BMC Genetics

BioMed Central

Proceedings

Open Access

Imputation methods for missing data for polygenic models

Brooke Fridley*[1,3], Kari Rabe[2] and Mariza de Andrade[2]

Address: [1]Department of Statistics, Iowa State University, Ames, Iowa, USA, [2]Division of Biostatistics, Department of Health Sciences Research, Mayo Clinic, Rochester, Minnesota, USA and [3]Department of Mathematics, University of Wisconsin-La Crosse, La Crosse, Wisconsin, USA

Email: Brooke Fridley* - Fridley.broo@uwlax.edu; Kari Rabe - Rabe.Kari@mayo.edu; Mariza de Andrade - mandrade@mayo.edu

* Corresponding author

from Genetic Analysis Workshop 13: Analysis of Longitudinal Family Data for Complex Diseases and Related Risk Factors
New Orleans Marriott Hotel, New Orleans, LA, USA, November 11–14, 2002

Published: 31 December 2003

BMC Genetics 2003, **4**(Suppl 1):S42

This article is available from: http://www.biomedcentral.com/1471-2156/4/s1/S42

Abstract

Methods to handle missing data have been an area of statistical research for many years. Little has been done within the context of pedigree analysis. In this paper we present two methods for imputing missing data for polygenic models using family data. The imputation schemes take into account familial relationships and use the observed familial information for the imputation. A traditional multiple imputation approach and multiple imputation or data augmentation approach within a Gibbs sampler for the handling of missing data for a polygenic model are presented.

We used both the Genetic Analysis Workshop 13 simulated missing phenotype and the complete phenotype data sets as the means to illustrate the two methods. We looked at the phenotypic trait systolic blood pressure and the covariate gender at time point 11 (1970) for Cohort 1 and time point 1 (1971) for Cohort 2. Comparing the results for three replicates of complete and missing data incorporating multiple imputation, we find that multiple imputation via a Gibbs sampler produces more accurate results. Thus, we recommend the Gibbs sampler for imputation purposes because of the ease with which it can be extended to more complicated models, the consistency of the results, and the accountability of the variation due to imputation.

Background

Methods to handle missing data have been a statistical area of research for many years [1]. Little has been done within the context of pedigree analysis. The goals of this paper are: 1) to present two imputation methods for missing phenotype information, and 2) to compare estimates of the additive polygenic effect using variance components or mixed models between the Genetic Analysis Workshop 13 (GAW13) simulated missing phenotype and the complete phenotype data sets for each imputation method [2-4]. In narrowing the focus of our topic, we only looked at the phenotypic trait systolic blood pressure and the covariate gender at time point 11 (1970) for Cohort 1 and time point 1 (1971) for Cohort 2. The methods for imputation described herein include traditional multiple imputation and multiple imputation (data augmentation) via a Gibbs sampler, with both methods accounting for the familial information in the imputation. In fitting the polygenic model to produce estimates of the overall mean effect, gender effect, additive genetic variance, and the residual error variance, we used the expectation-maximization (EM) algorithm program PolyEM [5] and a Bayesian analysis involving use of a Gibbs sampler.

Methods

Traditional multiple imputation method

Multiple imputation is carried out by using the conditional distribution of the missing values given the observed values. Let $\mathbf{Y}_i = (\mathbf{Y}^T_{mis}, \mathbf{Y}^T_{obs})^T$ be the quantitative

phenotype values for the i^{th} family with $\mathbf{Y}_{mis}$ representing the vector of missing values and $\mathbf{Y}_{obs}$ representing the vector of observed values. Similarly, we can partition the mean and covariance matrix. Thus, for the polygenic model (including an overall mean and a gender effect), the distribution of $\mathbf{Y}_{mis}$ given $\mathbf{Y}_{obs}$ is

$$\mathbf{Y}_{mis} \mid \mathbf{Y}_{obs} \sim N(\mu_1, \Sigma_1),$$

with $\mu_1 = (\mu + X_1\beta) + (\sigma^2 D_{12})(\sigma^2 D_{22} + \tau^2 I_{22})^{-1}(\mathbf{Y}_{obs} - (\mu + X_2 \beta))$

and $\Sigma_1 = (\sigma^2 D_{11} + \tau^2 I_{11}) - (\sigma^2 D_{12})(\sigma^2 D_{22} + \tau^2 I_{22})^{-1} (\sigma^2 D_{21})$.

