Multiple Imputation of Missing Blood Alcohol Concentration (BAC) Values in FARS

Rajesh Subramanian and Dennis Utter

National Highway Traffic Safety Administration, 400, 7th Street, S.W., Room 6201, Washington, DC 20590
rsubra@nhtsa.dot.gov, dutter@nhtsa.dot.gov

Introduction

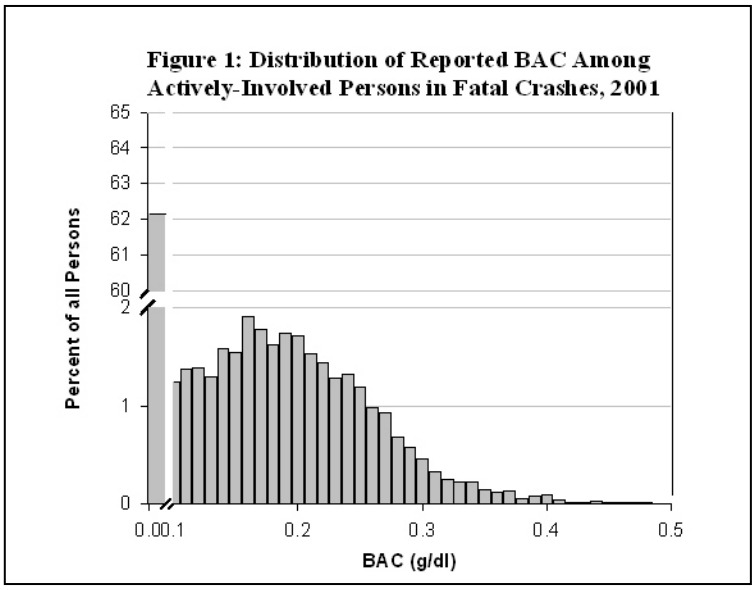
Alcohol involvement is a major contributing factor in the occurrence of motor vehicle traffic crashes. According to NHTSA's preliminary estimate for 2002, alcohol was involved in about 42 percent of all motor vehicle crashes where there was a fatality. The most direct measure of a driver's or nonoccupant's alcohol involvement is a BAC test result reported in NHTSA's Fatality Analysis Reporting System (FARS). These results are based on a variety of sources like breath-tests administered by police or a toxicology test from the Medical Examiner's Office. BAC is the grams of alcohol in a deciliter of blood and can have a plausible value between 0 and 0.94. However, in FARS, BAC results are not known for many of the persons involved in the fatal crashes. The significant number of missing BAC values (about 58 percent in 2001) greatly inhibits the ability to report the extent of alcohol involvement, to identify groups for targeting campaigns to reduce impaired driving, and to evaluate the effectiveness of existing impaired-driving programs.

In order to remedy the missing data problem, NHTSA has employed *Multiple Imputation (MI)* to estimate missing BACs in FARS. MI imputes ten values for each missing BAC value in FARS. NHTSA transitioned to MI in 2002 and has revised historical estimates of alcohol involvement back to the crash data for 1982 in order to provide a consistent database of alcohol estimates for trend-analysis, etc.

BAC in FARS

NHTSA's Fatality Analysis Reporting System (FARS) is an interwoven hierarchical dataset containing detailed information on all motor vehicle crashes where there was a fatality and all vehicles and persons involved in those crashes. Primary interest lies in BAC values for "actively involved persons", which comprise the drivers of vehicles and of any nonoccupants (pedestrians and pedalcyclists). Figure 1 depicts the distribution of BAC among actively-involved persons as reported to FARS in 2001.

As seen in Figure 1, the distribution of BAC may be regarded as *semicontinuous*; a substantial proportion of BAC values are zero, and the remaining responses are continuously distributed over the positive real number line within the plausible range (0 to 0.94) although responses above 0.4 are sparse.



Multiple Imputation

Multiple Imputation (Rubin, 1987; Schafer 1997) is a simulation-based approach to missing data in which each missing data is replaced by several plausible values drawn randomly from a probability distribution, reflecting the uncertainty with which the missing values can be predicted from the observed data. Each missing response is replaced by multiple simulated values. The multiple imputations, together with the non-missing responses, produce multiple complete versions of the variable, each of which may

be analyzed by standard complete-data techniques. Results from analyzing the ten versions will vary somewhat, and this variation is used to estimate the extra uncertainty in statistical summaries due to missing data.

