ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΟΜΕΑΣ ΜΑΘΗΜΑΤΙΚΩΝ

Διπλωματική Εργασία

Γενικευμένα Γραμμικά Μοντέλα & Παραγοντικοί Σχεδιασμοί

ΜΙΧΑΛΗ ΘΑΛΑΣΣΙΑ

Επιβλέπων:

Κουκουβίνος Χρήστος, Καθηγητής Ε.Μ.Π.

ΑΘΗΝΑ, ΜΑΡΤΙΟΣ 2011

ΠΕΡΙΛΗΨΗ

Στη Στατιστική, τα Γενικευμένα Γραμμικά Μοντέλα αποτελούν μια επέκταση του Γενικού Γραμμικού Μοντέλου . Τα Γενικευμένα Γραμμικά Μοντέλα προτάθηκαν από τους John Nelder , Robert Weddeburn ως μια ενοποίηση διαφόρων άλλων στατιστικών μοντέλων π.χ. Λογιστική ή Poisson παλινδρόμηση.

Αρχικά , παρουσιάζονται τα Γενικευμένα Γραμμικά Μοντέλα και η μέθοδος Μεγίστης Πιθανοφάνειας για την εύρεση εκτιμητριών των παραμέτρων. Στη συνέχεια, παρουσιάζονται δυο μοντέλα παλινδρόμησης . Ιδιαίτερη προσοχή δίνεται στην επιλογή βέλτιστου σχεδιασμού (βελτιστοποιήσεις A, D, D_S , T , T_E) άλλα και στις προσεγγίσεις που αναπτύχθηκαν επίσης όπως Μπεϋζιανοί σχεδιασμοί, Διαδοχικοί σχεδιασμοί, Πολυσταδιακοί σχεδιασμοί κτλ. Τέλος, παρουσιάζονται 2^k παραγοντικοί σχεδιασμοί με Δυαδική απόκριση .

Θα ήταν παράλειψη μου να μην ευχαριστήσω τον Καθηγητή ΕΜΠ κ. Χρήστο Κουκουβίνο αλλά και τον υποψήφιο διδάκτορα του Τομέα Μαθηματικών ΕΜΠ κ. Εμμανουήλ Ανδρουλάκη για την καθοδήγηση, το χρόνο τους αλλά και την πολύτιμη βοήθεια τους στην συγγραφή αυτής της εργασίας. Θα ήθελα να ευχαριστήσω επίσης την οικογένεια μου για την αμέριστη συμπαράσταση που έδειξαν στο πρόσωπο μου καθ' όλη τη διάρκεια των σπουδών μου.

Μάρτιος, 2011

Θαλασσία Μιχάλη

ABSTRACT

In Statistics Generalized Linear Models (GLMs) are an extention of the General Linear Model.Generalized linear models were formulated by John Nelder and Robert Wedderburn as a way of unifying various other statistical models, including logistic regression and Poisson regression.

Firstly, we introde the GLMs and the method of Maximum Likelihood for the estimation of the model's parameters. Then we introduce two regression models. Special attention is paid in the choice of optimal designs (optimalities A,D, D_S , T , T_E) and the approximations that were evolved such as Bayesian designs, Sequential designs, Multistage designs etc. Finally, 2^k factorial designs with Binary response are presented.

It would be a great omission if i didn't thank Professor of EMP Mr. C. Koukouvinos and the Phd Candidate of the Mathematics Department of EMP Mr. Emmanuel Androulakis for their guidance, time and valuable help while writing this thesis. I would also like to express my thanks to my family who has been very supportive throughout my studies.

March, 2011.

Thalassia Michali

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

ΚΕΦΑΛΑΙΟ 1 : ΕΙΣΑΓΩΓΗ 5
1.1 Ορισμοί 5
1.2 Εφαρμογές μοντέλων 6
1.3 Γενικευμένα Γραμμικά Μοντέλα 7
ΚΕΦΑΛΑΙΟ 2 : ΓΕΝΙΚΕΥΜΕΝΑ ΓΡΑΜΜΙΚΑ ΜΟΝΤΕΛΑ 10
2.1 Η Εκθετική Οικογένεια 10
2.2 Σημαντικές Κατανομές στα GLM 13
2.3 Μεθοδολογία για Γενικευμένα Γραμμικά Μοντέλα 16
2.4 Μέθοδος Μέγιστης Πιθανοφάνειας στα GLM 17
2.5 Μέθοδος Πιθανοφάνειας Quasi 20
2.6 Ομάδα Συναρτήσεων Σύνδεσης – Συνάρτηση Δύναμης 21
2.7 Έλεγχοι Καταλληλότητας και Υπόλοιπα στα GLM 22
ΚΕΦΑΛΑΙΟ 3 : ΜΟΝΤΕΛΑ ΠΑΛΙΝΔΡΟΜΗΣΗΣ ΣΤΑ GLM 26
3.1 Λογιστική και Poisson Παλινδρόμηση 26
3.2 Μέγιστη Πιθανοφάνεια στη Λογιστική Παλινδρόμηση 27
3.3 Ιδιότητες Διασποράς της Μέγιστης Πιθανοφάνειας 30
3.4 Έλεγχος Wald στη Λογιστική Παλινδρόμηση 31
3.5 Καταλληλότητα Μοντέλου (Απόκλιση (Deviance)) 33
3.6 Probit , Complementary Log –Log Μοντέλα για Δυαδικές Αποκρίσεις 34
3.7 Υπερ-διασπορά στο Λογιστικό Μοντέλο Παλινδρόμησης 35
3.8 Poisson Παλινδρόμηση: Μέγιστη Πιθανοφάνεια και Απόκλιση 37
ΚΕΦΑΛΑΙΟ 4: ΕΦΑΡΜΟΓΕΣ ΤΩΝ ΓΕΝΙΚΕΥΜΕΝΩΝ ΓΡΑΜΜΙΚΩΝ ΜΟΝΤΕΛΩΝ ... 40
4.1 Παραγοντικοί Σχεδιασμοί και Βελτιστοποίηση 40
4.2 Δυσκολίες στην Εύρεση Βέλτιστου Μοντέλου στα GLM 43
4.3 Επιλογή Σχεδιασμού 46
4.3.1 Λογιστικό Μοντέλο Παλινδρόμησης για Δυαδικά Δεδομένα 46
4.3.2 Poisson μοντέλο καταμέτρησης 56
4.4 Τύποι Σχεδιασμών 57
4.4.1 Ορθογώνιοι Σχεδιασμοί στα GLM 57
4.4.2 Διαδοχικοί Σχεδιασμοί 63
4.4.3 Πολυσταδιακοί Σχεδιασμοί 64

4.4.4 Μπεϋζιανοί Σχεδιασμοί και Προβλήματα 64
4.5 Ειδική Περίπτωση : Περισσότερες από μια Επεξηγηματικές Μεταβλητές 68
4.6 Ασυμπτωτικά αποτελέσματα 70
4.7 GLM, Πειράματα Κρησαρίσματος και Μετασχηματισμοί Δεδομένων 74
4.8 Μοντελοποίηση Μέσου και Διακύμανσης μέσω GLM 75
4.9 Δυαδικής Απόκρισης Μοντέλα 77
ΚΕΦΑΛΑΙΟ 5: ΣΧΕΔΙΑΣΜΟΙ ΜΕ ΔΥΑΔΙΚΗ ΑΠΟΚΡΙΣΗ & ΚΡΙΤΗΡΙΑ ΕΠΙΛΟΓΗΣ ΜΟΝΤΕΛΩΝ ΓΙΑ ΔΙΩΝΥΜΙΚΕΣ ΑΠΟΚΡΙΣΕΙΣ 79
5.1 Βέλτιστοι Σχεδιασμοί 2^k Παραγοντικών Σχεδιασμών για Δυαδικές Αποκρίσεις ... 79
5.1.1 2^2 Παραγοντικός Σχεδιασμός 80
5.1.2 Αναλυτικές λύσεις συγκεκριμένων περιπτώσεων 82
5.1.3 Ακριβείς λύσεις χρησιμοποιώντας την μέθοδο CAD 85
5.1.4 Αναλυτικές κατά προσέγγιση λύσεις 85
5.1.5 Ευρωστία 2^2 Σχεδιασμών 86
5.2 Γενική περίπτωση : 2^k πειράματα 89
ΠΑΡΑΡΤΗΜΑ 91
ΒΙΒΛΙΟΓΡΑΦΙΑ 105

ΚΕΦΑΛΑΙΟ 1 :
ΕΙΣΑΓΩΓΗ

1.1 Ορισμοί

Με τον όρο **μοντέλο** αναφερόμαστε σε μια απεικόνιση της πραγματικότητας η οποία μας παρέχει μια προσέγγιση μερικών πιο περίπλοκων θεμάτων. Τα μοντέλα μπορούν να χωριστούν σε δυο κατηγορίες:

α) ντετερμινιστικά

β) μη στοχαστικά

Στα **ντετερμινιστικά μοντέλα**, τα αποτελέσματα και οι αποκρίσεις είναι επακριβώς καθορισμένα συχνά από μια σειρά εξισώσεων π.χ νόμος του Ohm($E=I\cdot R$), νόμος των ιδανικών αερίων ($P\cdot V=n\cdot R\cdot T$), πρώτος νόμος Θερμοδυναμικής

($\int dW= J \int dQ$).

Στα **μη στοχαστικά**, τα αποτελέσματα και οι αποκρίσεις παρουσιάζουν μεταβλητότητα επειδή είτε το μοντέλο περιέχει τυχαίες μεταβλητές είτε επηρεάζεται κατά κάποιο τρόπο από τυχαίες δυνάμεις.

Ένα από τα πιο σημαντικά μη στοχαστικά μοντέλα είναι το γραμμικό μοντέλο:

$$y = \beta_0 + \beta_1 x_1 + \ldots\ldots\ldots\ldots\ldots + \beta_\kappa x_\kappa + \varepsilon \qquad (1.1)$$

όπου :

y: απόκριση ή αποτέλεσμα

$x_1, x_2, \ldots\ldots, x_k$: μεταβλητές παλινδρόμησης ή συντελεστές

$\beta_0, \beta_1, \ldots\ldots, \beta_k$: άγνωστες παράμετροι

ε : τυχαίο σφάλμα που ακολουθεί την κανονική κατανομή $N(0,\sigma^2)$

Η σχέση (1.1) συχνά καλείται **γραμμικό μοντέλο παλινδρόμησης.**

Από τα παραπάνω προκύπτει ότι η μέση τιμή του αποτελέσματος στο απλό γραμμικό μοντέλο είναι:

$$\mathrm{E}(y) = \beta_0 + \beta_1 x_1 + \ldots\ldots\ldots\ldots\ldots + \beta_\kappa x_\kappa \qquad (1.2)$$

Η σχέση (1.1) καλείται **απλό γραμμικό μοντέλο** επειδή η μέση τιμή της απόκρισης είναι γραμμική συνάρτηση των άγνωστων παραμέτρων $\beta_0, \beta_1, \ldots\ldots, \beta_k$.

Παραδείγματα γραμμικών μοντέλων αποτελούν τα παρακάτω:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \varepsilon$$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \varepsilon$$

$$y = \beta_0 + \beta_1 \sin\frac{2\pi x}{12} + \beta_2 \cos\frac{2\pi x}{12} + \varepsilon$$

1.2 Εφαρμογές μοντέλων

Τα γραμμικά μοντέλα παλινδρόμησης έχουν πολλές εφαρμογές:

- Πρώτον, αποτελούν προσεγγιστικά πολυώνυμα για πιο πολύπλοκες σχέσεις και για το λόγο αυτό αναφέρονται και ως **εμπειρικά μοντέλα**. Για παράδειγμα, στην περίπτωση E(y) = f(x) η σειρά Taylor πρώτης τάξης προσεγγίζει την παραπάνω σχέση στο σημείο x_0ως εξής :

$$\mathrm{E}(\mathrm{y}) = \mathrm{f}(x_0) + \left.\frac{df(x)}{dx}\right|_{x=x_0} (\mathrm{x} - x_0) + \mathrm{R} = \beta_0 + \beta_1 (x_1 - x_0) \qquad (1.3)$$

  Αν αγνοήσουμε το υπόλοιπο R στη σχέση (1.3) τότε η σχέση αυτή αποτελεί ένα γραμμικό μοντέλο με ένα παράγοντα.
- Δεύτερον, επειδή αποτελούν μια πιο σύντομη διαδικασία υπολογισμού των αγνώστων παραμέτρων $\beta_0, \beta_1, \ldots\ldots, \beta_k$.
- Τρίτον, υπάρχει μια πολύ καλά οργανωμένη στατιστική θεωρία γύρω από τα γραμμικά μοντέλα, η οποία δεδομένου ότι το τυχαίο σφάλμα ε ακολουθεί την κανονική κατανομή (~N(0,σ^2)), δίνει πολύ ικανοποιητικά αποτελέσματα για τα διαστήματα εμπιστοσύνης, τις προβλέψεις των αποκρίσεων y αλλά και των άγνωστων παραμέτρων β .

Στις περιπτώσεις όπου η απόκριση y δεν είναι γραμμική σχέση των αγνώστων παραμέτρων β τότε λέμε ότι έχουμε **μη γραμμικό μοντέλο** παλινδρόμησης π.χ. στον νόμο ψύξης του Νεύτωνα η απόκριση y δίνεται από τον εξής τύπο:

$$y = T_A + (T_1 - T_A)e^{\beta t} + \varepsilon \; y = T_A + (T_1 - T_A)e^{\beta t} + \varepsilon$$

όπου:

y : στιγμιαία θερμοκρασία αντικειμένου

T_A: θερμοκρασία περιβάλλοντος

T_1: αρχική θερμοκρασία αντικειμένου

ε : τυχαίο σφάλμα που ακολουθεί την κανονική κατανομή N(0, σ^2)

Μη γραμμικά μοντέλα συναντούμε στη Μηχανική για το λόγο αυτό αναφέρονται και ως **μηχανιστικά μοντέλα**. Όπως στα γραμμικά μοντέλα, έτσι και στα μη γραμμικά μοντέλα υπάρχει αντίστοιχη στατιστική θεωρία διότι είναι απαραίτητος ο υπολογισμός των άγνωστων παραμέτρων β και των διαστημάτων εμπιστοσύνης.

1.3 Γενικευμένα Γραμμικά Μοντέλα

Και στις δυο περιπτώσεις, γραμμικά και μη γραμμικά μοντέλα, κυρίαρχο ρόλο έπαιξε η Κανονική κατανομή $N(\mu, \sigma^2)$ την οποία υποθέσαμε ότι ακολουθεί η απόκριση.

Αυτό όμως δε συμβαίνει πάντοτε π.χ. η απόκριση μπορεί να ακολουθεί την Διωνυμική κατανομή δηλ. τα αποτελέσματα να είναι της μορφής 0 (=αποτυχία) ή 1 (=επιτυχία). Στην περίπτωση αυτή, έχουμε τα **Γενικευμένα Γραμμικά Μοντέλα(Generalized Linear Models ή GLM)** των οποίων τα δεδομένα ακολουθούν κατανομές της Εκθετικής Οικογένειας.

Η Εκθετική Οικογένεια αποτελείται από τις εξής κατανομές:

- Κανονική
- Διωνυμική
- Poisson

- Γεωμετρική
- Αρνητική Διωνυμική
- Εκθετική
- Γάμμα
- αντίστροφες της Κανονικής.

Για παράδειγμα, αν y_i οι αποκρίσεις το **GLM** δίνεται από τον τύπο :

$$g(\mu_i) = g[E(y_i)] = x_i'\beta$$

όπου:

x_i : διάνυσμα των μεταβλητών παλινδρόμησης

β : διάνυσμα των άγνωστων παραμέτρων

Κάθε γενικευμένο γραμμικό μοντέλο αποτελείται από τρία συστατικά :

α) την κατανομή της απόκρισης.

β) μια γραμμική παράμετρο πρόβλεψης που περιέχει τις μεταβλητές παλινδρόμησης x_i.

γ) την συνάρτηση σύνδεσης η οποία ενώνει τη γραμμική παράμετρο πρόβλεψης με τη μέση τιμή της απόκρισης.

Στην σχέση (1.1) για παράδειγμα, η κατανομή της απόκριση της είναι η Κανονική, η γραμμική παράμετρο πρόβλεψης είναι :

$$x'\beta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 ..... + \beta_\kappa x_\kappa + \varepsilon$$

ενώ η συνάρτηση σύνδεσης δίνεται από τον τύπο :

$$E(y) = \mu = \beta_0 + \beta_1 x_1 + \beta_2 x_2 ..... + \beta_\kappa x_\kappa$$

δηλ. το γραμμικό μοντέλο της σχέσης (1.1) είναι και ένα GLM.

Ωστόσο, GLM μπορεί να είναι και ένα μη γραμμικό μοντέλο. Αυτό εξαρτάται από τη συνάρτηση σύνδεσης. Για παράδειγμα, αν χρησιμοποιήσουμε ως συνάρτηση σύνδεσης g(a) = lna τότε έχουμε:

$$E(y) = \mu = e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 \ldots\ldots\ldots\ldots\ldots + \beta_\kappa x_\kappa}$$

Γενικότερα, τα GLM μπορούμε να τα βλέπουμε ως μια ενοποίηση γραμμικών και μη γραμμικών κατανομών απόκρισης που ενσωματώνουν αποκρίσεις μιας μεγάλης οικογένεια κανονικών ή μη κανονικών κατανομών

ΚΕΦΑΛΑΙΟ 2 :
ΓΕΝΙΚΕΥΜΕΝΑ ΓΡΑΜΜΙΚΑ ΜΟΝΤΕΛΑ

2.1 Η Εκθετική Οικογένεια

Στα GLM σημαντικό ρόλο παίζουν η κατανομή της απόκρισης και το μοντέλο που συνδέει τη μέση απόκριση με τις μεταβλητές παλινδρόμησης. Μάλιστα, οι δυο αυτοί παράγοντες δεν είναι εντελώς ανεξάρτητοι μεταξύ τους αλλά συσχετίζονται.

Όπως αναφέρθηκε προηγουμένως τα GLM εφαρμόζονται σε περιπτώσεις που οι αποκρίσεις ακολουθούν κατανομές της Εκθετικής Οικογένειας δηλ. έχουν συνάρτηση πυκνότητας πιθανότητας της μορφής :

$$f(y;\theta,\varphi) = \exp\left\{\frac{y\theta - b(\theta)}{\alpha(\varphi)} + c(y,\varphi)\right\} \qquad (2.1)$$

όπου:

α,b,c : συγκεκριμένες συναρτήσεις

θ : παράμετρος θέσης

φ : παράμετρος διασποράς

α(φ) : = φω με ω γνωστή σταθερά.

Για μερικές κατανομές μέλη της Εκθετικής Οικογένειας όπως η Διωνυμική, η Poisson ισχύει φ=1.0 εκτός από περιπτώσεις όπου έχουμε πολύ μεγάλη διασπορά όπου πρέπει να λάβουμε υπόψη ακόμη έναν παράγοντα.

Για παράδειγμα, για την Κανονική κατανομή όπου η απόκριση y έχει μέση τιμή μ και η διασπορά σ είναι :

$$f(y;\mu,\sigma) = \exp\left\{\frac{-[y-\mu]^2}{2\sigma^2}\right\}\frac{1}{\sqrt{2\pi}\sigma} = \exp\left\{\frac{-[y\mu-\mu^2/2]}{\sigma^2} - \frac{1}{2}[\frac{y^2}{\sigma^2} + \ln(2\pi\sigma^2)]\right\}$$

όπου:

$$\theta = \mu$$

$$b(\theta) = \mu^2 / 2$$

$$\alpha(\varphi) = \varphi$$

$$\varphi = \sigma^2$$

$$c(y,\varphi) = -\frac{1}{2}[\frac{y^2}{\sigma^2} + \ln(2\pi\sigma^2)]$$

Αντίστοιχα, στην κατανομή Poisson η συνάρτηση πυκνότητας πιθανότητας είναι:

$$f(y;\mu) = \frac{e^{-\mu}\mu^y}{y!} = \exp[y\ln\mu - \mu - \ln(y!)]$$

όπου:

$$\theta = \ln\mu,$$

$$b(\theta) = e^{\theta}$$

$$\varphi = 1.0$$

$$c(y,\varphi) = -\ln(y!).$$

Στην Διωνυμική κατανομή με παραμέτρους n και P έχουμε:

$$\theta = \ln(P/1-P)$$

$$b(\theta) = n\ln(1 + e^{\theta})$$

$$\varphi = 1.0$$

$$\alpha(\varphi) = 1.0$$

$$c(y,\varphi) = \ln\binom{n}{y}$$

Είναι εύκολο (Bickel and Doksom ,2001) να δείξουμε ότι για τα μέλη της Εκθετικής Οικογένειας ισχύει:

$$E(\frac{\partial \ln L(\theta)}{\partial \theta}) = 0$$

$$E(\frac{\partial^2 \ln L(\theta)}{\partial \theta^2}) + \mathrm{E}(\frac{\partial \ln L(\theta)}{\partial \theta})^2 = 0$$

Και έχουμε :

$$\mu = \mathrm{E}(y) = \frac{db(\theta)}{d(\theta)} = b'(\theta)$$

$$\mathrm{var}(y) = \frac{d^2 b(\theta)}{d\theta^2}\alpha(\varphi) = b''(\theta)\alpha(\varphi)$$

$$= \frac{d\mu}{d\theta}a(\varphi)$$

Η var μ διακύμανση της απόκρισης y είναι :

$$\mathrm{var}_{\mu}(y) = \frac{\mathrm{var}(y)}{\alpha(\varphi)} = \frac{d\mu}{d\theta}$$

Συγκεντρωτικά, ισχύουν :

- Για την Κανονική κατανομή :

$$\theta = \mu$$

$$b(\theta) = \mu^2 / 2$$

$$\alpha(\varphi) = \sigma^2$$

$$\mathrm{E}(y) = \frac{db(\theta)}{d(\theta)} = \mu$$

$$\mathrm{var}(y) = \frac{d^2 b(\theta)}{d\theta^2}\alpha(\varphi) = \sigma^2$$

- Για την Poisson κατανομή :

$$\theta = \ln \mu$$

$$\mu = \exp(\theta)$$

$$b(\theta) = \mu$$

$$\alpha(\varphi) = 1.0$$

$$c(y,\theta) = -\ln(y!)$$

$$\mathrm{E}(y) = \frac{db(\theta)}{d(\theta)} = \frac{db(\theta)}{d\mu}\frac{d\mu}{d\theta} = 1\exp(\theta) = \mu$$

$$\mathrm{var}(y) = \frac{d\mu}{d\theta} = \frac{dE(y)}{d\theta} = \exp(\theta) = \mu$$

2.2 Σημαντικές Κατανομές στα GLM

Δυο πολύ σημαντικές κατανομές των GLM είναι η Εκθετική και η Γάμμα. Ειδικότερα η Εκθετική είναι μια περίπτωση της Γάμμα αλλά κάθε μια έχει σημαντικές εφαρμογές. Θα μελετήσουμε την κάθε μια ξεχωριστά.

Η Εκθετική:

Η συνάρτηση πυκνότητας πιθανότητας της κατανομής είναι :

$$f(y) = \frac{1}{\lambda}e^{(-y/\lambda)} \qquad y>0;\ \lambda>0$$

ή σύμφωνα με την εξίσωση (2.1) :

$$f(y) = \exp\{(-1)[y(\frac{1}{\lambda}) + \ln\lambda]\ \}$$

όπου :

$$\alpha(\varphi) = -1$$

$$\theta = 1/\lambda$$

$$b(\theta) = \ln\theta$$

$$c(y,\theta) = 0$$

$$\mu = \lambda$$

$$\sigma^2 = \lambda^2$$

Υποθέτοντας την κανονική συνάρτηση συσχέτιση έχουμε $\eta_i = \theta_i$:

$$1/\lambda = x'\beta \rightarrow 1/\mu = x'\beta \rightarrow \mu = 1/x'\beta$$

Η Γάμμα:

Η κατανομή αυτή έχει εφαρμογή σε προβλήματα παλινδρόμησης όπου η απόκριση είναι συνεχής και η διακύμανση δεν είναι σταθερή αλλά ανάλογη του τετραγώνου του μέσου μ. Μια λύση στο πρόβλημα αυτό είναι να χρησιμοποιήσουμε λογαριθμικό μετασχηματισμό για να σταθεροποιηθεί η διακύμανση όπου όλοι οι συντελεστές είναι αμερόληπτοι και η απόκριση είναι λογαριθμική- κανονική. Η τομή είναι μεροληπτική κατά $(\sigma/\mu)^2/2$ αφού από την ανάλυση κατά Taylor έχουμε:

$$E[\ln y] = \ln\mu - (\sigma/\mu)^2/2$$

Η συνάρτηση πυκνότητας πιθανότητας της κατανομής είναι :

$$f(y) = \frac{1}{\Gamma(r)} \left(\frac{1}{\lambda}\right)' e^{(-y/\lambda)} y^{r-1} \qquad r>0,\ \lambda>0$$

η οποία βάσει της συνάρτησης (2.1) δίνει :

$$\theta = \frac{-1}{\lambda r} = \frac{-1}{\mu}$$

$$\mu = r\lambda$$

$$vary = \frac{\mu^2}{r} \rightarrow \frac{vary}{\mu^2} = r\lambda^2$$

$$\alpha(\varphi) = r^{-1}$$

$$b(\theta) = -\ln(-\theta)$$

$$c(\varphi) = r\ln r - \ln\Gamma(r) + (r-1)\ln y$$

όπου r είναι παράμετρος κλίμακας .Όταν r = 1 τότε έχουμε την Εκθετική. Στην περίπτωση όπου r δεν ποικίλει αλλά είναι σταθερό κατά την ανάλυση των δεδομένων τότε μέσω της παραμέτρου λ έχουμε αλλαγή του μέσου.

Κανονική Συνάρτηση Σύνδεσης : $\theta = x'\beta \rightarrow \mu^{-1} = x'\beta \rightarrow \mu = 1/x'\beta$

Ωστόσο, υπάρχουν περιπτώσεις λάθους π.χ. όταν έχουμε μη αρνητικές τιμές αποκρίσεων. Παρόλα αυτά, υπάρχει η πιθανότητα οι εκτιμήτριες b να δώσουν αρνητικές εκτιμήτριες αποκρίσεων έτσι η εκτιμήτρια του μ μπορεί να έχει και αρνητική τιμή που μπορεί να προκαλέσει πρόβλημα.

Λογαριθμική Συνάρτηση Σύνδεσης :

Συχνά αντί της Κανονικής σύνδεσης χρησιμοποιείται με αρκετή επιτυχία η Λογαριθμική η οποία δεν δημιουργεί αρνητικές εκτιμήτριες παραμέτρων και συνδέεται με το γραμμικό μοντέλο όπου η απόκριση είναι ln y.

Στην πρώτη περίπτωση έχουμε μετασχηματισμό των δεδομένων (επηρεάζει την κατανομή που ακολουθεί το σφάλμα) ενώ στη δεύτερη μετασχηματισμό του μέσου (δεν επηρεάζει την κατανομή που ακολουθεί το σφάλμα).

Στην περίπτωση όπου χρησιμοποιηθεί η Κανονική σύνδεση ο ασυμπτωτικός πίνακας διακύμανσης –συνδιακύμανσης είναι :

$$var(b) = (X'X)^{(-1)}\sigma^2$$

ενώ σε αυτή της μη Κανονικής σύνδεσης έχουμε :

$$var(b) = (X'\Delta V\Delta X)^{-1}[a(\varphi)]^2 \qquad (2.2)$$

Πιο συγκεκριμένα, η σχέση (2.2) με :

$$\ln \mu = x'\beta$$

$$\theta = -1/\mu = -\ e^{-x'\beta}$$

$$\Delta i = \frac{\partial \theta_i}{\partial (x_i'\beta)} = -e^{-x_i'\beta}$$

$$\text{var}(y) = \frac{\mu^2}{r} = \frac{e^{(2x_i'\beta)}}{r}$$

$$\Delta V \Delta = \text{diag}\ \{1/r,\ 1/r, \ldots, 1/r\}$$

δίνει $var(b) = (X'\Delta V \Delta X)^{-1}[a(\varphi)]^2 = (X'X)^{-1}(1/r)$.

2.3 Μεθοδολογία για Γενικευμένα Γραμμικά Μοντέλα

1. Οι ανεξάρτητες αποκρίσεις $y_1, y_2, ..., y_n$ έχουν μέσους $\mu_1, \mu_2, ..., \mu_n$ αντίστοιχα.
2. Οι παρατηρήσεις y_i ακολουθούν κατανομές που ανήκουν στην Εκθετική Οικογένεια.
3. Το μοντέλο περιέχει μεταβλητές παλινδρόμησης της μορφής $x_1, x_2, ..., x_k$.
4. Το μοντέλο κατασκευάζεται βάσει της γραμμικής παραμέτρου πρόβλεψης :

$$\eta = x'\beta = \beta_0 + \sum_{i=1}^{k} \beta_i x_i$$

5. Το μοντέλο κατασκευάζεται μέσω της συνάρτησης σύνδεσης :

$$\eta_i = g(\mu_i) \qquad \text{i} = 1,2,3,\ldots\ldots\ldots,\text{n}$$

η οποία δηλώνει ότι υπάρχει συσχέτιση μεταξύ του μέσου και της γραμμικής παραμέτρου πρόβλεψης με μέση τιμή απόκρισης :

$$\text{E}(y_i) = g^{-1}(\eta_i) = g^{-1}(x_i'\beta)$$

6. Η συνάρτηση σύνδεσης είναι μονότονη και διαφορίσιμη.
7. Η διακύμανση σ_i^2 (i = 1,2,3,…..,n) είναι συνάρτηση του μέσου μ_i.

Όσον αφορά τις συναρτήσεις σύνδεσης υπάρχει μεγάλη ποικιλία.

Αν επιλέξουμε όμως την $\eta_i = \theta_i$ τότε λέμε ότι η η_i είναι η **Κανονική Συνάρτηση Σύνδεσης.**

Στον πίνακα που ακολουθεί έχουμε τις κανονικές συσχετίσεις για τα GLM.

Κατανομές	***Κανονική Συνάρτηση Σύνδεσης***
Κανονική	$\eta_i = \mu_i$
Διωνυμική	η_i = ln (P/1-P)
Poisson	η_i = ln(μ_i)
Εκθετική	$\eta_i = 1/\mu_i$
Γάμμα	$\eta_i = 1/\mu_i$

2.4 Μέθοδος Μέγιστης Πιθανοφάνειας στα GLM

Η μέθοδος της Μέγιστης Πιθανοφάνειας χρησιμοποιείται στα GLM για την εκτίμηση των αγνώστων παραμέτρων β. Υποθέτοντας κανονική συσχέτιση η $\eta_i = g(\mu_i) = x_i'\beta$ στα GLM έχουμε :

$$L = \log l(y, \beta) = \sum_{i=1}^{n} \{[y_i\theta_i - b(\theta_i)]/\, a(\varphi) + c(y_i, \varphi)\}$$

$$\frac{\partial L}{\partial \beta} = \frac{\partial L}{\partial \theta_i}\frac{\partial \theta_i}{\partial \beta} = \sum_{i=1}^{n} \frac{1}{a(\varphi)}\{y_i - \frac{db(\theta_i)}{d\theta_i}\}x_i$$

$$= \sum_{i=1}^{n} \frac{1}{a(\varphi)}\{y_i - \mu_i\}x_i$$

Τώρα είναι εύκολο να υπολογίσουμε τις παραμέτρους β μηδενίζοντας την παραπάνω εξίσωση και αφού υποθέσουμε ότι α(φ) είναι σταθερό άρα παραλείπεται, έχουμε :

$$\sum_{i=1}^{n}\{y_i - \mu_i\}x_i = 0$$

ή με μορφή πινάκων $\mathbf{X}'(\mathbf{y}-\boldsymbol{\mu}) = \mathbf{0}$ όπου μ΄ = $[\mu_1, \mu_2 \ldots\ldots\ldots \mu_n]$. Οι εξισώσεις αυτές ονομάζονται **Εξισώσεις Αποτελέσματος Μέγιστης Πιθανοφάνειας.**

Υποθέτοντας ότι b η εκτιμήτρια του β ο πίνακας πληροφοριών των εκτιμητριών b είναι:

$$I(b) = var\{\frac{1}{a(\varphi)}\}[X'(y-\mu)] = \frac{(X'VX)}{[a(\varphi)]^2}$$

όπου : V = diag{ σ_i^2 } και σ_i^2 συνάρτηση του μέσου μ_i εξαρτάται από την κατανομή.

