CASE STUDY: IMPUTATION OF BAC IN FATAL COLLISIONS

Patrick Boily[1,2,3]

Abstract

Alcohol is often a factor in fatal collisions, but the presence of alcohol in the blood cannot always be confirmed until an autopsy is performed. In this case study, we present a two-stage multiple imputation algorithm that imputes the blood alcohol content levels of drivers involved in fatal collisions, based on a number of descriptive collision variables. We then provide an artificial example that illustrates the algorithm, as well as the result of the imputation for Ontario in 2007.

Keywords

Multiple imputation, logistic regression, BAC, case study.

Funding Acknowledgement

This work was conducted when the author was an employee of Transport Canada.

[1]Department of Mathematics and Statistics, University of Ottawa, Ottawa
[2]Data Action Lab, Ottawa
[3]Idlewyld Analytics and Consulting Services, Wakefield, Canada
Email: pboily@uottawa.ca

Contents

1. Introduction

When fatal collisions occur, it is frequently the case that at least one of the drivers (or one of the pedestrians/cyclists, as the case may be) involved in the collision was affected by alcohol [5,6]. Since breathalyzer tests cannot be conducted on deceased individuals, the presence of alcohol in the blood cannot be confirmed until the coroner's report becomes available. For various reasons, this report is not always immediately available: in certain cases, it can take up to a year before the Blood Alcohol Concentration (BAC) level makes its way to the collision databases [1]. Rather than waiting for this process to take place, data analysts often resort to imputation methods in order to make an informed guess as to the value of the BAC in fatal collisions. Once the imputed values are supplanted by the coroner's values, BAC-dependent preliminary analyses with the imputed values can easily be re-conducted with the actual values to obtain up-to-date results.

Policy makers require fast and reliable analysis results. If the method used to impute the BAC level is based on sound statistical techniques, the preliminary analysis using imputed values is likely to give results that are comparable to the eventual results obtained using the true data, saving precious time in the quest for road safety improvements.

In this case study, we present the algorithm used by the Evaluation & Data Systems Division of the Road Safety and Motor Vehicle Directorate at Transport Canada. It imputes the BAC level in fatal collisions based on a number of descriptive (or explanatory) variables linked to the collisions. Details are provided in the sections on Data Preparation and Methodology, together with an artificial example that illustrates the method. A discussion of the BAC imputation results for 2007 is then provided, together with some final comments regarding the algorithm and how it could be improved.

2. Statistical Imputation

Ideally, every record of a data set would be complete. In practice, this is not always the case: observation times may be missed, values may be unavailable, data may get corrupted by machine errors, etc. The more holes in a data set, the lesser its utility.

Imputation methods are processes by which missing values are substituted by reasonable "guesses". Statistical imputation uses probability theory to provide these "guesses." The number of imputation strategies is vast, ranging from classical hot-deck and cold-deck imputation to the more modern methods of logistic regression, nearest neighbours imputation and multiple imputation. Certain methods might give better results when adapted to certain types of data sets, but in general, we cannot speak of THE method for BAC imputation.

Two previously published imputation methods have influenced our approach: the routine used by the National Highway Traffic Safety Administration (NHTSA) to impute BAC in FARS [4], and the multivariate technique for imputation using a sequence of regression models of Raghunathan, Lepkowski, Van Hoewyk and Solenberger [3].

The NHTSA approach [4] uses a two-stage model where zero/non-zero BAC status is first imputed through some multivariate procedure, and, conditional on non-zero BAC, a general linear model (together with appropriate transformations) is used to impute ten BAC values for each missing value, allowing valid statistical inferences on variances and confidence intervals to be drawn. The main drawback of this method, however, is that the values of some explanatory variables are missing for a large number of records. For each variable, missing values were treated as belonging to a separate category: that of ’missing value’. As there may be many disparate reasons to explain why different records are missing a given variable, this may lead to a loss of information, which translates into a less powerful imputation method.

In the case of multiple missing values in the explanatory variables, [3] uses a sequence of regression models. The missing values for each explanatory variable are imputed as follows: first, the explanatory variable Y_1 with the fewest missing values is imputed to $\tilde{Y}_1$ using the explanatory variables X with no missing values. Then the explanatory variable Y_2 with the next fewest missing values is imputed to $\tilde{Y}_2$ using the explanatory variables $\{X, \tilde{Y}_1\}$. The process continues in sequence until the last remaining explanatory variable with missing values Y_m is imputed to $\tilde{Y}_m$ using $\{X, \tilde{Y}_1, \ldots, \tilde{Y}_{m-1}\}$. The main drawback of this method is that some information might be “hiding” in $\{Y_2, Y_3, \ldots, Y_m\}$ which, combined with the information found in X, could provide a better imputation for Y_1.

Transport Canada’s BAC Imputation Algorithm (TCBACIA) retains the two-stage model and multiple imputation of [4], as well as sequential regression from [3], but it does so in a manner that eliminates the drawbacks associated with either of the methods, as described above.

3. Data Preparation

TCBACIA imputes a likely BAC level for drivers and pedestrians involved in fatal collisions for a given year based on a number of variables from the National Collision Database (NCDB) as well as data from the Traffic Injury Research Foundation (TIRF) over a preceding five-year period. Once all records involving non-fatal collisions and all records involving non-drivers or non-pedestrians in fatal collisions have been removed, two BAC-linked dependent variables can clearly be identified (one categorical and one semi-continuous).

1. Was BAC equal to 0, or was it greater than 0? (TEST)
2. What was the BAC level? (P_BAC1F)

In a preliminary phase [2], a multivariate analysis of variance (MANOVA) identified the following independent (or explanatory) NCDB variables as having a significant effect on the dependant variables:

- whether the record identifies a driver or a pedestrian (P_PSN);
- the sex (P_SEX) and age (P_AGE) of the deceased;
- whether a safety device was worn (P_SAFE) by the deceased;
- the hour (C_HOUR) and weekday (C_WDAY) when the collision occurred;
- the number of vehicles/pedestrians involved in the collision, and (C_VEHS)
- various contributing factors (V_CF1—V_CF4) as determined by police officers on the scene.

Some of the explanatory variables classes were originally grouped in order to insure meaningful MANOVA. The actual data is thus categorical.

Variable	Classification
P_PSN_GR	1 = ’Driver’ 2 = ’Pedestrian/Cyclist’ . = ’Missing’
C_WDAY_GR	1 = ’Weekday’ 2 = ’Weekend’ . = ’Missing’
C_HOUR_GR	1 = ’00:00 to 05:59’ 2 = ’06:00 to 09:59’ 3 = ’10:00 to 15:59’ 4 = ’16:00 to 19:59’ 5 = ’20:00 to 23:59’ . = ’Missing’
C_VEHS_GR	1 = ’One vehicle involved’ 2 = ’Two vehicles involved’ 3 = ’Three or more vehicles involved’ . = ’Missing’
P_SEX_GR	1 = ’Male’ 2 = ’Female’ . = ’Missing’
P_AGE_GR	1 = ’<= 19’ 2 = ’20-29’ 3 = ’30-39’ 4 = ’40-49’ 5 = ’50-59’ 6 = ’>=60’ . = ’Missing’
P_SAFE_GR	1 = ’No Safety Device Used’ 2 = ’Safety Device Used’ 3 = ’Not Applicable’ . = ’Missing’
V_CF_GR	1 = ’Alcohol Deemed a Contributing Factor by Police Officer’ 2 = ’Alcohol not Deemed a Contributing Factor by Police Officer’ . = ’Missing’

One might think that V_CF_GR as defined above would be a very significant predictor of BAC, but preliminary analyses show that it is not any more significant when taken individually than any of the other explanatory variables that have been retained.

4. Methodology

So how does our algorithm differ from [3, 4]? Roughly speaking, TCBACIA inflates the original data set using replicates (analogues of multiple imputation), then uses sequential logistic regression on the entire data set in order to impute the missing values of explanatory variables upon which the two-stage model is built. The data set is eventually deflated down to its original size. The process is described in detail in this section.

Inflating the Data Set

Suppose the original data set contains n records. We start by replicating the data set k times, where $k \geq 1$ is some integer. The value of k is selected in order to create data sets which will be large enough for whatever imputation method was chosen to produce statistically meaningful results. If the original data set contained n records, the replicated data set contains kn records.

For data sets with n large or without systematic patterns in the missing values, small values of k can be used; when n is smaller, larger values of k must be used. For instance, using SAS 9.2's proc logit to impute BAC values (according to the method which will be described below) for real-life Ontario fatal collision data from 2000 to 2007 with $n \approx 10000$, a value of $k = 9$ was found to eliminate all parametric convergence problems.

Step 1—1: First First-Order Imputation

Let m be the number of explanatory variables. Amongst the m_1 explanatory variables with missing values, find the one with the fewest, and denote it by Y_{α_1}. (In the event of a tie, Y_{α_1} can be selected at random.)

Let W_{α_1} denote all records for which none of the non-Y_{α_1} values are missing. We can further subdivide W_{α_1} into $W^{\text{imp}}_{\alpha_1}$ and $W^{\text{train}}_{\alpha_1}$, depending on whether the value of Y_{α_1} is missing or not for those records.

Next, impute the missing values of Y_{α_1} in $W^{\text{imp}}_{\alpha_1}$ using $W^{\text{train}}_{\alpha_1}$ as a training set. Any acceptable imputation method can be used. Considering the categorical nature of the data points, generalised (or multinomial) logistic regression seems specially well-suited to the task.

Step 1—2: Second First-Order Imputation

Amongst the remaining explanatory variables, find the one with the next fewest number of missing values and denote it by Y_{α_2}.

Let W_{α_2} denote all records for which none of the non-Y_{α_2} values are missing; we can further subdivide W_{α_2} into $W^{\text{imp}}_{\alpha_2}$ and $W^{\text{train}}_{\alpha_2}$ as above. Impute the missing values of Y_{α_2} in $W^{\text{imp}}_{\alpha_2}$ using $W^{\text{train}}_{\alpha_2}$ as a training set.

Step 1—m_1: Last First-Order Imputation

This process is repeated until the imputation of missing values of the last remaining explanatory variable (and the one with the largest number of missing values in the original data set), denoted by $Y_{\alpha_{m_1}}$, in $W^{\text{imp}}_{\alpha_{m_1}}$ using $W^{\text{train}}_{\alpha_{m_1}}$ as a training set.

By construction, a record with two or more missing values will never be involved in the preceding steps; consequently, after first-order imputation, any record with missing values will have no fewer than two missing values.

Step 2—1: First Second-Order Imputation

We now alter the data set slightly by appending m_2 new variables, obtained by crossing all the distinct pairs of explanatory variables which still have missing values. Amongst those new explanatory variable, denote the one with the fewest number of missing values by Y_{α_1,β_1}.

Let W_{α_1,β_1} denote all records for which none of the non-Y_{α_1,β_1} values are missing. We can further subdivide W_{α_1,β_1} into $W^{\text{imp}}_{\alpha_1,\beta_1}$ and $W^{\text{train}}_{\alpha_1,\beta_1}$, depending on whether the Y_{α_1,β_1} values are missing or not for those records. Impute the missing values for Y_{α_1,β_1} in $W^{\text{imp}}_{\alpha_1,\beta_1}$ using $W^{\text{train}}_{\alpha_1,\beta_1}$ as a training set.

Step 2—2: Second Second-Order Imputation

Amongst the remaining crossed explanatory variables, find the one with the next fewest number of missing values and denote it by Y_{α_2,β_2}.

