1 Introduction

1.1 Observed and Latent Variables

Observed variables are those that can be directly measured, such as systolic blood pressure, diastolic blood pressure, waist hip ratio, body mass index, and heart rate. Measurements from observed variables provide data as the basic source of information for statistical analysis. In medical, social, and psychological research, it is common to encounter latent constructs that cannot be directly measured by a single observed variable. Simple examples are intelligence, health condition, obesity, and blood pressure. To assess the nature of a latent construct, a combination of several observed variables is needed. For example, systolic blood pressure and diastolic blood pressure should be combined to evaluate blood pressure; and waist hip ratio and body mass index should be combined to evaluate obesity. In statistical inference, a latent construct is analyzed through a latent variable which is appropriately defined by a combination of several observed variables.

For practical research in social and biomedical sciences, it is often necessary to examine the relationships among the variables of interest. For example, in a study that focuses on kidney disease of type 2 diabetic patients (see Appendix 1.1), we have data from the following observed key variables: plasma creatine (PCr), urinary albumin creatinine ratio (ACR), systolic blood pressure (SBP), diastolic blood pressure (DBP), body mass index (BMI), waist hip ratio (WHR), glycated hemoglobin (HbAlc), and fasting plasma glucose (FPG). From the basic medical knowledge about kidney disease, we know that the severity of this disease is reflected by both PCr and ACR. In order to understand the effects of the explanatory (independent) variables such as SBP, BMI etc, on kidney disease, one possible approach is to apply the well-known regression model by treating

PCr and ACR as outcome (dependent) variables and regressing them on the observed explanatory (independent) variables as follows:

$$\text{PCr} = \alpha_1\text{SBP} + \alpha_2\text{DBP} + \alpha_3\text{BMI} + \alpha_4\text{WHR} + \alpha_5\text{HbA1c} + \alpha_6\text{FPG} + \epsilon_1, \tag{1.1}$$

$$\text{ACR} = \beta_1\text{SBP} + \beta_2\text{DBP} + \beta_3\text{BMI} + \beta_4\text{WHR} + \beta_5\text{HbA1c} + \beta_6\text{FPG} + \epsilon_2. \tag{1.2}$$

From the estimates of α's and β's, we can assess the effects of the explanatory variables on PCr and ACR. For example, based on the estimates of α_1 and β_1, we can evaluate the effects of SBP on PCr and ACR, respectively. However, this result cannot provide a clear and direct answer to the question about the effect of SBP on kidney disease. Similarly, the effects of other observed explanatory variables on kidney disease cannot be directly assessed from results obtained from regression analysis of equations (1.1) and (1.2). The deficiency of the regression model when applying to this study is due to the fact that kidney disease is a latent variable (construct) rather than an observed variable. A better approach is to appropriately combine PCr and ACR to a latent variable 'kidney disease (KD)' and regress this latent variable on the explanatory variables. Moreover, one may be interested in the effect of blood pressure rather than in the seperate effects of SBP and DBP. Although the estimates of α_1 and α_2 can be used to examine the effects of SBP and DBP on PCr, respectively, they cannot provide a direct and clear assessment on the effect of blood pressure on PCr. Hence, it is desirable to group SBP and DBP together to form a latent variable that can be interpreted as 'blood pressure (BP)', and then use BP as an explanatory variable. Based on similar reasoning, {BMI, WHR} and {HbA1c, FPG} are appropriately grouped together to form latent variables that can be interpreted as 'obesity (OB)' and 'glycemic control (GC)', respectively. To study the effects of blood pressure, obesity, and glycemic control on kidney disease, we consider the following simple regression equation with latent variables:

$$\text{KD} = \gamma_1\text{BP} + \gamma_2\text{OB} + \gamma_3\text{GC} + \delta. \tag{1.3}$$

This simple regression equation can be generalized to the multiple regression equation with product terms. For example, the following regression model can be used to assess the additional interactive effects among blood pressure, obesity, and glycemic control on kidney disease:

$$\text{KD} = \gamma_1\text{BP} + \gamma_2\text{OB} + \gamma_3\text{GC} + \gamma_4(\text{BP} \times \text{OB}) + \gamma_5(\text{BP} \times \text{GC}) + \gamma_6(\text{OB} \times \text{GC}) + \delta. \quad (1.4)$$

Note that studying these interactive effects by using the regression equations with the observed variables (see (1.1) and (1.2)) is extremely tedious.

It is obvious from the above simple example that incorporating latent variables in developing models for practical research is advantageous. First, it can reduce the number of variables in the key regression equation. Comparing Equation (1.3) with (1.1) and (1.2), the number of explanatory variables is reduced from six to three. Second, as highly correlated observed variables are grouped into latent variables, the problem induced by multicollinearity is alleviated. For example, the multicollinearity induced by the highly correlated variables SBP and DBP in analyzing regression equation (1.1) or (1.2) does not exist in regression equation (1.3). Third, it gives better assessments on the interrelationships of latent constructs. For instance, direct and interactive effects among the latent constructs blood pressure, obesity, and glycemic control can be assessed through the regression model (1.4). Hence, it is important to have a statistical method that simultaneously groups highly correlated observed variables into latent variables and assesses interrelationships among latent variables through a regression model of latent variables. This strong demand is the motivation for the development of structural equation models.

