Using Proportional Hazard Regression for Building Imputation Models

Christèle Augard
Robert D. Small

1. Introduction

A critical step in carrying out a multiple imputation analysis [2] in the presence of missing data is the imputation model. Classically this is given some thought but it is often assumed that it is a simple step. A description or even an implementation of the method often concentrates on the other steps. Those include sampling from the imputed distributions, combining the multiply imputed groups and getting the resultant estimates and interpreting the results and how they compare to other more naïve methods. Here we focus on only the imputation step and only one method of imputation. The method is distribution free and robust against and number of deviations. It is also very general and does not make assumption about joint distributions or rely on ad hoc and arbitrary techniques when the data do not satisfy simple assumptions such as multivariate normality.

We describe the technique which is an application of proportional hazard regression [1], apply it to a representative data set, and compare it to an alternative approach that makes common but questionable assumptions about the distributions of the data. The method performs well on the data set we exhibit.

2. Background and Need for Alternative Methods

Usually when we have a data set with some missing data we have to pass through a sequence of steps. First we use the relations ships between the data to build a model that we can use to predict the missing data. This usually is a relationship between the observed data and the missing data and is used to predict the parameters of a distribution of the missing data. The imputed distributions are then sampled to give a set of imputed data. These imputed values are used to complete several copies of completed data sets. Each completed data set is analyzed in a standard way and then the multiple estimates are combined. We will focus on the imputation step.

It is common to assume some standard relationship between the observed data and the missing data. For example it is common to assume a multivariate normal relationship [4]. The usual relationships between the conditional means and the variance covariance matrix are then used to predict the means and variances of the missing data. The marginal normal distributions are then samples to complete the data set multiple times. When the missing data are not normal we resort to various common approaches. Transformations, logistic or Poisson regression, and sometimes truncation or dichotomization are used. When the independent, variables are not normal transformations or robustness are used to justify techniques. These methods have proven useful and sometimes robust [5].

Here we propose a method for use when the dependent variable is missing and the independent variables in the imputation model can be of any type. We fit the independent variables to the dependent variable using Proportional Hazard Regression. This method depends only on the ranks of the dependent variable and not its distribution. Thus any type of distribution can be used. Further the model once fit can be used to forecast the dependent variable with any type of independent variables. Later we will give an example of the procedure on some data and show that in that case the method does not lose power versus the utopian situation.

3. Description of the Method

Assume we have data with y the dependent variable to be analyzed. Let X be a matrix of independent covariates to be fitted to produce an imputation model. Assume that we use only the complete data. Fit the X to the y using the usual Proportional Hazard Regression method (ref). Some computer programs (e.g. SAS PHREG) which are designed to analyze time to event data will not accept negative values. This is easily solved by adding any constant making the smallest value positive. Since the PH fit depends only on the ranks of the dependent variable the coefficients will be the same. We can also replace the dependent variable with its ranks.

The resulting fit will give a basis survival function S_0 and that will give an estimated CDF of the dependent variable as a function of the independent variables as

$$F(y) = 1 - [S_0(y)]^{\exp(-X\beta)}$$

Thus for any row of X we have an imputed value. To carry out the imputation we need to take a sample from $\hat{F}(y\colon X\beta)$. This is easily done by generating a uniform random number between 0 and 1. The result is inversely mapped through $\hat{F}(y\colon X\beta)$. Often some interpolation will be needed to get the correct value of y. Nothing we have done makes any assumption on the distribution of the y data. The resulting regression equation imputes the ranks of the missing data. Any type of data can be used in the independent set X. The stochastic part of the imputation is implemented by generating uniform random numbers on [0 1].

In the next section we will use the method on a vaccine data set in several different ways and compare it to more common method of handling deviations from normality or other assumptions.

