Ανάλυση Κατηγορικών Δεδομένων
Πανεπιστημιακές παραδόσεις
Οικονομικό Πανεπιστήμιο Αθηνών
Τμήμα Στατιστικής

Πέτρος Δελλαπόρτας

Οκτώβριος 1999

Πρόλογος

Αυτό το τεύχος έχει σκοπό να βοηθήσει τους φοιτητές που παρακολουθούν τα μαθήματά μου. Το περιεχόμενό του, αφορά ορισμένες μόνο πτυχές της Θεωρίας των κατηγορικών δεδομένων, που παρουσιάζονται έτσι ώστε να αποτελούν απλά ένα βοήθημα και όχι ολοκληρωμένη παρουσίαση του θέματος. Ελπίζω ότι δεν θα παρεξηγηθεί ο σκοπός του και ότι οι φοιτητές δεν θα περιοριστούν για την κατάρτησή τους, στην ανάγνωση αυτού και μόνο του τεύχους.

Θέλω να ευχαριστήσω τους φοιτητές, που με τις κρίσεις και παρατηρήσεις τους, με βοήθησαν να δώσω στο μάθημα αυτή τη μορφή. Ειδικά, οι φοιτητές Μαρία Κατσικαβέλη και Μιχάλης Λιναρδάκης, διάβασαν μεγάλο μέρος των κειμένων μετά την πρώτη δακτυλογράφηση και έκαναν πολλές χρήσιμες παρατηρήσεις και υποδείξεις.

"Υπεύθυνη" για την γρήγορη και επιμελημένη δακτυλογράφηση είναι η Πόλυ Σάρρου, την οποία και ευχαριστώ πολύ.

Πέτρος Δελλαπόρτας

Τα τελευταία 80 χρόνια η στατιστική μεθοδολογία για τα κατηγορικά δεδομένα, έχει αναπτυχθεί με γοργούς ρυθμούς και έχει φθάσει τα επίπεδα περιπλοκότητας της μεθοδολογίας των συνεχών δεδομένων. Κύριος λόγος είναι ο γρήγορος ρυθμός ανάπτυξης των κοινωνικών και βιοϊατρικών επιστημών, με συνεπακόλουθη την ανάγκη για εξελιγμένες μορφές στατιστικής ανάλυσης.

Στις κοινωνικές επιστήμες, οι κατηγορικές μετρήσεις είναι απαραίτητες για την μέτρηση ποσότητας ή γνώμης σε θέματα και δημογραφικά χαρακτηριστικά όπως φύλο, φυλή, κοινωνική τάξη. Στις βιοϊατρικές επιστήμες, υπάρχει ανάγκη να μετρήσουμε παράγοντες όπως δριμύτητα ατυχήματος, βαθμός ανάρρωσης από χειρουργική επέμβαση, κατάσταση επιδημίας.

Στατιστικά μοντέλα αντίστοιχα αυτών των συνεχών δεδομένων (ανάλυση παλινδρόμησης, ανάλυση διακύμανσης και συνδιακύμανσης κ.λ.π.) έχουν αναπτυχθεί για κατηγορικές αποκρίσεις, για να ανταπεξέλθουν στην ανάγκη ανάλυσης πολυμεταβλητών διακριτών παρατηρήσεων. Σε αυτό το τεύχος, ειδική έμφαση δίνεται στα μοντέλα που προέρχονται από τα γενικευμένα γραμμικά μοντέλα (λογαριθμικά γραμμικά μοντέλα, μοντέλα για διωνυμικά δεδομένα).

Θα ξεκινήσουμε δίνοντας τρία παραδείγματα δεδομένων, που μπορούν να αναλυθούν με τις τεχνικές που θα αναπτυχθούν παρακάτω και θα μας χρησιμεύσουν σαν αναφορές στην συνέχεια:

1.1 Τύποι Δεδομένων

DATASET 1: Γόνιμα ποντίκια βάσει βιταμίνης Ε:

Δόση (mg)	Αριθμός ποντικιών	Αριθμός γόνιμων
3.75	5	0
5.00	10	2
6.25	10	4
7.50	10	8
10.00	11	10
15.00	11	11

DATASET 2: Αποτέλεσμα μετά από αυτοκινητιστικό ατύχημα:

	Αρχική Δριμύτητα του ατυχήματος	Απόκριση ΟΧΙ	Απόκριση ΝΑΙ
ΦΑΡΜΑΚΟ Α	Δριμύ	6	14
	Μέτριο	5	0
	Ελαφρύ	4	1
ΦΑΡΜΑΚΟ Β	Δριμύ	1	6
	Μέτριο	3	14
	Ελαφρύ	2	4

DATASET 3: Αποτέλεσμα μετά από τραυματισμό κεφαλιού:

	ΑΠΟΚΡΙΣΗ				
	Χειρότερο	Ίδιο	Καλύτερο	Πολύ καλύτερο	Σύνολο
A	13	13	12	22	60
ΦΑΡΜΑΚΟ B	4	24	28	34	90
C	3	8	15	24	50
	20	45	55	80	200

1.1.1 Σχόλια

DATASET 1: $p_i = P_r$ (ποντίκι γόνιμο | Δόση i)

Ερώτηση: Ποιά είναι η αναμενόμενη γονιμότητα στην δόση $= \hat{d}$;

DATASET 2: $p_{ij} = P_r$ (Ανάρρωση | Φάρμακο i και αρχική δριμύτητα j)

Ερώτηση: Είναι η απόκριση ανεξάρτητη του φαρμάκου και /ή αρχικής δριμύτητας;

DATASET 3: Ενδιαφερόμαστε για το ποιό φάρμακο δίνει την "βέλτιστη" απόκριση - αλλά τώρα η διακριτή απόκριση έχει 4 επίπεδα. Αν η απόκριση ήταν συνεχής, τότε θα μπορούσαμε να χρησιμοποιήσουμε Ανάλυση Διασποράς !

1.2 Μοντέλα / Επανάληψη

Στα γραμμικά μοντέλα θεωρήσαμε την

$$\mathbf{y} = \mathbf{z}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

$\mathbf{y}$ – n μονοδιάστατες αποκρίσεις

$\mathbf{z}$ – $n \times p$ πίνακας σταθερών (design matrix)

$\boldsymbol{\beta}$ – $p \times l$ διάνυσμα ΠΑΡΑΜΕΤΡΩΝ.

Στην απλούστερή τους μορφή, τα σφάλματα είναι:

(i) ανεξάρτητα

(ii) ταυτοτικά κατανεμημένα σαν $N(0, \sigma_i^2)$

(iii) ανεξάρτητα του i, δηλαδή $\sigma_i^2 = \sigma^2$

Έχουμε τις παρακάτω ειδικές περιπτώσεις γραμμικών μοντέλων:

α) Πολλαπλή παλινδρόμηση - συνεχείς μεταβλητές

β) Ανάλυση Διασποράς - παράγοντες (factors), διακριτές μεταβλητές

γ) Ανάλυση Συδιακύμανσης - συνεχείς και διακριτές μεταβλητές

Η παρακάτω στρατηγική για την επιλογή κατάλληλου μοντέλου θα χρησιμοποιηθεί και για τα κατηγορικά δεδομένα:

Α) "Αναγνωριστική " ανεπίσημη έρευνα των δεδομένων (exploratory data analysis)

Β) Πρόταση μοντέλου

Γ) Προσαρμογή μοντέλου

Δ) Προσδιορισμός επάρκειας μοντέλου

Ε) Προσαρμογή άλλων πιθανών μοντέλων και επιλογή "καλύτερου / καλύτερων"

Ζ) "Ερμηνεία" καλύτερου/καλύτερων μοντέλου / μοντέλων.

Με τον όρο "ερμηνεία" εννοούμε

- ΠΡΟΒΛΕΨΗ
- ΠΕΡΙΛΗΨΗ ΔΕΔΟΜΕΝΩΝ
- ΑΝΑΓΝΩΡΙΣΗ ΕΝΔΙΑΦΕΡΟΝΤΩΝ ΣΧΕΣΕΩΝ
- ΣΥΜΠΕΡΑΣΜΑΤΟΛΟΓΙΑ ΓΙΑ ΠΑΡΑΜΕΤΡΟΥΣ ΤΟΥ ΜΟΝΤΕΛΟΥ

1.3 Αναφορές

- Agresti A, (1990), Categorical Data Analysis, John Wiley and Sons.

- McCullagh and Nelder, (1989), Generalised Linear Models, Chapman and Hall.

- Dobson A.J. (1983), An introduction to Statistical Modelling, Chapman and Hall.

2.1 ΟΡΟΛΟΓΙΑ

2.1.1 Δεδομένα

Αποκτούμε δεδομένα από n πειραματικές μονάδες όπου κάθε μονάδα είναι διάνυσμα k τιμών, και οι τιμές είναι πραγματοποιήσεις (realizations) k τυχαίων μεταβλητών (τ.μ.). Αλλες μεταβλητές μπορούν να *ελεγχθούν* ενώ άλλες είναι αληθινές τ.μ.

Για παράδειγμα, στο DATASET 3 (αποτέλεσμα μετά από τραυματισμό στο κεφάλι), αν τα δεδομένα έχουν συλλεχθεί στην διάρκεια του χρόνου, τότε τίποτα δεν έχει σταθεροποιηθεί απο την αρχή. Εναλλακτικά, η μελέτη θα μπορούσε να σχεδιαστεί ώστε να συλλέξουμε δεδομένα από ένα προκαθορισμένο αριθμό ασθενών (π.χ. n=200). Μερικές φορές αποφασίζουμε να κρατήσουμε σταθερό τον αριθμό των ανθρώπων που δοκιμάζουν κάποιο φάρμακο, δηλ. να κρατήσουμε σταθερό το άθροισμα των γραμμών (Συμβατικά, οι γραμμές είναι *αγωγές*, οι στήλες είναι *αποκρίσεις*). Τέτοιες μελέτες καλούνται *προσδοκώμενες* μελέτες (Prospective studies).

Θα μπορούσαμε αντί προσδοκώμενη, να είχαμε σχεδιάσει *αναδρομική* μελέτη : να κρατήσουμε σταθερό το άθροισμα των στηλών.

Οι μεταβλητές μπορούν να μετρηθούν σε διάφορες κλίμακες:

1. Συνεχείς.
2. Διακριτές : δυαδικές ή διχοτόμες (π.χ. 0/1 ή M/F).
3. Πολυκατηγορικές.

Οι πολυκατηγορικές μπορεί να είναι:

(i) Ονομαστικές (nominal), π.χ. κόκκινο, πράσινο, μαύρο

(ii) Τακτικές (ordinal), π.χ. ελαφρύς, μέτριος, δριμύς.

Οι διακριτές μεταβλητές καλούνται συνήθως *παράγοντες* (factors). Οι επιτρεπόμενες τιμές ενός παράγοντα καλούνται *επίπεδα* (levels), π.χ. *SEX* είναι δυαδικός παράγοντας με δύο επίπεδα.

2.1.2 Πίνακες συναφείας

Πίνακας συναφείας είναι πίνακας συχνοτήτων: ουσιαστικά εκφράζουμε διαφορετικά ένα πίνακα όπου οι γραμμές είναι *περιπτώσεις* και οι στήλες *παράγοντες*.

ΠΙΝΑΚΑΣ Α

ΠΕΡΙΠΤΩΣΕΙΣ	**SEX** 1= MALE 2= FEMALE	**ΑΓΩΓΗ** 1=ΦΑΡΜΑΚΟ 1 2=ΦΑΡΜΑΚΟ 2	**ΑΠΟΚΡΙΣΗ** 1=ΑΠΟΤΥΧΙΑ 2=ΘΕΡΑΠΕΙΑ
1	1	1	2
2	1	1	1
3	2	2	2
4	1	2	2
5	1	1	2
⋮	⋮	⋮	⋮
n	1	2	1

Εκφράζεται σαν

		ΑΠΟΚΡΙΣΗ	
	SEX	Α	Θ
ΦΑΡΜΑΚΟ 1	M	n_{111}	n_{211}
	F	n_{112}	n_{212}
ΦΑΡΜΑΚΟ 2	M	n_{121}	n_{221}
	F	n_{122}	n_{222}

Δηλαδή τα n_{ijk}, όπου ΑΠΟΚΡΙΣΗ $=i$, ΑΓΩΓΗ $=j$, SEX $=k$, αναπαριστούν τα δεδομένα του ΠΙΝΑΚΑ Α.

Η διάσταση του πίνακα ισούται με τον αριθμό των παραγόντων που ορίζουν τον πίνακα. Για παράδειγμα, το DATASET 2 έχει διάσταση 3, ενώ το DATASET 3 έχει διάσταση 2.

Αν x_i είναι παράγοντες με L_i επίπεδα, $i = 1, \ldots, k$, τότε ο αντίστοιχος $k-$διάστατος πίνακας έχει

$$\prod_{i=1}^{k} L_i \qquad \text{κελλιά.}$$

π.χ. DATASET 3: $k = 3, L_1 = 2, L_2 = 2, L_3 = 3 \Rightarrow 12$ κελλιά.

