MULTIPLE IMPUTATION FOR THE FATAL ACCIDENT REPORTING SYSTEM

Daniel F. Heitjan, Penn State University and Roderick J. A. Little, UCLA
Daniel F. Heitjan, Penn State University College of Medicine, Hershey, PA 17033

1. INTRODUCTION.

The Fatal Accident Reporting System (FARS) gathers data on fatal traffic accidents in the United States. Data are collected at the state level for the National Highway Traffic Safety Administration (NHTSA), which publishes results in an Annual Report (NHTSA 1985). An important methodological concern with FARS data is the high level of nonresponse for certain variables, particularly blood alcohol content (BAC), an important variable in FARS analysis that is available only when a blood sample is taken. In 1987 NHTSA convened a panel of statisticians to provide the agency with guidance on imputation methods for FARS. The main goals articulated by the agency were a) to produce a public use data file that does not contain missing data; and b) to obtain a method for imputation of FARS variables that can provide valid standard errors for values thus imputed. This paper proposes a multiple imputation method for FARS data that largely meets these goals, and improves on existing practice. Some preliminary results on applying the method to 1985 FARS data are presented, and some further refinements are outlined.

With respect to the goals, we think that the imputation method adopted by NHTSA should aim to have the following properties (cf. Little 1988):

A) It should be appropriately *conditional*, in the sense that imputes for missing values in a record should condition on the values of observed variables for that case.

B) The method should take into account the *multivariate* nature of the nonresponse, that is, the fact that values are missing on more than one variable, with a general pattern of missing data.

C) Imputations should not distort marginal distributions and associations between the observed and imputed variables. To achieve this they should be *stochastic*, that is they should represent values from the predictive distribution of the missing variables, rather than means.

D) The imputed data set should allow the computation of *valid standard errors* of estimates of relevant parameters from complete-data methods applied to the filled-in data.

Note that the last goal differs somewhat from NHTSA's goal b), in that it concerns standard errors for parameter estimates rather than for the imputed values themselves, which are usually of secondary interest. The main limitation of these goals is the failure to account for so-called *nonignorable nonresponse*, which arises when the distribution of respondents and nonrespondents on an incomplete variable differs, even after conditioning on values of observed variables. Adjustments for nonignorable nonresponse are highly speculative, and are not considered here.

As a starting point we first consider Klein's (1986) discriminant analysis method, which represents a useful advance on previous approaches to the missing-data problem in the FARS. Klein fills in missing values in the BAC variable (which is in units of percent alcohol in the blood), by a) splitting BAC into three levels, BAC=0, 0<BAC<0.1 and BAC$\geq$0.1, and b) computing probabilities of falling into these three categories conditional on observed covariates, using discriminant analysis coefficients estimated from the complete cases. The splitting of BAC into three categories is motivated by the fact that the upper (0.1) limit is the legal limit in many states. The discriminant analysis provides a computationally straightforward method of incorporating knowledge about observed covariates into the imputations. That is, the imputations satisfy goal A) to some degree.

However, the method does have limitations. On a general level, the use of discriminant analysis to predict a categorical outcome requires multivariate normal assumptions for the predictors, which are not justified here since the majority of the predictors are categorical in nature. We propose a different treatment of the BAC variable based on logistic regression for the zero/non-zero dichotomy, and linear regression for the amount given that it is non-zero. This treatment avoids both the multivariate normal assumption in the discriminant analysis method and the grouping of the non-zero amounts into just two categories. We also attempt

to improve on the treatment of observed covariates in Klein's analysis by a more systematic treatment of interactions, and by some modeling of state differences, which are not adequately captured by the Klein models.

Turning to goal B), the Klein method is not multivariate in that it deals only with missing data in BAC, which is the most important of the FARS variables with substantial missing data, but is not the only one. In the 1985 FARS data three variables had substantial proportions of missing data: BAC (57.1%), police-reported alcohol involvement (DRINKING, 29.2%) and seatbelt use (MRESTR, 23.3%). Table 1 shows the pattern of missing data with respect to these three variables, based on a 10% sample of the data. The pattern does not exhibit a convenient monotonic form. Two other variables had small but non-negligible amounts of missing data, driving record (DRREC, 4.0%) and license status (LSTAT, 3.1%); these two variables tended to be missing or present together. Other variables had nonresponse rates of less than 2%.

