Επαναληπτικό μάθημα GLM

GLM: Πιθανοφάνεια, εκθετική οικογένεια κατανομών (1)

- Ο αριθμός των ατυχημάτων το έτος 2001 για 5 οδηγούς ήταν αντίστοιχα:
  3, 1, 5, 0 και 2.
  - Γράψτε τη likelihood των δεδομένων
  - Υπολογίστε τον εκτιμητή μέγιστης πιθανοφάνειας (MLE) του ρυθμού ατυχημάτων ανά έτος.
  - Εάν θα θέλατε να διερευνήσετε την πιθανή εξάρτηση του αριθμού των ατυχημάτων από άλλες παραμέτρους όπως την ημέρα της εβδομάδας (X1) πως θα γράφατε το αντίστοιχο μοντέλο; Ποιά είναι η συνδετική (link function) συνάρτηση; Ποια είναι η κανονική συνδετική συνάρτηση (Canonical link function); Εξηγείστε

GLM: Πιθανοφάνεια, εκθετική οικογένεια κατανομών (2)

- Η κατανομή είναι Poisson:

$$f(y;\mu) = \Pr(Y = y;\mu) = e^{-\mu}\frac{\mu^{y}}{y!}$$

- Η Likelihood των δεδομένων:

$$L = \prod_{i=1}^{5} f_i(\mu; y_i) = \prod_{i=1}^{5} e^{-\mu}\frac{\mu^{y_i}}{y_i!}$$

GLM: Πιθανοφάνεια, εκθετική οικογένεια κατανομών (3)

- Εκτιμητής μέγιστης πιθανοφάνειας (MLE) του ρυθμού ατυχημάτων ανά έτος
  - Πρέπει να βρεθεί η τιμή του «μ» που μεγιστοποιεί τη Likelihood (L) ή ισοδύναμα τη log Likelihood (l). Θα χρειαστεί να υπολογίσουμε τη log Likelihood και την 1η παράγωγό της ως προς «μ». Η τιμή εκείνη της «μ» (μMLE) που μηδενίζει αυτή την παράγωγο είναι ο εκτιμητής μέγιστης πιθανοφάνειας. (Πρέπει να ελέγξουμε ότι όντως αντιστοιχεί σε μέγιστο.)

GLM: Πιθανοφάνεια, εκθετική οικογένεια κατανομών (4)

$$l = \log(L) = \log(\prod_{i=1}^{5} e^{-\mu} \frac{\mu^{y_i}}{y_i!}) = \sum_{i=1}^{5} \log(e^{-\mu} \frac{\mu^{y_i}}{y_i!}) = \sum_{i=1}^{5} (y_i \log \mu - \mu) + c$$

$$\frac{dl}{d\mu} = \frac{d}{d\mu}[\sum_{i=1}^{5} (y_i \log \mu - \mu) + c] = \sum_{i=1}^{5} (\frac{y_i}{\mu} - 1) + 0 = \sum_{i=1}^{5} \frac{y_i}{\mu} - 5$$

$$\frac{dl}{d\mu} = 0 \Rightarrow \sum_{i=1}^{5} \frac{y_i}{\mu} - 5 = 0 \Rightarrow \mu = \frac{\sum_{i=1}^{5} y_i}{5} \Rightarrow \mu = \frac{3+1+5+0+2}{5} \Rightarrow \mu = 2.2$$

$$\frac{d^2 l}{d\mu^2} = \frac{d(\frac{dl}{d\mu})}{d\mu} = \frac{d(\sum_{i=1}^{5} \frac{y_i}{\mu} - 5)}{d\mu} = \frac{d(\sum_{i=1}^{5} \frac{y_i}{\mu})}{d\mu} - \frac{d(5)}{d\mu}$$

$$= \frac{d(\sum_{i=1}^{5} \frac{y_i}{\mu})}{d\mu} = \frac{d(\frac{1}{\mu} \sum_{i=1}^{5} y_i)}{d\mu} == \frac{(\sum_{i=1}^{5} y_i) d(\frac{1}{\mu})}{d\mu} = -\frac{1}{\mu^2} (\sum_{i=1}^{5} y_i) < 0$$

GLM: Πιθανοφάνεια, εκθετική οικογένεια κατανομών (5)

- Εξάρτηση του αριθμού των ατυχημάτων από ημέρα της εβδομάδας
  - Ορίζουμε 6 ψευδομεταβλητές x1…x6 για τις ημέρες Δευτέρα…Σάββατο αντίστοιχα αφήνοντας την Κυριακή ως κατηγορία αναφοράς. Το Poisson μοντέλο γράφεται:

$$\log(\mu_i) = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + b_4 x_4 + b_5 x_5 + b_6 x_6$$

