ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΣΧΟΛΗ ΕΠΙΣΤΗΜΩΝ ΥΓΕΙΑΣ

ΤΜΗΜΑ ΙΑΤΡΙΚΗΣ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

«ΜΕΘΟΔΟΛΟΓΙΑ ΒΙΟΙΑΤΡΙΚΗΣ ΕΡΕΥΝΑΣ,ΒΙΟΣΤΑΤΙΣΤΙΚΗ ΚΑΙ ΚΛΙΝΙΚΗ ΒΙΟΠΛΗΡΟΦΟΡΙΚΗ»

Μεταπτυχιακή Διπλωματική Εργασία

"Η Λογιστική Παλινδρόμηση ως μέθοδος της Διαχωριστικής Ανάλυσης"

"The use of Logistic Regression for Discriminant Analysis"

Επιβλέποντες:

***Απόστολος Μπατσίδης**, Επίκουρος Καθηγητής, Τμήμα Μαθηματικών, Πανεπιστήμιο Ιωαννίνων.*
***Χρήστος Χατζηχριστοδούλου**, Καθηγητής Υγιεινής και Επιδημιολογίας, Ιατρική Σχολή Λάρισας, Πανεπιστήμιο Θεσσαλίας*
***Γεώργιος Ραχιώτης**, Επίκουρος Καθηγητής Επιδημιολογίας και Επαγγελματικής Υγιεινής, Ιατρική Σχολή Λάρισας, Πανεπιστήμιο Θεσσαλίας*

Ευαγγελία Αστάρα

Λάρισα 2017

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 06:38:11 EEST - 18.216.131.50

ΠΕΡΙΕΧΟΜΕΝΑ

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 06:38:11 EEST - 18.216.131.50

Κεφάλαιο 1: ***Περίληψη***

Αντικείμενο αυτής της διπλωματικής διατριβής είναι δύο μέθοδοι, που σκοπός τους, μεταξύ άλλων, είναι να κατατάσσουν τις πειραματικές μονάδες σε γνωστές ομάδες βάσει των διαθέσιμων μετρήσεων σε κάποια χαρακτηριστικά γνωρίσματα (μεταβλητές). Αυτές οι μέθοδοι είναι η λογιστική παλινδρόμηση (logistic regression) και η διαχωριστική ή διαφοροποιούσα ανάλυση (discriminant analysis). Απώτερος στόχος της διπλωματικής διατριβής είναι να αναδείξει πώς η λογιστική παλινδρόμηση μπορεί να χρησιμοποιηθεί ως μέθοδος διαχωριστικής ανάλυσης. Στο πλαίσιο αυτό, στο δεύτερο κεφάλαιο («Λογιστική Παλινδρόμηση») το μοντέλο της λογιστικής παλινδρόμησης περιγράφεται εν συντομία, ενώ περιληπτικά αναφέρεται το αντικείμενο της Διαχωριστικής Ανάλυσης στο τρίτο κεφάλαιο («Διαχωριστική Ανάλυση»). Στο τέταρτο κεφάλαιο περιγράφονται οι ομοιότητες και οι διαφορές των δύο μεθόδων. Στο πέμπτο κεφάλαιο («Εφαρμογές») παρατίθενται τρεις εφαρμογές σε γνωστά σύνολα δεδομένων. Ειδικότερα στα τρία αυτά σύνολα δεδομένων η λογιστική παλινδρόμηση χρησιμοποιείται ως μέθοδος της διαχωριστικής ανάλυσης. Η υλοποίηση αυτή πραγματοποιείται με τη βοήθεια του στατιστικού πακέτου IBM SPSS. Τέλος, η διπλωματική διατριβή ολοκληρώνεται με την παράθεση χρήσιμων βιβλιογραφικών αναφορών.

Κεφάλαιο 2: ***Λογιστική Παλινδρόμηση***

Η Λογιστική Παλινδρόμηση (**logistic regression**, ή **logit regression**, ή **logit model)**, η οποία πρωτοπαρουσιάσθηκε στη στατιστική βιβλιογραφία από τον D. Cox (1958), είναι ένα μοντέλο παλινδρόμησης στο οποίο η εξαρτημένη μεταβλητή είναι κατηγορική, ενώ οι ανεξάρτητες μεταβλητές μπορεί να είναι είτε ποσοτικές συνεχείς, είτε κατηγορικές. Στο σημείο αυτό θα πρέπει να επισημανθεί ότι μια κατηγορική ανεξάρτητη μεταβλητή υπεισέρχεται στο μοντέλο της λογιστικής παλινδρόμησης με τη βοήθεια κ-1 το πλήθος δείκτριων μεταβλητών (indicators ή dummy), όπου με κ συμβολίζεται το πλήθος των επιπέδων της κατηγορικής ανεξάρτητης μεταβλητής. Δείκτρια είναι κάθε δίτιμη μεταβλητή με τιμές 1 και 0.

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 06:38:11 EEST - 18.216.131.50

Στόχος λοιπόν της Λογιστικής Παλινδρόμησης είναι η δημιουργία ενός μοντέλου πρόβλεψης των τιμών της υπό μελέτης κατηγορικής εξαρτημένης μεταβλητής χρησιμοποιώντας κάποιες ποσοτικές και ποιοτικές ανεξάρτητες μεταβλητές. Είναι εύκολα αντιληπτό λόγω αντικειμένου της Λογιστικής Παλινδρόμησης ότι βρίσκει εφαρμογή σε πλήθος επιστημονικών πεδίων. Για παράδειγμα στο χώρο της υγείας για την πρόβλεψη αν ένα έμβρυο θα γεννηθεί με βάρος λιγότερο ή περισσότερο από 2.5 κιλά, στο χώρο του Marketing για την πρόβλεψη αν ένας καταναλωτής προβεί στην αγορά ή όχι κάποιων προϊόντων, στο χώρο της παιδείας για την πρόβλεψη αν ένας μαθητής επιτύχει ή όχι στις εξετάσεις ενός μαθήματος.

Διακρίνονται τρεις τύποι λογιστικής παλινδρόμησης ανάλογα με την ιδιαίτερη φύση της εξαρτημένης κατηγορικής μεταβλητής, η Binary (Δίτιμη ή διχοτομική), η Multinomial (πολυωνυμική ή πολυχοτομική) και η Ordinal (Διατάξιμη). Ειδικότερα, στην Binary Logistic Regression, η εξαρτημένη κατηγορική μεταβλητή συνίσταται από δύο κατηγορίες, όπως π.χ. είναι οι εκβάσεις επιτυχία/αποτυχία, ΝΑΙ/ΟΧΙ, γεγονός απόν/παρόν. Στην Multinomial Logistic Regression η εξαρτημένη μεταβλητή έχει τρεις ή περισσότερες κατηγορίες, οι οποίες δεν έχουν κάποια φυσική διαβάθμιση. Για παράδειγμα όταν η εξαρτημένη μεταβλητή είναι ο χαρακτηρισμός του χρώματος αντικειμένων ως ερυθρού, πράσινου, κίτρινου κτλ. Τέλος, στην Ordinal Logistic Regression η εξαρτημένη μεταβλητή συνίσταται από δύο ή περισσότερες κατηγορίες μεταξύ των οποίων ισχύει η έννοια της ανισότητας, όπως π.χ. σε μια ερώτηση συμφωνίας/διαφωνίας με κλίμακα καθόλου, λίγο, μέτρια, αρκετά, πολύ ή στην κατάταξη ενός στρώματος υλικού ως λεπτού, μεσαίου, παχέος. Σε όσα ακολουθούν, χωρίς βλάβη της γενικότητας, το ενδιαφέρον περιορίζεται στην περίπτωση της Binary Logistic Regression, δηλαδή σε περιπτώσεις δίτιμων εξαρτημένων τυχαίων μεταβλητών, με την τιμή 1 της εξαρτημένης μεταβλητής να αντιστοιχεί σε "επιτυχία" και την τιμή 0 σε "αποτυχία".

