American Journal of Epidemiology
Copyright © 2003 by the Johns Hopkins Bloomberg School of Public Health
All rights reserved

Vol. 157, No. 4
Printed in U.S.A.
DOI: 10.1093/aje/kwf215

PRACTICE OF EPIDEMIOLOGY

Statistical Analysis of Correlated Data Using Generalized Estimating Equations: An Orientation

James A. Hanley[1,2], Abdissa Negassa[3], Michael D. deB. Edwardes[2], and Janet E. Forrester[4]

[1] Department of Epidemiology and Biostatistics, Faculty of Medicine, McGill University, Montreal, Quebec, Canada.
[2] Division of Clinical Epidemiology, Royal Victoria Hospital, Montreal, Quebec, Canada.
[3] Division of Epidemiology and Biostatistics, Department of Epidemiology and Social Medicine, Albert Einstein College of Medicine of Yeshiva University, Bronx, NY.
[4] Department of Family Medicine and Community Health, School of Medicine, Tufts University, Boston, MA.

Received for publication January 7, 2000; accepted for publication August 7, 2002.

The method of generalized estimating equations (GEE) is often used to analyze longitudinal and other correlated response data, particularly if responses are binary. However, few descriptions of the method are accessible to epidemiologists. In this paper, the authors use small worked examples and one real data set, involving both binary and quantitative response data, to help end-users appreciate the essence of the method. The examples are simple enough to see the behind-the-scenes calculations and the essential role of weighted observations, and they allow nonstatisticians to imagine the calculations involved when the GEE method is applied to more complex multivariate data.

correlation; epidemiologic methods; generalized estimating equation; longitudinal studies; odds ratio; statistics

Abbreviation: GEE, generalized estimating equations.

The generalized estimating equations (GEE) (1, 2) method, an extension of the quasi-likelihood approach (3), is being increasingly used to analyze longitudinal (4) and other (5) correlated data, especially when they are binary or in the form of counts. We are aware of only two articles which try to make the GEE approach more accessible to nonstatisticians. One focuses on software (6). The other, an excellent expository article (5) covering several approaches to correlated data, has limited coverage of GEE. Examples in most texts and manuals are too extensive, and the treatment too theoretical, to allow end-users to follow the calculations or fully appreciate the principles behind them. In this paper, we attempt to redress this. To illustrate the ideas, we use the data shown in table 1. They consist of the age- and sex-standardized heights (and data on the covariates gender and socioeconomic status) of 144 children in a sample of 54 randomly selected households in Mexico (7).

Textbooks all advise researchers *not* to treat observations from the same household (or "cluster") as if they were independent and thus not to calculate standard errors using n = 144 as the sample size. For example, Colton (8, pp. 41–43) warns against being misled by "great masses of observations, which upon closer scrutiny, may often vanish," and he uses as an example an n of *800* blood pressure *measurements*—10 taken each week over an 8-week treatment course, in 10 patients! He stresses that "appropriate conclusions regarding the drug's effect rely on subject-to-subject variation, so that the sample size of *10 subjects* is crucial to such analysis." However, few texts explain how one *is* to properly use all 800 (or, in our example, 144) data points, or how much each observation contributes statistically.

Although some articles do discuss how much statistical information is obtainable from observations on paired organs (9) or individuals in clusters such as classrooms or physicians'

Correspondence to Dr. James A. Hanley, Department of Epidemiology and Biostatistics, Faculty of Medicine, McGill University, 1020 Pine Avenue West, Montreal, Quebec H3A 1A2, Canada (e-mail: james.hanley@mcgill.ca).

TABLE 1. Heights (expressed as number of standard deviations above US age- and sex-specific norms) of 144 children in a sample of 54 Mexican households*

SES† of household	Household identifier	Heights‡ of children§			
3	1	–0.76	–0.90	–1.20	–0.93
3	2	**–1.61**			
3	3	–0.78	**–0.96**		
3	4	–3.12	–2.57		
3	5	–0.01	**–0.50**	–0.02	**–0.74**
3	6	**–1.36**	**–0.33**	–0.31	**–0.50**
3	7	–0.80	0.02		
3	8	**–1.03**	**–0.38**	**–1.05**	
3	9	**1.07**	**–1.02**	**–0.57**	**0.76**
3	10	**–1.35**	–1.14		
3	11	**–1.13**	–2.12	**–2.39**	
3	12	**–2.67**	–3.12	**–2.24**	
3	13	–0.53	–1.55		
4	14	**0.36**	**–2.54**		
4	15	**–2.87**	**–1.26**	**–1.22**	
4	16	**–1.51**	–2.68	**–2.24**	
4	17	0.71	**–1.21**	–0.03	
4	18	**–2.00**	–1.14	**–1.29**	
4	19	0.47	**–0.64**		
4	20	–0.92	**–1.64**		
4	21	**1.54**	**0.19**		
5	22	–1.22	–1.11	**–2.49**	
5	23	–2.38	–2.30	–1.24	**–1.96**
5	24	–1.06			
5	25	0.37	**0.29**		
5	26	**–1.61**	**–1.87**	–2.57	**–0.72**
5	27	–1.75	**–0.77**	–2.55	

Table continues

practices (10), investigators often take a conservative approach. In one example, where all eligible children in a household were randomized to the same treatment (11), statistics were computed as if the observations were independent but standard errors were based on the numbers of households. In another (12), where one fourth of the subjects had a sibling in the study, the authors excluded the data obtained from one of the two siblings.

In this expository article, we show how the GEE approach uses weighted combinations of observations to extract the appropriate amount of information from correlated data. We first motivate and introduce the approach using hand calculations on small hypothetical data sets. We use households as clusters, with the letter "h" (household) as a subscript. We use the Greek letters μ and σ and the uppercase letters P, B, and R when referring to a *parameter* (a mean, standard deviation, proportion, or regression or correlation coefficient); and we use the symbol $\bar{y}$ and the lowercase letters p, b, and r for the corresponding *statistic* (empirical value, calculated from a sample) which serves as an estimate of the parameter.

ELEMENTS

Variability of statistics formed from weighted sums or weighted averages of observations—the general case

The instability of a statistic is measured by its variance. Many statistics involve weighted sums of observations or random variables; weights that add to 1 produce weighted averages. In the general case, the variance of a weighted sum of n random variables y_1 to y_n is a sum of n^2 products. These involve 1) the n weights, w_1 to w_n; 2) the n standard deviations, σ_1 to σ_n, of the random variables; and 3) the $n \times n$ matrix of pairwise correlations, $R_{1,1}$ to $R_{n,n}$, of the random variables. As figure 1 illustrates, the variance of a weighted sum or average can be conveniently computed by placing w_1 to w_n and σ_1 to σ_n along both the row and column margins of the $n \times n$ correlation matrix, forming the product

$$w_{\text{row}} \times w_{\text{column}} \times \sigma_{\text{row}} \times \sigma_{\text{column}} \times R_{\text{row,column}}$$

for each {row, column} combination, and then summing these products over the n^2 row-column combinations.

