FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

MASTER THESIS

Václav Veselý

Structural Equation Models with Application in Social Sciences

Department of Probability and Mathematical Statistics

Supervisor:	RNDr. Michal Pešta, Ph.D.
Study programme:	Mathematics
Study branch:	Probability, Mathematical Statistics and Econometrics

Prague 2018

I declare that I carried out this master thesis independently, and only with the cited sources, literature and other professional sources.

I understand that my work relates to the rights and obligations under the Act No. 121/2000 Sb., the Copyright Act, as amended, in particular the fact that the Charles University has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 subsection 1 of the Copyright Act.

In Prague on May 8, 2018 signature of the author

Title: Structural Equation Models with Application in Social Sciences

Author: Václav Veselý

Department: Department of Probability and Mathematical Statistics

Supervisor: RNDr. Michal Pešta, Ph.D., Department of Probability and Mathematical Statistics

Abstract: We investigate possible usage of Errors-in-Variables estimator (EIV), when estimating structural equations models (SEM). Structural equations modelling provides framework for analysing complex relations among set of random variables where for example the response variable in one equation plays role of the predictor in another equation. First an overview of SEM and some common covariance based estimators is provided. Special case of linear regression model is investigated, showing that the covariance based estimators yield the same results as ordinary least squares.

A compact review of EIV models follows, Errors-in-Variables models are regression models where not only response but also predictors are assumed to be measured with an error. Main contribution of this paper then lies in defining modifications of the EIV estimator to fit in the SEM framework. General optimization problem to estimate the parameters of structural equations model with errors-in-variables si postulated. Several modifications of two stage least squares are also proposed for future research.

Equation-wise Errors-in-Variables estimator is proposed to estimate the coefficients of structural equations model. The coefficients of every structural equation are estimated separately using EIV estimator. Some theoretical conditions are proposed under which this method yields consistent estimates of the parameters. However the practical applications seem to be fairly limited due to difficulties with meeting the previously mentioned conditions. Psychological study of sexual and relationship satisfaction of women in young adulthood is used as an illustrative example.

Keywords: structural equation modelling, simultaneous equation models, SEM, LISREL, application in social sciences, Errors-in-Variables, total least squares

First motivation for the topic of this thesis origins from a practical problem that I have encountered thanks to PhDr. Eliška Kovaříková. That is only one among many reasons why she deserves my thanks. When I had just a foggy idea about the topic of my future thesis I was directed towards RNDr. Michal Pešta, Ph.D. who helped me by specifying the topic more precisely. For that and for the motivation and all the consultations that he provided during my work on this paper he also deserves my thanks.

Notation

$\overline{x}$	...	$\frac{1}{n}\sum x_i$
$\|A\|_F$	...	Frobenius norm of the matrix A
$\mathcal{M}(A)$	...	Column space of the matrix A
$\lvert A\rvert$	...	Determinant of the matrix A
$X \perp\!\!\!\perp Y$	...	X is independent of Y
I	...	Identity matrix of conformable dimension
$a.s.$	...	Almost surely
s.d.	...	Standard deviation
1st Qu.	...	First quartile

Contents

Notation **iv**

Introduction **3**

1 Structural equitation model **4**
1.1 Motivation . . . 4
1.2 General Structural Equation Model . . . 5
1.3 Model parameters . . . 6
1.3.1 Implied covariance matrix – Definition . . . 6
1.3.2 Special cases . . . 6
1.4 Parameter estimation . . . 8
1.4.1 Introduction . . . 8
1.4.2 Covariance based estimators . . . 8
1.4.3 Two stage least squares (2SLS) . . . 9
1.4.4 Other methods . . . 10
1.5 Comparison to OLS regression . . . 11

2 Error-in-Variables Model, Total least squares **12**
2.1 Motivation . . . 12
2.2 Partial Errors-in-Variables Model . . . 12
2.2.1 Definition . . . 12
2.2.2 Example . . . 14
2.3 Properties of EIV estimator . . . 15
2.4 Euler–Lagrange solution to constrained EIV problem . . . 17

3 Errors-in-Variables in context of structural equations model **18**
3.1 Motivation . . . 18
3.2 Problems specification . . . 18
3.3 Modified estimation methods . . . 19
3.3.1 Constrained total least squares . . . 19
3.3.2 Equation-wise Error-in-Variable estimation . . . 20
3.3.3 Modification of two stage least squares . . . 23

4 Application **25**
4.1 Introduction . . . 25
4.2 Methods . . . 25
4.3 Results . . . 26
4.4 Application conclusion . . . 32

Conclusion **33**

Bibliography **34**

List of Figures **36**

List of Tables **37**

A Appendix **38**
A.1 Additional theorems and definitions 38
A.2 Code used for computing PEIV estimator 39

Introduction

Linear regression is very useful tool when investigating association of one response variable Y_1 (for example sexual satisfaction in relationship) and a set of regressors $\mathbf{X}$ (for example psychological characteristic, big five personality traits). Imagine that we want to investigate such relation for more then one response variable (also investigate Y_2 relationship satisfaction). We could build two regression models separately but that could lead to loss of possibly important information about association of Y_1 and Y_2.

One maybe more suitable approach to such problem is via structural equitation modelling (SEM). Using SEM one can investigate relation between Y_1 and Y_2 trough structural coefficients while taking in account other variables that influence Y_1 or Y_2. In the first chapter introduction to structural equation modelling is provided. General form of LISREL model is postulated and some special cases are derived. So called observed variable model is the model on which is the main focus of this paper. Then a brief overview of classical covariance based estimation techniques follows and most commonly used two stage least squares (2SLS) is also described. At the end of the first chapter similarity between ordinary least squares (OSL) and covariance based estimators for linear regression model is described.

Basic theory of Error-in-Variables model is described in the second chapter. Error-in-variables model is linear regression model where covariates are assumed to be measured with an error. That might be practically well applicable when covariates are results of sociological or psychological questionnaire. We describe Error-in-variable estimator (EIV) that is strongly consistent under some conditions on the errors.

Finally the third chapter contains main contribution of this work. First we discuss that direct usage of EIV for some special cases of structural equations models might be sensible choice. Then we propose more complex modifications of error-in-variable estimator to fit structural equation model.

General optimization problem for estimating observed variable model with errors in variables is postulated. Lagrange function is defined but solution to this problem is yet to be investigated.

Then equation-wise Error-in-Variables estimator is proposed. This method estimates coefficients in every structural equation separately using EIV estimator. Some theoretical conditions for this method to yield consistent estimator are postulated. Practical difficulties with meeting this conditions are discussed.

In the last part several modifications of two stage least squares are proposed. Possible simulation study is proposed in order to distinguish those methods worth future attention.

The last chapter provides a numerical illustration of equation-wise Error-in-Variables applied to practical example from psychology. Models proposed in the original study of sexual and relationship satisfaction of women are reviewed and compared to the results of newly proposed technique of equation-wise Error-in-Variables.

1. Structural equitation model

1.1 Motivation

Simple linear regression problem is widely used example when introducing structural equitation modelling (SEM). Consider following model:

$$y = \alpha + \gamma\, x + \zeta. \tag{1.1}$$

Where random variable y is the response, random variable x is the regressor, $(\alpha, \gamma) = \Gamma$ are the regression coefficients and ζ is the random error uncorrelated with x. Assume that we have (x_i, y_i), $i = 1, \ldots, N$ independent observations of (x, y). We can estimate Γ using ordinary least squares (OLS), which means we minimize $S(\alpha, \gamma) = \sum_{i=1}^{N}(y_i - \alpha + \gamma\, x_i)^2$ with respect to α and γ. Leading to estimates (see [1] section 5.3):

$$\hat{\gamma} = \frac{\sum x_i y_i - \frac{1}{N}\sum x_i \sum y_i}{\sum x_i^2 - \frac{1}{N}(\sum x_i)^2} = \frac{\widehat{\operatorname{cov}(x,y)}}{\widehat{\operatorname{var}(x)}}, \qquad \hat{\alpha} = \overline{y} - \hat{\gamma}\,\overline{x}\;. \tag{1.2}$$

Another approach is to investigate correlation structure of observed variables y and x. Let σ_{yy} (resp. σ_{xx}) stand for variance of y (resp. x), denote covariance $\operatorname{cov}(x, y) = \sigma_{xy}$ and $\operatorname{var}(\zeta) = \psi$. Then we rewrite theoretical covariance matrix as follows,

$$\operatorname{Cov}\begin{pmatrix} y \\ x \end{pmatrix} = \begin{pmatrix} \sigma_{yy} & \sigma_{yx} \\ \sigma_{xy} & \sigma_{xx} \end{pmatrix} = \begin{pmatrix} \gamma^2\sigma_{xx} + \psi & \gamma\sigma_{xx} \\ \gamma\sigma_{xx} & \sigma_{xx} \end{pmatrix} = \Sigma\left(\gamma, \sigma_{xx}, \psi\right). \tag{1.3}$$

The idea of covariance based structural equations model (SEM) is to analyse how close is such covariance matrix implied by the model to the empirical covariance matrix $\mathcal{S}$. In other words, to estimate parameters $\gamma, \sigma_{xx}, \psi$ we find

$$\min F\Big(\Sigma\big(\underbrace{\gamma, \sigma_{xx}, \psi}_{\theta}\big), \mathcal{S}\Big), \tag{1.4}$$

where F is some suitable penalising function. One possible choice of the penalising function is $F(\Sigma(\theta), \mathcal{S}) = \frac{1}{2}\left(\|\Sigma(\theta) - \mathcal{S}\|_F\right)^2 = \frac{1}{2}\operatorname{tr}\left\{(\Sigma(\theta) - \mathcal{S})(\Sigma(\theta) - \mathcal{S})^\top\right\}$ which is commonly referred as Unweighted Least Squares (ULS). For details see section 1.4. In this case of simple linear regression both methods leads to the same estimate $\hat{\gamma}$ which is shown at the Section 1.5.

Although some estimator of coefficients of structural equations models rely on different rationale the idea of the second approach is easily generalizable for more complex set of equations including more general relations among variables which are the focus of this chapter. Response variable in some equation can play role of the regressor in another equation. One can also include latent unobserved variables etc.

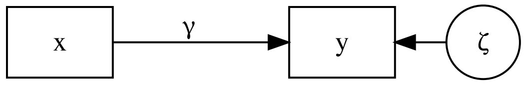

Figure 1.1: Path diagram of the model (1.1).

1.2 General Structural Equation Model

I this section we define General Structural Equation Model (SEM), also known as LISREL (Linear Structural Relations). In the generic form the LISREL model is set of random variables (bold symbols, see table 1.1) and linear relations among them consisting of **latent variable model**:

$$\boldsymbol{\eta} = B\boldsymbol{\eta} + \Gamma\boldsymbol{\xi} + \boldsymbol{\zeta}\,, \tag{1.5}$$

where there is one *structural equation* for each variable in random vector $\boldsymbol{\eta}$. And **measurement model**:

$$\mathbf{y} = \Lambda_y\boldsymbol{\eta} + \boldsymbol{\epsilon}_y \tag{1.6}$$

$$\mathbf{x} = \Lambda_x\boldsymbol{\xi} + \boldsymbol{\epsilon}_x\,. \tag{1.7}$$

Let us have N observations of the variables $\mathbf{y}, \mathbf{x}$ for which we will use notation:

$$\mathbb{Y} = \begin{pmatrix}\mathbf{y}_1 & \cdots & \mathbf{y}_N\end{pmatrix}^\top \qquad \mathbb{X} = \begin{pmatrix}\mathbf{x}_1 & \cdots & \mathbf{x}_N\end{pmatrix}^\top. \tag{1.8}$$

Definition 1 (LISREL)**.** *Let data* $[\mathbb{X}, \mathbb{Y}]$ *be independently sampled from distribution of generic random variables* $\mathbf{x}, \mathbf{y}$ *and let* (1.5), (1.6), (1.7) *hold for random variables* $\boldsymbol{\zeta}$, $\boldsymbol{\epsilon}_y$, $\boldsymbol{\epsilon}_x$, $\boldsymbol{\eta}$, $\boldsymbol{\xi}$, $\mathbf{y}$, $\mathbf{x}$ *then we say that data satisfies LISREL if following conditions hold*

I1 $\mathsf{E}[\boldsymbol{\epsilon}_x] = \mathsf{E}[\boldsymbol{\epsilon}_y] = 0$*;* $\boldsymbol{\epsilon}_x \perp\!\!\!\perp \boldsymbol{\epsilon}_y$ *and both are independent of* $\boldsymbol{\xi}, \boldsymbol{\eta}, \boldsymbol{\zeta}$*.*

I2 $\mathsf{E}[\boldsymbol{\zeta}] = 0$ *and* $\boldsymbol{\zeta} \perp\!\!\!\perp \boldsymbol{\xi}$*.*

N1 Each variable $\boldsymbol{\epsilon}_x, \boldsymbol{\epsilon}_y, \boldsymbol{\zeta}$ *has multivariate normal distribution.*

N2 Variables $\boldsymbol{\xi}$ *has multivariate normal distribution.*

It is theoretically possible to relax some conditions and define even more general LISREL model. Firstly the assumptions of normality N1 and N2 are not needed in most parts of Chapter 3. However we will stick with this definition unless explicitly stated otherwise. Note that we require the B to have zeros on the diagonal.

Table 1.1: Description of the variables, coefficients and their properties.

