Augustin:

Some Basic Results on the Extension of Quasi-Likelihood Based Measurement Error Correction to Multivariate and Flexible Structural Models

Sonderforschungsbereich 386, Paper 196 (2000)

Some Basic Results on the Extension of Quasi-Likelihood Based Measurement Error Correction to Multivariate and Flexible Structural Models

Th. Augustin

Seminar für Ökonometrie und Statistik

Universität München

D-80799 München, Germany

Abstract

Quasi-score equations derived from corrected mean and variance functions allow for consistent parameter estimation under measurement error. However, the practical use of some approaches relying on this general methodological principle was strongly limited by the assumptions underlying them: only one covariate was allowed to be measured with non-negligible error, and, additionally, this covariate had to be conditionally independent of the other covariates. This paper extends basic principles of this method to multivariate and flexible models in a way that, on the one hand, retains the neat statistical properties, but on the other hand, manages to do without the restrictive assumptions needed up to now.

Keywords: Measurement error, error-in-variables, quasi-likelihood, mixtures of normals

1 Introduction

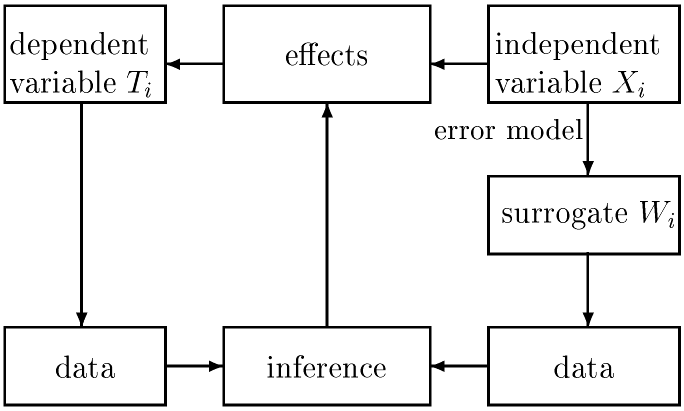
A typical problem in regression analysis is the presence of covariate measurement error. Often there are covariates X ('latent variables') of particular interest, which cannot be directly observed or measured correctly. However, if one ignores the measurement error by just plugging in substitutes or incorrect measurements W instead of X ('naive estimation'), then all the parameter estimates must be suspected to be severely biased. *Error-in-variables modeling* provides a methodology, which is serious about that fact and develops procedures to adjust for the measurement error. For the linear model many basic results had already been achieved until the eighties. They are summarized e.g. in the books by Schneeweiß, Mittag (1986) and by Fuller (1987). Recent developments in that area are covered by Cheng, van Ness (1999), while Caroll et. al. (1995) present the state of the art in nonlinear models up to the middle of the nineties.[1]

One general and powerful methodological principle to deal with measurement error is quasi-likelihood based measurement error correction: corrected mean and variance functions can be used to construct a measurement error corrected quasi-score equation, which produces consistent parameter estimates. In particular this idea underlies the work of Armstrong (1985), Liang, Lu (1991), Caroll et. al. (1995, Section 7.8 and Appendix A.4), and also the papers of Thamerus (1998A, 1998B) and Augustin (2000), which are closest to the development here.

The present paper discusses basic ingredients of this method in an extended context which does not suffer from severe restrictions inherent to some former approaches. Section 2 recalls a few essentials around the problem of measurement error and then states the model used throughout the paper. Special attention is paid to the question how to model the distribution of the unknown variables with sufficient flexibility. Section 3 is devoted to measurement error corrected quasi-likelihood estimation and demonstrates how the requirements this technique needs can be satisfied by the model introduced.

[1] According to the literature the term 'measurement error' is only applied to continuous variables. The corresponding problem for discrete variables ('misclassification') is not addressed here. This paper will also concentrate on covariate measurement error by assuming that the dependent variables are measured without error.

Figure 1: regression under covariate measurement error:



2 Measurement Error

2.1 Some Basic Considerations

Measurement error occurs in very different areas of application: often for all (or some of the) units $i = 1, \dots, n$ variables X_i of primary interest are not observable. Instead one has to be satisfied with so called surrogates W_i, i.e. with somehow related, but different variables. For instance, in physics or medical science these surrogates are typically inexact measurements of X_i. In sociology or psychology measurement error naturally arises by the insufficiency of operationalizations of complex theoretic constructs.

