U.S. Department
of Transportation
**National Highway
Traffic Safety
Administration**

NHTSA
People Saving People
http://www.nhtsa.dot.gov

DOT HS 808 816

October 1998

Technical Report

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

Technical Report Documentation Page

1. Report No.	2. Government Accession No.	3. Recipients's Catalog No.
DOT HS 808 816		

4. Title and Subtitle	5. Report Date
Multiple Imputation of Missing Blood Alcohol Concentration (BAC) values in FARS	October 1998
	6. Performing Organization Code NRD-31

7. Author(s)	8. Performing Organization Report No.
Dr. Donald B. Rubin, Dr. Joseph L. Schafer and Rajesh Subramanian	

9. Performing Organization Name and Address	10. Work Unit No. (TRAIS)n code
Rainbow Technology, Inc. 17106 Thatcher Court Olney, MD 20832	
	11. Contract or Grant No. DTNH22-94-D-07005

12. Sponsoring Agency Name and Address	13. Type of Report and Period Covered
Mathematical Analysis Division; National Center For Statistics and Analysis; Research and Development; National Highway Traffic Safety Administration; U.S. Department of Transportation 400 7th Street, S.W. Washington, D.C. 20590	NHTSA Technical Report
	14. Sponsoring Agency Code

15. Supplementary Notes

Dr. Donald B. Rubin of Harvard University, Dr. Joseph L. Schafer of Pennsylvania State University and Rajesh Subramanian, co-authors of report, are NCSA contractors employed by Rainbow Technology, Inc.

16. Abstract

NHTSA has undertaken several approaches to remedy the problem of missing blood alcohol test results in FARS. The approach currently in use employs a linear discriminant model that estimates the probability that a driver or nonoccupant has a BAC in grams per deciliter (g/dl) of 0.00, 0.01 to 0.09 or 0.10 and greater. The estimates are generated only for drivers and nonoccupants (pedestrians, pedalcylists) for whom alcohol test results were not reported.

The proposed methodology extends the current model by simulating specific values of BAC across the full range of possible values rather than estimating probabilities. By imputing ten values of BAC for each missing value, valid statistical inferences like variance, confidence intervals and deviation tests can be drawn. The estimation of discrete values also facilitates analysis by nonstandard boundaries of alcohol involvement (e.g., 0.08+).

17. Key Words	18. Distribution Statement
Blood Alcohol Concentration (BAC); Fatality Analysis Reporting System (FARS); Multiple Imputation; alcohol involvement	Document is available to the public through the National Technical Information Service, Springfield, VA 22161

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No of Pages	22. Price
Unclassified	Unclassified	34	

Form DOT F1700.7 (8-72)

List of Exhibits

Table of Contents

Executive Summary

The objective of this report is to present a new methodology that creates estimates of alcohol involvement for cases with missing Blood Alcohol Concentration (BAC) in the Fatality Analysis Reporting System (FARS). Missing data can result for a number of reasons, the most frequent of which is that persons are not always tested for alcohol. The degree of testing for drivers or nonoccupants varies among states. In 1996, the differences in testing between the states ranged from a low of 15 percent known BACs to a high of 75 percent known BACs.

NHTSA has undertaken several approaches to remedy the problem of missing blood alcohol test results in FARS. The approach currently in use employs a linear discriminant model that estimates the probability that a driver or nonoccupant has a BAC in grams per deciliter (g/dl) of 0.00, 0.01 to 0.09 or 0.10 and greater. The estimates are generated only for drivers and nonoccupants (pedestrians, pedalcylists) for whom alcohol test results were not reported.

The proposed methodology extends the current model by simulating specific values of BAC across the full range of possible values rather than estimating probabilities. By imputing ten values of BAC for each missing value, valid statistical inferences like variance, confidence intervals and deviation tests can be drawn. The estimation of discrete values also facilitates analysis by nonstandard boundaries of alcohol involvement (e.g., 0.08+).

This report documents several validation analyses by comparing the results from the existing discriminant model and the multiple imputation model across traditional categories of age, gender, time of the day, day of the week and crash type. The estimates presented include a mix of both known and estimated BACs.

Because of the strong similarities between the two methods, they are expected to produce similar results. In particular, it was expected that estimates of the rates of alcohol involvement (BAC>0) and rates in excess of the typical legal limit (BAC>.10) within important subclasses would be similar. In many respects, the pattern of results was indeed similar. One important difference, however, was that under the new method the estimated rates of alcohol involvement (BAC>0) were generally higher. Positive differences of about 1-2% in the rate of alcohol involvement appeared consistently across most vehicle classes and demographic subgroups, and across classifications of crashes by time of day and day of week. Differences in rates across subgroups, and trends in rates across time, were quite similar under the two methods as shown in the exhibit below. The only major difference was that the baseline rates of alcohol involvement were slightly but consistently higher under the new method.

Trend of Percentage Alcohol Involvement in Fatal Crashes

Methods	1982	1986	1990	1993	1995	1996	1997
Proposed	59.1	53.8	50.7	44.6	42.5	42.2	39.6
Current	56.7	51.7	49.4	43.5	41.3	40.9	38.5

1. Introduction

The purpose of this project was to create multiple imputations of missing values of Blood Alcohol Concentration (BAC) in the Fatality Analysis Reporting System (FARS) maintained by the National Highway Traffic Safety Administration (NHTSA).

Alcohol involvement is a major contributing factor in the occurrence of traffic crashes. Alcohol has been found to be more prevalent in fatal crashes than in personal injury and property-damage-only crashes. FARS collects information on all fatal motor vehicle crashes that occur on public roads, if the fatality occurs within 30 days of the date of the crash. The most direct measure of a driver's or a nonoccupant's (pedestrian's or pedalcyclist's) alcohol involvement is a known BAC test result, either based on breath tests administered by police, or blood tests. BAC results for many drivers and nonoccupants involved in fatal crashes are not known. The significant number of missing BAC values greatly inhibits the ability to describe the extent and trends of alcohol involvement in fatal crashes, to identify high-risk groups and times for targeting countermeasures, and to evaluate the effectiveness of anti-drunk driving programs.

NHTSA has undertaken several approaches to remedying the missing data problem. The most recent approach, and the one currently in use, is described in the report "*A Method for Estimating Posterior BAC Distributions for Persons Involved in Fatal Traffic Accidents* " (Klein, 1986). This method employs 3-level linear discriminant models to estimate the probability that a particular driver or nonoccupant has a BAC in grams per deciliter (g/dl) of 0.00 (no alcohol), 0.01-0.09 (some alcohol) or 0.10 and greater (generally considered legally intoxicated in most states). A limitation of this approach is that it does not support analyses regarding other classifications of BAC (e.g. 0.08+ or 0.20+).

The present project develops a new method for addressing missing values of BAC. The new method extends the approach of Klein (1986) in two important ways. First, rather than estimating probabilities that the unknown values of BAC fall into a small number of distinct categories, we now impute specific values of BAC across the full range of possible responses. The distribution of BAC may be regarded as *semicontinuous;* a substantial proportion of BAC values are zero, and the remaining responses can be modeled as continuously distributed over the positive real number line (although in practice BAC values are recorded only to two decimal places). The system of Klein (1986) is based on a coarsened version of Blood Alcohol Concentration which we shall call "BAC3" and which takes possible values 1 (when BAC=0), 2 (when 0<BAC<.10) and 3 (when BAC=.10 or greater). The extension to imputing actual values of BAC, rather than probabilities for BAC3, facilitates a greater variety of statistical analyses. Under the new system, for example, one can easily examine boundaries other than 0.10 (e.g. 0.05) to exploring possible consequences of different standards for driving under the influence of alcohol.

