Using logistic regression in perinatal epidemiology: an introduction for clinical researchers. Part 2: the logistic regression equation

Pediatric and Perinatal Epidemiology 1990; **4:** 39-52.

R. Brand, Department of Medical Statistics, Medical Faculty, State University, Leiden. The Netherlands

Summary. In Part I basic concepts were introduced as a preparation for an introductory explanation of logistic regression. Logistic regression is a statistical modelling technique, designed for the estimation of the simultaneous effects of predictors on the risk of a certain dichotomous outcome variable where each effect is estimated while adjusting for the effect of the other factors considered. The basic concepts-odds, odds ratio, confounding and interaction - were introduced in such a way that they naturally lead to the concept of logistic regression. In Part 2 the concepts are 'translated' into simple equations. By studying these equations the equivalence between such mathematical expressions and the underlying clinical assessment of risk will become clear.

2.0 The logistic regression equation

In ordinary linear regression, one models the mean (expected) value of some dependent variable, measured on a numerical scale, as a linear function of one or more other independent variables (predictors). For example, if one wants to calculate the expected value of birthweight knowing only gestational age, the following formula will be considered:

{expected value of birthweight} = $\mu + \alpha$ {a particular gestational age}

In this formula, μ and α are numbers chosen in such a way that this linear relationship is as close as possible a description of 'reality' (i.e. such that the equation describes as close as possible the relationship within the population investigated using only gestational age. Of course, trying to predict birthweight itself on the basis of gestational age alone is not possible: other factors would have to be introduced in order to obtain a fairly accurate estimate).

As an example, consider the following fictitious measurements of birthweight in a number of gestation age categories (Figure 1):

	Gestational age in completed weeks					
	24	25	26	27	28	29
Birthweight measurements in grams	700 800	700 900 1100	700 1000 1300	1400 1100 800	1300 1100 1200	1300 1400
Means	750	900	1000	1100	1200	1350

Corresponding regression equation:
(mean birthweight in grams) = -1600 + 100 (gestational age in weeks)

Figure 1.

The data in this table can be approximated by a straight line, the regression line, which describes the tendency of the mean birthweight as a function of gestational age. For this particular data, the equation would become:

{birthweight} = -1600 + 100 {gestational age}

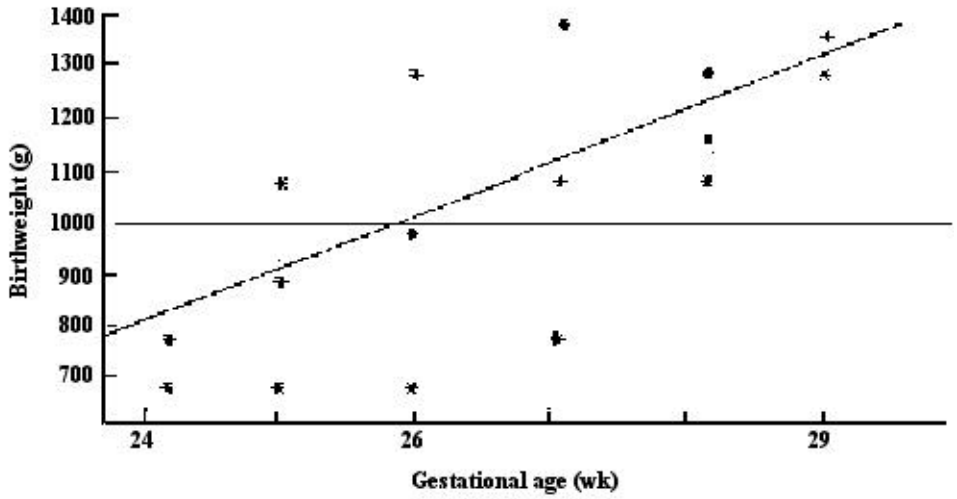


Figure 2.

Data and fitted regression line are depicted in Figure 2. Note that the fitted regression line applies only to the interval (24 - 29) on which it has been computed. It is obvious that extrapolation, both before 24 and beyond 29 weeks, leads to meaningless estimates of birthweight. In general one should never extrapolate a fitted regression model outside the scope of the variables on which it is based.

Logistic regression deals with dependent variables which are not measured on a numerical scale (as are length, age or weight), but which can have only two values (yes/no, with/without some disease, dead/alive) like for instance mortality. In a logistic regression equation, one does not model the mean of the outcome variable itself, but the probability that this outcome has one of two possible values. Of course the 'probability' is also, in a certain sense, a 'mean'. The logistic counterpart of the above linear regression equation, now modelling the dichotomous variable 'neonatal mortality' as an outcome, would look like:

log(**odds**(neonatal mortality)) = $\mu + \alpha$ {gestational age}

So, just as the straight line in the previous figure was a convenient way to depict the birthweight to be expected on the basis of gestational age alone, in the same way we can draw a straight line which depicts the risk of neonatal mortality (measured as log(odds)) based on gestational age.

