

# GENERALIZED LINEAR MODELS: Logistic Regression



Γιώτα Τουλούμη

Καθηγήτρια Βιοστατιστικής και Επιδημιολογίας  
Εργ. Υγιεινής, Επιδημιολογίας και Ιατρικής Στατιστικής  
Ιατρική Σχολή Πανεπιστημίου Αθήνας

gtouloum@med.uoa.gr

# Notes

$$E[\ln(Y) | \mathbf{X}] = b_0 + b_1 X_1 + b_2 X_2 + \dots + b_p X_p$$

---

$$\ln(Y_1) = b_0 + b_1 \text{age} \quad (1)$$

$$\ln(Y_2) = b_0 + b_1 (\text{age} + 1) = b_0 + b_1 \text{age} + b_1 \quad (2)$$

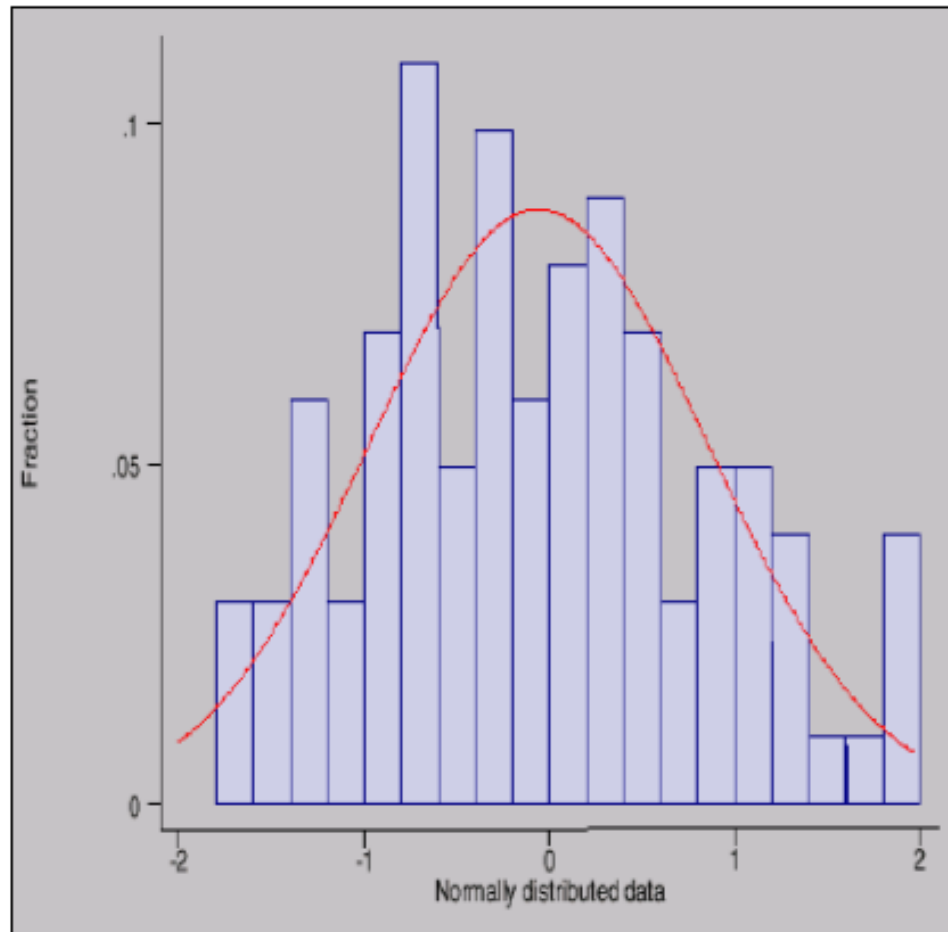
$$(2) - (1) \Rightarrow \ln(Y_2) - \ln(Y_1) = b_1 \Rightarrow \ln\left(\frac{Y_2}{Y_1}\right) \stackrel{E[\ln(Y) | \mathbf{X}] = b_0 + b_1 \text{age}}{=} b_1 \Rightarrow \frac{Y_2}{Y_1} = e^{b_1}$$

$$b_1 = 0.0161012 \quad e^{0.0161012} = 1.0162$$

Όταν η ηλικία αυξηθεί κατά ένα έτος τα επίπεδα ρετινόλης αναμένεται να αυξηθούν κατά μέσο 16% (ανεξάρτητα από τις υπόλοιπες μεταβλητές του μοντέλου)