TABLE 1. Continued

SES† of household	Household identifier	Heights‡ of children§			
5	28	–0.99	**0.19**		
5	29	–1.40	**–0.24**	**–2.28**	
5	30	**–2.80**	–2.30	**–2.18**	
5	31	1.10	0.77		
5	32	1.70	–0.31		
5	33	–0.64	–0.40		
5	34	**–1.02**	**–1.04**	**–1.03**	
6	35	0.47	0.56		
6	36	0.28	**–1.06**		
6	37	**–2.05**	–1.73		
6	38	**–1.44**	–2.37	**–2.29**	
6	39	–0.99	**–1.11**		
6	40	**–0.93**	**0.57**		
6	41	–1.93	–0.42	**–0.96**	
6	42	–0.15	–0.65	**–0.53**	
6	43	**–0.18**	–1.56	0.53	–0.33
7	44	–2.31			
7	45	**–1.47**	**0.81**	**1.03**	
7	46	0.93	1.10		
7	47	–0.90	**–1.93**	**–2.78**	–2.66
8	48	–1.22	–1.66	**–0.50**	**–2.70**
		–0.00	–2.26	–2.06	–1.80
		–2.48			
8	49	**0.38**			
8	50	**–1.86**	**–0.43**		
8	51	**–0.85**	2.04		
8	52	**–1.40**	**–2.88**		
8	53	–2.39	**–1.83**		
9	54	1.43	**–1.31**	**–2.59**	

* Source: Forrester et al. (7).

† SES, socioeconomic status.

‡ Expressed as number of standard deviations above US age- and sex-specific norms.

§ **Boldface** denotes female.

For the remainder of this section, we will assume that the σ's are all equal.

Variability of statistics derived from uncorrelated observations

When the observations are uncorrelated, the off-diagonal elements in the correlation matrix are zero. If each of the n weights equals $1/n$, then the weighted sum is the mean, $\bar{y}$. Its variance (the sum of the diagonal elements in part b of figure 1) is thus

$$\mathrm{Var}[\bar{y}] = \left(\frac{1}{n}\right)^2\sigma^2 + \left(\frac{1}{n}\right)^2\sigma^2 + \ldots + \left(\frac{1}{n}\right)^2\sigma^2 = \frac{\sigma^2}{n},$$

yielding the familiar formula $\mathrm{SD}[\bar{y}] = \sigma/\sqrt{n}$.

With equal statistical weights of 1 each, the variance of the simple sum is $\mathrm{Var}[\Sigma y] = n\ \sigma^2$, so that $\mathrm{SD}[\Sigma y] = \sqrt{n}\ \sigma$. Although our main example involves "physical" heights and "statistical" weights, a side example is instructive. Assume that the "physical" weights of elevator-taking adults vary from person to person by, for example, $\sigma = 10$ kg. Then elevators of 16 persons each (i.e., $n = 16$), randomly chosen from among these, will vary from load to load with a standard deviation of (only!) $\sqrt{16}\ (10) = 40$ kg, while the average per person in each load of 16 will vary with a standard deviation of only $10/\sqrt{16} = 2.5$ kg.

Variability of statistics derived from correlated observations

In the elevator example, the "$\sigma/\sqrt{n}$" and "$\sqrt{n}\ \sigma$" laws for the variability of the two statistics do not hold if the variable of interest on sampled individuals tends to be similar from one individual to the other ("co-related")—for example, if

(a)

$$\begin{array}{ccc|ccc}
 & & & y_1 & y_2 & y_3 \\
 & & & w_1 & w_2 & w_3 \\
 & & & \sigma_1 & \sigma_2 & \sigma_3 \\
\hline
y_1 & w_1 & \sigma_1 & R_{11} & R_{12} & R_{13} \\
y_2 & w_2 & \sigma_2 & R_{21} & R_{22} & R_{23} \\
y_3 & w_3 & \sigma_3 & R_{31} & \mathrm{R}_{32} & R_{33}
\end{array}$$

(b)

$$\begin{array}{ccc}
w_1{}^2\sigma_1{}^2 & w_1w_2\sigma_1\sigma_2R_{12} & w_1w_3\sigma_1\sigma_3R_{13} \\
w_1w_2\sigma_1\sigma_2R_{21} & w_2{}^2\sigma_2{}^2 & w_2w_3\sigma_2\sigma_3R_{23} \\
w_1w_3\sigma_1\sigma_3R_{31} & w_2w_3\sigma_2\sigma_3R_{32} & w_3{}^2\sigma_3{}^2
\end{array}$$

FIGURE 1. Calculating the variance of a weighted sum of three correlated random variables, y_1 to y_3. Var[$w_1y_1 + w_2y_2 + w_3y_3$] is a function of 1) the weights, w_1 to w_3, 2) the standard deviations, σ_1 to σ_3, and 3) the matrix of the pairwise correlations, R_{11} to R_{33}—all shown on the left side of the figure (part *a*) ($R_{11} = R_{22} = R_{33} = 1$). The variance of the weighted sum is the sum of the nine products (3 × 3 = 9) shown on the right (part *b*).

elevators are sometimes used by professional football teams and sometimes by ballet dance classes. The variance of a weighted combination of such observations now involves—in addition to the 1's on the diagonal—the pairwise nonzero off-diagonal elements of the correlation matrix.

When the y's of individuals in a cluster are *positively* correlated, as is typical, the additional off-diagonal elements in part b of figure 1 make the standard deviation of the unweighted average $\bar{y}$ *greater* than $\sigma/\sqrt{n}$.

Preamble to GEE: optimal combination of correlated observations

Suppose, for simplicity, that households have either one or two children and that the mean (μ) and standard deviation (σ) of the variable being measured are the same in both types of households (in some applications (see Hoffman et al. (13), p. 440), μ may vary systematically with cluster size, but that situation will not be considered here). Let the correlation of measurements within two-child households be R. Consider the estimation of μ using a measurement on each of three children (n = 3), one from a randomly chosen single-child household and two from a two-child household. The 3 × 3 correlation matrix (figure 2) for the three y's is made up of a 1 × 1 matrix for the response from the singleton, a 2 × 2 matrix for the two responses from the two siblings, and zeroes for pairs of responses from unrelated children. The y's for *some* actual pairs of unrelated children will both be above or below μ, but *on average*, across all possible such pairs, the *expected* product of deviations is zero.

The first three rows of figure 3 list different possible estimators of μ—each one a different weighted average of the three random variables. The first is the "straight" average of the three observations, using weights of one third each. The second estimator discards one of the related observations. The third uses all three, first creating an average of the two related observations and then averaging it and the unrelated observation.

