Chapter 1 Introduction

1.1 Background

From 1999 to 2000, a regular screening program was executed at Hsin-Yi rural area of Nantou County of Taiwan. This program targeted at several purposes. First, potential chronic disease patients could be traced and followed-up through this program since the resources and accompanied medical care are not well established in this area. Second, health-related interventions could be implemented via face-to-face interview and physical examinations. Third, an epidemiologic cohort of a health promotion plan and consecutive researches could be built on the basis of this cross-sectional sample. In this program, physical examinations, collection of serum and urine samples, as well as a questionnaire, were administrated on each person. The risk factors of several chronic diseases, including GOUT, hyperuricemia,..etc., have already been reported in Lai (2002).

Physiological phenomena and the motivation problem

As we know that many chronic diseases are related with each other in that there might be one or even several common factors or latent variables involve in the mechanism of these diseases. Investigation of the relationship between factors is thus appealing in several aspects. (1) A further causal structure can be clarified if prospective data are available. (2) Prediction models (for early detection and treatment) of diseases can be set-up. (3) Confounding and/or variance component structure of an extra variable, the genetic factor for example, can be readily added. With this concern, we implemented a structural equation model (SEM) to explore the interrelations of physiological indices since they are closely related to each other group-by-group, along with a set of ‘baseline’ variables.

Structural equation modeling

Structural equation model (SEM) is often used in psychometrics. It allows one to evaluate causal hypotheses on a set of inter-correlated non-experimental data. Mathematically, SEM can be thought of as a combination of classical path analysis (possibly with latent variables) and the ‘confirmatory factor analysis (‘CFA’, or ‘measurement model’). Recently, the CFA have been proved suitable for evaluating the quality of blood pressure measurements other than the psychiatric data in early researches.(Batista-Foguet) It summarizes the relationships between latent variables in a standard model or between risk factors and outcomes in ‘nonstandard model’. SEM techniques are distinguished by two characteristics: estimation of multiple and interrelated dependence relationships, and, more importantly, the ability to represent unobserved, conceptualized variables in these relationships and account for measurement errors in the estimation process.

1.2. Goal of this study

The collected data, which is described and explored in the next chapter, contains various serum indices (including urine, GOT, GPT,…, etc.), physiological variables (including systolic pressure, diastolic pressure, WBC, RBC, …, etc.), the baseline variables, and some variables representing personal lifestyle in cigarette smoking, drinking, and betel nut eating. (In addition, genetic factors including family data are being collected.) If further information about the status of several chronic diseases such as GOUT, hypertention, DM, was available, these variables/risk factors can be used as predictors of disease under the follow-up study framework. To this end, a number of linear regression or prospective logistic regression models are usually used. That is, the physiological and biochemical measurement of an individual may have power of prediction on several diseases. In the present study, however, no specific diseases status was diagnosed. So the main purpose of this article is to construct a primary model, an SEM, to build a possible inter-relation structure among these variables.

This thesis is organized as follows. Chapter 2 reviews the role and characteristics of the serum biochemical measurements, and the notations and expressions of structural equation model (SEM). Motivations contrasting with the conventional analysis procedure are also addressed. Chapter 3 gives exploratory data analyses for the whole dataset. The results of univariate analyses and pair-wise correlations are reported. Note that the preliminary correlation matrix offers a naïve perspective of the multi-collinearity structure among the covariables, it paves way to a later factor analysis with latent variables. Chapter 4 gives the main result of this thesis, which suggests a **two-stage** algorithm of model construction. First, The 'measurement model' is constructed using a data-dependent exploratory factor analysis (EFA). Second, the structure of measurement model is employed in a construction of the entire structural equation model. In this procedure, goodness-of-fit (GOF) indices are the main criteria to give a valid, or at least, reliable modeling. The complexity of model building task always involves the inclusion and elimination of some variable(s) and/or factor(s). As a first step, we use a multivariate regression analysis for each *primary* 'univariate' variables or for each *secondary* **factor score** consisting of several primary variables in the same factor. Different rules for the assessment of contribution to the significance of a 'factor' can be used to select some possible models. As a final step, the marginal correlation structure of all observed variables serves as a tool to give a 'final' model, in terms of the GOF indices. In Chapter 5, we give some discussion about our results.

Chapter 2 Notations and Literature Review

2.1 On the biochemical values

In the natural history of a chronic disease, pre-clinical symptoms are usually not apparent to be diagnosed. In spite of this, a public-health concern is to seek for suitable indicators obtained from serum samples in regular examinations. Laboratory tests are then used as overall physical assessments to detect some abnormal results. Routinely performed tests include hemoglobin, red blood cell count, cholesterol, triglycerides, total lymphocyte count, serum albumin, etc. Other tests such as pla telet, globulin, glucose, AST, ALT, BuN, creatinine, and uric acid also provide non-ignorable information. In this thesis, all these values are called physiological or biochemical indices. They are used as variables to be classified into several groups or factors.

In a general classification, white blood cell, red blood cell, hemoglobin, platelet are usually grouped together and treated as being related to the "**function of blood manufacturing**". A group of "**cardiovascular function**" includes systolic and diastolic blood pressures, cholesterol level, triglycerides, HDL-C/LDL-C. Another group related to "**liver function**" includes the synthesis of albumin and globulin, AST, and ALT. The other group of "**kidney function**" is composed of nitrogen balance, creatinine, serum uric acid and one about metabolism and nephritic absorption, blood sugar. In the following, we describe some characteristics and functions of these indicators.

Specific indicators

(1) **Uric acid** is synthesized in liver, and excreted from kidney and intestine. In blood, a part of ion of uric acid combine with albumin, some exists with an ion type, and most of them exist in body fluid outside of vessel. The rates of decomposition and synthesis of protein balance each other.

(2) **Blood pressure** (BP) is the force of the blood pushing against the side of

vessel wall. The systolic pressure is the maximum pressure felt on the artery during left ventricular contraction. The diastolic pressure is the elastic recoil, or resting, pressure that the blood exerts constantly between each contraction. These variations come from age, sex, race, rhythm, weight, exercise, emotions, stress, and so on.

(3) **Red blood cell** count is the total number of hemoglobin per cubic millimeter in blood. The hemoglobin is an important part in red blood cell, and carries oxygen around our bodies. If a person suffers from anaemia, their red blood cell count and hemoglobin will always be under a normal level.

(4) **White blood cell** count and type are the most commonly used tests of immune function. The number of white blood cells increases as a result of bacterial infection, bleeding, fever, inflammation, metabolism, and smoking and decreases due to antibodies which induce the autoimmune response.

(5) **Platelets** are very small cells in the blood, and their major function is blood clotting. The decrease of platelets number will increase the chance of bleeding, even without injury. The mechanism involves autoimmune, chemotherapy, leukemia, viral infection, anaemia. An increase of number will make more blood clots, this involves bone marrow, splenectomy, etc.

(6) **Serum albumin and globulin** make up most of proteins and their major functions are to provide nutrition for our body tissues. Hyperproteinemia is the major result of globulin. The albumin and globulin major synthesize in liver. In clinical aspects, decreases in concentration of albumin may due to hunger, malnutrition, synthesis velocity, liver cirrhosis, kidney syndrome, and infection, etc.

(7) In human body, there are two important kinds of aminotransferase: aspartate oxaloacetate transaminase (GOT/**AST**), and alanine pyruvate transaminase (GPT/**ALT**). They are used to detect the damage in liver. Note that GOT also exists in brain, heart, and blood cell, the increase in GOT-value may imply health problems of related organs. The major function of a liver relates to metabolism, storage, phagocytosis, and maintain plasma capacity and concentration.

(8) **Glycogen** is the important part of sugar, and it stored by high concentration in

liver and muscle, and supplies the large number of energy for all body tissue by blood circulation. The glucose dissolution produces most of energy for the demand of human body. A greater part of the glucose in lipocyte will become lipide and be stored in lipidic tissues. Insulin can promote the glucose to form lipide. A part of the glucose in lipocyte changes to glycogen in the muscle tissues, and the process is also affected by insulin. As what is known, diabetes-status is a risk factor for the cardiovascular disease, and the cardiovascular disease can also lead to diabetes. A high level of blood sugar thus indicates problems in these organs or in hormone.

(9) The **lipids** of human body contains **triglyceride** and **cholesterol**. Clinically, many diseases relate to lipoprotein change. Lipoprotein is a combination of lipid and protein (for example, triglyceride combines with alpha-globulin). Most of lipids combine with globulin. The sources of triglyceride and cholesterol are food and synthesized by liver. The causes of a high triglyceride level are due to diabetes, arteriosclerosis, kidney syndrome, hypothyroidism, hungry, diet, obesity, obstructive jaundice, acute/chronic pancreatitis, uremia, alcohol, hormone. The reasons for a low level are beta-lipoprotein deficiency, liver diseases, absorption deficiency syndrome, heparin use, or the problems in metabolism function. The value of serum triglyceride is age-dependent, and can be used for a screening of hyperlipidemia and to determine the risk of coronary artery disease. Moreover, total cholesterol is measured to evaluate fat metabolism and to assess the risk of cardiovascular disease. The normal range of a cholesterol level varies with age and gender.

(10) Nitrogen balance (**BuN**) is a basic item of **kidney** examination, it is also an index of protein nutritional status. Nitrogen is released with the metabolism of amino acids, and the final production is urea. The concentration of BuN in blood is determined by protein ingestion and excretive rate of kidney.

(11) **Creatinine** is derived from the breakdown of creatine through the synthesis of liver. It is not affected by protein ingestion and excreted unchanged in the urine at a constant rate. Thus the increase of concentration of creatinine in blood indicates the kidney function deficiency. The level of creatinine depends on individual weight,

height, gender, and age. It is sometimes viewed as an indicator of ***ageing***.

(12) **Uric acid** is released with metabolism of amino acid through purine. A high level of uric acid is due to hungry, obesity, hyperlipide ingestion, and alcohol. Hyperlipide produces ketone and alcohol restrains the excretive function of uric acid in kidney. Furthermore, diuretic, adrenalin, Niacin, Ethambutol, L-DOPA affect the uric acid le vel. The complication of hyperuricemia is an increase in red blood cell, leukemia, kidney function deficiency, or hypertension. The value of uric acid is also age-dependent.

The common ranges of the biochemical data

Blood pressure : 120 ~ 159mmHg in **SBP**; 80 ~ 90mmHg in **DBP**.

WBC (white blood cell) : 5000 ~9000/ul.

RBC (red blood cell) : 4.5 ~ 5.5×10^6/ul for male; 4.0 ~ 5.0×10^6/ul for female.

Hb (hemoglobin) : 14 ~ 18g/dl for male; 12 ~ 16g/dl for female.

PLA (platelet) : 140 ~ 350×10^3/ul.

ALB (albumin) : 3.7 ~ 5.2 mg/dl.

GLO (globulin) : about 2.4 mg/dl.

Liver function: **AST** or **ALT** is higher than 40 U/ml and it is defined abnormal.

BS (fasting blood sugar): 60 ~ 120 mg/dl.

CHO (cholesterol level) : 130 ~ 225 mg/dl.

Triglyceride **is 25 ~ 150 mg/dl. The definition of hypertriglyceridemia is higher than 200mg/dl.**

In kidney function **blood urea** nitrogen is higher than 22mg/dl or **creatinine** is higher than 1.2mg/dl.

Uric acid is 3.5 ~ 7.2mg/dl. It is abnormal when UA is higher than 7mg/dl for male or 6mg/dl for female.

2.2 Notations

Structural equation model (SEM)

Notations adopted in this thesis are the ‘**LISREL notation**’ system. The SEM is composed of two primary components: a structural model and a measurement model. The structural model is ? = G?+?. Two equations for the measurement model are y =? $_{y}$?+e and x =? $_{x}$?+d. There are two kinds of variables: endogenous variables and exogenous variables. ‘Endogenous’ refers to variables that are influenced by other variables in SEM and ”exogenous” describes variables that are determined outside of the model system. The ‘matrix’ expression of the SEM structure include three parts: the structural model, the measurement model, and the covariance matrices. In this section, we take Figure 2.1 as an illustration of our notations.

Structural model:

$$\begin{pmatrix} h_1 \\ h_2 \\ h_3 \\ h_4 \end{pmatrix} = \begin{pmatrix} g_{11} & g_{21} & g_{31} & 0 \\ 0 & 0 & 0 & g_{42} \\ g_{13} & 0 & 0 & 0 \\ g_{14} & 0 & 0 & g_{44} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} + \begin{pmatrix} V_1 \\ V_2 \\ V_3 \\ V_4 \end{pmatrix}$$

Measurement model:

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} + \begin{pmatrix} d_1 \\ d_2 \\ d_3 \\ d_4 \end{pmatrix} \text{(ds can be equal to zero.)}$$

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ M \\ y_{16} \end{pmatrix} = \begin{pmatrix} l_{11} & 0 & 0 & 0 \\ M & M & M & M \\ 0 & l_{52} & 0 & 0 \\ M & M & M & M \\ 0 & 0 & l_{83} & 0 \\ M & M & M & M \\ 0 & 0 & 0 & l_{164} \end{pmatrix} \begin{pmatrix} h_1 \\ h_2 \\ h_3 \\ h_4 \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \\ M \\ e_{16} \end{pmatrix}$$

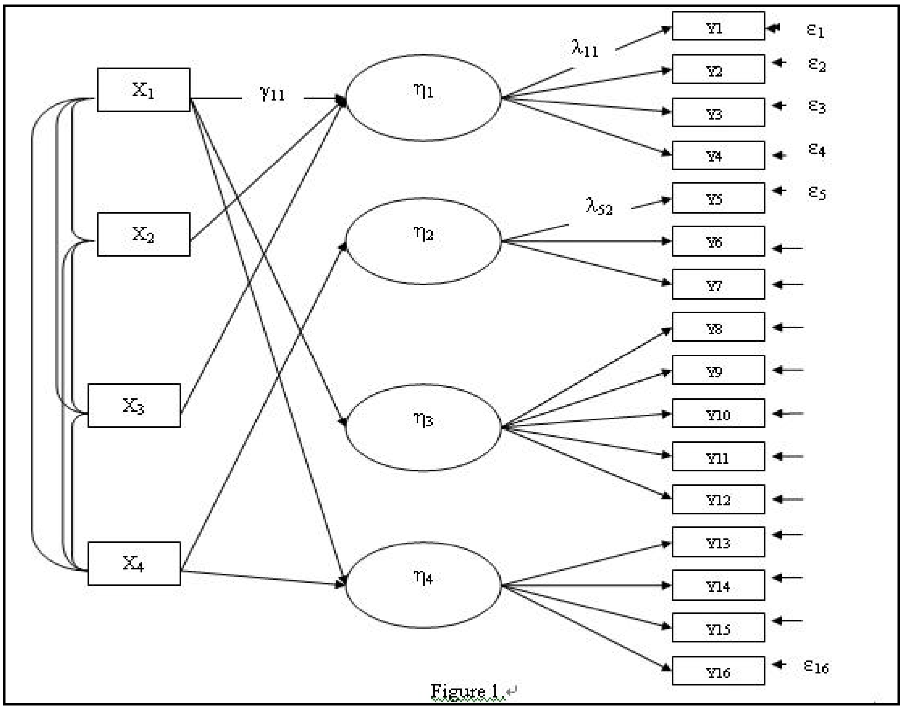
Covariance matrices:

$$\text{Var(X)=}\Phi = \begin{pmatrix} f_{11} & f_{12} & f_{13} & f_{14} \\ f_{21} & f_{22} & f_{23} & f_{24} \\ f_{31} & f_{32} & f_{33} & f_{34} \\ f_{41} & f_{42} & f_{43} & f_{44} \end{pmatrix}, \quad \Psi = \begin{pmatrix} y_{11} & y_{12} & y_{13} & y_{14} \\ y_{21} & y_{22} & y_{23} & y_{24} \\ y_{31} & y_{32} & y_{33} & y_{34} \\ y_{41} & y_{42} & y_{43} & y_{44} \end{pmatrix}$$

$$\mathrm{Var}(\delta)=\Theta_d=\begin{pmatrix}0&0&0&0\\0&0&0&0\\0&0&0&0\\0&0&0&0\end{pmatrix},\quad \Theta_e=\begin{pmatrix}\Theta_{e_{11}}&0&\mathrm{L}&0\\0&\Theta_{e_{22}}&\mathrm{L}&0\\\mathrm{M}&\mathrm{O}&\mathrm{M}&\mathrm{M}\\0&\mathrm{L}&0&\Theta_{e_{1616}}\end{pmatrix}$$

The y-variables are the 16 observed (measured) physiological/biochemical indices which are expressed as y=?×?+e. The ?-variables are the latent endogenous variables with errors ?s. The x-variables are the exogenous variables obtained with or without errors. If there is no error (d) in x, then ?=x. The parameters of major interests to be estimated are ? (the factor loading of a ‘y’ with respect to an ’?’) and ? (the effect of an ‘x’ or ‘?’ on a factor ’?’). In order to obtain valid estimates, the covariance parameters are usually solved simultaneously with the parameters of interests. Finally, F and ? are covariances of ? and ?, respectively.