The imputation is carried out by generating missing values from the conditional multivariate normal distribution, taking into account the family structure. This imputation is completed *m* times to produce *m* complete data sets to analyze. From these *m* analyses, the final point estimate would be the mean of the *m* estimates. The computation of the standard error can be done by first computing the between-imputation variation, B_m, and the within-imputation variation, W_m. Then, the total variability is $T_m = W_m + B_m (m + 1)/m$. Confidence intervals for parameters of interest use the *t*-distribution with degrees of freedom $(m - 1) (1 + 1/ (m + 1)*W_m/B_m)^2$ [6].

The one problem with this imputation scheme is that values for μ, β, σ^2, and τ^2 are required for the imputation. To address this issue, we ran the analysis on the observed data and used the resulting estimates to create *k* complete data sets, from which *k* sets of point estimates are found. Then, the average of the *k* sets of estimates is used for the *m* multiple imputations. Point estimates were found using the EM algorithm program PolyEM with $k = 5$ and $m = 25$ [5,7]. Variance estimates were found using the Fisher Information matrix and the large sample properties of maximum likelihood estimates (MLEs) [4,7,8].

Multiple imputation via Gibbs sampler

Another approach for the imputation of missing data is through a Bayesian analysis via a Gibbs sampler. The Gibbs sampler is a particular Markov chain algorithm that is useful when working with high dimensional problems. In addition to the traditional use of the Gibbs sampler, an imputation step can also be added to impute missing values. A multiple imputation scheme can be implemented by having an imputation step at the beginning of the Gibbs sampler. For each iteration, a new imputation is done, giving multiple imputations for each missing value [9-15]. For a detailed proof of the data augmentation methodology, see Tanner and Wong [12].

A Bayesian polygenic model with non-informative prior distributions for the i^{th} family is $\mathbf{Y}_i = X_i\beta + \mathbf{a}_i + \varepsilon_i$, where $\mathbf{Y}_i$ is a vector containing the individual responses in family i, X_i is a design matrix containing covariate information, β is a vector of covariate effects, $\mathbf{a}_i$ is a vector of random family effects where $\mathbf{a}_i \sim MVN(0, \sigma^2 D_i)$ and D_i is a known coefficient of relationship matrix, and $\varepsilon_i \sim MVN(\mathbf{0}, \tau^2 I)$. Non-informative priors were then placed on all other parameters in the model, i.e., $p(\beta)$ proportional to **1**, $p(\sigma^2)$ proportional to $1/\sigma^2$, and $p(\tau^2)$ proportional to $1/\tau^2$ [6].

The steps for implementing the Gibbs sampler with an imputation step for the Bayesian polygenic model follow.

1. Set starting values for $\beta^{(0)}$, $\sigma^{2(0)}$, $\tau^{2(0)}$, $\mathbf{a}_i^{(0)}$ for all i = 1,..., k, and set m = 1 (iteration).

2. If y_{ij} is missing, impute y_{ij} by simulating an observation from $N(\mathbf{X}_{ij} \beta^{(m-1)} + a_{ij}^{(m-1)}, \tau^{2(m-1)})$.

3. Generate $\beta^{(m)}$ from $MVN((X^TX)^{-1}X^T(\mathbf{y} - \mathbf{a}^{(m-1)}), \tau^{2(m-1)}(X^TX)^{-1})$.

4. Generate $\tau^{2(m)}$ from $INGAM(N/2, 1/2\ \Sigma(\mathbf{y}_i - X_i\beta^{(m)} - \mathbf{a}_i^{(m-1)})^T(\mathbf{y}_i - X_i\beta^{(m)} - \mathbf{a}_i^{(m-1)}))$.

