

Contents

1	Survey of Quantitative Methods	2
1.2	Data Preparation	3
1.2.1	General Principles	3
1.2.2	Data Quality	6
1.2.3	Missing Values	8
1.2.4	Anomalous Observations	12
1.2.5	Data Transformation	16
1.2.6	Case Study: Imputation of Blood Alcohol Content Levels	21

List of Figures

1	Accuracy as bias, precision as standard error.	7
2	An illustration of heaping	7
3	Dr. Vanderwhede's original <i>Advanced Retroencabulation</i> dataset	10
4	Imputed values for Dr. Vanderwhede's dataset	11
5	Tukey's boxplot test for outliers	14
6	Summary visualisations for a plant dataset	16
7	Visualisations for a service point dataset	16
8	Illustration of the curse of dimensionality	18
9	Examples of data transformations	19

List of Tables

1	Data cleaning bingo card.	5
2	Summary data for an (artificial) medical dataset.	8
3	Summary and visualisation for an appendage length dataset	17

1 Survey of Quantitative Methods

The bread and butter of quantitative consulting is the ability to apply quantitative methods to business problems in order to obtain actionable insight. Clearly, it is impossible (and perhaps inadvisable, in a more general sense) for any given individual to have expertise in every field of mathematics, statistics, and computer science.

We believe that the best consulting framework is reached when a small team of consultants possesses expertise in 2 or 3 areas, as well as a decent understanding of related disciplines, and a passing knowledge in a variety of other domains: this includes keeping up with trends, implementing knowledge redundancies on the team, being conversant in non-expertise areas, and knowing where to find detailed information (online, in books, or through external resources).

In this section, we present an introduction for 9 “domains” of quantitative analysis:

- survey sampling and data collection;
- data processing;
- data visualisation;
- statistical methods;
- queueing models;
- data science and machine learning;
- simulations;
- optimisation, and
- trend extraction and forecasting;

Strictly speaking, the domains are not free of overlaps. Large swaths of data science and time series analysis methods are quite simply statistical in nature, and it’s not unusual to view optimisation methods and queueing models as sub-disciplines of operations research. Other topics could also have been included (such as Bayesian data analysis or signal processing, to name but two), and might find their way into a second edition of this book.

Our treatment of these topics, by design, is brief and incomplete. Each module is directed at students who have a background in other quantitative methods, but not necessarily in the topic under consideration. Our goal is to provide a quick “reference map” of the field, together with a general idea of its challenges and common traps, in order to highlight opportunities for application in a consulting context. These subsections are emphatically NOT meant as comprehensive surveys: they focus on the basics and talking points; perhaps more importantly, a copious number of references are also provided.

We will start by introducing a number of motivating problems, which, for the most part, we have encountered in our own practices. Some of these examples are reported on in more details in subsequent sections, accompanied with (partial) deliverables in the form of charts, case study write-ups, report extract, etc.).

As a final note, we would like to stress the following: it is **IMPERATIVE** that quantitative consultants remember that acceptable business solutions are not always optimal theoretical solutions. Rigour, while encouraged, often must take a backseat to applicability. This lesson can be difficult to accept, and has been the downfall of many a promising candidate.

1.2 Data Preparation

Data Validation

Martin Kerdaniel: Data is messy, Alison.

Alison MacIntosh: Even when it's been cleaned?

Martin Kerdaniel: Especially when it's been cleaned.

– P. Boily, I. Kiewiet, *The Great Balancing Act*.

Once the raw data has been collected and stored in a dataset that is accessible to the quantitative consultants, the focus should shift to data cleaning and processing. This requires testing for **soundness** and fixing **errors**, designing and implementing strategies to deal with **missing values** and **outlying/influential observations**, as well as low-level **exploratory data analysis** and **visualisation** to determine what **data transformations** and **dimension reduction** approaches will be needed in the final analysis. Consultants should be prepared to spend up to 80% of their time processing and cleaning the data.

The following remarks must be taken to heart during this stage:

- Processing should **NEVER** be done on the original dataset – make copies along the way.
- **ALL** cleaning steps and procedures need to be documented.
- If **too much** of the data requires cleaning up, the data collection procedure might need to be **revisited**.
- An entire record should only be discarded as a **last resort**.

Another thing to keep in mind is that cleaning and processing may need to take place more than once depending on the type of data collection (one pass, batch, continuously).

Finally, note that we are assuming that the datasets of interest contain only numerical and/or categorical observations. Additional steps must be taken when dealing with unstructured data, such as text or images.

1.2.1 General Principles

Data Validation

Dilbert: I didn't have any accurate numbers, so I just made up this one. Studies have shown that accurate numbers aren't any more useful than the ones you make up.

Pointy-Haired Boss: How many studies showed that?

Dilbert: [*beat*] Eighty-seven.

– Scott Adams, **Dilbert**, 8 May 2008

Approaches to Data Cleaning There are two main **philosophical** approaches to data cleaning and validation, which we call

- methodical, and
- narrative.

The **methodical** approach consists in running through a **check list** of potential issues and flagging those that apply to the data. The **narrative** approach, on the other hand, consists in **exploring** the dataset while searching for unlikely or irregular patterns. Which approach the consultant opts to follow depends on a number of factors, not least of which is the client's needs and views on the matter – consultants have a responsibility to discuss this point with the clients.

Pros and Cons The methodical approach focuses on **syntax**; the check-list is typically **context-independent**, which means that it (or a subset) can be reused from one project to another, which makes data analysis pipelines **easy to implement** and **automate**. In the same vein, common errors are **easily identified**. On the flip side, the check list may be quite extensive and the entire process may prove **time-consuming**. The biggest disadvantage of this approach is that it makes it difficult to identify **new types of errors**.

The narrative approach focuses on **semantics**; even false starts may simultaneously produce **data understanding** prior to switching to a more mechanical approach. It is easy, however, to miss important sources of errors and invalid observations when the datasets have a **large number of features**. There is an additional downside: **domain expertise**, coupled with the narrative approach, may bias the process by neglecting “uninteresting” areas of the dataset.

Tools and Methods An non-exhaustive list of common data issues can be found in the *Data Cleaning Bingo Card* (see Table 1); there are obviously other possibilities. Other methods include

- **visualisations** – see Section ??;
- **data summaries** – # of missing observations; 5-pt summary, mean, standard deviation, skew, kurtosis, for numerical variables; distributional tables for categorical variables;
- **n-way tables** – counts for joint distributions of categorical variables;
- **small multiples** – tables/visualisations indexed along categorical variables, and
- **preliminary data analyses** – which may provide “huh, that's odd...” realisations.

IMPORTANT NOTE: there is nothing wrong with running a number of analyses to flush out data issues, but remember to label your initial forays as **preliminary** analyses. From the client's perspective, repeated analyses may create a sense of unease and distrust, even if they form a crucial part of the analytical process (doing so will also facilitate invoicing).

In our (admittedly biased and incomplete) experience, **computer scientists** and **programmers** tend to naturally favour the methodical approach, while **mathematicians** and **statisticians** tend to naturally favour the narrative approach (although we have met plenty of individuals with unexpected backgrounds in both camps). Quantitative consultants should be comfortable with **both** approaches.

The narrative approach is akin to working out a crossword puzzle with a pen and putting down potentially erroneous answers once in a while to try to open up the grid, so to speak. The mechanical approach, on the other hand, is similar to working out the puzzle with a pencil and a dictionary, only putting down answers when their correctness is guaranteed. More puzzles get solved when using the first approach, but mistakes tend to be spectacular. Not as many puzzles get solved the second way, but the trade-off is that it leads to fewer mistakes.

random missing values	outliers	values outside of expected range - numeric	factors incorrectly/inconsistently coded	date/time values in multiple formats
impossible numeric values	leading or trailing white space	badly formatted date/time values	non-random missing values	logical inconsistencies across fields
characters in numeric field	values outside of expected range - date/time	DCB!	inconsistent or no distinction between null, 0, not available, not applicable, missing	possible factors missing
multiple symbols used for missing values	???	fields incorrectly separated in row	blank fields	logical inconsistencies within field
entire blank rows	character encoding issues	duplicate value in unique field	non-factor values in factor	numeric values in character field

Table 1: Data cleaning bingo card.

1.2.2 Data Quality

The Importance of Validation

Calvin's Dad: OK Calvin. Let's check over your math homework.

Calvin: Let's not and say we did.

Calvin's Dad: Your teacher says you need to spend more time on it. Have a seat.

Calvin: More time?! I already spent 10 whole minutes on it! 10 minutes shot! Wasted! Down the drain!

Calvin's Dad: You've written here $8 + 4 = 7$. Now you know that's not right.

Calvin: So I was off a little bit. Sue me.

Calvin's Dad: You can't **add** things and come with **less** than you started with!

Calvin: I can do that! It's a free country! I've got my rights!

– Bill Watterson, *Calvin and Hobbes*, 15 September 1990.

The quality of the data has an important effect of the quality of the results: as the old computer science saying goes: “garbage in, garbage out.”

Data is said to be **sound** when it has as few issues as possible with

- **validity** – are observations sensible, given data type, range, mandatory response, uniqueness, value, regular expressions, etc. (e.g. a value that is expected to be text value is a number, a value that is expected to be positive is negative, etc.)?;
- **completeness** – are there missing observations (more on this in a subsequent section)?;
- **accuracy and precision** – are there measurement and/or data entry errors (e.g. an individual has -2 children, etc., see the target diagrams of Figure 1, linking accuracy to bias and precision to the standard error)?;
- **consistency** – are there conflicting observations (e.g. an individual has no children, but the age of one kid is recorded, etc.)?, and
- **uniformity** – are units used uniformly throughout (e.g. an individual is 6ft tall, whereas another one is 145cm tall)?

Finding an issue with data quality after the analyses are completed is a surefire way of losing the client's trust – check early and often!

Common Sources of Error If the analysts have some control over the data collection and initial processing, regular data validation tests are easier to set-up. When the analysts are dealing with **legacy**, **inherited**, or **combined** datasets, it can be difficult to recognise errors arising (among others) from

- missing data being given a code;
- 'NA'/'blank' entries being given a code;
- data entry errors;
- coding errors;
- measurement errors;
- duplicate entries, and
- heaping (see Figure 2 for an example).

