The Stata Journal

The *Stata Journal*, electronic version (ISSN 1536-8734) is a publication of Stata Press, and Stata is a registered trademark of StataCorp LP.

The Stata Journal (2006)
6, Number 1, pp. 40–57

Generalized least squares for trend estimation of summarized dose–response data

Nicola Orsini
Karolinska Institutet
Stockholm, Sweden
nicola.orsini@ki.se

Rino Bellocco
Karolinska Institutet
Stockholm, Sweden

Sander Greenland
UCLA School of Public Health
Los Angeles, CA

Abstract. This paper presents a command, `glst`, for trend estimation across different exposure levels for either single or multiple summarized case–control, incidence-rate, and cumulative incidence data. This approach is based on constructing an approximate covariance estimate for the log relative risks and estimating a corrected linear trend using generalized least squares. For trend analysis of multiple studies, `glst` can estimate fixed- and random-effects metaregression models.

Keywords: st0096, glst, dose–response data, generalized least squares, trend, meta-analysis, metaregression

1 Introduction

Epidemiological studies often assess whether the observed relationship between increasing (or decreasing) levels of exposure and the risk (or odds) of diseases follows a linear dose–response pattern. Methods for trend estimation of single and multiple summarized dose–response studies (Berlin, Longnecker, and Greenland 1993) are particularly useful when the full original data are not available.

To demonstrate these methods, our paper uses different types of dose–response data arising from published case–control, incidence-rate, and cumulative incidence data (also see [ST] **epitab**). Summarized data are typically reported as a series of dose-specific relative risks, with one category serving as the common referent group. The term *relative risk* (RR) will be used as a generic term for the risk ratio (cumulative incidence data), rate ratio (incidence-rate data), and odds ratio (case–control data).

Table 1 shows a summary of case–control data investigating the association between the consumption of alcohol and the risk of breast cancer, first presented by Rohan and McMichael (1988), in which it appears that risk of breast cancer increases with increasing levels of alcohol intake.

Table 1: Case–control data on alcohol and breast cancer risk (Rohan and McMichael 1988)

Alcohol (g/d)	Assigned dose (g/d)	No. of cases	No. of controls	Total subjects	Adjusted RR (95% CI)
0	0	165	172	337	1.0 (Referent)
< 2.5	2	74	93	167	0.80 (0.51–1.27)
2.5–9.3	6	90	96	186	1.16 (0.73–1.85)
> 9.3	11	122	90	212	1.57 (0.99–2.51)

Table 2 shows a summary of incidence-rate data investigating the association between the long-term intake of dietary fiber and risk of coronary heart disease among women, first presented by Wolk et al. (1999), which supports the hypothesis that higher fiber intake reduces the risk of coronary heart disease.

Table 2: Incidence-rate data on fiber intake and coronary heart disease risk (Wolk et al. 1999)

Quintile of fiber intake	Assigned dose (g/d)	No. of cases	Person-years	Adjusted RR (95% CI)
1	11.5	148	134,707	1.0 (Referent)
2	14.3	127	133,824	0.98 (0.77–1.24)
3	16.4	114	130,654	0.92 (0.71–1.18)
4	18.8	107	124,522	0.87 (0.66–1.15)
5	22.9	95	117,808	0.77 (0.57–1.04)

Table 3 shows a summary of cumulative incidence data investigating the association between high-fat dairy food intake and risk of colorectal cancer, first presented by Larsson, Bergkvist, and Wolk (2005), which suggests that more servings per day of high-fat dairy food reduces the risk of colorectal cancer.

(*Continued on next page*)

Table 3: Cumulative incidence data on high-fat dairy food and colorectal cancer risk (Larsson, Bergkvist, and Wolk 2005)

High-fat dairy (servings/d)	Assigned dose (servings/d)	No. of cases	Total subjects	Adjusted RR (95% CI)
<1.0	0.5	110	8,103	1.0 (Referent)
1.0– <2.0	1.5	212	17,538	0.75 (0.60–0.96)
2.0– <3.0	2.5	211	15,304	0.74 (0.58–0.95)
3.0– <4.0	3.5	132	9,078	0.68 (0.52–0.90)
≥ 4.0	6.5	133	10,685	0.59 (0.44–0.79)

For each of these summarized tables, we have adjusted relative risks and confidence limits for each nonreference exposure level. The usual approach to trend estimation, namely, the expected change of the log relative risks for a unit change of the exposure level, is to fit a linear regression through the origin, where the response variable is the log relative risks, the assigned dose is the covariate, and the log relative risks are weighted by the inverse of their variances. This method is known as weighted least-squares (WLS) regression (see [R] **vwls**), and it assumes that the log relative risks are independent—an assumption that is never satisfied in practice. The log relative risks are correlated given that they are estimated using a common referent group, and this standard approach underestimates the variance of the slope (Greenland and Longnecker 1992). This problem can be particularly relevant in a meta-analysis of summarized dose–response data where each study slope (trend) is weighted by the inverse of the variance (Shi and Copas 2004).

