Meta-analysis of epidemiological dose-response studies

2nd Italian Stata Users Group meeting

October 10-11, 2005

Nicola Orsini

Institute Environmental Medicine, Karolinska Institutet

Rino Bellocco

Dept. Medical Epidemiology and Biostatistics, Karolinska Institutet

Sander Greenland

Dept. Epidemiology, UCLA School of Public Health

Outline

- Motivating example - Case-control and Incidence Rate Data
- The statistical model and estimation method
- How to fit the variance-covariance matrix
- Analysis of multiple studies
- Modeling sources of heterogeneity

Meta-analysis

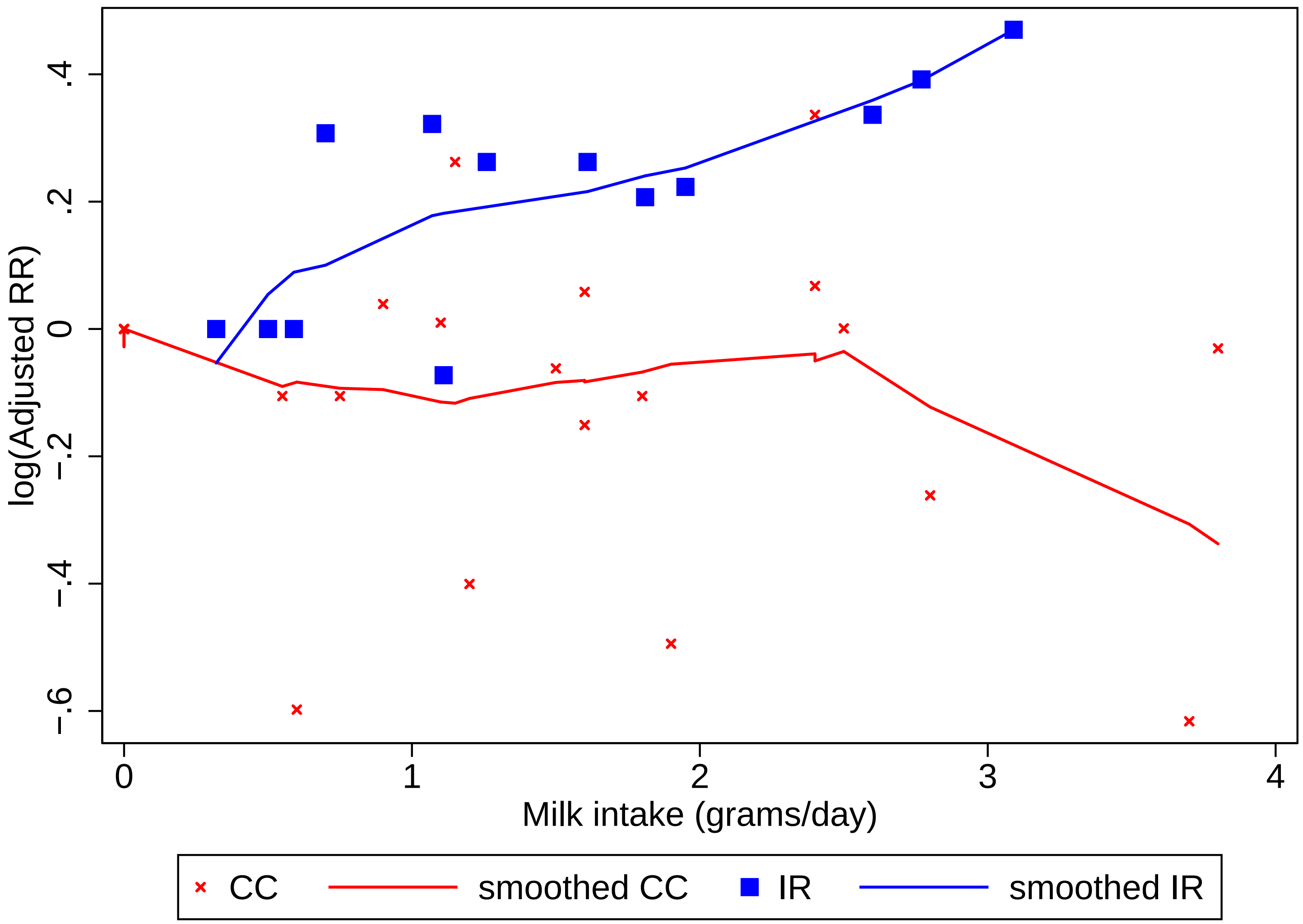
Larsson S.C., Orsini N., Wolk A., *Milk, milk products and lactose intake and ovarian cancer risk: A meta-analysis of epidemiological studies*, Int J Cancer, 2005.