1.2 Structural Equation Model (SEM)

Structural equation model (SEM) is a powerful multivariate tool for studying interrelationships among observed and latent variables. This statistical method is very popular in behavioral, educational, psychological, and social research. Recently, it has also re-

ceived a great deal of attention in biomedical research; see for example Bentler and Stein (1992), and Pugesek, Tomer and von Eye (2003).

The basic SEM, for example, the widely used LISREL model (Jöreskog and Sörbom, 1996) is formulated by two components. The first component is a confirmatory factor analysis (CFA) model which groups the highly correlated observed variables into latent variables and takes the measurement error into account. This component can be regarded as a regression model which regresses the observed variables on a smaller number of latent variables. As the covariance matrix of the latent variables is allowed to be non-diagonal, the correlations/covariances of the latent variables can be evaluated. However, various effects of the explanatory latent variables on the key outcome latent variables of interest cannot be assessed by the CFA model of the first component. Hence, a second component is needed. This component is again a regression type model, in which the outcome latent variables are regressed on the explanatory latent variables. As a result, SEM is conceptually formulated by the familiar regression type model. However, as latent variables in the model are random, the standard technique in regression analysis cannot be applied to analyze SEMs.

In many substantive research, it is often important to develop an appropriate model to evaluate a series of simultaneous hypotheses on the impacts of some explanatory observed and latent variables on the key outcome variables. Based on its particular formulation, SEM is very useful for achieving the above objective. Furthermore, it is easy to appreciate the key idea of SEM, and to apply the model to substantive research; one only needs to understand the basic concepts of latent variables and the familiar regression model. As a result, this model has been extensively applied to behavioral, educational, psychological, and social research. Due to the strong demand, more than a dozen user-friendly SEM software packages have been developed; typical examples are AMOS, EQS6, LISREL,

and Mplus. Recently, SEM becomes a popular statistical tool for biomedical and environmental research. For instance, it has been applied to the analysis of the effects of in utero methylmercury exposure on neurodevelopment (Sánchez, *et al.*, 2005), to the study of ecological and evolutionary biology (Pugesek, Tomer and von Eye, 2003), and to the evaluation of the interrelationships among latent domains in quality of life (Lee *et al.*, 2005; among others).

1.3 Objectives of the Book

Like most other statistical methods, the methodological developments of standard SEMs depend on crucial assumptions. More specifically, the most basic assumptions are: (i) The regression model in the second component is based on a simple linear regression equation in which higher-ordered product terms (such as quadratic terms or interaction terms) cannot be assessed. (ii) The observed random variables are assumed to be continuous, independently and identically distributed as a normal distribution. As these assumptions may not be valid in substantive research, they induce limitations in applying SEMs to the analysis of real data in relation to complex situations. Motivated by the need for overcoming the limitations, the growth of SEM has been very rapid in recent years. New models and statistical methods have been developed to relax various aspects of the crucial assumptions for better analyses of complex data structure in practical research. These include but are not limited to: (i) nonlinear SEMs with covariates (Schumacker and Marcoulides, 1998; Lee and Song, 2003a; among others), (ii) SEMs with mixed continuous, ordered and/or unordered categorical variables (Shi and Lee, 2000; Moustaki, 2003; Song and Lee, 2004; Song *et al.*, 2007; among others), multilevel SEMs (Lee and Shi, 2001; Rabe-Hesketh, Skrondal and Pickles, 2004; Song and Lee, 2004; Lee and Song, 2005; among others), mixture SEMs (Dolan and van der Maas, 1998; Zhu and Lee, 2001; Lee and Song, 2003b; among others), SEMs with missing data

(Jamshidian and Bentler, 1999; Lee and Tang, 2006; Song and Lee, 2006; among others), SEMs with variables from exponential family distributions (Wedel and Kamakura, 2001; Song and Lee, 2007; among others), longitudinal SEMs (Dunson, 2003; Song, Lee and Hser, 2008), semiparametric SEMs (Lee, Lu and Song, 2008; Song, Xia and Lee, 2009; Yang and Dunson, 2010; Song and Lu, 2010), and transformation SEMs (van Montfort, Mooijaart and Meijerink, 2009; Song and Lu, 2011). As the existing software packages in SEMs are developed on the basis of the covariance structure approach, and their primary goal is to analyze the standard SEM under usual assumptions, they cannot be effectively and efficiently applied to the analysis of the more complex models and/or data structures mentioned above. Blindly applying these software to complex situations has a very high chance to obtain questionable results and to draw misleading conclusions.

