Published by Oxford University Press on behalf of the International Epidemiological Association
International Journal of Epidemiology 2005;**34**:1089–1099
doi:10.1093/ije/dyi153

Regression models for twin studies: a critical review

John B Carlin,[1,2]* Lyle C Gurrin,[3] Jonathan AC Sterne,[4] Ruth Morley,[5] and Terry Dwyer[6]

Accepted 11 July 2005

Twin studies have long been recognized for their value in learning about the aetiology of disease and specifically for their potential for separating genetic effects from environmental effects. The recent upsurge of interest in life-course epidemiology and the study of developmental influences on later health has provided a new impetus to study twins as a source of unique insights. Twins are of special interest because they provide naturally matched pairs where the confounding effects of a large number of potentially causal factors (such as maternal nutrition or gestation length) may be removed by comparisons between twins who share them. The traditional tool of epidemiological 'risk factor analysis' is the regression model, but it is not straightforward to transfer standard regression methods to twin data, because the analysis needs to reflect the paired structure of the data, which induces correlation between twins. This paper reviews the use of more specialized regression methods for twin data, based on generalized least squares or linear mixed models, and explains the relationship between these methods and the commonly used approach of analysing within-twin-pair difference values. Methods and issues of interpretation are illustrated using an example from a recent study of the association between birth weight and cord blood erythropoietin. We focus on the analysis of continuous outcome measures but review additional complexities that arise with binary outcomes. We recommend the use of a general model that includes separate regression coefficients for within-twin-pair and between-pair effects, and provide guidelines for the interpretation of estimates obtained under this model.

Keywords Statistics, twin studies, regression models, data correlation

Epidemiological data analysis concerning risk factors relies heavily on regression models, which seek to 'explain' the variation in an outcome of interest in terms of differences in one or more risk factors (which we may refer to more generally as covariates). Regression modelling provides a flexible set of tools for examining such associations while allowing either for potentially confounding effects of other factors or for interaction effects (effect modification). Epidemiologists are well aware that despite the use of terms such as 'effect' for a regression coefficient, regression modelling only describes associations among variables, so that inferences about causation must be made with extreme caution.[1] A prominent example of the use of regression models in epidemiological studies of the early life origins of disease has been in relating cardiovascular outcomes in later childhood or adulthood to birth weight. Several studies have shown that people with lower birth weight tend to have slightly higher risk of coronary heart disease.[2,3]

The paired structure of data arising from twins presents a number of challenges in the construction, estimation, and interpretation of regression models. Although there has been an explosion of statistical literature on methods for modelling correlated data, there has been relatively little discussion of these methods in the epidemiological literature, and, so far as we are aware, none in the specific context of twins. At the same time, interest in using data from twins to illuminate

[1] Clinical Epidemiology and Biostatistics Unit, Murdoch Children's Research Institute, Royal Children's Hospital, Melbourne, Australia.

[2] Department of Paediatrics and School of Population Health, University of Melbourne, Australia.

[3] Epidemiology and Biostatistics Unit, School of Population Health, University of Melbourne, Australia.

[4] Department of Social Medicine, University of Bristol, UK.

[5] Department of Paediatrics, University of Melbourne, Australia.

[6] Murdoch Childrens Research Institute, Melbourne, Australia.

* Corresponding author. Clinical Epidemiology and Biostatistics Unit, Royal Children's Hospital, Parkville, VIC 3052, Australia. E-mail: jbcarlin@unimelb.edu.au

aetiological questions in life-course epidemiology continues to grow. For example, a recent systematic review identified 10 studies of twins, which examined the association between birth weight and later blood pressure.[4] This review failed to define clearly the meanings of 'unpaired' and 'paired' analyses, and statistical methods used under each of these names differed among these studies.

In this paper, we aim to explain in non-technical terms the distinctions and connections among different regression modelling approaches that may be adopted for twin data, and provide recommendations on how they should be used and interpreted. We focus primarily on the analysis of a continuous outcome measure where the covariate of interest is also continuous, for several reasons. First, this type of analysis has been especially prominent in recent epidemiological studies of twins.[4–10] Second, outcomes of this kind support a more detailed unpacking of associations that may differ when examined within twin pairs and between twin pairs compared with binary or ordinal outcomes. Third, the statistical issues associated with fitting some of the models of interest (using so-called random effects) become more complex with binary outcomes.

For expository purposes, we use an example in which we investigated shared (maternal) and individual (twin) factors that may be predictive of birth weight: thus birth weight is the outcome of interest, rather than a potential predictor of later cardiovascular outcomes as in much of the recent literature. In particular, there is interest in factors that may be associated with growth restriction and/or pre-term delivery. Twin pregnancies offer a unique opportunity to study the contributions of and interactions between shared and individual factors, for example, between maternal factors (common to both twins) and 'supply line' factors (e.g. placental, which may differ between twins). In the present example, we examined the association between cord blood erythropoietin (EPO) and birth weight.[11]

We begin by introducing the dataset, and then we provide a detailed description of three different regression-based approaches to the data, exploring the rationale for the models that underlie the methods. All three approaches have been used and recommended in various combinations in order to investigate the extent to which associations may reflect factors shared by two twins in a pair [e.g. fixed maternal factors such as diet and socioeconomic background, genes for monozygotic (MZ) twins] vs factors that could differ between twins [e.g. fetal nutrient supply line, genes for dizygotic twins (DZ)].[5,7–9] After discussing the results of these three analytic approaches for the example, extensions of the methods to handle a dichotomous outcome are considered. Finally, some general discussion and recommendations for practice are offered, along with further comments on other issues such as the potential use of zygosity information for assessing the importance of genetic effects.

Example: birth weight and EPO

In a recently published analysis,[11] we examined the association between birth weight standard deviation score (or 'Z-score', the birth weight standardized to a reference distribution for gestational age and sex), as outcome, and cord blood EPO, as independent variable. The rationale for modelling birth weight in terms of EPO was that the latter is a marker for hypoxic stress *in utero*, and the investigators were interested in the hypothesis that higher levels of EPO would be associated with growth restriction and, thus, lower birth weight for gestational age. All analyses used $\log_2$ EPO, so that regression coefficients can be interpreted as the mean difference in outcome per doubling of EPO concentration.