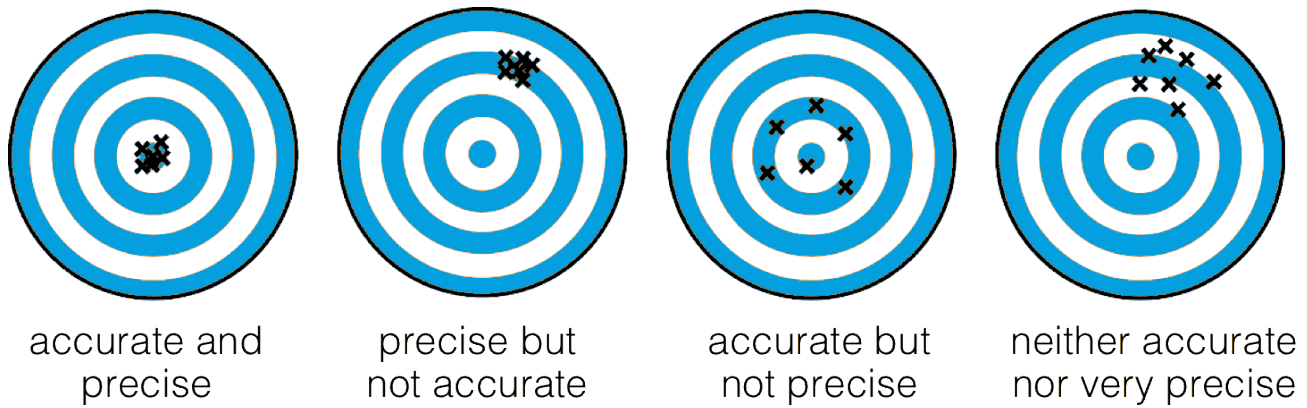


Figure 1: Accuracy as bias, precision as standard error.

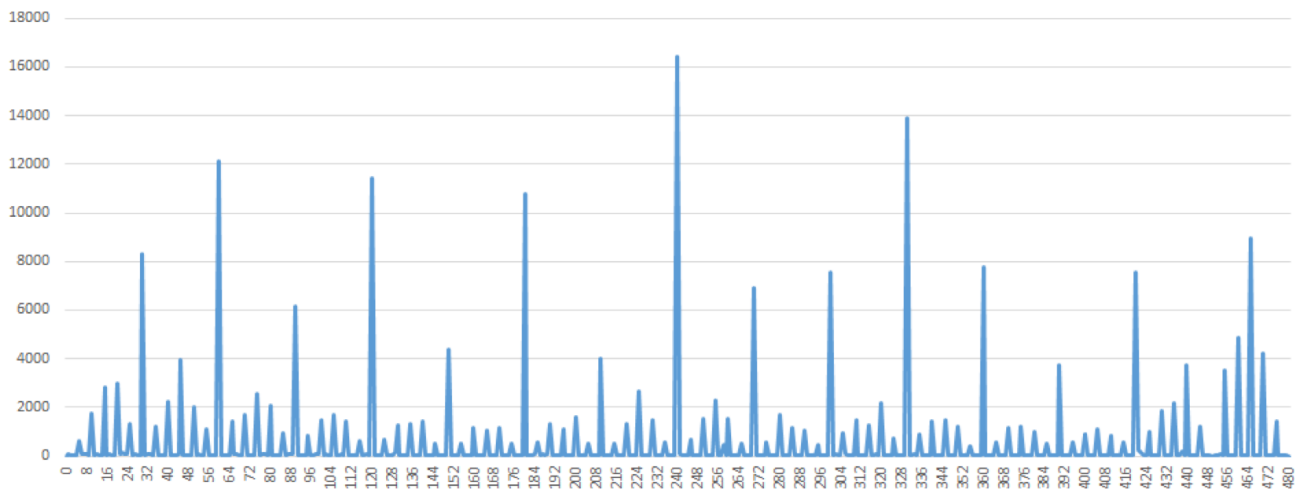


Figure 2: An illustration of heaping: self-reported time spent working in a day (personal file). Note the rounding off at various multiples of 5 minutes.

Detecting Invalid Entries Potentially invalid entries can be detected with the help of a number of methods:

- **univariate descriptive statistics** – count, range, z -score, mean, median, standard deviation, logic check, etc.;
- **multivariate descriptive statistics** – n -way tables and logic check, and
- **data visualisation** – scatterplot, histogram, joint histogram, etc.

We will briefly discuss these methods in Sections ?? and ?. For now, we simply point out that univariate tests do not always tell the whole story.

Consider, for instance, a medical dataset consisting of 38 patients' records, containing, among others, fields for the **sex** and the **pregnancy status** of the patients. A summary of the data of interest is afforded by the frequency counts (1-way tables) shown in Table 2a.

The analyst can quickly notice that some values are missing (in green) and that an entry has been miscoded as 99 (in yellow). Using only these univariate summaries, however, it is impossible to decide what to do with these invalid entries.

Sex	Male	19	Pregnant	Yes	7				
	Female	17		No	27				
	(blank)	2		99	1				
	Total	38		(blank)	3				
			Total		38				

		Pregnant				Total
		Yes	No	99	(blank)	
Sex	Male	1	17	1	0	19
	Female	6	9	0	2	17
	(blank)	0	1	0	1	2
	Total	7	27	1	3	38

(a) 1-way tables

(b) 2-way table

Table 2: Summary data for an (artificial) medical dataset.

The 2-way frequency counts of Table 2b shed some light on the situation, and uncover other potential issues with the data. One of the green entries is actually blank along the two variables; depending on the other information, this entry could be a candidate for **imputation** or outright **deletion** (more on these concepts in the next section). Three other observations are missing a value along exactly one variable, but the information provided by the other variables may be complete enough to warrant imputation. Of course, if more information is available about the patients, the analyst may be able to determine why the values were missing in the first place, although privacy concerns at the collection stage might muddy the waters. The miscoded information on the pregnancy status is linked to a male client, and as such re-coding it as 'No' is likely to be a reasonable decision (although not necessarily the correct one). A similar reasoning process might make the analyst question the validity of the entry shaded in red – the entry might very well be correct, but it is important to at the very least inquire about this data point, as this could lead to an eventual re-framing of the definitions and questions used at the collection stage.

In general, there is no universal or one-size-fits-all approach – a lot depends on the **nature of the data**. As always, domain expertise can help. Remember that a failure to detect invalid entries is **not a guarantee** that there are in fact no invalid entries in the dataset. It is important not to oversell this step to the client. When only a small number of invalid entries are detected, the general recommendation is to treat these values as **missing**, which we discuss presently.

1.2.3 Missing Values

Easier Said Than Done

Obviously, the best way to treat missing data is not to have any.

– T. Orchard, M. Woodbury, *A Missing Information Principle: Theory and Applications*, 1972

Why does it matter that some values may be **missing**? On top of potentially introducing bias into the analysis, most analytical methods can not easily accommodate missing observations. Consequently, when faced with missing observations, analysts have two options: they can either **discard** the missing observation (which is not typically recommended, unless the data is missing completely randomly), or they can **create a replacement value** for the missing observation (the **imputation** strategy has drawbacks since we can never be certain that the replacement value is the true value, but is often the best available option; information in this section is taken partly from [2–5]).

Blank fields come in 4 flavours:

- **nonresponse** – an observation was expected but none was entered;
- **data entry issues** – an observation was recorded but was not entered in the dataset;
- **invalid entries** – an observation was recorded but was considered invalid and has been removed, and
- **expected blanks** – a field has been left blank, but not unexpectedly so.

Too many missing values of the first three types can be indicative of **issues with the data collection process**, while too many missing values of the fourth type can be indicative of **poor questionnaire design** (see Section ?? for a brief discussion on these topics). Either way, missing values cannot simply be **ignored**.

Missing Value Mechanisms The relevance of an imputation method is dependent on the underlying **missing value mechanism**; values may be

- **missing completely at random** (MCAR) – the item absence is independent of its value or of the unit's auxiliary variables (e.g., an electrical surge randomly deletes an observation in the dataset);
- **missing at random** (MAR) – the item absence is not completely random, and could, in theory, be accounted by the unit's complete auxiliary information, if available (e.g., if women are less likely to tell you their age than men for societal reasons, but not because of the age values themselves), and
- **not missing at random** (NMAR) – the reason for nonresponse is related to the item value (e.g., if illicit drug users are less likely to admit to drug use than teetotalers).

The consultant's main challenge in that regard is that the missing mechanism cannot typically be determined with any degree of certainty.

Imputation Methods There are numerous statistical **imputation** methods. They each have their strengths and weaknesses; consequently, consultants should take care to select a method which is appropriate for the situation at hand. They work best under MCAR or MAR, but they all tend to produce **biased estimates**.

- In **list-wise deletion**, all units with at least one missing value are removed from the dataset. This straightforward imputation strategy assumes MCAR, but it can introduce bias if MCAR does not hold, and it leads to a reduction in the sample size and an increase in standard errors.
- In **mean or most frequent imputation**, the missing values are substituted by the average or most frequent value in the unit's subpopulation group (stratum). This approach also assumes MCAR is commonly used, but it can create distortions in the underlying distributions (such as a spike at the mean) and create spurious relationships among variables.
- In **regression or correlation imputation**, the missing values are substituted using a regression on the other variables. This model assumes MAR and trains the regression on units with complete information, in order to take full advantage of the auxiliary information when it is available. However, it artificially reduces data variability and produces over-estimates of correlations.

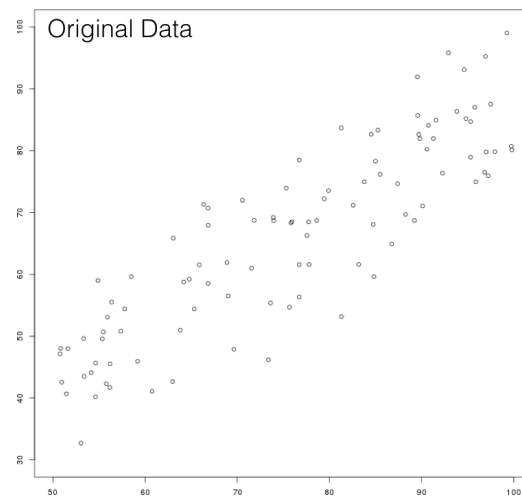


Figure 3: Dr. Vanderwhede’s original *Advanced Retroencabulation* dataset; mid-term grades on the x -axis, final exam grades on the y -axis.