Αν n είναι ο συνολικός αριθμός πειραματικών μονάδων τότε εάν το k είναι μεγάλο και το n δεν είναι τόσο μεγάλο, πολλά κελλιά στον πίνακα περιέχουν *μηδενικές* συχνότητες, - τέτοιοι πίνακες καλούνται *Διεσπαρμένοι* (sparse). Στην πράξη, πολύ σπάνια έχουμε $k > 5$.

Παράδειγμα διδιάστατου πίνακα συναφείας

$$
\begin{array}{ccccc|l}
n_{11} & n_{12} & \ldots & \ldots & n_{1J} & n_{1+} \\
n_{21} & n_{22} & \ldots & \ldots & n_{2J} & n_{2+} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
n_{I1} & n_{I2} & \ldots & \ldots & n_{IJ} & n_{I+} \\
\hline
n_{+1} & n_{+2} & \ldots & \ldots & n_{+J} & n_{++} = n
\end{array}
$$

Ο συμβολισμός $+$ συμβολίζει άθροισμα του αντίστοιχου δείκτη, π.χ. $n_{i+} = \sum_{j=1}^{J} n_{ij}$ (ΑΘΡΟΙΣΜΑΤΑ ΓΡΑΜΜΩΝ), $n = n_{++} = \sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}$ (ΟΛΙΚΟ ΣΥΝΟΛΟ).

Τα αθροίσματα γραμμών / στηλών ορίζουν πίνακες διάστασης 1 (one- way tables) που ονομάζονται *περιθώριοι* στον διδιάστατο πίνακα.

Για τρισδιάστατους και μεγαλύτερους πίνακες :

- Η παρουσίαση είναι υποκειμενικό θέμα
- Παράγουν περιθώριους πίνακες διάστασης $k-1$, όπου k η διάσταση του αρχικού πίνακα.

Δεσμευμένοι πίνακες αποκτούνται από τον αρχικό πίνακα ή από περιθώριο πίνακα κρατώντας σταθερά διάφορα επίπεδα διαφορετικών παραγόντων - μπορείτε να τους σκεφθείτε σαν "φέτες" του αρχικού πίνακα!

2.2 Διακριτές κατανομές

2.2.1 Κατανομή Bernoulli

Η X είναι τ.μ. Bernoulli εάν και μόνο εάν

$$
P_r(X = x) = \begin{cases} 1-p & \text{αν} \quad x = 0 \\ p & \text{αν} \quad x = 1 \\ 0 & \text{αλλιώς} \end{cases}
$$

2.2.2 Διωνυμική κατανομή

Έστω X μια τ.μ που περιγράφει τον αριθμό των "1" που αποκτούνται σε n ανεξάρτητες δοκιμές Bernoulli. Τότε

$$P_r(X=x)=\begin{pmatrix} n \\ x \end{pmatrix} p^x(1-p)^{n-x}, \quad x=0,\ldots,n$$

Γράφουμε $X \sim B(n,p)$.

2.2.3 Γενικευμένη κατανομή Bernoulli

Έστω X ένας παράγοντας που παίρνει k διακεκριμένες τιμές με πιθανότητες $p_1, p_2, \ldots, p_k$, όπου $\sum_{i=1}^{k} p_i = 1$. Έστω $\mathbf{T}$ μια διανυσματική τ.μ. $(k \times 1)$ με δειγματικό χώρο

$$\left\{ \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix}, \ldots, \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix} \right\}$$

όπου

$$P_r\left(T = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \right) = p_1 \qquad \text{κ.λ.π.}$$

Τότε η $\mathbf{T}$ έχει γενικευμένη κατανομή Bernoulli.

2.2.4 Πολυωνυμική κατανομή

Έστω $\mathbf{T_i}, i=1,\ldots,n$, n ανεξάρτητες πραγματοποιήσεις μιας γενικευμένης τ.μ.Bernoulli. Έστω $\mathbf{N} = \sum_{i=1}^{n} \mathbf{T_i}$. Τότε

$$\mathbf{P_r}(\mathbf{N}=\mathbf{n}) = \mathbf{P_r}(N_1=n_1, N_2=n_2, \ldots, N_k=n_k) = \frac{n!}{n_1! n_2! \ldots n_k!} p_1^{n_1} p_2^{n_2} \ldots p_k^{n_k}$$

Σημειώσεις

(i) Γράφουμε $\mathbf{N} \sim M_k(\mathbf{n}, \mathbf{p})$

(ii) Πολυδιάστατη κατανομή

(iii) $\sum_{i=1}^{k} p_i = 1$, άρα υπάρχουν $k-1$ διακεκριμένες παράμετροι, και έχουμε μία $(k-1)$-διάστατη παραμετρική κατανομή.

2.2.5 Κατανομή Poisson

Η τ.μ. N έχει κατανομή Poisson εάν

$$P_r(N=n) = \frac{e^{-m}m^n}{n!}, \quad n = 0, 1, 2, \ldots$$

Σημειώσεις:

(i) Γράφουμε $N \sim p(m)$. $E[N] = m, \quad V[N] = m$

(ii) Θυμηθείτε την ανέλιξη Poisson με ένταση λ. Στο διάστημα $(0, t)$ ο αριθμός των γεγονότων έχει κατανομή Poisson με $m = \lambda t$.

(iii) Έστω k ανεξάρτητες ανελίξεις Poisson, με εντάσεις $\lambda_1, \lambda_2, \ldots, \lambda_k$. Έστω N_i ο αριθμός των γεγονότων που παρατηρείται από την $i-$οστή ανέλιξη σε χρόνο t. Τότε

$$P_r(N_1 = n_1, N_2 = n_2, \ldots, N_k = n_k) = \prod_{i=1}^{k} \frac{e^{-m_i}m_i^{n_i}}{n_i!}$$

όπου $m_i = \lambda_i t$. Μια πολυδιάστατη κατανομή που εκφράζεται σαν γινόμενο μονοδιάστατων κατανομών.

(iv) Αν $N_i \sim p(m_i), \quad i = 1, \ldots, k,$ και N_i ανεξάρτητες, τότε

$$\sum_{i=1}^{k} N_i \sim p(\sum_{i=1}^{k} m_i) \quad .$$

2.3 Σχήματα δειγματοληψίας για πίνακα $I \times J$

Οταν δίνεται ένα σύνολο δεδομένων, είναι *πολύ σημαντικό* να ξέρουμε τον τρόπο με τον οποίο τα δεδομένα έχουν συλλεχθεί: να ξέρουμε δηλαδή το σχήμα δειγματοληψίας. Θα αναφερθούμε σε κάθε ένα πιθανό σχήμα στο DATASET 3:

2.3.1 Πολυωνυμική δειγματοληψία

Αν κρατήσουμε σταθερό το ολικό σύνολο, των πειραματικών μονάδων εκ των προτέρων, τότε έχουμε μια πολυωνυμική κατανομή για τις προκύπτουσες παρατηρήσεις:

$$P_r(\mathbf{N} = \mathbf{n} \mid \mathbf{p}) = \frac{n_{++}!}{n_1! \ldots n_{IJ}!} \prod_{i=1}^{I} \prod_{j=1}^{J} p_{ij}^{n_{ij}} \equiv M_{IJ}(n_{++}, \mathbf{p}), \quad \sum_i \sum_j p_{ij} = 1.$$

π.χ. στο DATASET 3 αποφασίσαμε να συλλέξουμε δεδομένα απο 200 ασθενείς.

2.3.2 Δειγματοληψία με γινόμενο πολυωνυμικών

Εναλλακτικά, έστω ότι αποφασίζουμε να συλλέξουμε για κάθε γραμμή ένα προκαθορισμένο αριθμό δεδομένων. Τότε για κάθε γραμμή έχουμε ένα διαφορετικό πολυωνυμικό πείραμα:

$$P_r\left(\mathbf{N} = \mathbf{n} \mid \mathbf{p} = \prod_{i=1}^{I} M_J(n_{i+}, p_i\right)$$

με $(J-1)I$ παραμέτρους, $p_i = (p_{i1}, p_{i2}, \ldots, p_{iJ})$ και $\sum_{j=1}^{J} p_{ij} = 1$.

Το σύνολο κάθε γραμμής είναι προκαθορισμένο από τον πειραματικό σχεδιασμό. Για παράδειγμα, στο DATASET 3, αποφασίζουμε να συλλέξουμε δεδομένα σε προκαθορισμένο αριθμό για το φάρμακο A, για το φάρμακο B και το φάρμακο C. Οταν συλλέξουμε αρκετά για το B, τότε απλώς περιμένουμε για A και C, δηλ. απορρίπτουμε τα B από εκείνο το σημείο και μετά. Αυτό σημαίνει ότι δεν μπορούν να εξαχθούν συμπεράσματα για τα σύνολα σε κάθε γραμμή (δηλ. δεν μπορούμε να βρούμε την πιθανότητα να έχει δοθεί το φάρμακο A). Το DATASET 1 μπορεί να είναι γινόμενο Διωνυμικών - ειδική περίπτωση. Προφανώς, μπορούμε να συλλέξουμε δεδομένα κρατώντας το σύνολο κάθε στήλης σταθερό.

2.3.3 Γινόμενο Poisson

Συλλέγουμε δεδομένα για προκαθορισμένο χρονικό διάστημα. Εδώ, το συνολικό άθροισμα n_{++} είναι επίσης τυχαίο. Συνεπώς έχουμε IJ παραμέτρους και ισχύει

$$P_r(\mathbf{N} = \mathbf{n} \mid \mathbf{m}) = \prod_{i=1}^{I} \prod_{j=1}^{J} \frac{e^{-m_{ij} m_{ij}^{n_{ij}}}}{n_{ij}!}$$

δηλαδή σε κάθε κελλί έχουμε κατανομή Poisson με μέσο m_{ij}.

2.3.4 Σχόλια

- Ολα τα μοντέλα γενικεύονται σε πίνακες μεγαλύτερης διάστασης.

- Οταν ξέρουμε το ιστορικό του πειράματος, μπορούμε να γράψουμε την *πιθανοφάνεια* των δεδομένων. Οπως με τα γραμμικά μοντέλα με κανονικά σφάλματα, ακολουθούμε μια διαδικασία προσαρμογής μοντέλου που (ελπίζουμε !) θα μας απαντήσει στις ερωτήσεις που μας ενδιαφέρουν. Για παράδειγμα, σε ένα πίνακα συναφείας $I \times J$ μια από τις πιο σπουδαίες ερωτήσεις είναι: "Είναι η απόκριση ανεξάρτητη της αγωγής;" $\equiv$ "είναι η κατηγοριοποίηση των γραμμών ανεξάρτητη της κατηγοριοποίησης των στηλών;"

- Οποιοδήποτε μοντέλο και να προσαρμόσουμε, ουσιαστικά εκφράζουμε την δομή των δεδομένων μέσω ενός αριθμού παραμέτρων. Οσο πιο περίπλοκη είναι, τόσο περισσότερες παραμέτρους έχουμε και αντίθετα. Το πιο περίπλοκο μοντέλο που μπορούμε να προσαρμόσουμε έχει τον ίδιο αριθμό παραμέτρων με τις ανεξάρτητες μονάδες δεδομένων : το ονομάζουμε *κορεσμένο* μοντέλο.

- Στρέφουμε τώρα την προσοχή μας στην εκτίμηση παραμέτρων. Μόλις εκτιμήσουμε τις παραμέτρους, πρέπει να εκτιμήσουμε την επάρκεια του μοντέλου.

2.4 Εκτιμήτριες μεγίστης πιθανοφάνειας (ΕΜΠ)

2.4.1 ΕΜΠ - επανάληψη

Έστω $L(\mathbf{X} \mid \mathbf{\Theta})$ η συνάρτηση πιθανοφάνειας, όπου με $\mathbf{X}$ συμβολίζουμε τα δεδομένα και με $\mathbf{\Theta} = (\Theta_1, \ldots, \Theta_p)$ το διάνυσμα των παραμέτρων. Αν η $L(\mathbf{X} \mid \mathbf{\Theta})$ αποκτά ολικό μέγιστο στο $\mathbf{\Theta} = \hat{\mathbf{\Theta}}$ τότε $\hat{\mathbf{\Theta}}$ είναι η ΕΜΠ του $\mathbf{\Theta}$ και $L(\mathbf{X} \mid \hat{\mathbf{\Theta}})$ είναι η μεγιστοποιημένη πιθανοφάνεια. Σημειώστε ότι $\hat{\mathbf{\Theta}}$ μεγιστοποιεί επίσης την $\log L(\mathbf{X} \mid \mathbf{\Theta})$. Οι ΕΜΠ έχουν εν γένει (όχι μόνο για κανονικές πιθανοφάνειες) επιθυμητές ιδιότητες:

(i) Συνέπεια ($\lim_{n\to\infty} p(\mid T_n - \Theta \mid \geq \varepsilon) = 0 \;\; \forall \varepsilon > 0$)

(ii) Ασυμπτοτικά κανονική ($\hat{\mathbf{\Theta}} \sim M \vee n_p(\mathbf{\Theta}, I(\mathbf{\Theta})^{-1})$, όπου ο $I(\mathbf{\Theta})$ έχει στοιχείο $(jk) = -E[\frac{\theta^2 \log L}{\theta \Theta_j \theta \Theta_k}]$.