The above equation relates the **log(odds)** to gestational age, not the **probability** or the **odds** as a measure of risk. One reason for the choice of log(odds) as a measure of risk is the fact that the log(odds) can vary (at least theoretically) between minus and plus infinity thus allowing the modelling of straight lines which (by their very nature) also extend between minus and plus infinity. If one tried to model the probability itself, the scope would have to be limited to the interval between 0% and 100% which precludes the use of straight lines or at least the interpretation of such a linear model when it predicts values which are theoretically impossible (below 0% or above 100%). The same holds for the odds which cannot assume negative values. The same argument applies when considering the risk ratio versus the odds ratio: any odds can be multiplied by any odds ratio and thus yield a valid (interpretable) odds; however, in terms of risks instead of odds, multiplying a risk of say 50% with a risk ratio of three would yield a risk of 150% which is clearly impossible. In other words, an estimated risk ratio has a valid interpretation only in a restricted interval of possible risks while the odds ratio

does not suffer from such a restriction. The other reason is that the parameters (values such as μ and α) in such an equation have a simple and *direct* clinical interpretation (which is highly desirable) as will be shown.

Now, applying a **logarithm** to a number is just like measuring a number on another *scale:* one can go back and forth between the two scales - in this case with the push of one button on a simple pocket calculator. The conversion between **odds** and **probability** is also a simple process (see Part 1). Hence one should not be distracted by a formula such as the above containing log and odds; rather one should read it as follows.

{some measure of neonatal mortality} = μ + α {gestational age}

The **log(odds[neonatal mortality])** is exactly the measure used in our diagrams in the sections of the preceding part on confounding and interaction. To simplify the discussion we used a multiplicative scale on the vertical axis while pretending that the axis measured the odds on mortality. In fact the vertical axis measures the log(odds) by virtue of which the relationship between mortality and gestational age became a straight line. We will now have to use the logarithm in our formulae to maintain the simple representation of a straight line, but for a qualitative *clinical interpretation* one could almost ignore it.

From now on we will discuss the logistic regression equations in relation to the Figures 8D, 9D, 10D and 12D presented in Part I which correspond to the Figures 4D, 5D, 6D and 8D in this part. This correspondence will be denoted by

(1:8D), . . ., (1:12D).

To simplify the computations we will assign to each gestational age, sex and mortality category a code as indicated in the following table. In the case of 'gestational age' we substract from each age the number 24 as if we start counting at that age. We will repeat the tables of Part 1 and explain how a straight regression line can describe the risk of mortality in the study cohort. The reader is encouraged to verify the equations himself on the basis of the data in the tables.

Variable	Short description	Value	Meaning
Neonatal mortality	MORT		0 = alive 1 = dead
Infant sex	SEX		0 = girl 1 = boy
Gestational age	GA		0 = 24 completed weeks 1 = 25 completed weeks etc. 5 = 29 completed weeks

Figure 3.

2.1 The odds

The equation which describes the relationship between mortality and gestational age in the first figure here (4D) is:

$$\log(\text{odds}[\text{MORT}]) = \mu + \alpha\ \text{GA} \qquad <0>$$

where in our cohort alpha is negative since the probability, and hence the odds, of dying (MORT) decreases with increasing GA (gestational age). The corresponding figure, (4T) is given below.

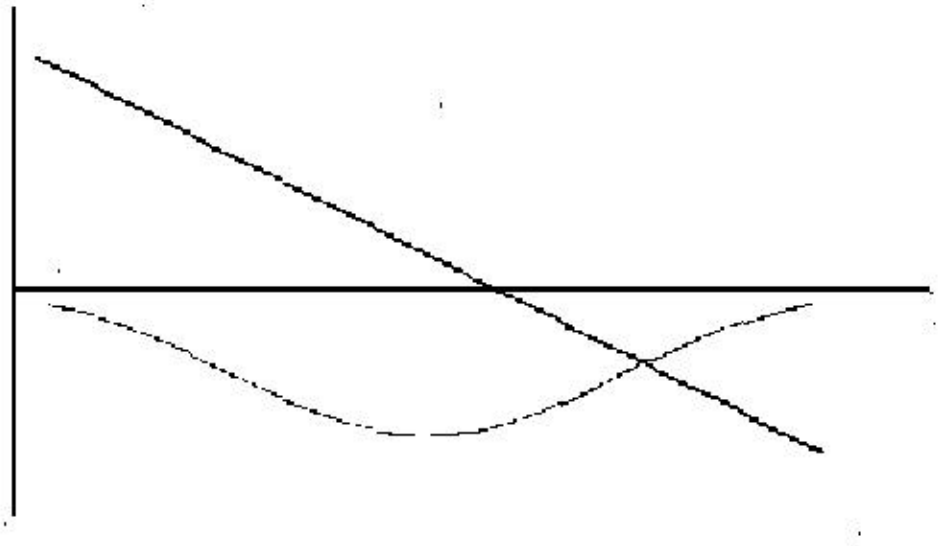


Figure 4D. (1:8D) Modelling the odds on mortality.