The second important extension to Klein's (1986) methodology is the use of multiple imputation. Multiple imputation (Rubin, 1987; Schafer, 1997) is a simulation-based approach to missing data in which each missing value 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 BAC result is replaced by ten simulated values. The ten imputations, together with the non-missing BAC values, produce ten apparently complete versions of BAC, 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. In particular, the ten sets of answers are combined with simple computational macros implementing rules given by Rubin (1987). Combining the ten answers according to these special rules produces statistical inferences that are valid (i.e., estimates of parameters that are consistent, nominal 95% confidence intervals that are in fact 95% confidence intervals, nominal 5% tests that are in fact 5% tests, and so on) under quite general conditions. Single imputation, even if done properly to allow consistent estimation of parameters, uniformly underestimates variability because one imputed value cannot possibly represent uncertainty.

Klein's (1986) estimated probabilities for BAC3, although limiting in ways discussed earlier, are preferable to single imputation because the probabilities do represent uncertainty about which category is correct. This approach was reasonable and appropriate, considering the computational and theoretical constraints of that era. Nevertheless, estimated probabilities for BAC3 are unattractive for many complete-data analyses, especially those pertaining to other categorizations of BAC. Moreover, they do not reflect any uncertainty about the estimated parameters in the discriminant models used to compute the probabilities. Multiple imputation of semicontinuous BAC addresses both of these issues simultaneously.

Aside from these two important issues, however, the new approach bears strong similarities to Klein's (1986) method. The explorations and computational work throughout this project consistently supported the wisdom of Klein's basic modeling approach. Some additional extensions (described below) that were originally planned were dropped because of their unforseen complexity and limited anticipated payoff in improving data quality.

2. Exploratory work and the development of an imputation strategy

FARS is an interwoven hierarchical dataset containing detailed information on fatal crashes and all vehicles and persons involved. Primary interest lies in BAC values for "actively involved persons," which include the drivers of vehicles and of any nonoccupants (e.g. pedestrians). Following the approach of Klein (1986), the actively involved person is used (rather than the crash or vehicle) as the basic unit of analysis, and construct statistical models to predict actively involved persons' BAC from other available covariates. Some of these covariates are characteristics of the crash itself, such as the day of the week and time of the crash, and the location of the crash in relation to the roadway. Other covariates include characteristics of the person (age, gender, possession of a valid driver's license, use of a seat belt, etc.) and the type of vehicle being driven, if any. In earlier work, Klein (1986) found that rates of alcohol involvement varied widely by vehicle class; for example, drivers of motorcycles were far more likely to have positive levels of BAC than drivers of medium and heavy trucks. Aside from vehicle class, the most powerful predictor of BAC was a variable called DRINKING, which records the opinion of law enforcement officials at the scene as to whether

alcohol may have been involved. Indeed, the initial exploratory work with 1993 FARS data confirmed the presence of strong relationships between BAC, vehicle class, and DRINKING. Building a useful statistical model to predict BAC from vehicle class, DRINKING, and other covariates requires special care, however, because of the semicontinuous nature of BAC, and because some of the covariates (especially DRINKING) are missing for many actively involved persons.

Over the last decade, dramatic advances have been made in the theory and practice of statistical analyses with missing data. New computational methods described by Schafer (1997) make it possible to generate multiple imputations of missing values in complex multivariate settings. In particular, chapter 9 of Schafer's (1997) book presents algorithms for imputation under the *general location model* (GLOM), a multivariate probability model for datasets containing both continuous and categorical variables. These algorithms have been implemented by Schafer in a free software library called MIX, which operates within the commercial statistical package S-PLUS (distributed by Mathsoft, Inc.). Early in this project, it became apparent that some version of the GLOM could be useful for imputing missing values of BAC. The semicontinuous BAC may be reexpressed as two variables: a dichotomous or binary indicator equal to 1 if BAC=0 and 2 if BAC>0, and a continuous variable indicating the actual level of BAC if BAC>0. (When BAC=0, the continuous variable is undefined and may be regarded as "missing.") By recoding BAC as two variables, it becomes possible to model the relationships between BAC and other covariates using a GLOM and impute the missing BAC values in a straightforward way.

Semicontinuous variables are not unique to this study; they arise in many areas of applied statistics. In econometric studies, for example, semicontinuous variables are often described by two-stage regression models (e.g., Manning et al., 1987). The first stage is a logit or probit regression to predict the probability of a nonzero response, and the second stage is a linear regression to predict the mean response (often after applying a log or power transformation) among the units for which the response is nonzero. Modeling the semicontinuous variable in two stages is desirable not only from a statistical viewpoint, but for scientific reasons as well; the covariates that predict the probability of nonzero response may be quite distinct from those that influence the level of response among the nonzero respondents. The strategy for modeling BAC mirrors the two-stage models of the econometric literature. The first step is to identify covariates that are significantly related to the dichotomized version of BAC, and then identify those related to the level of BAC among persons having BAC>0. Once these covariates are identified, they are incorporated into a GLOM that simultaneously estimates their relationships to the two-variable recoded version of BAC. Missing values of these variables are then multiply imputed using MIX software.

Initially, the model was to describe all actively involved persons for any given year of FARS by a single large GLOM. This proved to be impractical, however, for a number of reasons. First, the vastness and sparseness of the data caused by the large number of potential predictors made it difficult to fit such a model with the computer resources available to us. Second, to specify such a model would have been difficult because the meaning and importance of many potential predictors varies across vehicle classes. For example, the predictor RESTR (use of restraint device) is an indicator of seat belt use for drivers of cars and trucks, an indicator of helmet use for motorcyclists,

and irrelevant for nonoccupants. LSTAT (possession of a valid driver's license) is irrelevant for nonoccupants. Gender is significantly related to BAC within some vehicle classes (e.g., passenger cars), but within other classes (e.g., medium and heavy trucks) the rarity of female drivers makes the estimation of the relationship nearly impossible. Because of these complications, we retreated from a full simultaneous modeling effort across all classes to fit a model separately within each class. Splitting the data by vehicle class, which was consistent with the modeling work of Klein (1986), effectively allowed us to consider all potential interactions between vehicle class and all other covariates for predicting BAC.

An additional complication was that the missingness in the police report of alcohol involvement was clearly nonignorable. Most missing data methods currently in use, including the procedures described by Schafer (1997), assume that nonresponse 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 quantities that are observed). For the variable DRINKING, which is missing for a substantial proportion of cases, the meaning of a missing value seemed to vary dramatically from state to state. Sometimes a missing value probably indicated "no alcohol"; the form was left blank because there were no indications of alcohol involvement present. In other cases the police data may have been unreported for "policy" reasons; and sometimes it was probably simply missing for reasons not related to actual BAC after controlling for other variables. An imputation method that assumed ignorable nonresponse 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, although not fully satisfactory, is consistent with the modeling approach used by Klein (1986). A better solution would have been to develop a plausible probability model for the nonresponse that includes interactions between DRINKING and state. Developing and fitting such a model would have been a very substantial task, well outside the scope of this project.

A final issue that created difficulty was the sparseness of data caused by a large number of potential predictors. Within each vehicle class, the imputation model was to include BAC, DRINKING, and up to ten additional categorical predictors. Theoretically it would have been desirable to include all predictors in each model, but many of the vehicle classes simply did not have enough cases to support such a large number of predictors. It was found to be necessary to eliminate predictors that were not significantly related to BAC, using another standard statistical technique formerly employed by Klein (1986): stepwise fitting of models, bringing in important predictors and leaving out predictors that appear to have little predictive power. A theoretical problem with this approach is that the omitted variables are implicitly being treated as though they are precisely known to be unrelated to BAC, when in fact their associations with BAC may not be well estimated. Omitting these predictors, rather than acknowledging that there may be only weak evidence about their associations with BAC, results in multiple imputations that understate the actual levels of statistical uncertainty. The GLOM multiple imputation procedure does correctly reflect the uncertainty due to the estimation of parameters within the model, but ignores uncertainty for parameters omitted from the system as a result of stepwise variable selection. This understatement of uncertainty, however, is present to an even greater degree in the methodology of Klein (1986),

which does not represent parameter uncertainty of any kind. Adopting a stepwise variable selection approach was seen as a reasonable compromise, given the scope of this project and the limitations of existing models and software.