(iii) Ασυμπτοτικά επαρκής

2.4.2 Παράδειγμα:

Οι ΕΜΠ για τις παραμέτρους της πολυωνυμικής κατανομής έχουν "ιδιαίτερα" προβλήματα :

Έστω $\mathbf{N} \sim M_k(n, \mathbf{p})$, και παρατηρούμε $\mathbf{N} = y$.

Θέλουμε να βρούμε την ΕΜΠ $\hat{\mathbf{p}}$.

$$\text{Πιθανοφάνεια} = \frac{n!}{\prod_{i=1}^{k} n_i!} \prod_{i=1}^{k} p_i^{n_i} \propto \prod_{i=1}^{k} p_i^{n_i} \qquad (*)$$

Η (*) επιτυγχάνει μέγιστο για $p_1 = p_2 = \cdots = p_k = 1$, αλλά αυτό δεν ισχύει γιατί πρέπει $\sum_{i=1}^{k} p_i = 1 \cdot 0$. Επειδή έχουμε τον περιορισμό $\sum_{i=1}^{k} p_i = 1$, χρησιμοποιούμε τους πολλαπλασιαστές Lagrance, δηλαδή μεγιστοποιούμε την συνάρτηση

$$\underbrace{\sum_{i=1}^{k} n_i \log p_i}_{\log - \text{Πιθανοφάνεια}} \quad - \quad \underbrace{\lambda}_{\text{Πολ/στής Lagrance}} \quad \underbrace{\left(\sum_{i=1}^{k} p_i - 1\right)}_{\text{περιορισμός}}$$

Παραγωγίζοντας ως προς p_i, $i = 1, 2, \ldots, k$ και εξισώνοντας με το μηδέν έχουμε $\frac{d}{dp_i} = 0 \Rightarrow n_i/\hat{p}_i - \lambda = 0$. Με άθροισμα όλων των i έχουμε $\sum n_i = \lambda(\sum \hat{p}_i) = \lambda \Rightarrow \lambda = n \Rightarrow \hat{p}_i =$

$\dfrac{n_i}{n}$, $i = 1, 2, \ldots, k$, και η μεγιστοποιημένη πιθανοφάνεια δίνεται από την σχέση

$$\frac{n!}{\prod_{i=1}^{k} n_i!} \prod_{i=1}^{k} \left(\frac{n_i}{n}\right)^{n_i}$$

Άρα σε αυτή την περίπτωση έχουμε υπολογίσει τις ΕΜΠ του κορεσμένου μοντέ λου, δηλαδή οι προσαρμοσμένες τιμές δίνονται από $np_i = n_i =$ οι παρατηρούμενες τιμές.

Αυτό το μοντέλο είναι λίγης/καθόλου αξίας στην πράξη, γιατί δεν "απλουστεύει" καθόλου τα αρχικά δεδομένα.

Εν γένει, θα έχουμε:

$$\mathbf{p} = \psi(\phi),$$

όπου $dim(\phi) = r < k - 1$.

Αργότερα θα δούμε μοντέλα που δεν έχουν αναλυτική λύση αλλά μπορούν να λυθούν αριθμητικά (μέσω π.χ. κάποιου στατιστικού πακέτου).

Τώρα θα εξετάσουμε την προσαρμογή του μοντέλου.

2.5 Pearson X^2 - στατιστικό

Το Pearson X^2 στατιστικό είναι χρήσιμο για να μετρήσουμε την προσαρμογή ενός μοντέλου συγκρινόμενου με το κορεσμένο μοντέλο.

$$\text{ΟΡΙΣΜΟΣ :} \qquad X^2 = \sum_{i=1}^{k} \frac{(n_i - \hat{m}_i)^2}{\hat{m}_i} \ ,$$

όπου $n_i =$ παρατηρούμενη συχνότητα (που αντιστοιχεί στο κορεσμένο μοντέλο)
$\hat{m}_i =$ προσαρμοσμένη συχνότητα.
Έστω H_0: μειωμένο μοντέλο
H_1: κορεσμένο μοντέλο
Έστω $s + t =$ αριθμός παραμέτρων στο κορεσμένο μοντέλο
και $s =$ αριθμός παραμέτρων στο μειωμένο μοντέλο

Αν η H_0 είναι αληθής, τότε (ασυμπτωτικά)

$$X^2 \sim \mathcal{X}_t^2$$

2.6 Ελεγχος πηλίκου πιθανοφανειών

Έστω ότι ενδιαφερόμαστε να εκτιμήσουμε την επάρκεια ενός μοντέλου που περιγράφει ένα σύνολο δεδομένων. Αυτό μπορεί να γίνει συγκρίνοντας την πιθανοφάνεια που προκύπτει απο το κορεσμένο μοντέλο, δηλαδή το μοντέλο που έχει αριθμό παραμέτρων ίσο με αριθμό δεδομένων (άρα οι προσαρμοσμένες τιμές είναι ίσες με τις παρατηρούμενες τιμές).

Έστω $\hat{\mathbf{\Theta}}_S$ η ΕΜΠ των παραμέτρων στο κορεσμένο (Saturated) μοντέλο και $\hat{\mathbf{\Theta}}_R$ η ΕΜΠ των παραμέτρων στο μειωμένο (Reduced) μοντέλο. Επίσης, έστω $L(\hat{\mathbf{\Theta}}_S)$ και $L(\hat{\mathbf{\Theta}}_R)$ οι μεγιστοποιημένες πιθανοφάνειες αντίστοιχα. Αν το μοντέλο παρέχει καλή προσαρμογή στο μοντέλο, θα περιμέναμε η $L(\hat{\mathbf{\Theta}}_R)$ να είναι περίπου τόσο μεγάλη όσο η $L(\hat{\mathbf{\Theta}}_S)$.

Έστω H_0 : Μειωμένο μοντέλο H_1 : Κορεσμένο μοντέλο

Οι παραπάνω υποθέσεις ελέγχονται ορίζοντας το στατιστικό πηλίκο πιθανοφανειών:

$$\Omega = \frac{L(\hat{\mathbf{\Theta}}_R)}{L(\hat{\mathbf{\Theta}}_S)}$$

Τιμές του Ω κοντά στο 1 υποδηλώνουν ότι η H_0 μπορεί να παρέχει καλή περιγραφή των δεδομένων, ενώ τιμές του Ω κοντά στο 0 υποδηλώνουν σημαντική έλλειψη προσαρμογής κατά την μετακίνηση από το κορεσμένο στο μειωμένο μοντέλο.

Για να ελέγξουμε τις υποθέσεις πρέπει να γνωρίζουμε την κατανομή του Ω όταν η H_0 είναι αληθής.

ΟΡΙΣΜΟΣ: Ορίζουμε την *Deviance* σαν

$$\text{Deviance} = D = -2\log\Omega.$$

Ασυμπτωτικά, όταν η H_0 είναι αληθής, η Deviance έχει κατανομή $\mathcal{X}^2_{n-p}$, όπου n είναι ο αριθμός των παρατηρήσεων και p ο αριθμός των παραμέτρων που ορίζονται από την H_0. Σημειώστε ότι

$$D = -2\log\left(\frac{L(\hat{\mathbf{\Theta}}_R)}{L(\hat{\mathbf{\Theta}}_S)}\right) = 2(\log L(\hat{\mathbf{\Theta}}_S) - 2\log L(\hat{\mathbf{\Theta}}_R))$$

Το παραπάνω αποτέλεσμα ισχύει ΑΣΥΜΠΤΩΤΙΚΑ για πολυωνυμική, Poisson, Gamma κ.λ.π.

Άρα έχουμε μία στρατηγική για προσαρμογή μοντέλων :

1. Προτείνουμε ένα μοντέλο
2. Εκτιμούμε παραμέτρους του μοντέλου με ΕΜΠ
3. Ελέγχουμε την επάρκεια του μοντέλου με τη βοήθεια της Deviance.

2.6.1 Παράδειγμα: Deviance για πολυωνυμική πιθανοφάνεια

$\mathbf{N} \sim M_k(n_+, \mathbf{p})$, παρατηρούμε $\mathbf{N} = \mathbf{n}$. Έστω οτι προτείνεται μοντέλο με $r(< k-1)$ παραμέτρους για τα στοιχεία του $\mathbf{p}$. Γράφουμε $\mathbf{p} = \psi(\varphi)$, με $dim(\varphi) = r$. (π.χ. $k = 4, p_1 = p_2 = \phi_1, p_3 = \phi_2$. Σε αυτή την περίπτωση για το κορεσμένο μοντέλο έχουμε $k-1 = 3$ παραμέτρους και για το μειωμένο μοντέλο έχουμε 2 παραμέτρους).

Για να δούμε εάν το μειωμένο μοντέλο είναι ισχύουσα απλοποίηση πρέπει να υπολογίσουμε το στατιστικό ελέγχου $D = -2\log\Omega$, όπου

$$\Omega = \frac{\underset{\varphi}{max}\ \frac{n_+!}{\prod n_i!}\prod_{i=1}^{k}\psi_i(\hat{\varphi})^{n_i}}{\underset{\varphi}{max}\ \frac{n_+!}{\prod n_i!}\prod_{i=1}^{k}\hat{p}_i^{n_i}}$$

Έχουμε ήδη εξετάσει τον παρανομαστή στο παράδειγμα 3.4.2 (δηλ. $\hat{p}_i = \frac{n_i}{n_+}$). Έχουμε οτι

$$-2\log\Omega = -2\sum_{i=1}^{k} n_i \log\psi_i(\hat{\varphi}) + 2\sum_{i=1}^{k} n_i \log\hat{p}_i =$$

$$= \ -2\sum_{i=1}^{k} n_i \log\left(\frac{\psi_i(\hat{\varphi}_i)}{\hat{p}_i}\right) =$$

$$= \ 2\sum_{i=1}^{k} n_i \log\left(\frac{n_i}{n_+\psi_i(\hat{\phi})}\right)$$

Αν συμβολίσουμε $\hat{m}_i = n_+\psi_i(\hat{\phi})$ την $i-$οστή προσαρμοσμένη τιμή κάτω από το μειωμένο μοντέλο τότε

$$\text{Deviance} = D = 2\sum_{i=1}^{k} n_i \log\left(\frac{n_i}{\hat{m}_i}\right)$$

Συνεπώς, για ένα πολυωνυμικό σέτ δεδομένων, αποκτούμε τις προσαρμοσμένες τιμές $(\hat{m}_i)$ με βάση την υπόθεση του μειωμένου μοντέλου, και μετά υπολογίζουμε την Deviance όπως περιγράφεται παραπάνω.

Συγκρίνουμε την τιμή που αποκτούμε με την κατανομή $\mathcal{X}^2_{k-1-r}$. Αν η Deviance είναι μεγάλη, τότε (πρός το παρόν !) απορρίπτουμε την H_0 και συμπεραίνουμε ότι (ίσως) το μειωμένο μοντέλο δεν είναι ισχύουσα περίληψη των δεδομένων.

ΣΗΜΕΙΩΣΗ: Τα παραπάνω είναι μόνο οδηγίες - προσοχή στα μικρά σετ δεδομένων !

2.6.2 Έλεγχος πηλίκου πιθανοφανειών για δεδομένα Poisson

Έστω οτι έχουμε k κελλιά, και k μέσους Poisson $m_i, i = 1,\ldots,k$. Για τον έλεγχο πηλίκου πιθανοφανειών πρέπει να υπολογίσουμε $D = -2\log\Omega$. Σύμφωνα με την Άσκηση 2, για το κορεσμένο μοντέλο ισχύει οτι η ΕΜΠ των m_i είναι $\hat{m}_i = n_i$.