TABLE 1
Missing Data Pattern,
Variables BAC, DRINKING and MRESTR
10% Sample of 1985 FARS, Drivers Only

Pattern			Frequency	Percent
R	D	B		
1	1	1	1422	24.6
1	1	0	1887	32.6
1	0	1	504	8.7
1	0	0	625	10.8
0	1	1	260	4.5
0	1	0	597	10.3
0	0	1	234	4.0
0	0	0	260	4.5
Total			5789	100.0

Note: Variable codes are: R=MRESTR *(use of manual restraints);* D=DRINKING *(officer-reported);* B=BAC *(blood alcohol content). Pattern codes are:* 1=*Present;* 0=*Missing.*

With respect to goal C), the Klein method is not stochastic as currently implemented, supplying for each missing BAC value a three-dimensional vector of proportions representing estimates of probabilities of falling into each of the three BAC categories. A consequence of imputing proportions is that valid standard errors of estimates cannot be readily constructed from the predicted probabilities supplied in the Klein method.

To make the Klein method stochastic in the sense of goal C), the analyst might impute a specific value of the BAC variable by sampling from this trinomial distribution. This is not hard to do, but requires of the analyst knowledge about missing-data methods. A simple improvement of the method would be to replace the vector of probabilities by a vector of three draws from the trinomial distribution. Then the variability of estimates from three analyses of the filled-in data using a) the set of first draws, b) the set of second draws and c) the set of third draws, would give an indication of added imputation variance.

This is a crude form of multiple imputation (Rubin 1987). With refinements to allow for the fact that parameters in the model are estimated, multiple imputation allows valid estimates of variance to be computed from the filled-in data sets, as sought in goal D). Multiple imputation is an important feature of our proposed approach.

2. THE PROPOSED PROCEDURE.

2.1 The Basic Method.

Our method is an adaptation of *predictive mean matching*, first proposed by Rubin (1986) in the context of statistical matching, and developed in the missing-data setting by Little (1988). Each incomplete case is matched to five complete cases with similar predicted values for the BAC variable. Each matched complete case supplies an imputation for every missing value in the incomplete case. Thus a multiply-imputed data set is achieved with $M = 5$ imputes for each missing value. Appropriate inferences for any particular analysis are achieved by repeating the analysis five times, with each of the five imputed values substituted for each missing value in turn. The resulting estimates and standard errors are combined by the simple procedures for analyzing multiply imputed data sets discussed in Rubin (1987), as summarized in Section 2 below.

The metric for matching complete and incomplete cases is based on regressions on the following variables:

$$\text{TEST1} = \begin{cases} 1, & \text{if BAC} > 0 \\ 0, & \text{if BAC} = 0 \end{cases}; \text{ and}$$

$$\text{TEST2} = \begin{cases} \text{BAC}, & \text{if BAC} > 0 \\ \text{missing}, & \text{if BAC} = 0 \end{cases}.$$

(Cf. Herzog and Rubin 1983). Specifically, the method involves the following steps:

1) Estimate the regression of TEST1 on all variables except BAC, including interactions between variables with large main effects if they add to the fit, and selecting variables by stepwise methods. The regressions are estimated using the complete cases, that is, cases with all variables observed. Exploratory regressions to select regressors can be linear; given the binary nature of TEST1, final regressions on TEST1 are preferably logistic if software is available for the large sample sizes involved. However this is not an essential requirement, since the regression equation is only used to supply a metric for matching complete and incomplete cases.

2) Regress TEST2 on all variables except BAC, using complete records with non-zero BAC levels. This regression can be linear in form; the histogram of TEST2 from a 10% sample of 1985 data (not shown) does not suggest the need for transformation, although attention might be paid to a small number of cases with particularly large BAC values, to ensure that they do not unduly influence the least squares estimates. Results on some preliminary work on the regressions of steps 1) and 2) are given in Section 3 below.

3) Compute predicted means of TEST1 for each complete case, and predicted means of TEST2 for each complete case with positive BAC level. This is a standard computation.