3. Implementing multiple imputation 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 within a vehicle class. Because this GLOM is being used to predict missing values of BAC, most of the effort was devoted to creating a plausible conditional model for BAC given the predictors. In the FARS database, however, missing values are found not only in BAC but in many of the covariates as well. To handle arbitrary patterns of missing values in the covariates, GLOM specifies a joint probability distribution for all variables at once, and the MIX software imputes missing values for BAC and all other variables simultaneously. Because the model was not specifically designed to produce high-quality imputations for missing covariates, however, these imputed values are discarded from the final multiply-imputed data set and only the imputed values for BAC are retained.

The GLOM is most easily understood as a two-stage model that captures the variable's semicontinuous nature. In the 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. In the second stage, the actual level of *log(*BAC*)* among the cases for which BAC>0 is related to a subset of the same covariates by conventional linear regression. A regression model for the logarithm of BAC, rather than for BAC itself, is attractive because negative values of BAC are excluded. Moreover, distributions of nonzero BAC values tend to be positively skewed, and taking logs tends to reduce this skewness. In many vehicle classes, however, the logarithmic transformation was not powerful enough to create a plausible normal linear regression model; it was often necessary to apply another power transformation after taking logs. The actual steps of the model-fitting and imputation procedure are detailed below.

Step 1: Separating cases by vehicle class

Following the earlier work of Klein (1986), the population of actively involved persons in a given year of FARS is first separated by vehicle class, because experience has shown that the relationships between BAC and the covariates may vary substantially across the classes. Indeed, the actual definitions of some covariates vary by vehicle class, so describing the entire FARS dataset by a single model is conceptually and practically difficult. The model-fitting and imputation steps described below are carried out separately within each class, and the imputed datasets are merged back into a single dataset. The classification, which is given in Exhibit 1, is the same as that used by Klein (1986) but with one modification: Due to the increased popularity of minivans in the last decade, these vehicles are represented as a separate category. Preliminary analyses indicated that the numbers of motor homes and buses appearing in FARS in a single year are typically too small to

support individual models for these classes.[1] To overcome this problem, BUS and HOME are combined with MISC for purposes of modeling. To allow for the possibility that the rates of alcohol involvement might vary among these groups, however, a main effect to distinguish among the three groups (BUS, HOME, MISC) is considered for inclusion in the variable-selection procedures below. At this point, any missing values of BAC for persons under 12 years old are set to zero. Any observed nonzero values of BAC in this age group, however, are allowed to stand.

Exhibit 1
Vehicle classes used in BAC imputation model

Class	Description	FARS Body Types (1993)
BUS	buses	12, 24-25, 50-59
HOME	motor homes	23, 42, 65, 73
LTV	light trucks and vans (pickup trucks and standard vans)	20-22, 28-41, 45-49
MINIV	minivans	20
MISC	miscellaneous vehicles	13, 90 and above
MHT	medium and heavy trucks	60-64, 66-72, 78, 79
MOT	motorcycles	80-89
NOC	nonoccupants	-
PC	passenger cars	1-11
UTIL	utility vehicles	14-19

Step 2: First-stage model selection

After the data are separated by vehicle class, a first-stage loglinear model for each class is selected. The purpose of this model is to capture the essential relationships among the dichotomized BAC 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

[1]In the 1993 FARS data, for example, there were no observed positive values of BAC among bus drivers. Thus there were insufficient data to even fit a model without covariates, because the observed data provided no information about the distribution of positive values of BAC.

loglinear models (Goodman, 1970), models that include all possible associations among the covariates, and a simple association between the dichotomized BAC and each covariate were 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 that it is significant at the 0.1 level as judged by a deviance (likelihood-ratio) test. After it is entered, the significance of each term currently in the model is tested, and any terms that are no longer significant at the 0.1 level are 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.

Model fitting at this stage is carried out via an ECM algorithm described in Chapter 8 of Schafer (1997). The ECM algorithm and deviance procedures are implemented by Schafer in an S-PLUS software library called CAT. The basic functions in CAT are called by macros that carry out the stepwise variable selection automatically. There are two minor technical details associated with this procedure. First, to avoid potential problems associated with maximum likelihood (ML) estimates on the boundary, a small amount of prior information is added to smooth the estimates toward a uniform table. In effect, fractional counts are added to each cell adding up to 5% of the total sample size, distributed uniformly across the cells. Second, if the ECM algorithm fails to converge within 100 iterations while entering a term, that term is excluded from further consideration. Failure to converge by 100 iterations indicates that the rate of missing information for that term is very high, and including it in the imputation model could be risky.

The covariates considered for possible inclusion in the first-stage model are listed in Exhibit 2. With two exceptions, this same pool of covariates is used for each vehicle class. The two exceptions are: (a) The covariates DRREC, RESTR, LSTAT, and SSS are excluded for the NOC class because they are not defined for non-occupants, and (b) the combined model for the BUS, HOME, and MISC may also include an extra covariate to distinguish among the three classes.

Exhibit 2
Covariates considered in the first-stage model for dichotomized BAC

Covariate	Description	Levels
DRINKING	police reported drinking	1=no alcohol, 2=alcohol, 3=missing
AGE	age category	1=under 12, 2=12-20, 3=21-29, 4=30-39, 5=40-49, 6=50-59, 7=60 and over
SEX	gender	1=male, 2=female
RESTR[2]	use of restraint	1=no, 2=yes
SEV	injury severity	1=non fatal, 2=fatal
LSTAT	license status	1=no valid license, 2=valid license
DRREC	previous incidents	1=none, 2=1 incident, 3=2 or more incidents
DAY	day of week	1=Mon-Thurs, 2=Fri, 3=Sat, 4=Sun
HOUR	time of day	1=6:00-9:59, 2=10:00-15:59, 3=16:00-19:59, 4=20:00-23:59, 5=0:00-5:59
SSS	vehicle role	1=single vehicle, 2=multiple vehicle striking, 3=multiple vehicle struck
RDWY	relation to roadway	1=not on roadway, 2=on roadway

Step 3: Selection of a transformation for nonzero BAC

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 the distribution of *log(BAC)* was found to be negatively skewed. Preliminary

[2]Accounts only for the use of belts or helmets irrespective of the presence or absence of a supplemental restraint system like an airbag at that seating position.

analyses showed that normal linear regression models for *log(BAC)* could impute values for BAC that are implausibly high. Power transformations of the form $log(BAC)^{\lambda}$ for various $\lambda>1$ gave better results, but a value of λ that worked well for one vehicle class often did not work well for another. Choosing a value of λ for each class by an interactive, trial-and-error procedure was straightforward but tedious. Finally, an automatic procedure was devised based on the well-known maximum-likelihood method of Box and Cox. The Box-Cox method finds the power transformation that makes $log(BAC)^{\lambda}$ most nearly normal, as judged by the mode of the profile likelihood function---i.e., the likelihood maximized with respect to the mean and variance for any fixed value of λ. The resulting ML estimate tended to worked well for many vehicle classes, but for some classes it still produced a handful of imputed values for BAC that were implausibly high. Adding 1 to the ML estimate, however, appeared to solve that problem. The automatic transformation procedure proceeds as follows: (a) The Box-Cox estimate is found by a grid search over the values 0.1, 0.2, ..., 4.5. (b) The positive values of BAC are transformed to

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

(c) After imputation, the imputed values are transformed back to the original BAC scale using the back-transformation g^{-1}. The linear model occasionally imputes a negative value for g(BAC) which cannot be transformed back. On these rare occasions, the missing value of BAC is simply assigned to zero.

Step 4: Second-stage model selection

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 model, with the exception of the 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) on dummy indicators for the categorical predictors. For simplicity, the stepwise procedure adopts a complete-case (CC) approach, omitting from consideration any persons with missing covariates. The CC method is nearly efficient as the overall proportion of cases discarded at this stage tends to be small. It is to be noted that the CC method is applied only for purposes of variable selection; after the second-stage covariates are determined, the actual imputation procedure described below operates on the full set of observed data, including cases that are only partially observed.