Έστω οτι οι προσαρμοσμένες τιμές για το μειωμένο μοντέλο είναι $\tilde{m}_i$.

$$\Omega = \frac{\prod_{i=1}^{k} \frac{e^{-\tilde{m}_i - \tilde{m}_i{}^{n_i}}}{n_i!}}{\prod_{i=1}^{k} \frac{e^{-n_i} n_i^{n_i}}{n_i!}}$$

Συνεπώς

$$\begin{aligned} \log \Omega &= -\sum_i \tilde{m}_i + \sum_i n_i \log \tilde{m}_i - \sum_i \log n_i! + \sum_i n_i - \textstyle\sum_i n_i \log n_i + \\ &\quad + \sum_i \log n_i! = \\ &= \sum_i (-\tilde{m}_i + n_i) + \sum_i n_i (\log \tilde{m}_i - \log n_i) \quad . \end{aligned}$$

Άρα

$$\text{Deviance} = -2 \log \Omega = 2 \sum_i (\tilde{m}_i - n_i) + 2 \sum_i n_i \log \frac{n_i}{\tilde{m}_i}.$$

Παρατηρούμε οτι αν προσαρμόσουμε μοντέλα που διατηρούν το ολικό σύνολο καταλήγουμε στο

$$\text{Deviance} \quad = \quad 2 \sum_{i=1}^{k} n_i \log \frac{n_i}{\tilde{m}_i}$$

2.6.3 Σχέση μεταξύ ελέγχου πηλίκου πιθανοφανειών και Pearson's X^2.

- Έχουμε πολυωνυμική κατανομή:

$$D = -2 \log \Omega = 2 \sum_{i=1}^{k} n_i \log \left(\frac{n_i}{\hat{m}_i} \right)$$

Επίσης

$$X^2 = \sum_{i=1}^{k} \frac{(n_i - \hat{m}_i)^2}{\hat{m}_i}$$

Θεωρείστε την $f(x) = x \log(\frac{x}{y}) = x \log x - x \log y$
και έστω οτι αναπτύσσουμε την σειρά Taylor για

$$\begin{aligned} x &= y : f(x) \approx f(y) + (x - y) f'(y) + \tfrac{1}{2}(x - y)^2 f'' y \\ f(y) &= y \log(\tfrac{y}{y}) = 0 \\ f'(x) &= \log x + 1 - \log y \Rightarrow f'(y) = 1 \\ f''(x) &= \tfrac{1}{x} \Rightarrow f''(y) = \tfrac{1}{y} \end{aligned}$$

Άρα

$$f(x) \approx (x-y) + \frac{1}{2}\frac{(x-y)^2}{y} \Rightarrow$$

$$\Rightarrow \sum_{i=1}^{k} f(x_i) = \sum_{i=1}^{k} x_i \log(\frac{x_i}{y_i}) \approx \sum_{i=1}^{k}(x_i - y_i) + \frac{1}{2}\sum_{i=1}^{k}\frac{(x_i-y_i)^2}{y_i}$$

Μας ενδιαφέρει η ανάπτυξη του $D = 2\sum_{i=1}^{k} n_i \log\left(\frac{n_i}{\hat{m}_i}\right)$, και θέτοντας $x_i = n_i$, $y_i = \hat{m}_i$ έχουμε

$$D \approx 2\sum_{i=1}^{k}\left[(n_i - \hat{m}_i) + \frac{1}{2}\frac{(n_i-\hat{m}_i)^2}{\hat{m}_i}\right]$$

και επειδή για μοντέλα πολυωνυμικών δεδομένων έχουμε

$$\Sigma(n_i - \hat{m}_i) = 0 \Rightarrow D \approx \sum_{i=1}^{n}\frac{(n_i-\hat{m}_i)^2}{\hat{m}_i} = X^2.$$

2.6.4 Περίληψη σχημάτων δειγματοληψίας

Έστω, χωρίς βλάβη της γενικότητας, ένας IxJ πίνακας. Τα δεδομένα μπορεί να προέρχονται από ένα από τα παρακάτω σχήματα δειγματοληψίας:

ΣΧΗΜΑ ΔΕΙΓΜΑΤΟΛΗΨΙΑΣ	ΑΡ. ΠΑΡΑΜΕΤΡΩΝ
Πολυωνυμική κατανομή	$IJ-1$
Γινόμενο πολυωνυμικών κατανομών	$I(J-1)$ ή $J(I-1)$
Poisson κατανομή	IJ

Για κάθε κατανομή μπορούμε να χρησιμοποιήσουμε ΕΜΠ για την εκτίμηση των παραμέτρων του κορεσμένου μοντέλου:

ΣΧΗΜΑ ΔΕΙΓΜΑΤΟΛΗΨΙΑΣ	ΕΜΠ
Πολυωνυμική κατανομή	$\hat{p}_{ij} = \frac{n_{ij}}{n_{++}}$
Γινόμενο πολυωνυμικών κατανομών	$\hat{p}_{ij} = \frac{n_{ij}}{n_{i+}}$
Poisson	$\hat{m}_{ij} = n_{ij}$

Όμοια εκτιμούμε τις παραμέτρους οποιουδήποτε άλλου μοντέλου μας ενδιαφέ ρει με ΕΜΠ, και συγκρίνουμε τα δύο μοντέλα με έλεγχο πηλίκου πιθανοφα νειών:

ΣΧΗΜΑ ΔΕΙΓΜΑΤΟΛΗΨΙΑΣ	DEVIANCE
Πολυωνυμική κατανομή	$D = 2\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}\log(\frac{n_{ij}}{\hat{m}_{ij}})$
Γινόμενο πολυωνυμικών κατανομών	$D = 2\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}\log(\frac{n_{ij}}{\hat{m}_{ij}})$
Poisson κατανομή	$D = 2\sum_{i}^{I}\sum_{j}^{J}(n_{ij} - \hat{m}_{ij}) +$
	$+ \ 2\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}\log\left(\frac{n_{ij}}{\hat{m}_{ij}}\right)$

2.6.5 Σύγκριση εγκλωβισμένων (nested) μοντέλων με χρήση ελέγχου πηλίκου πιθανοφανειών.

Έστω ότι έχουμε την παρακάτω εγκλωβιστική ακολουθία μοντέλων:

$$M_0 \subset M_1 \subset M_S$$

όπου M_S είναι το κορεσμένο μοντέλο, και έστω ότι τα μοντέλα M_0 , M_1 και M_S έχουν αντίστοιχα $p, p+q$ και n παραμέτρους. Θέλουμε να ελέγξουμε τις υποθέσεις $H_0 : M_0$ κατά $H_1 : M_1$.

Επίσης, έστω D_0 η Deviance που αποκτούμε από την σύγκριση μεταξύ M_0 και M_s , D_1 η Deviance που αποκτούμε από τη σύγκριση μεταξύ M_1 και M_s, και $\hat{\mathbf{\Theta}}_0$, $\hat{\mathbf{\Theta}}_1$ και $\hat{\mathbf{\Theta}}_S$ οι ΕΜΠ για τα τρία μοντέλα αντίστοιχα. Μπορούμε να ελέγξουμε την H_0 κατά H_1 χρησιμοποιώντας την διαφορά των ελέγχων πηλίκου πιθανοφανειών:

$$\begin{aligned}
\Delta D &= D_0 - D_1 = \\
&= 2[\log L(\hat{\mathbf{\Theta}}_s;\mathbf{y}) - \log L(\hat{\mathbf{\Theta}}_0;\mathbf{y})] - 2[\log L(\hat{\mathbf{\Theta}}_s;\mathbf{y}) - \log L(\hat{\mathbf{\Theta}}_1;\mathbf{y})] = \\
&= 2[\log L(\hat{\mathbf{\Theta}}_1;\mathbf{y}) - \log L(\hat{\mathbf{\Theta}}_0;\mathbf{y})]
\end{aligned}$$

Αν και τα δύο μοντέλα που ορίζονται απο τις H_0 και H_1 περιγράφουν τα δεδομένα καλά, τότε $D_0 \sim X^2_{n-p}$ και $D_1 \sim X^2_{n-p-q}$. Συνεπώς (με την προϋπόθεση ότι ισχύουν ορισμένες υποθέσεις ανεξαρτησίας) $\Delta D \sim X^2_q$.

Το παραπάνω αποτέλεσμα σημαίνει ότι εάν η πραγματοποιηθείσα τιμή ΔD είναι συνεπής με την κατανομή X^2_q, θα επιλέγουμε (γενικά) το μοντέλο M_0 επειδή είναι απλούστερο. Εάν πάρουμε μια μεγάλη τιμή του ΔD, απορρίπτουμε την H_0 υπέρ της H_1, δηλαδή προτιμούμε το μοντέλο M_1 από το μοντέλο M_0. Σημειώστε ότι αυτό δεν σημαίνει απαραίτητα ότι το M_1 είναι καλή περιγραφή των δεδο μένων: αυτό μπορεί μόνο να εξακριβωθεί συγκρίνοντας την Deviance με αυτή του κορεσμένου μοντέλου.

Συμπέρασμα: Μπορούμε να προσαρμόσουμε το κορεσμένο μοντέλο, και μετά να αποσύρουμε σταδιακά όρους από το μοντέλο, ελέγχοντας σε κάθε στάδιο την μεταβολή της Deviance με τους αντίστοιχους βαθμούς ελευθερίας.

Σημειώστε επίσης ότι η προσθετική ιδιότητα της Deviance δεν μεταφέρεται στον έλεγχο Pearson's X^2: για κάθε μοντέλο ελέγχουμε την προσαρμογή κατά του κορεσμένου μοντέλου.

2.7 Μοντέλο ανεξαρτησίας για πολυωνυμική δειγματοληψία

Χωρίς βλάβη της γενικότητας, θεωρούμε πίνακα $I \times J$. Το μοντέλο ανεξαρτησίας είναι πολύ σπουδαίο, συνήθως το μόνο που ελέγχουμε. Ο συνήθης έλεγχος είναι
H_0: ανεξαρτησία της κατηγοριοποίησης γραμμών και στηλών
H_1: κορεσμένο μοντέλο.

Τί σημαίνει ανεξαρτησία γραμμών και στηλών σε σχέση με τις παραμέ τρους p_{ij} ;

$$P_r(\text{γραμμής } \ i \ \text{ και } \ \text{στήλης } \ j) = P_r(\text{γραμμής } \ i)P_r(\text{στήλης } \ j)$$

ή

$$p_{ij} = \theta_i \varphi_j \quad \text{όπου} \quad \sum_{i=1}^{I} \theta_i = \sum_{j=1}^{J} \phi_j = 1 \quad (1)$$

Για να υπολογίσουμε την D ή το X^2 πρέπει να εκτιμήσουμε $\hat{\theta}_i$ και $\hat{\phi}_j$ και μετά να υπολογίσουμε τις προσαρμοσμένες τιμές $\hat{m}_{ij} = n_{++}\hat{\theta}_i\hat{\phi}_j$. Τα $\hat{\theta}_i$ και $\hat{\phi}_j$ εκτιμώνται μεγιστοποιώντας την $\prod_{i=1}^{I}\prod_{j=1}^{J}(\theta_i\phi_j)^{n_{ij}}$ με την προϋπόθεση της (1).

Ισχύει ότι (Ασκηση 3)

$$\hat{m}_{ij} = \frac{n_{i+}n_{+j}}{n_{++}}$$

$\left(= n_{++}P_r(\text{γραμμή } \ i) \ \ P_r(\text{στήλη } \ j) = n_{++}\left(\frac{n_{i+}}{n_{++}}\right)\left(\frac{n_{+j}}{n_{++}}\right)\right)$

Οι βαθμοί ελευθερίας είναι $(IJ-1)-((I-1)+(J-1)) = (I-1)(J-1)$. Για γινόμενο πολυωνυμικών (με σταθερό το άθροισμα γραμμών), ανεξαρτησία συνεπάγεται $\mathbf{p}_1 = \mathbf{p}_2 = \ldots = \mathbf{p}_I$. Το κορεσμένο μοντέλο έχει $I(J-1)$ παραμέτρους, ενώ το μοντέλο ανεξαρτησίας έχει $J-1$ παραμέτρους, άρα οι βαθμοί ελευθερίας είναι $IJ-I-J+1 = (I-1)(J-1)$.

Αυτό το μοντέλο λέγεται μοντέλο *ομογένειας των γραμμών*. Ισχύει ότι η Deviace είναι η ίδια όπως και στην πολυωνυμική δειγματοληψία (Ασκηση 4). Αργότερα θα δείξουμε ότι και στη δειγματοληψία Poisson καταλήγουμε στην ίδια Deviance και στους ίδιους βαθμούς ελευθερίας.

3.1 Εισαγωγή

Ας αρχίσουμε θεωρώντας την απλούστερη περίπτωση, δηλαδή τον πίνακα 2x2:

Παράδειγμα:

	ΑΠΟΚΡΙΣΗ		
	ΝΑΙ	ΟΧΙ	
ΦΑΡΜΑΚΟ 1	n_{11}	n_{12}	n_{1+}
ΦΑΡΜΑΚΟ 2	n_{21}	n_{22}	n_{2+}
	n_{+1}	n_{+2}	n_{++}

Σημειώστε ότι όσον αφορά τον συμβολισμό, θα χρειαστεί να εναλλάσουμε μεταξύ του παραπάνω και της περίπτωσης που έχουμε N διωνυμικά πειράματα

	ΑΠΟΚΡΙΣΗ	
	Y	TOTAL
1	r_1	n_1
Πείραμα / Φάρμακο 2	r_2	n_2
$\vdots$	$\vdots$	$\vdots$
N	r_N	n_N

Ο παραπάνω πίνακας μπορεί να έχει αποκτηθεί μέσω προσδοκώμενης ή αναδρομικής μελέτης. Θα υποθέσουμε προς το παρόν, προσδοκώμενη δειγματοληψία. Η πιθανοφάνεια δίνεται από

$$P(n_{11}, n_{21} \mid p_1, p_2, n_{1+}, n_{2+}) \propto p_1^{n_{11}}(1-p_1)^{n_{12}} p_2^{n_{21}}(1-p_2)^{n_{22}}$$

όπου $p_i = P_r$ (Απόκριση | Φάρμακο i).