Since all three are "valid" (unbiased) estimates of μ, one can use their relative precision to choose among them. The variance of each (i.e., over *all possible* such samples of three) is given by $\Sigma_{\text{row}}\Sigma_{\text{column}}w_{\text{row}}w_{\text{column}}R_{\text{row,column}}\sigma^2$, where summation is over all nine pairs. Since four of these nine pairwise correlations, and thus the products involving them, are zero, and two others are identical, the variance simplifies to that shown in the third footnote of figure 3. The different sets of weights lead to the different variances shown in the third column of figure 3. From these, a number of lessons emerge: The greater the correlation, the greater the variability of the estimator that gives a weight of one third to each observation (first row); unless there is perfect correlation, the estimator that discards one of the two correlated observations is more variable (i.e., less "efficient") than the others; and the estimator formed by averaging the two correlated observations and then averaging this with the other observation (third row) is less variable than the others in high-correlation situations but more variable than the others in low-correlation situations.

For any given R, there is a less variable estimator than the three considered. Suppose that, relative to a weight of 1 for the observation on the singleton, the weight for the y for each sibling is w, yielding the weighted average

$$\bar{y}_{1:w:w} = \frac{1}{1+2w}y_{\text{singleton}} + \frac{w}{1+2w}y_{\text{sib1}}\frac{1}{1+2w}y_{\text{sib2.}}$$

	$y_{\text{singleton}}$	y_{sib1}	y_{sib2}
$y_{\text{singleton}}$	**1**	0	0
y_{sib1}	0	**1**	***R***
y_{sib2}	0	***R***	**1**

FIGURE 2. Expected (theoretical) correlations of three responses (n = 3), one from a randomly chosen single-child household ($y_{\text{singleton}}$) and two from a two-child household (y_{sib1} and y_{sib2}). The "block-diagonal" pattern is indicated in boldface.

Estimator*	Form	Variance of estimator†			
		In general‡	If $R = 0$	If $R = 0.5$	If $R = 1.0$
$\bar{y}_{1:1:1}$	$\frac{1}{3}\, y_{\text{singleton}} + \frac{1}{3}\, y_{\text{sib1}} + \frac{1}{3}\, y_{\text{sib2}}$	$\frac{3+2R}{9}\,\sigma^2$	$\frac{\sigma^2}{3}$	$\frac{4\sigma^2}{9}$	$\frac{5\sigma^2}{9}$
$\bar{y}_{1:1}$	$\frac{1}{2}\, y_{\text{singleton}} + \frac{1}{2}\, y_{\text{sib1 or sib2}}$	$\frac{1}{2}\,\sigma^2$	$\frac{\sigma^2}{2}$	$\frac{\sigma^2}{2}$	$\frac{\sigma^2}{2}$
$\bar{y}_{2:1:1}$	$\frac{2}{4}\, y_{\text{singleton}} + \frac{1}{4}\, y_{\text{sib1}} + \frac{1}{4}\, y_{\text{sib2}}$	$\frac{3+R}{8}\,\sigma^2$	$\frac{3\sigma^2}{8}$	$\frac{7\sigma^2}{16}$	$\frac{\sigma^2}{2}$
$\bar{y}_{\text{optimal}}$	$\frac{1+R}{3+R}\, y_{\text{singleton}} + \frac{1}{3+R}\, y_{\text{sib1}} + \frac{1}{3+R}\, y_{\text{sib2}}$	$\frac{1+R}{3+R}\,\sigma^2$	$\frac{\sigma^2}{3}$	$\frac{3\sigma^2}{7}$	$\frac{\sigma^2}{2}$

----

* In this example, $\bar{y}_{1:1:1}$ is the "classical" sampling theory estimator; while $\bar{y}_{\text{optimal}}$ is the generalized estimating equations (GEE) estimator.

† R, correlation of responses of pairs of observations from the same household.

‡ Using the expression $(w_{\text{singleton}}^2 + w_{\text{sib1}}^2 + w_{\text{sib2}}^2 + 2\, w_{\text{sib1}} w_{\text{sib2}} R)\, \sigma^2$.

FIGURE 3. Four estimators of μ and their associated variances. Estimators are based on a sample of three observations, one from a randomly chosen single-child household and two from a two-child household. Correlations between pairs of *y*'s are the same as in figure 2.

One can show that its variance, $\sigma^2(1 + 2w^2 + 2Rw^2)/(1 + 2w)^2$, is lowest when $w = 1/(1 + R)$. For example, if $R = 0.5$, the optimal (relative) weights are in the ratio 1:(2/3):(2/3), and the variance is $(3/7)\sigma^2$, smaller than that of the competitors.

More generally, suppose the sample of n consists of several sets of children from 1-, 2-, …, k-child households, with the same μ and the same pairwise within-household correlation R in all households, regardless of size. If the responses are ordered by household, then the $n \times n$ correlation matrix consists of several repetitions of various "block-diagonal" patterns, as in figure 2. One can show by calculus that the optimal weights for combining the responses of individual children from households of sizes 1, 2, 3, ..., k are 1, $1/(1 + R)$, $1/(1 + 2R)$, ..., $1/(1 + \{k - 1\}R)$. These values can be obtained by summing the entries in any row or column of the inverses of the 1 × 1, 2 × 2, ..., $k \times k$ submatrices in the overall $n \times n$ block-diagonal matrix used in the GEE equations (see next subsection).

With data from paired organs, all "clusters" are of size $k = 2$. Rosner and Milton (9) illustrate this idea of "effective sample size" using responses of a person's left and right eyes to the same treatment: If these have a correlation of 0.54, then 200 eyes, two from each of 100 persons, contribute the "statistical equivalent" of one-eye contributions from each of 130 persons ($200 \times 1/(1 + 0.54) = 130$). The closer the correlation is to 1, the closer the effective sample size is to 100.

Estimation by GEE: the "EE" in GEE

In the $n = 3$ example in figures 2 and 3, consisting of just one household of size $k = 1$ and one of size $k = 2$, each y is a separate legitimate (unbiased) estimator, $\hat{\mu}$, of μ (the circumflex or "hat" over μ denotes an estimate of it, calculated from data). As was the practice in the pre-least-squares era (14), one can combine the three separate *estimating equations*: $y_{\text{singleton}} - \hat{\mu} = 0$, $y_{\text{sib1}} - \hat{\mu} = 0$, and $y_{\text{sib2}} - \hat{\mu} = 0$, using the weights $w_{\text{singleton}}$, w_{sib1}, and w_{sib2}, to obtain a *single* estimating equation

$$w_{\text{singleton}}(y_{\text{singleton}} - \hat{\mu}) + w_{\text{sib1}}(y_{\text{sib1}} - \hat{\mu}) + w_{\text{sib2}}(y_{\text{sib2}} - \hat{\mu}) = 0.$$

In this simple case, $\hat{\mu} = \frac{\Sigma wy}{\Sigma w} = \sum \frac{w}{\Sigma w} y$.

