Author/s:
Lee, KJ;Carlin, JB

Title:
Multiple imputation in the presence of non-normal data

Date:
2017-02-20

Citation:
Lee, K. J. & Carlin, J. B. (2017). Multiple imputation in the presence of non-normal data. Statistics in Medicine, 36 (4), pp.606-617. https://doi.org/10.1002/sim.7173.

Persistent Link:
https://hdl.handle.net/11343/292136

Multiple imputation in the presence of non-normal data

Katherine J Lee[1,2] and John B Carlin[1,2].

[1] Clinical Epidemiology and Biostatistics Unit, Murdoch Childrens Research Institute, Melbourne, Victoria, Australia

[2] Department of Paediatrics, University of Melbourne, Melbourne, Victoria, Australia

Corresponding Author: Katherine Lee

Clinical Epidemiology and Biostatistics Unit

Murdoch Childrens Research Institute

Flemington Road

Parkville

Melbourne

Victoria

Australia

Tel: 03 93456549

Key words: multiple imputation, missing data, non-normal data, transformation, predictive mean matching

Words: 4,123

Abstract

Multiple imputation (MI) is becoming increasingly popular for handling missing data. Standard approaches for MI assume normality for continuous variables (conditionally on the other variables in the imputation model). However, it is unclear how to impute non-normally distributed continuous variables. Using simulation and a case study we compared various transformations applied prior to imputation, including a novel non-parametric transformation, to imputation on the raw scale and using predictive mean matching (PMM) when imputing non-normal data.

We generated data from a range of non-normal distributions, and set 50% to missing completely at random or missing at random. We then imputed missing values on the raw scale, following a zero-skewness log, Box-Cox) or non-parametric transformation and using PMM with both type 1 and 2 matching. We compared inferences regarding the marginal mean of the incomplete variable and the association with a fully observed outcome. We also compared results from these approaches in the analysis of depression and anxiety symptoms in parents of very preterm compared with term-born infants.

The results provide novel empirical evidence that the decision regarding how to impute a non-normal variable should be based on the nature of the relationship between the variables of interest. If the relationship is linear in the untransformed scale, transformation can introduce bias irrespective of the transformation used. However, if the relationship is non-linear, it may be important to transform the variable to accurately capture this relationship. A useful alternative is to impute the variable using PMM with type 1 matching.

Introduction

Missing data arise in most clinical and public health research studies and multiple imputation (MI) has become a popular approach for conducting statistical analysis of incomplete data [1]. MI is a two-stage approach whereby missing values are imputed a number of times based on the available data and then each of the completed datasets is analysed using a standard method of analysis with results combined across the completed datasets using Rubin's rules [2]. When missingness is in a single variable, missing values may be imputed using a simple univariate regression model, with "proper" imputation requiring the incorporation of estimation uncertainty via the redrawing of parameter values for each imputation. When there are multiple incomplete variables (with a non-monotone pattern of missingness) there are two commonly used methods of imputation that are widely available in standard statistical software [3, 4]. The first, multivariate normal imputation (MVNI) assumes a joint normal distribution for all of the variables in the imputation model [5]. The alternative is fully conditional specification (FCS) or "chained equations", where each variable is imputed sequentially using a univariate regression model conditional on the other variables in the imputation model [6, 7]. Both of these multivariate imputation approaches (and indeed univariate imputation) assume normality for continuous variables conditional on the other variables in the imputation model. In practice data are often non-normal, which raises a question around how best to impute such variables.

One option that has been suggested in the literature for imputing non-normal, continuous data is to impute missing values using alternative distributional models, such as Beta and Weibull distributions [8], the generalised lambda distribution [9], the gh distribution [10], and Fleishman polynomials [11]. Difficulties with such approaches are that most of these alternative distributions have only been considered in the univariate context and importantly none are readily accessible in standard statistical packages.

An alternative approach for imputing a non-normal variable is to transform the data to an approximately normal distribution prior to imputation, and then impute the missing data on the transformed scale using one of the standard approaches [12]. A recent paper by Von Hippel [13] compared imputation of raw data and imputation following a de-skewing transformation when imputing missing data generated from an exponential or chi-squared distribution, in both the univariate and multivariate context. He concluded that it is usually safer to impute without transformation unless you are interested in estimating quantiles and shape parameters, but his findings are somewhat limited. Firstly, incomplete data were generated from an exponential or a chi-squared distribution with 1, 2, 4 and 8 degrees of freedom, all of which represent positively skewed distributions. This limits the generaliz-ability of the findings. Secondly, the bias and mean squared error from the two approaches were summarised across all of the distributions for a range of missing data mechanisms and strengths of relationships. Given the amalgamation of results it is difficult to disentangle how far wrong you can go when imputing a non-normal variable, or whether there are scenarios where it may indeed be more appropriate to transform the data prior to imputation. Finally, in

all scenarios imputation was performed after the same de-skewing transformation despite varying degrees of skewness. This leaves open the question of what is the most appropriate transformation to use in any given scenario. White et al [12] suggested that the best choice is to "find a model that is congenial and a good representation of the data", although this was somewhat vague and not backed up by any empirical research.

Another option for imputing non-normal data within the FCS algorithm is to use predictive mean matching (PMM), where the missing values are replaced by an observed value from a donor pool of *k* candidate donors based on the distance between the expected mean from the linear prediction model [14]. This approach was developed as a method for drawing imputations that relaxes some of the assumptions of parametric imputation.

The aim of the current manuscript was to extend the study by Von Hippel to compare various transformations applied to data from a range of non-normal distributions, in order to provide a more detailed exploration of when transformations may or may not be appropriate, and which transformation to use. Like Von Hippel, we consider missing data in a key exposure variable and explore the effects on the estimation of the marginal mean of the exposure and the relationship with a completely observed continuous outcome. However, we also extend our investigation to a binary outcome. As part of this study we consider a range of de-skewing transformations including a novel non-parametric transformation where we

transform the data based on quantiles of the normal distribution, along with the PMM approach. We explore these aims through simulation and a case study.

Simulation Methods

Data generation

We generated 2000 datasets of 1000 observations. For each set of simulated datasets, we generated a covariate *X* from one of a range of distributions:

- a normal distribution,
- positively and negatively skewed distributions:
  - a gh distribution (based on a simple transformation of normal deviates using 2 parameters to control the heaviness of the tails (g) and the elongation (h)) [10]
  - a gamma distribution
  - a log-normal distribution, and
- a bimodal distribution generated from a mixture of normal distributions.