In Table 1 we present a range of results from various regression analyses of the birth weight–EPO data, from 110 DZ twin pairs. Initially we ignore the presence of several covariates of interest, although the full analysis focused on possible

Table 1 Estimates of regression coefficients representing linear association between birth weight *Z*-score and $\log_2$ EPO level for 220 DZ twin infants, under alternative regression approaches as described in the text

Parameter	Estimation method	Estimate	SE	*P*-value	95% CI
Model (1)					
β_C	Standard OLS	−0.194	0.0440	<0.001	(−0.281 to −0.107)
	OLS + 'robust' SE	−0.194	0.0396	<0.001	(−0.273 to −0.116)
	GLS/REML	−0.236	0.0454	<0.001	(−0.325 to −0.147)
	GLS/MLE	−0.235	0.0467	<0.001	(−0.327 to −0.144)
	GLS/GEE	−0.235	0.0413	<0.001	(−0.284 to −0.122)
Model (2)					
β_W	GLS/REML	−0.379	0.0691	<0.001	(−0.515 to −0.244)
β_B		−0.133	0.0588	0.024	(−0.248 to −0.018)
β_W	GLS/MLE	−0.379	0.0688	<0.001	(−0.514 to −0.247)
β_B		−0.133	0.0583	0.022	(−0.247 to −0.019)
β_W	GLS/GEE	−0.379	0.0573	<0.001	(−0.492 to −0.267)
β_B		−0.133	0.0453	0.003	(−0.222 to −0.044)
Model (3)					
β_W	OLS through origin	−0.379	0.0691	<0.001	(−0.516 to −0.243)

β_C = common ('twins-as-individuals') regression coefficient; β_W = within-pair regression coefficient; and β_B = between-pair regression coefficient.

effect modification by some of these covariates. We checked that it was reasonable to assume a linear relationship between birth weight and log EPO.

The regression models

To describe the models we introduce some notation. We use Y to denote the outcome and X the covariate of immediate interest, with i used to index twin pairs and j =1, 2 to index individual twins within pairs, so Y_{ij}, X_{ij}, represent, respectively, the outcome and covariate value for the jth twin of the ith pair. The expected value of Y_{ij}, given the value of X_{ij}, is denoted by $E(Y_{ij})$.

(1) Regression on *X* alone (treating twins 'as individuals,' ignoring paired nature of data)

The simplest approach is to use least squares to find the best-fitting values of β_0 and β_C in the following model:

$$E(Y_{ij}) = \beta_0 + \beta_C X_{ij}. \tag{1}$$

In this notation β_C represents the average rate of change in Y for every unit increase in X (based on variation between subjects, so not necessarily applying for an individual). The standard 'ordinary least squares' (OLS) method for fitting model (1) treats all the Y_{ij} values as independent given the corresponding covariate values; see Figure 1 for illustration using the birth weight–EPO data. This makes no allowance for the fact that outcome values from pairs of twins might well tend to be more similar than values from two unrelated individuals, so that they are not statistically independent. OLS estimates are, therefore, presented only as a point of comparison, not as an approach that should actually be used for twin data. Using standard OLS to fit a simple regression model of form (1) has three problems.

First, the standard method for calculating standard errors when using OLS is not correct in the context of twin data. Assuming independence wrongly treats each observation as providing the same information about the regression slope. This may lead to standard errors that are either too large or too small (depending on the underlying correlation structure) and, hence, to confidence intervals (CIs) and *P*-values that are invalid. Methods of estimation that explicitly allow for the correlation structure, such as the generalized least squares (GLS) approaches described below, will produce valid standard errors. For completeness, however, we note that it is possible to produce 'robust' (information-sandwich) standard errors and CIs for OLS estimates, where the correlation structure is acknowledged in calculating standard errors but not used in estimating the regression parameters[12] (see Table 1 for illustration).

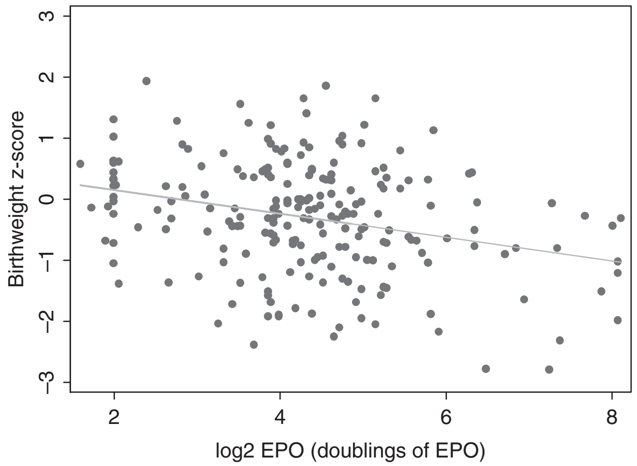

Figure 1 Scatter plot of birth weight *Z*-scores vs $\log_2$ EPO level, for 220 DZ twins, treating all twins as unrelated individuals, with OLS regression line superimposed

The second problem relates to the properties of the OLS estimates of the regression coefficients, not just the standard error estimation. In large samples the OLS estimates are unbiased, so the method is not strictly invalid, if the robust standard error calculation is used. However, it can be improved upon, since the paired structure of the data provides additional information that can contribute to finding a better straight-line fit. Formalization of this argument underlies the GLS method of estimation, which applies differential weighting to the data points in a manner that reflects the correlation structure. A simple model allowing a common correlation within each pair of twins and independence between unrelated individuals is the natural starting point. We use the term GLS to refer to the weighted estimation of regression parameters using particular values for the error variance and correlation—the parameters required to describe the variability and correlation around the mean values determined by Equation (1); see below. These variance–covariance parameters need to be estimated from the data, and particular implementations of 'GLS' differ according to how this estimation is performed, although the practical differences are minimal for twin data.

Two widely used approaches are (i) maximum likelihood estimation (MLE) (or a variation called 'restricted' ML or 'REML') and (ii) generalized estimating equations (GEEs). These approaches yield essentially the same result in the current context of continuous outcome and continuous exposure (e.g. Table 1). The ML approach is sometimes referred to as 'model-based' and in particular said to be based on a 'mixed model,' since the underlying model may be conceived as arising from inclusion of a 'random effect' along with the standard 'fixed' effects in the expected value specification (1). This is often presented as follows:

$$Y_{ij} = \beta_0 + \beta_C X_{ij} + \alpha_i + \varepsilon_{ij}, \tag{1a}$$

where α_i is a 'random' shift in the intercept applying to both twins in pair i and ε_{ij} is the usual individual-specific random error, and each of these is modelled as normally distributed (independently of each other), with zero mean, and variances σ^2_α and σ^2_ε, respectively. This model is completely equivalent to the alternative specification

$$Y_{ij} = \beta_0 + \beta_C X_{ij} + \delta_{ij}, \tag{1b}$$

where δ_{ij} is an error term that has variance $\sigma^2_\alpha + \sigma^2_\varepsilon$ and is not independently distributed within twin pairs; the within-pair correlation is $\sigma^2_\alpha/(\sigma^2_\alpha + \sigma^2_\varepsilon)$. With this normal linear model, the only difference between the MLE and GEE approaches to estimation is that with GEE it is usual to use the 'robust' method for standard error estimation, rather than a model-based method. If the data do not closely follow a normal distribution or the sample size is not large enough, then the resulting standard error estimates may be somewhat different. See Appendix for commands used in the Stata package (Release 9.0, Stata Corporation, College Station, TX) to obtain these and other estimates shown in Table 1. For a tutorial introduction to the concepts of weighted estimation and generalized estimating equations see Hanley *et al.*[13]

