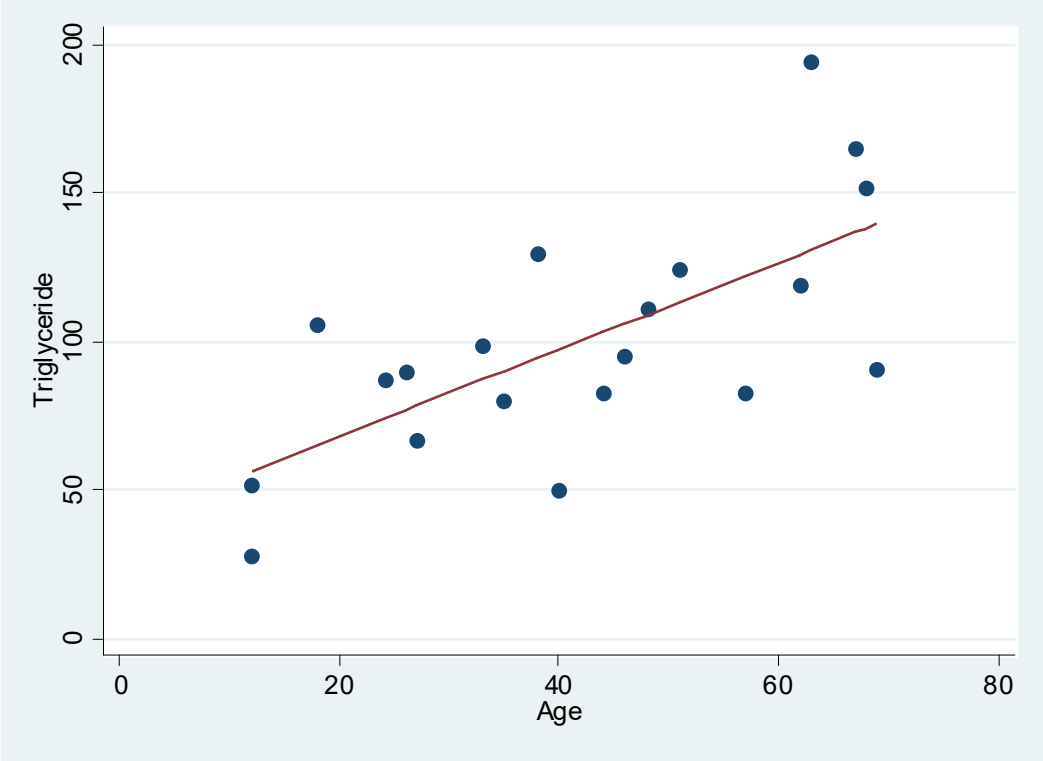
Λογαριθμιστική εξάρτηση

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

Εισαγωγή

- Έχουμε μάθει ότι αν θέλουμε να διερευνήσουμε την εξάρτηση ενός **ποσοτικού** μεγέθους από ένα άλλο παράγοντα μπορούμε να χρησιμοποιήσουμε την απλή γραμμική εξάρτηση
  - π.χ. Εξάρτηση επιπέδων τριγλυκεριδίων από την ηλικία



Γραμμική εξάρτηση

Η χρήση αυτής της μεθόδου επιτρέπει:

- Να αξιολογήσουμε αν η ηλικία επηρεάζει τα επίπεδα τριγλυκεριδίων
- Να κάνουμε προβλέψεις για τα αναμενόμενα επίπεδα τριγλυκεριδίων με βάση την ηλικία του ατόμου

Προϋπόθεση:

- Η εξαρτημένη μεταβλητή Y να έχει κανονική κατανομή
- Η ευθεία γραμμή να είναι ικανοποιητική προσέγγιση της σχέσης Y και X

Μοντέλα για δίτιμες μεταβλητές

Τι συμβαίνει αν η μεταβλητή την οποία θέλουμε να μελετήσουμε (η εξαρτημένη) δεν είναι ποσοτική αλλά **ποιοτική με 2 επίπεδα;**

π.χ

- Εμφάνιση καρκίνου του μαστού - Ναι ή Όχι
- Ανταπόκριση στη θεραπεία - Ναι ή Όχι

Παράδειγμα

- Ηλικία και παρουσία στεφανιαίας νόσου (CHD-Coronary Heart Disease) για N=33 άτομα

Age	CHD
22	0
23	0
24	0
27	0
28	0
30	0
30	0
32	0
33	0
35	1
38	0

Age	CHD
40	0
41	1
46	0
47	0
48	0
49	1
49	0
50	1
51	0
51	1
52	0

Age	CHD
54	0
55	1
58	1
60	1
60	0
62	1
65	1
67	1
71	1
77	1
81	1

Πώς μπορούμε να αναλύσουμε αυτά τα δεδομένα;

