LBD Synthesis Procedures

February 4, 2011

Appendix to "Towards Unrestricted Public Use Business Microdata: The Synthetic Longitudinal Business Database" by S. K. Kinney, J. P. Reiter, A. P. Reznek, J. Miranda, R. S. Jarmin, and J, M. Abowd

1 Generic synthesis procedures

1.1 Normal Method

A common approach for generating imputations for continuous variables is to model the posterior distribution using a normal linear regression model, possibly on transformed data. Given the highly skewed nature of payroll and employment data in the LBD, the kernel density estimation procedure of Abowd and Woodcock (2004) is used to transform the response variables so that marginally they approximately follow standard normal distributions, and a normal model can be used.

Using the normal approach with a KDE transform, a synthetic variable $\tilde{y}_k^{(i)}$ is generated from $(X, y_1, \ldots, y_{k-1}, \tilde{y}_1, \ldots, \tilde{y}_{k-1})$ by drawing from the posterior predictive distribution of y_k as follows:

1. Apply the KDE transform to the response variable and any needed tranformation functions to the predictors to satisfy approximately linear regression assumptions. For simplicity, the tranformations performed on the predictors are not notated here, though the models used are given in Section 5.1.2. When the KDE transforms were applied to the linear predictors, the observed-data correlations were not preserved in the synthetic data.

   For each observed value $y_{k,l}, l = 1, \ldots, n$, of response variable y_k, the transformed values $y^*_{k,l}$ are computed as $\Phi^{-1}(\hat{K}(y_{k,l}))$, where Φ denotes the standard normal cumulative distribution function and $\hat{K}(y_{k,l})$ is a kernel density estimate of $y_{k,l}$.

2. Fit a linear model, $f(y_k^*|Z, \beta, \sigma^2) = N(Z\beta, \sigma^2)$ to the observed data, where $Z = (X, y_1, \ldots, y_{k-1})$, and obtain estimates of β and σ^2.

3. For each imputation, draw new values $\tilde{\sigma}^{2(i)}$ and $\tilde{\beta}^{(i)}$ from the posterior distributions $f(\sigma^2|X, Y)$ and $f(\beta|\sigma^2, Y, X)$.

4. Draw $\tilde{y}_k^{(i)*}$ from $N(\tilde{Z}^{(i)}\tilde{\beta}^{(i)}, \tilde{\sigma}^{2(i)})$, where $\tilde{Z}^{(i)} = (X, \tilde{y}_1, \ldots, \tilde{y}_{k-1})$.

5. Apply the inverse KDE transform, $\tilde{y}_{k,l}^{(i)} = \hat{K}^{-1}(\Phi(\tilde{y}_{k,l}^{(i)*}))$, to return to the original scale of y_k.

Step 3 can be considered optional for census data as the parameters are considered to be known and can be computed from the data. Including this step can potentially reduce disclosure risks by increasing between-imputation variance. For speed and simplicity, this step is omitted in the imputation of the LBD. Disclosure risks are still limited by the smoothing in Step 2 and the random draws in Step 4. Similarly, the transformation function applied depends on the data, and thus contains uncertainty when imputing random samples; hence, Abowd and Woodcock (2004) draw a Bayesian bootstrap sample to estimate the transformation in each implicate to account for this additional uncertainty. This step is also skipped in the synthesis of the LBD.

1.2 Nonnormal Models

The normal approach can be modified for nonlinear models by replacing the normal model with a nonlinear one. For binary and categorical responses without very many categories, one can sample from binomial and multinomial distributions, using appropriate generalized linear models to obtain the sampling probabilities.

The synthetic variable $\tilde{y}_k$ for binary response y_k is generated by approximating draws from $f(y_k|X, y_1, \ldots, y_{k-1}, \tilde{y}_1, \ldots, \tilde{y}_{k-1})$ as follows:

1. Use the observed data to fit a logistic model, $\text{logit}(p(y_k = 1)) = Z\beta$, where $Z = (X, y_1, \ldots, y_{k-1})$, to obtain $\hat{p}_l(Z_l), l = 1, \ldots, N$.

