University of Southern Queensland

Faculty of Health, Engineering and Sciences

Statistical Methodology for Regression Model with Measurement Error

A thesis submitted by

Anwar A. Mohamad Saqr

B.Sci., M.Sci.

in fulfilment of the requirements for the degree of

Doctor of Philosophy

Submitted: May, 2013

Abstract

This thesis primarily deals with the estimation of the slope parameter of the simple linear regression model in the presence of measurement errors (ME) or error-in-variables in both the explanatory and response variables. It is a very old and difficult problem which has been considered by a host of authors since the third quarter of the nineteenth century. The ME poses a serious problem in fitting the regression line, as it directly impacts on estimators and their standard error (see eg Fuller, 2006, p. 3). The standard linear regression methods, including the least squares or maximum likelihood, work when the explanatory variable is measured without error. But in practice, there are many situations where the variables can only be measured with ME. For example, data on the medical variables such as blood pressure and blood chemistries, agricultural variables such as soil nitrogen and rainfall etc can hardly be measured accurately. The apparent observed data represents the *manifest variable* which measures the actual unobservable *latent variable*

with ME.

The ME model is divided into two general classifications, (i) functional model if the explanatory (ξ) is a unknown constant, and (ii) structural model if ξ is independent and identically distributed random variable (cf Kendall, 1950, 1952). The most important characteristic of the normal structural model is that the parameters are not identifiable without prior information about the error variances as the ratio of error variances (λ) (see Cheng and Van Nees, 1999, p. 6). However, the non-normal structural model is identifiable without any prior information. The normal and non-normal structural models with ME in both response and explanatory variables are considered in this research.

There are a number of commonly used methods to estimate the slope parameter of the ME model. None of these methods solves the estimation problem in varying situations. A summary of the well known methods is provided in Table 1.

The first two chapters of this thesis cover an introduction to the ME problem, background, and motivation of the study. From Chapter 3 we provide a new methodology to fit the regression line using the *reflection* of the explanatory variable about the fitted regression line with the manifest variables. The asymptotic consistency and the mean absolute error (MAE) criteria are used

Table 1: A summary of commonly used methods to handle the ME model problem

Methods	Model	Assumption	Criticism
Instrumental variable (IV)	Normal and non normal	High correlation with ξ. No correlation with ME	Difficult to fond valid IV
Maximum likelihood (Orthogonal regression)	Normal	λ known True points fall on a straight line	Misspecification λ. Large sample required
Fourth moments	Non normal	Model not close to normal. Large sample size	Difficult to satisfy these assumptions
Three moments	Non normal	Model not close to normal. Large sample size	Difficult to satisfy these assumptions
Grouping	Normal and non normal	Groups are independent of ME	Less efficiency
Geometric Mean	Normal and non normal	$\lambda = \beta_1^2$	Unrealistic assumption, too restrictive sensitive to error variances

to compare the new estimators and the relevant existing estimators under different conditions.

One of the most commonly used methods to deal with the ME model is the instrumental variable (IV) method. But it is difficult to find valid IV that is highly correlated to the explanatory but uncorrelated with the error term. Therefore, in Chapter 4 we propose a new method to find a good IV based on the reflection of explanatory variable. The new method is easy to implement, and performs much better than the existing methods. The superiority of this method is demonstrated both analytically and via numerical as well as graphical illustrations under certain assumptions.

In Chapter 5, a commonly used method to deal with the normal structural model, namely the orthogonal regression (OR) (which is the same the maximum likelihood solution when $\lambda = 1$) method under the assumption of known λ is discussed. But the OR method does not work well (inconsistent) if λ is misspecified and/or the sample size is small. We provide an alternative method based on the reflection method (RM) of estimation for measurement error model. The RM uses a new transformed explanatory variable which is derived from the reflection formula. This method is equivalent or asymptotically equivalent to the orthogonal regression method, and nearly asymptotically unbiased and efficient under the assumption that λ is equal to one and the sample size is large. If λ is misspecified the RM method is better than the OR method under the MAE criterion even if the sample size is small.

Chapter 6 considers the Wald method (two grouping method) which is still widely used, in spite of increasing criticism on the efficiency of the estimator. To address this problem, we introduce a new grouping method based on the reflection grouping (RG) approach. The proposed method provides new grouping process to modify Wald method in order to increase its efficiency. The RG method introduces a new way of dividing the data using the rank of the reflection of the explanatory variable. The method recommends different grouping criteria depending on the value of λ to be one or more/less than one. The RG method significantly increases the efficiency of Wald method, and it is more precise than the other competing methods and works well for different sample sizes and for different values of λ. Moreover, the RG method also removes the shortcomings of the maximum likelihood method when λ is misspecified and sample size is small.

The geometric mean (GM) regression is covered in Chapter 7. The GM method is widely used in many disciplines including medical, pharmacology, astrometry, oceanography, and fisheries researches etc. This method is known by many names such as reduced major axis, standardized major axis, line of organic correlation etc. We introduce a new estimator of the slope parameter when both variables are subject to ME. The weighted geometric mean (WGM) estimator is constructed based on the reflection and the mathematical relationship between the vertical and orthogonal distances of the observed points and the regression line of the manifest model. The WGM estimator possesses better statistical properties than the geometric mean estimator, and OLS-bisector estimator. The WGM estimator is stable and work well for different values of λ and for different sample sizes.

The properties of the proposed reflection estimators are investigated in Chapters 3-7. Also, these estimators are compared with the relevant existing estimators by simulation studies. The computer package Matlab is used for all computations and preparation of graphs. Based on the asymptotic consistency and MAE criteria the proposed reflection estimators perform better than the existing estimators, in some cases, even the standard assumption on λ and sample size are violated.

Chapter 8 provides some concluding summaries remarks.

Certification of Dissertation

I certify that the ideas, designs and experimental work, results, analyses and conclusions set out in this dissertation are entirely my own effort, except where otherwise indicated and acknowledged.

I further certify that the work is original and has not been previously submitted for assessment in any other course or institution, except where specifically stated.

Signature of Candidate________________________________

Signature of Principal Supervisor ________________________

Signature of Associate Supervisor ________________________

Acknowledgments

My first and foremost thanks to ALLAH for the opportunities that He has given to me throughout my life, especially those that have brought me to the position of finishing this thesis. I would like to express my thankfulness and gratitude to my principal supervisor Professor Shahjahan Khan for his invaluable assistance, support, patience and guidance during the period of my research, without his knowledge and assistance this study would not have been successful. My special thanks and gratitude go to my associate supervisor Dr Trevor Langlands for his advice, support, constructive feedback and invaluable assistance. I would like to thank all the staff in the faculty of health, engineering and sciences specially the staff of the school of agricultural, computational and environmental sciences and the library for providing a very good scientific environment for statistical research.

This thesis is dedicated to the souls of my mother and father (may ALLAH bless them with Jannah), that I wish them to be alive to see what I have achieved and to share my happiness for completing this thesis, who always supported, encouraged and directed me for higher education.

Anwar A. Mohamad Saqr

Contents

List of Figures

List of Tables

List of Notations

ξ_j	Unobserved explanatory variable (*latent variable*).
η_j	Unobserved response variable (*latent variable*).
x_j	Observed explanatory variable (*manifest variable*).
y_j	Observed response variable (*manifest variable*).
x_j^*	Reflection of the observed explanatory variable.
y_j^*	Reflection of the observed response variable (*manifest variable*).
δ_j	Measurement error in the explanatory variable.
ϵ_j	Measurement error in the response variable.
e_j	Equation error in the true model.
v_j	Equation error in the Measurement Error model.
ψ	Reflection angle about the unfitted regression line (*by manifest variables*).
θ	Reflection angle about the fitted regression line (*by latent variables*).

Chapter 1

Introduction

1.1 Introduction

Regression analysis forms an important part of the statistical tools for investigating the relationships between variables. For example, regression analysis may be used to investigate whether there is a relationship between the number of road accidents and the age of the driver. Linear regression is a common statistical data analysis technique in the fields of medical, agricultural, chemical, physical and economic studies (Gillard and Iles, 2009; Warton et al. 2006). The regression model may be used to predict body weight based on body fat, or the yield of a crop based on soil moisture or rainfall levels. However, measuring the explanatory variable, namely the body fat or soil moisture level, is likely to involve measurement error (Anderson 1984). The ordinary least squares (OLS) estimator of the regression parameters is inappropriate in the presence of measurement error (cf Fuller, 2006, p. 3). As a result, in real life, measurement error causes a serious problem as it directly impacts on estimators and their standard error. It is well known that the measurement error in the response variable is not as serious as it is in the explanatory variable. The error in the response variable can be absorbed in the error term of the model; however, the error in the explanatory variable causes various problems, and needs to be handled appropriately (Madansky 1959).

The measurement error (ME) or error-in-variables is a real problem and it has been considered by a host of authors since the late nineteenth century (Gillard, 2010). Adcock (1877, 1878) discussed the problem in the context of least squares method. Pearson (1901) suggested some estimators based on Adcock's work. The problem has been seriously considered by researchers from the last century. Wald (1940), Bartlet (1949), Durbin (1954), and Riggs et al. (1978), considered fitting the regression line when both variables are subject to error. Berkson (1950) noted that the error in the explanatory variable leads to bias in the estimated parameters of the regression line, regardless of the data being a random sample or the population. Burr (1988) considered error in the explanatory variable for the binary responses model.

Freedman et al. (2004) suggested a moment method to deal with error in the explanatory variable. The problem of error in both explanatory and response variables was considered by Madansky (1959) and Halperin (1961).

Degracie and Fuller (1972) considered estimation of the slope and covariance when the variable is measured with error. Grubbs (1973) discussed error of measurement, precision and the statistical inference. Aigner (1973) considered regression with a binary variables subject to the error of observation. Florens et al. (1974) considered Bayesian inference in error-in-variables models. Schneeweiss (1976) proposed consistent estimation of a regression with error in the variables. Bhargava (1977) introduced maximum likelihood estimation in a multivariate error-in-variables regression model with an unknown error covariance matrix. Garber and Klepper (1980) extended the classical normal error-in-variables model. Prentice (1982) dealt with covariant measurement error and parameters estimation.

Amemiya et al. (1984) proposed estimation of the multivariate error-in-variables model with estimated error covariance matrix. Klepper and Leamer (1984) provided consistent sets of estimates for regression with error in all variables. Stefanski and Carroll (1985) discussed covariant measurement error in logistic regression. Carroll et al. (1985) proposed comparison of least squares and measurement error model with randomized analysis of covariance. Armstrong (1985) dealt with the measurement error in the generalized linear model. Bekker (1986) provided comments on the identification issues in the measurement error model. Schafer (1986) combined information on the measurement error model. Carroll and Ruppert (1996) discussed the use and misuse of orthogonal regression in the measurement error model. Fuller (2006) covered various aspects of the measurement error model and related inferences. Carroll et al. (2006) summarized much of what is known about the consequences of measurement error for estimating the linear regression parameters. Recently McCartin (2010) has introduced a new concept of oblique linear least squares approximation. This thesis introduces a new methodology of fitting a straight regression line when both response and explanatory variables are subject to error. This has not been discussed previously in the literature of measurement error.

It is well known that the fitting of a straight line to bivariate data (ξ, η) is a common procedure and widely used in analysis of linear relationships. This procedure works under the standard linear regression theory where the explanatory variable is measured without error. The response variable η depends on the explanatory variable ξ according to the usual additive model

$$\eta_j = \beta_0 + \beta_1 \xi_j + e_j, \qquad j = 1, 2, \cdots, n, \tag{1.1}$$

where e_j is a random error representing the intrinsic scatter in η about the regression line, and (β_0, β_1) are the regression parameters. It is often assumed that the mean of the error term e_j is zero with a non-zero variance.

The main goal here is to estimate the parameters β_0 and β_1 of the model (1.1). One of the common techniques to estimate these parameters involves minimising the function of the random error term e_j. This technique, called the least squares theory, suggests minimising the sum of the squared error components, and was introduced by Carl Freidrich Gauss (1777-1855) and Adrien Marie Legendre (1752-1833). Here the regression line of η on ξ is obtained by minimising the sum of squares of the vertical distances from the points (ξ_j, η_j) to the regression line which is given by the estimated equation model $\hat{\eta}_j = \hat{\beta}_0 + \hat{\beta}_1 \xi_j$. This is given by

$$\sum_{j=1}^{n} e_j^2 = \sum_{j=1}^{n} (\eta_j - \beta_0 - \beta_1 \xi_j)^2,$$

where the least squares estimators of the parameters β_0 and β_1 can be obtained by differentiating $\sum_{j=1}^{n} e_j^2$ with respect to each of the parameters, and solving the equations which arise after setting the derivatives to zero to find

$$\begin{aligned} \hat{\beta}_1 &= \frac{S_{\eta\xi}}{S_\xi^2} \\ \hat{\beta}_0 &= \bar{\eta} - \hat{\beta}_1 \bar{\xi}, \text{ where} \end{aligned}$$

$$\begin{aligned} S_{\eta\xi} &= \frac{1}{n-1} \sum_{j=1}^{n} (\xi_j - \bar{\xi})(\eta_j - \bar{\eta}), \\ S_\xi^2 &= \frac{1}{n-1} \sum_{j=1}^{n} (\xi_j - \bar{\xi})^2, \\ S_\eta^2 &= \frac{1}{n-1} \sum_{j=1}^{n} (\eta_j - \bar{\eta})^2, \end{aligned}$$

where $\bar{\eta}$ and $\bar{\xi}$ are the sample means of the variables η and ξ respectively.

Note it is easy to show how to obtain of $\hat{\beta}_0$ and $\hat{\beta}_1$ by minimising the sum of squares $\sum_{j=1}^{n} e_j^2$ (see for example Johnston, 1971).

It is well known that the procedure of η on ξ regression requires some assumptions one of them being that the error is only present in the response variable η, while the explanatory variable ξ is measured without error. However, in some situations, it may be possible that there are errors in both variables. Indeed, as real data is seldom observed directly, and the common problem known as the errors in variables or measurement error model arises (Gillard, 2010). Casella and Berger (1990) pointed out that the measurement error model

“ is so fundamentally different from the simple linear regression $\cdots$ that it is probably best thought of as a different topic.”

This type of the measurement error model usually occurs when both the explanatory variable ξ and the response variable η are experimentally measured (Gillard and Iles, 2009). In fact, errors in variables causes the least squares estimator of the slope in η on ξ regression to be biased (Fuller, 2006, p. 3). The random measurement error artificially inflates the dispersion of observations of the independent variable ξ and biases least squares estimators. This thesis describes circumstances where simple linear regression models are significantly incorrect, when there are measurement errors in both the explanatory variable ξ and the response variables η.

1.2 The measurement error problem

This study deals with the commonly known problem of measurement error (ME) or error-in-variables (see for example Warton et al. 2006; Carroll et al. 2006). This problem occurs when variables are measured or observed with random error. The measurement error model could be linear or nonlinear, where at least one of the variables explanatory ξ_j or response η_j is measured with error. There are two different types of measurement error. The first is called *the classical additive error model*, and occurs when the observed variable is an unbiased measure of the true variable. The second is the *the error calibration model* where the observed variable is a biased measure of the unobserved variable (Carriquiry, 2001).

In general, measurement error potentially affects all statistical analysis, because it affects the probability distribution of the data (Chesher, 1991). To deal with the measurement error problem we should first distinguish and identify the variables of the model. Let ξ_j be the true explanatory variable which is unobserved and is called the latent variable. This unobserved variable does not include any measurement error. Let x_j be the observed explanatory variable which is called the manifest variable which is observed with measurement error. Similarly let η_j be the true response variable without any measurement error, and y_j be the observed response variable which includes random measurement error. Let δ_j be the measurement error in the observed explanatory variable, $\delta_j = x_j - \xi_j$, and ϵ_j be the measurement error in the observed response variable, $\epsilon_j = y_j - \eta_j$. When there is no measurement error in the variables then it is usually assumed that both response η_j and explanatory ξ_j variables are related by

$$\eta_j = \beta_0 + \beta_1 \xi_j, \tag{1.2}$$

where β_0 is the intercept, β_1 is the slope parameter, and $\xi_j' s$ are fixed in repeated sampling $j = 1, 2, ..., n$. Note that the model above is called standard measurement error model if it is not included the equation error (error term).

It is often assumed that the measurement error in the response variable ϵ_j is normally distributed $\epsilon_j \sim N(0, \sigma_\epsilon^2)$, and $E(\xi_j \epsilon_j) = 0$. When there is no measurement error, the ordinary least squares (OLS) estimator of the slope parameter β_1 for the model (1.2) is

$$\hat{\beta}_{1\xi} = \frac{\sum_{j=1}^{n}(\xi_j - \bar{\xi})(\eta_j - \bar{\eta})}{\sum_{j=1}^{n}(\xi_j - \bar{\xi})^2}.$$

This estimator is unbiased for β_1 and has the smallest variance among all unbiased linear estimators. This estimator $\hat{\beta}_1$ is the maximum likelihood estimator of β_1, if $\xi \sim N(\mu_\xi, \sigma_\xi^2)$ and $Cov(\xi, \epsilon) = 0$ (cf Fuller, 2006, p. 2). The theory of classical linear regression analysis assumes that the explanatory variable, ξ_j, is measured without error. In practice this assumption is often violated, particularly in social science, biological assay, and in economic data (Warton et al. 2006). Since the explanatory variable being measured with error, the ordinary least squares method is unable to produce unbiased estimators of parameters of the measurement error model.

However, when only the response variable includes measurement error, $y_j = \eta_j + \epsilon_j$, then the estimator is unbiased. This can be seen by replacing η_j to y_j in the model (1.2) as follows

$$y_j \quad = \quad \beta_0 + \beta_1 \xi_j + \epsilon_j. \tag{1.3}$$

The only negative consequence of the measurement error in the response variable is that it inflates the standard errors of the estimator of the regression coefficient (cf Chen, et al. 2007).

On the other hand, when the explanatory variable has measurement error the estimator becomes biased and inconsistent. This can be seen by rewriting (1.3) by using x_j instead of ξ_j, where $\xi_j = x_j - \delta_j$, as follows

$$y_j \quad = \quad \beta_0 + \beta_1 x_j + (\epsilon_j - \beta_1 \delta_j) = \beta_0 + \beta_1 x_j + v_j, \tag{1.4}$$

where $v_j = (\epsilon_j - \beta_1 u_j) \sim N(0, \sigma_v^2)$, and $E(x_j v_j) \neq 0$. Here x_j and δ_j are not independent, since

$$Cov(x_j, v_j) = Cov(x_j, \epsilon_j) - \beta_1 Cov(x_j, \delta_j) = -\beta_1 \sigma_\delta^2 \neq 0.$$

For the model (1.4), the least squares estimator of y_j on x_j is given by

$$\hat{\beta}_{1x} = \frac{\sum_{j=1}^{n}(x_j - \bar{x})(y_j - \bar{y})}{\sum_{j=1}^{n}(x_j - \bar{x})^2}.$$

The probability limit of $\hat{\beta}_{1x}$ is given by

$$plim\hat{\beta}_{1x} = \beta_1 + \frac{Cov(x_j, v_j)}{Var(x_j)} = \beta_1 - \frac{\beta_1 \sigma_\delta^2}{\sigma_\xi^2 + \sigma_\delta^2} = \beta_1 \frac{\sigma_\xi^2}{\sigma_\xi^2 + \sigma_\delta^2}.$$

Hence $\hat{\beta}_{1x}$ is a biased and inconsistent estimator for β_1. Obviously, when the explanatory variable as well as the response variable are subject to measurement error, the regression situation becomes considerably more complicated (Draper and Smith, 1981, p. 124).

1.3 Outline of the Thesis

In this thesis, attention is concentrated on introducing a new methodology for estimating the slope in a simple linear regression when both explanatory variable, ξ, and response variable, η, are measured with error. It is well known that the model fitting and parameter estimation of an measurement error model is notably different to fitting a simple linear regression model without measurement error.

Chapter 2 describes different methods that have been used to tackle the problem of error in both variables. Some of these solutions work under various assumptions about the underlying model and sampling plan to avoid the identifiability problem. One of these methods, known as the variance ratio method, is based on the assumed knowledge of the relative magnitude of measurement error in the response variable η and explanatory variable ξ. In fact, these assumptions are suggested to make the parameters of the normal structural model to be identifiable. In the literature, there are six assumptions required as extra information about the variances of errors, to make the normal structural model identifiable (Cheng and Van Ness, 1999, p. 6). In measurement error models it turns out that the method of maximum likelihood is only satisfactory when all random variables in the model ξ, ϵ and δ are normally distributed.

Chapter 3 provides a new methodology constructed based on the reflection technique and the regression line of the measurement error model, and introduces the proof of the following propositions not previously discussed before:

Proposition 1 *The squares of the unexplained variation of y by x can be partitioned in to the vertical and horizontal components as follows:*

$$(y_j - \hat{y}_j)^2 = (y_j^* - \hat{y}_j)^2 + (x_j^* - x_j)^2 \quad , \quad j = 1, 2, \cdots, n.$$

Then it can be shown that the sum of squares error can be written as

$$SSE_{yx} = \sum_{j=1}^{n}(y_j - \hat{y}_j)^2 = \sum_{j=1}^{n}(y_j^* - \hat{y}_j)^2 + \sum_{j=1}^{n}(x_j^* - x_j)^2 = SSE_y + SSE_x,$$

where x^* and y^* are transformed variables of the manifest variables x and y respectively (see equations (3.1) and (3.2)), SSE_y is the vertical unexplained variation in y, and SSE_x is the horizontal unexplained variation as a function of x.

Proposition 2 *The average of the manifest explanatory variable $\bar{x}$ equals that of the latent variable $\bar{\xi}$ and the reflection of manifest variable $\bar{x}^*$, that is*

$$\bar{x}^* = \bar{x} = \bar{\xi}.$$

Proposition 3 *The average of the manifest response variable $\bar{y}$ equals that of the latent variable $\bar{\eta}$ and the reflection of manifest variable $\bar{y}^*$, that is*

$$\bar{y}^* = \bar{y} = \bar{\eta}.$$

Proposition 4 *The estimator of the regression parameters of y on x equals the estimator of the regression parameters of y^* on x.*

$$\hat{\beta}_{1yx} = \hat{\beta}_{1y^*x}, \text{ and } \hat{\beta}_{0yx} = \hat{\beta}_{0y^*x},$$

where $\hat{\beta}_{1yx}$ is the slope estimator of ordinary least squares of y on x, and $\hat{\beta}_{1y^*x}$ is that of y^* on x. Also $\hat{\beta}_{0yx}$ is the intercept of ordinary least squares of y on x and $\hat{\beta}_{0y^*x}$ is that of y^* on x.

Proposition 5 *The sample variance of the response variable y is greater than that of its reflection y^*.*

$$S_y^2 = \frac{1}{n-1}\sum_{j=1}^{n}(y_j - \bar{y})^2 > S_{y^*}^2 = \frac{1}{n-1}\sum_{j=1}^{n}(y_j^* - \bar{y}^*)^2$$

Proposition 6 *The sample covariance of the manifest explanatory variable x and its reflection x^* is equal the sample variance of manifest explanatory* *variable x.*

$$\hat{cov}(x, x^*) = S_x^2,$$

where S_x^2 is the sample variance of x.

Proposition 7 *The sample covariance of the response variable y and the reflection variable x^* is greater than that of the response variable y and the manifest variable x.*

$$S_{yx^*} > S_{yx},$$

Proposition 8 *The sample variance of explanatory variable x is less than that of its reflection variable x^*.*

$$S_x^2 \leq S_{x^*}^2$$

Proposition 9 *The difference between the sum of squares of the reflection variable $S_{x^*}^2$ and the sum of squares of the manifest explanatory variable S_x^2 is given by*

$$SS_{x^*} - SS_x = SST_y - SSR_{yx} - SSE_y,$$

where SST is sum of squares total of y, SSR_{yx} is the explained variation of y by x.

Chapter 4 proposes a new instrumental variable to estimate the parameters of a simple linear regression model where the explanatory variable is subject to measurement error. The new instrumental variable is defined using reflection of the observed values of the explanatory variable. Like other instrumental variable estimators, it is unbiased and consistent, but over performs estimators proposed by Wald (1940), Bartlett (1949), and Durbin (1954) if the ratio of the error variances is equal to or less than one. The method is straightforward, easy to implement, and performs much better than the existing instrumental variable based estimators. The theoretical superiority of the proposed estimator over the existing instrumental variable based estimators is established by analytical results of simulation. Two illustrative examples for numerical comparisons of the results are also included.

Chapter 5 proposes an estimation method based on the reflection of the explanatory (*manifest*) variable to estimate the parameters of a simple linear regression model when both the response and the explanatory variables are subject to measurement error (ME). The *reflection method* (RM) uses all observed data points, and does not exclude or ignore part of the data or replace them by ranks. The RM is straightforward, and easy to implement. We show that the RM is equivalent or asymptotically equivalent to the orthogonal regression method. Simulation studies show that the RM produces estimators that are nearly asymptotically unbiased and efficient under the assumption that the ratio of the error variances $\lambda = \sigma_\epsilon^2 \sigma_\delta^{-2} = 1$. Moreover, it allows us to define the sum of squares error uniquely, the same way as in the case of no measurement error. The numerical comparisons of the results are also included.

Chapter 6 introduces a new slope estimator for regression model when both variables are subject to measurement errors and the model includes equation error. The main aim of the proposed method is to improve the efficiency of Wald's estimator under flexible assumption on the ratio of error variances (λ). It is well known that in the presence of equation error in regression models any estimator based on assumed knowledge of (λ) is biased. Although Wald's method could deal with models that include equation error, it lacks efficiency and is subject to identifiability problem. To compare the relative efficiency of the proposed estimator with the OLS, Wald's and Geary's estimators, simulation studies under various assumptions are undertaken. Moreover, a comparison of the new estimator with the method of moments estimator when λ is biased due to the presence of the equation error is included.

Chapter 7 introduces a new estimator to fit the regression line when both variables are subject to measurement errors and there is no prior information known about the variances of error. The proposed weighted reduced major axis (WG) is derived based on the mathematical relationship between the vertical and orthogonal distances of the observed points and the regression line. The geometric mean (GM) regression method is widely used in many disciplines as a solution to errors in variables model, although it lacks efficiency. To evaluate the geometric mean GM estimator method, this Chapter provides an alternative view on GM estimator. The common belief, which is not quite true, is that this method minimizes the vertical and horizontal distances between observed points and the best-fit straight line. We compare the performance of the proposed WG estimator with the GM and OLS-bisector estimators, and the sensitivity to the variation of the ratio of error variances (λ). The final chapter summarises the contents of this thesis, and indicate some further work in this area.

Chapter 2

Historical background of measurement error models

2.1 Introduction

Currently, there is a huge literature on measurement error (ME) models (Fuller, 2006; Carroll et al. 2006; Cheng and Van Ness, 1999; Gillard, 2010). The literature of ME has become widespread in diverse fields such as economics, medical science, agriculture, chemistry, physics, astronomy and particularly in epidemiology. Measurement error can introduce serious bias into the estimation of regression parameters and can strongly affect the statistical power of studies (Freudenheim and Marshall, 1988). This Chapter provides some of the common estimation techniques to deal with measurement error models, and discusses some interconnections between these methods. In fact, it is difficult to discuss the whole wealth of literature on measurement error models in this thesis, but the focus has been placed upon a few key developments and methods. Unfortunately the notation set of the measurement error models has not been standardised in the literature, so it will be carefully introduced at the beginning of the thesis. The measurement error problem is also known as error-in-variables or model II regression (cf Sokal and Rohlf, 1995, p. 541).

2.2 Major Axis Regression (Orthogonal)

The problem of fitting a simple linear regression model when both variables are subject to error was first considered by Adcock (1877), where he introduced the major axis regression (MAR) technique which is also known as orthogonal regression (OR). However, this method is equivalent to the bivariate case of principal components analysis (PCA) (Mohler et al. 1978). Geometrical exposition of this method is to minimise the squared perpendicular distances from the data points to the fitted regression line. The estimator of the true slope of the simple linear regression model $y = \beta_0 + \beta_1 x$, by this technique is given by

$$\hat{\beta}_{1OR} = \frac{(S_y^2 - S_x^2) + \sqrt{(S_y^2 - S_x^2)^2 + 4S_{yx}^2}}{2S_{yx}},$$

where S_y^2 is the sample variance of the manifest response variable y, S_x^2 is the sample variance of the manifest explanatory variable x and S_{yx} is the sample covariance of y and x.

An alternative form of this estimator is

$$\hat{\beta}_{1MA} = 0.5\left[(\hat{\beta}_2 - \hat{\beta}_1^{-1}) + sgn\{S_{yx}\}\sqrt{4 + (\hat{\beta}_2 - \hat{\beta}_1^{-1})^2}\ \right],$$

where $\hat{\beta}_1 = \dfrac{S_{yx}}{S_x^2}$, and $\hat{\beta}_2 = \dfrac{S_y^2}{S_{yx}}$.

Adcock dealt with a special case of the problem of estimating β_1 in the standard simple linear regression model, where there is no equation error. This case assumes that the variances of measurement error in both variables are equal, that is, $\sigma_\epsilon^2 = \sigma_\delta^2$, where ϵ_j is the measurement error in the manifest response variable y_j, $(y_j = \eta_j + \epsilon_j)$, and δ_j is the measurement error in the manifest explanatory variable x_j, $(x_j = \xi_j + \delta_j)$. Adcock defined the line of the best fit through the data as the line which minimises the sum of squares of the orthogonal distances from the observed points to the fitted line. Whereas the least squares method defines the line of the best fit which minimises the sum of squares vertical distances (residuals) as

$$\sum_{j=1}^{n}(y_j - \hat{\beta}_0 - \hat{\beta}_1\xi_j)^2.$$

Adcock mentioned that the best regression line should pass through the mean of the n points, where n is the sample size.

Kummell (1879) extended the work of Adcock, where he assumed that the ratio of error variances $\lambda = \dfrac{\sigma_\epsilon^2}{\sigma_\delta^2}$ was known instead of taking equal error variances, $\sigma_\epsilon^2 = \sigma_\delta^2$. He justified this assumption since it is realistic that most experienced practitioners will have sufficient knowledge about the spread of the measurement errors. Pearson (1901) suggested a very similar estimator to that proposed by Adcock (cf Fuller, 2006, p. 30). He showed that the fitted regression line of this method always lies between the regression line of ξ on y and that of y on ξ. In addition, this technique does not depend on which variable is treated as response variable and which is explanatory variable (cf Amman and Van Ness, 1988). Isobe et al. (1990) pointed out that the major axis regression is appropriate only for scale free variables, such as ratios of observable variables or logarithmical transformed variables.

2.3 Deming Regression Technique

The Deming regression technique is one of the most widely known techniques for fitting simple linear regression model when there are errors-in-variables (EIV). It is also known as the functional maximum likelihood estimator under the assumption that the ratio of error variances $\lambda = \dfrac{\sigma_\epsilon^2}{\sigma_\delta^2}$ is known (Gillard, 2010). It takes measurement errors for both variables into account, therefore it is more generally applicable than major axis regression technique (Linnet, 1998). In his book, Deming (1943) suggested this technique to minimise the common error simultaneously to obtain the best line that fits the data. The slope estimator of this technique is given by

$$\hat{\beta}_{1DEM} = \frac{(S_y^2 - \lambda S_x^2) + \sqrt{(S_y^2 - \lambda S_x^2)^2 + 4\lambda S_{yx}^2}}{2S_{yx}}.$$

Note that the slope estimator of Deming regression technique $\hat{\beta}_{1DEM}$ becomes the slope estimator of the orthogonal regression technique $\hat{\beta}_{1OR}$ when $\sigma_\epsilon^2 = \sigma_\delta^2$, $(\lambda = 1)$ (cf Gillard and Iles, 2009).

Fuller (2006, p. 30) stated the above estimator was first derived by Kummel (1879) for a general λ, but he did not formulate the model in precisely the same manner. In clinical chemistry literature this technique is attributed to Deming (1943) (cf Linnet, 1998). It is also called *orthogonal regression* in statistical literature. In his book, Fuller (2006, p. 30) called it a method of moments estimator (MOM), although it differs from the commonly used method of moments estimator (cf Carroll and Ruppert, 1996).

Linnet (1998) pointed out the value of λ describes the angle in which to project points onto the line to minimise the sum of squares deviations. The distance between the observed and predicted response values, with this angle, is aimed to minimise the error term. Deming regression solution is the major axis regression solution when the ratio of error variances is equal to one, $\lambda = 1$. The major axis regression method is a special case of Deming regression method when the variance of error terms are equal $\sigma_{\epsilon}^2 = \sigma_{\delta}^2$.

The fundamental problem of using the Deming regression method arises when the value of λ is not known with certainty (see Dunn, 2004). Carroll et al. (1995) and Carroll and Ruppert (1996) question if the assumption of the value of λ is at all correct, and they concluded that λ is frequently incorrect. An inappropriate λ leads to biased estimates of parameter β_1 and this is why these critics claim that many if not all examples of Deming regression are flawed. The inclusion of the equation error in linear regression is a common practice but this technique does not take it into account (see Carroll and Ruppert, 1996).

2.4 Grouping Method

Wald (1940) proposed an estimation method based on the grouping of the data. It divides the observations on both response and explanatory variables into two groups, G1 and G2, where G1 contains the first half of the ordered observations and G2 contains the second half. The grouping is made based on the explanatory manifest variable x_j. Wald showed that the slope of the line joining the group means provided consistent estimator for the slope parameter of the simple linear regression model. Properties of this estimator can be found in Gupta and Amanullah (1970).

To explain the grouping method, suppose that the variables ξ_j and η_j are related by the equation $\eta_j = \beta_0 + \beta_1 \xi_j$ with both variables are subject to random measurement error. Let $x_j = \xi_j + \delta_j$ and $y_j = \eta_j + \epsilon_j$, where x_j and y_j represent the manifest explanatory variable and the manifest response variable respectively. The measurement error in the manifest explanatory variable x_j is δ_j and in the manifest response variable y_j is ϵ_j (Gillard, 2010). It is well known, for both large and small sample cases, that the presence of measurement error in the explanatory variable makes the ordinary least squares (OLS) estimator inconsistent and biased (see Barnett, 1969). Furthermore, it makes the maximum likelihood estimator unacceptable (see Kendall and Stuart, 1961, p. 383). In 1940 Wald pointed out that a consistent estimator of β_1 may be calculated if the following assumptions are met:

1. The random variables $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ have the same distribution and they are uncorrelated, that is, $E(\epsilon_i \epsilon_j) = 0$ for $i \neq j$, and the variance of ϵ_j, $\sigma_\epsilon^2 = E(\epsilon_i \epsilon_j)$, for $i = j$, is finite.

2. The random variables $\delta_1, \delta_2, \ldots, \delta_n$ have the same distribution and they are uncorrelated, that is, $E(\delta_i \delta_j) = 0$ for $i \neq j$, and the variance of δ_j,

$\sigma_\delta^2 = E(\delta_i \delta_j)$, for $i = j$, is finite.

3. The random variables ϵ_j and δ_j are uncorrelated, that is, $E(\epsilon_j \delta_j) = 0$, for all j.

4. $\dfrac{\sum_{j=k+1}^{n} x_j - \sum_{j=1}^{k} x_j}{n} > 0$ or $\bar{x}_{k+1} > \bar{x}_k$, where $\bar{x}_{k+1}$ is the mean of the group G2, $\bar{x}_k$ is the mean of the group G1, n is even ($n = 2, 4, 6, \ldots, \infty$), and $k = \frac{n}{2}$. In other words, we can be sure that as $n \to \infty$, b_1 does not approach zero (cf Madansky, 1959).

The observations are then divided into two groups based on the ranks of the manifest explanatory variable x_j, those above the median of x_j into one group, G_1 and those below the median into another group, G_2. Then Wald's estimator of β_1 and β_0 are given by

$$\hat{\beta}_{1W} = \frac{a_1}{b_1} = \frac{(y_1 + \ldots + y_k) - (y_{k+1} + \ldots + y_n)}{(x_1 + \ldots + x_k) - (x_{k+1} + \ldots + x_n)} = \frac{\bar{y}_2 - \bar{y}_1}{\bar{x}_2 - \bar{x}_1},$$

where $(\bar{x}_1, \bar{y}_1)$ are the means of (x_j, y_j) into group $G1$, for $j = 1, 2, \cdots, k$, and $(\bar{x}_2, \bar{y}_2)$ are the means of (x_j, y_j) into group $G2$, for $j = k+1, k+2, \cdots, n$.