In substantive research, data obtained for evaluating hypotheses of complex diseases are usually very complicated. In analyzing these complicated data, more subtle models and rigorous statistical methods are important for providing correct conclusions. In view of this, there is an urgent need to introduce statistical sound methods that are recently developed to applied researches. This is the main objective of writing this book. At the moment, there is only a very limited number of references/textbooks in SEM. Bollen (1989) was devoted to standard SEMs and focused on the covariance structure approach. Compared to Bollen (1989), this book introduces more advanced SEMs and emphasizes on the Bayesian approach which is more flexible than the covariance structure approach in handling complex data and models. Lee (2007) provides a Bayesian approach for analyzing the standard and more subtle SEMs. Compared to Lee (2007), the first four chapters of this book provide less technical discussions and explanations of the basic ideas in addition to the more involved, theoretical developments of the statistical methods, so that they can be understood without much difficulty by applied researchers. Another

objective of this book is to introduce important models that are recently developed and have not been covered by Lee (2007), including innovative growth curve models and longitudinal SEMs for analyzing longitudinal data and for studying the dynamic changes of characteristics with respect to time; semiparametric SEMs for relaxing the normality assumption and for assessing the true distributions of explanatory latent variables; SEMs with a nonparametric structural equation for capturing the true general relationships among latent variables, and transformation SEMs for analyzing highly non-normal data. We believe that these advanced SEMs are very useful in substantive research.

1.4 The Bayesian Approach

A traditional method in analyzing SEMs is the covariance structure approach which focuses on fitting the covariance structure under the proposed model to the sample covariance matrix computed from the observed data. For simple SEMs, when the underlying distribution of the observed data is normal, this approach works fine with reasonably large sample sizes. However, some serious difficulties may be encountered in many complex situations in which deriving the covariance structure or obtaining an appropriate sample covariance matrix for statistical inferences is difficult.

Thanks to the recent advance of statistical computing, such as the development of various efficient Markov chain Monte Carlo (MCMC) algorithms, the Bayesian approach has been extensively applied to analyze many complex statistical models. Inspired by its wide applications in statistics, we will use the Bayesian approach to analyze the advanced SEMs that are useful for medical and social-psychological research. Moreover, in formulating and fitting the model, we emphasize on the raw individual random observations rather than the sample covariance matrix. The Bayesian approach coupled with the formulation based on raw individual observations has several advantages. First, the development of statistical methods is based on the first moment properties of the raw

individual observations which are simpler than the second moment properties of the sample covariance matrix. Hence, it has potential to be applied to more complex situations. Second, it produces a direct estimation of latent variables, which cannot be obtained with classical methods. Third, it directly models observed variables with their latent variables through the familiar regression equations; hence, it gives a more direct interpretation and can utilize the common techniques in regression such as outlier and residual analyses in conducting statistical analysis. Fourth, in addition to the information that is available in the observed data, the Bayesian approach allows the use of genuine prior information for producing better results. Fifth, the Bayesian approach provides more easily assessable statistics for goodness-of-fit and model comparison, and also other useful statistics such as the mean and percentiles of the posterior distribution. Sixth, it can give more reliable results for small samples (see Dunson, 2000; Lee and Song, 2004). For methodological researchers in SEMs, technical details that are necessary in developing the theory and the MCMC methods are given in the appendices of the chapters. Applied researchers who are not interested in the methodological developments can skip those appendices. For convenience, we will introduce the freely available software WinBUGS (Spiegelhalter, *et al.*, 2003) through analyses of simulated and real data sets. This software is able to produce reliable Bayesian statistics including the Bayesian estimates and their standard error estimates for a wide range of statistical models (Congdon, 2003) and for SEMs (Lee, 2007).

1.5 Real Data Sets and Notations

We will use several real data sets for motivating the models and for illustrating the proposed Bayesian methodologies. These data sets are respectively related to the studies about: (i) job and life satisfaction, work attitude, and other related social-political issues; (ii) effects of some phenotype and genotype explanatory latent variables on kidney disease

for type 2 diabetic patients; (iii) quality of life related to residents of several countries, and related to stroke patients; (iv) the development and findings from an AIDS preventative intervention for Filipino commercial sex workers; (v) the longitudinal characteristics of cocaine and polydrug use; (vi) the functional relationships between bone mineral density (BMD) and its observed and latent determinants for old men; and (vii) the academic achievement and its influential factors for American youth. Some information of these data sets is given in Appendix 1.1.

In the discussion of various models and their associative statistical methods, we will encounter different types of observations in relation to observable continuous and discrete variables or covariates; unobservable measurements in relation to missing data or continuous measurements underlying the discrete data; latent variables; as well as different types of parameters, such as thresholds, structural parameters in the model, and hyperparameters in the prior distributions. Hence, we do not have enough symbols. If the context is clear, some Greek alphabets may be used to serve different purposes. For example, α has been used to denote an unknown threshold in defining an ordered categorical variable, and to denote a hyperparameter in some prior distributions. Nevertheless, some general notations are given as in Table 1.1.

Appendix 1.1 Information of Real Data Sets

(i) Inter-university Consortium for Political and Social Research(ICPSR) data

ICPSR data set was collected in the project WORLD VALUES SURVEY 1981-1984 AND 1990-1993 (World Values Study Group, ICPSR Version). The whole data set was the answers to questionnaire survey about work attitude, job and family life, religious belief, interest in politics, attitude towards competition, etc. The items that have been used in the illustrative examples of this book are given below.

