Covariate measurement error in generalized linear models for longitudinal data: a latent Markov approach

Roberto Di Mari*[1], Antonio Punzo[1], and Antonello Maruotti[2,3]

[1]Department of Economics and Business, University of Catania, Italy
[2]Department of Law, Economics, Politics and Modern languages, LUMSA, Italy
[3]Centre for Innovation and Leadership in Health Sciences, University of Southampton, UK

Abstract

One common approach to handle covariate measurement error in Generalized Linear Models (GLM) is classical error modeling. In the past 20 years, classical error modeling has been brought to (Non-Parametric) Maximum Likelihood (NPML) estimation, by means of finite mixture modeling: the supposedly continuous true score is modeled as a discrete (multinomial) static latent variable, and is handled as a part of the model. Nonetheless, the true score is not allowed to vary over time: if the true score has own underlying dynamics, these are either unaccounted for or mistaken for measurement error, or possibly both. The aim of the present paper is to formulate a joint model for the outcome variable, the covariate observed with error (measurement model), and the true score model that accounts for the underlying dynamics in the true score. The true score and its dynamics are modeled non-parametrically as a first-order latent (hidden) Markov chain. Estimation is done extending the NPML approach, in a full maximum likelihood environment with a well-know modification of the EM algorithm (forward-backward algorithm). From an applied researcher perspective, our methodology can safely handle both the case where the latent underlying characteristic is stable over time, as well as providing a suitable framework even when changes across measurement occasions are substantial. Within a GLM framework, it is demonstrated, by means of extensive simulation studies, that this is crucial to get correct estimates of the regression coefficients, as well as good coverages. In the real-data application, the effect of heart rate on the occurrence of cardiovascular diseases in a sample of Chinese elderly patients is measured. Modeling the true (unobserved) heart rate and its dynamics - which, in elderly patients, are likely to be non negligible - will be showed to allow a correct assessment of risk factors of cardiovascular diseases occurrence.

Key-Words: covariate measurement error, errors–in–variables, longitudinal data, generalized linear models, latent Markov models.

*roberto.dimari@unict.it

1 Introduction

In all areas of scientific research, being able to collect high quality measures in order to assess a given phenomenon of interest is crucial. Unaccounted measurement error due to poor measures can severely distort the analysis, leading to unreasonable substantial conclusions.

A relevant type of measurement error is covariate measurement error - or errors in variables. Covariate measurement error issues have long history in epidemiological studies. For instance, Cotton et al. (2005) show how measurement error can affect diagnosis of developmental dyslexia in children with reading difficulties. Kipnis et al. (2003) find that dietary intake assessed with error in a biomarker study produces a severely attenuated estimate of disease relative risk. Guo & Little (2011) report, in pre-menopause women, a negative effect of carotenoids on progesterone, estimated to be zero if measurement error is not accounted for.

The two most common approaches for covariate measurement error modeling are classical error models - known also as error calibration models - and regression calibration models (for an extensive review, see Carroll et al., 2006). In the past 20 years, Aitkin (1996, 1999) and Aitkin & Rocci (2002) among other, have provided a way to bring classical error modeling to maximum likelihood estimation, allowing the user to make no parametric assumption on the true score. That is, nonparametric maximum likelihood (NPML) handles the distribution of the true score as a part of the model non-parametrically, by using finite mixture models (for a recent review, see for instance Alfó & Viviani, 2016. Whereby remaining in a fully ML setup, this approach, in many practical situations, can be more convenient than assuming normality of the true score - as is commonly done in the regression calibration literature. Nonetheless, the true score is not allowed to vary over time. If the true character has own underlying dynamics, these are either unaccounted for or mistaken for measurement error (Alwin, 2007), or possibly both.

Separating unreliability from change in the true score is possible if the true score dynamics are modeled. Quasi Simplex models, or Quasi Autoregressive/Markov Simplex models (Alwin, 2007) are used in survey methods literature to address the issue of changes in the true-score distribution (see, for instance, Uhrig & Watson, 2017). The model is fitted in a confirmatory factor-analytic Structural Equation Modeling (SEM) framework, by assuming a continuous dynamic latent variable with Gaussian error (one-factor model). A similar approach in epidemiology can be found in Sánchez et al. (2009), who study the effects of in-utero lead exposure on child development.

