Μοντέλα πρόβλεψης με κατηγορικά δεδομένα στην ψυχολογική έρευνα μέσω των στατιστικών μεθόδων της ανάλυσης προβλέψεων και της λογιστικής παλινδρόμησης

ΓΡΗΓΟΡΗΣ ΚΙΟΣΕΟΓΛΟΥ
Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης

ΠΕΡΙΛΗΨΗ Οι στατιστικές μέθοδοι που αποσκοπούν στην κατασκευή μοντέλων πρόβλεψης εφαρμόζονται σε ένα μεγάλο φάσμα περιοχών της εμπειρικής επιστημονικής έρευνας στην οποία ανήκουν και οι επιστήμες της συμπεριφοράς. Ειδική περίπτωση αποτελούν οι μέθοδοι εκείνες που χειρίζονται κατηγορικά δεδομένα και οι οποίες, παρόλο το ενδιαφέρον που παρουσιάζουν, είναι λιγότερο γνωστές λόγω του ότι είναι σχετικά πιο πρόσφατες. Δύο τέτοιες στατιστικές μέθοδοι είναι αυτές της Ανάλυσης Προβλέψεων και της Λογιστικής Παλινδρόμησης. Αν και οι μέθοδοι αυτοί είναι, από την άποψη της στατιστικής τεχνικής που χρησιμοποιούν, τελείως διαφορετικές, έχουν το κοινό γνώρισμα ότι επιτρέπουν την πρόβλεψη των καταστάσεων μιας εξαρτημένης κατηγορικής μεταβλητής από μια ή περισσότερες ανεξάρτητες κατηγορικές μεταβλητές. Στην παρούσα εργασία επιχειρείται η παρουσίαση των βασικών αρχών των δύο αυτών μεθόδων. Ταυτόχρονα, δίνονται παραδείγματα από ερευνητικά δεδομένα που προέρχονται από το χώρο της εξελικτικής ψυχολογίας, της ψυχολογίας της γλώσσας, και της κοινωνικής ψυχολογίας με στόχο την καλύτερη κατανόηση των μεθόδων και την αποτελεσματική εφαρμογή τους από τους ερευνητές των επιστημών της συμπεριφοράς.

Λέξεις κλειδιά: Ανάλυση προβλέψεων, λογιστική παλινδρόμηση.

Η έννοια της στατιστικής πρόβλεψης έχει απασχολήσει ερευνητές από διάφορους επιστημονικούς κλάδους από πολύ παλιά. Οι μέθοδοι που αρχικά προτάθηκαν αποσκοπούσαν στη δημιουργία μοντέλων πρόβλεψης χρησιμοποιώντας ποσοτικές μεταβλητές. Αργότερα το ενδιαφέρον στράφηκε και στη χρήση ποιοτικών μεταβλητών. Τέτοιου είδους μεταβλητές συναντώνται αρκετά συχνά στις επιστήμες της συμπεριφοράς, γεγονός που προκαλεί το ενδιαφέρον για την παρουσίαση μεθόδων οι οποίες αποβλέπουν στη δημιουργία μοντέλων πρόβλεψης με ποιοτικές μεταβλητές.

Η εργασία αυτή στοχεύει στην παρουσίαση των βασικών αρχών δύο στατιστικών μεθόδων, αυτών της Ανάλυσης προβλέψεων και της Λογιστικής παλινδρόμησης, οι οποίες μπορούν να χρησιμοποιηθούν με αποτελεσματικό τρόπο στην ψυχολογική έρευνα στην περίπτωση κατά την οποία το ενδιαφέρον του ερευνητή εστιάζεται στην αξιολόγηση της πρόβλεψης μιας εξαρτημένης κατηγορικής μεταβλητής από μια ή περισσότερες ανεξάρτητες κατηγορικές μεταβλητές.

Ανάλυση Προβλέψεων

Η μέθοδος της ανάλυσης προβλέψεων προτάθηκε από τους Hildebrand, Laing και Rosenthal

Διεύθυνση: Γ. Κοσεόγλου, Τμήμα Ψυχολογίας, Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης, 540 06, Θεσσαλονίκης. Τηλ. 031-997337, e-mail: kios@psy.auth.gr

το 1977 (βλέπε ακόμα Froman & Hubert, 1980) ως μέθοδος για την αξιολόγηση μοντέλων βάσει των οποίων επιχειρείται πρόβλεψη των κατηγοριών μιας κατηγορικής μεταβλητής από μια ή περισσότερες άλλες κατηγορικές μεταβλητές.

Στη μέθοδο της Ανάλυσης προβλέψεων (ΑΠ) τα μοντέλα πρόβλεψης διατυπώνονται από τον ερευνητή a priori δηλαδή πριν τη συλλογή των δεδομένων και συνήθως προκύπτουν από κάποια θεωρία. Η παρουσίαση της μεθόδου θα αφορά την εφαρμογή της σε μοντέλα που περιλαμβάνουν μια εξαρτημένη μεταβλητή Υ και μια ανεξάρτητη Χ (δισδιάστατη πρόβλεψη) καθώς και σε μοντέλα με δύο ανεξάρτητες μεταβλητές Χ και Ζ (τρισδιάστατη πρόβλεψη). Τα δύο παραδείγματα που θα παρουσιαστούν προέρχονται από το χώρο της εξελικτικής ψυχολογίας (βλέπε Κιοσέογλου & Δημητρίου, 1992).

Δισδιάστατη πρόβλεψη

Στη δισδιάστατη πρόβλεψη η μέθοδος χρησιμοποιείται σε δεδομένα που βρίσκονται υπό τη μορφή ενός πίνακα συμπτώσεων Υ x Χ με *m* γραμμές και *n* στήλες του οποίου οι γραμμές i (i=1... m) αποτελούν τις κατηγορίες της μιας ποιοτικής μεταβλητής Υ και οι στήλες j (j=1... n) τις κατηγορίες της δεύτερης ποιοτικής μεταβλητής Χ. Στον πίνακα αυτό, στη διασταύρωση της κατηγορίας i της μεταβλητής Υ και της κατηγορίας j της μεταβλητής Χ εμφανίζεται το πλήθος των ατόμων που ανήκουν ταυτόχρονα στις κατηγορίες i και j δηλαδή στην κυψέλη (i, j). Ορίζονται ακόμα η πιθανότητα p(i, j) ύπαρξης ατόμων στην κυψέλη (i, j) καθώς και οι πιθανότητες περιθωρίου p(i.) και p(.j) δηλαδή οι πιθανότητες του να υπάρχουν άτομα στις κατηγορίες i και j αντίστοιχα.

Στην ΑΠ μια συγκεκριμένη πρόβλεψη μπορεί να διατυπωθεί στον πίνακα Υ x Χ μέσω των κυψελών σφάλματος, δηλαδή εκείνων των κυψελών στις οποίες σύμφωνα με το μοντέλο (τις υποθέσεις) του ερευνητή δε θα πρέπει να ισχύει η πρόβλεψη (δηλαδή η κατηγορία x της Χ δεν προβλέπει την κατηγορία y της Υ). Στον Πίνακα 1 διασταυρώνονται δύο κατηγορικές μεταβλητές Υ και Χ με τρεις κατηγορίες η κάθε μια. Οι γραμμοσκιασμένες κυψέλες είναι οι κυψέλες σφάλματος στις οποίες θα πρέπει, σύμφωνα με την πρόβλεψη, να μην υπάρχουν άτομα.

