ΤΕΧΝΟΛΟΓΙΚΟ ΙΔΡΥΜΑ ΔΥΤΙΚΗΣ ΕΛΛΑΔΟΣ

ΣΧΟΛΗ ΔΙΟΙΚΗΣΗΣ ΚΑΙ ΟΙΚΟΝΟΜΙΑΣ

ΤΜΗΜΑ ΔΙΟΙΚΗΣΗ ΕΠΙΧΕΙΡΗΣΕΩΝ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

ΜΗ ΓΡΑΜΜΙΚΑ ΜΟΝΤΕΛΑ ΠΑΛΙΝΔΡΟΜΗΣΗΣ

NONLINEAR REGRESSION MODELS

Σπουδαστές: Τεφάνη Ιωάννα

Τσαντός Χαράλαμπος

Εποπτεύων καθηγητής: Μιχοπούλου Μαρία

Πάτρα 2014

Περιεχόμενα

Περίληψη:

Μη γραμμικά μοντέλα παλινδρόμησης

Παρακάτω θα ασχοληθούμε με μοντέλα που δεν ανταποκρίνονται στο γραμμικό πρότυπο, πολυωνυμικά μοντέλα, curve modeling, μοντέλα μη γραμμικά ως προς τις παραμέτρους. Επίσης θα αναπτύξουμε τη μέθοδο ελαχίστων τετραγώνων για τον προσδιορισμό μη γραμμικών μοντέλων. Θα αναφερθούμε σε παραδείγματα και εφαρμογές.

Στόχος μας είναι η μελέτη των μη γραμμικών μοντέλων παλινδρόμησης, πολυωνυμικών μοντέλων, exponential decay model, exponential growth model, two-term exponential model, Τέλος, θα εφαρμόσουμε τα αποτελέσματα σε πραγματικά δεδομένα με χρήση του στατιστικού πακέτου SPSS.

Abstract:

Non linear regression models

Below we will deal with models that do not satisfy the linear model, polynomial models, curve modeling, models nonlinear in the parameters. It will also develop the least squares method for the determination of non-linear models. We will refer to examples and applications.

Our goal is the study of nonlinear regression models, polynomial models, exponential decay model, exponential growth model, two-term exponential model. Finally, we will apply the results to real data using the statistical package SPSS.

Εισαγωγή

Για να ασχοληθούμε με την μελέτη των μοντέλων μη γραμμικής παλινδρόμησης θα πρέπει πρώτα να αναφερθούμε στο απλό γραμμικό μοντέλο.

Για την εφαρμογή στατιστικών μελετών και ιδιαίτερα στην εξέταση οικονομικών, κοινωνικών, πολιτικών, γεωγραφικών και άλλων φαινομένων θα πρέπει να αναγνωριστεί η ταυτόχρονη συμπεριφορά δύο ή περισσότερων ποσοτικών μεταβλητών. Αυτό σημαίνει ότι θα εξετάσουμε αν οι τιμές της μιας μεταβλητής επιδρούν στη διαμόρφωση των τιμών των υπολοίπων μεταβλητών. Αν αυτό είναι εφικτό προσδιορίζουμε μια μαθηματική σχέση, η οποία θα εκφράζει την αλληλεπίδραση αυτή.

Οπότε, σε ένα στατιστικό σύνολο, πληθυσμό ή δείγμα, δύο ποσοτικών μεταβλητών X,Y σε κάθε στατιστικό στοιχείο αντιστοιχεί ένα διατεταγμένο ζεύγος τιμών (x,y). Εδώ θα πρέπει να προσδιοριστεί η συναρτησιακή σχέση μεταξύ των δύο μεταβλητών, αν βέβαια υπάρχει. Στην περίπτωση που υπάρχει η συναρτησιακή αυτή σχέση καθορίζουμε τη μία από τις δύο μεταβλητές ως ανεξάρτητη και την άλλη ως εξαρτημένη μεταβλητή.

Η μαθηματική αυτή σχέση είναι της μορφής:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Σε αυτή την περίπτωση ορίζουμε ως:

Y_i την τιμή της εξαρτημένης μεταβλητής για την i περίπτωση

β_0, β_1 τις παραμέτρους

X_i τη γνωστή σταθερά

ε_i τα τυχαία σφάλματα με μέση τιμή μηδέν $E\{\varepsilon_i\} = 0$ και διασπορά $\sigma^2\{\varepsilon_i\} = \sigma^2$. Τα ε_i, ε_j είναι ασυσχέτιστα ώστε η συνδιασπορά να ισούται με το μηδέν $\sigma^2(\varepsilon_i, \varepsilon_j) = 0$, για κάθε i, j. Άρα τα $\varepsilon_i \sim N(0, \sigma^2)$.

Επίσης, $\sigma^2\{Y_i\} = \sigma^2$ και $E\{Y_i\} = \beta_0 + \beta_1 X_i$.

Στόχος όλων αυτών είναι η ερμηνεία της μεταβλητής Y με τη χρήση άλλων μεταβλητών X. Για την επίτευξη αυτού κατασκευάζουμε το παραπάνω μοντέλο. Η γενική μορφή των μοντέλων αυτών είναι της μορφής:

$$Y_i = f(X_i) + \varepsilon_i$$

(Τα ε_i είναι τυχαία σφάλματα και συμπεριλαμβάνουν όλες τις παραμέτρους που δεν λάβαμε υπόψη.) Αν η παραπάνω σχέση είναι γραμμική τότε και το μοντέλο μας είναι γραμμικό και τα μεγέθη X,Y θα σχετίζονται γραμμικώς αν όλα τα σημεία (X_i, Y_i) τείνουν να βρίσκονται σε ευθεία.

Όπως είναι γνωστό η εξίσωση μιας ευθείας του επιπέδου είναι:

$$Y = \beta_0 + \beta_1 X$$

Αλλά το μοντέλο αυτό συμπεριλαμβάνει και τα σφάλματα. Στόχος μας είναι η ελαχιστοποίηση των σφαλμάτων βρίσκοντας την ευθεία παλινδρόμησης.

Για να είναι εφικτό πρέπει να βρεθούν οι εκτιμήσεις των παραμέτρων β_0,β_1, το οποίο θα γίνει με τη μέθοδο των ελαχίστων τετραγώνων ή με τη μέθοδο μέγιστης πιθανοφάνειας. Οι εκτιμητές β_0,β_1, που είναι πληθυσμιακά μεγέθη, θα συμβολίζουμε με b_0, b_1, που είναι δειγματικά μεγέθη και εκτιμούν τα β_0,β_1.

Οπότε η λύση του μοντέλου με τη χρήση εκτιμητριών συναρτήσεων θα είναι της μορφής:

$Y = b_0 + b_1 X + e.$

Με αποτέλεσμα η ευθεία παλινδρόμησης να είναι:

$\hat{Y} = b_0 + b_1 X.$

Επίσης, μπορούμε να θεωρήσουμε γραμμικά μοντέλα όλα όσα μπορούν να αναχθούν στη γενική μορφή των γραμμικών.

Έχουμε επέκταση του γραμμικού μοντέλου αν αντί για μία μεταβλητή πρόβλεψης έχουμε ρ μεταβλητές. Αποτέλεσμα αυτού είναι το μοντέλο να παίρνει τη μορφή:

$$Y = \beta_0 X_0 + \beta_1 X_1 + \ldots\ldots + \beta_{\rho-1} X_{\rho-1} + \varepsilon$$

και η εξίσωση παλινδρόμησης γίνεται:

$\hat{Y} = b_0X_0 + b_1X_1 + \ldots\ldots + b_{\rho-1}X_{\rho-1}$.

Ένα οποιοδήποτε γραμμικό μοντέλο, ανεξάρτητα το πλήθος των μεταβλητών X που περιέχει, μπορεί να δηλωθεί ως εξής:

$$Y_i = f(X_i, \beta) + \varepsilon_i$$

Όπου X_i το διάνυσμα των παρατηρήσεων των ανεξάρτητων μεταβλητών για την i περίπτωση $X_\iota = (1, X_{i1}, X_{i2}, \ldots, X_{i.p-1})'$, β το διάνυσμα διάστασης (ρ x 1) των συντελεστών παλινδρόμησης και $f(X_i, \beta)$ η αναμενόμενη τιμή $E\{Y_i\}$ που ισούται με $f(X_i, \beta) = X_i'\beta$.

1. Κεφάλαιο 1

1.1. Εισαγωγή στη μη γραμμική παλινδρόμηση

1.1.1. Μη γραμμική παλινδρόμηση

Σε κάποια προβλήματα το διαγνωστικό γράφημα μπορεί να μας υποδείξει ότι η εξάρτηση μιας εξαρτημένης τυχαίας μεταβλητής Y από μια ανεξάρτητη μεταβλητή X είναι κάποιας συγκεκριμένης μη γραμμικής μορφής.

1.1.2. Μη γραμμικά μοντέλα παλινδρόμησης

Τα μη γραμμικά μοντέλα παλινδρόμησης είναι της μορφής:

$$Y_i = f(X_i, \gamma) + \varepsilon \qquad (1.1)$$

Ομοίως στα μη γραμμικά μοντέλα όπως και στα γραμμικά τα σφάλματα έχουν μαθηματική ελπίδα μηδέν, σταθερή διασπορά και είναι ασυσχέτιστα μεταξύ τους.

Εδώ το διάνυσμα των παραμέτρων θα δηλώνεται με γ αντί του β για να ξεχωρίζουμε ότι η συνάρτηση απόκρισης δεν είναι γραμμική ως προς τις παραμέτρους.

1.1.2.1. Εκθετικά μοντέλα παλινδρόμησης

Όταν στο μοντέλο που εξετάζουμε έχουμε μια ανεξάρτητη μεταβλητή τότε αυτό παίρνει τη μορφή:

$$Y_i = f(X_i, \gamma) + \varepsilon \qquad (1.2)$$

όπου γ_0 και γ_1 οι παράμετροι, X_i οι γνωστές σταθερές και ε_i τα ανεξάρτητα σφάλματα που ακολουθούν $N(0, \sigma^2)$.

Η συνάρτηση απόκρισης για το μοντέλο θα είναι :

$$\mathrm{f(X, \gamma)} = \gamma_0 \exp(\gamma_1 X) \qquad (1.3)$$

Να σημειωθεί πως αυτό το μοντέλο δεν είναι γραμμικό ως προς τις παραμέτρους g_0, g_1. Ένα πιο γενικό εκθετικό μοντέλο παλινδρόμησης με μια ανεξάρτητη μεταβλητή, θα έχει μορφή:

$$\mathrm{Y_i} = \gamma_0 + \gamma_1 \exp(\gamma_2 \mathrm{X_i}) + \varepsilon_\mathrm{i} \qquad (1.4)$$

και εδώ τα σφάλματα είναι ανεξάρτητα, ακολουθούν κανονική κατανομή και έχουν σταθερή διασπορά s^2.

Η συνάρτηση απόκρισης για αυτό το μοντέλο θα είναι:

$$\mathrm{f(X, \gamma)} = \gamma_0 + \gamma_1 \exp(\gamma_2 X) \qquad (1.5)$$

1.1.2.2. Λογιστικά Μοντέλα Παλινδρόμησης

Η γενική μορφή του μοντέλου όταν έχουμε μια ανεξάρτητη μεταβλητή και κανονικά κατανεμημένα σφάλματα είναι:

$$Y_{\mathrm{i}} = \frac{\gamma_0}{1 + \gamma_1 \exp(\gamma_2 X_i)} + \varepsilon_i \qquad (1.6)$$

όπου και εδώ τα σφάλματα είναι ανεξάρτητα, ακολουθούν κανονική κατανομή και έχουν σταθερή διασπορά σ^2. Η συνάρτηση απόκρισης θα είναι:

$$f(X, \gamma) = \frac{\gamma_0}{1 + \gamma_1 \exp(\gamma_2 \mathrm{X})} \qquad (1.7)$$

Επαναλαμβάνουμε πως η συνάρτηση απόκρισης δεν είναι γραμμική ως προς τις παραμέτρους $\gamma_0, \gamma_1, \gamma_2$.

Τέλος, αξίζει να αναφερθούμε στο ότι τα λογιστικά μοντέλα χρησιμοποιούνται και όταν η συνάρτηση απόκρισης είναι ποιοτική, με αυτή την περίπτωση θα ασχοληθούμε στο επόμενο κεφάλαιο.

1.1.2.3. Γενική Μορφή Μη Γραμμικών Μοντέλων

Όπως παρατηρήσαμε και στα προηγούμενα παραδείγματα, τα μη γραμμικά μοντέλα είναι σχεδόν ίδια, στη γενική τους μορφή, με τα γραμμικά μοντέλα. Κάθε Y παρατήρηση αξιώνουμε να είναι το άθροισμα των μέσων τιμών $f(X_i, \gamma)$**.**

Τα σφάλματα ε_i συχνά υποθέτουμε πως είναι ανεξάρτητες κανονικές, τυχαίες μεταβλητές με σταθερή διασπορά.

Μια σημαντική αλλαγή που έχουν τα μη γραμμικά μοντέλα με τα γραμμικά είναι πως το πλήθος των παραμέτρων παλινδρόμησης δεν είναι απαραιτήτως συνδεδεμένο με το πλήθος των ανεξάρτητων μεταβλητών του μοντέλου. Στα γραμμικά

μοντέλα παλινδρόμησης αν υπάρχουν $p-1$ ανεξάρτητες μεταβλητές στο μοντέλο, τότε θα υπάρχουν p συντελεστές παλινδρόμησης. Ενώ στο εκθετικό μοντέλο (1.4) υπάρχει μια ανεξάρτητη μεταβλητή αλλά τρεις συντελεστές παλινδρόμησης, ακριβώς το ίδιο συμβαίνει και στο λογιστικό μοντέλο (1.6). Για αυτό το λόγο από εδώ και πέρα, για τα μη γραμμικά μοντέλα, θα δηλώνουμε με q το πλήθος των ανεξάρτητων μεταβλητών, ενώ το πλήθος των παραμέτρων θα εξακολουθήσουμε να το δηλώνουμε με p. Για παράδειγμα στο εκθετικό μοντέλο (1.2) υπάρχουν $p = 2$ παράμετροι παλινδρόμησης και $q = 1$ ανεξάρτητες μεταβλητές.

Όπως προαναφέραμε, τα μη γραμμικά μοντέλα έχουν γενική μορφή:

$$Y_i = f(X_i, \gamma) + \varepsilon_i$$

Όπου $X_i = (X_{i1}, X_{i2}, \dots, X_{iq})'$

και $\gamma_0 = (\gamma_0, \gamma_1, \dots, \gamma_{p-1})'$

Να σημειώσουμε πως τα μη γραμμικά μοντέλα μπορούν να γίνουν γραμμικά χρησιμοποιώντας τον κατάλληλο μετασχηματισμό.

Ας δούμε για παράδειγμα την εκθετική συνάρτηση απόκρισης:

$$\mathrm{f}(\mathrm{X}, \gamma) = \gamma_0[\exp(\gamma_1 X)]$$

που μπορεί να γίνει γραμμική, αρκεί να την λογαριθμίσουμε. Τότε θα έχουμε

$$\log_e \mathrm{f}(\mathrm{X}, \gamma) = \log_e \gamma_0 + \gamma_1 X$$

Η μετασχηματιζόμενη συνάρτηση απόκρισης θα είναι:

$$g(X, \gamma) = \beta_0 + \beta_1 X \qquad (1.8)$$

Όπου $g(X, \gamma) = \log_e \mathrm{f}(\mathrm{X}, \gamma)$, $\beta_0 = \log_e \gamma_0$ και $\beta_1 = \gamma_1$.

1.1.3. Εκτίμηση παραμέτρων παλινδρόμησης

Η εκτίμηση των παραμέτρων μη γραμμικών μοντέλων παλινδρόμησης γίνεται είτε με την μέθοδο ελάχιστων τετράγωνων είτε με τη μέθοδο μέγιστης πιθανοφάνειας.

Δηλαδή δυο μεθόδους τις οποίες χρησιμοποιούμε και στα γραμμικά μοντέλα. Παρακάτω κάνουμε μια αναφορά και στις δυο μεθόδους.

1.1.3.1. Εκτίμηση Μέγιστης Πιθανοφάνειας

Ας θεωρήσουμε την συνάρτηση απόκρισης

$$f(X, \gamma) = \gamma_0[\exp(\gamma_1 X)]$$

Το κριτήριο ελάχιστων τετράγωνων σε αυτή την περίπτωση θα είναι:

$$Q = \sum_{i=1}^{n} [Y_i - \gamma_0 \exp(\gamma_1 X_i)]^2$$

Μπορούμε να διαπιστώσουμε ότι η μέθοδος μέγιστης πιθανοφάνειας μας οδηγεί στο ίδιο κριτήριο όταν τα σφάλματα είναι ανεξάρτητα. Η συνάρτηση πιθανοφάνειας θα είναι:

$$L(\gamma, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{\pi/2}} \exp[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} [[Y_i - \gamma_0 \exp(\gamma_1 X_i)]^2]] \quad (1.9)$$

Η πιο πάνω συνάρτηση είναι γνησίως φθίνουσα ως προς το μέτρο Q, άρα η συνάρτηση γίνεται μέγιστη όταν η ποσότητα Q γίνεται ελάχιστη. Οπότε η εκτίμηση μέγιστης πιθανοφάνειας είναι η ίδια με την εκτίμηση ελάχιστων τετράγωνων.