- In **stochastic regression imputation**, the regression estimates are augmented with with random error terms added. Just as in the previous case, the model assumes MAR; an added benefit is that it tends to produce estimates that “look” more realistic than regression imputation, but it comes with an increased risk of type I error (false positives) due to small standard errors.
- **Last observation carried forward (LOCF)** and its cousin **next observation carried backward (NOCB)** are useful for longitudinal data; a missing value can simply be substituted by the previous or next value. LOCF and NOCB can be used when the values do not vary greatly from one observation to the next, and when values are MCAR. Their main drawback is that they may be too “generous”, depending on the nature of study.
- Finally, in **k -nearest-neighbour imputation**, a missing entry in a MAR scenario is substituted by the average (or median, or mode) value from the subgroup of the k most similar complete respondents. This requires a notion of **similarity** between units (which is not always easy to define reasonably). The choice of k is somewhat arbitrary and can affect the imputation, potentially distorting the data structure when it is too large.

What would imputation look like in practice? Consider the following scenario (which is somewhat embarrassingly based on a real event). After marking the final exams of the 100 students who did not drop her course in *Advanced Retroencabulation* at State University, Dr. Helga Vanderwhede plots the final exam grades (y) against the mid-term exam grades (x) as in Figure 3.

She takes a quick look at the data and sees that high final exam grades are **correlated** with high mid-term exam grades, and *vice-versa*. She also sees that there is a fair amount of variability in the data: the noise is not very tight around the line of best fit. Furthermore, she realises that the final exam was harder than the students expected; she suspects that they just did not prepare for the exam seriously (and not that she made the exam too difficult, no matter what her ratings on RateMyProfessor.com suggest), as most of them could not match their mid-term exam performance.

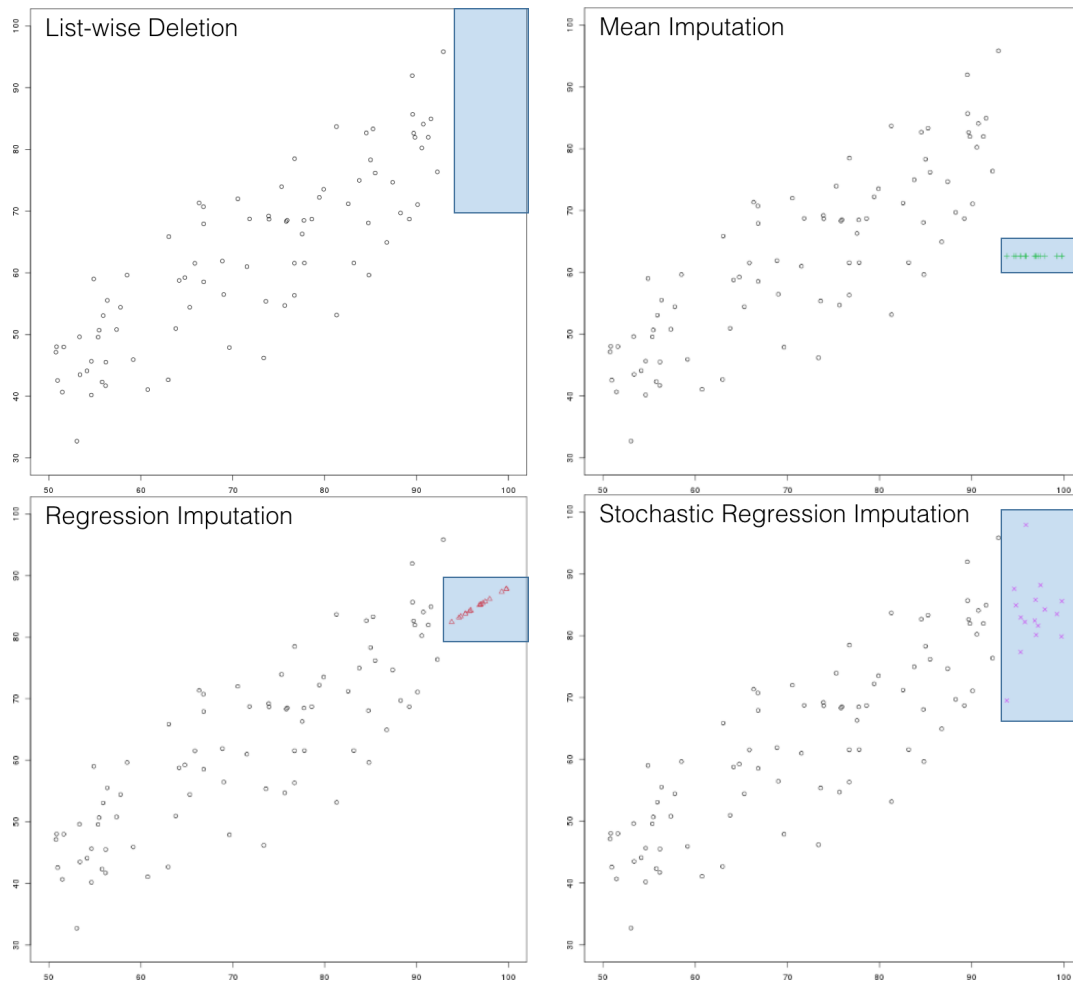


Figure 4: Imputed values for Dr. Vanderwhede's dataset.

As Dr. Vanderwhede comes to term with her disappointment, she decides to take a deeper look at the numbers, at some point sorting the dataset according to the mid-term exam grades. It looks like good old Mary Sue performed better on the final than on the mid-term (where performance was already superlative), scoring the only perfect score. What a fantastic student Mary Sue is! And such a good person – in spite of her superior intellect, she is adored by all of her classmates, thanks to her sunny disposition and willingness to help at all times. If only all students were like Mary Sue... She continues to toy with the spreadsheet, and the phone rings. After a long and exhausting conversation with Dean Bitterman about teaching loads and University's reputation, Dr. Vanderwhede returns to the spreadsheet and notices in horror that she has accidentally deleted the final exam grades of all students with a mid-term grade greater than 92. What is she to do?

A technically-savvy consultant would advise her to either undo her changes or to close the file without saving the changes (or better yet, to re-enter the final grades by comparing with the physical papers), but let's assume for the time being that, in full panic mode, the only solution that comes to her mind is to impute the missing values. She knows that the missing final grades are MAR (and not MCAR since she remembers sorting the data along the x values); she produces the imputations shown in Figure 4. She remembers what the data looked like originally, and concludes that the best imputation method is the stochastic regression model.

But this only applies to this specific example. In general, that might not be the case, however, due to various *No Free Lunch* results (we will discuss this important technical results and its ramifications in Section ??). The principal take-away from this example is that various imputation strategies lead to different outcomes, and perhaps more importantly, that even though the imputed data might “look” like the true data, we have no way to measure its **departure from reality**. Any single imputed value is likely to be completely off. Mathematically, this might not be problematic, as the average departure might be relatively small, but in a business or personal context, this might create gigantic problems – how is Mary Sue likely to feel about Dr. Vanderwhede’s solution in the previous example? How is Dean Bitterman likely to react, if he finds out about the imputation scenario from irrate students? Even though such questions are not quantitative in nature, they will have an effect on actionable solutions.

Multiple Imputation Another drawback of imputation is that it tends to increase the noise in the data, because the imputed data is treated as the *actual* data. In **multiple imputation**, the impact of that noise can be reduced by consolidating the analysis outcome from multiple imputed datasets. Once an imputation strategy has been selected on the basis of the (assumed) missing value mechanism,

1. the imputation process is repeated m times to produce m versions of the dataset;
2. each of these datasets is analyzed, yielding m outcomes, and
3. the m outcomes are pooled into a single result for which the mean, variance, and confidence intervals are known.

On the plus side, multiple imputation is **easy to implement**, **flexible**, as it can be used in a most situations (MCAR, MAR, even NMAR in certain cases), and it accounts for **uncertainty** in the imputed values. However, m may need to be quite **large** when the values are missing in large quantities from many of the dataset’s features, which can substantially slow down the analyses. There may also be additional technical challenges when the output of the analyses is not a single value but some more complicated object.

1.2.4 Anomalous Observations

The Good Doctor’s Take

The most exciting phrase to hear [...], the one that heralds the most discoveries, is not “Eureka!” but “That’s funny...”.

– Isaac Asimov (attributed)

Outlying observations are data points which are **atypical** in comparison to the unit’s remaining features (*within-unit*), or in comparison to the measurements for other units (*between-units*), or as part of a collective subset of observations. Outliers are thus observations which are **dissimilar to other cases** or which contradict **known dependencies** or rules. Outlying observations may be anomalous along any of the individual variables, or in combination (information in this section is taken partly from [11, 18, 19, 25]).

Consider, for instance, an adult male who is 6-foot tall. Such a man would fall in the 86th percentile in Canada [26], which, while on the tall side is not unusual; but in Bolivia, he would fall in the 99.9th percentile [26], which would mark him as extremely tall and quite dissimilar to the rest of the population. (Why is there such a large discrepancy in the two populations?)

The most common mistake that analysts make when dealing with outlying observations is to remove them from the dataset without careful studying whether there are good reasons to retain them.

Influential data points, meanwhile, are observations whose absence leads to **markedly different** analysis results. When influential observations are identified, remedial measures (such as data transformation strategies) may need to be applied to minimize any undue effect. Note that outliers may be influential, and influential data points may be outliers, but the conditions are neither necessary nor sufficient.

Detecting Anomalies Anomalies are by definition **infrequent**, and typically shrouded in **uncertainty** due to small (relative) sample sizes, which makes differentiating them from **noise** or **data entry errors** difficult. It could also be the case that the boundaries between a normal unit and a deviating unit is **fuzzy**; with the advent of e-shops, a purchase made at 3am local time does not necessarily ring alarm bells anymore. It is hard enough as it is to try to identify “honest” anomalies; when anomalies are associated with **malicious activities**, they are typically **disguised** to look like a normal observation, which muddies the picture even more.

Numerous methods exist to identify anomalous observations; **none of them are foolproof** and judgement must be used. Methods that employ graphical aids (such as box-plots, scatterplots, scatterplot matrices, and 2D tours, for outliers, say ??) are particularly easy to implement and interpret, especially in a low-dimensional setting. Analytical methods also exist (using Cooke’s or Mahalanobis’ distances, say), but in general some additional level of analysis must be performed, especially when trying to identify influential points (*cf.* **leverage**).