Η πρόβλεψη P διατυπώνεται ως εξής:

$$P: x_1 \rightarrow (y_1 \text{ ή } y_3),\ x_2 \rightarrow y_2 \ \&\ x_3 \rightarrow y_3$$

όπου το σύμβολο → σημαίνει προβλέπει.

Η πρόβλεψη μπορεί επίσης να οριστεί χρησιμοποιώντας το δείκτη κυψελών σφάλματος ω(i, j). Συγκεκριμένα, αν ω(i, j)=1 τότε η κυψέλη (y, x) θεωρείται κυψέλη σφάλματος. Αντίθετα, αν ω(i, j)=0 τότε η κυψέλη (y, x) δεν είναι κυψέλη σφάλματος. Σε ορισμένες περιπτώσεις είναι δυνατό ο δείκτης ω να πάρει, για κάποιες κυψέλες σφάλ-

Πίνακας 1
Κυψέλες σφάλματος στη δισδιάστατη πρόβλεψη

	x_1	x_2	x_3
y_1		■	■
y_2	■		■
y_3		■	

ματος, την τιμή 0.5 αντί της τιμής 1 όταν για τις κυψέλες αυτές δικαιολογείται η ύπαρξη κάποιων ελάχιστων ατόμων. Στην παραπάνω πρόβλεψη ο δείκτης κυψελών σφάλματος παίρνει τις τιμές: $\omega(2, 1)=\omega(1, 2)=\omega(3, 2)=\omega(1, 3)=\omega(2, 3)=1$ ενώ για όλες τις άλλες κυψέλες $\omega(i, j)=0$.

Η μέθοδος ΑΠ αξιολογεί μια συγκεκριμένη πρόβλεψη με το λεγόμενο μέτρο της επιτυχίας της πρόβλεψης (Hildebrand, Laing, & Rosenthal, 1977) το οποίο ορίζεται ως:

$$\nabla(YX)=1-\frac{K}{U}$$

όπου:

K είναι η πιθανότητα ύπαρξης ατόμων στις κυψέλες σφάλματος

$$K=\left\{\sum_i\sum_j \omega(i,j)p(i,j)\,|\, i=1...m, j=1...n\right\}$$

δηλαδή η πιθανότητα σφάλματος κάτω από το συγκεκριμένο μοντέλο, και U είναι η πιθανότητα με την οποία θα αναμέναμε να υπάρχουν άτομα στις κυψέλες σφάλματος αν υπήρχε στατιστική ανεξαρτησία μεταξύ των δύο μεταβλητών, αν δηλαδή κατά την πρόβλεψη της Y καμία απολύτως συνεισφορά δεν είχε η X.

$$U=\left\{\sum_i\sum_j \omega(i,j)p(i.)p(.j)\,|\, i=1...m, j=1...n\right\}$$

Το μέτρο $\nabla(YX)$ έχει εύρος τιμών από $-\infty$ έως $+1$. Η τιμή $+1$ επιτυγχάνεται στην περίπτωση ιδανικής πρόβλεψης όταν δηλαδή όλες οι κυψέλες σφάλματος είναι κενές που σημαίνει ότι για κάθε κυψέλη σφάλματος $p(i, j)=0$. Όσο η τιμή του μέτρου πλησιάζει το $+1$ τόσο έχουμε μεγαλύτερη αναλογική μείωση των σφαλμάτων βάσει του μοντέλου πρόβλεψης και συνεπώς η πρόβλεψη είναι επιτυχέστερη. Η τιμή 0 αποκτάται όταν $K = U$. Αυτό συμβαίνει όταν οι μεταβλητές X και Y είναι ανεξάρτητες, δηλαδή όταν για κάθε κυψέλη ισχύει $p(i, j)=p(i.)p(.j)$. Αρνητική τιμή του μέτρου σημαίνει πλήρη αποτυχία του προτεινόμενου μοντέλου γιατί τότε η πιθανότητα σφάλματος βάσει του μοντέλου θα είναι μεγαλύτερη από αυτήν του μοντέλου της ανεξαρτησίας μεταξύ των δύο κατηγορικών μεταβλητών.

Η ποσότητα U εκφράζει τη λεγόμενη ακρίβεια της πρόβλεψης και η τιμή της αυξάνει γενικά με το πλήθος των κυψελών σφάλματος. Όταν έχουμε μεγάλο πλήθος κυψελών σφάλματος τότε η κάθε κατηγορία x της μεταβλητής X θα προβλέπει έναν μάλλον μικρό αριθμό κατηγοριών της εξαρτημένης μεταβλητής Y και συνεπώς η πρόβλεψη θα είναι πιο ακριβής.

Στην ΑΠ χρησιμοποιούνται διάφοροι στατιστικοί έλεγχοι που αφορούν το μέτρο $\nabla(YX)$. Ειδικότερα, μπορούμε να διακρίνουμε τριών ειδών ελέγχους στους οποίους οι αντίστοιχες μηδενικές υποθέσεις διατυπώνονται ως εξής:

1) H_0: $\nabla(YX)=0$

2) H_0: $\nabla(YX)=\nabla'(YX)$ (μια πρόβλεψη για διαφορετικούς πληθυσμούς)

3) H_0: $\nabla(YX)=\nabla'(YX)$ (δύο διαφορετικές προβλέψεις για τον ίδιο πληθυσμό).

Η πρώτη περίπτωση ελέγχου είναι η πλέον χρησιμοποιούμενη επειδή στοχεύει στο να ελέγξει τη στατιστική σημαντικότητα του μέτρου $\nabla(YX)$. Αν η μηδενική υπόθεση δεν απορριφθεί, τότε οι δύο μεταβλητές είναι ανεξάρτητες. Η δεύτερη περίπτωση εφαρμόζεται όταν ελέγχουμε το ίδιο μοντέλο πρόβλεψης σε δύο διαφορετικούς και ανεξάρτητους πληθυσμούς οπότε είναι επιθυμητός ο στατιστικός έλεγχος της ισότητας των αντίστοιχων μέτρων $\nabla(YX)$ και $\nabla'(YX)$ για να διαπιστωθεί αν αυτά είναι ίσα ή διαφέρουν υπέρ κάποιου πληθυσμού. Στην τρίτη περίπτωση, τέλος, έχουμε δύο εναλλακτικά μοντέλα πρόβλεψης και δοκιμάζοντάς τα στον ίδιο πληθυσμό αποσκοπούμε, συγκρίνοντας τα αντίστοιχα μέτρα τους, να επιλέξουμε το καλύτερο από αυτά. Η περιγραφή των τύπων για την εφαρμογή των στατιστικών αυτών ελέγχων δεν παρουσιάζεται εδώ αλλά παραπέμπουμε τον αναγνώστη στο βασικό σύγγραμμα των Hildebrand et al. (1977), που αναφέρεται στη μέθοδο.

