Το υπόδειγμα της λογιστικής παλινδρόμησης στη διαγνωστική έρευνα

1. Εισαγωγή
2. Σχετικός κίνδυνος
3. Λόγος των odds
4. Logit
5. Υπόδειγμα μονομεταβλητής λογιστικής παλινδρόμησης
6. Υπόδειγμα πολυμεταβλητής λογιστικής παλινδρόμησης

ΑΡΧΕΙΑ ΕΛΛΗΝΙΚΗΣ ΙΑΤΡΙΚΗΣ 2004, 21(2):172–178
ARCHIVES OF HELLENIC MEDICINE 2004, 21(2):172–178

Θ. Κατοστάρας

Τμήμα Νοσηλευτικής, Πανεπιστήμιο Αθηνών, Αθήνα

The logistic regression model in diagnostic research

Abstract at the end of the article

Λέξεις ευρετηρίου

Διαγνωστική έρευνα
Ιατρική έρευνα
Λογιστική εξάρτηση
Λογιστική παλινδρόμηση
Πολυμεταβλητή ανάλυση

Υποβλήθηκε 3.7.2003
Εγκρίθηκε 16.7.2003

1. ΕΙΣΑΓΩΓΗ

Ένας άνδρας ηλικίας 50 ετών με μακρύ ιστορικό καπνίσματος προσέρχεται με αιμόπτυση. Με αυτό το ιστορικό και με το συγκεκριμένο κλινικό εύρημα, ποια είναι η πιθανότητα να πάσχει από καρκίνο του πνεύμονα;

Τέτοιου είδους ερωτήματα απασχολούν συνεχώς τους ιατρούς που έχουν την ευθύνη της διάγνωσης του νοσήματος, δηλαδή του εντοπισμού του, βασισμένοι σε μερικά χαρακτηριστικά. Αυτά τα χαρακτηριστικά ονομάζονται *προσδιοριστές* (determinant of occurrence) και αποτελούνται από τα συμπτώματα, τα σημεία της κλινικής εικόνας, τα εργαστηριακά ευρήματα που προκύπτουν από τον έλεγχο του ασθενούς, καθώς και από τα χαρακτηριστικά του ατόμου που συνδέονται με τη συχνότητα του νοσήματος. Η διάγνωση, μαζί με την αιτιόγνωση και την πρόγνωση, αποτελούν τις τρεις συνιστώσες του γνωστικού μέρους της φροντίδας υγείας (ΦΥ). Η ΦΥ είναι το τεχνολογικό μέρος της επιστημονικής Ιατρικής και περιλαμβάνει την πρόληψη, τη θεραπεία και την αποκατάσταση.

Το υπόδειγμα (πρότυπο, μοντέλο) της *λογιστικής ή λογαριθμιστικής παλινδρόμησης* (logistic regression model) είναι η σημαντικότερη στατιστική μέθοδος που χρησιμοποιείται κυρίως στην αιτιογνωστική, αλλά και στη διαγνωστική και στην προγνωστική έρευνα. Στην αιτιογνωστική έρευνα, εκτιμάται η πιθανότητα η ύπαρξη ενός νοσήματος (καρκίνος του στοματοφάρυγγα) να συσχετίζεται με την παρουσία διαφόρων προσδιοριστών (ένταση καπνίσματος, ποσότητα κατανάλωσης αλκοόλ) και, ακόμα, εκτιμάται ο βαθμός επικινδυνότητας ή προστασίας διαφόρων παραγόντων στην εμφάνιση ή μη μιας νόσου. Στη διαγνωστική έρευνα εκτιμάται η πιθανότητα της παρουσίας ενός νοσήματος (φυματίωση) όταν υπάρχουν μερικοί προσδιοριστές (βήχας, δεκατική πυρετική κίνηση από μηνός, αιμόφυρτα πτύελα από τριημέρου). Στην προγνωστική έρευνα εκτιμάται η πιθανότητα που έχει ένας ασθενής, ο οποίος πάσχει από ένα ορισμένο νόσημα (καρκίνο του μαστού) και έχει κάποιους προσδιοριστές (διόγκωση μασχαλιαίων λεμφαδένων), να εκδηλώσει μια συγκεκριμένη έκβαση (θάνατος) σε μια χρονική στιγμή της πορείας της νόσου.

Στη διαγνωστική έρευνα, η εκτίμηση με επιστημονικό τρόπο της πιθανότητας να υπάρχει ένα νόσημα, γίνεται με την εφαρμογή της μεθόδου της λογιστικής παλινδρόμησης. Η μέθοδος αυτή είναι η μόνη έγκυρη διαγνωστική μέθοδος και υπερτερεί της διαγνωστικής διαδικασίας που μπορεί να πραγματοποιηθεί με εφαρμογή του θεωρήματος του Bayes, γιατί, εκτός των άλλων, δεν απαιτεί την *a priori* πιθανότητα της ύπαρξης του νοσήματος, η οποία όμως είναι αναγκαία κατά την εφαρμογή της Μπαγεσιανής θεωρίας.

