Λογιστική Παλινδρόμηση - Logistic Regression

Δρ. Σωτήριος Δ. Νικολόπουλος

Big Data & Analytics

Πανεπιστήμιο Πελοποννήσου

Τμήμα Λογιστικής & Χρηματοοικονομικών

s.nikolopoulos@go.uop.gr

Λογιστική Παλινδρόμηση

Λογιστική Παλινδρόμηση

Λογιστική Παλινδρόμηση

Εισαγωγή

Η λογιστική παλινδρόμηση ερευνά το μη γραμμικό αποτέλεσμα μίας εξαρτημένης κατηγορικής μεταβλητής αναφορικά με τη δράση πολλών ανεξάρτητων μεταβλητών.

Χαρακτηρίζεται, αναλόγως της φύσης των κατηγοριών της εξαρτημένης μεταβλητής, από τρεις κατηγορίες μοντέλων,

τη **διωνυμική παλινδρόμηση** (με δυο μόνο κατηγορίες),

την **τακτική** (οι κατηγορίες διατάσσονται με αυξητική τάση) και

την **ονομαστική** (ποιοτικές κατηγορίες).

Λογιστική Παλινδρόμηση

Εισαγωγή

Η λογιστική παλινδρόμηση ερευνά το μη γραμμικό αποτέλεσμα μίας εξαρτημένης κατηγορικής μεταβλητής αναφορικά με τη δράση πολλών ανεξάρτητων μεταβλητών.

Χαρακτηρίζεται, αναλόγως της φύσης των κατηγοριών της εξαρτημένης μεταβλητής, από τρεις κατηγορίες μοντέλων,

τη **διωνυμική παλινδρόμηση** (με δυο μόνο κατηγορίες),

την **τακτική** (οι κατηγορίες διατάσσονται με αυξητική τάση) και

την **ονομαστική** (ποιοτικές κατηγορίες).

Λογιστική Παλινδρόμηση

Η λογιστική παλινδρόμηση (Logistic regression) αποτελεί στην ουσία ένα μοντέλο ταξινόμησης των τιμών μιας μεταβλητής απόκρισης Y με βάση τη θεωρία των πιθανοτήτων.

Στο μοντέλο αυτό όπου η μεταβλητή Y συνήθως έχει δυαδικό χαρακτήρα (λαμβάνει δύο τιμές) στοχεύεται η πρόβλεψη της έκβασης αυτής από ένα πλήθος προβλεπτικών μεταβλητών που μπορεί να είναι

ονομαστικές,

τακτικές ή

ποσοτικές.

Λογιστική Παλινδρόμηση

Η σημαντικότερη διαφοροποίηση μεταξύ λογιστικής και γραμμικής παλινδρόμησης βασίζεται στη φύση της επιλεγμένης μεταβλητής απόκρισης, η οποία στην μεν πρώτη μπορεί να είναι κατηγορική, (τακτική ή ονομαστική, στη δε δεύτερη αποκλειστικά ποσοτική.

Ενώ κατά την κλασική γραμμική παλινδρόμηση η εκτίμηση των παραμέτρων a και b_1 γίνεται με τη μέθοδο των ελάχιστων τετραγώνων, **κατά τη λογιστική παλινδρόμηση η εκτίμηση των παραμέτρων γίνεται με τη μέθοδο του λόγου πιθανοφάνειας** (maximum likelihood - μέθοδος συνήθως εφαρμοζόμενη στα γενικευμένα γραμμικά υποδείγματα), δηλαδή επιλέγονται οι πιο πιθανοφανείς τιμές των παραμέτρων, προκειμένου να οδηγήσουν στα παρατηρούμενα αποτελέσματα.

Ως επακόλουθο, η πρώτη παραδέχεται την ύπαρξη ομοιογένειας (ομοσκεδαστικότητας) στα υπολείμματα των αποκρίσεων ενώ στη δεύτερη **αναπτύσσεται πάντα ετεροσκεδαστικότητα σε κάθε προβλεπόμενη τιμή εξαιτίας του μεταβαλλόμενου ποσοστού διακύμανσης που αναλογεί σε αυτήν**.

