Advances in Methodology, Data Analysis, and Statistics
Anuška Ferligoj (Editor), Metodološki zvezki, 14
Ljubljana: FDV, 1998

Missing Binary Covariate Data and Imputation in Regression Models

Georg Heinze[1]

Abstract

This paper presents a simple way to handle missing values in categorical covariates, namely conditional probability imputation. Properties of this technique are given for various patterns of missing data in regression studies. An example shows its use in the proportional hazards model. The probability imputation technique is furthermore compared with multiple imputation and model-based approaches. It can be concluded that for certain patterns of missing data occuring typically in prognostic factor studies, the probability imputation technique has properties not inferior to more sophisticated but also more difficult-to-implement methods, and is outperforming standard techniques like complete case analysis or omission of covariates with missing values.

1 Introduction

1.1 A practical example

Suppose we are confronted with the following survival time data set (taken and modified from Andrews and Herzberg, 1985): We have 483 observations, each consisting of survival time and a censoring indicator as well as eight dichotomous covariates. Three of the covariates, HX, SG, and WT, have about 1/3 of their values missing independently from each other, leading to an effective sample size of 128, which is not much more than one fourth of the original sample.

The clinical partners expect all eight covariates to be strong prognostic factors, independently from each other, and we are forced to analyse all of them with respect to their effect on survival, e. g. using the proportional hazards model as described by Cox (1972). If we use a standard software package for statistical analysis and apply a proportional hazards regression analysis on the data, we yield the univariate and adjusted regression coefficients and significance levels presented in Table 1.

[1] Department of Medical Computer Sciences, Vienna University, Spitalgasse 23, A-1090 Vienna, Austria

Table 1: Regression coefficients (log hazard ratios) of a proportional hazards model fitted on 128 complete observations. Marginal coefficients refer to unadjusted models, partial coefficients to the model involving all covariates. * denotes a p-value ≤ 0.05, ** p ≤ 0.01, *** p ≤ 0.001, **** p ≤ 0.0001.

Factors	**marginal**	**partial**
TM	-0.20	-0.29
AG	0.56*	0.32
PF	0.58	0.39
HG	0.29	0.16
SZ	0.87**	0.64*
HX	0.59**	0.47*
SG	0.47*	0.28
WT	0.11	0.15

We see that in spite of all factors supposed to be relevant, only two of them turn out to be significant on the 5 %-level in the adjusted model, and none is significant on the 1 %-level. Do the clinicians fail with their expectations or are we somehow loosing power?

The standard software package uses information only from the 128 completely observed cases. So the important thing is to exploit the information contained in the other 75 % of the data set.

1.2 The Probability Imputation Technique (PIT)

A straightforward approach is to replace the missing values by plausible guesses. For dichotomous 0-1 coded covariates, Schemper and Smith (1990) suggested to *impute* the probability of a factor taking on the value 1 conditional on the values of the other factors in the model. This probability could be estimated by applying a logistic regression (Hosmer and Lemeshow, 1989) of the incomplete covariate on the complete covariates. The dichotomous covariate becomes a continous one then. If we have one factor subject to missing values, we estimate this probabilities from our complete observations. If we have more than one factor subject to missing values, several patterns of missing data can occur. Schemper and Heinze (1996) suggest an iterative procedure:

Algorithm 1: Computing imputed values in a data set with more than one variable subject to missing values.

```
for (each set of subjects with the same missingness pattern)
   {
   compute preliminary imputations for the factor to be imputed by
   logistic regression on all other non-missing factors of this pattern
   using all subjects of the dataset, where the factor to be imputed and
   the non-missing factors of the pattern are observed
   }
do
   {
   for (each variable subject to missing values)
      {
      compute imputations from logistic regression on all other variables,
      where missing values are replaced by previous imputations
      }
   }
while (convergence in the imputations not achieved);
```

Applying this algorithm to our data and using again standard software for proportional hazards regression, we get the results presented in Table 2.