Fig. 2.1 An illustration of notations using an SEM structure obtained in Chapter 4.



2.3 Literature review

The SEM approach to statistical analysis is largely studied in econometrical and psychometrical literatures as well as in behavioral sciences, clinical researches in nursing, and the field of hospital management, etc. General studies of development of SEM methodology include Bollen (1989), Mueller (1996), and Wan (2002). Bollen (2001) also provided a simple introduction to the theory, notations, and statistical issues of SEM. With SEM method, several systems of analysis packages (among others) have been developed:

(1) SAS PROC CALIS (SAS Inc., Version 8.2) contains **unconstrained** estimation of measurement model (CFA) as well as the entire SEM. It provides generalized least squares (GLS option), weighted least squares (ADF option), and maximum likelihood estimates (ML option) in the MODEL statement.

(2) LISREL (Jöreskog and Sörbom, 1992, LISREL 8) offers **constrained** estimation of CFA and SEM components. It is user friendly but suffers for *convergence problem* (in our experience!) if data analyzed is not suitably standardized in some situation.

(3) EQS (Bentler, 1989) is developed as a simple version of LISREL.

In order to obtain a final structural model, first one has to obtain a measurement model based on a confirmatory factor analysis (CFA). The CFA framework is usually executed with the knowledge of which variable(s) should be grouped together and which should not. And then, based on the construct of the measurement model obtained from CFA, an SEM is fitted. There are two scenarios played in the procedures of constructing an SEM: the **one-stage** and **two-stage** approaches. In a one-stage approach (or simultaneous estimation), parameters are estimated through maximum likelihood method, for example, in a simultaneous estimation procedure, in which, however, a problem of non-convergence is often encountered. In a two-stage approach, on the other hand, the parameter in a measurement model (CFA) is firstly estimated and the entire part is then used as **fixed** to undergo the construct of a likelihood in the coming estimation.

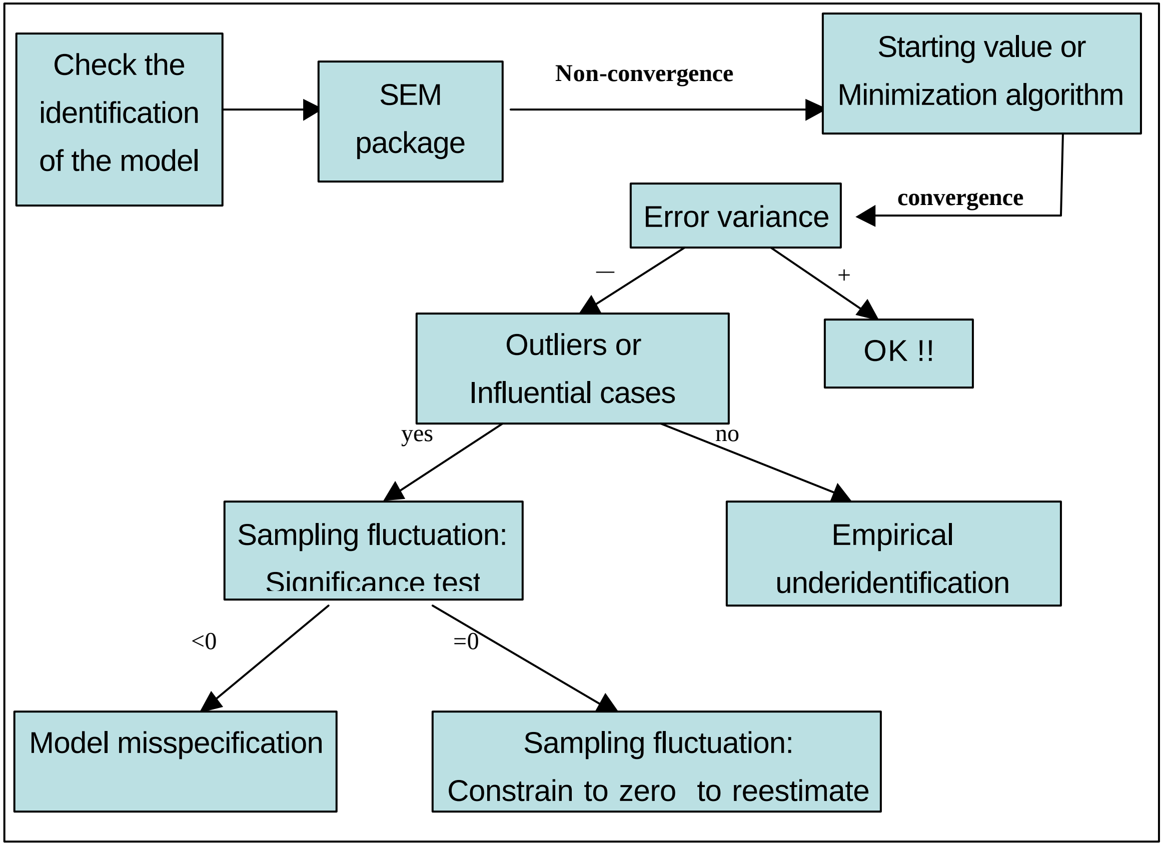
In our problem, however, there is no confirmatory part based on physiological reasons. It is thus appealing to use an exploratory factor analysis (EFA) and its corresponding result to build a measurement model. A question arises since the errors or residuals of an EFA are not correlated but those of a CFA are correlated. This induces a concern about the procedure of constructing a full SEM. In this thesis, we suggest a **hybrid** approach to the construction of the SEM by the following three steps: (I) To give an EFA analysis for the observed variables (y) in order to obtain a primary measurement model; (II) to construct the SEM according to some preliminary analyses on the inter-relations between the observed/latent variables; and (III) since the previous step introduces some correlations between the errors (or residuals), a correlation structure of the observed variables (y) is considered in a stepwise manner to improve the fit. We call this hybrid procedure a **two-stage construct of SEM with a 'simultaneous' (rather than 'two-stage') estimation.**

For our dataset and whole research structure, population and family data of disease status, genotype, and other variables are still in collection. Before we can try to implement an SEM analysis on a future (more complete) dataset, an application of the SEM method to the present cross-sectional data serves as a premise to further statistical analysis. For example, on the stand of population-level, heritability estimation based on population and/or family data is of interest. (Pausova et al.2001). On a lower but more structured level, Province et al.(2001) use path analysis modeling to estimate familial aggregation and heritability; and Williams (1999) use a variance component analysis, along with the knowledge of genetic segregation, to give a linkage analysis. These also motivate our present study of SEM (using our present and future datasets).

For the techniques of implementing a system of structural equations, several aspects of data characteristics need to be checked. For example, if the observed variables are seriously skewed, a robust approach via transformation of variables can be considered (K.H.Yuan et al.2000). Second, if **non-convergence** problem and/or **improper solution** are encountered, guidelines of a model-building procedure have to be taken.

In what follows, we use the suggestion of Chen et al. (1999), which is summarized and expressed in the following diagram (Figure 2.2).

Fig. 2.2 A diagram of possible chart for model fitting techniques.


Check the identification of the model
SEM package
Non-convergence
Starting value or Minimization algorithm
convergence
Error variance
—
+
Outliers or Influential cases
OK !!
yes
no
Sampling fluctuation: Significance test
Empirical underidentification
<0
=0
Model misspecification
Sampling fluctuation: Constrain to zero to reestimate

Chapter 3 Materials and Preliminary Analyses

3.1 Dataset

Our dataset is collected and offered by Dr. Li-Hsin Lai and his staff of health section of Hsin-Yi township, Nantou County of Taiwan. A cross-sectional community-based survey and screening program was conducted on 2,565 adult participants (aged 40 or older) during a health examination. Demographic data and variables concerning life-style are obtained from questionnaire; biochemical values are from blood extraction. The screening rate (or participation proportion) was 45.8%. There was no significant difference between the participants and non-participants in terms of the age, sex, and race structure/distributions of the whole township. In the sample, there were 1226 (47.8%) males and 1339 (52.2%) females, and 1318 (51.4%) aborigines and 1247 (48.6%) non-aborigines.

Covariables and measurements

In our study, we used age, life-style (smoking, drinking, betel nut chewing) and as risk factors. In particular, they were treated as ***exogenous*** variables in the SEM context. ***Endogenous*** variables mainly consisted of the physiological and biochemical measurements such as systolic blood pressure (SBP), diastolic blood pressure (DBP), white blood cell (WBC), red blood cell (RBC), hemoglobin (Hb), platelet (PLA), albumin (ALB), globulin (GLO), AST (GOT), ALT (GPT), blood sugar (BS), cholesterol (CHO), triglyceride (TRI), blood urea nitrogenk (BuN), creatinine (CRE), and uric acid level (UA). These instruments included Sysmex-100 in blood examination, Hitach 704 in biochemical examination, UA by enzymatic-color method, blood sugar by oxidase method and cholesterol and triglyceride by oxidase--peoxidase method and glycorokinas-glycerophosphate-oxidase-peroxidase method, respectively. For all of the above measurements, they was performed by a standard checkout of lab condition.

3.2 A preliminary of data analysis procedure

As the first step of exploratory data analysis (EDA), the characteristics of all variables/covariables should be analyzed. There were three types of variables in our data: continuous variables (age and all physiological/biochemical measurements), indicator variables (sex and race), and ordinal variables (concerning life-style). Basically, all these variables are ‘primary’ or ‘directly observed’. In some researches, ‘secondary’ variable such as BMI is also considered as a confounder which is defined by other primary variables. In our analysis, we only study the **linear** relationship among all these **directly observable variables** so the secondary variables are not included in the analysis. For discrete variables, we presented frequencies; for continuous variables, sample means and standard deviations are calculated and histograms (with smoothing loess curve estimates) are plotted. Because the relationships among these physiological/biochemical values are of interests, we presented the pairwise correlation matrix.

The second step is to construct the measurement model of an SEM. According to a previous context, there is no clear and evident classification for functional target of all the physiological/biochemical measurements of human body in clinical test, we used exploratory factor analysis (EFA) for the purpose of classification. The 16-item physiological/biochemical values were subjected to an exploratory factor analysis using the squared multiple correlations as prior communality estimates (L. Hatcher, 1998). The maximum likelihood (ML) method was used to extract the factors, and ploted the factor pattern before rotation. The **scree test** and the rule of ‘**eigenvalu-one**’ suggested a solution of **four factors** that will be retained for further analysis (L. Hatcher, 1998). As a result, factors 1 to 4 accounts for nearly 100% of the total sum of squares. This classification was then treated as being useful for further development of confirmatory factor analysis (CFA). The measurement model comprised of *latent* endogenous variable and *observed* endogenous variables; it was tested and renewed until a statistically acceptable model, in terms of 'good fits', was obtained.

In order to select a construct of SEM, we compared factor scores in EFA and CFA and carried out multiple linear regression analysis for each factor score. However, there were different choices of factor scores. For examples, Bartlett (1937,1938) and Thomson (1951) suggested $Y=X\hat{\Psi}^{-1}\hat{\Lambda}(\hat{\Lambda}'\hat{\Psi}^{-1}\hat{\Lambda})^{-1}$ and $Y=X\hat{\Psi}^{-1}\hat{\Lambda}(I+\hat{\Lambda}'\hat{\Psi}^{-1}\hat{\Lambda})^{-1}$, respectively. The Bartlett's score is hereafter referred to as a **naïve** method since $Y=V\hat{\Lambda}_0(\hat{\Lambda}'_0\hat{\Lambda}_0)^{-1}$. **($V= X\hat{\Psi}^{-1/2}$,$\hat{\Lambda}_0=\hat{\Psi}^{-1/2}\hat{\Lambda}$, $\hat{\Psi}$ is symmetric.)** So the factor scores produced by the naïve method could be compared to the Thomson's scores. Next, we carried out multiple regressions for factor score each in individual. That means we only considered one response (dependent variable) at a time, and the response variable can be the unobserved latent factor or the observed physiological or biochemical measurements. From the analysis, we recorded significant level and used some criteria to produce a construct of measurement model. In the procedure of model fitting, several goodness-of-fit indices were employed as indices of model adequacy.

3.3 Exploratory data analysis

Descriptive statistics

Descriptive statistics, as well as the distributions, of all exogenous and endogenous variables are reported in Table 3.1 and Figure 3.1. The joint distribution of sex and race of this sample is not significantly different from the distribution of the entire Hsin-Yi area for those aged 40 or older. Concerning the life-style variables, there are 76.22% of nonsmoking, 62.53% of non-drinking, and 72.28% of people without chewing betel nut. The mean age is 58.01 years old (standard deviation= 12.02); the mean and standard deviation for physiological/biochemical values are 134.17±22.18 mmHg (SBP), 80.45±12.99 mmHg (DBP), 6940.34±1971.14 /ul (WBC), 14.28±1.54 g/dl (Hb), 4.65±0.48 ×10^6µl (RBC), 232.96±70.07 (PLA), 4.36±0.27 mg/dl (ALB), 3.01±0.31 mg/dl (GLO), 34.57±26.36U/ml (AST), 34.97±28.30 U/ml (ALT), 103.58±48.05 mg/dl (BS), 193.28±42.02 mg/dl (CHO), 174.82±188.27 mg/dl (TRI), 15.42±4.82 mg/dl (BuN), 1.07±0.36 mg/dl (CRE), 7.01±2.05 mg/dl (UA), respectively. To compare with the common range, we found that most of the observations falls into the common range except for uric acid. The uric acid level is obviously higher than the general population. Health-related problems of this community such as hyperuricemia and gout are thus important.

