The International Journal of Biostatistics

Volume 8, *Issue* 1 2012 *Article* 20

Measures of Family Resemblance for Binary Traits: Likelihood Based Inference

Mohamed M. Shoukri, *Schulich School of Medicine and Dentistry*
Abdelmoneim ElDali, *King Faisal Specialist Hospital*
Allan Donner, *Schulich School of Medicine and Dentistry*

Recommended Citation:
Shoukri, Mohamed M.; ElDali, Abdelmoneim; and Donner, Allan (2012) "Measures of Family Resemblance for Binary Traits: Likelihood Based Inference," *The International Journal of Biostatistics*: Vol. 8: Iss. 1, Article 20.
DOI: 10.1515/1557-4679.1410

Measures of Family Resemblance for Binary Traits: Likelihood Based Inference

Mohamed M. Shoukri, Abdelmoneim ElDali, and Allan Donner

Abstract

Detection and estimation of measures of familial aggregation is considered the first step to establish whether a certain disease has genetic component. Such measures are usually estimated from observational studies on siblings, parent-offspring, extended pedigrees or twins. When the trait of interest is quantitative (e.g. Blood pressures, body mass index, blood glucose levels, etc.) efficient likelihood estimation of such measures is feasible under the assumption of multivariate normality of the distributions of the traits. In this case the intra-class and inter-class correlations are used to assess the similarities among family members. When the trail is measured on the binary scale, we establish a full likelihood inference on such measures among siblings, parents, and parent-offspring. We illustrate the methodology on nuclear family data where the trait is the presence or absence of hypertension.

KEYWORDS: family resemblance; bivariate exchangeable distributions; likelihood inference; clustered data; bootstrap technology

Author Notes: Partial support for this research by the Natural Sciences and Research Council (NSERC) of Canada to Allan Donner is greatly acknowledged. The comments made by two reviewers have substantially improved the paper.

1. Introduction

Classical epidemiology deals with disease patterns and factors associated with causation of disease with the ultimate aim of prevention and control. When the question is whether a disease has a genetic component, the detection and estimation of familial aggregation (e.g., higher occurrence rates in siblings or offspring) becomes very important. Fortunately, relevant information on familial aggregation may already be known from descriptive epidemiology studies. In particular, results from observational studies on sib-correlations, parent-offspring correlation, parent-parent correlation, and twin-concordance may suggest a genetic component in the etiology of a disease or trait. As was noted by Laird and Lange [1], "The general concepts used in aggregation and heritability analysis are widely accepted as useful measures of the degree to which traits are inherited; most researchers would not undertake genetic analysis without evidence of aggregation or heritability of the trait."

However familial aggregation of a trait is a necessary but not sufficient condition for inferring the importance of genetic susceptibility, since environmental and cultural influences can also play a role in familial clustering and excess familial risk. Note that in traditional societies the presence of consanguinity would increase the role of genetic factors in understanding variability of a trait. That is, when the parents are related, the measure of similarity would have a significant genetic component. On the other hand in the absence of consanguinity, as in developed countries, the genetic component of the measure of similarity will be negligible. For quantitative traits, the biometrical approach introduced by Morton [2], Rao et al [3], and Morton and McLean [4] to evaluate the degree of resemblance among family members has relied on the well developed multivariate normal theory. However in assessing the degree of family resemblance, clinical epidemiologists often prefer to report the disease status of individuals on a binary scale. Therefore analytic approaches established under the multivariate normal model are not useful. However several strategies have been developed to estimate the degree of familial resemblance in this case. The most widely used of these is the semi-parametric approach known as Generalized Estimating Equations (GEE) developed by of Liang and Zeger [5] and Zeger and Liang [6]. A limitation of this approach is that measures of family resemblance are treated as nuisance parameters while the modeling strategy focuses on the estimation of regression coefficients corresponding to the selected risk factors.

The chief objective of this paper is to derive fully efficient estimators of two sets of measures of family resemblance using maximum likelihood methods. The first is the set comprising sib-sib correlations, which requires the estimation of the intraclass correlation coefficient for binary traits. The second set comprises interclass correlations that provide a measure of resemblance among parents and

their siblings. We therefore introduce here a model that is suitable for the analysis of data with a specific structure that is characteristic of nuclear family data. Based on this model, likelihood inferences are developed that produce correlation estimates that are fully efficient.

The paper is structured as follows: In Section 2 we introduce the model. In Section 3 we present likelihood inferences for the model parameters followed by an illustration using arterial blood pressure data collected from nuclear families. For the purpose of comparison, we also discuss alternative estimators, and use the bootstrap method to derive their empirical standard errors.

