Copyright © Athens Medical Society
www.mednet.gr/archives
ARCHIVES OF HELLENIC MEDICINE: ISSN 11-05-3992

Πολυμεταβλητή ανάλυση επιδημιολογικών δεδομένων

ΑΡΧΕΙΑ ΕΛΛΗΝΙΚΗΣ ΙΑΤΡΙΚΗΣ 2009, 26(3):407–422
ARCHIVES OF HELLENIC MEDICINE 2009, 26(3):407–422

Π. Γαλάνης

Εργαστήριο Οργάνωσης και Αξιολόγησης Υπηρεσιών Υγείας, Τμήμα Νοσηλευτικής, Πανεπιστήμιο Αθηνών, Αθήνα

Multivariate analysis of epidemiological data

Abstract at the end of the article

1. Εισαγωγή
2. Μαθηματικά μοντέλα
3. Παλινδρόμηση
3.1. Απλή γραμμική παλινδρόμηση
3.2. Πολλαπλή γραμμική παλινδρόμηση
3.3. Μετασχηματισμός του μοντέλου της γραμμικής παλινδρόμησης
4. Λογιστική παλινδρόμηση
5. Επιλογή του κατάλληλου μοντέλου
6. Εκτίμηση του κινδύνου
7. Κατασκευή πολυμεταβλητών μοντέλων στην αιτιολογική έρευνα
7.1. «Κατασκευή των μεταβλητών»
7.2. Εφαρμογή της διαστρωματικής ανάλυσης
7.3. Επιλογή των συγχυτών
7.4. Εκτίμηση της σχέσης μεταξύ προσδιοριστή και πάθησης
7.5. Εκτίμηση της συνεπίδρασης
8. Σύνοψη

Λέξεις ευρετηρίου

Διαστρωματική ανάλυση
Λογιστική παλινδρόμηση
Παλινδρόμηση
Πολυμεταβλητή ανάλυση

Υποβλήθηκε 26.6.2008
Εγκρίθηκε 14.7.2008

1. ΕΙΣΑΓΩΓΗ

Η μέθοδος της *πολυμεταβλητής ανάλυσης** (multivariate analysis) αρχικά δημιουργεί την εντύπωση ότι πρόκειται για την πλέον ελκυστική μέθοδο ανάλυσης των δεδομένων μιας επιδημιολογικής μελέτης, καθώς επιτρέπει τόσο την εξουδετέρωση των συγχυτών όσο και την εκτίμηση των αλληλεπιδράσεων –ή καλύτερα συνεπιδράσεων– μεταξύ των ενδεικτικών κατηγοριών των προσδιοριστών και μάλιστα με σημαντική στατιστική αποτελεσματικότητα. Ο ηλεκτρονικός υπολογιστής πραγματοποιεί όλους τους απαραίτητους υπολογισμούς και εξάγει τα αποτελέσματα ταχύτατα. Μάλιστα, η ταχύτητα επεξεργασίας των δεδομένων και εξαγωγής των αποτελεσμάτων, συγκρινόμενη με τη μακροχρόνια και την επίμονη προσπάθεια συλλογής των δεδομένων, προκαλεί σε αρκετές περιπτώσεις απογοήτευση στους ερευνητές.

Όσο χρήσιμη και αποτελεσματική, πάντως, και αν είναι η πολυμεταβλητή ανάλυση, δεν αποτελεί «στατιστική πανάκεια». Το μεγαλύτερο μειονέκτημά της είναι ότι εισάγει εμπόδια, ενίοτε ανυπέρβλητα, ανάμεσα στον ερευνητή και τα «δεδομένα». Άλλες μέθοδοι ανάλυσης, όπως π.χ. η διαστρωματική ανάλυση, παρέχουν τη δυνατότητα εύκολης κατανόησης του τρόπου κατανομής των δεδομένων, οπότε μπορεί να γίνουν αντιληπτά εύκολα και άμεσα τυχόν μειονεκτήματα ή ελλείψεις. Μια τέτοια περίπτωση, για παράδειγμα, είναι η ύπαρξη μικρού αριθμού δεδομένων σε ορισμένα στρώματα, γεγονός που μειώνει την εγκυρότητα μιας μελέτης. Η πολυμεταβλητή ανάλυση δεν επιτρέπει την άμεση επαφή του ερευνητή με τα δεδομένα μιας μελέτης, ενώ τα αποτελέσματα δεν γίνονται άμεσα και εύκολα κατανοητά από τους αναγνώστες, καθώς οι περισσότεροι δεν είναι εξοικειωμένοι με τα μαθηματικά μοντέλα που χρησιμοποιούνται. Έτσι, τόσο οι αναγνώστες όσο και οι ίδιοι οι ερευνητές αντιλαμβάνονται ευκολότερα και σαφέστερα την κατανομή των δεδομένων όταν αυτά παρουσιάζονται με τη μορφή συχνοτήτων σε πίνακες (tabular analysis), όπως συμβαίνει στην περίπτωση της διαστρωματικής ανάλυσης.

* Ο όρος «ανάλυση δεδομένων» περιλαμβάνει (α) τα περιγραφικά στατιστικά μέτρα (descriptive statistics) που συνοψίζουν τα δεδομένα και (β) τα διαλογισμικά στατιστικά μέτρα (inferential statistics), τα οποία χρησιμοποιούνται στα στατιστικά υποδείγματα ή μοντέλα για την εξαγωγή συμπερασμάτων για τα αντικείμενα μιας μελέτης, που είναι, ουσιαστικά, η παράμετρος που ενδιαφέρει τους ερευνητές. Πρέπει να σημειωθεί ότι ο όρος «ανάλυση δεδομένων» (data analysis) είναι εσφαλμένος εννοιολογικά, αφού οι παρατηρήσεις δεν είναι δεδομένα και η επεξεργασία των παρατηρήσεων που είναι θεωρητικά φορτισμένες αποτελεί σύνθεση και όχι ανάλυση. Στην ουσία, η «ανάλυση των δεδομένων» είναι το σύνολο της ένδειξης (evidence), όπου η εμπειρική έρευνα παράγει για τον έλεγχο της υπόθεσης. Τα «δεδομένα» αποτελούν το μέσο επιλογής του κατάλληλου μαθηματικού υποδείγματος, δηλαδή του υποδείγματος που ταιριάζει καλύτερα στα δεδομένα.

Η ραγδαία ανάπτυξη των ηλεκτρονικών υπολογιστών και των στατιστικών προγραμμάτων είχε ως αποτέλεσμα την ευρεία χρήση της πολυμεταβλητής ανάλυσης ακόμη και στις περιπτώσεις εκείνες στις οποίες η χρήση άλλων μεθόδων (όπως π.χ. της διαστρωματικής ανάλυσης) είναι περισσότερο αποδοτική και ωφέλιμη. Υπάρχουν περιπτώσεις, εξάλλου, στις οποίες εφαρμόζονται αρκετά πολυμεταβλητά μοντέλα για την ανάλυση των δεδομένων, ενώ αρκούν μόλις ένα ή δύο μοντέλα. Τονίζεται ότι είναι προτιμότερο να εφαρμόζεται αρχικά η μέθοδος της διαστρωματικής ανάλυσης, εφόσον βεβαίως είναι δυνατή, επειδή παρέχει, τόσο στους ερευνητές όσο και στους αναγνώστες, τη δυνατότητα αμεσότερης επαφής με τα δεδομένα.

Αναμφίβολα, η πολυμεταβλητή ανάλυση είναι εξαιρετικά χρήσιμη για την ανάλυση δεδομένων, όταν η εφαρμογή της διαστρωματικής ανάλυσης[1–12] είναι πρακτικά αδύνατη. Η εφαρμογή της διαστρωματικής ανάλυσης εξαρτάται από τον αριθμό των στρωμάτων στα οποία στρωματοποιούνται τα δεδομένα μιας μελέτης. Πιο συγκεκριμένα, αν τα στρώματα που δημιουργούνται είναι πάρα πολλά, οπότε ο αριθμός των δεδομένων σε ορισμένα στρώματα είναι πολύ μικρός, τότε η διαστρωματική ανάλυση δεν είναι αποτελεσματική και τα αποτελέσματα που προκύπτουν δεν είναι έγκυρα. Η πολυμεταβλητή ανάλυση είναι μια μέθοδος με την οποία επιτυγχάνεται το κριτήριο της εγκυρότητας, καθώς χρησιμοποιείται ένα μαθηματικό μοντέλο που επιτρέπει στα δεδομένα να είναι περισσότερο αποτελεσματικά για την εκτίμηση πολλών χαρακτηριστικών (ή μεταβλητών) ταυτόχρονα. Όσες λιγότερες προϋποθέσεις απαιτεί η εφαρμογή ενός μαθηματικού μοντέλου τόσο μεγαλύτερη εγκυρότητα παρέχει η πολυμεταβλητή ανάλυση. Είναι προτιμότερο, πάντως, να συνδυάζεται η διαστρωματική ανάλυση με την πολυμεταβλητή ακόμη και όταν ελέγχονται πολλές μεταβλητές* ταυτόχρονα. Στην περίπτωση αυτή, αρχικά, θα πρέπει να εφαρμόζεται η διαστρωματική ανάλυση για τον έλεγχο των πιο σημαντικών μεταβλητών και στη συνέχεια η πολυμεταβλητή για τον έλεγχο όλων των μεταβλητών.[8,9] Η διαδικασία αυτή είναι χρονοβόρα και σε ορισμένες περιπτώσεις επίπονη, αλλά δεν αποτελεί σπατάλη χρόνου, καθώς προσφέρει πολύτιμα συμπεράσματα.

Η πολυμεταβλητή ανάλυση,** με την ευρεία έννοια, αφορά σε κάθε ανάλυση δεδομένων που λαμβάνει υπόψη της ένα μεγάλο αριθμό μεταβλητών. Σε κάθε περίπτωση, χρησιμοποιείται ένα μαθηματικό μοντέλο για να ερμηνευτεί η συσχέτιση μεταξύ των μεταβλητών μιας μελέτης. Οι «ισότητες πιθανοφάνειας» (likelihood equations) που χρησιμοποιούνται για την εκτίμηση των επιδημιολογικών μέτρων σχέσης*** με ταυτόχρονο έλεγχο των συγχυτών αποτελούν χαρακτηριστικό παράδειγμα μαθηματικού μοντέλου που χρησιμοποιεί η πολυμεταβλητή ανάλυση.[8] Αρκετά άλλα πολυμεταβλητά μαθηματικά μοντέλα χρησιμοποιούνται σε συγκεκριμένες περιπτώσεις. Για παράδειγμα, η «ανάλυση παραγόντων» (factor analysis) χρησιμοποιείται για τη μετατροπή ενός μεγάλου αριθμού προβλεπτικών**** μεταβλητών σε ένα μικρότερο αριθμό «παραγόντων» που αντιπροσωπεύει συσχετισμένες υποομάδες από την αρχική «αυθεντική» ομάδα των μεταβλητών. Επιπλέον, η «ανάλυση χρονικών σειρών» (time series analysis) είναι μια μέθοδος που χρησιμοποιείται για την εκτίμηση της σχέσης μεταξύ μεταβλητών, ενώ παράλληλα λαμβάνει υπόψη της τη μεταβλητότητα που οφείλεται στο χρόνο και τη συσχέτιση μεταξύ επιλεγμένων μεταβλητών. Στην επιδημιολογία, η πολυμεταβλητή ανάλυση που παρουσιάζει το μεγαλύτερο ενδιαφέρον και τις περισσότερες εφαρμογές είναι η *πολυμεταβλητή* (multivariate) ή *πολλαπλή ανάλυση παλινδρόμησης* (multiple regression analysis).

2. ΜΑΘΗΜΑΤΙΚΑ ΜΟΝΤΕΛΑ

Στην εικόνα 1 φαίνεται η σχέση μεταξύ της προτυπο-

* Τονίζεται ότι οι μεταβλητές δεν υπάρχουν στη φύση, αλλά είναι στατιστικές έννοιες και σχεδιάζονται από τον ερευνητή. Το φύλο, για παράδειγμα, δεν είναι μεταβλητή, αλλά μετατρέπεται σε μεταβλητή Χ και οι πραγματοποιήσεις του είναι οι αριθμοί 1 (για τους άνδρες) και 0 (για τις γυναίκες). Πρέπει να σημειωθεί, εξάλλου, ότι ο αγγλικός όρος για τη μεταβλητή είναι "variate" και όχι "variable".

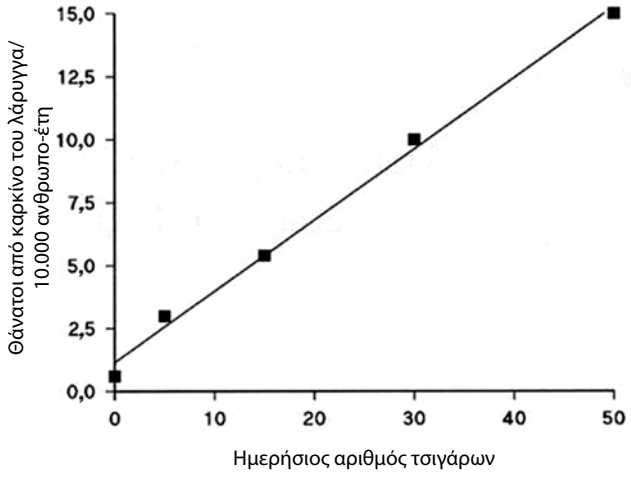
** Ο όρος "multivariate analysis" αποδίδεται ορθά με τον όρο «πολυμεταβλητή ανάλυση» και όχι με τον όρο «πολυπαραγοντική ανάλυση».

*** Τα μέτρα αποτελέσματος (effect measures) υπολογίζονται στις περιπτώσεις εκείνες κατά τις οποίες η έκθεση και η μη έκθεση στην ενδεικτική κατηγορία ενός προσδιοριστή αφορούν στον ίδιο πληθυσμό. Τα μέτρα αποτελέσματος εκτιμούν την αιτιακή σχέση μεταξύ της ενδεικτικής κατηγορίας του προσδιοριστή και της συχνότητας εμφάνισης της πάθησης. Αντιθέτως, τα μέτρα σχέσης (measures of association) υπολογίζονται στις περιπτώσεις εκείνες κατά τις οποίες η έκθεση και η μη έκθεση στην ενδεικτική κατηγορία ενός προσδιοριστή αφορούν σε δύο διαφορετικούς πληθυσμούς. Επιπλέον, εκτιμούν την πιθανολογική σχέση μεταξύ της ενδεικτικής κατηγορίας ενός προσδιοριστή και της συχνότητας εμφάνισης της πάθησης. Ο υπολογισμός των μέτρων αποτελέσματος απαιτεί τα μελετώμενα άτομα να βρίσκονται ταυτόχρονα τόσο στην εκτεθειμένη όσο και στη μη εκτεθειμένη ομάδα. Αυτό βεβαίως είναι αδύνατο και γι' αυτό σε μια μελέτη υπολογίζονται πάντοτε τα μέτρα σχέσης.