See the Appendix for details of these distribution. We selected parameters for the distributions to give 11 different distributional shapes for *X*. For the gh and the mixture of normal distributions we included a shift to ensure that all simulated data were positive, to avoid having to deal with negative values when using power transformations. In all cases we standardised the scale of *X* relative to the interquartile range of the distribution to ensure a

similar scale for *X* across each of the distributions. Densities for the selected distributions are shown in Figure 1.

< Figure 1 about here >

Next we generated an outcome, *Y*, according to 4 different scenarios:

1. We generated a continuous outcome from a regression model linear in *X*:

$$Y_i = \alpha + \beta X_i + e_i \quad (1)$$

   We set α=1 and β=1 and generated e_i from a $\text{Normal}(0, 1)$ distribution.

2. We generated a continuous outcome from a regression model linear in log(*X*)

$$Y_i = \alpha + \beta \log(X_i) + e_i \quad (2)$$

   Again we set α=1 and β=1, and generated e_i from a $\text{Normal}(0, 1)$ distribution. Given the log transformation is only valid for positive values of *X*, and that this transformation is only sensible for positively skewed data, we restricted our exploration of this scenario to the cases where *X* was generated from a gamma or log-normal distribution.

3. We generated a binary outcome from a binomial distribution with probability of a positive outcome determined by a logistic regression model on *X:*

$$Y_i \sim \text{Bin}(1, \pi_i) \quad (3)$$

$$\text{logit}(\pi_i) = \alpha + \beta X_i$$

In each case we set α = -2 and β = 0.693 (corresponding to an odds ratio of 2). This resulted in a different proportion of observations with $Y = 1$ across the different distributions for X (ranging from 0.22 to 0.64).

4. Finally, we generated a binary outcome from a binomial distribution with probability of a positive outcome determined by a logistic regression model on log(X):

$$Y_i \sim \text{Bin}(1, \pi_i) \tag{4}$$

$$\text{logit}(\pi_i) = \alpha + \beta \log(X_i) + e_i$$

Again we set α = -2 and β = 0.693 (corresponding to an odds ratio of 2) corresponding to the proportion of observations with $Y = 1$ ranging from 0.10 to 0.25 across the different distributions for X. As with scenario 2, we restricted our exploration of this scenario to the cases in which X was generated from a gamma or log-normal distribution.

Finally, we set 50% of values for covariate X to missing. The missingness in X was either missing completely at random (MCAR), by setting a random 50% of the X values to missing, or missing at random (MAR), where the probability of missingness was determined by a logistic regression model dependent on a standardised version of Y, defined as $Y_s = (Y - \text{E}(Y))/\text{SD}(Y)$ where E(Y) and SD(Y) are the expected value and the standard deviation of Y respectively:

$$\text{logit}(\pi_{miss}) = \gamma + \delta Y_s \tag{5}$$

We set δ to 0.693 (corresponding to an odds ratio of 2, a reasonably strong, but potentially realistic association with missingness), and γ was set to -0.945 to ensure approximately 50% missingness for each scenario. It was important that *X* be set to missing dependent on the standardised value of *Y* (Y_s), as opposed to *Y* itself, as the scale of *Y* varied depending on the distribution of *X*. We focussed primarily on scenarios where *X* is MCAR in order to isolate issues due to the non-normality of the data from those arising due to the MAR mechanism. However we also considered an extension where *X* was MAR (scenario 1), to demonstrate that similar findings were obtained in this more realistic setting.

Analysis

For each scenario (each combination of the generation of *X* and *Y* described above and the two missingness mechanisms), analysis was carried out using MI to handle the missing data, through the `mi impute regress` command in Stata: Release 12 [4] (since missingness was in a single variable). Missing values of *X* were imputed either:

- on the raw scale;
- following a zero-skewness log transformation, using the `lnskew0` command in Stata;
- following a Box-Cox transformation, using the `bcskew0` command in Stata;
- following a non-parametric transformation where we equated quantiles of the variable to quantiles of a normal distribution based on deciles, percentiles or with the quantiles

defined by single observations, i.e. using the cumulative distribution function as is done in standard normal quantile-quantile plots.

- using PMM applied to the raw data, where a linear prediction model is used to obtain the expected mean of the variable to be imputed, and missing values are imputed from a pool of *k* "matching" donors defined as the observed cases whose predicted values are closest to the predicted value for the incomplete observation. We consider two versions of PMM, "type 1" and "type 2" matching. Type 1 matching uses the linear predictor based on best-fitting parameter values for the observed cases but uses a draw of the parameters for the linear prediction model for the incomplete observation, while type 2 matching, uses a draw of the parameters for the linear prediction model for both the observed and incomplete cases [15]. PMM with type 1 matching was implemented using the `ice` command in Stata (as this form of matching is not available within the mi suite of commands), and PMM with type 2 matching was implemented using `mi impute pmm`. In both cases we set *k*=10.

This resulted in 8 imputation approaches for each scenario. In each case the imputation model included just the outcome *Y*. For each approach, 50 imputed datasets were generated. Following imputation, where a transformation had been applied, imputed values were back transformed onto the raw scale for analysis. For the zero-skewness log transformation, values that were imputed below the shift on the transformed scale were replaced by the smallest observed value above the shift prior to back-transformation. Similarly, when the analysis model was the regression of *Y* on log(*X*) (scenarios 2 and 4) values that were imputed as less than 0 were replaced by the smallest observed value of *X* prior to analysis. Analysis following

MI was carried out using the `mi estimate command` in Stata, using the same model that was used to generate the data (equations 1-4).

Inference from each of the imputation approaches was compared using the bias, root mean squared error (RMSE) and the coverage of the 95% confidence interval (CI) of the estimated marginal mean of *X*, and the estimated regression coefficient of *Y* regressed on *X* or log(*X*), β, compared with the true value used to generate the data (i.e. mean of the distribution used to generate *X* obtained from a pseudo-population of 1,000,000 observations, and $\beta = 1$ for continuous *Y* and 0.693 for binary *Y*), averaged across the 2000 simulated datasets. Occasionally the analysis models did not converge when fitting a logistic regression model for *Y* on *X* or log(*X*) following imputation for a particular simulated dataset (<1% of simulated datasets). In these cases, the simulated dataset was removed from all summaries. It is not possible to compare inference across the 11 distributions for *X* or the 4 data generation scenarios or 2 missingness mechanisms, since the marginal mean and range of *X* and hence the distribution of *Y* varies for each example. However, inference can be compared across the 8 imputation methods for any given scenario. A well-performing approach is characterised by small bias and RMSE and approximately 95% coverage across all of the scenarios considered.