	Gestational age in completed weeks						
Mortality	24	25	26	27	28	29	Total
alive	3	14	34	50	44	28	173
dead	24	56	68	50	22	7	227
Odds	8	4	2	1	0.5	0.25	1.31

Figure 4T.

Just as we depicted in the first figure of this section the relationship between birthweight and gestational age, we now want to depict the relationship between the **odds** on mortality and **gestational age,** based on the data in Figure 4T and w want to compute the exact formula which describes that picture. We claim that the actual formula by which we can compute the odds in the cohort as described by Figure 4T is:

$$\log(\text{odds}[\text{MORT}]) = \log(8) — \log(2).GA$$

We will now verify this formula in detail. First one should know a *general*, property of logarithms: if two logarithms are subtracted, the result is the same as the logarithm of the quotient of the original values. For example: logarithm(8) - logarithm(2) = logarithm(4). This holds for any pair of numbers. Of course, when we add two logarithms, the result is the logarithm of the product; e.g. log(2) + log(5)= log(10).

We are now going to verify that the equation above exactly describes the data in Figure 4T, i.e. exactly describes the mortality odds in each gestational age category. To that end we compute the right-hand side of the equation for each gestational age:

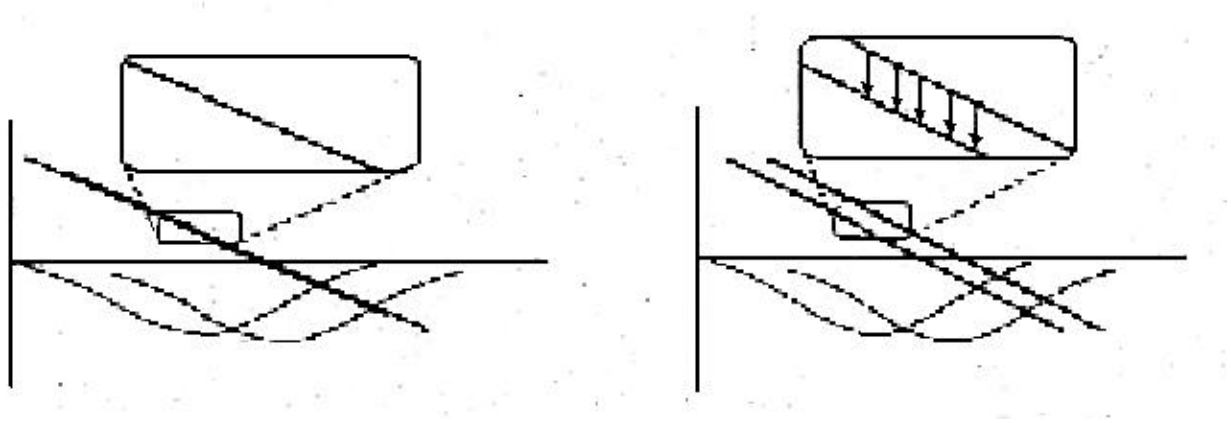
age 24, coded 0: log(8) - log(2).GA = log(8) - log(2).0 = = log(8) = log($odds_{24}$)
age 25, coded 1: log(8) - log(2).GA = log(8) - log(2).1 = log(8/2) = log(4) = log($odds_{25}$)
age 26, coded 2: log(8) - log(2).GA = log(8) - log(2).2 = log(8/4) = log(2) = log($odds_{26}$)
age 27, coded 3: log(8) - log(2).GA = log(8) - log(2).3 = log(8/8) = log(l) = log($odds_{27}$)
age 28, coded 4: log(8) - log(2).GA = log(8) - log(2).4 = log(8/16) = log(l/2) = log($odds_{28}$)
age 29, coded 5: log(8) - log(2).GA = log(8) - log(2).5 = log(8/32) = log(1/4) = log($odds_{29}$)

Hence the formula describes *exactly* what the odds are in each category.

In the forthcoming sections, the reader should try to perform such verifications himself. By *substituting* a single gestational age into the formula one can calculate the expected risk (as a log[odds]) and verify that the calculated number is an approximation of the risk stated in the appropriate table.

Of course, the expression <0> is a mathematical equation; it says that adding some number μ to α times the gestational age (expressed in weeks) one obtains an estimate of the log of the odds to die for an infant of that particular gestational age. But one can and should look at it in a more clinical way.

Apart from the log function, which is just a mathematical way of scaling the risk, the above relation tells us that one can view the odds for a particular infant as arising from some overall odds (represented by μ) which in its turn is modified by the particular gestational age we are considering (and a represents the 'weight' of the impact of gestational age on this overall odds μ). If α is negative, the odds diminish with increasing age, if α is positive, the odds increase.



