A Comparison of Multiple Regression to Two Latent Variable Techniques for Estimation and Prediction

RUIFENG LI 1 AND MELANIE M. WALL 2

ABSTRACT

In the areas of epidemiology, psychology, sociology, and other social and behavioral sciences, researchers often encounter situations where there are not only many variables contributing to a particular phenomenon, but also there are strong relationships among many of the predictor variables of interest. By using the traditional multiple regression on all the predictor variables, it is possible to have problems with interpretation and multicollinearity. As an alternative to multiple regression, we explore the use of a latent variable model which can address the relationship among the predictor variables. We consider two different methods for estimation and prediction for this model: one that uses multiple regression on factor score estimates and the other that uses structural equation modeling. The first method uses multiple regression but on a set of predicted underlying factors (i.e., factor scores) and the second method is a full multivariate maximum likelihood technique that incorporates the complete covariance structure of the data. In this paper, we will explain the model and each estimation method including how to do prediction. A data example will be used for demonstration, where respiratory disease death rates by county in Minnesota are predicted by five county level census variables. A simulation study is performed to evaluate the efficiency of prediction using the two latent variable modeling techniques compared to multiple regression.

keywords Multiple regression, factor analysis, structural equation modeling, respiratory disease.

1 Introduction

Multiple regression is one of the most widely used of all statistical methods. The two main uses of multiple regression are: estimating the effect that certain predictors have on the outcome when “adjusting” for other variables and predicting the outcome given a set of predictors. Despite its popularity, there are some disadvantages to this method. First, in order to have better estimation and prediction, it is very common to have more than 3 predictor variables, which makes identification and interpretation of the inter-relationships less straightforward, since our ability to visualize relationships is limited to 2 or 3 dimensions. One way to address this problem is to choose a smaller set of variables by using model selection methods. But, since most model selection criteria are highly data dependent, this does not allow the model to reflect the subject-matter knowledge. Another problem that can occur when using multiple

1Ruifeng Li is Biostatistician in the Department of Epidemiology, School of Public Health at Harvard University
2Melanie M. Wall is Assistant Professor in the Division of Biostatistics, School of Public Health at the University of Minnesota, Minneapolis, MN, 55455. This work is supported by National Center for Health Statistics grant NCHS-UR6/CCU517477-01.
regression stems from multicollinearity. When the predictor variables are highly correlated with one another, standard errors are inflated and typically most of the predictors are not statistically significant. Multicollinearity can commonly arise in fields where latent variables or factors are prominent, such as in epidemiology, psychology, sociology, and other social and behavioral sciences. In practice, when a variable is latent and thus not directly measurable, one chooses a variety of indicators of the latent variables which can be measured. Because they are all basically measuring the same latent variable, these observed variables are expected to be highly correlated. If all of these observed variables are put together into one large multiple regression, there are often too many predictor variables to interpret and also problems from multicollinearity.

A way of overcoming these shortcomings is by exploiting the highly correlated data structure and explicitly modeling it. Thus, the researcher can incorporate knowledge about the way in which predictor variables may be related to each other. Also, from an empirical point of view, it is possible to examine the correlation structure and hypothesize that variables which are highly correlated amongst themselves (but have relatively small correlations with other variables) are indicators of the same latent variable.

In this paper we will compare multiple regression with two different approaches that incorporate a latent variable model (i.e. incorporate information about how the predictors are related to one another). The first method, which we refer to as multiple regression on factor scores, is a 2-step technique where in the first step the underlying latent variables are estimated, then in the second step, multiple regression is performed on these newly estimated predictors. The other method, structural equation modeling, estimates the covariance structure of the outcome and all the observed predictors simultaneously while incorporating a model describing the way the predictors relate to one another.

In Section 2 we introduce the latent variable model and describe the two estimation and prediction techniques. Section 3 presents an example involving respiratory disease mortality by county in Minnesota and five census variables where both of the latent variable techniques will be used and their ability to describe the data compared to the results one would obtain with multiple regression. A simulation study comparing the efficiency of prediction using the latent variable methods to multiple regression is presented in Section 4. Conclusions and discussion
are found in Section 5.

2 A Latent Variable Model

Let $Y$ be some response of interest and $X$ be a vector of $p$ predictor variables. It is common to find research problems where several of the predictor variables in $X$ are in fact measuring the same latent variable. In such cases the impact of any one of the observed variables on the outcome is less of interest than the impact that the latent variable they are each measuring has on the outcome. Let $f$ be a vector of $k$ underlying latent variables or “factors” where $k < p$. Thus instead of estimating the relationship between $Y$ and $X$, we estimate the relationship between $Y$ and $f$. Consider the following

$$
X = \mu + \Lambda f + u \tag{1}
$$

$$
Y = \alpha_0 + \alpha' f + \epsilon, \tag{2}
$$

where $\mu$ $(p \times 1)$ and $\Lambda$ $(p \times k)$ are fixed coefficients of known or unknown scalars relating the $p$ observed variables $X$ to the $k$ underlying factors $f$, $u$ is a $p \times 1$ vector of i.i.d. random errors independent of $f$ with $E(u) = 0$ and $\text{Var}(u) = \Psi$, where $\Psi$ is a diagonal matrix. Assuming diagonality for $\Psi$ implies that the variables in $X$ are correlated solely because they are related to the same $k$ underlying factors. In general, this diagonality assumption can be relaxed although we do not consider it here. The “regression” parameters of interest are $\alpha_0$ and $\alpha$ which describe how the latent factors are related to the outcome of interest $Y$. The equation error $\epsilon$ is assumed to be i.i.d. random error independent of both $f$ and $u$ with $\text{Var}(\epsilon) = \sigma^2_\epsilon$.

