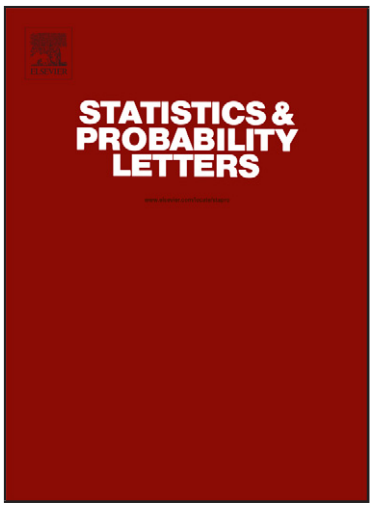
A note on distortions induced by truncation with applications to linear regression systems

Giovanni M. Marchetti, Elena Stanghellini



www.elsevier.com/locate/stapro

A note on distortions induced by truncation with applications to linear regression systems

Giovanni M. Marchetti, Elena Stanghellini

PII: S0167-7152(07)00338-0
DOI: doi:10.1016/j.spl.2007.09.050
Reference: STAPRO 4784

To appear in: *Statistics & Probability Letters*

Received date: 16 April 2007
Revised date: 2 August 2007
Accepted date: 25 September 2007

Cite this article as: Giovanni M. Marchetti and Elena Stanghellini, A note on distortions induced by truncation with applications to linear regression systems, *Statistics & Probability Letters* (2007), doi:10.1016/j.spl.2007.09.050

A note on distortions induced by truncation with applications to linear regression systems

Giovanni M. Marchetti, Dipartimento di Statistica "G. Parenti", 50134 Florence, Italy
Elena Stanghellini, Dipartimento di Economia Finanza e Statistica, 06100 Perugia, Italy*

We explore the effects of truncation on the joint distribution of the observable random variables. A general formula for the distortion induced by truncation in the least-squares coefficients is presented. The implications of our derivations are illustrated with an example.

Keywords: Conditional independence model; directed acyclic graphs; incidental truncation; instrumental variables; selection bias; truncated normal distribution.

1 Introduction

We study the effects induced by truncation in multivariate systems, with particular attention to the distortion induced on linear regression coefficients. Truncation mechanisms may be used to model some forms of selection. As an instance, in econometric studies on income, only observations with income above a threshold may appear in the sample. In biometric studies on the effect of an expensive treatment using health care claims data, there might be worries that only the sickest patients are given the treatment.

It is well-known that, in linear regression modelling, the distribution for any fixed set of units depends on the covariate values, so selection on the basis of covariate values affects the distribution but does not affect the specification of the model. If there is selection on the basis of variables other than the covariates, we may have two well-known situations, such as as censoring or truncation.

In this paper we focus on truncation. This implies that the population consists of all units satisfying the selectivity condition. We do not address issues of estimation, but we explore the effects of truncation on the distribution of the observable variables and give an explicit formula of the distortion induced in the least-squares regression coefficients of interest. We will assume that some knowledge either on the conditional

*Corresponding author. *Email addresses: elena.stanghellini@stat.unipg.it (E. Stanghellini), giovanni.marchetti@ds.unifi.it (G.M. Marchetti).*

independence structure or on zero constraints on partial regression coefficients is available *a priori*. This allows to make connections to graphical models, see e.g. Lauritzen (1996). For a related approach, where different sources of selection bias are illustrated using graphical models, see Hernán *et al.* (2004). The situation here considered forms the basis of the widely studied Heckman's model (Heckman, 1979). There is an extensive literature on the topic of selection bias. A broad account can be found in Copas and Li (1997), with discussion, while a general review on estimation is in and Vella (1998).

2 Some general results on truncation

Let $X = (X_j)$, $j = 1, \ldots, d$, be a $d \times 1$ vector of random variables indexed by the set $V = \{1, \ldots, d\}$. We assume X to have a joint density function f_V. After partitioning $V = N \cup S$, suppose that the variables X_S are truncated, i.e. X_j, $j \in S$, is observed only if it belongs to the interval $\mathcal{I}_j = (a_j, b_j)$ which is a subset of the support. Let $\mathcal{I}_S$ be the Cartesian product of all intervals $\mathcal{I}_j$ for $j \in S$. The density function of X after truncation on X_S is $\tilde{f}_V(x) = \alpha^{-1} f_V(x) \mathbb{I}[x_S \in \mathcal{I}_S]$ where $\mathbb{I}[x_S \in \mathcal{I}_S]$ is the indicator function of the set in square brackets and

$$\alpha = P(X_S \in \mathcal{I}_S). \tag{1}$$

In the following, we make use of three disjoint subsets of V called R, T, and C, which may be interpreted as indexes of responses, treatments and covariates, respectively. We shall denote with F the set of truncated variables outside R, T and C. With the notation $R \perp\!\!\!\perp T | C$ we indicate that the random vectors X_R and X_T are conditionally independent given X_C. We shall look at marginal and conditional densities before and after truncation on S. Using the tilde, we use the convention that truncation occurs always before marginalization or conditioning.