Figures 5D and 6D. (1:9D and 1:10D) The odds ratio and confounding

2.2 The odds ratio

Consider the logistic model:

$$\log(\text{odds}(\text{MORT})) = \mu + \beta\ \text{SEX} + \alpha\ \text{GA} \qquad <1>$$

in which the risk factor SEX can have a value of either 0 (girl) or 1 (boy), and MORT is the outcome variable. In general coding a risk factor (and in particular the 'exposure' under investigation) as O (e.g. absent) and 1 (e.g. present) is a very convenient way of coding. It makes the equation very easy to interpret. Note also that a change from 0 to 1 is in a way a change of 1 unit which makes it comparable to changes in continuous (exposure) variables of 1 unit (like grams, weeks, etc.) as discussed in section 1.2.

	24			25			26			27			28			29					
Mortality	G	B	T	G	B	T	G	B	T	G	B	T	G	B	T	G	B	T	G	B	T
alive	1	2	3	5	9	14	14	20	34	30	20	50	26	18	44	20	8	28	96	77	173
dead	8	16	24	20	36	56	28	40	68	30	20	50	13	9	22	5	2	7	104	123	227
Odds	8	8	8	4	4	4	2	2	2	1	1	1	1/2	1/2	1/2	1/4	1/4	1/4	1.08	1.60	1.31
OR (B:G)	1.0			1.0			1.0			1.0			1.0			1.0			1.47		

Figure 5T.

We will now try to describe Table 5T in terms of an equation, just like we did in the case of Figure 4T, although there we had not yet considered SEX as a predictor of mortality. Consider the following equation:

$$\log(\text{odds}(\text{MORT})) = \log(8) + \beta\ \text{SEX} - \log(2).\text{GA}$$

in which β *is* a number still to be determined (it will turn out to be a measure of the *impact* SEX has on the resulting mortality odds). Now the variable SEX can assume two values: **zero** for girls and **one** for boys.

Substituting both values in this equation leads to two formulae for the calculation of the odds on mortality:

for girls: log(odds[MORT]) = log(8) + - log(2).GA <1.0>
for boys: log(odds[MORT]) = log(8) + β - log(2).GA <1.1>

As we have seen, Figure 5T reflects the situation where the odds for boys and girls are identical within each gestational age category. In what way will the above formulae reflect that situation? Clearly, if β equals 0, both equations are the same; both lines would coincide (have a distance O) and be equal to the one line describing Table 5T and hence there would be **no sex effect.** This corresponds exactly to Figure 5D.

On the other hand, if β is different from 0, both lines would he parallel - they would have the same slope (-log[2]), i.e. GA has the same impact - and would only differ in height by an amount of (log[8] + β) - (log[8]) = β. Note that this result is obtained by simply subtracting equation <1.0> from equation <1.1>.

First we will verify that a β of log(O.75) will give us a fair description of the mortality risk for boys in Figure 6T. We will calculate the odds in the category of 26 weeks using formula <1.1>.

age 26, coded 2: log(8) + log(0.75) - log(2).2 =
log(8) + log(3/4) - log(2) - log(2) =
log(8 * 3/4 /2 /2) =
log(1.5) = $\log(\text{odds}_{26})$

In the same way the reader can verify that log(O.75) yields a fair (though not always exact) approximation to the odds ratios in the table (Figure 6T).

	Gestational age in completed weeks, broken down by sex														
	24		25		26		27		28		29				
Mortality	G	B	G	B	G	B	G	B	G	B	G	B	G	B	T
alive	1	3	5	11	14	24	30	23	26	20	20	8	96	89	185
dead	8	15	20	34	28	36	30	17	13	7	5	2	104	111	215
Odds	8	5	4	3.1	2	1.5	1	0.74	0.5	0.35	0.25	0.25	1.08	1.25	1.16
OR (B:G)	0.63		0.77		0.75		0.74		0.7		1.0		1.15		

Figure 6T.

In Figure 6D the coefficient β equals log(O.75), corresponding to the constant odds ratio of 0.75. This coefficient β is exactly what we were looking for. The number β *is* the distance between the two fitted lines, measured as log(odds), so β *is an adjusted measure* for *a difference in probability of the outcome between boys and* girls. Note that the distance between the lines is independent of the gestational age.

This is *the crux of the regression approach.* Although gestational age has a great influence on the log(odds[MORT]), by considering only the difference between the lines, this influence is accounted for and no longer influences the estimation of the exposure effect. Incorporating a term GA in the regression equation is exactly the same as what a clinician would call 'adjusting or correcting for differences in gestational age'. Or, more intuitively, equation <1> shows us that the actual risk *any* infant has, can be divided into three 'parts': part μ, representing the risk any infant has, independent of its gender or gestational age; part β the risk associated with gender (adjusted for gestational age); and part α, the risk associated with a particular gestational age (adjusted for gender). It is the **model** which states that the total risk is the **sum** of those three partial risks. Of course the applicability of the model should be thoroughly investigated, both on clinical and statistical grounds. The actual values of these 'parameters' do however depend on the actual coding of the risk factors and cannot be interpreted in an absolute sense.

