Social Science Research 53 (2015) 118–136

Contents lists available at ScienceDirect

Social Science Research

journal homepage: www.elsevier.com/locate/ssresearch

Multicollinearity in hierarchical linear models ☆

Han Yu [a], Shanhe Jiang [b], Kenneth C. Land [c,*]

[a] Department of Mathematics, Computer Science and Information System, Northwest Missouri State University, USA
[b] Department of Criminal Justice, The University of Toledo, USA
[c] Department of Sociology, Duke University, USA

ARTICLE INFO

Article history:
Received 28 August 2014
Revised 18 February 2015
Accepted 24 April 2015
Available online 19 May 2015

Keywords:
Multicollinearity
Hierarchical linear models
Top-down diagnosis
Singular value decomposition
Covariate pool

ABSTRACT

This study investigates an ill-posed problem (multicollinearity) in Hierarchical Linear Models from both the data and the model perspectives. We propose an intuitive, effective approach to diagnosing the presence of multicollinearity and its remedies in this class of models. A simulation study demonstrates the impacts of multicollinearity on coefficient estimates, associated standard errors, and variance components at various levels of multicollinearity for finite sample sizes typical in social science studies. We further investigate the role multicollinearity plays at each level for estimation of coefficient parameters in terms of shrinkage. Based on these analyses, we recommend a top-down method for assessing multicollinearity in HLMs that first examines the contextual predictors (Level-2 in a two-level model) and then the individual predictors (Level-1) and uses the results for data collection, research problem redefinition, model re-specification, variable selection and estimation of a final model.

1. Introduction

In the past two decades, hierarchical linear modeling has been increasingly applied in the social sciences in general, and in criminal justice in particular, as well as in other scientific fields. Although multicollinearity has been treated extensively in the literature on general multiple linear regression models and generalized linear regression models (McCullagh and Nelder, 1989), a literature review shows that little attention has been devoted to the effects of multicollinearity on parameter estimation in the context of hierarchical linear models (HLMs). However, the effect of multicollinearity on parameter estimates in HLMs is not trivial. It is more complicated than that in the classical linear models since the two kinds of effects (fixed and random) in HLMs are different and treated differently. On the one hand, HLMs have the following characteristics under well-conditioned situations: they are often applied to data with clustered structures so that they have more complex model components; they have conditional and marginal distributions and they are considerably more difficult to fit than classical linear models, typically requiring iterative optimization. On the other hand, HLMs also have the following problematic characteristics under ill-conditioned situations: multicollinearity at different levels might have different impacts on estimation; multicollinearity among the weighted predictors may occur at the maximum likelihood estimates of coefficients, resulting in the detrimental effect of inflated variances of the estimated coefficients due to many more levels of predictors and residuals in a HLM (cf. Belsley, 1991; Lesaffre and Marx, 1993); and the degree of multicollinearity of a HLM depends on the particular value of a coefficient. These characteristics complicate the impacts of multicollinearity in HLMs and help to account for the fact that comparatively little work on this topic has been done for HLMs.

☆ The authors would like to thank the four anonymous referees for their helpful comments and suggestions, which led to this revised version.
* Corresponding author.
E-mail address: kland@soc.duke.edu (K.C. Land).

http://dx.doi.org/10.1016/j.ssresearch.2015.04.008
0049-089X/

Based on our teaching experience, readings of articles in criminology and criminal justice journals, and communications with colleagues in the field, the effects of multicollinearity in HLMs are still not clear to scholars who use these models in empirical analyses. Popular textbooks on hierarchical linear models (Raudenbush and Bryk, 2002) and statistical packages such as HLM7 and SPSS do not have clear, detailed guidelines for diagnosing the presence of multicollinearity, the effect of multicollinearity on estimation of fixed and random effects, and remedies to it. Multicollinearity diagnostics are only available for linear regression in most software packages. Yet the problem of multicollinearity in HLMs is of increasingly great interest to applied quantitative researchers. The issues of where the presence of multicollinearity is and how to diagnose and remedy it in HLMs is specifically more relevant to applied researchers. Moreover, the issue of multicollinearity effects for a small sample size is more salient to applied researchers than large sample (asymptotic distribution theory) that still is being explored in the theoretical statistics literature.