**** Οι προβλεπτικές (predictors) μεταβλητές ονομάζονται και ανεξάρτητες (independents) ή απαντητικές (responses) ή εξηγητικές (explanatory) μεταβλητές. Πρέπει, πάντως, να σημειωθεί ότι οι στατιστικοί δεν προτιμούν τους όρους «εξαρτημένες» και «ανεξάρτητες» μεταβλητές, έτσι ώστε να αποφεύγεται η σύγχυση με τους αντίστοιχους όρους της θεωρίας των πιθανοτήτων (θεωρία της στατιστικής ανεξαρτησίας),[13,14] όπου δύο (π.χ. Α και Β) ενδεχόμενα (events) είναι ανεξάρτητα όταν (α) η πιθανότητα πραγματοποίησης του Α δεν επηρεάζεται από την πληροφορία ότι έχει ήδη πραγματοποιηθεί το Β, δηλαδή $P(A/B)=P(A)$ και (β) η πιθανότητα πραγματοποίησης του Β δεν επηρεάζεται από την πληροφορία ότι έχει ήδη πραγματοποιηθεί το Α, δηλαδή $P(B/A)=P(B)$. Οι όροι «εξαρτημένες» και «ανεξάρτητες» μεταβλητές, όπως χαρακτηριστικά αναφέρει ο Strike, είναι ένα «παλιό ρούχο» ή, αλλιώς, αποφόρι προερχόμενο από τη γεωμετρία, που δεν πρέπει να χρησιμοποιείται.

ποιημένης κατά ηλικία θνησιμότητας από καρκίνο του λάρυγγα και του αριθμού των τσιγάρων που καταναλώνονται καθημερινά.[15] Η ευθεία γραμμή στην εικόνα 1 αποτελεί παράδειγμα ενός απλού μαθηματικού μοντέλου. Πρόκειται για μοντέλο, καθώς χρησιμοποιείται η μαθηματική ισότητα της ευθείας γραμμής που αντιστοιχεί στα δεδομένα της μελέτης για να περιγραφεί η σχέση μεταξύ των δύο μεταβλητών του γραφήματος, δηλαδή του αριθμού των τσιγάρων και της θνησιμότητας από καρκίνο του λάρυγγα. Στην επιδημιολογία, τα μαθηματικά μοντέλα χρησιμοποιούνται για διάφορους σκοπούς, κυριότεροι από τους οποίους είναι η πρόβλεψη και η εξουδετέρωση των συγχυτών.[9] Τα προβλεπτικά μοντέλα χρησιμοποιούνται για την πρόβλεψη ή, με άλλη διατύπωση, την αποτίμηση του κινδύνου* (risk assessment) και στηρίζονται στην πληροφορία που προέρχεται από προβλεπτικούς προσδιοριστές** ή αλλιώς παράγοντες κινδύνου. Έτσι, ένα μαθηματικό προβλεπτικό μοντέλο εξάγει μια μαθηματική ισότητα που στη συνέχεια χρησιμοποιείται για να εκτιμηθεί, π.χ., ο κίνδυνος ενός ατόμου να εμφανίσει έμφραγμα του μυοκαρδίου στα επόμενα 10 έτη με βάση την ηλικία, το φύλο, το ιατρικό ιστορικό, την αρτηριακή πίεση, το ιστορικό καπνισματικής συνήθειας, το βάρος, το ύψος, τη σωματική άσκηση και το οικογενειακό ιστορικό του ατόμου. Οι τιμές για τον καθέναν από τους προβλεπτικούς αυτούς προσδιοριστές είναι δυνατόν να εισέλθουν σε μια μαθηματική ισότητα που προβλέπει τον κίνδυνο εμφάνισης εμφράγματος του μυοκαρδίου με βάση το συνδυασμό όλων αυτών των προσδιοριστών. Το μοντέλο πρέπει να περιλαμβάνει όρους για όλους τους μελετώμενους προσδιοριστές.

Η επιδημιολογική έρευνα, πάντως, εστιάζεται όχι τόσο στην αποτίμηση του κινδύνου για συγκεκριμένα άτομα (περιγραφικές σχέσεις), αλλά κυρίως στην αναζήτηση του αιτιακού ρόλου συγκεκριμένων προσδιοριστών (αιτιακές σχέσεις) στην πρόκληση μιας πάθησης.[9] Στην αιτιολογική έρευνα, τα μαθηματικά μοντέλα χρησιμοποιούνται για να εκτιμηθεί ο αιτιολογικός ρόλος ενός προσδιοριστή με ταυτόχρονη εξουδετέρωση των πιθανών συγχυτών. Η χρήση των πολυμεταβλητών μοντέλων στην αιτιολογική έρευνα διαφέρει σημαντικά από τη χρήση των αντίστοιχων μοντέλων για την εκτίμηση του κινδύνου σε συγκεκριμένα άτομα και γι' αυτό σε καθεμιά περίπτωση η κατασκευή του κατάλληλου μοντέλου απαιτεί ορισμένες προϋποθέσεις. Αρκετές φορές, δυστυχώς, δεν διαχωρίζεται η χρήση των πολυμεταβλητών μοντέλων για την εκτίμηση του κινδύνου σε συγκεκριμένα άτομα από τη χρήση των αντίστοιχων μοντέλων στην αιτιολογική έρευνα.



Εικόνα 1. Προτυποποιημένη, ως προς την ηλικία, θνησιμότητα από καρκίνο του λάρυγγα σε σχέση με τον ημερήσιο αριθμό τσιγάρων.[10]

3. ΠΑΛΙΝΔΡΟΜΗΣΗ

Ο όρος "regression" (παλινδρόμηση) εισήχθη το 1877 από τον Francis Galton*** (1822–1911) στην προσπάθειά

* Η έννοια του κινδύνου συνδέεται άμεσα με την έννοια της επίπτωσης-ποσοστού.[8–11,16,17] Με τον όρο κίνδυνος νοείται η επίπτωση-ποσοστό σε επίπεδο ατόμου. Ο κίνδυνος (risk) ορίζεται ως η πιθανότητα ενός ατόμου να εμφανίσει ένα ανεπιθύμητο συμβάν (έναρξη πάθησης ή θάνατος) σε ένα συγκεκριμένο χρονικό διάστημα. Η επίπτωση-ποσοστό είναι εμπειρικό ή θεωρητικό μέτρο συχνότητας, ενώ ο κίνδυνος είναι θεωρητικό μέτρο και το μέγεθός του δεν υπολογίζεται, αλλά έχει a priori μια συγκεκριμένη, αλλά άγνωστη τιμή. Η τιμή αυτή εκτιμάται με βάση τα εμπειρικά μέτρα συχνότητας που διαπιστώνονται σε έναν ορισμένο τομέα. Εάν η εγχειρητική θνητότητα μιας πάθησης, για παράδειγμα, είναι 30%, τότε ο κίνδυνος ενός ατόμου να πεθάνει κατά τη διάρκεια της εγχείρησης είναι 0,30 πριν και μετά από την εγχείρηση, αδιακρίτως έκβασης.

** Παράγοντας κινδύνου (risk factor) ή έκθεση (exposure) ή προσδιοριστής (determinant), όπως τελικά επικράτησε να λέγεται σήμερα, είναι το χαρακτηριστικό των ατόμων από το οποίο εξαρτάται (σχετίζεται ή συναρτάται) η συχνότητα της μελετώμενης έκβασης.[5,11,17,18] Ο προσδιοριστής της συχνότητας μιας μελετώμενης έκβασης περιλαμβάνει δύο κατηγορίες, την ενδεικτική κατηγορία (index category) και την κατηγορία αναφοράς (reference category). Η επιλογή της κατηγορίας αναφοράς του προσδιοριστή είναι μείζονος σημασίας στο σχεδιασμό της μελέτης. Είναι κρίσιμης σημασίας να γίνει αντιληπτό ότι η κατηγορία αναφοράς δεν είναι συμφυές χαρακτηριστικό του προσδιοριστή, αλλά αποτέλεσμα εκλογής του ερευνητή, άρα συνέπεια του σχεδιασμού της μελέτης. Από την επιλογή της κατηγορίας αναφοράς εξαρτάται η ερμηνεία των αποτελεσμάτων μιας μελέτης. Για παράδειγμα, προσδιοριστής της συχνότητας της νεφρικής πάθησης δεν είναι η αρτηριακή υπέρταση, αλλά η αρτηριακή πίεση. Η αρτηριακή υπέρταση είναι μια κατηγορία και συνήθως η ενδεικτική κατηγορία του προσδιοριστή, στην οποία μελετάται η συχνότητα της νεφρικής πάθησης, σε σχέση πάντοτε με τη συχνότητα της νεφρικής πάθησης στην κατηγορία αναφοράς, εν προκειμένω στην κατηγορία των ατόμων που δεν έχουν αρτηριακή πίεση.

*** Ο Francis Galton το 1885 προέβη στην εξήγηση του όρου "regression", στηριζόμενος στις αρχές της κανονικής κατανομής. Στην προσπάθειά του αυτή σημαντική ήταν η συμβολή του Hamilton Dickson, καθηγητή μαθηματικών στο Cambridge. Ο Galton το 1889 εξέδωσε το "natural inheritance" συμπεριλαμβάνοντας την εξήγησή του για τον όρο "regression".[19] Ο Galton θεωρείται ως ο θεμελιωτής της βιομετρίας και σε συνεργασία με τους Pearson, Edgeworth και Fisher συνέβαλε σημαντικά στην πρόοδο των μαθηματικών και της στατιστικής. Αν και ολοκλήρωσε τις σπουδές του στην Ιατρική, δεν την άσκησε ποτέ, προτιμώντας τα μαθηματικά και τη βιομετρία και εκδίδοντας μάλιστα, το 1901, σε συνεργασία με τους Pearson και Weldon το περιοδικό "*Biometrika*".

του να ερμηνεύσει τη σχέση του ύψους μεταξύ γονιών και παιδιών. Διαπίστωσε ότι γονείς υψηλού αναστήματος συνήθως αποκτούν παιδιά υψηλού αναστήματος, αλλά το μέσο ύψος των παιδιών αυτών τείνει να είναι μικρότερο από το μέσο ύψος των γονιών τους, πλησιάζοντας μάλιστα το μέσο ύψος του πληθυσμού υπό παρακολούθηση. Επιπλέον, διαπίστωσε ότι οι γονείς μικρού αναστήματος αποκτούν, συνήθως, παιδιά μικρού αναστήματος, αλλά το μέσο ύψος των παιδιών αυτών τείνει να είναι μεγαλύτερο από το μέσο ύψος των γονιών τους πλησιάζοντας το μέσο ύψος του πληθυσμού υπό παρακολούθηση. Ο Galton όρισε το φαινόμενο αυτό ως «παλινδρόμηση προς τη μετριότητα» ("regression towards mediocrity"), ενώ σήμερα οι στατιστικοί το ονομάζουν «παλινδρόμηση προς τη μέση τιμή» ("regression towards the mean").[13,19] Πρέπει, πάντως, να σημειωθεί ότι σήμερα η χρήση του όρου «παλινδρόμηση» στη στατιστική θεωρείται ως «ιστορική ανωμαλία» που δημιουργεί σύγχυση.

Όπως προαναφέρθηκε, στην ανάλυση δεδομένων ο όρος «παλινδρόμηση» (regression*) σήμαινε αρχικά την παλινδρόμηση προς τη μέση τιμή, οπότε τα «μοντέλα παλινδρόμησης» (regression models) και η «ανάλυση παλινδρόμησης» (regression analysis) αναφέρονται στη σχέση μεταξύ μιας εξαρτημένης παραμέτρου (π.χ. του μέσου μιας ποσότητας) και ενός συνόλου ποσοτικών ανεξάρτητων μεταβλητών (προσδιοριστών).[18] Αντιθέτως, στην «ανάλυση διασποράς» (analysis of variates) οι ανεξάρτητες μεταβλητές είναι ενδεικτικές** μεταβλητές (indicator variables), ενώ στην «ανάλυση συνδιασποράς» (analysis of covariance) οι ανεξάρτητες μεταβλητές είναι τόσο ποσοτικές όσο και ενδεικτικές. Σήμερα, πάντως, ο όρος «παλινδρόμηση» αναφέρεται στη σχέση μιας παραμέτρου*** με ένα σύνολο προσδιοριστών ανεξάρτητα από τη φύση τους.

Ένα μοντέλο παλινδρόμησης θεωρείται γραμμικό (linear) όταν είναι της μορφής:

$$P = A_0 + \sum_i A_i X_i$$

Στην παραπάνω ισότητα, η τιμή της εξαρτημένης παραμέτρου P θεωρείται ως γραμμικός συνδυασμός των ανεξάρτητων συντελεστών A_0, A_1, A_2…

3.1. Απλή γραμμική παλινδρόμηση

Το βασικό μαθηματικό μοντέλο που περιγράφει τη σχέση μεταξύ δύο μεταβλητών είναι η ευθεία γραμμή.[8] Το γραμμικό μοντέλο για δύο μεταβλητές αποτελεί τη βάση για τη δημιουργία πιο σύνθετων μοντέλων μεταξύ περισσότερων μεταβλητών.