6 Case-control studies

3 Cohort studies

```
. use http://nicolaorsini.altervista.org/2ISM/ovcancer, clear
```

```
. list if id < 4 | id == 9, clean

       id    author   year   study   adjrr     lb     ub   dose   case        n
  1.    1     Engle   1991      CC       1      1      1      0     15       50
  2.    1     Engle   1991      CC      .9     .4    2.2    5.5     21       56
  3.    1     Engle   1991      CC     1.3     .6    2.9   11.5     35       54
  4.    1     Engle   1991      CC      .9     .4      2     18     16       52
  5.    2     Risch   1994      CC       1      1      1      0     97      232
  6.    2     Risch   1994      CC    1.04    .71   1.53      9    107      250
  7.    2     Risch   1994      CC     .86    .58   1.28     16    102      243
  8.    2     Risch   1994      CC    1.07    .72   1.59     24    143      284
  9.    3      Webb   1998      CC       1      1      1      0    128      292
 10.    3      Webb   1998      CC    1.01    .71   1.43     11    133      297
 11.    3      Webb   1998      CC    1.06    .74   1.51     16    134      296
 12.    3      Webb   1998      CC     1.4    .98      2     24    177      328
 13.    3      Webb   1998      CC     .97    .67   1.41     38    149      317
 34.    9   Larsson   2005      IR       1      1      1    5.9     54   227238
 35.    9   Larsson   2005      IR     1.3     .9   1.88   12.6     68   219977
 36.    9   Larsson   2005      IR    1.23    .86   1.76   18.1     74   222101
 37.    9   Larsson   2005      IR    1.48   1.05   2.09   27.7     92   225412
```


log(Adjusted RR)
−.6
−.4
−.2
0
.2
.4
0
1
2
3
4
Milk intake (grams/day)
CC
smoothed CC
IR
smoothed IR

Fixed-effects Dose-Response Model

$$\mathbf{y} = \mathbf{X}\beta + \epsilon$$

where

$\mathbf{y}$ is a $n \times 1$ vector of beta coefficients (log odds ratios, log rate ratios, log risk ratios)

$\mathbf{X}$ is a $n \times p$ fixed-effects design matrix (no intercept). x_{i1} is assumed to be the exposure variable, where $i = 1, 2, ..., n$ identifies non-reference exposure levels

β is a $p \times 1$ vector of unknown coefficients

ϵ is a $n \times 1$ vector of random errors, such that $\epsilon \sim N(\mathbf{0}, \mathbf{\Sigma})$

Generalized Least Squares

Suppose for now that the variance-covariance matrix of the error $\boldsymbol{\Sigma}$ is known.

This method involves minimizing $(\mathbf{y} - \mathbf{X}\beta)'\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \mathbf{X}\beta)$ with respect to β.

The resulting estimator $\widehat{\beta}$ of the regression coefficients β is

$$\widehat{\beta} = (\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{y}$$

and the estimated covariance matrix $\mathbf{V}$ of $\widehat{\beta}$ is

$$\mathbf{V} = \text{Cov}(\widehat{\beta}) = (\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}$$

Variance-Covariance Matrix

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_{11} & & & & \\ \vdots & \ddots & & & \\ \sigma_{i1} & & \sigma_{ij} & & \\ \vdots & & & \ddots & \\ \sigma_{n1} & \dots & \sigma_{nj} & \dots & \sigma_{nn} \end{bmatrix}$$

- In Weighted Least Square (WLS) the off-diagonal elements of $\mathbf{\Sigma}$ are set to zeros ($\mathbf{y}$ are **independent**).

