第五章 结构方程模型

Structural Equation Modeling

- A structural equation model (SEM) represents a natural extension of a measurement model and is a mature statistical modeling technique.
- It deals with difficult modeling challenges, including variables unobservable or latent and measured using exogenous variables, endogeneity among variables, and complex underlying social phenomena.
- SEMs lend themselves to graphical representations, and these graphical representations have become the standard means for presenting and communicating.

Structural Equation Modeling

- The SEM framework resolves potential problems by explicitly incorporation measurement errors into the modeling framework.
- The SEM model can accommodate a latent variable as a dependent variable, which can not be done in the traditional regression analysis.
- SEMs allow for direct, indirect, and associative relationships to be explicitly modeled, unlike ordinary regression, which implicitly model associations.
- SEMs rely on information contained in the variance-covariance matrix.

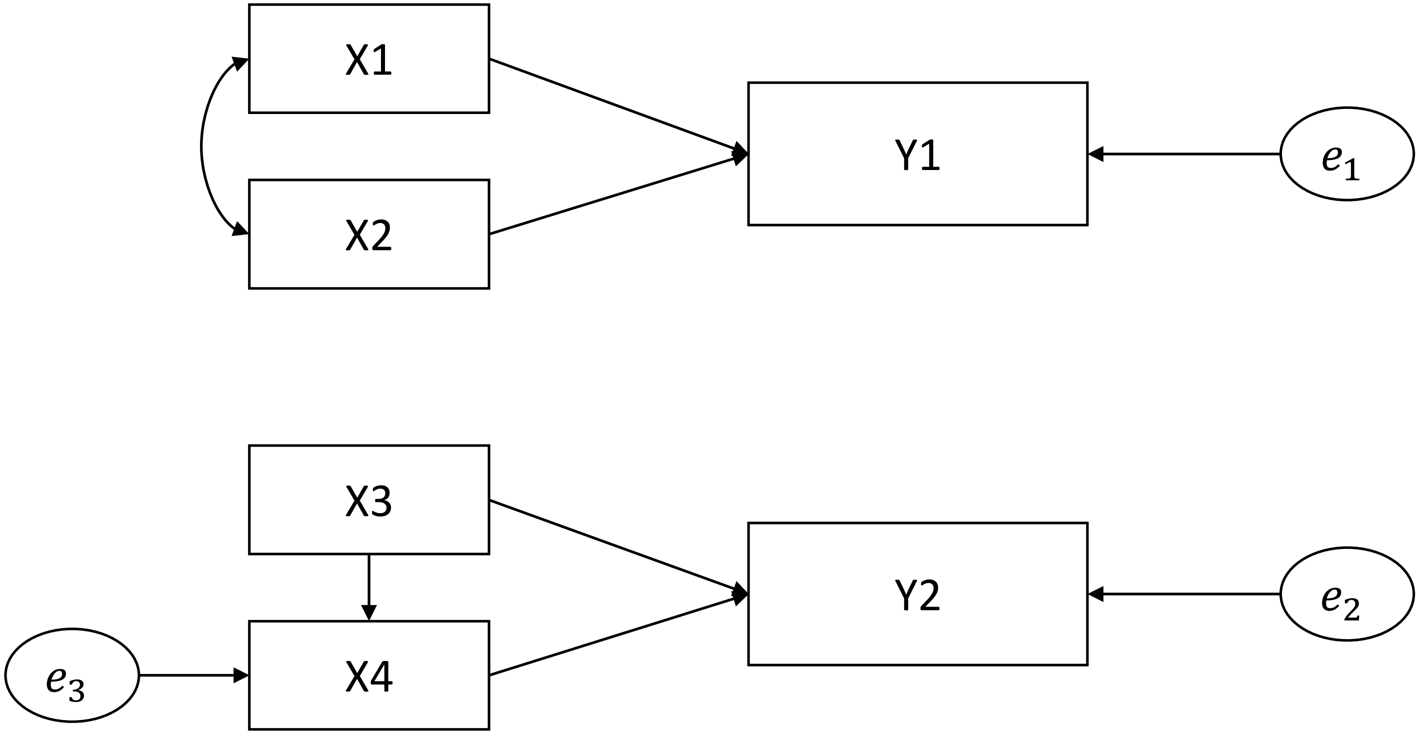
Basic concepts in SEM

- Structure: two components, a measurement model and a structural model.
- The measurement model: how well various measured exogenous variables measure latent variables & estimate of measurement errors of exogenous variables. (eg. A classical factor analysis)
- The structural model: how the model variables are related to one another. (similar to a system of simultaneous equations)

Basic concepts in SEM

- Two components: observed variables and unobserved variables. Observed variables are measured, whereas unobserved variables are similar to factors in factor analysis, which represent underlying unobserved constructs.
- Need to distinguish two parameters: fixed parameters are set by the analyst and free parameters are estimated from data. The collection of these two parameters will imply a variance-covariance matrix.

Graphical representation



SEM example

To examine simple path models, consider the relationships between an attitudinal variable in the HOV survey and two behavioral variables. Table 9.4 shows the results of two SEMs.

