Missing Data Estimation Examples with Categorical Variables

This example extends the multiple logistic regression example from the handout "Logistic Regression." As in the earlier example, I used data from the Late Life Study of Social Exchanges (LLSSE; Sorkin & Rook, 2004) to predict self-reported heart disease. This example uses the report of heart disease from the third wave of the study (`w3hheart`) in order to provide an illustration with more missing values.[1] Predictors include sex (`w1sex`), vigorous physical activity (`w1active`), depression symptomatology from the brief 9-item version (`w1cesd9`; Santor & Coyne, 1997) of the Center for Epidemiologic Studies-Depression scale (Radloff, 1977), and a measure of negative social exchanges (`w1neg`; Newsom, Rook, Nishishiba, Sorkin, & Mahan), which assesses the frequency of interpersonal conflicts. The report of heart disease from the first wave (`w1hheart`) is used as an additional predictor.

R

The first illustration is with the R package called `mice` (multiple imputation with multivariate imputation by chained equation; van Buuren & Groothuis-Oudshoorn, 2011). This package can be used for continuous, binary, ordinal, or nominal categorical outcome variables. I use the `glm` function for the logistic in this example.

```
 > library(psych)
> describe(d)
         vars   n mean   sd median trimmed  mad min max range  skew kurtosis   se
w3hheart    1 739 0.18 0.38   0.00    0.10 0.00   0   1     1  1.69     0.85 0.01
w1hheart    2 905 0.20 0.40   0.00    0.12 0.00   0   1     1  1.52     0.32 0.01
w1sex       3 916 0.62 0.49   1.00    0.65 0.00   0   1     1 -0.49    -1.76 0.02
w1cesd9     4 910 5.10 4.99   4.00    4.28 4.45   0  26    26  1.50     2.30 0.17
w1activ     5 732 1.77 2.23   0.00    1.46 0.00   0   6     6  0.80    -0.93 0.08
w1neg       6 872 0.41 0.57   0.17    0.30 0.25   0   4     4  1.99     5.17 0.02
```

For comparison, rhw following analysis uses listwise deletion, which is the default for logistic regression procedures. 350 cases were eliminated due to missing values on one or more of the variables.

```
> logmod <- glm(w3hheart ~ w1hheart + w1sex + w1activ + w1cesd9 + w1neg, data = d, family = binomial)
> summary(logmod)

Call:
glm(formula = w3hheart ~ w1hheart + w1sex + w1activ + w1cesd9 + 
    w1neg, family = binomial, data = d)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept) -2.48487    0.35423  -7.015  2.3e-12 ***
w1hheart     3.91776    0.32607  12.015  < 2e-16 ***
w1sex       -0.45853    0.32896  -1.394   0.1634    
w1activ     -0.07141    0.07216  -0.990   0.3224    
w1cesd9      0.04934    0.03688   1.338   0.1809    
w1neg       -0.64390    0.34222  -1.882   0.0599 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 499.16  on 565  degrees of freedom
Residual deviance: 293.65  on 560  degrees of freedom
  (350 observations deleted due to missingness)
AIC: 305.65
```

The following analyses use `mice` with 20 imputations, which is often the recommended number; Enders (2022) recommends using as many as 100.

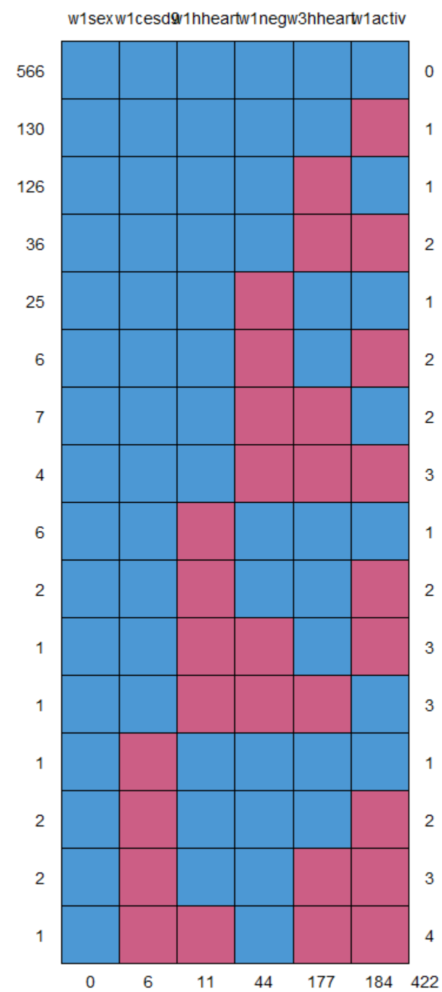
```
>library(mice)
> #I-step: impute 20 data sets (m=20), maxit is maximum iterations, method is type used (pmm is 
predictive mean matching),
> #seed sets a random number generator start for replication
> impdata <- mice(d, m=20, maxit = 50, seed = 500, print=FALSE)


> #look at missing data patterns
> #present data indicated by 1, missing data indicated by 0
> md.pattern(d)
```