Then $\hat{\beta}_{0W} = \bar{y} - \hat{\beta}_1 \bar{x}$,

where $\bar{y} = \dfrac{\sum_{j=1}^{n} y_j}{n}$, $\bar{x} = \dfrac{\sum_{j=1}^{n} x_j}{n}$, and

$$\begin{aligned} a_1 &= \frac{(x_1 + \ldots + x_k) - (x_{k+1} + \ldots + x_n)}{n}, \text{ and} \\ b_1 &= \frac{(y_1 + \ldots + y_k) - (y_{k+1} + \ldots + y_n)}{n}. \end{aligned}$$

The Wald's technique was further developed by Bartlett (1949). It has been suggested as a simple method to handle the problem of measurement error when both variables are subject to imprecision, and no knowledge of the error of measurement is available. Instead of dividing the ordered observations into two groups, he proposed that greater efficiency would be obtained by dividing it into three groups, $G1$, $G2$ and $G3$. $G1$ and $G3$ are the outer groups, and G2 is the middle group.

Lindley (1947) proposed another grouping technique based on four groups. It requires the calculation of two slope estimates, where the first estimator uses the first and third quarters as the two groups and the second estimator makes use of the second and fourth groups. The proposed estimator of this technique is given by the mean of these two slope estimates.

Generally these grouping methods are designed to counter the problem of inconsistency. However, the groups are not independent of the error terms if they are not based on the order of the true values. But Wald proved that the grouping by the observed values is the same as grouping with respect to the true values. There are some criticisms in the literature about Wald estimator but these lack consensus. Neyman and Scott (1951) pointed out that the Wald estimator is consistent for β_1 in the structural relation situation if and only if

$$Pr[x_{p_1} - \epsilon < \xi \leq x_{p_1} - \mu] = Pr[x_{1-p_2} - \epsilon < \xi < x_{p_1} - \mu] = 0,$$

where x_{p_1} and x_{1-p_2} are the p_1 and $(1-p_2)$ percentile points of the distribution function of x, and $\epsilon - \mu$ is the range of δ. This condition means that we must know the range of the error in x, and in order to satisfy the condition the range should be finite, otherwise the condition becomes $Pr[-\infty < \xi < \infty] = 0$ which is never satisfied. Madansky (1959) pointed out that this condition relies on the central limit theorem and assumes that δ is normally distributed. But this has an infinite range, and so the above condition remains unsatisfied when the errors δ_j are normally distributed (cf Madansky, 1959).

Wald's estimator is consistent under very general conditions except where the errors are not normally distributed (cf Gupta and Amanullah, 1970). Pakes (1982) claimed that the work of Gupta and Amanullah (1970) is needless given that Wald's estimator is inconsistent. However, according to Theil (1956), Wald's method is valuable though there is a loss of efficiency. Johnston (1972, p. 284) stated "Under fairly general conditions the Wald estimator is consistent but likely to have a large sampling variance". Moreover, Fuller (2006, p. 74) mentioned that the Wald's method was often interpreted improperly.

In fact, there are many discussions on improving the efficiency of the grouping method by dividing the observations to more than two groups and to groups of unequal size (see Nair and Banerjee, 1942; Bartlett, 1949; Dorff and Gurland, 1961; Ware, 1972). In practice, the grouping method is still important and the grouping estimator is the maximum likelihood estimator under the normality assumption (see Chang and Huang, 1997; Cheng and Van Ness, 1999, p. 130).

2.5 Reduced Major Axis

One of the simplest approaches to handle the error in variables is the geometric mean (GM) functional relationship, initially proposed by Teissier (1948) and later by Barker et al. (1988), and Draper and Yang (1997). This estimator has frequently been mentioned in the literature for two cases. First is when there is no basis for distinguishing between the response and explanatory variables, and the second is to handle the errors-in-variables when the additional information is not available. The geometric mean functional relationship is widely used in fisheries studies. It has received much attention, and has been suggested that it is more useful than ordinary least squares (OLS) estimator for comparing the lean body proportions (Sprent and Dolby, 1980).

This approach defines the estimator as the geometric mean of the slope of y

on x regression line, and the reciprocal of the slope of x on y regression line where x and y both are random (see Leng et al. 2007).

The GM estimator of the slope is given as

$$\hat{\beta}_{1G} = sgn(SP_{xy}) \sqrt{\frac{SS_y}{SS_x}} = sgn(Sp_{xy}) \left(\frac{S_y}{S_x} \right) ,$$

where $SS_x = \sum_{j=1}^{n}(x_j - \bar{x})^2$, $SS_y = \sum_{j=1}^{n}(y_j - \bar{y})^2$,

$SP_{yx} = \sum_{j=1}^{n}(x_j - \bar{x})(y_j - \bar{y})$, and S_y and S_x are the standard deviation of y and x respectively.

In the literature of biology and allometry the geometric mean method is known as the standardized major axis (MA) regression (Warton et al. 2006). It is also known as reduced major axis (RMA), or the line of organic correlation (see Tessier, 1948; Kermack and Haldane, 1950; Ricker, 1973). Moreover, in physics it is known as a type of standard weighting model (see Machonald and Thompson, 1992), while the astronomers call it as Strömberg's Impartial Line (Feigelson and Babu, 1992).

A host of recent publications indicate that using the GM or RMA is necessary and sufficient to fit the straight line when the response and explanatory variables are both subject to error (cf Levinton and Allen, 2005; Zimmerman et al. 2005; Sladek et al. 2006; Vincent and Lailvaux, 2006). While Jolicoeur (1975) and Sprent and Dolby (1980) pointed out that the GMFR estimator is unbiased if and only if

$$\lambda = \frac{\sigma_{\epsilon}^2}{\sigma_{\delta}^2} = \frac{\sigma_y^2}{\sigma_x^2}.$$

But several studies indicate that this assumption is unrealistic (cf Sprent and Dolby, 1988).

There is a common recommendation to use GM estimator but is often employed without mentioning the reason of using it (cf Smith, 2009). Jolicoeur (1975) stated that is difficult to interpret the meaning of the slope estimated by the geometric mean method. However, the common perception is that the geometric mean method seeks to minimise the vertical and horizontal distances between the observed points and the fitted line (Halfon, 1985; Draper and Yang, 1997). But this is not quite true, given that the GM minimises the orthogonal distance of the observed points (x_j, y_j) from the unfitted line rather than the fitted line $(\hat{\eta}_j = \hat{\beta}_{0\xi} + \hat{\beta}_{1\xi}\xi_j)$.

2.6 Moments estimators

2.6.1 Estimators based on the first and second moments

The main problem in fitting the measurement error model using the method of moments is that of identifiability. Therefore, this method is based on the assumption that some prior knowledge about the error variances is available. Under this assumption the method of moments equations can easily be solved. Otherwise, it can be seen from equations (2.1-2.6) below that a unique solution cannot be found for the parameters since there are five equations with six unknown parameters (Gillard, 2010). The expressions for population moments are

$$\begin{aligned} E[x] &= E[\xi] = \mu, \text{ and} \\ E[y] &= E[\eta] = \beta_0 + \beta_1 \xi. \end{aligned}$$

The variances and covariance of the manifest variables are

$$\begin{aligned} var(x) &= \sigma_\xi^2 + \sigma_\delta^2, \\ var(y) &= \beta_1^2 \sigma_\xi^2 + \sigma_\epsilon^2, \text{ and} \end{aligned}$$

$$cov(y, x) = \beta_1 \sigma_\xi^2.$$

The estimating equations of the method of moments are found by equating the above population moments to their sample equivalents as follows

$$\begin{aligned}
\bar{x} &= \hat{\mu}_x = \frac{1}{n}\sum_{j=1}^{n} x_j, &(2.1)\\
\bar{y} &= \hat{\beta}_0 + \hat{\beta}_1\hat{\mu}_x, &(2.2)\\
S_x^2 &= \hat{\sigma}_\xi^2 + \hat{\sigma}_\delta^2, &(2.3)\\
S_y^2 &= \hat{\beta}_1^2\hat{\sigma}_\xi^2 + \hat{\sigma}_\epsilon^2, &(2.4)\\
S_{yx} &= \hat{\beta}_1\hat{\sigma}_\xi^2. &(2.5)
\end{aligned}$$

Van Montfort (1989) introduced the hyperbolic relationship between the method of moments estimator for σ_ϵ^2 and σ_δ^2 which is called the Frisch hyperbola, given as

$$(S_x^2 - \hat{\sigma}_\delta^2)(S_y^2 - \hat{\sigma}_\epsilon^2) = (S_{yx})^2. \qquad (2.6)$$

This equation relates pairs of estimates $(\hat{\sigma}_\delta^2, \hat{\sigma}_\epsilon^2)$, and it is also a useful equation to derive another pairs of parameters such as

$$S_y^2 = \hat{\beta}_1 S_{yx} + \hat{\sigma}_\epsilon^2.$$

However, to use the first and second moment estimating equations we should specify which assumptions of the parameter space is likely to suit the purpose in order to avoid the identifiability problem. Kendall and Stuart (1973), Hood et al. (1999) and Dunn (2004) described these assumptions in the context of the method of moments as follows.

The first assumption is that the intercept β_0 is known. An estimator for the slope parameter β_1 can be derived by using equations (2.1) and (2.2), and the estimator of the slope β_1 is

$$\hat{\beta}_1 = \frac{\bar{y} - \beta_0}{\bar{x}}.$$

Dunn (2004) considered this assumption when $\beta_0 = 0$, and he noted that this assumption is extremely unsafe. It is clear that the problem occurs with this estimator when $\bar{x} \approx 0$. Therefore there are specific admissibility conditions for the estimator based on assumption that the intercept β_0 is known which are

$$\begin{aligned} \bar{x} \neq 0&, \\ S_x^2 &> \sigma_\xi^2, \\ S_y^2 &> \frac{\bar{y} - \beta_0}{\bar{x}} S_{yx}. \end{aligned}$$

In fact, the assumption that the intercept β_0 is known does not make the normal model of more than one explanatory variable identifiable (cf Cheng and Van Ness, 1999, p. 6).

The second assumption is that the ratio of error variances $\lambda = \sigma_\epsilon^2 \sigma_\delta^{-2}$ is known, and that $\sigma_\xi^2 > 0$. Then equations (2.3), (2.4) and (2.5) yield the following quadratic equation in $\hat{\beta}_1$

$$\hat{\beta}_1^2 S_{yx} - \hat{\beta}_1 (S_y^2 - \lambda S_x^2) - \lambda S_{yx} = 0.$$

The positive root of this equation is the maximum likelihood solution which can be expressed as

$$\hat{\beta}_1 = \frac{(S_y^2 - \lambda S_x^2) + \sqrt{(S_y^2 - \lambda S_x^2)^2 + 4\lambda S_{yx}^2}}{2S_{yx}}. \tag{2.7}$$

There are some equivalent forms of (2.7) such as

$$\hat{\beta}_1 = \frac{2\lambda S_{yx}}{(S_y^2 - \lambda S_x^2) + \sqrt{(S_y^2 - \lambda S_x^2)^2 + 4\lambda S_{yx}^2}},$$

and

$$\hat{\beta}_1 = \varphi(\lambda) + sgn\{S_{yx}\}(\varphi^2(\lambda) + \lambda)^{\frac{1}{2}},$$

where $\varphi(\lambda) = \frac{S_y^2 - \lambda S_x^2}{2S_{yx}}$.

Note all these forms are equivalent, and this solution is the same as that in Deming regression (cf Cheng and Van Ness, 1999, p. 17).

Riggs et al. (1978) recommended the use of this solution based on their results of simulation studies, but they emphasized the importance of having a reliable prior knowledge of the ratio of error variances, λ. Edland (1996) pointed out that slope estimator of linear measurement error models based on assumed knowledge of the ratio of error variances is biased if the underlying linear relationship is anything other than a completely deterministic, law-like relationship. Lakshminarayanan and Gunst (1984) discussed the performance of estimator when the ratio of error variances (λ) is incorrectly specified.

The third assumption is the reliability ratio $\kappa = \sigma_\xi^2 \sigma_x^{-2}$ is known. Then it is possible to obtain an unbiased estimator of the slope parameter β_1. In fact, for some specific disciplines information about reliability ratio κ is available, particularly in psychology, and sociology literature. For example, studies of community loyalty, social consciousness, willingness to adopt new practices, managerial ability (cf Fuller, 2006, p. 5).

The slope estimator of y on x is biased when there is measurement error in x, and the magnitude of this bias is the reliability ratio κ. The estimator based on the assumption that the reliability ratio is known is considered as a correction of the bias of the slope estimator for y on x regression. It is well known that

$$\begin{aligned} plim\hat{\beta}_{1x} &= \beta_1 + \frac{Cov(x_j, v_j)}{Var(x_j)} \\ &= \beta_1 - \frac{\beta_1 \sigma_\delta^2}{\sigma_\xi^2 + \sigma_\delta^2} \\ &= \beta_1 \frac{\sigma_\xi^2}{\sigma_\xi^2 + \sigma_\delta^2} = \beta_1 \frac{\sigma_\xi^2}{\sigma_x^2} = \beta_1 \kappa. \end{aligned}$$

Then if the reliability ratio κ is known, the unbiased estimator becomes

$$\hat{\beta}_1 = \hat{\beta}_{1x} \kappa^{-1}.$$

This estimator could be obtained from the first and second moment equations by dividing equation (2.5) by equation (2.3) (cf Gillard, 2010). A more general reliability definition was introduced by Gleser (1992). The reliability ratio κ is called attenuation factor and its inverse is called the linear correction for attenuation.

The fourth assumption is that the error variance σ_ϵ^2 is known. Equation (2.4) and (2.5) immediately give

$$\hat{\beta}_1 = \frac{(S_y^2 - \sigma_\epsilon^2)}{S_{yx}}.$$

The estimator based on the known error variance is a modification of the reciprocal of the slope of the x on y regression. This modification is to subtract the known error variance σ_ϵ^2 from S_y^2 in the numerator of the estimator (cf Cheng and Van Ness, 1999, p. 18).

The fifth assumption is that the error variance σ_δ^2 is known. Equations (2.3) and (2.5) are used to obtain an estimator for the slope parameter β_1. The equation (2.3) can be written in terms of σ_ξ^2, if the error variance σ_δ^2 is known, as follows

$$\hat{\beta}_1 = \frac{S_{yx}}{S_x^2 - \sigma_\delta^2}.$$

In fact, there are five restrictions that should be satisfied to obtain the maximum likelihood estimator. These five restrictions are

$$
\begin{aligned}
S_x^2 &\geq \frac{S_{yx}}{\hat{\beta}_1}, && (2.8)\\
S_y^2 &\geq \hat{\beta}_1 S_{yx}, && (2.9)\\
S_x^2 &\geq \hat{\sigma}_\delta^2, && (2.10)\\
S_y^2 &\geq \hat{\sigma}_\epsilon^2, && (2.11)\\
sgn(S_{yx}) &= sgn(\hat{\beta}_1). && (2.12)
\end{aligned}
$$

If any one or all these conditions are not satisfied then the maximum likelihood solution is

$$\hat{\beta}_1 = \frac{S_y^2}{S_{yx}}.$$

This estimator is just the reciprocal of the slope estimator of the inverse regression of y on x (Cheng and Ness, 1999, p. 18).

The sixth assumption is that both variances σ_δ^2 and σ_ϵ^2 are known. In this case, any four of the moment equations (2.1) to (2.5) can be used to derive unique estimator. Based on this assumption the possible solutions of estimating equations (2.1) to (2.5) are

1. If both error variances σ_δ^2, and σ_ϵ^2 are known, then the ratio of the error variances is also known. This yields

$$\hat{\beta}_1 = \frac{(S_y^2 - \lambda S_x^2) + \sqrt{(S_y^2 - \lambda S_x^2)^2 + 4\lambda S_{yx}^2}}{2S_{yx}}.$$

2. Substituting (2.4) into equation (2.5) yields the same estimator as when σ^2_ϵ is known

$$\hat{\beta}_1 \quad = \quad \frac{(S^2_y - \sigma^2_\epsilon)}{S_{yx}}.$$

3. Another estimator for the slope parameter β_1 is obtained by rearranging equation (2.4) in terms of $\beta^2_1\sigma^2_\xi$ and dividing by equation (2.3) which yields

$$\hat{\beta}_1 \quad = \quad sgn\{S_{yx}\}\sqrt{\frac{S^2_y - \sigma^2_\epsilon}{S^2_x - \sigma^2_x}}.$$

4. Substituting (2.3) into equation (2.5) yields the same estimator as when σ^2_δ is known:

$$\hat{\beta}_1 \quad = \quad \frac{S_{yx}}{S^2_x - \sigma^2_\delta}.$$

All of the estimators outlined above are obtained by restricting the parameter space. If a restriction is unsatisfied, then the method of moment equations are not useful. This is a problem due to having six unknown parameters, but only five moment estimating equations. However, the conditions basically suggests that the fitted regression line lies between the OLS line of y on x and OLS line of x on y. Otherwise, there may be negative estimates for some or all of the variances in the model (Gillard, 2010).

2.6.2 The method of higher-order moments

One of the most widely known techniques for slope estimator of the simple linear regression model with measurement error is the method of higher-order moments. Many statistical books have referred to this method and describe that the method of higher-order moments corresponding sample moments to parameter estimates such as Casella and Berger (1990). The method of higher-order moments has a long history, yet it is still an effective tool, because it is easily implemented (see Bowman and Shenton, 1988). Commonly, many statistical texts give greater attention to the method of higher-order moments. The estimators of the method of higher-order moments are not uniquely defined and it is necessary to choose amongst possible estimates to find the best estimator to data. This may lead to the cases where the method is used in measurement error models.

The first person to recognize the potential of population moments as a basis of estimation is Karl Pearson (1890). He introduced the method of moments (MM) estimation in a series of his papers published after 1890s. This method has two fundamental features:

1. It is based upon the empirical distribution that approximates the true distribution when the sample size is large enough. Thus the MM estimator relies on asymptotic theory to justify its usefulness.

2. It does not need to specify any sort of distribution and does not use any information about the population distribution other than its moments.

Indeed, using the method of moments to handle the measurement error problem requires that some information regarding a parameter must be assumed to be known, or more estimating equations have to be derived by the higher moments (Cheng and Van Ness, 1999). Scott (1950) introduced an estimator based on the third moments for the structural model, and showed that if the third central moment of ξ exists and is non-zero, then the equation is given by

$$f_{n,1}(\hat{\beta}_{1\xi}) \quad = \quad \frac{\sum_{j=1}^{n}[(y_j - \bar{y}) - \hat{\beta}_{1\xi}(\xi_j - \bar{\xi})]^3}{n} = 0,$$

where, he mentioned that, $\hat{\beta}_{1\xi}$ is a consistent estimator of β_1. This is because

$$\lim_{n\to\infty} f_{n,1}(\hat{\beta}_{1\xi}) \quad = \quad (\beta_1 - \hat{\beta}_{1\xi})^3 \mu_{\xi}^3,$$

where μ_{ξ}^3 denotes the third central moment of ξ. He showed that the estimator of the slope is a function of the third order sample moments. Scott (1950) mentioned that estimators based on the lower order moments may be more accurate than those based on higher order moments. He introduced the estimator without a method of extracting the root which would provide the consistent estimator (Gillard, 2010).

Drion (1951), Pal (1980), Van Montfort et al. (1987), Van Montfort (1989) and Cragg (1997) used the higher order moment estimating equations, and discussed some large sample properties. Drion (1951) mentioned that an estimator could be derived through the third-order non-central moment equations for a functional model. Moreover, he introduced the variances of all the sample moments, and showed that his estimator of the slope is consistent, and the sample moments are given by

$$M_{rs}(x,y) = \frac{\sum_{j=1}^{n}(x_j-\bar{x})^r(y_j-\bar{y})^s}{n}, \text{ and}$$
$$M'_{rs}(x,y) = \frac{\sum_{j=1}^{n} x_j^r y_j^s}{n},$$

where r and s are order of moments, and $\bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_j$, and $\bar{y} = \frac{1}{n}\sum_{j=1}^{n} y_j$.

Pal (1980) and Van Montfort et al. (1987) introduced a treatment for the structural relationship model under the assumption that the latent variable ξ is not normally distributed and the moments exist. It is also assumed that the latent variable ξ_j, the measurement error in the response variable ϵ_j, and the measurement error in the explanatory variable δ_j are independent of one another. The equation error (q) is allowed in this approach by absorbing the equation error term q_j into the measurement error of the response variable as $e_j = q_j + \epsilon_j$. The non-normal structural model could be written as

$$y_j = \beta_0 + \beta_1\xi_j + (q_j + \epsilon_j) = \beta_0 + \beta_1\xi_j + e_j.$$

Drion (1950) considered the following five equations based upon the second-order moments:

1. $M'_1(x) = \mu'_1(\xi)$,

2. $M_1'(y) = \beta_0 + \beta_1 \mu_1'(\xi)$,

3. $M_2'(x) = \mu_2'(\xi) + \sigma_\delta^2$,

4. $M_2'(y) = \beta_0^2 + 2\beta_0\beta_1\mu_1'(\xi) + \beta_1^2\mu_2'(\xi) + \sigma_\epsilon^2$,

5. $M_{11}'(x, y) = \beta_0\mu_1'(\xi) + \beta_1\mu_2'(\xi)$.

Similarly the equations which are based on the third-order moments are

1. $M_3(x) = \mu_3(\xi)$,

2. $M_3(y) = \beta_1^3\mu_3(\xi)$,

3. $M_{21}(x, y) = \beta_1\mu_3(\xi)$,

4. $M_{12}(x, y) = \beta_1^2\mu_3(\xi)$.

Drion (1950) introduced an estimator of the slope β_1 given by

$$\hat{\beta}_1 = \pm \left(\frac{M_3(y)}{M_3(x)} \right)^{\frac{1}{3}}.$$

This estimator is consistent under the mild condition that

$$\lim_{n \to \infty} M_3(x) \neq 0.$$

There are estimators for each pair of the last four equations. There are other five choices for estimating the slope parameter β_1:

1. $\hat{\beta}_2 = \dfrac{M_{03}}{M_{12}}$,

2. $\hat{\beta}_3 = \dfrac{M_{12}}{M_{21}}$,

3. $\hat{\beta}_4 = \dfrac{M_{21}}{M_{30}}$,

4. $\hat{\beta}_5 = \pm\sqrt{\dfrac{M_{03}}{M_{21}}}$,

5. $\hat{\beta}_6 = \pm\sqrt{\dfrac{M_{12}}{M_{30}}}$.

The signs of $\hat{\beta}_5$, and $\hat{\beta}_6$, are given by the sign of the sum of product of y and x (SP_{yx}) or the sign of the sample correlation coefficient $(\hat{\rho})$. These estimators need to assume that $\mu_3(\xi) \neq 0$ and $\beta_1 \neq 0$. But there is no method to find out the particular estimator which is the consistent estimator of the slope parameter β_1.

Pal (1980) noted many estimators can be found by using the weighted arithmetic or geometric mean of the above estimators. Moreover, Scott (1950) pointed out that the root of

$$M_{03} - 3bM_{12} + 3b^2M_{21} - b^3M_{30} = 0,$$

itself will be a consistent estimator of the slope parameter β_1.

These estimators could be consistent for the slope β_1, because the sample moments are consistent estimators of the true moments. But if M_{12} and M_{21} are close to zero, then these estimators will probably be very unstable unless the sample size is very large. It seems that the lower-order moments generally have lower variations (cf Cheng and Van Ness, 1999, p. 121). Van Montfort et al. (1987) noted that the estimator $\hat{\beta}_3 = M_{12}/M_{21}$ is optimal in the sense of minimal asymptotic variance of all consistent estimators that are based on the moments up to order three, when the errors are nonsymmetric.

The asymptotic efficiency of a suitable estimator is provided by Pal (1980) with respect to the least squares estimator for different distributions of ξ. He showed that three of the above estimators are functions of the other slope estimators as follows:

$$(1)\ \hat{\beta}_1 = (\hat{\beta}_2.\hat{\beta}_3.\hat{\beta}_4)^{\frac{1}{3}}, \qquad (2)\ \hat{\beta}_5 = \pm(\hat{\beta}_2.\hat{\beta}_3)^{\frac{1}{2}}, \qquad (3)\ \hat{\beta}_6 = \pm(\hat{\beta}_3\hat{\beta}_4)^{\frac{1}{2}}.$$

Moreover, Van Montfort et al. (1987) introduced an optimal estimator of the slope which is a function of three slope estimators, and they discussed the estimators based on third order moments. Furthermore, they pointed out that if the variance covariance matrix Σ of the third-order moments is not known, then they should be estimated in order to obtain the optimal estimator of the slope parameter. The optimal estimator is based on moments up to order three since moments of order higher than three appear in the estimation of the variance covariance matrix (Van Montfort et al. 1987). Gillard (2010) mentioned that through a simulation study, Van Montfort et al. (1987) demonstrated that the optimal estimator works well for a sample size about 50, and it is superior to any other estimator based on third moment, under the standard assumption that the errors δ_j and ϵ_j are independent. The same study was replicated for a sample size of 25 but the third moment estimators performed badly.

Van Montfort (1989) gave alternative approaches to errors-in-variables modelling, including estimation based on third order moments, extensions to polynomial regressions, and using characteristic functions with the factor analysis model. Also he provided details on the asymptotic variances and covariances of the third order moment slope estimators. The optimal estimator is the weighted mean if both errors δ_j and ϵ_j are symmetric, and it is given by

$$\hat{\beta}_{1opt} = \frac{\kappa'\Sigma^{-1}\tau}{\kappa'\Sigma^{-1}\kappa},$$

where $\tau' = (\hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4)$, $\kappa' = (1, 1, 1)$, and Σ is the asymptotic covariance matrix of τ.

Note usually the asymptotic covariance matrix Σ is unknown and has a complicated form, so one needs to replace the asymptotic covariance matrix Σ by a consistent estimate of Σ. Thus the result is asymptotically equivalent to the optimal estimator $\hat{\beta}_{1opt}$, and it is no longer a function of moments up to the third order. However, Van Montfort et al. (1987) noted that the asymptotic variance of the optimal estimator $\hat{\beta}_{1opt}$ has the form $(\kappa'\Sigma^{-1}\kappa)^{-1}$. Moreover, they gave the asymptotic covariance matrix Σ of τ and a consistent estimator of Σ (cf Van Montfort, 1988, Ch. 1).

The method of higher-order moments lacks the desirable optimality properties of the maximum likelihood and ordinary least squares estimators (Cramer, 1946). The method of higher-order moments can be criticized because it is not uniquely defined. But when this method is applicable then it has the advantage of simplicity, as it only needs the low order moments of the distribution describing the observations to exist. Often, it assumes that these distributions are mutually uncorrelated. Cramer (1946) gave theoretical asymptotic variances and covariances of the estimators of the method of moments using the Delta Method. It can be used to fit the line and calculate approximate confidence intervals for the associated parameters after making particular distributional assumptions.

There are simpler ways of estimating the slope that are available if the error terms δ and ϵ are from a symmetric distribution, where the additional assumption $\mu_3(u) = \mu_3(e) = 0$ holds. For the third order sample moments of y and x to be sufficiently different from zero, the distribution of the latent variable ξ has to be sufficiently skewed. It is also necessary that the regression line of the third order estimator lie between the ordinary least squares regression lines of y on x and x on y.

On the other hand, using the fourth order moment estimating equations does not require the assumption that the distribution of the latent variable ξ has to be sufficiently skewed. However, the sample size should be larger to ensure that the estimator of the slope parameter whether using the third or fourth order moments is a stable estimator. But for the fourth order moments to be significantly different from zero, we need the distributions of y and x to have sufficiently large kurtosis. Then one can obtain the fourth order central moments as follows:

$$\begin{aligned}
M_4(m) &= \frac{1}{n}\sum_{j=1}^{n}(x_j-\bar{x})^4, && (2.13)\\
M_{31}(m,y) &= \frac{1}{n}\sum_{j=1}^{n}(x_j-\bar{x})^3(y_j-\bar{y}), && (2.14)\\
M_{22}(m,y) &= \frac{1}{n}\sum_{j=1}^{n}(x_j-\bar{x})^2(y_j-\bar{y})^2, && (2.15)\\
M_{13}(m,y) &= \frac{1}{n}\sum_{j=1}^{n}(x_j-\bar{x})(y_j-\bar{y})^3, && (2.16)\\
M_4(y) &= \frac{1}{n}\sum_{j=1}^{n}(y_j-\bar{y})^4. && (2.17)
\end{aligned}$$

Note the fourth order moment equations can be derived in a similar way to that used to derive the third order moment equations. As a result if there does not exist a unique estimator for the slope parameter then some of these equations are not needed. In this case, we use the equations which avoid the higher order moments of the error terms.

2.6.3 Estimation with cumulants

Another method of estimation is a method based on product cumulants proposed by Geary (1942). This method is closely related to the method of higher-order moments, and both methods lead to similar estimators. The series estimators of the method of cumulants are obtained by simple formula. The cumulants can be defined as follows. Let the explanatory variable x and the response variable y be jointly distributed random variables. Then provided the expansions are valid in the given domain, the natural logarithm of the joint characteristic function is

$$\begin{aligned}\Psi(t_1, t_2) &= log_e \phi(t_1, t_2) = log_e[E(e^{it_1 m + it_2 y})] \\ &= \sum_{r,s=0}^{\infty} k(r,s) \frac{(it_1)^r}{r!} \frac{(it_2)^s}{s!}, \end{aligned} \tag{2.18}$$

where Ψ is called the joint cumulant generating function. If $r \neq 0$ and $s \neq 0$ then $k(r, s)$ is called the r, s product cumulant of x and y. Using the method of cumulants the slope of the classical structural model (without equation error) can be estimated as follows. Let

$$\eta_j = \beta_0 + \beta_1 \xi_j, \quad y_j = \eta_j + \epsilon_j, \quad x_j = \xi_j + \delta_j,$$

then the joint characteristic function of (ξ_j, η_j) is

$$\phi(t_1, t_2) = E(e^{it_1 \xi + it_2 \eta}). \tag{2.19}$$

If the true values of η and ξ are centered with respect to their true mean, then the intercept vanishes, and the structural relationship could be rewritten as

$$\beta_1 \xi - \eta = 0. \tag{2.20}$$

According to Stuart and Ord (1994, Ch. 12) the important properties of bivariate cumulants are:

(1) The cumulant of a sum of independent random variables is the sum of the cumulants.

(2) The bivariate cumulant of independent random variables is zero.

(3) Cumulants are invariant under the change of origin, except for the first cumulant.

Based on the property (1) the joint cumulant generating function Ψ of the observed points (x_j, y_j) is the sum of the joint cumulant generating function of the unobserved points (ξ_j, η_j) and the measurement error in both response and explanatory variables (ϵ_j, δ_j). Due to the property (2) the bivariate cumulants of the measurement error in both variables (ϵ_j, δ_j) are zero at any order s, r, where the order s, r is positive. Furthermore, by property (3) the centering of the mean does not effect the estimation (see Cheng and Van Ness, 1999, p. 125).

Letting $k(x, y)$ denote the cumulants of (x, y), and $k(\xi, \eta)$ denote the cumulants of (ξ, η) then

$$k_{(m,y)}(r, s) = k_{(\xi,\eta)}(r, s).$$

From (2.19) and (2.20)

$$\beta_1 \frac{\partial \phi}{\partial i t_1} - \frac{\partial \phi}{\partial i t_2} = E[(\beta_1 \xi - \eta) e^{i t_1 \xi + i t_2 \eta}] = 0,$$

for more details see for example Cheng and Van Ness, 1999, p. 125; and Pal (1980).

If we replace the joint characteristic function ϕ by the cumulant generating function Ψ, which yields

$$\beta_1 \frac{\partial \Psi}{\partial it_1} - \frac{\partial \Psi}{\partial it_2} = \frac{1}{\phi}\left(\beta_1 \frac{\partial \phi}{\partial it_1} - \frac{\partial \phi}{\partial it_2}\right) = 0. \tag{2.21}$$

For all $r, s > 0$ and from (2.18) and (2.21) we then have

$$\beta_1 k(r+1, s) - k(r, s+1) = 0.$$

So if $k(r+1, s) \neq 0$ an estimator of the slope parameter is

$$\hat{\beta}_{1C} \quad = \quad \frac{k(r, s+1)}{k(r+1, s)}. \tag{2.22}$$

For any $r, s > 0$ the consistent estimators of the slope parameter are obtained by replacing the cumulants of the population $k(r, s)$ by the corresponding sample cumulants of order (r, s). That is, $K(r, s)$, of the empirical distribution P_n of the observed points (x_j, y_j), where the probability function of this distribution is

$$P_n(m, y) = \frac{1}{n} \sum_{j=1}^{n} I_{(-\infty, m]}(x_j) I_{(-\infty, y]}(y_j),$$

where I_Ω is the indicator function of the set Ω.

In fact, sample cumulants could be obtained by using moment estimates. For example,

$$K(3, 1) = M_{31} - 3M_{20}M_{11} = M_{31} - 3S_x^2 S_{yx},$$

where $M_{rs} = \frac{1}{n}\sum_{j=1}^{n}(m - \bar{x})^r (y - \bar{y})^s$. For more details see Stuart and Ord (1994, Section 3.29).

However, the cumulant method for estimating the slope parameter β_1 does not work if the manifest variables (x_j, y_j) are jointly normally distributed. Because all odd cumulants of order greater than or equal to three are zero in normal system, thus the estimator (2.22) is useless in the case of normal model. Moreover, the cumulant estimators become more and more unstable if a non normal model gets closer and closer to a normal model (cf Cheng and Van Ness, 1999, p. 127).

2.7 Instrumental Variables

One of the most popular methods in the econometrics literature of obtaining consistent estimator of the slope parameter, β_1, which has received extensive consideration, is a method based on the use of instrumental variables. The method of instrumental variables (IV) has become the most standard approach to measurement error problems. It produces consistent estimates of the slope parameter if a suitable instrumental variable exists. An instrumental variable is suitable if it is uncorrelated with the measurement error and the equation error. However, if correlated with the correctly measured variable, then it provides a consistent estimator only under the conditions to make the model identifiable. The estimation via instrumental variable was coined by Reiersol (1950), and for a historical review of this method see

Goldberger (1972). In fact, use of the instrumental variable was considered as a different type of auxiliary information to make the normal model identifiable (Fuller, 2006, p. 50). Assume that a structural measurement error model is given by

$$y_j = \beta_0 + \beta_1 x_j + \varepsilon_j, \quad x_j = \xi_j + \delta_j,$$

for $j = 1, 2, \ldots, n$ where ε_j are independent random variables and $\varepsilon_j \sim N(0, \sigma_\varepsilon^2)$. Suppose there is a third variable, denoted by z_j, known to be correlated with the latent explanatory variable ξ_j, and uncorrelated with the measurement error δ_j. It is often assumed that both ε_j and δ_j are independent of z_j. The instrumental variable z_j is valid to use if it satisfies the following conditions

$$\text{(a)} \qquad E\left[\frac{1}{n}\sum_{j=1}^{n}(z_j - \bar{z})(\varepsilon_j, \delta_j)\right] = (0, 0) \tag{2.23}$$

$$\text{(b)} \qquad E\left[\frac{1}{n}\sum_{j=1}^{n}(z_j - \bar{z})\xi_j\right] \neq 0. \tag{2.24}$$

Notice that

$$\begin{aligned} \frac{1}{n}\sum(z_j - \bar{z})y_j &= \frac{1}{n}\sum(z_j - \bar{z})\beta_0 + \frac{1}{n}\sum(z_j - \bar{z})\beta_1 x_j + \frac{1}{n}\sum(z_j - \bar{z})\varepsilon_j \\ &= \frac{1}{n}\sum(z_j - \bar{z})\beta_1 x_j + \frac{1}{n}\sum(z_j - \bar{z})\varepsilon_j \\ &= \frac{1}{n}\sum(z_j - \bar{z})\beta_1 x_j, \text{ as } n \to \infty. \end{aligned}$$

Then the instrumental variable estimator of β_1 is

$$\hat{\beta}_{1IV} = \frac{\sum(z_j - \bar{z})y_j}{\sum(z_j - \bar{z})x_j}. \tag{2.25}$$

The instrumental variable estimator of the intercept β_0 is then

$$\hat{\beta}_{0IV} = \bar{y} - \hat{\beta}_{1IV}\bar{x}. \tag{2.26}$$

Johnston (1972, p. 284) showed how to use Wald's grouping method and the method based on ranks as an instrumental variable. The instrumental variable estimator of the slope parameter is consistent and asymptotically normal, if and only if the following conditions are met

(1) The instrumental variable must be correlated with the latent explanatory variable ξ_j.

