Properties of the Null Distribution of the GLS Test Statistic Applied to the Haseman-Elston Procedure

Richard M. Single, St. Olaf College, Stephen J. Finch, S.U.N.Y. at Stony Brook
Richard M. Single, St. Olaf College, Northfield, MN 55057 (single@stolaf.edu)

Key Words: sib-pair linkage analysis, correlated data.

Abstract:

A method for detecting linkage between a genetic marker and a quantitative trait in sibship data is presented that models the dependence structure within families. A computationally efficient algorithm is given for the application of generalized least squares (GLS) to the regression procedure developed by Haseman and Elston (1972).

The null distribution of the test statistic based on the GLS estimator of the Haseman-Elston regression coefficient is studied. The distribution is significantly skewed for studies with a small number of families when only a baseline correlation between sib pairs that share a sibling is incorporated. However, the observed significance levels for the test are not far from nominal levels when this correlation is estimated using an intraclass correlation coefficient.

1. Introduction

Nonparametric sib-pair methods of linkage analysis, such as the Haseman-Elston (H-E) sib-pair procedure (Haseman and Elston, 1972) and its extensions (Amos and Elston, 1989; Olson and Wijsman, 1993; Olson, 1995), continue to play an important role in the study of complex diseases. Several methods of accounting for the non-independence of sib pairs from sibships with three or more siblings have been proposed (Hodge, 1984; Wilson and Elston, 1993; Collins and Morton, 1995). These methods consider all possible sib pairs and estimate the equivalent number of independent sib pairs by either identifying an effective sample size or by using weights.

Current implementations of the H-E regression procedure use either ordinary least squares (OLS) or weighted least squares (WLS) in which weights are reciprocals of conditional variances of the squared sib-pair differences. WLS methods are not considered in this paper since they do not model the correlation between sib pairs and since the conditional variance is constant under the null hypothesis of no linkage.

Using all possible sib pairs from sibships of size $s > 2$ in the H-E test leads to a more powerful test for linkage than one based upon the same total number of sibs in independent pairs (Blackwelder, 1977). Blackwelder and Elston (1982) investigated different methods of allowing for the dependence between sib pairs in the H-E test for sibships of size three. They found that using all possible sib pairs and assuming independence led to a test with roughly the same power as tests which allowed for the dependence among sib pairs from the same family.

Amos and Elston (1989) and Olson and Wijsman (1993) suggested that modeling the dependence among sib pairs would lead to a gain in efficiency. Single and Finch (1995) numerically calculated the expected gain in efficiency from using GLS for studies with a small number of families and found that it asymptotically approaches 11%, 25%, and 36% for studies with 3, 4, and 5 siblings per family, respectively.

In this paper, we discuss the computational aspects of implementing GLS in the H-E procedure for studies with a large number of families. In addition, we describe the null distribution of the resulting test statistic.

2. The Genetic Model

The genetic model for a quantitative trait, x_{ik}, is

$$x_{ik} = \mu + g_{ik} + e^*_{ik}, \qquad (1)$$

where μ is an overall mean, g_{ik} is a major gene effect, and e^*_{ik} is a normally distributed random environmental effect. If the i^{th} family has s_i siblings with observed trait values x_{ik}, $k = 1, \ldots, s_i$, there will be $J_i = s_i(s_i - 1)/2$ possible different sib pairs. Letting j index the J_i pairs and $e_{ij} = (e^*_{ik} - e^*_{ik'})$, $k \neq k', j = 1, ..., J_i$, we have $e_{ij} \sim N(0, {\sigma_e}^2)$.

The squared sib pair differences are denoted $y_{ij} = (x_{ik} - x_{ik'})^2$, $k \neq k', j = 1, ..., J_i$, and the proportion of alleles shared identically by descent (IBD) at the

The authors thank Nancy R. Mendell and Robert C. Elston for their helpful comments and suggestions.

marker locus for the j^{th} sib pair in the i^{th} family is denoted π_{ij}. Haseman and Elston (1972) showed that for a codominant marker locus, the expected value of y_{ij} is a linear function of π_{ij}

$$E(y_{ij}) = \alpha + \beta\pi_{ij} \tag{2}$$

Under the null hypothesis of no linkage, the correlation between two squared sib pair differences is zero when there is no sibling common to the two pairs. However, when the squared differences involve a shared sibling, $corr(y_{ik}, y_{il}) = \rho \geq 0.25$ for most combinations of the genetic parameters (Blackwelder, 1977). We will refer to $\rho = .25$ as the baseline value.

