ΓΕΝΙΚΕΥΜΕΝΑ ΓΡΑΜΜΙΚΑ ΜΟΝΤΕΛΑ

ΜΕΡΟΣ Α

ΕΙΣΑΓΩΓΗ

ΛΟΓΙΣΤΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ

Γ. ΤΖΑΒΕΛΑΣ

1. Εισαγωγή

1.1. Σκοπός

Ο σκοπός του Μαθήματος αυτού είναι να εισάγει τον αναγνώστη σε μία τάξη στατιστικών μοντέλων που είναι φυσική γενίκευση των κλασσικών γραμμικών μοντέλων. Τα Γενικευμένα Γραμμικά Μοντέλα (Generalized Linear Models) περιλαμβάνουν σαν ειδική περίπτωση, την γραμμική παλινδρόμηση, την ανάλυση διασποράς, τα logit και probit μοντέλα, τα λογαριθμογραμμικά και τα πολυωνιμικά μοντέλα, καθώς και κάποια μοντέλα της ανάλυσης επιβίωσης. Αποδεικνύεται ότι αυτά τα μοντέλα μοιράζονται κάποιες κοινές ιδιότητες, καθώς και ότι έχουν κοινή μέθοδο εκτίμησης παραμέτρων. Αυτές οι κοινές ιδιότητες μας επιτρέπουν να μελετήσουμε μέσω των Γενικευμένων Γραμμικών Μοντέλων (Γ.Γ.Μ.) μία ευρεία ομάδα στατιστικών μοντέλων παρά το καθένα από αυτά χωριστά, Βασική προϋπόθεση είναι ο αναγνώστης να είναι εξοικειωμένος με τις βασικές στατιστικές έννοιες και μεθοδολογίες όπως, κατανομές δειγματοληψίας, έλεγχος υποθέσεων και απλή γραμμική παλινδρόμηση. Επιπλέον κάποιες γνώσεις γραμμικής άλγεβρας και απειροστικού λογισμού θεωρούνται απαραίτητες.

1.2 Ορολογία

Οι στατιστικές μέθοδοι που αναπτύσσονται στο Μάθημα αυτό έχουν σκοπό την ανίχνευση και μελέτη σχέσεων μεταξύ μετρήσεων που έγιναν σε διάφορες ομάδες ανθρώπων ή αντικειμένων. Για παράδειγμα οι μετρήσεις μπορεί να είναι το ύψος ή το βάρος αγοριών και κοριτσιών, ή το ύψος της σοδιάς κάποιων καλλιεργειών κάτω από διάφορες συνθήκες καλλιέργειας. Χρησιμοποιούμε τον όρο **εξαρτημένη μεταβλητή** για τις μετρήσεις που θεωρούμε σαν τυχαίες μεταβλητές. Οι μεταβλητές αυτές θεωρούνται ότι μεταβάλλονται ελεύθερα εν αντιθέσει με τις **ανεξάρτητες μεταβλητές** οι οποίες θεωρούνται σαν μη τυχαίες

μεταβλητές (δηλαδή παίρνουν συγκεκριμένες τιμές ανάλογα με τον σχεδιασμό του πειράματος).

Οι μεταβλητές μπορεί να ταξινομηθούν σαν

- **κατηγορικές ή ποιοτικές** π.χ. χρώμα ματιών, φύλο, τύπος αίματος, κ.λ.π.

  ειδικότερα για τις **δυαδικές μεταβλητές** υπάρχουν μόνο δυο κατηγορίες.
- **διάταξης** για τις οποίες υπάρχει κάποια φυσική διάταξη μεταξύ των κατηγοριών : π.χ. όταν η ηλικία καταγράφεται σαν νέος, μεσήλικας, γέρος; η όταν η αρτηριακή πίεση καταγράφεται σαν $\leq$70, 71-90, 91-110, $\geq$111 mm Hg.
- **Συνεχής μεταβλητές** όπου οι παρατηρήσεις μπορούν να πάρουν οποιαδήποτε τιμή σε κάποιο διάστημα.

Μια ποιοτική ανεξάρτητη μεταβλητή καλείται **παράγοντας (factor)** και οι κατηγορίες λέγονται **επίπεδα (level)** του παράγοντα. Οι συνεχείς ανεξάρτητες μεταβλητές λέγονται **covariates**.

1.3 Το γραμμικό μοντέλο

Το γραμμικό μοντέλο

$$y = Xb + e$$

περιγράφεται με τη βοήθεια πινάκων ως εξής

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_p \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

όπου $y = (y_1, y_2, ..., y_n)^T$ είναι η στήλη των παρατηρήσεων της εξαρτημένης μεταβλητής, ο πίνακας **X** διάστασης $n \times p$ είναι ο πίνακας των τιμών των ανεξάρτητων μεταβλητών $X_1, X_2, ...X_p$. Κάθε γραμμή αναφέρεται σε μια διαφορετική στατιστική μονάδα ή παρατήρηση και κάθε στήλη σε διαφορετική

ανεξάρτητη μεταβλητή. Η στήλη των παραμέτρων $b=(b_1,b_2,...,b_p)^T$ περιλαμβάνει τους συντελεστές των ανεξάρτητων μεταβλητών οι οποίοι θεωρούνται άγνωστοι και πρέπει να εκτιμηθούν. Η στήλη των υπολοίπων (residuals) $e=(e_1,e_2,...,e_n)^T$ είναι η στήλη των τυχαίων σφαλμάτων (random error terms).

Η υπόθεση που υιοθετούμε στο παραπάνω γραμμικό μοντέλο είναι ότι τα $e_1,e_2,....,e_n$ είναι ανεξάρτητα με την ίδια κατανομή $N(0,\sigma^2)$.

Το μοντέλο αυτό μπορεί να γενικευθεί με πολλούς τρόπους. Θα ασχοληθούμε με τις εξής δύο:

- Οι ανεξάρτητες παρατηρήσεις ακολουθούν κατανομή διαφορετική της κανονικής. Θα μπορούσε να είναι ακόμα και διακριτή.
- Η σχέση μεταξύ εξαρτημένων και ανεξάρτητων μεταβλητών να μην είναι γραμμική.

Η πρώτη γενίκευση βασίζεται στο γεγονός ότι πολλές από τις ``καλές`` ιδιότητες της κανονικής κατανομής απαντώνται σε μια μεγαλύτερη κλάση κατανομών την **Εκθετική Οικογένεια κατανομών**.

Η δεύτερη γενίκευση αναφέρεται στη σχέση $\mu=\mathrm{E}Y=X\beta$ η οποία μπορεί να αντικατασταθεί από την σχέση $\mu=g(Xb)$. Να σημειωθεί εδώ ότι η γραμμική έκφραση δεν εξαλείφεται πλήρως αλλά βρίσκεται μέσα σε κάποια άλλη συνάρτηση.

Ορισμός 1.1:Λέμε ότι η κατανομή μιας τ.μ. Υ ανήκει στην Εκθετική Οικογένεια κατανομών όταν μπορεί να γραφτεί στη μορφή

$$f(y;\theta)=\exp[\sum b_i(\theta)T_i(y)+c(\theta)+h(y)] \qquad (1.1)$$

όπου b_i,T_i,c,h θεωρούνται γνωστές συναρτήσεις.

Η b(θ) λέγεται **φυσική παράμετρος**. Στη συνέχεια θα γίνει φανερό ότι η παραμετρικοποίηση της (1.1) με τη βοήθεια της φυσικής παραμέτρου είναι πολύ βολική.

Τις περισσότερες φορές η μορφή η οποία θα μας απασχολήσει εδώ είναι η απλούστερη μορφή

$$f(y;\theta) = \exp[b(\theta)a(y) + c(\theta) + h(y)] \ . \qquad (1.2)$$

Η μορφή

$$f(y;\theta) = \exp[b(\theta)\, y + c(\theta) + h(y)]$$

λέγεται **κανονική μορφή .**

Αν επιπλέον η (1.2) μπορεί να γραφεί στη μορφή

$f(y;\theta) = \exp y\theta + k(\theta) + d(y)$ τότε λέμε ότι η Υ είναι γραμμένη στην **<u>τυπική της μορφή</u>.**

Το σύνολο $\Theta = \{\theta : \int f(y;\theta)dy < \infty\}$ **λέγεται παραμετρικός χώρος της f.**

Πολλές γνωστές κατανομές ανήκουν στην εκθετική οικογένεια κατανομών. Για παράδειγμα η Poisson η Δυωνυμική, η Κανονική κατανομή μπορούν να γραφτούν στην κανονική της μορφή. Πράγματι

<u>Για την Poisson</u>

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

όπου $a(y) = y, \quad b(\lambda) = \ln\lambda, c(\lambda) = -\lambda, h(y) = y!$

Προφανώς η παραπάνω κατανομή είναι στην κανονική της μορφή. Αν θα την παραμετρικοποιήσουμε σε σχέση με τη φυσική παράμετρο

$\theta = \ln\lambda$, η Posson κατανομή μπορεί να γραφτεί στη τυπική μορφή

$$f(y;\lambda) = \exp[y\theta - \exp(\theta) - \log y!]$$

με παραμετρικό χώρο Θ=ℜ

<u>Για την κανονική κατανομή</u>

$$f(y;\mu) = \frac{1}{(2\pi\sigma^2)^{1/2}}\exp[-\frac{1}{2\sigma^2}(y-\mu)^2$$

- Αν η μ είναι άγνωστη και η διακύμανση σ^2 θεωρείται γνωστή, τότε η κανονική κατανομή μπορεί να γραφτεί στη μορφή

$$f(y;\mu) = \exp[-\frac{y^2}{2\sigma^2} + \frac{y\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)] .$$

Αυτή είναι η κανονική της μορφή με φυσική παράμετρο $b(\mu) = \frac{\mu}{\sigma^2}$.

Επιπλέον $c(\mu) = -\frac{\mu}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)$ και $h(y) = -\frac{y^2}{2\sigma^2}$.

Παραμετρικοποιώντας την, σε σχέση με την φυσική παράμετρο έχουμε την τυπική της μορφή

$$f(y;\theta) = \exp[\theta y - \frac{\sigma^2}{2}\theta^2 - \frac{1}{2}\log(2\pi\sigma^2) - \frac{y^2}{2\sigma^2}]$$

Επιπλέον Θ=$\Re$.

- Για την περίπτωση όπου και το μ και το σ είναι άγνωστα τότε

$$f(y;\mu,\sigma) = \frac{1}{(2\pi\sigma^2)^{1/2}}\exp[-\frac{1}{2\sigma^2}(y-\mu)^2$$

$$= \exp[-\frac{y^2}{2\sigma^2} + \frac{\mu}{\sigma^2}y - \frac{\mu^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)]$$

Στην περίπτωση αυτή

$$T_1(y) = y^2, T_2(y) = y, d_1(\mu,\sigma) = -\frac{1}{2\sigma^2}, d_2(\mu,\sigma) = \frac{\mu}{\sigma^2}, c(\mu,\sigma) = -\frac{1}{2}\log[2\pi\sigma^2]$$

<u>Για την Διωνυμική κατανομή</u>

$$f(y;\pi) = \binom{n}{y}\pi^y(1-\pi)^{n-y} = \exp[\ln\pi y - y\ln(1-\pi) + n\ln(1-\pi) + \ln\binom{n}{y}]$$

$$= \exp[\log\frac{\pi}{1-\pi}y + n\ln(1-\pi) + \ln\binom{n}{y}]$$

Εχουμε $a(y) = y,\ c(\pi) = \ln\frac{\pi}{1-\pi},\ d(y) = \ln\binom{n}{y}$.

