ΓΕΝΙΚΕΥΜΕΝΑ ΓΡΑΜΜΙΚΑ ΜΟΝΤΕΛΑ, 3-6-2002

Άσκηση 1 (Δίτιμα δεδομένα). Έντομα εκτέθηκαν για 5 ώρες σε 7 διαφορετικές συγκεντρώσεις ενός εντομοκτόνου (gaseous carbon disulphine) και μετρήθηκε το αρχικό πλήθος των εντόμων και το πλήθος των νεκρών εντόμων στο τέλος του πειράματος. Τα αποτελέσματα του πειράματος ήταν

Δόση x_i ($\log_{10}CS_2 mgl^{-1}$)	**1.6907**	**1.7242**	**1.7552**	**1.7842**	**1.8113**	**1.8369**	**1.8610**	**1.8839**
Πλήθος εντόμων m_i	**59**	**60**	**62**	**56**	**63**	**59**	**62**	**60**
Πλήθος νεκρών εντόμων y_i	**6**	**13**	**18**	**28**	**52**	**53**	**61**	**60**

Η γραμμική συνάρτηση πρόβλεψης είναι $\eta_i = \beta_0 + \beta_1 x_i$.

1) Να γίνει η γραφική παράσταση (x_i, y_i / m_i).

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 i) Logistic, ii) Probit.

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

4) Είναι αυτά τα μοντέλα ικανοποιητικά; (α=0.05). Ποιο είναι το καλύτερο μοντέλο;

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.

6) Να κάνετε τη γραφική παράσταση $\hat{\pi}(x)$, $x \in (1.6,1.9)$ σε κάθε μοντέλο.

7) Να υπολογιστούν τα κατάλοιπα Pearson και deviance.

8) Βελτιώνεται σημαντικά η προσαρμογή του κάθε μοντέλου αν πάρουμε συνάρτηση πρόβλεψης $\eta_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2$. Να κάνετε τη γραφική παράσταση $\hat{\pi}(x)$, $x \in (1.5,2)$ σε κάθε μοντέλο. Τι παρατηρείτε για τη συνάρτηση πρόβλεψης πριν το $x < 1.6$;

9) Να εκτιμηθούν οι παράμετροι β_0, β_1 και οι διασπορές τους με την επαναληπτική μέθοδο στη logistic παλινδρόμηση (χρησιμοποιήστε το πολύ 4 επαναλήψεις).

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);

Άσκηση 3. Τα παρακάτω δεδομένα δίνουν το πλήθος y_{jk} των φυτών που διατήρησαν μια ιδιότητα όταν n_{jk} φυτά βρέθηκαν κάτω από διαφορετικές συνθήκες.

Ένας ποιοτικός παράγοντας (αγωγή) ήταν η αποθήκευση σε θερμοκρασία 3^0C για 48 ώρες ή η αποθήκευση σε κανονικές συνθήκες. Μία άλλη συμμεταβλητή ήταν η χρησιμοποίηση φυγόκεντρης δύναμης σε τρεις διαφορετικές τιμές 40, 150, 350.

Συνθήκη αποθήκευσης		*Φυγόκεντρος* 40	150	350
Κανονική	y_{1k}	55	52	57
	n_{1k}	102	99	108
Αγωγή	y_{2k}	55	50	50
	n_{2k}	76	81	90

Πάρε ως συμμεταβλητές τις $x_1 = \log 40 = 3.689$, $x_2 = \log 150 = 5.011$ και $x_3 = \log 350 = 5.858$

α) Κάνε γραφική παράσταση των αναλογιών $p_{jk} = y_{jk} / n_{jk}$ ως προς x_k.

β) Εξέτασε τα τρία μοντέλα (εκτίμησε παραμέτρους, αποκλίσεις, εκτιμημένες αναλογίες):

i) logit $\pi_{jk} = \alpha_j + \beta_j x_k$, $j = 1$ κανονικές συνθήκες, $j = 2$ αγωγή.

ii) logit $\pi_{jk} = \alpha_j + \beta x_k$,

iii) logit $\pi_{jk} = \alpha + \beta x_k$,

γ) Κάνε στο μοντέλο (i) τον έλεγχο $\beta_1 = \beta_2$ σε στάθμη 0.05. Επίσης κάνε τον έλεγχο $\alpha_1 = \alpha_2$ στο μοντέλο (ii) σε στάθμη 0.05.

Απαντήσεις (Άσκηση 1)

Θεωρούμε ότι οι μεταβλητές απόκρισης Y_1, Y_2, ...,Y_n (πλήθος νεκρών εντόμων στα πειράματα 1,2,...,n αντίστοιχα) προέρχονται από διωνυμική κατανομή με παραμέτρους (m_i, p_i), i = 1, 2, ..., n. Συγκεκριμένα,

$$f(y_i;\pi_i)=\binom{m_i}{y_i}\pi_i^{y_i}(1-\pi_i)^{m_i-y_i}=\exp\left\{y_i\ln\frac{\pi_i}{1-\pi_i}+m_i\ln(1-\pi_i)+\ln\binom{m_i}{y_i}\right\}$$

όπου υποθέτουμε ότι τα $\pi_1,\pi_2,...,\pi_n$ εξαρτώνται από τη συγκέντρωση εντομοκτόνου x ως εξής:

$$\pi_i=\frac{e^{\eta_i}}{1+e^{\eta_i}}=\frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}} \quad \text{στο } \textbf{Logit} \text{ μοντέλο} \quad (\eta_i=g(\pi_i)=\ln\frac{\pi_i}{1-\pi_i}),$$

$$\pi_i=\Phi(\eta_i)=\Phi(\beta_0+\beta_1 x_i), \quad \text{στο } \textbf{Probit} \text{ μοντέλο,} \quad (\eta_i=g(\pi_i)=\Phi^{-1}(\pi_i))$$

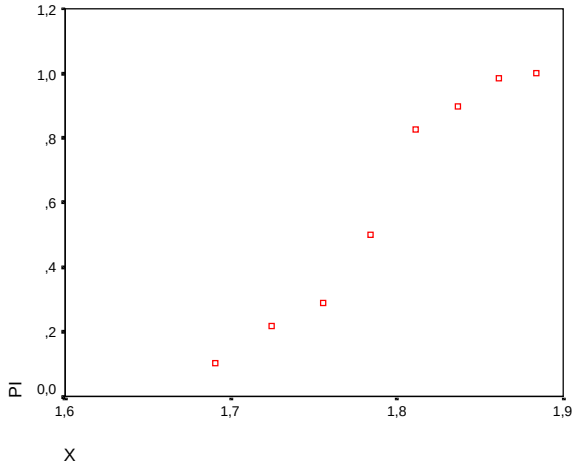
$$\pi_i=1-\exp\{-e^{\eta_i}\}=1-\exp\{-e^{\beta_0+\beta_1 x_i}\} \quad \text{στο } \textbf{c-log-log} \text{ μοντέλο} \quad (\eta_i=g(\pi_i)=\ln(-\ln(1-\pi_i)))$$

για κάποιες παραμέτρους β_0, β_1. Σκοπός μας είναι η εκτίμηση των β_0, β_1 και η εξέταση της καταλληλότητας του κάθε μοντέλου. Αρχικά εισάγουμε τα δεδομένα στο SPSS ως εξής:

x	m	y
1.6907	59	6
1.7242	60	13
1.7552	62	18
1.7842	56	28
1.8113	63	52
1.8369	59	53
1.8610	62	61
1.8839	60	60

1) Να γίνει η γραφική παράσταση (x_i = συγκέντρωση εντομοκτόνου, y_i/m_i = ποσοστό νεκρών εντόμων).

Ως συνήθως κατασκευάζουμε την νέα στήλη p = y/m (Transform/Compute) και στη συνέχεια χρησιμοποιούμε τη διαδικασία Graphs/scatter/Simple, X Axis: x, Y Axis: p.