Step 5: Imputation

Once the first and second-stage covariates have been selected, multiple imputations of the missing data are created under a general location model (GLOM). The GLOM is simply a combination of the first-stage loglinear model with the second-stage linear regression procedure. The imputation proceeds as follows. First, ML estimates of the model parameters are found using an ECM algorithm described in Chapter 9 of Schafer (1997), implemented in the software library MIX. Then, using the ML estimates as starting values, new values of the parameters are simulated from their

posterior distribution by a Markov-Chain Monte Carlo (MCMC) algorithm, also described in Chapter 9 of Schafer (1997). Experience has shown that the number of steps required for ECM to converge is a conservative estimate of the number of steps required for the MCMC method 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 the simulated values of the parameters. Repeating the process ten times results in ten imputations of the missing data as ten random starting values are provided for the simulation by the random number generator function. The imputed values of g(BAC) are then transformed back to the BAC scale, and rounded to the same number of decimal places as found in the original data.

In the initial rounds of imputation, the MCMC algorithm behaved erratically for two of the largest vehicle classes, namely the PC and the LTV classes; the algorithm halted due to numeric overflows. After an extensive investigation, it was discovered that the overflows were caused by unduly large simulated values of the residual variance in the second-stage regression model. The posterior distribution, it turned out, was nonexistent. The "Jeffreys prior" that was being used assigns uniform density over the real line to all values of $\log(\sigma^2)$. This is not a proper probability distribution, and in some cases it may result in an improper posterior. To remedy the problem, the Jeffreys prior was replaced with a scaled inverted-chi square,

$$\sigma^{-2} \sim c\chi^2_{\nu}.$$

In this prior distribution, c/ν can thought of as a prior estimate of σ^2, and ν can thought of as a prior number of degrees of freedom on which this estimate is based. To distort the inferences as little as possible, ν was chosen to be equal to 3 and c was taken to be the ML estimate of $3\sigma^2$ obtained from the ECM algorithm. Under this prior, the MCMC algorithm exhibited no more numerical difficulties.

4. Analyzing the multiply-imputed datasets

In this section, two methods for analyzing the multiply-imputed datasets are reviewed: Rubin's (1987) method for scalar estimands, and the method of Meng and Rubin (1992) for combining likelihood-ratio test statistics.

Method for scalar estimands

Let Q be a one-dimensional quantity of interest --- a population proportion, a coefficient from a linear or logistic regression model, etc. The goal is to find a confidence interval or test a hypothesis about Q. Let Y denote the data from FARS that are necessary to estimate Q, which is partitioned into observed and missing parts,

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

where Y_{obs} is seen, but the missing part Y_{mis} is unknown and has to be multiply imputed. Let $\hat{Q}$ be the complete-data point estimate for Q, the estimate that one would use 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 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 the sample size sufficiently large for the asymptotic normal approximation

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

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 tth set of imputed data, $t=1,2,\ldots,m$. The multiple-imputation point estimate for Q is simply the average of the complete-data point estimates.

$$\bar{Q} = \frac{1}{m}\sum_{t=1}^{m} \hat{Q}^{(t)}. \tag{2}$$

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

$$\bar{U} = \frac{1}{m}\sum_{t=1}^{m} U^{(t)}. \tag{3}$$

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

$$B = \frac{1}{m-1}\sum_{t=1}^{m} (\hat{Q}^{(t)} - \bar{Q})^2. \tag{4}$$

and inferences are based on the approximation

$$T^{-1/2}(Q - \bar{Q}) \sim t_{\nu}, \tag{6}$$

where the degrees of freedom are given by

$$\nu = (m-1)[1+\frac{\bar{U}}{(1+m^{-1})B}]^2. \tag{7}$$

Thus a 100(1-α)% interval estimate for Q is

$$\bar{Q} \pm t_{\nu,\, 1-\alpha/2}\sqrt{T}, \tag{8}$$

and a p-value for testing the null hypothesis

$$Q = Q'$$

against a two-sided alternative is

$$2P(t_{\nu} \geq T^{-1/2}|\bar{Q}-Q'|)$$

or, equivalently,

$$P[F_{1,\nu} \geq T^{-1}(\bar{Q}-Q')^2]. \tag{9}$$

The degrees of freedom (7) depend not only on m, but also on the ratio

$$r = \frac{(1+m^{-1})B}{\bar{U}}. \tag{10}$$

Rubin (1987) calls r the *relative increase in variance due to nonresponse*, because $\bar{U}$ represents the estimated total variance when there is no missing information about Q (i.e., when B=0). When m is large and/or r is small, the degrees of freedom will be large and (6) will be approximately normal. Moreover,

$$\begin{aligned} \hat{\lambda} &= (\bar{U}^{-1} - (v+1)(v+3)^{-1}T^{-1})\bar{U} \\ &= \frac{(r+2)/(v+3)}{r+1} \end{aligned} \tag{11}$$

is an estimate of the *fraction of missing information* about Q. In applications, examination of r and $\hat{\lambda}$ is highly recommended, as they are interesting and useful diagnostics for assessing how the missing data contribute to inferential uncertainty about Q. An implementation of this method for scalar estimands is part of the accompanying software packages CAT and MIX, in an S-PLUS function called `mi.inference` (see Appendix). More information is provided by the S-PLUS command `help(mi.inference)`. The three examples described below illustrate sample analyses of multiply-imputed datasets.

Example 1: Population proportions

This example finds a confidence interval for a population proportion. Under simple random sampling, the natural complete-data estimate $\hat{Q}$ is the sample proportion, and the natural estimate of its variance is

$$U = \frac{\hat{Q}(1-\hat{Q})}{n}$$

where n is the sample size. To obtain the confidence interval, calculate $\hat{Q}^{(1)}, \ldots \hat{Q}^{(m)}$ and $U^{(1)}, \ldots, U^{(m)}$ from the multiply-imputed dataset and apply (2)-(8). The S-PLUS file `analyze.s` contains examples of how to carry this out within S-PLUS.

Example 2: Logistic-regression coefficients

This example performs a logistic-regression analysis on the multiply-imputed dataset. In this case, $\hat{Q}$ is an estimated coefficient and U is its squared standard error, both of which are provided by standard logistic-regression software. The regression model is fit m times, once for each imputation, storing the m sets of coefficients and standard errors. The S-PLUS function `mi.inference` will apply the combination rules for all coefficients at once, automatically producing a table of combined estimates, standard errors, and p-values.

Combining likelihood-ratio tests

The methods described above aid in making inferences about scalar parameters one at a time. Occasionally, however, it is necessary to make a joint inference about a group of parameters at once. For example, if one is interested in testing the null hypothesis that a particular regression model is true, versus the alternative that a larger model is true, where (a) the smaller model is a special case of the larger one, and (b) the two models differ by k>1 degrees of freedom. The most straightforward way of obtaining a p-value in this situation is to combine the results of m likelihood-ratio tests, using the method of Meng and Rubin (1992). Let ψ denote the vector of unknown parameters in the analyst's model, and $Q=Q(\psi)$ a k-dimensional function of ψ that is of interest; specifically, we wish to test the hypothesis that $Q=Q'$ for a given Q'. Let $l(\psi|Y_{obs}, Y_{mis})$ denote the complete-data log-likelihood function, $\hat{\psi}$ the MLE or maximizer of $l(\psi|Y_{obs}, Y_{mis})$, and $\hat{\psi}_0$ the maximizer of $l(\psi|Y_{obs}, Y_{mis})$ subject to the constraint $Q(\psi)=Q'$. In regular problems, the complete-data likelihood-ratio test statistic

