An Improved Nonlinear Imputation/Transformation Method

Frank E. Harrell Jr
Clinical Biostatistics
Division of Biometry and The Heart Center
Duke University Medical Center
Box 3363 Durham NC 27710
feh@biostat.mc.duke.edu

Biometric Society – ENAR

12 April 1994

AN IMPROVED NONLINEAR IMPUTATION/TRANSFORMATION METHOD

Frank E Harrell Jr
Division of Biometry, Duke University Medical Center, Durham NC USA

Regression modeling is often plagued by at least two problems: too many parameters per observation and missing predictor variables. Multiple parameters arise from numerous predictors and nonlinear and interaction terms. One method for reducing d.f. is to pre–transform the predictors without reference to the response. To solve the missing data problem, incomplete records may be deleted, but the estimates are biased and inefficient. Therefore, some kind of imputation is needed.

Young, Takane, and de Leeuw (Psychometrika 43:279; 1978) developed the maximum total variance (MTV) method for simultaneously estimating transformations of a series of variables of mixed types (continuous, polytomous, ordinal). The MTV method solves for transformations that maximize the variation explained by the first m principal components. Kuhfeld extended MTV by incorporating simultaneous imputation in the SAS `PRINQUAL` procedure (SAS/STAT User's Guide, Vol. 2). Kuhfeld also implemented the maximum generalized variance algorithm (MGV) due to Sarle, where variables are transformed to maximize the first canonical correlation between the terms representing the variable and the set of all other variables. MTV and MGV are useful even without imputation, as the transformations they produce are generally closer to the true transformation than are linear representations, without costing d.f.

When the fraction of missing values is not tiny, MTV and MGV's imputation strategy does not converge well. Multiple variables with missing values on the same observation are especially problematic. This talk will cover an extension and modification of MGV that is stable even with a large fraction of missing values. The new algorithm, `transcan` (*t*ransformations using *c*anonical variates) is available in the `statlib` repository for S–functions. `transcan` works well with multiple non–monotonically–related predictors, such as lab values from critically ill patients.

Outline

1. Need for Imputing Missing Data
2. Multivariate Transformation Techniques
3. New Method
4. Robustness to Fraction of Missing Data
5. Example of Transforming Multiple Predictors
6. Summary

1 Need for Imputing Missing Data

- Almost always discard obs. with missing Y
- Discard obs. with missing X only if X is extremely important and can't be imputed from other predictors
- Discarding many obs. with missing $X \rightarrow \uparrow$ variance and bias [3]
- Continuous X unrelated to other Xs $\rightarrow$ substitute median or mean [3]

- Otherwise better to use individual predictive model for each X based on other Xs [2, 6–8]
- Non–monotonically transformed $X \rightarrow$ general transformation procedure should be used while imputing
- Most methods assume missing at random

2 Multivariate Transformation Techniques

- Mixture of qualitative and continuous variables: qualitative principal components
- Maximum total variance (MTV) of Young, Takane, de Leeuw [10]
  1. Compute PC_1 of variables using correlation matrix
  2. Use regression (with splines, dummies, etc.) to predict PC_1 from each X — expand each X_j and regress it separately on PC_1 to get working transformations
  3. Recompute PC_1 on transformed Xs
  4. Repeat 3–4 times until variation explained by PC_1 plateaus and transformations stabilize
- Maximum generalized variance (MGV) method of Sarle [5, pp. 1267-1268]
  1. Predict each variable from (current transformations of) all other variables
  2. For each variable, expand it into linear and non–linear terms or dummies, compute first

canonical variate

3. For example, if there are only two variables X_1 and X_2 represented as quadratic polynomials, solve for a, b, c, d such that $aX_1 + bX_1^2$ has maximum correlation with $cX_2 + dX_2^2$.
4. Goal is to transform each var. so that it is most similar to predictions from other transformed variables
5. Does not rely on PCs or variable clustering