Παρατηρούμε ότι όσο αυξάνεται η συγκέντρωση του εντομοκτόνου, τόσο αυξάνεται και το ποσοστό των νεκρών εντόμων. Αν εκτιμήσουμε τα β_0, β_1 (και το μοντέλο που χρησιμοποιούμε είναι σωστό) τότε θα έχουμε και την καμπύλη που δίνει την σχέση μεταξύ της πιθανότητας $\pi = \pi(x)$ εξουδετέρωσης ενός εντόμου και της συγκέντρωσης εντομοκτόνου x που χρησιμοποιούμε. Για παράδειγμα αν δεχτούμε το logit μοντέλο ως επαρκές, θα είναι

$$\hat{\pi}(x) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x}}.$$

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 i) Logistic, ii) Probit.
i) Αρχικά μελετάμε το μοντέλο με το Logit link. Επιλέγουμε τη διαδικασία Analyze / regression / probit (options: no fiducial c.i.) με Response: y, Total: m, Covariates: x, Model logit από όπου προκύπτει ο εξής πίνακας:

```
* * * * * * * * * * * *  P R O B I T    A N A L Y S I S  * * * * * * * * * * * *
DATA  Information

          8 unweighted cases accepted.
          0 cases rejected because of missing data.
          0 cases are in the control group.
MODEL Information

      ONLY Logistic Model is requested.

* * * * * * * * * * * *  P R O B I T    A N A L Y S I S  * * * * * * * * * * * *
 Parameter estimates converged after 16 iterations.
 Optimal solution found.

 Parameter Estimates (LOGIT model:  (LOG(p/(1-p))) = Intercept + BX):

           Regression Coeff.  Standard Error     Coeff./S.E.

   X                34,27033         2,91214        11,76809


                   Intercept  Standard Error  Intercept/S.E.

                   -60,71747         5,18071       -11,71991

  Pearson  Goodness-of-Fit  Chi Square =     10,027    DF = 6   P =  ,124

  Since Goodness-of-Fit Chi square is significant, a heterogeneity
  factor is used in the calculation of confidence limits.
```

Σύμφωνα με τα αποτελέσματα της παραπάνω διαδικασίας, οι εκτιμήσεις των β_0 β_1 μετά από 16 βήματα της επαναληπτικής διαδικασίας (Fisher's scoring method) είναι

$$\hat{\beta}_0 \approx -60.72,\ \hat{\beta}_1 \approx 34.27,\ s(\hat{\beta}_0) = \sqrt{\hat{V}(\hat{\beta}_0)} \approx 5.18,\ \ s(\hat{\beta}_1) = \sqrt{\hat{V}(\hat{\beta}_1)} \approx 2.91$$

ενώ επίσης το χι-τετράγωνο του Pearson είναι ίσο με 10.027 (με $n–p = 6$ β.ε.). Το p-value που δίνεται αφορά τον έλεγχο της υπόθεσης H_0: g(**μ**) = **Xβ** (το μοντέλο είναι σωστό) και είναι ίσο με

$$p - value = P(X^2 > 10.027 \mid X^2 \sim \chi^2_6) \approx 0.124\,.$$

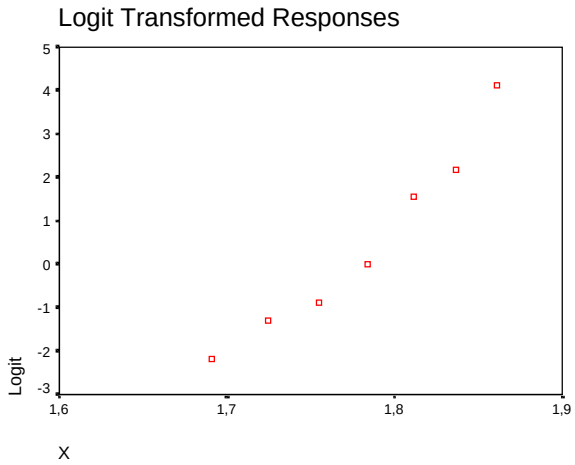
Σύμφωνα λοιπόν με το χι-τετράγωνο του Pearson, δεν έχουμε αρκετά στοιχεία ώστε να απορρίψουμε (0.124 > 0.05, 0.1) ότι το μοντέλο είναι σωστό (δηλ. το μοντέλο γίνεται αποδεκτό).

```
* * * * * * * * * * * *  P R O B I T    A N A L Y S I S  * * * * * * * * * * * *

 Observed and Expected Frequencies

              Number of    Observed    Expected
            X  Subjects   Responses   Responses    Residual     Prob

         1,69       59,0         6,0       3,457       2,543   ,05860
         1,72       60,0        13,0       9,842       3,158   ,16403
         1,76       62,0        18,0      22,451      -4,451   ,36212
         1,78       56,0        28,0      33,898      -5,898   ,60532
         1,81       63,0        52,0      50,096       1,904   ,79517
         1,84       59,0        53,0      53,291       -,291   ,90324
         1,86       62,0        61,0      59,222       1,778   ,95520
         1,88       60,0        60,0      58,743       1,257   ,97905
```

Οι τρεις πρώτες στήλες του παραπάνω πίνακα είναι οι x, m, y αντίστοιχα. Η 4η στήλη (expected responses) περιέχει τις εκτιμημένες (fitted) τιμές $\hat{\mu}_i = \hat{y}_i$ οι οποίες (λόγω του logit link) θα είναι:

$$\hat{\mu}_i = \hat{y}_i = m_i \hat{\pi}_i, \quad \hat{\pi}_i = \frac{e^{\hat{\eta}_i}}{1+e^{\hat{\eta}_i}} = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 x_i}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 x_i}}$$

(πρόκειται για τον εκτιμημένο αναμενόμενο αριθμό νεκρών εντόμων σε κάθε μία από τις $n = 8$ συγκεντρώσεις εντομοκτόνου x_i). Η στήλη με τα κατάλοιπα (residuals) περιέχει τις διαφορές $y_i - \hat{y}_i$ (η διαφορά μεταξύ του παρατηρούμενου και του εκτιμημένου αναμενόμενου αριθμού νεκρών εντόμων). Τέλος, η στήλη Prob αποτελείται από τα $\hat{\pi}_i$ (εκτιμημένες πιθανότητες εξουδετέρωσης στις $n = 8$ συγκεντρώσεις εντομοκτόνου). Επίσης το SPSS παρέχει και το ακόλουθο γράφημα



το οποίο αποτελείται από τα σημεία

$$(x_i, \ln\frac{y_i/m_i}{1-y_i/m_i}).$$

Το γράφημα αυτό χρησιμοποιείται για έναν πρόχειρο έλεγχο καταλληλότητας της συνάρτησης σύνδεσης. Συγκεκριμένα, όσο πιο «κοντά» βρίσκονται τα σημεία σε μία ευθεία, τόσο καλύτερα ερμηνεύονται τα δεδομένα από το μοντέλο με τη συγκεκριμένη συνάρτηση σύνδεσης. Αυτό συμβαίνει διότι, σύμφωνα με τα παραπάνω, αν το logit link είναι σωστό θα πρέπει για κάποια β_0, β_1 να είναι

$$\beta_0 + \beta_1 x_i = g(\pi_i) = \ln\frac{\pi_i}{1-\pi_i}$$

ή με άλλα λόγια θα πρέπει να υπάρχει γραμμική σχέση μεταξύ των x_i και των $\ln\pi_i/(1-\pi_i)$. Μία απλή εκτίμηση των π_i η οποία δεν βασίζεται στην υπόθεση ότι όντως υπάρχει μία τέτοια σχέση δίνεται από τα y_i/m_i. (Υπενθυμίζεται ότι η εκτίμηση των π_i μέσω των $\hat{\beta}_0, \hat{\beta}_1$ γίνεται με την υπόθεση ότι το link είναι σωστό και άρα δεν είναι λογικό να χρησιμοποιηθεί για τον έλεγχο του link).