An efficient estimation method for the slope of a single study is therefore proposed and implemented in the command `glst`, as described by Greenland and Longnecker (1992). This method is then incorporated in the estimation of fixed and random-effects metaregression models for the analysis of multiple studies.

The rest of the article is organized as follows: section 2 introduces the dose–response model and the estimation method; section 3 describes the syntax of the command `glst`; section 4 presents some practical examples based on published data; section 5 compares the corrected and uncorrected methods for trend estimation; and section 6 contains final comments.

2 Method

2.1 Log-linear dose–response model for a single study

It is possible to analyze the shape of the dose–response relationship between reported log relative risks and the exposure levels by estimating a log-linear dose–response regression model (Greenland and Longnecker 1992; Berlin, Longnecker, and Greenland 1993;

Shi and Copas 2004). Assuming that the exposure variable takes value 0 in the reference category, the estimated log relative risk in the reference category is set to zero (log 1); therefore, no intercept models are used. The matrix notation is

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{e} \tag{1}$$

$$\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ \vdots & \vdots & & \vdots \\ x_{i1} & x_{i2} & & x_{ip} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix} \quad \beta = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \quad \mathbf{e} = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_i \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

where $\mathbf{y}$ is an $n \times 1$ vector of (reported) estimated log relative risks; $i = 1, 2, \dots, n$ identifies nonreference exposure levels; $\mathbf{X}$ is an $n \times p$ matrix of nonstochastic covariates, where the first column, denoted by x_{i1}, identifies the exposure variable, and the remaining $p-1$ columns, for instance, may represent transformations of x_{i1}; β is a $p \times 1$ vector of unknown regression coefficients; and $\mathbf{e}$ is an $n \times 1$ vector of random errors, with expected value $E(\mathbf{e}) = 0$ and variance–covariance matrix $\text{Cov}(\mathbf{e}) = E(\mathbf{ee}')$ equal to the following symmetric matrix given by

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

Thus the response variable $\mathbf{y}$ has expected value $E(\mathbf{y}) = \mathbf{X}\beta$ and covariance matrix $\text{Cov}(\mathbf{y}) = \mathbf{\Sigma}$.

2.2 Generalized least squares

We use generalized least squares (GLS) to efficiently estimate the β vector of regression coefficients in (1). Assuming that the variance–covariance matrix of $\mathbf{e}$ is $\text{Cov}(\mathbf{e}) = \mathbf{\Sigma}$, this method involves minimizing $(\mathbf{y} - \mathbf{X}\beta)'\mathbf{\Sigma}^{-1}(\mathbf{y} - \mathbf{X}\beta)$ with respect to β. Suppose initially that the variance–covariance matrix $\mathbf{\Sigma}$ is known. In matrix notation, the resulting estimator $\mathbf{b}$ of the regression coefficients β is

$$\mathbf{b} = (\mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{y} \tag{2}$$

and the estimated covariance matrix $\mathbf{v}$ of $\mathbf{b}$ is

$$\mathbf{v} = \widehat{\text{Cov}}(\mathbf{b}) = (\mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{X})^{-1} \tag{3}$$

A remarkable property of the GLS estimator is that for any choice of $\mathbf{\Sigma}$, the GLS estimate of β is unbiased; that is, $E(\mathbf{b}) = \beta$.

GLS estimation imposes no distributional assumption for the random errors, $\mathbf{e}$, whereas maximum likelihood (ML) estimation assumes a distribution, and the log-likelihood of the sample observed is then maximized. Under the assumption that random errors are normally distributed with zero mean and variance–covariance matrix $\mathbf{\Sigma}$, i.e., $\mathbf{e} \sim N(0,\, \mathbf{\Sigma})$, the log-likelihood function can be written as the following:

$$l = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{\Sigma}| - \frac{1}{2}\Big\{\big(\mathbf{y} - \mathbf{X}\beta\big)'\mathbf{\Sigma}^{-1}\big(\mathbf{y} - \mathbf{X}\beta\big)\Big\} \tag{4}$$

Maximizing (4) with respect to β is equivalent to solving $\partial l/\partial\beta = 0$. The solution is the ML estimator of β, which under the normality assumption turns out to be the same as the GLS estimator given by (2).

2.3 Statistical inference

To construct confidence intervals and tests of hypotheses about β, we can make direct use of the GLS estimate, $\mathbf{b}$, and its estimated covariance matrix, $\mathbf{v}$. When the normality assumption of the random error $\mathbf{e}$ is introduced, the distributional properties of $\mathbf{y}$ and functions of $\mathbf{y}$ follow at once.