Pairwise correlation matrix

Conventional factor analysis and principal component analysis rely heavily on the structure of inter-correlations among the variables studied. By calculating the pairwise correlations of physiological/biochemical data, it offers insight into the factor analysis. For example, SBP and DBP are both used to check the *blood pressure* and they surely have a high correlation. Similarly, AST and ALT, Hb and RBC, BuN and CRE, are used to check the *liver function*, *blood function/anemia*, and *kidney function* respectively. All pairs have high correlations. We draw the color with dark or light to represent different levels of correlations. By a suitable **alignment**, the pattern of **clusters** could be determined from the correlation matrix. Nonetheless, some mathematical techniques is yet developed (at least in this thesis) and compared to the conventional principal component analysis (PCA) or factor analysis. From this matrix, on the other hand, we only distinguished (**roughly**) four clusters from PCA (Table 3.2). The variables SBP, DBP, WBC, and CHO are treated as from a factor connected with **cardiovascular function**; the variables GLO, AST, and ALT are grouped and thought to be associated with **liver function**. We continued this process to group Hb, RBC, PLA, ALB, and TRI, connected with **manufacture blood function**, or **quality of blood**. Finally, BS, BuN, CRE, and UA are combined into one group correlated with **metabolism, excrete, and kidney function**. These groupings will be further checked and confirmed by exploratory factor analysis reported in the next chapter.

Table 3.1 Descriptive statistics of the exogenous and endogenous variables

Predictors		Frequency	Percent (%)	
Sex	Male	1226	47.80	
	Female	1339	52.20	
Race	Non	1247	48.62	
	Aborigine	1318	51.38	
Smoke	Never	1955	76.22	
	Sometimes	88	3.43	
	Often	12	0.47	
	Everyday	510	19.88	
Drink	Never	1604	62.53	
	Sometimes	725	28.27	
	Often	37	1.44	
	Everyday	199	7.76	
Betel nut	Never	1854	72.28	
	Sometimes	465	18.13	
	Often	8	0.31	
	Everyday	238	9.28	
Variable	**Mean**	**Std Dev**	**Minimum**	**Maximum**
AGE	58.01	12.02	39.43	93.71
SBP	134.17	22.18	75.00	244.00
DBP	80.45	12.99	32.00	155.00
WBC	6940.34	1971.14	3100.00	29500.00
Hb	14.28	1.54	4.90	18.90
RBC	4.65	0.48	2.86	6.99
PLA	232.96	70.07	25.00	536.00
ALB	4.36	0.27	2.20	5.60
GLO	3.01	0.31	2.10	4.50
AST	34.57	26.36	11.00	612.00
ALT	34.97	28.30	10.00	391.00
BS	103.58	48.05	54.00	538.00
CHO	193.28	42.02	78.00	391.00
TRI	174.82	188.27	32.00	3063.00
BuN	15.42	4.82	9.00	78.50
CRE	1.07	0.36	0.70	8.30
UA	7.01	2.05	2.10	17.20

Fig. 3.1 The distributions of eighteen continuous variables

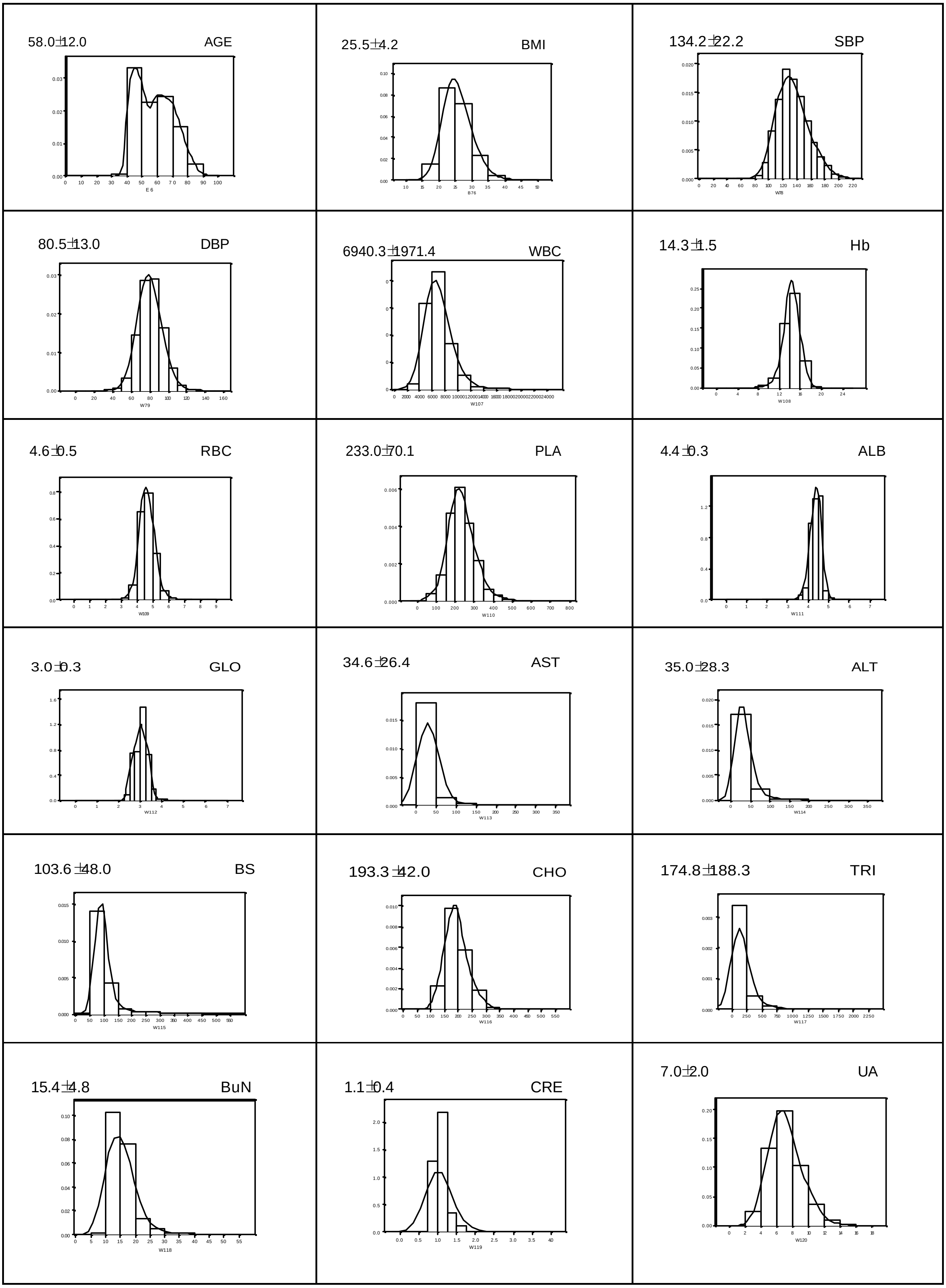


Table3.2 Pairwise correlation matrix for 16 physiological/biochemical variables

variables	SBP	DBP	WBC	CHO	GLO	AST	ALT	Hb	RBC	PLA	ALB	TRI	BS	BuN	CRE	UA
SBP																
DBP	0.7695 <.0001															
WBC	0.1095 <.0001	0.1195 <.0001														
CHO	0.1009 <.0001	0.0934 <.0001	0.0970 <.0001													
GLO	0.0907 <.0001	0.0780 <.0001	0.1046 <.0001	0.0337 0.0882												
AST	0.0205 0.2996	0.0143 0.4680	-.0161 0.4153	-.0631 0.0014	0.2267 <.0001											
ALT	0.0336 0.0889	0.0356 0.0712	0.0573 0.0037	-.0051 0.7947	0.1830 <.0001	0.7549 <.0001										
Hb	0.0575 0.0036	0.1228 <.0001	0.1116 <.0001	0.1243 <.0001	-.0316 0.1098	0.0729 0.0002	0.1487 <.0001									
RBC	0.0304 0.1234	0.0940 <.0001	0.0871 <.0001	0.0903 <.0001	-.0914 <.0001	-.0660 0.0008	0.0403 0.0411	0.6044 <.0001								
PLA	-.0294 0.1360	0.0154 0.4360	0.2075 <.0001	0.1125 <.0001	0.0136 0.4918	-.0982 <.0001	-.0787 <.0001	-.1423 <.0001	-.0286 0.1477							
ALB	0.0353 0.0743	0.0667 0.0007	0.0603 0.0022	0.1953 <.0001	-.0338 0.0872	-.0921 <.0001	0.0225 0.2552	0.2713 <.0001	0.3096 <.0001	0.0994 <.0001						
TRI	0.0849 <.0001	0.1263 <.0001	0.0744 0.0002	0.3105 <.0001	0.1643 <.0001	0.1267 <.0001	0.1233 <.0001	0.1287 <.0001	0.0072 0.7163	0.0751 0.0001	0.0805 <.0001					
BS	0.0762 0.0001	0.0576 0.0035	0.0415 0.0357	0.0759 0.0001	0.1094 <.0001	0.0462 0.0193	0.0933 <.0001	0.0509 0.0099	0.0140 0.4774	-.0363 0.0658	0.0115 0.5622	0.2709 <.0001				
BuN	0.0557 0.0048	0.0049 0.8058	0.1629 <.0001	0.1383 <.0001	0.0227 0.2496	-.0712 0.0003	-.0497 0.0118	-.1474 <.0001	-.1286 <.0001	-.0256 0.1949	-.0125 0.5267	-.0034 0.8653	0.0696 0.0004			
CRE	0.0726 0.0002	0.0668 0.0007	0.0870 <.0001	0.0506 0.0103	0.0031 0.8768	-.0175 0.3761	-.0189 0.3396	-.0235 0.2347	-.0485 0.0140	-.0330 0.0950	-.0227 0.2503	0.0305 0.1220	0.0541 0.0062	0.5837 <.0001		
UA	0.1008 <.0001	0.1410 <.0001	0.2197 <.0001	0.0401 0.0424	0.1981 <.0001	0.1752 <.0001	0.1605 <.0001	0.2028 <.0001	0.0699 0.0004	0.0653 0.0009	-.0010 0.9592	0.2496 <.0001	-.0132 .5043	0.1751 <.0001	0.2061 <.0001	

Chapter 4 Statistical Analysis and Main Results

In sociological researches, often there is a hypothetical structure, which is referred to as a 'model', and data or information collection is processed through a (structured) questionnaire and accompanied statistical analysis to validate the model. The most common adopted statistical method is the structural equation modeling (SEM). However, it should not be a *paradigm* since this 'model' or 'procedure' *a priori* of scientific research tells no more story than a result obtained from that without a model assumption. That means, a scientific model could also be set-up by a procedure with a '**meta**'-sense in that we can still build a *posterior* model after the analysis is completed, if there is some **interpretability** concerning the results.

For our health-related data, it is still lack of theoretical support and physiological/pathological evidence or reasons that which variables should be grouped together and, likewise, the mechanism and causal results of a variables/index on the other(s) and their recursive relationships are also unknown. For a set of variables collected from a cross-sectional sample consisted of aged people, the inter-relations between variables may have different pattern from the reasoning of physiological/pathological aspects. For example, SBP, DBP, and WBC grouped in a factor cannot be over-interpreted in that they do have causal relationship in the **formation** of chronic diseases. On the other hand, they should be viewed as being common **results** related to an unknown, unobservable pathway through a latent factor. In this regard, a confirmatory factor based on medical knowledge and an exploratory factor reflected from a '**prevalence** data' give no confliction.

4.1 The measurement model in SEM

There is a question about the 'adequacy' of giving a factor analysis before it is executed. To this end, a **Kaiser's** pre-analysis measure can serve to judge the 'level' of adequacy. For our data, the overall **MSA** (measure of sampling adequacy) is 0.563 and variable-specific measures are:

SBP	DBP	WBC	CHO	GLO	AST	ALT	Hb
0.52	0.53	0.63	0.58	0.77	0.53	0.54	0.58
RBC	PLA	ALB	TRI	BS	BuN	CRE	UA
0.60	0.44	0.71	0.59	0.52	0.52	0.55	0.65

As an usually experience, the level of 0.5 is recognized as an lowest acceptable threshold; and the level of 0.7 or above as being promising for a good factor analysis (H-J Chiou). It is noted from the above result that most of the ‘adequacy’ level lie within 0.5 to 0.6, the level of PLA is 0.44, and those of GLO and ALB are greater than 0.7. It is treated as being feasible to undergo a factor analysis. Another measure of sampling adequacy is the *communality*, which will be discussed later for specific models.

In an exploratory factor analysis, each observed variable y_1, y_2, …, y_p of a *centered* random vector y is assumed to be a linear combination of m factors f_1, f_2, …, f_m :

$$\mathbf{y_1-\mu_1=\lambda_{11}f_1+\lambda_{12}f_2+\cdots+\lambda_{1m}f_m+\varepsilon_1}$$

$$\mathbf{y_2-\mu_2=\lambda_{21}f_1+\lambda_{22}f_2+\cdots+\lambda_{2m}f_m+\varepsilon_2}$$

…………

$$\mathbf{y_p-\mu_p=\lambda_{p1}f_1+\lambda_{p2}f_2+\cdots+\lambda_{pm}f_m+\varepsilon_p.}$$

The coefficients λ_{ij} is referred to as the loading of factor j (f_j) on the i-th observed variable y_i. If λ_{i1} is close to zero, for example, it means that the level of y_i which is attributable to factor 1 (f_1) is nearly zero or at least very non-significant. Moreover, with the above expressions, each y_i represents a ‘point’ in the space spanned by factors (f_1,f_2,…,f_m). A suitable presentation of the ‘position’ of the point with respect to an (f_i,f_j)-pair can reveal comparative factor loadings between f_i and f_j of the variable concerned. If one plots the point of a variable y_k in the f_i-f_j plane, for example, and if the position of y_k is very close to the f_i-axis, then the factor loading of y_k with respect to f_i is much larger than that of f_j. In this case we believe that the ‘path’ from f_i

to y_k should be considered, and the path from f_j to y_k may not be an important one. However, if the location of y_k is just between the two axes (or lies around the line f_i=f_j), it means that both of the pathways from the two factors to the variable y_k should be considered.

The exploratory factor analysis (EFA) suggested four factors according to at least one of the following four criteria: **eigenvalue-one** criterion, the **scree-plot** diagnostics, the attributable **proportion** of (centered) sum of squares, and the '**interpretability**' criterion. Of course, more factors also could be considered if the 'interpretability' criterion is greatly emphasized. In an analysis not reported in the context, most of measurement models based on more factors with a great physiological assortment and interpretability do not converge even under a low decimal criterion. The resultant classification of the 16 physiological/biochemical values into four factors is as follows.

Table 4.1 A classification according to exploratory factor analysis.

	Variables (y)
Factor 1 (f_1)	GLO, AST, ALT
Factor 2 (f_2)	SBP, DBP, WBC, CHO
Factor 3 (f_3)	Hb, RBC, PLA, ALB, TRI
Factor 4 (f_4)	BS, BuN, CRE, UA

The plots of each 'y'-variables on various f_i-f_j planes are shown in Figure 4.1. By careful examinations, one can check the above classification through the six plots in that which variable (y) is reasonably classified into which factor (f), as well as the co-influence of two or more factors on the same variable (y). For example, the points A and B (SBP and DBP) are almost no doubt to be classified as Factor 2 (f_1), but the point C (WBC) can be attributed by Factor 2 (f_1) and Factor 4 (f_4). The latter suggested a possible path from Factor 4 to WBC in the measurement model or SEM analysis. Similarly, the points D and E (Hb and RBC) both can be co-attributed by Factors 3 and 4, and so paths from Factor 4 to Hb and RBC are then possible.