$$d_L=(\hat{\psi},\hat{\psi}_0|Y_{obs},Y_{mis})$$

$$=2[l(\hat{\psi}|Y_{obs},Y_{mis})-l(\hat{\psi}_0|Y_{obs},Y_{mis})]$$

is asymptotically distributed as ψ^2 under the null hypothesis. Let

$$d_L^{(t)} = d_L(\hat{\psi}^{(t)},\hat{\psi}_0^{(t)}|Y_{obs},Y_{mis}^{(t)}),$$

be the likelihood-ratio test statistic from the tth imputed dataset, t=1,2, ... ,m, where $\hat{\psi}^{(t)}$ is the maximizer of $l(\psi|Y_{obs}, Y_{mis}^{(t)})$ and $\hat{\psi}_0^{(t)}$ is the maximizer of $l(\psi|Y_{obs}, Y_{mis}^{(t)})$ subject to $Q(\psi)=Q'$. Let

$$\bar{d}_L = \frac{1}{m}\sum_{(t=1)}^{m} d_L^{(t)}$$

be the average of these likelihood-ratio statistics, and

$$\bar{\hat{\psi}} = \frac{1}{m}\sum_{t=1}^{m} \hat{\psi}^{(t)}, \qquad (12)$$

be the averages of the complete-data estimates of ψ across imputations. Finally, let

$$\bar{d}_L = \frac{1}{m}\sum_{t=1}^{m} d_L(\bar{\Psi}_0, \bar{\Psi} | Y_{obs}, Y_{mis}^{(t)})$$

$$\bar{\Psi}_0 = \frac{1}{m}\sum_{t=1}^{m} \hat{\Psi}_0^{(t)} \tag{13}$$

be the average of the likelihood-ratio statistics evaluated at Ψ_0 and Ψ, rather than at the imputation-specific parameter estimates. The test statistic proposed by Meng and Rubin (1992) is

$$D = \frac{\bar{d}_L}{k(1+r)}, \tag{14}$$

where

$$r = \frac{m+1}{k(m-1)}(\bar{d}_L - \tilde{d}_L) \tag{15}$$

is an estimate of the average relative increase due to nonresponse across the components of ψ. The p-value associated with D is

$$p = P(F_{k,v} \geq D) \tag{16}$$

with degrees of freedom calculated as

$$v = \begin{cases} 4+(t-4)[1+(1-2t^{-1})r^{-1}]^2 & \text{if } t = k(m-1)>4, \\ t(1+k^{-1})(1+r^{-1})^2/2 & \text{otherwise.} \end{cases}$$

In addition to the usual likelihood-ratio test statistics for each imputed dataset, this procedure also requires evaluation of the complete-data likelihood ratio at ($\bar{\Psi}$, $\bar{\Psi}_0$) for each dataset.

Implementation of this procedure thus requires code for evaluating the complete-data loglikelihood at user-specified values of the parameter, something which is not typically provided in standard statistical software. For many commonly used models, however, the complete-data log-likelihood is straightforward to derive and compute, and with a little effort on the part of the analyst, the procedure can often be implemented without difficulty.

Example 3: Testing the joint significance of a group of logistic-regression coefficients

A logistic-regression model is

$$\log\frac{\pi_i}{1-\pi_i} = x_i^T\beta, \tag{17}$$

where π_i denotes the probability of "success" for subject I, x_i is a vector of covariates for subject I, and β a vector of unknown coefficients. With complete data, the loglikelihood function for the logistic model (17) may be written as

$$l(\beta|Y_{obs},Y_{mis}) = \sum_{i=1}^{n}[z_i\log\frac{e^{x_i^T\beta}}{1+e^{x_i^T\beta}} + (1-z_i)\log\frac{1}{1+e^{s_i^T\beta}}],$$

where z_i=1 if individual I has a success, and z_i=0 otherwise (e.g. McCullagh and Nelder, 1989). Suppose one wants to test the null hypothesis that a group of k>=1 coefficients is simultaneously zero. The usual likelihood-ratio test with complete data requires one to fit (a) the full model with all variables, and (b) the reduced model with all variables except those whose coefficients are zero under the null hypothesis. The ML estimates of β under the full and reduced models are denoted by $\hat{\beta}$ and $\tilde{\beta}$, respectively. For notational convenience, it is assumed that $\hat{\beta}$ and $\tilde{\beta}$ are of the same length, with the elements of $\tilde{\beta}$ corresponding to the omitted variables set to zero. The likelihood-ratio test statistic is

$$d_L(\hat{\beta},\tilde{\beta}|Y_{obs},Y_{mis}) = 2[l(\hat{\beta}|Y_{obs},Y_{mis}) - l(\tilde{\beta}|Y_{obs},Y_{mis})],$$

which, under the reduced model, is approximately distributed as χ^2_k because the reduced model differs from the full model by k parameters. The method of Meng and Rubin (1992) requires two passes through the imputed data. Let $\hat{\beta}^{(t)}$ and $\tilde{\beta}^{(t)}$ denote the ML estimates for the full and reduced models, respectively, fit to the t^{th} imputed dataset. In the first pass, the likelihood-ratio statistic is calculated for each imputed dataset and their average computed as,

$$\bar{d}_L = \frac{1}{m}\sum_{t=1}^{m} d_L(\hat{\beta}^{(t)}|Y_{obs}, Y_{mis})],$$

In the second pass, the average of the likelihood-ratio test statistics is calculated with $\hat{\beta}^{(t)}$ and $\tilde{\beta}^{(t)}$ replaced by their averages,

$$\bar{d}_L = \frac{1}{m}\sum_{t=1}^{m} d_L(m^{-1}\sum_{t=1}^{m} \tilde{\beta}^{(t)}|Y_{obs}, Y_{mis}^{(t)}).$$

The test statistic D and p-value are then found by (14)-(16).

5. Comparing the results of multiple imputation and the previous method

Multiple imputations of missing BAC have been created for the FARS databases for 1982, 1986, 1990, 1993 and 1995. In 1997, an extensive study was carried out at NHTSA to compare the results of the new imputation method to those of the previous system implemented by Klein (1986). Klein's method differs from the new system in that it relies on a linear discriminant model rather than a general location model (GLOM). Moreover, Klein's system estimates probabilities for the three-level categorization BAC3, rather than imputing random values of BAC. In many other respects, however, the two methods are quite similar. In both methods, the FARS data are split by vehicle class prior to modeling, and both employ stepwise variable-selection procedures using nearly identical pools of covariates. Finally, both methods regard DRINKING as a three-level (yes, no, missing) categorical predictor, in recognition of this variable's unusual pattern of nonresponse.

Because of the strong similarities between the two methods, they are expected to produce similar results. In particular, it was expected that estimates of the rates of alcohol involvement (BAC>0) and rates in excess of the typical legal limit (BAC>.10) within important subclasses would be similar. In many respects the pattern of results was indeed similar. One important difference, however, was that under the new method the estimated rates of alcohol involvement (BAC>0) were generally higher. Positive differences of about 2% in the rate of alcohol involvement appeared consistently across most vehicle classes and demographic subgroups, and across classifications of crashes by time of day and day of week. Differences in rates across subgroups, and trends in rates across time, were quite similar under the two methods. The only major discrepancy was that the baseline rates of alcohol involvement were slightly but consistently higher under the new method.

Exhibits 3 through 6 present a comparison of the BAC estimates from the multiple imputation method to those of Klein's (1986) discriminant model under various categories traditionally reported and released by NHTSA in its annual fact sheets. The estimates from multiple imputation are printed

in each cell accompanied by the corresponding estimates from the discriminant method in parenthesis. The values in each cell are the percentage of all crashes or drivers in that category for two levels of BAC, namely 0.01[3] and greater (0.01+) as well as 0.10 and greater (0.10+). For example, an entry of 36.2 for a given year for drivers killed in the 0.01+ category means that for that particular year, 36.2 percent of all the drivers killed in fatal crashes had a BAC of 0.01 or greater.