2.1 Multiple Regression on Factor Scores

Since $f$ is unobservable, one way to proceed in estimating $\alpha_0$ and $\alpha$ is to first “measure” $f$ by using the information in $X$. The following details the method which first obtains the factor score estimates from (1) and then uses them in the regression for (2). Methods using factor scores have been popular in the educational and psychosocial literature for several decades, Horn (1965), Fleming (1981), Skrondal and Laake (2001), and see Grice (2001) for a comprehensive list of articles.
We first estimate the parameters in the factor analysis model (1). Because (1) is not identifiable without additional restrictions, we use model (3) instead with the common parameterization technique called errors-in-variables parameterization, i.e., we define \( k \) of the observed variables \( X_2 \) to be equal to the underlying factor plus measurement error and then the remaining \( (p-k) \) variables \( X_1 \) have the errors-in-variables regression relationship with \( X_2 \); see Fuller (1987). That is,

\[
X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} \beta_0 \\ 0_k \end{pmatrix} + \begin{pmatrix} \beta_1 \\ I_k \end{pmatrix} f + u, \tag{3}
\]

where \( 0_k \) is a \( k \times 1 \) vector of zeros and \( I_k \) is a \( k \times k \) identity matrix. Assume the \( f \) and \( u \) are i.i.d. normal and that \( E(f) = \mu_f \) and \( Var(f) = \Phi \), we then can proceed via normal maximum likelihood. Note, that the maximum likelihood estimators under the normality of \( f \) and \( u \) are known to produce \( \sqrt{n} \)-consistent estimators with good properties for virtually any type of \( f \) and any distribution of \( u \) (see, e.g. Anderson and Amemiya 1988; Browne and Shapiro 1988). Thus the estimates are the values maximizing the following:

\[
F_{ML} = tr(S \Sigma_{XX}^{-1}) + ln(|\Sigma_{XX}|),
\]

where \( S \) is the sample covariance matrix of \( X \), \( \Sigma_{XX} \) is the model covariance matrix of \( X \) with all parameters in it to be estimated, i.e.,

\[
\Sigma_{XX} = \begin{pmatrix} \beta_1 \\ I_k \end{pmatrix} \Phi(\beta_1', I_k) + \Psi. \tag{4}
\]

The parameter estimates \( \hat{\beta}_0, \hat{\beta}_1 \) and \( \hat{\Psi} \) can be obtained in any canned software package which performs factor analysis. Then, treating the \( \hat{\beta}_0, \hat{\beta}_1 \) and \( \hat{\Psi} \) as if they are known, we can use generalized least squares to obtain estimates for each \( f_i \), called the factor scores. That is, for \( X_i, i=1, \ldots, n \) (\( i^{th} \) observation of \( X \)), the corresponding \( f_i \) is estimated as the following:

\[
\hat{f}_i = \left( (\beta_1', I_k) \hat{\Psi}^{-1} \left( \begin{pmatrix} \beta_1 \\ I_k \end{pmatrix} \right) \right)^{-1} \left( (\beta_1', I_k) \hat{\Psi}^{-1} \left( X_i - \begin{pmatrix} \hat{\beta}_0 \\ 0_k \end{pmatrix} \right) \right), \tag{5}
\]

or the following more general formula which can be used when \( \hat{\Psi} \) is singular:

\[
\hat{f}_i = X_{2i} - \hat{\Sigma}_{uv} (\hat{\Sigma}_{vv})^{-1} \hat{V}_i, \tag{6}
\]
where $\mathbf{V}_i = \mathbf{X}_{1i} - \hat{\beta}_0 - \hat{\beta}_1 \mathbf{X}_{2i}$, $\hat{\Sigma}_{vv} = -\mathbf{\hat{\Psi}}_2 \hat{\beta}_1'$, and $\hat{\Sigma}_{iv} = \mathbf{\hat{\Psi}}_1 + \hat{\beta}_1 \hat{\Psi}_2 \hat{\beta}_1'$, where $\mathbf{\hat{\Psi}}_1$ is a diagonal matrix of the first $(p - k) \times (p - k)$ elements of $\mathbf{\hat{\Psi}}$, $\mathbf{\hat{\Psi}}_2$ is a diagonal matrix of the last $k \times k$ elements of $\mathbf{\hat{\Psi}}$.