In this didactic example, the value of R used to construct the weights was considered "known"; in practice, it must be estimated, along with μ. The process is illustrated in figure 4, using a total of five observations ($n = 5$) from two clusters. Beginning with $R = 0$, one calculates five weights and, from them, an estimate of μ; from the degree of similarity of the within-cluster residuals, one obtains a new estimate, r, of R. The cycle is repeated until the estimates stabilize—that is, until "convergence" is achieved.

The estimating equation for the parameter μ has an obvious form. Equations for multiple regression parameters—representing absolute or relative differences in means, proportions, and rates—are formed by adapting the (iteratively re)weighted least squares equations used to obtain maximum likelihood parameter estimates from uncorrelated responses (3).

Estimation of a proportion (or odds) rather than a mean: the "G" in GEE

Figure 5 shows the GEE estimation of the expected proportion P from 0/3 and 4/5 positive responses in eight

```
                                  Residuals† and their
                                  squares and products

  w*        y                    4      2     −1     −2     −3           (v)‡     (c)§     (r)¶

 0.2       15          4    |   16      8
 0.2       13          2    |           4
 0.2       10         −1    |                  1      2      3
 0.2        9         −2    |                         4      6
 0.2        8         −3    |                                9          8.50     6.33
  μ̂        =11#                                                                  r =  0.74

                                3.6    1.6   −1.4   −2.4   −3.4
 0.57      15        3.6    |  12.96   5.76
 0.57      13        1.6    |          2.56
 0.40      10       −1.4    |                1.96   3.36   4.76
 0.40       9       −2.4    |                       5.76   8.16
 0.40       8       −3.4    |                             11.56         8.70     7.35
  μ̂        =11.4#                                                                r =  0.84
 . . .     . . .                                                     . . .    . . .    . . .
 . . .     . . .                                                     . . .    . . .    . . .
  μ̂        =11.47#    SE**(model-based) = 2.0††                      8.77      r =  0.86
                      SE(empirical)     = 1.8‡‡
---------------------
```

* Weights are in the ratio $1/(1 + r)$: $1/(1 + r)$: $1/(1 + 2r)$: $1/(1 + 2r)$: $1/(1 + 2r)$; first cycle uses $r = 0$. Final weights: 0.54 × 2 (subtotal 1.08) and 0.37 × 3 (subtotal 1.11); effective sample size: 2.19.

† Residual = $y - \hat{\mu}$.

‡ Estimated variance = sum of 5 squared deviations (on diagonal), divided by 4 (= 5 – 1).

§ Estimated covariance = sum of 4 off-diagonal products, divided by 3 (= 4 – 1).

¶ Estimated correlation = estimated covariance divided by estimated variance.

$\hat{\mu} = \Sigma wy/\Sigma w$.

** SE, standard error.

†† Model-based SE = (estimated unit variance/effective sample size)$^{1/2}$ = $(8.77/2.19)^{1/2}$.

‡‡ Cluster means: 14 and 9; empirical SE = $[\{1.08^2(14 - 11.47)^2 + 1.11^2(9 - 11.47)^2\}/2.19^2]^{1/2}$.

FIGURE 4. Generalized estimating equations estimation of a mean μ and correlation R in a simplified hypothetical example with $n = 2 + 3$ from clusters of size 2 and 3. Shown are the first two cycles and the results of the final cycle. To simplify the display, numbers were rounded after each calculation. See the Appendix for the SAS and Stata statements used to produce these estimates.

subjects in two households. The weights are $1/(1 + 2R)$ and $1/(1 + 4R)$ for the individuals in households of size three and five, respectively. The final estimate of P is $p = 0.42$, corresponding to $r = 0.45$. It is a weighed average of the eight 0's and 1's, with weights of $1/(1 + 2r) = 0.53$ for each of the three responses in household 1 and $1/(1 + 4r) = 0.36$ for each of the five responses in household 2; that is,

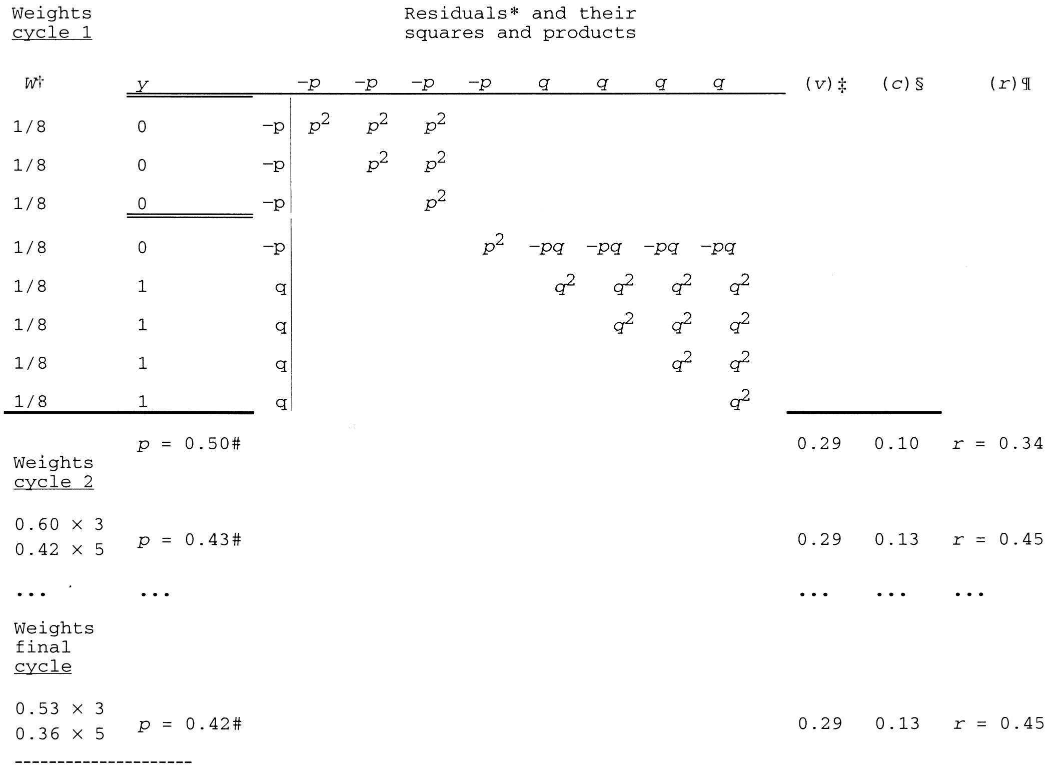
$$p = \frac{0 \times 0.53 \times 3 + \{0 \times 0.36 \times 1 + 1 \times 0.36 \times 4\}}{0.53 \times 3 + 0.36 \times 5}$$

$$= \frac{1.44}{3.39} = 0.42.$$

The sum of the eight weights, 0.53 each for the three persons in household 1 and 0.36 each for the five persons in household 2, can be viewed as the "effective" sample size of 3.39. Estimation of logit[P] = log[$P/(1 - P)$] involves the same core calculations.