Ο ασυμπτωτικός πίνακας διακύμανσης- συνδιακύμανσης του b γενικά είναι :

$$\text{var}(b) = I^{(-1)}(b) = (X'VX)^{-1}[a(\varphi)]^2 \qquad (2.3)$$

Για την Κανονική κατανομή έχουμε $\sigma_i^2 = \sigma^2$,α(φ) = σ^2 η σχέση (2.3) μας δίνει :

$$\text{var}(b) = (X'X)^{-1}\sigma^2$$

ενώ για τις κατανομή Poisson όπου σ_i^2 = exp (x_i'β), α (φ) = 1.0 η σχέση (2.3) μας δίνει :

$$\text{var}(b) = (X'VX)^{-1}$$

Ωστόσο, πρέπει να επισημάνουμε ότι ενώ η κανονική συνάρτηση σύνδεσης είναι η πιο συνηθισμένη που χρησιμοποιείται αυτό δε σημαίνει ότι μια μη κανονική συνάρτηση σύνδεσης πρέπει να αποκλεισθεί αμέσως. Στην περίπτωση αυτή όπου

$\eta_i \neq \theta_i$ έχουμε :

$$\frac{\partial L}{\partial \beta} = \sum_{i=1}^{n} \frac{\partial L}{\partial \theta_i} \frac{\partial \theta_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta}$$

Λαμβάνοντας υπόψη ότι :

$$\frac{\partial L}{\partial \theta_t} = \sum_{i=1}^{n} \frac{1}{a(\varphi)} \{ y_i - \frac{db(\theta_i)}{d\theta_i} \} = \frac{1}{a(\varphi)} \sum_{i=1}^{n} (y_i - \mu_i)$$

$$\frac{\partial \eta_i}{\partial \beta} = x_i$$

έχουμε :

$$\frac{\partial L}{\partial \beta} = \sum_{i=1}^{n} \frac{1}{a(\varphi)} (y_i - \mu_i) \frac{\partial \theta_i}{\partial \eta_i} x_i$$

Μηδενίζοντας την παραπάνω συνάρτηση και υποθέτοντας ότι α(φ) σταθερό άρα παραλείπεται έχουμε με μορφή πινάκων:

$$\mathbf{X'\Delta\ (y - \mu)\ =\ 0}$$

όπου Δ = diag $\{ \frac{\partial \theta_i}{\partial \eta_i} \}$. Άρα ο πίνακας πληροφοριών και ο ασυμπτωτικός πίνακας διακύμανσης- συνδιακύμανσης με τις εκτιμήσεις b των αγνώστων παραμέτρων β είναι αντίστοιχα :

$$I(b) = \frac{X'\Delta V \Delta X}{[a(\varphi)]^2}$$

$$\text{var}(b) = (X'\Delta V \Delta X)^{-1} [a(\varphi)]^2$$

Γενικά, στα GLM παρατηρούμε τα εξής :

1. Τυπικά, οι αναλυτές δεδομένων όταν χρησιμοποιούν μετασχηματισμό επιλέγουν συχνότερα τη μέθοδο των Ελαχίστων Τετραγώνων να εφαρμόσουν στο μοντέλο.
2. Η διακύμανση της απόκρισης y δεν είναι σταθερή για το λόγο αυτό χρησιμοποιούμε τη μέθοδο των Σταθμισμένων Ελαχίστων Τετραγώνων για την εκτίμηση των παραμέτρων.
3. Αυτό σημαίνει ότι τα GLM πρέπει να υπερέχουν βασικών αναλύσεων που βασίζονται σε μετασχηματισμούς όταν ένα πρόβλημα εξακολουθεί να έχει σταθερή διακύμανση μετά το μετασχηματισμό.
4. Η απόκλιση του μοντέλου μπορεί να χρησιμοποιηθεί στον έλεγχο της καταλληλότητας του μοντέλου. Τα αποτελέσματα του ελέγχου του Wald μπορούν να χρησιμοποιηθούν στον έλεγχο της υπόθεσης αλλά και στην κατασκευή διαστημάτων εμπιστοσύνης για κάθε παράμετρο του μοντέλου ξεχωριστά.

2.5 Μέθοδος Πιθανοφάνειας Quasi

Για την εύρεση του διανύσματος παραμέτρων β χρησιμοποιήσαμε τη μέθοδο της Μέγιστης Πιθανοφάνειας με τις προϋποθέσεις ότι :

α) οι κατανομές των αποκρίσεων είναι μέλη της Εκθετικής Οικογένειας

β) οι αποκρίσεις είναι ανεξάρτητες.

Παρόλα αυτά υπάρχουν περιπτώσεις όπου :

- είτε οι αποκρίσεις y_i είναι ανεξάρτητες αλλά δεν ακολουθούν κατανομή της Εκθετικής Οικογένειας
- είτε οι αποκρίσεις y_i έχουν διακυμάνσεις που είναι συναρτήσεις του μέσου αλλά είναι εξαρτημένες.

Η τελευταία περίπτωση εμφανίζεται συχνά σε βιοϊατρικές μελέτες ή μελέτες βιομηχανικών διεργασιών.

Η Πιθανοφάνεια Quasi απορρέει από τα Σταθμισμένα Ελάχιστα Τετράγωνα ή από τα Γενικευμένα Ελάχιστα Τετράγωνα για την περίπτωση όπου οι αποκρίσεις

είναι εξαρτημένες. Ο Wedderburn (1974) ανέπτυξε τη θεωρία της Πιθανοφάνειας Quasi η οποία εκμεταλλεύεται το γεγονός ότι οι συναρτήσεις αποτελέσματος εμπεριέχουν τις κατανομές των αποκρίσεων τις πρώτες δυο στιγμές. Επίσης, το έργο του υποδεικνύει ότι οι πληροφορίες από τη χρήση Γενικευμένων Ελαχίστων Τετραγώνων είναι παρόμοιες με αυτές της Μέγιστης Πιθανοφάνειας.

Στη γενική περίπτωση θεωρούμε ότι ο πίνακας V είναι θετικά ορισμένος αλλά όχι απαραίτητα διαγώνιος. Η μέθοδος των Γενικευμένων Ελαχίστων Τετραγώνων εστιάζει στη συνάρτηση $(y-\mu)'V^{-1}(y-\mu)$ η οποία δίνει ως συνάρτηση αποτελέσματος την :

$$D'V^{-1}(y-\mu)=0 \qquad (2.4)$$

με D πίνακα των παραγώγων dμ / dβ. Ακόμη, αν οι αποκρίσεις είναι ανεξάρτητες αλλά όχι απαραίτητα μέλη της Εκθετικής Οικογένειας και V= $\{\sigma_i{}^2\}$ η εξίσωση (2.4) παίρνει τη μορφή :

$$\sum_{i=1}^{n}\frac{(y_i-\mu_i)}{\sigma_i^2}\frac{d\mu_i}{d\beta}=0$$

όπου $\sigma_i^2=\alpha(\varphi)\,\mathrm{var}(\mu_i)$ και επιλύεται ως προς β.

Στην ειδική περίπτωση όπου η συνάρτηση σύνδεσης περιέχει τη γραμμική παράμετρο πρόβλεψης x΄β η εξίσωση (2.4) παίρνει τη μορφή :

$$\sum_{i=1}^{n}\frac{(y_i-\mu_i)}{\sigma_i^2}\frac{d\mu_i}{d\eta_i}x_i=0$$

ή με μορφή πινάκων : **X΄ΓV (y- μ) = 0** όπου $\Gamma=diag\{\frac{\partial\mu_i}{\partial\eta_i}\}$.

2.6 Ομάδα Συναρτήσεων Σύνδεσης – Συνάρτηση Δύναμης

Υπάρχουν περιπτώσεις όπου η Κανονική σύνδεση μπορεί να μην είναι η καταλληλότερη π.χ. η Λογαριθμική συσχέτιση είναι η Κανονική σύνδεση για την Poisson κατανομή αλλά αρκετά ικανοποιητική και για την Γάμμα κατανομή. Η διαδικασία εύρεσης της καταλληλότερης σύνδεσης και του μοντέλου είναι αρκετά

χρονοβόρα γι' αυτό συχνά χρησιμοποιείται η ομάδα συναρτήσεων δύναμης όπου ο μετασχηματισμός γίνεται στο μέσο μ ως εξής :

$$\mu^{\lambda} = x'\beta \qquad \lambda \neq 0$$

$$\ln \mu = x'\beta \qquad \lambda = 0$$

Ωστόσο, για να υπάρχει συνέχεια στο λ = 0 η δεύτερη εξίσωση γράφεται ως :

$$x'\beta = \frac{\mu^{\lambda} - 1}{\lambda}$$

Μέσω της προσέγγισης Box- Cox μπορούμε να βρούμε το διάστημα εμπιστοσύνης του λ και να αποφασίσουμε τον κατάλληλο μετασχηματισμό για τα δεδομένα. Συνήθως, είμαστε πιο ευχαριστημένοι με τις "φυσικές " τιμές του λ π.χ. αν λ = 0 τότε επιλέγουμε την λογαριθμική σύνδεση.

Η προτεινόμενη από τον Pregibon (1980) επαναληπτική διαδικασία για την εύρεση του βέλτιστου λ βασίζεται στην ανάλυση κατά Taylor του μ^{λ} γύρω από μια αρχική τιμή του :

$$\mu^{\lambda} = \mu^{\lambda_0} + (\lambda - \lambda_0)\mu^{\lambda_0} \ln \mu \rightarrow \mu^{\lambda_0} = \mu^{\lambda} - (\lambda - \lambda_0)\mu^{\lambda_0} \ln \mu$$
$$\rightarrow \mu^{\lambda_0} = x'\beta - (\lambda - \lambda_0)\mu^{\lambda_0} \ln \mu$$

αφού $\mu^{\lambda} = x'\beta$.

2.7 Έλεγχοι Καταλληλότητας και Υπόλοιπα στα GLM

Οι στατιστικοί έλεγχοι ή έλεγχοι καταλληλότητας χρησιμοποιούνται για την επιλογή των στατιστικά σημαντικών μεταβλητών σε ένα μοντέλο γραμμικής παλινδρόμησης. Οι έλεγχοι αυτοί αφορούν τους συντελεστές β των μεταβλητών στη συνάρτηση παλινδρόμησης που προσαρμόζουμε στα δεδομένα του προβλήματος. Αν μια μεταβλητή δεν είναι στατιστικά σημαντική, τότε η συνεισφορά της στη συνάρτηση παλινδρόμησης είναι ελάχιστη ή μηδενική.

Ο έλεγχος του Wald είναι όπως και κάθε στατιστικός έλεγχος αποτελείται από τα ακόλουθα στοιχεία :

α) την μηδενική υπόθεση H_0 και την εναλλακτική υπόθεση H_1

β) το προκαθορισμένο επίπεδο σημαντικότητας

γ) τη στατιστική συνάρτηση ελέγχου, η οποία ακολουθεί μια συγκεκριμένη κατανομή υπό την H_0.

Με τον έλεγχο Wald έχουμε τις υποθέσεις :

$$\mathrm{H}_0 : \beta_j = 0$$
$$\mathrm{H}_1 : \beta_j \neq 0$$

με στατιστική συνάρτηση :

$$(\frac{b_j}{se(b_j)}) \sim \chi_1^2$$

Τα διαστήματα εμπιστοσύνης των αποκρίσεων και των παραμέτρων θα βρεθούν αργότερα. Η ελεγχοσυνάρτηση Deviance μετρά την απώλεια προσαρμογής όταν επιβάλουμε συνάρτηση σύνδεσης. Η ελεγχοσυνάρτηση Deviance δίνεται από τον τύπο :

$$\mathrm{D}(\beta) = -2\ln[\frac{L(\beta)}{L(\mu)}]$$

όπου :

L(β): πιθανοφάνεια μοντέλου

L(μ): πιθανοφάνεια "κορεσμένου" μοντέλου δηλ. μοντέλου που έχει τόσες παραμέτρους όσες και οι παρατηρήσεις.

Η ελεγχοσυνάρτηση Deviance για διάφορες κατανομές δίνεται από τον παρακάτω πίνακα :

Κατανομές	***Deviance***
Κανονική	$\sum_{i=1}^{n}(y-\mu)^2$
Διωνυμική	$2\sum_{i=1}^{n}[y\ln(\frac{y}{\mu})-(y-\mu)]$
Poisson	$2\sum_{i=1}^{n}[y\ln(\frac{y}{\mu})+(m-y)\ln(\frac{m-y}{m-\mu})]$
Γάμμα	$2\sum_{i=1}^{n}[-\ln(\frac{y}{\mu})+(\frac{y-\mu}{\mu})]$

Υπάρχουν τρία είδη υπολοίπων κατάλληλα για τα GLM :

α) τα συνηθισμένα υπόλοιπα $y_i-\mu_i$ τα οποία δεν αποτελούν την καλύτερη επιλογή αφού η διακύμανση των αποκρίσεων (var (y_i)) δεν είναι σταθερή.

β) τα υπόλοιπα Pearson $r_p=\frac{y_i-\mu_i}{\sqrt{\text{var}(y_i)}}$

γ) τα υπόλοιπα απόκλισης (Deviance) με τις εξής ιδιότητες:

i) $i=\text{sgn}(y_i-\mu_i)\sqrt{d_i}$

ii) $\sum_{i=1}^{n}{d_{i,r}}^2=D(\beta)$

Στην ερώτηση ποια υπόλοιπα είναι τα καταλληλότερα οι Pierce και Schafer (1986) προτείνουν υπόλοιπα απόκλισης για την κατασκευή διαγνωστικών σχεδιαγραμμάτων.

ΚΕΦΑΛΑΙΟ 3 :
ΜΟΝΤΕΛΑ ΠΑΛΙΝΔΡΟΜΗΣΗΣ ΣΤΑ GLM

3.1 Λογιστική και Poisson Παλινδρόμηση

Στα Γενικευμένα Γραμμικά Μοντέλα τα δυο πιο σημαντικά μοντέλα που συναντούμε είναι αυτά της Λογιστικής και της Poisson παλινδρόμησης τα οποία έχουν πολλές εφαρμογές σε βιολογικά, βιοϊατρικά, περιβαλλοντικά και βιομηχανικά προβλήματα.

Πιο συγκεκριμένα, η Λογιστική παλινδρόμηση έχει μεγάλη απήχηση στην καμπύλη απόκρισης-δόσης η οποία προσδιορίζει τις σχέσεις όπου αναπτύσσουν οι τοξικολόγοι και βιολόγοι που ενδιαφέρονται να μοντελοποιήσουν την απόκριση π.χ. ποσοστά αποβολής σε μια χημειοθεραπεία. Στην Λογιστική παλινδρόμηση, όπως και σε όλα τα GLM έχουμε συσχέτιση μεταξύ του μέσου και της διακύμανσης της απόκρισης.

Στην πρώτη περίπτωση μπορούμε να λάβουμε υπόψη μας το πείραμα όπου η απόκριση είναι μια Bernoulli τυχαία μεταβλητή και έχουμε :

$$E(y_i) = P_i = P(x_i) \qquad i = 1, 2, \ldots\ldots\ldots, n \qquad (3.1)$$

$$\text{var}(y_i) = P_i(1 - P_i)$$

όπου P_i = πιθανότητα σε μια Bernoulli διαδικασία είναι συνάρτηση των μεταβλητών παλινδρόμησης x_i άρα και η διακύμανση είναι συνάρτηση του μέσου.

Στη δεύτερη περίπτωση, μπορούμε να λάβουμε υπόψη μας το πείραμα όπου η απόκριση είναι ο αριθμός των ελαττωμάτων σε μια μικροηλεκτρονική συσκευή με :

$$E(y_i) = \mu(x_i) \quad i = 1, 2, \ldots\ldots\ldots, n \qquad (3.2)$$

$$\text{var}(y_i) = E(y_i) = \mu(x_i)$$

όπου μ = παράμετρος του μοντέλου Poisson και μ_i είναι συνάρτηση των μεταβλητών παλινδρόμησης x_i.

Από τα προηγούμενα συμπεραίνουμε ότι η Λογιστική παλινδρόμηση αναφέρεται σε γεγονότα των οποίων το αποτέλεσμα μπορεί να αναπαρασταθεί μέσω των ψηφίων 0 και 1 π.χ. ανταπόκριση ασθενούς σε φαρμακευτική αγωγή με παραμέτρους ηλικία, βάρος κ.λ.π ή δυνατότητα αποπληρωμής δόσεων πιστωτικών καρτών με οικονομικές, κοινωνικές παραμέτρους.

Αν υποθέσουμε ότι y=1 επιτυχία και y=0 αποτυχία τότε μοντελοποιούμε την μέση απόκριση P(x_i) όπου P(x_i) είναι η πιθανότητα επιτυχίας το λογιστικό μοντέλο γίνεται:

$$P(x_i) = \frac{1}{1+\exp(-x_i'\beta)} \tag{3.3}$$

όπου $x_i'\beta$ είναι η γραμμική παράμετρος πρόβλεψης και $0 \leq P(x_i) \leq 1$. Το μοντέλο αυτό χρησιμοποιείται περισσότερο από οποιοδήποτε άλλο της οικογένειας των GLM. Στην περίπτωση όπου έχουμε μια μεταβλητή η σχέση (3.3) παίρνει τη μορφή της σχέσης (3.1).

3.2 Μέγιστη Πιθανοφάνεια στη Λογιστική Παλινδρόμηση

Ας υποθέσουμε ότι έχουμε μια κλινική μελέτη στην οποία λαμβάνουν μέρος n_i αντικείμενα ή ζώα στα οποία χορηγείται μια δόση ενός φαρμάκου ή συνδυασμού φαρμάκων. Επίσης, υποθέτουμε ότι τα δεδομένα είναι ομαδοποιημένα. Η εξίσωση (3.3) γίνεται :

$$\mathrm{E}(y_i) = n_i P(x_i) = n_i \frac{1}{1+\exp(-x_i'\beta)}$$

$$\mathrm{var}(y_i) = n_i P(x_i)[1-P(x_i)]$$

όπου $y_1, y_{2,\ldots,} y_m$ οι παρατηρούμενες τιμές των ανεξάρτητων δυωνυμικών τυχαίων μεταβλητών με $\sum_{i=1}^{m} n_i = n$ το συνολικό μέγεθος. Έτσι η μέθοδος Μέγιστης Πιθανοφάνειας μας δίνει :

$$ln[L(P;y)] = \sum_{i=1}^{m} \{y_i ln[P(x_i)/1-P(x_i)] + n_i ln[1-P(x_i)]\} \tag{3.4}$$

με $ln[P(x_i)/1-P(x_i)] = \beta_0 + \sum_{\iota=1}^{n} x_{ij}\beta_j$ i = 1,2, …, m m ≥ k+1

$$\ln[L(\beta; y)] = \sum_{i=1}^{m}\sum_{i=1}^{k} y_i x_{ij}\beta_j - \sum_{i=1}^{m} n_i \ln(1+\exp\sum_{i=1}^{k} x_{ij}\beta_j) \quad (3.5)$$

ή με μορφή πινάκων $\ln[L(\beta; y)] = \beta' \mathrm{Xy} - \sum_{i=1}^{n} n_i \ln[1+\exp(x_i'\beta)]$ (3.6)

Διαφορίζοντας τη σχέση (3.6) έχουμε :

$$\frac{\partial \ln L(\beta_i; y)}{\partial \beta} = \mathrm{X}'\mathrm{y} - \sum_{i=1}^{m}[\frac{n_i}{1+\exp(x_i'\beta)}]\exp(x_i'\beta)x_i$$

όμως $\exp(x_i'\beta)/1+\exp(x_i'\beta) = 1/1+\exp(-x_i'\beta) = P(x_i)$ άρα :

$$\frac{\partial \ln L(\beta_i; y)}{\partial \beta} = \mathrm{X}'\mathrm{y} - \sum_{i=1}^{m} n_i P(x_i)x_i$$

ή με μορφή πινάκων : $\partial \mathbf{lnL}(\boldsymbol{\beta}_i; \mathbf{y}) \;/\; \partial\, \boldsymbol{\beta} = \mathbf{X}'(\mathbf{y} - \boldsymbol{\mu})$ όπου $\mu_i = n_i P(x_i)$.

Επομένως η συνάρτηση αποτελέσματος είναι : $\mathbf{X}'(\mathbf{y} - \boldsymbol{\mu}) = \mathbf{0}$ (3.7)

Η παραπάνω εξίσωση δεν είναι ασήμαντη αφού οι παράμετροι β εμφανίζονται στο μέσο μ όπου $\mu_i = n_i \frac{1}{1+\exp(-x_i'\beta)}$.

Όμως, η μη σταθερή διακύμανση της Λογιστικής παλινδρόμησης μας οδηγεί στη χρήση των Σταθμισμένων Ελαχίστων Τετραγώνων η οποία μας δίνει :

$$S = \sum_{i=1}^{m}[\frac{(y_i - \mu_i)^2}{\sigma_i^2}] \quad (3.8)$$

όπου :

$$\mu_i = n_i P(x_i)$$

$$\sigma_i^2 = n_i P(x_i)[1-P(x_i)] = \frac{n_i \exp(-x_i'\beta)}{[1+\exp(-x_i'\beta)]^2}$$

Διαφορίζοντας την σχέση (3.8) και θέτοντας τη ίση με μηδέν έχουμε :

$$2\sum_{i=1}^{m}\frac{(y_i-\mu_i)}{\sigma_i^2}\frac{\partial\mu_i}{\partial\beta}=0$$

όμως $\frac{\partial\mu_i}{\partial\beta}=n_iP(x_i)[1-P(x_i)]x_i=\sigma_i{}^2x_i$ άρα :

$$\sum_{i=1}^{m}(y_i-\mu_i)x_i=0$$

η οποία είναι όμοια με την εξίσωση (3.7) .

Στην περίπτωση όπου οι υπό εξέταση μονάδες είναι σχετικά ομογενείς π.χ. έρευνα ζώων η Λογιστική παλινδρόμηση να πάρει τη μορφή του μοντέλου απόκρισης δόσης με k = 1 και p = 2 :

$$P(x_i)=\frac{1}{1+\exp(-\beta_0-\beta_1x_i)} \qquad (3.9)$$

Ιδιαίτερη προσοχή δίνεται στον υπολογισμό των συντελεστών μεμονωμένα αφού συχνά υπάρχει η ανάγκη να αναφέρουμε τις περιττές αναλογίες .Για παράδειγμα, η ανεξάρτητη μεταβλητή x είναι **κατηγορηματική** και πιθανότατα μια ομάδα αντικειμένων χωρίζεται σε αυτά που πήραν μεγάλη δόση βιταμίνης C (x = 0) και σε αυτά χωρίς θεραπεία (x = 1) ενώ η απόκριση αντιπροσωπεύει τη λοίμωξη του αναπνευστικού συστήματος (y = 1) ή όχι (y = 0). Η ιδέα υπολογισμού των περιττών αναλογιών είναι αποτέλεσμα της χρήσης του logit (P / 1 -P) . Επίσης είναι εύκολο να δούμε ότι η σχέση (3.3) δίνει :

$$\ln[\,P(x_i)/1-P(x_i)\,]=x_i'\beta \qquad (3.10)$$

και έτσι ο μετασχηματισμός του P γραμμικοποιεί τη λογιστική συνάρτηση. Αν συνδυάσουμε τις σχέσεις (3.9) και (3.10) για την συσχέτιση βιταμίνης C με τη λοίμωξη του αναπνευστικού συστήματος έχουμε:

$$\ln[\,P(x_i)/1-P(x_i)\,]=\beta_0+\beta_1x_i$$

Για την περίπτωση μέγιστης δόσης βιταμίνης C (x = 0) η παράμετρος exp(β_0) μπορεί να ερμηνευθεί ως η αναλογία συχνότητας μολυσμένων προς μη μολυσμένων αντικειμένων του συνόλου υπό εξέταση. Αντίθετα, στην περίπτωση όπου x = 1 δηλ. έχουμε "αθεράπευτη ομάδα" ισχύει :

$$\ln [\ P(x_i) / 1 - P(x_i)\ ] = \beta_0 + \beta_1$$

Έτσι για την ομάδα με x = 1 έχουμε:

$$\ln[\Pr(Y=1 \ / x=1) / \Pr(Y=0 \ / x=1)\ ] = \ln[\Pr(Y=1 \ / x=0) / \Pr(Y=0 \ / x=0)\ ] + \beta_1$$

με το exp (β_1) να είναι η αναλογία "αθεράπευτης ομάδας" προς "θεραπευμένης". Είναι εμφανές ότι ο παρατηρητής ερμηνεύει ότι ένα $\beta_0 << 0$ είναι πιο ευνοϊκό για τη θεραπεία όπως και ένα $\beta_1 >> 0$.

Συγκεντρωτικά, θα μπορούσαμε να πούμε ότι οι συντελεστές είναι εξίσου σημαντικοί στον υπολογισμό πιθανοτήτων όσο οι έλεγχοι καταλληλότητας και τα διαστήματα εμπιστοσύνης. Παρόλα αυτά στις βιομηχανικές εφαρμογές η μέθοδος των συντελεστών δεν είναι η επικρατούσα ενώ θα περιμέναμε να έχουν πιο ευρεία χρήση αφού συχνά ο μεγάλος αριθμός μεταβλητών απαιτεί τη χρήση ελέγχων καταλληλότητας.

3.3 Ιδιότητες Διασποράς της Μέγιστης Πιθανοφάνειας

Είναι γνωστό ότι οι εκτιμήτριες της Μέγιστης Πιθανοφάνειας έχουν ιδιότητες ασυμπτωτικής διακύμανσης – συνδιακύμανσης οι οποίες δίνονται ως συνάρτηση του πίνακα πληροφορίας ο οποίος στην περίπτωση γραμμικού μοντέλου με σφάλμα που ακολουθεί την κανονική κατανομή είναι :

$$I(b) = \frac{(X'X)}{\sigma^2}$$

ενώ ο πίνακας διακύμανσης – συνδιακύμανσης είναι : $I^{(-1)}(b) = (X'X)^{-1}\sigma^2$.

Ο πίνακας πληροφορίας είναι μια ποσοτική ένδειξη της ποιότητας των πληροφοριών των εκτιμητριών . Ένας "μεγάλος" πίνακας πληροφορίας έχει σαν αποτέλεσμα τις μικρές διακυμάνσεις των εκτιμούμενων συντελεστών. Ο πίνακας

πληροφορίας προκύπτει από αρκετές διαφορετικές μεθόδους με ευκολότερη αυτή που περιέχει τη συνάρτηση αποτελέσματος:

$$I\left(b\right) = \text{var}\left(\text{score}\right) = \text{var}\left[X'(y-\mu)\right]$$

$$= X'\text{var}[(y-\mu)]X$$

$$= X'VX$$

όπου V= diag $\{\sigma_1^2,\sigma_2^2,\ldots,\sigma_n^2\}$ασυμπτωτικός πίνακας διακύμανσης – συνδιακύμανσης είναι :

$$\text{var}(b)=(X'VX)^{-1}$$

3.4 Έλεγχος Wald στη Λογιστική Παλινδρόμηση

Εκτός της μεθόδου της Μέγιστης Πιθανοφάνειας στα GLM χρησιμοποιείται και ο έλεγχος του Wald ο οποίος έχει ως θέμα τον έλεγχο των υποθέσεων των παραμέτρων στη λογιστική παλινδρόμηση δηλ. θέλουμε να ελέγξουμε αν :

$$H_0 : \beta_j = 0$$
$$H_1 : \beta_j \neq 0$$

Σε μια βιομηχανική μελέτη μας ενδιαφέρει να έχουμε τον έλεγχο των μεταβλητών σε αντίθεση σε μια βιοστατιστική μελέτη επικεντρωνόμαστε στην επίδραση ενός συντελεστή σε ένα πρόβλημα δόσης – απόκρισης. Για την εκτιμήτρια b_j της παραμέτρου β_j έχουμε την :

$$z_j = \frac{b_j - \beta_j}{\sigma b_j}$$

να είναι ασυμπτωτικά κανονική κατανομή N(0,1) ενώ $z_j^2 = (\frac{b_j}{\sigma b_j})^2$ ακολουθεί την κατανομή χ_1^2 υπό την υπόθεση με H_0: σ b_j να είναι το κατάλληλο διαγώνιο στοιχείο του πίνακα διακύμανσης- συνδιακύμανσης του b.

Μια επιπλέον χρήση του ελέγχου του Wald είναι ο υπολογισμός των διαστημάτων εμπιστοσύνης μιας δυωνυμικής πιθανότητας σε συγκεκριμένα ή

αυθαίρετα δεδομένα θέσης το οποίο αντιστοιχεί σε διαστήματα εμπιστοσύνης της μέσης απόκρισης.

Για τη Λογιστική παλινδρόμηση γνωρίζουμε ότι η μέση απόκριση σε ένα σημείο x = x_i δίνεται από τον τύπο:

$$P = \frac{1}{1+\exp(-x'\beta)}$$

Σε μια βιομηχανική εφαρμογή ο μηχανικός μπορεί να ζητήσει ένα 95% διάστημα εμπιστοσύνης για τη πιθανότητα εμφάνισης ελαττώματος στο σημείο x = x_i όπου η εκτίμηση θα δίνεται από y_i = P(x_i) . Μια μέθοδος που χρησιμοποιείται συχνά σε τέτοιες περιπτώσεις είναι η Δέλτα αλλά υπάρχει και μια εναλλακτική μέθοδος που βασίζεται στην ύπαρξη της γραμμικής παραμέτρου πρόβλεψης στο λογιστικό μοντέλο:

$$P = \frac{1}{1+\exp(-x'\beta)}$$

η οποία είναι μονότονη συνάρτηση ως προς το x΄β βάσει του οποίου μπορούμε να προσαρμόσουμε ένα 100 (1-α)% διάστημα εμπιστοσύνης του P. Έτσι, γνωρίζοντας την κατανομή που ακολουθεί ο παράγοντας x΄b ~ $N[x'\beta, x'(X'VX)^{-1}x]$ έχουμε :

$$x'b \pm z_{(a/2)}\sqrt{x'(X'VX)^{-1}x}$$

Σε ένα βιολογικό ή χημικό παράδειγμα όπου τα δεδομένα είναι ομαδοποιημένα η απόκριση y_i είναι δυωνυμική με παραμέτρους P(x_i) και n_i θέλουμε να υπολογίσουμε διάστημα εμπιστοσύνης για την y_i. Με P = P(x_i) από τη μέθοδο Δέλτα έχουμε:

$$\text{var}[P(x_i)] = \left(\frac{\partial P(x_i)}{\partial b}\right)'(X'VX)^{-1}\left(\frac{\partial P(x_i)}{\partial b}\right)$$

όμως $\frac{\partial P(x_i)}{\partial b} = n_i P(x_i)[1-P(x_i)]x_i$ ή $\frac{\partial \mu_i}{\partial b} = [\text{var}\,(y_i)]\,x_i$ άρα :

$$\text{var}[P(x_i)] = [\text{var}\,(y_i)]^2 x_i'(X'VX)^{-1}x_i$$

Έτσι, το 100 (1-α)% διάστημα εμπιστοσύνης δίνεται ως εξής :

$$P(x_i) \pm z_{(a/2)}\{n_i P(x_i)[1-P(x_i)]\}\sqrt{1+x_i'(X'VX)^{-1}x_i} \qquad \text{i = 1,2,……,m}$$

3.5 Καταλληλότητα Μοντέλου (Απόκλιση (Deviance))

Πολλά υπολογιστικά πακέτα που υποστηρίζουν τη Λογιστική παλινδρόμηση έχουν τουλάχιστον ένα τεστ καταλληλότητας όπου οι αποκρίσεις είναι δυαδικές ή δυωνυμικές. Ένα από αυτά τα τεστ χρησιμοποιεί το κριτήριο αναλογιών Πιθανοφάνειας για να οριστεί η απόκλιση η οποία μας δίνει τη δυνατότητα να αποφασίσουμε αν το μοντέλο που προσαρμόζουμε είναι σημαντικά χειρότερο από το κορεσμένο μοντέλο:

$$\text{E}\,(y_i) = P_i \qquad \text{i = 1,2,……,m}$$

όπου $P_i \neq P(x_i)$ και το κορεσμένο μοντέλο απαιτεί τον υπολογισμό m ανεξάρτητων παραμέτρους παλινδρόμησης. Στο κορεσμένο μοντέλο η απόκριση y_i είναι η εκτιμήτρια $\widehat{P}_i$ με αποτέλεσμα να μην υπάρχουν βαθμοί ελευθερίας των υπολοίπων μετά τον υπολογισμό. Σε ένα μη ομαδοποιημένο παράδειγμα η εκτίμηση θα είναι 0 ή 1. Το κορεσμένο μοντέλο θα επιδείξει πιθανοφάνεια όχι μικρότερη από αυτή του προσαρμοσμένου μοντέλου παλινδρόμησης το οποίο είναι ανάλογο με κορεσμένο μοντέλο με σφάλμα που ακολουθεί την Κανονική κατανομή και το άθροισμα τετραγώνων του σφάλματος είναι μηδέν.