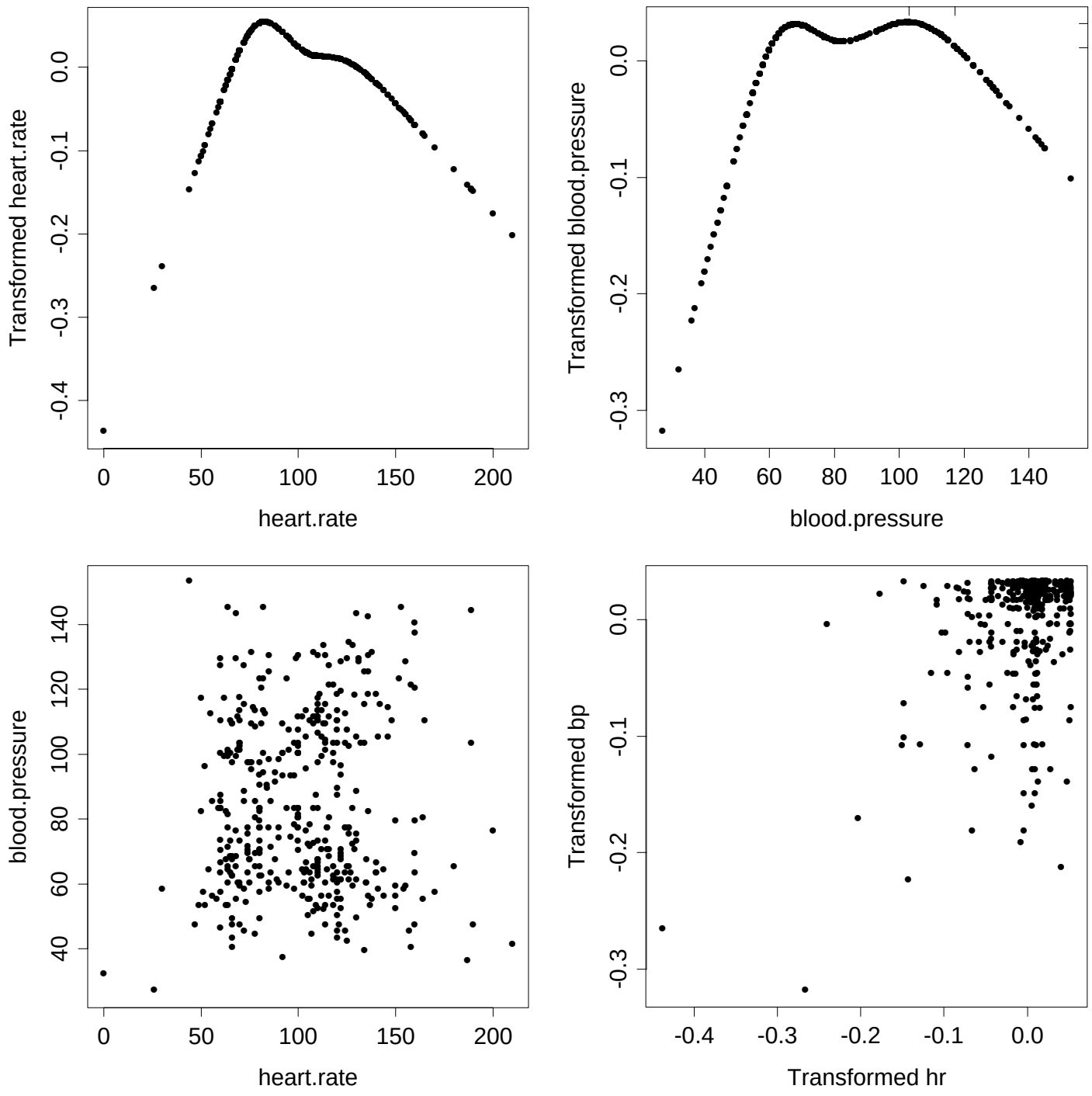

Figure 1: Transformations fitted using `transcan`*. Tick marks indicate the two imputed values for blood pressure. The lower left plot contains raw data (Somers' $D = 0.02$); the lower right is a scatterplot of the corresponding transformed values ($D = 0.14$). Data courtesy of the SUPPORT study*

- MTV and MGV implemented in SAS `PROC PRINQUAL` [5]
  1. Allows flexible transformations including monotonic splines
  2. Does not allow restricted cubic splines, so may be unstable unless monotonicity assumed
  3. Allows simultaneous imputation but often yields wild estimates
- ACE (Alternating Conditional Expectation) of Breiman and Friedman [1]
  1. Uses nonparametric "super smoother" [4]
  2. Allows monotonicity constraints, categorical vars.
  3. Does not handle missing data

3 New Method

- Initialize missings to medians or most frequent categories
- Initialize transformations to original variables

- Take each variable in turn as Y
- Exclude obs. missing on Y
- Expand Y (spline or dummy variables)
- Score (transform Y) using first canonical variate
- Missing $Y \rightarrow$ predict canonical variate from Xs
- Constrain imputed values to be in range of non–imputed ones
- Convergence based on maximum change in fitted transformation
- Optionally shrink imputed values using Van Houwelingen and Le Cessie [9]
- Imputations on original scale
  1. Continuous $\rightarrow$ back–solve with linear interpolation
  2. Categorical
     (a) Use category whose canon. score is closest to prediction
     (b) Alternative: classification tree (most freq. cat.)

