Ανάλυση Κατηγορικών Δεδομένων

Στέλιος Ζήμερας
Τμήμα Στατιστικής και Αναλογιστικών – Χρηματοοικονομικών Μαθηματικών
Σάμος
2020

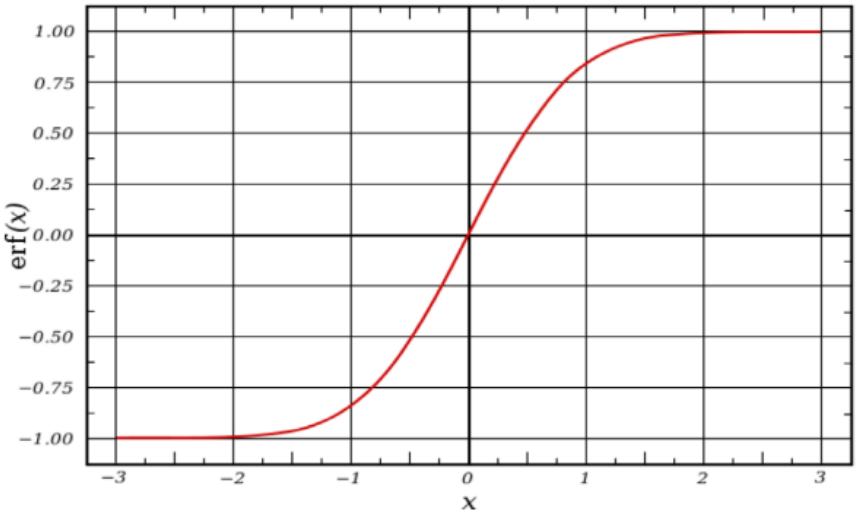
ΛΟΓΙΣΤΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ

Η λογιστική παλινδρόμηση **(Logistic regression)** αποτελεί στην ουσία ένα μοντέλο ταξινόμησης των τιμών μιας μεταβλητής απόκρισης Υ με βάση τη θεωρία των πιθανοτήτων. Στο μοντέλο αυτό όπου η μεταβλητή Υ 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.

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.

ΛΟΓΙΣΤΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ



η λογιστική παλινδρόμηση χρησιμοποιείται για την πρόβλεψη της πιθανότητας εμφάνισης ενός γεγονότος προσαρμόζοντας τα δεδομένα της μελέτης στην εξίσωση της λογιστικής καμπύλης

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).

ΛΟΓΙΣΤΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ

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.

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

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

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

ΛΟΓΙΣΤΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ

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. 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.

ΛΟΓΙΣΤΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ

Οι πιθανότητες που συγκλίνουν υπέρ της εμφάνισης ενός γεγονότος ή πρόθεσης εκφράζονται ως λόγος ζεύγους ακέραιων τιμών (odds) όπου ο αριθμητής προσδιορίζει την πιθανότητα που έχει το προσδοκώμενο γεγονός να συμβεί και ο παρονομαστής την πιθανότητα να μη συμβεί. Έτσι, αν p είναι η πιθανότητα να εμφανιστεί το γεγονός και 1-p η πιθανότητα να μη συμβεί τότε ο λόγος των πιθανοτήτων θα είναι p/(1-p). Για παράδειγμα, η πιθανότητα να ανασυρθεί μια κάρτα σπαθί από μια τράπουλα 52 φύλλων είναι 25% δηλαδή μία στις τέσσερις ή αριθμητικά 13/52=1:4 ή και ¼. Με ανάλογο τρόπο εκφράζεται και η πιθανότητα μη εμφάνισης μιας κάρτας σπαθί η οποία ισούται με 4:1, αντιστρέφοντας απλώς τους όρους του κλάσματος, (1-p)/p.