Link function

Linear predictor

Λογιστική παλινδρόμηση (1)

- Τα δεδομένα προέρχονται από μια μελέτη σε 2159 παιδιά, με κύριο σκοπό τον εντοπισμό διαφορών στα ποσοστά πλήρους (ανάλογα με την ηλικία) εμβολιασμού ανάλογα με την εθνικότητα. Η μελέτη έγινε στη Γουατεμάλα και οι εθνικότητες που συγκρίνονται είναι α) Λατινογενείς (Latin-1) β) Ιθαγενείς που μιλούν Ισπανικά (Ind. Spanish-2) και γ) Ιθαγενείς που δε μιλούν Ισπανικά (Ind. No Spanish-3).

```
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
kid             int    %8.0g                  child id
immun           byte   %8.0g       ynlbl      whether fully immunized
ethn            float  %15.0g      ethlbl     ethnicity
-------------------------------------------------------------------------------
```

Λογιστική παλινδρόμηση (2)

- Με τη βοήθεια του πίνακα που ακολουθεί υπολογίστε τον αριθμό των πλήρως εμβολιασμένων παιδιών που θα αναμένατε για κάθε εθνικότητα αν οι πιθανότητες για πλήρη εμβολιασμό ήταν ίδιες και στις τρεις εθνικότητες.

	whether fully immunized	
ethnicity	No	Yes	Total
Latin	655	628	1283
Ind. Spanish	295	207	502
Ind. no Spanish	245	129	374
Total	1195	964	2159

Λογιστική παλινδρόμηση (3)

- Υπόθεση: Δεν υπάρχει σχέση μεταξύ εθνικότητας και ποσοστού εμβολιασμού (Ηο: μηδενική υπόθεση).

Εφαρμόζουμε απλή μέθοδο των τριών:

Στα 2159 παιδιά 1283 Λατινογενείς

Στα 1195 χωρίς εμβολιασμό Χ; Λατινογενείς

$$X=1283*\frac{1195}{2159}=$$

$$=\frac{\alpha\nu\tau\iota\sigma\tau οιχο\, ορι\zeta.\alpha\theta ροισμα*\alpha\nu\tau\iota\sigma\tau οιχο\, \kappa\alpha\theta.\alpha\theta ροισμα}{\Sigma\upsilon\acute{\nu}ολο}$$

Πίνακας παρατηρηθέντων και αναμενόμενων συχνοτήτων

	Εθνικότητα		
Εμβολιαμός	Λατιν.	Ιθαγ.Ισπ.	Ιθαγ. Οχι Ισπ.
No	655	295	245
	710,137	**277,855**	**207,008**
Yes	628	207	129
	572,863	**224,145**	**166,992**

Λογιστική παλινδρόμηση (4)

- Χρησιμοποιώντας το κατάλληλο τεστ διερευνείστε αν διαφέρουν σε βαθμό στατιστικά οι συχνότητες εμβολιασμού ανά εθνικότητα; (χ2(0.95, 2) ≈ 5.99)

$$X^2 = \sum_{i=1}^{6} \frac{(O-E)^2}{E} = \frac{(655-710{,}137)^2}{710{,}137} + \frac{(295-277{,}855)^2}{277{,}855} + \ldots + \frac{(129-166{,}992)^2}{166{,}992}$$

B.E=(k-1)(l-1)=(3-1)(2-1)=2

X^2=27,573> $X2_{(0.95,2)}$=5,99

Συμπέρασμα: Οι συχνότητες εμβολιασμού διαφέρουν σε βαθμό ισχυρά στατιστικά σημαντικό ανάλογα με την εθνικότητα.

Ερμηνεία: Φαίνεται ότι το ποσοστό εμβολιασμού είναι μεγαλύτερο από το αναμενόμενο στους Λατινογενείς (628/1283*100=48,95%) μικρότερο στους ισπανόφωνους ιθαγενείς (207/502*100=41,24%) και ακόμα μικρότερο στους μη-ισπανόφωνους ιθαγενείς (129/374*100=34,49%). Να σημειωθεί ότι το αναμενόμενο ποσοστό κάτω από τη μηδενική υπόθεση θα ήταν 964/2159*100=44,65%.