Given these factor score estimates $\hat{\mathbf{f}}$, we can treat the $\hat{\mathbf{f}}$ as the predictor variables and fit the following model,

$$Y = \alpha_0 + \alpha \hat{\mathbf{f}} + \epsilon.$$  \hspace{1cm} (7)

Let $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)'$ and $\hat{\mathbf{f}} = (\hat{f}_1, \hat{f}_2, \ldots, \hat{f}_n)'$, then using ordinary least squares we obtain

$$[\hat{\alpha}_0, \hat{\alpha}]' = ([1_n, \hat{\mathbf{f}}]', [1_n, \hat{\mathbf{f}}])^{-1}([1_n, \hat{\mathbf{f}}]' \mathbf{Y}),$$

where $1_n$ is a vector of ones used to estimate the intercept. Note that because the $\hat{\mathbf{f}}$ are treated as if they are measuring the underlying factors without error, that the estimators for $\alpha_0$ and $\alpha$ in (8) are biased as in the typical error-in-variables regression model (Fuller, 1987).

Given a sample of $n$ observed $\mathbf{Y}$ and $\mathbf{X}$, we can estimate $\hat{\alpha}_0$, and $\hat{\alpha}$ as in (8). Then given a new observation $\mathbf{X}_{n+1}$, the predictor of $Y_{n+1}$ using this method is

$$\hat{Y}_{n+1} = \hat{\alpha}_0 + \hat{\alpha} \hat{f}_{n+1},$$

where $\hat{f}_{n+1}$ is the estimated factor score for the new observation estimated from formula (5) or (6), in which the corresponding parameter estimates $\hat{\beta}_1$, $\mathbf{\hat{\Psi}}$, and $\hat{\beta}_0$ can be estimated from the given sample of size $n$ and $\mathbf{X}_{n+1}$ is plugged in for $\mathbf{X}_i$. Simply plugging $\hat{f}_{n+1}$ into (9) as if it were the observed factor is similar to what others have suggested for prediction in the presence of measurement error, see e.g. Buonaccorsi (1995). The problem here differs though because $\hat{f}_{n+1}$ is not simply a new independently observed predictor with measurement error, it is also a function of the original data because it is a function of $\hat{\beta}_1$, $\mathbf{\hat{\Psi}}$, and $\hat{\beta}_0$.

### 2.2 Structural Equation Modeling

To avoid the problem of biased estimates and possible lack of efficiency due to two step estimation encountered when using regression on factor scores, we consider using structural equation modeling (SEM) to fit model (1)-(2). Although model (1)-(2) may itself be considered a structural equation model (Bollen, 1989), we reserve the term structural equation modeling for the
estimation method that considers simultaneously the complete structure of covariance between predictor variables and latent variables, as well as between response variable and latent variables. So the structural equation modeling method fits model (2) simultaneously with model (3) and performs estimation in one step. We consider the covariance structure \( \Sigma \) for \( (Y,X)'/1 \) simultaneously, i.e.,

\[
\text{Var} \begin{pmatrix} Y \\ X \end{pmatrix} = \Sigma = \begin{pmatrix} \Sigma_{YY} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{XX} \end{pmatrix},
\]

(10)

where \( \Sigma_{XX} \) is as before in (4), \( \Sigma_{YY} = \alpha' \Phi \alpha + \sigma^2_e \), and \( \Sigma_{XY} = \Sigma_{YX} = (\beta_1', I_k)^/\Phi \alpha \).

Assuming \( (Y,X)^/ \) is \((p + 1)\) multivariate normally distributed, we proceed via normal maximum likelihood and the parameter estimates are the values maximizing the following:

\[
F_{ML} = tr(S^{-1}S) + \ln(\left| \Sigma \right|),
\]

where \( S \) is the sample covariance matrix of \( (Y,X)^/ \), and \( \Sigma \) is the covariance matrix in (10). Any software package which performs factor analysis or more generally structural equation modeling can be used to obtain the maximum likelihood estimators, we call them \( \hat{\Psi}, \hat{\Phi}, \hat{\beta}_1, \) and \( \hat{\alpha} \). Note that the parameters appearing only in the mean structure, i.e., \( \alpha_0, \beta_0 \) and \( \mu_f \), can be estimated by subtraction once the other parameters are estimated. That is, following the notation for \( X_1 \) and \( X_2 \) as in (3), \( \hat{\alpha}_0 = \hat{Y} - \alpha' \hat{X}_2, \hat{\beta}_0 = \hat{X}_1 - \beta_1, \hat{X}_2, \) and \( \hat{\mu}_f = \hat{X}_2 \).

The prediction of an unobserved \( Y_{n+1} \) given a new observation \( X_{n+1} \) using this method can then be obtained as the conditional mean of \( Y_{n+1} \) given \( X_{n+1} \) with the maximum likelihood estimators based on a sample of size \( n \) plugged in, i.e.,

\[
\hat{Y}_{n+1} = \hat{\alpha}_0 + \alpha' \mu = \hat{\Sigma}_{YX} \hat{\Sigma}_{XX}^{-1} \left[ X_{n+1} \right] - \left( \begin{array}{c} \hat{\beta}_0 \\ 0 \end{array} \right) - \left( \begin{array}{c} \hat{\beta}_1 \\ I \end{array} \right) \mu_f'.
\]

(11)

Although very straightforward, this type of prediction is rarely done by users of structural equation modeling. This is because usually the results of most interest to SEM researchers are the parameter estimates and in particular if the covariance structure implied by the model matches that observed in the data. Traditionally, model checking has almost exclusively focused on how closely the fitted model covariance matches the sample covariance rather than how closely the observed individual data matches the predicted individual data. More recently, the
formation of individual level residuals has been proposed by Raykov and Penev (2001) where a similar predicted value to (11) is used.