Let W_{α_2,β_2} denote all records for which none of the non-Y_{α_2,β_2} values are missing; we can further subdivide W_{α_2,β_2} into $W^{\text{imp}}_{\alpha_2,\beta_2}$ and $W^{\text{train}}_{\alpha_2,\beta_2}$ as above. Impute the missing values for Y_{α_2,β_2} in $W^{\text{imp}}_{\alpha_2,\beta_2}$ using $W^{\text{train}}_{\alpha_2,\beta_2}$ as a training set.

Step 2—m_2: Last Second-Order Imputation

This process is repeated until the imputation of missing values of the last remaining crossed explanatory variable, denoted by $Y_{\alpha_{m_2},\beta_{m_2}}$, in $W^{\text{imp}}_{\alpha_m,\beta_{m_2}}$ using $W^{\text{train}}_{\alpha_m,\beta_{m_2}}$ as a training set. By construction, a record with three or more missing values will never be involved in the preceding steps; consequently, after second-order imputation, any record with missing values will have no fewer than three such missing values.

Continuation

This process is repeated with triplets of explanatory variables, then quadruplets, and so on, until the data set contains no record with missing values of the explanatory variables.

Imputation of the Dependent Variables Z_1 and Z_2

Denote the two dependent variables described in the previous section by Z_1 (BAC > 0 or not) and Z_2 (BAC level).

Let $W_{Z_1}^{\text{imp}}$ and $W_{Z_1}^{\text{train}}$ denote the records for which the value of Z_1 is missing and the records for which it is available, respectively. The missing values of the categorical variable Z_1 in $W_{Z_1}^{\text{imp}}$ can be imputed as above, using $W_{Z_1}^{\text{train}}$ as a training set.

The variable Z_2 is seen as semi-continuous because a substantial proportion of BAC values are zero while the non-zero responses are continuously distributed over the positive real number line within some acceptable range, say $(0,A)$, where $A > 0$ is some upper BAC limit.

Our model is thus a two-stage model where zero/non-zero BAC status (i.e. the value of Z_1) is first imputed through some procedure (*e.g.* logistic regression), and, conditional on $Z_1 = 1$ (*i.e.* BAC > 0), some other model (such as a general linear model) can be used to impute the actual BAC level.

For all records with $Z_1 = 1$, let $W_{Z_1=1,Z_2}^{\text{imp}}$ and $W_{Z_1=1,Z_2}^{\text{train}}$ denote the records for which value of Z_2 is missing and the records for which it is available, respectively. The missing values of the semi-continuous variable Z_2 in $W_{Z_1=1,Z_2}^{\text{imp}}$ can be imputed using some general linear model built upon $W_{Z_1=1,Z_2}^{\text{train}}$.

Deflating the Data Set

At this stage, for each of the n original records, we have k values of Z_1 and Z_2; let us denote the j^{th} replicate of the i^{th} record by $Z_1^{j,i}$ and $Z_2^{j,i}$. Pick some threshold $a \in (0,1)$ and define

$$\overline{Z_1^i} = \frac{1}{n}\sum_{j=1}^{l} Z_1^{j,i} \quad \text{and} \quad \overline{Z_2^i} = \frac{\sum_{j=1}^{k} Z_1^{j,i} Z_2^{j,i}}{n\overline{Z_1^i}}.$$

Then the actual imputed values for the i^{th} record are

$$Z_1^i = \begin{cases} 1 & \text{if } \overline{Z_1^i} > a \\ 0 & \text{else} \end{cases} \quad \text{and} \quad Z_2^i = \begin{cases} \overline{Z_2^i} & \text{if } \overline{Z_1^i} > a \\ 0 & \text{else} \end{cases}$$

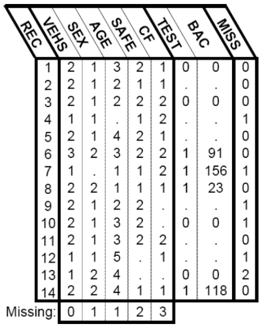
The threshold value a has the following interpretation: if more than $100a\%$ of the replicates for a given record have been imputed to have non-zero BAC, that record is reported to have non-zero BAC, and its BAC level is the average of the BAC levels taken over all its non-zero BAC replicates. If the "cost" associated with false positives (imputed BAC > 0 but actual BAC = 0) is the same as that of a false negative (imputed BAC = 0 but actual BAC > 0), then $a = 0.5$ is a good choice.

5. Artificial Example

The following simplified artificial example will be used to illustrate the method presented in the previous section.

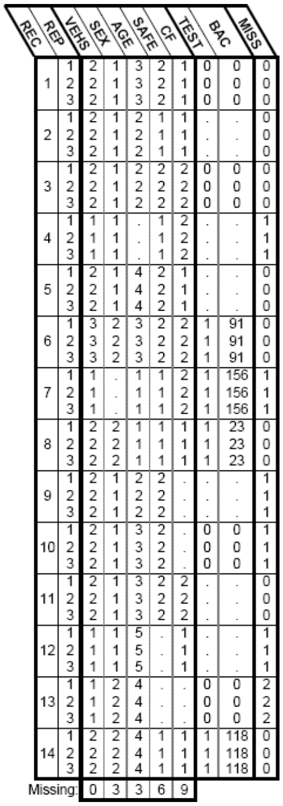
Inflating the Data Set

The database consists of the $n = 14$ records shown in the table below.



REC	VEHS	SEX	AGE	SAFE	CF	TEST	BAC	MISS
1	2	1	3	2	1	0	0	0
2	2	1	2	1	1	.	.	0
3	2	1	2	2	2	0	0	0
4	1	1	.	1	2	.	.	1
5	2	1	4	2	1	.	.	0
6	3	2	3	2	2	1	91	0
7	1	.	1	1	2	1	156	1
8	2	2	1	1	1	1	23	0
9	2	1	2	2	.	.	.	1
10	2	1	3	2	.	0	0	1
11	2	1	3	2	2	.	.	0
12	1	1	5	.	1	.	.	1
13	1	2	4	.	.	0	0	2
14	2	2	4	1	1	1	118	0
Missing:	0	1	1	2	3			

In the example, each record is replicated $k = 3$ times. The replicated records $X_{i,j}$, $i = 1,\ldots,14$, $j = 1,\ldots,3$, have five categorical explanatory variables: Y_1 (`VEHS`), Y_2 (`SEX`), Y_3 (`AGE`), Y_4 (`SAFE`) and Y_5 (`CF`), as well as a categorical dependant variable Z_1 (`TEST`) and a semi-continuous dependent variable Z_2 (`BAC`). The replicated values are given in the second table from the left. Missing values are indicated by a `'.'` (see below).



REC	REP	VEHS	SEX	AGE	SAFE	CF	TEST	BAC	MISS
1	1	2	1	3	2	1	0	0	0
	2	2	1	3	2	1	0	0	0
	3	2	1	3	2	1	0	0	0
2	1	2	1	2	1	1	.	.	0
	2	2	1	2	1	1	.	.	0
	3	2	1	2	1	1	.	.	0
3	1	2	1	2	2	2	0	0	0
	2	2	1	2	2	2	0	0	0
	3	2	1	2	2	2	0	0	0
4	1	1	1	.	1	2	.	.	1
	2	1	1	.	1	2	.	.	1
	3	1	1	.	1	2	.	.	1
5	1	2	1	4	2	1	.	.	0
	2	2	1	4	2	1	.	.	0
	3	2	1	4	2	1	.	.	0
6	1	3	2	3	2	2	1	91	0
	2	3	2	3	2	2	1	91	0
	3	3	2	3	2	2	1	91	0
7	1	1	.	1	1	2	1	156	1
	2	1	.	1	1	2	1	156	1
	3	1	.	1	1	2	1	156	1
8	1	2	2	1	1	1	1	23	0
	2	2	2	1	1	1	1	23	0
	3	2	2	1	1	1	1	23	0
9	1	2	1	2	2	.	.	.	1
	2	2	1	2	2	.	.	.	1
	3	2	1	2	2	.	.	.	1
10	1	2	1	3	2	.	0	0	1
	2	2	1	3	2	.	0	0	1
	3	2	1	3	2	.	0	0	1
11	1	2	1	3	2	2	.	.	0
	2	2	1	3	2	2	.	.	0
	3	2	1	3	2	2	.	.	0
12	1	1	1	5	.	1	.	.	1
	2	1	1	5	.	1	.	.	1
	3	1	1	5	.	1	.	.	1
13	1	1	2	4	.	.	0	0	2
	2	1	2	4	.	.	0	0	2
	3	1	2	4	.	.	0	0	2
14	1	2	2	4	1	1	1	118	0
	2	2	2	4	1	1	1	118	0
	3	2	2	4	1	1	1	118	0
	Missing:	0	3	3	6	9			

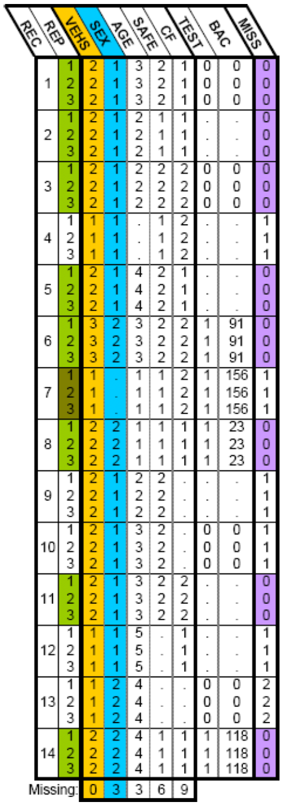
The number of missing values for each explanatory variables is shown at the bottom of each table; the number of missing explanatory variables by record is found in the last column. Ultimately, we are looking to impute the values of Z_1 and Z_2 for the six records for which these values are missing. Along the way, we will also impute the missing values of the explanatory variables.

Step 1—1

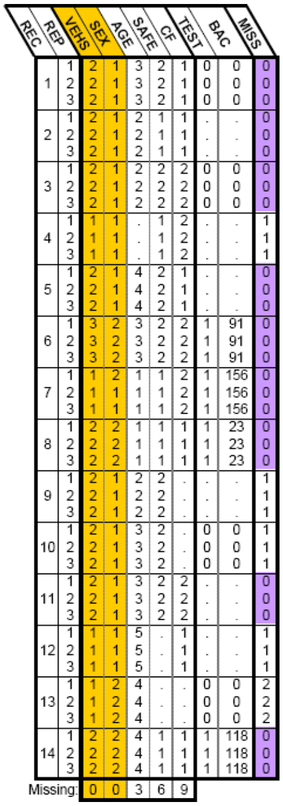
In this case, there are $m = 5$ explanatory variables, $m_1 = 4$ such variables with missing values and the one with the fewest number of missing values is $Y_{\alpha_1} = Y_2$, which is highlighted in blue in the (pre-imputation) table below (on the left). The set $W_{\alpha_1}^{\text{imp}} = \{X_{7,1}, X_{7,2}, X_{7,3}\}$ is shown in brown; the training set

$$W_{\alpha_1}^{\text{train}} = \{X_{1,j}, X_{2,j}, X_{3,j}, X_{5,j}, X_{6,j}, X_{8,j}, X_{11,j}, X_{14,j}\}_{j=1}^{3}$$

is in light green. The (artificial) results of the imputation are shown in the (post-imputation) table on the right. Explanatory variables shown in yellow indicates that this variable will no longer be imputed for the current imputation order.