Results

Scenario 1 – *Y* continuous and linear in *X*, *X* MCAR

When imputation was carried out on the raw scale, there was little bias in the estimation of the mean of *X* irrespective of its underlying distribution. Bias was, however, introduced when a zero-skewness log or Box-Cox transformation was applied prior to imputation for some distributions of *X* (Figure 2a). Using a non-parametric transformation prior to imputation produced an unbiased estimate of the mean of *X* when the quantiles were defined at the observation level (i.e. small bins), but small levels of bias were introduced when wider bins were used. Similarly there were small biases in the results obtained using PMM. The RMSE was broadly similar for the 8 analysis methods across all distributions for *X*, although there was under-coverage when the non-parametric approach was applied using deciles or when imputation was carried out using type 2 PMM (Figure A1).

When estimating the association between *X* and *Y*, imputing on the raw scale or using PMM produced unbiased estimates irrespective of the distribution of *X*. Imputation following a zero-skewness log or Box-Cox transformation introduced bias for the majority of distributions for *X*, in particular when *X* was normally distributed or negatively skewed (Figure 2b). Applying a non-parametric transformation prior to imputation also introduced bias for the majority of distributions for *X*, with the smallest bias when wider bins were used (e.g. deciles). The poor behaviour of the normalising transformations was also apparent in the RMSE and the coverage (Figure A2).

Scenario 2 - *Y* continuous and linear in log(*X*), *X* MCAR

When estimating the marginal mean of *X* for scenario 2, imputation on the raw scale, following a zero-skewness log or Box-Cox transformation or using PMM resulted in unbiased estimates across all distributions for *X* (Figure 3a). As with scenario 1, imputation following a non-parametric transformation produced an unbiased estimate of the marginal mean of *X* when carried out using narrow bins, but introduced bias when wider bins were used. All analyses resulted in similar RMSE and coverage with the exception of the non-parametric transformation applied using deciles and using type 2 PMM (Figure A3). When *Y* was linearly associated with log(*X*), imputation on the raw scale introduced some bias in the estimation of the association between *Y* and log(*X*) irrespective of the distribution of *X*, which was reduced by transforming *X* prior to imputation. The smallest bias was seen following a non-parametric transformation, with similar results irrespective of the underlying distribution of *X* and the width of the bins used in the non-parametric transformation, and using PMM. A similar pattern of results was seen in the RMSE and coverage (Figure A4).

Scenarios 3 and 4 - *Y* binary and logistic-linear in *X* / log(*X*), *X* MCAR

The pattern of results was broadly similar when *Y* was a binary outcome. There was slight bias in the estimation of the mean of *X* from all most imputation approaches for both scenarios, with the exception of when imputation was carried out using PMM. The largest bias was observed when the non-parametric transformation was applied using deciles (Figure A5a and A7a). When *Y* was linearly dependent on *X* there was more consistent bias in the

estimation of association across all of the imputation approaches compared with scenario 1 (Figure A6a). However when Y was dependent on log(*X*), imputation on the raw scale or using type 2 PMM resulted in much larger bias in the estimation of the association than when imputation was carried out after a normalising transformation or using type 1 PMM (Figure A8a). A similar pattern of results was seen in the RMSE and the coverage (Figures A5-A8).

Scenario 1 - *Y* continuous and linear in *X*, *X* MAR

Finally, when *X* was MAR and we generated a continuous outcome linear in *X*, there was some bias in estimating the mean of *X* using all of the approaches, with a much larger bias when imputation was carried out following a non-parametric transformation with wide bins (Figure 4). Although the RMSEs were reasonably similar across the approaches, there was some under-coverage in the estimation of the mean when imputation was carried out on the raw scale, following a non-parametric transformation with wide bins or using PMM with type 2 matching (Figure A9).

There was also slight bias from all imputation approaches when estimating the association between *Y* and *X*, although somewhat larger bias when imputation was carried out following a zero-skewness log or a Box-Cox transformation. The smallest bias was observed when imputation was carried out using PMM. A similar pattern of results was seen in the RMSE

and the coverage, although there was also some under-coverage in the estimation of the association when imputation was carried out on the raw scale (Figures A9 and A10).

Case Study

The 8 imputation approaches were applied in an analysis comparing the occurrence of parental anxiety and depression at term-equivalent age in mothers of very preterm and term-born infants [16]. Families with very preterm (<30 weeks' gestational age, n=113) and full terms infants (≥37 weeks' gestational age with a birth weight >2499g, n=112) were recruited between January 2011 and March 2014 at the Royal Women's Hospital (RWH), Melbourne, Australia. Depression and anxiety symptoms were assessed in parents at regular intervals following the birth of their child using the Centre for Epidemiological Studies Depression Scale (CES-D)[13] and the anxiety scale from the Hospital Anxiety and Depression Scale (HADS)[16] respectively. The levels of anxiety and depression were compared between mothers of very preterm and term-born infants at term-equivalent age using linear regression adjusted for plurality (singleton vs multiple), other siblings (no versus yes) and social risk as potential confounders. Social risk was measured using a composite score based on six components including family structure, education of primary caregiver, occupation of primary income earner, employment status of primary income earner, language spoken at home, and maternal age at birth. This variable was positively skewed and was missing for 13 infants (6%). For the sake of this analysis, social risk was imputed using the 8 strategies

described above using the variables in the analysis model (gestational age group, depression and anxiety) as well as sex, gestational age and birthweight of the child.

Table 1 shows the estimates for the regression coefficient for the group difference (very preterm versus term) and for social risk from these 8 analyses. There are some slight differences in the estimates from the various approaches but the overall conclusions remain the same.