Η εκτίμηση της διαγνωστικής πιθανότητας γίνεται με τη μελέτη του επιπολασμού του νοσήματος «σε μια χρονι-

κή στιγμή» σε ένα μελετώμενο πληθυσμό, ο οποίος είναι δυνατό να είναι κλειστός ή ανοικτός.

Κλειστός ή σταθερός ή στατικός πληθυσμός ή κοόρτη (cohort) είναι κάθε προκαθορισμένο σύνολο ατόμων, στο οποίο επιτρέπεται η είσοδος και όπου η ιδιότητα του μέλους καθορίζεται από ένα συμβάν σε μια συγκεκριμένη περιοχή και η οποία ιδιότητα δεν χάνεται ούτε με το θάνατο του ατόμου. *Ανοικτός ή δυναμικός πληθυσμός* (dynamic) είναι εκείνος, στον οποίο η ιδιότητα του μέλους προσδιορίζεται από μια κατάσταση και διαρκεί όσο διαρκεί η κατάσταση.

Κάθε μέλος ενός μελετώμενου πληθυσμού είναι δυνατό να έχει (N+) ή να μην έχει (N–) ένα νόσημα και να έχει ή να μην έχει έναν προσδιοριστή. Οι διαφορετικές καταστάσεις της ύπαρξης ή μη, καθώς και της έντασης του προσδιοριστή, αποδίδονται με τις τιμές μιας μεταβλητής κατά την πραγματοποίηση ενός διαγνωστικού ελέγχου. Είναι δυνατό να οριστεί ένα διαχωριστικό όριο στις τιμές της μεταβλητής και να οριστεί ένα σημείο, πέραν του οποίου οι τιμές της μεταβλητής να θεωρούνται «παθολογικές» ή «θετικές» (T+) εάν σχετίζονται με υπάρχουσα νόσο και κάτω του οποίου οι τιμές να θεωρούνται «φυσιολογικές» ή «αρνητικές» (T–). Ο ορισμός του σημείου είναι κρίσιμης σημασίας, γιατί καθορίζει την ικανότητα της δοκιμασίας να διακρίνει τα άτομα που πάσχουν από ένα συγκεκριμένο νόσημα, από τα άτομα που δεν πάσχουν από το νόσημα.

Μεταξύ των ατόμων του πληθυσμού που έχουν έναν προσδιοριστή, έστω ότι το P είναι ο επιπολασμός ενός νοσήματος, δηλαδή το P είναι το ποσοστό των ατόμων που πάσχουν από το συγκεκριμένο νόσημα, και το 1–P είναι το ποσοστό των ατόμων που δεν πάσχουν από το νόσημα. Το P είναι επίσης η πιθανότητα να επιλεγεί ένα άτομο που πάσχει από το νόσημα, μεταξύ των μελών του πληθυσμού που έχουν τον προσδιοριστή.

Αν $P=\kappa(1-P)$, δηλαδή $P/(1-P)=\kappa$
το κ ονομάζεται *οτz* του *P* και συμβολίζεται και με *odds* του *P* ή *οτz (P)* ή *odds (P)*.

Για $\kappa=1$ ισχύει $P=1-P$. Δηλαδή, μεταξύ των ατόμων στον πληθυσμό που έχουν τον προσδιοριστή, το ποσοστό των ατόμων που έχουν το συγκεκριμένο νόσημα είναι ίσο με το ποσοστό των ατόμων που δεν το έχουν. Αν $\kappa<1$ είναι $P<1-P$, δηλαδή μεταξύ των ατόμων στον πληθυσμό που έχουν τον προσδιοριστή, το ποσοστό των ατόμων που έχουν το συγκεκριμένο νόσημα είναι μικρότερο από το ποσοστό των ατόμων που δεν το έχουν και, επομένως, ο προσδιοριστής είναι δυνατό να σχετίζεται αρνητικά με την εμφάνιση του νοσήματος. Αν $\kappa>1$ είναι $P>1-P$, δηλαδή μεταξύ των ατόμων στον πληθυσμό που έχουν τον προσδιοριστή, το ποσοστό των ατόμων που έχουν το συγκεκριμένο νόσημα είναι μεγαλύτερο από το ποσοστό των ατόμων που δεν το έχουν και, επομένως, ο προσδιοριστής είναι δυνατό να συσχετίζεται θετικά με την ύπαρξη του νοσήματος.

Προφανώς, *οτz(1–P)=1/κ*.

2. ΣΧΕΤΙΚΟΣ ΚΙΝΔΥΝΟΣ

Τα κελιά του πίνακα 1 περιέχουν τις απόλυτες συχνότητες της ύπαρξης ή μη ενός νοσήματος και του αποτελέσματος ενός διαγνωστικού ελέγχου.