The aim of this work is to formulate a joint model for the outcome variable, the covariate observed with error (measurement model), and the true score model that accounts for the underlying dynamics in the true score. The true score and its dynamics are modeled non-parametrically as a first-order latent (hidden) Markov chain (Bartolucci et al., 2012; Collins & Lanza, 2010; Rabe-Hesketh & Skrondal, 2008; J. K. Vermunt et al., 1999; Wiggins, 1973; Zucchini et al., 2016). Model estimation is done in a fully maximum likelihood environment, with a well-know modification of the EM algorithm (Baum et al., 1970; Welch, 2003).

Our approach is closely related to Aitkin & Rocci (2002)'s, in that we make no distributional assumption on the true score, whereby the key novelty is in modeling the true-score

dynamics. From an applied researcher perspective, our methodology can safely handle both the case where the latent underlying characteristic is stable over time, as well as providing a suitable framework even when changes across measurement occasions are substantial. Within a generalized linear modeling (GLM) framework, we demonstrate that this is crucial to get correct estimates of the regression coefficients, as well as good coverages. Although confirmatory factor-analytic/SEM methodologies allow for dynamics in the true score, estimation relies on identifying restrictions (for instance on the true score variance) and distributional assumptions (normality of each regression errors), which might be restrictive in certain practical situations. In the methodology we propose, we need no identifying restrictions, and we handle non-parametrically a possibly continuous underlying latent variable, modeling it as a dynamic discrete trait (Catania & Di Mari, 2018; Di Mari & Bakk, 2017; J. Vermunt & Magidson, 2004).
We illustrate the proposed methodology by analyzing data from the Chinese Longitudinal Healthy Longevity Survey, where a sample of n Chinese old patients is observed T times, and information on cardiovascular diseases for each person is reported alongside demographics and well-known risk factors, among which heart rate. The aim is to measure the effect of heart rate on the occurrence of a cardiovascular disease, controlling for demographic characteristics and dietary habits.
The structure of the paper is as follows. In Section 2 we will give details on the modeling specification, and describe how the model parameters can be estimated with our latent Markov approach (Section 2. In Section 4 we will summarize the results from the simulation study and the empirical application.

2 Outcome, measurement and error components in common error correction modeling

Let Y_t, W_t and $\mathbf{Z}_t$ be respectively the outcome variable, a continuous covariate corresponding to the true score X_t and a k-vector of error–free covariates, for $t = 0, \dots, T$. In addition, we let $\mathbf{Y}$ be the full vector of outcomes, $\mathbf{Z}$ the full set of available covariates, and $\mathbf{W}$ the full vector of covariate values with corresponding true scores $\mathbf{X}$, observed for the $T + 1$ time occasions. We assume that, given the true score, Y_t and W_t are conditionally independent - non–differential measurement error model.
As it is typical in GLM context, we assume Y_t to have distribution belonging to the exponential family, with the following linear predictor

$$\eta_t(\boldsymbol{\theta}) = \alpha + \beta\,\mathrm{X}_t + \boldsymbol{\gamma}'\,\mathbf{Z}_t, \tag{1}$$

where $\eta_t(.)$ is an appropriate link function, $\boldsymbol{\gamma}$ and β are respectively k-vector and scalar regression coefficients, α is an intercept term and $\boldsymbol{\theta} = \{\alpha, \beta, \boldsymbol{\gamma}\}$.
Equation (1) defines the outcome model in terms of its linear predictor.
As for the classical measurement error model, we assume

$$\mathrm{W}_t = \mathrm{X}_t + \xi_t, \tag{2}$$

where $\xi_t \sim N(0, \sigma^2_{\mathrm{W}})$. The classical additive model can be also applied to variables transformed in log-scale, in order to model multiplicative rather than additive error.

The true score and its dynamics can be modeled extending the usual assumption of (conditional) normality of the true score (given the exogenous covariates $\mathbf{Z}_t$; Aitkin & Rocci, 2002), by assuming the score follows an AR(1) process, such that

$$\begin{aligned} \mathrm{X}_0 &= \epsilon_0 + \boldsymbol{\lambda}' \mathbf{Z}_0, \\ \mathrm{X}_t &= \mathrm{X}_{t-1}\, \rho + \boldsymbol{\lambda}' \mathbf{Z}_t + \epsilon_t. \end{aligned} \tag{3}$$

