Blood Alcohol Content Imputation

Practical Data Processing

P. Boily
Centre for Quantitative Analysis and Decision Support
Idlewyld Analytics and Consulting Services

Contents

Project Description

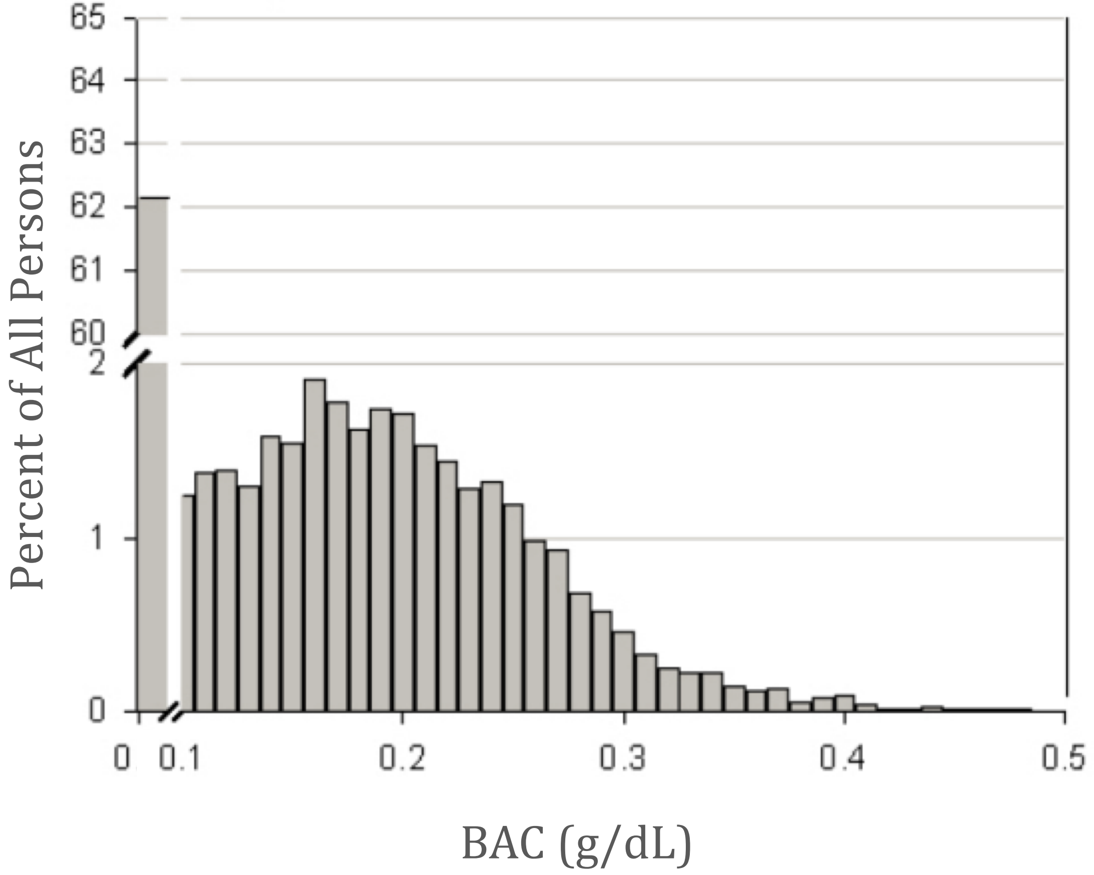
Project Description
BAC Imputation

- Fatal collisions often involve **alcohol** (driver, pedestrian, cyclist).
- Breathalyzer tests cannot be conducted on deceased individuals, so the presence of alcohol in the blood cannot be confirmed until the coroner's report is available.
- For various reasons, these reports can take **up to a year** to produce.
- The **blood alcohol concentration** (BAC) levels may not make their ways to interested parties in a **timely fashion**.
- This can cause **delays** in policy implementation and could possibly lead to otherwise preventable deaths.
- Data analysts often resort to **imputation methods** in order to make an informed guess as to the BAC level in fatal collisions.
- This is what the *Ministry of Transportation of Ontario* (MTO) was looking for in 2007: using a small number of features (many of which are themselves missing values), is it possible to
  - predict whether alcohol was involved, and if so,
  - predict the BAC level?