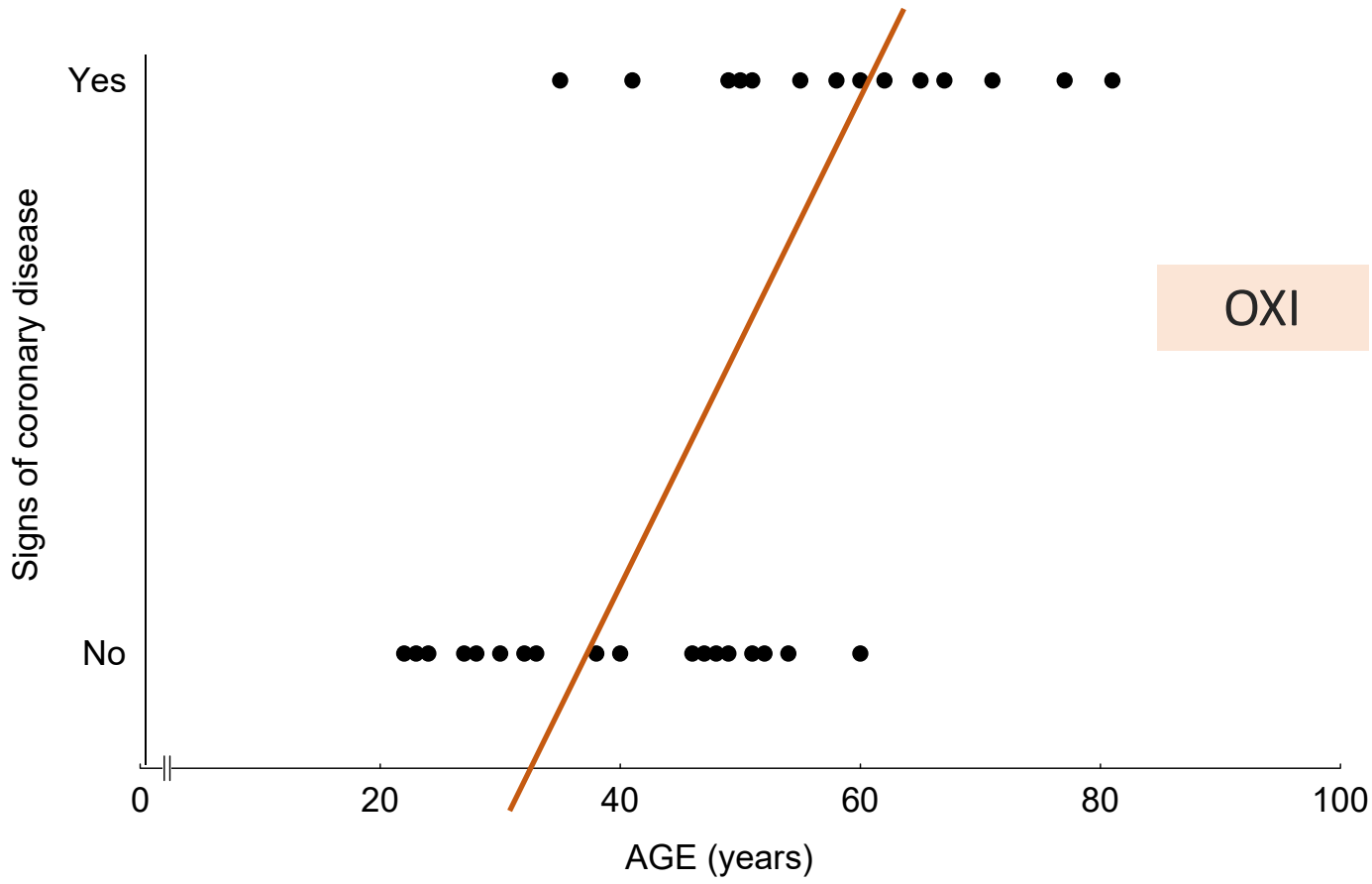
- Συγκρίνοντας τη μέση ηλικία ασθενών με τη μέση ηλικία των υγιών

  - Υγιείς: 38.6 έτη
  - Ασθενείς: 58.7 έτη

  } t-test: $p<0.001$

- Μπορώ να χρησιμοποιήσω γραμμική εξάρτηση για να διερευνήσω αν η ηλικία επηρεάζει την εμφάνιση στεφανιαίας νόσου;

Διάγραμμα ηλικίας και CHD



Μοντέλα για δίτιμες εξαρτημένες μεταβλητές

- Όταν η εξαρτημένη είναι **συνεχής** μεταβλητή, μας ενδιαφέρει να εκτιμήσουμε πως μεταβάλλονται τα επίπεδα της κατά μέσο όρο αν αυξηθεί η ανεξάρτητη κατά μία μονάδα, π.χ.
  - αύξηση της εβδομαδιαίας φυσικής δραστηριότητας κατά 1 ώρα σημαίνει μείωση του βάρους κατά 500 gr κατά μέσο όρο

- Όταν η εξαρτημένη είναι δίτιμη μεταβλητή (π.χ. απουσία/ παρουσία νόσου) μας ενδιαφέρει αν η αύξηση της ανεξάρτητης μεταβλητής (π.χ. αριθμός τσιγάρων) σχετίζεται με **αύξηση ή μείωση της πιθανότητας να έχει το άτομο τη νόσο**

Μοντέλα για δίτιμες εξαρτημένες μεταβλητές

- Σε τέτοια δεδομένα η εξαρτημένη μεταβλητή (Υ) παίρνει τιμές 0 και 1
  - Π.χ. θεραπεία **0=αποτυχία, 1=επιτυχία**
    νόσος **0=απουσία, 1=παρουσία**

- Συμβολίζουμε
  - πιθανότητα παρουσίας της νόσου: **P(Y=1)= π**
  - πιθανότητα απουσίας της νόσου: **P(Y=0)=1- π**

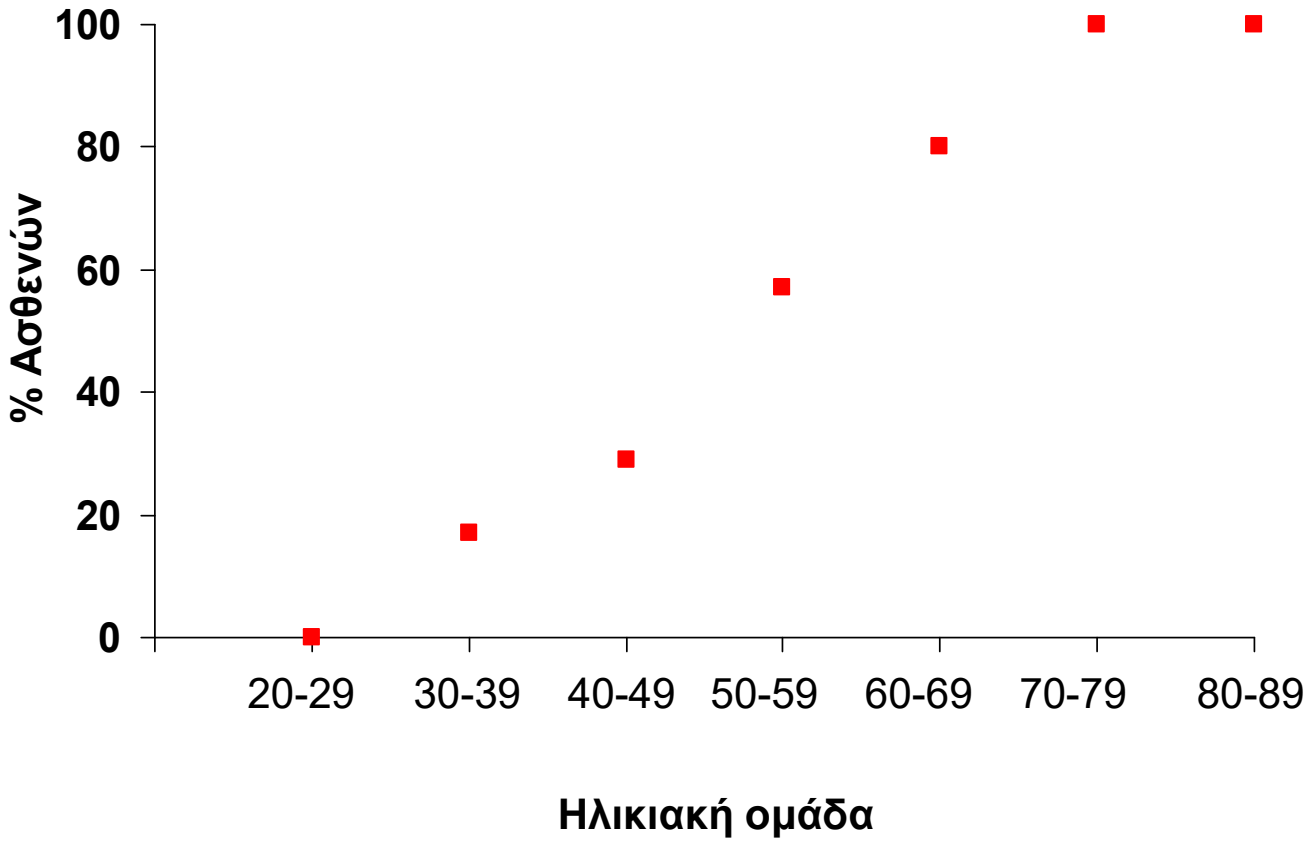
- Μας ενδιαφέρει να ανιχνεύσουμε τις μεταβλητές (ανεξάρτητες μεταβλητές) που ενδέχεται να σχετίζονται με την πιθανότητα π εμφάνισης της νόσου.