Λογιστική Παλινδρόμηση

Διακρίνονται τρεις τύποι λογιστικής παλινδρόμησης ανάλογα με την ιδιαίτερη φύση της εξαρτημένης κατηγορικής μεταβλητής η οποία μπορεί να είναι:

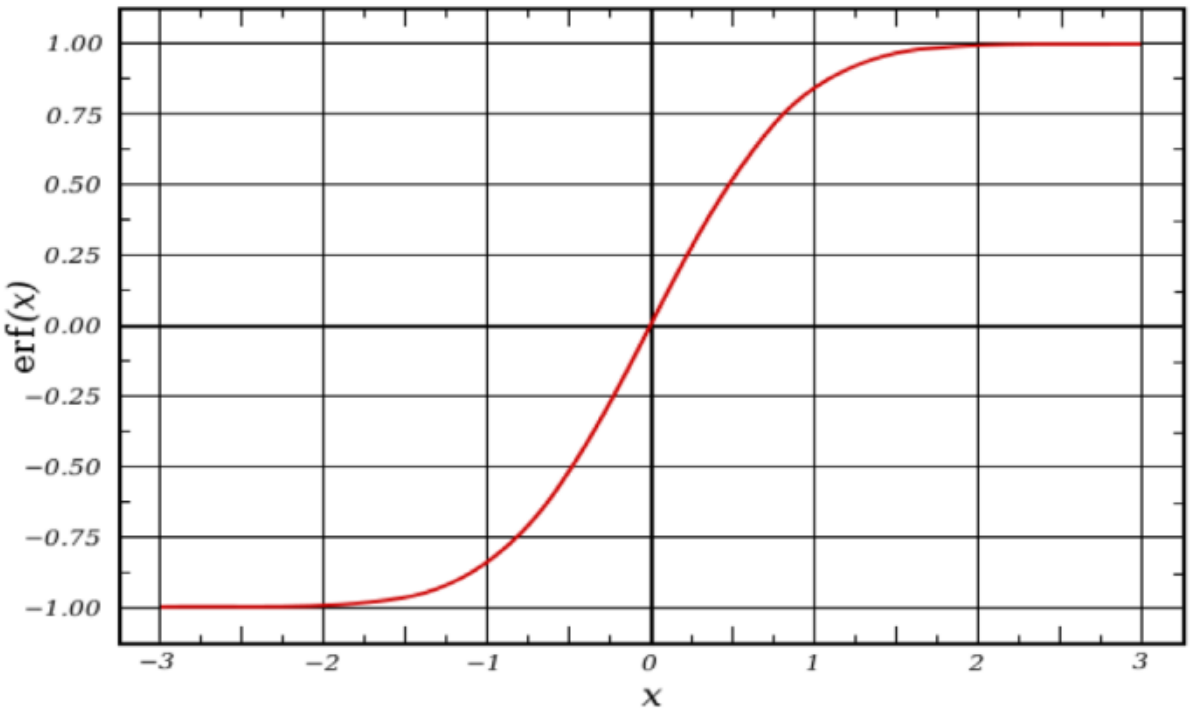
1. Δίτιμη ή δυαδική ή διχοτομική (binary) ή διμερής εξαρτημένη μεταβλητή. Συνίσταται από δύο κατηγορίες, όπως π.χ. είναι οι εκβάσεις επιτυχία/αποτυχία, ΝΑΙ/ΟΧΙ, γεγονός απόν/παρόν.

2. Τακτική (ordinal) μεταβλητή. Η εξαρτημένη μεταβλητή συνίσταται από τρεις ή περισσότερες κατηγορίες μεταξύ των οποίων ισχύει η έννοια της ανισότητας, όπως π.χ. σε μια ερώτηση της κλίμακας διαφωνώ καθόλου, λίγο, μέτρια, αρκετά, πολύ, στην κατάταξη ενός στρώματος υλικού ως λεπτού, μεσαίου, παχέος.

3. Ονομαστική (Nominal) ή πολυωνυμική (polynomial) ή πολυχοτομική (polychotomus) ή κατηγορική αδιαβάθμητη (non-ordered categorical) ή πολυμερής μεταβλητή απόκρισης. Περιέχει τρεις ή περισσότερες κατηγορίες χωρίς κάποια φυσική διαβάθμιση, όπως π.χ. ο χαρακτηρισμός ενός τροφίμου ως τραγανού, μαλακού, εύθρυπτου ή του χρώματος αντικειμένων ως ερυθρού, πράσινου, κίτρινου κτλ.