Η απόκλιση για το προσαρμοσμένο ορίζεται ως εξής :

$$\text{D}(\beta) = -2\ln[\frac{L(\beta)}{L(\mu)}] \sim \chi^2_{m-p}$$

όπου:

L (β): η πιθανοφάνεια του προσαρμοσμένου μοντέλου όπου αντικαθιστούμε το β με την εκτιμήτρια του

L(P) : η πιθανοφάνεια του κορεσμένου μοντέλου.

Μια ασήμαντη τιμή της D(β) υποδηλώνει ότι η προσαρμογή του μοντέλου είναι σημαντικά χειρότερη από αυτή του κορεσμένου μοντέλου έτσι μια σχετικά μικρή τιμή της D(β) είναι ευνοϊκή για το προσαρμοσμένο μοντέλο.

3.6 Probit , Complementary Log –Log Μοντέλα για Δυαδικές Αποκρίσεις

Θα μελετήσουμε άλλα δυο μοντέλα παλινδρόμησης εκτός του Λογιστικού που χρησιμοποιούνται για δυαδικές αποκρίσεις το μοντέλο probit και το συμπληρωματικό μοντέλο log-log. Για να καταλάβουμε τα μοντέλα αυτά πρέπει να δούμε από που παράγεται το Λογιστικό μοντέλο παλινδρόμησης το οποίο βασίζεται στην ανοχή (tolerance) της κατανομής. Η καλύτερη περίπτωση είναι όταν έχουμε μια ανεξάρτητη μεταβλητή π.χ. η δόση z ενός συστατικού σε έντομα όπου τα έντομα παρουσιάζουν διαφορετικές ανοχές στη δόση. Κάποια από αυτά πεθαίνουν με μικρή δόση του συστατικού ενώ άλλα με μεγαλύτερη.

Αν η ανοχή ακολουθεί τη Λογιστική κατανομή τότε η συνάρτηση πυκνότητας πιθανότητας είναι :

$$f(z) = \frac{\exp[(z-\mu)/\tau]}{c[1+\exp\{(z-\mu)/\tau\}]^2}$$

με μέσο μ και διακύμανση $\pi^2\ \tau^2/\ 3$. Η πιθανότητα θανάτου είναι η συνάρτηση κατανομής:

$$P = \int_{-\infty}^{x} f(z)dz = \frac{\exp[(x-\mu)/\tau]}{1+\exp\{(x-\mu)/\tau\}}$$

η οποία ισούται με (Collett, 1991) :

$$P = \frac{\exp[\beta_0+\beta_1 x]}{1+\exp[\beta_0+\beta_1 x]} = \frac{1}{1+\exp[-(\beta_0+\beta_1 x)]}$$

όπου $\beta_0 = -\mu\ /\tau, \beta_1 = 1/\tau$ δηλ. υπάρχει συσχέτιση των παραμέτρων με την ανοχή της κατανομής. Το παραπάνω μοντέλο μπορεί να γραφτεί σε γραμμική μορφή ως :

$$\ln(\mathrm{P}/\ 1-\mathrm{P})\ = \beta_0+\beta_1 x$$

ή αν έχουμε πολλές μεταβλητές παλινδρόμησης:

$$\ln(\mathrm{P}/\ 1-\mathrm{P}) = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k$$

δηλ. $\mathbf{logit(P)} = \mathbf{x}'\boldsymbol{\beta}$.

Το δεύτερο μοντέλο είναι πιο κοντά στον αναγνώστη αφού η Κανονική κατανομή του είναι πιο οικεία . Το μοντέλο probit παράγεται από την υπόθεση ότι η ανοχή ακολουθεί την Κανονική κατανομή με μέση τιμή μ και διακύμανση σ^2. Η πιθανότητα P ενός γεγονότος με μια μεταβλητή παλινδρόμησης είναι :

$$\mathrm{P} = \Phi\ (\beta_0 + \beta_1 x)$$

με β_0= μ /σ και β_1 = 1/ σ. Έτσι, μπορούμε να πούμε ότι:

$$\text{Probit}\ (\mathrm{P}) = \Phi^{-1}(\mathrm{P}) = \beta_0 + \beta_1 x$$

Ωστόσο, πρέπει να αναφερθεί ότι στην περίπτωση μιας μεταβλητής η χρήση της log x παράγει πιο επιθυμητά αποτελέσματα αφού οι λογάριθμοι της ανοχής έχουν πιο συμμετρικές κατανομές.

Η μεγαλύτερη εφαρμογή είναι στα μοντέλα επιβίωσης όπου η γραμμικοποιημένη μορφή του μοντέλου είναι :

$$\log[\ -\log(1\ -\mathrm{P})] = \mathrm{x}'\beta$$

Εκτός όμως από τη μέθοδο της απόκλισης που συγκρίνει το προσαρμοσμένο μοντέλο με το κορεσμένο στα GLM χρησιμοποιείται η μέθοδος Pearson χ^2 :

$$\chi^2 = \sum_{i=1}^{m} \frac{(y_i - n_i P_i)^2}{n_i P_i (1 - P_i)}$$

η οποία ακολουθεί την κατανομή χ^2_{m-p}. Μάλιστα σε αρκετές περιπτώσεις και οι δυο μέθοδοι δίνουν παρόμοια αποτελέσματα και σπάνια διαψεύδουν η μια την άλλη . Ωστόσο, το 1989 οι Hosmer και Lemeshaw ανέπτυξαν ένα εναλλακτικό μοντέλο.

3.7 Υπερ-διασπορά στο Λογιστικό Μοντέλο Παλινδρόμησης

Υπάρχει η πιθανότητα το μοντέλο Λογιστικής παλινδρόμησης που υποθέτουμε να μην είναι το καταλληλότερο. Αυτό αποδίδεται στους εξής λόγους :

1. Η διωνυμική κατανομή της απόκρισης είναι λάθος.
2. Η επιλογή του μοντέλου logit είναι ακατάλληλη.
3. Η δομή που χρησιμοποιήθηκε στη γραμμική παράμετρο πρόβλεψης είναι λάθος π.χ. υπάρχουν επιπλέον παράγοντες που δεν έχουν συμπεριληφθεί.
4. Υπάρχουν ακραίες τιμές στα δεδομένα.

Ακόμη κι αν υποθέσουμε ότι η κατανομή και το μοντέλο είναι κατάλληλα και δεν υπάρχουν ακραίες τιμές στα δεδομένα μπορεί να υπάρξει πρόβλημα με τη μέση απόκλιση το οποίο ονομάζεται **υπερ-διασπορά.** Αυτό συμβαίνει όταν η διακύμανση της Δυωνυμικής κατανομής nP(1-P) δεν είναι αρκετή ή όταν σ^2 είναι μεγαλύτερο της μονάδας και η nP(1-P) γίνεται nP(1-P)σ^2. Στην περίπτωση όπου έχουμε σ^2 μικρότερο της μονάδας λέμε ότι **υπο-διασπορά.**

Οι αναλυτές ωστόσο δεν πρέπει να συμπεραίνουν ότι υπάρχει υπερ-διασπορά αν δεν προσπαθήσουν πάρα πολύ να βρουν το κατάλληλο μοντέλο. Ένα λάθος υποτιθέμενο μοντέλο έχει τα ίδια αποτελέσματα με την υπερ-διασπορά η οποία μπορεί να οφείλεται σε μεγάλη μέση απόκλιση ή ύπαρξη ακραίων τιμών στα δεδομένα. Αξίζει λοιπόν να μελετήσουμε τους παράγοντες που προκαλούν υπερ-διασπορά και πως επιδρούν στη Λογιστική παλινδρόμηση.

Κυριότερη αιτία της υπερ-διασποράς είναι η ανομοιογένεια των πειραματικών μονάδων. Στην περίπτωση γραμμικής παλινδρόμησης η ανομοιογένεια αυτή οδηγεί σε λάθος F-τέστ ή σε ακατάλληλη εκτίμηση της διακύμανσης των υπολοίπων. Στη λογιστική παλινδρόμηση η ανομοιογένεια μπορεί να οδηγήσει σε περιπτώσεις όπου ενώ οι μονάδες εκτίθενται στο ίδιο πείραμα έχουν διαφορετικές δυωνυμικές πιθανότητες.

Ένα παράδειγμα υπερ-διασποράς έχουμε αν υποθέσουμε μεταβλητότητα της δυωνυμικής παραμέτρου p η οποία σε σταθερές πειραματικές συνθήκες έχει μέση τιμή μ και διακύμανση φ > 0.

Υποθέτοντας Y τυχαία δυωνυμική μεταβλητή έχουμε:

$$E(Y) = E[E(Y/p)] = nE(p) = p$$

$$var(Y) = var(E(Y/p)) + E[var(Y/p)]$$

με $\text{var}(\text{E}(\text{Y}/\text{p})) = \text{var}[np] = n^2\varphi$,

$$\text{E}[\text{var}(\text{Y}/\text{p})] = n\text{E}[\text{p(1-p)}] = n[\mu - (\varphi - \mu)^2]$$

η παραπάνω σχέση γίνεται :

$$var(Y) = n^2\varphi + n\mu - n\varphi - n\,\mu^2 = n\mu(1-\mu) + n\varphi(n-1) > n\mu(1-\mu)$$

Στην υπερ-διασπορά της λογιστικής παλινδρόμησης η παράμετρος $\sigma^2 > 1$ εισέρχεται στον πίνακα διακύμανσης – συνδιακύμανσης : $var(b) = (\text{X}'VX)^{-1}\,\sigma^2$

Τέλος, το φαινόμενο της υπερ-διασποράς παρατηρείται σε βιολογικές ή βιοϊατρικές εφαρμογές όπου οι πειραματικές μονάδες είναι ζώα.

3.8 Poisson Παλινδρόμηση: Μέγιστη Πιθανοφάνεια και Απόκλιση

Εκτός της Λογιστικής παλινδρόμησης έχουμε και την Poisson όπου οι αποκρίσεις είναι Poisson μετρήσεις που ακολουθούν ανεξάρτητες Poisson κατανομές με $\text{E}(\text{y}_\text{i}) = \mu_\text{i}$ και $\text{var}(\text{y}_\text{i}) = \mu_\text{i}$. Μια σειρά από μεταβλητές $y_1, ..., y_n$ επηρεάζουν το μ μέσω του μοντέλου :

$$\mu_i = \exp(x_i'\beta) \qquad \text{i} = 1, 2, 3, \ldots\ldots\ldots,\text{n} \qquad (3.11)$$

με $x_i' = [1, x_{i1}, ..., x_{ik}]$. Ωστόσο, πρέπει να αναφέρουμε ότι όπως διάφορα μοντέλα αντί της Λογιστικής είναι κατάλληλα για δυωνυμικές αποκρίσεις το ίδιο ισχύει και για την Poisson παλινδρόμηση.

Η μέθοδος της Μεγίστης Πιθανοφάνειας για τις αποκρίσεις $y_1, ..., y_n$ είναι η εξής:

$$\ln \text{L}(\beta;\ \text{y}) = \log\prod_{i=1}^{n}\left[\frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!}\right]$$

$$= \sum_{i=1}^{n}[-\exp(x_i'\beta) + y_i x_i'\beta - \ln y_i!]$$

Παραγωγίζοντας την παραπάνω σχέση και θέτοντας τη ίση με μηδέν έχουμε:

$$\frac{\partial \ln L(\beta; y)}{\partial \beta} = 0$$

$$\sum_{i=1}^{n}[y_i x_i - \exp(x_i'\beta)x_i] = 0$$

ή όπως αλλιώς μπορεί να γραφτεί η παραπάνω εξίσωση : $\sum_{i=1}^{n}(y_i - \mu_i)x_i = 0$.

Η σχέση αυτή με μορφή πινάκων παίρνει τη μορφή $X'(y - \mu) = 0$ η οποία είναι ίδια με αυτή που προέρχεται από τη Μέγιστη Πιθανοφάνεια της Λογιστικής παλινδρόμησης.

Το ίδιο ισχύει και για τη σταθμισμένη Poisson παλινδρόμηση η οποία συσχετίζεται με αυτή της Λογιστικής. Για το μοντέλο της σχέσης (3.11) το τετράγωνο του αθροίσματος των σταθμισμένων υπολοίπων είναι :

$$S = \sum_{i=1}^{m}[\frac{(y_i - \mu_i)^2}{\sigma_i^2}]$$

ενώ η ελαχιστοποίηση του παραπάνω μας δίνει :

$$\partial S / \partial \beta = 0$$

$$2\sum_{i=1}^{n}[y_i - \exp(x_i'\beta)]x_i = 0$$

$$\rightarrow X'(y - \mu) = 0$$

Υπάρχει δηλ. μια ενδιαφέρουσα ομοιότητα μεταξύ Λογιστικής και Poisson παλινδρόμησης η οποία οφείλεται στην ύπαρξη της σχέσης $\partial \mu_i / \partial \beta = x_i$ var (y_i).

Επίσης, όπως υπάρχει η υπερ-διασπορά στο Λογιστικό μοντέλο έτσι γίνεται και στο Poisson και μάλιστα οφείλεται στους ίδιους παράγοντες.

Οι εφαρμογές της Poisson παλινδρόμησης συμπεριλαμβάνουν προβλήματα όπου η απόκριση είναι θετική τιμή : 0,1,2,3....Οι ακέραιες τιμές περιγράφουν την επίδοση ενός βιολογικού ή κοινωνικού συστήματος ή την παραγωγική διαδικασία μιας βιομηχανικής εφαρμογής.

Στην περίπτωση του μοντέλου αυτού η μέθοδος της Μεγίστης Πιθανοφάνειας δίνει ως εκτίμηση του μ_i στο κορεσμένο μοντέλο το y_i επειδή στο κορεσμένο μοντέλο υποθέτουμε ότι οι n παρατηρήσεις είναι ανεξάρτητες και ανεπηρέαστες από τις μεταβλητές παλινδρόμησης x_i. Έτσι στο κορεσμένο μοντέλο έχουμε :

$$\ln L(\mu) = -\sum_{i=1}^{n} y_i + \sum_{i=1}^{n} y_i \ln y_i - \sum_{i=1}^{n} y_i !$$

Για το Poisson μοντέλο έχουμε :

$$\ln L(\beta) = -\sum_{i=1}^{n} \mu_i + \sum_{i=1}^{n} y_i \ln \mu_i - \sum_{i=1}^{n} y_i !$$

όπου $\widehat{\mu_i}$ προέρχεται από τη Μέγιστη Πιθανοφάνεια . Η απόκλιση είναι :

$$\mathrm{D}(\beta) = -2\ln[\mathrm{L}(\beta) / \mathrm{L}(\lambda)]$$

$$= 2[-\sum_{i=1}^{n}(y_i - \mu_i) + \sum_{i=1}^{n} y_i(\ln y_i - \ln \mu_i)]$$

$$= 2[-\sum_{i=1}^{n}(y_i - \mu_i) + \sum_{i=1}^{n} y_i[\ln(y_i / \mu_i)]$$

η οποία αν y_i είναι κοντά στο $\widehat{\mu_i}$ πλησιάζει το μηδέν .Επίσης, αν θέλουμε να απλοποιήσουμε τη σχέση της απόκλισης πρέπει να λάβουμε υπόψη τη σχέση αποτελέσματος X΄(y - μ) = 0, η οποία μας δίνει ότι $\sum_{i=1}^{n}(y_i - \mu_i) = 0$ και η τελική μορφή της απόκλισης είναι:

$$D(\beta) = 2\sum_{i=1}^{n} y_i \ln \frac{y_i}{\mu_i}$$

ΚΕΦΑΛΑΙΟ 4:
ΕΦΑΡΜΟΓΕΣ ΤΩΝ ΓΕΝΙΚΕΥΜΕΝΩΝ ΓΡΑΜΜΙΚΩΝ ΜΟΝΤΕΛΩΝ

4.1 Παραγοντικοί Σχεδιασμοί και Βελτιστοποίηση

Οι παραγοντικοί σχεδιασμοί χρησιμοποιούνται ευρύτατα σε πειράματα που περιλαμβάνουν αρκετούς παράγοντες όπου είναι αναγκαία η μελέτη της κοινής επίδρασης των παραγόντων στην απόκριση. Για το λόγο αυτό, είναι χρήσιμοι σε βιομηχανικές εφαρμογές οι οποίες περιλαμβάνουν διαδικασίες ανάπτυξης, βελτίωσης και ελέγχου.

Για την ανάλυση αυτών των σχεδιασμών υποθέτουμε τα εξής:

- οι παράγοντες είναι σταθεροί
- οι σχεδιασμοί είναι πλήρως τυχαιοποιημένοι
- ισχύουν οι συνήθεις υποθέσεις κανονικότητας.

Ο σχεδιασμός 2^k είναι ιδιαίτερα χρήσιμος αφού μας παρέχει τον μικρότερο αριθμό εκτελέσεων με τις οποίες οι k παράγοντες πρέπει να μελετηθούν σε έναν πλήρη παραγοντικό σχεδιασμό. Πιο συγκεκριμένα, όσο περισσότερο αυξάνει ο αριθμός παραγόντων σε ένα 2^k σχεδιασμό τόσες περισσότερες εκτελέσεις απαιτούνται για μια πλήρη επανάληψη του σχεδιασμού.

Αν ο πειραματιστής μπορεί να υποθέσει ότι ορισμένες αλληλεπιδράσεις είναι αμελητέες τότε μπορούμε να αντλήσουμε πληροφορίες για τις κύριες επιδράσεις και αλληλεπιδράσεις εκτελώντας ένα κλάσμα μόνο του πλήρους παραγοντικού πειράματος. Έχουμε δηλαδή **κλασματικούς παραγοντικούς σχεδιασμούς** που έχουν ευρεία χρήση σε πειράματα κρησαρίσματος (screening experiments), στις επιδράσεις και προσαρμογές του εμπειρικού μοντέλου για την πρόβλεψη της απόκρισης και τη βελτιστοποίηση της διαδικασίας.

Η επιτυχία των κλασματικών παραγοντικών σχεδιασμών οφείλεται σε τρεις παράγοντες:

- η αρχή της σποραδικότητας των επιδράσεων . Η ύπαρξη αρκετών μεταβλητών έχει σαν αποτέλεσμα η διαδικασία να οδηγείται αρχικά από μερικές από τις κύριες επιδράσεις και τις αλληλεπιδράσεις.
- η προβολική ιδιότητα. Οι κλασματικοί παραγοντικοί σχεδιασμοί μπορούν αν προβάλλονται σε μεγαλύτερους σχεδιασμούς με αντικείμενο τους σημαντικούς παράγοντες.
- ακολουθιακός πειραματισμός. Είναι δυνατόν να συνδυάσουμε τις εκτελέσεις δύο ή περισσοτέρων κλασματικών παραγοντικών σχεδιασμών για να συγκεντρώσουμε ακολουθιακά έναν μεγαλύτερο σχεδιασμό για να εκτιμήσουμε τις επιδράσεις και αλληλεπιδράσεις των παραγόντων που ενδιαφέρουν.

Μεγάλη προσοχή ωστόσο πρέπει να δοθεί στη διαδικασία εύρεσης του βέλτιστου σχεδιασμού κατά τη χρήση των GLM. Είναι εμφανές ότι αν μπορούμε να ελέγξουμε τις μεταβλητές παλινδρόμησης τότε επηρεάζονται και οι λοιπές μεταβλητές. Για παράδειγμα, σε ένα πρόβλημα δόσης-απόκρισης ελέγχοντας τον αριθμό των δόσεων αλλά και τις θέσεις που γίνονται αυτές επηρεάζονται τα αποτελέσματα κυρίως δηλ. οι εκτιμήτριες των άγνωστων συντελεστών β.

Σε ένα 2^k κλασματικό παραγοντικό σχεδιασμό σημαντικό ρόλο παίζει η ορθογωνιότητα (Box and Hunter 1961, Montgomery 2001, Myers and Montgomery 1995). Έτσι, θεωρώντας ένα μοντέλο παλινδρόμησης της μορφής :

$$\mathrm{y} = \mathrm{X}\beta + \varepsilon$$

περιμένουμε δηλ. τις στήλες του πίνακα X να είναι ορθογώνιες μεταξύ τους. Για παράδειγμα, σε ένα 2^2 σχεδιασμό με συντελεστές x_1, x_2, x_1x_2 ο πίνακας X είναι:

$$\mathrm{X} = \begin{pmatrix} & x_1 & x_2 & x_1x_2 \\ 1 & -1 & -1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & 1 & 1 & 1 \end{pmatrix}$$

με X΄X = 4 I_4 λόγω ορθογωνιότητας και σφάλμα σ / 2 όπου σ: απόκλιση. Αυτό συμβαίνει σε κάθε ορθογώνιο σχεδιασμό με τιμές ± 1 που συμπεριλαμβάνουν τις κύριες επιδράσεις x_i και τις αλληλεπιδράσεις $x_i\ x_j$.(Myers and Montgomery 1995).

Το μοντέλο αυτό είναι γνωστό σαν **βέλτιστο μοντέλο διακύμανσης** το οποίο βασίζεται στην ορθογωνιότητα και τις τιμές ± 1 του πίνακα X. Ένα μοντέλο τέτοιου είδους είναι και το D-βέλτιστο μοντέλο το οποίο βασίζεται στην ελαχιστοποίηση της γενικευμένης διακύμανσης των συντελεστών. Έτσι, το μοντέλο δίνεται ως μεγιστοποίηση της :

$$D = \left|\frac{X'X}{n}\right|$$

Ο πίνακας $\frac{X'X}{n}$ ονομάζεται **πίνακας ροπών** ενώ το D-βέλτιστο μοντέλο έχει ευρεία χρήση στη βιομηχανία σαν κριτήριο επιλογής. Επίσης, κάθε κριτήριο το οποίο έχει ως βάση τη διακύμανση των συντελεστών συμπεριλαμβάνει τον πίνακα πληροφορίας και για τα γραμμικά μοντέλα με σφάλμα που ακολουθεί την κατανομή N(0,σ) ο πίνακας αυτός είναι ανάλογος με τον X΄X.

Η λογική της μεγιστοποίησης της ορίζουσας D βασίζεται στο γεγονός ότι σε ένα μοντέλο με γραμμικές παραμέτρους ο πίνακας X΄X δεν έχει παραμέτρους με αποτέλεσμα να απαιτείται μόνο η γνώση του μοντέλου για την εύρεση βέλτιστου μοντέλου. Ακόμη, και αν το μοντέλο περιέχει όρους πέραν των x_i, x_ix_j με i≠j είναι εύκολο να βρούμε ένα βέλτιστο μοντέλο.

Υπάρχει μια επιπλέον εξήγηση για την ιδιότητα της ορθογωνιότητας η οποία βασίζεται στην πολυσυγγραμικότητα. Η έννοια αυτή σχετίζεται με τη μελέτη εξάρτησης των μεταβλητών παλινδρόμησης x_i στην οποία συμμετέχουν οι ιδιοτιμές του πίνακα συσχέτισης καθώς και η διάφοροι παράγοντες πληθωρισμού.

Αν το μοντέλο είναι ορθογώνιο τότε δεν υπάρχουν γραμμικές εξαρτήσεις ανάμεσα στις μεταβλητές του μοντέλου. Πράγματι, ο πίνακας συσχέτισης είναι ταυτοτικός πίνακας με όλες τις ιδιοτιμές και τους διάφορους παράγοντες πληθωρισμού.

4.2 Δυσκολίες στην Εύρεση Βέλτιστομυ Μοντέλου στα GLM

Όπως στην περίπτωση των μοντέλων με κανονικό σφάλμα είναι λογικό τα βέλτιστα μοντέλα να απορρέουν από τον πίνακα πληροφορίας. Γενικά στα GLM ο πίνακας πληροφορίας είναι X΄ΔVΔX = X΄WX όπου W: Εσσιανός πίνακας ενώ στην περίπτωση της ταυτοτικής συνάρτησης σύνδεσης και των κανονικών σφαλμάτων ο πίνακας πληροφορίας είναι X΄X.

Επίσης, για άλλα GLM τα διαγώνια στοιχεία του πίνακα V είναι συναρτήσεις του μέσου και έτσι του x΄β ενώ τα διαγώνια στοιχεία του Δ είναι συναρτήσεις του x΄β με εξαίρεση την περίπτωση της κανονικής σύνδεσης όπου Δ = I. Τα παραπάνω έχουν σαν αποτέλεσμα να καθιστούν δύσκολη την εύρεση του βέλτιστου μοντέλου αν δεν γνωρίζουμε τις παραμέτρους.

Παρά τις δυσκολίες που συναντώνται στην εφαρμογή βέλτιστων μοντέλων στα GLM υπάρχει ιδιαίτερο ενδιαφέρον για τη Λογιστική παλινδρόμηση σε μοντέλο με μια μεταβλητή (Abdelbasit and Plackett 1983, Sitter 1992, Heise and Myers 1996, Myers and Carter 1994, Minkin 1987, Kalish and Rosenberger 1978) .

Για μοντέλα με δυο παραμέτρους υπάρχουν παραδείγματα τόσο στην Poisson όσο και την Λογιστική παλινδρόμηση (Branden, Vidmar and McKean 1988, Jia and Myers 2001, Van Mullekom and Myers 2001, Sitter and Torsney 1995). Όπως είναι φανερό η προϋπόθεση της γνώσης των παραμέτρων ή των εκτιμήσεων τους επιφέρει ακόμα περισσότερες δυσκολίες όταν κάποιος ασχολείται με μοντέλα με πάνω από μια μεταβλητή.

Εφαρμογές τέτοιου τύπου έχουμε όταν οι αποκρίσεις είναι Poisson τυχαίες μεταβλητές ενός συστήματος και οι μεταβλητές του μοντέλου αντιπροσωπεύουν δόσεις φαρμάκων ή ρύπων οι οποίες δυσκολεύουν την ανάπτυξη του συστήματος.

Συγκεντρωτικά, θα λέγαμε ότι η χρήση βέλτιστων σχεδιασμών πρέπει να αντικατασταθεί από διαδοχικούς σχεδιασμούς ή από Μπευζιανά μοντέλα. Οι δυο όμως αυτές προσεγγίσεις απαιτούν την ανάπτυξη νέων τεχνολογιών. Για το λόγο αυτό, στρεφόμαστε σε πιο καθιερωμένους σχεδιασμούς με ομοιόμορφη διακύμανση όπως τους 2^k κλασματικούς παραγοντικούς σχεδιασμούς.

Οι σχεδιασμοί αυτοί αναπτύχθηκαν για κανονικές συνθήκες σε αντίθεση με τους μη γραμμικούς συνδυασμούς στα GLM και την ταυτόχρονη επιπλοκή της μη ομογενούς διακύμανσης.

Όπως γνωρίζουμε, η ιδιότητα της ορθογωνιότητας παράγει σε κανονικές συνθήκες μοντέλα με τη μικρότερη διακύμανση αλλά απορρέει από τον πίνακα πληροφορίας :

$$I(b) = X'WX \tag{4.1}$$

με $W=\Delta V\Delta$ ένα διαγώνιο πίνακα με στάθμες $w_1, w_2, \ldots\ldots, w_n$ ο οποίος ονομάζεται **σταθμισμένος Εσσιανός πίνακας.** Αν τώρα υποθέσουμε τον πίνακα Z n× p:

$$Z = W^{1/2}$$

με $W^{1/2}$ διαγώνιο πίνακα με i-οστό διαγώνιο στοιχείο $w_i^{1/2}$ η σχέση (4.1) γίνεται :

$$I(b) = Z'Z \tag{4.2}$$

δηλ. περιέχει παραμέτρους και δεν μπορεί να ελεγχθεί. Αν τώρα θεωρήσουμε την περίπτωση των 2^k σχεδιασμών με τιμές ± 1 τα διαγώνια στοιχεία είναι ίσα και μας δίνουν:

$$I(b)_{ii} = \sum_{j=1}^{N} w_j \qquad j = 1,2,3,\ldots\ldots p \tag{4.3}$$

με w_j τη j-οστή Εσσιανή στάθμη και j δηλ. το j-οστό στοιχείο του W.

Γεννιέται λοιπόν το ερώτημα αν οι 2^k κλασματικοί παραγοντικοί σχεδιασμοί μπορούν να κατασταθούν ορθογώνιοι για τα GLM. Η ορθογωνιότητα εδώ υποδηλώνει ότι οι στήλες του πίνακα Z είναι αμοιβαία ορθογώνιες και έχει σαν αποτέλεσμα την διαγωνοποίηση του πίνακα πληροφορίας αλλά και την ασυμπτωτική ανεξαρτησία των εκτιμητριών των συντελεστών του μοντέλου.

Η επιλογή του σχεδιασμού των GLM είναι σημαντική για την κατασκευή του κατάλληλου μοντέλου . Παρόλα αυτά, το σημαντικότερο πρόβλημα στην κατασκευή των GLM είναι η εξάρτηση των αγνώστων παραμέτρων του προσαρμοσμένου μοντέλου. Το πρόβλημα αυτό απασχολεί τους ερευνητές τα τελευταία 25 χρόνια μόνο

που έχουν βρεθεί μερικές λύσεις για μεμονωμένες περιστάσεις. Για το λόγο αυτό, παρέχονται διάφορες τεχνικές σχετικά με το πρόβλημα της εξάρτησης των παραμέτρων.

Στις περισσότερες περιπτώσεις η μοντελοποίηση της απόκρισης οφείλεται σε μεθόδους παλινδρόμησης μέσω των οποίων γίνεται η επιλογή των παραγόντων που επιδρούν στην απόκριση. Όλη η μεθοδολογία για την κατάλληλη επιλογή , προσαρμογή και αξιολόγηση του σχεδιασμού είναι η γνωστή ως **Μεθοδολογία Επιφάνειας Απόκρισης (Response Surface Methology RSM) .**

Η χρήση των RSM ξεκίνησε από τη χημική βιομηχανία για τον καθορισμό των βέλτιστων συνθηκών λειτουργίας αλλά επεκτάθηκε στις φυσικές και μηχανικές επιστήμες καθώς και στις βιολογικές, κλινικές και κοινωνικές μελέτες.

Οι περισσότερες μέθοδοι σχεδιασμών για RSM μοντέλα αναπτύχθηκαν γύρω από αγροτικά, βιομηχανικά, εργαστηριακά πειράματα. Σημειωτέον, για τους σχεδιασμούς αυτούς υποθέσαμε ότι οι αποκρίσεις είναι συνεχείς, συνήθως ακολουθούν την Κανονική κατανομή, με ασυσχέτιστα σφάλματα και ομογενείς διακυμάνσεις.