Όμοια για το Probit μοντέλο:

```
* * * * * * * * * * * *  P R O B I T    A N A L Y S I S  * * * * * * * * * * * *
DATA  Information

          8 unweighted cases accepted.
          0 cases rejected because of missing data.
          0 cases are in the control group.
MODEL Information

      ONLY Normal Sigmoid is requested.
```

```
* * * * * * * * * * * *  P R O B I T    A N A L Y S I S  * * * * * * * * * * * *
 Parameter estimates converged after 13 iterations.
 Optimal solution found.

 Parameter Estimates (PROBIT model:  (PROBIT(p)) = Intercept + BX):

           Regression Coeff.  Standard Error     Coeff./S.E.

   X               19,72798         1,48406        13,29324


                  Intercept  Standard Error  Intercept/S.E.

                  -34,93535         2,63951       -13,23555

  Pearson  Goodness-of-Fit  Chi Square =      9,513    DF = 6   P =  ,147

  Since Goodness-of-Fit Chi square is significant, a heterogeneity
  factor is used in the calculation of confidence limits.
```

Σε αυτό το μοντέλο μετά από 13 επαναλήψεις προκύπτει ότι,

$$\hat{\beta}_0 \approx -34.93,\ \hat{\beta}_1 \approx 19.73,\ s(\hat{\beta}_0) = \sqrt{\hat{V}(\hat{\beta}_0)} \approx 2.63,\quad s(\hat{\beta}_1) = \sqrt{\hat{V}(\hat{\beta}_1)} \approx 1.48$$

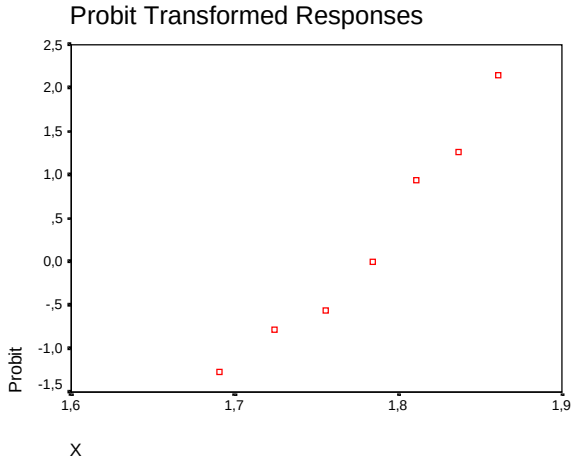
ενώ το χι-τετράγωνο του Pearson είναι ίσο με 9.513 (με $n–p = 6$ β.ε.). Και εδώ δεν έχουμε αρκετά στοιχεία ώστε να απορρίψουμε (p-value = 0.147 > 0.05, 0.1) την H_0: g(**μ**) = **Xβ** δηλ. ότι το μοντέλο είναι σωστό.

```
* * * * * * * * * * * *  P R O B I T    A N A L Y S I S  * * * * * * * * * * * *
 Observed and Expected Frequencies

            X   Number of Subjects   Observed Responses   Expected Responses     Residual      Prob

         1,69              59,0                 6,0                3,358        2,642    ,05691
         1,72              60,0                13,0               10,722        2,278    ,17869
         1,76              62,0                18,0               23,482       -5,482    ,37874
         1,78              56,0                28,0               33,816       -5,816    ,60385
         1,81              63,0                52,0               49,616        2,384    ,78755
         1,84              59,0                53,0               53,319         -,319   ,90371
         1,86              62,0                61,0               59,665        1,335    ,96233
         1,88              60,0                60,0               59,228         ,772    ,98713
```

Ο παραπάνω πίνακας είναι αντίστοιχος με την περίπτωση που είχαμε logit link. Η διαφορά εδώ είναι στο ότι η 4η στήλη (expected responses) περιέχει τις εκτιμημένες (fitted) τιμές $\hat{\mu}_i = \hat{y}_i$ οι οποίες (λόγω του probit link) είναι :

$$\hat{y}_i = m_i\,\hat{\pi}_i = m_i \Phi(\hat{\eta}_i) = m_i \Phi(\hat{\beta}_0 + \hat{\beta}_1 x_i)$$

Αντίστοιχα, η στήλη με τα κατάλοιπα περιέχει τις διαφορές $y_i - \hat{y}_i$ ενώ η στήλη Prob αποτελείται από τα $\hat{\pi}_i = \Phi(\hat{\beta}_0 + \hat{\beta}_1 x_i)$. Τέλος, το παρακάτω γράφημα



αποτελείται από τα σημεία (Probit link)

$$(x_i, \Phi^{-1}(y_i / m_i)) \text{ όπου } \Phi \text{ είναι η συνάρτηση κατανομής της } \mathrm{N}(0,1)$$

Όμοια με το Logit μοντέλο, το γράφημα αυτό χρησιμοποιείται για έναν πρόχειρο έλεγχο καταλληλότητας της συνάρτησης σύνδεσης. Όσο πιο «κοντά» βρίσκονται τα σημεία σε μία ευθεία, τόσο καλύτερα ερμηνεύονται τα δεδομένα μέσω του Probit link (Αν το Probit link είναι σωστό θα πρέπει για κάποια β_0, β_1 να είναι $\beta_0 + \beta_1 x_i = g(\pi_i) = \Phi^{-1}(\pi_i)$).

Τα διαστήματα εμπιστοσύνης (συντελεστού 95%) 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). Π.χ. για την περίπτωση όπου έχουμε logit link, ένα (προσεγγιστικό) δ.ε. συντ. 95% για το β_0 θα είναι το

$$\left(\hat{\beta}_0 \pm s(\hat{\beta}_0) \cdot Z_{0.025}\right) = (-60.72 \pm 5.173 \cdot 1.96) = (-70.87, -50.57) .$$

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

$$\left(\hat{\beta}_1 \pm s(\hat{\beta}_1) \cdot Z_{0.025}\right) = (28.57, 39.97)$$

Για το Probit μοντέλο αντίστοιχα λαμβάνουμε τα δ.ε. 95% :

$$\left(\hat{\beta}_0 \pm s(\hat{\beta}_0) \cdot Z_{0.025}\right) = (-40.12, -29.74), \quad \left(\hat{\beta}_1 \pm s(\hat{\beta}_1) \cdot Z_{0.025}\right) = (16.81, 22.64)$$

3) Δώστε την απόκλιση (deviance) και το X^2 του Pearson σε κάθε μοντέλο.
Το X^2 του Pearson δίνεται απευθείας από το πακέτο:

$$X^2 = \sum_{i=1}^{n} \frac{m_i (y_i - \hat{\mu}_i)^2}{\hat{\mu}_i (m_i - \hat{\mu}_i)} = 10.027 \text{ , } 9.513 \quad \text{(Logit, Probit link)}$$

Η απόκλιση δεν δίνεται από το SPSS και θα πρέπει να υπολογιστεί μέσω των $\hat{\pi}_i$ (ο υπολογισμός της ζητείται για λόγους εκπαιδευτικούς. Το X^2 του Pearson αρκεί για τους ελέγχους καταλληλότητας του μοντέλου). Συγκεκριμένα γνωρίζουμε ότι

$$D_{c,f}(\mathbf{y}, \hat{\boldsymbol{\mu}}) = 2\sum_{i=1}^{n} \left(y_i \ln \frac{y_i}{m_i \hat{\pi}_i} + (m_i - y_i) \ln \frac{m_i - y_i}{m_i - m_i \hat{\pi}_i} \right), \quad \hat{\pi}_i = g^{-1}(\hat{\beta}_0 + \hat{\beta}_1 x_i) .$$

Ο υπολογισμός του μπορεί π.χ. να γίνει χρησιμοποιώντας το αποτέλεσμα παρακάτω ερωτήματος στο οποίο υπολογίζονται τα κατάλοιπα απόκλισης. Βρίσκουμε ότι D = 11.232 και 10.120 για το Logit και Probit μοντέλο αντίστοιχα. (βλ. απάντηση ερωτήματος 7 για λεπτομέρειες).