3 Example Data

The example data used here are Minnesota county-level respiratory disease death rates based on death record data from 1990 to 1998. In Minnesota, the respiratory disease death rate on average over the years 1990-1998 is around 7 cases per 10,000 people. The total number of cases in this period is 31,374, which accounts for slightly over 10% of all deaths. An ecological data analysis of interest is to relate these death rates to county level census variables. The response variable, $RESP$, is the log age-adjusted respiratory disease death rate for each of the 87 Minnesota counties over the years 1990-1998. We consider 5 predictor variables available from the 1990 census, of which 3 of them ($eduhs$, $medhhin$, and $percapit$) could be considered measures of social economic status (SES) of each county, and 2 of them ($pubwater$ and $wood$) could be considered measures of the ruralness of each county. Table 1 provides the basic characteristics about the six variables of interest.

Applying multiple regression to these variables with $RESP$ as the outcome and the five census variables as predictors we have,

$$
RESP_i = \beta_0 + \beta_1 eduhs_i + \beta_2 medhhin_i + \beta_3 percapit_i + \beta_4 pubwater_i + \beta_5 wood_i + \epsilon_i,
$$

where $i=1, \ldots, 87$ counties in Minnesota. The results are given in Table 2.

From Table 2, we find that when we include all 5 predictors in the multiple regression, only the estimate for $\beta_5$ (coefficient of $wood$) is statistically significant, and all others are non-significant with very large p-values. Thus the interpretation is that after adjusting for all the other variables, only the percent of households using wood has a significant effect on respiratory disease death rate. But if we throw out $wood$, then $pubwater$ actually becomes significant. And, if we throw $medhhin$ and $percapit$ out but leave both $wood$ and $pubwater$ in, then $eduhs$ becomes significant. These changing significant effects caused by the high correlation among the variables, make interpreting the effect that each of these variables has on the outcome very difficult. If we use the variance inflation factor as an index of multicollinearity, we have $eduhs$
is 4.3, medhhin is 8.1, percapit is 10.4, pubwater is 5.1, and wood is 4.6 from model (12) which are all considerably high. Table 3 and the scatter plot matrix in Figure 1 display the strong correlations (r), especially pubwater with wood (r = -0.87), eduhs with medhhin (r = 0.84), and also eduhs and medhhin with percapit (r = .86, .93) respectively.

This correlation structure and what we know substantively about the observed variables suggest the existence of two latent factors: ruralness (measured by wood and pubwater) and SES (measured by eduhs, medhhin, and percapit). Thus, it is reasonable to consider a latent variable model where we are interested in regressing the outcome RESP on the factors ruralness and SES rather than on the observed indicators.

Hence we consider the following latent variable model which takes into account the existence of the two latent factors ruralness and SES,

\[
\begin{align*}
eduhs &= \beta_{10} + \beta_{11} SES + u_1 \\
medhhin &= \beta_{20} + \beta_{21} SES + u_2 \\
percapit &= \beta_{30} + \beta_{32} ruralness + u_3 \\
pubwater &= \beta_{30} + \beta_{32} ruralness + u_4 \\
wood &= \beta_{30} + \beta_{32} ruralness + u_5 \\
RESP &= \alpha_0 + \alpha_1 SES + \alpha_2 ruralness + \epsilon.
\end{align*}
\]

Using the notation in (3), we have column vector \(X_1\) as \((eduhs, medhhin, pubwater)'\) and \(X_2\) as \((percapit, wood)'\). Note that \(X_2\) is chosen so that each factor has one observed indicator identifying the scale. This latent variable model has \(k = 2\) factors \(f = (SES, ruralness)'\) and we make assumptions for the \(u\) and \(\epsilon\) as in Section 2. The estimation method described in Section 2.1 requires the prediction of the underlying factors ruralness and SES. Focusing only on the model for the \(X\) variables, we obtain using SAS PROC CALIS, the following estimates \(\hat{\beta}_0, \hat{\beta}_1,\) and \(\hat{\Psi}\):

\[
\begin{align*}
\hat{\beta}_0' &\equiv (\hat{\beta}_{10}, \hat{\beta}_{20}, \hat{\beta}_{30}) = (47.6, -4734.8, 80.3) \\
\hat{\beta}_1 &\equiv \begin{pmatrix}
\hat{\beta}_{11} & 0 \\
\hat{\beta}_{21} & 0 \\
0 & \hat{\beta}_{32}
\end{pmatrix} = \begin{pmatrix}
0.002 & 0 \\
2.65 & 0 \\
0 & -2.38
\end{pmatrix} \\
\hat{\Psi} &\equiv \text{var}(u) = \text{diag}(8.59, 4761551.76, 0, 2.17, 20.43).
\end{align*}
\]
Since $\mathbf{\Psi}$ is singular, formula (6) is used to estimate the underlying factors $\mathbf{\hat{f}}_i = (\mathbf{SES}_i, \mathbf{ruralness}_i)$ for each county, and we can then use these as “observations” of the latent factors and fit the model

$$\text{RESP} = \alpha_0 + \alpha_1 \ast \mathbf{SES} + \alpha_2 \ast \mathbf{ruralness} + \epsilon$$

(13)

using ordinary least squares. The resulting $(\hat{\alpha}_0, \hat{\alpha}_1, \hat{\alpha}_2)$ and their associated inference are given in Table 4.