2. Update model parameters by taking draws from their posterior distributions. As before, this step is omitted in the LBD synthesis.

3. Use the observed data model to obtain $\hat{p}_l(Z_l^{(i)})$, where $Z_l^{(i)} = (X_l, \tilde{y}_{1,l}^{(i)}, \ldots, \tilde{y}_{k-1,l}^{(i)})$.

4. Obtain $\tilde{y}_k^{(i)}$ by sampling from $Bin(1, \hat{p}_l(\tilde{Z}_l^{(i)})), l = 1, \ldots, N$.

For categorical responses, the same approach can be used, with a generalized logit model used in place of the logistic model to obtain the posterior probabilities, $\hat{p}_{lj}(x_l, y_{1,l}, \ldots, y_{k-1,l}), l = 1, \ldots, n; j = 1, \ldots, c$, where c is the number of categories in the response. A multinomial distribution is used in place of the binomial.

1.3 Dirichlet-multinomial method

When there are many categories in the response, and many categorical predictors, the generalized logit model can become computationally infeasible. The simpler and faster Dirichlet-multinomial approach provides a convenient framework for sampling from the posterior predictive distribution for a categorical y_k when the predictors in $X, y_1, \ldots, y_{k-1}$ are categorical.

Let c be the number of categories in the response y_k. Let l be the number of unique categories determined by the predictors in $X, y_1, \ldots, y_{k-1}$. Assuming a flat prior on the cell probabilities, $\tilde{y_k}$ is generated as follows:

1. Use the observed data to determine the cell counts $n_j = (n_j^1, \ldots, n_j^c)$, $j = 1, \ldots, l$.

2. Draw new values of the cell probabilities $p_j = (p_j^1, \ldots, p_j^c)$ from a Dirichlet(n_j).

3. For each unit in the synthetic data, look up the appropriate cell probabilities p_j based on the values of X and $\tilde{y}_1, \ldots, \tilde{y}_k$.

4. Sample from a multinomial distribution with cell probabilities p_j.

As in the previous methods, we skip Step 2. In Step 3, if an exact cell match is not found in the observed data, a possibility depending on the disclosure control applied to $y_1, \ldots, y_{k-1}$, then the cell is collapsed untill a match is found. Hence, in Step 1, cell counts must be determined for one or more sets of broader categories as well.

This approach is very fast computationally and appears to yield good predictions with sufficient disclosure control when used in the LBD synthesis. With sufficient variability in the observed data, disclosure control is provided by sampling from the multinomial distribution and by the disclosure control methods applied to any predictors. In some cases, this method can fail to provide sufficient disclosure protection. When there are a large number of categories and categorical variables, numerous units are uniquely determined by their values of the categorical predictors, yielding predictions that are "too good." For example, let C be a unique category determined by categorical predictors in X and let y_C be the observed values of a categorical response variable corresponding to the n_C units in C. If $n_C = 1$, or $y_{Ci}, i = 1, \ldots, n_C$ all have the same value, then a categorical model will impute

synthetic values $\tilde{y}_C$ for y_C such that $\tilde{y}_C = y_C$ in each implicate. This creates a high risk of re-identification of y_C.

Disclosure control in this case is improved by using an informative prior distribution to add a positive probability that, for a given category C, the $\tilde{y}_C$ generated may contain values not present in y_C. The prior is estimated by replacing one of the categorical predictors with a coarsened version and using this to determine the prior cell counts. For example, if County is a predictor, the prior could be obtained from state-level cell counts. The prior counts are normalized to represent a small number of units to reduce the sensitivity to the prior; however these can be scaled to increase the noise if desired. This serves to add noise in a controlled fashion, meeting the goal of reducing disclosure risks with minimal loss of utility.

Let c be the number of categories in the response y_1. Let l be the number of unique categories determined by the predictors in X and let p be the number of unique categories in a coarsened version of X, X_p, i.e., with one or more of the predictors dropped or coarsened, so that $l > p$ and X_p has fewer categories than X and larger cell counts. Generate draws from $f(y_1|X)$ as follows:

1. Using the observed data, determine the cell counts $n_k = (n_k^1, \ldots, n_k^c)$, $k = 1, \ldots, p$. Normalize these cell counts so that $\sum_{i=1}^{p} n_k^i = a$, where a is a small number. Larger values will give more weight to the prior.