We do not recommend the general use of **automated detection/removal** – as tempting as this might get when the dataset is large. This stems partly from the fact that once the “anomalous” observations have been removed from the data set, previously “regular” observations can become anomalous in turn in the smaller dataset; it is not clear when the runaway train will stop.

In the early stages, **simple data analyses** (such as descriptive statistics, 1- and 2-way tables, and traditional visualisations) may be performed to help identify anomalous observations, or to obtain insights about the data, which could eventually lead to modifications of the analysis plan.

Outlier Tests So how do we *actually* detect outliers? Most methods come in one of two flavours: **supervised** and **unsupervised** (we will discuss those concepts – and others – in Section ??).

Supervised methods use a historical record of **labeled** (that is to say, previously identified) anomalous observations to build a **predictive classification or regression model** which estimates the

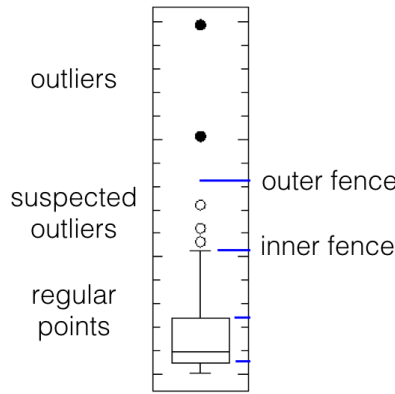


Figure 5: Tukey's boxplot test; suspected outliers are marked by white disks, outliers by black disks.

probability that a unit is anomalous; domain expertise is required to tag the data. Since anomalies are typically **infrequent**, these models often have to accommodate the rare occurrence problem (more on this in Section ??). Unsupervised methods, on the other hand, use no previous information or data; the following traditional methods and tests of outlier detection fall into this category (note that **normality** of the underlying data is an assumption for most tests; how robust these tests are against departures from this assumption depends on the situation).

Perhaps the most commonly known such test is **Tukey's boxplot test**: for normally distributed data, regular observations typically lie between the **inner fences**

$$Q_1 - 1.5(Q_3 - Q_1) \quad \text{and} \quad Q_3 + 1.5(Q_3 - Q_1).$$

Suspected outliers lie between the inner fences and the **outer fences**

$$Q_1 - 1.5(Q_3 - Q_1) \quad \text{and} \quad Q_3 + 1.5(Q_3 - Q_1).$$

Points beyond the outer fences are identified as **outliers** (Q_1 and Q_3 represent the data's 1st and 3rd quartile, respectively; see Figure 5). The **Grubbs test** is another univariate test, which takes into consideration the number of observations in the dataset. Let x_i be the value of feature X for the i^{th} unit, $1 \leq i \leq N$, (\bar{x}, s_x) be the mean and standard deviation of feature X , α be the significance level, and $T(\alpha, N)$ be the critical value of the Student t -distribution at significance $\alpha/2N$. Then, the i^{th} unit is an **outlier along feature X** if

$$|x_i - \bar{x}| \geq \frac{s_x(N-1)}{\sqrt{N}} \sqrt{\frac{T^2(\alpha, N)}{N-2+T^2(\alpha, N)}}.$$

Other common tests include:

- the **Dixon Q test**, which is used in the experimental sciences to find outliers in (extremely) small datasets – it is of dubious validity;
- the **Mahalanobis distance**, which is linked to the leverage of an observation (a measure of influence), can also be used to find multi-dimensional outliers, when all relationships are linear (or nearly linear);

- the **Tietjen-Moore** test, which is used to find a specific number of outliers;
- the **generalized extreme studentized deviate**, if the number of outliers is unknown;
- the **chi-square** test, when outliers affect the goodness-of-fit, as well as
- DBSCAN and other unsupervised outlier detection methods.

What do we do when the data is not normally distributed? We will discuss one possible approach after we present three more examples illustrating the basics of visual outlier and anomaly detection.

On a specific day, the height of several plants in a nursery are measured. The records also show each plant's age (the number of days since the seed has been planted). Histograms of the data are shown in Figures 6a and 6b. Very little can be said about the data at that stage: the age of the plants (controlled by the nursery staff) seems to be somewhat haphazard, as does the response variable (height). A scatter plot of the data (see Figure 6c), however, reveals that growth is strongly correlated with age during the early days of a plant's life for the observations in the dataset; most points clutter around a linear trend. But one point (in yellow) is easily identified as an **outlier**. There are at least two possibilities: either that measurement was botched or mis-entered in the database (representing an invalid entry), or that one specimen has experienced unusual growth (outlier). Either way, the analyst has to investigate further.

A government department has 11 service points in a jurisdiction. Service statistics are recorded: in particular, the monthly average arrival rates per teller and monthly average service rates per teller for each service point are available. A scatter plot of the service rate per teller (y axis) against the arrival rate per teller (x axis), with linear regression trend, is shown in Figure 7a. The trend is seen to inch upwards with increasing x values. A similar graph, but with the left-most point removed from consideration, is shown in Figure 7b. The trend still slopes upward, but the fit is significantly improved suggesting that the removed observation is unduly **influential** – a better understanding of the relationship between arrivals and services is afforded if it is set aside. Any attempt to fit that data point into the model must take that information into consideration. Note, however, that influential observations depend on the analysis that is ultimately being conducted – a point may be influential for one analysis, but not for another.

Measurements of the length of the appendage of a certain species of insect have been made on 71 individuals. Descriptive statistics have been computed; the results are shown in Figure 3a. Analysts who are well-versed in statistical methods would recognise the tell-tale signs that the distribution of appendage lengths is likely to be asymmetrical (since the skewness is non-negligible) and to have a “fat” tail (due to the kurtosis being commensurate with the mean and the standard deviation, the range being so much larger than the interquartile range, and the maximum value being so much larger than the third quartile). The mode, minimum, and first quartile values belong to individuals without appendages, so there would appear to be two sub-groups in the population (perhaps split along the lines of juveniles/adults, or males/females). The maximum value has already been seen to be quite large compared to the rest of the observations, which at

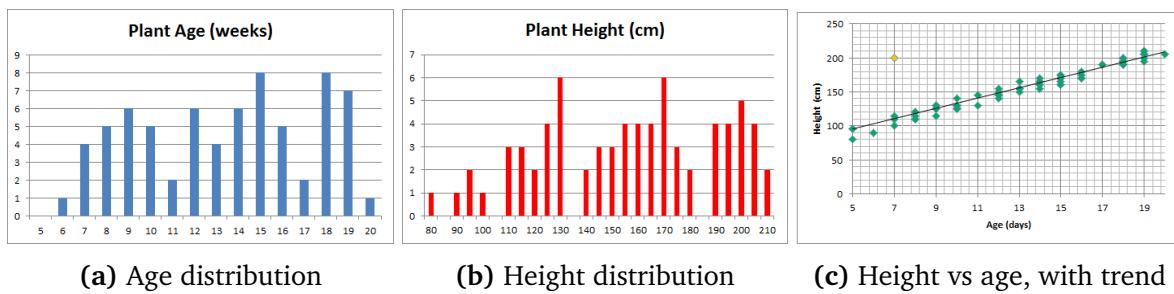


Figure 6: Summary visualisations for an (artificial) plant dataset.

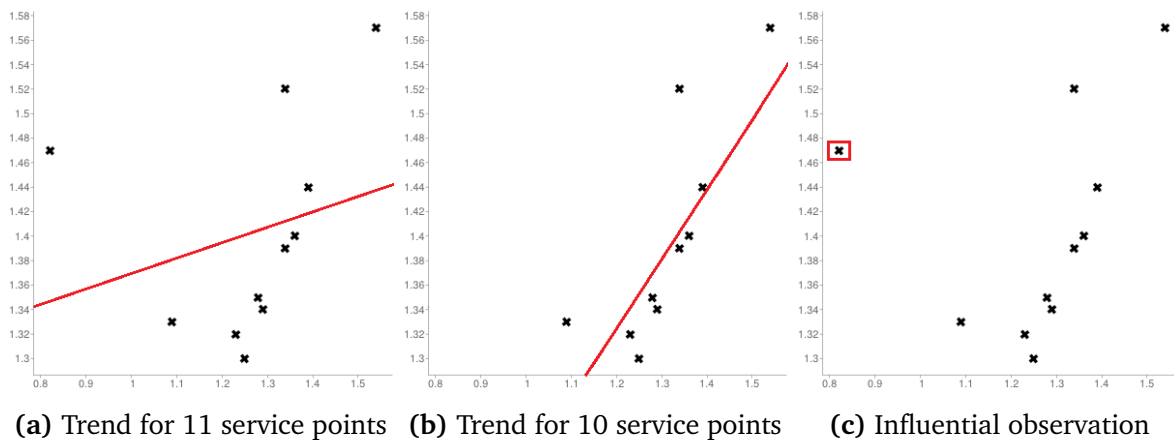


Figure 7: Visualisations for an (artificial) service point dataset.

first suggests that it might belong to an **outlier** or **invalid entry**. The histogram of the measurements, however, shows that there are 3 individuals with very long appendages (see Figure 3b): it now becomes plausible for these anomalous entries to belong to individuals from a different species altogether who were **erroneously added** to the dataset. This does not, of course, constitute a proof of such an error, but it raises the possibility, which is often the best that a consultant can do for a client.

1.2.5 Data Transformation

It's Also True of Data

History is the transformation of tumultuous conquerors into silent footnotes.