Using the structural equation model method in Section 2.2, we fit the model for all the variables simultaneously (again using SAS PROC CALIS) and obtain the estimates and inference for $(\alpha_0, \alpha_1, \alpha_2)$ shown in Table 4. The $R^2$ value using the SEM technique is calculated by taking $(\bar{\text{Var}}(\text{RESP}) - \hat{\sigma}_i)/\bar{\text{Var}}(\text{RESP})$.

From Table 4, we see that for both latent variable methods the coefficients of $\text{SES}$ and $\mathbf{ruralness}$ are significant, which means there is a statistically significant impact of county level $\text{SES}$ and $\mathbf{ruralness}$ on the respiratory mortality rate. The multicollinearity problem present in the multiple regression has been accounted for by the presence of underlying latent variables. Notice that the signs of the effects match what we find in the simple correlations in Table 3 as well. That is, as $\text{SES}$ increases there is an increase in death rate due to respiratory disease (note each of eduhs, medhhin, and percapi is positively related with RESP in Table 3), and as $\mathbf{ruralness}$ increases there is also an increase in death rate due to respiratory disease (note in Table 3 wood which is the percent of households that heat their home with wood is positively correlated with RESP, and pubwater which is in the opposite direction as ruralness is negatively related with RESP). The scales of the latent variables are in the same units as the observed variables they were identified with, i.e. the units of $\text{SES}$ are the same as per capita income and for $\mathbf{ruralness}$, the units are percent of households with wood heat. Although we present unstandardized estimates, it is common that standardized coefficients be used with latent variable models in order to avoid confusion about scale.

4 Simulation Study

We have seen in the previous section through the data example that there are benefits in terms of estimation and interpretation in using the latent variable techniques instead of multiple re-
gression. In this section, we will examine whether these methods can perform well for prediction as compared to multiple regression.

We will use two criteria for comparison of the predictability of the different methods: empirical expected mean squared error (EMSE) and empirical expected mean squared prediction error (EMSPE). Details of how these are computed are given below. In general terms, the EMSE measures how well the model fits a given data set, and the EMSPE examines how well the fitted model predicts future observations.

We consider samples generated from the following model

\[
x_1 = 3 + 0.5 f_1 + 0.5 e_1 \\
x_2 = 3 + 0.7 f_1 + 0.5 e_2 \\
x_3 = f_1 + 0.5 e_3 \\
x_4 = 2 + 0.8 f_2 + 0.5 e_4 \\
x_5 = 3 + 0.7 f_2 + 0.5 e_5 \\
x_6 = f_2 + 0.5 e_6 \\
y = 3 + 4 f_1 + 2.5 f_2 + \zeta e_7,
\]

where \(e_i \sim iid N(0,1), i = 1, \ldots, 7\). Data are generated from two different assumptions for the underlying factors, i.e. \(f_1\) and \(f_2\), are distributed either as standard normal or standardized \(\chi^2(1)\) with mean 0 and variance 1; and from two different values for \(\zeta\), i.e. it is either 1 or 4. We choose \(\chi^2(1)\) because it represents a highly skewed distribution (highly non-normal). We choose two different levels for \(\zeta\) to represent a high and low level of reliability for the \(y\) variable. The reliability for \(y\) (i.e. \(1 - Var(\epsilon)/Var(y)\)) equals .96 when \(\zeta = 1\) and .58 when \(\zeta = 4\).

We consider the fit of the following latent variable model to the generated data,

\[
x_1 = \beta_{10} + \beta_{11} f_1 + \beta_{12} f_2 + u_1 \\
x_2 = \beta_{20} + \beta_{21} f_1 + \beta_{22} f_2 + u_2 \\
x_3 = f_1 + u_3 \\
x_4 = \beta_{30} + \beta_{31} f_1 + \beta_{32} f_2 + u_4 \\
x_5 = \beta_{40} + \beta_{41} f_1 + \beta_{42} f_2 + u_5 \\
x_6 = f_2 + u_6
\]
\[ y = \alpha_0 + \alpha_1 f_1 + \alpha_2 f_2 + \epsilon. \]

Notice we do not assume that the \( \beta_{12} = \beta_{22} = \beta_{31} = \beta_{41} = 0 \) although the generated data fix them to zero. That is, we do not give the latent variable techniques an unfair advantage in knowing the true model, except for the fact that two underlying factors are assumed. Both latent variable methods discussed in Section 2 are considered for fitting this model. Using the same generated data, we also fit of the following multiple regression model:

\[ y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \delta, \quad (15) \]

where \( \delta \) is assumed to be i.i.d. and ordinary least squares is used for estimation.