[1] This question was not included in the second wave. The study collected data every six months, so the third wave is for the one-year follow-up.

```
    w1sex w1cesd9 w1hheart w1neg w3hheart w1activ
566     1       1        1     1        1       1   0
130     1       1        1     1        1       0   1
126     1       1        1     1        0       1   1
36      1       1        1     1        0       0   2
25      1       1        1     0        1       1   1
6       1       1        1     0        1       0   2
7       1       1        1     0        0       1   2
4       1       1        1     0        0       0   3
6       1       1        0     1        1       1   1
2       1       1        0     1        1       0   2
1       1       1        0     0        1       0   3
1       1       1        0     0        0       1   3
1       1       0        1     1        1       1   1
2       1       0        1     1        1       0   2
2       1       0        1     1        0       0   3
1       1       0        0     1        0       0   4
        0       6       11    44      177     184 422
```

The `md.pattern` function also generates a figure. The variables are listed as columns across the top and each row is a different missing data pattern with the frequency or nonmissing cases for each pattern on the left of the figure and number of missing cases for each pattern on the right. At the bottom is the number of cases missing for each variable total (which is also available from any descriptive analysis).




```
> summary(impdata)
Class: mids
Number of multiple imputations:  20
Imputation methods:
w3hheart w1hheart    w1sex  w1cesd9  w1activ    w1neg
   "pmm"    "pmm"       ""    "pmm"    "pmm"    "pmm"
PredictorMatrix:
         w3hheart w1hheart w1sex w1cesd9 w1activ w1neg
w3hheart        0        1     1       1       1     1
w1hheart        1        0     1       1       1     1
w1sex           1        1     0       1       1     1
w1cesd9         1        1     1       0       1     1
```

```
w1activ         1        1     1       1       0     1
w1neg           1        1     1       1       1     0
```

The following fits the logistic model for each data set and then combines the results.[2]

```
> #build predictive model
> fit <- with(data = impdata, glm(w3hheart == 2 ~ w1hheart + w1sex + w1activ + w1cesd9 + w1neg, family
= binomial))
>
> #combine results of all 20 models
> combine <- pool(fit)
> #this option turns off scientific notation, so p-values are a little easier to read
> options(scipen = 999)
> summary(combine)
         term  estimate std.error statistic      df                                 p.value
1 (Intercept) -2.560051  0.330028  -7.75706 155.041 0.00000000000108459890231318777 1605192
2    w1hheart  3.891440  0.275282  14.13618 242.246 0.00000000000000000000000000000000017268
3       w1sex -0.404729  0.281330  -1.43863 235.855 0.15158220962078478022405647607 22918436
4     w1activ -0.068609  0.073033  -0.93942  93.980 0.34992289992724057157147399266 25043154
5     w1cesd9  0.043377  0.033332   1.30138  81.966 0.19677424462151971162882091448 40948284
6       w1neg -0.520230  0.296883  -1.75231 109.197 0.08252792140033934908949220243 81230585
```

Notice that, compared with the listwise deletion analysis, the statistical conclusions do not change for this particular example, although the p-values have changed in some cases.

