Controlling omitted variables and measurement errors by means of constrained autoregression and structural equation modeling

Suparman, Yusep

Chapter 2 Hedonic Price Models with Omitted Variables and Measurement Errors: A Constrained Autoregression - Structural Equation Modeling Approach with Application to Urban Indonesia[1]

Abstract

Omitted variables and measurement errors in explanatory variables frequently occur in hedonic price models. Ignoring these problems leads to biased estimators. In this paper we develop a constrained autoregression - structural equation model (ASEM) to handle both types of problems. Standard panel data models to handle omitted variables bias are based on the assumption that the omitted variables are time-invariant. ASEM allows handling of both time-varying and time-invariant omitted variables by constrained autoregression. In the case of measurement error, standard approaches require additional external information which is usually difficult to obtain. ASEM exploits the fact that panel data are repeatedly measured which allows decomposing the variance of a variable into the true variance and the variance due to measurement error. We apply ASEM to estimate a hedonic housing model for urban Indonesia. To get insight into the consequences of measurement error and omitted variables, we compare the ASEM estimates with the outcomes of (i) a standard SEM, which does not account for omitted variables, (ii) a constrained autoregression model, which does not account for measurement error, and (iii) a fixed effects hedonic model which ignores measurement error and time-varying omitted variables. The differences between the ASEM estimates and the outcomes of the three alternative approaches are substantial.

[1] This chapter has been published as Suparman Y, Folmer H, Oud JHL (2014) Hedonic price models with omitted variables and measurement errors: a constrained autoregression - structural equation modeling approach with application to urban Indonesia. *Journal of Geographical System* 16(1):49-70.

Keywords: hedonic housing price model, panel model, structural equation model, constrained autoregression, measurement error, omitted variable bias, urban Indonesia.

1. Introduction

The hedonic price (HP) model is one of the most basic tools in spatial sciences. It is based on the notion that a market good is composed of a bundle of attributes. For instance, a house is made up of structural attributes like size, number of rooms, wall and floor material, availability of facilities like toilets and electricity, and neighborhood characteristics like vicinity of schools, parks, shopping centers, safety, neighborhood wealth, and so on. The price of a good is a function of the prices of the attributes. Based on this notion, the HP approach decomposes the price of a good like a house into the prices of its attributes.

The first application of HP modeling dates back to Waugh (1926) who analyzed the impacts of quality factors on vegetable price. Since the mid 1960's application of HP modeling started increasing. For instance, Ridker and Henning (1967), Nourse (1967), and Barrett and Waddell (1970) estimated the effect of air pollution on housing prices, while Dornbusch and Barrager (1973), and Gamble et al. (1973) analyzed the impacts of water pollution and highway noise on property values. The foundation of HP modeling was developed by Rosen (1974). Since then, HP models have been widely used to value both public and market goods. Particularly, the application of housing HP models has grown in popularity.

Although housing HP studies have become popular and are routinely applied, they frequently encounter under-specification problems, i.e. the variables that actually belong to the true or population model are missing. Under-specification may be due to

data collection problems, such as time or resource constraints (Clarke 2005) or to methodological considerations. For instance, the inclusion of all relevant characteristics in a housing HP model increases multicollinearity. Accordingly, in practice several of the systematic house characteristics are omitted from the model. This practice leads to omitted variable bias (Greene 2008; Butler 1982; Ozanne and Malpezzi 1985). Particularly, the estimators of the parameters of the included variables are biased in a complicated fashion. In addition, the directions of the biases are difficult to assess, since they depend on the correlations among the included and omitted variables, and on the signs of the impacts of the latter on the dependent variable (Greene 2008). Note that although researchers have started considering spatial spill-over and heterogeneity in HP models to increase prediction accuracy (Páez 2007; Pace and Lesage 2004), little attention is still being paid to omitted variables.

Another source of bias in HP studies is measurement error. Amongst others Gujarati (2004) and Greene (2008) show that measurement error in an explanatory variable leads to attenuation (bias toward zero) of the estimator of the coefficient of the explanatory variable measured with error, while the biases of the estimators of the coefficients of the other variables can be in different directions. In housing HP studies measurement errors in the explanatory variables are very common. Many house characteristics are latent variables[2] that cannot be measured directly. To give empirical meaning to a latent variable indicators are used. However, the relationship between a latent variable and its indicator(s) is partial (Hempel 1970). That is, the latent variable is measured with error. For instance, drinking water quality is often measured by type of

[2] A latent variable refers to a phenomenon that is supposed to exist but cannot be observed directly. Examples are welfare, quality of life, socioeconomic status. A latent variable is given empirical meaning by means of a correspondence statement or operational definition. Such a statement connects a latent variable with a set of observables. For instance, the latent variable socio-economic status is operationalized (measured) by observed variables like income, education, and profession. For further details, see Folmer (1986) and the references therein.

domestic water supply (tap, well, river or lake) (see e.g. Yusuf and Koundouri 2004). However, an indicator like type of water supply only partly measures water quality aspects like the concentration of heavy metals or bacteria coli. Accordingly, the variance of the indicator "type of water supply" is made up of the variance of the latent variable "water quality" and the variance of the measurement error. In a similar vein, a latent variable like neighborhood quality is measured by indicators like crime rate, quality of schools, and socioeconomic status of the population. Each of the indicators partly measures a dimension of neighborhood quality which again leads to measurement errors, and thus to variances of the observed indicators consisting of the variance of the latent variable on the one hand and the variances of the measurement errors on the other. Even in the case of "technical measures" such as floor area, measurement is likely to be subject to error. In this case the source is the accuracy of the measurement instrument. For most technical measures the measurement error variance is small and therefore often ignored which nevertheless biases the estimators of the model coefficients.

In this paper we exploit panel data characteristics to account for both omitted variables bias and measurement error bias. Omitted variables bias is corrected via the constrained autoregression option that panel data offers. To account for measurement error we propose structural equation models (SEM). A SEM estimated on the basis of panel data makes it possible to repeatedly measure a variable and thus to decompose its variance into a component related to the latent variable on the one hand and measurement error variance on the other. We denote the combined model as the constrained autoregression - structural equation model (ASEM).

2. The Constrained Autoregression-Structural Equation Model (ASEM)

2.1. Constrained autoregression

As argued above, a housing HP model is based on the view that a good is a bundle of attributes. Hence, the price or rent of a house can be decomposed into the prices of the individual attributes (Malpezzi 2008). That is, let the price of house i at time t (p_{it}) be determined by $a+b$ systematic house characteristics $q_{1it} \cdots q_{(a+b)it}$ according to the linear function:

$$p_{it} = \pi_{0t} + \sum_{j=1}^{a+b} \pi_{jt} q_{jit} + o_{it}, \tag{1}$$

with π_{0t} the intercept, π_{jt} the marginal price for the-j^{th} characteristic, and o_{it} an independent-identically-distributed (iid) error term for which the zero conditional mean assumption holds, i.e. the expected value of the error term does not depend on the $a+b$ characteristics. If b characteristics which are correlated with the a characteristics are omitted model (1) reduces to

$$p_{it} = \pi^{\bullet}_{0t} + \sum_{j=1}^{a} \pi^{\bullet}_{jt} q_{jit} + o^{\bullet}_{it}. \tag{2}$$

The estimators of the coefficients of the a house characteristics in model (2) will usually be biased and their variances will be incorrect (omitted variables bias, see e.g. Greene 2008). Note that some of the omitted variables may be constant over time (unobserved or individual heterogeneity).

Standard panel approaches to omitted variables bias are based on the assumption that the omitted house characteristics are constant over time. In that case the unobserved effects of the omitted variables can be removed by, for instance, differencing the data in

adjacent time periods and applying a standard estimator, particularly OLS, to the differences. Alternatively, fixed effects transformation can be applied. Under strict exogeneity of the explanatory variables (i.e. for each *t*, the expected value of the idiosyncratic error given the explanatory variables in all time periods is zero) this estimator is unbiased. Moreover, compared with first differencing, it is efficient. When the unobserved variables are uncorrelated with all the observed explanatory variables, they can be captured by the error term and the resulting serial correlation can be handled by generalized least squares estimation. When there is interest in the effects of time-invariant variables, instrumental variables may be used as an alternative to the fixed or random effects approaches (Hausman and Taylor 1981). [3]

When the omitted house characteristics vary over time (e.g. neighborhood characteristics), the above methods cannot be applied. For this case we propose the following alternative autoregression procedure.

Let[4]

$$\nu_t = \pi_{0t} + \sum_{j=a+1}^{a+b} \pi_{jt} q_{jt} + o_t \,. \tag{3}$$

i.e. the right hand side of (3) is equal to the intercept, plus the sum of the *a* omitted variables, plus the error term for which the zero conditional mean assumption applies. (The assumption implies that the error term is not correlated with the *a* observed systematic characteristics.)

