Η ΧΡΗΣΗ ΤΩΝ ΜΟΝΤΕΛΩΝ ΛΟΓΑΡΙΘΜΙΚΗΣ ΠΑΛΙΝΔΡΟΜΗΣΗΣ ΣΤΙΣ ΕΜΠΕΙΡΙΚΕΣ ΕΡΕΥΝΕΣ ΤΩΝ ΚΟΙΝΩΝΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

Χαράλαμπος Γναρδέλλης

Τεχνολογικό Εκπαιδευτικό Ίδρυμα Μεσολογγίου

Περίληψη: Ένα ζήτημα που τίθεται συχνά στις ποσοτικές προσεγγίσεις των κοινωνιολογικών ερευνών αφορά τη διερεύνηση της σχέσης που υπάρχει μεταξύ μιας διχοτομικής κατηγορικής μεταβλητής και ενός συνόλου άλλων ερμηνευτικών μεταβλητών. Η συνήθης περίπτωση αυτών των διερευνήσεων εστιάζει στο ενδεχόμενο πραγματοποίησης ή μη ενός γεγονότος αναφορικά με τις επιδράσεις ενός συνόλου άλλων παραγόντων οι οποίοι δυνητικά επηρεάζουν την έκβασή του. Για παράδειγμα, η χρήση ή μη ενδοφλέβιων ναρκωτικών από κρατούμενους σωφρονιστικών ιδρυμάτων σε σχέση με τον ποινικό χαρακτηρισμό τους, το συνολικό χρόνο εγκλεισμού τους και το ενδεχόμενο προηγούμενης καταδίκης τους για ναρκωτικά. Στις περιπτώσεις αυτές, οι συνήθεις αναλύσεις γραμμικής παλινδρόμησης δεν είναι κατάλληλες για τη μελέτη του διχοτομικού αποτελέσματος, διότι οι προϋποθέσεις χρήσης τους δεν αντιστοιχούν στο επίπεδο μετρησιμότητας της διερευνώμενης εξαρτημένης μεταβλητής. Αναγκαία προϋπόθεση για τη χρήση ενός γραμμικού μοντέλου σε μια τέτοια περίπτωση είναι ο μετασχηματισμός της εξαρτημένης μεταβλητής σε ένα λόγο συμπληρωματικών πιθανοτήτων, δηλαδή το λόγο της πιθανότητας πραγματοποίησης του γεγονότος προς την πιθανότητα μη πραγματοποίησης. Ο λογάριθμος αυτού του λόγου (ο οποίος ονομάζεται logit) παίρνει τιμές στο σύνολο των πραγματικών αριθμών και μπορεί να αποτελέσει την εξαρτημένη μεταβλητή ενός συνήθους γραμμικού μοντέλου παλινδρόμησης.

Λέξεις κλειδιά: Γενικευμένα γραμμικά μοντέλα, Κοινωνιολογικές έρευνες, Λογαριθμική παλινδρόμηση.

Διεύθυνση: Χαράλαμπος Γναρδέλλης, Τεχνολογικό Εκπαιδευτικό Ίδρυμα Μεσολογγίου, Νέα Κτίρια, 302 00 Μεσολόγγι. Τηλ.: 210-7512651, 6972148215. E-mail: hgnardellis@yahoo.gr

ΕΙΣΑΓΩΓΗ

Συχνά στις πολυμεταβλητές αναλύσεις των εμπειρικών ερευνών το ενδιαφέρον εστιάζεται στην επίδραση που ασκεί ένα σύνολο ανεξάρτητων μεταβλητών $X_1, X_2,..., X_k$ στις τιμές μιας άλλης (δυνητικά) εξαρτημένης από αυτές μεταβλητής Y, όπως φαίνεται στην παρακάτω σχέση:

$$Y \longleftarrow (X_1, X_2,..., X_k)$$

Διερευνήσεις αυτού του τύπου, οι οποίες είναι συνήθεις στην ποσοτική ανάλυση, κατά κανόνα πραγματοποιούνται με την εφαρμογή τεχνικών μιας ευρύτερης "οικογένειας" θεωρητικών μοντέλων, τα οποία ονομάζονται *γενικευμένα γραμμικά μοντέλα*[1] (Dobson, 1990. McCullagh & Nelder, 1989). Το ενδιαφέρον και η χρησιμότητα των γενικευμένων γραμμικών μοντέλων επικεντρώνεται τόσο στην επάρκεια της προσαρμογής τους στα εμπειρικά δεδομένα όσο και στην ερμηνεία των συντελεστών τους, οι οποίοι ουσιαστικά εκφράζουν τη συμβολή κάθε ανεξάρτητης μεταβλητής στον προσδιορισμό των τιμών της εξαρτημένης. Γενικότερα, η θεωρία των γενικευμένων γραμμικών μοντέλων παρέχει ένα ενιαίο πλαίσιο μέσα από το οποίο μπορούν να κατασκευάζονται και να ελέγχονται μοντέλα των οποίων η εξαρτημένη μεταβλητή μπορεί να είναι οποιουδήποτε τύπου (π.χ., μια συνεχής ποσοτική μεταβλητή, μια *μεταβλητή συχνοτήτων*[2], η πιθανότητα πραγματοποίησης ενός γεγονότος, ή ακόμη και ο *ρυθμός*[3] πραγματοποίησης ενός γεγονότος στη μονάδα του χρόνου.

Τα γενικευμένα γραμμικά μοντέλα που χρησιμοποιούνται συχνότερα στην εμπειρική έρευνα, ανάλογα με το είδος της εξαρτημένης μεταβλητής, είναι: (α) η *πολλαπλή γραμμική παλινδρόμηση*[4], όταν η εξαρτημένη μεταβλητή είναι ποσοτική συνεχής, (β) τα *λογαριθμικά γραμμικά μοντέλα*[5], όταν η εξαρτημένη μεταβλητή είναι μία συχνότητα, (γ) η *παλινδρόμηση Poisson*[6], όταν η εξαρτημένη μεταβλητή είναι ο ρυθμός πραγματοποίησης ενός γε-

[1] Generalized linear models.
[2] Counts.
[3] Rate.
[4] Multiple linear regression.
[5] Log-linear models.
[6] Poisson regression.

γονότος στη μονάδα του χρόνου, (δ) τα *μοντέλα αναλογικής διακινδύνευσης*[7], όταν μελετάται η πιθανότητα έκβασης ενός γεγονότος στη μονάδα του χρόνου και, τέλος, (ε) η *λογαριθμική παλινδρόμηση*[8], όταν η εξαρτημένη μεταβλητή εκφράζει την πιθανότητα πραγματοποίησης ή μη ενός γεγονότος.

Η χρήση και η ερμηνεία των αποτελεσμάτων της λογαριθμικής παλινδρόμησης, στην οποία επικεντρώνεται η συγκεκριμένη εργασία, όπως και των υπόλοιπων γενικευμένων γραμμικών μοντέλων, μπορεί να γίνει αφού πρώτα αναφερθούν οι βασικές ιδιότητες της πολλαπλής γραμμικής παλινδρόμησης. Μέσα από τις γενικές ιδιότητες της πολλαπλής γραμμικής παλινδρόμησης μπορούν να ερμηνεύονται και τα υπόλοιπα μοντέλα, εφόσον πρώτα μετασχηματιστούν κατάλληλα ώστε να πάρουν τη μορφή ενός απλού γραμμικού μοντέλου.

Πολλαπλή γραμμική παλινδρόμηση

Στο μοντέλο της πολλαπλής γραμμικής παλινδρόμησης υποθέτουμε ότι η σχέση μιας ποσοτικής μεταβλητής Y και μιας σειράς k ανεξάρτητων μεταβλητών $X_1, X_2, ..., X_k$ είναι γραμμική. Η εκτίμηση των τιμών της Y για τα δεδομένα του δείγματος γίνεται από την εξίσωση $Y = b_0 + b_1X_1 + b_2X_2 + ... + b_kX_k + e$.

Το πρώτο μέρος της εξίσωσης που εκφράζει το μοντέλο συνοψίζει τη γραμμική σχέση της Y με τις μεταβλητές $X_1, X_2, ..., X_k$ ενώ το e είναι το σφάλμα που αφορά την απόκλιση από τη γραμμικότητα. Ο συντελεστής b_i (όπου $i = 1, 2, ..., k$)[9] εκτιμά τη μεταβολή της εξαρτημένης μεταβλητής Y για μια μονάδα αύξησης της ανεξάρτητης μεταβλητής X_i όταν οι τιμές των άλλων ανεξάρτητων μεταβλητών παραμένουν σταθερές ή, ισοδύναμα, όταν οι γραμμικές επιδράσεις των υπόλοιπων μεταβλητών στη X_i έχουν απομακρυνθεί από το μοντέλο (Γναρδέλλης, 2003). Στην περίπτωση κατά την οποία μια ανεξάρτητη μεταβλητή X_i του μοντέλου είναι *δίτιμη*[10], εκφρασμένη αριθμητικά με τις τιμές 0 και 1 (π.χ., 0 = γυναίκες και 1 = άνδρες), η ερμηνεία του συντελεστή της παραμένει ανάλογη με αυτή των ποσοτικών μεταβλητών. Δηλαδή, η ποσότητα b_i που αφορά τη δίτιμη μεταβλητή

---

[7] Proportional hazards models.
[8] Logistic regression.
[9] Regression coefficient
[10] Binary.

X_i εξακολουθεί να ορίζει τη μεταβολή της Y για μια μονάδα αύξησης της X_i (δηλαδή πόσο διαφέρει η τιμή της Y στις δύο κατηγορίες της X_i) όταν οι τιμές των άλλων ανεξάρτητων μεταβλητών παραμένουν σταθερές.