For convenience, we define the transformed true score $\mathrm{X}_0^* = \mathrm{X}_0 - \boldsymbol{\lambda}' \mathbf{Z}_0$, $\mathrm{X}_t^* = \mathrm{X}_t - \boldsymbol{\lambda}' \mathbf{Z}_t$, and $\boldsymbol{\gamma}^* = \boldsymbol{\gamma} + \beta\lambda$. By dropping the stars, the measurement model is now transformed as

$$\mathrm{W}_t = \mathrm{X}_t + \boldsymbol{\lambda}' \mathbf{Z}_t + \xi_t, \tag{4}$$

and the linear predictor of the outcome model becomes

$$\eta_t(\boldsymbol{\theta}) = \alpha + \beta\, \mathrm{X}_t + \boldsymbol{\gamma}' \mathbf{Z}_t\,. \tag{5}$$

We can now express, using Aitkin & Rocci (2002)'s notation, the following joint model for $(\mathrm{Y}_t, \mathrm{W}_t, \mathrm{X}_t \,|\, \mathbf{Z}_t)$

$$P(\mathrm{Y}_t, \mathrm{W}_t, \mathrm{X}_t \,|\, \mathbf{Z}_t) = P(\mathrm{Y}_t \,|\, \mathrm{X}_t, \mathbf{Z}_t) m(W_t | \mathrm{X}_t, \mathbf{Z}_t) \pi(X_t), \tag{6}$$

where $P(\mathrm{Y}_t \,|\, \mathrm{X}_t, \mathbf{Z}_t)$ is the density or pmf of the outcome, $m(W_t | \mathrm{X}_t, \mathbf{Z}_t)$ is the measurement model density, and $\pi(X_t)$ is the true score density.
We can now define the following joint marginal distribution for (Y_t, W_t)

$$P(\mathrm{Y}_t, \mathrm{W}_t, \mathrm{X}_t \,|\, \mathbf{Z}_t) = \int P(\mathrm{Y}_t \,|\, \mathrm{X}_t) m(W_t | \mathrm{X}_t) \pi(X_t) dX_t, \tag{7}$$

Let $\{(\mathrm{Y}_{it}, \mathrm{W}_{it}, \mathbf{Z}_{it})\}_n = \{(\mathrm{Y}_{1t}, \mathrm{W}_{1t}, \mathbf{Z}_{1t}), \ldots, (\mathrm{Y}_{nt}, \mathrm{W}_{nt}, \mathbf{Z}_{nt})\}$ be a sample of n independent observations, observed for $t = 0, \ldots, T$ time points. The sample log-likelihood - corresponding to the model of Equation (7) for $t = 0, \ldots, T$ - can be defined as

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log \left\{ \int P(\mathbf{Y}_i \,|\, \mathbf{X}_i, \mathbf{Z}_i) m(\mathbf{W}_i \,|\, \mathbf{X}_i, \mathbf{Z}_i) \pi(\mathbf{X}_i) d\, \mathbf{X}_i \right\}, \tag{8}$$

where

$$P(\mathbf{Y}_i, | \, \mathbf{X}_i, \mathbf{Z}_i) m(\mathbf{W}_i \,|\, \mathbf{X}_i, \mathbf{Z}_i) = \prod_{t=0}^{T} P(\mathrm{Y}_{it} \,|\, \mathrm{X}_{it}, \mathbf{Z}_{it}) m(W_{it} | \mathrm{X}_{it}, \mathbf{Z}_{it}), \tag{9}$$

due to local independence of the distributions of outcome and the covariate measured with error across time given the true score. The assumption on the distribution of X_t involves also assumptions on its dynamics. With a relatively simple AR(1) specification, estimating the model parameters by maximixing the (log) likelihood of Equation (8) requires evaluating an integral over a $(T+1)$-dimensional space. This can be done by using a (nonlinear) filtering algorithm, known in the time series literature (Heiss, 2008), which is based on the sequential application of Gaussian quadrature rules (see also Bartolucci et al., 2014).