2. Using the observed data, determine the cell counts $n_j = (n_j^1, \ldots, n_j^c)$, $j = 1, \ldots, l$. Add each n_j to its corresponding normalized n_k to obtain the posterior counts $m_j, j = 1, \ldots, l$.

3. Draw new values of the cell probabilities $p_j = (p_j^1, \ldots, p_j^c)$ from a Dirichlet$(m_j = (m_j^1, \ldots, m_j^c))$.

4. For each unit in the synthetic data, look up the appropriate cell probabilities p_j based on the values of X.

5. Sample from a multinomial distribution with cell probabilities p_j.

As before, for the LBD, we omit drawing parameters in Step 3, and sample from a multinomial distribution with cell probabilities given by $m_j / \sum_{i=1}^{c} m_j^i$.

2 Unit information prior

The normal method is modified for small subgroups by incorporating an informative prior for the vector of regression coefficients β. For a given 3-digit SIC group, a comparable subgroup

is found in the corresponding 2-digit SIC group, which is used to estimate the prior, and 4-digit SIC is dropped from the imputation model. For example, if there are too few single-unit nonzero births in a given year in the 3-digit SIC group being imputed, the prior is estimated from all single-unit nonzero births in the same year from the corresponding 2-digit SIC group. This is analogous to the common practice of using information from previous experiments, external surveys, and censuses to determine prior values.

Using a unit information prior allows for all of the available data to be used to estimate a prior mean and variance for the regression coefficient without overwhelming the data. The unit information prior has the same amount of information about β as contained in a single observation. In this case the information is in the sample used to estimate the prior, which has the form

$$\begin{aligned} p(\beta|\sigma^2) &= N(\beta_0, \sigma^2\Sigma_0) \\ p(\sigma^2) &= \chi^{-2}(n_0 - k, s_0^2) \end{aligned}$$

where $\beta_0 = (X_0'X_0)^{-1}X_0'Y_0$, $\Sigma_0 = n_0(X_0'X_0)^{-1}$, X_0 and Y_0 are the prior data for X and Y, s_0^2 is the sample variance $(Y_0 - X_0\beta_0)'(Y_0 - X_0\beta_0)/(n_0 - k)$, and n_0 is the prior data sample size.

The resulting posterior $(\beta, \sigma^2|Y, X)$, used to draw from the posterior predictive distribution, is given by

$$\begin{aligned} p(\beta|\sigma^2, Y, X) &= N(\hat{\beta}, \hat{\Sigma}) \\ p(\sigma^2|Y, X) &= \chi^{-2}(n + n_0 - k, s^2) \end{aligned}$$

where $\hat{\beta} = \hat{\Sigma}(\Sigma_0^{-1}\beta_0 + X'Y)$, $\hat{\Sigma} = (\Sigma_0^{-1} + X'X)^{-1}$, and $s^2 = \{(n_0 - k)s_0^2 + (y - X\hat{\beta})'(y - X\hat{\beta}) + (\hat{\beta} - \beta_0)\Sigma_0^{-1}(\hat{\beta} - \beta_0)\}/(n + n_0 - k)$.

When $n_0 \geq k$, this gives a full-rank model for drawing from the posterior predictive distribution under an informative prior. In addition to providing a full-rank model for small subgroups where $n < k$, this provides a degree of disclosure protection by using information from external data to build the model. If additional noise is desired, more weight can be given to the prior by replacing n_0 in the prior specification with $n_p < n_0$. If $n_0 < k$, predictors may be dropped to obtain a full-rank model, or the group used to estimate the prior may be expanded.

References

Abowd, J. M. and Woodcock, S. D. (2004), "Multiply-Imputing Confidential Characteristics and File Links in Longitudinal Linked Data," in *Privacy in Statistical Databases*, eds. Domingo-Ferrer, J. and Torra, V., New York: Springer-Verlag, pp. 290–297.