Ανάπτυξη του μοντέλου

Στη γλώσσα της στατιστικής, η λογιστική παλινδρόμηση χρησιμοποιείται για την πρόβλεψη της πιθανότητας εμφάνισης ενός γεγονότος προσαρμόζοντας τα δεδομένα της μελέτης στην εξίσωση της λογιστικής καμπύλης όπως αυτή παρουσιάζεται στο σχήμα



Σχήμα: Τυπική ανάπτυξη σιγμοειδούς καμπύλης.

Ανάπτυξη του μοντέλου

Η καμπύλη αυτή έχει σιγμοειδή μορφή και χαρακτηρίζεται από ένα στάδιο εκθετικής ανάπτυξης στο οποίο ο ρυθμός αύξησης επιβραδύνεται βαθμιαία και περατώνεται στο ασυμπτωτικό στάδιο κορεσμού της ανάπτυξης (η ευθεία βαίνει τελικά παράλληλα στον άξονα Χ).

Η δυαδική λογιστική παλινδρόμηση αποτελεί μια διωνυμική εξίσωση στην οποία η μεταβλητή απόκρισης Υ είναι το τυχαίο αποτέλεσμα εμφάνισης μιας από δύο δυνητικές εκβάσεις του τύπου επιτυχία ή αποτυχία όπως π.χ. είναι το αποτέλεσμα της ρίψης ενός νομίσματος δύο διαφορετικών όψεων (κορώνα-γράμματα), η ρίψη ενός ζαριού όπου το αποτέλεσμα εμφάνισης του αριθμού 6 θεωρείται επιτυχία και των λοιπών αριθμών αποτυχία, η θετική ψήφος εκλογής ενός πολιτικού εκπροσώπου κτλ.

Ανάπτυξη του μοντέλου

Η δίτιμη λογιστική παλινδρόμηση έχει τη μορφή

$$f(z) = \frac{e^z}{1+e^z} = \frac{1}{1+e^{-z}}$$

όπου z είναι η μεταβλητή εισόδου και f(z) το αποτέλεσμα αυτής. Στα πλεονεκτήματα της εξίσωσης συγκαταλέγεται και το γεγονός ότι η μεταβλητή εισόδου λαμβάνει θετικές και αρνητικές τιμές ενώ το αποτέλεσμα αυτής f(z) περιορίζεται σε εύρος τιμών μεταξύ 0 και 1.

Αναλυτικότερα, η μεταβλητή z εκπροσωπεί τη δράση μιας ομάδας ανεξάρτητων μεταβλητών ενώ η $f(z)$ προσδιορίζει την πιθανότητα ενός συγκεκριμένου αποτελέσματος λόγω της δράσης της ομάδας αυτής.

Ανάπτυξη του μοντέλου

Η μεταβλητή z (λογιστική) εκφράζει επίσης το μέτρο της ολικής συνεισφοράς όλων των συμμετεχουσών ανεξάρτητων μεταβλητών στο μοντέλο και ορίζεται ως

$$z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$$

όπου β_0 είναι το ύψος της κλίσης της γραμμής παλινδρόμησης και ισούται με την τιμή z όταν οι τιμές όλων των ανεξάρτητων μεταβλητών ισούνται με 0, ενώ β_i είναι οι συντελεστές παλινδρόμησης καθένας των οποίων εκφράζει το μέγεθος συνεισφοράς της αντίστοιχης μεταβλητής.

Θετική τιμή του συντελεστή δηλώνει ότι η επεξηγηματική μεταβλητή αυξάνει την πιθανότητα της επιτυχημένης έκβασης (να συμβεί δηλαδή το γεγονός), αρνητική τιμή σημαίνει ότι η μεταβλητή μειώνει την πιθανότητα αυτής της έκβασης.

Ανάπτυξη του μοντέλου

Υψηλή τιμή του συντελεστή σημαίνει ότι η ανεξάρτητη μεταβλητή επηρεάζει πολύ ισχυρά την πιθανότητα να συμβεί το γεγονός ή μη, ενώ χαμηλή τιμή δηλώνει μικρή επίδραση της ανεξάρτητης μεταβλητής στην πιθανότητα εμφάνισης της ανάλογης έκβασης.

Συνοψίζοντας, η λογιστική παλινδρόμηση χρησιμεύει στην περιγραφή της σχέσης που αναπτύσσεται μεταξύ μιας ή περισσότερων ανεξάρτητων μεταβλητών (π.χ. ηλικία, φύλο, τοξική συγκέντρωση ουσίας) και μιας δυαδικής μεταβλητής απόκρισης εκφρασμένης ως πιθανότητα δυνάμενη να πάρει μία από δύο τιμές, όπως π.χ. θετική (1) αρνητική (0), παρόν ενδεχόμενο (1) απόν ενδεχόμενο (0), επιζών (1) θανών (0), αρεστός (1) δυσάρεστος (0).

