Γενικευμένα Γραμμικά Μοντέλα (GLM)

Επισκόπηση

Γενική μορφή

$$g(E[\mathbf{Y} \mid X]) = X\mathbf{b}$$

- Κατανομή της Y στην εκθετική οικογένεια
- Ανεξάρτητες παρατηρήσεις
- ‘Ενας όρος για το σφάλμα

$g(.)$ Συνδετική συνάρτηση (link function)

π.χ. Normal regression g(y)=y,

Poisson regression g(y)=log(y)

Εκτίμηση

- Maximum Likelihood Methods
  - Εύρεση $\hat{\mathbf{b}}$ τέτοιου ώστε να μεγιστοποιείται η πιθανοφάνεια του μοντέλου $L(\mathbf{b} \mid X)$
  - 100(1-a)% Διαστήματα εμπιστοσύνης για τα $\mathbf{b}_i$:

$$\hat{b}_i \pm z_{1-a/2} SE(\hat{b}_i)$$

Standard Errors για τις εκτιμήσεις των b_i : τετραγωνική ρίζα της αντίστοιχης Variance [από πίνακα Variance-Covariance των εκτιμήσεων των b_i – δίνει και Covariances π.χ. Cov(b_1,b_2)]

Έλεγχος υποθέσεων

- Likelihood ratio test

  $2[\text{loglikelihood}(M_{p+k}) - \text{loglikelihood}(M_p)] \sim X^2_k$

- Wald test $\hat{\mathbf{b}}' V(\hat{\mathbf{b}})^{-1} \hat{\mathbf{b}} \sim X^2_k$

  k: διάσταση του **b**, για k=1

$$\frac{\hat{b}^2}{Var(\hat{b})} \sim X^2_1 \Leftrightarrow \frac{\hat{b}}{SE(\hat{b})} \sim N(0,1)$$

Γραμμικοί συνδυασμοί των b_i

- Για να κατασκευάσουμε διαστήματα εμπιστοσύνης για γραμμικούς συνδυασμούς (π.χ. b_1+b_2, b_2-b_1) ή για αντίστοιχο έλεγχο υποθέσεων χρειαζόμαστε τη Variance του γραμμικού συνδυασμού
  - $Var(b_1+b_2)=Var(b_1)+Var(b_2)+2Cov(b_1,b_2)$
  - $Var(b_1-b_2)=Var(b_1)+Var(b_2)-2Cov(b_1,b_2)$
- Γενικά

$$Var(c'\hat{\mathbf{b}}) = c'Var(\hat{\mathbf{b}})c$$

Έλεγχος γραμμικότητας (1)

- Έλεγχος για σημαντικότητα πολυονυμικών όρων π.χ.