Από τα παραπάνω γίνεται αντιληπτό ότι η Λογιστική Παλινδρόμηση έχει σκοπό παρόμοιο με εκείνον της κλασικής Γραμμικής Παλινδρόμησης με τη διαφοροποίηση ότι στην περίπτωση της Λογιστικής Παλινδρόμησης η εξαρτημένη

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 06:38:11 EEST - 18.216.131.50

μεταβλητή είναι κατηγορική και όχι ποσοτική. Επισημαίνεται ότι σε τέτοιες περιπτώσεις η υιοθέτηση ενός μοντέλου κλασικής γραμμικής παλινδρόμησης είναι εσφαλμένη, καθώς δεν πληρούνται βασικές υποθέσεις για τα σφάλματα (ομοσκεδαστικότητα, κανονικότητα) καθώς επίσης παραβιάζεται και η υπόθεση της γραμμικότητας-ορθότητας του μοντέλου. Για να ξεπεραστούν αυτά τα προβλήματα, στην περίπτωση των δίτιμων εξαρτημένων κατηγορικών μεταβλητών, το πιο διαδεδομένο μοντέλο Binary Logistic Regression υιοθετεί τον λεγόμενο λογιστικό μετασχηματισμό (logit), ο οποίος ορίζεται ως:

$$\ln(odds) = \beta_0 + \beta_1 X_{1i} + ... + \beta_m X_{mi}$$

ή ισοδύναμα

$$P(Y_i = 1/X_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X_{1i} + ... + \beta_m X_{mi}$$

όπου $\beta_0 + \beta_1 X_{1i} + ... + \beta_m X_{mi}$ είναι ένας γραμμικός συνδυασμός των m το πλήθος ανεξάρτητων μεταβλητών $X_{1,}..., X_m$ που συμμετέχουν στο μοντέλο της λογιστικής παλινδρόμησης, ενώ

$$\text{odds} = p/(1-p),$$

με p να συμβολίζεται η πιθανότητα να συμβεί το γεγονός που έχει ορισθεί ως επιτυχία (Y=1) του πειράματος, με Y να συμβολίζει την κατηγορική δίτιμη εξαρτημένη μεταβλητή. Τα odds ουσιαστικά παριστάνουν τη σχετική συχνότητα με την οποία διαφορετικά ενδεχόμενα πραγματοποιούνται, ενώ οι συντελεστές ερμηνεύονται ως η αλλαγή στο Log odds για μοναδιαία αύξηση στην τιμή της ανεξάρτητης μεταβλητής, όταν οι υπόλοιπες ανεξάρτητες μεταβλητές παραμένουν σταθερές.

Από την παραπάνω σχέση εύκολα προκύπτει με αλγεβρικές πράξεις ότι

$$p_i = \frac{\exp(\beta_0 + \beta_1 X_{1i} + ... + \beta_m X_{mi})}{1 + \exp(\beta_0 + \beta_1 X_{1i} + ... + \beta_m X_{mi})} = \frac{1}{1 + \exp(-(\beta_0 + \beta_1 X_{1i} + ... + \beta_m X_{mi}))}$$

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 06:38:11 EEST - 18.216.131.50

Επομένως για κάθε πειραματική μονάδα μπορεί να προκύψει μια (προβλεπτική) πιθανότητα πραγματοποίησης του γεγονότος που έχει ορισθεί ως επιτυχία. Αν η πιθανότητα αυτή ξεπερνά το 50% τότε η πειραματική μονάδα κατατάσσεται στο γκρουπ με τιμή στην εξαρτημένη μεταβλητή=1 (επιτυχία), ενώ διαφορετικά στο άλλο γκρουπ.

Χρησιμοποιώντας την παραπάνω χαρακτηριστική ιδιότητα προκύπτει ότι κάποιος μπορεί να χρησιμοποιήσει τη λογιστική παλινδρόμηση ως μέθοδο Διαχωριστικής ή Διαφοροποιούσας Ανάλυσης, ως μια μέθοδο δηλαδή για να κατατάσσει μία νέα πειραματική μονάδα σε ένα από τα δύο γκρουπ (τιμές της εξαρτημένης μεταβλητής Υ) με βάση τις τιμές σε ένα πλήθος χαρακτηριστικών, όπως αυτά δηλώνονται στις ανεξάρτητες μεταβλητές. Αυτό απαιτεί τον υπολογισμό για τη νέα πειραματική μονάδα της προβλεπόμενης πιθανότητας, η οποία θα προκύπτει από το μοντέλο της Λογιστικής Παλινδρόμησης. Για να υπολογιστούν αυτές οι πιθανότητες απαραίτητη προϋπόθεση είναι η εκτίμηση των άγνωστων παραμέτρων, συντελεστών (β_0,.... β_m). Οι συντελεστές των ανεξάρτητων μεταβλητών στην εξίσωση της παλινδρόμησης εκτιμώνται βάσει της μεθόδου της Μέγιστης Πιθανοφάνειας. Σε αντίθεση με την κλασική γραμμική παλινδρόμηση, όπου υπό την υπόθεση της κανονικότητας των σφαλμάτων, οι Εκτιμητές Μέγιστης Πιθανοφάνειας (ΕΜΠ) προσδιορίζονται σε κλειστή αναλυτική μορφή, στο μοντέλο της Λογιστικής Παλινδρόμησης οι ΕΜΠ δεν προσδιορίζονται σε κλειστή μορφή και χρησιμοποιούνται αριθμητικές μέθοδοι, όπως ο αλγόριθμος Newton (Hosmer and Lemeshow, 2000).

Οι υποθέσεις της Λογιστικής Παλινδρόμησης είναι οι ακόλουθες (Tabachnick and Fidell, 2012):

Υπόθεση 1 LR. Δεν πρέπει να υπάρχουν ακραίες τιμές στα δεδομένα. Η υπόθεση αυτή μπορεί να ικανοποιηθεί με τη μετατροπή των συνεχών μεταβλητών σε τυποποιημένες τιμές και την απομάκρυνση των πειραματικών μονάδων με τιμές μικρότερες από -3,29 ή μεγαλύτερες από 3,29.

Υπόθεση 2 LR. Δεν πρέπει να υπάρχει υψηλή συσχέτιση μεταξύ των ανεξάρτητων μεταβλητών (συγγραμμικότητα). Κάτι τέτοιο μπορεί να ελεγχθεί είτε από τον πίνακα συσχετίσεων είτε με την εφαρμογή διαγνωστικών ελέγχων.

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 06:38:11 EEST - 18.216.131.50

Υπόθεση 3 LR. Πρέπει να υπάρχει γραμμική σχέση μεταξύ του λογαρίθμου των Odds και των ανεξάρτητων μεταβλητών.