[3] Note that the the spatial spatial error model introduced by Cliff and Ord (1969) arises because of spatially correlated omitted variables. We furthermore refer to Anselin and Gracia (2008) and Kelejian and Prucha (2007) who present nonparametric approaches towards estimating covariance matrices affected by omitted variables. The approach presented in this paper is different in that it accounts for omitted variables in the regression equation and thus addresses both omitted variables bias of the estimator of the regression coefficients and of the covariance matrix of the estimators.

[4] Since it is not needed for the remainder of this subsection, we suppress the index *i*.

Let π_{0t} include the expected value of the house characteristics captured by o_t.

Accordingly, the expected value of (3) is:

$$\mathrm{E}(\nu_t) = \pi_{0t} + \sum_{j=a+1}^{a+b} \pi_{jt}\, \mathrm{E}(q_{jt}) = \pi^*_{0t}. \tag{4}$$

Let

$$\nu^*_t = \nu_t - \pi^*_{0t}. \tag{5}$$

Combining (5) and (1) gives:

$$p_t = \pi^*_{0t} + \sum_{j=1}^{a} \pi_{jt} q_{jt} + \nu^*_t \tag{6}$$

or

$$\nu^*_t = p_t - \pi^*_{0t} - \sum_{j=1}^{a} \pi_{jt} q_{jt}\ . \tag{7}$$

We approximate the model of the time-varying omitted house characteristics including possible unobserved heterogeneity (equation 7) by the following first-order autoregression

$$\nu^*_t = \rho_{0t} + \rho_{1t} \nu^*_{t-1} + \upsilon_t\,, \tag{8}$$

with υ_t an iid error term.

Substituting the right hand of (7) into the left hand side of (8) and its lag into the right hand side for the $(T+1)$-waves $(t = 0,1,\ldots,T)$ of observations gives:

$$p_t - \pi^*_{0t} - \sum_{j=1}^{a} \pi_{jt} q_{jt} = \rho_{0t} + \rho_{1t}\left(p_{t-1} - \pi^*_{0t-1} - \sum_{j=1}^{p} \pi_{jt-1} q_{jt-1} \right) + \upsilon_t\,. \tag{9}$$

Finally, by rearranging (9) we obtain the following constrained autoregressive price model:

$$p_t = \left(\rho_{0t} + \pi^*_{0t} - \rho_{1t}\pi^*_{0t-1}\right) + \rho_{1t} p_{t-1} + \sum_{j=1}^{a} \pi_{jt} q_{jt} - \rho_{1t} \sum_{j=1}^{a} \pi_{jt-1} q_{jt-1} + \upsilon_t\,, \tag{10}$$

for $t = 1,\ldots,T$.

From (10) it follows that apart from the intercept, the omitted variables bias is corrected for by the difference between (i) the lagged dependent variable effect and (ii) the total effect of the lagged observed explanatory variables weighted by the autoregression coefficient. Note that the correction for omitted variables bias depends on the accuracy of the autoregression (8) to capture the effects of the omitted variables.

Without constraints on π^{*}_{0t-1} for $t = 1$, the ρ_{0t} for $t = 1,...,T$ and the π_{jt-1} for $t = 1$ and $j = 1,...,a$, the model is not identified. The simplest way to solve this identification problem for the first term is by combining the three intercept components in (10) into a single parameter π^{**}_{0t} . We do not have to identify those intercept components, instead of a single intercept π^{**}_{0t} for $t = 1,\ldots,T$. Hence, (10) reduces to

$$p_t = \pi^{**}_{0t} + \rho_{1t} p_{t-1} + \sum_{j=1}^{a} \pi_{jt} q_{jt} - \rho_{1t} \sum_{j=1}^{a} \pi_{jt-1} q_{jt-1} + \upsilon_{t_i}, \text{ for } t = 1,\ldots,T . \qquad (11)$$

For the identification of π_{jt-1} for $t = 1$ and $j = 1,...,a$ in the fourth term, we may impose equality constraints for π_{jt} over t. Note that equality of the π_{jt}'s over time means that the marginal prices are constant over time.

2.2. SEM

Now we turn to measurement errors in explanatory variables. A standard econometric approach to measurement error bias is instrumental variables (Greene 2008). However, obtaining adequate instruments may be difficult. In addition, the adequacy of the instrument can usually not be empirically validated (Verbeek 2000). Fuller (1986) suggests alternatives to the instrumental variables approach. However, these alternatives also require external information, i.e. information on measurement error variances. In

addition, these approaches become quite complicated if there is measurement error in more than one explanatory variable. Due to these difficulties, in practice (spatial) econometricians tend to ignore the measurement error problem (Verbeek 2000).

We propose structural equation modeling (SEM) with latent variables to handle measurement error in the explanatory variables. A SEM (Jöreskog 1973; Jöreskog and Sörbom 1996) can be specified in different ways, particularly, by different numbers of parameter matrices (Oud and Delsing 2010). Here we specify SEM by the following three equations with twelve parameter matrices and vectors:[5]

$$\boldsymbol{\eta} = \boldsymbol{\alpha} + \boldsymbol{\Gamma}\boldsymbol{\xi} + \mathbf{B}\boldsymbol{\eta} + \zeta, \text{ with } \operatorname{cov}(\zeta) = \boldsymbol{\Psi} \tag{12}$$

and

$$\mathbf{x} = \boldsymbol{\tau}_x + \boldsymbol{\Lambda}_x\boldsymbol{\xi} + \boldsymbol{\delta}, \text{ with } \operatorname{cov}(\boldsymbol{\delta}) = \boldsymbol{\Theta}_\delta, \tag{13}$$

$$\mathbf{y} = \boldsymbol{\tau}_y + \boldsymbol{\Lambda}_y\boldsymbol{\eta} + \boldsymbol{\varepsilon}, \text{ with } \operatorname{cov}(\boldsymbol{\varepsilon}) = \boldsymbol{\Theta}_\varepsilon. \tag{14}$$

In the structural equation (12), $\boldsymbol{\eta}$ and $\boldsymbol{\xi}$ are vectors of latent endogenous and exogenous variables, respectively. The vector $\boldsymbol{\alpha}$ contains the intercepts for the endogenous latent variables. The matrix $\boldsymbol{\Gamma}$ specifies the relationships between the latent exogenous and the latent endogenous variables and the matrix $\mathbf{B}$ the relationships among the latent endogenous variables. The vector ζ contains structural equation errors with covariance matrix $\boldsymbol{\Psi}$. They are assumed to be uncorrelated with $\boldsymbol{\xi}$. The expectation vector and covariance matrix of $\boldsymbol{\xi}$ are $\boldsymbol{\kappa}$ and $\boldsymbol{\Phi}$, respectively.

In the measurement equations (13) and (14), the vector $\mathbf{x}$ and $\mathbf{y}$ contain observed exogenous and endogenous variables, respectively. The former are indicators of the exogenous latent variables and the latter of the endogenous latent variables. The

[5] The standard SEM model, as in *inter alia* Jöreskog and Sörbom (1996), explains covariance structures in terms of eight parameter matrices. The inclusion of the means requires four additional parameter vectors (see e.g. Jöreskog and Sörbom 1996).

relationships between the observed variables $\mathbf{x}$ and $\mathbf{y}$ and their respective latent variables $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ are specified in the loading matrices $\boldsymbol{\Lambda}_x$ and $\boldsymbol{\Lambda}_y$, respectively. The elements of $\boldsymbol{\tau}_x$ and $\boldsymbol{\tau}_y$ are the intercepts of the measurement models. $\boldsymbol{\delta}$ and $\boldsymbol{\varepsilon}$ are the measurement error vectors with covariance matrices $\boldsymbol{\Theta}_\delta$ and $\boldsymbol{\Theta}_\varepsilon$, respectively. Often, but not necessarily, these covariance matrices are specified diagonal. The elements of $\boldsymbol{\delta}$ and $\boldsymbol{\varepsilon}$ are assumed uncorrelated with their corresponding latent variables.

The measurement models decompose the variance of an observed variable into the variance explained by the latent variable(s) and the variance of the corresponding measurement error.[6] Hence, the parameters of the structural model are estimated such that the variances of the latent variables are free from measurement errors and hence are not attenuated. In addition, multicollinearity is mitigated by subsuming highly correlated variables under one and the same latent variable in the structural model (Oud and Folmer 2008).