Παράδειγμα:

Στον παρακάτω πίνακα δεδομένων (Πίνακας 2) παρουσιάζεται η κατανομή 372 νεαρών ατόμων 9-16 χρόνων ως προς τα επίπεδα μέσω των

οποίων εξελίσσεται η ικανότητα να εκτελεί κανείς τις τέσσερις πράξεις της αριθμητικής και ως προς τα επίπεδα μέσω των οποίων εξελίσσεται η αλγεβρική ικανότητα. Με Α, Β, Γ, Δ, Ε συμβολίζονται τα επίπεδα της ικανότητας για εκτέλεση των αριθμητικών πράξεων και με Ι, ΙΙ, ΙΙΙ, IV, V τα επίπεδα της αλγεβρικής ικανότητας.

Στα δεδομένα εφαρμόστηκαν δύο μοντέλα που επιχειρούν να προβλέψουν τον τρόπο εξέλιξης των δύο αυτών μαθηματικών ικανοτήτων. Το πρώτο μοντέλο (Πίνακας 2, μοντέλο 1) προβλέπει ότι η ικανότητα για εκτέλεση αριθμητικών πράξεων εξελίσσεται πριν την αλγεβρική ικανότητα ή συγχρόνως με αυτήν. Το δεύτερο μοντέλο (Πίνακας 2, μοντέλο 2) προβλέπει ότι η ικανότητα για εκτέλεση αριθμητικών πράξεων εξελίσσεται πριν την αλγεβρική ικανότητα εκτός του 5ου επιπέδου όπου η εξέλιξη είναι σύγχρονη.

Η εφαρμογή της ΑΠ έδωσε για το πρώτο μοντέλο $\nabla(YX)=0.597$ ($p<0.001$), $U=0.227$ ενώ για το δεύτερο μοντέλο $\nabla'(YX)=0.246$ ($p<0.001$), $U'=0.317$. Επίσης βρέθηκε ότι οι δύο δείκτες επιτυχίας της πρόβλεψης διαφέρουν σημαντικά ($p<0.001$). Το δεύτερο μοντέλο επιτυγχάνει μια κάπως μεγαλύτερη ακρίβεια της πρόβλεψης ($U'=0.317$) λόγω του ότι περιλαμβάνει μεγαλύτερο αριθμό κυψελών σφάλματος. Ο δείκτης όμως της επιτυχίας της πρόβλεψης στο δεύτερο μοντέλο είναι στατιστικά μικρότερος από αυτόν του πρώτου, και συνεπώς το πρώτο μοντέλο είναι αυτό που πρέπει τελικά να προτιμηθεί.

Τρισδιάστατη πρόβλεψη

Στην τρισδιάστατη πρόβλεψη το μοντέλο περιλαμβάνει μια εξαρτημένη μεταβλητή Y με m κατηγορίες και δύο ανεξάρτητες μεταβλητές: τη X με n κατηγορίες και τη Z με q κατηγορίες. Τα δεδομένα στην τρισδιάστατη πρόβλεψη παρουσιάζονται σε έναν τρισδιάστατο πίνακα συμπτώσεων Y x X x Z με m x n x q κυψέλες. Στην κυψέλη (i, j, k) του πίνακα αυτού εμφανίζεται ο αριθμός των ατόμων που ανήκουν ταυτόχρονα στην κατηγορία i (i=1... m) της Y, στην κατηγορία j (j=1... n) της X και στην κατηγορία k (k=1...q) της μεταβλητής Z. Μπορούν επίσης να οριστούν οι πιθανότητες p(i, j, k) καθώς και οι διάφορες πιθανότητες περιθωρίου του τρισδιάστατου πίνακα. Η περίπτωση της τρισδιάστατης πρόβλεψης μπορεί να αναχθεί στην περίπτωση της δισδιάστατης πρόβλεψης αρκεί να θεωρήσουμε τους συνδυασμούς των κατηγοριών των δύο ανεξάρτητων μεταβλητών ως τις κατηγορίες μιας νέας σύνθετης

Πίνακας 2
Μοντέλο εξελικτικής προτεραιότητας της ικανότητας εκτέλεσης αριθμητικών πράξεων έναντι της αλγεβρικής ικανότητας. Πίνακας δεδομένων και κυψέλες σφάλματος για τα δύο μοντέλα πρόβλεψης

	Δεδομένα						Μοντέλο 1						Μοντέλο 2				
	Α	Β	Γ	Δ	Ε		Α	Β	Γ	Δ	Ε		Α	Β	Γ	Δ	Ε
Ι	5	11	4	3	5	Ι	0	0	0	0	0	Ι	1	0	0	0	0
ΙΙ	6	18	12	30	14	ΙΙ	1	0	0	0	0	ΙΙ	1	1	0	0	0
ΙΙΙ	3	11	19	21	30	ΙΙΙ	1	1	0	0	0	ΙΙΙ	1	1	1	0	0
IV	1	1	5	13	53	IV	1	1	1	0	0	IV	1	1	1	1	0
V	0	1	2	4	100	V	1	1	1	1	0	V	1	1	1	1	0

Σημ.: Τα Α, Β, Γ, Δ, Ε παριστούν τα επίπεδα μέσω των οποίων εξελίσσεται η ικανότητα να επιτελεί κανείς τις τέσσερις πράξεις της αριθμητικής. Τα Ι, ΙΙ, ΙΙΙ, IV, V παριστούν τα επίπεδα μέσω των οποίων εξελίσσεται η αλγεβρική ικανότητα.

κατηγορικής μεταβλητής W.

Στον Πίνακα 3 διασταυρώνονται τρεις κατηγορικές μεταβλητές Y, X και Z που αποτελούνται από δύο κατηγορίες η κάθε μια. Οι κυψέλες με τιμή 1 δηλώνουν τις κυψέλες σφάλματος.

Η πρόβλεψη P εκφράζεται ως εξής:

P: (w_1) x_1 & z_1 → y_1
(w_2) x_1 & z_2 → $(y_1$ ή $y_2)$
(w_3) x_2 & z_1 → y_1
(w_4) x_2 & z_2 → y_2

όπου με w_1, w_2, w_3, w_4 συμβολίζονται οι τέσσερις κατηγορίες της μεταβλητής W που είναι οι συνδυασμοί των κατηγοριών των δύο ανεξάρτητων μεταβλητών X και Z. Χρησιμοποιώντας το δείκτη κυψελών σφάλματος έχουμε ότι: $\omega(2,1,1)=\omega(2,2,1)=\omega(1,2,2)=1$. Για όλες τις άλλες κυψέλες $\omega(i,j,k)=0$.

Στην τρισδιάστατη πρόβλεψη μπορούμε να ορίσουμε το μέτρο επιτυχίας της πρόβλεψης ∇(YXZ) κατά τρόπο ανάλογο της δισδιάστατης πρόβλεψης.

$$\nabla(YXZ) = 1 - \frac{\sum_i \sum_j \sum_k \omega(i,j,k)p(i,j,k)}{\sum_i \sum_j \sum_k \omega(i,j,k)p(i..)p(.jk)}$$

όπου

$$p(i..)=\sum_j \sum_k p(i,j,k) \quad \text{και}$$

$$p(.jk)=\sum_i p(i,j,k) \quad i=1...m,\ j=1...n,\ k=1...q$$

Το μέτρο αυτό είναι ουσιαστικά το μέτρο για τη δισδιάστατη πρόβλεψη μεταξύ της εξαρτημένης μεταβλητής Y και της σύνθετης μεταβλητής W. Το εύρος των τιμών του μέτρου ∇(YXZ) είναι πάλι από $-\infty$ έως $+1$. Επίσης, ο παρονομαστής του παραπάνω λόγου αποτελεί, όπως και στη δισδιάστατη πρόβλεψη, μέτρο για την ακρίβεια της πρόβλεψης.