Thinking about your reasons for doing voluntary work, please use the following five-point scale to indicate how important each of the reasons below have been in your own case (1 is unimportant and 5 is very important).

V 62 Religious beliefs 1 2 3 4 5

During the past few weeks, did you ever feel $\cdots$ (Yes: 1; No: 2)

V 89 Bored 1 2

V 91 Depressed or very unhappy 1 2

V 93 Upset because somebody criticized you 1 2

V 96 All things considered, how satisfied are you with your life as a whole these days?

1 2 3 4 5 6 7 8 9 10

Dissatisfied Satisfied

Here are some aspects of a job that people say are important. Please look at them and tell me which ones you personally think are important in a job. (Mentioned: 1; Not Mentioned: 2)

V 99	Good Pay	1	2
V 100	Pleasant people to work with	1	2
V 102	Good job security	1	2
V 103	Good chances for promotion	1	2

V 111 A responsible job 1 2

V 115 How much pride, if any, do you take in the work that you do?

1 A great deal, 2 Some, 3 Little, 4 None

V 116 Overall, how satisfied or dissatisfied are you with your job?

1 2 3 4 5 6 7 8 9 10

Dissatisfied Satisfied

V 117 How free are you to make decisions in your job?

1 2 3 4 5 6 7 8 9 10

Not at all A great deal

V 129 When job are scarce, people should be forced to retire early,

1 Agree, 2 Neither, 3 Disagree

V 132 How satisfied are you with the financial situation of your household?

1 2 3 4 5 6 7 8 9 10

Dissatisfied Satisfied

V 176 How important is God in your life? 10 means very important and

1 means not at all important.

1 2 3 4 5 6 7 8 9 10

V 179 How often do you pray to God outside of religious services? Would you say ...

1 Often 2 Sometimes

3 Hardly ever 4 Only in times of crisis 5 Never

V 180 Overall, how satisfied or dissatisfied are you with your home life?

1 2 3 4 5 6 7 8 9 10

Dissatisfied Satisfied

V 241 How interested would you say you are in politics?

1 Very interested 2 Somewhat interested

3 Not very interested 4 Not at all interested

Now I'd like you to tell me your views on various issues. How would you place your views on this scale? 1 means you agree completely with the statement on the left, 10 means you agree completely with the statement on the right, or you can choose any number in between.

V 252

1 2 3 4 5 6 7 8 9 10

Individual should take more responsibility for providing for themselves.

The state should take more responsibility to ensure that everyone is provided for.

V 253

1 2 3 4 5 6 7 8 9 10

People who are unemployed should have to take any job available or lose their unemployment benefits.

People who are unemployed should have the right to refuse a job they do not want.

V 254

1 2 3 4 5 6 7 8 9 10

Competition is good. It stimulates people to work hard and develop new ideas.

Competition is harmful. It brings out the worst in people.

V 255

1 2 3 4 5 6 7 8 9 10

In the long run, hard work usually brings a better life.

Hard work doesn't generally bring success — it's more a matter of luck and connections.

Please tell me for each of the following statements whether you think it can always be justified, never be justified, or something in between.

V 296 Claiming government benefits which you are not entitled to

1	2	3	4	5	6	7	8	9	10
Never									Always

V 297 Avoiding a fare on public transport

1	2	3	4	5	6	7	8	9	10
Never									Always

V 298 Cheating on tax if you have the chance

1	2	3	4	5	6	7	8	9	10
Never									Always

V 314 Failing to report damage you've done accidentally to a parked vehicle

1	2	3	4	5	6	7	8	9	10
Never									Always

I am going to read out some statements about the government and the economy. For each one, could you tell me how much you agree or disagree?

V 336 Our government should be made much more open to the public

1	2	3	4	5	6
Agree Completely					Disagree Completely

V 337 We are more likely to have a healthy economy if the government allows more freedom for individuals to do as they wish

1	2	3	4	5	6
Agree Completely					Disagree Completely

V 339 Political reform in this country is moving too rapidly

1	2	3	4	5	6
Agree Completely					Disagree Completely

(ii) Type 2 diabetic patients data

The data set was collected from an applied genomics program conducted by the Institute of Diabetes, the Chinese University of Hong Kong. It aims to examine the clinical and molecular epidemiology of type 2 diabetes in Hong Kong Chinese, with particular emphasis on diabetic nephropathy. A consecutive cohort of 1,188 type 2 diabetic patients was enrolled into the Hong Kong Diabetes Registry. All patients underwent a structured 4-hour clinical and biochemical assessment including renal function measured by plasma creatine (PCr) and urinary albumin creatinine ratio (ACR); continuous phenotype variables: systolic blood pressure (SBP), diastolic blood pressure (DBP), body mass index (BMI), waist hip ratio (WHR), glycated hemoglobin (HbA1c), fasting plasma glucose (FPG), non-high density lipoprotein cholesterol (non-HDL-C), lower density lipoprotein cholesterol (LDL-C), plasma triglyceride (TG); and multinomial genotype variables: beta3 adrenergic receptor (ADRβ3), beta2 adrenergic receptor SNP1 (ADRβ21), beta2 adrenergic receptor SNP2 (ADRβ22), angiotensin converting enzyme (DCP1 intro 16 del/ins (DCP1)), and angiotensin II receptor type 1 AgtR1 A1166C (AGTR1), etc.