Από μελέτη 96 ασθενών με στρεπτοκοκκική (β-αιμολυτικός στρεπτόκοκκος ομάδας Α) ή μη κυνάγχη, με σκοπό την εκτίμηση της διαγνωστικής ποιότητας της καλλιέργειας του φαρυγγικού επιχρίσματος στη διάγνωση της στρεπτοκοκκικής κυνάγχης, προέκυψαν τα αποτελέσματα των καλλιεργειών που περιλαμβάνονται στον πίνακα 1. Ως μέθοδος αναφοράς για τη διάκριση των ατόμων σε πάσχοντες και μη από στρεπτοκοκκική κυνάγχη χρησιμοποιήθηκε ο τίτλος των αντιστρεπτολυσινών του ορού.

$P1=P(N+/T+)=a/(a+b)=27/47=0,574$ είναι το ποσοστό των ατόμων που είχαν το νόσημα, από τα άτομα που είχαν θετική την εξέταση, και *ονομάζεται θετική διαγνωστική αξία* (ΘΔΑ).

$Q1=1-P1=P(N-/T+)=b/(a+b)=20/47=0,426$ είναι το ποσοστό των ατόμων που δεν είχαν το νόσημα, από τα άτομα που είχαν θετική την εξέταση, και ονομάζεται *θετικό διαγνωστικό σφάλμα* (ΘΔΣ).

Πίνακας 1. Αποτελέσματα του ελέγχου ύπαρξης στρεπτοκοκκικής κυνάγχης.

	Παρουσία κυνάγχης (N+)	**Απουσία κυνάγχης (N–)**	
Θετικό αποτέλεσμα καλλιέργειας (T+)	a=27	b=20	a+b=47
Αρνητικό αποτέλεσμα καλλιέργειας (T–)	c=3	d=46	c+d=49
	a+c=30	b+d=66	a+b+c+d=96

P2=P(N+/T–)=c/(c+d)=3/49=0,061 είναι το ποσοστό των ατόμων που είχαν το νόσημα, από τα άτομα που είχαν αρνητική την εξέταση, και ονομάζεται *αρνητικό διαγνωστικό σφάλμα* (ΑΔΣ).

Q2=1–P2=P(N–/T–)=d/(c+d)=46/49=0,939 είναι το ποσοστό των ατόμων που δεν είχαν το νόσημα, από τα άτομα που είχαν αρνητική την εξέταση, και ονομάζεται *αρνητική διαγνωστική αξία* (ΑΔΑ).

Ο *σχετικός κίνδυνος* (relative risk) στη διαγνωστική έρευνα ορίζεται ως το πηλίκο της θετικής διαγνωστικής αξίας προς το αρνητικό διαγνωστικό σφάλμα, συμβολίζεται με *RR* και υπολογίζεται ως εξής:

$$RR=P1/P2=P(N+/T+)/P(N+/T–)=$$
$$a/(a+b)/c/(c+d)=0,574/0,061=9,38$$

Το αποτέλεσμα σημαίνει ότι η συχνότητα της στρεπτοκοκκικής κυνάγχης βρέθηκε 9 φορές μεγαλύτερη στους πάσχοντες από το νόσημα, οι οποίοι είχαν θετικό το αποτέλεσμα της καλλιέργειας του φαρυγγικού επιχρίσματος, σε σχέση με τους πάσχοντες που είχαν αρνητικό το αποτέλεσμα της καλλιέργειας.

Ο σχετικός κίνδυνος στη διαγνωστική έρευνα αποτελεί δείκτη της διακριτικής ικανότητας του προσδιοριστή.

Αν *RR=r=1*, τότε το ποσοστό των ατόμων που πάσχουν από το νόσημα είναι το ίδιο στα άτομα με θετική την εξέταση και στα άτομα με αρνητική την εξέταση. Για *RR=r<1*, το ποσοστό των ατόμων που πάσχουν από το νόσημα είναι μικρότερο κατά r φορές στα άτομα που έχουν τον προσδιοριστή, σε σχέση με τα άτομα που δεν τον έχουν. Όταν *RR=r>1*, το ποσοστό ατόμων που πάσχουν από το νόσημα είναι μεγαλύτερο κατά r φορές στα άτομα που έχουν τον προσδιοριστή, σε σχέση με τα άτομα που δεν τον έχουν. Επομένως, είναι σημαντικό να γνωρίζει ο ερευνητής το σχετικό κίνδυνο για διάφορους προσδιοριστές, οι οποίοι και εκφράζουν τα διάφορα συμπτώματα και σημεία της κλινικής εικόνας, καθώς και τα αποτελέσματα των εργαστηριακών δοκιμασιών.