(2) Multiple regression: including the co-twin *X* value in the model

The third problem with OLS regression goes beyond issues relating to the methods used for fitting the model: it may be that the specification for the expected value of Y_{ij} is inadequate. Model (1) assumes that the average difference between outcome (Y) values for a given difference between covariate (X) values is the same whether we are comparing two unrelated individuals or two twins. This assumption may be untrue, and to investigate it we need to fit a more general regression model that allows the covariate effect to differ within and between twin pairs. Another way of expressing this is to say that there may be information about the value of Y_{ij} in the co-twin's X value as well as in the subject's own X. Note that this issue, concerning the appropriate specification of the model for the expected value (mean) of Y_{ij} given the X values, is quite distinct from the matter of the statistical technique used to obtain estimates of the model parameters.

An acceptable formulation for a more general model is as follows:

$$E\left(Y_{ij}\right) = \beta_0 + \beta_W\left(X_{ij} - \bar{X}_{i.}\right) + \beta_B\bar{X}_i. \qquad (2)$$

where $\bar{X}_i$ represents the mean value of X for twin pair i. The within-pair coefficient β_W gives the expected change in Y for a one-unit change in the difference between the individual X and the twin-pair average X value, while holding the latter constant. The between-pair coefficient β_B gives the expected change in Y for a one-unit change in the twin-pair average X, while holding the individual deviation from the average constant. Because the two covariates $X_{ij} - \bar{X}_i$ and $\bar{X}_i$ are algebraically derived from the original X values for the ith twin pair, there are many different ways in which this regression model can be expressed and these lead to different possible interpretations.

Some insight into the interpretation of formulation (2) may be obtained by considering alternative possible values of the two coefficients, β_W and β_B. To begin with, suppose that $\beta_W = \beta_B$. In this case the model reduces to model (1) since the effect of the twin-pair mean cancels out, and thus the expected difference in Y for a given difference in X is the same whether we compare within or between twin pairs.

We illustrate this model and other possibilities with a diagram using a small number of 'randomly generated' hypothetical twin pairs, with X and Y distributions similar to those of X = log EPO and Y = birth weight Z-score in the example (except that the strength of the linear trends has been increased to make the distinctions between models clearer). The first row of Figure 2 illustrates model (1), which is the special case of model (2) where β_W and β_B are identical. In the row labelled 2a, the between-pair coefficient β_B is zero and the only systematic source of variation in Y_{ij} is within-pair variation in X: an individual's outcome can be predicted by how different their X value is from their co-twin's, but once this is accounted for the outcome does not depend on the value of X. If this model were to hold it would appear to rule out hypotheses that involve X representing common causal factors (maternal, environmental, and genetic) shared between twins. Row 2b illustrates the other extreme possibility, where the within-pair coefficient β_W is zero and the only systematic variation relates to the 'common' factor represented by the average X value of the two twins. If this model were true, then it would suggest that outcome variation is explained only by between-pair variation in X, although it is important to note that there could well be other within-pair factors operating through pathways unrelated to X. Row 2c illustrates a general case, where both coefficients are non-zero and of the same sign, but not equal.

Given a set of (X, Y) data on twin pairs, there is again a variety of estimation methods that could be applied to fit model (2) and obtain estimates of the two parameters β_W and β_B, along with a test statistic to assess the null hypothesis that β_W and β_B are equal. It turns out that all of the standard methods, including OLS, lead to the same (point) estimates for these regression coefficients.[14] Again, however, valid standard errors are only obtained by methods that respect the paired structure of the data, for instance MLE under a model allowing correlation within pairs, or the robust sandwich variance method. Before trusting the results of any such analysis one should check that the usual assumptions of linear regression are satisfied, in particular that the scale of measurement of Y is such that the additive model specification for between-pair and within-pair effects is reasonable.

(3) Regression using twin-pair difference values

A third, widely used, approach to regression with twin data is based on analysing paired-difference values. We define the differences between X and Y values within each pair by ordering the twins according to birth order, say, leading to $D_i^Y = Y_{i1} - Y_{i2}$ and $D_i^X = X_{i1} - X_{i2}$. If we assume that model (2) holds, then, by subtracting on left-hand and right-hand sides, we find

$$E\left(D_i^Y\right) = \beta_W D_i^X. \qquad (3)$$

Both the constant term β_0 and the between-pair effect $\beta_B\bar{X}_i$ are cancelled out by the subtraction. The resulting difference values are independent of each other so β_W can be estimated by performing OLS regression on the differences, with the constraint that the fitted regression line must pass through the origin, i.e. the point where $D_i^Y = D_i^X = 0$. The same result is obtained irrespective of the basis on which the twins are ordered within pairs (e.g. Figure 3). It is also important to note that the same simple linear regression relationship of difference in Y vs difference in X holds if model (1), viewed as a special case of model (2), is true, i.e. if $\beta_B = \beta_W = \beta_C$. It follows that if the analysis is based solely on within-pair differences then we cannot distinguish between models (1) and (2): the difference analysis only provides information on the within-pair covariate effect.