1.1.3.2. Εκτίμηση Ελάχιστων Τετραγώνων

Όπως ξέρουμε από την γραμμική παλινδρόμηση, σκοπός μας είναι η ελαχιστοποίηση της ποσότητας Q όπου:

$$Q = \sum_{i=1}^{n} [Y_i - \beta_0 + \beta_1 X_i]^2$$

Οι τιμές των β_0 και β_1 που ελαχιστοποιούν το Q είναι οι εκτιμητές ελάχιστων τετράγωνων και δηλώνονται με b_0 και b_1. Χρησιμοποιώντας αριθμητικές διαδικασίες είναι ένας τρόπος για να βρούμε τους εκτιμητές ελάχιστων τετράγωνων. Ένας δεύτερος

τρόπος για να βρούμε τους εκτιμητές ελάχιστων τετράγωνων είναι τα ελάχιστα τετράγωνα κανονικών εξισώσεων, εδώ τα ελάχιστα τετράγωνα κανονικών εξισώσεων, βρίσκονται αναλυτικά παραγωγίζοντας το Q ως προς το β_0 και β_1 και θέτοντας τις παραγώγους ίσες με μηδέν. Η λύση των κανονικών εξισώσεων δίνει τις εκτιμήσεις ελάχιστων τετράγωνων. Ας δούμε όμως πιο αναλυτικά αυτή τη διαδικασία.

1.1.3.3. Λύση Κανονικών Εξισώσεων

Χρειαζόμαστε να ελαχιστοποιήσουμε το κριτήριο ελάχιστων τετράγωνων Q

$$Q = \sum_{i=1}^{n} [Y_i - f(X_i, \gamma)]^2$$

με εκτίμηση των $\gamma_0, \gamma_1, \dots, \gamma_{p-1}$.

Η μερική παράγωγος του Q θα είναι :

$$\frac{\partial Q}{\partial \gamma_k} = \sum_{i=1}^{n} -2[Y_i - f(X_i, \gamma)] [\frac{\partial f(X_i, \gamma)}{\partial \gamma_k}] \qquad (1.10)$$

Αν οι μερικοί παράγωγοι είναι η κάθε μια ίση με το μηδέν και οι παράμετροι γ_k αντικατασταθούν από τους εκτιμητές ελάχιστων τετράγωνων g_k, πετυχαίνουμε μετά από απλοποίηση τις p κανονικές εξισώσεις.

$$\sum_{i=1}^{n} Y_i [\frac{\partial f(X_i, \gamma)}{\partial \gamma_k}]_{\gamma=g} - \sum_{i=1}^{n} f(X_i, g)[\frac{\partial f(X_i, \gamma)}{\partial \gamma_k}]_{\gamma=g} = 0 \qquad (1.11)$$

Όπου $k = 0,1, \dots, p-1$.

Όπου g είναι το διάνυσμα των εκτιμώμενων ελάχιστων τετράγωνων $g = (g_o, g_1, \dots, g_{p-1})'$

Λόγω της μη γραμμικής φύσης της συνάρτησης απόκρισης, το σύστημα των κανονικών εξισώσεων είναι μη γραμμικό και συνήθως δεν υπάρχει λύση κλειστής μορφής. Οπότε για να επιλυθεί επιστρατεύονται αριθμητικές μέθοδοι για επαναληπτική εύρεση των εκτιμητών ελάχιστων τετραγώνων.

1.1.3.4. Αριθμητικές διαδικασίες (Μέθοδος Gauss - Newton)

Σε πολλά από τα προβλήματα της μη γραμμικής παλινδρόμησης είναι πιο πρακτικό να βρούμε τις εκτιμήσεις ελάχιστων τετράγωνων με άμεσες αριθμητικές διαδικασίες εύρεσης, από το να βρίσκουμε πρώτα τις κανονικές εξισώσεις και μετά να χρησιμοποιούμε αριθμητικές μεθόδους για να βρούμε τις λύσεις για τις εξισώσεις αυτές. Τα ποιο σημαντικά στατιστικά πακέτα υπολογιστών χρησιμοποιούν μια ή περισσότερες άμεσες αριθμητικές διαδικασίες, για την επίλυση μη γραμμικών προβλημάτων παλινδρόμησης, μια αριθμητική διαδικασία θα αναλύσουμε παρακάτω.

Η μέθοδος Gauss Newton που λέγεται και μέθοδος γραμμικοποίησης χρησιμοποιεί επέκταση σε σειρά Taylor για να προσεγγίσει το μη γραμμικό μοντέλο με γραμμικούς όρους και μετά χρησιμοποιεί τα ελάχιστα τετράγωνα για να εκτιμήσει τις παραμέτρους.

Η μέθοδος Gauss Newton ξεκινάει με αρχικές τιμές για τις παραμέτρους παλινδρόμησης $\gamma_0,\ \gamma_1, \dots, \gamma_{p-1}$ τις οποίες δηλώνουμε με $g_0^{(0)}, g_1^{(0)}, \dots, g_{p-1}^{(0)}$ όπου ο άνω δείκτης δηλώνει τον αριθμό των επαναλήψεων. Οι αρχικές τιμές μπορούν να αποκτηθούν από προηγούμενες ή σχετικές μελέτες, θεωρητικές προσδοκίες ή προκαταρκτική εξέταση για τιμές παραμέτρων που οδηγούν σε συγκριτικά χαμηλή τιμή της ποσότητας Q. Αφού αποκτήσουμε τις αρχικές τιμές για τις παραμέτρους προσεγγίζουμε τις μέσες τιμές για τις n περιπτώσεις από τους γραμμικούς όρους του αναπτύγματος Taylor γύρω από τις αρχικές τιμές. Για την i περίπτωση θα έχουμε:

$$f(X_i,\gamma) = f\left(X_i, g^{(0)}\right) + \sum_{k=0}^{p-1}\left[\frac{\partial f(X_i,\gamma)}{\partial \gamma_k}\right]_{\gamma=g^{(0)}} (\gamma_k - g_k^{(0)}) \quad (1.12)$$

Όπου:

$$g^{(0)} = \begin{pmatrix} g_0^{(0)} \\ g_1^{(0)} \\ : \\ . \\ g_{p-1}^{(0)} \end{pmatrix} \tag{1.13}$$

Σημειώνουμε πως $g^{(0)}$είναι το διάνυσμα των αρχικών τιμών των παραμέτρων. Στην παραπάνω σχέση οι όροι στις παρενθέσεις είναι οι ίδιες μερικές παράγωγοι της συνάρτησης παλινδρόμησης που αντιμετωπίσαμε νωρίτερα στις κανονικές εξισώσεις, αλλά εδώ υπολογίζονται στο $\gamma_k = g_k^{(0)}$, για $k = 0,1,\dots,p-1$.

Ας απλοποιήσουμε, τους συμβολισμούς με τον εξής τρόπο:

$$f_i^{(0)} = f(X_i, g^{(0)}) \tag{1.14}$$

$$\beta_k^{(0)} = \gamma_k - g_k^{(0)} \tag{1.15}$$

$$D_{ik}^{(0)} = \left[\frac{\partial f(X_i,\gamma)}{\partial \gamma_k}\right]_{\gamma=g^{(0)}} \tag{1.16}$$

Η προσέγγιση Taylor (1.12) για τη μέση τιμή για την i περίπτωση θα γίνεται:

$$f(X_i,\gamma) = f_i^{(0)} + \sum_{k=0}^{p-1} D_{ik}^{(0)}\beta_k^{(0)}$$

Και μια προσέγγιση για το μη γραμμικό $Y_i = f(X_i,\gamma) + \varepsilon_i$ είναι

$$Y_i = f_i^{(0)} + \sum_{k=0}^{p-1} D_{ik}^{(0)}\beta_k^{(0)} + \varepsilon_i \tag{1.17}$$

όταν μετατοπίζουμε τον όρο $f_i^{(0)}$ προς τα αριστερά και θεωρήσουμε τη διάφορα $Y_i - f_i^{(0)}$ αποκτούμε την ακόλουθη προσέγγιση :

$$Y_i = \sum_{k=0}^{p-1} D_{ik}^{(0)} \beta_k^{(0)} + \varepsilon_i \qquad (1.18)$$

όπου $Y_i^{(0)} \sim Y_i - f_i^{(0)}$.

Να υπογραμμίσουμε ότι η προσέγγιση γραμμικού μοντέλου παλινδρόμησης είναι της μορφής:

$$Y_i = \beta_0 X_{i0} + \beta_1 X_{i1} + \cdots + \beta_{p-1} X_{i.p-1} + \varepsilon_i$$

Οι αποκρίσεις στην (1.18) είναι υπόλοιπα, συγκεκριμένα οι αποκλίσεις των παρατηρήσεων γύρω από την συνάρτηση μη γραμμικής παλινδρόμησης με τις παραμέτρους να έχουν αντικατασταθεί από τις αρχικές εκτιμήσεις. Οι παρατηρήσεις $D_{ik}^{(0)}$ των X μεταβλητών είναι οι μερικές παράγωγοι της μέσης τιμής εκτιμώμενη για κάθε μια από τις n περιπτώσεις με τις παραμέτρους να έχουν αντικατασταθεί από τις αρχικές εκτιμήσεις.

Κάθε συντελεστής παλινδρόμησης $\beta_k^{(0)}$ αντιπροσωπεύει την διαφορά μεταξύ της πραγματικής παραμέτρου παλινδρόμησης και της αρχικής εκτίμησης της παραμέτρου, έτσι οι συντελεστές παλινδρόμησης αναλογούν στα ποσά προσαρμογής με τα οποία πρέπει να διορθωθούν οι αρχικοί συντελεστές παλινδρόμησης.

Πλέον θα απεικονίσουμε την προσέγγιση γραμμικού μοντέλου παλινδρόμησης σε μορφή πινάκων ως:

$$Y^0 = D^0 \beta^0 + \varepsilon \qquad (1.19)$$

Όπου:

$$Y^{(0)} = \begin{pmatrix} Y_1 - f_1^{(0)} \\ Y_2 - f_2^{(0)} \\ \vdots \\ Y_n - f_n^{(0)} \end{pmatrix} \qquad D^{(0)} = \begin{pmatrix} D_{1,0}^{(0)}, \dots, D_{1,p-1}^{(0)} \\ D_{2,0}^{(0)}, \dots, D_{2p-1}^{(0)} \\ \vdots \\ D_{n,0}^{(0)}, \dots, D_{n,p-1}^{(0)} \end{pmatrix}$$

$$\beta^{(0)} = \begin{pmatrix} \beta_0^{(0)} \\ \beta_1^{(0)} \\ \vdots \\ \beta_{p-1}^{(0)} \end{pmatrix} \qquad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Να λάβουμε υπόψη μας ότι το μοντέλο προσέγγισης στην (1.19) είναι ακριβώς στην μορφή του γενικού μοντέλου γραμμικής παλινδρόμησης με τον D πίνακα μερικών παραγώγων να παίζει τον ρόλο του X πίνακα αλλά χωρίς μια στήλη μονάδων για την παρεμβολή. Μπορούμε λοιπόν να υπολογίσουμε τις παραμέτρους $\beta^{(0)}$με ελάχιστα τετράγωνα και να πάρουμε:

$$b^{(0)} = (D^{(0)\prime}D^{(0)})^{-1}D^{(0)\prime}Y^{(0)} \qquad (1.20)$$

Όπου $b^{(0)}$ είναι το διάνυσμα των εκτιμωμένων συντελεστών παλινδρόμησης.

Μπορούμε να μεταχειριζόμαστε ένα πρόγραμμα παλινδρόμησης στον υπολογιστή για να υπολογίσουμε τους εκτιμώμενους συντελεστές παλινδρόμησης $b_k^{(0)}$.

Στην συνέχεια χρησιμοποιούμε τις εκτιμήσεις ελάχιστων τετραγώνων για να βρούμε τους επαναλαμβανόμενους εκτιμωμένους συντελεστές παλινδρόμησης $g_k^{(1)}$ μέσω της (1.15).

$$g_k^{(1)} = g_k^{(0)} + b_k^{(0)} \qquad (1.21)$$

Σε μορφή πινάκων φανερώνουμε την επαναληπτική διαδικασία ως εξής :

$$g^{(1)} = g^{(0)} + b^{(0)} \qquad (1.22)$$

Αυτή την στιγμή μπορούμε να συγκεντρώσουμε πληροφορίες κατά πόσο οι επαναλαμβανόμενοι συντελεστές παλινδρόμησης αντιπροσωπεύουν ρυθμίσεις στην κατάλληλη κατεύθυνση. Θα παραστήσουμε το κριτήριο ελάχιστων τετραγώνων Q, εκτιμώμενο για τους αρχικούς συντελεστές παλινδρόμησης $g^{(0)}$, με $SSE^{(0)}$, θα είναι :

$$SSE^{(0)} = \sum_{i=1}^{n}[Y_i - f(X_i, g^{(0)})]^2 = \sum_{i=1}^{n}(Y_i - f_i^{(0)})^2 \qquad (1.23)$$

Στο τέλος της πρώτης επανάληψης, οι επαναλαμβανόμενοι συντελεστές παλινδρόμησης θα είναι $g^{(1)}$ και το κριτήριο ελάχιστων τετραγώνων θα συμβολίζεται με $SSE^{(1)}$.

$$SSE^{(1)} = \sum_{i=1}^{n} [Y_i - f(X_i, g^{(1)})]^2 = \sum_{i=1}^{n} (Y_i - f_i^{(1)})^2$$

Αν η μέθοδος GAUSS – NEWTON δουλεύει αποτελεσματικά στην πρώτη επανάληψη το $SSE^{(1)}$ θα πρέπει να είναι μικρότερο από το $SSE^{(0)}$.

Στη συνεχεία διεξάγεται εκ νέου το κεντρικό τμήμα του αλγορίθμου με καινούριες τιμές παραμέτρων παλινδρόμησης αυτές που υπολογίστηκαν στο τέλος της πρώτης επανάληψης. Έτσι δημιουργείται ένα καινούριο ζεύγος εκτιμήσεων και μέτρου ελάχιστων τετραγώνων. Ο αλγόριθμος τερματίζει, όταν η διαφορά δύο διαδοχικών εκτιμήσεων συντελεστών είναι αμελητέα ή όταν η διαφορά μεταξύ δύο διαδοχικών μέτρων κριτηρίου ελάχιστων τετραγώνων είναι αμελητέα. Δηλαδή ανάλογα με την επιθυμητή ακρίβεια προσέγγισης καθορίζουμε ένα κατώφλι ακρίβειας, επανάληψη του αλγορίθμου πέρα από αυτό το σημείο απλά αυξάνει την υπολογιστική πολυπλοκότητα χωρίς ουσιαστικά αποτελέσματα. Πιο παραστατικά:

$$\left|g^{(s+1)} - g^{(s)}\right| \leq THRESHOLD$$
$$\left|SSE^{(s+1)} - SSE^{(s)}\right| \leq THRESHOLD'$$

1.1.3.5. Έλεγχοι – κατασκευή διαστήματος εμπιστοσύνης

Αυστηρή τήρηση στις διαδικασίες διεξαγωγής συμπερασμάτων σχετικά με τις παραμέτρους παλινδρόμησης είναι διαθέσιμες για κάθε μέγεθος δείγματος στην περίπτωση των γραμμικών μοντέλων με σφάλματα κανονικής κατανομής. Δε συμβαίνει το ίδιο και με τα μη γραμμικά μοντέλα κανονικών σφαλμάτων, όπου οι εκτιμητές ελάχιστων τετραγώνων και μέγιστης πιθανοφάνειας δεν είναι κανονικά κατανεμημένοι, unbiased και έχουν ελάχιστη απόκλιση για κάθε μέγεθος δείγματος. Συνεπώς, η διεξαγωγή συμπερασμάτων βασίζεται αποκλειστικά στη λεγόμενη θεωρία μεγάλου δείγματος (large-sample theory). Σύμφωνα με αυτήν οι εκτιμητές για μη γραμμικά μοντέλα με κανονικά σφάλματα, όταν το μέγεθος δείγματος είναι μεγάλο, είναι περίπου

κανονικά κατανεμημένοι, unbiased και έχουν ελάχιστη απόκλιση Συνεπώς, η διεξαγωγή συμπερασμάτων βασίζεται αποκλειστικά στη λεγόμενη θεωρία μεγάλου δείγματος (large sample theory). Σύμφωνα με αυτήν οι εκτιμητές για μη γραμμικά μοντέλα με κανονικά σφάλματα, όταν το μέγεθος δείγματος είναι μεγάλο, είναι περίπου κανονικά κατανεμημένοι, unbiased και έχουν ελάχιστη απόκλιση.

Βασικό ερώτημα είναι το πότε η large-sample theory είναι εφαρμόσιμη, δηλαδή πότε το μέγεθος δείγματος είναι αρκετά μεγάλο ώστε για κάθε συνάρτηση παλινδρόμησης να είναι κατάλληλη η ασυμπτωτική προσέγγιση του θεωρήματος. Δυστυχώς, δεν υπάρχει γενική λύση καθώς υπάρχει εκτός των άλλων και έντονη εξάρτηση από το ποσοστό μη γραμμικότητας της συνάρτησης παλινδρόμησης. Υπάρχει, όμως, ένας αριθμός κανόνων και οδηγιών που αναπτύχθηκαν ώστε να εκτιμηθεί πόσο κατάλληλη είναι η large-sample theory για κάθε εφαρμογή:

1. Γρήγορη σύγκλιση της επαναληπτικής μεθόδου Gauss-Newton στην εύρεση εκτιμήσεων των παραμέτρων σημαίνει ότι έχουμε σωστή γραμμική προσέγγιση άρα είναι εφαρμόσιμες οι ασυμπτωτικές ιδιότητες των εκτιμητών παλινδρόμησης.
2. Χρήση διαφόρων ποσοτικών μέτρων ελέγχου της καταλληλότητας της large-sample theory. Έχουν εισαχθεί μέτρα μη γραμμικότητας της συνάρτησης παλινδρόμησης, μέτρα bias των υπολογισμένων συντελεστών παλινδρόμησης, μέτρα κυρτότητας (skewness) των κατανομών δειγματοληψίας των υπολογισμένων συντελεστών παλινδρόμησης. Εάν τα μέτρα αυτά έχουν σχετικά μικρές τιμές τότε είναι δυνατή η αξιόπιστη χρήση της large-sample theory.
3. Χρήση bootstrap-sampling μεθόδων για εξέταση του εάν οι κατανομές δειγματοληψίας είναι περίπου κανονικές, εάν οι αποκλίσεις των κατανομών δειγματοληψίας είναι κοντά στις αποκλίσεις για το μοντέλο γραμμικής προσέγγισης, εάν το bias σε κάθε εκτιμητή παραμέτρου είναι σχετικά μικρό. Στην περίπτωση που ισχύουν τα παραπάνω, η δειγματοληπτική συμπεριφορά των μη γραμμικών εκτιμητών παλινδρόμησης λέγεται ότι είναι κοντά στη γραμμική (close-to-linear) και μπορεί να χρησιμοποιηθεί η large-sample theory.