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

ΛΟΓΙΣΤΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ

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

$$L=\prod_{i=1}^{n} f(x_i\theta) \quad \Rightarrow \quad L=\sum_{i=1}^{n} \log_e f(x_i\theta)$$

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

Τα λογαριθμικά μοντέλα λόγου (συμπληρωματικών) πιθανοτήτων ή λογιστικά μοντέλα (logit models) χρησιμοποιούνται όταν η εξαρτημένη μεταβλητή είναι δυαδική. Η λογιστική παλινδρόμηση είναι μη γραμμικής μορφής γιατί αναγκάζει τις προβλεπόμενες τιμές να κυμαίνονται μεταξύ 0 και 1.

Τα λογιστικά μοντέλα μοντέλα εκτιμούν την πιθανότητα της εξαρτημένης μεταβλητής να λαμβάνει την τιμή 1 (Y=1), δηλαδή την πιθανότητα ότι κάποιο γεγονός συμβαίνει. Τα μοντέλα αυτά υπακούουν στη συνθήκη πιθανότητας εμφάνισης Pr,

$$Pr\ (Y=1\,|\,X_1, X_2, \ldots X_k)=\frac{1}{1+e^{-z}}$$

ΛΟΓΙΣΤΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ

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.

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,

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

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

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

ΛΟΓΙΣΤΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ

Τόσο τα logit όσο και τα probit μοντέλα παρέχουν παρόμοια αποτελέσματα, διαφέρουν μόνο ως προς την κατανομή των στοιχείων. Τα logit ακολουθούν την αθροιστική τυπική λογιστική κατανομή (F) και τα probit την αθροιστική ομαλή κατανομή (Φ).

Η λέξη logit προέρχεται συγκοπτόμενη από τη φράση logarithmic unit (λογαριθμομονάδα) σε ακολουθία με την πρωταρχικά αποδοθείσα ονοματολογία της λέξης probit (probability unit - πιθανομονάδα) και έχει την έννοια της νεπέρειας (φυσικής) λογαρίθμησης ενός αριθμού p ($\log_e$ ή ln) που εκπροσωπεί πιθανότητα (αναλογία) και άρα τιμές μεταξύ 0 και 1:

$$logit(p) = \log_e\left(\frac{p}{1-p}\right) = \log_e p - \log_e(1-p)$$

Αν p δηλώνει κάποια πιθανότητα τότε η σχέση p/(1-p) αντιστοιχεί στην επιτυχημένη πιθανότητα έκβασης (odds) και κατ΄αντιστοιχία ο λογάριθμος της p στον λογάριθμο της επιτυχημένης πιθανότητας. Με τον ίδιο συλλογισμό, η διαφορά μεταξύ των λογαρίθμων δυο ευνοϊκών πιθανοτήτων p_1 και p_2 αποτελεί και το λογάριθμο του λόγου των ευνοϊκών πιθανοτήτων R σύμφωνα με την ακολουθία των εξισώσεων:

$$\log_e R = \log_e \frac{p_1/(1-p_1)}{p_2/(1-p_2)} = \log_e \frac{p_1}{1-p_1} - \log_e \frac{p_2}{(1-p_2)} = logit(p_1) - logit(p_2)$$

ΛΟΓΙΣΤΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ

Πιθανότητα (επιτυχημένης) έκβασης

Ονομάζεται και προβλεπόμενη πιθανότητα p. Αν οι αποκρίσεις αποτιμώνται ως 0 (αποτυχία) και 1 (επιτυχία), τότε p_j είναι η πιθανότητα όπου η i ανεξάρτητη μεταβλητή (ποσοτική ή κατηγορική) δίνει απόκριση 1,

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

Λογαριθμική πιθανότητα έκβασης

Εφαρμόζεται για την άριστη εκτίμηση των συντελεστών της παλινδρόμησης και επίσης για τη σύγκριση δύο μοντέλων που διαφέρουν ως προς το σύνολο των ανεξάρτητων μεταβλητών σε καθένα από αυτά:

$$L_{(\beta)} = \sum_j [y_j \log_e p_j + (m_j - y_j) \log_e (1 - p_j)]$$

όπου p_j= πιθανότητα επιτυχημένης έκβασης, y_j= απόκριση, m_j= αριθμός προσπαθειών ή παρατηρήσεων σχετιζόμενων με την j ανεξάρτητη μεταβλητή. Αν τα στοιχεία περιέχουν μια παρατήρηση ανά ανεξάρτητη μεταβλητή τότε m_j=1.

ΛΟΓΙΣΤΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ

Συντελεστές β_i

Ο εκτιμώμενος συντελεστής για κάθε ανεξάρτητη μεταβλητή εκφράζει τη μεταβολή του λογαρίθμου του λόγου p/(1-p) για κάθε μονάδα μεταβολής της αντίστοιχης ανεξάρτητης μεταβλητής, ενώ οι λοιπές παραμένουν σταθερές (το αποτέλεσμά τους είναι υπό έλεγχο). Για να βρεθεί η τιμή του β η οποία μεγιστοποιεί τη λογαριθμική πιθανότητα έκβασης $L_{(\beta)}$, η τελευταία διαφορίζεται ως προς β_0 και β_i έτσι ώστε να προκύπτει,

$$\sum_j (y_j - m_j p_j) = 0$$

$$\sum_j X_{ji} (y_j - m_j p_j) = 0$$

Οι εξισώσεις αυτές επειδή δεν είναι γραμμικές, χρησιμοποιείται για την επίλυσή τους η εκτίμηση της μέγιστης πιθανοφάνειας (MLE).

ΛΟΓΙΣΤΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ

Ο παρακάτω πίνακας ενδεχομένων καταρτίστηκε ύστερα από έρευνα αγοράς σχετικής με την προτίμηση (πρόθεση αγοράς) 80 καταναλωτών ως προς δύο προϊόντα Α και Β. Οι κωδικοί αριθμοί 1 και 0 αναφέρονται στην προτίμηση η μη των καταναλωτών και οι τιμές (συχνότητες εμφάνισης), ανάλογα με την επιλογή τους στα συνδυασμένα επίπεδα (κελιά):

		Είδος προϊόντος		
		Α	Β	Σύνολο
Πρόθεση αγοράς	1	36	20	56
	0	14	10	24
Σύνολο		50	30	80

Το είδος προϊόντος στον πίνακα αυτόν αποτελεί το αίτιο πρόκλησης (μεταβλητή Χ, με δύο κατηγορίες (Α και Β)) και η πρόθεση αγοράς την έκβαση του αποτελέσματος (απόκριση Υ), ευνοϊκή (1) ή μη (0).

Οι υπολογισμοί βασίζονται στη χρήση της εξίσωσης,

$$logit(p) = \log_e\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X$$

ΛΟΓΙΣΤΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ

Αν επιθυμούμε την εξέταση μόνο της Y τότε η εξίσωση τροποποιείται σε logit(p)=β_0, και η ολική πιθανότητα εμφάνισης της πρόθεσης αγοράς των προϊόντων θα ισούται με p=56/80=0,70. Η ευνοϊκή έκβαση της απόκρισης Y για όλο τον πληθυσμό υπολογίζεται ως p/(1-p)=0,7/(1-0,7)=2,333, ενώ η λογαριθμική μορφή της θα δίνει logit(p)=$\log_{e(}$2,333)=0,847.

Όταν το ενδιαφέρον στρέφεται στην ολοκληρωμένη εξίσωση logit(p) τότε οι πιθανότητες αγοράς για καθένα από τα δύο προϊόντα (ευνοϊκές εκβάσεις (1)) έχουν ως εξής:

Προϊόν A: p=(36/50)/(14/50)=0,72/0,28=2,57, δηλαδή πιθανότητα 2,57:1 να προτιμηθεί το προϊόν A

Προϊόν B: p=(20/30)/(10/30)=0,67/0,33=2,00, δηλαδή πιθανότητα 2:1 να αγοραστεί το προϊόν B

Ο λόγος θ των δύο ευνοϊκών πιθανοτήτων προκύπτει ως θ=(2,57)/(2,00)=1,285 η οποία δείχνει ότι το προϊόν A έχει 1,285 φορές μεγαλύτερη πιθανότητα (ή 28,5%) να αγοραστεί από τους καταναλωτές από ό,τι το προϊόν B.