Στο παράδειγμα της στεφανιαίας νόσου (CHD)

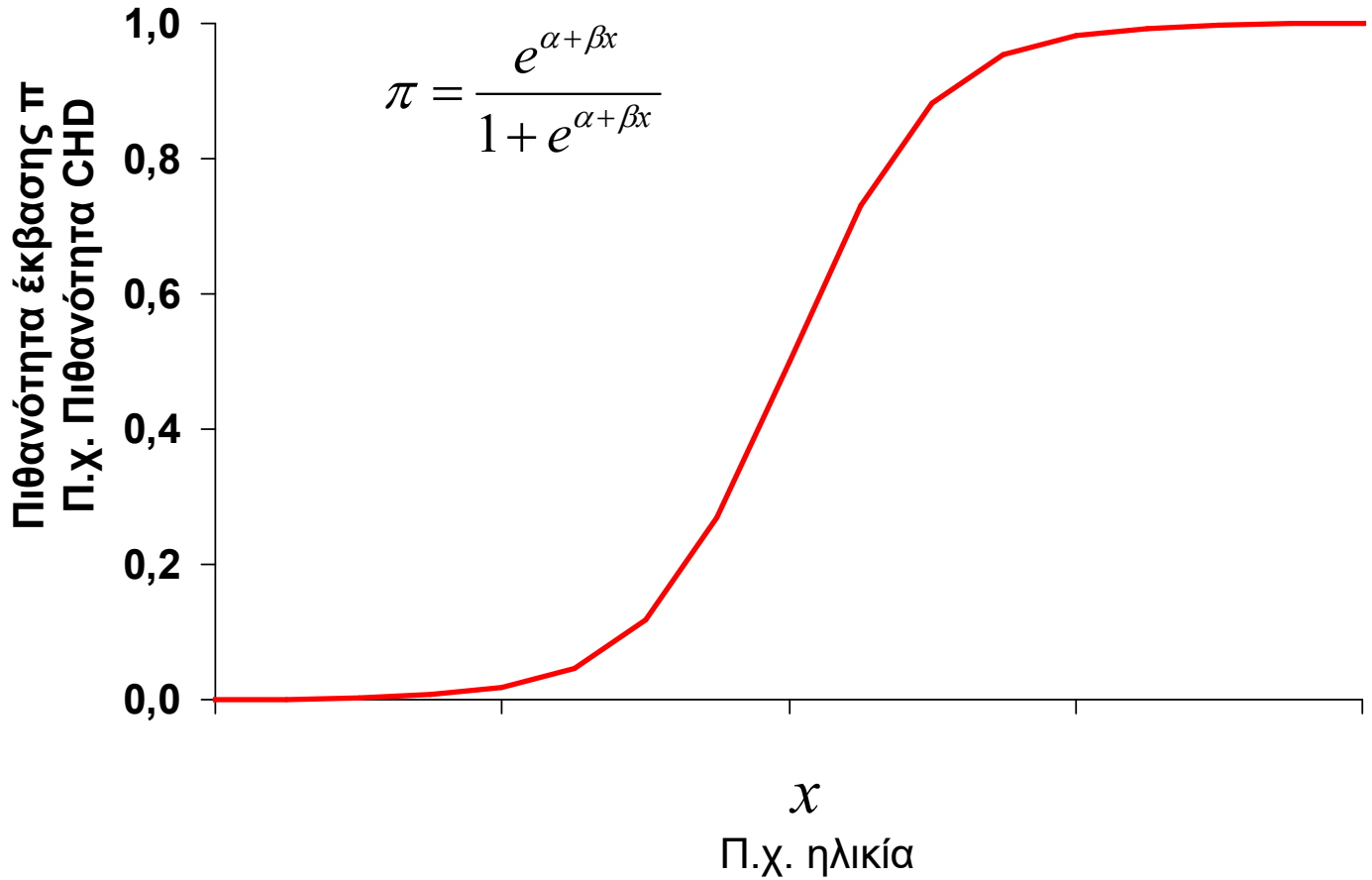
Αρχική διερεύνηση για το πως εξαρτάται η πιθανότητα CHD από την ηλικία στο δείγμα μας → **υπολογισμός % ασθενών ανά ηλικιακή ομάδα**

		Diseased	
Age group	**# in group**	**#**	**%**
20 - 29	**5**	**0**	**0**
30 - 39	**6**	**1**	**17**
40 - 49	**7**	**2**	**29**
50 - 59	**7**	**4**	**57**
60 - 69	**5**	**4**	**80**
70 - 79	**2**	**2**	**100**
80 - 89	**1**	**1**	**100**

Αντίστοιχο σχήμα



Λογαριθμιστική (logistic) συνάρτηση



Λογαριθμιστική (logistic) συνάρτηση

- Η **πιθανότητα της νόσου π** εξαρτάται από **την ηλικία (Χ)** μέσω της λογαριθμιστικής συνάρτησης

$$\pi = \frac{e^{\alpha+\beta X}}{1+e^{\alpha+\beta X}}$$

- Πώς μπορώ να μετασχηματίσω το παραπάνω σε μία πιο «οικεία» μορφή όπως αυτή της γραμμικής εξάρτησης (Υ=α+βΧ) ;

Μετασχηματισμός

$$\pi = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$$

$$\Rightarrow \frac{\pi}{1-\pi} = e^{\alpha+\beta x}$$

$$\Rightarrow \underbrace{\ln\left(\frac{\pi}{1-\pi}\right)}_{\text{Logit}(\pi)} = \alpha + \beta x$$