If P is different for different covariate patterns or strata, then the "unit" variance $\sigma^2 = P(1 - P)$ is no longer homogeneous. Nonconstant variances can be allowed for by incorporating a function of σ^2 into the weight for each observation (this is the basis of the iteratively reweighted least squares algorithm used with the usual logistic regression for uncorrelated responses).

Indeed, using different weights for each of n uncorrelated outcomes allows a unified approach to the maximum likeli-



Weights cycle 1			Residuals* and their squares and products										
W†	y		−p	−p	−p	−p	q	q	q	q	(v)‡	(c)§	(r)¶
1/8	0	−p	p^2	p^2	p^2								
1/8	0	−p		p^2	p^2								
1/8	0	−p			p^2								
1/8	0	−p				p^2	−pq	−pq	−pq	−pq			
1/8	1	q					q^2	q^2	q^2	q^2			
1/8	1	q						q^2	q^2	q^2			
1/8	1	q							q^2	q^2			
1/8	1	q								q^2			
	p = 0.50#										0.29	0.10	r = 0.34
Weights cycle 2													
0.60 × 3 0.42 × 5	p = 0.43#										0.29	0.13	r = 0.45
...	...										...	...	...
Weights final cycle													
0.53 × 3 0.36 × 5	p = 0.42#										0.29	0.13	r = 0.45

----

* Residual = $y - p = 1 - p = q$ when $y = 1$; $= 0 - p = -p$ when $y = 0$.

† Weights for the first 3 and last 5 observations are in the ratio $1/(1 + 2R)$: $1/(1 + 4R)$; in the first cycle, $R = 0$.

‡ Estimated variance = sum of 8 squared deviations (on diagonal), divided by 7 (= 8 − 1); that is, var = $(4p^2 + 4q^2)/7$.

§ Estimated covariance = sum of 13 off-diagonal products, divided by 12 (= 13 − 1); that is, covariance = $(3p^2 - 4pq + 6q^2)/12$.

¶ Estimated correlation = estimated covariance divided by estimated variance.

$p = \Sigma wy/\Sigma w$; after the final iteration, $\Sigma w = 0.53 \times 3 + 0.36 \times 5 = 3.39$ = "effective" sample size.

FIGURE 5. Generalized estimating equations estimation of a proportion *P* and correlation *R* in a simplified hypothetical example with two clusters of size 3 and 5, in which proportions of positive responses are 0/3 and 4/5. To simplify the display, numbers were rounded after each calculation.

hood estimation of a family of "generalized linear models" (15, 16). Parameters are fitted by minimizing the weighted sum of squared residuals, using functions of the σ^2's as weights. For binomial and Poisson responses, where σ^2 is a function of the mean, weights are reestimated after each iteration. With GLIM software (17), Wacholder (18) illustrated how the risk difference, risk ratio, and odds ratio are estimated using the identity, log, and logit "links," respectively. This unified approach to *un*correlated responses has since become available in most other statistical packages. GEE implementations for *correlated* data use this same unified approach but use a quasi-likelihood rather than a full likelihood approach (3). Since correct specification of the mean and variance functions is sufficient for unbiased estimates, the model used does not fully specify the distribution of the responses in each cluster.

Standard errors: model-based or data-based (empirical)?

Two versions of the standard error are available for accompanying GEE estimates. The difference between them can be illustrated using the previously cited estimate, p = 0.42, of the parameter *P*. The "model-based" standard error is based on the estimated (exchangeable) correlation r = 0.45. This in turn implies the "effective sample size" of 3.39 ($\Sigma w = 3 \times 0.53 + 5 \times 0.36 = 1.59 + 1.80 = 3.39$) shown above

	(a)			(b)		
Method/model	$\hat{\mu}$	r	SE*($\hat{\mu}$)	p	r	SE(p)
SRS*	–1.06		0.11	55.6%		5.2%
Cluster sample†	–1.06		0.16	55.6%		7.2%
GEE*	–1.02‡	0.50‡	0.16$^{\underline{m}}$§ 0.16$^{\underline{e}}$¶	54.2%	0.45	7.0%$^{\underline{m}}$§ 7.2%$^{\underline{e}}$¶

* SE, standard error; SRS; simple random sample (i.e., ignoring the clusters); GEE, generalized estimating equations.

† Estimators from sampling theory (22).

‡ Obtained from Stata xtgee (version 3.3.0; Stata Corporation, College Station, Texas, October 10, 1997) with Gaussian and binomial variation, identity link, exchangeable correlation, and model-based standard errors. Some software implementations (e.g., Stata and SAS GENMOD (SAS Institute, Inc., Cary, North Carolina)) allow a choice of links for a given distribution, so that one can directly obtain p by using the Identity link. Other GEE solvers (e.g., SPlus (MathSoft, Cambridge, Massachusetts)) limit the choice to the canonical link (Identity for Gaussian, Logit for binomial, and Log for Poisson). Thus, for example, if b is an estimate of logit[P], then the estimate of P is $p = \exp[b]/(1 + \exp[b])$.

§ SE$^{\underline{m}}$, model-based SE.

¶ SE$^{\underline{e}}$, empirical SE.

FIGURE 6. Estimates of (part *a*) mean height μ (measured as the number of standard deviations above US norms) and (part *b*) the proportion *P* of short children calculated using data from households with a socioeconomic status index of 5 or lower (see table 1).

and in the last footnote of figure 5. Thus, based on the binomial model,

$$\begin{aligned} \mathrm{SE}_{\text{model-based}}(p) &= \{p \times (1 - p)/\Sigma w\}^{1/2} \\ &= \{0.42 \times 0.58/3.39\}^{1/2} = 0.27. \end{aligned}$$

The "empirical" or "robust" standard error uses the actual variations in the cluster-level statistics, that is, the $p_1 = 0/3 = 0$ and $p_2 = 4/5 = 0.8$, and the "effective sizes" of the subsamples

$$\begin{aligned} \mathrm{SE}_{\text{empirical}}(p) &= \{[1.59^2(0 - 0.42)^2 \\ &\quad + 1.80^2(0.8 - 0.42)^2]/3.39^2\}^{1/2} = 0.28. \end{aligned}$$