4) Είναι αυτά τα μοντέλα ικανοποιητικά; (α=0.05). Ποιο είναι το καλύτερο μοντέλο;
Είναι γνωστό ότι απορρίπτουμε την H_0: $g(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta}$ όταν

$$D_{c,f}(\mathbf{y}, \hat{\boldsymbol{\mu}}) > \chi^2_{n-p;a} \quad \text{ή} \quad X^2_{Pearson} > \chi^2_{n-p;a}$$

Παρατηρούμε ότι και τα δύο μοντέλα θεωρούνται ικανοποιητικά για την περιγραφή των δεδομένων. Ειδικότερα, δεν είμαστε σε θέση να απορρίψουμε την υπόθεση ότι το εκάστοτε μοντέλο είναι σωστό σε επίπεδο σημαντικότητας 0.05 επειδή $D = 11.232$ και $10.120 < \chi^2_{6,0.05} \approx 12.59$ αντίστοιχα, ή ισοδύναμα επειδή το p-value του κάθε μοντέλου

p-value(logit model)= $P(X^2 > 11.232 \,|\, X^2 \sim \chi^2_6) \approx 0.0815$,

p-value(probit model)= $P(X^2 > 10.120 \,|\, X^2 \sim \chi^2_6) \approx 0.1197$,

είναι μεγαλύτερο του 0.05. Ελάχιστα καλύτερο μοντέλο όμως μπορεί να θεωρηθεί το probit εφ'όσον παρουσιάζει το μικρότερο Deviance ή ισοδύναμα το μεγαλύτερο p-value (Οι αποκλίσεις μπορούν να συγκριθούν άμεσα γιατί έχουν τους ίδιους βαθμούς ελευθερίας).

Επίσης, σύμφωνα και με το X^2 του Pearson (X^2=10.022, 9.513 αντίστοιχα) τα δύο μοντέλα είναι αποδεκτά ($X^2_{Pearson} < \chi^2_{6,0.05} \approx 12.59$) και μάλιστα το καλύτερο είναι και πάλι το probit.

5) Να γίνει ο έλεγχος της υπόθεσης $H_0 : \beta_1 = 0$, με εναλλακτική $H_1 : \beta_1 \neq 0$, δηλαδή ότι το αποτέλεσμα δεν εξαρτάται από την δόση.

Θα χρησιμοποιήσουμε το δ.ε. 95% του β_1 που υπολογίστηκε παραπάνω:

(28.57,39.97) (logit model), (16.81,22.64) (probit model)

Επειδή το 0 δεν ανήκει στα δ.ε. 95% απορρίπτουμε την υπόθεση $H_0 : \beta_1 = 0$ με εναλλακτική την $H_1 : \beta_1 \neq 0$, και στα δύο μοντέλα 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 $(x_i, y_i/m_i)$

Εδώ ζητείται η γραφική παράσταση της καμπύλης που εκφράζει την σχέση μεταξύ της εκτιμημένης πιθανότητας $\hat{\pi}(x)$ εξουδετέρωσης ενός εντόμου και της συγκέντρωσης εντομοκτόνου x που χρησιμοποιούμε. Στο logit και probit μοντέλο αντίστοιχα ισχύει ότι

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

(τα $\hat{\beta}_0, \hat{\beta}_1$ προφανώς είναι διαφορετικά στα δύο μοντέλα). Εφόσον τα $\hat{\beta}_0, \hat{\beta}_1$ είναι γνωστά σε κάθε μοντέλο, οι παραπάνω γραφικές παραστάσεις (άξονας x: x, άξονας y: $\hat{\pi}(x)$) είναι εύκολο να γίνουν χρησιμοποιώντας κατάλληλο λογισμικό Η/Υ. Ας δούμε πως γίνεται δυνατή η κατασκευή ενός τέτοιου γραφήματος 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

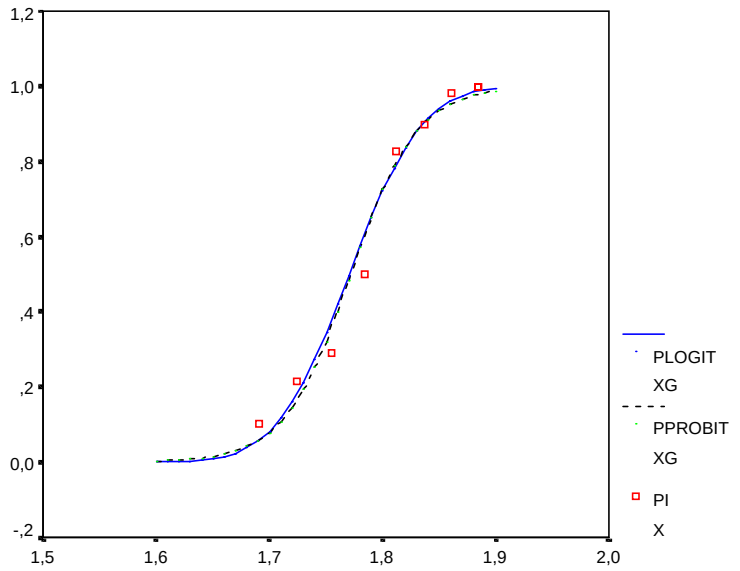
```
xg = i/30*0.3+1.6
```

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)
```

οι οποίες μας δίνουν τις τιμές της $\hat{\pi}(x)$ στα 31 σημεία x = xg για το logit και probit μοντέλο αντίστοιχα. Το γράφημα της $\hat{\pi}(x)$ (και για τα δύο 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) -- x, pprobit -- xg, plogit -- xg. (Θα χρειαστεί 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).

Στη συνέχεια ανοίγουμε τον SPSS Chart Editor κάνοντας διπλό κλικ στο σχήμα που προέκυψε από την παραπάνω διαδικασία και επιλέγουμε τα 31 σημεία που αφορούν το plogit και τα ενώνουμε με Format / interpolation / straight line (παράλληλα μικραίνουμε τα 31 σημεία με Format / Marker). Επαναλαμβάνουμε το ίδιο και για τα 31 σημεία που αφορούν το pprobit (επιπλέον, με το Format/line style μπορούμε να κάνουμε διακεκομμένες τις συγκεκριμένες γραμμές). Ως αποτέλεσμα λαμβάνουμε το επόμενο γράφημα το οποίο μας δίνει τις καμπύλες που παριστούν την σχέση μεταξύ της εκτιμημένης **πιθανότητας** $\hat{\pi}(x)$ **εξουδετέρωσης** ενός εντόμου και της **συγκέντρωσης εντομοκτόνου** x στα μοντέλα logit και probit μαζί με τα 8 παρατηρούμενα σημεία (ποσοστό των νεκρών εντόμων σε σχέση με τη συγκέντρωση του εντομοκτόνου).



Παρατηρούμε ότι για τιμές συγκέντρωσης του εντομοκτόνου μικρότερες του 1.7 η πιθανότητα εξουδετέρωσης αυξάνεται πολύ αργά (κάτω από 1.6 η αποτελεσματικότητα του εντομοκτόνου είναι μηδενική). Για τιμές μεταξύ του 1.7 και 1.85 η αύξηση της αποτελεσματικότητας του εντομοκτόνου (δηλ. πιθανότητας εξουδετέρωσης) γίνεται ραγδαία. Τέλος για τιμές του εντομοκτόνου μεγαλύτερες του 1.9 η πιθανότητα εξουδετέρωσης είναι σχεδόν 1. Αν π.χ. τώρα ένα εργοστάσιο παραγωγής εντομοκτόνων επιθυμεί να βρεί τη βέλτιστη συγκέντρωση του εντομοκτόνου που πρέπει να περιέχει ένα σκεύασμα (σταθμίζοντας μεταξύ κόστους και αποτελεσματικότητας) θα πρέπει να βασιστεί στην παραπάνω καμπύλη.

7) 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 Logit link: Αρχικά υπολογίζουμε (με compute) τα $\hat{\pi}_i$:

```
ep = Exp(-60.72+34.27*x)/(1+ Exp(-60.72+34.27*x))
```

(ή μπορούμε να τα αντιγράψουμε από τον πίνακα που δίνει το πακέτο). 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_p	r_d
1,6907	59	6	0,1017	0,0584	1,4168	1,2898
1,7242	60	13	0,2167	0,1636	1,1112	1,0689

1,7552	62	18	0,2903	0,3614	-1,1650	-1,1845
1,7842	56	28	0,5000	0,6046	-1,6004	-1,5826
1,8113	63	52	0,8254	0,7947	0,6039	0,6160
1,8369	59	53	0,8983	0,9030	-0,1208	-0,1200
1,8610	62	61	0,9839	0,9551	1,0950	1,2556
1,8839	60	60	1,0000	0,9790	1,1349	1,5965

Η στήλη r_p περιέχει τα κατάλοιπα Pearson, ενώ η στήλη r_d περιέχει τα κατάλοιπα απόκλισης. (logit 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 $X^2 = 10,02686725905$ και D =11,23280219627 διότι

$$D_{c,f} = \sum_{i=1}^{n} (r_i^D)^2, \quad X^2_{Pearson} = \sum_{i=1}^{n} (r_i^P)^2$$

Ανάλογα εργαζόμαστε και στο μοντέλο Probit.

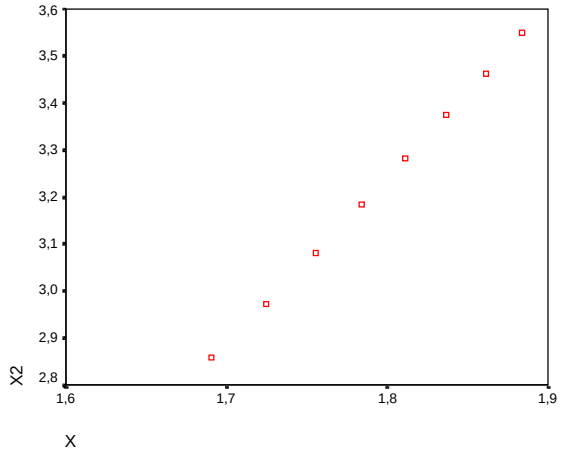
8) Αρχικά κατασκευάζουμε τη νέα μεταβλητή **x2** (= x^2).

y	**m**	**x**	**x2 (=x^2)**
6	59	1.6907	2,8585
13	60	1.7242	2,9729
18	62	1.7552	3,0807
28	56	1.7842	3,1834
52	63	1.8113	3,2808
53	59	1.8369	3,3742
61	62	1.8610	3,4633
60	60	1.8839	3,5491

Εκτελώντας τη διαδικασία Analyze/regression/probit/Response Frequency:y, Total Observed:m, Covariates: x, x2 προκύπτει ο παρακάτω πίνακας ο οποίος μας πληροφορεί ότι το SPSS δεν μπορεί να υπολογίσει τον πίνακα συνδιασπορών λόγω υψηλής συσχέτισης μεταξύ των συμμεταβλητών x, x2 (ίσως χρειάζεται να αυξήσουμε τον αριθμό των maximum iterations στα options)

```
* * * * * * * * * * P R O B I T   A N A L Y S I S * * * * * * * * * *

 Parameter estimates converged after 27 iterations.

>Warning # 13532
>Covariance matrix of parameter estimates cannot be computed, possibly due
>to linear dependencies among covariates.  Statistics will not be reported.
```

Πράγματι, τα σημεία (**x**, **x2**) έχουν υψηλή γραμμική συσχέτιση όπως αυτό φαίνεται στο παρακάτω γράφημα (συντελεστής συσχέτισης ίσος με 1).



Correlations

		X	X2
X	Pearson Correlation	1,000	1,000**
	Sig. (2-tailed)	,	,000
	N	8	8
X2	Pearson Correlation	1,000**	1,000
	Sig. (2-tailed)	,000	,
	N	8	8

**. Correlation is significant at the 0.01 level

Το γεγονός αυτό συμβαίνει διότι η x παίρνει τιμές σε μία μικρή περιοχή (1.69, 1.89) στην οποία η συνάρτηση $f(x) = x^2$ είναι σχεδόν γραμμική. Για να αποφύγουμε το φαινόμενο αυτό θα πρέπει να μετασχηματίσουμε τα δεδομένα x (αλλαγή κλίμακας) ώστε να παίρνουν τιμές σε ένα καταλληλότερο διάστημα, π.χ. στο (–3,3). Θέτουμε λοιπόν $z_i = (x_i - \bar{x})/s_x$ (θα μπορούσαμε π.χ. να είχαμε θεωρήσει και τον μετασχηματισμό $z_i = (x_i - \bar{x})/\max_j\{|x_j - \bar{x}|\}$ ώστε $z \in [-1,1]$).

Εκτελούμε λοιπόν τη διαδικασία Analyze/Descriptive Statistics/Descriptives (save standardized values as variables) από όπου λαμβάνουμε την τυποποιημένη $z = (x - \bar{x})/s_x$ όπου, όπως φαίνεται και από τον παρακάτω πίνακα, $\bar{x} = 1.793425, s_x \approx 0.0674563$.

Descriptive Statistics

	N	Mean	Std. Deviation
X	8	1,793425	6,74563E-02
Valid N (listwise)	8		

Στη συνέχεια κατασκευάζουμε την z2 = z^2 και εκτελούμε τη διαδικασία Analyze / regression / probit / Response Frequency:y, Total Observed:m, Covariates: z, z2. Από τη διαδικασία αυτή εκτιμώνται οι παράμετροι γ_0, γ_1, γ_2 για τις οποίες ισχύει ότι

$$\eta_i = \gamma_0 + \gamma_1 z_i + \gamma_2 z_i^2 = \gamma_0 + \gamma_1 \frac{x_i - \bar{x}}{s_x} + \gamma_2 \frac{(x_i - \bar{x})^2}{s_x^2}$$

$$= \gamma_0 + \gamma_1 \frac{x_i}{s_x} - \gamma_1 \frac{\bar{x}}{s_x} + \gamma_2 \frac{x_i^2}{s_x^2} + \gamma_2 \frac{\bar{x}^2}{s_x^2} - \gamma_2 \frac{2x_i\bar{x}}{s_x^2}$$

$$= \left(\gamma_0 - \gamma_1 \frac{\bar{x}}{s_x} + \gamma_2 \frac{\bar{x}^2}{s_x^2}\right) + \left(\gamma_1 \frac{1}{s_x} - \gamma_2 \frac{2\bar{x}}{s_x^2}\right) x_i + \left(\gamma_2 \frac{1}{s_x^2}\right) x_i^2$$

$$= \beta_0 + \beta_1 x_i + \beta_2 x_i^2,$$

δηλαδή,

$$\beta_0 = \gamma_0 - \gamma_1 \frac{\bar{x}}{s_x} + \gamma_2 \frac{\bar{x}^2}{s_x^2}, \quad \beta_1 = \gamma_1 \frac{1}{s_x} - \gamma_2 \frac{2\bar{x}}{s_x^2}, \quad \beta_2 = \gamma_2 \frac{1}{s_x^2}$$

Επομένως, οι εκτιμήσεις των $\beta_0, \beta_1, \beta_2$ υπολογίζονται μέσω των παραπάνω σχέσεων από τις εκτιμήσεις των $\gamma_0, \gamma_1, \gamma_2$ που δίνει το SPSS. Συγκεκριμένα λαμβάνουμε τους πίνακες (Logit link)

```
* * * * * * * * * *  P R O B I T    A N A L Y S I S  * * * * * * * * * *
DATA  Information
          8 unweighted cases accepted.
          0 cases rejected because of missing data.
          0 cases are in the control group.
MODEL Information
      ONLY Logistic Model is requested.
```

```
* * * * * * * * * * P R O B I T   A N A L Y S I S * * * * * * * * * * *

 Parameter estimates converged after 18 iterations.
 Optimal solution found.