Λογιστική παλινδρόμηση (5)

Χρησιμοποιώντας αποτελέσματα από το output που ακολουθεί ποιο τεστ απαντά στο προηγούμενο ερώτημα;. Σχολιάστε το αποτέλεσμά του συγκριτικά με το αποτέλεσμα που τεστ που εφαρμόσατε προηγουμένως. Υπολογίστε τη loglikelihood του αντίστοιχου μηδενικού (null) μοντέλου.

```
. xi:logit immun i.ethn,nolog
Logit estimates                                   Number of obs   =       2159
                                                  LR chi2(2)      =      27.89
                                                  Prob > chi2     =     0.0000
Log likelihood = -1470.1781                       Pseudo R2       =     0.0094
------------------------------------------------------------------------------
       immun |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    _Iethn_2 |  -.3121615   .1064886    -2.93   0.003    -.5208753   -.1034477
    _Iethn_3 |  -.5993507   .1222805    -4.90   0.000     -.839016   -.3596855
       _cons |  -.0420951   .0558487    -0.75   0.451    -.1515565    .0673663
------------------------------------------------------------------------------
. mat li e(V)
symmetric e(V)[3,3]
            _Iethn_2    _Iethn_3       _cons
_Iethn_2   .01133982
_Iethn_3   .00311907   .01495251
   _cons  -.00311907  -.00311907   .00311907
```

Λογιστική παλινδρόμηση (6)

Στο output το τεστ που απαντά στο προηγούμενο ερώτημα είναι το **LR Chi(2)**, δηλαδή το τεστ μέγιστης πιθανοφάνειας (επίσης στους δύο Β.Ε.).

Η ελάχιστη διαφορά στη τιμή τους προέρχεται από το ότι τα δύο τεστ (**Pearson X2** test και **Likelihood Ratio** test) είναι δυο διαφορετικές προσεγγίσεις (approximations).

LRTest=-2(*l1*-*l2*), όπου *l1* η πιθανοφάνεια του null (μηδενικού) μοντέλου και *l2* η πιθανοφάνεια του τρέχοντος μοντέλου.

Άρα έχουμε:

28.89=-2*[*l1*-(-1470,1781)]

$$\Rightarrow 1470,1781 + l_1 = -13,9450 \Rightarrow$$

$$\Rightarrow l_1 = -1470,1781 - 13,9450 \Rightarrow$$

$$\Rightarrow l_1 = -1484,1231$$

Λογιστική παλινδρόμηση (7)

- Ερμηνεύστε τα αποτελέσματα του παραπάνω μοντέλου (ερμηνεία όλων των συντελεστών, δώστε τα κατάλληλα 95% CI).
  - Το ποσοστό εμβολιασμού διαφέρει σε βαθμό στατιστικά σημαντικό ανάμεσα στις διαφορετικές εθνικότητες (και οι δύο συντελεστές είναι διαφορετικοί του μηδενός σύμφωνα με το Wald τεστ, ενώ το μοντέλο έχει συνολικά σημαντική εφαρμογή).
  - Ιθαγενείς που μιλούν Ισπανικά έχουν 27% ($1-e^{-0,3122}$ =1-0,7319=0,2681) μικρότερη πιθανότητα εμβολιασμού συγκριτικά με τους Λατινογενείς με 95% όρια αξιοπιστίας 9,8% μέχρι 40,6% ($1-e^{-0,1034477}$, $1-e^{-0,5208753}$).
  - Οι Ιθαγενείς που δεν μιλούν Ισπανικά έχουν περίπου τη μισή πιθανότητα ($e^{-0,5993507}$=0,5491681) να εμβολιαστούν συγκριτικά με τους Λατινογενείς (95% Ο.Α.: $e^{-0,839016}$=0,4321355 μέχρι $e^{-0,3596855}$=0,6978958).
  - **Σημείωση**: Τα συμπεράσματα είναι αντίστοιχα με αυτά που είχαμε εφαρμόζοντας το Χ2 ετερογένειας.

Λογιστική παλινδρόμηση (8)

- Εκφράστε το παραπάνω μοντέλο στη μαθηματική του μορφή

$$\ln(odds) = \ln(\frac{\pi}{1-\pi}) = a + b_1(I_{\text{Ι}\theta\alpha\gamma.-\text{Ι}\sigma\pi\alpha\nu o\phi.}) + b_2(I_{\text{Ι}\theta\alpha\gamma.-o\chi\iota\ \text{Ι}\sigma\pi\alpha\nu o\phi.})$$

link function: ln(odds) π: πιθανότητα εμβολιασμού

Λογιστική παλινδρόμηση (9)

- Υπολογίστε τα Odds πλήρους εμβολιασμού και τα αντίστοιχα 95% CI's και για τις τρεις εθνικότητες.