- In Generalized Least Squares (GLS) the off-diagonal elements $\mathbf{\Sigma}$ may not be zeros ($\mathbf{y}$ are **dependent**).

Statistical problems using WLS

Because the relative risks are estimated using a common referent group they are not independent. The WLS method would lead to

- Inefficiency of the slope estimator

- Inconsistency of the variance estimator

In a meta-analysis of summarized dose-response data underestimation of the variance of the slope leads to overestimation of the weight.

How to calculate the variances

The diagonal element σ_{ij} of $\mathbf{\Sigma}$, with $i = j$, and simply denoted by σ_i, the variance of the beta coefficient y_i, is calculated from the normal-theory-based confidence limits

$$\sigma_i = [(\log(u_b) - \log(l_b))/(2 \times z_{\alpha/2})]^2$$

where

u_b and l_b are, respectively, the upper and lower bounds of the reported $\exp(y_i)$,

$z_{\alpha/2}$ denotes the $(1 - \alpha/2)$-level standard normal deviate (e.g. use 1.96 for 95% confidence interval).

Information required to estimate covariances

As described by Greenland and Longnecker (1992), for each exposure levels, $i = 1, 2, ..., n$, we need to know the

- number of cases

and, according to the type of study

- number of controls in Case-Control (CC) Data
- number of person-time in Incidence-Rate (IR) Data
- number of non-cases in Cumulative Incidence (CI) Data

How to calculate the covariances in CC

	Exposure levels						Total
	x_{01}	x_{11}	$\dots$	x_{i1}	$\dots$	x_{n1}	
Cases	A_0	A_1	$\dots$	A_i	$\dots$	A_n	$M_1 = \sum_{i=0}^{n} A_i$
Controls	B_0	B_1	$\dots$	B_i	$\dots$	B_n	$M_0 = \sum_{i=0}^{n} B_i$
Total	N_0	N_1	$\dots$	N_i	$\dots$	N_n	$M_1 + M_0$

1. Fit cell counts to the interior of the $2 \times (n+1)$ summary table (which has margin M_1 and N_i), such that

$$(A_i \times B_0)/(A_0 \times B_i) = \exp(y_i)$$

2. Estimate the asymptotic correlation, r_{ij}, by

$$r_{ij} = s_0/(s_i s_j)^{1/2}$$

where $s_0 = (1/A_0 + 1/B_0)$ and $s_i = (1/A_i + 1/B_i + 1/A_0 + 1/B_0)$.

3. Estimate the off-diagonal elements, σ_{ij}, of the asymptotic covariance matrix $\mathbf{\Sigma}$ by

$$\sigma_{ij} = r_{ij} \times (\sigma_i \sigma_j)^{1/2}$$

where σ_i and σ_j are the variances of y_i and y_j.

How to calculate the covariances in IR

	Exposure levels						Total
	x_{01}	x_{11}	$\ldots$	x_{i1}	$\ldots$	x_{n1}	
Cases	A_0	A_1	$\ldots$	A_i	$\ldots$	A_n	$M_1 = \sum_{i=0}^{n} A_i$
Person-time	N_0	N_1	$\ldots$	N_i	$\ldots$	N_n	$M_0 = \sum_{i=0}^{n} N_i$

1. Fit cell counts such that $(A_i \times N_0)/(A_0 \times N_i) = \exp(y_i)$

2. Estimate the correlations $r_{ij} = s_0/(s_i s_j)^{1/2}$ where $s_0 = (1/A_0)$ and $s_i = (1/A_i + 1/A_0)$

3. Estimate the covariances $\sigma_{ij} = r_{ij} \times (\sigma_i \sigma_j)^{1/2}$

Heterogeneity

The analysis of the estimated residual vector $\widehat{\epsilon} = \mathbf{y} - \mathbf{X}\widehat{\beta}$ is useful to evaluate how close reported and fitted beta coefficients are at each exposure level.