Στην εικόνα 1, τα δεδομένα αναπαριστούν (δηλώνουν ή εκφράζουν) μια σχεδόν τέλεια γραμμική σχέση μεταξύ του αριθμού των τσιγάρων που καταναλώνονται καθημερινά και της προτυποποιημένης για την ηλικία θνησιμότητας από καρκίνο του λάρυγγα.[9] Σπάνια, πάντως, τα δεδομένα μιας επιδημιολογικής μελέτης αναπαριστούν μια τόσο εντυπωσιακή γραμμική σχέση. Η ευθεία γραμμή που δημιουργείται διαμέσου των σημείων που αντιστοιχούν στα δεδομένα μιας μελέτης ονομάζεται *γραμμή παλινδρόμησης* (regression line) και εκτιμά τις μέσες τιμές για τη μεταβλητή στον κάθετο άξονα (y) σύμφωνα με τις τιμές της μεταβλητής στον οριζόντιο άξονα (x). Στην περίπτωση αυτή, το μαθηματικό μοντέλο που χρησιμοποιείται είναι η *απλή γραμμική παλινδρόμηση* (simple linear regression), καθώς η σχέση μεταξύ των δύο μεταβλητών περιγράφεται από μια ευθεία γραμμή σύμφωνα με την ακόλουθη ισότητα:[8,9,13,19,21]

$$Y = A_0 + A_1X_1 + \varepsilon \qquad (1)$$

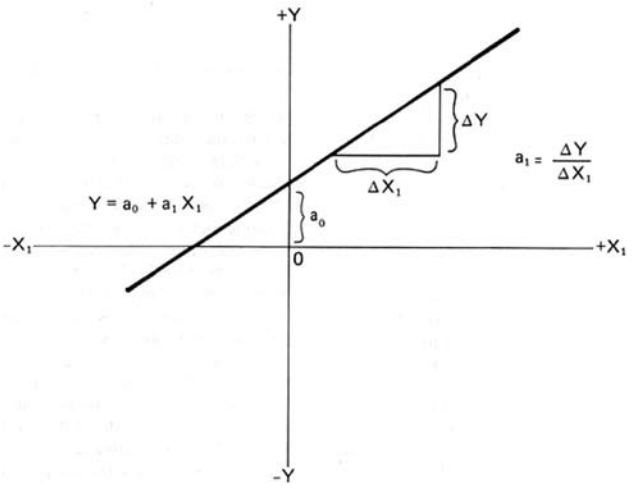
Στην ισότητα 1, το Y είναι η εξαρτημένη μεταβλητή της απλής γραμμικής παλινδρόμησης, ενώ το X_1 είναι η ανεξάρτητη μεταβλητή. Ουσιαστικά, το Y αντιστοιχεί στη μελετώμενη έκβαση, ενώ το X_1 αντιστοιχεί στο μελετώμενο προσδιοριστή. Το A_0 είναι η σταθερά της απλής γραμμικής παλινδρόμησης και είναι η μέση τιμή που λαμβάνει η μεταβλητή Y, όταν η μεταβλητή X_1 ισούται με 0. Το A_1 περιγράφει την κλίση της ευθείας γραμμής που συσχετίζει το X_1 με το Y. Το A_1 είναι ο αριθμός των μονάδων που μεταβάλλεται το Y κάθε φορά που η τιμή του X_1 μεταβάλλεται κατά μία μονάδα. Το ε είναι το τυχαίο σφάλμα που αντιπροσωπεύει την τυχαία απόκλιση από την αναμενόμενη τιμή της εξαρτημένης μεταβλητής Y. Η μέση τιμή του τυχαίου σφάλματος, γενικά, θεωρείται ίση με 0. Οι τιμές των A_0 και A_1 κυμαίνονται θεωρητικά από -∞ έως +∞. Στην εικόνα 2 απεικονίζεται η απλή γραμμική σχέση μεταξύ δύο οποιωνδήποτε μεταβλητών X_1 και Y.

Όπως προαναφέρθηκε, είναι σπάνιο η πραγματική σχέση μεταξύ δύο μεταβλητών να είναι τελείως γραμμική. Η υπόθεση που ισχύει για την εφαρμογή ενός μαθηματικού μοντέλου σε μια ανάλυση είναι ότι το μοντέλο αυτό

* Ο όρος "regression" σε ορισμένες περιπτώσεις αποδίδεται με τον όρο «εξάρτηση».[20]

** Πρέπει να σημειωθεί ότι στην περίπτωση των ενδεικτικών μεταβλητών, τόσο η έκβαση (εξαρτημένη μεταβλητή) όσο και οι προσδιοριστές (ανεξάρτητες μεταβλητές) δεν αποτελούν στατιστικές ποσότητες, δεν αποτελούν δηλαδή μεταβλητές. Στατιστικά, ωστόσο, αναπαρίστανται (εκφράζονται ή δηλώνονται) με τη μορφή μεταβλητών, κάτι που αποτελεί επιλογή του ερευνητή. Η εμφάνιση ή μη μιας πάθησης, π.χ., δεν αποτελεί μεταβλητή, αλλά εκφράζεται με τη μορφή μιας ενδεικτικής μεταβλητής όπου η εμφάνιση ή μη της πάθησης δηλώνεται με αριθμούς, συνήθως 1 και 0, αντίστοιχα. Στη συνέχεια, αναφέρεται αναλυτικά ο τρόπος με τον οποίο λαμβάνονται υπόψη σε ένα μαθηματικό υπόδειγμα (ή μοντέλο) οι μεταβλητές που έχουν τουλάχιστον τρεις κατηγορίες (π.χ. η ομάδα αίματος).

*** Στην επιδημιολογία, παράμετροι δεν είναι οι μεταβλητές (π.χ. X_1, X_2... X_n), αλλά οι συντελεστές παλινδρόμησης (A_0, A_1, A_2...A_n).



Εικόνα 2. Γραφική αναπαράσταση της απλής γραμμικής σχέσης μεταξύ δύο μεταβλητών X_1 και Y.

αποτελεί μια απλοποιημένη περιγραφή της σχέσης μεταξύ δύο μεταβλητών και δεν συμμορφώνεται αναγκαστικά με την πραγματική τους σχέση. Προσεγγίζει, ωστόσο, αρκετά την πραγματική σχέση μεταξύ των δύο μεταβλητών, με αποτέλεσμα να δικαιολογείται η χρήση του. Αποκλίσεις των δεδομένων από το εφαρμοζόμενο μαθηματικό μοντέλο απεικονίζουν την ασυμφωνία μεταξύ του μοντέλου και της φύσης, καθώς επίσης και πηγές ανακρίβειας στη συλλογή των πληροφοριών. Επιλέγεται ένα μοντέλο που, λογικά, ταιριάζει όσο το δυνατόν περισσότερο με τα δεδομένα, έτσι ώστε το μεγαλύτερο μέρος της απόκλισης των δεδομένων από το μοντέλο αυτό να οφείλεται στην ανακρίβεια που προέρχεται από τα ίδια τα δεδομένα και όχι στην ακαταλληλότητα του μοντέλου.

Στην εικόνα 1, το Y αναπαριστά την προτυποποιημένη για την ηλικία θνησιμότητα από καρκίνο του λάρυγγα, ενώ το X_1 αναπαριστά τον αριθμό των τσιγάρων που καταναλώνονται καθημερινά. Η ισότητα για τη γραμμή παλινδρόμησης της εικόνας 1 είναι $Y=1,15+0,282X_1$. Οι τιμές αυτές αναφέρονται σε θανάτους από καρκίνο του λάρυγγα ανά 10.000 ανθρωπο-έτη παρακολούθησης. Η σταθερά ($A_0=1,15$) δηλώνει τον αριθμό των θανάτων από καρκίνο του λάρυγγα που θα συνέβαιναν ανά 10.000 ανθρωπο-έτη σε περίπτωση απουσίας του καπνίσματος. Έτσι, για τους μη καπνιστές, η θνησιμότητα από καρκίνο του λάρυγγα, σύμφωνα με το μοντέλο αυτό της απλής γραμμικής παλινδρόμησης, είναι 1,15 θάνατοι ανά 10.000 ανθρωπο-έτη. Σύμφωνα με τα δεδομένα της μελέτης, η αντίστοιχη θνησιμότητα για τους μη καπνιστές βρέθηκε ίση με 0,6 θανάτους ανά 10.000 ανθρωπο-έτη. Επομένως, η πραγματική θνησιμότητα $\left(\frac{\text{0,6 θάνατοι}}{\text{10.000 ανθρωπο-έτη}}\right)$ για τους μη καπνιστές είναι λίγο μικρότερη από την αντίστοιχη θνησιμότητα $\left(\frac{\text{1,15 θάνατοι}}{\text{10.000 ανθρωπο-έτη}}\right)$ που υπολογίζεται με το εφαρμοζόμενο μοντέλο της απλής γραμμικής παλινδρόμησης. Το γεγονός αυτό οφείλεται στο ότι η ευθεία γραμμή της εικόνας 1 εκτιμάται έπειτα από το συνδυασμό και των πέντε σημείων που αντιστοιχούν στα δεδομένα της μελέτης και όχι μόνο από το σημείο που αντιστοιχεί στην απουσία του καπνίσματος. Η κλίση της γραμμής παλινδρόμησης ισούται με 0,282, που σημαίνει ότι ο αριθμός των θανάτων ανά 10.000 ανθρωπο-έτη αυξάνεται κατά 0,282 για κάθε επιπλέον τσιγάρο που καταναλώνεται καθημερινά. Έτσι, για εκείνους που καπνίζουν καθημερινά 50 τσιγάρα η θνησιμότητα από καρκίνο του λάρυγγα ισούται με $1,15+0,282\times50=15,2$ θανάτους ανά 10.000 ανθρωπο-έτη.

Με την προϋπόθεση ότι έχουν εξουδετερωθεί οι συγχυτές,* καθώς και τα υπόλοιπα συστηματικά σφάλματα της μελέτης, η τιμή ή, αλλιώς, το μέγεθος της κλίσης ($A_1=0,282$) ποσοτικοποιεί το αποτέλεσμα του καπνίσματος στη θνησιμότητα από καρκίνο του λάρυγγα. Η ισότητα για τη γραμμή παλινδρόμησης ($Y=1,15+0,282X_1$), εξάλλου, παρέχει τη δυνατότητα εκτίμησης των λόγων θνησιμοτήτων σε διαφορετικά επίπεδα καπνισματικής συνήθειας. Για παράδειγμα, για εκείνους που καταναλώνουν καθημερινά 50 τσιγάρα η θνησιμότητα από καρκίνο του λάρυγγα είναι $\frac{\text{15,2 θάνατοι}}{\text{10.000 ανθρωπο-έτη}}$. Συγκρίνοντας τη θνησιμότητα αυτή με τη θνησιμότητα για τους μη καπνιστές $\left(\frac{\text{1,15 θάνατοι}}{\text{10.000 ανθρωπο-έτη}}\right)$, προκύπτει ότι ο λόγος των θνησιμοτήτων για εκείνους που καπνίζουν καθημερινά 50 τσιγάρα σε σχέση με εκείνους που δεν καπνίζουν είναι $\frac{15,2}{1,15}=13,3$. Έτσι, η καπνισματική συνήθεια αποτελεί προσδιοριστή της συχνότητας του θανάτου από καρκίνο του λάρυγγα.

* Ο συγχυτής (confounder)[1,2,4,5,8–12,18,22] είναι το χαρακτηριστικό εκείνο που (α) σχετίζεται με τη συχνότητα εμφάνισης της πάθησης, αποτελώντας έτσι εξωγενή προσδιοριστή της συχνότητας εμφάνισης της πάθησης, (β) σχετίζεται με το μελετώμενο προσδιοριστή, ανισοκατανέμεται δηλαδή στις δύο κατηγορίες του προσδιοριστή και (γ) δεν αποτελεί αποτέλεσμα του μελετώμενου προσδιοριστή. Οι εξωγενείς προσδιοριστές (extraneous determinants) ονομάζονται δυνητικοί συγχυτές (potential confounders), καθώς σχετίζονται με τη συχνότητα εμφάνισης της μελετώμενης πάθησης. Όταν οι εξωγενείς αυτοί προσδιοριστές ανισοκατανέμονται κιόλας στις κατηγορίες του μελετώμενου προσδιοριστή, τότε ονομάζονται πραγματικοί συγχυτές (actual confounders). Οι συγχυτές είναι συστηματικά σφάλματα που μπορούν, εφόσον διαπιστωθούν, να εξουδετερωθούν (ή ελεγχθούν) κατά την ανάλυση των δεδομένων, κάτι που δεν μπορεί να γίνει στα συστηματικά σφάλματα επιλογής ή πληροφορίας.

3.1.1. Πολλαπλή γραμμική παλινδρόμηση

Η ισότητα 1 χρησιμοποιείται στην περίπτωση της απλής γραμμικής παλινδρόμησης επειδή υπάρχει μόνο μία ανεξάρτητη μεταβλητή. Το μοντέλο, ωστόσο, της γραμμικής παλινδρόμησης μπορεί να επεκταθεί, περιλαμβάνοντας περισσότερες από μία ανεξάρτητες μεταβλητές, οπότε προκύπτει το μοντέλο της πολλαπλής γραμμικής παλινδρόμησης:[8,9,13,19,21]

$$Y = A_0 + A_1X_1 + A_2X_2 + A_3X_3 + \text{.....} \quad (2)$$

Το μοντέλο της πολλαπλής γραμμικής παλινδρόμησης, το οποίο στη στατιστική ονομάζεται *γενικό γραμμικό μοντέλο* (general linear model), και γενικότερα τα μοντέλα που περιλαμβάνουν περισσότερες από μία ανεξάρτητες μεταβλητές ταυτόχρονα, χρησιμοποιούνται για την εξουδετέρωση των συγχυτών αντί της διαστρωματικής ανάλυσης. Στα πολυμεταβλητά μαθηματικά μοντέλα περιλαμβάνονται αρκετοί προσδιοριστές, αλλά το αποτέλεσμα του καθενός δεν συγχέεται από τη δράση των υπολοίπων και γι' αυτό τα μοντέλα αυτά είναι ιδανικά για την εξουδετέρωση των συγχυτών μιας μελέτης όταν, βεβαίως, δεν είναι δυνατή η εφαρμογή της διαστρωματικής ανάλυσης.

Η ισότητα 2 αντιστοιχεί και πάλι σε μια ευθεία γραμμή, αλλά η γραμμή αυτή διαγράφει στο χώρο ένα διάστημα με περισσότερες από δύο διαστάσεις, καθώς σε κάθε μεταβλητή αντιστοιχεί μία διάσταση. Το Y αναπαριστά τη θνησιμότητα από καρκίνο του λάρυγγα, ενώ υπάρχουν δύο ανεξάρτητες μεταβλητές X_1 και X_2 στο μοντέλο της πολλαπλής γραμμικής παλινδρόμησης. Το X_1 είναι ο αριθμός των τσιγάρων που καταναλώνονται καθημερινά, ενώ το X_2 είναι τα γραμμάρια οινοπνεύματος που καταναλώνονται καθημερινά. Η κατανάλωση ή μη οινοπνεύματος είναι, επίσης, προσδιοριστής της συχνότητας του θανάτου από καρκίνο του λάρυγγα. Όταν υπάρχουν δύο ανεξάρτητες μεταβλητές και μία εξαρτημένη, τότε απαιτείται η τρισδιάστατη απεικόνιση των δεδομένων στο χώρο, καθώς απαιτούνται δύο διαστάσεις για τις δύο ανεξάρτητες μεταβλητές και μία διάσταση για την εξαρτημένη μεταβλητή.