Several estimators have been developed for SEM, i.e. unweighted least squares, generalized least squares, instrumental variables, generally weighted least squares, diagonally weighted least squares and maximum likelihood (ML). Here, we restrict ourselves to the ML estimator which minimizes the fit function (15) in terms of the unknown parameters in the parameter matrices in (12)-(14) for a given data set $\mathbf{Z}$ of N observations on observed variables vector $\mathbf{z} = (\mathbf{y} \quad \mathbf{x})$:

$$\ell(\boldsymbol{\omega}|\mathbf{Z}) = \ln|\boldsymbol{\Sigma}| + \mathrm{tr}(\mathbf{S}\boldsymbol{\Sigma}^{-1}) - \log|\mathbf{S}| - T(a+1) + (\bar{\mathbf{z}} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\bar{\mathbf{z}} - \boldsymbol{\mu}). \qquad (15)$$

where $\boldsymbol{\omega}$ is the vector of unknown parameters in $\boldsymbol{\Gamma}$, $\mathbf{B}$, $\boldsymbol{\Phi}$, $\boldsymbol{\Psi}$, $\boldsymbol{\Lambda}_x$, $\boldsymbol{\Lambda}_y$, $\boldsymbol{\Theta}_\delta$, $\boldsymbol{\Theta}_\varepsilon$, $\boldsymbol{\alpha}$, $\boldsymbol{\kappa}$, $\boldsymbol{\tau}_x$ and $\boldsymbol{\tau}_y$, $\mathbf{S}$ is the sample covariance matrix, $\bar{\mathbf{z}}$ is the sample mean vector of

[6] An indicator may be related to more than one latent variable.

$T(a+1)$ observed variables (including the unit vector). $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu}$ are the model-implied covariance matrix and means vector (Jöreskog et al. 2006):

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Lambda}_y \mathbf{B}_* \left(\boldsymbol{\Gamma}\boldsymbol{\Phi}\boldsymbol{\Gamma}' + \boldsymbol{\Psi}\right)\mathbf{B}_*'\boldsymbol{\Lambda}_y' + \boldsymbol{\Theta}_\varepsilon & \boldsymbol{\Lambda}_y \mathbf{B}_* \boldsymbol{\Gamma}\boldsymbol{\Phi}\boldsymbol{\Lambda}_x' \\ \boldsymbol{\Lambda}_x \boldsymbol{\Phi}\boldsymbol{\Gamma}'\mathbf{B}_*'\boldsymbol{\Lambda}_y' & \boldsymbol{\Lambda}_x \boldsymbol{\Phi}\boldsymbol{\Lambda}_x' + \boldsymbol{\Theta}_\delta \end{bmatrix} \text{ and} \tag{16a}$$

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\tau}_y + \boldsymbol{\Lambda}_y \mathbf{B}_* \left(\boldsymbol{\alpha} + \boldsymbol{\Gamma}\boldsymbol{\kappa}\right) \\ \boldsymbol{\tau}_x + \boldsymbol{\Lambda}_x \boldsymbol{\kappa} \end{bmatrix} \text{ with } \mathbf{B}_* = \left(\mathbf{I} - \mathbf{B}\right)^{-1}. \tag{16b}$$

The ML method assumes that the sample is drawn from a multivariate normal distribution. Application of ML, if the distribution actually deviates from normality, may be defended on the basis of the fact that the effect of non-normality on the estimator is often negligible (Finch et al. 1997). Note that ML is robust, and usually is an estimator with acceptable properties for a wide class of distributions (Chou et al. 1991). However, in the case of deviation from normality, the standard errors produced by SEM programs should be interpreted with caution (Bentler 1989; Bollen 1989).[7]

Before estimating a SEM, identification should be checked, i.e. whether each parameter is uniquely determined by the population covariance matrix and means vector. The presence of latent variables requires that their measurement scales are fixed to render the model identified. This can be done by fixing for each latent variable one of its factor loadings or by fixing the latent variance. Fixing a loading at 1 implies that the latent variable gets the same scale as the corresponding indicator. Fixing the variance of a latent variable (usually at 1), gives a standardized latent variable. In a cross-sectional SEM analysis, identification requires that each latent variable has at least two indicators (Bollen 1989). However, in the case of panel data, identification can be achieved by way of one indicator under the restrictions that its loadings and error variances are the same

[7] When the sample size increases, the asymptotic properties of the ML estimator start becoming effective and the impacts of deviation from normality start decreasing. Nevertheless, under non-normality, as reflected by amongst others the skewness and kurtosis of the data, and showing up in implausibly large standard errors, one may turn to robust standard error estimates (Jöreskog et al. 2000) or the bootstrap.

over time (Jöreskog and Sörbom 1996). Repeated measurement thus allows decomposing the variance of an observed, repeatedly measured, variable into the variance of the latent variable across time and the measurement error variance.

In addition to the scale condition, the usual order and rank identification conditions for systems of equations apply. In large models these conditions are tedious to check. However, indications about unidentified parameters can be obtained from the information matrix. If all parameters are identified, then this matrix will almost always be positive definite. Most SEM programs like LISREL 8 and Mx provide warnings for non-identification (i.e. a singular information matrix).

There exist several statistical tests and fit indices for SEM models, particularly z statistics for the individual parameters and the chi-square statistic for the overall model fit. The chi-square test is sensitive to deviation from normality and to sample size, however. Therefore, a poor chi-square statistic does not necessarily indicate poor model fit (Jöreskog and Sörbom 1996). Most SEM programs provide several alternative fit indices, particularly the Root Mean Square Error of Approximation (RMSEA). Browne and Cudeck (1993) show that values of RMSEA up to 0.08 indicate a reasonable overall model fit. The fit in the sense of high predictability of individual structural and measurement equations can be evaluated by means of R^2 values. A poorly fitting model can be improved by using the modification indices. Such an index, reported by several SEM programs, is the expected decrease in the chi-square value, if the corresponding constrained or fixed parameter is 'freed'. Absence of modification indices larger than 8 is another indication of reasonable model fit (Jöreskog and Sörbom 1996).

3. An ASEM Housing HP for Urban Indonesia

3.1. Conceptual Model

We now apply ASEM to estimate an HP housing model for urban Indonesia. As a first step, we present the conceptual model, i.e. the endogenous and exogenous variables and the relationships among them. We will use the SEM notation as introduced in (12)-(14).

The standard dependent variable in HP models is actual house price or rent. However, in Indonesia, as in many other developing countries, appropriate actual house price or rent data is difficult to obtain. Often recorded sales in the real estate market do not reflect actual prices due to, for example, high transaction costs. Similarly for rent which is often artificially low due to rent control. When recorded sales or rents are absent or unreliable, one may resort to owner appraisal data (e.g. Anselin et al. 2008; North and Griffin 1993; Yusuf and Koundouri 2004, 2005). In the present paper we follow this procedure. That is, the dependent variable is house owner monthly *Rent Appraisal* (see also Yusuf and Koundouri 2005). It will be explained by the following nine observed explanatory variables measuring three latent variables.

The first explanatory variable is *Median Household Monthly Expenditure on food and nonfood in the neighborhood* (abbreviated as *Median Household Monthly Expenditure*) where the house is located. It (or a proxy for it) is included in most housing HP studies to indicate the socio-economic status of the neighborhood (e.g. Dinan 1989; Kim et al. 2003; McMillen 2004; Yusuf and Koundouri 2004, 2005). Note, that *Median Household Monthly Expenditure* is a neighborhood indicator that ignores many important neighborhood characteristics such as environmental quality (Minguez et al. 2012), safety, quality of schools, presence of amenities, accessibility, quality of surrounding neighborhoods and so on. In addition, it is not so much the objective statuses of these

characteristics that affect an individual's evaluation but rather their perceptions (Minguez et al. 2012; Tang et al. 2013). Finally, these characteristics tend to change over time. Hence, neighborhood characteristic is a typical case of time-varying omitted variables.

The next explanatory variable is *Floor Area* which is also used in most hedonic housing models (e.g. Arimah 1992; Chattopadhyay 1999; Kim et al. 2003; Tyrväinen 1997). The third explanatory variable is the composite variable *House Condition*. It is the sum of seven dummy variables representing the following house characteristics: *Floor* and *Wall Material* (see e.g. Arimah 1992; Jimenez 1982; McMillen 2004; North and Griffin 1993; Tyrväinen 1997), presence of one or more *Toilets* (e.g. Arimah 1992; Gross 1988; Jimenez 1982; Tiwari and Parikh 1997), *Sewage Connection* (Engle 1985), *Electricity Connection* (Arimah 1992; Tiwari and Parikh 1997), *Piped Water Connection* and availability of *Well Water* (Anselin et al. 2008; Arimah 1992; Tiwari and Parikh 1997; Yusuf and Koundouri 2004). If *Floor Material* is ceramic or tiles, it takes the value 1, otherwise it is 0. If *Wall* material is brick or cement, it takes the value 1, otherwise 0. The dummy variables for presence of a *Toilet*, *Well Water*, *Piped Water*, *Sewage* and *Electricity* connection are defined in the usual way: 1 for presence and 0 for absence. The composite variable takes values between 0 and 7. We combine the dummies, which are strongly correlated, into a composite variable to reduce multicollinearity.[8] In addition, we do not specify them as reflective indicators[9] of a latent variable *House Condition*, since, conceptually, they are formative indicators. Specifying them as reflective indicators would lead to bias due to causality misspecification (Diamantopoulos et al. 2008). We hypothesize that all three explanatory variables *Median Household Monthly Expenditure*,

[8] The composite variable represents the number of positive house attributes. Its coefficient is the average marginal price for an additional attribute, or improvement in one of the house materials.