Ωστόσο, υπάρχουν και περιπτώσεις όπου δεν ικανοποιούνται οι υποθέσεις αυτές π.χ. κλινικές ή βιολογικές μελέτες όπου οι αποκρίσεις μετρώνται για το ίδιο αντικείμενο άρα είναι συσχετισμένες και θεωρώντας τις ανεξάρτητες οδηγούμαστε σε εσφαλμένα συμπεράσματα. Για το λόγο αυτό αντί των γραμμικών μοντέλων χρησιμοποιούνται τα GLM που είναι πιο κατάλληλα.

Γενικά, ο σκοπός των σχεδιασμών είναι ο καθορισμός των μεταβλητών παλινδρόμησης με στόχο την παροχή ικανοποιητικών αποτελεσμάτων. Τα GLM είναι μια ενοποιημένη κλάση μοντέλων παλινδρόμησης για διακριτές και συνεχείς μεταβλητές απόκρισης π.χ. λογιστική παλινδρόμηση για Δυωνυμική απόκριση.

Εφαρμογές των GLM συναντούμε σε οικονομικά, έλεγχο ποιότητας, απλές μελέτες, οικονομετρία, επιδημιολογία σχετικά με χρόνιες ασθένειες.

4.3 Επιλογή Σχεδιασμού

Η επιλογή σχεδιασμού αναφέρεται στον προσδιορισμό των μεταβλητών ελέγχου με αποτέλεσμα προβλεπόμενες αποκρίσεις με επιθυμητές ιδιότητες. Ένας καλός σχεδιασμός είναι αυτός που ελαχιστοποιεί το τετράγωνο του μέσου σφάλματος $\hat{\mu}(x)$:

$$\mathrm{MSE}[\mu(x)] = \mathrm{E}[\mu(x) - \mu(\mathrm{x})]^2 = \mathrm{var}[\mu(x)] + \{\mathrm{Bias}\,\mu(x)\}^2$$

η οποία όμως εξαρτάται από το άγνωστο διάνυσμα των παραμέτρων β . Το πρόβλημα αυτό εξάρτησης σχεδιασμού από τις άγνωστες παραμέτρους εμφανίζεται στους A, D, E, G βέλτιστους ελέγχους αν και είναι κριτήρια βασισμένα στη διακύμανση.

Οι προτεινόμενες λύσεις του προβλήματος είναι οι εξής :

- ο προσδιορισμός αρχικών τιμών ή υποθέσεων για τις παραμέτρους ο οποίος οδηγεί σε τοπικούς βέλτιστους σχεδιασμούς
- η διαδοχική προσέγγιση που επιτρέπει στον χρήστη την παροχή εκτιμητριών των παραμέτρων σε επιτυχημένα στάδια ξεκινώντας από τις αρχικές τιμές του πρώτου σταδίου
- η μπεϋζιανή προσέγγιση όπου υποθέτουμε μια πρωταρχική κατανομή των παραμέτρων η οποία ενσωματώνεται σε ένα κατάλληλο σχεδιασμό ενσωματώνοντας το στην αρχική κατανομή
- η χρήση σχεδιαγραμμάτων διασποράς που επιτρέπει τη σύγκριση

Πιο γνωστά παραδείγματα τοπικών βέλτιστων σχεδιασμών είναι η ανάλυση δυαδικών δεδομένων με λογιστικό μοντέλο παλινδρόμησης και οι Poisson καταμετρήσεις.

4.3.1 Λογιστικό Μοντέλο Παλινδρόμησης για Δυαδικά Δεδομένα

Για μια δυαδική απόκριση y με τιμές 0 και 1 σημαντικό ρόλο παίζει η μη στοχαστική μεταβλητή X που παίρνει συγκεκριμένες τιμές. Για X= x η y ισούται με 1 με πιθανότητα :

$$P(x_i) = \frac{1}{1 + \exp(-a - \beta x_i)}$$

με α , β άγνωστες παράμετροι με β >0. Οι παράμετροι αλλά και κάποιες παραμετρικές συναρτήσεις όπως $\frac{\alpha}{\beta}$ είναι σημαντικές για τον ερευνητή. Σκοπός είναι να δημιουργήσουμε συνεχείς βέλτιστους σχεδιασμούς ώστε να έχουμε βέλτιστες τιμές για το $\alpha + \beta x$ γι΄αυτό χρειάζονται αρχικές τιμές για τις α, β. Αν υποθέσουμε τις διακριτές δόσεις x_1, x_2,….. x_s και θέλουμε να έχουμε n_i παρατηρήσεις στις αποκρίσεις y με $\sum_{i=1}^{s} n_i = n$.

Το πρόβλημα του βέλτιστου σχεδιασμού έγκειται στην επιλογή βέλτιστου αριθμού s διακριτών επιπέδων δόσεων για ένα συγκεκριμένο n. Βέβαια, μεγάλο ενδιαφέρον παρουσιάζει ο υπολογισμός :

- των β, $\frac{\alpha}{\beta}$, $P(x_i)$
- των κοινών ζευγών εκτιμητριών παραμέτρων όπως α και β, β και $\frac{\alpha}{\beta}$, β και $P(x_i)$

Για τον υπολογισμό οποιασδήποτε παραπάνω μεταβλητής αρχικά θεωρούμε τον ασυμπτωτικό πίνακα διακύμανσης- συνδιακύμανσης της Μεγίστης Πιθανοφάνειας εκτιμητριών των α και β και στη συνέχεια επιλέγουμε τα x_i και τις στάθμες p_i ελαχιστοποιώντας την κατάλληλη συνάρτηση βασιζόμενοι στην φύση του προβλήματος αλλά και στο κριτήριο βελτιστοποίησης που εφαρμόζεται.

Για το λόγο αυτό ακολουθούμε την εξής διαδικασία:

- θεωρούμε τον πίνακα πληροφορίας των δυο παραμέτρων ως σταθμισμένο συνδυασμό πινάκων πληροφορίας βασισμένους στους x_i χρησιμοποιώντας τις p_i ως στάθμες.
- θεωρούμε ότι ο ασυμπτωτικός πίνακας διακύμανσης-συνδιακύμανσης των εκτιμητριών Μεγίστης Πιθανοφάνειας των α και β είναι ο αντίστροφος του σταθμισμένου πίνακα πληροφορίας που υπολογίστηκε προηγουμένως .
- αυτό ισοδυναμεί με την ελαχιστοποίηση της κατάλληλης συνάρτησης του πίνακα διακύμανσης – συνδιακύμανσης των εκτιμητριών Μεγίστης Πιθανοφάνειας των παραμέτρων.

Τα βέλτιστα κριτήρια A και D είναι παραδείγματα τέτοιου τύπου.

Ο βέλτιστος έλεγχος D λαμβάνει περισσότερης προσοχής ενώ ο βέλτιστος έλεγχος A έχει χρησιμοποιηθεί από κάποιους συγγραφείς (Sitter and Wu, 1993). Τα βέλτιστα επίπεδα δόσης εξαρτώνται από τις άγνωστες παραμέτρους α και β. Στην πραγματικότητα, λύσεις για τον βέλτιστο σχεδιασμό που προαναφέρθηκε είναι οι βέλτιστες τιμές της παράστασης $\alpha + \beta x_i$. Έτσι, για να εφαρμόσουμε στην πράξη το μοντέλο χρειαζόμαστε καλές αρχικές τιμές.

Παρότι έχουμε κατορθώσει να κατασκευάσουμε κυρίως τους A και D βέλτιστους σχεδιασμούς μπορέσαμε να βρούμε τον E βέλτιστο σχεδιασμό σε μερικές περιπτώσεις. Τις περισσότερες φορές όμως οι A βέλτιστοι σχεδιασμοί παράγονται σε κλάσεις συμμετρικών σχεδιασμών.

Υποθέτοντας ότι $\xi_i = n_i / n$ με $\xi_i > 0$ και $\sum_{i=1}^{m} \xi i = 1$ και ο σχεδιασμός συμβολίζεται με $D = \{ (x_i, \xi_i)\ i = 1,2,\ldots,m \}$ με $0 < x < \infty$.

Πρέπει να σημειώσουμε ότι με I(α, β) αναφερόμαστε στον πίνακα πληροφορίας των α, β και αν θ_1 και θ_2 συναρτήσεις των α, β τότε ο πίνακας πληροφορίας τους είναι J I(α, β)J΄ όπου ο πίνακας J δεν εξαρτάται από τα επίπεδα της δόσης με αποτέλεσμα ο D βέλτιστος σχεδιασμός να είναι ίδιος για τον υπολογισμό των α και β.

Παρακάτω παρουσιάζονται υπολογισμοί για το λογιστικό μοντέλο παλινδρόμησης μέσω βελτιστοποιήσεων.

Α. ΥΠΟΛΟΓΙΣΜΟΣ ΤΩΝ α ΚΑΙ β

Αρχικά πρέπει να λάβουμε υπόψη το παρακάτω λήμμα.

Λήμμα 1:

Αν ξ_i θετικοί πραγματικοί αριθμοί με $\sum_{i=1}^{m} \xi i = 1$ τότε για οποιαδήποτε ομάδα διακριτών πραγματικών αριθμών $\alpha_1, \alpha_2, \ldots \alpha_m$ υπάρχει c τέτοιο ώστε να ισχύει :

$$\sum_{i=1}^{m} \xi_i \frac{e^{a_i}}{(1+e^{a_i})^2} = \frac{e^{c}}{(1+e^{c})^2} \qquad (4.4)$$

$$\sum_{i=1}^{m} \xi_i a_i^2 \frac{e^{a_i}}{(1+e^{a_i})^2} \le c^2 \frac{e^c}{(1+e^c)^2}$$

Η πλήρης απόδειξη του λήμματος 1 βρίσκεται στο παράρτημα του βιβλίου.

Υποθέτουμε ότι $\alpha_i = \alpha + \beta x_i$ τότε ο πίνακας πληροφορίας για τον κοινό υπολογισμό των α και β είναι :

$$I(\alpha, \beta) = \begin{pmatrix} \sum_{i=3}^{m} \xi_i \frac{e^{-a_i}}{(1+e^{-a_i})^2} & \sum_{i=3}^{m} \xi_i x_i \frac{e^{-a_i}}{(1+e^{-a_i})^2} \\ \sum_{i=3}^{m} \xi_i x_i \frac{e^{-a_i}}{(1+e^{-a_i})^2} & \sum_{i=3}^{m} \xi_i x_i^2 \frac{e^{-a_i}}{(1+e^{-a_i})^2} \end{pmatrix}$$

D –βελτιστοποίηση :

Πρέπει να μεγιστοποιήσουμε την ορίζουσα του κοινού πίνακα πληροφορίας των α και β ως εξής :

$$\beta^2 |I(\alpha, \beta)| = [\sum_{i=1}^{m} \xi_i \frac{e^{-a_i}}{(1+e^{-a_i})^2}][\sum_{i=1}^{m} \xi_i a_i^2 \frac{e^{-a_i}}{(1+e^{-a_i})^2}] - [\sum_{i=1}^{m} \xi_i a_i \frac{e^{-a_i}}{(1+e^{-a_i})^2}]$$

Επειδή ο D - βέλτιστος σχεδιασμός πρέπει να είναι συμμετρικός στα α_i και $-\alpha_i$ δηλ. συμβαίνουν στην ίδια στάθμη. Έτσι, η παραπάνω εξίσωση απλοποιείται ως :

$$\beta^2 \sum_{i=1}^{m} \xi_i \frac{e^{-a_i}}{(1+e^{-a_i})^2} = \sum_{i=1}^{m} \xi_i \frac{e^{-a_i}}{(1+e^{-a_i})^2} \sum_{i=1}^{m} \xi_i a_i^2 \frac{e^{-a_i}}{(1+e^{-a_i})^2} \qquad (4.5)$$

αφού ισχύει $\frac{e^{a_i}}{(1+e^{a_i})^2} = \frac{e^{-a_i}}{(1+e^{-a_i})^2}$.

Από το λήμμα 1 και τη σχέση (4.5) για οποιοδήποτε σχεδιασμό υπάρχει c που ικανοποιεί την :

$$\beta^2 |I(\alpha, \beta)| \le \frac{e^c}{(1+e^c)^2} c^2 \frac{e^c}{(1+e^c)^2} = c^2 \frac{e^{2c}}{(1+e^c)^4}$$

Με άλλα λόγια, ο συμμετρικός σχεδιασμός {(c ,1/2) , (-c,1/2)} μεγιστοποιεί την $|I(\alpha, \beta)|$ όπου το c προέρχεται από τη μεγιστοποίηση του $c^2 \frac{e^{2c}}{(1+e^c)^4}$.

Η μέγιστη της c είναι $c_D = 1.5434$.

Έτσι, ο D βέλτιστος σχεδιασμός αποτελείται από τα σημεία x_{1D} και x_{2D} με στάθμες ½ το καθένα και ικανοποιούν τις σχέσεις :

$$\alpha + \beta x_{1D} = -c_D \; και \; \alpha + \beta x_{2D} = c_D$$

A-βελτιστοποίηση:

Για να αποκτήσουμε τον A βέλτιστο σχεδιασμό πρέπει να ελαχιστοποιήσουμε την σχέση Var ($\hat{a}$) + Var ($\hat{\beta}$) όπου $\hat{a} \; και \; \hat{\beta}$ εκτιμήτριες των α και β και η διακύμανση που υπολογίζουμε είναι η ασυμπτωτική διακύμανση.

$$\text{Var}(\alpha) + \text{Var}(\beta) = \sum_{i=1}^{m} \xi_i \frac{e^{-a_i}}{(1+e^{-a_i})^2}[1+x_i^2] / \left|\text{I}(\alpha,\beta)\right|$$

$$= \sum_{i=1}^{m} \xi_i \frac{e^{-a_i}}{(1+e^{-a_i})^2}[1+\frac{(a_i-a)^2}{\beta^2}] / \left|\text{I}(\alpha,\beta)\right| \qquad (4.6)$$

Δεν έχουμε πλήρη λύση του A-βέλτιστου προβλήματος παρόλα αυτά αν περιοριστούμε σε συμμετρικά μοντέλα τότε ο A-βέλτιστος σχεδιασμός μπορεί να αποκτηθεί εφαρμόζοντας το λήμμα 1.

Η σχέση (4.6) απλοποιείται ως εξής:

$$\text{Var}(\alpha) + \text{Var}(\beta) = \frac{\frac{1}{\beta^2}\sum_{i=1}^{m} \xi_i \frac{e^{-a_i}}{(1+e^{-a_i})^2}[\alpha^2+\beta^2+\alpha_i^2]}{\frac{1}{\beta^2}\sum_{i=1}^{m} \xi_i \frac{e^{-a_i}}{(1+e^{-a_i})^2} \sum_{i=1}^{m} \xi_i a_i^2 \frac{e^{-a_i}}{(1+e^{-a_i})^2}}$$

$$= \frac{\alpha^2+\beta^2}{\sum_{i=1}^{m} \xi_i a_i^2 \frac{e^{-a_i}}{(1+e^{-a_i})^2}} + \frac{1}{\sum_{i=1}^{m} \xi_i \frac{e^{-a_i}}{(1+e^{-a_i})^2}}$$

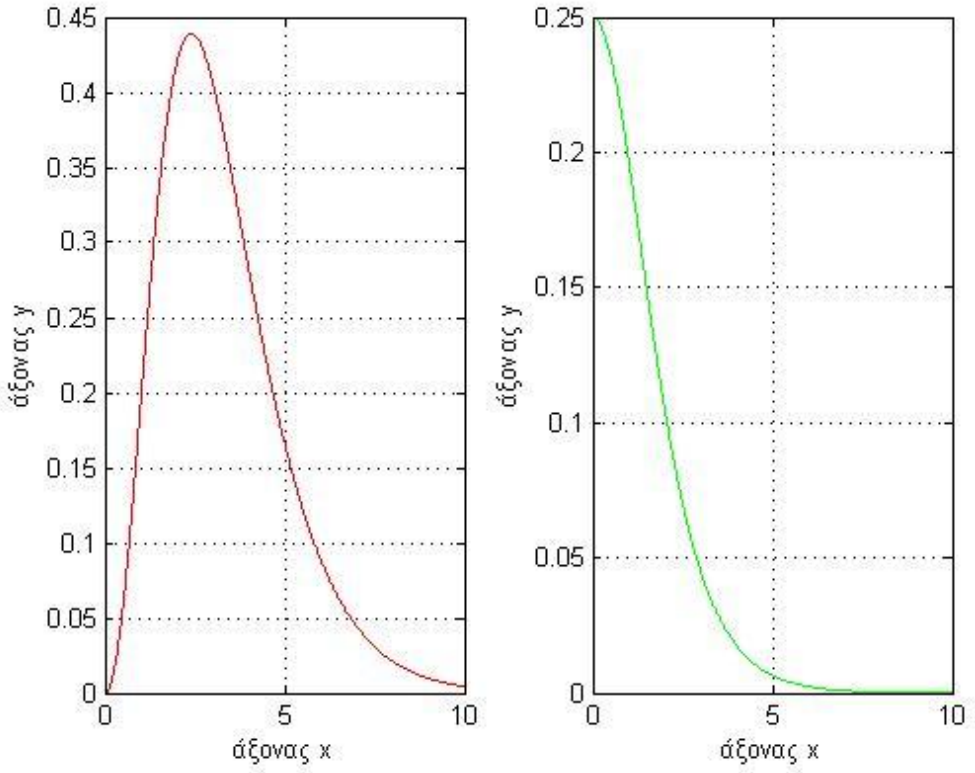
Από τη σχέση (4.4) προκύπτει το εξής :

$$\text{Var}(\alpha) + \text{Var}(\beta) \geq \frac{\alpha^2+\beta^2}{c^2 \frac{e^c}{(1+e^c)^2}} + \frac{1}{\frac{e^c}{(1+e^c)^2}}$$

Με άλλα λόγια για την κλάση των συμμετρικών σχεδιασμών ο A- βέλτιστος σχεδιασμός δίνεται από το { (c,1/2) , (-c, 1/2)} όπου η c ελαχιστοποιεί την σχέση :

$$\frac{\alpha^2+\beta^2}{c^2\dfrac{e^c}{(1+e^c)^2}}+\frac{1}{\dfrac{e^c}{(1+e^c)^2}} \qquad (4.7)$$

Έχοντας τις αρχικές τιμές των α και β η A- βέλτιστη επιλογή του c ας πούμε c_A μπορεί να βρεθεί αριθμητικώς ελαχιστοποιώντας την εξίσωση (4.7).



Στο παραπάνω σχεδιάγραμμα , έχουμε τις συναρτήσεις $c^2\dfrac{e^c}{(1+e^c)^2}$ και $\dfrac{e^c}{(1+e^c)^2}$ με $c \geq 0$.

- Η συνάρτηση $\frac{e^c}{(1+e^c)^2}$ είναι γνησίως φθίνουσα συνάρτηση της c με c ≥ 0 με μέγιστη τιμή το 1/4 στο σημείο c = 0.
- Αντίστοιχα για c $\geq$ 0, η συνάρτηση $c^2\frac{e^c}{(1+e^c)^2}$ είναι αύξουσα ,παίρνει μέγιστη τιμή στο σημείο c = 2.399 (περίπου) και μετά είναι φθίνουσα.

Αποτέλεσμα αυτών είναι αν ελαχιστοποιήσουμε μια εκ των δυο φθινουσών συναρτήσεων $c^2\frac{e^c}{(1+e^c)^2}$ και $\frac{e^c}{(1+e^c)^2}$ τότε η ελάχιστη τιμή της c δεν θα υπερβαίνει την τιμή 2.399 .

Αν και κατασκευάσαμε ένα A-βέλτιστο σχεδιασμό στις κλάσεις συμμετρικών σχεδιασμών είναι ξεκάθαρο από τη σχέση (4.6) ότι ο A- βέλτιστος σχεδιασμός στην κλάση των σχεδιασμών μπορεί να μην είναι συμμετρικός αν α = 0.

Είναι λοιπόν δύσκολο να χαρακτηριστεί ένας A-βέλτιστος σχεδιασμός για όλους τους σχεδιασμούς.

Αν περιοριστούμε σε ένα σχεδιασμό δυο σημείων {(c_1, ξ_1) , (c_2, ξ_2) } μπορούμε να αποκτήσουμε αριθμητικώς έναν A-βέλτιστο σχεδιασμό μέσα σε κλάση όλων των σχεδιασμών. Αν περιοριστούμε σε ένα δυο σημείων σχεδιασμό, το πρόβλημα του A- βέλτιστου σχεδιασμού περιορίζεται στην ελαχιστοποίηση του :

$$\frac{\xi_1 C_1\left[\beta^2+(c_1-a)^2\right]+\xi_2 C_2\left[\beta^2+(c_2-a)^2\right]}{(\xi_1 C_1+\xi_2 C_2)(\xi_1 c_1^2 C_1+\xi_2 c_2^2 C_2)-(\xi_1 c_1 C_1+\xi_2 c_2 C_2)^2}$$

με

$$C_i=\frac{e^{-c_i}}{(1+e^{-c_i})^2} \quad \text{i= 1,2.}$$

Β. ΥΠΟΛΟΓΙΣΜΟΣ ΤΩΝ $\theta_1 = \alpha/\beta$ ΚΑΙ β

Ο πίνακας πληροφορίας τώρα δίνεται από τον τύπο :

$$I(\theta_1, \beta) = \begin{pmatrix} \beta^2 \sum_{i=1}^{m} \xi_i \frac{e^{-a_i}}{(1+e^{-a_i})^2} & \sum_{i=1}^{m} \xi_i \alpha_i \frac{e^{-a_i}}{(1+e^{-a_i})^2} \\ \sum_{i=1}^{m} \xi_i \alpha_i \frac{e^{-a_i}}{(1+e^{-a_i})^2} & \frac{1}{\beta^2} \sum_{i=1}^{m} \xi_i \alpha_i^2 \frac{e^{-a_i}}{(1+e^{-a_i})^2} \end{pmatrix}$$

όπου $\alpha_i = \alpha + \beta x_i = \beta (x_i + \theta_1)$

<u>Λήμμα 2 :</u>

Αν $I(\theta_1, \beta)$ πίνακας όπως αναφέρθηκε πριν και $I_c(\theta_1, \beta)$ ο πίνακας πληροφορίας του σχεδιασμού $\{(c, 1/2), (-c, 1/2)\}$ όπου η c ικανοποιεί το λήμμα 1 τότε το διάνυσμα των ιδιοτιμών του $I(\theta_1, \beta)$ κυριαρχεί του διανύσματος των ιδιοτιμών του $I_c(\theta_1, \beta)$.

Η πλήρης απόδειξη του Λήμματος 2 βρίσκεται στο παράρτημα.

<u>Υποσημείωση</u> :

Για δυο διανύσματα x , y $\in R^n$ για τα οποία ισχύει $x_1 \geq x_2 \geq \ldots \geq x_n, y_1 \geq y_2 \ldots \geq y_n$

λέμε ότι το y κυριαρχεί του x και συμβολίζουμε $x \geq y$ αν :

$$\sum_{i=1}^{k} x_i \geq \sum_{i=1}^{k} y_i \qquad i=1, \ldots, k$$

Α- Βελτιστοποίηση:

Ο Α –βέλτιστος σχεδιασμός ελαχιστοποιεί την :

$$\frac{(1+e^{-c})^2}{e^{-c}} [\frac{1}{\beta^2} + \frac{\beta^2}{c^2}]$$

Η αριθμητική ελαχιστοποίηση της εξίσωσης αυτής είναι εύκολη όταν η τιμή της β είναι διαθέσιμη.

Οι Sitter and Wu (1993) ασχολήθηκαν με το πρόβλημα βέλτιστου σχεδιασμού για τον υπολογισμό των θ_1 και β. Παρόλα αυτά, ο A- βέλτιστος σχεδιασμός που κατασκεύασαν είναι για τον υπολογισμό των θ_1 και 1/ β με τη διαφορά ότι η συνάρτηση προς ελαχιστοποίηση είναι η εξής :

$$\frac{(1+e^{-c})^2}{e^{-c}}[1+\frac{1}{c^2}]$$

Ενώ με τον *E- βέλτιστο σχεδιασμό* το πρόβλημα τώρα είναι η ελαχιστοποίηση της :

$$\max[\frac{(1+e^{-c})^2}{\beta^2 e^{-c}}, \frac{\beta^2(1+e^{-c})^2}{c^2 e^{-c}}]$$

Γ. ΥΠΟΛΟΓΙΣΜΟΣ ΤΩΝ β ΚΑΙ P(x)

Υποδηλώνουμε με δ το 100γ-οστό ποσοστό της P (x) με :

$\delta = \frac{l-a}{\beta}$ με $l = \ln \frac{100\,\gamma}{100(1-\gamma)}$

Ο πίνακας πληροφορίας των δ και β δίνεται ως εξής:

$$I(\delta, \beta) = \begin{pmatrix} \beta^2 \sum_{i=1}^{m} \xi_i \frac{e^{-a_i}}{(1+e^{-a_i})^2} & -\sum_{i=1}^{m} \xi_i(\alpha_i - 1)\frac{e^{-a_i}}{(1+e^{-a_i})^2} \\ -\sum_{i=1}^{m} \xi_i(\alpha_i - 1)\frac{e^{-a_i}}{(1+e^{-a_i})^2} & \frac{1}{\beta^2}\sum_{i=1}^{m} \xi_i(\alpha_i - 1)^2 \frac{e^{-a_i}}{(1+e^{-a_i})^2} \end{pmatrix}$$

με $\alpha_i = \alpha + \beta x_i = l + \beta (x_i - \delta)$.

Είναι προφανές ότι :

$$\left|I(\delta,\ \beta)\right| = [\sum_{i=1}^{m} \xi_i \frac{e^{-a_i}}{(1+e^{-a_i})^2}][\sum_{i=1}^{m} \xi_i a_i^2 \frac{e^{-a_i}}{(1+e^{-a_i})^2}] - [\sum_{i=1}^{m} \xi_i a_i \frac{e^{-a_i}}{(1+e^{-a_i})^2}]^2$$

Ο A- βέλτιστος σχεδιασμός ελαχιστοποιεί την :

$$\text{Var}\,(\delta)\ +\ \text{Var}\,(\beta) = \sum_{i=1}^{m} \xi_i \frac{e^{-a_i}}{(1+e^{-a_i})^2}[\beta^2 + \frac{(\alpha_i - l)^2}{\beta^2}] / \left|\mathrm{I}(\delta,\ \beta)\right|$$

όπου $\hat{\delta}$ *και* $\hat{\beta}$ οι εκτιμήτριες Μεγίστης Πιθανοφάνειας και οι διακυμάνσεις είναι οι ασυμπτωτικές διακυμάνσεις.

Αν λάβουμε υπόψη μας την :

$$\text{Var}\,(\delta)\ +\ \text{Var}\,(\beta) = \frac{1}{\beta^2}[\frac{\beta^4 + l^2}{\sum_{i=1}^{m} \xi_i \alpha_i^2 \frac{e^{-a_i}}{(1+e^{-a_i})^2}} + \frac{1}{\sum_{i=1}^{m} \xi_i \frac{e^{-a_i}}{(1+e^{-a_i})^2}}]$$

και ο σχεδιασμός δίνεται από το {(c ,1/2), (-c, 1/2)} όπου το c ελαχιστοποιεί την παράσταση:

$$\frac{1}{\beta^2}[\frac{\beta^4 + l^2}{c^2 \frac{e^c}{(1+e^c)^2}} + \frac{1}{\frac{e^c}{(1+e^c)^2}}]$$

Δ. ΥΠΟΛΟΓΙΣΜΟΣ ΔΥΟ P(x)

Έστω δ_1 και δ_2 αντίστοιχα τα $100\gamma_1$ και $100\gamma_2$ των P(x) υποθέτοντας ότι $\gamma_1 > \gamma_2$.

Ορίζουμε ότι :

$l_1 = \ln(\frac{100\gamma_1}{100\,(1-\gamma_1)})$, $l_2 = \ln(\frac{100\gamma_2}{100\,(1-\gamma_2)})$

$\delta_1 = \frac{l_1 - \alpha}{\beta}$, $\delta_2 = \frac{l_2 - \alpha}{\beta}$

Ο πίνακας πληροφορίας των (δ_1, δ_2) δίνεται ως :

$\mathrm{I}\left(\delta_1, \delta_2\right) =$

$$\frac{1}{(\delta_1 - \delta_2)^2}\begin{pmatrix} \sum_{i=1}^{m} \xi_i (\alpha_i - l_2)^2 \frac{e^{-\alpha_i}}{(1+e^{-\alpha_i})^2} & -\sum_{i=1}^{m} \xi_i (\alpha_i - l_1)(\alpha_i - l_2) \frac{e^{-\alpha_i}}{(1+e^{-\alpha_i})^2} \\ -\sum_{i=1}^{m} \xi_i (\alpha_i - l_1)(\alpha_i - l_2) \frac{e^{-\alpha_i}}{(1+e^{-\alpha_i})^2} & \sum_{i=1}^{m} \xi_i (\alpha_i - l_1)^2 \frac{e^{-\alpha_i}}{(1+e^{-\alpha_i})^2} \end{pmatrix}$$

όπου $\alpha_i = \alpha + \beta x_i = \frac{1}{\delta_1 - \delta_2}[(l_2 - l_1) + (l_1 - l_2)x_i]$.

Τότε :

$$\frac{(\delta_1 - \delta_2)^4}{(l_1 - l_2)^2}\left|I(\delta_1, \delta_2)\right| = [\sum_{i=1}^{m} \xi_i \frac{e^{-a_i}}{(1+e^{-a_i})^2}][\sum_{i=1}^{m} \xi_i a_i^2 \frac{e^{-a_i}}{(1+e^{-a_i})^2}] - [\sum_{i=1}^{m} \xi_i a_i \frac{e^{-a_i}}{(1+e^{-a_i})^2}]$$

αν όμως λάβουμε υπόψη μας συμμετρικούς σχεδιασμούς τότε η εξίσωση αυτή ισοδυναμεί με :

$$\frac{2}{\sum_{i=1}^{m} \xi_i \frac{e^{-a_i}}{(1+e^{-a_i})^2}} + \frac{l_1^2 + l_2^2}{\sum_{i=1}^{m} \xi_i a_i^2 \frac{e^{-a_i}}{(1+e^{-a_i})^2}}$$

Εφαρμόζοντας το λήμμα 1 βλέπουμε ότι ο A- βέλτιστος σχεδιασμός που δίνεται ως {(c, 1/2), (-c, 1/2)} όπου το c ελαχιστοποιεί την :

$$\frac{2}{\frac{e^c}{(1+e^c)^2}} + \frac{l_1^2 + l_2^2}{c^2 \frac{e^c}{(1+e^c)^2}}$$

Έτσι λοιπόν για προβλήματα δυο παραμέτρων που είναι συναρτήσεις των α και β έχουμε δημιουργήσει μια προσέγγιση για την κατασκευή D και A βέλτιστους σχεδιασμούς εφαρμόζοντας το λήμμα 1.

Σε μερικές περιπτώσεις, έχουμε καταφέρει να κατασκευάσουμε A – βέλτιστους σχεδιασμούς μόνο στην κλάση συμμετρικών σχεδιασμών. Ωστόσο, τέτοιοι περιορισμοί για τον D- βέλτιστο σχεδιασμό δεν υπάρχουν.

Πρέπει να σημειώσουμε ότι όλοι οι σχεδιασμοί είναι σχεδιασμοί είναι δυο σημείων. Αριθμητικά αποτελέσματα δείχνουν ότι ο A – βέλτιστος σχεδιασμός είναι πιθανά συμμετρικός κατά σημείο αλλά όχι συμμετρικός σχεδιασμός δηλ. αν ο A – βέλτιστος σχεδιασμός είναι πάντα σχεδιασμός δυο σημείων είναι ερώτημα.

