

Data Quality Control Based on Metric Data Models

Veit Köppen¹ and Hans - J. Lenz²

¹Institute of Production, Information Systems and Operations Research, Freie Universität Berlin, Garystr. 21, D-14195 Berlin, Germany, koeppen@wiwiss.fu-berlin.de

²Institute of Statistics and Econometrics, Freie Universität Berlin, Garystr. 21, D-14195 Berlin, Germany, hjlenz@wiwiss.fu-berlin.de

Summary. We consider statistical edits defined on a metric data space spanned by the non-key attributes (variables) of a given database. Integrity constraints are defined on this data space based on definitions, behavioral equations or a balance equation system. As an example think of a set of business or economic indicators. The variables are linked by the four basic arithmetic operations only. Assuming a multivariate Gaussian distribution and an error in the variables model estimation of the unknown (latent) variables can be carried out by a generalized least-squares (GLS) procedure. The drawback of this approach is that the equations form a non-linear equation system due to multiplication and division of variables, and that generally one assumes independence between all variables due to a lack of information in real applications. As there exists no finite parameter density family which is closed under all four arithmetic operations we use MCMC-simulation techniques, cf. Smith and Gelfand (1992) and Chib (2004) to derive the “exact” distributions in the non-normal case and under cross-correlation. The research can be viewed as an extension of Köppen and Lenz (2005) in the sense of studying the robustness of the GLS approach with respect to non-normality and correlation.

1 Introduction

Fellegi and Holt (1976) published a break-through paper on automatic editing and imputation. Wetherill and Gerson (1987) put together the methodology about edits as validation rules on symbolic, logical, probabilistic and relational data sets. Lenz and Rödel (1991) extended validation rules to statistical edits, and Lenz and Müller (2000) to fuzzy edits in the case of metric data spaces. Liepins and Uppuluri (1990) published their view and pragmatics on data quality control. Aitchison (1986) considered non-negative measurements that sum up to unity.

Recently, Batini and Scannapieco (2006) put together the methodology about data quality known in both areas ‘Statistics’ and ‘Database Theory’.

In the following we are concerned with statistical edits, i.e. validation rules based on a fully specified model and defined on a metric data space. The model is assumed to be correctly specified. As an example think of sales = profit + costs as a linear equation which is true due to definition. Note, that the fundamental economic equation “sales = sold_quantity * unit_price” is a non-linear relation. Such relations can be represented by an error-in-the-variables model. Let ξ be a p -dimensional vector of error-free variables and x the corresponding observation vector with superimposed measurement errors u , i.e. we have $\xi = x + u$ as state space vector. The available knowledge about definitions and balance equations $\zeta = H(\xi)$ is encapsulated in the observation equation system with fixed dimension $q \in \mathbb{N}$. It is modeled as $z = H(\xi) + v$ where z is a q -dimensional observation vector and v an additive noise vector independent of u . If all state equations (definitions and balance equations) are linear, then H is a $(q \times p)$ observation matrix. Generally, some equations are non-linear leading to $H: \text{dom}(\xi) \rightarrow \text{dom}(z)$. Lenz and Rödel (1991) showed for linear models that, given the data (x, z) and the observation model H , ξ and ζ can be estimated by a generalized least-squares (GLS) approach by $\hat{\xi} = x + K(z - Hx)$ and $\hat{\zeta} = Hx$ where $K = PH'(HPH' + R)^{-1}$ and the covariance matrices of the errors u, v are given by $P = \sum_{uu}$ and $R = \sum_{vv}$. The estimators are best (UMVUE) for a quadratic loss function if u, v are jointly Gaussian distributed.

In the following we shall relax the assumption of a joint Gaussian distribution and, moreover, assume cross-correlation between the variables according to some prior information. This implies to substitute GLS estimation by MCMC simulation, cf. Chib (2004) and Köppen and Lenz (2005). In this sense the study can be viewed as a study of robustness with respect to non-normality and dependencies between the variables. First, we introduce a simple model, and then we present the simulation approach and close with various scenarios showing the main effects of deviations between GLS estimates and estimates based upon our MCMC simulation.