Exhibit 3 presents the overall rates of alcohol involvement and intoxication in fatal crashes as estimated by the two methods. Alcohol is said to be involved in a fatal traffic crash if either the driver of any vehicle or a nonoccupant (pedestrian or pedalcyclist) involved in the crash has a BAC of 0.01 or greater. A driver or nonoccupant is deemed "intoxicated" if the BAC is 0.10 or greater. Exhibit 4 presents similar comparisons for drivers only, i.e., for any driver with the relevant BAC level in any vehicle involved in a fatal crash. The statistics have been broken down by categories of sex, age and for the four major categories of vehicle type, namely, passenger cars, light trucks and vans, medium and heavy trucks, and motorcycles. Exhibits 5 and 6 present the corresponding results for drivers killed and nonoccupants killed, respectively.

Exhibit 3: Percentage Alcohol involvement and intoxication in fatal crashes.

	1982		1986		1990		1993		1995	
	0.01+	0.10+	0.01+	0.10+	0.01+	0.10+	0.01+	0.10+	0.01+	0.1+
Fatal Crashes	59.1 (56.7)	49.2 (46.1)	53.8 (51.7)	42.9 (40.8)	50.7 (49.4)	41.4 (39.7)	44.6 (43.5)	36.3 (35.0)	42.5 (41.3)	34.2 (32.8)
Total Fatalities	59.5 (57.3)	49.3 (46.3)	54.1 (52.2)	42.9 (41.1)	50.7 (49.6)	41.2 (39.6)	44.6 (43.5)	36.0 (34.9)	41.9 (41.3)	33.5 (32.5)

Note: Cell entries represent Multiple Imputation (Klein Method)

[3]A BAC level of 0.01 implies that the alcohol content is 0.01 grams/deciliter. Because BAC is reported to two decimal places, the 0.01+ category includes all positive values of BAC.

Exhibit 4: Alcohol involvement for drivers in fatal crashes.

Drivers Involved in Fatal Crashes	1982		1986		1990		1993		1995	
	0.01+	0.10+	0.01+	0.10+	0.01+	0.10+	0.01+	0.10+	0.01+	0.10+
Total Drivers										
Total	40.8 (38.9)	32.3 (30.0)	36.0 (34.3)	27.4 (25.8)	33.2 (32.1)	26.0 (24.7)	28.1 (27.3)	21.9 (21.0)	26.8 (25.5)	20.7 (19.3)
Total Drivers Involved by Sex										
Male	44.0 (41.8)	35.0 (32.4)	39.6 (37.6)	30.3 (28.5)	37.2 (35.7)	29.4 (27.7)	31.6 (30.5)	24.7 (23.7)	29.9 (28.5)	23.2 (21.8)
Female	26.9 (25.7)	20.4 (18.9)	21.8 (20.9)	15.7 (14.8)	20.1 (19.2)	14.9 (13.8)	17.2 (16.5)	12.9 (12.1)	17.3 (15.7)	13.0 (11.2)
Total Drivers Involved by Age										
16-20	44.8 (44.0)	32.4 (31.1)	37.2 (36.4)	24.4 (23.7)	32.6 (31.7)	22.3 (21.1)	24.7 (24.5)	16.5 (16.2)	21.8 (20.6)	14.1 (12.7)
21-24	53.5 (51.5)	42.6 (40.0)	49.0 (47.2)	37.6 (36.1)	46.1 (44.9)	36.3 (34.7)	40.2 (39.4)	31.5 (30.7)	37.8 (37.2)	28.5 (27.8)
25-34	46.3 (43.9)	37.9 (35.1)	43.6 (41.5)	34.9 (33.0)	42.8 (41.3)	34.7 (33.0)	37.2 (36.1)	29.8 (28.5)	35.6 (33.9)	28.7 (26.8)
35-44	37.6 (34.9)	31.0 (27.9)	33.0 (30.6)	26.7 (24.5)	33.7 (32.0)	27.8 (25.8)	30.9 (29.3)	25.2 (23.6)	30.6 (28.6)	24.8 (22.8)
45-64	28.7 (26.4)	23.2 (20.7)	21.5 (20.2)	17.8 (16.2)	19.7 (18.5)	17.0 (15.5)	20.2 (18.4)	15.7 (14.6)	19.8 (18.4)	15.8 (14.3)
Over 64	14.9 (13.7)	10.7 (9.9)	11.7 (10.5)	7.3 (6.8)	10.6 (9.7)	6.8 (6.3)	8.3 (8.0)	5.5 (5.4)	8.4 (7.4)	5.7 (4.9)

Exhibit 4 (Continued..) : Alcohol involvement for drivers in fatal crashes.

Drivers Involved in Fatal Crashes	1982		1986		1990		1993		1995	
	0.01+	0.10+	0.01+	0.10+	0.01+	0.10+	0.01+	0.10+	0.01+	0.10+
Total Drivers Involved by Vehicle Type										
Pass. Cars	41.9 (39.9)	33.2 (30.6)	36.3 (34.7)	27.4 (25.8)	34.1 (32.0)	26.6 (24.3)	28.5 (27.3)	22.0 (20.7)	26.8 (25.7)	20.7 (19.2)
Light Trucks	44.4 (43.0)	36.6 (34.4)	38.9 (37.2)	30.6 (29.5)	36.9 (36.5)	(29.9) (29.3)	31.7 (31.1)	25.9 (24.9)	31.7 (28.7)	25.5 (22.8)
Medium/Heavy Trucks	10.3 (8.2)	5.5 (4.5)	7.2 (5.8)	4.3 (3.2)	4.9 (5.1)	2.7 (2.5)	3.4 (3.3)	1.7 (1.7)	3.4 (3.1)	1.7 (1.4)
Motorcycles	55.5 (53.5)	42.6 (40.5)	55.7 (54.4)	42.3 (40.9)	52.5 (52.1)	39.8 (39.3)	45.4 (44.0)	33.9 (32.9)	41.6 (40.7)	29.8 (29.2)

Exhibit 5: Alcohol involvement for drivers killed in fatal crashes.

Drivers Killed in Fatal Crashes	1982		1986		1990		1993		1995	
	0.01+	0.10+	0.01+	0.10+	0.01+	0.10+	0.01+	0.10+	0.01+	0.10+
Total Drivers Killed										
Total	55.5 (53.1)	46.2 (43.8)	49.8 (48.2)	40.0 (38.7)	46.3 (45.9)	38.2 (38.0)	41.0 (39.7)	33.8 (32.8)	38.7 (37.9)	31.5 (30.9)
Total Drivers Killed by Crash Type and Time of Day										
Single Vehicle (SV)	70.9 (68.7)	61.4 (59.6)	66.2 (64.3)	55.8 (54.5)	63.2 (62.9)	54.2 (54.2)	57.7 (56.5)	49.7 (48.9)	54.9 (54.9)	46.3 (46.7)
SV-Daytime	42.2 (39.7)	34.4 (32.1)	36.5 (34.3)	28.7 (27.2)	33.4 (33.2)	26.9 (26.8)	28.9 (27.8)	23.3 (23.1)	28.5 (27.2)	22.7 (21.8)
SV-Nighttime	82.9 (80.8)	72.7 (71.1)	80.3 (78.5)	68.6 (67.5)	78.1 (77.9)	67.9 (68.1)	74.4 (73.0)	64.9 (63.9)	70.2 (71.3)	60.1 (61.6)
Multiple Vehicle (MV)	39.9 (37.4)	30.7 (27.9)	33.9 (32.6)	24.8 (23.5)	29.8 (29.4)	22.6 (22.2)	25.4 (24.1)	19.1 (18.0)	23.7 (22.3)	17.8 (16.4)
MV-Daytime	20.3 (18.4)	14.3 (12.1)	17.8 (16.8)	11.5 (10.5)	14.4 (14.0)	9.5 (9.1)	12.0 (10.8)	7.9 (7.0)	12.7 (11.3)	8.2 (7.0)
MV-Nighttime	58.9 (55.8)	46.6 (43.4)	54.1 (52.4)	41.4 (39.8)	50.6 (50.1)	40.4 (39.8)	46.4 (44.8)	36.6 (35.1)	42.4 (41.1)	34.1 (32.4)
Total Drivers Killed by Day of Week										
Weekday	46.0 (43.5)	37.7 (35.3)	40.6 (39.0)	32.3 (31.1)	36.0 (35.5)	29.2 (28.8)	31.1 (30.1)	25.4 (24.8)	29.6 (28.7)	23.8 (23.1)
Weekend	66.8 (64.6)	56.3 (54.1)	61.8 (60.1)	50.1 (48.7)	59.5 (59.3)	49.8 (49.8)	54.9 (53.1)	45.7 (44.4)	51.8 (51.3)	42.6 (42.3)
Total Drivers Killed by Time of Day										
Daytime	28.6 (26.4)	21.9 (19.6)	24.5 (23.0)	17.6 (16.5)	21.3 (21.0)	15.8 (15.5)	18.0 (17.0)	13.3 (12.8)	18.3 (16.9)	13.3 (12.2)
Nighttime	72.8 (70.3)	61.8 (59.5)	69.5 (67.8)	57.5 (56.1)	67.0 (66.7)	56.7 (56.6)	63.0 (61.6)	53.4 (52.3)	59.2 (59.1)	50.0 (50.0)

Exhibit 5 (contd.): Alcohol involvement for drivers killed in fatal crashes.