4.3.2 Poisson μοντέλο καταμέτρησης

Μελετούμε το Poisson μοντέλο καταμέτρησης και εξηγούμε τα βέλτιστα αποτελέσματα ακολουθώντας τους Minkin(1993) και Liski et al. (2002) υποθέτοντας

ότι η απόκριση y ακολουθεί την Poisson κατανομή με μέση τιμή μ(x) = c(x)exp[θ(x)] όπου θ(x) = αx + β, β>0. Επίσης σύμφωνα με τους Liski et al (2002) ισχύουν :

$$E(y) = \alpha + \beta x$$
$$V(y) = v(x)\,\sigma^2$$
$$v(x) = \exp(x)\ \mu\varepsilon\ 0 \le x \le \infty$$

Όμως οι Das, Mandal and Sinha (2003) γενίκευσαν το αποτέλεσμα αλλάζοντας το , v(x) σε :

- $v(x) = k^x$, $k \ge 1$
- $v(x) = (1+ x)^{(1+\gamma)/2}$ $\gamma \ge -1$

Στη συνέχεια, υποθέτουμε τη συνάρτηση διακύμανσης v(x) με v(0) =1 και η v(x) αυξάνεται στον x από 0 ως ∞. Ακολούθως , ξεκινάμε ένα σχεδιασμό 2 σημείων της μορφής [(α, p) , (b , q)] με 0< α< b < ∞ και 0 < p, q = 1-p < 1.

Δυο σημαντικές συναρτήσεις είναι οι ψ(x) = 1 / v(x) και φ(x) = (v(x) - 1) / x για κάθε x > 0 για τις οποίες ισχύουν :

- φ(0) = 1
- φ(x) αυξάνεται στον x
- υπάρχουν 0<s<1, c>0 τέτοια ώστε :

  $1 - s = [\ p(1 - \psi(\alpha)) + q\,(1- \psi(b))\ ] / [1 - \psi(c)]$

  με $\psi(c) < p\,\psi(\alpha) + q\,\psi(b)$

Οι Das, Mandal and Sinha (2003) απέδειξαν ότι και για τις δυο μορφές του v(x) οι παραπάνω προϋποθέσεις ικανοποιούνται.

4.4 Τύποι Σχεδιασμών

4.4.1 Ορθογώνιοι Σχεδιασμοί στα GLM

Είναι σαφές ότι οι τιμές της σχέσης (4.2) βασίζονται στις παραμέτρους οι οποίες είναι συναρτήσεις της κατανομής αλλά και της συνάρτησης σύνδεσης. Στην πραγματικότητα, δοσμένης μιας κατανομής η επιλογή συγκεκριμένης συνάρτησης σύνδεσης μπορεί να αλλάξει τις ιδιότητες του μοντέλου μέσω του πίνακα

πληροφορίας. Στην περίπτωση της Poisson απόκρισης, το μοντέλο της ομογενούς διακύμανσης είναι:

$$E(y^{1/2}) = x'\beta \tag{4.8}$$

$$\mu^{1/2} = x'\beta \tag{4.9}$$

Για το μοντέλο της εξίσωσης (4.8) τα δεδομένα μετασχηματίζονται ενώ για το μοντέλο της εξίσωσης (4.9) δηλώνει ένα μετασχηματισμό στο μέσο του πληθυσμού. Για το Poisson μοντέλο ο τετραγωνικής ρίζας μετασχηματισμός (4.9) ονομάζεται **σύνδεση σταθεροποιημένης διακύμανσης.**

Η σειρά Taylor $1^{ης}$ τάξεως του $y^{1/2}$ στο $y = \mu_y$ μας δίνει:

$$y^{1/2} = \mu_y^{1/2} + [\frac{\partial(y^{1/2})}{\partial y}]_{y=\mu_y}(y-\mu_y)$$

$$= \mu_y^{1/2} + \frac{1}{2}\mu_y^{-1/2}(y-\mu_y)$$

Συνεπώς, αν var (y) = $\mu_y^{1/2}$ (Poisson κατανομή) τότε: var ($y^{1/2}$) = $\frac{1}{4}$. (4.10)

Ομοίως , δείχνουμε ότι για την εκθετική κατανομή η σύνδεση σταθεροποιημένης διακύμανσης είναι η Λογιστική σύνδεση.

Ωστόσο, πρέπει να διευκρινιστεί ότι ο όρος σύνδεση σταθεροποιημένης διακύμανσης δεν συνεπάγεται ότι η διακύμανση της κατανομής της y είναι σταθερή. Οι κατανομές Poisson, εκθετική και άλλες έχουν διακυμάνσεις που είναι συναρτήσεις του μέσου. Η έννοια "σταθεροποιημένη διακύμανση " σημαίνει ότι η εφαρμογή της συνάρτησης στα δεδομένα δηλ. στην απόκριση y έχει σαν αποτέλεσμα την προσεγγιστική σταθεροποίηση των μεταβλητών.

Σε ένα 2^k κλασματικό παραγοντικό σχεδιασμό η χρήση της σύνδεσης σταθεροποιημένης διακύμανσης έχει σαν αποτέλεσμα ένα ορθογώνιο μοντέλο. Αυτό αποδεικνύεται εύκολα αν θυμηθούμε ότι W = ΔVΔ και

$$w_{ii} = \frac{\partial\theta_i}{\partial(x_i'\beta)} \tag{4.11}$$

$$v_i = \text{var}(y_i)\frac{\partial \mu_i}{\partial \theta_i} \qquad (4.12)$$

Επίσης υποθέτουμε ως συνάρτηση σύνδεσης : g (μ) = x΄β. Αν ο μετασχηματισμός γίνει στο y τότε ιδανικό μοντέλο είναι το E[g(y)] = x΄β. Επεκτείνοντας τη g(y) σε σειρά Taylor στο $y = \mu_y$:

$$g(y) = g(\mu_y) + [g'(\mu_y)]\,[y - \mu]$$

με $g'(\mu_y) = [\partial g / \partial y]_{y=\mu y}$ και $\text{var}[g(y)] = [g'(\mu_y)]^2 \text{var}(y)]$Έτσι, αν g(y) είναι σύνδεση σταθεροποιημένης διακύμανσης τότε $[g'(\mu_y)]^2$ είναι ανάλογη με 1/ [var (y)]. Ο σχεδιασμός είναι ορθογώνιος αν :

$$W = \Delta V \Delta = kI$$

όπου k σταθερά αν όλες οι εσσιανές στάθμες είναι ίσες. Επίσης, αν $w_{ii} = \delta_i^2$ var (y_i) με

$$\delta_i = \partial \theta_i / \partial (x_i' \beta)$$

$$w_{ii} = \partial \theta_i / \partial (x_i' \beta)\, \partial \mu_{y_i} / \partial (x_i' \beta) \partial \theta_i / (x_i' \beta)$$

ή βάσει του κανόνα αλυσίδας :

$$w_{ii} = \partial \theta_i / \partial (x_i' \beta)\, \partial \mu_{y_i} / \partial (x_i' \beta) \qquad (4.13)$$

Ακόμη, αν υποθέσουμε ότι χρησιμοποιούμε την σύνδεση σταθεροποιημένης διακύμανσης τότε έχουμε:

$$g'(\mu_{yi})\ \text{var}\left(y_i\right) = k^* \qquad (4.14)$$

όμως : $g(\mu) = x'\beta$ άρα $\dfrac{\partial (x_i'\beta)}{\partial \mu_{y_i}} \dfrac{\partial (x_i'\beta)}{\partial y_i} \dfrac{\partial \mu_i}{\partial \theta_i} = k^*$

$$k^* = \frac{\partial (x_i'\beta)}{\partial \mu_{y_i}} \frac{\partial (x_i'\beta)}{\partial \theta_i} \qquad (4.15)$$

Από τις εξισώσεις (4.13) και (4.15) έχουμε ότι η σταθερά k* δεν περιέχει παραμέτρους του μοντέλου,

$$w_{ii} = \frac{\partial \theta_i}{\partial (x_i' \beta)} \frac{\partial \mu_{y_i}}{\partial (x_i' \beta)}$$

$$= \{ \frac{\partial (x_i' \beta)}{\partial \mu_{y_i}} \frac{\partial (x_i' \beta)}{\partial \theta_i} \}^{-1}$$

$$= \frac{1}{k^*} = k \qquad (4.16)$$

Έτσι, όλα οι Εσσιανές στάθμες είναι ίσες και ο πίνακας πληροφορίας της εξίσωσης (4.2) είναι της μορφής :

$$Z'Z = X'WX = k(X'X) \rightarrow (Z'Z)^{-1} = \frac{1}{k} (X'X)^{-1} \qquad (4.17)$$

Στην περίπτωση λοιπόν της σύνδεσης σταθεροποιημένης διακύμανσης ένα 2^k κλασματικό παραγοντικό σχεδιασμό που είναι ορθογώνιος για το γραμμικό μοντέλο με τιμές ± 1 τότε θα είναι ορθογώνιο και για τα GLM με τη σύνδεση σταθεροποιημένης διακύμανσης όπου ισχύει (X΄X) = NI με N συνολικό αριθμό πειραμάτων.

Οι ασυμπτωτικές διακυμάνσεις των συντελεστών του μοντέλου είναι τα διαγώνια στοιχεία του πίνακα $(Z'Z)^{-1}$ που δίνονται από τον τύπο :

$$\text{var}(b_i) = \frac{1}{kN}$$

όπου k είναι η κοινή Εσσιανή στάθμη της σχέσης (4.12). Εδώ η συνδιακύμανση είναι ασυμπτωτικά μηδενική.

Η χρήση της σύνδεσης σταθεροποιημένης διακύμανσης σε συνδυασμό με 2^k κλασματικό παραγοντικό σχεδιασμό δίνει πολύ ικανοποιητικά αποτελέσματα. Η ορθογωνιότητα διατηρείται στα GLM μέσω του πίνακα πληροφορίας παράγοντας ασυμπτωτικά ανεξάρτητες εκτιμήσεις για τους συντελεστές ίσες διακυμάνσεις.

Μήπως όμως απαιτούνται άλλες συναρτήσεις σύνδεσης?

Υπάρχουν αρκετές ενδείξεις ότι αυτό συμβαίνει σε αρκετά παραδείγματα. Επίσης, υπάρχουν περιπτώσεις όπου ένας 2^k κλασματικός παραγοντικός σχεδιασμός είναι αρκετά ικανοποιητικός χωρίς τη σύνδεση σταθεροποιημένης διακύμανσης. Ας μελετήσουμε την περίπτωση της σύνδεσης σταθεροποιημένης διακύμανσης όπου ο πίνακας πληροφορίας της εξίσωσης (4.1) περιέχει έναν Εσσιανό σταθμισμένο πίνακα $W = \text{diag}\{w_1, w_{2,\ldots\ldots,}w_n\}$

$$I(b) = X'WX = k(X'X)$$

έτσι είναι εμφανές ότι αν $X'X$ είναι διαγώνιος τότε οι διακυμάνσεις των παραμέτρων είναι ίσες και ασυσχέτιστες. Αν τώρα, η σύνδεση που επιλέγεται δεν είναι αυτή της σταθεροποιημένης διακύμανσης τότε τα w_i δεν είναι ίσα και $I(b) = Z'Z$ με $Z = W^{1/2} X$. Για παράδειγμα σε έναν 2^3 παραγοντικό σχεδιασμό με μεταβλητές x_1, x_2, x_3 και το GLM δίνεται από τον τύπο :

$$g(\mu) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

Ο πίνακας πληροφορίας δίνεται ως εξής:

$$Z = \begin{pmatrix} w_1^{1/2} & w_1^{1/2}x_{11} & w_1^{1/2}x_{12} & w_1^{1/2}x_{13} \\ w_2^{1/2} & w_2^{1/2}x_{21} & w_2^{1/2}x_{22} & w_2^{1/2}x_{23} \\ \vdots & \vdots & \vdots & \vdots \\ w_8^{1/2} & w_8^{1/2}x_{81} & w_8^{1/2}x_{82} & w_8^{1/2}x_{83} \end{pmatrix}$$

με x_{ij} την i-οστή στάθμη και j-οστή μεταβλητή. Οι στήλες του πίνακα X είναι ορθογώνιες και οι στήλες του Z θα είναι ορθογώνιες αν οι Εσσιανές στάθμες είναι ίσες.

Όσον αφορά τον πίνακα πληροφορίας $I(b) = Z'Z$ τα διαγώνια στοιχεία θα είναι $\sum_{i=1}^{8} w_i$ ενώ τα μη διαγώνια θα είναι οι αντιθέσεις των σταθμών δηλ.

$$I(b) = \begin{pmatrix} \sum w & contr\ A & contr\ B & contr\ C \\ & \sum w & contr\ AB & contr\ AC \\ & & \sum w & contr\ BC \\ & & & \sum w \end{pmatrix}$$

ο οποίος είναι καλά ορισμένος αν αυτές οι επιδράσεις είναι σχεδόν μηδέν. Οι αλληλεπιδράσεις παρουσιάζονται ακόμα και αν δεν υπάρχουν αλληλεπιδράσεις στο μοντέλο και έτσι για τον πίνακας πληροφορίας είναι καλύτερα αν κυριαρχούν τα διαγώνια στοιχεία. Αυτό εξαρτάται από τη φύση του Εσσιανού πίνακα.

Στα περισσότερα παραδείγματα θα περιμέναμε το άθροισμα των Εσσιανών σταθμών να κυριαρχεί των αντιθέσεων ενώ αν δεν υπάρχουν αλληλεπιδράσεις στο μοντέλο θα περιμέναμε τις αντιθέσεις των αλληλεπιδράσεων να είναι πολύ μικρές. Πρέπει όμως να έχουμε υπόψη μας ότι οι στάθμες είναι συναρτήσεις του μέσου και στην περίπτωση της κανονικής κατανομής οι στάθμες είναι ίσες με τις διακυμάνσεις.

Πολλές φορές στην καθημερινότητα μας η χρήση ενός 2^k κλασματικού παραγοντικού σχεδιασμού δίνει ικανοποιητικά αποτελέσματα ακόμη και αν δεν χρησιμοποιείται η σύνδεση σταθεροποιημένης διακύμανσης. Στην πραγματικότητα πολλά μοντέλα θα είναι σχεδόν ορθογώνια και οι διακυμάνσεις των συντελεστών σχεδόν ίσες. Θα μελετήσουμε την περίπτωση της δυωνυμικής απόκρισης με χρήση της Λογιστικής παλινδρόμησης. Αρχικά, το μοντέλο περιέχει τους όρους x_1, x_2,x_3, x_1x_2, x_1x_3, x_2x_3. Τα μη διαγώνια στοιχεία παρουσιάζουν μεγάλο ενδιαφέρον ενώ οι αντιθέσεις των Εσσιανών σταθμών είναι οι γενικευμένες αντιθέσεις αλληλεπιδράσεων όπως φαίνεται παρακάτω.

$$
\text{I(b)}=\begin{array}{cccccc}
\sum w & contr\ x_1 & contr\ x_2 & contr\ x_3 & contr\ x_1x_2 & contr\ x_1x_3 \\
 & \sum w & contr\ x_1x_2 & contr\ x_1x_3 & contr\ x_2 & contr\ x_3 \\
 & & \sum w & contr\ x_2x_3 & contr\ x_1 & contr\ x_1x_2x_3 \\
 & & & \sum w & contr\ x_1x_2x_3 & contr\ x_1 \\
 & & & & \sum w & contr\ x_2\ x_3 \\
 & & & & & \sum w
\end{array}
\qquad
\begin{array}{c}
contr\ x_2x_3 \\
contr\ x_1x_2x_3 \\
contr\ x_3 \\
contr\ x_3 \\
contr\ x_1x_3 \\
contr\ x_1x_2 \\
\\
\sum w
\end{array}
$$

Οι αντιθέσεις που εμφανίζονται στα μη διαγώνια στοιχεία είναι συχνά μικρές σε σχέση με τα διαγώνια στοιχεία. Επίσης, οι επιδράσεις στον Εσσιανό πίνακα είναι επιδράσεις στο μέσο αφού ο μέσος είναι ίσος με τη διακύμανση.

Στην πραγματικότητα, οι μεγάλες επιδράσεις φέρνουν έλλειψη αποτελεσματικότητας. Πρέπει να επισημάνουμε ότι το άθροισμα των Εσσιανών σταθμών είναι συνάρτηση του μοντέλου αλλά και της συνάρτησης σύνδεσης .

Αυτή η σχέση υποδηλώνει τη σημαντική διαφορά μεταξύ ελέγχου της αποδοτικότητας του μοντέλου για τα γραμμικά και τα μη γραμμικά μοντέλα.

Για τα γραμμικά μοντέλα, τα διαγώνια στοιχεία του πίνακα X΄X ελέγχονται από τις τιμές ± 1 τα μοντέλα αν είναι ορθογώνια τότε είναι και τα βέλτιστα. Παρόλα αυτά, στα GLM δεν μπορούμε να ελέγξουμε τα διαγώνια στοιχεία του πίνακα πληροφορίας I(b) = X΄WX αφού εξαρτώνται από τις άγνωστες παραμέτρους.

4.4.2 Διαδοχικοί Σχεδιασμοί

Οι διαδοχικοί σχεδιασμοί προτάθηκαν από τους Wu (1985), Sitter and Forbes(1997), Sitter and Wu (1997).Γνωρίζουμε ότι οι αρχικές τιμές των παραμέτρων ή "υποθέσεις" καθορίζουν ένα τοπικό βέλτιστο σχηματισμό.

Στον διαδοχικό σχεδιασμό οι πειραματισμοί δεν σταματούν στο αρχικό στάδιο αλλά αναπτύσσουν εκτιμήτριες των παραμέτρων και χρησιμοποιούνται στον προσδιορισμό πρόσθετων σημείων του σχεδιασμού σε διαδοχικά στάδια. Η διαδικασία αυτή συνεχίζεται έως ότου επιτευχθεί η σύγκλιση σε σχέση με ένα κριτήριο βελτιστοποίησης π.χ. D βέλτιστος έλεγχος.

Στόχος των βέλτιστων ελέγχων είναι η επιλογή σημείων σχεδιασμού που ελαχιστοποιούν αντικειμενική συνάρτηση που εκλαμβάνεται ως απώλεια. Σε έναν διαδοχικό σχεδιασμό μας ενδιαφέρει να αποφασίσουμε αν πρέπει να προσθέσουμε παρατηρήσειςή όχι και να σταματήσουμε τη δειγματοληψία .Αν χρειαστούν επιπλέον παρατηρήσεις και υπάρχει η δυνατότητα επιλογής από περισσότερους από ένα πληθυσμούς είναι αναγκαίος ένας κανόνας δειγματοληψίας.

Έτσι οι βέλτιστοι σχεδιασμοί και η διαδοχική ανάλυση αποτελούν αυτό που ονομάζουμε **Διαδοχικοί Πειραματικοί Σχεδιασμοί** οι οποίοι κατηγοριοποιούνται σε πολλές ομάδες ανάμεσα των οποίων και η Στοχαστική Προσέγγιση (Robbins and Monro 1951).

Στην προσέγγιση ισχύουν τα εξής :

- η απόκριση y είναι τυχαία μεταβλητή εξαρτώμενη από τη x
- η συνάρτηση κατανομής της y είναι H(y / x)
- μέση τιμή της y είναι E(y / x) = M(x)

- διακύμανση της y είναι $V(y / x) = \sigma^2(x)$.

Όμως η M(x) είναι άγνωστη στον ερευνητή, γνησίως αύξουσα τέτοια ώστε η εξίσωση $M(x) = \alpha$ έχει μοναδική ρίζα έστω ζ.

Επιθυμούμε να υπολογίσουμε την ζ κάνοντας εύστοχες παρατηρήσεις στο y π.χ. $y_1, y_2, \ldots y_n$ στα επίπεδα $x_1, x_2, \ldots x_n$ η επιλογή των οποίων είναι και το βασικότερο πρόβλημα. Για το λόγο αυτό θεωρούμε τη μη σταθερή Μαρκοβιανή αλυσίδα με αυθαίρετη αρχική τιμή x_1 και θεωρούμε αναδρομικά :

$$x_{n+1} = x_n - \alpha_n(y_n - \alpha) \qquad n = 1,2,\ldots.$$

με α_n ακολουθία θετικών σταθερών που ικανοποιεί την εξής ανισότητα:

$$0 < \sum_{n=1}^{\infty} a_n^2 < \infty$$

4.4.3 Πολυσταδιακοί Σχεδιασμοί

Στις πιο πρακτικές περιπτώσεις, επιλέγουμε ένα πολυσταδιακό σχεδιασμό παρά ένα διαδοχικό αφού οι ενημερώσεις της διαδικασίας με κάθε νέα παρατήρηση μπορεί να μην είναι εφικτές. Εφαρμογή τέτοιου τύπου είναι η Φάση I σε κλινικές μελέτες (Storer 1989) όπου στόχος είναι ο υπολογισμός της μέγιστης υποφερτής δόσης ενός νέου φαρμάκου.

Από κλινικής απόψεως, ο βέλτιστος σχεδιασμός είναι αυτός που καθορίζει τη δόση στην οποία σταματάει η δοκιμή. Αρχικά, ο Storer εισάγει τέσσερις σχεδιασμούς με ένα στάδιο ενώ αργότερα προτείνει δυο σχεδιασμούς με δυο στάδια.

4.4.4 Μπεϋζιανοί Σχεδιασμοί και Προβλήματα

Η χρήση Μπεϋζιανών σχεδιασμών στα GLM είναι μια υποσχόμενη μέθοδος αποφυγής του προβλήματος εξάρτησης των παραμέτρων του σχεδιασμού . Μια λύση του προβλήματος αυτού είναι η χρήση πειράματος με συγκεκριμένο αριθμό υποθέσεων για τις παραμέτρους οδηγώντας σε ένα τοπικό βέλτιστο σχεδιασμό (Chernoff, 1953).

Άλλη προτεινόμενη φυσική μέθοδος είναι να εκφράσουμε την αβεβαιότητα των παραμέτρων μέσω μιας εκ των προτέρων κατανομής των παραμέτρων (Box and Lucas, 1959). Το Μπεϋζιανό πρόβλημα σχεδιασμών για κανονικά γραμμικά μοντέλα

έχει συζητηθεί από τους Owen (1970), Brooks (1972, 1974, 1976, 1977), Chaloner (1984), Pilz (1991) και DasGupta(1996). Οι Atkinson and Haines(1996) επικεντρώνονται σε τοπικούς και Μπεϋζιανούς σχεδιασμούς για μη γραμμικά γενικευμένα γραμμικά μοντέλα.

Τα Μπεϋζιανά κριτήρια σχεδιασμών βασίζονται κυρίως σε κανονικές προσεγγίσεις εκ των προτέρων κατανομών του διανύσματος παραμέτρων β. Η πιο κοινή μορφή τέτοιας κανονικής προσέγγισης δηλώνει ότι υπό κανονικές συνθήκες η εκ των προτέρων κατανομή του β είναι :

$$N(\ \beta,[nI(\beta,\xi)]^{-1})$$

με $\widehat{\beta}$ την εκτιμήτρια Μεγίστης Πιθανοφάνειας του β και για οποιεσδήποτε μετρήσεις ξ ο αναμενόμενος πίνακας πληροφορίας Fischer δηλώνεται ως I(β , ξ).

Επίσης υποθέτουμε ότι οι μετρήσεις του σχεδιασμού ξ βάζουν σχετικές στάθμες ($p_1, p_2, \ldots, p_n$) σε k διακριτά σημεία ($x_1, x_2, \ldots, x_n$) αντίστοιχα με $\sum_{i=1}^{k} p_i = 1$

Πρώτο κριτήριο ανάλογο του D βέλτιστου ελέγχου είναι :

$$\varphi_1(\xi) = \mathrm{E}_{\beta}[\ \log\det \mathrm{I}(\beta,\xi)]$$

Μεγιστοποίηση αυτής της εξίσωσης είναι ισοδύναμη με μεγιστοποίηση της αναμενόμενης αύξησης της πληροφορίας Shannon ή μεγιστοποίηση της αναμενόμενης Kullback-Leibler απόστασης μεταξύ της εκ των προτέρων και της εκ των υστέρων κατανομής (Lindley, 1965; DeGroot, 1986; Bernardo, 1979).

Το επόμενο κριτήριο είναι ενδιαφέρον όταν η μόνη ποσότητα προς υπολογισμό είναι συνάρτηση του β ας πούμε h(β). Σε τέτοιες περιπτώσεις η ασυμπτωτική διακύμανση που προέρχεται από τη Μέγιστη Πιθανοφάνεια του h(β) είναι :

$$\mathrm{c}(\beta)^{\mathrm{T}}\left[\mathrm{I}\ (\beta\ ,\ \xi)\right]^{-1}\mathrm{c}(\beta)$$

όπου το i-οστό στοιχείο του c (β) είναι c_i (β) = $\frac{\partial h\ (\beta)}{\partial\ \beta_i}$.Το Μπεϋζιανό κριτήριο c-βελτιστοποίησης προσεγγίζει την εκ των υστέρων αναμενόμενη χρησιμότητα ως :

$$\varphi_2(\xi) = -\mathrm{E}_{\beta}[\mathrm{c}(\beta)^{\mathrm{T}}\left[\mathrm{I}\ (\beta\ ,\ \xi)\right]^{-1}\mathrm{c}(\beta)\ ]$$

Αν κάποιος ενδιαφέρεται να υπολογίσει διάφορες συναρτήσεις του β με πιθανότητα διαφορετικές στάθμες συνδεδεμένες σε αυτές και αν B(β) είναι το σταθμισμένο μέσο των ξεχωριστών πινάκων της μορφής c (β)c (β)^{T} τότε το κριτήριο προς μεγιστοποίηση είναι :

$$\varphi_3(\xi) = -\mathrm{E}_{\beta}\{\mathrm{trB}(\beta)\ \left[\mathrm{I}\ (\beta\ ,\ \xi)\right]^{-1}\}$$

Οι συναρτήσεις c (β), B(β) μπορεί να μην εξαρτάται από το β αν θεωρήσουμε μόνο γραμμικές συναρτήσεις του β. Άλλοι σχεδιασμοί κριτηρίων που μπορούν να συσχετιστούν με την Μπεϋζιανή προοπτική συναντώνται στους Tsukawa (1972, 1980), Zacks (1977), Pronzato and Walter(1987). Όμως δεν είναι γνωστό πόσο καλά τα Μπεϋζιανά κριτήρια έχουν μεγάλη χρησιμότητα σε μικρά δείγματα. Μερικά παραδείγματα παρουσιάζονται στους Atkinson et al (1993), Clyde (1993), Sun, Tsutakawa and Lu (1996).

Στη συνέχεια αναλύουμε διάφορα προβλήματα των Μπεϋζιανών σχεδιασμών.

Α. Μοντέλα με πολλές αποκρίσεις

Οι Draper και Hunter (1967) θεώρησαν μη γραμμικά πειράματα με πολλές αποκρίσεις και υιοθέτησαν τοπικούς βέλτιστους ή διαδοχικούς σχεδιασμούς ως βάση των σχεδιασμών. Όμως, τα πειράματα με πολλές αποκρίσεις για GLM βάσει Μπεϋζιανών σχεδιασμών δεν έχουν αναπτυχθεί αρκετά.

Οι Hatzis και Larntz (1992) θεωρούν μη γραμμικά πειράματα με πολλές αποκρίσεις με κατανομή πιθανότητας για τις αποκρίσεις που ακολουθούν τυχαίες Poisson διαδικασίες. Θεωρούν τοπικούς D βέλτιστους σχεδιασμούς που ελαχιστοποιούν τη γενικευμένη διακύμανση των εκτιμητριών των παραμέτρων για συγκεκριμένες τιμές των παραμέτρων.

Συζητούν, επίσης την περίπτωση όπου τους ενδιαφέρει μια υποομάδα των παραμέτρων με αποτέλεσμα τη χρήση τοπικών D_S βέλτιστων κριτηρίων. Χρησιμοποιούν ένα γενικευμένο αλγόριθμο μαζί με τον αλγόριθμο Simplex των Nelder και Mead (1965).

Η απόκτηση Μπεϋζιανών βέλτιστων σχεδιασμών για μη γραμμικά πειράματα με πολλές αποκρίσεις θέτει κάποιες υπολογιστικές προκλήσεις μιας και πρέπει να γίνει αριθμητική ολοκλήρωση σε μεγαλύτερους χώρους διαστάσεων.

Β. Διαδοχικά Μπεϋζιανά Μοντέλα

Ο Ridout (1995) θεωρεί ένα μοντέλο αραίωσης για τη δυαδική απόκριση σε ένα πείραμα σποράς. Υποθέσεις του Ridout είναι:

- n δείγματα σποράς
- το i-οστό δείγμα (i = 1,2,3,….,n) να περιέχει x_i σπόρους
- φέρουν δυαδική μεταβλητή απόκρισης y_i με τιμές 0 αν το δείγμα είναι ελεύθερο μολύνσεων και 1 στην αντίθετη περίπτωση
- π_i = P (με το i-οστό δείγμα ελεύθερο μόλυνσης)
- θ = αναλογία μολυσμένων σπόρων του δείγματος
- για την απόκριση y_i ισχύει $y_i \sim$ Bernoulli (π_i) με $\pi_i = 1 - (1-\theta)^{x_i}$ i = 1,2,…,n

Ο συγγραφέας εισάγει την παράμετρο λ = log {- log(1-θ)} και τα κριτήρια σχεδιασμού βασίζονται σε εκτιμήσεις των σχέσεων της πληροφορίας Fischer για το λ. Σχεδιασμοί με ένα στάδιο και τρία στάδια αντίστοιχα αναπτύχθηκαν και συγκρίθηκαν μεταξύ τους όταν τα δείγματα είναι μικρά. Οι σχεδιασμοί με τρία στάδια βρέθηκαν πιο αποτελεσματικοί από αυτούς με ένα στάδιο.

Γ. Αριθμός Σημείων

Στους κλασικούς σχεδιασμούς υπάρχει ένα ανώτερο όριο σημείων ενός βέλτιστου σχεδιασμού από το Θεώρημα του Καραθεοδωρή. Οι D βέλτιστοι σχεδιασμοί έχουν τόσα σημεία όσα και οι άγνωστες παράμετροι του μοντέλου με ίσες στάθμες σε κάθε σημείο (Silvey, 1980; Pukelsheim, 1993). Το μόνο μειονέκτημα είναι ότι δεν υπάρχουν αρκετά σημεία που να επιτρέπουν τον έλεγχο καταλληλότητας του μοντέλου. Αυτά τα ανώτερα όρια εφαρμόζονται στα τοπικά βέλτιστα κριτήρια και στα Μπεϋζιανά βέλτιστα κριτήρια για γραμμικά μοντέλα. Για μη γραμμικά μοντέλα με συνεχή εκ των προτέρων κατανομή δεν υπάρχει αντίστοιχο άνω φράγμα. Για τα περισσότερα κοίλα κριτήρια βελτιστοποίησης αν υπάρχουν k σημεία της εκ των προτέρων κατανομής τότε υπάρχει Μπεϋζιανός βέλτιστος

σχεδιασμός με $k^{p(p+1)/2}$ διαφορετικά σημεία (Dette and Neugebauer, 1996) με p να είναι ο αριθμός των παραμέτρων .

Δ. Ευαισθησία ως προς τις προηγούμενες κατανομές

Ευρωστία του σχεδιασμού ως προς την προηγούμενη κατανομή είναι μια επιθυμητή ιδιότητα. Αναπτύχθηκε από τους DasGupta and Studden (1991) , DasGupta, Mukhopadhyay and Studden (1992) , Toman and Gastwirth (1993, 1994) ένα πλαίσιο για εύρωστους πειραματικούς σχεδιασμούς στα γραμμικά μοντέλα. Όμως, χρειάζονται προσπάθειες για την πρόταση ενός εύρωστου Μπεϋζιανού πειραματικού σχεδιασμού για μη γραμμικά μοντέλα και για GLM.

Ε. Στατιστικά πακέτα

Η εύρεση βέλτιστων Μπεϋζιανών βέλτιστων σχεδιασμών για μη γραμμικά προβλήματα με πολλές παραμέτρους είναι πολύ δύσκολη και μπορεί να αποκτηθεί μόνο αριθμητικά. Ωστόσο για ιστορικούς λόγους αναφέρουμε τις παρακάτω ημερομηνίες.