Because $\mathbf{y} \sim N(\mathbf{X}\beta, \mathbf{\Sigma})$, the vector $\mathbf{b}$, which is a linear function of $\mathbf{y}$, is therefore approximately normally distributed $\mathbf{b} \sim N(\beta, \mathbf{v})$.

A test of the null hypothesis, H_0: $\mathbf{b}_j = 0$ versus H_A: $\mathbf{b}_j \neq 0$, can be based on the following Wald statistic,

$$Z = \frac{\mathbf{b}_j}{\sqrt{\mathbf{v}_j}}$$

where $\mathbf{b}_j$ denotes the jth element of the vector $\mathbf{b}$ and $\mathbf{v}_j$ denotes the jth diagonal element of $\mathbf{v}$, with $j = 1, 2, \ldots, p$. The Z statistic can be compared with a standard normal distribution.

Wald test–type confidence intervals of β are computed using the large-sample approximation, the z distribution rather than the t distribution, because the estimates, $\mathbf{b}$, are based on a collection of n presumably large groups of subjects rather than n subjects (Grizzle, Starmer, and Koch 1969; Greenland 1987).

2.4 Covariances

In summarized dose–response data, the log relative risks, $\mathbf{y}$, are estimated using a common reference group. Therefore, the elements of $\mathbf{y}$ are not independent and the off-diagonal elements of $\mathbf{\Sigma}$ are not zero (Greenland and Longnecker 1992). This section describes the method and formulas needed to estimate all the elements of $\mathbf{\Sigma}$.

The diagonal element σ_{ii} of $\mathbf{\Sigma}$, the variance of the log relative risk y_i, is estimated from the normal theory–based confidence limits

$$\sigma_{ii} = \Big[\big\{ \log(u_b) - \log(l_b)\big\}/(2 \times z_{\alpha/2})\Big]^2 \tag{5}$$

where u_b and l_b are, respectively, the upper and lower bounds of the reported relative risks, $\exp(y_i)$, and $z_{\alpha/2}$ denotes the $(1-\alpha/2)$-level standard normal deviate (e.g., use 1.96 for 95% confidence interval).

Following the method proposed by Greenland and Longnecker (1992), one way to estimate the off-diagonal elements σ_{ij} of $\mathbf{\Sigma}$, with $i \neq j$, is to assume that the correlations between the unadjusted log relative risks are approximately equal to those of the adjusted log relative risks. Here, besides the log relative risks, their variances, and exposure levels, we also need to know for each exposure level the number of cases and the number of controls for case–control data (table 4), or the number of cases for incidence-rate data (table 5), or the number of cases and noncases for cumulative incidence data (table 6)—information usually available from the publication.

Table 4: Summary of case–control data

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

The off-diagonal elements of $\mathbf{\Sigma}$ can be estimated using the following three-step procedure, where formulas used for steps 1 and 2 change according to the study type: case–control, incidence-rate, or cumulative incidence data.

For case–control data, where we model log odds ratios, the off-diagonal elements σ_{ij} of $\mathbf{\Sigma}$ are computed as follows:

1. Fit cell counts A_i and B_i as modeled in table 4 (which has margin M_1 and N_i), such that

   $$(A_i \times B_0)/(A_0 \times B_i) = \exp(y_i) \tag{6}$$

   where A_i is the fitted number of cases and B_i is the fitted number of controls at each exposure level (see iterative algorithm described in Greenland and Longnecker 1992, appendix 2).

2. For $i \neq j$, estimate the asymptotic correlation, r_{ij}, of y_i and y_j by

$$r_{ij} = s_0/(s_i s_j)^{1/2} \tag{7}$$

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, estimated using (5).

The above method can be easily extended to the analysis of incidence-rate and cumulative incidence data, upon redefinition of terms in (6) and (7).

Table 5: Summary of incidence-rate data

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

For instance, for incidence-rate data, where we model log incidence-rate ratios, fit cell counts A_i as modeled in table 5 such that $(A_i \times N_0)/(A_0 \times N_i) = \exp(y_i)$. In (7), we redefine $s_0 = (1/A_0)$ and $s_i = (1/A_i + 1/A_0)$.

Table 6: Summary of cumulative incidence data

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

Then, for cumulative incidence data, where we model log risk ratios, fit cell counts A_i as modeled in table 6 such that $(A_i \times N_0)/(A_0 \times N_i) = \exp(y_i)$. In (7), again s_0 and s_1 need to be computed differently: $s_0 = (1/A_0 - 1/N_0)$ and $s_i = (1/A_i - 1/N_i + 1/A_0 - 1/N_0)$.