To calculate EMSE, we generate 1000 datasets from (14) each with \( n = 200 \) observations. Let \( (y_{ij}, x_{1ij}, x_{2ij}, x_{3ij}, x_{4ij}, x_{5ij}, x_{6ij}) = (y_{ij}, x_{ij}) \) be the \( j^{th} \) observation \((j = 1, \ldots, 200)\) within the \( i^{th} \) dataset \((i = 1, \ldots, 1000)\). Then for the \( i^{th} \) dataset, we predict \( y_{ij} \) using either (9), (11), or the usual OLS prediction based on (15). Let \( \hat{\theta}_i \) represent the set of parameters estimated using the \( i^{th} \) data set for any of the three methods. Then we let \( \hat{g}(\hat{\theta}_i, x_{ij}) \) denote the predicted value of \( y_{ij} \) given \( \hat{\theta}_i \) and \( x_{ij} \). Since there are \( j = 1, \ldots, 200 \) observations, we obtain 200 predicted values using each of the three methods (OLS, regression on factor scores, and SEM) and obtain for dataset \( i \), \( MSE_i = \sum_{j=1}^{200}(\hat{g}(\hat{\theta}_i, x_{ij}) - y_{ij})^2 / 200. \) Table 5 presents the \( \sum_{i=1}^{1000} MSE_i / 1000 \) or the empirical expected mean squared error (EMSE) for each of the methods and each data scenario.

The empirical EMSE examines how well each of the three methods predicts \( Y \) values that were observed. The empirical expected mean square prediction error (EMSPE) will allow us to consider true prediction, i.e. how well the different methods perform when predicting future \( Y \) values based on a fitted model from a training dataset. To compute the empirical EMSPE we first use each of the three methods to fit models to 50 randomly generated “training data” sets of size 200. For the \( k^{th} \) training data set, we obtain the estimated parameters \( \hat{\theta}_k \) using each of the three methods. Then we see how well these three fitted models each with their own \( \hat{\theta}_k \) can predict the future \( y \) values in 1000 data sets of size 200 with newly generated predictor variables. For each of the training data sets and each of the 1000 newly generated data sets we calculate \( MSPE_{ik} = \sum_{j=1}^{200}(\hat{g}(\hat{\theta}_k, x_{ij}) - y_{ij})^2 / 200 \) where we emphasize that unlike the \( MSE_i \) above, that \( \hat{\theta}_k \) has been estimated using a separate data set. Note it was not necessary to choose the size of the newly generated data sets to be the same size as the original training data sets.
Then the empirical EMSPE is calculated as \( \sum_{k=1}^{50} \sum_{i=1}^{1000} MSPE_{ik} / 50,000 \). The results for the EMSPE are shown in Table 5.

From Table 5, we see that the structural equation modeling method is always best in terms of EMSE and EMSPE for the four different scenarios. These differences translate into an average improvement in relative efficiency for SEM over OLS of 1.8% for EMSE and 2.3% for EMSPE. The factor score prediction method is performing slightly worse than OLS except for EMSPE in the cases where the reliability is low. As is expected the scale of the EMSE and EMSPE is larger when the reliability is lower because the error variance in the model is bigger. Note that each of the methods is quite robust to the non-normality. That is, the results for the \( \chi^2 \) data are very similar to the normal data.

In order to examine the effect of sample size on these results, we also consider the same model set up with the \( f \) normally distributed and the reliability of \( y \) equal to .96, but we let \( n = 50, 200, \) and \( 1000 \). Table 6 presents the results. The increase in relative efficiency of SEM compared to OLS when \( n = 50 \) is 8.2% for EMSE and 8.5% for EMSPE. As mentioned above, when \( n = 200 \), the relative gain in efficiency is approximately 2%, and when \( n = 1000 \), there is practically no gain, i.e. .03%. Thus we see that the improvement of SEM over OLS is decreasing with increasing sample size. Also in Table 6 we see that the inferiority of the factor score method compared to OLS lessens as the sample size increases.

## 5 Conclusion and Discussion

We have demonstrated the use of two latent variable modeling techniques in the respiratory disease death rate example and have shown that they both yield similar straightforward interpretation of effects on outcomes as compared to multiple regression. This is accomplished because of the natural way they incorporate the inherent collinearity in the predictor variables. We have also shown via a simulation study that in terms of prediction, the SEM method outperforms the ordinary least squares method for data where the latent variable model is appropriate.

A possible criticism of the simulation study we present may be that we are somehow cheating because we generate data from a two factor latent variable model and then we fit the “correct” two factor model using the latent variable techniques. In practice one does not know the correct
number of factors. Part of any sound latent variable model building is first testing for the number of underlying factors using a simple chi-squared test statistic based on the likelihood ratio. Hence in practice, one first decides on the number of factors and then proceeds as in Section 2. It is certainly possible that no number of identifiable factors fit the data well according to the likelihood ratio test, in which case we would not recommend fitting a latent variable model.