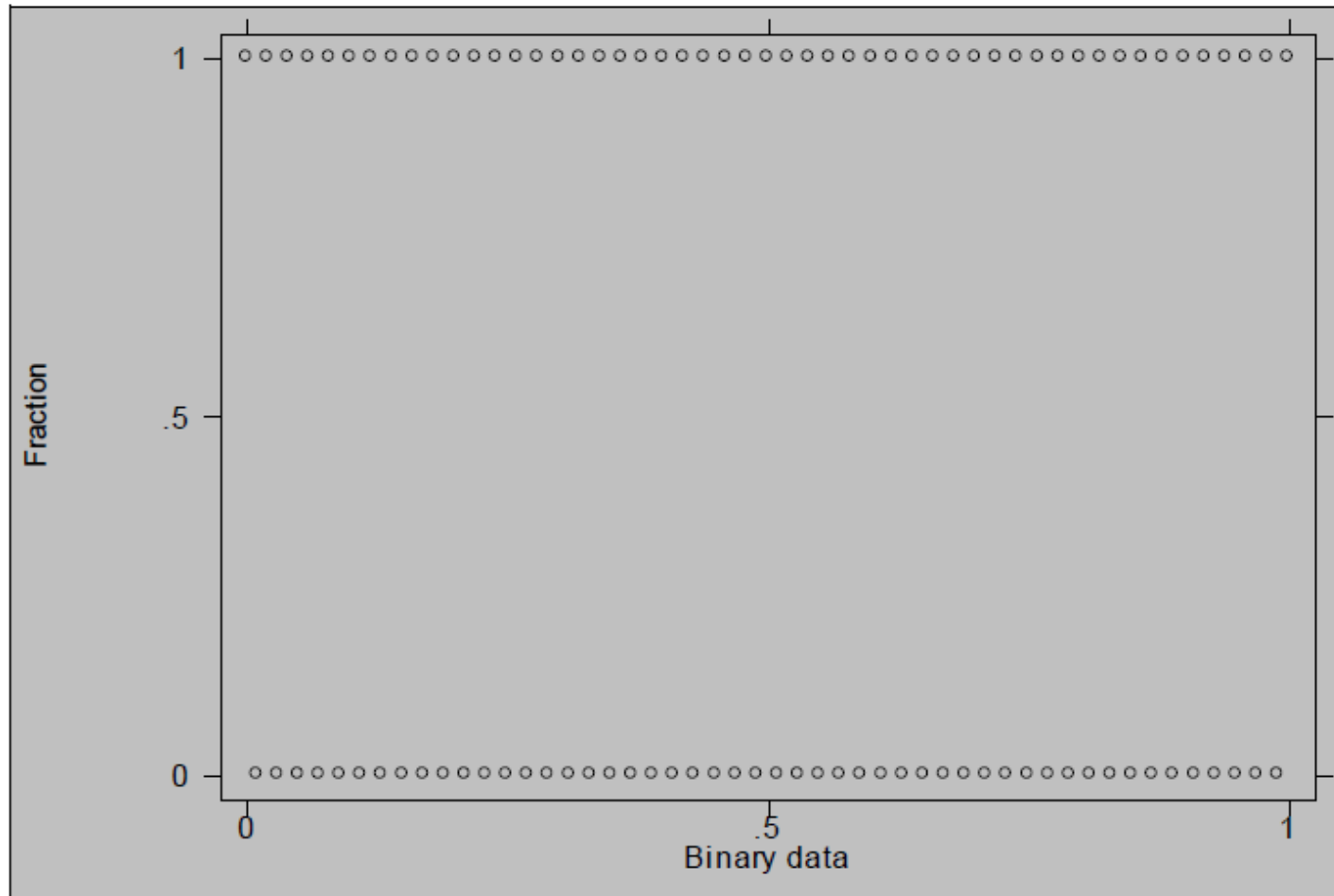
# General linear Model

This model deals with data that look as follows:



# Models for binary data

Dichotomous (zero/one or binary) data look like this:



## Binary data – covariate classes

Suppose that for each individual or experimental unit, the response,  $Y$ , can take only one of the two possible values 0,1. We can write

$$\Pr(Y_i=0)=1-\pi_i ; \quad \Pr(Y_i=1)= \pi_i$$

For the probability of “failure” and “success” respectively.

In most observational studies we have, associated with each individual, a vector of covariates or explanatory variables  $(x_1, x_2, \dots, x_p)$ . The vector of covariates consists of measure variables thought likely to influence the probability of positive response. The principal objective of the a statistical analysis, therefore, is to investigate the relationship between the response probability  $\pi=\pi(\mathbf{x})$  and the explanatory variables  $\mathbf{x}=(x_1, x_2, \dots, x_p)$ . Suppose that of the  $N=m_1+m_2+\dots+m_n$  individuals under study  $m_i$  share the covariate vector  $(x_{i1}, x_{i2}, \dots, x_{ip})$ . These individuals are said to form a **covariate class**.

## Alternative ways of presenting the same data

### a) Data listed by subject ID

<u>Subject No</u>	<u>Covariate (<math>x_1, x_2</math>)</u>	<u>Response Y</u>
1	1,1	0
2	1,2	1
3	1,2	0
4	2,1	0
5	2,2	1
6	1,2	1
7	1,1	1

### b) Data listed by covariate class

<u>Covariate (<math>x_1, x_2</math>)</u>	<u>Class size m</u>	<u>Response Y</u>
1,1	2	1
1,2	3	2
2,1	1	0
2,2	1	1

---

The difference in **a)** and **b)** is that in **b)** some information is lost.

## The contraceptive-use data (Little, 1998, Rodriguez, 2000)

This is data from the Fiji fertility survey (1975) derived from Little (1998) and presented in Rodriguez (2000).

Age	Education	Desires more Children?	Contraceptive use		Total
			No	Yes	
<25	Lower	Yes	53	6	59
<25		No	10	4	14
<25	Higher	Yes	212	52	264
<25		No	50	10	60
25-29	Lower	Yes	60	14	74
25-29		No	19	10	29
25-29	Higher	Yes	155	54	209
25-29		No	65	27	92
30-39	Lower	Yes	112	33	145
30-39		No	77	80	157
30-39	Higher	Yes	118	46	164
30-39		No	68	78	146
40-49	Lower	Yes	35	6	41
40-49		No	46	48	94
40-49	Higher	Yes	8	8	16
40-49		No	12	31	43
Total			1100	507	1607

## Individual versus grouped data

In the example above we have data on 1607 women. Each woman uses ( $Y_i=1$ ) or does not use ( $Y_i=0$ ) contraceptives. The resulting *Bernoulli* probability of contraceptive use *for each woman* is:

$$P(Y_i = y_i) = \pi_i^{y_i} (1 - \pi_i)^{1 - y_i}$$

and the log-likelihood is

$$l(\boldsymbol{\pi}; \mathbf{Y}) = \sum_{i=1}^n \left\{ y_i \log(\pi_i) + (1 - y_i) \log(1 - \pi_i) \right\}$$

Consider the case of the 16 different groups of women according to the explanatory variables above. The *counts*  $y_m$  of group members in each group have a *Binomial* distribution, i.e.,

$$P(Y_m = y_m) = \binom{n_m}{y_m} \pi_m^{y_m} (1 - \pi_m)^{n_m - y_m}$$

with associated likelihood

$$l(\boldsymbol{\pi}; \mathbf{Y}) = \sum_{m=1}^k \left\{ y_m \log(\pi_m) + (n_m - y_m) \log(1 - \pi_m) \right\} + C$$

where  $k$  is the number of categories ( $k=16$  in this example) and  $C = \log \binom{n_m}{y_m}$ .