SEM example

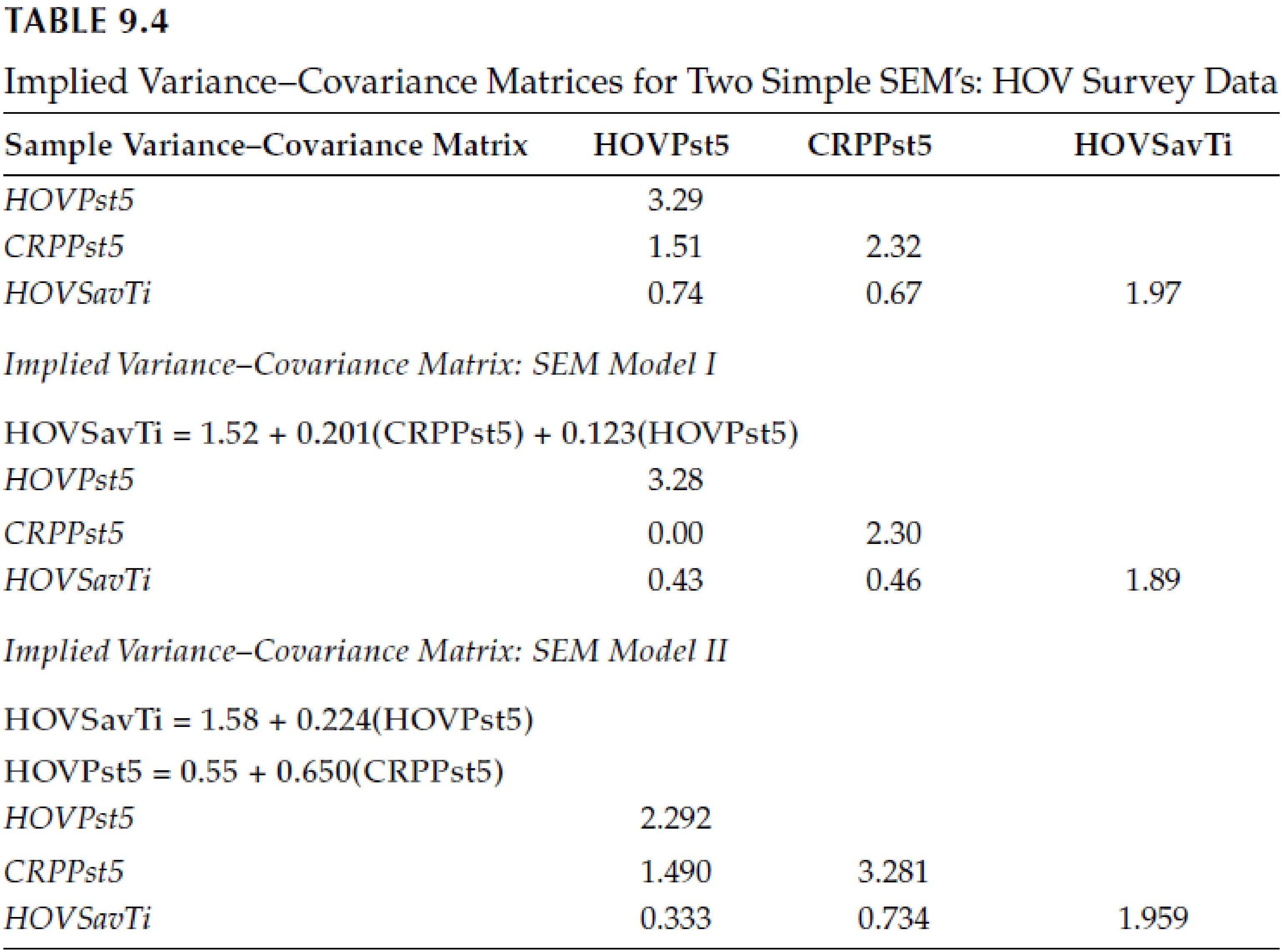


TABLE 9.4

Implied Variance–Covariance Matrices for Two Simple SEM's: HOV Survey Data

Sample Variance–Covariance Matrix	HOVPst5	CRPPst5	HOVSavTi
HOVPst5	3.29		
CRPPst5	1.51	2.32	
HOVSavTi	0.74	0.67	1.97
Implied Variance–Covariance Matrix: SEM Model I			
HOVSavTi = 1.52 + 0.201(CRPPst5) + 0.123(HOVPst5)			
HOVPst5	3.28		
CRPPst5	0.00	2.30	
HOVSavTi	0.43	0.46	1.89
Implied Variance–Covariance Matrix: SEM Model II			
HOVSavTi = 1.58 + 0.224(HOVPst5)			
HOVPst5 = 0.55 + 0.650(CRPPst5)			
HOVPst5	2.292		
CRPPst5	1.490	3.281	
HOVSavTi	0.333	0.734	1.959

A general framework of SEM

- SEMs are used to evaluate theories or hypotheses using empirical data.
- The empirical data are contained in a $p \times p$ variance-covariance matrix **S**, which is an unstructured estimator of the population variance-covariance matrix **Σ**. A SEM is hypothesized to be a function of *Q*, a vector of unknown structural parameters, **θ**, which in turn will generate a model-implied variance-covariance matrix $\Sigma(\boldsymbol{\theta})$.
- All variables in the model are classified as either independent (endogenous) or dependent (exogenous).