Drivers Killed in Fatal Crashes	1982		1986		1990		1993		1995	
	0.01+	0.10+	0.01+	0.10+	0.01+	0.10+	0.01+	0.10+	0.01+	0.10+
Total Drivers Killed by Day of Week and Time of Day										
Weekday Daytime	24.7 (22.3)	18.6 (16.2)	20.9 (19.5)	14.9 (13.8)	17.6 (17.2)	12.6 (12.3)	14.7 (13.7)	10.8 (10.4)	15.2 (13.7)	11.0 (9.8)
Weekday Nighttime	68.8 (66.1)	58.1 (55.6)	65.2 (63.4)	54.0 (52.7)	62.1 (61.6)	52.6 (52.3)	57.0 (56.1)	48.3 (47.2)	53.5 (53.5)	44.8 (45.1)
Weekend Daytime	38.5 (36.9)	30.3 (28.3)	33.6 (32.1)	24.5 (23.3)	31.0 (30.9)	24.1 (24.0)	27.0 (25.4)	20.2 (19.1)	26.8 (25.5)	19.5 (18.7)
Weekend Nighttime	75.9 (73.6)	64.6 (62.4)	73.0 (71.4)	60.3 (58.9)	70.7 (70.5)	59.9 (59.9)	67.8 (66.0)	57.5 (56.2)	63.9 (63.9)	53.9 (53.9)

Exhibit 6: Alcohol involvement for nonoccupants killed in fatal crashes.

Drivers Killed in Fatal Crashes	1982		1986		1990		1993		1995	
	0.01+	0.10+	0.01+	0.10+	0.01+	0.10+	0.01+	0.10+	0.01+	0.10+
Pedestrian Fatalities by Age Group										
16 - 20	57.4 (58.8)	44.8 (47.0)	43.4 (47.4)	31.6 (35.6)	42.1 (46.6)	32.4 (35.6)	36.3 (42.0)	29.1 (33.8)	32.1 (38.0)	25.3 (29.9)
21 - 24	64.8 (62.4)	56.0 (52.1)	64.3 (62.6)	52.6 (50.8)	61.6 (60.1)	53.5 (51.7)	60.4 (59.6)	51.4 (50.4)	56.2 (55.9)	46.7 (45.7)
25 - 34	63.9 (61.6)	56.6 (53.3)	63.9 (61.9)	55.3 (53.0)	65.4 (64.4)	58.1 (56.8)	62.4 (61.2)	55.1 (53.2)	59.4 (58.1)	51.1 (49.8)
35-44	61.4 (55.8)	55.4 (48.5)	61.0 (57.7)	53.6 (49.4)	59.7 (57.8)	52.6 (50.3)	58.3 (57.2)	51.8 (50.6)	58.8 (56.5)	50.2 (48.3)
45-64	52.3 (48.3)	45.7 (40.7)	48.1 (46.2)	41.7 (39.4)	44.0 (43.5)	36.9 (35.8)	41.6 (40.7)	36.5 (35.0)	44.4 (43.3)	38.9 (37.6)
Over 64	20.1 (19.2)	14.9 (13.8)	17.3 (17.0)	11.2 (10.8)	12.9 (12.9)	9.4 (9.0)	18.0 (15.7)	13.6 (10.6)	14.6 (15.3)	10.8 (10.8)
Total	45.7 (40.7)	40.3 (33.8)	39.3 (38.7)	32.5 (31.5)	37.8 (37.7)	32.2 (31.6)	37.2 (37.5)	32.0 (31.8)	36.6 (36.5)	30.9 (30.4)
Pedalcyclists	21.5 (20.3)	16.1 (14.3)	17.0 (17.7)	12.3 (12.2)	20.6 (21.0)	16.6 (15.9)	21.4 (23.3)	16.1 (17.1)	23.5 (24.2)	19.0 (18.9)

Upon further investigation, the generally higher rates of alcohol involvement under the new imputation system could be attributed to a fundamental difference between the GLOM and the linear discriminant models for BAC. In classical linear discriminant analysis, the joint distribution of the predictor variables is assumed to have constant variances and covariances across classification groups. The loglinear model within GLOM, however, makes no such assumption about the predictors. When this covariance assumption is (at least approximately) satisfied, the two approaches will predict essentially identical rates of alcohol involvement. When the assumption is violated, however, slight biases can be introduced by a linear discriminant approach relative to loglinear modeling. In Klein's (1986) discriminant models, the covariate DRINKING is expressed in the form of two dummy indicator variables to distinguish among the three groups (yes, no, missing). Because of the strong correlation between DRINKING and BAC, and because of the inherent relationship between the mean and variance of a binary variable, the variance of the DRINKING variable differed sharply among cases observed to have BAC=0 and those with BAC>0. Empirical investigations showed that this discrepancy would be sufficient to introduce a downward bias in discriminant-based estimates of alcohol involvement of approximately 2%.

6. Multiple imputation of known BAC values set to 'missing'

Because of the complexity of the new multiple imputation method, additional validation tests were conducted to ensure that the method produced plausible estimates. The most convincing evidence that the new procedure is performing properly came from an experiment in which multiple imputations were created for 'known' values of BAC in the FARS files. A subset of crash records with known values of BAC was extracted from the FARS files. Twenty-five percent of these values were chosen at random and set to be "missing". BAC values for these missing data were then imputed by the multiple imputation procedure, and the results from imputation were compared to the known figures. This validation experiment was carried out on data drawn from 1982, 1986, 1990, 1993, and 1995. Results from this experiment are summarized in Exhibits 7 and 8. For each year, these Exhibits report the known true rates of alcohol involvement and intoxication among the 25% of the cases that had been set to missing, and the rates among the same cases estimated from the multiple imputation method. Exhibit 7 presents results for drivers involved in fatal crashes subdivided by time of day, and Exhibit 8 presents the corresponding results for drivers killed. The imputed results track the known true values quite closely. Similar good behavior is seen when the data are subdivided by other crucial variables (e.g., gender and age).

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. The value of such tests would be dubious, however, because the nonresponse in FARS is not completely at random. We have strong evidence that missing values of BAC in FARS are more likely to be zero than are the observed variables, because of the relationships between missingness and many covariates that are strongly related to BAC. Nevertheless, the data in these Exhibits 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. A similar experiment was performed on the Discriminant method and is documented in the Model

Validation and Maintenance section of the report of Klein (1986).