2 Business Indicators Model

It is sufficient for our purposes to consider a simplified business indicators model M based on two equations and five variables. We have the structural equation system

$$\text{Sales} = \text{Profit} + \text{Cost}$$

$$\text{Return-on-Investment (ROI)} = \text{Profit} / \text{Capital}.$$

Evidently, there are two endogenous and three exogenous random variables. The only assumption about M we need in the following is that each equation fulfills the separability condition. This means that each equation of M is uniquely resolvable for each variable showing up on its right hand side (RHS). In order to simplify the notation we write $x \sim N(\mu, \sigma^2)$ instead of $x = \mu + u$ with $u \sim N(0, \sigma^2)$. For instance, $\text{costs} \sim N(80, 8^2)$. The mean is either estimated from the observed values or is known. The variance is assumed to be known too. The distributions considered include the normal (Gaussian), a skewed multivariate normal, exponential, gamma and the Dirichlet distribution. The correlation coefficients between pairs of variables can vary from ± 0.7 , ± 0.6 , ± 0.4 to 0.0 .

3 Estimation by MCMC Simulation Technique

Each of the random variables x, z is described by its density function. We use the Metropolis-Hastings algorithm, cf. Hastings (1970), for the MCMC simulation of the random variables which are transformed according to M . In the first step we consider each of the p variables and assign that subset of equations to it, where it shows up either as a LHS or RHS variable. In the later case the corresponding equations are to be solved for the given variable to make it a LHS variable. For example, resolving for profit in M we get $\text{profit} = \text{sales} - \text{costs}$ and $\text{profit} = \text{ROI} * \text{capital}$. The next step is to start MCMC sampling of all RHS variables. Note that the distribution of each LHS variable is either fully specified or unknown. In the last case it will be estimated from the corresponding equations where it shows up as a LHS or RHS variable. If a random variable shows up in $1 < k \leq q$ equations the k simulations must be “fused”. Therefore k pairs of lower (\underline{q}_k) and upper (\overline{q}_k) $\alpha/2$ -quantiles for that variable are computed. Let us define $\underline{q}_{\max} = \max\{\underline{q}_1, \underline{q}_2, \dots, \underline{q}_k\}$ and $\overline{q}_{\min} = \min\{\overline{q}_1, \overline{q}_2, \dots, \overline{q}_k\}$. Then we have

Definition 1: A data set of an equation system M is called M -inconsistent (contradictive) if at least for one variable it is true that $\overline{q}_{\min} \leq \underline{q}_{\max}$.

In Fig.1 we illustrate M -inconsistency for the case of a variable x , say, sampled as x_1 and x_2 from two equations of M . Evidently, in the upper case the overlap I is empty, i.e. the data set is of bad quality while in the lower case the overlap I_q is non empty and the data are (weak) consistent. The final step is to project the joint distribution on the subspace spanned by $x_1 - x_2 = 0$ getting the density $f_{x_1, x_2}(x, x)$ for all $x \notin I_q = [\underline{q}_{\max}, \overline{q}_{\min}]$. The algorithm SamPro – sampling and projection – summarizes the procedure.