Odds και πιθανότητα

- **Πιθανότητα**: π

- **Odds** (**λόγος συμπληρωματικών πιθανοτήτων**)

$$odds = \frac{\text{πιθανότητα να συμβεί}}{\text{πιθανότητα να μη συμβεί}} = \frac{\pi}{1-\pi}$$

- π.χ. πιθανότητα 80% να κερδίσει μία ομάδα

→ odds=0.80/(1-0.80=4/1 → για κάθε 4 νίκες, μία ήττα

- Η έννοια του odds χρησιμοποιείται για να εκφράσει γενικά την πιθανότητα ή τον κίνδυνο (χωρίς να έχει ακριβώς την ίδια ερμηνεία με τις έννοιες αυτές)

Λογαριθμιστική εξάρτηση

Αν X η ανεξάρτητη μεταβλητή:

$$\overbrace{\ln\left(\frac{\pi}{1-\pi}\right)}^{Y} = \alpha + \beta X$$

$$\ln(odds) = \alpha + \beta X$$

- Εξετάζουμε μέσω του β την εξάρτηση της «πιθανότητας» της νόσου (για την ακρίβεια του log-odds) από την X
- Με το μετασχηματισμό αυτό μεταφερόμαστε από εύρος τιμών (0,1) που είχε η πιθανότητα π σε εύρος $(-\infty, +\infty)$

Μέθοδος εκτίμησης συντελεστών

Μέθοδος διαδοχικών προσεγγίσεων της πιθανοφάνειας (iterative maximum likelihood method)

Γραμμική και λογαριθμιστική εξάρτηση

	Γραμμική εξάρτηση	**Λογαριθμιστική εξάρτηση**
Είδος εξαρτημένης μεταβλητής	Ποσοτική με κανονική κατανομή	**Ποιοτική με 2 επίπεδα**
Τι εκφράζει η εξαρτημένη μεταβλητή	Τα επίπεδα της ποσοτικής μεταβλητής (Υ)	**Την πιθανότητα παρουσίας (ή απουσίας) του ποιοτικού χαρακτηριστικού (π)**
Μοντέλο	$\hat{Y}=\alpha+\beta X$	$\ln(\frac{\pi}{1-\pi}) = \alpha + \beta X$

Ερμηνεία του β

$$\ln\left(\frac{\pi}{1-\pi}\right) = a + \beta X$$

- Με αυτή την παραμετροποίηση το μοντέλο «μοιάζει» με αυτό της απλής γραμμικής εξάρτησης

- **Ερμηνεία β**: Αύξηση της X κατά μία μονάδα συνεπάγεται αύξηση του log-odds κατά β

- Πως μπορεί να ερμηνευτεί αυτό με πιο φυσικό τρόπο;
  - Π.χ. θα μας ενδιέφερε αν η αύξηση της ηλικίας συνεπάγεται αύξηση του odds (κινδύνου, πιθανότητας) για CHD (και όχι του log-odds)

Στο παράδειγμα της CHD: Δύο άτομα με ηλικία που διαφέρει κατά 1 έτος

$$\ln\left(\frac{\pi_1}{1-\pi_1}\right) = a + \beta X$$

$$\ln\left(\frac{\pi_2}{1-\pi_2}\right) = a + \beta(X+1)$$

$$\ln\left(\frac{\pi_2}{1-\pi_2}\right) - \ln\left(\frac{\pi_1}{1-\pi_1}\right) = a + \beta(X+1) - a - \beta X$$

$$\Rightarrow \ln\left(\frac{\pi_2}{1-\pi_2}\right) - \ln\left(\frac{\pi_1}{1-\pi_1}\right) = \beta$$

$$\Rightarrow \ln\left(\frac{\frac{\pi_2}{1-\pi_2}}{\frac{\pi_1}{1-\pi_1}}\right) = \beta$$

Odds να έχει CHD το 2ο άτομο ←

Odds να έχει CHD το 1ο άτομο ←

$$\Rightarrow \frac{\frac{\pi_2}{1-\pi_2}}{\frac{\pi_1}{1-\pi_1}} = \text{Odds Ratio (OR)} = e^{\beta}$$

Στο παράδειγμα της CHD: Δύο άτομα με ηλικία που διαφέρει κατά 1 έτος

$$\text{Odds}_2 \longleftarrow \frac{\pi_2}{1-\pi_2} = OR = e^{\beta}$$

$$\text{Odds}_1 \longleftarrow \frac{\pi_1}{1-\pi_1}$$

Π.χ. $e^{\beta}=1.2 \rightarrow \text{odds}_2=1.2*\text{odds}_1$

→Αύξηση της ηλικίας κατά μία μονάδα συνεπάγεται **1.2 φορές μεγαλύτερο odds CHD**

(ένα άτομο έχει 1.2 φορές μεγαλύτερη «πιθανότητα» CHD σε σχέση με ένα άλλο που είναι ένα έτος μικρότερο)

Ερμηνεία του OR (e^{β})

- Αν η ηλικία δεν παίζει ρόλο (β=0)

$$\frac{\frac{\pi_2}{1-\pi_2}}{\frac{\pi_1}{1-\pi_1}} = OR = e^{\beta} \cong 1$$

- Αν όσο αυξάνει η ηλικία αυξάνει ο κίνδυνος CHD (β>0):

$$\frac{\frac{\pi_2}{1-\pi_2}}{\frac{\pi_1}{1-\pi_1}} = OR = e^{\beta} > 1$$

- Αν όσο αυξάνει η ηλικία μειώνεται ο κίνδυνος CHD (β<0):

$$\frac{\frac{\pi_2}{1-\pi_2}}{\frac{\pi_1}{1-\pi_1}} = OR = e^{\beta} < 1$$

Ερμηνεία του OR (e^{β})

- **Αν OR κοντά στο 1**: παρόμοιος κίνδυνος παρουσίας της νόσου σε άτομα που διαφέρει η ηλικία τους κατά 1 έτος

- **Αν OR>1**, π.χ. OR=1.2: ένα άτομο έχει 1.2 φορές μεγαλύτερη «πιθανότητα» CHD σε σχέση με ένα άλλο που είναι ένα έτος μικρότερο
  - **Εναλλακτικά**: ένα άτομο έχει 20% μεγαλύτερη πιθανότητα CHD σε σχέση με ένα άλλο που είναι ένα έτος μικρότερο