Ανάπτυξη του μοντέλου

Η φύση των ανεξάρτητων μεταβλητών εισόδου στην εξίσωση της πολλαπλής λογιστικής παλινδρόμησης μπορεί να είναι ποσοτική, τακτική ή ονομαστική (αδιαβάθμητη κατηγορική).

Για παράδειγμα, η πιθανότητα ένα άτομο να υποστεί καρδιακό επεισόδιο σε συγκεκριμένο χρονικό διάστημα (εξαρτημένη μεταβλητή) μπορεί να προβλεφθεί από ένα πλήθος ανεξάρτητων μεταβλητών όπως είναι η ηλικία, το φύλο, ο δείκτης μάζας σώματος, η φυσική αγωγή, το ιστορικό του ασθενούς κτλ.

Η ηλικία ενδέχεται να ενταχθεί στη εξίσωση είτε ως ποσοτική (με την πραγματική τιμή της) είτε ως τακτική: 15-25 ετών, 25-40, 40-60, >60.

Η φυσική αγωγή ως ονομαστική μεταβλητή (άθληση ή μη), ο δείκτης μάζας σώματος (Body Mass Index - BMI) ως ποσοτική ή τακτική (<25, 25-30, >30), και το ιστορικό ως ονομαστική (ύπαρξη προδιάθεσης ή μη). Επιστήμες όπως η ιατρική, οι κοινωνικές επιστήμες και το marketing καταφεύγουν συχνά στην εφαρμογή της πολυωνυμικής λογιστικής παλινδρόμησης.

Ανάπτυξη του μοντέλου

Οι συντελεστές της παλινδρόμησης υπολογίζονται με τη βοήθεια της εκτίμησης της μέγιστης πιθανοφάνειας (Maximum Likelihood Estimate – MLE), ως

$$L = \prod_{i=1}^{n} f(x_i, \theta)$$

ή προτιμότερο με τη λογαριθμική έκδοση αυτής,

$$L = \sum_{i=1}^{n} log_e(x_i, \theta)$$

όπου θ είναι μια παράμετρος της μεταβλητής η οποία μπορεί να μεταβάλλεται ελεύθερα.

Η προβλεπόμενη τιμή για κάθε παρατήρηση θα ισούται με

$$\hat{l} = \frac{1}{n} log_e L$$

Ανάπτυξη του μοντέλου

Η συνάρτηση της πιθανοφάνειας έκβασης ενός γεγονότος (likelihood) δείχνει πόσο κατάλληλα ένα παρατηρούμενο δείγμα περιγράφεται από κάποιες τιμές παραμέτρων π.χ. μέσος όρος, τυπική απόκλιση.

Άρα, η μεγιστοποίηση της συνάρτησης της πιθανότητας έκβασης καθορίζει τις παραμέτρους εκείνες που είναι οι πλέον ικανές να παράγουν τα παρατηρούμενα στοιχεία.

Από άποψη στατιστικής βαρύτητας, η MLE προτείνεται για εφαρμογές σε μεγάλα δείγματα καθόσον είναι ευέλικτη, προσαρμόζεται εύκολα στην παραγωγή πολλών διαφορετικού τύπου μοντέλων, το χειρισμό διαφορετικής φύσης στοιχείων και περιέχει ακριβέστερες μετρήσεις.

Ανάπτυξη του μοντέλου

Η αξιοπιστία των αποτελεσμάτων της λογιστικής παλινδρόμησης επηρεάζεται κατά πολύ από το δειγματοληπτικό μέγεθος της έρευνας.

Ένας χρυσός κανόνας υπαγορεύει την αντιστοιχία του αριθμού των επιθυμητών εκβάσεων προς τον αριθμό των ανεξάρτητων μεταβλητών να προσδιορίζεται από τη σχέση 10:1.

Εάν υπάρχουν ονομαστικές ανεξάρτητες μεταβλητές, όπως, για παράδειγμα, διχοτομικές, ο παραπάνω κανόνας θα ισχύει για το μέγεθος των παρατηρήσεων της ολιγοπληθέστερης κατηγορίας.