Discussion

This paper compared model-based imputation with various prior transformations and PMM when imputing non-normal data. The results from our simulations suggest that across a variety of distributions for the non-normal variable with missing data, imputation on the raw scale or using PMM provides a reasonable estimate of the marginal mean of this variable, although there was under-coverage in the estimation of the mean when imputation was carried out using type 2 PMM. When estimating the marginal mean, applying a normalising transformation prior to imputation generally introduced bias. The exception was when a non-parametric transformation with narrow bins was used to transform the variable towards normality prior to imputation. In contrast, our case study demonstrated little difference between the approaches in this practical example, presumably due mainly to the small amount of missing data.

When estimating the association between a completely observed outcome and the incomplete variable, the performance of the various imputation approaches depended on the nature of the relationship between the incomplete covariate and the outcome. Imputation on the raw scale resulted in an unbiased estimate of association when the outcome was linearly related to the incomplete covariate. However, when the relationship between the outcome and the incomplete covariate was log-linear, imputing the covariate on the raw scale introduced bias. In contrast, transforming the incomplete variable prior to imputation introduced bias when the outcome was linearly related to the incomplete covariate, but less so when the relationship was log-linear. Imputation using PMM with type 1 matching resulted in minimal bias across all scenarios, although using type 2 matching resulted in bias when the outcome was binary (scenarios 3 and 4). In practice the true relationship between the outcome and covariate is unknown, but one would hope that the analysis model would reflect the true relationship. Potential non-linear relationships can be explored by either categorising the covariate into e.g. deciles and plotting the average outcome within each group, or using methods such as fractional polynomials [17]. Once a well-fitting analysis model has been selected, our findings suggest that the decision regarding whether to transform an incomplete variable prior to imputation should depend on the nature of the relationship between the incomplete variable and the outcome in the analysis model, rather than the distribution of the incomplete variable. Alternatively, PMM with type 1 matching can be used to impute the non-normal variable. These novel empirical findings are consistent with the recommendation by White et al [12]. Importantly our findings indicate that blindly using a transformation to

improve the normality of an incomplete covariate can introduce bias. This adds to the growing body of literature that MI can in fact introduce bias if not carried out appropriately [18]. In extreme cases inappropriate specifications in the imputation model could result in the identification of spurious relationships in the analysis.

The fact that imputing on the raw scale performs well when estimating the linear relationship between a fully observed outcome and an incomplete covariate is consistent with findings from Von Hippel [13]. However, in contrast to Von Hippel, we have demonstrated that there are scenarios where it may be important to transform an incomplete variable prior to imputation. This latter finding is in line with recent literature on the importance of compatibility between the imputation and analysis models [19]. Two conditional models (in our case the imputation and analysis models) are said to be compatible if there exists a joint model for which the conditionals of the joint model are the same as the two conditional models of interest [19]. Briefly, Bartlett et al [19] demonstrated that imputing incomplete covariates using an imputation model that is compatible with the analysis model gave unbiased estimates for a range of substantive models, which was not the case when an incompatible imputation model was used.

In terms of comparing the performance of the various transformations, the non-parametric transformation presented in this paper generally resulted in less bias and smaller RMSE compared with the zero-skewness log and Box-Cox transformations. The proposed non-

parametric transformation is a very flexible approach that can be used to handle different forms of non-normality. In contrast, both the Box-Cox and zero-skewness log transformation can only handle skewed data, with the Box-Cox transformation really designed for positively skewed data that are greater than 0 (although this transformation could in principle be applied with a reflection and a shift). The difficulty with applying the non-parametric transformation is deciding how many bins to use, with our results suggesting that narrow bins are preferable when estimating a marginal mean but wider bins are more appropriate when estimating an association. The reasons for this are unclear, and additional research is needed to explore this phenomenon further. A similar non-parametric approach based on normal quantiles was suggested by Hussain et al [20], although they used a "functional relationship" between the observed data and the corresponding normal quantiles. It was difficult to determine what the authors meant by a "functional relationship", nor why it was necessary, so we did not include an evaluation of this approach in our work.

In this study, PMM with type 1 matching was found to perform better than PMM with type 2 matching. This is in line with previous studies by Morris et al. [15] and van Buuren [21] who have advised against the use of type 2 matching since if the number of covariates in the prediction model is small the uncertainty in the missing data may not be adequately represented. More generally, Morris et al comment that PMM can lead to bias if there are few donors in the vicinity of an incomplete case e.g. if there are large proportions of missing data, under MAR and in the tails of distributions [15]. A disadvantage with type 1 matching is that is it not currently available within the standard suite of commands in Stata. Another challenge

with PMM in general is that it is not clear how many donors to use within the algorithm. This would be an interesting area for further work.

The strength of the current study is that we have compared a number of different approaches for imputing non-normal data across a range of scenarios. Although it is difficult to draw definitive conclusions from a specific set of simulation studies, we have discovered that the decision regarding the imputation of non-normal variables should be made on the basis of the intended analysis model rather than the distribution of the incomplete variable. What is not clear is what part (if any) the other variables in the imputation and analysis models should play in decisions about the use of transformations. In this paper, we restricted our attention to missing data in a single variable. When analysing real data, there are likely to be multiple variables with missing values and complex inter-relationships between variables. In this context it is likely to be even more important that relationships between the variables are modelled accurately in order to obtain valid inference. The simulations presented here are fairly extreme in that 50% of observations had missing data. In practice there is often less missing data and hence the decision regarding whether or not to transform a variable prior to imputation may be less important, as demonstrated in our case study. The majority of this paper focussed on the scenario where data are MCAR, although we illustrated a similar pattern of results in our MAR example. There is no reason to expect a different pattern of results when data are MAR although further investigation would be useful. Finally, throughout this paper we apply the correct analysis model for Y i.e. the model used to generate the simulated data. In practice the true underlying model is not known and the

analyst chooses a model to apply to the data. This raises a further question around how these approaches would fare if the analysis was carried out using a poorly fitting model.

Conclusions

In summary, in this study we have provided novel empirical evidence that decisions about how to impute a non-normal variable should depend on the parameter of interest. In estimating the relationship between the incomplete covariate and a fully observed outcome, the decision regarding transformation should be made based on the nature of the relationship between the variables rather than the distribution of the incomplete covariate. Importantly, the imputation model should be compatible with the analysis model. Automatically transforming a variable to improve normality prior to imputation can introduce bias when estimating associations. A useful alternative is to use PMM using type 1 matching.