[9] Indicators can be categorized on the basis of the causal relationships to their latent constructs. A reflective indicator is the effect of a latent construct; a formative indicator is the cause (Bollen 1989, pp 64-65).

Floor Area and *House Condition* have positive impacts on *Rent Appraisal*, since they reflect house quality dimensions.

3.2. Data

We analyze the Indonesia Family Life Survey (IFLS)[10] data set to estimate the housing HP model. The longitudinal data set consists of three waves. The first (IFLS1) was administered in 1993. For IFLS2 and IFLS3 the same respondents were re-interviewed in 1997 and 2000, respectively (Strauss et al. 2004).[11] The study covers a panel of 2259 non-rented houses in urban areas. Due to attrition, the effective panel size is 1562. [12]

Attrition may lead to bias when the missingness depends on the values of variables in the data set (Little 1988; Verbeek and Nijman 1992). To test for this, we apply Little's (1988) test. The null hypothesis is that the missing values are missing completely at random (MCAR), i.e. that the missingness does not depend on the values of variables in the data set. We obtained a chi-square value of 18.8 with 9 degrees of freedom, corresponding to a significance level of $p = 0.023$. Hence, the null hypothesis should be rejected at the 5%-level. Note, however, that in the case of large samples the test tends to reject the null hypothesis when it is true (Morrison 2004). Since our sample size is very large and the *p*-value is not far from 5%, we accept the null hypothesis and conclude that the missing values are MCAR. For SEM, this implies that the available-case method can be applied (Little and Rubin 1987). This approach uses the sample units

[10] The IFLS is a longitudinal socio-economic and health survey of Indonesian individuals and households. It was conducted by the RAND Institute (Strauss et al. 2004).

[11] The data set relates to urban and rural residents. In this paper we analyze the former only.

[12] The effective sample size is the number of sample units with complete measurement, i.e. without missing values.

with a complete set of observations, i.e. observations for all waves, for the calculation of the means, variances and covariances as input for the analysis.

Table 2.1. Descriptive statistics

Variables	Urban Indonesia		
	Wave-1	Wave -2	Wave- 3
Rent Appraisal (Rp100,000)*	0.55	0.72	0.57
	(0.90)	(1.03)	(1.12)
Median Household Monthly Expenditure (Rp100,000) *	2.20	2.78	2.33
	(1.02)	(1.35)	(1.02)
Floor area ($10m^2$)	5.49	7.15	7.38
	(3.76)	(3.96)	(4.63)
House Condition	3.89	4.31	4.50
	(1.68)	(1.52)	(1.41)
floor (1=ceramic/tiles, 0=others)	0.39	0.48	0.48
	(0.49)	(0.50)	(0.50)
wall (1=brick/cement, 0=others)	0.71	0.77	0.80
	(0.46)	(0.42)	(0.40)
toilet (1=yes, 0=no)	0.67	0.75	0.79
	(0.47)	(0.43)	(0.41)
sewer (1=yes, 0=no)	0.76	0.77	0.80
	(0.43)	(0.42)	(0.40)
electricity (1=yes, 0=no)	0.90	0.98	0.99
	(0.30)	(0.15)	(0.12)
piped water (1=yes, 0=no)	0.17	0.23	0.28
	(0.38)	(0.42)	(0.45)
well water (1=yes, 0=no)	0.29	0.34	0.36
	(0.45)	(0.47)	(0.48)

Notes: *in 1993 values; Standard deviations in brackets

The variables, their means and standard deviations (in brackets) are presented in Table 2.1. Observe that the means of *Rent Appraisal* and *Median Household Monthly Expenditure* in wave 3 are lower than in wave 2. This is due to the economic crisis which occurred at the end of 1997, just after the IFLS2 survey was completed.

3.3. ASEM specification

Before going into detail, we make the following observations. First, we use a linear model with *Rent Appraisal*, *Median Household Monthly Expenditure* and *Floor Area* measured in natural logarithms (log). The choice of a linear model (in the present case SEM) is supported by Cropper et al. (1988), who suggest that a linear HP model consistently outperforms alternative functional forms, particularly the quadratic Box-Cox model, when some variables are not observed or replaced by proxies. Secondly, since each observed variable in each wave is measured by the same question or questions, we assume that the reliabilities of the measurements and the functional relationships between each indicator and its corresponding latent variable are equal over the waves. For a given measurement equation this assumption implies that the same measurement model applies to the three waves. Hence, a given λ and the variance of a given δ or ε are taken equal over the waves (*constraint 1*). Thirdly, following Flores and Carson (1997), we assume that the valuation of the housing attributes is proportional to income. This assumption translates into linear equality constraints such that for each observation the coefficients of the explanatory variables in equation (11) are proportional to the mean of *Household Monthly Expenditure* (*constraint 2*). (The constraints are presented below.) This constraint increases the efficiency of the estimator by increasing the degrees of freedom. Finally, to economize on the degrees of freedom and to enable identification of the measurement error variance of a single indicator measurement equation, we define auxiliary autoregressive models for the three latent explanatory variables (log(*Median Household Monthly Expenditure*), log(*Floor Area*) and *House Condition*).[13] The auxiliary

[13] Without the auxiliary autoregressions, all of the variances and covariances of a single indicator over time are used for identification of variances and covariances of its latent variable over time. By specifying the auxiliary autoregressions, the latent variables beyond the initial time period become endogenous and the parameters related to them are the autoregressive parameters and error model variances only. For instance, with three observations over time, there are three different variances and three different covariances of,

autoregressive model renders the measurement error variance of a single indicator identified, since it adds the covariances of this indicator over different waves into the identification process of the measurement error variance (see Jöreskog and Sörbom 1996, pp 230-234). Observe that these autoregressions turn the house characteristic variables at $t = 1,2$ in (11) into endogenous variables.

We now present ASEM by first specifying the above variables of the HP housing model in SEM terms. Note that log(*Rent Appraisal*) at wave-0 in (11) is a predetermined endogenous variable. Thus, we denote log(*Rent Appraisal*), log(*Median Household Monthly Expenditure*), log(*Floor Area*), and *House Condition* at wave-0 as x_1, x_2, x_3, and x_4, respectively, and the corresponding latent variables as ξ_1, ξ_2, ξ_3, and ξ_4. For wave-1, we denote log(*Rent Appraisal*), log(*Median Household Monthly Expenditure*), log(*Floor Area*), and *House Condition* as y_1 - y_4 and in wave-2 as y_5 - y_8. The corresponding latent variables are denoted η_1 - η_4 and η_5 - η_8, respectively. Accordingly, the exogenous and endogenous observed and latent vectors are $\mathbf{x}' = [x_1 \quad \cdots \quad x_4]$, $\mathbf{y}' = [y_1 \quad \cdots \quad y_8]$, $\boldsymbol{\xi}' = [\xi_1 \quad \cdots \quad \xi_4]$ and $\boldsymbol{\eta}' = [\eta_1 \quad \cdots \quad \eta_8]$.

The structural model consists of the constrained autoregression (11) with the observed variables replaced by the latent variables in $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ and the three autoregressive models of the house characteristics. The parameter matrices of the effects of $\boldsymbol{\xi}$ on $\boldsymbol{\eta}$ ($\boldsymbol{\Gamma}$) and of the effects among the $\boldsymbol{\eta}$-variables mutually ($\mathbf{B}$) are:

---

say, *House condition* which can be used to identify six SEM parameters. For the auxiliary autoregression, however, only four of the six variances plus covariances are needed (i.e. two autoregressive parameters and two error term variances). Hence, there are two moments left that are available for identification of the time invariant measurement error variance of the observed *House Condition* at the three time points.

$$\boldsymbol{\Gamma} = \begin{bmatrix} \gamma_{11} & \gamma_{12} & \gamma_{13} & \gamma_{14} \\ 0 & \gamma_{22} & 0 & 0 \\ 0 & 0 & \gamma_{33} & 0 \\ 0 & 0 & 0 & \gamma_{44} \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \text{ and } \mathbf{B} = \begin{bmatrix} 0 & \beta_{12} & \beta_{13} & \beta_{14} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \beta_{51} & \beta_{52} & \beta_{53} & \beta_{54} & 0 & \beta_{56} & \beta_{57} & \beta_{58} \\ 0 & \beta_{62} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \beta_{73} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \beta_{84} & 0 & 0 & 0 & 0 \end{bmatrix}$$

The first row of $\boldsymbol{\Gamma}$ represents the impacts of log(*Rent Appraisal*), log(*Median Household Monthly Expenditure*), log(*Floor Area*), and *House Condition* at wave-0 on log(*Rent Appraisal*) at wave-1 (lagged coefficients), and the first row of $\mathbf{B}$ the impacts of log(*Median Household Monthly Expenditure*), log(*Floor Area*), and *House Condition* at wave-1 on log(*Rent Appraisal*) at wave-1 (current coefficients). The second row of $\boldsymbol{\Gamma}$ represents the autoregression coefficient of log(*Median Household Monthly Expenditure*) at wave-1; the third and fourth rows the autoregression coefficients of log(*Floor Area*), and *House Condition* at wave-1, respectively. Row 5, columns 1-4, of matrix $\mathbf{B}$, contains the impacts on log(*Rent Appraisal*) at wave-2 of log(*Rent Appraisal*), log(*Median Household Monthly Expenditure*), log(*Floor Area*), and *House Condition* at wave-1 (lagged coefficients), and columns 6-8 those of current log(*Median Household Monthly Expenditure*), log(*Floor Area*), and *House Condition* (current coefficients). Rows 6-8 of $\mathbf{B}$ contain the autoregression coefficients at wave-2 of log(*Median Household Monthly Expenditure*), log(*Floor Area*) and *House Condition* at wave-2, respectively.

