Probability and Statistics 2 (GCED)
Models for Binary response

Marta Pérez-Casany and Jordi Valero Baya

Department of Statistics and Operations Research
Technicat University of Catalonia

Facultat d'Informàtica de Barcelona, First Semester

Example 1

One is interested in comparing different doses of an insecticide, with respect to the mortality of a given insect. One has m different groups of n_i insects each one. To each group a different dose is administrated, denoted by x_i. One observes the total mortality Y_i produced by the i-thm dose.

x_i	n_i	y_i
0.75	90	0
1.5	80	2
3	90	4
6	60	13
7.5	85	27

x_i	n_i	y_i
10	60	32
15	90	55
20	60	44
50	50	47
100	40	38

Example 1

Experimental units: insects

Variables:

- Y number of deaths for a given dose (response variable)
- X insecticide dose level (explanatory variable)

Y is a **discrete** variable and X is a ontinuous variable.

Experimental conditions: Each one of the insecticide dose considered.

Example 1

We want to know:

- Do it exists differences between mortatily levels due to the different doses?
- Does it exists a particular recomended dose for a particular level of mortality?

The model

$$g_1(p_i) = g_2(\mu_i) = \beta_0 + \beta_1 x_i, \;\; i = 1, \cdots m.$$

where p_i is the probability of death receiving a dose equal to x_i.

Observe that the model is **defined in terms of the expectation**, that's why it doesn't appear an error term.

Example 2

Current Use of contraception Among Married women by Age, Education and Desire for More children. Fiji fertility survey, 1975

Experimental units: Women

Variables:

- Y Contraceptive Use (Yes, No) (response variable) **Binary variable**
- X_1 Age (explanatory variable) categorical with four levels.
- X_2 Education level (explanatory variable) categorical with two levels.
- X_3 Desires more children? (explanatory variable) categorical with two levels.

Y is a **discrete** variable. more precisely it is a Binary variable

Experimental conditions: Each one of the possible combinations of the three explanatory variables. We have a total of 16 different experimental conditions.

Example 2

		Desires More	Contraceptive use		
Age	Education	children?	Yes	No	Total
¡ 25	Lower	Yes	53	6	59
		No	10	4	14
	Upper	Yes	212	52	264
		No	50	10	60
25-29	Lower	Yes	60	14	74
		No	19	10	29
	Upper	Yes	155	54	209
		No	65	27	92
30-39	Lower	Yes	112	33	145
		No	77	80	157
	Upper	Yes	118	46	164
		No	68	78	146
40-49	Lower	Yes	35	6	41
		No	46	48	94
	Upper	Yes	8	8	16
		No	112	31	43

Example 2

Some question to answer:

- Does the Age have any influence in the use of contraceptive ?
- Does the Education level have any influence in the use of contraceptive?
- Does the desire of more children have any influence in the use of contraceptive?
- Has the Education level the same influence in the contraceptive use in all the ages?
- Has the Age the same influence in the use of contraceptive independently if the woman desires more children or not?

The model

$$g_1(p_i) = g_2(\mu_i) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 X_3, \;\; i = 1, \cdots m.$$

where p_i is the probability of death receiving a dose equal to x_i.

Bernouïlli and Binomial distributions

A r.v. $Y \sim \mathrm{B}(p)$ (*Bernouilli*), $0 \leq p \leq 1$ if, and only if, takes only values 0 y 1 with probabilities:

$$\Pr\{Y = 1\} = p \ y \ \Pr\{Y = 0\} = 1 - p.$$

A r.v. $Y \sim \mathrm{Bin(n,p)}$ (*Binomial*) with parameter $n \in \mathbb{N}$ and $0 \leq p \leq 1$, if, and only if, takes bvalues in $\{0, 1, 2, \cdots, n\}$ with probabilities:

$$\Pr\{Y = k\} = \binom{n}{k} p^k (1 - p)^{n-k}, \quad \forall k \in \{0, 1, \cdots, n\}.$$

In the later case:

$$E(Y) = n\,p \text{ y } Var(Y) = n\,p\,(1 - p).$$

If y is a realization of Y, $\hat{p} = y/n$.