Unfortunately, however, later analyses of measurement model and SEM with these 'inter-factor relationship' usually resulted in **improper** solutions, **non-convergence** estimates, and/or worse fits. At a request of parsimony when improper solutions are encountered (Chen, Bollen, et al., 1999), hereafter we will not take the case of 'inter-factor relationship' into consideration in our presentation and analysis. Figure 4.2 gives the estimates associated with the EFA of Table 4.1. When the biggest marginal correlation (CHO and TRI) among the observed variables is considered, goodness-of-fit indices (GFI and AGFI) were improved to a satisfactory level; though, substantial changes in the estimates of each factor loadings are not observed. This construct will be used as an initial 'guess model' for later modeling of SEM except for that more of the inter-variable correlations could be included to improve the fits.

Finally, it is worth noting that the pairwise correlation matrix of Table 3.2 gives a contrast with the results of exploratory factor analysis of Table 4.1.

[Put Figures 4.1 and 4.2 about here.]

4.2 The structural model in SEM

As a primary hypothesis, we assumed that the physiological/biochemical mechanism is not different between races and genders. With this assumption, we did not take race and gender as exogenous variables to make things simple and then the whole dataset was used to pursuit a reasonable model-building procedure in SEM. When it is believed that there is different among genders and/or races, however, more complicated fitting can be considered. For example, a 'stratification' on gender or race is possible.

Full model

Since the inter-relationship between the four factors, reduced fom 16 observed variables, and four risk factors or risk taking behaviors is of major concern, first we draw all possible paths as an initial construct of SEM. It is hereafter referred to as the *full model*. The fit of full model is not a good one (GFI=0.6808, NFI=0.0452, CFI=0.0424). Standardized parameter estimates are presented in Figure 4.3.

In order to improve the fit (in terms of goodness-of-fit indices), significant paths from the exogenous variables to the latent factors (terms as the γ-path) need to be identified and non-significant paths to be deleted. Traditional method concerning this ‘model-selection’ procedure is to use the Lagrange multiplier test (a parallel of Rao’s score test in regression set-up when there exists a likelihood) or the Wald test in a **stepwise** manner. However, when the likelihood of an SEM is written, it involves the **whole** structure of the SEM, including all the covariance parameters, γ-paths, and λ-paths (from factors to the observed y variables). This implies that the stepwise procedure involves a simultaneous estimation of all parameters, not only the γ-paths (or γ-parameters). In this thesis, we propose that the construct of an SEM as a **two-stage** procedure, but the estimation is a **simultaneous** one. In this regard, an alternative (but **naïve**) algorithm based on the two-stage thinking is proposed based on the building stone of univariate-multiple linear regression. ‘Univariate’ means that the outcome can be (i) the univariate observed variable, y, or (ii) a combined factor score; ‘multiple’ indicates that the explanatory variables are the set of risk factors (AGE, SMK, DRI, and PEA). For (ii), we use the naïve score proposed by Bartlett (1937) in which factor loadings are substituted by those parameter estimates obtained from the ML estimation of measurement model. There is another factor score suggested by the SAS system and, for the current dataset, scatter-plots (Figure 4.4) of these two factor scores shows that these two scores are high surrogates to each other.

[Put Figures 4.3 and 4.4 about here.]

Univariate linear regression with observed dependent variables

We used the physiological/biochemical variables as outcome variables and risk factors as predictors and proceeded regression analyses. In this process, we considered one responser (dependent variable) at a time, and the model could have many predictors (independent variables), so it is called a *univariate multiple regression* analysis. According to the analysis, we recorded the significant level and provided some criteria to decide the structural model. The double asterisks represented the p-value less than 0.01, and one asterisk indicated a p-value within 0.01 to 0.05. In each cell, a 'double asterisks' is treated as being a *full mark*. The result of univariate multiple regression analysis was reported in Table 4.2.

Next, total numbers of asterisks of x's (age, smoking, drinking, and betel nut eating) on every observed variable (y) were counted for each factors (f). After this, various criterion rules can be used. (Note that each criterion rule corresponds to a construct.) Examples of considerations on the rules and their interpretation are as follows.

(1) Additive-1/2 rule: The total number of asterisks is greater than or equal to a half of the possible number of asterisks. In this case, the corresponding γ-path is identified as being important. For example, from Table 4.2, since Factor 1 (f_1) consisted of 4 variables, thus there must be 8 possible asterisks in 4 cells for each of the 4 risk factors. As a result, the age-f_1 relation has 6 asterisks, reveals that the γ-path from AGE to Factor 1 should be considered. Similarly, the γ-paths from SMOKE and DRINK to Factor 1 are both important, but that from BETEL NUT to Factor 1 is not. This criterion relies on the additive effect of significance attributed from the relationship between risk factors (x) and **distinct** observed variables (y).

(2) Relative significance rule: If the number of cells (which equals the number of variables related to a factor) with two-asterisks significance level exceeds, or equals to, the total number of cells, the γ-path is considered. This rule is very strict in asking for parsimony in the construct of γ-path.

(3) Strict additive-1/2 rule: Like the rule of (1) except for that the 'equal to'-requirement is cancelled.

(4) Absolute significance rule: When the number of cells with two-asterisks significance level exceeds 2, it is also reasonable to treat the factor to be highly attributable to the x variables in the sense that there are genuine contributions from x to the *combined* observed variables (y) which consisted of the factor (f).

It is important to note that some variants of (1)~(4) or their configurations are also possible. (For details, please refers to the results of Table 4.4.)

Univariate linear regression with latent factor scores

When the naïve factor scores were used as outcome variables, the case of p-value less than 0.01 were further partitioned into two sub-cases: $0.001<p< 0.01$ and $p< 0.001$. We set 3 asterisks to the case of p-value<0.001, 2 asterisks to the case of $0.001<p< 0.01$, and 1 asterisk to that of $0.01<p< 0.05$. The result was presented in Table 4.3.

[Put Tables 4.2 and 4.3 about here.]

According to the results of multiple regression analyses, we borrowed the criterion of **relative significance rule**. (i) If the number of cells (which equals the number of variables related to a factor) with 2 or more asterisks, or, (ii) if, in a more restrict sense, the number of cells with 3 asterisks exceeds or equals to the total number of cells, the γ -path is considered. The first consideration gives the following goodness-of-fit indices: GFI=0.7853, NFI=0.4782, CFI=0.4828; The second one gives GFI=0.8200, NFI=0.5500, CFI=0.5558.

As a final construct by combining the above results, we obtained an SEM shown in Figure 4.5 with the best goodness-of-fit indices with GFI=0.8445 and AGFI=0.7920.

[Put Table 4.4 and Figure 4.5 about here.]

Adding/deleting correlated error terms stepwisely

In order to obtain a satisfactory fit, in terms of goodness-of-fit indices, we tried to add the path of correlations between error terms of observed variables (y) in the order from high to low level of **marginal** correlations (though **partial** correlation also could be considered). We had the following order of correlations: SBP/DBP (0.769), AST/ALT (0.755), Hb/RBC (0.604), BuN/CRE (0.584), CHO/TRI (0.311), RBC/ALB (0.310), Hb/ALB (0.271), TRI/BS (0.271), TRI/UA (0.250), GLO/AST (0.227), WBC/UA (0.220), WBC/PLA (0.208), CRE/UA (0.206), Hb/UA (0.203). According to this order, we added the path between variables (y) one at a time. This is why we called it a '**stepwise**' manner. It is very different from the standard procedure suggested by most of the statistical packages that the selection of path is decided by a Wald test or a Lagrange multiplier test. The reason is, as stated previously, we seek to add the path(s) by a 'two-stage' manner. After the global structure is constructed, we can add the correlation terms by considering the 'correlations' between the observed variables (y) to raise the goodness-of-fit indices (GFI or AGFI, etc.) to a acceptable level. As suggested by this thesis, the primary pairwise (marginal) correlations can be used in a *descending* order. The result is shown in Table 4.5, in which adding the marginal correlation greater than 0.2 will finally give a GFI index greater than 0.90. On the other hand, if the Lagrange multiplier test is used *from this stage* (as suggested by the statistical packages) *without regards to the parts other than the correlations between error terms*, a model-building procedure can also be adopted. We contrasted these two procedures, in terms of the GFI/AGFI index, by Figures 4.6 and 4.7, respectively. It demonstrates the growth rate of GFI/AGFI and tells the betterment of our procedure at the early inclusion of higher correlations. Nonetheless, the Lagrange multiplier test still gives better fits from some step although it still falls into the framework of 'two-stage' modeling. Before a 'final' model is obtained, we can still investigate the '**lack-of-fit**' problem in what the change is when deleting a correlation between the observed variables (y). The results are reported in Table 4.6 and Figure 4.8.

Finally, the magnitude of GFI-change when deleting one path of correlation in a '**backward'** manner is shown in Figure 4.8; and a 'final' model is given in Figure 4.9.

[Put Tables 4.5 and 4.6 about here.]

[Put Figures 4.6, 4.7, 4.8, and 4.9 about here.]

4.3 Enhancing the model

As stated in a previous context, there may be sex and/or race difference in the distributions of indicator variables of serum sample, but the mechanism or structure among all variables discussed is believed to be the same. There are at least two ways to deal with the effect of causal or confounding effect introduced by sex and race. They are very similar to the discussion in regression setting of statistical and epidemiological fields in which we consider two ways of treating confounding effect. Hereafter, we will further consider variables age and race to enhance the power and feasibility of an SEM model. That is, to add the sex/race variable into the structure as an exogenous variable or to use sex/race as a stratification variable. *We illustrate the case of using race as an exogenous variable and sex as a stratification variable*. Model building procedure follows what has been taken in this chapter except for the third step of adding correlations between observed variables (y) is neglected. The results are shown in Figures 4.10 and 4.11. Comparison of parameter estimates of different genders are attached at the end of these two figures using SAS PROC CALIS (unconstrained estimates).

[Put Figures 4.10 and 4.11 about here.]

Table 4.2. Univarate linear regression analysis

Y=ß$_0$+ß$_1$(age)+ ß$_2$(smoke)+ ß$_3$(drink)+ ß$_4$(Betel nut)

Y	p-value of $\hat{b}$ (t-test)				F-test
	AGE (x_1)	SMOKE (x_2)	DRINK (x3)	BETEL NUT (x_4)	$x_1+x_2+x_3+x_4$
SBP	**	**	**	**	**
DBP		**	**	*	**
WBC	**				*
CHO	**				**
GLO		**	*	**	**
AST			**	**	**
ALT	**	*		**	**
Hb	**	**	**		**
RBC	**	**			**
PLA	**				**
ALB	**	*			**
TRI	**		**	**	**
BS		**	**		**
BuN	**				**
CRE	**	**	*	**	**
UA			**	**	**

Table 4.3 Univarate linear regression analysis

f=ß$_0$+ß$_1$(age)+ ß$_2$(smoke)+ ß$_3$(drink)+ ß$_4$(Betel nut)

Y	p-value of $\hat{b}$ (t-teast)				F-test
	AGE (x_1)	SMOKE (x_2)	DRINK (x3)	BETEL NUT (x4)	$x_1+x_2+x_3+x_4$
FACTOR1	***	***	***	*	***
FACTOR2		**	***	***	***
FACTOR3	***	***			***
FACTOR4	***			***	***

***：p<0.001

**：0.001<p<0.01

*：0.01<p<0.05

Table 4.4 To determine the **γ** -paths based on univariate linear regressions and several rules

Criteria Y: observed	Total no.(**&*)≧ 1/2 no. (**&*)	No. of cell(**)≧ 1/2 no. of cell	Total no.(**&*)＞ a half no. (**&*)	No. of cell “**”≧2
Fit function	2.5260	1.8968	6.6842	3.7064
$?^2$	6476.7421	4863.3846	17138.2794	9503.2163
$?^2$/df	6476.7421/152	4863.3846/155	17138.2794/160	9503.2163/153
GFI	0.7992	0.8321	0.7389	0.7120
AGFI	0.7226	0.7725	0.6574	0.6047
NFI	0.4691	0.6014	-0.4048	0.2210
NNFI	0.3417	0.5194	-0.6788	0.0332
CFI	0.4734	0.6080	-0.4137	0.2214
PGFI	0.6394	0.6788	0.6223	0.5733
Criteria: FS	** &***	***	Combine “No. of cell(**)≧ a half no. cell” and “***”	
Fit function	2.4826	2.1412	1.6573	
$?^2$	6365.4117	5490.0760	4249.2828	
$?^2$/df	6365.4117/154	5490.0760/155	4249.2828/157	
GFI	0.7853	0.8200	0.8445	
AGFI	0.7072	0.7561	0.7920	
NFI	0.4782	0.5500	0.6517	
NNFI	0.3619	0.4555	0.5876	
CFI	0.4828	0.5558	0.6593	
PGFI	0.6365	0.6689	0.6978	

Table 4.5 The model-fit indices for the cases of adding correlated error terms in a stepwise (**forward**) manner

Criteria: correlation	Add SBP/DBP (0.769)model A	A+AST/ALT (0.755)model B	B+Hb/RBC (0.604)model C	C+BuN/CRE (0.584)model D	D+CHO/TRI (0.311)model E	E+RBC/ALB (0.310) model F	F+Hb/ALB (0.271)model G
Fit function	1.4284	1.4288	1.3318	1.3489	1.2698	1.2293	1.1644
?^{2}	3662.4104	3663.3917	3414.8134	3458.5558	3255.6965	3152.0024	2985.4060
?^{2}/df	3662.4104/156	3663.3917/155	3414.8134/154	3458.5558/153	3255.6965/152	3152.0024/151	2985.4060/150
GFI	0.8572	0.8572	0.8810	0.8799	0.8842	0.8886	0.8930
AGFI	0.8078	0.8065	0.8377	0.8351	0.8400	0.8451	0.8502
NFI	0.6998	0.6997	0.7201	0.7165	0.7331	0.7416	0.7553
NNFI	0.6444	0.6419	0.6650	0.6582	0.6770	0.6856	0.7009
CFI	0.7080	0.7079	0.7285	0.7248	0.7416	0.7501	0.7639
PGFI	0.7038	0.6993	0.7141	0.7085	0.7073	0.7062	0.7050
Criteria: correlation	G+TRI/BS (0.271)model H	H+TRI/UA (0.250)model I	I+GLO/AST (0.227)model J	J+WBC/UA (0.220)model K	K+WBC/PLA (0.208)model L	L+CRE/UA (0.206) model M	M+Hb/UA (0.203) model N
Fit function	1.1628	1.1038	1.1729	1.1378	1.0955	1.0828	1.0583
?^{2}	2981.2940	2830.0509	3007.3030	2917.2612	2808.8118	2776.2852	2713.4437
?^{2}/df	2981.2940/149	2830.0509/148	3007.3030/147	2917.2612/146	2808.8118/145	2776.2852/144	2713.4437/143
GFI	0.8931	0.8979	0.8927	0.8963	0.8987	0.9009	0.9028
AGFI	0.8493	0.8552	0.8467	0.8508	0.8534	0.8554	0.8573
NFI	0.7556	0.7680	0.7535	0.7609	0.7698	0.7724	0.7776
NNFI	0.6993	0.7133	0.6922	0.6997	0.7094	0.7108	0.7156
CFI	0.7642	0.7767	0.7618	0.7692	0.7782	0.7808	0.7860
PGFI	0.7004	0.6994	0.6906	0.6887	0.6859	0.6828	0.6795

Table 4.6 The model-fit indices for the cases of deleting correlated error terms in a **backward** manner considering **only one step**