Variable	var	Coefficients: Structural	Coefficients: Loadings	Variable description
$\underset{m\times 1}{\boldsymbol{\zeta}}$	Ψ			structural errors
$\underset{p\times 1}{\boldsymbol{\epsilon}_y}$	Θ_y			measurement errors in $\mathbf{y}$
$\underset{q\times 1}{\boldsymbol{\epsilon}_x}$	Θ_x			measurement errors in $\mathbf{x}$
$\underset{m\times 1}{\boldsymbol{\eta}}$		$\underset{m\times m}{B}$	$\underset{p\times m}{\Lambda_y}$	latent endogenous variables
$\underset{n\times 1}{\boldsymbol{\xi}}$	Φ	$\underset{m\times n}{\Gamma}$	$\underset{q\times n}{\Lambda_x}$	latent exogenous variables
$\underset{p\times 1}{\mathbf{y}}$				indicators of latent endogenous variables
$\underset{q\times 1}{\mathbf{x}}$				indicators of latent exogenous variables

1.3 Model parameters

1.3.1 Implied covariance matrix – Definition

If we specify some LISREL model using above notation we can define theoretical covariance matrix of observed variables, so called **implied covariance matrix**, as a function of model parameters:

$$\Sigma(\theta) = \Sigma(B, \Gamma, \Lambda_y, \Lambda_x, \Phi, \Psi, \Theta_y, \Theta_x). \tag{1.9}$$

When the assumptions I1 and I2 in Definition 1 hold we can rewrite the implied covariance matrix (see [3, p. 235]):

$$\Sigma(\theta) = \begin{bmatrix} \Sigma_{yy}(\theta) & \Sigma_{yx}(\theta) \\ \Sigma_{xy}(\theta) & \Sigma_{xx}(\theta) \end{bmatrix} = \\ \begin{bmatrix} \Lambda_y(I-B)^{-1}(\Gamma\Phi\Gamma^\top + \Psi)\left[(I-B)^{-1}\right]^\top \Lambda_y^\top + \Theta_y & \Lambda_y(I-B)^{-1}\Gamma\Phi\Lambda_x^\top \\ \Lambda_x\Phi\Gamma^\top\left[(I-B)^{-1}\right]^\top \Lambda_y^\top & \Lambda_x\Phi\Lambda_x^\top + \Theta_x \end{bmatrix}. \tag{1.10}$$

Now when the implied covariance is defined we remind the definition of the empirical covariance matrix $\mathcal{S}$ which we are going to compare with the $\Sigma(\theta)$.

$$\mathcal{S}^* = \frac{1}{N}\left[\mathbb{Y}, \mathbb{X}\right]^\top \underbrace{\left(I_N - \frac{1}{N}\mathbf{1}_N\mathbf{1}_N^\top\right)}_{\mathcal{H}}\left[\mathbb{Y}, \mathbb{X}\right]; \quad \mathcal{S} = \frac{N}{N-1}\mathcal{S}^* \tag{1.11}$$

Note that for large samples the difference between $\mathcal{S}$ and $\mathcal{S}^*$ becomes negligible.

1.3.2 Special cases

As convenient it can be to have general results applicable for all possible settings in some cases simplification provides better intuitive insight. There are some special cases of LISREL models (and therefore special structures of implied covariance matrix) that are particularly useful. We state special cases in form of definitions for easier referencing.

Definition 2 (Observed variables model with error in variables)**.** *We say that particular LISREL model si observed variable model (possibly with errors in variables) if* Λ_y, Λ_x *are identity matrices.*

If the variables are measured without error (i.e. $\boldsymbol{\epsilon}_y, \boldsymbol{\epsilon}_x$ are absent) then we get $\mathbf{y} = \boldsymbol{\eta}$ and $\mathbf{x} = \boldsymbol{\xi}$, which can be called as observed variable model without errors in variables or commonly just observed variable model. It is straight forward to show that for observed variables model it holds that:

$$\Sigma(\theta) = \begin{bmatrix} (I-B)^{-1}(\Gamma\Phi\Gamma^\top + \Psi)\left[(I-B)^{-1}\right]^\top + \Theta_y & (I-B)^{-1}\Gamma\Phi \\ \Phi\Gamma^\top\left[(I-B)^{-1}\right]^\top & \Phi + \Theta_x \end{bmatrix}. \tag{1.12}$$

We will mainly focus on observed variables models even though models with latent unobserved variables (see further) are very popular in social science. Observed variables models can bee further distinguished based on structure of matrix B, respectively $B_* = I - B$.

Definition 3. *Let us call an observed variables model without errors in variables:*

1. *Recursive if B_* is lower-triangular and Ψ is diagonal.*
2. *Block recursive if B_* is block lower-triangular and Ψ is block diagonal.*
3. *Nonrecursive if model is neither recursive nor partially recursive.*

We finish off this section with obvious definition of measurement model. Confirmatory factor analysis model is the most known example of the measurement model (see [8, Section 1.4]).

Definition 4. *(Measurement model) We say that particular LISREL model is measurement model if the latent variable model is absent.*

Focus on observed variables model

Even though we provide an overview of the methods usable for general LISREL model the main contribution of this paper is made regarding observed variables model with errors in variables. To that end it is convenient to develop following notation. We say that observations $[\mathbb{X}, \mathbb{Y}]$ follow observed variable SEM if

$$\mathbb{Y} - \boldsymbol{\varepsilon}_y = (\mathbb{Y} - \boldsymbol{\varepsilon}_y)\boldsymbol{\beta} + (\mathbb{X} - \boldsymbol{\varepsilon}_x)\boldsymbol{\Gamma} + \Xi\,, \tag{1.13}$$

where some columns of error terms $\boldsymbol{\varepsilon}$ and some elements of matrices $\boldsymbol{\beta} = B^\top$, $\boldsymbol{\Gamma} = \Gamma^\top$ can be restricted according to specific model in hand. Note that $\boldsymbol{\beta}$ is always restricted to have zeros on the diagonal. This notation will be further extended and modified according to our needs in Chapter 3.

For now just note that the role of the error variables $\boldsymbol{\zeta}$ and $\boldsymbol{\epsilon}_y$ is sometime somehow indistinguishable when analysing observed variable model. More precisely assume $\mathbf{y} - \boldsymbol{\epsilon}_y = \boldsymbol{\eta}$ and $\mathbf{x} = \boldsymbol{\xi}$ plug it in (1.5) and assume $B = 0$. Then we get

$$\mathbf{y} - \boldsymbol{\epsilon}_y = \Gamma\mathbf{x} + \boldsymbol{\zeta},$$

where the role of the measurement error $\boldsymbol{\epsilon}_y$ and equation (structural) error $\boldsymbol{\zeta}$ is clearly undistinguishable. More generally we encounter same problem when separating measurement error in endogenous variable y_j and equation error in j-th structural equation. If we rewrite model in so called reduced form

$$\mathbb{Y} - \boldsymbol{\varepsilon}_y = (\mathbb{X} - \boldsymbol{\varepsilon}_x)\boldsymbol{\Gamma}(I - \boldsymbol{\beta})^{-1} + \Xi(I - \boldsymbol{\beta})^{-1}, \tag{1.14}$$

then distinguishing errors in equations and errors in endogenous variables becomes even more complicated. Later on we will discuss when is it possible to merge errors in equations and errors in variables.

Last note about our specific needs is made regrading categorical variables. Unfortunately covariance based estimators that are proposed later bring some complications when categorical variables are involved (see [3, p. 433–446]) and we did not have enough space to address such problem.

1.4 Parameter estimation

1.4.1 Introduction

There exist several estimation procedures. As we already mentioned some common procedures are based on solving optimization problem

$$\min_{\theta} F\left(\Sigma(\theta), \mathcal{S}\right),$$

where F is some suitable penalising function. However for some models other techniques as OLS or two stage least square (2SLS) are commonly used.

Before we try to estimate the parameters we should investigate if it is even possible to uniquely estimate the parameters. Such problem is in context of structural equation modelling is commonly refered as identification problem. We will include this topic in form of reference to [4, Section 4.3] and [3, Chapter 4, Section Identification].

1.4.2 Covariance based estimators

Generalized least squares (GLS, ULS)

This technique was mentioned in the beginning of this chapter. It is very straight forward and intuitive to define penalizing function to be minimized as

$$F_{GLS_{\mathcal{W}}}\left(\Sigma(\theta), \mathcal{S}\right) = \frac{1}{2}\operatorname{tr}\left\{\left[\left(\Sigma(\theta) - \mathcal{S}\right)\mathcal{W}^{-1}\right]\left[\left(\Sigma(\theta) - \mathcal{S}\right)\mathcal{W}^{-1}\right]^{\top}\right\}. \tag{1.15}$$

As we know from properties of Frobenius norm, this approach takes as core of penalizing function sum of scaled squared differences of elements of implied and observed covariance matrices. One specific choice $\mathcal{W} = \mathcal{S}$ leads to generalized least squares:

$$F_{GLS}\left(\Sigma(\theta), \mathcal{S}\right) = \frac{1}{2}\operatorname{tr}\left\{\left[\Sigma(\theta)\mathcal{S}^{-1} - I\right]\left[\Sigma(\theta)\mathcal{S}^{-1} - I\right]^{\top}\right\} \tag{1.16}$$

and $\mathcal{W} = I$ leads to unweighted least squares:

$$F_{ULS}\left(\Sigma(\theta), \mathcal{S}\right) = \frac{1}{2}\operatorname{tr}\left\{\left[\Sigma(\theta) - \mathcal{S}\right]\left[\Sigma(\theta) - \mathcal{S}\right]^{\top}\right\}. \tag{1.17}$$

Maximum likelihood (ML)

This method is based on assumptions of multivariate normality N1 and N2. Some authors [3] define

$$F_{ML}\left(\Sigma(\theta), \mathcal{S}\right) = \log|\Sigma(\theta)| + \operatorname{tr}\left\{\mathcal{S}\Sigma^{-1}(\theta)\right\} - \log|\mathcal{S}| - (p+q) \tag{1.18}$$

to be minimized while others [4] use directly log-likelihood function which is to be maximized (or one can minimize minus the log-likelihood):

$$l\left(\Sigma(\theta), \mathcal{S}\right) = -\frac{N(p+q)}{2}\log(2\pi) - \frac{N}{2}\left[\log|\Sigma(\theta)| + \operatorname{tr}\left\{\mathcal{S}^{*}\Sigma^{-1}(\theta)\right\}\right]. \tag{1.19}$$

The difference between 1.18 and 1.19 vanishes for large samples, discussion is provided at [3, p. 131–135]. The maximum likelihood method was proposed quite early in [9].

Properties

The above methods are in the literature usually described as optimization problems with few attention to constrains that are imposed on some elements of the implied covariance matrix. One should of course take in account nonnegativity of variance elements.

Under some distributional conditions (N1, N2) one could apply asymptotic theory of maximal likelihood to investigate asymptotic properties of the maximum likelihood estimator. However we now only discuss some distribution free properties of covariance based estimators. They arise straight forward from properties of covariance. The estimates are not affected by scaling whole data by a constant or shifting them. One can also change the order of the variables because reordering columns and rows of implied covariance matrix and empirical covariance matrix does not affect either F_{GLS} nor F_{ML}

1.4.3 Two stage least squares (2SLS)

Unlike previous methods two stage least squares is limited information instrumental variable technique. Limited information because we estimate each structural equation separately. We define 2SLS in context of observed variables model. The 2SLS is one of the most practically used methods because of its simplicity [4]. The rationale of this method is that we replace variable $\mathbb{Y}$ by a variable $\hat{\mathbb{Y}}$ that is highly correlated with $\mathbb{Y}$ and uncorrelated with the error term Ξ.

First formulation

For this method it is useful to partition matrix $\boldsymbol{\beta}$ on vectors of coefficients according to specific model of relations among elements of $\mathbf{y}$:

$$\boldsymbol{\beta} = \left[\beta_1^*, \cdots, \beta_p^*\right],$$

where the vector of coefficient β_i^* is coefficient from equation $y_j = \beta_j^\top \mathbf{y}_j + \cdots$ with added zeros on corresponding positions so $\dim(\beta_i^*) = p$. We handle $\boldsymbol{\Gamma}$ in similar manner. For observed data we also rewrite

$$\mathbb{Y} = [\mathbf{Y}_1, \cdots, \mathbf{Y}_p], \qquad \mathbb{X} = [\mathbf{X}_1, \cdots, \mathbf{X}_q]$$

and denote $\mathbb{Y}_{(j)}$ matrix that contains columns corresponding to endogenous variables from right hand side of j-th structural equation and $\mathbb{X}_{(j)}$ contains exogenous variables from j-th structural equation (with additional intercept also included). And now for every structural equation $j = 1, \ldots, p$ we have

$$\mathbf{Y}_j = \mathbb{Y}_{(j)}\beta_j + \mathbb{X}_{(j)}\gamma_j + \Xi_j. \tag{1.20}$$

At this moment we could jut try to apply simple OLS for every equation to estimate parameters (β_j, γ_j). That would however bring a problem. When estimating nonrecursive model endogenous variables $\mathbb{Y}_{(j)}$ might be correlated with the error term. Instead first consider model:

$$\mathbb{Y}_j = \mathbb{X}_{(j)}\Pi_j + \Delta_j$$

and estimate parameter Π_j using OLS, note fitted values as $\hat{\mathbb{Y}}_{(j)}$. Now let us plugin $\mathbb{Y}_{(j)} = \hat{\mathbb{Y}}_{(j)} + \hat{\Delta}_j$ in right hand side of (1.20) and get

$$\mathbf{Y}_j = (\hat{\mathbb{Y}}_{(j)} + \hat{\Delta}_j)\beta_j + \mathbb{X}_{(j)}\gamma_j + \Xi_j \tag{1.21}$$

$$\mathbf{Y}_j = \hat{\mathbb{Y}}_{(j)}\beta_j + \mathbb{X}_{(j)}\gamma_j + (\hat{\Delta}\beta_j + \Xi_j)$$

$$\mathbf{Y}_j = \left[\hat{\mathbb{Y}}_{(j)}, \mathbb{X}_{(j)}\right] \begin{pmatrix} \beta_j \\ \gamma_j \end{pmatrix} + \Xi_j^* \tag{1.22}$$

and finally we can estimate $[\beta_j, \gamma_j]$ using OLS. In little different setup and under some additional conditions [2] shows consistency of this method. Justification of equivalence to the described method can be found in [4, p. 253–260]. It is essential to argue about the structure of $\Xi_j^* = (\hat{\Delta}_j\beta_j + \Xi_j)$. Because fitted values $\hat{\mathbb{Y}}_j$ and errors $\hat{\Delta}_j$ are uncorrelated (conditionally on $\mathbb{X}$) and $\hat{\Delta}_j$ are uncorrelated with exogenous variables $\mathbb{X}$ (as assumption of OLS in the first step) one can say that Ξ_j^* is in the limit uncorrelated with $[\hat{\mathbb{Y}}_{(j)}, \mathbb{X}]$ which justifies usage of OLS in the second step.

