Lecture 8: Logistic Regression

Maochao Xu

Department of Mathematics
Illinois State University
mxu2@ilstu.edu

Binomial Distribution

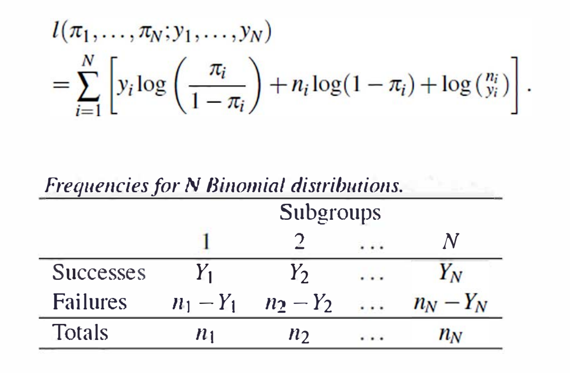
The random variable Y has the distribution Bin(n, π):

$$P(Y = y) = \binom{n}{y}\pi^y(1-\pi)^{n-y}, \quad y = 0, 1, \ldots, n.$$

Consider the general case of N independent random variables, $Y_1, \ldots, Y_N$ corresponding to the numbers of successes in N different subgroups. If

$$Y_i \sim \mathrm{Bin}(n_i, \pi)$$

the log-likelihood function is



$$l(\pi_1, \ldots, \pi_N; y_1, \ldots, y_N)$$
$$= \sum_{i=1}^{N}\left[y_i \log\left(\frac{\pi_i}{1-\pi_i}\right) + n_i \log(1-\pi_i) + \log\binom{n_i}{y_i}\right].$$

Frequencies for N Binomial distributions.

	Subgroups			
	1	2	...	N
Successes	Y_1	Y_2	...	Y_N
Failures	$n_1 - Y_1$	$n_2 - Y_2$	...	$n_N - Y_N$
Totals	n_1	n_2	...	n_N

Generalized linear models

We model the probabilities

$$g(\pi_i) = \mathbf{x}_i^T \boldsymbol{\beta},$$

where xi is a vector of explanatory variables (dummy variables for factor levels and measured values for covariates), β is a vector of parameters and g is a link function.

- Probit model

  $$g(\pi) = \Phi^{-1}(\pi).$$

  Probit models are used in several areas of biological and social sciences in which there are natural interpretations of the model.
- Logit model

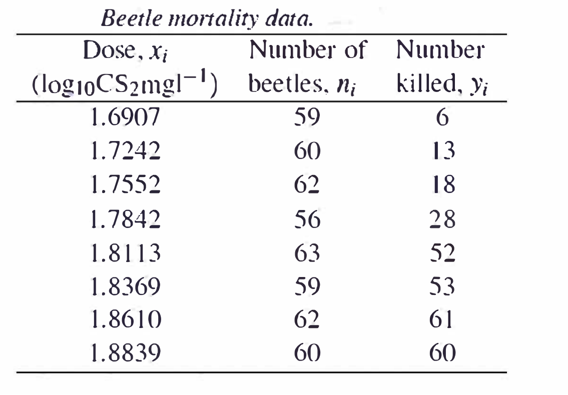
  $$g(\pi) = \log\left(\frac{\pi}{1-\pi}\right).$$

  The logistic model is widely used for Binomial data and is implemented in many statistical programs.
- Complementary log-log model

  $$g(\pi) = \log(-\log(1-\pi)).$$

Example: Beetle mortality data

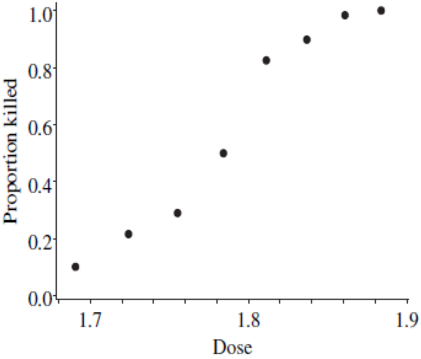
The following Table shows numbers of beetles dead after five hours of exposure to gaseous carbon disulphide at various concentrations



Beetle mortality data.