REC	REP	VEHS	SEX	AGE	SAFE	CF	TEST	BAC	MISS
1	1	2	1	3	2	1	0	0	0
	2	2	1	3	2	1	0	0	0
	3	2	1	3	2	1	0	0	0
2	1	2	1	2	1	1	.	.	0
	2	2	1	2	1	1	.	.	0
	3	2	1	2	1	1	.	.	0
3	1	2	1	2	2	2	0	0	0
	2	2	1	2	2	2	0	0	0
	3	2	1	2	2	2	0	0	0
4	1	1	1	.	1	2	.	.	1
	2	1	1	.	1	2	.	.	1
	3	1	1	.	1	2	.	.	1
5	1	2	1	4	2	1	.	.	0
	2	2	1	4	2	1	.	.	0
	3	2	1	4	2	1	.	.	0
6	1	3	2	3	2	2	1	91	0
	2	3	2	3	2	2	1	91	0
	3	3	2	3	2	2	1	91	0
7	1	1	.	1	1	2	1	156	1
	2	1	.	1	1	2	1	156	1
	3	1	.	1	1	2	1	156	1
8	1	2	2	1	1	1	1	23	0
	2	2	2	1	1	1	1	23	0
	3	2	2	1	1	1	1	23	0
9	1	2	1	2	2	.	.	.	1
	2	2	1	2	2	.	.	.	1
	3	2	1	2	2	.	.	.	1
10	1	2	1	3	2	.	0	0	1
	2	2	1	3	2	.	0	0	1
	3	2	1	3	2	.	0	0	1
11	1	2	1	3	2	2	.	.	0
	2	2	1	3	2	2	.	.	0
	3	2	1	3	2	2	.	.	0
12	1	1	1	5	.	1	.	.	1
	2	1	1	5	.	1	.	.	1
	3	1	1	5	.	1	.	.	1
13	1	1	2	4	.	.	0	0	2
	2	1	2	4	.	.	0	0	2
	3	1	2	4	.	.	0	0	2
14	1	2	2	4	1	1	1	118	0
	2	2	2	4	1	1	1	118	0
	3	2	2	4	1	1	1	118	0
Missing:		0	3	3	6	9			



REC	REP	VEHS	SEX	AGE	SAFE	CF	TEST	BAC	MISS
1	1	2	1	3	2	1	0	0	0
	2	2	1	3	2	1	0	0	0
	3	2	1	3	2	1	0	0	0
2	1	2	1	2	1	1	.	.	0
	2	2	1	2	1	1	.	.	0
	3	2	1	2	1	1	.	.	0
3	1	2	1	2	2	2	0	0	0
	2	2	1	2	2	2	0	0	0
	3	2	1	2	2	2	0	0	0
4	1	1	1	.	1	2	.	.	1
	2	1	1	.	1	2	.	.	1
	3	1	1	.	1	2	.	.	1
5	1	2	1	4	2	1	.	.	0
	2	2	1	4	2	1	.	.	0
	3	2	1	4	2	1	.	.	0
6	1	3	2	3	2	2	1	91	0
	2	3	2	3	2	2	1	91	0
	3	3	2	3	2	2	1	91	0
7	1	1	2	1	1	2	1	156	0
	2	1	1	1	1	2	1	156	0
	3	1	1	1	1	2	1	156	0
8	1	2	2	1	1	1	1	23	0
	2	2	2	1	1	1	1	23	0
	3	2	2	1	1	1	1	23	0
9	1	2	1	2	2	.	.	.	1
	2	2	1	2	2	.	.	.	1
	3	2	1	2	2	.	.	.	1
10	1	2	1	3	2	.	0	0	1
	2	2	1	3	2	.	0	0	1
	3	2	1	3	2	.	0	0	1
11	1	2	1	3	2	2	.	.	0
	2	2	1	3	2	2	.	.	0
	3	2	1	3	2	2	.	.	0
12	1	1	1	5	.	1	.	.	1
	2	1	1	5	.	1	.	.	1
	3	1	1	5	.	1	.	.	1
13	1	1	2	4	.	.	0	0	2
	2	1	2	4	.	.	0	0	2
	3	1	2	4	.	.	0	0	2
14	1	2	2	4	1	1	1	118	0
	2	2	2	4	1	1	1	118	0
	3	2	2	4	1	1	1	118	0
Missing:		0	0	3	6	9			

Step 1—2

After the first first-order imputation, we have $Y_{\alpha_2} = Y_3$, $W_{\alpha_2}^{\text{imp}} = \{X_{4,1}, X_{4,2}, X_{4,3}\}$, and

$$W_{\alpha_2}^{\text{train}} = \{X_{1,j}, X_{2,j}, X_{3,j}, X_{5,j}, X_{6,j}, X_{7,j}, X_{8,j}, X_{11,j}, X_{14,j}\}_{j=1}^{3}.$$

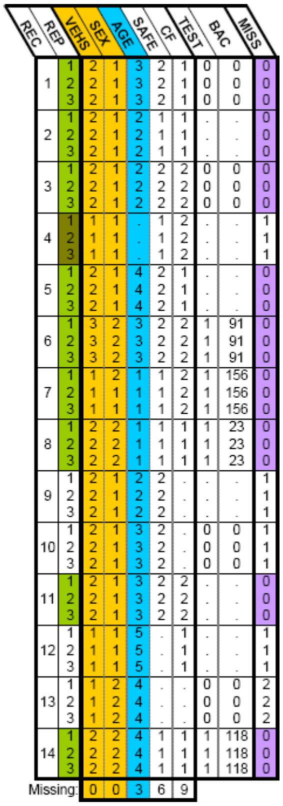
The (pre-imputation) table is the top left entry in the next column; the (artificial) post-imputation results are found in the top right entry.

Step 1—3

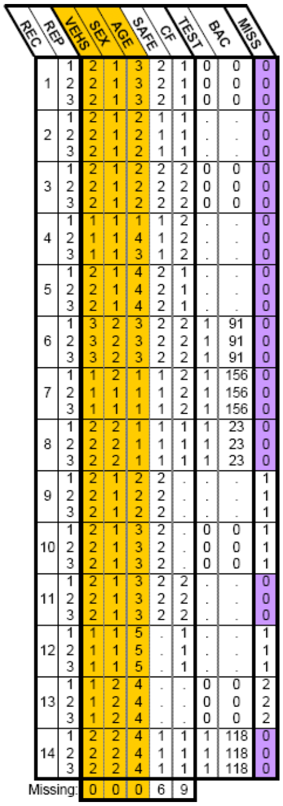
After the second first-order imputation, we have $Y_{\alpha_3} = Y_4$, $W_{\alpha_3}^{\text{imp}} = \{X_{4,1}, X_{4,2}, X_{4,3}\}$, and

$$W_{\alpha_3}^{\text{train}} = \{X_{1,j}, X_{2,j}, X_{3,j}, X_{5,j}, X_{6,j}, X_{7,j}, X_{8,j}, X_{11,j}, X_{14,j}\}_{j=1}^{3}.$$

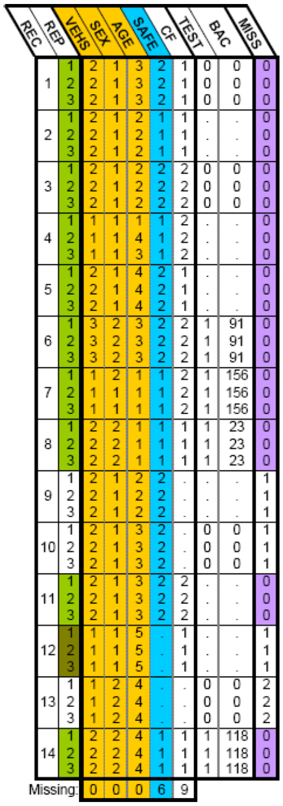
The (pre-imputation) table is the bottom left entry below; the (artificial) post-imputation results are found in the bottom right table.



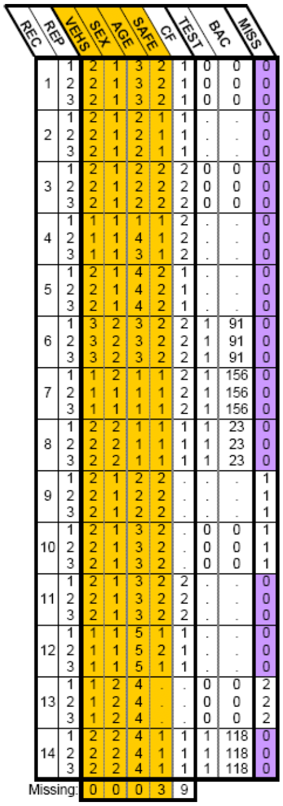
REC	REP	VEHS	SEX	AGE	SAFE	CF	TEST	BAC	MISS
1	1	2	1	3	2	1	0	0	0
	2	2	1	3	2	1	0	0	0
	3	2	1	3	2	1	0	0	0
2	1	2	1	2	1	1	.	.	0
	2	2	1	2	1	1	.	.	0
	3	2	1	2	1	1	.	.	0
3	1	2	1	2	2	2	0	0	0
	2	2	1	2	2	2	0	0	0
	3	2	1	2	2	2	0	0	0
4	1	1	1	.	1	2	.	.	1
	2	1	1	.	1	2	.	.	1
	3	1	1	.	1	2	.	.	1
5	1	2	1	4	2	1	.	.	0
	2	2	1	4	2	1	.	.	0
	3	2	1	4	2	1	.	.	0
6	1	3	2	3	2	2	1	91	0
	2	3	2	3	2	2	1	91	0
	3	3	2	3	2	2	1	91	0
7	1	1	2	1	1	2	1	156	0
	2	1	1	1	1	2	1	156	0
	3	1	1	1	1	2	1	156	0
8	1	2	2	1	1	1	1	23	0
	2	2	2	1	1	1	1	23	0
	3	2	2	1	1	1	1	23	0
9	1	2	1	2	2	.	.	.	1
	2	2	1	2	2	.	.	.	1
	3	2	1	2	2	.	.	.	1
10	1	2	1	3	2	.	0	0	1
	2	2	1	3	2	.	0	0	1
	3	2	1	3	2	.	0	0	1
11	1	2	1	3	2	2	.	.	0
	2	2	1	3	2	2	.	.	0
	3	2	1	3	2	2	.	.	0
12	1	1	1	5	.	1	.	.	1
	2	1	1	5	.	1	.	.	1
	3	1	1	5	.	1	.	.	1
13	1	1	2	4	.	.	0	0	2
	2	1	2	4	.	.	0	0	2
	3	1	2	4	.	.	0	0	2
14	1	2	2	4	1	1	1	118	0
	2	2	2	4	1	1	1	118	0
	3	2	2	4	1	1	1	118	0
Missing:		0	0	3	6	9			



