Topics in Advanced Statistics - Individual Assignment

C.A. Vriends (440435)

March 2020

1. Introduction

In this paper, which is a simulation study, the robustness of the MM-regression estimator in the cases where missing data and cellwise contamination are present will be explored. Each case has its own implications for the robustness of the estimates. Although, it is possible to encounter both cases simultaneously in real-life applications, in this paper each case will be treated separately as it is a more effective approach to isolate the effects of each issue. Furthermore, two methods to impute missing data are considered: *multiple imputation* and the *(nonparametric) bootstrap*. Each method has its own imputation technique, multiple imputation is based on Iterative Robust Model-Based Imputation (IRMI) (Templ et al., 2011) and the bootstrap method is based on k-Nearest Neighbors imputation (kNN) (Kowarik and Templ, 2016), both available in VIM.

The paper proceeds as follows. Firstly, there is more in depth discussion of the missing data mechanisms and the cellwise contamination model and the challenges that each case entails. Subsequently, there is a description of the imputation methods. Lastly, there is an explanation of the simulation setup and a short discussion of the (interesting) results.

2. Methodology & Methods

2.1. Missing data mechanisms

Missing data is a practical issue that one has to take into consideration before delving further into the statistical analysis. This is in part due to the fact that the majority of estimators rely on complete data (there is no inherent mechanism to cope with missing data). A naive approach of deleting all the observations that have missing data causes a severe loss of information, thus a loss in efficiency, if one is willing to make assumptions about the mechanisms that caused the missingness, it is possible to retrieve some of the lost information. Furthermore, a naive deletion of these observations can lead to biased estimates if the missing data is not distributed at random within the data. Rubin (1976) made a distinction between three different mechanisms of missing data, Completely Missing At Random (MCAR), Missing At Random (MAR) and Missing Not At Random (MNAR).

If the data is missing according to the MCAR mechanism it is missing in a completely random fashion and the missingness is evenly distributed among the data, in a more formal sense this means that the probability of missing is not dependent on the data (e.g. a scanner in a supermarket is equally faulty, generating missing values, among all products in the supermarket, not for a specific category).

If the data is missing according to the MAR mechanism, it is still missing in a random manner, but it is completely random (or MCAR) within a subset of the data (e.g. the supermarket scanner is equally faulty among the vegetables, but not for other products). Furthermore, MAR is broader than MCAR, more formal, a MCAR situation is also a MAR situation, but not vice versa.

If the data is missing according to the MNAR mechanism, it is not missing in a random manner, the missingness is dependent on the data and on the missing data (e.g. the supermarket scanner is faulty among the meat products, but the probability of a faulty scan is higher for the larger meats than for the smaller meats).

Although the distinction of missing data mechanisms is a useful trichotomy, each one differs in how realistic it is in practice. MCAR is the most unrealistic mechanism as it is quite unlikely to occur in real-life applications. MAR and MNAR are more realistic. MNAR has its own difficulties as it requires external context on the missingness of data (e.g. in the case of the supermarket example, it requires one to actually scan the different meats to detect the pattern that is present). With regards to the robustness of different estimates, using an imputation method, such as bootstrap and multiple imputation, theoretical guarantees concerning the consistency are more often obtainable for MCAR and MAR mechanisms. While some guarantees might be possible if the MNAR mechanism is explicitly modeled in the process of imputation.

2.2. Cellwise contamination

Another issue that might arise in practice is cellwise contamination, instead of a complete observation that could be an outlier, it might occur that a specific variable or set of variables has outliers. This could, for example be the result of faulty measurement equipment, such that it records positively biased or negatively biased data (at random). Cellwise contamination is a more recent development in the field of (applied) statistics and it also requires a different contamination model than the more traditional (rowwise) Tukey-Huber contamination model. Alqallaf et al. (2009) suggest the following contamination model

$$x_i = \big(I_p - B_\varepsilon\big)y_i + B_\varepsilon z_i \tag{1}$$

$$\begin{aligned} y_i &\sim F && F \text{ is the } \textbf{model} \text{ distribution} \\ z_i &\sim G && G \text{ is the } \textbf{outlier} \text{ generating distribution} \\ B_\varepsilon &= diag\big(B_1, \ldots, B_p\big) && \text{where } B_j \sim Bin(1, \varepsilon) \end{aligned}$$

In this model the cellwise outliers occur at random within each variable with equal probability (ε). Cellwise contamination poses a specific problem as a small percentage of cellwise outliers might result in a high percentage of rowwise outliers, thus making the use of rowwise robust estimators not feasible or at the grave loss of efficiency (as the complete observation will be considered an outlier (thus discarded in the robust estimates of location and scatter), while it might be an outlier in just one specific variable). One approach to address the issue of cellwise contamination is to tackle the problem in three steps. First, attempt to detect the cellwise outliers and remove the values from these cells (impute missing values, in a sense). Second, if the cellwise contamination is true to the above described model, the missing cells are missing according to the MCAR mechanism, one can use different imputation methods to impute the missing cells. Third, use a (robust) estimator for the location and scatter, these can in turn be used for different statistical methods (e.g. PCA).