Το οτz της θετικής διαγνωστικής αξίας είναι:

$$οτz\ (P1)=P1/(1–P1)=a/b=27/20=1,350$$

Το οτz του αρνητικού διαγνωστικού σφάλματος είναι:

$$οτz\ (P2)=P2/(1–P2)=c/d=3/46=0,065$$

3. ΛΟΓΟΣ ΤΩΝ ODDS

Ο λόγος του οτz (odds ratio) της θετικής διαγνωστικής αξίας προς το οτz του αρνητικού διαγνωστικού σφάλματος είναι:

$$OR=οτz(P1)/οτz(P2)=ad/bc=20,7$$

Δηλαδή, το οτz της θετικής διαγνωστικής αξίας είναι 21 περίπου φορές μεγαλύτερο από το οτz του αρνητικού διαγνωστικού σφάλματος. Όταν το νόσημα είναι σπάνιο, ο επιπολασμός του νοσήματος είναι μικρός και ο λόγος των οτz λαμβάνεται ως προσέγγιση του σχετικού κινδύνου να υπάρχει το νόσημα αν υπάρχει ή όχι ο προσδιοριστής, δηλαδή ο σχετικός κίνδυνος *RR* προσεγγιστικά λαμβάνεται ίσος με το πηλίκο των γινομένων των διαγωνίων του πίνακα των απόλυτων συχνοτήτων:

$$OR=ad/bc=RR$$

Η τελευταία προσεγγιστική σχέση δικαιολογεί τη χρησιμοποίηση του λόγου των οτz για την εκτίμηση της πιθανότητας ύπαρξης ενός νοσήματος από την ύπαρξη ενός προσδιοριστή. Με την εφαρμογή του υποδείγματος της λογιστικής παλινδρόμησης υπολογίζεται ο *OR* για κάθε προσδιοριστή και προσεγγιστικά εκτιμάται ο σχετικός κίνδυνος της ύπαρξης του νοσήματος, από τη συνύπαρξη διαφόρων προσδιοριστών.

Το *πιθανό (τυπικό) σφάλμα* του λογάριθμου του λόγου των οτz είναι:

$$SE(ln(OR)) = \sqrt{((1/\alpha)+(1+b)+(1/c)+(1/d))} =$$
$$= \sqrt{((1/27)+(1/20)+(1/3)+(1/46))} = 0,665$$

και ένα 95% διάστημα εμπιστοσύνης του δίνεται από τον τύπο:

$$(ln(OR)–1,96\ SE(ln(OR)),\ ln(OR)+1,96\ SE(ln(OR))=$$
$$(ln(20,7)–1,96*0,665,\ ln(20,7)+1,96*0,665)=(1,73,\ 4,33)$$

Με απολογαρίθμηση προκύπτουν 95% όρια εμπιστοσύνης του λόγου των οτz, που προσεγγιστικά λαμβάνονται ως 95% όρια εμπιστοσύνης του σχετικού κινδύνου:

$$(exp(1,73),\ exp(4,33))=(5,62,\ 76,17)$$

4. LOGIT

Ο *φυσικός λογάριθμος* (natural logarithm) του οτz της P συμβολίζεται με:

λότζιτ (P) ή *logit (P)*.

Επομένως:

$$Logit(P1)=ln[(P1)/(1–P1)]=ln(1,35)=0,300$$

$$Logit(P2)=ln\ [(P2)/(1–P2)]=ln(0,065)=–2,73$$

5. ΥΠΟΔΕΙΓΜΑ ΜΟΝΟΜΕΤΑΒΛΗΤΗΣ ΛΟΓΙΣΤΙΚΗΣ ΠΑΛΙΝΔΡΟΜΗΣΗΣ

Το απλούστερο μαθηματικό υπόδειγμα της λογιστικής παλινδρόμησης έχει έναν προσδιοριστή και ονομάζεται *μονομεταβλητή λογιστική παλινδρόμηση* (univariate logistic regression). Η μορφή του είναι:

$$P=exp(A+BX)/(1+exp(A+BX)),$$

όπου P είναι ο επιπολασμός του νοσήματος και A και B είναι οι παράμετροί του που θα εκτιμηθούν.

Μετά από αλγεβρικές πράξεις γίνεται:

$$P/(1-P)=exp(A+BX)$$

όπου *exp(t)* είναι ο αριθμός $e=2,718$ ... υψωμένος στη δύναμη t. Είναι:

$$odds(P)=exp(A+BX)$$

Με λογαρίθμηση των όρων του γίνεται:

$$ln(P/(1-P))=ln(odds(P))=Logit(P)$$ και:
$$Logit(P)=A+BX$$

Η μεταβλητή X είναι δίτιμη, με $X=1$ όταν υπάρχει ο προσδιοριστής και $X=0$ όταν δεν υπάρχει, ενώ το P εκτιμάται με τη *μέση τιμή p* μιας μεταβλητής Υ, με $Y=1$ όταν υπάρχει το νόσημα και $Y=0$ όταν δεν υπάρχει.