Παρατηρήσεις

1. Η Λογιστική Παλινδρόμηση, για την σωστή εφαρμογή της απαιτεί μεγάλο δείγμα, προκειμένου να παράγει αξιόπιστο αποτέλεσμα. Ένας εμπειρικός κανόνας αναφέρει ότι το δείγμα θα πρέπει να είναι 30 φορές μεγαλύτερο από τον αριθμό των παραμέτρων που εκτιμά το μοντέλο

2. Τα βήματα κατασκευής ενός μοντέλου Λογιστικής Παλινδρόμησης είναι ανάλογα αυτών της γραμμικής παλινδρόμησης. Αρχικά προσδιορίζεται η κατηγορική εξαρτημένη μεταβλητή και το σύνολο των ανεξάρτητων μεταβλητών που θα συμμετέχουν στην παλινδρόμηση. Έπειτα, διερευνούμε τα δεδομένα για ύπαρξη ασυνήθιστων πειραματικών μονάδων (ύπαρξη ακραίων τιμών), ελέγχουμε την ικανοποίηση των υποθέσεων της Λογιστικής Παλινδρόμησης και διερευνούμε αν υπάρχει κάποια συγκεκριμένη τιμή η οποία επηρεάζει υπερβολικά τα αποτελέσματα. Τέλος, ερμηνεύουμε τα αποτελέσματα που αφορούν την επίδραση κάθε ανεξάρτητης μεταβλητής στο μοντέλο. Θα ήταν παράλειψη να μην αναφερθεί ότι στο εξεταζόμενο μοντέλο παλινδρόμησης δύναται να εφαρμοστούν τεχνικές βέλτιστης επιλογής ανεξάρτητων μεταβλητών προς ένταξη και διαγνωστικά κριτήρια εγκυρότητας και αξιοπιστίας του μοντέλου (Δημητροπουλάκης, 2017)

Κεφάλαιο 3: *Διαχωριστική Ανάλυση*

Η Διαχωριστική ή Διαφοροποιούσα Ανάλυση (Discriminant Analysis) είναι κλάδος της Πολυμεταβλητής Ανάλυσης, που αναπτύχθηκε από τον R. A. Fisher (1936) ο οποίος εισήγαγε την έννοια της γραμμικής διαχωριστικής συνάρτησης σε μία από τις κλασικές εφαρμογές της διαχωριστικής ανάλυσης, για την ταξινόμηση λουλουδιών. Η μέθοδος της διαχωριστικής ανάλυσης έχει διττό στόχο. Ο ένας της

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 06:38:11 EEST - 18.216.131.50

στόχος είναι η διάκριση ενός πληθυσμού σε ευδιάκριτα σύνολα (ομάδες – υποπληθυσμούς) και ο απώτερος στόχος της είναι να κατατάξει, να ταξινομήσει μια νέα παρατήρηση, μια νέα πειραματική μονάδα σε έναν από δύο ή περισσότερους εκ των προτέρων πλήρως προσδιορισμένους πληθυσμούς. Παραδείγματα προβλημάτων τέτοιου τύπου είναι η ταξινόμηση ενός εμβρύου στην ομάδα των βρεφών που θα γεννηθούν με φυσιολογικό βάρος ή στην ομάδα αυτών που θα γεννηθούν ολιγοβαρή, η ταξινόμηση ενός υποψήφιου μεταπτυχιακού φοιτητή σε αυτούς που θα ολοκληρώσουν με επιτυχία τις μεταπτυχιακές σπουδές ή σε αυτούς που θα αποτύχουν. Ένα άλλο παράδειγμα αποτελεί ο διαχωρισμός των πρωτεϊνών σε κατηγορίες με βάση τις γεωμετρικές τους αποστάσεις, διαχωρισμός που επιπρόσθετα θα βοηθήσει στην εύρεση των χαρακτηριστικών που θα μπορέσουν να ομαδοποιήσουν και οποιαδήποτε άλλη πρωτεΐνη. Είναι προφανές ότι η Διαχωριστική Ανάλυση βρίσκει εφαρμογή σε αρκετούς επιστημονικούς τομείς. Ειδικότερα, έχει εφαρμογές για παράδειγμα στην Ιατρική για τη διάγνωση ασθενειών με βάση τα συμπτώματα, στις πολιτικές επιστήμες για τη μελέτη της εκλογικής συμπεριφοράς, στον τραπεζικό τομέα για την ανάλυση του προφίλ των δανειοληπτών, στην Πληροφορική για την αναγνώριση προτύπων (pattern recognition).

Η πρώτη βασική προϋπόθεση για την εφαρμογή της διαχωριστικής ανάλυσης είναι ότι οι παρατηρήσεις πρέπει να είναι μέλη δύο ή περισσότερων αμοιβαία αποκλειόμενων ομάδων, δηλαδή οι ομάδες πρέπει να ορίζονται με τέτοιο τρόπο, ώστε κάθε παρατήρηση να ανήκει σε μία και μόνη ομάδα. Οι μεταβλητές στη διαχωριστική ανάλυση είναι είτε ποσοτικές είτε δείκτριες. Οι μεταβλητές αυτές ονομάζονται και διακριτικές μεταβλητές (discriminating variables). Πρέπει να υπάρχει όμως στα δεδομένα μία ποιοτική μεταβλητή, με βάση την οποία να ορίζονται οι ομάδες των παρατηρήσεων, όπως στην ανάλυση διασποράς. Σύμφωνα με την Ηλιοπούλου (2015) η διαχωριστική ανάλυση μπορεί να θεωρηθεί ως μια τεχνική παλινδρόμησης, όπου εξαρτημένη μεταβλητή είναι η ποιοτική μεταβλητή, η οποία ορίζει τις ομάδες των παρατηρήσεων και ανεξάρτητες μεταβλητές οι ποσοτικές μεταβλητές, οι οποίες διαχωρίζουν τις ομάδες (Klecka, 1980).

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 06:38:11 EEST - 18.216.131.50

Όπως και σε άλλες στατιστικές μεθοδολογίες, έτσι και στη Διαχωριστική Ανάλυση, υπάρχουν κάποιοι περιορισμοί στις στατιστικές ιδιότητες που πρέπει να έχουν οι διακριτικές μεταβλητές. Στη συνέχεια παρατίθενται αυτές (Ηλιοπούλου, 2015):

Υπόθεση 1 DA. Κάθε ομάδα έχει προκύψει από πληθυσμό ο οποίος ακολουθεί την πολυδιάστατη (πολυμεταβλητή) κανονική κατανομή. Δηλαδή κάθε μεταβλητή έχει κανονική κατανομή για συγκεκριμένες τιμές των άλλων μεταβλητών.

Υπόθεση 2 DA. Δεν πρέπει να χρησιμοποιούνται διακριτικές μεταβλητές με μεγάλους συντελεστές συσχέτισης μεταξύ τους (πολυσυγγραμμικότητα).

Υπόθεση 3 DA. Στην περίπτωση του γραμμικού διαχωριστικού κανόνα, οι πίνακες διακύμανσεων συνδιακυμάνσεων (variance-covariance matrices) των ομάδων είναι ίσοι. Αυτή η προϋπόθεση διευκολύνει τον υπολογισμό των διακριτικών συναρτήσεων, επειδή οι συναρτήσεις αυτές υπολογίζονται συνήθως ως γραμμικοί συνδυασμοί των διακριτικών μεταβλητών.

Τέλος, το δείγμα πρέπει να είναι μεγάλο ώστε να είναι ανθεκτική (robust) η διαδικασία σε περίπτωση μη κανονικών κατανομών. Επιπλέον, το πλήθος των πειραματικών μονάδων πρέπει να υπερβαίνει τον αριθμό των μεταβλητών και για αυτό το λόγο συνήθως προτείνεται ο κανόνας ο αριθμός των πειραματικών μονάδων να είναι εικοσαπλάσιος του αριθμού των διακριτικών μεταβλητών (Ηλιοπούλου, 2015).