 Parameter Estimates (LOGIT model:  (LOG(p/(1-p))) = Intercept + BX):

          Regression Coeff.  Standard Error     Coeff./S.E.
   Z                2,72588          ,28792         9,46744
   Z2                ,71173          ,26330         2,70314

                  Intercept  Standard Error  Intercept/S.E.
                     ,49936          ,16592         3,00974

  Pearson  Goodness-of-Fit  Chi Square =      3,004    DF = 5   P =  ,699

  Since Goodness-of-Fit Chi square is NOT significant, no heterogeneity
  factor is used in the calculation of confidence limits.
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
 Covariance(below) and Correlation(above) Matrices of Parameter Estimates

                     Z         Z2
 Z              ,08290     ,71348
 Z2             ,05409     ,06933
```

Άρα,

$$\hat{\gamma}_0 \approx 0.49936,\ \hat{\gamma}_1 \approx 2.72588,\ \hat{\gamma}_2 \approx 0.71173,\ \ s(\hat{\gamma}_0) \approx 0.16592,\ \ s(\hat{\gamma}_1) \approx 0.28792,\ \ s(\hat{\gamma}_2) \approx 0.26330$$

και επομένως,

$$\hat{\beta}_0 = \hat{\gamma}_0 - \hat{\gamma}_1 \frac{\bar{x}}{s_x} + \hat{\gamma}_2 \frac{\bar{x}^2}{s_x^2} \approx 0.49936 - 2.72588 \frac{1.793425}{0.0674563} + 0.71173 \frac{1.793425^2}{0.0674563^2} \approx 431.1$$

$$\hat{\beta}_1 = \hat{\gamma}_1 \frac{1}{s_x} - \hat{\gamma}_2 \frac{2\bar{x}}{s_x^2} \approx \frac{2.72588}{0.0674563} - 2 \cdot 0.71173 \frac{1.793425}{0.0674563^2} \approx -520.6$$

$$\hat{\beta}_2 = \hat{\gamma}_2 \frac{1}{s_x^2} \approx \frac{0.71173}{0.0674563^2} \approx 156.4$$

με

$$V(\hat{\beta}_1) = \frac{1}{s_x^2} V(\hat{\gamma}_1) + \frac{4\bar{x}^2}{s_x^4} V(\hat{\gamma}_2) - \frac{2\bar{x}}{s_x^3} Cov(\hat{\gamma}_1, \hat{\gamma}_2) \Rightarrow$$

$$s^2(\hat{\beta}_1) \approx \frac{1}{0.0674563^2} 0.08290 + \frac{4 \cdot 1.793425^2}{0.0674563^4} 0.06933 - \frac{2 \cdot 1.793425}{0.0674563^3} 0.05409 \approx 42464.35$$

και

$$V(\hat{\beta}_2) = \frac{1}{s_x^4} V(\hat{\gamma}_2) \Rightarrow s^2(\hat{\beta}_2) \approx \frac{0.06933}{0.0674563^4} \approx 3348.3$$

Οι τιμές του $X^2_{Pearson}$ και των fitted values που δίνει το πακέτο από την παραπάνω διαδικασία (y, m, z, z2) είναι οι σωστές διότι

$$g(\hat{\pi}_i) = \hat{\eta}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_2 x_i^2 = \hat{\gamma}_0 + \hat{\gamma}_1 z_i + \hat{\gamma}_2 z_i^2 .$$

```
* * * * * * * * * * * P R O B I T   A N A L Y S I S * * * * * * * * * * *
 Observed and Expected Frequencies
              Number of    Observed    Expected
            Z  Subjects   Responses   Responses    Residual     Prob
       -1,52       59,0         6,0       7,026      -1,026   ,11908
       -1,03       60,0        13,0      10,519       2,481   ,17531
        -,57       62,0        18,0      19,000      -1,000   ,30645
        -,14       56,0        28,0      29,955      -1,955   ,53492
         ,26       63,0        52,0      49,205       2,795   ,78103
         ,64       59,0        53,0      54,734      -1,734   ,92769
        1,00       62,0        61,0      60,822        ,178   ,98100
        1,34       60,0        60,0      59,740        ,260   ,99566
```

Όμοια, για Probit link προκύπτει ότι:

```
* * * * * * * * * * * *  P R O B I T    A N A L Y S I S  * * * * * * * * * * * *
DATA  Information
          8 unweighted cases accepted.
          0 cases rejected because of missing data.
          0 cases are in the control group.
MODEL Information

      ONLY Normal Sigmoid is requested.
```

```
* * * * * * * * * * * *  P R O B I T    A N A L Y S I S  * * * * * * * * * * * *
 Parameter estimates converged after 17 iterations.
 Optimal solution found.

 Parameter Estimates (PROBIT model:  (PROBIT(p)) = Intercept + BX):

            Regression Coeff.  Standard Error     Coeff./S.E.

   Z                 1,52996          ,14433        10,60040
   Z2                 ,35164          ,13979         2,51544

                   Intercept  Standard Error  Intercept/S.E.

                      ,29141          ,09801         2,97315

  Pearson  Goodness-of-Fit  Chi Square =      2,978    DF = 5   P =  ,703

  Since Goodness-of-Fit Chi square is NOT significant, no heterogeneity
  factor is used in the calculation of confidence limits.
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
 Covariance(below) and Correlation(above) Matrices of Parameter Estimates

                       Z         Z2
 Z                ,02083     ,67889
 Z2               ,01370     ,01954
```

Άρα,

$$\hat{\gamma}_0 \approx 0.29141,\ \hat{\gamma}_1 \approx 1.52996,\ \hat{\gamma}_2 \approx 0.35164,\ \ s(\hat{\gamma}_0) \approx 0.09801,\ \ s(\hat{\gamma}_1) \approx 0.14433,\ \ s(\hat{\gamma}_2) \approx 0.13979$$

και επομένως,

$$\hat{\beta}_0 = \hat{\gamma}_0 - \hat{\gamma}_1 \frac{\bar{x}}{s_x} + \hat{\gamma}_2 \frac{\bar{x}^2}{s_x^2} \approx 0.29141 - 1.52996 \frac{1.793425}{0.0674563} + 0.35164 \frac{1.793425^2}{0.0674563^2} \approx 208.17$$

$$\hat{\beta}_1 = \hat{\gamma}_1 \frac{1}{s_x} - \hat{\gamma}_2 \frac{2\bar{x}}{s_x^2} \approx \frac{1.52996}{0.0674563} - 2 \cdot 0.35164 \frac{1.793425}{0.0674563^2} \approx -254.5$$

$$\hat{\beta}_2 = \hat{\gamma}_2 \frac{1}{s_x^2} \approx \frac{0.35164}{0.0674563^2} \approx 77.27$$

με

$$s^2(\hat{\beta}_1) \approx \frac{1}{0.0674563^2} 0.02083 + \frac{4 \cdot 1.793425^2}{0.0674563^4} 0.01954 - \frac{2 \cdot 1.793425}{0.0674563^3} 0.01370 \approx 11985.7$$

και

$$s^2(\hat{\beta}_2) \approx \frac{0.01954}{0.0674563^4} \approx 943.7$$

```
* * * * * * * * * * * *  P R O B I T    A N A L Y S I S  * * * * * * * * * * * *
 Observed and Expected Frequencies