- Το 1988 οι Chaloner και Larntz έκαναν την πρώτη προσπάθεια προς αυτή την κατεύθυνση εισάγοντας FORTRAN77 προγράμματα για τους Μπεϋζιανούς βέλτιστους σχεδιασμούς για τη λογιστική παλινδρόμηση με μια επεξηγηματική μεταβλητή.
- Το 1997 οι Spears, Brown και Atkinson παρουσίασαν το πρόγραμμα SINGLE με την λογιστική αλλά και την log-log συνάρτηση σύνδεσης.
- Το 1998 οι Smith και Ridout παρουσίασαν μια βελτιωμένη εκδοχή του προγράμματος τους με την ονομασία DESIGNV1 με μια μεγαλύτερη ποικιλία συναρτήσεων σύνδεσης εκτός της λογιστικής.

Το 2003 οι Smith και Ridout παρουσίασαν το λογισμικό DESIGNV2 με δυο επεξηγηματικές μεταβλητές

4.5 Ειδική Περίπτωση : Περισσότερες από μια Επεξηγηματικές Μεταβλητές

Οι Atkinson et al (1995) υποθέτουν ένα πείραμα δόσης –απόκρισης όταν αρσενικά και θηλυκά έντομα αντιδρούν διαφορετικά βάσει του μοντέλου :

$$\log(\pi(x,z) / 1-\pi(x,z)) = \alpha + \theta x + \gamma z$$

με π(x ,z) να είναι η πιθανότητα θανάτου του εντόμου με δόση x και z είναι 0 για τα αρσενικά και 1 για τα θηλυκά υποθέτοντας ότι η αναλογία αρσενικών και θηλυκών είναι ίση. Όσο μεγαλύτερος είναι ο διαχωρισμός μεταξύ των δυο ομάδων τόσο περισσότερο προτιμώνται οι τοπικοί D- βέλτιστοι σχεδιασμοί.

Επίσης, θεωρούν μια Μπεϋζιανή εκδοχή του προβλήματος αυτού επιβάλλοντας κανονική κατανομή με τρεις παραμέτρους και σημειώνουν ότι ο σχεδιασμός είναι ακμαίος σε σχέση με την αβεβαιότητα των παραμέτρων.

Οι Sitter και Torsney (1995) θεωρούν τοπικούς D βέλτιστους σχεδιασμούς όταν το μοντέλο περιέχει δυο ποσοτικές μεταβλητές . Οι Burridge και Sebastiani (1992) θεωρούν ένα γενικευμένο γραμμικό μοντέλο με δυο μεταβλητές και ένα γραμμικό προγνωστικό της μορφής $\eta = \alpha x_1 + \theta x_2$ και αποκτούν τοπικούς D-βέλτιστους σχεδιασμούς .

Οι Burridge και Sebastiani (1994) αποκτούν τοπικούς D- βέλτιστους σχεδιασμούς για τα γενικευμένα γραμμικά μοντέλα όταν οι παρατηρήσεις έχουν διακυμάνσεις ανάλογες του τετραγώνου του μέσου ενώ επιτρέπουν οποιοδήποτε αριθμό πιθανών προγνωστικών. Παρόλα αυτά, τα αποτελέσματα τους περιορίζονται σε συναρτήσεις σύνδεσης δύναμης. Αποδεικνύουν ότι υπό ορισμένες συνθήκες των παραμέτρων του μοντέλου η μετατροπή "ενός παράγοντα κάθε φορά" είναι D-βέλτιστος σχεδιασμός.

Επίσης, διεξάγουν μια αριθμητική μελέτη για να συγκρίνουν τις αποδοτικότητες των κλασικών παραγοντικών σχεδιασμών με τους βέλτιστους σχεδιασμούς και συνιστούν κάποιους αποδοτικούς συμβιβαστικούς σχεδιασμούς.

Οι Sebastiani και Settimi (1998) αποκτούν D-βέλτιστους σχεδιασμούς για μια ποικιλία μη γραμμικών μοντέλων με έναν αυθαίρετο αριθμό συντελεστών υπό συγκεκριμένες συνθήκες στον Fisher πίνακα πληροφορίας.

Στο πιο πρόσφατο άρθρο των Smith και Ridout (2003) οι βέλτιστοι Μπεϋζιανοί σχεδιασμοί αποκτώνται για βιο-δοκιμασίες με δυο παράλληλες σχέσεις δόσης-απόκρισης όπου το κύριο ενδιαφέρον είναι στον υπολογισμό της δραστικότητας ενός τεστ ναρκωτικών ή ενός τεστ χρήσης ουσιών. Το μοντέλο που θεωρείται στην περίπτωση αυτή είναι :

$$\pi(x,z) = F(\alpha + \theta(x - \rho z))$$

με z να παίρνει τις τιμές 0 και 1 για τις δυο ουσίες και F^{-1} η συνάρτηση σύνδεσης και ρ παράμετρος.

Για να αποδείξουν τους ισχυρισμούς τους οι Smith και Ridout (2003) χρησιμοποιούν μια ομάδα δεδομένων από τον Ashton (1972, p. 79) που παρέχει τον αριθμό των παρασιτικών των χρυσάνθεμων που σκοτώνονται από διαφορετικές δόσεις δυο ουσιών.

Οι Smith και Ridout (2005) εφαρμόζουν την θεωρία τους σε πολύπαραμετρικά προβλήματα δόσης-απόκρισης για να αποκτήσουν Μπεϋζιανούς βέλτιστους σχεδιασμούς σε ένα δυαδικό μοντέλο δόσης-απόκρισης με τρεις παραμέτρους με παράμετρο τον έλεγχο θνησιμότητας. Σε περιπτώσεις βιο-δοκιμασίας όπου η απόκριση είναι συχνά ο θάνατος, όμως ο θάνατος μπορεί να προέρχεται από φυσικά αίτια. Το μοντέλο που παρουσίασαν το 2005 ,που περιλαμβάνει τον έλεγχο θνησιμότητας σε ένα μοντέλο δόσης- απόκρισης , είναι :

$$\pi(x) = \lambda + (1-\lambda)F[\theta(x - \mu)]$$

με λ παράμετρο ελέγχου θνησιμότητας.

4.6 Ασυμπτωτικά αποτελέσματα

Πολλές φορές γίνεται λόγος για ασυμπτωτικά αποτελέσματα τα οποία είναι ικανοποιητικά στις περισσότερες περιπτώσεις της καθημερινότητας μας στις οποίες γίνεται χρήση των GLM. Ωστόσο, δεν είναι εύκολο να μελετήσουμε όλες τις κατανομές, τις συναρτήσεις σύνδεσης και τα μεγέθη του δείγματος. Αρχικά, θα μελετήσουμε την κατασκευή διαστημάτων εμπιστοσύνης με τη μέθοδο του Wald για τη μέση απόκριση σε περιπτώσεις με 8,16,32 επαναλήψεις.

Υπενθυμίζουμε ότι το διάστημα εμπιστοσύνης για τη μέση απόκριση στο σημείο $x_0' = [1, x_{10}, x_{20}, \ldots x_{k0}]$ είναι:

$$g^{-1}[x_0' b \pm z_{\alpha/2}\sqrt{x_0'(D'V^{-1}D)^{-1}x_0)}]$$

όπου g: συνάρτηση σύνδεσης .

Υπάρχει όμως και άλλος τρόπος να αναπτύξουμε ένα Wald διάστημα εμπιστοσύνης για τη μέση απόκριση στα GLM. Υπενθυμίζουμε ότι στην θεωρία κανονικής γραμμικής παλινδρόμησης το διάστημα εμπιστοσύνης στο E(y/ x= x_0) για μοντέλο με p παραμέτρους δίνεται ως εξής :

$$y(x_0) \pm t_{\alpha/2,n-p} s \sqrt{x_0'(X'X)^{-1}x_0)} \tag{4.18}$$

Οι Myers και Montgomery (1997) παρουσιάζουν μια ανάλογη έκφραση για τα GLM. Το ασυμπτωτικό 100(1-α) % διάστημα εμπιστοσύνης της μέσης απόκρισης στο σημείο x_0 είναι :

$$\mu(x_0) \pm z_{\alpha/2} \sqrt{d_0'(D'V^{-1}D)^{-1}d_0)} \tag{4.19}$$

όπου D είναι πίνακας με i-οστή σειρά $(\frac{\partial \mu_i}{\partial \beta})'$, V= diag {var($y_i$)},$d_0$ είναι το διάνυσμα των παραγώγων αυτών στο σημείο x_0 και $\widehat{\mu(x_0)}$ η εκτίμηση της μέσης απόκρισης στο σημείο x_0.

Από τους McCullagh και Nelder (1989,p. 28) η μορφή της Πιθανοφάνειας για την εκθετική κατανομή είναι :

$$L(\theta) = \{ (y\theta - b(\theta)) / \alpha(\varphi) + c(y;\varphi) \}$$

με μ = b΄(θ) και var(y) = b΄΄(θ) α(φ) . Για τη σύνδεση της μορφής g (μ) = x΄β έχουμε $\hat{\mu}$ = g^{-1}(x΄b) όπου b είναι η εκτιμήτρια μεγίστης πιθανοφάνειας. Γενικά, το μ είναι μη γραμμική συνάρτηση του β. Ο πίνακας πληροφορίας είναι:

$$I(\beta) = D'V^{-1}D$$

ενώ από την εκθετική κατανομή γνωρίζουμε ότι $\frac{\partial \mu}{\partial \beta} = \frac{\partial \mu}{\partial \theta}\frac{\partial \theta}{\partial x'\beta}$ x η οποία σχέση δίνει ότι $D = V\Delta X / \alpha(\varphi)$.

Έτσι $D'V^{-1}D = \frac{X'\Delta V\Delta X}{[\alpha(\varphi)]2} = \frac{X'WX}{[\alpha(\varphi)]2}$ με πίνακα $\Delta = \frac{\partial \theta_i}{\partial x_{i'\beta}}$.

Για την κανονική σύνδεση θ = x΄β έχουμε I (b) $= \frac{X'VX}{[\alpha(\varphi)]2}$.

Για το διάστημα εμπιστοσύνης του $\mu(x_0)$ πρέπει να χρησιμοποιήσουμε τη μέθοδο Δέλτα για να προσεγγίσουμε τη var $(\widehat{\mu(x_0)})$. Η μέθοδος Δέλτα επιτρέπει την προσέγγιση της διακύμανσης μιας ποσότητας που είναι μη γραμμική συνάρτηση τυχαίων μεταβλητών με άγνωστες διακυμάνσεις . Η $\widehat{\mu(x_0)}$ είναι μη γραμμική συνάρτηση των εκτιμητριών αγνώστων παραμέτρων b έτσι η διακύμανση της είναι:

$$\text{var}\,[\mu(x_0)]=\mathrm{d}_0'\text{var}(\mathrm{b})\ \mathrm{d}_0$$

με $\mathrm{d}_0 = \frac{\partial\widehat{\mu(x_0)}}{\partial b}$ και var($\hat{b}$) ισούται με τον ασυμπτωτικό πίνακα διακύμανσης – συνδιακύμανσης του b που δίνεται ως εξής: $\mathrm{I(b)}^{-1}= (\mathrm{D'V}^{-1}\mathrm{D})^{-1}$. Η ασυμπτωτική κατανομή

$$\frac{\mu(x_0)-\mu(x_0)}{\sqrt{\mathrm{d}_0'(\mathrm{D'V}^{-1}\mathrm{D})^{-1}\ \mathrm{d}_0}}$$

ακολουθεί την κατανομή N(0,1) και έτσι το 100(1-α)% διάστημα εμπιστοσύνης του $\mu(x_0)$ είναι :

$$\mu(x_0)\pm z_{\alpha/2}\sqrt{d_0'(D'V^{-1}D)^{-1}d_0)}$$

Λόγω του κανόνα αλυσίδας αλλά και των σχέσεων της εκθετικής κατανομής έχουμε:

$$\frac{\partial\mu}{\partial\beta}=\frac{\partial\mu}{\partial\theta}\frac{\partial\theta}{\partial(x'\beta)}\frac{\partial(x'\beta)}{\partial\beta}=\frac{\text{var(y)}\Delta\text{x}}{a(\varphi)}$$

Για την κανονική σύνδεση δ=1 και $\mathrm{D}=\frac{VX}{a(\varphi)}$ $(\mathrm{D'V}^{-1}\mathrm{D})^{-1}=(\mathrm{X'VX})^{-1}[a(\varphi)]^2$,

$\mathrm{d}_0=\frac{\text{var(y)}\mathrm{x}_0}{a(\varphi)}$. Έτσι, έχουμε το διάστημα εμπιστοσύνης :

$$\mu(x_0)\pm z_{\alpha/2}\text{var}(y_0)\sqrt{x_0'(D'V^{-1}D)^{-1}x_0)}$$

Στα GLM σημαντικό ρόλο παίζει η επιλογή της συνάρτησης σύνδεσης. Στις περισσότερες περιπτώσεις επιλέγεται η κανονική σύνδεση πιθανότατα για ευκολία στην επεξήγηση των αποτελεσμάτων. Ωστόσο, σε άλλες περιπτώσεις μπορεί να μην είναι τόσο εμφανής η επιλογή της κατάλληλης συνάρτησης σύνδεσης. Για το λόγο

αυτό πρέπει να αναλογιστούμε την επίδραση της επιλογής λανθασμένης συνάρτησης σύνδεσης στα διαστήματα εμπιστοσύνης αλλά και την ακρίβεια.

Ένα παράδειγμα τέτοιου τύπου είναι ένα πείραμα όπου η απόκριση είναι δυωνυμική μεταβλητή. Θα δούμε την επίδραση της κανονικής σύνδεσης αλλά και της λογιστικής σύνδεσης. Για να προσομοιάσουμε την κατάσταση αυτή επιλέγουμε ένα 2^2 παραγοντικό σχεδιασμό. Για το κανονικό μοντέλο Lewis, Montgomery, Myers (2001b) επιλέγουν $\mu_i = g^{-1}(x_i'\beta) = g^{-1}(10 + 5x_{i1} + 3x_{i2})$. Αντί της εφαρμογής της σωστής συνάρτησης σύνδεσης χρησιμοποιείται το μοντέλο με τη λογιστική σύνδεση.

Το πείραμα διεξήχθη 5000 φορές για n = 8,16,32 και τα αποτελέσματα του φαίνονται στον παρακάτω πίνακα.

Επαναλήψεις	***Σωστή συνάρτηση σύνδεσης***	***Λάθος συνάρτηση σύνδεσης***
8	95.48%	81.24%
16	93.98%	69.54%
32	94.51%	53.58%

Η χρήση της λάθος συνάρτησης σύνδεσης οδηγεί σε ελάττωση κάλυψης αν και ο αριθμός των επαναλήψεων αυξάνεται δηλ. καθώς αυξάνονται οι επαναλήψεις τα διαστήματα εμπιστοσύνης γίνονται μικρότερα και περιορίζονται γύρω από τη λάθος μέση τιμή.

Γενικότερα, τα διαστήματα εμπιστοσύνης παρέχουν μια πιο αποτελεσματική μέθοδο για τον υπολογισμό της μέσης απόκρισης συγκεκριμένου μοντέλου αλλά διευκολύνουν και τις συγκρίσεις μεταξύ των διαφορετικών μοντέλων. Μεγάλη σημασία στην προσαρμογή του μοντέλου παίζει η επιλογή της συνάρτησης σύνδεσης. Αναφέραμε ήδη ότι η επιλογή λανθασμένης συνάρτησης σύνδεσης οδηγεί σε ελάττωση κάλυψης.

Επίσης, στην πραγματικότητα τις περισσότερες φορές επιλέγεται η λάθος συνάρτηση σύνδεσης. Παρόλα αυτά, τα αποτελέσματα της ανάλυσης δίνουν ένα ικανοποιητικό μοντέλο υποδηλώνοντας ότι η κάλυψη είναι κοντά στην πραγματική

άσχετα με το αν το μοντέλο είναι το σωστό ή όχι π.χ. η ανάλυση της γάμμα κατανομής απόκρισης συχνά ανέχεται την κακώς προσδιοριζόμενη συνάρτηση σύνδεσης.

4.7 GLM, Πειράματα Κρησαρίσματος και Μετασχηματισμοί Δεδομένων

Οι 2^k παραγοντικοί και 2^{k-p} κλασματικοί παραγοντικοί σχεδιασμοί χρησιμοποιούνται ευρέως ως πειράματα κρησαρίσματος (screening experiments) για την εύρεση της ομάδας των παραγόντων με τη μεγαλύτερη επίδραση στην απόκριση.

Μια τυπική προσέγγιση των 2^k σχεδιασμών σε ένα πείραμα κρησαρίσματος είναι ο σχεδιασμός των εκτιμητριών των επιδράσεων (ή των εκτιμητριών των συντελεστών του μοντέλου b) σε ένα γραφικό έλεγχο της υπόθεσης της κανονικότητας των υπολοίπων (normal probability plot).

Η προσέγγιση αυτή είναι αρκετά αποδοτική αφού οι εκτιμήτριες έχουν ίσες διακυμάνσεις και είναι ασυσχέτιστες. Στα GLM, οι συντελεστές του μοντέλου δεν είναι γενικά ασυσχέτιστοι και δεν έχουν ίσες διακυμάνσεις. Παρόλα αυτά ένας γραφικός έλεγχος της υπόθεσης της κανονικότητας των υπολοίπων διαιρούμενων με τα τυπικά σφάλματα συχνά προσφέρει χρήσιμη καθοδήγηση για την επιλογή των ενεργών παραγόντων εκτός αν οι συσχετίσεις των εκτιμητριών είναι πολύ μεγάλες.

Όπως είδαμε πριν υπάρχουν μετασχηματισμοί που εφαρμόζονται στην απόκριση y με σκοπό την σταθεροποίηση της διακύμανσης .Σε άλλες περιπτώσεις χρησιμοποιείται μετασχηματισμός όταν τα τυχαία σφάλματα δεν ακολουθούν την κανονική κατανομή ή όταν η διακύμανση είναι συνάρτηση του μέσου. Ωστόσο, δεν πρέπει να συγχέονται οι μετασχηματισμοί με τις συναρτήσεις σύνδεσης αφού οι πρώτοι αφορούν μετασχηματισμούς των δεδομένων y ενώ οι δεύτεροι μετασχηματισμοί του μέσου του πληθυσμού μ. Πρακτικά, πιο χρήσιμοι είναι οι μετασχηματισμοί των δεδομένων.

Τα πλεονεκτήματα και η υπεροχή των GLM έναντι των μετασχηματισμών δεδομένων συνοψίζονται ως εξής:

- Πρώτον, αν η απόκριση είναι μη κανονική μπορεί να είναι αδύνατο για τον ίδιο μετασχηματισμό να δημιουργήσει τυχαία σφάλματα με κανονική κατανομή, να σταθεροποιήσει την διακύμανση και να οδηγήσει σε γραμμικό

μοντέλο. Για το λόγο αυτό τα GLM εκμεταλλεύονται την κατανομή των δεδομένων.

- Δεύτερον, η σταθερότητα της διακύμανσης δεν αποτελεί θέμα για τα GLM καθώς βασίζουν την ανάλυση τους στη φυσική διακύμανση της κατανομής των δεδομένων .
- Τρίτον, η επιλογή συνάρτησης σύνδεσης παρέχει στον αναλυτή μεγάλη ευελιξία στα μη γραμμικά μοντέλα που χρησιμοποιούνται για την προσαρμογή των δεδομένων.
- Τέλος, με τα GLM ο ερευνητής δε χάνει κανένα από τα βασικά στοιχεία της ανάλυσης δεδομένων γραμμικών μοντέλων.

4.8 Μοντελοποίηση Μέσου και Διακύμανσης μέσω GLM

Σημαντικό πρόβλημα στην βιομηχανία αποτελεί η μοντελοποίηση του μέσου και της διακύμανσης μιας διαδικασίας . Η επίλυση αυτού του προβλήματος έγκειται στη χρήση GLM με δυο προσεγγίσεις :

- όταν υπάρχουν επαναλήψεις
- όταν δεν υπάρχουν επαναλήψεις , όπου χρησιμοποιούμε τα υπόλοιπα ως βάση για τη μοντελοποίηση της διακύμανσης.

Η ύπαρξη επαναλήψεων μας επιτρέπει να δημιουργήσουμε πληροφορίες για τη διακύμανση ανεξάρτητα από τη δομή του μοντέλου αφού μπορούμε να δημιουργήσουμε διακυμάνσεις s_i^2 στα σημεία επανάληψης. Η ανάλυση βασίζεται στην κατανομή των δεδομένων. Παρόλα αυτά, αν τα δεδομένα ακολουθούν την κανονική κατανομή ο μέσος και η διακύμανση του δείγματος στα σημεία επανάληψης είναι ανεξάρτητα μεταξύ τους.

Η διαδικασία για την κατασκευή μοντέλου για το μέσο είναι η εξής :

- αρχικά χρησιμοποιούμε τα GLM για να δημιουργήσουμε ένα μοντέλο για τις διακυμάνσεις
- έπειτα, μπορούμε να χρησιμοποιήσουμε το μοντέλο αυτό για να δημιουργήσουμε τις κατάλληλες στάθμες για να κατασκευάσουμε τη γενικευμένη μέθοδο ελαχίστων τετραγώνων

- τέλος, από τη μέθοδο ελαχίστων τετραγώνων δημιουργούμε το κατάλληλο μοντέλο για το μέσο.

Στην αντίθετη περίπτωση, υπάρχουν δυσκολίες για την κατασκευή μοντέλων διακύμανσης. Η δυσκολία αυτή έγκειται στο γεγονός ότι το μοντέλο της διακύμανσης είναι πολύ εξαρτώμενο από την επιλογή του κατάλληλου μοντέλου για το μέσο μιας και η πηγή της μεταβλητότητας είναι τα υπόλοιπα.

Για το βασικό μοντέλο διακύμανσης θα ασχοληθούμε με τη γραμμική λογιστική δομή. Για πιθανή χρήση των GLM πρέπει να λάβουμε υπόψη το μοντέλο σφάλματος

$\varepsilon_i = y_i - x_i'\beta$ και σαν αρχικό σημείο υποθέστε ότι

$$\sigma_i^2 = \mathrm{E}(\varepsilon_i)^2 = \exp(x_i'\gamma)$$

όπου $(\varepsilon_i)^2 \sim \sigma_i^2 \chi_i^2$.

Έτσι, στην περίπτωση έλλειψης επαναλήψεων το πρόβλημα εστιάζεται στην ταυτόχρονη αποτελεσματικότητα των εκτιμήσεων των συντελεστών β του μοντέλου μέσου αλλά και των συντελεστών γ του μοντέλου διακύμανσης μέσω Μέγιστης Πιθανοφάνειας. Το μοντέλο του μέσου μπορεί να γραφτεί ως $y = X\beta + \varepsilon$ με $\mathrm{var}(\varepsilon) = V_{n\,x\,n}$ και $v_i = \exp(u_i'\gamma)$ με u_i το i-οστό σετ παλινδρομήσεων που χρησιμοποιούνται στο μοντέλο διακύμανσης.Το μοντέλο της Μέγιστης Πιθανοφάνειας για το β είναι : $b = (X'V^{-1}X)^{-1}X'V^{-1}y$.

Αν κάνουμε χρήση του τυχαίου διανύσματος $s' = (\varepsilon_{1,}^2\, \varepsilon_{2,}^2 \ldots\ldots, \varepsilon_{n,}^2)$ έχουμε μια σειρά από τις ανεξάρτητες χ_1^2 τυχαίες μεταβλητές που ακολουθούν την Γάμμα κατανομή με 2 παραμέτρους. Η εκτιμήτρια Μεγίστης Πιθανοφάνειας για το β περιέχει το γ μέσω του πίνακα V και αντίστοιχα η εκτιμήτρια Μεγίστης Πιθανοφάνειας για το γ περιέχει το β μέσω των $\varepsilon_{i.}$

Για το λόγο αυτό απαιτείται μια επαναληπτική διαδικασία (Aitkin 1987) η οποία είναι η εξής:

- χρησιμοποιούμε την μέθοδο ελαχίστων τετραγώνων για να αποκτήσουμε το b_0 για το μοντέλο μέσου $y_i = x_i'\beta + \varepsilon_i$

- χρησιμοποιούμε το b_0 για να υπολογίσουμε n υπόλοιπα $e_i = y_i - x_i^{'}\beta_0$ για i = 1,2,..
- χρησιμοποιούμε τα e_i^2 ως δεδομένα για να προσαρμόσουμε τη διακύμανση του μοντέλου με συντελεστές παλινδρομήσεις u και τη λογιστική σύνδεση με 2 παραμέτρους.
- χρησιμοποιούμε τις εκτιμήσεις των παραμέτρων $\widehat{\gamma_i}$ για να κατασκευάσουμε τον πίνακα V
- χρησιμοποιούμε τον V με τα σταθμισμένα ελάχιστα τετράγωνα για να update το b_0 σε b_1.
- πηγαίνουμε πίσω στο βήμα 2 και αντικαθιστούμε το b_1 με το b_0.
- συνεχίζουμε τις μετατροπές.

Ο αλγόριθμος αυτός δεν είναι δύσκολος να κατασκευαστεί και είναι πολύ εύχρηστος για τον ταυτόχρονο υπολογισμό του μοντέλου του μέσου και της διακύμανσης. Το μόνο μειονέκτημα της μεθόδου αυτής είναι ότι η διακύμανση του μοντέλου θα είναι μεροληπτική αφού η εκτιμήτρια Μεγίστης Πιθανοφάνειας για το γ δεν μετράει για την εκτιμήτρια του β. Γι' αυτό χρησιμοποιούμε την **Περιορισμένη Μέγιστη Πιθανοφάνεια (REMIL).**

4.9 Δυαδικής Απόκρισης Μοντέλα

Οι Μπεϋζιανοί σχεδιασμοί εκτός των κανονικών γραμμικών μοντέλων είναι πολύ χρήσιμοι σε δυαδικής απόκρισης μοντέλα. Οι Tsutakawa (1972, 1980), Owen (1975) και Zachs (1977) έχουν θεωρήσει βέλτιστους σχεδιασμούς για δυαδικής απόκρισης μοντέλα υπό την Μπεϋζιανή οπτική.

Πολλά από αυτά τα μοντέλα περιορίζονται σε ίσα κατανεμημένα σημεία με ίσες στάθμες σε κάθε σημείο. Οι βέλτιστοι σχεδιασμοί αποκτώνται από την εφαρμογή του αλγορίθμου Simplex των Nelder και Mead(1965) οι οποίοι υποθέτουν ομοιόμορφες εκ των προτέρων κατανομές των παραμέτρων και αξιολογούν την προσδοκία των εκ των προτέρων κατανομών μέσω αριθμητικών μεθόδων. Μεγάλο ενδιαφέρον στα δυαδικής απόκρισης μοντέλα υπάρχει λόγω της χρήσης τους σε βιο-δοκιμασίες δόσης-απόκρισης .

Οι Zhu, Ahn, Wong (1998) και Zhu,Wong (2001) μελετούν τους βέλτιστους σχεδιασμούς για τον υπολογισμό διαφόρων ποσοστών του λογιστικού μοντέλου με διαφορετικές στάθμες το καθένα. Πιο συγκεκριμένα, χρησιμοποιούν ένα αντικειμενικό κριτήριο το οποίο είναι συνδυασμός ξεχωριστών αντικειμενικών κριτηρίων αλλάζοντας το logit σχεδιασμό των Chaloner , Larntz (1989) για να βελτιώσουν το κριτήριο αυτό.

Οι Zhu and Wong (2001) ισχυρίζονται ότι οι διαδοχικοί σχεδιασμοί είναι συγκρίσιμοι με τα Μπεϋζιανά μοντέλα. Όταν συγκρίνονται με τους τοπικούς βέλτιστους σχεδιασμούς , όπως είναι αναμενόμενο, οι Μπεϋζιανοί σχεδιασμοί είναι καλύτεροι από τους τοπικούς βέλτιστους σχεδιασμούς αν συγκεκριμένες παράμετροι είναι μακριά από τις πραγματικές παραμέτρους.

Διάφορες κλινικές μελέτες που χρησιμοποιούν τις Μπεϋζιανές ιδέες εκτός από τις μελέτες απόκρισης – δόσης παρουσιάζονται από τους Berry and Fristedt (1985), Berry and Pearson (1985), Parmigianni (1993), Parmigianni and Berry (1994).

ΚΕΦΑΛΑΙΟ 5:
ΣΧΕΔΙΑΣΜΟΙ ΜΕ ΔΥΑΔΙΚΗ ΑΠΟΚΡΙΣΗ & ΚΡΙΤΗΡΙΑ ΕΠΙΛΟΓΗΣ ΜΟΝΤΕΛΩΝ ΓΙΑ ΔΙΩΝΥΜΙΚΕΣ ΑΠΟΚΡΙΣΕΙΣ

5.1 Βέλτιστοι Σχεδιασμοί 2^k Παραγοντικών Σχεδιασμών για Δυαδικές Αποκρίσεις

Αντικείμενο του πρώτου μέρους του κεφαλαίου αυτού είναι οι τοπικοί D-βέλτιστοι σχεδιασμοί για παραγοντικούς σχεδιασμούς με 2 παράγοντες με δυαδική απόκριση.

Θα παρουσιαστούν προβλήματα 2^2 παραγοντικού σχεδιασμού με τους βέλτιστους σχεδιασμούς τους αλλά και η **Κυλινδρική Αλγεβρική Αποσύνθεση (Cylindrical Algebraic Decomposition,CAD).**

Σημαντικό ρόλο παίζει η επιλογή αρχικών συνθηκών .Αν δεν υπάρχει κάποια βάση για την επιλογή τους συνήθως ακολουθούμε τους ομοιόμορφους σχεδιασμούς π.χ. ίσος αριθμός παρατηρήσεων σε κάθε τέσσερα σημεία.

Σκοπός κάθε μελέτης είναι σχεδιασμών με δυαδική απόκριση είναι οι ποσοτικοί παράγοντες βάσει βέλτιστων σχεδιασμών για γενικευμένα γραμμικά μοντέλα. Αν η απόκριση είναι ποσοτική και οι παράγοντες είναι δυαδικοί χρησιμοποιούμε τη θεωρία παραγοντικών σχεδιασμών.

Παραδείγματα παραγοντικών σχεδιασμών με δυαδικές αποκρίσεις είναι τα εξής:

- χρήση ορρού κατά του πνευμονόκοκκου όπου η επεξηγηματική μεταβλητή είναι οι δόσεις του ορρού και η απόκριση είναι η επιβίωση ή μη πέρα των επτά ημερών.
- πειράματα επιβίωσης σπερματοζωαρίων (Myers, Montgomery and Vinning (2002).
- πειράματα επιβίωσης σπόρων (Crowder 1978).

Οι υποθέσεις που κάνουμε για να προχωρήσουμε στην κατασκευή βέλτιστου σχεδιασμού είναι οι εξής:

- η χρήση γενικευμένων γραμμικών μοντέλων με οποιαδήποτε συνάρτηση σύνδεσης είναι κατάλληλη για τη μελέτη μας. Συνήθως χρησιμοποιούνται οι συναρτήσεις σύνδεσης: logit, probit, log-log και η συμπληρωματική log-log.
- ο βέλτιστος σχεδιασμός θα αποκτηθεί μέσω του D- κριτηρίου το οποίο μεγιστοποιεί την ορίζουσα του πίνακα πληροφορίας.
- χρησιμοποιούμε την τοπική προσέγγιση βελτιστοποίησης του Chernoff (1953) για να ξεπεράσουμε το πρόβλημα εξάρτησης των παραμέτρων.
- κάθε παράγοντας έχει δυο επίπεδα και αυτό είναι σημαντικό από πλευράς τεστ κρησαρίσματος (screening tests) αφού μας ενδιαφέρουν οι πλήρεις 2^k σχεδιασμοί δηλ. αυτοί που υποστηρίζονται από 2^k σημεία.
- το μοντέλο μπορεί να περιέχει κύριες επιδράσεις και αλληλεπιδράσεις.

