

ΓΕΝΙΚΕΥΜΕΝΑ ΓΡΑΜΜΙΚΑ ΜΟΝΤΕΛΑ

Άσκηση 2. Να απαντήσετε στα ερωτήματα 2-7 της άσκησης 1 χρησιμοποιώντας c-log-log link (χρησιμοποιήστε το πολύ 4 επαναλήψεις της επαναληπτικής μεθόδου). Ποιο είναι το καλύτερο μοντέλο από τα 3 (logit, probit, c-log-log);

Λύσεις Άσκησης 2.

2) Να εκτιμηθούν (σημειακά και με δ.ε. 95%) οι παράμετροι β_0, β_1 και να δοθούν οι εκτιμημένες τιμές των y_i όταν το μοντέλο που χρησιμοποιείται είναι c-log-log

Η εκτίμηση των β_i γίνεται χρησιμοποιώντας την επαναληπτική διαδικασία

$$\beta^{(m)} = (X^T W^{(m-1)} X)^{-1} X^T W^{(m-1)} Z^{(m-1)}$$

όπου στη διωνυμική κατανομή,

$$Z_i^{(m-1)} = \left[\frac{y_i - m_i \pi_i^{(m-1)}}{m_i} g'(\pi_i^{(m-1)}) + \eta_i^{(m-1)} \right], \quad W^{(m-1)} = \text{diag} \left[\frac{m_i}{\pi_i^{(m-1)} (1 - \pi_i^{(m-1)}) g'(\pi_i^{(m-1)})^2} \right]_i$$

Αν χρησιμοποιήσουμε c-log-log, τότε $g(\pi) = \ln(-\ln(1 - \pi)) \Rightarrow g'(\pi) = \frac{1}{-(1 - \pi) \ln(1 - \pi)}$ και α-

ντικαθιστώντας στον παραπάνω αναγωγικό τύπο τελικά θα έχουμε ότι

$$Z_i^{(m-1)} = \left[-\frac{y_i - m_i \pi_i^{(m-1)}}{m_i (1 - \pi_i^{(m-1)}) \ln(1 - \pi_i^{(m-1)})} + \eta_i^{(m-1)} \right], \quad W^{(m-1)} = \text{diag} \left[\frac{m_i (1 - \pi_i^{(m-1)}) (\ln(1 - \pi_i^{(m-1)}))^2}{\pi_i^{(m-1)}} \right]_i$$

και

$$\eta_i^{(m-1)} = \mathbf{x}_i^T \beta^{(m-1)}, \quad \pi_i^{(m-1)} = g^{-1}(\eta_i^{(m-1)}) = 1 - \exp(-\exp(\eta_i^{(m-1)})).$$

Ως αρχικό σημείο επιλέγουμε:

$$\pi_i^{(0)} = \begin{cases} y_i / m_i, & y_i \neq 0, m_i \\ 0.5 / m_i, & y_i = 0 \end{cases}, \quad \eta_i^{(0)} = g(\pi_i^{(0)}).$$

Θα πραγματοποιήσουμε 4 βήματα της παραπάνω επαναληπτικής διαδικασίας.

1^ο βήμα:

$p\theta = y/m$ (εκτός της 8^{ης} παρατήρησης που αντι 1 βάζουμε 0.99)

$\text{pre_}\theta = \text{LN}(-\text{LN}(1 - p\theta))$

$z\theta = \text{pre_}\theta - (y - m * p\theta) / (m * (1 - p\theta) * \text{LN}(1 - p\theta))$

$w\theta = m * (1 - p\theta) * (\text{LN}(1 - p\theta))^{**2} / p\theta$

$\text{pre_}1 = \text{unstandardized predicted values (Linear Regression: Dep->z\theta, Indep->x, WLS weights->w\theta)}$

Coefficients^{a,b}

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
1 (Constant)	-38,343	2,275		-16,851	,000
X	21,354	1,259	,990	16,956	,000

a. Dependent Variable: Z0

b. Weighted Least Squares Regression - Weighted by W0

MSE = 0,6183,

2^ο βήμα:

$p1 = 1 - \text{Exp}(-\text{Exp}(\text{pre_}1))$

$z1 = \text{pre_}1 - (y - m * p1) / (m * (1 - p1) * \text{LN}(1 - p1))$

$w1 = m * (1 - p1) * (\text{LN}(1 - p1))^{**2} / p1$

$\text{pre_}2 = \text{unstandardized predicted values (Linear Regression: Dep->z1, Indep->x, weights->w1)}$