Project Description
NHTSA Imputation Algorithm

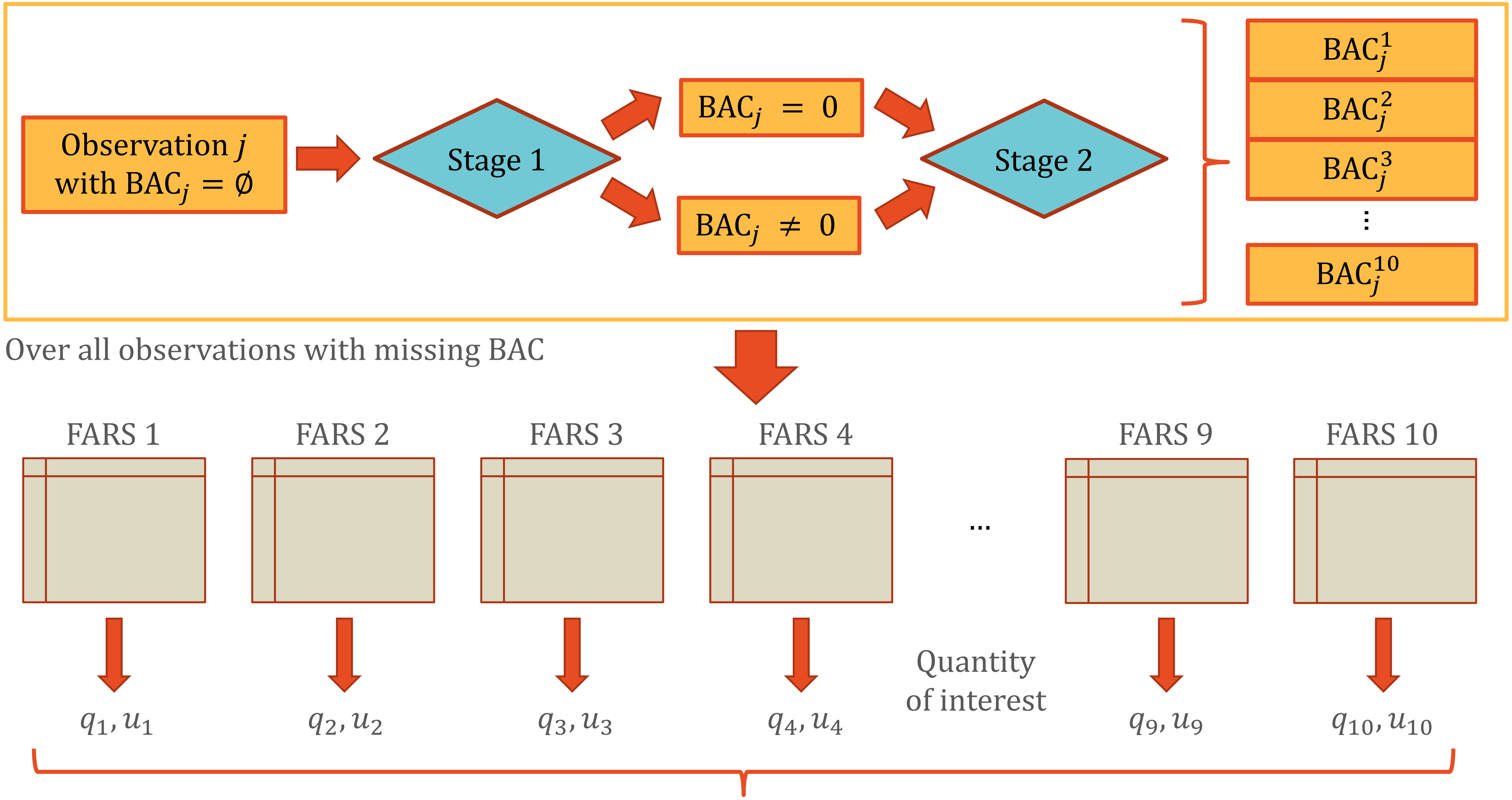
- According to preliminary estimates for 2002, alcohol was involved in about **42%** of all motor vehicle crashes where there was a fatality in the United States.
- BAC levels were **missing from 58%** of fatality reports in 2001.
- The distribution of BAC levels for observations for which it was provided is **semi-continuous**; about 62% of the units have BAC=0, and 38% fall in the range 0 < BAC < 0.94.
- Responses above 0.4 are sparse.