Παραμετρικοποιώντας την παραπάνω κατανομή με τη φυσική παράμετρο

$\theta = \ln\frac{\pi}{1-\pi}$ έχουμε την μορφή

$$f(y;\pi) = \exp[\ y\theta - n\ln(1+e^{\theta}) + \ln\binom{n}{y}]$$

Δεν είναι όλες οι κατανομές εκθετικής μορφής. Για παράδειγμα η ομοιόμορφη κατανομή με διάστημα (θ, 2) με θ<2 δεν μπορεί να γραφεί στη μορφή (*). Ένα άλλο παράδειγμα μη εκθετικής οικογένειας κατανομής είναι η Cauchy κατανομή.

Ασκηση 1 Να συμπληρωθεί ο ακόλουθος πίνακας

Κατανομή	b(.)	T(.)	c(.)	h(.)	Θ
NB(r,θ), r=γνωστό					
G(α,θ) α=γνωστό					
G(α,θ) θ=γνωστό					
IG(μ,1)					

Στη συνέχεια για να παράγουμε κάποιες ιδιότητες της εκθετικής οικογένειας κατανομών θα χρειαστούμε την ακόλουθη πρόταση.

<u>Πρόταση 1.1</u>. Αν f(x,θ) είναι μία οικογένεια κατανομών (όχι κατ' ανάγκη εκθετικής μορφής) για την οποία επιτρέπεται η παραγώγηση ως προς θ κάτω από το ολοκλήρωμα ως προς x τότε ισχύουν τα ακόλουθα.

$$1.\ E(\log f(x,\theta))' = E\frac{f'(x,\theta)}{f(x;\theta)} = 0 \qquad (1.3)$$

$$2.\ \mathrm{var}((\log f(x,\theta))') = -E[\log f(x,\theta)''] \qquad (1.4)$$

Εδώ η παραγώγιση λαμβάνεται ως προς την μεταβλητή θ.

Απόδειξη

1. $E([\log f(x,\theta)]') = E\frac{f'(x,\theta)}{f(x;\theta)} = \int f'(x,\theta)dv(x) = [\int f(x,\theta)dv(x)]' = 0$

2. Παραγωγίζοντας την τελευταία σχέση ακόμα μια φορά ως προς θ, έχουμε

$$\int (\frac{f'(x,\theta)}{f(x;\theta)} f(x,\theta))' dx = \int (\frac{f'(x,\theta)}{f(x;\theta)})' f(x,\theta)dx + \int (\frac{f'(x,\theta)}{f(x;\theta)})^2 f(x,\theta)dx = 0.$$

Τελικά έχουμε

$$E([\log f(x,\theta)]'') + E([\log f(x,\theta)]')^2 = 0$$

Από την (1.2) τελικά έχουμε

$$\mathrm{var}((\log f(x,\theta))') = -E[\log f(x,\theta)]''$$

Πρόταση 1.2

Αν κατανομή τ.μ. Y είναι εκθετικής μορφής (1) τότε ισχύουν τα εξής

1 $\mathrm{E}[\alpha(\mathrm{Y})] = -\frac{c'(\theta)}{b'(\theta)}$ (1.5)

2. $\mathrm{var}[a(Y)] = \frac{b''(\theta)c'(\theta) - c''(\theta)b'(\theta)}{[b'(\theta)]^3}$ (1.6)

Απόδειξη

1. Αν η f(y,θ) είναι εκθετικής μορφής τότε

$$\log f(y,\theta) = a(y)b(\theta) + c(\theta) + h(y)$$

και παραγωγίζοντας έχουμε

$$[\log f(x,\theta)]' = a(y)b'(\theta) + c'(\theta) \quad . \qquad (1.7)$$

Μπορεί να αποδειχθεί ότι μια οικογένεια κατανομών ικανοποιεί τις συνθήκες της Πρότασης 1.1. Ετσι από (1.3) έχουμε

$$E[\log f(x,\theta)]' = b'(\theta)E[a(y)] + c'(\theta) = 0.$$

Λύνοντας ως προς $E[a(y)]$

Έχουμε

$$\mathrm{E}[\alpha(\mathrm{Y})] = -\frac{c'(\theta)}{b'(\theta)}.$$

2. Παρομοίως

$$[\log f(x,\theta)]'' = a(y)b''(\theta) + c''(\theta)\ .$$

Από την (1.7) έχουμε

$$\mathrm{var}([\log f(x,\theta)]') = \mathrm{var}(a(y))[b'(\theta)]^2,$$

και από την (1.5)

$$E[\log f(x,\theta)]'' = E[a(y)]b''(\theta) + c''(\theta)$$
$$= -\frac{c'(\theta)}{b'(\theta)}b''(\theta) + c''(\theta).$$

Από την (1.6) έχουμε

$$\mathrm{var}(a(y))[b'(\theta)]^2 = \frac{c'(\theta)}{b'(\theta)}b''(\theta) - c''(\theta)\ .$$

Λύνοντας ως προς var(a(y)) τελικά πέρνουμε

$$\mathrm{var}[a(Y)] = \frac{b''(\theta)c'(\theta) - c''(\theta)b'(\theta)}{[b'(\theta)]^3}.$$

Όταν η εκθετική οικογένεια είναι στην τυπική της μορφή τότε οι παραπάνω δυο σχέσεις γράφονται ως

$$1 \quad \mathrm{E}[\mathrm{Y}] = -c'(\theta)$$
$$2.\ \mathrm{var}[Y] = -c''(\theta)$$

(Ασκηση 2)

Μπορεί επιπλέον να αποδεχθεί ότι για κάθε τάξης cumulant κ_r της τ.μ. Υ έχουμε ότι

$$k_r = -c^{(r)}(\theta).$$

Αν $Y_1, Y_2, ..., Y_n$ είναι ανεξάρτητες τυχαίες μεταβλητές που ακολουθούν την ίδια κατανομή (1.1) τότε η από κοινού κατανομή είναι

$$f(y_1, y_2, ..., y_n) = \prod_{i=1}^{n} \exp[b(\theta)a(y_i) + c(\theta) + d(y_i)]$$

$$= \exp[b(\theta)\sum_{i=1}^{n} a(y_i) + nc(\theta) + \sum_{i=1}^{n} d(y_i)]$$

Ο όρος $\sum_{i=1}^{n} a(y_i)$ είναι η επαρκής στατιστική συνάρτηση για την ποσότητα *b(θ).*

Αυτό σημαίνει ότι το άθροισμα $\sum_{i=1}^{n} a(y_i)$ συγκεντρώνει όλη την πληροφορία για την παράμετρο θ.

Σημαντικές ιδιότητες των κατανομών εκθετικής μορφής.

1) Ο παραμετρικός χώρος Θ για τις κατανομές τυπικής εκθετικής μορφής είναι κυρτό.

(Ασκηση 3)

2) Για τις ροπές $\int f(x)\exp[\sum \theta_j T_j(x)]d\mu(x)$ η παράγωγος ως προς θ_j μπορεί να περάσει κάτω από το σύμβολο της ολοκλήρωσης.

3) Για τις κατανομές της μορφής (1.2) παραμετρικοποιημένη με τη φυσική παράμετρο, ισχύει ότι

$$Cov(T_i(x), T_j(x)) = -\frac{\partial^2 c(\theta)}{\partial\theta_i \partial\theta_j}$$

(άσκηση 4)

4) Για τις κατανομές τυπικής εκθετικής μορφής η διακύμανση είναι μια αυστηρώς μονότονη συνάρτηση της μέσης τιμής.

(άσκηση 5)

Μπορούν να γίνουν διάφοροι χαρακτηρισμοί με βάση τη συνάρτηση διακύμανσης, όπως για παράδειγμα ότι η μόνη τυπική εκθετική οικογένεια κατανομών με σταθερά συνάρτηση διακύμανσης είναι η κανονική κατανομή.

5) Ο εκτιμητής μέγιστης πιθανοφάνειας για την φυσική παράμετρο θ υπάρχει πάντα και είναι μεροληπτικός εκτός από την κανονική κατανομή.

6) Η κανονική οικογένεια κατανομών έχει την ιδιότητα του μονότονου λόγου πιθανοφάνειας, μια ιδιότητα πολύ σημαντική για τον έλεγχο υποθέσεων.

1.4 Τα στοιχεία του Γενικευμένου γραμμικού μοντέλου.

Το Γ.Γ.Μ ορίζεται σε σχέση με ένα σύνολο από ανεξάρτητες τυχαίες μεταβλητές

$Y_1, Y_2 ..., Y_n$ όπου κάθε μία από αυτές ακολουθεί κατανομή εκθετικής μορφής με τις εξής ιδιότητες:

1. Η κατανομή κάθε μιας από τις Y_i είναι κανονικής μορφής και εξαρτάται από μια μόνο παράμετρο θ_t , έτσι

$$f(y_t) = \exp[b(\theta_t) y_i + c(\theta_t) + d(y_i)]$$

Η από κοινού πυκνότητα πιθανότητας είναι

$$f(y_1, y_2, ..., y_n; \theta_1, ..., \theta_n) = \exp[\sum_{i=1}^{n} y_i b_i(\theta_i) + \sum_{i=1}^{n} c(\theta_i) + \sum_{i=1}^{n} d(y_i)].$$

Οι παράμετροι που μας ενδιαφέρουν δεν είναι οι θ_t αφού μπορεί να είναι ένα θ για κάθε παρατήρηση. Για τα Γ.Γ.Μ. θεωρούμε ένα μικρότερο σύνολο παραμέτρων $\beta^{\mathrm{T}} = (\beta_1, ..., \beta_p)$ όπου φυσικά p<n.

Ενώ στο κλασσικό γραμμικό μοντέλο θεωρούμε ότι ισχύει η σχέση

$\mu_t = \mathrm{E}(\mathrm{Y}_t) = x_i^T \beta$ όπου $x_i = (x_{i1}, x_{i2}, ...x_{ip})$ το διάνυσμα των ανεξάρτητων μεταβλητών, στο Γ.Γ.Μ. θεωρούμε ότι υπάρχει μια μονότονη και διαφορίσημη συνάρτηση *g*, τέτοια ώστε

$$g(\mu_t) = x_i^T \beta \qquad (1.8)$$

Η συνάρτηση αυτή λέγεται συνάρτηση σύνδεσης (link function).

1.5 Λίγα περισσότερα για την συνάρτηση σύνδεσης

Όπως έχουμε πει η link συνάρτηση σχετίζει το γραμμικό κομμάτι $x_i^T \beta$ με τη μαθηματική ελπίδα $\mu_t = \mathrm{E}(\mathrm{Y}_t)$ μέσω της σχέσης (1.8).