Other studies have compared the predictability of models that reduce the dimension of the data. Morris (1980) compared regression of several different types of factor scores including the factor scores in (5), to ordinary least squares in a simulation study. Morris's study which focused on relatively small sample sizes and an outcome with low reliability equal to .33 consistently found the prediction using factor score regression to beat the ordinary least squares prediction. The results of that study should be viewed skeptically though as the simulation was based on resampling of the same 1000 observations for all cases. Also, Morris only examined the correlation between observations and predictions rather than a statistic which could also account for bias in the prediction such as EMSPE. Helland and Ahmey (1994) examined the asymptotic properties and did a simulation study to compare principle component regression and partial least squares regression and hybrids of them. They found mixed results with no clear winner for prediction.

The difference between the SEM technique and dimension reduction techniques such as factor score, principle component, or partial least squares regression is that it is not really reducing the dimension of the data; it is only modeling more of it by modeling the relationships in the $X$. It is the fact that SEM is not reducing the dimension or information in the data that makes it beat the factor score method, and it is this modeling of the $X$ variables that gives SEM the advantage over ordinary least squares.

How is it possible for the SEM technique to beat ordinary least squares in terms of prediction? It is well known that the prediction from ordinary least squares regression is the best linear unbiased predictor when the errors in the equation are i.i.d. While the predictor using factor score estimates (9) is a linear predictor (i.e. a linear function of $Y$), the SEM predictor (11) is not since the parameter estimators are nonlinear function of both the $Y$ and $X$ variables. This may help to explain how it performs more efficiently.

When the i.i.d. data $(Y, X)$ is jointly normally distributed, it is also commonly understood
that the ordinary least squares predictor is, in fact, the best unbiased predictor (not just best linear). So how is it that the SEM technique can beat the OLS procedure when the data is normal? We know that $E(Y|X)$ is the best predictor in terms of expected mean squared error and when the $(Y,X)$ is jointly normal, with $E(Y,X) = (\mu_Y, \mu_X)$, $Var(X) = \Sigma_{XX}$ and $Cov(Y,X) = \Sigma_{YX}$, then $E(Y|X) = \mu_Y + \Sigma_{YX} \Sigma^{-1}_{XX}(X - \mu_X)$. The best predictor given a particular data set is then equal to $E(Y|X)$ with the maximum likelihood estimates plugged in for $\mu_Y$, $\mu_X$, $\Sigma_{YX}$, and $\Sigma_{XX}$. When nothing is assumed about the $p(p+1)/2$ unique elements of the symmetric matrix $\Sigma_{XX}$, the maximum likelihood estimator for $\Sigma_{XX}$ is simply the sample covariance matrix of $X$, i.e. every element is estimated independently, and the $E(Y|X)$ with maximum likelihood estimators plugged in yields the OLS predictor. But, if we have some model for the elements of $\Sigma_{XX}$ as is the case in the latent variable model where $\Sigma_{XX}$ is a function of fewer parameters than $p(p+1)/2$, and these parameters also appear in $\Sigma_{YX}$, then maximum likelihood estimators based on the modeled $\Sigma_{XX}$ and $\Sigma_{YX}$ plugged into $E(Y|X)$ such as (11) should be best with respect to mean squared prediction error.

A final comment should be made about the restriction of the results shown in this paper to the cases where all the latent variables are linearly related to the outcomes. Traditionally latent variable modeling techniques have focused on models that contained only linear latent variables. Although new techniques have been proposed for estimating parameters in polynomial latent variable models, Wall and Amemiya (2000, 2001), it is not clear how these techniques would perform in terms of prediction.

References


<table>
<thead>
<tr>
<th>Variable</th>
<th>Interpretation (all on county-level)</th>
<th>Mean</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>RESP</td>
<td>log age-adjusted respiratory disease death rate</td>
<td>-7.3</td>
<td>-7.8</td>
<td>-6.8</td>
</tr>
<tr>
<td>eduhs</td>
<td>percent with high school education</td>
<td>75.2</td>
<td>64.3</td>
<td>90.7</td>
</tr>
<tr>
<td>medhhin</td>
<td>median household income (in dollars)</td>
<td>25,052</td>
<td>16,924</td>
<td>44,122</td>
</tr>
<tr>
<td>percapit</td>
<td>per capita income (in dollars)</td>
<td>11,227</td>
<td>7,737</td>
<td>18,496</td>
</tr>
<tr>
<td>pubwater</td>
<td>percent of households with access to public water</td>
<td>56.4</td>
<td>11.2</td>
<td>97.1</td>
</tr>
<tr>
<td>wood</td>
<td>percent of households using wood to heat the home</td>
<td>10.0</td>
<td>0.3</td>
<td>38.4</td>
</tr>
</tbody>
</table>