Η καπνισματική συνήθεια και η κατανάλωση ή μη οινοπνεύματος συσχετίζονται μεταξύ τους, οπότε αναμένεται καθένα από τα δύο αυτά χαρακτηριστικά να είναι συγχυτής της σχέσης του άλλου με τη μελετώμενη έκβαση, όπου στη συγκεκριμένη περίπτωση είναι ο θάνατος από καρκίνο του λάρυγγα. Η διαστρωματική ανάλυση αποτελεί την καλύτερη μέθοδο εξουδετέρωσης της σύγχυσης όταν ο αριθμός των πιθανών συγχυτών είναι μικρός, αλλά μπορεί να χρησιμοποιηθεί και η πολλαπλή γραμμική παλινδρόμηση, οπότε εφαρμόζεται η ισότητα 2. Στο συγκεκριμένο παράδειγμα, η εφαρμογή της ισότητας 2 περιλαμβάνει δύο προσδιοριστές (ή αλλιώς προβλεπτικούς δείκτες), την καπνισματική συνήθεια (X_1) και την κατανάλωση ή μη οινοπνεύματος (X_2), οπότε υπολογίζονται δύο συντελεστές, A_1 και A_2, αντίστοιχα. Οι συντελεστές A_1 και A_2 εκτιμούν το αποτέλεσμα της δράσης του καπνίσματος και του οινοπνεύματος, αντίστοιχα, εξουδετερώνοντας ταυτόχρονα τη σύγχυση που τυχόν υπάρχει στη σχέση μεταξύ της ενδεικτικής κατηγορίας του κάθε προσδιοριστή ξεχωριστά και της συχνότητας του θανάτου από καρκίνο του λάρυγγα. Μαθηματικά, τουλάχιστον, δεν υπάρχει περιορισμός στον αριθμό των ανεξάρτητων μεταβλητών που μπορούν να χρησιμοποιηθούν στην πολλαπλή γραμμική παλινδρόμηση, αν και μικρός αριθμός δεδομένων επιτρέπει να χρησιμοποιηθούν λίγες μόνο ανεξάρτητες μεταβλητές.

3.1.2. Μετασχηματισμός του μοντέλου της γραμμικής παλινδρόμησης

Μαθηματικά, η εξαρτημένη μεταβλητή σε ένα μοντέλο παλινδρόμησης μπορεί να λάβει οποιαδήποτε τιμή. Στην επιδημιολογία, ωστόσο, η εξαρτημένη μεταβλητή σε αρκετές περιπτώσεις περιορίζεται σε ορισμένες μόνο τιμές. Αν η εξαρτημένη μεταβλητή, π.χ., είναι η αρτηριακή πίεση, τότε μπορεί να λάβει μόνο θετικές τιμές. Στην περίπτωση που η εξαρτημένη μεταβλητή είναι η συχνότητα εμφάνισης μιας πάθησης, τότε η μελετώμενη έκβαση μπορεί είτε να εμφανιστεί είτε όχι, οπότε η μεταβλητή λαμβάνει τιμές 1 και 0, αντίστοιχα. Συνήθως, η εμφάνιση της μελετώμενης έκβασης, όταν η μεταβλητή είναι ενδεικτική, ονομάζεται «επιτυχία» και συμβολίζεται με 1, ενώ η μη εμφάνιση της μελετώμενης έκβασης ονομάζεται «αποτυχία» και συμβολίζεται με 0. Στις περιπτώσεις αυτές απαιτείται ο μετασχηματισμός του μοντέλου της γραμμικής παλινδρόμησης, έτσι ώστε η εξαρτημένη μεταβλητή να λαμβάνει τιμές που είναι δυνατές. Στην εικόνα 1, π.χ., η σταθερά (A_0) της ευθείας γραμμής είναι 1,15 θάνατοι ανά 10.000 ανθρωπο-έτη, αλλά αν τα δεδομένα της μελέτης ήταν λίγο διαφορετικά, τότε η ευθεία γραμμή της γραμμικής παλινδρόμησης θα μπορούσε να κινηθεί στον άξονα των y σε μια τιμή <0. Στην περίπτωση αυτή η θνησιμότητα για τους μη καπνιστές θα ήταν αρνητική, κάτι που είναι αδύνατο. Το μοντέλο της γραμμικής παλινδρόμησης επιτρέπει στην εξαρτημένη μεταβλητή να λάβει οποιαδήποτε τιμή.

Είναι εφικτός, πάντως, ο μετασχηματισμός του μοντέλου της γραμμικής παλινδρόμησης, έτσι ώστε η εξαρτημένη μεταβλητή να λαμβάνει μόνο θετικές τιμές. Ένας τρόπος για να επιτευχθεί αυτό είναι η προσαρμογή της ευθείας γραμμής στο λογάριθμο της θνησιμότητας από καρκίνο του λάρυγγα παρά στη θνησιμότητα:

$$\ln(Y) = A_0 + A_1X_1 + A_2X_2 \quad (3)$$

Ο ln(Y) είναι ο φυσικός λογάριθμος του Y. Στην ισότητα 3, τόσο το αριστερό όσο και το δεξιό μέρος της ισότητας μπορούν να λάβουν τιμές από -∞ έως +∞. Το Y, ωστόσο, μπορεί να λάβει μόνο θετικές τιμές, καθώς δεν υφίσταται ο λογάριθμος αρνητικών αριθμών. Η ισότητα 3 επιλύεται ως προς Y, λαμβάνοντας τον αντιλογάριθμο και των δύο μελών:

$$Y=e^{A_0+A_1X_1+A_2X_2} \quad (4)$$

Η ισότητα 4 επιτρέπει στο Y να λάβει μόνο θετικές τιμές. Πρέπει, πάντως, να σημειωθεί ότι για να επιτευχθεί η διευκόλυνση αυτή δεν χρησιμοποιείται, πλέον, ένα απλό γραμμικό μοντέλο, αλλά ένα εκθετικό μοντέλο, όπως φαίνεται στην ισότητα 4.

Η χρησιμοποίηση ενός εκθετικού μοντέλου αντί ενός γραμμικού έχει ως αποτέλεσμα και τη διαφορετική ερμηνεία των συντελεστών του μοντέλου. Στην εικόνα 1, η κλίση της ευθείας γραμμής, ο συντελεστής δηλαδή της απλής γραμμικής παλινδρόμησης, ισούται με 0,282 θανάτους ανά 10.000 ανθρωπο-έτη και είναι ένα απόλυτο μέτρο της αύξησης της θνησιμότητας από καρκίνο του λάρυγγα για κάθε επιπλέον τσιγάρο που καταναλώνεται καθημερινά. Εάν το μοντέλο της απλής γραμμικής παλινδρόμησης εφαρμοστεί στην περίπτωση όπου η μελετώμενη έκβαση είναι ενδεικτική, με την τιμή X=0 να αντιστοιχεί στους μη εκτεθειμένους (ή, αλλιώς, στην κατηγορία αναφοράς του μελετώμενου προσδιοριστή) και την τιμή X=1 να αντιστοιχεί στους εκτεθειμένους (ή, αλλιώς, στην ενδεικτική κατηγορία), τότε ο συντελεστής A_1 αντιστοιχεί στη διαφορά συχνοτήτων μεταξύ εκτεθειμένων και μη εκτεθειμένων:

$$\text{Εκτεθειμένοι: } Y_e=A_0+A_1X_1=A_0+A_11=A_0+A_1$$

$$\text{Μη εκτεθειμένοι: } Y_0=A_0+A_1X_1=A_0+A_10=A_0$$

$$\text{Διαφορά συχνοτήτων: } Y_e-Y_0=A_0+A_1-A_0=A_1$$

Εάν η ισότητα για τους μη εκτεθειμένους (όταν X=0) αφαιρεθεί από την ισότητα για τους εκτεθειμένους (όταν X=1), τότε ο συντελεστής A_1 ισούται με τη διαφορά των συχνοτήτων μεταξύ εκτεθειμένων και μη εκτεθειμένων. Έτσι, χωρίς τον οποιοδήποτε μετασχηματισμό, ο συντελεστής της απλής γραμμικής παλινδρόμησης μπορεί να ερμηνευτεί ως η διαφορά των συχνοτήτων μεταξύ εκτεθειμένων και μη εκτεθειμένων. Εάν χρησιμοποιηθεί, ωστόσο, ο λογαριθμικός μετασχηματισμός που φαίνεται στις ισότητες 3 και 4, τότε ο συντελεστής A_1 στο εκθετικό μοντέλο που προκύπτει δεν ερμηνεύεται ως διαφορά συχνοτήτων:

$$\text{Εκτεθειμένοι: } \ln(Y_e)=A_0+A_1X_1=A_0+A_11=A_0+A_1$$

$$\text{Μη εκτεθειμένοι: } \ln(Y_0)=A_0+A_1X_1=A_0+A_10=A_0$$

$$\text{Διαφορά συχνοτήτων: } \ln(Y_e)-\ln(Y_0)=A_0+A_1-A_0=A_1$$

$$\text{Λόγος συχνοτήτων: } \frac{Y_e}{Y_0}=e^{A_1}$$

Έτσι, στην περίπτωση του εκθετικού μοντέλου, ο αντιλογάριθμος του συντελεστή A_1 που λαμβάνεται όταν υψωθεί η σταθερά e στη δύναμη του συντελεστή A_1 ισούται με το λόγο των συχνοτήτων στους εκτεθειμένους σε σχέση με τους μη εκτεθειμένους. Φαίνεται, λοιπόν, ότι ο μετασχηματισμός του γραμμικού μοντέλου της απλής γραμμικής παλινδρόμησης σε εκθετικό μοντέλο έχει ως συνέπεια όχι μόνο την αποφυγή των αρνητικών τιμών της εξαρτημένης μεταβλητής, αλλά και τη διαφορετική ερμηνεία του συντελεστή παλινδρόμησης. Χωρίς το μετασχηματισμό του γραμμικού μοντέλου, ο συντελεστής παλινδρόμησης εκτιμά διαφορές συχνοτήτων, ενώ με το μετασχηματισμό σε εκθετικό μοντέλο εκτιμά λόγους συχνοτήτων.

4. ΛΟΓΙΣΤΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ

Στην περίπτωση του γραμμικού μοντέλου της απλής γραμμικής παλινδρόμησης, η εξαρτημένη μεταβλητή μπορεί, θεωρητικά, να λάβει τιμές από το -∞ έως το +∞, ενώ όταν το μοντέλο αυτό μετασχηματιστεί σε εκθετικό, η εξαρτημένη μεταβλητή μπορεί να λάβει τιμές από 0 έως +∞. Έτσι, το μετασχηματισμένο μοντέλο της γραμμικής παλινδρόμησης μπορεί να χρησιμοποιηθεί όταν η εξαρτημένη

---

* Η επίπτωση-πυκνότητα (incidence-density)[5,8–11,16,23] είναι ένα μέτρο συχνότητας που ποσοτικοποιεί την εμφάνιση ενός συμβάντος με τάξη αναφοράς μια πεπερασμένη ποσότητα πληθυσμο-χρόνου (population-time). Ο ιδιόμορφος αυτός χρόνος, δηλαδή ο πληθυσμο-χρόνος (που δεν είναι ημερολογιακός ή ηλικιακός), προκύπτει κατά την κίνηση ενός πληθυσμού (ανοικτού ή κλειστού) στον ημερολογιακό χρόνο και συνίσταται από άπειρο αριθμό προσωπο-στιγμών. Αποτελεί το άθροισμα των χρονικών περιόδων παρακολούθησης των μελών του πληθυσμού. Με δεδομένο ότι τάξη αναφοράς σε αυτό το μέτρο συχνότητας είναι ο πληθυσμο-χρόνος, η συχνότητα αφορά μόνο στην εμφάνιση συμβάντων και όχι καταστάσεων. Η επίπτωση-πυκνότητα δεν είναι καθαρός αριθμός, αλλά έχει αφενός αριθμητική τιμή και αφετέρου μονάδα μέτρησης που είναι το αντίστροφο του χρόνου [π.χ. (έτη)$^{-1}$].

** Η επίπτωση-ποσοστό (incidence proportion)[8–11,16,23] είναι το ποσοστό των προσωπο-στιγμών στην αρχή της παρακολούθησης (T_0=επιστημονικός χρόνος) που εμφάνισε τις περιπτώσεις πάθησης κατά τη διάρκεια μιας ορισμένης χρονικής περιόδου. Ο όρος προσωπο-στιγμή (person-moment ή instance) εισήχθη για να δηλώσει την ύπαρξη ενός συγκεκριμένου προσώπου σε μια συγκεκριμένη χρονική στιγμή σε ένα συγκεκριμένο τόπο. Οι περιπτώσεις της πάθησης της οποίας μελετάται η συχνότητα, είναι συμβάντα που παρατηρούνται κατά τη διάρκεια μιας περιόδου παρακολούθησης και ονομάζονται συμβάντα περιόδου (period events). Η επίπτωση-ποσοστό εφαρμόζεται μόνο σε κλειστούς πληθυσμούς και εφόσον ο αριθμός των συμβάντων περιόδου δεν είναι σχετικά μεγάλος. Αυτό το μέτρο συχνότητας έχει νόημα εφόσον η διάρκεια παρακολούθησης εμπεριέχεται στην έννοια της επίπτωσης-ποσοστού, όπως η βρεφική ή η νεογνική νοσηρότητα, αλλιώς θα πρέπει να εκφράζεται με σαφήνεια.

μεταβλητή είναι συχνότητα (επίπτωση-πυκνότητα*), καθώς η συχνότητα λαμβάνει τιμές από 0 έως +∞. Το μοντέλο αυτό, ωστόσο, δεν μπορεί να χρησιμοποιηθεί όταν η εξαρτημένη μεταβλητή είναι ποσοστό (επίπτωση-ποσοστό** ή επιπολασμός***), καθώς το ποσοστό (proportion, P) μπορεί να λάβει τιμές από 0–1. Για κάθε ευθεία γραμμή ενός γραμμικού μοντέλου, στο οποίο η κλίση δεν ισούται με 0, η εξαρτημένη μεταβλητή Y μπορεί να λάβει τιμές από -∞ έως +∞ και όχι από 0–1. Επομένως, ένα γραμμικό μοντέλο, χωρίς τον απαιτούμενο μετασχηματισμό, μπορεί να οδηγήσει σε τιμές κινδύνου για ένα άτομο είτε αρνητικές είτε >1. Ο συνηθέστερος μετασχηματισμός που πραγματοποιείται, με σκοπό οι προβλεπόμενες τιμές κινδύνου για ένα άτομο να κυμαίνονται στο επιτρεπτό εύρος (0–1), είναι το μοντέλο της λογιστικής παλινδρόμησης.