2.5 Heterogeneity

The analysis of the estimated residual vector $\widehat{\mathbf{e}} = \mathbf{y} - \mathbf{Xb}$ is useful to evaluate how close reported and fitted log relative risks are at each exposure level. A statistic for the goodness of fit of the model is

$$Q = (\mathbf{y} - \mathbf{Xb})'\mathbf{\Sigma}^{-1}(\mathbf{y} - \mathbf{Xb}) \tag{8}$$

where Q has approximately, under the null hypothesis that the fitted model is correct, a χ^2 distribution with $n - p$ degrees of freedom. If the p-value derived from this statistic is small, we may infer that there is some problem with the model; e.g., perhaps heterogeneity is present or there is some unaccounted-for bias. If, however, the p-value is large, we can conclude only that the test did not detect a problem with the model, not that there is no problem. The Q statistic (like most fit statistics) has low power; i.e., its sensitivity to model problems is limited.

2.6 Log-linear dose–response model for multiple studies

The method discussed in the previous section can be applied to estimate the underlying trend from multiple summarized data. When dealing with multiple studies and multiple exposure levels, a more flexible method of trend estimation requires pooling the study data before estimating the dose–response model (Greenland and Longnecker 1992).

In a meta-analysis of dose–response studies, heterogeneity means that the shape or slope of the dose–response relationship varies among studies (Berlin, Longnecker, and Greenland 1993). The pool-first method increases the number of the log relative risks and dose values available for the analysis and it allows either to get a better fit of the dose–response relationship, by including fractional polynomials and splines in $\mathbf{X}$, or to identify sources of heterogeneity across studies, by including effect modifiers in $\mathbf{X}$.

Fixed-effects dose–response metaregression model

Let $\mathbf{y}_k$ be the $n_k \times 1$ response vector and let $\mathbf{X}_k$ be the $n_k \times p$ covariates matrix for the kth study, with $k = 1, 2, \ldots, S$. The number of nonreference exposure levels, n_k, for the kth study might vary among the S studies. We pool the data by concatenating the matrices $\mathbf{y}_k$ and $\mathbf{X}_k$

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

so the outcome $\mathbf{y}$ will be an $n \times 1$ vector, where $n = \sum_{k=1}^{S} n_k$, and the linear predictor $\mathbf{X}$ will be an $n \times p$ matrix.

Using the pool-first method, the log-linear model

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{e} \tag{9}$$

becomes a fixed-effects dose–response metaregression model, where now the vector of random errors, $\mathbf{e}$, has expected value $E(\mathbf{e}) = 0$ and covariance $\mathrm{Cov}(\mathbf{e}) = E(\mathbf{e}\mathbf{e}')$ equal to the following symmetric $n \times n$ block-diagonal matrix,

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_1 & & & & \\ \vdots & \ddots & & & \\ \mathbf{0} & & \boldsymbol{\Sigma}_k & & \\ \vdots & & & \ddots & \\ 0 & \ldots & \mathbf{0} & \ldots & \boldsymbol{\Sigma}_S \end{bmatrix} \tag{10}$$

where $\boldsymbol{\Sigma}_k$ is the $n_k \times n_k$ estimated covariance matrix for the kth study. We assume that the log relative risks are correlated within each study but uncorrelated across different studies.

The GLS estimators are given by (2) and (3), where the variance–covariance matrix is now given by (10). The summary slope (trend) across studies is a weighted average of each study slope with weighting matrix given by the inverse of $\boldsymbol{\Sigma}$.

A test for heterogeneity is again given by (8), where the variance–covariance matrix is given by (10). The Q statistic has approximately, under the null hypothesis, a χ^2 distribution with $n - p$ degrees of freedom.

The assumption implicit in a fixed-effects metaregression model is that each study is estimating the same underlying trend. If heterogeneity is detected then it means that we could fit a better dose–response model, namely, one closer to the observed log relative risks, by either including in the linear predictor transformations of the dose variable and/or interaction terms between exposure dose levels and additional covariates, such as the study design. If important residual heterogeneity is still present after accounting for all known effect modifiers, a random-effect metaregression dose–response model will be necessary to estimate a summary trend across studies (Berlin, Longnecker, and Greenland 1993).