Coefficients^{a,b}

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-39,526	2,292		-17,242	,000
	X	22,015	1,272	,990	17,301	,000

a. Dependent Variable: Z1

b. Weighted Least Squares Regression - Weighted by W1

MSE = 0,5362,

3^ο βήμα:

p2=1-Exp(-Exp(pre_2))

z2= pre_2-(y-m*p2)/(m*(1-p2)*LN(1-p2))

w2= m*(1-p2)*(LN(1-p2))**2/p2

pre_3=unstandardized predicted values (Linear Regression: Dep->z2, Indep->x, weights->w2)

Coefficients^{a,b}

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-39,573	2,396		-16,513	,000
	X	22,041	1,331	,989	16,564	,000

a. Dependent Variable: Z2

b. Weighted Least Squares Regression - Weighted by W2

MSE = 0,5484,

4^ο βήμα:

p3=1-Exp(-Exp(pre_3))

z3= pre_3-(y-m*p3)/(m*(1-p3)*LN(1-p3))

w3= m*(1-p3)*(LN(1-p3))**2/p3

pre_4=unstandardized predicted values (Linear Regression: Dep->z3, Indep->x, weights->w3)

Coefficients^{a,b}

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-39,572	2,401		-16,481	,000
	X	22,041	1,333	,989	16,530	,000

a. Dependent Variable: Z3

b. Weighted Least Squares Regression - Weighted by W3

MSE = 0,5491,

Ο πίνακας των δεδομένων περιέχει τις μεταβλητές:

x	m	y	p0	pre 0	z0	w0	pre 1	p1	z1	w1	pre 2
1,6907	59	6	,1017	-2,23	-2,23	5,99	-2,23925	,10	-2,23	5,96	-2,30506
1,7242	60	13	,2167	-1,41	-1,41	12,94	-1,52388	,20	-1,40	11,70	-1,56756
1,7552	62	18	,2903	-1,07	-1,07	17,82	-,86190	,34	-1,06	21,04	-,88509
1,7842	56	28	,5000	-,37	-,37	26,91	-,24262	,54	-,36	28,93	-,24665
1,8113	63	52	,8254	,56	,56	40,59	,33608	,75	,54	40,41	,34996
1,8369	59	53	,8983	,83	,83	34,90	,88275	,91	,82	33,75	,91355
1,8610	62	61	,9839	1,42	1,42	17,31	1,39739	,98	1,42	18,08	1,44411
1,8839	60	60	,9900	1,53	1,74	12,85	1,88641	1,00	2,04	3,57	1,94826

p2	z2	w2	pre 3	p3	z3	w3	pre 4
,09	-2,23	5,60	-2,30732	,09	-2,23	5,58	-2,30730
,19	-1,40	11,25	-1,56894	,19	-1,40	11,24	-1,56893
,34	-1,06	20,67	-,88566	,34	-1,06	20,66	-,88565
,54	-,36	28,87	-,24646	,54	-,36	28,87	-,24646
,76	,55	40,49	,35086	,76	,55	40,49	,35086
,92	,82	33,04	,91512	,92	,82	33,00	,91511
,99	1,42	16,31	1,44631	,99	1,42	16,23	1,44631
1,00	2,09	2,65	1,95106	1,00	2,09	2,61	1,95105

Συγκεντρωτικά θα είναι

m	$\hat{\beta}_0^{(m)}$	$\hat{\beta}_1^{(m)}$	$s(\hat{\beta}_0^{(m)})$	$s(\hat{\beta}_1^{(m)})$	$s^2 = \text{MSE}^{(m)}$
1	-38.343	21.354	2.275	1.259	0.6183
2	-39.526	22.015	2.292	1.272	0.5362
3	-39.573	22.041	2.396	1.331	0.5484
4	-39.572	22.041	2.401	1.333	0.5491

$$s(\hat{\beta}_0) = \frac{s(\hat{\beta}_0^{(m)})}{\sqrt{\text{MSE}^{(m)}}} \approx \frac{2.401}{\sqrt{0.5491}} \approx 3.24, \quad s(\hat{\beta}_1) = \frac{s(\hat{\beta}_1^{(m)})}{\sqrt{\text{MSE}^{(m)}}} \approx \frac{1.333}{\sqrt{0.5491}} \approx 1.799$$

Και εδώ, τα διαστήματα εμπιστοσύνης υπολογίζονται κάτω από την παραδοχή ότι τα β_i ακολουθούν κανονική κατανομή με μέση τιμή β_i και διασπορά $s^2(\hat{\beta}_i)$ (κάτι που ισχύει προσεγγιστικά για μεγάλο δείγμα). Ένα (προσεγγιστικό) δ.ε. συντ. 95% για το β_0 θα είναι το

$$(\hat{\beta}_0 \pm s(\hat{\beta}_0) \cdot Z_{0.025}) = (-39.572 \pm 3.24 \cdot 1.96) = (-45.92, -33.22).$$

Όμοια βρίσκουμε ένα (προσεγγιστικό) δ.ε. συντ. 95% για το β_1 :

$$(\hat{\beta}_1 \pm s(\hat{\beta}_1) \cdot Z_{0.025}) = (22.041 \pm 1.799 \cdot 1.96) = (18.515, 25.567)$$

οι εκτιμημένες τιμές των y_i υπολογίζονται μέσω του τύπου

$$y_est = m * (1 - \text{Exp}(-\text{Exp}(-39.572 + 22.041 * x)))$$

και είναι 5.59, 11.28, 20.95, 30.37, 47.78, 54.14, 61.11, 59.95.