Το μοντέλο της πολλαπλής γραμμικής παλινδρόμησης μπορεί αρχικά να αξιολογηθεί ως προς την προσαρμογή του στα δειγματικά δεδομένα, συνήθως με τη χρήση του *συντελεστή πολλαπλού προσδιορισμού*[11] (R^2), ο οποίος εκφράζει το ποσοστό της μεταβλητότητας της Y που ερμηνεύεται από το μοντέλο. Στη συνέχεια, όμως, μπορεί να χρησιμοποιηθεί και επαγωγικά, ελέγχοντας αν οι σχέσεις που προσδιορίστηκαν περιγραφικά (μέσω της δειγματικής εξίσωσης) ισχύουν και σε πληθυσμιακό επίπεδο. Η επαγωγική αξιολόγηση του μοντέλου της πολλαπλής γραμμικής παλινδρόμησης αφορά (α) τον έλεγχο της γραμμικής σχέσης μεταξύ της εξαρτημένης μεταβλητής και των ανεξαρτήτων (μέσω της αξιολόγησης του R^2), (β) τον έλεγχο της επίδρασης κάθε ανεξάρτητης μεταβλητής στην εξαρτημένη, αφού απομακρυνθούν οι επιδράσεις των υπόλοιπων μεταβλητών, και (γ) την κατασκευή *διαστημάτων εμπιστοσύνης*[12] (ΔΕ) για τους αντίστοιχους πληθυσμιακούς συντελεστές της παλινδρόμησης.

Για παράδειγμα, χρησιμοποιώντας ένα μοντέλο πολλαπλής γραμμικής παλινδρόμησης σε μια έρευνα καθημερινών συνηθειών (Γναρδέλλης, Λάγιου, Χλόπτσιος, Μπενέτου, & Τριχοπούλου, 1999), εκτιμήθηκε ο χρόνος βραδινού ύπνου ενήλικων Ελλήνων αναφορικά με το φύλο, την ηλικία, τις ώρες ημερήσιας εργασίας και την περιοχή διαμονής τους. Το μοντέλο που προέκυψε εκφράσθηκε με μια εξίσωση της μορφής:

$$\text{Ώρες ύπνου} = 9{,}15 + 0{,}682(\text{φύλο}) + 0{,}004(\text{ηλικία}) - 0{,}224(\text{εργασία}) - 0{,}111(\text{περιοχή})$$

Στην παραπάνω εξίσωση το φύλο και η περιοχή είναι δίτιμες μεταβλητές κωδικοποιημένες αριθμητικά με 0 και 1 (0 = γυναίκες και 1 = άνδρες για το φύλο, 0 = λοιπή χώρα και 1 = Λεκανοπέδιο Αττικής για την περιοχή), ενώ η ηλικία και η εργασία είναι εκφρασμένες σε έτη και ώρες καθημερινής εργασίας αντιστοίχως. Τα αποτελέσματα της συγκεκριμένης ανάλυσης, έτσι όπως παρήχθησαν από το SPSS, εμφανίζονται στους Πίνακες 1 έως 3.

---

[11] Coefficient of multiple determination (R^2) or squared multiple correlation.

[12] Confidence intervals.

Πίνακας 1. Συντελεστής πολλαπλού προσδιορισμού της γραμμικής παλινδρόμησης

Model Summary				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,424[a]	,179	,177	1,76325

Σημείωση: [a] Predictors: (Constant), Ημερήσιος χρόνος εργασίας, Περιοχή, Φύλο, Ηλικία

Πίνακας 2. Έλεγχος γραμμικότητας του μοντέλου της πολλαπλής γραμμικής παλινδρόμησης

ANOVA[b]						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	962,400	4	240,600	77,387	,000[a]
	Residual	4399,324	1415	3,109		
	Total	5361,724	1419			

Σημείωση: [a] Predictors: (Constant), Ημερήσιος χρόνος εργασίας, Περιοχή, Φύλο, Ηλικία.
[b] Dependent Variable: Ώρες ύπνου

Πίνακας 3. Συντελεστές του μοντέλου της πολλαπλής γραμμικής παλινδρόμησης

Coefficients[a]						
		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
Model		B	Std. Error	Beta	B	Std. Error
1	(Constant)	9,150	,284		32,217	,000
	Φύλο	,682	,105	,173	6,507	,000
	Ηλικία	,004	,005	,022	,784	,433
	Ημερήσιος χρόνος εργασίας	-,224	,016	-,434	-14,362	,000
	Περιοχή	-,111	,099	-,028	-1,131	,258

Σημείωση: [a] Dependent Variable: Ώρες ύπνου.

Στον Πίνακα 1 δίνεται το R^2 του μοντέλου (R Square = 0,179). Από την τιμή του προκύπτει ότι περίπου 18% της μεταβλητότητας του χρόνου βραδινού ύπνου ερμηνεύεται από το φύλο, την ηλικία, την περιοχή διαμονής και τον ημερήσιο χρόνο εργασίας των ατόμων. Η τιμή αυτή διορθώνεται ως προς την αμεροληψία της για την εκτίμηση του πληθυσμιακού συντελεστή πολλαπλού προσδιορισμού και εμφανίζεται στη διπλανή στήλη του πίνακα (Adjusted R Square = 0,177).[13]

Στο Πίνακα 2 των αποτελεσμάτων γίνεται ο έλεγχος της γραμμικότητας του μοντέλου (μέσω της αξιολόγησης του R^2). Η αξιολόγηση γίνεται με τη βοήθεια της ανάλυσης διακύμανσης. Αρχικά δίνεται το *άθροισμα τετραγώνων*[14] της παλινδρόμησης και των σφαλμάτων, ενώ στη συνέχεια δίνονται οι βαθμοί ελευθερίας και τα αντίστοιχα μέσα τετράγωνα. Η τιμή του κριτηρίου F, το οποίο χρησιμοποιείται για τον έλεγχο της *μηδενικής υπόθεσης*[15], είναι ο λόγος του μέσου τετραγώνου της παλινδρόμησης προς το μέσο τετράγωνο των σφαλμάτων. Η πιθανότητα του ελέγχου, δηλαδή η πιθανότητα να ισχύει η μηδενική υπόθεση είναι πολύ μικρή (Sig. < 0,0005), γεγονός που μας επιτρέπει να ισχυριστούμε ότι το φύλο, η ηλικία, η περιοχή διαμονής και ο ημερήσιος χρόνος εργασίας συνδυασμένα γραμμικά (με τη μορφή της εξίσωσης της μοντέλου) συμβάλλουν σημαντικά στην ερμηνεία του χρόνου βραδινού ύπνου (Γναρδέλλης, 2006).

Στον Πίνακα 3 της ανάλυσης περιλαμβάνονται οι συντελεστές της παλινδρόμησης και οι έλεγχοι που γίνονται επ' αυτών. Ο συντελεστής του φύλου, υποδηλώνει ότι ο χρόνος βραδινού ύπνου είναι στους άνδρες κατά 0,682 της ώρας μεγαλύτερος από ό,τι στις γυναίκες, και αυτό ανεξαρτήτως ηλικίας, περιοχής διαμονής και ημερήσιου χρόνου εργασίας. Επιπλέον, για κάθε μια ώρα αύξησης του ημερήσιου χρόνου εργασίας ο χρόνος βραδινού ύπνου ελαττώνεται κατά 0,224 της ώρας. Δηλαδή δύο άτομα του ίδιου φύλου, της ίδιας ηλικίας και της ίδιας περιοχής διαμονής, τα οποία διαφέρουν ως προς το χρόνο ημερήσιας εργασίας κατά μία ώρα, διαφέρουν ως προς το χρόνο βραδινού ύπνου, κατά μέσο όρο 0,224 της ώρας. Οι μεταβλητές του φύλου και του ημερήσιου χρόνου εργασίας είναι, με βάση την

---

[13] Adjusted R^2 είναι ο *προσαρμοσμένος συντελεστής πολλαπλού προσδιορισμού*. Αποτελεί αμερόληπτη εκτίμηση του πληθυσμιακού συντελεστή πολλαπλού προσδιορισμού.

[14] Sum of squares.

[15] H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$, όπου β_i $(i = 1, 2, ..., k)$ είναι οι πληθυσμιακοί συντελεστές παλινδρόμησης.

επαγωγική αξιολόγηση των συντελεστών τους, οι δύο σημαντικότερες κατά τον προσδιορισμό του χρόνου βραδινού ύπνου σύμφωνα με το μοντέλο της πολλαπλής γραμμικής παλινδρόμησης που χρησιμοποιήθηκε.

Λογαριθμική παλινδρόμηση

Σε περιπτώσεις διερευνήσεων όπου η εξαρτημένη μεταβλητή Y είναι δίτιμη και υποδηλώνει την πραγματοποίηση ή μη ενός γεγονότος (βλ. την παρακάτω σχέση) το μοντέλο της πολλαπλής γραμμικής παλινδρόμησης δεν είναι επαρκές για την εκτίμηση των τιμών της εξαρτημένης μεταβλητής (η οποία πλέον δεν είναι συνεχής), διότι δεν ισχύουν οι προϋποθέσεις χρήσης του.

$$Y = \begin{pmatrix} \nu\alpha\iota \\ \acute{o}\chi\iota \end{pmatrix} \longleftarrow (X_1, X_2, ..., X_k)$$

Σε μια τέτοια περίπτωση, θα μπορούσε να χρησιμοποιηθεί ως εξαρτημένη μεταβλητή του μοντέλου η πιθανότητα p πραγματοποίησης του γεγονότος. Όπως εκτιμάται η τιμή μιας συνεχούς μεταβλητής Y με τη βοήθεια ενός μοντέλου πολλαπλής γραμμικής παλινδρόμησης, έτσι μπορεί να εκτιμηθεί (με τη χρήση κατάλληλου μοντέλου) και η πιθανότητα p πραγματοποίησης ενός γεγονότος (ή αλλιώς η πιθανότητα "επιτυχίας" μιας δίτιμης μεταβλητής) για ένα σύνολο τιμών μίας ή περισσοτέρων ανεξάρτητων μεταβλητών. Η τεχνική που χρησιμοποιείται σε αυτές τις περιπτώσεις ονομάζεται *λογαριθμική παλινδρόμηση* (Γναρδέλλης, 2003).