First, notice that since the marginal distribution of X_S after truncation, is $\tilde{f}_S(x_S) = \alpha^{-1} f_S(x_S) \cdot \mathbb{I}[x_S \in \mathcal{I}_S]$, the conditional distribution of $X_N | X_S$ is unaffected by truncation. A more general condition under which the conditional densities before and after truncation are equal is provided by the following proposition.

Proposition 1. *Let F be non empty. If the response variables R are not truncated, i.e. $R \subset N$, then there is no distortion, $\tilde{f}_{R|TC} = f_{R|TC}$, whenever $R \perp\!\!\!\perp F \mid T \cup C$, that is the responses are independent of the truncated variables outside the covariates given the covariates.*

Proof. Assume, without loss of generality, that $V = R \dot\cup T \dot\cup C \dot\cup F$. Then, if $R \perp\!\!\!\perp F \mid T \cup C$, $f_V = f_{RTC} f_{FTC} / f_{TC}$ and thus, $\tilde{f}_V = \alpha^{-1} f_{RTC} f_{FTC} / f_{TC} \cdot \mathbb{I}(x_S \in \mathcal{I}_S)$ with α as in (1). Marginalizing over X_F we get

$$\tilde{f}_{RTC} = f_{RTC}/f_{TC} \left[\frac{1}{\alpha} \int_{\mathcal{I}_F} f_{FTC} \cdot \mathbb{I}(x_S \in \mathcal{I}_S) d_F \right] = f_{RTC}/f_{TC} \cdot g_{TC}.$$

The result follows by noting that g_{TC} is a function not depending on X_R. □

The following Proposition establishes a relationship between the conditional independence structure before and after truncation.

Proposition 2. *Let T, F and C be three disjoint subsets of the node set V such that, before truncation, $T \perp\!\!\!\perp F \mid C$. Then, the conditional independence is preserved if truncation is on $S \subseteq F \cup C \cup T$.*

Proof. By assumption, the marginal density after truncation is

$$\begin{aligned} \tilde{f}_{TFC} &= \int_{\mathcal{I}_U} \alpha^{-1} f_{TFCU} \cdot \mathbb{I}(x_T \in \mathcal{I}_T, x_F \in \mathcal{I}_F, x_C \in \mathcal{I}_C) d_U \\ &= \alpha^{-1} f_{FC} f_{TC} / f_C \cdot \mathbb{I}(x_T \in \mathcal{I}_T) \mathbb{I}(x_F \in \mathcal{I}_F) \mathbb{I}(x_C \in \mathcal{I}_C), \end{aligned}$$

with $U = V \setminus (T \cup C \cup F)$ and α as in (1). Therefore, $\tilde{f}_{TFC}$ can be factorized into $g_{FC} g_{TC}$ and the result follows. □

Corollary 1. *All conditional independencies $T \perp\!\!\!\perp F | C$, such that $V = T \cup F \cup C$, continue to hold after truncation on $S \subseteq V$.*

Corollary 1 has implications for undirected graphical models (see Lauritzen, 1996, Section 3.2), that define a class of distributions that obeys to the pairwise Markov property, such that a conditional independence between X_i and X_j given all the remaining variables holds whenever there is a missing edge in the associated undirected graph. By Corollary 1, the conditional independence graph before truncation matches the one after truncation.

3 The truncated multivariate normal distribution

We show the effects of truncation on the concentration matrix. We assume that X follows a multivariate normal distribution with mean μ and positive definite covariance matrix Σ partitioned as

$$\Sigma = \begin{pmatrix} \Sigma_{SS} & \Sigma_{SN} \\ . & \Sigma_{NN} \end{pmatrix}, \quad \Sigma^{-1} = \begin{pmatrix} \Sigma^{SS} & \Sigma^{SN} \\ . & \Sigma^{NN} \end{pmatrix}$$

where the dot is the usual shortcut for symmetric matrices. In the following, we indicate with $\tilde{\Sigma}$ and $\tilde{\Sigma}^{-1}$ the covariance and concentration matrix of the joint distribution of X after truncation on X_S. Then, the covariance matrix after truncation is, see Johnson and Kotz (1972, p. 70),