3. The GLS Estimator

For the i^{th} family in an n family study, let $Y_i = (y_{i1}, ..., y_{iJ_i})^T$ be the vector of squared sib-pair differences, $\Pi_i = (\pi_{i1}, ..., \pi_{iJ_i})^T$ be the vector containing the proportion of alleles IBD, and 1_i represent a vector of ones. Let V_i denote the variance-covariance matrix of the squared differences. For a single family with $s = 4$ siblings, the corresponding correlation matrix is as given below (for $s = 3$ the upper left 3 $\times$ 3 portion is used):

$$\begin{pmatrix} 1 & \rho & \rho & \rho & \rho & 0 \\ \rho & 1 & \rho & \rho & 0 & \rho \\ \rho & \rho & 1 & 0 & \rho & \rho \\ \rho & \rho & 0 & 1 & \rho & \rho \\ \rho & 0 & \rho & \rho & 1 & \rho \\ 0 & \rho & \rho & \rho & \rho & 1 \end{pmatrix}$$

The variance-covariance matrix V for the n family study is a block diagonal matrix with blocks V_i. The entries in the second column of the design matrix Π will be the Π_i, and the corresponding 1_i will be the first column entries. Y is a column vector with entries Y_i.

$$V = \begin{pmatrix} V_1 & 0 & \dots & 0 \\ 0 & V_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & V_n \end{pmatrix}, \Pi = \begin{pmatrix} 1_1 & \Pi_1 \\ \vdots & \vdots \\ 1_n & \Pi_n \end{pmatrix},$$

$$Y = (Y_1, ..., Y_n)^T$$

Under the null hypothesis, the GLS estimator of the regression slope is:

$$\hat{\beta}_{GLS} = \left[(\Pi^T V^{-1} \Pi)^{-1} \Pi^T V^{-1} Y\right]_{2,1} . \tag{3}$$

where the subscript indicates the (2,1) entry of the matrix product.

3.1 Computational Aspects

The structure of the variance-covariance matrix for a multifamily study facilitates the computation of the matrix product involved in calculating the GLS estimate. Due to the block diagonal form of the matrix V, the GLS estimate of the regression coefficient given in Eq. (3) can be written as follows:

$$\hat{\beta}_{GLS} = \frac{\sum 1_i{}^T V_i{}^{-1} 1_i \sum \Pi_i{}^T V_i{}^{-1} Y_i - \sum \Pi_i{}^T V_i{}^{-1} 1_i \sum 1_i{}^T V_i{}^{-1} Y_i}{\sum 1_i{}^T V_i{}^{-1} 1_i \sum \Pi_i{}^T V_i{}^{-1} \Pi_i - (\sum \Pi_i{}^T V_i{}^{-1} 1_i)^2} \tag{4}$$

where i indexes the n families in the study.

The null variance of the GLS estimate can be written as follows:

$$\widehat{\text{var}}(\hat{\beta}_{GLS}) = \left[(\Pi^T V^{-1} \Pi)^{-1}\right]_{2,2} \hat{\sigma}^2_{GLS} = \frac{\sum 1_i{}^T V_i{}^{-1} 1_i \hat{\sigma}^2_{GLS}}{\sum 1_i{}^T V_i{}^{-1} 1_i \sum \Pi_i{}^T V_i{}^{-1} \Pi_i - (\sum \Pi_i{}^T V_i{}^{-1} 1_i)^2} . \tag{5}$$

where $\hat{\sigma}^2_{GLS}$ is defined in Eq. (8)

Each of the five different summations in Eq. (4) and Eq. (5) is a sum of quadratic polynomials in either π_{ij}, y_{ij}, or both. Letting m index the possible family sizes, the polynomials ${1_m}^T {V_m}^{-1} 1_m$, ${\Pi_m}^T {V_m}^{-1} 1_m$, ${\Pi_m}^T {V_m}^{-1} \Pi_m$, ${1_m}^T {V_m}^{-1} Y_m$, and ${\Pi_m}^T {V_m}^{-1} Y_m$ can be calculated symbolically, for example using Maple V (Char et. al., 1991), and stored for each of the possible family sizes.