Στην τρισδιάστατη πρόβλεψη μπορεί να οριστεί ένα ακόμα μέτρο που επιτρέπει τη μέτρηση της συνεισφοράς της δεύτερης ανεξάρτητης μεταβλητής Z στην πρόβλεψη της εξαρτημένης μεταβλητής Y, πέραν της όποιας συνεισφοράς έχει η πρώτη ανεξάρτητη μεταβλητή X. Το μέτρο αυτό δίνεται από τον τύπο:

$$\nabla(YXZX)=1-\frac{\sum_i \sum_j \sum_k \omega(i,j,k)p(i,j,k)}{\sum_i \sum_j \sum_k \omega(i,j,k)[p(ij.)/p(.j.)]p(.jk)}$$

όπου:

$$p(.j.)=\sum_i \sum_k p(i,j,k) \quad \text{και}$$

$$p(ij.)=\sum_k p(i,j,k) \quad i=1...m,\ j=1...n,\ k=1...q$$

Οι τιμές του μέτρου αυτού που ονομάζεται μέτρο μερικής επιτυχίας εκτείνονται επίσης στο διάστημα από $-\infty$ έως $+1$. Μικρή τιμή του μέτρου σημαίνει ότι πέραν της συνεισφοράς της πρώτης ανεξάρτητης μεταβλητής X, ο ρόλος της δεύτερης μεταβλητής Z είναι ασήμαντος.

Επισημαίνουμε τέλος ότι, όπως στη δισδιάστατη πρόβλεψη έτσι και στην τρισδιάστατη, στατιστικοί έλεγχοι μπορούν να πραγματοποιη-

Πίνακας 3
Κυψέλες σφάλματος στην τρισδιάστατη πρόβλεψη

	x_1 & z_1	x_1 & z_2	x_2 & z_1	x_2 & z_2
y_1	0	0	0	1
y_2	1	0	1	0

θούν για να ελέγξουν τη στατιστική σημαντικότητα των μέτρων ∇(YXZ) και ∇(YXZIX).

Παράδειγμα

Τα δεδομένα βασίζονται σε έρευνα στην οποία συμμετείχαν 338 άτομα νεαρής ηλικίας. Τα άτομα κατατάχθηκαν σε τρία επίπεδα (που συμβολίζονται με 0, 1, 2) γνωστικής ανάπτυξης όσον αφορά την ποσοτική-συσχετιστική τους ικανότητα. Κατόπιν επακολούθησε άσκηση των ατόμων με στόχο τη βελτίωση της συγκεκριμένης ικανότητας. Η άσκηση αξιολογήθηκε ως επιτυχημένη (1) ή αποτυχημένη (2). Μετά την άσκηση τα άτομα κατατάχθηκαν εκ νέου στα τρία επίπεδα γνωστικής ανάπτυξης. Τα τρία επίπεδα γνωστικής ανάπτυξης μετά την άσκηση αποτελούν τις κατηγορίες της εξαρτημένης μεταβλητής Y, ενώ τα αντίστοιχα επίπεδα πριν την άσκηση αποτελούν τις κατηγορίες της πρώτης ανεξάρτητης μεταβλητής X. Οι δύο καταστάσεις επιτυχία ή αποτυχία που αναφέρονται στο αποτέλεσμα της άσκησης συνιστούν τις κατηγορίες της δεύτερης ανεξάρτητης μεταβλητής Z. Τα δεδομένα περιέχονται σε έναν τρισδιάστατο πίνακα συμπτώσεων 3x3x2 (Πίνακας 4).

Ο Πίνακας 5 δίνει τα δεδομένα της δισδιάστατης πρόβλεψης της μεταβλητής Y από την X. Το μοντέλο που χρησιμοποιήθηκε γι' αυτήν την πρόβλεψη προέβλεπε ότι μετά την άσκηση τα άτομα θα πρέπει να βρίσκονται σε επίπεδο ίσο ή ανώτερο από αυτό που βρίσκονταν πριν την άσκηση (βλέπε Πίνακα 5 κυψελών σφάλματος).

Στην τρισδιάστατη πρόβλεψη το προτεινόμενο μοντέλο (Πίνακας 6) υπέθετε ότι τα άτομα που πέτυχαν στην άσκηση, θα έπρεπε μετά την άσκηση να βρίσκονται σε επίπεδο ίσο ή ανώτερο από αυτό που βρίσκονταν πριν την άσκηση (δηλαδή το ίδιο σκεπτικό με αυτό του μοντέλου της δισδιάστατης πρόβλεψης). Όσον αφορά στα άτομα που απέτυχαν στην άσκηση, το τρισδιάστατο μοντέλο υποστήριζε ότι τα άτομα αυτά δε θα πρέπει μετά την άσκηση να έχουν αλλάξει επίπεδο με εξαίρεση τα άτομα του μεσαίου επιπέδου 1, μέρος των οποίων θα μπορούσαν να βρίσκονται μετά την άσκηση και στο προηγούμενο επίπεδο 0.

Για το δισδιάστατο μοντέλο η ΑΠ έδωσε $\nabla(YX)=0.42$ ($p<0.01$). Για το τρισδιάστατο μοντέλο βρέθηκε ότι $\nabla(YXZ)=0.56$ ($p<0.001$). Επίσης $\nabla(YXZIX)=0.32$ ($p<0.01$). Συνεπώς από τα αποτελέσματα φάνηκε ότι ο συνδυασμός των δύο ανεξάρτητων μεταβλητών προβλέπει καλύτερα τις κατηγορίες της εξαρτημένης μεταβλητής από ότι μόνη της η πρώτη ανεξάρτητη μεταβλητή. Επίσης αποδείχτηκε ότι πέρα από την προβλεπτική αξία της πρώτης ανεξάρτητης μεταβλητής, αξιόλογο ρόλο παίζει και η συμμετοχή της δεύτερης ανεξάρτητης μεταβλητής που είναι το αποτέλεσμα της άσκησης.

Πίνακας 4
Κατανομή των ατόμων σύμφωνα με το επίπεδο της γνωστικής τους ανάπτυξης και το αποτέλεσμα της άσκησης

	0 & 1	0 & 2	1 & 1	1 & 2	2 & 1	2 & 2
0	37	58	3	27	0	10
1	4	18	6	13	2	3
2	41	15	24	9	24	44

Σημ.: Οι γραμμές του πίνακα αντιστοιχούν στα επίπεδα γνωστικής ανάπτυξης (0, 1, 2) μετά την άσκηση. Οι στήλες αντιστοιχούν στους συνδυασμούς των γνωστικών επιπέδων πριν την άσκηση (0, 1, 2) και των αποτελεσμάτων (επιτυχία=1, αποτυχία=2) της άσκησης.