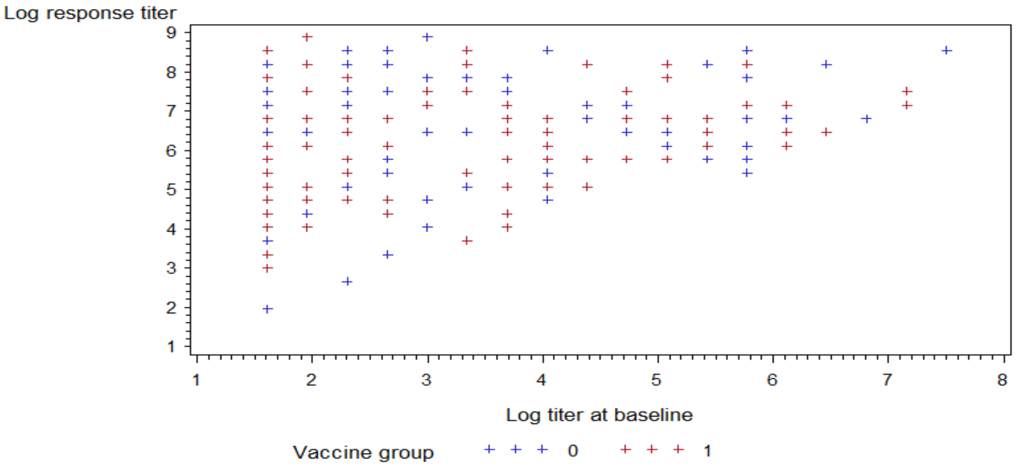
4. Example and Comparison to Standard Methods

We are going to give an example of the method on a vaccine trial comparing the response titers to vaccination 30 days after vaccination. The trial was carried out in Belgium, France and Germany. Subjects were randomized to one of the three groups – two controls and an investigational group. The subjects were between 16 and 60 years old. The subset of data chosen for this analysis was complete and we randomly removed 10% of the 30 day titers for the missing data.

We analyzed the data three ways. We used our proposed method. We also carried out a standard method of Multiple Imputation assuming that the joint distribution of the dependent variable and the covariates was multivariate normal. One of the covariates was sex and we used separate imputation models for each gender. Finally we carried out an analysis on all of the data without removing the random sample that we declared missing. Since we had the data we did this for a comparison.

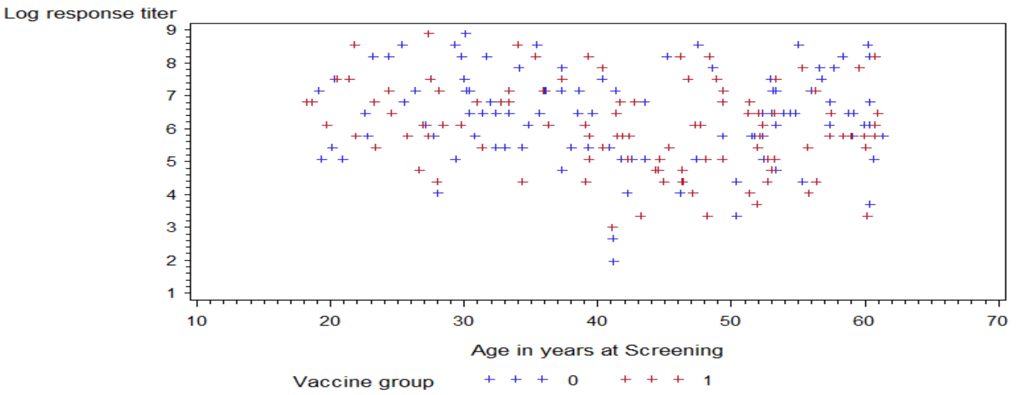
The covariates used in building the imputation model were sex, age and baseline titer taken immediately before vaccination. The same covariates were used in the analysis model along with the treatment group.

This is a plot of the response titer versus baseline titer with the treatment group identified

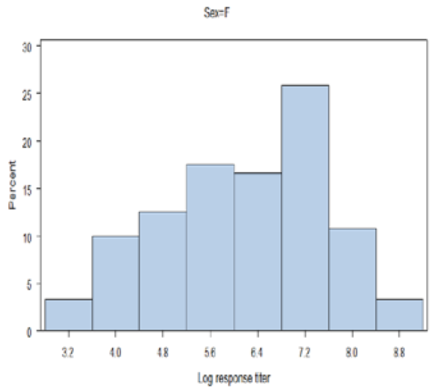


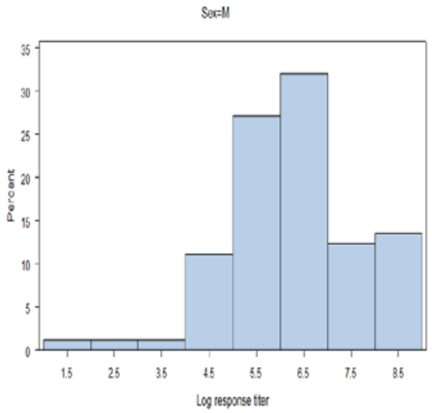
Though noisy it has some predictive value in both groups. The relationship does not seem to be different in the two groups.