$$\ln(odds) = \ln(\frac{\pi}{1-\pi}) = a = -0,0420951$$

$$\frac{\pi}{1-\pi} = e^{a} = e^{-0,0420951} = 0,9587786$$

$$95\%\ C.I.: e^{-0,1515565} - e^{0,0673663} = 0,8594 - 1,0697$$

Λογιστική παλινδρόμηση (10)

Ιθαγενείς που μιλάνε Ισπανικά

$$\ln(odds) = \ln(\frac{\pi}{1-\pi}) = a + b_1(I_{\text{Ιθαγ.-Ισπανοφ.}}) = -0,0420951 - 0,3121615 = -0,3532566$$

$$Var(a+b_1) = Var(a) + \text{var}(b_1) + 2*\text{cov}(a,b_1) =$$

$$= 0,00311907 + 0,01133982 - 2*0,00311907 = 0,0082$$

$$SE(a+b_1) = \sqrt{0,0082} = 0,0907$$

$$95\%CI : (a+b_1) \pm \{1,96*SE(a+b_1)\} = -0,3542566 \pm (1,96*0,0907) =$$

$$= -0,53196 \; to - 0,1765499$$

$$\frac{\pi}{1-\pi} = e^{a+b_1} = e^{-0,0420951-0,3121615} = 0,70169489$$

$$95\;CI : e^{-0,5319632} - e^{-0,1765499} = 0,58745055 - 0,83815695$$

Λογιστική παλινδρόμηση (11)

- Σχολιάστε τα αποτελέσματα σε σχέση με τα τρία p-values που σχετίζονται με τους εκτιμηθέντες συντελεστές του μοντέλου. Ποιες είναι οι αντίστοιχες μηδενικές υποθέσεις;

•Η σταθερά του μοντέλου ΔΕΝ ειναι στατιστικά σημαντική. Κατά αντιστοιχία τα 95% Ο.Α. του αντίστοιχου Odds περιλαμβάνουν την μονάδα (ή του συντελεστή περιλαμβάνουν το 0). Αντίθετα, οι δύο άλλοι συντελεστές είναι στατιστικά σημαντικοί και κατ'αντιστοιχία τα 95% Ο.Α. των Odds δεν περιλαμβάνουν την μονάδα.

•Οι αντίστοιχες μηδενικές υποθέσεις είναι:

$$\mathrm{H}_0^1 : a = 0$$

$$\mathrm{H}_0^2 : b_1 = 0$$

$$\mathrm{H}_0^3 : b_2 = 0$$

Λογιστική παλινδρόμηση (12)

- Από τι είδους έλεγχο έχουν προκύψει αυτά τα p-values και πως υπολογίζονται τα αντίστοιχα z-κριτήρια;
  - P-values / Tests
  - Γενική του μορφή το Wald test

$$z = \frac{b_i}{SE(b_i)}$$

$$W = (\hat{\theta} - \theta_0)' \hat{I}(\theta)(\hat{\theta} - \theta_0) \sim X^2_n$$

$$\text{ό}\pi\text{ου}: \hat{\theta}\ \text{το διάνυσμα των εκτιμητών}$$

$$\hat{\text{I}}(\theta)\ \text{ο πίνακας πληροφορίας}\ (observed\ \text{inf}.matrix)$$

(II)

```
. xi:logit case i.female*iron,nolog
i.female          _Ifemale_0-1        (naturally coded; _Ifemale_0 omitted)
i.female*iron     _IfemXiron_#        (coded as above)
Logistic regression                               Number of obs   =        908
                                                  LR chi2(3)      =      99.60
                                                  Prob > chi2     =     0.0000
Log likelihood = -549.60912                       Pseudo R2       =     0.0831
------------------------------------------------------------------------------
        case |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  _Ifemale_1 |  -2.587112   .5391018    -4.80   0.000    -3.643733   -1.530492
        iron |  -.0318812   .6897189    -0.05   0.963    -1.383705    1.319943
_IfemXiron_1 |   3.624367   1.739636     2.08   0.037     .2147439     7.03399
       _cons |  -.0902706   .2369655    -0.38   0.703    -.5547145    .3741733
. mat li e(V)
symmetric e(V)[4,4]
                        case:         case:         case:         case:
                   _Ifemale_1          iron  _IfemXiron_1         _cons
  case:_Ifemale_1     .29063074
        case:iron     .15369011     .47571212
case:_IfemXiron_1    -.88154339    -.47571212      3.026332
       case:_cons    -.05615266    -.15369011     .15369011     .05615266
```