Criteria: correlation	N- SBP/DBP (0.769)model I	N- AST/ALT (0.755)model II	N-Hb/RBC (0.604) III	N-BuN/CRE (0.584)VI	N-CHO/TRI (0.311) V	N-RBC/ALB (0.310)VI	N-Hb/ALB (0.271)VII
Fit function	NON-CONVERGE	1.0583	NON-CONVERGE	NON-CONVERGE	1.1460	1.1559	1.1285
?^{2}		2713.4318			2938.2705	2963.80033	2893.5676
?^{2}/df		2713.4318/144			2938.2705/144	2963.8033/144	2893.5676/144
GFI		0.9028			0.8974	0.8935	0.8973
AGFI		0.8583			0.8503	0.8446	0.8502
NFI		0.7776			0.7592	0.7571	0.7628
NNFI		0.7177			0.6930	0.6902	0.6979
CFI		0.7861			0.7673	0.7652	0.7711
PGFI		0.6843			0.6801	0.6771	0.6801
Criteria: correlation	N-TRI/BS (0.271) VIII	N-TRI/UA (0.250) IX	N-GLO/AST (0.227) X	N-WBC/UA (0.220) XI	N-WBC/PLA (0.208) XII	N-CRE/UA (0.206) XIII	N-Hb/UA (0.203) XIIII
Fit function	1.0599	1.0491	1.0589	1.0943	1.0982	1.0674	1.0828
?^{2}	2717.5263	2689.8768	2715.0217	2805.8833	2815.8276	2736.8329	2776.2852
?^{2}/df	2717.5263/144	2689.8768/144	2715.0217/144	2805.8833/144	2815.8276/144	2736.8329/144	2776.2852/144
GFI	0.9027	0.9036	0.9028	0.8984	0.9004	0.9005	0.9009
AGFI	0.8581	0.8595	0.8583	0.8519	0.8548	0.8549	0.8554
NFI	0.7772	0.7795	0.7775	0.7700	0.7692	0.7757	0.7724
NNFI	0.7173	0.7203	0.7175	0.7076	0.7065	0.7151	0.7108
CFI	0.7857	0.7880	0.7859	0.7784	0.7775	0.7841	0.7808
PGFI	0.6842	0.6849	0.6842	0.6809	0.6824	0.6825	0.6828

Criteria:	SBP/DBP, Hb/RBC,	N-TRI/UA,	N-TRI/UA,	N-TRI/UA,	N-AST/ALT,	N-AST/ALT,	N-TRI/BS
correlation	BuN/CRE	AST/ALT	TRI/BS	GLO/AST	TRI/BS	GLO/AST	GLO/AST
Fit function	1.3489	1.0491	1.0507	1.0496	1.0599	1.0589	1.0402
?^{2}	3458.5586	2689.8872	2693.9674	2691.1847	2717.5759	2714.9895	2667.0988
?^{2}/df	3458.5586/154	2689.8872/145	2693.9674/145	2691.1847/145	2717.5759/145	2714.9895/145	2667.0988/145
GFI	0.8799	0.9036	0.9035	0.9036	0.9027	0.9028	0.9049
AGFI	0.8362	0.8604	0.8602	0.8604	0.8591	0.8593	0.8623
NFI	0.7165	0.7795	0.7792	0.7794	0.7772	0.7775	0.7814
NNFI	0.6605	0.7223	0.7219	0.7222	0.7193	0.7196	0.7248
CFI	0.7248	0.7881	0.7878	0.7880	0.7858	0.7860	0.7900
PGFI	0.7132	0.6896	0.6895	0.6896	0.6889	0.6890	0.6906

Criteria:	N-TRI/UA	N-TRI/UA,	N-AST/ALT,	N- TRI/BS, AST/ALT,
correlation	AST/ALT, TRI/BS	AST/ALT, GLO/AST	TRI/BS, GLO/AST	GLO/AST, TRI/UA
Fit function	1.0507	1.0487	1.0605	1.0512
? ^{2}	2693.9418	2688.9534	2719.1212	2695.2408
? ^{2}/df	2693.9418/146	2688.9534/146	2719.1212/146	2695.2407/147
GFI	0.9035	0.9036	0.9027	0.9035
AGFI	0.8612	0.8613	0.8600	0.8622
NFI	0.7792	0.7796	0.7771	0.7791
NNFI	0.7239	0.7244	0.7212	.0.7258
CFI	0.7878	0.7883	0.7857	0.7878
PGFI	0.6943	0.6943	0.6937	0.6990

Fig. 4.1 The factor-to-factor position of each variable (y)

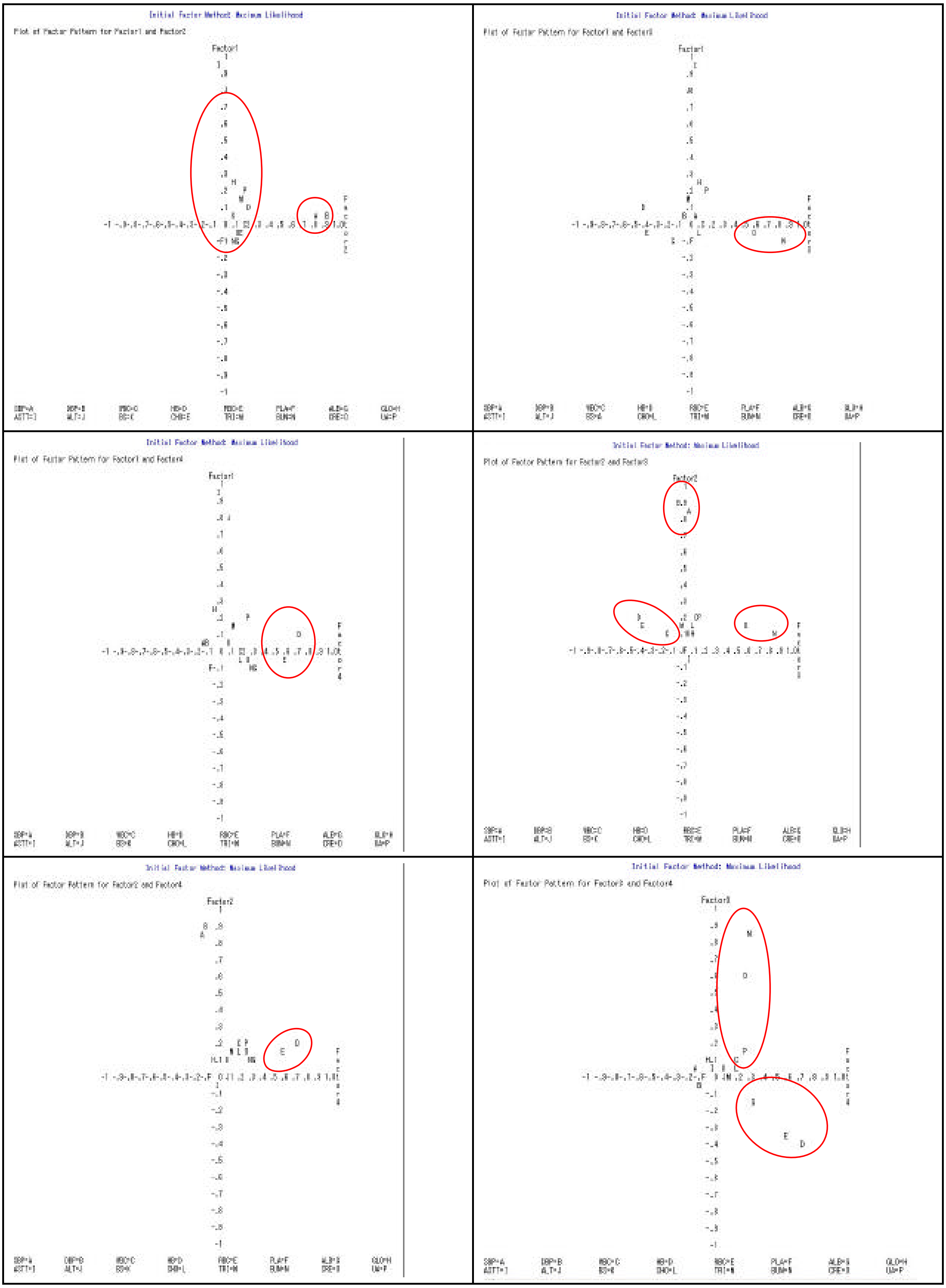
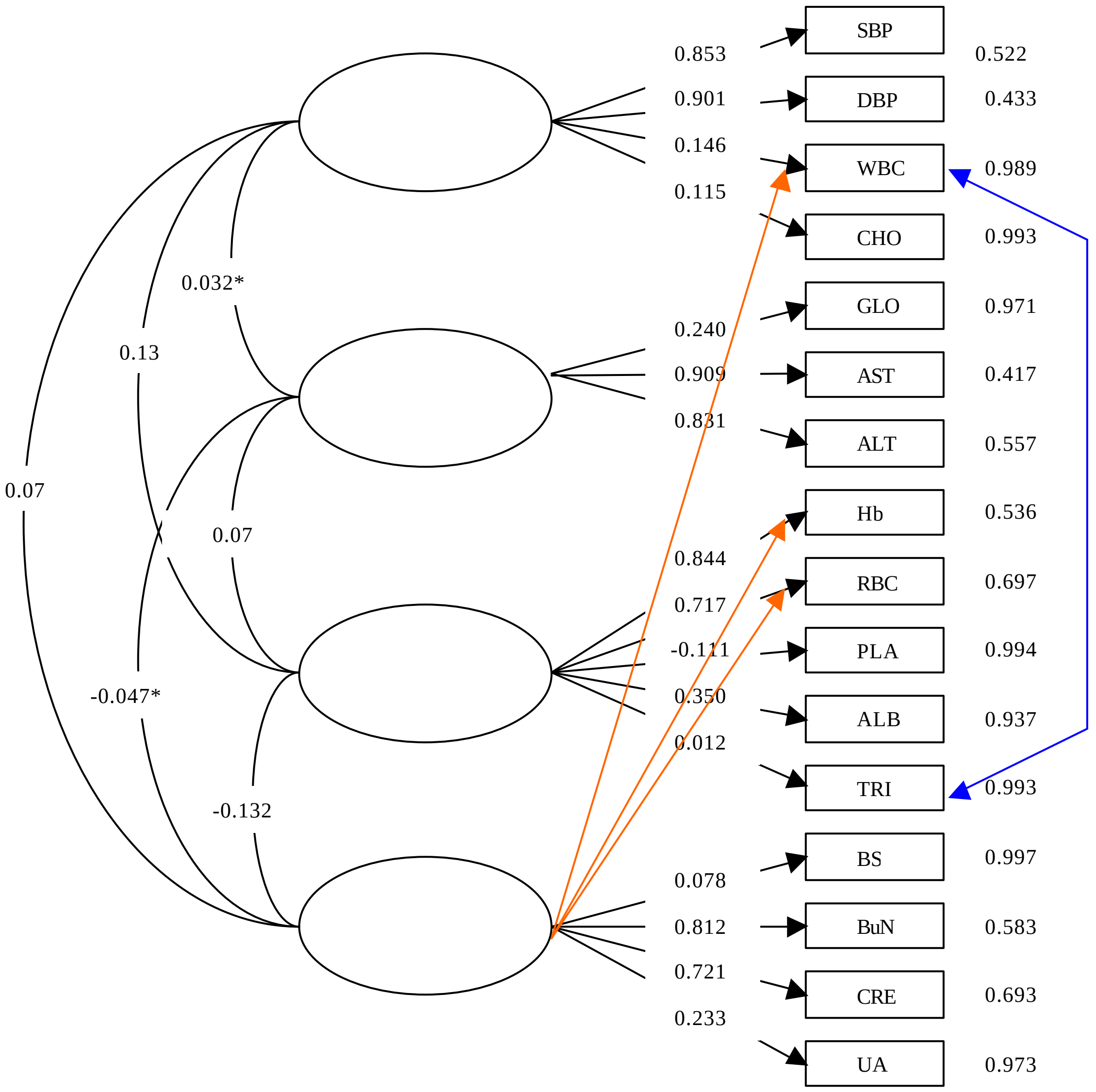
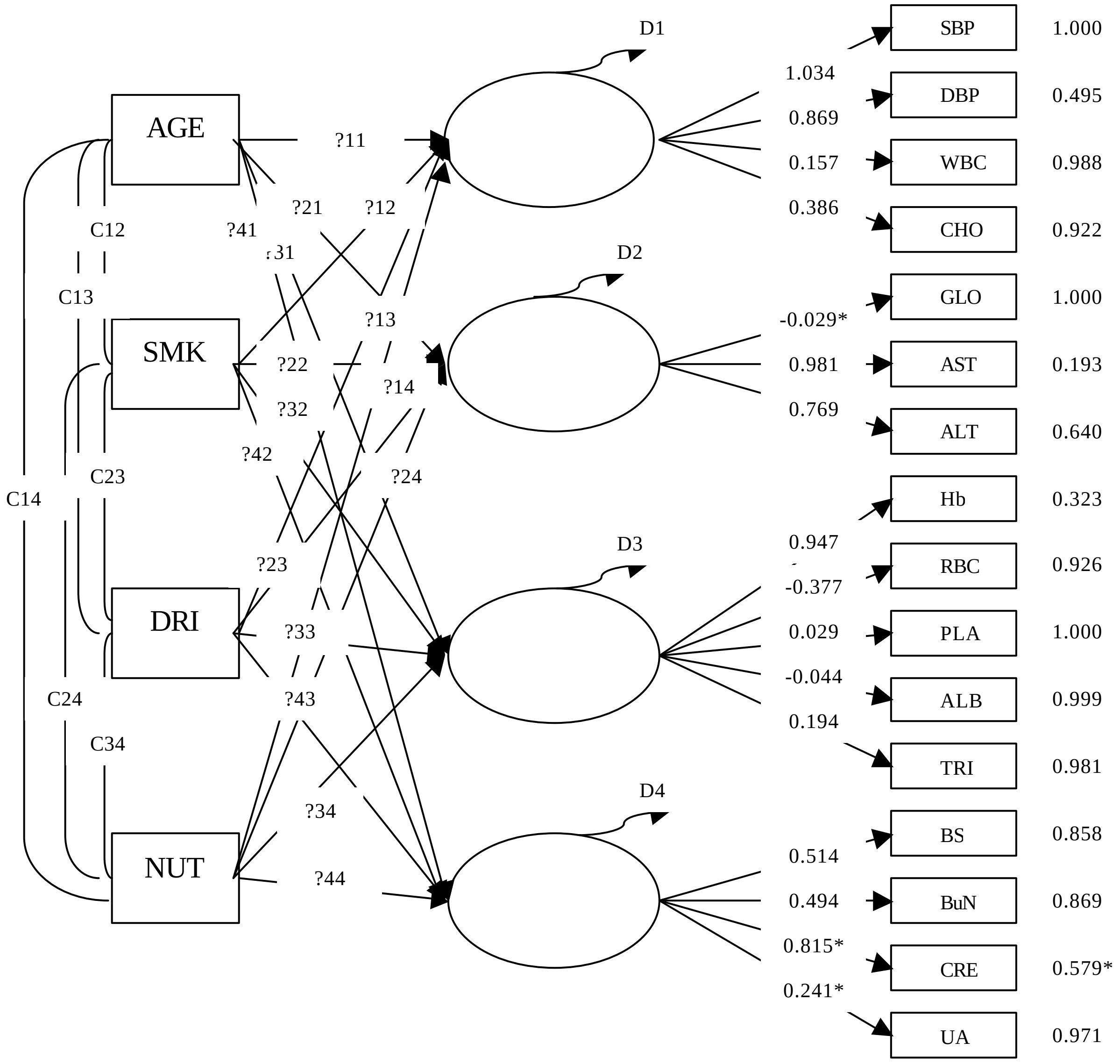