MI and FARS

The imputation strategy to estimate missing BAC in FARS uses the actively-involved person as the basic unit of analysis and statistical models are constructed to predict actively involved persons' BAC from other available covariates. Some of these covariates are characteristics of the crash itself and other covariates include characteristics of the person (age, gender, use of a safety-belt, etc.) and the type of vehicle being driven. Rates of alcohol involvement vary widely by vehicle class; for example, operators of motorcycles are far more likely to have positive alcohol as compared to drivers of Large Trucks. Aside from the type of the vehicle being driven, the most powerful predictor of BAC was the variable **DRINKING**, which records the opinion of law enforcement officials at the scene as to whether alcohol may have been involved.

Challenges in Developing Imputation Strategy

Semicontinuous Nature of BAC
A significant proportion of BACs are clustered around 0 and the positive responses are distributed over the plausible range.

Algorithms (Schafer, 1997) for imputation under the *General Location Model (GLOM)* were found to be useful for imputing missing BAC. The semicontinuous BAC was re-expressed as two variables: a **dichotomous** or binary indicator (BAC2) expressed as:

BAC2=1 if BAC=0
BAC2=2 if BAC>0,

and a **continuous** variable indicating the actual level of BAC, conditional on BAC>0 (when BAC=0, the continuous variable is undefined and may be regarded as "missing"). Recoding BAC as two variables made it possible to model the relationship between BAC and other covariates using a GLOM and impute the missing BAC in a straightforward way.

Missing DRINKING Values
Police-reported DRINKING is missing for many actively involved persons. The procedures in GLOM assume that non-response is ignorable (Rubin, 1976 or Little and Rubin, 1987) in the sense that the probability that a data value is missing does not depend on that value (although it may depend on other variables that are reported). For DRINKING, the meaning of non-response varied significantly from state to state.

Table 1: Covariates used in predicting missing BAC in FARS

Covariate	Description
DRINKING	Police reported Drinking
AGE	Age Category
GENDER	Male/Female
RESTR	Use of Safety-belt/Helmets
SEV	Fatal/Survived
LSTAT	Valid/Invalid License
DRREC	Prior Traffic Convictions
DAY	Day of the Week
HOUR	Time of the Day
ROLE	Striking/Struck Vehicle
RDWY	On/Off Roadway

Sometimes, a missing value probably indicated "no alcohol"; the field in the crash report was left blank because there were no indications of alcohol involvement present. In other cases, the field may have been left blank for "policy" reasons. A method to impute missing BAC that assumed ignorable non-response for DRINKING might have introduced serious biases into estimates of alcohol involvement, particularly at the state level. To address this problem, DRINKING was treated as a fully observed three-level covariate, with "missing" regarded as a substantive category. This treatment, though not fully satisfactory, is consistent with the earlier modeling approach to estimate missing BAC (Klein, 1986). A better solution would have been to develop a plausible probability model for the non-response that includes interactions between DRINKING and state. Developing and fitting such a model would have been a substantial task and could change over time.

Implementing MI in FARS

The GLOM at the heart of the multiple imputation procedure is a multivariate statistical model describing the entire joint distribution of BAC, DRINKING and other significant predictors [Table 1] within each vehicle class. GLOM specifies a joint probability distribution for all the covariates at once. The GLOM is most easily understood as a two-stage model:

First Stage:
A dichotomized version of BAC (i.e., a binary indicator for BAC>0 versus BAC=0) is related to categorical covariates by a conventional loglinear model for cross-classified categorical data. The model is fitted for each vehicle class and its purpose is to capture essential relationships between BAC2 and the other covariates. If the other covariates had no missing values, then this first-stage model could be regarded simply as a logistic regression for predicting dichotomized BAC. The fact that covariates are sometimes missing, however, makes it necessary to model their full joint distribution at this stage. Capitalizing on the well-known relationship between logistic regression and loglinear models, a simple association between dichotomized BAC and each covariate was examined. This model is selected by an automated stepwise procedure beginning with a null model of no predictors. At each step, the significance of each term not in the model is tested. The most significant term is entered into the model, provided it is significant at the 0.1 level by a deviance (likelihood-ratio) test. After it is entered, the significance of each term currently in the model is tested, and any term that is no longer significant at the 0.1 level is discarded. This discarding is performed one term at a time, beginning with the least significant term. The whole process is repeated until there are no more terms outside of the model that are significant at the 0.1 level, and every term in the model is significant at the 0.1 level.