Ερμηνεύστε τα αποτελέσματα του μοντέλου (II) (Ποιοι παράγοντες επηρεάζουν την εμφάνιση στεφανιαίας νόσου και πώς;). Χρησιμοποιώντας τα αποτελέσματα του προηγούμενου output [μοντέλο (II)] υπολογίστε το Odds Ratio για εμφάνιση στεφανιαίας νόσου και το αντίστοιχο 95% διάστημα εμπιστοσύνης που σχετίζεται με αύξηση στην ημερήσια πρόσληψη σιδήρου κατά 1 γραμμάριο, στις γυναίκες. (**5**)

Προσέξτε ότι έχει προηγηθεί μετατροπή των τιμών της μεταβλητής iron έτσι ώστε η μονάδα μέτρησης να είναι τώρα το γραμμάριο (g) και όχι το milligram (mg).

$$\text{Log[Odds]}=\beta_0+\beta_1*\text{Fem}+\beta_2*\text{Iron}+\beta_3*\text{Fem}*\text{Iron}$$

Log[Odds]=	Iron=i	Iron=i+1
Male (Fem=0)	$\beta_0+\beta_1*\text{Fem}+\beta_2*\text{Iron}+\beta_3*\text{Fem}*\text{Iron}=$ **$\beta_0+\beta_2*i$**	$\beta_0+\beta_1*\text{Fem}+\beta_2*\text{Iron}+\beta_3*\text{Fem}*\text{Iron}=\beta_0+\beta_2*(i+1)=$ **$\beta_0+\beta_2*i+\beta_2$**
Female (Fem=1)	$\beta_0+\beta_1*\text{Fem}+\beta_2*\text{Iron}+\beta_3*\text{Fem}*\text{Iron}=\beta_0+\beta_1+\beta_2*i+\beta_3*i=$ **$\beta_0+\beta_1+(\beta_2+\beta_3)*i$**	$\beta_0+\beta_1*\text{Fem}+\beta_2*\text{Iron}+\beta_3*\text{Fem}*\text{Iron}=\beta_0+\beta_1+\beta_2*(i+1)+\beta_3*(i+1)=\beta_0+\beta_1+(\beta_2+\beta_3)*(i+1)=$ **$\beta_0+\beta_1+(\beta_2+\beta_3)*i+(\beta_2+\beta_3)$**

Παλινδρόμηση Poisson (1)

Τα πιο κάτω δεδομένα αφορούν νέες περιπτώσεις μελανώματος στις ΗΠΑ κατά το χρονικό διάστημα 1969-1991, μεταξύ λευκών ανδρών ανά ηλικιακή ομάδα και περιοχή. Η στήλη N δείχνει τον αντίστοιχο πληθυσμό.

Region	Age	Cases	N
North	0–35	61	2880262
	35–44	76	564535
	45–54	98	592983
	55–64	104	450740
	65–74	63	270908
	75+	80	161850
South	0–35	64	1074246
	35–44	75	220407
	45–54	68	198119
	55–64	63	134084
	65–74	45	70708
	75+	27	34233

Παλινδρόμηση Poisson (2)

- Γράψτε με μαθηματική μορφή ένα κατάλληλο μοντέλο που να λαμβάνει υπόψη του όλα τα δεδομένα του πιο πάνω πίνακα

$$\ln(\mu) = \ln(N) + \beta_o + \sum_{i=1}^{5} \beta_i I_{age_i} + \beta_6 I_{South}$$

$$\ln(rate) = \ln(\frac{\mu}{N}) = \ln(\mu) - \ln(N) = \beta_0 + \sum_{i=1}^{5} \beta_i I_{age_i} + \beta_6 I_{South}$$

$$\ln(\mu) : Link\ Function\ (Kanonical\ \gamma\iota\alpha\ Poisson)$$

$$\ln(N) : Offset$$

$$\beta_0 + \sum_{i=2}^{6} \beta_i I_{age_i} + \beta_1 I_{South} = n : Linear\ \Pr edictor$$

Παλινδρόμηση Poisson (3)