Επιχειρώντας να εκτιμήσουμε την πιθανότητα p πραγματοποίησης ενός γεγονότος με ένα μοντέλο που έχει τη μορφή $p = b_0 + b_1X_1 + b_2X_2 + ... + b_kX_k$, το κύριο πρόβλημα που συναντάμε είναι ότι, αν και οι τιμές της p θεωρητικά δεν μπορούν να βρίσκονται εκτός του διαστήματος [0, 1], οι τιμές της ποσότητας $(b_0 + b_1X_1 + b_2X_2 + ... + b_xX_x)$ μπορούν να διακυμαίνονται σε όλο το εύρος των πραγματικών αριθμών. Σε μια προσπάθεια διευθέτησης του προβλήματος θα μπορούσαμε να αντικαταστήσουμε στο προηγούμενο μοντέλο την πιθανότητα p πραγματοποίησης του γεγονότος με τη *σχετική_πιθανότητα*[16] πραγματοποίησης, δη-

[16] Odds.

λαδή, με το λόγο της πιθανότητας πραγματοποίησης προς την πιθανότητα μη πραγματοποίησης, $\frac{p}{1-p}$.

Ο συγκεκριμένος λόγος, αν και θεωρητικά μπορεί να διακυμαίνεται μέχρι το $+\infty$, δεν μπορεί να παίρνει τιμές μικρότερες του 0. Οι τιμές του, δηλαδή, είναι θετικές ή ίσες με το 0. Άρα, και στην περίπτωση αυτή δεν είναι επαρκές για την εκτίμηση της p ένα γραμμικό μοντέλο που έχει τη μορφή:

$$\frac{p}{1-p} = b_0 + b_1X_1 + b_2X_2 + ... + b_kX_k$$

Ένας επιπλέον μετασχηματισμός της σχετικής πιθανότητας επιλύει το πρόβλημα. Αν αντί του λόγου $\frac{p}{1-p}$ χρησιμοποιηθεί ο φυσικός του λογάριθμος, $\ln\left[\frac{p}{1-p}\right]$, τότε οι τιμές του μετασχηματισμένου λόγου, οι οποίες διακυμαίνονται πλέον στο διάστημα $(-\infty, +\infty)$, μπορούν να εκτιμηθούν με τη βοήθεια ενός λογαριθμικού μοντέλου που έχει τη μορφή:

$$\ln\left[\frac{p}{1-p}\right] = b_0 + b_1X_1 + b_2X_2 + ... + b_kX_k$$

Η *συνάρτηση* $\ln\left[\frac{p}{1-p}\right]$, *η οποία συνδέει*[17] την πιθανότητα πραγματοποίησης του γεγονότος με τις ανεξάρτητες μεταβλητές $X_1, X_2, ..., X_k$, στην ορολογία των λογαριθμικών γραμμικών μοντέλων συμβολίζεται ως $\text{logit}(p)$, δηλαδή:

$$\text{logit}(p) = \ln\left[\frac{p}{1-p}\right] = b_0 + b_1X_1 + b_2X_2 + ... + b_kX_k$$

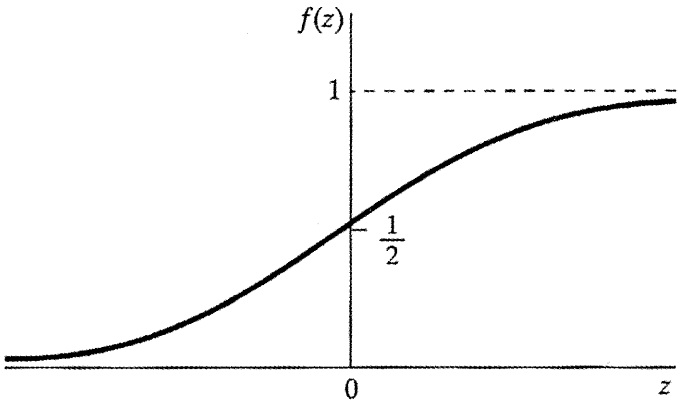
[17] Linking function.

Αντιλογαριθμίζοντας τα δύο μέλη της προηγούμενης εξίσωσης προκύπτει $\frac{p}{1-p} = e^{b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_k X_k}$, ενώ θέτοντας $Z = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_k X_k$ προκύπτει $\frac{p}{1-p} = e^Z$. Επιλύοντας την τελευταία εξίσωση ως προς p, παίρνουμε

$$p = e^Z - p e^Z \quad \text{ή} \quad p = \frac{e^Z}{1+e^Z}$$

Η τελευταία εξίσωση, διατυπωμένη επίσης με τη μορφή $p = \frac{1}{1+e^{-Z}}$, αποτελεί την εκτίμηση της πιθανότητας p πραγματοποίησης του γεγονότος.

Η διαγραμματική απεικόνιση της συνάρτησης $f(z) = \frac{1}{1+e^{-z}}$, η οποία εκτιμά την p, είναι σιγμοειδής (Σχήμα 1), ενώ οι τιμές της κυμαίνονται στο διάστημα [0, 1] εφόσον οι τιμές της Z μεταβάλλονται στο διάστημα (- ∞, +∞). Η συναρτησιακή έκφρασή της είναι, επομένως, κατάλληλη να χρησιμοποιηθεί ως μοντέλο για την εκτίμηση μιας πιθανότητας. Από τη διαγραμματική απεικόνισή της προκύπτει ότι η σχέση των ανεξάρτητων μεταβλητών $X_1, X_2, \ldots, X_k$ και της πιθανότητας πραγματοποίησης του γεγονότος είναι μη γραμμική.



Σχήμα 1. Διαγραμματική απεικόνιση της συνάρτησης $f(z) = \frac{1}{1+e^{-z}}$.

Είναι εμφανές από όλα τα προηγούμενα ότι από τις τρεις ποσότητες που επιχειρήθηκε να εκτιμηθούν με ένα γραμμικό μοντέλο, δηλαδή (α) την πιθανότητα του γεγονότος της επιτυχίας p, (β) τη σχετική πιθανότητα του γεγονότος της επιτυχίας $\frac{p}{1-p}$ και (γ) το λογάριθμο της σχετικής πιθανότητας, $\ln\left[\frac{p}{1-p}\right]$, αυτή η οποία μας "διευκολύνει" περισσότερο όσον αφορά το γραμμικό προσδιορισμό της, είναι ο λογάριθμος της σχετικής πιθανότητας. Η ποσότητα αυτή, όπως ήδη έχουμε δει, μπορεί να προσδιοριστεί με τη βοήθεια ενός γραμμικού μοντέλου, εφόσον οι δυνατές τιμές της κυμαίνονται στο διάστημα (-∞, +∞). Ακολουθώντας τη λογική της γραμμικής παλινδρόμησης, ο λογαριθμικός μετασχηματισμός της σχετικής πιθανότητας θα μπορούσε να ιδωθεί ως μια απαραίτητη μετατροπή της εξαρτημένης μεταβλητής σε ένα μοντέλο πολλαπλής γραμμικής παλινδρόμησης, προκειμένου η μεταβλητή αυτή να προσαρμοστεί στις απαιτήσεις του συγκεκριμένου μοντέλου. Μόνο που στην περίπτωση της λογαριθμικής παλινδρόμησης, η εξαρτημένη μεταβλητή που μετασχηματίζεται λογαριθμικά δεν είναι μια οποιαδήποτε ποσότητα, αλλά η σχετική πιθανότητα πραγματοποίησης ενός γεγονότος, δηλαδή της "επιτυχίας" μιας δίτιμης μεταβλητής (Everitt & Dunn, 1992).

Η χρήση της μεθόδου των ελάχιστων τετραγώνων, με την οποία εκτιμώνται οι συντελεστές του μοντέλου της πολλαπλής γραμμικής παλινδρόμησης, δεν μπορεί να χρησιμοποιηθεί και για την εκτίμηση των συντελεστών του λογαριθμικού μοντέλου. Για το μοντέλο της λογαριθμικής παλινδρόμησης, αντί της συνήθους μεθόδου των ελάχιστων τετραγώνων, χρησιμοποιείται η μέθοδος των *εκτιμήσεων μέγιστης πιθανοφάνειας*[18].

Ερμηνεία των συντελεστών του μοντέλου της λογαριθμικής παλινδρόμησης

Η ερμηνεία των συντελεστών του μοντέλου της λογαριθμικής παλινδρόμησης μπορεί να γίνει σε εναρμόνιση με την ερμηνεία των συντελεστών του γραμμικού μοντέλου. Έτσι από την εξίσωση της λογαριθμικής παλινδρόμησης

$$\ln\left[\frac{p}{1-p}\right] = b_0 + b_1X_1 + b_2X_2 + \ldots + b_kX_k$$

[18] Maximum likelihood estimations.

