countimp 1.0 – A multiple imputation package for incomplete count data

Kristian Kleinke & Jost Reinecke
University of Bielefeld, Faculty of Sociology and Centre for Statistics

Overview – Functions of the `countimp` package

The `countimp` package contains functions to multiply impute incomplete

ordinary count data
- Poisson imputation

overdispersed count data
- Quasi-Poisson imputation
- Negative Binomial imputation

Zero-inflated (and overdispersed) count data
- Zero Inflated Poisson imputation (ZIP)
- Zero Inflated Negative Binomial imputation (ZINB)

Multilevel count data
- two-level Poisson imputation
- two-level NB imputation
- two-level imputation for zero-inflated (and overdispersed) data

Overview – Theoretical Background

Our count data imputation procedures
- are based on the *multiple imputation by chained equations* framework (Raghunathan, Lepkowski, van Hoewyk, & Solenberger, 2001; van Buuren & Groothuis-Oudshoorn, 2011)
- work as add-ons to the `mice` software in R (van Buuren & Groothuis-Oudshoorn, 2011)
- follow either Rubin's (1987) Bayesian regression approach, or
- follow a bootstrap regression approach.

Multiple Imputation

3 Steps:
1. Impute each missing value m times (with different but equally plausible values) and obtain m complete datasets.
2. Analyze each of these m complete datasets separately and obtain m statistical results.
3. Combine these m results into an overall result using Rubin's rules for MI inference. These rules take variation **within** and **between** these imputed datasets into account. This additional variation reflects uncertainty in parameter estimation due to missing data.

Multiple Imputation by Chained Equations

- A separate conditional model $P(Y_j|Y_s, \theta_j)$ is specified for each incompletely observed variable Y_j in the dataset.
  Y_j denotes a variable with missing values.
  Y_s is a subset of Y containing some or all of the variables in the dataset except from Y_j that is used to model Y_j.
- Imputations of missing values in Y_j are then generated from $P(Y_j|Y_s, \theta_j)$.
- Missing data are assumed to be MAR. Furthermore the assumptions of the respective regression model should be met.

Bayesian Regression Variant

1. Fit regression model and get posterior distribution of model parameters θ based on the observed data $P(\theta|Y_{obs})$.
2. Introduce between imputation variability: Draw new parameters θ^* from $P(\theta|Y_{obs})$.
3. Impute missing data Y^* from $P(Y_{mis}|Y_{obs}, \theta^*)$.
4. Repeate steps 2 and 3 m times to obtain the m imputations.

Bootstrap Regression Variant

1. Fit regression model to bootstrap sample and get model parameters θ.
2. Predict missing data based on these parameters.
3. Repeate steps 2 and 3 m times to obtain the m imputations.

Evaluation

Monte Carlo Simulations assessing the procedures' quality may be found in
- Kleinke, de Jong, Spiess, & Reinecke (2011) – imputation of ordinary and overdispersed count data
- Kleinke & Reinecke (2013a) – imputation of zero-inflated count data
- Kleinke & Reinecke (2013b) – imputation of two-level Poisson data

Example

Multiple Imputation of zero-inflated and overdispersed count data based on a two-level hurdle NB model: The zero model is a binomial GLM, the count model is a zero-truncated NB model.

```
R> require( "countimp" )
R> require( "glmmADMB" )
R> data( "MZINB.data.Rdata" )
R> ini <- mice( MZINB.data, maxit = 0 )
R> pred <- ini$predictorMatrix
R> pred[1,] <- c( 0, 0, 2, 0, -2, 0, 1 )
R> meth <- ini$method; meth[1] <- "2l.zihnb"
R> imp <- mice( MZINB.data, maxit = 1, method = meth,
    + predictorMatrix = pred, seed = 1234)
R> result <- do.mira( imp, DV = "Y",
    + fixedeff = "X1+Z1",
    + randeff = "X1", fam = "truncnbinom", grp = "GRP",
    +  id = "ID" )
R> summary( result )

Pooled Fixed Effects Coefficients:
                      est  std.err t.value     df p.value
(Intercept).zero   0.0649   0.0755  0.8602   28.9    0.40
X1.zero            0.4429   0.0399 11.1048 2754.8 < 2e-16 ***
Z1.zero            0.1049   0.0733  1.4319   19.5    0.17
(Intercept).count  0.8663   0.0780 11.1049  429.2 < 2e-16 ***
X1.count           0.7959   0.0390 20.3958   77.7 < 2e-16 ***
Z1.count           0.4456   0.0717  6.2133  139.2 5.6e-09 ***
alpha.count        0.9083   0.0615 14.7678   36.5 < 2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
                        lower     upper          r       fminf
(Intercept).zero  -0.08945051 0.2192930 0.59223598 0.41131303
X1.zero            0.36465704 0.5210512 0.03961445 0.03880251
Z1.zero           -0.04817742 0.2580075 0.82803987 0.50160149
(Intercept).count  0.71297064 1.0196306 0.10685837 0.10072310
X1.count           0.71816528 0.8735436 0.29354952 0.24609907
Z1.count           0.30378615 0.5873607 0.20415673 0.18122715
alpha.count        0.78362897 1.0329870 0.49473412 0.36484799

 Pooled Random Effects SD(s):
 (Intercept).zero           X1.zero     Residual.zero
        0.3897336         0.2342796         0.9952241
(Intercept).count          X1.count
        0.4812296         0.1961808

 Pooled Random Effects Correlation(s):
                 (Intercept).zero X1.zero
(Intercept).zero            1.000   0.027
X1.zero                     0.027   1.000
```

The example is explained in detail in the `countimp` user's manual.

Download and Requirements

The `countimp` package and the `countimp` user's manual are available from our website:

`http://www.uni-bielefeld.de/soz/kds/software.html.`

Requires: R (>= 2.15.0), MASS, mice (>= 2.14), aster, pscl, glmmADMB

References

Kleinke, K., de Jong, R., Spiess, M., & Reinecke, J. (2011). *Multiple imputation of incomplete ordinary and overdispersed count data* [unpublished manuscript, available upon request].

Kleinke, K., & Reinecke, J. (2013a). Multiple imputation of incomplete zero-inflated count data. *Statistica Neerlandica*, *67*(3), 311–336. doi: 10.1111/stan.12009

Kleinke, K., & Reinecke, J. (2013b). Multiple imputation of multilevel count data. Manuscript submitted for publication.

Raghunathan, T. E., Lepkowski, J. M., van Hoewyk, J., & Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology*, *27*(1), 85–96.

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

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). MICE: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*(3), 1–67.

Kristian Kleinke & Jost Reinecke – University of Bielefeld, Faculty of Sociology and Centre for Statistics, Postfach 10 01 31, D-33501 Bielefeld
kristian.kleinke@uni-bielefeld.de; jost.reinecke@uni-bielefeld.de