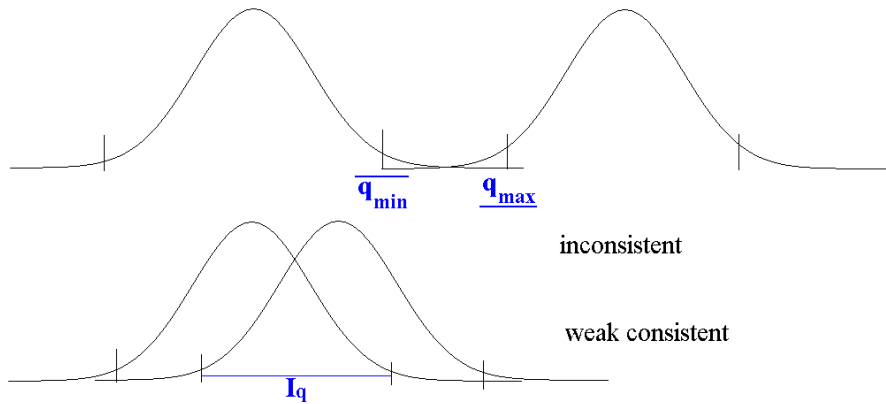


Figure 1: M-inconsistency and M-consistency

SamPro-Algorithm

input: a stochastic equation system M , one observation per variable (missing values allowed), error probability α

output: estimates for all variables, i.e. densities, means and standard deviations

begin

resolve (set $LHS^1 \equiv RHS$) for all variables in all equations

sample from the joint density function of all RHS variables for LHS

estimate all LHS variables

estimate $\alpha/2, 1-\alpha/2$ - quantiles $\underline{q}_{\max}, \overline{q}_{\min}$ for each variable, $i=1,2,\dots,p$.

if $\overline{q}_{\min} > \underline{q}_{\max}$ then “M -inconsistency found” and stop

else **compute** the distribution \hat{f}_{xz} restricted to the subspace $x-z = 0$

end

4 Scenarios and Robustness Analysis

Let us start with generating scenarios given the two-equation model M , i.e. sales = profit + cost as a linear equation and ROI = profit / capital as a nonlinear relation.

¹ $z = x_1 + x_2$ then z is LHS and the x_1 and x_2 are RHS.

In all of our experiments with up to 2.5 million replications each, we assume that a realization of the random variables profit, cost and capital is at hand. Moreover, their types of distributions are varying. This implies that at least the mean of the various distributions can be determined. Furthermore, prior information is available about the standard deviation or variance of the measurement errors. The other two variables, i.e. sales and ROI, are handled in experimental group A as variables with missing values (null values) and later in group B as (correctly or noisy) observed values. If missing values of variables exist, they must be estimated, i.e. imputed.

The MCMC simulation results are compared with GLS estimation approach using the software package QUANTOR, originated as PRTI by Schmid (1979). In experimental group A the first three experiments analyze the effect of skewed distributions compared with Gaussian distributions, if all variables are not correlated. The next two experiments separately investigate the effects of cross-correlation. Finally, the interaction between non-normality and correlation is of concern.

In experimental group B the data set is complete and Gaussian distributions are assumed, that means no missing values exist. The effects of negative, zero and positive correlation are studied for two cases: The measurement of the variables fulfill (“M -inconsistent variables”) or do not fulfill the balance equation system.

Experimental Group A: Effects of non-normality and / or correlation in the case of missing values

4.1 Scenario 1: Normality; no correlation

Specification of the distributions:

Profit $\sim N(20, 2^2)$; Cost $\sim N(80, 8^2)$; Capital $\sim N(60, 6^2)$

Missing values:

Sales, ROI unknown

Result: The MCMC simulation and the GLS estimation result for sales as part of a linear relation are compatible with respect to mean and standard deviation (sd). The same is true for the imputation of ROI as a nonlinear relationship. Note, that the Gaussian hypothesis about the distribution of all variables is valid.

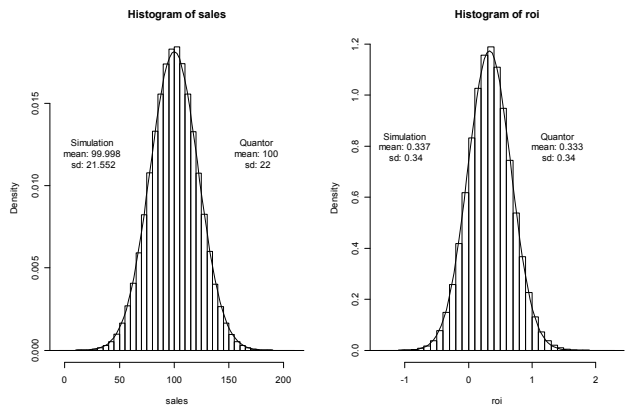


Figure 2: Normal distributed variables with no correlation

4.2 Scenario 2: Effect of Skewness; no correlation

Specification of the distributions:

Profit $\sim \text{Exp}(1/20)$ vs. $N(20, 20^2)$; Cost $\sim \text{Gamma}(8, 0.1)$ vs. $N(80, 28^2)$; Capital $\sim \text{Gamma}(15, 0.25)$ vs. $N(60, 8.6^2)$

Missing values:

Sales, ROI unknown

Result: In the linear case the mean and the standard deviation are similar, as the experiments for the variable sales show. In the nonlinear case the mean and the standard deviation (sd) are overestimated by about 15%.

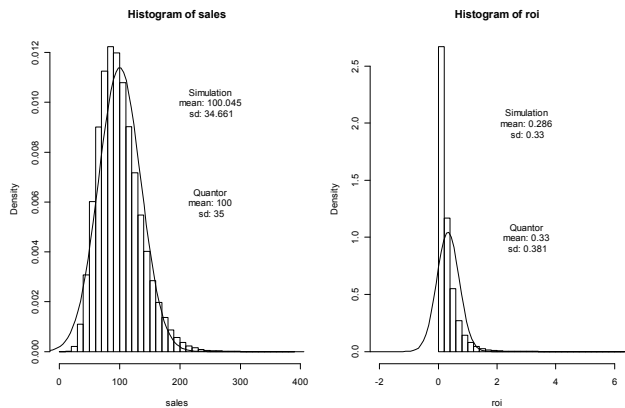


Figure 3: Effect of skewed distributions of all (observed) variables; no correlation

4.3 Scenario 3: Normality and negative cross-correlation

Specification of the distributions:

Profit $\sim N(20, 2^2)$; Cost $\sim N(80, 8^2)$; Capital $\sim N(60, 6^2)$

Missing values:

Sales, ROI unknown

Correlation used for simulation:

$\rho(\text{profit, cost}) = -0.7$; $\rho(\text{profit, capital}) = -0.7$

Result: While the means are nearly equal, the standard deviations differ between +18% and -16%.

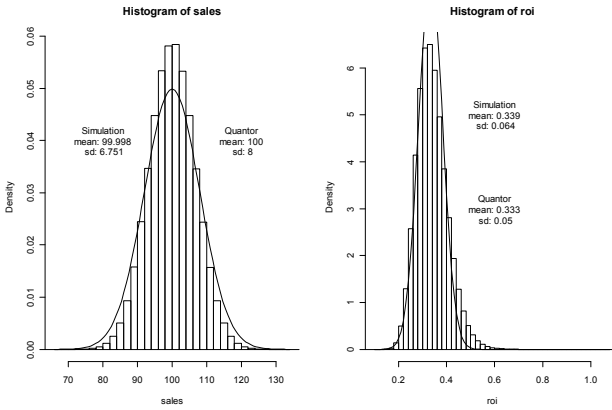


Figure 4: Normality and negative correlation ($\rho = -0.7$)

4.4 Scenario 4: Normality and positive cross-correlation

Specification of the distributions:

Profit $\sim N(20, 2^2)$; Cost $\sim N(80, 8^2)$; Capital $\sim N(60, 6^2)$

Missing values:

Sales, ROI unknown

Correlation used for simulation:

$\rho(\text{profit, cost}) = 0.7$; $\rho(\text{profit, capital}) = 0.7$

The positive sign of the correlation coefficients is contra intuitive from a manager's point of view. Nevertheless, it is used here more formally as an opposite case to negative correlation in scenario 3.

Result: While the means have nearly the same values, the percentage of differences changes sign: -19% for sales vs. +16% for ROI.