It is defined the **ODDS** of a Binomial r.v. as $ODDS = \frac{p}{1-p} \in (0, +\infty)$, and it verifies:

$$ODDS \begin{array}{ll} = 1 & \text{si } p = 1/2 \\ > 1 & \text{si } p > 1/2 \\ < 1 & \text{si } p < 1/2 \end{array}$$

If Y is measured in two different populations, it is defined the **ODDS Ratio** as:

$$OR = \frac{\frac{p_1}{1-p_1}}{\frac{p_2}{1-p_2}} = \frac{p_1\,(1-p_2)}{p_2\,(1-p_1)} \in (0, +\infty)$$

$$OR \begin{array}{ll} = 1 & \text{si } p_1 = p_2 \\ > 1 & \text{si } p_1 > p_2 \\ < 1 & \text{si } p_1 < p_2 \end{array}$$

	$Y=1$	$Y=0$	
A	a	b	n_1
B	c	d	n_2

Given that $\hat{p}_1 = \frac{a}{n_1}$ y $\hat{p}_2 = \frac{c}{n_2}$ one has that:

$$\hat{OR} = \frac{ad}{cb},$$

that's why it is called (*cross-product ratio*).

Binary response and covariates

Question: Why it has no sense to consider:

$$E(Y_i) = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{ip-1}\beta_{p-1}$$

when $Y_i \sim Bin(n_i, p_i)$?

and it neither has sense to consider: $p_i = E(Y_i/n_i)$?

Three important reasons:

1) $Var(\frac{Y_i}{n_i}) = \frac{p_i\,(1-p_i)}{n_i}$,
2) we do not have normality,
3) $(X\beta)_i \in \mathbb{R}$ while $p_i \in (0,1)$.

Possible link functions

Observation: It has no sense to think that p, i. e. the mean of Y/m, is linear in the covariates, given that it takes values in the interval $(0, 1)$ and $X\beta$ takes values in the real line.

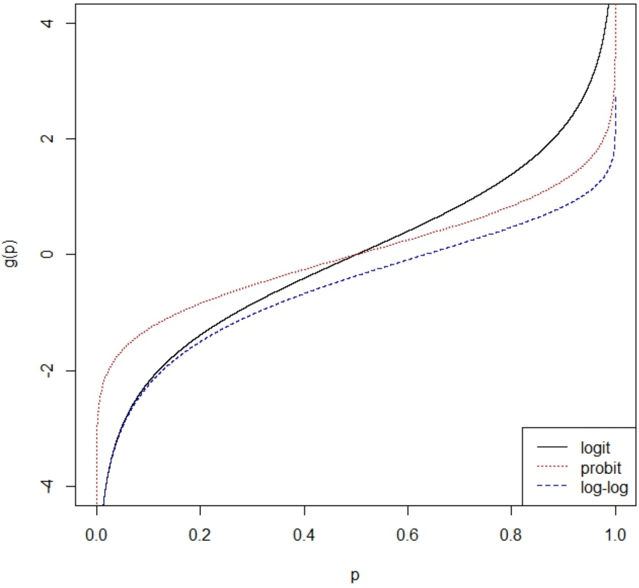
The following are functions of p that have sense to be linear in the covariates

- Función **logit**: $\mathbf{g_1(p) = \log(p/(1-p))}$;
- Función **probit**: $\mathbf{g_2(p) = \Phi^{-1}(p)}$, donde Φ es la función de distribución de la Normal tipificada;
- Función **complementario log-log**: $\mathbf{g_3(p) = \log(-\log(1-p))}$ o $\mathbf{\log(-\log(p))}$

All of them go from $(0, 1)$ to the entire real line.

The model:

$$\mathbf{g(p) = X}\beta$$


g(p)
4
2
0
-2
-4
0.0
0.2
0.4
0.6
0.8
1.0
p
logit
probit
log-log

To take into account:

1) Probit y logit are simetrical with respect to $p = 1/2$ and it is not the c-log-log.
2) Logit and c-log-log are very difficult to distinguish for p values near zero.
3) Parameter interpretation in the logistic case. Given that

$$\log(p_i/(1-p_i)) = \beta_0 + \beta_1\, d_i, \tag{1}$$

β_0 is the logit value when $d_i = 0$ (*baseline*).

Moreover, if p_{i+1} is the probability associated to a dose equal to d_{i+1}, one has that:

$$\log(p_{i+1}/(1-p_{i+1})) - \log(p_i/(1-p_i)) = \beta_0 + \beta_1\,(d_i+1) - \beta_0 - \beta_1\, d_i = \beta_1$$

from where $\beta_1 = \log(OR)$.