Second Stage:
The second-stage model is a normal linear regression for predicting the actual level of BAC among the cases for which BAC is positive. It would have been very convenient to fit a linear model to *log(BAC)*, because the logarithmic transformation maps the positive real numbers to the entire real line; a linear regression on the log scale would never predict a negative value of BAC. Unfortunately, for many vehicle classes, *log (BAC)* was negatively skewed. Preliminary analyses showed that normal linear regression models for *log(BAC)* could impute implausibly high values of BAC.

Power Transformation

Power transformation of the form

$$\log(BAC)^{\lambda}; \lambda \geq 2$$

gave better results, but a value of λ that worked well for one vehicle class did not work well for another. An automatic procedure based on the maximum-likelihood method of Box and Cox was devised to find the power transformation *g(BAC)* that makes $log(BAC)^{\lambda}$ most nearly normal. The resulting Maximum Likelihood (ML) estimate tended to work well for many vehicle classes, but still produced implausible BAC values for other vehicle classes. Adding 1 to the ML estimate, however, appeared to solve the problem. The automatic transformation procedure proceeds as follows:

- The Box-Cox estimate is found by a grid search over the values 0.1, 0.2,…, 4.5
- The positive values are transformed :

$$g(BAC) = \log(BAC)^{\lambda+1}$$

After imputation, the imputed values are transformed back to the original BAC scale using the back transformation g^{-1}.

After an appropriate transformation is selected, a set of covariates is chosen to serve as linear predictors in the second-stage regression model. All covariates in the first-stage loglinear model, with the exception of dichotomized BAC , are eligible for inclusion in the second stage. From this pool, a subset of significant predictors is chosen by ordinary least-square stepwise regression of *g(BAC)*.

Imputation

Once the first and second-stage covariates have been selected, multiple imputations of missing BAC are created under GLOM. First ML estimates of the model parameters are found using an ECM algorithm (Schafer, 1997). Using these ML estimates as starting values, new parameters are simulated from their posterior distribution by a Markov-Chain Monte Carlo (MCMC) algorithm. Usually, the number of steps required for ECM to converge is a conservative estimate of the number of steps required by the MCMC to achieve approximate stationarity, especially if the chain is started at the Maximum Likelihood Estimate (MLE). Beginning at the MLE, the chain is allowed to run for this many steps and the missing data are imputed under simulated values of the parameters. Repeating this value ten times results in ten imputations of the missing BACs. The imputed values of *g(BAC)* are then transformed back to the BAC scale.

Analyzing the Multiply-Imputed Data

When assessing the extent of alcohol involvement in traffic crashes, the quantity of interest is usually the proportion of a population that shows the involvement of alcohol (e.g., percent of drivers killed that were intoxicated, percent of fatally injured nonoccupants, etc). This proportion is the percentage of the standard population of the stratum of interest that has alcohol involvement. Alcohol involvement is determined jointly from the known set of alcohol test results as well as the imputed values for unknown BAC. Under multiple imputation, each missing BAC value is replaced by ten imputed values. In order to estimate population proportions, the results (proportions) from each of the ten sets of values have to be combined by standard computational macros.

Rubin's method of scalar estimands (Rubin, [6]) is used to estimate quantities of interest. Let Q be a one-dimensional quantity of interest – a *proportion* of crashes or persons that showed a positive alcohol test result in a universe of crashes or people or a *coefficient* from a linear or logistic regression model. The goal is to find a confidence interval or test a hypothesis about Q. Let $\boldsymbol{Y}$ denote the data from FARS that are necessary to estimate Q. $\boldsymbol{Y}$ is partitioned into observed and missing parts,

$$Y = (Y_{obs}, Y_{mis})$$

where Y_{obs} is known and Y_{mis} is unknown and has been multiply-imputed. Let $\hat{Q}$ be the complete-data point estimate for Q, the estimate to be used if no data were missing. Let U be the variance estimate associated with $\hat{Q}$, so that $\sqrt{U}$ is the complete-data standard error. As U and $\hat{Q}$ are both functions of $Y = (Y_{obs}, Y_{mis})$, they may be rewritten as $\hat{Q}(Y_{obs}, Y_{mis})$ and $U(Y_{obs}, Y_{mis})$, respectively. Multiple Imputation inference assumes that the complete data problem is sufficiently regular and sample size sufficiently large for the asymptotic normal approximation

$$U^{-1/2}(Q - \hat{Q}) \sim N(0,1)$$

to work well. With m imputations, m different versions of $\hat{Q}$ and U can be calculated.