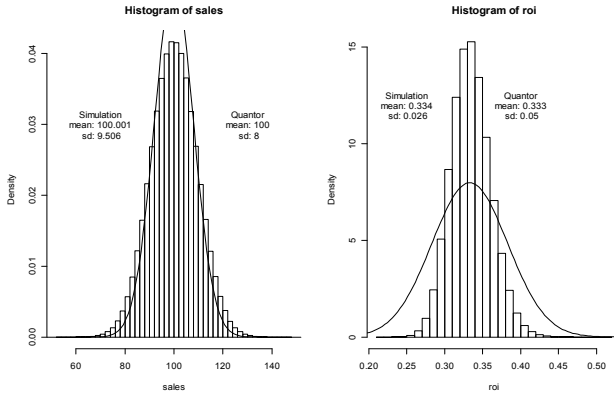


Figure 5: Normality and positive correlation ($\rho = 0.7$)

4.5 Scenario 5: Skewness and positive cross-correlation

Specification of the distributions:

$$(Profit, Costs, Capital) \sim MSN \left(\xi = (2, 8, 6), \Omega = \begin{pmatrix} 1 & 0.6 & 0.7 \\ 0.6 & 1 & 0.2 \\ 0.7 & 0.2 & 1 \end{pmatrix}, \alpha = (2500, 80, 300) \right) v.s.$$

$$Profit \sim N(20, 4.3^2)$$

$$Costs \sim N(80, 8.3^2)$$

$$Capital \sim N(60, 7.3^2)$$

Missing values:

Sales, ROI unknown

Correlation used for simulation: This correlation is imposed by multivariate skewed normal distribution (MSN) (see Azzalini and Valle (1996) with the above given parameters. Note that the parameterisation is adopted from Azzalini and Capitanio (1999).

$$\rho(\text{profit}, \text{cost}) = 0.4; \rho(\text{profit}, \text{capital}) = 0.5$$

Result: While the means have nearly the same value for both variables, this is not true for the standard deviation (sd). For sales we get $sd = 10.87$ by simulation and $sd' = 9.4$ under the Gaussian assumption. The corresponding values for ROI are $sd = 0.063$ vs. $sd' = 0.069$.

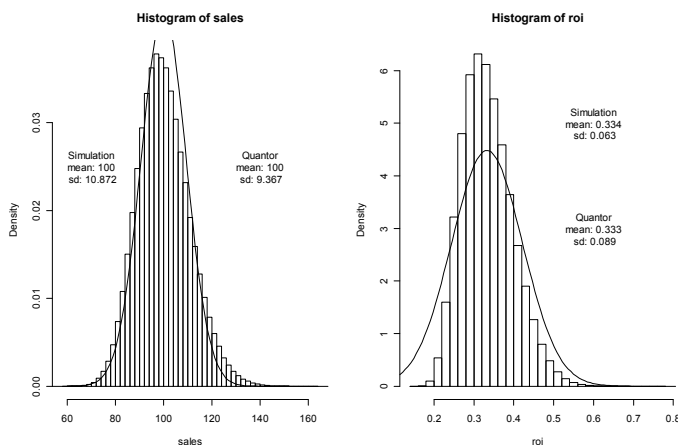


Figure 6: Skewness and positive cross-correlation

4.6 Scenario 6: Skewness and negative cross-correlation

Specification of the distributions:

(Profit, Cost, Capital) \sim Dir(30, 40, 8) vs.

Profit \sim N(20, 2.8²); Cost \sim N(80, 8.8²); Capital \sim N(60, 20²)

Missing values:

Sales, ROI unknown

Correlation used for simulation imposed by Dirichlet (Dir) distribution:

$\rho(\text{profit, cost}) = -0.8$; $\rho(\text{profit, capital}) = -0.3$

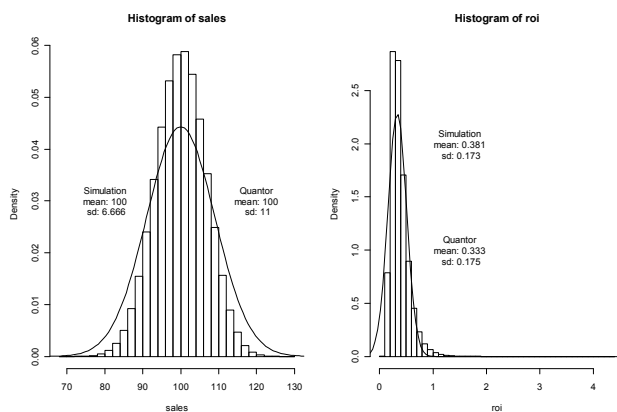


Figure 7: Non-Normality and negative cross-correlation

Result: While the means of the variable ‘sales’ are identical, this is not true for ROI. The GLS approach over-estimated the simulated (exact) values about 12%. Quite opposite, the standard deviation of sales is over-estimated by GLS by about 65% while sd of ROI is about the same.

Experimental Group B: Normality, cross-correlation, no missing values

Scenario 1: Normality, M - consistent observations of sales and ROI, and cross-correlation

Specification of the distributions:

Profit $\sim N(20, 2^2)$; Cost $\sim N(80, 8^2)$; Capital $\sim N(60, 6^2)$; Sales $\sim N(100, 10^2)$; ROI $\sim N(0.333, 0.333^2)$

Note that as above the distribution of profit is threefold determined by the prior $N(20, 2^2)$, by profit = sales – cost and by profit = ROI * capital.

Missing values: no

Correlation Matrix specified as: $\rho \in \{-0.4, 0.0, 0.4\}$. Economic reasoning leads to different signs of the correlation coefficient ρ . This expert knowledge might vary from company or business sector and represents in this example only one possible specification out of many. The lower and upper bounds of ρ are necessary to ensure a positive definite correlation matrix **R**:

$$R = \begin{bmatrix} & \text{profit} & \text{cost} & \text{capital} & \text{sales} & \text{roi} \\ \text{profit} & 1 & \rho & 0 & -\rho & -\rho \\ \text{cost} & \rho & 1 & 0 & -\rho & \rho \\ \text{capital} & 0 & -\rho & 1 & 0 & \rho \\ \text{sales} & -\rho & -\rho & 0 & 1 & -\rho \\ \text{roi} & -\rho & \rho & \rho & -\rho & 1 \end{bmatrix}$$

Results: The variance of costs, capital, sales and ROI is proportional to ρ . The variance of the estimated profit is non monotonic in ρ and has its maximum at $\rho = 0.2$. The means of all variables are more or less constant.

Variable	Mean	Sd	Mean	Sd
	Prior		Posterior	
Profit	20.00	2	19.87	1.58
Costs	80.00	8	79.95	6.29
capital	60.00	6	59.81	4.89
sales	100.00	10	99.96	6.37
ROI	0.33	0.33	0.33	0.03

Table 1: Means and Standard Deviations (Sd) of all observed and simulated variables, Gaussian distributions, no correlation, no missing values, M-consistent observations.

We close scenario 1 by presenting three three-dimensional scatter plots showing the simulated values of the variable profit determined from the prior distribution and the two RHS of the model equations for $\rho = -0.4, 0.0, 0.4$.

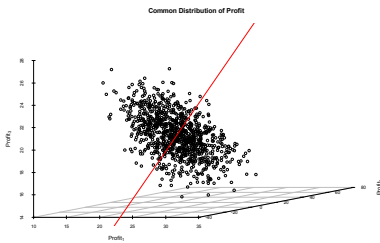


Fig. 9a: Scatter plot of simulated profit values for $\rho = -0.4$

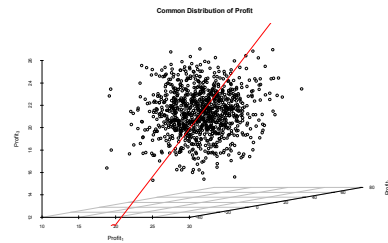


Fig. 9b: Scatter plot of simulated profit values for $\rho = 0.0$

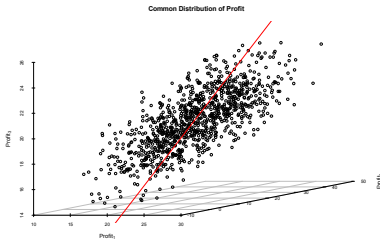


Fig. 9c: Scatter plot of simulated profit values for $\rho = +0.4$

Scenario 2: Normality, M-inconsistent observation of sales and ROI, and cross-correlations as in scenario 1