(iii) WHOQOL-BREF quality of life assessment data

The WHOQOL-100 assessment was developed by the WHOQOL group in 15 international field centers for assessing quality of life (QOL). The WHOQOL-BREF instrument is a shorten version of WHOQOL-100 by selecting 24 ordinal categorical items out of the 100 items. This instrument was established to evaluate four domains: physical health, mental health, social relationships, and environment. The instrument also includes two ordinal categorical items for the overall QOL and the health-related QOL, giving a total of 26 items. All of the items are measured with a 5-point scale (1 = 'not at all/very dissatisfied'; 2 = 'a little/dissatisfied'; 3 = 'moderate/neither'; 4 = 'very much/satisfied'; 5 = 'extremely/very satisfied'). The frequencies of the ordinal scores of the items:

WHOQOL Items	Ordinal Scores					Number of incomplete obs.
	1	2	3	4	5	
Q1 Overall QOL	3	41	90	233	107	1
Q2 Overall health	32	127	104	154	58	0
Q3 Pain and discomfort	21	65	105	156	127	1
Q4 Medical treatment dependence	21	57	73	83	239	2
Q5 Energy and fatigue	15	57	166	111	118	8
Q6 Mobility	16	36	58	120	243	2
Q7 Sleep and rest	28	87	95	182	83	0
Q8 Daily activities	7	73	70	224	100	1
Q9 Work capacity	19	83	88	191	91	3
Q10 Positive feeling	2	30	141	241	59	2
Q11 Spirituality/personal beliefs	13	45	149	203	61	4
Q12 Memory and concentration	4	40	222	184	21	4
Q13 Bodily image and appearance	9	46	175	137	106	2
Q14 Self-esteem	13	72	130	210	50	0
Q15 Negative feeling	4	54	137	239	39	2
Q16 Personal relationship	8	46	68	218	134	1
Q17 Sexual activity	25	55	137	149	76	33
Q18 Social support	2	23	84	228	136	2
Q19 Physical safety and security	2	25	193	191	62	2
Q20 Physical environment	4	29	187	206	43	6
Q21 Financial resources	27	56	231	105	54	2
Q22 Daily life information	5	27	176	194	70	3
Q23 Participation in leisure activity	10	99	156	163	47	0
Q24 Living condition	9	27	53	235	151	0
Q25 Health accessibility and quality	0	17	75	321	61	1
Q26 Transportation	8	38	61	253	113	2

(iv) AIDS preventative intervention data

The data set was collected from female commercial sex workers (CSWs) in 95 establishments (bars, night clubs, Karaoke TV and massage parlours) in Philippine cities. The whole questionnaire consists of 134 items on areas of demographics knowledge, attitudes, beliefs, behaviors, self-efficacy for condom use, and social desirability. The primary concern is finding an AIDS preventative intervention for Filipino CSWs. Questions are:

(1) How much of a threat do you think AIDS is to the health of people?

no threat at all/very small/moderate/strong/very great

(2) What are the chances that you yourself might get AIDS?

none/very small/moderate/great/very great

(3) How worried are you about getting AIDS?

not worried/slightly/moderate/very/extremely

How great is the risk of getting AIDS or the AIDS virus from sexual intercourse with someone:

(4) Who has the AIDS virus using a condom?

none/very small/moderate/great/very great

(5) Whom you don't know very well without using a condom?

none/very small/moderate/great/very great

(6) Who injects drugs?

none/very small/moderate/great/very great

(7) How often did you perform vaginal sex in the last 7 days?

(8) How often did you perform manual sex in the last 7 days?

(9) How often did you perform oral sex in the last 7 days?

(10) Have you ever used a condom? Yes/No

(11) Did you use a condom the last time you have sex? Yes/No

(12) Have you ever put a condom on a customer? Yes/No

(13) Do you agree or disagree that condoms make sex less enjoyable?

strongly agree/agree/neutral/disagree/strongly disagree

(14) Do you agree or disagree that condoms cause a man to lose his erection?

strongly agree/agree/neutral/disagree/strongly disagree

(15) Do you agree or disagree that condoms cause pain or discomfort?

strongly agree/agree/neutral/disagree/strongly disagree

(16) Are condoms available at your establishment for the workers who work there? Yes/No

(17) How much do you think you know the disease called AIDS?

nothing/a little/somewhat/moderate/a great deal

(18) Have you ever had an AIDS test? Yes/No

(v) Polydrug use and treatment retention data

This is a longitudinal study of polydrug use conducted in five California counties in 2004. Data were collected from self-reported and administrative questionnaires about the retention of drug treatment (i.e., the days of stay in treatment), drug use history, drug-related crime history, and service and test received for 1,588 participants at intake, 3-month, and 12-month follow-up interviews. In addition, variables about treatment motivation (Mtsum01, Mtsum02, and Mtsum03) were collected at intake. Variables include:

(1) Drgplm30: Drug problems in past 30 days at intake, which ranges from 0 to 30.