- Ερμηνεύστε τους συντελεστές “_Iage2”, “_Iregion_2” και “_cons” που φαίνονται στο output που ακολουθεί. Χρησιμοποιείστε κατάλληλους μετασχηματισμούς ώστε να δώσετε μια κατανοητή ερμηνεία στους παραπάνω αναφερθέντες συντελεστές.

```
. xi:poisson cases i.age i.reg,exposure(n) nolog
i.age             _Iage_1-6           (_Iage_1 for age==0-35 omitted)
i.region          _Iregion_1-2        (_Iregion_1 for region==North omitted)
Poisson regression                               Number of obs   =         12
                                                 LR chi2(6)      =     889.60
                                                 Prob > chi2     =     0.0000
Log likelihood = -39.219909                      Pseudo R2       =     0.9190
------------------------------------------------------------------------------
       cases |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     _Iage_2 |   1.797375   .1209263    14.86   0.000     1.560364    2.034386
     _Iage_3 |   1.913088   .1184391    16.15   0.000     1.680951    2.145224
     _Iage_4 |   2.241802   .1183364    18.94   0.000     2.009867    2.473738
     _Iage_5 |   2.365724   .1315178    17.99   0.000     2.107954    2.623494
     _Iage_6 |   2.944679   .1320473    22.30   0.000     2.685871    3.203487
  _Iregion_2 |   .8194846   .0710279    11.54   0.000     .6802724    .9586968
       _cons |  -10.65831   .0951846  -111.98   0.000    -10.84487   -10.47175
           n | (exposure)
```

Παλινδρόμηση Poisson (4)

- Για άτομα ηλικίας 0-34 ετών που μένουν στις βόρειες (North) περιοχές.

$$\ln(\frac{\mu}{N}) = -10.65831$$

Ο μέσος αριθμός νέων περιπτώσεων ανά άτομο για την περίοδο 1969-1991 (22 έτη) σε άτομα ηλικίας 0-35 ετών και που διαμένουν στις βόρειες περιοχές είναι
$e^{-10.66}=2.35*10^{-5}$, ή ο μέσος αριθμός νέων περιπτώσεων ανά 1.000.000 ανθρωποέτη : $2.35*10^{-5}*1.000.000/22=0.96$

Παλινδρόμηση Poisson (5)

- Ελέγχοντας για την περιοχή, ο σχετικός κίνδυνος ανάπτυξης μελανώματος σε άτομα ηλικίας 35-44 συγκριτικά με άτομα ηλικίας 0-35 ετών είναι: 6.04 ($=e^{1.797375}$).
- Ελέγχοντας για διαφορές στην ηλικία, τα άτομα που διαμένουν στις Νότιες περιοχές έχουν υπερδιπλάσιο ($=e^{0.8194846}=2,2693$) κίνδυνο ανάπτυξης μελανώματος συγκριτικά με αυτά που διαμένουν σε βόρειες περιοχές.

Παλινδρόμηση Poisson (6)

- Θεωρώντας αποδεκτό τον μετασχηματισμό των ηλικιακών ομάδων (age) σε συνεχή μεταβλητή (“contage”) όπως φαίνεται παρακάτω, εφαρμόστηκαν 2 μοντέλα (Μοντέλο 2 και Μοντέλο 3). Στο πρώτο από αυτά εισάγεται η ηλικία ως συνεχής μεταβλητή (Μοντέλο 2) ενώ στο δεύτερο εισάγεται επιπλέον και το τετράγωνο της ηλικίας (“cont2”, Μοντέλο 3). Συγκρίνεται τα 3 μοντέλα (κατηγορική ηλικία, συνεχής- γραμμική, συνεχής-παραβολοειδής). Ποιο θα επιλέγατε και γιατί;
- (Λάβετε υπόψη σας τα output και το γράφημα που ακολουθεί. Δίνονται:
- $\chi2(0.95, 4) \approx 9.45$, $\chi2(0.95, 1) \approx 3.84$ $\chi2(0.95, 3) \approx 7.81$).

Παλινδρόμηση Poisson (7)

```
. li age contage in 1/6
          age    contage
  1.      0-35         25
  2.     35-44         40
  3.     45-54         50
  4.     55-64         60
  5.     65-74         70
  6.       75+         80
```