Αυτός ο μετασχηματισμός είναι μία αναγκαιότητα γιατί σε πολλές εφαρμογές οι δυο αυτές συναρτήσεις δεν έχουν το ίδιο πεδίο τιμών.

Για παράδειγμα όταν ασχολούμαστε με απαριθμήσεις (counts) η y_i ακολουθεί την κατανομή Poisson και στην περίπτωση αυτή $\mu_t = \mathrm{E}(\mathrm{Y}_t) > 0$. Ετσι δεν μπορούμε να έχουμε $\mu_t = x_i^T \beta$ αφού στο αριστερό μέρος έχουμε τον περιορισμό $\mu_t > 0$ και στο δεξί ο περιορισμός αυτός δεν υπάρχει. Στην

περίπτωση αυτή η *g* πρέπει να απεικονίζει το διάστημα $\Re^+$ σε όλο την ευθεία $\Re$. Μία τέτοια συνάρτηση είναι η $g(\mu) = \ln(\mu)$.

Το κλασσικό γραμμικό μοντέλο $y = Xb + e$ είναι προφανώς μια απλή περίπτωση του Γ.Γ.Μ. αφού όλα τα y_i του διανύσματος *y* ακολουθούν την κατανομή $N(\mu_i,\sigma^2)$και $\mu_t = x_i \beta^{\mathrm{T}}$ Προφανώς $g(\mu_t) = \mu_t$.

Για την Δυωνυμική κατανομή, προφανώς δεν μπορούμε να γράψουμε $p_t = x_i \beta^{\mathrm{T}}$. Χρειαζόμαστε μια συνάρτηση g από το $\Re$ στο (0,1). (Ποια;)

1.5 . Εκτιμητική

Δυο από τις συνηθέστερες μεθόδους εκτίμησης είναι η εκτίμηση με τη μέθοδο της μέγιστης πιθανοφάνειας (ΕΜΠ) και η μέθοδο των ελαχίστων τετραγώνων.(Μ.Ε.Τ.)

Μέθοδος μέγιστης πιθανοφάνειας

Σύμφωνα με τη μέθοδο αυτή, ο εκτιμητής μέγιστης πιθανοφάνειας (ΕΜΠ) $\hat{\theta}$ της παραμέτρου *θ* είναι εκείνες οι τιμές οι οποίες μεγιστοποιούν την συνάρτηση πιθανοφάνειας

$L(\theta; y_1, y_2, ..., y_n) = \prod f(y_i; \theta)$ ή ισοδύναμα την λογαριθμική συνάρτηση πιθανοφάνειας

. $l(\theta; y_1, y_2, ..., y_n) = \sum_{i=1}^{n} \ln f(y_i; \theta)$

Συνήθως ο εκτιμητής $\hat{\theta}$ βρίσκεται παραγωγίζοντας την συνάρτηση $l(\theta; y_1, y_2, ..., y_n)$ σε σχέση με κάθε στοιχείο θ_t του *θ* και λύνοντας το σύστημα εξισώσεων

$$\frac{\partial l(\theta; y)}{\partial \theta_j} = 0 \ \gamma\iota\alpha \quad j = 1,2,..., p$$

Είναι πάντα αναγκαίο να ελέγξουμε ότι ο Hessian πίνακας των δευτέρων παραγώγων της $l(\theta; y_1, y_2, ..., y_n)$ είναι αρνητικά ορισμένος.

Ο εκτιμητής αυτός έχει ιδιότητες που τον κάνουν να υπερέχει έναντι των άλλων εκτιμητών. Μερικές από αυτές .είναι οι εξής

1. Αν $g(\theta)$ είναι μία συνάρτηση του θ τότε ο εκτιμητής μέγιστης πιθανοφάνειας του $g(\theta)$ είναι $g(\hat{\theta})$.
2. Συνέπεια (consistency)
3. Επάρκεια (Sufficiency)
4. Ασυμπτωτική αποτελεσματικότητα (Asymptotic efficiency)

Μέθοδος των ελαχίστων τετραγώνων

Εστω $Y_1, Y_2, ...Y_n$ είναι τυχαίες μεταβλητές τέτοιες ώστε $E(Y_i) = \mu_i$ για i=1,2,…n, και τα μ_i είναι συναρτήσεις των παραμέτρων $\beta^{\mathrm{T}} = (\beta_1, ..., \beta_p)$. Τότε για το γραμμικό μοντέλο $Y_i = \mu_i + e_i$ i=1,2,…n, η μέθοδος των ελαχίστων τετραγώνων ορίζεται σαν την τεχνική με την οποία επιχειρείται να εκτιμηθεί η παράμετρος β ελαχιστοποιώντας την ποσότητα

$$S = \sum e_i^2 = \sum (Y_i - \mu_i(\beta))^2 . \qquad (1.9)$$

Με τη βοήθεια των πινάκων η παραπάνω σχέση γράφεται στη μορφή

$$S = (y - \mu)^{\mathrm{T}} (y - \mu)$$

όπου

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \quad \text{και} \quad \mu = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_n \end{pmatrix}.$$

Συνήθως ο εκτιμητής $\hat{\beta}$ βρίσκεται παραγωγίζοντας τη συνάρτηση S σε σχέση με κάθε στοιχείο β_j και στη συνέχεια λύνοντας το σύστημα των εξισώσεων

$$\frac{\partial S}{\partial \beta_j} = 0 \quad j = 1, ..., p .$$

Φυσικά πρέπει να ελέγξουμε ότι η λύση αντιστοιχεί σε ελάχιστο (Δηλαδή ότι ο Hessian πίνακας των δευτέρων παραγώγων είναι θετικά ορισμένος). Στην πράξη

μπορεί να υπάρχει επιπλέον πληροφορία για τις τιμές του Y_i για παράδειγμα ότι κάποιες παρατηρήσεις είναι λιγότερο αξιόπιστες από κάποιες άλλες. Στην περίπτωση αυτή ίσως να χρειασθεί να σταθμίσουμε τους όρους στο άθροισμα (1.9) και αντί αυτού του αθροίσματος να ελαχιστοποιήσουμε το άθροισμα

$$S_W = \sum w_i (Y_i - \mu_i(\beta))^2$$

όπου τα w_i αποτελούν βάρη. Θα μπορούσαν για παράδειγμα να είναι

$$w_i = [\mathrm{Var}(Y_i)]^{-1}$$

Στη γενικότερη περίπτωση τα Y_i μπορεί να είναι συσχετισμένα. Αν V είναι ο πίνακας συνδιακύμανσης των Y_i , τότε ο εκτιμητής σταθμισμένων ελαχίστων τετραγώνων (ΣΤΕΕΤ) είναι το διάνυσμα β το οποίο ελαχιστοποιεί τη συνάρτηση

$$S_W = (y - \mu)^{\mathrm{T}} V^{-1} (y - \mu).$$

Αν $\mu = \mathrm{X}\beta$ για κάποιον πίνακα X διαστάσεων Nxp τότε

$$S_W = (y - X\beta)^{\mathrm{T}} V^{-1} (y - \mathrm{X}\beta)$$

Το διάνυσμα των παραγώγων της S_w ως προς το διάνυσμα β είναι το

$$\frac{\partial S_W}{\partial \beta} = -2X^{\mathrm{T}} V^{-1} (y - \mathrm{X}\beta)$$

(Ασκηση 6)

Ετσι ο ΣΤΕΕΤ b είναι η λύση της <u>κανονικής εξίσωσης</u>

$$X^{\mathrm{T}} V^{-1} \mathrm{X}\beta = X^{\mathrm{T}} V^{-1} y$$

Μπορεί εύκολα να αποδειχθεί ότι ο Hessian πίνακας είναι θετικά ορισμένος (Ασκηση 7). Το πλεονέκτημα του ΣΤΕΕΤ είναι ότι υπάρχει πάντα και αντιστοιχεί σε τοπικό ελάχιστο. Επιπλέον η εύρεσή του απαιτεί τη γνώση μόνο των δυο πρώτων ροπών του διανύσματος Y, και καμιά άλλη υπόθεση για την κατανομή του. Τουναντίον ο ΕΜΠ απαιτεί τη γνώση της κατανομής της Y και μερικές φορές η εύρεσή του είναι αδύνατη γιατί απαιτεί τη λύση ενός πολύπλοκου μη γραμμικού συστήματος εξισώσεων. Ο ΕΜΠ πλεονεκτεί έναντι του ΣΤΕΕΤ στο ότι είναι ασυμπτωτικά αποτελεσματικός.

Στη συνέχεια αποδεικνύομε ότι για τα ΓΓΜ οι δυο αυτοί εκτιμητές ταυτίζονται.

Θα χρησιμοποιήσουμε τους εξής συμβολισμούς

$$X=\begin{pmatrix} x_{11} & \cdots & x_{1i} & \cdots & x_{1N} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{j1} & \cdots & x_{ji} & \cdots & x_{jN} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{p1} & \cdots & x_{pj} & \cdots & x_{pN} \end{pmatrix}, \qquad Y-\mu=\begin{pmatrix} y_1-\mu_1(\theta) \\ \vdots \\ y_j-\mu_j(\theta) \\ \vdots \\ y_N-\mu_N(\theta) \end{pmatrix},$$

$$V=\begin{pmatrix} Var(Y_1) & \cdots & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & Var(Y_k) & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & 0 & \cdots & Var(Y_N) \end{pmatrix} \quad \text{και} \quad D=\begin{pmatrix} \frac{\partial\mu_1}{\partial\eta_1} & \cdots & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & \frac{\partial\mu_i}{\partial\eta_i} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & 0 & \cdots & \frac{\partial\mu_N}{\partial\eta_{Ni}} \end{pmatrix}$$

Επιπλέον αν $U_j=\frac{\partial l(\theta;y)}{\partial\beta_j}=\sum_{i=1}^{N}\frac{\partial l_i}{\partial\beta_j}$ τότε η score συνάρτηση γράφεται σαν

$$U=\begin{pmatrix} U_1 \\ \vdots \\ U_j \\ \vdots \\ U_p \end{pmatrix}$$

Λήμμα 1.

1) Η score για την εκτίμηση των παραμέτρων $\beta=(\beta_1,\beta_2,....\beta_p)^T$ είναι η

$$U=XVD(Y-\mu)$$

2) Ο πληροφοριακός πίνακας του Fisher $I(\theta)$ είναι

$$I(\theta)=\mathrm{X}VD^2VX^T$$

Απόδειξη

1)Αν $Y_1,Y_2,...Y_N$ είναι παρατηρήσεις των ανεξαρτήτων μεταβλητών τότε η λογαριθμική συνάρτηση πιθανοφάνειας είναι

$$l(\theta; y) = \sum y_i b(\theta_i) + \sum c(\theta_i) + \sum d(y_i)$$

όπου από την Πρόταση 1.2 έχουμε ότι

$$E(Y_i) = \mu_i = -c'(\theta_i) / b'(\theta_i) .$$

επίσης από την υπόθεση του μοντέλου

$$g(\mu_i) = x_i^T \beta = \sum_j^p x_{ij} \beta_j = \eta_i \qquad (1.10)$$

όπου g είναι μία μονότονη και διαφορήσιμη συνάρτηση.