Μπορούμε, πλέον, να αναφέρουμε το θεώρημα μεγάλου δείγματος, όπως αυτό διατυπώνεται στην περίπτωσή μας. Όταν οι όροι σφάλματος είναι ανεξάρτητοι μεταξύ

τους, ακολουθούν κανονική κατανομή μηδενικής μέσης τιμής και σταθερής απόκλισης και το μέγεθος του δείγματος είναι αρκετά μεγάλο, τότε:

1. Η κατανομή δειγματοληψίας του διανύσματος g είναι προσεγγιστικά κανονική και η αναμενόμενη τιμή του μέσου διανύσματος είναι περίπου: $E\{g\} = \gamma$.
2. Ο προσεγγιστικός πίνακας απόκλισης των παραμέτρων παλινδρόμησης εκτιμάται από τη σχέση: $s^2[g] = MSE(D'D)^{-1}$. Όπου ο D είναι ο πίνακας μερικών παραγώγων της συνάρτησης παλινδρόμησης υπολογισμένων πάνω στις τελικές εκτιμήσεις ελάχιστων τετραγώνων.

Οπότε, όταν το μέγεθος δείγματος είναι μεγάλο και οι όροι σφάλματος ανεξάρτητοι, κανονικοί, με σταθερή απόκλιση τότε οι εκτιμητές ελάχιστων τετραγώνων είναι περίπου unbiased και κανονικά κατανεμημένοι. Επίσης, έχουν ελάχιστη απόκλιση καθώς αυτή προκύπτει από τον παραπάνω τύπο. Αξίζει να σημειωθεί ότι το θεώρημα αυτό ισχύει ακόμα και όταν οι όροι σφάλματος δεν είναι κανονικά κατανεμημένοι.

Βασική συνέπεια του θεωρήματος μεγάλου δείγματος είναι ότι συμπεράσματα για μη γραμμικές παραμέτρους παλινδρόμησης εξάγονται με τον ίδιο τρόπο με αυτά της γραμμικής παλινδρόμησης. Έτσι χρησιμοποιούνται οι γνωστοί τύποι διαστήματος εμπιστοσύνης και ελέγχου υποθέσεων. Βέβαια, εδώ οι διαδικασίες συμπεράσματος είναι μόνο προσεγγιστικές, όταν εφαρμόζονται σε μη γραμμική παλινδρόμηση, αλλά συχνά η προσέγγιση μπορεί να είναι αρκετά καλή (ανάλογα με το μέγεθος του δείγματος και τη μη γραμμικότητα της συνάρτησης παλινδρόμησης).

Προτού προχωρήσουμε στα διαστήματα εμπιστοσύνης αξίζει να αναφερθούμε στις ενέργειες που μπορούμε να κάνουμε σε περίπτωση που η θεωρία μεγάλου δείγματος δεν είναι εφαρμόσιμη.

1. Μετασχηματισμός των παραμέτρων της συνάρτησης παλινδρόμησης, δηλαδή αλλαγή της μαθηματικής της μορφής, ώστε να συμπεριφέρεται καλύτερα κατά την αλγοριθμική διαδικασία.

Π.χ.

$$Y_i = \frac{\gamma_o X_i}{\gamma_1 + X_i} + \varepsilon_i = \frac{X_i}{\frac{\gamma_1}{\gamma_o} + \frac{1}{\gamma_o} X_i} + \varepsilon_i = \frac{X_i}{\theta_1 X_i + \theta_2} + \varepsilon_i$$

με τις κατάλληλες αντικαταστάσεις. Ενώ η πρώτη μορφή εμφανίζει προβλήματα για μέσο αριθμό δείγματος, η ισοδύναμη δεύτερη μορφή δεν εμφανίζει τέτοια προβλήματα για τον ίδιο αριθμό δείγματος.

2. Χρήση bootstrap εκτιμητών διαστημάτων εμπιστοσύνης αντί για χρήση συμπερασμάτων μεγάλου δείγματος. Ενδεχομένως να εμφανιστούν προβλήματα όπως η αργή αλγοριθμική σύγκλιση ή η δυσκολία εύρεσης διαστημάτων.
3. Αύξηση, αν αυτό είναι δυνατόν, του μεγέθους του δείγματος.

Τέλος, για να εξαχθούν συμπεράσματα για τις παραμέτρους μη γραμμικής παλινδρόμησης απαιτείται εκτίμηση της απόκλισης του όρου σφάλματος. Όπως και στη γραμμική παλινδρόμηση πρόκειται για το μέσο όρο των τετραγώνων των υπολοίπων:

$$MSE = \frac{SSE}{n-p} = \frac{\sum(Y_i - \hat{Y}_i)^2}{n-p} = \frac{\sum(Y_i - f(X_i, g))^2}{n-p}$$

όπου g είναι η τελευταία εκτίμηση του αλγορίθμου, $n-p$ είναι οι βαθμοί ελευθερίας καθώς πρέπει να εκτιμηθούν p παράμετροι. Μάλιστα για τη μη γραμμική παλινδρόμηση το MSE δεν είναι unbiased εκτιμητής της απόκλισης, αλλά το bias είναι μικρό όταν έχουμε μεγάλο δείγμα.

1.1.3.6. Κατασκευή διαστήματος εμπιστοσύνης

Έχοντας ως δεδομένο πως τα σφάλματα ακολουθούν κανονική κατανομή και πως το δείγμα μας είναι μεγάλο, θα βρούμε ένα διάστημα εμπιστοσύνης για την παράμετρο γ_k. Θεωρώντας ότι έχουμε επίπεδο εμπιστοσύνης $1-a$ και πως $g_k \sim N(\gamma_k, s^2\{g_k\})$ αν στην συνέχεια τυποποιήσουμε θα έχουμε $\frac{g_k-\gamma_k}{s\{g_k\}} \sim t(n-p)$ όπου $t(n-p)$ είναι μια t μεταβλητή με $n-p$ βαθμούς ελευθερίας. Το διάστημα εμπιστοσύνης για την παράμετρο γ_k θα είναι:

$$\gamma_k : g_k \pm t\left(1-\frac{a}{2}, n-p\right) s\{g_k\}$$

Πριν ολοκληρώσουμε την αναφορά μας στα διαστήματα εμπιστοσύνης θα αναφέρουμε μια μέθοδο εύρεσης από κοινού διαστήματος εμπιστοσύνης για διαφορετικές παραμέτρους, αυτή βασίζεται στην διαδικασία Bonferroni. Εάν έχουμε m παραμέτρους να εκτιμήσουμε, με κατά προσέγγιση συντελεστή εμπιστοσύνης $1-a$, τα από κοινού όρια εμπιστοσύνης θα είναι :

$$\gamma_k: g_k \pm Bs\{g_k\}$$

όπου: $B = t(1-\frac{a}{2m}, n-p.$

1.1.3.7. Έλεγχος που αφορά την παράμετρο γ_k

Ο έλεγχος μας θα είναι :

$$H_0: \gamma_k = \gamma_{k0}$$
$$H_a: \gamma_k \neq \gamma_{k0}$$

Θα χρησιμοποιήσουμε το στατιστικό t* test, με την προϋπόθεση ότι το n είναι μεγάλο:

$$t^* = \frac{g_k - \gamma_{k0}}{s\{g_k\}}$$

Ο κανόνας απόφασης θα είναι:

Αν: $|t^*| \leq t(1-\frac{a}{2}, n-p)$ διάλεξε την H_0

Αν: $|t^*| > t(1-\frac{a}{2}, n-p)$ διάλεξε την H_a

1.1.3.8. Εγγενής γραμμική συνάρτηση παλινδρόμησης

Εγγενής γραμμική συνάρτηση είναι η μη γραμμική συνάρτηση που με κατάλληλο μετασχηματισμό γίνεται γραμμική.

Γνωστές εγγενής γραμμικές συναρτήσεις:

Εγγενής συνάρτηση	Μετασχηματισμός	Γραμμική συνάρτηση
Εκθετική: $y = ae^{\beta x}$	$y' = \ln y$	$y' = \ln a + \beta x$
Δύναμης: $y = ae^{\beta}$	$y' = \log y, x' = \log x$	$y' = \log a + \beta x'$
$y = a + \beta \log x$	$x' = \log x$	$y = a + \beta x'$
Αντίστροφη: $y = a + \beta \frac{1}{x}$	$x' = \frac{1}{x}$	$y = a + \beta x'$

Στην περίπτωση που γνωρίζουμε ότι η μορφή της συνάρτησης δεν είναι μια οποιαδήποτε μη γραμμική συνάρτηση αλλά εγγενής γραμμική, έχουμε το βασικό πλεονέκτημα ότι μπορούμε να εκτιμήσουμε τις παραμέτρους της συνάρτησης με τη μέθοδο των ελαχίστων τετραγώνων όπως και στη γραμμική συνάρτηση. Αυτό γίνεται διότι η συνάρτηση του αθροίσματος των τετραγώνων των σφαλμάτων είναι γραμμική ως προς τις παραμέτρους.

1.1.3.9. Πολυωνυμική συνάρτηση

Κοινό χαρακτηριστικό των μη γραμμικών μοντέλων παλινδρόμησης που παίρνουμε από εγγενείς γραμμικές συναρτήσεις της εξαρτημένης μεταβλητής y προς την ανεξάρτητη μεταβλητή x είναι ότι οι συναρτήσεις αυτές είναι μονότονες, αύξουσες ή φθίνουσες. Συνήθως η θεωρητική προσέγγιση ή το διάγραμμα διασποράς συνιστά ότι η συνάρτηση έχει ένα ή περισσότερα σημεία καμπής. Σε αυτή την περίπτωση η πολυωνυμική συνάρτηση κάποιου βαθμού k μπορεί να αποτελεί ικανοποιητική προσέγγιση της πραγματικής συνάρτησης παλινδρόμησης.

1.1.3.10. Μοντέλο πολυωνυμικής γραμμικής παλινδρόμησης βαθμού k

(k-th degree polynomial regression model)

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_k x^k + \varepsilon$$

Έτσι και εδώ όπως και στη γραμμική παλινδρόμηση υποθέτουμε ότι τα σφάλματα της παλινδρόμησης ακολουθούν κανονική κατανομή με μέση τιμή 0 και διασπορά σ^2. Μέσω αυτής της υπόθεσης μπορούμε να εκτιμήσουμε διαστήματα εμπιστοσύνης και να ελέγξουμε τις παραμέτρους του μοντέλου και να εκτιμήσουμε διαστήματα πρόβλεψης. Όμως, η μέθοδος ελαχίστων τετραγώνων δεν προϋποθέτει κανονικότητα των σφαλμάτων για να πάρουμε τις καλύτερες εκτιμήσεις των παραμέτρων.

Όπως και στο γραμμικό μοντέλο έτσι και εδώ η εκτίμηση των παραμέτρων γίνεται με τη μέθοδο ελαχίστων τετραγώνων. Αυτό συμβαίνει γιατί ενώ η πολυωνυμική συνάρτηση παλινδρόμησης είναι μη γραμμική ως προς την ανεξάρτητη μεταβλητή x είναι γραμμική ως προς τους συντελεστές $\beta_0, \beta_{1,\ldots,} \beta_k$. Το άθροισμα των τετραγώνων των σφαλμάτων για κάποιο διμεταβλητό δείγμα μεγέθους n των (X, Y) δίνεται από τον τύπο:

$$f(\beta_0, \beta_1, \ldots, \beta_k) = \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_k x_i^k)]^2$$

Με το σύστημα κανονικών εξισώσεων, που δίνεται από τους μερικούς παράγωγους της συνάρτησης αυτής ως προς κάθε παράμετρο $\beta_0, \beta_1, \ldots, \beta_k$ βρίσκουμε τις εκτιμήσεις $b_0, b_1, \ldots, b_k$.

$$b_0 n + b_1 \sum x_i + b_2 \sum x_i^2 + \cdots b_k \sum x_i^k = \sum y_i$$

$$b_0 \sum x_i + b_1 \sum x_i^2 + b_2 \sum x_i^3 + \cdots + b_k \sum x_i^{k+1} = \sum x_i y_i$$

$$\begin{array}{ccc} . & . & . \\ . & . & . \\ . & . & . \end{array}$$

$$b_0 \sum x_i^k + b_1 \sum x_i^{k+1} + b_2 \sum x_i^{k+2} + \cdots + b_k \sum x_i^{2k} = \sum x_i^k y_i$$

$e_i = y_i - \hat{y}_i$, τα σφάλματα του μοντέλου πολυωνυμικής παλινδρόμησης που εκτιμήθηκε με τη μέθοδο ελαχίστων τετραγώνων.

Όπου $\hat{y}_i = b_0 + b_1 x_i + b_2 x_i^2 + \cdots + b_k x_i^k$.

Η εκτίμηση της διασποράς των σφαλμάτων e_i δίνεται από τον τύπο:

$s_e^2 = \frac{1}{n-(k+1)} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$.

Τον συντελεστή του πολλαπλού προσδιορισμού R^2, που δηλώνει την αναλογία της μεταβλητότητας που εξηγείται από το μοντέλο, παίρνουμε από τον τύπο:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

1.1.3.11. Άλλα μη-γραμμικά μοντέλα

Υπάρχουν και άλλες κλάσεις μοντέλων που δεν έχουν κάποια γνωστή αναλυτική μορφή αλλά δίνονται σαν άθροισμα διαφορετικών βασικών συναρτήσεων, όπως τα νευρωνικά δίκτυα (neural networks). Τέλος υπάρχουν και μη παραμετρικά μοντέλα που κάνουν εκτίμηση ή πρόβλεψη για τις δεδομένες τιμές των ανεξάρτητων μεταβλητών χρησιμοποιώντας από τα υπάρχοντα δεδομένα αυτά που είναι «γειτονικά». Τέτοια είναι τα μοντέλα πυρήνων (kernels).

2. Κεφάλαιο 2

2.1.Λογιστική παλινδρόμηση

Στην ενότητα αυτή θα ασχοληθούμε με μοντέλα μη γραμμικής παλινδρόμησης όπου οι τιμές της συνάρτησης απόκρισης είναι διακριτές και τα σφάλματα δεν είναι κατανεμημένα κανονικά.

Αρχικά, θα ασχοληθούμε με το λογιστικό μη γραμμικό μοντέλο παλινδρόμησης.

Η ανεξάρτητη μεταβλητή σε αυτό το μοντέλο είναι ποιοτική με δυο πιθανά αποτελέσματα. Π.χ. η πίεση του αίματος που μπορεί να είναι υψηλή ή χαμηλή. Αυτό το μοντέλο μπορεί να επεκταθεί όταν η ποιοτική μεταβλητή έχει πάνω από δυο πιθανά αποτελέσματα, για παράδειγμα η πίεση του αίματος μπορεί να ταξινομηθεί ως υψηλή, κανονική, χαμηλή. επιπροσθέτως θα μιλήσουμε και για το Poisson μοντέλο παλινδρόμησης, όπου και αυτό είναι μη γραμμικό μοντέλο και η μεταβλητή απόκρισης είναι μια ποσοτική μεταβλητή, όπου μεγάλες τιμές είναι ένα σπάνιο γεγονός.

Τα προαναφερθέντα μη γραμμικά μοντέλα παλινδρόμησης χρησιμοποιούνται πολύ στην ανάλυση δεδομένων που προκύπτουν είτε από μελέτες παρατήρησης είτε από μελέτες πειράματος που βασίζονται σε ένα απολύτως τυχαίο σχεδιασμό.

Η μέθοδος της λογιστικής παλινδρόμησης χρησιμεύει στην ανάπτυξη σχέσης μεταξύ μιας δυαδικής ανεξάρτητης τυχαίας μεταβλητής, η οποία είναι ποιοτική με δύο πιθανά αποτελέσματα, και συνεχών ή διακριτών ανεξάρτητων μεταβλητών. Με αυτή τη μέθοδο γενικεύονται τα γραμμικά μοντέλα ώστε η εξαρτημένη μεταβλητή να ακολουθεί την εκθετική κατανομή.

2.1.1. Μοντέλα παλινδρόμησης όπου η μεταβλητή Y είναι δυαδική

Μόνο δυο πιθανά ποιοτικά αποτελέσματα έχει η εξαρτημένη μεταβλητή στις περισσότερες εφαρμογές της παλινδρόμησης και μπορεί να αντιπροσωπευθεί από μια δυαδική δείκτρια μεταβλητή που θα παίρνει τις τιμές 0,1. Ένα χαρακτηριστικό

παράδειγμα είναι τα καρδιακά προβλήματα συναρτήσει της ηλικίας, του βάρους, της χοληστερίνης και της πίεσης του αίματος. Η συνάρτηση απόκρισης μας προσφέρει 2 πιθανά αποτελέσματα. Το άτομο ανέπτυξε καρδιακά προβλήματα ή όχι κατά τη διάρκεια της μελέτης, αυτά δηλώνονται με 1,0. Ένα άλλο είναι, εάν ένας εμπορικός οίκος έχει βιομηχανικό τμήμα σύμφωνα με το μέγεθος της φίρμας και πάλι η συνάρτηση απόκρισης θα έχει δυο πιθανά αποτελέσματα έχει ή δεν έχει βιομηχανικό τμήμα. Όλα αυτά τα αποτελέσματα θα εκφράζονται με τους αριθμούς 0 και 1.

Υπάρχουν αρκετές εφαρμογές όπου συναντάμε δυαδική συνάρτηση απόκρισης.

Στην αρχή θα ασχοληθούμε με την έννοια-ερμηνεία της συνάρτησης απόκρισης, όταν αυτή είναι δυαδική και στη συνέχεια θα αναφερθούμε σε ειδικά προβλήματα που δημιουργούνται σε αντίστοιχες περιπτώσεις.