Στο πλαίσιο αυτό στη βιβλιογραφία έχουν εμφανιστεί διάφορες μεθοδολογίες για την κατασκευή ενός κανόνα διαχωρισμού, διάκρισης δύο πληθυσμών με τη βοήθεια των τιμών που παίρνουν κάποιες μεταβλητές (τις οποίες τις καθορίζουμε κάθε φορά εξαρχής με βάση το διαχωρισμό που θέλουμε να κάνουμε). Στη συνέχεια εν συντομία αναφέρουμε τις τέσσερεις σημαντικότερες και πιο ευρέως χρησιμοποιούμενες μεθόδους.

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 06:38:11 EEST - 18.216.131.50

α) ***Το κριτήριο Μέγιστης Πιθανοφάνειας***

Είναι ένας απλός τρόπος για να αποφασίσουμε σε ποια ομάδα θα κατατάξουμε μια καινούρια παρατήρηση. Η λογική του κριτηρίου είναι να βρεθεί η τιμή της πιθανοφάνειας που έχει αυτή η παρατήρηση στην καθεμία ομάδα, και όπου έχουμε τη μεγαλύτερη πιθανοφάνεια θα είναι και η πιο πιθανή περιοχή για να κατατάξουμε την παρατήρησή μας.

β) Ο κανόνας του Bayes

Καθώς ο πιο πάνω κανόνας δε λαμβάνει υπόψη του εάν οι ομάδες μας έχουν διαφορετικά μεγέθη, θέλουμε να βρούμε έναν κανόνα στον οποίο θα χρησιμοποιούμε και την πιθανότητα να πάρουμε μια παρατήρηση από την κάθε ομάδα. Για να γίνει αυτό πρέπει να χρησιμοποιήσουμε τον κανόνα απόφασης του Bayes, οπότε πρέπει να βρούμε τις εκ των υστέρων πιθανότητες οι οποίες για να βρεθούν χρειαζόμαστε τις τιμές της πιθανοφάνειας αλλά και τις εκ των προτέρων πιθανότητες.

γ) Ο κανόνας ελαχιστοποίησης του κόστους λανθασμένης κατάταξης

Όπως είναι κατανοητό οι κανόνες διαχωρισμού κάποιες φορές κάνουν λάθος στην κατάταξη. Όμως σε μερικές περιπτώσεις, ενδέχεται η κατάταξη μιας παρατήρησης που θα έπρεπε να είναι στην $1^{η}$ ομάδα, αλλά τοποθετείται στην $2^{η}$ ομάδα, να έχει σοβαρότερη επίπτωση από την αντίθετη περίπτωση. Επομένως θα πρέπει να λαμβάνουμε υπόψη και το κόστος που θα έχουμε από την κάθε μια λανθασμένη κατάταξη ώστε να βρεθεί ένας βέλτιστος κανόνας. Η λογική αυτή κρύβεται πίσω από τον κανόνα ελαχιστοποίησης του κόστους λανθασμένης ταξινόμησης

δ) Ο κανόνας ελαχιστοποίησης της συνολικής πιθανότητας λανθασμένης κατάταξης

Είναι ένας κανόνας που χρησιμοποιείται για το διαχωρισμό δύο πληθυσμών, ο οποίος όμως δε λαμβάνει υπόψη το κόστος. Σκοπός του είναι να ελαχιστοποιήσει

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 06:38:11 EEST - 18.216.131.50

τη συνολική πιθανότητα λανθασμένων κατατάξεων (TPM - Total Probability of Misclassification).

Κεφάλαιο 4: *Ομοιότητες και Διαφορές Λογιστικής Παλινδρόμησης και Διαχωριστικής Ανάλυσης*

Στο Κεφάλαιο αυτό περιγράφονται, συνοπτικά, οι ομοιότητες και οι διαφορές της διαχωριστικής ανάλυσης και της λογιστικής παλινδρόμησης.

Ομοιότητες: Από την περιγραφή του στόχου των δύο μεθόδων που έγινε στα προηγούμενα κεφάλαια είναι κατανοητό ότι οι δύο μέθοδοι χρησιμοποιούνται ή μπορούν να χρησιμοποιηθούν για να κατατάξουν, ταξινομήσουν μία ή περισσότερες νέες πειραματικές μονάδες σε γνωστές ομάδες. Επιπλέον πρόκειται για διαδεδομένες μεθοδολογίες, οι οποίες είναι διαθέσιμες σε όλα τα στατιστικά πακέτα. Τόσο οι ανεξάρτητες μεταβλητές όσο και οι διακριτικές μεταβλητές δεν πρέπει να έχουν υψηλούς συντελεστές συσχέτισης (βλέπε Υπόθεση 3 DA και Υπόθεση 3 LR).

Διαφοροποιήσεις: Στη λογιστική παλινδρόμηση δε χρειάζονται οι περίπλοκες υποθέσεις που απαιτούνται να γίνουν στη διαχωριστική ανάλυση. Δηλαδή, στη λογιστική παλινδρόμηση δε μας ενδιαφέρει αν όλες οι ανεξάρτητες μεταβλητές ακολουθούν κανονική κατανομή ή και αν έχουν ίσες διασπορές για τον κάθε ένα πληθυσμό όπως συμβαίνει στη διαχωριστική ανάλυση. Από την άλλη μεριά η διαχωριστική ανάλυση στηρίζεται σε πιο ρεαλιστικές μεθόδους και υπολογιστικά είναι πιο εύκολη. Αν σε κάποια περίπτωση ισχύουν: α) η υπόθεση της κανονικότητας και β) οι πίνακες διακυμάνσεων συνδιακυμάνσεων για κάθε πληθυσμό είναι ίσοι, η διαχωριστική ανάλυση δίνει καλύτερα αποτελέσματα σε σχέση με τη λογιστική παλινδρόμηση (Ξενή, 2012)

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 06:38:11 EEST - 18.216.131.50