A general framework of SEM

- □ Dependent variables are collected into a vector $\boldsymbol{\eta}$, while independent variables are collected in the vector ξ. $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are estimated vectors that contain regression parameters for variables. And $\boldsymbol{\epsilon}$ is a vector of regression disturbances.

$$\boldsymbol{\eta} = \boldsymbol{\beta\eta} + \boldsymbol{\gamma\xi} + \boldsymbol{\epsilon}$$

- □ The exogenous factor covariance matrix is $\boldsymbol{\Phi} = COV[\boldsymbol{\epsilon}, \boldsymbol{\epsilon^T}]$, and the error covariance matrix as$\boldsymbol{\theta} = COV[\boldsymbol{\epsilon}, \boldsymbol{\epsilon^T}]$.

A general framework of SEM

- The variance-covariance matrix for the model is

$$\Sigma(\boldsymbol{\theta})=\mathbf{G}(\mathbf{I}-\boldsymbol{\beta})^{-1}\boldsymbol{\gamma}\boldsymbol{\Phi}\boldsymbol{\gamma}^{T}(\mathbf{I}-\boldsymbol{\beta})^{-1^{T}}\mathbf{G}^{T}$$

where G is a selection matrix containing either 0 or 1 to select the observed variables from all the dependent variables in η.