Πίνακας 5
Δεδομένα και κυψέλες σφάλματος για την πρόβλεψη των επιπέδων γνωστικής ανάπτυξης μετά την άσκηση από τα επίπεδα γνωστικής ανάπτυξης πριν την άσκηση

Δεδομένα				Κυψέλες σφάλματος			
	0	1	2		0	1	2
0	95	30	10	0	0	1	1
1	22	19	5	1	0	0	1
2	56	33	68	2	0	0	0

Σημ.: Οι γραμμές του πίνακα αντιστοιχούν στα επίπεδα γνωστικής ανάπτυξης (0, 1, 2) μετά την άσκηση ενώ οι στήλες στα επίπεδα γνωστικής ανάπτυξης (0, 1, 2) πριν την άσκηση.

Πίνακας 6
Κυψέλες σφάλματος της τρισδιάστατης πρόβλεψης

	0 & 1	0 & 2	1 & 1	1 & 2	2 & 1	2 & 2
0	0	0	1	0	1	1
1	0	1	0	0	1	1
2	0	1	0	1	0	0

Λογιστική παλινδρόμηση

Η μέθοδος της λογιστικής παλινδρόμησης ανήκει στην ευρύτερη οικογένεια των μεθόδων παλινδρόμησης που αποσκοπούν, ως γνωστό, στη δημιουργία μοντέλων βάσει των οποίων επιχειρείται η πρόβλεψη των τιμών μιας εξαρτημένης μεταβλητής από ένα σύνολο από ανεξάρτητες μεταβλητές. Το χαρακτηριστικό της λογιστικής παλινδρόμησης (ΛΠ) που τη διακρίνει από τις άλλες μεθόδους παλινδρόμησης είναι ότι δίνει την εκτίμηση της πιθανότητας πραγματοποίησης ενός ενδεχομένου Α βάσει ενός συνόλου από ανεξάρτητες μεταβλητές.

Υπενθυμίζεται ότι το μοντέλο της πολλαπλής γραμμικής παλινδρόμησης με k ανεξάρτητες μεταβλητές διατυπώνεται ως εξής:

$$Y = B_0 + B_1X_1 + B_2X_2 + ... + B_kX_k$$

όπου Y είναι η εξαρτημένη μεταβλητή, X_i (i=1...k) είναι οι ανεξάρτητες μεταβλητές, B_i (i=1...k) είναι οι συντελεστές παλινδρόμησης, και B_0 είναι η σταθερά του μοντέλου. Η μέθοδος αυτή εφαρμόζεται γενικά όταν όλες οι μεταβλητές είναι ποσοτικές. Ειδικότερα, μπορεί κάποιες από τις ανεξάρτητες μεταβλητές να είναι κατηγορικές, οπότε και απαιτείται η κωδικοποίησή τους υπό τη μορφή ψευδομεταβλητών.

Ενώ στο μοντέλο της πολλαπλής γραμμικής παλινδρόμησης η σχέση που συνδέει την εξαρτημένη μεταβλητή με τις ανεξάρτητες είναι γραμμική, στην περίπτωση της ΛΠ το μοντέλο είναι μη γραμμικό και συγκεκριμένα εκφράζεται ως εξής (Cox & Snell, 1989):

$$P(A) = \frac{e^{B_0+B_1X_1+B_2X_2+...+B_kX_k}}{1+e^{B_0+B_1X_1+B_2X_2+...+B_kX_k}} \quad (1)$$

όπου P(A) είναι η πιθανότητα του ενδεχομένου Α και e η βάση των φυσικών λογαρίθμων (e =

2.718). Σημειώνεται ότι η πιθανότητα $P(A^0)$ δηλαδή η πιθανότητα του συμπληρωματικού ενδεχομένου του A ισούται με 1 - P(A). Ισοδύναμα το μοντέλο γράφεται και ως εξής (Cox & Snell, 1989):

$$\ln\left(\frac{P(A)}{1-P(A)}\right)=B_0+B_1X_1+B_2X_2+\ldots+B_kX_k \qquad (2)$$

Στην ΛΠ η εξαρτημένη μεταβλητή είναι δυαδικής μορφής όπου η μια κατηγορία αναφέρεται στο ενδεχόμενο A (το οποίο ορίζει ο ερευνητής) ενώ η άλλη κατηγορία στο συμπληρωματικό ενδεχόμενο A^0. Οι ανεξάρτητες μεταβλητές X_1, X_2 ... X_k μπορεί να είναι ποσοτικές, αλλά και κατηγορικές με δύο ή περισσότερες κατηγορίες.

Στην παρούσα εργασία θα αναφερθούμε στην ειδική περίπτωση που οι ανεξάρτητες μεταβλητές είναι κατηγορικές. Όπως στη μέθοδο της πολλαπλής γραμμικής παλινδρόμησης έτσι και στη ΛΠ, ο χειρισμός των κατηγορικών μεταβλητών απαιτεί την κωδικοποίησή τους σε μια σειρά από ψευδομεταβλητές. Υπάρχουν διάφοροι τρόποι κωδικοποίησης, οι οποίοι οδηγούν σε διαφορετική ερμηνεία των αποτελεσμάτων. Ο πλέον χρησιμοποιούμενος τρόπος κωδικοποίησης είναι αυτός που αποσκοπεί στη δυνατότητα σύγκρισης των διάφορων κατηγοριών μιας κατηγορικής μεταβλητής με κάποια κατηγορία της μεταβλητής αυτής, που εμείς επιλέγουμε, και η οποία παίζει το ρόλο της κατηγορίας αναφοράς (Hosmer & Lemeshow, 1989). Συγκεκριμένα, αν C είναι το πλήθος των κατηγοριών μιας κατηγορικής μεταβλητής, για την κωδικοποίησή της δημιουργούνται C-1 ψευδομεταβλητές με τιμές (0, 1). Για την κατηγορία αναφοράς όλες οι ψευδομεταβλητές παίρνουν την τιμή 0. Αντίθετα, για κάθε μια από τις υπόλοιπες C-1 κατηγορίες η αντίστοιχη ψευδομεταβλητή παίρνει την τιμή 1 και οι άλλες ψευδομεταβλητές παίρνουν την τιμή 0. Στον Πίνακα 7 περιγράφεται ο τρόπος δημιουργίας των τριών ψευδομεταβλητών D_1, D_2, D_3 που αντιστοιχούν σε μια κατηγορική μεταβλητή με τέσσερις κατηγορίες A_1, A_2, A_3 και A_4. Η κατηγορία A1 επιλέχτηκε ως η κατηγορία αναφοράς.

Στο παράδειγμα που ακολουθεί θα παρουσιαστούν τα βασικά αποτελέσματα από την εφαρμογή της μεθόδου καθώς και η ερμηνεία των συντελεστών του μοντέλου.

Παράδειγμα

Σε δείγμα 163 ατόμων, από τα οποία 72 ήταν κληρονομικά αριστερόχειρα και 91 κληρονομικά δεξιόχειρα, δόθηκε λεκτικό τεστ που αφορούσε στη συμπλήρωση προτάσεων[1]. Οι επιδόσεις των ατόμων αποτέλεσαν τις τρεις κατηγορίες μιας κατηγορικής μεταβλητής X στην οποία με 1 συμβολίστηκαν οι χαμηλές επιδόσεις, με 2 οι μεσαίες

Πίνακας 7
Κωδικοποίηση μεταβλητής με τέσσερις κατηγορίες

	D_1	D_2	D_3
A_1	0	0	0
A_2	1	0	0
A_3	0	1	0
A_4	0	0	1

Σημ.: Οι ψευδομεταβλητές D_1, D_2 και D_3 εκφράζουν τη σύγκριση των κατηγοριών A_2, A_3 και A_4 αντίστοιχα με την κατηγορία αναφοράς A_1.