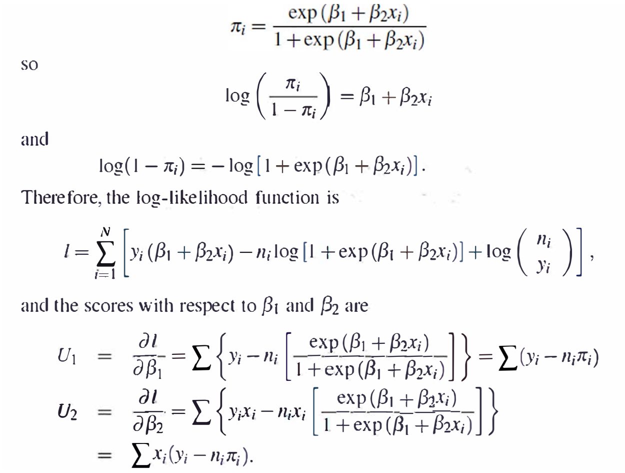
Dose, x_i	Number of	Number
$(\log_{10}\mathrm{CS}_2\mathrm{mgl}^{-1})$	beetles, n_i	killed, y_i
1.6907	59	6
1.7242	60	13
1.7552	62	18
1.7842	56	28
1.8113	63	52
1.8369	59	53
1.8610	62	61
1.8839	60	60

Example: Beetle mortality data



Logistic regression

Fitting the logistic model



$$\pi_i = \frac{\exp(\beta_1 + \beta_2 x_i)}{1 + \exp(\beta_1 + \beta_2 x_i)}$$

so

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_1 + \beta_2 x_i$$

and

$$\log(1-\pi_i) = -\log\left[1 + \exp(\beta_1 + \beta_2 x_i)\right].$$

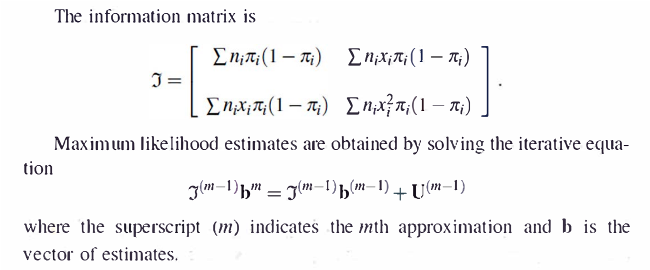
Therefore, the log-likelihood function is

$$l = \sum_{i=1}^{N}\left[y_i(\beta_1 + \beta_2 x_i) - n_i \log\left[1 + \exp(\beta_1 + \beta_2 x_i)\right] + \log\binom{n_i}{y_i}\right],$$

and the scores with respect to β_1 and β_2 are

$$\begin{aligned} U_1 &= \frac{\partial l}{\partial \beta_1} = \sum\left\{y_i - n_i\left[\frac{\exp(\beta_1 + \beta_2 x_i)}{1 + \exp(\beta_1 + \beta_2 x_i)}\right]\right\} = \sum(y_i - n_i\pi_i) \\ U_2 &= \frac{\partial l}{\partial \beta_2} = \sum\left\{y_i x_i - n_i x_i\left[\frac{\exp(\beta_1 + \beta_2 x_i)}{1 + \exp(\beta_1 + \beta_2 x_i)}\right]\right\} \\ &= \sum x_i(y_i - n_i\pi_i). \end{aligned}$$

MLE



The information matrix is

$$\mathfrak{J} = \begin{bmatrix} \sum n_i\pi_i(1-\pi_i) & \sum n_i x_i \pi_i(1-\pi_i) \\ \sum n_i x_i \pi_i(1-\pi_i) & \sum n_i x_i^2 \pi_i(1-\pi_i) \end{bmatrix}.$$

Maximum likelihood estimates are obtained by solving the iterative equation

$$\mathfrak{J}^{(m-1)}\mathbf{b}^m = \mathfrak{J}^{(m-1)}\mathbf{b}^{(m-1)} + \mathbf{U}^{(m-1)}$$

where the superscript (m) indicates the mth approximation and $\mathbf{b}$ is the vector of estimates.