Είναι ευκολότερο να αντιληφθεί κάποιος το λογιστικό μετασχηματισμό εάν σκεφθεί ότι πραγματοποιούνται δύο διαδοχικοί μετασχηματισμοί. Ο πρώτος μετασχηματισμός επιτρέπει τη μετατροπή του ποσοστού σε ένα μέτρο, η τιμή του οποίου κυμαίνεται από 0 έως +∞ και όχι από 0–1, όπως συμβαίνει με το ποσοστό. Ο μετασχηματισμός αυτός επιτυγχάνεται, λαμβάνοντας το λόγο συμπληρωματικών πιθανοτήτων (odds***) του ποσοστού $\left(\frac{\mathrm{P}}{1-\mathrm{P}}\right)$ και όχι το ποσοστό αυτό καθεαυτό. Όταν η τιμή του ποσοστού πλησιάζει το 0, τότε η τιμή της ποσότητας σχεδόν ταυτίζεται με την τιμή του ποσοστού, ενώ όταν η τιμή του ποσοστού πλησιάζει το 1, τότε ο παρονομαστής της ποσότητας $\frac{P}{1-P}$ πλησιάζει το 0, οπότε η τιμή της ποσότητας $\frac{P}{1-P}$ προσεγγίζει το +∞. Ο δεύτερος μετασχηματισμός μετατρέπει το λόγο των συμπληρωματικών πιθανοτήτων του ποσοστού σε ένα μέτρο, η τιμή του οποίου κυμαίνεται από 0–1. Ο μετασχηματισμός αυτός είναι ίδιος με εκείνον που πραγματοποιήθηκε στην περίπτωση των συχνοτήτων, οπότε λαμβάνεται ο λογάριθμος του λόγου των συμπληρωματικών πιθανοτήτων. Το μέτρο που προκύπτει, έπειτα και από τους δύο μετασχηματισμούς, είναι ο $\ln\left[\frac{\mathrm{P}}{1-\mathrm{P}}\right]$ και ονομάζεται λότζιτ (logit). Επομένως, το logit είναι ο λογάριθμος του λόγου των συμπληρωματικών πιθανοτήτων. Ο μετασχηματισμός αυτός που περιλαμβάνει τα δύο παραπάνω βήματα ονομάζεται λογιστικός μετασχηματισμός. Το λογιστικό μοντέλο είναι εκείνο στο οποίο το logit είναι η εξαρτημένη μεταβλητή μιας ισότητας που αντιστοιχεί σε μια ευθεία γραμμή, της ισότητας δηλαδή 1:

$$\ln\left(\frac{P}{1-P}\right) = A_0 + A_1X_1 \qquad (5)$$

Η ισότητα 5 αντιστοιχεί σ' ένα μοντέλο παλινδρόμησης, στο οποίο η εξαρτημένη μεταβλητή είναι ποσοστό. Εφόσον η εξαρτημένη μεταβλητή Y είναι ποσοστό, η ισότητα 5 μπορεί να λάβει και την παρακάτω μορφή:

$$\ln\left(\frac{Y}{1-Y}\right) = A_0 + A_1X_1 \qquad (6)$$

Η ισότητα 6 είναι το μοντέλο της απλής λογιστικής παλινδρόμησης, καθώς περιλαμβάνει μόνο μία ανεξάρτητη μεταβλητή (X_1). Προφανώς, όπως και σε άλλα γραμμικά μοντέλα, είναι δυνατόν να συμπεριληφθούν περισσότερες από μία ανεξάρτητες μεταβλητές (X_1, X_2, X_3 ...), οπότε προκύπτει το μοντέλο της πολλαπλής λογιστικής παλινδρόμησης:

$$\ln\left(\frac{Y}{1-Y}\right) = A_0 + A_1X_1 + A_2X_2 + A_3X_3 + \ldots \qquad (7)$$

Σύμφωνα με τον ορισμό του logit, η τιμή του Y κυμαίνεται πάντοτε από 0–1, χωρίς να έχει σημασία η τιμή που λαμβάνει το δεξιό μέλος της ισότητας 6. Το γεγονός αυτό αποτελεί και το κύριο πλεονέκτημα της λογιστικής παλινδρόμησης, καθώς μπορεί να χρησιμοποιηθεί όταν η εξαρτημένη μεταβλητή μετράται ως ποσοστό.

Ιδιαίτερα σημαντική είναι η ερμηνεία των συντελεστών της λογιστικής παλινδρόμησης (logistic regression). Όταν η εξαρτημένη μεταβλητή X_1 είναι ενδεικτική (X_1=1 για τους εκτεθειμένους και X_1=0 για τους μη εκτεθειμένους), τότε ο συντελεστής A_1 ισούται με το λόγο των logits των εκτεθειμένων σε σχέση με τους μη εκτεθειμένους:

$$\ln\left(\frac{P_1}{1-P_1}\right) - \ln\left(\frac{P_0}{1-P_0}\right) = \ln\left(\frac{\frac{P_1}{1-P_1}}{\frac{P_0}{1-P_0}}\right) =$$

$$\ln\left[\frac{P_1\,(1-P_0)}{P_0\,(1-P_1)}\right] = A_1 \qquad (8)$$

** Όταν η συχνότητα της μελετώμενης πάθησης δεν αφορά σε ενάρξεις του νοσήματος (νέες περιπτώσεις), που αποτελούν σημεία στο χρόνο (συμβάντα), αλλά σε καταστάσεις που έχουν διάρκεια, τότε το μέτρο συχνότητας ονομάζεται επιπολασμός και είναι ποσοστό που προκύπτει από ένα κοινό κλάσμα, το οποίο αριθμητή έχει τις υπάρχουσες (νέες και παλαιές) περιπτώσεις πάθησης και παρονομαστή (τάξη αναφοράς) μια σειρά προσωπο-στιγμών.[8–11,16] Η βάση μελέτης στην περίπτωση αυτή είναι η τομή ενός πληθυσμού και ο υπολογισμός γίνεται σε μια ορισμένη στιγμή του ημερολογιακού χρόνου. Ο επιπολασμός μιας πάθησης αντιστοιχεί στο ποσοστό του πληθυσμού που έχει την πάθηση σε ένα ορισμένο σημείο στο χρόνο.

** Το odds[11,16] μιας πιθανότητας είναι ο λόγος των συμπληρωματικών πιθανοτήτων, οπότε εάν η πιθανότητα εμφάνισης ενός ενδεχομένου συμβολιστεί με p και η πιθανότητα μη εμφάνισης με (1–p), τότε το odds υπέρ του ενδεχομένου είναι $\frac{p}{1-p}$ ως προς 1.

Έτσι, στη λογιστική παλινδρόμηση, ο αντιλογάριθμος (e^{A_1}) του συντελεστή παλινδρόμησης (A_1) μιας ενδεικτικής ανεξάρτητης μεταβλητής (X_1) αποτελεί εκτίμηση του λόγου των odds στους εκτεθειμένους σε σχέση με τους μη εκτεθειμένους:

$$e^{A_1} = \frac{P_1\,(1-P_0)}{P_0\,(1-P_1)} \qquad (9)$$

Το μοντέλο της λογιστικής παλινδρόμησης παρέχει εκτιμήσεις του λόγου των odds και γι' αυτό χρησιμοποιείται συχνά στις μελέτες «ασθενών-μαρτύρων», καθώς και στις μελέτες όπου χρησιμοποιούνται δεδομένα επιπολασμού για τον υπολογισμό των κατάλληλων επιδημιολογικών μέτρων σχέσης.

Ένας περιορισμός της πολλαπλής λογιστικής παλινδρόμησης που δεν πρέπει να αγνοείται είναι η πολλαπλασιαστική σχέση των ανεξάρτητων μεταβλητών μεταξύ τους.[8] Καθώς η συνεισφορά κάθε ανεξάρτητης μεταβλητής στο συνολικό αποτέλεσμα είναι ο λογάριθμος του λόγου των odds, οι διάφορες ανεξάρτητες μεταβλητές στο μοντέλο της πολλαπλής λογιστικής παλινδρόμησης έχουν μια πολλαπλασιαστική σχέση μεταξύ τους αναφορικά με τη συχνότητα εμφάνισης της μελετώμενης έκβασης. Η πολλαπλασιαστική αυτή σχέση ισοδυναμεί με την υπόθεση ότι το υπολογιζόμενο μέτρο σχέσης είναι σταθερό στα επιμέρους στρώματα που δημιουργούνται με βάση τους πιθανούς συγχυτές. Η υπόθεση αυτή, που συνιστά και προϋπόθεση για την εφαρμογή της διαστρωματικής ανάλυσης, δεν αποτελεί ιδιαίτερο πρόβλημα, αλλά πρέπει να σημειωθεί ότι, σε αντίθεση με τη διαστρωμάτωση, το μοντέλο της λογιστικής παλινδρόμησης δεν επιτρέπει άμεση και εύκολη εκτίμηση του αν ισχύει η ομοιομορφία του μέτρου σχέσης στα επιμέρους στρώματα. Είναι δυνατόν, πάντως, να ελεγχθεί η υπόθεση της πολλαπλασιαστικής σχέσης μεταξύ των ανεξάρτητων μεταβλητών, εξετάζοντας το μέγεθος των συντελεστών για τους όρους εκείνους που αντιστοιχούν στο αποτέλεσμα της αλληλεπίδρασης μεταξύ δύο ανεξάρτητων μεταβλητών. Η χρήση του γενικού μοντέλου σχετικού κινδύνου του Thomas παρέχει τη δυνατότητα άμεσου ελέγχου της πολλαπλασιαστικότητας.[24]

Ανεξάρτητα, πάντως, από το αν η υπόθεση της πολλαπλασιαστικότητας αντιστοιχεί σε μια κατάλληλη μαθηματική περιγραφή των δεδομένων, η εκτίμηση της βιολογικής –σε αντίθεση με τη στατιστική– αλληλεπίδρασης ή, καλύτερα, βιολογικής συνεπίδρασης[6,8,10,11] μεταξύ των ενδεικτικών κατηγοριών των προσδιοριστών –ή αλλιώς μεταξύ των συνιστωσών αιτιών– γίνεται ιδιαίτερα πολύπλοκη εξαιτίας της πολλαπλασιαστικής φύσης του μοντέλου της λογιστικής παλινδρόμησης.[8,25]

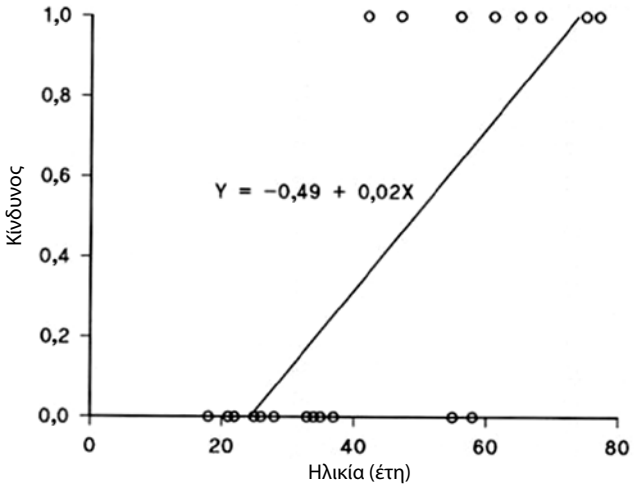
5. ΕΠΙΛΟΓΗ ΤΟΥ ΚΑΤΑΛΛΗΛΟΥ ΜΟΝΤΕΛΟΥ

Από μαθηματικής πλευράς, τα πλεονεκτήματα των παραπάνω μετασχηματισμών του γραμμικού μοντέλου εξισώνονται με τη μαθηματική συμπεριφορά των υπολογιζόμενων μέτρων, εξασφαλίζοντας ότι οι ατομικές εκτιμήσεις που προκύπτουν από τα μαθηματικά μοντέλα κυμαίνονται εντός των επιτρεπόμενων αριθμητικών ορίων.[8] Από πρακτική άποψη, ωστόσο, οι μαθηματικοί μετασχηματισμοί υποδηλώνουν το μέτρο σχέσης που εκτιμάται από τους συντελεστές του χρησιμοποιούμενου μοντέλου. Εάν τα δεδομένα μιας μελέτης αφορούν σε κινδύνους και αντικείμενο είναι η εκτίμηση της διαφοράς των κινδύνων, τότε το μοντέλο της λογιστικής παλινδρόμησης δεν είναι κατάλληλο, καθώς εκτιμά λόγους συμπληρωματικών πιθανοτήτων. Όταν όμως αντικείμενο είναι η εκτίμηση του κινδύνου εμφάνισης μιας πάθησης σ' ένα άτομο, τότε το πλέον κατάλληλο μοντέλο είναι εκείνο της λογιστικής παλινδρόμησης. Και αυτό γιατί ο κίνδυνος είναι πιθανότητα, οπότε δεν μπορεί να λάβει αρνητικές τιμές ή τιμές >1, κάτι που επιτυγχάνεται με την εφαρμογή του μοντέλου της λογιστικής παλινδρόμησης. Εάν το αντικείμενο μιας μελέτης, εξάλλου, είναι η εκτίμηση του συνολικού αποτελέσματος ενός προσδιοριστή, με βάση το συντελεστή παλινδρόμησης, τότε το ενδιαφέρον επικεντρώνεται, κυρίως, στο μέτρο σχέσης που εκτιμά το κάθε μοντέλο και λιγότερο στο αν όλες οι ατομικές εκτιμήσεις λαμβάνουν τιμές εντός των επιτρεπόμενων ορίων. Έτσι, σε αρκετές επιδημιολογικές μελέτες, η επιλογή του μοντέλου στηρίζεται ουσιαστικά στο υπολογιζόμενο μέτρο σχέσης.

Στον πίνακα 1[8] αναφέρεται η εμφάνιση ή μη μιας πάθησης σε 5 έτη και η ηλικία του ατόμου κατά την έναρξη της πενταετούς παρακολούθησης. Παρακολουθούνται 20 άτομα για 5 έτη και εμφανίζουν ή όχι τη μελετώμενη πάθηση. Τα δεδομένα της μελέτης αυτής απεικονίζονται στο διάγραμμα σημείων δύο κατευθύνσεων (scatter plot) της εικόνας 3. Σε ένα διάγραμμα σημείων δύο κατευθύνσεων, όταν η εξαρτημένη μεταβλητή είναι ενδεικτική, λαμβάνει τιμές είτε 1 (όταν εμφανίζεται η μελετώμενη έκβαση) είτε 0 (όταν δεν εμφανίζεται η μελετώμενη έκβαση), οπότε όλες οι παρατηρήσεις στον κάθετο άξονα πρέπει να «συμπίπτουν» στο 1 ή στο 0, αντίστοιχα. Στην εικόνα 3, εξάλλου, απεικονίζεται η γραμμή παλινδρόμησης και η αντίστοιχη ισότητα για τα δεδομένα του πίνακα 1 όταν εφαρμόζεται το μοντέλο της απλής γραμμικής παλινδρόμησης με εξαρτημένη μεταβλητή τον κίνδυνο εμφάνισης της πάθησης και ανεξάρτητη μεταβλητή την ηλικία. Η σταθερά (A_0 – 0,45) της γραμμικής παλινδρόμησης είναι η εκτιμώμενη τιμή του κινδύνου για εκείνους που έχουν ηλικία 0 έτη. Η τιμή της

Πίνακας 1. Εμφάνιση ή μη μιας πάθησης στη διάρκεια 5 ετών σε 20 άτομα.[2]

Άτομο	Ηλικία (έτη)	Πάθηση
1	18	Όχι
2	21	Όχι
3	22	Όχι
4	25	Όχι
5	26	Όχι
6	28	Όχι
7	33	Όχι
8	34	Όχι
9	35	Όχι
10	37	Όχι
11	42	Ναι
12	47	Ναι
13	55	Όχι
14	56	Ναι
15	58	Όχι
16	61	Ναι
17	65	Ναι
18	68	Ναι
19	75	Ναι
20	77	Ναι

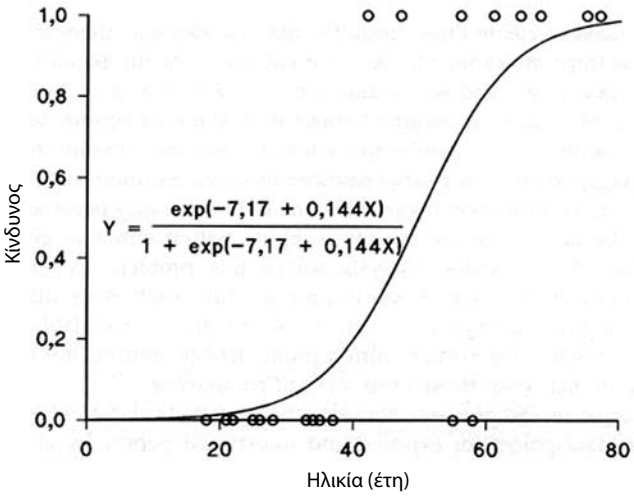


Εικόνα 3. Διάγραμμα σημείων δύο κατευθύνσεων για τα δεδομένα του πίνακα 1, όπου απεικονίζεται η ευθεία γραμμή της απλής γραμμικής παλινδρόμησης.