2. **Models**

Let $Y_{ij} = 1(0)$ denote the presence (absence) of a trait in the j^{th} sibling from the i^{th} family $(j = 1,2,...n_i; i = 1,2,...k)$. Similarly, let $X_{if} = 1(0)$ and $X_{im} = 1(0)$ denote the presence (absence) of the condition, respectively in the mother and father. Let $\lambda_i = P(Y_{ij} = 1 \mid \lambda_i)$ denote the probability that a randomly selected sibling from the i^{th} family is classified as having the condition, and let $1-\lambda_i = P(Y_{ij} = 0 \mid \lambda_i)$. Moreover let $P(X_{if} = 1 \mid \pi_i) = P(X_{im} = 1 \mid \pi_i) = \pi_i$, and $P(X_{if} = 0 \mid \pi_i) = P(X_{im} = 0 \mid \pi_i) = 1-\pi_i$. For the time being we shall assume that the distribution of offspring scores is conditionally independent of the distribution of their parents score. To introduce the correlation between parents within the i^{th} family we shall assume that π_i is an element of a random sample from a beta distribution with parameters (α, β) so that $\mu_p = E(\pi_i) = \dfrac{\alpha}{\alpha+\beta}$, and $\text{var}(\pi_i) = \dfrac{\alpha\beta}{(\alpha+\beta)^2(1+\alpha+\beta)} = \rho_p \mu_p (1-\mu_p)$, where $\rho_p = (1+\alpha+\beta)^{-1}$ [7, 8, 9]. The choice of the beta distribution is also justified in the context of Bayesian inference as it is the conjugate prior for the event probability in Bernoulli trials. Therefore the unconditional distribution of the sum $X_i = X_{if} + X_{im}$ is that of a beta binomial distribution with:

$$E(X_i) = 2\mu_p, \text{ and } \text{var}(X_i) = 2\mu_p(1-\mu_p)[1+\frac{\vartheta_p}{1+\vartheta_p}], \text{ where } \vartheta_p = \rho_p/(1-\rho_p)$$

and where the parameter ρ_p is in fact the intraclass correlation between (X_{if}, X_{im}). It is therefore a population average measure of similarity between parents within the same family.

The probability distribution of $X_i = X_{if} + X_{im}$, known as the extended beta-binomial distribution is given by:

$$P(X_i = x_i) = \binom{2}{x_i} \frac{\prod_{j=0}^{x_i-1}(\mu_p + j\theta_p)\prod_{j=0}^{1-x_i}(1-\mu_p + j\theta_p)}{\prod_{j=0}^{1}(1+j\theta_p)} \quad x_i = 0,1,2, \tag{1}$$

with the convention $\prod_{j=0}^{-1}(\alpha_j) = 1$. In particular:

$P(X_i = 0) = (1-\mu_p)^2 + \mu_p(1-\mu_p)\rho_p$

$P(X_i = 1) = \mu_p(1-\mu_p)(1-\rho_p)$ and

$P(X_i = 2) = \mu_p^2 + \mu_p(1-\mu_p)\rho_p$

Similar to the above set-up, we assume that the offspring scores are conditionally independent. Then due to the exchangeability assumption, the conditional distribution of $Y_i = \sum_{j=1}^{n_i} Y_{ij}$ has a binomial distribution with parameters (n_i, p_i). Note that the above results have been obtained previously within the context of interrater agreement by Bloch and Kraemer [10]. To model the correlation among offspring within the same family we assume that λ_i has a beta distribution with parameters (δ, γ). The unconditional distribution of Y_i is beta binomial with mean and variance given respectively as:

$E(Y_i) = n_i\mu_s$ and $\text{var}(Y_i) = n_i\mu_s(1-\mu_s)[1 + (n_i - 1)\frac{\vartheta_s}{1+\vartheta_s}]$, where

$\mu_s = \frac{\delta}{\delta+\gamma}$, $\vartheta_s = \rho_s/(1-\rho_s)$, and $\rho_s = (1+\delta+\gamma)^{-1}$ is the intraclass correlation between pairs of siblings within the same family. The marginal probability distribution of Y_i is given by:

$$P(Y_i = y_i) = \binom{n_i}{y_i} \frac{\prod_{j=0}^{y_i-1}(\mu_s + j\theta_s)\prod_{j=0}^{n_i-y_i-1}(1-\mu_s + j\theta_s)}{\prod_{j=0}^{n_i-1}(1+j\theta_s)} \qquad y_i = 0,1,2,....n_i \tag{2}$$

Setting $n_i = 2$ in equation (2), we obtain equation (1). In order to obtain the familial correlations, we follow an approach known as Positive Dependence by Mixture (PDM) [11, 12]. That is, by mixing the Bernoulli distribution with a beta distribution we can obtain the intracluster correlation. A different approach however is needed to construct the bivariate distribution of the parent and offspring scores, characterized by the interclass correlation. This approach was developed by Sarmanov [11] and Lancaster [12] and is known as Positive Dependence by Expansion (PDE). Danaher [13] proposed a simplified and flexible form of this distribution given by:
$P(x_i, y_i) = P(X_i = x_i).P(Y_i = y_i)[1 + \rho_{12}u_{1i}u_{2i}]$, where

$$u_{1i} = \frac{x_i - 2\mu_p}{\sigma_p},\ u_{2i} = \frac{y_i - n_i\mu_s}{\sigma_{si}},\ \sigma_p = \sqrt{\text{var}(x_i)},\ \sigma_{si} = \sqrt{\text{var}(y_i)}, \tag{3}$$

and $\rho_{12} = Corr(x_i, y_i)$ is the interclass correlation.

It is clear that X_i and Y_i are statistically independent if and only if $\rho_{12} = Corr(x_i, y_i) = 0$.

In Table 1 we depict the data layout, which shows that the data structure is similar to that obtained under split-cluster sampling.