REC	REP	VEHS	SEX	AGE	SAFE	CF	TEST	BAC	MISS
1	1	2	1	3	2	1	0	0	0
	2	2	1	3	2	1	0	0	0
	3	2	1	3	2	1	0	0	0
2	1	2	1	2	1	1	.	.	0
	2	2	1	2	1	1	.	.	0
	3	2	1	2	1	1	.	.	0
3	1	2	1	2	2	2	0	0	0
	2	2	1	2	2	2	0	0	0
	3	2	1	2	2	2	0	0	0
4	1	1	1	1	1	2	.	.	0
	2	1	1	4	1	2	.	.	0
	3	1	1	3	1	2	.	.	0
5	1	2	1	4	2	1	.	.	0
	2	2	1	4	2	1	.	.	0
	3	2	1	4	2	1	.	.	0
6	1	3	2	3	2	2	1	91	0
	2	3	2	3	2	2	1	91	0
	3	3	2	3	2	2	1	91	0
7	1	1	2	1	1	2	1	156	0
	2	1	1	1	1	2	1	156	0
	3	1	1	1	1	2	1	156	0
8	1	2	2	1	1	1	1	23	0
	2	2	2	1	1	1	1	23	0
	3	2	2	1	1	1	1	23	0
9	1	2	1	2	2	.	.	.	1
	2	2	1	2	2	.	.	.	1
	3	2	1	2	2	.	.	.	1
10	1	2	1	3	2	.	0	0	1
	2	2	1	3	2	.	0	0	1
	3	2	1	3	2	.	0	0	1
11	1	2	1	3	2	2	.	.	0
	2	2	1	3	2	2	.	.	0
	3	2	1	3	2	2	.	.	0
12	1	1	1	5	.	1	.	.	1
	2	1	1	5	.	1	.	.	1
	3	1	1	5	.	1	.	.	1
13	1	1	2	4	.	.	0	0	2
	2	1	2	4	.	.	0	0	2
	3	1	2	4	.	.	0	0	2
14	1	2	2	4	1	1	1	118	0
	2	2	2	4	1	1	1	118	0
	3	2	2	4	1	1	1	118	0
Missing:		0	0	0	6	9			



REC	REP	VEHS	SEX	AGE	SAFE	CF	TEST	BAC	MISS
1	1	2	1	3	2	1	0	0	0
	2	2	1	3	2	1	0	0	0
	3	2	1	3	2	1	0	0	0
2	1	2	1	2	1	1	.	.	0
	2	2	1	2	1	1	.	.	0
	3	2	1	2	1	1	.	.	0
3	1	2	1	2	2	2	0	0	0
	2	2	1	2	2	2	0	0	0
	3	2	1	2	2	2	0	0	0
4	1	1	1	1	1	2	.	.	0
	2	1	1	4	1	2	.	.	0
	3	1	1	3	1	2	.	.	0
5	1	2	1	4	2	1	.	.	0
	2	2	1	4	2	1	.	.	0
	3	2	1	4	2	1	.	.	0
6	1	3	2	3	2	2	1	91	0
	2	3	2	3	2	2	1	91	0
	3	3	2	3	2	2	1	91	0
7	1	1	2	1	1	2	1	156	0
	2	1	1	1	1	2	1	156	0
	3	1	1	1	1	2	1	156	0
8	1	2	2	1	1	1	1	23	0
	2	2	2	1	1	1	1	23	0
	3	2	2	1	1	1	1	23	0
9	1	2	1	2	2	.	.	.	1
	2	2	1	2	2	.	.	.	1
	3	2	1	2	2	.	.	.	1
10	1	2	1	3	2	.	0	0	1
	2	2	1	3	2	.	0	0	1
	3	2	1	3	2	.	0	0	1
11	1	2	1	3	2	2	.	.	0
	2	2	1	3	2	2	.	.	0
	3	2	1	3	2	2	.	.	0
12	1	1	1	5	.	1	.	.	1
	2	1	1	5	.	1	.	.	1
	3	1	1	5	.	1	.	.	1
13	1	1	2	4	.	.	0	0	2
	2	1	2	4	.	.	0	0	2
	3	1	2	4	.	.	0	0	2
14	1	2	2	4	1	1	1	118	0
	2	2	2	4	1	1	1	118	0
	3	2	2	4	1	1	1	118	0
Missing:		0	0	0	6	9			



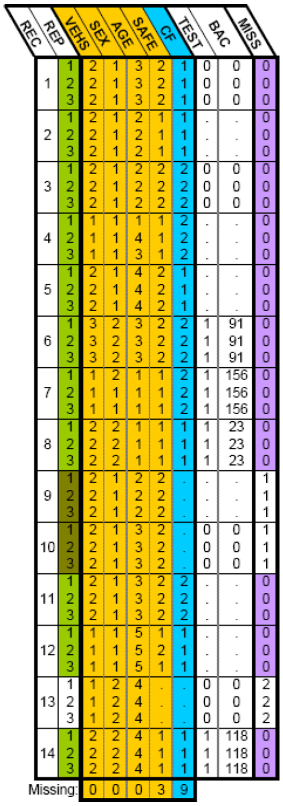
REC	REP	VEHS	SEX	AGE	SAFE	CF	TEST	BAC	MISS
1	1	2	1	3	2	1	0	0	0
	2	2	1	3	2	1	0	0	0
	3	2	1	3	2	1	0	0	0
2	1	2	1	2	1	1	.	.	0
	2	2	1	2	1	1	.	.	0
	3	2	1	2	1	1	.	.	0
3	1	2	1	2	2	2	0	0	0
	2	2	1	2	2	2	0	0	0
	3	2	1	2	2	2	0	0	0
4	1	1	1	1	1	2	.	.	0
	2	1	1	4	1	2	.	.	0
	3	1	1	3	1	2	.	.	0
5	1	2	1	4	2	1	.	.	0
	2	2	1	4	2	1	.	.	0
	3	2	1	4	2	1	.	.	0
6	1	3	2	3	2	2	1	91	0
	2	3	2	3	2	2	1	91	0
	3	3	2	3	2	2	1	91	0
7	1	1	2	1	1	2	1	156	0
	2	1	1	1	1	2	1	156	0
	3	1	1	1	1	2	1	156	0
8	1	2	2	1	1	1	1	23	0
	2	2	2	1	1	1	1	23	0
	3	2	2	1	1	1	1	23	0
9	1	2	1	2	2	.	.	.	1
	2	2	1	2	2	.	.	.	1
	3	2	1	2	2	.	.	.	1
10	1	2	1	3	2	.	0	0	1
	2	2	1	3	2	.	0	0	1
	3	2	1	3	2	.	0	0	1
11	1	2	1	3	2	2	.	.	0
	2	2	1	3	2	2	.	.	0
	3	2	1	3	2	2	.	.	0
12	1	1	1	5	1	1	.	.	0
	2	1	1	5	2	1	.	.	0
	3	1	1	5	1	1	.	.	0
13	1	1	2	4	.	.	0	0	2
	2	1	2	4	.	.	0	0	2
	3	1	2	4	.	.	0	0	2
14	1	2	2	4	1	1	1	118	0
	2	2	2	4	1	1	1	118	0
	3	2	2	4	1	1	1	118	0
Missing:		0	0	0	3	9			

Step 1—4

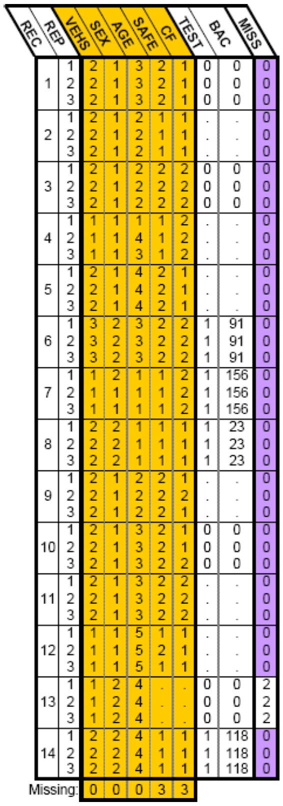
The last first-order imputation is the imputation of $Y_{\alpha_4} = Y_5$, where $W^{\text{imp}}_{\alpha_4} = \{X_{4,1}, X_{4,2}, X_{4,3}\}$,

$$W^{\text{train}}_{\alpha_4} = \{X_{1,j}, X_{2,j}, X_{3,j}, X_{5,j}, X_{6,j}, X_{7,j}, X_{8,j}, X_{11,j}, X_{12,j}, X_{14,j}\}_{j=1}^{3}.$$

The (pre-imputation) table is the left entry below; the (artificial) post-imputation results are found next to it.



REC	REP	VEHS	SEX	AGE	SAFE	CF	TEST	BAC	MISS
1	1	2	1	3	2	1	0	0	0
	2	2	1	3	2	1	0	0	0
	3	2	1	3	2	1	0	0	0
2	1	2	1	2	1	1	.	.	0
	2	2	1	2	1	1	.	.	0
	3	2	1	2	1	1	.	.	0
3	1	2	1	2	2	2	0	0	0
	2	2	1	2	2	2	0	0	0
	3	2	1	2	2	2	0	0	0
4	1	1	1	1	1	2	.	.	0
	2	1	1	4	1	2	.	.	0
	3	1	1	3	1	2	.	.	0
5	1	2	1	4	2	1	.	.	0
	2	2	1	4	2	1	.	.	0
	3	2	1	4	2	1	.	.	0
6	1	3	2	3	2	2	1	91	0
	2	3	2	3	2	2	1	91	0
	3	3	2	3	2	2	1	91	0
7	1	1	2	1	1	2	1	156	0
	2	1	1	1	1	2	1	156	0
	3	1	1	1	1	2	1	156	0
8	1	2	2	1	1	1	1	23	0
	2	2	2	1	1	1	1	23	0
	3	2	2	1	1	1	1	23	0
9	1	2	1	2	2	.	.	.	1
	2	2	1	2	2	.	.	.	1
	3	2	1	2	2	.	.	.	1
10	1	2	1	3	2	.	0	0	1
	2	2	1	3	2	.	0	0	1
	3	2	1	3	2	.	0	0	1
11	1	2	1	3	2	2	.	.	0
	2	2	1	3	2	2	.	.	0
	3	2	1	3	2	2	.	.	0
12	1	1	1	5	1	1	.	.	0
	2	1	1	5	2	1	.	.	0
	3	1	1	5	1	1	.	.	0
13	1	1	2	4	.	.	0	0	2
	2	1	2	4	.	.	0	0	2
	3	1	2	4	.	.	0	0	2
14	1	2	2	4	1	1	1	118	0
	2	2	2	4	1	1	1	118	0
	3	2	2	4	1	1	1	118	0
Missing:		0	0	0	3	9			



REC	REP	VEHS	SEX	AGE	SAFE	CF	TEST	BAC	MISS
1	1	2	1	3	2	1	0	0	0
	2	2	1	3	2	1	0	0	0
	3	2	1	3	2	1	0	0	0
2	1	2	1	2	1	1	.	.	0
	2	2	1	2	1	1	.	.	0
	3	2	1	2	1	1	.	.	0
3	1	2	1	2	2	2	0	0	0
	2	2	1	2	2	2	0	0	0
	3	2	1	2	2	2	0	0	0
4	1	1	1	1	1	2	.	.	0
	2	1	1	4	1	2	.	.	0
	3	1	1	3	1	2	.	.	0
5	1	2	1	4	2	1	.	.	0
	2	2	1	4	2	1	.	.	0
	3	2	1	4	2	1	.	.	0
6	1	3	2	3	2	2	1	91	0
	2	3	2	3	2	2	1	91	0
	3	3	2	3	2	2	1	91	0
7	1	1	2	1	1	2	1	156	0
	2	1	1	1	1	2	1	156	0
	3	1	1	1	1	2	1	156	0
8	1	2	2	1	1	1	1	23	0
	2	2	2	1	1	1	1	23	0
	3	2	2	1	1	1	1	23	0
9	1	2	1	2	2	1	.	.	0
	2	2	1	2	2	2	.	.	0
	3	2	1	2	2	1	.	.	0
10	1	2	1	3	2	1	0	0	0
	2	2	1	3	2	1	0	0	0
	3	2	1	3	2	2	0	0	0
11	1	2	1	3	2	2	.	.	0
	2	2	1	3	2	2	.	.	0
	3	2	1	3	2	2	.	.	0
12	1	1	1	5	1	1	.	.	0
	2	1	1	5	2	1	.	.	0
	3	1	1	5	1	1	.	.	0
13	1	1	2	4	.	.	0	0	2
	2	1	2	4	.	.	0	0	2
	3	1	2	4	.	.	0	0	2
14	1	2	2	4	1	1	1	118	0
	2	2	2	4	1	1	1	118	0
	3	2	2	4	1	1	1	118	0
Missing:		0	0	0	3	3			

By construction, at the end of first-order imputation, all the records are either complete or they contain no fewer than 2 missing values.