Το μοντέλο ανεξαρτησίας δίνεται από $p_1 = p_2 = p$. Πιο γενικά, ενδιαφερόμαστε για τη διαφορά $p_1 - p_2 = c$, οπότε για $c = 0$ παίρνουμε το μοντέλο ανεξαρτησίας.

Ασυμπτωτικά ισχύει ότι

$$\hat{p}_i \sim N\left(p_i, \frac{p_i(1-p_i)}{n_{i+}}\right),$$

όπου $\hat{p}_i = \frac{n_{i1}}{n_{i+}}$ η ΕΜΠ των p_i. Αν θεωρήσουμε $H_0 : p_1 - p_2 = p$ έχουμε συνεπώς

$$\frac{\frac{n_{11}}{n_{1+}} - \frac{n_{21}}{n_{2+}}}{\sqrt{\left(\frac{1}{n_{1+}} + \frac{1}{n_{2+}}\right) p(1-p)}} \sim N(0,1)$$

Αλλά το p είναι άγνωστο, και χρησιμοποιούμε την ΕΜΠ του p:

$$\hat{p} = \frac{n_{11} + n_{21}}{n_{1+} + n_{2+}} = \frac{n_{+1}}{n_{++}}.$$

Συνεπώς, μεγάλες τιμές του

$$\frac{\frac{n_{11}}{n_{1+}} - \frac{n_{21}}{n_{2+}}}{\sqrt{\left(\frac{1}{n_{1+}} + \frac{1}{n_{2+}}\right) \frac{n_{+1}}{n_{++}} \cdot \frac{n_{+2}}{n_{++}}}}$$

(π.χ. > 2) αποτελούν ένδειξη για μή αληθή H_0.

Σημειώστε ότι αυτό το αποτέλεσμα είναι ασυμπτωτικό, δηλαδή περιμένουμε καλή συμπεριφορά για “ μεγάλες ” συχνότητες και $p \in (0.2, 0.8)$, αλλά εάν το p είναι μικρό ή μεγάλο η υπόθεση της σταθερής διασποράς $\mathrm{var}(\hat{p}_1)$ είναι “ύποπτη”.

3.2 Κίνητρο για λογιστική παλινδρόμηση

Έστω ότι η απόκριση μετριέται σε συνεχή κλίμακα. Πώς θα αναλύαμε δεδομένα y_{ij} όπου $i = 1, 2$ δηλώνει την αγωγή και $j = 1, \ldots, n_i$ δηλώνει την επανάληψη;

Αν τα δεδομένα $\mathbf{y}$ ακολουθούσαν κανονική κατανομή θα χρησιμοποιούσαμε ένα πίνακα ANOVA κατά ένα παράγοντα:

$$\begin{aligned} y_{1j} &= \mu + \varepsilon_{1j} \ , \quad j = 1, \ldots, n_1 \\ y_{2j} &= \mu + \alpha_2 + \varepsilon_{2j} \ , \quad j = 1, \ldots, n_2 \end{aligned}$$

Σημειώστε ότι $\mu \in (-\infty, \infty)$ και $\alpha_2 \in (-\infty, \infty)$.

Θα μπορούσαμε να προσπαθήσουμε μια παρόμοια προσέγγιση με τα p:

$$\begin{aligned} E(p_1) &= \mu \\ E(p_2) &= \mu + \alpha_2 \end{aligned}$$

Αλλά λόγω του $0 \leq r_i/n_i \leq 1$ υπάρχουν περιορισμοί στα μ, α_2. Έστω ότι έχουμε N διωνυμικές προσπάθειες και συμμεταβλητή X (DATASET 1)

$$p_i = \alpha + \beta x_i$$

Χρειαζόμαστε

Προφανή μειονεκτήματα είναι η δυσκολία εξαγωγής συμπερασμάτων (extrapolate) και η περιορισμένη ερμηνεία. Σημειώστε ακόμα ότι η διασπορά του $r_i = n_i p_i$ εξαρτάται από το p_i, άρα, αν χρειαζόμαστε κανονικούς ελέγχους, η υπόθεση της σταθερής διασποράς δεν ισχύει.

3.3 Ανεπίσημη εισαγωγή στα γενικευμένα γραμμικά μοντέλα

Στα γραμμικά μοντέλα υποθέσαμε

(1) $y_i = \mu_i + \varepsilon_i$, $\varepsilon_i \sim N(0, \sigma^2)$, ανεξάρτητα

(2) Γραμμική προβλέπουσα $\mathbf{n} = \sum_{j=1}^{p} \beta_j X_j$

(3) $\mathbf{n} = \boldsymbol{\mu}$: ταυτοτικός σύνδεσμος.

Εισάγουμε τώρα μια νέα κατηγορία μοντέλων, τα *γενικευμένα γραμμικά μοντέλα*. Αυτά τα μοντέλα υποθέτουν ότι το (2) είναι πάντα αληθές, αλλά όχι τα (1) και (3).

Στην (3) ο σύνδεσμος μεταξύ των αναμενόμενων δεδομένων και της συνάρτησης του μοντέλου είναι ο ταυτοτικός. Ορίζουμε τώρα την συνάρτηση συνδέσμου (link function) g τέτοια ώστε $\mathbf{n} = g(\boldsymbol{\mu})$. Είδαμε ήδη τα προβλήματα που παρουσιάζονται στα διωνυμικά δεδομένα όταν υποθέτουμε ταυτοτικό σύνδεσμο. Τώρα, στην περίπτωση των διωνυμικών δεδομένων, θέλουμε να συνδέσουμε με κάποιο μοντέλο την σχέση μεταξύ των p_i και την πιθανή παρουσία συμμεταβλητών, δηλαδή αν $E(\frac{r}{n}) = p$ θέλουμε $g(p) = \sum_{j=1}^{p} \beta_j X_j$.

3.3.1 Συναρτήσεις συνδέσμου για δυωνυμικά δεδομένα

Ας δούμε πάλι το DATASET 1, και έστω ότι σχεδιάζουμε τα ποσοστά p_i κατά των συμμεταβλητών x_i.

ΙΔΕΑ: Μήπως υπάρχει ομοιότητα με το γράφημα της αθροιστικής συνάρτησης κατανομής ;

Θα μπορούσαμε να ορίσουμε την συνάρτηση συνδέσμου να είναι $g(p) = \Phi^{-1}(p) = x$, αλλά αυτό είναι πολύ περιοριστικό, δεν μας δίνει ελαστικότητα γιατί η γραμμή είναι σταθερή. Θεωρείστε την

$$g(p) = \Phi^{-1}(p) = \alpha + \beta x$$

όπου το α μας επιτρέπει να κινούμαστε στον άξονα των x και το β ρυθμίζει την κλίση, σε αντιστοιχία με την απλή γραμμική παλινδρόμηση. Αυτός ο σύνδεσμος είναι γνωστός σαν probit σύνδεσμος. Σημειώστε την συμμετρία επειδή $g(p) = -g(1-p)$.

3.4 Ερμηνεία των συναρτήσεων συνδέσμου για διωνυμικά δεδομένα

Έστω ότι εκτελούμε ένα πείραμα για να καθορίσουμε την τοξικότητα κάποιας ουσίας. Δίνουμε επίπεδα συγκέντρωσης X_i, $i = 1, \ldots, N$ της ουσίας σε n_i πειραματικές μονάδες, και παρατηρούμε κάποια διωνυμική απόκριση (π.χ. θάνατος). Για κάθε μονάδα υπάρχει ένα επίπεδο συγκέντρωσης κάτω από το οποίο συμβαίνει η απόκριση και πάνω από το οποίο η απόκριση δεν συμβαίνει. Αυτή η τιμή ονομάζεται *ανοχή* (Tolerance), και την συμβολίζουμε με λ. Άρα

$$\begin{array}{ll} y = 0 & \text{αν} \quad \lambda < x_i \\ y = 1 & \text{αν} \quad \lambda \geq x_i \end{array}$$

Θεωρούμε τώρα τον πληθυσμό όλων των μονάδων, και προφανώς μέσα σε αυτόν τον πληθυσμό η ανοχή ποικίλλει. Έστω $f(\lambda)$ η σ.π.π. των ανοχών. Στον πληθυσμό, το ποσοστό που αποκρίνεται σε δόση μεγέθους x_i είναι συνεπώς

$$p_i = P_r(\lambda \leq x_i) = \int_0^{x_i} f(\lambda) d\lambda$$

(Δηλαδή εάν δεν υπήρχε μεταβλητότητα των ανοχών στον πληθυσμό, όλα τα x_i θα ήταν ή μικρότερα της ανοχής και θα είχαμε $p_i = 0$ ή μεγαλύτερα και θα είχαμε $p_i = 1$).

Τώρα θεωρείστε επίπεδα συγκέντρωσης $x_1 < x_2 < \ldots < x_N$ (πιθανώς μετά από κάποιο μετασχηματισμό, π.χ. log) και έστω ότι η κατανομή ανοχής είναι $\lambda \sim N(\mu, \sigma^2)(= f(\lambda))$. Για $i = 1, \ldots, N$, n_i πειραματικές μονάδες επιλέγονται από τον πληθυσμό και λαμβάνουν επίπεδα συγκέντρωσης x_i. Η πιθανότητα της απόκρισης δίνεται από

$$\begin{aligned} p_i &= P_r(\lambda \leq x_i) = \\ &= \int_{-\infty}^{x_i} \frac{1}{\sqrt{2\pi\sigma^2}} \ exp \ \left[-\frac{1}{2\sigma^2}(\lambda - \mu)^2\right] d\lambda = \Phi\left(\frac{x_i - \mu}{\sigma}\right) \end{aligned}$$

Ο αριθμός των αποκρίσεων Y_i έχει διωνυμική κατανομή με παραμέτρους n_i και p_i, δηλ. $Y_i \sim B(n_i, P_i)$, $i = 1, \ldots, N$. Συνεπώς

$$\Phi^{-1}(p_i) = \frac{x_i}{\sigma} - \frac{\mu}{\sigma} = \beta x_i + \alpha$$

όπου $\alpha = \frac{-\mu}{\sigma}$ και $\beta = \frac{1}{\sigma}$. Άρα η συνάρτηση συνδέσμου $g(\ .\ )$ είναι $\Phi^{-1}(\ .\ )$, και αυτό το μοντέλο είναι ένα παράδειγμα του μοντέλου probit. Η τιμή $x = \mu$ καλείται η διάμεσος θανάσιμη δόση, (median lethal dose), συμβολίζεται LD50, και αντιστοιχεί στην απαιτούμενη δόση να σκοτώσει κατά μέσο όρο τις μισές πειραματικές μονάδες.

Εάν η κατανομή ανοχής $f(\lambda)$ είναι ομοιόμορφη στο διάστημα $[\lambda_1, \lambda_2]$, τότε

$$p_i = \int_{\lambda_1}^{x_i} f(\lambda) d\lambda = \frac{x_i - \lambda_1}{\lambda_2 - \lambda_1} \quad \text{για} \quad \lambda_1 \leq x_i \leq \lambda_2$$

που είναι της μορφής $p_i = \alpha + \beta x_i$ με

$$\alpha = -\frac{\lambda_1}{\lambda_2 - \lambda_1} \quad \text{και} \quad \beta = \frac{1}{\lambda_2 - \lambda_1}$$

- Θυμηθείτε την συζήτηση της ανεπάρκειας αυτού του μοντέλου στην παράγραφο 3.2.

Ένα μοντέλο που δίνει αριθμητικά αποτελέσματα παρόμοια του μοντέλου probit, αλλά που είναι υπολογιστικά ευκολότερο, είναι το λογιστικό (logistic ή logit) μοντέλο. Η κατανομή ανοχής είναι η λογιστική κατανομή.

$$f(\lambda) = \frac{\beta\ exp(\alpha + \beta\lambda)}{[1 + exp\ (\alpha + \beta\lambda)]^2}, \quad -\infty < \lambda < \infty.$$

Σε αυτή τη περίπτωση

$$\begin{aligned} p_i &= P_r(\lambda \leq x_i) \\ &= \textstyle\int_{-\infty}^{x_i} \frac{\beta\ exp\ (\alpha + \beta\lambda)}{[1 +\ exp\ (\alpha + \beta\lambda)]^2}\ d\lambda \end{aligned}$$

$$= \left[\frac{exp\,(\alpha + \beta\lambda)}{1 + exp(\alpha + \beta\lambda)}\right]_{-\infty}^{x_i} = \frac{exp(\alpha + \beta x_i)}{1 + exp(\alpha + \beta x_i)}$$

που δίνει την συνάρτηση συνδέσμου

$$\log\left(\frac{p_i}{1 - p_i}\right) = \alpha + \beta x_i$$

δηλαδή ένα γραμμικό μοντέλο σε σχέση με τα $\log\left(\frac{p_i}{1 - p_i}\right)$ (= logits). Αυτό το μοντέλο χρησιμοποιείται πολύ συχνά στα διωνυμικά δεδομένα.