Κεφάλαιο 5: ***Εφαρμογές***

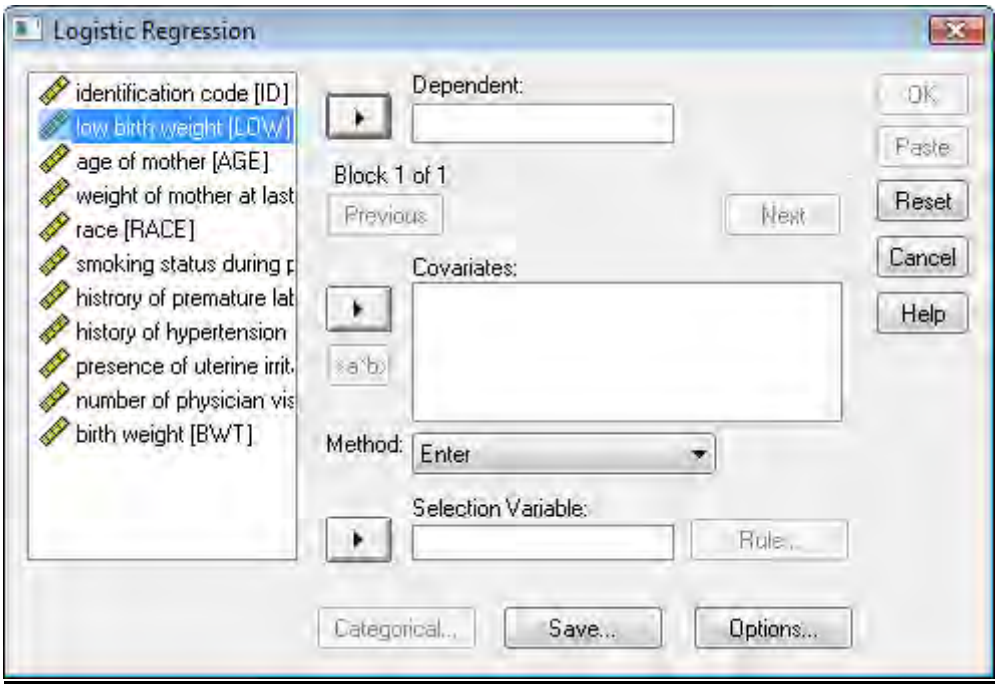
Εφαρμογή 1

Στο αρχείο lowbirthweight.sav (πηγή Hosmer and Lemeshow, 2000, http://www.umass.edu/statdata/statdata/stat-logistic.html) καταγράφονται πληροφορίες για 189 γεννήσεις καθώς και για τις μητέρες των νεογνών. Το ενδιαφέρον επικεντρώνεται στη μελέτη του φαινομένου της γέννησης νεογνών με βάρος μικρότερου των 2.500 γραμμαρίων. Το ενδιαφέρον εξηγείται διότι η θνησιμότητα των νεογνών σε τέτοιες περιπτώσεις είναι πολύ υψηλή. Οι πληροφορίες που μεταξύ άλλων καταγράφονται παρατίθενται στον πίνακα που ακολουθεί. Σκοπός μας είναι να αναδείξουμε πως μπορεί να χρησιμοποιηθεί η λογιστική παλινδρόμηση ως μέθοδος διαχωριστικής ανάλυσης.

identification code	Κωδικός αριθμός συμμετέχοντα
Low birth weight	Αν το νεογνό ζυγίζει λιγότερο από 2500γρ
Age of mother	Ηλικία της μητέρας σε έτη
Weight of mother at last menstrual period	Βάρος της μητέρας την τελευταία εμμηνορροϊκή περίοδο
Race	Φυλή, γένος.
smoking status during pregnancy	Κάπνισμα κατά την κυοφορία
History of premature labor	Ιστορικό πρόωρων γεννών
History of hypertension	Ιστορικό υπέρτασης
Presence of uterine irritability	Παρουσία ερεθιστικότητας
Number of physician visits during the first trimester	Αριθμός επισκέψεων ιατρού κατά το πρώτο τρίμηνο
birth weight	Βάρος νεογνού

Προσαρμόζουμε το μοντέλο της Binary λογιστικής παλινδρόμησης μέσω της διαδικασίας Analyze Regression Binary Logistic

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 06:38:11 EEST - 18.216.131.50



Στο παράθυρο διαλόγου που προκύπτει τοποθετούμε στο πλαίσιο **Dependent** την κατηγορική (δίτιμη αφού επιλέξαμε Binary Logistic) μεταβλητή, ενώ στο πλαίσιο **Covariates** τις ανεξάρτητες μεταβλητές, οι οποίες είναι είτε κατηγορικές με k το πλήθος δυνατές τιμές είτε συνεχείς. Πατώντας το πλαίσιο Categorical δηλώνουμε τις όποιες ανεξάρτητες μεταβλητές είναι κατηγορικές. Τότε δημιουργούνται k-1 δείκτριες μεταβλητές.

Στο αρχείο των αποτελεσμάτων ο πίνακας **Dependent Variable Encoding** μας πληροφορεί για την κωδικοποίηση της εξαρτημένης μεταβλητής, ενώ ο πίνακας Categorical Variables Encoding δίνει τις συχνότητες για τις κατηγορικές μεταβλητές του μοντέλου καθώς και την κωδικοποίηση των δείκτριων μεταβλητών που θα χρησιμοποιηθούν. Έτσι προκύπτει ότι race (1) είναι για τους λευκούς κοκ.

Dependent Variable Encoding

Original Value	Internal Value
>=2500g	0
<2500g	1

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 06:38:11 EEST - 18.216.131.50

Categorical Variables Codings

		Frequency	Parameter coding (1)	Parameter coding (2)
race	White	96	1,000	,000
	Black	26	,000	1,000
	Other	67	,000	,000
presence of uterine irritability	no	161	1,000	
	yes	28	,000	
history of hypertension	no	177	1,000	
	yes	12	,000	
smoking status during pregnancy	no	115	1,000	
	yes	74	,000	

Στον πίνακα Omnibus Tests of Model Coefficients μας δίνεται η τιμή καθώς και η αντίστοιχη p- τιμή του X^2 στατιστικού για τον έλεγχο ότι το συνολικό μοντέλο είναι στατιστικά σημαντικό. Δηλαδή είναι το αντίστοιχο του F-τεστ της γραμμικής παλινδρόμησης. Παρατηρούμε ότι η p-τιμή είναι μικρότερη του 0.05 επομένως το μοντέλο είναι στατιστικά σημαντικό.

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	33,387	9	,000
	Block	33,387	9	,000
	Model	33,387	9	,000

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 06:38:11 EEST - 18.216.131.50

Στον πίνακα Variables in the Equation έχουμε το προσαρμοζόμενο μοντέλο. Πληροφορούμαστε λοιπόν για την σχέση των ανεξάρτητων μεταβλητών με την εξαρτημένη μεταβλητή, η οποία είναι στην κλίμακα Logit. Από τις τιμές των Wald στατιστικών και τις αντίστοιχες p- τιμές κρίνουμε ποιες μεταβλητές είναι στατιστικά σημαντικές. Διαπιστώνουμε ότι οι μεταβλητές age, ptl, ui ftv δεν συνεισφέρουν στατιστικά σημαντικά στο μοντέλο.

Αν σκοπός μας είναι να αποκτήσουμε ένα μοντέλο που προσαρμόζεται όσο γίνεται καλύτερα και επιπλέον ελαχιστοποιείται ο αριθμός των παραμέτρων το επόμενο λογικό βήμα είναι η προσαρμογή ενός μοντέλου που περιέχει τις μεταβλητές που είναι στατιστικά σημαντικές και η σύγκρισή του με το πλήρες μοντέλο.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	AGE	-,030	,037	,637	1	,425	,971
	LWT	-,015	,007	4,969	1	,026	,985
	RACE			7,116	2	,028	
	RACE(1)	-,880	,441	3,990	1	,046	,415
	RACE(2)	,392	,538	,531	1	,466	1,480
	SMOKE(1)	-,939	,402	5,450	1	,020	,391
	PTL	,543	,345	2,474	1	,116	1,722
	HT(1)	-1,863	,698	7,136	1	,008	,155
	UI(1)	-,768	,459	2,793	1	,095	,464
	FTV	,065	,172	,143	1	,705	1,067
	Constant	4,931	1,493	10,908	1	,001	138,506

a. Variable(s) entered on step 1: AGE, LWT, RACE, SMOKE, PTL, HT, UI, FTV.

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 06:38:11 EEST - 18.216.131.50

Προσαρμόζοντας το νέο μοντέλο προκύπτει ότι

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	LWT	-,018	,007	6,937	1	,008	,982
	RACE			8,095	2	,017	
	RACE(1)	-,944	,423	4,968	1	,026	,389
	RACE(2)	,344	,536	,411	1	,521	1,411
	SMOKE(1)	-1,072	,388	7,646	1	,006	,342
	HT(1)	-1,749	,691	6,411	1	,011	,174
	Constant	4,116	1,252	10,817	1	,001	61,339

a. Variable(s) entered on step 1: LWT, RACE, SMOKE, HT.