7) Να υπολογιστούν τα κατάλοιπα Pearson και deviance.

3) Δώστε την απόκλιση (deviance) και το X^2 του Pearson.

Υπενθυμίζεται ότι τα κατάλοιπα Pearson είναι τα

$$r_i^P = \frac{y_i - m_i \hat{\pi}_i}{\sqrt{m_i \hat{\pi}_i (1 - \hat{\pi}_i)}}$$

ενώ τα κατάλοιπα απόκλισης δίνονται από τον τύπο

$$r_i^D = \text{sgn}(y_i - m_i \hat{\pi}_i) \sqrt{2} \sqrt{y_i \ln \frac{y_i}{m_i \hat{\pi}_i} + (m_i - y_i) \ln \frac{m_i - y_i}{m_i - m_i \hat{\pi}_i}}$$

Για το c-log-log link: Αρχικά υπολογίζουμε (με compute) τα $\hat{\pi}_i$:

$$ep = 1 - \text{Exp}(-\text{Exp}(pre_4))$$

Στη συνέχεια σχηματίζουμε τα κατάλοιπα Pearson r_i^P :

$$r_p = (y - m * ep) / (m * ep * (1 - ep)) * 0.5$$

και τα κατάλοιπα Deviance r_i^D :

$$r_d = (y-m*ep)/ABS(y-m*ep)*(2**0.5)*ABS(y*LN(y/(m*ep)) + (m-y)*LN((m-y)/(m-m*ep)))*0.5$$

Από τα παραπάνω προκύπτει ο ακόλουθος πίνακας στον SPSS Data Editor:

x	m	y	p	ep	r_d	r_p
1,6907	59	6	0,1017	,0947	,1806	,1825
1,7242	60	13	0,2167	,1880	,5577	,5681
1,7552	62	18	0,2903	,3380	-,8033	-,7932
1,7842	56	28	0,5000	,5423	-,6344	-,6355
1,8113	63	52	0,8254	,7584	1,2888	1,2430
1,8369	59	53	0,8983	,9177	-,5237	-,5413
1,8610	62	61	0,9839	,9857	-,1188	-,1212
1,8839	60	60	1,0000	,9991	,3250	,2298

Η στήλη r_p περιέχει τα κατάλοιπα Pearson, ενώ η στήλη r_d περιέχει τα κατάλοιπα απόκλισης (c-log-log link). Με βάση τα παραπάνω αποτελέσματα μπορούμε να υπολογίσουμε και τα X^2 του Pearson και το Deviance ως εξής: Υπολογίζουμε τα $r_p2 = r_p**2$, $r_d2 = r_d**2$ και εκτελούμε τη διαδικασία Analyze/Descriptive Statistics/Descriptives/ r_p2, r_d2 (options: sum) από όπου (sums) προκύπτει ότι

$$X^2 = 3,294693723703 \quad \text{και} \quad D = 3,446438733025$$

$$\text{διότι } D_{c,f} = \sum_{i=1}^n (r_i^D)^2, \quad X_{Pearson}^2 = \sum_{i=1}^n (r_i^P)^2.$$

4) Είναι το μοντέλο ικανοποιητικό; ($\alpha=0.05$). Είναι γνωστό ότι απορρίπτουμε την H_0 : $g(\mu) = X\beta$ όταν

$$D_{c,f}(y, \hat{\mu}) > \chi_{n-p;a}^2 \quad \text{ή} \quad X_{Pearson}^2 > \chi_{n-p;a}^2$$

Παρατηρούμε το μοντέλο θεωρείται ικανοποιητικό για την περιγραφή των δεδομένων διότι $X^2 = 3,29$ και $D = 3,44 < \chi_{6,0.05}^2 \approx 12.59$ αντίστοιχα.

5) Να γίνει ο έλεγχος της υπόθεσης $H_0 : \beta_1 = 0$, με εναλλακτική $\beta_1 \neq 0$, δηλαδή ότι το αποτέλεσμα δεν εξαρτάται από την δόση.