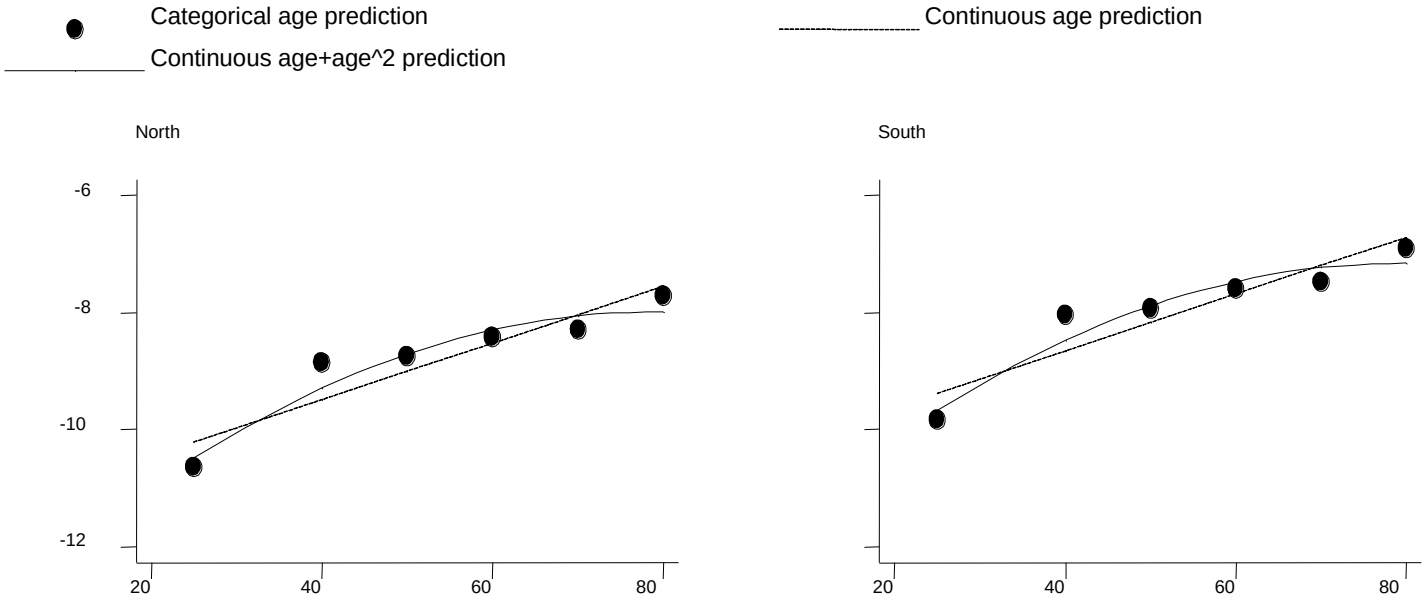
Μοντέλο 2

```
. xi:poisson cases contage i.reg,exp(n) nolog
Log likelihood = -89.222205                     Pseudo R2       =     0.8157
------------------------------------------------------------------------------
       cases |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     contage |   .0484561   .0018081    26.80   0.000     .0449122        .052
  _Iregion_2 |   .8413212   .0710087    11.85   0.000     .7021467    .9804956
       _cons |  -11.43071   .1084527  -105.40   0.000    -11.64328   -11.21815
           n | (exposure)
------------------------------------------------------------------------------
```

Μοντέλο 3

```
.  xi:poisson  cases contage cont2 i.reg,exp(n) nolog
Log likelihood = -61.626313                     Pseudo R2       =     0.8727
------------------------------------------------------------------------------
       cases |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     contage |   .1349716    .012069    11.18   0.000     .1113167    .1586265
       cont2 |  -.0008507   .0001173    -7.25   0.000    -.0010806   -.0006209
  _Iregion_2 |   .8248273   .0709744    11.62   0.000       .68572    .9639346
       _cons |  -13.33605   .2913078   -45.78   0.000      -13.907    -12.7651
           n | (exposure)
------------------------------------------------------------------------------
```

Παλινδρόμηση Poisson (8)



Παλινδρόμηση Poisson (9)

Θεωρώντας το γραμμικό και το παραβολοειδές μοντέλο εμφωλιασμένο (nested) στο μοντέλο με κατηγορική την ηλικία, τα τρία μοντέλα μπορούν να συγκριθούν με το έλεγχο πηλίκου μέγιστης πιθανοφάνειας.

Κατηγ. προς γραμμικό:

LRT=-2(log likelihoodmodel2-log likelihoodmodel1=

-2{-89,222205-(-39,219909)}=

-2*(-50,0023)=100,0046

Κάτω από την μηδενική υπόθεση (μη-διαφοράς των μοντέλων) το LRT ακολουθεί την X2 κατανομή με β.ε. τη διαφορά των παραμέτρων στα δύο μοντέλα.

Άρα:

LRT=100,0046>>9.45=X2(0,95,4)

Επομένως το μοντέλο με κατηγορική την ηλικία έχει σημαντικά καλύτερη εφαρμογή από το μοντέλο με γραμμική την ηλικία.

Παλινδρόμηση Poisson (10)

Μοντέλο 3/ μοντέλο 2:

LRT=55.1918>>3.84=X2(0,95,1)

Άρα το παραβολοειδές μοντέλο εφαρμόζει καλύτερα του γραμμικού.