Σύμφωνα με το απλό γραμμικό μοντέλο:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad (2.1)$$

Όπου $Y_i = 0{,}1$ στην περίπτωση μας, η αναμενόμενη τιμή $E(Y_i)$ έχει ιδιαίτερη σημασία.

Αφού $E(\varepsilon_i) = 0$ θα έχω:

$$E\{Y_i\} = \beta_0 + \beta_1 X_i \qquad (2.2)$$

Αν η Y_i είναι BERNOULLI τυχαία μεταβλητή ορίζουμε την κατανομή της ως εξής:

Y_i	Πιθανότητα
1	$P(Y_i) = \pi_i$
0	$P(Y_i) = 1 - \pi_i$

Συνεπώς π_i είναι η πιθανότητα ότι η $Y_i = 1$ και $1 - \pi_i$ είναι η πιθανότητα ότι $Y_i = 0$. Από τον ορισμό της αναμενόμενης τιμής μιας τυχαίας μεταβλητής θα έχουμε:

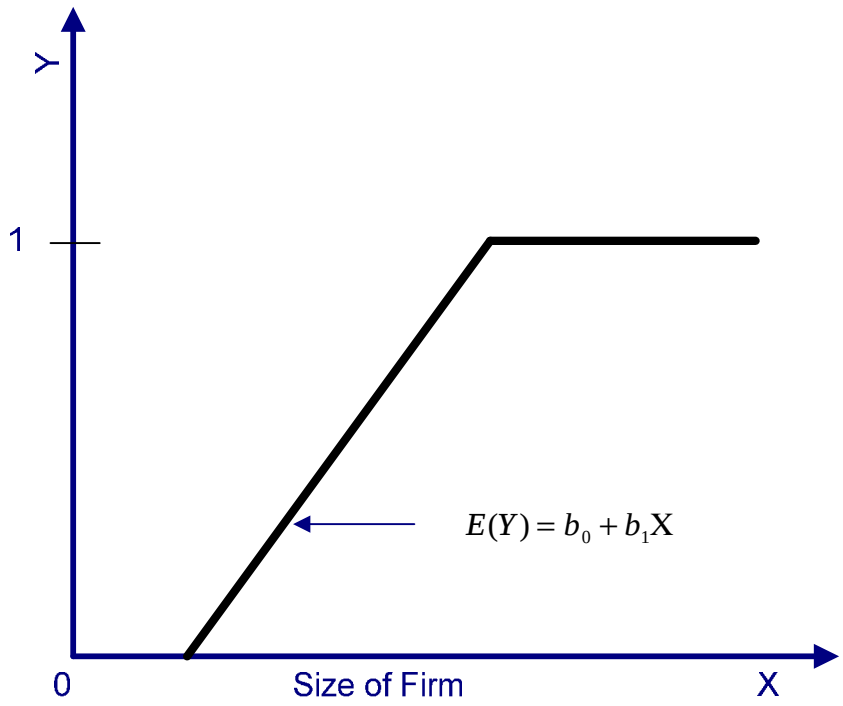
$$E\{Y_i\} = 1\pi_i + 0(1 - \pi_i) = \pi_i \qquad (2.3)$$

Εξισώνοντας τις (2.2) και (2.3) προκύπτει:

$$E\{Y_i\} = \beta_0 + \beta_1 X_i = \pi_i \qquad (2.4)$$

Η μέση τιμή $E(Y_i) = \beta_0 + \beta_1 X_i$ είναι η πιθανότητα ότι η $Y_i = 1$ όταν η τιμή της ανεξάρτητης μεταβλητής είναι X_i. Αυτή είναι η ερμηνεία της μέσης τιμής είτε όταν η συνάρτηση απόκρισης είναι απλή γραμμική είτε πολλαπλή.

Το παρακάτω σχήμα μας δείχνει μια απλή γραμμική συνάρτηση απόκρισης με μια ανεξάρτητη μεταβλητή, και αναφέρεται στο προαναφερθέν παράδειγμα, όπου ο κατακόρυφος άξονας εκφράζει την πιθανότητα ότι η φίρμα έχει βιομηχανικό τμήμα και ο οριζόντιος το μέγεθος της φίρμας.



2.1.2. Ειδικά προβλήματα όταν η συνάρτηση απόκριση είναι δυαδική

Όταν η μεταβλητή απόκρισης είναι δυαδική μεταβλητή δημιουργούνται κάποια προβλήματα. Θα μελετήσουμε κάποια είδη προβλημάτων, και θα χρησιμοποιήσουμε το απλό γραμμικό μοντέλο παλινδρόμησης για επεξήγηση.

1. Όχι κανονικά σφάλματα

Όταν έχουμε δυαδική μεταβλητή απόκρισης, κάθε σφάλμα

$$\varepsilon_i = Y_i - (\beta_0 + \beta_1 X_i)$$

μπορεί να πάρει μόνο δυο τιμές:

- Όταν $Y_i = 1$ $\varepsilon_i = 1 - \beta_0 - \beta_1 X_i)$ (2.5.α)

- Όταν $Y_i = 0$ $\varepsilon_i = -\beta_0 - \beta_1 X_i)$ (2.5.β)

Προφανώς, το μοντέλο παλινδρόμησης κανονικού σφάλματος που υποθέτει πως τα ε_i είναι κατανεμημένα κανονικά δεν είναι κατάλληλο.

2. Μη σταθερή διασπορά σφάλματος

Ένα άλλο πρόβλημα με τα σφάλματα ε_i είναι ότι δεν έχουν ίσες διασπορές όταν η μεταβλητή απόκρισης είναι δείκτρια μεταβλητή.

$$\sigma^2\{Y_i\} = E\{(Y_i - E\{Y_i\})^2\} = (1-\pi_i)^2\pi_i + (0-\pi_i)^2(1-\pi_i)$$
$$\sigma^2\{Y_i\} = \pi_i(1-\pi_i) = (E\{Y_i\})(1-E\{Y_i\}) \quad (2.6)$$

η διασπορά της ε_i είναι η ίδια με της Y_i γιατί $\varepsilon_i = Y_i - \pi_i$ και η π_i είναι σταθερή

$$\sigma^2\{\varepsilon_i\} = \pi_i(1-\pi_i) = (E\{Y_i\})(1-E\{Y_i\}) \quad (2.7)$$
$$\sigma^2\{\varepsilon_i\} = (\beta_0 + \beta_1 X_i)(1-\beta_0-\beta_1 X_i) \quad (2.7.\alpha)$$

Να σημειώσουμε ότι η $\sigma^2\{\varepsilon_i\}$ εξαρτάται από το X_i. Για αυτό το λόγο η διασπορά σφάλματος θα είναι διαφορετική για διαφορετικά επίπεδα του X και τα συνηθισμένα ελάχιστα τετράγωνα δεν θα είναι πια βέλτιστα.

3. Περιορισμοί στην συνάρτησης απόκρισης Y

Αφού η Y αντιπροσωπεύει πιθανότητες που έχουν αποτέλεσμα 0 ή 1 οι μέσες τιμές θα πρέπει να περιορίζονται ως εξής:

$$0 \leq E(Y) \leq 1 \qquad (2.8)$$

Ο παραπάνω περιορισμός δημιουργεί σοβαρές δυσκολίες, χρησιμοποιώντας την μέθοδο σταθμισμένων ελάχιστων τετράγωνων μπορούμε να αντιμετωπίσουμε το πρόβλημα των άνισων διασπορών σφάλματος. Με δείγμα μεγάλου μεγέθους η μέθοδος ελάχιστων τετράγωνων εξασφαλίζει εκτιμητές οι οποίοι είναι ασυμπτωτικά κανονικοί κάτω από γενικές προϋποθέσεις ακόμα και αν η κατανομή που ακολουθούν τα σφάλματα δεν είναι κανονική. Ωστόσο ο περιορισμός των μέσων τιμών να βρίσκονται ανάμεσα στο 0 και 1 συχνά αποκλείει μια γραμμική συνάρτηση απόκρισης .

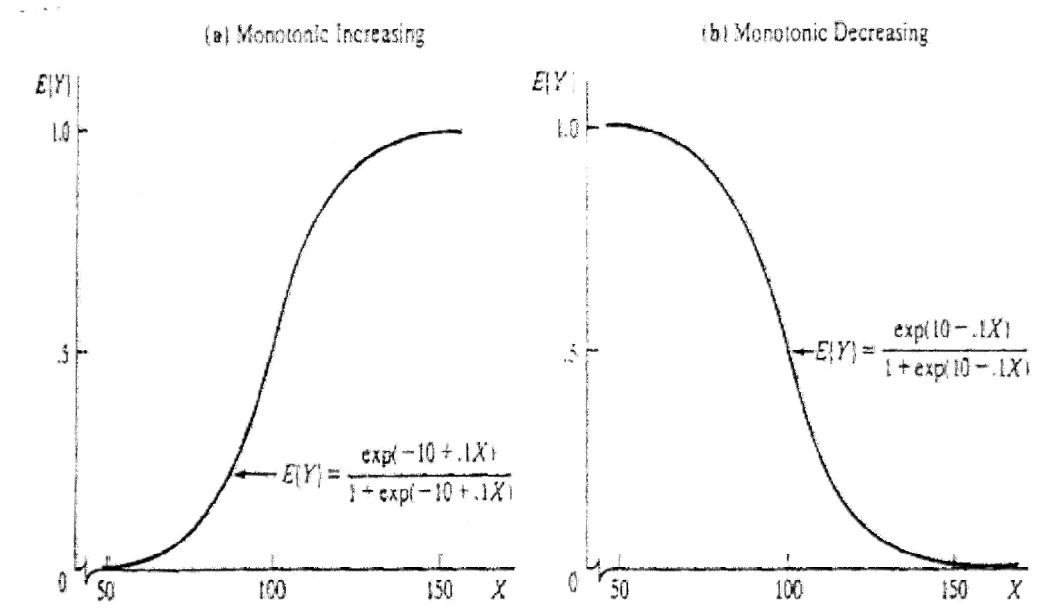
2.2. Απλή λογιστική συνάρτηση απόκρισης

Εμπειρικές και θεωρητικές μελέτες καταλήγουν στο συμπέρασμα πως όταν η μεταβλητή απόκρισης είναι δυαδική η γραφική της απεικόνιση είναι μια καμπύλη γραμμή. Οι συναρτήσεις, αυτές έχουν το σχήμα είτε ενός πλάγιου S είτε ενός ανάστροφου πλάγιου S και είναι προσεγγιστικά γραμμικές αν εξαιρέσουμε τα άκρα τους, τις ονομάζουμε σιγμοειδείς. Αυτές έχουν ασύμπτωτες στο 0 και στο 1 και έτσι ικανοποιούν αυτόματα τους περιορισμούς της μέσης τιμής. Οι συναρτήσεις αυτές ονομάζονται λογιστικές συναρτήσεις απόκρισης και δίνονται από τον τύπο:

$$\mathrm{E\{Y\}} = \frac{\exp(\beta_0 + \beta_1 \mathrm{X})}{1 + \exp(\beta_0 + \beta_1 \mathrm{X})} \qquad (2.9)$$

ή από τον τύπο:

$$\mathrm{E\{Y\}} = \left[1 + \exp\left(-(\beta_0 - \beta_1 \mathrm{X})\right)\right]^{-1} \qquad (2.9.\alpha)$$



2.2.1. Ιδιότητες της λογιστικής συνάρτησης

Μελετώντας τις γραφικές παραστάσεις συμπεραίνουμε ότι η λογιστική συνάρτηση απόκρισης είναι μονότονη, αύξουσα στη μία περίπτωση και φθίνουσα στην άλλη. Επίσης, η Y μπορεί να γίνει εύκολα γραμμική.

Αυτό γίνεται αν πάρουμε ως δεδομένο ότι $E\{Y\} = \pi$, αφού η μέση τιμή είναι η πιθανότητα όταν η συνάρτηση απόκρισης είναι 0,1 δείκτρια μεταβλητή και κάνουμε το μετασχηματισμό:

$$\pi' = log_e\left(\frac{\pi}{1-\pi}\right) \qquad (2.10)$$

Που λέγεται logit μετασχηματισμός της πιθανότητας π και ο λόγος $\frac{\pi}{1-\pi}$ λέγετε odds.

Ως γνωστόν

$$E\{Y\} = \frac{exp(\beta_0 + \beta_1 X)}{1 + exp(\beta_0 + \beta_1 X)}$$

Και

$$E\{Y\} = \pi$$

Οπότε με την βοήθεια των παραπάνω έχουμε:

$$\pi' = log_e\left(\frac{\pi}{1-\pi}\right)$$

$$\pi' = log_e\left(\frac{\frac{exp(\beta_0 + \beta_1 X)}{1 + exp(\beta_0 + \beta_1 X)}}{1 - \frac{exp(\beta_0 + \beta_1 X)}{1 + exp(\beta_0 + \beta_1 X)}}\right) = log_e\left(\frac{\frac{exp(\beta_0 + \beta_1 X)}{1 + exp(\beta_0 + \beta_1 X)}}{\frac{1}{1 + exp(\beta_0 + \beta_1 X)}}\right)$$

$$= log_e\, exp(\beta_0 + \beta_1 X) = \beta_0 + \beta_1 X$$

Άρα

$$\pi' = \beta_0 + \beta_1 X \qquad (2.11)$$

Η μετασχηματιζόμενη αυτή συνάρτηση απόκρισης λέγεται logit συνάρτηση απόκρισης και η π’ logit μέση τιμή.

2.2.2. Χρήσεις της λογιστικής συνάρτησης

Η λογιστική συνάρτηση απόκρισης όπως και οι άλλες συναρτήσεις απόκρισης χρησιμοποιείται για να περιγράψουμε την σχέση μεταξύ της μέσης τιμής και των ανεξάρτητων μεταβλητών και για να κάνουμε προβλέψεις. Για περιγραφικούς σκοπούς, αλλά και για λόγους πρόβλεψης πρώτα χρειάζεται να εκτιμήσουμε τις παραμέτρους της λογιστικής συνάρτησης απόκρισης.

2.3. Απλή λογιστική παλινδρόμηση

Όταν η μεταβλητή απόκρισης είναι δυαδική και παίρνει τιμές 0,1 με πιθανότητες 1-π είναι τυχαία μεταβλητή Bernoulli με $E(Y) = \pi$ το μοντέλο θα είναι της μορφής $Y_i = E(Y_i) + \varepsilon_i$.

Επειδή η κατανομή των σφαλμάτων e_i εξαρτάται από την κατανομή Bernoulli της συνάρτησης απόκρισης δηλώνουμε το απλό λογιστικό μοντέλο παλινδρόμησης στην ακόλουθη μορφή, όπου Y_i είναι ανεξάρτητες μεταβλητές που ακολουθούν κατανομή BERNOULLI, με αναμενόμενες τιμές $E(Y_i) = p_i$, έχουμε:

$$E\{Y_i\} = \pi_{i=} \frac{exp(\beta_0 + \beta_1 X_i)}{1 + exp(\beta_0 + \beta_1 X_i)} \qquad (2.12)$$

Όπου οι X_i παρατηρήσεις είναι σταθερές που είναι γνωστές, εναλλακτικά σε περίπτωση που οι X_i παρατηρήσεις είναι τυχαίες η $E\{Y_i\}$ είναι υποθετική μέση τιμή που δίνει τις τιμές των X_i.

Συνάρτηση πιθανοφάνειας

Αφού η κάθε παρατήρηση Y_i είναι κανονική τυχαία μεταβλητή Bernoulli όπου

$P((Y_i = 1)) = \pi_i, \quad P(Y_i = 0) = 1 - \pi_i$

Η πιθανότητα κατανομής γράφεται ως εξής:

$$f_i = (Y_i) = \pi_i^{Y_i}(1 - \pi_i)^{1-Y_i} \qquad (2.13)$$

Όπου $Y_i = 0,1$, $i = 0,1, \dots, n$, $f_i(1) = \pi_i$ και , $f_i(0) = 1 - \pi_i$.

Αφού οι Y_i παρατηρήσεις είναι ανεξάρτητες, η από κοινού συνάρτηση πιθανοφάνειας τους είναι:

$$g(Y_1, \dots, Y_n) = \prod_{i=1}^{n} f_i(Y_i) = \prod_{i=1}^{n} \pi_i^{Y_i} (1 - \pi_i)^{1-Y_i} \qquad (2.14)$$

Για την εύρεση των εκτιμήσεων μέγιστης πιθανοφάνειας είναι πιο εύκολο να χρησιμοποιήσουμε λογάριθμους, όποτε η παραπάνω συνάρτηση θα γίνει:

$$log_e\, g\,(Y_1, \dots, Y_n) = log_e \prod_{i=1}^{n} \pi_i^{Y_i} (1 - \pi_i)^{1-Y_i} = \sum_{i=1}^{n} \left[Y_i\, log_e \left(\frac{\pi_i}{1 - \pi_i} \right) \right] + \sum_{i=1}^{n} (1 - \pi_i) \quad (2.15)$$

Επειδή $E(Y_i) = \pi_i$ και $E[Y_i] = [1 + exp(-\beta_o - \beta_1 X)]^{-1}$ παίρνουμε το εξής:

$$1 - \pi_i = [1 + exp(-\beta_o - \beta_1 X)]^{-1} \qquad (2.16)$$

Από τις (2.10) και (2.11) πετυχαίνουμε:

$$log_e\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 X_i \qquad (2.17)$$

Άρα

$$log_e\, L(\beta_0, \beta_1) = \sum_{i=1}^{n} Y_i(\beta_0 + \beta_1 X_i) - \sum_{i=1}^{n} log_e[1 + exp\,(\beta_0 + \beta_1 X_i)] \qquad (2.18)$$

Όπου η $L(\beta_0, \beta_1)$ αντικαθιστά την $g(Y_1, \dots, Y_n)$.