3 Handling covariate measurement error with latent Markov modeling

The idea underlying the NPML approach (Laird, 1978; Lindsay, 1983a,b) is that finding the MLE of $\pi(\cdot)$, say $\widehat{\pi}(\cdot)$, involves a standard convex optimization problem and, as long as the model likelihood is bounded, $\widehat{\pi}(\cdot)$ is concentrated over at most as many support points as the number of sample units, and is uniquely identified by locations and related masses. We let S, with $S \leq n$, be the state space of the concentrated distribution at time t, X_{it}^{s} be the realized discretized true score for the i-th observation at time t corresponding to the s-th location x_s, with time–varying mass π_{st}, for $s = 1, \ldots, S$, at time t. We propose to model the time–varying masses by using the properties of first–order homogeneous Markov chains. In particular, by letting $\boldsymbol{\delta} = \{\delta_s\}_S$ be the common initial probabilities, where $\delta_s = P(\mathrm{X}_{i0}^{s} = s)$, and $\mathbf{Q}$ the common transition matrix, with elements $\{\mathrm{q}_{rs}\}$, where $\mathrm{q}_{rs} = P(\mathrm{X}_{t}^{s} = \mathrm{x}_s \,|\, \mathrm{X}_{t-1}^{s} = \mathrm{x}_r)$ with $1 < s \leq S$, and $1 < r \leq S$, we can approximate the log likelihood function of Equation (8) as follows

$$\ell(\boldsymbol{\theta}) \approx \sum_{i=1}^{n} \log \left\{ \prod_{t=0}^{T} \sum_{s=1}^{S} P(\mathrm{Y}_{it} \,|\, \mathrm{X}_{it}^{s}, \mathbf{Z}_{it}) m(\mathrm{W}_{it} \,|\, \mathrm{X}_{it}^{s}, \mathbf{Z}_{it}) \pi_{st} \right\}, \tag{10}$$

where, thanks to the properties of Markov chains, $\pi_{s0} = \delta_s$, and π_{st} is the s-th element of the vector $\boldsymbol{\pi}_t = \boldsymbol{\delta}' * \mathbf{Q}^t$.

The elements of the initial state probabilities and the transition probabilities can be parametrized according to logistic parametrizations as follows

$$\log \frac{P(X_0 = s)}{P(X_0 = 1)} = \beta_{s0}, \tag{11}$$

with $1 < s \leq S$, for the initial state probability, and

$$\log \frac{P(\mathrm{X}_t = s | X_{t-1} = r)}{P(\mathrm{X}_t = 1 | X_{t-1} = r)} = \gamma_{0s} + \gamma_{0rs}, \tag{12}$$

with $1 < s \leq S$, and $1 < r \leq S$ for the transitions probabilities. We take the first category as reference - setting to zero the related parameters. For the transition model, this means that parameters related to the elements in the first row and column of the transition matrix are set to zero.

Iterative procedures, like the EM algorithm (Dempster et al., 1977) can be used to maximize Equation (10) in order to estimate the model parameters in one step. However, when using the standard EM, the time and storage required for parameter estimation of latent Markov models increase exponentially with the number of time points (Vermunt, Langeheine, & Böckenholt, 1999). For this reason, the forward–backward algorithm (Baum et al., 1970; Welch, 2003) is typically implemented: this is a special version of the standard EM in which the size of the problem increases only linearly with the number of time occasions (Zucchini et al., 2016).

4 Results and conclusions

We have assessed the proposed latent Markov approach for parameters estimation of generalized linear models with covariate(s) measured with error under a broad set of scenarios, resulting from combinations of different sample size (100, 500 and 1000 for each time point, with $T = 5$), measurement error size ($\sigma^2_{\mathrm{W}} = (1, 1.5, 2)$), and size of the effect of the true score on the outcome $\beta = (1, 1.5)$, for both continuous and dichotomous outcome variables. We have generated the continuous true score from a continuous dynamic model. We found that our latent Markov approach with a number of states between 3 and 5 is enough to approximate the continuous underlying distribution of the true score. The results have showed that the proposed method yields correct parameter estimates in all conditions, except for small sample size (100 observations), as well as good coverages.

In the empirical application on the Chinese Longitudinal Healthy Longevity Survey data, we were able to find, modeling the true (unobserved) heart rate and its dynamics, risk factors for the cardiovascular disease occurrence consistent with the medical literature.

References

Aitkin, M. (1996). A general maximum likelihood analysis of overdispersion in generalized linear models. *Statistics and Computing*, *6*(3), 251–262.

Aitkin, M. (1999). A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics*, *55*(1), 117–128.

Aitkin, M., & Rocci, R. (2002). A general maximum likelihood analysis of measurement error in generalized linear models. *Statistics and Computing*, *12*(2), 163–174.