Second formulation

New compact description of this method is provided. Assume that data follow observed variable model $\mathbf{y} = B\mathbf{y} + \Gamma\mathbf{x} + \boldsymbol{\zeta}$. We rewrite model for N observations as

$$\mathbb{Y} = \mathbb{Y}\boldsymbol{\beta} + \mathbb{X}\boldsymbol{\Gamma} + \Xi. \tag{1.23}$$

All exogenous variables are taken as instruments in the way that in the **first step** we consider linear model:

$$\mathbb{Y} = \mathbb{X}\Pi + \Delta,$$

where parameter Π is estimated using OLS. More importantly we take fitted values $\hat{\mathbb{Y}}$ and rewrite observations as $\mathbb{Y} = \hat{\mathbb{Y}} + \hat{\Delta}$, what is now plugged in right hand side of (1.23) and rewritten as

$$\mathbb{Y} = (\hat{\mathbb{Y}} + \hat{\Delta})\boldsymbol{\beta} + \mathbb{X}\boldsymbol{\Gamma} + \Xi$$

$$\mathbb{Y} = \hat{\mathbb{Y}}\boldsymbol{\beta} + \mathbb{X}\boldsymbol{\Gamma} + (\hat{\Delta}\boldsymbol{\beta} + \Xi) \tag{1.24}$$

and in the **second step** estimate $[\boldsymbol{\beta}, \boldsymbol{\Gamma}]$ using constrained OLS (see [10] section 6.3). Constraints are simply setting some parameters to zero according to specified model. As was mentioned before, the matrix $\boldsymbol{\beta}$ should always have zeros on the diagonal.

It might seem unnecessary to even state the second formulation when the first approach gives simple algorithm for estimating 2SLS. However for our future purpose (Section 3.3.3) the first approach would not suffice hence the second compact formulation.

1.4.4 Other methods

Let us mention some other estimation techniques. Starting with the least absolute deviation (LAD) as relatively new technique proposed at [22], being based on

$$F_{LAD}\left(\Sigma(\theta), \mathcal{S}\right) = \mathbf{1}_h^\top \left|\sigma(\theta) - \boldsymbol{s}\right| \tag{1.25}$$

where $\sigma(\theta)$ (respective $\boldsymbol{s}$) is vectorized lower triangular part of $\Sigma(\theta)$ (respective $\mathcal{S}$) and the $\mathbf{1}_h^\top$ is vector of ones of suitable dimension.

There have been developed techniques as three stage least squares 3SLS [24] and some authors [12] discuss Bayesian approach to SEM. Other more recent methods are discussed in [13]. Diversity of proposed methods for estimating SEM seems to be quite large. It is likely result of SEM being class of models rather than one specific type of model.

1.5 Comparison to OLS regression

It is straight forward to see that model (1.1) presented as motivation example is an observed variable LISREL model. We will show that covariance based approach to linear regression problem gives the same result as OLS.

Theorem 1. *Assume standard linear regression model $y = \Gamma\mathbf{x}_* + \zeta$. If there is the intercept included in the model i.e. $\mathbf{x}_*^\top = (1, \mathbf{x}^\top)$, then OLS estimate of non-intercept part of Γ is the same as the covariance based SEM estimate using F_{GLS} or F_{ML}.*

Proof. First investigate implied covariance matrix, note that second column and row are by definition filled wit zeros, because of the intercept.

$$\Sigma(\theta) = \begin{bmatrix} \Gamma\Phi\Gamma^\top + \Psi & \Gamma\Phi \\ \Phi\Gamma^\top & \Phi \end{bmatrix}$$

Take the OLS estimate $\hat{\Gamma}^\top = (\mathbb{X}^\top\mathbb{X})^{-1}\mathbb{X}^\top\mathbb{Y}$. Note that linear regression does not restrict covariance structure among regressors so we can estimate:

$$\hat{\Phi}_{ols} = \frac{1}{N-1}\mathbb{X}^\top\mathcal{H}\mathbb{X}\,,$$

where $\mathcal{H}$ is the centring matrix. Now plug the parameters estimated via OLS in implied covariance matrix. Starting with:

$$\hat{\Phi}_{ols}\hat{\Gamma}^\top = \frac{1}{N-1}\mathbb{X}^\top\mathcal{H}\mathbb{X}(\mathbb{X}^\top\mathbb{X})^{-1}\mathbb{X}^\top\mathbb{Y}.$$

Note that it can be rewritten as estimate of covariances between y and elements of $\mathbf{x}_*$. Now we use property of liner regression model that $SS_T = SS_e + SS_R$ (see Theorem 9) to handle $\Gamma\Phi\Gamma^\top + \Psi$. The Ψ is straight forward, by definition it holds $\hat{\Psi}_{ols} = SS_e$. When plugging OLS estimates in the $\Gamma\Phi\Gamma^\top$, we get

$$\hat{\Gamma}\hat{\Phi}_{ols}\hat{\Gamma}^\top = \frac{1}{N-1}\underbrace{\mathbb{Y}^\top\mathbb{X}(\mathbb{X}^\top\mathbb{X})^{-1}\mathbb{X}^\top}_{\hat{\mathbb{Y}}^\top}\mathcal{H}\underbrace{\mathbb{X}(\mathbb{X}^\top\mathbb{X})^{-1}\mathbb{X}^\top\mathbb{Y}}_{\hat{\mathbb{Y}}} = SS_R$$

Therefore $\hat{\Gamma}\hat{\Phi}_{ols}\hat{\Gamma}^\top + \hat{\Psi}_{ols}$ equals to empirical variance of y. Now it is clear that $\Sigma(\hat{\theta}_{ols}) = \mathcal{S}$ therefore F_{GLS} is zero and because it is nonnegative we found its minimum. Similar reasoning hold for F_{ML}.

□

This result is not very surprising and to some readers it might seem obvious. However some authors [7] place not negligible importance (especially for educational purposes) on viewing linear regression as structural equation modelling.

2. Error-in-Variables Model, Total least squares

2.1 Motivation

Let us return to simple linear regression once more. But this time we assume that regressor x is measured with an error. In other words our model is:

$$y = \alpha + \gamma \xi + \epsilon_y \ , \quad x = \xi + \epsilon_x. \tag{2.1}$$

Note the similarity with structural equations modelling. The first part can be seen as latent variable model and the second part as measurement model. Let us have data $\mathbb{X}, \mathbb{Y}$ consisting of N independent observations of random variables (x, y). Note that the first column of $\mathbb{X} = [\mathbf{1}, \mathbf{X}]$ is the intercept which is unlike x assumed to be without any error. The idea is to minimize errors in both variables y, x. More precisely:

$$\min_{[\boldsymbol{\varepsilon}_x, \boldsymbol{\varepsilon}_y] \in \mathbb{R}^{N \times 2}, \alpha, \gamma} \|[\boldsymbol{\varepsilon}_x, \boldsymbol{\varepsilon}_y]\|_F \quad \text{substitute to} \quad \mathbb{Y} - \boldsymbol{\varepsilon}_y = (\mathbb{X} - [\mathbf{0}, \boldsymbol{\varepsilon}_x]) \begin{pmatrix} \alpha \\ \gamma \end{pmatrix}$$

Such estimation technique is called total least squares. We further develop this idea in more general setup of partial Errors-in-Variables model. One can also incorporate non-random intercept and categorical variables in such model.

2.2 Partial Errors-in-Variables Model

2.2.1 Definition

We start with definition of error-in-variable estimator in setting where all regressors are subject to an error. Let the data $[\mathbb{X}, \mathbb{Y}]$ consist of independent observations of the random variables $\mathbf{y} \in \mathbb{R}^p$ and $\mathbf{x} \in \mathbb{R}^q$. We define:

Definition 5 (EIV model)**.** *We say that data $[\mathbb{X}, \mathbb{Y}]$ follow errors-in-variables (EIV) model if*

$$\mathbb{Y} = \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\varepsilon}_y \quad \text{and} \quad \mathbb{X} = \mathbf{Z} + \boldsymbol{\varepsilon}_x, \tag{2.2}$$

where $\boldsymbol{\beta}$ is matrix of regression coefficients, $\mathbb{X}$ and $\mathbb{Y}$ are observed random variables and $\mathbf{Z}$ is full rank matrix of unknown constants.

Definition 6 (EIV estimator)**.** *Suppose that $\|\cdot\|$ is an unitary invariant matrix norm. Consider optimization problem:*

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^n, \, [\boldsymbol{\varepsilon}_x, \boldsymbol{\varepsilon}_y] \in \mathbb{R}^{N \times (q+p)}} \|[\boldsymbol{\varepsilon}_x, \boldsymbol{\varepsilon}_y]\| \quad s.t. \quad \mathbb{Y} - \boldsymbol{\varepsilon}_y = (\mathbb{X} - \boldsymbol{\varepsilon}_x)\boldsymbol{\beta}. \tag{2.3}$$

If there exists a solution $\{\hat{\boldsymbol{\beta}}, [\hat{\boldsymbol{\varepsilon}}_x, \hat{\boldsymbol{\varepsilon}}_y]\}$ to (2.3) then any $\hat{\boldsymbol{\beta}}$, satisfying

$$\mathbb{Y} - \hat{\boldsymbol{\varepsilon}}_y = (\mathbb{X} - \hat{\boldsymbol{\varepsilon}}_x)\hat{\boldsymbol{\beta}}$$

is called an errors-in-variables estimator (EIV estimator). If (2.3) has no solution then the estimator is a fixed matrix.

As is shown at [14] it is generally possible to use any unitary invariant matrix norm and still arrive to the same results as when using Frobenius norm $\|\cdot\|_F$. We will mostly use Frobenius norm $\|\cdot\|_F$ unless explicitly stated otherwise. Next theorem provides solution to errors-in-variables problem.

Theorem 2 (EIV solution)**.** *To solve errors-in-variables problem first compute singular value decomposition (SVD) of the data matrix:*

$$[\mathbb{X}, \mathbb{Y}] = [\underbrace{U_{11}, U_{12}}_{U_1}, U_2] \begin{bmatrix} \Sigma \\ 0 \end{bmatrix} V^\top = U_1 \Sigma V^\top,$$

where

$$V = \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix}, \quad V_{11} \in \mathbb{R}^{q \times q},\ V_{22} \in \mathbb{R}^{p \times p},\ V_{12} \in \mathbb{R}^{q \times p}$$

and analogously

$$\Sigma = \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{bmatrix}, \quad \Sigma_1 \in \mathbb{R}^{q \times q},\ \Sigma_2 \in \mathbb{R}^{p \times p}.$$

If V_{22} is nonsingular, then errors-in-variables estimator is

$$\hat{\boldsymbol{\beta}} = -V_{12} V_{22}^{-1}$$

and fitted values are

$$[\hat{\mathbb{X}}, \hat{\mathbb{Y}}] = [\mathbb{X} - \hat{\boldsymbol{\varepsilon}}_x, \mathbb{Y} - \hat{\boldsymbol{\varepsilon}}_y] = U_{11} \Sigma_1 [V_{11}^\top, V_{21}^\top].$$

Proof. This estimation procedure is nicely summarized and justified for example in [20] as Algorithm 1.

□

Now we assume that some regressors are measured precisely. Comprehensive summary of following method is provided at [15]. Let us denote $\mathbb{X} = [\mathbf{W}, \mathbf{X}]$ so we have separated regressors subjected to an error in $\mathbf{X}$ and regressors measured without any error $\mathbf{W}$. Partial errors-in-variables model is:

Definition 7 (PEIV model)**.** *We say that data $[\mathbb{X}, \mathbb{Y}]$ follow partial error-in-variable (PEIV) model if*

$$\mathbb{Y} = \mathbf{W}\boldsymbol{\alpha} + \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\varepsilon}_y \quad \text{and} \quad \mathbf{X} = \mathbf{Z} + \boldsymbol{\varepsilon}_x. \tag{2.4}$$

Definition 8 (PEIV estimator)**.** *Suppose that $\|\cdot\|$ is an unitary invariant matrix norm. Consider optimization problem:*

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^n,\, [\boldsymbol{\varepsilon}_x, \boldsymbol{\varepsilon}_y] \in \mathbb{R}^{N \times (q+p)}} \|[\boldsymbol{\varepsilon}_x, \boldsymbol{\varepsilon}_y]\| \quad s.t. \quad \mathbb{Y} - \boldsymbol{\varepsilon}_y = [\mathbf{W}, \mathbf{X} - \boldsymbol{\varepsilon}_x] \begin{pmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{pmatrix}. \tag{2.5}$$

If there exists a solution $\{(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}), [\hat{\boldsymbol{\varepsilon}}_x, \hat{\boldsymbol{\varepsilon}}_y]\}$ to (2.5) then any $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})$ satisfying

$$\mathbb{Y} - \hat{\boldsymbol{\varepsilon}}_y = [\mathbf{W}, \mathbf{X} - \hat{\boldsymbol{\varepsilon}}_x] \begin{pmatrix} \hat{\boldsymbol{\alpha}} \\ \hat{\boldsymbol{\beta}} \end{pmatrix}$$

is called a partial errors-in-variables estimator (PEIV estimator). If (2.5) has no solution then the estimator is a fixed matrix.