2.3.1. Εκτίμηση μέγιστης πιθανοφάνειας

Αφού έχουμε βρει τους εκτιμητές μέγιστης πιθανοφάνειας b_0, b_1 κάνουμε αντικατάσταση τις τιμές τους στην (2.7) προς εξασφάλιση της προσαρμοσμένης συνάρτησης απόκρισης. Στη συνέχεια θα χρησιμοποιήσουμε τον συμβολισμό $\widehat{\pi_i}$ για να δηλώσουμε τις προσαρμοσμένες τιμές για την i περίπτωση

$$\pi_i = \frac{\exp\,(b_0 + b_1 X_i)}{1 + \exp(b_0 + b_1 X_i)} \qquad (2.19)$$

Η προσαρμοσμένη λογιστική συνάρτηση απόκρισης θα είναι:

$$\widehat{\pi_i} = \frac{exp\,(b_0 + b_1 X)}{1 + exp(b_0 + b_1 X)} \qquad (2.20)$$

Βάση της (2.11) και της (2.20) προκύπτει:

$$\widehat{\pi}' = b_0 + b_1 X \qquad (2.21)$$

Όπου $\widehat{\pi}' = log_e\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right)$ (2.21.α) και λέγεται fitted logit συνάρτηση απόκρισης.

2.3.2. Ερμηνεία του $\mathbf{b_1}$

Η ερμηνεία των εκτιμώμενων συντελεστών παλινδρόμησης b_1 στην προσαρμοσμένη λογιστική συνάρτηση απόκρισης δεν είναι η ακριβής ερμηνεία της κλίσης όπως στο γραμμικό μοντέλο παλινδρόμησης. Αυτό συμβαίνει επειδή η επίδραση της μοναδιαίας αύξησης του X ποικίλει για το λογιστικό μοντέλο παλινδρόμησης ανάλογα με τη θέση του αρχικού σημείου στην X κλίμακα. Μια ερμηνεία του b_1 βρίσκεται στην ιδιότητα της προσαρμοσμένης λογιστικής συνάρτησης, βάση της οποίας οι εκτιμώμενες odds $\frac{\hat{\pi}}{1-\hat{\pi}}$ πολλαπλασιάζονται με το $exp(b_1)$ για κάθε μοναδιαία αύξηση του X.

Για να το δούμε αυτό θεωρούμε την τιμή της προσαρμοσμένης logit συνάρτησης απόκρισης στο $X = X_j : \widehat{\pi}'(X_j) = b_0 + b_1 X_j$.

Ο συμβολισμός $\widehat{\pi}'(X_j)$ δείχνει συγκεκριμένα το επίπεδο των X που σχετίζεται με την προσαρμοσμένη τιμή. Επιπροσθέτως θεωρούμε την τιμή της προσαρμοσμένης logit συνάρτησης απόκρισης για $X = X_j + 1$, $\widehat{\pi}'(X_j + 1) = b_0 + b_1(X_j + 1)$.

Η διαφορά ανάμεσα στις δυο προσαρμοσμένες τιμές είναι: $\widehat{\pi}'(X_j + 1) - \widehat{\pi}'(X_j) = b_1$.

Τώρα σύμφωνα με την (2.21α) το $\widehat{\pi}'(X_j)$ είναι ο λογάριθμος των εκτιμώμενων odds όταν $X = X_j$. Θα το δηλώσουμε με $\log_e(odds_1)$ ομοίως $\widehat{\pi}'(X_j + 1)$ είναι ο λογάριθμος των εκτιμώμενων odds όταν $X = X_j + 1$ θα το δηλώνουμε με $\log_e(odds_2)$. Για αυτό το λόγο η διαφορά μεταξύ των fitted logit θα δηλώνεται με:

$$\log_e(odds_2) - \log_e(odds_1) = \log_e\left(\frac{odds_2}{odds_1}\right) = b_1$$

Αν θεωρήσουμε την εκθετική συνάρτηση απόκρισης exp (b_1)

$$\widehat{OR} = \frac{odds_2}{odds_1} = \exp(b_1) \qquad (2.22)$$

2.3.3. Επαναλαμβανόμενες παρατηρήσεις

Σε κάποιες περιπτώσεις, ιδιαίτερα στα πειράματα σχεδιασμού ένας αριθμός επαναλαμβανόμενων παρατηρήσεων επιτυγχάνεται σε διαφορετικά επίπεδα της ανεξάρτητης μεταβλητής X.

Όταν έχουμε επαναλαμβανόμενες παρατηρήσεις η λογαριθμική συνάρτηση απόκρισης (2.19) μπορεί να απλουστευθεί. Θα δηλώσουμε τα X επίπεδα στα οποία επιτυγχάνονται οι επαναλαμβανόμενες παρατηρήσεις ως $X_1, \dots, X_c$.

Ο αριθμός των παρατηρήσεων στα επίπεδα X_j θα δηλώνεται από το n_j $j = (1,2, \dots, c)$ και τα R_j δηλώνουν τον αριθμό της πρώτης από τα X_j.

Άρα

$$p_j = \frac{R_j}{n_j} \qquad (2.23)$$

Οπότε η λογαριθμισμένη συνάρτηση πιθανοφάνειας (2.18) μπορεί να μετασχηματιστεί:

$$log_e L(\beta_0, \beta_1) = \sum_{j=1}^{c} \left\{ \log_e \binom{n_j}{R_j} + R_j(\beta_0 + \beta_0 X_j) - n_j \log_e[1 + exp(\beta_0 + \beta_0 X_j)] \right\} \qquad (2.24)$$

Όπου $\binom{n_j}{R_j} = \frac{n_j!}{R_j!(n_j - R_j)!}$

2.4. Πολλαπλή λογιστική παλινδρόμηση

Το απλό λογιστικό μοντέλο παλινδρόμησης μπορεί να επεκταθεί σε περισσότερες από μία μεταβλητές. Πολλές φορές διαφορετικές ανεξάρτητες μεταβλητές απαιτούνται για την εξασφάλιση ικανοποιητικής περιγραφής και χρήσιμων προβλέψεων. Όσο περισσότερες μεταβλητές έχουμε τόσο μικρότερο το σφάλμα που θα έχουμε. Σε επέκταση του απλού λογιστικού λογιστικού μοντέλου παλινδρόμησης απλά αντικαθιστούμε στην (2.15) το $\beta_0 + \beta_1 X_1$ με $\beta_0 + \beta_1 X_1 + \cdots + \beta_{\rho-1} X_{\rho-1}$.

Η απλοποίηση των τύπων θα γίνει με τη χρήση πινάκων και τριών διανυσμάτων:

$$\beta = \left(\beta_0, \beta_1, \ldots \beta_{\rho-1}\right)'$$

$$X = \left(1, X_1, X_2, \ldots X_{\rho-1}\right)'$$

$$X_i = \left(1, X_{i1}, X_{i2}, \ldots X_{i,\rho-1}\right)' \qquad (2.25)$$

Οπότε θα έχουμε:

$$\beta' X = \beta_0 + \beta_1 X_1 + \cdots + \beta_{\rho-1} X_{\rho-1} \qquad (2.26.\alpha)$$

$$\beta' X_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_{\rho-1} X_{i,\rho-1} \qquad (2.26.\beta)$$

Αναλογικά με την (2.9) για το πολλαπλό γραμμικό μοντέλο έχουμε:

$$E\{Y\} = \frac{\exp\left(\beta' \mathrm{X}\right)}{1 + \exp\left(\beta' \mathrm{X}\right)} \qquad (2.27)$$

Καθώς και

$$E\{Y\} = [1 + exp\ (\beta' X)]^{-1} \qquad (2.28)$$

Ομοίως ο logit μετασχηματισμός θα είναι:

$$\pi' = log_e\left(\frac{\pi}{1-\pi}\right) \qquad (2.29)$$

Και προκύπτει $\pi' = \beta' X$ (2.30)

Το πολλαπλό λογιστικό μοντέλο παλινδρόμησης ορίζεται ως εξής:

Οι Y_i είναι ανεξάρτητες Bernoulli τυχαίες μεταβλητές όπου $E\{Y_i\} = \pi_i$ οπότε έχουμε τη δυνατότητα να γράψουμε:

$$E\{Y_i\} = \pi_i = \frac{\exp(\beta' X_i)}{1 + \exp(\beta' X_i)} \qquad (2.31)$$

Οι X είναι γνωστές σταθερές αλλιώς αν οι X είναι τυχαίες μεταβλητές τότε η E{Y} είναι μια υποθετική μέση τιμή. Όπως η απλή λογιστική συνάρτηση απόκρισης έτσι και η πολλαπλή λογιστική συνάρτηση απόκρισης είναι μονότονη και σιγμοειδής καθώς και γραμμική όταν το π παίρνει τιμές από 0.2 έως 0.8

2.4.1. Προσαρμογή του μοντέλου

Για την εκτίμηση των παραμέτρων της πολλαπλής λογιστικής συνάρτησης απόκρισης χρησιμοποιείτε η μέθοδος μέγιστης πιθανοφάνειας. Η λογαριθμική συνάρτηση πιθανοφάνειας για την απλή λογιστική παλινδρόμηση επεκτείνεται στην πολλαπλή λογιστική παλινδρόμηση.

$$\log_e L(\beta) = \sum_{i=1}^{n} Y_i(\beta' X_i) - \sum_{i=1}^{n} \log_e[1 + \exp(\beta' X_i)] \qquad (2.32)$$

Χρησιμοποιούμε αριθμητικές έρευνες για να βρούμε τις τιμές των $\beta_0, \beta_1, \dots, \beta_{p-1}$ οι οποίες μεγιστοποιούν την $\log_e L(\beta)$. Τις εκτιμήσεις μέγιστης πιθανοφάνειας θα δηλώνουμε με $\mathrm{b}_0, \mathrm{b}_1, \dots, \mathrm{b}_{p-1}$.

Το b δηλώνει το διάνυσμα των εκτιμήσεων αυτών άρα:

$$b = \begin{pmatrix} \mathrm{b}_0 \\ \mathrm{b}_1 \\ : \\ . \\ \mathrm{b}_{p-1} \end{pmatrix} \qquad (2.33)$$

Η προσαρμοσμένη λογιστική συνάρτηση απόκρισης εκφράζεται:

$$\hat{\pi} = \frac{\exp(b'X)}{1 + \exp(b'X)} = \left[1 + \exp(-b'X)\right]^{-1} \qquad (2.34.\alpha)$$

$$\hat{\pi}_i = \frac{\exp(b'X_i)}{1 + \exp(b'X_i)} = \left[1 + \exp(-b'X_i)\right]^{-1} \qquad (2.34.\beta)$$

Όπου $\mathrm{b}'\mathrm{X} = \mathrm{b}_0 + \mathrm{b}_1\mathrm{X}_1 + \cdots + \mathrm{b}_{\mathrm{p}-1}\mathrm{X}_{\mathrm{p}-1}$

Και $b'X_i = \mathrm{b}_0 + \mathrm{b}_1\mathrm{X}_{\mathrm{i}1} + \cdots + \mathrm{b}_{\mathrm{p}-1}\mathrm{X}_{\mathrm{i,p}-1}$

2.4.2. Εναλλακτική διαδικασία εύρεσης εκτιμητών

Οι εκτιμήσεις μέγιστης πιθανοφάνειας της παραμέτρου b για το λογιστικό μοντέλο παλινδρόμησης μπορεί να εξασφαλιστούν από επαναληπτικά reweighed (ανασταθμισμένα) ελάχιστα τετράγωνα. Η διαδικασία είναι ακριβής όμως απαιτείται εντατική χρήση του υπολογιστή. Ας δούμε τα βήματα αυτής της διαδικασίας.

1. Χρησιμοποιούμε αρχικές τιμές για τις παραμέτρους παλινδρόμησης δηλώνοντας τις με $b(0)$. Συχνά λογικές αρχικές τιμές μπορούν να αποκτηθούν από παλινδρόμηση ελάχιστων τετράγωνων του Y πάνω στις μεταβλητές πρόβλεψης $X_1, X_2, \dots, X_{p-1}$ χρησιμοποιώντας γραμμικό μοντέλο πρώτης τάξης .
2. Με τη χρήση αυτών των αρχικών τιμών πετυχαίνουμε:

$$\hat{\pi}_i'(0) = [b(0)]'X_i \qquad (2.35)$$

$$\hat{\pi}_i(0) = \frac{\exp[\hat{\pi}_i'(0)]}{1 + \exp[\hat{\pi}_i'(0)]} \qquad (2.36)$$

3. Υπολογίζουμε τη νέα μεταβλητή απόκρισης

$$Y_i'(0) = \hat{\pi}_i(0) + \frac{Y_i - \hat{\pi}_i(0)}{\hat{\pi}_i(0)[1 - \hat{\pi}_i(0)]} \qquad (2.37)$$

Και τα βάρη

$$w_i(0) = \hat{\pi}_i(0)[1 - \hat{\pi}_i(0)] \qquad (2.38)$$

4. Παλινδρομούμε την $Y'(0)$ στην (2.37) πάνω στις μεταβλητές πρόβλεψης $X_1, X_2, \dots, X_{p-1}$ με τη χρήση του γραμμικού μοντέλου πρώτης τάξης με τα παραπάνω βάρη για να αποκτήσουμε καινούριες εκτιμήσεις συντελεστών παλινδρόμησης, που τις συμβολίζουμε με $b(1)$.
5. Επαναλαμβάνουμε τα βήματα 1 έως 4 χρησιμοποιώντας τις τελευταίες προσεγγίσεις συντελεστών μέχρι να υπάρχει λίγη ή καμία αλλαγή στις τιμές τους. Συνήθως για να έχουμε σύγκλιση κάνουμε τρείς ή τέσσερις επαναλήψεις.

2.5. Δημιουργία μοντέλου

2.5.1. Επιλογή μεταβλητών πρόβλεψης

Πολλές φορές στη δημιουργία ενός μοντέλου είναι ενδιαφέρον ο προσδιορισμός του πότε οι μεταβλητές X, σε ένα πολλαπλό λογιστικό μοντέλο μπορούν να περιοριστούν. Αυτό συμβαίνει όταν κάνουμε χρήση των ελέγχων για $\beta_\kappa = 0$.

Για να κάνουμε τους ελέγχους αυτούς χρησιμοποιούμε την εκτίμηση μέγιστης πιθανοφάνειας, όπως και στους γραμμικούς ελέγχους. Ο έλεγχος αυτός λέγεται likelihood ratio test και απαιτείται μεγάλο δείγμα. Βασίζεται σε μια στατιστική που καλείται model deviance.

2.5.2. Model deviance

Η στατιστική αυτή συγκρίνει την λογαριθμισμένη συνάρτηση πιθανοφάνειας του προσαρμοσμένου μοντέλου με αυτήν ενός μοντέλου με n παραμέτρους που έχουν οριστεί τέλεια. Αυτό ονομάζεται saturated model (κορεσμένο μοντέλο).

Θεωρώντας την λογαριθμική συνάρτηση πιθανοφάνειας (2.15) όπου οι n μεταβλητές ακολουθούν κατανομή Bernoulli. Αν αφήσουμε το π_i να είναι απεριόριστο έτσι ώστε η κάθε Y_i παρατήρηση να παίρνει την τιμή 1 με διαφορετική πιθανότητα π_i, χωρίς τον περιορισμό του π_i, τότε έχουμε n παραμέτρους για τις n παρατηρήσεις και εξασφαλίζουμε τέλεια προσαρμογή. Η πιο πάνω συνάρτηση μεγιστοποιείται όταν $\pi_i = Y_i$.

Ο εκτιμητής μέγιστης πιθανοφάνειας π_i για το saturated model δηλώνεται ως $\widehat{\pi_{\imath s}}$, όπου $\widehat{\pi_{\imath s}} = Y_i$.

Με τη χρήση του $\widehat{\pi_{\imath s}} = Y_i$ μπορούμε να δείξουμε ότι η πιθανότητα των παρατηρήσεων του δείγματος $L(\widehat{\pi_{1s}}, \widehat{\pi_{2s}}, \dots, \widehat{\pi_{ns}})$ ισούται με 1 έτσι ώστε η λογαριθμισμένη πιθανοφάνεια να είναι ίση με το 0.

$$\log_e L(\widehat{\pi_{1s}}, \widehat{\pi_{2s}}, \dots, \widehat{\pi_{ns}}) = \sum_{i=1}^{n} [Y_i \log_e(Y_i) + (1 - Y_i) \log_e(1 - Y_i)] = 0 \quad (2.39)$$

Μπορούμε να κάνουμε τη σύγκριση μεταξύ της τιμής της λογαριθμισμένης μέγιστης πιθανοφάνειας για το κορεσμένο μοντέλο και της τιμής της λογαριθμισμένης μέγιστης πιθανοφάνειας για το προσαρμοσμένο μοντέλο. Θεωρούμε μια λογαριθμισμένη πιθανοφάνεια για το λογιστικό μοντέλο όπου

$\pi_i' = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_{p-1} X_{i,p-1}$.

Η παραπάνω συνάρτηση δίνεται στην:

$$\log_e L(\beta) = \sum_{i=1}^{n} Y_i(\beta' X_i) - \sum_{i=1}^{n} \log_e[1 + \exp(\beta' X_i)]$$

Οι τιμές για τις παρατηρήσεις όταν οι εκτιμητές μέγιστης πιθανοφάνειας χρησιμοποιούνται στην log-πιθανοφάνειας δηλώνεται ως εξής $L(b_0, b_1, \dots, b_{p-1})$ και

$$\log_e L(b_0, b_1, \dots, b_{p-1}) = \sum_{i=1}^{n} Y_i(\beta' X_i) - \sum_{i=1}^{n} \log_e[1 + \exp(\beta' X_i)] \quad (2.40)$$

Η τιμή της log-πιθανοφάνειας για το προσαρμοσμένο μοντέλο δεν μπορεί να είναι μεγαλύτερη από την τιμή της log-πιθανοφάνειας για το κορεσμένο μοντέλο γιατί το προσαρμοσμένο μοντέλο έχει λιγότερες παραμέτρους.

Η deviance βασίζεται στην διαφορά των δυο λογαριθμισμένων πιθανοφανειών. Δηλώνουμε την deviance για το fitted μοντέλο με $Dev(X_0, X_1, \dots, X_{p-1})$ όπου $X_0 = 1$.