## Individual versus grouped data

When binary data are grouped by covariate class, the responses have the form  $y_1/m_1, \dots, y_n/m_n$ , where  $0 \leq y_i \leq m_i$  is the number of successes out of  $i$ th covariate class. The vector of covariate class sizes  $\mathbf{m} = (m_1, m_2, \dots, m_n)$  is called the binomial index vector or **binomial denominator** vector.

Ungrouped data, or data listed by individual subjects, can be considered as a special case for which  $m_1 = m_2 = \dots = m_n = 1$ .

The distinction between grouped and ungrouped data is important for at least two reasons:

- Some methods of analysis appropriate to grouped data, particularly those involving Normal approximation, are not applicable to ungrouped data
- Asymptotic approximations for models applied to grouped data can be based on either of two distinct asymptots, either  $\mathbf{m} \rightarrow \infty$  or  $N \rightarrow \infty$ . Only the latter limit is appropriate for ungrouped data.

## Models for binary data

We suppose that the dependence of  $\pi$  on  $(x_1, x_2, \dots, x_p)$  occurs through the linear combination :

$$\eta = \sum_{j=1}^p x_j \beta_j$$

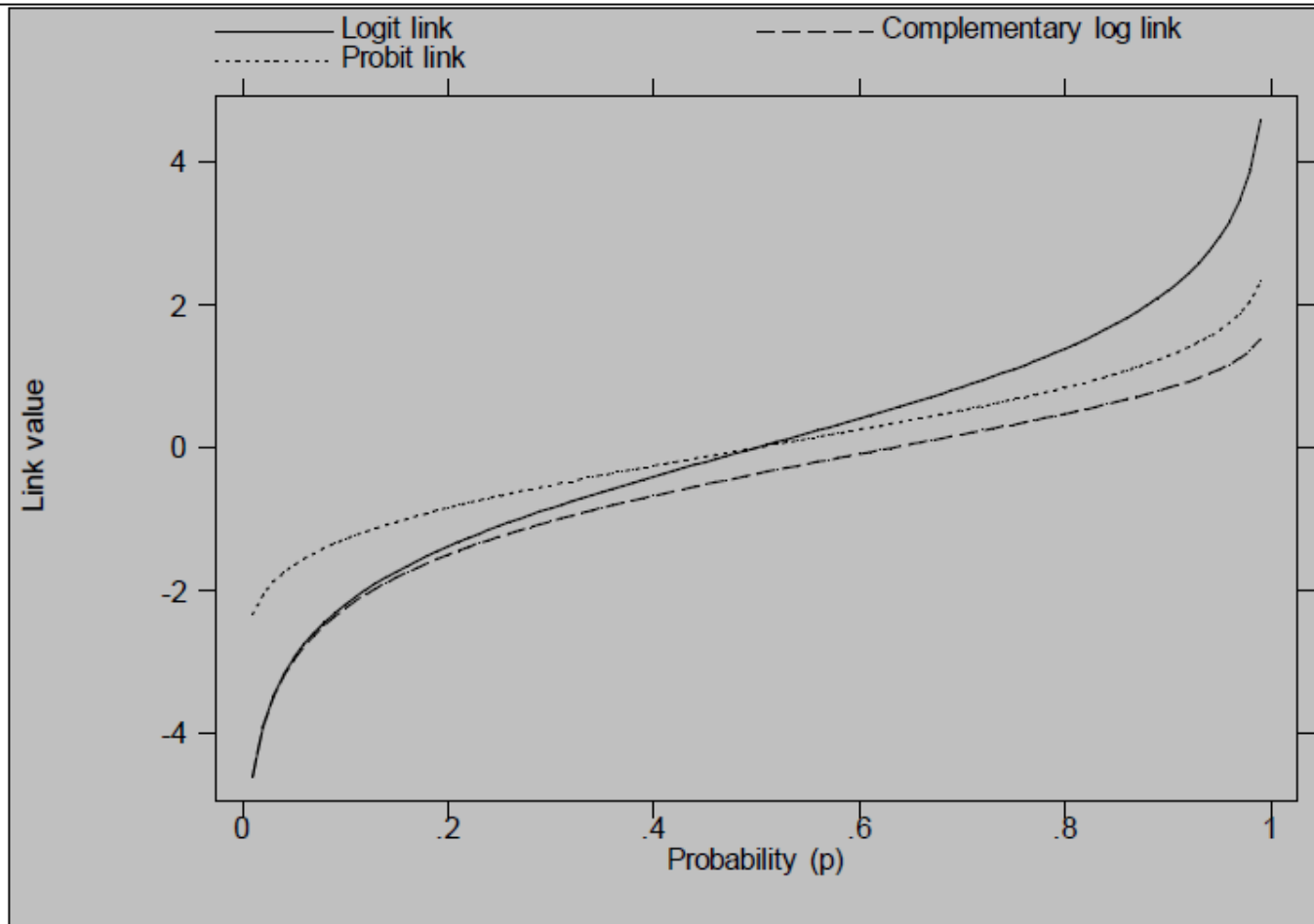
for unknown coefficients  $\beta_1, \beta_2, \dots, \beta_p$ . Unless restrictions are imposed on  $\beta$  we have  $-\infty < \eta < \infty$ . Thus to express  $\pi$  the linear combination would be inconsistent with the laws of probability. So, we need to use a transformation  $g(\pi)$  that maps the interval  $(0,1)$  to the whole line  $(-\infty, \infty)$ . This leads to GLM in which the systematic part is:

$$g(\pi_i) = \eta_i = \sum_{j=1}^p x_{ij} \beta_j \quad i = 1, 2, \dots, n$$

The most common link functions are:

- The **logit** or **logistic** function  $g_1(\pi) = \log\{\pi/(1-\pi)\}$
- The **probit** or **inverse Normal** function  $g_2(\pi) = \Phi^{-1}(\pi)$
- The **complementary log-log** function  $g_3(\pi) = \log\{-\log(1-\pi)\}$

## The logistic, probit and c-log-log links



The logistic regression and probit functions are almost linearly related over the interval 0.1 to 0.9. Difficult to discriminate between the two. For small values of  $\pi$  c-log-log approaches the logistic. As  $\pi$  approaches 1, the c-log-log approaches infinity much more slowly than the other two functions.