(2) The instrumental variable must be independent of the measurement errors δ and ϵ, and also independent of the equation error e.

Indeed, there is practical difficulty in how to find a variable that is correlated with the latent variable ξ and independent of the measurement error. Carroll et al. (1995, p. 107) noted:

" One possible source of an instrumental variable is a second measurement on ξ obtained by an independent method. This second measurement need not be unbiased for ξ. Thus the assumption that a variable is an instrument is weaker than the assumption that it is a replicate measurement".

The other difficulty that the instrumental variable may not be unique, then how to use the instrumental variable based on the concept of consistency only. Since the consistency means that the estimator converges to the parameter with increase the sample size, whereas the unbiasedness means that the estimator has a sampling distribution centered on the parameter for any sample size. Angrist and Krueger (2001) stated “The instrumental variables estimates are consistent, but not unbiased, researchers using instrumental variables should aspire to work with large samples”.

2.8 Method based on ranks

The method of ranks is proposed by Theil (1950), it can be viewed as being related to the instrumental variable approach. This method is based on ordering the data according to either the manifest explanatory variable x_j or the manifest response variable y_j. He proposed a linear regression procedure with no special assumptions regarding the distribution of the data. The parameters are estimated by a nonparametric principle, and there are no assumptions of error distributions. Furthermore, this method does not presume Gaussian distributions of the true values, but only with regards to the error distributions. Moreover, the jackknife principle, used for estimation of standard errors in the Deming, Non-parametric, and Weighted Deming procedures (cf Saracli et al. 2009). The method of ranks takes measurement errors of both variables x, y into account, but the method presumes that the ratio between analytical standard deviations is related to the slope, otherwise the estimator will be biased. If one assumes that the data is ordered based on the explanatory variable x to obtain the order statistics $x_{(1)}, \ldots, x_{(n)}$, then the instrumental variable can be taken as $z_j = i$, where i is the order of the statistics of the manifest variable x_j (cf Cheng and Van Ness, 1999, p. 119).

However, this method requires a strong assumption that the values of the latent explanatory variable ξ are so spread out compared with the variance of error σ_δ^2 that the series of observed x_j is in the same order as the latent variable ξ_j. This assumption is equivalent to

$$x_i \leq x_j \quad \Leftrightarrow \quad \xi_i \leq \xi_j. \tag{2.27}$$

Moreover, the standard assumptions on the error structure are made as follows:

Assume that $\delta_1, \ldots, \delta_n$, $\epsilon_1, \ldots, \epsilon_n$ all have finite variances and uncorrelated, and have mean zero, that is,

$$\begin{aligned} E(\delta_j) &= E(\epsilon_j) = 0 \qquad \text{for all } j, & (2.28)\\ cov(\delta_i, \delta_j) &= cov(\epsilon_i, \epsilon_j) = 0 \qquad \text{for all } j \neq i, & (2.29)\\ cov(\delta_j, \epsilon_j) &= 0 \qquad \text{for all } j, & (2.30) \end{aligned}$$

and n is even such that, $n = 2k$, and that $x_i \neq x_j$ for all $i \neq j$, then the lagged slopes are

$$b_i = \frac{y_{(k+i)} - y_{(i)}}{x_{(k+i)} - x_{(i)}}, \qquad i = 1, \ldots, k, \tag{2.31}$$

and the paired slopes,

$$b_{i,j} = \frac{y_{(j)} - y_{(i)}}{x_{(j)} - x_{(i)}}, \qquad i = 1, \ldots, j-1;\ j = 2, \ldots, n. \tag{2.32}$$

Then one can form estimators of the slope parameter β_1 using the arithmetic mean or median of either from the lagged slopes or the paired slopes.

By the lagged slopes we have

$$\hat{\beta}_{1La} = \frac{1}{k}\sum_{i=1}^{k} b_i, \tag{2.33}$$

$$\hat{\beta}_{1Lm} = median(b_i), \tag{2.34}$$

or by the paired slopes we have

$$\hat{\beta}_{1Pa} = \frac{2}{n(n-1)}\sum_{j=2}^{n}\sum_{i=1}^{k} b_{i,j}, \tag{2.35}$$

$$\hat{\beta}_{1Pm} = median(b_{i,j}). \tag{2.36}$$

Cheng and Van Ness (1999, p. 119) stated that the assumption of this method is considered as a new identifiability side condition, and it might hold for the structural model except for very small sample sizes and either the latent variable ξ_j is not normal or σ_δ^2 is very small. They also mentioned that if the sample size n is large and $\sigma_\delta^2 > 0$ then the assumption above will not be satisfied for the normal structural model.

2.9 Maximum likelihood approach

The application of the maximum likelihood method to measurement error models has been considered by many authors. The common approach to finding the maximum likelihood estimator is by minimizing the likelihood function by differentiating it with respect to the parameters, and setting the resulting derivatives to zero. In fact, the likelihood equations can have one or more solutions, which might be a saddle point, a local maximum, or a local minimum of the likelihood function. Hood et al. (1999) pointed out that it is unlikely that theoretical results concerning the asymptotic variances of the estimators can be derived for anything other than the normal structural model.

The first author to use maximum likelihood estimation for the errors-in-variables model is Lindley (1947). He mentioned that the likelihood equations are consistent if there is some prior information available on the parameters. He pointed out that the most common assumption is to assume that the ratio of error variances λ is known. Kendall and Stuart (1973) showed the estimation in measurement error model using the maximum likelihood approach. They pointed out that the sample means, variances and covariances form sufficient statistics to derive the estimates from the familiar maximum likelihood estimates for means, variances, and covariances for a bivariate normal distribution (see for example, Kendall and Stuart, 1979, ch. 18). Kendall and Stuart (1973) introduced various cases based on a different assumption regarding a subset of the parameters. For each of these cases they derived estimators for the parameters. Moreover, they advise on how to construct confidence intervals. Also they introduced a brief survey on cumulants, instrumental variables and grouping methods.

Barnett (1970) considered the fitting of a functional model with applications on the importance of measurement error models in the medical and biological areas. He used the maximum likelihood approach for estimating the slope parameter. He commented that the maximum likelihood approach tended to run into computational problems, because of the awkward nature of the likelihood equations. Barnett showed the inherent difficulties in using the maximum likelihood method, and he examined alternative error structures which could be applicable to biological and medical data, but no closed form solution could be found.

Wong (1989) focused on the likelihood equations when both error variances were assumed to be known and equal. In fact, the situation of when both error variances are known and equal has received much attention. Wong used an orthogonal parametrisation in which the slope parameter is orthogonal to the remaining parameters. Which also included approximate confidence intervals for the parameters, information on testing hypotheses of the slope, and the density function for the slope estimator.

Solari (1969) considered the maximum likelihood method when the observations of the variables are normally distributed with unknown means and unknown variances. She pointed out that the maximum likelihood solution for the linear functional equations was a saddle point, and not a maximum. Solari (1969) concluded that a maximum likelihood solution for the linear functional model exists if only there is some prior distribution to place on a parameter. The maximum likelihood estimator of the slope parameter introduced by Solari is

$$\hat{\beta}_1 \quad = \quad sgn\{S_{yx}\}\sqrt{\frac{S_y^2}{S_x^2}}.$$

Sprent (1970) and Copas (1972) examined Solari's work and the practical implications of her findings. Copas pointed out that the likelihood surface becomes bounded when the rounding-off errors are considered in the observation. This situation allows for a different consideration of the likelihood surface. Copas mentioned that it is possible to find a solution which is approximately the maximum likelihood in the sense that the value of the likelihood at that solution is close to the global supremum. His solution for the slope is equivalent to using the x on y estimate and the y on x estimate.

The modified likelihood equation introduced by Copas is

$$\begin{aligned} L &= \prod P_j(x_j)Q_j(y_j), \text{ where} \\ P_j(x_j) &= P\left(x - \frac{1}{2}h \le \xi_j < x + \frac{1}{2}h\right), \text{ and} \\ Q_j(y_j) &= P\left(y - \frac{1}{2}h \le \eta_j < y + \frac{1}{2}h\right), \end{aligned}$$

when $\sigma_\xi > 0$, then

$$P_j(x_j) = \Phi\left(\frac{x - \mu_j + \frac{1}{2}h}{\sigma_\xi}\right) - \Phi\left(\frac{x - \mu_j - \frac{1}{2}h}{\sigma_\xi}\right),$$

where Φ is the standard normal distribution function, and it is approximately given by

$$\Phi = \frac{h}{\sqrt{2\pi\sigma_\xi^2}} \exp\left\{-\frac{(x - \mu_j)^2}{2\sigma_\xi^2}\right\}.$$

Note that Copas's model did not include an intercept, and the value of h was introduced to allow a discrepancy when $(\xi_j; \beta_1\xi_i)$ were recorded or measured. The direct consequence of the saddle point of the modified likelihood equation of Copas is

$$\begin{aligned} A &= \{\beta_1, \sigma_\delta, \sigma_\epsilon, \xi : \sum(x_j - \xi_j)^2 = 0, \sigma_\delta = 0\}, \text{ and} \\ B &= \{\beta_1, \sigma_\delta, \sigma_\epsilon, \xi : \sum(y_j - \beta_1\xi_j)^2 = 0, \sigma_\epsilon = 0\}. \end{aligned}$$

Within sets A and B the modified likelihood equation reduces to the likelihood equation for OLS(y/x) regression and OLS(x/y) regression (cf Copas 1972). It is clear that Copas's method is equivalent to using y on ξ regression if σ_δ^2 is close to zero. But if $\beta_1\xi_j$ is close to the manifest response variable

y then Copas's method is equivalent to using x on y regression (cf Gillard, 2010).

However, if all the distributions describing variation in the data are assumed to be normal then the maximum likelihood method handles the measurement error problem only. Gillard (2010) stated, “a unique solution is available only if additional information about certain parameters of the model are available, which often includes information regarding the variances of error”.

2.10 Structural equation modelling

Structural equation modelling (SEM) is used to describe a large number of statistical models used to evaluate the validity of substantive theories with empirical data. Structural equation modelling is also known as covariance structure analysis. It could be considered as an extension of general linear modeling (GLM) procedures, such as the ANOVA and multiple regression analysis. There are computer packages that fit the models through the structural equation modelling such as LISREL (Linear Structural Relationships) (see for example Skrondal and Rabe-Hesketh, 2004). In fact, there is a common belief, which is not quite true, that the structural equation models have been successfully applied to handle the measurement error problem, specially in the behavioral and social sciences in modelling.

But in order to estimate the parameters of normal structural models something further assumed to be known to avoid the problem of identifiability. Madansky (1959) stated “To use standard statistical techniques of estimation to estimate β_1, one needs additional information about the variance of the estimator”. Sanchez et al. (2005) stated “the structural equation models are particularly susceptible to identifiability problems. The SEM estimate for the exposure effect is the same as the estimate from an instrumental variables approach to measurement error”. He also mentioned that many statisticians and researchers in other areas of application are relatively unfamiliar with SEM.

Grewal et al. (2004) stated “In the past, researchers have often assumed that because SEM takes into account measurement error and corrects paths for attenuation, measure unreliability is less of a problem. Our findings clearly show that this assumption is not warranted”. However SEM, which sometimes called covariance structure modeling, is a complicated model and many researchers prefer to ignore the measurement error problem rather than using SEM. Moreover, Brannick (1995) pointed out that the covariance structure modeling is unlikely to produce scientific progress. Furthermore, Barrett (2007) stated “ the structural equation model fit has recently become a confusing and contentious area of evaluative methodology”.

2.11 Other contributions

Degracie and Fuller (1972) considered estimation of the slope and covariance when the concomitant variable is measured with error. Grubbs (1973) discussed errors of measurement, precision, accuracy and the statistical inference. Aigner (1973) considered regression with a binary variable subject to errors of observation. Florens et al. (1974) considered Bayesian inference in error-in-variables models.

Schneeweiss (1976) proposed consistent estimator of the regression model with errors in the variables. Bhargava (1977) introduced maximum likelihood estimation in a multivariate errors-in-variables regression model with unknown error covariance matrix. Prentice (1982) dealt with covariant measurement errors and parameter estimation in a failure time regression model. Amemiya et al. (1984) proposed estimation of the multivariate errors-in-variables model with estimated error covariance matrix.

Klepper and Leamer (1984) provided consistent sets of estimators for regression with errors in all variables. Stefanski and Carroll (1985) discussed covariant measurement error in logistic regression. Carroll et al (1985) proposed comparison of least squares and errors-in-variables regression with special reference to randomized analysis of covariance. Armstrong (1985) dealt with the measurement error in the generalized linear model. Bekker (1986) commented on identification in the linear errors-in-variables model. Schafer (1986) combined information on measurement error in errors-in-variables model. Kim and Saleh (2002, 2003, 2005) concentrated on the test and improve the estimation of the parameters of the simple linear models with measurement error. Recently Fuller (2006) covered various aspects of the measurement error models and related inferences.

The next chapter introduces a new methodology based on the mathematical transformation of the manifest variables. This methodology relies on the combination of the reflection and ordinary least squares techniques. Moreover, it includes some theorems to help interpret vertical, orthogonal, and horizontal distances between the observed points and regression line. In order to fit the regression line when both the explanatory and the response variables are subject to error, we will be using the *reflection* of the explanatory variable about the regression line. The asymptotic consistency and the mean absolute error (MAE) criteria are used to compare the new estimator with the relevant existing estimators under different conditions.

Chapter 3

The reflection approach to measurement error model

3.1 Introduction

This chapter introduces a mathematical transformation of the manifest variables. It is an algebraic transform of the manifest data of both response and explanatory variables. The reflection technique has some useful geometrical properties to interpret the problem of measurement error in the linear regression model. It is well known that the reflection technique transforms the points whilst preserving midpoint, collinearity, betweenness, distances, and angles. For example, the distance of the observed point (x, y) from the line of reflection is the same as the distance of the reflection point (x^*, y^*) from the line of reflection.

The reflection based regression line bisects the distance between the observed point (x, y) and its reflection point (x^*, y^*). The line between the observed point (x, y) and its reflection point (x^*, y^*) is perpendicular with the reflection line. Based on the reflection technique and ordinary least squares (OLS) method, this chapter provides a set of theorems related to the linear regression models. These theorems are applied in the following chapters to define proposed estimator and study its properties.

3.2 Methodology

The proposed methodology relies on the combination of the reflection and ordinary least squares techniques. The incorporation of both ordinary least squares and reflection techniques help interpret vertical, orthogonal, and horizontal distances between the observed points and regression line. The transformation formulas are resulted from overlap between the ordinary least squares (OLS) and the reflection technique. These formulas produced transformed variables (reflection variables), x^* and y^*, for both the explanatory and response variables, x and y respectively.

The transformed variables are obtained by considering the OLS regression line of y on x as a reflection line. All the observed points (x, y) are reflected about the OLS regression line of y on x using the following reflection formulas

$$x_j^* = x_j cos2\psi + (y_j - \hat{\beta}_{0x}) sin2\psi, \tag{3.1}$$

$$y_j^* = x_i sin2\psi - (y_j - \hat{\beta}_{0x}) cos2\psi + \hat{\beta}_{0x}. \tag{3.2}$$

Here x^* and y^* are transformed (reflection) variables of the manifest variables x and y respectively, $\hat{\beta}_{0x}$ and $\hat{\beta}_{1x}$ are the least squares estimator of the intercept and slope when both response and explanatory variables are subject to measurement error, and $\psi = tan^{-1}\hat{\beta}_{1x}$. For the definition of reflection of points on the Cartesian plane readers may see Vaisman (1997, p. 164-169).

3.3 Residuals analysis by reflection technique

It is well known that the regression line does not pass through all the data points on the scatter plot unless the correlation coefficient is ±1. Often the data points are scattered around the regression line. Points not falling on the regression line, have a vertical distance from the fitted line. This distance is known as the residual representing unexplained variation in the regression. The length of the vertical residual vary from point to point.

This section introduces a new method to analyse the vertical residuals of the measurement error model by Theorem 1 below. The reflection based regression allows to partition the squared residuals (unexplained variation) in to vertical and horizontal parts.

Theorem 1 *The square of the unexplained variation of y by x can be partitioned in to the vertical and horizontal components as follows:*

$$(y_j - \hat{y}_j)^2 = (y_j^* - \hat{y}_j)^2 + (x_j^* - x_j)^2, \qquad j = 1, 2, ..., n.$$

Then it can be shown that the sum of squared residuals is

$$SSE_{yx} = \sum_{j=1}^{n}(y_j - \hat{y}_j)^2 = \sum_{j=1}^{n}(y_j^* - \hat{y}_j)^2 + \sum_{j=1}^{n}(x_j^* - x_j)^2 = SSE_y + SSE_x$$

and total sum of squares is

$$\begin{aligned} TSS &= SSR_{yx} + SSE_{yx} = \sum_{j=1}^{n}(y_j - \bar{y})^2 = \sum_{j=1}^{n}(\hat{y}_j - \bar{y})^2 + \sum_{j=1}^{n}(y_j - \hat{y})^2 \\ &= \sum_{j=1}^{n}(\hat{y}_j - \bar{y})^2 + \sum_{j=1}^{n}(y_j^* - \hat{y}_j)^2 + \sum_{j=1}^{n}(x_j^* - x_j)^2. \end{aligned}$$

Here SSR_{yx} is the squared sum of explained variation of y by x, SSE_{yx} is the squared sum of unexplained variation of y by x, SSE_y is the squared sum of vertical unexplained variation in y, and SSE_x is the squared sum of horizontal unexplained variation as a function of x.

Proof From (3.1), (3.2) and for each $j = 1, 2, 3, ..., n$,

$$(y_j^* - \hat{y}_j)^2 + (x_j^* - x_j)^2 = (y_j^{*2} - 2y_j^*\hat{y}_j + \hat{y}_j^2) + (x_j^{*2} - 2x_j^*x_j + x_j^2)$$

$$\begin{aligned} = \quad & x_j^2 \sin^2 2\psi - 2y_j x_j \sin 2\psi \cos 2\psi + y_j^2 \cos^2 2\psi + 2\hat{\beta}_{0x} x_j \sin 2\psi \cos^2 \psi \\ & -4\hat{\beta}_{0x} y_j \cos 2\psi \cos^2 \psi + 4\hat{\beta}_{0x}^2 \cos^4 \psi - 2\hat{\beta}_{0x} x_j \sin 2\psi + 2\hat{\beta}_{0x} y_j \cos 2\psi \\ & -2\hat{\beta}_{0x}^2 \cos 2\psi - 2\hat{\beta}_{1x} x_j^2 \sin 2\psi + 2\hat{\beta}_{1x} y_j x_j \cos 2\psi - 2\hat{\beta}_{0x}\hat{\beta}_{1x} x_j \cos 2\psi \\ & +\hat{\beta}_{0x}^2 + 2\hat{\beta}_{0x}\hat{\beta}_{1x} x_j + \hat{\beta}_{1x}^2 x_j^2 + x_j^2 \cos^2 2\psi + 2y_j x_j \cos 2\psi \sin 2\psi \\ & -2\hat{\beta}_{0x} x_j \cos 2\psi \sin 2\psi + y_j^2 \sin^2 2\psi - 2\hat{\beta}_{0x} y_j \sin^2 2\psi + \hat{\beta}_{0x}^2 \sin^2 2\psi \\ & -2x_j^2 \cos 2\psi - 2y_j x_j \sin 2\psi + 2\hat{\beta}_{0x} x_j \sin 2\psi + x_j^2. \end{aligned}$$

In order to simplify the algebraic expression above, we separate the terms to three parts as follows:

Part(I) collect all terms involving $\hat{\beta}_{0x}$:

$\hat{\beta}_{0x}[2x_j(\sin 2\psi \cos^2 \psi - \sin 2\psi + \hat{\beta}_{1x}(1 - \cos 2\psi) - \cos 2\psi \sin 2\psi + \sin 2\psi)$

$-2y_j(2\cos^2 \psi \cos 2\psi - \cos 2\psi + \sin^2 2\psi)]$

$= \hat{\beta}_{0x}[2x_j(\sin 2\psi + 2\hat{\beta}_{1x} \sin^2 \psi) - 2y_j]$

$= 2\hat{\beta}_{0x}\hat{\beta}_{1x} x_j - 2\hat{\beta}_{0x} y_j.$

Note that $(2\sin 2\psi \cos^2 \psi - \cos 2\psi \sin 2\psi) = \sin 2\psi$ and

$(2\cos^2 \psi \cos 2\psi - \cos 2\psi + \sin^2 2\psi) = 1.$

Part(II) collecting all terms multiplied by $\hat{\beta}_{0x}^2$:

$\hat{\beta}_{0x}^2(4\cos^4\psi - 2\cos 2\psi + 1 + \sin^2 2\psi)$

$$= \hat{\beta}_{0x}^2(4\cos^4\psi - \cos 2\psi + 3\sin^2\psi + 4\cos^2\psi\sin^2\psi)$$

$$= \hat{\beta}_{0x}^2\cos^2\psi(4\cos^2\psi - 1 + 3\hat{\beta}_{1x}^2 + 4\sin^2\psi)$$

$$= \hat{\beta}_{0x}^2\cos^2\psi(1 + \hat{\beta}_{1x}^2) = \hat{\beta}_{0x}^2\cos^2\psi\left(1 + \frac{\sin^2\psi}{\cos^2\psi}\right)$$

$$= \hat{\beta}_{0x}^2\cos^2\psi\left(\frac{\cos^2\psi + \sin^2\psi}{\cos^2\psi}\right) = \hat{\beta}_{0x}^2.$$

Note that $(4\cos^4\psi - 2\cos 2\psi + 1 + \sin^2 2\psi) = 3$ for any value of ψ.

Part(III) collecting all terms multiplied by x_j^2:

$$\begin{aligned}
& x_j^2(\sin^2 2\psi - 2\hat{\beta}_{1x}\sin 2\psi + \hat{\beta}_{1x}^2 + \cos^2 2\psi - 2\cos 2\psi + 1) \\
= \; & x_j^2(\sin^2 2\psi - 2\hat{\beta}_{1x}\sin 2\psi + \hat{\beta}_{1x}^2 + (\cos 2\psi - 1)^2) \\
= \; & x_j^2(4\sin^2\psi\cos^2\psi - 4\sin^2\psi + \frac{\sin^2\psi}{\cos^2\psi} + 4\sin^4\psi) \\
= \; & x_j^2\sin^2\psi(4\cos^2\psi + 4\sin^2\psi + \frac{1}{\cos^2\psi} - 4) \\
= \; & x_j^2\sin^2\psi(\frac{1}{\cos^2\psi}) = \hat{\beta}_{1x}^2 x_j^2,
\end{aligned}$$

where $\psi = tan^{-1}\hat{\beta}_{1x}$, $\hat{\beta}_{0x} = \bar{y} - \hat{\beta}_{1x}\bar{x}$, and

$(\sin^2 2\psi - 2\hat{\beta}_{1x}\sin 2\psi + \hat{\beta}_{1x}^2 + \cos^2 2\psi + 2\cos 2\psi + 1) = \hat{\beta}_{1x}^2.$

After simplifying the expressions we get

$$
\begin{aligned}
&(y_j^{*2} - 2y_j^*\hat{y}_j + \hat{y}_j^2) + (x_j^{*2} - 2x_j^*x_j + x_j^2) \\
&= y_j^2 - 2\hat{\beta}_{0x}y_j - 2\hat{\beta}_{1x}x_jy_j + \hat{\beta}_{0x}^2 + 2\hat{\beta}_{0x}\hat{\beta}_{1x}x_j + \hat{\beta}_{1x}^2x_j^2 \\
&= y_j^2 - 2y_j(\hat{\beta}_{0x} + \hat{\beta}_{1x}x_j) + (\hat{\beta}_{0x} + \hat{\beta}_{1x}x_j)^2 \\
&= y_j^2 - 2y_j\hat{y}_j + \hat{y}_j^2 = (y_j - \hat{y}_j)^2,
\end{aligned}
$$

hence

$$(y_j^* - \hat{y}_j)^2 + (x_j^* - x_j)^2 \;\; = \;\; (y_j - \hat{y}_j)^2. \tag{3.3}$$

Now by inserting sum to the both sides of (3.3) we get

$$\sum_{j=1}^{n}(y_j^* - \hat{y}_j)^2 + \sum_{j=1}^{n}(x_j^* - x_j)^2 = \sum_{j=1}^{n}(y_j - \hat{y}_j)^2 = SSE. \tag{3.4}$$

Since $\sum_{j=1}^{n}(y_j^* - \hat{y}_j)^2 = SSE_y$ and $\sum_{j=1}^{n}(x_j^* - x_j)^2 = SSE_x$, then

$$SSE \;\; = \;\; SSE_y + SSE_x, \text{ and} \tag{3.5}$$

$$TSS \;\; = \;\; SSR + SSE = SSR + SSE_y + SSE_x. \tag{3.6}$$

3.3.1 An alternative proof

Note that we can alternatively prove this theorem geometrically based on the properties of the reflection of point as follows: In Figure 3.1, let A be the point (x, y), D be its reflection point (x^*, y^*), $B = (x, \hat{y})$, $C = (x, y^*)$, and consider that the regression line of OLS of y on x be the reflection line.

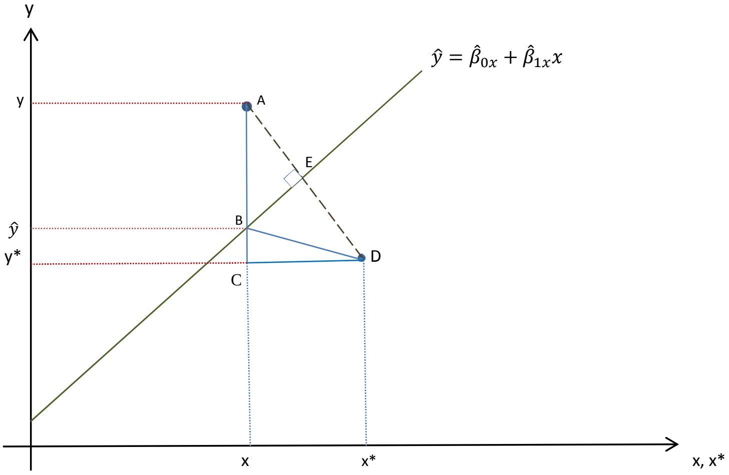

Figure 3.1: Graph of a reflection point about the OLS regression line of y on x.

Then $\overline{AE} = \overline{ED}$, and BE is a common side between the triangles ΔABE and ΔDBE. Based on that, triangles ΔABE and ΔDBE are identical, hence $\overline{AB} = \overline{BD}$. From the triangle ΔBCD that $\overline{BD}^2 = \overline{BC}^2 + \overline{CD}^2$. Then

$$\overline{AB}^2 = \overline{BC}^2 + \overline{CD}^2,$$

where $\overline{AB} = (y_j - \hat{y}_j)$, $\overline{BC} = (y_j^* - \hat{y}_j)$, and $\overline{CD} = (x_j^* - x_j)$. Hence

$$(y_j - \hat{y}_j)^2 = (y_j^* - \hat{y}_j)^2 + (x_j^* - x_j)^2.$$

By summing both sides over j from 1 to n, we get

$$\sum_{j=1}^{n}(y_j - \hat{y}_j)^2 = \sum_{j=1}^{n}(y_j^* - \hat{y}_j)^2 + \sum_{j=1}^{n}(x_j^* - x_j)^2.$$

Then it follows:

$$\begin{aligned} SSE &= SSE_y + SSE_x, \text{ and} \\ TSS &= SSR + SSE = SSR + SSE_y + SSE_x. \end{aligned}$$

3.4 Advantages of using reflection

Theorem 2 *The mean of the reflection of manifest variable $\bar{x}^*$ equals the arithmetic mean of manifest explanatory variable $\bar{x}$ and the mean of latent variable $\bar{\xi}$. That is,* $\bar{x}^* = \bar{x} = \bar{\xi}$.

Proof From (3.1)

$$\begin{aligned}\sum_{j=1}^{n} x_j^* &= \sum_{j=1}^{n}(x_j \cos 2\psi + y_j \sin 2\psi - \hat{\beta}_{0x} \sin 2\psi)\\ &= \sum_{j=1}^{n} x_j \cos 2\psi + \sum_{j=1}^{n} y_j \sin 2\psi - n\hat{\beta}_{0x} \sin 2\psi,\end{aligned}$$

since $\hat{\beta}_{0x} = \bar{y} - \hat{\beta}_{1x}\bar{x}$, then

$$\begin{aligned}\sum_{j=1}^{n} x_j^* &= \sum_{j=1}^{n} x_j \cos 2\psi + \sum_{j=1}^{n} y_j \sin 2\psi - \sum_{j=1}^{n} y_j \sin 2\psi + \hat{\beta}_{1x} \sum_{j=1}^{n} x_j \sin 2\psi\\ &= \sum_{j=1}^{n} x_j(\cos 2\psi + \hat{\beta}_{1x} \sin 2\psi)\\ &= \sum_{j=1}^{n} x_j(\cos^2\psi - \sin^2\psi + \frac{\sin\psi}{\cos\psi} 2\cos\psi \sin\psi)\\ &= \sum_{j=1}^{n} x_j(\cos^2\psi - \sin^2\psi + 2\sin^2\psi)\\ &= \sum_{j=1}^{n} x_j(\cos^2\psi + \sin^2\psi) = \sum_{j=1}^{n} x_j.\end{aligned}$$

Thus

$$\frac{1}{n}\sum_{j=1}^{n} x_j^* = \frac{1}{n}\sum_{j=1}^{n} x_j$$

It is well know that the mean of manifest explanatory variable $\bar{x}$ equals the mean of latent variable $\bar{\xi}$, because there is a common assumption in the literature of the error in variables that the population mean of measurement error equals zero, hence

$$\bar{x}^* = \bar{x} = \bar{\xi}.$$

Theorem 3 *The mean of the manifest response variable $\bar{y}$ equals that of the latent variable $\bar{\eta}$ and the reflection of manifest variable $\bar{y}^*$.That is,*

$$\bar{\eta} = \bar{y} = \bar{y}^*$$

. **Proof** From (3.2)

$$\begin{aligned}
\sum_{j=1}^{n} y_j^* &= \sum_{j=1}^{n}(x_j \sin 2\psi - y_j \cos 2\psi + \hat{\beta}_{0x}\cos 2\psi + \hat{\beta}_{0x}) \\
&= \sum_{j=1}^{n}(x_j \sin 2\psi - y_j \cos 2\psi + \hat{\beta}_{0x}(\cos 2\psi + 1)) \\
&= \sum_{j=1}^{n}(x_j \sin 2\psi - y_j \cos 2\psi + \hat{\beta}_{0x}(\cos^2\psi - \sin^2\psi + \cos^2\psi + \sin^2\psi)) \\
&= \sum_{j=1}^{n}(x_j \sin 2\psi - y_j \cos 2\psi + 2\hat{\beta}_{0x}\cos^2\psi) \\
&= \sum_{j=1}^{n}(x_j \sin 2\psi - y_j \cos 2\psi + 2(\bar{y} - \hat{\beta}_{1x}\bar{x})\cos^2\psi) \\
&= \sum_{j=1}^{n} x_j \sin 2\psi - \sum_{j=1}^{n} y_j \cos 2\psi + 2\sum_{j=1}^{n} y_j \cos^2\psi - 2\hat{\beta}_{1x}\sum_{j=1}^{n} x_j \cos^2\psi \\
&= \sum_{j=1}^{n} x_j \sin 2\psi + \sum_{j=1}^{n} y_j(2\cos^2\psi - \cos 2\psi) - 2\sum_{j=1}^{n} x_j \frac{\sin\psi}{\cos\psi}\cos^2\psi \\
&= \sum_{j=1}^{n} x_j \sin 2\psi + \sum_{j=1}^{n} y_j(2\cos^2\psi - \cos^2\psi + \sin^2\psi) - \sum_{j=1}^{n} x_j \sin 2\psi \\
&= \sum_{j=1}^{n} y_j.
\end{aligned}$$

Then $\frac{1}{n}\sum_{j=1}^{n} y_j^* = \frac{1}{n}\sum_{j=1}^{n} y_j$.

Based on the assumption that the population mean of measurement error equals zero, the mean of the manifest response variable $\bar{y}$ equals the mean of the latent response variable $\bar{\eta}$, hence

$$\bar{y}^* = \bar{y} = \bar{\eta}.$$

Theorem 4 *The estimator of slope for the regression model of y on x equals the estimator of slope for the regression model of y^* on x, that is,*

$$\hat{\beta}_{1yx} = \hat{\beta}_{1y^*x}, \qquad \hat{\beta}_{0yx} = \hat{\beta}_{0y^*x},$$

where $\hat{\beta}_{1yx}$ is the OLS estimator of slope of y on x, and $\hat{\beta}_{1y^*x}$ is that of y^* on x. Also $\hat{\beta}_{0yx}$ is the estimator of intercept of OLS y on x and $\hat{\beta}_{0y^*x}$ is that of y^* on x.

Proof

It is well known that $\hat{\beta}_{1yx} = \frac{S_{yx}}{S_x^2}$, and $\hat{\beta}_{1y^*x} = \frac{S_{y^*x}}{S_x^2}$, where S_x^2 is the sample variance of the manifest explanatory variable x. Then in order to prove that $\hat{\beta}_{1yx} = \hat{\beta}_{1y^*x}$ we only need to prove that $S_{yx} = S_{y^*x}$.

$$\begin{aligned}
S_{y^*x} &= \frac{1}{n-1}\sum_{j=1}^{n}(x_j - \bar{x})(y_j^* - \bar{y}^*) \\
&= \frac{1}{n-1}\sum_{j=1}^{n}(x_j(y_j^* - \bar{y}^*) - \bar{x}(y_j^* - \bar{y}^*)) \\
&= \frac{1}{n-1}\sum_{j=1}^{n}(x_j y_j^* - x_j\bar{y}^* - y_j^*\bar{x} + \bar{y}^*\bar{x})
\end{aligned}$$

$$\begin{aligned}
S_{y^*x} &= \frac{1}{n-1}\left(\sum_{j=1}^{n} x_j y_j^* - \sum_{j=1}^{n} x_j \bar{y}^* - \sum_{j=1}^{n} y_j^* \bar{x} + \sum_{j=1}^{n} \bar{y}^* \bar{x}\right) \\
&= \frac{1}{n-1}\left(\sum_{j=1}^{n} x_j y_j^* - n\bar{y}^*\bar{x} - n\bar{y}^*\bar{x} + n\bar{y}^*\bar{x}\right) \\
&= \frac{1}{n-1}\left(\sum_{j=1}^{n} x_j y_j^* - n\bar{y}^*\bar{x}\right) \qquad (3.7)
\end{aligned}$$

Note that $\bar{y}^* = \bar{y}$ from Theorem 3.

