A multiple imputation procedure of censored values in family-based genetic association studies

Fabiola Del Greco M., Cristian Pattaro, Cosetta Minelli, Peter P. Pramstaller, John R. Thompson

Abstract Biological quantitative data are subjected to censoring when a portion of values cannot be quantified because they are smaller or greater than the limit of detection (LOD) of the laboratory assay. In genetic association studies of quantitative trait, the handling of censored data has received little attention and often the solutions are unsatisfactory. However, the approach used to deal with such data can have a substantial impact on the results of the analysis. While the Tobit model represents an appropriate method for independent data, there is no evidence on its performance in the presence of non-independent observations, typical of family- or pedigree-based studies. In the context of a family-based study, we propose a Bayesian approach which takes into account the uncertainty of the imputation procedure using several imputations for each censored value. In particular, assuming vague (uninformative) priors for all hyper-parameters, the imputation based on Gibbs sampling is applied to variance-components linear regression models, where the primary outcome is related to a secondary outcome. Through simulation, we describe the behavior of the Tobit model in the presence of different degrees of censoring and heritability of the trait compared with the Bayesian model and the naïve approach of replacing all censored values with the LOD value.

Key words: Censored data; Tobit model; Multiple imputation; Gibbs sampling; Genetic association studies

F. Del Greco M. · C. Pattaro · C. Minelli · P.P. Pramstaller
Center for Biomedicine, EURAC research, Viale Druso 1, Bolzano, Italy
e-mail: fabiola.delgreco@eurac.edu

J.R. Thompson
Department of Health Sciences, Room 241F, Adrian Building, University of Leicester, Leicester LE1 7RH, UK

1 Introduction

In biomedical research, laboratory assay's limitations often do not enable the full range measurement of biomarker levels and a portion of values cannot be detected (NDs) because they fall outside the instrumental limit of detection (LOD). In particular, biological data are mainly subjected to left censoring (i.e. NDs fall below LOD), due to the difficulty of quantifying very low concentrations of the biological parameter.

To handle censored data, several approaches have been proposed. They can be classified as (i) single imputation, (ii) distributional, or (iii) robust methods [6]. Despite being quite arbitrary and leading to unsatisfactory results, single imputation methods, that replace NDs with a constant value (e.g. LOD; LOD$/2$), are widely used. The second group includes multiple imputation (MI) procedure, which creates several sets of replacements sampling values from an underlying distribution [8]. The third group includes methods that fit the data to a distribution, by either MLE or probability plot procedures, which is only used to extrapolate NDs. This class includes the Tobit [9] and censored quantile regression models [7] and all traditional nonparametric methods [2].

Genetic association studies aim to test whether some genetic markers are associated with specific biomarker levels [1], often based on related individuals. A single imputation or the omission of NDs are generally used in this context. The main purpose of this work is to understand how estimates could be affected by the strategy followed to handle censoring in the presence of dependent observations. We develop a Bayesian MI procedure which uses correlated biomarkers which are measured together for practical and economic reasons, and takes into account the family structure of the study sample. The new method is compared with single imputation and Tobit models through an exhaustive set of simulations.

2 Bayesian multiple imputation procedure

Let $\mathbf{Y} = (\mathbf{y}_i)_{i=1}^n$ denote a censored biomarker collected in n family trios (three relatives). Suppose that each $\mathbf{y}_i$ is distributed like a $N(\mu_i, \mathbf{V}_i)$, where $\mu_i = \beta \mathbf{x}_i$ is a linear function of a completely observed biomarker $\mathbf{X} = (\mathbf{x}_i)_{i=1}^n$ and $\mathbf{V}_i = \sigma_g^2 \mathbf{K}_i + \sigma_e^2 \mathbf{F} + \sigma^2 \mathbf{I}$ is the variance-covariance matrix among relatives, where σ_g^2 is the additive genetic variance, σ_e^2 is the environmental variance, σ^2 is the residual variance in the model, $\mathbf{K}_i$ is the matrix of relatedness (kinship) coefficients each describing in terms of probabilities the pairwise genetic similarities, $\mathbf{F}$ is a matrix whose entries are equal to one and $\mathbf{I}$ is the identity matrix. The ratio $\sigma_g^2/(\sigma_g^2 + \sigma_e^2)$ can be interpreted as the genetic heritability of a biomarker, that is the degree to which it is transmitted from one generation to the next [5].

Our aim is to test the association between $\mathbf{Y}$ and a biallelic marker $\mathbf{S} \sim Bi(p, 2)$, i.e. the number of alleles each selected independently with probability p, using a lin-

ear mixed effects model (LMM) with random individual effect correlated according to the kinship coefficients [4], that is $\mathbf{y}_i = \mathbf{z}_i + \alpha\mathbf{s}_i + \beta\mathbf{x}_i + \varepsilon$, where $\mathbf{z}_i \sim N(0, \sigma_g^2\mathbf{K}_i)$ and $\varepsilon \sim N(0, \mathbf{I})$.

In this context we provide a new MI procedure. The key idea is to create m plausible sets of replacements from a Bayesian model, and to combine the results of the m analyses, in order to take into account the uncertainty in the imputed values [10]. At each family level i, NDs are independently generated from the posterior density $P(y_{c,i}|y_{o,i}, \theta)$, where $y_{c,i}$ and $y_{o,i}$ are censored and observed values respectively, and θ is the vector of unknown parameters $(\beta, \sigma_e^{-2}, \sigma_g^2/\sigma_e^2)$, which are randomly drawn from their conditional distributions $P(\theta|y_{o,i})$ assuming vague priors. This method of drawing follows the Gibbs sampling procedure [3]. The convergence of the chains is usually diagnosed by graphical tools, i.e. trace plot and auto-correlogram.