Table 1: Data layout depicting the hierarchical structure of the family data

		Family		
Score	1	2	*i*	*K*
Parents	x_{1f}	x_{2f}	x_{if}	x_{kf}
	x_{1m}	x_{2m}	x_{im}	x_{km}
Total score of both parents	x_1	x_2	x_i	x_k
Offspring	y_{11}	y_{21}	y_{i1}	y_{k1}
	y_{12}	y_{22}	y_{i2}	y_{k2}
	y_{1n_1}	y_{2n_2}	$y_{in_{1.}}$	y_{kn_k}
Total scores of offspring	y_1	y_2	y_i	y_k

3. Parameters Estimation

In this section we consider two methods of estimation. The first is the commonly used method of maximum likelihood. The second method is non-parametric in nature and was proposed by Karlin et al [14]

3.1. Likelihood Representation

Suppose that we have a sample of k randomly selected families, where the i^{th} family has the scores for both parents and n_i children. The likelihood of the sample may be written as:

$$\begin{aligned} L &= \prod_{i=1}^{k} P(x_i, y_i) \\ &= \prod_{i=1}^{k} \left[\frac{P(x_i, y_i)}{P(X_i = x_i)} \cdot P(X_i = x_i) \right] \\ &= \prod_{i=1}^{k} \left[P(y_i | x_i) \cdot P(x_i) \right] = \prod_{i=1}^{k} P(y_i \mid x_i) \prod_{i=1}^{k} P(x_i) = L_2 . L_1 \end{aligned} \tag{4}$$

Thus it is seen that the likelihood function can be expressed as the product of two functions, the first, L_1, depends on the parameters that characterize the parent parameters (μ_p, ρ_p), while the second, L_2 based on the conditional distribution of the offspring scores given the parents' scores. We write the log-likelihood function as the sum of two components: $\log L = l_1 + l_2$, where

$$l_1 = \sum_{i=1}^{k} \log P(x_i) \text{ and } l_2 = \sum_{i=1}^{k} \log P(y_i | x_i).$$

Note that l_1 can be maximized with respect to μ_p and ρ_p, with their maximum likelihood estimators then obtained in closed forms as :

$$\hat{\mu}_p = x_0 / 2k, \text{ and } \hat{\rho}_p = \frac{MSB - MSW}{MSB + MSW}, \text{ where } x_0 = \sum_{i=1}^{k} x_i,$$

$$MSW = \frac{1}{k}[x_0 - \frac{\sum_{i=1}^{k} x_i^2}{2}], \text{ and } MSB = \frac{1}{2(k-1)}[\sum_{i=1}^{k} x_i^2 - \frac{x_0^2}{k}]$$

Once the estimators of the parameters in l_1 are obtained, we substitute them in the function l_2 which is maximized using the Newton-Raphson at the solutions of the equations: $\frac{\partial l_2}{\partial \mu_s} = 0, \ \frac{\partial l_2}{\partial \rho_s} = 0, \ \frac{\partial l_2}{\partial \rho_{12}} = 0.$ This approach was proposed by Richards [15]. Briefly, let the parameter vector θ_{px1} be partitioned so that $\theta_{px1} = (\theta_1, \theta_2)^T$, where θ_1 and θ_2 have dimensions r and $(p-r)$ respectively. Let $\hat{\theta}_1(\theta_2)$ be the MLE of θ_1 for fixed values of θ_2, and $\hat{\theta}_2$ be the MLE of θ_2.

Solving $\frac{\partial l(\theta)}{\partial \theta_i} = 0, \ i = 1,2,\ldots r$ and substituting $\hat{\theta}_1(\theta_2)$ for θ_1 in $\frac{\partial l}{\partial \theta_2}$, we obtain a modified likelihood function, which is a function of θ_2 only. Using this modified likelihood function we find the MLE $\hat{\theta}_2 = (\tilde{\theta}_{r+1}, \ldots \hat{\theta}_p)^r$ by solving $\frac{\partial l(\theta)}{\partial \theta_i} = 0, \ i = r+1, \ldots p.$

The advantage of this approach is that instead of maximizing the full likelihood function for five parameters, the proposed partition enables us to perform the calculations with greater accuracy and less time.

The variance-covariance matrix Σ is obtained by inverting the negative of the matrix of the second partial derivatives, and then substituting the values of the estimated parameters.

$$\hat{\Sigma} = -[\frac{\partial^2 l}{\partial \theta_i \partial \theta_j} \Big|_{\hat{\theta}_i, \hat{\theta}_j}] \text{ where, } \theta_1 \equiv \mu_p, \ \theta_2 \equiv \rho_p, \ \theta_3 \equiv \mu_s, \ \vartheta_4 \equiv \rho_s, \ \theta_5 \equiv \rho_{12}.$$

3.2. Karlin's estimators:

Karlin et al. [14] used correlations to assess the similarity among family members. Although their development was aimed primarily at continuously distributed traits, the proposed estimators can also be used to assess similarities with respect to binary outcomes. In this section we introduce alternative moment estimators for the sib-sib correlation and the parent-offspring correlation.