Οι δυσκολίες της μεθοδολογίας αυτής είναι οι εξής :

- ο σταθερός αριθμός συνολικών παρατηρήσεων δημιουργεί πρόβλημα στον καθορισμό της αναλογίας των παρατηρήσεων που κατανέμονται σε κάθε ένα από τα 2^k σημεία.
- αν η απόκριση είναι απλό γραμμικό μοντέλο τότε το μοντέλο είναι ομοιόμορφο στα 2^k σημεία και καθολικά βέλτιστο.
- η βελτιστοποίηση του ομοιόμορφου σχεδιασμού μέσω της ελαχιστοποίησης των διακυμάνσεων κάθε εκτιμήτριας παραμέτρου (Rao 1971).

5.1.1 2^2 Παραγοντικός Σχεδιασμός

Αν και υπάρχει δυσκολία στον υπολογισμό των αναλυτικών λύσεων για τους γενικούς 2^2 παραγοντικούς υπάρχουν χαρακτηρισμοί για μερικές συγκεκριμένες περιπτώσεις. Για να υπερνικήσουμε το πρόβλημα των τοπικών D-βέλτιστων σχεδιασμών από τις αρχικές τιμές των παραμέτρων εκτελούμε συνεχείς προσομοιώσεις θα αποδείξουμε ότι ο σχεδιασμός συνήθως αυτοδύναμος όσον αφορά την επιλογή αυτών των τιμών.

Ποια η σχέση μεταξύ κριτηρίων βελτιστοποίησης και διακυμάνσεων ?

Τα κριτήρια βελτιστοποίησης μπορούν να γραφτούν υπό τη μορφή διακυμάνσεων που εξαρτώνται από τις παραμέτρους μέσω της συνάρτησης σύνδεσης ή πληροφοριών. Ωστόσο, πρέπει να επισημάνουμε ότι ο D-βέλτιστος σχεδιασμός είναι αρκετά διαφορετικός από τον ομοιόμορφο σχεδιασμό ιδιαίτερα αν τουλάχιστον δυο διακυμάνσεις διαφέρουν αρκετά μεταξύ τους. Γενικά, ισχύει ο εξής κανόνας:

- αν ο ερευνητής έχει περίπου ιδέα των διακυμάνσεων τότε το μοντέλο που χρησιμοποιεί αυτές τις τιμές στον τοπικό D-βέλτιστο σχεδιασμό είναι πολύ ικανοποιητικός σχεδιασμός.
- αν η διακύμανση σε ένα σημείο είναι αρκετά μεγαλύτερη από τις άλλες τότε κατάλληλος είναι ο ομοιόμορφος σχεδιασμός. Αυτή η επιλογή ακολουθείται στις περιστάσεις όπου χρησιμοποιούμε ως συναρτήσεις σύνδεσης τις logit, probit, log-log και η συμπληρωματική log-log .

Υπόδειξη :

Αν και ο D-βέλτιστος σχεδιασμός μπορεί να μην χρησιμοποιείται για συγκεκριμένες εφαρμογές μπορεί να θεωρηθεί ως σημείο αναφοράς για άλλους σχεδιασμούς.

Στη συνέχεια, παραθέτουμε κάποιες προκαταρτικές γνώσεις.

Έστω 2^k πείραμα με k επεξηγηματικές μεταβλητές με 2 επίπεδα η κάθε μια και n_i παρατηρήσεις κατανεμημένες στην i-οστή πειραματική κατάσταση τέτοιες ώστε $n_i \geq 0$ για i = 1,2,3, ….,2^k και $n_1 + n_2 + \ldots\ldots\ldots\ldots + n_{2^k} = n$

Για καθορισμένο n έχουμε το πρόβλημα του προσδιορισμού των βέλτιστων n_i αλλά συνήθως χρησιμοποιείται η μορφή των αναλογιών :

$$p_i = \frac{n_i}{n} \qquad i = 1,2,3, \ldots,2^k$$

και το πρόβλημα επικεντρώνεται στην εύρεση των βέλτιστων p_i.

Έστω η γραμμική παράμετρος πρόβλεψης η που περιλαμβάνει τις κύριες επιδράσεις και τις αλληλεπιδράσεις του υποτιθέμενου μοντέλου π.χ. σε έναν 2^2 παραγοντικό σχεδιασμό με κύριες επιδράσεις έχουμε : $\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ με $\beta = (\beta_0, \beta_1, \beta_2)'$. Υπενθυμίζουμε ότι για την απόκριση y ισχύουν :

$$E(y) = \mu$$
$$\eta = g(\mu)$$

όπου g η συνάρτηση σύνδεσης. Για την εκτιμήτρια Μεγίστης Πιθανοφάνειας ο ασυμπτωτικός πίνακας διακύμανσης είναι ο αντίστροφος του n X΄W X με

$$W= \text{diag}(w_1 p_1 + w_2 p_2 + ... + w_{2^k} p_{2^k}) , \; w_i = \frac{(\frac{\partial \mu_i}{\partial \eta_i})^2}{\mu_i(1-\mu_i)} \text{ και}$$

$$X= \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & -1 \\ 1 & -1 & 1 \\ 1 & -1 & -1 \end{pmatrix}$$

Ο D-βέλτιστος σχεδιασμός μεγιστοποιεί την :

$$|X'WX|$$

ή όπως αλλιώς μπορεί να γραφτεί η παραπάνω ορίζουσα δηλ. με την εξής μορφή :

$$\det(w, p) = G(p) = p_2 w_2 p_3 w_3 p_4 w_4 + p_1 w_1 p_3 w_3 p_4 w_4 + p_1 w_1 p_2 w_2 p_4 w_4 + p_1 w_1 p_2 w_2 p_3 w_3$$

με $w = (w_1, w_2, w_3, w_4)'$ και $p = (p_1, p_2, p_3, p_4)'$.

Επίσης ισχύουν οι σχέσεις $p_i \geq 0$ και $\sum_i p_i = 1$.

5.1.2 Αναλυτικές λύσεις συγκεκριμένων περιπτώσεων

Διακρίνουμε τις εξής περιπτώσεις :

- αν τα w_i είναι ίσα τότε το ομοιόμορφο μοντέλο ($p_1=p_2= p_3= p_4= 1/4$) είναι D-βέλτιστος σχεδιασμός (Kiefer 1975).
- αν ένα και μόνο ένα από τα w_i είναι μηδέν τότε ο βέλτιστος σχεδιασμός είναι ο ομοιόμορφος στα σημεία όπου τα w_i διάφορα του μηδενός.
- αν δυο ή περισσότερα w_i είναι μηδενικά τότε G (p) = 0

Ορίζουμε ότι :

$$L = G(p) \; / w_1 w_2 w_3 w_4 , \; v_i = 1/ w_i \; i=1,2,3,4$$

και αντί της μεγιστοποίησης της G (p) έχουμε την μεγιστοποίηση της :

$$L(p) = \nu_4 p_1p_2p_3 + \nu_3 p_1p_2p_4 + \nu_2 p_1p_4p_3 + \nu_1 p_4p_2p_3 \qquad (5.1)$$

<u>Θεώρημα 1</u>:

Η L (p) έχει ένα μοναδικό μέγιστο το p=(0,1/3,1/3,1/3) αν και μόνο αν $\nu_1 \geq \nu_2 + \nu_3 + \nu_4$.

<u>Υποσημείωση :</u>

Το παραπάνω θεώρημα δεν ανταποκρίνεται σε πλήρεις 2^2 σχεδιασμούς αλλά μόνο σε σχεδιασμούς με τρία σημεία ο οποίος είναι κορεσμένος για το μοντέλο κύριων επιδράσεων $\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2$.

Η πλήρης απόδειξη του Θεωρήματος 1 βρίσκεται στο παράρτημα.

<u>Λήμμα 1 :</u>

Αν $\nu_1 > \nu_2$ τότε οποιαδήποτε λύση της (5.1) πρέπει να ικανοποιεί την $p_1 \leq p_2$, αν $\nu_1 = \nu_2$ τότε οποιαδήποτε λύση της (5.1) πρέπει να ικανοποιεί την $p_1 = p_2$.

Η πλήρης απόδειξη του Λήμματος 1 βρίσκεται στο παράρτημα.

<u>Θεώρημα 2</u>:

Αν $\nu_1 \geq \nu_2$, $\nu_3 = \nu_4 = \nu$ και $\nu_1 < \nu_2 + 2\nu$ τότε η λύση που μεγιστοποιεί την L(p) είναι :

$$p_1 = \frac{1}{2} - \frac{\nu_1 - \nu_2 + 4\nu}{2\,(-2\delta + D)}$$

$$p_2 = \frac{1}{2} + \frac{\nu_1 - \nu_2 - 4\nu}{2\,(-2\delta + D)}$$

$$p_3 = p_4 = \frac{2\nu}{(-2\delta + D)}$$

με $L = 2\nu^2(\delta^2 + 4\nu_1\nu_2 - \delta D)/(-2\delta + D)^3$, $\delta = \nu_1 + \nu_2 - 4\nu$ και $D = \sqrt{12\nu_1\nu_2 + \delta^2}$.

Η πλήρης απόδειξη του Θεωρήματος 2 βρίσκεται στο παράρτημα.

Πόρισμα 1:

Αν $\nu_3 = \nu_4 = \nu_2 = \nu$ και $\nu_1 < 3\nu$ τότε η λύση που μεγιστοποιεί την L(p) είναι :

$$p_1 = \frac{3\nu - \nu_1}{9\nu - \nu_1}$$

$$p_2 = p_3 = p_4 = \frac{2\nu}{9\nu - \nu_1}$$

με μέγιστο $L = 4\nu^3 / (9\nu - \nu_1)^2$.

Πόρισμα 2:

Αν $\nu_1 = \nu_2 = \nu$, $\nu_3 = \nu_4 = c$ και $\nu > c$ τότε η λύση που μεγιστοποιεί την L(p) είναι :

$$p_1 = p_2 = \frac{2\nu - c - d}{6(\nu - d)}$$

$$p_3 = p_4 = \frac{\nu - 2c + d}{6(\nu - d)}$$

με μέγιστο $L = \frac{(2\nu - c - d)(\nu - 2c + d)(\nu + c + d)}{108(\nu - c)^2}$ όπου $d = \sqrt{\nu^2 - \nu c + c^2}$

Το Θεώρημα 1 αποκαλύπτει ότι ο D- βέλτιστος σχεδιασμός είναι κορεσμένος αν και μόνο αν : $\frac{2}{\min\{w_1, w_2, w_3, w_4\}} \geq \frac{1}{w_1} + \frac{1}{w_2} + \frac{1}{w_3} + \frac{1}{w_4}$

Η παραπάνω ανισότητα είναι γνωστή ως **συνθήκη κορεσμού**.

Θεώρημα 3:

Για τη συνάρτηση σύνδεσης logit η συνθήκη κορεσμού ισχύει αν και μόνο αν:

$$\beta_0 \neq 0$$

$$|\beta_1| > \frac{1}{2}\log\left(\frac{e^{2|\beta_0|}+1}{e^{2|\beta_0|}-1}\right)$$

$$|\beta_2| \geq \log\left(\frac{\sqrt{(e^{4|\beta_0|}-1)(e^{4|\beta_1|}-1)}+2e^{|\beta_0|+|\beta_1|}}{(e^{2|\beta_0|}-1)(e^{2|\beta_1|}-1)-2}\right)$$

Για τη συνάρτηση σύνδεσης logit τα β_0 , β_1, β_2 είναι συμμετρικά σε σχέση με τα w_1 , w_2, w_3 , w_4. Αν μεταθέσουμε τα β_0 , β_1, β_2 τότε πρέπει να αλλάξουμε τη σειρά των w_1 , w_2, w_3 αντίστοιχα αφού δεν επηρεάζεται από το αν ικανοποιείται ή όχι η συνθήκη κορεσμού.

Η πλήρης απόδειξη του Θεωρήματος 3 βρίσκεται στο παράρτημα.

5.1.3 Ακριβείς λύσεις χρησιμοποιώντας την μέθοδο CAD

Η μέθοδος της κυλινδρικής αλγεβρικής αποσύνθεσης χρησιμοποιείται σε γενικές περιπτώσεις προβλημάτων βελτιστοποίησης για την εύρεση ακριβών λύσεων και περιγράφεται πλήρως από τους Fotiou et al (2005). Έτσι η μεγιστοποίηση της L(p) μπορεί να γραφτεί μέσω της Boolean έκφρασης :

$$\left(L_3 - f \geq 0\right) \Lambda \left(p_1 \geq 0\right) \Lambda \left(p_2 \geq 0\right) \Lambda \left(p_3 \geq 0\right) \Lambda \left(p_1 + p_2 + p_3 \leq 1\right)$$

όπου $L_3 = L\ (p_1, p_2\ ,\ p_3, 1- p_1- p_2 - p_3)$ και f παράμετρος που επιδεικνύει την αξία της αντικειμενικής συνάρτησης.

5.1.4 Αναλυτικές κατά προσέγγιση λύσεις

Θα μελετήσουμε τις αναλυτικές κατά προσέγγιση λύσεις αλλά και ένα άνω φράγμα για το σφάλμα της προσέγγισης αυτής. Για λόγους απλοποίησης γράφουμε :

$L\ (\nu_1, \nu_2\ ,\ \nu_3) = \max_p L$

$$= \max_{x,y,z}\{\ 2\nu xyz + \nu_2 xz^2 + \nu_1 yz^2\ \}$$

με τους εξής περιορισμούς:

$$x \geq 0 \ , \ y \geq 0, \ z \geq 0$$

$$x + y + 2z = 1$$

Χωρίς βλάβη γενικότητας, υποθέτουμε ότι: $v_1 < v_2 < v_3 < v_4$ και $v_4 < v_1 + v_2 + v_3$ και ορίζουμε :

$$L_{34} = L[v_1, v_2 , \ (v_3 + v_4) \ / 2 \ , \ (v_3 + v_4) \ / 2 \ ].$$

Ομοίως ορίζονται οι L_{12} , $L_{13,}$ L_{14} , L_{23} , L_{24} και χρησιμοποιούμε το :

$$\max \{ L_{12} , L_{13,} L_{14} , L_{23} , L_{24}, L_{34} \}$$

για να προσεγγίσουμε την $\max_p$ L χρησιμοποιώντας το παρακάτω Θεώρημα.

<u>Θεώρημα 4:</u>

Υποθέτοντας ότι $v_1 < v_2 < v_3 < v_4$ και $v_4 < v_1 + v_2 + v_3$ τότε :

$$\max \{ L_{13,} L_{14}, L_{24} \} \ \leq \max \{ L_{12,} L_{23}, L_{34} \}$$

$$\max_p L - \max \{ L_{12,} L_{23}, L_{34} \} \ \leq \min \{ \frac{v_2 - v_1}{216} \ , \frac{v_3 - v_2}{96\sqrt{3}} \ , \frac{v_4 - v_3}{54} \ \}$$

Η πλήρης απόδειξη του Θεωρήματος 4 βρίσκεται στο παράρτημα.

5.1.5 Ευρωστία 2^2 Σχεδιασμών

Μιας και οι τοπικοί βέλτιστοι σχεδιασμοί βασίζονται σε υποθετικές τιμές των παραμέτρων είναι βασικό να μελετήσουμε τον μη επηρεασμό των σχεδιασμών από τις τιμές αυτές. Αν δεν υπάρχει κάποια βάση για την επιλογή των αρχικών τιμών τότε επιλέγουμε τον ομοιόμορφο σχεδιασμό. Επίσης, θα μελετήσουμε τον μη επηρεασμό του βέλτιστου σχεδιασμού από την εσφαλμένη επιλογή συνάρτησης σύνδεσης αλλά και από την επιλογή του ομοιόμορφου σχεδιασμού.

Για τον μη επηρεασμό του βέλτιστου σχεδιασμού από την εσφαλμένη επιλογή συνάρτησης σύνδεσης θα μελετήσουμε το παράδειγμα της logit συνάρτησης σύνδεσης.

Η μεθοδολογία που θα ακολουθήσουμε είναι η εξής :

- επιλέγουμε 1000 διανύσματα $w_i = (w_{i1}, w_{i2}, w_{i3}, w_{i4})$ i=1,2,3,….,1000 όπου $0.05 \le w_{ij} \le 0.25$
- $w_t = (w_{t1}, w_{t2}, w_{t3}, w_{t4})$ i= 1,2,…,1000 είναι η αληθινή w
- $w_c = (w_{c1}, w_{c2}, w_{c3}, w_{c4})$ i= 1,2,…,1000 είναι η υποτιθέμενη w
- χρησιμοποιούμε την CAD για να καθορίσουμε τους βέλτιστους σχεδιασμούς $p_t = (p_{t1}, p_{t2}, p_{t3}, p_{t4})$ και $p_c = (p_{c1}, p_{c2}, p_{c3}, p_{c4})$ σε αντιστοιχία με τις w_t, w_c
- η σχετική απώλεια αποδοτικότητας ορίζεται ως :

$$R(t,c) = \frac{\det(w_t, p_t)^{1/3} - \det(w_c, p_c)^{1/3}}{\det(w_t, p_t)^{1/3}}$$

- η μέγιστη σχετική απώλεια αποδοτικότητας ορίζεται ως :

$$R_{max}(c) = \max_t \{R(t,c)\}$$

η οποία δηλώνει την άσχημη απόδοση του σχεδιασμού λόγω του λάνθασμένου w.

- Για τον μη επηρεασμό του βέλτιστου σχεδιασμού από την επιλογή του ομοιόμορφου σχεδιασμού η σχετική απώλεια αποδοτικότητας ορίζεται ως :

$$R_u(w) = \frac{\det(w_t, p_t)^{1/3} - \det(w_c, p_c)^{1/3}}{\det(w_t, p_t)^{1/3}}$$

$$= 1 - \frac{1}{4}\left(\frac{v_1 + v_2 + v_3 + v_4}{L(p_t)}\right)^{1/3}$$

η μέγιστη σχετική απώλεια αποδοτικότητας ορίζεται ως :

$$R_{max}(u) = \max_w \{R_u(w)\}$$

Θεώρημα 5 :

Χωρίς βλάβη γενικότητας, υποθέτουμε ότι $v_1 \geq v_2 \geq v_3 \geq v_4$ τότε :

$$R_{max} = \begin{cases} 1 - \frac{3\,(1+\frac{3\alpha}{b})^3}{4} & \alpha\nu\ \nu_1 \geq \nu_2 + \nu_3 + \nu_4 \\ 1 - \frac{3\,2^{1/3}}{4} & \alpha\nu\ \nu_1 < \nu_2 + \nu_3 + \nu_4 \end{cases}$$

Παράδειγμα :

Στον πίνακα που ακολουθεί παρουσιάζεται η έρευνα της αναπαραγωγής των δαμασκηνιών. Μεταξύ του Οκτωβρίου 1931 και Φεβρουαρίου 1932 έγιναν τομές στις ρίζες των παλιών δαμασκηνιών. Έτσι το πείραμα έχει 2 παράγοντες με 2 επίπεδα ο καθένας. Πιο συγκεκριμένα :

- ο πρώτος παράγοντας είναι ο χρόνος εμφύτευσης δηλ. αν φυτεύτηκε κατευθείαν μετά την κοπή ή αν ενσωματώθηκε στην άμμο και φυτεύτηκε την επόμενη άνοιξη
- ο δεύτερος παράγοντας είναι το μήκος των ριζών που κόβονται δηλ. 6cm ή 12cm

Η κατάσταση του φυτού (ζωντανό ή πεθαμένο) τον Οκτωβρίου 1932 είναι η απόκριση y. Το δείγμα των 240 τομών δέντρων από κάθε ένα από τους τέσσερις σχεδιασμούς και τον ομοιόμορφο σχεδιασμό μελέτησαν οι Hoblyn και Palmer.

Μήκος Τομής	**Χρόνος εμφύτευσης**	**Αριθμός επιζώντων**
6 cm	Κατευθείαν	107
	Την Άνοιξη	31
12 cm	Κατευθείαν	156
	Την Άνοιξη	84

Μετά την προσαρμογή του logit μοντέλου έχουμε :

$\hat{\beta}$ = (- 0.5088, -0.5088,0.7138) ΄

w = (0.224 , 0.128, 0.221, 0.221) ΄

p_0 = (0.2818, 0.1686, 0.2748, 0.2748) ΄

$|X'WX|$ = 8. 197 10^{-3}

Η ορίζουσα του πίνακα πληροφορίας του ομοιόμορφου σχεδιασμού είναι: 7. 975 10^{-3}

Η αποδοτικότητα του μοντέλου αυτού είναι : $\left(\frac{7.975\ 10^{-3}}{8.197\ 10^{-3}} \right)^{1/3}$ = 99.1 %

5.2 Γενική περίπτωση : 2^k πειράματα

Η γενική περίπτωση των 2^k σχεδιασμών είναι αρκετά πολύπλοκη όμως θα μελετήσουμε την περίπτωση ομοιόμορφου σχεδιασμού η οποία έχει την μέγιστη – ελά χιστη βέλτιστη ιδιότητα δηλ. μεγιστοποιεί το κάτω φράγμα του D βέλτιστου κριτηρίου.

Έστω q παράμετροι (κύριες επιδράσεις και αλληλεπιδράσεις) στο επιλεγμένο μοντέλο για 2^k. Άρα ο β είναι διάνυσμα q x 1 και ο πίνακας X είναι 2^k x q. Επίσης, ο πίνακας X μπορεί να επεκταθεί σε έναν 2^k x 2^k πίνακα F του πλήρους μοντέλου. Όμως ο πίνακας F είναι ένας πίνακας Hadamard δηλ. F΄F = $2^k I_{2^k}$.

$$\left|\mathrm{F'WF}\right| = \left|\mathrm{X'}WX\right|\left|R'WR - R'WX\left(X'WX\right)^{-1}X'WR\right|$$

$$\leq \left|\mathrm{X'}WX\right|\left|R'WR\right|$$

$$\left|\mathrm{X'}WX\right| \geq \frac{\left|F'WF\right|}{\left|R'WR\right|}$$

Όμως $|\mathrm{F'WF}| = 2^{kq}\prod w_i \prod p_i$ και $\left|R'WR\right| \leq (w_M)^{2^k-q}\ |R'PR|$ με :

$$w_M = \max\{w_1, \ldots, w_{2^k}\}$$

$$\mathrm{P} = \mathrm{diag}\{p_1, \ldots, p_{2^k}\}.$$

Επομένως έχουμε για την $|X'WX|$:

$$|X'WX| \geq \frac{2^{kq}\prod w_i \prod p_i}{(w_M)^{2^k - q}}$$

αυτό είναι το κατώτερο όριο που μεγιστοποιείται όταν $p_1 = p_2 = \ldots = p_{2^k} = \frac{1}{2^k}$.

Η απώλεια αποδοτικότητας του ομοιόμορφου σχεδιασμού είναι :

$$R_u(w) = 1 - \left(\frac{\left|X'diag\left(\frac{w_1}{2^k},\ldots,\frac{w_{2^k}}{2^k}\right)X\right|}{\max|X'WX|}\right)^{1/q}$$

Τέλος, μπορεί να δειχθεί ότι :

$$R_u(w) \leq 1 - \frac{|X'W_0X|^{1/q}}{2^k w_M}$$

$$\leq 1 - \frac{w_m}{w_M}$$

όπου $w_m = \min\{w_1,\ldots,w_{2^k}\}$ και $W_0 = \text{diag}(w_1,\ldots,w_{2^k})$.

Για παράδειγμα, η αποτελεσματικότητα ενός σχεδιασμού δεν είναι μικρότερη του 70% αν $0.14 \leq w_i \leq 0.20$ ανεξάρτητα από τις τιμές των k και του πίνακα X.

ΠΑΡΑΡΤΗΜΑ

Λήμμα 1(Σελίδα 48):

Αν ξ_i (i=1,2,…,m) θετικοί πραγματικοί αριθμοί με $\sum_{i=1}^{m} \xi i = 1$ τότε για οποιαδήποτε ομάδα διακριτών πραγματικών αριθμών $\alpha_1, \alpha_{2,\ldots} \alpha_m$ υπάρχει c τέτοιο ώστε να ισχύει :

$$\sum_{i=1}^{m} \xi_i \frac{e^{a_i}}{(1+e^{a_i})^2} = \frac{e^c}{(1+e^c)^2}$$

$$\sum_{i=1}^{m} \xi_i a_i^2 \frac{e^{a_i}}{(1+e^{a_i})^2} \le c^2 \frac{e^c}{(1+e^c)^2}$$

Απόδειξη:

Χωρίς βλάβη γενικότητας υποθέτουμε ότι $\alpha_i \ge 0$ αφού $\frac{e^{a_i}}{(1+e^{a_i})^2} = \frac{e^{-a_i}}{(1+e^{-a_i})^2}$

Θα αποδείξουμε ότι το λήμμα ισχύει για m = 2 για το λόγο αυτό υποθέτουμε ότι υπάρχουν μη αρνητικοί πραγματικοί αριθμοί p και q που ικανοποιούν την εξίσωση:

$$p + q = 1.$$

Για $\alpha, z \ge 0$ πρέπει να δείξουμε ότι υπάρχει $c \ge 0$ που ικανοποιεί τις :

$$p\frac{e^{\alpha}}{(1+e^{\alpha})^2} + q\frac{e^{z}}{(1+e^{z})^2} = \frac{e^c}{(1+e^c)^2} \qquad \text{(A.1)}$$

$$pa^2\frac{e^{\alpha}}{(1+e^{\alpha})^2} + qz^2\frac{e^{z}}{(1+e^{z})^2} \le c^2\frac{e^c}{(1+e^c)^2} \qquad \text{(A.2)}$$

Χωρίς βλάβη γενικότητας υποθέτουμε ότι $\alpha < z$ και ότι :

$$A(z) = p\frac{e^{\alpha}}{(1+e^{\alpha})^2} + q\frac{e^{z}}{(1+e^{z})^2} \qquad \text{(A.3)}$$

Αφού για κάθε x ισχύει $0 < \frac{e^x}{(1+e^x)^2} < ¼$ τότε έχουμε $0 < A(z) < ¼$.

Γράφοντας $w = e^c$ έχουμε ότι $A(z) = \frac{w}{(1+w)^2}$ με w >1.

Η λύση είναι :

$$w = \frac{\left(1-2A\left(z\right)\right)+\sqrt{1-4A(z)}}{2A(z)}$$ άρα το c ισούται με :

$$c = c\,(z) = \ln w = \ln \frac{\left(1-2A\left(z\right)\right)+\sqrt{1-4A(z)}}{2A(z)} \qquad (A.4)$$

Από την εξίσωση (A.2) έχουμε ότι η $\frac{e^z}{(1+e^z)^2}$ είναι φθίνουσα και ότι :

$$\alpha \le c(z) \le z \qquad (A.5)$$

Τώρα χρειάζεται να δείξουμε ότι το c (z) ικανοποιεί την (A.2) για $z \ge \alpha$. Έστω

$$g(z) = z^2 \frac{e^z}{(1+e^z)^2} \qquad (A.6)$$

$$f(z) = (c(z))^2 \frac{e^{c(z)}}{\left(1+e^{c(z)}\right)^2} - pa\left(\frac{e^{\alpha}}{(1+e^{\alpha})^2}\right)^2 - qz\,\frac{e^z}{(1+e^z)^2}$$

$$= g\,(c\,(z)) - [pg(\alpha) + qg(z)] \qquad (A.7)$$

Για να δείξουμε ότι η c (z) ικανοποιεί την (A.2) πρέπει να δείξουμε ότι $f(z) \ge 0$ για $z \ge \alpha$. Από την (A.4) προκύπτει ότι $c(\alpha) = \alpha$ και $f(\alpha) = 0$. δηλ. αρκεί να δείξω ότι η η f(z) είναι αύξουσα συνάρτηση του z.

$$\frac{d\,f(z)}{d\,z} = \frac{d\,g\,(c(z))}{d\,z} - q\frac{d\,g\,(z)}{d\,z}$$

$$= \frac{d\,g\,(c(z))}{d\,z}\frac{d\,c(z)}{d\,A(z)}\frac{d\,A(z)}{d\,z} - q\frac{d\,g\,(c(z))}{dz} \qquad (A.8)$$

$$=\frac{\mathrm{c}(\mathrm{z})}{e^{c(z)}-1}[\ \{\mathrm{c}(\mathrm{z})+2\}-e^{c(z)}\{\mathrm{c}(\mathrm{z})-2\}]q\frac{(e^{z}-1)\ e^{z}}{(1+e^{z})^{3}}$$

$$-q\frac{ze^{z}}{(1+e^{z})^{3}}[(z+2)-e^{z}(z-2)]$$

Η παραπάνω παράγωγος είναι θετική αν και μόνο αν :

$$\frac{\mathrm{c}(\mathrm{z})}{e^{c(z)}-1}[\ \{\mathrm{c}(\mathrm{z})+2\}-e^{c(z)}\{\mathrm{c}(\mathrm{z})-2\}]\ \geq\frac{\mathrm{z}}{e^{z}-1}[(z+2)-e^{z}(z-2)] \qquad (A.9)$$

Από τις (A.5) και (A.9) έχουμε ότι $\mathrm{c}(\mathrm{z})\leq z$ αν δείξουμε ότι η συνάρτηση :

$$\mathrm{h}(\mathrm{z})=\frac{\mathrm{z}}{e^{z}-1}[(z+2)-e^{z}(z-2)]$$

είναι φθίνουσα συνάρτηση του z δηλ. η παράγωγος της είναι αρνητική ως προς z
Έχουμε :

$$\frac{d\ \mathrm{h}(\mathrm{z})}{d\ z}=-2\ \frac{(z+1)+e^{2z}\ (\mathrm{z}-1)}{(e^{z}-1)^{2}}$$

η οποία είναι αρνητική για κάθε z > 0.

Αφού αποδείχθηκε το λήμμα για m=2 πρέπει να δειχθεί για m ≥ 3 υποθέτουμε ότι :

$$\xi_{1*}=\ \xi_{1}/(1-\sum_{i=3}^{m}\xi_{i})\quad \text{και}\quad \xi_{2*}=\ \xi_{2}/(1-\sum_{i=3}^{m}\xi_{i}) \qquad (A.10)$$

με $\xi_{1*}+\xi_{2*}=1.$ Θέλουμε να αποδείξουμε την ύπαρξη του c που ικανοποιεί τις (2.1) και (2.2) ισοδυναμεί με το να δείξουμε ότι :

$$(1-\sum_{i=3}^{m}\xi_{i})[\xi_{1*}\frac{e^{\alpha_1}}{(1+\ e^{\alpha_1})^{2}}+\xi_{2*}\frac{e^{\alpha_2}}{(1+\ e^{\alpha_2})^{2}}]+\sum_{i=3}^{m}\xi_{i}\frac{e^{\alpha_i}}{(1+\ e^{\alpha_i})^{2}}=\frac{e^{c}}{(1+e^{c})^{2}}$$

$$(1-\sum_{i=3}^{m}\xi_{i})[\xi_{1*}a_1^{2}\frac{e^{\alpha_1}}{(1+\ e^{\alpha_1})^{2}}+\xi_{2*}a_2^{2}\frac{e^{\alpha_2}}{(1+\ e^{\alpha_2})^{2}}]+\sum_{i=3}^{m}\xi_{i}a_i^{2}\frac{e^{\alpha_i}}{(1+\ e^{\alpha_i})^{2}}\leq c^{2}\frac{e^{c}}{(1+e^{c})^{2}} \qquad (A.11)$$

Υποθέτουμε ότι υπάρχει c_* που ικανοποιεί τις :

$$\xi_{1*}\frac{e^{\alpha_1}}{(1+\ e^{\alpha_1})^2}+\xi_{2*}\frac{e^{\alpha_2}}{(1+\ e^{\alpha_2})^2}=\frac{e^{c_*}}{(1+\ e^{c_*})^2}$$

$$\xi_{1*}a_1^2\frac{e^{\alpha_1}}{(1+\ e^{\alpha_1})^2}+\xi_{2*}a_2^2\frac{e^{\alpha_2}}{(1+\ e^{\alpha_2})^2}=c_*^2\frac{e^{c_*}}{(1+\ e^{c_*})^2} \qquad \text{(A.12)}$$

Λαμβάνοντας υπόψη της Α.11 και Α.12 μπορούμε να δείξουμε ότι :

$$(1-\sum_{i=3}^{m}\xi_i)\frac{e^{c_*}}{(1+\ e^{c_*})^2}+\sum_{i=3}^{m}\xi_i\frac{e^{\alpha_i}}{(1+\ e^{\alpha_i})^2}=\frac{e^{c}}{(1+e^{c})^2}$$

$$(1-\sum_{i=3}^{m}\xi_i)c_*^2\frac{e^{c_*}}{(1+\ e^{c_*})^2}+\sum_{i=3}^{m}\xi_i a_i^2\frac{e^{\alpha_i}}{(1+\ e^{\alpha_i})^2}\le c^2\frac{e^{c}}{(1+e^{c})^2} \qquad \text{(A.13)}$$

Έτσι αποδείχθηκε το ζητούμενο.