Looking at it in still another way, we have two equations which model the log(odds) as a function of gestational age, one for girls and one for boys (both derived from the equation <0>). *Subtract* equation <1.0> from <1.1>. The left-hand side becomes:

$$\log(\text{odds}[\text{MORT}|\text{boys}]) - \log(\text{odds}[\text{MORT}|\text{girls}])$$

which equals

$$\log \frac{\text{odds}(\text{MORT}|\text{boys})}{\text{odds}(\text{MORT}|\text{girls})} = \log(OR_E)$$

The right-hand side becomes:

$$(\log[8] + \beta - \log[2]\ GA) - (\log[8] - \log[2]\ GA) = \beta$$

for any value of GA.

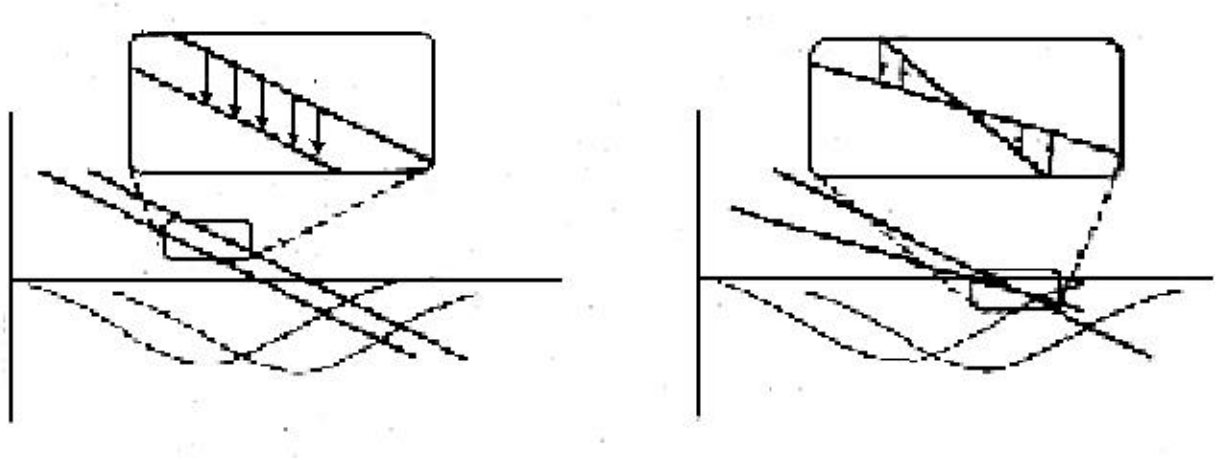
Hence the coefficient β is quite simply the logarithm of the odds ratio with respect to the risk factor SEX and the outcome MORT and thus we have:

$$OR_E\,(MORT) = e^{\beta}$$

where e is a mathematically defined number (appr. 2.718) which is the 'inverse' a the (natural) logarithm: if a = **log**(b) then b = e^a (e to the power a) and can be found on any simple pocket calculator.

In conclusion, a statistical test for the null hypothesis that the coefficient equals O is exactly an answer to the (clinical) question of whether there is evidence in the data that SEX is associated with MORT after correction for possible differences in the distribution of gestational age in both groups.

The logistic regression framework provides the statistician with the appropriate tests to see whether such a coefficient is more different from O than one might expect on the basis of random variation. This is exactly the property of the logistic regression equation which makes it such a useful tool for both statistician and clinician: the numbers (coefficients) emerging from the fitting of a simple straight line to data obtained from a clinical study have a direct clinical interpretation a measuring the association between 'exposure' and 'outcome'; their being zero a not can be both statistically tested and clinically interpreted.



Figures 7D and 8D. (1:11D and 1:12D) The odds ratio and interaction

2.3 Interaction

The figure with the data on which Figure 8D was based is given below.

	Gestational age in completed weeks, broken down by sex														
	24		25		26		27		28		29				
Mortality	G	B	G	B	G	B	G	B	G	B	G	B	G	B	T
alive	1	4	5	15	14	26	30	21	26	17	20	7	96	90	186
dead	8	14	20	30	28	34	30	19	13	10	5	3	104	110	214
Odds	8	3.5	4	2	2	1.3	1	0.90	0.5	0.59	0.25	0.43	1.08	1.22	1.15
OR (B:G)	0.44		0.50		0.65		0.90		1.18		1.71		1.13		

Figure 8T.

We return to Figure 7D and 8D and consider the following equations:

girls: log(odds[MORT]) = log(8) + (-log[2]) GA <3.0>
boys: log(odds[MORT]) = log(8) + β + (-log[2] + δ) GA <3.1>

We repeat that the coefficient of GA is, in fact. the slope of the regression line (which measures how fast the odds increase or decrease per unit change in GA).

therefore, the equation tells us that for the girls there is a line with slope (-log[2]) (which means, as we know now, that the odds are reduced by a factor 2 with increasing gestational age) and for the boys there is another line with slope (-log[2] + δ) for some value of δ (which might be negative, zero or positive).

If δ is zero, the lines are parallel which means that the rate by which the odds decrease with increasing gestational age (per unit change) is the **same** for boys and girls (which has nothing to do with the actual odds themselves: if δ does not equal zero, boys and girls do have different odds when of the same gestational age; however as does not measure the odds themselves but the way the odds change when gestational age is changed).