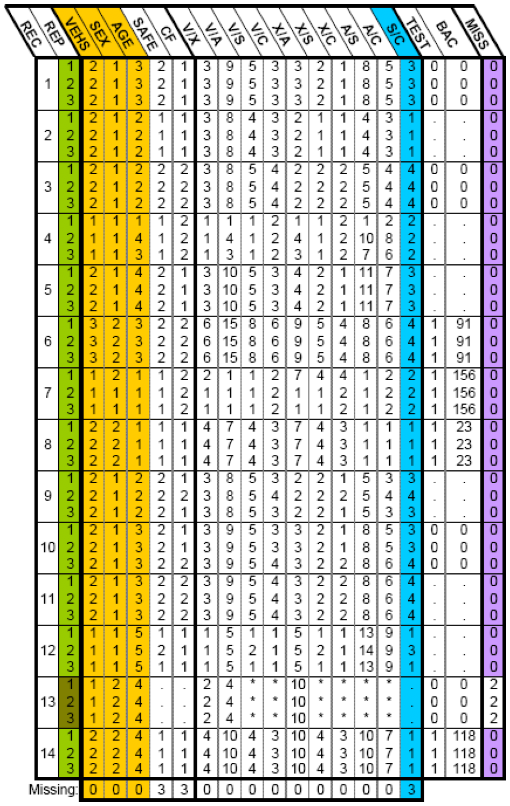
Step 2—1

We have $m_2 = 1$ since the only pair of distinct explanatory variables with missing values is $\{Y_4, Y_5\}$ – crossing them yields Y_{α_1,β_1}, highlighted in blue in the larger table in the next column. The set $W^{\text{imp}}_{\alpha_1,\beta_1} = \{X_{13,1}, X_{13,2}, X_{13,3}\}$ is shown in brown; the training set

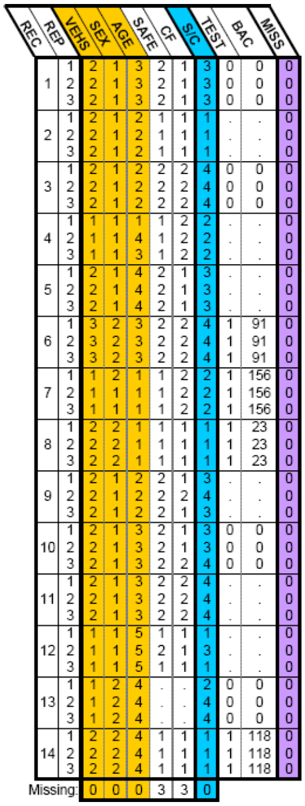
$$W^{\text{imp}}_{\alpha_1} = \{X_{i,j} : i \neq 13, j = 1,2,3\}$$

is in light green. The (artificial) results of the imputation are shown in the (post-imputation) bottom left table. The last table shows the result of “de-crossing” Y_{α_1,β_1} into its constituent variables Y_4 and Y_5.

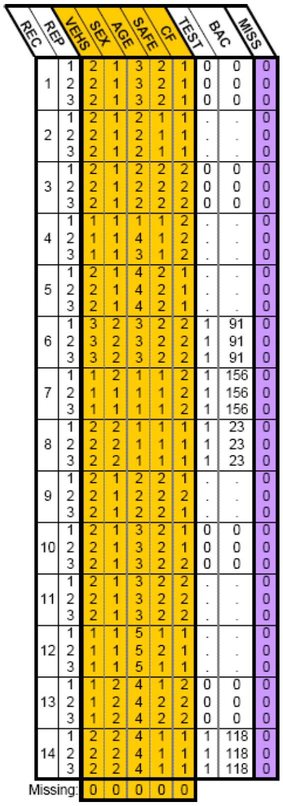
Since all explanatory variables have been imputed, we can now conduct the imputation of the dependent variables.



REC	REP	VEHS	SEX	AGE	SAFE	CF	V/X	V/A	V/S	V/C	X/A	X/S	X/C	A/S	A/C	S/C	TEST	BAC	MISS
1	1	2	1	3	2	1	3	9	5	3	3	2	1	8	5	3	0	0	0
	2	2	1	3	2	1	3	9	5	3	3	2	1	8	5	3	0	0	0
	3	2	1	3	2	1	3	9	5	3	3	2	1	8	5	3	0	0	0
2	1	2	1	2	1	1	3	8	4	3	2	1	1	4	3	1	.	.	0
	2	2	1	2	1	1	3	8	4	3	2	1	1	4	3	1	.	.	0
	3	2	1	2	1	1	3	8	4	3	2	1	1	4	3	1	.	.	0
3	1	2	1	2	2	2	3	8	5	4	2	2	2	5	4	4	0	0	0
	2	2	1	2	2	2	3	8	5	4	2	2	2	5	4	4	0	0	0
	3	2	1	2	2	2	3	8	5	4	2	2	2	5	4	4	0	0	0
4	1	1	1	1	1	2	1	1	1	2	1	1	2	1	2	2	.	.	0
	2	1	1	4	1	2	1	4	1	2	4	1	2	10	8	2	.	.	0
	3	1	1	3	1	2	1	3	1	2	3	1	2	7	6	2	.	.	0
5	1	2	1	4	2	1	3	10	5	3	4	2	1	11	7	3	.	.	0
	2	2	1	4	2	1	3	10	5	3	4	2	1	11	7	3	.	.	0
	3	2	1	4	2	1	3	10	5	3	4	2	1	11	7	3	.	.	0
6	1	3	2	3	2	2	6	15	8	6	9	5	4	8	6	4	1	91	0
	2	3	2	3	2	2	6	15	8	6	9	5	4	8	6	4	1	91	0
	3	3	2	3	2	2	6	15	8	6	9	5	4	8	6	4	1	91	0
7	1	1	2	1	1	2	2	1	1	2	7	4	4	1	2	2	1	156	0
	2	1	1	1	1	2	1	1	1	2	1	1	2	1	2	2	1	156	0
	3	1	1	1	1	2	1	1	1	2	1	1	2	1	2	2	1	156	0
8	1	2	2	1	1	1	4	7	4	3	7	4	3	1	1	1	1	23	0
	2	2	2	1	1	1	4	7	4	3	7	4	3	1	1	1	1	23	0
	3	2	2	1	1	1	4	7	4	3	7	4	3	1	1	1	1	23	0
9	1	2	1	2	2	1	3	8	5	3	2	2	1	5	3	3	.	.	0
	2	2	1	2	2	2	3	8	5	4	2	2	2	5	4	4	.	.	0
	3	2	1	2	2	1	3	8	5	3	2	2	1	5	3	3	.	.	0
10	1	2	1	3	2	1	3	9	5	3	3	2	1	8	5	3	0	0	0
	2	2	1	3	2	1	3	9	5	3	3	2	1	8	5	3	0	0	0
	3	2	1	3	2	2	3	9	5	4	3	2	2	8	6	4	0	0	0
11	1	2	1	3	2	2	3	9	5	4	3	2	2	8	6	4	.	.	0
	2	2	1	3	2	2	3	9	5	4	3	2	2	8	6	4	.	.	0
	3	2	1	3	2	2	3	9	5	4	3	2	2	8	6	4	.	.	0
12	1	1	1	5	1	1	1	5	1	1	5	1	1	13	9	1	.	.	0
	2	1	1	5	2	1	1	5	2	1	5	2	1	14	9	3	.	.	0
	3	1	1	5	1	1	1	5	1	1	5	1	1	13	9	1	.	.	0
13	1	1	2	4	.	.	2	4	*	*	10	*	*	*	*	.	0	0	2
	2	1	2	4	.	.	2	4	*	*	10	*	*	*	*	.	0	0	2
	3	1	2	4	.	.	2	4	*	*	10	*	*	*	*	.	0	0	2
14	1	2	2	4	1	1	4	10	4	3	10	4	3	10	7	1	1	118	0
	2	2	2	4	1	1	4	10	4	3	10	4	3	10	7	1	1	118	0
	3	2	2	4	1	1	4	10	4	3	10	4	3	10	7	1	1	118	0
Missing:		0	0	0	3	3	0	0	0	0	0	0	0	0	0	3			



REC	REP	VEHS	SEX	AGE	SAFE	CF	S/C	TEST	BAC	MISS
1	1	2	1	3	2	1	3	0	0	0
	2	2	1	3	2	1	3	0	0	0
	3	2	1	3	2	1	3	0	0	0
2	1	2	1	2	1	1	1	.	.	0
	2	2	1	2	1	1	1	.	.	0
	3	2	1	2	1	1	1	.	.	0
3	1	2	1	2	2	2	4	0	0	0
	2	2	1	2	2	2	4	0	0	0
	3	2	1	2	2	2	4	0	0	0
4	1	1	1	1	1	2	2	.	.	0
	2	1	1	4	1	2	2	.	.	0
	3	1	1	3	1	2	2	.	.	0
5	1	2	1	4	2	1	3	.	.	0
	2	2	1	4	2	1	3	.	.	0
	3	2	1	4	2	1	3	.	.	0
6	1	3	2	3	2	2	4	1	91	0
	2	3	2	3	2	2	4	1	91	0
	3	3	2	3	2	2	4	1	91	0
7	1	1	2	1	1	2	2	1	156	0
	2	1	1	1	1	2	2	1	156	0
	3	1	1	1	1	2	2	1	156	0
8	1	2	2	1	1	1	1	1	23	0
	2	2	2	1	1	1	1	1	23	0
	3	2	2	1	1	1	1	1	23	0
9	1	2	1	2	2	1	3	.	.	0
	2	2	1	2	2	2	4	.	.	0
	3	2	1	2	2	1	3	.	.	0
10	1	2	1	3	2	1	3	0	0	0
	2	2	1	3	2	1	3	0	0	0
	3	2	1	3	2	2	4	0	0	0
11	1	2	1	3	2	2	4	.	.	0
	2	2	1	3	2	2	4	.	.	0
	3	2	1	3	2	2	4	.	.	0
12	1	1	1	5	1	1	1	.	.	0
	2	1	1	5	2	1	3	.	.	0
	3	1	1	5	1	1	1	.	.	0
13	1	1	2	4	.	.	2	0	0	0
	2	1	2	4	.	.	4	0	0	0
	3	1	2	4	.	.	4	0	0	0
14	1	2	2	4	1	1	1	1	118	0
	2	2	2	4	1	1	1	1	118	0
	3	2	2	4	1	1	1	1	118	0
Missing:		0	0	0	3	3	0			



