Goodness-of-fit tests for fitted RRs

Author : P.N. Lee
Date : 11th June 2012

1. Prospective studies

We have an observed table of pseudo-numbers:

Level	Cases	At risk
Baseline	A_0	N_0
Low exposure	A_1	N_1
High exposure	A_2	N_2
Total	A_S	N_S

We have fitted a set of RRs, $1 : R_1 : R_2$ $(R_0 = 1)$

We wish to derive a set of fitted cases F_0, F_1, F_2

We have the following formulae:

$F_0 + F_1 + F_2 = A_S$ (marginal totals stay the same) (1)

$R_1 = (F_1N_0)/(F_0N_1)$ (2)

$R_2 = (F_2N_0)/(F_0N_2)$ (3)

From (2) $F_1 = F_0N_1R_1/N_0$ (4)

From (3) $F_2 = F_0N_2R_2/N_0$ (5)

From (1,4,5) $F_0 + \dfrac{F_0 N_1 R_1}{N_0} + \dfrac{F_0 N_2 R_2}{N_0} = A_S$ (6)

so $F_0 N_0 + F_0 N_1 R_1 + F_0 N_2 R_2 = A_S N_0$ (7)

or $F_0 = (A_S N_0 R_0) / \sum_{i=0}^{2} (N_i R_i)$ (8)

From (4,8) $F_1 = (A_S N_1 R_1) / \sum_{i=0}^{2} (N_i R_i)$ (9)

From (5,8) $F_2 = (A_S N_2 R_2) / \sum_{i=0}^{2} (N_i R_i)$ (10)

This allows one to derive fitted values and is clearly generalisable to multiple exposure levels (k).

A chisquared test of goodness-of-fit may be derived in the usual way from the formula:

$$\chi^2 = \sum_{i=0}^{k} (A_i - F_i)^2 / F_i$$

The degrees of freedom is k – 1.

2. Case-control studies

Here the observed table of pseudo-numbers is:

Level	Cases	Controls	Total
Baseline	A_0	B_0	C_0
Level 1	A_1	B_1	C_1
Level 2	A_2	B_2	C_2
Total	A_S	B_S	C_S

The expected table of fitted numbers is:

Level	Cases	Controls
Baseline	F_0	G_0
Level 1	F_1	G_1
Level 2	F_2	G_2

We have fitted RRs, $1 : R_1 : R_2$ $(R_0 = 1)$

2.1 A first attempt to solve the equations

The reader may prefer to skip to section 2.2.

We can write down the following formulae based on the marginal totals and the RRs:

$$F_0 + G_0 = C_0 \tag{1}$$

$$F_1 + G_1 = C_1 \tag{2}$$

$$F_2 + G_2 = C_2 \tag{3}$$

$$F_0 + F_1 + F_2 = A_S \tag{4}$$

$$R_1 = F_1G_0/(F_0G_1) \tag{5}$$

$$R_2 = F_2G_0/(F_0G_2) \tag{6}$$

From (5) $$G_1 = F_1G_0/(F_0R_1) \tag{7}$$

From (6) $$G_2 = F_2G_0/(F_0R_2) \tag{8}$$

From (2,5) $$F_1 + \frac{F_1G_0}{F_0R_1} = C_1 \tag{9}$$

or $$F_0F_1R_1 + F_1G_0 = C_1F_0R_1 \tag{10}$$

From (1) $$F_0F_1R_1 + F_1(C_0 - F_0) = C_1F_0R_1 \tag{11}$$

or $F_1 = C_1F_0R_1/(F_0R_1 + C_0 - F_0)$ (12)

Similarly $F_2 = C_2F_0R_2/(F_0R_2 + C_0 - F_0)$ (13)

From (4,12,13) $F_0 + \dfrac{C_1F_0R_1}{(F_0R_1 + C_0 - F_0)} + \dfrac{C_2F_0R_2}{(F_0R_2 + C_0 - F_0)} = A_S$ (14)

This is an equation in F_0 only, which hopefully can be solved.

Formula (12) gives F_1 in terms of F_0

Formula (13) gives F_2 in terms of F_0

Formulae (1,2,3) then give G_i in terms of F_i

This gives the whole table of fitted numbers. Here we can calculate a goodness-of-fit statistic as:

$$\chi^2 = \sum_{i=0}^{k} (A_i - F_i)^2 / F_i + \sum_{i=0}^{k} (B_i - G_i)^2 / G_i$$

The degrees of freedom is 2k – 1.

The problem is solving formula (14). For k = 2 it is a cubic (as can be seen by multiplying through by the denominators). For general k it involves powers of k + 1 presumably, and I don't know how to solve that. I suppose one could search starting at $F_0 = A_0$.

2.2 An alternative approach

This is illustrated in T:/PNLEE/FITFORCCSTUDIES.XLSX. Here we have an observed table of pseudo-numbers of

Level	Cases	Controls	Total
Baseline	$A_0 = 40$	$B_0 = 110$	$C_0 = 150$
Level 1	$A_1 = 90$	$B_1 = 110$	$C_1 = 200$
Level 2	$A_2 = 220$	$B_2 = 80$	$C_2 = 300$
Total	$A_S = 350$	$B_S = 300$	$C_S = 650$

We have fitted RRs at 1 : 2 : 4

In line 7 of the spreadsheet we first calculate B_iR_i for each level and in total, and then derive a first estimate of fitted cases using the formula

$$F_i = (A_S B_i R_i) / \sum_{i=0}^{e} (B_i R_i)$$

Taking the first estimate of fitted controls (G_i) as the original pseudo number (B_i), we then have a set of cases and controls which produce the fitted RRs but with the marginal totals not adding up over level.

In line 8, we scale the estimates of F_i and G_i by the same factor so that their sum is equal to C_i. Now the marginal totals over all cases and over all controls do not equal A_S and B_S. So in line 9, we scale them so that they do sum to A_S and B_S. However, the totals over levels, C_i, are then wrong.

We therefore repeat the two steps until they converge to the right answer, which is

Level	Cases	Controls	Total
Baseline	$F_0 = 50$	$G_0 = 100$	$C_0 = 150$
Level 1	$F_1 = 100$	$G_1 = 100$	$C_1 = 200$
Level 2	$F_2 = 200$	$G_2 = 100$	$C_2 = 300$
Total	$A_5 = 350$	$B_5 = 300$	$C_5 = 650$

This happens rapidly as can be seen and by 5 goes (line 16) the fit is very good.

Lines 22-35 in the spreadsheet correspond to lines 7 – 20 and show the chisquared statistic.