- Option to insert constants as imputed values (ignored during transformation estimation)
- Easy out–of–data transformation/imputation
- A function (`Function.transcan`) creates S functions that analytically transform each variable
- These methods find marginal transformations
- Check adequacy of transformations using Y
  1. Graphical
  2. Nonparametric smoothers (X vs. Y)
  3. Expand original variable using spline, test additional predictive information over original transformation

4 Robustness to Fraction of Missing Data

- Let X_1 be a vector of $n = 500$ random normal values
- Define

$$X_2 = 1 - \exp(-\max(X_1, -2.5)) + \epsilon_1$$
$$X_3 = X_1 + 0.4\epsilon_2,$$

  where ϵ_1, ϵ_2 are independent random standard normal vectors.
- Let f denote fraction of X_1 to set to missing at random
- Obs. with missing $X_1 \rightarrow$ set $\frac{1}{4}$ of X_3 to missing
- Obs. with non–missing $X_1 \rightarrow$ set $\frac{1}{4}f$ of X_3 to missing
- Vary f from 0 to 0.95

5 Example of Transforming Multiple Predictors

- Subset of APACHE III database: 1195 ICU patients with sepsis
- Follow–up to 28d
- Predict time to death

Variable Name	Meaning
age	age (y)
icuday	day in ICU before at qualification
prelos	days in hospital before ICU admission
bili	bilirubin (mg/dl)
wblc	white blood count ($\times 10^3/\text{mm}^3$)
temp	temperature, (°C)
hrat	heart rate (/min)
resr	respiration rate
seph	serum pH
crea	serum creatinine (mg/dl)
gluc	glucose
hect	hematocrit
pao2	$\text{P}_a\text{O}_2/\text{F}_i\text{O}_2$
pco2	PCO_2
sbun	BUN
sena	sodium (meq/L)
uout	urine output
mblp	mean arterial blood pressure

- Bilirubin most frequently missing ($n = 498$)

- In a separate study, model using all transformed predictors validated better than a model which derived transformations from response variable
- This is due to data reduction

6 Summary

- Modification of MGV method of SAS `PROC PRINQUAL` result in more stability
  1. Constrain imputed values to be in range of actual values
  2. Different convergence criteria
  3. Temporarily discard observations with missings while variable is being predicted
  4. Shrinkage prevents "over–imputing"
- In many cases (especially where there is a common pathway to the endpoint), the non–response–variable transformations are adequate
- In most cases, the transformations are better than assuming linearity
- The reduction in the number of parameters to estimate (with respect to Y) can result in better predictive validation
- This scaling/transformation method provides diagnostics for "cononlinearity"

References

[1] L. Breiman and J. H. Friedman. Estimating optimal transformations for multiple regression and correlation (with discussion). *Journal of the American Statistical Association*, 80:580–619, 1985.

[2] S. F. Buck. A method of estimation of missing values in multivariate data suitable for use with an electronic computer. *Journal of the Royal Statistical Society B*, 22:302–307, 1960.

[3] A. Donner. The relative effectiveness of procedures commonly used in multiple regression analysis for dealing with missing values. *American Statistician*, 36:378–381, 1982.

[4] J. H. Friedman. A variable span smoother. Technical Report 5, Laboratory for Computational Statistics, Department of Statistics, Stanford University, 1984.

[5] W. F. Kuhfeld. The PRINQUAL procedure. In *SAS/STAT User's Guide*, volume 2, chapter 34, pages 1265–1323. SAS Institute, Inc., Cary NC, Fourth edition, 1990.

[6] J. S. Roberts and G. M. Capalbo. A SAS macro for estimating missing values in multivariate data. In *Proceedings of the twelfth annual SAS Users Group International Conference*, pages 939–941, Cary NC, 1987. SAS Institute, Inc.

[7] M. Schemper and T. L. Smith. Efficient evaluation of treatment effects in the presence of missing covariate values. *Statistics in Medicine*, 9:777–784, 1990.

[8] N. H. Timm. The estimation of variance–covariance and correlation matrices from incomplete data. *Psychometrika*, 35:417–437, 1970.

[9] J. C. Van Houwelingen and S. Le Cessie. Predictive value of statistical models. *Statistics in Medicine*, 8:1303–1325, 1990.

[10] F. W. Young, Y. Takane, and J. de Leeuw. The principal components of mixed measurement level multivariate data: an alternating least squares method with optimal scaling features. *Psychometrika*, 43:279–281, 1978.