Fig. 4.2 The measurement model obtained from EFA. The inter-factor paths (orange arrows) are not included in later analyses. The attached table gives model-fit indices for measurement model based on EFA of Figure 4.2 and Table 4.1



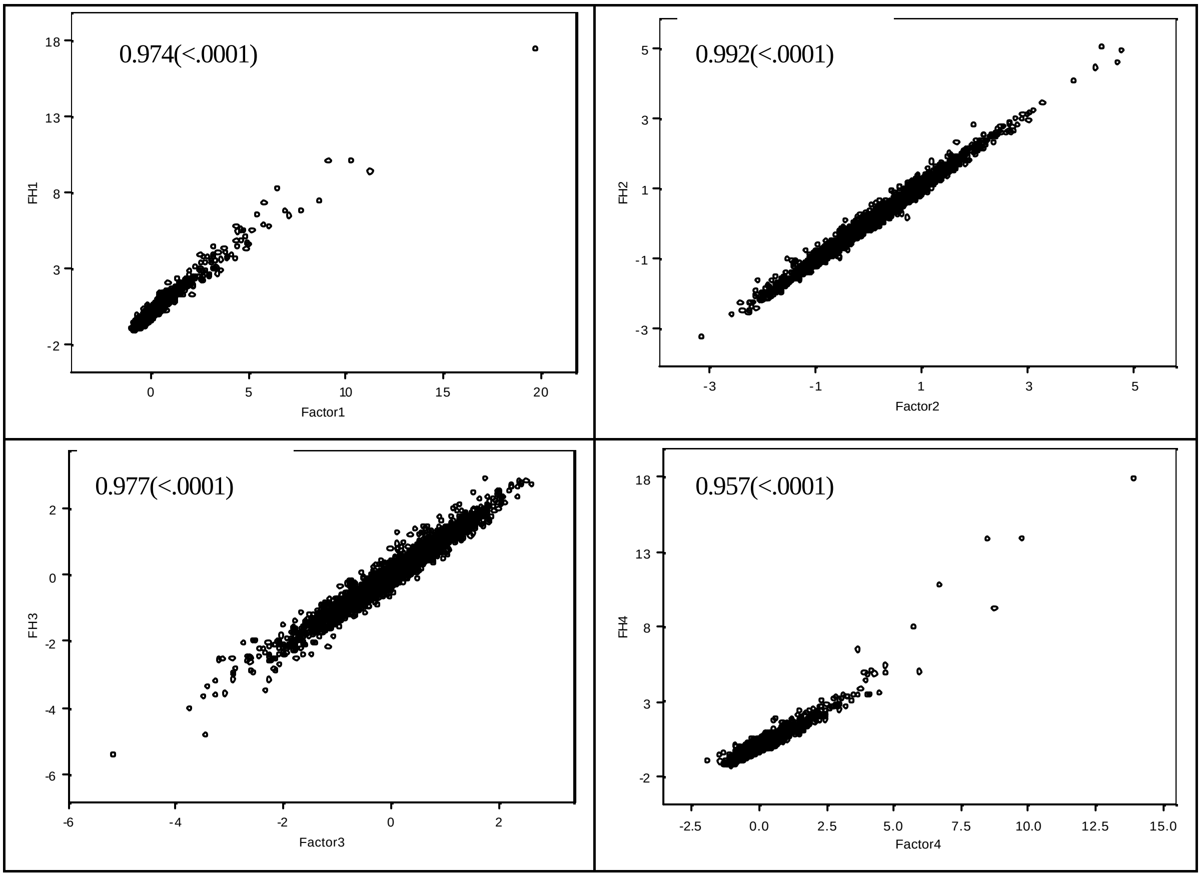
Fit function	0.8765
?^{2}	2247.4033
?^{2}/df	2247.4033/98
GFI	0.8989 (**0.9066**)
AGFI	0.8597 (**0.8690**)
NFI	0.7678
NNFI	0.7247
CFI	0.7751
PGFI	0.7341

Fig. 4.3 The full structural model estimates and model-fit indices for the full structural model in SEM without correlations among observed (y) variables



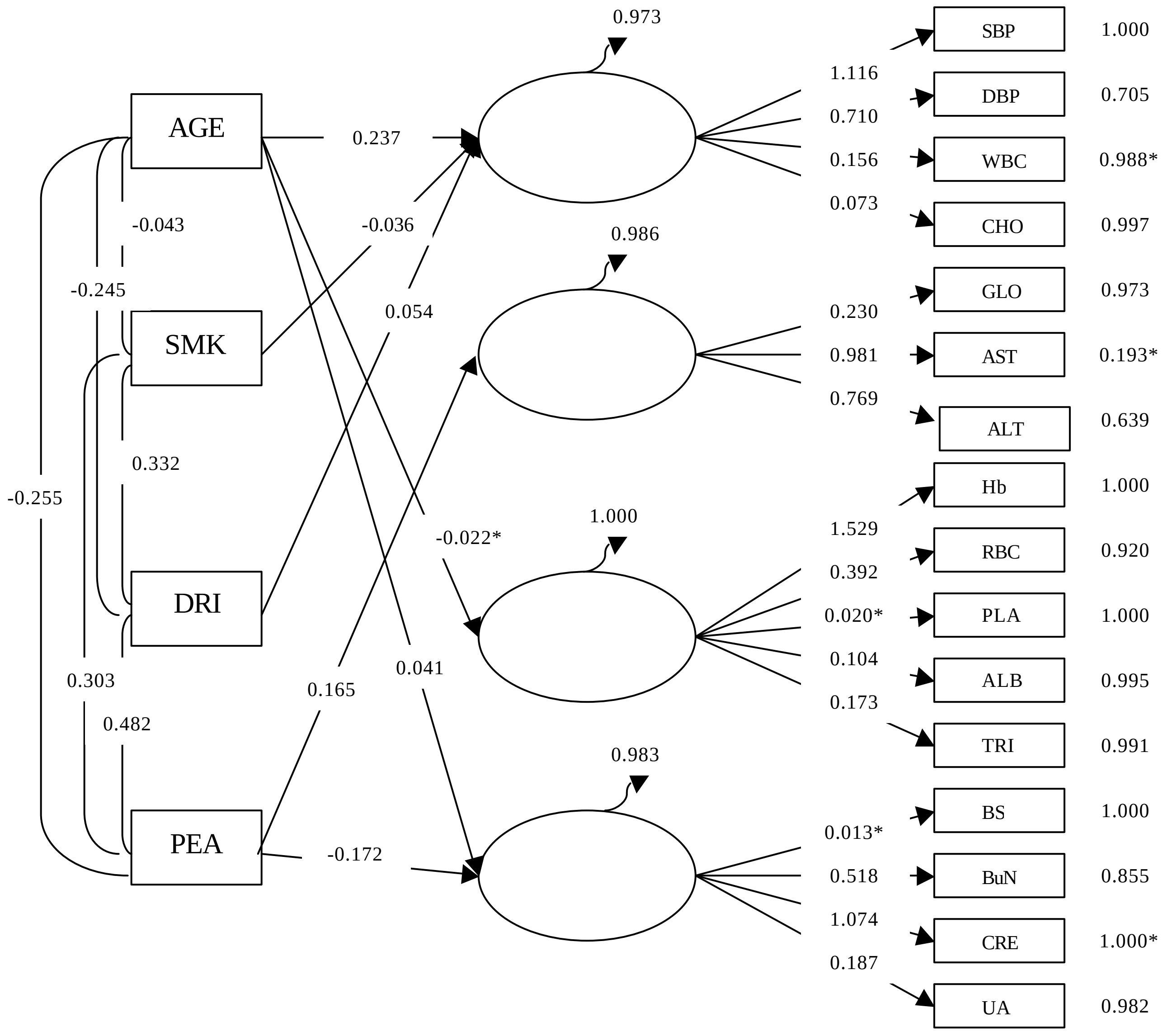
Fit function	4.5432	$^{?11}$ = -0.218	^{D2} = 0.980*	$^{?44}$= -0.044
$?^2$	11648.8043	$^{?12}$ = -0.055	$^{?31}$ = -0.052*	^{D4} = 0.991
$?^2$/df	11648.8043/148	$^{?13}$ = 0.057	$^{?32}$ = 0.151*	^{C12}= -0.043
GFI	0.6808	$^{?14}$ = 0.026	$^{?33}$ = 0.403	^{C13}= -0.245
AGFI	0.5470	^{D1} = 0.969	$^{?34}$ = -0.289*	^{C14}= -0.255
NFI	0.0452	$^{?21}$= -0.014*	^{D3} = 0.908	^{C23}= 0.332
NNFI	-0.2294	$^{?22}$= -0.024*	$^{?41}$= -0.057*	^{C24}= 0.303
CFI	0.0424	$^{?23}$= 0.028*	$^{?42}$= 0.101	^{C34}= 0.482
PGFI	0.5303	$^{?24}$= 0.185*	$^{?43}$= -0.110	

Figure 4.4 The comparison of the standardized factor scores in factor analysis with naï ve Bartlett factor scores and the factor scores used by SAS System.



The x-axis shows Bartlett’s score and y-axis shows score of SAS .

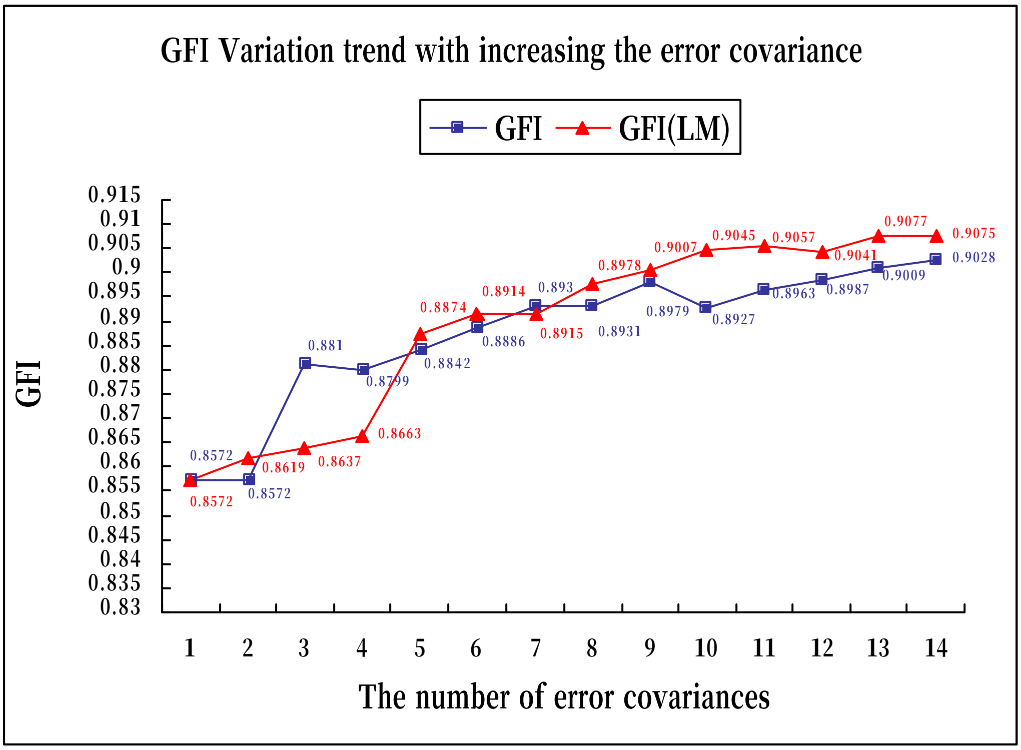
Fig. 4.5 The SEM obtained from a combination of the constructs corresponding with criteria in Table 4.4 (without considering error-covariance of y-variables.)



Fit function	1.6573
?^{2}	4249.2828
?^{2}/df	4249.2828/157
GFI	0.8445
AGFI	0.7920
NFI	0.6517
NNFI	0.5876
CFI	0.6593
PGFI	0.6978

Fig.4.6 and **Fig. 4.7**

A comparison of the trend in GFI (the upper panel) and AGFI (the bottom panel) at the step of adding covariance path in a **forward** manner based on marginal correlations (**blue**) and Lagrange Multiplier test (**red**)



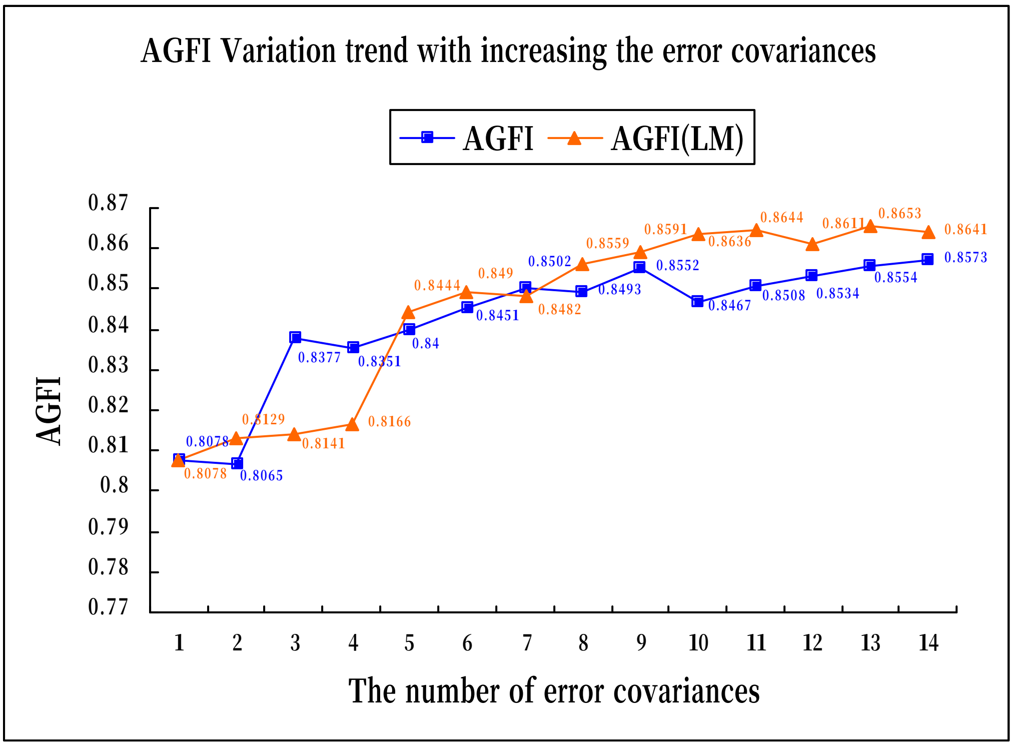


Fig. 4.8 The model fit index, GFI, of deleting a covariance path from a ‘ final’ Model (Model N) with error covariance paths in Figure 4.6 with GFI=0.9028