Table 1: The basic characteristics of the data set n=87 counties in Minnesota

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Standard Error</th>
<th>T value</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_0$ : intercept</td>
<td>-8.03</td>
<td>0.40</td>
<td>-20.22</td>
<td>0.00</td>
</tr>
<tr>
<td>$\beta_1$ : eduhs</td>
<td>0.005</td>
<td>0.007</td>
<td>0.71</td>
<td>0.47</td>
</tr>
<tr>
<td>$\beta_2$ : medhhin</td>
<td>0.000007</td>
<td>0.00001</td>
<td>0.72</td>
<td>0.47</td>
</tr>
<tr>
<td>$\beta_3$ : percapit</td>
<td>-0.00000002</td>
<td>0.00003</td>
<td>-0.05</td>
<td>0.95</td>
</tr>
<tr>
<td>$\beta_4$ : pubwater</td>
<td>0.001</td>
<td>0.002</td>
<td>0.46</td>
<td>0.64</td>
</tr>
<tr>
<td>$\beta_5$ : wood</td>
<td>0.013</td>
<td>0.005</td>
<td>2.77</td>
<td>0.01</td>
</tr>
</tbody>
</table>

R²: 0.23

Table 2: The estimated coefficients of model (12) using ordinary least squares
### Table 3: The correlation (lower diagonal) and covariance (upper diagonal), and variance (diagonal) of the six variables in the model (12)

<table>
<thead>
<tr>
<th></th>
<th>RESP</th>
<th>eduhs</th>
<th>medhhin</th>
<th>percapit</th>
<th>pubwater</th>
<th>wood</th>
</tr>
</thead>
<tbody>
<tr>
<td>RESP</td>
<td>0.04</td>
<td>0.19</td>
<td>124.77</td>
<td>31.42</td>
<td>-1.15</td>
<td>0.73</td>
</tr>
<tr>
<td>eduhs</td>
<td>0.16</td>
<td>32.96</td>
<td>27820.76</td>
<td>9916.99</td>
<td>40.26</td>
<td>-16.48</td>
</tr>
<tr>
<td>medhhin</td>
<td>0.10</td>
<td>0.84</td>
<td>33165283.45</td>
<td>10705435.33</td>
<td>49235.00</td>
<td>-23081.00</td>
</tr>
<tr>
<td>percapit</td>
<td>0.07</td>
<td>0.86</td>
<td>0.93</td>
<td>4034904.53</td>
<td>21171.14</td>
<td>-8881.00</td>
</tr>
<tr>
<td>pubwater</td>
<td>-0.28</td>
<td>0.36</td>
<td>0.43</td>
<td>0.53</td>
<td>389.40</td>
<td>-162.44</td>
</tr>
<tr>
<td>wood</td>
<td>0.37</td>
<td>-0.30</td>
<td>-0.43</td>
<td>-0.47</td>
<td>-0.87</td>
<td>88.58</td>
</tr>
</tbody>
</table>

### Table 4: The estimated coefficients incorporating the latent variable model

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Regression on factor scores</th>
<th>SEM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimate (s.e.)</td>
<td>P-value</td>
</tr>
<tr>
<td>$\alpha_0 : \text{intercept}$</td>
<td>-7.68 (0.14)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>$\alpha_1 : \text{SES}$</td>
<td>0.00003 (0.00001)</td>
<td>0.0100</td>
</tr>
<tr>
<td>$\alpha_2 : \text{ruralness}$</td>
<td>0.010 (0.003)</td>
<td>0.0006</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.14</td>
<td></td>
</tr>
<tr>
<td>simulated scenarios distribution</td>
<td>EMSE</td>
<td>EMSPE</td>
</tr>
<tr>
<td>----------------------------------</td>
<td>------</td>
<td>-------</td>
</tr>
<tr>
<td>of f reliability</td>
<td>OLS</td>
<td>FS</td>
</tr>
<tr>
<td>Normal</td>
<td>3.80</td>
<td>3.94</td>
</tr>
<tr>
<td>$\chi^2$</td>
<td>3.76</td>
<td>3.91</td>
</tr>
<tr>
<td>Normal</td>
<td>18.76</td>
<td>18.89</td>
</tr>
<tr>
<td>$\chi^2$</td>
<td>18.87</td>
<td>19.01</td>
</tr>
</tbody>
</table>

Table 5: Empirical expected mean squared error and expected mean square prediction error for samples of size 200 for the three prediction techniques: Ordinary least squares (OLS), regression of factor scores (FS) and structural equation modeling (SEM) under four different generated data scenarios. EMSE are averages of 1000 simulated datasets and EMSPE are averages of 50 $\times$ 1000 simulated data sets.

<table>
<thead>
<tr>
<th>sample size</th>
<th>EMSE</th>
<th>EMSPE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>OLS</td>
<td>FS</td>
</tr>
<tr>
<td>50</td>
<td>3.79</td>
<td>4.36</td>
</tr>
<tr>
<td>200</td>
<td>3.80</td>
<td>3.94</td>
</tr>
<tr>
<td>1000</td>
<td>3.81</td>
<td>3.84</td>
</tr>
</tbody>
</table>

Table 6: Empirical expected mean squared error and expected mean square prediction error for samples generated with normal $f$ and reliability = .96 for $y$. The three prediction techniques: Ordinary least squares (OLS), regression of factor scores (FS) and structural equation modeling (SEM) were used with three different sample sizes. EMSE are averages of 1000 simulated datasets and EMSPE are averages of 50 $\times$ 1000 simulated data sets.
Figure 1: Scatter Plot: RESP vs. Predictor Variables