2.3. Multiple imputation (MI)

Before the advent of multiple imputation, single imputation, such as "best-fit" imputation, was often used (Rubin, 1987). It is merely the imputation of the missing data in a dataset one time, or one round. Multiple imputation, as the name suggests, uses multiple rounds of imputation and as a result has multiple complete imputed datasets. According to Rubin (1987) there are several benefits to use multiple rounds of imputation as opposed to one round. As there is uncertainty in the imputation of the missing value, if one uses a stochastic imputation technique, the variability in the imputations results in a more efficient estimation as opposed to single imputation and accounts for the uncertainty of imputation. In essence, multiple imputation estimates the variance of an estimator by combining the variability between imputed datasets with the variability within each imputed dataset. As it seems agnostic to the imputation technique used (one can insert domain knowledge in this manner), it requires the imputation technique to be proper, it has to draw from the correct empirical conditional (given the data that one observes) distribution. Furthermore, it is not limited to ignorable nonresponse, it can be extended to non-ignorable nonresponse as long as one specifies the conditional distribution in a correct manner. The `IRMI` imputation is a proper imputation method according to Templ et al. (2011). The stochastic component mentioned earlier is part of the requirements before an imputation technique is considered proper (Rubin, 1976).

2.4. *(Nonparametric) bootstrap (BS)*

The nonparametric bootstrap is a general method to approximate the empirical distribution of an estimator. In short, a population is generated according to a certain distribution F, often one is interested in a certain population estimate $\theta = g(F)$. As one samples data from the population, the best one can do is approximate F, this is the empirical distribution $\hat{F}$ and the estimate is based on this distribution $\hat{\theta} = g(\hat{F})$. It might be the case that $g(\cdot)$ is an estimator that has an exotic distribution, it is problematic to make any inferential statements as one is not certain about the quality of $\hat{\theta}$, is it a proper approximation of θ? The nonparametric bootstrap method attempts to solve this problem (Efron and Tibshirani, 1993).

In the case of missing data, the general bootstrap approach can be interpreted in the following manner. In essence, it is a paraphrased version of the idea presented in Efron (1994). Therefore, the same notation is used. The distribution F is concealed according to a certain mechanism, let suppose that this concealed distribution is $G = c(F)$. As one is interested in, an although unfeasible, $\theta_F = s(F)$, it is necessary to resort to a decent approximation, that is $\theta = t(G)$. The imputation mechanism is implicit in $t(\cdot)$, as an imputation is often quite involved, the resulting distribution is quite complex and one is not able to distill it to an elegant analytical expression. Therefore, to approximate $\hat{\theta} = t(\hat{G})$ one uses the nonparametric bootstrap. $t(G)$ is an accurate approximation (it is Fisher consistent) of $s(F)$, in this setting, when F is multivariate normal and the missing data is of the type ignorable nonresponse according to Efron (1994).

3. Simulation Study

3.1. *Setup*

In this section the setup of the simulation study will be outlined. The bias and variance for the estimator under consideration, the MM-estimator, in a regression context will be assessed. The different scenarios entail the robustness in the case of three different missing data mechanisms (MCAR, MAR and MNAR) and the robustness in case of the presence of cellwise contamination at different contamination levels. Due to the computational burden of the earlier described methods, the nonparametric bootstrap in particular, the number of simulation replications will be limited to 200 (except for the third scenario, the number of replications is set to 50).

Due to the aformentioned computational burden the DGP is restricted to a one-dimensional dependent variable and two independent variables $y_i = 2x_{1,i} + 3x_{2,i} + \varepsilon_i$. Although it is quite simple, it still is sufficient for the different scenarios that are considered. The model distribution of $y_i|x_{1,i}, x_{2,i}$ is a standard normal distribution, as ε_i is drawn from a standard normal distribution. $x_{1,i}$, $x_{2,i}$ are drawn, simultaneously, from a multivariate normal distribution with $\mu = (0, 0)'$ and a covariance matrix $\Sigma_{1,1} = 1.76$, $\Sigma_{2,2} = 1.87$, $\Sigma_{2,1} = \Sigma_{1,2} = 0.81$. The number of observations N is set to 1000, regardless of the scenario.