A statistic for the goodness of fit of the model is

$$Q = (\mathbf{y} - \mathbf{X}\widehat{\beta})' \mathbf{\Sigma}^{-1} (\mathbf{y} - \mathbf{X}\widehat{\beta})$$

where

Q has approximately, under the null hypothesis, a χ^2 distribution with $n - p$ degrees of freedom.

Example: WLS trend for a single study

```
. vwls logrr dose if id == 9 & logrr != 0, sd(se) nocons

Variance-weighted least-squares regression           Number of obs   =          3
Goodness-of-fit chi2(2)    =    0.27                 Model chi2(1)   =       7.95
Prob > chi2                =  0.8728                 Prob > chi2     =     0.0048
------------------------------------------------------------------------------
       logrr |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        dose |   .1423799   .0505126     2.82   0.005      .043377    .2413827
------------------------------------------------------------------------------
```

Example: GLS trend for a single study

```
. glst logrr dose if id == 9, se(se) cov(n case) ir

Generalized least-squares regression                    Number of obs   =        3
Goodness-of-fit chi2(2)      =     0.56                 Model chi2(1)   =     4.49
Prob > chi2                  =   0.7553                 Prob > chi2     =   0.0340
------------------------------------------------------------------------------
       logrr |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        dose |   .1309131   .0617632     2.12   0.034     .0098594    .2519669
------------------------------------------------------------------------------

. mat list e(Sigma)

symmetric e(Sigma)[3,3]
           c1          c2          c3
r1  .03531387
r2   .0189355   .03337611
r3  .01899974   .01828257   .03083846
```

Meta-analysis of multiple studies with fixed-effects models

Let's define the matrices $\mathbf{y}_k$ and $\mathbf{X}_k$, respectively, the $n_k \times 1$ response vector and the $n_k \times p$ covariates matrix for the k^{th} study, with $k = 1, 2, ..., S$.

The number of non-reference exposure levels n_k for the k^{th} study might varies among the S studies.

Let's pool the data by appending the matrices $\mathbf{y}_k$ and $\mathbf{X}_k$ underneath each other,

$$\mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_k \\ \vdots \\ \mathbf{y}_S \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_k \\ \vdots \\ \mathbf{X}_S \end{bmatrix}$$

The outcome variable $\mathbf{y}$ the of dose-response model will be a $T \times 1$ vector, with $T = \sum_{k=1}^{S} n_k$; and the linear predictor $\mathbf{X}$ will be a $T \times p$ matrix.

Let $\mathbf{\Sigma}$ be a symmetric $T \times T$ block-diagonal matrix,

$$\mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma}_1 & & & & \\ \vdots & \ddots & & & \\ \mathbf{0} & & \mathbf{\Sigma}_k & & \\ \vdots & & & \ddots & \\ \mathbf{0} & \ldots & \mathbf{0} & \ldots & \mathbf{\Sigma}_S \end{bmatrix}$$

where $\mathbf{\Sigma}_k$ is the $n_k \times n_k$ estimated covariance matrix for the k^{th} study.

Example: Trend for multiple studies

```
. glst logrr dose , se(se) cov(n case) pfirst(id study)

Generalized least-squares regression                 Number of obs    =       28
Goodness-of-fit chi2(27)    =   40.25                Model chi2(1)    =     1.11
Prob > chi2                 =  0.0486                Prob > chi2      =   0.2925
------------------------------------------------------------------------------
       logrr |      Coef.   Std. Err.       z     P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        dose |   .0254944   .0242201     1.05   0.293    -.0219761    .0729648
------------------------------------------------------------------------------
```

Overall, there is no evidence of association between milk intake and risk of ovarian cancer. However, the goodness-of-fit test (Q=40.25, $p = 0.0486$) suggests that we should take into account potential sources of heterogeneity.

Example: Trend estimate for case-control studies

```
. glst logrr dose if study == 1, se(se) cov(n case) pfirst(id study)

Generalized least-squares regression                   Number of obs   =        18
Goodness-of-fit chi2(17)   =    24.02                  Model chi2(1)   =      1.22
Prob > chi2                =   0.1190                  Prob > chi2     =    0.2699
------------------------------------------------------------------------------
       logrr |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        dose |  -.0340478   .0308599    -1.10   0.270     -.094532    .0264365
------------------------------------------------------------------------------
```

No association between milk intake and risk of ovarian cancer was found among 6 case-control studies.

