Computer Syntax

SAS code is provided for the logistic regression imputation described in this article. This code is listed in parts, with description provided for each part, however, in a given analysis these parts would be combined into a single syntax file. This syntax file can be obtained at www.uic.edu/∼hedeker/long.html. To start, the code below reads in the dataset used in this article.

```
DATA one; INFILE 'c:\smoke.dat';
INPUT id smk miss smk0 grp;
```

Here, `id` is the subject identifier, `smk` is the smoking status at the final timepoint (0=abstinent, 1=smoking, .=missing), `miss` is the missing indicator (0=observed or 1=missing at the final timepoint), `smk0` is the $t0$ smoking status (0=abstinent, 1=smoking), and `grp` is the grouping variable (0=control, 1=treatment). Uppercase letters are used to designate SAS syntax, and lowercase letters are used to designate user-defined entities.

Multiple Imputation Several variables need to be defined for the subsequent logistic regression multiple imputation model. First, we designate the cell frequencies for the observed cells in the crosstab of `smk0` by `miss` by `smk` (*i.e.*, the cell indicators forming the suffix of the `n` variables reflect the values of these three binary variables in this particular order). Note that these frequencies could be obtained in the present analysis, but, for simplicity, here they have been gotten from a previous analysis of these data and simply typed in. Comments are provided on each line using the /* ... */ convention of SAS.

```
n111 = 42; /* number of abstainers - smk0=abstinent */
n112 = 71; /* number of smokers - smk0=abstinent */
n211 = 36; /* number of abstainers - smk0=smoking */
n212 = 223; /* number of smokers - smk0=smoking */
```

Now, the mean values of the coefficients for the logistic regression are obtained based

on the above frequencies and the assumed level of the odds ratio for missing and smoking.

```
orat = 2;
beta0m = LOG(n112/n111);
beta1m = LOG(orat);
beta2m = LOG(n212/n211);
beta3m = LOG(orat);
seed = 974677743;
```

The regression coefficients `beta0m` and `beta2m` are based on the smoking rates for $t0$ non-smokers and $t0$ smokers, respectively. Similarly, the coefficients `beta1m` and `beta3m` are set depending on the assumed odds ratio of smoking and missing, for $t0$ non-smokers and smokers, respectively. Here, an odds ratio of 2 is assumed for both. The SAS function `LOG` computes the natural logarithm of the argument. The variable `seed`, which is an integer value that starts the subsequent random number functions, is also set.

The cell frequencies for the missing cells are now based on the observed cell frequencies, the numbers of missing subjects (these numbers are simply typed in), and the assumed odds ratio. Again, these are for the crosstab of `smk0` by `miss` by `smk`.

```
n12dot = 37; /* number of missing for smk0=abstinent */
p122 = (orat*(n112/n111))/(1 + (orat*(n112/n111)));
n122 = p122*n12dot;
n121 = (1 - p122)*n12dot;

n22dot = 80; /* number of missing for smk0=smoking */
p222 = (orat*(n212/n211))/(1 + (orat*(n212/n211)));
n222 = p222*n22dot;
n221 = (1 - p222)*n22dot;
```

Based on the cell sample sizes, the variances associated with the regression coefficients are calculated (these formulas can be found in Agresti [2002]). These will be used in the imputation below in order to add sampling variation into the process (*i.e.*, because

the regression coefficients above are estimates of the population parameters, sampling variation needs to be added to these regression coefficients in the imputation).

```
beta0v = (n111 + n112)/(n111*n112);
beta1v = 1/n111 + 1/n112 + 1/n121 + 1/n122;
beta2v = (n211 + n212)/(n211*n212);
beta3v = 1/n211 + 1/n212 + 1/n221 + 1/n222;
```