Let

$$\hat{Q}^{(t)} = \hat{Q}(Y_{obs}, Y_{mis}^{(t)})$$

and

$$U^{(t)} = U(Y_{obs}, Y_{mis}^{(t)})$$

be the point and variance estimates using the t-*th* set of imputed data, t=1,2,...,10. The multiple imputation point-estimate for Q is simply the average of the complete-data point estimates.

$$\overline{Q} = \frac{1}{10}\sum_{i=1}^{10} \hat{Q}^{(t)}$$

$\overline{Q}$ is the final quantity of interest, for example, the proportion of drivers involved in fatal crashes whose BAC was .01 or above.

The variance estimate associated with $\overline{Q}$ has two components. The *within-imputation* variance is the average of the complete-data variance estimates,

$$\overline{U} = \frac{1}{10}\sum_{t=1}^{10} U^{(t)}$$

and the *between-imputation* variance is the variance of the complete-data point estimates,

$$B = \frac{1}{9}\sum_{t=1}^{10} (\hat{Q}^{(t)} - \overline{Q})^2$$

The *total-variance* is defined as

$$T = \overline{U} + (1 + m^{-1})B$$

Validation

Validation tests were conducted to ensure that the multiple imputation procedure produced plausible estimates of alcohol involvement. The most convincing evidence that multiple imputation performed properly came from an experiment in which multiple imputations were created for 'known' values of BAC in the FARS files. A set of all crash records with known BAC values was extracted from the FARS files. Twenty-five percent of these records were randomly sampled and their BAC values were intentionally set to 'missing'. BAC values for these records were then estimated using the multiple imputation procedure, and the results were compared to the original 'known' BAC values as shown in Table 2.

Table 2: Validation Tests: Extent of Non-Sober Drivers (BAC=0.01+) Computed from all Drivers with Known BAC Results, and Computed from Imputing for 25 Percent of these Known Results Randomly set to Missing

Year	Known	MI
1982	64%	63%
1986	57%	56%
1990	51%	51%
1993	46%	46%
1995	44%	44%

If this experiment were replicated a large number of times, it would be possible to conduct formal tests of unbiasedness of the imputation method under this completely random missingness mechanism. However, the value of such tests would be dubious, because the nonresponse of BAC in FARS is not completely at random. There is strong evidence that missing BAC in FARS are more likely to be zero than are the observed values, because of the relationships between missingness and many covariates that are strongly related to BAC. Nevertheless, the data in Table 2 do suggest that the GLOM that underlies the multiple imputation procedure is capable of preserving essential features of the BAC distribution, both in a marginal sense and conditionally upon important covariates.

Implementing MI

Estimates from Multiple Imputation replaced those from a prior methodology (Klein, 1986), beginning with the 2001 data year. However, NHTSA frequently reports alcohol involvement going back to the 1982 data year. The multiple imputation procedure was hence applied back to the 1982 data in order to provide consistent datasets for trend analyses and reporting.

Conclusion

The multiply imputed estimates of missing BAC represent a substantial improvement over prior methods to estimate missing BAC. The new procedure facilitates a wider variety of analyses as compared to prior imputation methodologies.

References

Klein, T.M. (1986) *A Method for estimating posterior BAC distributions for persons involved in fatal traffic accidents*, Report DOT HS 807 094, NHTSA, USDOT.

Rubin, D.B., Schafer, J.L. and Subramanian, R. (1998) *Multiple Imputation of Missing BAC. Values in FARS* Report DOT HS 808 816, NHTSA, USDOT.

Klein, T.M. (1986) *A Method for estimating posterior BAC distributions for persons involved in fatal traffic accidents*, Report DOT HS 807 094, NHTSA, USDOT.

Schafer, J.L. (1997) *Analysis of Incomplete Multivariate Data.* Chapman & Hall, London.