4) For each incomplete case, fill in missing values of regressor variables from 1) by their means from complete cases. These are not the final imputes for these missing values, but are introduced so that the single estimated regression equations for TEST1 and TEST2 can be applied to cases with missing regressor variables.

5) Compute predicted means for TEST1 and TEST2 for each incomplete case, by applying the regression equations computed in steps 1) and 2).

6) Find for each incomplete case i the five complete cases j with the smallest values of the distance function

$$d^2 = (\widehat{\text{TEST1}}(i) - \widehat{\text{TEST1}}(j))^2/\text{V}_1 + (\widehat{\text{TEST2}}(i) - \widehat{\text{TEST2}}(j))^2/\text{V}_2, \quad (1)$$

where $\widehat{\text{TEST1}}(k)$ and $\widehat{\text{TEST2}}(k)$ are predicted means for record k from steps 3) and 5), and V_1 and V_2 are the sample variances of the predictions of TEST1 and TEST2 respectively, based on the complete cases. The idea is to get matches that are close to the incomplete case with respect to predicted means of TEST1 and TEST2. Alternative choices of metric are discussed in Section 5.

7) From the five matched complete cases, randomly select with replacement a sample of size five. The five imputes for each missing value are the observed values of the missing variables from these five complete cases. These values overwrite the imputed means for missing regressors from Step 4).

Inference from multiply-imputed data is straightforward using the methods described in Rubin (1987). In particular, confidence intervals are computed in the following manner. Some number M (in our case, $M = 5$) of analyses are performed, where for analysis m the mth imputation of each missing value is substituted. For the mth analysis, let $\hat{\theta}_m$ be the estimate of a particular parameter θ of interest and $\hat{U}_m$ be its estimated variance, ignoring the effect of imputation. The final estimate of θ is $\hat{\theta} = \sum \hat{\theta}_m/M$, with estimated variance

$$T = \bar{U} + (1 + 1/M)B, \quad (2)$$

where $\bar{U} = \sum \hat{U}_m/M$ is the average variance within imputed data sets, and $B = \sum(\hat{\theta}_m - \hat{\theta})^2/(M-1)$ represents between-imputation variance. Large-sample inference for θ is based on comparing $(\hat{\theta} - \theta)/\sqrt{T}$ with a standard normal distribution. The fraction of information missing due to nonresponse is estimated as

$$\gamma = \frac{r}{r+1}, \text{ where } r = (1 + 1/M)B/\bar{U}. \quad (3)$$

Note that r is roughly the relative increase in variance due to nonresponse.

A refinement of this analysis replaces the normal reference distribution by a t-distribution with ν degrees of freedom, where

$$\nu = (M-1)(1 + 1/r)^2 \quad (4)$$

is based on a Satterthwaite approximation (Rubin and Schenker 1986; Rubin 1987, Section 3.1).

2.2 A Refinement.

A problem with the basic method described in Section 2.1 is that it does not account for uncertainty in the estimates of the parameters used in the regression equations that define the matching metric. Methods that ignore this source of variability have been termed "improper", and it is known that in general they lead to interval estimates that are too small (Rubin 1987). Thus we have developed an adaptation of predictive mean matching that accounts for this additional source of variability.

The basic idea of our method is to use different sets of regression parameter estimates to compute the matching metric. Each set of parameter estimates will be used to generate a single set of imputed data. The parameter estimates should be drawn from their posterior distribution given the data, which could be done approximately by drawing from a multivariate normal with mean at the MLE and covariance matrix based on the observed information matrix. Another alternative — and the one we have elected to use in this application — is to achieve draws of the parameters by drawing bootstrap samples from the complete cases and refitting the regression models for each. This second method is easier to implement, as it does not involve computing the incomplete data information matrix. The method can be viewed as a generalization of the approximate Bayesian Bootstrap, as described in Example 4.4 of Rubin (1987).

Specifically, a single set of imputations is created in the following four steps:

1) Draw a bootstrap sample from the set of cases having complete data on all variables.

2) Use the bootstrap sample to compute estimates of the parameters in the regressions of TEST1 and TEST2 on the selected predictors.

3) For each case with missing data, find the five complete cases that match it most nearly using the metric of equation (1).

4) Select one of these five matches at random, and use it as the donor for imputing values for all the variables that are missing on that incomplete case.