– Paul Eldridge, American educator

This **crucial** last step is often neglected or omitted altogether when consultants embark on complex data analysis projects. Various transformation methods are available, depending on the analysts' needs and data types, including:

- **standardization** and **unit conversion**, which put the dataset's variables on an equal footing – a requirement for basic comparison tasks and more complicated problems of clustering and similarity matching;

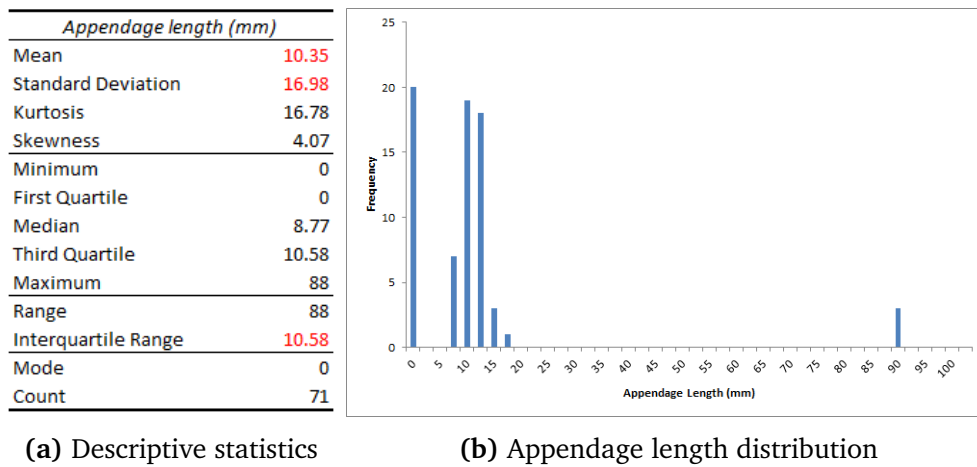


Table 3: Summary and visualisation for an (artificial) appendage length dataset.

- **normalization**, which attempts to force a variable into a normal distribution – an assumption which must be met in order to use a number of traditional analysis methods, such as regression analysis or ANOVA, and
- **smoothing methods**, which help remove unwanted noise from the data, but at a price – perhaps removing natural variance in the data.

Another type of data transformation is pre-occupied with the concept of **dimensionality reduction**. There are many advantages to working with low-dimensional data:

- **visualisation methods** of all kinds are available to extract and present insights out of such data (see Section ??);
- high-dimensional datasets are subject to the so-called **curse of dimensionality**, which asserts (among other things) that multi-dimensional spaces are vast, and when the number of features in a model increases, the number of observations required to maintain predictive power also increases, but at a **substantially higher rate** (see Figure 8);
- another consequence of the curse is that in high-dimension sets, all observations are roughly **dissimilar** to one another – observations tend to be nearer the dataset’s boundaries than they are to one another.

Dimension reduction techniques such as the ubiquitous **principal component analysis**, **independent component analysis**, and **factor analysis** (for numerical data), or **multiple correspondence analysis** (for categorical data) project multi-dimensional datasets onto low-dimensional but high-information spaces (the so-called **Manifold Hypothesis**). Some information is necessarily lost in the process, but in many instances the drain can be kept under control and the gains made by working with smaller datasets can offset the losses of completeness. We will touch on this topic briefly in Section ??.

Common Transformations Models often require that certain data assumptions be met. For instance, ordinary least square regression assumes:

- that the response variable is a **linear combination** of the predictors;

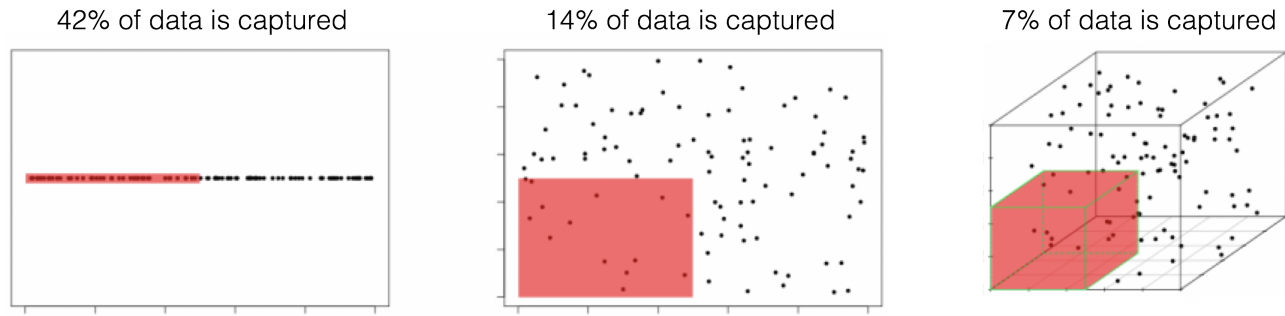


Figure 8: Illustration of the curse of dimensionality; $N = 100$ observations are uniformly distributed on the unit hypercube $[0, 1]^d$, $d = 1, 2, 3$. The red regions represent the smaller hypercubes $[0, 0.5]^d$, $d = 1, 2, 3$. The percentage of captured datapoints is seen to decrease with an increase in d [28].

- **constant** error variance;
- **uncorrelated residuals**, which may or may not be statistically independent;
- etc.

In reality, it is rare that raw data meets the requirements, but that does not necessarily mean that we need to abandon the model – an **invertible** sequence of data transformations may produce a derived data set which *does* meet the requirements, allowing the consultant to draw conclusions about the original data.

In the regression context, invertibility is guaranteed by **monotonic** transformations: identity, logarithmic, square root, inverse (all members of the power transformations), exponential, etc. (illustrations are provided in Figure 9). There are rules of thumb and best practices to transform data, but consultants should not discount the importance of explore the data visually before making a choice.

Transformations on the predictors X may be used to achieve the **linearity assumption**, but they usually come at a price – correlations are not preserved by such transformations, for instance. Transformations on the target Y can help with **non-normality** of residuals and **non-constant variance** of error terms. Note that transformations can be applied **both** to the target variable or the predictors: as an example, if the linear relationship between two variables X and Y is expressed as $Y = a + bX$, then a unit increase in X is associated with an average of b units in Y . But a better fit might be afforded by either of

$$\log Y = a + bX, \quad Y = a + b \log X, \quad \text{or} \quad \log Y = a + b \log X,$$

for which:

- a unit increase in X is associated with an average $b\%$ increase in Y ;
- a 1% increase in X is associated with an average $0.01b$ unit increase in Y , and
- a 1% increase in X is associated with a $b\%$ increase in Y , respectively.

Box-Cox Transformation The choice of transformation is often as much of an art as it is a science. There is a common framework, however, that provides the optimal transformation, in a sense. Consider the task of predicting the target Y with the help of the predictors X_j , $j = 1, \dots, p$.

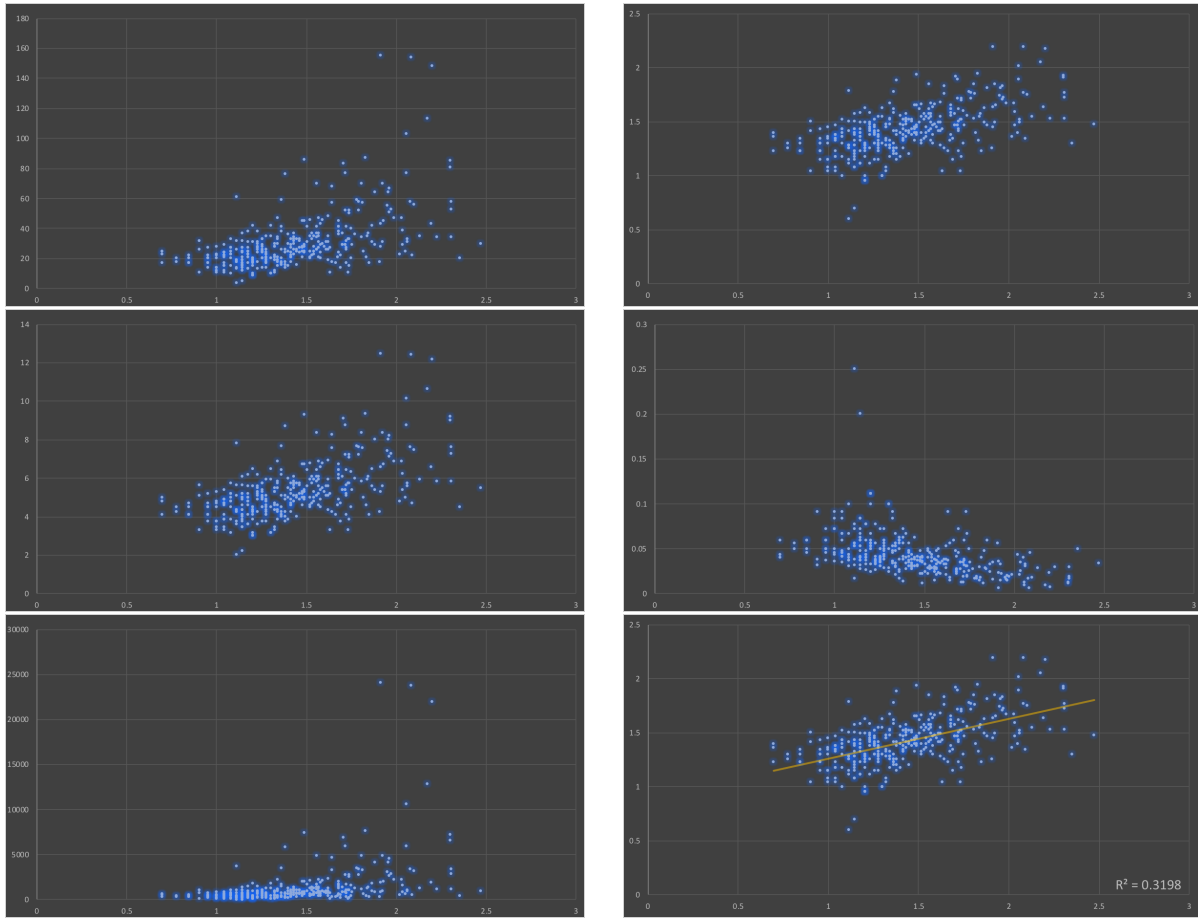


Figure 9: Examples of data transformations, for a subset of the BUPA liver dataset [27]. From left to right, top to bottom: original data, $Y' = \log Y$, $Y' = \sqrt{Y}$, $Y' = \frac{1}{Y}$, $Y' = Y^2$, and Box-Cox best choice ($\approx \log$).

The usual model takes the form

$$y_i = \sum_{j=1}^p \beta_j X_{x,i} + \varepsilon_i, \quad i = 1, \dots, n.$$

Perhaps the residuals are skewed, or their variance is not constant, or the trend itself does not appear to be linear. A power transformation might be preferable, but if so, which one?