Πολλαπλή Διωνυμική Παλινδρόμηση

Αποτελεί μέρος κατηγορικών στατιστικών μοντέλων γνωστών ως Γενικευμένα Γραμμικά μοντέλα (McCullagh & Nelder, 1989), τα οποία περιλαμβάνουν τη γνωστή κλασική παλινδρόμηση, την ανάλυση διακύμανσης και συνδιακύμανσης και τη λογαριθμογραμμική παλινδρόμηση.

Η μέθοδος αυτή επιτρέπει την πρόβλεψη των τιμών εξαρτημένης διμερούς μεταβλητής μορφής από ένα πλήθος ανεξάρτητων μεταβλητών, οι οποίες μπορεί να είναι ποσοτικές, διχοτομικές ή πολυμερείς ή και συνδυασμοί αυτών.

Αντίποδας της λογιστικής παλινδρόμησης είναι η διακριτική ανάλυση κατηγοριών με τη διαφορά ότι στη δεύτερη συμμετέχουν μόνο ποσοτικές ανεξάρτητες μεταβλητές.

Η μεταβλητή απόκρισης στην αλγεβρική της έκδοση λαμβάνει την τιμή 1 με πιθανότητα επιτυχίας p και την τιμή 0 με πιθανότητα αποτυχίας 1-p και καλείται δυαδική (Binary) ή δυωνυμική (Binomial) ή μεταβλητή του Bernouli.

Πολλαπλή Διωνυμική Παλινδρόμηση

Λόγω της φύσης των συμμετεχουσών μεταβλητών, απουσιάζουν οι προϋποθέσεις της ομαλής κατανομής των τιμών και της ομοιογένειας των διακυμάνσεών τους, η δε έλλειψη της γραμμικότητας μεταξύ της Y και των ανεξάρτητων μεταβλητών βελτιώνεται με τη χρήση της λογαριθμικής εξίσωσης ως εξής,

$$p = \frac{e^z}{1 + e^z}$$

όπου

$$z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

Ακολούθως χρησιμοποιείται η σχέση: $log_e \left(\frac{p}{1-p} \right)$

ή λογαριθμώντας προκύπτει

$$\left(\frac{p}{1-p} \right) = e^z$$

Λογιστική Παλινδρόμηση: Παράδειγμα στην R

Σε αυτή την ενότητα θα δούμε την μέθοδο της λογιστικής παλινδρόμησης η οποία χρησιμεύει στο να αναπτύξουμε σχέση μίας δίτιμης ανεξάρτητης τυχαίας μετα□λητής και συνεχών η διακριτών ανεξάρτητων μεταβλητών.

Ουσιαστικά η μέθοδος αυτή γενικεύει τα γραμμικά μοντέλα, έτσι ώστε η εξαρτημένη μεταβλητή να ακολουθεί την εκθετική οικογένεια κατανομών.

Λογιστική Παλινδρόμηση: Παράδειγμα στην R

Περιγραφή των Δεδομένων

Έρευνα με εργάτες της Αμερικάνικης Βιομηχανίας Βαμβακιού θέλει να εξετάσει αν κάποιος εργάτης πάσχει από κάποια συγκεκριμένη ασθένεια του πνεύμονα.

Επίσης, συγκεντρώθηκαν οι τιμές για τις ακόλουθες πέντε μεταβλητές :

1. φυλή (race) (1=λευκός, 2=άλλο)
2. φύλο (sex) (1=άρρεν, 2=θήλυ)
3. κάπνισμα (1=καπνιστής, 2=μη καπνιστής)
4. διάρκεια εργασίας
   (1= λιγότερο από 10 χρόνια, 2=10–22 χρόνια, 3= περισσότερο από 20 χρόνια)
5. σκόνη : ποσοστό σκόνης στον εργασιακό χώρο (1=ψηλό, 2=μέτριο 3=χαμηλό)

Το πρόβλημα για αυτά τα δεδομένα είναι το να εξακριβωθεί κατά πόσο οι επεξηγηματικές μεταβλητές είναι σημαντικές στην εμφάνιση αυτής της ασθένειας.

Λογιστική Παλινδρόμηση: Παράδειγμα στην R

Με άλλα λόγια, ποιες από αυτές τις μεταβλητές μπορούν να χρησιμοποιηθούν για να προβλέψουν κατά πόσο ένας εργάτης πάσχει από ασθένεια του πνεύμονα.