    1.2-1=0.2=20%

- **Αν OR<1,** π.χ. OR=0.75: ένα άτομο έχει 0.75 φορές μικρότερη «πιθανότητα» CHD σε σχέση με ένα άλλο που είναι ένα έτος μικρότερο
  - **Εναλλακτικά**: ένα άτομο έχει 25% μικρότερη πιθανότητα CHD σε σχέση με ένα άλλο που είναι ένα έτος μικρότερο

    0.75-1=-0.25=-25%

Ερμηνεία του συντελεστή β στη γραμμική και λογαριθμιστική εξάρτηση

	Σχέση με ανεξάρτητη μεταβλητή Χ		
	Απουσία σχέσης	**Θετική σχέση**	**Αρνητική σχέση**
Γραμμική εξάρτηση	**$\beta=0$**	**$\beta>0$**	**$\beta<0$**
Λογαριθμιστική εξάρτηση	**$\beta=0$ ή $e^{\beta}=1$**	**$\beta>0$ ή $e^{\beta}>1$**	**$\beta<0$ ή $e^{\beta}<1$**

Πολλαπλή λογαριθμιστική εξάρτηση

Αν $X_1, X_2, \ldots, X_p$ p ανεξάρτητες μεταβλητές

Μοντέλο:

$$\ln\left(\frac{\pi}{1-\pi}\right) = \log(odds) = \alpha + b_1 X_1 + b_2 X_2 + \ldots + b_p X_p$$

- Παρόμοια ερμηνεία με πολλαπλή γραμμική εξάρτηση:

  Αύξηση του log-odds για μία μονάδα αύξησης της X_ι όταν οι υπόλοιπες μεταβλητές είναι σταθερές (ανεξάρτητα δηλαδή από τις υπόλοιπες μεταβλητές)

Ποιοτικές ανεξάρτητες μεταβλητές

- Όπως στην πολλαπλή γραμμική παλινδρόμηση, οι ανεξάρτητες μεταβλητές μπορεί να είναι
  - Ποσοτικές
  - Ποιοτικές

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

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. παρουσία CHD σε άνδρες σε σχέση με γυναίκες → **κατηγορία αναφοράς:** γυναίκες

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

- Ο συντελεστής e^b για το φύλο εκφράζει **πόσο αυξημένο (ή μειωμένο) κίνδυνο έχουν οι άνδρες σε σχέση με τις γυναίκες**

Ποιοτικές μεταβλητές με >2 επίπεδα

Όπως στην πολλαπλή γραμμική εξάρτηση: Δημιουργία ψευδομεταβλητών (dummy variables- indicator variables)

Παράδειγμα: Για την επίδραση της φυλής στον κίνδυνο CHD

1: White / 2: Black / 3: Hispanic / 4: Other

Η μεταβλητή έχει 4 επίπεδα → 3 ψευδομεταβλητές στο μοντέλο

Π.χ. αν κατηγορία αναφοράς: White

X_1 = 1 black

= 0 not black

X_2 = 1 hispanic

= 0 not hispanic

X_3 = 1 other

= 0 not other

Αν γνωρίζουμε τις 3 αυτές τις μεταβλητές, μπορούμε να προσδιορίσουμε πλήρως τη φυλή

(π.χ. αν $X_1=X_2=X_3=0$ →λευκή)

Πως εισάγουμε ποιοτικές μεταβλητές στο μοντέλο;

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

  - **Black vs. white** → e^{b_1} δείχνει πόσο αυξημένο/μειωμένο κίνδυνο CHD έχουν οι μαύροι σε σχέση με λευκούς

  - **Hispanic vs. white** → e^{b_2} δείχνει πόσο αυξημένο/μειωμένο κίνδυνο CHD έχουν οι hispanic σε σχέση με λευκούς

  - **Other vs. white** → e^{b_3} δείχνει πόσο αυξημένο/μειωμένο κίνδυνο CHD έχουν οι άλλοι σε σχέση με λευκούς

Παράδειγμα: Φυλή και CHD

Race	Odds ratio (e^b)
White	Κατηγορία αναφοράς
Black	8.0
Hispanic	6.0
Other	4.0

Εναλλακτικά μπορούμε να παρουσιάσουμε τον πίνακα όπως παρακάτω:

Race	Odds ratio
Black/White	8.0
Hispanic/White	6.0
Other/White	4.0

- Π.χ. οι μαύροι έχουν 8 φορές μεγαλύτερο κίνδυνο CHD σε σχέση με τους λευκούς

Στατιστική αξιολόγηση συντελεστών λογαριθμιστικής εξάρτησης

- Το β μπορεί να αξιολογηθεί μέσω του $\frac{b}{SE_b}$

- Αξιολογείται στους πίνακες της **κανονικής κατανομής** ή **t κατανομής** στους **άπειρους** Β.Ε.

- Μηδενική υπόθεση

H_0: $\beta=0$
H_1: $\beta\neq0$

Ή εναλλακτικά

H_0: $e^{\beta}=1$
H_1: $e^{\beta}\neq1$

- Η αξιολόγηση επίσης μπορεί να γίνει και μέσω των 95% ορίων αξιοπιστίας του b ή του odds ratio e^b
- 95% όρια αξιοπιστίας του OR $e^{b\pm1.96SE_b}$

Παράδειγμα

Race	Odds ratio (e^b)	95% CI
White	Κατηγορία αναφοράς	
Black	8.0	(2.3, 27.6)
Hispanic	6.0	(1.7, 21.3)
Other	4.0	(1.1, 14.9)

- Απουσία συσχέτισης: b=0 → e^b=1
- Αν το 95% ΔΕ του odds ratio περιλαμβάνει το 1 → όχι στατιστικά σημαντικό

Παράδειγμα

Race	Odds ratio (e^b)	95% CI	p
White	Κατηγορία αναφοράς		
Black	8.0	(2.3, 27.6)	0.001
Hispanic	6.0	(1.7, 21.3)	0.006
Other	4.0	(1.1, 14.9)	0.039

Αποτελέσματα πολλαπλής λογαριθμιστικής εξάρτησης για τη διερεύνηση του ρόλου ορισμένων μεταβλητών στην εμφάνιση εμφράγματος του μυοκαρδίου. Μελέτη 234 ασθενών και 1742 μαρτύρων

Ανεξάρτητη μεταβλητή	Συντελεστής μερικής εξάρτησης (b_i)	Πιθανό σφάλμα $SE(b_i)$	Πηλίκο $b_i/SE(b_i)$	OR $Exp(b_i)$
X_1: Χρήση αντισυλληπτικών (0: Όχι, 1: Ναι)	b_1: 1,188	0,261	4,552	3,281
X_2: Ηλικία (έτη)	b_2: **0,152**	0,014	10,857	1,164
X_3: Κάπνισμα 1-24 τσιγ./ημ. (0: Όχι, 1: Ναι)	b_3: 1,125	0,210	5,357	3,080
X_4: Κάπνισμα ≥25 τσιγ./ημ. (0: Όχι, 1: Ναι)	b_4: 2,137	0,209	10,225	8,474