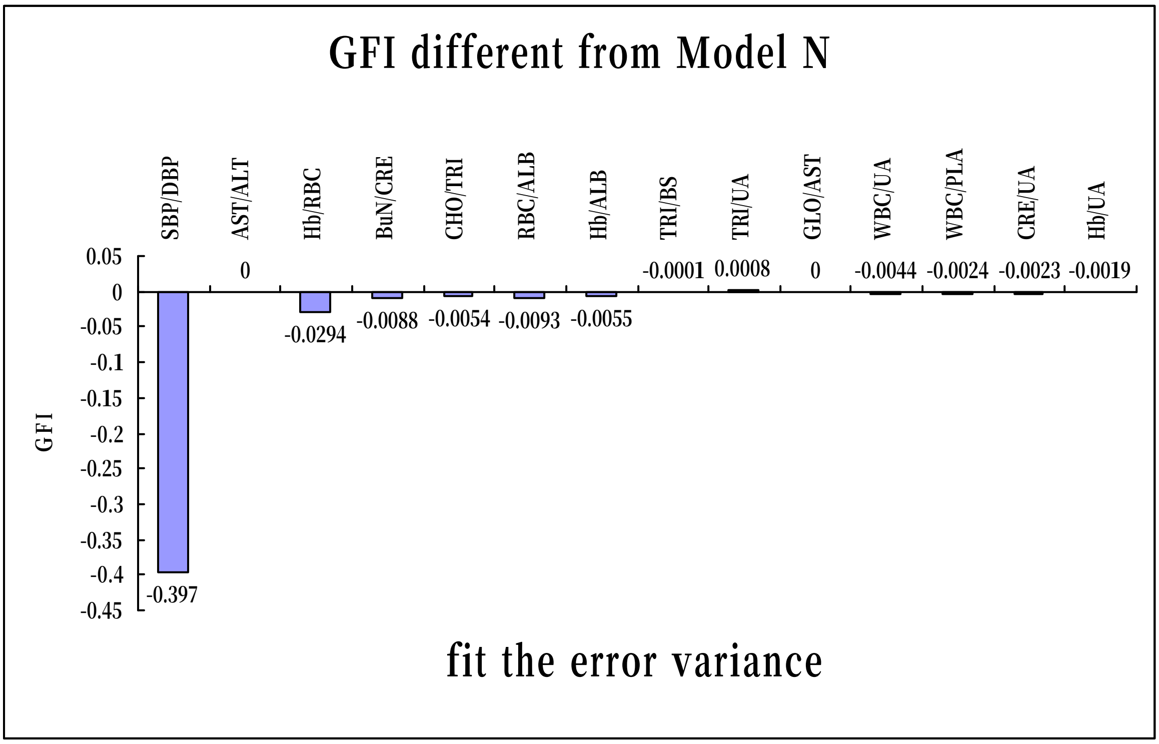
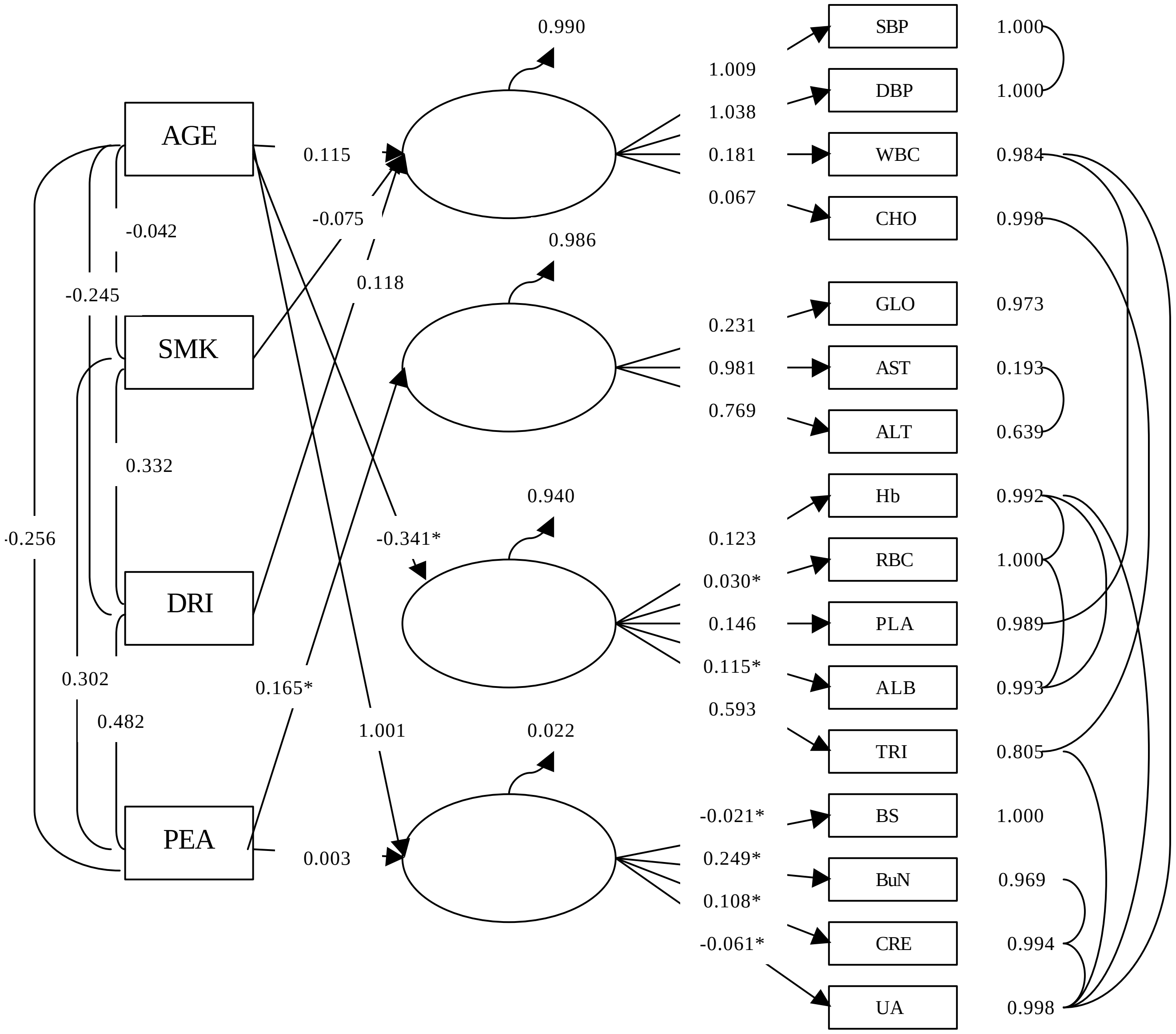
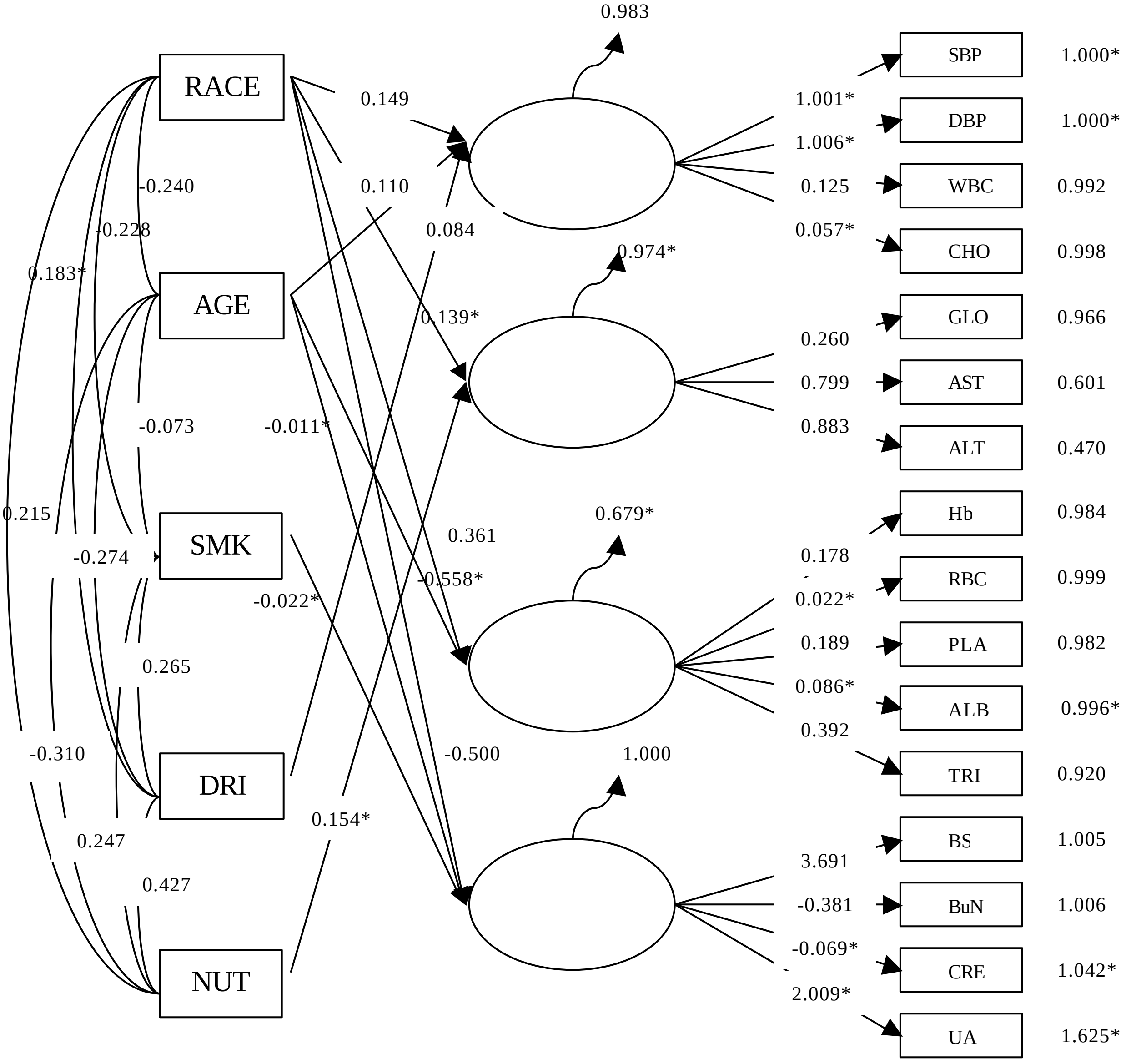


Fig. 4.9 An ‘final’ construct of SEM with covariance paths



Fit function	1.6573	Add correlated errors：	1.0402
?^{2}	4249.2828	SBP/DBP(0.000), AST/ALT(-0.001),	2667.0988
?^{2}/df	4249.2828/157	Hb/RBC (0.602), WBC/PLA (0.196),	2667.0988/145
GFI	0.8445	Hb/ALB (0.263), RBC/ALB (0.308),	0.9049
AGFI	0.7920	CHO/TRI(0.287), BuN/CRE (0.566),	0.8623
NFI	0.6517	WBC/UA (0.179), Hb/UA (0.130),	0.7814
NNFI	0.5876	TRI/UA (0.047), CRE/UA (0.098) .	0.7248
CFI	0.6593		0.7900
PGFI	0.6978		0.6906

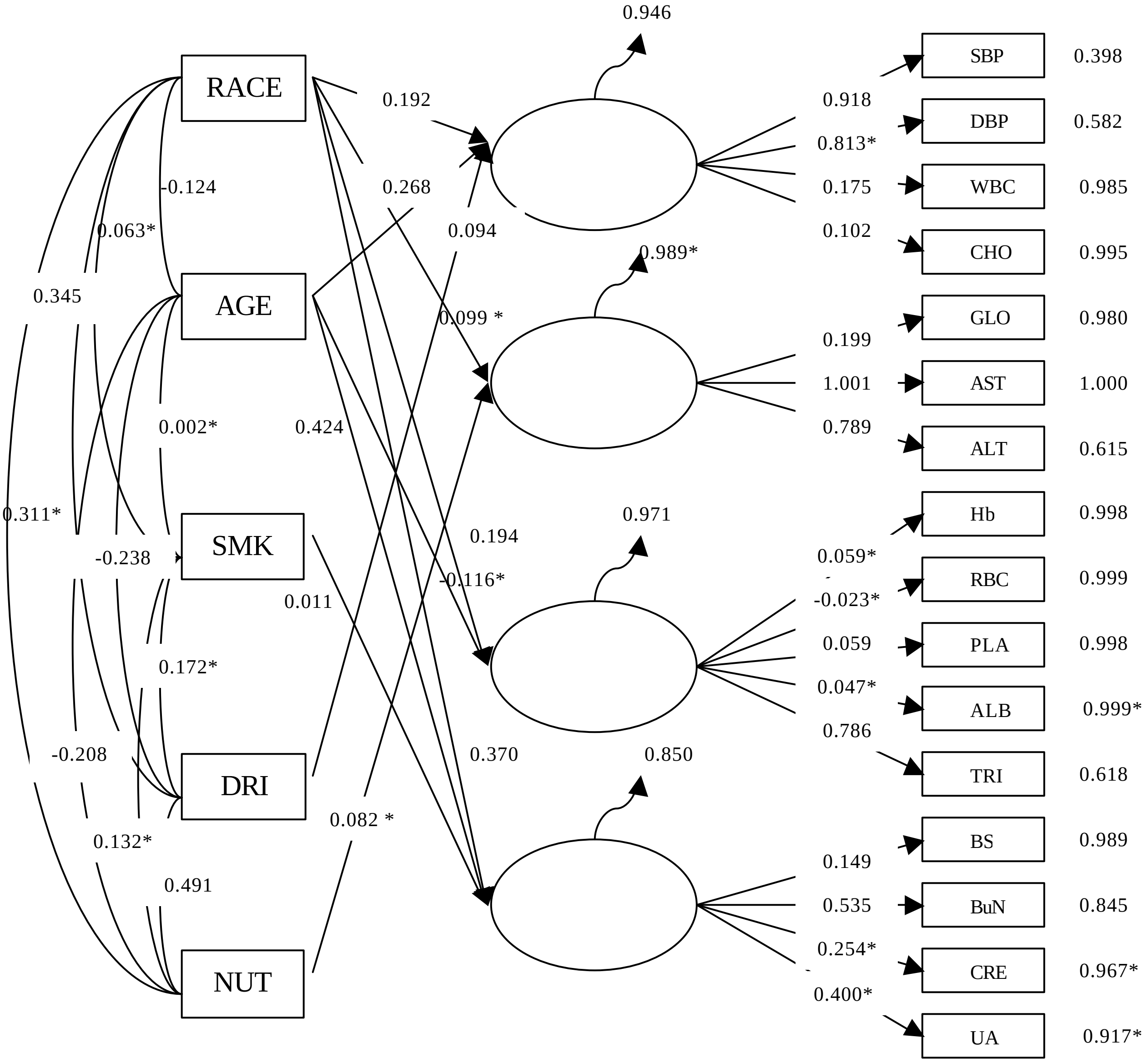
Fig. 4.10. An ‘final’ construct of SEM for **male** with race as an exogenous variable but without considering covariance paths



Fit function	1.5905	Add correlated errors： SBP/DBP(0.000), Hb/RBC (0.571), WBC/PLA (0.189), Hb/ALB (0.316), RBC/ALB (0.360*), CHO/TRI(0.303), BuN/CRE (0.534), WBC/UA (0.130), Hb/UA (0.040), CRE/UA (-0.131*) .GLO/AST (0.051)
?^{2}	1948.4221	
?^{2}/df	1948.4221/159	
GFI	0.8789	
AGFI	0.8240	
NFI	0.6816	
NNFI	0.6001	

CFI	0.6972	
PGFI	0.6654	

Fig. 4.11. An ‘final’ construct of SEM for **female** with race as an exogenous variable but without considering covariance paths