Acknowledgment of support

This work was supported by funding from the National Health and Medical Research Council: Career Development Fellowship ID#1053609 (KJL) and Centre of Research Excellence grant, ID#1035261, for the Victorian Centre for Biostatistics (ViCBiostat). Research at the Murdoch Childrens Research Institute is supported by the Victorian Government's Operational Infrastructure Support Program.

References

1. Klebanoff MA, Cole SR. Use of Multiple Imputation in the Epidemiologic Literature. *American Journal of Epidemiology* 2008; **168**: 355-357.
2. Rubin DB. *Multiple imputation for nonresponse in surveys.* Wiley: New York, 1987.
3. SAS Institute inc. PROC MI. In PROC MI. SAS Institute Inc: Cary, NC, 2008.
4. StataCorp. *Stata Statistical Software: Release 12.* StataCorp LP: College Station, TX, 2011.
5. Schafer JL. *Analysis of Incomplete Multivariate Data.* Chapman & Hall: London, 1997.
6. Raghunathan TE, Siscovick DS. A multiple imputation analysis of a case–control study of the risk of primary cardiac arrest among pharmacologically treated hypertensives. *Applied Statistics* 1996; **45**: 335-352.
7. VanBuuren S, Boshuizen HC, Knook DL. Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine* 1999; **18**: 681-694.
8. Demirtas H, Hedeker D. Imputing continuous data under some non-Gaussian distributions. *Statistica Neerlandica* 2008; **62**: 193-205.
9. Demirtas H. Multiple imputation under the generalised lambda distribution. *Journal of Biopharmaceutical Statistics* 2009; **19**: 77-89.
10. He Y, Raghunathan TE. Tukey's gh distribution for multiple imputation. *The American Statistician* 2006; **60**: 251-256.
11. Demirtas H, Hedeker D. Multiple Imputation under power polynomials. *Communication in Statistics - Simulation and Computation* 2008; **37**: 1682-1695.

12. White IR, Royston P, Wood AM. Multiple imputation using chained equations: issues and guidance for practice. *Statistics in Medicine* 2011; **30**: 377-399.
13. von Hippel PT. Should a normal imputation model be modified to impute skewed variables? *Sociological Methods and Research* 2013; **42**: 105-138.
14. Little RJA. Missing-data adjustments in large surveys. *J Business & Econ Stat* 1988; **6**: 287-296.
15. Morris TP, White IR, Royston P. Tuning multiple imputation by predictive mean matching and local residual draws. *BMC Medical Research Methodology* 2014; **14**: 75.
16. Pace CC, Spittle AJ, Molesworth CM, Lee KJ, Northam EA, Cheong JLY, Davis PG, Doyle LW, Treyvaud K, Anderson PJ. Evolution of depression and anxiety symptoms in mothers and fathers of very preterm infants during the newborn period – a longitudinal study. *JAMA pediatrics* 2016; **170** 863-870.
17. Royston P, Altman DG. Regression using fractional polynomials of continuous covariates: Parsimonious parametric modelling (with discussion). *Applied Statistics* 1994; **43**: 429–467.
18. Lee KJ, Carlin JB. Recovery of information from multiple imputation: a simulation study. *Emerg Themes Epidemiol* 2012; **9**: 3.
19. Bartlett JW, Seaman SR, White IR, Carpenter JR, for the Alzheimer's Disease Neuroimaging Initiative. Multiple imputation of covariates by fully conditional specification: accomodating the substantive model. *Statistical Methods in Medical Research* 2015; **24**: 462-487.

20. Hussain S, Mohammed MA, Haque MS, Holder R, Macleod J, Hobbs R. A simple method to ensure plausible multiple imputation for continuous multivariate data. *Communication in Statistics - Simulation and Computation* 2010; **39**: 1779-1784.

21. Van Buuren S. *Flexible Imputation of Missing Data.* CRC Press: Hoboken, 2012.

Table 1: Estimates of the regression estimates from the case study

	Regression coefficient for preterm vs term group		**Regression coefficient for social risk**	
	Estimate (95% CI)	p-value	Estimate (95% CI)	p-value
DEPRESSION				
Raw	8.14 (4.82, 11.47)	<0.001	-0.24 (-1.44, 0.97)	0.70
Zero-skewness log	8.10 (4.78, 11.42)	<0.001	-0.18 (-1.38, 1.02)	0.76
Box-Cox	8.10 (4.78, 11.42)	<0.001	-0.18 (-1.38, 1.02)	0.76
NP deciles	8.13 (4.82, 11.44)	<0.001	-0.23 (-1.45, 0.98)	0.71
NP percentiles	8.15 (4.83, 11.47)	<0.001	-0.25 (-1.44, 0.95)	0.68
NP per obs	8.10 (4.78, 11.43)	<0.001	-0.19 (-1.41, 1.03)	0.75
PMM Type 1	8.07 (4.75, 11.39)	0.000	-0.19 (-1.39, 1.01)	0.76
PMM Type 2	8.11 (4.80, 11.42)	0.000	-0.21 (-1.41, 1.00)	0.74
ANXIETY				
Raw	2.04 (0.69, 3.39)	0.003	0.09 (-0.37, 0.55)	0.69
Zero-skewness log	2.04 (0.70, 3.39)	0.003	0.10 (-0.36, 0.55)	0.68
Box-Cox	2.05 (0.69, 3.40)	0.003	0.09 (-0.37, 0.55)	0.70
NP deciles	2.05 (0.70, 3.39)	0.003	0.09 (-0.37, 0.55)	0.70
NP percentiles	2.05 (0.69, 3.40)	0.003	0.09 (-0.37, 0.55)	0.70
NP per obs	2.05 (0.70, 3.40)	0.003	0.08 (-0.38, 0.55)	0.72
PMM Type 1	2.05 (0.70, 3.40)	0.003	0.09 (-0.37, 0.55)	0.71
PMM Type 2	2.04 (0.70, 3.39)	0.003	0.10 (-0.36, 0.55)	0.68

Results are from linear regression models for depression and anxiety score fitted using generalised estimating equations with an exchangeable correlation matrix to allow for the clustering of multiple births. CI = confidence interval; NP = non-parametric transformation; PMM = predictive mean matching.

Figure 1: Range of distributions for covariate X. a) gh distributions, b) gamma distributions, c) mixture of normal distributions, and d) log-normal distributions. Graphs are kernel densities based on 1,000,000 simulated observations (see online Appendix for definitions of the distributions).