Imputation is now done for the logistic regression model. As described in the article, it is important to perform this imputation multiple times. The code below can be used for this purpose, in this case it is done 100 times. To get a random draw from a standard logistic distribution, we use the fact that this distribution can be obtained as the natural logarithm of the ratio of two independent standard exponential distributions (see McCullagh and Nelder [1989], page 20). Random draws from standard exponential distributions are obtained using the SAS function `RANEXP`. Random draws from standard normal distributions are obtained using the SAS function `RANNOR`. As described in the article, the Cholesky factorization, or matrix square root, of the variance-covariance matrix associated with the regression coefficients is used, and applied to the standard random normal deviates that are obtained using `RANNOR`. For this, the SAS function `SQRT` is used to give the square root of the indicated argument.

```
DATA sim; SET one;
ARRAY smks(100) smks1-smks100;
DO i = 1 TO 100;
  IF miss EQ 0 THEN smks(i) = smk;
  IF miss EQ 1 THEN DO;
    exp1 = RANEXP(seed); exp2 = RANEXP(seed); stdl = LOG(exp1/exp2);
    ran0 = RANNOR(seed); ran1 = RANNOR(seed);
    /* the next lines incorporate the covariance between beta0 and beta1
    (likewise for beta2 and beta3) using the Cholesky factorization */
    beta0 = beta0m + ran0*SQRT(beta0v);
```

```
    beta1 = beta1m - ran0*SQRT(beta0v) + ran1*SQRT(beta1v - beta0v);
    ran2 = RANNOR(seed); ran3 = RANNOR(seed);
    beta2 = beta2m + ran2*SQRT(beta2v);
    beta3 = beta3m - ran2*SQRT(beta2v) + ran3*SQRT(beta3v - beta2v);
    ystar = (1 - smk0)*(beta0 + beta1*miss) + smk0*(beta2 + beta3*miss) + stdl;
    smks(i) = 0; IF ystar > 0 THEN smks(i) = 1;
  END;
END;
```

Here, a new dataset `sim` is created which will contain 100 smoking variables named `smks1` to `smks100`. A `DO` loop is used to create these 100 variables, and the `ARRAY` statement is used to specify a vector named `smks` containing the 100 smoking repetitions. These are set to the original variable `smk` for observed individuals, and imputed otherwise.

Analysis of Multiply-Imputed Data To analyze the multiply-imputed data, we first have to adjust the data so that they are in the "long" format. Namely, in the file `sim`, which is in the "wide" format, each of the 100 smoking variables are associated with one case, whereas, for the analyses to be performed, each needs to be a separate case, with a variable indicating the imputation number. The code below does this translation, yielding a variable `smki` that is the smoking variable, and the variable `_imputation_` that is the imputation number. These variables, and `grp`, are saved in the dataset `unisim`.

```
DATA unisim (KEEP = id grp smki _imputation_); SET sim;
ARRAY smks(100) smks1-smks100;
DO _imputation_ = 1 TO 100;
  smki = smks(_imputation_);
  OUTPUT;
END;
```

The data are now sorted by `_imputation_`.

```
PROC SORT; BY _imputation_ ;
```

The logistic regression analysis is performed, stratified by `_imputation_` (*i.e.*, performed 100 times), and the results from each analysis are saved in the dataset `outlreg`.

```
PROC LOGISTIC DESCENDING NOPRINT OUTEST=outlreg COVOUT;
MODEL smki = grp / LINK = LOGIT;
BY _imputation_ ;
```

The results corresponding to the regression coefficients (*i.e.*, for `intercept` and `grp`) from the 100 logistic regression analyses are combined using `PROC MIANALYZE`.

```
PROC MIANALYZE DATA=outlreg;
VAR intercept grp;
```

`PROC MIANALYZE` prints out the results of the multiple imputation process for the two logistic regression parameters `intercept` and `grp`. The test statistic and p-value of the latter are reported in Table 2 of the article.

References

A. Agresti. *Categorical Data Analysis, 2nd edition.* Wiley, New York, 2002.

P. McCullagh and J. A. Nelder. *Generalized linear models (2nd edition).* Chapman and Hall, New York, 1989.