Equation (11) specifies constraints between the coefficients of the lagged house characteristics and the autoregressive coefficient of the dependent variable log(*Rent Appraisal*). These autoregression constraints translate into constraints on the coefficients in the matrices $\boldsymbol{\Gamma}$ and $\mathbf{B}$. To specify the constraints for wave-1 we would need the wave-0 coefficients which, however, are not estimated. To obtain them, we use proportional

Constraint 2 which translates into $0.7892\beta_{1k}$ for $k = 2,3,4$ where 0.7892 is the wave-0/wave-1 household expenditure means ratio. Accordingly for the wave-1 model, the autoregression constraint translates into $\gamma_{1k} = -\gamma_{11}(0.7892\beta_{1k})$ for $k = 2,3,4$. For the wave-2 model, the autoregression constraint translates into $\beta_{5k} = -\beta_{51}\beta_{1k}$ for $k = 2,3,4$.

We furthermore apply *Constraint* 2 to improve efficiency by reducing the number of free parameters. For that purpose we specify constraints among the wave-1 and wave-2 coefficients as follows: $\beta_{1k} = 1.1604\beta_{5l}$ for $(k,l) = \{(2,6),(3,7),(4,8)\}$ with 1.1604 the wave-1/wave-2 household expenditure means ratio. In addition, we specify and test the constraint that the coefficients are time-invariant. For that purpose we specify a model with time-invariant coefficients, i.e. we replace *Constraint* 2 by the identity equality constraints $\gamma_{1k} = -\gamma_{11}(\beta_{1k})$ for wave-1 and $\beta_{1k} = \beta_{5l}$ for wave 2.

The intercepts of the structural equations are given in the vector $\boldsymbol{\alpha}' = [\alpha_1 \quad \cdots \quad \alpha_8]$, and the variances of the structural errors in the diagonal matrix $\boldsymbol{\Psi} = \text{diag}[\psi_{11} \quad \cdots \quad \psi_{88}]$. It is possible to relax the constraint of a diagonal matrix to allow covariances among the errors.

The means and covariances of the exogenous latent variables are defined in the parameter vector $\boldsymbol{\kappa}' = [\kappa_1 \quad \cdots \quad \kappa_4]$ and in the symmetric covariance matrix

$$\boldsymbol{\Phi} = \begin{bmatrix} \phi_{11} & & & \\ \phi_{21} & \phi_{22} & & \\ \phi_{31} & \phi_{32} & \phi_{33} & \\ \phi_{41} & \phi_{42} & \phi_{43} & \phi_{44} \end{bmatrix}.$$

As pointed out above, all of the latent variables are indicated by a single observed variable, measured three times. Thus, each observed variable serves as a reference variable for the underlying latent variable which results in a unit loading and zero intercept for each observed variable. Accordingly, we have the loading matrix $\boldsymbol{\Lambda}_x = \mathbf{I}_{(4\times4)}$

and $\boldsymbol{\Lambda}_y = \mathbf{I}_{(8\times 8)}$ and the intercept vector $\boldsymbol{\tau}_x = \mathbf{0}_{(4)}$ and $\boldsymbol{\tau}_y = \mathbf{0}_{(8)}$. Furthermore, the measurement error covariance matrices are the diagonal matrix $\boldsymbol{\Theta}_\delta = \text{diag}\begin{bmatrix} 0 & \theta_{22}^{\delta} & \theta_{33}^{\delta} & \theta_{44}^{\delta} \end{bmatrix}$

and

$\boldsymbol{\Theta}_\varepsilon = \text{diag}\begin{bmatrix} 0 & \theta_{22}^{\varepsilon} & \theta_{33}^{\varepsilon} & \theta_{44}^{\varepsilon} & 0 & \theta_{44}^{\varepsilon} & \theta_{55}^{\varepsilon} & \theta_{66}^{\varepsilon} \end{bmatrix}$.

Constraint 1 translates into $\theta_{kk}^{\delta} = \theta_{kk}^{\varepsilon} = \theta_{ll}^{\varepsilon}$ for $(k,l) = \{(2,6),(3,7),(4,8)\}$. Note that the variances of the log(*Rent Appraisal*) measurement errors need not be estimated, since they will not influence the estimator of the coefficients. They are absorbed in the structural model error variance (Greene 2008).

In addition to ASEM, we estimate a constrained autoregression HP (AUT), a SEM HP without omitted variables correction (SEM) and a fixed effect panel HP model (FE). Comparison of ASEM and AUT provides information about the consequences of measurement errors; comparison of ASEM and SEM about omitted variables bias. FE is the usual procedure applied in practice when analyzing panel data. Therefore, comparison of ASEM and FE provides information about the bias due to ignoring measurement error and time-varying omitted variables. Note that the comparisons merely give partial information. To gain full insight Monte Carlo simulations are required. The specifications of the three alternative models are given in Appendix 2.1.

3.4. Empirical results

ASEM and the alternative models are estimated by means of the ML estimator in the LISREL 8 software program (Jöreskog and Sörbom 1996). We first discuss the two ASEM models (i.e. the model under *Constraint 2* and the model with time-invariant coefficients) and next compare ASEM under *Constraint 2* to the alternative models in a

bid to illustrate that ASEM adequately controls for omitted variables and measurement error. The results for the *Constraint 2* ASEM measurement model is presented in Table 2.2 and its structural model in Table 2.3. The results for the constant coefficients ASEM are given in Appendix 2.2.

The chi-square value for overall fit of ASEM under *Constraint 2* is 273.10 with 48 degree of freedom which gives an almost zero *p*-value. However, as mentioned in the preceding section, the Chi-square test is sensitive to deviation from normality and large sample size. Since the variables deviate from normality and the sample size is large, the chi- square test is not appropriate in this case. Instead, we use the Root Mean Square Error of Approximation (RMSEA). It equals 0.06 which is well below 0.08, which is usually taken as upper limit of a reasonable fit (Browne and Cudeck 1993). In addition, the matrix of modification indices[14] does not contain elements larger than 8 which is another indication of a reasonable fit. Finally, the R^2s of log (*Rent Appraisal*) are 0.70 for wave-1 and 0.76 wave 2, respectively.

The time-invariant coefficient ASEM (Appendix B) has slightly better fit indices than the *Constraint 2* ASEM (Table 2.3). The Chi-square value is 273.10 with 48 degrees of freedom, the RMSEA is 0.054 and the R^2s of the individual wave models are 0.70 for wave-1 and 0.77 for wave 2. However, compared to SEM, the time-invariant coefficient model produces an unreasonable omitted variable bias correction (see below for details). Therefore, we restrict the discussion below to the *constraint* 2 ASEM.

ASEM has a better RMSEA than its alternatives. AUT, SEM, and FE have RMSEA > 0.08, ASEM 0.06 only. Particularly, FE has a very poor overall fit. Furthermore, the ASEM R^2s are higher than those of its alternatives, except FE. For all models, except FE, the modification indices are smaller than 8. Especially the

[14] The matrices of modification indices are available at http://blogs.unpad.ac.id/yusepsuparman/

modification indices of the structural parameters related to log(*Rent Appraisal*) are small and in most cases equal to zero.

In Table 2.2, we present the estimated measurement error variances, which, as observed above, are constrained to be equal over the waves. In addition, for each observed variable we present its reliability R^2 (defined as one minus (measurement error variance divided by the observed variable variance)). These statistics show that the indicators are highly reliable with R^2s > 0.95.