σταθεράς όμως είναι αρνητική, κάτι που είναι αδύνατο. Ουσιαστικά, με βάση το γραμμικό μοντέλο, η τιμή του κινδύνου είναι αρνητική για τα άτομα ηλικίας <24 ετών και >1 για τα άτομα ηλικίας >74 ετών.

Οι μη αποδεκτές αυτές τιμές, ωστόσο, του κινδύνου είναι δυνατόν να αποφευχθούν αν αντί του γραμμικού μοντέλου της εικόνας 3 χρησιμοποιηθεί το λογιστικό μοντέλο της εικόνας 4. Το σιγμοειδές σχήμα στην εικόνα 4 είναι χαρακτηριστικό της καμπύλης της λογιστικής παλινδρόμησης. Το σιγμοειδές αυτό σχήμα διατηρεί την τιμή του κινδύνου εντός των επιτρεπόμενων ορίων (0–1) για κάθε ηλικία, αποφεύγοντας έτσι τις μη αποδεκτές τιμές που προέρχονται από το γραμμικό μοντέλο της εικόνας 3.

Συγκρίνοντας το γραμμικό μοντέλο με το λογιστικό προκύπτει ότι το λογιστικό είναι καταλληλότερο όταν το υπολογιζόμενο μέτρο συχνότητας είναι ο κίνδυνος. Το παράδειγμα, όμως, του πίνακα 1 υποδηλώνει ότι το λογιστικό μοντέλο δεν είναι πάντοτε το πλέον κατάλληλο. Πιο συγκεκριμένα, ο συντελεστής παλινδρόμησης της ηλικίας, στο γραμμικό μοντέλο της εικόνας 3, ερμηνεύεται ως η διαφορά κινδύνων για κάθε επιπλέον ηλικιακό έτος. Ο συντελεστής παλινδρόμησης της ηλικίας υποδηλώνει ότι ο κίνδυνος αυξάνει κατά 2% για κάθε επιπλέον ηλικιακό έτος. Στην πραγματικότητα, πάντως, δεν αναμένεται η ευθεία γραμμή της εικόνας 3 να προσαρμόζεται με τον καλύτερο δυνατό τρόπο στα δεδομένα της μελέτης, εκτός από την κεντρική περιοχή του γραφήματος. Επιπλέον, για την κεντρική αυτή περιοχή, η ευθεία γραμμή παρέχει έναν απλό και χρήσιμο τρόπο για την εκτίμηση της διαφοράς των κινδύνων για κάθε επιπλέον ηλικιακό έτος. Αντιθέτως, το λογιστικό μοντέλο της εικόνας 4 δεν επιτρέπει την εκτίμηση της διαφοράς των κινδύνων, αλλά την εκτίμηση του λόγου των odds που σχετίζεται με την αύξηση της ηλικίας κατά ένα έτος. Ο λόγος αυτός ισούται με τον αντιλογάριθμο του συντελεστή παλινδρόμησης της ηλικίας στο λογιστικό



Εικόνα 4. Διάγραμμα σημείων δύο κατευθύνσεων για τα δεδομένα του πίνακα 1, όπου απεικονίζεται η καμπύλη της απλής λογιστικής παλινδρόμησης.

μοντέλο που είναι $e^{0,144}$=1,15.

Έτσι, η επιλογή του μοντέλου εξαρτάται σε σημαντικό βαθμό από το μέτρο σχέσης που επιθυμεί ο ερευνητής να υπολογίσει. Το λογιστικό μοντέλο χρησιμοποιείται όταν το αντικείμενο της μελέτης είναι η εκτίμηση του λόγου των κινδύνων, ενώ το γραμμικό μοντέλο χρησιμοποιείται όταν το αντικείμενο είναι η εκτίμηση της διαφοράς των κινδύνων. Όπως προαναφέρθηκε, η λογιστική παλινδρόμηση είναι ιδιαίτερα χρήσιμη στις μελέτες «ασθενών-μαρτύρων», επειδή ο υπολογιζόμενος λόγος των odds, που είναι ουσιαστικά ο λόγος των οιονεί επιπτώσεων-πυκνοτήτων, μπορεί να χρησιμοποιηθεί για την εκτίμηση του πραγματικού λόγου των επιπτώσεων-πυκνοτήτων.

6. ΕΚΤΙΜΗΣΗ ΤΟΥ ΚΙΝΔΥΝΟΥ

Αρκετές είναι οι περιπτώσεις εκείνες, στις οποίες κατασκευάζονται μοντέλα με σκοπό την εκτίμηση του κινδύνου εμφάνισης μιας έκβασης σε κάθε μελετώμενο άτομο ξεχωριστά. Οι Murabito et al[26] κατέληξαν σε ένα μοντέλο πολλαπλής λογιστικής παλινδρόμησης που εκτιμούσε τον κίνδυνο εμφάνισης διαλείπουσας χωλότητας σε χρονικό διάστημα 4 ετών. Το λογιστικό αυτό μοντέλο φαίνεται στον πίνακα 2 και περιγράφεται από την ισότητα 7. Λύνοντας την ισότητα 7 ως προς την εξαρτημένη μεταβλητή προκύπτουν τα εξής:

$$Y=\frac{e^{A_1+A_1X_1+A_2X_2+A_3X_3+...}}{1+e^{A_1+A_1X_1+A_2X_2+A_3X_3+...}} \quad (10)$$

Έτσι, χρησιμοποιώντας την ισότητα 10 είναι εύκολο να υπολογιστεί ο κίνδυνος εμφάνισης διαλείπουσας χωλότητας σε ένα συγκεκριμένο άτομο σε χρονικό διάστημα 4 ετών. Εάν π.χ. ένας άνδρας είναι 70 ετών, μη καπνιστής, με φυσιολογική αρτηριακή πίεση, διαβητικός, πάσχει από στεφανιαία νόσο και το επίπεδο χοληστερόλης είναι ίσο με 250 mg/dL, τότε, με βάση την ισότητα 10, ο κίνδυνος να εμφανίσει διαλείπουσα χωλότητα σε διάστημα 4 ετών ισούται με:

$$Y=\text{κίνδυνος}=\frac{e^{-8,915+0,503\cdot1=0,037\cdot70+0,000\cdot0+0,950\cdot1+0,031\cdot0+0,005\cdot250+0,994\cdot1}}{1+e^{-8,915+0,503\cdot1=0,037\cdot70+0,000\cdot0+0,950\cdot1+0,031\cdot0+0,005\cdot250+0,994\cdot1}}=$$

$$=\frac{e^{-2,628}}{1+e^{-2,628}}=0,067$$

Άρα, ο κίνδυνος εμφάνισης διαλείπουσας χωλότητας στο χρονικό διάστημα των 4 ετών για το συγκεκριμένο άνδρα είναι 6,7%. Εάν ο άνδρας αυτός είχε υπέρταση σταδίου 2, αντί για φυσιολογική αρτηριακή πίεση, τότε ο κίνδυνος εμφάνισης διαλείπουσας χωλότητας θα ήταν 13,8%.

Το μοντέλο της πολλαπλής λογιστικής παλινδρόμησης του πίνακα 2 χρησιμοποιείται για την εκτίμηση του κινδύνου σε κάθε άτομο ξεχωριστά, λαμβάνοντας υπόψη τα ιδιαίτερα χαρακτηριστικά του. Για την εκτίμηση του κινδύνου δεν είναι απαραίτητο όλοι οι προβλεπτικοί προσδιοριστές του πίνακα 2 να σχετίζονται αιτιακά με τη μελετώμενη έκβαση, με την εμφάνιση δηλαδή διαλείπουσας χωλότητας. Στο μοντέλο του πίνακα 2 ορισμένοι από τους προβλεπτικούς προσδιοριστές δεν μπορούν να θεωρηθούν ως αιτίες. Η ηλικία και η στεφανιαία νόσος, για παράδειγμα, θεωρούνται μη αιτιακοί προβλεπτικοί προσδιοριστές της διαλείπουσας χωλότητας, αλλά αποτελούν και οι δύο σημαντικούς προβλεπτικούς προσδιοριστές για την εκτίμηση του κινδύνου εμφάνισης της μελετώμενης πάθησης. Για το λόγο αυτόν περιλαμβάνονται στο προβλεπτικό μοντέλο της πολλαπλής λογιστικής παλινδρόμησης. Άλλοι προβλεπτικοί προσδιοριστές, όπως το κάπνισμα, η αρτηριακή πίεση και ο διαβήτης, μπορεί να αποτελούν αιτίες της διαλείπουσας χωλότητας.

Πίνακας 2. Μοντέλο πολυμεταβλητής λογιστικής παλινδρόμησης για την εκτίμηση του κινδύνου εμφάνισης διαλείπουσας χωλότητας σε χρονικό διάστημα 4 ετών.[8]

Προβλεπτικοί προσδιοριστές	Συντελεστής
Σταθερά (α)	-8,915
Φύλο (γυναίκα=0, άνδρας=1)	0,503
Ηλικία	0,037
Αρτηριακή πίεση	
Φυσιολογική	0,000
Υψηλή φυσιολογική	0,262
Υπέρταση σταδίου 1	0,407
Υπέρταση σταδίου 2+	0,798
Διαβήτης	0,950
Τσιγάρα/ημέρα	0,031
Χοληστερόλη (mg/dL)	0,005
Στεφανιαία νόσος (όχι=0, ναι=1)	0,994

7. ΚΑΤΑΣΚΕΥΗ ΠΟΛΥΜΕΤΑΒΛΗΤΩΝ ΜΟΝΤΕΛΩΝ ΣΤΗΝ ΑΙΤΙΟΛΟΓΙΚΗ ΕΡΕΥΝΑ

Στην επιδημιολογία, τα πολυμεταβλητά μοντέλα κατασκευάζονται, συνήθως, για την εκτίμηση των επιδημιολογικών μέτρων σχέσης με ταυτόχρονη εξουδετέρωση των συγχυτών, ενώ χρησιμοποιούνται και για την εκτίμηση της αλληλεπίδρασης, είτε στατιστικής είτε βιολογικής.[8,9] Η εξαρτημένη μεταβλητή είναι, συνήθως, ενδεικτική –για παράδειγμα η εμφάνιση ή μη μιας πάθησης– οπότε λαμβάνει τιμές 0–1. Με 1 συμβολίζεται, συνήθως, η εμφάνιση της μελετώμενης έκβασης και με 0 η μη εμφάνισή της. Ο μελετώμενος προσδιοριστής και όλοι οι πιθανοί συγχυτές περιλαμβάνονται στο μοντέλο ως ανεξάρτητες, ή προβλεπτι-

κές, μεταβλητές. Αν και η κατασκευή ενός πολυμεταβλητού μοντέλου απαιτεί ιδιαίτερη ανάλυση, στη συνέχεια θα γίνει αναφορά στα σημαντικότερα σημεία.

7.1. «Κατασκευή των μεταβλητών»

Στην ανάλυση των επιδημιολογικών δεδομένων, μια μεταβλητή δεν είναι απαραίτητο να αντιστοιχεί σε έναν απλό όρο σε ένα πολυμεταβλητό μοντέλο. Για κάθε μεταβλητή απαιτείται ένας τουλάχιστον όρος ή και περισσότεροι.[8] Ο απλούστερος τύπος μεταβλητών είναι οι ενδεικτικές (π.χ. το φύλο) που έχουν μόνο δύο κατηγορίες (άνδρες και γυναίκες). Μια τέτοια μεταβλητή περιλαμβάνεται στο πολυμεταβλητό μοντέλο ως ένας απλός όρος. Οι τιμές 0 και 1 αντιστοιχούν στις δύο κατηγορίες του χαρακτηριστικού και αποτελούν την αριθμητική απεικόνιση της μεταβλητής «φύλο» στο πολυμεταβλητό μοντέλο. Στην περίπτωση των μεταβλητών που έχουν τουλάχιστον τρεις κατηγορίες (όπως για παράδειγμα η ομάδα αίματος ενός ατόμου, που μπορεί να είναι Α, Β, ΑΒ ή Ο) δεν μπορούν να δοθούν αριθμητικές τιμές στις τέσσερις κατηγορίες και να περιληφθεί η μεταβλητή «ομάδα αίματος» ως ένας απλός όρος. Και αυτό επειδή οι τέσσερις κατηγορίες δεν αναπαριστούν ποσοτικές πλευρές ενός απλού μέτρου, οπότε ένας και μόνο όρος δεν επιτρέπει την πλήρη περιγραφή του αποτελέσματος μιας μεταβλητής. Όταν μια μεταβλητή έχει τέσσερις κατηγορίες, όπως η ομάδα αίματος, τότε απαιτούνται τρεις όροι στο πολυμεταβλητό μοντέλο. Μία από τις τέσσερις κατηγορίες, π.χ. η ομάδα αίματος Ο, επιλέγεται ως κατηγορία αναφοράς (reference category). Οι τρεις όροι στο μοντέλο δηλώνουν την παρουσία μίας από τις υπόλοιπες τρεις ομάδες αίματος, Α, Β ή ΑΒ. Οι τρεις όροι που δηλώνουν τη μεταβλητή «ομάδα αίματος» συμβολίζονται με X_1, X_2 και X_3. Το X_1 αναπαριστά την ομάδα αίματος Α, το X_2 την ομάδα αίματος Β και το X_3 την ομάδα αίματος ΑΒ. Εάν ένα άτομο ανήκει στην ομάδα αίματος Ο, δηλαδή στην κατηγορία αναφοράς, τότε οι όροι X_1, X_2 και X_3 λαμβάνουν την τιμή Ο. Εάν ένα άτομο ανήκει στην ομάδα αίματος Α, τότε ο όρος X_1 λαμβάνει την τιμή 1 και οι όροι X_2 και X_3 λαμβάνουν την τιμή Ο κ.λπ. Οι συντελεστές παλινδρόμησης του εφαρμοζόμενου μοντέλου για τους όρους X_1, X_2 και X_3 αντιστοιχούν στα μέτρα σχέσης των ομάδων αίματος Α, Β και ΑΒ ως προς την ομάδα αίματος Ο. Δεν έχει ιδιαίτερη σημασία, συνήθως, ποια κατηγορία επιλέγεται ως κατηγορία αναφοράς. Είναι προτιμότερο, πάντως, να επιλέγεται η κατηγορία εκείνη που έχει τα περισσότερα άτομα, οπότε αυξάνεται η στατιστική σταθερότητα των συντελεστών παλινδρόμησης ενός μοντέλου. Εάν η μεταβλητή έχει μια «φυσική» κατηγορία αναφοράς, τότε πρέπει να επιλέγεται η κατηγορία αυτή ως κατηγορία αναφοράς, ακόμη και αν δεν έχει τα περισσότερα άτομα. Έτσι, σε μια μελέτη εκτίμησης του κινδύνου από τη συμμετοχή σε διάφορες αθλητικές δραστηριότητες, εκείνα τα άτομα που δεν συμμετέχουν στις δραστηριότητες αυτές αποτελούν τη «φυσική» κατηγορία αναφοράς, ανεξάρτητα από τον αριθμό των ατόμων σε κάθε κατηγορία.