προκύπτει ότι κάθε ένας από τους συντελεστές b_i (όπου $i = 1, 2, ..., k$) εκφράζει τη μεταβολή της εξαρτημένης μεταβλητής $\ln\left[\frac{p}{1-p}\right]$ για μία μονάδα αύξησης της αντίστοιχης ανεξάρτητης μεταβλητής X_i, εφόσον οι τιμές των υπόλοιπων ανεξάρτητων μεταβλητών παραμένουν σταθερές. Επειδή, όμως, είναι περισσότερο κατανοητό και ενδιαφέρον από απόψεως ερμηνείας των δεδομένων να αναφερόμαστε στη σχετική πιθανότητα πραγματοποίησης ενός γεγονότος και όχι στο λογάριθμο της σχετικής πιθανότητας, είναι προτιμότερο η παραπάνω εξίσωση να γραφεί με αντιλογαρίθμιση των δύο μερών της ως εξής:

$$\frac{p}{1-p} = e^{b_0 + b_1 X_1 + b_2 X_2 + ... + b_k X_k} = e^{b_0} \cdot e^{b_1 X_1} \cdot e^{b_2 X_2} \cdots e^{b_k X_k}$$

Από την τελευταία εξίσωση προκύπτει ότι η ποσότητα e^{b_i} (δηλαδή, ο αντιλογάριθμος του b_i) είναι ο παράγοντας επί τον οποίο πολλαπλασιάζεται η σχετική πιθανότητα πραγματοποίησης του γεγονότος, όταν η ανεξάρτητη μεταβλητή X_i αυξηθεί κατά μία μονάδα (και εφόσον, βέβαια, οι υπόλοιπες μεταβλητές παραμένουν σταθερές). Ειδικότερα:

(α) Αν ο συντελεστής b_i *είναι θετικός*, ο παράγοντας e^{b_i} *είναι μεγαλύτερος από το 1*, γεγονός που σημαίνει ότι η σχετική πιθανότητα, δηλαδή η εξαρτημένη μεταβλητή $\frac{p}{1-p}$, αυξάνει. Για παράδειγμα, e^{b_i}=1,30 σημαίνει αύξηση της σχετικής πιθανότητας πραγματοποίησης ενός γεγονότος κατά 130%-100%=30%, για κάθε μονάδα αύξησης της ανεξάρτητης μεταβλητής X_i, ενώ e^{b_i}=2, σημαίνει αύξηση της σχετικής πιθανότητας κατά 100% ή, αλλιώς, διπλασιασμό της.

(β) Αν ο συντελεστής b_i *είναι αρνητικός*, ο παράγοντας e^{b_i} *είναι μικρότερος της μονάδας*, δηλαδή η σχετική πιθανότητα μειώνεται. Για παράδειγμα, e^{b_i}=0,85, σημαίνει ελάττωση της σχετικής πιθανότητας πραγματοποίησης ενός γεγονότος κατά 100%-85% = 15% για κάθε μονάδα αύξησης της ανεξάρτητης μεταβλητής X_i.

(γ) Τέλος, όταν ο συντελεστής b_i *είναι μηδέν*, ο παράγοντας e^{b_i} *γίνεται ίσος με τη μονάδα* και η σχετική πιθανότητα παραμένει αμετάβλητη. Αν, δηλαδή e^{b_i}=1, τότε η σχετική πιθανότητα παραμένει αμετάβλητη για κάθε μονάδα αύξησης της X_i. Άρα, η μεταβλητή X_i δεν επηρεάζει το ενδεχόμενο πραγματοποίησης του γεγονότος.

Χρήση ανεξάρτητων κατηγορικών μεταβλητών

Στην περίπτωση που μια ανεξάρτητη μεταβλητή του μοντέλου είναι δίτιμη (π.χ., το φύλο), εκφρασμένη αριθμητικά με τις τιμές 0 και 1 (π.χ., 0 = γυναίκες και 1 = άνδρες), η ερμηνεία των συντελεστών της λογαριθμικής παλινδρόμησης παραμένει η ίδια. Δηλαδή, η ποσότητα b_i που αφορά τη δίτιμη μεταβλητή (το φύλο) εξακολουθεί να ορίζει τον παράγοντα επί τον οποίο πολλαπλασιάζεται η σχετική πιθανότητα πραγματοποίησης του γεγονότος στους άνδρες αναφορικά με τη σχετική πιθανότητα πραγματοποίησης του γεγονότος στις γυναίκες (πόσο μεγαλύτερη ή μικρότερη δηλαδή είναι η πιθανότητα πραγματοποίησης του γεγονότος στους άνδρες σε σχέση με τις γυναίκες). Με πιο αναλυτική διατύπωση, συμβολίζοντας με $\frac{p_A}{1-p_A}$ τη σχετική πιθανότητα πραγματοποίησης του γεγονότος στους άνδρες και $\frac{p_\Gamma}{1-p_\Gamma}$ τη σχετική πιθανότητα πραγματοποίησης του γεγονότος στις γυναίκες, το λογαριθμικό μοντέλο –στην απλή μορφή του– γράφεται:

$$\ln\left[\frac{p}{1-p}\right] = b_0 + b_1 X = b_0 + b_1(\varphi\acute{\upsilon}\lambda o)$$

ή

$$\frac{p}{1-p} = e^{b_0 + b_1 X} = e^{b_0 + b_1(\varphi\acute{\upsilon}\lambda o)}$$

Εφόσον το φύλο εκφράζεται αριθμητικά με τις τιμές 0 (γυναίκες) και 1 (άνδρες), η προηγούμενη διατύπωση του λογαριθμικού μοντέλου εξειδικεύεται σε

$$\frac{p_A}{1-p_A} = e^{b_0 + b_1 \cdot 1} = e^{b_0 + b_1} \qquad \text{για τους άνδρες}$$

και

$$\frac{p_\Gamma}{1-p_\Gamma} = e^{b_0 + b_1 \cdot 0} = e^{b_0} \qquad \text{για τις γυναίκες.}$$

Διαιρώντας τις δύο εκφράσεις της σχετικής πιθανότητας, προκύπτει:

$$\frac{p_A/(1-p_A)}{p_\Gamma/(1-p_\Gamma)} = \frac{e^{b_0 + b_1}}{e^{b_0}} = e^{b_1}$$

Το πρώτο μέρος της τελευταίας εξίσωσης είναι ο *λόγος των σχετικών πιθανοτήτων*[19] πραγματοποίησης του γεγονότος για τις δύο κατηγορίες του φύλου (πόσο μεγαλύτερη ή μικρότερη δηλαδή είναι η πιθανότητα πραγματοποίησης του γεγονότος στους άνδρες σε σχέση με τις γυναίκες).

Όταν χρησιμοποιείται ως ανεξάρτητη μεταβλητή ενός μοντέλου μια κατηγορική μεταβλητή με περισσότερες των δύο κατηγοριών, θα πρέπει να κωδικοποιηθεί με τρόπο αντίστοιχο των δίτιμων μεταβλητών. Αν δηλαδή η μεταβλητή περιλαμβάνει k κατηγορίες, πρέπει να δημιουργηθούν k-1 νέες *ψευδομεταβλητές*[20] με τιμές 0 και 1. Για το σκοπό αυτό, μία από τις κατηγορίες της μεταβλητής επιλέγεται ως κατηγορία αναφοράς, ενώ δημιουργούνται k-1 ψευδομεταβλητές που αντιπροσωπεύουν τις υπόλοιπες κατηγορίες. Η τιμή κάθε μιας από τις οριζόμενες ψευδομεταβλητές είναι 1, αν μια παρατήρηση ανήκει στην αντίστοιχη κατηγορία, και 0, αν δεν ανήκει. Αν, για παράδειγμα, χρησιμοποιηθεί μια μεταβλητή που ορίζει το επίπεδο εκπαίδευσης, με κατηγορίες *αναλφάβητος/η*, *απόφοιτος/η στοιχειώδους εκπαίδευσης*, *απόφοιτος/η μέσης εκπαίδευσης* και *απόφοιτος/η τριτοβάθμιας εκπαίδευσης*, πρέπει αρχικά να οριστεί μία κατηγορία αναφοράς (π.χ., η κατηγορία αναλφάβητος/η). Στη συνέχεια να δημιουργηθούν τρεις ψευδομεταβλητές που αντιπροσωπεύουν αντιστοίχως τους απόφοιτους στοιχειώδους εκπαίδευσης (X_1), μέσης εκπαίδευσης (X_2) και τριτοβάθμιας εκπαίδευσης (X_3). Κάθε μία από αυτές τις ψευδομεταβλητές παίρνει την τιμή 1, αν κάποιος είναι του ίδιου επιπέδου εκπαίδευσης με αυτήν που αντιπροσωπεύει η ψευδομεταβλητή, και 0, αν δεν είναι. Έτσι, για έναν απόφοιτο στοιχειώδους εκπαίδευσης οι δημιουργούμενες ψευδομεταβλητές θα πάρουν τις τιμές $X_1 = 1, X_2 = 0, X_3 = 0$, για έναν απόφοιτο μέσης εκπαίδευσης, τις τιμές $X_1 = 0, X_2 = 1, X_3 = 0$, και για έναν απόφοιτο τριτοβάθμιας εκπαίδευσης, τις τιμές $X_1 = 0, X_2 = 0, X_3 = 1$.

Εφόσον οι k-1 ψευδομεταβλητές εισαχθούν σε ένα μοντέλο λογαριθμικής παλινδρόμησης, οι συντελεστές που θα προκύψουν εκφράζουν πόσο μεγαλύτερη ή μικρότερη είναι η πιθανότητα πραγματοποίησης του γεγονότος στην κατηγορία που αντιπροσωπεύει κάθε ψευδομεταβλητή σε σχέση με την κατηγορία αναφοράς. Αν στο παράδειγμα της εκπαίδευσης η ποσότητα e^{b_k} για την κατηγορία *απόφοιτος/η τριτοβάθμιας εκπαίδευσης* είναι π.χ., $e^{b_k} = 0{,}76$ αυτό πρακτικά σημαίνει ότι η πιθανότητα πραγματοποίησης

[19] Odds ratio.

[20] Dummy variables.

του γεγονότος που διερευνάμε για τους απόφοιτους τριτοβάθμιας εκπαίδευσης είναι κατά 100% - 76% = 24% μικρότερη από αυτήν των αναλφάβητων (κατηγορία αναφοράς).

