Mechanisms for multivariate outliers and missing values

Beat Hulliger, Tobias Schoch
University of Northwestern Switzerland FHNW, e-mail: beat.hulliger@fhnw.ch

Abstract

In any survey on economic data from enterprises or households outliers and missing values may occur. Usually several quantitative variables are collected and therefore the outlyingness and missingness is multivariate. In addition the survey design may be complex. In this paper the well-known mechanisms for missingness (MCAR, MAR) are combined with mechanisms for outlyingness (OCAR, OAR) and contamination (CCAR, CAR). The joint mechanism needed for inference is OMAR, a joint "at random" condition on the outlyingness and the missingness. In addition, a condition is needed which ensures that the outlyingness, which is latent, is detected properly. This condition, called separable contamination, is discussed.

Keywords: Missing at random, contamination, observed likelihood, robust estimation, outlier detection, imputation

Acknowledgements: This research was partially supported through the European Union FP7 research project AMELI (FP7-SSH-2007-217322).

1. A model for the outlier, contamination and missingness mechanisms

Outliers and missing values often occur in survey data. Both problems alone are difficult to treat and jointly the pose a challenge to inference. In addition survey data usually is multivariate. In the project AMELI (Münnich et al. 2011) a model for the mechanisms covering this mixture of problems is developed (Hulliger et al. 2011a). The well-known mechanisms for missingness (Little and Rubin 2002) are combined with mechanisms for outlyingness and contamination.

The proposed mechanism is shown in Figure 1. Many paths are omitted in the path-diagram. This reflects that certain multivariate dependencies must be excluded to make the treatment possible. The covariates X (known covariates) and Z (unknown) covariates may influence all other variables. However, the sample design is supposed not influenced by any other variable than known covariates X and later the influence of unknown covariates Z must be excluded to allow any inference.
2. Types of mechanisms

The full mechanism that creates the data observed in the sample $Y^-$ starts with a finite population $Y^\emptyset$ of which certain observations (the outliers) are contaminated to become representative outliers. The binary variable $O^r$ determines whether an observation is a representative outlier and if this is the case the representative contamination $Y^r$ is plugged into the data instead of the original data. The resulting population $Y^*$ is the population of inference or true population. Thus all estimands are functions of $Y^*$, which includes representative outliers. The notion of representative outliers introduced by Chambers (1986) is re-interpreted by assuming that these outliers are part of the true population. Non-representative outliers are contaminating the true population through $O^n$ and $Y^n$. Missing values through the response indicator matrix $R$ are added to the observable population $Y^+$ either as missing at random (MAR) or completely at random (MCAR) or not at random (MNAR). Finally the sampling process $S$ follows. The result is the sampled observed data $Y^-$. Outliers are created as completely at random (OCAR), as in classical robust statistics, or at random (OAR), i.e. conditionally on observed covariates $X$, or not at random (ONAR). The outlier mechanism determines, which observations are outliers. The contamination distribution determines the outlier value $Y^n$. Also the contamination may be completely at random (CCAR), at random (CAR) or not at random (CNAR).

The decomposition of the full likelihood shows the interplay between outlyingness, contamination and missingness. The roles of outlyingness and missingness can thus be clarified. Phenomena like masking of outlyingness by missingness can be discussed in the model. The model for the mechanisms has proved its usefulness in the simulations of the AMELI project (Hulliger et al. 2011b). Different scenarii may be simulated which come closer to the reality of the data generating process than the classical contamination
model in robust statistics which assumes OCAR-CCAR. The model is described and discussed fully in (Hulliger and Schoch 2013).

3. Conditions for inference

Since unlike missingness outlyingness is latent their roles are very distinct when it comes to inference. Assuming first, that outlyingness is observed (like missingness) a condition for a likelihood based inference on the distribution of $Y^*$ can be derived: joint outlyingness and missingness at random (OMAR). The only way to remedy the latency of outlyingness is to assume error-free detection in the sense that there is a non-empty subset of the sample space where only true observations lie, a safe support say, and where not all the observations are missing. This condition is called separable contamination. It refers to non-representative outliers because representative outliers in the safe support will be taken into account for inference on the true population. The conditions needed for inference on $Y^*$ thus are OMAR and separable contamination. In addition the outlier detection method must be able to determine at least a subset of the safe support. This capability depends largely on the distribution of the non-representative outliers and comes down to a classification problem (Hulliger 2013).

The BACON algorithm (Billor et al. 2000) and its extension to incomplete survey data BACON-EEM (Béguin and Hulliger 2008) build on separable contamination because these algorithms assume that there is an outlier-free subset where the algorithms can start.

The conditions for inference on the true population in the presence of missing values and outliers make clear, that contamination is a more difficult problem than missingness. Under appropriate assumptions about the joint outlyingness and missingness mechanism and on the nature of the contamination inference on the target population is possible. However, the conditions are very strong and may rarely be met in practice.

References