Table 2: Regression coefficients (log hazard ratios) of a proportional hazards model fitted on 483 observations imputed using PIT. Marginal coefficients refer to unadjusted models, partial coefficients to the model involving all covariates.
* denotes p-value ≤ 0.05, ** $p \leq 0.01$, *** $p \leq 0.001$, **** $p \leq 0.0001$.

Factors	marginal	partial
TM	-0.20	-0.14
AG	0.43****	0.29*
PF	0.75****	0.31
HG	0.51****	0.27*
SZ	0.79****	0.61***
HX	0.60****	0.54****
SG	0.65****	0.48***
WT	0.39**	0.21

Now, in the adjusted model, five of the factors are significant, and all but one in the unadjusted models. A question now arising is, *how much can we trust in these results*?

2 Properties of Probability Imputation

Recent investigations (Heinze, 1995; Schemper and Heinze, 1997; Vach, 1994; Vach and Schumacher, 1993) show that after applying PIT, regression coefficients cannot be estimated consistently in logistic and proportional hazards models. The magnitude of this asymptotic bias is strongly dependent on the distribution of the missing values. Table 3 gives an impression of what we can expect in typical settings of a prognostic factor study involving logistic regression. In this table, Y denotes the binary outcome, X_1 denotes a completely observed binary variable, X_2 refers to a binary variable subject to about 33 % missing values, and Z refers to factors not included - and not correlated with any variables - in the model. „Small bias" denotes situations where the estimated asymptotic bias of the parameter estimation after applying PIT is within an interval of [+/- 0.05] in all settings, „moderate bias" means an estimated asymptotic bias which at least once exceeds [+/- 0.05] but is still within [+/- 0.20], „large bias" denotes a bias occuring in at least one setting that exceeds the „moderate bias" interval.

Table 3: Estimated asymptotic bias of regression coefficients in logistic regression after using PIT based on a simulation study. Detailled specifications of the simulation study can be found in the appendix.

Distribution of missing values depending on	**Estimation of complete factor (X_1)**	**Estimation of incomplete factor (X_2)**
Z, X_1	small bias	small bias
X_2	small bias	small bias
Y	small bias	moderate bias
X_1 and Y	moderate bias	moderate bias
X_2 and Y	large bias	large bias
X_1 and X_2	moderate bias	small bias

As long as we are not concerned with a distribution of missing values depending on the outcome Y, which we can expect from prognostic factor studies, we are not confronted with large bias. However, in case-control studies, where the cases and the controls are sampled from different sources, we should be aware of different missingness generating processes, which causes more bias when PIT is applied.

Note that even if the distribution of the missing values is depending only on the true unobservable values, this situation is denoted by „nonignorable missingness" in the literature as compared to „missing at random", we have acceptable performance of PIT.

In the simulation we also studied the effect of imputation on inference about the parameter estimates. For the situations where Probability Imputation is reasonable with respect to asymptotic bias, the results are in brief:

- the residual variance is increased by 2 - 3 % compared to full data (no missingness) analysis
- 95 %-confidence intervals around a parameter estimated by PIT covers the true value in 94 - 95 %, this means that tests are valid

- tests on the effect of a complete variable achieve the power of full data analysis if there is no correlation in the covariates, and some 5 % less power than full data analysis if covariates are correlated
- tests on the effect of the imputed variable achieve 1 - 3 % less power than analysis of the completely observed cases

The sensitivity of the analysis on the assumption about the distribution of the missing values made by the Probability Imputation Technique can be examined by several analyses with lower or higher imputation values. For our example, we could decrease and increase the imputed values of HX to see the effect on, say, the strongest factors SZ and SG.

Table 4: Analysis of sensitivity of different assumptions about the distribution of the missing values of HX on parameter estimates of SZ and SG in the survival time data set.

Imputed value of HX increased by	**SZ**	**SG**
-0.2	0.61366	0.45606
-0.1	0.61239	0.46978
+0.0	0.61186	0.48189
+0.1	0.61200	0.49126
+0.2	0.61288	0.49495

As we see, the adjusted parameter estimate of SZ does not change with different imputed values of HX. The changes in the parameter estimate of SG indicate that a misspecification of the distribution of the missing values of HX, if it occurred, would have altered our conclusions on SG; but the changes are small, so we will not worry about them.