Η score συνάρτηση σε σχέση με το β_j ορίζεται σαν

$$U_j = \frac{\partial l(\theta; y)}{\partial \beta_j} = \sum_{i=1}^{N} \frac{\partial l_i}{\partial \beta_j}$$

όπου

$l_i = y_i b(\theta_i) + c(\theta_i) + d(y_i)$.

Για την εύρεση της U_j χρησιμοποιούμε τη σχέση

$$\frac{\partial l_i}{\partial \beta_j} = \frac{\partial l_i}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \beta_j} \qquad (1.11)$$

Στη συνέχεια θα βρούμε κάθε έναν από τους τρεις όρους του γινομένου στο δεξί μέλος.

Παραγωγίζοντας την l_i συνάρτηση ως προς θ_i , βρίσκουμε

$$\frac{\partial l_i}{\partial \theta_i} = y_i b'(\theta_i) + c'(\theta_i) = b'(\theta_i)(y - \mu_i) .$$

Στη συνέχεια για την παράγωγο $\frac{\partial \theta_i}{\partial \mu_i}$

αρκεί να παραγωγίσουμε την συνάρτηση μ_i ως προς θ_i . Ετσι

$$\frac{\partial \mu_i}{\partial \theta_i} = -\frac{c''(\theta_i)}{b'(\theta_i)} + \frac{c'(\theta_i) b''(\theta_i)}{[b'(\theta_\iota)]^2} = b'(\theta_i) \operatorname{var}(Y_i) .$$

Παραγωγίζοντας την συνάρτηση (1.10) έχουμε

$$\frac{\partial \mu_i}{\partial \beta_j} = \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j} = x_{ij} \frac{\partial \mu_i}{\partial \eta_i}$$

Αντικαθιστώντας στο δεξί μέλος της (1.11) τις αντίστοιχες παραγώγους έχουμε

$$\frac{\partial l_i}{\partial \beta_j} = \frac{\partial l_i}{\partial \theta_i}\frac{\partial \mu_i}{\partial \beta_j} / (\frac{\partial \mu_i}{\partial \theta_i}) = \frac{(y_i - \mu_i)x_{ij}}{\text{var}(Y_i)}(\frac{\partial \mu_i}{\partial \eta_i})$$

και τελικά

$$U_j = \sum_{i=1}^{N} \frac{(y_i - \mu_i)x_{ij}}{\text{var}(Y_i)}(\frac{\partial \mu_i}{\partial \eta_i}) \qquad \text{j=1,2,…..p.}$$

Τελικά το σύστημα των εκτιμητριών συναρτήσεων γράφεται υπό μορφή πίνακα

$$U = XDV^{-1}(Y - \mu)$$

(2) Για την εύρεση του Πληροφοριακού πίνακα του Fisher χρησιμοποιούμε την σχέση

$$I = E(UU^T) = XDV^{-1}E(Y-\mu)(\mathrm{Y}-\mu)^{\mathrm{T}}VD^T X^T$$
$$= XDVD^T X^T \quad .$$

Οι εξισώσεις U_j =0 είναι μη γραμμικές και γι' αυτό το λόγο επιλύονται με αριθμητικές μεθόδους. Η συνηθέστερη μέθοδος είναι η Newton-Rapshon της οποίας το επαναληπτικό σχήμα είναι το εξής¨

$$b^{(m)} = b^{(m-1)} - \left[\frac{\partial^2 l}{\partial \beta_i \partial \beta_k}\right]^{-1}_{\beta = b^{(m-1)}} U^{(m-1)}$$

όπου $\left[\frac{\partial^2 l}{\partial \beta_i \partial \beta_k}\right]_{\beta = b^{(m-1)}}$

είναι ο πίνακας των δευτέρων παραγώγων της l στην τιμή $\beta = b^{(m-1)}$ και $U^{(m-1)}$ είναι το διάνυσμα των πρώτων παραγώγων $(U_1, U_2, ..., U_p)^T$ υπολογισμένο στην τιμή $\beta = b^{(m-1)}$

Μια έκδοση της παραπάνω μεθόδου μερικές φορές απλούστερη είναι η μέθοδος scoring στην οποία ο Hessian πίνακας των δευτέρων παραγώγων αντικαθίσταται από τον πληροφοριακό πίνακα του Fisher

$$I = E[UU^T]$$

με στοιχεία $I_{kj} = E[U_{kj}] = E[\frac{\partial l}{\partial b_k}\frac{\partial l}{\partial bj}] = -E[\frac{\partial^2 l}{\partial b_k \partial b_j}]$

δηλαδή ο αλγόριθμος της μεθόδου scoring είναι

$$b^{(m)} = b^{(m-1)} + \left[I^{(m-1)}\right]^{-1} U^{(m-1)}$$

Αν πολλαπλασιάσουμε και τις δυο πλευρές της τελευταίας εξίσωσης με $I^{(m-1)}$ Έχουμε

$$I^{(m-1)} b^{(m)} = I^{(m-1)} b^{(m-1)} + U^{(m-1)}.$$

Για τα γενικευμένα γραμμικά μοντέλα ισχύει ότι

Ετσι ο πίνακας I μπορεί να γραφεί στη μορφή

$$I=X^{T}WX$$

Όπου W είναι ο NxN διαγώνιος πίνακας με στοιχεία

$$\frac{1}{\mathrm{var}(Y_i)}\left(\frac{\partial \mu_i}{\partial \eta_t}\right)^2$$

Ετσι η εξίσωση της score συνάρτησης γίνεται

$$X^T W X b^{(m)} = X^T W z$$

Όπου z είναι στήλη με στοιχεία

$$z_i = \sum_k x_{ik} b_k^{(m-1)} + (y_i - \mu_i)\frac{\partial \mu_i}{\partial \eta_t}$$

Η τελευταία μορφή είναι των σταθμισμένων ελαχίστων τετραγώνων.

<u>ΚΑΤΑΝΟΜΗ ΤΟΥ ΕΜΠ</u>

Υποθέτουμε ότι η λογαριθμική συνάρτηση πιθανοφάνειας έχει μοναδικό μέγιστο το b και ότι αυτός ο εκτιμητής είναι κοντά στην παράμετρο β. Το ανάπτυγμα Taylor πρώτης τάξης του διανύσματος U(β) στο σημείο β=b είναι

$$U(\beta) = U(b) + H(b)(\beta - b)$$

Όπου H(b) είναι ο πίνακας των δευτέρων παραγώγων της λογαριθμικής συνάρτησης πιθανοφάνειας στο σημείο β= b. Ασυμπτωτικά ο πίνακας H ισούται με τον πληροφοριακό πίνακα

$$I = E[UU^T] = E[-H]$$

Για τον λόγο αυτό για μεγάλα δείγματα έχουμε

$$U(\beta) \cong U(b) - I(\beta - b)$$

Αλλά U(b)=0 γιατί b μεγιστοποιεί την λογαριθμική συνάρτηση πιθανοφάνειας άρα μηδενίζει την παράγωγό της. Προσεγγιστικά λοιπόν

$$(b - \beta) \cong \mathrm{I}^{-1}U$$

Υπό την προυπόθεση ότι ο πίνακας I είναι μη-μηδενικός, συμπεραίνουμε ότι ασυμπτωτικά

$$E(b - \beta) = \mathrm{I}^{-1}\mathrm{E}(U) = 0 \ ,$$

καθώς και

$$E(b - \beta)(b - \beta)^T \cong \mathrm{I}^{-1}\mathrm{E}(UU^T)I^{-1} = I^{-1}$$

Ετσι για μεγάλα n

$$(b - \beta) \ \cong N(0, \mathrm{I}^{-1})$$

Καθώς και

$$(b - \beta)I(b - \beta)^T \cong \chi_p^2$$

Η στατιστική $(b - \beta)I(b - \beta)^T$ καλείται στατιστική του Wald.

<u>Η επάρκεια του μοντέλου</u>

Ας υποθέσουμε ότι θέλουμε να ελέγξουμε την επάρκεια της προσαρμογής ενός μοντέλου σε ένα σύνολο δεδομένων. Αυτό μπορεί να γίνει συγκρίνοντας την συνάρτηση πιθανοφάνειας αυτού του μοντέλου αυτού με τη συνάρτηση πιθανοφάνειας του μέγιστου μοντέλου το οποίο περιγράφεται ως εξής

1. Το μέγιστο μοντέλο είναι ένα γενικευμένο γραμμικό μοντέλο με την ίδια κατανομή όπως το μοντέλο που μας ενδιαφέρει.
2. Το μέγιστο μοντέλο έχει την ίδια συνάρτηση σύνδεσης με το μοντέλο που μας ενδιαφέρει.

3. Ο αριθμός των παραμέτρων στο μέγιστο μοντέλο ισούται με τον αριθμό των παρατηρήσεων.

Λόγω της 3 μπορεί να θεωρηθεί ότι το μέγιστο μοντέλο περιγράφει πλήρως τα δεδομένα.

Οι συναρτήσεις πιθανοφάνειας υπολογίζονται στον εκτιμητή μέγιστης πιθανοφάνειας $b_{\max}$ και b αντίστοιχα και λαμβάνουμε $L(b_{\max}; y)$ και $L(b; y)$ αντίστοιχα. Αν το μοντέλο που μας ενδιαφέρει περιγράφει τα δεδομένα ικανοποιητικά, τότε $L(b; y)$ πρέπει να είναι κοντά στο $L(b_{\max}; y)$. Τουναντίον αν το μοντέλο δεν είναι ικανοποιητικό τότε το $L(b; y)$ πρέπει να είναι μικρότερο από το $L(b_{\max}; y)$. Αυτό μας οδηγεί στην χρήση του Γενικευμένου λόγου πιθανοφάνειας

$$\lambda = \frac{L(b_{\max}; y)}{L(b; y)}$$

ή ισοδύναμα τον λογάριθμο αυτής

$$\log \lambda = \log(L(b_{\max}; y) - \log(L(b_{\max}; y)) = l(b_{\max}; y) - l(b_{\max}; y)$$

σαν ένα μέτρο καλής προσαρμογής του μοντέλου.

Μεγάλες τιμές του logλ είναι ένδειξη μη καλής προσαρμογής του μοντέλου. Για να βρούμε την κριτική περιοχή του logλ πρέπει να βρούμε τη δειγματική κατανομή του.

Δειγματική κατανομή της λογαριθμικής συνάρτησης πιθανοφάνειας

Η λογαριθμική συνάρτηση πιθανοφάνειας ορίζεται ως

$$D = 2[l(b_{\max}; y) - l(b; y)$$

Οι Nelder και Wedderburn (1972) κάλεσαν την συνάρτηση αυτή **deviance.**

Η συνάρτηση αυτή μπορεί να γραφτεί στη μορφή

$$\begin{aligned} D = 2\{&[l(b_{\max}; y) - l(\beta_{\max}; y)] \\ &-[l(b; y) - l(\beta; y)] \\ &+[l(\beta_{\max}; y) - l(\beta; y)]\} \end{aligned}$$

Ο πρώτος όρος στις αγκύλες ακολουθεί την X^2_N και ο δεύτερος την X^2_p κατανομή. Ο τρίτος όρος είναι μία θετική σταθερά που είναι κοντά στο μηδέν όταν

το μοντέλο με p παραμέτρους περιγράφει το μοντέλο όπως το μέγιστο μοντέλο. Σε γενικές γραμμές μπορούμε να πούμε ότι όταν οι δυο πρώτοι όροι είναι ανεξάρτητοι και ο τρίτος όρος είναι κοντά στο μηδέν τότε

$D \sim X^2_{N-p}$.