REC	REP	VEHS	SEX	AGE	SAFE	CF	TEST	BAC	MISS
1	1	2	1	3	2	1	0	0	0
	2	2	1	3	2	1	0	0	0
	3	2	1	3	2	1	0	0	0
2	1	2	1	2	1	1	.	.	0
	2	2	1	2	1	1	.	.	0
	3	2	1	2	1	1	.	.	0
3	1	2	1	2	2	2	0	0	0
	2	2	1	2	2	2	0	0	0
	3	2	1	2	2	2	0	0	0
4	1	1	1	1	1	2	.	.	0
	2	1	1	4	1	2	.	.	0
	3	1	1	3	1	2	.	.	0
5	1	2	1	4	2	1	.	.	0
	2	2	1	4	2	1	.	.	0
	3	2	1	4	2	1	.	.	0
6	1	3	2	3	2	2	1	91	0
	2	3	2	3	2	2	1	91	0
	3	3	2	3	2	2	1	91	0
7	1	1	2	1	1	2	1	156	0
	2	1	1	1	1	2	1	156	0
	3	1	1	1	1	2	1	156	0
8	1	2	2	1	1	1	1	23	0
	2	2	2	1	1	1	1	23	0
	3	2	2	1	1	1	1	23	0
9	1	2	1	2	2	1	.	.	0
	2	2	1	2	2	2	.	.	0
	3	2	1	2	2	1	.	.	0
10	1	2	1	3	2	1	0	0	0
	2	2	1	3	2	1	0	0	0
	3	2	1	3	2	2	0	0	0
11	1	2	1	3	2	2	.	.	0
	2	2	1	3	2	2	.	.	0
	3	2	1	3	2	2	.	.	0
12	1	1	1	5	1	1	.	.	0
	2	1	1	5	2	1	.	.	0
	3	1	1	5	1	1	.	.	0
13	1	1	2	4	1	2	0	0	0
	2	1	2	4	2	2	0	0	0
	3	1	2	4	2	2	0	0	0
14	1	2	2	4	1	1	1	118	0
	2	2	2	4	1	1	1	118	0
	3	2	2	4	1	1	1	118	0
Missing:		0	0	0	0	0			

Imputation of the Dependent Variable Z_1

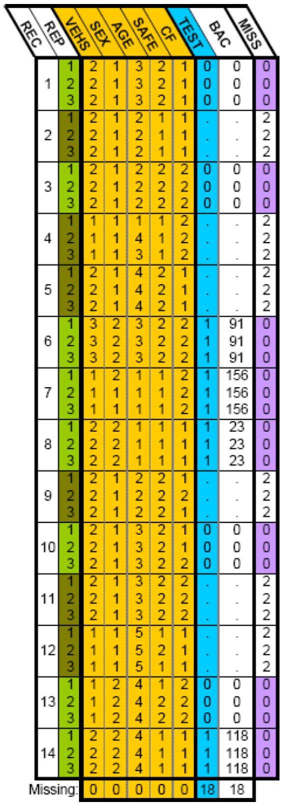
From this point on, it is the number of dependent variables by record which is found in the last (magenta) column. The set

$$W_{Z_1}^{\text{imp}} = \{X_{2,j},X_{4,j},X_{5,j},X_{9,j},X_{11,j},X_{12,j}\}_{j=1}^{3}$$

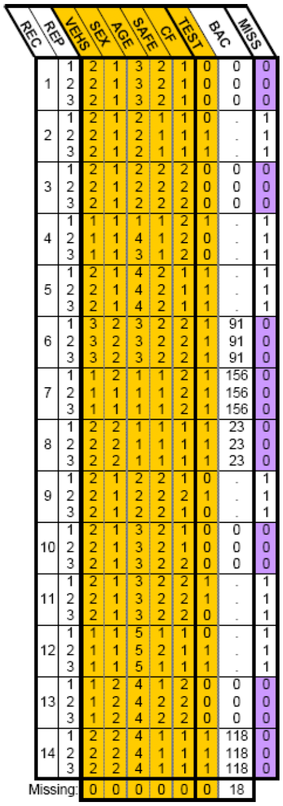
is shown in brown and the training set

$$W_{\alpha_1}^{\text{train}} = \{X_{1,j},X_{3,j},X_{6,j},X_{7,j},X_{8,j},X_{10,j},X_{13,j},X_{14,j}\}_{j=1}^{3}$$

is in light green in the table on the left below. The (artificial) results of the imputation are shown in the (post-imputation) table on the right.



REC	REP	VEHS	SEX	AGE	SAFE	CF	TEST	BAC	MISS
1	1	2	1	3	2	1	0	0	0
	2	2	1	3	2	1	0	0	0
	3	2	1	3	2	1	0	0	0
2	1	2	1	2	1	1	.	.	2
	2	2	1	2	1	1	.	.	2
	3	2	1	2	1	1	.	.	2
3	1	2	1	2	2	2	0	0	0
	2	2	1	2	2	2	0	0	0
	3	2	1	2	2	2	0	0	0
4	1	1	1	1	1	2	.	.	2
	2	1	1	4	1	2	.	.	2
	3	1	1	3	1	2	.	.	2
5	1	2	1	4	2	1	.	.	2
	2	2	1	4	2	1	.	.	2
	3	2	1	4	2	1	.	.	2
6	1	3	2	3	2	2	1	91	0
	2	3	2	3	2	2	1	91	0
	3	3	2	3	2	2	1	91	0
7	1	1	2	1	1	2	1	156	0
	2	1	1	1	1	2	1	156	0
	3	1	1	1	1	2	1	156	0
8	1	2	2	1	1	1	1	23	0
	2	2	2	1	1	1	1	23	0
	3	2	2	1	1	1	1	23	0
9	1	2	1	2	2	1	.	.	2
	2	2	1	2	2	2	.	.	2
	3	2	1	2	2	1	.	.	2
10	1	2	1	3	2	1	0	0	0
	2	2	1	3	2	1	0	0	0
	3	2	1	3	2	2	0	0	0
11	1	2	1	3	2	2	.	.	2
	2	2	1	3	2	2	.	.	2
	3	2	1	3	2	2	.	.	2
12	1	1	1	5	1	1	.	.	2
	2	1	1	5	2	1	.	.	2
	3	1	1	5	1	1	.	.	2
13	1	1	2	4	1	2	0	0	0
	2	1	2	4	2	2	0	0	0
	3	1	2	4	2	2	0	0	0
14	1	2	2	4	1	1	1	118	0
	2	2	2	4	1	1	1	118	0
	3	2	2	4	1	1	1	118	0
Missing:		0	0	0	0	0	18	18	



REC	REP	VEHS	SEX	AGE	SAFE	CF	TEST	BAC	MISS
1	1	2	1	3	2	1	0	0	0
	2	2	1	3	2	1	0	0	0
	3	2	1	3	2	1	0	0	0
2	1	2	1	2	1	1	0	.	1
	2	2	1	2	1	1	1	.	1
	3	2	1	2	1	1	1	.	1
3	1	2	1	2	2	2	0	0	0
	2	2	1	2	2	2	0	0	0
	3	2	1	2	2	2	0	0	0
4	1	1	1	1	1	2	1	.	1
	2	1	1	4	1	2	0	.	1
	3	1	1	3	1	2	0	.	1
5	1	2	1	4	2	1	1	.	1
	2	2	1	4	2	1	1	.	1
	3	2	1	4	2	1	1	.	1
6	1	3	2	3	2	2	1	91	0
	2	3	2	3	2	2	1	91	0
	3	3	2	3	2	2	1	91	0
7	1	1	2	1	1	2	1	156	0
	2	1	1	1	1	2	1	156	0
	3	1	1	1	1	2	1	156	0
8	1	2	2	1	1	1	1	23	0
	2	2	2	1	1	1	1	23	0
	3	2	2	1	1	1	1	23	0
9	1	2	1	2	2	1	0	.	1
	2	2	1	2	2	2	1	.	1
	3	2	1	2	2	1	0	.	1
10	1	2	1	3	2	1	0	0	0
	2	2	1	3	2	1	0	0	0
	3	2	1	3	2	2	0	0	0
11	1	2	1	3	2	2	1	.	1
	2	2	1	3	2	2	1	.	1
	3	2	1	3	2	2	0	.	1
12	1	1	1	5	1	1	0	.	1
	2	1	1	5	2	1	1	.	1
	3	1	1	5	1	1	1	.	1
13	1	1	2	4	1	2	0	0	0
	2	1	2	4	2	2	0	0	0
	3	1	2	4	2	2	0	0	0
14	1	2	2	4	1	1	1	118	0
	2	2	2	4	1	1	1	118	0
	3	2	2	4	1	1	1	118	0
Missing:		0	0	0	0	0	0	18	

Imputation of the Dependent Variable Z_2

In light of the two-stage model described in the Methodology, when $Z_1 = 0$, Z_2 is automatically 0, which is illustrated in the table on top in the next column.

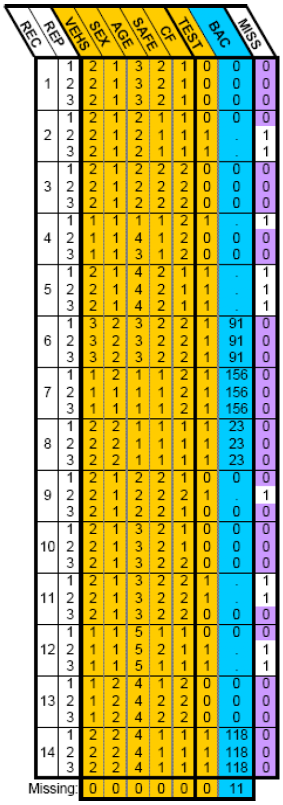
We now have

$$W_{Z_1=1,Z_2}^{\text{imp}} = \{X_{2,2},X_{2,3},X_{4,1},X_{5,1},X_{5,2},X_{5,3},X_{9,2},\\ X_{11,1},X_{11,2},X_{12,2},X_{12,3}\}$$

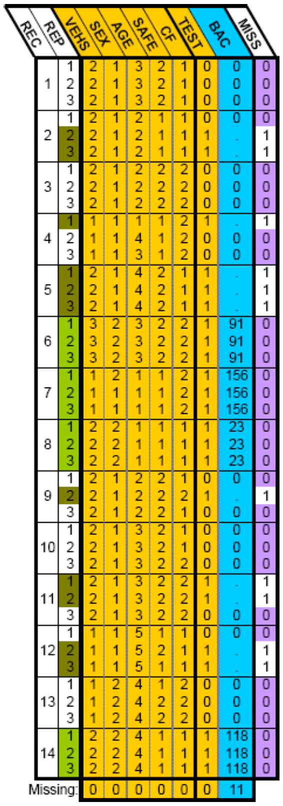
in brown and

$$W_{Z_1=1,Z_2}^{\text{train}} = \{X_{6,j},X_{7,j},X_{8,j},X_{14,j}\}_{j=1}^{3}$$

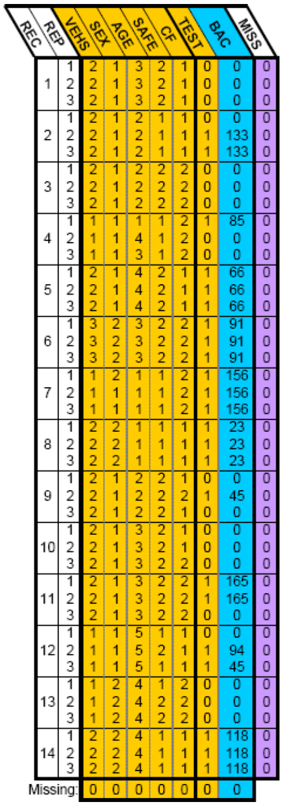
in light green in the table on the bottom left in the next column; the (artificial) results of the imputation are shown in the (post-imputation) table bottom right.