(2) Drgday30: Drug use in past 30 days at intake, which ranges from 0 to 30.

(3) DrgN30: The number of kinds of drugs used in past 30 days at intake, which ranges from 1 to 8.

(4) Incar: The number of incarcerations in lifetime at intake, which ranges from 0 to 216.

(5) ArrN: The number of arrests in lifetime at intake, which ranges from 1 to 115.

(6) Agefirstarrest: The age of first arrest, which ranges from 6 to 57.

(7) Retent: Days of stay in treatment or retention, which ranges from 0 to 365.

(8) M12drg30: Primary drug use in past 30 days at 12 month interview, which ranges from 1 to 5.

(9) Servicem: Services received in past 3 months at TSI 3 month interview.

(10) DrugtestTX: The number of drug tests by TX in past 3 months at TSI 3 month interview, which ranges from 0 to 36.

(11) DrugtestCJ: The number of drug tests by criminal justice in past 3 months at TSI 3 month interview, which ranges from 0 to 12.

(12) Mtm01: Motivation subscale 1 at intake, which ranges from 1 to 5.

(13) Mtm02: Motivation subscale 2 at intake, which ranges from 1 to 5.

(14) Mtm03: Motivation subscale 3 at intake, which ranges from 1 to 5.

(vi) Quality of life for stroke survivors data

The setting of this study was in the Prince of Wales Hospital (PWH) in Hong Kong which is a regional university hospital with 1,500 beds serving a population of 0.7 million people. Patients with acute stroke within 2 days of admission were identified and followed up at three, six, and twelve months post-stroke. All patients included in the study were ethnic Chinese. As this study aimed to study those with a first disabling stroke, patients were excluded if they had moderate or severe premorbid handicap level (Rankin Scale score > 2). Outcome measures are obtained from questionnaires, which respectively measure respondents' functional status, depression, health-related quality of life, and handicap situation, including (1) the modified Barthel Index (MBI) score, (2) Geriatric Depression Scale (GDS) score, (3) Chinese Mini-Mental State Examination (MMSE) score, (4) World Health Organization Quality of Life measure (abbreviated Hong Kong version) (WHOQOL BREF (HK)) scores, and (5) the London Handicap Scale (LHS) score.

(vii) Cocaine use data

This data set was obtained from a longitudinal study about cocaine use conducted at the UCLA Center for Advancing Longitudinal Drug Abuse Research. The UCLA Center collected various measures from patients admitted in 1988-89 to the West Los Angeles Veterans Affairs Medical Center and met the DSM III-R criteria for cocaine dependence. The cocaine-dependent patients were assessed at baseline, one year after treatment, two years after treatment, and 12 years after treatment in 2002. Measures at each time point include

(1) cocaine use (CC), an ordered categorical variable with codings 1 to 5 to denote days of cocaine use per month that are fewer than 2 days, between 2-7 days, between 8-14 days, between 15-25 days, and more than 25 days, respectively.

(2) Beck inventory (BI), an ordered categorical variable with codings 1 to 5 to denote scores that are less than 3.0, between 3.0 to 8.0, between 9.0 to 20.0, between 21 to 30, and larger than 30.

(3) Depression (DEP), an ordered categorical variable based on the Hopkins Symptom Checklist-58 scores, with codings 1 to 5 to denote scores that are less than 1.1, between 1.1 and 1.4, between 1.4 and 1.8, between 1.8 and 2.5, and larger than 2.5.

(4) Number of friends (NF), an ordered categorical variable with codings 1 to 5 to denote no friend, 1 friend, 2-4 friends, 5-8 friends, more than 9 friends.

(5) 'Have someone to talk to about problem (TP)', $\{0, 1\}$ for {No, Yes}.

(6) 'Currently employed (EMP)', $\{0, 1\}$ for {No, Yes}.

(7) 'Alcohol dependence (AD) at baseline', $\{0, 1\}$ for {No, Yes}.

(viii) Bone mineral density (BMD) data

This data set was collected from a partial study on osteoporosis prevention and control. The study concerned the influence of serum concentration of sex hormones, their

precursors and metabolites on bone mineral density in older men. It was part of a multi-center prospective cohort study of risk factors of osteoporotic fractures in older people. A total of 1,446 Chinese men aged 65 years and above were recruited using a combination of private solicitation and public advertising from community centers and public housing estates.

The observed variables include: spine BMD, hip BMD, estrone (E1), estrone sulphate (E1-S), estradiol (E2), testosterone (TESTO), 5-Androstenediol (5-DIOL), dihydrotestosterone (DHT), androstenedione (4-DIONE), dehydroepiandrosterone (DHEA), DHEA sulphate (DHEA-S), androsterone (ADT), ADT glucuronide (ADT-G), 3α-diol-3G (3G), and 3α-diol-17G (17G). Moreover, weight and age were also measured.