Distribution of Reported BAC Among Actively-Involved Persons in Fatal Crashes, 2001



Project Description
NHTSA Imputation Algorithm

- The U.S.A.'s *National Highway Traffic Safety Administration* (NHTSA) uses a **two-stage model**:
  1. impute **zero/non-zero** BAC status through a multivariate procedure (details can be found in Subramanian and Utter's paper), and
  2. **conditional** on non-zero BAC, they impute 10 BAC levels for each missing BAC value *via* a general linear model (for zero BAC, the 10 BAC levels are all set at 0).
- This creates 10 (potentially *different*) versions of the dataset with **no missing BAC values**.
- The analysis of interest is conducted **10 separate times**, once on each of the distinct versions
- This allows for **valid statistical inferences** and for **confidence intervals** to be drawn.
- The main drawback of this method is that the **values of some explanatory variables may be missing** for a large number of records; these missing values are treated as belonging to a separate category (one for each variable): that of '***missing value***'.
- As there may be many disparate reasons to explain why different records are missing a given variable's value, this may lead to a **loss of information**, which translates into a **less powerful** imputation method.



$$\bar{Q}, T, \quad \text{where } T = \bar{U} + \frac{11}{10}\hat{\sigma}_Q^2$$

Project Description
NHTSA Imputation Algorithm

- **Validation:** for 5 years in the FARS data base, 25% of observations for which BAC was known were removed.
- Removed BAC values were estimated using the 2-stage algorithm.
- Comparison with known values are shown in the table.
- Assumed missing mechanism: MCAR
- Evidence suggests that this is not an appropriate assumption – observations with missing BAC levels are **much more likely to be 0**, everything else being equal.

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

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

Project Description
Regression Sequences

- In the case of multiple missing values in the **explanatory variables**, a possible solution is to use a **sequence of regression models**.

- Missing values for each explanatory variable are imputed as follows:
  1. the explanatory variable Y_1 with the **fewest missing values** is imputed to $\tilde{Y}_1$, using the explanatory variables $\boldsymbol{X}$ with **no missing values** ($\tilde{Y}_1$ contains no missing values).
  2. the explanatory variable Y_2 with the **next fewest missing values** is imputed to $\tilde{Y}_2$ using the explanatory variables $\{\boldsymbol{X}, \tilde{Y}_1\}$ ($\tilde{Y}_2$ contains no missing values).
  3. …
  4. the process continues in sequence **until the last remaining explanatory variable with missing values** Y_m is imputed to $\tilde{Y}_m$ using $\{\boldsymbol{X}, \tilde{Y}_1, \dots, \tilde{Y}_{m-1}\}$. At this point, there are no more missing values in the dataset.

- The main drawback of this method is that some information might be "**hiding**" in $\{Y_2, \dots, Y_m\}$ which, combined with the information found in $\boldsymbol{X}$, could provide a better imputation for Y_1 than $\tilde{Y}_1$.

Objective: combine two approaches while removing their respective drawbacks… but with the caveat that there is no future use: the MTO simply wanted a predicted BAC.

Data Preparation and Methodology

NCDB Data

- Our algorithm 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
- Start by removing all records involving **non-fatal collisions** and all records involving **non-drivers** or **non-pedestrians**
- There are two BAC-linked target variables (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, a MANOVA identified a subset of NCDB variables as having a significant effect on the target variables.

Imputation Variables

- Retained (and ***binned***) 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** by the deceased (P_SAFE);
  - the **hour** (C_HOUR) & **weekday** (C_WDAY) when the collision occurred;
  - the **number of vehicles/pedestrians** involved in the collision (C_VEHS), and
  - **various contributing factors** as determined by police officers on the scene (V_CF1–V_CF4).
- V_CF_GR might be expected to be a more significant predictor of BAC, but preliminary analyses show that it is not **any more significant** than other retained variables.