Random-effects dose–response metaregression model

We extend the fixed-effect dose–response model (9) to incorporate residual heterogeneity by including an additive random effect

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\eta + \mathbf{e}$$

where $\mathbf{Z}$ is an $n \times 1$ vector containing the dose variable, first column of $\mathbf{X}$, and η is a random effect with expected value $E(\eta) = 0$ and variance $E(\eta\eta') = \tau^2$, and the random variables η and $\mathbf{e}$ are independent. The τ^2 represents a between-study variance component and quantifies the amount of spread about an overall slope (trend) of the dose variable in the reference category of all covariates specified in $\mathbf{X}$. We estimate the between-study variance using the moment estimator

$$\widehat{\tau}^2 = \frac{Q-(n-p)}{\mathtt{tr}(\mathbf{\Sigma}^{-1}) - \mathtt{tr}\{\mathbf{\Sigma}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Sigma}^{-1}\}}$$

where `tr` denotes the trace of a matrix. A revised variance–covariance matrix, $\mathbf{\Sigma}$, is obtained by replacing the matrices $\mathbf{\Sigma}_k = \mathbf{\Sigma}_k + \widehat{\tau}^2\mathbf{Z}_k\mathbf{Z}'_k$ in the block diagonal matrix (10). The revised matrix $\mathbf{\Sigma}$ is plugged into the GLS estimators $\mathbf{b}$ and $\mathbf{v}$, defined by (2, 3), and into the Q statistic, defined by (8). To get a fully efficient estimator, this procedure is repeated until the difference between successive estimates of $\widehat{\tau}^2$ is less than 10^{-5}. Whenever $\widehat{\tau}^2$ is negative, because $Q < n-p$, it is set to zero. The above iterative GLS method is approximately equivalent to first estimating the slope for each study and then pooling the slopes with a random-effects model (DerSimonian and Laird 1986).

3 The glst command

The estimation command `glst` is written for Stata 9.1, and it uses several inline Mata functions (see [M-5] **intro**).

3.1 Syntax of glst

`glst` *depvar dose* [*indepvars*] [*if*] [*in*], `se(`*stderr*`)` `cov(`*n cases*`)` [[`cc`|`ir`|`ci`] `pfirst(`*id study*`)` `random` `level(`*#*`)` `eform`]

where *depvar*, the outcome variable, contains log relative risks; *dose*, a required covariate, contains the exposure levels; and *indepvars* may contain other covariates, such as transformations of *doses* or interaction terms.

3.2 Options

`se(`*stderr*`)` specifies an estimate of the standard error of *depvar*. `se()` is required.

`cov(`*n cases*`)` specifies the variables containing the information required to fit the covariances among correlated log relative risks. At each exposure level, according to the study type, *n* is the number of subjects (controls plus cases) for case–control data (`cc`); or the total person-time for incidence-rate data (`ir`); or the total number of persons (cases plus noncases) for cumulative incidence data (`ci`). The variable *cases* contains the number of cases at each exposure level.

`cc` specifies case–control data. It is required for trend estimation of a single study unless the option `pfirst(`*id study*`)` is specified.

`ir` specifies incidence-rate data. It is required for trend estimation of a single study unless the option `pfirst(`*id study*`)` is specified.

`ci` specifies cumulative incidence data. It is required for trend estimation of a single study unless the option `pfirst(`*id study*`)` is specified.

`pfirst(`*id study*`)` specifies the pool-first method with multiple summarized studies. The variable *id* is a numeric indicator variable that takes the same value across correlated log relative risks within a study. The variable *study* must take value 1 for case–control, 2 for incidence-rate, and 3 for cumulative incidence study. Within each group of log relative risks, the first observation is assumed to be the referent.

`random` specifies the iterative generalized least squares method to estimate a random-effect metaregression model. Between-study variability of the *dose* coefficient is estimated with the moment estimator.

`level(`#`)` specifies the confidence level, as a percentage, for confidence intervals. The default is `level(95)` or as set by `set level`.

`eform` reports coefficient estimates as $\exp(b)$ rather than b. Standard errors and confidence intervals are similarly transformed.

3.3 Saved results

`glst` saves in `e()`:

Scalars			
`e(N)`	number of observations	`e(df_gf)`	goodness-of-fit degrees of freedom
`e(chi2)`	model χ^2 statistic		
`e(ll)`	log likelihood	`e(chi2_gf)`	goodness-of-fit test
`e(tau2)`	between-study variance τ^2	`e(S)`	number of studies
`e(df_m)`	model degrees of freedom		
Macros			
`e(cmd)`	`glst`	`e(properties)`	`b V`
`e(depvar)`	name of dependent variable		
Matrices			
`e(b)`	coefficient vector	`e(V)`	variance–covariance matrix of the estimators
`e(Sigma)`	$\widehat{\boldsymbol{\Sigma}}$ matrix		
Functions			
`e(sample)`	marks estimation sample		

4 Examples

4.1 Case–control data: Alcohol and breast cancer risk

Consider the case–control data shown in table 1 on alcohol and breast cancer (Rohan and McMichael 1988). We use the dataset containing the summarized information, and we calculate the standard errors of the log relative risks from the reported 95% confidence intervals using (5).

```
. use cc_ex
. gen double se = (logub - loglb)/(2*invnormal(.975))
```