a. Sib-sib correlation

Similar to the case of continuous phenotypes, Karlin et al. suggested on ANOVA type estimator:

$$\widetilde{\rho}_{ss} = \frac{B.S.V. - W.S.V}{B.S.V + (n_0 - 1)W.S.V}, \text{ where } n_0 = \frac{1}{k-1}[N - \frac{\sum_{i=1}^{k} n_i^2}{N}], \text{ and } N = \sum_{i=1}^{k} n_i,$$

$$W.S.V. = \frac{1}{N-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2, \; B.S.V. = \frac{1}{k-1}\sum_{i=1}^{k} n_i(\bar{y}_i - \bar{y})^2,$$

$$\bar{y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}, \text{ and } \bar{y} = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}.$$

b. Parent – Offspring Correlation

Let $\widetilde{\mu}_p = \sum_{i=1}^{k} n_i x_i / N, \; \widetilde{\mu}_s = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}.$

The proposed pairwise estimator of the parent-offspring correlation is given by:

$$\widetilde{\rho}_F = \frac{\sum_{i=1}^{k}(x_i - \widetilde{\mu}_p)\sum_{j=1}^{n_i}(y_{ij} - \widetilde{\mu}_s)/n_i}{\left[\left(\sum_{i=1}^{k}(x_i - \widetilde{\mu}_p)^2\right)\left(\sum_{i=1}^{k}\frac{1}{n_i}\sum_{j=1}^{n_i}(y_{ij} - \widetilde{\mu}_s)^2\right)\right]^{1/2}} \tag{5}$$

Since the standard error for $\widetilde{\rho}_{ss}$ is difficult to obtain, we shall use the bootstrap method to find its standard error.

3.3. Bootstrap standard error and the delta method:

In a recent paper Field and Walsh [16] suggested several approaches to implementing the bootstrap method with clustered data. In a random sample of k clusters each of size n, they considered the observations as fixed with inferences constructed with respect to the random sampling mechanism. In this case their

main concern was with the accommodation of different forms of cluster sampling. One of the simplest approaches suggested was the so-called "cluster bootstrap". Roberts and Fan [17] implemented a specific form of bootstrap cluster sampling which they named "nested bootstrap" using the PROC MIXED procedure in SAS [18]. Note that the PROC MIXED was developed to fit normally hierarchical data, and is not appropriate for the analysis of binary response data. We therefore modified the bootstrap macro so that PROC GLM [18] is used to calculate the between and within mean squares of the appropriate analysis of variance (see Appendix). F this we obtain the bootstrap replications, and hence the bootstrap estimator, which we denote by $\widetilde{\rho}_{ssboot}$.

4. Regression estimator of parent-sib correlation:

4.1. Least square regression

The bivariate representation given in (3) has an interesting property that can be used to construct an alternative estimator of the parent-sib correlation. It can be shown that the regression of Y_i on x_i is linear in ρ_{12}. That is:

$$E(Y_i|x_i) = n_i\mu_s + \rho_{12}\frac{\sigma_{si}}{\sigma_p}(x_i - 2\mu_p), \text{ or } E\left(\frac{Y_i}{n_i}|x_i\right) = \mu_s + \rho_{12}\frac{\sigma_{si}}{n_i\sigma_p}(x_i - 2\mu_p).$$

The regression equation may then be written as:

$$\bar{y}_{i\cdot} = \frac{y_i}{n_i} = \mu_s + \rho_{12}\widetilde{Z}_i + \varepsilon_i \qquad (6)$$

In equation (6), $\widetilde{Z}_i = \frac{\hat{\sigma}_{s_i}}{n_i\hat{\sigma}_p}(x_i - 2\hat{\mu}_p)$, and ε_i has mean zero and unknown variance γ, whose estimator may be obtained from the mean square of the regression residuals. Therefore, the parameter ρ_{12} can be estimated by the method of least squares, obtaining

$$\widetilde{\rho}_{12} = \sum_{i=1}^{k}(\bar{y}_{i\cdot} - \mu_s)\widetilde{Z}_i \Big/ \sum_{i=1}^{k}\widetilde{Z}_i.$$

The estimated variance of this estimator is given approximately by

$$\hat{\text{var}}(\widetilde{\rho}_{12}) = \hat{\gamma} \Big/ \sum_{i=1}^{k}\widetilde{Z}_i^2.$$

With, $\hat{\gamma}$ given by

$$\hat{\gamma}=\frac{1}{k-2}\sum_{i=1}^{k}\left(\tilde{y}_{i\cdot}-\tilde{\rho}_{12}\tilde{Z}_i\right)^2,\ \text{where}\ \tilde{y}_{i\cdot}=\bar{y}_{i\cdot}-\hat{\mu}_s.$$

A $(1-\alpha)100\%$ confidence interval for ρ_{12} can be constructed using either the standard Wald formulation $\tilde{\rho}_{12}\pm t_{\alpha/2,k-2}\sqrt{\text{var}(\tilde{\rho}_{12})},$ or alternatively, using the well-known Fisher's Z transformation for the Pearson correlation coefficient.

4.2. The Generalized Estimating Equations (GEE)

The (GEE) approach may be used to estimate both the sib-sib correlation and a measure of parent-sib similarity, such as the odds ratio. The GEE approach requires only correct specification of the mean and variance of the response, with the sib-sib correlation estimated using a "working correlation". When we model the sibling binary response as a function of the total parent score, we obtain an estimate of the log-odds ratio and a robust sandwich estimator of its standard error. The estimated odds ratio, which has a population-averaged estimate interpretation, can be used as a measure of parent-sib similarity in place of the interclass correlation, while the standard error is obtained using the delta method.