The method of estimating σ^2 depends on the function of the residuals minimized in determining the parameter estimates. Using generalized least squares, $\hat{\beta}$ is obtained by minimizing the following polynomial in the residuals (SSE_{GLS}):

$$\begin{aligned} SSE_{GLS} &= (Y - \hat{Y})^T V^{-1} (Y - \hat{Y}) \\ &= [Y^T V^{-1} Y] - [Y^T V^{-1} \Pi (\Pi^T V^{-1} \Pi)^{-1} \Pi^T V^{-1} Y]. \end{aligned} \tag{6}$$

Taking advantage of the structure of the variance-covariance matrix V, this can be rewritten in terms of summations over the n families as follows:

$$\sum Y_i{}^T V_i{}^{-1} Y_i - \left[\frac{1}{\sum 1_i{}^T V_i{}^{-1} 1_i \sum \Pi_i{}^T V_i{}^{-1} \Pi_i - (\sum \Pi_i{}^T V_i{}^{-1} 1_i)^2}\right] \times \left[(\sum 1_i{}^T V_i{}^{-1} Y_i)^2 \sum \Pi_i{}^T V_i{}^{-1} \Pi_i\right.$$

$$-2\sum \Pi_i{}^T V_i{}^{-1} 1_i \sum 1_i{}^T V_i{}^{-1} Y_i \sum \Pi_i{}^T V_i{}^{-1} Y_i$$

$$+\sum 1_i{}^T V_i{}^{-1} 1_i (\sum \Pi_i{}^T V_i{}^{-1} Y_i)^2 \Bigg] . \quad (7)$$

The GLS estimate of σ^2 is then given by,

$$\hat{\sigma}^2_{GLS} = \frac{SSE_{GLS}}{N-2}, \quad (8)$$

where

$$N = \sum_{i=1}^{n} s_i(s_i - 1)/2 \quad (9)$$

and s_i is the number of siblings in the i^{th} family.

The OLS estimate of σ^2 is obtained by dividing the sum of squared residuals by its degrees of freedom:

$$\hat{\sigma}^2_{OLS} = \frac{SSE_{OLS}}{N-2} = \frac{(Y-\hat{Y})^T(Y-\hat{Y})}{N-2}. \quad (10)$$

The estimated variance of the OLS regression coefficient is then given by the following:

$$\widehat{\mathrm{var}}(\hat{\beta}_{OLS}) = [(\Pi^T\Pi)^{-1}\Pi^T V \Pi (\Pi^T\Pi)^{-1}]_{2,2}\hat{\sigma}^2_{OLS}. \quad (11)$$

Assuming independence among sib pairs within families leads to the following estimate of the variance of the regression coefficient:

$$\widehat{\mathrm{var}_{\mathrm{I}}}(\hat{\beta}_{OLS}) = [(\Pi^T\Pi)^{-1}]_{2,2}\hat{\sigma}^2_{OLS}. \quad (12)$$

When there are only two siblings per family, the variance-covariance matrix of the squared sib-pair differences is diagonal and the three estimates of the variance of the regression coefficient, Eq.(5), Eq. (11), and Eq. (12), are identical.

The calculation of the correct null variance of the OLS estimator is computationally intensive. Unlike the variance of the GLS estimator given by Eq. (5), the overall matrix product given in Eq. (11) does not decompose into sums of within family matrix products.

4. The Null Distribution of the GLS Test Statistic

Earlier results on the estimated variance of the GLS regression coefficient indicated some dependence on the sibship size s and on the genetic parameters (that is, the gene frequency p and the heritability h^2 of the trait). Therefore, we expected this parameter dependence to manifest itself in the null distribution of the GLS test statistic,

$$t_{GLS} = \frac{\hat{\beta}_{GLS}}{\sqrt{\widehat{\mathrm{var}}(\hat{\beta}_{GLS})}}. \quad (13)$$

Thus, separate simulations were done for each combination of the following genetic parameters and no dominance at the trait locus: p = .3 and .5; h^2 = .1, .5, and .9. For each combination of the parameters, 5,000 replications were used.

Table 1: Empirically Determined Number of Sib-Pairs Needed in order that | Skewness |< .1 when ρ is Fixed at Baseline.

p	h^2	$s=3$	$s=4$	$s=5$
.3	.1	30	52	63
.3	.5	32	55	62
.3	.9	40	73	80
.5	.1	27	55	60
.5	.5	33	56	67
.5	.9	35	63	74

4.1 Mean, Variance, and Skewness of the Distribution with ρ Fixed at Baseline

The mean of the t_{GLS} statistic was not significantly different from zero (range = -0.045 to 0.051). Its variance was approximately equal to one (range = 0.95 to 1.10). There appeared to be no dependence on the sibship size or the number of families with regard to the mean or variance of t_{GLS}.

The null distribution of the GLS test statistic was found to be significantly negatively skewed for studies with a small number of families. As the number of families is increased, this skewness becomes negligible. The rate at which these asymptotic results are reached depends upon p, h^2, and the sibship size. Higher heritability, more extreme values of p, and larger sibship sizes lead to larger negative skewness values.