Άλλες συναρτήσεις συνδέσμου προέρχονται από διαφορετικές αθροιστικές συναρτήσεις κατανομών. Για παράδειγμα, από την κατανομή ακροτάτων (extreme value distribution)

$$f(\lambda) = \beta exp[(\alpha + \beta\lambda) - exp(\alpha + \beta\lambda)]$$

οδηγούμαστε στην

$$p_i = 1 - exp(1 - e^{\alpha + \beta\chi_i}) \Rightarrow\Rightarrow \log[-\log(1 - p_i)] = \alpha + \beta x_i.$$

Ο παραπάνω σύνδεσμος είναι γνωστός σαν *συμπληρωματικός log-log σύνδεσμος* (Complementary log - log link function). Αυτό το μοντέλο είναι παρόμοιο με το λογιστικό και probit για τιμές του p κοντά στο 1/2, αλλά διαφέρει κοντά στο 0 και 1. Σημειώστε επίσης ότι είναι συμμετρικό γύρω από το $p = 1/2$ σαν τις συναρτήσεις probit και λογιστική.

3.4.1 Σχόλια

(1) Πολλοί επιστήμονες, κυρίως επιδημιολόγοι, προτιμούν να σκέφτονται χρησιμοποιώντας λόγους συμπληρωματικών πιθανοτήτων (odds):

$$O_i = \frac{p_i}{1-p_i}, \quad i = 1, \ldots, N$$

π.χ. 3 προς 1 υπέρ $\Rightarrow \frac{p_i}{1-p_i} = 3 \Rightarrow p_i = \frac{3}{4}$
3 προς 1 κατά $\Rightarrow \frac{p_i}{1-p_i} = \frac{1}{3} \Rightarrow p_i = \frac{1}{4}$

Επειδή $0 \leq O_i \leq \infty$ ο λογικός σύνδεσμος λύνει πιθανά προβλήματα. Προσέξτε τώρα ότι

$$\begin{aligned}
\log\left(\tfrac{p_i}{1-p_i}\right) = \operatorname{logit}(p_i) &\in (-\infty, \infty) \ \text{και} \\
\log\left(\tfrac{p_i}{1-p_i}\right) &= \log 3 = 1.1 \\
\log\left(\tfrac{p_i}{1-p_i}\right) &= \log(\tfrac{1}{3}) = -\log 3 = -1.1
\end{aligned}$$

(2) Για 2x2 πίνακες, η χρήση του logit συνδέσμου προτρέπει σε μία προφανή ποσότητα που είναι ενδιαφέρουσα, την διαφορά των logit:

$$\begin{aligned}
\Delta &= \log\left(\tfrac{p_1}{1-p_1}\right) - \log\left(\tfrac{p_2}{1-p_2}\right) \\
&= \log\left(\tfrac{p_1(1-p_2)}{p_2(1-p_1)}\right)
\end{aligned}$$

Θα ονομάζουμε το Δ "Διαφορά λόγου συμπληρωματικών πιθανοτήτων". Προσέξτε ότι το Δ μας δηλώνει απλώς την *διαφορά* μεταξύ των λόγων συμπληρωματικών πιθανοτήτων, δεν μας δίνει πληροφορίες για τα p_i. Για παράδειγμα, έστω $\Delta = 1$. Τότε εάν $p_1 = 0.1 \Rightarrow p_2 = 0.04$, ενώ εάν $p_1 = 0.5 \Rightarrow p_2 = 0.27$. Αυτό βέβαια είναι λογικό, μια και χρειαζόμαστε τουλάχιστον δύο ποσότητες από τις p_1, p_2 και Δ για να καθορίσουμε εντελώς την κατανομή.

3.5 Εμπειρικός λογιστικός μετασχηματισμός

Έχουμε δεί ότι ο κύριος σύνδεσμος για τα διωνυμικά δεδομένα είναι ο λογιστικός σύνδεσμος $g(p_i) = \log(\frac{p_i}{1-p_i})$. Συνήθως αρχίζουμε την ανάλυση αποκτώντας εκτιμήτριες των logit δια μέσου των εκτιμητριών του δείγματος. Αυτές είναι χρήσιμες για τον σχεδιασμό κατά των συμμεταβλητών ή παραγόντων για την εξαγωγή συμπερασμάτων για πιθανές σχέσεις. Επειδή εκτιμούμε p_i με $\hat{p}_i = \frac{r_i}{n_i}$, μια προφανής εκτιμήτρια θα ήταν

$$Z_i = \log\left(\frac{\frac{r_i}{n_i}}{1-\frac{r_i}{n_i}}\right) = \log\left(\frac{r_i}{n_i - r_i}\right)$$

Δυστυχώς υπάρχει ένα πιθανό πρόβλημα, στις περιπτώσεις που $r_i = 0$ ή $r_i = n_i$. Για αυτό ορίζουμε τον *τροποποιημένο εμπειρικό logit* , με

$$Z_i = \log\left(\frac{r_i + \frac{1}{2}}{n_i - r_i + \frac{1}{2}}\right)$$

3.6 Συμπερασματολογία για λογιστική παλινδρόμηση

Αν εκτιμήσουμε τις παραμέτρους (έστω α, β) με ΕΜΠ, τότε οι εκτιμήτριες είναι ασυμπτωτικά κανονικές. Συνεπώς διαστήματα εμπιστοσύνης για τις παραμέτρους έχουν την μορφή

$$\hat{\beta} \ \pm \ Z_{\alpha/2} \times \text{τυπ.σφάλμα} \ \ (\hat{\beta})$$

όπου

$$Z_{\alpha/2} = \int_{-\infty}^{Z_{\alpha/2}} \phi(t)dt = 1 - \alpha/2.$$

Η παραπάνω σχέση επιτρέπει προσεγγιστικούς ελέγχους, όπως $H_0 : \beta = 0$ κ.λ.π. Μπορούμε επίσης να χρησιμοποιήσουμε τον έλεγχο πηλίκου πιθανοφανειών, π.χ. να βρούμε την Deviance μεταξύ δύο εγκλωβισμένων μοντέλων και να την συγκρίνουμε με την X_a^2 κατανομή με βαθμούς ελευθερίας ίσους με την διαφορά διαστάσεων μεταξύ των δύο μοντέλων.

3.7 ΕΜΠ για λογιστική παλινδρόμηση

Θεωρείστε το μοντέλο

$$logit\left(\frac{p_i}{1-p_i}\right) = \alpha + \beta x_i, \quad i = 1, \ldots, N$$

Έστω ότι ενδιαφερόμαστε για τις ΕΜΠ για α και β και έστω $L(\alpha, \beta)$ η πιθανοφάνεια.

$$\begin{array}{rcl} L(\alpha,\beta) & \propto & \prod_{i=1}^{N} p_i^{r_i}(1-p_i)^{n_i-r_i} \Rightarrow \\ \Rightarrow \log L & = & \sum_{i=1}^{N}\left[r_i \log p_i + (n_i - r_i)\log(1-p_i)\right] \end{array}$$

Τώρα,

$$p_i = \frac{e^{\alpha+\beta x_i}}{1+e^{\alpha+\beta x_i}} \quad .$$

Άρα,

$$\begin{array}{rcl} \log L & = & \sum_{i=1}^{N} r_i \log\left(\frac{e^{\alpha+\beta x_i}}{1+e^{\alpha+\beta x_i}}\right) + (n_i - r_i)\log\left(\frac{1}{1+e^{\alpha+\beta x_i}}\right) = \\ & = & \sum_{i=1}^{N} r_i(\alpha+\beta x_i) - n_i \log(1+e^{\alpha+\beta x_i}) \quad . \end{array}$$

Άρα,

$$\begin{array}{rcl} \frac{\theta \log L}{\theta \alpha} & = & 0 \Rightarrow \sum_{i=1}^{N}\left(r_i - \frac{n_i e^{\hat{\alpha}+\hat{\beta}x_i}}{1+e^{\hat{\alpha}+\hat{\beta}x_i}}\right) = 0 \\ \frac{\theta \log L}{\theta \beta} & = & 0 \Rightarrow \sum_{i=1}^{N}\left(r_i x_i - \frac{n_i x_i e^{\hat{\alpha}+\hat{\beta}x_i}}{1+e^{\hat{\alpha}+\hat{\beta}x_i}}\right) = 0 \end{array}$$

Οι παραπάνω εξισώσεις δεν είναι δυνατόν να λυθούν αναλυτικά, αλλά λύνονται διαμέσου στατιστικών πακέτων αριθμητικά.

3.8 Τύποι μοντέλων λογιστικής παλινδρόμισης

Σε αναλογία με τα κανονικά γραμμικά μοντέλα, μπορούμε να προσαρμόσουμε τα παρακάτω λογιστικά μοντέλα:

$$\begin{array}{llcll} (A) & \log\left(\frac{p_i}{1-p_i}\right) & = & \alpha + \beta x_i \quad , & i = 1, \ldots, N \\ (B) & \log\left(\frac{p_i}{1-p_i}\right) & = & \left\{ \begin{array}{ll} \mu & , \quad i = 1 \\ \mu + \alpha_i & \quad\;\; i = 2, \ldots, N \end{array} \right\} & \end{array}$$

Το (A) αντιστοιχεί στην γραμμική παλινδρόμηση και το (B) στην ανάλυση διασποράς κατά ένα παράγοντα. Έστω ότι έχουμε πειράματα με μια συμμεταβλητή (π.χ. επίπεδο δοσοληψίας ή ηλικία) και ένα παράγοντα (π.χ. φύλο). Έστω ότι $i = 1, \ldots, N$ συμβολίζει τις συμμεταβλητές και $j = 1, 2$ τα επίπεδα του παράγοντα.

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \left\{ \begin{array}{ll} \alpha_1 + \beta x_i \;\; , \;\; i = 1, \ldots, N \;\; , \;\; j = 1 \\ \alpha_2 + \beta x_i \;\; , \;\; i = 1, \ldots, N \;\; , \;\; j = 2 \end{array} \right.$$

Εδώ έχουμε ένα σαφή συσχετισμό με μοντέλα ANCOVA (ανάλυση συνδιακύμανσης). Μπορούμε επίσης να συμπεριλάβουμε άλλες περιπτώσεις όπως "μή παράλληλες ευθείες" κ.λ.π.

3.9 Μοντέλα λογιστικής παλινδρόμησης για πίνακες $2 \times I \times J$.

Έστω ότι έχουμε διωνυμική απόκριση με δύο παράγοντες A και B που έχουν I και J επίπεδα αντίστοιχα. Έστω επίσης p_{ij} η πιθανότητα επιτυχίας για το $(i, j)-$στό δυωνυμικό πείραμα. Θεωρούμε τα παρακάτω μοντέλα:

(1) M_1: Μηδενικό μοντέλο, (Null model)

$$logit(p_{ij}) = \mu, \quad i = 1, \ldots, I, \quad j = 1, \ldots, J$$

Το παραπάνω μοντέλο συνεπάγεται ότι η πιθανότητα επιτυχίας είναι ανεξάρτητη των δύο παραγόντων, και έχει μια παράμετρο.

(2) M_2: Η πιθανότητα επιτυχίας εξαρτάται μόνο από τον παράγοντα A :

$$logit(p_{ij}) = \left\{ \begin{array}{llll} \mu & , & i = 1 & ; \;\; j = 1, \ldots, J \\ \mu + \alpha_i & , & i = 2, \ldots, I & ; \;\; j = 1, \ldots, J \end{array} \right.$$

Οι παράμετροι εδώ είναι I.

(3) M_3: Η πιθανότητα επιτυχίας εξαρτάται από τον παράγοντα B μόνο:

$$logit(p_{ij}) = \left\{ \begin{array}{llll} \mu & , & i = 1, \ldots, I & ; \;\; j = 1 \\ \mu + \beta_j & , & i = 1, \ldots, I & ; \;\; j = 2, \ldots, J \end{array} \right.$$

Εδώ έχουμε J παραμέτρους.

(4) M_4 : Η πιθανότητα επιτυχίας εξαρτάται και από τους δύο παράγοντες:

$$logit(p_{ij}) = \begin{cases} \mu & i=1 \quad ; \quad j=1 \\ \mu+\alpha_i & i=2,\ldots,I \quad ; \quad j=1 \\ \mu+\beta_j & i=1 \quad ; \quad j=2,\ldots,J \\ \mu+\alpha_i+\beta_j & i=2,\ldots,I \quad ; \quad j=2,\ldots,J \end{cases}$$

Εδώ έχουμε $(I-1)+(J-1)+1$ παραμέτρους.