The contribution of this paper is to exposit complicated multicollinearity problems in HLMs in such a way as to facilitate easy understanding and accessibility to those who have knowledge of classical linear models. After examining multicollinearity in hierarchical linear models as compared to classical linear models and proposing an intuitive, effective approach for diagnosing the presence of multicollinearity in HLMs by constructing a pool of predictors at the different levels and cross-levels, we suggest remedies to it. A simulation study is presented for insights into the multicollinear effects of finite sample sizes on coefficient estimates, associated standard errors and variance components under varying degrees of multicollinearity as a guide for prediction (predictive analysis) or for interpretation (prescriptive analysis/true model recovery), which are somewhat different tasks. We also show the role of different levels of multicollinearity in the estimation of coefficient parameters in terms of shrinkage, which is of more interest if predictive analysis is of major concern.

In the next section, we describe the well-known classical multiple linear model in a way that makes the extension to the class of hierarchical linear models appear natural for examining multicollinearity in HLMs and briefly review the nature, diagnostics and remedies of multicollinearity in classical multiple linear regression. The multicollinearity defined in classical multiple linear regression models then is carried naturally over to HLMs in Section 3 where we compare the multicollinearity in these models with that in classical linear models to facilitate better understanding and shed some new insight into the connection between hierarchical linear models and shrinkage regression on multicollinearity. After that, an empirical example illustrates methods for the diagnosis of multicollinearity. A simulation study in Section 4 is used to investigate how multicollinearity for various finite sample sizes affects fixed effects parameter estimates, associated standard errors and random effects parameter estimates under varying degrees of multicollinearity in HLMs. We also show the role multicollinearity plays at each hierarchical level in estimation of coefficient parameters in terms of shrinkage. A summary and conclusion is provided in Section 5.

2. Multicollinearity and its remedies in classical multiple linear models

The problem of multicollinearity (usually *near multicollinearity*) is a data problem in Multiple Regression that is surprisingly common (Greene, 2011; Bingham and Fry, 2010). Mandel (1982) claimed that the greatest source of difficulties in using least squares is the existence of "multicollinearity" in many sets of data. The rationale of multicollinearity is that there often exists a set of underlying *latent interrelationships* or *collinearities* among observed variables that span the columns of the data matrix. Certain subsets of regressors in a linear regression model may be connected by unobserved auxiliary regression equations that must be identified and accounted for if one is to generate reliable estimates of the regression model. Although the presence of substantial near multicollinearity is not a violation of the assumptions of the regression model—OLS remains unbiased and consistent and the standard errors accurate in the face of multicollinearity as long as it is not perfect collinearity, the existence of substantial correlation among a set of explanatory variables creates difficulties, namely, numerical instability, a reason why automated computer procedures such as the R commands *step* and *update* produce different outcomes depending on the order in which variables are selected in the model, and problems of identification and interpretation of the separate effects of the explanatory variables on the response variable of interest. Perfect multicollinearity further makes estimation methods that use the inverse of the predictor cross-product matrix such as Least Squares (LS) crash. It is important to keep in mind that multicollinearity refers not only to the situation where a pair of predictor variables has a substantial correlation with each other. It is also possible to have relationships between multiple predictors at once (see Belsley et al., 1980; Fox, 1991; Bingham and Fry, 2010; Kuhn and Johnson, 2013). It must be emphasized that some degree of multicollinearity is always present in most datasets; what is important is not the presence or absence of multicollinearity per se but the degree of multicollinearity in a dataset. In this regard, we are primarily interested in the cases in which the regressors in a model are highly (although not perfectly) collinear, which is termed *near multicollinearity*.

Classical linear regression methods are known to fail or, at least, perform sub-optimally, in the presence of multicollinearity. The use of collinear predictors can lead to serious statistical problems in parameter estimation: coefficient estimates may be unstable and have very high associated standard errors, coefficient estimates may have the "wrong" sign or implausible magnitudes or low significance levels even though they are jointly significant and the R^2 for the regression is quite high. As a result, the model selection process may proceed on a wrong trajectory and statistical inference is not reliable: the confidence interval becomes wide, a false null hypothesis is not rejected, and the individual coefficient effect is impossible to interpret separately from its correlated coefficient effects (i.e. the individual coefficient effect cannot be explained by fixing its correlated coefficient effects).