To estimate coefficients in PEIV model, we firs define projection matrix to space orthogonal to $\mathcal{M}(\mathbf{W})$ i.e. to space $\mathcal{M}(\mathbf{W})^{\perp}$:

$$\mathbf{R} = I - \mathbf{W}\left(\mathbf{W}^{\top}\mathbf{W}\right)^{-1}\mathbf{W}^{\top}$$

and apply it on (2.4) which leads to

$$\mathbf{R}\mathbb{Y} = \mathbf{R}\mathbf{Z} + \mathbf{R}\boldsymbol{\varepsilon}_y \quad \text{and} \quad \mathbf{R}\mathbf{X} = \mathbf{R}\mathbf{Z} + \mathbf{R}\boldsymbol{\varepsilon}_x\,, \tag{2.6}$$

now we apply EIV estimator (Theorem 2) on this transformed model. Once the estimate $\hat{\boldsymbol{\beta}}$ is found, we subtract $\mathbf{X}\hat{\boldsymbol{\beta}}$ from the observations and estimate another model:

$$\mathbb{Y} - \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{W}\boldsymbol{\alpha} + \tilde{\boldsymbol{\varepsilon}}_y, \tag{2.7}$$

where we use OLS to estimate the parameter $\boldsymbol{\alpha}$ resulting in:

$$\hat{\boldsymbol{\alpha}} = \left(\mathbf{W}^{\top}\mathbf{W}\right)^{-1}\mathbf{W}^{\top}\left(\mathbb{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right).$$

2.2.2 Example

As an illustrative example we investigate association of two variables x and y where both variables are subjected to an error. It may not be clear which one, we should consider as the response variable and which one as the regrresor, especially when both variables are measured by psychological questionnaire.

One could of course consider correlation analysis, but that is not focus of this paper. Let us compare 4 models. First consider $y = \alpha + \beta x$ and estimate parameter using OLS, then $x = \alpha + \beta_* y$ and again use OLS, then inverse estimated

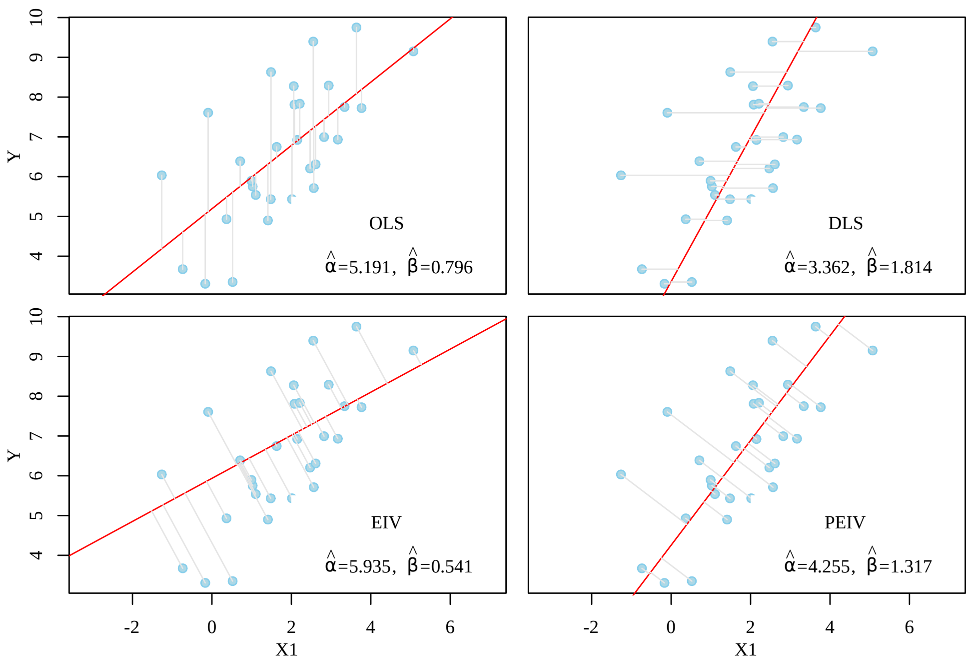

Figure 2.1: Comparison of OLS, DLS, EIV, PEIV, estimators. The true values of the parameters are $\alpha = 4, \beta = 1.3$.

coefficient. This method can be called data least squares (DLS). Then we use errors-in-variables model and finally (in this case most suitable model) partial errors-in-variables model.

At the figure 2.1 we can see that PEIV estimation seems to perform relatively well in this case. On the other hand EIV model with random intercept does not seem to perform well. We should be careful when deciding which regressors are assumed to be measured with an error. It does not make much sense to consider error in the intercept.

Advantage of EIV when analysing sociological or psychological study is that the variables are treated symmetrically. We get the same regression line regardless of the role of the variables, unlike when using OLS. This is to some extend advantageous but also imposes restrictive assumptions to meet in order to get consistent estimator.

2.3 Properties of EIV estimator

We note some of the properties of EIV estimator starting with strong consistency. To that end we firstly specify more precisely the error structure. We will impose fairly strong assumptions on our data.

Let rows of the data $[\mathbb{X}, \mathbb{Y}, \boldsymbol{\varepsilon}_x, \boldsymbol{\varepsilon}_y]$ be independent identically distributed with generic distribution of random variables[1] $[\mathbf{y}^\top, \mathbf{x}^\top, \boldsymbol{\epsilon}_x^\top, \boldsymbol{\epsilon}_y^\top]$. In the following theorem we impose additional assumption on error variables $[\boldsymbol{\epsilon}_x, \boldsymbol{\epsilon}_y]$.

Theorem 3 (Strong consistency of EIV estimator)**.** *Let data* $[\mathbb{X}, \mathbb{Y}]$ *of size* N *follow errors-in-variables model. If following hold*

A1 The errors have zero mean and are homoscedastic i.e.

$$var\begin{pmatrix}\boldsymbol{\epsilon}_x \\ \boldsymbol{\epsilon}_y\end{pmatrix} = \sigma^2 I_{p+q}\,.$$

A2 The $\lim_{N\to\infty} \frac{1}{N}\mathbf{Z}^\top\mathbf{Z}$ *exists and is positive definite.*

Then EIV estimator proposed in Theorem 2 is strongly consistent estimator of the true parameter $\boldsymbol{\beta}$*.*

$$\lim_{N\to\infty} \hat{\boldsymbol{\beta}}_N = \boldsymbol{\beta} \text{ } a.s.$$

Proof. See [6] Lemma 3.3.

□

We will not proceed much further with statistical properties of EIV as main focus of this paper lies elsewhere. For a summary of statistical properties one can see [23] Chapter 8. Conditions for consistency of PEIV can be split. Let (2.6) fulfil the conditions of Theorem 3 and let (2.7) with the true value of $\boldsymbol{\beta}$ fulfil the conditions for strong consistency of OLS (see Theorem 8).

Now we address other properties. The following theorems 4 and 5 show that the estimates are not affected by scaling whole data by a constant or changing the order of the variables. This properties are similar as covariance based SEM estimators. Let us denote applying the EIV estimator on the data by $EIV([\mathbb{X}, \mathbb{Y}])$.

[1]Note difference between $\boldsymbol{\varepsilon}$ and $\boldsymbol{\epsilon}$

Theorem 4 (scale invariance)**.** *EIV estimator is scale invariant i.e. for all $a > 0$, it holds that*

$$EIV([\mathbb{X}, \mathbb{Y}]) = EIV(a[\mathbb{X}, \mathbb{Y}]).$$

Proof. Straightforward generalization of [14] Theorem 3 (i). □

Theorem 5 (interchange equivariance)**.** *EIV estimator is interchange equivariant i.e. for any permutation π and its permutation matrix P_π it holds*

$$EIV([\mathbb{X}, \mathbb{Y}]\tilde{P}_\pi) = P_{\pi^{-1}} EIV([\mathbb{X}, \mathbb{Y}]), \quad \text{where} \quad \tilde{P}_\pi = \begin{bmatrix} P_\pi & 0 \\ 0 & I \end{bmatrix}.$$

Proof. Straightforward generalization of [14] Theorem 3 (ii). □

Theorem 6 (direction equivariance)**.** *EIV estimator is direction equivariant i.e. for any diagonal matrix D which has only ± 1 on diagonal it holds*

$$EIV([\mathbb{X}, \mathbb{Y}]\tilde{D}) = -D\, EIV([\mathbb{X}, \mathbb{Y}]), \quad \text{where} \quad \tilde{D} = \begin{bmatrix} D & 0 \\ 0 & I \end{bmatrix}.$$

Proof. Straightforward generalization of [14] Theorem 3 (iii). □

Theorem 7 (rotation equivariance)**.** *EIV estimator is rotation equivariant i.e. for any rotation matrix R it holds*

$$EIV([\mathbb{X}, \mathbb{Y}]\tilde{R}) = R^\top EIV([\mathbb{X}, \mathbb{Y}]), \quad \text{where} \quad \tilde{R} = \begin{bmatrix} R & 0 \\ 0 & I \end{bmatrix}.$$

Proof. Straightforward generalization of [14] Theorem 3 (iv). □

Remark. Note that estimating linear model with multivariate response using standard OLS without any restrictions on parameters is equivalent to estimating univariate response model for each response variable separately. Such equivalence does not hold when using TLS.

2.4 Euler–Lagrange solution to constrained EIV problem

Let us again consider errors-in-variables problem (2.2). There is another method (see for example [20]) for estimating coefficients of such model. It is based on solving optimization problem 2.3 using Lagrange function:

$$\mathcal{L}(\boldsymbol{\varepsilon}_y, \boldsymbol{\varepsilon}_x, \Lambda^\top, \boldsymbol{\beta}) = \operatorname{tr}\left\{\boldsymbol{\varepsilon}_y\boldsymbol{\varepsilon}_y^\top\right\} + \operatorname{tr}\left\{\boldsymbol{\varepsilon}_x\boldsymbol{\varepsilon}_x^\top\right\} + 2\operatorname{tr}\left\{\Lambda^\top\left[\mathbb{Y} - \boldsymbol{\varepsilon}_y - (\mathbb{X} - \boldsymbol{\varepsilon}_x)\boldsymbol{\beta}\right]\right\}. \quad (2.8)$$

One can also quite easily modify the Lagrange function to include constraint $G\boldsymbol{\beta} = W$, see for example [25]. Modified Lagrange function is then

$$\mathcal{L}(\boldsymbol{\varepsilon}_y, \boldsymbol{\varepsilon}_x, \Lambda^\top, \boldsymbol{\beta}) =$$
$$\operatorname{tr}\left\{\boldsymbol{\varepsilon}_y\boldsymbol{\varepsilon}_y^\top\right\} + \operatorname{tr}\left\{\boldsymbol{\varepsilon}_x\boldsymbol{\varepsilon}_x^\top\right\} + 2\operatorname{tr}\left\{\Lambda_2^\top\left[\mathbb{Y} - \boldsymbol{\varepsilon}_y - (\mathbb{X} - \boldsymbol{\varepsilon}_x)\boldsymbol{\beta}\right]\right\} + 2\operatorname{tr}\left\{\Lambda_2^\top(W - G\boldsymbol{\beta})\right\}.$$

Constrained problem is important because, as was already said, we can not simply partition matrix of coefficients on vectors as we could do with linear regression. It is discussed in [21] Remark 2.

Partial error in variable model could be treated as in previous the chapter by projecting out the exact variables or one could formulate

$$\mathcal{L}(\boldsymbol{\varepsilon}_y, \boldsymbol{\varepsilon}_x, \Lambda^\top, \boldsymbol{\beta}) = \operatorname{tr}\left\{\boldsymbol{\varepsilon}_y\boldsymbol{\varepsilon}_y^\top\right\} + \operatorname{tr}\left\{\boldsymbol{\varepsilon}_x\boldsymbol{\varepsilon}_x^\top\right\} +$$
$$2\operatorname{tr}\left\{\Lambda_2^\top\left[\mathbb{Y} - \boldsymbol{\varepsilon}_y - [\mathbf{W}, \mathbf{X} - \boldsymbol{\varepsilon}_x]\boldsymbol{\beta}\right]\right\} + 2\operatorname{tr}\left\{\Lambda_2^\top(W - G\boldsymbol{\beta})\right\}.$$

Once the Lagrange function is defined we need to solve the necessary optimality conditions, meaning we set all partial derivatives of the Lagrange function to zero and solve resulting system. In case of (2.8) we get partial derivatives:

$$\frac{1}{2}\frac{\partial \mathcal{L}}{\partial \boldsymbol{\varepsilon}_y} = \boldsymbol{\varepsilon}_y - \Lambda \quad (2.9)$$
$$\frac{1}{2}\frac{\partial \mathcal{L}}{\partial \boldsymbol{\varepsilon}_x} = \boldsymbol{\varepsilon}_x + \Lambda\boldsymbol{\beta}^\top \quad (2.10)$$
$$\frac{1}{2}\frac{\partial \mathcal{L}}{\partial \Lambda} = \mathbb{Y} - \boldsymbol{\varepsilon}_y - \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}_x\boldsymbol{\beta} \quad (2.11)$$
$$\frac{1}{2}\frac{\partial \mathcal{L}}{\partial \boldsymbol{\beta}} = (-\mathbb{X}^\top + \boldsymbol{\varepsilon}_x^\top)\Lambda \quad (2.12)$$

Now it remains to solve corresponding nonlinear system of equations. For that one can use Algorithm 2 or Algorithm 3 [20].