<u>Λήμμα 2 (Σελίδα 50):</u>

Αν $I(\theta_1, \beta)$ πίνακας όπως αναφέρθηκε πριν και $I_c(\theta_1, \beta)$ ο πίνακας πληροφορίας του σχεδιασμού {(c, 1/2), (-c, 1/2)} όπου η c ικανοποιεί το λήμμα 1 τότε το διάνυσμα των ιδιοτιμών του $I(\theta_1, \beta)$ κυριαρχεί του διανύσματος των ιδιοτιμών του $I_c(\theta_1, \beta)$.

<u>Απόδειξη :</u>

Όπως είναι γνωστό το διάνυσμα των ιδιοτιμών ενός πραγματικού συμμετρικού πίνακα κυριαρχεί του διανύσματος των διαγώνιων στοιχείων . Με άλλα λόγια, το διάνυσμα των διαγώνιων στοιχείων του πίνακα $I(\theta_1, \beta)$ είναι ασθενώς μεγαλύτερο από το διάνυσμα των ιδιοτιμών του $I(\theta_1, \beta)$. Όταν η c ικανοποιεί το λήμμα 1 τότε το διάνυσμα ιδιοτιμών του $I_c(\theta_1, \beta)$ δηλ. το διάνυσμα

$$(\beta^2\sum_{i=1}^{m}\xi_i\frac{e^{-\alpha_i}}{(1+e^{-\alpha_i})^2},\frac{1}{\beta^2}\sum_{i=1}^{m}\xi_i\,\alpha_i^2\frac{e^{-\alpha_i}}{(1+e^{-\alpha_i})^2})'$$

κυριαρχεί του διανύσματος

$$(\beta^2 \frac{e^{-c}}{(1+e^{-c})^2}, \frac{1}{\beta^2} c^2 \frac{e^{-c}}{(1+e^{-c})^2})'$$

Αφού λοιπόν το δεύτερο διάνυσμα είναι το διάνυσμα των διαγώνιων στοιχείων του $I(\theta_1, \beta)$ η απόδειξη του λήμματος 2 είναι πλήρης.

Θεώρημα 1(Σελίδα 83):

Η L(p) έχει μοναδικό μέγιστο στο σημείο p = (0, 1/3, 1/3, 1/3) αν και μόνο αν :

$$v_1 \geq v_2 + v_3 + v_4$$

Απόδειξη :

A. Αν $v_1 \geq v_2 + v_3 + v_4$:

$$L = v_4(p_1p_2p_3 + p_2p_3p_4) + v_3(p_1p_2p_4 + p_2p_3p_4) + v_2(p_1p_3p_4 + p_2p_3p_4)$$
$$+ (v_1 - v_2 - v_3 - v_4)\, p_2p_3p_4$$
$$= v_4(p_1 + p_4)\, p_2p_3 + v_3(p_1 + p_3)\, p_2p_4 + v_2(p_1 + p_2)\, p_3p_4$$
$$+ (v_1 - v_2 - v_3 - v_4)\, p_2p_3p_4$$
$$\leq v_4(\frac{(p_1 + p_4) + p_2 + p_3}{3})^3 + v_3(\frac{(p_1 + p_3) + p_2 + p_4}{3})^3 + v_2(\frac{(p_1 + p_2) + p_3 + p_4}{3})^3$$
$$+ (v_1 - v_2 - v_3 - v_4)\, p_2\, p_3\, p_4 \qquad \text{(B.1)}$$
$$= \frac{v_4}{27} + \frac{v_3}{27} + \frac{v_2}{27} + (v_1 - v_2 - v_3 - v_4)\, p_2\, p_3\, p_4$$
$$\leq \frac{v_2 + v_4 + v_3}{27} + (v_1 - v_2 - v_3 - v_4)\,(\frac{p_2 + p_3 + p_4}{3})^3$$
$$+ (v_1 - v_2 - v_3 - v_4)\, p_2p_3p_4 \qquad \text{(B.2)}$$

$$\leq \frac{\nu_{2+}\, v_4 \,+\nu_3}{27} + \left(v_1 - v_2 - v_3 - v_4\right)\left(\frac{p_1+p_2+p_3+p_4}{3}\right)^3 \qquad \text{(B.3)}$$

$$= \frac{1}{27}\, v_1$$

Δηλαδή η σχέση (B.1) ισχύει αν και μόνο αν :

$$p_1 + p_4 = p_2 = p_3$$
$$p_1 + p_3 = p_2 = p_4$$
$$p_1 + p_2 = p_3 = p_4$$

Η λύση του συστήματος αυτού είναι :

$$p_1 = 0,\ p_2 = p_3 = p_4 = 1/3$$

Η σχέση (B.2) ισχύει αν και μόνο αν : $p_2 = p_3 = p_4$ ενώ η (B.3) ισχύει αν και μόνο αν : $p_1 = 0$

B. Αν το σημείο p= (0, 1/3, 1/3,1/3) μεγιστοποιεί την L(p) ισχυριζόμαστε ότι :

$$v_1 \geq v_2 + v_3 + v_4$$

διαφορετικά αν $v_1 < v_2 + v_3 + v_4$ η λύση είναι της μορφής:

$$p_\varepsilon = (\varepsilon, (1-\varepsilon)/3\ ,\ (1-\varepsilon)/3\ , (1-\varepsilon)/3\ )$$

για μικρές τιμές του ε >0. Δείχνετε εύκολα η ότι $L(p_\varepsilon) > L(p)$ αν και μόνο αν :

$$3d_1 > \varepsilon(3 + 6d_1 - 2\varepsilon - 3d_1\varepsilon)$$

όπου $d_1 = \frac{1}{v_1}(-v_1 + v_2 + v_3 + v_4) > 0$

<u>Θεώρημα 2(Σελίδα 83) :</u>

Αν $v_1 \geq v_2$, $v_3 = v_4 = v$ και $v_1 < v_2 + 2v$ τότε η λύση που μεγιστοποιεί την L(p) είναι :

$$p_1 = \frac{1}{2} - \frac{\nu_1 - \nu_2 + 4\nu}{2\,(-2\delta + D)},\ p_2 = \frac{1}{2} + \frac{\nu_1 - \nu_2 - 4\nu}{2\,(-2\delta + D)},\ p_3 = p_4 = \frac{2\nu}{(-2\delta + D)} \qquad (\Gamma.1)$$

με $L = 2\nu^2(\delta^2 + 4\nu_1\nu_2 - \delta D)\ /(-2\delta + D)^3$, $\delta = \nu_1 + \nu_2 - 4\nu$ και $D = \sqrt{12\nu_1\nu_2 + \delta^2}$

<u>Απόδειξη:</u>

Μπορεί να δειχθεί ότι αν $\nu_3 = \nu_4$ τότε αυτό συνεπάγεται $p_3 = p_4$ σε κάθε λύση της L (p). Αν θεωρήσουμε $x = p_1$, $y = p_2$ και $z = p_3 = p_4$ τότε $x = 1 - y - 2z$. Το πρόβλημα της L(p) μπορεί να γραφτεί ως μεγιστοποίηση της :

$$L(y,z) = z\left[\, 2\nu y - 2\nu y^2 + \nu_2 y + (\nu_1 - \nu_2 - 4\nu)\, yz - 2\,\nu_2 z^2 \right]$$

με τους εξής περιορισμούς :

$$y \geq 0, z \geq 0 \text{ και } y + 2z \leq 1$$

Πρώτα πρέπει να λάβουμε υπόψη μας τις εξής περιπτώσεις :

α) αν $z = 0$ τότε $L = 0$

β) $y = 0$ τότε $L = \nu_2 x z^2$ με μέγιστη τιμή την $\nu_2 / 27$ στο σημείο $p_1 = p_3 = p_4 = 1/3$, $p_2 = 0$.

γ) αν $y + 2z = 1$, $x=0$ τότε $L = \nu_1 y z^2$ με μέγιστη τιμή την $\nu_1 / 27$ στο σημείο $p_1 = 0$,

$p_1 = p_3 = p_4 = 1/3$

Εφαρμόζοντας όμως το Θεώρημα 1 παρατηρούμε ότι καμία από αυτές τις περιπτώσεις δεν αποτελεί λύση.

Υποθέτουμε ότι $y > 0$, $z > 0$ και $y + 2z < 1$ και παραγωγίζοντας την L (y,z) έχουμε :

$$\frac{\partial L}{\partial y} = 0 \Rightarrow z\left[\, (\nu_1 - \nu_2 - 4\nu)\, z + 2\nu - 4\nu y \right] = 0$$

$$\Rightarrow y = y_* = z(\nu_1 - \nu_2 - 4\nu)\ /\ 4\nu + 1/2$$

Για $y = y_*$ η L (y,z) μας δίνει :

$$L(z) = z(\frac{\Delta}{8\nu} z^2 + \frac{\nu_1 + \nu_2 - 4\nu}{2} z + \frac{\nu}{2})$$

όπου $\Delta = (\nu_1 - \nu_2 - 4\nu)^2 - 4\nu_1\nu_2$.

Άρα χρειάζεται να μεγιστοποιήσουμε μόνο την L (z) υπό τον περιορισμό $0 \leq z \leq z_*$:

$$\frac{\partial L(z)}{\partial z} = 0 \Rightarrow \frac{3}{8}\frac{\Delta}{\nu} z^2 + \frac{\nu_1 + \nu_2 - 4\nu}{2} 2z + \frac{\nu}{2} = 0$$

$$\Rightarrow z = z_* = 2\nu / (\nu_1 - \nu_2 - 4\nu)$$

Υποθέτοντας ότι $\nu_1 > \nu_2$ διακρίνουμε τις εξής περιπτώσεις :

i) Δ = 0

$$\frac{\partial L(z)}{\partial z} = 0 \Rightarrow \frac{\nu_1 + \nu_2 - 4\nu}{2} 2z + \frac{\nu}{2} = 0$$

$$\Rightarrow z = z_0 = \nu / 2(4\nu - \nu_1 - \nu_2)$$

όμως $\nu_1 < \nu_2 + 2\nu$ τότε $\sqrt{\nu_1} = 2\sqrt{\nu} - \sqrt{\nu_2}$ συνεπώς $9\nu_2 > \nu_1 > \nu > \nu_2$ και $\nu_1 - \nu_2 - 4\nu < 0$.Στο σημείο z_0 έχουμε τη μέγιστη τιμή της L (z) με τιμή. $(\sqrt{\nu_1} + \sqrt{\nu_2})^4 / 256\sqrt{\nu_1\nu_2}$.Η λύση μεγιστοποίησης της L(p) είναι :

$$p_1 = \frac{3\sqrt{\nu_2} - \sqrt{\nu_1}}{8\sqrt{\nu_2}}, p_2 = \frac{3\sqrt{\nu_1} - \sqrt{\nu_2}}{8\sqrt{\nu_1}}, p_3 = p_4 = \frac{(\sqrt{\nu_1} + \sqrt{\nu_2})^2}{16\sqrt{\nu_1\nu_2}}$$

(Γ.2)

Η σχέση (Γ.2) είναι εύκολο να δειχθεί ότι είναι μια ειδική περίπτωση της (Γ. 1) όπου

$$\sqrt{\nu_1} = 2\sqrt{\nu} - \sqrt{\nu_2}$$

Από εδώ και πέρα υποθέτουμε ότι $\Delta = \delta^2 - 4\nu_1\nu_2 \neq 0$, $\delta = \nu_1 + \nu_2 - 4\nu$ και οι λύσεις του $\frac{\partial L(z)}{\partial z} = 0$ είναι οι :

$$z_1 = \frac{2\nu}{2\delta + \sqrt{\delta^2 + 12\nu_1\nu_2}}, z_2 = \frac{2\nu}{-2\delta + \sqrt{\delta^2 + 12\nu_1\nu_2}}$$

ενώ $\frac{\partial L(z)}{\partial z} = \frac{3}{8}\frac{\Delta}{\nu}(z - z_1)(z - z_2)$.

ii) Δ > 0

Τότε $v_1 < v_2 + 2v$, $\sqrt{v_1} < 2\sqrt{v} - \sqrt{v_2}$ και $0 < z_2 < z_* < z_1$. Η L (z) αποκτά μέγιστο στο σημείο z_2 και η λύση αντιστοιχεί στην (Γ.1) η οποία μεγιστοποιεί την L (p).

iii) Δ < 0

Τότε $\nu < 4\nu_2$, $z_1 < 0 < z_2 < z_*$ και η L (z) αποκτά μέγιστο στο σημείο z_2 οπότε είναι ίδιο με το ii.

Επομένως η λύση (Γ.1) μεγιστοποιεί την L (p) όταν $v_1 > v_2, v_4 = v_3 = \nu$ και $v_1 < \nu + v_2$.

<u>Θεώρημα 3(Σελίδα 85) :</u>

Για τη συνάρτηση σύνδεσης logit η συνθήκη κορεσμού ισχύει αν και μόνο αν:

$$\beta_0 \neq 0$$

$$|\beta_1| > \frac{1}{2}\log\left(\frac{e^{2|\beta_0|}+1}{e^{2|\beta_0|}-1}\right)$$

$$|\beta_2| \geq \log\left(\frac{\sqrt{(e^{4|\beta_0|}-1)(e^{4|\beta_1|}-1)} + 2e^{|\beta_0|+|\beta_1|}}{(e^{2|\beta_0|}-1)(e^{2|\beta_1|}-1) - 2}\right)$$

<u>Απόδειξη :</u>

Υποθέτουμε ότι $\alpha = e^{\beta_0}$, $b = e^{\beta_1}$, $c = e^{\beta_2}$ και

$$\frac{1}{w_1} \propto (\alpha + bc)^2, \frac{1}{w_2} \propto (b + \alpha c)^2, \frac{1}{w_3} \propto (c + b\alpha)^2, \frac{1}{w_4} \propto (1 + \alpha bc)^2$$

Διακρίνουμε τις εξής περιπτώσεις :

α) αν $\beta_1 = 0$ τότε $b = 1$ με $w_1 = w_3, w_2 = w_4$

β) αν $\beta_2 = 0$ τότε $w_1 = w_2, w_3 = w_4$

γ) αν $\beta_0 = 0$ τότε $w_1 = w_4, w_3 = w_2$

Η συνθήκη κορεσμού δεν είναι αληθής στις παραπάνω περιπτώσεις αλλά μπορεί να αναφερθεί ως εξής :

1) για $(1+b^2)(1+c^2) < 2$

$$\text{αν } 0 < \alpha \le \alpha_1 \text{ τότε } \frac{1}{w_4} \ge \frac{1}{w_1} + \frac{1}{w_2} + \frac{1}{w_3}$$

$$\text{αν } \alpha \ge \alpha_4 \text{ τότε } \frac{1}{w_1} \ge \frac{1}{w_4} + \frac{1}{w_2} + \frac{1}{w_3}$$

όπου $a_1 = \dfrac{-2bc + \sqrt{(1-b^4)(1-c^4)}}{2-(1-b^2)(1-c^2)}, a_4 = \dfrac{2bc + \sqrt{(1-b^4)(1-c^4)}}{2-(1+b^2)(1+c^2)}$.

Στην περίπτωση αυτή b <1, c <1, $\alpha_4 > \alpha_1 > 0$.

Στην περίπτωση αυτή η σχέση $|\beta_0| \ge \log(\alpha_4)$ μας εγγυάται τη συνθήκη κορεσμού $(1+ b^2)(1+ c^2) < 2$

2) για $(1-b^2)(1-c^2) > 2$

$$\text{αν } 0 < \alpha \le \alpha_3 \text{ τότε } \frac{1}{w_1} \ge \frac{1}{w_4} + \frac{1}{w_2} + \frac{1}{w_3}$$

$$\text{αν } \alpha \ge \alpha_2 \text{ τότε } \frac{1}{w_4} \ge \frac{1}{w_1} + \frac{1}{w_2} + \frac{1}{w_3}$$

όπου $a_3 = \dfrac{-2bc + \sqrt{(1-b^4)(1-c^4)}}{(1+b^2)(1+c^2)-2}, a_2 = \dfrac{2bc + \sqrt{(1-b^4)(1-c^4)}}{(1-b^2)(1-c^2)-2}$.

Στην περίπτωση αυτή b >1, c >1, $\alpha_2 > \alpha_3 > 0$.

Στην περίπτωση αυτή η σχέση $(1-b^2)(1-c^2) > 2$ δηλώνει ότι $\beta_2 > \frac{1}{2}\log(\frac{e^{2\beta_1}+1}{e^{2\beta_1}-1})$.

3) για $(1+b^2)(1-c^2) > 2$

$$\text{αν } 0 < \alpha \le \alpha_5 \text{ τότε } \quad \frac{1}{w_2} \ge \frac{1}{w_4} + \frac{1}{w_1} + \frac{1}{w_3}$$

$$\text{αν } \alpha \ge \alpha_8 \text{ τότε } \quad \frac{1}{w_3} \ge \frac{1}{w_4} + \frac{1}{w_1} + \frac{1}{w_2}$$

όπου $a_5 = \dfrac{-2bc + \sqrt{-\left(1-b^4\right)(1-c^4)}}{2-\left(1-b^2\right)\left(1+c^2\right)}, a_8 = \dfrac{2bc + \sqrt{-\left(1-b^4\right)(1-c^4)}}{\left(1+b^2\right)\left(1-c^2\right)-2}.$

Στην περίπτωση αυτή b >1, c >1, $\alpha_8 > \alpha_5 > 0$.

Στην περίπτωση αυτή η σχέση $(1-\mathrm{b}^2)(1-\mathrm{c}^2) > 2$ δηλώνει ότι $\beta_2 < -\frac{1}{2}\log(\frac{e^{2\beta_1}+1}{e^{2\beta_1}-1})$.

Επίσης οι περιπτώσεις (2) και (3) συνοψίζονται ως εξής:

$$|\beta_2| > \frac{1}{2}\log(\frac{e^{2\beta_1}+1}{e^{2\beta_1}-1})$$

$$|\beta_0| > \frac{1}{2}\log a_2(\beta_1, |\beta_2|)$$

4) για $(1-\mathrm{b}^2)(1+\mathrm{c}^2) > 2$

$$\text{αν } 0 < \alpha \le \alpha_7 \text{ τότε } \quad \frac{1}{w_3} \ge \frac{1}{w_4} + \frac{1}{w_1} + \frac{1}{w_2}$$

$$\text{αν } \alpha \ge \alpha_6 \text{ τότε } \quad \frac{1}{w_2} \ge \frac{1}{w_4} + \frac{1}{w_1} + \frac{1}{w_3}$$

όπου $a_7 = \dfrac{-2bc + \sqrt{-\left(1-b^4\right)(1-c^4)}}{2-\left(1+b^2\right)\left(1-c^2\right)}, a_8 = \dfrac{2bc + \sqrt{-\left(1-b^4\right)(1-c^4)}}{\left(1-b^2\right)\left(1+c^2\right)-2}.$

Στην περίπτωση αυτή b <1, c >1, $\alpha_6 > \alpha_7 > 0$.

Επίσης μπορεί να αποδειχθεί ότι $\alpha_1\ \alpha_4 = 1$, $\alpha_2\ \alpha_3 = 1$, $\alpha_5\ \alpha_8 = 1$ και $\alpha_6\ \alpha_7 = 1$. Τέλος αντικαθιστώντας το β_1 στην $\frac{1}{2}\log(\frac{e^{2\beta_1}+1}{e^{2\beta_1}-1})$ και στην $\log a_2(\beta_1, |\beta_2|)$ έχουμε τις περιπτώσεις (1) και (4).

Θεώρημα 4(Σελίδα 86):

Υποθέτοντας ότι $v_1 < v_2 < v_3 < v_4$ και $v_4 < v_1+v_2+v_3$ τότε :

$$\max\{L_{13,}L_{14},L_{24}\} \le \max\{L_{12,}L_{23},L_{34}\}$$

$$\max_p L - \max\{L_{12,}L_{23},L_{34}\} \le \min\{\frac{v_2-v_1}{216}, \frac{v_3-v_2}{96\sqrt{3}}, \frac{v_4-v_3}{54}\}$$

Απόδειξη :

Πρώτα αποδεικνύουμε ότι $L_{24} \le L_{23}$ αν $v_4+v_2 \le 2v_3$ ενώ $L_{24} \le L_{34}$ αν $v_2+v_4 > 2v_3$.

Έστω η λύση της L_{24} είναι $x = p_1, y = p_3, z = p_2 = p_4$.

- Αν $v_4+v_2 \le 2v_3$ τότε:

$$p_2 \ge p_3 \text{ και } L_{23} - L_{24} \ge (v_4 - v_3)xz(z-y)\big|_{x=p_1, y=p_3, z=p_2} \ge 0$$

- Αν $v_4+v_2 > 2v_3$ τότε:

$$p_2 \le p_3 \text{ και } L_{34} - L_{24} \ge (v_3 - v_2)xz(y-z)\big|_{x=p_1, y=p_3, z=p_2} \ge 0$$

Ομοίως δείχνουμε ότι :

- $L_{13} \le L_{23}$ $\alpha\nu$ $v_1+v_3 \ge 2v_2$
- $L_{13} \le L_{12}$ $\alpha\nu$ $v_1+v_3 < 2v_2$
- $L_{14} \le L_{13}$ $\alpha\nu$ $v_1+v_4 \le 2v_3$
- $L_{14} \le L_{24}$ $\alpha\nu$ $v_1+v_4 \ge 2v_2$

Επομένως $\max\{L_{13,}L_{14},L_{24}\} \le \max\{L_{12,}L_{23},L_{34}\}$.

Το αρχικό πρόβλημα είναι:

$$\max_p L = \max_p\{v_4p_1p_2p_3 + v_3p_1p_2p_4 + v_2p_1p_3p_4 + v_1p_4p_2p_3\}$$

ενώ η προσέγγιση του είναι:

$$L_{34} = \max_p\{v_4' p_1 p_2 p_3 + v_3' p_1 p_2 p_4 + v_2 p_1 p_3 p_4 + v_1 p_4 p_2 p_3\}$$

όπου $v_4' = v_3' = (v_3 + v_4)/2$.

Η διαφορά $\max_p L - L_{34} \le (v_4 - \nu_3) p_1^* p_2^* (p_3^* - p_4^*)/2$ όπου $(\ p_1^* p_2^* p_3^* p_4^*\ )'$ μεγιστοποιεί την L. Επίσης ισχύουν οι σχέσεις :

$$p_1^* \ge p_2^* \ge p_3^* \ge p_4^* \ge 0 \text{ και } p_1^* + p_2^* + p_3^* + p_4^* = 1$$

οπότε :

$$\max_p L - \max\{L_{12}, L_{23}, L_{34}\} = \min\{\max_p L - L_{12}, \max_p L - L_{23}, \max_p L - L_{34}\}$$

$$\le \min\{\frac{\nu_2 - \nu_1}{2} p_3^* p_4^* (p_1^* - p_2^*), \frac{\nu_3 - \nu_2}{2} p_1^* p_4^* (p_2^* - p_3^*), \frac{\nu_4 - \nu_3}{2} p_1^* p_2^* (p_3^* - p_4^*)\}$$

$$\le \min\{\frac{\nu_2 - \nu_1}{216}, \frac{\nu_3 - \nu_2}{96\sqrt{3}}, \frac{\nu_4 - \nu_3}{54}\}$$

<u>Θεώρημα 5(Σελίδα 88) :</u>

Χωρίς βλάβη γενικότητας, υποθέτουμε ότι $v_1 \ge v_2 \ge v_3 \ge v_4$ τότε :

$$R_{max} = \begin{cases} 1 - \frac{3\,(1+\frac{3\alpha}{b})^3}{4} & \alpha\nu\ \nu_1 \ge \nu_2 + \nu_3 + \nu_4 \\ 1 - \frac{3\,2^{1/3}}{4} & \alpha\nu\ \nu_1 < \nu_2 + \nu_3 + \nu_4 \end{cases}$$

<u>Απόδειξη :</u>

Η μεγιστοποίηση της R_u (w) είναι ισοδύναμη με την ελαχιστοποίηση της :

$$Q(\nu_1, \nu_2, \nu_3, \nu_4) = \frac{\nu_1 + \nu_2 + \nu_3 + \nu_4}{\nu_1 p_2 p_3 p_4 + \nu_2 p_1 p_3 p_4 + \nu_3 p_2 p_1 p_4 + \nu_4 p_1 p_2 p_3}$$

όπου (p_1, p_2, p_3, p_4) η βέλτιστη κατανομή των $(\nu_1, \nu_2, \nu_3, \nu_4)$.

i) αν $\nu_1 \ge \nu_2 + \nu_3 + \nu_4$ τότε $p_1 = 0,\ p_2 = p_3 = p_4 = 1/3$ και

$$R_u(w) = 1 - \frac{3}{4}(1 + \frac{\nu_2 + \nu_3 + \nu_4}{\nu_1})^{1/3}$$

Υποθέτοντας ότι $0 < \alpha \le \nu_i \le$ b για i = 1,2,3,4, $\nu_1 \ge \nu_2 + \nu_3 + \nu_4$ αποδεικνύεται ότι

$b \ge 3\alpha$ και $Q(\nu_1, \nu_{2,} \nu_{3,} \nu_4) = 27[1 + \frac{\nu_2 + \nu_3 + \nu_4}{\nu_1}] \ge 27\ [1 + \frac{3\alpha}{b}]$

Το ελάχιστο της συνάρτησης Q είναι στο $\nu_1 = b, \nu_2 = \nu_3 = \nu_4 = a$ και

$$R_u(w) = 1 - \frac{3}{4}(1 + \frac{3a}{b})^3$$

ενώ αν $\nu_1 = \nu_2 + \nu_3 + \nu_4$ τότε $Q(\nu_1, \nu_{2,} \nu_{3,} \nu_4) = 54$ και $R_u(w) = 1 - \frac{3}{4}2^{1/3}$.

ii) αν $\nu_1 < \nu_2 + \nu_3 + \nu_4$ τότε $p_1 > 0$. Υποθέτουμε δ>0 τέτοιο ώστε

$p_1' = p_1 - \delta(1 - p_1)$ δηλ. $p_i' = (\delta + 1)p_i$ i = 2,3,4 και $p_1' + p_2' + p_3' + p_4' = 1$. Για οποιαδήποτε $\nu_1' > \nu_1$ ισχύει :

$$Q(\nu_1', \nu_{2,} \nu_{3,} \nu_4) \le \frac{\nu_1' + \nu_2 + \nu_3 + \nu_4}{\nu_1' p_2' p_3' p_4' + \nu_2 p_1' p_3' p_4' + \nu_3 p_2' p_1' p_4' + \nu_4' p_1' p_2' p_3'} < Q(\nu_1, \nu_{2,} \nu_{3,} \nu_4)$$

για αρκετά μικρό δ >0. Επίσης επιβεβαιώνεται ότι

$\lim_{\nu_1 \uparrow \nu_2 + \nu_3 + \nu_4} Q(\nu_1, \nu_{2,} \nu_{3,} \nu_4) = Q(\nu_2 + \nu_3 + \nu_4, \nu_{2,} \nu_{3,} \nu_4) = 54$

ανεξάρτητα των τιμών v_2 , v_3 , v_4 .

ΒΙΒΛΙΟΓΡΑΦΙΑ

1. Myers R.M – Montgomery D.C. – Vining G.C. (2002). "*Generalized Linear Models with Applications in Engineering and the Science"*. Wiley Interscience, New York.

2. Khuri A.I. – Mukherjee B. – Sinha B.K. – Ghosh M (2006). *" Design Issues for Generalized Linear Models : A Review "*. Statistical Sciences Journal, Institute of Mathematical Statistics U.S.A.

3. Mathew T. - Sinha B.K. (2001) *"Optimal Designs for Binary Data under Logistic Regression"*. Journal of Statistical Planning and Inference, Elsevier, Holland.

4. Mandal A.- Yang J.- Majumdar D. (2010) *" Optimal Designs for two –Level Factorial Experiments with Binary Response".* Arxiv : 1003.1557v1, New York.

5. Κουκουβίνος Χ. (2005) Σημειώσεις Μαθήματος-*"Γραμμικά Μοντέλα και Σχεδιασμοί"*, Εθνικό Μετσόβειο Πολυτεχνείο,Αθήνα.

6. Οικονόμου Π.-Καρώνη Χ.(2007). Σημειώσεις Μαθήματος-*"Ανάλυση Παλινδρόμησης "* Εθνικό Μετσόβιο Πολυτεχνείο, Αθήνα.

7. Waterhouse T.H.-Woods D.C.- Eccleston J.A.-Lewis S.M. (2007) "*Design Selection Criteria for Discrimination / Estimation for Nested Models and a Binomial Response"*. Journal of Statistical Planning and Inference, Elsevier, Holland.

8. Berry D.A.- Fristedt B. (1985) *"Bandit Problems : Sequential Allocation of Experiments"* .Chapman and Hall, London.

9. Box G.E. – Drapper N.R (1987) *"Empirical Model- Building and Response Surfaces"*. Wiley, New York.

10. Chaloner K.- Larntz K. (1988) *" Optimal Bayesian Design Applied to Logistic Regression Experiments "*. Journal of Statistical Planning and Inference, Elsevier, Holland.

11. Chernoff H. (1953) *"Local Optimal Designs for Estimating Parameters".* Annals of Mathematical Statistics, Institute of Mathematical Statistics U.S.A.

12. McCullagh P.- Nelder J. (1989)" *Generalized Linear Models*". Second Edition Chapman and Hall, Boca Raton.

13. Atkinson A.C. (2007) *"DT- Optimum Designs for Model Discrimination and Parameter Estimation"* Journal of Statistical Planning and Inference, Elsevier, Holland.

14. Agresti A. (2002) "Categorical Data Analysis". Sec. Edition Wiley, New York

15. Wijesinha M.C. – Khuri A.I. (1987). *"Construction of Optimal Designs to Increase the Power of the Multiresponse Lack of Fit Test".* Journal of Statistical Planning and Inference, Elsevier, Holland.

16. Zocchi S.S.- Atkinson A.C. (1999) *"Optimum Experimental Designs for Multinomial Logistic Models".* Biometrics 55, 437- 444. MR 1705110

17. Wu C. F.J.(1985) *"Efficient Sequential Designs with Binary Data"*. Journal of American Statistical Association 80, 974-984. MR 0819603

18. Ponce de Leon A.C.M. – Atkinson A.C. (1991) *"Optimal Experimental Design for discriminating Between Two Rival Models in the Presence of Prior Information* ". Biometrika 78, 601-608

19. Ponce de Leon A.C.M. – Atkinson A.C. (1992) "The Design of Experiments to Discriminate Between *Two Rival Generalized Models* ". Advances in GLM and Statistical Modeling .Springer, New York.

20. Abdelbasit K.M.- Plackett R.L. (1983) "*Experimental Design for Binary Data"* Journal of Statistical Planning and Inference, Elsevier, Holland.

21. Dette H.- Haines L.M. (1994) *"E-optimal Designs for Linear and Non Linear Models with two Parameters"*. Biometrika,81, 739-754.