Γενικά, εάν μια μεταβλητή έχει n κατηγορίες, τότε απαιτούνται n–1 όροι σε ένα μοντέλο για να περιγραφεί το αποτέλεσμα καθεμιάς από τις n–1 κατηγορίες της μεταβλητής σε σχέση με την κατηγορία αναφοράς. Εάν περιλαμβάνονται όροι και για τις n κατηγορίες, τότε το μοντέλο εμφανίζει πλεονασμό και είναι αδύνατη η επίλυση του συστήματος των εξισώσεων που προκύπτει χωρίς την προσθήκη επιπλέον μαθηματικών προϋποθέσεων. Οι κατηγορίες μιας μεταβλητής μπορούν να θεωρηθούν ως μια ομάδα μεταβλητών και το αποτέλεσμα της καθεμιάς εκτιμάται ξεχωριστά σε σχέση με μια αυθαίρετα επιλεγμένη κατηγορία αναφοράς. Οι κατηγορίες πρέπει να είναι αμοιβαία αποκλειόμενες και το σύνολό τους να καλύπτει το δειγματικό χώρο.

Ανεξάρτητα από τη φύση του μελετώμενου προσδιοριστή, αρχικά πρέπει να εκτιμηθεί το είδος της καμπύλης που συσχετίζει το μελετώμενο προσδιοριστή με το υπολογιζόμενο μέτρο σχέσης. Για παράδειγμα, στη λογιστική παλινδρόμηση όταν ο μελετώμενος προσδιοριστής είναι μια συνεχής μεταβλητή, τότε πρέπει οπωσδήποτε να σχετίζεται εκθετικά με το λόγο των odds. Ένας απλός τρόπος να επιλυθεί το πρόβλημα που αφορά στις συνεχείς μεταβλητές είναι να πραγματοποιηθεί η κατηγοριοποίησή τους, οπότε στη συνέχεια λαμβάνονται υπόψη ως κατηγορικές μεταβλητές. Στην περίπτωση αυτή, είναι δυνατόν να συμπεριληφθούν περισσότεροι από ένας όροι στο πολυμεταβλητό μοντέλο. Έτσι, αν η συνεχής μεταβλητή είναι η ημερήσια κατανάλωση τσιγάρων, τότε μπορούν να δημιουργηθούν, για παράδειγμα, τέσσερις κατηγορίες: 0 τσιγάρα/ημέρα (μη καπνιστές), 1–20 τσιγάρα/ημέρα, 21–40 τσιγάρα/ημέρα και >40 τσιγάρα/ημέρα. Με τον τρόπο αυτόν, η συνεχής μεταβλητή «ημερήσια κατανάλωση τσιγάρων» μετατρέπεται σε μια μεταβλητή με τέσσερις κατηγορίες. Είναι προφανές ότι η κατηγοριοποίηση και ο αριθμός των κατηγοριών αποτελούν αυθαίρετες επιλογές του ερευνητή. Θεωρώντας ως κατηγορία αναφοράς τους μη καπνιστές και συμβολίζοντας την κατηγορία αυτή με 0, οι υπόλοιπες κατηγορίες συμβολίζονται με 1, 2 και 3, αντίστοιχα, οπότε στη συνέχεια εφαρμόζεται η μεθοδολογία που προαναφέρθηκε εκτενώς για τις μεταβλητές που έχουν τουλάχιστον τρεις κατηγορίες.

7.2. Εφαρμογή της διαστρωματικής ανάλυσης

Πριν από την κατασκευή ενός πολυμεταβλητού μοντέλου

απαιτείται η διαστρωματική ανάλυση των δεδομένων.[1–5,7–9,11] Το σημαντικότερο πλεονέκτημα της πολυμεταβλητής ανάλυσης είναι η ταυτόχρονη εξουδετέρωση της συγχυτικής δράσης πολλών χαρακτηριστικών. Η διαστρωματική ανάλυση, αντίθετα, επιτρέπει την ταυτόχρονη εξουδετέρωση μικρού αριθμού συγχυτών. Ακόμη, πάντως, και όταν οι συγχυτές είναι πολλοί, απαιτείται η εφαρμογή της διαστρωματικής ανάλυσης για δύο, τουλάχιστον, χαρακτηριστικά που θεωρείται ότι προκαλούν το μεγαλύτερο ποσοστό της σύγχυσης.

7.3. Επιλογή των συγχυτών

Αρχικά, αναλύεται η σχέση κάθε εξωγενούς προσδιοριστή (ή δυνητικού συγχυτή) ξεχωριστά με τη μελετώμενη έκβαση και με βάση τη στατιστική σημαντικότητα επιλέγεται, τελικά, ένα σύνολο δυνητικών συγχυτών. Έπειτα κατασκευάζεται ένα μοντέλο με τη διαδοχική προσθήκη ενός δυνητικού συγχυτή κάθε φορά. Μετά την προσθήκη ενός δυνητικού συγχυτή εξετάζεται ο βαθμός της μεταβολής του συντελεστή παλινδρόμησης του μελετώμενου προσδιοριστή. Εάν ο συντελεστής παλινδρόμησης του προσδιοριστή μεταβληθεί σημαντικά (οι περισσότεροι θεωρούν σημαντική μια μεταβολή της τάξης του 10%), τότε ο δυνητικός συγχυτής προστίθεται στο μοντέλο ως πραγματικός συγχυτής. Η επιλογή των συγχυτών με τον τρόπο αυτόν προϋποθέτει ότι ο προσδιοριστής εισάγεται στο μοντέλο ως ένας απλός όρος. Εάν, για παράδειγμα, ο προσδιοριστής είναι η ημερήσια κατανάλωση τσιγάρων, τότε πρέπει να εισαχθεί στο μοντέλο ένας απλός όρος, ο οποίος ποσοτικοποιεί την κατανάλωση τσιγάρων, και όχι περισσότεροι του ενός όροι για διάφορα επίπεδα κατανάλωσης τσιγάρων.

Στην πολυμεταβλητή ανάλυση είναι ιδιαίτερα συχνή η χρήση των «μοντέλων διαδοχικών σταδίων» (stepwise models). Η διαδοχική δημιουργία ενός πολυμεταβλητού μοντέλου χρησιμοποιεί έναν αλγόριθμο, ο οποίος αυτόματα επιλέγει τους όρους που περιλαμβάνονται στο τελικό μοντέλο. Η επιλογή των όρων με βάση τον αλγόριθμο στηρίζεται, ουσιαστικά, στο επίπεδο της στατιστικής σημαντικότητας του συντελεστή παλινδρόμησης του κάθε όρου ξεχωριστά. Η προσέγγιση των διαδοχικών σταδίων, πάντως, είναι περισσότερο λογικό να χρησιμοποιείται για τη δημιουργία ενός προβλεπτικού μοντέλου παρά ενός αιτιακού. Ο στατιστικός έλεγχος δεν επιτρέπει να εκτιμηθεί ξεχωριστά το μέγεθος μιας σχέσης και η ακρίβεια μιας εκτίμησης, καθώς αναμειγνύει τις δύο αυτές έννοιες. Δεν συνιστάται η χρήση των επιπέδων της στατιστικής σημαντικότητας για την εισαγωγή των συγχυτών σε ένα μοντέλο είτε χρησιμοποιείται ένας λογάριθμος διαδοχικών σταδίων είτε όχι. Και αυτό γιατί ο βαθμός της σύγχυσης εξαρτάται από δύο σχέσεις, τη σχέση μεταξύ του δυνητικού συγχυτή και του μελετώμενου προσδιοριστή και τη σχέση μεταξύ του δυνητικού συγχυτή και της μελετώμενης έκβασης, ενώ ο συντελεστής παλινδρόμησης, ο οποίος ελέγχεται για σημαντικότητα σε ένα μοντέλο διαδοχικών σταδίων, εκτιμά τη σχέση μόνο μεταξύ του δυνητικού συγχυτή και της μελετώμενης έκβασης και αγνοεί τη σχέση μεταξύ του δυνητικού συγχυτή και του μελετώμενου προσδιοριστή. Έτσι, με τον τρόπο αυτόν μπορεί να συμπεριληφθούν στο πολυμεταβλητό μοντέλο συγχυτές που στην πραγματικότητα δεν είναι, ενώ μπορεί να μη συμπεριληφθούν συγχυτές οι οποίοι στην πραγματικότητα είναι, αλλά η σχέση τους με τη μελετώμενη έκβαση δεν είναι στατιστικά σημαντική.

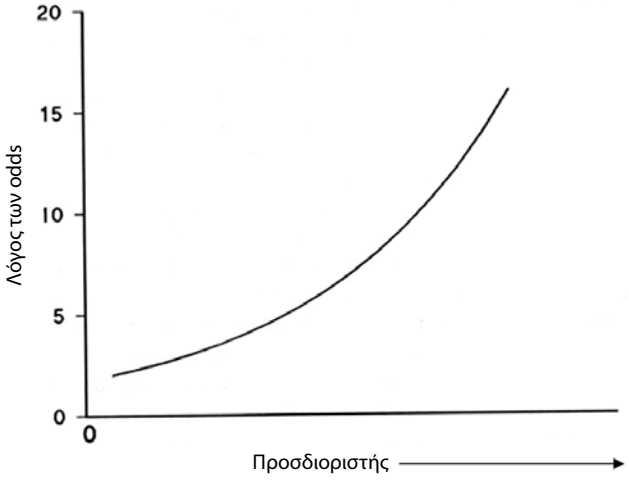
7.4. Εκτίμηση της σχέσης μεταξύ προσδιοριστή και πάθησης

Εάν ο μελετώμενος προσδιοριστής είναι ενδεικτική μεταβλητή, τότε, μετά από την προσθήκη των συγχυτών στο πολυμεταβλητό μοντέλο, είναι δυνατή η απευθείας εκτίμηση του αποτελέσματος του προσδιοριστή από το συντελεστή του όρου του προσδιοριστή. Εάν όμως ο προσδιοριστής είναι συνεχής μεταβλητή, όπως π.χ. ο αριθμός των τσιγάρων που καταναλώνονται καθημερινά, τότε ο όρος που αφορά στον προσδιοριστή πρέπει να καθοριστεί εκ νέου, έπειτα από την προσθήκη των συγχυτών στο μοντέλο. Ο λόγος για τον οποίο απαιτείται ο επανακαθορισμός του όρου που αφορά στον προσδιοριστή είναι ότι ο απλός όρος του προσδιοριστή που υπήρχε στο μοντέλο για την εκτίμηση της σύγχυσης δεν αποκαλύπτει την καμπύλη που σχετίζει την πάθηση με τον προσδιοριστή όταν ο μελετώμενος προσδιοριστής είναι συνεχής μεταβλητή. Εάν το εφαρμοζόμενο μοντέλο περιλαμβάνει λογαριθμικό μετασχηματισμό, όπως η πλειονότητα των μοντέλων που χρησιμοποιούνται στην ανάλυση των επιδημιολογικών δεδομένων, τότε ένας απλός όρος για μια συνεχή μεταβλητή (το μελετώμενο προσδιοριστή) μπορεί να περιοριστεί μαθηματικά ώστε να λάβει την καμπύλη που υπαγορεύει το μαθηματικό μοντέλο. Σε ένα μοντέλο λογιστικής παλινδρόμησης, ο συντελεστής του προσδιοριστή είναι ο λογάριθμος του λόγου των odds για κάθε μονάδα μεταβολής στη μεταβλητή που αφορά στον προσδιοριστή. Εάν ο προσδιοριστής είναι ο αριθμός των τσιγάρων που καταναλώνονται καθημερινά, τότε ο συντελεστής ενός απλού όρου που αντιστοιχεί στην ημερήσια κατανάλωση τσιγάρων ισούται με το λογάριθμο του λόγου των odds για κάθε επιπλέον τσιγάρο που καταναλώνεται. Επειδή υπάρχει μόνο ένας απλός όρος, το μοντέλο υπαγορεύει ότι το αποτέλεσμα του κάθε τσιγάρου πολλαπλασιάζει το λόγο των odds με μια σταθερή ποσότητα. Το αποτέλεσμα

είναι μια εκθετική καμπύλη μεταξύ της πάθησης και του μελετώμενου προσδιοριστή που εκφράζεται μέσω του λόγου των odds (εικ. 5).

Η εκθετική αυτή καμπύλη προσαρμόζει τα δεδομένα της μελέτης, ανεξάρτητα από την πραγματική καμπύλη που σχετίζει την πάθηση με τον προσδιοριστή, εφόσον η μεταβλητή που αφορά στον προσδιοριστή είναι συνεχής και περιορίζεται σε έναν απλό όρο σε ένα μοντέλο που χρησιμοποιεί λογαριθμικό μετασχηματισμό. Στα γραμμικά μοντέλα, μια γραμμική σχέση και όχι μια εκθετική μπορεί να αποτελέσει την καλύτερη λύση. Το πρόβλημα υπάρχει στις περιπτώσεις εκείνες, στις οποίες η πραγματική σχέση που ορίζεται από τη φύση δεν αντιστοιχεί στη σχέση που ορίζεται από την καμπύλη του εφαρμοζόμενου μοντέλου.