## Use of the different link functions

**C-log-log function** is used in *Dilution assays* to help the estimation of the number of infective organisms per unit volume. For more details refer to McCullagh and Nelder, *Generalized Linear Models*, pages: 11-12.

### **Probit function**

In toxicology experiments test animals or insects are divided into sets, usually but not necessarily, of equal sizes. Each set of animal is subjected to a known level  $x$  of a toxin. The dose varies from set to set but is assumed to be uniform within each set. For the  $j$ th set, the number  $y_j$  surviving out of the original  $m_j$  is recorded, together with the dose  $x_j$  administered. It is required to model the proportion surviving  $\pi_x$ , at dose  $x$  as a function of  $x$ . The probit model is  $\pi_x = \Phi(\alpha + \beta x)$

where  $\Phi(\cdot)$  is the cumulative Normal distribution. Note that if  $\beta > 0$ , the surviving probability is monotonely increasing in the applied dose. Otherwise, if  $\beta < 0$ , the survival probability is monotonely decreasing in the dose.

## Logistic regression for grouped data

All asymptotic and approximate theory presented here applies regardless of the choice of the link function. However, we will be concerned with the logistic function mainly because of its simple interpretation as the logarithm of the **odds ratio**.

*REMEMBER:* The **logit** is the **canonical** link. Therefore, the log likelihood depends on  $y$  only through the linear combinations  $\mathbf{X}'\mathbf{y}$ . These  $p$  combinations are sufficient for  $\beta$ .

### A. Grouped data

Consider the data of the potency of an insecticide (Finney, DJ 1971):

Dose	Number exposed	Deaths
0	49	0
2.6	50	6
3.8	48	16
5.1	46	24
7.7	49	42
10.2	50	44

Here we have grouped data: for an  $i$  dose we have  $y_i$  deaths out of  $m_i$  exposed individuals (Binomial denominator)

## Logistic regression for grouped data

All asymptotic and approximate theory presented here applies regardless of the choice of the link function. However, we will be concerned with the logistic function mainly because of its simple interpretation as the logarithm of the **odds ratio**.

*REMEMBER:* The **logit** is the **canonical** link. Therefore, the log likelihood depends on  $y$  only through the linear combinations  $\mathbf{X}'\mathbf{y}$ . These  $p$  combinations are sufficient for  $\beta$ .

### A. Grouped data

Consider the data of the potency of an insecticide (Finney, DJ 1971):

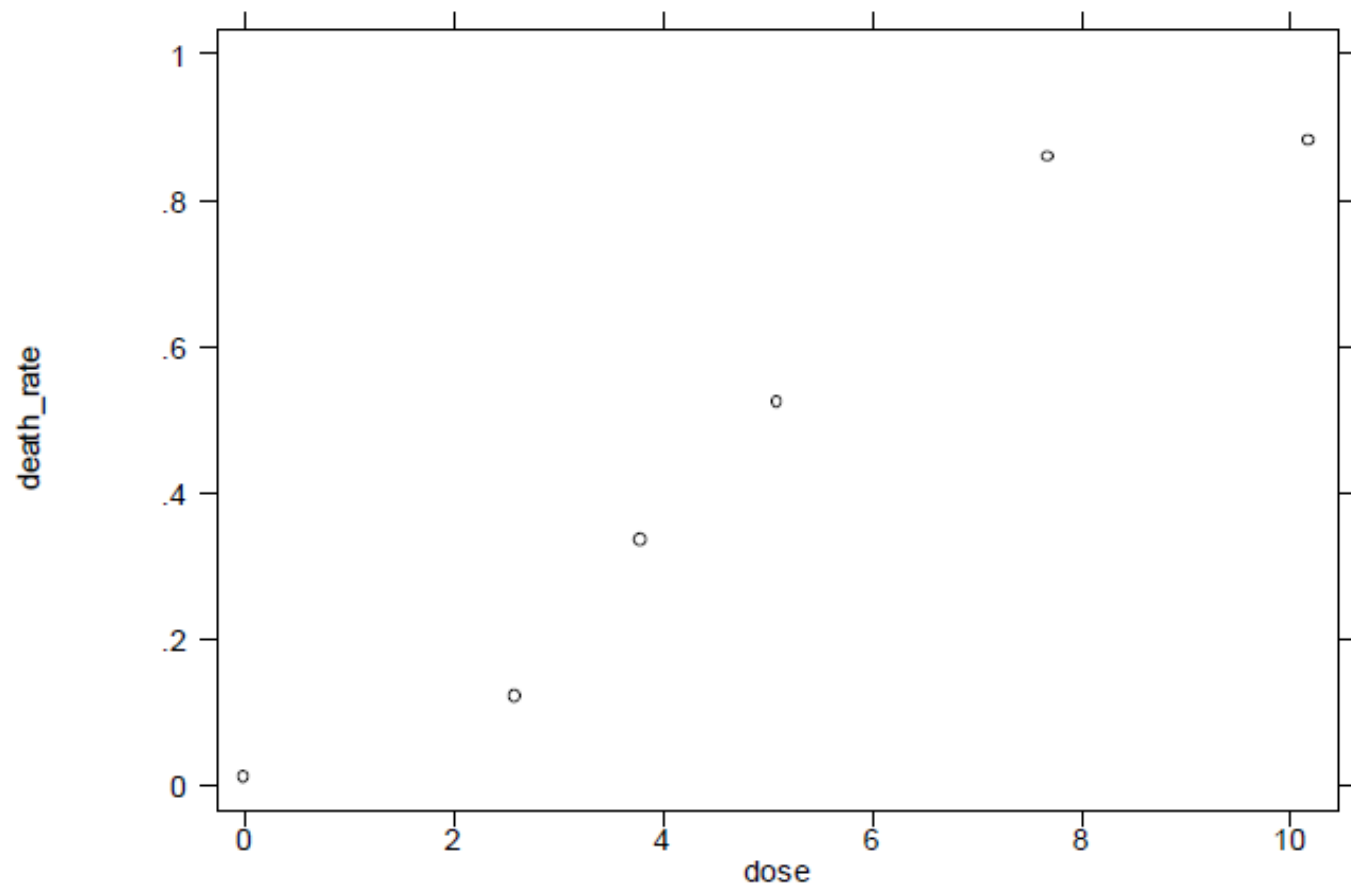
Dose	Number exposed	Deaths	D/Exp	$\pi_i$
0	49	0	0/49	0
2.6	50	6	6/50	0.12
3.8	48	16	16/48	0.33
5.1	46	24		
7.7	49	42		
10.2	50	44		