REC	REP	VEHS	SEX	AGE	SAFE	CF	TEST	BAC	MISS
1	1	2	1	3	2	1	0	0	0
	2	2	1	3	2	1	0	0	0
	3	2	1	3	2	1	0	0	0
2	1	2	1	2	1	1	0	0	0
	2	2	1	2	1	1	1	.	1
	3	2	1	2	1	1	1	.	1
3	1	2	1	2	2	2	0	0	0
	2	2	1	2	2	2	0	0	0
	3	2	1	2	2	2	0	0	0
4	1	1	1	1	1	2	1	.	1
	2	1	1	4	1	2	0	0	0
	3	1	1	3	1	2	0	0	0
5	1	2	1	4	2	1	1	.	1
	2	2	1	4	2	1	1	.	1
	3	2	1	4	2	1	1	.	1
6	1	3	2	3	2	2	1	91	0
	2	3	2	3	2	2	1	91	0
	3	3	2	3	2	2	1	91	0
7	1	1	2	1	1	2	1	156	0
	2	1	1	1	1	2	1	156	0
	3	1	1	1	1	2	1	156	0
8	1	2	2	1	1	1	1	23	0
	2	2	2	1	1	1	1	23	0
	3	2	2	1	1	1	1	23	0
9	1	2	1	2	2	1	0	0	0
	2	2	1	2	2	2	1	.	1
	3	2	1	2	2	1	0	0	0
10	1	2	1	3	2	1	0	0	0
	2	2	1	3	2	1	0	0	0
	3	2	1	3	2	2	0	0	0
11	1	2	1	3	2	2	1	.	1
	2	2	1	3	2	2	1	.	1
	3	2	1	3	2	2	0	0	0
12	1	1	1	5	1	1	0	0	0
	2	1	1	5	2	1	1	.	1
	3	1	1	5	1	1	1	.	1
13	1	1	2	4	1	2	0	0	0
	2	1	2	4	2	2	0	0	0
	3	1	2	4	2	2	0	0	0
14	1	2	2	4	1	1	1	118	0
	2	2	2	4	1	1	1	118	0
	3	2	2	4	1	1	1	118	0
Missing:		0	0	0	0	0	0	11	



REC	REP	VEHS	SEX	AGE	SAFE	CF	TEST	BAC	MISS
1	1	2	1	3	2	1	0	0	0
	2	2	1	3	2	1	0	0	0
	3	2	1	3	2	1	0	0	0
2	1	2	1	2	1	1	0	0	0
	2	2	1	2	1	1	1	.	1
	3	2	1	2	1	1	1	.	1
3	1	2	1	2	2	2	0	0	0
	2	2	1	2	2	2	0	0	0
	3	2	1	2	2	2	0	0	0
4	1	1	1	1	1	2	1	.	1
	2	1	1	4	1	2	0	0	0
	3	1	1	3	1	2	0	0	0
5	1	2	1	4	2	1	1	.	1
	2	2	1	4	2	1	1	.	1
	3	2	1	4	2	1	1	.	1
6	1	3	2	3	2	2	1	91	0
	2	3	2	3	2	2	1	91	0
	3	3	2	3	2	2	1	91	0
7	1	1	2	1	1	2	1	156	0
	2	1	1	1	1	2	1	156	0
	3	1	1	1	1	2	1	156	0
8	1	2	2	1	1	1	1	23	0
	2	2	2	1	1	1	1	23	0
	3	2	2	1	1	1	1	23	0
9	1	2	1	2	2	1	0	0	0
	2	2	1	2	2	2	1	.	1
	3	2	1	2	2	1	0	0	0
10	1	2	1	3	2	1	0	0	0
	2	2	1	3	2	1	0	0	0
	3	2	1	3	2	2	0	0	0
11	1	2	1	3	2	2	1	.	1
	2	2	1	3	2	2	1	.	1
	3	2	1	3	2	2	0	0	0
12	1	1	1	5	1	1	0	0	0
	2	1	1	5	2	1	1	.	1
	3	1	1	5	1	1	1	.	1
13	1	1	2	4	1	2	0	0	0
	2	1	2	4	2	2	0	0	0
	3	1	2	4	2	2	0	0	0
14	1	2	2	4	1	1	1	118	0
	2	2	2	4	1	1	1	118	0
	3	2	2	4	1	1	1	118	0
Missing:		0	0	0	0	0	0	11	

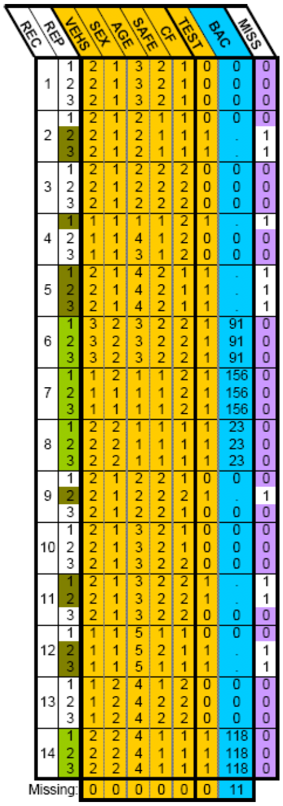


REC	REP	VEHS	SEX	AGE	SAFE	CF	TEST	BAC	MISS
1	1	2	1	3	2	1	0	0	0
	2	2	1	3	2	1	0	0	0
	3	2	1	3	2	1	0	0	0
2	1	2	1	2	1	1	0	0	0
	2	2	1	2	1	1	1	133	0
	3	2	1	2	1	1	1	133	0
3	1	2	1	2	2	2	0	0	0
	2	2	1	2	2	2	0	0	0
	3	2	1	2	2	2	0	0	0
4	1	1	1	1	1	2	1	85	0
	2	1	1	4	1	2	0	0	0
	3	1	1	3	1	2	0	0	0
5	1	2	1	4	2	1	1	66	0
	2	2	1	4	2	1	1	66	0
	3	2	1	4	2	1	1	66	0
6	1	3	2	3	2	2	1	91	0
	2	3	2	3	2	2	1	91	0
	3	3	2	3	2	2	1	91	0
7	1	1	2	1	1	2	1	156	0
	2	1	1	1	1	2	1	156	0
	3	1	1	1	1	2	1	156	0
8	1	2	2	1	1	1	1	23	0
	2	2	2	1	1	1	1	23	0
	3	2	2	1	1	1	1	23	0
9	1	2	1	2	2	1	0	0	0
	2	2	1	2	2	2	1	45	0
	3	2	1	2	2	1	0	0	0
10	1	2	1	3	2	1	0	0	0
	2	2	1	3	2	1	0	0	0
	3	2	1	3	2	2	0	0	0
11	1	2	1	3	2	2	1	165	0
	2	2	1	3	2	2	1	165	0
	3	2	1	3	2	2	0	0	0
12	1	1	1	5	1	1	0	0	0
	2	1	1	5	2	1	1	94	0
	3	1	1	5	1	1	1	45	0
13	1	1	2	4	1	2	0	0	0
	2	1	2	4	2	2	0	0	0
	3	1	2	4	2	2	0	0	0
14	1	2	2	4	1	1	1	118	0
	2	2	2	4	1	1	1	118	0
	3	2	2	4	1	1	1	118	0
Missing:		0	0	0	0	0	0	0	

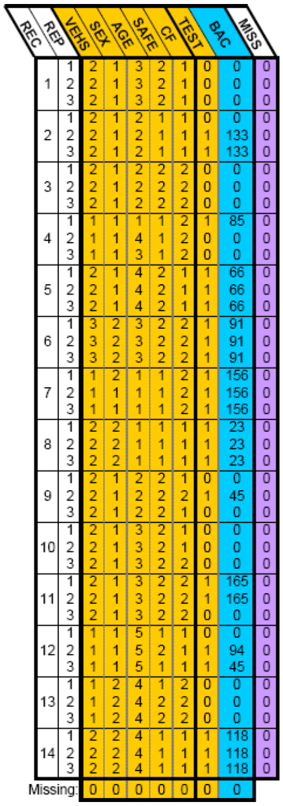
Deflating the Data Set

In this example, we assume that the threshold a is 0.5: if more than 50% of the replicates for a given record have $Z_1 = 1$, the record has $Z_1 = 1$ and its value for Z_2 is the

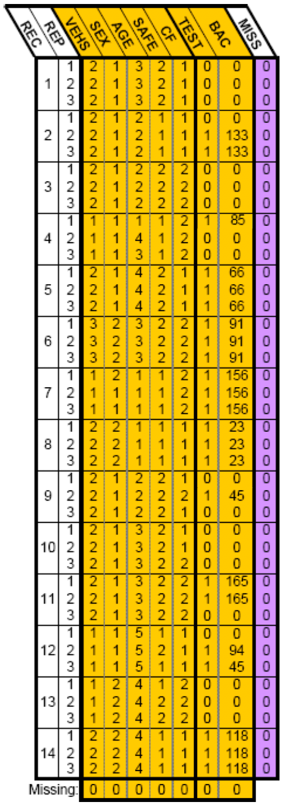
BAC average taken over all its non-zero Z_1 replicates. The final results are shown in the last two tables below: red entries indicate records for which alcohol was deemed to have played a factor.



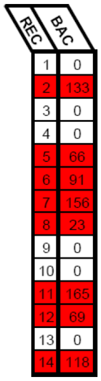
REC	REP	VEHS	SEX	AGE	SAFE	CF	TEST	BAC	MISS
1	1	2	1	3	2	1	0	0	0
	2	2	1	3	2	1	0	0	0
	3	2	1	3	2	1	0	0	0
2	1	2	1	2	1	1	0	0	0
	2	2	1	2	1	1	1	.	1
	3	2	1	2	1	1	1	.	1
3	1	2	1	2	2	2	0	0	0
	2	2	1	2	2	2	0	0	0
	3	2	1	2	2	2	0	0	0
4	1	1	1	1	1	2	1	.	1
	2	1	1	4	1	2	0	0	0
	3	1	1	3	1	2	0	0	0
5	1	2	1	4	2	1	1	.	1
	2	2	1	4	2	1	1	.	1
	3	2	1	4	2	1	1	.	1
6	1	3	2	3	2	2	1	91	0
	2	3	2	3	2	2	1	91	0
	3	3	2	3	2	2	1	91	0
7	1	1	2	1	1	2	1	156	0
	2	1	1	1	1	2	1	156	0
	3	1	1	1	1	2	1	156	0
8	1	2	2	1	1	1	1	23	0
	2	2	2	1	1	1	1	23	0
	3	2	2	1	1	1	1	23	0
9	1	2	1	2	2	1	0	0	0
	2	2	1	2	2	2	1	.	1
	3	2	1	2	2	1	0	0	0
10	1	2	1	3	2	1	0	0	0
	2	2	1	3	2	1	0	0	0
	3	2	1	3	2	2	0	0	0
11	1	2	1	3	2	2	1	.	1
	2	2	1	3	2	2	1	.	1
	3	2	1	3	2	2	0	0	0
12	1	1	1	5	1	1	0	0	0
	2	1	1	5	2	1	1	.	1
	3	1	1	5	1	1	1	.	1
13	1	1	2	4	1	2	0	0	0
	2	1	2	4	2	2	0	0	0
	3	1	2	4	2	2	0	0	0
14	1	2	2	4	1	1	1	118	0
	2	2	2	4	1	1	1	118	0
	3	2	2	4	1	1	1	118	0
Missing:		0	0	0	0	0	0	11	