The deviance can be written as

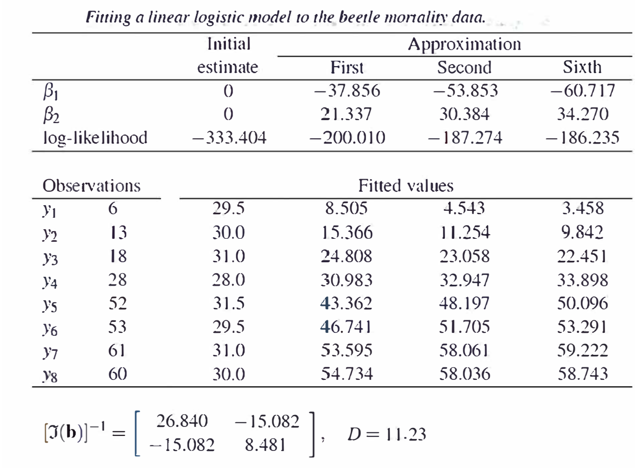
$$D = 2\sum_{i=1}^{N}\left[y_i \log\left(\frac{y_i}{\widehat{y}_i}\right) + (n_i - y_i)\log\left(\frac{n-y_i}{n-\widehat{y}_i}\right)\right]$$

Note that

$$D \sim \chi^2(N-p),$$

where p is the number of parameters estimated and N the number of covariate patterns.

Results



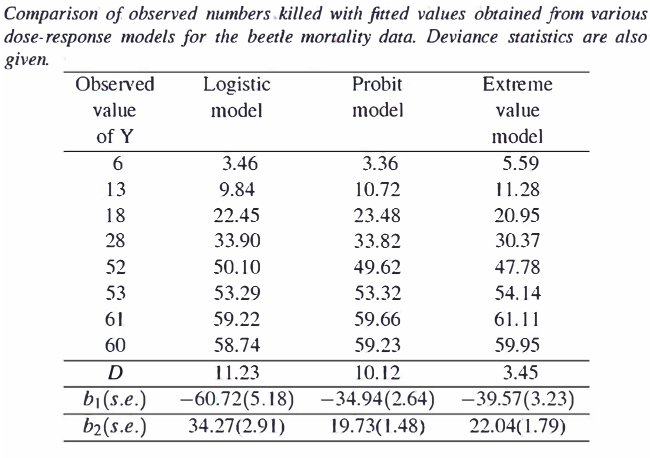
Fitting a linear logistic model to the beetle mortality data.

	Initial	Approximation		
	estimate	First	Second	Sixth
β_1	0	−37.856	−53.853	−60.717
β_2	0	21.337	30.384	34.270
log-likelihood	−333.404	−200.010	−187.274	−186.235

Observations		Fitted values			
y_1	6	29.5	8.505	4.543	3.458
y_2	13	30.0	15.366	11.254	9.842
y_3	18	31.0	24.808	23.058	22.451
y_4	28	28.0	30.983	32.947	33.898
y_5	52	31.5	43.362	48.197	50.096
y_6	53	29.5	46.741	51.705	53.291
y_7	61	31.0	53.595	58.061	59.222
y_8	60	30.0	54.734	58.036	58.743

$$[\mathfrak{I}(\mathbf{b})]^{-1} = \begin{bmatrix} 26.840 & -15.082 \\ -15.082 & 8.481 \end{bmatrix}, \quad D = 11.23$$