5. Example: Mial and Oldham's blood pressures data

The data used for illustration here are obtained from a survey that aimed at assessing the levels of similarity in systolic and diastolic blood pressure among family members living within 25 miles of Rhonda Fach Valley in South Wales and published by Miall and Oldham [19]. Observations were made on parents and their offspring, with each observation consisting of systolic and diastolic blood pressures measured to the nearest 5mm Hg. However among 250 sampled families, only 204 contained information on brothers and sisters. Furthermore, because of the impossibly low systolic blood pressure (15mm Hg) for one daughter, another family was omitted .leaving 203 families for the analysis. Since these data were given on a continuous scale, we dichotomized the observations as described below:

For an individual whose blood level was above 130/80, the assigned binary score was 1, otherwise it was 0. The results of the data analysis are summarized in Table 2:

Table 2: Analysis of Miall and Oldham data.

Parameter	MLE	SE	
$\hat{\mu}_p$	0.258	.0369	
$\hat{\rho}_p$	0.0159	.0288	
$\hat{\mu}_s$	0.1134	0.0145	
$\hat{\rho}_s$	0.208	.0438	
$\hat{\rho}_{12}$	0.027	0.053	Regression $\tilde{\rho}_{12} = 0.163(0.1033)$

$\text{cov}(\hat{\rho}_s, \hat{\rho}_p) = 0.0003$, $\text{cov}(\hat{\rho}_s, \hat{\rho}_{12}) = 0.0010$, and $\text{cov}(\hat{\rho}_p, \hat{\rho}_{12}) = 0.0008$.

Note that the estimators proposed by Karlin et al and the MLE's are numerically very similar. Using the data in the example, we have $\tilde{\rho}_{ssboot} = 0.194$, with the bootstrap empirical standard error given by 0.01711. The parent-sib correlation is estimated as $\tilde{\rho}_F = 0.188$, with standard error 0.023. We also note that the estimator of the parent-sib correlation estimator obtained using maximum likelihood differs from that obtained by the least squares estimators. This is not unusual, as two estimators of the same parameters can be numerically different and have different distributional properties. But it requires further investigation.

Remark:

One of the advantages of using least square for estimating the parent-offspring correlation is that the estimator and its standard error are available in closed forms. Moreover, if we assume that the error term in (6) is normally distribution then one may perform exact statistical inferences on the model parameters, although this approach should be preceded by residual diagnostics [20]. If the required assumptions for the least squares are not satisfied, the MLE approach is a reasonable alternative because it yields estimators that have asymptotically optimal properties [21].

Using PROC GENMOD in SAS [22] the estimated odds ratio, which may be used as a measure of similarity between parents and their offspring, was given by 0.847 and its standard error based on the robust sandwich estimator by 0.140.

The p-value for testing the null hypothesis that the population odds ratio equals one is given by 0.274. The working correlation when specified within the GEE methodology as "exchangeable" was 0.173 and can be regarded as a population averaged measure of sib-sib correlation. Note that this correlation

estimate is quiet close to its maximum likelihood counterpart. A good summary of the GEE methodology can be found in many texts, for example chapters 12 and 13 in [23]. The SAS code for the *GEE model is:*

Proc Genmod;
Class familyid;
Model $y_{ij} = x_i$ */ dist= bin link = logit;*
Repeated subject=familyid / type=exch corrw;
Run;

Another approach to deriving the standard error of the bootstrap estimator of the sib-sib correlation is to use the delta method. In general if $\hat{\theta}_1$ and $\hat{\theta}_2$ are two statistics, and $g(\hat{\theta}_1, \hat{\theta}_2)$ is a non-linear differentiable function, then the first order approximation of the variance of g is given by:

$$\text{var}(g(\hat{\theta}_1, \hat{\theta}_2)) \approx (\frac{\partial g}{\partial \theta_1})^2 \text{var}(\hat{\theta}_1) + (\frac{\partial g}{\partial \theta_2})^2 \text{var}(\hat{\theta}_2) + 2(\frac{\partial g}{\partial \theta_1})(\frac{\partial g}{\partial \theta_2}) \text{cov}(\hat{\theta}_1, \hat{\theta}_2)$$

In the case of the sib-sib correlation we may take $\hat{\theta}_1 \equiv B.S.V.$, and $\hat{\theta}_2 \equiv W.S.V.$ The empirical variances and covariance of $\hat{\theta}_1$ and $\hat{\theta}_2$ are obtained from the bootstrap replications.

Application of the delta method gives a standard error of $SE(\tilde{\rho}_{ssboot}) = 0.0336.$ The 2000 replication bootstrap estimate of Karlin's parent-sib correlation is $\tilde{\rho}_{Fboot} = 0.189,$ and its empirical bootstrap standard error is 0.102. Figures 1 and 2 give the histograms of the 2000 bootstrap replications of the sib-sib and parent-sib correlations respectively.