Επαγωγικοί έλεγχοι για τους συντελεστές του λογαριθμικού μοντέλου

Το μοντέλο της λογαριθμικής παλινδρόμησης μπορεί αρχικά να χρησιμοποιηθεί σε περιγραφικό επίπεδο, συνοψίζοντας τη σχέση της εξαρτημένης μεταβλητής με τις ανεξάρτητες υπό τη μορφή της εξίσωσης:

$$\ln\left[\frac{p}{1-p}\right] = b_0 + b_1X_1 + b_2X_2 + ... + b_kX_k$$

Στη συνέχεια, όμως, πρέπει να αξιολογηθεί και επαγωγικά ελέγχοντας αν οι σχέσεις που προσδιορίστηκαν περιγραφικά (μέσω της δειγματικής εξίσωσης) ισχύουν και σε πληθυσμιακό επίπεδο. Η επαγωγική αξιολόγηση του λογαριθμικού μοντέλου αφορά: (α) Τον έλεγχο των συντελεστών του, κατά πόσο δηλαδή, οι συντελεστές που προσδιορίστηκαν κατά την εκτίμηση του μοντέλου αφορούν όχι μόνο τα δειγματικά δεδομένα για τα οποία έγινε η εκτίμηση, αλλά και τον πληθυσμό από τον οποίο προήρθε το δείγμα. (β) Την κατασκευή διαστημάτων εμπιστοσύνης (ΔΕ) για τους συντελεστές. (γ) Τον έλεγχο της προσαρμογής του μοντέλου στα δειγματικά δεδομένα.

Έλεγχος των συντελεστών του μοντέλου. Οι συντελεστές του δειγματικού μοντέλου της λογαριθμικής παλινδρόμησης αποτελούν σημειακές εκτιμήσεις των συντελεστών του αντίστοιχου πληθυσμιακού μοντέλου, όπως φαίνεται στην παρακάτω σχέση:

$$\ln\left[\frac{p}{1-p}\right] = b_0 + b_1X_1 + b_2X_2 + ... + b_kX_k \quad \textit{δειγματικό μοντέλο}$$

$$\downarrow \quad \downarrow \quad \downarrow \quad \downarrow$$

$$\ln\left[\frac{p}{1-p}\right] = \beta_0 + \beta_1X_1 + \beta_2X_2 + ... + \beta_kX_k \quad \textit{πληθυσμιακό μοντέλο}$$

Ο προσδιορισμός τους γίνεται με τη βοήθεια της μεθόδου των εκτιμήσεων μέγιστης πιθανοφάνειας, μιας διαδικασίας η οποία εκτιμά τις πλη-

θυσμιακές παραμέτρους του μοντέλου με κριτήριο τη μεγιστοποίηση της πιθανότητας τα διαθέσιμα δειγματικά δεδομένα να έχουν παραχθεί από τις εκτιμώμενες παραμέτρους (Aggresti, 1990. Hosmer & Lemeshow, 1989).

Εκτός των συντελεστών, $b_0, b_1, b_2, \ldots, b_k$ του δειγματικού μοντέλου η μέθοδος των εκτιμήσεων μέγιστης πιθανοφάνειας εκτιμά και τα αντίστοιχα τυπικά σφάλματα αυτών. Αποδεικνύεται ότι ένας οποιοσδήποτε δειγματικός συντελεστής b_i του λογαριθμικού μοντέλου, εφόσον το μέγεθος του δείγματος είναι επαρκώς μεγάλο, ακολουθεί κατά προσέγγιση την κανονική κατανομή με μέση τιμή τον αντίστοιχο πληθυσμιακό συντελεστή β_i και εκτιμώμενο (με τη βοήθεια της συνάρτησης μέγιστης πιθανοφάνειας) τυπικό σφάλμα $se(b_i)$. Επομένως, όταν το μέγεθος του δείγματος είναι επαρκώς μεγάλο, το πηλίκο $\frac{b_i - \beta_i}{se(b_i)}$ ακολουθεί κατά προσέγγιση την τυπική κανονική κατανομή. Ο έλεγχος δηλαδή της μηδενικής υπόθεσης $H_0 : \beta_i = 0$ (ή ισοδύναμα) $e^{\beta_i} = 1$ έναντι της εναλλακτικής $H_A : \beta_i \neq 0$ (ή ισοδύναμα $e^{\beta_i} \neq 1$) μπορεί να γίνει με τη βοήθεια του κριτηρίου $z = \frac{b_i}{se(b_i)}$, το οποίο (εφόσον ισχύει η $H_0 : \beta_i = 0$) ακολουθεί την τυπική κανονική κατανομή. Η ποσότητα $\frac{b_i}{se(b_i)}$ με βάση την οποία γίνεται ο έλεγχος ονομάζεται *κριτήριο του Wald* (Hauck & Donner, 1977. Rao, 1973). Ένας ισοδύναμος έλεγχος γίνεται με τη βοήθεια της ποσότητας $\left[\frac{b_i}{se(b_i)}\right]^2$, η οποία, όταν το μέγεθος του δείγματος είναι μεγάλο, και, εφόσον ισχύει η μηδενική υπόθεση $H_0 : \beta_i = 0$, ακολουθεί την κατανομή χ^2 με 1 βαθμό ελευθερίας.

Διαστήματα εμπιστοσύνης για τους συντελεστές του λογαριθμικού μοντέλου. Εκτός των επαγωγικών ελέγχων που μπορούν να γίνουν επί των συντελεστών του λογαριθμικού μοντέλου, μπορούν να κατασκευαστούν και τα αντίστοιχα διαστήματα εμπιστοσύνης. Έτσι το 95%ΔΕ για τον πληθυσμιακό συντελεστή β_i είναι $(b_i - 1{,}96se(b_i),\ b_i + 1{,}96se(b_i))$ εφόσον, όπως αναφέραμε, η κατανομή του δειγματικού συντελεστή b_i είναι κατά προσέγγιση κανονική.

Αντίστοιχα διαστήματα εμπιστοσύνης μπορούν να κατασκευαστούν και για τις ποσότητες e^{β_i} (όπου $i = 1, 2, \ldots, k$) με αντιλογαρίθμηση των προ-

ηγούμενων ορίων. Δηλαδή το 95%ΔΕ για τον αντιλογάριθμο του β_i είναι $\left(e^{[b_i - 1{,}96se(b_i)]}, e^{[b_i + 1{,}96se(b_i)]}\right)$.

Αξιολόγηση της προσαρμογής του λογαριθμικού μοντέλου. Όπως στην πολλαπλή γραμμική παλινδρόμηση, έτσι και στη λογαριθμική, μετά την εκτίμηση των συντελεστών του λογαριθμικού μοντέλου εναπομένει η αξιολόγηση της προσαρμογής του στα δειγματικά δεδομένα. Ένα μέτρο της καλής προσαρμογής του λογαριθμικού μοντέλου είναι η μέγιστη τιμή της συνάρτησης πιθανοφάνειας. Όσο μεγαλύτερη είναι η τιμή της συνάρτησης πιθανοφάνειας L ή, αντιστοίχως, όσο μικρότερη είναι η τιμή της συνάρτησης λογαριθμο-πιθανοφάνειας $-2\ln L$, τόσο καλύτερη είναι και η προσαρμογή του μοντέλου στα δειγματικά δεδομένα. Προκειμένου οι δύο αυτές ποσότητες να χρησιμοποιηθούν για την αξιολόγηση ενός μοντέλου, πρέπει να συγκρίνονται με τις αντίστοιχες ποσότητες ενός άλλου *απλούστερου μοντέλου*[21] που εκτιμάται για τα ίδια δειγματικά δεδομένα. Η σύγκριση γίνεται με τον υπολογισμό του λόγου των μέγιστων τιμών της *συνάρτησης πιθανοφάνειας*[22].

Αν, για παράδειγμα, πρόκειται να αξιολογηθεί η προσαρμογή του μοντέλου $\text{logit}(p) = b_0 + b_1X_1 + b_2X_2$ (με μέγιστη τιμή πιθανοφάνειας L_2) σε σχέση με το απλό μοντέλο $\text{logit}(p) = b_0 + b_1X_1$ (με μέγιστη τιμή πιθανοφάνειας L_1) αυτό μπορεί να γίνει με τη βοήθεια του λόγου:

$$-2\ln\left(\frac{L_1}{L_2}\right) = -2\ln L_1 - \left(-2\ln L_2\right)$$

Υπό τη μηδενική υπόθεση, $H_0 : e^{\beta_2} = 1$ (γεγονός που σημαίνει ότι η μεταβλητή X_2 δεν επηρεάζει την πιθανότητα πραγματοποίησης του γεγονότος), ο λόγος $-2\ln\left(\frac{L_1}{L_2}\right)$ ακολουθεί την κατανομή χ^2 με 1 βαθμό ελευθερίας.

Γενικότερα, οι βαθμοί ελευθερίας του λόγου των τιμών της συνάρτησης πιθανοφάνειας είναι ίσοι με τη διαφορά του αριθμού των ανεξάρτητων μεταβλητών που περιλαμβάνονται στα δύο συγκρινόμενα κάθε φορά μοντέλα. Όταν αξιολογείται συνολικά η προσαρμογή του μοντέλου

[21] Baseline model.
[22] Likelihood ratio statistic.

$\text{logit}(p) = b_0 + b_1X_1 + b_2X_2 + ... + b_kX_k$ αυτό συγκρίνεται με το μοντέλο που περιλαμβάνει μόνο το σταθερό όρο, δηλαδή με το μοντέλο $\text{logit}(p) = b_0$.

Θεωρώντας ότι για το πληθυσμιακό μοντέλο ισχύει η μηδενική υπόθεση $H_0 : e^{\beta_1} = e^{\beta_2} = ... = e^{\beta_k} = 1$, η ποσότητα $-2\ln\left(\frac{L_0}{L_F}\right)$ (όπου L_F είναι η τιμή της συνάρτησης πιθανοφάνειας για το δειγματικό μοντέλο και L_0 η τιμή για το μοντέλο που περιλαμβάνει μόνο τον σταθερό όρο) ακολουθεί την κατανομή χ^2 με k βαθμούς ελευθερίας. Η απόρριψη της μηδενικής υπόθεσης H_0 σε αυτή την περίπτωση οδηγεί στην εναλλακτική υπόθεση H_A, δηλαδή ότι τουλάχιστον μία από τις ποσότητες e^{β_i} (όπου $i = 1,2,...k$) είναι διάφορη του 1.