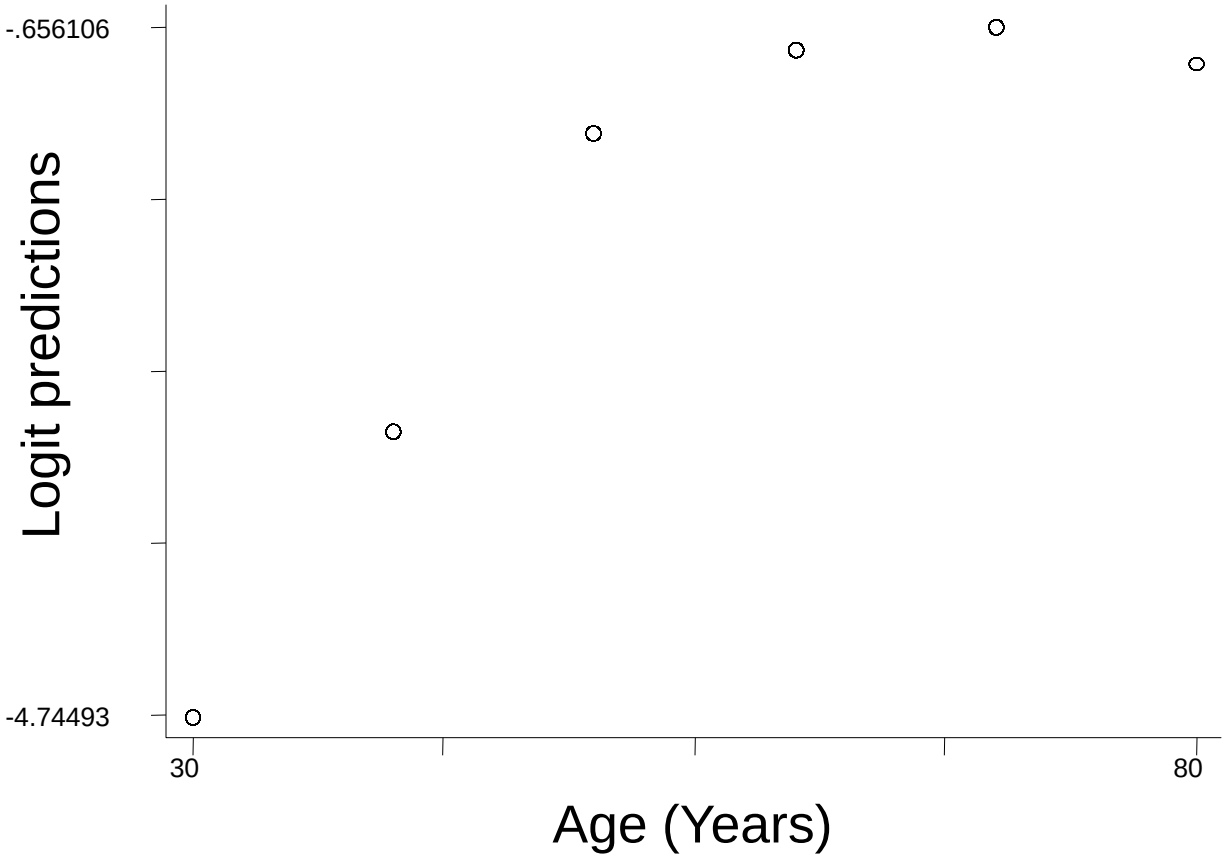
```
. xi:logit  outcome contage
Log likelihood = -451.09778
------------------------------------------------------------------------------
     outcome |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     contage |   .0574208   .0066054     8.69   0.000     .0444745    .0703671
       _cons |  -4.558457   .3956411   -11.52   0.000    -5.333899   -3.783015
------------------------------------------------------------------------------
```

```
. xi:logit  outcome contage contage2
Log likelihood = -435.34809
------------------------------------------------------------------------------
     outcome |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     contage |   .4018381   .0706098     5.69   0.000     .2634455    .5402307
    contage2 |  -.0029861   .0005969    -5.00   0.000    -.0041559   -.0018163
       _cons |   -14.0894    2.05425    -6.86   0.000    -18.11565   -10.06314
------------------------------------------------------------------------------
```

Έλεγχος γραμμικότητας (2)

- Γραφικός έλεγχος (logit predictions από κατηγορικό μοντέλο)

```
qui xi:logit  outcome i.age
predict logcat,xb
gr7 logcat contage
```



Logistic regression (1)

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = b_0 + \sum_{j=1}^{p} b_j x_{ij} \ , \quad i = 1, 2, ..., n$$

$$b_0 : \log(Odds) \, \gamma\iota\alpha \, x_j = 0, \forall \ j$$

$$b_j : \log(Odds\ Ratio)$$

- Για αύξηση του x_j κατά 1 μονάδα
- Αν το x_j είναι ψευδομεταβλητή για συγκεκριμένη κατηγορία κ κατηγορικής μεταβλητής το Odds Ratio αντιστοιχεί σε:
Κατηγορία κ vs. Κατηγορία αναφοράς

Logistic regression (2)

Ερμηνεία όρων αλληλεπίδρασης:

Έστω **$\log[\pi/1-\pi]=\beta_0+\beta_1$*Γυναίκες+$\beta_2$*Ηλικία+$\beta_3$*(Ηλικία*Γυναίκες)**

Άνδρες: **$\log[\pi/1-\pi]=\beta_0+\beta_2$*Ηλικία**

Γυναίκες: **$\log[\pi/1-\pi]=\beta_0+\beta_1$*Γυναίκες+($\beta_2+\beta_3$)*Ηλικία**

Odds Ratio για αύξηση ηλικίας κατά 1 χρόνο

Άνδρες: **$\exp(\beta_2)$**

Γυναίκες: **$\exp(\beta_2+\beta_3)$**

$\exp(\beta_3)$: Πόσο μεγαλύτερο (ή μικρότερο) είναι το OR που σχετίζεται με αύξηση της ηλικίας κατά ένα χρόνο στις γυναίκες σε σχέση με το αντίστοιχο των αντρών

The model

$E[\log\{\pi/(1-\pi)\}] = \beta_0+\beta_1*\texttt{Age\_2}+\beta_2*\texttt{Age\_3}+\beta_3*\texttt{Age\_4}+\beta_4*\texttt{More\_1}$
$+\beta_5*\texttt{Age\_2*More\_1}+\beta_6*\texttt{Age\_3*More\_1}+\beta_7*\texttt{Age\_4*More\_1}$

	Age			
More Children	<25 (Age_1)	25-29 (Age_2)	30-39 (Age_3)	40-49 (Age_4)
Yes (More_1)	$\exp(\beta_0+\beta_4)$	$\exp(\beta_0+\beta_1+\beta_4+\beta_5)$	$\exp(\beta_0+\beta_2+\beta_4+\beta_6)$	$\exp(\beta_0+\beta_3+\beta_4+\beta_7)$
No (More_0)	$\exp(\beta_0)$	$\exp(\beta_0+\beta_1)$	$\exp(\beta_0+\beta_2)$	$\exp(\beta_0+\beta_3)$

OR_1

Odds Ratio for contraceptive use (More=Yes vs. More=No | Age=25-29)= $\exp(\beta_4+\beta_5)$

OR_2

Odds Ratio for contraceptive use (More=Yes vs. More=No | Age=<25)= $\exp(\beta_4)$

$OR_1 / OR_2 = \exp(\beta_5)$

OR_1

Odds Ratio for contraceptive use (Age=25-29 vs. Age<25 | More=Yes)= $\exp(\beta_1+\beta_5)$

OR_2

Odds Ratio for contraceptive use (Age=25-29 vs. Age<25 | More=No)= $\exp(\beta_1)$

$OR_1 / OR_2 = \exp(\beta_5)$

Logistic regression (3)

Κυριώτερα διαγνωστικά

Overall goodness of fit: **Pearson ή Deviance** χ^2 . Απαιτεί μεγάλο αριθμό παρατηρήσεων ανά covariate pattern. Αν αυτή η πρϋπόθεση δεν ικανοποιείται: **Hosmer-Lemeshow goodness of fit**

Residuals: Pearson, Deviance

Γραφήματα: ΔX^2 ή ΔD vs. estimated probabilities

Influence: $\Delta\beta$

Conditional Logistic regression

- Ανάλυση δεδομένων από matched case-control studies (με οποιοδήποτε matching σχήμα)
- Ερμηνεία συντελεστών όπως στην απλή logistic regression
- ΠΡΟΣΟΧΗ: Το output του Stata δεν δίνει εκτίμηση του β_0, το οποίο δεν έχει νόημα να ερμηνευτεί
- Δεν μελετάμε την επίδραση μεταβλητών που έχουν χρησιμοποιηθεί για matching ως main effects. Αντίθετα **μπορούμε** να μελετήσουμε αλληλεπιδράσεις matching μεταβλητών με άλλες μεταβλητές