Hence Figure 7D corresponds to the situation in which δ equal 0, and Figure 8D to the one in which $\delta > 0$. If both lines coincide, then both β and δ equal 0. Furthermore, δ being greater than zero reflects the fact that the regression line for boys has a smaller (less steep) negative slope than the line for girls (since a positive amount is added to the slope when SEX changes from 'girl' to 'boy').

We will now show how *interaction* can be captured in a logistic regression equation. It is more or less a kind of trick, but a simple one. We combine these two equations into one equation by introducing SEX:

$$\log(\text{odds}[\text{MORT}]) = \log(8) + \beta\ \text{SEX} - \log(2).\text{GA} + \delta\ \text{SEX GA} \qquad <3>$$

Do verify that substituting a '1' for SEX renders equation <3.1> while substituting a '0' renders equation <3.0>. The crucial point is that this equation differs from equation if only by the term δSEX GA: the product of the parameter δ, the risk factor SEX (1 or 0), and the confounder GA (gestational age). This parameter is the one we are looking for, i.e. the parameter which measures how the influence of gestational age on the odds is *modified* by SEX.

Hence the question of whether δ is zero or not will give the answer to the clinical question of *whether the odds ratio varies with* GA. In statistical terminology: the test for the null hypothesis $\delta = 0$ is the same as a test of interaction between the risk factor 'gender' and the confounder 'gestational age', or, clinically speaking, test of whether gestational age *modifies* the effect gender has on the odds of mortality.

If we subtract equation <3.0> from equation <3.1>, as we did before, we get:

$$\log(OR_E\ [\text{MORT}]) = \beta + \delta\ \text{GA}$$

because both the constant (base line) odds of log(8) and the term (-log[2] GA) simply disappear when subtracting the two equations which differ *only* by the value of SEX substituted. But the mathematical process of subtracting the two equations means clinically *comparing the odds for boys and girls* by computing the

difference in their respective probabilities of dying. Since this comparison yield the simple expression ' $\beta + \delta$ GA', this means that the OR amounts to some value to which one should add δ GA for each gestational age considered. *If* δ *equals 0, the OR no longer varies with* varying *GA and is just one number,* β. If δ can be shown to be significantly different from zero, this means that there is no such thing as *the* odds ratio of gender, but that such an odds ratio depends on the value of gestational age.

As an exercise, the reader should verify that equation <3> and the corresponding equation for the $\log(OR_E)$ in the following specific form describe the situation depicted by Figure 8D and the corresponding cross-tabulation very well:

log(odds[MORT]) = 2.08 - 0.95 SEX - 0.69 GA + 0.28 SEX GA
$\log(OR_E$ [MORT]) = -0.95 + 0.28 GA

Simply substitute specific values for SEX and GA and verify that the result of the equation is indeed the logarithm of the odds presented in Figure 8T; in the same way the formula for the odds ratio can be verified. Note that the values (2.08, 0.95 etc.) have been obtained by using a computer programme for logistic regression and cannot be computed from the associated table. Note also that one obtains the odds from the log(odds) by exponentiation.

The different possibilities in equation <3>

log(odds[MORT])= log(8) + β SEX - log(2) GA + δ SEX GA <3>

are summarised in Figures 9, 10, 11.

Figure 9. $\beta = 0$, $\delta = 0$; the odds depend only on gestational age; given the same age, boys and girls have the same risk: the odds ratio equals 1.

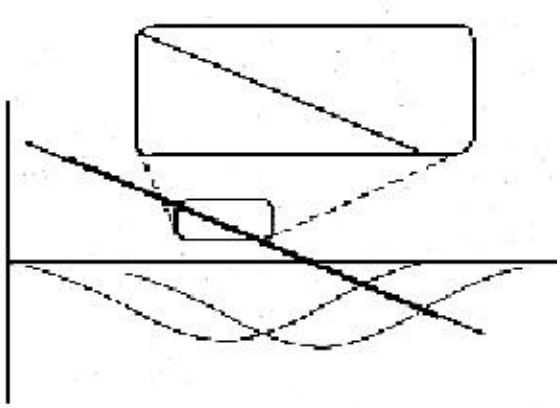


Figure 10. $\delta = 0$: odds depend both on gender and on gestational age; the odds ratio associated with gender however does not depend on gestational age and can he summarised in one number which measures the ratio of the risks (odds) for a boy versus a girl having the same gestational age, independent of the particular value of that age.