The convergence to an asymptotic distribution that is not skewed was fastest for p = 0.5 and h^2 = 0.1. In this case between 30 and 60 families were needed before the skewness was not significantly negative. For the various simulation parameter settings, Table 1 reports an estimate of the number of families needed in order to have a skewness value that is greater that -0.1.

4.2 Selected Percentiles of t_{GLS} with ρ Fixed at Baseline

The fact that the GLS test statistic should follow a Student's t distribution led to a specific strategy for modeling the percentiles of the null distribution of t_{GLS}. The simulated percentile points were regressed on different functions of the number of families in a study. Since the sibship size was found to affect the rate at which asymptotic results were reached, the percentiles of the null distribution of t_{GLS} were modeled separately for the different sibship sizes considered. Also, graphical inspection and analysis of variance techniques indicated that the genetic parameters under which the simulation was performed were significant predictors of the empirical percentiles of the null distribution for $s \geq 4$. However, this was not the case for h^2 with $s = 3$.

The model used was

$$E(Y_{qns}) = \alpha_{qsk} + \beta_{qsk}(1/n)^k, k > 0 \qquad (14)$$

where Y_{qns} denotes the q^{th} percentile of the distribution from a study with n families of sibship size s at a particular combination of the genetic parameters. Results were calculated for values of k in the range from $\frac{1}{8}$ to 3.

Using this model, the intercept, $\hat{\alpha}_{qsk}$, is an estimate of the q^{th} percentile of the asymptotic distribution (Thode et. al., 1988). The following strategy was used to model the percentiles of the simulated distributions: First, the intercept was fixed at the value of the q^{th} percentile of the standard normal distribution. Then, k was chosen such that the regression on $(1/n)^k$ explained the highest proportion of the variability in the percentiles. This procedure suggested itself because R^2 was a convex function of the transformation parameter k, with a maximum in the selected range of k values. Three separate simulations, each with 5,000 replications, were conducted which enabled the calculation of a lack of fit F statistic. The associated test for lack of fit, at the .05 level of significance, was significant for less than 8% of the regressions. This occurred only when modeling percentiles from studies with $s = 3$ and $s = 4$ siblings per family. For studies with $s \geq 4$, the results suggested that convergence of the percentiles may be a power series in $(1/n)^{\frac{1}{2}}$. The convergence was faster for $s = 3$, with $k \geq .75$.

With $s \geq 4$, there was a trend in h^2 for a given percentile. Higher heritability corresponded to a larger fitted percentile in absolute value. For a fixed value of the gene frequency, p, this trend was strongest for the larger sibships.

The effect of the slower convergence for larger sibship sizes is apparent in Figure 1 which shows the observed significance levels for studies of $n = 100$ families at the nominal .01 significance level.

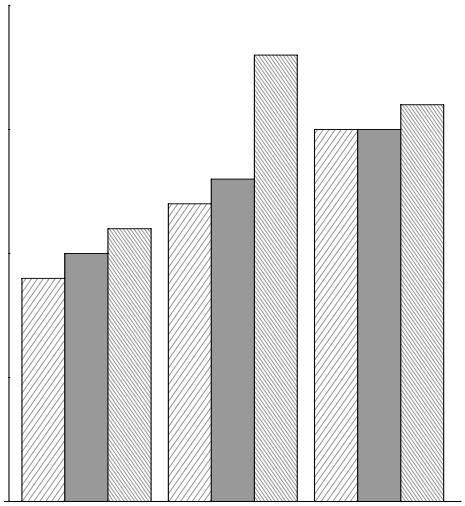

Figure 1: Observed Significance Levels for Studies of $n = 100$ families at the .01 level ($p = .5$ and ρ is fixed at baseline).

4.3 Estimating ρ

For studies with $s = 3$ siblings per family, each sib pair shares a sibling in common. The intraclass correlation coefficient $\hat{\rho}_I$ provides a measure of the correlation between different sib pairs that share a sibling. In this situation, the families serve as the groups in an analysis of variance.