But we still need to prove that $\sum_{j=1}^{n} x_j y_j^* = \sum_{j=1}^{n} x_j y_j$ as follows

$$\begin{aligned}
\sum_{j=1}^{n} x_j y_j^* &= \sum_{j=1}^{n} x_j(x_j \sin 2\psi - y_j \cos 2\psi + \hat{\beta}_{0x} \cos 2\psi + \hat{\beta}_{0x}) \\
&= \sum_{j=1}^{n} (x_j^2 \sin 2\psi - y_j x_j \cos 2\psi + \hat{\beta}_{0x} \cos 2\psi x_j + \hat{\beta}_{0x} x_j) \\
&= \sum_{j=1}^{n} x_j^2 \sin 2\psi - \sum_{j=1}^{n} y_j x_j \cos 2\psi + \sum_{j=1}^{n} \hat{\beta}_{0x} \cos 2\psi x_j + \sum_{j=1}^{n} \hat{\beta}_{0x} x_j \\
&= \sum_{j=1}^{n} x_j^2 \sin 2\psi - \sum_{j=1}^{n} y_j x_j \cos 2\psi + n\hat{\beta}_{0x} \cos 2\psi \bar{x} + n\hat{\beta}_{0x}\bar{x} \\
&= \sum_{j=1}^{n} x_j^2 \sin 2\psi - \sum_{j=1}^{n} y_j x_j \cos 2\psi + n\bar{y} \cos 2\psi \bar{x} - n\hat{\beta}_{1x} \cos 2\psi \bar{x}^2 \\
&\quad + n\bar{y}\bar{x} - n\hat{\beta}_{1x}\bar{x}^2 \\
&= \sum_{j=1}^{n} x_j^2 \sin 2\psi - \sum_{j=1}^{n} y_j x_j (2\cos^2\psi - 1) + n\bar{y}\bar{x}(\cos 2\psi + 1) \\
&\quad - n\hat{\beta}_{1x}\bar{x}^2(\cos 2\psi + 1) \\
&= \sum_{j=1}^{n} x_j^2 \sin 2\psi - 2\sum_{j=1}^{n} y_j x_j \cos^2\psi + \sum_{j=1}^{n} y_j x_j + 2n\bar{y}\bar{x}\cos^2\psi \\
&\quad - 2n\hat{\beta}_{1x}\bar{x}^2\cos^2\psi
\end{aligned}$$

$$\begin{aligned}
\sum_{j=1}^{n} x_j y_j^* &= \sum_{j=1}^{n} x_j^2 \sin 2\psi + \sum_{j=1}^{n} y_j x_j - \left(\sum_{j=1}^{n} y_j x_j - n\bar{y}\bar{x} \right) 2\cos^2\psi \\
&\quad -2n\hat{\beta}_{1x}\bar{x}^2 \cos^2\psi \\
&= \sum_{j=1}^{n} y_j x_j - \left(\sum_{j=1}^{n} y_j x_j - n\bar{y}\bar{x} \right) 2\cos^2\psi + \sum_{j=1}^{n} x_j^2 \sin 2\psi \\
&\quad -2n\bar{x}^2 \frac{\sin\psi}{\cos\psi} \cos^2\psi \\
&= \sum_{j=1}^{n} y_j x_j - 2SP_{yx}\cos^2\psi + \sum_{j=1}^{n} x_j^2 \sin 2\psi - n\bar{x}^2 \sin 2\psi \\
&= \sum_{j=1}^{n} y_j x_j - 2SP_{yx}\cos^2\psi + SS_x \sin 2\psi.
\end{aligned}$$

Note that $\hat{\beta}_{1x} = \dfrac{\sin\psi}{\cos\psi} = \dfrac{SP_{yx}}{SS_x}$, then

$$\begin{aligned}
\sum_{j=1}^{n} x_j y_j^* &= \sum_{j=1}^{n} y_j x_j - 2SS_x \frac{\sin\psi}{\cos\psi} \cos^2\psi + SS_x \sin 2\psi \\
&= \sum_{j=1}^{n} y_j x_j - SS_x \sin 2\psi + SS_x \sin 2\psi \\
&= \sum_{j=1}^{n} y_j x_j.
\end{aligned}$$

Hence

$$S_{y^*x} = \frac{1}{n-1} \sum_{j=1}^{n} (x_j - \bar{x})(y_j^* - \bar{y}^*) = \frac{1}{n-1} \sum_{j=1}^{n} (x_j - \bar{x})(y_j - \bar{y}) = S_{yx}.$$

Thus

$$\hat{\beta}_{1y^*x} = \frac{S_{y^*x}}{S_x^2} = \frac{S_{y^*x}}{S_x^2} = \hat{\beta}_{1y^*x}.$$

Note that $\bar{x} = \bar{x^*}$ and $\bar{y} = \bar{y^*}$, so

$$\hat{\beta}_{0yx} \;=\; \bar{y} - \hat{\beta}_{1yx}\bar{x} = \bar{y^*} - \hat{\beta}_{1y^*x}\bar{x^*} = \hat{\beta}_{0y^*x}.$$

Then

$$\hat{\beta}_{0yx} = \hat{\beta}_{0y^*x}.$$

Obviously, based on the theorems above, the ordinary least squares (OLS) estimator of y^* on x is better than OLS estimator of y on x, where the estimators of slope and intercept are equal, but there is difference in the sum squares errors as follow

$$\sum_{j=1}^{n}(y_j^* - \hat{y}_j)^2 \leq \sum_{j=1}^{n}(y_j - \hat{y}_j)^2,$$

where y_j^* is the reflection of y_j.

Theorem 5 *The sample variance of the response variable y is greater than that of its reflection y^*:*

$$S_y^2 = \frac{1}{n-1}\sum_{j=1}^{n}(y_j - \bar{y})^2 > S_{y^*}^2 = \frac{1}{n-1}\sum_{j=1}^{n}(y_j^* - \bar{y}^*)^2.$$

Proof

$$\begin{aligned} S_y^2 &= \frac{1}{n-1}\sum_{j=1}^{n}(y_j - \bar{y})^2 \\ &= \frac{1}{n-1}\sum_{j=1}^{n}(y_j - y_j^* + y_j^* - \bar{y})^2 \\ &= \frac{1}{n-1}\sum_{j=1}^{n}((y_j - y_j^*) + (y_j^* - \bar{y}))^2 \\ &= \frac{1}{n-1}\left[\sum_{j=1}^{n}(y_j - y_j^*)^2 - 2\sum_{j=1}^{n}(y_j - y_j^*)(y_j^* - \bar{y}^*) + \sum_{j=1}^{n}(y_j^* - \bar{y})^2\right]. \end{aligned}$$

Note that $\sum_{j=1}^{n}(y_j - y_j^*)(y_j^* - \bar{y}^*)$ is given by

$\sum_{j=1}^{n}(y_j - y_j^*)(y_j^* - \bar{y}^*) = \sum_{j=1}^{n} y_j y_j^* - \sum_{j=1}^{n} y_j^{*2} - n\bar{y}\bar{y}^* + n\bar{y}^{*2},$

where from Theorem 3 $\bar{y} = \bar{y}^*$, then

$$\sum_{j=1}^{n}(y_j - y_j^*)(y_j^* - \bar{y}^*) \;=\; \sum_{j=1}^{n} y_j y_j^* - \sum_{j=1}^{n} y_j^{*2}.$$

Now from the equation (3.2)

$$\begin{aligned}
\sum_{j=1}^{n} y_j y_j^* - \sum_{j=1}^{n} y_j^{*2} &= \sum_{j=1}^{n} y_j(x_j \sin 2\psi - y_j \cos 2\psi + 2\hat{\beta}_{0x}\cos^2\psi) - \sum_{j=1}^{n} y_j^{*2} \\
&= \sum_{j=1}^{n} y_j x_j \sin 2\psi - \sum_{j=1}^{n} y_j^2 \cos 2\psi + 2n\hat{\beta}_{0x}\bar{y}\cos^2\psi - \sum_{j=1}^{n} y_j^{*2} \\
&= \sum_{j=1}^{n} y_j x_j \sin 2\psi - \sum_{j=1}^{n} y_j^2 \cos 2\psi + 2n\bar{y}^2\cos^2\psi \\
&\quad -2n\hat{\beta}_{1x}\bar{y}\bar{x}\cos^2\psi - \sum_{j=1}^{n} y_j^{*2} \\
&= \sum_{j=1}^{n} y_j x_j \sin 2\psi - 2\sum_{j=1}^{n} y_j^2 \cos^2\psi + \sum_{j=1}^{n} y_j^2 \\
&\quad +2n\bar{y}^2\cos^2\psi - n\bar{y}\bar{x}\sin 2\psi - \sum_{j=1}^{n} y_j^{*2} \\
&= SP_{xy}\sin 2\psi - 2SS_y\cos^2\psi + \sum_{j=1}^{n} y_j^2 - \sum_{j=1}^{n} y_j^{*2}.
\end{aligned}$$

By adding and subtracting $n\bar{y}^2$ we have

$$\begin{aligned}
\sum_{j=1}^{n}(y_j - y_j^*)(y_j^* - \bar{y}^*) &= SP_{xy}\sin 2\psi - 2SS_y\cos^2\psi \\
&\quad +\sum_{j=1}^{n} y_j^2 - n\bar{y}^2 - \sum_{j=1}^{n} y_j^{*2} + n\bar{y}^2 \\
&= SP_{xy}\sin 2\psi - 2SS_y\cos^2\psi + SS_y - SS_{y^*}.
\end{aligned}$$

Here $SP_{xy}\sin 2\psi = SP_{xy}2\sin\psi\cos\psi\frac{\cos^2\psi}{\cos^2\psi} = 2\hat{\beta}_{1x}SP_{xy}\cos^2\psi$ then we have

$$\begin{aligned}\sum_{j=1}^{n}(y_j - y_j^*)(y_j^* - \bar{y}^*) &= 2\hat{\beta}_{1x}SP_{xy}\cos^2\psi - 2SS_y\cos^2\psi + SS_y - SS_{y^*}\\ &= 2(\hat{\beta}_{1x}SP_{xy} - SS_y)\cos^2\psi + SS_y - SS_{y^*}\\ &= SS_y - SS_{y^*} - 2SS_v\cos^2\psi,\end{aligned}$$

where $SS_v = SSE = \sum_{j=1}^{n}(y_j - \hat{y}_j)^2$ is the sum of squares residuals of the manifest model y on x.

Now and after analysing the terms above, the sample variance S_y^2 of response variable is given by

$$\begin{aligned}S_y^2 &= \frac{1}{n-1}\left[\sum_{j=1}^{n}(y_j - y_j^*)^2 - 2\sum_{j=1}^{n}(y_j - y_j^*)(y_j^* - \bar{y}^*) + \sum_{j=1}^{n}(y_j^* - \bar{y})^2\right]\\ &= \frac{1}{n-1}\left[\sum_{j=1}^{n}(y_j - y_j^*)^2 - SS_y + SS_{y^*} + 2SS_v\cos^2\psi + \sum_{j=1}^{n}(y_j^* - \bar{y})^2\right]\\ &= \left[\frac{1}{n-1}\sum_{j=1}^{n}(y_j - y_j^*)^2\right] - S_y^2 + S_{y^*}^2 + 2S_v^2\cos^2\psi + S_{y^*}^2.\end{aligned}$$

Rearranging we find

$$\begin{aligned}2S_y^2 &= \left[\frac{1}{n-1}\sum_{j=1}^{n}(y_j - y_j^*)^2\right] + 2S_v^2\cos^2\psi + 2S_{y^*}^2,\\ S_y^2 &= S_{y^*}^2 + S_v^2\cos^2\psi + \left[\frac{1}{2(n-1)}\sum_{j=1}^{n}(y_j - y_j^*)^2\right].\end{aligned}$$

Note that all terms on the right hand side of this equation are always positive, so

$$S_y^2 \geq S_{y^*}^2.$$

Theorem 6 *The sample covariance of the manifest explanatory variable x and its reflection x^* is equal to the sample variance of manifest explanatory variable x.*

$$\hat{cov}(x, x^*) = S_x^2,$$

where S_x^2 is the sample variance of x.

Proof

$$\begin{aligned}
\hat{cov}(x, x^*) &= \frac{1}{n-1}\left[\sum_{j=1}^{n}(x_j - \bar{x})(x_j^* - \bar{x}^*)\right] \\
&= \frac{1}{n-1}\left[\sum_{j=1}^{n}(x_j x_j^* - \bar{x}x_j^* - x_j\bar{x}^* + \bar{x}\bar{x}^*)\right], \\
&= \frac{1}{n-1}\left[\sum_{j=1}^{n} x_j x_j^* - n\bar{x}\bar{x}^* - n\bar{x}\bar{x}^* + \bar{x}\bar{x}^*\right].
\end{aligned}$$

From Theorem 2, $\bar{x}^* = \bar{x}$, and so

$$\hat{cov}(x, x^*) = \frac{1}{n-1}\left[\sum_{j=1}^{n} x_j x_j^* - n\bar{x}^2\right].$$

From equation (3.1)

$$\begin{aligned}
\hat{cov}(x, x^*) &= \frac{1}{n-1}\left[\sum_{j=1}^{n} x_j(x_j \cos 2\psi + y_j \sin 2\psi - \hat{\beta}_{0x} \sin 2\psi) - n\bar{x}^2\right] \\
&= \frac{1}{n-1}\left[\sum_{j=1}^{n} x_j(x_j \cos 2\psi + y_j \sin 2\psi - \hat{\beta}_{0x} \sin 2\psi) - n\bar{x}^2\right] \\
&= \frac{1}{n-1}\left[\sum_{j=1}^{n}(x_j^2 \cos 2\psi + y_j x_j \sin 2\psi - \hat{\beta}_{0x} x_j \sin 2\psi) - n\bar{x}^2\right]
\end{aligned}$$

$$\begin{aligned}
&= \frac{1}{n-1}\left[\sum_{j=1}^{n} x_j^2 \cos 2\psi + \sum_{j=1}^{n} y_j x_j \sin 2\psi - n\hat{\beta}_{0m}\bar{x}\sin 2\psi - n\bar{x}^2\right] \\
&= \frac{1}{n-1}\left[\sum_{j=1}^{n} x_j^2 \cos 2\psi + \left(\sum_{j=1}^{n} y_j x_j - n\bar{y}\bar{x}\right)\sin 2\psi\right] \\
&\quad + \frac{1}{n-1}\left[n\hat{\beta}_{1x}\bar{x}^2 \sin 2\psi - n\bar{x}^2\right] \\
&= \frac{1}{n-1}\left[\sum_{j=1}^{n} x_j^2 \cos 2\psi + SP_{xy}\sin 2\psi + 2n\bar{x}^2\sin^2\psi - n\bar{x}^2\right] \\
&= \frac{1}{n-1}\left[\sum_{j=1}^{n} x_j^2 \cos 2\psi + SP_{xy}\sin 2\psi + n\bar{x}^2(2\sin^2\psi - 1)\right] \\
&= \frac{1}{n-1}\left[\sum_{j=1}^{n} x_j^2 \cos 2\psi + SP_{xy}\sin 2\psi - n\bar{x}^2\cos 2\psi\right] \\
&= \frac{1}{n-1}(SS_x\cos 2\psi + SP_{xy}\sin 2\psi) \\
&= S_x^2\cos 2\psi + S_{xy}\sin 2\psi,
\end{aligned}$$

where $\hat{\beta}_{1x} = \dfrac{S_{xy}}{S_x^2} = \dfrac{\sin\psi}{\cos\psi}$, $\sin 2\psi = 2\cos\psi\sin\psi$ and $\cos 2\psi = \cos^2\psi - \sin^2\psi$ then

$$\begin{aligned}
\hat{cov}(x,x^*) &= S_x^2\cos 2\psi + 2S_x^2\sin^2\psi \\
&= S_x^2(\cos 2\psi + 2\sin^2\psi) \\
&= S_x^2(\cos^2\psi - \sin^2\psi + 2\sin^2\psi) = S_x^2(\cos^2\psi + \sin^2\psi) = S_x^2.
\end{aligned}$$

Based on the theorem above one could conclude that the sample coefficient correlation between the manifest explanatory variable x and its reflection x^* is given by

$$r_{xx^*} = \frac{\hat{cov}(x,x^*)}{S_x S_{x^*}} = \frac{S_x^2}{S_x S_{x^*}} = \frac{S_x}{S_{x^*}}.$$

Moreover, the ordinary least squares estimator of slope parameter of x on its reflection x^* is given by

$$\hat{\beta}_{1xx*} = \frac{S_{xx*}}{S^2_{x^*}} = \frac{S^2_x}{S^2_{x^*}} = r^2_{xx^*}.$$

Theorem 7 *The sample covariance of the response variable y and the reflection variable x^* is greater than that of the response variable y and the manifest variable x.*

$$\mid S_{x^*y} \mid \geq \mid S_{xy} \mid$$

Proof From (3.1) and by subtracting x_j we get

$$\begin{aligned}
(x^*_j - x_j) &= x_j \cos 2\psi + (y_j - \hat{\beta}_{0x}) \sin 2\psi - x_j \\
&= x_j(\cos 2\psi - 1) + y_j \sin 2\psi - \hat{\beta}_{0x} \sin 2\psi \\
&= -x_j(2\sin^2\psi) + y_j \sin 2\psi - \bar{y} \sin 2\psi + \bar{x}2\sin^2\psi \\
&= (y_j - \bar{y}) \sin 2\psi - (x_j - \bar{x})2\sin^2\psi,
\end{aligned}$$

where x^* is the reflection of x. Multiplying both sides of the above equation by y and taking the sum over j, we obtain

$$\begin{aligned}
\sum_{j=1}^{n}(x^*_j - x_j)y_j &= \sum_{j=1}^{n}(y_j - \bar{y})y_j \sin 2\psi - \sum_{j=1}^{n}(x_j - \bar{x})y_j 2\sin^2\psi \\
\sum_{j=1}^{n} y_j x^*_j - \sum_{j=1}^{n} y_j x_j &= \sum_{j=1}^{n}(y_j - \bar{y})y_j \sin 2\psi - \sum_{j=1}^{n}(x_j - \bar{x})y_j 2\sin^2\psi,
\end{aligned}$$

by adding and subtracting $n\bar{y}\bar{x}$ to the left hand side we then have

$$\begin{aligned}
\sum_{j=1}^{n} y_j x^*_j - \sum_{j=1}^{n} y_j x_j &= \left(\sum_{j=1}^{n} y_j x^*_j - n\bar{y}\bar{x}\right) - \left(\sum_{j=1}^{n} y_j x_j - n\bar{y}\bar{x}\right) \\
&= SP_{x^*y} - SP_{xy}.
\end{aligned}$$

Then

$$\begin{aligned} SP_{x^*y} - SP_{xy} &= \sum_{j=1}^{n}(y_j - \bar{y})y_j \sin 2\psi - \sum_{j=1}^{n}(x_j - \bar{x})y_j 2\sin^2\psi \\ &= SS_y \sin 2\psi - SP_{xy} 2\sin^2\psi, \end{aligned}$$

where $\sum_{j=1}^{n}(y_j - \bar{y})y_j = SS_y$, and $\sum_{j=1}^{n}(x_j - \bar{x})y = SP_{xy}$.

Hence

$$SP_{x^*y} - SP_{xy} = SS_y \sin 2\psi - SP_{xy} 2\sin^2\psi.$$

Now dividing both sides by $n-1$ yields

$$S_{x^*y} - S_{xy} = S_y^2 \sin 2\psi - 2S_{xy}\sin^2\psi.$$

Note that

$$\frac{2\sin^2\psi}{\sin 2\psi} = \tan\psi = \hat{\beta}_{1x},$$

and

$$2\sin^2\psi = \hat{\beta}_{1x}\sin 2\psi.$$

Then we obtain

$$S_{x^*y} - S_{xy} = (S_y^2 - \hat{\beta}_{1x}S_{yx})\sin 2\psi,$$

$$\begin{aligned} S_{x^*y} - S_{xy} &= S_v^2 \sin 2\psi \\ &= \frac{2S_v^2 S_{yx}\cos^2\psi}{S_x^2}. \end{aligned} \tag{3.8}$$

Hence

$$S_{x^*y} = S_{xy} + \frac{2S_v^2 S_{yx}\cos^2\psi}{S_x^2} = S_{yx}\left(1 + \frac{2S_v^2\cos^2\psi}{S_x^2}\right) \tag{3.9}$$

where S_v^2 is the sum of squares residuals, $\hat{\beta}_{1x} = \dfrac{S_{yx}}{S_x^2} = \dfrac{\sin\psi}{\cos\psi}$ and

$\sin\psi = \dfrac{S_{yx}\cos\psi}{S_x^2}$. Obviously, from (3.5) we see that $\mid S_{x^*y} \mid \geq \mid S_{xy} \mid$.

Theorem 8 *The sample variance S_x^2 of manifest explanatory variable x is less than that of its reflection variable $S_{x^*}^2$, and the difference between both of them multiplied by $\dfrac{1}{4\sin^2\psi}$ is equal the sum of squares orthogonal distances $\dfrac{S_v^2}{1+\hat{\beta}_{1x}^2}$.*

$$S_x^2 \leq S_{x^*}^2, \quad \textit{and}$$

$$\frac{S_{x^*}^2 - S_x^2}{4\sin^2\psi} = \frac{S_v^2}{1+\hat{\beta}_{1x}^2}$$

Proof By definition

$$S_{x^*}^2 = \frac{1}{n-1}\sum_{j=1}^{n}(x_j^* - \bar{x}^*)^2.$$

Now from equation (3.1)

$$x_j^* = x_j\cos 2\psi + y_j\sin 2\psi - \hat{\beta}_{0x}\sin 2\psi \tag{3.10}$$

By multiplying both sides of (3.10) by x^* and taking the summation over j we then have

$$\begin{aligned}
\sum_{j=1}^{n} x_j^{*2} &= \sum_{j=1}^{n} x_j x_j^*\cos 2\psi + \sum_{j=1}^{n} y_j x_j^*\sin 2\psi - \hat{\beta}_{0x}\sin 2\psi\sum_{j=1}^{n} x_j^* \\
&= \sum_{j=1}^{n} x_j x_j^*\cos 2\psi + \sum_{j=1}^{n} y_j x_j^*\sin 2\psi - n\bar{y}\bar{x}\sin 2\psi + n\hat{\beta}_{1x}\bar{x}^2\sin 2\psi \\
&= \sum_{j=1}^{n} x_j x_j^*\cos 2\psi + \sum_{j=1}^{n} y_j x_j^*\sin 2\psi - n\bar{y}\bar{x}\sin 2\psi + 2n\bar{x}^2\sin^2\psi
\end{aligned}$$

Note that $2\sin^2\psi = 1 - \cos 2\psi$ and from theorem 2 that $\bar{x} = \bar{x}^*$ then

$$\begin{aligned}\sum_{j=1}^{n} x_j^{*2} &= \sum_{j=1}^{n} x_j x_j^* \cos 2\psi + (\sum_{j=1}^{n} y_j x_j^* - n\bar{y}\bar{x}) \sin 2\psi - n\bar{x}^2 \cos 2\psi + n\bar{x}^2 \\ &= (\sum_{j=1}^{n} x_j x_j^* - n\bar{x}^2) \cos 2\psi + (\sum_{j=1}^{n} y_j x_j^* - n\bar{y}\bar{x}) \sin 2\psi + n\bar{x}^2,\end{aligned}$$

which gives

$$\sum_{j=1}^{n} x_j^{*2} - n\bar{x}^2 = (\sum_{j=1}^{n} x_j x_j^* - n\bar{x}^2) \cos 2\psi + (\sum_{j=1}^{n} y_j x_j^* - n\bar{y}\bar{x}) \sin 2\psi,$$

where $\sum_{j=1}^{n} x_j x_j^* - n\bar{x}^2 = SP_{xx^*}$ then

$$SS_{x^*} = SP_{xx^*} \cos 2\psi + SP_{yx^*} \sin 2\psi,$$

by dividing both sides by $n-1$, then we get

$$S_{x^*}^2 = S_{xx^*} \cos 2\psi + S_{yx^*} \sin 2\psi.$$

From Theorem 7, $S_{xx^*} = S_x^2$ then

$$S_{x^*}^2 = S_x^2 \cos 2\psi + S_{yx^*} \sin 2\psi.$$

From the equation (3.4) that

$$S_{x^*}^2 = S_x^2 \cos 2\psi + S_{yx} \sin 2\psi + S_v^2 \sin^2 2\psi.$$

Note that $S_{yx} \sin 2\psi = 2S_{yx} \sin\psi \cos\psi = 2\hat{\beta}_{1x} S_x^2 \sin\psi \cos\psi = 2S_x^2 \sin^2\psi$. Then

$$\begin{aligned}S_{x^*}^2 &= S_x^2 \cos 2\psi + 2S_x^2 \sin^2\psi + S_v^2 \sin^2 2\psi \\ &= S_x^2(\cos 2\psi + 2\sin^2\psi) + S_v^2 \sin^2 2\psi \\ &= S_x^2(\cos^2\psi - \sin^2\psi + 2\sin^2\psi) + S_v^2 \sin^2 2\psi \\ &= S_x^2 + S_v^2 \sin^2 2\psi. \qquad (3.11)\end{aligned}$$

It is then clear that $S_x^2 \leq S_{x^*}^2$.

Note that the sample variance of explanatory variable x equals the sample variance of its reflection variable x^* if and only if the sum squares residuals S_v^2 equals zero. In fact, that means all the data points must be on the straight line.

From (3.11) we then have

$$\begin{aligned} S_{x^*}^2 - S_x^2 &= S_v^2 \sin^2 2\psi \\ &= 4S_v^2 \sin^2 \psi \cos^2 \psi. \end{aligned}$$

Then

$$\begin{aligned} \frac{S_{x^*}^2 - S_x^2}{4\sin^2 \psi} &= S_v^2 \cos^2 \psi \\ &= \frac{S_v^2}{1 + \hat{\beta}_{1x}^2}, \end{aligned}$$

where $\sin^2 2\psi = 4\sin^2 \psi \cos^2 \psi$ and $\cos^2 \psi = \dfrac{1}{1 + \hat{\beta}_{1x}^2}$.

Theorem 9 *The difference between the sum of squares of the reflection variable $SS_{x^*}^2$ and the sum of squares of the manifest explanatory variable SS_x^2 is given by*

$$SS_{x^*} - SS_x = TSS - SSR - SSE_y.$$

Proof From Theorem 1 we have

$$SSE = \sum_{j=1}^{n}(y_j - \hat{y}_j)^2 = \sum_{j=1}^{n}(y_j^* - \hat{y}_j)^2 + \sum_{j=1}^{n}(x_j^* - x_j)^2,$$

this means

$$\sum_{j=1}^{n}(x_j^* - x_j)^2 \;=\; SSE - \sum_{j=1}^{n}(y_j^* - \hat{y}_j)^2,$$

by adding and subtracting $\bar{x}$ to the left hand side of the equation, we have

$$\begin{aligned}\sum_{j=1}^{n}(x_j^* - x_j)^2 &= \sum_{j=1}^{n}\big((x_j^* - \bar{x}) - (x_j - \bar{x})\big)^2 \\ &= \sum_{j=1}^{n}(x_j^* - \bar{x})^2 - 2\sum_{j=1}^{n}(x_j^* - \bar{x})(x_j - \bar{x}) + \sum_{j=1}^{n}(x_j - \bar{x})^2.\end{aligned}$$

Note from Theorem 7, $\sum_{j=1}^{n}(x_j^* - \bar{x})(x_j - \bar{x}) = SP_{xx^*} = SS_x$ then

$$\begin{aligned}\sum_{j=1}^{n}(x_j^* - x_j)^2 &= \sum_{j=1}^{n}(x_j^* - \bar{x})^2 - 2SP_{xx^*} + \sum_{j=1}^{n}(x_j - \bar{x})^2 \\ &= SS_{x^*} - 2SS_x + SS_x \\ &= SS_{x^*} - SS_x.\end{aligned}$$

Hence from (3.4) and (3.6) we have the final result

$$SS_{x^*} - SS_x = SSE - SSE_y = TSS - SSR - SSE_y.$$

The above theorems are not only valid to apply on sample data but also to population data. The reflection of all points (x_j, y_j) about the OLS regression line of y on x produces reflection points (x_j^*, y_j^*). In this chapter, we obtained the transformed reflection variables x^* and y^* for both x and y respectively. It allows us to define the sum of squares error uniquely, in the same way as in the case of no measurement error.

3.5 Concluding remarks

This chapter explains that the proposed methodology relies on the combination of the reflection and ordinary least squares techniques. We defined the reflection variables for both the explanatory and the response variables. The proposed methodology has useful properties which allows the analysis of the mathematical relationships between the manifest variables and the transformed variables. The proposed methodology will be used to develop some statistical methods to deal with the estimation of regression parameters when both response and explanatory variables are subject to measurement error. The theorems of this chapter help interpret vertical, orthogonal, and horizontal distances between the observed points and regression line. The applications of the theorems will be included in the next chapters to define proposed estimator and study its properties.

Chapter 4

Instrumental variable estimator for measurement error model

4.1 Introduction

This chapter proposes an instrumental variable (IV) estimator for the parameters of a simple linear regression model which includes an equation error and the explanatory variable is subject to measurement error. Based on the previous chapter, the instrumental variable is defined using reflection of the observed values of the explanatory variable. Like other instrumental variable estimators, it is unbiased and consistent but under one assumption mentioned in Section 4.4 about the ratio of the vertical and horizontal error.

The proposed modified method uses the reflection of the *manifest* values of the explanatory variable to define IV estimator. The use of the reflections of the observed values of the explanatory variable in defining the IV method provides a much better estimator of the slope and intercept parameters. It also reduces the mean sum of squares error. The analysis of variance and regression inferences based on the reflections have much better statistical properties than any other form of the IV estimator.

In the next section the measurement error regression model is introduced. Section 4.3 covers the existing estimation methods for the measurement error model. The proposed modified estimator based on the reflections of the observed values of the explanatory variable is provided in Section 4.4. The superior properties of the modified estimator are discussed in Section 4.5. Two numerical illustrations are provided in Section 4.6, and some concluding remarks are given in Section 4.7.

4.2 Measurement error models

In the conventional notation, let ξ_j denote the true measurement on the explanatory variable. This is also called the *latent* explanatory variable. In the presence of measurement error the actual observations are different from ξ_j. Let x be the observable, or *manifest* variable of the explanatory variable.

When the true value of the *latent* variable ξ_j is observed, the commonly used classical simple linear regression model is represented by

$$\eta_j \quad = \quad \beta_{0\xi} + \beta_{1\xi}\xi_j + e_j, \quad j = 1, 2, \ldots, n, \tag{4.1}$$

where η_j is the jth realisation of the latent response variable, ξ_j is the fixed jth value of the explanatory variable, and e_j is the equation error for $j = 1, 2, \ldots, n$. It is assumed that the equation error e_j is independently distributed with constant but unknown variance, that is, $e_j \sim N(0, \sigma_e^2)$.

If there is error in the explanatory variable, the actual observed value, x_j, is not the 'true' value of the explanatory variable. The observed value of the explanatory variable contains measurement error given as

$$x_j = \xi_j + \delta_j, \quad j = 1, 2, \ldots, n, \tag{4.2}$$

where δ_j is the measurement error, and is assumed to be distributed as $N(0, \sigma_\delta^2)$. Note that, unlike ξ_j, x_j is a random variable which is assumed to be distributed as $N(\mu_x, \sigma_x^2)$. The model with the fixed ξ_j is called the *functional model*, and the model with the random or stochastic x is called the *structural model.*

The simple regression model with measurement error in the explanatory variable can be expressed as

$$\eta_j \quad = \quad \beta_{0\xi} + \beta_{1\xi}x_j + v_j, \qquad j = 1, 2, \ldots, n, \tag{4.3}$$

where $v_j = e_j - \beta_{1\xi}\delta_j$. Note in equation (4.1) ξ_j and e_j are independent, but in equation (4.3), x_j and v_j are not independent. So the application of least squares method is not valid for the models with measurement error. Thus, unlike for the model in (4.1), the validity of the estimator of the slope and intercept of the model in (4.3) is not obvious. However, Fuller (2006, p. 3) assumes that δ_j, ξ_j and e_j are mutually independent for the estimation of the parameters. It also assumes that the *reliability ratio*, $k_{x\xi} = \sigma_x^{-2}\sigma_\xi^2$ is known, where σ_x^2 is the variance of the *manifest* variable x_j, and σ_ξ^2 is the variance of the *latent* variable ξ_j.

4.3 Existing Estimators of parameters

The ordinary least squares (OLS) estimator of the regression parameters for the *functional model* are

$$\hat{\beta}_{1\xi} = \frac{S_{\xi\eta}}{S_\xi^2}, \text{ and } \hat{\beta}_{0\xi} = \bar{\eta} - \hat{\beta}_{1\xi}\bar{\xi}, \tag{4.4}$$

where

$$S_{\xi\eta} = \frac{1}{n-1}\sum_{j=1}^{n}(\xi_j - \bar{\xi})(\eta_j - \bar{\eta}), \qquad S_\xi^2 = \frac{1}{n-1}\sum_{j=1}^{n}(\xi_j - \bar{\xi})^2, \tag{4.5}$$

in which $\bar{\eta} = \frac{1}{n}\sum_{j=1}^{n}\eta_j$ and $\bar{\xi} = \frac{1}{n}\sum_{j=1}^{n}\xi_j$. The estimators of slope and intercept parameters are linear functions of the responses, and they are well known to be the best linear unbiased estimators if there is no measurement error in the variables.

The sampling distribution of the estimator of the regression parameters is given by

$$\begin{pmatrix} \hat{\beta}_{0\xi} \\ \\ \hat{\beta}_{1\xi} \end{pmatrix} \sim N_2 \left[\begin{pmatrix} \beta_{0\xi} \\ \\ \beta_{1\xi} \end{pmatrix}, \sigma_e^2 \begin{pmatrix} \frac{1}{n} + \frac{\bar{\xi}^2}{S_\xi^2} & \frac{-\bar{\xi}}{S_\xi^2} \\ \\ \frac{-\bar{\xi}}{S_\xi^2} & \frac{1}{S_\xi^2} \end{pmatrix} \right]. \tag{4.6}$$

The unbiased estimator of the error variance σ_e^2 is given by

$$\hat{\sigma}_e = (n-2)^{-1} SSE_e = S_e^2,$$

where

$$SSE_e = \sum_{j=1}^{n} (\eta_j - \hat{\eta}_j)^2,$$

in which $\hat{\eta}_j = \hat{\beta}_{0\xi} + \hat{\beta}_{1\xi}\xi_j$ is the estimated value of η_j. Also, $\sigma_e^{-2} SSE_e$ follows a χ^2 distribution with $(n-2)$ degrees of freedom.

In the presence of measurement error, the x values are observed instead of ξ_j, then the least squares method yields the estimator of the slope as

$$\hat{\beta}_{1x} = \frac{S_{x\eta}}{S_x^2}, \text{ and } \hat{\beta}_{0x} = \bar{\eta} - \hat{\beta}_{1x}\bar{x}. \tag{4.7}$$

It can be easily shown that $\hat{\beta}_{1x}$ is a biased estimator of $\beta_{1\xi}$. Also, the above estimator is not a consistent estimator of $\beta_{1\xi}$.

Note that the regression parameters are different for the model with the *manifest* variable than the model with the *latent* variable. Even though the aim is to estimate and test $\beta_{0\xi}$ and $\beta_{1\xi}$, but in reality one may end up estimating and testing β_{0x} and β_{1x} if one fully relies upon x, and over looks the presence of the measurement error.

4.3.1 Instrumental variable (IV) estimator

In the presence of measurement error in the explanatory variable the IV estimator for the regression parameters is defined as

$$\hat{\boldsymbol{\beta}} = (z'x)^{-1}z'\eta, \tag{4.8}$$

where $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1)'$ is the vector of estimator of the intercept and slope parameters of the model where

$$x = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ \\ x_1 & x_2 & \cdots & x_n \end{pmatrix} \text{ and } z = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ \\ z_1 & z_2 & \cdots & z_n \end{pmatrix},$$

in which z_j's are the values of the second row of the instrumental variable z. The selection of the values of z_j's require that it is highly correlated with the explanatory variable but uncorrelated with the model errors. The variance-covariance of the above estimator vector is given by

$$\text{var}(\hat{\boldsymbol{\beta}}) = \sigma_\delta^2(\mathrm{z'x})^{-1}(\mathrm{z'z})(\mathrm{z'x})^{-1}. \tag{4.9}$$

Obviously the value of the estimator and the variance depend on the choice of z (see Johnson, 1972). For instance, the Wald method, as suggested by

Maddala (1988), defines z by assigning z_j to be -1 or +1 depending upon if x_j is smaller or larger than the median value of the *manifest* variable. The estimator of slope under this choice of IV is

$$\hat{\beta}_{1W} = \frac{\bar{\eta}_2 - \bar{\eta}_1}{\bar{x}_2 - \bar{x}_1},$$

where $\bar{\eta}_1$ is the mean of η-values associated with the values of x less than its median, and $\bar{\eta}_2$ is for the mean values larger than the median value of η. Bartlett (1949) followed the same selection criterion of z_j's but suggested the exclusion of the middle 1/3 of the values, and his estimator is based on the lower and upper 1/3 of the values of x and the associated $\eta's$. The estimator is expressed as

$$\hat{\beta}_{1B} = \frac{\bar{\eta}_3 - \bar{\eta}_1}{\bar{x}_3 - \bar{x}_1},$$

where $\bar{\eta}_1$ is the mean of η-values associated with the smallest 1/3 of the values of x, and $\bar{\eta}_3$ is that for the largest 1/3. Durbin (1954) proposed to use the rank of x as z_j's. His method yields the following estimator of the slope parameter

$$\hat{\beta}_{1D} = \left[\sum_{j=1}^{n} j\eta_j\right] / \left[\sum_{j=1}^{n} jx_j\right],$$

but does not define the estimator of the intercept.