The ASEM structural parameter estimates are presented in Table 2.3, column 1. Due to *Constraint 2* the parameter estimates and their estimated standard errors for $t = 2$ are proportional to the ones for $t = 1$ while their z-values are equal. Accordingly, we only present the estimates for $t = 1$, i.e. β_{12} - β_{14} in Table 3.[15] All the estimates of β_{12} - β_{14} are significant at the 5%-level. The elasticity of log(*Median Household Monthly Expenditure*) is slightly larger than one (1.13) and of log(*Floor Area*) slightly smaller (0.9). Furthermore, *House Condition* has a positive impact. An increase by one unit leads to an increase of average price by 31.96%.

Table 2.2. The ASEM measurement model under *Constraint* 2

Variable	Measurement error variance	Reliability
log (*Median Household Monthly Expenditure*)	0.03* (0.00)	0.97
log (*Floor area*)	0.01 (0.01)	0.99
House condition	0.41* (0.02)	0.98

Note: The first line: estimate; the second (in brackets): standard error.
* significant at least at a 0.01 level for a two sided test.

[15] To economize on space, we do not present the estimates of the lagged coefficients. They are available at http://blogs.unpad.ac.id/yusepsuparman/.

Table 2.3. ASEM, AUT, SEM, and FE under *Constraint* 2

Variable	Dependent variable: log (*Rent Appraisal*) Model			
	ASEM	AUT	SEM	FE
Lagged log(*Rent Appraisal*)	0.26 (0.03)	0.28 (0.02)	n.a. n.a.	n.a. n.a.
log(*Median Household Monthly Expenditure*)	1.13 (0.06)	0.80 (0.05)	1.32 (0.05)	0.84 (0.04)
log(*Floor Area*)	0.09 (0.03)	0.11 (0.03)	0.12 (0.03)	0.11 (0.03)
House condition	0.32 (0.02)	0.26 (0.01)	0.33 (0.01)	0.27 (0.01)
Constant	-2.79 (0.12)	-2.29 (0.09)	-3.85 (0.07)	-3.09 (0.09)
1997's R^2	0.70	0.64	0.70	0.72
2000's R^2	0.76	0.73	0.74	0.78
RMSEA	0.06	0.16	0.08	0.56

Note: First line: estimate; second line (in brackets): standard error.
All parameters are significant at least at a 0.01 level for a right sided test. The estimate of a coefficient of a lagged variable in ASEM is he negative of the autoregression coefficient multiplied by the stimates of its current coefficient. They are not given here. The full sets of estimates are available at http://blogs.unpad.ac.id/yusepsuparman/.

We now compare ASEM and AUT, which both control for omitted variables, but differ with respect to accounting for measurement error. Table 3 shows that the AUT estimates of the impacts of log(*Median Household Monthly Expenditure*) and *House Condition* are 30% and 19% lower than the corresponding ASEM estimates. This result is in line with the expectation that controlling for measurement error reduces attenuation. For log(*Floor Area*) the AUT estimate is 15% higher than the ASEM estimate. This latter result is due to the fact that the biases of the AUT estimators of the coefficients of log(*Median Household Monthly Expenditure*) and *House Condition* also bias the estimator of the coefficient of log(*Floor Area*). This bias is a mixture of all the parameters in the model such that the sizes and even the directions are not easily derived

(Greene 2008). Hence, the AUT estimator of log(*Floor Area*) may have an upward bias and thus exceed the ASEM estimator.

Next, we compare ASEM and SEM to get insight into ASEM's capability of correcting omitted variable bias. Note that omitted variables like neighborhood characteristics such as accessibility, quality of schools, safety, etc. are positively correlated with the observed variables and have a positive impact on the dependent variable log(*Rent Appraisal*). Moreover, the observed house characteristics in the model are positively correlated. Accordingly, the omitted variables cause the estimators of the coefficients of the observed variables to be biased upwards. This is confirmed by the SEM estimates which are higher than the ASEM estimates. In particular, the SEM estimates of the coefficients of log(*Median Household Monthly Expenditure*), log(*Floor Area*), and *House Condition* are 17%, 33% and 4% higher than the corresponding ASEM estimates. Hence, *constraint 2* ASEM reduces biases due to omitted variables. In contrast, some estimates in the time-invariant coefficient ASEM and SEM do not confirm the omitted variable bias premise. Particularly, the SEM estimate for *House Condition* is lower than the ASEM estimate (Appendix B). This result provides some evidence that *constraint 2* ASEM is preferable to constant coefficient ASEM.

Finally, we compare ASEM and FE. As mentioned above, FE accounts for unobserved heterogeneity (omitted variables that are constant over time) but not for time-varying omitted variables. Nor does it control for measurement error. FE was estimated as a structural equation model with observed exogenous variables, and a latent variable representing unobserved heterogeneity with fixed coefficient equal to 1 (Bollen and Brand 2010).

We first note that the modification indices for the fixed coefficient of the unobserved heterogeneity variable were 7.08 for wave-1 and 28.29 for wave-3,

respectively.[16] These modification indices indicate that the individual effect is not constant over time (Bollen and Brand, 2010). Accordingly, the FE estimator is subject to omitted variables misspecification in addition to measurement error.

The FE estimate of the coefficient of log(*Floor Area*) is 20% larger than the ASEM estimate while the estimates of the coefficients of log(*Median Household Monthly Expenditure*) and *House Condition* are 26% and 16% lower than the ASEM estimates. These opposite outcomes are due to the opposite effects of measurement error and omitted variables. While the former leads to attenuation, the latter causes upward bias. In addition, both types of errors interact. Compared with the AUT estimates, FE improves the attenuated estimates of the coefficients of log(*Median Household Monthly Expenditure* and *House Condition*) but worsens the upward biased of the coefficient of log(*Floor Area*). These upward corrections can be explained from the expected change of the fixed coefficients of individual effect (i.e. -0.25 at wave-1, 0.06 at wave-2, and 0.47 at wave-3) if the parameters are freed. Their positive sum indicates that the total constant individual effect is underestimated. Since the individual effect is positively correlated with the other explanatory variables, unabsorbed total positive individual effect (under time-varying individual effect) is distributed among the other explanatory variables as positive changes from their respective in AUT. The ultimate outcome of FE, which depends on all the parameters in the model, may be downward bias for one set of coefficients and upward bias for another.

As noted above, to get a full insight into the biases of SEM, AUT and FE is by means of a simulation study.

[16] The full set of modification indices can be obtained at http://blogs.unpad.ac.id/yusepsuparman/.

4. Conclusions

Obtaining accurate estimates is of greatest importance in any empirical (spatial) analysis including hedonic price studies. Two frequently encountered problems are underspecification or omitted variables and measurement errors in explanatory variables. Both problems can lead to substantial bias whose size and even direction are not easily derived because they depend on all the parameters in the model. Ignoring either or both problems thus invalidates inference.

In practice both problems are frequently ignored. In this paper we present a constrained autoregression-structural equation model (ASEM) as a device that can routinely be applied to control for both types of misspecification. One important feature of ASEM is that it allows handling of time-variant missing variables and thus supplements standard econometric procedures like differencing that can be applied to time-invariant missing variables only. Another important characteristic is that ASEM requires no external information to handle measurement error. The application to urban Indonesia presented in this paper shows that omitted variables measurement and measurement errors in explanatory variables should be handled simultaneously, as done by ASEM.

The ASEM model presented here is a micro model that does not require spatial dependence to be accounted for. However, Oud and Folmer (2008) and Folmer and Oud (2008) show that SEM can be extended to control spatial spillover effects and to model spatial dependence as a latent variable, respectively.

References

Anselin L, Gracia NL (2008) Error in variables and spatial effects in hedonic house price models of ambient air quality. Empir Econ 34:5-34.

Anselin L, Gracia NL, Deichmann U, Lall S (2008) Valuing access to water: A spatial hedonic approach applied to Indian cities. Policy Research Working Paper, WPS4533, the World Bank.

Arimah BC (1992) An empirical analysis of the demand for housing aAttributes in a third world city. Land Econ 68:366-379.

Bentler PM (1989) EQS structural equations program manual. BMDP Statistical Software, Los Angeles.

Bollen KA (1989) Structural equation with latent variables. Wiley, New York.

Bollen KA, Brand JE (2010) A general panel model with random and fixed effects: A structural equations approach. Soc Forces 89:1-34.

Browne MW, Cudeck R (1993) Alternative ways of assessing model fit. In: Bollen KA, Long JS (Eds) Testing structural equation models. Sage, Newbury Park, pp 136-162.

Butler RV (1982) The specification of hedonic indexes for urban housing. Land Econ 58:96-108.

Chattopadhyay S (1999) Estimating the demand for air quality: New evidence based on the chicago housing market. Land Econ 75:22-38.

Chou CP, Bentler P, Satorra A (1991) Scaled test statistics and robust standard errors for non-normal data in covariance structure analysis: A Monte Carlo study. Br J Math and Stat Psychol 44:347–357.

Clarke KA (2005) The phantom menace: Omitted variable bias in econometric research. Confl Manag Peace Sci 22:341-352.