OR ανά **5-ετή αύξηση** της ηλικίας:
exp(0,152*5)=exp(0.760)=2.138

Λογαριθμιστική εξάρτηση και πίνακες συνάφειας

- Έχουμε δει ότι η σχέση 2 ποιοτικών μεταβλητών μπορεί να διερευνηθεί με την εφαρμογή του X^2 τεστ

- Επίσης, αν υπάρχει μία τρίτη μεταβλητή που είναι πιθανός συγχυτικός παράγοντας (π.χ. κάπνισμα στη σχέση καφέ-καρκίνου πνεύμονα), κάνουμε διάστρωση, δηλ. πίνακες για τη σχέση καφέ-καρκίνου χωριστά για καπνιστές και μη καπνιστές

- Τα παραπάνω μπορούν να υπολογιστούν και με τη λογαριθμιστκή εξάρτηση

Λογαριθμιστική εξάρτηση και πίνακες συνάφειας

	Exposed	Unexposed	Total
Disease	a	b	a+b
No disease	c	d	c+d
Total	a+c	b+d	*n*

- Υπάρχει σχέση μεταξύ ασθένειας και έκθεσης σε κάποιον παράγοντα; (π.χ. καρκίνος και κάπνισμα)

- Το χ^2 τεστ δεν ποσοτικοποιεί τη σχέση ούτε μας δείχνει την κατεύθυνση.

  → Σχετικός λόγος ή odds ratio.

Odds ratio (σχετικός λόγος)

	Exposed	Unexposed	Total
Disease	a	b	a+b
No disease	c	d	c+d
Total	a+c	b+d	*n*

Εκτίμηση σχετικού λόγου: $\mathrm{OR} = \dfrac{\text{odds of disease in exposed}}{\text{odds of disease in unexposed}}$

$$\Rightarrow \mathrm{OR} = \frac{\dfrac{\pi_{\exp}}{1-\pi_{\exp}}}{\dfrac{\pi_{\text{un}\exp}}{1-\pi_{\text{un}\exp}}} \quad \Rightarrow \mathrm{OR} = \frac{\dfrac{a/(a+c)}{c/(a+c)}}{\dfrac{b/(b+d)}{d/(b+d)}} \quad \boxed{\Rightarrow \mathrm{OR} = \frac{ad}{bc}}$$

Παράδειγμα:

Διερεύνηση παραγόντων που σχετίζονται με αυξημένο κίνδυνο γέννησης λιποβαρούς νεογνού (<2500 kg)

Hosmer & Lemeshow

Συλλογή δεδομένων από 189 γυναίκες και τα νεογνά τους

VARIABLE	VARIABLE LABEL
id	identification code
low	Birth weight<2500g (Yes/No)
age	age of mother (years)
race	Race (White, Black, Other)
smoke	smoked during pregnancy (Yes/No)

Υπό μελέτη έκβαση (ποιοτική μεταβλητή)

Ανεξάρτητες μεταβλητές

Πιθανά ερωτήματα

- Σχέση ηλικίας με πιθανότητα γέννησης λιποβαρούς νεογνού
- Σχέση φυλής με πιθανότητα γέννησης λιποβαρούς νεογνού
- Σχέση καπνίσματος με πιθανότητα γέννησης λιποβαρούς νεογνού

Καθώς και άλλα ερωτήματα, π.χ.

- Η σχέση φυλής με πιθανότητα γέννησης λιποβαρούς νεογνού παραμένει ακόμα και αν λάβουμε υπόψη το κάπνισμα;

Κάπνισμα κατά την εγκυμοσύνη και πιθανότητα γέννησης λιποβαρούς νεογνού: Πίνακας συνάφειας

Κάπνισμα	**Λιποβαρές νεογνό**	
	Ναι	**Όχι**
Ναι	30 (40.5%)	44 (59.5%)
Όχι	29 (25.2%)	86 (74.8%)

- Αξιολόγηση μέσω X^2: $X^2=4.93$

Βαθμοί ελευθερίας: (K-1)(L-1)= (2-1)*(2-1)=1

Αξιολόγηση στους πίνακες της X^2 κατανομής:

ΒΕ	10%	5%	1%	1‰
1	**2,71**	**3,84**	**6,64**	**10,83**

$X^2=4.93>3.84$ (και <6.64)

Άρα $0.01<p<0.05$ → στατιστικά σημαντική σχέση

Από στατιστικό πρόγραμμα: p=0.026

Άρα, οι γυναίκες που καπνίζουν είναι πιο πιθανό να γεννήσουν λιποβαρή νεογνά σε σχέση με τις μη καπνίστριες (40.5% έναντι 25.2%, p<0.05)

Πως θα ερμηνεύσω καλύτερα τη σχέση που υπάρχει;

Κάπνισμα	**Λιποβαρές νεογνό**	
	Ναι	**Όχι**
Ναι	30 (40.5%)	44 (59.5%)
Όχι	29 (25.2%)	86 (74.8%)

$$OR = \frac{30 \bullet 86}{29 \bullet 44} = 2,02$$

Οι καπνίστριες έχουν διπλάσια πιθανότητα γέννησης λιποβαρούς νεογνού σε σχέση με τις μη καπνίστριες

95% ΔΕ του OR: (1.03, 3,97)

Ανάλυση με λογαριθμιστική εξάρτηση

- π: πιθανότητα γέννησης λιποβαρούς νεογνού
  → π/(1-π): odds

- Χ: κάπνισμα (ναι/όχι)
  - Θέλω να συγκρίνω τις γυναίκες που καπνίζουν σε σχέση με τις γυναίκες που δεν καπνίζουν
  - Χ=1 → καπνίζουν
  - Χ=0 → δεν καπνίζουν (κατηγορία αναφοράς)

- Μοντέλο: $\ln(\frac{\pi}{1-\pi}) = \alpha + bX$ και εκτιμάται από

  τα δεδομένα:

$$\ln(\frac{\pi}{1-\pi}) = -1.087 + 0.704X$$

Σύγκριση γυναικών που καπνίζουν(Χ=1) σε σχέση με τις γυναίκες που δεν καπνίζουν (Χ=0)

Ο συντελεστής b = 0.704

$e^{0.704}$ = 2.02 → Όμοια με το OR που υπολογίσαμε από τον πίνακα συνάφειας

Επίσης, από τον έλεγχο υποθέσεων H_0: β=0 → **p=0.028**

(αντίστοιχο με το P-value του X^2)

Αξιολόγηση συντελεστών

- Από το μοντέλο δίνεται SE_b=0.320, άρα:

$$\frac{|b|}{SE_b} = \frac{0.704}{0.320} = 2.20$$

t κατανομή στους άπειρους Β.Ε.