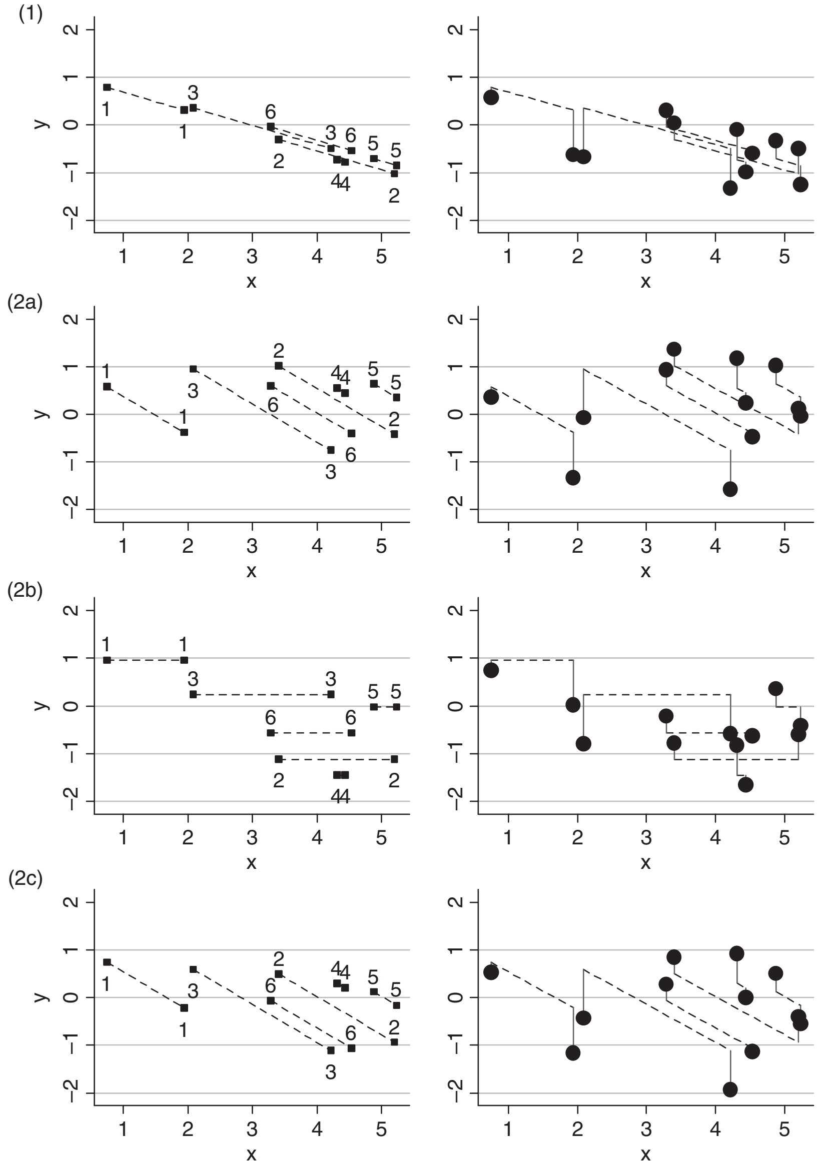

Figure 2 Graphs of expected *Y* vs *X* for 6 hypothetical twin pairs, with a fixed set of *X* values. In each row, the left-hand panel shows expected (mean) value of *Y* according to regression model: (1) simple linear decline with *X*, $\beta_C = -0.4$ [model (1), with jitter added so the lines for each twin can be distinguished from each other]; (2a) $\beta_W = -0.8$, $\beta_B = 0$; (2b) $\beta_W = 0$, $\beta_B = -0.4$; (2c) $\beta_W = -0.8$, $\beta_B = -0.4$. Right-hand panel for each case shows hypothetical (simulated) data where correlated random errors with between-subject and within-subject standard deviation both 0.5 has been added to the mean values in the left-hand panel. To maximize comparability between panels, the same simulated errors are used in each case. Note that by chance some twin pairs have very similar *X* values (e.g. pair 4) while others have quite different values (pairs 2, 3)

Results and interpretation for the example

We display estimates using the three approaches for the EPO-birth weight data from 110 DZ twin pairs in Table 1. In this example the estimated value of β_W is considerably larger (more negative) than that of β_B. A test that the two coefficients are equal reveals strong evidence against the null hypothesis ($P = 0.007$ by likelihood ratio or Wald test), so the adjustment for pair-level factors (by way of the pair mean term) clearly improves the fit of the model. Further insight can be gained from the fact that the twin-specific deviation from the pair mean, $X_{i1} - \bar{X}_i$, can be re-expressed as half the difference between the two twin's *X* values: $X_{i1} - \bar{X}_i = (X_{i1} - X_{i2})/2$, so we are using both twins' values in the model to estimate β_W with a constraint that the effects are equal and opposite for each twin in a pair. The model-fitting results show that using $X_{ij} - \bar{X}_i$ rather than simply the twin's own value X_{ij} as a single predictor provides a better fit to the data, indicating that the relative value of *X* within a twin pair is more useful in predicting *Y* than the absolute level of *X*.

In summary, fitting model (1) for the expected value of *Y* essentially ignores the pairing between the twins, fitting a standard linear regression that treats the twins 'as individuals'. In this example we find an estimated regression coefficient ($\hat{\beta}_C$) that varies from −0.19 to −0.24, depending on the

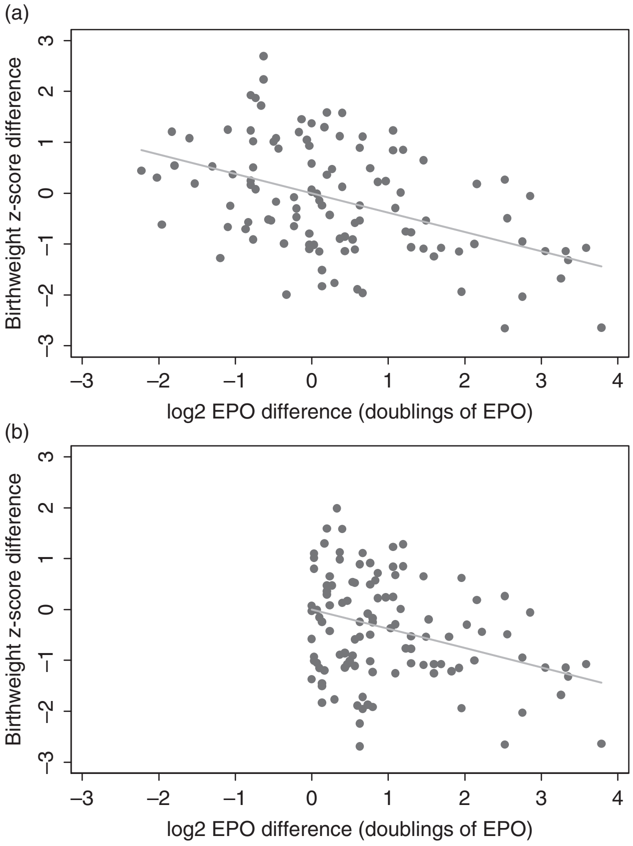

Figure 3 Scatter plot of difference values for birth weight *Z*-score vs $\log_2$ EPO, where difference values were obtained (a) as first-born twin minus second-born twin, and (b) ordered according to EPO value (so the difference in EPO values is always positive). Despite the apparently different scatter, the estimated regression slope, for the line constrained through the origin, is the same from both sets of difference values

estimation method. Approach (2) estimates two regression coefficients, a 'within-pair' and a 'between-pair' effect. For this model the different estimation methods produce the same coefficient estimates, but the key point of interest (in this example) is that the within-pair coefficient is much larger (in absolute value) than the between-pair, while the original single coefficient from Model (1) is midway between the two.[15] This second approach has only appeared recently in twin data analyses.[9,16,17] The third approach (3 in Table 1) is based on reducing the data to paired differences between twins, and regressing the difference in outcome against the difference in the independent variable, with the intercept constrained to zero. Note that the resulting regression coefficient is exactly the same as the within-pair effect from Model (2). The paired-difference method has been much used in the twin research literature.[7,8,18]

