U.S. Department
of Transportation

**National Highway
Traffic Safety
Administration**

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

Research Note

October 1998

MULTIPLE IMPUTATION OF MISSING BLOOD ALCOHOL CONCENTRATION (BAC) VALUES IN FARS

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. The Fatality Analysis Reporting System (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 test 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. This research note describes the implementation of a new methodology that generates multiple imputations of missing Blood Alcohol Concentration values in FARS.

BACKGROUND

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). These probabilities, in conjunction with known alcohol test results, have been used by NHTSA since 1982 to estimate the level of alcohol involvement in fatal crashes across various crash, vehicle and person-level characteristics (e.g., alcohol was involved in 38.5 percent of all fatal crashes in 1997). The underlying reason for the selection of the cutpoints for the BAC groups (0.01 and 0.10) was to account for the general standards of intoxication while driving in the U.S. NHTSA defines a fatal traffic crash as being alcohol-related if either a driver or a nonoccupant (e.g., pedestrians or pedalcyclists) had a BAC of 0.01 or greater in a police-reported crash. Persons with BAC of 0.10 or greater involved in fatal crashes are considered intoxicated.

ISSUE AT HAND

While this approach is sufficient to report levels of alcohol involvement across the three categories of BAC, it does not provide a solution for simulating specific values of BAC across the full range of possible values. Estimation of discrete values of BAC, rather than probabilities of alcohol involvement, facilitates analyses by nonstandard boundaries of alcohol involvement (e.g., 0.08+). This issue assumes greater importance now that some states have reduced the legal level of intoxication to 0.08 and 0.02 for youthful drivers (under 21 years of age). The proposed methodology, developed by Dr. Donald B. Rubin (Harvard University) and Dr. Joseph L. Schafer (Penn State University), extends the current model by imputing ten values of BAC for each missing value. These ten values can be combined using simple computational macros to provide valid statistical inferences like variance, confidence intervals and deviation tests.

MULTIPLE IMPUTATION METHODOLOGY

The primary interest lies in simulating values for missing BACs of *actively involved* persons (drivers and nonoccupants). Following the earlier approach, the actively involved person is used as the basic unit of analysis and statistical models are constructed using other predictors (covariates) of BAC to estimate missing BAC as shown in Exhibit 1.

IMPUTATION STRATEGY

New computational methods under the *General Location Model (GLOM),* a multivariate probability model for data sets containing both continuous and categorical variables, have been applied to FARS to generate multiple imputations of missing BAC. The distribution of BAC may be regarded as *semicontinuous*: a substantial portion of BAC values are zero, and the remaining responses can be modeled as continuously distributed over the positive real number line. In order to fit

Exhibit 1
FARS 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[1]	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
RDWY	relation to roadway	1=not on roadway, 2=on roadway

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

semicontinuous BAC into the available GLOMs, BAC was reexpressed as two variables:
a dichotomous or binary indicator (BAC2):

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

and a *continuous variable* if BAC >0 equal to:

the *actual level* of BAC if BAC is nonzero and
missing if BAC=0.

By recoding the BAC as two variables, it is possible to model the relationships between BAC and the other covariates listed in Exhibit 1 using the GLOM. Semicontinuous BAC is then described by two-stage regression models. The first stage is a logit regression model to predict the probability of BAC=0. The second stage is a linear regression model to predict the mean response (often applying a log or power transformation) among those cases for which BAC>0. Modeling semicontinuous BAC in two stages is desirable not only from a statistical viewpoint, but also scientific reasons as the covariates that predict the probability of BAC being zero may be quite distinct from those that influence the actual value of BAC given that BAC>0.

It is well known from earlier work (Klein, 1986) that rates of alcohol involvement varied widely by vehicle class. The modeling procedure was done within each vehicle class and the imputed datasets from each vehicle class were combined to produce the completed dataset of multiply-imputed BAC values. Exhibit 2 lists the vehicle classes across which the imputation procedure was carried out.

MODEL SPECIFICATIONS

The imputation procedure consists of a two-stage model selection procedure and the actual imputation process that incorporates both these processes.

First Stage Model Selection

- First-stage loglinear model
- Model-fitting done by Schafer's ECM algorithm
- Select set of covariates that are significant in predicting dichotomous BAC.

Second Stage Model Selection

- Subset of 1st stage predictors is chosen by ordinary least-square stepwise regression of $g(BAC)=\log(BAC)^{\lambda+1}$ using the complete-case (CC) approach
- Box-Cox algorithm to estimate λ for the power transformation of $\log(BAC)^{\lambda+1}$
- Selects subset of covariates that are strong indicators of semicontinuos BAC.

Exhibit 2
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)	21-22, 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

Multiple Imputation

- Multiple Imputations created under GLOM
- Maximum Likelihood (ML) estimates are found using Schafer's ECM algorithm
- The ML estimates are used as starting values. New values of parameters are found using the Markov-Chain Monte Carlo (MCMC) algorithm
- The imputed values of g(BAC) are transformed back to their original scale
- Imputed datasets for each vehicle class are combined into SAS data sets and alcohol involvement estimates are generated using simple computational macros in SAS.

EFFECT ON ESTIMATES OF ALCOHOL INVOLVEMENT

The shift from the older to the newer method will involve some revision of historical alcohol estimates. 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>0.10) within important subclasses would be similar. In many respects, the pattern of results was indeed similar. 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 3 below. The only major difference was that the baseline rates of alcohol involvement were slightly but consistently higher under the new method**.** By 1990, this difference had declined to about 1.2 percent. At this time NCSA has used the discriminant model-based estimates for calendar year 1997. Multiply imputed datasets exist for calendar years 1982, 1986, 1990, 1993, 1995 and 1997. All the remaining intervening years' data will be added to the existing datastets. In preparation for the shift from the Discriminant Method to Multiple Imputation estimates, which is planned for the Summer of 1999 (covering calendar year 1998 crashes), NCSA will develop and distribute materials containing commonly used measures of alcohol involvement under both methods.

Exhibit 3: Trend of Percentage Alcohol Involvement in Fatal Crashes

Methods	1982	1986	1990	1993	1995	1996	1997
Mult Imptn.	59.1	53.8	50.7	44.6	42.5	42.2	39.6
Klein	56.7	51.7	49.4	43.5	41.3	40.9	38.5

For additional copies of this research note, please call Terry Klein at (202) 366-0328 or fax your request to (202) 366-7078. For questions regarding the data reported in this research note, contact Terry Klein [202-366-0328] of the National Center for Statistics and Analysis.

This research note and other general information on highway traffic safety may be accessed by Internet users at *http://www.nhtsa.dot.gov/people/ncsa.*