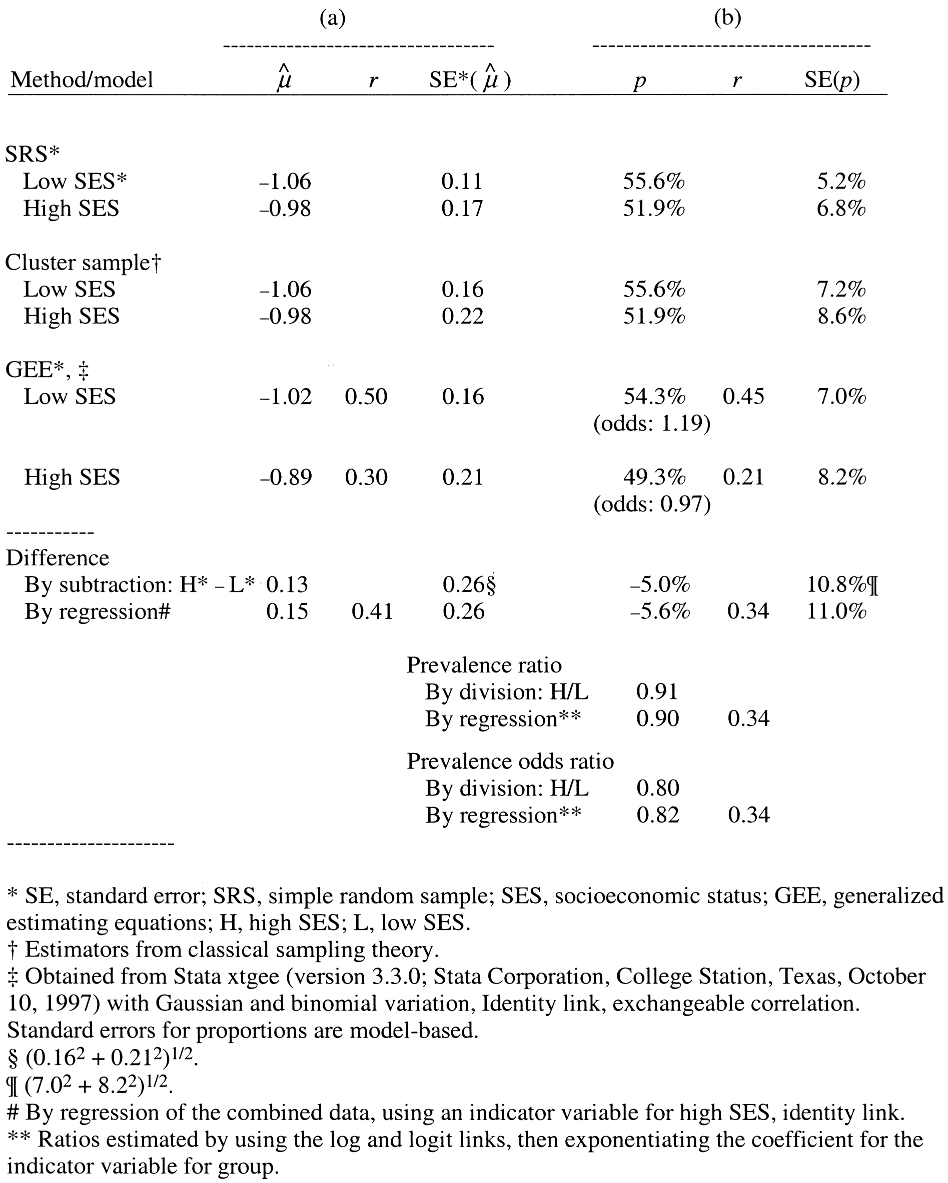
Unless data are sparse, the empirical standard error is likely to be more trustworthy than the model-based one. Agreement between the model-based and empirical standard errors suggests that the assumed correlation structure is reasonable. However, the robust variance estimator, also known as the "sandwich" estimator, was developed for uncorrelated observations, and its theoretical behavior with correlated data has only recently received attention (19). Methods designed to improve on the poor performance in small samples (20) include bias-correction and explicit small-sample adjustments, that is, use of t rather than z (21). A second concern has been the case in which the cluster size itself is related to the outcome and so is "nonignorable." In such instances, within-cluster resampling, coupled with the use of a generalized linear model for uncorrelated data (13), provides more valid confidence intervals than GEE.

APPLICATION

Figure 6 shows—for the lower socioeconomic status group in table 1—the various estimates of $\hat{\mu}$, the average z score, and p, the proportion of children with z scores less than –1. The GEE estimate $\hat{\mu} = -1.02$, based on an estimated r of 0.50, is a weighted average, with heights of children in households of size 1, 2, 3, ... receiving weights of 1, 0.67, 0.50, ... Thus, for estimating μ, the 90 children (in 34 households) constitute an "effective sample size" of 48.3 "unrelated" individuals. The proportion $p = 0.542$ is obtained similarly, with weights calculated from $r = 0.45$.

The model-based and empirical standard errors agree to two decimal places in the case of $\hat{\mu}_{\text{GEE}}$ and differ only slightly (7.0 percent vs. 7.2 percent) in the case of the proportion p. Of interest is the fact that the empirical standard error of 7.2 percent is identical to that calculated from the variance formula for a proportion estimated from a cluster sample in the classic survey sample textbook (22).

Figure 7 contrasts the Low and High socioeconomic status groups with respect to μ, mean height, and P, the proportion of children with z scores less than –1. We can estimate a difference by subtracting the specific estimates, and we can estimate its standard error from the rules for the variance of a difference between two independent estimates. Alternatively, the difference can be estimated as the coefficient of



	(a)			(b)		
Method/model	$\hat{\mu}$	r	SE*($\hat{\mu}$)	p	r	SE(p)
SRS*						
Low SES*	–1.06		0.11	55.6%		5.2%
High SES	–0.98		0.17	51.9%		6.8%
Cluster sample†						
Low SES	–1.06		0.16	55.6%		7.2%
High SES	–0.98		0.22	51.9%		8.6%
GEE*, ‡						
Low SES	–1.02	0.50	0.16	54.3% (odds: 1.19)	0.45	7.0%
High SES	–0.89	0.30	0.21	49.3% (odds: 0.97)	0.21	8.2%

Difference						
By subtraction: H* – L*	0.13		0.26§	–5.0%		10.8%¶
By regression#	0.15	0.41	0.26	–5.6%	0.34	11.0%
			Prevalence ratio			
			By division: H/L	0.91		
			By regression**	0.90	0.34	
			Prevalence odds ratio			
			By division: H/L	0.80		
			By regression**	0.82	0.34	
---------------------						

* SE, standard error; SRS, simple random sample; SES, socioeconomic status; GEE, generalized estimating equations; H, high SES; L, low SES.
† Estimators from classical sampling theory.
‡ Obtained from Stata xtgee (version 3.3.0; Stata Corporation, College Station, Texas, October 10, 1997) with Gaussian and binomial variation, Identity link, exchangeable correlation. Standard errors for proportions are model-based.
§ $(0.16^2 + 0.21^2)^{1/2}$.
¶ $(7.0^2 + 8.2^2)^{1/2}$.
By regression of the combined data, using an indicator variable for high SES, identity link.
** Ratios estimated by using the log and logit links, then exponentiating the coefficient for the indicator variable for group.