(5) M_5: Κορεσμένο μοντέλο.

$$logit(p_{ij}) = \begin{cases} \mu & i=1 \quad ; \quad j=1 \\ \mu+\alpha_i & i=2,\ldots,I \quad ; \quad j=1 \\ \mu+\beta_j & i=1 \quad ; \quad j=2,\ldots,J \\ \mu+\alpha_i+\beta_j+(\alpha\beta)_{ij} & i=2,\ldots,I \quad ; \quad j=2,\ldots,J \end{cases}$$

Οι παράμετροι του κορεσμένου μοντέλου είναι

$$(I-1)+(J-1)+1+(I-1)(J-1)=IJ.$$

Τα εγκλωβισμένα μοντέλα μπορούν να περιγραφούν γραφικά.

Ανάλογα μοντέλα μπορούμε να προτείνουμε και για άλλες συναρτήσεις συνδέσμου (π.χ. probit κ.λ.π.).

3.10 Σχόλια για την προσαρμογή των μοντέλων

(1) **Όροι αλληλεπίδρασης**
Αλληλεπίδραση πρώτου βαθμού : Ερμηνεύεται σαν την "δομή πέραν αυτής που δίνεται από την απλή προσθετικότητα των παραγόντων όταν δρούν χωριστά ".

Αλληλεπίδραση δεύτερου βαθμού: "η δομή επιπλέον της απλής προσθετικότητας των παραγόντων και της αλληλεπίδρασης πρώτου βαθμού".

Όλα τα μοντέλα που έχουμε θεωρήσει ικανοποιούν την *αρχή της περιθωριοποίησης* (principle of marginality). Εάν ένα μοντέλο περιέχει ένα συγκεκριμένο όρο, τότε περιέχει όλους τους *περιθώριους* όρους προς αυτόν: π.χ. εάν $A \cdot B$ είναι παρόν, τότε A και B είναι επίσης παρόντες. Συνήθως έχουμε συμβολισμό της μορφής

$$A*B = A+B+A\cdot B \Rightarrow A*B - A\cdot B \equiv A+B$$

(2) **Επίσημοι έλεγχοι**
Ο έλεγχος της επάρκειας του τρέχοντος μοντέλου κατά του κορεσμένου γίνεται με την γνωστή σύγκριση της Deviance με την κατάλληλη X^2 κατανομή. Σημειώστε ότι, αφαιρώντας όρους από διαφορετικά επίπεδα οδηγούμαστε σε διαφορετικές μεταβολές της Deviance. Για παράδειγμα, έστω ότι έχουμε 3 παράγοντες A, B και C. Εάν περνούμε από το μοντέλο $A+B+A\cdot B$ στο μοντέλο $A+B$, η Deviance δεν θα είναι απαραίτητα η ίδια με αυτή της σύγκρισης των μοντέλων $A+B+C+A\cdot B$ και $A+B+C$.

(3) **Εγκλωβισμένα μοντέλα για 3 παράγοντες**

Θα αναφερθούμε σε πίνακα $I \times J$ συχνοτήτων Poisson. Υπενθύμιση:

$$P_r(N_{ij} = n_{ij}) = \frac{e^{-m_{ij}} m_{ij}^{n_{ij}}}{n_{ij}!}, \quad i = 1, \ldots, I; j = 1, \ldots, J \; .$$

4.1 Μοντέλο ανεξαρτησίας για πίνακα $I \times J$

Συχνά το πρώτο μοντέλο που θεωρούμε είναι το μοντέλο ανεξαρτησίας, δηλ. αυτό που υποθέτει ότι οι κατηγορίες γραμμών και στηλών είναι ανεξάρτητες. Ορίζουμε την ανεξαρτησία σαν

Αναμενόμεν. αριθμός στο κελλί $(i, j) = k_i \times$ Αναμενόμεν. αριθμός στην στήλη j

δηλαδή

$$E(N_{ij}) = m_{ij} = k_i E(N_{+j})$$

αλλά

$$E(N_{+j}) = E(\sum_{i=1}^{I} N_{ij}) = \sum_{i=1}^{I} E(N_{ij}) = m_{+j}$$

(γιατί εάν $N_i \sim \text{Poisson}(m_i) \Rightarrow \sum_{i=1}^{k} N_i \sim \text{Poisson}(\sum_{i=1}^{k} m_i))$

$$\begin{aligned} &\Rightarrow \quad m_{ij} = k_i m_{+j} \quad \Rightarrow \\ &\Rightarrow \quad \textstyle\sum_{j=1}^{J} m_{ij} = \sum_{j=1}^{J} k_i m_{+j} \Rightarrow m_{i+} = k_i m_{++} \Rightarrow \\ &\Rightarrow \quad k_i = \tfrac{m_{i+}}{m_{++}} \quad . \end{aligned}$$

Συνεπώς, η ανεξαρτησία για πίνακα Poisson $I \times J$ συνεπάγεται

$$m_{ij} = \frac{m_{i+} m_{+j}}{m_{++}}.$$

Μπορεί να αποδειχθεί (αλλά είναι κοπιαστικό) ότι οι ΕΜΠ για δεδομένα Poisson είναι

$$\hat{m}_{ij} = \frac{n_{i+} n_{+j}}{n_{++}}.$$

Ο συνολικός αριθμός των παραμέτρων είναι $(I-1)+(J-1)+1$ (γιατί $\sum_i m_{i+} = \sum_j m_{+j} = m_{++}$ σύν την m_{++}) και η διαφορά από το κορεσμένο (residual degrees of freedom) είναι

$$IJ - \{(I-1)+(J-1)+1\} = (I-1)(J-1).$$

Επίσης, η Deviance είναι ίδια σύμφωνα με την 2.6.2. Άρα για πολυωνυμική, γινόμενο πολυωνυμικών και Poisson σχήματα δειγματοληψίας έχουμε τις ίδιες προσαρμοσμένες τιμές, τις ίδιες Deviance και ίδιους βαθμούς ελευθερίας. Αυτό εξηγεί γιατί ο " απλοϊκός " Pearson X^2 έλεγχος ανεξαρτησίας χρησιμοποιείται ανεξαρτήτως του σχήματος δειγματοληψίας.

Εκτός από την ανεξαρτησία, υπάρχουν πολλές άλλες σχέσεις που θα θέλαμε να εισάγουμε σε κάποιο μοντέλο δεδομένων Poisson. Θυμηθείτε ότι στα δεδομένα διωνυμικής κατανομής ενδιαφερόμαστε για μοντέλα που καθορίζουν σχέσεις των πιθανοτήτων "επιτυχίας " p_i. Όταν έχουμε δεδομένα Poisson προσπαθούμε να συσχετίσουμε τους μέσους των κελλιών m_{ij}. Έχουμε $m_{ij} > 0$ συνεπώς, σε σχέση με τα γενικευμένα γραμμικά μοντέλα, θέλουμε μια συνάρτηση συνδέσμου g τέτοια ώστε $-\infty < g(m_{ij}) < \infty$ και μετά προσαρμόζουμε μοντέλα της μορφής $g(m_{ij}) = \sum_k X_k \beta_k$. Τα μοντέλα που θα μελετήσουμε στην συνέχεια είναι ιδιαίτερα χρήσιμα γιατί εκτός από δεδομένα Poisson, μπορούν επίσης να χρησιμοποιηθούν για δεδομένα πολυωνυμικής κατανομής ή γινομένου πολυωνυμικών κατανομών.

4.2 Λογαριθμικά γραμμικά μοντέλα (log-linear models)

Η ανεξαρτησία είναι ισοδύναμη με $m_{ij} = \frac{m_{i+}m_{+j}}{m_{++}}$. Αυτό συνεπάγεται ότι για $i = 1, \ldots, I; j = 1, \ldots, J$,

$$\log m_{ij} = \log m_{i+} + \log m_{+j} - \log m_{++} \tag{4.1}$$

και ακολούθως στις παρακάτω εξισώσεις:

$$I^{-1}\sum_i \log m_{ij} = I^{-1}\sum_i \log m_{i+} + \log m_{+j} - \log m_{++} \tag{4.2}$$

$$J^{-1}\sum_j \log m_{ij} = \log m_{i+} + J^{-1}\sum_j \log m_{+j} - \log m_{++} \tag{4.3}$$

$$(IJ)^{-1}\sum_i\sum_j \log m_{ij} = I^{-1}\sum_i \log m_{i+} + J^{-1}\sum_j \log m_{+j} - \log m_{++} \tag{4.4}$$

$$\text{Από } (2) \Rightarrow \log m_{i+} = J^{-1}\sum_j \log m_{ij} - J^{-1}\sum_j \log m_{+j} + \log m_{++} \tag{4.5}$$

$$\text{Από } (1) \Rightarrow \log m_{+j} = I^{-1}\sum_i \log m_{ij} - I^{-1}\sum_i \log m_{i+} + \log m_{++} \tag{4.6}$$

Αντικαθιστώντας (3.5) και (3.6) στην (3.1) έχουμε:

$$
\begin{aligned}
\log m_{ij} &= \log m_{i+} + \log m_{+j} - \log m_{++} = \\
&= \left(J^{-1}\sum_j \log m_{ij} - J^{-1}\sum_j \log m_{+j} + \log m_{++}\right) + \\
&+ \left(I^{-1}\sum_i \log m_{ij} - I^{-1}\sum_i \log m_{i+} + \log m_{++}\right) - \log m_{++} \quad =
\end{aligned}
$$

$$
= \quad J^{-1}\sum_j \log m_{ij} + I^{-1}\sum_i \log m_{ij} - \left[I^{-1}\sum_i \log m_{i+} + J^{-1}\sum_j \log m_{+j} - \log m_{++}\right]
$$

Η σχέση μέσα στις αγκύλες λόγω της (3.4) ισούται με $(IJ)^{-1}\sum_i\sum_j \log m_{ij}$. Συνεπώς

$$
\log m_{ij} = J^{-1}\sum_j \log m_{ij} + I^{-1}\sum_i \log m_{ij} - (IJ)^{-1}\sum_i\sum_j \log m_{ij} \tag{4.7}
$$

Αν γράψουμε $\mu_{ij} = \log m_{ij}$, η (3.7) γίνεται

$$
\begin{aligned}
\mu_{ij} &= \bar{\mu}_{+j} + \bar{\mu}_{i+} - \bar{\mu}_{++} = \\
&= \bar{\mu}_{++} + (\bar{\mu}_{i+} - \bar{\mu}_{++}) + (\bar{\mu}_{+j} - \bar{\mu}_{++})
\end{aligned} \tag{4.8}
$$

Συνεπώς, έχουμε εκφράσει το μοντέλο ανεξαρτησίας σαν αθροιστικό στην $\log m_{ij}$ κλίμακα. Μπορούμε να εκφράσουμε την (3.8) σαν

$$
\log m_{ij} = \begin{cases}
\mu & i = 1 \quad ; \quad j = 1 \\
\mu + \alpha_i & i = 2, \ldots, I \quad ; \quad j = 1 \\
\mu + \beta_j & i = 1 \quad ; \quad j = 2, \ldots, J \\
\mu + \alpha_i + \beta_j & i = 2, \ldots, I \quad ; \quad j = 2, \ldots, J
\end{cases}
$$

και έχουμε αντιστοιχία με μοντέλο ANOVA χωρίς αλληλεπιδράσεις κατά δύο παράγοντες για συνεχή δεδομένα.

Σημειώστε ότι έχουμε $(I-1)+(J-1)+1$ παραμέτρους.

4.3 Σχέση μεταξύ μοντέλων λογιστικής παλινδρόμησης και λογαριθμικά γραμμικά μοντέλα.

Έστω ότι έχουμε διωνυμική απόκριση με I επίπεδα αγωγής:

		ΑΠΟΚΡΙΣΗ		
		ΝΑΙ	ΟΧΙ	ΣΥΝΟΛΟ
	1	n_{11}	n_{12}	n_{1+}
ΕΠΙΠΕΔΑ	2	n_{21}	n_{22}	n_{2+}
ΑΓΩΓΗΣ	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	I	n_{I1}	n_{I2}	n_{I+}

Αν υποθέσουμε ότι διεξήχθη ένα διωνυμικό πείραμα για κάθε επίπεδο αγωγής (δηλ. τα n_{i+} είναι τα σταθερά), τότε μπορούμε να αναλύσουμε τα δεδομένα χρησιμοποιώντας μοντέλα λογιστικής παλινδρόμησης. Σε αυτή την οικογένεια των μοντέλων, το μοντέλο ανεξαρτησίας δίνεται από

$$\log\left(\frac{p_i}{1-p_i}\right) = \mu, \quad i = 1, \ldots, I.$$

Το μοντέλο

$$\log\left(\frac{p_i}{1-p_i}\right) = \begin{cases} \mu & i = 1 \\ \mu + \alpha_i & i = 2, \ldots, I. \end{cases}$$

(κορεσμένο μοντέλο) συνεπάγεται ότι η πιθανότητα θετικής απόκρισης είναι διαφορετική σε διαφορετικά επίπεδα αγωγής.