3. Errors-in-Variables in context of structural equations model

3.1 Motivation

As was mentioned before the model (2.1) can be viewed as structural equations model with the latent variable model:

$$\underbrace{y - \epsilon_y}_{\eta} = \alpha + \gamma\xi + \zeta \tag{3.1}$$

and the measurement model:

$$\begin{aligned} y &= \eta + \epsilon_y \\ x &= \xi + \epsilon_x \,. \end{aligned} \tag{3.2}$$

For simplicity of notation the measurement model for intercept was left out as it is constant 1 ”measured” without any error and does not influence rest of the covariance structure. Variances of the variables $\xi, \epsilon_y, \epsilon_x, \zeta$ are:

$$\mathrm{var}(\xi) = \phi, \quad \mathrm{var}(\epsilon_y) = \vartheta_{\epsilon_y}, \quad \mathrm{var}(\epsilon_x) = \vartheta_{\epsilon_x}, \quad \mathrm{var}(\zeta) = \psi$$

and coefficients are: $B = 0$, $\Gamma = (\gamma)$, $\Lambda_x = I_1$, $\Lambda_y = I_1$. One can write the implied covariance matrix in the following form:

$$\Sigma(\theta) = \begin{bmatrix} (\gamma^2\phi + \psi) + \vartheta_{\epsilon_y} & \phi\gamma \\ \phi\gamma & \phi + \vartheta_{\epsilon_x} \end{bmatrix}. \tag{3.3}$$

Using covariance based SEM approach to estimate parameters of such model bears some complications. We have just 3 non-degenerated covariances of the observed variables and 5 parameters to estimate. Even if we consider $\psi + \vartheta_{\epsilon_y} = \psi_*$ as one parameter we still have one more parameter to estimate than we have sample covariances.

One possible solution of the identification problem is based on using some complementary information about measurement error, for more details see [18]. Or one could use EIV estimator instead. In this chapter we will try to utilize EIV estimator even for more complex SEM. Brief discussion of some modifications of the covariance based estimation is also provided.

If there was no error in measurement of x the mentioned rewriting $\psi + \vartheta_{\epsilon_y} = \psi_*$ would help to identify the problem. In such case one can not distinguish the error in the variable ϵ_y and the error in the equation ζ.

3.2 Problems specification

In this chapter we address problem of errors in measurement and we will try to find conditions under which the EIV estimator is sensible approach to SEM model. We also propose estimation technique for the observed variables model based on generalization of two stage least squares method. But first let us review the notation in convenient manner. Trough this whole chapter we consider the observed variable model

$$\mathbf{y} - \boldsymbol{\epsilon}_y = B(\mathbf{y} - \boldsymbol{\epsilon}_y) + \Gamma(\mathbf{x} - \boldsymbol{\epsilon}_x) + \boldsymbol{\zeta}. \tag{3.4}$$

The observations of the endogenous variables are $\mathbb{Y}$, observations of the exogenous variables measured without any error are $\mathbf{W}$ and $\mathbf{X}$ are those with an error. The model can be rewritten in terms of observations as

$$\mathbb{Y} - \boldsymbol{\varepsilon}_y = (\mathbb{Y} - \boldsymbol{\varepsilon}_y)\boldsymbol{\beta} + [\mathbf{W}, \mathbf{X} - \boldsymbol{\varepsilon}_x]\begin{bmatrix}\boldsymbol{\Gamma}_1\\ \boldsymbol{\Gamma}_2\end{bmatrix} + \Xi. \tag{3.5}$$

Note that one can restrict some columns of $\boldsymbol{\varepsilon}_y$ to be zero if the variable is measured without error. We will discuss this in more detail later on.

In some special cases it might seem appropriate to use EIV (or constrained EIV) estimation directly. That is the case when we are not really interested in relations among the endogenous variables i.e. the structure of $\boldsymbol{\beta}$ is not of interest and therefore $\boldsymbol{\beta}$ is not included in the model. As was mentioned before estimating model with multivariate response using OLS is the same as estimating multiple models separately. Not so when using EIV estimator.

However when we consider structural equations modelling we are often specifying relations among endogenous variables more precisely and the coefficient $\boldsymbol{\beta}$ is of main interest. That will be the focus of the following sections.

3.3 Modified estimation methods

3.3.1 Constrained total least squares

We try to modify and generalize the constrained EIV estimator [25, 21, 19] and develop Lagrange function based approach for solving SEM problem. To that end we define optimization problem in full generality:

$$\min_{\boldsymbol{\varepsilon}_y, \boldsymbol{\varepsilon}_x, \boldsymbol{\beta}, \boldsymbol{\Gamma}, \Xi} \operatorname{tr}\left\{\boldsymbol{\varepsilon}_y \boldsymbol{\varepsilon}_y^\top\right\} + \operatorname{tr}\left\{\boldsymbol{\varepsilon}_x \boldsymbol{\varepsilon}_x^\top\right\} + \operatorname{tr}\left\{\Xi\Xi^\top\right\} \tag{3.6}$$

$$s.t.$$

$$\begin{gathered}[\mathbf{Y}^0, \mathbf{Y}^\epsilon - \boldsymbol{\varepsilon}_y] = [\mathbf{Y}^0, \mathbf{Y}^\epsilon - \boldsymbol{\varepsilon}_y]\begin{bmatrix}\boldsymbol{\beta}_0\\ \boldsymbol{\beta}_\epsilon\end{bmatrix} + \Xi; +[\mathbf{W}, \mathbf{X} - \boldsymbol{\varepsilon}_x]\begin{bmatrix}\boldsymbol{\Gamma}_1\\ \boldsymbol{\Gamma}_2\end{bmatrix} + \Xi\,; \\ D\boldsymbol{\beta} = 0\,; \quad G\boldsymbol{\Gamma} = 0\,.\end{gathered} \tag{3.7}$$

Structure of constraint matrices D, G is determined by the assumed model. When there is an error in observations of all endogenous variables the we can simplify

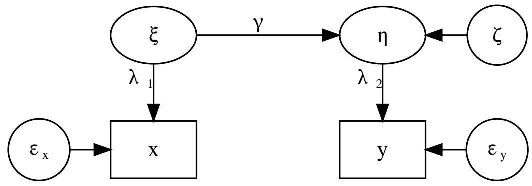

Figure 3.1: Path diagram of model (3.1)

$[\mathbf{Y}^0, \mathbf{Y}^\epsilon - \boldsymbol{\varepsilon}_y] = \mathbb{Y} - \boldsymbol{\varepsilon}_y$. Note that G can be partitioned on G_1, G_2 similarly as corresponding $\boldsymbol{\Gamma}$ is partitioned on $\boldsymbol{\Gamma}_1, \boldsymbol{\Gamma}_2$. Now let us define Lagrange function:

$$\begin{gathered}\mathcal{L}(\boldsymbol{\varepsilon}_y, \boldsymbol{\varepsilon}_x, \Lambda, \Lambda_1, \Lambda_2, \boldsymbol{\beta}, \boldsymbol{\Gamma}) = \operatorname{tr}\left\{\boldsymbol{\varepsilon}_y \boldsymbol{\varepsilon}_y^\top\right\} + \operatorname{tr}\left\{\boldsymbol{\varepsilon}_x \boldsymbol{\varepsilon}_x^\top\right\} + \operatorname{tr}\left\{\Xi\Xi^\top\right\} \\ +2\operatorname{tr}\left\{\Lambda^\top\Big([\mathbf{Y}^0, \mathbf{Y}^\epsilon - \boldsymbol{\varepsilon}_y] - [\mathbf{Y}^0, \mathbf{Y}^\epsilon - \boldsymbol{\varepsilon}_y]\boldsymbol{\beta} - [\mathbf{W}, \mathbf{X} - \boldsymbol{\varepsilon}_x]\boldsymbol{\Gamma} - \Xi\Big)\right\} \\ +2\operatorname{tr}\left\{\Lambda_1^\top D\boldsymbol{\beta}\right\} + 2\operatorname{tr}\left\{\Lambda_2^\top G\boldsymbol{\Gamma}\right\}.\end{gathered}$$

We shall proceed by solving the necessary optimality conditions i.e. solving system where we take all partial derivatives of Lagrange function to be zero (similarly as in [25]). That procedure leads to complicated non-linear system of equations that needs to be solved numerically.

Solving the system in full generality (or just determining existence of solution) could be potentially very tedious therefore it seems convenient to consider some simplification. For now we will leave this general approach opened to future research.

First simplification, that one can propose, leaves the Ξ error in equation out and assume that all endogenous variables are measured with an error. But note that this approach may lead to wrong model specification, on the other hand may be taken advantage of in some special cases in following section.

3.3.2 Equation-wise Error-in-Variable estimation

Observed variable model (3.5) can be rewritten in terms of p equations as in (1.20). In following we will justify using method where we estimate for each j

$$\mathbf{Y}_j \approx \mathbb{Y}_{(j)}\beta_j + \mathbb{X}_{(j)}\gamma_j \tag{3.8}$$

using PEIV estimator where $\mathbb{Y}_{(j)}$ is considered to be measured with an error. We start with an example. In the following example we assume that there are no measurement errors in exogenous variables. Consider observed variable model in the generic form as

$$\begin{aligned} y_1 &= \beta_{12}y_2 + \gamma_{11}x_1 + \gamma_{13}x_3 + \zeta_1 \\ y_2 &= \beta_{21}y_1 + \gamma_{22}x_2 + \zeta_2 \end{aligned} \tag{3.9}$$

and denote observations of the exogenous variables by $\mathbb{X} = [\mathbf{1}, \mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3]$ and observations of the endogenous variables by $\mathbb{Y} = [\mathbf{Y}_1, \mathbf{Y}_2]$. Besides previously mentioned disadvantages of OLS, we can say that, if we estimate each equation separately with 0LS, we are treating y_j somehow differently based on if it is on left hand or right hand side of the equation. To some extend one can see the rationale of this method as fairly similar to 2SLS[1]. Recall the second step of the two stage least squares. Plugging in fitted values from the first step to (1.20) gives

$$\mathbf{Y}_1 = (\hat{\mathbf{Y}}_2 + D)\beta_{12} + \mathbb{X}_j\gamma_j + \Xi_j\,. \tag{3.10}$$

Now we can view the problem as error-in-variable model where $\mathbf{Y}_2 - D = \hat{\mathbf{Y}}_2$ plays role of the unobserved $\mathbf{Z}$. We proceed using PEIV, where $\mathbf{Y}_2$ is considered to be with an error, instead of plugging in the values from the first step.

[1]Indeed I got the idea while studying the two stage least squares.

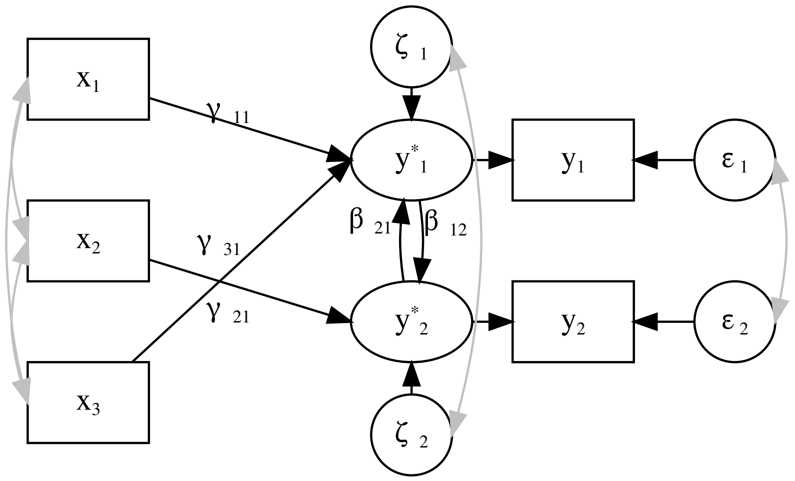

Figure 3.2: Path diagram of the model (3.9).

Let us investigate a modification of the model (3.9) before proceeding further. One could say that it is preferable to assume that there are errors in measurements of the endogenous variables and rewrite the model as

$$\begin{aligned} y_1 - \epsilon_1 &= \beta_{12}(y_2 - \epsilon_2) \quad + \gamma_{11} x_1 \qquad\qquad + \gamma_{13} x_3 + \zeta_1 \\ \underbrace{y_2 - \epsilon_2}_{y_2^*} &= \beta_{21} \underbrace{(y_1 - \epsilon_1)}_{y_2^*} \qquad\quad + \gamma_{22} x_2 \qquad\qquad + \zeta_2 . \end{aligned} \tag{3.11}$$

The model is visualized at Figure 3.2. This model provides more symmetrical view of the endogenous variables regardless of their position. Note that the error terms ζ_j and ϵ_j are somehow inseparable when estimating each equation separately.

The main idea becomes even clearer when we rewrite the model in terms of the observations $\mathbb{X}$ and $\mathbb{Y}$. As was mentioned, we assume that the observations of the endogenous variables are affected by an error, so we write $\mathbf{Y}_j = \mathbf{Y}_j^* + \boldsymbol{\varepsilon}_i$ where $\mathbf{Y}_j^*$ are the true values. Let us rewrite the model (3.11) in the following form

$$(\mathbf{Y}_1 - \boldsymbol{\varepsilon}_1) = (\mathbf{Y}_2 - \boldsymbol{\varepsilon}_2)\beta_{12} + [\mathbf{X}_1, \mathbf{X}_3] \begin{pmatrix} \gamma_{11} \\ \gamma_{13} \end{pmatrix} + \Xi_1 \tag{3.12}$$

$$(\mathbf{Y}_2 - \boldsymbol{\varepsilon}_2) = (\mathbf{Y}_1 - \boldsymbol{\varepsilon}_1)\beta_{21} + \mathbf{X}_2 \gamma_{22} + \Xi_2 \, . \tag{3.13}$$

When estimating both (3.12) and (3.13) separately using PEIV we practically merge the errors Ξ_j and $\boldsymbol{\varepsilon}_j$ in one $\dot{\boldsymbol{\varepsilon}}_j$. However due to this fact and reciprocal nature of relation between y_1 and y_2 we can not generally consider $\dot{\epsilon}_j, \epsilon_{j+(j \mod 2)}$ to be uncorrelated. Yet with some extra effort we could overcome this complication and describe conditions under which use of the PEIV estimator is appropriate. Also note that the intercept was omitted until now for simplification of the notation.