Inclusion of other covariates, effect modification

In many twin studies the research question of interest relates not simply to the strength of linear association between outcome and covariate considered across all twins, but to whether a linear association is modified by twin characteristics, including environmental and genetic factors. For example, Morley *et al.*[11] examined whether the negative birth size–EPO association was stronger in infants delivered by elective Caesarean section than by other delivery modes. They focussed on the within-pair association measured by β_W since this is free of any confounding due to shared maternal, genetic, or environmental factors. These analyses can be performed by including interaction terms involving other covariates in model (2) or in the paired-difference approach of model (3), but the more general model (2) approach allows examination of between-pair effects and related interactions as well.

Models for dichotomous outcomes

Many epidemiological studies use dichotomous outcomes rather than continuous measures. We have concentrated most of this paper on continuous outcomes because this has been the focus of most of the recent interest in this type of twin study regression modelling. One reason for this may be that there is more scope to distinguish between different sources of variation when examining a continuously varying measure than a simple dichotomy. In addition, however, although the same general considerations carry over from the previous discussion to fitting logistic regression models to binary outcomes, there is an important additional complication. We will outline this briefly in the context of a further illustrative analysis involving the association of EPO level with the risk of the baby being born at low birth weight. For the purpose of this illustration we define low birth weight as being >1 SD below expected weight, based on gender and gestational age at birth; 47/220 (21%) of twins met this definition.

Parallel to model (1), for the binary outcome it is natural to start with the logistic regression specification:

$$\operatorname{logit}(\pi_{ij}) = \log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \beta_0 + \beta_C X_{ij}, \qquad (1D)$$

where π_{ij} denotes the probability that Y_{ij} is positive on the dichotomous outcome (i.e. the baby has low birth weight in this example). Note that π_{ij} is also the expected value of Y_{ij}, if Y_{ij} is coded 1 for 'yes' and 0 for 'no', but importantly the model no longer relates this mean value itself in a linear fashion to the regression parameters. Along with the fact that the errors around the expected value for the dichotomous Y_{ij} cannot be assumed to be normal (a binomial model applies), this means that constructing a full model that allows for correlation between responses within twin pairs, along the lines of (1a) above, is not straightforward. Essentially, the problem is that if one adds a random effect to the linear predictor in (1D) then one obtains a conditionally specified log-odds for the event of interest:

$$\operatorname{logit}(\pi_{ij}) = \log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \beta_B^* + \beta_C^* X_{ij} + \alpha_i, \qquad (1Da)$$

which involves parameters β_0^* and β_C^* that are different from the β_0 and β_C defined by the marginal specification (1D). This is the distinction between what have been called 'population-averaged' and 'subject-specific' (in this case, twin-pair-specific) parameters.[19,20] Space does not permit a full discussion of this issue here but the essential point is that the conditional log-odds represented by the pair-specific (starred) parameters in (1Da) represents association within a twin-pair, and when one averages out over twin pairs the relationship in the log-odds scale is attenuated. The degree of attenuation increases with the strength of intra-pair correlation.[21,22]

When we now consider the central concern of this paper, that the 'twins-as-individuals' model of type (1) should be considered only as a special case of the more general model (2), in which within-pair and between-pair effects may be clearly separated, the extension to binary outcomes is complicated by the distinction between marginal and conditional models. It is natural to consider the analogue of (2):

$$\text{logit}(\pi_{ij}) = \beta_0 + \beta_W(X_{ij} - \bar{X}_i) + \beta_B\bar{X}_i, \qquad (2D)$$

but there are now two versions of this model, one using the marginal (population-averaged) specification, and the other the conditional (pair-specific) specification. Each of these models may be fitted readily with modern software, using the GEE approach for the marginal specification, and a ML approach (requiring numerical integration) for the pair-specific model; see Appendix for details in Stata. However, the two different approaches estimate different quantities and require different interpretations, although assessments of statistical significance are generally consistent between the two approaches.[21]

Results for the EPO example are displayed in Table 2. The standard logistic regression and GEE approaches produce only slightly different estimates of the common odds ratio parameter [$\exp(\beta_C)$], while the ML estimate of $\exp(\beta_C^*)$ is somewhat larger. These differences are consistent with a modest intra-pair correlation in the dichotomous outcome (estimated at 0.2 by the GEE approach, 0.37 by the random effects/MLE approach). Comparing with the results for model (2), we see (as for the continuous outcome) that the model (1) estimates are intermediate between the estimates for the between-pair and within-pair coefficients, whether we consider the marginal or conditional versions. There is moderate evidence for differences between β_B^* and β_W^* (random-effects model; likelihood ratio test $P = 0.022$, Wald test $P = 0.037$) and likewise between β_B and β_W (marginal model; Wald $P = 0.016$). Either version of model (2) could be used to assess evidence for differential between-pair and within-pair effects, but the parameter of primary interest is likely to be the conditional within-pair effect, β_W^*, since this measures the change in the log-odds of the outcome per unit change in $X_{ij} - \bar{X}_i$, the deviation of the twin's X value from the pair mean (or equivalently, from their co-twin's X value). The population-averaged version of this parameter, β_W, is of less interest since it measures the average change in the log-odds across all twin pairs. Conversely, it is the population-averaged version of the between-pair effect that appears to be more readily interpretable.