Blimp

The Blimp package (Keller & Enders, 2023) is a free standalone package that conducts multiple imputation for a variety of different data types and tests a variety of different models. The algorithmic approach is a bit different from the mice package because it uses an MCMC Bayesian approach. For categorical variables this involves the latent response y^* process briefly described in the “Bayesian Models for Categorical Variables” handout. Thus, burn-in (number of initial iterations ignored) and convergence specifications are involved. The input must be a .csv, a comma separated text file and the missing values must be given a discrete number or NA.[3] Below I test the same logistic model as above. `NIMPS` is the number of imputed data sets (I use 20 here), `OPTIONS: psr` specifies the potential scale reduction factor for checking convergence (values less than 1.05 or 1.1 are considered good; Gelman et al., 2014), `CHAINS` is the number of MCMC chains (best to use a few).

```
DATA: c:/jason/spsswin/cdaclass/heart missing.csv; #input is comma separated
VARIABLES: w3hheart w1hheart w1sex w1cesd9 w1activ w1neg;
ORDINAL: w3hheart w1hheart w1sex ; #identifies binary or ordinal variables;
MISSING: -99; #missing values must be discrete number or NA
MODEL: logit(w3hheart) ~ w1hheart w1sex w1cesd9 w1activ w1neg;
NIMPS: 20;
BURN: 3000;
THIN: 1;
SEED: 90291;
OUTFILE: c:/jason/blimp/heart missing out.csv;
OPTIONS: psr;
ITER: 1000;
CHAINS: 10 processors 2;
```

```
---------------------------------------------------------------------------

                                  Blimp
                                 3.2.22

             Blimp was developed with funding from Institute of
           Education Sciences awards R305D150056 and R305D190002.

            Craig K. Enders, P.I. Email: cenders@psych.ucla.edu
           Brian T. Keller, Co-P.I. Email: btkeller@missouri.edu
                 Han Du, Co-P.I. Email: hdu@psych.ucla.edu
                 Roy Levy, Co-P.I. Email: roy.levy@asu.edu
```