---

Here we have grouped data: for an  $i$  dose we have  $y_i$  deaths out of  $m_i$  exposed individuals (Binomial denominator)

$p=y/m$  (proportion dying; for the zero count, replace it with  $0.5/m_i$ )

Graph proportions versus dose



It appears to follow some sigmoid curve.

Analyse the data with glm command:

**. glm deaths dose, family (binomial noexp) link (logit)**

Iteration 0: log likelihood = -15.088325

Iteration 1: log likelihood = -14.739768

Iteration 2: log likelihood = -14.739357

Iteration 3: log likelihood = -14.739357

Generalized linear models

No. of obs = 6

Optimization : ML: Newton-Raphson

Residual df = 4

Scale param = 1

Deviance = 10.25831651

(1/df) Deviance = 2.564579

Pearson = 9.697151289

(1/df) Pearson = 2.424288

Variance function:  $V(u) = u*(1-u/noexp)$

[Binomial]  $N\pi(1-\pi)=Y[1-(Y/N)]$

Link function :  $g(u) = \ln(u/(noexp-u))$

[Logit]  $\pi/1-\pi=Y/N-Y (\pi=Y/N)$

Standard errors : OIM

Log likelihood = -14.73935665

AIC = 5.579786

BIC = 6.67479757



---

deaths	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
dose	.6051256	.0678129	8.92	0.000	.4722148	.7380363
_cons	-3.225663	.3699181	-8.72	0.000	-3.950689	-2.500637

---

**Note:** For grouped data, the response variable is the number of failures (deaths) and in the family we have also to define the binomial denominator (noexp: no exposed).

**Interpretation:**

$$\text{logit}(x+1) - \text{logit}(x) = b \Rightarrow \log(\text{odds}_{x+1}) - \log(\text{odds}_x) = b \Rightarrow \log\left(\frac{\text{odds}_{x+1}}{\text{odds}_x}\right) = b \Rightarrow$$

$$\frac{\frac{\hat{\pi}_{x+1}}{1 - \hat{\pi}_{x+1}}}{\frac{\hat{\pi}_x}{1 - \hat{\pi}_x}} = \frac{\text{odds}_{x+1}}{\text{odds}_x} = \text{odds ratio} = e^b = 1.831$$

*The odds of dying is increased by 83% per each unit increase of dose.*

$$e^{-3.23} = 0.039 \approx 0$$

Analyse data with *blogit*:

*blogit deaths noexp dose*

Logit estimates                  Number of obs =     292  
    LR chi2(1)       =   153.49  
    Prob > chi2      =   0.0000  
 Log likelihood = -124.31132      Pseudo R2       =   0.3817

-----  
 \_outcome |    Coef.   Std. Err.    z   P>|z|   [95% Conf. Interval]  
 -----+-----  
   dose |   .6051256   .0678099   8.92   0.000   .4722207   .7380304  
 \_cons | -3.225663   .3699052  -8.72   0.000  -3.950664  -2.500662  
 -----

The option `or` gives the odds ratio straightforward.

**Note:** Check results if alternatively the `glogit` (group logistic via weighted least squares) command is used.

*Check linearity assumption:*

**blogit deaths noexp dose dose2**

Logit estimates	Number of obs =	292
	LR chi2(2) =	162.67
	Prob > chi2 =	0.0000
Log likelihood = -119.71879	Pseudo R2 =	0.4045

```

-----
  _outcome |   Coef.  Std. Err.   z  P>|z|   [95% Conf. Interval]
-----+-----
    dose |  1.513806   .359191   4.21  0.000   .8098049   2.217808
   dose2 | -.0764208   .0277264  -2.76  0.006  -0.1307635  -0.0220782
    _cons | -5.466344   1.023386  -5.34  0.000  -7.472143  -3.460545
-----

```

The square of dose is significant ( $P=0.006$ ) according to Wald test. Alternatively can be checked by the likelihood ratio test (blogit) or the deviance test (glm).

## B. Individual data


Consider the example of the contraceptive use data

### Contingency tables

Consider the following table:

<b>Contraceptive use</b>	<b>Desires more children?</b>		<b>Total</b>
	<b>No</b>	<b>Yes</b>	
<b>Yes</b>	288	219	507
<b>No</b>	347	753	1100
<b>Total</b>	635	972	1607

This means that 972 out of the 1607 women desire more children, while 635 do not and 507 out of 1607 use contraception while 1100 do not. Furthermore, 288 women that do not desire any more children out of 635 use contraception, while 219 women that desire more children out of 972 use contraception out.



## Contingency tables (continued)

Now suppose that you wanted to determine whether there is any *association* between desire for children and contraceptive use rate. You can perform a the *Pearson chi-square* test. This is based on the  $\chi^2$  distribution which we will cover momentarily.