Table 2 Estimates of odds ratios from logistic regression models for the association between risk of low birth weight (defined as Z-score < -1) and $\log_2$ EPO level for 220 DZ twin infants, under alternative regression approaches as described in the text

Parameter	Estimation method	Estimate	*P*-value	95% CI
Model (1)				
β_C	Standard GLM	1.41	0.005	(1.11–1.79)
β_C	GLM + 'robust' SE	1.41	0.008	(1.09–1.82)
β_C^*	MLE (pair-specific)	1.66	0.010	(1.13–2.44)
β_C	GEE	1.48	0.002	(1.15–1.91)
Model (2)				
β_W^*	MLE (pair-specific)	3.00	0.002	(1.49–6.05)
β_B^*		1.27	0.255	(0.84–1.92)
β_W	GEE	2.29	<0.001	(1.52–3.44)
β_B		1.21	0.249	(0.88–1.66)
Model (3)				
β_W^*	Conditional ML	3.20	0.020	(1.21–8.50)

β_C, β_C^* = common ('twins-as-individuals') logistic regression coefficients; β_W, β_W^* = within-pair coefficients; and β_B, β_B^* = between-pair coefficients; each marginal and pair-specific, respectively.

Finally, it is worth noting that the conditional likelihood method may be used to fit the analogue of model (3) for a dichotomous outcome: see last line of Table 2. Unlike the continuous paired-difference method, the analogy is not exact, and this method is not as efficient as likelihood estimation of model (2D).[23]

Discussion

How should regression analysis be applied in twin studies? Although some authors have recently recommended the use of the within-pairs and between-pairs model (2),[9,16] practice in the twin literature has generally followed the paired-difference approach (3), which we have shown can be subsumed under model (2).

A prominent recent strand of the literature has fitted the 'twins-as-individuals' model (1), and suggested that substantive conclusions should be based on comparing the results with those from within-pair analysis, i.e. comparing estimates of β_C and β_W.[5,7,8,24] However, if the more general model (2) holds, i.e. if $\beta_W \neq \beta_B$, then it will be difficult to interpret the estimated value of β_C. In this regard, it can be shown that the GLS estimate of β_C is a weighted average of the estimates of β_W and β_B where the weights depend on the within-pair correlation in both the outcome and the covariate.[17,25] Model (1) may be seen as addressing the question: how does Y vary with X (in the twin population) when one treats each twin separately? If we ignore the co-twin and ask how predictable is the outcome simply on the basis of the current twin's covariate value, then fitting model (1) may be appropriate. It may be, however, that the resulting association is wholly or partially confounded by factors that we know about, at least indirectly, because of the data we have on the co-twin.

To investigate an association after removing all confounding by factors shared between twins, one should examine the within-pair effect β_W. We have shown how this can be done either by performing the traditional paired-differences analysis [approach (3)], or by fitting the more general model (2). Advantages of fitting model (2) are that it allows (i) direct examination of whether a common-effect model ($\beta_W = \beta_B$) fits the data and (ii) simultaneous examination of other confounding effects due to factors that may be constant or variable within twin pairs. (Constant confounders will not affect β_W but may affect β_B while variable confounders may affect both.)

How then do we interpret the results of fitting model (2)? The within-pair effect β_W is quite readily interpretable: the equivalence with the paired-difference analysis reflects the fact that this parameter represents an association that is free of confounding due to factors that are common to the two twins. As pointed out elsewhere,[24,26] this supports interpretation in terms of possible causal mechanisms related to individual-specific factors such as different fetal nutrient supply lines: large values indicating for instance that whatever fetal factors underlie within-pair differences in X may also be driving the associated differences in Y.

The between-pair effect β_B reflects further variation in Y that can be explained by variation in the twin-pair mean of X. It may be unwise, however, to invest too much meaning in this latter parameter. Although the parameterization used in expression (2) has attractive features, there are many other possible re-expressions of the model. For example a recent paper has suggested that analysts might prefer to use the alternative version of the model expressed as

$$E(Y_{ij}) = \beta_0 + \beta_W X_{ij} + \beta_{B^*}\bar{X}_i,$$

where $\beta_{B^*} = \beta_B - \beta_W$.[27] Note that this model is simply a reparametrization of model 2 and, therefore, its fit to the data will be identical. The parameter β_{B^*} may be interpreted as the expected difference in outcome for a unit change in pair-mean X while holding the individual X value fixed. This parameter may be used to predict the difference in outcome between two individuals with the same X value but different pair-mean values (which obviously can only arise from the individuals having different co-twin values). Whether this is a more useful quantity than β_B, which predicts difference in outcome per unit of pair average between two individuals who have the same deviation from their pair average, is a matter of opinion. Clearly the two values will in general be quite different, and indeed β_{B^*} may be substantial in the presence of a negligible β_B [as in the EPO example, where $\hat{\beta}_{B^*} = 0.27$ (95% CI 0.11–0.43)].

This duality of possible interpretations between regression parameters occurs whenever covariates in a multiple regression specification are functionally related to each other. Another example is in the fitting of models for later-life outcomes such as blood pressure using standardized body size measures at birth and at current age, where the model can be interpreted either in terms of parameters for birth weight and current weight or in terms of parameters for birth weight and growth or change in weight between the two points.[28,29]

The primary attraction of the original parameterization used in (2) may be the technical feature that the covariates $X_{ij} - \bar{X}_i$ and $\bar{X}_i$ are 'orthogonal' to each other. This means that the coefficients and their estimates are well-defined independently of each other: separately regressing the individual twin values Y_{ij} on either the twin-pair mean or twin-pair difference of the covariate yields β_B or β_W, respectively. In fact the same point estimates are obtained using OLS as with the GLS methods, although as noted above correct standard errors require estimation based on a model incorporating correlation between twins.

So far we have considered only the regression of Y on X, that is, we have tried to explain variation in Y conditional on the observed pattern of X values, which was assumed to be fixed or 'given' in the usual sense of regression analysis. However, there may be little variation in X, especially within pairs, in which case there will be little scope for discerning related variation in Y and the corresponding estimated regression coefficient will have a large standard error. In order to explain the joint variation of two measures it is possible to develop a joint or bivariate model for X and Y, rather than relying on the regression approach of pretending that one variable is 'given' and that we are only interested in modelling the variation of the other one given the first. A natural approach is to build a variance components models for each of X and Y, where different components are included to represent sources of variation common to either or both the measurements on X and Y within a twin pair.[15,30] Such a model is capable of representing a range of underlying biological sources of variability, and in particular can be used in the manner of traditional twin study analysis to address the question of whether the within-pair correlations of either X or Y, and the regression relationship between them, appear to be explicable by genetic factors.