Fit function	1.1272	Add correlated errors：
?^{2}	1508.1554	SBP/DBP(0.040),
?^{2}/df	1508.1554/159	Hb/RBC (0.516), WBC/PLA (0.215),
GFI	0.9052	Hb/ALB (0.249*), RBC/ALB (0.277),
AGFI	0.8623	CHO/TRI(0.301), BuN/CRE (0.619),
NFI	0.7821	WBC/UA (0.117), Hb/UA (0.103),

NNFI	0.7345	CRE/UA (0.082) .GLO/AST (0.000)
CFI	0.7990	
PGFI	0.6854	

Parameters	All	Male	Female	Parameters	All	Male	Female
?^{11}	1.006	1.001*	0.918	E1	1.000	1.000*	0.398
?^{21}	1.037	1.006*	0.813*	E2	1.000	1.000*	0.582
?^{31}	0.091	0.125	0.175	E3	0.996	0.992	0.985
?^{42}	0.071*	0.057*	0.102	E4	0.998	0.998	0.995
?^{52}	0.249	0.260	0.199	E5	0.968	0.966	0.980
?^{62}	0.981	0.799	1.002	E6	0.193	0.601	1.000
?^{73}	0.770	0.883	0.789	E7	0.638	0.470	0.615
?^{83}	0.092	0.178	0.059*	E8	0.996	0.984	0.998
?^{93}	-0.049*	0.022*	-0.023*	E9	0.999	0.999	0.999
?^{103}	0.126	0.189	0.059	E10	0.992	0.982	0.998
?^{113}	0.054*	0.086*	0.047*	E11	0.999	0.996	0.999*
?^{123}	0.604	0.392	0.786	E12	0.797	0.920	0.618
?^{134}	0.125	3.691	0.149	E13	0.992	1.005	0.989
?^{144}	0.369	-0.381	0.535	E14	0.929	1.006	0.845
?^{154}	0.158*	-0.069*	0.254*	E15	0.987	1.042*	0.967*
?^{164}	0.335*	2.009*	0.400*	E16	0.942	1.625*	0.917*
?^{11}	0.129	0.149	0.192	D1	0.986	0.983	0.946
?^{12}	0.124	0.110	0.268				
?^{14}	0.046	0.084	0.094				
?^{21}	0.102*	0.139*	0.099*	D2	0.981*	0.974*	0.989*
?^{25}	0.142*	0.154*	0.082*				
?^{31}	0.280	0.361	0.194	D3	0.911	0.679*	0.971
?^{32}	-0.257*	-0.558*	-0.116				
?^{41}	0.541	0.500	0.370	D4	0.808	1.000	0.850
?^{42}	0.361	-0.011*	0.424				
?^{43}	0.045	-0.022*	0.011				
C12	-0.179	-0.240	-0.124	C24	-0.245	-0.274	-0.238
C13	-0.147	-0.228	0.063	C25	-0.255	-0.310	-0.208
C14	0.215	0.183	0.345	C34	-0.332	0.265	0.172*
C15	0.221	0.215	0.311*	C35	0.303	0.247	0.132*
C23	-0.043	-0.073	0.002*	C45	0.482	0.427	0.491
CE1E2	0.000	0.000	0.040	CE8E11	0.268	0.316	0.249*
CE3E10	0.196	0.189	0.215	CE8E16	0.121	0.040	0.103
CE3E16	0.188	0.130	0.117	CE9E11	0.313	0.360*	0.277
CE4E12	0.287	0.303	0.301	CE14E15	0.567	0.534	0.619
CE5E6	-0.096	0.051	0.000	CE15E16	0.126	-0.131*	0.082
CE8E9	0.606	0.571	0.516				

Chapter 5 Discussion

Since our data came from a cross-sectional (which may be a biased) survey, it is difficult to check the causal-effect relationship among observed/latent variables. In a population-based study, this may be due to systematic errors and selection biases of the sample. In summary, sample fluctuations may exist and it is not possible to release it. Nevertheless, to embed the analysis into a framework of follow-up study is our forthcoming effort. This research offers a chance to explore a model with prediction ability for disease development, and serves as a statistical tool for screening programs of multiple chronic diseases with considerations on genetic/familial factors.

Our goal of this thesis

Since an EFA is used in place of the CFA for a construction of the measurement model, we seek to offer a **hybrid algorithm** for a cross-sectional dataset without resorts to a confirmatory structure of the observed endogenous variable. There may be some **drawbacks** in the model building process. It relies too much on the statistical tool of exploratory factor analysis and thus, sometimes, it is difficult to address the mechanism with physiological feasibility. On the other hand, our study renders a simple and easy treatment of how to build an acceptable model, in terms of the goodness-of-fit indices.

Constrained vs. unconstrained estimates. (SAS PROC CALIS vs. LISREL or EQS)

The LISREL software offers constrained estimates for the measurement model and the entire structural equation model (SEM). When improper solutions are encountered, we followed the guidelines of Chapter 2 to solve it. With the present dataset, the factor loading of SBP (with respect to Factor 1) is improper in any case. We have two ways to deal with this problem. First, the error terms may be set to zero. By doing this, since the first factor loading of each factor is reasonably set to be *one* in LISREL estimation, it means that in this case Factor 1 is recognized as being totally equal to SBP. This is an unavoidable identification when there are improper solutions appeared in the estimated model. It also reveals that Factor 1 needs more amendment. Second, we considered a possibility to delete the SBP variable and re-estimate the model. The result is coherent in other factors except for Factor 1 in which only three variables are retained. (See Figures 5.1 and 5.2.)

[Put Figures 5.1 and 5.2 here.]

Fig. 5.1 To deal with improper solutions: the error terms set to zero

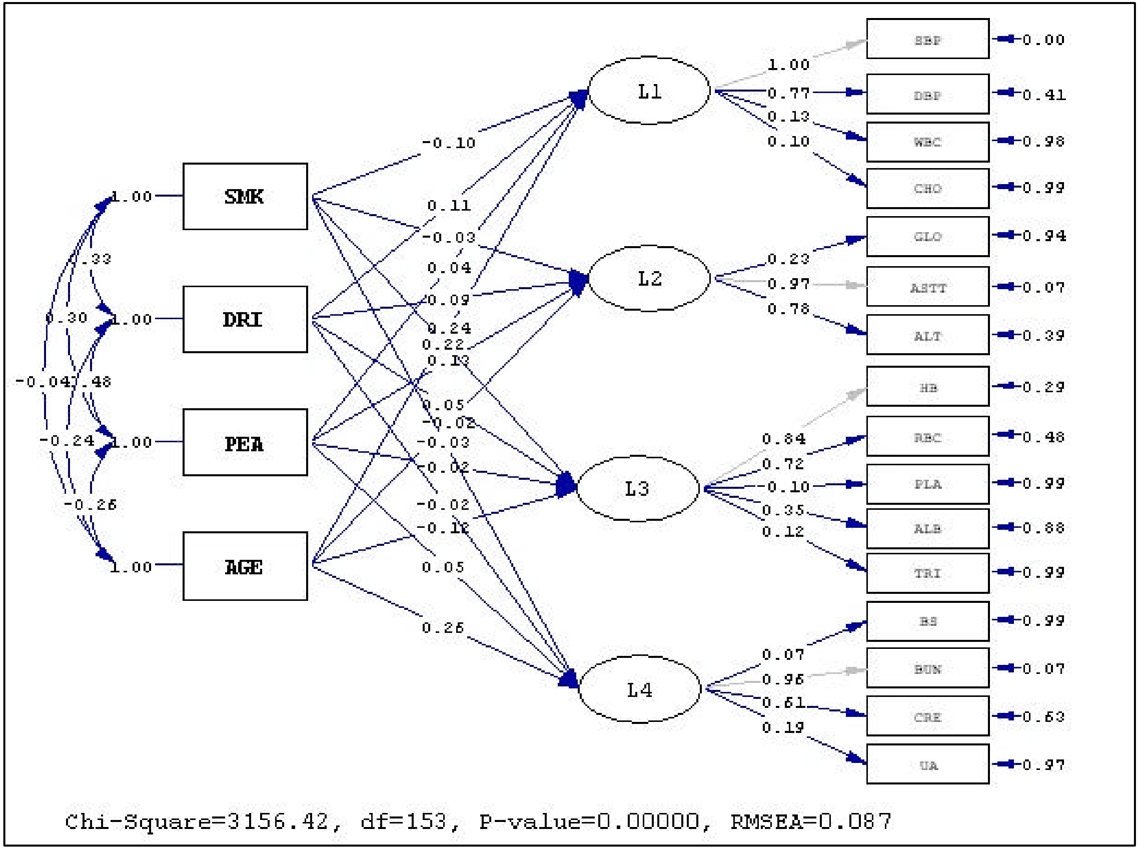
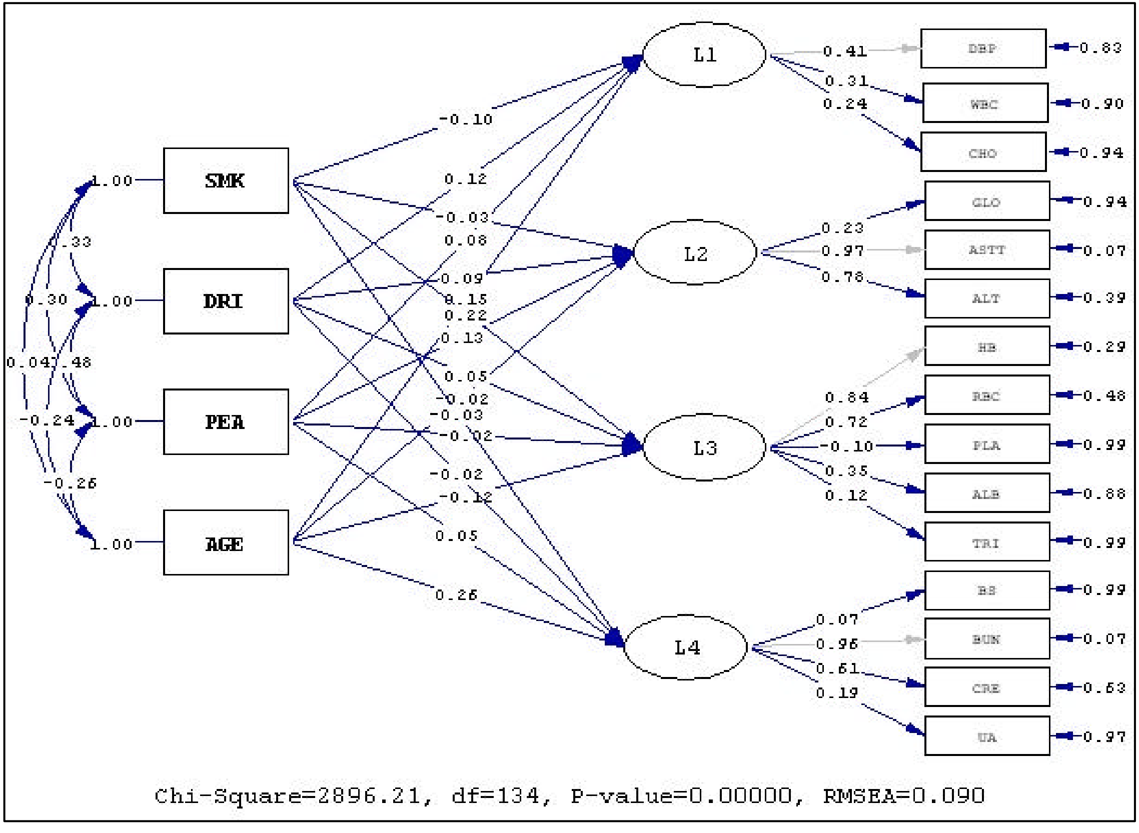


Fig. 5.2 To deal with improper solutions: to delete the SBP variable



Reference

1. C. M. Musil, S. L. Jones, C. D. Warner. (1998) Structural equation modeling and its relationship to multiple regression and factor analysis. Research in Nursing & Health, Vol.21 p271-281.
2. D. F. Morrison. (1976) Multivariate Statistical methods 2nd. New York: McGraw-Hill p334-336
3. F. Chen, K. A. Bollen, P. Paxton, P. J. Curran, J. B. Kirby. (2001) Improper Solutions in structural equation models: causes, consequences, and strategies. Sociological Methods & Research, Vol.29 No.4, p468-508.
4. J. M. Batista-Foguet, G. Coenders & M. A. Ferragud. (2001) Using structural equation models to evaluate the magnitude of measurement error in blood pressure. Statistics in Medicine.Vol.20 p2351-2368. John Wiley & Sons, Ltd.
5. J. T. Williams, P. Van Eerdewe gh, L. Almasy, & J. Blangero. (1999a) Joint multipoint linkage analysis of multivariate qualitative and quantitative traits. I. Likelihood formulation and simulation results. American Journal of Human Genetics. Vol. 65,p1134-1147.
6. J. T. Williams, H. Begleiter, B. Porjesz, H. J. Edenberg, T. Foroud, T. Reich, A. Goate, P. Van Eerdewegh, L. Almasy, & J. Blangero. (1999b) Joint multipoint linkage analysis of multivariate qualitative and quantitative traits. II. Alcoholism and event-related potentials. American Journal of Human Genetics. Vol. 65, p1148-1160.
7. K. A. Bollen. (1989). Structural Equations with Latent Variables. Wiley, New York.
8. K. A. Bollen. (1998). P. Aruitage & T. Cotton (editors in chief).Structural equation models. Encyclopedia of Biostatistics. Sussex, England. John Wiley. p4363-4372.
9. K-H Yuan, W. Chan, P. M. Bentler. (2000) Robust transformation with applications to structural equation modeling. British Journal of Mathematical and Statistical psychology. Vol. 53, p31-50.
10. K-H Yuan, P. M. Bentler. (1998) Normal theory based test statistics in structural equation modeling. British Journal of Mathematical and Statistical psychology. Vol. 51, p289-309.
11. K. Jöreskog, & D. Sörbom. (1993). LISREL 8: Structural equation modeling with the SIMPLIS command language. Chicago: Scientific Software International.
12. L. Hatcher (1998) A step-by-step approach to using the SAS system for

factor analysis and structural equation modeling 3th, Cary, NC: SAS Institute Inc.

13. L-H Lai. (2001) Association between blood lead levels and hyperuricemia in Hsin-Yi township, Nautou county (in Chinese). Institute of Environmental Health, China Medical College. (Master's Thesis:IEH-1212)
14. M. A. Province. (2001) D. C. Rao (editor) Linkage and association with structural relationships. Genetic Dissection of Complex Traits. Academic Press. P183-190.
15. P. M. Bentler & J. A. Stein.(1992) Structural equation models in medical research. Statistical Methods in Medical Research. Vol.1 No.2, p159-181
16. R. O. Mueller. (1996). Basic Principles of Structural Equation Modeling: An Introduction to LISREL and EQS. Springer-Verlag New York, Inc.
17. T. H. Wan, Ph.D.(2002) Evidence-Based Health Care Management: Multivariate Modeling Approaches. Kluwer Academic Publishers.
18. Z. Pausova, F. Gossard, D. Gaudet, J. Tremblay, T. A. Kotchen, A. W. Cowley, P. Hamet. (2001) Hritability estimates of obesity measures in siblings with and without Hypertension. Hypertension. Vol. 38, p41-47.
19. **謝博生著**(1982)**：臨床數據-判讀與運用**(**上下冊**)**。醫文社印行。**
20. H-J Chiou(2000)Quantitative research and statistical analysis in social and behavioral sciences p15-4~15-12.(in Chinese)

    邱皓政著(2000)**：社會與行為科學的量化研究與統計分析：**SPSS **中文視窗版資料分析範例與解析**