3 Comparison with other approaches

3.1 Multiple Imputation

In Probability Imputation, we impute the expectation of the incomplete covariate conditional on the other, independent factors. The imputed values can assume all values between 0 and 1.

In Multiple Imputation as proposed by Rubin (1987), we impute draws from the distribution of the incomplete covariate conditional on all other factors and the dependent variable. These draws can assume the values 0 or 1 only.

Having completed the data set by draws, it is analysed, and the parameters estimated are stored. The imputation/analysis procedure is repeated m times, where m should lie between 3 and 10, say. After m steps, we compute the mean of the parameter estimates over the analyses to get a multiple imputation estimate of the regression coefficients. The variance of the estimates can be computed by adding the within-step variance W and the between-step variance B, corrected by 1+1/m. For inference about the parameters, we use the t-distribution with (m-1)(1+m/(m+1))W/B degrees of freedom.

As in PIT, we have to estimate the probability of the missing value being 1. This probability now depends also on the outcome variable. Why not use the outcome variable in PIT? Little (1992) states that when imputing means, including the dependent variable in the computation of the conditional mean introduces bias. The conditional expectation is then too related to the observed outcome data, leading to a wrong direction in the estimation of the regression coefficients.

In the simulation study mentioned above, Multiple Imputation was compared to Probability Imputation. We found that Multiple Imputation has little less power than Probability Imputation, but it produces asymptotically unbiased estimates in all missing at random situations. Therefore it is useful also in case-control studies. For complex relationships between dependent and independent variables, e. g. in survival studies, where there are censored outcomes, the conditional distribution of the missing values is not easy to compute, and a satisfactory Multiple Imputation has not been yet published.

3.2 Maximum Likelihood Estimation

Recent research (Vach and Schumacher, 1993; Vach, 1994) has concentrated on building a model for the distribution of the missing values, and estimating the parameters of this model along with the parameters of interest. In practise, the log-likelihood function has summands refering to missing values of an incomplete variable, and in these summands we integrate over the conditional distribution of this variable. There are some special cases, where this procedure can be done „quite easily“, but in general software for ML-estimation with missing values is not yet available.

Maximum Likelihood estimation in logistic regression with missing dichotomous covariates is exhaustively studied in Vach (1994). If the distribution of the missing values is independent of the outcome, as it is expected in prognostic factor studies, Maximum Likelihood yields the same results as Probability Imputation, with the advantage of estimating asymptotically unbiased. The bias of PIT, however, is very small, and only relevant if the factor subject to missing values has a true regression coefficient greater than 1.5. If the missing-at-random assumption is violated, PIT and Maximum Likelihood show a comparable sensitivity.

Screening of marginal and partial effects is easily done in Probability Imputation and Multiple Imputation, and the procedure to obtain marginal regression coefficients in Maximum Likelihood estimation is not so intuitive.

In case-control-studies, however, PIT is not recommended, and one should consider Maximum Likelihood or Multiple Imputation, or estimate using the completely recorded observations only.

Acknowledgements

The author would like to thank Michael Schemper for making important contributions to this work, an anonymous referee for making helpful suggestions, and the committee of the 1996 International Conference on Methodology and Statistics.

References

[1] Andrews, D. F. and Herzberg, A. M. (1985): *Data - A Collection of Problems from Many Fields for the Student and Research Worker*. New York: Springer.

[2] Cox, D. R. (1973): Regression models and life tables (with discussions). *Journal of the Royal Statistical Society*, Series B, **34**, 187-220.

[3] Heinze G. (1995): *Imputationsverfahren in der Logistischen Regression mit dichotomen Kovariablen*, Unpublished Master's thesis, Department of Medical Computer Sciences, Vienna University, Vienna.

[4] Hosmer, D. W. and Lemeshow, S. (1989): *Applied Logistic Regression*. New York: Wiley.