The IV method of estimation of the regression parameters does not require any strict assumptions like the ratio of error variances is known. But the actual estimator depends on how the IV is defined, as the definition of z

affects both the estimator and its variance. In general, the available methods of defining IV causes a significant loss of sample information (data) either by replacing the observed values of the explanatory variable by -1 or +1, or exclusion of some data, or due to ranking of data. But the proposed definition of the IV does not lose any information. Furthermore, the method produces a more precise IV estimator than those proposed by Wald, Bartlett, and Durbin.

4.4 Proposed new IV estimator

To avoid the unwanted and troublesome influence of the measurement error in the explanatory variable, the idea of *reflection* of the manifest variable is used for all values of the explanatory variable. The *reflection* of the points is taken about the fitted regression line. This is essentially done by a transformation of the observed values of the explanatory variable to their reflection on the Euclidean plane. In the conventional notation, the reflection of the explanatory variable $x_j = \xi_j + \delta_j$ (with measurement error δ_j) for $j = 1, 2, \ldots, n$, can be defined as

$$x^* = x \cos 2\psi + (\eta - \hat{\beta}_{0x}) \sin 2\psi, \tag{4.10}$$

where $\hat{\beta}_{0x}$ is the least squares estimate of the intercept parameter, ψ is the angle measure defined as $\psi = \arctan \hat{\beta}_{1x}$ in which $\hat{\beta}_{1x}$ is the least squares estimate of the slope parameter in the *manifest* model, and cos and sin are the usual trigonometric cosine and sine functions respectively. For the definition of *reflection* points on the Cartesian plane readers may see Vaisman (1997, p. 164-169).

The proposed reflection method requires to compute the reflection of all data points, and the use of the transformed values of x, i.e. x^*, in defining the IV to fit the regression line of η. The IV estimator of the slope parameter under the proposed modified method is

$$\hat{\beta} \;=\; (z_r'x)^{-1}z_r'\eta, \text{ and } \hat{\beta}_{1R} \;=\; \frac{S_{x^*\eta}}{S_x^2},$$

where

$$Z_r = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ \\ x_1^* & x_2^* & \cdots & x_n^* \end{pmatrix} \text{ and } S_{x^*x} = S_x^2,$$

in which $S_{x^*x} = \sum_{j=1}^{n}(x_j^* - \bar{x^*})(x_j - \bar{x})$.

The proposed estimator of the slope parameter of the simple regression model using IV based on the *reflection* of x is

$$\hat{\beta}_{1R} \;=\; \frac{S_{x^*\eta}}{S_x^2},$$

$$\hat{\beta}_{1\xi} = \frac{S_{\xi\eta}}{S_\xi^2} = \frac{S_{x\eta}}{S_\xi^2} \quad \text{and} \quad \hat{\beta}_{1R} = \frac{S_{x^*\eta}}{S_x^2}. \tag{4.11}$$

From (4.11), it is easy to show that $S_{xy} = S_{\xi y}$ and $S_x^2 = S_\xi^2 + S_\delta^2$. It can be found that

$$S_{x^*y} - S_{xy} = SSE_x \sin 2\psi, \tag{4.12}$$

where ψ is as defined in equation (4.10), and SSE_x is the sum of squares error for the *manifest* model. The above result follows from the fact that

$$\begin{aligned} x_j^* - x_j &= x_j \cos 2\psi + (\eta_j - \hat{\beta}_{0x}) \sin 2\psi - x_j \\ &= x_j(\cos 2\psi - 1) + \eta_j \sin 2\psi - \hat{\beta}_{0x} \sin 2\psi \\ &= -x_j(2\sin^2\psi) + \eta_j \sin 2\psi - \bar{\eta} \sin 2\psi + \bar{x} 2 \sin^2 \psi \\ &= (\eta_j - \bar{\eta}) \sin 2\psi - (x_j - \bar{x}) 2 \sin^2 \psi, \end{aligned} \tag{4.13}$$

where x_j^* is the reflection of x_j. Multiplying both sides of the above equation by η_j and taking sum over j, yields

$$\begin{aligned} \sum (x_j^* - x_j)\eta_j &= \sum (\eta_j - \bar{\eta})\eta_j \sin 2\psi - \sum (x_j - \bar{x})\eta_j 2 \sin^2 \psi \\ S_{x^*\eta} - S_{x\eta} &= S_\eta^2 \sin 2\psi - S_{x\eta} 2 \sin^2 \psi \\ \frac{S_{x^*\eta} - S_{x\eta}}{\sin 2\psi} &= SST - SSR_x = SSE_x, \end{aligned} \tag{4.14}$$

where $S_\eta^2 = SST$ is the sum of squares total, SSR_x is the sum of squares regression, and SSE_x is the sum of squares error for the regression of η on x. Note that $\frac{2\sin^2\psi}{\sin 2\psi} = \tan\psi = \hat{\beta}_{1x}$.

Then using equation (4.11), it can be written as

$$\begin{aligned} \hat{\beta}_{1\xi} &= \frac{S_{\xi\eta}}{S_\xi^2} = \frac{S_{x\eta}}{S_\xi^2} = \frac{S_{x^*\eta} - SSE_x \sin 2\psi}{S_x^2 - S_\delta^2} \\ \hat{\beta}_{1R} &= \frac{S_{x^*\eta}}{S_x^2} = \frac{S_{x\eta} + SSE_x \sin 2\psi}{S_\xi^2 + S_\delta^2} = \frac{S_{\xi\eta} + SSE_x \sin 2\psi}{S_\xi^2 + S_\delta^2}. \end{aligned}$$

Let λ^* be the ratio of the vertical v_j and horizontal δ_j error variances, where $\lambda^* = \frac{\sigma_v^2}{\sigma_\delta^2}$.

Based on the assumption $\lambda^* = \frac{S_{x^*\eta}}{S_x \sin 2\psi}$, then

$$\hat{\beta}_{1R} = \frac{S_{x^*\eta}}{S_x^2} = \frac{S_{x^*\eta} - S_{\xi\eta}}{S_x^2 - S_\xi^2} \tag{4.15}$$

$$S_{x^*\eta}(S_x - S_\xi) \quad = \quad S_x^2(S_{x^*\eta} - S_{\xi\eta})$$

which leads to

$$S_{x^*\eta} S_\xi^2 = S_{\xi\eta} S_x, \tag{4.16}$$

and finally simplification yields

$$\frac{S_{x^*\eta}}{S_x^2} = \frac{S_{\xi\eta}}{S_\xi^2}, \text{ hence } \hat{\beta}_{1R} = \hat{\beta}_{1\xi}. \tag{4.17}$$

4.4.1 Geometric Explanation

The presence of measurement error in the explanatory variable and its impact on the estimator of the slope as well as how the proposed method ‘treats’ the measurement error can be explained by graphs. The graphical representation also explains how the actual estimator of the slope is recovered by the new method.

Figure 4.1 represents the sum of squares and sum of products associated with the definition of the estimators of slope both for the *latent* and *manifest* variables. This graph represents the presence of measurement error in the explanatory variable as well as the two estimators of the slope parameter. On the other hand Figure 4.2 displays the same along with that of the reflection of the *manifest* variable and three estimators of the slope parameter.

From Figure 4.1, the true estimator of the slope when the *latent* variable is available, that is, $\hat{\beta}_{1\xi}$ is represented by the tan of $\angle BAC$ of $\triangle ABC$. In the absence of the values of the *latent* variable this is unavailable. But for the *manifest* variable one can find the estimator of the slope to be $\hat{\beta}_{1x}$ which is represented by the tan of $\angle DAE$ of $\triangle ADE$. Note that here DC (or equivalently BE) represents the sum of squares of measurement error (S^2_δ). Furthermore, under the assumptions of $E[\eta\delta] = 0$ and $E[\xi\delta] = 0$, we have $BC = DE$ or $S_{\xi\eta} = S_{x\eta}$. Finally, $\hat{\beta}_{1\xi} = \frac{S_{\xi\eta}}{S^2_\xi} = \frac{BC}{AC}$, and $\hat{\beta}_{1x} = \frac{S_{xy}}{S^2_x} = \frac{ED}{AD}$.

The introduction of the reflection of the manifest variable changes $\triangle ADE$ of Figure 4.1 to $\triangle ADF$ in Figure 4.2. In fact the main difference between the two Figures is that Figure 4.2 has the small $\triangle BEF$ added to Figure 4.1. This triangle represents the effect of the reflection of the manifest variable. From Figure 4.2 the estimates of the slope are

$$\hat{\beta}_{1x} = \frac{S_{x\eta}}{S^2_x}\left(= \frac{DE}{DA}\right) \tag{4.18}$$

$$\hat{\beta}_{1\xi} = \frac{S_{\xi\eta}}{S^2_\xi}\left(= \frac{BC}{AC}\right) \tag{4.19}$$

$$\hat{\beta}_{1R} = \frac{S_{x^*\eta}}{S^2_x}\left(= \frac{FD}{AD}\right). \tag{4.20}$$

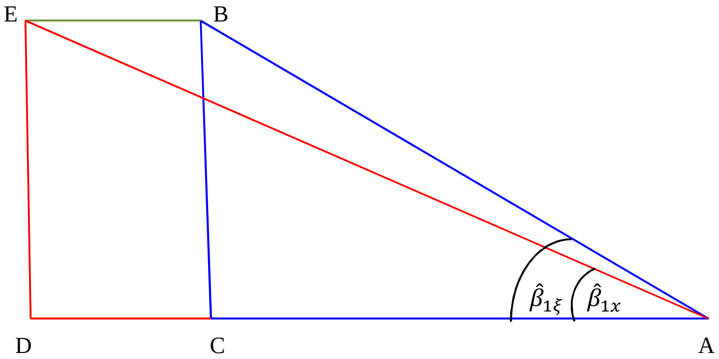

Figure 4.1: Graph representing the sum of squares and products in the presence of measurement error in the explanatory variable.

Since the tan of $\angle BAC$ represents the estimator $\hat{\beta}_{1\xi}$ and tan of $\angle DAF$ represents $\hat{\beta}_{1R}$, then $\hat{\beta}_{1\xi} = \hat{\beta}_{1R}$ because $\angle BAC = \angle DAF$.

4.5 Some properties and relationships

The estimated regression lines based on the OLS, and three IV methods are summarised in the following way:

$$\hat{\eta}_{\xi} = \hat{\beta}_{0\xi} + \hat{\beta}_{1\xi}\xi, \tag{4.21}$$

$$\hat{\eta}_{R} = \hat{\beta}_{0R} + \hat{\beta}_{1R}x, \tag{4.22}$$

$$\hat{\eta}_{W} = \hat{\beta}_{0W} + \hat{\beta}_{1W}x, \tag{4.23}$$

$$\hat{\eta}_{B} = \hat{\beta}_{0B} + \hat{\beta}_{1B}x. \tag{4.24}$$

Obviously, in the absence of ξ_j, the fitted model in (4.21) is unavailable. The other fitted lines are obtainable since the manifest variable x is always

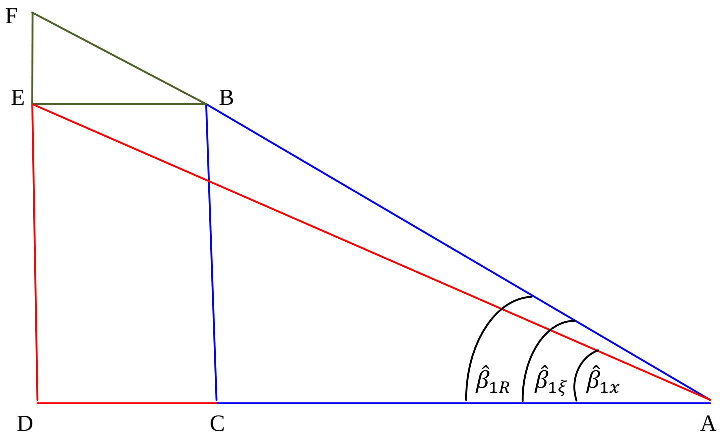

Figure 4.2: Graph representing the sum of squares and products when the measurement error in the explanatory variable is 'treated' by reflection.

observed along with the response η. Furthermore, even though the regression parameters are the same, the estimated models are different since the observed x is different from the true value of the explanatory variable ξ_j. Thus

$$\hat{\beta}_{0\xi} + \hat{\beta}_{1\xi}\xi \quad \neq \quad \hat{\beta}_{0\xi} + \hat{\beta}_{1\xi}x.$$

Another useful fact is that the sum of squares total is the same for regression of η_j on ξ_j and that on x_j. That is,

$$SS_\eta \quad = \quad SSR_\xi + SSE_\xi = SSR_x + SSE_x.$$

Similarly, the following relationship of the regression sum of squares for models using ξ_j, x_j, and x^* under the above assumption are observed:

$$\begin{aligned} SSR_\xi \quad &= \quad \hat{\beta}_{1\xi}SS_{\eta\xi} = \hat{\beta}_{1R}SS_{x\eta} \\ &= \quad \hat{\beta}^2_{1R}SS_{xx} = \hat{\beta}_{1R}SS_{x^*y} = SSR_R. \end{aligned}$$

Finally, the coefficient of determination is noted to be

$$R_\xi^2 = \frac{SSR_\xi}{SST} = \frac{SSR_R}{SST}.$$

4.6 Illustration

In this section, two illustrative examples based on two real life datasets are provided. The first dataset has measurement error in the explanatory variable only, but the second dataset has measurement error in both the response and explanatory variables. For the second example it is assumed that the ratio of error variances, $\lambda = \frac{\sigma_\epsilon^2}{\sigma_\delta^2} < 1$ is known, where σ_ϵ^2 is the error variance of the response variable and σ_δ^2 is the error variance of the explanatory variable.

4.6.1 Yield of Corn Data

The dataset of the first example deals with the yield of corn (η) for different levels of soil nitrogen (x), and is taken from Fuller (2006, p. 18). Here the explanatory variable x, soil nitrogen level, has been found with measurement error. Fuller has analysed the data under assumption that the *reliability ratio*, ($k_{\xi x}$), is known. We provide the regression analyses of the data for

Table 4.1: Fitted regression models for the corn yield data

Method	Fitted regression equation	MS Error	R^2
OLS	$\hat{Y}_x = 73.153 + 0.344x$	57.321	0.412
Wald	$\hat{Y}_W = 75.91 + 0.305x$	60.98	0.364
Bartlett	$\hat{Y}_B = 72.38 + 0.355x$	56.05	0.425
Reflection	$\hat{Y}_R = 65.8164 + 0.4479x$	45.224	0.536
σ_δ^2 Known	$\hat{Y}_V = 67.561 + 0.423x$	48.125	0.506

both with (a) the measurement error in the explanatory variable x, and (b) the instrumental variables including one defined by x^*, the reflection of the observed explanatory variable x. Comparison of estimators of the regression parameters and related results from different methods are provided below. Table 4.1 below shows the fitted regression lines, mean sum of squares error, and the coefficient of determination based on the OLS and various IV methods including the *reflection* method.

The OLS regression of η on x produces the estimated regression line, $\hat{Y}_x = 73.153 + 0.344x$ with mean sum of squares error, $MSE_x = 57.321$ (see the first regression line in Table 4.1) and $R_x^2 = 0.421$. This analysis does not take into account the presence, and hence the effect, of the measurement errors in the explanatory variable. As such these results are not based on any sound statistical method and hence unacceptable.

Fuller (2006, p. 18-19) assumes that $\sigma_\delta^2 = 57$, and that the *reliability ratio*, $k_{\xi x}$, is known. Under the assumption the estimated regression line is reported to be $\hat{\eta}_V = 67.561 + 0.423x$ with modified MSE, $MSE_V = \hat{\sigma}_\epsilon^2 = 48.125$, and $R_V^2 = 0.506$. Clearly, there has been an improvement in the proportion of variability in η that is explained by x under the method used by Fuller (2006). The MSE has also decreased (from $MSE_x = 57.321$ to $MSE_V = 48.125$) under the Fuller method. Thus the Fuller method is not only a better method than the OLS, but also provides a much better fit.

The use of the reflection of x in the specification of the instrumental variable leads to the fitted regression line, $\hat{Y}_R = 65.8164 + 0.4479x$ with a mean sum of squares error, $MSE_R = 45.224$ (see second last row of Table 4.1) and $R_R^2 = 0.536$. Unlike Fuller's method, these results are obtained without additional assumptions on any of the parameters of the model or the reliability ratio. However, the modified IV estimator obtained by the reflection of x are fairly close to those obtained by Fuller under the previously stated assumption. The regression line produced by the modified IV method provides a much better fit than that obtained by Fuller. Obviously, the MSE_R under the modified IV is much smaller than $\hat{\sigma}_\epsilon^2$ obtained by Fuller's method. Moreover, under the proposed method the value of the coefficient of determination is 53.6%, compared to only 50.60% given by the Fuller's method.

The estimates of the regression parameters of the *manifest* model are $\hat{\beta}_{1x} =$

0.344 and $\hat{\beta}_{0x} = 73.153$, and that of the proposed *instrumental variable* model are $\hat{\beta}_{1R} = 0.4479$ and $\hat{\beta}_{0R} = 65.8164$, then $\hat{\beta}_{1R} = 0.4479 > \hat{\beta}_{1x} = 0.344$, and $\hat{\beta}_{0R} = 65.8164 < \hat{\beta}_{0x} = 73.153$.

It is important to compare the results of the new IV estimator with other IV estimators such as the Wald's and Bartlett's methods specified earlier. The Wald method yields, $\hat{Y}_W = 75.91 + 0.305x$ with $MSE_W = 60.98$ and $R^2_W = 0.364$. Moreover, using Bartlett's definition of the IV, we get

$$\hat{\eta}_B = 72.38 + 0.355x, \text{ with } MSE_B = 56.05 \text{ and } R^2_B = 0.425.$$

Practically both methods are inefficient, although Bartlett's method produces a better fit (larger R^2_B) than that of Wald R^2_W. The reliability ratio method provides a much better fit than the OLS, Wald's and Bartlett's methods. However, the reflection based IV fitted model has the largest R^2. At the same time the regression estimates of the slope and intercept for the Fuller method is much close to that of the reflection based estimator. Thus the IV estimator based on the reflection of x provides the best model.

4.6.2 Hen Pheasants Data

The dataset for the second example is also taken from Fuller (2006, p. 34). The data deal with the number of hen pheasants in Iowa at two different seasons/times of the year, and were collected by the Iowa Conservation Com-

Table 4.2: Fitted regression models for the Hen peasants data

Method	Fitted regression equation	MS Error	R^2
OLS	$\hat{\eta}_x = 2.142 + 0.649m$	0.347	0.826
Wald	$\hat{\eta}_W = 2.498 + 0.614m$	0.44	0.78
Bartlett	$\hat{\eta}_B = 2.036 + 0.66m$	0.32	0.84
Reflection	$\hat{\eta}_R = 1.323 + 0.731m$	0.14	0.93
Moments	$\hat{\eta}_{MO} = 1.116 + 0.751m$	0.09	0.95

mission. This data is based on the average number of birds sighted by trained observers traveling a number of specific routes in late April and early May, and again in August. Both variables are subject to error for two reasons. First, the routes are a sample of all possible routes in Iowa. Second, observers cannot be expected to sight all pheasants along the route. The response variable η is the average number of hens in August, and the explanatory variable x is the average number of hens in Spring, where the ratio of error variances $\lambda < 1$. On the basis of previous analyses, it has been estimated that the error variance for the Spring count is about six times larger than that in August. The fitted regression models and associated statistics are provided in the Table 4.2.

The first regression equation and the associated statistics in Table 4.2, $\hat{\eta}_x = 2.142 + 0.649x$, $MSE_x = 0.347$, and $R^2_x = 0.826$, are obtained by the OLS method using x which is subject to the measurement error. The method of moments (MOM) estimator, under the assumption that the ratio $\delta = \sigma_{\delta}^{-2}\sigma_{\epsilon}^{2}$ is known. Following Fuller (2006, p. 35), for $\delta = \frac{1}{6}$, the fitted regression equation becomes $\hat{\eta}_{MO} = 1.1158 + 0.7516x$ with $MSE_{MO} = 0.09$ and $R_{MO}^{2} = 0.95$. This is a much better fitted model, with an increased value of R^2, than that obtained by the OLS method.

The second last row of Table 4.2 represents the regression line and other statistics produced by the proposed modified method based on the reflection of x: $\hat{\eta}_R = 1.323 + 0.731x$, $MSE_R = 0.139$ and $R_R^2 = 0.93$.

The IV estimator based on Wald's method yields $\hat{\eta}_W = 2.498 + 0.614x$ with $MSE_w = 0.44$ and $R_w^2 = 0.78$. Similarly, Bartlett's IV method gives $\hat{\eta}_B = 2.036 + 0.66x$, $MSE_B = 0.32$ and $R_B^2 = 0.84$.

In terms of the R^2 value, the Wald's method is the worst, followed by the OLS method. Thus Wald's IV method may produce a worse fit than the OLS method. The Bartlett's method gives a similar R^2 as the OLS method. However, the MOM estimation produces the largest R^2, although it is not too far from that produced by the proposed reflection based method. It is important to note that the MOM is based on the assumption that the value of σ_{δ}^{2} is known. Furthermore, due to the nature of the definition of the IV, it is only 'treated' the measurement error in the explanatory variable. We should mention that the preference between the estimators can not be generalised but it is only valid for the dataset of this example.

Among the IV estimators the proposed reflection based IV estimator performs much better than the others in terms of providing the best fitted model with largest R^2. This is not surprising due to the fact that IVs proposed by Wald and Bartlett fails to use part of the information of the sample data to define the IV. Although the MOM estimator provides slightly better fit than the proposed reflection based IV method, the former is dependent on the assumption that σ_δ^2 is know, which is not always available.

4.7 Concluding Remarks

This chapter considers the simple regression model with measurement error in the explanatory variable. It also proposes a new estimation procedure based on the idea of a new instrumental variable which is defined from reflection of the *manifest* variable. It compares the existing methods with proposed modified method. Unlike, some of the existing methods it does not lose information.

The illustrative examples demonstrate the fact that the proposed method significantly reduces the mean sum of squares error than the currently used

IV methods. As such, the coefficient of determination of the proposed method is higher than that of the existing IV methods.

Surprisingly, the proposed IV method recovers the true estimator of the slope, $\hat{\beta}_{1\xi}$, from the *manifest* variable and *stochastic* model even if the true values of the *latent* explanatory variable is unobservable. The same comment would apply for the estimator of the intercept.

Chapter 5

Reflection method of estimation for measurement error models

5.1 Introduction

In this chapter we provide an alternative method to the orthogonal regression approach. Moreover, we conduct a comparison using simulation to examine and demonstrate the superior performance of the proposed method.

This chapter introduces the reflection method (RM) of estimation to estimate the parameters of the simple regression model with measurement error (ME) in both variables. Furthermore, theoretical analysis and simulation studies are used to demonstrate the performance of the proposed estimator of the slope parameter for both normal and non-normal structural models. Also we illustrate that the RM estimator has a smaller mean absolute error (MAE) than the orthogonal regression (OR) estimator even if the sample size is small and/or the ratio of error variances (λ) is far from one.

In fact, there is a technical criticism of the assumption that the ratio of error variances (λ) is known. According to Carroll and Ruppert (1996) often we do not have an accurate value of λ. One of the main reasons for that is the presence of the equation error. The maximum likelihood estimator or the orthogonal regression (OR) estimator under the constraint of known λ may over or underestimate the parameter. Weisberg (1985, p. 6) stated, “Real data almost never fall exactly on a straight line”. Lakshminarayanan and Gunst (1984) stated,“Incorrect selection of λ, especially the selection of too small a value when λ is large, compromises the effectiveness of the structural model estimator relative to least squares estimator”.

Geary (1943) introduced the slope estimators for the non-normal structural model which are given by

$$\hat{\beta}_{1G_a} = \frac{k(1,3)}{k(2,2)}, \quad \text{and } \hat{\beta}_{1G_b} = \frac{k(2,2)}{k(3,1)},$$

where $k(\cdot,\cdot)$ represents an appropriate cumulants (see Fuller 2006, p. 72, for details). The cumulants based estimators become unstable if the non-normal model is too close to the normal model (see Cheng and Ness, 1999, p. 127).

Wald (1940) proposed an estimation method based on the grouping of the data. He divides the observations on both the response and explanatory variables into two groups, G1 and G2, where G1 contains the first half of the ordered observations and G2 contains the second half. Wald's estimator of β_1 is given by

$$\hat{\beta}_{1W} = \frac{a_1}{b_1} = \frac{(y_1 + \ldots + y_k) - (y_{k+1} + \ldots + y_n)}{(x_1 + \ldots + x_k) - (x_{k+1} + \ldots + x_n)} = \frac{\bar{y}_2 - \bar{y}_1}{\bar{x}_2 - \bar{x}_1},$$

where $\bar{x}_1$ and $\bar{y}_1$ are the means of x_j and y_j in group $G1$, for $j = 1, 2, \cdots, k$, and $\bar{x}_2$ and $\bar{y}_2$ are the means of x_j and y_j in group $G2$, for $j = k+1, k+2, \cdots, n$. Then

$$\hat{\beta}_{0W} = \bar{y} - \hat{\beta}_1 \bar{x},$$

where $\bar{y} = \frac{\sum_{j=1}^{n} y_j}{n}$, $\bar{x} = \frac{\sum_{j=1}^{n} x_j}{n}$, and

$$a_1 = \frac{(x_1 + \ldots + x_k) - (x_{k+1} + \ldots + x_n)}{n}, \text{ and}$$

$$b_1 = \frac{(y_1 + \ldots + y_k) - (y_{k+1} + \ldots + y_n)}{n}.$$

In the literature of ME there are some criticisms of the Wald's estimator but these lack consensus (Gillard, 2010).

In the next section the orthogonal regression method, which is a special case of the maximum likelihood method, or Deming regression approach when the ratio of error variance equals one, is included. The proposed estimator is based on the reflection of the observed values of the explanatory variable as discussed in Section 5.3. Two numerical illustrations are provided in Section 5.4. These examples compare the proposed RM estimator with orthogonal regression method estimator under correct and incorrect specification of λ, two grouping method, method of moments using third-order moments, and Geary's methods. Some concluding remarks are included in Section 5.5.

5.2 Orthogonal regression method

One of the techniques suggested to overcome the problem of measurement error is the orthogonal regression. This technique is also known as the major axis, principal component regression or the perpendicular distance method. The reason that the orthogonal regression (OR) approach was adopted, instead of the ordinary least squares regression, is that both variables are subject to error. This method considers a bivariate case of principal components analysis. The basic idea of this method is to minimise the squared perpendicular distances of the data points from the fitted regression line. The orthogonal regression estimator of the true slope is given by

$$\hat{\beta}_{OR} = \frac{(S_y^2 - S_x^2) + \sqrt{(S_y^2 - S_x^2)^2 + 4S_{yx}^2}}{2S_{yx}}.$$

An alternative form of this estimator is given by

$$\hat{\beta}_{OR} = 0.5\left[(\hat{\beta}_2 - \hat{\beta}_1^{-1}) + sgn\{S_{yx}\}\sqrt{4 + (\hat{\beta}_2 - \hat{\beta}_1^{-1})^2}\;\right],$$

where $\hat{\beta}_1 = \dfrac{S_{yx}}{S_x^2}$, $\hat{\beta}_2 = \dfrac{S_y^2}{S_{yx}}$, S_y^2 is the sample variance of the manifest response variable y, S_x^2 is the sample variance of the manifest explanatory variable x and S_{yx} is the sample covariance of y and x.

The orthogonal regression method is an appropriate solution to the measurement error problem if the following assumptions are met:

1. There is no equation error in the model which means that all the points (ξ_j, η_j) fall exactly on a straight line.
2. The ratio of error variances (λ) equals one, this means that the variance of the measurement error in the response variable equals the variance of the measurement error in the explanatory variable, that is, $\sigma_\epsilon^2 = \sigma_\delta^2$.

Indeed, the first assumption is unlikely to be satisfied because most of the variables are not related by mathematical or physical laws. For instance, Warton, et al. (2006) stated “In allometry, equation error is often large compared to measurement error, in which case it would be more reasonable to assume there is no measurement error than to assume no equation error”. Moreover, the second assumption is also viewed as a strict assumption and is rarely met. Despite the above criticisms the orthogonal regression method is still used in many disciplines. In fact, these criticisms were behind the motivation to provide an alternative method with flexible assumptions and better performance than orthogonal regression method.

5.3 Proposed reflection method of estimation

To avoid the unwanted and troublesome influence of the measurement error in both the explanatory and the response variables, the idea of reflection of the manifest variable is used for all the values of the manifest explanatory variable x_j. The reflection of the points is taken about the fitted regression line of the manifest variables. This is essentially done by a transformation of the observed values of the explanatory variable to their reflection on the Euclidean plane. In the conventional notation, the reflection of the explanatory variable $x_j = \xi_j + \delta_j$ (with measurement error δ_j) for $j = 1, 2, \cdots, n$, can be defined as

$$x_j^* \quad = \quad x_j \cos 2\psi + (y_j - \hat{\beta}_{0x}) \sin 2\psi, \tag{5.1}$$

where $\hat{\beta}_{0x}$ is the least squares estimate of the intercept parameter, ψ is the angle measure defined as $\psi = \arctan \hat{\beta}_{1x}$ in which $\hat{\beta}_{1x}$ is the least squares estimate of the slope parameter in the manifest model, and cos and sin are the usual trigonometric cosine and sine functions respectively. For the definition of the reflection of points on the Euclidean plane see Vaisman (1997, p. 164).

It is well known, that the least squares criterion for estimating β_0 and β_1 is to choose estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimises the sum of squared distances of the observed points from the estimated line (see Fuller, 2006, p. 37). The maximum likelihood approach for the normal ME regression model is an orthogonal regression method when the ratio of the error variances is one, $\lambda = \dfrac{\sigma_\epsilon^2}{\sigma_\delta^2} = 1$ (see Cheng and Van Ness, 1999, p. 9). The orthogonal regression estimators of β_0 and β_1 are obtained by minimising the following weighted sum of squares:

$$min\, \sigma_\epsilon^2 \;=\; min\left(\frac{\sigma_v^2}{1+\beta_1^2}\right). \tag{5.2}$$

Note that the equation (5.2) is correct only when $\lambda = \dfrac{\sigma_\epsilon^2}{\sigma_\delta^2} = 1$ and $\sigma_{\epsilon\delta} = 0$. Because from the equation (1.4) in Chapter 1, the error term v_j of the measurement error model is $v_j = \epsilon_j - \beta_1\delta_j$, then the variance of v_j is given by

$$\begin{aligned}\sigma_v^2 &= \sigma_\epsilon^2 + \beta_1^2\sigma_\delta^2 = \sigma_\epsilon^2(1+\beta_1^2), \text{ then}\\ \sigma_\epsilon^2 &= \frac{\sigma_v^2}{(1+\beta_1^2)}.\end{aligned}$$

In this case when both the response and explanatory variables are subject to measurement error and the ratio of error variances equal one ($\lambda = 1$), then the distance between the true point (ξ, η) and the observed point (x, y) is the perpendicular distance, instead of vertical distance, of the fitted regression line. That is why this case requires methods for minimising the orthogonal distance. But the orthogonal regression method works well only when ($\lambda =$

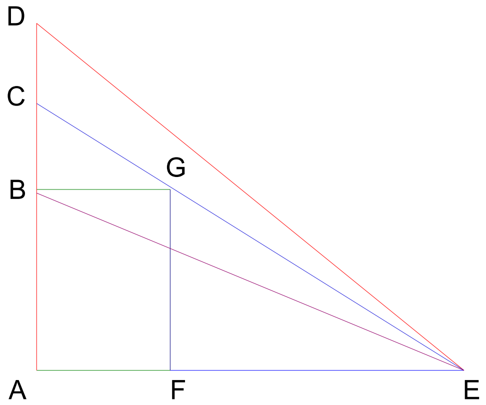

Figure 5.1: Graph of the sum of squares and products of the latent and manifest variables in the presence of ME in the variables.

1), and there is no equation error in the model. About the equation error Warton et al. (2006) pointed out that in practice it is rare to find a good regression model without including an equation error. This chapter provides a new estimator based on minimising the perpendicular distance for linear regression model with or without equation error.

5.3.1 Geometric explanation

It can be easily explained geometrically that the presence of measurement error in both response and explanatory variables impacts on the estimator of the slope parameter. The graphical representation also explains how the actual estimator of the slope is recovered by the new method.

Figure 5.1 represents the sum of squares and sum of products associated with the definition of the estimators of slope for the latent, manifest and reflection variables. This graph represents the measurement error in the variables as well as the three estimators of the slope parameter. These estimators are the OLS of y on ξ, the proposed method, and the OLS of y on x estimators of the slope.

From Figure 5.1, the true estimator of the slope $\hat{\beta}_{1\xi}$ when the latent explanatory variable is available, is represented by $\tan\hat{\theta}$, where $\hat{\theta} = \angle FEG$ of $\triangle FEG$. In the absence of the values of the latent variable this is unavailable. But for the manifest variable one can find the estimator of the slope to be $\hat{\beta}_{1x}$ which is represented by $\tan\psi$, where $\psi = \angle AEB$ of $\triangle AEB$. The latter $\mid \hat{\beta}_{1x} \mid$ is an underestimate of the former $\mid \hat{\beta}_{1\xi} \mid$. Note that here AF (or equivalently BG) represents the sum of squares of measurement error in the explanatory variable (SS_δ). Furthermore, under the assumptions of $E[y\delta] = 0$ and $E[\xi\delta] = 0$, we have $AB = FG$ or $SP_{\xi y} = SP_{xy}$.

Finally, $\hat{\beta}_{1\xi} = \dfrac{SP_{\xi y}}{SS_\xi} = \dfrac{FG}{FE}$, and $\hat{\beta}_{1x} = \dfrac{SP_{xy}}{SS_x} = \dfrac{AB}{AE}$.

The introduction of the reflection of the manifest explanatory variable changes $\triangle AEB$ of Figure 5.1 to $\triangle AEC$. This triangle represents the effect of the reflection of the manifest variable. From Figure 5.1 the estimates of the slope are

$$\begin{aligned}
\hat{\beta}_{1x} &= \frac{SP_{xy}}{SS_x}\left(=\frac{AB}{AE}\right), && (5.3)\\
\hat{\beta}_{1\xi} &= \frac{SP_{\xi y}}{SS_\xi}\left(=\frac{FG}{FE}\right), && (5.4)\\
\hat{\beta}_{1RM} &= \frac{SP_{d_1 y}}{SS_x}\left(=\frac{AC}{AE}\right) && (5.5)
\end{aligned}$$

Note d_1 in equation (5.5) is the mean value of x and x^*, that is $d_{1j} = \dfrac{x_j^* + x_j}{2}$, in which x_j^* is the reflection of the explanatory variable x_j.