Αν η εκτίμηση του υποδείγματος είναι η $Logit(p)=a+\beta X$, χρησιμοποιώντας κατάλληλο στατιστικό πρόγραμμα στα περιεχόμενα του πίνακα 1, προκύπτουν:

$$a=-2,73 \text{ και } \beta=3,03$$

Από αυτές τις εκτιμήσεις προκύπτει η εκτίμηση του σχετικού κινδύνου της ύπαρξης του νοσήματος, σε σχέση με το αποτέλεσμα της καλλιέργειας, καθώς και το αρνητικό διαγνωστικό σφάλμα της δοκιμασίας, γιατί ισχύουν:

$$OR=exp(\beta) \text{ και } ΑΔΣ=exp(a)/(1+exp(a))$$

Είναι:

$$OR=οτz(P1)/οτz(P2)=οτz(p)/οτz(1-p)=ad/bc=20,7$$ και:
$$lnOR=ln20,7=3,03=\beta$$ και $$OR=exp(\beta)=exp(3,03)=20,7$$

Επίσης, είναι:

$$ln(οτz(P2))=ln(οτz(ΑΔΣ))=ln(0,065)=-2,73=a$$ και:
$$ΑΔΣ=exp(-2,73)/(1+exp(-2,73))=0,61$$

Επομένως, από την εκτίμηση του υποδείγματος προκύπτει η εκτίμηση του λόγου των οτz, η οποία, όπως προαναφέρθηκε, αποτελεί εκτίμηση του σχετικού κινδύνου. Επίσης, εκτιμάται το αρνητικό διαγνωστικό σφάλμα της δοκιμασίας.

6. ΥΠΟΔΕΙΓΜΑ ΠΟΛΥΜΕΤΑΒΛΗΤΗΣ ΛΟΓΙΣΤΙΚΗΣ ΠΑΛΙΝΔΡΟΜΗΣΗΣ

Γενικά, το υπόδειγμα της *πολυμεταβλητής λογιστικής (λογαριθμιστικής) παλινδρόμησης* (multiple logistic regression) είναι ένα μη γραμμικό στατιστικό υπόδειγμα με περισσότερους από έναν προσδιοριστές, που έχει τη στοχαστική μορφή:

$$P=exp(B_0+B_1X_1+ \dots +B_κX_κ+ε)/[1+exp(B_0+B_1X_1+ \dots +B_κ X_κ)]$$

Το P είναι το ποσοστό εμφάνισης ενός συγκεκριμένου νοσήματος στον πληθυσμό, δηλαδή το P είναι η πιθανότητα επιλογής ενός ατόμου του πληθυσμού με το υπό μελέτη νόσημα. Οι $X_1, X_2, \dots X_κ$ είναι δίτιμες μεταβλητές προσδιοριστών (ερμηνευτικές μεταβλητές ή ανεξάρτητες μεταβλητές) και μπορεί να προέρχονται από ποσοτικές ή μη ποσοτικές μεταβλητές. Οι ποσοτικές μεταβλητές μετατρέπονται σε δίτιμες, με διαχωριστικό όριο τη διάμεσο ή τη μέση τιμή τους ή με οποιοδήποτε άλλο όριο. Το $ε$ είναι τυχαία μεταβλητή που ακολουθεί την κανονική κατανομή με μέσο ίσο με το μηδέν και σταθερή διακύμανση και η οποία ονομάζεται *τυχαίο σφάλμα* ή *τυχαία απόκλιση ή διαταρακτικός όρος*. Με το τυχαίο σφάλμα συνεκτιμάται η επίδραση επί της πιθανότητας P να υπάρχει το νόσημα, όλων των άλλων άγνωστων ή και μη μετρήσιμων μεταβλητών, εκτός από τις $X_1, X_2, \dots X_κ$. Τα $B_0, B_1, B_2, \dots B_κ$ ονομάζονται παράμετροι του υποδείγματος. Τα $B_1, B_2 \dots B_κ$ ονομάζονται και συντελεστές του, ενώ ο B_0 ονομάζεται σταθερός όρος. Μετά από αλγεβρικές πράξεις, το υπόδειγμα λαμβάνει την ισοδύναμη μορφή:

$$P/(1-P)=exp(B_0+B_1X_1+ \dots +B_κX_κ+w)$$ ή
$$οτz(P)=exp(B_0+B_1X_1+ \dots +B_κX_κ+w)$$

Με αυτή τη μορφή εκφράζεται το οτz της πιθανότητας εμφάνισης ενός νοσήματος ως συνάρτηση πολλών προσδιοριστών. Το w είναι τυχαίο σφάλμα.

Αν λογαριθμιστούν οι όροι του, το υπόδειγμα παίρνει τη μορφή:

$$Logit(P)=B_0+B_1X_1+B_2X_2+ \dots +B_κX_κ+u$$

όπου $Ln(P/(1-P))=Logit(P)$ είναι ο φυσικός λογάριθμος του οτz της P, δηλαδή του λόγου των συμπληρωματικών πιθανοτήτων, της πιθανότητας P να υπάρχει το νόσημα προς την πιθανότητα $1-P$ να μην υπάρχει, και $u=Lnw$ είναι ο φυσικός λογάριθμος του διαταρακτικού όρου w.

