Available online at www.sciencedirect.com
ScienceDirect
Computational Statistics & Data Analysis 52 (2007) 1183–1195
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

Effects of measurement errors in predictor selection of linear regression model

Kimmo Vehkalahti[a,*], Simo Puntanen[b], Lauri Tarkkonen[a]

[a] *Department of Mathematics and Statistics, University of Helsinki, P.O. Box 54, FI-00014 Helsinki, Finland*
[b] *Department of Mathematics, Statistics and Philosophy, University of Tampere, FI-33014 Tampere, Finland*

Available online 13 May 2007

Abstract

Measurement errors may affect the predictor selection of the linear regression model. These effects are studied using a measurement framework, where the variances of the measurement errors can be estimated without setting too restrictive assumptions about the measurement model. In this approach, the problem of measurement is solved in a reduced *true score space*, where the latent true score is multidimensional, but its dimension is smaller than the number of the measurable variables. Various measurement scales are then created to be used as predictors in the regression model. The stability of the predictor selection as well as the estimated predicted validity and the reliability of the prediction scales is examined by Monte Carlo simulations. Varying the magnitude of the measurement error variance four sets of predictors are compared: all variables, a stepwise selection, factor sums, and factor scores. The results indicate that the factor scores offer a stable method for predictor selection, whereas the other alternatives tend to give biased results leading more or less to capitalizing on chance.

Keywords: Measurement error; Regression; Reliability; Validity; Factor analysis

1. Introduction

The predictor selection of the linear regression model is affected, not only by the sampling variation, but also by the measurement errors. Let us assume that a predictor, say, x is measured with error. We can express this as $x = \tau + \varepsilon$, where τ is the true value of the predictor and ε is the random measurement error. It is reasonable to assume that ε is uncorrelated with τ, and hence we can write the variance of x as $\mathrm{var}(x) = \mathrm{var}(\tau) + \mathrm{var}(\varepsilon)$, where $\mathrm{var}(\tau)$ represents the sampling variation and $\mathrm{var}(\varepsilon)$ represents the measurement error variation. Either of these may dominate in a given study. If the measurements are unreliable, we cannot improve the situation by increasing the sample size. Instead, we should have more accurate measurements. In many applications it would be preferable to reduce the effects of the measurement errors in the predictor selection, and hence make the models more stable. However, the measurement errors are often neglected in the statistical models, including perhaps the most widely applied one, the linear regression model.

* Corresponding author. Tel.: +358 9 191 24863; fax: +358 9 191 51400.
E-mail address: Kimmo.Vehkalahti@helsinki.fi (K. Vehkalahti).

0167-9473/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2007.05.005