Η deviance δηλώνεται ως εξής:

$$Dev(X_0, X_1, \dots, X_{p-1}) = 2 \log_e L(\widehat{\pi_{1s}}, \widehat{\pi_{2s}}, \dots, \widehat{\pi_{ns}}) - 2 \log_e L(b_0, b_1, \dots, b_{p-1}) \quad (2.41)$$

Για το πολλαπλό λογιστικό μοντέλο (2.31) έχω:

$$Dev(X_0, X_1, \dots, X_{p-1}) = -2\sum_{i=1}^{n}[Y_i \log_e(\widehat{\pi_\iota}) + (1 - Y_i)\log_e(1 - \widehat{\pi_\iota})] \qquad (2.42)$$

Όπου $\widehat{\pi_\iota}$ είναι η τιμή i για το (2.34.β). Όσο μικρότερη είναι η διάφορα των τιμών των log πιθανοφανειών τόσο πιο κοντά είναι το προσαρμοσμένο μοντέλο στο κορεσμένο. Τέλος μπορούμε να πούμε ότι όσο μεγαλύτερη είναι η deviance του μοντέλου τόσο χειρότερη είναι η προσαρμογή.

Η deviance για το normal-error linear δηλώνεται:

$$SSE = \sum(Y_i - \widehat{Y_\iota})^2$$

2.5.3. Partial deviance

Για κάθε προσαρμοσμένο μοντέλο μπορεί να υπολογιστεί η παρέκκλιση του. Η διαφορά ανάμεσα στις deviance για 2 προσαρμοσμένα μοντέλα καλείται Partial Deviance και μας βοήθα να εξετάσουμε αν κάποιες μεταβλητές πρόβλεψης μπορούν να απαλειφθούν από το μοντέλο.

Το πλήρες λογιστικό μοντέλο με συνάρτηση απόκρισης είναι το εξής:

$$\pi = [1 + \exp(-\beta_F' X)]^{-1} \qquad (2.43)$$

Όπου

$$\beta_F' X = \beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1} \qquad (2.43.\alpha)$$

Ο συμβολισμός του διανύσματος των συντελεστών παλινδρόμησης για το πλήρες μοντέλο είναι β_F. Οι εκτιμήσεις μέγιστης πιθανοφάνειας για το πλήρες μοντέλο συμβολίζεται με b_F και η deviance με $Dev(X_0, X_1, \dots, X_{p-1})$.

Ο έλεγχος υποθέσεων θα είναι:

$$H_o: \beta_q = \beta_{q+1} = \cdots = \beta_{p-1} = 0$$
$$H_\alpha: \textit{όχι όλα τα } \beta_k \textit{ στην } H_o \textit{ να είναι μηδέν.} \qquad (2.44)$$

Για ευκολία εξετάζουμε τους τελευταίους p-q συντελεστές του μοντέλου.

Το μειωμένο λογιστικό μοντέλο θα έχει συνάρτηση απόκρισης:

$$\pi = [1 + \exp(-\beta_R' X)]^{-1} \qquad (2.45)$$

Όπου $\beta_R' X = \beta_0 + \beta_1 X_1 + \cdots + \beta_{q-1} X_{q-1}$

Στη συνέχεια θα βρούμε την εκτίμηση μέγιστης πιθανοφάνειας b_R για το Reduce model και το Deviance του.

Αν η (αποκλίνουσα συμπεριφορά) deviance του μειωμένου μοντέλου δεν είναι αρκετά μεγαλύτερη από την deviance του Full model τότε το Reduce με τις λιγότερες παραμέτρους είναι κοντά στο προσαρμοσμένο όποτε μπορούμε να δεχτούμε την αρχική υπόθεση και να παραλείψουμε τα $(X_q, \dots, X_{p-1})$ από το λογιστικό μοντέλο παλινδρόμησης. Μεγάλη διάφορα ανάμεσα στις deviance των προαναφερθέντων μοντέλων φανερώνει πως θα πρέπει να κρατήσουμε τις μεταβλητές πρόβλεψης $(X_q, \dots, X_{p-1})$. Η διαφορά ανάμεσα στις 2 Deviance είναι η partial deviance και θα δηλώνεται με $Dev(X_q, X_1, \dots, X_{p-1}/X_0, X_1, \dots, X_{q-1})$.

$$Dev(X_q, X_1, \dots, X_{p-1}/X_0, X_1, \dots, X_{q-1}) = Dev(X_0, X_1, \dots, X_{q-1}) - Dev(X_0, X_1, \dots, X_{p-1}) \qquad (2.46)$$

Μπορούμε να δείξουμε ότι αν η H_o δεν απορρίπτεται και τα n είναι μεγάλα τότε η (2.46) ακολουθεί chi-square κατανομή με p-q βαθμούς ελευθερίας. Οι βαθμοί ελευθερίας ανταποκρίνονται στη διαφορά των βαθμών ελευθερίας για τα σφάλματα των δύο προσαρμοσμένων μοντέλων: (n-q)-(n-p)=(p-q).

Ο κανόνας απόφασης για να εξετάσουμε την εναλλακτική H_a στην (2.44) είναι :

$$\begin{aligned} &\text{Αν } Dev(X_q, X_1, \dots, X_{p-1}/X_0, X_1, \dots, X_{q-1}) \le \chi^2(1-\alpha, p-q) \text{ αποδοχή της } H_0 \\ &\text{Αν } Dev(X_q, X_1, \dots, X_{p-1}/X_0, X_1, \dots, X_{q-1}) \ge \chi^2(1-\alpha, p-q) \text{ αποδοχή της } H_\alpha \end{aligned} \qquad (2.47)$$

2.5.4. Τρεις διευκρινίσεις για τον έλεγχο partial deviance

1. Το λογιστικό μοντέλο παλινδρόμησης που περιέχει τις X_1, X_2, X_3 είναι προσαρμοσμένο με $Dev(X_0, X_1, X_2, X_3)$. Ελέγχουμε την $H_0: \beta_2 = \beta_3 = 0$. Το λογιστικό μοντέλο παλινδρόμησης θα έχει μόνο μια ανεξάρτητη μεταβλητή, την X_1, και θα είναι προσαρμοσμένο. Η απαιτούμενη partial deviance θα είναι:

$$Dev(X_2, X_3 / X_0, X_1) = Dev(X_0, X_1) - Dev(X_0, X_1, X_2, X_3)$$

και η κατάλληλη chi – square κατανομή έχει (n-2)-(n-4) = 2 βαθμούς ελευθέριας.

2. Τώρα θα ελέγξουμε την $H_0: \beta_1 = \beta_2 = \beta_3 = 0$. Η partial deviance θα είναι: $Dev(X_1, X_2, X_3 / X_0) = Dev(X_0) - Dev(X_0, X_1, X_2, X_3)$

και η chi – square κατανομή έχει (n-1)-(n-4)=3 βαθμούς ελευθέριας.

3. Αν ελέγξουμε την $H_0: \beta_1 = 0$ απαιτούμε

$$Dev(X_1 / X_0, X_2, X_3) = Dev(X_0, X_2, X_3) - Dev(X_0, X_1, X_2, X_3)$$

Η κατάλληλη chi – square κατανομή έχει 1 βαθμό ελευθέριας.

2.5.5. Έλεγχος λόγου πιθανοτήτων

Ο έλεγχος της partial deviance (2.47) είναι ο ίδιος με αυτόν του λόγου πιθανοτήτων. Οι εναλλακτικές της (2.44) ελέγχονται αν προσαρμόσουμε το πλήρες μοντέλο στην (2.43). Αποκτώντας τις εκτιμήσεις μέγιστης πιθανοφάνειας b_F και υπολογίζοντας την συνάρτηση πιθανοφάνειας για $\beta = b_F$.

Η τιμή της πιθανοφάνειας θα δηλώνεται με $L(F) = L(b_0, b_1, \dots, b_{p-1})$ στη συνέχεια προσαρμόζουμε στο μειωμένο μοντέλο και έχουμε $L(R) = L(b_0, b_1, \dots, b_{p-1})$.

Η αναλογία των δυο πιθανοτήτων $\frac{L(R)}{L(F)}$ είναι ο λόγος πιθανοτήτων.

Ο στατιστικός έλεγχος για το λόγο πιθανοτήτων θα δηλώνεται X^2 και είναι:

$$X^2 = -2\log_e \frac{L(R)}{L(F)} = 2\log_e L(F) - 2\log_e L(R)$$
$$= 2\log_e L(b_0, b_1, \dots, b_{p-1}) - 2\log_e L(b_0, b_1, \dots, b_{q-1}) \quad (2.48)$$

Αλλά αυτό είναι η partial deviance $Dev(X_q, X_1, \dots, X_{p-1}/X_0, X_1, \dots, X_{q-1})$ μπορούμε να το καταλάβουμε, εάν αντικαταστήσουμε τις deviance από τους ορισμούς τους.

2.5.6. Διαγνωστικοί έλεγχοι

2.5.6.1. Πρακτική εξέταση καλής προσαρμογής

Σημαντικό σημείο της στατιστικής ανάλυσης είναι η εξέταση της καταλληλότητας του προσαρμοσμένου λογιστικού μοντέλου παλινδρόμησης για την περιγραφή ενός φαινόμενου δυαδικής απόκρισης. Συγκεκριμένα, πρέπει να ελέγξουμε αν η εκτιμώμενη συνάρτηση απόκρισης των δεδομένων είναι μονότονη και σιγμοειδής στη μορφή, χαρακτηριστικά που είναι βασικά για την λογιστική συνάρτηση απόκρισης.

Μια διαδικασία πρακτικής εξέτασης καλής προσαρμογής είναι η εξής:

1. Για κάθε περίπτωση, δηλαδή κάθε τιμή της ανεξάρτητης μεταβλητής X, βρίσκουμε τις προσαρμοσμένες τιμές $\widehat{\pi_i}$.
2. Χωρίζουμε τον συνολικό αριθμό περιπτώσεων σε c κλάσεις με κριτήριο το πόσο παρόμοιες είναι οι $\widehat{\pi_i}$ τιμές τους, φροντίζοντας ο αριθμός των περιπτώσεων σε κάθε κλάση να είναι περίπου ίδιος και ο αριθμός των κλάσεων να είναι μικρός σε σχέση με το δείγμα.
3. Για κάθε κλάση γράφουμε το διάστημα στο οποίο ανήκουν οι $\widehat{\pi_i}$ και βρίσκουμε το midpoint, βρίσκουμε την αναλογία εκτιμώμενης μοναδιαίας απόκρισης που συμβολίζουμε με p_j, έχουμε τη δυνατότητα αντί των προσαρμοσμένων τιμών να χρησιμοποιήσουμε τις προσαρμοσμένες logit τιμές $\widehat{\pi_i'}$, λόγω της μονότονης φύσης του logit μετασχηματισμού.

4. Κατασκευάζουμε το διάγραμμα των p_j συναρτήσει των midpoints και συνδέουμε με συνεχή γραμμή τα σημεία. Εάν η καμπύλη που προκύψει είναι μονότονη και σιγμοειδής, τότε έχουμε ένα δείγμα καταλληλότητας του λογιστικού μοντέλου, χωρίς, όμως, μεγάλη ακρίβεια.

2.5.6.2. Έλεγχος chi – square καλής προσαρμογής

Ο συγκεκριμένος έλεγχος αξιώνει μόνο, πως οι Y παρατηρήσεις είναι ανεξάρτητες και πως το μέγεθος του δείγματος είναι μεγάλο. Ο έλεγχος μπορεί να ανακαλύψει μεγάλες αποκλίσεις από μια λογιστική συνάρτηση απόκρισης , όχι όμως και μικρές αποκλίσεις.

Οι υποθέσεις που μας ενδιαφέρουν είναι:

$$H_o: E\{Y\} = [1 + \exp(-\beta' X)]^{-1}$$
$$H_\alpha: E\{Y\} \neq [1 + \exp(-\beta' X)]^{-1} \qquad (2.49)$$

Συμφωνά με την διαγνωστική διαδικασία τα δείγματα ομαδοποιούνται σε κλάσεις, με κατά προσέγγιση ίσο αριθμό περιπτώσεων σε κάθε κλάση. Ο αριθμός κλάσεων θα δηλώνεται με c. Ο αριθμός των περιπτώσεων στην j κλάση θα δηλώνεται με n_j.

$$\sum_{j=1}^{c} n_j = n \qquad (2.50)$$

Ο αριθμός των περιπτώσεων στην j κλάση με αποτέλεσμα 1 δηλώνεται με O_{j1} και με αποτέλεσμα 0 δηλώνεται με O_{j0}, επειδή η συνάρτηση απόκρισης Y_i είναι Bernoulli με 0,1 , τα O_{j1} και O_{j0} θα δηλώνονται ως εξής:

$$O_{j1} = \sum Y_i \qquad (2.51.\alpha)$$

$$O_{j0} = \sum (1 - Y_i) = n_j - O_{j1} \qquad (2.51.\beta)$$

Αν η λογιστική συνάρτηση απόκρισης είναι κατάλληλη η αναμενόμενη τιμή της Y_i θα δίνεται από τον τύπο:

$$E\{Y\} = \pi_i = [1 + \exp(-\beta' X)]^{-1} \qquad (2.52)$$

Και εκτιμάται από την:

$$\widehat{\pi_i} = [1 + \exp(-b' X)]^{-1} \qquad (2.53)$$

Οι αναμενόμενες τιμές όταν $Y_i = 0$ και $Y_i = 1$ είναι:

$$E_{j1} = \sum \widehat{\pi_i} \qquad (2.54)$$

$$E_{j0} = \sum (1 - \widehat{\pi_i}) = n_j - E_{j1} \qquad (2.55)$$

Όπου E_{j1} είναι η εκτιμώμενη μέση τιμή με $Y_i = 1$ για την κλάση j. Ομοίως για το E_{j0} όπου η εκτιμώμενη μέση τιμή με $Y_i = 0$ για την κλάση j.

Ο στατιστικός chi – square έλεγχος θα είναι:

$$X^2 = \sum_{j=1}^{c} \sum_{k=0}^{1} \frac{(O_{jk} - E_{jk})^2}{E_{jk}} \qquad (2.56)$$

Αν η λογιστική συνάρτηση απόκρισης είναι κατάλληλη η X^2 ακολουθεί κατά προσέγγιση, χ^2 κατανομή με c-2 βαθμούς ελευθερίας, όταν το n είναι μεγάλο και το p μικρότερο του c.

Αν οι τιμές του στατιστικού ελέγχου X^2 είναι μεγάλες τότε αυτό μα δείχνει ότι η λογιστική συνάρτηση απόκρισης δεν είναι κατάλληλη. Ο κανόνας απόφασης για να ελέγξουμε τις υποθέσεις στην (2.49) είναι:

$$\begin{aligned} &\text{Αν } X^2 \leq \chi^2 \text{ (1-α,c-2) δέξου την } H_o \\ &\text{Αν } X^2 > \chi^2 \text{ (1-α,c-2) δέξου την } H_\alpha \end{aligned} \qquad (2.57)$$

2.5.6.3. Έλεγχος έλλειψης καλής προσαρμογής

Ένας ακόμη έλεγχος μπορεί να βασιστεί στη $Dev(X_0, X_1, \dots, X_{p-1})$. Αν η λογιστική συνάρτηση απόκρισης είναι η κατάλληλη και το μέγεθος του δείγματος μεγάλο, τότε η deviance ακολουθεί X^2 κατανομή με n-p βαθμούς ελευθερίας. Μεγάλες τιμές της deviance δείχνουν ότι η προσαρμογή δεν είναι καλή.

Για να κάνουμε των έλεγχο των υποθέσεων στην (2.49) ο κατάλληλος κανόνας απόφασης είναι:

$$\begin{aligned} &\text{Αν } Dev(X_0, X_1, \dots, X_{p-1}) \leq \chi^2 \text{ (1-α, n-p) συμπεριέλαβε την } H_o \\ &\text{Αν } Dev(X_0, X_1, \dots, X_{p-1}) > \chi^2 \text{ (1-α, n-p) συμπεριέλαβε την } H_\alpha \end{aligned} \quad (2.58)$$

2.5.6.4. Υπόλοιπα απόκλισης για τη λογιστική παλινδρόμηση

Η ανάλυση υπόλοιπων για τη λογιστική παλινδρόμηση είναι πιο δύσκολη από αυτή των μοντέλων γραμμικής παλινδρόμησης επειδή οι αποκρίσεις Y παίρνουν μόνο τις τιμές 0,1, συνεπώς τα υπόλοιπα δεν θα είναι κανονικά κατανεμημένα και όντως η κατανομή τους, υποθέτοντας πως το προσαρμοσμένοι μοντέλο είναι σωστό, δεν είναι γνωστή.

Μια χρήσιμη μορφή τέτοιων υπολοίπων είναι το deviance residual. Για την i περίπτωση θα συμβολίζεται με dev_i και δίνεται από την σχέση:

$$dev_i = \pm\{-2[Y_i \log_e(\widehat{\pi_i}) + (1 - Y_i) \log_e(1 - \widehat{\pi_i})]\}^{\frac{1}{2}} \quad (2.59)$$

Χρησιμοποιούμε το (+) όταν $Y_i \geq \widehat{\pi_i}$ και το (-) όταν $Y_i < \widehat{\pi_i}$.