Εφαρμόζοντας τη μέθοδο των *διαδοχικών προσεγγίσεων μέγιστης πιθανοφάνειας* (iterative maximum likelihood method) με τα στοιχεία ενός δείγματος λαμ-

βάνεται η εκτίμηση του υποδείγματος, που έχει τη μορφή:

$$p=\exp(b_0+b_1X_1+ \dots +b_\kappa X_\kappa)/(1+\exp(b_0+b_1X_1+ \dots +b_\kappa X_\kappa))$$

ή τη μορφή:

$$logit(p)=b_0+b_1X_1+b_2X_2+\dots+b_\kappa X_\kappa$$

Τα *p* και *logit(p)* αποτελούν αντίστοιχες εκτιμήσεις των *P* και *Logit(P)*.

Τα $b_0, b_1, b_2, \dots b_\kappa$ αποτελούν κατάλληλες εκτιμήσεις των παραμέτρων του υποδείγματος $B_0, B_1, B_2, \dots B_\kappa$.

Κάθε $b_j, j=1, 2, \dots k$ εκτιμά την κατά μέσο όρο μεταβολή του *Logit(P),* δηλαδή του λογάριθμου του λόγου των συμπληρωματικών πιθανοτήτων της ύπαρξης του νοσήματος όταν ο προσδιοριστής είναι παρών ($Xj=1$) και οι υπόλοιποι προσδιοριστές είναι απόντες. Αν αντικατασταθούν οι συγκεκριμένες τιμές των ερμηνευτικών μεταβλητών στο υπόδειγμα που εκτιμήθηκε, εκτιμάται η πιθανότητα εμφάνισης του χαρακτηριστικού.

Όπως και στη μονομεταβλητή λογαριθμιστική παλινδρόμηση, σε κάθε προσδιοριστή με μεταβλητή $Xj=1$, όταν ο προσδιοριστής είναι παρών, και $Xj=0$ όταν δεν είναι, η τιμή *exp(bj)* αποτελεί εκτίμηση του λόγου *OR* των συμπληρωματικών πιθανοτήτων, ο οποίος αποτελεί εκτίμηση του σχετικού κινδύνου *RR* να υπάρχει το νόσημα όταν στο άτομο υπάρχει μόνο ο συγκεκριμένος προσδιοριστής και στην περίπτωση που η έκβαση είναι σπάνια. Αν $bj>0$, είναι και $exp(bj)>1$ και τότε είναι αυξημένη η πιθανότητα να υπάρχει το νόσημα όταν υπάρχει ο προσδιοριστής, ενώ αν $bj<0$, είναι $exp(bj)<1$ και τότε είναι μειωμένη η πιθανότητα να υπάρχει το νόσημα όταν υπάρχει ο προσδιοριστής. Αν $bj=0$, είναι $exp(bj)=1$ και η πιθανότητα να υπάρχει το νόσημα δεν συσχετίζεται με την ύπαρξη του προσδιοριστή.

Στη μονομεταβλητή λογαριθμιστική παλινδρόμηση, ο εκτιμημένος λόγος των συμπληρωματικών πιθανοτήτων του παράγοντα ονομάζεται και *πρωτογενής* (crude) και συμβολίζεται και με *COR*, ενώ στην πολυμεταβλητή λογαριθμιστική παλινδρόμηση ο εκτιμημένος λόγος των συμπληρωματικών πιθανοτήτων κάθε παράγοντα ονομάζεται και *προσαρμοσμένος* (adjusted) και συμβολίζεται και με *AOR*.

Η αξιοπιστία της εκτίμησης *bj* κάθε συντελεστή *Bj* του υποδείγματος ελέγχεται με διάφορες στατιστικές δοκιμασίες. Ένας τέτοιος στατιστικός έλεγχος μπορεί να γίνει με το Wald-test. Αν η τιμή της Wald-statistic είναι σημαντικά μεγαλύτερη του 1, ο συντελεστής είναι στατιστικά σημαντικός και ο αντίστοιχος προσδιοριστής συσχετίζεται με την ύπαρξη ή τη μη ύπαρξη του νοσήματος, κάτω από σχετική βεβαιότητα. Ένα μειονέκτημα του Wald-test είναι ότι, πολλές φορές, χαρακτηρίζει ως στατιστικά μη σημαντικούς προσδιοριστές που είναι στατιστικά σημαντικοί, δηλαδή προκύπτει στατιστικό σφάλμα πρώτου είδους. Αυτό μπορεί να συμβεί όταν η εκτίμηση της διακύμανσης του *bj* είναι μεγάλη. Χρησιμοποιώντας την τιμή της Wald-statistic σε μοντέλα πολυμεταβλητής λογαριθμιστικής παλινδρόμησης υπολογίζεται ο συντελεστής μερικής συσχέτισης *Rj* για κάθε προσδιοριστή *Xj*. Ο συντελεστής αυτός εκτιμά την ένταση της συνεισφοράς κάθε προσδιοριστή στη διαμόρφωση της πιθανότητας εμφάνισης του νοσήματος, ανεξάρτητα από τη συνεισφορά των άλλων προσδιοριστών. Ισχύει: $-1<Rj<1$ και ο *Rj* έχει το πρόσημο του *bj*. Αν $Rj>0$ (αντίστοιχα, $Rj<0$), υπάρχει θετική (αντίστοιχα, αρνητική) συνεισφορά του προσδιοριστή, που γίνεται εντονότερη όσο ο *Rj* πλησιάζει στο *1* (αντίστοιχα, στο *–1*). Αν $Rj=0$, δεν υπάρχει αντίστοιχη συνεισφορά.