3.2. *Correctness of the estimates*

To give an indication of bias in the case of the coefficient estimates and its standard errors, the (mean) RMSE of the simulation will be used. In the case of the coefficient estimates, it is quite straightforward what the comparison should be, the true values of the data generating process. In the case of the standard errors, it is not that trivial, one could consider the estimates (of the MM-estimator) on the unaltered data, the issue with this approach is that it does not take into account the uncertainty that is associated with each imputation technique (MI with 10 or 20 imputations might have a different standard error, as one could be the more efficient choice in a certain situation, e.g. 10% versus 20% missing). Therefore, in a similar vein as the general bootstrap approach, one could use the standard deviation of the within-simulation estimates of the coefficients as an approximation to the true standard error (that is different for MI or BS). That is also the case in this simulation study, the standard deviation of the within-simulation coefficient estimates will be used as the true value for the RMSE.

3.2.1. *Scenario 1: Missing data mechanisms*

In this scenario the three different missing data mechanisms are considered. The missing values occur in x_1 and it depends on each data mechanisms how the values are set to NA (or missing). In each missing data mechanism the missing percentage is set to 20%.

In the MCAR mechanism, a random selection of 20% of the observations is made that is set to NA, this is independent of the distribution of either x_1 or x_2. In the MAR mechanism the values in x_1 are set to NA if the value in x_2 is in the 80th percentile, thus if a value is missing in x_1 it depends on the observed data x_2. In the MNAR mechanism the values in the 80th percentile of x_1 are set to NA, thus if a value is missing in x_1 it solely depends on the distribution of x_1.

Both the imputation methods will be used, in the case of multiple imputation the number of imputed datasets, or m, is set to 20, as this is in line with the rule of thumb suggested by Bodner (2008). The number of bootstraps, R, is set to 100. It is clear to the author that this is a low number compared to what is suggested in the literature, it is however gravely impractical to increase this number, as the computational burden of the bootstrap approach is severe.

3.2.2. Scenario 2: Cellwise contamination

In this scenario three different contamination levels, $\varepsilon = (10\%, 20\%, 30\%)$, will be investigated, the cells, there are $N * 2$ possible cells, are selected completely at random for the four different contamination levels. The model distribution is the standard multivariate normal distribution as mentioned earlier, the outlier generating distribution is a normal distribution with $\mu = 10$ and $\sigma = 1$. The outlier generating distribution is sufficiently different from the model distribution that the Detect Deviating Cells (or `DDC`) (Rousseeuw and Bossche, 2017) method should be able to sufficiently separate the outlying cells from the non-outlying cells, although there is no guarantee of no false-positives nor false-negatives.

3.2.3. Scenario 3: Cellwise contamination & increasing extremity

As mentioned in previous scenario, the biasedness of the estimates is dependent on the ability of `DDC` to properly distinct the outlying cells from the non-outlying cells. In a sense, having perfect sensitivity and perfect specificity. This scenario will investigate the biasedness of the estimates if the extremity of the outlier is (stepwise) increased. The outliers will be generated from a normal distribution with $\mu = 1$ and $\sigma = 1$ and the different levels of extremity $\delta = (1,\ 2.5,\ 5)$. The outlier sampled from the normal distribution will be multiplied by the extremity factor δ. For the level of contamination, that has a meaningful effect, $\varepsilon = 30\%$ will be used.

3.3. Results

3.3.1. Scenario 1: Missing data mechanisms

The results for the different missing data mechanisms in conjunction with the two imputation methods are presented below. As one can observe the meam RMSE, for the coefficients, as well as for the standard errors, is the lowest for both imputation methods if the missing data mechanism is MCAR. This makes sense, especially for the BS, the MM-estimator inherits the asymptotic properties from the M-estimator (it inherits the breakdown point from the S-estimator) and it is thus asymptotically normal (Yohai, 1987). In these circumstances, ignorable nonresponse and asymptotically normality, BS is consistent. One slightly surprising result is the relatively high mean RMSE for BS in the case of MAR, MAR still is a case of ignorable nonresponse, therefore it should also exert the same results as in the case of MCAR. This result, and the difference in mean RMSE between MI and BS, might be due to the relatively low number of bootstrap samples compared to what is commonly advised.