Justification of generalization of the equation-wise EIV for more complex models is theoretically straightforward (One can easily include errors in exogenous variables) yet practice may bear many difficulties with specifying correct relations among the variables and meeting conditions imposed on the errors.

Theorem 8 (Consistency). *Assume that the data* $[\mathbb{X}, \mathbb{Y}]$ *(consisting of N independent observations) follow observed variable model with errors in endogenous variables and without errors in equation (structural errors). Let j-th structural equation be*

$$\mathbf{Y}_j - \boldsymbol{\varepsilon}_j = (\mathbb{Y}_{(j)} - \boldsymbol{\varepsilon}_{(j)})\boldsymbol{\beta}_j + \mathbb{X}_{(j)}\boldsymbol{\Gamma}_j \,.$$

Assume there exist unobserved constants $\mathbf{Z}_{(j)}$ *such that* $\mathbb{Y}_{(j)} = \mathbf{Z}_{(j)} + \boldsymbol{\varepsilon}_{\mathbb{Y}_j}$. *Denote the projection matrix to* $\mathcal{M}\left(\mathbb{X}_{(j)}\right)^{\perp}$ *by* $\mathbf{R}$ *and the generic random variable that generates rows of the* $\boldsymbol{\varepsilon}$ *by* $\boldsymbol{\epsilon}$. *If the following holds*

1. $\mathsf{E}[\epsilon_j, \boldsymbol{\epsilon}_{(j)}] = \mathbf{0}$ *and* $var[\epsilon_j, \boldsymbol{\epsilon}_{(j)}] = \sigma^2 I$.
2. $\lim_{N\to\infty} \frac{1}{n}\mathbf{Z}^\top\mathbf{R}\mathbf{Z}$ *exists and is positive definite.*
3. *The transformed model* $\mathbf{Y}_j - \mathbb{Y}_{(j)}\boldsymbol{\beta} = \mathbb{X}_j\boldsymbol{\Gamma} + \tilde{\boldsymbol{\varepsilon}}_y$ *fulfils the conditions of Theorem 10 for OLS estimator of* $\boldsymbol{\Gamma}$ *to be strongly consistent.*

Then PEIV estimator of the parameters $\boldsymbol{\beta}, \boldsymbol{\Gamma}$ *is strongly consistent i.e.*

$$\lim_{N\to\infty} \begin{pmatrix} \hat{\boldsymbol{\beta}}_j^{(N)} \\ \hat{\boldsymbol{\Gamma}}_j^{(N)} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\beta}_j \\ \boldsymbol{\Gamma}_j \end{pmatrix} \quad a.s.$$

Proof. It is enough to realize that the first two assumptions correspond to the assumptions for strong consistency of EIV estimator in [6] Lemma 3.3 and corollary 3.1. And the third condition assures consistency of $\hat{\boldsymbol{\Gamma}}_j^N$.

□

Generalization of the previous theorem for models, where some exogenous variables are subjected to errors is straightforward. To make use of the previous theorem let us now rewrite the observed variable model in so called reduced form. As was mentioned before the structural errors and the errors in the endogenous variables might be merged in some special cases. We will formulate conditions under which one can general observed variable model rewrite into the observed variables model without the structural errors.

$$\mathbb{Y} - \boldsymbol{\varepsilon}_y = (\mathbb{X} - \boldsymbol{\varepsilon}_x)\boldsymbol{\Gamma}(I - \boldsymbol{\beta})^{-1} + \Xi(I - \boldsymbol{\beta})^{-1} \tag{3.14}$$

$$\mathbb{Y} - \underbrace{\left(\boldsymbol{\varepsilon}_y + \Xi(I - \boldsymbol{\beta})^{-1}\right)}_{\ddot{\varepsilon}_y} = (\mathbb{X} - \boldsymbol{\varepsilon}_x)\boldsymbol{\Gamma}(I - \boldsymbol{\beta})^{-1}. \tag{3.15}$$

Now we want the errors $\ddot{\boldsymbol{\varepsilon}}_y = \boldsymbol{\varepsilon}_y + \Xi(I - \boldsymbol{\beta})^{-1}$ and $\boldsymbol{\varepsilon}_x$ to fulfil the conditions proposed in Theorem 8. Remind notation from Table 1.1 and write $\boldsymbol{\beta}_* = I - \boldsymbol{\beta}$. Now we propose conditions:

$$\mathrm{var}(\boldsymbol{\epsilon}_x) = \sigma^2 I_q \tag{♩}$$

$$\mathrm{var}(\ddot{\boldsymbol{\epsilon}}_y) = \Theta_y + (\boldsymbol{\beta}_*^{-1})^\top \Psi \boldsymbol{\beta}_*^{-1} = \sigma^2 I_p \,. \tag{♫}$$

If there are no errors in the exogenous variables then it is sufficient that the condition (♫) holds for the rewritten model to fulfil the conditions of Theorem 8. More generally we need also (♩) to hold. On the other hand, if there are no errors in endogenous variables, one can still rewrite the model using $\ddot{\boldsymbol{\epsilon}}_y = \Xi(I - \boldsymbol{\beta})^{-1}$ and then the condition (♫) simplifies to $(\boldsymbol{\beta}_*^{-1})^\top \Psi \boldsymbol{\beta}_*^{-1} = \sigma^2 I_p$.

If we omit the coefficient $\boldsymbol{\beta}$ from the model, we get every structural equation in form of errors-in-variables model. In that case the condition (♫) is reduced to $\Theta_y + \Psi = \sigma^2 I$. When also (♩) holds for the same σ^2 then one can say that the conditions coincide with conditions imposed on the error structure of errors-in-variables model in the previous chapter. However the condition is unnecessarily strong when we estimate each structural equation separately. It would be enough to require matrix $\mathrm{var}(\ddot{\boldsymbol{\epsilon}}_y)$ to have σ^2 on the diagonal.

We have shown that using PEIV to estimate coefficients in j-th structural equation may under some rather restrictive assumptions yield consistent estimates. Generalization when some exogenous variables are affected with error was discussed.

It must be noted that the assumptions of Theorem 8 will not be in most cases fulfilled in practice. Some techniques how to address general covariance structure of the errors of EIV model might be found at Section 3.6.2. [23]. However the solution of this problem is highly case specific. Therefore one can not claim that using PEIV to estimate individual structural equations is generally consistent technique.

3.3.3 Modification of two stage least squares

The idea is to modify 2SLS by replacing OLS estimation in both steps by PEIV estimation in suitable way. In this case assume that the data follow observed variable model with errors in variables. We propose several ways how to employ the modifications. We will use following $\approx$ instead of $=$ in the notation like this

$$\mathbb{Y} \approx \mathbb{X}\boldsymbol{\Gamma} + \Xi, \tag{3.16}$$

to write down an error in variable model without specifying which variables are with an error.

Modification 1

We proceed in similar manner as we did in the first equation wise procedure when defining two stage least squares in section 1.4.3. In the first step we estimate following model for:

$$\mathbb{Y}_{(j)} \approx \mathbb{X}_{(j)}\Pi_j + \Delta_j$$

for each j using partial error-in-variable estimator (PEIV), where some chosen (according to the model in hand) variables are considered to be measured with an error. Note fitted values as $\hat{\mathbb{Y}}_{(j)}$. Now as with 2SLS plugin $\mathbb{Y}_{(j)} = \hat{\mathbb{Y}}_{(j)} + \hat{\Delta}_j$ in right hand side of (1.20). This time we will proceed in the second step by estimating

$$\mathbf{Y}_j \approx \hat{\mathbb{Y}}_{(j)}\beta_j + \mathbb{X}_{(j)}\gamma_{(j)} \tag{3.17}$$

using PEIV where $\hat{\mathbb{Y}}_{(j)}$ is considered without any error[2] and some chosen exogenous variables are considered with an error.

[2]See discussion

Modification 2

In this modification we take advantage of one interesting property of EIV estimator. In the first step we consider errors-in-variables model with multivariate response:

$$\mathbb{Y} \approx \mathbb{X}\Pi + \Delta.$$

Denote fitted values by $\hat{\mathbb{Y}}$. Now let us take corresponding columns of fitted values $\hat{\mathbb{Y}}_{(j)}$ and for each j, plug them in right hand side of (1.20) and get models to estimate:

$$\mathbf{Y}_j \approx \hat{\mathbb{Y}}_{(j)}\boldsymbol{\beta}_{(j)} + \mathbb{X}_{(j)}\gamma_{(j)} \tag{3.18}$$

as in Modification 1 use PEIV where $\hat{\mathbb{Y}}_{(j)}$ is considered without an error and some chosen exogenous variables are considered to be with an error.

Modification 3

This modification is very similar to Modification 2 but in the second step we use constrained TLS estimator to estimate model

$$\mathbb{Y} \approx \hat{\mathbb{Y}}\beta + \mathbb{X}\mathbf{\Gamma} \tag{3.19}$$

with constraints according to given model.

Variants of modifications and discussion

All proposed modifications should be applied according to the specific problem in hand because every modification requires a little bit different assumptions about the model. All modifications can be further modified.

Firs straightforward modification arises when we can assume that the exogenous variables are measured without any error. That should simplify our method to 2SLS. For that reason we treat the fitted values of endogenous variables $\hat{\mathbb{Y}}$ in the second step as variables without any error. There is a question if we should always consider them to be without any error. Especially when combining with further modifications.

The fact that, when we address fitted values in EIV model we also have fitted values of the regressors $\hat{\mathbb{X}}$ generally different from $\mathbb{X}$, is another thing to consider. This fact could suggest pluging in also $\hat{\mathbb{X}}$ in second step of the modified procedures.

This section is to be taken as proposition of some ideas for future research. Some simulated examples seemed to be promising however no rigorous simulation study was performed. Author suggests performing fairly large simulation study to narrow set of possible modifications which should be then approached theoretically.

4. Application

4.1 Introduction

We apply some of the methods for estimating structural equations on modelling relations among Relationship satisfaction (RS), Sexual satisfaction (SS) and other psychological and sociological characteristics of women in young adulthood. Brief introduction follows and deeper theoretical background (in Czech) can be found in the original study [11].

We build on top of the previously mentioned study. Taking in account previous result we have hypothesis that Sexual satisfaction is positively associated with Relationship satisfaction.

For Sexual satisfaction they found positive association with Self esteem, Passion, Equality and Sexual intercourse frequency. Negative association with Contraception type hormonal. Other considered variables: Relationship type, Relationship duration, Intimacy, Commitment were not proven to have significant effect on Sexual satisfaction.

For Relationship satisfaction they found positive association with Relationship type Dating, and Relationship type Living together, compared to Open relationship. They also found positive association with Self esteem, Intimacy, Commitment, Passion and Equality. Other considered variables: couple Heterogamy, Sexual frequency, Extraversion, Openness, Conscientiousness and Neuroticism were not proven to have significant effect.

To summarize. We have two endogenous variables SS and RS, we suppose they have an association. For each of them we have set of exogenous variables that we suppose to have an association with corresponding endogenous variable. First we will replicate the results of the previous study with some minor changes. Then we will model this relations as two structural equations.

The goal is to illustrate numerical differences among OLS, two stage least squares and EIV estimator applied to practical example rather than performing new statistical analysis of given problem. As convenient it is to know the true coefficients when estimating an artificially simulated example, a real problem in hand brings different framework to think in.

4.2 Methods

The data were collected using an online survey. We had access to cleaned data that we used in almost the same format as in original study. Only transformation was performed on variables Sexual satisfaction, Relationship Satisfaction, Intimacy, Commitment and Equality to eliminate Skewness. Transformation function was

$$f_t(x) = 4\frac{\exp(x) - \exp(1)}{\exp(5) - \exp(1)} + 1.$$

First of all marginal association of RS and SS was investigated using OLS and EIV estimators. Association was investigated in both ways, changing the role of regressor and response. For completeness we also replicated the models used in the original study using standard linear regression.

As was mentioned before we have two structural equations, one for each exogenous variable. Note Sexual satisfaction as y_s and Relationship satisfaction as y_r. Categorical covariates are parametrized using reference group pseudo contrast. The structural equations are

$$y_s - \epsilon_1 = \beta_s(y_r - \epsilon_2) + \sum_{j \in J_1} \gamma_{1j} x_j + \zeta_1 \tag{4.1}$$

$$y_r - \epsilon_2 = \beta_r(y_s - \epsilon_1) + \sum_{j \in J_2} \gamma_{2j} x_j + \zeta_2 \,, \tag{4.2}$$

where x_j are the exogenous variables (parametrized in case of the categorical variables) and J_i is set of corresponding indexes. We label the coefficients β, γ by the names of corresponding variables when displaying the estimated values.

We directly compare five possible estimators. Both structural equations were estimated using: OLS while omitting the endogenous variable from covariates. This is the same as in the previous study and we denote it by lm_1. Then we use just OLS and denote it by lm_2. Two stage least squares (2SLS) estimator denote by 2SLS. Denote by EIV_1 the PEIV estimator while taking all but categorical variables with an error, and denote by EIV_2 taking only endogenous variables with an error.