Any number of imputations may be created by repeating these steps. Such data sets should be analyzed using the methods of Section 2.1.

3. IMPLEMENTATION.

3.1 Regression Modeling.

To give some idea of how the methods described in Section 2 work we present results from an analysis based on a 10% sample of records on drivers from the 1985 FARS data set. Stepwise linear regressions for TEST1 and TEST2 were estimated for the limited set of regressors defined in Table 2. Results are displayed for the models with the lowest C_p statistic, in Table 3 for the regression of TEST1 and in Table 4 for the regression of TEST2. Conclusions can be summarized as follows:

1) For the binary variable TEST1, the R^2 for the "best" (lowest C_p) model is 0.668 (Table 3). As one might expect, police-reported alcohol involvement (DRINKING) is a very good predictor; with this variable missing the R^2 drops to 0.254. Adding all remaining variables to these models results in negligible increases in R^2.

2) For the amount variable TEST2, the R^2 for the best model is 0.200 (Table 4). When DRINKING is missing the R^2 drops to 0.135. Thus (as one might expect) the ability to predict amounts of blood alcohol is not as good as the ability to predict presence. Despite this result, many of the regressors had substantial and highly significant effects on the TEST2 means, so the regression is far from a waste of time. Again, adding the remaining variables resulted in a negligible increase in R^2.

3) The recoded state dummies 1 and 3 are highly significant in the TEST1 regressions, suggesting some differentials in presence of alcohol between states. State dummies are not very predictive in the TEST2 regressions, however. In fitting the models we used a different state coding system than we had used in the variable selection stage, and we included all state code dummies in the regression equation. The newer code system consists of ten codes based on the NHTSA administrative regions system.

4) The effects of accident severity and age are particularly significant. The effects of age are nonlinear, since the quadratic terms as well as the linear terms are important.

5) Other regressions (not shown) that added all two-way interaction terms between DRINKING, LSTAT, AGE, MLDA, SEV, SSS2, and

CLASS2 produced a small improvement in fit for TEST1, and a negligible improvement in fit for TEST2. (Interactions with state dummies were not tried, however). Hence while some modeling of interactions is advisable, preliminary analysis suggests that their impact will not be very great. An additive linear regression model without disaggregation of the sample may suffice, with only modest modifications to include significant interactions.

6) More work is needed to refine the variable definitions (particularly with regard to state groupings) and to incorporate insights from the earlier Klein analyses.

3.2 Inferences from the Imputed Data.

Five sets of imputations were created under each of the proper and improper predictive mean matching methods. Table 5 displays distributions of observed and imputed values of the five variables with non-negligible missing data, from the first imputed data set under the proper imputation method. The remaining nine data sets had the same distribution of observed values, but slightly different distributions of imputed values. Note that the observed and imputed distributions for the BAC and DRINKING variables are quite different, reflecting the strong effect of the covariates; in particular predicted BAC for cases when BAC is missing is much lower than observed BAC when BAC is present, presumably since the DRINKING variable is more frequently zero when BAC is missing (Table 5A). On the other hand, the proportion with a positive response on officer-reported drinking is higher among the imputed cases than among the cases with data observed for this variable (Table 5B).

In Table 5 we have presented only results of the proper imputations, because we noticed very little difference between the data sets imputed using the two methods. This was expected, since the objective of using the proper imputation method is not to change location estimates but to increase their variability to more accurately reflect posterior uncertainty. Thus the differences between the two methods should not be apparent when comparing the results of single imputations.

Table 6 displays 95% confidence intervals for six summary statistics based on the entire sample, computed by seven methods: a) standard analysis using only available cases, i.e. those for which the variable is available (Method AC); b) standard analysis based on the first imputed data set using the improper method (I1); c) the multiple imputation analyses of the first three (I3) and all five (I5) imputations using the improper method; d) the analogous inferences based on the imputations using the proper method (P1, P3 and P5). A column for the percentage of missing data is included; for method AC this is simply the fraction of missing cases for each variable, and for the other methods it is the percentage of missing information, computed using Equation (3). The last column shows the degrees of freedom for Rubin's t correction, computed using Equation (4). The t correction was used in computing confidence intervals for the analyses using three or five imputations.