Το τυπικό σφάλμα εκτιμάται ως

$$SE=\sqrt{(1/a)+(1/b)+(1/c)+(1/d)}=\sqrt{(1/36)+(1/20)+(1/14)+(1/10)}=\sqrt{0{,}249}=0{,}499$$

και τα 95% όρια εμπιστοσύνης του λόγου θ

$$e^{\log_e(\theta)\pm 1{,}96\cdot SE}=e^{\log_e(1{,}285)\pm 1{,}96\cdot 0{,}499}=e^{0{,}25\pm 0{,}98}$$

ΛΟΓΙΣΤΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ

Τυπικό σφάλμα των συντελεστών

Υπολογίζεται το ασυμπτωτικό τυπικό σφάλμα το οποίο όσο μικρότερη τιμή παρέχει τόσο ακριβέστερη θεωρείται η εκτίμηση. Η στατιστική σημαντικότητα των συντελεστών των ανεξάρτητων μεταβλητών ελέγχεται με δύο κριτήρια:

α) Το κριτήριο του Wald,

$$z = \frac{\beta_i}{SE}$$

Η τιμή z συγκρίνεται με την τιμή 1,96 ή υψούμενη στο τετράγωνο με τη θεωρητική τιμή χ^2 (3,841).

Τιμές του z μεγαλύτερες από 1,96 δείχνουν στατιστική σημαντικότητα της μεταβλητής. Τα 95% όρια εμπιστοσύνης κάθε συντελεστή β_i εξάγονται ως $\beta_i \pm z_{0,05/2} \cdot SE$ και τα αντίστοιχα όρια εμπιστοσύνης του λόγου επιτυχημένης έκβασης υπολογίζονται αντιλογαριθμίζοντας το ανώτερο και κατώτερο της παραπάνω σχέσης. Εντός του εύρους των ορίων εμπιστοσύνης, ο λόγος των πιθανοτήτων αντιπροσωπεύεται πλήρως και ισοδύναμα από οποιαδήποτε τιμή.

Το κριτήριο του Wald προκαλεί διεύρυνση του τυπικού σφάλματος όταν οι συγκρινόμενοι συντελεστές έχουν υψηλή τιμή, μία ιδιότητα καθόλου επιθυμητή, διότι οδηγεί σε πολύ μικρή τιμή του στατιστικού Wald και στην λανθασμένα αποδοχή της σημαντικότητας του εξεταζόμενου συντελεστή (Hauck and Donner 1977). Στις περιπτώσεις αυτές, είναι προτιμότερη η απόπειρα ανάπτυξης κάποιου υποδείγματος με και χωρίς τη συγκεκριμένη - με υψηλό συντελεστή - μεταβλητή και ο έλεγχος της υπόθεσης να στηρίζεται στη μεταβολή -2LL, δηλαδή του λογάριθμου πιθανοφάνειας όπως περιγράφεται παρακάτω.

ΛΟΓΙΣΤΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ

β) Το κριτήριο του λόγου ή λογάριθμου πιθανοφάνειας -2LL (Likelihood ratio statistic), το οποίο ελέγχει ένα μικρότερο μοντέλο S με s συντελεστές και πιθανοφάνεια L_s προς ένα μεγαλύτερο μοντέλο L με l συντελεστές (συνήθως ένα παραπάνω) και πιθανοφάνεια L_l και με τον περιορισμό ότι οι παράμετροι s αποτελούν μέρος του συνόλου των παραμέτρων l:

$$-2\log_e\left(\frac{L_s}{L_l}\right)=-2\left[\log_e(L_s)-\log_e(L_l)\right]=-2(L_s-L_l)$$

Η τιμή του κριτηρίου συγκρίνεται με τη θεωρητική τιμή χ^2