Not all sib pairs share a sibling in common in sibships of size $s > 3$. If we let $\hat{\rho}_{Ij}$ denote the estimate computed using only sib pairs that share sibling j, then an overall intraclass correlation coefficient estimate can be computed as the average of these s estimates:

$$\hat{\rho}_I = \frac{\hat{\rho}_{I1} + \ldots + \hat{\rho}_{Is}}{s}. \qquad (15)$$

Table 2 lists the observed significance levels for the t_{GLS} test statistic with the correlation between sib pairs that share a sibling estimated by $\hat{\rho}_I$. Results

Table 2: Observed Significance Levels of t_{GLS} with ρ estimated by $\hat{\rho}_I$ for studies of $n = 100$ families

			Nominal Level		
p	h^2	s	1%	2.5%	5%
.3	.1	3	.010	.026	.048
.3	.5	3	.011	.026	.056
.3	.9	3	.009	.026	.055
.5	.1	3	.011	.029	.055
.5	.5	3	.009	.024	.046
.5	.9	3	.010	.030	.056
.3	.1	4	.013	.031	.058
.3	.5	4	.010	.028	.053
.3	.9	4	.020	.035	.060
.5	.1	4	.011	.026	.051
.5	.5	4	.015	.029	.056
.5	.9	4	.018	.037	.066
.3	.1	5	.016	.033	.057
.3	.5	5	.014	.030	.055
.3	.9	5	.022	.040	.071
.5	.1	5	.012	.029	.056
.5	.5	5	.015	.032	.053
.5	.9	5	.015	.033	.065

NOTE: Each entry is based on 5,000 replications for each parameter combination.

are given for studies with $n = 100$ families and are based on 5,000 replications for each combination of parameters.

5. Discussion

In this paper we have investigated the null distribution of the GLS test statistic applied to the H-E regression procedure. The GLS test statistic that incorporates the baseline correlation among sib pairs that share a sibling has a negatively skewed distribution for studies with a small number of families. The degree of the skewness depends on the underlying genetic model. This same dependence on the genetic parameters is demonstrated by the fitted percentiles of the null distribution. The convergence of the percentiles is fastest for smaller sibship sizes.

Presumably, restricting the correlation between sib pairs that share a sibling to the value of .25 is the source of much of the parameter dependence. Thus, simply modeling the baseline correlation is not adequate. The value of .25 is correct when there is no major gene causing the observed trait (i.e., $h^2 = 0$). The correlation between sib pairs sharing a sibling increases with increasing heritability.

Observed significance levels were calculated for the various parameter settings when ρ is estimated rather than fixed at the baseline value. Table 2 indicates that empirical type I error rates are close to nominal levels when ρ is estimated by $\hat{\rho}_I$.

6. References

Amos, C.I., Elston, R.C., Wilson, A.F., Bailey-Wilson, J.E. (1989). A More Powerful Robust Sib-Pair Test of Linkage for Quantitative Traits. *Genetic Epidemiology,* 6, 435-449.

Blackwelder, W.C. (1977). *Statistical Methods for Detecting Genetic Linkage from Sibship Data.* Dissertation from the University of North Carolina, Chapel Hill. Institute of Statistics Mimeo Series Number 1114.

Blackwelder, W.C., Elston, R.C. (1982). Power and robustness of sib-pair linkage tests and extension to larger sibships. *Communications in Statistics, Theory and Methods,* 11, 449-484.

Char, B.W., Keith, O.G., Gaston H.G., Benton, L.L., Monagan, M.B., Watt, S.M. (1991): *Maple V Library Reference Manual.* New York: Springer-Verlag.

Draper, N.R., Smith, H. (1981). *Applied Regression Analysis, Second Edition.* New York: John Wiley and Sons.

Haseman, J.K., Elston, R.C. (1972). The Investigation of Linkage Between a Quantitative Trait and a Marker Locus. *Behavior Genetics,* 2, 3-19.

Hodge, S.E. (1984). The Information Contained in Multiple Sibling Pairs. *Genetic Epidemiology,* 1, 109-122.

Nick, T.G., Varghese G., Elston, R.C., Wilson, A.F. (1995). Statistical Validity for Testing Associations Between Genetic Markers and Quantitative Traits in Family Data. *Genetic Epidemiology,* 12, 145-161.

Olson, J.M. (1995). Robust Multipoint Linkage Analysis: An Extension of the Haseman-Elston Method *Genetic Epidemiology,* 12, 177-193.

Olson, J.M., Wijsman, E. (1993). Linkage Between Quantitative Trait and Marker Loci: Methods Using All Relative Pairs. *Genetic Epidemiology,* 10, 87-102.

Single, R.M., Finch, S.J. (1995): Gain in Efficiency from Using Generalized Least Squares in the Haseman-Elston Test. *Genetic Epidemiology*, 12, 889-894.

Wilson, A.F., Elston, R.C. (1993). Statistical Validity of the Haseman-Elston Sib-Pair Test in Small Samples. *Genetic Epidemiology*, 10, 593-598.

Tran, L.D., Elston, R.C., Keats, B.J.B., Wilson, A.F., Sib-Pair Linkage Program Version 2.6, Part of S.A.G.E. Release 2.2 documentation (1994).