(ix) National Longitudinal Surveys of Youth (NLSY) data

The four-decade-long NLSY is one of the most comprehensive longitudinal studies of youths conducted in North America. The NLSY data include a nationally representative sample of youths who were 14-21 years old in 1979 and 29-36 years in 1994.

The data set derived for the illustrative examples of this book includes 1,660 observations and the following measures: the Peabody Individual Achievement Tests (PIAT) with continuous scales in three domains: math, reading recognition, and reading comprehension; the Behavior Problem Index (BPI) with an ordinal scale in five domains: antisocial, anxious, dependent, headstrong, and hyperactive behavior; home environment in three domains: cognitive stimulation, emotional support, and household conditions; and friendship in two domains: the number of boyfriends and the number of girlfriends. The instruments for measuring these constructs were taken from a short form of Home Observation for Measurement of the Environment (HOME) Inventory.

References

Bentler, P. M. and Stein, J. A. (1992) Structural equation models in medical research. *Statistical Methods in Medical Research*, **1**, 159-181.

Bollen, K. A. (1989) *Structural Equation Models with Latent Variables*. NJ: John Wiley & Sons, Inc.

Chow, S. M., Tang, N. S., Yuan, Y., Song, X. Y. and Zhu, H. T. (2011) Bayesian estimation of semiparametric nonlinear dynamic factor analysis models using the Dirichlet process prior. *British Journal of Mathematical and Statistical Psychology*, **64**, 69-106.

Congdon, P. (2003) *Applied Bayesian Modelling*. NJ: John Wiley & Sons, Ins.

Dolan, C. V. and van der Maas, H. L. J. (1998) Fitting multivariage normal finite mixtures subject to structural equation modeling. *Psychometrika*, **63**, 227-253.

Dunson, D. B. (2000) Bayesian latent variable models for clustered mixed outcomes. *Journal of the Royal Statistical Society, Series B*, **62**, 355-366.

Dunson, D. B. (2003) Dynamic latent trait models for multidimensional longitudinal data. *Journal of the American Statistical Association*, **98**, 555-563.

Jöreskog, K. G. and Sörbom, D. (1996) LISREL 8: *Structural Equation Modeling with the SIMPLIS Command Language*. Scientific Software International.

Jamshidian, M. and Bentler, P. M. (1999) ML estimation of mean and covariance structures with missing data using complete data routines. *Journal of Educational and Behavioral Statistics*, **24**, 21-41.

Lee, S. Y. (2007) *Structural Equation Modeling: A Bayesian Approach*. UK: John Wiley & Sons, Ltd.

Lee, S. Y. and Shi, J. Q. (2001) Maximum likelihood estimation of two-level latent variable models with mixed continuous and polytomous data. *Biometrics*, **57**, 787-794.

Lee, S. Y. and Song X. Y. (2003a) Model comparison of nonlinear structural equation models with fixed covariates. *Psychometrika*, **68**, 27-47.

Lee, S. Y. and Song, X. Y. (2003b) Maximum likelihood estimation and model comparison for mixtures of structural equation models with ignorable missing data. *Journal of Classification*, **20**, 221-255.

Lee, S. Y. and Song, X. Y. (2004) Evaluation of the Bayesian and maximum likelihood approaches in analyzing structural equation models with small sample sizes. *Multivariate Behavioral Research*, **39**, 653-686.

Lee, S. Y. and Song, X. Y. (2005) Maximum likelihood analysis of a two-level nonlinear structural equation model with fixed covariates. *Journal of Educational and Behavioral Statistics*, **30**, 1-26.

Lee, S. Y. and Tang, N. S. (2006) Bayesian analysis of nonlinear structural equation models with nonignorable missing data. *Psychometrika*, **71**, 541-564.

Lee, S. Y., Lu, B. and Song, X. Y. (2008) Semiparametric Bayesian analysis of structural equation models with fixed covariates. *Statistics in Medicine*, **27**, 2341-2360.

Lee, S. Y., Song, X. Y., Skevington, S. and Hao, Y. T. (2005) Application of structural equation models to quality of life. *Structural Equation Modeling - A Multidisciplinary Journal*, **12**, 435-453.

Moustaki, I. (2003) A general class of latent variable models for ordinal manifest variables with covariate effects on the manifest and latent variables. *British Journal of Mathematical and Statistical Psychology*, **56**, 337-357.

Pugesek, B. H., Tomer, A. and von Eye, A. (2003) *Structural Equation Modeling: Applications in Ecological and Evolutionary Biology*, Cambridge: Cambridge University Press.

Rabe-Hesketh, S., Skrondal, A. and Pickles, A. (2004) Generalized multilevel structural equation modeling. *Psychometrika*, **69**, 167-190.

Sánchez, B. N., Budtz-Jorgensen, E., Ryan L, M. and Hu, H. (2005) Structural equation models: a review with applications to environmental epidemiology. *Journal of the American Statistical Association*, **100**, 1443-1455.