ΒΕ	10%	5%	1%	1‰
∞	**1,65**	**1,96**	**2,58**	**3,29**

2,20>1,96
p<0.05

Επομένως, το β διαφέρει στατιστικά σημαντικά από το 0

(ή αντίστοιχα το odds ratio e^{β} διαφέρει από το 1)

→ στατιστικά σημαντική σχέση μεταξύ καπνίσματος & γέννησης λιποβαρούς νεογνού

95% ΟΑ του odds ratio

95%CI:

$$e^{b \pm 1.96 SE_b}$$

exp(0.704±1.96*0.320)

exp(0.078) → **1.08**

exp(1.33) → **3.78**

Επομένως OR (95% CI): 2.02 (1.08, 3.78)
(δεν περιλαμβάνεται η μονάδα)

Η επίδραση της φυλής στη γέννηση λιποβαρούς νεογνού

Η επίδραση του παράγοντα “φυλή” μπορεί να διαπιστωθεί με παρόμοιο τρόπο θεωρώντας τον παρακάτω 3×2 πίνακα

race	birth weight<2500g 0	1	Total
white	73	23	96
	76.04	23.96	100.00
black	15	11	26
	57.69	42.31	100.00
other	42	25	67
	62.69	37.31	100.00

- Αξιολόγηση μέσω X^2: X^2=5.00 με 2 βαθμούς ελευθερίας

→ p=0.082 (οριακά στατιστικά σημαντική σχέση)

Ερμηνεία με τη βοήθεια odds ratios

race	birth weight<2500g 0	1	Total
white	73	23	96
	76.04	23.96	100.00
black	15	11	26
	57.69	42.31	100.00
other	42	25	67
	62.69	37.31	100.00

Φυλή	**Low birth weight**	
	Yes	**No**
Μαύρη	11	15
Λευκή	23	73

Φυλή	**Low birth weight**	
	Yes	**No**
Άλλο	25	42
Λευκή	23	73

$$OR_{Μαύρη/Λευκή} = \frac{11*73}{15*23} = 2.33$$

$$OR_{Άλλο/Λευκή} = \frac{25*73}{42*23} = 1.89$$

Η επίδραση της φυλής (με λογαριθμιστική εξάρτηση)

- Εφόσον η φυλή είναι ένας παράγοντας με τρία επίπεδα, για να εισαχθεί στο μοντέλο απαιτείται η δημιουργία ψευδομεταβλητών. Έτσι, θα δημιουργήσουμε 2 ψευδομεταβλητές X_1 και X_2 όπως φαίνεται παρακάτω (εδώ η φυλή «λευκή» χρησιμοποιείται ως κατηγορία αναφοράς).

Μαύρη φυλή (X_1)
= 1 αν μαύρη φυλή
= 0 διαφορετικά

Άλλη φυλή (X_2)
= 1 αν άλλη φυλή
= 0 διαφορετικά

Το μοντέλο είναι της μορφής: $$\log\left(\frac{\pi}{1-\pi}\right) = a + b_1 X_1 + b_2 X_2$$

Αποτελέσματα και ερμηνεία

$$\ln\left(\frac{\pi}{1-\pi}\right) = -1.15 + 0.844X_1 + 0.636X_2$$

Ερμηνεία:

exp(0.844)=2.33

Οι γυναίκες μαύρης φυλής είναι 2.33 φορές πιο πιθανό να γεννήσουν λιποβαρές νεογνό σε σχέση με τις λευκές γυναίκες.

exp(0.636)=1.89

Οι γυναίκες άλλης φυλής είναι 1.89 φορές πιο πιθανό να γεννήσουν λιποβαρές νεογνό σε σχέση με τις λευκές γυναίκες.

Ίδια odds ratios με αυτά που υπολογίσαμε πριν από τους πίνακες συνάφειας

Αποτελέσματα εφαρμογής Logistic regression για την επίδραση της φυλής στην πιθανότητα γέννησης λιποβαρούς νεογνού

low	Coef.	Std. Err.	z*	P>|z|	[95% Conf.	Interval]
race						
black	.8448103	.4634141	1.82	0.068	-.0634647	1.753085
other	.6361714	.347831	1.83	0.067	-.0455648	1.317908
_cons	-1.154965	.2391169	-4.83	0.000	-1.623626	-.6863047

* z=coef/SE

Ή με odds ratio (αντί για b):

low	Odds Ratio	Std. Err.	z*	P>|z|	[95% Conf.	Interval]
race						
black	2.327536	1.078613	1.82	0.068	.9385072	5.772385
other	1.889234	.6571342	1.83	0.067	.9554577	3.735597
_cons	.3150685	.0753382	-4.83	0.000	.1971825	.503433

* z=coef/SE

Η σχέση της φυλής με τη γέννηση λιποβαρούς νεογνού είναι ανεξάρτητη από τυχόν διαφορές σε άλλα χαρακτηριστικά που μπορεί να υπάρχουν μεταξύ γυναικών διαφορετικών φυλών;

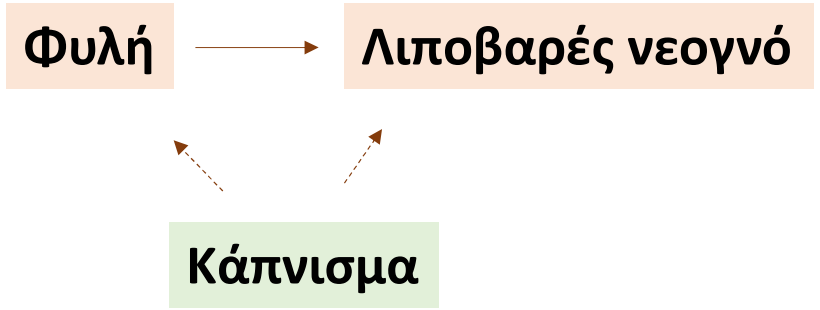
Π.χ. Οι καπνιστικές συνήθειες διαφέρουν μεταξύ γυναικών διαφορετικών φυλών

race	smoked during pregnancy 0	1	Total
white	44 45.83	52 54.17	96 100.00
black	16 61.54	10 38.46	26 100.00
other	55 82.09	12 17.91	67 100.00

p<0.001

Οι λευκές γυναίκες καπνίζουν σε μεγαλύτερο ποσοστό

Αλλά είδαμε πριν ότι το κάπνισμα αυξάνει την πιθανότητα γέννησης λιποβαρους νεογνού [OR (95% CI): 2.02 (1.08, 3.78)]