Conclusion

If the investigator is only interested in within-pair effects, that is, in the association of Y with X after removing all shared factors (covariates that are constant between twins), then the simple paired-difference analysis is sufficient. If one wishes to estimate an average relationship across the twin population without specifically using information on twin's shared characteristics, then model (1) provides an appropriate approach, except that the model should be fitted using a GLS method that allows for the correlation between twins. However, the more general model (2) is recommended since it allows simultaneous examination of both within-pair and between-pair effects, and if these effects are different then the results for model (1) are of less interest. We, therefore, recommend fitting model (2) using GLS ('mixed model' estimation or GEE), with the following general interpretations of the possible results:

- If $\beta_B \approx 0$ then we conclude that the only association is a dependence of the outcome on within-pair difference in covariate, not on the absolute level of the covariate. This would be consistent with causal mechanisms that relate to individual-specific factors that vary between twins in a pair, reflected in differences in the covariate and related differences in the outcome.
- If $\beta_W \approx 0$ we conclude that any association of Y with X, which would then only appear in the form of a non-zero β_B, can be explained by shared twin-pair factors. Once these are adjusted for, the association disappears, so an association found in a crude (unpaired) association between Y and X [an analysis assuming model (1)] could be explained as being due to confounding by these shared factors.
- If $\beta_B \approx \beta_W$ we may conclude that expected change in the outcome for a given change in the covariate is the same irrespective of whether the comparison is made between two twins or between two unrelated individuals in the twin population. It seems reasonable to infer from this that if there is a causal mechanism underlying the association

then it cannot rely entirely on shared, twin-pair level (e.g. maternal), factors. For one thing, such factors are by definition similar if not constant within twin pairs and so should not lead to differences in X, let alone to related differences in Y. On the other hand, given a non-zero β_B, the effect cannot be purely due to within-pair mechanisms.

- If both coefficients are non-zero but not equal, the interpretation becomes more complex—a compromise between the previously described three extremes.

Most of this paper has concentrated on continuous outcome measures, for which direct linear modelling of the mean, with normal error terms, is a natural approach (possibly after transformation of the outcome variable). Much epidemiological analysis is concerned with dichotomous outcomes, where the standard approach of logistic regression is to model the log-odds of the parameter of interest (the risk or probability of the outcome), with a binomial error distribution. As we have outlined, many of the concepts discussed in this paper carry over to the logistic regression framework, but there are some important further complications, in particular the fact that the ML and GEE approaches to model estimation diverge, as commonly applied—the GEE method is used to estimate marginally specified models while ML is used for a conditionally specified random-effects model. These two types of models involve different parameter definitions that require different interpretations.[19,20]

A further issue that requires comment is handling dichotomous covariates. These provide a special case, since the deviation from the pair mean, or half the difference between twins, can be only −½ or ½, when the twins are discordant, or 0 when the twins are concordant (assuming the dichotomous X is coded 0/1). The pair mean can only take the values 0, ½, or 1. With dichotomous covariates the concept of between-pair regression effects may not be useful and it appears best to analyse such covariates by examining within-pair differences amongst discordant pairs.

Returning to the application of the continuous normal linear model, many studies seek to draw conclusions about possible causal pathways by comparing regression parameters between various strata in the twin population, with the classic example being the comparison between MZ and DZ twins. These inferences need to be approached with caution, for reasons discussed in the previous section, but certain cases may be reasonably compelling. For instance, in the birth weight–EPO example, the comparison of the estimated value of β_W between twin pairs born by elective Caesarean section and those born vaginally indicated a stronger within-pair association in the Caesarean-born twins. This suggests that the twin-specific factors underlying this association may be blunted by various extraneous changes in EPO levels related to exposure to labour, which one would not expect to be predictive of birth weight.

Similarly, one might argue that a smaller within-pair regression coefficient for MZ twins than for DZ twins should be taken as evidence that the association is due to genetic factors, which are ‘matched out’ for the MZ pairs. It is possible, however, for the within-pair regression coefficients for MZ and DZ pairs to be similar even if the observed within-pair correlation structure for the outcome and exposure is consistent with a genetic influence.[15] More work is needed to clarify these issues but we believe it is not feasible to use regression methods to derive conclusions about the role of genetic factors in explaining associations.

As always, epidemiologists need to keep in mind the limitations of regression methodology,[31] remembering that it is fundamentally only a tool for assessing patterns of variation in an outcome conditional on one or more covariates. With twin data, the difficulties are increased, since one needs to consider whether the modelling should be conditional just on the individual’s covariate(s) or on their correlated twin’s as well. Explaining the causes of systematic variation identified by the regression model will in general require information from beyond the data, especially in the absence of randomization to control unknown confounders.

References

[1] Greenland S, Brumback B. An overview of relations among causal modelling methods. *Int J Epidemiol* 2002;**31:**1030–37.

[2] Leon DA, Lithell HO, Vagero D *et al*. Reduced fetal growth rate and increased risk of death from ischaemic heart disease: cohort study of 15 000 Swedish men and women born 1915–29. *BMJ* 1998;**317:** 241–45.

[3] Rich-Edwards JW, Stampfer MJ, Manson JE *et al*. Birth weight and risk of cardiovascular disease in a cohort of women followed up since 1976. *BMJ* 1997;**315:**396–400.

[4] McNeill G, Tuya C, Smith W. The role of genetic and environmental factors in the association between birthweight and blood pressure: evidence from meta-analysis of twin studies. *Int J Epidemiol* 2004;**33:**995–1001.

[5] Dwyer T, Blizzard L, Morley R, Ponsonby A. Within pair association between birth weight and blood pressure at age 8 in twins from a cohort study. *BMJ* 1999;**319:**1325–29.

[6] Zhang J, Brenner RA, Klebanoff MA. Differences in birth weight and blood pressure at age 7 years among twins. *Am J Epidemiol* 2001;**153:**779–82.

[7] Nowson CA, MacInnis RJ, Hopper JL *et al*. Association of birth weight and current body size to blood pressure in female twins. *Twin Res* 2001;**4:**378–84.

[8] Christensen K, Stovring H, McGue M. Do genetic factors contribute to the association between birth weight and blood pressure? *J Epidemiol Community Health* 2001;**55:**583–87.

[9] Johansson-Kark M, Rasmussen F, De Stavola B, Leon DA. Fetal growth and systolic blood pressure in young adulthood: the Swedish Young Male Twins Study. *Paediatr Perinat Epidemiol* 2002;**16**:200–9.

[10] McNeill G, Tuya C, Campbell DM *et al*. Blood pressure in relation to birth weight in twins and singleton controls matched for gestational age. *Am J Epidemiol* 2003;**158:**150–55.

[11] Morley R, Moore VM, Dwyer T, Owens JA, Umstad MP, Carlin JB. Association between erythropoietin in cord blood of twins and size at birth: does it relate to gestational factors or to factors during labor or delivery? *Pediatr Res* 2005;**57:**680–84.