Based on the foregoing explanations we can rewrite the formula of the true estimator $\hat{\beta}_{1\xi}$ in order to find a formula for the true slope estimator based on the manifest explanatory variable x. This can be done by finding the value of AC in Figure 5.1, since the solution relies on finding the value of AC or BC, where $AC = \hat{\beta}_{1\xi} SS_x$ and $BC = \hat{\beta}_{1\xi} SS_\delta$. In order to find the true slope estimator $\hat{\beta}_{1\xi} = \dfrac{AC}{SS_x}$ we need to know the distance AC as shown in Figure 5.1. Therefore, we suggest to use the variable $d_{1j} = \dfrac{x_j^* + x_j}{2}$ to provide an estimator for the slope which minimises the orthogonal distance as follows:

From (5.1) $x_j^* = x_j \cos 2\psi + (y_j - \hat{\beta}_{0x}) \sin 2\psi$, where

$$\begin{aligned}
\tan\psi &= \frac{SP_{yx}}{SS_x},\\
\cos 2\psi &= \cos^2\psi - \sin^2\psi = \frac{SS_x^2 - SP_{yx}^2}{SS_x^2 + SP_{yx}^2},\\
\sin 2\psi &= 2\cos\psi\sin\psi = \frac{2SS_x SP_{yx}}{SS_x^2 + SP_{yx}^2}.
\end{aligned}$$

Then

$$\begin{aligned}
x_j^* &= x_j\left(\frac{SS_x^2 - SP_{yx}^2}{SS_x^2 + SP_{yx}^2}\right) + (y_j - \hat{\beta}_{0x})\left(\frac{2SS_xSP_{yx}}{SS_x^2 + SP_{yx}^2}\right) \\
x_j^*SS_x^2 + x_j^*SP_{yx}^2 &= x_jSS_x^2 - x_jSP_{yx}^2 + 2y_jSS_xSP_{yx} - 2\hat{\beta}_{0x}SS_xSP_{yx} \\
2y_jSS_xSP_{yx} &= x_j^*SS_x^2 + x_j^*SP_{yx}^2 - x_jSS_x^2 + x_jSP_{yx}^2 + 2\hat{\beta}_{0x}SS_xSP_{yx}
\end{aligned}$$

$$\begin{aligned}
2y_jSS_xSP_{yx} &= (x_j^* - x_j)SS_x^2 + (x_j^* + x_j)SP_{yx}^2 + 2\hat{\beta}_{0x}SS_xSP_{yx} \\
y_j &= \hat{\beta}_{0x} + (x_j^* + x_j)\frac{SP_{yx}^2}{2SS_xSP_{yx}} + (x_j^* - x_j)\frac{SS_x^2}{2SS_xSP_{yx}} \\
y_j &= \hat{\beta}_{0x} + \hat{\beta}_{1x}\frac{(x_j^* + x_j)}{2} + \frac{(x_j^* - x_j)}{2\hat{\beta}_{1x}}.
\end{aligned}$$

Then let $d_{1j} = \dfrac{(x_j^* + x_j)}{2}$ and $t_j = \dfrac{(x_j^* - x_j)}{2}$ so

$$y_j = \bar{y} - \hat{\beta}_{1x}\bar{x} + \hat{\beta}_{1x}d_{1j} + \frac{t_j}{\hat{\beta}_{1x}}.$$

Based on Theorem 2 in Chapter 3 we get $\bar{d}_1 = \bar{x}$ and $\bar{t} = 0$, then

$$(y_j - \bar{y}) = \hat{\beta}_{1x}(d_{1j} - \bar{d}_1) + \frac{t_j}{\hat{\beta}_{1x}}.$$

Multiplying both sides by y_j and taking sum over j, we get

$$\begin{aligned}
\sum_{j=1}^{n}(y_j - \bar{y})y_j &= \hat{\beta}_{1x}\sum_{j=1}^{n}(d_{1j} - \bar{d}_1)y_j + \frac{\sum_{j=1}^{n} t_jy_j}{\hat{\beta}_{1x}} \\
SS_y &= \hat{\beta}_{1x}SP_{yd_1} + \frac{SP_{yt}}{\hat{\beta}_{1x}}.
\end{aligned}$$

Note from Theorem 7 in Chapter 3

$$\begin{aligned}
SP_{yt} &= \frac{SSE_v \sin 2\psi}{2}, \text{ because} \\
SP_{yt} &= \frac{SP_{yx^*} - SP_{yx}}{2} \text{ then} \\
SS_y &= \hat{\beta}_{1x} SP_{yd_1} + \frac{SSE_v \sin 2\psi}{2\hat{\beta}_{1x}}, \\
\text{where } \frac{\sin 2\psi}{2\hat{\beta}_{1x}} &= \cos^2 \psi, \text{ and hence} \\
SS_y &= \hat{\beta}_{1x} SP_{yd_1} + SSE_v \cos^2 \psi \\
&= \frac{SP_{yd_1} SP_{yx}}{SS_x} + SSE_v \cos^2 \psi.
\end{aligned}$$

So the new proposed estimator for the slope parameter β_1 is $\frac{SP_{yd_1}}{SS_x} = \hat{\beta}_{1RM}$, and so

$$SS_y = \hat{\beta}_{1RM} SP_{yx} + SSE_v \cos^2 \psi, \tag{5.6}$$

where SS_y is the sum of squares of y, SP_{yx} is the sum of products of y and x, $\hat{\beta}_{1x}$ is the OLS estimator of the slope, SSE_v is the sum of squares of error of the OLS estimator for the measurement error model.

$$\begin{aligned}
\text{Note when } SSE_{d_1} &= SSE_v \cos^2 \psi, \text{ then} \qquad (5.7) \\
SS_y &= \hat{\beta}_{1RM} SP_{yx} + SSE_{d_1}, \qquad (5.8)
\end{aligned}$$

where SSE_{d_1} is the sum of squares of error of the RM estimator for the measurement error model.

Obviously, the proposed estimator $\hat{\beta}_{1RM}$ has minimised the sum of squared residuals SSE_v, because $SSE_v \cos^2 \psi \leq SSE_v$, where $0 \leq \cos^2 \psi \leq +1$. That means, the sum of squared residuals SSE_v reduced by $SSE_v \sin^2\psi$, where $SSE_v - SSE_{d_1} = SSE_v - SSE_v \cos^2\psi = SSE_v \sin^2\psi$. Here we show what we have stated previously that the proposed estimator $\hat{\beta}_{1RM}$ minimises the orthogonal distances. Therefore, we seek to prove that the proposed estimator $\hat{\beta}_{1RM}$ works as the orthogonal regression and the maximum likelihood solution to minimise the sum of squared perpendicular distances from the data points to the regression line even when λ is misspecified. We can show that the sum of squared residuals SSE_{d_1} is the sum of squared perpendicular distances as follows:

$$SSE_{d_1} = SSE_v \cos^2\psi = \frac{SSE_v}{1+\hat{\beta}_{1x}^2},\ \textit{then}$$

$$\begin{aligned}\cos^2\psi &= \frac{1}{\frac{1}{\cos^2\psi}} = \frac{1}{\frac{\cos^2\psi+\sin^2\psi}{\cos^2\psi}}\\ &= \frac{1}{\frac{\cos^2\psi}{\cos^2\psi}+\frac{\sin^2\psi}{\cos^2\psi}} = \frac{1}{1+\frac{\sin^2\psi}{\cos^2\psi}}\\ &= \frac{1}{1+\hat{\beta}_{1x}^2},\end{aligned}$$

where $\dfrac{\sin^2\psi}{\cos^2\psi} = \hat{\beta}_{1x}^2$, and $SSE_v = \sum_{j=1}^{n}(y_j - \hat{\beta}_{0x} - \hat{\beta}_{1x}x_j)^2$, then

$$SSE_{d_1} = \frac{\sum_{j=1}^{n}(y_j - \hat{\beta}_{0x} - \hat{\beta}_{1x}x_j)^2}{1+\hat{\beta}_{1x}^2}.$$

Consequently, the sum of squared residuals SSE_{d_1} is the sum of squared perpendicular distances of the data points from the regression line.

It can be proved from (5.6) that the proposed estimator $\hat{\beta}_{1RM}$ is greater than the OLS estimator $\hat{\beta}_{1x}$, as follows:

It is well known that

$$SS_y \quad = \quad \hat{\beta}_{1x} SP_{yx} + SSE_v. \tag{5.9}$$

From (5.6) and (5.9) we get

$$SS_y = \hat{\beta}_{1RM} SP_{yx} + SSE_v \cos^2\psi \quad = \quad \hat{\beta}_{1x} SP_{yx} + SSE_v. \tag{5.10}$$

Hence

$$\begin{aligned}(\hat{\beta}_{1RM} - \hat{\beta}_{1x}) SP_{yx} &= (SSE_v - SSE_v \cos^2\psi) \\ &= SSE_v(1 - \cos^2\psi) \\ &= SSE_v \sin^2\psi. \qquad (5.11)\end{aligned}$$

From (5.10) and (5.11),

$$\mid \hat{\beta}_{1x} \mid \leq \mid \hat{\beta}_{1RM} \mid \leq \frac{S_y^2}{\mid S_{yx} \mid}, \tag{5.12}$$

where the right hand side of (5.11) $SSE_v \sin^2\psi$ is always positive.

5.4 Relationship between $\hat{\beta}_{1RM}$ and $\hat{\beta}_{1RMA}$

This section introduces the relationship between the reflection estimator $\hat{\beta}_{1RM}$ and the reduced major axis estimator $\hat{\beta}_{1RMA}$ as follows

$$\begin{aligned} d_{1j} &= \frac{(x_j^* + x_j)}{2} \\ 2d_{1j} &= x_j^* + x_j = x_j \cos 2\psi + y_j \sin 2\psi - \hat{\beta}_{0x} \sin 2\psi + x_j \\ &= x_j(\cos 2\psi + 1) + y_j \sin 2\psi - \hat{\beta}_{0x} \sin 2\psi \end{aligned}$$

where $\hat{\beta}_{0x} = \bar{y} - \hat{\beta}_{1x}\bar{x}$. Now

$$\begin{aligned}
2d_{1j} &= 2x_j \cos^2\psi + y_j \sin 2\psi - (\bar{y} - \hat{\beta}_{1x}\bar{x}) \sin 2\psi \\
&= 2x_j \cos^2\psi + (y_j - \bar{y}) \sin 2\psi + 2\bar{x} \sin^2\psi \\
&= 2x_j \cos^2\psi + (y_j - \bar{y}) \sin 2\psi + 2\bar{x} - 2\bar{x} \cos^2\psi \\
&= 2(x_j - \bar{x}) \cos^2\psi + (y_j - \bar{y}) \sin 2\psi + 2\bar{x}, \qquad (5.13)
\end{aligned}$$

where $(\cos 2\psi + 1) = 2\cos^2\psi$, $\hat{\beta}_{1x} = \dfrac{\sin\psi}{\cos\psi}$, $\hat{\beta}_{1x} \sin 2\psi = 2\sin^2\psi$.

Multiplying both sides of the equation (5.13) by y_j, and taking the sum over j, we obtain

$$2\sum_{j=1}^{n} y_j d_{1j} = 2\sum_{j=1}^{n}(x_j - \bar{x})y_j \cos^2\psi + \sum_{j=1}^{n}(y_j - \bar{y})y_j \sin 2\psi + 2\bar{x}\sum_{j=1}^{n} y_j$$

$$2\sum_{j=1}^{n} y_j d_{1j} - 2n\bar{x}\bar{y} = 2\sum_{j=1}^{n}(x_j - \bar{x})y_j \cos^2\psi + \sum_{j=1}^{n}(y_j - \bar{y})y_j \sin 2\psi. \qquad (5.14)$$

Note based on Theorem 2 in Chapter 3 we have $\bar{d_{1j}} = \bar{x} = \bar{x}^* = \bar{\xi}$, and by dividing both sides of the equation (5.14) by $2(n-1)$, we find

$$\frac{\sum_{j=1}^{n} y_j d_{1j} - n\bar{x}\bar{y}}{n-1} = \frac{\sum_{j=1}^{n}(x_j - \bar{x})y_j}{n-1} \cos^2\psi + \frac{\sum_{j=1}^{n}(y_j - \bar{y})y_j}{2(n-1)} \sin 2\psi, \qquad (5.15)$$

which gives

$$S_{yd_1} = S_{yx} \cos^2\psi + S_y^2 \sin\psi \cos\psi. \qquad (5.16)$$

Now by dividing both sides of the equation (5.16) over S_x^2, then

$$\begin{aligned}
\frac{S_{yd_1}}{S_x^2} &= \frac{S_{yx}}{S_x^2}\cos^2\psi + \frac{S_y^2}{S_x^2}\sin\psi\cos\psi \\
\hat{\beta}_{1RM} &= \hat{\beta}_{1x}\cos^2\psi + \hat{\beta}_{1RMA}^2\cos\psi\sin\psi \\
&= \cos\psi\sin\psi + \hat{\beta}_{1RMA}^2\cos\psi\sin\psi \\
&= (1+\hat{\beta}_{1RMA}^2)\cos\psi\sin\psi \\
&= (1+\hat{\beta}_{1RMA}^2)\frac{S_x^2 S_{yx}}{S_x^4+S_{yx}^2}, \qquad (5.17)
\end{aligned}$$

where $\hat{\beta}_{1RMA}$ is the reduced major axis estimator of the slope parameter β_1, and $\cos\psi = \dfrac{S_x^2}{\sqrt{S_x^4+S_{yx}^2}}$ and $\sin\psi = \dfrac{S_{yx}}{\sqrt{S_x^4+S_{yx}^2}}$.

Note that the equation (5.17) refers to the relationship between the reflection estimator $\hat{\beta}_{1RM}$ and the reduced major axis estimator $\hat{\beta}_{1RMA}$.

5.5 Simulation studies

In this section, two illustrative examples using simulated data are provided, where both the response and explanatory variables are subject to error. These examples reveal that the proposed new estimator works well under the assumption $\lambda = 1$.

Example 1 The dataset is taken from the example 4.14 of Cheng and Van Ness (1999, p. 127). The purpose of this example is to compare the sample mean absolute errors of the proposed estimator with other methods namely: the method of moments using third-order moments, Geary's methods (a,b) using fourth-order cumulates, and the grouping method with two groups (for more details see Chapter 2). The dataset is based on 500 replications of non-normal data of the latent explanatory variable ξ, and normal data of the measurement error ϵ and δ as follows:

1. Generate 100 independent values $\xi_1, \ldots, \xi_{100}$ of $\xi \sim$ uniform on $[-5, 5]$.
2. Generate 100 independent values $\delta_1, \ldots, \delta_{100}$ of $\delta \sim N(0, 1)$.
3. Generate 100 independent values $\epsilon_1, \ldots, \epsilon_{100}$ of $e \sim N(0, 1)$.

We calculated the generated values of the response and explanatory variables for preselected values of β_0, β_1, and n. Then compute values of the estimators from the simulated data and find their means and mean absolute errors (MAE) for each of the five estimators. The simulated mean of the estimators and the MAE when $\beta_1 = 1$, $\beta_0 = 0$, $n = 100$ are recorded in Table 5.1. Table 5.2 shows the simulated mean of the estimators and the MAE for $\beta_1 = 2$, $\beta_0 = 0$, and $n = 100$.

Table 5.1: The simulated mean of five different estimators and the MAE when $\beta_1 = 1$, $\beta_0 = 0$, $n = 100$.

Method	Mean $\hat{\beta}_0$	MAE $\hat{\beta}_0$	Mean $\hat{\beta}_1$	MAE $\hat{\beta}_1$
Two groups	-0.010	0.114	0.919	0.084
Method of moments	-0.440	0.747	-2.796	4.860
Geary method(a)	-0.010	0.117	0.997	0.055
Geary method(b)	-0.010	0.118	1.001	0.056
Reflection method	-0.0007	0.009	0.998	0.039

Table 5.2: The simulated mean of five different estimators with the MAE when $\beta_1 = 2$, $\beta_0 = 0$, $n = 100$.

Method	Mean $\hat{\beta}_0$	MAE $\hat{\beta}_0$	Mean $\hat{\beta}_1$	MAE $\hat{\beta}_1$
Two groups	-0.002	0.165	1.845	0.157
Method of moments	-0.211	0.541	2.053	2.200
Geary method(a)	-0.002	0.172	2.002	0.082
Geary method(b)	-0.002	0.173	2.003	0.090
Reflection method	-0.001	0.014	1.996	0.063

Both Tables 5.1 and 5.2 show that the RM works well, and it is better than other estimators. It is clear that the MAE of the RM estimator is the smallest compared to the others. In general, the proposed RM estimator is superior to the other estimators for non-normal model. The next example shows the comparison between the RM estimator, orthogonal regression (maximum likelihood) estimator, and the slope estimator of the OLS of y on x for the normal structural model.

Example 2 The purpose of this example is to compare the mean absolute error of the reflection estimator with the orthogonal regression estimator for normal structural model, when both variables are subject to error with correct and incorrect selection of λ. Based on $100,000$ replications of symmetric normal data for $\beta_1 = 1$, $\beta_0 = 0$, we calculated the OR, OLS, and RM estimators.

The estimates of the slope and their mean absolute errors are shown in Figure 5.2 when the sample size is small $10 < n < 30$ and correct selection of $\lambda = 1$. Figure 5.3 shows the slope estimates and their mean absolute errors when the sample size is large $10 < n < 120$ and incorrect selection of $\lambda = 1.44$.

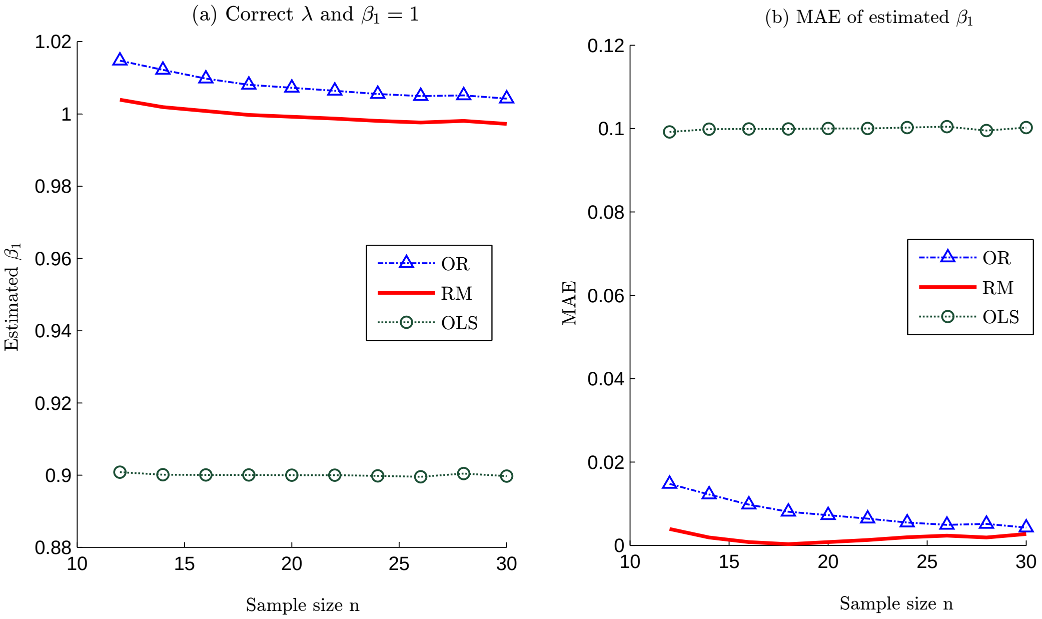

Figure 5.2: (a) Graph of the slope estimated and (b) Graph of MAE of the RM, OR, and OLS estimators of β_1, when $\lambda = 1$ is correct and small sample sizes $10 < n < 30$.

Figure 5.2a shows that the RM estimator is closer to the true slope β_1 than other estimators under correct value of $\lambda = 1$ and small sample sizes for the normal structural model. The results of the mean absolute error in Figure 5.2b demonstrate the superiority of the RM estimator compared to the other estimators.

Under the misspecification of the value of $\lambda (= 1.44)$ and increased sample sizes, Figure 5.3a reveals that the RM estimator remains closer to the true slope β_1 than the other estimators. Note all these estimators are biased when λ is incorrect or misspecified. It is noteworthy that the results of the mean

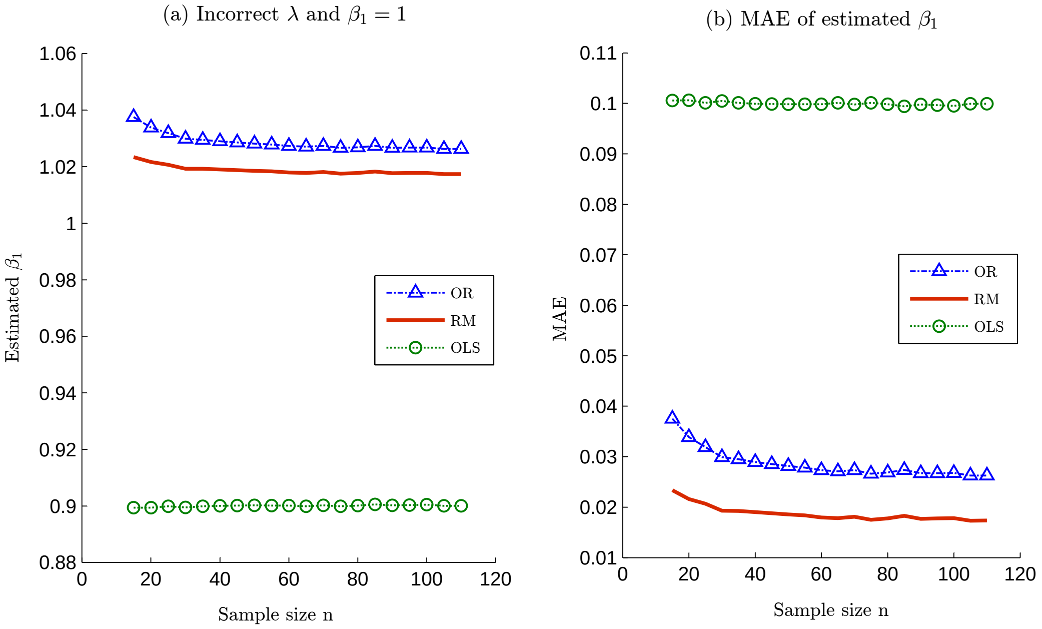

Figure 5.3: (a) Graph of the slope estimated and (b) Graph of MAE of the RM, OR, and OLS estimators for β_1, when $\lambda\,(= 1.44)$ is incorrect and larger sample sizes $10 < n < 120$.

absolute error in Figure 5.3b indicate that the RM estimator is less sensitive to the misspecification of λ.

5.6 Concluding remarks

This chapter considers the simple regression model with measurement error in both response and explanatory variables. It proposes a new estimation procedure based on the reflection of the explanatory variable. We have shown that the RM estimator is equivalent or asymptotically equivalent to the orthogonal regression estimator, and nearly asymptotically unbiased under the assumption of $\lambda = \dfrac{\sigma_\epsilon^2}{\sigma_\delta^2} = 1$. Moreover, even if the ratio of error variances $\lambda \neq 1$ and the sample size is not large, the mean absolute error of the RM estimator is lower than that of the orthogonal regression and OLS estimators. The simulated results in Tables 5.1-5.2 and Figures 5.2-5.3 clearly demonstrate that the RM estimator performs better than its competitors in both normal and non-normal models and under correct and incorrect specification of λ regardless of the sample size.

Chapter 6

Reflection in grouping method estimation

6.1 Introduction

The main aim of this chapter is to propose a new grouping method for Wald's IV approach based on the reflection of the explanatory variable. The new reflection grouping (RG) method is a modification of the Wald's estimator. The second aim of the chapter is to deal with the situation when the assumption of known λ is violated. The proposed method assumes a very flexible range of values of λ, namely (1) $\lambda = 1$, (2) $\lambda > 1$, or (3) $\lambda < 1$. This is a much weaker and a more realistic assumption than knowing the value of λ.

In addition, we provide a performance comparison between the RG method estimator and several existing estimators such as the OLS, Geary (a and b), Wald's, and ML estimators using the dataset of the Example 4.12 of Cheng and Van Ness (1999, p. 123). Moreover, we perform large scale simulation studies to illustrate that the proposed estimator is asymptotically unbiased and consistent under both non-normal and normal distributions of ξ and the flexible assumption on the value of λ.

Section 6.2 provides the summary of Wald's grouping method. The proposed reflection grouping method is introduced in Section 6.3. Simulation study and comparison of estimators are provided in Section 6.4. The final section contains the concluding remarks.

6.2 Wald's grouping method

The Wald's grouping method is also known as two grouping method or average grouping method (see Gillard, 2010). In 1940 Wald pointed out that a consistent estimator of β_1 may be calculated if the following assumptions are met:

1. The random variables $\epsilon_1, \cdots, \epsilon_n$ have the same distribution and they are uncorrelated, that is, $E(\epsilon_i \epsilon_j) = 0$ for $i \neq j$. The variance of ϵ_j is

finite.

2. The random variables $\delta_1, \cdots, \delta_n$ have the same distribution and they are uncorrelated, that is, $E(\delta_i \delta_j) = 0$ for $i \neq j$. The variance of δ_j is finite.

3. The random variables ϵ_j and δ_j are uncorrelated, that is, $E(\epsilon_j \delta_j) = 0$ for all $j = 1, 2, \cdots, n$.

4. $\frac{\sum_{j=k+1}^{n} x_j - \sum_{j=1}^{k} x_j}{n} > 0$ or $\bar{x}_{k+1} > \bar{x}_k$, where $\bar{x}_{k+1}$ is the mean of the group G2, $\bar{x}_k$ is the mean of the group G1, n is even $(n = 2, 4, 6, \ldots, \infty)$, and $k = \frac{n}{2}$. In other words, we can be sure that as $n \to \infty$, b_1 does not approach zero (cf Madansky, 1959).

The method divides the observations into two groups based on the ranks of the manifest explanatory variable x_j, those above the median of x_j into one group, G_1, and those below the median into another group, G_2. Wald's estimator of β_1 is given by

$$\hat{\beta}_{1W} \quad = \quad \frac{a_1}{b_1} = \frac{(y_1 + \cdots + y_k) - (y_{k+1} + \cdots + y_n)}{(x_1 + \cdots + x_k) - (x_{k+1} + \cdots + x_n)} = \frac{\bar{y}_2 - \bar{y}_1}{\bar{x}_2 - \bar{x}_1},$$

where $\bar{x}_1$ and $\bar{y}_1$ are the means of x_j and y_j in group $G1$, for $j = 1, 2, \cdots, k$, and $\bar{x}_2$ and $\bar{y}_2$ are the means of x_j and y_j in group $G2$, for $j = k+1, k+2, \cdots, n$. Then

$$\hat{\beta}_{0W} \quad = \quad \bar{y} - \hat{\beta}_{1W}\bar{x},$$

where $\bar{y} = \dfrac{\sum_{j=1}^{n} y_j}{n}$, $\bar{x} = \dfrac{\sum_{j=1}^{n} x_j}{n}$, and

$$\begin{aligned} a_1 &= \frac{(x_1 + \cdots + x_k) - (x_{k+1} + \cdots + x_n)}{n}, \\ b_1 &= \frac{(y_1 + \cdots + y_k) - (y_{k+1} + \cdots + y_n)}{n}. \end{aligned}$$

This method can also be put into the context of instrumental variables. Johnston (1972, p. 284) showed how to express Wald's grouping method as an instrumental variable method. If the number of sample observations is even then define a z matrix as

$$z' = \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ -1 & -1 & -1 & \cdots & -1 \end{bmatrix},$$

where the second row included minus or plus one according to the value of the manifest explanatory variable x_j is below or above the median of x.

If we rewrite the estimated model $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ in matrix form as $y = x'\beta$, where

$$x' = \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ x_1 & x_2 & x_3 & \cdots & x_n \end{bmatrix},$$

and $\beta = (\beta_{0W}, \beta_{1W})'$, then the instrumental variable estimator of β is defined by

$$\hat{\beta} = (z'x)^{-1} z'y = \begin{bmatrix} n & 0 \\ 0 & \frac{n}{2}(\bar{x}_2 - \bar{x}_1) \end{bmatrix}^{-1} \begin{bmatrix} n\bar{y} \\ \frac{n}{2}(\bar{y}_2 - \bar{y}_1) \end{bmatrix} = \begin{bmatrix} \bar{y} \\ \frac{\bar{y}_2 - \bar{y}_1}{\bar{x}_2 - \bar{x}_1} \end{bmatrix}.$$

Then the Wald's estimator of the slope is

$$\hat{\beta}_{1W} = \frac{\bar{y}_2 - \bar{y}_1}{\bar{x}_2 - \bar{x}_1}.$$

According to Johnston (1972, p. 284) $\bar{y} = \frac{\sum_{j=1}^{n} y_j}{n}$ is the estimator of $\beta_0 + \beta_1 E(x)$, and hence

$$\hat{\beta}_{0W} = \bar{y} - \hat{\beta}_{1W}\bar{x}.$$

It is suggested that one should omit the central observation of the ordered array before computations if n is odd.

Wald's estimator has seen some criticisms in the literature of measurement error model, but these criticisms lack consensus. For instance Gupta and Amanullah (1970) pointed out that the Wald's estimator is consistent under very specific conditions except that the errors are not normally distributed. Pakes (1982) claimed that the work of Gupta and Amanullah (1970) is needless, when the Wald's estimator is inconsistent. Under the normality assumption the grouping estimator is the maximum likelihood estimator (see Chang and Huang, 1997). In practice, the grouping method is still important, and the grouping estimator is the maximum likelihood estimator under the normality assumption (see Chang and Huang, 1997; Cheng and Van Ness, 1999, p. 130). Neyman and Scott (1951) pointed out that the Wald's estimator is consistent for β_1 in the structural relationship if and only if

$$Pr[x_{p_1} - e < \xi \leq x_{p_1} - \mu] = Pr[x_{1-p_2} - e < \xi < x_{p_1} - \mu] = 0,$$

where x_{p_1} and x_{1-p_2} are the p_1 and $(1-p_2)$ percentile points of $F(x)$, the distribution function of x (see Madansky, 1959).

This condition means that we must know the range of the error in x, and in order to satisfy the condition the range should be finite, otherwise the condition becomes $Pr[-\infty < \xi < \infty] = 0$ which is never satisfied.

Theil and Yzeren (1956) mentioned that the Wald's method is valuable, though there is a loss of efficiency. Johnston (1972, p. 284) stated "Under fairly general conditions the Wald's estimator is consistent but likely to have a large sampling variance". Moreover, Fuller (2006, p. 74) mentioned that the Wald's method was often interpreted improperly. In fact, there are many discussions on improving the efficiency of the grouping method by dividing the observations into more than two groups and groups of unequal size (see Nair and Banerjee, 1942; Bartlett, 1949; Dorff and Gurland 1961; and Ware, 1972).

6.3 Proposed reflection grouping method

Based on the idea of the proposed estimator into the previous chapter, we can extend the role of the transformed variable d_{1j} to derive two other transformed variables d_{2j} and d_{3j}. The proposed (RG) method suggest grouping based on the ranks of the transformed variables d_{1j}, d_{2j} and d_{3j} which are calculated as the mean of the manifest explanatory variable x_j and its *reflection* x_j^*. To avoid the unwanted and troublesome influence of the measurement error in the explanatory variable, the idea of *reflection* of the manifest variable is used here for all the values of the explanatory variable. In the conventional notation, the *reflection* formula of the manifest explanatory variable $x_j = \xi_j + \delta_j$ (with measurement error δ_j) for $j = 1, 2, \cdots, n$, can be defined as

$$x_j^* = x_j \cos 2\psi + (y_j - \hat{\beta}_{0x}) \sin 2\psi, \tag{6.1}$$

where $\hat{\beta}_{0x}$ is the least square estimate of the intercept parameter, ψ is the angle measure defined as $\psi = \arctan \hat{\beta}_{1x}$ in which $\hat{\beta}_{1x}$ is the least square estimate of the slope parameter in the manifest model, and cos and sin are the usual trigonometric cosine and sine functions respectively. For the definition of *reflection* of points on the Euclidean plane see Vaisman (1997, p. 164-169).

The main difference between the RG method and Wald's original method is the use of the ranks of the transformed variable d_{1j} to divide the observations into two groups instead of using the ranks of the manifest explanatory variable.

The general motivation for using the reflection of x is that the true value of the latent explanatory variable is located at the middle of the observed value of the manifest variable x and its reflection x^*, if the ratio of error variances is $\lambda = 1$. We use Theorems 1 and 6 given in Chapter 3 to introduce the basic of the proposed (RG) method in this chapter.

The reflection group estimator takes a different form depending on the value of λ. There are three cases (1) $\lambda = 1$, (2) $\lambda > 1$, and (3) $\lambda < 1$. Therefore we suggest the grouping critera as follows.

$$\text{The grouping critera} = \begin{cases} \text{Case I} & d_{1j} = \frac{x_j + x_j^*}{2} \quad \text{if } \lambda = 1 \\ \text{Case II} & d_{2j} = \frac{d_{1j} + x_j}{2} \quad \text{if } \lambda > 1 \\ \text{Case III} & d_{3j} = \frac{d_{1j} + x_j^*}{2} \quad \text{if } \lambda < 1. \end{cases} \tag{6.2}$$

The main reason of using the transformed variables d_{1j}, d_{2j} and d_{3j} is that the latent explanatory variable ξ is located somewhere between the manifest explanatory variable x and its reflection variable x^*. Therefore, we suggest these variables in order to be close to the true variable ξ, where they are located between the manifest explanatory variable x and its reflection variable x^*. Moreover, it can be shown that the transformed variables d_{1j}, d_{2j} and d_{3j} produce estimators closer to the true slope parameter β_1 than the OLS estimator $\hat{\beta}_{1x}$. If we used the transformed variables d_{1j}, d_{2j} and d_{3j} as instrumental variables, then we get:

(1) The transformed variable $\mathbf{d_{1j}}$

Let d_{1j} be an instrumental variable, then the slope estimator is given by

$$\hat{\beta}_{1d_1} = \frac{\sum_{j=1}^{n}(y_j - \bar{y})d_{1j}}{\sum_{j=1}^{n}(x_j - \bar{x})d_{1j}} = \frac{S_{yd_1}}{S_{xd_1}}.$$

Note from Theorem 6 in Chapter 3 the sample covariance (S_{xx^*}) of x and x^* equals the sample variance (S_x^2) of x, then the sample covariance

(S_{xd_1}) of x and d_1 equals the sample variance (S_x^2) of x. That is,

$$\begin{aligned} c\hat{o}v(x, d_1) &= c\hat{o}v(x, \frac{(x+x^*)}{2}) \\ &= \frac{1}{2} c\hat{o}v(x, (x+x^*)) \\ &= \frac{1}{2}(c\hat{o}v(x, x) + c\hat{o}v(x, x^*)). \end{aligned}$$

From Theorem 6 in Chapter 3 $c\hat{o}v(x, x^*) = c\hat{o}v(x, x) = S_x^2$, then

$$\begin{aligned} S_{xd_1} &= c\hat{o}v(x, d_1) = \frac{1}{2}(c\hat{o}v(x, x) + c\hat{o}v(x, x)) \\ &= \frac{1}{2}(2c\hat{o}v(x, x)) \\ &= c\hat{o}v(x, x) = S_x^2. \end{aligned} \tag{6.3}$$

Then the slope estimator is given by

$$\hat{\beta}_{1d_1} = \frac{\sum_{j=1}^{n}(y_j - \bar{y})d_{1j}}{\sum_{j=1}^{n}(x_j - \bar{x})d_{1j}} = \frac{S_{yd_1}}{S_{xd_1}} = \frac{S_{yd_1}}{S_x^2}.$$

Note that this estimator has been introduced and examined in the previous chapter and was denoted as $\hat{\beta}_{1RM}$ (See Chapter 5, p. 124-131). From equation (5.12) of the previous chapter

$$\mid \hat{\beta}_{1x} \mid \leq \mid \hat{\beta}_{1RM} \mid = \mid \hat{\beta}_{1d1} \mid \leq \frac{S_y^2}{\mid S_{yx} \mid}. \tag{6.4}$$

(2) The transformed variable $\mathbf{d_{2j}}$

Similar to the transformed variable d_{1j}, it can be shown that the other estimator of slope $\hat{\beta}_{1d_2}$ using the transformed variable d_{2j}, is as follows

$$\hat{\beta}_{1d_2} = \frac{\sum_{j=1}^{n}(y_j - \bar{y})d_{2j}}{\sum_{j=1}^{n}(x_j - \bar{x})d_{2j}} = \frac{S_{yd_2}}{S_{xd_2}}.$$

Note from Theorem 6 in Chapter 3 and equation (6.3) that the sample covariance (S_{xd_2}) of x and d_2 equals the sample variance (S_x^2) of x.