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'

Methodology
Inflating the Data Set

- Original data set contains n records.
- **Replicate** the data set $k \geq 1$ times, where k is selected in order to create a large enough data set to produce statistically meaningful results.
- Replicated data set contains kn records.
- If $n \gg 1$ or if there is no **systematic pattern** in the missing values, small values of k can be used.
- When n is smaller, larger values of k must be used
- **Aim:** impute `TEST` and `P_BAC1F`

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			

$k = 3$

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			

Methodology

Step 1–1: 1st Order Imputation

- ❑ If there are explanatory variables that have **no missing value**, they do not need to be processed – yellow in the example
- ❑ Among the remaining explanatory variables, find the one with the **fewest missing values** (tie: pick at random) – **blue** in the example
- ❑ The records for which that value is missing will be **imputed** – **brown** in the example
- ❑ The records for which the values of the other explanatory variables are not missing constitute the **training set** for imputation – **green** in the example
- ❑ If the training set is **too small**, there might be issues with the quality of imputation.

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			

imputation

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			

Methodology

Step 1–2: 1st Order Imputation

- If there are explanatory variables that have **no missing value**, they do not need to be processed – yellow in the example
- Among the remaining explanatory variables, find the one with the fewest missing values (tie: pick at random) – blue in the example
- The records for which that value is missing will be **imputed** – **brown** in the example
- The records for which the values of the other explanatory variables are not missing constitute the **training set** for imputation – **green** in the example
- The imputation method is left to the analyst – it could even vary from one step to the next.

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			

imputation →

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			

Methodology

Step 1–3: 1st Order Imputation

- If there are explanatory variables that have **no missing value**, they do not need to be processed – yellow in the example
- Among the remaining explanatory variables, find the one with the **fewest missing values** (tie: pick at random) – blue in the example
- The records for which that value is missing will be **imputed** – **brown** in the example
- The records for which the values of the other explanatory variables are not missing constitute the **training set** for imputation – **green** in the example
- Note that some of the missing values may end up not being imputed (why?) – see **red** box

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			

imputation

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			

Methodology

Step 1–4: 1st Order Imputation

- The processed explanatory variables are shown in yellow in the example
- In general, more than one record will be imputed at every step – see **red** box.
- **At most** m_1 first-order imputations can be conducted; $m_1 = \#$ of explanatory variables
- 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.

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

imputation

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

Methodology

Crossing and Uncrossing Variables

- Two variables X_1 and X_2 are **crossed** into $X_{1,2}$ as follows:
  - assume that X_1's levels are $\{1, \dots, n_1\}$
  - assume that X_2's levels are $\{1, \dots, n_2\}$
  - there are $n_1 \times n_2$ distinct crossed levels

$$\mathcal{A} = \{(1,1), \dots, (n_1, 1), (1,2), \dots, (n_1, 2), \dots, (1, n_2), \dots, (n_1, n_2)\}$$

  - construct a bijection $f_{1,2}\colon \mathcal{A} \to \{1, \dots, n_1 \times n_2\}$
  - (there are many such bijections)
  - if $X_1 = i$ and $X_2 = j$, then $X_{1,2} = f_{1,2}(i, j)$

- The variable $X_{1,2}$ is **uncrossed** into X_1 and X_2 as follows:
  - if $X_{1,2} = \alpha$, then $(X_1, X_2) = f_{1,2}^{-1}(\alpha)$

- There is no need to cross variables for which **there are no missing records**

- Imputation proceeds as before (**training set**, **imputing set**, **imputed variable**, etc.)

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	*	*	*	*	3			

Methodology

Step 2: Second-Order Imputation

- This process is repeated until the imputation of missing values of the last remaining crossed explanatory variable

- Imputation of the explanatory variables requires **uncrossing** of the imputed crossed variable

- 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.