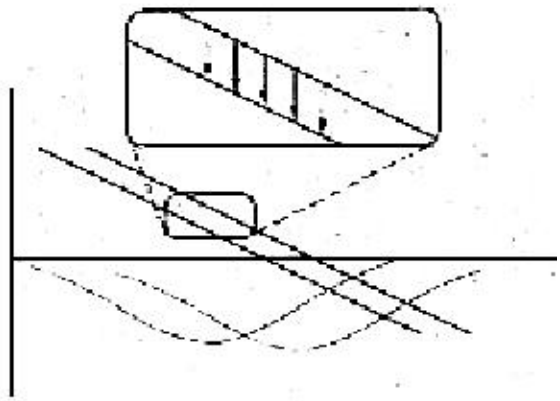
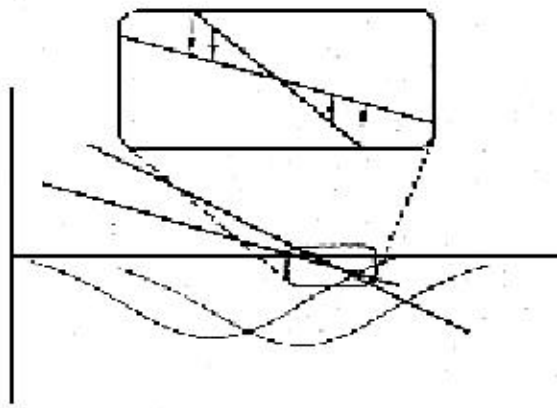


Figure 11. β, δ not 0; both the odds and the odds ratio depend on gestational age; no single number summarises the effect of gender on the odds; gestational age modifies the effect of gender.



We have omitted in our discussion one important aspect, the relaxation of the assumption of a linear relationship between the log(odds) and the confounder (i.e. gestational age). This was done because the introduction of such a non-linear relationship and the possibility of confounders being discrete or even unordered does not contribute at all to the understanding of the concept of confounding itself but would rather obscure the essence for the reader new to this field. It is however very important to realise that in the*practical application* of the logistic regression technique a *confounder can be any type of variable and can have a non-linear relationship with, the log(odds).* How one should deal with these situations is however beyond the scope of this introduction.

Such a situation is illustrated by the three figures below which are exactly analogous to the ones displayed above, hut assume a non-linear relationship between log(odds) and a confounder which assumes discrete values only.

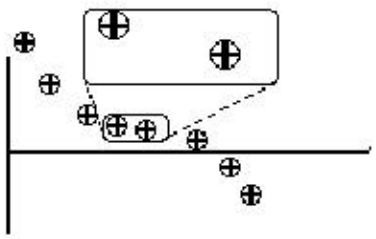


Figure 12. $\beta = 0$, $\delta = 0$.

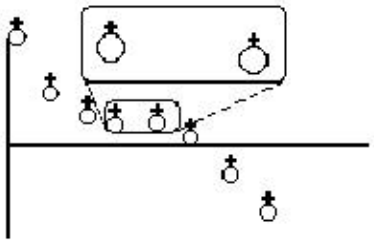


Figure 13. $\delta = 0$.

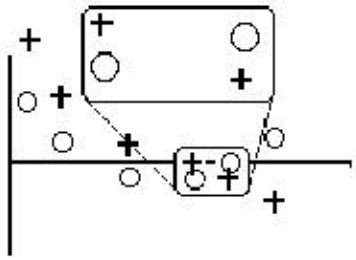


Figure 14. β, δ not 0.

This concludes the introduction to logistic regression in this paper. The following generalisation takes account of more than one confounder and presents the formulae in a somewhat more abstract way.

2.4 Summary and generalisation

A logistic regression equation which models the effect of an exposure factor E on an outcome variable OUT, in which we make adjustments for (possible) confounders C_1, C_2, . . . C_n, has the following general form:

$$\log(\text{odds}[OUT]) = \mu + \beta E + \alpha_1 C_1 + ... + \alpha_n C_n + \delta_1 E C_1 + ... + \delta_n E C_n$$

If all δ_i, can be supposed to equal 0 (e.g. by some statistical testing method), the equation reduces to:

$$\log(\text{odds}[OUT]) = \mu + \beta E + \alpha_1 C_1 + ... + \alpha_n C_n$$

and hence we have:

$$\log(OR_E) = \beta$$

because subtracting the two equations emerging when E is 1 and 0 respectively, yields an equation in which the left-hand side is *by definition* the log of the odds ratio and in the right-hand side everything not related to E cancels out.

If, for example, δ_1 and δ_2 cannot be supposed to equal 0, the equation would be:

$$\log(OR_E) = \beta + \delta_1 C_1 + \delta_2 C_2$$

and hence the OR can only he computed from this model by specifying specific combinations of values of both C_1 and C_2.

Broadly speaking then, we can use the logistic regression equation in two situations:

(1) one is interested in the *prediction* of the actual odds for a certain category of patients on the basis of .the values observed for each of the risk factors in the equation- in this case, one might view the logistic regression as a kind of 'smoothing procedure';

(2) one is interested in the *effect of a* certain *exposure while other risk factors figure as 'confounders' to be adjusted for in order to get as valid an estimate as possible.*

2.5 A final example

Suppose we are studying a cohort of very preterm infants who were liveborn and were neonatally transported to a neonatal intensive care unit (NICU). It is of interest to know whether the actual distance covered in the neonatal transport is associated with the risk of mortality during the first hospitalisation period following the transportation Suppose for the sake of simplicity, that we have a kind of 'index of illness' which measures the degree of illness for each child (this index might be constructed from several clinical indices) on a numerical scale.