Model comparison

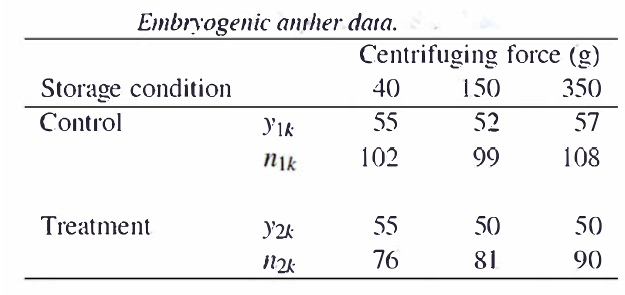


Comparison of observed numbers killed with fitted values obtained from various dose-response models for the beetle mortality data. Deviance statistics are also given.

Observed value of Y	Logistic model	Probit model	Extreme value model
6	3.46	3.36	5.59
13	9.84	10.72	11.28
18	22.45	23.48	20.95
28	33.90	33.82	30.37
52	50.10	49.62	47.78
53	53.29	53.32	54.14
61	59.22	59.66	61.11
60	58.74	59.23	59.95
D	11.23	10.12	3.45
$b_1(s.e.)$	−60.72(5.18)	−34.94(2.64)	−39.57(3.23)
$b_2(s.e.)$	34.27(2.91)	19.73(1.48)	22.04(1.79)

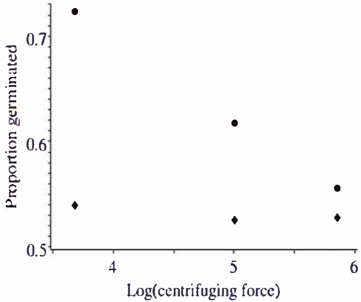
Example: Embryogenic anther data

They are numbers y_{jk} of embryogenic anthers of the plant species Datura innoxia Mill. obtained when numbers n_{jk} of anthers were prepared under several different conditions. We will compare the treatment and control effects on the proportions after adjustment (if necessary) for centrifuging force.

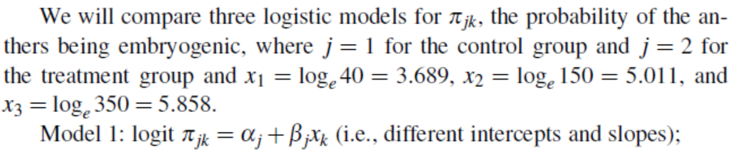


Embryogenic anther data.

		Centrifuging force (g)		
Storage condition		40	150	350
Control	y_{1k}	55	52	57
	n_{1k}	102	99	108
Treatment	y_{2k}	55	50	50
	n_{2k}	76	81	90



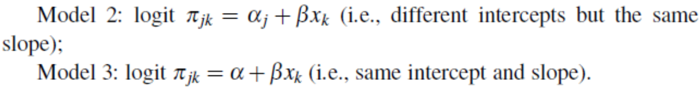
Anther data from : proportion that germinated $p_{jk} = y_{jk}/n_{jk}$ *plotted against* $\log_e$*(centrifuging force); dots represent the treatment condition and diamonds represent the control condition.*



We will compare three logistic models for π_{jk}, the probability of the anthers being embryogenic, where $j = 1$ for the control group and $j = 2$ for the treatment group and $x_1 = \log_e 40 = 3.689$, $x_2 = \log_e 150 = 5.011$, and $x_3 = \log_e 350 = 5.858$.

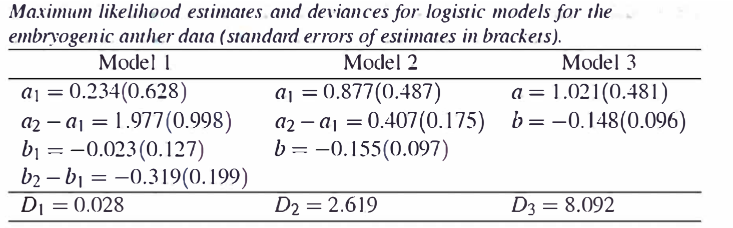
Model 1: logit $\pi_{jk} = \alpha_j + \beta_j x_k$ (i.e., different intercepts and slopes);

and



Model 2: logit $\pi_{jk} = \alpha_j + \beta x_k$ (i.e., different intercepts but the same slope);

Model 3: logit $\pi_{jk} = \alpha + \beta x_k$ (i.e., same intercept and slope).