This test is set-up as follows:

1.  $H_0$ : Contraceptive use rate is not associated with desire for more children
2.  $H_a$ : There is an association between desire for more children and use of contraceptions


### Observed versus expected counts

Now let us consider the implication of the null hypothesis:

If the distinction between the two groups (women that desire more children versus those that do not) is an artificial one, then the rate of contraception use is estimated by  $\hat{p} = \frac{507}{1607} = 0.3155$ , the overall contraception-use rate.

Under this assumption:

- 1) The *expected* number of women that use contraceptives among those that desire more children is  $(0.3155)(972) = \mathbf{306.67}$  on average (versus the *observed* 219)
- 2) The *expected* number of women that do not use contraceptives among those that desire more children is  $(1-0.3155)(972) = \mathbf{665.34}$  (versus the observed 753)
- 3) The *expected* number of women that use contraceptives among those that desire no more children is  $(0.3155)(635) = \mathbf{200.34}$  (versus the *observed* 288)
- 4) The *expected* number of women that do not use contraceptives among those that desire no more children is  $(1-0.3155)(635) = \mathbf{434.66}$  (versus the *observed* 347).



## The Pearson chi-square test

Our test should be based on quantifying whether deviations from these two *expected* numbers are serious enough to warrant rejection of the null hypothesis.

In general, the chi square test looks like this:

$$\chi^2 = \sum_{i=1}^{rc} \frac{(O_i - E_i)^2}{E_i}$$

$E_i$  is the *expected* number,  $O_i$  is the *observed* number,  $r$  is the number of rows, and  $c$  is the number of columns. Then,  $\chi^2$  is distributed according to the chi square distribution with  $df=(r-1)(c-1)$  degrees of freedom. Critical percentiles of the chi-square distribution can be found in the appendix of your textbook.

### Example (continued)

Returning to the above example, the chi square test is (notice the continuity correction that improves the approximation to the chi-square distribution):

$$\begin{aligned}\chi^2 &= \sum_{i=1}^{rc} \frac{(|O_i - E_i| - 0.5)^2}{E_i} \\ &= \frac{(|219 - 306.67| - 0.5)^2}{306.67} + \frac{(|753 - 665.34| - 0.5)^2}{665.34} + \frac{(|288 - 200.34| - 0.5)^2}{200.34} + \frac{(347 - 434.66 - 0.5)^2}{434.66} \\ &= 24.78 + 11.42 + 38.36 + 17.48 \\ &= 92.04\end{aligned}$$

Comparing this value to 3.84, the right tail of the chi-square distribution with  $(2-1)(2-1)=1$  degree of freedom, the null hypothesis is strongly rejected.



## Computer implementation:

Carrying out the test using STATA is as follows:

```
. tabulate cuse more [freq=N], chi
```

Contraceptive use (Yes/No)	Desires more children?		Total
	No	Yes	
No	347	753	1100
Yes	288	219	507
Total	635	972	1607

Pearson chi2(1) = 92.6442 Pr = 0.000

Columns 1 and 2 and rows 1 and 2 correspond to “No” and “Yes” (Desires more children?) and (Desires no more children?) respectively. Since the p-value of the test is  $0.0000 < 0.05$ , we reject the null hypothesis. There is a strong association between desire for more children and use of contraceptives. Note that STATA calculates the Pearson  $\chi^2$  a bit differently (92.64 instead of 92.04 from our hand-calculations).



## The odds ratio

The chi square test of association answers only the question of association. It does not comment on the nature or direction of the association. For a further investigation of this hypothesis one must rely on a different test.

We define as the odds of having the disease if exposed is  $P(\text{disease}|\text{exposed})/[1-P(\text{disease}|\text{exposed})]$ .

The odds of having the disease if unexposed is  $P(\text{disease}|\text{unexposed})/[1-P(\text{disease}|\text{unexposed})]$ .

The *odds ratio* (**OR**) is defined as:


$$\text{OR} = \frac{P(\text{disease} | \text{exposed})/[1 - P(\text{disease} | \text{exposed})]}{P(\text{disease} | \text{unexposed})/[1 - P(\text{disease} | \text{unexposed})]}$$

Consider the following 2×2 table:

	<b>Exposed</b>	<b>Unexposed</b>	<b>Total</b>
<b>Disease</b>	a	b	a+b
<b>No disease</b>	c	d	c+d
<b>Total</b>	a+c	b+d	<i>n</i>

An estimate of the odds ratio is

$$\text{OR} = \frac{[a/(a+c)]/[c/(a+c)]}{[b/(b+d)]/[d/(b+d)]} = \frac{a/c}{b/d} = \frac{ad}{bc}$$



If the odds of having the disease in the exposed and unexposed groups are equal, then the odds ratio should be close to 1. A test of this is constructed as follows:

$H_0$ : There is no association between exposure and disease

$H_a$ : There *is* an association between exposure and disease.

If the null hypothesis is true, the *odds ratio* should be close to 1. The test will answer the question:

“How far from 1 is too far to warrant rejection of the null hypothesis?”

The OR itself is not distributed normally. But its logarithm is. In fact, the statistic

$Z = \frac{\ln(ad/bc)}{\sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}}$  is approximately distributed according to the standard normal distribution. Tests

and confidence intervals are derived as usual.

## Testing the hypothesis of no association (using the odds ratio)

1.  $H_0$ : There is no association between exposure and disease ( $OR=1$ )
- 2a.  $H_a$ : There *is* a positive or negative association between exposure and disease ( $OR>1$  or  $OR<1$  respectively)
- b.  $H_a$ : There *is* an association between exposure and disease ( $OR \neq 1$ )
3. The test statistic is  $Z$
- 4 **Rejection rule:**
  - a. Reject the null hypothesis, if  $Z \geq z_{0.05}$ , or  $Z \leq -z_{0.05}$  (one-sided tests)
  - b. Reject the null hypothesis, if  $|Z| > z_{0.025}$  (two-sided tests)

### Confidence intervals (Woolf method)

A  $(1-\alpha)\%$  confidence interval of the *log-odds ratio* is given by

$$\left[ \ln(OR) - z_{\alpha/2} \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}, \ln(OR) + z_{\alpha/2} \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}} \right]$$

Thus, the  $(1-\alpha)\%$  confidence interval of the true *odds ratio* is given by

$$\left[ e^{\ln(OR) - z_{\alpha/2} \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}}, e^{\ln(OR) + z_{\alpha/2} \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}} \right]$$

This confidence interval can also be used to perform a hypothesis test by inspecting whether it covers 1 (the OR hypothesized value under the null hypothesis).