Επειδή η ανεξάρτητη μεταβλητή είναι δυαδική, θα χρησιμοποιηθεί η λογιστική παλινδρόμηση για την ανάλυση.

Λογιστική Παλινδρόμηση: Παράδειγμα στην R

Λογιστική Παλινδρόμηση

Αντί να χρησιμοποιηθεί ένα γραμμικό μοντέλο για να εξεταστεί η εξάρτηση της πιθανότητας εμφάνισης της ασθένειας του πνεύμονα από τις επεξηγηματικές μεταβλητές, χρησιμοποιείται ο λογιστικός μετασχηματισμός, ο οποίος ορίζεται ως

$$log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

Στο παράδειγμα, p είναι η πιθανότητα ένας εργάτης να πάσχει από ασθένεια του πνεύμονα.

Στο μοντέλο υπάρχουν k (στο παράδειγμα 5) επεξηγηματικές μεταβλητές.

Οι συντελεστές παλινδρόμησης εκτιμούνται με τη μέθοδο της μέγιστης πιθανοφάνειας με την υπόθεση ότι η εξαρτημένη μεταβλητή ακολουθεί τη διωνυμική κατανομή.

Από την εξίσωση, το p μπορεί να υπολογιστεί από

$$p = \frac{exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k)}{1 + exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k)}$$

Λογιστική Παλινδρόμηση: Παράδειγμα στην R

Ανάλυση στην R

Οι πρώτες δύο στήλες των δεδομένων καταγράφουν τη συχνότητα των εργατών με ή χωρίς την ασθένεια για τις αντίστοιχες τιμές (κατηγορίες) των επεξηγηματικών μεταβλητών.

Η ανάλυση της λογιστικής παλινδρόμησης γίνεται με την εντολή **glm** με ανάλογο τρόπο με τη εντολή **lm** για τη γραμμική παλινδρόμησης δίνοντας και την συνάρτηση σύνδεσης (**link function**) με το όρισμα **family**.

```
out1<-glm( cbind(Yes, No)~dust+race+sex+smoking+Empleng,family=binomial, data=df)
> out1
Call:  glm(formula = cbind(Yes, No) ~ dust + race + sex + smoking +
Empleng, family = binomial, data = df)

Coefficients:
(Intercept)        dust        race         sex     smoking     Empleng
    -0.4898     -1.3747      0.2474     -0.2585     -0.6286      0.3861

Degrees of Freedom: 63 Total (i.e. Null);  58 Residual
Null Deviance:      322.3
Residual Deviance: 69.46        AIC: 188.1
```

Το αποτέλεσμα δίνει τις εκτιμήσεις των συντελεστών των παραμέτρων, την απόκλιση (deviance) του μηδενικού μοντέλου και των υπολοίπων μαζί με τους βαθμούς ελευθερίας τους αλλά και την τιμή του κριτηρίου AIC.

Πιο λεπτομερή ανάλυση των συντελεστών των παραμέτρων δίνεται με την εντολή **summary**, ενώ η εντολή **anova** παρουσιάζει τον πίνακα ανάλυσης της απόκλισης.

Λογιστική Παλινδρόμηση: Παράδειγμα στην R

Ανάλυση στην R

```
summary(out1)
Call:
glm(formula = cbind(Yes, No) ~ dust + race + sex + smoking +
Empleng, family = binomial, data = df)

Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.4898     0.6067  -0.807 0.419442
dust         -1.3747     0.1156 -11.896  < 2e-16 ***
race          0.2474     0.2062   1.200 0.230112
sex          -0.2585     0.2116  -1.222 0.221841
smoking      -0.6286     0.1931  -3.256 0.001130 **
Empleng       0.3861     0.1070   3.610 0.000306 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 322.341  on 63  degrees of freedom
Residual deviance:  69.457  on 58  degrees of freedom
AIC: 188.13

Number of Fisher Scoring iterations: 5
```

Λογιστική Παλινδρόμηση: Παράδειγμα στην R

Ανάλυση στην R

```
anova(out1)
Analysis of Deviance Table

Model: binomial, link: logit

Response: cbind(Yes, No)

Terms added sequentially (first to last)


         Df Deviance Resid. Df Resid. Dev
NULL                        63     322.34
dust      1  221.826        62     100.51
race      1    1.048        61      99.47
sex       1    5.959        60      93.51
smoking   1   10.716        59      82.79
Empleng   1   13.335        58      69.46
```

Λογιστική Παλινδρόμηση: Παράδειγμα στην R

Ανάλυση στην R

Από τα πιο πάνω συμπεραίνεται ότι οι μεταβλητές **dust**, **smoking** και **Empleng** είναι οι πιο σημαντικές για την πρόβλεψη ασθένειας του πνεύμονα, ενώ □αίνεται ότι οι άλλες δύο μεταβλητές δεν είναι τόσο σημαντικές.