Για κάθε παράμετρο του υποδείγματος υπολογίζονται όρια εμπιστοσύνης *(CI)* και, επίσης, υπολογίζονται όρια εμπιστοσύνης για το λόγο *OR* των συμπληρωματικών πιθανοτήτων κάθε προσδιοριστή, τα οποία αποτελούν προσεγγιστικά και τα όρια εμπιστοσύνης του σχετικού κινδύνου *RR* εμφάνισης του νοσήματος αν ο προσδιοριστής είναι παρών.

Αν το *1* περιέχεται στο διάστημα εμπιστοσύνης του *OR*, τότε –κάτω από σχετική βεβαιότητα (συνήθως 95%)– το νόσημα δεν σχετίζεται με τον προσδιοριστή. Αν τα όρια του διαστήματος είναι μεγαλύτερα από το *1* (αντίστοιχα, αν είναι μικρότερα από το *1*), ο προσδιοριστής έχει θετική συσχέτιση (αντίστοιχα, αρνητική συσχέτιση) με την εμφάνιση του νοσήματος, που γίνεται εντονότερη όσο μεγαλώνει η απόσταση του *1* από τα όρια του διαστήματος, σύμφωνα με την ερμηνεία που δόθηκε για τον *RR=r*, και κάτω από σχετική βεβαιότητα.

Υπάρχουν διάφοροι τρόποι για να ελεγχθεί αν το υπόδειγμα της λογιστικής παλινδρόμησης που εκτιμήθηκε είναι κατάλληλο για χρησιμοποίηση. Ένας τρόπος είναι η μελέτη της ικανότητάς του για σωστή ταξινόμηση, υπολογίζοντας διάφορα ποσοστά ταξινόμησης. Αν αντικατασταθούν οι τιμές των μεταβλητών κάθε μονάδας του δείγματος στο υπόδειγμα που εκτιμήθηκε, υπολογίζεται η πιθανότητα να υπάρχει το νόσημα.

Με την πιθανότητα αυτή, οι μονάδες του δείγματος κατατάσσονται ως εξής:

Κατάταξη	Νόσημα	
	Ναι	Όχι
Ναι	F11	F12
Όχι	F21	F22

Ποσοστό σωστά θετικά ταξινομημένων=F11/(F11+F12)
Ποσοστό σωστά αρνητικά ταξινομημένων=F22/(F21+F22)
Ποσοστό σωστά ταξινομημένων=(F11+F22)/(F11+F12+F21+F22)
Ποσοστό λανθασμένα ταξινομημένων=(F12+F21)/(F11+F12+F21+F22)

Αν τα ποσοστά σωστής ταξινόμησης είναι ικανοποιητικά, το υπόδειγμα κρίνεται κατάλληλο.

Όταν το υπόδειγμα της λογιστικής παλινδρόμησης χρησιμοποιείται για τη διάγνωση ενός νοσήματος σε διαφορετικό τόπο-χρόνο, η μεταβλητότητα της συχνότητας του νοσήματος που οφείλεται στη μεταβλητότητα του τόπου-χρόνου (τοπικός επιπολασμός) δεν επηρεάζει την εκτίμηση των συντελεστών του υποδείγματος, αλλά μόνο το σταθερό όρο b_0. Αν έχει εκτιμηθεί ένα υπόδειγμα λογιστικής παλινδρόμησης σε έναν τόπο όπου ο επιπολασμός του νοσήματος εκτιμήθηκε ίσος με p, αυτή η εκτίμηση του υποδείγματος μπορεί να χρησιμοποιηθεί για διάγνωση του νοσήματος σε έναν άλλο τόπο, όπου το νόσημα έχει επιπολασμό p^* και του οποίου το οτz είναι f^* φορές μεγαλύτερο από το οτz του επιπολασμού της μελέτης, δηλαδή $f^*=p^*/(1-p^*)/p/(1-p)$. Αν και στους δύο τόπους ο επιπολασμός των διαφοροδιαγνωστικώς συναφών νοσημάτων είναι ίδιος, τότε ο σταθερός όρος της λογιστικής παλινδρόμησης b_0 θα πρέπει να αντικατασταθεί από τον $b_0^*=b_0+ln(f^*)$.