1. Τα ερευνητικά, δεδομένα του παραδείγματος προέρχονται από τον καθηγητή Ψυχολογίας κ. Δ. Νατσόπουλο.

και με 3 οι υψηλές επιδόσεις. Η εξαρτημένη μεταβλητή Υ αποτελούνταν από το ενδεχόμενο Α που θεωρήθηκε η δεξιοχειρία (κωδικός 1) και το συμπληρωματικό ενδεχόμενο που ήταν η αριστεροχειρία (κωδικός 0). Η κωδικοποίηση της ανεξάρτητης μεταβλητής έγινε βάσει δύο ψευδομεταβλητών (Πίνακας 8). Ως κατηγορία αναφοράς επιλέχτηκε η κατηγορία των χαμηλών επιδόσεων. Με τον τρόπο αυτό η ψευδομεταβλητή D_1 εκφράζει τη σύγκριση των μεσαίων επιδόσεων με τις χαμηλές, ενώ η D_2 τη σύγκριση των υψηλών επιδόσεων με τις χαμηλές.

Ο Πίνακας 9 δίνει τα αποτελέσματα από την εφαρμογή της ΛΠ.

Στη στήλη Β δίνονται οι τιμές των συντελεστών B_1 και B_2 των δύο ψευδομεταβλητών D_1 και D_2 αντίστοιχα καθώς και η τιμή της σταθεράς του μοντέλου B_0 ενώ στη στήλη S.E. δίνονται τα αντίστοιχα τυπικά σφάλματα των συντελεστών. Στη στήλη με τίτλο Wald δίνονται οι τιμές του στατιστικού του ελέγχου Wald (Hosmer & Lemeshow, 1989) βάσει του οποίου ελέγχεται η σημαντικότητα των διάφορων συντελεστών του μοντέλου (δηλαδή αν αυτοί διαφέρουν από το 0). Το στατιστικό αυτό ακολουθεί, για μια μεταβλητή με C κατηγορίες, κατανομή x - τετράγωνο με C-1 βαθμούς ελευθερίας. Συνεπώς στις διάφορες ψευδομεταβλητές αντιστοιχεί 1 βαθμός ελευθερίας. Οι βαθμοί ελευθερίας εμφανίζονται στη στήλη D.F. Τέλος, στην τελευταία στήλη παρουσιάζεται η στατιστική σημαντικότητα των συντελεστών. Παρατηρούμε ότι μόνο ο συντελεστής B_2 είναι στατιστικά σημαντικός, που σημαίνει (λόγω και του θετικού προσήμου που έχει) ότι τα άτομα με τις υψηλές επιδόσεις συγκρινόμενα με τα άτομα των χαμηλών επιδόσεων έχουν μεγαλύτερη πιθανότητα να είναι δεξιόχειρα.

Ας υπολογίσουμε τώρα τις πιθανότητες του να είναι δεξιόχειρο κάποιο άτομο με υψηλές επιδόσεις όπως και κάποιο άτομο με χαμηλές επιδόσεις.

Το μοντέλο της ΛΠ (βλέπε τύπο 1) για το παράδειγμα μας είναι το εξής:

$$P(A)=\frac{e^{B_0+B_1D_1+B_2D_2}}{1+e^{B_0+B_1D_1+B_2D_2}}$$

Ας θεωρήσουμε ένα άτομο με υψηλές επιδόσεις. Για το άτομο αυτό θα είναι D_1: 0, D_2: 1 οπότε αντικαθιστώντας τις τιμές των B_0 και B_2 έχουμε ότι η πιθανότητα να είναι δεξιόχειρο είναι

Πίνακας 8
Κωδικοποίηση των επιδόσεων βάσει δύο ψευδομεταβλητών

Επιδόσεις	D_1	D_2
Χαμηλές (X_1)	0	0
Μεσαίες (X_2)	1	0
Υψηλές (X_3)	0	1

Πίνακας 9
Αποτελέσματα της λογιστικής παλινδρόμησης

Επιδόσεις	B	S.E.	Wald	D.F.	p
Μεσαίες - Χαμηλές (D_1)	0.6082	0.3904	2.4269	1	μ.σ.
Υψηλές - Χαμηλές (D_2)	1.8593	0.4507	17.0198	1	**
Σταθερά του μοντέλου	-0.5436	0.2963	3.3653	1	μ.σ.

** : $p < .01$, μ.σ. : μη στατιστικά σημαντικό.

$$P(A)=\frac{e^{B_0+B_1D_1+B_2D_2}}{1+e^{B_0+B_1D_1+B_2D_2}}=$$
$$=\frac{e^{-0.5436+1.8593D_2}}{1+e^{-0.5436+1.8593D_2}}=0.7884$$

Έστω τώρα άτομο χαμηλών επιδόσεων. Για το άτομο αυτό είναι D_1: 0, D_2: 0, οπότε η αντίστοιχη πιθανότητα είναι

$$P(A)=\frac{e^{B_0+B_1D_1+B_2D_2}}{1+e^{B_0+B_1D_1+B_2D_2}}=$$
$$=\frac{e^{-0.5436}}{1+e^{-0.5436}}=0.3673$$

Παρατηρούμε ότι η πιθανότητα να είναι δεξιόχειρο ένα άτομο με υψηλές επιδόσεις είναι σαφώς μεγαλύτερη από την πιθανότητα ενός ατόμου με χαμηλές επιδόσεις.

Χρησιμοποιώντας τώρα την έκφραση (2) του μοντέλου της ΛΠ έχουμε για το παράδειγμά μας την εξής διατύπωση:

$$\ln\left(\frac{P(A)}{1-P(A)}\right)=B_0+B_1D_1+B_2D_2$$

Η ποσότητα αριστερά του ίσον (δηλαδή ο λογάριθμος του λόγου P(A)/(1-P(A)) ονομάζεται logit. Στην περίπτωση ατόμου με υψηλές επιδόσεις το logit ισούται με

$$\ln\left(\frac{P(A)}{1-P(A)}\right)=1.3153 \quad \text{όπου} \quad P(A)=0.7884$$

ενώ στην περίπτωση ατόμου με χαμηλές επιδόσεις το logit είναι

$$\ln\left(\frac{P(A)}{1-P(A)}\right)=-0.5438 \quad \text{όπου} \quad P(A)=0.3673$$

Η μεταβολή στα logit όταν συγκρίνουμε τις υψηλές επιδόσεις με τις χαμηλές ισούται με: 1.3153-(-0.5438)=1.859.

Παρατηρούμε ότι η τιμή 1.859 δεν είναι άλλη από την τιμή του συντελεστή B_2 της ψευδομεταβλητής D_2 που εκφράζει τη σύγκριση μεταξύ υψηλών και χαμηλών επιδόσεων (βλέπε Πίνακα 9). Ενώ λοιπόν στη γραμμική παλινδρόμηση οι συντελεστές του μοντέλου εκφράζουν, ως γνωστό, το βαθμό μεταβολής της εξαρτημένης μεταβλητής όταν η ανεξάρτητη μεταβάλλεται (αυξάνεται ή μειώνεται) κατά μια μονάδα, στη ΛΠ οι συντελεστές εκφράζουν τη μεταβολή που επέρχεται στα logit όταν κάποια κατηγορία συγκρίνεται με την κατηγορία αναφοράς.