Ακόμη μία ιδιότητα των υπολοίπων είναι:

$$\sum_{i=1}^{n} (dev_i)^2 = Dev(X_0, X_1, \dots, X_{p-1}) \quad (2.60)$$

2.6. Συμπεράσματα για τις παραμέτρους της λογιστικής παλινδρόμησης

Τα συμπεράσματα που προσπαθούμε να εξάγουμε στην λογιστική παλινδρόμηση είναι σαν αυτά της γραμμικής παλινδρόμησης δηλαδή συμπεράσματα για τους συντελεστές παλινδρόμησης, για την εκτίμηση της μέσης τιμής και προβλέψεις νέων παρατηρήσεων. Υπενθυμίζουμε πως για μεγάλα δείγματα και κάτω από κατάλληλες προϋποθέσεις οι εκτιμητές μέγιστης πιθανοφάνειας, για την λογιστική παλινδρόμηση, ακολουθούν κανονική κατανομή με κατά προσέγγιση διασπορές-συνδιασπορές που είναι συναρτήσεις των δεύτερης τάξης μερικών παραγώγων του λογάριθμου της συνάρτησης πιθανοφάνειας

Συγκεκριμένα ας δηλώσουμε με **G** τον πίνακα των δεύτερης τάξης μερικών παράγωγων της συνάρτησης

$$\log_e L(\beta) = \sum_{i=1}^{n} Y_i(\beta' X_i) - \sum_{i=1}^{n} \log_e[1 + \exp(\beta' X_i)]$$

$G = [g_{ij}]$ όπου $i = 0,1,\dots,p-1$ και $j = 0,1,\dots,p-1$

$$g_{00} = \frac{\partial^2 \log_e L(\beta)}{\partial \beta_0^2} \quad g_{01} = \frac{\partial^2 \log_e L(\beta)}{\partial \beta_o \partial \beta_1}$$

Ο παραπάνω πίνακας ονομάζεται HESSIAN.

Ακόμη ο εκτιμώμενος πίνακας διασπορών – συνδιασπορών μπορεί να δηλωθεί ως εξής:

$$s^2\{b\} = [(-g_{ij})_{\beta=b}]^{-1}$$

Τα συμπεράσματα για τους συντελεστές παλινδρόμησης για το απλό και το πολλαπλό μοντέλο παλινδρόμησης βασίζονται στο:

$$\frac{b_k - \beta_k}{s\{b_k\}} = z \text{ όπου } k = 0,1,\dots,p-1 \qquad (2.61)$$

2.6.1. Διάστημα εμπιστοσύνης της $\boldsymbol{\beta_k}$

Από τα παραπάνω πετυχαίνουμε 1-α όριο εμπιστοσύνης για την

$$\beta_k : b_k \pm z\left(1-\frac{a}{2}\right) s\{b_k\} \qquad (2.62)$$

Όπου $z\left(1-\frac{a}{2}\right)$ είναι το $\left(1-\frac{a}{2}\right)100$ ποσοστό της κανονικής κατανομής.

Τα αντίστοιχα όρια εμπιστοσύνης για την αναλογία των $odds$ $\exp(\beta_k)$ είναι:

$$\exp[b_k \pm z\left(1-\frac{a}{2}\right) s\{b_k\}] \qquad (2.63)$$

2.6.2. Ταυτόχρονη εκτίμηση διαστήματος

Η εύρεση κοινού διαστήματος εμπιστοσύνης για διαφορετικές παραμέτρους λογιστικής παλινδρόμησης μπορεί να γίνει μέσω της διαδικασίας Bonferroni, το οποίο έχουμε συναντήσει και στην γραμμική παλινδρόμηση. Αν πρόκειται να εκτιμηθούν g παράμετροι με συντελεστή εμπιστοσύνης 1-α τα κοινά όρια εμπιστοσύνης Bonferroni θα είναι:

$$b_k \pm Bs\{b_k\} \text{ όπου } B = z(1-\frac{a}{2g}) \qquad (2.64)$$

Αυτό θα είναι το όριο εμπιστοσύνης Bonferroni.

2.6.3. Έλεγχος που αφορά μόνο ένα $\boldsymbol{\beta_k}$

Ένας άλλος έλεγχος που μπορεί να γίνει για την παράμετρο παλινδρόμησης β_k, εκτός αυτών που βασίζονται στην partial deviance (2.47), είναι ένας εναλλακτικός, μεγάλου μεγέθους, έλεγχος που βασίζεται στην $\frac{b_k - \beta_k}{s\{b_k\}} = z$.

Με τις υποθέσεις:

$$H_0: \beta_k = 0$$
$$H_a: \beta_k \neq 0 \qquad (2.65)$$

Ο κατάλληλος έλεγχος είναι:

$$z^* = \frac{b_k}{s\{b_k\}} \qquad (2.66)$$

και ο κανόνας απόφασης είναι:

$$\text{Αν } |z^*| \leq z(1 - \tfrac{a}{2}) \text{ δέξου την } H_0$$
$$\text{Αν } |z^*| > z(1 - \tfrac{a}{2}) \text{ δέξου την } H_a \qquad (2.67)$$

2.6.4. Συμπεράσματα για τη μέση τιμή

Συχνά απαιτείται εκτίμηση της πιθανότητας π για ένα ή περισσότερα διαφορετικά σύνολα τιμών των μεταβλητών πρόβλεψης .

2.6.5. Σημειακή εκτίμηση

Ως συνήθως, δηλώνεται το διάνυσμα των επιπέδων των X μεταβλητών, για τις οποίες το π πρόκειται να εκτιμηθεί από X_h.

$$X_h = \begin{pmatrix} 1 \\ X_{h1} \\ X_{h2} \\ . \\ : \\ X_{h,p-1} \\ \end{pmatrix}$$

Και την μέση τιμή με π_h, όπου $\pi_h = [1 + \exp(-\beta' X_h)]^{-1}$.

Ο εκτιμητής σημείου της π_h δηλώνεται με $\widehat{\pi_h}$ και είναι: $\widehat{\pi_h} = [1 + \exp(-b' X_h)]^{-1}$.

Όπου b είναι το διάνυσμα των εκτιμώμενων συντελεστών παλινδρόμησης.

2.6.6. Υπολογισμός διαστήματος

Το διάστημα εμπιστοσύνης για το π_h εξασφαλίζεται σε δύο βήματα.

1. Υπολογίζουμε τα όρια εμπιστοσύνης για την logit mean response, π'_h.
2. Με τη βοήθεια της $E\{Y\} = [1 + \exp(-\beta' X)]^{-1}$ πετυχαίνουμε όρια εμπιστοσύνης για τη μέση τιμή π_h.

Αν στην παραπάνω σχέση υποθέσουμε ότι $X = X_h$ τότε προκύπτει:

$$E\{Y_h\} = [1 + \exp(-\beta' X_h)]^{-1}$$

Και αν θέσουμε $E\{Y_h\} = \pi_h$ και $\beta' X_h = \pi'_h$ έχουμε:

$$\pi_h = [1 + \exp(-\pi'_h)]^{-1}$$

Με τη χρήση της σχέσης αυτής κάνουμε την μετατροπή των ορίων του διαστήματος εμπιστοσύνης για το π'_h σε διαστήματα εμπιστοσύνης για το π_h.

Ο εκτιμητής σημείου της logit μέσης τιμής $\pi'_h = \beta' X_h$ είναι $\widehat{\pi'_h} = b' X_h$. Υποθέτοντας ότι $b' X_h = X'_h b$ επειδή είναι scalar έτσι ώστε η κατά προσέγγιση εκτιμώμενη διασπορά του $\widehat{\pi'_h} = b' X_h = X'_h b$ θα είναι:

$$s^2\{\widehat{\pi'_h}\} = s^2\{X'_h b\} = X'_h s^2\{b\} X_h$$

Κατά προσέγγιση 1-α, μεγάλου δείγματος, όρια εμπιστοσύνης για τη logit μέση τιμή παίρνουμε από

$$L = \widehat{\pi_h'} - z\left(1 - \frac{a}{2}\right) s(\widehat{\pi_h'})$$

$$U = \widehat{\pi_h'} + z\left(1 - \frac{a}{2}\right) s(\widehat{\pi_h'})$$

Τα L και U είναι αντίστοιχα τα πάνω και κάτω όρια εμπιστοσύνης για το π_h'. Με την χρήση της μονότονης σχέσης μεταξύ των π_h και π_h' μετασχηματίζουμε τα όρια εμπιστοσύνης L και U για το π_h' σε όρια εμπιστοσύνης L^* και U^* για τη μέση τιμή π_h.

$$L^* = [1 + \exp(-L)]^{-1}$$

$$U^* = [1 + \exp(-U)]^{-1}$$

3. Κεφάλαιο 3

3.1. Poisson Παλινδρόμηση

Η κατανομή poisson χρησιμεύει στην ανάλυση δεδομένων συχνοτήτων όπως τον αριθμό περιπτώσεων στα κελιά ενός πίνακα συνάφειας, τον αριθμό πελατών ενός τραπεζικού καταστήματος, τον αριθμό σεισμικών δονήσεων μεγαλύτερων των 6,5R σε μια περιοχή.

Η κατανομή poisson δίνεται από τον τύπο:

$$P(Y = y) = \frac{\mu^y e^{-\mu}}{y!} \qquad (3.1)$$

$y = 0,1,2,\dots,v$ και $\mu > 0$

Για να δηλώσουμε ότι μια τυχαία μεταβλητή έχει κατανομή poisson γράφουμε:

$$Y \sim P(\mu)$$

Όπου Y μια διακριτή τυχαία μεταβλητή που μετρά τη συχνότητα εμφάνισης σπανίων ενδεχομένων σε προκαθορισμένο χρονικό διάστημα.

Η παράμετρος μ είναι ο μέσος αριθμός των εμφανίσεων του ενδεχόμενου σε διάστημα προκαθορισμένου πλάτους.

3.1.1. Κατανομή poisson

Κατανομή poisson έχουμε:

- Αν τα δεδομένα είναι ανεξάρτητα.

Δηλαδή αν η εμφάνιση ενός ενδεχομένου σε προκαθορισμένο χρονικό διάστημα σταθερού πλάτους δεν επηρεάζεται και δεν επηρεάζει την εμφάνιση κάποιου άλλου ενδεχόμενου.

- Τα ενδεχόμενα είναι ισοπίθανα σε μικρά διαστήματα ίσου πλάτους.

Δηλαδή η πιθανότητα εμφάνισης ενός ενδεχόμενου κατά τη διάρκεια μικρού χρονικού διαστήματος π.χ. Δt παραμένει σταθερή για όλα τα διαστήματα πλάτους ίσου με το Δt.

- Τα ενδεχόμενα εμφανίζονται μεμονωμένα σε διαστήματα μικρού πλάτους.

Δηλαδή ένα μόνο ενδεχόμενο είναι δυνατόν να εμφανιστεί κατά την διάρκεια μικρού χρονικού διαστήματος Δt.

Χαρακτηριστική ιδιότητα της κατανομής poisson είναι ότι αν η τυχαία μεταβλητή έχει κατανομή poisson με παράμετρο μ τότε η αναμενόμενη τιμή της και η διακύμανσή της είναι ίσες και ισούνται με μ.

$$E[Y] = \mu \text{ και } Var[Y] = \mu \qquad (3.2)$$

3.1.2. Poisson παλινδρόμηση

Έστω $Y_1, Y_2, \dots, Y_n$ ανεξάρτητες μεταβλητές από την Poisson μ_i. Υποθέτουμε ότι $E(Y_i) = \mu_i = n_i\theta_i$. Αν σαν παράδειγμα πάρουμε Y_i τον απαιτούμενο αριθμό πωλήσεων ενός προϊόντος Α σε μια αλυσίδα supermarket τότε n_i θα είναι το σύνολο πωλήσεων του συγκεκριμένου προϊόντος και σαν θ_i μπορούμε να ορίσουμε την περιοχή του καταστήματος, την ηλικία του αγοραστικού κοινού κ.ο.κ.. Για να αναλύσουμε τέτοια δεδομένα συνήθως χρησιμοποιούμε το μοντέλο:

$$\theta^i = e^{x_i^T \beta} \qquad (3.3)$$

Σε αυτή την περίπτωση το αντίστοιχο γενικευμένο μοντέλο είναι της μορφής:

$$E(Y_i) = \mu_i = n_i e^{x_i^T \beta} \qquad (3.4)$$

Αν για παράδειγμα πάρουμε $x_i = 0{,}1$ τότε:

$$E(Y_i|X_i = 0) = n_i, E(Y_i|X_i = 0) = n_i e^{x_i^T \beta} \quad (3.5)$$

Άρα το ποσοστιαίο πηλίκο δίνεται από τον τύπο:

$$RR = \frac{E(Y_i|X_i = 1)}{E(Y_i|X_i = 0)} = e^{\beta} \quad (3.6)$$

Και μας δείχνει την αλλαγή στην αναμενόμενη τιμή. Την εκτίμηση της παραμέτρου β παίρνουμε μέσα από τη θεωρία πιθανοφάνειας για γενικευμένα γραμμικά μοντέλα.

Αν $\hat{\beta}$ είναι ο εκτιμητής μέγιστης πιθανοφάνειας (Ε.Μ.Π.) τότε μπορούμε να ελέγξουμε τις υποθέσεις $H_0: \beta = \beta_0$ μέσω score test, Wald test και έλεγχο πηλίκου πιθανοφάνειας.

$$\hat{Y}_i = \hat{\mu}_i = n_i e^{x_i \hat{\beta}} \quad (3.7)$$

Τα υπόλοιπα Pearson δίνονται από

$$r_i = \frac{O_i - E_i}{\sqrt{E_i}} \quad (3.8)$$

Όπου $O_i = Y_i$ και $E_i = \hat{Y}_i$.

Έτσι προκύπτει

$$X^2 = \sum r_i^2 = \sum \left(\frac{O_i - E_i}{\sqrt{E_i}} \right)^2 \quad (3.9)$$

Η συνάρτηση Deviance δίνεται από τον τύπο:

$$D = 2 \sum \left\{ O_i \log\left(\frac{O_i}{E_i}\right) - (O_i - E_i) \right\} \quad (3.10)$$

Και τα υπόλοιπα Deviance από τον τύπο:

$$d_i = \text{sign}(O_i - E_i)\sqrt{2\left[O_i \log\left(\frac{O_i}{E_i}\right) - (O_i - E_i)\right]} \quad (3.11)$$

Οπότε $D = \sum_i^n d_i^2$ έτσι απορρίπτω το μοντέλο, αν σε επίπεδο σημαντικότητας α, το D ή το X^2 είναι μεγαλύτερο του X^2_{N-p}.

3.2.Παράδειγμα

Τα δεδομένα του πίνακα αναφέρονται σε μία μελέτη όπου γιατροί στη Βρετανία απάντησαν σε ένα ερωτηματολόγιο σχετικά με το αν είναι καπνιστές ή όχι. Ο πίνακας δείχνει τον αριθμό θανάτων από στεφανιαία νόσο μετά από 10 χρόνια. Δείχνει επίσης και τον ολικό πληθυσμό.

Ηλικίες	Καπνιστές		Μη-καπνιστές	
	Θάνατοι	Πληθυσμός	Θάνατοι	Πληθυσμός
35-44	32	52407	2	18790
45-54	104	43248	12	10673
55-64	206	28612	28	5710
65-74	186	12663	28	2585
75-84	102	5317	31	1462

Θα εξετάσουμε τρία ερωτήματα:

1. Είναι τα ποσοστά θανάτου πιο υψηλά στους καπνιστές;
2. Αν ναι, κατά πόσο;
3. Υπάρχει διαφοροποίηση λόγω ηλικίας;

Θα μοντελοποιήσουμε τα δεδομένα του παραπάνω πίνακα με δεδομένο ότι οι θάνατοι είναι ανάλογοι των ανθρωποετών (αριθμός ανθρώπων από τους οποίους έγινε η λήψη των δεδομένων τα χρόνια που διήρκησε η έρευνα).

Για να γίνει η μοντελοποίηση πρέπει να λάβουμε υπόψη ότι πρέπει να απαντήσουμε στα παραπάνω ερωτήματα.

Η εξαρτημένη μεταβλητή θα είναι ο αριθμός θανάτων. Ενώ οι παράγοντες ηλικία και αν ο γιατρός που μελετάται είναι καπνιστής ή όχι είναι οι επεξηγηματικές μεταβλητές.

Θα κατηγοριοποιήσουμε την ηλικία ως εξής:

Ηλικία

35-44 ⟶ 1

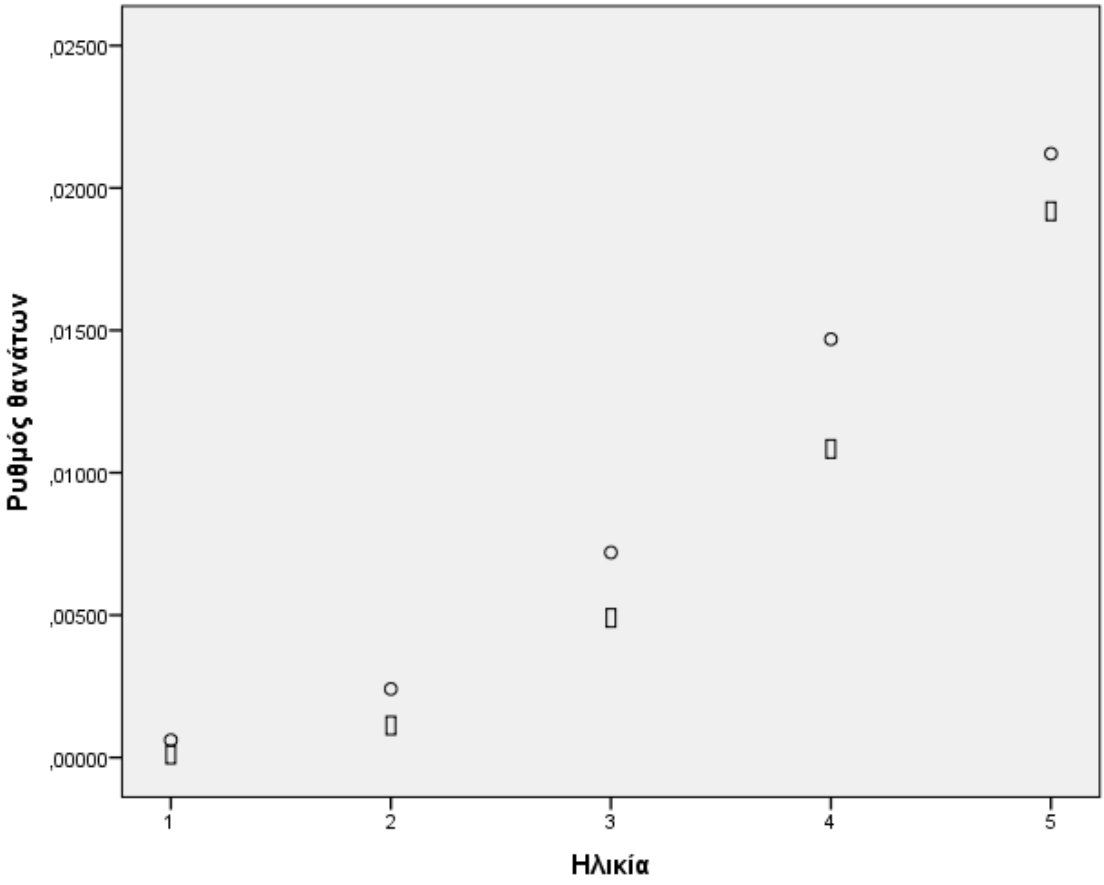
45-54 → 2

55-64 → 3

65-74 → 4

75-84 → 5

Καθώς θα θεωρήσουμε και τη ψευδομεταβλητή κάπνισμα=1 αν είναι καπνιστής και κάπνισμα=0 αν δεν είναι.