* Transformation of normal deviates using 2 parameters (g,h) controlling the heaviness of the tails (g) and the elongation (h).

† mix(σ_1^2, σ_2^2) from a mixture of a Normal(2, σ_1^2) + Normal(6, σ_2^2) distribution.

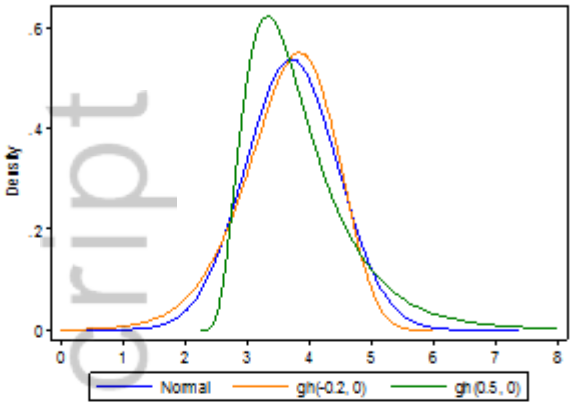
Figure 2: Bias in a) the estimated mean of X and b) the estimated regression coefficient for the linear regression of Y on X, compared with the true values in the simulation model, where Y was a continuous outcome generated from a linear regression model conditional on X (see Methods: Scenario 1), and X was missing completely at random, for various underlying distributions of X. Monte Carlo error <0.001 for the mean of X, and <0.002 for the regression coefficient across all of the distributions for X. NP = non-parametric transformation; PMM = predictive mean matching.

Figure 3: Bias in a) the estimated mean of X and b) the estimated regression coefficient for the linear regression of Y on log(X), compared with the true values in the simulation model, where Y was a continuous outcome generated from a linear regression model conditional on log(X) (see Methods: Scenario 2), and X was missing completely at random, for various underlying distributions of X. Monte Carlo error <0.002 for both the mean of X and the regression coefficient across all of the distributions for X. NP = non-parametric transformation; PMM = predictive mean matching.

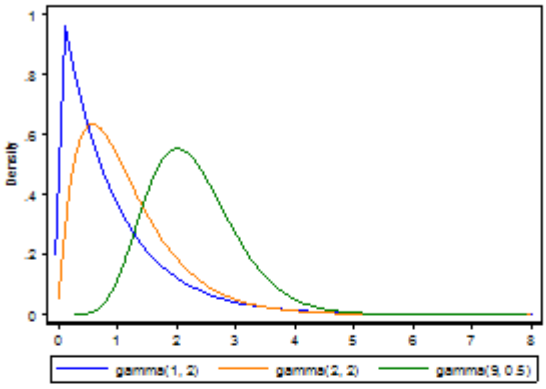
Figure 4: Bias in a) the estimated mean of X and b) the estimated regression coefficient for the linear regression of Y on X, compared with the true values in the simulation model, where Y was a continuous outcome generated from a linear regression model conditional on X (see Methods: Scenario 1), and X was missing at random, for various underlying distributions of X. Monte Carlo error <0.001 for the mean of X and <0.002 for the regression coefficient across all of the distributions for X. NP = non-parametric transformation; PMM = predictive mean matching.

Lee – Figure 1

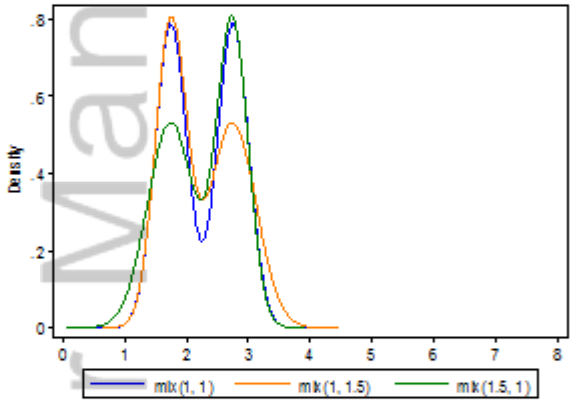
a)



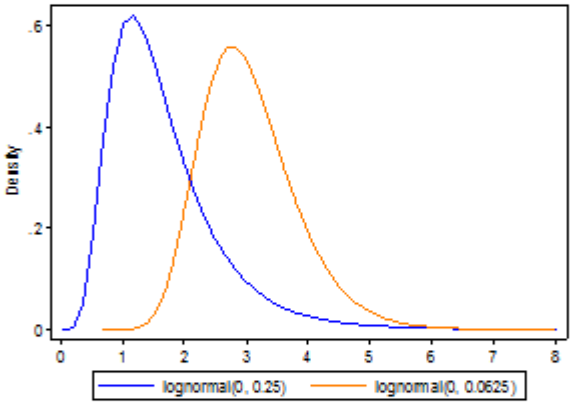
b)



c)

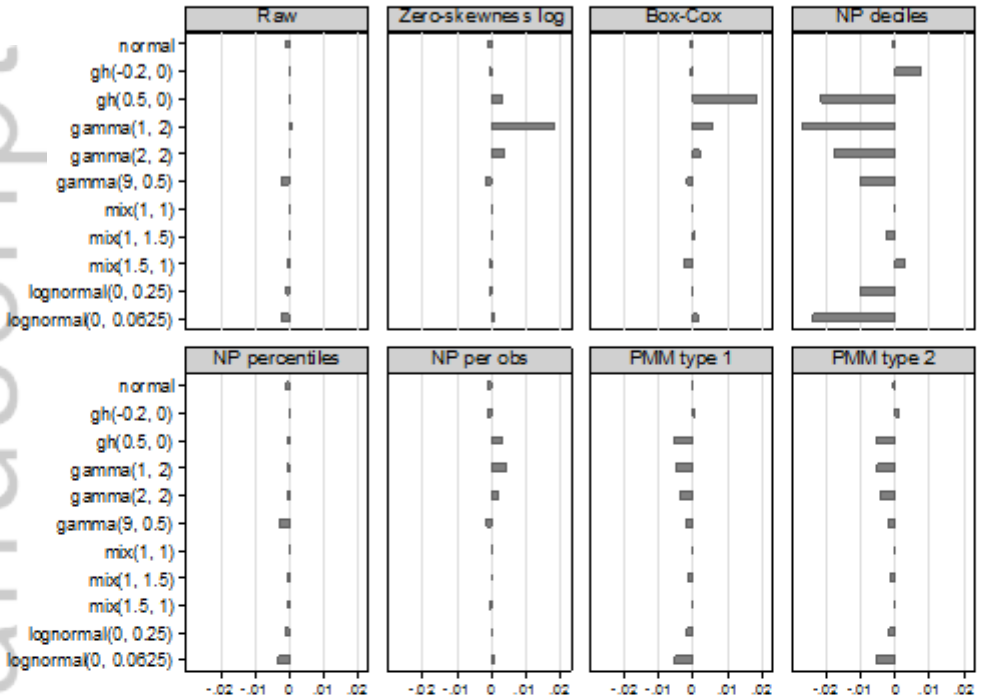


d)