Furthermore, in the case of MI, the imputation technique is not modified to cope with the case of non-ignorable nonresponse (or MNAR), and in the case of BS, it is similar, the MNAR missing data mechanism is not incorporated in the nonparametric BS method (the so called *full-mechanism bootstrap* (Efron, 1994)). Hence, one would expect the coefficients to be biased. For both imputation methods, the mean RMSE of the coefficients as well as of the standard errors is higher than in the case of MAR and especially in the case of MCAR.

Scenario 1: Missing data mechanisms (20% missing)

	Multiple Imputation (MI)					
	MCAR		MAR		MNAR	
		$\hat{\sigma}$		$\hat{\sigma}$		$\hat{\sigma}$
$\hat{\beta}_1$	2.0036	0.0293	2.0013	0.0293	2.1814	0.0344
$\hat{\beta}_2$	3.0009	0.0288	2.9515	0.0310	3.0185	0.0303
Mean RMSE	0.0010	$9.18 * 10^{-6}$	0.0024	$1.21 * 10^{-5}$	0.018	$2.17 * 10^{-5}$

	Bootstrap (BS)					
	MCAR		MAR		MNAR	
		$\hat{\sigma}$		$\hat{\sigma}$		$\hat{\sigma}$
$\hat{\beta}_1$	2.0682	0.0326	2.0083	0.0410	2.2776	0.0459
$\hat{\beta}_2$	2.9644	0.0325	2.7605	0.0624	3.0856	0.0388
Mean RMSE	0.0040	$9.42 * 10^{-6}$	0.0310	$1.03 * 10^{-5}$	0.0438	$2.019 * 10^{-5}$

3.3.2. Scenario 2: Cellwise contamination

The varying probabilities of contamination have an effect if it is not taken into consideration. This can be observed in the contaminated results, one interesting result is the fact that the regular robust MM-estimator, is still estimating the coefficients in quite a correct manner for the 10% and 20% contamination levels. Therefore, it is not that surprising that the estimates for MI and BS with `DDC` are not far off the true values (biased) and as a result have quite low mean RMSEs.

However, if the cellwise contamination is not taken into account, at the 30% contamination level the MM-estimator is severely biased. Even if the contamination level is quite high, the MI and BS, in combination with `DDC`, manage to limit the negative effect. Furthermore, one indication that `DDC` is working quite well, is the fact that there does not seem to be a clear pattern in the RMSE, it does not monotically increase with the increase in the level of contamination.

Scenario 2: Cellwise contamination (ε = 10%, 20% & 30%)

	Multiple Imputation (MI) with `DDC`					
	10%		20%		30%	
		$\hat{\sigma}$		$\hat{\sigma}$		$\hat{\sigma}$
$\hat{\beta}_1$	2.0033	0.0298	2.0075	0.0312	2.0226	0.0340
$\hat{\beta}_2$	3.0046	0.0291	3.0023	0.0303	2.9931	0.0330
Mean RMSE	$8.81 * 10^{-4}$	$4.41 * 10^{-6}$	$1.08 * 10^{-3}$	$8.43 * 10^{-6}$	$2.1 * 10^{-3}$	$1.20 * 10^{-4}$

	Bootstrap (BS) with `DDC`					
	10%		20%		30%	
		$\hat{\sigma}$		$\hat{\sigma}$		$\hat{\sigma}$
$\hat{\beta}_1$	2.0157	0.0329	2.0380	0.0388	2.0158	0.0329
$\hat{\beta}_2$	3.0122	0.0322	3.0132	0.0380	3.0123	0.0323
Mean RMSE	$1.13 * 10^{-3}$	$1.65 * 10^{-5}$	$2.05 * 10^{-3}$	$2.33 * 10^{-5}$	$1.10 * 10^{-3}$	$1.65 * 10^{-5}$

	Contaminated estimates (*solely the MM-estimator*)					
	10%		20%		30%	
		$\hat{\sigma}$		$\hat{\sigma}$		$\hat{\sigma}$
$\hat{\beta}_1$	2.0246	0.0230	2.0006	0.0332	0.0894	0.0196
$\hat{\beta}_2$	2.9501	0.0298	2.9879	0.0318	3.8462	0.0709

3.3.3. Scenario 3: Cellwise contamination & increasing extremity

It is apparent that the first outlier extremity is hard to identify by the `DDC` algorithm, it has a quite low specificity. This is to be expected, as the outlier generation distribution does not differ that severely from the distribution that generates the independent variables (except for the correlation structure, as the outliers are independent and identically distributed). It does bias the coefficient estimates, but not in a drastic sense, as the RMSE is still quite low.