Από το παραπάνω γράφημα που παρουσιάζει το ρυθμό θανάτων (θάνατοι ανά ανθρωποέτη) σε συνάρτηση της ηλικιακής ομάδας για τους καπνιστές και τους μη καπνιστές, παρατηρούμε ότι ο ρυθμός θανάτου είναι μεγαλύτερος στους καπνιστές από αυτό στους μη καπνιστές, με εξαίρεση την ηλικιακή ομάδα 5 (75-84), ενώ όσο μεγαλώνει η ηλικία μεγαλώνει και η διαφορά του ρυθμού θανάτων ανάμεσα στις δύο ομάδες.

Η σχέση μεταξύ ρυθμού θανάτου και ηλικίας δεν είναι γραμμική και για αυτό το λόγο θα συμπεριλάβουμε ως επεξηγηματική μεταβλητή και το τετράγωνο της ηλικίας,

ηλικία2, και όρο αλληλεπίδρασης μεταξύ του καπνίσματος και της ηλικίας. Θα προσαρμόσουμε μοντέλο Poisson με offset το λογάριθμο των ανθροποετών. Το μοντέλο θα είναι της μορφής:

$$\log\left(\frac{\theta\acute{\alpha}νατοι}{\alpha\nu\theta\rho\omega\pi o\acute{\epsilon}τη}\right) = \beta_o + \beta_1 κάπνισμα + \beta_2 ηλικία + \beta_3 ηλικία^2 + \beta_4(ηλικία * κάπνισμα)$$

$$\log\left(\frac{\theta\acute{\alpha}νατοι}{\alpha\nu\theta\rho\omega\pi o\acute{\epsilon}τη}\right) = \beta_o + \beta_1 κάπνισμα + \beta_2 ηλικία + \beta_3 ηλικία^2 + \beta_4(ηλικία * κάπνισμα) + offset$$

Model Information

Dependent Variable	Θάνατοι
Probability Distribution	Poisson
Link Function	Log
Offset Variable	ln(ανθρωποέτη)

Case Processing Summary

	N	Percent
Included	10	100,0%
Excluded	0	0,0%
Total	10	100,0%

Continuous Variable Information

		N	Minimum	Maximum	Mean	Std. Deviation
Dependent Variable	Θάνατοι	10	2	206	73,10	73,422
Covariate	Ηλικία	10	1	5	3,00	1,491
	Ηλικία2	10	1	25	11,00	9,117
	Κάπνισμα	10	0	1	,50	,527
	Ηλικία*Κάπνισμα	10	0	5	1,50	1,900
Offset	ln(ανθρωποέτη)	10	7	11	9,27	1,186

Μέτρα καλής προσαρμογής του μοντέλου:

Goodness of Fit[a]

	Value	df	Value/df
Deviance	1,635	5	,327
Scaled Deviance	1,635	5	
Pearson Chi-Square	1,550	5	,310
Scaled Pearson Chi-Square	1,550	5	

Log Likelihood[b]	-28,352		
Akaike's Information Criterion (AIC)	66,703		
Finite Sample Corrected AIC (AICC)	81,703		
Bayesian Information Criterion (BIC)	68,216		
Consistent AIC (CAIC)	73,216		

Dependent Variable: Θάνατοι

Model: (Intercept), Ηλικία, Ηλικία2, Κάπνισμα, Ηλικία * Κάπνισμα, offset = lnέτη

a. Information criteria are in small-is-better form.

b. The full log likelihood function is displayed and used in computing information criteria.

Το μοντέλο που εκτιμήθηκε βελτιώνει την ικανότητα πρόβλεψης σε σχέση με μοντέλο όπου όλοι οι συντελεστές πλην του σταθερού όρου είναι μηδέν:

Omnibus Test[a]

Likelihood Ratio Chi-Square	df	Sig.
933,432	4	,000

Dependent Variable: Θάνατοι

Model: (Intercept), Ηλικία, Ηλικία2, Κάπνισμα, Ηλικία * Κάπνισμα, offset = lnέτη

a. Compares the fitted model against the intercept-only model.

Tests of Model Effects

Source	Type III		
	Wald Chi-Square	df	Sig.
(Intercept)	1056,847	1	,000
Ηλικία	129,648	1	,000
Ηλικία2	52,173	1	,000
Κάπνισμα	14,989	1	,000
Ηλικία * Κάπνισμα	10,044	1	,002

Dependent Variable: Θάνατοι

Model: (Intercept), Ηλικία, Ηλικία2, Κάπνισμα, Ηλικία * Κάπνισμα, offset = lnέτη

Οι συντελεστές είναι στατιστικά σημαντικοί, συμπεριλαμβανομένου και του όρου αλληλεπίδρασης των παραγόντων:

Parameter Estimates

Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test			Exp(B)	95% Wald Confidence Interval for Exp(B)	
			Lower	Upper	Wald Chi-Square	df	Sig.		Lower	Upper
(Intercept)	-9,351	,2876	-9,915	-8,787	1056,847	1	,000	8,690E-005	4,945E-005	,000
Ηλικία	2,069	,1817	1,713	2,425	129,648	1	,000	7,916	5,544	11,303
Ηλικία2	-,198	,0274	-,251	-,144	52,173	1	,000	,821	,778	,866
Κάπνισμα	-1,441	,3722	-2,170	-,711	14,989	1	,000	,237	,114	,491
Ηλικία * Κάπνισμα	,308	,0970	,117	,498	10,044	1	,002	1,360	1,125	1,645
(Scale)	1[a]									

Dependent Variable: Θάνατοι

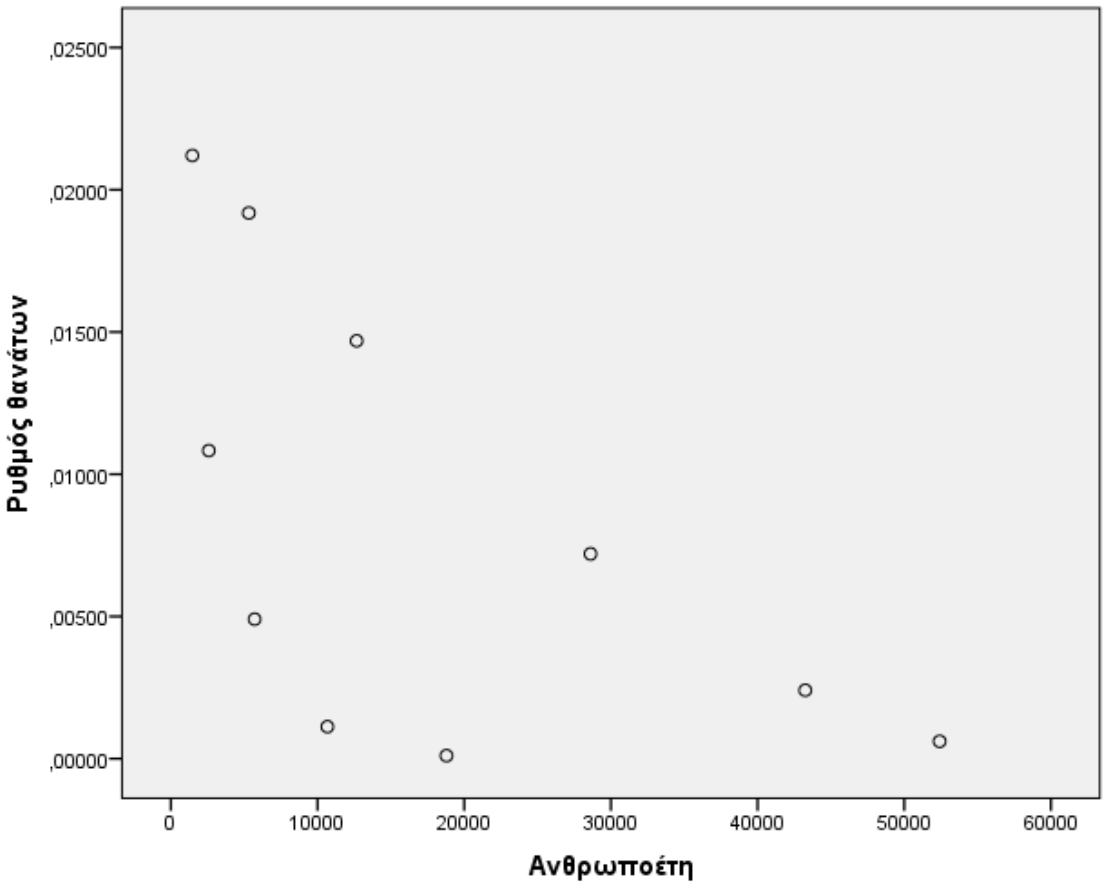
Model: (Intercept), Ηλικία, Ηλικία2, Κάπνισμα, Ηλικία * Κάπνισμα, offset = lnέτη

a. Fixed at the displayed value.

Το μοντέλο που εκτιμήθηκε είναι:

$$\log(\theta\acute{\alpha}\nu\alpha\tau o\iota) = -10{,}792 + 1{,}441\kappa\acute{\alpha}\pi\nu\iota\sigma\mu\alpha + 2{,}376\eta\lambda\iota\kappa\acute{\iota}\alpha - 0{,}198\eta\lambda\iota\kappa\acute{\iota}\alpha^2 - 0{,}308(\eta\lambda\iota\kappa\acute{\iota}\alpha * \kappa\acute{\alpha}\pi\nu\iota\sigma\mu\acute{\alpha}) + offset$$

Όλοι οι συντελεστές είναι στατιστικά σημαντικοί. Οπότε η πιθανότητα θανάτου από καρδιακή ανεπάρκεια επηρεάζεται από το αν είναι κάποιος καπνιστής ή όχι, καθώς και από την ηλικία του. Από τη στήλη Exp(B) του πίνακα των εκτιμημένων συντελεστών βγάζουμε το συμπέρασμα ότι όταν οι υπόλοιπες μεταβλητές είναι σταθερές το ρίσκο θανάτου για τους καπνιστές είναι 4,2 φορές μεγαλύτερο από αυτό των μη καπνιστών.


,02500
,02000
,01500
,01000
,00500
,00000
Ρυθμός θανάτων
0
10000
20000
30000
40000
50000
60000
Ανθρωποέτη

4. Επίλογος

4.1. Γενικευμένα γραμμικά μοντέλα

Όλα τα μοντέλα παλινδρόμησης, γραμμικά και μη γραμμικά, ανήκουν σε μια κατηγορία που λέγετε γενικευμένα γραμμικά μοντέλα. Τα εισήγαγαν αρχικά οι Nelder και Wedderburn και περιγράφουν γραμμικά μοντέλα κανονικού σφάλματος και μη γραμμικά εκθετικά λογιστικά και Poisson μοντέλα παλινδρόμησης.

Η δομή των γενικευμένων γραμμικών μοντέλων συνίσταται από τα εξής τρία στοιχεία:

- Το στοιχείο τυχαιότητας (random component) που καθορίζει την υποθετική κατανομή της μεταβλητής απόκρισης Y_i δοθέντων των εκτιμητών. Οι $Y_i, \dots, Y_n$ είναι ανεξάρτητες αποκρίσεις που ακολουθούν μια κατανομή που ανήκει στην εκθετική οικογένεια με αναμενόμενη τιμή $E\{Y_i\} = \mu_i$.

- Μια γραμμική συνάρτηση των συντελεστών παλινδρόμησης, από την οποία εξαρτάται η μέση τιμή του Y_i που καλείται γραμμικός εκτιμητής (linear predictor) και βασίζεται στις $X_{i1}, \dots, X_{i,p-1}$ (μεταβλητές πρόβλεψης). Θα δηλώνεται με $X_i'\beta$ όπου: $X_i'\beta = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_{p-1} X_{i,p-1}$
- Μια αντιστρέψιμη συνάρτηση σύνδεσης (link function) g η οποία μετασχηματίζει τη μέση τιμή της απόκρισης στην γραμμική εκτίμηση, δηλαδή $X_i'\beta = g(\mu_i)$. Η αντίστροφη συνάρτηση της συνάρτησης σύνδεσης λέγεται και συνάρτηση μέσης τιμής (mean function) και προφανώς θα ισχύει $g^{-1}(X_i'\beta) = \mu_i$.

5. Σχόλια:

Τα γενικευμένα γραμμικά μοντέλα είναι πιθανό να έχουν μη σταθερές διασπορές σ_i^2 για τις αποκρίσεις Y, αλλά η διασπορά σ_i^2 θα πρέπει να είναι μια συνάρτηση των μεταβλητών πρόβλεψης μέσω της μέσης τιμής μ_i. Για να επεξηγήσουμε

την έννοια της link function θεωρούμε το απλό λογιστικό μοντέλο παλινδρόμησης εκεί ο logit μετασχηματισμός της χρησιμοποιείται για να συνδέσει τον linear predictor με την μέση τιμή $\mu_i = \pi_i$

$$g((\mu_i)) = g((\pi_i)) = \log_e \frac{\pi_i}{1-\pi_i} = X_i'\beta$$

Σαν δεύτερο παράδειγμα ας θεωρήσουμε το μοντέλο Poisson, εκεί θεωρήσαμε διάφορες συναρτήσεις απόκρισης. Από την συνάρτηση απόκρισης $(\mu_i) = \exp(X_i'\beta)$ η link function θα είναι $g((\mu_i)) = \log_e((\mu_i)) = X_i'\beta$.

Παρατηρούμε από τα μοντέλα Poisson ότι μπορεί να υπάρχουν πολλές διαφορετικές πιθανές link functions που μπορούν να χρησιμοποιηθούν αρκεί να είναι μονότονες και διαφορίσιμες.

Πλέον μπορούμε να εξηγήσουμε ποια είναι η αιτία αυτής της κατηγοριοποίησης, περιγράφοντας συνοπτικά τον κοινό τρόπο ανάλυσης των γενικευμένων γραμμικών μοντέλων. Κάθε μοντέλο παλινδρόμησης που ανήκει στην οικογένεια των γενικευμένων γραμμικών μοντέλων μπορεί να αναλυθεί, μελετηθεί με ένα ενοποιημένο τρόπο. Οι εκτιμητές μέγιστης πιθανοφάνειας των παραμέτρων παλινδρόμησης μπορούν να εξασφαλιστούν από επαναληπτικά ανασταθμισμένα ελάχιστα τετράγωνα. Έλεγχοι για ανάπτυξη μοντέλου, για να καθοριστεί εάν κάποιες μεταβλητές πρόβλεψης μπορούν να απορριφθούν από το μοντέλο, μπορούν να διεξαχθούν χρησιμοποιώντας partial deviances. Τέλος, έλεγχος καλής προσαρμογής μπορεί να διεξαχθεί με χρήση model deviance.

6. Βιβλιογραφία

- Applied Linear Regression Models, Neter – Kutner – Nachtsheim – Wasserman , Third Edition
- Generalized Linear Models: An Introduction, York Summer Programme in Data Analysis, John Fox – May 2005 – Mc Master University (http://socserv.mcmaster.ca/jfox/courses/SPIDA/GLMs-handout.pdf)
- Linear Regression and Least Squares Estimation, Guy Lebanon – April 25, 2007 (http://www.stat.purdue.edu/~lebanon/notes/linReg.pdf)
- Bias, Variance and MSE of Estimator, Guy Lebanon – February 14, 2006 (http://www.stat.purdue.edu/~lebanon/notes/estimators1.pdf)
- Confidence Intervals, Guy Lebanon – February 23, 2006 (http://www.stat.purdue.edu/~lebanon/notes/confInt.pdf)
- Maximum Likelihood Estimation, Guy Lebanon – May 13, 2006 (http://www.stat.purdue.edu/~lebanon/notes/mle.pdf)
- Sampling Distributions, Guy Lebanon – February 14, 2006 (http://www.stat.purdue.edu/~lebanon/notes/samplingDist.pdf)
- Asymptotic Efficiency of the Maximum Likelihood Estimator, Guy Lebanon – January 5, 2008

(http://www.stat.purdue.edu/~lebanon/notes/mleEfficiency.pdf)

- Logistic Regression Model, Professor Jon E. Anderson

(cda.morris.umn.edu/~anderson/math3611/notes/logistic.pdf)

- Ordinary Least Squares and Poisson Regression Models,

Luc Anselin, University of Illinois

(http://www.icpsr.umich.edu/CRIMESTAT/files/CrimestatAppendix.C.pdf)

- Chi-Square: Testing for Goodness of Fit, Peter Scott

(http://physics.ucsc.edu/~drip/133/ch4.pdf)

- Chi Square Goodness-of-Fit Test, Engineering Statistics

Handbook

(www.itl.nist.gov/div898/handbook/eda/section3/eda35f.htm

- http://labs.fme.aegean.gr/decision/files/docs/Odigos-SPSS-Pramaggioulis.pdf
- http://users.auth.gr/dkugiu/Teach/DataAnalysis/Chp5.pdf
- http://194.42.1.1/~fokianos/GreekRbook/poissondata.pdf