Είναι ανάγκη η καμπύλη που σχετίζει την πάθηση με τον προσδιοριστή να μην υπαγορεύεται από το μαθηματικό μοντέλο, αλλά από τα υπάρχοντα δεδομένα. Για να επιτευχθεί αυτό, απαιτείται ο ερευνητής να καθορίσει εκ νέου τον όρο του προσδιοριστή στο μοντέλο. Η συνηθέστερη μέθοδος είναι να κατηγοριοποιηθεί ο προσδιοριστής και να αποδοθούν ξεχωριστοί όροι σε κάθε κατηγορία του προσδιοριστή, εκτός από μια αυθαίρετα επιλεγμένη κατηγορία που αποτελεί την κατηγορία αναφοράς. Για παράδειγμα, η κατανάλωση τσιγάρων μπορεί να κατηγοριοποιηθεί στις εξής κατηγορίες: 0 τσιγάρα/ημέρα (μη καπνιστές), 1–9 τσιγάρα/ημέρα, 10–19 τσιγάρα/ημέρα κ.λπ. Στο μοντέλο θα υπάρχει ένας ξεχωριστός όρος για κάθε κατηγορία, εκτός από την κατηγορία των μη καπνιστών (0 τσιγάρα/ημέρα), που αποτελεί την κατηγορία αναφοράς. Η μεταβλητή που αντιστοιχεί σε κάθε όρο είναι πλέον ενδεικτική, φανερώνοντας απλά την κατηγορία στην οποία ανήκει ένα άτομο. Έτσι, ένα άτομο θα έχει την τιμή 1 για την κατηγορία στην οποία ανήκει και την τιμή 0 για όλες τις υπόλοιπες κατηγορίες, ενώ ένας μη καπνιστής θα έχει την τιμή 0 για όλες τις κατηγορίες. Οι συντελεστές παλινδρόμησης που προκύπτουν, αποδίδουν το ξεχωριστό αποτέλεσμα για κάθε επίπεδο του καπνίσματος που καθορίζεται από τα δεδομένα και όχι από τις μαθηματικές ιδιότητες του εφαρμοζόμενου μοντέλου.



Εικόνα 5. Διάγραμμα λογιστικής παλινδρόμησης που απεικονίζει τη θετική σχέση μεταξύ πάθησης και προσδιοριστή, εφόσον ο προσδιοριστής είναι συνεχής μεταβλητή.

7.5. Εκτίμηση της συνεπίδρασης

Για την κατάλληλη εκτίμηση της αλληλεπίδρασης ή, καλύτερα, της συνεπίδρασης μεταξύ των ενδεικτικών κατηγοριών δύο προσδιοριστών απαιτείται να καθοριστούν εκ νέου οι δύο προσδιοριστές και να δημιουργηθεί μια νέα σύνθετη μεταβλητή που να αφορά σε ένα νέο προσδιοριστή. Για δύο ενδεικτικούς προσδιοριστές Α και Β, η νέα σύνθετη μεταβλητή θα πρέπει να έχει τέσσερις κατηγορίες: (α) μη έκθεση στους Α και Β, (β) έκθεση στον Α και όχι στο Β, (γ) έκθεση στο Β και όχι στον Α, (δ) ταυτόχρονη έκθεση και στον Α και στο Β. Κάθε άτομο θα ανήκει σε μία από τις κατηγορίες που προκύπτουν από την κοινή έκθεση και στους δύο προσδιοριστές Α και Β. Χρησιμοποιώντας ως κατηγορία αναφοράς τη μη έκθεση στους Α και Β, το μαθηματικό μοντέλο παρέχει εκτιμήσεις του σχετικού αποτελέσματος για καθεμιά από τις άλλες τρεις κατηγορίες. Η μέθοδος αυτή επιτρέπει τη χρήση του πολυμεταβλητού μοντέλου για την εκτίμηση των αποκλίσεων από το προσθετικό μοντέλο (additive model) χωρίς να επιβάλλει την πολλαπλασιαστική σχέση που υπαγορεύεται από το χρησιμοποιούμενο μοντέλο.[27]

8. ΣΥΝΟΨΗ

Το κυριότερο πλεονέκτημα των πολυμεταβλητών μαθηματικών μοντέλων στην ανάλυση των επιδημιολογικών δεδομένων είναι η ευκολία με την οποία εξουδετερώνουν πολλούς συγχυτές ταυτόχρονα. Η ένταξη πολλών μεταβλητών σε ένα πολυμεταβλητό μοντέλο επιτρέπει το αποτέλεσμα της κάθε μεταβλητής να μη συγχέεται από τη δράση των υπόλοιπων μεταβλητών. Με τον τρόπο αυτόν επιτυγχάνεται η εξουδετέρωση της σύγχυσης που προκαλείται από πολλά χαρακτηριστικά ταυτόχρονα, κάτι ανέφικτο με τη μέθοδο της διαστρωματικής ανάλυσης. Εάν υπάρχουν, για παράδειγμα, 5 συγχυτές και καθένας έχει 3 κατηγορίες, τότε με την εφαρμογή της διαστρωμάτωσης προκύπτουν 3×3×3×3×3=243 στρώματα. Όταν ο αριθμός των στρωμάτων είναι τόσο μεγάλος, τότε ο αριθμός των δεδομένων σε ορισμένα στρώματα ενδεχομένως να είναι πολύ μικρός, οπότε η χρήση της διαστρωματικής ανάλυσης

οδηγεί σε μη έγκυρα συμπεράσματα. Στην περίπτωση αυτή είναι προτιμότερη η χρήση της πολυμεταβλητής ανάλυσης. Η διαστρωματική ανάλυση, πάντως, χρησιμοποιείται για έναν ή δύο, τουλάχιστον, συγχυτές που θεωρούνται οι σημαντικότεροι. Πρέπει να σημειωθεί, εξάλλου, ότι η πολυμεταβλητή ανάλυση είναι περισσότερο ευαίσθητη στο σφάλμα σε σχέση με τη διαστρωματική.

Η διαστρωματική ανάλυση παρουσιάζει ορισμένα πλεονεκτήματα έναντι της πολυμεταβλητής. Πιο συγκεκριμένα, με τη διαστρωματική ανάλυση, τόσο ο ερευνητής όσο και οι αναγνώστες αντιλαμβάνονται εύκολα και άμεσα την κατανομή των δεδομένων με βάση τις σημαντικότερες, τουλάχιστον, μεταβλητές, εφόσον βεβαίως τα στρωματοποιημένα δεδομένα παρουσιάζονται γραπτώς. Στην πολυμεταβλητή ανάλυση δεν συμβαίνει κάτι αντίστοιχο, με αποτέλεσμα οι αναγνώστες και σε αρκετές περιπτώσεις και οι ίδιοι οι ερευνητές να έχουν περιορισμένες γνώσεις όσον αφορά στην κατανομή των δεδομένων. Για το λόγο αυτόν η πολυμεταβλητή ανάλυση συνιστάται ως συμπλήρωμα της διαστρωματικής, η οποία πρέπει να είναι το πρωταρχικό εργαλείο στην ανάλυση των επιδημιολογικών δεδομένων. Ακόμη και όταν οι πιθανοί συγχυτές μιας μελέτης είναι αρκετοί, απαιτείται η εφαρμογή της διαστρωματικής ανάλυσης για τους δύο, τουλάχιστον, σημαντικότερους συγχυτές, οι οποίοι συνήθως είναι το φύλο και η ηλικία. Δυστυχώς, όμως, αρκετοί ερευνητές εστιάζονται μόνο στην εφαρμογή της πολυμεταβλητής ανάλυσης, αδιαφορώντας έτσι για τη χρησιμότητα της διαστρωματικής. Στις περισσότερες περιπτώσεις, μάλιστα, οι αναγνώστες έχουν στη διάθεσή τους μόνο τους συντελεστές παλινδρόμησης του πολυμεταβλητού μοντέλου, γεγονός που δεν επιτρέπει την εξαγωγή ασφαλών συμπερασμάτων.

Συμπερασματικά, η πολυμεταβλητή ανάλυση είναι εξαιρετικά χρήσιμη μόνον όταν χρησιμοποιείται ως συμπλήρωμα της διαστρωματικής και όχι ως πρωταρχικό εργαλείο για την ανάλυση επιδημιολογικών δεδομένων.

ABSTRACT

Multivariate analysis of epidemiological data

P. GALANIS

Center for Health Services Management and Evaluation, Department of Nursing, University of Athens, Athens, Greece

Archives of Hellenic Medicine 2009, 26(3):407–422

In epidemiology, mathematical models are used for various purposes. The two primary purposes are prediction and control for confounding. Prediction models are used to estimate risk based on information from risk predictors. In contrast to the goal of risk prediction for specific individuals, much epidemiologic research is aimed at learning about the causal role of specific characteristics (or determinants) for disease. In causal research, multivariate mathematical models are used to evaluate the causal role of one or more characteristics while simultaneously controlling for possible confounding effects of other characteristics. In a multivariate model, the inclusion of several variates results in each item being unconfounded by the other items. This is an easy and efficient approach to controlling confounding by several variates at once, something that might be difficult to achieve through a stratified analysis. However, with stratified analysis, both the investigator and the readers (when the stratified data are presented in a paper) are aware of the distribution of the data according to the key study variates. For this reason, a multivariate analysis should be used as a supplement to a stratified analysis, rather than as the primary analytical tool. In epidemiology, the most frequently used models are the general linear model and the logistic regression model. The outcome (or dependent variate) in the general linear model is continuous, while in logistic regression it is the indicator variate.

Key words: Logistic regression, Multivariate analysis, Regression, Stratified analysis

Βιβλιογραφία

1. ΓΑΛΑΝΗΣ Π, ΣΠΑΡΟΣ Λ. Διαστρωματική ανάλυση δεδομένων. *Αρχ Ελλ Ιατρ* 2005, 21:378–384
2. ΓΑΛΑΝΗΣ Π, ΣΠΑΡΟΣ Λ. Τροποποίηση του μέτρου αποτελέσματος και σύγχυση στην εφαρμοσμένη ιατρική έρευνα. *Αρχ Ελλ Ιατρ* 2005, 22:170–177
3. ΓΑΛΑΝΗΣ Π, ΣΠΑΡΟΣ Λ. Στατιστική αλληλεπίδραση και τροποποίηση του μέτρου αποτελέσματος. *Αρχ Ελλ Ιατρ* 2005, 21:137–147

4. AHLBOM A, NORELL S. *Εισαγωγή στη σύγχρονη επιδημιολογία*. Ιατρικές Εκδόσεις Λίτσας, Αθήνα, 1992
5. ΣΠΑΡΟΣ Λ, ΓΑΛΑΝΗΣ Π, ΖΑΧΟΣ Ι, ΤΣΙΛΙΔΗΣ Κ. *Επιδημιολογία Ι*. Εκδόσεις ΒΗΤΑ, Αθήνα, 2004
6. ΓΑΛΑΝΗΣ Π, ΣΠΑΡΟΣ Λ. Βιολογική και στατιστική αλληλεπίδραση. *Αρχ Ελλ Ιατρ* 2004, 21:123–136
7. ΓΑΛΑΝΗΣ Π. Αλληλεπίδραση παραγόντων κινδύνου στην επιδημιολογία. Μεταπτυχιακή διπλωματική εργασία, Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών, Τμήμα Νοσηλευτικής, Αθήνα, 2003
8. ROTHMAN KJ. *Modern epidemiology*. 1st ed. Little, Brown & Co, Boston, 1986
9. ROTHMAN KJ. *Epidemiology: An introduction*. Oxford University Press, New York, 2002
10. ROTHMAN KJ, GREENLAND S. *Modern epidemiology*. 2nd ed. Lippincott Williams & Wilkins, Philadelphia, 1998
11. ΣΠΑΡΟΣ Λ, ΓΑΛΑΝΗΣ Π. *Δοκίμια επιδημιολογίας*. Εκδόσεις Παρισιάνου, Αθήνα, 2006
12. ΓΑΛΑΝΗΣ Π, ΣΠΑΡΟΣ Λ. Διαστρωμάτωση επιδημιολογικών δεδομένων. *Αρχ Ελλ Ιατρ* 2006, 23:626–637
13. STRIKE PW. *Statistical methods in laboratory medicine*. Butterworth Heinemann, Cambridge, 1991
14. BLAND M. *An introduction to medical statistics*. 2nd ed. Oxford Medical Publications, Oxford, 1996
15. ROTHMAN KJ, CANN CI, FLANDERS D, FRIED MP. Epidemiology of laryngeal cancer. *Epidemiol Rev* 1980, 2:195–209
16. ΓΑΛΑΝΗΣ Π, ΣΠΑΡΟΣ Λ. Μέτρα συχνότητας των νοσημάτων. *Αρχ Ελλ Ιατρ* 2005, 22:178–191
17. ΣΠΑΡΟΣ Λ. *Μετα-επιδημιολογία*. Εκδόσεις ΒΗΤΑ, Αθήνα, 2001
18. MIETTINEN OS. *Theoretical epidemiology. Principles of occurrence research in medicine*. John Wiley & Sons, New York, 1985
19. GALTON F. *Natural inheritance*. McMillan, London, 1889
20. ΤΡΙΧΟΠΟΥΛΟΣ Δ. *Γενική και κλινική επιδημιολογία. Εγχειρίδιο επιδημιολογίας και αρχών κλινικής έρευνας*. Εκδόσεις Παρισιάνου, Αθήνα, 2002
21. ΔΡΑΚΑΤΟΣ ΚΓ. *Στατιστική*. 2η έκδοση. Εκδόσεις Παπαζήση, Αθήνα, 1984
22. ΓΑΛΑΝΗΣ Π, ΣΠΑΡΟΣ Λ. Συστηματικά σφάλματα στις επιδημιολογικές μελέτες. *Αρχ Ελλ Ιατρ* 2007, 24:373–388
23. ΓΑΛΑΝΗΣ Π, ΣΠΑΡΟΣ Λ. Στατιστικά μοντέλα για την ανάλυση των επιδημιολογικών δεδομένων. *Αρχ Ελλ Ιατρ* 2006, 23:404–417
24. THOMAS DC. General relative risk models for survival time and matched case-control analysis. *Biometrics* 1981, 37:673–676
25. GREENLAND S. Limitations of the logistic analysis of epidemiologic data. *Am J Epidemiol* 1979, 110:693–698
26. MURABITO JM, D'AGOSTINO RB, SIBERSHATZ H, WILSON WF. Intermittent claudication. A risk profile from the Framingham Heart Study. *Circulation* 1997, 96:44–49
27. ASSMAN SF, HOSMER DW, LEMESHOW S, MUNDT KA. Confidence intervals for measures of interaction. *Epidemiology* 1996, 7:286–290

Corresponding author:

P. Galanis, 14 Dikis street, GR-157 73 Athens, Greece
e-mail: pegalan@nurs.uoa.gr