6. Discussion

The likelihood inference procedure discussed in this paper allows parametric estimation of familial correlations when the responses are binary. The early work by mathematical statisticians to develop distributions that allow for dependency in the presence of multiple levels of hierarchy made it possible for us to construct the likelihood function, and hence to derive the maximum likelihood estimates of the correlation parameters. The application of maximum likelihood estimation to familial data having two levels of structure, under the model proposed by Sarmanov, is novel. Obtaining the solutions of the likelihood equations using the quasi Newton method was simplified here by the presence of only a few

parameters. However, the model is flexible enough to allow for inclusion of family level covariates, although the number of parameters to be estimated increases with the number of covariates included in the model. It should be mentioned that because the of the exchangeability assumption we can model the aggregate of the responses of both parents as well as the aggregate of responses of their siblings. For this reason the model cannot accommodate covariates measured at the individual level. There are however models that can accommodate covariates measured at both levels (family level and within family individual level) such as the Generalized Linear Mixed Model (GLIMMIX) and the Generalized Estimating Equations (GEE) based models. However, care must be taken when fitting these models to correlated data since these models consider the correlation parameters as nuisance, focusing instead on the estimation of regression coefficients. As a consequence of [5, 6], specification of the correlations between measures made on the units within a cluster is not required. However Crowder [24] demonstrated that the parameters involved in working correlation matrix are subject to "uncertainty of definition which can lead to a breakdown of the asymptotic properties of the estimators". A clear advantage of our proposed model is that it can be used to analyze data arising from more complex designs, such as those constructed using twins with more than two measured phenotypes for each twin (see; Shoukri and Donner [25]). Similarly, in the absence of parental responses, and when siblings are characterized by gender, then the analysis can focus on sex-specific correlations. For example one may test the equality of intraclass correlations for males and females, leading to the possible conclusion that gender is a risk factor in disease clustering. Testing this hypothesis can be carried out using either a likelihood ratio test or a score test. However the main focus of this paper is on estimation rather than hypothesis testing.

We conclude the discussion with two important remarks. First, it would be desirable to extend the models designed for the analysis of clustered binary data so that they can facilitate likelihood inferences if there is interest in analyzing multiple phenotypes. In addition to the intraclass and the interclass correlations for each phenotype, there may be interest in estimating several cross correlations including correlations between relatives on different phenotypes. This means that the extended model should accommodate a general interclass correlation structure, of which the present model is a special case. However, the estimation problem then requires future investigation. Second, although the use of the beta distribution is ubiquitous in the analysis of clustered binary data, a sensitivity analysis may be needed to assess the robustness of the resulting estimators when other distributions are used as priors for the response probability. In this regard a full Bayesian inference procedure may be adopted, where through Gibbs sampling several competing priors may be compared.

Appendix: SAS cod for the nested bootstrapping of blood pressures family data.

```
dm 'log;clear;output;clear';
/* SAS Program for Bootstrapping Individuals within each Family */

option pageno=1 nodate;
libname in  'd:\Shoukri';
libname out 'd:\Shoukri';

data sibsib;
 set in.karlinSibSib;
run;
** Direct the SAS log to a disk file to avoid SAS LOG Window becoming full;
proc printto log='D:\logfile.temp';
run;
%macro BTRAP;          ** start of bootdtrap macro 'BTRAP';
 %do Btrap=1 %to 2000;  ** 2000 bootstrapped samples;
 %do A=1 %to 203;       ** select each family sequentially;
   Data D1;
   set sibsib;          ** 223 families in the data set, unequal N in each family;
   if famid=&A;

  ** sampling with replacement within each selected family;
  ** bootstrapped sample size equal to the original sample size in each family;
  ** bootstrapped sample within each family is named BTDATA_n;
 Data btdata;
   drop I;
  do I=1 to N;
   IOBS=INT(ranuni (0) *N) + 1;
   set D1 point=IOBS nobs=N;
   output;
  end;
 stop;
     ** assign a unique random number for later combining data sets;
 Data BTdata_&A;
    set btdata;
    unique=rannor (0);
   %if &A=1 %then %do;
     Data BTdata_all;
       set btdata_&A;
   %end;
```

```
 %if &A>1 %then %do;
   proc sort data=btdata_all; by unique; run;
   proc sort data=btdata_&A;  by unique; run;

      ** combining bootstrapped samples from each family;
  Data BTdata_all;
   update btdata_all btdata_&A;
     by unique;
      run;
 %end;
%end;

 ** direct Proc GLM output to a file on disk;
 ** avoids potential problem of SAS output window becoming full;
Filename glmout 'D:\GlmFile';
proc printto print=glmout new;
run;

 ** direct GLM ODS tables to 2 datasets;
ods output
        OverallANOVA =TestAnova
        ModelANOVA=Subject;
ods listing close;
proc glm;
 class famid;
 model y=famid;
 random famid;
ods listing;

  ** extract the required variables from the 2 datasets;
 data result1 (keep= SIGERROR);
  set testanova;
  if Source='Error' then   SIGERROR=MS;
  if sigerror ne .;
 data result2(keep= Fam);
  set Subject;
  if Source='FamId' then   Fam=MS;
  if Fam ne .;
 run;
  ** re-direct the output to SAS output window;
 proc printto print=print;
```

```
 run;
  ** combine the two data sets to have one observation for;
  ** each bootstrapped sample;
 Data Both;
  set result1; set result2;
  ** append estimates from each bootstrap iteration;
  ** to a permanent SAS dataset on disk: Results;
 proc append base=out.Results force;
 run;
%end;              ** end bootstrap iterations;
%mend BTRAP;       ** end of bootstrap macro;
%Btrap;            ** execute the BTRAP macro;

  ** read the data of bootstrapped results;
  ** (2000 observations from 2000 bootstrap iterations);
 data temp;
 set out.Results;

  ** calculate SigmaB and RO variables;
 SigmaB=(fam-sigerror)/3.07;
 RO=sigmaB/(sigerror+sigmaB);

  ** direct all results to ODS file;
ods rtf body='d:\Families.rtf' style=minimal;
 proc print;
  ** obtain some basic descriptive statistics for;
  ** the bootstrapped distributions of the estimates;
 proc means n mean std min max maxdec=4;
  title1 'Bootstrap Individuals within each Family';
 run;
ods rtf close;
```