As symbolized in Figure 1, the problem caused by measurement error is that one is interested in estimating effects of the variable X_i, while the data are realizations of a different variable W_i. In estimating regression parameters, however, this difference has to be taken into account: neglecting it by just plugging in W_i instead of X_i in the estimating procedures will typically lead to estimates with a considerable bias.

The theory of *measurement error correction* or *error-in-variables modeling* provides a framework which aims at deriving nevertheless consistent parameter estimates. It develops procedures to make sound conclusions from realizations of $W_1, \dots, W_n$ on the effects of $X_1, \dots, X_n$. As is also suggested by Figure 1, this can only be possible if one takes some relationship between the Xs and the Ws into account. In the case of validation data, i.e. simultaneous observation of the Ws and the Xs in a sub-sample, this relationship can be estimated from the data. Otherwise, one has to model it as flexible as possible. Here the following flexible model is used. (i=1,...,n)

2.2 The error model I

- Assume all covariates X_i to be continuous.
- Additive measurement error:[2] $W_i = X_i + U_i$.
- U_i is independent of T_i, X_i and U_j, $j \neq i$.
- Normal measurement error: $U_i \sim \mathcal{N}(0, \Sigma_U)$ with Σ_U known.
- Structural model: X_i is stochastic. $X_1, \ldots, X_n$ are independently and identically distributed.

These assumptions imply that the measurement error is *nondifferential*: T_i and W_i are conditionally independent given X_i, i.e. W_i possess no information with respect to T_i which is not contained in X_i. So, knowing X_i would make knowledge of W_i superfluous.

2.3 The Error Model II – the Distribution of X

In addition to the assumptions listed above, an appropriate class of parametric distributions for X has to be chosen. For sake of mathematical convenience there is a strong temptation to take a normal distribution as the distribution law P_X of X. Then, by additivity of normally distributed random variables, also the W_is would be normal. However, in many applications the empirical marginal distributions of W are heavily skewed and/or possess several modes, which makes the assumption of normality for P_X rather questionable.

To account for multi-modality and skewness turning to *mixtures of normals* proves to be successful. The main idea is to allow for heterogeneity: one takes the population to be divided into m different groups, where in principle m need not be known a priori. Conditional on being in group j now normality is assumed with group specific parameters: $X_i \sim \mathcal{N}(\mu_j, \Sigma_j)$. With

[2]Note that this formulation also covers the case of correctly measured components of the vector of covariates. If $X_i[j]$ is correctly measured then one puts $U_i[j] \equiv 0$.

κ_j as the unknown probability to belong to group j the overall distribution is a so-called *mixture of normals* or *mixed normal distribution*

$$X_i \sim \mathcal{MIXN}\left(m;\, \kappa_1, \ldots, \kappa_m;\, \mu_1, \ldots, \mu_m;\, \Sigma_1, \ldots, \Sigma_m\right). \tag{1}$$

This models is highly flexible[3], but will nevertheless prove to be sufficiently tractable from the mathematical point of view.

3 Quasi-Likelihood Based Correction for Covariate Measurement Error

3.1 A Look on Previous Work

As already discussed and also illustrated in Figure 1 the parameter estimation has to take into account that the data are not realizations of the variables of interest but are steaming from surrogate variables. So the likelihood relevant for parameter estimation is the so-to-say *data-based likelihood*, i.e. the likelihood $\mathrm{Lik}(\theta \| T_i, W_i)$ of the unknown parameter vector θ given $W_1, \ldots, W_m$. For many models of interest it is however not manageable to calculate this expression from the *ideal likelihood*, i.e. the likelihood $\mathrm{Lik}(\theta \| T_i, X_i)$ derived from the regression model formulated in terms of the unobservable quantities $X_1, \ldots, X_n$.

Then one is forced to search for another general estimation principle. Here a successful choice will be quasi-score estimation based on mean and variance functions. The basic ideas of this approach were introduced in Wedderburn (1974) and developed further especially by McCullagh (1983, 1991). In the meanwhile they are embedded into the considerably extended framework of general estimation functions (see Heyde (1997) for a comprehensive monograph on this topic).