[12] Carlin JB, Wolfe R, Coffey C, Patton GC. Analysis of binary outcomes in longitudinal studies using weighted estimating equations and discrete-time survival methods: prevalence and incidence of smoking in an adolescent cohort. *Stat Med* 1999;**18:**2655–79.

[13] Hanley JA, Negassa A, Edwardes MD, Forrester JE. Statistical analysis of correlated data using generalized estimating equations: an orientation. *Am J Epidemiol* 2003;**157:**364–75.

[14] Scott AJ, Holt D. The effect of two-stage sampling on ordinary least squares methods. *J Am Stat Assoc* 1982;**77:**848–54.

[15] Gurrin LC, Carlin JB, Sterne JAC, Dite GS, Hopper JL. Using bivariate models to understand between- and within-cluster regression coefficients, with application to twin data. *Biometrics* (in press).

[16] Iliadou A, Cnattingius S, Lichtenstein P. Low birthweight and type 2 diabetes: a study on 11 162 Swedish twins. *Int J Epidemiol* 2004;**33:** 948–53.

[17] Mann V, De Stavola BL, Leon DA. Separating within and between effects in family studies: an application to the study of blood pressure in children. *Stat Med* 2004;**23:**2745–56.

[18] Hopper JL, Seeman E. The bone density of female twins discordant for tobacco use. *N Engl J Med* 1994;**330:**387–92.

[19] Hu FB, Goldberg J, Hedeker D, Flay BR, Pentz MA. Comparison of population-averaged and subject-specific approaches for analyzing repeated binary outcomes. *Am J Epidemiol* 1998;**147:**694–703.

[20] Carlin JB, Wolfe R, Brown CH, Gelman A. A case study on the choice, interpretation and checking of multilevel models for longitudinal binary outcomes. *Biostatistics* 2001;**2:**397–416.

[21] Zeger SL, Liang K-Y, Albert PS. Models for longitudinal data: a generalized estimating equation approach. *Biometrics* 1988;**44:** 1049–60.

[22] Neuhaus JM. Statistical methods for longitudinal and clustered designs with binary responses. *Stat Methods Med Res* 1992;**1:**249–73.

[23] Neuhaus JM, Lesperance ML. Estimation efficiency in a binary mixed-effects model setting. *Biometrika* 1996;**83:**441–46.

[24] Dwyer T, Morley R, Blizzard L. Twins and fetal origins hypothesis: within-pair analyses. *Lancet* 2002;**359:**2205–06.

[25] Neuhaus JM, Kalbfleisch JD. Between- and within-cluster covariate effects in the analysis of clustered data. *Biometrics* 1998;**54:**638–45.

[26] Leon DA. The foetal origins of adult disease: interpreting the evidence from twin studies. *Twin Res* 2001;**4:**321–26.

[27] Begg MD, Parides MK. Separation of individual-level and cluster-level covariate effects in regression analysis of correlated data. *Stat Med* 2003;**22:**2591–602.

[28] Lucas A, Fewtrell MS, Cole TJ. Fetal origins of adult disease—the hypothesis revisited. *BMJ* 1999;**319:**245–49.

[29] Cole TJ. Modeling postnatal exposures and their interactions with birth size. *J Nutr* 2004;**134:**201–04.

[30] Neale MC, Cardon LR. *Methodology for Genetic Studies of Twins and Families*. Dordrecht: Kluwer, 1992.

[31] Berk RA. *Regression Analysis: A Constructive Critique*. Thousand Oaks: Sage Publications, 2004.

Appendix

We display commands used in the Stata package (Release 9, Stata Corporation, College Station, TX, 2005) to obtain the results shown in Tables 1 and 2.

```
/* Variables used are
   bwtsds    birth weight SD-score or z-score
   lowbwt    low birth weight indicator (=1 if SD-score < −1, 0 otherwise)
   epolog2   log(base 2)of EPO concentration in cord blood
   pair_no   unique identifier for each twin pair
   twin_no   birth order of twins(1,2)                                  */

* first create mean and difference variables:
by pair_no: egen epolmn=sum(epolog2)
replace epolmn=epolmn/2
generate epold=epolog2-epolmn
sort pair_no twin_no
by pair_no:  generate epoldif=epolog2[_n]-epolog2[_n-1]
by pair_no:  generate bwtdif=bwtsds[_n]-bwtsds[_n-1]

* TABLE 1
* model(1)
* standard OLS:
regress bwtsds epolog2
* OLS with ‘‘robust’’ SEs:
regress bwtsds epolog2, cl(pair_no)
* GLS estimation (REML method for variance parameters):
xtmixed bwtsds epolog2 ‖ pair_no:, reml
* GLS estimation (ML method for variance parameters):
xtmixed bwtsds epolog2 ‖ pair_no:, mle
est store A
* GEE estimation
xtgee bwtsds epolog2, i(pair_no) robust

* model(2)
* GLS estimation (REML method for variance parameters):
xtmixed bwtsds epold epolmn ‖ pair_no:, reml
* GLS estimation (ML method for variance parameters):
```

```
xtmixed bwtsds epold epolmn ‖ pair_no:, reml
est store B
* LR test for difference between beta_W and beta_B:
lrtest A B
* Wald test for difference between beta_W and beta_B:
lincom epolmn-epold
* GEE estimation
xtgee bwtsds epold epolmn, i(pair_no) rob
lincom epolmn-epold

* model (3)
reg bwtdif epoldif, nocons

* TABLE 2
* model (1D)
* standard GLM fit ignoring pairing:
logit lowbwt epolog2, or
* standard GLM fit with ‘‘robust’’ SEs
* (= GEE with independence working correlation):
logit lowbwt epolog2, cl(pair) or
* ML estimation of random effects model (pair-specific ORs)
xtlogit lowbwt epolog2, i(pair) or quad(24) nolog
est store A
* GEE estimation (i.e. with non-zero working correlation):
xtgee lowbwt epolog2, fam(binom) i(pair) eform robust nolog

* model (2)
* ML estimation of random effects model (pair-specific ORs)
xtlogit lowbwt epold epolmn, i(pair) or quad(24) nolog
est store B
quadchk
* LR test for difference between beta*_W and beta*_B:
lrtest A B
* Wald test for difference between beta*_W and beta*_B:
lincom epolmn-epold
* GEE estimation (i.e. with non-zero working correlation):
xtgee lowbwt epold epolmn, fam(binom) i(pair) eform robust nolog
* Wald test for difference between beta_W and beta_B:
lincom epolmn-epold

* model (3)
* conditional logistic regression estimation of beta*_W:
clogit lowbwt epold, group(pair) or
```