Maximum likelihood estimates and deviances for logistic models for the embryogenic anther data (standard errors of estimates in brackets).

Model 1	Model 2	Model 3
$a_1 = 0.234(0.628)$	$a_1 = 0.877(0.487)$	$a = 1.021(0.481)$
$a_2 - a_1 = 1.977(0.998)$	$a_2 - a_1 = 0.407(0.175)$	$b = -0.148(0.096)$
$b_1 = -0.023(0.127)$	$b = -0.155(0.097)$	
$b_2 - b_1 = -0.319(0.199)$		
$D_1 = 0.028$	$D_2 = 2.619$	$D_3 = 8.092$

Goodness of fit statistics

- Pearson chi-squared statistic

$$X^2 = \sum \frac{(O-e)^2}{e}.$$

- For the binomial regression

$$X^2 = \sum_{i=1}^{N} \frac{(y_i - n_i\pi_i)^2}{n_i\pi_i(1-\pi_i)}.$$

- Estimated expected frequencies

$$X^2 = \sum_{i=1}^{N} \frac{(y_i - n_i\hat{\pi}_i)^2}{n_i\hat{\pi}_i(1-\hat{\pi}_i)}.$$

- Asymptotically equivalent to the deviances

$$D = 2\sum_{i=1}^{N} \left[y_i \log\left(\frac{y_i}{n_i\hat{\pi}_i}\right) + (n_i - y_i)\log\left(\frac{n_i - y_i}{n_i - n_i\hat{\pi}_i}\right)\right].$$

- Asymptotic distribution
$$X^2 \sim \chi^2(N - p).$$
There is some evidence to suggest that X^2 is often better than D because D is unduly influenced by very small frequencies.
- If each observation has a different covariate pattern, so y_i is zero or one, then neither D nor X^2 provides a useful measure of fit. This can happen if the explanatory variables are continuous.
- Hosmer and Lemeshow (HL) Statistic: Group observations into g categories on the basis of their predicted probabilities. Typically about 10 groups are used with approximately equal numbers of observations in each group.
$$X^2_{HL} \sim \chi^2(g - 2).$$

Residuals

- Pearson, or chi-squared, residual

$$X_k = \frac{y_k - n_k\hat{\pi}_k}{\sqrt{n_k\hat{\pi}_k(1-\hat{\pi}_k)}}.$$

- Standardized Pearson residuals

$$r_{Pk} = \frac{X_k}{1-h_k},$$

  where h_k is the leverage.

- Deviance residuals

$$d_i = \mathrm{sign}(y_i - n_i\hat{\pi}_i)\left(2\left[y_i \log\left(\frac{y_i}{n_i\hat{\pi}_i}\right) + (n_i - y_i)\log\left(\frac{n_i - y_i}{n_i - n_i\hat{\pi}_i}\right)\right]\right)^{1/2}.$$