It can be shown that the slope estimator $\hat{\beta}_{1d_2}$ is a greater than the OLS estimator $\hat{\beta}_{1x}$ as follows

$$\begin{aligned}
d_{2j} &= \frac{d_{1j}+x_j}{2} \\
2d_{2j} &= d_{1j}+x_j = \frac{x_j^*+x_j}{2}+x_j = \frac{1}{2}(x_j^*+3x_j) \\
4d_{2j} &= x_j^*+3x_j = x_j\cos 2\psi + y_j\sin 2\psi - \hat{\beta}_{0x}\sin 2\psi + 3x_j \\
&= x_j(1-2\sin^2\psi) + y_j\sin 2\psi - \hat{\beta}_{0x}\sin 2\psi + 3x_j \\
4d_{2j}-4x_j &= y_j\sin 2\psi - \hat{\beta}_{0x}\sin 2\psi - 2x_j\sin^2\psi \\
&= (y_j-\bar{y})\sin 2\psi + 2\bar{x}\sin^2\psi - 2x_j\sin^2\psi \\
&= (y_j-\bar{y})\sin 2\psi - 2(x_j-\bar{x})\sin^2\psi. \qquad (6.5)
\end{aligned}$$

Multiplying both sides of equation (6.5) by y_j, and taking the sum over j, we obtain

$$\begin{aligned}
4\sum_{j=1}^{n} y_j d_{2j} - 4\sum_{j=1}^{n} y_j x_j &= \sum_{j=1}^{n}(y_j-\bar{y})y_j\sin 2\psi \\
&\quad -2\sum_{j=1}^{n}(x_j-\bar{x})y\sin^2\psi. \qquad (6.6)
\end{aligned}$$

Based on Theorem 2 in Chapter 3 we have $\bar{d}_{2j} = \bar{d}_{1j} = \bar{x} = \bar{x}^* = \bar{\xi}$, and then by dividing both sides of the equation (6.6) by $(n-1)$, and adding $(4n\bar{x}\bar{y})$, $(-4n\bar{x}\bar{y})$ to the left side, we then have

$$\frac{(4\sum_{j=1}^{n} y_j d_{2j} - 4n\bar{x}\bar{y})}{(n-1)} - \frac{(4\sum_{j=1}^{n} y_j x_j - 4n\bar{x}\bar{y})}{(n-1)}$$

$$= \frac{\sum_{j=1}^{n}(y_j-\bar{y})y_j \sin 2\psi}{(n-1)} - \frac{2\sum_{j=1}^{n}(x_j-\bar{x})y \sin^2\psi}{(n-1)}.$$

Then

$$\begin{aligned}
4S_{yd_2} - 4S_{yx} &= S_y^2 \sin 2\psi - 2S_{yx}\sin^2\psi \\
S_y^2 \sin 2\psi &= 2S_{yx}\sin^2\psi + 4(S_{yd_2} - S_{yx}) \\
S_y^2 &= \frac{2S_{yx}\sin^2\psi}{\sin 2\psi} + \frac{4}{\sin 2\psi}(S_{yd_2} - S_{yx}) \\
S_y^2 &= \hat{\beta}_{1x}S_{yx} + \frac{4S_x^2}{\sin 2\psi}(\frac{S_{yd_2}}{S_x^2} - \frac{S_{yx}}{S_x^2}),
\end{aligned}$$

where $\hat{\beta}_{1d_2} = \frac{S_{yd_2}}{S_x^2}$, $\hat{\beta}_{1x} = \frac{2\sin^2\psi}{\sin 2\psi}$, and $S_x^2 = \frac{S_{yx}\cos\psi}{\sin\psi}$, and

$$\begin{aligned}
S_y^2 &= \hat{\beta}_{1x}S_{yx} + \frac{4S_x^2}{\sin 2\psi}(\frac{S_{yd_2}}{S_x^2} - \frac{S_{yx}}{S_x^2}) \\
&= \hat{\beta}_{1x}S_{yx} + \frac{2S_{yx}}{\sin^2\psi}(\hat{\beta}_{1d_2} - \hat{\beta}_{1x}). \qquad (6.7)
\end{aligned}$$

From (6.7) and when $S_{yx} > 0$, we then have

$$|\ \hat{\beta}_{1x}\ | \leq |\ \hat{\beta}_{1d_2}\ | \leq \frac{S_y^2}{|\ S_{yx}\ |}. \qquad (6.8)$$

(3) The transformed variable $\mathbf{d_{3j}}$

Similar to the transformed variables d_{1j} and d_{2j}, we can introduce another estimator of slope, $\hat{\beta}_{1d_3}$ using the transformed variable d_{3j} as

$$\hat{\beta}_{1d_3} = \frac{\sum_{j=1}^{n}(y_j-\bar{y})d_{3j}}{\sum_{j=1}^{n}(x_j-\bar{x})d_{3j}} = \frac{S_{yd_3}}{S_{xd_3}}.$$

Note from Theorem 6 in Chapter 3 and equation (6.3) the sample covariance (S_{xd_2}) of x and d_2 is equal to the sample variance (S_x^2) of x, and

$$\hat{\beta}_{1d_3} = \frac{S_{yd_3}}{S_x^2}.$$

Similarly, it can show that the slope estimator $\hat{\beta}_{1d_3}$ is greater than the OLS estimator $\hat{\beta}_{1x}$ as shown below

$$\begin{aligned}
d_{3j} &= \frac{d_{1j} + x_j^*}{2} \\
2d_{3j} &= d_{1j} + x_j^* = \frac{x_j^* + x_j}{2} + x_j^* = \frac{1}{2}(3x_j^* + x_j) \\
4d_{3j} &= 3x_j^* + x_j = 3x_j \cos 2\psi + 3y_j \sin 2\psi - 3\hat{\beta}_{0x} \sin 2\psi + x_j \\
&= 3x_j(1 - 2\sin^2\psi) + 3y_j \sin 2\psi - 3\hat{\beta}_{0x} \sin 2\psi + x_j \\
4d_{3j} - 4x_j &= 3y_j \sin 2\psi - 3\hat{\beta}_{0x} \sin 2\psi - 6x_j \sin^2\psi \\
&= 3(y_j - \bar{y}) \sin 2\psi + 6\bar{x} \sin^2\psi - 6x_j \sin^2\psi \\
&= 3(y_j - \bar{y}) \sin 2\psi - 6(x_j - \bar{x}) \sin^2\psi. \qquad (6.9)
\end{aligned}$$

Multiplying both sides of the equation (6.9) by y_j, and taking the sum over j, we obtain

$$\begin{aligned}
4\sum_{j=1}^{n} y_j d_{3j} - 4\sum_{j=1}^{n} y_j x_j &= 3\sum_{j=1}^{n}(y_j - \bar{y})y_j \sin 2\psi \\
&\quad -6\sum_{j=1}^{n}(x_j - \bar{x})y \sin^2\psi. \qquad (6.10)
\end{aligned}$$

Note from Theorem 2 in Chapter 3 we have

$$\bar{d}_{3j} = \bar{d}_{2j} = \bar{d}_{1j} = \bar{x} = \bar{x}^* = \bar{\xi},$$

then by dividing both sides of equation (6.10) by $(n-1)$, and adding $(4n\bar{x}\bar{y})$, $(-4n\bar{x}\bar{y})$ to the left side, we have

$$\begin{aligned}&\frac{(4\sum_{j=1}^{n} y_j d_{3j} - 4n\bar{x}\bar{y})}{(n-1)} - \frac{(4\sum_{j=1}^{n} y_j x_j - 4n\bar{x}\bar{y})}{(n-1)}\\ &= \frac{3\sum_{j=1}^{n}(y_j-\bar{y})y_j \sin 2\psi}{(n-1)} - \frac{6\sum_{j=1}^{n}(x_j-\bar{x})y \sin^2\psi}{(n-1)}.\end{aligned}$$

Then

$$\begin{aligned}4S_{yd_3} - 4S_{yx} &= 3S_y^2 \sin 2\psi - 6S_{yx}\sin^2\psi\\ 3S_y^2\sin 2\psi &= 6S_{yx}\sin^2\psi + 4(S_{yd_3} - S_{yx})\\ S_y^2 &= \frac{6S_{yx}\sin^2\psi}{3\sin 2\psi} + \frac{4}{3\sin 2\psi}(S_{yd_3} - S_{yx})\\ &= \hat{\beta}_{1x}S_{yx} + \frac{4S_x^2}{3\sin 2\psi}(\frac{S_{yd_3}}{S_x^2} - \frac{S_{yx}}{S_x^2})\\ &= \hat{\beta}_{1x}S_{yx} + \frac{4S_x^2}{3\sin 2\psi}(\hat{\beta}_{1d_3} - \hat{\beta}_{1x}). \qquad (6.11)\end{aligned}$$

From (6.11) and when $S_{yx} > 0$, then

$$|\hat{\beta}_{1x}| \leq |\hat{\beta}_{1d_3}| \leq \frac{S_y^2}{|S_{yx}|}. \qquad (6.12)$$

In order to examine the performance of the estimators $\hat{\beta}_{1d}$, $\hat{\beta}_{2d}$, and $\hat{\beta}_{3d}$ we should refer to the general property of the measurement error model that the true regression line always lies between the OLS line of y on x and the OLS line of x on y. That means the maximum likelihood (ML) estimator of $\hat{\beta}_1$ is often located in the following range (see for example Cheng and Van Ness, 1999, p. 11)

$$|\hat{\beta}_{1x}| \leq |\hat{\beta}_1| \leq \frac{S_y^2}{|S_{yx}|}. \qquad (6.13)$$

Then from equations (6.4), (6.8), and (6.12) the estimators $\hat{\beta}_{1d}$, $\hat{\beta}_{2d}$, and $\hat{\beta}_{3d}$ of the true slope all lie in the same range (6.13) as the ML estimator $\hat{\beta}_1$. Obviously, these estimators are greater than the OLS estimator $\mid \hat{\beta}_{1x} \mid$ of the slope. Note that we obtained these estimators ignoring the independent condition between the instrumental variable and measurement error as mentioned above. But in order to satisfy this condition we use the ranks of the transformed variables $(d's)$ above instead the rank of the manifest explanatory variable x in the Wald's method. As shown above that the estimators of the slope based on the transformed variables are closer to the true slope than the OLS estimator of the slope via x. The ranks of these variables could be very close to the rank of the true variable ξ when the values of the ratio of error variances (λ) are (1) $\lambda = 1$, (2) $\lambda > 1$, or (3) $\lambda < 1$, as shown in the next sections.

6.3.1 Propose modifications to Wald's method

(1) Case I when $\lambda = 1$

Let the second row of the instrumental variable matrix T_1' be based on the ranks of the transformed variable $d_{1j} = \dfrac{x_j + x_j^*}{2}$. The entries in the second row of T_1' is -1 if the value of d_{1j} is less then the median of d_{1j}, and $+1$

otherwise. A typical representation of T_1' is

$$T_1' = \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ -1 & -1 & -1 & \cdots & -1 \end{bmatrix}.$$

Then the first RG estimator (RG1) of β_1 and β_0 is given by

$$\hat{\beta}_{RG1} = (T_1'x)^{-1}T_1'y = \begin{bmatrix} n & 0 \\ 0 & \frac{n}{2}(\bar{x}_{12} - \bar{x}_{11}) \end{bmatrix}^{-1} \begin{bmatrix} n\bar{y} \\ \frac{n}{2}(\bar{y}_{12} - \bar{y}_{11}) \end{bmatrix} = \begin{bmatrix} \bar{y} \\ \frac{\bar{y}_{12}-\bar{y}_{11}}{\bar{x}_{12}-\bar{x}_{11}} \end{bmatrix}$$

Then

$$\hat{\beta}_{1RG1} = \frac{\bar{y}_{12} - \bar{y}_{11}}{\bar{x}_{12} - \bar{x}_{11}} \quad \text{and} \quad \hat{\beta}_{0RG1} = \bar{y} - \hat{\beta}_{1RG1}\bar{x},$$

where $\bar{y}_{11}$ is the mean of the first group of y, $\bar{y}_{12}$ is the mean of the second group of y, $\bar{x}_{11}$ is the mean of the first group of x, and $\bar{x}_{12}$ is the mean of the second group of x.

(2) Case II when $\lambda > 1$

Similarly, let the second row of the instrumental variable matrix T_2' be based on the ranks of the transformed variable $d_{2j} = \dfrac{d_{1j} + x_j}{2}$. The entries in the second row of T_2' is -1 if the value of d_{2j} is less than the median of d_{2j}, and $+1$ otherwise. The second RG estimator (RG2) of β_1 and β_0 is obtained as

$$\hat{\beta}_{1RG2} = \frac{\bar{y}_{22} - \bar{y}_{21}}{\bar{x}_{22} - \bar{x}_{21}} \quad \text{and} \quad \hat{\beta}_{0RG2} = \bar{y} - \hat{\beta}_{1RG2}\bar{x}.$$

where $\bar{y}_{21}$ is the mean of the first group of y, $\bar{y}_{22}$ is the mean of the second group of y, $\bar{x}_{21}$ is the mean of the first group of x, and $\bar{x}_{22}$ is the mean of the second group of x, these means are constructed based on the ranks of the transformed variable d_{2j}.

(3) Case III when $\lambda < 1$

Finally, let the second row of the instrumental variable matrix T_3' be defined based on the ranks of the transformed variable $d_{3j} = \dfrac{d_{1j} + x_j^*}{2}$. The entries in the second row of T_3' is -1 if the value of d_{3j} is less than the median of d_{3j}, and $+1$ otherwise. Then the third RG estimator (RG3) of β_0 and β_1 is defined as

$$\hat{\beta}_{1RG3} = \frac{\bar{y}_{32} - \bar{y}_{31}}{\bar{x}_{32} - \bar{x}_{31}} \quad \text{and} \quad \hat{\beta}_{0RG3} = \bar{y} - \hat{\beta}_{1RG3}\bar{x}.$$

where $\bar{y}_{31}$ is the mean of the first group of y, $\bar{y}_{32}$ is the mean of the second group of y, $\bar{x}_{31}$ is the mean of the first group of x, and $\bar{x}_{32}$ is the mean of the second group of x, these means are constructed based on the ranks of the transformed variable d_{3j}.

To implement the method, we omit the central ordered observation before computing $d's$ if n is odd. Although the second row of T_1', T_2', T_3' (that is, the sequence of -1 and $+1$) may appear similar, they will be different when the method is applied to any real dataset.

6.3.2 Example

To check the performance of the RG method estimator it would be useful to compare its performance with those of the Wald's, Geary's (a, b), OLS(y/x), and ML methods when the ratio of the error variances λ is known. This is done using the data set of Example 4.12 of Cheng and Van Ness (1999, p. 123). Here we set $\beta_1 = 1$ and $\beta_0 = 0$, and assume that the latent explanatory variable is distributed as $\xi_j \sim \chi^2_{(4)}$, $\lambda = 1$, and sample size $n = 36$. The results of the estimators are recorded in Table 6.1 below.

Table 6.1: Estimated β_1 and β_0 for different estimators when both variables are subject to measurement error, and $\lambda = 1$.

Methods	Slope estimate, $\hat{\beta}_1$	Intercept estimate, $\hat{\beta}_0$
RG1	1.0123	0.2186
Wald	0.8048	1.0630
Geary (a)	0.589	1.943
Geary (b)	0.721	1.406
OLS yx	0.8534	0.8653
ML	0.9277	0.5628

It is clear from Table 6.1 that the proposed RG1 estimator works well and its performance is much better than the other five estimators. Clearly Wald's estimator is biased and underestimates the slope parameter. But the RG1 method has improved the Wald's method significantly for the dataset of this example and also for another dataset as shown in the next section. Although the ML estimator is biased and underestimates the slope parameter, it is closer to the RG1 estimator, and slightly better than the others. It is not surprising, that both estimators of Geary (a) and (b) are strongly biased and underestimate the slope since it is well known that the estimators of fourth-order cumulants are unstable when the latent explanatory variable ξ is close to being symmetric (see Cheng and Van Ness, 1999, p. 127).

6.4 Simulation studies

We performed large scale simulations to illustrate that the proposed RG estimator is asymptotically unbiased and consistent whether the latent explanatory variable ξ has non-normal or normal distribution with flexible assumption about the knowledge of λ. Moreover, we demonstrate that the proposed RG method increases the efficiency of the Wald's method. The simulations are conducted for both study for non-normal distributions of the latent variable ξ, and a second study for normal distributions of the latent variable ξ.

6.4.1 First study: Non-normal distributions of ξ

Here we consider the case when the latent variable ξ has non-normal distribution. For this study we select ξ assuming uniform distribution within a specified interval. The parameters settings for the simulation studies are $\beta_1 = 1$ and $\beta_0 = 0$. We compare the estimated values and the mean absolute error of the proposed RG, Wald's, Geary's (a, b), and OLS of y on x estimators for selected arbitrary sample sizes $n = 20, 30, \cdots, 110$. The simulation is based on 10,000 replications, where ξ is assumed to follow an uniform distribution in the interval $[-5, 5]$:

(1) Case I when $\lambda = 1$ (d_{1j} is used for $RG1$), $\delta \sim N(0, 1)$, and $\epsilon \sim N(0, 1)$.

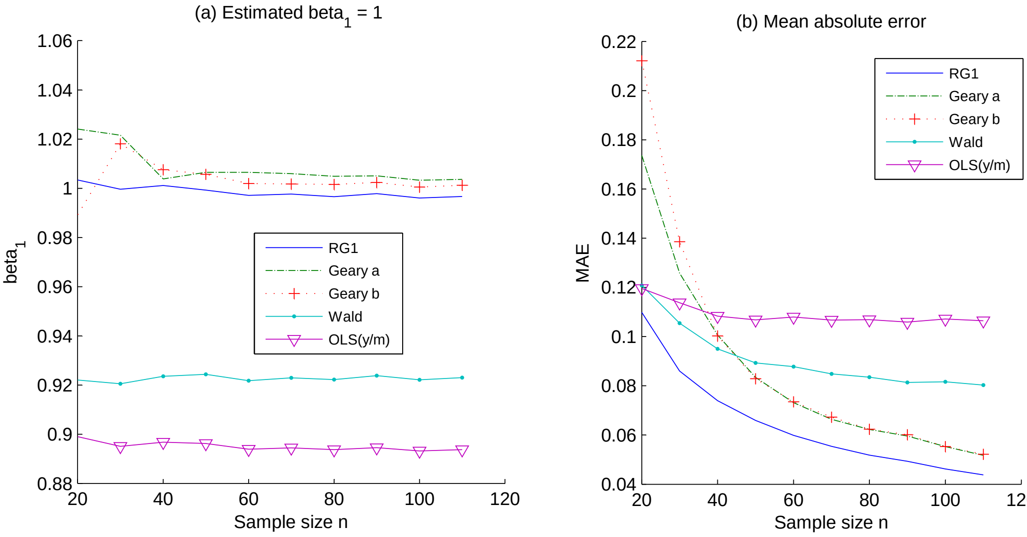

Figure 6.1: Graph of the estimated slope (a) and the mean absolute error (b) for five different estimators when $\lambda = 1$, and $\beta_1 = 1$.

(2) Case II when $\lambda > 1$ (d_{2j} is used for $RG2$), $\delta \sim N(0,1)$, and $\epsilon \sim N(0,2.25)$.

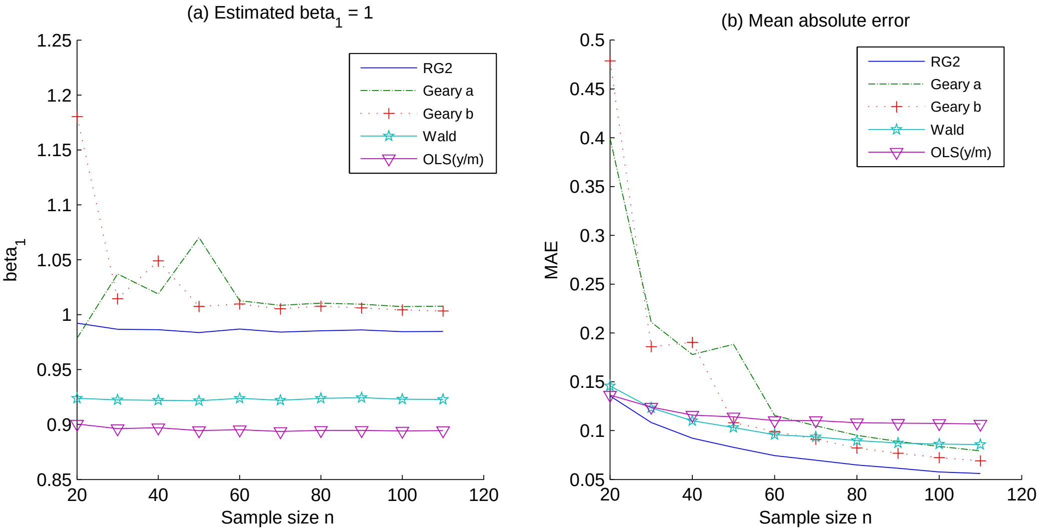

Figure 6.2: Graph of the estimated slope (a) and the mean absolute error (b) for five different estimators when $\lambda > 1$, and $\beta_1 = 1$.

From Figures 6.1(a), 6.2(a), and 6.3(a) the values of the OLS estimator for the slope are the lowest and far below the true value of $\beta_1 = 1$. The values of Wald's estimator are also away from the true value of β_1, but they appear to be slightly closer to the true value of β_1 than those predicted by the OLS estimator. The values of the two estimators of Geary are both close to the true value of the slope if the sample size is large ($n = 40$ or more), but they fluctuate significantly if n is small. Clearly the RG estimator is much closer to the true value of β_1 than the Geary estimators. In fact, the proposed RG

(3) Case III when $\lambda < 1$ (d_{3j} is used for $RG3$), $\delta \sim N(0, 2.25)$, and $\epsilon \sim N(0, 1)$.

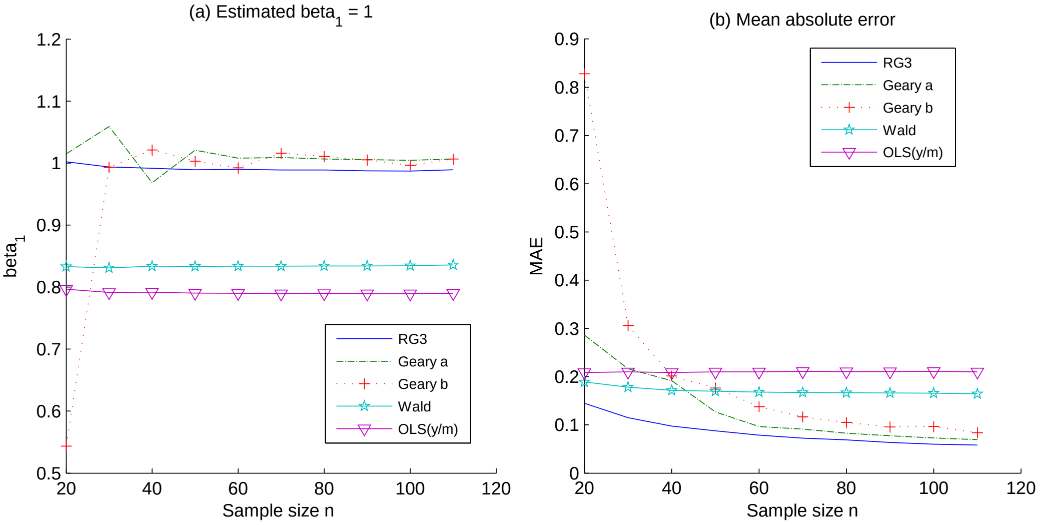

Figure 6.3: Graph of the estimated slope and the mean absolute error for five different estimators when $\lambda < 1$, and $\beta_1 = 1$.

method estimator is consistently closest to the true value of β_1 for all sample sizes.

It is clear from Figures 6.1(b), 6.2(b), and 6.3(b) that the presence of the measurement error makes the mean absolute error of the OLS estimator the largest. While the mean absolute error of the Wald's estimator appears to be smaller than that of the OLS estimator, but it is not the smallest. The mean absolute error for the Geary's estimators (or cumulant method estimators) are smaller than that of Wald's and OLS estimators. But if the sample size is small, then the mean absolute error of the Geary estimators become unstable. Obviously, the mean absolute error of the RG method estimator is the smallest compared to the other estimators, and is stable over the range of selected sample sizes. Thus the RG method estimator performs better than the other estimators in terms of having smallest MAE when ξ follows a non-normal distribution.

6.4.2 Second study: Normal distributions of ξ

Here we now assume that the latent variable ξ follows a normal distribution.This simulation study compares the RG method estimator with the ML, Wald's, and OLS estimators of y on x when λ is misspecified. The simulation is based on $10,000$ replications using MATLAB software. We use different sample sizes to show the behavior of the above estimators for selected sample sizes $50, 100, 150, 200, \cdots, 450$. In the simulation study we consider misspecification of λ as (1) using the incorrect value of $\lambda = 0.5$, instead of the correct value of 1, (2) using the incorrect value of $\lambda = 2$, instead of the correct value of 2.89, and (3) using the incorrect value of $\lambda = 1.5$, instead of the correct value of 0.5.

(1) Case I uses the incorrect value of $\lambda = 0.5$, for $\xi \sim N(0, 49)$, $\beta_0 = 0$, and $\beta_1 = -0.8$, if the correct value of $\lambda = 1 = \dfrac{9}{9}$ is unavailable. The graph of the estimated slopes is given in Figure 6.4.

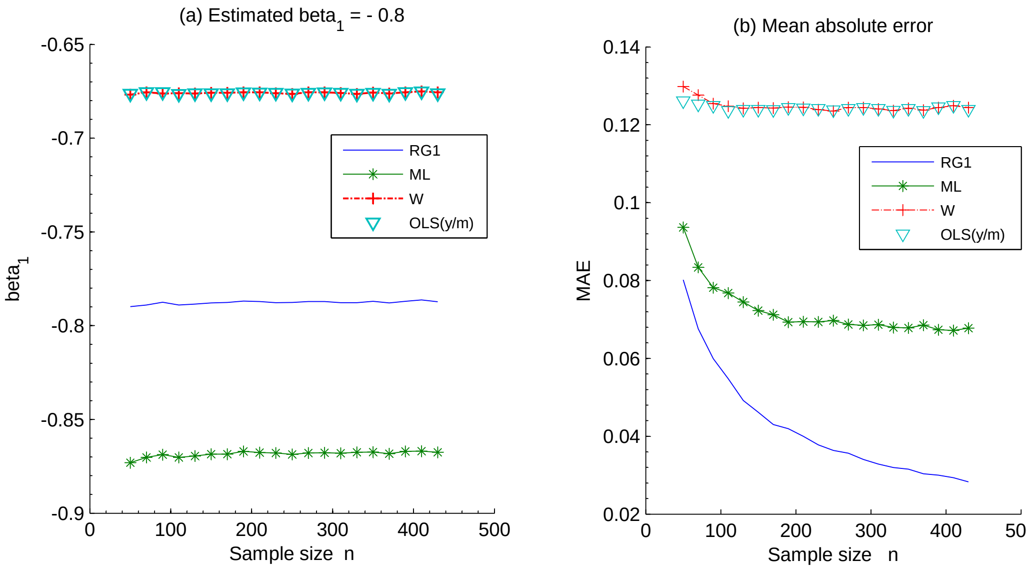

Figure 6.4: Graphs of the estimated slope (a) and the mean absolute error (b) for four different estimators RG_1, ML, W, and OLS for case I.

(2) Case II uses the incorrect value of $\lambda = 2$, for $\xi \sim N(0, 49)$, $\beta_0 = 0$, and $\beta_1 = 0.6$, if the correct value of $\lambda = 2.89 = \dfrac{72.25}{25}$ is unavailable. The graph of the estimated slopes is given in Figure 6.5.

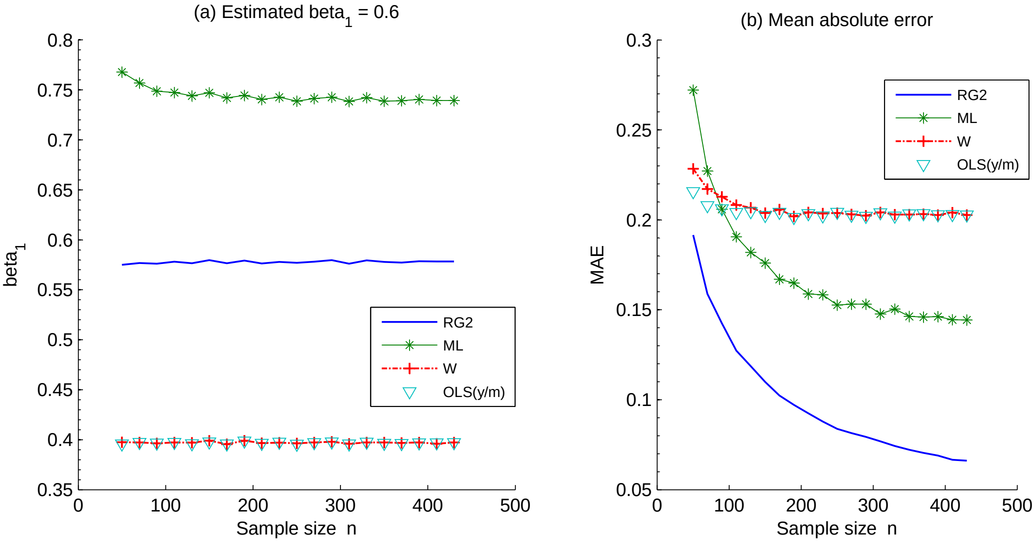

Figure 6.5: Graphs of the estimated slope (a) and the mean absolute error (b) for four different estimators RG_2, ML, W, and OLS for case II.

(3) **Case III** uses the incorrect value of $\lambda = 1.5$, for $\xi \sim N(0, 36)$, $\beta_0 = 0$, and $\beta_1 = 1.4$, if the correct value of $\lambda = 0.5 = \dfrac{8}{16}$ is unavailable. The graph of the estimated slopes is given in Figure 6.6.

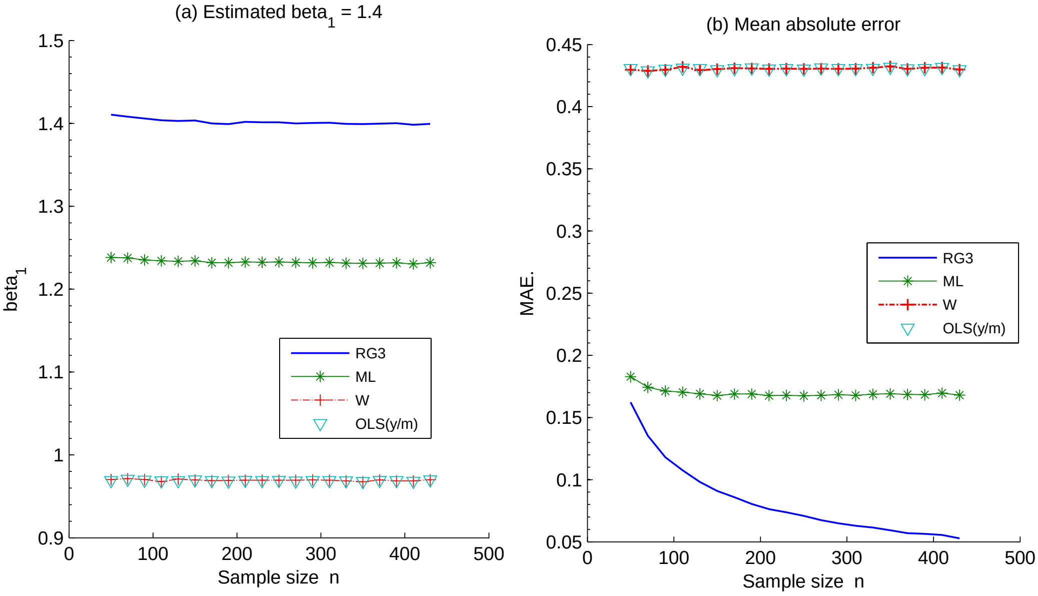

Figure 6.6: Graphs of the estimated slope and the mean absolute error for four different estimators RG_3, ML, W, and OLS for case III.

Figures 6.4-6.6 show the estimated slope and the mean absolute error for four different estimators. From each of the above graphs it is evident that the RG method estimator (RG1, RG2, RG3) is consistently better than the other three estimators. This superior performance of the RG method estimator does not depend on the accuracy of selecting the value of λ beyond the knowledge of less than or greater than or equal to 1.

6.5 Concluding remarks

This chapter proposes a new grouping method based on the rank of the reflection of the manifest explanatory variable as an improvement to Wald's estimator. It proposes specific modifications to Wald's grouping method of fitting a straight line when both variables are subject to measurement errors. The RG estimator works under the same assumptions as Wald's method, but without requiring the restriction that the error terms to be too small or large. The main difference of the RG method from Wald's original method is the use of ranks of a new transformed variable d_{1j} to divide the observations of the manifest explanatory variable x into two groups instead of using the ranks of x itself.

Extensive simulation studies were conducted to compare the RG estimator with existing alternative estimators when the latent variable ξ follows both non-normal and normal distributions. The comparison was done in terms of the estimated values of the slope parameter as well as the mean absolute error of the estimators. The graphical and numerical analyses provide clear evidence that the RG method estimator is more precise than the other competing estimators. Therefore, the proposed RG estimator possesses better statistical proprieties than the OLS estimator, as well as the grouping method proposed by Wald's, and cumulant based estimators introduced by Geary. The new method is stable and works well for different sample sizes and for different values of λ. It is clear, from the forgoing discussion that the reflection grouping method significantly increases the efficiency of Wald's method. The simulation study also confirms that the RG method estimator performs much better than the maximum likelihood estimator when ξ follows a normal distribution. This superior performance occurs even when the exact value of the ratio of error variances λ is unavailable. Simulations with other choices of the parameters (slope and intercept), sample sizes and number of replications produced similar results demonstrating the consistency and superior performance of the proposed RG estimator.

Chapter 7

Weighted geometric mean estimator

7.1 Introduction

This chapter introduces a new estimator to fit regression line when both variables are subject to measurement error. It provides an alternative view on the geometric mean estimator. The proposed estimator is based on the mathematical relationship between the vertical and orthogonal distances of the observed points and the fitted regression line. It minimises the orthogonal distance and is less sensitive to the ratio of error variances (λ). The simulation results show that the proposed estimator is more consistent and efficient than the geometric mean and OLS-bisector estimators.

Dent (1935) suggested the geometric mean functional relationship estimator to be as a solution of the likelihood equations when there is no additional information in the case of the normal functional model (cf Cheng and Ness, 1999, p. 43). This estimator is called geometric mean (GM) estimator, because it is the geometric mean of the least squares estimators of the slope for the regression of y on x and the reciprocal of that of x on y. This technique has been introduced many times under different names such as the reduced major axis, or the least products regression (cf Ludbrook, 2010).

Halfon (1985) and Draper and Yang (1997) pointed out that the geometric mean estimator minimises the vertical and horizontal distances between the observed points and the regression line. Richard (2009) criticised that the geometric mean (GM) estimator is widely used in the literature without explaining why it was selected. Jolicoeur (1975) stated that it is difficult to interpret the meaning of the slope of the geometric mean regression. Isobe et al. (1990) examined five linear methods, and pointed out that the OLS bisector (OLS-b) estimator is the best method to use, when there is no basis to distinguish between the explanatory and response variables.

The next section presents the mathematical relationship between the vertical and orthogonal distances of the observed points from both the fitted and unfitted regression lines. The geometric mean estimator, and an alternative way to derive this estimator, are provided in Sections 7.3 and 7.4. The proposed weighted geometric mean estimator is introduced in Section 7.5. The simulation studies, and the concluding remarks are included in Sections 7.6 and 7.7.

7.2 Relationship between the vertical and orthogonal distances

It is well known that there are different approaches to minimise the vertical, horizontal, orthogonal, or both orthogonal and horizontal, distances in regression analysis. The ordinary least squares method works on the basis of minimising the vertical distance when there are no measurement errors. Inverse least squares method minimises the horizontal distance when there is measurement error only in the explanatory variable (cf Leng et al. 2007). The orthogonal regression approach minimises the orthogonal distance under the assumption that the ratio of error variances is equal to one, that is, $\lambda = \sigma_\epsilon^2 \sigma_\delta^{-2} = 1$. The maximum likelihood estimator minimises both the horizontal and orthogonal distances when λ is known (cf Leng et al. 2007).

It is crucial to note the difference between the distance from the observed point and the fitted line, the unfitted line, and the unobserved point. Although, many authors use distance between the observed point and regression line without being specific. This issue is crucial when there are measurement errors in both variables. This section introduces the mathematical relationship between the vertical and orthogonal distances of the observed points and the fitted regression line.

Let (x_j, y_j) be the observed point and (ξ_j, η_j) be the associated unobserved point. Then the fitted line is given by

$$\eta_j = \beta_0 + \beta_1 \xi_j, \quad j = 1, 2, \cdots, n. \tag{7.1}$$

Note that all the true points (ξ_j, η_j) are on the fitted line (7.1), because there is no equation error in the model.

Now we will have two different reflection points for the observed point (x_j, y_j) one about the fitted line and other about the unfitted line. Therefore, we define the reflection point (A_j, B_j) of the observed point (x_j, y_j) about the fitted line (7.1) as follows:

$$A_j = x_j \cos 2\theta + (y_j - \beta_0) \sin 2\theta, \tag{7.2}$$

$$B_j = x_j \sin 2\theta - (y_j - \beta_0) \cos 2\theta + \beta_0, \tag{7.3}$$

where $\theta = \tan^{-1} \beta_1$, and β_0, and β_1 are the regression parameters. For details on reflection of points please see Vaisman (1997, p. 164-169). For simplicity, we consider the relationships between the orthogonal and vertical distance of the observed point (x_j, y_j) and the fitted line $(\eta_j = \beta_0 + \beta_1 \xi_j)$ as a first case. While the second case is related to the relationship between the observed point (x_j, y_j) and the unfitted line $(y_j = \beta_0 + \beta_1 x_j)$.