The centers of the confidence intervals for the imputation methods I1, I3, I5, P1, P3 and P5 are very similar, and are close to method AC except for the alcohol variables, for which method AC yields much higher estimates. The widths of the confidence intervals for $M = 3$ and $M = 5$ are always greater than the widths for $M = 1$, reflecting the fact that multiple imputation includes the component of between-imputation variance. For the BAC variables and DRINKING the widths of the multiple imputation intervals (except for P3) are narrower than the intervals from available cases, reflecting the gain of information from using incomplete cases. There are discrepancies from the general patterns, however, which we expect would disappear if more imputations were done. The degrees of freedom and percentage information losses are quite variable for these small values of M. Notice that the variables that do not directly measure alcohol consumption are less affected by the imputation than are the alcohol variables. We believe that two factors are responsible for this: First, the proportions of missing data are smaller for these variables, so that the effect of imputation is necessarily smaller. Second, these variables were not used as responses in computing the distances, so the regression models appear to predict these variables only indifferently.

4. REMARKS AND REFINEMENTS.

1) In determining matches, the proposed method gives weight to variables that are highly predictive of BAC. As such, the method places primary emphasis on the BAC variables in supplying the best possible imputations. This seems appropriate given the extent of nonresponse of this variable and its importance for analysis. The method does supply imputes for the other variables that are in line with the stated goals in Section 1. One possible variation of the method would be to also compute regressions on seatbelt use, and then include predicted seatbelt use in the metric (1). The impact of this change would depend on the extent to which the prediction equation for seatbelt use differed from the prediction equations for TEST1 and TEST2.

2) The choices of five matches for each imputed value and five imputes for each missing value are somewhat arbitrary. There is a law of diminishing returns as the number of imputes increases; thus two imputes are much better than one, three are somewhat better than two, and gains from more than three are somewhat smaller. In fact, the decision to create five imputes was based on instabilities observed in an earlier version of Table 6, in which only $M = 1$, 2 and 3 were considered. Knowledge of γ can be used to guide the choice of M (cf. Rubin, 1987). However estimates of γ and the degrees of freedom are very unreliable if M is small, although the sampling properties of multiple imputation procedures are quite good even if M is only 2 or 3 (Rubin and Schenker 1986).

3) The metric of Equation (1) has no strong theoretical basis, and might be profitably modified, for example by allowing for covariation between the predicted means or by replacing V_1 and V_2 by residual variances from the regressions. A more refined alternative to matching on TEST1 and TEST2 simultaneously is the following two-stage procedure: a) fill in $\widehat{\text{TEST1}} = 0$ or 1 for TEST1 by Bernoulli sampling with probability $P(\text{TEST1} = 1|x)$ estimated from the regression of TEST1 on x. b) If $\widehat{\text{TEST1}} = 1$, TEST2 is imputed by matching to a case with TEST1=1 based on the predicted means $\widehat{\text{TEST2}}$. If $\widehat{\text{TEST1}} = 0$, then TEST2 is imputed as zero, and other missing values are imputed by matching to a complete case with TEST1= 0. The metric could be based on predicted means from regression on another variable with a substantial fraction of missing values, such as seatbelt use.

This method has theoretical advantages over the basic method in Section 3, but note that the predicted probabilities from the TEST1 regression are used directly, so that this regression needs more careful modeling than in our procedure, where the regression model is used only to determine a metric for matching. Also in repeating the procedure to obtain several imputes, a complete case can be matched to cases with TEST1=1 and TEST1=0, so the matching procedure is a bit more complicated to implement than the single step procedure based on Equation (1).

5. SUMMARY.

We have proposed a method for multivariate multiple imputation for missing values in FARS data sets. The method involves only standard regression software and a fairly easily programmed matching algorithm, makes good use of available information on incomplete records, fills in all the missing values, and largely meets the goals of a good imputation procedure outlined in Section 1. In particular the method provides a means for assessing the impact of imputation by presenting a set of five imputes for each missing value. A preliminary implementation on a subsample of FARS data illustrated the fact that observed covariates have some predictive power for imputing missing BAC values, particularly when police reported alcohol involvement is available.

ACKNOWLEDGEMENTS.

