο καλύτερα
- Διαφορά 3 βαθμών ή παραπάνω είναι σημαντική

# Hosmer and Lemenshow's τεστ για έλεγχο προσαρμογής

- Συγκρίνουμε παρατηρούμενα και προσδοκώμενα γεγονότα (σύμφωνα με τις προβλέψεις του μοντέλου)
- Αν το μοντέλο είναι καλό, παρατηρούμενα και προσδοκώμενα γεγονότα πρέπει να συμφωνούν
- Μέθοδος
  - Χωρίζουμε τα δεδομένα σε ομάδες, π.χ. 10 ομάδες
  - Εκτιμούμε τον αριθμό γεγονότων από το μοντέλο
  - Για  $H_0$  : 'Το μοντέλο ταιριάζει στα δεδομένα' υπολογίζουμε ένα  $\chi^2$  τεστ
  - Εάν  $p\text{-value} > 0.05$  Το μοντέλο δεν είναι αταίριαστο με τα δεδομένα



Table 2: Regression of LBWT on SES: Controlling for Sociodemographics, Health Behaviors, Health-Related Variables, and Mother's BWT

<i>Variable</i>	<i>Model 1</i>	<i>Model 2</i>	<i>Model 3</i>	<i>Model 4</i>
Income (thousands)	− .0445**			− .0437**
Mom's Education (years)	− .0259	− .0529**	− .0486**	− .0244
Dad's Education (years)	− .0166	− .0382**	− .0293 <sup>†</sup>	− .0141
Occupational Grade			− .0067 <sup>†</sup>	− .0014
Income Inequality (Gini)		3.0692		2.0935
Constant	− 1.6536	− 2.7799	− 1.4554	− 2.6686
Wald Chi-Square ( <i>df</i> )	152.83 (8)	72.23 (8)	73.49 (6)	156.77 (14)
N-size	12,814	12,814	12,814	12,814

*Note:* <sup>†</sup> $p < .10$ , \* $p < .05$ , \*\* $p < .01$ ; Normal BWT is the reference category—high BWT results are modeled simultaneously to improve efficiency of the estimates, but results are not presented.

$$\text{Logit}(\pi) = -1.65 - 0.045\text{Income} - 0.03\text{YearsM} - 0.02\text{YearsF} + \dots$$

$$\text{OR} = \exp(-0.045) = 0.96 \text{ για διαφορά } 1000\$$$

# Πρόγνωση

- $\text{Logit}(\pi) = -1.65 - 0.045 \text{ Income} - 0.03 \text{ YearsM} - 0.02 \text{ YearsF} + \dots$
- Για Income 2,000 με 5 και 1 έτη σπουδών στην μητέρα και πατέρα αντίστοιχα
  - $\text{Logit}(\pi) = -1.91$
  - $\text{logit}^{-1}(\pi) = \exp(-1.91) / [1 + \exp(-1.91)] \rightarrow \pi = 15\%$
- Για Income 10,000 με 5 και 1 έτη σπουδών στην μητέρα και πατέρα αντίστοιχα
- $\text{Logit}(\pi) = -2,3$ 
  - $\pi = 10\%$

Figure 1: Predicted Probabilities for LBWT by Income: Separate Unadjusted and Adjusted Models