Alfó, M., & Viviani, S. (2016). Finite mixtures of structured models. In H. C, M. Meila, F. Murtagh, & R. Rocci (Eds.), *Handbook of Cluster Analysis* (pp. 217–240). Chapman & Hall: Boca Raton, FL.

Alwin, D. F. (2007). *Margins of Error: A Study of Reliability in Survey Measurement* (Vol. 547). John Wiley & Sons.

Bartolucci, F., Bacci, S., & Pennoni, F. (2014). Longitudinal analysis of self-reported health status by mixture latent auto-regressive models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *63*(2), 267–288.

Bartolucci, F., Farcomeni, A., & Pennoni, F. (2012). *Latent markov models for longitudinal data.* Chapman and Hall / CRC Press.

Baum, L. E., Petrie, T., Soules, G., & Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *The Annals of Mathematical Statistics*, *41*(1), 164–171.

Carroll, R. J., Ruppert, D., Stefanski, L. A., & Crainiceanu, C. M. (2006). *Measurement Error in Nonlinear Models: a Modern Perspective.* CRC press.

Catania, L., & Di Mari, R. (2018). Hierarchical hidden markov models for multivariate integer-valued time-series with application to crime data. *Under review*.

Collins, L. M., & Lanza, S. T. (2010). *Latent Class and Latent Transition Analysis: With Applications in the Social, Behavioral, and Health Sciences* (Vol. 718). Wiley.

Cotton, S. M., Crewther, D. P., & Crewther, S. G. (2005). Measurement error: Implications for diagnosis and discrepancy models of developmental dyslexia. *Dyslexia*, *11*(3), 186–202.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1–38.

Di Mari, R., & Bakk, Z. (2017). Mostly harmless direct effects: a comparison of different latent markov modeling approaches. *Structural Equation Modeling: A Multidisciplinary Journal*.

Guo, Y., & Little, R. J. (2011). Regression analysis with covariates that have heteroscedastic measurement error. *Statistics in Medicine*, *30*(18), 2278–2294.

Heiss, F. (2008). Sequential numerical integration in nonlinear state space models for microeconometric panel data. *Journal of Applied Econometrics*, *23*(3), 373–389.

Kipnis, V., Subar, A. F., Midthune, D., Freedman, L. S., Ballard-Barbash, R., Troiano, R. P., … Carroll, R. J. (2003). Structure of dietary measurement error: results of the open biomarker study. *American Journal of Epidemiology*, *158*(1), 14–21.

Laird, N. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association*, *73*(364), 805–811.

Lindsay, B. G. (1983a). The geometry of mixture likelihoods: a general theory. *The Annals of Statistics*, *11*(1), 86–94.

Lindsay, B. G. (1983b). The geometry of mixture likelihoods, part ii: the exponential family. *The Annals of Statistics*, *11*(3), 783–792.

Rabe-Hesketh, S., & Skrondal, A. (2008). Classical latent variable models for medical research. *Statistical Methods in Medical Research*, *17*(1), 5–32.

Sánchez, B. N., Budtz-Jørgensen, E., & Ryan, L. M. (2009). An estimating equations approach to fitting latent exposure models with longitudinal health outcomes. *The Annals of Applied Statistics*, 830–856.

Uhrig, S. C. N., & Watson, N. (2017). The impact of measurement error on wage decompositions: Evidence from the british household panel survey and the household, income and labour dynamics in australia survey. *Sociological Methods & Research*.

Vermunt, J., & Magidson, J. (2004). Factor analysis with categorical indicators: A comparison between traditional and latent class approaches. In L. van der Ark, M. Croon, & K. Sijtsma (Eds.), *New developments in categorical data analysis for the social and behavioral sciences* (p. 41-63). Mahwah, NJ: Erlbaum.

Vermunt, J. K., Langeheine, R., & B ockenholt, U. (1999). Discrete-time discrete-state latent markov models with time-constant and time-varying covariates. *Journal of Educational and Behavioral Statistics*, *24*, 179-207.

Welch, L. R. (2003). Hidden markov models and the baum-welch algorithm. *IEEE Information Theory Society Newsletter*, *53*(4), 10–13.

Wiggins, L. M. (1973). *Panel analysis: latent probability models for attitude and behaviour processes.* Elsevier, Amsterdam.

Zucchini, W., MacDonald, I. L., & Langrock, R. (2016). *Hidden Markov Models for Time Series: an Introduction Using R.* Chapman and Hall / CRC Press.