3 Simulations and results

The performance of the MI procedure is evaluated in the context of family-based genetic association analysis. We randomly generated 100 family trios and assessed several scenarios varying the censoring rate (.2, .4) and the biomarkers correlation $(.8, .5, .2)$. Assuming a genetic heritability of .35 and an explained genetic variance of 5%, which corresponds to a genetic effect of .29, for each scenario, we generated $1,000$ datasets from these trivariate distributions $\mathbf{x}_i \sim N(0, \mathbf{V}_i)$; $\mathbf{s}_i \sim Bi(.4, 2)$; $\mathbf{y}_i \sim N(\beta\mathbf{x}_i + .29\mathbf{s}_i, \mathbf{V}_i)$. We left censored each dataset and followed these strategies: (i) a single imputation, replacing NDs with the LOD; (ii) the Tobit model with robust standard error estimation, randomly drawning residuals of NDs from a Normal distribution centered on their predicted values; (iii) the MI procedure repeating the imputation 5 times. After dealing with the censoring problem we tested the genetic association using a LMM. In the analysis, for (ii) we tested the Tobit residuals; for (iii) 5 new datasets were analyzed separately and combined $2-5$ at a time, taking the mean of genetic effect estimates [10].

The genetic effect estimates are reported in Table 1 with the corresponding accuracy measures. The choice of the strategy followed when handling NDs could dilute the genetic effect. A single imputation or a Tobit method gave biased results with a quite small coverage probability, especially with large censoring rate. The MI procedure seems to have better performance with estimates very closed to the true value. As previously shown, a very small number of imputations $(2 \leq m \leq 5)$ is sufficient in order to judge the efficiency of the estimators [10].

4 Conclusions

Despite being computationally intensive, a Bayesian MI procedure which borrows strength from other, usually observed, biomarkers can outperform the more sim-

Table 1 Genetic association analysis with a censored biomarker: genetic effect estimates (Estimate) and standard error (se) with Tobit model (Tobit), single imputation (Naïve) and new multiple imputation (MI) procedure with m equal to 3 and 5. The true genetic effect is .29.

		c = .2					c = .4				
ρ	Method	Estimate	se	rmse	cp(%)	Lc	Estimate	se	rmse	cp(%)	Lc
.8	Tobit	.24	.08	.05	86	.79	.20	.07	.09	72	.49
	Naïve	.23	.08	.06	87	.74	.18	.07	.11	61	.39
	MI $m=3$	.27	.14	.02	96	.46	.24	.14	.05	95	.42
	MI $m=5$	.27	.14	.02	93	.46	.24	.14	.05	89	.42
.5	Tobit	.24	.08	.05	89	.79	.19	.07	.10	64	.49
	Naïve	.24	.08	.05	88	.77	.18	.06	.11	52	.41
	MI $m=3$	.28	.10	.01	94	.82	.26	.10	.03	94	.73
	MI $m=5$	.28	.10	.01	97	.82	.26	.10	.01	94	.74
.2	Tobit	.24	.07	.05	88	.79	.19	.06	.10	61	.44
	Naïve	.24	.08	.05	88	.78	.18	.06	.11	52	.39
	MI $m=3$	.28	.09	.01	95	.95	.26	.09	.03	95	.83
	MI $m=5$	.28	.09	.01	98	.95	.26	.09	.03	96	.84

For each method, censoring rate (c), correlation between biomarkers (ρ), root mean square error (rmse), coverage probability (cp), Lin's concordance correlation coefficient between estimates and true value (Lc).

plistic approaches in terms of both accuracy and precision. The advantage is more evident for large censoring rates. The behavior of the MI procedure under different causal relationships between the outcome **Y**, a covariate **X** and a genetic marker **S** are being investigated.

References

1. Cordell H., Clayton D.G.: Genetic association studies. Series Genet Epidemiol 3, Lancet **336**, 1121–1131 (2005)
2. Desu M.M., Raghavarao: Nonparametric statistical methods for complete and censored data. Chapman and Hall / CRC (2004)
3. Gelman A., Carlin J.B., Stern H.S., Rubin D.B.: Bayesian data analysis. 2th edition, Chapman and Hall (2004)
4. Kennedy B.W., Quinton M., van Arendonk J.A.: Estimation of effects of single genes on quantitative traits, J Anim Sci, **70**, 2000–2012 (1992)
5. Lange K.: Mathematical and statistical methods for genetic analysis. 2th edition, Springer (2003)
6. Lubin J.H., Colt J.C., Camann D., Davis S., Cerhan J.R., Severson R.K., Bernstein L., Hartge P.: Epidemiologic evaluation of measurement data in the presence of detection limits. Eviron Health Perspect, **112 (17)**, 1691–1696 (2004)
7. Portnoy S.: Censored quantile regression, JASA, **98**, 1001–1012 (2003)
8. Rubin D.B.: Inference and missing data (with discussion), Biometrika, **63**, 581–592 (1976)
9. Tobin J.: Estimation of relationship for limited dependent variables, Econometrica, **26 (1)**, 24–36 (1958)
10. Shafer J.L.: Multiple imputation: a primer, Stat Methods Med Res, **8**, 3–15 (1999)