- There are p^2 elements or simultaneous equations in the above equation, one for each element in $\Sigma(\boldsymbol{\theta})$. Some of the p^2 equations are redundant, leaving $p^* = \frac{p(p-1)}{2}$ independent equations. The p^* independent equations are used to solve for unknown parameters $\boldsymbol{\theta}$, which consist of $\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\Phi}$.

SEM identification

- There are Q unknown model parameters (comprising $\boldsymbol{\theta}$), which must be solved using p^* simultaneous independent equations.
- Two conditions necessary and sufficient: one is that the number of simultaneous equations must be equal to or greater than the number of unknown model parameters, $Q \leq p^*$. The second is that each and every free model parameter must be identified, which can be difficult.

SEM identification

- Parameters are estimated using a discrepancy function, where the difference between the sample variance-covariance matrix and the model-implied variance-covariance matrix are minimized. The function is
$$F = F(\mathbf{S}, \Sigma(\hat{\boldsymbol{\theta}}))$$
- The discrepancy function of maximum likelihood estimation:
$$F_{MLE} = LN|\Sigma(\boldsymbol{\theta})| + \mathrm{TRACE}[\Sigma(\boldsymbol{\theta})^{-1}\boldsymbol{S}] - LN|\boldsymbol{S}| - p$$
- Discrepancy functions can also used to test the null hypothesis $H_0 : \Sigma(\boldsymbol{\theta}) = \boldsymbol{\Sigma}$, where
$$X^2 = F(n-1) \approx \chi^2(\alpha, p^* - Q)$$

Non-ideal conditions in SEM

- Ideal conditions in SEM: multivariate normality of independent variables, the correct model functional form, independent and dependent variables measured on the interval or ratio scale, and a sufficiently large sample size.
- However, not so many researches are conducted under ideal conditions.
- Non-normality can arise from poorly distributed continuous variables or coarsely categorized continuous variables. Interactions and nonlinear effects arise frequently in the modeling of real data.

Remedies for non-normality in SEM

- The asymptotically distribution free estimator (ADF) is a generalized least squares estimation approach that does not rely on multivariate normality.
- The scaled X^2 test statistic corrects or rescales the X^2 test statistic so that it approximates the referenced χ^2 distribution.
- Bootstrapping deals with non-normal samples.
- A continuous/ categorical variable methodology (CVM) weighted least squares estimator and discrepancy function, which results in unbiased, consistent, and efficient parameter estimates when variables are measured on nominal and ordinal scales.

Remedies for non-normality in SEM

- Two general approaches to handling interactions and nonlinear effects: the indicant product approach and the multisample approach.
- The indicant product approach is well developed only for multiplicative cases and requires a centering transformation.
- The multisample approach is more flexible, avoids some multilinearity and distributional problems associated with the product indicant approach, and is suitable under the widest range of conditions.

Model GOF measures in SEM

- Model goodness-of-fit measures are an important part of any statistical model assessment.
- The saturated model is as complex as the original data, it does not summarize the data into succinct and useful relationships.
- The independence model is constrained such that no relationships exist in the data and all variables in the model are independent of each other.
- The saturated and independence models can be viewed as two extremes within which the best model will lie.

Model GOF measures in SEM

- □ The first class of GOF indices include measures of parsimony. Models with few parameters are preferred to those with many. Three simple measures of parsimony are the number of model parameters Q, the degrees of freedom of the being tested df=p^*-Q, and the parsimony ratio is

$$PR = \frac{d}{d_i}$$

Where d is the degree of freedom of the estimated model, and $d_i$is the degrees of the independence model. A higher PR is preferred.

Model GOF measures in SEM

□ The second GOF indices based on the X^2 test statistic. The X^2 statistic is the minimum value of the discrepancy function *F* times tis degrees of freedom. The statistic X^2/(model degrees of freedom) has been suggested as a useful GOF measure. Rules of thumb have suggested that this measure (except under ULS and SLS estimation) should be close to 1 for correct models. In general, researchers have recommended this statistic lie somewhere less than 5, with values close to 1.