Κάπνισμα: πιθανός συγχυτικός παράγοντας

Πώς εξετάζω τη σχέση 2 μεταβλητών λαμβάνοντας υπόψη ένα πιθανό συγχυτικό παράγοντα;

- **Στρωματοποιημένη ανάλυση (Stratified analysis):** Εξετάζω τη σχέση φυλής-γέννησης λιποβαρούς νεογνού χωριστά σε καπνίστριες και σε μη καπνίστριες
  - **Ένα odds ratio** για τη σχέση σχέση φυλής (other vs. white)-γέννησης λιποβαρούς νεογνού **για καπνίστριες** και **ένα odds ratio για μη καπνίστριες**
  - Σύνθεση των αποτελεσμάτων για εκτίμηση ενός «συνολικού» odds ratio το οποίο λαμβάνει υπόψη (διορθώνει) την ηλικία → Mantel-Haenszel:

$$OR_{\mathrm{M-H}} = \frac{\sum a_i d_i / N_i}{\sum b_i c_i / N_i} = 3.05$$

(ενώ χωρίς τη διαστρωμάτωση: other vs. white → Odds ratio = 1.89)

Αξιολόγηση H_0: $OR_{MH} = 1$ μέσω του τεστ M-H.

Ανάλυση μέσω λογαριθμιστικής παλινδρόμησης

Ανάλυση μέσω λογαριθμιστικής παλινδρόμησης: για να ελέγξω τη σχέση φυλής και γέννησης λιποβαρούς νεογνού διορθώνοντας για το κάπνισμα, αρκεί να προσθέσω στο μοντέλο και το κάπνισμα

$$\ln\left(\frac{\pi}{1-\pi}\right) = -1.84 + 1.08 * black + 1.10 * other + 1.12 * smoke$$

→Διορθωμένος OR για τη σχέση φυλής (other vs. white):

exp(1.10)=3.00

→όπως και μέσω στρωματοποιημένης ανάλυσης κατά M-H.

Άρα:

<u>Αδρό (unadjusted)</u>	**OR = 1.89, 95% CI: (0.96, 5.77), p=0.067**
<u>Διορθωμένο ως προς το κάπνισμα</u>	
(adjusted for smoking)	**OR = 3.00, 95% CI: (1.38, 6.64), p=0.006**

Σύγκριση αποτελεσμάτων της λογαριθμιστικής εξάρτησης για τη σχέση φυλής –πιθανότητας γέννησης λιποβαρούς νεογνού χωρίς και με το κάπνισμα

Χωρίς το κάπνισμα

low	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
race						
black	2.327536	1.078613	1.82	0.068	.9385072	5.772385
other	1.889234	.6571342	1.83	0.067	.9554577	3.735597
_cons	.3150685	.0753382	-4.83	0.000	.1971825	.503433

.

Διορθώνοντας για τις διαφορές ως προς το κάπνισμα

low	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
race						
black	2.956742	1.448759	2.21	0.027	1.131716	7.724838
other	3.030001	1.212927	2.77	0.006	1.382616	6.64024
smoke	3.052631	1.127112	3.02	0.003	1.480432	6.294487
_cons	.1587319	.0560108	-5.22	0.000	.0794888	.3169732

Ερμηνεία

- Οι γυναίκες άλλης φυλής έχουν 1.89 φορές μεγαλύτερη πιθανότητα να γεννήσουν λιποβαρές νεογνό από τις λευκές

- **Όμως**, οι λευκές γυναίκες κάπνιζαν σε μεγαλύτερο ποσοστό από τις other & το κάπνισμα είναι παράγοντας κινδύνου

- Στη σύγκριση μεταξύ φυλών, εμπλέκονται και οι διαφορές που έχουν οι 2 φυλές ως προς τις καπνιστικές συνήθειες

- Αν λάβουμε υπόψη τις διαφορές στις καπνιστικές συνήθειες, οι other έχουν ακόμα μεγαλύτερο κίνδυνο από τις λευκές

Παράδειγμα:
Διερευνώντας την επίδραση της κατανάλωσης καφέ και του καπνίσματος στον καρκίνο του πνεύμονα

Καφές & καρκίνος του πνεύμονα

coffee	lung No	lung Yes	Total
No	72 69.23	32 30.77	104 100.00
Yes	33 32.67	68 67.33	101 100.00
Total	105 51.22	100 48.78	205 100.00

Pearson chi2(1) = 27.4077 Pr = 0.000

Odds ratio για καρκίνο πνεύμονα (κατανάλωση καφέ vs. όχι καφές):
4.64 95% CI: (2.47, 9.73), p<0.001

Με διάστρωση ανάλογα με το κάπνισμα

Μη καπνιστές

coffee	lung: No	Yes	Total
No	64 79.01	17 20.99	81 100.00
Yes	18 72.00	7 28.00	25 100.00
Total	82 77.36	24 22.64	106 100.00

Pearson chi2(1) = 0.5363 Pr = 0.464

Πρώην καπνιστές

coffee	lung: No	Yes	Total
No	5 41.67	7 58.33	12 100.00
Yes	8 50.00	8 50.00	16 100.00
Total	13 46.43	15 53.57	28 100.00

Pearson chi2(1) = 0.1915 Pr = 0.662

Καπνιστές

coffee	lung: No	Yes	Total
No	3 27.27	8 72.73	11 100.00
Yes	7 11.67	53 88.33	60 100.00
Total	10 14.08	61 85.92	71 100.00

Pearson chi2(1) = 1.8709 Pr = 0.171

Mantel-Haneszel OR
1.39
95% CI: (0.66, 2.93)

Με λογαριθμιστική εξάρτηση

Χωρίς διόρθωση για κάπνισμα

lung	Odds ratio	Std. err.	z	P>|z|	[95% conf.	interval]
coffee	4.636364	1.392051	5.11	0.000	2.573993	8.351174
_cons	.4444445	.0944263	-3.82	0.000	.2930704	.6740049

Διορθώνοντας για το κάπνισμα

lung	Odds ratio	Std. err.	z	P>|z|	[95% conf.	interval]
coffee	1.412601	.5477896	0.89	0.373	.6605924	3.020683
smoking						
Ex	3.534626	1.627404	2.74	0.006	1.433625	8.714682
Current	17.08341	7.919247	6.12	0.000	6.886338	42.37997
_cons	.2682142	.0683825	-5.16	0.000	.1627286	.4420788