Exhibit 7: BAC Estimates for Total Drivers Involved for the 25% of Known values randomly set to Missing

Methods	1982		1986		1990		1993		1995	
	0.01+	0.1+	0.01+	0.1+	0.01+	0.1+	0.01+	0.1+	0.01+	0.1+
Total Drivers Involved										
Known	64.3	51.9	57.1	43.8	51.2	40.6	46.0	36.9	43.7	34.7
MI	63.2	53.2	56.2	45.6	50.7	41.9	45.7	37.7	43.6	35.7
Total Drivers Involved (Daytime)										
Known	33.8	24.5	30.1	21.1	24.5	17.7	20.8	15.5	20.8	15.1
MI	36.2	30.2	29.7	23.7	24.9	20.3	21.2	17.0	20.9	16.7
Total Drivers Involved (Nighttime)										
Known	78.4	64.7	73.3	57.3	68.3	55.2	64.8	52.9	62.5	50.8
MI	75.7	64.0	72.1	58.9	67.5	56.0	64.2	53.2	62.2	51.3

Exhibit 8: BAC Estimates for Total Drivers Killed for the 25% of Known values randomly set to Missing

Methods	1982		1986		1990		1993		1995	
	0.01+	0.1+	0.01+	0.1+	0.01+	0.1+	0.01+	0.1+	0.01+	0.1+
Total Drivers Killed										
Known	63.0	53.3	56.0	44.6	51.8	43.3	47.3	39.7	45.3	37.7
MI	62.2	54.3	55.2	47.0	50.5	42.9	47.0	40.4	45.2	38.6
Total Drivers Killed (Daytime)										
Known	31.3	23.4	28.5	20.0	24.0	18.0	20.6	15.9	21.0	15.5
MI	31.0	29.1	28.5	23.6	24.0	20.1	21.5	18.0	21.5	17.7
Total Drivers Killed (Nighttime)										
Known	79.3	68.8	74.3	61.0	70.9	60.7	68.4	58.6	66.3	56.9
MI	77.0	67.3	73.1	62.7	69.0	58.7	67.2	58.3	65.7	56.6

7. Discussion

The new multiple imputation procedure for missing values of BAC in FARS represents a substantial improvement over the previous method based on three-class linear discriminant models. One major advantage of the new procedure is that the imputed datasets facilitate a wider variety of analyses. It is now possible, for example, to estimate intoxication rates under definitions of intoxication other than BAC=0.1+. Another important advantage is that uncertainty due to missing data in virtually any summary statistic can now be assessed in a straightforward way, using the rules for multiple-imputation inference developed by Rubin (1987) and Meng and Rubin (1992).

Because of inherent differences between loglinear and linear discriminant models, rates of alcohol involvement (BAC>0) estimated under the new method will tend to be slightly higher than those previously published. Increases of about 2% in base rates should be expected overall and within important subgroups. Important differences in rates (e.g., trends over time), however, should be relatively unaffected by the change in the missing-data procedure.

8. References

Goodman, L.A. (1970) The multivariate analysis of qualitative data: interactions among multiple classifications. *Journal of the American Statistical Association*, 65, 226-256.

Klein, T.M. (1986) A method for estimating posterior BAC distributions for persons involved in fatal traffic accidents. Report DOT-HS-807-094, National Highway Traffic Safety Administration, Department of Transportation.

Manning, W.G., Newhouse, J.P., Duan, N., Keeler, E.B., Leibowitz, A., and Marquis, M.S. (1987) Health insurance and the demand for medical care: evidence from a randomized experiment. *American Economic Review*, 77, 251-277.

McCullagh, P. and Nelder, J. (1989) *Generalized Linear Models*, Second edition. Chapman and Hall, London.

Meng, X.L. and Rubin, D.B. (1992) Performing likelihood ratio tests with multiply-imputed data sets. *Biometrika*, 79, 103--111.

Rubin, D.B. (1987) *Multiple Imputation for Nonresponse in Surveys.* J. Wiley & Sons, New York.

Rubin, D.B. (1996) Overview of multiple imputation for BAC in FARS. Unpublished report to National Highway Traffic Safety Administration.

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

Appendix: Implementation of the Multiple Imputation System at NHTSA

The software for imputing the missing values of BAC in FARS is written in the Fortran and S-PLUS programming languages. The Fortran code was compiled in NHTSA's DEC Alpha System using the standard Fortran-77 Compiler. The S-PLUS v3.3 compiler by MathSoft was used to process the S-PLUS codes. The programs are compatible with the Digital Unix v4.0 base operating system. The S-PLUS routines access the Fortran routines using standard object links to the compiled routines. BATCH jobs of the imputation runs can be executed remotely from any site that has access to the NHTSA's R&D Unix Cluster.

NHTSA began providing imputed values of BAC from 1982. The FARS data for the years 1982, 1993 and 1995 were chosen to test and validate the multiple imputation software. The FARS ASCII files were transmitted to the UNIX system from the LAN of NCSA using standard File Transfer Protocols (FTPs). BAC values are imputed only for drivers (FARS Person Type=01) in the vehicle classes as well as pedestrians and pedalcyclists (05<=FARS Person Type <=08) .

The programs to perform the multiple imputations are written in a combination of S-PLUS and Fortran programming languages. The software has been implemented on NHTSA's Digital Unix Alpha Server System. The CAT and MIX software packages comprise the two main components of the imputation programs.

CAT and MIX libraries

CAT and MIX are general-purpose packages that have been written for analyzing incomplete CATegorical and MIXed (continuous and categorical) data, respectively. They contain a large number of S-PLUS functions, S-PLUS help files, and Fortran source code. Basic versions of CAT and MIX are distributed at http://www.stat.psu.edu/~jls/misoftwa.html. These versions were extended with specially written S-PLUS functions to carry out the more tedious aspects of the procedures (e.g. stepwise modeling) automatically.

S-PLUS imputation programs

The five files listed below contain S-PLUS code for multiple imputation of BAC in FARS. The files must be executed in the stated order.

`bacvars.s`

This file reads in the FARS data, edits variables, and splits the data by vehicle class. The input FARS data is organized into the person, vehicle and crash level variables. The format and record layout for each of these files is listed in the appendix. The three input files are analytical files in the standard ASCII format that can be created from the appropriate SAS data sets. `Bacvars.s` reads the data from these files and extracts the information pertaining to drivers and non-occupants (pedestrians and pedalcyclists). The categorical variables are then coded into different levels for analyses by the next modules of the software.

`model1.s`

The set of functions in this program perform the stepwise variable selection of the first-stage loglinear model for each vehicle class. The input for this module is the S-PLUS data object created by `bacvars.s`. This program creates an S-PLUS file that contains the statistics for the selected variables for each of the vehicle classes.

`model2.s`

The power transformation for *log(BAC)* and variable selection for the second-stage regression model for each vehicle class are performed in this stage. The input file for this program is the S-PLUS object containing the selected variables and the significance-statistics for each of these variables. This module selects the variables needed to impute the non-zero values of BAC. The set of variables selected by these set of routines is a subset of the variables selected by `model1.s`.

`impute.s`

This is the program that performs the actual imputations of BAC given that BAC2 is not zero. The imputation procedure is performed for each of the vehicle classes. The ten imputations are then linked with the identifiers and an analytical file containing the ten imputations is created.

In addition, the file `sfuncs.s` contains S-PLUS functions needed by the other files. The function to perform the stepwise model selection is part of `sfuncs.s`. This program calculates the deviance test-statistics and the p-values for the variables that are entered into the forward-selection/backward-elimination stepwise procedure.

A users-manual has been documented to illustrate the program steps in detail. Sample analyses of the imputed data can be performed using the program `analyze.s`. This program contains functions that compare the estimates by the multiple imputation method and the estimates by the discriminant method (Klein, 1986).

Sample analyses in S-PLUS

The file `analyze.s` contains sample analyses of a multiply-imputed FARS dataset. Three types of analyses are presented:

(a) tabulations of proportions within the three traditional BAC categories (0, 0.01-0.09, 0.10+) by age of the driver/nonoccupant and vehicle class;

(b) tabulations of proportions within two additional categories (0.0-0.04, 0.05+) by age and vehicle class; and

(c) logit modeling of BAC>0.

Analysis (a) will facilitate comparisons of the new imputation method with the older method of Klein (1986), whereas (b) and (c) are examples of analyses that cannot easily performed under the older imputation method.

People Saving People

http://www.nhtsa.dot.gov