FIGURE 7. Comparison of (part *a*) estimated mean height μ (measured as the number of standard deviations above US norms) and (part *b*) the proportion *P* of short children among children of lower and higher socioeconomic status.

the indicator variable I_{High} (1 if high socioeconomic status, 0 if not) in a regression model applied to the combined data. For height measured quantitatively, the intercept represents the mean of low socioeconomic status children, and the coefficient of I_{High} represents the L-H difference in means.

GEE estimates of the proportions are shown in the right half of figure 7. The proportions are compared using various regression forms applied to the combined data. The slight discrepancy between the difference of the separately estimated group-specific proportions and the difference obtained directly from the regression model stems from the fact that the latter uses a common covariance rather than two separate covariances. The 11.0 percent standard error of the difference in proportions, calculated using the pooled covariances, and the $(7.02 + 8.22)^{1/2} = 10.8$ percent obtained from the two separate standard errors are nearly identical.

In the above examples, groups can be compared directly. However, to assess trends in responses over levels of one or more quantitative variables measured at a cluster level (here, household level), a regression approach is more practical. Since GEE analysis is carried out at the child level, it can also include covariates, such as illness histories, that differ from child to child within a household.

DISCUSSION

This orientation focused on correlated data arising from the relatedness of several individuals in the same cluster,

TABLE 2. Meaning of parameters in models fitted by means of generalized estimating equations: comparisons of response proportions, P_0 and P_1, in a hypothetical example with extreme variation in P_0 from some clusters to others*

	Response proportion (%) in those with the factor absent (P_0)	Response proportion (%) in those with the factor present (P_1)	Comparative parameter	
			Difference (%)	Odds ratio
Clusters†				
1 to N/2	10	50	40	9.0
N/2 + 1 to N	50	90	40	9.0
Summary measures				
"Crude" (from aggregated data)	30	70	40	5.4
Mantel-Haenszel			40	9.0
Logistic regression‡				9.0
Generalized estimating equations§			40	5.4

* This example was modified from that of Gail et al. (27).

† For the sake of illustration, all clusters were taken to be of equal size; the factor is present in half of the individuals in each cluster.

‡ Obtained via the unconditional model logit(P) = $B_0 + B_1 \times$ factor + $C_1 \times I_1 \ldots + C_h \times I_h \ldots + C_N \times I_N$, where I_h is an indicator variable for cluster h and C_1 to C_N are the corresponding regression coefficients (or via conditional logistic regression if cluster sizes are small).

§ Logit (P) = $B_0 + B_1 \times$ factor, with clusters identified as such.

rather than several "longitudinal" observations in the same individual. We chose examples that 1) could also be handled by classical methods and 2) were small enough to hand-calculate the weights induced by the correlations. These weights are used both to generate parameter estimates and to calculate standard errors. Although they are nuisance parameters, the correlations do provide for efficient estimates of the primary parameters and for accurate quantification of their precision.

The GEE approach differs in a fundamental conceptual way from the techniques included under the rubric of "random-effects," "multilevel," and "hierarchical" models (e.g., the MIXED and NLMIXED procedures in SAS, MLn (23, 24), or other software described in the paper by Burton et al. (5)). Besides the seeking of more efficient estimators of regression parameters, the main benefit of GEE is the production of reasonably accurate standard errors, hence confidence intervals with the correct coverage rates. The procedures in the other set of techniques explicitly model and estimate the between-cluster variation and incorporate this, and the residual variance, into standard errors. The GEE method does not explicitly model between-cluster *variation*; instead it focuses on and estimates its counterpart, the within-cluster *similarity* of the residuals, and then uses this estimated correlation to reestimate the regression parameters and to calculate standard errors. With GEE, the computational complexity is a function of the size of the largest cluster rather than of the number of clusters—an advantage, and a source of reliable estimates, when there are many small clusters.

However, because the GEE approach does not contain explicit terms for the between-cluster variation, the resulting parameter estimates for the contrast of interest do not have the usual "keeping other factors constant" interpretation. To appreciate this, consider the (admittedly extreme) situation in table 2. If all N clusters are sufficiently large, one can fit an unconditional logistic regression model to the data. If clusters are small, one can avoid fitting one nuisance parameter per cluster (and the consequent bias in the estimated parameter of interest) and instead fit a more economical conditional logistic regression model, using each cluster as a "set." The appropriate logistic regression model "recovers" the common within-cluster ratio of 9, as does the nonregression Mantel-Haenszel approach. However, the GEE approach, with clusters identified as such, yields an odds ratio of only 5.4. The 5.4 contrasts the P_1 for an individual selected randomly from the *population* with the P_0 for another individual selected randomly from the *population*, that is, without "matching" on cluster. In addition, this "population averaged" measure, from the *marginal* model (5) used in the GEE approach, is specific to the mix of clusters studied. In contradistinction to this, the odds ratio of 9 contrasts the P_1 for an individual with the P_0 for another individual from the *same cluster*.

The subtleties of combining ratios, where the rules for "collapsibility" vary with the comparative measure (25), have long been recognized; indeed, the example of combining a 1 percent versus 5 percent contrast in one stratum (odds ratio ~ 5) and a 95 percent versus 99 percent contrast in the other (again, odds ratio ~ 5) was used by Mantel and Haenszel (26, p. 736). Gail et al. (27) used the even more extreme example with odds ratios of 9 (as in table 2) to show how a covariate omitted from a regression analysis can lead to attenuated estimates of what the authors call a "nonlinear" comparative parameter (such as the odds ratio and the hazard ratio), even if—as in table 2—it is "balanced" across the compared levels of the factor.

The above extreme examples are quite hypothetical. In practice, with much less variation in P_0 across clusters, the discrepancy is usually relatively minor. The discrepancy does *not* arise with *absolute differences*, since, with balanced sample sizes, the difference in an aggregate is the aggregate of the within-cluster differences. Table 2 confirms this, showing that the GEE approach, with the *identity* link, accurately recovers the common 40 percent "risk difference" within each cluster.

Unfortunately, as currently implemented in most software, the GEE approach cannot handle several levels of clustering/hierarchy, such as households selected from randomly selected villages that in turn were selected from selected counties. For binary responses, it is possible to use alternating logistic regression (28), an extension of GEE, implemented in S-PLUS, to model different correlations at different levels, but this procedure is not yet available in SPSS, Stata, and SAS implementations of GEE. Likewise, unlike multilevel models, the GEE approach cannot accommodate both cluster-specific intercepts and slopes in longitudinal data.

In our height example, several children within the household are measured *cross-sectionally*, that is, just once, each at a different age. Consider a different study, in which (*unrelated*) children are followed and their heights and covariates are measured at several different ages (times). In such *longitudinal* data, now with the child as the "cluster," *unless the model includes at least a separate intercept for each child*, the successive residual heights of a child will be correlated, with stronger correlations among residuals that are closer together in time. Autoregressive correlation structures are commonly used for longitudinal data. The main analytical challenges are accounting appropriately for missing data and dealing with observations spaced unevenly in time. The reader is referred to the work of Liang and Zeger and colleagues (1, 2, 4) for a treatment of the GEE analysis of quantitative longitudinal data.