Cliff AD, Ord JK (1969) The problem of spatial autocorrelation. In: Scott AJ (Ed) London Papers in Regional Sciences 1, Studies in Regional Science. Pion, London, pp 25-55.

Cropper ML, Deck L, McConnell KE (1988) On the choice of functional forms for hedonic price functions. Rev Econ Stat 70:668-675.

Diamantopoulos A, Riefler P, Roth KP (2008) Advancing formative measurement models. J Bus Res 61:1203-1218.

Dinan TM, Miranowski JA (1989) Estimating the implicit price of energy efficiency improvement in the residential housing market: A hedonic approach. J Urban Econ 25:52-67.

Engle RF, Lilien DM, Watson M (1985) A dynamic model of housing price determination. J Econom 28:307-326.

Finch JF, West SG, MacKinnon DP (1997) Effects of sample size and nonnormality on the estimation of mediated effects in latent variable models. Struct Equ Model 2:87–105.

Flores NE, Carson RT (1997) The relationship between the income elasticities of demand and willingness to pay. J Environ Econ and Manag 33:287-295.

Folmer H (1986) Regional economy policy. Measurement of its effects. Kluwer, Dordrecht.

Folmer H, Oud JHL (2008). How to get rid of W: a latent variable approach to modeling spatially lagged variables. Environ Plan A 40:2526-2538.

Fuller WA (1986) Measurement error models. Wiley, New York.

Greene WH (2008) Econometric Analysis. Prentice Hall, New Jersey.

Gross DJ (1988) Estimating willingness to pay for housing characteristics: An application of the ellickson bid-rent model. J Urban Econ 24:95-112.

Gujarati DN (2004) Basic econometrics. McGraw-Hill, Singapore.

Hausman JA, Taylor W E (1981) Panel data and unobserved individual effects. Econom 49:1377-1398.

Hempel CG (1970) On the standard conception of scientific theories. In: Radner M, Winokur S (Eds) Minnesota Studies in the Philosophy of Science. University of Minnesota Press, Minneapolis.

Jimenez E (1982) The value of squatter dwellings in developing countries. Econ Dev Cult Chang 30:739-752.

Jöreskog KG (1973) A general method for estimating a linear structural equation system. In: Goldberger A S, Duncan O D (Eds) Structural equation model in the social sciences. Freeman, San Fransisco, pp 1-56.

Jöreskog K, Sörbom D (1996) LISREL 8: User's reference guide. Scientific Software International, Chicago.

Jöreskog K, Sörbom D, du Toit S, du Toit M (2000) LISREL 8: New statistical features. Scientific Software International, Chicago.

Keil KA, Zabel JE (1997) Evaluating the usefulness of the American housing survey for creating price indices. J Real Estate Financ Econ 14:189-202.

Kelejian HH, Prucha IR (2007) HAC estimation in a spatial framework. J Econom 140:131-154.

Kim CW, Phipps TT, Anselin L (2003) Measuring the benefit of air quality improvement: A spatial hedonic approach. J of Environ Econ Manag 45:24-39.

Little RJA (1988) A test of missing completely at random for multivariate data with missing values. J Am Stat Assoc, 83:1198-1202.

Little RJA, Rubin DB (1987) Statistical aAnalysis with missing data. Wiley, New York.

Malpezi S (2008) Hedonic pricing models: A selective and applied review. In: O'Sullivan T, Gibb K (eds) Housing economics and public policy. Wiley, New York, pp 67-89

McMillen DP (2004) Airport expansions and property values: The case of Chicago O'Hare Airport. J Urban Econ 55:627-640.

Minguez R, Montero JM, Fernandex-Avilez G (2012) Measuring the impact of pollution on property prices in Madrid: Objective versus subjective pollution indicators in spatial models. J Geogr Syst. doi:10.1007/s10109-012-016

Morrison DF (2004) Multivariate statistical methods. McGraw-Hill, New York.

North J, Griffin C (1993) Water source as a housing characteristic: Hedonic property valuation and willingness to pay for water. Water Resour Res 29:1923-1929.

Oud JHL, Delsing JMH (2010) Continuous time modeling of panel data by means SEM. In: van Montfort K, Oud JHL, Satorra A (Eds) Longitudinal research with latent variables. Springer, Heidelberg, pp 201-244.

Oud JHL, Folmer H (2008) A structural equation approach to models with spatially dependence. Geogr Anal 40:152-66.

Ozanne L, Malpezzi S (1985) The efficacy of hedonic estimation with the annual housing survey: Evidence from the demand experiment. J Econ Soc Meas 13:152-172.

Pace RK, LeSage JP (2004) Spatial statistics and real estate. J Real Estate Financ Econ 29:147-148.

Pace RK, LeSage JP (2010) Omitted Variable Biases of OLS and Spatial Lag Models. In: Páez A, Le Gallo J, Buliung R, Dallerba S (Eds.) Progress in spatial analysis: Methods and applications. Springer, Heidelberg, pp 17-28.

Páez A (2009) Recent research in spatial real estate hedonic analysis. J Geogr Syst 11:311-316.

Ridker RG, Henning JA (1967) The determinant of residential property value with special reference to air pollution. Rev Econ Stat 49:246-257.

Rosen S (1974) Hedonic prices and implicit markets: Product differentiation in price competition. J Polit Econ 82:34-55.

Strauss J, Beegle K, Sikoki B, Dwiyanto A, Herawati Y, Witoelar F (2004) The third wave of the Indonesia Family Life Survey (IFLS3): Overview and field report. Working paper, WR-144/1-NIA/NICHD, RAND.

Tang J, Folmer H, Xue J (2013) Estimation of awareness and perception of water scarcity among farmers in the Guangzhong Plain, China, by mean of a structural equation model. J Environ Manag (forthcoming).

Tiwari P, Parikh J (1997) Demand for housing in the Bombay metropolitan region. J Policy Model 19:295-321.

Tyrväinen L (1997) The aminity value of the urban forest: An application of the hedonic pricing method. Landsc Urban Plan 37:211-22.

Verbeek M (2000) A guide to modern econometrics. Wiley, Chichester.

Verbeek M, and Nijman T (1992) Testing for selectivity bias in panel data model. Int Econ Rev 33:681-703.

Waugh F (1926) Quality factor influencing vegetable price. J Farm Econ 10:185-196.

Yusuf AA, Koundouri P (2004) Household valuation of domestic water in Indonesia: revisiting the supply driven approach. In: Koundouri P (ed) Econometrics informing natural resources management: Selected empirical analyses. Edward Elgar, Cheltenham, pp 127-142.

Yusuf AA, Koundouri P (2005) Willingness to pay for water and location bias in hedonic price analysis: Evidence from the Indonesian housing market. Environ Dev Econ 10:821-836.

Appendix 2.1 Model Specifications

FE

In SEM notation the FE HP housing model reads as follows. For each wave log(*Rent Appraisal*) is the only endogenous variable while all the house characteristics are exogenous variables. Time-invariant unobserved heterogeneity is represented by the latent exogenous variable (ξ_{10}) which is correlated with the other exogenous variables and has a fixed unit regression coefficient in the three waves. The model does not account for measurement error, and hence the relationships between the observed and the latent variables are identity relationships. The measurement models thus read:

$$\mathbf{\Lambda}_y = \mathbf{I}_{(3\times3)},\ \mathbf{\Lambda}_x = \left[\mathbf{I}_{(9\times9)}\mathbf{0}_{(9)}\right],\ \boldsymbol{\tau}_y = \mathbf{0}_{(3)},\ \boldsymbol{\tau}_x = \mathbf{0}_{(9)},\ \mathbf{\Theta}_\varepsilon = \mathbf{0}_{(3\times3)} \text{ and } \mathbf{\Theta}_\delta = \mathbf{0}_{(9\times9)}.$$

The structural model parameter matrices are: $\boldsymbol{\alpha}' = \begin{bmatrix}\alpha_1 & \alpha_2 & \alpha_3\end{bmatrix}$, $\mathbf{B} = \mathbf{0}_{(3\times3)}$,

$$\mathbf{\Gamma} = \begin{bmatrix} \gamma_{11} & \gamma_{12} & \gamma_{13} & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & \gamma_{24} & \gamma_{25} & \gamma_{26} & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & \gamma_{37} & \gamma_{38} & \gamma_{39} & 1 \end{bmatrix}, \qquad \mathbf{\Psi} = \operatorname{diag}\begin{bmatrix}\psi_{11} & \psi_{22} & \psi_{33}\end{bmatrix},$$

$$\boldsymbol{\kappa}' = \begin{bmatrix}\kappa_1 & \cdots & \kappa_9 & 0\end{bmatrix} \text{ and } \Phi = \begin{bmatrix} \phi_{11} & & & \\ \phi_{21} & \phi_{22} & & \\ \vdots & \vdots & \ddots & \\ \phi_{10,1} & \phi_{10,2} & \cdots & \phi_{10,10} \end{bmatrix}.$$

Under *Constraint 2* $\gamma_{1k} = 0.7892\gamma_{2l}$ and $\gamma_{2l} = 1.1606\gamma_{3m}$ for $(k,l,m) = \{(1,4,7),(2,5,8),(3,6,9)\}$, while under the time-invariant coefficients assumption $\gamma_{1k} = \gamma_{2l}$ and $\gamma_{2l} = \gamma_{3m}$ for $(k,l,m) = \{(1,4,7),(2,5,8),(3,6,9)\}$.