Model GOF measures in SEM

□ The third GOF indices based on the population discrepancy. These measures rely on the notion of a population discrepancy function (as opposed to the sample discrepancy function) to estimate GOF measure, including the noncentrality parameter (NCP), the root mean square error of hypothesis test of RMSEA=0.5.

Model GOF measures in SEM

□ Information theoretic measures are designed primarily for use with MLE methods, and are meant to provide a measure of the amount of information contained in a given model. The forth class of GOF - the Akaike information criterion is used to assess fit in this class.

$$AIC = 2Q - 2LL(\boldsymbol{\theta})$$

where Q is the number of parameters and $LL(\boldsymbol{\theta})$ is the log-likelihood at convergence.

Model GOF measures in SEM

- The fifth class of GOF measures is designed to compare the fitted model to baseline models. The normed fit index is given as

$$NFI = 1 - \frac{X^2}{X_b^2}$$

where the X_b^2 is the minimum discrepancy of the baseline model comparison, usually the saturated or independence model. NFI will take on values between 0 and 1. rules of thumb for this measure suggest that models with a NFI less than 0.9 can be substantially improved.

Guidelines for SEM

- Reasons for using SEMs:

  1) SEMs handle measurement problems well.

  2) SEMs provide a way to check the entire structure of data assumptions, not just whether or not the dependent variable predictions fit observations well.

  3) SEMs cope with endogeneity among variables well.

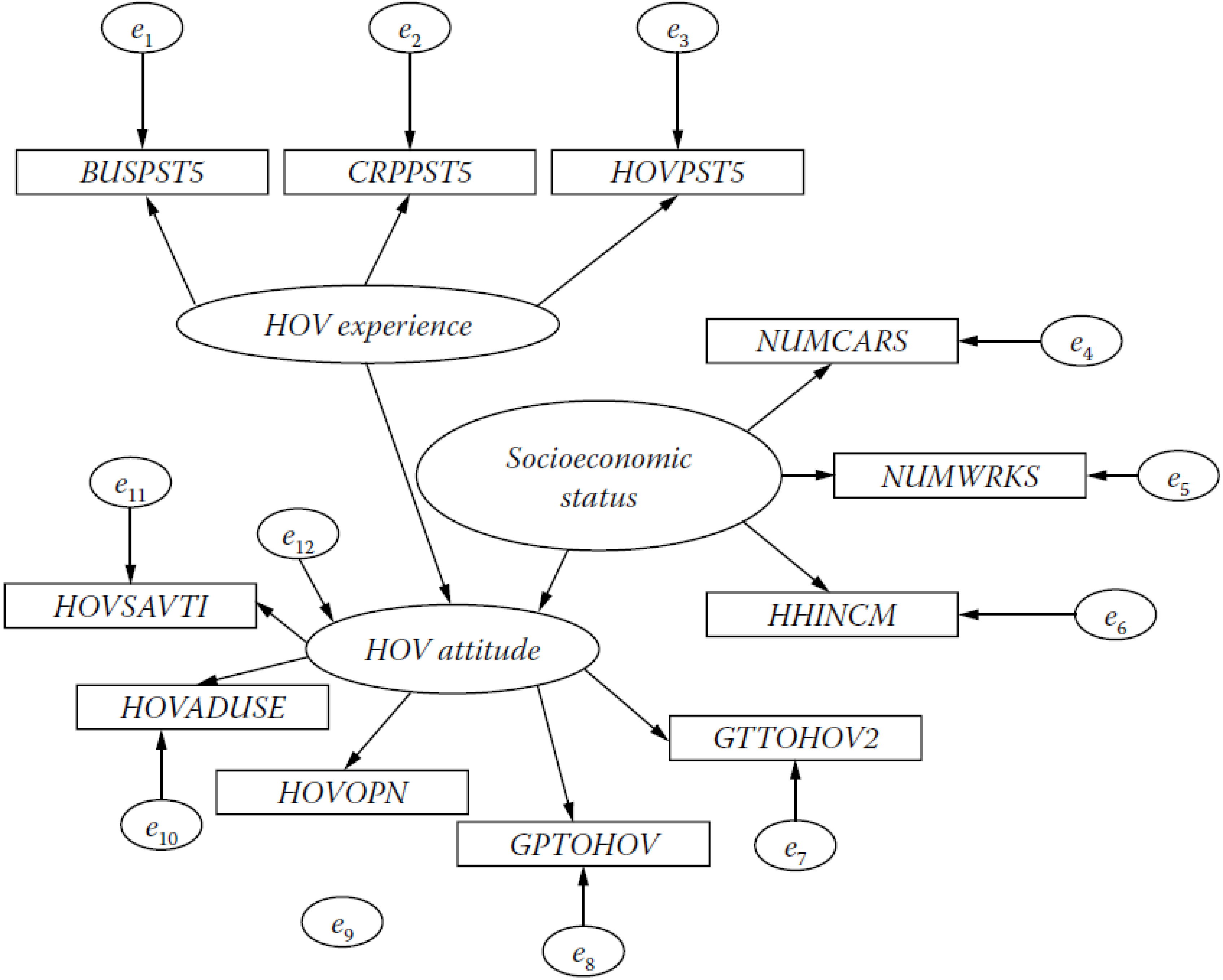
- Assessment of a SEM should take into account many criteria, including theoretical appeal of the model specification, overall X^2 GOF between observed and implied variance-covariance matrices, individual variable coefficients and their standard errors, and GOF indices.