$$\tilde{\Sigma} = \begin{pmatrix} \tilde{\Sigma}_{SS} & \tilde{\Sigma}_{SS}\Pi_{N|S}^{\mathrm{T}} \\ . & \Sigma_{NN.S} + \Pi_{N|S}\tilde{\Sigma}_{SS}\Pi_{N|S}^{\mathrm{T}} \end{pmatrix} \tag{2}$$

where $\Pi_{N|S} = \Sigma_{NS}\Sigma_{SS}^{-1} = -(\Sigma^{NN})^{-1}\Sigma^{NS}$, is the matrix of least squares regression coefficients and $\Sigma_{NN.S} = (\Sigma^{NN})^{-1}$ is the partial covariance matrix. The expression for $\tilde{\Sigma}_{SS}$ can be found from the cumulant generating function of the truncated normal, see Tallis (1961) and Finney (1962). For $|S| = 1$ and $b_S = +\infty$, the marginal distribution of X_N after truncation is an extended skew-normal, see Capitanio *et al.* (2003).

Proposition 3. *The concentration matrix of vector X after truncation on X_S is*

$$\tilde{\Sigma}^{-1} = \begin{pmatrix} \tilde{\Sigma}^{SS} & \Sigma^{SN} \\ . & \Sigma^{NN} \end{pmatrix},$$

where $\tilde{\Sigma}^{SS} = \tilde{\Sigma}_{SS}^{-1} + \Pi_{N|S}^{\mathrm{T}}\Sigma^{NN}\Pi_{N|S}$.

Proof. The covariance matrix in (2) can decomposed as

$$\tilde{\Sigma} = \begin{pmatrix} I & 0 \\ \Pi_{N|S} & I \end{pmatrix} \begin{pmatrix} \tilde{\Sigma}_{SS} & 0 \\ 0 & \Sigma_{NN.S} \end{pmatrix} \begin{pmatrix} I & \Pi_{N|S}^{\mathrm{T}} \\ 0 & I \end{pmatrix}.$$

Taking the inverse and multiplying, the result follows. □

This result is valid outside the Gaussian case provided that the conditional distributions have linear regressions and homoscedastic covariance matrices. Notice that the matrices Σ^{-1} and $\tilde{\Sigma}^{-1}$ must share the same sets of structural zeros in blocks (S, N) and (N, N). For i and j in S, if $\sigma^{ij} = 0$ then, from Corollary 1, $X_i \perp\!\!\!\perp X_j | rest$ also after truncation.

4 Distortion in linear regression coefficients induced by truncation

In this section we give an explicit formula for the distortion induced by truncation on linear regression coefficients. Let $\Pi_{R|T.C}$ the partial least-squares regression coefficient

of X_T when regressing X_R on X_T and X_C, before truncation, and let $\tilde{\Pi}_{R|T.C}$ be the same coefficient computed in the distribution resulting after truncation on X_S. Therefore,

$$\Pi_{R|T.C} = \Sigma_{RT.C}\Sigma_{TT.C}^{-1} \text{ and } \tilde{\Pi}_{R|T.C} = \tilde{\Sigma}_{RT.C}\tilde{\Sigma}_{TT.C}^{-1}.$$

We have that $\Pi_{R|TC} = (\Pi_{R|T.C}\ \Pi_{R|C.T})$. We will make use of the matrix extension of Cochran's (1938) recursive formula, Cox and Wermuth (2004),

$$\Pi_{R|T} = \Pi_{R|T.W} + \Pi_{R|W.T}\Pi_{W|T}. \tag{3}$$

Proposition 4. *(i) If the response variables R are not truncated and $S \subseteq T$, then*

$$\tilde{\Pi}_{R|T} = \Pi_{R|T}.$$

(ii) Let $R \subseteq S$ and $T \subseteq N$, then

$$\tilde{\Pi}_{R|T} = \Lambda_{R|T}\Pi_{R|T}$$

where $\Lambda_{R|T} = \tilde{\Sigma}_{RR.T}(\Sigma_{RR.T})^{-1}$.