**Example:** Consider the previous example:

```
. cci 219 288 753 347
```

	Exposed	Unexposed	Total	Proportion Exposed
Cases	219	288	507	0.4320
Controls	753	347	1100	0.6845
Total	972	635	1607	0.6049
	Point estimate		[95% Conf. Interval]	
Odds ratio	.3504178		.2804118	.4379295 (exact)
Prev. frac. ex.	.6495822		.5620705	.7195882 (exact)
Prev. frac. pop	.4446686			
chi2(1) =			92.64	Pr>chi2 = 0.0000

The odds ratio is 0.350 with a 95% confidence interval (0.280, 0.438). Thus, the null hypothesis of no association is rejected as both limits of the confidence interval are below 1.0.

$$\text{Prev. Frac.} = 1 - \text{RR} = 1 - 0.35 = 0.65$$

$$\text{Prev. Frac. Pop.} = \frac{P_C(1 - \text{RR})}{P_C(1 - \text{RR}) + \text{RR}} = \frac{0.432(1 - 0.35)}{0.432(1 - 0.35) + 0.35} = 0.445$$

$$P_C = \% \text{ of cases that are exposed} = \frac{219}{507}$$

## Analysis using logistic regression

Consider now the analysis of the same table using logistic regression. That is, fitting the model

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 X_i, \text{ where } i=1, \dots, n \text{ and } X_i = \begin{cases} 1 & \text{if woman desires children} \\ 0 & \text{if woman wants no more children} \end{cases}$$

The model probabilities in the 2x2 table are defined as follows:

Contraceptive use	Desires more children?	
	Yes ( $X=1$ )	No ( $X=0$ )
Yes ( $Y=1$ )	$\hat{\pi}(1) = \frac{e^{\beta_0 + \beta_1}}{1 + e^{\beta_0 + \beta_1}}$	$\hat{\pi}(0) = \frac{e^{\beta_0}}{1 + e^{\beta_0}}$
No ( $Y=0$ )	$1 - \hat{\pi}(1) = \frac{1}{1 + e^{\beta_0 + \beta_1}}$	$1 - \hat{\pi}(0) = \frac{1}{1 + e^{\beta_0}}$
Total	1.0	1.0

where,  $\hat{\pi}(x) = P(Y=1|X=x)$ . Note that in his notes Rodriguez (2000) uses the group of women that desire more children as the reference group (that results in estimates that have the opposite sign).



## Analysis using logistic regression (continued)

Using STATA this analysis looks as follows (note that we use “No use” as the reference cell):

```
. xi: logit cuse i.more [freq=N], nolog
i.more          Imore_0-1      (naturally coded; Imore_0 omitted)

Logit estimates                               Number of obs   =       1607
                                                LR chi2(1)      =       91.67
                                                Prob > chi2     =       0.0000
Log likelihood = -956.00957                    Pseudo R2      =       0.0458

-----+-----
      cuse |          Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-----+-----
  Imore_1 |   -1.048629     .110672    -9.475  0.000    -1.265542    -.831716
    _cons |   -.1863643    .0797124   -2.338  0.019    -.3425977    -.0301309
-----+-----

.lrtest, saving(1)
```

The estimated coefficients are  $\hat{\beta}_0 \approx -0.186$  (s.e. 0.080) and  $\hat{\beta}_1 \approx -1.049$  (s.e. 0.111). The likelihood-ratio test for this model is 91.67, which is distributed asymptotically according to a chi-square distribution with one degree of freedom. This is very close to the Pearson chi-square statistic presented above. The likelihood of this model is saved with the `lrtest` command.

## Interpretation of the logistic regression coefficients

From the above table we can see that the odds ratio is

$$\hat{\psi} = \frac{\hat{\pi}(1)/1 - \hat{\pi}(1)}{\hat{\pi}(0)/1 - \hat{\pi}(0)} = \frac{\frac{e^{\hat{\beta}_0 + \hat{\beta}_1}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1}}}{\frac{e^{\hat{\beta}_0}}{1 + e^{\hat{\beta}_0}}} \bigg/ \frac{\frac{1}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1}}}{\frac{1}{1 + e^{\hat{\beta}_0}}} = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1}}{e^{\hat{\beta}_0}} = e^{\hat{\beta}_1}$$

that is, the estimate  $\hat{\beta} = \log(\hat{\psi})$ .

Thus, the above estimate of  $\hat{\beta}$  results in an estimated odds ratio  $\hat{\psi} = e^{-1.049} = 0.350$ . This means that

women that desire more children (i.e.,  $X=1$ ) have 35% odds (or alternatively are  $\frac{1}{0.35} = 2.85$  times less

likely) to be using contraceptives compared to women that do not want any more children ( $X=0$ ). Notice

that this estimate is produced by multiplying the diagonals of the  $2 \times 2$  table, i.e.,  $\hat{\psi} = \frac{(219)(347)}{(288)(753)} \approx 0.350$

STATA produces estimates of the odds ratios in two identical ways: Either by including the option `or` after the `logit` statement, or by using the `logistic` command.

```
. logit , or
Logit estimates                               Number of obs   =       1607
                                                LR chi2(1)      =       91.67
                                                Prob > chi2     =       0.0000
Log likelihood = -956.00957                    Pseudo R2      =       0.0458

-----+-----
      cuse | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-----+-----
      more |   .3504178   .0387814   -9.475  0.000     .2820863   .4353017
```

```
. xi: logistic cuse i. more [freq=N]
i.more          Imore_0-1      (naturally coded; Imore_0 omitted)
Logit estimates                               Number of obs   =       1607
                                                LR chi2(1)      =       91.67
                                                Prob > chi2     =       0.0000
Log likelihood = -956.00957                    Pseudo R2      =       0.0458

-----+-----
      cuse | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-----+-----
      more |   .3504178   .0387814   -9.475  0.000     .2820863   .4353017
```