References

1- Laird, N., and Lange, C. *The fundamentals of Modern Statistical Analysis Springer,* NY. 2012.

2- Morton, N.E. Analysis of family resemblance. I. Introduction. *American Journal of Human Genetics*, 26:318-330, 1974.

3- Rao, D.C., Volger, P.G., McGue, M., and Russell, J.M. Maximum likelihood estimation of familial correlations from multivariate quantitative data on pedigrees: A general method and examples. *American Journal of Human Genetics*, 41:1104-1116, 1987.

4- Morton, N.E., and MacLean, C.J. Analysis of family resemblance. III Complex segregation analysis of quantitative traits. *American Journal of Human Genetics*, 26:489-503, 1974.

5- Liang, K-Y. and Zeger, S. Longitudinal data analysis using generalized linear models. *Biometrika*, 73:13-22, 1986.

6- Zeger, S., and Liang, K-Y. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, 42:121-130, 1986.

7- Cox, D.R. and Snell, E. *Analysis of Binary Data*, Chapman & Hall, London, UK, 1989.

8- Prentice, R. Correlated binary regression with covariates specific to each binary observation. *Biometrics*, 44: 1033-1048.

9- Crowder, M, Beta- binomial ANOVA for proportions. *Applied Statistics*, 27:34-37.

10- Bloch, D. And Kraemer C. 2 x 2 Kappa coefficient: Measure of agreement or association. Biometrics, 45: 269-287, 1989.

11- Sarmanov, O.V. Generalized normal correlation and two-dimensional Frechet classes. *Doklady (Soviet Mathematics)*, 168: 596-599, 1966.

12- Lancaster, H.O. *The chi-squared distribution*. New York: Wiley 1969.

13- Danaher, P.J. A canonical expansion model for multivariate media exposure distributions: generalization of the duplication of viewing law. *Journal of Marketing Research*, 28, 361-367, 1991.

14- Karlin, S., Cameron, E.C., and Williams, P.T. Sibling and parent-offspring correlation estimation with variable family size. *Proceedings of the National Academy of Science*, Vol. 78, No.5, 2664-2668, 1981.

15- Richards, F.G. A method of maximum likelihood estimation. *Journal of the Royal Statistical Society*, B 23, 469-475, 1961.

16- Field C.A., and Welsh, A.H. Bootstrapping clustered data. *Journal of the Royal Statistical Society, B*: 69, Part 3, 369-390, 2007.

17- Roberts, J.K. and Fan, Xitao. Bootstrapping within the Multilevel/ Hierarchical linear modeling framework: A primer for use with SAS and SPLUS. *Multiple Linear Regression Viewpoints*, Vol. 30(1): 23-34, 2004.

18- Littell, R., Milliken, G., Stroup, W., Wolfinger, R., Schabenberger, O. *SAS for mixed models*. 2nd edition, SAS Institute. Cary, NC, US 2006.

19- Miall, WE, and Oldham, PO. A study of arterial blood pressure and its inheritance in a sample of the general population. *Clinical Science*, 14: 459-487, 1955.

20- Cook, RD, and Weisberg S. *Residuals and influence in regression*. London, Chapman and Hall, 1982.
21- Cox, DR, and Hinkley DV. *Theoretical statistics.* London, Chapman and Hall, 1974.
22- SAS/STAT 9.2 User's Guide. *The GENMOD Procedure*. SAS Institute, Cary, NC, US 2006.
23- Fitzmaurice, G., Laird, N., Ware, J. *Applied longitudinal analysis*, 2nd edition, Wiley New York, 2011.
24- Crowder, M. On the use of a working correlation matrix in using generalized linear models for repeated measures. *Biometrika*, 82; 2:407-410 1995.
25- Shoukri, M.M., and Donner, A. Bivariate models for co-aggregation of dichotomous traits in twins. *Statistics in Medicine*, 26-2, 336-21, 2007.