Majority of the computations was performed using software R [16]. The function for calculating 2SLS was taken from package [5]. The function for fitting PEIV models is in Appendix A.2. We use this opportunity to point out other R package for fitting structural equations models [17], that was also used for some complementary calculations.

4.3 Results

Descriptive statistics

The data description is of course almost the same as in [11]. The data consist of 290 Czech and Slovakian women in age 16–40 years. Mean age is 23.9 years (s.d. = 3.4). Most women had university education (55.9 %) then 39 % of women had high school education and only 5.1 % had lower than high school education. Other characteristic of the data sample are very deeply described in the original paper. We provide descriptive statistics juts for the variables that were used in our models.

Descriptive statistics of the continuous variables are in Table 4.1 and descriptive statistics of the categorical variables are in Table 4.2. All numerical variables were scaled to interval $[1, 5]$. Most of the variables have mean around 3 ranging from 2.43 (Equality) to 3.74 (Openness) and standard deviation (s.d.) around 0.7 ranging from 0.54 (Openness) to 1.13 (Commitment). The variables are unit-free.

There was 134 women using Hormonal Contraception and also 134 using Nonhormonal Contraception, only 22 women were not using any Contraception. Women in relationship type Dating and Living together were in similar proportion and only 24 women were in Open relationship. Most women (217) had sexual intercourse about 1-3 times a week. Also most frequent relationship duration was 25 months or longer.

Table 4.1: Descriptive statistics for the continuous variables.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	s.d.
Sexual sat.	1.20	2.30	2.79	2.80	3.44	4.51	0.71
Relationship sat.	1.09	2.48	3.33	3.32	4.26	5.00	1.12
Self esteem	1.53	3.13	3.53	3.60	4.03	5.00	0.67
Extraversion	1.58	2.92	3.42	3.38	3.83	4.92	0.64
Openness	1.77	3.46	3.77	3.74	4.08	4.92	0.54
Conscientiousness	1.25	2.92	3.33	3.35	3.92	4.92	0.74
Neuroticism	1.38	2.54	3.15	3.12	3.69	4.54	0.71
Intimacy	1.00	2.27	3.02	3.00	3.84	5.00	1.08
Commitment	1.00	2.13	3.02	2.99	3.84	5.00	1.13
Passion	1.05	2.11	2.65	2.79	3.46	5.00	0.99
Equality	1.01	1.91	2.30	2.43	2.85	5.00	0.79

Table 4.2: Descriptive statistics for the categorical variables.

Relationship type		Contraception		Heterogamy	
Open	24	None	22	Heterogamous	116
Dating	128	Nonhormonal	134	Nonheterogamous	174
Living together	138	Hormonal	134		

Sex frequency		Relationship duration	
Once a month max	31	0-3 months	19
1-3 per week	217	4-24 months	101
4+ per week	42	25+ months	170

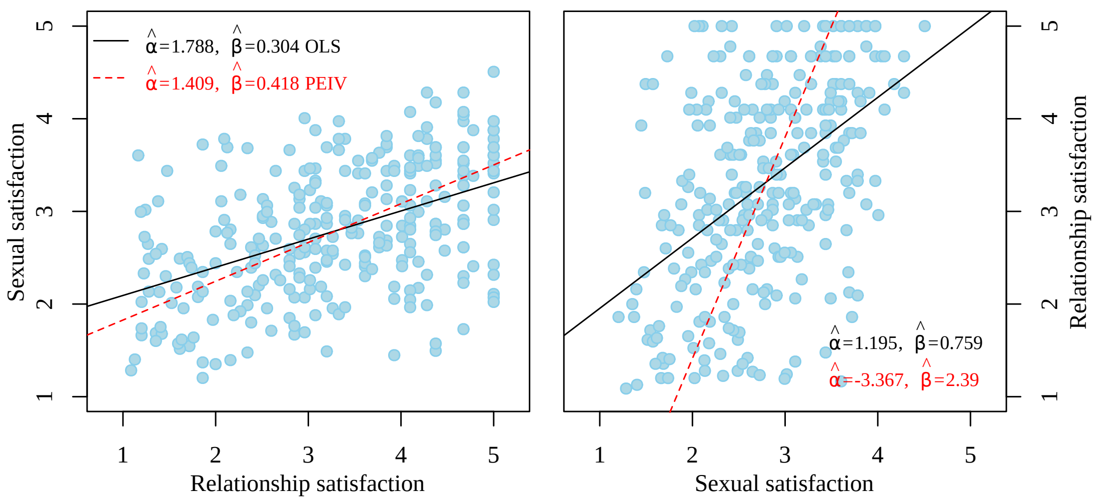

Figure 4.1: Relation between Relationship satisfaction and Sexual satisfaction estimated using OLS and PEIV.

Comparison of estimation methods

First we compare OLS and PEIV estimates of association of Sexual and Relationship satisfaction (see Figure 4.1). We have estimated the coefficients of the models

$$y_s \approx \alpha_s + \beta_s y_r$$
$$y_r \approx \alpha_r + \beta_r y_s$$

using OLS and PEIV. The estimates were fairly similar for model with Sexual satisfaction (y_s) as the response. When Relationship satisfaction (y_r) was taken as the response then the difference was much larger. OLS estimate of β_s was 0.759 and and PEIV was 2.39. Note that for the PEIV estimates of the coefficients β it holds that $\hat{\beta}_s = 1/\hat{\beta}_r$.

Next we review the linear regression models used in the original study. One can see the summary Tables 4.3 and 4.4 of the estimated coefficients. We can confirm the main findings of the original study. The study was focused on the effect of the hormonal contraception method. The whole categorical variable Contraception type was significant (p-val=0.047) predictor of Sexual satisfaction. Results are consistent with original study i.e. women using Nonhormonal Contraception method had sexual satisfaction on average higher that women using Hormonal Contraception.

In the following step we applied mentioned estimators on models where the exogenous variables were the same as in the original study. In the Table 4.5 one can see comparison of five different sets of estimates of the coefficients of the structural equation for Sexual satisfaction. One can see that estimates are not dramatically varying depending on chosen method. It is worth noting that negative estimates of the effect of Relationship satisfaction on Sexual satisfaction yielded by 2SLS and EIV_1 does not make much sense.

The Table 4.6 provides estimates of the coefficients of the structural equation for Relationship satisfaction. In this case the differences among the estimates were much more noticeable. Especially EIV_1 and EIV_2 gave opposite results and both gave rather nonsensible estimates of β_s, namely -4.32 and 12.06.

Finally we left out the variables that were not significant in the original linear regression models. Test on the submodel with omitted covariates was also performed and was not significant. The same comparison of the reduced models as before was performed.

In the Table 4.7 one can see that all estimators but EIV_1 gave very similar and quite sensible results when estimating structural equation for Sexual satisfaction. On the other hand there were still quite contradictory results of EIV_1 and EIV_2 when estimating structural equation for Relationship satisfaction, see Table 4.8.

Because the variables Passion and Equality seemed to have relatively strong association with Sexual satisfaction we tried to remove them from structural equation for Relationship satisfaction to avoid possible multicolinearity problem. This time all methods gave more or less sensible results as one can see in the following Table 4.9.

Table 4.3: Summary table of the model from the original study with Sexual satisfaction as the response.

	γ	s.d.	t-val	p-val
Contraception, ref. is None		whole factor p-val:		0.047
Nonhormonal	*0.00*	*0.12*	*0.03*	*0.977*
Hormonal	*-0.15*	*0.12*	*-1.24*	*0.216*
Relationship type, ref. is Open		whole factor p-val:		0.210
Dating	*-0.11*	*0.13*	*-0.84*	*0.401*
Living together	*0.00*	*0.14*	*0.02*	*0.987*
Relationship duration, ref. is 0-3 months		whole factor p-val:		0.627
4-24 months	*0.11*	*0.13*	*0.89*	*0.376*
25+ months	*0.08*	*0.14*	*0.56*	*0.579*
Sex frequency, ref. is Once a month max		whole factor p-val:		0.141
1-3 per week	*0.20*	*0.10*	*1.92*	*0.055*
4+ per week	*0.23*	*0.13*	*1.74*	*0.083*
Self esteem	0.19	0.05	4.17	0.000
Intimacy	-0.01	0.04	-0.18	0.858
Commitment	0.01	0.04	0.24	0.810
Passion	0.34	0.04	8.68	0.000
Equality	0.25	0.04	6.24	0.000

Table 4.4: Summary table of the model from the original study with Relationship satisfaction as the response.

	γ	s.d.	t-val	p-val
Heterogamy, ref. is Heterogamous		whole factor p-val:		0.108
Nonheterogamous	*0.13*	*0.08*	*1.61*	*0.108*
Relationship type, ref. is Open		whole factor p-val:		0.000
Dating	*0.80*	*0.17*	*4.79*	*0.000*
Living together	*0.80*	*0.18*	*4.42*	*0.000*
Sex frequency, ref. is Once a month max		whole factor p-val:		0.523
1-3 per week	*0.15*	*0.14*	*1.04*	*0.298*
4+ per week	*0.19*	*0.18*	*1.08*	*0.282*
Self esteem	0.23	0.08	2.73	0.007
Extraversion	-0.04	0.07	-0.60	0.546
Openness	0.07	0.08	0.78	0.439
Conscientiousness	0.05	0.06	0.90	0.367
Neuroticism	-0.04	0.08	-0.50	0.620
Intimacy	0.26	0.05	4.94	0.000
Commitment	0.22	0.05	4.56	0.000
Passion	0.28	0.06	4.99	0.000
Equality	0.19	0.05	3.52	0.001

Table 4.5: Comparison of the point estimates of the coefficients in the structural equation for Sexual satisfaction.

	lm_1	lm_2	2SLS	EIV_1	EIV_2
Relationship sat.	–	0.04	-0.10	-0.13	0.08
Contraception, ref. is None					
Nonhormonal	*0.00*	*0.01*	*0.00*	*-0.05*	*0.01*
Hormonal	*-0.15*	*-0.15*	*-0.14*	*-0.20*	*-0.15*
Relationship type, ref. is Open					
Dating	*-0.11*	*-0.14*	*-0.02*	*0.12*	*-0.17*
Living together	*0.00*	*-0.03*	*0.09*	*0.27*	*-0.07*
Relationship duration, ref. is 0-3 months					
4-24 months	*0.11*	*0.11*	*0.13*	*0.14*	*0.11*
25+ months	*0.08*	*0.07*	*0.08*	*0.20*	*0.07*
Sex frequency, ref. is Once a month max					
1-3 per week	*0.20*	*0.19*	*0.21*	*0.09*	*0.19*
4+ per week	*0.23*	*0.22*	*0.24*	*0.01*	*0.21*
Self esteem	0.19	0.18	0.22	0.40	0.17
Intimacy	-0.01	-0.02	0.02	-0.11	-0.03
Commitment	0.01	0.00	0.03	-0.01	-0.01
Passion	0.34	0.33	0.37	0.57	0.32
Equality	0.25	0.24	0.27	0.35	0.24

Table 4.6: Comparison of the point estimates of the coefficients in the structural equation for Relationship satisfaction.

	lm_1	lm_2	2SLS	EIV_1	EIV_2
Sexual sat.	–	0.07	-0.12	-4.32	12.06
Heterogamy, ref. is Heterogamous					
Nonheterogamous	*0.13*	*0.13*	*0.13*	*-0.30*	*0.45*
Relationship type, ref. is Open					
Dating	*0.80*	*0.81*	*0.79*	*1.33*	*1.99*
Living together	*0.80*	*0.80*	*0.80*	*3.91*	*0.48*
Sex frequency, ref. is Once a month max					
1-3 per week	*0.15*	*0.13*	*0.17*	*-0.83*	*-2.29*
4+ per week	*0.19*	*0.17*	*0.22*	*-2.29*	*-2.62*
Self esteem	0.23	0.22	0.24	5.20	-0.95
Extraversion	-0.04	-0.05	-0.03	-4.20	-1.19
Openness	0.07	0.07	0.06	-6.18	0.46
Conscientiousness	0.05	0.05	0.06	0.73	-0.17
Neuroticism	-0.04	-0.03	-0.05	-0.95	0.89
Intimacy	0.26	0.26	0.26	-1.23	0.29
Commitment	0.22	0.22	0.22	-0.10	0.20
Passion	0.28	0.25	0.32	5.03	-3.82
Equality	0.19	0.18	0.22	1.08	-2.80

Table 4.7: Comparison of the point estimates of the coefficients in the reduced structural equation for Sexual satisfaction.

	lm_1	lm_2	2SLS	EIV_1	EIV_2
Relationship sat.	–	0.04	0.03	-0.11	0.05
Contraception, ref. is None					
Nonhormonal	*0.01*	*0.01*	*0.01*	*-0.02*	*0.01*
Hormonal	*-0.15*	*-0.16*	*-0.16*	*-0.17*	*-0.16*
Self esteem	0.21	0.20	0.20	0.39	0.19
Passion	0.36	0.34	0.34	0.49	0.33
Equality	0.27	0.26	0.26	0.35	0.25

Table 4.8: Comparison of the point estimates of the coefficients in the reduced structural equation for Relationship satisfaction.

	lm_1	lm_2	2SLS	EIV_1	EIV_2
Sexual sat.	–	0.07	0.01	-6.44	11.53
Relationship type, ref. is Open					
Dating	*0.88*	*0.89*	*0.88*	*1.11*	*1.81*
Living together	*0.89*	*0.89*	*0.89*	*2.29*	*0.18*
Self esteem	0.27	0.26	0.27	2.67	-1.94
Intimacy	0.27	0.27	0.27	-0.47	0.28
Commitment	0.22	0.22	0.22	0.00	0.23
Passion	0.28	0.26	0.28	3.66	-3.92
Equality	0.21	0.19	0.21	2.28	-2.86

Table 4.9: Comparison of the point estimates of the coefficients in the reduced and modified structural equation for Relationship satisfaction.

	lm_1	lm_2	2SLS	EIV_1	EIV_2
Sexual sat.	–	0.35	0.71	1.39	1.63
Relationship type, ref. is Open					
Dating	*0.75*	*0.84*	*0.93*	*0.99*	*1.16*
Living together	*0.65*	*0.74*	*0.83*	*0.85*	*1.06*
Self esteem	0.29	0.21	0.14	-0.07	-0.05
Intimacy	0.41	0.35	0.28	0.21	0.11
Commitment	0.30	0.26	0.23	0.19	0.13

4.4 Application conclusion

We have compared only the point estimates because we have not developed appropriate theory to compare interval estimates. Comparing only the point estimates could potentially lead to serious misinterpretations. But as the point of this chapter author consider only illustration of some proposed techniques rather than statical modelling of given practical example.