            Number of    Observed    Expected
         Z   Subjects   Responses   Responses    Residual     Prob
     -1,52       59,0         6,0       6,529       -,529   ,11066
     -1,03       60,0        13,0      10,911       2,089   ,18185
      -,57       62,0        18,0      19,952      -1,952   ,32181
      -,14       56,0        28,0      29,980      -1,980   ,53536
       ,26       63,0        52,0      48,176       3,824   ,76470
       ,64       59,0        53,0      54,440      -1,440   ,92271
      1,00       62,0        61,0      61,086       -,086   ,98526
      1,34       60,0        60,0      59,912        ,088   ,99854
```

Παρατηρούμε και εδώ ότι τα δύο μοντέλα θεωρούνται ικανοποιητικά για την περιγραφή των δεδομένων σε επίπεδο σημαντικότητας 0.05 (X^2(logit) = 3.004, X^2(probit) = 2.978 < $\chi^2_{5,0.05} \approx 11.07$). Ελάχιστα καλύτερο θεωρείται το Probit μοντέλο με το μικρότερο X^2 του Pearson's.

Για να εξετάσουμε αν βελτιώνεται σημαντικά η προσαρμογή του κάθε μοντέλου θα ελέγξουμε την υπόθεση

$$H_0 : \eta_i = \beta_0 + \beta_1 x_i$$

$$H_1 : \eta_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2$$

Θέλουμε δηλαδή να ελέγξουμε αν ισχύει η H_0: $g(\boldsymbol{\mu}) = \mathbf{X}\cdot\boldsymbol{\beta}$ έναντι του μεγαλύτερου μοντέλου H_1: $g(\boldsymbol{\mu}) = \mathbf{X}'\cdot\boldsymbol{\beta}'$. Ως γνωστό, για έλεγχο αυτής της μορφής απορρίπτουμε την H_0 όταν

$$X^2_{Pearson} - X'^2_{Pearson} > \chi^2_{p'-p;a} = \chi^2_{3-2;a} = \chi^2_{1,a}$$

Αν α=0.05 ισχύει ότι $X^2_{1,\,0.05} = 3.84$ και άρα για τα παραπάνω μοντέλα είναι

$X^2_{Pearson} - X'^2_{Pearson} = 10.027 - 3.004 = 7.02 > 3.84$ (logit) : βελτιώνεται η προσαρμ. του μοντ.

$= 9.513 - 2.978 = 6.53 > 3.84$ (Probit), : βελτιώνεται η προσαρμ. του μοντ.

(Ισοδύναμος έλεγχος μπορεί να γίνει χρησιμοποιώντας και το Deviance). Εναλλακτικά, θα μπορούσαμε να ελέγξουμε την υπόθεση H_0: $\beta_2 = 0$ έναντι της H_1: $\beta_2 \neq 0$ εξετάζοντας αν το 0 ανήκει στο δ.ε. του β_2 σε κάθε μοντέλο (όμοια με παραπάνω ερώτημα). Συγκεκριμένα τα προσεγγιστικά δ.ε. (συντελεστού 95%) θα είναι

$$\left(\hat{\beta}_2 \pm s(\hat{\beta}_2)\cdot Z_{0.025}\right) = (43.0,\ 269.8)\ \text{(logit)},\ (17.3,\ 137.3)\ \text{(probit)}$$

και επομένως απορρίπτουμε ότι $\beta_2 = 0$ για το Logit και Probit μοντέλο.

Τέλος, και πάλι ζητείται η γραφική παράσταση της καμπύλης που εκφράζει την σχέση μεταξύ της εκτιμημένης πιθανότητας $\hat{\pi}(x)$ εξουδετέρωσης ενός εντόμου και της συγκέντρωσης εντομοκτόνου x που χρησιμοποιούμε ($x \in [1.5,\ 2]$). Στο logit και probit μοντέλο αντίστοιχα ισχύει τώρα ότι

$$\hat{\pi}(x) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2}}, \qquad \hat{\pi}(x) = \Phi(\hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2)$$

(τα $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ είναι διαφορετικά στα δύο μοντέλα). Εργαζόμαστε παρόμοια με παραπάνω ερώτημα:

- Σχηματίζουμε τη στήλη **i** με 51 αριθμούς από το 0 έως το 50, και σχηματίζουμε τη νέα μεταβλητή **xg = i/50*0.5+1.5** η οποία θα παίρνει 51 τιμές από το 1.5 έως το 2.
- Σχηματίζουμε τις νέες μεταβλητές

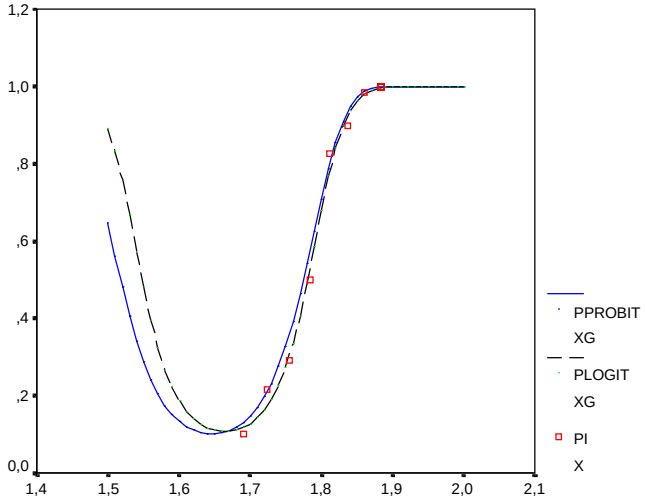
```
plogit = Exp(431.1-520.6*xg+156.4*xg**2)/(1+Exp(431.1-520.6*xg+156.4*xg**2))
pprobit = CDF.NORMAL(208.2-254.5*xg+77.3*xg**2,0,1)
```

(οι τιμές της $\hat{\pi}(x)$ στα 51 σημεία x = xg για το logit και probit μοντέλο αντίστοιχα).

- Εκτελούμε τη διαδικασία Scatterplot/overlay με Y-X pairs: p(=y/m) -- x, pprobit -- xg, plogit -- xg (θα χρειαστεί και πάλι να προσθέσουμε στις μεταβλητές x, p άλλες 51−8 τιμές).
- Ανοίγουμε τον SPSS Chart Editor και ενώνουμε τα 51 σημεία που αφορούν το plogit με Format/interpolation/straight line (παράλληλα τα μικραίνουμε με Format/Marker/Tiny) και επαναλαμβάνουμε το ίδιο και για τα σημεία που αφορούν το pprobit (κάνοντας διακεκομμένες τις συγκεκριμένες γραμμές).

Από τα παραπάνω προκύπτει το επόμενο γράφημα το οποίο μας δίνει τις καμπύλες που παριστούν την σχέση μεταξύ της εκτιμημένης **πιθανότητας** $\hat{\pi}(x)$ **εξουδετέρωσης** ενός εντόμου

και της **συγκέντρωσης εντομοκτόνου** x ($x \in [1.5, 2]$) στα μεγαλύτερα logit και probit μοντέλα μαζί με τα 8 παρατηρούμενα σημεία.



Βλέπουμε ότι οι εκτιμημένες καμπύλες του $\pi(x)$ είναι αρκετά κοντά στις παρατηρούμενες τιμές $(x_i, y_i/m_i)$ και στα δύο μοντέλα (έχουμε ήδη δει ότι έχουν περίπου ίδιο X^2 του Pearson's). Παρατηρούμε όμως ότι οι εκτιμημένες καμπύλες παρουσιάζουν μία μη αποδεκτή συμπεριφορά για $x <$ 1.69 (διότι δεν είναι λογικό να δεχτούμε ότι η μείωση της συγκεντρώσεως του εντομοκτόνου αυξάνει την πιθανότητα εξουδετέρωσης ενός εντόμου). Επομένως οι καμπύλες είναι αποδεκτές μόνο στην περιοχή των παρατηρήσεων $x > 1.69$.

9) Υπενθυμίζεται ότι η εκτίμηση των β_i γίνεται χρησιμοποιώντας την επαναληπτική διαδικασία

$$\boldsymbol{\beta}^{(m)} = (\mathbf{X}^T \mathbf{W}^{(m-1)} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}^{(m-1)} \mathbf{z}^{(m-1)}$$

όπου στη διωνυμική κατανομή,

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

Αν χρησιμοποιήσουμε logit link, τότε $g(\pi) = \log \frac{\pi}{1-\pi} \Rightarrow g'(\pi) = \frac{1}{\pi(1-\pi)}$ και αντικαθιστώντας στον παραπάνω αναγωγικό τύπο τελικά θα έχουμε ότι

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

και

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

Ως αρχικό σημείο επιλέγουμε:

$$\pi_i^{(0)} = \begin{cases} y_i / m_i, y_i \neq 0, m_i \\ 0.5/m_i, y_i = 0 \\ 1 - 0.5/m_i, \ y_i = m_i \end{cases}, \quad \eta_i^{(0)} = g(\pi_i^{(0)}).$$

(ή θα μπορούσαμε να είχαμε ξεκινήσει επιλέγοντας $\boldsymbol{\beta}^{(0)} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{p}$, $p_i = y_i / m_i$). Θα πραγματοποιήσουμε 4 βήματα της παραπάνω επαναληπτικής διαδικασίας. Επειδή

$$\boldsymbol{\beta}^{(m)} = (\mathbf{X}^T \mathbf{W}^{(m-1)} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}^{(m-1)} \mathbf{z}^{(m-1)}$$

τα $\boldsymbol{\beta}^{(m)}$ υπολογίζονται χρησιμοποιώντας WLS στη Linear Regression με dependent: $\mathbf{z}^{(m-1)}$, independent:$\mathbf{X}$, weights: $\mathbf{W}^{(m-1)}$. Υπενθυμίζεται ότι αν έχουμε το μοντέλο $\mathbf{Y} = \mathbf{X} \cdot \boldsymbol{\beta} + \mathbf{e}$, $\mathbf{e} \sim N(\mathbf{0}, \sigma^2\mathbf{V} = \sigma^2 diag[v_i]_i)$ (δηλαδή, οι διασπορές των σφαλμάτων $V(e_i) = \sigma^2 v_i$ είναι σταθμισμένες με βάρη $\mathbf{V}^{-1} = diag[1/v_i]_i$), τότε $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{Y}$ με $V(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}$)

1ο βήμα:

p0=y/m (εκτός της 8ης παρατήρησης που αντι 1 βάζουμε 0.99)