Οι εκτιμητές (στήλη B) μας δίνουν την αύξηση (ή αντίστοιχα μείωση αν το πρόσημο είναι αρνητικό), στα προβλεπόμενα log odds της low birth weight=1 όταν θα έχουμε μοναδιαία αύξηση στην αντίστοιχη ανεξάρτητη μεταβλητή διατηρώντας τις υπόλοιπες σταθερές. Έτσι αύξηση του βάρους της μητέρας κατά την τελευταία εμμηνορροϊκή περίοδο κατά ένα κιλό αναμένεται να επιφέρει ελάττωση κατά 0.018 στα log odds της low birth weight=1. Στη στήλη Exp(B) δίνονται τα odds ratio για τις ανεξάρτητες μεταβλητές.

Επομένως το νέο μοντέλο που προκύπτει είναι το ακόλουθο:

$$\ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = 4.116 - 0.018LWT - 0.944RACE(1) + 0.344RACE(2) - 1.072SMOKE(1) - 1.749HT(1)$$

Έτσι για μία μητέρα με βάρος κατά την τελευταία εμμηνορροϊκή περίοδο 60 κιλά που είναι λευκή, δεν κάπνιζε κατά την εγκυμοσύνη ενώ είχε ιστορικό υπέρτασης προκύπτει η ακόλουθη πρόβλεψη:

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 06:38:11 EEST - 18.216.131.50

$$\ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = 4.116 - 0.018*50 - 0.944*1 + 0.344*0 - 1.072*1 - 1.749*0,$$

δηλαδή

$$\ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = 4.116 - 0.9 - 0.944 - 1.072 = 1.2.$$

Επομένως αφού Ln(odds)=1.2 μετά από πράξεις έχουμε ότι

$$\hat{p} = \frac{\exp(1.2)}{1+\exp(1.2)} = 0,77.$$

Επομένως αυτή η γυναίκα έχει 77% πιθανότητα να γεννήσει παιδί με βάρος μικρότερο των 2500 γραμμαρίων και καθώς ξεπερνά το 50% θα την ταξινομούσαμε στο πρώτο γκρουπ.

Αποτίμηση της προβλεπτικής ικανότητας της λογιστικής παλινδρόμησης

Το συνολικό ποσοστό ορθής ταξινόμησης είναι 73,5%.

Classification Table[a]

			Predicted		
			low birth weight		
	Observed		>=2500g	<2500g	Percentage Correct
Step 1	low birth weight	>=2500g	123	7	94,6
		<2500g	43	16	27,1
	Overall Percentage				73,5

a. The cut value is ,500

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 06:38:11 EEST - 18.216.131.50

Επιπλέον παρατηρούμε ότι το ποσοστό ορθής ταξινόμησης εντός των νεογνών με βάρος μεγαλύτερο των 2500 γραμμαρίων είναι ίσο με 94,6% (specificity) ενώ το ποσοστό ορθής ταξινόμησης εντός των νεογνών με βάρος μικρότερο των 2500 γραμμαρίων είναι 27,1% (sensitivity).

Ισχύουν οι ακόλουθοι ορισμοί:

Accuracy: $\frac{\text{Correctly Classified}}{\text{Total observed}} = \frac{123+16}{123+16+7+43}$

Sensitivity ή true positive fraction: $\frac{\text{Correctly Classified as } Y=1}{\text{Total observed as } Y=1} = \frac{16}{59}$

Specificity ή false positive fraction: $\frac{\text{Correctly Classified as } Y=0}{\text{Total observed as } Y=0} = \frac{123}{130}$

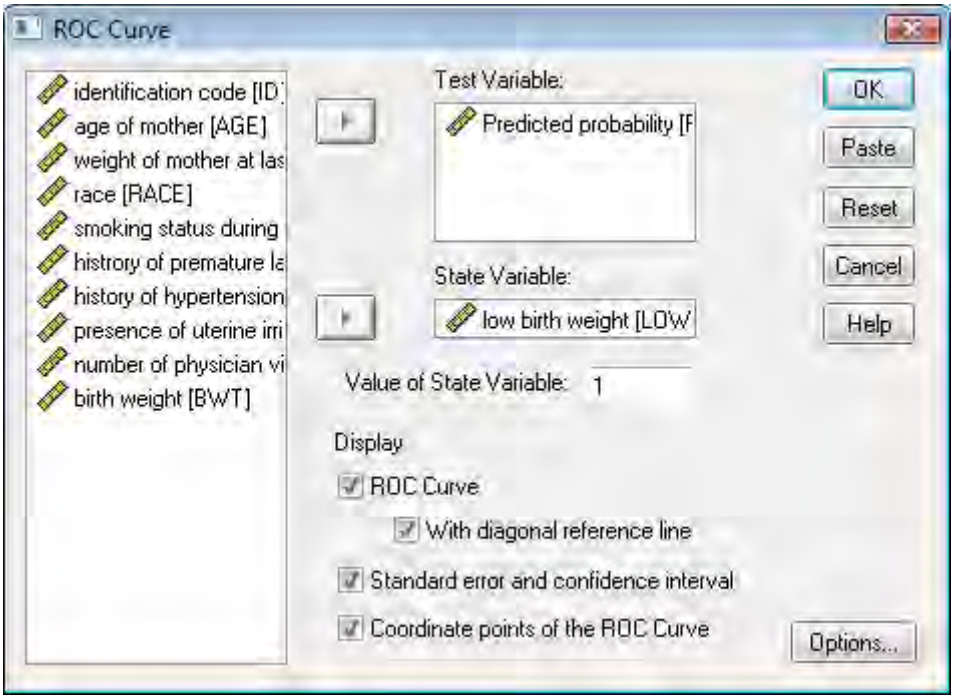
Positive predictive value: $\frac{\text{Correctly Classified as } Y=1}{\text{Total classified as } Y=1} = \frac{16}{23}$

Negative predictive value: $\frac{\text{Correctly Classified as } Y=0}{\text{Total classified as } Y=0} = \frac{123}{166}$.

Οι δείκτες sensitivity και specificity χρησιμοποιούνται για την αποτίμηση της ακρίβειας της πρόβλεψης. Μας δίνουν πόσα καλά γίνεται η ταξινόμηση μεταξύ πειραματικών μονάδων που ικανοποιούν ή όχι μία συγκεκριμένη συνθήκη. Η επιλογή του cut-off είναι εκείνη που καθορίζει το πλήθος των σωστών και λανθασμένων ταξινομήσεων. Εύκολα γίνεται αντιληπτό ότι καθώς αυξάνεται η sensitivity ταυτόχρονα μειώνεται η specificity. Μία πιο πλήρη περιγραφή της ακρίβειας της ταξινόμησης επιτυγχάνεται με την λεγόμενη Receiver Operating Characteristic (ROC) Curve. Είναι το γράφημα της (1-specificity) στον άξονα των X και sensitivity στον άξονα των Y για τις διάφορες τιμές του cut off point. Το εμβαδό του χωρίου κάτω από την καμπύλη (αναφέρεται και ως index of accuracy A ή *concordance index*) αποτελεί έναν δείκτη της ακρίβειας. Όσο μεγαλύτερο τόσο καλύτερη η ισχύς της πρόβλεψης. Η ROC αν είναι στη διαγώνιο σημαίνει τυχαίος ταξινομικός κανόνας (ισοδύναμος με το να ρίχνουμε ένα νόμισμα).

