Multiple Imputation of Squared Terms

Gerko Vink   Stef van Buuren

XVIII ISA World Congress of Sociology

Universiteit Utrecht   TNO innovation for life   cbs

Some introduction

- Multiple imputation (MI)[1]
  - A general and valid approach for dealing with missing data
  - Widely accepted and well documented for a variety of problems and situations.
  - Can be very flexible
- MI requires **correct** specification of the imputation model
  - Straightforward when the analysis model contains main effects
  - Less clear when nonlinear terms are included

[1] Rubin, D. B. (1987). Multiple imputation for nonresponse in surveys. New York: Wiley.

Adding nonlinear terms

- Take the following model of scientific interest

$$Y = \alpha + X\beta_1 + X^2\beta_2 + \epsilon$$

- If we want to predict Y from X and its square X^2 , then both X and X^2 should be included in the imputation model.
- Leaving the term X^2 out of the imputation model will result in a downward bias of the slopes when we perform a regression analysis on the imputed data.

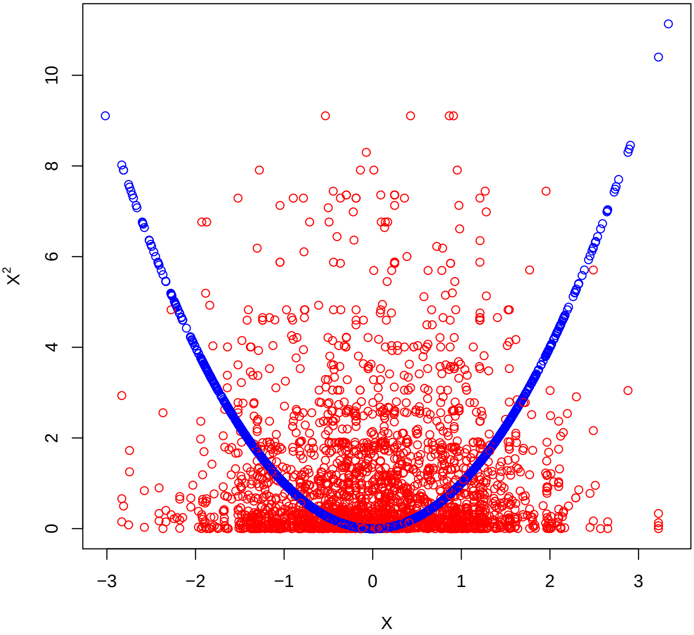
Imputing squares: Option[2] 1

- Transform, then impute
  - Transform X in the incomplete data and impute X^2 as just another variable.
  - Yields unbiased regression estimates under MCAR
  - Heavily distorts the relation between X and X^2 after imputation.

	Missingness Mechanism				
	MCAR	MARleft	MARmid	MARtail	MARright
Transform, then impute					
Intercept (α)	0	0.19	-0.13	0.01	-0.05
Slope of X (β_1)	1	0.91	0.97	1.14	1.32
Slope of X^2 (β_2)	1	0.91	0.95	1.14	1.32
Residual SD (σ_ϵ)	1	0.95	1	1.06	1.15
R^2	0.75	0.77	0.75	0.72	0.67

[2] These options have been studied by Paul von Hippel in Von Hippel, P. (2009) How to Impute Interactions, Squares, and Other Transformed Variables. Sociological Methodology 39:265-91.

Transform, then impute



Imputing squares: Option 2&3

- Impute, then transform
  - Transform X to X^2 in the imputed data.
  - Preserves the relation between X and X^2 after imputation.
  - Yields heavily biased regression estimates (also under MCAR)
- Passive imputation
  - Computes X^2 each time X is imputed.
  - Equivalent to 'Impute, then transform'

