Editorial

Open Access

Missing data in longitudinal covariates in building a cox prediction model: overview and some practical guidance

Abstract

Missing covariates can be commonly observed in human subject studies. Consider a clinical example in which a medical prediction needs to be made to determine a patient's mortality at a certain point in time after a critical surgery under the condition that personal risk factor covariates have missing values. If missing values are small or completely at random, subjects with incomplete covariates can be removed from analysis. However, if the missingness occurred substantially with at least 30% of subjects having incomplete data, their deletion will change the description of an analysis sample. Then what kind of practice can be done to build a prediction model that has missing covariates? Missing data methods for such cases have been extensively studied and developed in the past decades. This article over views and recommends practical guidance for such missing covariates in building a prediction model.

Volume 2 Issue 3 - 2015

Joseph Kang

Department of Preventive Medicine, Northwestern University, USA

Correspondence: Joseph kang, Department of Preventive Medicine, Northwestern University, Feinberg School of Medicine, USA, Tel: 773-850-2141, Fax 312-908-9588, Email joseph-kang@northwestern.edu

Received: April 07, 2015 | **Published:** April 16, 2015

Categorization of missing values

To identify missing information, the rate of missing data will first have to be described before applying any complex statistical methods such as imputation. After identifying missing rates, missing values can be categorized for an initial analysis that does not involve imputation or deletion. For example, suppose patients' BMI (body mass index) values were missing. Then BMI values can be categorized to have low, medium, high and missing so that the missing category can be still used in descriptive and inferential statistical analysis. Other variables, which do not have clinical thresholds, can be categorized with respect to their observed quantiles. Categorical variables are straightforward to handle because a missing data category can be simply added. Such categorization is an initial step to the overall assessment of covariates that are associated with the patient mortality outcome. For this particular example, a Cox model with time-dependent covariates, which have categorized missing data as described previously, can be initially used to build a prediction model for patient mortality.

Machine learning methods that automatically categorize missing values

Because the initial Cox model, which was described in the previous section, may poorly convey information regarding the nonlinearity of categorized continuous variables, the boosting algorithm[1,2] can be used to independently build another Cox model for predicting the mortality outcome. The boosting algorithm was implemented in the R package gbm and has been used in observational studies.[3] The boosting algorithm uses regression and the classification tree (CART)[4] which treats missing values as categorical values without imputing or deleting them. The boosting algorithm is often compared with random forests method[5] which fundamentally uses the CART technology. Random forests were implemented in the R package random Forest. Both the boosting method and random forests method can be employed to build a nonparametric Cox prediction model for mortality outcome. While these two methods are expected to provide a reasonably good prediction, the influence of longitudinal variables in the Cox model will not be easily interpreted because the functional form of covariates in these two machine learning methods are nonparametric, without having any structure. Such feature of these machine learning methods make them appear as "black box" tools, which, to researchers, means that no effect size such as hazard ratio will be directly produced from those models. Thus if it is a research interest to know the effect of a certain covariate that has missing values in a Cox model, the imputation of missing values will make the analysis more interpretable.

Imputation of missing values

The imputation of missing values to generate complete data set will enable us to assess the Cox model parameters and the effects of time-dependent covariates. Missing covariates can be easily imputed with the multiple imputation (MI) method[6,7] via the R package mice. MI imputes multivariate missing values under the assumption that imputation models are correct. MI has gained its popularity since its first inception in the 1980's. For example, MI was used to impute missing values in income reports in the National Health Interview Survey data sets managed by Centers of Diseases Control and Prevention (CDC). Various forms of MI are available in many commercial software programs. However, a blind application of MI needs to be cautioned because MI fundamentally depends on the assumption that imputation models for missing values are correct. Therefore sensitivity analysis with various imputation model options needs to be conducted to evaluate the impact of missing values. Apart from MI, weighting of observed data has also been extensively studied. The weighting method uses the inverse of probability weighting (IPW) method which is also known as the Horvitz-Thompson estimator.[8] Though theoretically reasoned well, the practical implementation and usage of the IPW methods do not seem as popular as the MI method.

Final model selection

The three approaches for dealing with missing data to build predictive Cox models with 1) categorized covariates, 2) boosting algorithm and 3) the multiple imputation method can be compared with respect to AUC values which assess predictive utility. In particular, estimating AUC values in the Cox model is available in the

R package survival ROC. Comparing AUC values, the best model will be chosen and used as the final predictive model.

Acknowledgement

None.

Conflict of interest

None.

References

1. Friedman JH. Stochastic Gradient Boosting. *Computational Statistics and Data Analysis*. 2002;38(4):367–378.
2. Hastie T, Tibshirani R, Friedman JH. *The elements of statistical learning : data mining, inference, and prediction : with 200 full-color illustrations*. Corrected print. 2nd ed. Springer, New York: 2002; pp 533.
3. Daniel F Mc Caffrey, Lane F Burgette, et al. *Toolkit for Weighting and Analysis of Nonequivalent Groups*. 2015.
4. Breiman L, Jerome F, Olshen R, et al. *Classification and Regression Trees*. 1st Ed. New York Chapman & Hall. 1984.
5. Breiman L, Last M, Rice J. Random Forests: finding quasars . In: Eric D. Feigelson, G. Jogesh Babu, editors. *Statistical challenges in astronomy: Third Statistical Challenges in Modern Astronomy (SCMA III) Conference*. University Park, PA, USA, 2003; p.243–254.
6. Rubin DB, Hoboken NJ. *Multiple imputation for nonresponse in surveys*. 2004; 320 p.
7. Buuren Sv, Boca Raton. *Flexible imputation of missing data*. CRC Press; 2012; 316 p.
8. Horvitz DG, Thompson DJ. A Generalization of Sampling without Replacement from a Finite Universe. *Journal of the American Statistical Asso*. 1952;47(260):663–85.

Citation: Kang J. Missing data in longitudinal covariates in building a cox prediction model: overview and some practical guidance. *Biom Biostat Int J.* 2015;2(3):85–86. DOI: 10.15406/bbij.2015.02.00030