We fit the log-linear dose–response model (1) to regress the log relative risks on the exposure level. The command **glst** fits the covariances and uses the GLS estimator to provide a correct estimate of the linear trend.

```
. glst logrr dose, se(se) cov(n case) cc
Generalized least-squares regression              Number of obs   =       3
Goodness-of-fit chi2(2)    =    1.93              Model chi2(1)   =    4.83
Prob > chi2                =  0.3816              Prob > chi2     =  0.0279

-------------+----------------------------------------------------------------
       logrr |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        dose |   .0454288   .0206639     2.20   0.028     .0049284    .0859293
-------------+----------------------------------------------------------------
```

The command **glst** stores the fitted covariance matrix of the log relative risks in **e(Sigma)**

```
. matrix list e(Sigma)
symmetric e(Sigma)[3,3]
           c1          c2          c3
r1  .05417235
r2  .01881768   .05627467
r3  .01943145   .02068682   .05632754
```

The exponentiated linear trend for a change of 11 g/d of alcohol level is 1.65 (95% CI = 1.06, 2.57).

```
. lincom dose*11, eform
 ( 1)  11 dose = 0
-------------+----------------------------------------------------------------
       logrr |     exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   1.648255   .3746524     2.20   0.028     1.055709    2.573384
-------------+----------------------------------------------------------------
```

The goodness-of-fit p-value ($Q = 1.93$, $\Pr = 0.3816$) is large. Thus this test detected no problems with the fitted model.

4.2 Incidence-rate data: Fiber intake and coronary heart disease

Consider now the incidence-rate data shown in table 2 on long-term intake of dietary fiber and risk of coronary heart disease among women (Wolk et al. 1999). As we did for case–control data, we use the command **glst** to get an efficient estimate of the slope.

```
. use ir_ex
. gen double se = (logub - loglb)/(2*invnormal(.975))
. glst logrr doser, se(se) cov(n case) ir
Generalized least-squares regression                 Number of obs   =       4
Goodness-of-fit chi2(3)    =    0.18                 Model chi2(1)   =    3.47
Prob > chi2                =  0.9809                 Prob > chi2     =  0.0626

       logrr        Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]

       doser    -.0232086   .0124649    -1.86   0.063    -.0476394    .0012221

. lincom doser*10, eform
 ( 1)  10 doser = 0

       logrr       exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]

         (1)     .7928775   .0988316    -1.86   0.063     .6210185    1.012296
```

For a 10-g/d increase in total fiber intake, the rate of coronary heart disease decreased by 21% (RR = 0.79, 95% CI = 0.62, 1.01). The linear trend estimated with the `glst` command on summarized data is very close to the linear trend estimated on full data (68,782) reported in the abstract of the paper (RR = 0.81, 95% CI = 0.66, 0.99).

4.3 Cumulative incidence data: High-fat dairy food intake and colorectal cancer risk

Finally, let's consider now the cumulative incidence data shown in table 3 on high-fat dairy food intake and colorectal cancer risk (Larsson, Bergkvist, and Wolk 2005).

```
. use ci_ex
. gen double se = (logub - loglb)/(2*invnormal(.975))
. glst logrr dose, se(se) cov(n case) ci
Generalized least-squares regression                 Number of obs   =       4
Goodness-of-fit chi2(3)    =    2.56                 Model chi2(1)   =   11.84
Prob > chi2                =  0.4648                 Prob > chi2     =  0.0006

       logrr        Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]

        dose     -.073636   .0214036    -3.44   0.001    -.1155863   -.0316857

. lincom dose*2, eform
 ( 1)  2 dose = 0

       logrr       exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]

         (1)     .8630591   .0369452    -3.44   0.001     .7936024    .9385948
```

Each increment of two servings per day of high-fat dairy foods corresponded to a 14% reduction in the risk of colorectal cancer (RR = 0.86, 95% CI = 0.79, 0.94). Once again, the linear trend estimated with the `glst` command on summarized data is very close to the linear trend estimated on full data (60,708) reported in the abstract of the paper (RR = 0.87, 95% CI = 0.78, 0.96).

4.4 Meta-analysis: Lactose intake and ovarian cancer risk

Earlier we showed how to estimate a linear trend for a single study. Here we show how to use the command `glst` to estimate a summary linear trend across multiple studies. We consider as a motivating example a meta-analysis of epidemiological studies (six case–control and three cohort studies) investigating the association between lactose intake and ovarian cancer risk (Larsson, Orsini, and Wolk 2005).