The **Box-Cox transformation** $y_i \mapsto y'_i(\lambda)$, $y_i > 0$ is defined by

$$y'_i(\lambda) = \begin{cases} (y_1 \dots y_n)^{1/n} \ln y_i, & \text{if } \lambda = 0 \\ \frac{y_i^\lambda - 1}{\lambda} (y_1 \dots y_n)^{\frac{1-\lambda}{n}}, & \text{if } \lambda \neq 0 \end{cases} ;$$

variants allow for the inclusion of a shift parameter $\alpha > 0$, which extends the transformation to $y_i > -\alpha$. The **suggested** choice of λ is the value that maximises the log-likelihood

$$\mathcal{L} = -\frac{n}{2} \log \left(\frac{2\pi\hat{\sigma}^2}{(y_1 \dots y_n)^{2(\lambda-1)/n}} + 1 \right).$$

There might be theoretical rationales which favour a particular choice of λ – these are not to be ignored. It is also important to produce a residual analysis, as the best Box-Cox choice does not necessarily meet all the least squares assumptions. Finally, it is important to remember that the resulting parameters have the least squares property **only with respect to the transformed data points**.

Scaling Numeric variables may have different scales (weights and heights, for instance). Since the variance of a large-range variable is typically greater than that of a small-range variable, leaving the data **unscaled** may introduce biases, especially when using unsupervised methods. It could also be the case that it is the relative positions/rankings which is of importance, in which case it could become important to look at relative distances between levels:

- **standardisation** creates a variable with mean 0 and standard deviations 1:

$$Y_i = \frac{X_i - \bar{X}}{s_X};$$

- **normalization** creates a new variable in the range $[0,1]$:

$$Y_i = \frac{X_i - \min X}{\max X - \min X}.$$

These are not the only options. Different schemes can lead to different outputs.

Discretising To reduce computational complexity, a numeric variable may need to be replaced with an **ordinal** variable (*height* values could be replaced by the qualitative “*short*”, “*average*”, and “*tall*”, for instance. Of course, what these terms represent depend on the context: Canadian short and Bolivian tall may be fairly commensurate. It is far from obvious how to determine the bins’ limits – **domain expertise** can help, but it could introduce unconscious bias to the analyses. In the absence of such expertise, limits can be set so that either

- the bins each contain the same **number of observations**;
- the bins each have the same **width**, or
- the performance of some modeling tool is maximised.

Again, various choices may lead to different outputs.

Creating Variables Finally, it is possible that new variables may need to be introduced (in contrast with dimensionality reduction). These new variables may arise

- as **functional relationships** of some subset of available features (introducing powers of a feature, or principal components, say);
- because modeling tool may require **independence of observations** or **independence of features** (in order to remove multicollinearity, for instance), or
- to simplify the analysis by looking at **aggregated summaries** (often used in text analysis).

There is no limit to the number of new variables that can be added to a dataset – but consultants should strive for **relevant additions**.

1.2.6 Case Study: Imputation of Blood Alcohol Content Levels

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. 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.

In large jurisdictions, distances to the coroner's office may take a while to traverse. A large volume of such fatalities may also slow down the process. For these (and other) reasons, it can take up to a year for the missing **blood alcohol concentration** (BAC) levels to make their way to various interested parties (policy makers, analysts, etc.). 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 prediction is made on the basis of a number of auxiliary variables, such as the age of the driver. 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.

In 2007, *Ministry of Transportation of Ontario* (MTO) faced such a situation: using a small number of features (many of which have missing values themselves), is it possible to

1. predict whether alcohol was involved, and if so,
2. predict the BAC level?

The problem is easily stated, but the existence of an actionable solution is not clear. There may simply be no link between the available features and the BAC level. For instance, how strong can the connection between the deceased's handedness and their BAC level (assuming we even have access to that information).

Another issue, which we have broached in Section ??, is the question of the data's representativeness: is it possible that whatever link might have existed in 2007 is simply not going to be present in the future, perhaps as a result of implemented policies? If that is the case, how useful would a general model prove to be?

The paper that describes the two-stage multiple imputation model used by *Transport Canada* to solve the MTO's problem is presented after the references – note how the flow is broken when the tables are not labeled.

References

- [1] Chapman, A. [2005], Principles and Methods of Data Cleaning - Primary Species and Species-Occurrence Data, Report for the Global Biodiversity Information Facility, Copenhagen.
- [2] van Buuren, S. [2012], Flexible Imputation of Missing Data, CRC Press, Boca Raton.
- [3] Hagiwara, S. [2012], Nonresponse Error in Survey Sampling - Comparison of Different Imputation Methods, Honours Thesis, Carleton University, Ottawa.

- [4] Raghunathan, T., Lepkowski, J., Van Hoewyk, J. and Solenberger, P. [2001], A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models, *Survey Methodology*, v.27, n.1, pp.85-95, Statistics Canada, Catalogue no. 12-001.
- [5] Rubin, D.B. [1987], *Multiple Imputation for Nonresponse in Surveys*, Wiley, New York.
- [6] Kutner, M., Nachtsheim, C., Neter, J. and Li, W. [2004], *Applied Linear Statistical Models*, 5th ed., McGraw-Hill/Irwin, New York.
- [7] Green, S. and Salkind, N. [2011], *Using SPSS for Windows and Macintosh - Analyzing and Understanding Data*, 6th ed., Prentice Hall, Upper Saddle River.
- [8] Wikipedia entry for Data Cleansing
- [9] Wikipedia entry for Imputation
- [10] Wikipedia entry for Outliers
- [11] Torgo, L. [2017], *Data Mining with R (2nd edition)*, CRC Press.
- [12] McCallum, Q.E. [2013], *Bad Data Handbook*, O'Reilly.
- [13] Kazil, J., Jarmul, K. [2016], *Data Wrangling with Python*, O'Reilly
- [14] de Jonge, E., van der Loo, M. [2013], *An Introduction to Data Cleaning with R*, Statistics Netherlands.
- [15] Pyle, D. [1999], *Data Preparation for Data Mining*, Morgan Kaufmann Publishers.
- [16] Weiss, S.M., Indurkha, I. [1999], *Predictive Data Mining: A Practical Guide*, Morgan Kaufmann Publishers.
- [17] Buttrey, S.E. [2017], *A Data Scientist's Guide to Acquiring, Cleaning, and Managing Data in R*, Wiley.
- [18] Aggarwal, C.C. [2013], *Outlier Analysis*, Springer.
- [19] Chandola, V., Banerjee, A., Kumar, V. [2007], *Outlier detection: a survey*, Technical Report TR 07-017, Department of Computer Science and Engineering, University of Minnesota.
- [20] Hodge, V., Austin, J. [2004], A survey of outlier detection methodologies, *Artif.Intell.Rev.*, 22(2):85-126.
- [21] Feng, L., Nowak, G., Welsh, A.H., O'Neill, T. [2014], *imputeR: a general imputation framework in R*.
- [22] Steiger, J.H. , *Transformations to Linearity*, lecture notes.
- [23] Wood, F., *Remedial Measures Wrap-Up and Transformations*, lecture notes.
- [24] Dougherty, J., Kohavi, R., Sahami, M. [1995], Supervised and unsupervised discretization of continuous features, in *Machine Learning: Proceedings of the Twelfth International Conference*, Frieditis, A., Russell, S. (eds), Morgan Kaufmann Publishers.
- [25] Orchard, T., Woodbury, M. [1972], *A Missing Information Principle: Theory and Applications*, Berkeley Symposium on Mathematical Statistics and Probability, University of California Press.
- [26] Height Percentile Calculator, by Age and Country, <https://tall.life/height-percentile-calculator-age-country/>
- [27] Dua, D., Karra Taniskidou, E. [2017], *Liver Disorders dataset*, UCI Machine Learning Repository.
- [28] <https://simplystatistics.org/2014/10/24/an-interactive-visualization-to-teach-about-the-curse-of-dimensionality/>

An Imputation Algorithm of Blood Alcohol Content Levels for Drivers and Pedestrians in Fatal Collisions

Patrick Boily¹


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 report, 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

¹ Evaluation & Data Systems Division, Road Safety and Motor Vehicles, Transport Canada
Email: patrick.boily@tc.gc.ca [inactive]



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 article, 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.

Contents	
1	Statistical Imputation
2	Data Preparation
3	Methodology
4	Artificial Example
5	Results for Ontario (2007)
6	Conclusion

1. 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 $\{X, \tilde{Y}_1, \dots, \tilde{Y}_{m-1}\}$. The main drawback of this method is that some information might be "hiding" in $\{Y_2, Y_3, \dots, Y_m\}$ 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.

2. 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.

3. 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 $W_{\alpha_1}^{\text{imp}}$ and $W_{\alpha_1}^{\text{train}}$, 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 $W_{\alpha_1}^{\text{imp}}$ using $W_{\alpha_1}^{\text{train}}$ 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 $W_{\alpha_2}^{\text{imp}}$ and $W_{\alpha_2}^{\text{train}}$ as above. Impute the missing values of Y_{α_2} in $W_{\alpha_2}^{\text{imp}}$ using $W_{\alpha_2}^{\text{train}}$ 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 $W_{\alpha_{m_1}}^{\text{imp}}$ using $W_{\alpha_{m_1}}^{\text{train}}$ 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 $W_{\alpha_1, \beta_1}^{\text{imp}}$ and $W_{\alpha_1, \beta_1}^{\text{train}}$, 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 $W_{\alpha_1, \beta_1}^{\text{imp}}$ using $W_{\alpha_1, \beta_1}^{\text{train}}$ 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 $W_{\alpha_2, \beta_2}^{\text{imp}}$ and $W_{\alpha_2, \beta_2}^{\text{train}}$ as above. Impute the missing values for Y_{α_2, β_2} in $W_{\alpha_2, \beta_2}^{\text{imp}}$ using $W_{\alpha_2, \beta_2}^{\text{train}}$ 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 $W_{\alpha_{m_2}, \beta_{m_2}}^{\text{imp}}$ using $W_{\alpha_{m_2}, \beta_{m_2}}^{\text{train}}$ 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. $\text{BAC} > 0$), 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

$$\bar{Z}_1^i = \frac{1}{n} \sum_{j=1}^k Z_1^{j,i} \quad \text{and} \quad \bar{Z}_2^i = \frac{\sum_{j=1}^k Z_1^{j,i} Z_2^{j,i}}{n \bar{Z}_1^i}.$$

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

$$Z_1^i = \begin{cases} 1 & \text{if } \bar{Z}_1^i > a \\ 0 & \text{else} \end{cases} \quad \text{and} \quad Z_2^i = \begin{cases} \bar{Z}_2^i & \text{if } \bar{Z}_1^i > a \\ 0 & \text{else} \end{cases}$$

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 $\text{BAC} > 0$ but actual $\text{BAC} = 0$) is the same as that of a false negative (imputed $\text{BAC} = 0$ but actual $\text{BAC} > 0$), then $a = 0.5$ is a good choice.