Όταν το μοντέλο δεν είναι ικανοποιητικό τότε η Deviance ακολουθεί προσεγγιστικά τη μη κεντρική X^2 κατανομή.

Παράδειγμα 1

Υποθέτουμε ότι οι μεταβλητές $Y_1, Y_2, ..., Y_N$ είναι ανεξάρτητες και ακολουθούν την κανονική κατανομή με μέσο μ_i και κοινή τυπική απόκλιση σ. Η λογαριθμική συνάρτηση πιθανοφάνειας είναι

$$l(\beta; y) = -\frac{1}{2\sigma^2}\sum_{\iota=1}^{N}(y_i - \mu_\iota)^2 - \frac{1}{2}\mathrm{N}\log(2\pi\sigma^2)$$

Για το μέγιστο μοντέλο $E(Y_i)= \mu_i$ όπου i=1,2,…N. Παραγωγίζοντας και ακολουθώντας την συνήθη διαδικασία βρίσκουμε ότι $\hat{\mu}_i = y_i$.

Για το λόγο αυτό

$$l(b_{\max}; y) = -\frac{1}{2}N\log(2\pi\sigma^2).$$

Θεωρούμε τώρα το μοντέλο στο οποίο όλα τα Y_i έχουν τον ίδιο μέσο μ. Τότε $\hat{\mu} = \bar{y}$ και

$$l(b; y) = -\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - \bar{y})^2 - \frac{1}{2}N\log(2\pi\sigma^2).$$

Τελικά

$$D = 2[l(b_{\max}; y) - l(b; y) = \frac{1}{\sigma^2}\sum_{i=1}^{N}(y_i - \bar{y})^2$$

Η στατιστική συνάρτηση D σχετίζεται με τη δειγματική διασπορά

$S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{y})^2$. Αν το μοντέλο στο οποίο τα Y_i ακολουθούν την $N(\mu,\sigma^2)$ είναι το σωστό τότε η $D = \frac{(N-1)}{\sigma^2}S^2$ ακολουθεί την X^2_{N-1} κατανομή.

Παράδειγμα 2. Αν $Y_1, Y_2, ..., Y_N$ είναι ανεξάρτητες και ακολουθούν την κατανομή Poisson με παράμετρο λ_i.

Η Deviance για το μονοπαραμετρικό μοντέλο είναι

$$D = 2\sum y_i \log(y_i / \bar{y})$$

(Ασκηση 8)

Ελεγχος υποθέσεων

Ο έλεγχος υποθέσεων για την παράμετρο β μπορεί να γίνει με την βοήθεια της ασυμπτωτικής δειγματικής κατανομής των εκτιμητών b για την οποία έχουμε δει ότι $b \sim N(\beta, I^{-1})$ η ισοδυνάμως με τη στατιστική συνάρτηση του Wald $(b-\beta)I(b-\beta)$ η οποία έχει την X_p^2 κατανομή. Μια διαφορετική προσέγγιση του προβλήματος είναι με τη βοήθεια της συνάρτησης Deviance. Εστω ότι η μηδενική υπόθεση H_0 και Εναλλακτική H_α είναι οι

$$H_0 : \beta = \beta_0 = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_q \end{bmatrix} \quad \text{και} \quad H_a : \beta = \beta_a = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \text{ όπου } q<p<N.$$

Μπορούμε να ελέγξουμε την H_0 έναντι της H_α χρησιμοποιώντας την διαφορά στις στατιστικές Deviance. Δηλαδή

$$\Delta D = D_0 - D_\alpha = 2[l(b_{max}; y) - l(b_0; y)] - 2[l(b_{max}; y) - l(b_1; y)]$$
$$= 2[l(b_1; y) - l(b_0; y)]$$

Αν και τα δυο μοντέλα περιγράφουν τα δεδομένα καλά τότε $D_0 \sim \chi^2_{N-q}$ και $D_0 \sim \chi^2_{N-p}$ Ετσι ώστε κάτω από κάποιες συνθήκες ανεξαρτησίας $\Delta D \sim \chi^2_{p-q}$ Αν η τιμή της ΔD είναι συνεπής με την χ^2_{p-q} κατανομή επιλέγουμε το μοντέλο H_0 γιατί είναι το απλούστερο (με τις λιγότερες μεταβλητές).

2. ΔΥΑΔΙΚΕΣ ΜΕΤΑΒΛΗΤΕΣ ΚΑΙ ΛΟΓΙΣΤΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ

Στο κεφάλαιο αυτό μελετάμε ΓΓΜ στα οποία τα αποτελέσματα μετρώνται σε δυαδική κλίμακα. Τέτοια δεδομένα εμφανίζονται σε ιατρικά πειράματα όπου στο τέλος κάθε πειράματος ο ασθενής είτε ανένηψε (Y=1) είτε κατέληξε (Y=0). Μπορούμε λοιπόν να γράψουμε

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

για τις πιθανότητες της 'Αποτυχίας ' και 'Επιτυχίας ' αντίστοιχα.

Για παράδειγμα τα αποτελέσματα που μας ενδιαφέρουν μπορεί να είναι ``ζωντανός ή νεκρός `` και γενικότερα επιτυχία ή αποτυχία.

Στην περίπτωση αυτή για να συσχετίσουμε την πιθανότητα π_i με τη γραμμική έκφραση

$$\eta_i = \sum_{j=1}^{p} x_{ij}\beta_j$$

Πρέπει να χρησιμοποιήσουμε ένα γραμμικό μετασχηματισμό g(π) ο οποίος απεικονίζει το διάστημα (0,1) σε όλη την ευθεία (-∞, ∞). Υπάρχει μια μεγάλη ποικιλία από τέτοιες συναρτήσεις σύνδεσης. Τρεις όμως είναι αυτές που χρησιμοποιούνται στην πράξη.

Η λογιστική συνάρτηση

$g_1(\pi) = \log \frac{\pi}{1-\pi}$,

η probit ή αντίστροφη κανονική συνάρτηση

$g_2(\pi) = \Phi^{-1}(\pi)$,

και η συμπληρωματική log-log συνάρτηση

$g_3(\pi) = \log\{-\log(1-\pi)\}$.

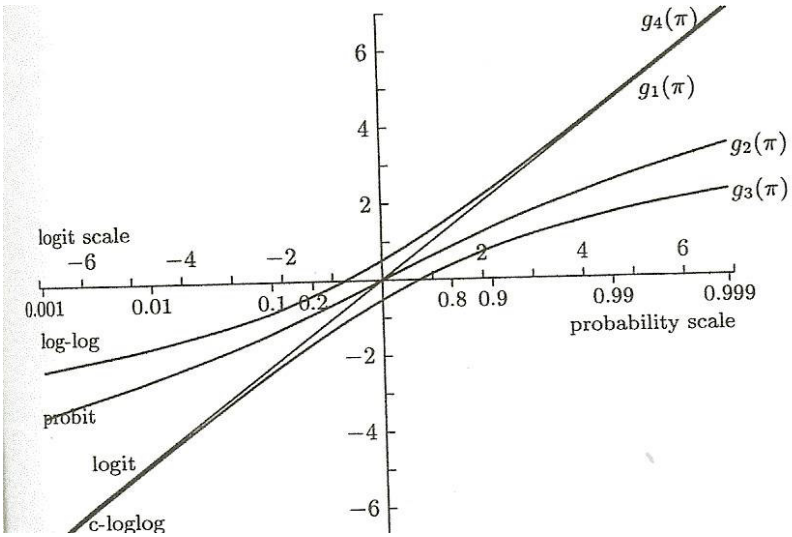
Η συνάρτηση

$g_4(\pi) = -\log\{-\log(\pi)\}$

Δεν χρησιμοποιείται συχνά γιατί δεν συμπεριφέρεται καλά για π<1/2 .

Και οι τέσσερεις συναρτήσεις είναι αντίστροφες συναρτήσεις γνωστών αθροιστικών συναρτήσεων κατανομών (ποιών;).

Σχήμα 1



Από το παραπάνω σχήμα παρατηρούμε τα εξής

- Η logit και η Probit σχετίζονται σχεδόν γραμμικά για τιμές του π στο διάστημα 0,1≤π≤0,9. Για τον λόγο αυτό είναι δύσκολη η διάκριση μεταξύ των δυο αυτών συναρτήσεων όταν πρόκειται για ζητήματα καλής προσαρμογής.
- Για μικρές τιμές του π, η συμπληρωματική log-log συνάρτηση είναι κοντά στην λογιστική συνάρτηση.

- Όταν π τείνει στο 1 τότε η συμπληρωματική log-log συνάρτηση τείνει στο άπειρο πολύ πιο αργά σε σύγκριση με τις άλλες τρεις συναρτήσεις.
- Παρομοίως η πιο αργή συνάρτηση στην περιοχή του 0 είναι η log-log.

Όλα τα ασυμπτωτικά αποτελέσματα που θα παρουσιαστούν εδώ ισχύουν ανεξαρτήτως της επιλογής της συνάρτησης σύνδεσης. Θα επικεντρωθούμε κυρίως στη λογιστική συνάρτηση για δυο κυρίως λόγους.

- Τα αποτελέσματα της ανάλυσης ερμηνεύονται εύκολα.
- Μπορούμε να ενσωματώσουμε στην ανάλυσή μας και στοιχεία που επιλέγονται retrospectively.

2.1. Παραμετρική εκτίμηση

Υιοθετώντας το λογιστικό μοντέλο με δυο covariates έχουμε τη σχέση

$$\log(\frac{\pi}{1-\pi}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \qquad (2.1)$$

Ο λόγος των πιθανοτήτων της επιτυχίας π προς την πιθανότητα της αποτυχίας 1-π, δηλαδή ο $\frac{\pi}{1-\pi}$ είναι σημαντικός στην λογιστική ανάλυση και λέγεται odds.