SEM example

The results form the principal components and factor analyses are used to develop the SEM specification. Three important latent variables are thought to define the structure in the HOV survey data.

The relationship between latent variables and the measured variables in shown in the figure.

Table shows the estimation results for the SEM depicted in the figure.



e_1 e_2 e_3

BUSPST5 *CRPPST5* *HOVPST5*

HOV experience

NUMCARS e_4

Socioeconomic status

NUMWRKS e_5

e_{11} e_{12}

HHINCM e_6

HOVSAVTI

HOV attitude

HOVADUSE

GTTOHOV2

HOVOPN

e_{10}

GPTOHOV

e_7

e_9

e_8

Regression Parameters			
HOV Attitude ← Socioeconomic Status	–0.001	0.080	–0.017
HOV Attitude ← HOV Experience	0.507	0.078	6.509
CRPPST5 ← HOV Experience	0.848	0.085	10.001
HOVPST5 ← HOV Experience	1.724	0.079	21.908
BUSPST5 ← HOV Experience	0.386	0.061	6.324
HOVSAVTI ← HOV Attitude	0.905	0.068	13.333
HOVADUSE ← HOV Attitude	0.661	0.056	11.875
HOVOPN ← HOV Attitude	–1.164	0.075	–15.505
GTTOHOV2 ← HOV Attitude	0.542	0.065	8.283
NUMCARS ← Socioeconomic Status	0.832	0.185	4.497
NUMWRKS ← Socioeconomic Status	0.374	0.091	4.114
HHINCM ← Socioeconomic Status	0.431	0.126	3.410
GPTOHOV ← HOV Attitude	0.847	0.067	12.690
Intercepts			
HOVPST5	0.973	0.105	9.234
CRPPST5	0.643	0.089	7.259
BUSPST5	0.290	0.061	4.785
HOVADUSE	1.268	20.053	0.063
HOVOPN	1.733	0.089	19.458
GPTOHOV	1.757	0.077	22.914
HOVSAVTI	1.787	0.079	22.742
GTTOHOV2	1.892	0.071	26.675
NUMCARS	2.418	0.061	39.524
HHINCM	6.220	0.088	70.537
NUMWRKS	1.849	0.046	40.250
Goodness-of-Fit Measures			
Degreews of freedom = $P^* - Q$			77 – 34 = 43
Chi-square			129.66
Parsimony ratio			0.652
Chi-square/degrees of freedom			3.02