4. 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	MIS
1	2	1	3	2	1	0	0	0
2	2	1	2	1	1	0	0	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 $X_{i,j}$, $i = 1, \dots, 14$, $j = 1, \dots, 3$, 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	MIS
1	1	2	1	3	2	1	0	0	0
1	2	2	1	3	2	1	0	0	0
1	3	2	1	3	2	1	0	0	0
2	1	2	1	2	1	1	.	.	0
2	2	2	1	2	1	1	.	.	0
2	3	2	1	2	1	1	.	.	0
3	1	2	1	2	2	2	0	0	0
3	2	2	1	2	2	2	0	0	0
3	3	2	1	2	2	2	0	0	0
4	1	1	1	.	1	2	.	.	1
4	2	1	1	.	1	2	.	.	1
4	3	1	1	.	1	2	.	.	1
5	1	2	1	4	2	1	.	.	0
5	2	2	1	4	2	1	.	.	0
5	3	2	1	4	2	1	.	.	0
6	1	3	2	3	2	2	1	91	0
6	2	3	2	3	2	2	1	91	0
6	3	3	2	3	2	2	1	91	0
7	1	1	.	1	1	2	1	156	1
7	2	1	.	1	1	2	1	156	1
7	3	1	.	1	1	2	1	156	1
8	1	2	2	1	1	1	1	23	0
8	2	2	2	1	1	1	1	23	0
8	3	2	2	1	1	1	1	23	0
9	1	2	1	2	2	.	.	.	1
9	2	2	1	2	2	.	.	.	1
9	3	2	1	2	2	.	.	.	1
10	1	2	1	3	2	.	0	0	1
10	2	2	1	3	2	.	0	0	1
10	3	2	1	3	2	.	0	0	1
11	1	2	1	3	2	2	.	.	0
11	2	2	1	3	2	2	.	.	0
11	3	2	1	3	2	2	.	.	0
12	1	1	1	5	.	1	.	.	1
12	2	1	1	5	.	1	.	.	1
12	3	1	1	5	.	1	.	.	1
13	1	1	2	4	.	.	0	0	2
13	2	1	2	4	.	.	0	0	2
13	3	1	2	4	.	.	0	0	2
14	1	2	2	4	1	1	1	118	0
14	2	2	2	4	1	1	1	118	0
14	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	1	2	1	3	2	1	0	0	0
3	1	2	1	3	2	1	0	0	0
4	1	1	1	1	1	2	0	0	1
5	1	2	1	4	2	1	0	0	0
6	1	3	2	3	2	2	1	91	0
7	1	2	1	1	1	2	1	156	1
8	1	2	2	1	1	1	1	23	0
9	1	2	1	2	2	0	0	0	1
10	1	2	1	3	2	0	0	0	1
11	1	2	1	3	2	2	0	0	0
12	1	1	1	5	1	1	0	0	1
13	1	1	2	4	0	0	0	0	2
14	1	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	1	2	1	3	2	1	0	0	0
3	1	2	1	3	2	1	0	0	0
4	1	1	1	1	1	2	0	0	1
5	1	2	1	4	2	1	0	0	0
6	1	3	2	3	2	2	1	91	0
7	1	2	1	1	1	2	1	156	0
8	1	2	2	1	1	1	1	23	0
9	1	2	1	2	2	0	0	0	1
10	1	2	1	3	2	0	0	0	1
11	1	2	1	3	2	2	0	0	0
12	1	1	1	5	1	1	0	0	1
13	1	1	2	4	0	0	0	0	2
14	1	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	1	2	1	3	2	1	0	0	0
3	1	2	1	3	2	1	0	0	0
4	1	1	1	1	1	2	0	0	1
5	1	2	1	4	2	1	0	0	0
6	1	3	2	3	2	2	1	91	0
7	1	2	1	1	1	2	1	156	0
8	1	2	2	1	1	1	1	23	0
9	1	2	1	2	2	0	0	0	1
10	1	2	1	3	2	0	0	0	1
11	1	2	1	3	2	2	0	0	0
12	1	1	1	5	1	1	0	0	1
13	1	1	2	4	0	0	0	0	2
14	1	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	1	2	1	3	2	1	0	0	0
3	1	2	1	3	2	1	0	0	0
4	1	1	1	1	1	2	0	0	1
5	1	2	1	4	2	1	0	0	0
6	1	3	2	3	2	2	1	91	0
7	1	2	1	1	1	2	1	156	0
8	1	2	2	1	1	1	1	23	0
9	1	2	1	2	2	0	0	0	1
10	1	2	1	3	2	0	0	0	1
11	1	2	1	3	2	2	0	0	0
12	1	1	1	5	1	1	0	0	1
13	1	1	2	4	0	0	0	0	2
14	1	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	1	2	1	3	2	1	0	0	0
3	1	2	1	3	2	1	0	0	0
4	1	1	1	1	1	2	0	0	1
5	1	2	1	4	2	1	0	0	0
6	1	3	2	3	2	2	1	91	0
7	1	2	1	1	1	2	1	156	0
8	1	2	2	1	1	1	1	23	0
9	1	2	1	2	2	0	0	0	1
10	1	2	1	3	2	0	0	0	1
11	1	2	1	3	2	2	0	0	0
12	1	1	1	5	1	1	0	0	1
13	1	1	2	4	0	0	0	0	2
14	1	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	1	2	1	3	2	1	0	0	0
3	1	2	1	3	2	1	0	0	0
4	1	1	1	1	1	2	0	0	1
5	1	2	1	4	2	1	0	0	0
6	1	3	2	3	2	2	1	91	0
7	1	2	1	1	1	2	1	156	0
8	1	2	2	1	1	1	1	23	0
9	1	2	1	2	2	0	0	0	1
10	1	2	1	3	2	0	0	0	1
11	1	2	1	3	2	2	0	0	0
12	1	1	1	5	1	1	0	0	1
13	1	1	2	4	0	0	0	0	2
14	1	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.

Step 1–4

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

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

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

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

REC	REP	VEIS	SEX	AGE	SAFE	CF	VIA	VIS	VIC	X/A	X/C	X/S	A/C	A/S	S/C	TEST	BAC	MIS
1	1	2	2	2	2	2	1	1	9	5	2	2	1	8	5	0	0	0
	2	2	1	1	3	2	1	1	9	5	3	2	1	8	5	0	0	0
	2	2	1	1	3	2	1	1	9	5	3	2	1	8	5	0	0	0
	2	2	1	1	3	2	1	1	9	5	3	2	1	8	5	0	0	0
2	1	2	2	2	2	2	3	3	8	4	3	2	1	1	4	3	3	0
	2	2	2	2	2	2	3	3	8	4	3	2	1	1	4	3	3	0
	2	2	2	2	2	2	3	3	8	4	3	2	1	1	4	3	3	0
	2	2	2	2	2	2	3	3	8	4	3	2	1	1	4	3	3	0
3	1	2	1	2	2	2	2	3	8	5	4	2	2	2	5	4	4	0
	2	2	1	2	2	2	2	3	8	5	4	2	2	2	5	4	4	0
	2	2	1	2	2	2	2	3	8	5	4	2	2	2	5	4	4	0
	2	2	1	2	2	2	2	3	8	5	4	2	2	2	5	4	4	0
4	1	1	1	1	2	1	2	1	1	2	1	1	1	2	2	2	2	0
	2	1	1	1	1	2	1	1	4	1	2	4	1	2	10	8	6	0
	2	1	1	1	1	2	1	1	4	1	2	4	1	2	10	8	6	0
	2	1	1	1	1	2	1	1	4	1	2	4	1	2	10	8	6	0
5	1	2	1	4	2	1	3	10	5	3	4	2	1	11	7	3	3	0
	2	2	1	4	2	1	3	10	5	3	4	2	1	11	7	3	3	0
	2	2	1	4	2	1	3	10	5	3	4	2	1	11	7	3	3	0
	2	2	1	4	2	1	3	10	5	3	4	2	1	11	7	3	3	0
6	1	3	2	3	2	2	6	15	8	6	9	5	4	8	6	4	1	91
	2	3	2	3	2	2	6	15	8	6	9	5	4	8	6	4	1	91
	2	3	2	3	2	2	6	15	8	6	9	5	4	8	6	4	1	91
	2	3	2	3	2	2	6	15	8	6	9	5	4	8	6	4	1	91
7	1	1	2	1	2	2	2	1	1	2	7	4	4	1	2	2	1	156
	2	1	1	1	1	2	2	1	1	2	1	2	1	2	2	2	1	156
	2	1	1	1	1	2	2	1	1	2	1	2	1	2	2	2	1	156
	2	1	1	1	1	2	2	1	1	2	1	2	1	2	2	2	1	156
8	1	2	2	1	1	1	4	7	4	3	7	4	3	1	1	1	1	23
	2	2	2	1	1	1	4	7	4	3	7	4	3	1	1	1	1	23
	2	2	2	1	1	1	4	7	4	3	7	4	3	1	1	1	1	23
	2	2	2	1	1	1	4	7	4	3	7	4	3	1	1	1	1	23
9	1	2	1	2	2	2	3	8	5	3	2	2	1	5	3	4	3	0
	2	2	1	2	2	2	3	8	5	3	2	2	2	5	3	4	3	0
	2	2	1	2	2	2	3	8	5	3	2	2	2	5	3	4	3	0
	2	2	1	2	2	2	3	8	5	3	2	2	2	5	3	4	3	0
10	1	2	1	3	2	2	3	9	5	3	3	2	1	8	5	3	0	0
	2	2	1	3	2	2	3	9	5	3	3	2	1	8	5	3	0	0
	2	2	1	3	2	2	3	9	5	3	3	2	1	8	5	3	0	0
	2	2	1	3	2	2	3	9	5	3	3	2	1	8	5	3	0	0
11	1	2	1	3	2	2	3	9	5	4	3	2	2	8	6	4	1	0
	2	2	1	3	2	2	3	9	5	4	3	2	2	8	6	4	1	0
	2	2	1	3	2	2	3	9	5	4	3	2	2	8	6	4	1	0
	2	2	1	3	2	2	3	9	5	4	3	2	2	8	6	4	1	0
12	1	1	1	5	1	1	1	5	1	1	5	1	1	13	9	1	1	0
	2	1	1	5	2	1	1	5	2	1	5	2	1	14	9	1	1	0
	2	1	1	5	1	1	1	5	1	1	5	1	1	13	9	1	1	0
	2	1	1	5	1	1	1	5	1	1	5	1	1	13	9	1	1	0
13	1	1	2	4	-	-	2	4	-	-	10	-	-	-	-	0	0	2
	2	1	2	4	-	-	2	4	-	-	10	-	-	-	-	0	0	2
	2	1	2	4	-	-	2	4	-	-	10	-	-	-	-	0	0	2
	2	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
	2	2	2	4	1	1	4	10	4	3	10	4	3	10	7	1	1	118
	2	2	2	4	1	1	4	10	4	3	10	4	3	10	7	1	1	118
	2	2	2	4	1	1	4	10	4	3	10	4	3	10	7	1	1	118
Missing	0	0	0	0	3	3	0	0	0	0	0	0	0	0	0	3		