Fixed-effects dose–response metaregression model

We can easily pool trend estimates across studies with the option `pfirst()`, which specifies the variable names identifying the correlated log relative risks and the type of study (case–control or incidence-rate data).

```
. use ma_ex
. glst logrr dose, se(se) cov(n case) pfirst(id study) eform
Fixed-effects dose-response model               Number of studies   =       9
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      exb(b)   Std. Err.       z    P>|z|      [95% Conf. Interval]

        dose    1.025822   .0248455     1.05   0.293     .9782636    1.075693
```

Overall, there is no evidence of association between milk intake (10 g/d) and risk of ovarian cancer (RR = 1.03, 95% CI = 0.98, 1.08). However, the goodness-of-fit test ($Q = 40.25$, Pr = 0.0486) suggests that we should take into account potential sources of heterogeneity. The estimated association of lactose intake with ovarian cancer risk might depend on the study design. Therefore, we create a product (interaction) term between the type of study (1 for incidence-rate and 0 for case–control data) and the dose variable, and we include it in the model. An alternative would be to stratify the meta-analysis by study design.

```
. gen types = study == 2
. gen doseXtypes = dose*types
```

(*Continued on next page*)

```
. glst logrr dose doseXtypes, se(se) cov(n case) pfirst(id study)
Fixed-effects dose-response model               Number of studies   =       9

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 |   .1550466   .0497982     3.11   0.002     .0574439    .2526492
------------------------------------------------------------------------------

. lincom dose + doseXtypes*0, eform
 ( 1)  dose = 0

------------------------------------------------------------------------------
       logrr |     exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   .9665253   .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
------------------------------------------------------------------------------
```

No association between milk intake and risk of ovarian cancer was found among six case–control studies (RR = 0.97, 95% CI =0.91, 1.03). A positive association between milk intake and risk of ovarian cancer was found among three cohort studies (RR = 1.13, 95% CI = 1.05, 1.22). A systematic difference in slopes related to study design might result, for instance, from the existence of recall bias in the case–control studies that would not be present in the cohort studies. Now the goodness-of-fit test (Q = 30.55, Pr = 0.2453) detects no further problems with the fitted model.

Random-effects dose–response metaregression model

We can also check residual heterogeneity across linear trend estimates by fitting a random-effects model.

```
. glst logrr dose doseXtypes, se(se) cov(n case) pfirst(id study) random
Random-effects dose-response model              Number of studies   =       9

Iterative Generalized least-squares regression      Number of obs   =      28
Goodness-of-fit chi2(26)    =   28.37               Model chi2(2)   =    7.29
Prob > chi2                 =  0.3407               Prob > chi2     =  0.0261

------------------------------------------------------------------------------
       logrr |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        dose |  -.0443064   .0394422    -1.12   0.261    -.1216116    .0329988
  doseXtypes |   .1654426    .063171     2.62   0.009     .0416297    .2892555
------------------------------------------------------------------------------
Moment-based estimate of between-study variance of the slope: tau2  =   0.0026
```

The trend estimates for case–control and cohort studies are quite close to the previous ones under fixed-effects models. The between-study standard deviation is close to zero ($\widehat{\tau} = 0.0026^{1/2} = 0.05$), which implies that the study-specific trends have only a small spread around the average trend (-0.044) for case–control studies. Furthermore, if we model heterogeneity directly with a random-effects model, without considering any effect modifiers, the results of the meta-analysis briefly described above could not be achieved at all.

```
. glst logrr dose, se(se) cov(n case) pfirst(id study) eform random
Random-effects dose-response model                Number of studies   =        9

Iterative Generalized least-squares regression          Number of obs   =       28
Goodness-of-fit chi2(27)    =   32.17                   Model chi2(1)   =     0.20
Prob > chi2                 =  0.2259                   Prob > chi2     =   0.6519

       logrr       exb(b)   Std. Err.       z    P>|z|      [95% Conf. Interval]

        dose     1.016753   .0374417     0.45   0.652      .9459546    1.092851

Moment-based estimate of between-study variance of the slope: tau2  =    0.0059
```

We would simply conclude that, overall, there is no association between lactose intake on ovarian cancer risk (RR = 1.02, 95% CI = 0.95, 1.09).

5 Empirical comparison of the WLS and GLS estimates

Here we compare and evaluate the uncorrected (WLS) and corrected (GLS) estimates of the linear trend, `b`, its standard error, `se` $= \sqrt{\mathbf{v}}$, and the heterogeneity statistic, Q. Table 7 summarizes the results for single (sections 4.1–4.3) and multiple studies (section 4.4)