[5] Little, R. J. A. (1992): Regression with missing X's: A Review. *Journal of the American Statistical Association*, **87**, 1227-1237.

[6] Rubin, D. B. (1987): *Multiple Imputation for nonresponse in Surveys*. New York: Wiley.

[7] Schemper, M. and Heinze, G. (1997): Probability Imputation Revisited for Prognostic Factor Studies. *Statistics in Medicine*, **16**, 73-80.

[8] Schemper, M. and Smith, T. J. (1990): Efficient evaluation of treatment effects in the presence of missing covariate values. *Statistics in Medicine*, **9**, 777-784.

[9] Vach, W. (1994): *Logistic Regression with Missing Values*, Lecture Notes in Statistics 86, New York: Springer.

[10] Vach, W. and Schumacher, M. (1993): Logistic regression with incompletely observed categorical covariates: A comparison of three approaches. *Biometrika* **80**, 353-362.

Appendix: Specifications of simulation study

Tables A.1 and A.2 show the constant and varied parameters, respectively, of the simulation study. Table A.3 shows in detail the missing value distributions used.

Table A.1: Constant parameters of simulation study.

Simulation parameter	**Value**
Number of simulation for each setting	1000
Number of observations per simulated data set	200
Distribution of X_1	binary, $Pr(X_1=0) = Pr(X_1=1)=0.5$
Distribution of X_2	binary, $Pr(X_{2true}=0) = Pr(X_{2true}=1)=0.5$
Distribution of Y	$Pr(Y=1) = 1/(1+exp(-b_0-b_1X_1-b_2X_{2true}))$, $Pr(Y=0) = 1-Pr(Y=1)$
b_0 = Intercept of logistic regression model	-1

Table A.2: Varied parameters of simulation study.

Simulation parameter	**Values**
b_1 = true logistic regression parameter value of X_1	0, 1
b_2 = true logistic regression parameter value of X_2	0, 1
Pearson correlation between X_1 and X_2	0, 0.6
Distribution of missing values of X_2 depending on	Z, X_1, X_{2true}, Y, X_1 and Y, X_{2true} and Y, X_1 and X_{2true}

Table A.3: Missing value generating mechanisms

Missing value distribution of X_2 depending on	**Missing value distribution of X_2**
Z	Pr(X_2=?) = 0.33
X_1	Pr(X_2=?|X_1=0) = 0.17, Pr(X_2=?|X_1=1) = 0.5
X_{2true}	Pr(X_2=?|X_{2true}=0) = 0.17, Pr(X_2=?|X_{2true}=1) = 0.5
Y	Pr(X_2=?|Y=0) = 0.17, Pr(X_2=?|Y=1) = 0.5
X_1 and Y	Pr(X_2=?|X_1=0 and Y=0) = 0.086, Pr(X_2=?|X_1=1 and Y=0) = 0.254, Pr(X_2=?|X_1=0 and Y=1) = 0.254, Pr(X_2=?|X_1=1 and Y=1) = 0.746
X_{2true} and Y	Pr(X_2=?|X_{2true}=0 and Y=0) = 0.086, Pr(X_2=?|X_{2true}=1 and Y=0) = 0.254, Pr(X_2=?|X_{2true}=0 and Y=1) = 0.254, Pr(X_2=?|X_{2true}=1 and Y=1) = 0.746
X_1 and X_{2true}	Pr(X_2=?|X_1=0 and X_{2true}=0) = 0.086, Pr(X_2=?|X_1=1 and X_{2true}=0) = 0.254, Pr(X_2=?|X_1=0 and X_{2true}=1) = 0.254, Pr(X_2=?|X_1=1 and X_{2true}=1) = 0.746

Each combination of the simulation parameters (there were 2*2*2*7 = 56) was used to construct 1000 data sets each containing 200 observations. For each dataset, in a first step feasible values of X_2 were generated and in a second step they were set to "missing" according to the distribution of missing values. In tables A.1, A.2 and A.3, the preliminary values of X_2 are denoted by X_{2true}.