Αν π είναι η πιθανότητα επιτυχίας τότε ο λόγος $\frac{\pi}{1-\pi}$ είναι ο λόγος των πιθανοτήτων επιτυχίας προς αποτυχίας. Δηλαδή όταν λέμε ότι τα odds είναι 2 εννοούμε ότι η πιθανότητα επιτυχίας είναι διπλάσια της πιθανότητας αποτυχίας.
Επιπλέον ο λόγος αυτός παίζει σημαντικό ρόλο στην ερμηνεία των συντελεστών της λογιστικής παλινδρόμησης.
Η ερμηνεία των συντελεστών είναι η εξής :
Κρατώντας τη μεταβλητή x_1 σταθερή τότε μία μεταβολή της x_2 κατά μία μονάδα αυξάνει τον λογάριθμο του odd κατά μια ποσότητα β_2 .
Πράγματι αν στην παραπάνω εξίσωση (2.1)
αυξηθεί η μεταβλητή x_2 κατά μία μονάδα έχουμε

$$\log(\frac{\pi'}{1-\pi'}) = \beta_0 + \beta_1 x_1 + \beta_2 (x_2 + 1) .$$

Αφαιρώντας την (2.2) από την (2.2) έχουμε ότι

$$\beta_2 = \log(\frac{\pi'}{1-\pi'}) - \log(\frac{\pi}{1-\pi}) = \log(\frac{\pi'/1-\pi'}{\pi/1-\pi})$$

Δηλαδή κρατώντας τις υπόλοιπες μεταβλητές σταθερές και αυξάνοντας την x_2 κατά μία μονάδα ο λογάριθμος των odds αυξάνει κατά β_2

Όταν η μεταβλητή x_2 είναι ψευδομεταβλητή τότε η ερμηνεία του συντελεστή β_2 είναι παρεμφερής με αυτή για τα γραμμικά μοντέλα. Δηλαδή η μετάβαση από την κατάσταση x_2 =0 στην κατάσταση x_2 =1 με όλες οι άλλες μεταβλητές σταθερές, αυξάνει τον λογάριθμο των odds κατά β_2 μονάδες.

Η ερμηνεία σε σχέση με την πιθανότητα π δεν είναι εξίσου εύκολη. Λύνοντας ως προς π έχουμε

$$\pi = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2)}$$

Παρατηρούμε ότι τα αποτελέσματα στο π μιας μοναδιαίας μεταβολής του x_2 εξαρτάται τόσο από το x_2 όσο και από το x_1. Η παράγωγος όμως της π ως προς x_2 δίνει

$\frac{\partial \pi}{\partial x_2} = \pi(1-\pi)\beta_2$. Από εδώ βλέπουμε ότι μια μικρή μεταβολή στη x_2 έχει τη μεγαλύτερη μεταβολή στη π για τιμές της π κοντά στο 0.5 και τη μικρότερη μεταβολή κοντά στα άκρα 0 και 1.

<u>Ειδικές περιπτώσεις</u>

Όταν υπάρχει μία ανεξάρτητη dummy μεταβλητή τότε το μοντέλο μπορεί να περιγραφεί με τον επόμενο πίνακα συνάφειας

	X=1	X=0
Y=1	$\pi(1) = \frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}$	$\pi(0) = \frac{e^{\beta_0}}{1+e^{\beta_0}}$
Y=0	$1-\pi(1) = \frac{1}{1+e^{\beta_0+\beta_1}}$	$1-\pi(0) = \frac{1}{1+e^{\beta_0}}$
Σύνολο	1.0	1.0

Ο λόγος των odds είναι

$g(1) = \log\{\pi(1)/[1-\pi(1)]\}$

και

$$g(0) = \log\{\pi(0)/[1-\pi(0)]\}$$

Ο λόγος των odds ορίζεται ως

$$\psi = \frac{\pi(1)/[1-\pi(1)]}{\pi(0)/[1-\pi(0)]}$$

Και ο λογάριθμος αυτών είναι

$$\log(\psi) = \log[\frac{\pi(1)/[1-\pi(1)]}{\pi(0)/[1-\pi(0)]}$$

Χρησιμοποιώντας τις εκφράσεις του πίνακα συνάφειας έχουμε

$$\psi = \frac{(\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}})(\frac{1}{1+e^{\beta_0}})}{(\frac{e^{\beta_0}}{1+e^{\beta_0}})(\frac{1}{1+e^{\beta_0+\beta_1}})}$$

$$= \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = e^{\beta_1}$$

Ετσι για τη λογιστική παλινδρόμιση με μια διχότομη ανεξάρτητη μεταβλητή έχουμε

$$\psi = e^{\beta_1}$$

Ή ισοδύναμα

$$\log(\psi) = \beta_1$$

Με άλλα λόγια όταν η ανάλυση παλινδρόμησης έχει σαν ανεξάρτητη μεταβλητή μία μόνο κατηγορική μεταβλητή, τότε η λογιστική παλινδρόμηση στην ουσία ισοδυναμεί με ανάλυση πίνακα κατηγοριών.

Η σπουδαιότητα της παράστασης ψ είναι εμφανής. Αν η τιμή του Ψ είναι κοντά στο 1 (ή ισοδύναμα η τιμή λογΨ κοντά στο 0 τότε η μεταβλητή Χ δεν έχει μεγάλη προβλεπτική ικανότητα αφού στις δυο ομάδες Χ=0 και Χ=1 η λόγος πιθανότητας επιτυχίας προς αποτυχία είναι ο ίδιος.

Τα βασικά αποτελέσματα τα οποία παράγονται από όλα τα στατιστικά πακέτα παρουσιάζονται στο επόμενο παράδειγμα.

Για τη μελέτη της επίδρασης της ηλικίας (Χ) στην αρτηριακή πίεση Υ μελετώνται 100 άτομα. Ο πίνακας συνάφειας είναι ο εξής

	Ηλικία (έτη)		
Πίεση	55>=	55<	Σύνολο
Υ=1(παρουσία)	21	22	43
Υ=0(απουσία)	6	51	57
Σύνολο	27	73	100

Τα αποτελέσματα της προσαρμογής της Λογιστικής παλινδρόμισης στα δεδομένα είναι τα εξής

Μεταβλητές	Εκτιμητές συντελεστών	Τυπική απόκλιση	Εκτιμητές /τυπική απόκλιση	$\widehat{\Psi}$
Ηλικία	2.094	0.529	3.96	8.1
Σταθερά	-0.841	0.255	-3.30	

Η ποσότητα στη στήλη $\widehat{\Psi}$ είναι το εκτιμητής μέγιστης πιθανοφάνειας του Λόγου odd $\widehat{\Psi}$ =$e^{2.094}$=8.1.

Ο λόγος Ψ είναι ένας σημαντικός παράγοντας στη λογιστική παλινδρόμιση. Η δειγματική κατανομή της $\widehat{\Psi}$ όμως είναι λοξή επειδή είναι φραγμένη μακριά από το 0. Από θεωρητικής πλευράς, για μεγάλα n η κατανομή της $\widehat{\Psi}$ προσεγγίζει την κανονική κατανομή. Δυστυχώς όμως το δείγμα που απαιτείται για μια τέτοια προσέγγιση είναι πολύ μεγάλο. Για τον λόγο αυτό οτιδήποτε συμπεράσματα εξάγονται για την $\widehat{\Psi}$ είναι μέσω της $\ln\widehat{\Psi}=\hat{\beta}_1$ η οποία είναι κοντά στη κανονική κατανομή για πολύ μικρότερα δείγματα. Για το 95% διάστημα εμπιστοσύνης της $\widehat{\Psi}$ βρίσκουμε πρώτα το διάστημα εμπιστοσύνης $\hat{\beta}_1 \pm z_{1-a/2}\times S\hat{E}(\hat{\beta}_1)$ της β_1 από το οποίο παίρνουμε το διάστημα εμπιστοσύνης $\exp[\hat{\beta}_1 \pm z_{1-a/2}\times S\hat{E}(\hat{\beta}_1)]$ της $\widehat{\Psi}=\exp(\hat{\beta}_1)$. Ετσι για το παράδειγμα το 95% διάστημα εμπιστοσύνης της $\widehat{\Psi}$ είναι $\exp[2{,}094\pm1{,}96\times0{,}529]$=(2,9 , 22,9).

Ο έλεγχος της υπόθεσης $H_0:\beta_1=0 \quad vs \quad H_1:\beta_1\neq 0$ είναι ισοδύναμος με το αν το διάστημα εμπιστοσύνης περιέχει το 0. Αναφορικά με το $\widehat{\Psi}$ μας ενδιαφέρει η υπόθεση $H_0:\widehat{\Psi}=1\, vs \quad H_1:\widehat{\Psi}\neq 1$ ο οποίος είναι ισοδύναμος με το αν το διάστημα εμπιστοσύνης περιέχει το 1.

2.3. Μέθοδοι καλής προσαρμογής για την λογιστική παλινδρόμηση

Όπως έχουμε πει και στο γενικό μέρος ένα μέτρο καλής προσαρμογής για το μοντέλο είναι η Deviance

$$D = 2[l(\hat{\pi}_{\max}; y) - l(\hat{\pi}; y)]$$

όπου $\hat{\pi}_{\max}$ ο ΕΜΠ για το μέγιστο μοντέλο και $\hat{\pi}$ ο ΕΜΠ για το μοντέλο που μας ενδιαφέρει.

Στην περίπτωση αυτή μπορεί να αποδειχθεί ότι

$$D = 2\sum_{i=1}^{N}[y_i \log(\frac{y_i}{n_i\hat{\pi}_t}) + (n_i - y_i)\log(\frac{n_i - y_i}{n_i - n_i\hat{\pi}_i})]$$

(Ασκηση 10)

Ετσι η D γράφεται

$$D = 2\sum o \log\frac{o}{e}$$

όπου ο είναι οι παρατηρούμενες συχνότητες y_i και (n_i-y_i) οι αναμενόμενες συχνότητες $n_i\hat{\pi}_t$ και $n - n_i\hat{\pi}_t$ και η υπόθεση μπορεί να ελεγχθεί μέσω της προσέγγισης $D \sim \chi^2_{N-p}$ όπου p ο αριθμός των β παραμέτρων.

Αντί να χρησιμοποιήσουμε τον ΕΜΠ μπορούμε να εκτιμήσουμε τις παραμέτρους ελαχιστοποιώντας το σταθμισμένο άθροισμα τετραγώνων

$$S_w = \sum_{i=1}^{N}\frac{(y_i - n_i\pi_t)^2}{n_i\pi_t(1-\pi_t)} \ .$$

Αυτό είναι ισοδύναμο με την ελαχιστοποίηση της X^2 στατιστικής του Pearson

$$X^2 = \sum\frac{(0-e)^2}{e} \ .$$

Πράγματι

$$X^2 = \sum\frac{(y_i - n_i\pi_t)^2}{n_i\pi_t} + \sum\frac{[(n_i - y_i) - n_i(1-\pi_t)]^2}{n_i(1-\pi_t)}$$

$$= \sum\frac{(y_i - n_i\pi_t)^2}{n_i\pi_t(1-\pi_t)}(1 - \pi_t + \pi_t) = S_w.$$

Οταν η X^2 υπολογίζεται στις αναμενόμενες συχνότητες τότε η στατιστική είναι η

$$X^2 = \sum_{i=1}^{N}\frac{(y_i - n_i\hat{\pi}_t)^2}{n_i\hat{\pi}_t(1-\hat{\pi}_t)}$$

η οποία είναι ασυμπτωτικά ισοδύναμα με την στατιστική D.