The second outlier extremity is better identifiable by the `DDC` algorithm, as the specificity increases, this is to be expected. There is a stronger signal that a cell is an outlier, as it is more extreme. The effect on the coefficient estimates is stronger, as there are still some outliers that are not ommitted from the contaminated data, these bias the results in a more severe manner than when the outliers were less detected, but also less severe, this can be observed from the (mean) RMSE, it is a steep increase from the first outlier extremity. This outlier extremity causes issues for the estimation on the contaminated dataset, as the estimates are gravely biased.

The third outlier extremity is even better identifiable by the `DDC` algorithm. As is clear from the specificity (and also general accuracy), one consequence of this fact is that the BS method seems to have less biased estimates compared to the previous outlier extremity, this is also the case for the MI method, although it is not a decline of similar significance (in (mean) RMSE). Again, this outlier extremity causes issues for the estimation on the contaminated dataset, as the estimates are gravely biased.

Scenario 3: Cellwise contamination ($\varepsilon = 30\%$) & increasing extremity (δ = (1, 2.5, 5))

Multiple Imputation (MI) with `DDC`

	$\delta = 1$		2.5		5	
		$\hat{\sigma}$		$\hat{\sigma}$		$\hat{\sigma}$
$\hat{\beta}_1$	1.721	0.0988	1.1580	0.0312	1.1916	0.1687
$\hat{\beta}_2$	2.980	0.0995	2.6580	0.0303	2.8180	0.2164
Mean RMSE	0.0483	$1.12 * 10^{-4}$	0.5251	$2.21 * 10^{-2}$	0.3746	$3.23 * 10^{-3}$

Bootstrap (BS) with `DDC`

	$\delta = 1$		2.5		5	
		$\hat{\sigma}$		$\hat{\sigma}$		$\hat{\sigma}$
$\hat{\beta}_1$	1.7500	0.1004	1.2302	0.1768	1.7420	0.1582
$\hat{\beta}_2$	2.976	0.1014	3.0000	0.1848	3.066	0.0957
Mean RMSE	0.0390	$4.66 * 10^{-4}$	0.3221	$2.27 * 10^{-3}$	0.0464	$2.30 * 10^{-3}$

`DDC` *identification characteristics*

Sensitivity	98.36%	99.92%	99.98%
Specificity	11.89%	31.10%	60.38%
Accuracy	72.42%	79.28%	88.10%

Contaminated estimates (*solely the MM-estimator*)

	$\delta = 1$		2.5		5	
		$\hat{\sigma}$		$\hat{\sigma}$		$\hat{\sigma}$
$\hat{\beta}_1$	1.7270	0.0919	0.4347	0.0700	0.2751	0.0601
$\hat{\beta}_2$	2.9740	0.0899	1.8470	0.2263	0.6257	0.1063

References

Alqallaf, F., Aelst, S. V., Yohai, V. J., Zamar, R. H., 2009. Propagation of outliers in multivariate data. The Annals of Statistics 37, 311–331.

Bodner, T. E., 2008. What improves with increased missing data imputations? Structural Equation Modeling: A Multidisciplinary Journal 15, 651–675.

Efron, B., 1994. Missing data, imputation, and the bootstrap. Journal of the American Statistical Association 89, 463–475.

Efron, B., Tibshirani, R. J., 1993. An Introduction to the Bootstrap. No. 57 in Monographs on Statistics and Applied Probability, Chapman & Hall/CRC, Boca Raton, Florida, USA.

Kowarik, A., Templ, M., 2016. Imputation with the R Package VIM. Journal of Statistical Software 74.

Rousseeuw, P. J., Bossche, W. V. D., 2017. Detecting deviating data cells. Technometrics 60, 135–145.

Rubin, D. B., 1976. Inference and missing data. Biometrika 63, 581–592.

Rubin, D. B., 1987. Multiple Imputation for Nonresponse in Surveys. Wiley.

Templ, M., Kowarik, A., Filzmoser, P., 2011. Iterative stepwise regression imputation using standard and robust methods. Comput. Stat. Data Anal. 55, 2793–2806.

Yohai, V. J., 1987. High breakdown-point and high efficiency robust estimates for regression. The Annals of Statistics 15, 642–656.