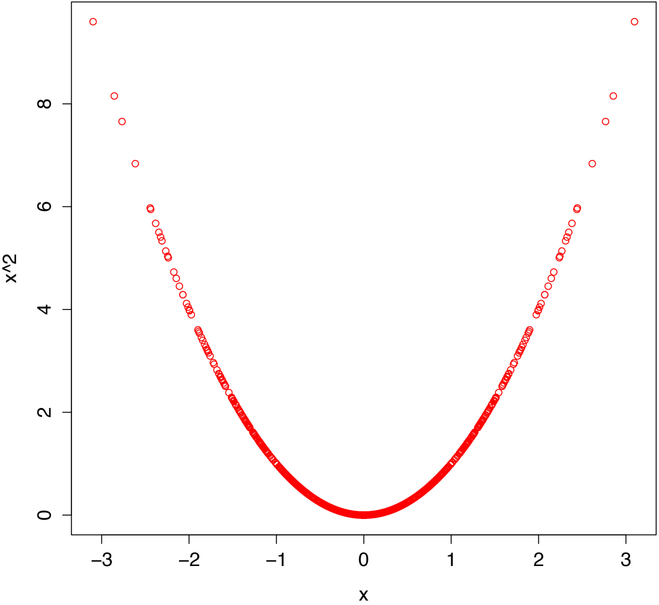
	MCAR	MARleft	MARmid	MARtail	MARright
Impute, then transform					
Intercept (α)	0.39	0.29	0.26	0.52	0.56
Slope of X (β_1)	0.93	0.94	0.87	1.01	1.06
Slope of X^2 (β_2)	0.61	0.60	0.67	0.56	0.66
Residual SD (σ_ϵ)	1.48	1.44	1.41	1.56	1.62
R^2	0.45	0.48	0.5	0.39	0.34

Proposal: bivariately impute X and X^2

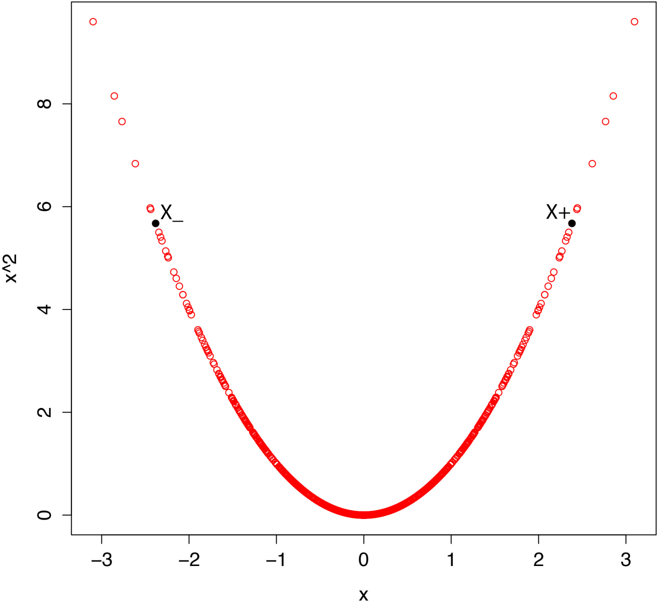
- Polynomial combination imputation
  - Do not impute X and X^2 but rather impute the linear combination $Z = X\beta_1 + X^2\beta_2$
  - Decompose the imputed linear combination Z into the distinct real roots

$$X_- = -\frac{1}{2\beta_2}\left(\sqrt{4\beta_2 Z + \beta_1^2} + \beta_1\right)$$
$$X_+ = \frac{1}{2\beta_2}\left(\sqrt{4\beta_2 Z + \beta_1^2} - \beta_1\right)$$