One possibly fatal misspecification of the model is that only Sexual satisfaction and Relationship satisfaction were taken as the endogenous variables. The model might be more complicated. For example there might be more reciprocal paths among the considered variables. One must also note that the variable choice for the reduced models based on their significance in the original linear regression model is questionable in context of structural equations model. Also note that one could also choose different categorizations of categorical variables. One should contact an expert on Psychology and social science and discuss possibly better specification of the model to address given practical problem in more refined manner.

In conclusion the EIV estimator yielded quite sensible estimates of coefficients from the structural equation for Sexual satisfaction. On the other hand for the coefficients from the structural equation for Relationship satisfaction the EIV estimator provided contradicting results based on what other exogenous variables were considered to be with an error.

Conclusion

Fundamentals of structural equations modelling and Error-in-Variables models are summarized in the first two chapters. While summarizing structural equations models we had used the opportunity to shown that the covariance based generalized least squares estimator (GLS) and maximal likelihood estimator (ML) yield the same estimate as standard OLS regression when estimating standard linear model. In the rest of this thesis our focus was tuned towards observed variable model with errors in variables. Then main result are in the third chapter.

General optimization problem to estimate the parameters of observed variable structural equations model with errors-in-variables si postulated. Several possible modifications of two stage least squares are also proposed for future research.

Other proposed method equation-wise EIV is based on using EIV estimator to estimate each structural equation in the model separately. If the model is imposed with some very restrictive conditions then the method yields strongly consistent estimates of the coefficients. The conditions are described in Theorem 8. The theorem can be loosely translated in the way, that if there are not any errors in equations (structural errors) and the errors in the endogenous variables has zero mean, are uncorrelated and have common variance, then equation-wise EIV is consistent. Proposition of some conditions under which one can rewrite more general model to suit the conditions of Theorem 8 follows. However the conditions might be practically difficult to meet in practice.

In the last chapter practical illustration of the equation-wise EIV method is given. The illustrative example investigate Relationship satisfaction, Sexual satisfaction and other psychological and sociological characteristic of women in young adulthood. The results were quite ambiguous.

According to our research the topic of possible usage of Errors-in-Variables estimator for estimating coefficients of structural equations model is barely addressed in the literature. As we have not found many relevant sources we tried to layout some fundamentals to build on in the future research. For that reason we placed more focus on the ideas and their presentation rather than on the technical aspects that are of course no less important. The author feels that some sections (for example Section 3.3.1) needs more attention and further development. On the other hand the author hopes that this thesis will be thought provoking and one could hopefully view that as the main contribution of this work.

Bibliography

[1] Anděl, J. *Základy matematické statistiky*, 3. ed. Matfyzpress, Praha, 2011.

[2] Basmann, R. L. A generalized classical method of linear estimation of coefficients in a structural equation. *Econometrica 25*, 1 (1957), 77–83.

[3] Bollen, K. A. *Structural Equations with Latent Variables.* John Wiley & Sons, Inc., New York, 1989.

[4] Fox, J. *Linear Statistical Models and Related Methods: With Applications to Social Research.* Wiley series in probability and mathematical statistics: Applied probability and statistics. Wiley, 1984.

[5] Fox, J., Nie, Z., and Byrnes, J. *sem: Structural Equation Models*, 2017. R package version 3.1-9.

[6] Gleser, L. J. Estimation in a multivariate "errors in variables" regression model: Large sample results. *The Annals of Statistics 9*, 1 (1981), 24–44.

[7] Graham, J. M. The general linear model as structural equation modeling. *Journal of Educational and Behavioral Statistics 33*, 4 (2008), 485–506.

[8] Jöreskog, K., Sörbom, D., and SPSS Inc. *LISREL 8: User's Reference Guide.* Scientific Software International, 1996.

[9] Jöreskog, K. G. A general method for analysis of covariance structures. *Biometrika 57*, 2 (1970), 239–251.

[10] Komárek, A. NMSA407 linear regression course notes. `http://msekce.karlin.mff.cuni.cz/~komarek/vyuka/2017_18/nmsa407/2017-NMSA407-notes.pdf#page=100`. Accessed: 2018-03-12, Last updated: 2017-12-20.

[11] Kovaříková, E. Souvislosti užívání antikoncepce a vztahové a sexuální spokojenosti. Master thesis, Masarykova univerzita, Filozofická fakulta, Brno, 2017.

[12] Lee, S.-Y., Song, X.-Y., and Tang, N.-S. Bayesian methods for analyzing structural equation models with covariates, interaction, and quadratic latent variables. *Structural Equation Modeling: A Multidisciplinary Journal 14*, 3 (2007), 404–434.

[13] Lu, I. R., Kwan, E., Thomas, D. R., and Cedzynski, M. Two new methods for estimating structural equation models: An illustration and a comparison with two established methods. *International Journal of Research in Marketing 28*, 3 (2011), 258 – 268.

[14] Pešta, M. Unitarily invariant errors-in-variables estimation. *Statistical Papers 57*, 4 (Dec 2016), 1041–1057.

[15] Pešta, M. *Modern Asymptotic Perspectives on Errors-in-variables Modeling.* Doctoral thesis, Charles University, Faculty of Mathematics and Physics (MFF), Prague, 2011.

[16] R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2018.

[17] Rosseel, Y. lavaan: An R package for structural equation modeling. *Journal of Statistical Software 48*, 2 (2012), 1–36.

[18] SAS Institute Inc. Illustration of model identification: Spleen data. `http://documentation.sas.com/?docsetId=statug&docsetTarget=statug_introcalis_sect010.htm&docsetVersion=14.3&locale=en`, 2017. Accessed: 2018-05-01, Last updated: 2017-12-01.

[19] Schaffrin, B. A note on constrained total least-squares estimation. *Linear Algebra and its Applications 417*, 1 (2006), 245 – 258. Special Issue devoted to papers presented at the Workshop on Matrices and Statistics.

[20] Schaffrin, B., and Felus, Y. A. On the multivariate total least-squares approach to empirical coordinate transformations. three algorithms. *Journal of Geodesy 82*, 6 (Jun 2008), 373–383.

[21] Schaffrin, B., and Felus, Y. A. An algorithmic approach to the total least-squares problem with linear and quadratic constraints. *Studia Geophysica et Geodaetica 53*, 1 (Jan 2009), 1–16.

[22] Siemsen, E., and Bollen, K. A. Least absolute deviation estimation in structural equation modeling. *Sociological Methods & Research 36*, 2 (2007), 227–265.

[23] Van Huffel, S., and Vandewalle, J. *The Total Least Squares Problem: Computational Aspects and Analysis.* Frontiers in Applied Mathematics. Society for Industrial and Applied Mathematics (SIAM, 3600 Market Street, Floor 6, Philadelphia, PA 19104), 1991.

[24] Zellner, A., and Theil, H. Three-stage least squares: Simultaneous estimation of simultaneous equations. *Econometrica 30*, 1 (1962), 54–78.

[25] Zhang, S., and Zhang, K. On a basic multivariate eiv model with linear equality constraints. *Applied Mathematics and Computation 236* (2014), 247 – 252.

List of Figures

1.1 Path diagram of the model (1.1). 4

2.1 Comparison of OLS, DLS, EIV, PEIV, estimators. The true values of the parameters are $\alpha = 4, \beta = 1.3$. 14

3.1 Path diagram of model (3.1) 19
3.2 Path diagram of the model (3.9). 21

4.1 Relation between Relationship satisfaction and Sexual satisfaction estimated using OLS and PEIV. 27

List of Tables

1.1 Description of the variables, coefficients and their properties. . . . 5

4.1 Descriptive statistics for the continuous variables. 27
4.2 Descriptive statistics for the categorical variables. 27
4.3 Summary table of the model from the original study with Sexual satisfaction as the response. 29
4.4 Summary table of the model from the original study with Relationship satisfaction as the response. 29
4.5 Comparison of the point estimates of the coefficients in the structural equation for Sexual satisfaction. 30
4.6 Comparison of the point estimates of the coefficients in the structural equation for Relationship satisfaction. 30
4.7 Comparison of the point estimates of the coefficients in the reduced structural equation for Sexual satisfaction. 31
4.8 Comparison of the point estimates of the coefficients in the reduced structural equation for Relationship satisfaction. 31
4.9 Comparison of the point estimates of the coefficients in the reduced and modified structural equation for Relationship satisfaction. . . 31

A. Appendix

A.1 Additional theorems and definitions

Note that Frobenius norm is $\|A\|_F = \sqrt{\sum_{i=1}^m \sum_{j=1}^n |a_{ij}|^2} = \sqrt{\operatorname{tr}\{AA^\top\}}$.

Definition 9 (Unitary matrix)**.** *A square matrix A is called unitary*

$$A^\top A = AA^\top = I$$

Definition 10 (Unitary invariant matrix norm)**.** *A matrix norm $\|\cdot\|$ is unitary invariant if*

$$\|UAV\| = \|A\|$$

for all $A \in \mathbb{R}^{n\times p}$ and all unitary matrices U, V.

Theorem 9 (Break down of SS_T)**.** *Let $\mathbf{Y} = \mathbb{X}\boldsymbol{\beta} + \sigma^2 I_N$ and let $\mathbf{1}_N \in \mathcal{M}(\mathbb{X})$ then*

$$\underbrace{\sum_{i=1}^N (Y_i - \bar{Y})^2}_{SS_T} = \underbrace{\sum_{i=1}^N (Y_i - \hat{Y}_i)^2}_{SS_e} + \underbrace{\sum_{i=1}^N (\hat{Y}_i - \bar{Y})^2}_{SS_R}.$$

Proof. See [10] Theorem 5.3. □

Theorem 10 (Strong consistency of OLS)**.** *Let following hold*

LS1 The data $[\mathbb{X}, \mathbb{Y}]$ consist of iid realizations of generic random vector $(\mathbf{X}, \mathbf{Y})$. And there exist $\boldsymbol{\beta}$ unknown parameters such that $Y = \mathsf{E}[\mathbf{Y}|\mathbf{X}] = \mathbf{X}^\top \beta$.

LS2 $\mathsf{E}\,|X_j, X_l| < \infty$ and $\mathsf{E}[\mathbf{X}\mathbf{X}^\top]$ is positive definite matrix.

LS3 $\mathsf{E}\,(var[\mathbf{Y}|\mathbf{X}]X_j X_l) < \infty$.

Then

$$\lim_{n\to\infty} \hat{\boldsymbol{\beta}}_n = \boldsymbol{\beta} \text{ a.s.}$$

Proof. See [10] Theorem 14.2. □

A.2 Code used for computing PEIV estimator

```
PEIV<-function(formula,fixed=.~1, data, ex=1, values=TRUE){
 m<-model.matrix(formula,data)
 namesorder<-colnames(m)
 noerror<-colnames(model.matrix(update(formula,fixed),data))
 Y<-as.matrix( model.response(model.frame(formula,data)))
 if(is.null(noerror)){
  X<-m
  M<-cbind(X,Y)
  rm(m)
 }else
 {
  X<-m[,!(colnames(m) %in% noerror),drop=FALSE]
  W<-m[,noerror, drop=FALSE]
  rm(m)
  M<-cbind(lm.fit(W,X)$residuals,lm.fit(W,Y)$residuals)
 }
 SVD<-svd(M)
 V<-SVD$v
 mm<-dim(X)[2]
 dd<-dim(Y)[2]
 vv<-mm+dd
 SG1<-diag(SVD$d)[1:mm,1:mm]
 U11<-SVD$u[,1:mm, drop=FALSE]

 V11<-V[1:mm,1:mm, drop=FALSE ]
 V12<-V[1:mm,(mm+1):vv ,drop=FALSE ]
 V21<-V[(mm+1):vv,1:mm ,drop=FALSE ]
 V22<-V[(mm+1):vv,(mm+1):vv, drop=FALSE ]

 beta<- -V12 %*% solve(V22)
 rownames(beta)<-colnames(X)

 alpha<-NULL
 if(!is.null(noerror)){alpha<- as.matrix(lm.fit(W,Y-  X%*%beta)$coeff) }
  coeff<-rbind(alpha, beta)
 if(values){ AY<-U11%*%SG1%*% cbind(t(V11),t(V21) )
  colnames(AY)<-colnames(M)
  if(ex[1]==0){
   RET<-list(coeff= coeff, Xhat = AY[,(1:mm),drop=FALSE],
   Yhat= AY[,(1:mm), drop=FALSE]%*%coeff)
  }
  else{
   RET<-list(coeff= coeff[namesorder,],
   Xhat = cbind(W, lm.fit(W,X)$fitted+ AY[,(1:mm), drop=FALSE] ),
   Yhat= cbind(W, lm.fit(W,X)$fitted+AY[,(1:mm), drop=FALSE])%*%coeff)}}
 else{ RET<-list(coeff= coeff[namesorder,], Xhat = NULL, Yhat=NULL) }
return( RET )
}
```