This article is based in part on an earlier paper that was partially supported by NHTSA. Initial computations in that paper were carried out by Hong-Lin Su of the Department of Biostatistics, UCLA.

REFERENCES.

Herzog, T. N. and Rubin, D. B. (1983), "Using Multiple Imputations to Handle Nonresponse in Sample Surveys", in *Incomplete Data in Sample Surveys, Volume 2: Theory and Bibliographies*, W. G. Madow, I. Olkin and D. B. Rubin, editors, 210–248, New York: Academic Press.

Klein, T. M. (1986), "A Method for Estimating Posterior BAC Distributions for Persons Involved in Fatal Traffic Accidents", Technical Report, National Highway Traffic Safety Administration.

Little, R. J. A. (1988), "Missing Data Adjustments in Large Surveys", *Journal of Business and Economic Statistics*, 6, 287–297 (with discussion).

Little, R. J. A. and Rubin, D. B. (1987), *Statistical Analysis with Missing Data*, New York: John Wiley.

NHTSA (1985), *Fatal Accident Reporting System 1985*, National Highway Traffic Safety Administration, U. S. Department of Transportation.

Rubin, D. B. (1986), "Statistical Matching Using File Concatenation with Adjusted Weights and Multiple Imputations", *Journal of Business and Economic Statistics*, 4, 87–94.

Rubin, D. B. (1987), *Multiple Imputation for Nonresponse in Surveys*, New York: John Wiley.

Rubin, D. B. and Schenker, N. (1986), "Multiple Imputation for Interval Estimation from Simple Random Samples with Nonignorable Nonresponse", *Journal of the American Statistical Association*, 81, 361–374.

TABLE 2
Regressor Variable Definitions

Factor	Type	Variables in Regression
Age	Continuous	AGE and AGE^2
Driver Record	Continuous	DRREC=Driving Record Score
Day/Night	Binary	HOUR: 1=11-23 (Day); 0=0-10 (Night)
Junction	Binary	JUNCTION: 1=At a Junction; 0=Not at a Junction
Land Use	Binary	LAND_USE: 1=Urban: 0=Rural
License Status	Binary	LSTAT: 1=Valid: 0=Not Valid
Manual Restraint	Binary	MRESTR: 1=Used, 0=Not Used
Minimum Age	Binary	MLDA: 1=Above Minimum Legal Driving Age; 0=Other
Non-Occupant Accident	Binary	NOC_ACC: 1=Non-Occupant Involved: 0=Other
Position	Binary	ROADWAY: 1=On Road; 0=Off Road
Severity	Binary	SEV: 1=Fatal to Driver; 0=Not Fatal
Sex	Binary	SEX: 1=Male: 0=Female
State	Nine Categories, Defined at Foot of Table	
Driver's Vehicle Class	Four Categories	Reference=Passenger Car; CLASS1=Dummy for Motorcycles; CLASS2=Dummy for LTV's; CLASS3=Dummy for Other Vehicles
Driver's Vehicle Role 1	Three Categories	Reference=No Collision; IMPACT1=Dummy for Striking; IMPACT2=Dummy for Struck
Driver's Vehicle Role 2	Three Categories	Reference=Single Vehicle; SSS1=Dummy for Multiple Vehicle Striking; SSS2=Dummy for Multiple Vehicle Struck
Weekend	Binary	WK: 1=Weekend; 0=Weekday

Original state codes (used in the variable selection phase):

Reference: California, Hawaii.
STATE1: Texas.
STATE2: Florida.
STATE3: DC, Maryland, N. Carolina, Virginia.
STATE4: Connecticut, Delaware, New Jersey, New York, Pennsylvania, Rhode Island, W. Virginia.
STATE5: Illinois, Indiana, Iowa, Michigan, Minnesota, Ohio, Oregon, Washington, Wisconsin.
STATE6: Alabama, Arkansas, Georgia, Kansas, Louisiana, Mississippi, Missouri, Oklahoma, S. Carolina, Tennessee.
STATE7: Maine, Massachusetts, New Hampshire, Vermont.
STATE8: Arizona, Alaska, Colorado, Idaho, Montana, Nebraska, Nevada, New Mexico, N. Dakota, S. Dakota, Utah, Wyoming.