The quasi-score function which will prove to be successful in the context considered here uses the data-based means $\mathbb{E}\left[T_i|W_i;\theta\right]$ and (co)variances $\mathbb{V}\left[T_i|W_i;\theta\right]$. In contrast to the full data-based likelihood these quantities will prove to be obtainable from the ideal model formulated in terms of the

[3]See, for instance, Everitt & Hand (1981, p. 28f.), who give an impression of the quite different shapes which can be produced by even only the mixture of two normals.

unobservable variables. The resulting quasi-score equation reads as

$$\sum_{i=1}^{n} \frac{\partial \, I\!E\,[T_i|W_i;\theta]}{\partial \, \theta} \cdot I\!V\,[T_i|W_i;\theta]^{-1} \cdot \{T_i - I\!E\,[T_i|W_i;\theta]\} \;=\; 0\,. \qquad (2)$$

To the author's knowledge Armstrong (1985) was the first to recognize the power this principle possesses for measurement error correction. Also Caroll et. al. (1995; Section 7.8 and Appendix A.4) briefly mention the importance of this idea.

Thamerus (1998A, 1998B) and Augustin (2000) worked with simpler versions of the model used here letting some of the main aspects of the arguments given below already shine up. For modeling the distribution of the latent variable, Thamerus (1998B) and Augustin (2000) do only allow for a single normal distribution, but not for mixtures. Even more important, all three papers just quoted had to concentrate on the case where only one dimension, $X_i[1]$ say, of the covariate vector is measured with error. This assumption may not only be unrealistic in many empirical situations, but it is also responsible for an additional requirement which may be even more tricky: to enable the calculation of measurement error corrected mean and covariance functions along the lines below, the conditional distribution of $X_i[1]$ given the surrogate $W_i[1]$ and other dimensions $X_i[2], X_i[3], \ldots$ of the vector of covariates de facto has to be independent of $X_i[2], X_i[3], \ldots$.

3.2 The Main Idea

The central observation of quasi-likelihood based measurement error correction is that, via the theorem of iterated expectation and the nondifferentiality of the measurement error, the conditional moments $I\!E\,(T_i^r|W_i;\theta)$ with respect to the observable quantities can be derived from their counterparts $I\!E\,(T_i^r|X_i;\theta)$ based on the unobservable quantities:

$$\begin{aligned} I\!E\,[T_i^r|W_i;\theta] &= I\!E\left(I\!E\,[T_i^r|X_i,W_i;\theta]\,\Big|W_i;\theta\right) \\ &= I\!E\Big(\underbrace{I\!E\,[T_i^r|X_i;\theta]}_{\text{ideal model}}\Big|\;\underbrace{W_i}_{\text{observable}}\;;\theta\Big) \end{aligned} \qquad (3)$$

Similar arguments hold for the covariance matrix $I\!V\,(T_i|W_i;\theta)$.

Relation (3) is very helpful for calculating the corrected mean and variance functions. It separates the problem into two distinct steps:

- Firstly, determine the 'ideal moments' of first and second order of the ideal model.
- Secondly, integrate over these moments with respect to the conditional distribution of X_i given W_i.

The first step is an easy exercise for most models.[4]

The second step is prepared by the following proposition applying some basic properties of mixtures of normals in the context under consideration.

Proposition. Let

$$X_i \sim \mathcal{MIXN}\left(m;\, \kappa_1, \ldots, \kappa_m; \mu_1, \ldots, \mu_m;\, \Sigma_1, \ldots, \Sigma_m\right)\,,$$

and denote the density of the $j-th$s component by $\varphi(\cdot\,\|\mu_j, \Sigma_j)$. Furthermore, let $U_i \sim \mathcal{N}(0,\, \Sigma_U)$, and U_i be independent of X_i. Define $W_i := X_i + U_i$. Then

$$\text{a)} \qquad W_i \sim \mathcal{MIXN}\big(m;\, \kappa_1, \ldots, \kappa_m; \mu_1, \ldots, \mu_m;\, \Sigma_1 + \Sigma_U, \ldots, \Sigma_m + \Sigma_U\big)$$

$$\text{b)} \quad X_i \mid W_i \sim \mathcal{MIXN}\big(m;\, \bar{\kappa}_{i,1}, \ldots, \bar{\kappa}_{i,m}; \bar{\mu}_{i,1}, \ldots, \bar{\mu}_{i,m};\, \bar{\Sigma}_1, \ldots, \bar{\Sigma}_m\big)$$

with $(j = 1, \ldots, m)$

$$\begin{aligned}
\bar{\kappa}_{i,j} &= \frac{\kappa_j \cdot \varphi(W_i\,\|\mu_j, \Sigma_j + \Sigma_U)}{\sum_{l=1}^{m} \kappa_l \cdot \varphi(W_i\,\|\mu_l, \Sigma_l + \Sigma_U)} \\
\bar{\mu}_{i,j} &= \mu_j + \Sigma_j \cdot (\Sigma_j + \Sigma_U)^{-1} \cdot (W_i - \mu_j) \\
\bar{\Sigma}_j &= \Sigma_j - \Sigma_j \cdot (\Sigma_j + \Sigma_U)^{-1}\, \Sigma_j\,.
\end{aligned}$$