- **No more than** $0.5m_1(m_1+1)$ second-order imputations will be conducted

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

uncrossing

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

Methodology
Continuation

- This process is repeated with **triplets** of explanatory variables, then **quadruplets**, and so on, until the dataset contains no record with missing values of the explanatory variables
- There is a danger: at every new step, we (potentially) use imputed values as if they were **actual** values, and these imputed values are in turn used to impute new values.
- Like all imputation methodologies, this procedure works best when the number of missing values is **small** relative to the number of total observations.
- A potential solution is to set k **large enough**, but that might be accompanied by an increase in computational time.
- **The proof of the eating is in the pudding**: in this application, the goal is to predict the presence/absence of BAC and its accompanying levels. How well does the procedure perform?

Methodology
Step 3: Target Variables

- At this stage there are **no missing values** in the explanatory variables – yellow in the example
- The categorical variable `TEST` (Z_1) is imputed in the **same manner** as the explanatory variables

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

imputation

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

Methodology

Step 3: Target Variables

- At this stage there are **no missing values** in the explanatory variables – yellow in the example
- The numerical variable `P_BAC1F` (Z_2) requires a different imputation framework, perhaps a general linear model (after an appropriate transformation)

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	

imputation

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	

Methodology

Deflating the Data Set

- At this stage, for each of the n original records, there are k values of for each of Z_1 and Z_2.
- Pick some **threshold** $a \in (0,1)$
- Let $p_{1,i}$ be the **proportion** of the i^{th} record's k replicates for which $Z_1 = 1$.
- Set $Z_1^i = \begin{cases} 1, \text{if } p_{1,i} > a \\ 0, \text{else} \end{cases}$
- Let $\overline{Z_2^i}$ be the average of the i^{th} record's Z_2 values, weighted by their Z_1 values.
- Set $Z_2^i = \begin{cases} \overline{Z_2^i}, \text{if } p_{1,i} > a \\ 0, \text{else} \end{cases}$

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

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

Results

Data

- We impute BAC levels for those fatal collisions occurring in **Ontario** during the year **2007** for which data is not available (**587 records** in total).
- The data set also contains the collisions from 2000 to 2005
- Missing values of categorical variables are imputed using SAS 9.2's **`proc logit`**.
- There were $\boldsymbol{n = 9689}$ records in the combined databases.
- Early trials confirmed that $k > 9$ replications eliminated all convergence errors in the logistic regression routine used by SAS. We use $\boldsymbol{k = 10}$.
- Furthermore, analysis of existing BAC levels determined that $\boldsymbol{A = 500}$ **mg/dL** is a reasonable upper limit for BAC levels.
- By comparison, a BAC level of **80 mg/dL** is the threshold for impaired driving in Ontario.

Data

- 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

Univariate Frequency Counts for Explanatory Variables

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

Distribution of Records with 0, 1, 2, 3, and 4 Missing Explanatory Variables Values.

Imputation

- 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, we used the threshold $\boldsymbol{a = 0.5}$ 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, the record itself was assumed to have non-zero BAC
- The existing BAC levels were first transformed according to

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

  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.

Results and Validation (Z_1)

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%
Negative Predictive Value	81.92%	100.00%	84.93%
Sensitivity	58.23%	100.00%	65.08%
Specificity	94.92%	87.95%	93.47%
False Positive Rate (α)	5.08%	12.05%	6.53%
False Negative Rate (β)	41.77%	0.00%	34.92%
Positive Likelihood Ratio	11.46	8.30	9.96
Negative Likelihood Ratio	0.44	0.00	0.37
F-score	0.69	0.86	0.73

Consulting Post-Mortem

Consulting Post-Mortem

- Client needed results **quickly**
  - didn't leave much time to fine-tune the model (playing around with various predictive models and transformations, etc)
- More emphasis was placed on Z_1 than Z_2, at the client's behest, but Z_2 would have been a **more important quantity to impute** (a certain amount of BAC is legally allowed)
  - numerical values harder to impute
- Client put a lot of faith in the idea that BAC absence/presence should be easy to impute **accurately**
  - felt that accuracy should have been in high 90s, in spite of small number of explanatory variables available
- The threshold value provides an **estimate of the variance in Z_1**, but in general, uncertainty was not going to be used in ulterior analyses – this simplified the algorithm design.
- **Overfitting** issues? No performance evaluation was conducted until validation – **risky**.
- In retrospect, while the algorithm did what was asked of it, I feel that it is neither robust enough or sophisticated enough. I lucked out.