Για την αξιολόγηση της προσαρμογής του λογαριθμικού μοντέλου, εκτός από το λόγο των μέγιστων τιμών της συνάρτησης πιθανοφάνειας, χρησιμοποιείται ένα επιπλέον μέτρο καλής προσαρμογής, αντίστοιχο με το συντελεστή προσδιορισμού R^2 της γραμμικής παλινδρόμησης. Το μέτρο αυτό ονομάζεται R^2 *των Cox και Snell* (1989) και ισούται με:

$R^2 = 1 - \left[\frac{L_0}{L_F}\right]^{2/n}$, όπου n το μέγεθος του δείγματος.

Το πρόβλημα με το συγκεκριμένο συντελεστή προσδιορισμού είναι ότι ποτέ δεν καταλήγει να πάρει μέγιστη τιμή το 1. Ο Nagelkerke (1991) πρότεινε μια τροποποίηση του συντελεστή R^2 *των Cox και Snell*, για να παρακάμψει το συγκεκριμένο πρόβλημα. Ο συντελεστής που πρότεινε ο Nagelkerke είναι ο

$\tilde{R}^2 = \frac{R^2}{R^2_{\max}} \in (0,1)$, όπου $R^2_{\max} = 1 - [L_0]^{2/n}$.

Στο παράδειγμα που ακολουθεί παρουσιάζεται η χρήση και η ερμηνεία των συντελεστών ενός μοντέλου λογαριθμικής παλινδρόμησης σε δεδομένα από έρευνα στις ελληνικές φυλακές ανδρών.

Χρήση ενδοφλέβιων ναρκωτικών στις ελληνικές φυλακές ανδρών

Σε έρευνα που έγινε στις ελληνικές φυλακές ανδρών (Koulierakis, Gnardellis, Agrafiotis, & Power, 2000), και η οποία αφορούσε τις γνώσεις και στάσεις των κρατουμένων στο θέμα του AIDS, ετέθη, με τη χρήση ανώνυμου ερωτηματολογίου, το ερώτημα της χρήσης ενδοφλέβιων ναρκωτι-

κών μέσα στις φυλακές. Η πιθανότητα χρήσης ενδοφλέβιων ναρκωτικών εκτιμήθηκε με τη βοήθεια ενός μοντέλου λογαριθμικής παλινδρόμησης, με εξαρτημένη μεταβλητή τη χρήση (ναι, όχι) και ανεξάρτητες μεταβλητές την ηλικία των κρατουμένων (σε χρόνια), τον ποινικό χαρακτηρισμό (υπόδικοι, κατάδικοι), το συνολικό χρόνο εγκλεισμού (σε χρόνια) και το ενδεχόμενο προηγούμενης καταδίκης για ναρκωτικά (ναι, όχι). Στο αρχείο των δεδομένων η μεταβλητή *prisinj* αφορά τη χρήση (0 = όχι, 1 = ναι), οι μεταβλητές *age* και *total* την ηλικία και το συνολικό χρόνο εγκλεισμού των κρατουμένων αντιστοίχως, ενώ οι μεταβλητές *inmatype* (0 = υπόδικοι, 1 = κατάδικοι) και *drugconv* (0 = όχι, 1 = ναι) τον ποινικό χαρακτηρισμό των κρατουμένων και το ενδεχόμενο καταδίκης για ναρκωτικά στο παρελθόν, αντιστοίχως. Το μοντέλο που προέκυψε από τα δειγματικά δεδομένα για τη χρήση ή μη ενδοφλέβιων ναρκωτικών εντός της φυλακής έχει τη μορφή:

$$\ln\left[\frac{p}{1-p}\right] = -1{,}972 - 0{,}14\,(age) + 0{,}416(inmatype) + 0{,}656(drugconv) + 0{,}45\,(total)$$

Στην παραπάνω εξίσωση p είναι η εκτιμώμενη πιθανότητα χρήσης εντός της φυλακής, *inmatype* είναι η μεταβλητή που υποδηλώνει τους κατάδικους (έναντι των υποδίκων) και *drugconv* η μεταβλητή που υποδηλώνει αυτούς που έχουν καταδικαστεί στο παρελθόν για ναρκωτικά (έναντι των υπολοίπων). Τα αποτελέσματα της ανάλυσης, έτσι όπως αυτά παρήχθησαν από το SPSS (Γναρδέλλης, 2006), εμφανίζονται στους Πίνακες 4-10.

Πίνακας 4. Έγκυρες παρατηρήσεις κατά την ανάλυση της λογαριθμικής παλινδρόμησης

Case Processing Summary			
Unweighted Cases		N	Percent
Selected Cases	Included in Analysis	273	94,1
	Missing Cases	17	5,9
	Total	290	100,0
Unselected Cases		0	,0
Total		290	100,0

Πίνακας 5. Εσωτερική κωδικοποίηση της εξαρτημένης μεταβλητής

Dependent Variable Encoding	
Original Value	Internal Value
Όχι	0
Ναι	1

Πίνακας 6. Εσωτερική κωδικοποίηση των ανεξάρτητων κατηγορικών μεταβλητών

Categorical Variables Codings			
		Frequency	Parameter Coding
Καταδίκη για ναρκωτικά	Όχι	158	,000
	Ναι	115	1,000
Ποινικός χαρακτηρισμός	Υπόδικοι	88	,000
	Κατάδικοι	185	1,000

Στον πρώτο πίνακα (Πίνακας 4) των αποτελεσμάτων δίνεται ο αριθμός των έγκυρων παρατηρήσεων της ανάλυσης καθώς και ο αριθμός των παρατηρήσεων με *ελλείπουσες τιμές*[23]. Στους δύο επόμενους πίνακες (Πίνακες 5 και 6) δίνεται η εσωτερική κωδικοποίηση της εξαρτημένης μεταβλητής (χρήση, 0 = όχι, 1 = ναι), καθώς και των ανεξάρτητων κατηγορικών μεταβλητών του μοντέλου (ποινικός χαρακτηρισμός, καταδίκη για ναρκωτικά στο παρελθόν) μαζί με τις αντίστοιχες συχνότητες των κατηγοριών τους.

Ο επόμενος πίνακας (Πίνακας 7) αφορά το αρχικό μοντέλο της ανάλυσης, το οποίο αποτελείται μόνο από το σταθερό όρο χωρίς άλλη ανεξάρτητη μεταβλητή στην εξίσωση της λογαριθμικής παλινδρόμησης, δηλαδή το μοντέλο $\text{logit}(p) = b_0$. Για το συγκεκριμένο μοντέλο, δίνεται η τιμή του σταθερού όρου (B) και οι αντίστοιχοι έλεγχοι επ' αυτού (έλεγχος του Wald).

[23] Missing values. Είναι οι παρατηρήσεις που έχουν ελλείπουσες τιμές σε μια τουλάχιστον από τις μεταβλητές της ανάλυσης.

Πίνακας 7. Αρχικό μοντέλο της ανάλυσης λογαριθμικής παλινδρόμησης

Variables in the Equation							
		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	,393	,123	10,158	1	,001	1,482

Πίνακας 8. Αξιολόγηση της προσαρμογής του τελικού λογαριθμικού μοντέλου με τη βοήθεια του λόγου των μέγιστων τιμών της συνάρτησης πιθανοφάνειας (Η σύγκριση γίνεται ως προς το αρχικό μοντέλο)

Omnibus Tests of Model Coefficients				
		Chi-square	df	Sig.
Step 1	Step	29,395	4	,000
	Block	29,395	4	,000
	Model	29,395	4	,000

Οι τρεις επόμενοι πίνακες αφορούν τη μορφή του τελικού μοντέλου με όλες τις ανεξάρτητες μεταβλητές της ανάλυσης καθώς και την αξιολόγησή του. Η αξιολόγηση της προσαρμογής του μοντέλου στα δειγματικά δεδομένα (Πίνακας 8) γίνεται με το λόγο των μέγιστων τιμών της συνάρτησης πιθανοφάνειας για το πλήρες μοντέλο (L_F) και το μοντέλο που περιλαμβάνει μόνο το σταθερό όρο (L_0). Η τιμή του λόγου είναι, $-2\ln\left(\frac{L_0}{L_F}\right) = 29{,}395$ ενώ η πιθανότητα να προκύψει μια τιμή τόσο μεγάλη για την κατανομή χ^2 με 4 βαθμούς ελευθερίας είναι Sig. < 0,0005. Επομένως, η μηδενική υπόθεση $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ απορρίπτεται. Μπορούμε δηλαδή να θεωρήσουμε ότι οι τέσσερεις ανεξάρτητες μεταβλητές *age* (ηλικία), *inmatype* (ποινικός χαρακτηρισμός), *drugconv* (προηγούμενη καταδίκη για ναρκωτικά) και *total* (συνολικός χρόνος εγκλεισμού) συνδυαζόμενες μεταξύ τους με τη μορφή του λογαριθμικού μοντέλου, συμβάλλουν σημαντικά στην πρόγνωση των τιμών της εξαρτημένης μεταβλητής.

Στον επόμενο πίνακα των αποτελεσμάτων (Πίνακας 9) δίνεται η τιμή της συνάρτησης λογαριθμο-πιθανοφάνειας (-2Log likelihood = 288,334) για το τελικό μοντέλο μαζί με το συντελεστή προσδιορισμού των Cox και

Πίνακας 9. Συνάρτηση λογαριθμο-πιθανοφάνειας του τελικού μοντέλου και συντελεστές προσδιορισμού των Cox & Snell και Nagelkerke

Model Summary			
Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	288,334	,119	,159

Πίνακας 10. Συντελεστές του τελικού λογαριθμικού μοντέλου και επαγωγικοί έλεγχοι επ' αυτών

Variables in the Equation									
		B	S.E.	Wald	df	Sig.	Exp(B)	95,0% C.I. for EXP(B)	
								Lower	Upper
Step 1	Age	-,140	,174	,645	1	,422	,869	,618	1,223
	inmatype	,416	,306	1,846	1	,174	1,516	,832	2,761
	drugconv	,656	,291	5,099	1	,024	1,927	1,090	3,407
	total	,450	,122	13,508	1	,000	1,568	1,234	1,994
	Constant	-1,972	,638	9,539	1	,002	,139		

Snell (0,119) και το συντελεστή προσδιορισμού του Nagelkerke (0,159). Δηλαδή, 16% της μεταβλητότητας της εξαρτημένης μεταβλητής ερμηνεύεται από τις τέσσερις ανεξάρτητες μεταβλητές του μοντέλου.