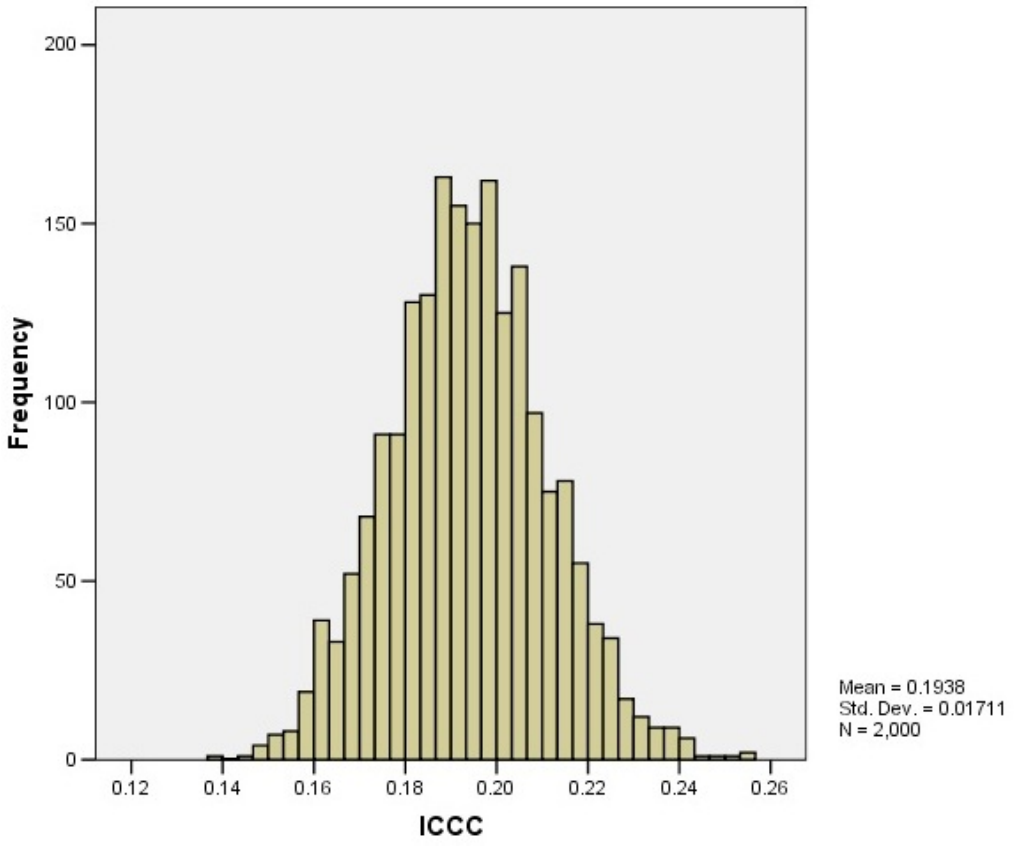

Figure 1: Histogram of 2000 bootstrap samples of Karlin's sib-sib intraclass correlation coefficient (ICCC).

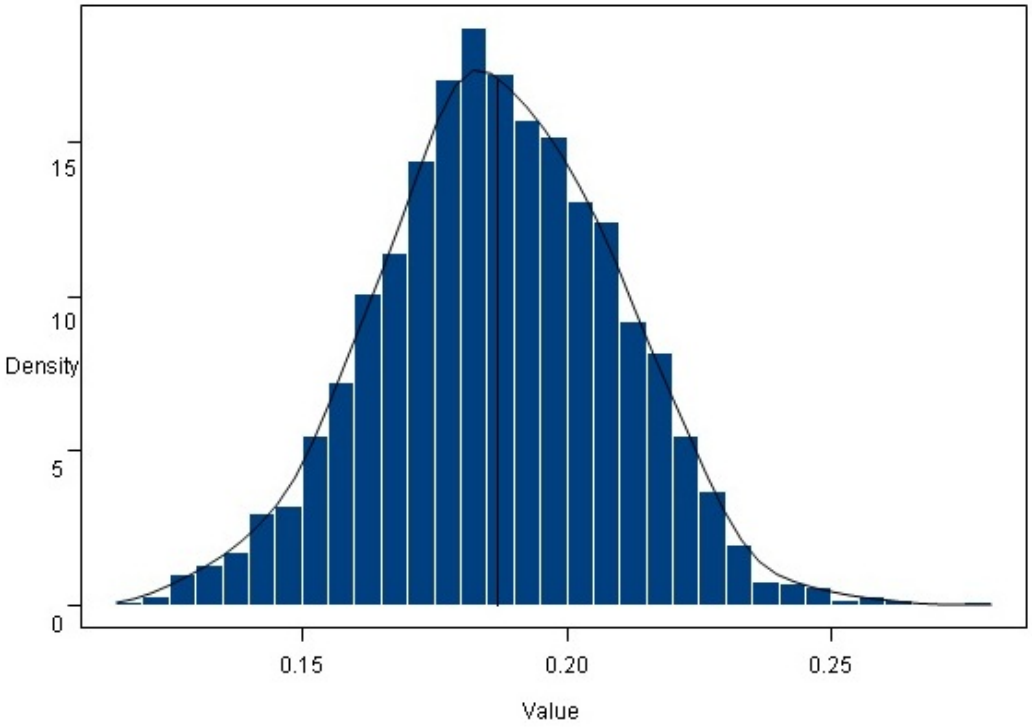

Figure 2: Histogram of 2000 bootstrap samples of Karlin's parent-sib correlation. The mean is 0.187, and the bootstrap standard error is 0.023.