Schumacker, R. E. and Marcoulides, G. A. (1998) *Interaction and Nonlinear Effects in Structural Equation Modeling*. Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Shi, J. Q. and Lee, S. Y. (2000) Latent variable models with mixed continuous and polytomous data. *Journal of the Royal Statistical Society, Series B*, **62**, 77-87.

Song, X. Y. and Lee, S. Y. (2004) Bayesian analysis of two-level nonlinear structural equation models with continuous and polytomous data. *British Journal of Mathematical and Statistical Psychology*, **57**, 29-52.

Song, X. Y. and Lee, S. Y. (2006) A maximum likelihood approach for multisample nonlinear structural equation models with missing continuous and dichotomous data. *Structural Equation Modeling - A Multidisciplinary Journal*, **13**, 325-351.

Song, X. Y. and Lee, S. Y. (2007) Bayesian analysis of latent variable models with non-ignorable missing outcomes from exponential family. *Statistics in Medicine*, **26**, 681-693.

Song, X. Y. and Lu, Z. H. (2010) Semiparametric latent variable models with Bayesian P-splines. *Journal of Computational and Graphical Statistics*, **19**, 590-608.

Song, X. Y. and Lu, Z. H. (2011) Semiparametric transformation models with Bayesian P-splines. *Statistics and Computing*, to appear.

Song, X. Y., Lee, S. Y. and Hser, Y. I. (2008) A two-level structural equation model approach for analyzing multivariate longitudinal responses. *Statistics in Medicine*, **27**, 3017-3041.

Song, X. Y., Xia, Y. M. and Lee, S. Y. (2009) Bayesian semiparametric analysis of structural equation models with mixed continuous and unordered categorical variables. *Statistics in Medicine*, **28**, 2253-2276.

Song, X. Y., Lee, S. Y., Ng, M. C. Y., So, W. Y. and Chan, J. C. N. (2007) Bayesian analysis of structural equation models with multinomial variables and an application to type 2 diabetic nephropathy. *Statistics in Medicine*, **26**, 2348-2369.

Spiegelhalter, D. J., Thomas, A., Best, N. G. and Lunn, D. (2003) *WinBugs User Manual. Version 1.4*. Cambridge, UK: MRC Biostatistics Unit.

van Montfort, K., Mooijaart, A. and Meijerink, F. (2009) Estimating structural equation models with non-normal variables by using transformations. *Statistica Neerlandica*, **63**, 213-226.

Wedel, M. and Kamakura, W. A. (2001) Factor analysis with (mixed) observed and latent variables in the exponential family. *Psychometrika*, **66**, 515-530.

Yang, M. G. and Dunson, D. B. (2010) Bayesian semiparametric structural equation models with latent variables. *Psychometrika*, **75**, 675-693.

Zhu, H. T. and Lee, S. Y. (2001) A Bayesian analysis of finite mixtures in the LISREL model. *Psychometrika*, **66**, 133-152.

Table 1.1: Typical Notations

Symbols	Meaning
$\boldsymbol{\omega}$	Latent vector in the measurement equation.
$\boldsymbol{\eta}$	Outcome (dependent) latent vector in the structural equation.
$\boldsymbol{\xi}$	Explanatory (independent) latent vector in the structural equation.
$\boldsymbol{\epsilon}$, $\boldsymbol{\delta}$	Random vectors of measurement errors.
$\boldsymbol{\Lambda}$	Factor loading matrix in the measurement equation.
$\mathbf{B}$, $\boldsymbol{\Pi}$, $\boldsymbol{\Gamma}, \boldsymbol{\Lambda}_\omega$	Matrices of regression coefficients in the structural equation.
$\boldsymbol{\Phi}$	Covariance matrix of explanatory latent variables.
$\boldsymbol{\Psi}_\epsilon$, $\boldsymbol{\Psi}_\delta$	Diagonal covariance matrices of measurement errors, with diagonal elements $\psi_{\epsilon k}$ and $\psi_{\delta k}$, respectively.
$\alpha_{0\epsilon k}, \beta_{0\epsilon k}, \alpha_{0\delta k}, \beta_{0\delta k}$	Hyperparameters in the Gamma distributions of $\psi_{\epsilon k}$ and $\psi_{\delta k}$.
$\mathbf{R}_0$, ρ_0	Hyperparameters in the Wishart distribution related to the prior distribution of $\boldsymbol{\Phi}$.
$\boldsymbol{\Lambda}_{0k}$, $\mathbf{H}_{0yk}$	Hyperparameters in the multivariate normal distribution related to the prior distribution of the kth row of $\boldsymbol{\Lambda}$ in the measurement equation.
$\boldsymbol{\Lambda}_{0\omega k}$, $\mathbf{H}_{0\omega k}$	Hyperparameters in the multivariate normal distribution related to the prior distribution of the kth row of $\boldsymbol{\Gamma}$ in the structural equation.
$\mathbf{I}_q$	A $q \times q$ identity matrix. Sometimes we just use $\mathbf{I}$ to denote an identity matrix if its dimension is clear.