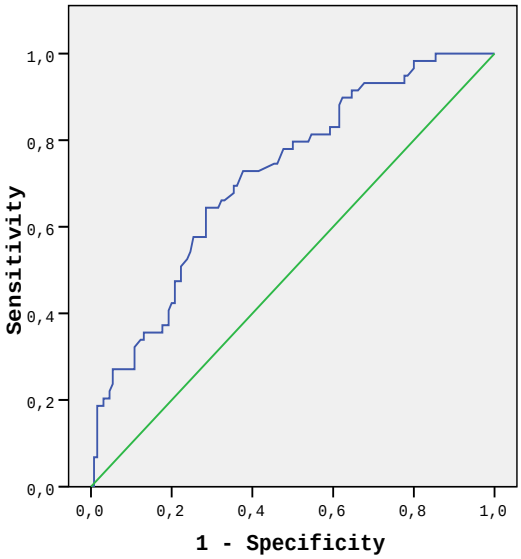
Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 06:38:11 EEST - 18.216.131.50

Χρησιμοποιώντας τις αποθηκευμένες προβλεπόμενες πιθανότητες πραγματοποίησης του ενδεχομένου κατασκευάζουμε την ROC Curve. Από το κεντρικό μενού επιλέγουμε Analyze Roc Curve και στη συνέχεια τα ακόλουθα:



Το αποτέλεσμα είναι το ακόλουθο

ROC Curve



Diagonal segments are produced by ties.

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 06:38:11 EEST - 18.216.131.50

Area Under the Curve

Test Result Variable(s): Predicted probability

Area	Std. Error(a)	Asymptotic Sig.(b)	Asymptotic 95% Confidence Interval	
			Upper Bound	Lower Bound
,718	,039	,000	,641	,794

The test result variable(s): Predicted probability has at least one tie between the positive actual state group and the negative actual state group. Statistics may be biased.

a Under the nonparametric assumption

b Null hypothesis: true area = 0.5

Στο πλαίσιο Area έχουμε ότι η περιοχή κάτω από την ROC είναι ίση με 0.718, που σημαίνει ότι 71,8% των πιθανών ζευγαριών όπου κάποιο νεογνό έχει βάρος μικρότερο των 2500 γραμμαρίων και το άλλο μεγαλύτερο το μοντέλο θα επιφορτίσει με μεγαλύτερη πιθανότητα αυτό με βάρος μεγαλύτερο των 2500. Επιπλέον, η p-τιμή είναι μικρότερη από 0.05, το οποίο σημαίνει ότι η χρησιμοποίηση του μοντέλου είναι καλύτερη από το στρίψιμο ενός νομίσματος

Εφαρμογή 2

Το δεύτερο σύνολο δεδομένων προέρχεται από το σύγγραμμα των Hosmer and Lemeshow (2000, p. 3). Καταγράφεται η ηλικία (age), η κατάταξη σε ηλικιακό γκρουπ (agerp) και η παρουσία ή όχι σημαντικών στεφανιαίων διαταραχών. Το ενδιαφέρον επικεντρώνεται στη μελέτη της σχέσης της ηλικίας με την παρουσία ή όχι διαταραχών.Εφαρμόζοντας λοιπόν το μοντέλο της Λογιστικής Παλινδρόμησης μέσω της διαδικασίας Binary Logistic επιλέγουμε την ηλικία (age) ως ανεξάρτητη μεταβλητή.

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 06:38:11 EEST - 18.216.131.50

Στον πίνακα Omnibus Tests of Model Coefficients μας δίνεται η τιμή καθώς και η αντίστοιχη p- τιμή του X^2 στατιστικό για τον έλεγχο ότι το συνολικό μοντέλο είναι στατιστικά σημαντικό Δηλαδή είναι το αντίστοιχο του F-τεστ της γραμμικής παλινδρόμησης. Παρατηρούμε ότι η p-τιμή είναι μικρότερη του 0.05 επομένως το μοντέλο είναι στατιστικά σημαντικό.

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	29,310	1	,000
	Block	29,310	1	,000
	Model	29,310	1	,000

Στον πίνακα Variables in the Equation έχουμε το τελικό μοντέλο. Έτσι πληροφορούμαστε ότι το μοντέλο πρόβλεψης είναι:

$$\log\left(\frac{p}{1-p}\right) = -5,309 + 0,111age\,.$$

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1(a)	age	,111	,024	21,254	1	,000	1,117
	Constant	-5,309	1,134	21,935	1	,000	,005

a Variable(s) entered on step 1: age.

Πληροφορούμαστε λοιπόν για την σχέση των ανεξάρτητων μεταβλητών με την εξαρτημένη μεταβλητή, η οποία είναι στην κλίμακα Logit.

Οι εκτιμητές (στήλη B) μας δίνουν την αύξηση (ή αντίστοιχα μείωση αν το πρόσημο είναι αρνητικό), στα προβλεπόμενα log odds της chd=1 όταν θα έχουμε μοναδιαία αύξηση στην αντίστοιχη ανεξάρτητη μεταβλητή διατηρώντας τις υπόλοιπες σταθερές. Έτσι αύξηση της ηλικία κατά ένας έτος αναμένεται να επιφέρει αύξηση κατά 0.111 στα log odds της chd. Στη στήλη Exp(B) δίνονται τα odds ratio για τις ανεξάρτητες μεταβλητές.

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 06:38:11 EEST - 18.216.131.50

Έτσι αν έχουμε μία αύξηση της ηλικίας κατά 10 έτη έχουμε τότε την ακόλουθη πρόβλεψη για το Odds ratio:

$$OR(10) = \exp(10 * 0.111) = 3.03$$

Αυτό υποδηλώνει ότι αύξηση της ηλικίας κατά 10 έτη επιφέρει τριπλάσιο ρίσκο ως προς την πρόκληση στεφανιαίων διαταραχών.

Έτσι για ένα άτομο ηλικίας 40 ετών προκύπτει η ακόλουθη πρόβλεψη:

Ln(odds)=-5.309+0.111*40=-0.869.

Επομένως η προβλεπόμενη πιθανότητα είναι ίση με 0.2955 άρα μικρότερη από 50% και θα τον ταξινομούσαμε στον γκρουπ αυτών που δεν διατρέχουν κίνδυνο.

Το συνολικό ποσοστό ορθής ταξινόμησης είναι 74%.