Proof. Let the marginal covariance matrix of (R, T) be Ω. (i) The regression coefficients after truncation is $\tilde{\Pi}_{R|T} = -(\tilde{\Omega}^{RR})^{-1}\tilde{\Omega}^{RT}$. But, as $S \subseteq T$, from Proposition 3, we have $-(\tilde{\Omega}^{RR})^{-1}\tilde{\Omega}^{RT} = -(\Omega^{RR})^{-1}\Omega^{RT} = \Pi_{R|T}$. (ii) Proposition 3 implies that $\tilde{\Omega}^{RT} = \Omega^{RT}$ only and $\tilde{\Pi}_{R|T} = -(\tilde{\Omega}^{RR})^{-1}\Omega^{RT} = (\tilde{\Omega}^{RR})^{-1}\Omega^{RR}\Pi_{R|T}$. The result follows by noting that $\Omega^{RR} = \Sigma_{RR.T}^{-1}$. □

Proposition 5. *Let R, T and C be three disjoint subsets of V with $R \subseteq N$. Let $F = S \setminus (T \cup C)$.*

(i) Let F be a non-empty set. Then:

$$\tilde{\Pi}_{R|T.C} = \Pi_{R|T.C} - \Pi_{R|F.TC}\{\Pi_{F|T.C} - \tilde{\Pi}_{F|T.C}\}. \tag{4}$$

(ii) Let F be the empty set. Then:

$$\tilde{\Pi}_{R|T} = \Pi_{R|T} - \Pi_{R|C.T}\{\Pi_{C|T} - \tilde{\Pi}_{C|T}\}. \tag{5}$$

Proof. (i) By (3), we have

$$\begin{aligned}
\tilde{\Pi}_{R|T.C} &= \tilde{\Pi}_{R|T.CF} + \tilde{\Pi}_{R|F.TC}\tilde{\Pi}_{F|T.C} \\
&= \Pi_{R|T.CF} + \Pi_{R|F.TC}\tilde{\Pi}_{F|T.C}, \quad \text{by Proposition 4(i)} \\
&= \Pi_{R|T.C} - \Pi_{R|F.TC}\{\Pi_{F|T.C} - \tilde{\Pi}_{F|T.C}\} \quad \text{using again (3).}
\end{aligned}$$

(ii) It follows analogously. □

Corollary 2. *If the response variables R are not truncated, then*

$$\tilde{\Pi}_{R|T.C} = \Pi_{R|T.C} \tag{6}$$

if either (i) $\Pi_{R|F.TC} = 0$ *or (ii)* $\Pi_{F|T.C} = \tilde{\Pi}_{F|T.C} = 0$.

Notice that, if a stronger condition holds, such as $R \perp\!\!\!\perp F | T \cup C$, from Proposition 1, it follows that conditional densities $\tilde{f}_{R|TC}$ and $f_{R|TC}$ are equal. In the linear case, this implies that $\Pi_{R|T.C}$ can be estimated using ordinary least-squares, with a loss of efficiency due to the restrictions on the range of admissible values of X_T and X_C. Equation (4) is the multivariate extension of Golberger's (1981) equation (37) for incidental truncation.

5 Some implications for linear recursive regressions

Sometimes a full ordering of the variables X can be determined such that the joint density of the variables in X can be factorized into a product of univariate densities. In that case, we say that the distribution is generated over a directed acyclic graph, see Lauritzen, 1996, Section 3.2.2. When truncation occurs on more than one variable of the univariate recursive process, then the problem arises on whether truncation on univariate densities in a stepwise fashion is equivalent to truncation on the joint distribution. Provided that truncations on each variable are independent, the two mechanisms lead to the same truncated distribution.

Particular cases of distributions generated over a directed acyclic graph are linear recursive regression systems with independent residuals, where it is assumed that the random variables X are mean-centred such that $AX = \varepsilon$ where A is a unit upper triangular matrix and the errors ε have zero means and are uncorrelated, see Wermuth and Cox (2004). In this framework X may contain latent variables which induce correlated residuals in the equations for the observed variables. We discuss an example illustrating the implications of the previous results in these systems.

Figure 1 about here

Example. The situation known as incidental truncation, Goldberger (1981), can be represented with a linear structural model

$$\begin{aligned} Y_1 &= \beta_{1X} X + \eta_1 \\ Y_2 &= \beta_{2X} X + \eta_2, \end{aligned}$$

where η_1 and η_2 are correlated error terms, marginally independent of X, and the unit are selected according to $Y_2 \in \mathcal{I}_2$. The model is related to Heckman's (1979) model. By adding a latent variable L inducing correlation among η_1 and η_2, an equivalent linear recursive regression model with independent residuals is derived. The associated directed acyclic graph is in Figure 1(a). The interest is in estimating $\Pi_{Y_1|X.L}$ after truncation on Y_2. Since $Y_1 \perp\!\!\!\perp Y_2 | X \cup L$, by Proposition 5

$$\tilde{\Pi}_{Y_1|X.L} = \Pi_{Y_1|X.L} - \Pi_{Y_1|Y_2.XL}(\Pi_{Y_2|X.L} - \tilde{\Pi}_{Y_2|X.L}) = \Pi_{Y_1|X.L}.$$