Στο συμπέρασμα αυτό καταλήγουμε από το **p-value** τους για τον **t έλεγχο**, αλλά και από την συνεισφορά της κάθε μεταβλητής στην απόκλιση όταν αυτή προστεθεί στο μοντέλο, η οποία παρουσιάζεται στο πίνακα ανάλυσης της απόκλισης.

Συνεπώς, εφαρμόζεται ένα νέο μοντέλο λογιστικής παλινδρόμησης με τις τρεις σημαντικές μεταβλητές και το συγκρίνεται με το προηγούμενο.

Λογιστική Παλινδρόμηση: Παράδειγμα στην R

Ανάλυση στην R

```
out2<-glm( cbind(Yes, No)~dust+smoking+Empleng, family=binomial, data=df)
anova(out2,out1)
Analysis of Deviance Table

Model 1: cbind(Yes, No) ~ dust + smoking + Empleng
Model 2: cbind(Yes, No) ~ dust + race + sex + smoking + Empleng
Resid. Df Resid. Dev Df Deviance
1        60     72.517
2        58     69.457  2   3.0603

1-pchisq(3.053,2)
[1] 0.2172949
```

Ο έλεγχος σύγκρισης μοντέλου έχει για μηδενική υπόθεση H_0 ότι το νέο μοντέλο εφαρμόζει καλύτερα τα δεδομένα.

Ο έλεγχος είναι X^2 και αφού το p-value (1-pchisq(3.053,2)) είναι μεγαλύτερο από 0.05, δεν απορρίπτεται η μηδενική υπόθεση.

Ποιο κάτω παρουσιάζεται η ανάλυση για του συντελεστές του μικρότερου μοντέλου.

Λογιστική Παλινδρόμηση: Παράδειγμα στην R

Ανάλυση στην R

```
summary(out2)

Call:
glm(formula = cbind(Yes, No) ~ dust + smoking + Empleng, family = binomial,
data = df)

Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.14310    0.34129  -0.419 0.675001
dust        -1.46542    0.10578 -13.853  < 2e-16 ***
smoking     -0.67726    0.18871  -3.589 0.000332 ***
Empleng      0.33327    0.08861   3.761 0.000169 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 322.341  on 63  degrees of freedom
Residual deviance:  72.517  on 60  degrees of freedom
AIC: 187.19

Number of Fisher Scoring iterations: 5
```

Λογιστική Παλινδρόμηση: Παράδειγμα στην R

Ανάλυση στην R

Θεωρώντας τις τιμές των συντελεστών από πιο πάνω, το μοντέλο λογιστικής παλινδρόμησης που εφαρμόζει καλύτερα τα δεδομένα δίνεται από

$$log\left(\frac{\hat{p}}{1-\hat{p}}\right) = -0.14310 - 1.46542 \times dust - 0.67726 \times smoking + 0.33327 \times Em$$

και είναι δυνατόν να υπολογιστεί η εκτιμώμενη τιμή της πιθανότητας κάποιος εργάτης να πάσχει από ασθένεια του πνεύμονα για κάθε συνδυασμό τιμών από τις τρεις επεξηγηματικές μεταβλητές.

Για παράδειγμα, αν ένας εργάτης δουλεύει σε εργασιακό χώρο με ψηλό ποσοστό σκόνης (dust=1), καπνίζει (smoking=1) και δουλεύει για περισσότερο από 20 χρόνια (Empleng=3), η εξίσωση δίνει το αποτέλεσμα log ($\hat{p}$/(1 − $\hat{p}$)) = −1.286, και συνεπώς $\hat{p}$ = 0.2165.