Αν όμως στους δύο τόπους ο επιπολασμός των διαφοροδιαγνωστικώς συναφών νοσημάτων διαφέρει και το οτz του επιπολασμού καθενός από τα διαφοροδιαγνωστικώς συναφή νοσήματα στο νέο τόπο είναι κατά f^{**} φορές μεγαλύτερο από το αντίστοιχο οτz στον τόπο της μελέτης, τότε ο σταθερός όρος της λογιστικής παλινδρόμησης της μελέτης θα πρέπει να αντικατασταθεί από τον $b_0^*=b_0+ln(f^*/f^{**})$.

ABSTRACT

The logistic regression model in diagnostic research

T. KATOSTARAS

Faculty of Nursing, University of Athens, Athens, Greece

Archives of Hellenic Medicine 2004, 21(2):172–178

The logistic regression model is the most important statistical method used, mainly in etiological research, but also in diagnostic and prognostic research. In diagnostic research with the use of this model, the probability of a disease is estimated according to the presence of other characteristics, called determinants. The logistic regression model is applied in studies on cohort and open populations. Between the members of the population, the ratio of the percentage of the population displaying a specific disease to the percentage of the population without the disease (ratio of the complementary probabilities) is called the odds percentage of the existence of the disease. Given that the determinant is present, the ratio of the odds of the existence probability of a rare disease to the odds of the absence probability of the disease, the odds ratio (OR) is taken approximately as the estimation of the relative risk (RR) of the appearance of the disease, for the given factor. This approximating equation justifies the use of the OR in order to correlate the presentation of a disease with the presence of a determinant. The logistic regression model is a non-linear statistical model used to estimate the probability of a specific characteristic (disease) in the members of a population when specific determinants are present. With the use of the estimated coefficients of the model, the OR is estimated and the reliability limits are calculated for each determinant, which are the corresponding estimations of the RR for each determinant. There are various ways of controlling whether the logistic regression model that has been estimated can be used in a specific case. One way is to study its capability for proper classification, with calculation of the different percentages of classification. When the values of the variables of each unit are substituted in the estimated model, the probability of the disease can be calculated. Using this probability, the individuals examined are classified into those estimated to have and those to have not the disease. If the percentages of accurate classification are high, the model is suitable for use in the diagnostic procedure.

Key words: Diagnostic research, Logistic regression, Medical research, Multiple logistic model, Multivariate analysis

Βιβλιογραφία

1. BLAND M. *An introduction to medical statistics.* Oxford University Press Inc, New York, 1996
2. GREENLAND S. Introduction to regression model. In: Rothmans KJ, Greenland S (eds) *Modern epidemiology*. Lippincott-Raven, Philadelphia, 1998
3. HOSMER D, LEMESHOW S. *Applied logistic regression*. Wiley, New York, 1989
4. PAGANO M, GAUVREAU K. *Principles of biostatistics*. Duxbury Press: An International Thomson Publ Co, Boston, 1996
5. ΚΑΤΟΣΤΑΡΑΣ Θ. *Εισαγωγή στη στατιστική*. Έκδοση ιδίου, Αθήνα, 1997
6. ΠΑΠΑΕΥΑΓΓΕΛΟΥ Γ, ΚΑΤΟΣΤΑΡΑΣ Θ. *Βιοστατιστική και μεθοδολογία έρευνας*. Εκδόσεις Ζήτα, Αθήνα, 1995
7. ΣΠΑΡΟΣ Λ. Η έννοια του λόγου των οτζ στην αιτιολογική έρευνα. *Αρχ Ελλην Ιατρ* 1997, 14:373–374
8. ΣΠΑΡΟΣ Λ. *Μετα-επιδημιολογία ή εφαρμοσμένη ιατρική έρευνα. Αιτιο-γνωστική, δια-γνωστική, προ-γνωστική*. Εκδόσεις ΒΗΤΑ, Αθήνα, 2001
9. ΣΠΑΡΟΣ Λ. *Θεωρία της λήψης κλινικών αποφάσεων*. Εκδόσεις ΒΗΤΑ, Αθήνα, 1999
10. ΤΡΙΧΟΠΟΥΛΟΣ Δ, ΤΖΩΝΟΥ Α, ΚΑΤΣΟΥΓΙΑΝΝΗ Κ. *Βιοστατιστική*. Επιστημονικές Εκδόσεις Μαρία Γ. Παρισιάνου, Αθήνα, 2000

Corresponding author:

T. Katostaras, 123 Papadiamantopoulou street, GR-115 27 Athens, Greece
e-mail: tkatos@nurs.uoa.gr