5. Generate $\sigma^{2(m)}$ from $INGAM(N/2, 1/2\ \Sigma\ \mathbf{a}_i^{(m-1)T} D_i^{-1} \mathbf{a}_i^{(m-1)})$.

6. Generate $\mathbf{a}_i^{(m)}$ from $MVN(\mu_a^{(m)}, V_a^{(m)})$, where $\mu_a^{(m)} = (1/\sigma^{2(m)} * D_i^{-1} + 1/\tau^{2(m)} * I_i)^{-1} * (1/\ \tau^{2(m)} * I_i\ (\mathbf{y}_i - X_i\beta^{(m)}))$ and $V_a^{(m)} = (1/\sigma^{2(m)} * D_i^{-1} + 1/\tau^{2(m)} * I_i)^{-1}$.

7. After one iteration of the algorithm, you have $(\beta^{(m)}, \sigma^{2(m)}, \tau^{2(m)}, \mathbf{a}^{(m)})$. Set m = m + 1 and repeat steps 1 through 6.

Approximate $(1 - \alpha)$% posterior confidence intervals are found by taking the $\alpha/2$ and the $1-\alpha/2$ percentiles of the simulated marginal posterior distributions for parameters of interest. The simulated posterior distributions will reflect the uncertainty in the estimation of the parameter along with the uncertainty due to the imputation of the missing data.

Results

The two imputation methods were run for three replicates of the GAW13 simulated complete and missing data sets for the phenotypic trait of systolic blood pressure and the covariate of gender at time point 11 (1970) for Cohort 1 and time point 1 (1971) for Cohort 2. We limited our analysis to two odd-numbered replicates and one even-numbered replicate to demonstrate the methods due to time and computational restrictions. Table 1 displays the results for the GAW13 missing and complete data sets, in

Table 1: 95% Confidence intervals using "traditional" multiple imputation within a likelihood analysis for the complete and missing simulated data

	Missing Data		Complete Data	
Rep	**Polygenic VC**	**Error VC**	**Polygenic VC**	**Error VC**
003	(77.68, 90.99)	(152.25, 189.97)	(86.08, 133.07)	(145.59, 188.12)
004	(71.29, 96.34)	(151.42, 190.75)	(81.58, 127.39)	(151.10, 189.15)
019	(75.86, 88.50)	(154.40, 188.91)	(65.23, 107.04)	(154. 90, 191.57)

Table 2: 95% Approximate posterior intervals using data augmentation within a Bayesian/Model for the complete and missing simulated data

	Missing Data		Complete Data	
Rep	**Polygenic VC**	**Error VC**	**Polygenic VC**	**Error VC**
003	(90.38, 129.60)	(148.30, 182.30)	(88.47, 131.90)	(151.50, 186.30)
004	(80.30, 122.50)	(147.90, 183.20)	(86.75, 126.70)	(153.40, 185.40)
019	(67.49, 95.44)	(157.50, 188.10)	(71.13, 104.80)	(158.90, 189.30)

which a likelihood analysis using an EM algorithm using the traditional imputation approach for the missing data. Table 2 displays the results for the GAW13 missing and complete data sets, in which a Bayesian model was used involving a data augmentation step for missing data.

The confidence intervals show little difference between the complete and missing data set analysis within each type of imputation method with regards to the estimates for the error variance component. The data augmentation within a Gibbs sampler gives intervals similar to those of the complete data set, while the traditional imputation approach for missing data produced differing intervals when compared with the complete data set analysis. Not only are the confidence intervals narrower after the imputation of missing data, but the point estimates for the polygenic variance component are much smaller (underestimated) as compared to the results of the complete data sets for the traditional imputation method. These intervals show that the traditional imputation method produced inaccurate intervals as compared with the intervals for the complete data set. The inaccuracy may be due to the choice of the parameter values for the imputation or the degrees of freedom used for the intervals [16].