Multinomial Logistic regression

Ανάλυση δεδομένων όπου η εξαρτημένη μεταβλητή είναι κατηγορική (μη διατεταγμένη). Για Y=0, 1, 2 …k και Y=0 η κατηγορία αναφοράς:

$$\log\left[\frac{P(Y=j|\mathbf{x})}{P(Y=0|\mathbf{x})}\right]=\beta_{j0}+\beta_{j1}x_1+\cdots+\beta_{jp}x_p \quad \text{για } j=1,2,\ldots k$$

$$\pi_j(\mathbf{x})=P(Y=j\,|\,\mathbf{x})=\frac{e^{g_j(\mathbf{x})}}{\sum_{j=0}^{k} e^{g_j(\mathbf{x})}} \qquad g_o(\mathbf{x})=0$$

- Οι συντελεστές ερμηνεύονται ως log(Relative Risk Ratios)
- Διαγνωστικά όπως στην απλή logistic regressions για τις επιμέρους συγκρίσεις (Y=1 vs. Y=0, …, Y=k vs. Y=0/Begg & Gray, Biometrika, 1984)

Ordinal Logistic regression

$$\log\{\gamma_j(x)/(1-\gamma_j(x))\} = \kappa_j - \beta^T x \quad , j = 1,...,I-1 \qquad I = \#\,of\ categories \tag{1}$$

where $\gamma_j = pr(Y \le j|\mathrm{x})$ cumulative probability up to and including category j

Proportional-odds model: Ratio of the odds of the event $Y \le j$ at $x=x_1$ and $x=x_2$ is:

$$\frac{\gamma_j(x_1)/(1-\gamma_j(x_1))}{\gamma_j(x_2)/(1-\gamma_j(x_2))} = \exp\{-\beta^T(x_1-x_2)\}$$

FEMALES

$P_1=1/[1+\exp(\beta_1-k_1)]$

$P_1+P_2=1/[1+\exp(\beta_1-k_2)]$

$P_1+P_2+P_3=1/[1+\exp(\beta_1-k_3)]$

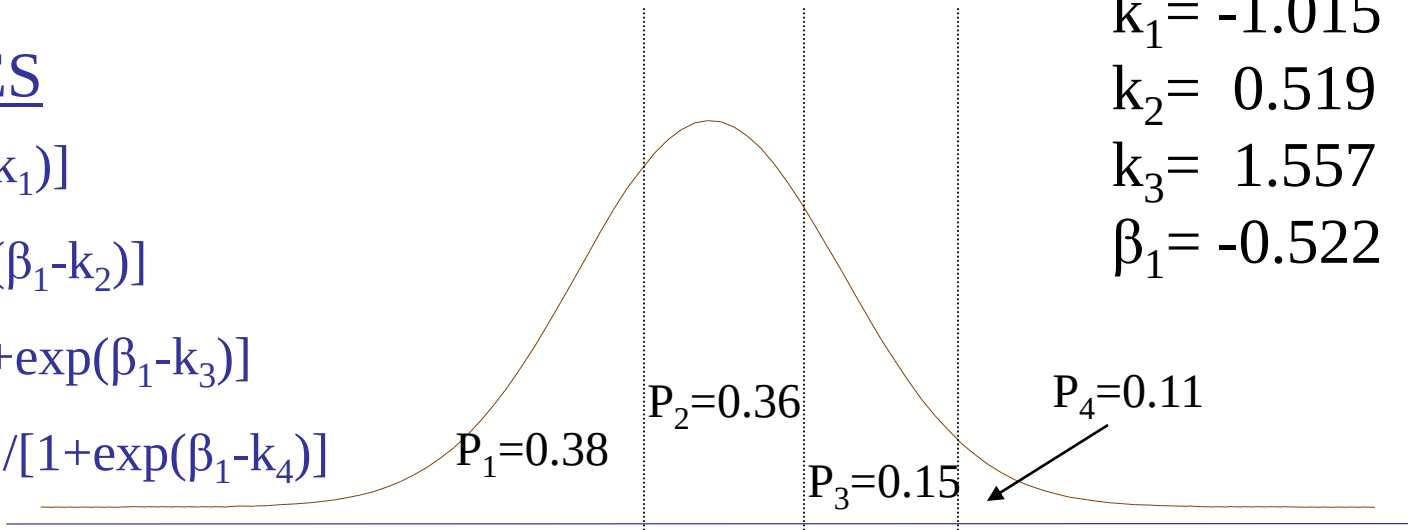
$P_1+P_2+P_3+P_4=1/[1+\exp(\beta_1-k_4)]$