Example: Trend estimate for cohort studies

```
. glst logrr dose if study == 2, se(se) cov(n case) pfirst(id study)

Generalized least-squares regression                  Number of obs   =        10
Goodness-of-fit chi2(9)     =     6.54                Model chi2(1)   =      9.58
Prob > chi2                 =   0.6852                Prob > chi2     =    0.0020
------------------------------------------------------------------------------
       logrr |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        dose |   .1209988   .0390836     3.10   0.002     .0443964    .1976012
------------------------------------------------------------------------------
```

A positive association between milk intake and risk of ovarian cancer was found among 3 cohort studies.

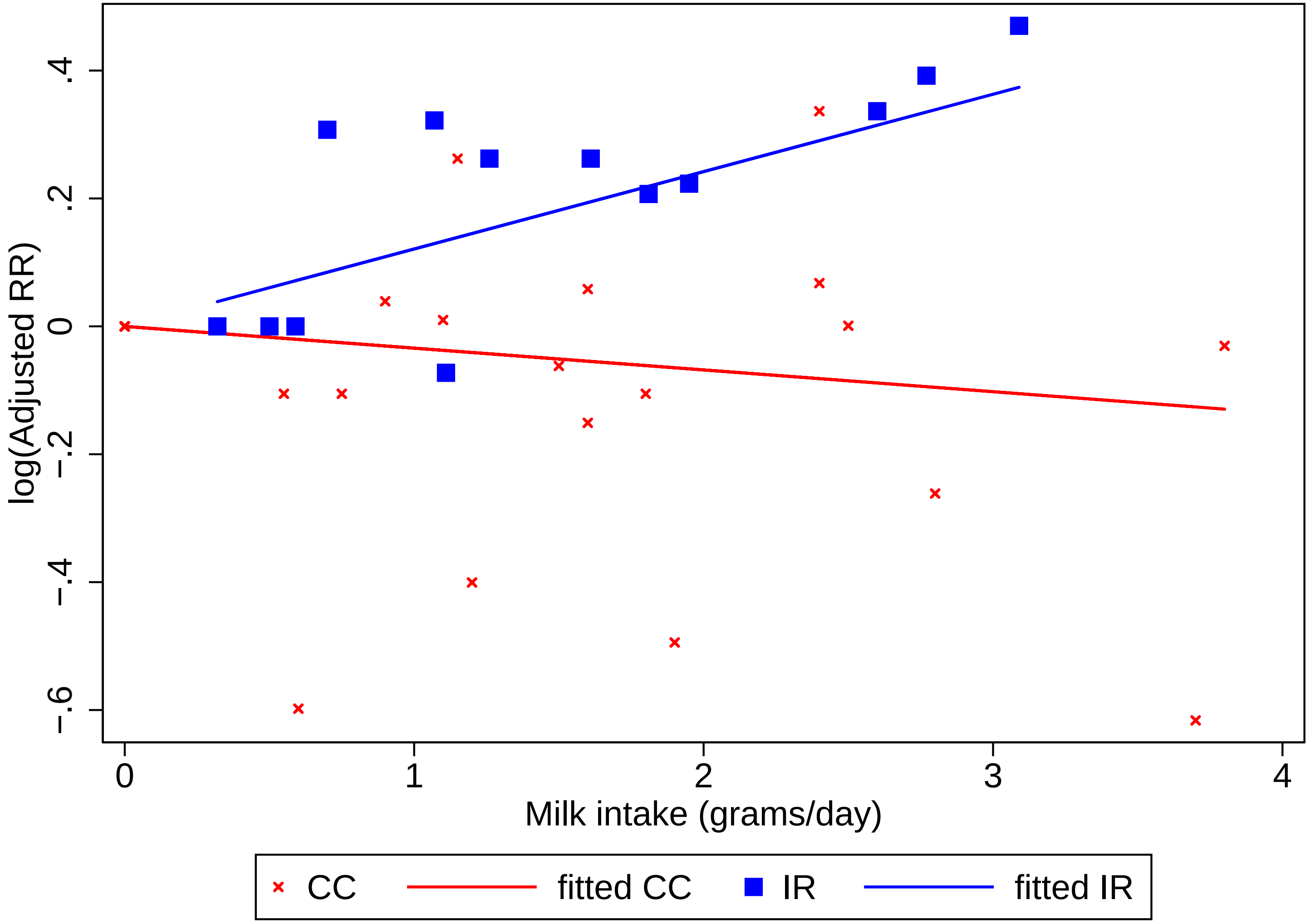
Modeling sources of heterogeneity

```
. gen types = study == 2

. gen doseXtypes = dose*types

. glst logrr dose doseXtypes, se(se) cov(n case) pfirst(id study)

Generalized least-squares regression                 Number of obs   =      28
Goodness-of-fit chi2(26)    =   30.55                Model chi2(2)   =   10.80
Prob > chi2                 =  0.2453                Prob > chi2     =  0.0045
------------------------------------------------------------------------------
       logrr |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        dose |  -.0340478   .0308599    -1.10   0.270     -.094532    .0264365
  doseXtypes |   .1550465   .0497982     3.11   0.002     .0574439    .2526492
------------------------------------------------------------------------------
```