The revised state codes are equivalent to the NHTSA administrative regions system.

TABLE 3
Regressions on TEST1, Variable DRINKING Present:
Model with Smallest C_p Statistic

$R^2 = 0.66761287$, $C_p = 17.49568206$

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	17	262.74013707	15.45530218	185.49	0.0001
Error	1570	130.81150021	0.08331943		
Total	1587	393.55163728			

	B Value	Standard Error	Type II SS	F	Prob>F
Intercept	0.08264459				
STATE2	-0.06552386	0.02269629	0.69444101	8.33	0.0039
STATE3	-0.08646418	0.02689448	0.86117725	10.34	0.0013
STATE4	0.06607334	0.02221719	0.73692051	8.84	0.0030
STATE8	0.05212950	0.02564012	0.34440813	4.13	0.0422
AGE	0.00565553	0.00249281	0.42885861	5.15	0.0234
CLASS2	-0.04017512	0.01955167	0.35179760	4.22	0.0401
CLASS3	-0.09553556	0.02898996	0.90485927	10.86	0.0010
WK	0.05702176	0.01539734	1.14270980	13.71	0.0002
HOUR	-0.03155930	0.01542381	0.34883303	4.19	0.0409
SEV	0.05390508	0.01620904	0.92149125	11.06	0.0009
JUNCTION	0.03589487	0.01724526	0.36097013	4.33	0.0376
LSTAT	-0.05810747	0.02182053	0.59085340	7.09	0.0078
ROADWAY	-0.08620318	0.01725056	2.08058966	24.97	0.0001
AGE2	-0.00007736	0.00002699	0.68422268	8.21	0.0042
SEX	0.09704962	0.01890089	2.19669358	26.36	0.0001
MLDA	-0.04539412	0.02628374	0.24852556	2.98	0.0844
DRINKING	0.71818565	0.01657084	156.50590006	1878.38	0.0001

TABLE 4
Regressions on TEST2, Variable DRINKING Present:
Model with Smallest C_p Statistic

$R^2 = 0.20005701$, $C_p = 3.70199046$

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	11	12708.56258801	1155.32387164	19.46	0.0001
Error	856	50816.14248112	59.36465243		
Total	867	63524.70506912			

	B Value	Standard Error	Type II SS	F	Prob>F
Intercept	-4.25698123				
STATE2	-1.59636755	0.89248209	189.93049380	3.20	0.0740
STATE7	-4.49034810	2.16581740	255.17861413	4.30	0.0384
AGE	0.81996554	0.10214391	3825.55246095	64.44	0.0001
CLASS1	-3.36222002	0.83729277	957.25029695	16.12	0.0001
HOUR	0.98034058	0.54493750	192.12704212	3.24	0.0724
SSS2	-3.54121634	0.99117888	757.75491994	12.76	0.0004
SEV	4.18831011	0.61252676	2775.59090007	46.75	0.0001
LSTAT	-2.65762513	0.68715818	887.97742554	14.96	0.0001
NOC_ACC	-1.87882174	1.05341285	188.84342175	3.18	0.0748
AGE2	-0.00916787	0.00125448	3170.56723096	53.41	0.0001
DRINKING	6.28134810	0.89772369	2906.35249032	48.96	0.0001

TABLE 5
Frequency Distributions of Observed and Imputed Values of Incomplete Variables from First Imputed Data Set

Table 5A — BAC × 100

	Counts			Percentages		
Value	Imputed	Observed	Total	Imputed	Observed	Total
0	2585	1167	3752	76.7	48.2	64.8
1-5	168	159	327	5.0	6.6	5.6
6-10	135	165	300	4.0	6.8	5.2
11-15	180	258	438	5.3	10.7	7.6
16-20	147	318	465	4.4	13.1	8.0
21-25	87	177	264	2.6	7.3	4.6
26-30	43	110	153	1.3	4.5	2.6
31-35	22	49	71	0.7	2.0	1.2
36-40	1	12	13	0.0	0.5	0.2
GT40	1	5	6	0.0	0.2	0.1
Total	3369	2420	5789	100.0	100.0	100.0