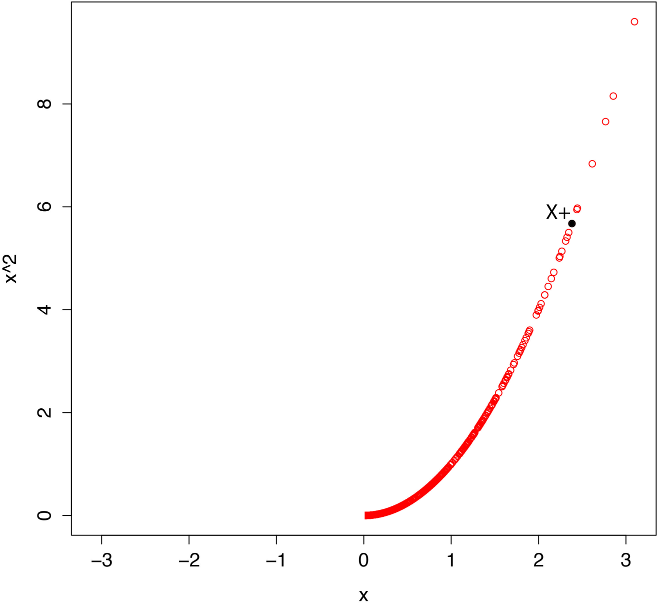
Choosing a root



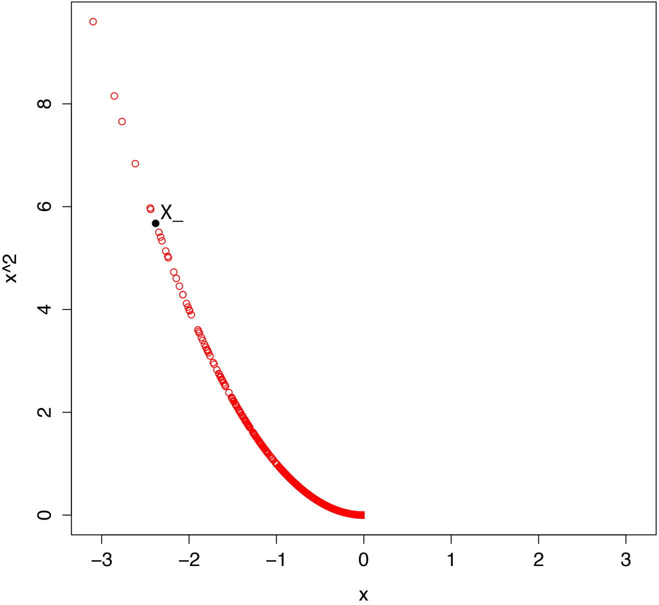
Choosing a root



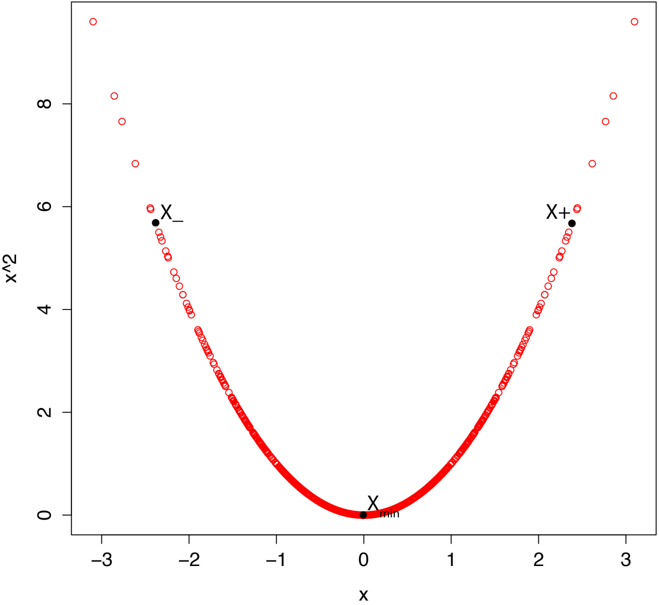
Choosing a root



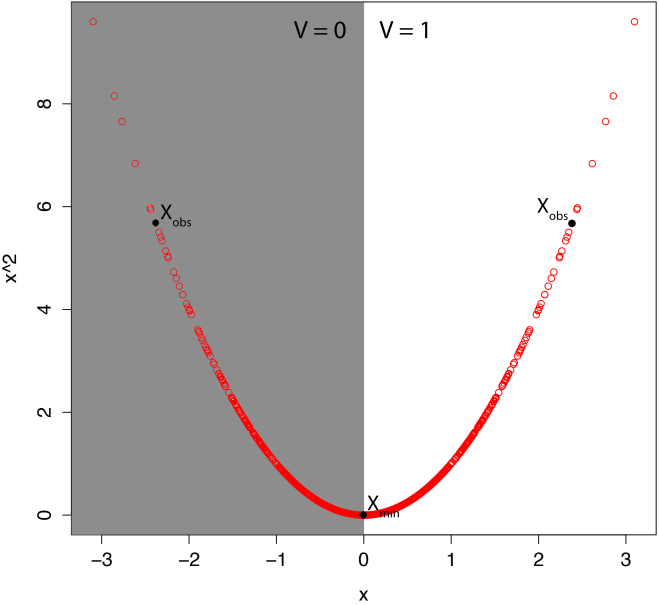
Choosing a root



The minimum is located at $X_{min} = -\beta_1 / -2\beta_2$



Create binary variable $V = (V_{obs}, V_{mis})$



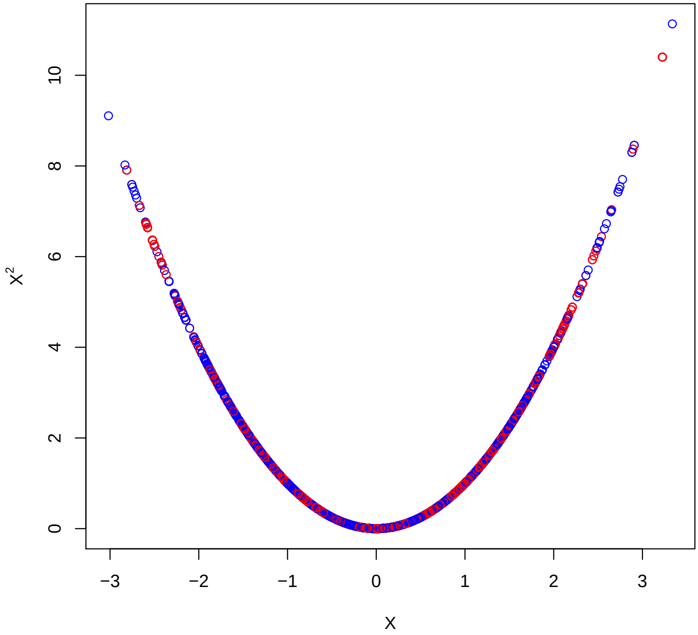
Proposal: drawing roots with some probability

- We can model the probability V_{obs} for either arm on the observed Z as

$$\text{logit} P(V = 1) = Y\beta_Y + Z\beta_Z + YZ\beta_{YZ}.$$

- We can expand this model to obtain estimated (or imputed) probabilities V_{mis} for imputed Z.
- Sample either X_- or X_+ by randomly drawing from the binomial distribution using the above probability.
- Calculate X^2 from the imputed X

Polynomial Combination



Polynomial Combination

- Yields unbiased regression estimates under MCAR and MAR
- Preserves the relation between X and X^2 after imputation.
- Easily applicable and already available in `mice`[3] in `R`[4]

	MCAR	MARleft	MARmid	MARtail	MARright
Polynomial combination					
Intercept (α)	0	-0.01	-0.01	-0.05	-0.07
Slope of X (β_1)	1	1	1	0.96	0.96
Slope of X^2 (β_2)	1	1	1.01	1.06	1.09
Residual SD (σ_ϵ)	1	1	1	1.03	1.05
R^2	0.75	0.75	0.75	0.73	0.73

[3] Stef van Buuren, Karin Groothuis-Oudshoorn (2011). mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software, 45(3), 1-67.

[4] R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Published as Vink, G., & van Buuren, S. (2013). Multiple Imputation of Squared Terms. Sociological Methods & Research, 42(4), 598-607.