Conclusions

Missing data complicates statistical analysis. One way to deal with missing data is through the use of imputation. In the case of family data, the information provided by the dependence structure can be utilized in the imputation of missing data. We have discussed two methods of imputation and illustrated their application through the GAW13 simulated data sets. We assumed the missing data mechanism in the GAW13 data set was ignorable. The Gibbs sampler and the traditional multiple imputation method can be easily applied to non-ignorable missing data, such as in cases involving censored phenotype data. In the case of non-ignorable missingness, not adjusting for the missing/censored values will lead to biased estimates. Based on results from the three replicates, the Gibbs sampler approach seems to give more accurate confidence intervals, as opposed to the traditional multiple imputation approach. In addition, the traditional imputation method has the drawback of needing parameter values for the completion of the imputation. The Gibbs sampler approach does not encounter this problem, because any set of estimates may be used as starting values for the algorithm. In conclusion, we recommend the Gibbs sampler for imputation purposes because of the ease with which it can be extended to more complicated models, the consistency of the results, and the accountability of the variation due to imputation. Future work is planned to investigate the use of data augmentation for missing phenotype data in longitudinal studies and quantitative trait locus analysis.

Acknowledgments

We thank Curtis Olswold for his help. This research was partially founded by NIH grant R01HL71917.

References

1. Little RJA, Rubin D: **Statistical Analysis with Missing Data.** *New York, Wiley* 22002.
2. Hopper J, Mathews J: **Extensions to multivariate normal models for family analysis.** *Ann Hum Genet* 1982, **46:**373-383.
3. de Andrade M, Amos C, Thiel T: **Methods to estimate genetic components of variance for quantitative traits in family studies.** *Genet Epidemiol* 1999, **17:**64-76.
4. Lange K, Westlake J, Spence M: **Extensions to pedigree analysis. III. Variance components by scoring method.** *Ann Hum Genet* 1976, **39:**485-491.
5. Thompson E, Shaw R: **Pedigree analysis for quantitative traits: variance components without matrix inversion.** *Biometrics* 1990, **46:**399-413.
6. Gelman A, Carlin J, Stern H, Rubin D: **Bayesian Data Analysis.** *New York, Chapman & Hall* 1995.
7. Dempster A, Laird N, Rubin D: **Maximum likelihood from incomplete data via the EM algorithm (with discussion).** *J R Stat Soc B* 1977, **39:**1-38.
8. Rao C: **Linear Statistical Inference and Its Applications.** *New York, Wiley* 21973.
9. Hopke P, Liu C, Rubin D: **Multiple imputation for multivariate data with missing and below-threshold measurements: time-series concentrations of pollutants in the arctic.** *Biometrics* 2001, **57:**22-33.
10. Gelfand A, Smith A: **Sampling-based approaches to calculating marginal densities.** *J Am Stat Assoc* 1990, **85:**398-409.
11. Geman S, Geman D: **Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images.** *IEEE Trans Pattern Analysis Machine Intelligence* 1984, **6:**721-741.
12. Tanner M, Wong W: **The calculation of posterior distributions by data augmentation.** *J Am Stat Assoc* 1987, **82:**528-540.
13. Li K: **Imputation using Markov chains.** *J Stat Comput Simul* 1988, **30:**57-79.
14. Van Dyk D, Meng X: **The art of data augmentation.** *J Comput Graph Stat* 2001, **10:**1-50.
15. Schafer J: **Analysis of Incomplete Multivariate Data.** *New York, Chapman & Hall* 1997.
16. Barnard J, Rubin D: **Small-sample degrees of freedom with multiple imputation.** *Biometrika* 1999, **86:**948-955.