Ας δούμε τώρα τα αποτελέσματα από μια ακόμα έρευνα που προέρχεται από το χώρο της κοινωνικής ψυχολογίας (Dikaiou & Kiosseoglou, 1993).

Παράδειγμα

Η έρευνα πραγματοποιήθηκε σε δύο ομάδες εφήβων ηλικίας 13-15 χρόνων. Η πρώτη ομάδα αποτελούνταν από 100 άτομα (50 αγόρια, 50 κορίτσια) και προέρχονταν από τη μειονότητα των Τσιγγάνων. Η δεύτερη ομάδα αποτελούνταν από 100 άτομα (50 αγόρια, 50 κορίτσια) που ήταν μη Τσιγγάνοι Έλληνες.

Ζητήθηκε από τους ερωτώμενους να περιγράψουν 3 σημαντικά για αυτούς προβλήματα καθώς και τους τρόπους αντιμετώπισης που εφαρμόζουν για κάθε ένα από αυτά. Από τις απαντήσεις δημιουργήθηκαν 5 κατηγορίες προβλημάτων και 5 στρατηγικές αντιμετώπισης σύμφωνα με την κωδικοποίηση της J. Gibson (1996):

Τύποι προβλημάτων

1. Φτώχεια.
2. Οικογενειακά θέματα.
3. Σχολικά θέματα.
4. Προσωπική ταυτότητα/αντίληψη του εαυτού.
5. Θέματα κοινωνικοποίησης εκτός σχολείου και οικογένειας.

Στρατηγικές αντιμετώπισης

1. Αναζήτηση βοήθειας και/ή κοινωνικής στήριξης.

2. Διαπροσωπικές στρατηγικές αντιμετώπισης του προβλήματος.
3. Ατομική επίλυση του προβλήματος.
4. Διαχείριση άγχους.
5. Παραίτηση.

Συνολικά μελετήθηκαν 6 ανεξάρτητες μεταβλητές (με 5 κατηγορίες η κάθε μια), από τις οποίες οι 3 αφορούσαν τους τύπους των προβλημάτων και οι άλλες 3 αναφέρονταν στις στρατηγικές αντιμετώπισής τους. Κάθε μεταβλητή εκφράστηκε από ένα σύστημα 4 ψευδομεταβλητών. Συνεπώς στη ΛΠ πήραν μέρος συνολικά 24 ψευδομεταβλητές. Η εξαρτημένη μεταβλητή ήταν δυαδική όπου ως ενδεχόμενο Α θεωρήθηκε αυθαίρετα η κατηγορία όχι Τσιγγάνος (κωδικός 1), οπότε το συμπληρωματικό ενδεχόμενο αποτέλεσε η κατηγορία Τσιγγάνος (κωδικός 0). Οι Πίνακες 10 και 11 δίνουν τον τρόπο κατασκευής των ψευδομεταβλητών. Στα προβλήματα, ως κατηγορία αναφοράς θεωρήθηκε η κατηγορία φτώχεια ενώ στις στρατηγικές αντιμετώπισης η κατηγορία παραίτηση, για λόγους που σχετίζονταν με τις υποθέσεις της έρευνας.

Στα δεδομένα, εφαρμόστηκε βήμα προς βήμα ΛΠ της οποίας τα αποτελέσματα συνοψίζονται στον Πίνακα 12.

Από τα αποτελέσματα φαίνεται ότι σημαντικότερο ρόλο παίζουν τα προβλήματα παρά οι

Πίνακας 10
Κατασκευή του συστήματος των ψευδομεταβλητών που αντιστοιχούν στους τύπους των προβλημάτων

Τύποι προβλημάτων	D_1	D_2	D_3	D_4
Φτώχεια	0	0	0	0
Οικογενειακά θέματα	1	0	0	0
Σχολικά θέματα	0	1	0	0
Προσωπική ταυτότητα	0	0	1	0
Κοινωνικοποίηση εκτός σχολείου και οικογένειας	0	0	0	1

Σημ.: Οι ψευδομεταβλητές D_1, D_2, D_3 και D_4 δίνουν τη σύγκριση των διάφορων κατηγοριών προβλημάτων με την κατηγορία Φτώχεια. D_1: Οικογενειακά θέματα - Φτώχεια, D_2: Σχολικά θέματα - Φτώχεια, D_3: Προσωπική ταυτότητα - Φτώχεια, D_4: Κοινωνικοποίηση εκτός σχολείου και οικογένειας - Φτώχεια.

Πίνακας 11
Κατασκευή του συστήματος των ψευδομεταβλητών που αντιστοιχούν στις στρατηγικές αντιμετώπισης

Στρατηγικές αντιμετώπισης	D_1	D_2	D_3	D_4
Παραίτηση	0	0	0	0
Αναζήτηση βοήθειας	1	0	0	0
Διαπροσωπικές στρατηγικές	0	1	0	0
Ατομική επίλυση προβλήματος	0	0	1	0
Διαχείριση άγχους	0	0	0	1

Σημ.: Οι ψευδομεταβλητές D_1, D_2, D_3 και D_4 δίνουν τη σύγκριση των διαφόρων στρατηγικών αντιμετώπισης με την κατηγορία Παραίτηση. D_1: Αναζήτηση βοήθειας και/ή κοινωνικής στήριξης - Παραίτηση, D_2: Διαπροσωπικές στρατηγικές αντιμετώπισης - Παραίτηση, D_3: Ατομική επίλυση προβλήματος - Παραίτηση, D_4: Διαχείριση άγχους - Παραίτηση.

στρατηγικές αντιμετώπισής τους λόγω του ότι στο τελικό μοντέλο (που δημιουργήθηκε με τη διαδικασία επιλογής βήμα προς βήμα) συμπεριλήφθηκαν και τα τρία προβλήματα ενώ οι στρατηγικές αντιμετώπισης αφορούσαν μόνο το τρίτο κατά σειρά πρόβλημα. Στα δύο πρώτα προβλήματα παρατηρούμε ότι υπάρχει η ίδια εικόνα ως προς τη στατιστική σημαντικότητα των διάφορων ψευδομεταβλητών. Τα αποτελέσματα έδειξαν ότι τα σχολικά προβλήματα καθώς και αυτά της προσωπικής ταυτότητας συνδέονται περισσότερο με τους μη Τσιγγάνους σε αντίθεση με τους Τσιγγάνους που φαίνεται να τους απασχολεί το πρόβλημα της φτώχειας. Στο τρίτο πρόβλημα, το θέμα της κοινωνικοποίησης εκτός σχολείου και οικογένειας είναι αυτό που διακρίνει τους μη Τσιγγάνους. Όσον αφορά στις στρατηγικές αντιμετώπισης του τρίτου προβλήματος, οι

Πίνακας 12

Αποτελέσματα της λογιστικής παλινδρόμησης στους τύπους προβλημάτων και στις στρατηγικές αντιμετώπισης