## The “null” model

Consider the following model:

```
. xi: logit cuse [freq=N], nolog
```

```
Logit estimates                               Number of obs   =       1607
                                                LR chi2(0)      =         0.00
                                                Prob > chi2     =         .
Log likelihood = -1001.8468                    Pseudo R2      =       0.0000
```

```
-----+-----
      cuse |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-----+-----
      _cons |  -.7745545   .0536794   -14.429   0.000   -1.8797641   -0.6693448
-----+-----
```

This is the “null” model  $\log\left(\frac{\pi_i}{1-\pi_i}\right)=\beta_o$ . The likelihood for this model is saved using the `lrtest` command. The estimate  $\hat{\beta}_o \approx -0.775$  results in the odds of using contraception (in general) in the study population, i.e.,  $e^{\hat{\beta}_o} \approx e^{-0.775} = 0.461 = 507/1100$ .

## Hypothesis testing

- a. The likelihood-ratio statistic compares the model with `more` as the covariate against the null model. Its value is 91.67, which compared to a chi-square distribution with one degree of freedom is extremely significant.
- b. The test of significance of the estimate  $\hat{\beta}$  is a Wald test of the form  $z = \frac{\hat{\beta}}{\text{s.e.}(\hat{\beta})} = -9.475$ . The asymptotic distribution of this statistic is standard normal.
- c. The 95% confidence interval is  $\hat{\beta} \pm (1.96)\text{s.e.}(\hat{\beta}) = [-1.265, -0.832]$ , with 95% confidence interval for the odds ratio  $[e^{-1.265}, e^{-0.832}] = [0.282, 0.435]$ . Women that desire children are 2.30 times ( $=1/0.435$ ) to 3.55 times ( $=1/0.282$ ) less likely to use contraceptive methods.

## Hypothesis testing (continued)

The  $z$  statistics above is consistent to the likelihood-ratio one presented above. In fact, the square of this statistic is 89.9, which is very close to the 91.67 likelihood ratio statistic. This (Wald) chi-square statistic can be obtained by the `test` command in STATA as follows:

```
. test Imore_1  
  
( 1)  Imore_1 = 0.0  
  
      chi2( 1) =    89.78  
      Prob > chi2 =    0.0000
```

This is very close to our calculations (within round-off error).

### The effect of age on use of contraception

The effect of the factor age can be ascertained in a similar manner, considering the following 2×4 table in this instance).

Contraceptive use	Age				Total
	<25	25-29	30-39	40-49	
Yes	72	105	237	93	507
No	325	299	375	101	1100
Total	397	404	612	194	1607

## The effect of age on use of contraception

Since age is a factor with four levels, an explicit factorization is necessary. Thus, we will construct three dummy variables  $X_1$ ,  $X_2$  and  $X_3$  as follows (here the group of women less than 25 years old is used as the reference group):

Dummy variable	Age factor			
	<25	25-29	30-39	40-49
$X_1$	0	1	0	0
$X_2$	0	0	1	0
$X_3$	0	0	0	1

The fitted model is as follows:

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$



## Analysis with STATA

The analysis with STATA is presented below. Notice that the first age group (<25) is the default reference group in STATA:

```
. xi: logit cuse i.age [freq=N] ,nolog
i.age          _Iage_1-4          (naturally coded; _Iage_1 omitted)

Logit estimates                               Number of obs   =       1607
                                                LR chi2(3)      =       79.19
                                                Prob > chi2     =       0.0000
Log likelihood = -962.25091                    Pseudo R2      =       0.0395

-----+-----
      cuse |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-----+-----
    _Iage_2 |   .4606758   .1727254     2.67   0.008     .1221403   .7992114
    _Iage_3 |   1.048293   .1544404     6.79   0.000     .7455955   1.350991
    _Iage_4 |   1.424638   .1939573     7.35   0.000     1.044489   1.804787
     _cons |  -1.507159   .1302527    -11.57  0.000    -1.76245  -1.251868
-----+-----

. lrtest, saving(2)
```

### Interpretation of estimated coefficients

The likelihood ratio test statistic is 79.19, which compared to a chi-square distribution with 3 degree of freedom is extremely significant. The model as fitted produces estimates of the odds ratios of each age group compared to the reference group (<25). For example, women 25-29 years old are, 58.5% ( $e^{\hat{\beta}} = e^{0.461} = 1.585$ ) more likely to be using contraceptives compared to women less than 25 years old.

This can be derived from the 2x2 table

```
. tab cuse age [freq=N] if age==1 | age==2
```

Contraceptive use (Yes/No)	Age		Total
	<25	25-29	
No	325	299	624
Yes	72	105	177
Total	397	404	801

and deriving the odds ratio as  $\hat{\psi} = \frac{(325)(105)}{(72)(299)} \approx 1.585$ .

## Interpretation of estimated coefficients (continued)

By similar arguments, 30-39 year-old women are 2.85 ( $= e^{1.048}$ ) times and 40-49 year-olds are 4.16 ( $= e^{1.425}$ ) times more likely to use contraceptive methods compared to women less than 25 years of age.

The individual Wald tests for the significance of the age-related coefficients  $\hat{\beta}_1, \hat{\beta}_2$  and  $\hat{\beta}_3$  are given in the STATA output above. A global test for the significance of all three is the Wald chi-square test:

```
. test Iage_2 Iage_3 Iage_4

( 1)  Iage_2 = 0.0
( 2)  Iage_3 = 0.0
( 3)  Iage_4 = 0.0

      chi2( 3) =    74.36
      Prob > chi2 =    0.0000
```

The value of this test is 74.4 and it is asymptotically distributed as a chi-square distribution with three degrees of freedom. The value of 74.4 is extremely significant as indicated by the p value.