[2] The `==` 2 specification was needed to avoid errors. If reading data in haven, you may need to convert variables to numeric or factor.
[3] My code for creating this data file was:

```
recode w3hheart w1hheart w1sex w1cesd9 w1activ w1neg (missing=-99).

save translate outfile='c:\jason\spsswin\cdaclass\heart missing.csv'
    /type=csv /map /replace
   /keep=w3hheart w1hheart w1sex w1cesd9 w1activ w1neg.
execute.
```

```
              Programming and Blimp Studio by Brian T. Keller

                   There is no expressed license given.
---------------------------------------------------------------------------
ALGORITHMIC OPTIONS SPECIFIED:

  Imputation method:                 Fully Bayesian model-based
  MCMC algorithm:                    Full conditional Metropolis sampler with
                                       Auto-Derived Conditional Distributions
  Default Priors:
    Multilevel covariance matrices:  Not applicable, single-level imputation
    Outcome covariance matrices:     Zero matrix, df = -(p + 1) (PRIOR2)
    Predictor model variances:       Unit sum of squares, df = 2 (XPRIOR1)
  Chain Starting Values:             Random starting values
  NOTE: The default prior for regression coefficients
        in categorical models is 'normal( 0.0, 5.0)'

  NOTE: For complete list of priors include 'pinfo' in the options.

BURN-IN POTENTIAL SCALE REDUCTION (PSR) OUTPUT:

  NOTE: Split chain PSR is being used. This splits each chain's
        iterations to create twice as many chains.

  Comparing iterations across 10 chains    Highest PSR   Parameter #
                          76 to 150           1.123             5
                         151 to 300           1.065             4
                         226 to 450           1.032            23
                         301 to 600           1.033            23
                         376 to 750           1.028            11
                         451 to 900           1.021             5
                         526 to 1050          1.016            11
                         601 to 1200          1.010             1
                         676 to 1350          1.010             5
                         751 to 1500          1.011             1
                         826 to 1650          1.010            36
                         901 to 1800          1.008            15
                         976 to 1950          1.010             5
                        1051 to 2100          1.007             2
                        1126 to 2250          1.007             8
                        1201 to 2400          1.006             5
                        1276 to 2550          1.008             5
                        1351 to 2700          1.006             5
                        1426 to 2850          1.006            18
                        1501 to 3000          1.005            23

METROPOLIS-HASTINGS ACCEPTANCE RATES:

  Chain 1:

    Variable                           Type    Probability   Target Value
    w1hheart                     imputation          0.503          0.500
    w1cesd9                      imputation          0.478          0.500
    w1activ                      imputation          0.501          0.500
    w1neg                        imputation          0.507          0.500

  NOTE: Suppressing printing of 9 chains.
        Use keyword 'tuneinfo' in options to override.

DATA INFORMATION:

  Sample Size:             916
  Missing Data Info:

                     Pattern #:      1     2     3     4     5     6     7     8     9    10
                         miss % ------------------------------------------------------------
               w3hheart = 19.3       -     -     M     M     -     M     -     -     M     -
               w1hheart =  1.2       -     -     -     -     -     -     M     -     -     M
                  w1sex =  0.0       -     -     -     -     -     -     -     -     -     -
                w1cesd9 =  0.7       -     -     -     -     -     -     -     -     -     -
                w1activ = 20.1       -     M     -     M     -     -     -     M     M     M
                  w1neg =  4.8       -     -     -     -     M     M     -     M     M     -
                                ------------------------------------------------------------
                                %  61.8  14.2  13.8   3.9   2.7   0.8   0.7   0.7   0.4   0.2
```

```
  NOTE: Suppressing printing of 6 missing data patterns.
        These patterns consist of 0.9% of the data.
        Use keyword 'allpatterns' in options to print all patterns.

MODEL INFORMATION:

  NUMBER OF PARAMETERS
    Outcome Models:          6
    Predictor Models:        28

MODEL FIT:

  INFORMATION CRITERIA

    Conditional Likelihood
      DIC2                   543.293
      WAIC                   606.116

IMPUTED DATA SUMMARIES:

  NOTE: Imputed data summaries are averaged across 1000 data sets.

Variables                        Mean     StdDev        25%        50%        75%        Min        Max
                         ----------------------------------------------------------------------------

  w3hheart                      0.186      0.389      0.000      0.000      0.000      0.000      1.000
  w1hheart                      0.199      0.400      0.000      0.000      0.000      0.000      1.000
  w1sex                         0.619      0.486      0.000      1.000      1.000      0.000      1.000
  w1cesd9                       5.099      4.985      1.000      4.000      7.000     -2.303     26.000
  w1activ                       1.732      2.231      0.000      0.817      3.673     -4.451      7.621
  w1neg                         0.415      0.571      0.000      0.167      0.667     -0.802      4.000

                         ----------------------------------------------------------------------------

CORRELATIONS AMONG RESIDUALS:

  No residual correlations.

OUTCOME MODEL ESTIMATES:

  Summaries based on 1000 iterations using 10 chains.
  NOTE: Estimate column based on posterior median.

Outcome Variable:  logit(w3hheart)

Parameters                   Estimate     StdDev       2.5%      97.5%       ChiSq     PValue      N_Eff
                         ----------------------------------------------------------------------------

Coefficients:
  Intercept                    -2.563      0.349     -3.287     -1.917      54.830      0.000    166.295
  w1hheart                      3.904      0.288      3.313      4.466     183.167      0.000    215.454
  w1sex                        -0.440      0.277     -0.990      0.102       2.647      0.104    247.750
  w1cesd9                       0.039      0.032     -0.024      0.102       1.400      0.237    167.846
  w1activ                      -0.071      0.069     -0.210      0.068       1.051      0.305    170.392
  w1neg                        -0.468      0.274     -0.992      0.044       3.041      0.081    171.936

Odds Ratio:
  Intercept                     0.077      0.028      0.037      0.147         ---        ---    174.616
  w1hheart                     49.587     15.207     27.472     86.973         ---        ---    236.320
  w1sex                         0.644      0.186      0.372      1.108         ---        ---    265.693
  w1cesd9                       1.040      0.034      0.977      1.107         ---        ---    170.048
  w1activ                       0.932      0.064      0.810      1.070         ---        ---    169.752
  w1neg                         0.626      0.178      0.371      1.045         ---        ---    164.605

Proportion Variance Explained
  by Coefficients               0.151      0.018      0.117      0.190         ---        ---    215.147
  by Residual Variation         0.849      0.018      0.810      0.883         ---        ---    215.147

                         ----------------------------------------------------------------------------
```

References
Enders, C.K. (2022). *Applied missing data analysis, second edition.* New York: Guilford Press.
Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2014). *Bayesian data analysis, third edition.* CRC.
Keller, B. T., & Enders, C. K. (2023). *Blimp user's guide (Version 3)*. Retrieved from www.appliedmissingdata.com/blimp
van Buuren S and Groothuis-Oudshoorn K (2011). mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software, 45(3), 1-67. doi:10.18637/jss.v045.i03