Table 7: Empirical comparison of GLS and WLS estimates

	GLS			WLS			Difference (%)		
	`b`	`se`	Q	`b`	`se`	Q	`b`	`se`	Q
Single study									
Case–control	0.045	0.021	1.93	0.033	0.019	1.72	26.4	9.5	10.5
Incidence-rate	−0.008	0.006	1.61	−0.007	0.004	0.93	14.6	33.7	42.2
Cumulative incidence	−0.073	0.021	2.57	−0.098	0.018	2.20	−33.2	15.6	14.1
Multiple studies									
Case–control	−0.034	0.031	24.02	−0.042	0.026	30.48	−23.1	17.2	−26.9
Incidence-rate	0.121	0.039	6.54	0.142	0.033	3.24	−17.0	15.0	50.5
Overall	0.025	0.024	40.25	0.026	0.020	52.90	−3.2	16.4	−31.4

The relative differences, expressed as percentages, between the GLS and WLS estimates are calculated as (GLS − WLS)/GLS ×100. The GLS estimates of the linear trend, `b`, could be higher or lower than the WLS estimates, and the small differences are not surprising because both estimators are consistent (Greenland and Longnecker 1992). The Q statistic based on GLS estimates could be higher or lower than the one based on WLS estimates. In the WLS procedure the off-diagonal elements of $\mathbf{\Sigma}$, covariances among log relative risks, are set to zeros, whereas in the GLS the covariances are not zeros (see section 2.4). Therefore, the weighting matrix, $\mathbf{\Sigma}^{-1}$, in the Q statistic depends both on variances and covariances of the log relative risks. As expected, the GLS estimates of the standard errors, `se`, are always higher than the WLS estimates of the standard errors for single and multiple studies. The underestimation of the standard error of the uncorrected WLS method somewhat overstates the precision of the trend estimate. Further empirical comparisons between the corrected and uncorrected methods can be found in Greenland and Longnecker (1992).

6 Conclusion

We presented a command, `glst`, to efficiently estimate the trend from summarized epidemiological dose–response data. As shown with several examples, the method can be applied for published case–control, incidence-rate, and cumulative incidence data, from either a single study or multiple studies. In the latter case, the command `glst` fits fixed-effects and random-effects metaregression models to allow a better fit of the dose–response relation and the identification of sources of heterogeneity. Adjusting the standard error of the slope for the within-study covariance is just one of the statistical issues arising in the synthesis of information from different studies. Other important issues, not considered in this paper, are the exposure scale, publication bias, and methodologic bias (Berlin, Longnecker, and Greenland 1993; Shi and Copas 2004; Greenland 2005). A limitation of the method proposed by Greenland and Longnecker (1992) is the assumption that the correlation matrices of the unadjusted and adjusted log relative risks are approximately equal. In future developments of the command, upper and lower bounds of the covariance matrix will be implemented to assess the sensitivity of the GLS estimators, as pointed out by Berrington and Cox (2003).

7 References

Berlin, J. A., M. P. Longnecker, and S. Greenland. 1993. Meta-analysis of epidemiologic dose–response data. *Epidemiology* 4: 218–228.

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

DerSimonian, R., and N. Laird. 1986. Meta-analysis in clinical trials. *Controlled Clinical Trials* 7.

Greenland, S. 1987. Quantitative methods in the review of epidemiologic literature. *Epidemiologic Reviews* 9: 1–30.

———. 2005. Multiple-bias modeling for analysis of observational data (with discussion). *Journal of the Royal Statistical Society, Series A* 168: 267–308.

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

Grizzle, J. E., C. F. Starmer, and G. G. Koch. 1969. Analysis of categorical data by linear models. *Biometrics* 25: 489–504.

Larsson, S. C., L. Bergkvist, and A. Wolk. 2005. High-fat dairy food and conjugated linoleic acid intakes in relation to colorectal cancer incidence in the Swedish Mammography Cohort. *American Journal of Clinical Nutrition* 82: 894–900.

Larsson, S. C., N. Orsini, and A. Wolk. 2005. Milk, milk products and lactose intake and ovarian cancer risk: A meta-analysis of epidemiological studies. *International Journal of Cancer* 118: 431–441.

Rohan, T. E., and A. J. McMichael. 1988. Alcohol consumption and risk of breast cancer. *International Journal of Cancer* 41: 695–699.

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

Wolk, A., J. E. Manson, M. J. Stampfer, G. A. Colditz, F. Hu, F. E. Speizer, C. H. Hennekens, and W. C. Willett. 1999. Long-term intake of dietary fiber and decreased risk of coronary heart disease among women. *Journal of the American Medical Association* 281: 1998–2004.

About the authors

Nicola Orsini is a Ph.D. student, Division of Nutritional Epidemiology, the National Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden.

Rino Bellocco is Associate Professor of Biostatistics, Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden, and Associate Professor of Biostatistics, Department of Statistics, University of Milano Bicocca, Milan, Italy.

Sander Greenland is Professor of Epidemiology, UCLA School of Public Health, and Professor of Statistics, UCLA College of Letters and Science, Los Angeles, CA.