Θα χρησιμοποιήσουμε το δ.ε. 95% του β_1 που υπολογίστηκε παραπάνω: (18.515, 25.567). Επειδή το 0 δεν ανήκει στο δ.ε. 95% απορρίπτουμε την υπόθεση $H_0 : \beta_1 = 0$ με εναλλακτική την $H_1 : \beta_1 \neq 0$, σε ε.σ. 5%.

6) Να κάνετε τη γραφική παράσταση $\hat{\pi}(x)$, $x \in (1.6, 1.9)$ σε κάθε μοντέλο μαζί με τα $(x_i, y_i/m_i)$

Στο logit, probit και c-log-log μοντέλο αντίστοιχα ισχύει ότι

$$\hat{\pi}(x) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x}}, \quad \hat{\pi}(x) = \Phi(\hat{\beta}_0 + \hat{\beta}_1 x), \quad \hat{\pi}_i(x) = 1 - \exp\{-e^{\hat{\beta}_0 + \hat{\beta}_1 x}\}$$

(τα $\hat{\beta}_0, \hat{\beta}_1$ προφανώς είναι διαφορετικά στα τρία μοντέλα). Με το SPSS:

Αρχικά σχηματίζουμε μία νέα στήλη **i** με π.χ. 31 αριθμούς από το 0 έως το 30. (0, 1, 2, ..., 30). Στη συνέχεια σχηματίζουμε (compute) τη νέα μεταβλητή

$$xg = i/30*0.3+1.6$$

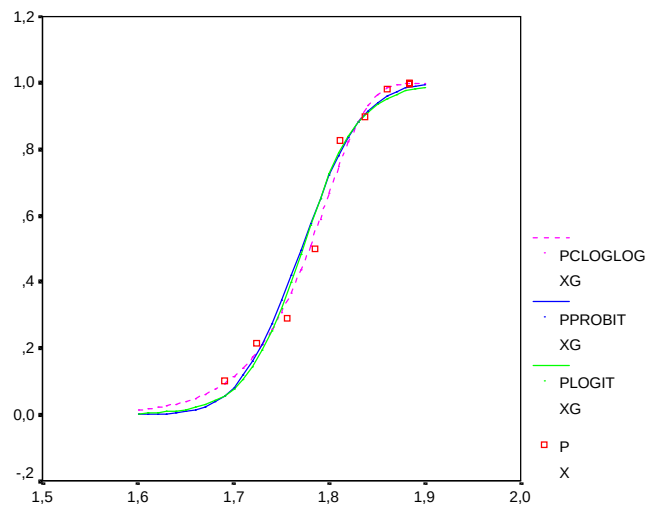
ή οποία θα παίρνει 31 τιμές από το 1.6 έως το 1.9. Στη συνέχεια σχηματίζουμε (compute) τις νέες μεταβλητές

plogit = Exp(-60.72+34.27*xg)/(1+Exp(-60.72+34.27*xg))

pprobit = CDF.NORMAL(-34.93+19.73*xg,0,1)

pcloglog = 1- Exp(-Exp(-39.572+22.041*xg))

οι οποίες μας δίνουν τις τιμές της $\hat{\pi}(x)$ στα 31 σημεία $x = xg$ για τα τρία μοντέλα. Το γράφημα της $\hat{\pi}(x)$ (και για τα τρία μοντέλα) πάνω στα 31 αυτά σημεία μαζί με τα 8 σημεία $(x_i, y_i/m_i)$ δίνεται από την διαδικασία Scatterplot/overlay με Y-X pairs: $p(=y/m) \text{ -- } x$, pprobit -- xg, plogit -- xg. (Θα χρειαστεί και πάλι να προσθέσουμε στις μεταβλητές x, p άλλες 31-8 γραμμές (π.χ. επαναλαμβάνοντας την 8^η παρατήρηση 23 ακόμη φορές) για να μην θεωρεί το πακέτο τις γραμμές 9 έως 31 στα xg, plogit, pprobit ως missing values). Μετά από κατάλληλη επεξεργασία λαμβάνουμε το επόμενο γράφημα το οποίο μας δίνει τις καμπύλες που παριστούν την σχέση μεταξύ της εκτιμώμενης **πιθανότητας $\hat{\pi}(x)$ εξουδετέρωσης** ενός εντόμου και της **συγκέντρωσης εντομοκτόνου x** στα τρία μοντέλα μαζί με τα 8 παρατηρούμενα σημεία (ποσοστό των νεκρών εντόμων σε σχέση με τη συγκέντρωση του εντομοκτόνου).



8) Ποιο είναι το καλύτερο μοντέλο από τα 3 (logit, probit, c-log-log);

Καλύτερο μοντέλο όμως μπορεί να θεωρηθεί το c-log-log εφ'όσον παρουσιάζει το μικρότερο Deviance και X^2 του Pearson.