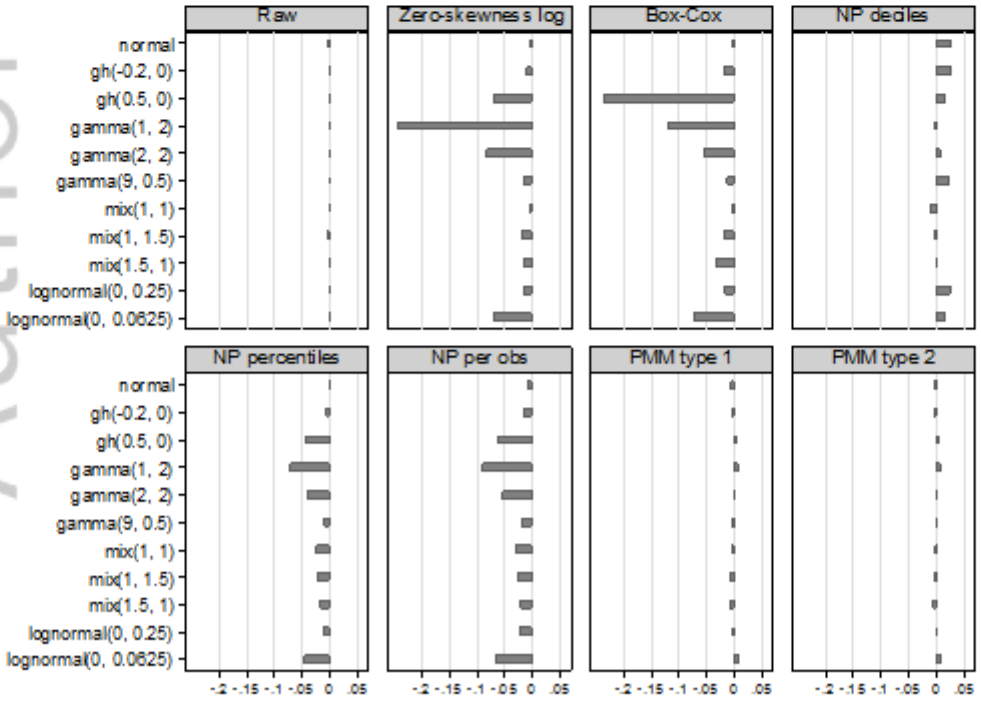


Lee – Figure 2

a)

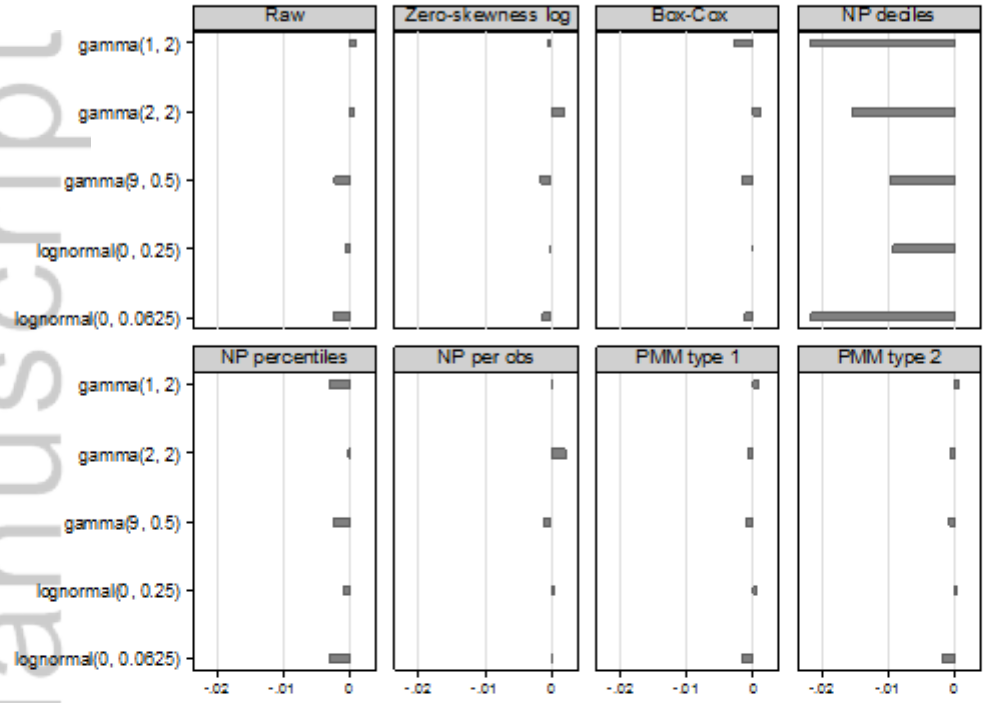


b)