In order to assess the true effect of the distance covered, we must make sure that there is no confounding effect from the severity of the infant's condition on the distance/mortality relationship. Such a confounding effect might occur if for example there was a tendency to prefer a short distance over a long one for very ill children, although the mother lives near a remote NICU. In Figure 15, an extreme example of confounding is depicted. We grouped the distance covered in four categories, increasing by 25 km. Obviously the odds on mortality is associated with the condition of the infant. Due to the confounding effect, the crude risks seem to diminish with increasing transportation distance (because the mean index of illness shifts towards a healthier infant in the 'far' transportations). But actually the risk for comparable infants increases distinctly in a consistent way ('dose-response relationship').

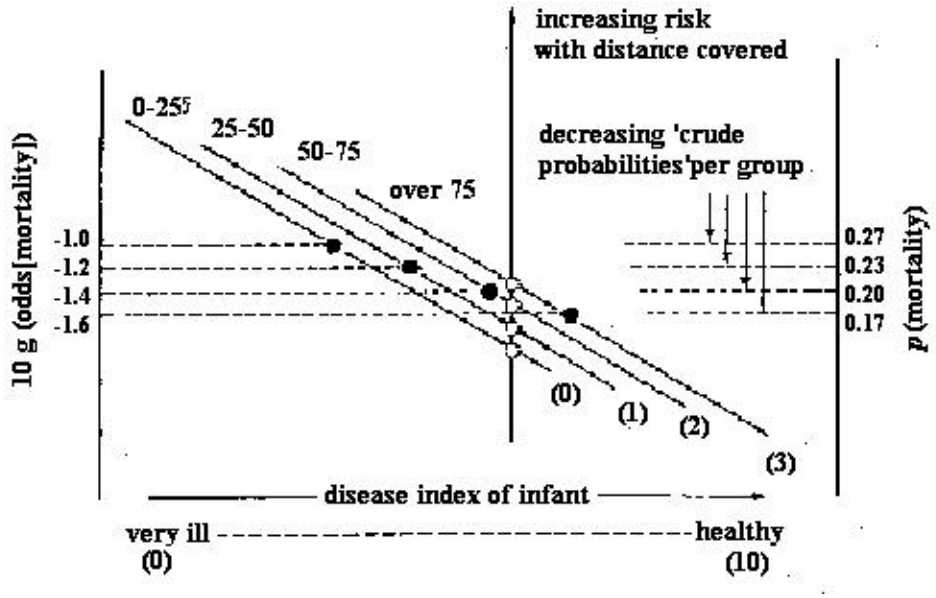


Figure 15. Effect of transportation distance (km)

The corresponding formula describing the regression lines in Figure 15 is:

$$\log(\text{odds[mortality]}) = 1 + 0.3 \times \text{DIST} - 0.5 \times \text{INDEX}$$

The variable DIST is coded as 0, 1, 2 or 3 and the variable INDEX runs from 0 to 10. The mean value for INDEX for each distance group separately is 4, 5, 6 and 7 respectively. Then for an 'average' infant in each of the groups:

Category	log(odds)	odds	risk

DIST = 0	1 + 0.3 x 0 - 0.5 x 4 = -1.0	0.37	27%
DIST = 1	1 + 0.3 x 1 - 0.5 x 5 = -1.2	0.30	23%
DIST = 2	1 + 0.3 x 2 - 0.5 x 6 = -1.4	0.25	20%
DIST = 3	1 + 0.3 x 3 - 0.5 x 7 = -1.6	0.20	17%

So in each group the risk to an infant with an INDEX equalling the mean index of that group, decreases. However, comparing like with like (infants with the same INDEX), we see the reverse:

Category	log(odds) - log(odds|DIST = 0)	OR (category vs. baseline)
DIST = 1	(1 + 0.3 x 1 - 0.5 x INDEX) - (1 + 0.3 x 0 - 0.5 x INDEX) = 0.3	$e^{0.3} = 1.35$
DIST = 2	(1 + 0.3 x 2 - 0.5 x INDEX) - (1 + 0.3 x 0 - 0.5 x INDEX) = 0.6	$e^{0.6} = 1.82$
DIST = 3	(1 + 0.3 x 3 - 0.5 x INDEX) - (1 + 0.3 x 0 - 0.5 x INDEX) = 0.9	$e^{0.9} = 2.46$

Here we have (arbitrarily) designated the category 'DIST = 0' as baseline category. Hence we see that when we adjust for the disease index there exists an increasing risk associated with increasing transportation distance covered.

[Acknowledgement and references: see Part 1]

Source:

Brand R, Keirse MJNC. Using logistic regression in perinatal epidemiology: an introduction for clinical researchers. Part 1: basic concepts. *Paediatrics and Perinatal Epidemiology*, 1990; **4**: 22-38, 221-235.