Τύποι Προβλημάτων	B	Wald	D.F.	p
Πρόβλημα 1				
Οικογενειακά - Φτώχεια	0.216	0.053	1	μ.σ.
Σχολικά - Φτώχεια	2.146	6.089	1	*
Προσωπική ταυτότητα - Φτώχεια	2.506	7.380	1	**
Κοινωνικοποίηση - Φτώχεια	1.328	1.533	1	μ.σ.
Πρόβλημα 2				
Οικογενειακά - Φτώχεια	0.031	0.002	1	μ.σ.
Σχολικά - Φτώχεια	1.872	4.090	1	*
Προσωπική ταυτότητα - Φτώχεια	2.964	7.590	1	**
Κοινωνικοποίηση - Φτώχεια	0.818	0.739	1	μ.σ.
Πρόβλημα 3				
Οικογενειακά - Φτώχεια	-0.542	0.484	1	μ.σ.
Σχολικά - Φτώχεια	0.234	0.065	1	μ.σ.
Προσωπική ταυτότητα - Φτώχεια	1.481	2.208	1	μ.σ.
Κοινωνικοποίηση - Φτώχεια	-2.391	7.859	1	**
Στρατηγικές Αντιμετώπισης				
Πρόβλημα 3				
Βοήθεια/κοινωνική στήριξη - Παραίτηση	2.888	6.044	1	*
Διαπροσωπικές στρατηγικές - Παραίτηση	0.184	0.025	1	μ.σ.
Ατομική επίλυση - Παραίτηση	2.319	6.786	1	**
Διαχείριση άγχους - Παραίτηση	2.217	4.829	1	*

* : $p<.05$, ** : $p<.01$, μ.σ. : μη στατιστικά σημαντικό.

Τσιγγάνοι χαρακτηρίζονται από την επιλογή της παραίτησης ενώ οι μη Τσιγγάνοι χρησιμοποιούν άλλες στρατηγικές όπως την αναζήτηση βοήθειας/κοινωνικής στήριξης, την ατομική επίλυση του προβλήματος και τη διαχείριση του άγχους.

Δείκτες καλής προσαρμογής του μοντέλου

Δύο κυρίως τρόποι χρησιμοποιούνται για τον έλεγχο της καλής προσαρμογής του μοντέλου της ΛΠ. Ο πρώτος τρόπος αφορά το ποσοστό της σωστής ταξινόμησης των ατόμων του δείγματος με βάση τις ανεξάρτητες μεταβλητές που συμπεριλήφθηκαν στο τελικό μοντέλο. Έτσι αν με βάση τον συνδυασμό των ανεξάρτητων μεταβλητών του μοντέλου, η πιθανότητα του να ανήκει κάποιο άτομο στο ενδεχόμενο Α όχι Τσιγγάνος είναι μεγαλύτερη του 0.5, το άτομο κατατάσσεται στην κατηγορία όχι Τσιγγάνος, ενώ αν η πιθανότητα αυτή είναι μικρότερη του 0.5, κατατάσσεται στην κατηγορία Τσιγγάνος που αποτελεί το συμπληρωματικό ενδεχόμενο. Στο παραπάνω παράδειγμα το 91% των Τσιγγάνων και το 90% των μη Τσιγγάνων ταξινομήθηκαν σωστά μέσω του συγκεκριμένου μοντέλου που σημαίνει ότι το μοντέλο προσαρμόζεται στα δεδομένα σε πολύ ικανοποιητικό βαθμό.

Ο δεύτερος τρόπος που χρησιμοποιείται για την εξακρίβωση του βαθμού της προσαρμογής αφορά τον υπολογισμό του στατιστικού που ονομάζεται *απόκλιση* (Weisberg, 1985). Στην περίπτωση ενός ιδανικού μοντέλου (με τέλεια προσαρμογή) η τιμή του στατιστικού αυτού είναι 0. Το στατιστικό αυτό ακολουθεί προσεγγιστικά χ - τετράγωνο κατανομή με n-r βαθμούς ελευθερίας όπου n είναι το μέγεθος του δείγματος και r είναι το πλήθος των υπό εκτίμηση παραμέτρων του μοντέλου. Αν συνεπώς η τιμή του στατιστικού αποκλίνει σημαντικά από το 0, το μοντέλο δεν προσαρμόζεται ικανοποιητικά. Στο παραπάνω παράδειγμα είναι n = 200 και r = 17 επειδή τόσος είναι ο αριθμός των συντελεστών Β (συμπεριλαμβανομένης και της σταθεράς B_0). Για n-r = 183 βαθμούς ελευθερίας η τιμή του στατιστικού είναι 125.8 και η αντίστοιχη πιθανότητα p = 0.99. Συνεπώς, επειδή ο έλεγχος δεν είναι στατιστικά σημαντικός δεν μπορούμε να απορρίψουμε την υπόθεση της καλής προσαρμογής του μοντέλου.

Βιβλιογραφία

Cox, D. R., & Snell, E. J. (1989). *Analysis of binary data* (p. 26). London: Chapman & Hall.

Dikaiou, M., & Kiosseoglou, G. (1993). Identified problems and coping strategies: Gypsy minority versus non-minority adolescents. *International Migration, XXXI,* 473-495.

Froman, T., & Hubert, L .J. (1980). Application of prediction analysis to developmental priority. *Psychological Bulletin, 87,* 136-146.

Gibson, J. (Ed.)(1996). *Adolescence from crisis to coping. A thirteen nation study.* Oxford: Butterworth-Heinnemann.

Hildebrand, D. K., Laing, J. D., & Rosenthal, H. (1977). *Prediction analysis of cross-classifications.* New York: Wiley.

Hosmer, D., & Lemeshow, S. (1989). *Applied logistic regression* (pp. 16-17, 47-50). New York: Wiley.

Κιοσέογλου, Γ., & Δημητρίου, Α. (1992). Ανάλυση προβλέψεων: Στατιστική μεθοδολογία για την αξιολόγηση μοντέλων πρόβλεψης με ποιοτικά δεδομένα στις κοινωνικές επιστήμες. *Σπουδαί, 42,* 44-66.

Weisberg, S. (1985). *Applied linear regression* (p. 270). New York: Wiley.

Prediction models with categorical data in psychological research via the statistical methods of prediction analysis and logistic regression

GRIGORIS KIOSSEOGLOU
Aristotle University of Thessaloniki, Greece

ABSTRACT The statistical methods used in the creation of prediction models are applied in a wide range of domains of the empirical scientific research to which the behavioral sciences belong. A special case are those methods dealing with categorical data which in spite of being very interesting they are less known due to the fact that they are relatively recent. The statistical methods of Prediction Analysis and Logistic Regression belong to this case. Although these methods, from the point of view of the statistical technique used, are completely different, they have a common feature that of permitting the prediction of the states of a dependent categorical variable from one or more independent categorical variables. In this paper the basic principles of these two methods are presented. Examples from research data from developmental psychology, psychology of language and social psychology for a better understanding of the methods and successful applications by researchers from behavioral sciences are analyzed.

Key words: Logistic regression, prediction analysis.

Address: Grigoris Kiosseoglou, School of Psychology, Aristotle University of Thessaloniki, 540 06 Thessaloniki, Greece. E-mail: kios@psy.auth.gr