Since L is usually not known $\Pi_{Y_1|X.L}$, cannot be estimated. Suppose now that X can be partitioned into X_1 and X_2 such that $X_1 \perp\!\!\!\perp Y_2 \mid X_2$ and $X_2 \perp\!\!\!\perp Y_1 | X_1$ as in Figure 1(b). The interest is now on $\Pi_{Y_1|X_1.L}$. By direct calculations, or using the results in Spirtes *et al.* (1998) Section 4.4, we have $\Pi_{Y_1|X_1.L} = \Pi_{Y_1|X_1.X_2Y_2}$. Then, from Proposition 1, $f_{Y_1|X_1.X_2Y_2} = \tilde{f}_{Y_1|X_1.X_2Y_2}$. The implication is that, in the linear case, the least-squares regression coefficient of interest can be estimated from a sample drawn from the truncated distribution as the OLS coefficient of X_1 in the linear regression of Y_1 against X_1,X_2 and Y_2. Notice that the derivations here do not make use of the information on the censoring mechanism induced by the truncation process.

6 Concluding remarks

This note details the effects of truncation with particular references to linear recursive systems. Sufficient conditions for the absence of distortion are given, based on conditional independencies or zero partial correlations before truncation. A formula for the distortion in linear regression coefficient is also provided. When some conditional independencies can be either postulated or induced via an adequate design, then rules are given to check which associations are not distorted after truncation. In the linear case, when the distribution is generated over a directed acyclic graph, this leads to useful conditions to find adjusting covariates that allow the identification of the least-squares coefficients of interest. Furthermore, as the example shows, there might be a gain in the estimation of the least-squares parameters of interest by conditioning on the truncation variables, when measured, even if not explicitly appearing in the equation.

Acknowledgement

We thank Nanny Wermuth and Richard Sheines for helpful comments. The work was partially supported by MIUR, Rome, under the project PRIN 2005132307.

References

Capitanio, A., Azzalini, A. and Stanghellini, E. (2003), Graphical models for skew-normal variates. *Scandinavian Journal of Statistics*, **30**, 129–144.

Cochran, W. G. (1938), The omission or addition of an independent variate in multiple linear regression. *J. R. Statist. Soc.*, suppl., 5, 171-176.

Copas, J. B. and Li, H. G. (1997), Inference for non-random samples. *J. of the Royal Statist. Soc.*, B, **59**, 55–95.

Finney, D.J. (1962), Cumulants of truncated multinormal distributions. *J. R. Statist. Soc.*, B, **24**, 535–536.

Goldberger, A.S. (1981), Linear regression after selection. *Journal of Econometrics*, **15**, 357–366.

Heckman, J. J. (1979), Sample bias as a specification error. *Econometrica*, **47**, 153–161.

Hernán, M., A., Hernández-Dìaz, S. and Robins, J., M. (2004), A structural approach to selection bias. *Epidemiology*, **5**, 615–625.

Johnson, N. L. and Kotz, S. (1972), *Distributions in Statistics: Continuous Multivariate Distributions*. New York: Wiley.

Lauritzen, S. L. (1996), *Graphical Models*. Oxford: Oxford University Press.

Spirtes, P., Richardson, P., Meek, C., Scheines, R. and Glymour, C. (1998), Using path diagrams as a structural equation modeling tool. *Sociological Methods Research*, **27**, 182–225.

Tallis, G.M. (1961), The moment generating function of the truncated multi-normal distribution. *J. R. Statist. Soc.*, B, **23**, 223–229.

Vella, F. (1998), Estimating models with sample selection bias: a survey. *The Journal of Human Resources*, **1**, 127–169.

Wermuth, N. and Cox, D.R. (2004), Joint response graphs and separation induced by triangular systems. *J. R. Statist. Soc.*, B **66**, 686–717.

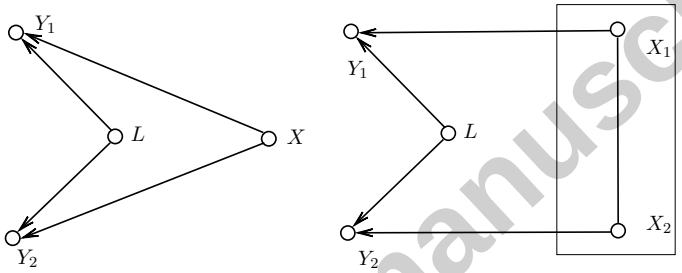


Figure 1: *Two conditional independence graphs for the incidental truncation problem (a) without and (b) with exclusion restrictions.*