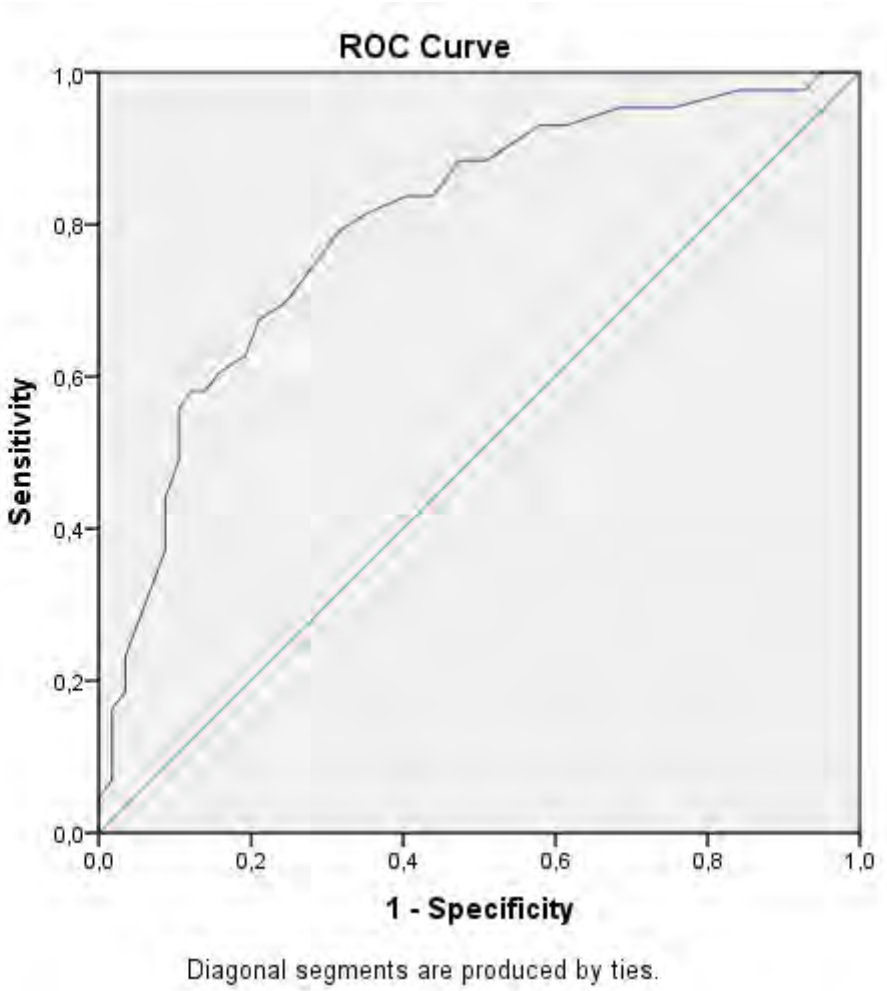
Classification Table(a)

Observed			Predicted		
			chd		Percentage Correct
			absence	presence	absence
Step 1	chd	Absence	45	12	78,9
		Presence	14	29	67,4
	Overall Percentage				74,0

a The cut value is ,500

Χρησιμοποιώντας τις αποθηκευμένες προβλεπόμενες πιθανότητες πραγματοποίησης του ενδεχόμενου κατασκευάζουμε την ROC Curve και το αποτέλεσμα είναι το ακόλουθο:

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 06:38:11 EEST - 18.216.131.50



Area Under the Curve

Test Result Variable(s): age

Area	Std. Error[a]	Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
,800	,045	,000	,712	,888

The test result variable(s): age has at least one tie between the positive actual state group and the negative actual state group. Statistics may be biased.

a. Under the nonparametric assumption

b. Null hypothesis: true area = 0.5

Στο πλαίσιο Area έχουμε ότι η περιοχή κάτω από τη ROC είναι ίση με 0,8, που σημαίνει ότι 80% των πιθανών ζευγαριών όπου κάποιος άνθρωπος που εμφανίζει στεφανιαία νόσο με κάποιον που δεν εμφανίζει θα επιφορτίσει με

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 06:38:11 EEST - 18.216.131.50

μεγαλύτερη πιθανότητα αυτόν που δεν εμφανίζει στεφανιαία νόσο. Επιπλέον, η p-τιμή είναι μικρότερη από 0,05, το οποίο σημαίνει ότι η χρήση του μοντέλου είναι πάρα πολύ καλή.

Εφαρμογή 3

Στην τρίτη εφαρμογή θα θεωρηθεί ένα μέρος του συνόλου δεδομένων που χρησιμοποιήθηκε στο πλαίσιο του μαθήματος «Προηγμένα Στατιστικά Μοντέλα» για την κατανόηση της μεθόδους της Διαχωριστικής ή Διαφοροποιούσας Ανάλυσης (βλέπε διαφάνειες Ζιντζαράς, 2016). Στο σύνολο αυτό των δεδομένων που θα χρησιμοποιηθεί καταγράφονται οι μετρήσεις 20 πρωτεϊνών που ανήκουν σε 2 διαφορετικές οικογένειες, ομάδες πρωτεϊνών (10 σε κάθε ομάδα) σε 4 μεταβλητές ενδιαφέροντος. Ειδικότερα,

X1= distances between the centres of two secondary structures.

X2= intercentroid tilt angles (radius) between each two structures.

X3= distances between the ends of secondary structures.

X4= connecting loop lengths given in amino acid residues between secondary structures.

Παρόμοια με την προηγούμενη εφαρμογή θα εξετάσουμε πως η λογιστική παλινδρόμηση μπορεί να χρησιμοποιηθεί ως μέθοδος διαχωριστικής ανάλυσης. Αρχικά παρατηρούμε ότι οι μεταβλητές X1, X3, X4 έχουν υψηλό συντελεστή συσχέτισης. Επιπλέον, το μοντέλο που προκύπτει ύστερα από διαδοχικές εφαρμογές του μοντέλου της λογιστικής παλινδρόμησης είναι το

$$\ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = -0.454\mathrm{X}_1 + 1.704\mathrm{X}_2$$

ή ισοδύναμα

$$\hat{p} = \frac{\exp(-0.454\mathrm{X}_1 + 1.704\mathrm{X}_2)}{1+\exp(-0.454\mathrm{X}_1 + 1.704\mathrm{X}_2)}.$$

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 06:38:11 EEST - 18.216.131.50

Αν αυτή η προβλεπόμενη πιθανότητα είναι μεγαλύτερη από 50% τότε ταξινομείται στην πρώτη ομάδα πρωτεϊνών. Η απόδοση αυτού του κανόνα δίνεται στον πίνακα **Classification Table,** με το εμβαδό κάτω από την καμπύλη ROC να είναι 83%.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	x1	-,454	,195	5,406	1	,020	,635
	x2	1,704	,729	5,455	1	,020	5,494

a. Variable(s) entered on step 1: x1, x2.

Classification Table[a]

			Predicted		
			group		Percentage Correct
	Observed		,00	1,00	
Step 1	group	,00	8	2	80,0
		1,00	2	8	80,0
	Overall Percentage				80,0

a. The cut value is ,500

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 06:38:11 EEST - 18.216.131.50

Βιβλιογραφία

1. Δημητροπουλάκης Π., Διδακτικές Σημειώσεις. Ιστότοπος lib.teiher.gr/webnotes/seyp/SPSS/Kef12.pdf 2.
2. Ηλιοπούλου Π., (2015). Γεωγραφική Ανάλυση. [ηλεκτρ. βιβλ.] Αθήνα:Σύνδεσμος Ελληνικών Ακαδημαϊκών Βιβλιοθηκών. Διαθέσιμο στο: http://hdl.handle.net/11419/2059
3. Ζιντζαράς Ηλίας, Διαφάνειες Μαθήματος «Προηγμένα Στατιστικά Μοντέλα» στα πλαίσια του ΜΔΕ «ΜΕΘΟΔΟΛΟΓΙΑ ΒΙΟΙΑΤΡΙΚΗΣ ΕΡΕΥΝΑΣ,ΒΙΟΣΤΑΤΙΣΤΙΚΗ ΚΑΙ ΚΛΙΝΙΚΗ ΒΙΟΠΛΗΡΟΦΟΡΙΚΗ», 2015-2016.
4. Hosmer D.W. & Lemeshow S. (2000). Applied Logistic Regression John Wiley & Sons, N. Jersey.
5. Klecka, W. (1980). Discriminant Analysis. Sage.
6. Tabachnick, B. G. & Fidell, L. S. (2012). Using multivariate statistics. Boston. MA: Pearson.
7. Ξενή Μαρία, (2012). Λογιστική Παλινδρόμηση & Διαχωριστική Ανάλυση. Διπλωματική Διατριβή, Πανεπιστήμιο Πατρών.

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 06:38:11 EEST - 18.216.131.50