A systematic difference in slopes related to study design might results, for instance, from the existence of recall bias in the case-control studies that would not be present in the cohort studies.

Interpretation of the slopes (trend)

```
. lincom dose + doseXtypes*0 , eform

 ( 1)  dose = 0

------------------------------------------------------------------------------
       logrr |     exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   .9665254   .0298269    -1.10   0.270     .9097986    1.026789
------------------------------------------------------------------------------

. lincom dose + doseXtypes*1 , eform

 ( 1)  dose + doseXtypes = 0

------------------------------------------------------------------------------
       logrr |     exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   1.128624   .0441106     3.10   0.002     1.045397    1.218476
------------------------------------------------------------------------------
```



log(Adjusted RR)
.4
.2
0
−.2
−.4
−.6
0
1
2
3
4
Milk intake (grams/day)
CC
fitted CC
IR
fitted IR

Conclusions

- The findings of case-control studies do not provide evidence of positive associations between dairy food and lactose intakes with risk of ovarian cancer.

- In contrast, the 3 cohort studies are consistent and show significant positive associations between intakes of total dairy foods, low-fat milk, and lactose and risk of ovarian cancer.

- The summary estimate of the relative risk for a daily increase of 10 g/day in lactose intake (the approximate amount in 1 glass of milk) was 1.13 (95% CI = 1.05-1.22) for cohort studies.

About the command

The command `glst` is written for Stata 9. It uses in-line Mata functions, the new matrix programming language (`help mata`) for the

- Iterative fitting algorithm (Newton's method) to get $\mathbf{\Sigma}$

- Generalized Least Squares estimator

- Confidence bounds of the covariances $\mathbf{\Sigma}$

Download

To install the `glst` command and run the do-file with the examples, type at the Stata command line

```
.  do http://nicolaorsini.altervista.org/2ISM/glst_exs.do
```

`glst` is downloadable from Nicola's website

```
.  net from http://nicolaorsini.altervista.org/stata
```

or from Statistical Software Components (SSC) archive

```
.  ssc install glst
```

References

S. Greenland and M. P. Longnecker, *Methods for trend estimation from summarized dose-reponse data, with applications to meta-analysis*, AJE, 135, 1301-1309, 1992

A. Berrington and D. R. Cox, *Generalized least squares for the synthesis of correlated information*, Biostatistics, 4, 423-431, 2003

J. Q. Shi and J. B. Copas, *Meta-analysis for trend estimation*, Statistics in Medicine, 23, 3-19, 2004

N. Orsini, R.Bellocco and S. Greenland, *Generalized Least Squares for trend estimation of summarized dose-response data*, submitted, 2005