- Deviance residual

$$r_{D_i} = \frac{d_i}{1-h_i}.$$

Residuals diagnostics

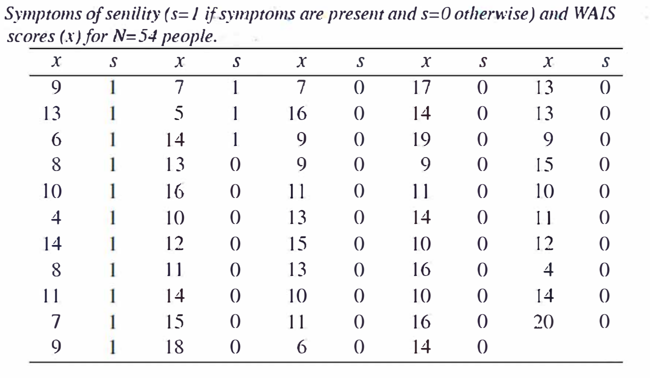
- Plotted against each continuous explanatory variable in the model to check if the assumption of linearity is appropriate and against other possible explanatory variables not included in the model.
- They should be plotted in the order of the measurements, if applicable, to check for serial correlation.
- Normal probability plots provided the numbers of observations for each covariate pattern are not too small.
- If the data are binary, or if n_k is small for most covariate patterns, then there are few distinct values of the residuals and the plots may be relatively uninformative. It may be necessary to rely on the aggregated goodness of fit statistics.
- Overdispersion. Variance is very large. This could be due to inadequate specification of the model. One approach is to include an extra parameter ϕ in the model so that

  $$\mathrm{Var}(Y_i) = n_i\pi_i(1-\pi_i)\phi.$$

  Quasibinomial distribution.

Example: Senility and WAIS

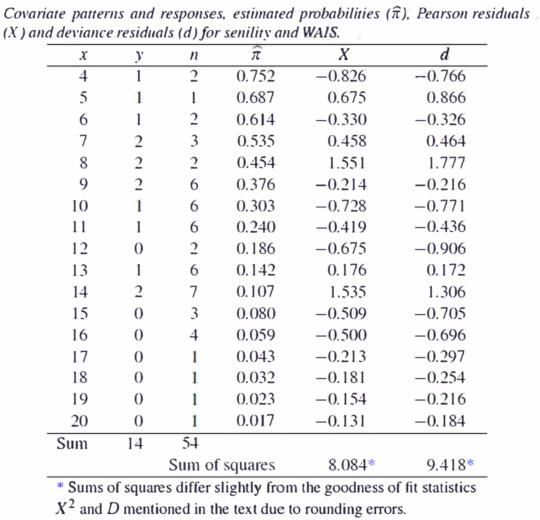
A sample of elderly people was given a psychiatric examination to determine whether symptoms of senility were present. Other measurements taken at the same time included the score on a subset of the Wechsler Adult Intelligent Scale (WAIS).



Symptoms of senility (s=1 if symptoms are present and s=0 otherwise) and WAIS scores (x) for N=54 people.

x	s	x	s	x	s	x	s	x	s
9	1	7	1	7	0	17	0	13	0
13	1	5	1	16	0	14	0	13	0
6	1	14	1	9	0	19	0	9	0
8	1	13	0	9	0	9	0	15	0
10	1	16	0	11	0	11	0	10	0
4	1	10	0	13	0	14	0	11	0
14	1	12	0	15	0	10	0	12	0
8	1	11	0	13	0	16	0	4	0
11	1	14	0	10	0	10	0	14	0
7	1	15	0	11	0	16	0	20	0
9	1	18	0	6	0	14	0		

Example: Senility and WAIS

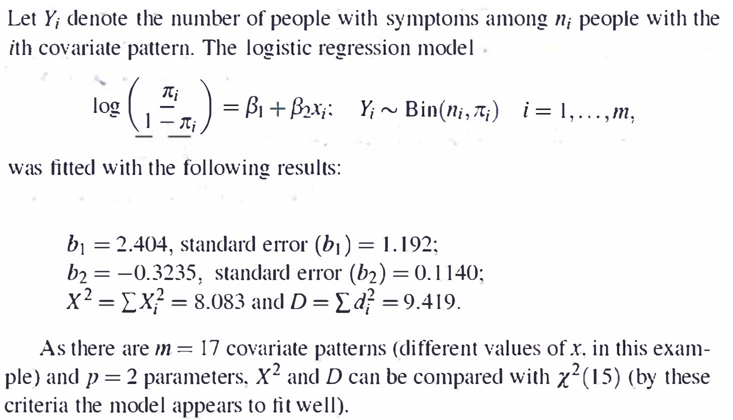


Covariate patterns and responses, estimated probabilities ($\hat{\pi}$), Pearson residuals (X) and deviance residuals (d) for senility and WAIS.