According to Part a) of this proposition, $W_1, \ldots, W_n$ follow a mixture of normals with the same set of unknown parameters as $X_1, \ldots, X_n$. Therefore, these unknown nuisance parameters can be estimated from the observable quantities $W_1, \ldots, W_n$ by any algorithm suitable for parameter estimation under mixtures of multivariate normals. These estimates can then be plugged in, and one obtains the conditional distribution of X_i given W_i along the lines of Part b).

[4] One interesting exception is the case of censored survival times, see Augustin (2000, Section 6) for details.

Another basic result from the theory of mixture distributions states that an expectation with respect to a mixture is just a weighted average of the expectations with respect to the single components. Therefore, the final integration with respect to the conditional distribution of X_i given W_i consists only of the evaluation of m integrals with respect to multivariate normals and their summing up weighted by the $\bar{\kappa}_j$s derived in Part b) of this proposition.

Solving the corresponding quasi-score equation (2) yields the measurement error corrected quasi-likelihood estimates. Under quite mild regularity conditions they can be shown to have appealing asymptotic properties. In particular they are consistent: the bias caused by the measurement error is eliminated.

4 Concluding Remarks

Though he was never concerned with measurement error modeling, this field provides a vivid example supporting McCullagh's (1991, p. 265) claim that "[...via quasi-likelihood] useful inferences are possible even in problems for which a full likelihood-based analysis is either intractable or impossible with the given assumptions". Quasi-likelihood provides an easy to handle tool for measurement error correction; at least in the extended version presented here, it promises to be widely applicable in many different models.

An area where the quasi-likelihood approach is particularly elegant is the case of parametric survival models without censoring. The concept of accelerated failure time models can serve as a superstructure which enables one to handle the commonly used models in a unified way. The approach can be extended to cover also the case of measurement error in the dependent variables, i.e. in the lifetimes themselves. The arguments given in Augustin (2000) carry over to the extended situation studied here.

Acknowledgement I am grateful to Helmut Küchenhoff and Hans Schneeweiß for helpful discussions and comments.

References

ARMSTRONG, B. (1985): Measurement error in the generalized linear model. *Communications in Statistics, Part B – Simulation and Computation, 14, 529–544.*

AUGUSTIN, T. (2000): Correcting for measurement error in parametric duration models by quasi-likelihood. Accepted for publication in *Biometrical Journal.*

CAROLL, R. J., RUPPERT, D. and STEFFANSKI, L. A. (1995): Measurement Error in Nonlinear Models. Chapman and Hall, London.

CHENG, C.-L. and VAN NESS, J.W (1999): Statistical Regression with Measurement Error. Arnold, London.

EVERITT, B. S. and HAND, D. J. (1981): Finite Mixture Distributions. Chapman and Hall, London.

FULLER, W. A. (1987): Measurement Error Models. Wiley, New York.

HEYDE, C. F. (1997): Quasi-Likelihood and its Application. A General Approach to Parameter Estimation. Springer, New York.

LIANG, K.-Y. and LIU, X.-H. (1991): Estimating equations in generalized linear models with measurement error. In: Godambe, V. P. (Ed.): Estimating Functions. Clarendon Press, Oxford, 47–63.

McCULLAGH, P. (1983): Quasi-likelihood functions. *The Annals of Statistics, 11, 59–67.*

McCULLAGH, P. (1991): Quasi-likelihood and estimating functions. In: Hinkley, D. V., Reid, N. and Snell, E. J. (Eds.): Statistical Theory and Modelling. Chapman and Hall, London, 265–286.

SCHNEEWEISS, H. and MITTAG, H. J. (1986): Lineare Modelle mit fehlerbehafteten Daten. Physica, Heidelberg.

THAMERUS, M. (1998A): Nichtlineare Regressionsmodelle mit heteroskedastischen Meßfehlern. Logos, Berlin.

THAMERUS, M. (1998B): Different nonlinear regression models with incorrectly observed covariates. In: Galata, R. and Küchenhoff, H. (Eds.): Econometrics in Theory and Practice. Physika. Heidelberg.

WEDDERBURN, R. W. M. (1974): Quasi-likelihood functions, generalised linear models and the Gauss-Newton method. *Biometrika, 61, 439–447.*