$k_1= -1.015$

$k_2= 0.519$

$k_3= 1.557$

$\beta_1= -0.522$



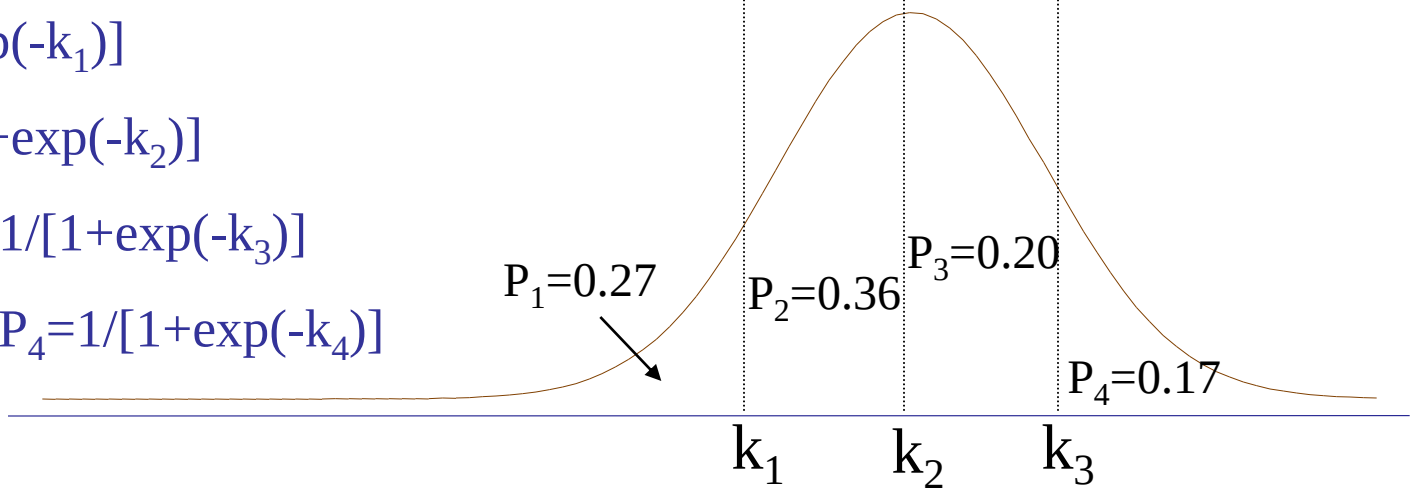
MALES

$P_1=1/[1+\exp(-k_1)]$

$P_1+P_2=1/[1+\exp(-k_2)]$

$P_1+P_2+P_3=1/[1+\exp(-k_3)]$

$P_1+P_2+P_3+P_4=1/[1+\exp(-k_4)]$



Poisson regression

Εξαρτημένη Y: counts

$$\log(\lambda) = \beta_o + \beta_1 X_1 + \cdots + \beta_p X_p = \eta$$

Ερμηνεία συντελεστών:

$$\frac{\lambda_i}{\lambda_j} = e^{\beta_k} \Leftrightarrow \frac{\lambda_i}{\lambda_j} - \frac{\lambda_j}{\lambda_j} = e^{\beta 4} - 1 \Leftrightarrow \frac{\lambda_i - \lambda_j}{\lambda_j} = e^{\beta 4} - 1$$

Overdispersion (scaled deviance ή scaled Pearson chi-square

- Διόρθωση standard errors [scale(x2)]
- Negative binomial regression