x	y	n	$\hat{\pi}$	X	d
4	1	2	0.752	−0.826	−0.766
5	1	1	0.687	0.675	0.866
6	1	2	0.614	−0.330	−0.326
7	2	3	0.535	0.458	0.464
8	2	2	0.454	1.551	1.777
9	2	6	0.376	−0.214	−0.216
10	1	6	0.303	−0.728	−0.771
11	1	6	0.240	−0.419	−0.436
12	0	2	0.186	−0.675	−0.906
13	1	6	0.142	0.176	0.172
14	2	7	0.107	1.535	1.306
15	0	3	0.080	−0.509	−0.705
16	0	4	0.059	−0.500	−0.696
17	0	1	0.043	−0.213	−0.297
18	0	1	0.032	−0.181	−0.254
19	0	1	0.023	−0.154	−0.216
20	0	1	0.017	−0.131	−0.184
Sum	14	54			
			Sum of squares	8.084*	9.418*

* Sums of squares differ slightly from the goodness of fit statistics X^2 and D mentioned in the text due to rounding errors.

Example: Senility and WAIS



Let Y_i denote the number of people with symptoms among n_i people with the ith covariate pattern. The logistic regression model

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_1 + \beta_2 x_i; \quad Y_i \sim \text{Bin}(n_i, \pi_i) \quad i = 1, \ldots, m,$$

was fitted with the following results:

$$b_1 = 2.404, \text{ standard error } (b_1) = 1.192;$$
$$b_2 = -0.3235, \text{ standard error } (b_2) = 0.1140;$$
$$X^2 = \sum X_i^2 = 8.083 \text{ and } D = \sum d_i^2 = 9.419.$$

As there are $m = 17$ covariate patterns (different values of x, in this example) and $p = 2$ parameters, X^2 and D can be compared with $\chi^2(15)$ (by these criteria the model appears to fit well).

Example: Senility and WAIS

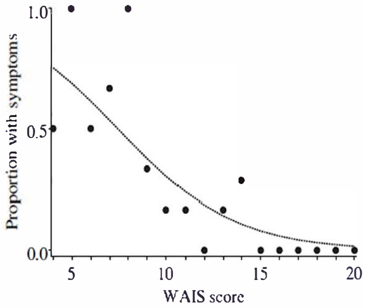


Figure: *dots represent observed proportions and the dotted line repre-sents estimated probabilities.*

Example: Senility and WAIS

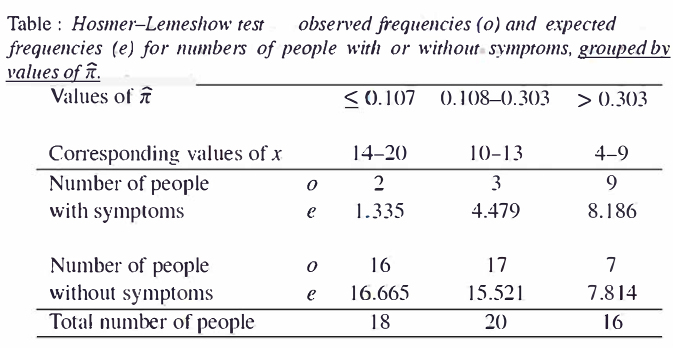


Table : *Hosmer–Lemeshow test observed frequencies (o) and expected frequencies (e) for numbers of people with or without symptoms, grouped by values of $\hat{\pi}$.*

Values of $\hat{\pi}$		≤ 0.107	0.108–0.303	> 0.303
Corresponding values of x		14–20	10–13	4–9
Number of people with symptoms	o	2	3	9
	e	1.335	4.479	8.186
Number of people without symptoms	o	16	17	7
	e	16.665	15.521	7.814
Total number of people		18	20	16