Specification of the distributions:

Profit $\sim N(30, 3^2)$; Cost $\sim N(80, 8^2)$; Capital $\sim N(60, 6^2)$; Sales $\sim N(100, 10^2)$; ROI $\sim N(0.333, 0.333^2)$

Note: As in scenario 1 the distribution of profit is threefold determined by the prior $N(30, 3^2)$, and by the equations profit = sales – costs and profit = ROI * capital. The mean and standard deviation of profit are increased from $N(20, 2^2)$ to $N(30, 3^2)$. This implies M-inconsistency of the observed values (means) of profit = sales - costs and profit = ROI * capital.

Missing values: no

Correlation Matrix as above with $\rho \in \{-0.4, 0.0, 0.4\}$

Results: The variance of costs, capital, sales and ROI is proportional to ρ . The variance of profit is non monotonic and gets a maximum at $\rho = 0.2$. The mean of profit is monotonically increasing, the means of the remaining variables are more or less constant. The case $\rho = 0.4$ leads to incoherency of profit, cf. Fig 9c, thus implying the M-incoherency of the whole equation system with the data set. Note that the observed value of each variable is equal to its corresponding (estimated) mean. To ensure that the first moments fulfill the equation system in a case of weak consistency, it might be necessary to iterate the SamPro algorithm. But after a few iterations this is achieved. In Tab. 2 the results of the 5th iteration are given. The first moments fulfill the equation system up to a small error.

Variable	Mean	Sd	Mean	Sd
	Prior		Posterior	
profit	30	3	24.93	0.85
costs	80	8	78.51	1.84
capital	60	6	67.03	2.33
sales	100	10	103.46	1.84
ROI	0.333	0.333	0.372	0.013

Table 2: Means and Standard Deviations of all observed and simulated variables, Gaussian distributions, no correlation, no missing values

Finally, we present three scatter plots in Fig. 9a-c for the variable profit with $\rho \in \{-0.4, 0.0, 0.4\}$. Note the effect of a “too large” mean of profit, i.e. $N(30,3^2)$, on the overlap of the point cloud and the (linear) subspace spanned by simulated values of profit, profit_1 and profit_2 .

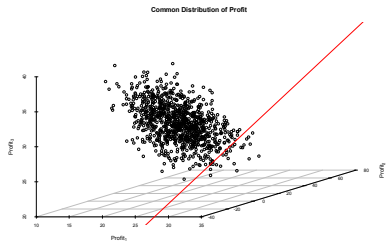


Fig. 10a: Scatter plot of simulated profit values for $\rho = -0.4$

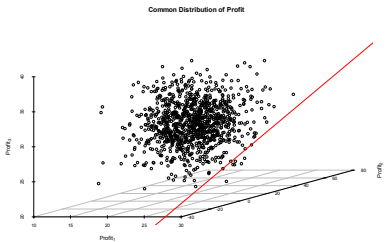


Fig. 10b: Scatter plot of simulated profit values for $\rho = 0.0$

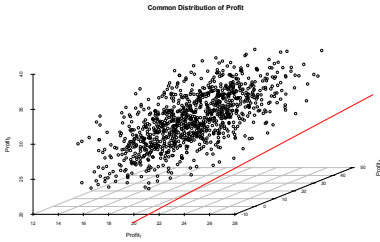


Fig. 10c: Scatter plot of simulated profit for $\rho = +0.4$

It is worthwhile mentioning that the simulated values of an M-consistent variable should lie on the straight line as in Fig. 9c, i.e. should fulfil all balance equations. As mentioned above the empty intersection of variable profit is caused by a “too large” (estimated) mean of the related distribution.