Ο τελευταίος πίνακας της ανάλυσης (Πίνακας 10) είναι και ο πλέον σημαντικός, διότι μας δίνει τους συντελεστές του τελικού μοντέλου μαζί με τους αντίστοιχους επαγωγικούς ελέγχους και τα διαστήματα εμπιστοσύνης των αντιλογαριθμισμένων τιμών τους, δηλαδή των $e^{\beta i}$. Με βάση το κριτήριο του Wald, σημαντική επίδραση στη διαμόρφωση των τιμών της εξαρτημένης μεταβλητής έχουν οι μεταβλητές *total* (Sig. < 0,0005) και *drugconv* (Sig. = 0,024), δηλαδή ο συνολικός χρόνος εγκλεισμού των κρατουμένων και η καταδίκη τους για ναρκωτικά στο παρελθόν. Σύμφωνα με τους εκτιμώμενους συντελεστές, για κάθε έναν επιπλέον χρόνο φυλάκισης των κρατουμένων (*total*), η σχετική πιθανότητα χρήσης ενδοφλέβιων ναρκωτικών αυξάνει περίπου κατά 57% ($e^{0,450} = 1,568$), ανεξάρτητα από την ηλικία, τον ποινικό χαρακτηρισμό των κρατουμένων και το ενδεχόμενο καταδίκης για

ναρκωτικά στο παρελθόν. Επιπλέον, ο σχετικός λόγος των πιθανοτήτων χρήσης εντός της φυλακής αυτών που έχουν καταδικαστεί για ναρκωτικά στο παρελθόν έναντι των υπολοίπων (*drugconv*) είναι ίσος με 1,93 ($e^{0,656}$ = 1,927), ανεξάρτητα από την ηλικία, τον ποινικό χαρακτηρισμό και το συνολικό χρόνο φυλάκισής τους. Δηλαδή, τα άτομα που έχουν καταδικαστεί στο παρελθόν για αδίκημα σχετικό με τα ναρκωτικά έχουν κατά 93% περίπου μεγαλύτερη πιθανότητα χρήσης εντός της φυλακής έναντι των υπολοίπων (που δεν έχουν καταδικαστεί για αδίκημα σχετικό με τα ναρκωτικά). Το 95%ΔΕ για την αύξηση της σχετικής πιθανότητας χρήσης ανά χρόνο φυλάκισης είναι (1,234, 1,994), ενώ το 95%ΔΕ του σχετικού λόγου χρήσης ναρκωτικών αυτών που έχουν καταδικαστεί για ναρκωτικά έναντι των υπολοίπων είναι (1,090, 3,407) (δηλαδή, με πιθανότητα 95% ο σχετικός κίνδυνος χρήσης ενδοφλέβιων ναρκωτικών αυξάνει από 23% έως 99% περίπου για κάθε επιπλέον χρόνο φυλάκισης ενός κρατουμένου, ενώ ο αντίστοιχος σχετικός κίνδυνος για τα άτομα που έχουν καταδικαστεί για ναρκωτικά μπορεί να είναι μέχρι τριπλάσιος έναντι του σχετικού κινδύνου των υπολοίπων).

Πίνακας και διάγραμμα ταξινόμησης των παρατηρήσεων

Μια περιγραφική διαδικασία για την αξιολόγηση της προσαρμογής του λογαριθμικού μοντέλου είναι η κατασκευή ενός πίνακα συνάφειας, στον οποίο οι παρατηρήσεις ταξινομούνται διαξονικά ως προς την πραγματοποίηση του γεγονότος με βάση τα δειγματικά δεδομένα (δηλαδή, το *παρατηρούμενο αποτέλεσμα*[24]) και ως προς την πραγματοποίηση του γεγονότος με βάση τις εκτιμήσεις του μοντέλου (δηλαδή, το *προβλεπόμενο αποτέλεσμα*[25]). Για να εκτιμηθεί, σύμφωνα με το μοντέλο, ότι το γεγονός θα συμβεί σε μία παρατήρηση, πρέπει η εκτιμώμενη πιθανότητα πραγματοποίησης του γεγονότος για την παρατήρηση να είναι μεγαλύτερη ή ίση από 0,5. Ένας πίνακας αυτού του τύπου (έτσι όπως παράγεται στα αποτελέσματα του SPSS) είναι ο Πίνακας 11 ο οποίος αφορά τα δεδομένα του παραδείγματος των φυλακών.

[24] Observed outcome.
[25] Predicted outcome.

Πίνακας 11. Πίνακας ταξινόμησης των παρατηρήσεων

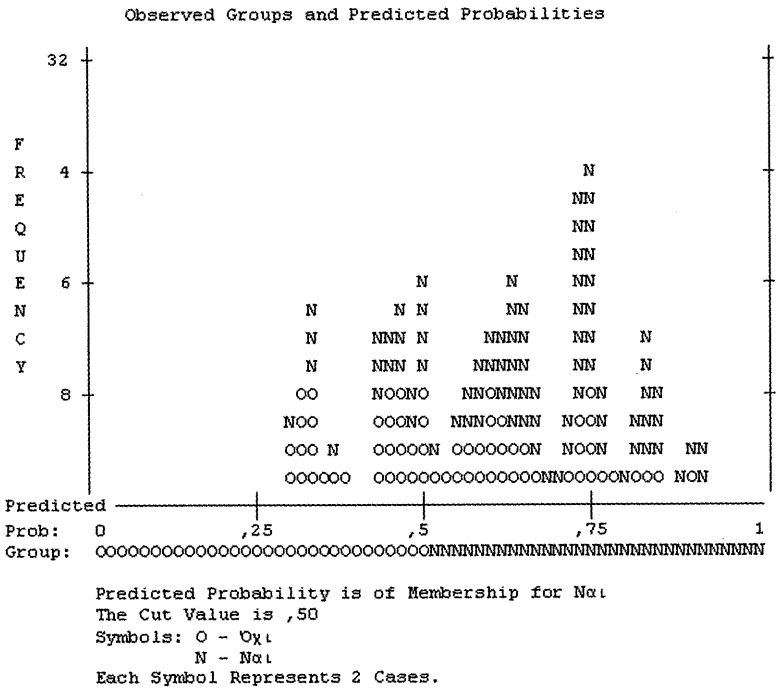
Classification Table[a]

			Predicted		
			Χρήση ενδοφλέβιων ναρκωτικών στη φυλακή		
	Observed		Όχι	Ναι	Percentage Correct
Step 1	Χρήση ενδοφλέβιων ναρκωτικών στη φυλακή	Όχι	52	58	47,3
		Ναι	38	125	76,7
	Overall Percentage				64,8

Σημείωση: [(a)] The cut value is .500.

Σε έναν πίνακα ταξινόμησης θα πρέπει οι παρατηρούμενες και οι εκτιμώμενες τιμές να συμφωνούν κατά το δυνατόν περισσότερο. Στον Πίνακα 11 η συμφωνία αυτή προσεγγίζει το 65% περίπου του συνόλου των παρατηρήσεων. Τα διαγώνια κελιά του πίνακα περιέχουν τις παρατηρήσεις που συμφωνούν ως προς τις δύο ταξινομήσεις (την παρατηρούμενη και την εκτιμώμενη από το μοντέλο), ενώ τα εκτός διαγωνίου κελιά περιέχουν τις παρατηρήσεις με ασυμφωνία. Το πρόβλημα με έναν πίνακα ταξινόμησης είναι ότι δεν παρέχει καμιά πληροφορία για το μέγεθος των εκτιμώμενων πιθανοτήτων στις περιπτώσεις που έχουμε εσφαλμένες εκτιμήσεις. Για παράδειγμα, δε γνωρίζουμε ποιες είναι οι εκτιμώμενες πιθανότητες για τα 38 άτομα τα οποία, ενώ κάνουν χρήση ενδοφλέβιων ναρκωτικών στη φυλακή, από το μοντέλο εκτιμώνται ότι δεν κάνουν. Μια ερώτηση, για παράδειγμα, που θα μπορούσε να τεθεί είναι αν οι εκτιμώμενες πιθανότητες για τα άτομα αυτά είναι κοντά στο όριο του 0,5 ή διαφοροποιούνται σημαντικά από αυτό (κάτι που θα μείωνε ακόμη περισσότερο την αξιοπιστία του μοντέλου). Ερωτήματα αυτού του τύπου μπορούν να απαντηθούν από το ιστόγραμμα των εκτιμώμενων πιθανοτήτων, ή αλλιώς *διάγραμμα ταξινόμησης*[2] των παρατηρήσεων, το οποίο επίσης μπορεί να κατασκευαστεί κατά την ανάλυση της λογαριθμικής παλινδρόμησης. Το διάγραμμα ταξινόμησης

[26] Classification plot.



Σχήμα 2. Διάγραμμα ταξινόμησης των παρατηρήσεων.

για τα δεδομένα του παραδείγματος των φυλακών είναι αυτό που εμφανίζεται στο Σχήμα 2.