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	MIS
1	1	2	1	3	2	1	0	0	0
1	2	2	1	3	2	1	0	0	0
1	3	2	1	3	2	1	0	0	0
2	1	2	1	2	1	1	-	-	2
2	2	2	1	2	1	1	-	-	2
2	3	2	1	2	1	1	-	-	2
3	1	2	1	2	2	2	0	0	0
3	2	2	1	2	2	2	0	0	0
3	3	2	1	2	2	2	0	0	0
4	1	1	1	1	1	2	-	-	2
4	2	1	1	4	1	2	-	-	2
4	3	1	1	3	1	2	-	-	2
5	1	2	1	4	2	1	-	-	2
5	2	2	1	4	2	1	-	-	2
5	3	2	1	4	2	1	-	-	2
6	1	3	2	3	2	2	1	91	0
6	2	3	2	3	2	2	1	91	0
6	3	3	2	3	2	2	1	91	0
7	1	1	2	1	1	2	1	156	0
7	2	1	1	1	1	2	1	156	0
7	3	1	1	1	1	2	1	156	0
8	1	2	2	1	1	1	1	23	0
8	2	2	2	1	1	1	1	23	0
8	3	2	2	1	1	1	1	23	0
9	1	2	1	2	2	1	-	-	2
9	2	2	1	2	2	1	-	-	2
9	3	2	1	2	2	1	-	-	2
10	1	2	1	3	2	1	0	0	0
10	2	2	1	3	2	1	0	0	0
10	3	2	1	3	2	1	0	0	0
11	1	2	1	3	2	2	-	-	2
11	2	2	1	3	2	2	-	-	2
11	3	2	1	3	2	2	-	-	2
12	1	1	1	5	1	1	-	-	2
12	2	1	1	5	2	1	-	-	2
12	3	1	1	5	1	1	-	-	2
13	1	1	2	4	1	2	0	0	0
13	2	1	2	4	2	2	0	0	0
13	3	1	2	4	2	2	0	0	0
14	1	2	2	4	1	1	1	118	0
14	2	2	2	4	1	1	1	118	0
14	3	2	2	4	1	1	1	118	0
Missing	0	0	0	0	0	0	18	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	MIS
1	1	2	1	3	2	1	0	0	0
1	2	2	1	3	2	1	0	0	0
1	3	2	1	3	2	1	0	0	0
2	1	2	1	2	1	1	1	0	1
2	2	2	1	2	1	1	1	0	1
2	3	2	1	2	1	1	1	0	1
3	1	2	1	2	2	2	0	0	0
3	2	2	1	2	2	2	0	0	0
3	3	2	1	2	2	2	0	0	0
4	1	1	1	1	1	2	1	0	1
4	2	1	1	4	1	2	0	0	0
4	3	1	1	3	1	2	0	0	0
5	1	2	1	4	2	1	1	0	1
5	2	2	1	4	2	1	1	0	1
5	3	2	1	4	2	1	1	0	1
6	1	3	2	3	2	2	1	91	0
6	2	3	2	3	2	2	1	91	0
6	3	3	2	3	2	2	1	91	0
7	1	1	2	1	1	2	1	156	0
7	2	1	1	1	1	2	1	156	0
7	3	1	1	1	1	2	1	156	0
8	1	2	2	1	1	1	1	23	0
8	2	2	2	1	1	1	1	23	0
8	3	2	2	1	1	1	1	23	0
9	1	2	1	2	2	1	0	0	0
9	2	2	1	2	2	1	0	0	0
9	3	2	1	2	2	1	0	0	0
10	1	2	1	3	2	1	0	0	0
10	2	2	1	3	2	1	0	0	0
10	3	2	1	3	2	1	0	0	0
11	1	2	1	3	2	2	1	0	1
11	2	2	1	3	2	2	1	0	1
11	3	2	1	3	2	2	1	0	1
12	1	1	1	5	1	1	0	0	0
12	2	1	1	5	2	1	1	0	1
12	3	1	1	5	1	1	1	0	1
13	1	1	2	4	1	2	0	0	0
13	2	1	2	4	2	2	0	0	0
13	3	1	2	4	2	2	0	0	0
14	1	2	2	4	1	1	1	118	0
14	2	2	2	4	1	1	1	118	0
14	3	2	2	4	1	1	1	118	0
Missing	0	0	0	0	0	0	0	11	

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

REC	REP	VEHS	SEX	AGE	SAFE	CF	TEST	BAC	MISC
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	4	2	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	2	1	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	0

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
1	2	2	1	3	2	1	0	0	0
1	3	2	1	3	2	1	0	0	0
2	1	2	1	2	1	1	1	133	0
2	2	2	1	2	1	1	1	133	0
2	3	2	1	2	1	1	1	133	0
3	1	2	1	2	2	2	0	0	0
3	2	2	1	2	2	2	0	0	0
3	3	2	1	2	2	2	0	0	0
4	1	1	1	1	1	2	1	85	0
4	2	1	1	1	1	2	0	0	0
4	3	1	1	1	1	2	0	0	0
5	1	2	1	4	2	1	1	66	0
5	2	2	1	4	2	1	1	66	0
5	3	2	1	4	2	1	1	66	0
6	1	3	2	3	2	2	1	91	0
6	2	3	2	3	2	2	1	91	0
6	3	3	2	3	2	2	1	91	0
7	1	1	2	1	1	2	1	156	0
7	2	1	1	1	1	2	1	156	0
7	3	1	1	1	1	2	1	156	0
8	1	2	2	1	1	1	1	23	0
8	2	2	2	1	1	1	1	23	0
8	3	2	2	1	1	1	1	23	0
9	1	2	1	2	2	1	0	0	0
9	2	2	1	2	2	1	0	0	0
9	3	2	1	2	2	1	0	0	0
10	1	2	1	3	2	1	0	0	0
10	2	2	1	3	2	1	0	0	0
10	3	2	1	3	2	1	0	0	0
11	1	2	1	3	2	2	1	165	0
11	2	2	1	3	2	2	1	165	0
11	3	2	1	3	2	2	1	165	0
12	1	1	1	5	1	1	0	0	0
12	2	1	1	5	2	1	1	94	0
12	3	1	1	5	1	1	1	45	0
13	1	1	2	4	1	2	0	0	0
13	2	1	2	4	2	2	0	0	0
13	3	1	2	4	2	2	0	0	0
14	1	2	2	4	1	1	1	118	0
14	2	2	2	4	1	1	1	118	0
14	3	2	2	4	1	1	1	118	0
Missing	0	0	0	0	0	0	0	11	0

REC	REP	VEHS	SEX	AGE	SAFE	CF	TEST	BAC	MISS
1	1	2	1	3	2	1	0	0	0
1	2	2	1	3	2	1	0	0	0
1	3	2	1	3	2	1	0	0	0
2	1	2	1	2	1	1	1	133	0
2	2	2	1	2	1	1	1	133	0
2	3	2	1	2	1	1	1	133	0
3	1	2	1	2	2	2	0	0	0
3	2	2	1	2	2	2	0	0	0
3	3	2	1	2	2	2	0	0	0
4	1	1	1	1	1	2	1	85	0
4	2	1	1	1	1	2	0	0	0
4	3	1	1	1	1	2	0	0	0
5	1	2	1	4	2	1	1	66	0
5	2	2	1	4	2	1	1	66	0
5	3	2	1	4	2	1	1	66	0
6	1	3	2	3	2	2	1	91	0
6	2	3	2	3	2	2	1	91	0
6	3	3	2	3	2	2	1	91	0
7	1	1	2	1	1	2	1	156	0
7	2	1	1	1	1	2	1	156	0
7	3	1	1	1	1	2	1	156	0
8	1	2	2	1	1	1	1	23	0
8	2	2	2	1	1	1	1	23	0
8	3	2	2	1	1	1	1	23	0
9	1	2	1	2	2	1	0	0	0
9	2	2	1	2	2	1	0	0	0
9	3	2	1	2	2	1	0	0	0
10	1	2	1	3	2	1	0	0	0
10	2	2	1	3	2	1	0	0	0
10	3	2	1	3	2	1	0	0	0
11	1	2	1	3	2	2	1	165	0
11	2	2	1	3	2	2	1	165	0
11	3	2	1	3	2	2	1	165	0
12	1	1	1	5	1	1	0	0	0
12	2	1	1	5	2	1	1	94	0
12	3	1	1	5	1	1	1	45	0
13	1	1	2	4	1	2	0	0	0
13	2	1	2	4	2	2	0	0	0
13	3	1	2	4	2	2	0	0	0
14	1	2	2	4	1	1	1	118	0
14	2	2	2	4	1	1	1	118	0
14	3	2	2	4	1	1	1	118	0
Missing	0	0	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	89
13	0
14	118

5. 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	
IMPUTED	BAC>0	92	16
	BAC=0	66	299

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

COMBINED		CORONER	
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

6. Conclusion

In this article, we have presented the BAC imputation algorithm used by the Evaluation & Data Systems Division of the Road Safety and Motor Vehicle Directorate at Transport Canada. It loosely based on the two-stage approach and multiple imputation of [4], and the sequential regression of [3]; however, it is hoped that some of the drawbacks of these methods can be overcome by introducing replicates in the observations before imputation proper starts.

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].

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.