REC	REP	VEHS	SEX	AGE	SAFE	CF	TEST	BAC	MISS
1	1	2	1	3	2	1	0	0	0
	2	2	1	3	2	1	0	0	0
	3	2	1	3	2	1	0	0	0
2	1	2	1	2	1	1	0	0	0
	2	2	1	2	1	1	1	133	0
	3	2	1	2	1	1	1	133	0
3	1	2	1	2	2	2	0	0	0
	2	2	1	2	2	2	0	0	0
	3	2	1	2	2	2	0	0	0
4	1	1	1	1	1	2	1	85	0
	2	1	1	4	1	2	0	0	0
	3	1	1	3	1	2	0	0	0
5	1	2	1	4	2	1	1	66	0
	2	2	1	4	2	1	1	66	0
	3	2	1	4	2	1	1	66	0
6	1	3	2	3	2	2	1	91	0
	2	3	2	3	2	2	1	91	0
	3	3	2	3	2	2	1	91	0
7	1	1	2	1	1	2	1	156	0
	2	1	1	1	1	2	1	156	0
	3	1	1	1	1	2	1	156	0
8	1	2	2	1	1	1	1	23	0
	2	2	2	1	1	1	1	23	0
	3	2	2	1	1	1	1	23	0
9	1	2	1	2	2	1	0	0	0
	2	2	1	2	2	2	1	45	0
	3	2	1	2	2	1	0	0	0
10	1	2	1	3	2	1	0	0	0
	2	2	1	3	2	1	0	0	0
	3	2	1	3	2	2	0	0	0
11	1	2	1	3	2	2	1	165	0
	2	2	1	3	2	2	1	165	0
	3	2	1	3	2	2	0	0	0
12	1	1	1	5	1	1	0	0	0
	2	1	1	5	2	1	1	94	0
	3	1	1	5	1	1	1	45	0
13	1	1	2	4	1	2	0	0	0
	2	1	2	4	2	2	0	0	0
	3	1	2	4	2	2	0	0	0
14	1	2	2	4	1	1	1	118	0
	2	2	2	4	1	1	1	118	0
	3	2	2	4	1	1	1	118	0
Missing:		0	0	0	0	0	0	0	



REC	REP	VEHS	SEX	AGE	SAFE	CF	TEST	BAC	MISS
1	1	2	1	3	2	1	0	0	0
	2	2	1	3	2	1	0	0	0
	3	2	1	3	2	1	0	0	0
2	1	2	1	2	1	1	0	0	0
	2	2	1	2	1	1	1	133	0
	3	2	1	2	1	1	1	133	0
3	1	2	1	2	2	2	0	0	0
	2	2	1	2	2	2	0	0	0
	3	2	1	2	2	2	0	0	0
4	1	1	1	1	1	2	1	85	0
	2	1	1	4	1	2	0	0	0
	3	1	1	3	1	2	0	0	0
5	1	2	1	4	2	1	1	66	0
	2	2	1	4	2	1	1	66	0
	3	2	1	4	2	1	1	66	0
6	1	3	2	3	2	2	1	91	0
	2	3	2	3	2	2	1	91	0
	3	3	2	3	2	2	1	91	0
7	1	1	2	1	1	2	1	156	0
	2	1	1	1	1	2	1	156	0
	3	1	1	1	1	2	1	156	0
8	1	2	2	1	1	1	1	23	0
	2	2	2	1	1	1	1	23	0
	3	2	2	1	1	1	1	23	0
9	1	2	1	2	2	1	0	0	0
	2	2	1	2	2	2	1	45	0
	3	2	1	2	2	1	0	0	0
10	1	2	1	3	2	1	0	0	0
	2	2	1	3	2	1	0	0	0
	3	2	1	3	2	2	0	0	0
11	1	2	1	3	2	2	1	165	0
	2	2	1	3	2	2	1	165	0
	3	2	1	3	2	2	0	0	0
12	1	1	1	5	1	1	0	0	0
	2	1	1	5	2	1	1	94	0
	3	1	1	5	1	1	1	45	0
13	1	1	2	4	1	2	0	0	0
	2	1	2	4	2	2	0	0	0
	3	1	2	4	2	2	0	0	0
14	1	2	2	4	1	1	1	118	0
	2	2	2	4	1	1	1	118	0
	3	2	2	4	1	1	1	118	0
Missing:		0	0	0	0	0	0	0	



REC	BAC
1	0
2	133
3	0
4	0
5	66
6	91
7	156
8	23
9	0
10	0
11	165
12	69
13	0
14	118

6. Results for Ontario (2007)

In this section, we show the results of our BAC imputation algorithm for fatal collisions occurring in Ontario during the year 2007. The data set also contains the collisions from 2000 to 2005 (which were the only data available when the algorithm was originally conceived).

Throughout, missing values of categorical variables are imputed using SAS 9.2's proc logit.

There were $n = 9689$ records in the combined databases. Early trials confirmed that $k = 9$ replications eliminated all convergence errors in the logistic regression routine used by SAS. Since using more replicates can only improve the method, we use $k = 10$ in order to conform with [4]. Furthermore, analysis of existing BAC levels determined that $A = 500$ would be a reasonable upper limit to use. By comparison, a BAC level of 80 is the threshold for impaired driving in Ontario.

The frequency tables for the explanatory variables in the replicated records are shown below.

P_11	Frequency	Percent
1	87940	90.76
2	8950	9.24

C_WDAY_GR	Frequency	Percent
1	50470	52.09
2	46420	47.91

C_HOUR_GR	Frequency	Percent
1	13310	13.78
2	13490	13.97
3	30230	31.31
4	25100	25.99
5	14430	14.94

Frequency Missing = 330

C_VEHS_GR	Frequency	Percent
1	30260	31.23
2	46730	48.23
3	19900	20.54

P_SEX_GR	Frequency	Percent
1	73790	76.55
2	22600	23.45

Frequency Missing = 500

P_AGE_GR	Frequency	Percent
1	9170	9.72
2	19750	20.92
3	17240	18.26
4	18490	19.59
5	13260	14.05
6	16480	17.46

Frequency Missing = 2500

P_SAFE_GR	Frequency	Percent
1	10560	11.68
2	62380	69.00
3	17460	19.31

Frequency Missing = 6490

V_CF_GR	Frequency	Percent
1	12290	13.20
2	80820	86.80

Frequency Missing = 3780

The number of replicated records with specific numbers of missing explanatory variables indicate that first-, second-, third- and fourth-order imputation of explanatory variables will be necessary.

vari	Frequency	Percent
0	84830	87.55
1	10750	11.10
2	1100	1.14
3	190	0.20
4	20	0.02

This means that 10750 first-order imputations, 1100 second-order imputations, 190 third-order and 20 fourth-order imputations were needed to obtain a complete set of replicated records.

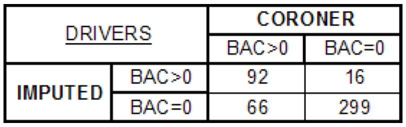
Once the values of Z_1 were imputed (using an extensive SAS program, written to implement the BAC Imputation Algorithm described above), we used a threshold $a = 0.5$ to determine whether a record had zero or non-zero BAC: if more than 50% of the replicates for a given record had $Z_1 = 1$, the record itself was assumed to have non-zero BAC, which was then imputed as follows.

The existing BAC levels were first transformed according to

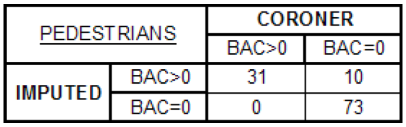
$$\hat{Z}_2 = \tan\left(\frac{\pi}{500}Z_2 - \frac{\pi}{2}\right),$$

in effect carrying the range of Z_2 from $(0, 500)$ to $(-\infty, \infty)$. SAS 9.2's proc glm was then used to impute $\hat{Z}_2$ for the missing values, and the inverse transformation provided the imputed Z_2 values.

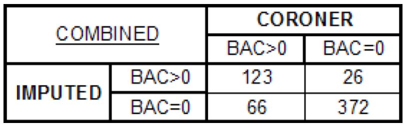
It is impossible to present the specific results of the imputation due to spatial considerations. It is however possible to compare the results of the imputation with validated data, that is, with the actual BAC value provided by the Coroner's report once those became available. Only the imputation results for Z_1 are presented as validation data for the actual BAC level Z_2 was not made available to the author at the time this paper was written. As can be seen, the performance for pedestrian fatalities was slightly better than the performance for driver fatalities when imputing BAC for fatal collisions occurring in Ontario during 2007.



DRIVERS		CORONER	
		BAC>0	BAC=0
IMPUTED	BAC>0	92	16
	BAC=0	66	299



PEDESTRIANS		CORONER	
		BAC>0	BAC=0
IMPUTED	BAC>0	31	10
	BAC=0	0	73



COMBINED		CORONER	
		BAC>0	BAC=0
IMPUTED	BAC>0	123	26
	BAC=0	66	372

Metric	Drivers	Pedestrians	Combined
Accuracy	82.66%	91.23%	84.33%
Precision (PPV)	85.19%	75.61%	82.55%
Neg Pred Value	81.92%	100.00%	84.93%
Sensitivity	58.23%	100.00%	65.08%
Specificity	94.92%	87.95%	93.47%
False Pos Rate (α)	5.08%	12.05%	6.53%
False Neg Rate (β)	41.77%	0.00%	34.92%
Pos L'hood Ratio	11.46	8.30	9.96
Neg L'hood Ratio	0.44	0.00	0.37
$F-$score	0.69	0.86	0.73

7. Conclusion

We used "naive" logistic regression and a basic general linear model for the categorical variables and the continuous BAC level variable, respectively. More sophisticated or better-suited imputation methods could no doubt improve the power of our algorithm. And while we were able to obtain various metrics for our algorithm when applied to the 2007 Ontario data, it would be beneficial to compare those results with those that would be obtained using other methods, specifically those of [3, 4].

Consulting Post-Mortem

- The client needed results **quickly**, which did not leave much time to fine-tune the model (playing around with various models and transformations, etc);
- at the client's request, more emphasis was placed on Z_1 than Z_2, but perhaps Z_2 would have been a **more important quantity to impute**, since a low amount of BAC is legally allowed (although that is a more difficult imputation problem);
- the client put a lot of faith in the idea that BAC absence/presence should be easy to impute **accurately** (in the high 90%s, in spite of small number of explanatory variables available);
- the risk of **overfitting** was high, given that no performance evaluation was conducted until validation.

In retrospect, while the algorithm did what was asked of it, we suspect that it was neither robust or sophisticated enough to be useful in a more general setting.

References

[1] Chouinard, A. [2010], Personal conversation.

[2] Michaud, I. and Gough, H. [2008], *Documentation of a Multiple Imputation Methodology For Transport Canada and the Ontario Ministry of Transportation*, Statistical Consultation Group, Statistics Canada, Ottawa.

[3] Raghunathan, T.E., Lepkowski, J.M., van Hoewyk, J. and Solenberger, P. [2001], 'A multivariate technique for multiply imputing missing values using a sequence of regression models', in *Survey Methodology*, 27(1):85-95.

[4] Rubin, D.B., Schafer, J.L. and Subramanian, R. [1998], *Multiple Imputation of Missing Blood Alcohol Concentration (BAC) Values in FARS*, NHTSA, DOT HS 808 816, Springfield, VA.

[5] Russell, R. [2010], 'Sobering stats on drunk drivers' in the *Globe and Mail* (online).

[6] TIRF [2006], Transport Canada leaflet on alcohol use by drivers.