Πράγματι χρησιμοποιώντας την ταυτότητα

$s\,\mathrm{Log}\,{}^{s}\!/_{t} = (s-t)+\frac{1}{2}\frac{(s-t)^2}{t}+\ldots$ έχουμε

$$D = 2\sum_{i=1}^{N}[y_i \log(\frac{y_i}{n_i\widehat{\pi}_t}) + (n_i - y_i)\log(\frac{n_i - y_i}{n_i - n_i\widehat{\pi}_i})]$$

$$= 2\sum_{i=1}^{N}[(y_i - n_i\widehat{\pi}_i) + \frac{1}{2}\frac{[(n_i - n_i\widehat{\pi}_i)^2}{n_i\widehat{\pi}_i} + [(n_i - y_i) - (n_i - n_i\widehat{\pi}_i)]$$

$$+\frac{1}{2}\frac{[(n_i - y_i) - (n_i - n_i\widehat{\pi}_i)]^2}{n_i - n_i\widehat{\pi}_i} + ...$$

$$\cong \sum_{i=1}^{N}\frac{(n_i - n_i\widehat{\pi}_i)^2}{n_i\widehat{\pi}_i(1-\widehat{\pi}_i)} = X^2$$

Δηλαδή η ασυμπτωτική κατανομή της D κάτω από την υπόθεση ότι το μοντέλο είναι σωστό είναι D~ x^2_{N-p} δηλαδή προσεγγιστικά X^2_{N-p}.

Υπάρχουν ενδείξεις ότι η X^2 είναι συχνά καλύτερη από την από την D γιατί η D εξαρτάται από μικρές συχνότητες. Και οι δυο προσεγγίσεις δεν θεωρούνται ικανοποιητικές όταν οι αναμενόμενες συχνότητες είναι πολύ μικρές (π.χ. μικρότερες της μονάδας).

2.4 Ο χειρισμός των κατηγορικών μεταβλητών σαν ανεξάρτητες μεταβλητές

Στην περίπτωση κατά την οποία πρέπει να χρησιμοποιηθεί μια κατηγορική μεταβλητή X με ν επίπεδα (X=0,1,2,…,v) σαν ανεξάρτητη μεταβλητή τότε πρέπει πρέπει αντί αυτής να εισάγουμε στο μοντέλο της λογιστικής παλινδρόμισης ν-1 το πλήθος δυωνυμικές μεταβλητές.

Παράδειγμα

Εστω ότι θέλουμε να εισάγουμε τη μεταβλητή X= Οικογενειακή κατάσταση,η οποία μπορεί να πάρει τις τιμές

$$X = \begin{cases} 0 & \text{Αγαμος/η} \\ 1 & \text{Παντρεμένος/η} \\ 2 & \text{Διαζευγμένος/η} \\ 3 & \text{Χήρος/α} \end{cases}$$

Αντί της παραπάνω μεταβλητής υπεισέρχονται στο μοντέλο οι τρεις μεταβλητές

$$X_1 = \begin{cases} 1 & \text{Παντρεμένος} \\ 0 & \text{Αλλο} \end{cases}$$

$$X_2 = \begin{cases} 1 & \text{Διαζευγμένος} \\ 0 & \text{Αλλο} \end{cases}$$

$$X_3 = \begin{cases} 1 & \text{Χήρος} \\ 0 & \text{Αλλο} \end{cases}$$

Προφανώς μια τέταρτη μεταβλητή με τιμή 1 για την επιλογή Αγαμος και 0 αλλού δεν μπορεί να μπεί στο μοντέλο γιατί θα είχαμε πολυσυγραμμικότητα .

Το σύνολο το οποίο ‘’αγνοείται ‘’ λαμβάνεται σαν βάση αναφοράς. Για το παραπάνω παράδειγμα, ο λογάριθμος του λόγου των odds των παντρεμένων, των διαζευγμένων, και των χήρων λαμβάνεται ως προς την ομάδα Αγαμος που είναι η βάση αναφοράς.

Στην πράξη αυτή η μετατροπή πολύτομων κατηγορικών ανεξάρτητων μεταβλητών σε ένα σύνολο δυαδικών γίνεται αυτόματα από όλα τα στατιστικά πακέτα.

2.5 Κατασκευή του μοντέλου

Εως τώρα έχουμε ασχοληθεί με τη εκτίμηση, τον έλεγχο και την ερμηνεία των συντελεστών των μοντέλων. Σε πολλές περιπτώσεις, από ένα μεγάλο σύνολο ανεξάρτητων μεταβλητών επιθυμούμε να επιλέξουμε εκείνες τις μεταβλητές οι οποίες οδηγούν σε ένα βέλτιστο (με κάποια έννοια μοντέλο). Για να πετύχουμε αυτόν τον στόχο, πρέπει να έχουμε α) μια βασική μέθοδο επιλογής των μεταβλητών και β) ένα σύνολο μεθόδων για να ελέγχουμε την επάρκεια των μεθόδων.

Η απλούστερη και συνηθέστερη διαδικασία είναι η enter όπου όλες οι μεταβλητές εισέρχονται σαν μία ομάδα.

Άλλες διαδικασίες είναι οι

Forward : όπου η διαδικασία ξεκινά με την ‘’καλύτερη ‘’ μεταβλητή , στη συνέχεια προσθέτει την καλύτερη από τις υπόλοιπες κ.λ.π. μέχρις ότου προσθέτοντας μία νέα μεταβλητή η αύξηση της λογαριθμικής συνάρτησης πιθανοφάνειας δεν είναι στατιστικώς σημαντική.

Backwards: Ξεκινά με όλο το σύνολο των μεταβλητών και απορρίπτει διαδοχικά τη χειρότερη από τις εναπομείναντες.

Stepwise: Είναι ανάλογη της Forward με τη διαφορά ότι κάθε φορά που μια μεταβλητή εισέρχεται στο μοντέλο, ελέγχει αν στο νέο σύνολο όλες οι μεταβλητές είναι στατιστικώς σημαντικές.

Η stepwise δεν είναι διαθέσιμη στο SPSS.

2.5 Η λογιστική παλινδρόμηση στο SPSS

Θα αναλύσουμε το αρχείο logistic.sav στο οποίο καταγράφονται παρατηρήσεις 420 παιδιών λυκείου . Ο σκοπός της μελέτης είναι να βρεθούν οι παράγοντες που επηρεάζουν την εμφάνιση άσθματος.

Οι μεταβλητές που κατεγράφησαν είναι

asdiagn95 =1 για παρουσία άσθματος , 0 για μη παρουσία άσθματος

fvc95= χωρητικότητα των πνευμόνων.

mef5095=πίεση εξόδου του αέρα από τα πνευμόνια

whz1295=αν υπάρχει συριγμός κατά την έξοδο του αέρα.

couev95=αν υπάρχει βήχας

heyfev95= παρουσία εαρινής ρινοεπιπεφυκίτιδας.

areagroup=1 για αστική περιοχή, 0 για επαρχία

smoking=αν καπνίζει το ίδιο το παιδί

mosmpreg=αν κάπνιζε η μητέρα του κατά την διάρκεια της εγκυμοσύνης

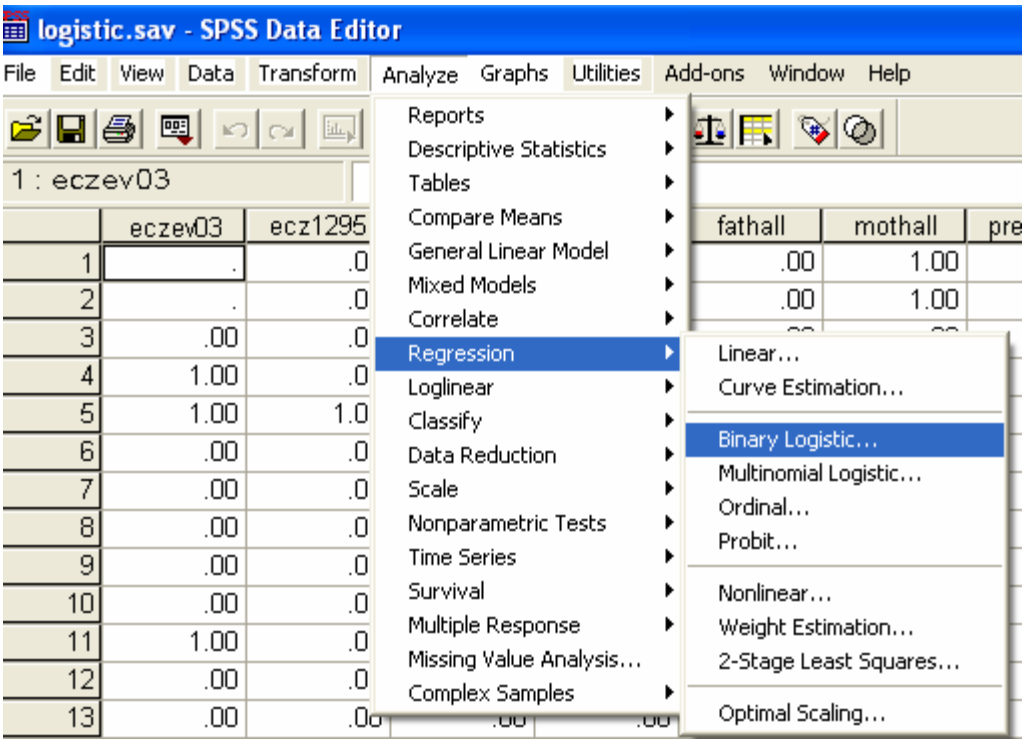
pets=ύπαρξη κατοικίδιων

catpets=επαφή με γάτες

outbask=εξωτερικές αθλητικές δραστηριότητες.

Για να χρησιμοποιήσουμε το SPSS στη λογιστική παλινδρόμιση κάνουμε τα εξής βήματα.