Τώρα ας υποθέσουμε ότι κατά την διάρκεια μιας προκαθορισμένης χρονικής περιόδου συλλέγουμε δεδομένα για επίπεδα αγωγής. Εδώ το σύνολο N_{++} είναι τυχαία μεταβλητή και έχουμε παρατηρήσεις σε $2I$ τυχαίες μεταβλητές $N_{11}, N_{12}, N_{21}, N_{22}, \ldots, N_{I1}, N_{I2}$. Μπορούμε να εξετάσουμε σχέσεις στα δεδομένα χρησιμοποιώντας λογαριθμικά γραμμικά μοντέλα. Θεωρείστε τα επόμενα μοντέλα, με m_{ij} να συμβολίζει τον μέσο του $(i, j)-$στού κελιού:

(1) $M_0 : \log m_{ij} = \mu, \quad i = 1, \ldots, I, j = 1, 2.$

Αυτό το μοντέλο θεωρεί ότι ο αναμενόμενος αριθμός παρατηρήσεων σε κάθε κελλί είναι σταθερός ($= e^{\mu}$).

$$(2) \qquad M_1 : \log m_{ij} = \begin{cases} \mu & i = 1 \;\; ; \;\; j = 1, 2 \\ \mu + \alpha_i & i = 2, \ldots, I \;\; ; \;\; j = 1, 2 \end{cases}$$

Αυτό το μοντέλο είναι κατάλληλο εάν ο αριθμός παρατηρήσεων που αποκρίνονται θετικά και αρνητικά είναι ο ίδιος σε κάθε γραμμή, αλλά διαφέρει για διαφορετικές γραμμές:

$$\begin{array}{cc} e^{\mu} & e^{\mu} \\ e^{\mu+\alpha_2} & e^{\mu+\alpha_2} \\ \vdots & \vdots \\ e^{\mu+\alpha_I} & e^{\mu+\alpha_I} \end{array}$$

Αυτό το μοντέλο διατηρεί το περιθώριο των γραμμών.

$$(3) \qquad M_2 : \log m_{ij} = \begin{cases} \mu & i = 1, \ldots, I \;\; ; \;\; j = 1 \\ \mu + \beta_j & i = 2, \ldots, I \;\; ; \;\; j = 2 \end{cases}$$

Αυτό το μοντέλο συνεπάγεται ότι η αναμενόμενη τιμή για κάθε γραμμή είναι η ίδια στην στήλη 1 και η αναμενόμενη τιμή για κάθε γραμμή είναι η ίδια στην στήλη 2 :

$$\begin{array}{cc} e^{\mu} & e^{\mu+\beta_2} \\ e^{\mu} & e^{\mu+\beta_2} \\ \vdots & \vdots \\ e^{\mu} & e^{\mu+\beta_2} \end{array}$$

Αυτό το μοντέλο διατηρεί το περιθώριο των στηλών.

(4) M_3 : Μοντέλο ανεξαρτησίας

$$\log m_{ij} = \begin{cases} \mu & i = 1 \qquad\quad ; \;\; j = 1 \\ \mu + \alpha_i & i = 2, \ldots, I \;\; ; \;\; j = 1 \\ \mu + \beta_j & i = 1 \qquad\quad ; \;\; j = 2 \\ \mu + \alpha_i + \beta_j & i = 2, \ldots, I \;\; ; \;\; j = 2. \end{cases}$$

Αυτό είναι το μοντέλο στο οποίο καταλήξαμε νωρίτερα:

$$\begin{array}{ll} e^{\mu} & e^{\mu+\beta_2} \\ e^{\mu+\alpha_2} & e^{\mu+\alpha_2+\beta_2} \\ e^{\mu+\alpha_3} & e^{\mu+\alpha_3+\beta_2} \\ \vdots & \vdots \\ e^{\mu+\alpha_I} & e^{\mu+\alpha_I+\beta_2} \end{array}$$

Εδώ διατηρούνται και οι περιθώριες γραμμών και οι περιθώριες στηλών.

(5) M_4: Κορεσμένο μοντέλο

$$\log m_{ij} = \begin{cases} \mu & i = 1 \qquad\quad ; \;\; j = 1 \\ \mu + \alpha_i & i = 2, \ldots, I \;\; ; \;\; j = 1 \\ \mu + \beta_j & i = 1 \qquad\quad ; \;\; j = 2 \\ \mu + \alpha_i + \beta_j + \gamma_{ij} & i = 2, \ldots, I \;\; ; \;\; j = 2. \end{cases}$$

όπου γ_{ij} δηλώνει την αλληλεπίδραση μεταξύ απόκρισης και αγωγής.

4.4 Σχόλια

(1) **Θεώρημα 1:** Αν $N_i \sim p(m_i), i = 1, \ldots, k$ και $\mathbf{N} = [N_1, \ldots, N_k]^T$, $\mathbf{m} = [m_1, \ldots, m_k]^T$ τότε

$$\mathbf{N} \mid N_+ = n_+ \sim M_k(n_+, \frac{1}{m_+}\mathbf{m})$$

(2) **Θεώρημα 2:** Αν $\mathbf{N} \sim M_k(n, \mathbf{p})$ και $\mathbf{N}_1 = [N_1, \ldots, N_r]^T (r < k)$, $\mathbf{N}_2 = [N_{r+1}, \ldots, N_k]^T$, και όμοια ορίσουμε $\mathbf{p}_1, \mathbf{p}_2, \mathbf{n}_1, \mathbf{n}_2, p_{1+}, p_{2+}, n_{1+}, n_{2+}, N_{1+}, N_{2+}$ τότε

(i) $N_{1+} \sim B(n, p_{1+})$

(ii) $\mathbf{N} \mid \left(N_{1+} = n_{1+} \text{και} N_{2+} = n_{2+}\right) \sim M_r(n_{1+}, \frac{\mathbf{p}_1}{p_{1+}}) \times M_{k-r}\left(n_{2+}, \frac{\mathbf{p}_2}{p_{1+}}\right)$

(3) **Δεσμευμένη συμπερασματολογία :** Αν ένας πίνακας $I \times J$ αποκτιέται από δειγματοληψία Poisson αλλά δεν ενδιαφερόμαστε για την τυχαία μεταβλητή N_{++}, τότε η ανάλυση μπορεί να προχωρήσει "δεχόμενοι ότι" N_{++} ήταν γνωστό από την αρχή ότι είναι n_{++}. Δηλαδή, από το Θεώρημα 1, έχουμε ότι:

$$P_r(\mathbf{N} = \mathbf{n} \mid N_{++} = n_{++}) \sim M_{IJ}(n_{++}, \frac{1}{m_{++}}\mathbf{m})$$

Επίσης, εάν δεν ενδιαφερόμαστε για τις περιθώριες των γραμμών $\mathbf{N}_{i+}$, τότε μπορούμε να προχωρήσουμε "δεχόμενοι ότι" $N_{i+} = n_{i+}$ ήταν γνωστό από την αρχή. Συνεπώς, από τη γενίκευση του Θεωρήματος 2,

$$P_r\left(\mathbf{N} = \mathbf{n} \mid N_{1+} = n_{1+}, \ldots, N_{I+} = n_{I+}\right) \quad \sim \quad \prod_{i=1}^{I} M_J\left(n_{i+}, \frac{\mathbf{p}_i}{p_{i+}}\right)$$

δηλαδή έχουμε κατανομή γινομένου πολυωνυμικών.

Τα κίνητρα για τα παραπάνω σχόλια είναι ότι αφενός μία υπόθεση μπορεί να είναι ευκολότερο να ελεγχθεί κάτω από ένα πιο περιορισμένο σχήμα δειγματοληψίας, αφετέρου τα μοντέλα των πιο περιορισμένων σχημάτων ερμηνεύονται ευκολότερα (π.χ. τα "πολυωνυμικά" p_{ij} είναι ευκολότερα από τα "Poisson" m_{+j}).

(4) Πολυωνυμικά δεδομένα με λογαριθμικά γραμμικά μοντέλα. Αν ένας πίνακας συνάφειας προέρχεται από πολυωνυμική δειγματοληψία, μπορεί να αναλυθεί με λογαριθμικά γραμμικά μοντέλα *με την προϋπόθεση* ότι χρησιμοποιούμε μοντέλα που διατηρούν το ολικό σύνολο. Παρόμοιο αποτέλεσμα ισχύει και για το σχήμα δειγματοληψίας του γινομένου των πολυωνυμικών.

5.1 Στατιστικά πακέτα

Συναφείς εντολές για στατιστικά πακέτα που υπάρχουν στο Υπολογιστικό Κέντρο του Ο.Π.Α. είναι

(i) SAS: CATMOD, FREQ, LOGISTIC, MODEL, LOGIST, MLOGIT, MPROBIT.

(ii) $SPSS^x$: LOGLINEAR, DESIGN, HILOGLINEAR, CROSSTABS, PROBIT.

5.2 Ασκήσεις

(1) Έστω Y_i ανεξάρτητες και ταυτοτικά κατανεμημένες τυχαίες μεταβλητές με

$$E(Y_i) = \mu_i \text{και} \quad y_i = \mu_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2).$$

Χρησιμοποιείστε τον έλεγχο πηλίκου πιθανοφανειών να ελέγξετε εάν η υπόθεση $H_0 : \mu_i = \mu$ είναι ικανοποιητική απλούστευση του κορεσμένου μοντέλου. Σχολιάστε το αποτέλεσμα - είναι γνώριμο ;

(2) Έστω ότι έχουμε σχήμα δειγματοληψίας Poisson. Αποδείξτε ότι για το κορεσμένο μοντέλο ισχύει ότι η ΕΜΠ για τα $m_i, i = 1, \ldots, k$ δίνεται από $\hat{m}_i = n_i$.

(3) Έστω ότι το σχήμα δειγματοληψίας είναι γινόμενο πολυωνυμικών. Αποδείξτε ότι η ΕΜΠ του κορεσμένου μοντέλου δίνεται από $\hat{p}_{ij} = \frac{n_{ij}}{n_{i+}}$ και η Deviance είναι η ίδια με αυτή της πολυωνυμικής κατανομής.

(4) Να βρείτε τις ΕΜΠ για το μοντέλο ανεξαρτησίας στην πολυωνυμική δειγματοληψία, σε πίνακα $I \times J$. Έτσι, γράψτε τις προσαρμοσμένες τιμές.

(5) Στο DATASET 3 να διεξάγετε ένα απλό έλεγχο X^2 ανεξαρτησίας χρησιμοποιώντας τις προσαρμοσμένες τιμές της άσκησης 4. Ποιοί είναι οι βαθμοί ελευθερίας ;
Επίσης υπολογίστε την Deviance και διεξάγετε έναν έλεγχο που βασίζεται σ'αυτή την στατιστική συνάρτηση. Συγκρίνατε τα αποτελέσματα στις δύο περιπτώσεις.

(6) Για το διωνυμικό μοντέλο

$$r_i = n_i y_i \sim \text{Διωνυμική}(n_i, p_i), i = 1, \ldots, N$$

αποδείξτε ότι η Deviance δίνεται από

$$D = 2 \sum \left\{ r_i \log \left(\frac{r_i}{n_i \hat{p}_i} \right) + (n_i - r_i) \log \left(\frac{n_i - r_i}{n_i - n_i \hat{p}_i} \right) \right.$$

όπου τα $\hat{p}_i$ είναι οι εκτιμήτριες μεγίστης πιθανοφάνειας που αποκτήθηκαν μεγιστοποιώντας την $p_i = g(Z_{i\beta})$ ως προς β σε κάποιο μη- κορεσμένο γενικευμένο γραμμικό μοντέλο.

5.3 Εργασία.

Για τα DATASET 1, DATASET 2 και DATASET 3 προσαρμόστε κατάλληλα μοντέλα και γράψτε τα αποτελέσματά σας σε δύο μέρη: Στο πρώτο, υποθέτοντας ότι απευθύνεστε στον αντίστοιχο επιστήμονα που διεξήγαγε το πείραμα και δεν ξέρει καθόλου στατιστική. Στο δεύτερο, υποθέτοντας ότι η εργασία σας απευθύνεται στο στατιστικό τμήμα του Οργανισμού που έκανε το πείραμα, όπου οι απαιτήσεις για τεχνικές (Στατιστικές) λεπτομέρειες είναι πολλές.

Μπορείτε να χρησιμοποιήσετε όποιο στατιστικό πακέτο θέλετε. Ημερομηνία παράδοσης: ημέρα των εξετάσεων.