Ο οριζόντιος άξονας του διαγράμματος ταξινόμησης αντιστοιχεί στις εκτιμώμενες από το μοντέλο πιθανότητες. Επάνω στον άξονα τοποθετούνται οι παρατηρήσεις της ανάλυσης υπό μορφή ιστογράμματος. Οι παρατηρήσεις οι οποίες έχουν εκτιμώμενη πιθανότητα πραγματοποίησης του γεγονότος μεγαλύτερη του 0,5 τοποθετούνται στο δεξιό μέρος του άξονα, ενώ οι παρατηρήσεις με πιθανότητα μικρότερη του 0,5 στο αριστερό. Επιπλέον η κάθε παρατήρηση ορίζεται από το αρχικό γράμμα της κατηγορίας που αντιστοιχεί στην πραγματοποίηση ή μη του γεγονότος έτσι όπως προκύπτει από τα δειγματικά δεδομένα. Στο παράδειγμα του Σχήματος 2, όπου το γεγονός που μελετάται είναι η χρήση ενδοφλέβιων ναρκωτικών στη φυλακή, τα άτομα (οι παρατηρήσεις) που έχουν εκτιμώμενη πιθανότητα χρή-

σης μεγαλύτερη του 0,5 τοποθετούνται στο δεξιό άκρο του οριζόντιου άξονα ενώ τα υπόλοιπα στο αριστερό. Επιπλέον κάθε άτομο συμβολίζεται με Ν (Ναι) ή Ο (Όχι) ανάλογα, αν με βάση τα δειγματικά δεδομένα κάνει ή δεν κάνει χρήση ενδοφλέβιων ναρκωτικών. Λογικό είναι στο διάγραμμα αυτό να αναμένουμε μεγάλη συσσώρευση ατόμων που κάνουν χρήση (συμβολιζόμενα με Ν) στο δεξιό άκρο του οριζόντιου άξονα με αντίστοιχη συσσώρευση ατόμων που δεν κάνουν χρήση (συμβολιζόμενα με Ο) στο αριστερό άκρο. Σε μια ιδανική περίπτωση θα έπρεπε και οι δύο αυτές ομάδες των ατόμων να είναι όσο το δυνατόν πιο απομακρυσμένες στα δύο άκρα του άξονα (κάτι που θα υποδήλωνε μικρές διαφοροποιήσεις των εκτιμώμενων τιμών από τις πραγματικές). Βέβαια η ιδανική αυτή κατάσταση δεν εμφανίζεται στο διάγραμμα του Σχήματος 2, εμφανίζεται όμως ένας αρκετά μεγάλος αριθμός παρατηρήσεων με (ορθά) εκτιμώμενες πιθανότητες χρήσης σημαντικά μεγαλύτερες του ορίου 0,5 (αρκετές από αυτές είναι μεγαλύτερες της τιμής 0,7). Αντιθέτως, ο αριθμός των παρατηρήσεων που έχουν (ορθά) εκτιμώμενες πιθανότητες μη χρήσης είναι αρκετά μικρότερος και μάλιστα με τις τιμές των αντίστοιχων πιθανοτήτων να μη διαφοροποιούνται πολύ από το 0,5.

ΣΥΖΗΤΗΣΗ

Η λογαριθμική παλινδρόμηση, ως μια ειδική περίπτωση των γενικευμένων γραμμικών μοντέλων, παρέχει τη δυνατότητα πρόβλεψης της έκβασης ενός γεγονότος, εκτιμώντας τη σχετική πιθανότητα πραγματοποίησής του, μέσα από μια εξίσωση γραμμικής παλινδρόμησης. Η σχετική πιθανότητα πραγματοποίησης του γεγονότος μετασχηματισμένη λογαριθμικά, μπορεί να αποτελέσει την εξαρτημένη μεταβλητή ενός γραμμικού μοντέλου, το οποίο θεωρητικά ικανοποιεί τις ελάχιστες προϋποθέσεις που απαιτούνται κατά την ανάλυση της γραμμικής παλινδρόμησης. Το ισχυρό πλεονέκτημα της λογαριθμικής παλινδρόμησης έναντι των άλλων τεχνικών που χρησιμοποιούνται κατά την εκτίμηση δίτιμων αποτελεσμάτων είναι οι σχετικά ελαστικές προϋποθέσεις χρήσης της, οι οποίες ταυτίζονται με αυτές της γραμμικής παλινδρόμησης, καθώς και η ερμηνεία των συντελεστών της, οι οποίοι εκφράζουν τη συμβολή κάθε ανεξάρτητης μεταβλητής στην αύξηση ή ελάττωση της σχετικής πιθανότητας πραγματοποίησης του διερευνώμενου γεγονότος. Σε αντίθεση με την απλότητα των προϋποθέσεων της λογαριθ-

μικής παλινδρόμησης η *διακρίνουσα ανάλυση*[27], η οποία χρησιμοποιείτο παλαιότερα σε περιπτώσεις μοντέλων με δίτιμες εξαρτημένες μεταβλητές, απαιτεί πολυμεταβλητή κανονικότητα των ανεξάρτητων μεταβλητών καθώς και ισότητα των πινάκων διακύμανσης-συνδιακύμανσης στις δύο κατηγορίες της εξαρτημένης μεταβλητής. Επιπλέον, η ερμηνεία των συντελεστών της δεν έχει την ευθύτητα της πιθανολογικής ερμηνείας που έχουν οι συντελεστές του λογαριθμικού μοντέλου.

Η χρήση της λογαριθμικής παλινδρόμησης έχει διευρυνθεί σημαντικά τις τελευταίες δύο δεκαετίας τόσο στις εμπειρικές έρευνες των κοινωνικών επιστημών αλλά κυρίως στις αιτιολογικές έρευνες της αναλυτικής επιδημιολογίας. Στις τελευταίες αυτές μελέτες, όπου η κεντρική υπόθεση που διερευνάται αφορά την αιτιολογική σχέση ενός νοσήματος με ένα σύνολο παραγόντων, η εφαρμογή της λογαριθμικής παλινδρόμησης συναρτάται ευθέως με την ευρέως χρησιμοποιούμενη έννοια του *σχετικού κινδύνου*[28]. Δηλαδή, της σχετικής πιθανότητας εμφάνισης του νοσήματος στους εκτεθέντες σε έναν παράγοντα κινδύνου προς την πιθανότητα εμφάνισης του νοσήματος στους μη εκτεθέντες. Σε αυτές τις περιπτώσεις, οι συντελεστές ενός μοντέλου λογαριθμικής παλινδρόμησης με ανεξάρτητες μεταβλητές τους διερευνώμενους παράγοντες, αποτελούν σημειακές εκτιμήσεις των αντίστοιχων σχετικών κινδύνων των παραγόντων (Breslow & Day, 1980, 1987).

ΒΙΒΛΙΟΓΡΑΦΙΑ

Aggresti, A. (1990). *Categorical data analysis*. New York: Wiley.

Breslow, N. E., & Day, N. E. (1980). *Statistical methods in cancer research: Volume I. The analysis of case-control studies*. Lyon, France: IARC.

Breslow, N. E., & Day, N. E. (1987). *Statistical methods in cancer research: Volume II. The design and analysis of cohort studies*. Lyon, France: IARC.

Cox, D. R., & Snell, E. J. (1989). *Analysis of binary data.* London: Chapman and Hall.

Γναρδέλλης, Χ. (2003). *Εφαρμοσμένη Στατιστική*. Αθήνα: Εκδόσεις Παπαζήση.

Γναρδέλλης, Χ. (2006). *Ανάλυση δεδομένων με το SPSS 14.0 for Windows*. Αθήνα: Παπαζήσης.

Γναρδέλλης, Χ., Λάγιου, Α., Χλόπτσιος, Ι., Μπενέτου, Β., & Τριχοπούλου, Α. (1999). Εκτίμηση φυσικής δραστηριότητας σε επιδημιολογικές μελέτες. Η ελληνική εμπειρία στο πλαίσιο του προγράμματος ΕΠΙΚ. *Ιατρική, 75*, 551-558.

[27] Discriminant analysis.

[28] Relative risk.

Dobson, A. J. (1990). *An introduction to generalized linear models*. London: Chapman and Hall.

Everitt, B. S., & Dunn, G. (1992). *Applied multivariate data analysis*. New York: Oxford University Press.

Hauck, W. W., & Donner, A. (1977). Wald's test as applied to hypotheses in logit analysis. *Journal of the American Statistical Association, 72*, 851-853.

Hosmer, D. W., & Lemeshow, S. (1989). *Applied logistic regression*. New York: Wiley.

Koulierakis, G., Gnardellis, Ch., Agrafiotis, D., & Power, K. G. (2000). HIV risk behaviour correlates among injecting drug users in Greek prisons. *Addiction, 95,* 1207-1216.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models*. London: Chapman and Hall.

Nagelkerke, N. J. D. (1991). A note on general definition of the coefficient of determination. *Biometrika, 78,* 691-692.

Rao, C. R. (1973). *Linear statistical inference and its application*. New York: Wiley.

THE USE OF LOGISTIC REGRESSION MODELS IN THE EMPIRICAL RESEARCH OF SOCIAL SCIENCES

Charalambos Gnardellis

Technological Educational Institute of Messolonghi, Greece

Abstract: An issue that is often posed in the quantitative approaches of sociological research regards the relation of a dichotomous categorical variable with a set of other explanatory variables. The usual case of these investigations focuses on the probability that an event occurs (the "success" of the event or not) in regard to a set of other factors which potentially influence the probability of success. For instance, the injective drug use within prison in regard to the penal history of inmates, their total time in prison and a previous drug conviction. In these cases, usual models of linear regression are not suitable for the study of the dichotomous outcome, because the assumptions of the models do not correspond in the measurement level of the response binary variable. Necessary condition for the use of a linear model in these cases is the transformation of the outcome variable in terms of the odds of the success, that is, the ratio of the probability of the success to the probability of failure. The logarithm of this ratio (named logit), which takes any real value, can constitute the dependent variable in a usual linear regression model.

Key words: Generalized linear models, Logistic regression, Sociological research.

Address: Charalambos Gnardellis, Technological Educational Institute (T.E.I.) of Messolonghi, Nea Ktiria, 302 00 Messolonghi, Greece. Phone: +30-210-7512651, +30-6972148215. E-mail: hgnardellis@yahoo.gr