SEM

The SEM HP structural model consists of the standard multiple regression model (2) in terms of latent variables, supplemented with the auxiliary autoregression models of

the house characteristics for identification of the measurement error variances. The exogenous variables in this model are the house characteristics in wave-0. The exogenous and endogenous observed and latent variables are $\mathbf{x}' = \begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix}$, $\mathbf{y}' = \begin{bmatrix} y_1 & \cdots & y_3 \end{bmatrix}$, $\boldsymbol{\xi}' = \begin{bmatrix} \xi_1 & \xi_2 & \xi_3 \end{bmatrix}$, $\boldsymbol{\eta}' = \begin{bmatrix} \eta_1 & \cdots & \eta_9 \end{bmatrix}$. Note that the observed log(*Rent Appraisal*) variables are y_1, y_2 and y_6, because there is no lagged dependent variable in the structural model. The structural parameter matrices are

$$\mathbf{B} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \beta_{23} & \beta_{24} & \beta_{25} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \beta_{67} & \beta_{68} & \beta_{69} \\ 0 & 0 & \beta_{73} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \beta_{84} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \beta_{95} & 0 & 0 & 0 & 0 \end{bmatrix} \quad \boldsymbol{\Gamma} = \begin{bmatrix} \gamma_{11} & \gamma_{12} & \gamma_{13} \\ 0 & 0 & 0 \\ \gamma_{31} & 0 & 0 \\ 0 & \gamma_{42} & 0 \\ 0 & 0 & \gamma_{53} \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad \boldsymbol{\alpha} = \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_9 \end{bmatrix},$$

$$\boldsymbol{\kappa} = \begin{bmatrix} \kappa_1 \\ \kappa_2 \\ \kappa_3 \end{bmatrix} \quad \boldsymbol{\Psi} = \operatorname{diag}\begin{bmatrix} \phi_{11} \\ \vdots \\ \phi_{99} \end{bmatrix} \text{ and } \boldsymbol{\Phi} = \begin{bmatrix} \phi_{11} & & \\ \phi_{21} & \phi_{22} & \\ \phi_{31} & \phi_{32} & \phi_{33} \end{bmatrix}.$$

Under C*onstraint 2*, $\gamma_{1k} = 0.7892\beta_{2l}$ and $\beta_{2l} = 1.1604\beta_{6m}$ for $(k,l,m) = \{(1,3,7),(2,4,8),(3,5,9)\}$, while under time-invariant assumption $\gamma_{1k} = \beta_{2l}$ and $\beta_{2l} = \beta_{6m}$ for $(k,l,m) = \{(1,3,7),(2,4,8),(3,5,9)\}$.

The parameter matrices in the measurement models are $\boldsymbol{\Lambda}_y = \mathbf{I}_{(9\times 9)}$, $\boldsymbol{\Lambda}_x = \mathbf{I}_{(3\times 3)}$, $\boldsymbol{\tau}_y = \mathbf{0}_{(9)}$, $\boldsymbol{\tau}_x = \mathbf{0}_{(3)}$, $\boldsymbol{\Theta}_\varepsilon = \operatorname{diag}\begin{bmatrix} 0 & 0 & \theta^{\varepsilon}_{33} & \theta^{\varepsilon}_{44} & \theta^{\varepsilon}_{55} & 0 & \theta^{\varepsilon}_{77} & \theta^{\varepsilon}_{88} & \theta^{\varepsilon}_{99} \end{bmatrix}$ and. $\boldsymbol{\Theta}_\delta = \operatorname{diag}\begin{bmatrix} \theta^{\delta}_{11} & \theta^{\delta}_{11} & \theta^{\delta}_{11} \end{bmatrix}$. *Constraint* 1 is $\theta^{\delta}_{kk} = \theta^{\varepsilon}_{ll} = \theta^{\varepsilon}_{mm}$ $(k,l,m) = \{(1,3,7),(2,4,8),(3,5,9)\}$.

AUT

The AUT model is (11). The endogenous variables in the model are log(*Rent Appraisal*) in wave-1 and 2, while all other variables are exogenous. The exogenous and endogenous observed and latent vectors are $\mathbf{x}' = [x_1 \quad \cdots \quad x_{10}]$, $\mathbf{y}' = [y_1 \quad y_2]$, $\boldsymbol{\xi}' = [\xi_1 \quad \cdots \quad \xi_{10}]$ and $\boldsymbol{\eta}' = [\eta_1 \quad \eta_2]$. Because of the absence of latent variables, the measurement models are $\boldsymbol{\Lambda}_y = \mathbf{I}_{(2\times 2)}$, $\boldsymbol{\Lambda}_x = \mathbf{I}_{(10\times 10)}$, $\boldsymbol{\tau}_y = \mathbf{0}_{(2)}$, $\boldsymbol{\tau}_x = \mathbf{0}_{(10)}$, $\boldsymbol{\Theta}_\varepsilon = \mathbf{0}_{(2\times 2)}$ and. $\boldsymbol{\Theta}_\delta = \mathbf{0}_{(10\times 10)}$.

The structural model matrices are $\boldsymbol{\alpha} = \begin{bmatrix} \alpha_1 \\ \alpha_1 \end{bmatrix}$, $\mathbf{B} = \begin{bmatrix} 0 & 0 \\ \beta_{21} & 0 \end{bmatrix}$,

$\boldsymbol{\Gamma} = \begin{bmatrix} \gamma_{11} & \gamma_{12} & \gamma_{13} & \gamma_{14} & \gamma_{15} & \gamma_{16} & \gamma_{17} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \gamma_{25} & \gamma_{26} & \gamma_{27} & \gamma_{28} & \gamma_{29} & \gamma_{210} \end{bmatrix}$, $\boldsymbol{\Psi} = \text{diag}[\psi_{11} \quad \psi_{22}]$,

$\boldsymbol{\kappa}' = [\kappa_{11} \quad \cdots \quad \kappa_{1010}]$, and $\boldsymbol{\Phi} = \begin{bmatrix} \phi_{11} & & & \\ \phi_{21} & \phi_{22} & & \\ \vdots & \vdots & \ddots & \\ \phi_{101} & \phi_{102} & \cdots & \phi_{1010} \end{bmatrix}$.

Under *Constraint 2* $\gamma_{1k} = 1.1604\gamma_{2l}$ for $(k,l) = \{(5,8),(6,9),(7,10)\}$, $\gamma_{1k} = -\gamma_{11}(0.7892\gamma_{1l})$ and $\gamma_{2l} = -\beta_{21}\gamma_{1l}$ for $(k,l) = \{(2,5),(3,6),(4,7)\}$. Under the time-invariant coefficients assumption $\gamma_{1k} = \gamma_{2l}$ for $(k,l) = \{(5,8),(6,9),(7,10)\}$, $\gamma_{1k} = -\gamma_{11}\gamma_{1l}$ and $\gamma_{2l} = -\beta_{21}\gamma_{1l}$ for $(k,l) = \{(2,5),(3,6),(4,7)\}$.

Appendix 2.2 Time-invariant coefficients models

ASEM, AUT, SEM and FE under the assumption of time-invariant coefficients.

Dependent variable: log(*Rent Appraisal*)

Variable	ASEM	AUT	SEM	FE
Lagged log(*Rent Appraisal*)	0.26	0.25	n.a.	n.a.
	(0.03)	(0.02)	n.a.	n.a.
log (*Median Household Monthly Expenditure*)	1.16	0.82	1.20	0.72
	(0.05)	(0.05)	(0.04)	(0.04)
log(*Floor area*)	0.09	0.10	0.10	0.09
	(0.03)	(0.03)	(0.02)	(0.03)
House condition	0.32	0.26	0.29	0.23
	(0.02)	(0.01)	(0.01)	(0.01)
Constant	-2.69	-2.27	-3.52	-3.09
	(0.12)	(0.08)	(0.06)	(0.09)
1997's R^2	0.70	0.63	0.68	0.79
2000's R^2	0.77	0.73	0.74	0.78
RMSEA	0.05	0.16	0.08	0.57

Note: First line: estimate; second line (in brackets): standard error.
All estimates are significant at least at a 0.01 level for a right sided test.
The estimate of a coefficient of a lagged variable in ASEM is the negative of the autoregression coefficient multiplied by the estimates of its current coefficient. The full sets of estimates are available at http://blogs.unpad.ac.id/yusepsuparman/