ACKNOWLEDGMENTS

This research was supported by the Natural Sciences and Engineering Research Council of Canada (J. H.), grant CA 70269 from the US National Institutes of Health (A. N.), the Fonds de la Recherche en Santé du Québec (M. DeB. E.), and grants 5P01/DK45734-05 and R01 DA/KK11598-01 from the US National Institutes of Health (J. F.).

The authors are grateful to Rolf Heinmueller, Machelle Wilchesky, and several other students for comments on and reactions to the manuscript.

REFERENCES

1. Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika 1986;73:13–22.
2. Zeger SL, Liang KY. The analysis of discrete and continuous longitudinal data. Biometrics 1986;42:121–30.
3. Wedderburn RW. Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. Biometrika 1974; 61:439–47.
4. Diggle PJ, Liang KY, Zeger SL. Analysis of longitudinal data. Oxford, United Kingdom: Oxford University Press, 1994.
5. Burton P, Gurrin L, Sly P. Extending the simple linear regression model to account for correlated responses: an introduction to generalized estimating equations and multi-level mixed modelling. Stat Med 1998;17:1261–91.
6. Horton NJ, Lipsitz SR. Review of software to fit generalized estimating equation regression models. Am Stat 1999;53:160–9.
7. Forrester JE, Scott ME, Bundy DA, et al. Predisposition of individuals and families in Mexico to heavy infection with *Ascaris lumbricoides* and *Trichuris trichiura*. Trans R Soc Trop Med Hyg 1990;84:272–6.
8. Colton T. Statistics in medicine. Boston, MA: Little, Brown and Company, 1974.
9. Rosner B, Milton RC. Significance testing for correlated binary outcome data. Biometrics1988;44:505–12.
10. Donner A, Klar N. Methods for comparing event rates in intervention studies when the unit of allocation is a cluster. Am J Epidemiol 1994;140:279–89.
11. Stansfield SK, Pierre-Louis M, Lerebours G, et al. Vitamin A supplementation and increased prevalence of childhood diarrhoea and acute respiratory infections. Lancet 1993;342:578–82.
12. Gillman MD, Rifas-Shiman SL, Frazier AL, et al. Family dinner and diet quality among older children and adolescents. Arch Fam Med 2000;9:235–40.
13. Hoffman EB, Sen PK, Weinberg CR. Within-cluster resampling. Biometrika 2001;61:439–47.
14. Stigler SM. Least squares and the combination of observations. In: The history of statistics: the measurement of uncertainty before 1900. Cambridge, MA: Harvard University Press, 1986:11–61.
15. Nelder JA, Wedderburn RW. Generalized linear models. J R Stat Soc Ser A 1972;135:370–84.
16. Armitage P, Berry G. Statistical methods in medical research. 3rd ed. Oxford, United Kingdom: Blackwell Scientific Publications, 1994.
17. Numerical Algorithms Group Ltd. GLIM (Generalised Linear Interactive Modelling) software. Oxford, United Kingdom: Numerical Algorithms Group Ltd, 1995.
18. Wacholder S. Binomial regression in GLIM: estimating risk ratios and risk differences. Am J Epidemiol 1986;123:174–84.
19. Edwardes MD. Risk ratio and rate ratio estimation in case-cohort designs: hypertension and cardiovascular mortality. (Letter). Stat Med 1995;14:1609–10.
20. Breslow N. Tests of hypotheses in overdispersed Poisson regression and other quasi-likelihood models. J Am Stat Assoc 1990;85:565–71.
21. Pan W, Wall M. Small-sample adjustments in using the sandwich variance estimator in generalized estimating equations. Stat Med 2002;21:1429–41.
22. Cochran WG. Sampling techniques. New York, NY: John Wiley and Sons, Inc, 1953:124–7,202–5.
23. Goldstein H. Multilevel statistical models. 2nd ed. London, United Kingdom: Edward Arnold, 1995.
24. Breslow N, Leroux B, Platt R. Approximate hierarchical modelling of discrete data in epidemiology. Stat Methods Med Res 1998;7:49–62.
25. Boivin JF, Wacholder S. Conditions for confounding of the risk ratio and of the odds ratio. Am J Epidemiol 1985;121:152–8.
26. Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. J Natl Cancer Inst 1959;22:719–48.
27. Gail MH, Wieand S, Piantadosi S. Biased estimates of treatment effect in randomized experiments with non-linear regressions and omitted covariates. Biometrika 1984;71:431–44.
28. Katz J, Carey VJ, Zeger SL, et al. Estimation of design effects and diarrhea clustering within households and villages. Am J Epidemiol 1993;138:994–1006.

APPENDIX

Stata* and SAS† software statements used to produce the estimates in figures 4–7.

	Stata	SAS (proc GENMOD)

Figure 4
and
Figure 6 (μ) ,
where "y" is height
	XTGEE y,
	LINK(identity) ‡
	Family(normal) ‡
	I(house)
	CORR(exchangeable) ‡
	XTCORR¶
		CLASS house;
		MODEL y = /
		LINK = identity‡
		DIST = normal‡;
		REPEATED SUBJECT=house /
		CORR = exch‡
		MODELSE§
		CORRW;

Figure 5
and
Figure 6 (b)
"y" = 1 if short, 0 if not
	XTGEE y,
	LINK(identity) #
	FAMILY(binomial)
	I(house)
	XTCORR¶
		CLASS house;
		MODEL y = /
		LINK = identity#
		DIST = binomial;
		REPEATED SUBJECT=house /
		CORR = exch
		MODELSE CORRW;

Figures 6 and 7
(a) Means
(b) Proportions
	XTGEE y Ihighses < Imale>,
	I(house)
	XTGEE y Ihighses,
	LINK(identity) ‡
	(b) FAMILY(binomial)
	I(house)
	(b)
	LINK(identity/log/logit)
		CLASS house;
		MODEL y = Ihighses ;
		REPEATED SUBJECT=house /
		CORR = exch
		MODELSE CORRW;
		(a) DIST = normal LINK = identity
		(b) DIST = binomial
		LINK = identity/log/logit;

* Stata Corporation, College Station, Texas.

† SAS Institute, Inc., Cary, North Carolina.

‡ If omitted, identity link, Guassian(normal) variation, and exchangeable correlation assumed by default.

§ If omitted, assumed by default; in Stata, model-based standard errors displayed, unless user specifies "robust."

¶ Used, after fitting, to display fitted correlation matrix.

Use identity link to estimate risk (*P*), logit link for log[*P*/(1-*P*)], log link for log(*P*).