The following is a graph of the response titer versus age. It is a little surprising that the relationship is not stronger. Often there is a significant reduction in response with age but it is very slight here.



Finally the following give histogram of the response titers by sex. They are different and though the female group might pass as normal the male one does not seem to.





We also did a second round of fitting where we replace the dependent variable with its rankit score. This should make those values more nearly normal.

The following table gives the result of the analyses after imputation. We generated 10 imputations for each missing value. All of the programs were run in SAS 9.2. Proc MI and Proc MIANALYZE [3] were used for the multivariate regression. Proc PHREG was used to implement the proposed method.

Results of analyses using PH regression and Normal Methods									
		PHREG				Multivariate Regression			
		EST.	SE	CI	p	EST.	SE	CI	p
Raw Data	TRT	-232.7	198.8	(-625.0, 159.6)	.24	-256.1	190.0	(-631, 119)	.18
	AGE	-12.0	8.4	(-28.6, 4.58)	.15	-13.5	8.14	(-29.6, 2.56)	.10
	SEX	51.6	203.5	(-349.8, 453.0)	.80	91.0	199.1	(-301 ,484)	.65
	BASE	.87	.42	(.04, 1.69)	.04	.91	.41	(.10, 1.72)	.03
Normal Trans.	TRT	-.20	.18	(-.57, .16)	.27	-.23	.18	(-.58, .13)	.21
	AGE	-.01	.01	(-.03, .001)	.07	-.14	.09	(-.33, .04)	.13
	SEX	-.002	.19	(-.38, .38)	.99	.07	.19	(-.29, .44)	.70
	BASE	.24	.06	(.13, .35)	<..001	.27	.06	(.15, .38)	<..001

The above table indicates that the results are very similar. The multivariate regression assumed multivariate normality and stratified on sex. The PHREG simply used the values as they were in a regression equation yet the results, both for the strength of the covariates and the treatment effects were similar. The normal transformation made the dependent variable as normal as possible and again the results were similar.

The following table shows the estimates of the treatment effect and the length of the 95% confidence intervals. Again it shows that the two approaches give similar results. The result without the missing data is also given to show by comparison how much information is lost by the missing data.

The lengths of the confidence intervals for the treatment effect are given in the table. The lengths are a good measure of the estimate of uncertainty resulting from each method.

	Raw Data			Normalized Transform		
	EST.	95% CI	CI Length	EST.	95% CI	CI Length
PHREG	-232.7	(-625.0, 159.6)	784.6	-.20	(-.57, .16)	.73
MULT REG	-256.1	(-631, 119.)	750.0	-.23	(-.58, .13)	.71
PP	-297.1	(-684.1, 89.9)	774.0	-.25	(-.61, .11)	.72

Here again we see that there is very little difference in the two approaches. They both tend to lengthen the confidence interval as one would expect due to the loss of data and the do so by the same amount.

5. Summary and Conclusions

We have introduced a general method for building imputation models that can be used for any Multiple Imputation situation. It can be used in any situation with missing dependent data. It is independent of the distribution of the dependent data and does not depend on the type (continuous, categorical, ordered, etc.) of the covariates. The method is based on Proportional Hazard Regression which has proven to be very powerful and robust in the analysis of time to event data when the distribution of the dependent variable is seldom known. It is easy to implement with widely available software.

We have demonstrate its use on a common data set and compared it to other commonly use methods. It did as well as other methods, including one that had to stratify on a binomial covariate (sex). This method simply included that binary covariate in the regression equation.

The method relies on the proportion hazard assumption. This seems to be a robust assumption based on tremendous experience with time to event data. Even when violated the assumption can be mitigated by stratification or transformation.

The method and resulting model also provide a possible approach to estimating the sensitivity due to non-random missingness. The form of the estimated CDF is the same as a Lehman Alternative and by varying the size of coefficients of the covariates it would be easy to investigate a wide range of alternative distributions resulting from non-random missingness. The aspect was not investigated yet.

6. References.

1. Kalbfleisch, J. D. & Prentice, R. L. (1980). The statistical analysis of failure time data. John Wiley & Sons. New York.
2. Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons. New York.
3. SAS Institute. (2013). *Statistical Analysis System*. Available at http:// www.sas.com.
4. Schafer, J. L. (1997) Analysis of Incomplete Multivariate Data. Chapman & Hall. London.
5. Van Buuren, Stef (2012) *Flexible Imputation of Missing Data*. CRC Press. Boca Raton, FL.