Table 5B — DRINKING

	Counts			Percentages		
Value	Imputed	Observed	Total	Imputed	Observed	Total
0	843	2938	3781	51.9	70.5	65.3
1	780	1228	2008	48.1	29.5	34.7
Total	1623	4166	5789	100.0	100.0	100.0

Table 5C — LSTAT

	Counts			Percentages		
Value	Imputed	Observed	Total	Imputed	Observed	Total
0	24	644	668	12.5	11.5	11.5
1	168	4953	5121	87.5	88.5	88.5
Total	192	5597	5789	100.0	100.0	100.0

Table 5D — DRREC

	Counts			Percentages		
Value	Imputed	Observed	Total	Imputed	Observed	Total
0	147	3120	3267	61.5	56.2	56.4
1	46	1180	1226	19.2	21.3	21.2
2	24	600	624	10.0	10.8	10.8
3	4	283	287	1.7	5.1	5.0
4	7	167	174	2.9	3.0	3.0
5	1	72	73	0.4	1.3	1.3
6–7	9	85	94	3.8	1.5	1.6
8–9	1	26	27	0.4	0.5	0.5
10–12	0	12	12	0.0	0.2	0.2
13–15	0	3	3	0.0	0.1	0.1
16–18	0	0	0	0.0	0.0	0.0
GT19	0	2	2	0.0	0.0	0.0
Total	239	5550	5789	100.0	100.0	100.0

Table 5E — MRESTR

	Counts			Percentages		
Value	Imputed	Observed	Total	Imputed	Observed	Total
0	1066	3541	4607	78.9	79.8	79.6
1	285	897	1182	21.1	20.2	20.4
Total	1351	4438	5789	100.0	100.0	100.0

TABLE 6
Confidence Intervals for Means and Proportions of Incomplete
Variables from Complete Cases and Imputed Data Sets

Table 6A — 100 × Mean BAC Level

Method	Mean	±	% Missing	DF
AC	8.403	0.407	58.2	—
I1	5.524	0.230	—	—
I3	5.547	0.285	27.9	25
I5	5.531	0.267	23.8	70
P1	5.322	0.228	—	—
P3	5.465	0.723	73.3	3
P5	5.492	0.398	57.9	11

Table 6B — Percent with BAC Level ≥ 0.10

Method	Mean	±	% Missing	DF
AC	39.92	1.95	58.2	—
I1	26.41	1.14	—	—
I3	26.41	1.29	18.8	56
I5	26.43	1.24	14.7	186
P1	25.31	1.12	—	—
P3	25.94	3.52	72.8	3
P5	25.99	1.89	55.9	12

Table 6C — Percent with Officer-Reported Drinking

Method	Mean	±	% Missing	DF
AC	29.48	1.38	28.0	—
I1	33.94	1.22	—	—
I3	34.08	1.25	4.8	874
I5	34.14	1.25	4.9	1643
P1	34.69	1.23	—	—
P3	34.55	1.25	4.3	1071
P5	34.46	1.38	19.0	110

Table 6D — Percentage of Drivers Having Valid Licenses

Method	Mean	±	% Missing	DF
AC	88.49	0.84	3.3	—
I1	88.53	0.82	—	—
I3	88.47	0.85	6.4	487
I5	88.51	0.84	4.8	1729
P1	88.46	0.82	—	—
P3	88.50	0.84	4.1	1178
P5	88.48	0.83	2.2	8066

Table 6E — Mean Driving Record

Method	Mean	±	% Missing	DF
AC	0.94	0.04	4.1	—
I1	0.94	0.04	—	—
I3	0.94	0.04	4.9	831
I5	0.94	0.04	4.5	1992
P1	0.94	0.04	—	—
P3	0.94	0.04	5.8	601
P5	0.94	0.04	5.3	1438

Table 6F — Percentage Using Manual Restraints

Method	Mean	±	% Missing	DF
AC	20.21	1.18	23.3	—
I1	20.64	1.04	—	—
I3	20.53	1.13	13.9	103
I5	20.50	1.09	7.9	636
P1	20.42	1.04	—	—
P3	20.60	1.11	11.2	160
P5	20.54	1.10	9.8	419

Note: Method AC=available cases only; Method IM=Improper, M imputations;
Method PM=Proper, M imputations.