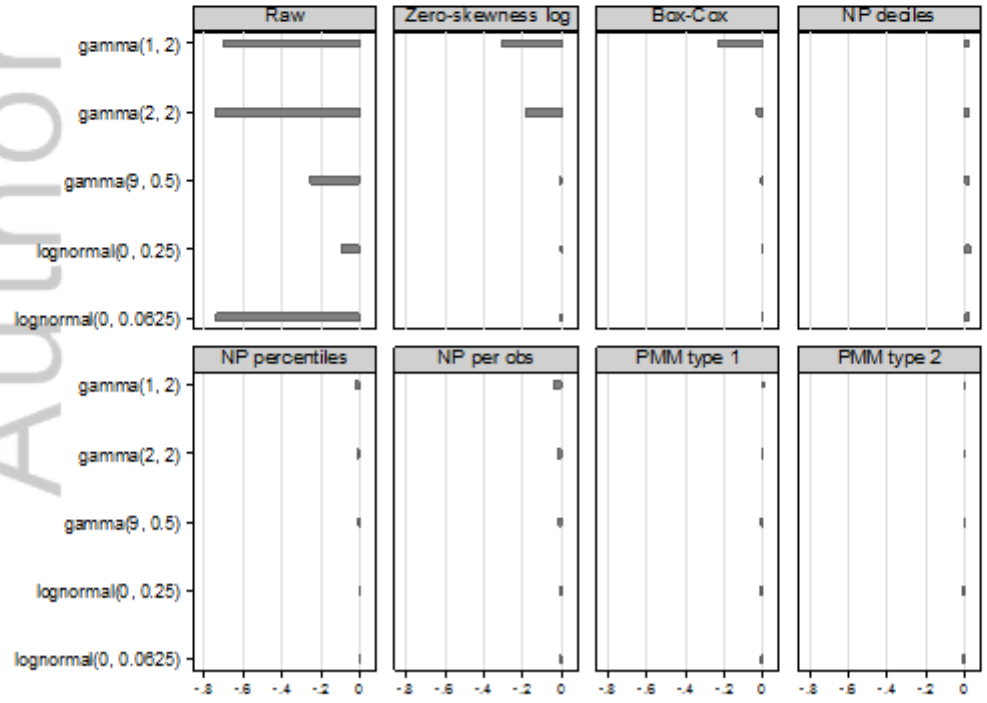


Lee – Figure 3

a)

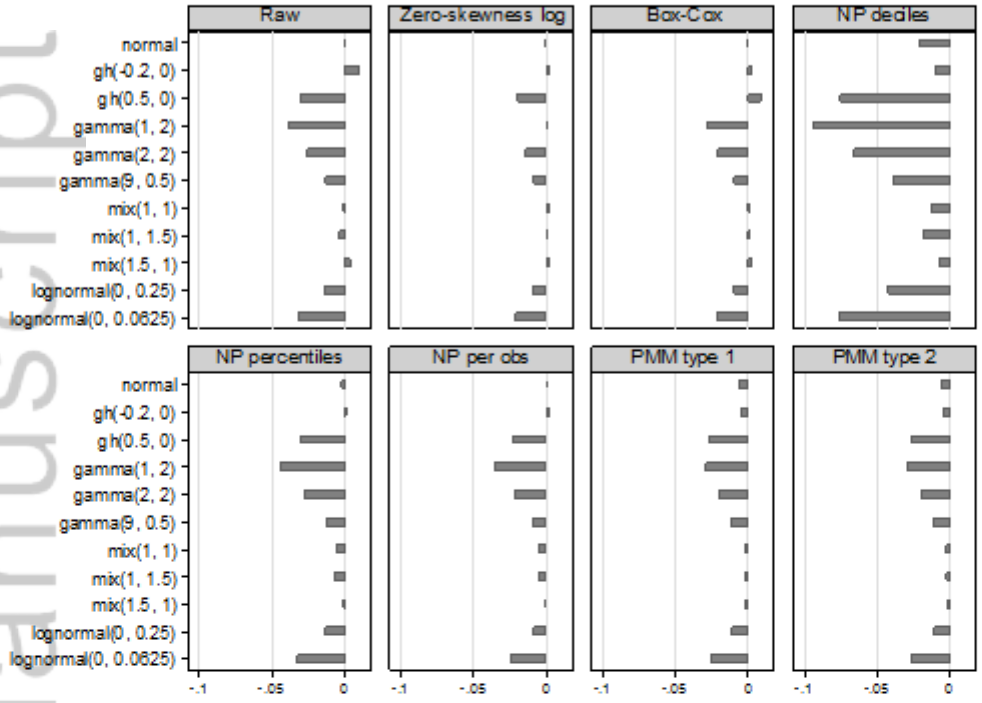


b)

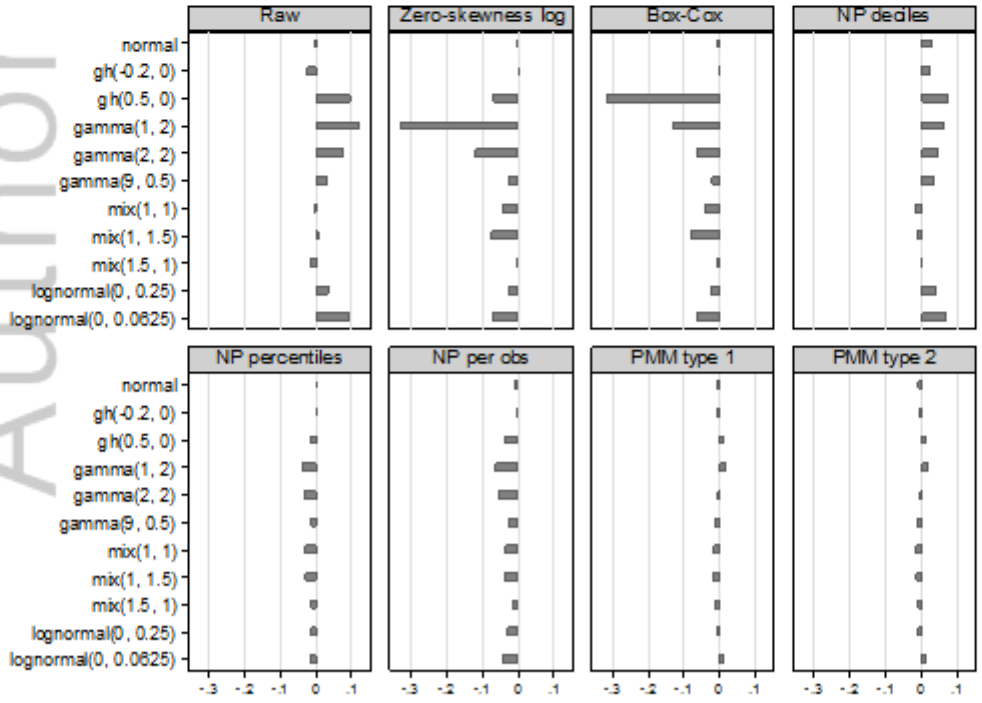


Lee – Figure 4

a)



b)



SIM_7173_F4.tif