Μοντέλο 3/ μοντέλο 1:

LRT=44.8128>>7.81=X2(0,95,3)

Άρα το μοντέλο με κατηγορική την ηλικία εφαρμόζει καλύτερα από τα άλλα δύο μοντέλα, και παρά τον μεγαλύτερο αριθμό παραμέτρων πρέπει να προτιμηθεί.

Παλινδρόμηση Poisson (11)

- Λαμβάνοντας υπόψην τα output βλέπουμε ότι το τετράγωνο της ηλικίας συνεισφέρει σημαντικά (και σύμφωνα με το Wald test). Επιπλέον σύμφωνα με το γράφημα, αν και το μοντέλο 3 (παραβολοειδής) πλησιάζει καλύτερα τα αποτελέσματα από το κατηγορικό μοντέλο, αδυνατεί να περιγράψει επαρκώς τις κατηγορίες τα δεδομένα, κυρίως στη 2η και 4η ηλικιακή ομάδα.

Παλινδρόμηση Poisson (12)

- Ανεξαρτήτως από την ορθότητα του γραμμικού μοντέλου (Μοντέλο 2) πώς ερμηνεύονται στο μοντέλο αυτό οι συντελεστές “_cons”, “contage” και “_Iregion_2”; Αν παρατηρήσετε τα διαστήματα εμπιστοσύνης στα δύο μοντέλα (με την ηλικία ως κατηγορική και με την ηλικία ως συνεχή μεταβλητή) θα δείτε ότι τα αποτελέσματα σχετικά με τις διαφορές ανά περιοχή είναι συμβατά μεταξύ τους, ενώ οι σταθερές (“_cons”) διαφέρουν σημαντικά στα δύο μοντέλα. Πού νομίζετε ότι οφείλεται αυτό;

Παλινδρόμηση Poisson (13)

- **Σταθερά**:
- Ο μέσος αριθμός νέων περιπτώσεων ανά άτομο για την περίοδο 1969-1991 σε άτομα ηλικίας 0 ετών και που διαμένουν στις βόρειες περιοχές είναι $e^{-11,43071}$=1,086*10-5. Παρατηρείστε ότι το μοντέλο αυτό κάνει προβολή σε ηλικία 0 ετών. Ο συντελεστής της σταθεράς είναι στο γράφημα για τις Βόρειες περιοχές το σημείο που η ευθεία τέμνει τον άξονα y
- **Contage**: Ελέγχοντας για την περιοχή, ο σχετικός κίνδυνος ανάπτυξης μελανώματος ανά έτος αύξησης της ηλικίας είναι: 1.0496 (=$e^{0,0484561}$). Ή αύξηση της ηλικίας κατά μία δεκαετία συνδέεται με άυξηση του κινδύνου ανάπτυξης μελανώματος κατά 62% (=$e^{0,0484561*10}$-1).
- Τα διαστήματα εμπιστοσύνης για την περιοχή είναι αντίστοιχα γιατί εκφράζουν αντίστοιχα πράγματα στα δύο μοντέλα. Αντίθετα, τα 95% Ο.Α. για τη σταθερά του μοντέλου 1 δεν περιέχουν την εκτίμηση της σταθεράς του μοντέλου 2. Αυτό συμβαίνει γιατί στο μοντέλο 1 η σταθερά αναφέρεται σε ηλικία 0-35 ετών, ενώ στο μοντέλο 2 σε ηλικία 0 ετών (προβολή στον άξονα y: συμβουλευτείτε επίσης το γράφημα).

Παλινδρόμηση Poisson (14)

- Στο πιο πάνω γράφημα οι δύο γραμμές που αφορούν προβλέψεις του μοντέλου με την συνεχή μεταβλητή για την ηλικία στις δύο περιοχές είναι παράλληλες. Το γεγονός αυτό δίνει κάποια πληροφορία για την ύπαρξη αλληλεπίδρασης μεταξύ περιοχής και ηλικίας; Δικαιολογήστε
  - Σε έλλειψη αλληλεπίδρασης, οι δύο γραμμές θα ήταν παράλληλες. Όμως εδώ στο μοντέλο 2 ΔΕΝ έχουμε προβλέψει για αλληλεπίδραση, άρα ΥΠΟΘΕΤΟΥΜΕ παραλληλία, και επομένως το γράφημα αντανακλά την προϋπόθεση του μοντέλου και ΔΕΝ δινει κάποια πληροφορία για την ύπαρξη αλληλεπίδρασης.