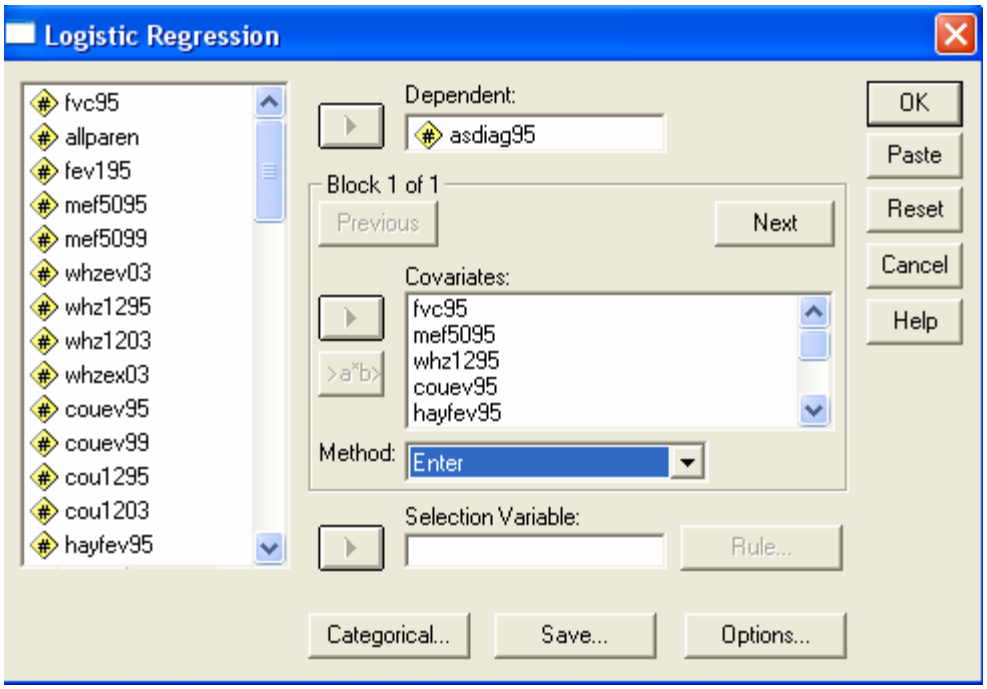
Analyze-→Regression-→Binary Logistic



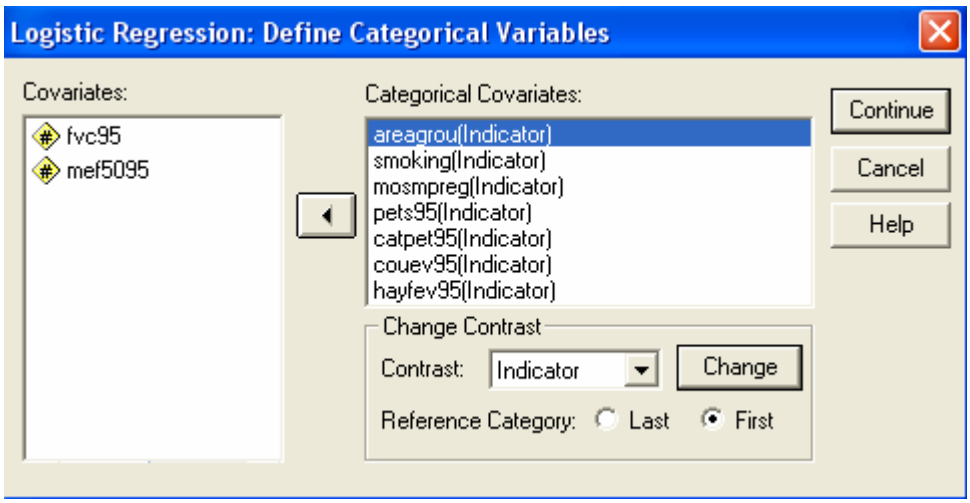
Στη συνέχεια στην επιλογή Binary Logistic θέτουμε στη θέση dependent την εξαρτημένη μεταβλητή **asdiag95.**

Στη θέση covariates θέτουμε τις ανεξάρτητες μεταβλητές.

Σαν μέθοδο επιλέγουμε την enter.



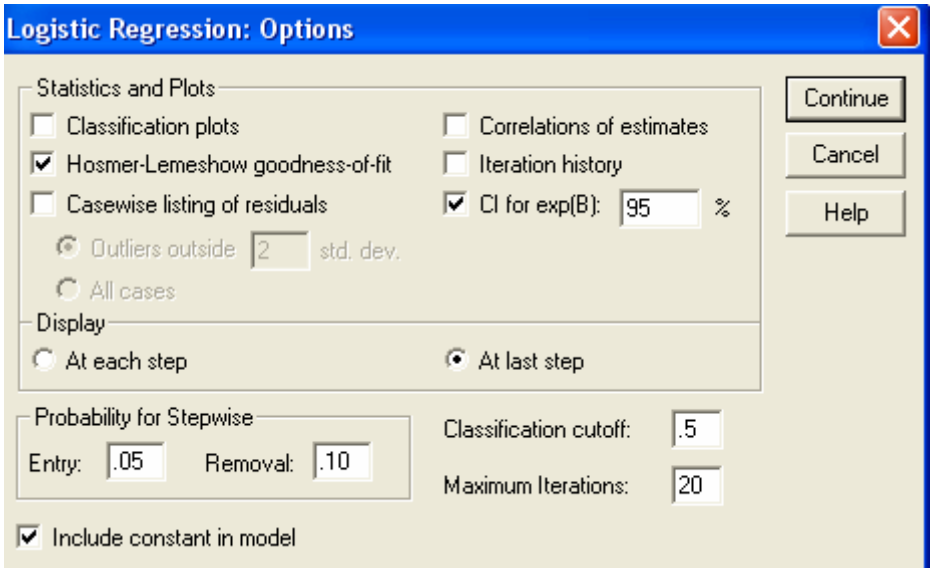
Στη συνέχεια πρέπει να ορίσουμε τις κατηγορικές μεταβλητές. Επιλέγουμε λοιπόν την θέση categorical και σύρουμε μέσα τις κατηγορικές μεταβλητές. Η επιλογή reference category έχει νόημα κυρίως όταν κατηγορική μεταβλητή έχει πάνω από δύο επίπεδα όπου όλα τα άλλα επίπεδα συγκρίνονται ως προς την reference category που επιλέγουμε.



Στη συνέχεια, στην επιλογή option οι πιο σημαντικές επιλογές είναι το Hosmer and Lemeshov goodness-of-fit test, το διάστημα εμπιστοσύνης του exp(β) Η επιλογή probability for stepwise αναφέρεται στην περίπτωση όπου η μέθοδος δεν είναι η

enter αλλά κάποια από τις forward ή backward όπου οι μεταβλητές εισέρχονται ή εξέρχονται στο μοντέλο με κάποιες προκαθορισμένες πιθανότητες.

Δεν πρέπει να ξεχνάμε ότι η εξαρτημένη μεταβλητή είναι δυαδική με τιμές 0 ή 1. Η αντίστοιχη πρόβλεψη είναι η $\pi=P(Y=1)$. Με το classification cutoff=0.5 ορίζουμε ότι για $\pi<0.05$ η προβλεπόμενη τιμή του μοντέλου είναι 0 και για $\pi\geq0.5$ είναι 1.



Τα αποτελέσματα είναι τα εξής

Logistic Regression

Case Processing Summary

Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	820	100,0
	Missing Cases	0	,0
	Total	820	100,0
Unselected Cases		0	,0
Total		820	100,0

a. If weight is in effect, see classification table for the total number of cases.

Στον παραπάνω πίνακα έχουμε τον αριθμό των παρατηρήσεων καθώς και των περιπτώσεων με missing values

Dependent Variable Encoding

Original Value	Internal Value
,00	0
1,00	1

Ο παραπάνω πίνακας μας ενημερώνει το πως αντιλαμβάνεται το SPSS κωδικοποίηση της εξαρτημένης μεταβλητής.

Categorical Variables Codings

		Frequency	Paramete (1)
outbaske	,00	673	1,000
	1,00	147	,000
couev95	,00	765	1,000
	1,00	55	,000
hayfev95	,00	717	1,000
	1,00	103	,000
areagrou	,00	478	1,000
	1,00	342	,000
smoking	,00	330	1,000
	1,00	490	,000
mosmpreg	,00	717	1,000
	1,00	103	,000
catpet95	,00	497	1,000
	1,00	323	,000
pets95	,00	375	1,000
	1,00	445	,000
whz1295	,00	762	1,000
	1,00	58	,000

Ο παραπάνω πίνακας είναι πληροφοριακός στο πως κωδικοποιεί το SPPP τις κατηγορικές μεταβλητές

Στη συνέχεια στο Block 0 μελετάται το τετριμμένο μοντέλο

$\log(\frac{\pi}{1-\pi}) = \beta_0$.

Block 0: Beginning Block

Classification Table[a,b]

Observed			Predicted		
			asdiag95		Percentage Correct
			,00	1,00	
Step 0	asdiag95	,00	724	0	100,0
		1,00	96	0	,0
	Overall Percentage				88,3

a. Constant is included in the model.

b. The cut value is ,500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	-2,020	,109	346,010	1	,000	,133

Ο παραπάνω πίνακας δείχνει ο σταθερός όρος δεν είναι 0.
Ο επόμενος πίνακας δείχνει ποία από τις μεταβλητές είναι η πιο σημαντική και η οποία θα μπει πρώτη στο μοντέλο. Είναι χρήσιμος όταν έχουμε επιλέξει κάποιας μορφής stepwise διαδικασία.

Variables not in the Equation

			Score	df	Sig.
Step 0	Variables	fvc95	,384	1	,535
		mef5095	1,616	1	,204
		whz1295(1)	275,954	1	,000
		couev95(1)	34,673	1	,000
		hayfev95(1)	267,916	1	,000
		areagrou(1)	1,232	1	,267
		smoking(1)	,556	1	,456
		mosmpreg(1)	1,005	1	,316
		pets95(1)	,172	1	,678
		catpet95(1)	,163	1	,687
		outbaske(1)	1,840	1	,175
	Overall Statistics		378,140	11	,000

Στη συνέχεια λόγω της μεθόδου που έχουμε επιλέξει εισάγει όλες τις μεταβλητές σαν ένα σύνολο (block)

Block 1: Method = Enter

Στον παρακάτω πίνακα ελέγχεται η υπόθεση ότι το μοντέλο με όλες τις μεταβλητές είναι το ίδιο με το τετριμμένο μοντέλο. Η υπόθεση αυτή απορρίπτεται.

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	256.920	11	.000
	Block	256.920	11	.000
	Model	256.920	11	.000

Στον παρακάτω πίνακα δίνονται κάποια στατιστικά χαρακτηριστικά του μοντέλου.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	335.207[a]	.269	.523

a. Estimation terminated at iteration number 6 because parameter estimates changed by less than .001.

Classification Table[a]

			Predicted		
			asdiag95		Percentage Correct
	Observed		.00	1.00	
Step 1	asdiag95	.00	712	12	98.3
		1.00	50	46	47.9
	Overall Percentage				92.4

a. The cut value is .500

Στον παρακάτω πίνακα δίνει ποιες μεταβλητές είναι στατιστικώς σημαντικές για το μοντέλο.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	fvc95	.005	.022	.052	1	.820	1.005
	mef5095	.004	.016	.063	1	.801	1.004
	whz1295(1)	-3.070	.421	53.226	1	.000	.046
	couev95(1)	-1.259	.450	7.819	1	.005	.284
	hayfev95(1)	-2.939	.331	79.066	1	.000	.053
	areagrou(1)	-.393	.377	1.083	1	.298	.675
	smoking(1)	.037	.311	.014	1	.906	1.037
	mosmpreg(1)	-.371	.463	.642	1	.423	.690
	pets95(1)	-.585	.453	1.664	1	.197	.557
	catpet95(1)	.305	.426	.513	1	.474	1.357
	outbaske(1)	-.178	.409	.189	1	.664	.837
	Constant	3.771	2.082	3.282	1	.070	43.429

a. Variable(s) entered on step 1: fvc95, mef5095, whz1295, couev95, hayfev95, areagrou, smoking, mosmpreg, pets95, catpet95, outbaske.

Σαν αποτέλεσμα έχουμε ότι οι παράγοντες που σχετίζονται με την εμφάνιση άσθματος είναι η παρουσία βήχα, το σύριγμα κατά την εκπνοή, και η παρουία εαρινής ρινοεπιπεφυκίτιδας.

Βιβλιογραφία

1. Dobson A.J. (1990) *An Introduction to Generalized Linear .Models* 2nd Ed.(University of Newcastle) Chapman and Hall
2. McCullagh P. and Nelder J.A. (1989) *Generalized Linear Models*. 2nd Ed. London Chapman and Hall.