5 Conclusion

We can summarize our study of a non-Gaussian non linear equation model as follows:

1. In the uncorrelated case, (the means of) the simulated quantities are about the same as the GLS estimates.
2. Skewness of distributions mostly has only a small effect on the estimates.
3. Positive cross-correlations of the variables can lead to severe problems: The equation system may become M-inconsistent with respect to a given data set, i.e. the overlap of the sets of simulated values of at least one variable determined from all equations, where it is part of, may become empty. Under a Gaussian regime with infinite domains and under the independence assumption this effect cannot happen.

Using a GLS approach is relative to MCMC simulation computational cost-effective. But skewness and correlation may lead to quite different estimates and the introduction of robust estimators, like median, improves all estimates. In the case of M-inconsistency it may be necessary to iterate the simulation algorithm several times for satisfying a given balance equation system. Of course, any iteration increases the computational efforts. Furthermore, note that if a given data set is contradictory to the corresponding equation system M-inconsistency is revealed by our MCMC simulation algorithm quite in contrast to the GLS approach used by QUANTOR which assumes a Gaussian regime.

6 References

- John Aitchison. *The Statistical Analysis of Compositional Data*. Kluwer, 1986.
- Adelchi Azzalini and Antonella Capitanio. Statistical Applications of the Multivariate Skew Normal Distribution, *Journal of the Royal Statistical Society. Series B*, 61, 579-602, 1999.
- Adelchi Azzalini and Alessandra Dalla Valle. The Multivariate Skew-Normal Distribution, *Biometrika*, 83, 715-726, 1996.
- Carlo Batini and Monica Scannapieco. *Data Quality Concepts, Methodologies and Techniques*, Springer, 2006.
- Siddhartha Chib. *Handbook of Computational Statistics - Concepts and Methods*, chapter Markov Chain Monte Carlo Technology, pages 71–102. Springer, 2004.
- I. P. Fellegi and D. Holt. A Systematic Approach to Automatic Edit and Imputation, *JASA*, 71, 17-35, 1976.
- W. Keith Hastings. Monte Carlo sampling methods using markov chains and their applications. *Biometrika*, 57:97–109, 1970.
- Veit Köppen and Hans-J. Lenz. Simulation of non-linear stochastic equation systems. In A.N. Pepelyshev, S.M. Ermakov, V.B. Melas, eds., *Proceeding of the Fifth Workshop on Simulation*, pages 373–378, St. Petersburg, Russia, July 2005. NII Chemistry Saint Petersburg University Publishers.
- Hans-J. Lenz and Roland M. Müller. On the solution of fuzzy equation systems. In G. Della Riccia, H-J. Lenz, and R. Kruse, eds., *Computational Intelligence in Data Mining*, CISM Courses and Lectures. Springer, New York, 2000.
- Hans-J. Lenz and Egmar Rödel. Statistical quality control of data. In Peter Gritzmam, Rainer Hettich, Reiner Horst, and Ekkehard Sachs, editors, *16th Symposium on Operations Research*, pages 341–346. Physica Verlag, Heidelberg, 1991.
- Gunar E. Liepins and V.R.R. Uppuluri. *Data Quality Control Theory and Pragmatics*, Marcel Dekker, 1991.
- Beat Schmid, (1979). Bilanzmodelle. Simulationsverfahren zur Verarbeitung unscharfer Teilinformationen, ORL-Bericht No. 40, ORL Institut, ETH Zürich, 1979.
- Adian F. M. Smith and Alan E. Gelfand. Bayesian statistics without tears: A sampling-resampling perspective. *The American Statistician*, 46(2):84–88, may 1992.
- G.Barrie Wetherill and Marion E. Gerson. *Computer Aids to Data Quality Control*, The Statisticians, 36, 598-592, 1987.