Λογιστική Παλινδρόμηση: Παράδειγμα στην R

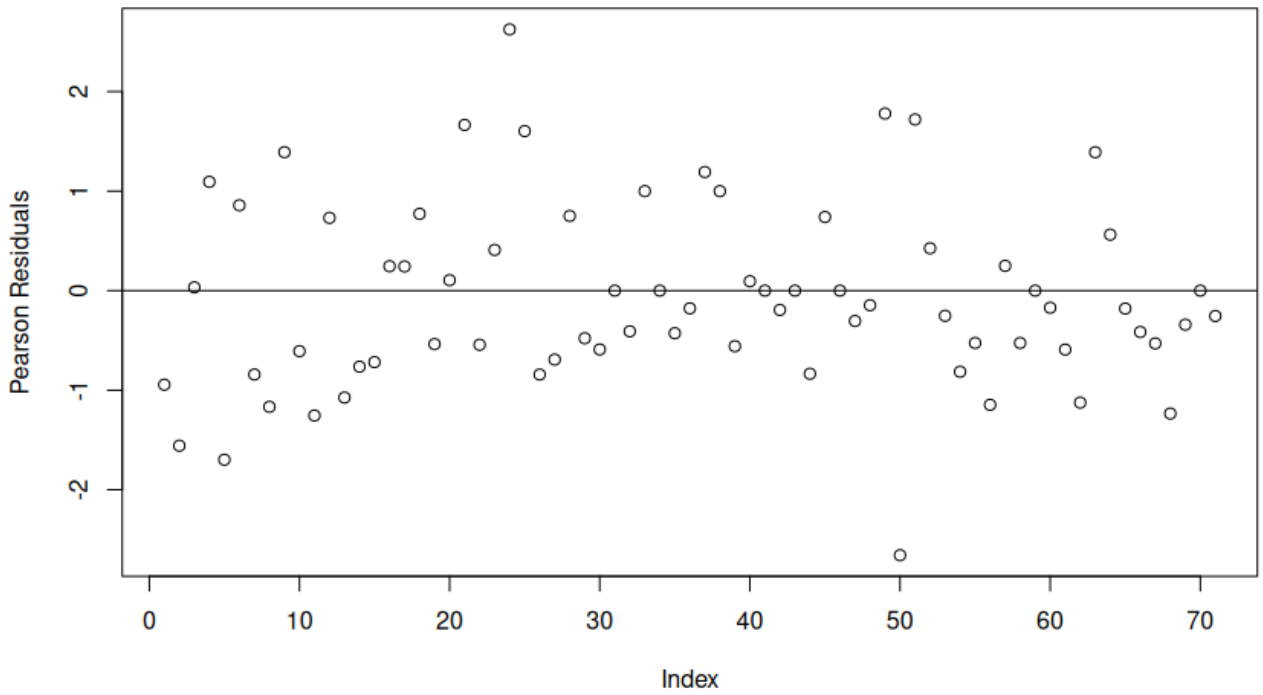
Ανάλυση στην R

Στη συνέχεια υπολογίζονται δύο είδη υπολοίπων της λογιστικής παλινδρόμησης, τα υπόλοιπα απόκλισης και τα υπόλοιπα Pearson, και κατασκευάζεται το γράφημά τους (Βλέπε επόμενα Σχήματα).

```
residuals(out2, type="pear")
plot(residuals(out2, type="pear"), xlab="Index",
ylab="Pearson Residuals")
abline(h=0)
```

Λογιστική Παλινδρόμηση: Παράδειγμα στην R

Ανάλυση στην R



Σχήμα: Υπόλοιπα απόκλισης.

Λογιστική Παλινδρόμηση: Παράδειγμα στην R

Ανάλυση στην R

Στη συνέχεια υπολογίζονται τα υπόλοιπα της λογιστικής παλινδρόμησης, (υπόλοιπα Pearson), και κατασκευάζεται το γράφημά τους.

Η μεθοδολογία της ανάλυσης υπολοίπων είναι παρόμοια με εκείνης της πολλαπλής γραμμικής παλινδρόμησης.

Από το γράφημα βλέπουμε ότι η 50η παρατήρηση είναι λίγο προβληματική.

Η αρνητική τιμή του υπολοίπου υποδεικνύει ότι η εκτιμώμενη τιμή είναι μεγαλύτερη από την παρατηρούμενη τιμή.

Εξετάζοντας τα δεδομένα, παρατηρείται ότι η 50η παρατήρηση αναφέρεται στους εργάτες με μέτριο ποσοστό σκόνης στον εργασιακό τους χώρο (dust=2), καπνίζουν (smoking=1) και εργάζονται για περισσότερα από 20 χρόνια (Empleng=3) και άρα $\hat{p}$ = 0.059.