```
pre_0=LN(p0/(1-p0))
z0= pre_0+(y-m*p0)/(m*p0*(1-p0))
w0=m*p0*(1-p0)
pre_1=unstandardized predicted values (Linear Regression: Dep->z0, Indep->x, WLS weights->w0)
```

Coefficients[a,b]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-58,104	6,737		-8,625	,000
	X	32,795	3,805	,962	8,620	,000

a. Dependent Variable: Z0

b. Weighted Least Squares Regression - Weighted by W0

MSE = 1.684,

2ο βήμα:

```
p1=Exp(pre_1)/(1+Exp(pre_1))
z1=pre_1+(y-m*p1)/(m*p1*(1-p1))
w1=m*p1*(1-p1)
pre_2=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)	-60,608	6,219		-9,745	,000
	X	34,209	3,495	,970	9,788	,000

a. Dependent Variable: Z1

b. Weighted Least Squares Regression - Weighted by W1

MSE = 1.567,

3ο βήμα:

```
p2=Exp(pre_2)/(1+Exp(pre_2))
z2= pre_2+(y-m*p2)/(m*p2*(1-p2))
w2=m*p2*(1-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)	-60,717	6,676		-9,094	,000
	X	34,270	3,753	,966	9,132	,000

a. Dependent Variable: Z2

b. Weighted Least Squares Regression - Weighted by W2

MSE = 1.667,

4ο βήμα:

```
p3=Exp(pre_3)/(1+Exp(pre_3))
z3= pre_3+(y-m*p3)/(m*p3*(1-p3))
w3=m*p3*(1-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)	-60,717	6,697		-9,066	,000
	X	34,270	3,765	,966	9,103	,000

a. Dependent Variable: Z3

b. Weighted Least Squares Regression - Weighted by W3

MSE = 1.6711,

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	0,1017	-2,18	-2,18	5,39	-2,65736	0,07	-2,07	3,61	-2,77174
1,7242	60	13	0,2167	-1,29	-1,29	10,18	-1,55873	0,17	-1,26	8,62	-1,62575
1,7552	62	18	0,2903	-0,89	-,89	12,77	-0,54209	0,37	-0,87	14,41	-0,56529
1,7842	56	28	0,5000	0,00	0,00	14,00	0,40896	0,60	-0,01	13,43	0,42676
1,8113	63	52	0,8254	1,55	1,55	9,08	1,29770	0,79	1,53	10,62	1,35381
1,8369	59	53	0,8983	2,18	2,18	5,39	2,13725	0,89	2,18	5,57	2,22955
1,8610	62	61	0,9839	4,11	4,11	0,98	2,92761	0,95	3,65	2,99	3,05397
1,8839	60	60	0,9900	4,60	5,61	0,59	3,67861	0,98	4,70	1,44	3,83735

p2	z2	w2	pre_3	p3	z3	w3	pre_4
0,06	-2,00	3,27	-2,77661	0,06	-2,00	3,25	-2,77661
0,16	-1,25	8,24	-1,62855	0,16	-1,24	8,23	-1,62856
0,36	-0,88	14,32	-0,56618	0,36	-0,88	14,32	-0,56618
0,61	-0,01	13,38	0,42766	0,61	-0,01	13,38	0,42766
0,79	1,54	10,28	1,35638	0,80	1,54	10,26	1,35639
0,90	2,18	5,17	2,23370	0,90	2,18	5,16	2,23371
0,95	3,73	2,67	3,05961	0,96	3,73	2,65	3,05962
0,98	4,86	1,24	3,84440	0,98	4,87	1,23	3,84441

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	–58.104	32.795	6.737	3.805	1.684
2	–60.608	34.209	6.219	3.495	1.567
3	–60.717	34.270	6.676	3.753	1.667
4	–60.717	34.270	6.697	3.765	1.6711

Υπενθυμίζεται ότι χρησιμοποιώντας τη διαδικασία Analyze/Regression/Probit για Logit link το SPSS μας είχε αυτόματα δώσει:

$$\hat{\beta}_0 \approx -60.72,\ \hat{\beta}_1 \approx 34.27,\ s(\hat{\beta}_0) \approx 5.18,\ \ s(\hat{\beta}_1) \approx 2.91$$

Παρατηρούμε ότι οι παραπάνω τιμές ταιριάζουν απόλυτα με τις τιμές που πήραμε από την επαναληπτική διαδικασία με μόλις 4 βήματα. Οι εκτιμήσεις των $s(\hat{\beta}_0),\ s(\hat{\beta}_1)$ προκύπτουν από τις $s(\hat{\beta}_0^{(m)}),\ s(\hat{\beta}_1^{(m)})$ παρατηρώντας ότι:

Οι εκτιμήσεις των διασπορών των $\hat{\beta}_0, \hat{\beta}_1$ στα GLM ως γνωστό δίνονται από τα διαγώνια στοιχεία του πίνακα $(\mathbf{X}^T\hat{\mathbf{W}}\mathbf{X})^{-1}$, όπου $\hat{\mathbf{W}}$ είναι ο τελευταίος πίνακας $\mathbf{W}^{(m-1)}$ που χρησιμοποιήθηκε στην επαναληπτική διαδικασία ($\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta},(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1})$). Επίσης, όπως αναφέρθηκε και παραπάνω, στα σταθμισμένα ελάχιστα τετράγωνα ο πίνακας διασπορών των παραμέτρων είναι ο $V(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}$ και συνεπώς η Linear Regression με WLS στο τελευταίο βήμα της επαναληπτικής διαδικασίας δίνει ως εκτιμήσεις των τυπικών αποκλίσεων των παραμέτρων στον πίνακα

με τα coefficients (standard errors of coefficients) τις ρίζες των διαγωνίων στοιχείων του πίνακα $s^2(\mathbf{X}^T\mathbf{W}^{(m-1)}\mathbf{X})^{-1} = s^2(\mathbf{X}^T\hat{\mathbf{W}}\mathbf{X})^{-1}$ όπου $s^2 = \text{MSE}^{(m)}$. Άρα τελικά,

$$s(\hat{\beta}_0) = \frac{s(\hat{\beta}_0^{(m)})}{\sqrt{MSE^{(m)}}} \approx \frac{6.697}{\sqrt{1.6711}} \approx 5.18\,, \quad s(\hat{\beta}_1) = \frac{s(\hat{\beta}_1^{(m)})}{\sqrt{MSE^{(m)}}} \approx \frac{3.765}{\sqrt{1.6711}} \approx 2.91$$