Observe that (1) is equivalent to:

$$p_i = \frac{e^{\beta_0+\beta_1\, d_i}}{1+e^{\beta_0+\beta_1\, d_i}},$$

4) If succes is changed by failure, what happens with the parameters of the logistic model?

$$\log\left(\frac{1-p}{p}\right) = \log\left(\frac{p}{1-p}\right)^{-1} = -\log\left(\frac{p}{1-p}\right) = -\beta_0 - \beta_1\, d_i$$

The same model keeps being good.

5) The logit makes easier the parameter interpretation.

Parameter vector Estimation

Given that $Y_i \sim \mathrm{Bin}(n_i, p_i)$ and assuming that

$$g(p_i) = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots x_{ip-1}\beta_{p-1} \quad i = 1, \cdots, n$$

One has that:

the likelihood function is equal to:

$$L(\beta; y) = \Pi_{i=1}^{m} \binom{n_i}{y_i} \left(g^{-1}(\sum_{j=0}^{p-1} x_{ij}\beta_j) \right)^{y_i} (1 - g^{-1}(\sum_{j=0}^{p-1} x_{ij}\beta_j))^{n_i - y_i};$$

and the log-likelihood equal to:

$$l(\beta; y) = \sum_{i=1}^{m} \{ y_i \log \left(g^{-1}(\sum_{j=0}^{p-1} x_{ij}\beta_j) \right) + (n_i - y_i) \log(1 - g^{-1}(\sum_{j=0}^{p-1} x_{ij}\beta_j)) \}.$$

In the particular case of the logistic model,

$$l(p;y) = \sum_{i=1}^{m} y_i \log\left(\frac{p_i}{1-p_i}\right) + \sum_{i=1}^{m} n_i \log(1-p_i)$$

from where

$$l(\beta;y) = \sum_{i=1}^{m} y_i \left(\sum_{j=0}^{p-1} x_{ij}\beta_j\right) - \sum_{i=1}^{m} n_i \log\left(1 + e^{\sum_{j=0}^{p-1} x_{ij}\beta_j}\right).$$

Thus,

$\frac{\partial l}{\partial \beta} = 0 \iff X^t(Y-\mu) = 0$; and the mle is equivalent to the moment estimator applied to X^tY.

Observation: X^ty is a minimal and sufficient estatistic for β.

Analysis of a 2 x 2 contingence table:

	$Y=1$	$Y=0$	
A	a	b	n_1
B	c	d	n_2

may be perform assuming a logistic regresion of the form:

$$\begin{pmatrix} \log\left(\frac{p_1}{1-p_1}\right) \\ \log\left(\frac{p_2}{1-p_2}\right) \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$$

Given that $\log(OR) = \beta_2$,

$$p_1 = p_2 \iff \beta_2 = 0 \iff OR = 1$$

Goodness of fit I

Pearson χ^2-square statistic

$$X^2 = \sum_{i=1}^{N} \frac{(o_i - e_i)^2}{e_i^2} = \sum_{i=1}^{n} \frac{(y_i - n_i\hat{p}_i)^2}{n_i\hat{p}_i(1-\hat{p}_i)} = \sum_{i=1}^{N} r_i^2$$

If the model is correct, X^2 assymtotically follows a χ^2_{N-p}.

Thus, we can reject our model when $X^2 \geq \chi^2_{\alpha,N-p}$.

The values signed r_i are called Pearson residuals and when plotted they should follow approximately a standarized Normal distribution.

Goodness of fit II

Deviance

It is defined as $D = 2[l(\hat{p}_{i,fullm}; y) - l(\hat{p}_{i,ourm}, y)]$

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

Obs: if for some i $y_i = 0$ or $y_i = n_i$ then the corresponding term in D is taken to be equal to zero.

Under the hypothesis that our model is correct, $D \sim \chi^2_{N-p}$, and we reject our model then $D \geq \chi^2_{\alpha,N-p}$.

The values signed d_i are known as deviance residuals and asymptotically follow a standarized normal distribution.

Goodness of fit III

DEFINITION:

Given two models (mod1, mod2), it is said mod1 is **nested** in mod2 if, and only if, mod2 contains all the parameters in mod1 and some more.

Denoting by p_i the number of parameters of mod$i$, and by D_i its corresponding scaled deviance,

to compare

$$H_0 : mod1 \quad vs \quad H_1 : mod2$$

one has that under H_0, assimptotically

$$D_1 - D_2 \sim \chi^2_{p_2 - p_1}$$

and we reject H_0 when $D_1 - D_2 \geq \chi^2_{\alpha, p_2 - p_1}$