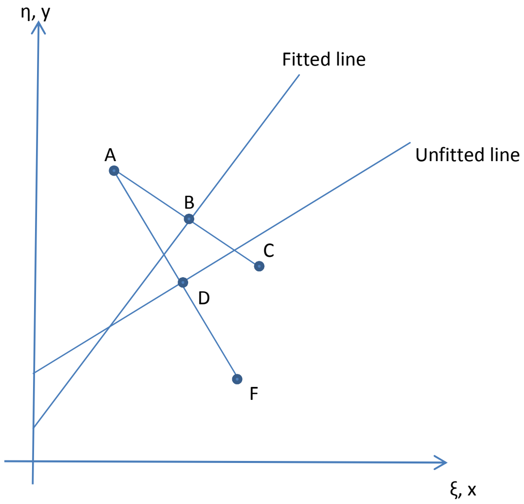

Figure 7.1: Graph of two orthogonal distances ($\overline{AB} = Od$, and $\overline{AD} = Ox$) between the observed point and the fitted and unfitted lines.

There are potentially two orthogonal distances of any observed point, one from the *fitted line* (here represented by Od) and the other from the *unfitted line* (Ox). In principle, the GM method should minimise Od, but in practice it minimises Ox. Figure 7.1 shows the reflection of $A = (x_j, y_j)$ about the *fitted line* $C = (A_j, B_j)$ with the orthogonal distance $Od = \overline{AB}$, and the reflection of $A = (x_j, y_j)$ about the *unfitted line* $F = (x_j^*, y_j^*)$ with the orthogonal distance $Ox = \overline{AD}$.

7.2.1 Fitted line case

From the properties of the reflection the fitted line (the reflection line) is a bisector and perpendicular on the distance between the observed point A, (x_j, y_j), and its reflection point C, (A_j, B_j). Then the half of the square distance between the observed point (x_j, y_j) and its reflection point (A_j, B_j) will equal the orthogonal square distance (Od_j^2) between the observed point (x_j, y_j) and the fitted line. The orthogonal distance is given by

$$Od_j = \frac{1}{2}\left(\sqrt{(A_j - x_j)^2 + (B_j - y_j)^2}\right). \tag{7.4}$$

Then from (7.2) and (7.3) the square orthogonal distance (Od_j^2) is given by

$$\begin{aligned} Od_j^2 &= \frac{1}{4}((2x_j \sin^2\theta + y_j \sin 2\theta - \beta_0 \sin 2\theta)^2 \\ &\quad +(x_j \sin 2\theta - 2y_j \cos^2\theta + 2\beta_0 \cos^2\theta)^2). \end{aligned}$$

Since $x_j = \xi_j + \delta_j$, $y_j = \eta_j + \epsilon_j$ and $\beta_1 = \dfrac{\sin\theta}{\cos\theta}$ so

$$\begin{aligned}
Od_j^2 &= \frac{1}{4}\Bigg((-2x_j \sin^2\theta + \beta_1 \xi_j \sin 2\theta + \epsilon_j \sin 2\theta)^2 \\
&\quad +(x_j \sin 2\theta - 2\beta_1 \xi_j \cos^2\theta - 2\epsilon_j \cos^2\theta)^2\Bigg) \\
&= \frac{1}{4}\Bigg((2\delta_j \sin^2\theta - \epsilon_j \sin 2\theta)^2 + (\delta_j \sin 2\theta - 2\epsilon_j \cos^2\theta)^2\Bigg) \\
&= \frac{1}{4}\Bigg(\delta_j^2(4\sin^4\theta + \sin^2 2\theta) - 4\delta_j\epsilon_j(\sin 2\theta \sin^2\theta + \sin 2\theta \cos^2\theta) \\
&\quad +e_j^2(4\cos^4\theta + \sin^2 2\theta)\Bigg) \\
&= u_j^2 \sin^2\theta - \delta_j\epsilon_j \sin 2\theta + e_j^2 \cos^2\theta.
\end{aligned}$$

From (7.3) $E(\delta_j) = E(\epsilon_j) = 0$ and $E(\delta_j\epsilon_j) = 0$, then

$$E(Od_j^2) \quad = \quad E(\delta_j^2)\sin^2\theta + E(\epsilon_j^2)\cos^2\theta.$$

From Theorems 2 and 3 in Chapter 3, $E(A_j - x_j) = E(B_j - y_j) = 0$, then the variance of Od is

$$\sigma_{Od}^2 \quad = \quad \sigma_\delta^2 \sin^2\theta + \sigma_\epsilon^2 \cos^2\theta.$$

From (7.5), and noting $\beta_1^2 = \sin^2\theta\ \cos^{-2}\theta$, the above variance becomes

$$\sigma_{Od}^2 \quad = \quad (\sigma_\epsilon^2 + \sigma_\delta^2 \sin^2\theta \cos^{-2}\theta)\cos^2\theta = (\sigma_\epsilon^2 + \beta_1^2\sigma_\delta^2)\cos^2\theta.$$

Then the relationship between the variance of the orthogonal distance and the variance of vertical distance is given by

$$\sigma_{Od}^2 = \sigma_v^2 \cos^2\theta = \frac{\sigma_v^2}{1+\beta_1^2}, \tag{7.5}$$

where $\beta_1^2 = \sin^2\theta / \cos^2\theta$ then

$$\begin{aligned} \frac{1}{1+\beta_1^2} &= \frac{1}{1+\frac{\sin^2\theta}{\cos^2\theta}} \\ &= \frac{1}{\frac{\cos^2\theta+\sin^2\theta}{\cos^2\theta}} \\ &= \frac{1}{\frac{1}{\cos^2\theta}} \\ &= cos^2\theta. \end{aligned}$$

Note that both vertical and orthogonal distances measure the distance between the observed point (x_j, y_j) and the fitted line, but it does not measure the distance between the observed point (x_j, y_j) and the unobserved point (ξ_j, η_j). Under certain assumptions such as $\lambda = 1$ or $\beta_1 = 1$ the distance between the observed point and the unobserved point is equal to twice that of the orthogonal distance, where the distance between the observed point and the unobserved point (Pd) is given by

$$Pd = \sqrt{(x_j - \xi_j)^2 + (y_j - \eta_j)^2} = \sqrt{\delta_j^2 + e_j^2},$$

where δ_j, ϵ_j are the measurement error in the explanatory and response variables respectively. From (7.3) the variance of the (Pd) distances is

$$\sigma_{Pd}^2 = \sigma_\epsilon^2 + \sigma_\delta^2.$$

From (7.5) and when $\lambda = 1$,

$$\sigma_{Pd}^2 = 2\sigma_\epsilon^2.$$

7.2.2 Unfitted line case

In order to find the relationship between the observed point (x_j, y_j) and the unfitted line we follow the similar steps of the first case except replacing the parameters of the fitted line, $\theta = \tan^{-1}\beta_1$, β_0, and β_1 with the coefficients of the unfitted line $\psi = \tan^{-1}\hat{\beta}_{1x}$, $\hat{\beta}_{0x}$, and $\hat{\beta}_{1x}$ respectively. Then we get

$$x_j^* = x_j \cos 2\psi - (y_j - \hat{\beta}_{0x}) \sin 2\psi$$

$$y_j^* = x_j \sin 2\psi - (y_j - \hat{\beta}_{0x}) \cos 2\psi + \hat{\beta}_{0x},$$

where (x_j^*, y_j^*) is the reflection point of the observed point (x_j, y_j) about the unfitted line.

The relationship between the sample variance of the orthogonal distance (Ox) and vertical distance (v) of the model (1.4) in Chapter 1, is given by

$$S_{Ox}^2 \quad = \quad S_v^2 \cos^2 \psi = \frac{S_v^2}{1+\hat{\beta}_{1x}^2}. \tag{7.6}$$

Also the relationship between observed point and unfitted line of the population becomes

$$\sigma_{Ox}^2 \quad = \quad \sigma_v^2 \cos^2 \psi = \frac{\sigma_v^2}{1+\beta_{1x}^2}. \tag{7.7}$$

From (7.5) and (7.7) the relationship between the orthogonal distance of the two cases is

$$\sigma_{Od}^2 \quad = \quad \sigma_{Ox}^2 \, \frac{\cos^2 \theta}{\cos^2 \psi} = \sigma_{Ox}^2 \left(\frac{1+\beta_{1x}^2}{1+\beta_1^2} \right). \tag{7.8}$$

Note that in general, $\sigma_{Od}^2 < \sigma_{Ox}^2$, because of $\beta_{1x}^2 < \beta_1^2$, and $\sigma_{Od}^2 = \sigma_{Ox}^2$ if and only if there is no measurement error, in which case $\beta_{1x}^2 = \beta_1^2$. Therefore, there is a difference between the method which aims to minimise σ_{Ox}^2 and that minimises σ_{Od}^2. The next section will show that the GM method is minimising σ_{Ox}^2, rather than σ_{Od}^2.

7.3 Geometric mean estimator

One of the simple approaches to handle the measurement error in the regression analysis is the geometric mean (GM) functional relationship, initially proposed by Teissier (1948) and later by Barker et al. (1988) (cf Draper and Yang, 1997). This estimator has frequently been mentioned in the literature for two reasons. First, when there is no basis for distinguishing between the response and explanatory variables. Second, to handle the measurement error when no prior information is available. The geometric mean regression method is widely used in fisheries studies (cf Richard, 2009). It has received much attention from the experts, and some have suggested that it is more useful than the ordinary least squares method (see Sprent and Dolby, 1980).

The geometric mean estimator of the slope is the geometric mean of the slope of y on x regression line, and the reciprocal of the slope of x on y regression line, where x and y both are random (see Leng et al. 2007). It is given by

$$\hat{\beta}_{1G} = sgn(SP_{xy})\sqrt{\frac{SS_y}{SS_x}} = sgn(Sp_{xy})\left(\frac{S_y}{S_x}\right) ,$$

where $SS_x = \sum_{j=1}^{n}(x_j - \bar{x})^2$, $SS_y = \sum_{j=1}^{n}(y_j - \bar{y})^2$,

$SP_{xy} = \sum_{j=1}^{n}(x_j - \bar{x})(y_j - \bar{y})$, and S_y and S_x are the sample standard deviations of y and x respectively.

In the literature, the geometric mean regression is known as the standardised major axis (MA) estimator (cf Warton et al. 2006). It is also known as the reduced major axis (RMA), or the line of organic correlation (cf Tessier, 1948; Kermack and Haldane, 1950; Ricker, 1973). In physics it is known as a type of standard weighting model (see Machonald and Thompson, 1992), while astronomers call it as Strömberg's impartial line (see Feigelson and Babu, 1992).

A host of recent publications indicate that using the GM or RMA method is necessary and sufficient to fit the straight line when both the response and explanatory variables are subject to errors (see Levinton and Allen, 2005, Zimmerman et al. 2005, Sladek et al. 2006, and Vincent and Lailvaux, 2006). While Jolicoeur (1975), and Spernt and Dolby (1980) pointed out that the GM estimator is unbiased if and only if

$$\lambda = \frac{\sigma_y^2}{\sigma_x^2} \quad \text{or} \quad \lambda = \beta_1^2 \ .$$

But several other studies indicate that this assumption is unrealistic (cf Sprent and Dolby, 1980).

It is commonly recommended to use the geometric mean estimator without mentioning the justifications (Smith, 2009). Jolicoeur (1975) stated that it is difficult to interpret the meaning of the slope of the geometric mean regression. However, the common belief is that the geometric mean regression minimises the vertical and horizontal distances between the observed points and the fitted line (Halfon, 1985; and Draper and Yang, 1997). But it is not quite true, because it could be interpreted that the GM or RMA estimator minimises the orthogonal error of the observed points with the unfitted regression line instead of the fitted regression line as shown in the next section.

7.4 Alternative view on the geometric mean estimator

From (7.7) the variance of the orthogonal distance between the observed point (x_j, y_j) and the unfitted line $(\hat{y}_j = \hat{\beta}_{0x} + \hat{\beta}_{1x} x_j)$ can be derived for the geometric mean estimator as follows:

$$\begin{aligned} SS_{Ox} &= SS_v \cos^2 \psi = \frac{1}{n-1} \sum_{j=1}^{n} (y_j - \hat{\beta}_{0x} - \hat{\beta}_{1x} x_j)^2 \cos^2 \psi \\ &= \sum_{j=1}^{n} ((y_j - \bar{y}) - \hat{\beta}_{1x}(x_j - \bar{x}))^2 \cos^2 \psi \\ &= \sum_{j=1}^{n} ((y_j - \bar{y}) \cos \psi - (x_j - \bar{x}) \sin \psi)^2. \end{aligned} \tag{7.9}$$

Let $L_1 = \sin \psi$, and $L_2 = \cos \psi$. Then

$$SS_{Ox} = \sum_{j=1}^{n} ((y_j - \bar{y}) L_2 - (x_j - \bar{x}) L_1)^2.$$

Differentiating of SS_{Ox} w.r.t. L_1, and L_2 and setting them equal to zero, we get

$$\begin{aligned} \frac{\partial SS_{Ox}}{\partial L_1} &= 2 \sum_{j=1}^{n} ((y_j - \bar{y}) L_2 - (x_j - \bar{x}) L_1)(-(x_j - \bar{x})) = 0, \\ \text{which gives} \quad & L_1 S_x^2 = L_2 S_{yx}, \text{ and} \qquad (7.10) \\ \frac{\partial SS_{Ox}}{\partial L_2} &= 2 \sum_{j=1}^{n} ((y_j - \bar{y}) L_2 - (x_j - \bar{x}) L_1)(y_j - \bar{y}) = 0, \\ \text{to give} \quad & L_2 S_y^2 = L_1 S_{yx}. \qquad (7.11) \end{aligned}$$

From (7.10), (7.11), and $\hat{\beta}_{1x} = \frac{L_1}{L_2}$ we get two estimators of the slope

$$\hat{\beta}_1 = \frac{S_{yx}}{S_x^2} \text{ and } \hat{\beta}_2 = \frac{S_y^2}{S_{yx}}. \tag{7.12}$$

Then the geometric mean of the estimators in (7.12) is the GM estimator, that is,

$$\hat{\beta}_{1G} = sgn\{S_{yx}\} \sqrt{\frac{S_y^2}{S_x^2}}.$$

Obviously, the above GM estimator is derived by minimising the orthogonal distance between the observed point (x_j, y_j) and unfitted line. Therefore, it does not minimise the distance between the observed point (x_j, y_j) and the fitted regression line.

7.5 Proposed estimator

The proposed estimator minimises the orthogonal distance between the observed point (x_j, y_j) and the unfitted regression line. This estimator is based on the relationship (7.7) between the vertical and orthogonal distances of the observed points and the unfitted regression line, then derives the proposed estimator from both equations (7.10) and (7.11) as follows:

Multiply equation (7.10) by S_y^2, and equation (7.11) by S_{yx}, then we get

$$L_1 S_x^2 S_y^2 = L_2 S_{yx} S_y^2, \tag{7.13}$$

$$L_1 S_{yx}^2 = L_2 S_{yx} S_y^2. \tag{7.14}$$

From equation (7.13) and adding equation (7.14) we get

$$\begin{aligned} L_1(S_x^2 S_y^2 + S_{yx}^2) &= L_2 2 S_{yx} S_y^2, \\ (S_x^2 S_y^2 + S_{yx}^2) sin\psi &= 2 S_{yx} S_y^2 cos\psi. \end{aligned}$$

Hence the proposed weighted geometric (WG) mean estimator is given by

$$\hat{\beta}_{1WG} = \frac{\sin\psi}{\cos\psi} = \frac{2S_{yx}S_y^2}{S_y^2S_x^2 + S^2yx}. \tag{7.15}$$

This estimator could be simplified as follows

$$\begin{aligned}\hat{\beta}_{1WG} &= \frac{2S_y^2S_x^{-2}}{S_y^2S_{yx}^{-1} + S_{yx}S_x^{-2}} \\ &= \frac{2\hat{\beta}_{1G}^2}{(\hat{\beta}_1 + \hat{\beta}_2)} \\ &= \mathcal{W}\ \hat{\beta}_{1G}, \end{aligned} \tag{7.16}$$

where $\mathcal{W} = \dfrac{\hat{\beta}_{1G}}{\hat{\beta}_{OLS-mean}}$, in which $\hat{\beta}_{OLS-mean}$ is obtained by taking the arithmetic mean of the slopes of the two ordinary least squares regression lines of OLS(yx) and OLS(xy). Note if the geometric mean (GM) estimator is equal to the OLS-mean estimator, then the proposed (WG) estimator is equal to both the geometric mean and the OLS-mean estimators, where $\mathcal{W}$ is equal to one.

The reasons for suggesting the weighted geometric mean estimator instead of the geometric mean estimator, or the OLS-bisector estimator will become apparent from the results of the next section.

7.6 Simulation studies

In this section we compare the proposed (WG) estimator with the geometric mean estimator and the OLS-bisector estimator for a wide range of values of λ $(0.08 \leq \lambda \leq 100)$.

We perform large scale simulations to illustrate that the proposed estimator

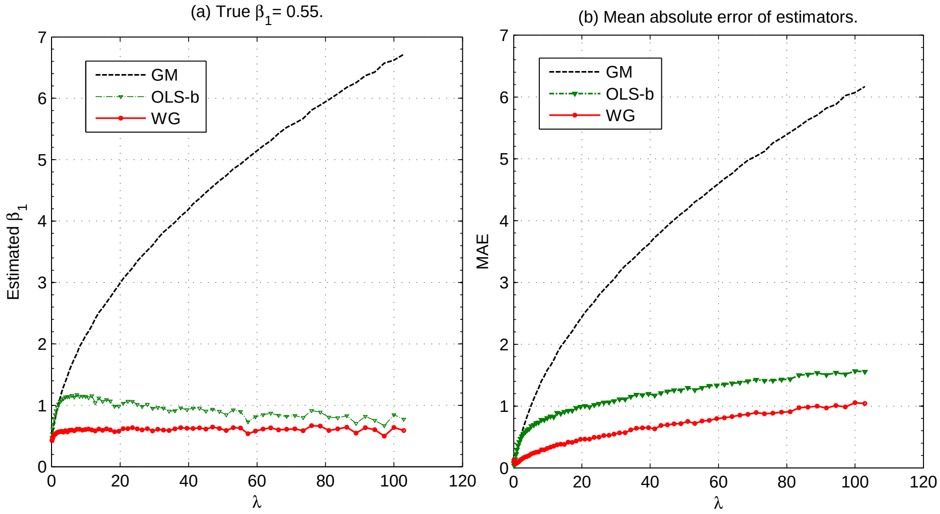

Figure 7.2: Graph of three estimators of the slope, and the mean absolute error when $\beta_0 = 20, \beta_1 = 0.55$ and $0.08 \leq \lambda \leq 100$.

is asymptotically unbiased and consistent compared to the geometric mean estimator and the OLS-bisector estimator. The latter estimator is given by

$$\hat{\beta}_{1OLS-b} = (\hat{\beta}_1 + \hat{\beta}_2)^{-1} \left[\hat{\beta}_1\hat{\beta}_2 - 1 + \sqrt{(1+\hat{\beta}_1^2)(1+\hat{\beta}_2^2)} \right],$$

where $\hat{\beta}_1 = \dfrac{S_{yx}}{S_x^2}$, and $\hat{\beta}_2 = \dfrac{S_y^2}{S_{xy}}$ (see Isobe et al. 1990).

This study demonstrates that the WG estimator is not sensitive to the ratio of error variances λ, whereas the geometric mean estimator grows larger as the value of λ increases.

The dataset is based on 1000 replications of samples size 100 of normal structural model simulated as follows:

1. Generate 100 independent values $\xi_1, \cdots, \xi_{100}$ of $\xi \sim N(0,\, 8)$.

2. Generate 100 independent values $\delta_1, \cdots, \delta_{100}$ of $\delta \sim N(0,\, 7)$.

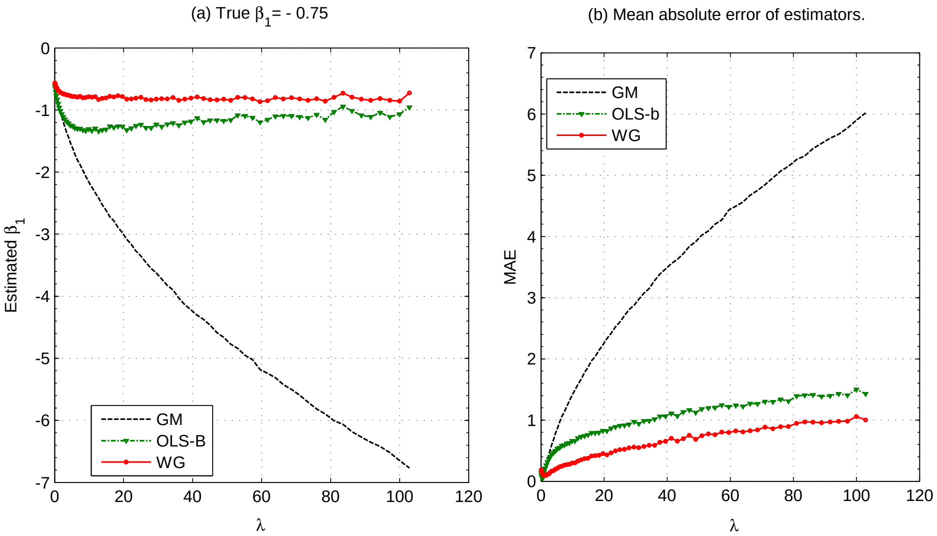

Figure 7.3: Graph of three estimators of the slope, and the mean absolute error when $\beta_0 = 27, \beta_1 = -0.75$ and $0.08 \leq \lambda \leq 100$.

3. Generate 100 independent values $\epsilon_1, \cdots, \epsilon_{100}$ of $\epsilon \sim N(0, \sigma_\epsilon)$, where $2 \leq \sigma_\epsilon \leq 71$, for each 1000 replications it is increased by 1.

4. The estimators of the slope and the mean absolute error are calculated using the MATLAB software.

From Figures 7.2(a)-7.4(a), the values of the OLS-bisector estimator are not the same as the true values of β_1, but are much better than those of the geometric mean estimator. The values of the GM estimator are far above the true values of β_1. The GM estimator appears to be an overestimate of the slope and so far away from the true value β_1. Clearly the proposed WG estimator is much closer to the true values of β_1 than the other two estimators. It is clear, from Figures 7.2(b)-7.4(b) that the measurement error makes the mean absolute error of the GM estimator the highest. While the mean absolute error of the OLS-bisector estimator appears to be better

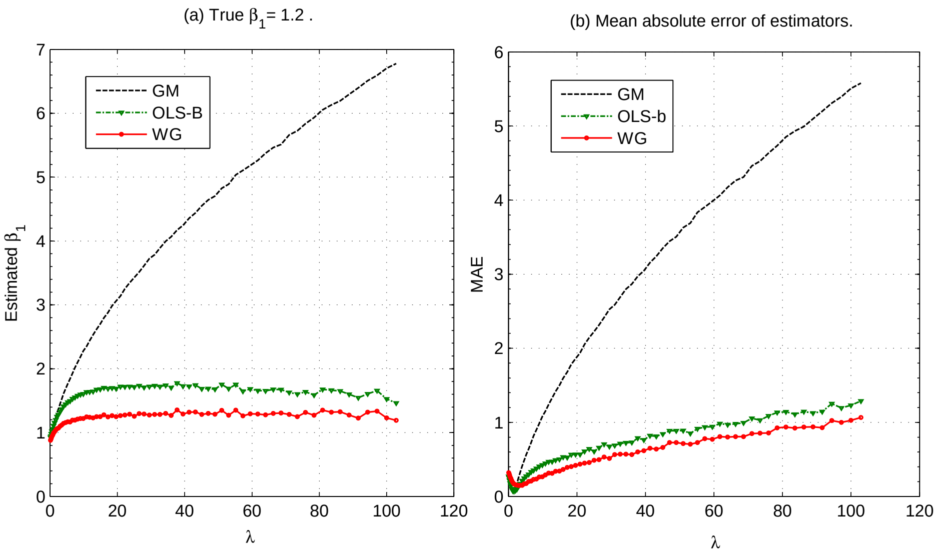

Figure 7.4: Graph of three estimators of the slope, and the mean absolute error when $\beta_0 = -15, \beta_1 = 1.2$ and $0.08 \leq \lambda \leq 100$.

than those of the GM estimator, though they are not small. Obviously, the mean absolute error of the WG estimator is the smallest compared to the other two estimators, and it seems to be stable over the range of selected ratio of error variances $0.08 \leq \lambda \leq 100$. Table 7.1 summarises the results of the simulation studies which indicate that the proposed estimator is more precise than the other competing estimators. Sarach and Celik (2011) discussed eight different regression techniques, and pointed out that the OLS-bisector estimator is near to the real value than all other estimators, and the mean squares error of the OLS-bisector is smaller than all other estimators. The current study reveals that the proposed WG estimator is consistently better than the OLS-bisector estimator in term of the closeness of $\hat{\beta}_{1WG}$ to β_1, and the size of the mean absolute error.

Table 7.1: Simulated mean values of the estimated slope and the mean absolute error for various selected values of the true intercept and slope when $0.08 \leq \lambda \leq 100$.

True slope	GM	OLS-B	WG	True model
0.55	3.4981	0.9340	0.5904	$\eta_j = 20 + 0.55\xi_j$
(MAE)	(2.9527)	(0.9989)	(0.5341)	
-0.75	-3.5299	-1.1455	-0.7857	$\eta_j = 27 - 0.75\xi_j$
(MAE)	(2.7910)	(0.8780)	(0.5328)	
1.2	3.6321	1.5622	1.2213	$\eta_j = -15 + 1.2\xi_j$
(MAE)	(2.4676)	(0.6548)	(0.5302)	

7.7 Concluding remarks

This chapter proposes a new estimator based on the mathematical relationship between the vertical and orthogonal distances of the observed points and the regression line. This estimator is appropriate to fitting a straight line when both variables are subject to measurement errors, especially when there is no basis for distinguishing between response and explanatory variables. In addition, this chapter presents an alternative view on the geometric mean estimator.

Extensive simulation studies are conducted to compare the three alternative estimators. The comparison is done in terms of the estimated value of the slope parameter under a wide range of values of the ratio of error variances λ. All the graphs in Figures 7.2-7.4 provide clear evidence that the WG estimator is more precise than the other competing estimators. The values of the proposed slope estimator are nearer to real value than the OLS-bisector, and the mean absolute error of the WG estimator is smaller than that of the OLS-bisector and GM estimators. Therefore, the proposed estimator possesses better statistical proprieties than the GM estimator, and the OLS-bisector estimator. The new method is stable and works well for different sample sizes and for different values of λ.

Chapter 8

Conclusions

In this study, we considered estimating the slope of a simple linear regression model when both the explanatory and the response variables are measured with error. The ordinary least squares estimator of the regression parameters is inappropriate when the variables are subject to error. In general, there remains the impression that the measurement error problem is rather intractable because no generally consistent estimator exists.

We aimed to introduce a new methodology based on a mathematical transformation called reflection technique as covered in Chapter 3. It is an algebraic transform of the manifest data of both response and explanatory variables. The proposed methodology relies on the combination of the reflection and ordinary least squares techniques. We provide some theorems to help interpret vertical, orthogonal, and horizontal distances between the observed points and regression line.

8.1 Conclusions and Summary

Here we provide a brief summary of the main results of the previous chapters of this thesis. The introductory chapter sets the scene for the remainder of the thesis, including an introduction of the measurement error, identifiability problems, and the outline of the thesis. Chapter 2 provides some of the common estimation techniques to deal with the simple linear regression model when both the response and the explanatory variables are subject to measurement error. It also discusses some interconnections amongst these methods.

The proposed methodology was used in various places throughout this thesis. Chapter 3 introduces the reflection method and applies to define the OLS estimator. It also derives a set of results related to regression analysis within the OLS framework.

Chapter 4 proposes a new estimation procedure based on the idea of a modified instrumental variable (IV) which is defined from reflection of the manifest variable. It also compares the related existing methods with the proposed modified method. The analytical results and the illustrative examples demonstrate the fact that the proposed method significantly reduces the mean sum of squares error compared with currently used methods.

Chapter 5 provides a new slope estimator based on the idea of reflection of the manifest explanatory variable. It compares the orthogonal regression estimator with the proposed RM estimator, and those of Geary's and third-order moments methods. The analytical results and the simulation studies demonstrate that the RM estimator significantly reduces the mean absolute error. Also if the ratio of error variances $\lambda \neq 1$ and the sample size is not large, the mean absolute error of the RM estimator is lower than that of the orthogonal regression and OLS estimators.

Chapter 6 considers a new grouping method based on the rank of the reflection of the explanatory variable. It proposes specific modifications to Wald's method for fitting a straight line when both variables are subject to measurement errors. The RG method is using the ranks of a transformed variable d_j to divide the observations of the manifest explanatory variable into two groups. Extensive simulation studies were conducted to compare the five alternative estimators. The comparison is done in terms of the estimated values of the slope parameter as well as the mean absolute error of the estimators.

Chapter 7 proposes a new estimator based on the mathematical relationship between the vertical and orthogonal distances of the observed points and the regression line. This estimator is appropriate to fitting a straight line when both variables are subject to measurement errors, especially when there is no basis for distinguishing between response and explanatory variables. In addition, this chapter presents an alternative view on the geometric mean estimator. Extensive simulation studies are conducted to compare the two alternative estimators. The comparison is done in terms of the estimated value of the slope parameter under a wide range of values of the ratio of error variances. The new method is stable and works well for different sample sizes and for different values of the ratio of error variances.

There is no universally agreed method to solve the ME problem. Any method to diagnose or detect ME in the data will be an welcome addition to statistics literature. This will help researchers to avoid making misleading inferences without knowing that the data actually is contaminated by ME. Next step will be to develop a new statistical method that would allow valid regression analysis for the data with measurement errors.

Bibliography

Adcock, J. (1877). Note on the method of least squares. *Analyst,* **4**, *183-184.*

Adcock, J. (1878). A problem in least squares. *Analyst* **5**, *53-54.*

Aigner, J. (1973). Regression with a binary variable subject to errors of observation. *Journal of Econometrics* **1**, *49-60.*

Akritas, G. and Bershady, A. (1996). Linear Regression for Astronomical Data with Measurement Errors and Intrinsic Scatter. *Astrophysical Journal,* **470**, *706-714.*

Amemiya, Y. and Fuller, A. (1984). Estimation of the multivariate errors-in-variables model with estimated error covariance matrix. *Ann. Statist* **12**, *497-509.*

Ammann, P. and Van Ness, W. (1988). A routine for converting regression algorithms into corresponding orthogonal regression algorithms. *ACM Trans. Math. Software,* **14**, *76-87.*

Anderson, W. (1984). Estimating linear statistical relationships. *Ann. Statist.* **12**, *1-45.*

Anderson, W. and Rubin (1956). Statistical Inference in Factor Analysis. In Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability. *Berkeley and Los Angeles: University of California, 111-150.*

Angrist, J. and Krueger, A. (2001). Instrumental variables and the search for identification: From supply and demand to natural experiments. *Journal of Economic Perspectives,* **15**, *69-85.*

Armstrong, B. (1985). Measurement error in the generalized linear model. *Commun. Statist. Part B,* **14**, *529-544.*

Barker, F., Soh, C. and Evans, J. (1988). Properties of the geometric mean functional relationship. *Biometrics,* **44***(1), 279-281.*

Barnett, D. (1970). Fitting straight lines - the linear functional relationship with replicated observations. *Journal of the Statistical Society of London,* **19**, *135-144.*

Bartlet, S. (1949). Fitting a straight line when both variables are subject to error. *Biometrics* **5**, *207-212.*

Brannick, T. (1995). Critical Comments on applying covariance structure modeling. *Journal of Organizational Behavior,* **16***(3), 201-213.*

Bekker, A. (1986). Comments on identification in the linear errors-in-variables model. *Econometrica,* **54**, *215-217.*

Berkson, J. (1950). Are there two regressions? *Journal of Am. Statist. Assoc,* **45**, *164-180.*

Bhargava, K. (1977). Maximum likelihood estimation in a multivariate errors-in-variables regression model with unknown error covariance matrix. *Commun. Statist. Part A,* **6**, *587-601.*

Bowman, O. and Shenton, R. (1988). Properties of Estimators for the Gamma Distribution. Marcel Dekker: New York.

Burr, D. (1988). On Errors-in-variables in binary regression Berkson case. *Journal of Am. Statist. Assoc,* **83**, *739-743.*

Carriquiry, A. (2001). Measurement Error Models. *International Encyclopedia of the Social and Behavioral Sciences, 9435-42.*

Carroll, J., Ruppert, D., and Stefanski, A. (2006). Measurement

Error in Nonlinear Models. Chapman and Hall: London.

Carroll, J. and Ruppert, D. (1996). The use and misuse of orthogonal regression in linear error-in-variables models. *Journal of Am. Statist. Assoc* **50**, *1-6.*

Carroll, J. Gallo, P., and Gleser, J. (1985). Comparison of least squares and errors-in-variable regression with special reference to randomized analysis of covariance. *Journal of Am. Statist. Assoc,* **80**, *929-932.*

Casella, G., and Berger, R. (1990). Statistical Inference. Wadsworth and Brooks, Pacific Grove: Canada.

Chen, Q. Yang, J. Cheng, S. and Brooks, B. (2007). Estimating a treatment effect with repeated measurements accounting for varying effectiveness duration. *Biometrika,* **94**, *387-402.*

Chang, P., Huang, T. (1997). Inferences for the linear errors-in-variables with change point model. *Journal of the American Statistical Association,* **92**, *171-178.*

Cheng, L., and Van Ness, W. (1999). Statistical regression with measurement error, Kendall's library Of statistics 6. Wiley: New York.

Cheng, H., and Iles, C. (1990). Embedded models in three-parameter distributions and their estimation. *Journal . Roy. Statist. Soc. Ser. B,* **52**, *135-149.*

Chesher, A. (1991). The effect of measurement error. *Biometrika,* **78***(3), 451-462.*

Copas, B. (1972). The likelihood surface in the linear functional relationship problem. *Journal of Roy. Statist. Soc. Ser. B,* **34**, *274-278.*

Cragg, G. (1997). Using higher moments to estimate the simple errors-in-variables model. *The RAND Journal of Economics,* **28**, *71-91.*

Cramer, H. (1946). Mathematical Methods of Statistics. Princeton Mathematical Series. Princeton University Press, Princeton.

Dagenais, G., and Dagenais, L. (1997). Higher moment estimators for linear regression models with errors in the variables. *Journal of Econometrics,* **76**, *193-221.*

Davidov, O. (2005). Estimating the slope in measurement error models - a different perspective. *Statistics and Probability Letters,* **71**, *215-223.*

Degracie, S., and Fuller, A. (1972). Estimation of the slope and covariance when the concomitant variable is measured with error. *Journal of Am. Statist. Assoc,* **67**, *930-937.*

Deming, E. (1931). The application of least squares. Philos. Mag: Ser. 7, **11**, 146-158.

Dent, M. (1935). On observation of points connected by a linear relation. *Proc. Physical Soc. London,* **47***, 92-108.*

Dorff, M., and Gurland, J. (1961). Small sample behavior of slope estimates in a linear functional relation. *Biometrics* **17***, 283-298.*

Draper, R. and Yang, Y. (1997). Generalization of the geometric mean functional relationship. *Computational Statistics and Data Analysis,* **23***, 355-372.*

Drion, F. (1951). Estimation of the parameters of a straight line and of the variances of the variables, if they are both subject to error. *Indagationes Math,* **13***, 256-260.*

Draper, R., and Smith, H. (1981). Applied Regression Analysis. (2nd ed.). John Wiley: New York.

Dunn, G. (2004). Statistical Evaluation of Measurement Errors. Second ed. Arnold: London.

Durbin, J. (1954). Errors-in-variables. *Int. Statist. Rev,* **22***, 23-32.*

Edland, S. (1996). Bias in slope estimates for the linear errors in variables model by the variance ratio method. *Biometrics,* **52***, 243-248.*

Feigelson, D., and Babu J. (1992). Linear regression in astronomy II. *Astrophys Journal,* **397***, 55-62.*

Florens, P., Moucharrt, M., and Richard, F. (1974). Bayesian inference in error-in-variables models. *Journal of Multivariate Anal,* **4***, 419-452.*

Freedman, S., Fainberg, V., Kipnis, V., Midthune, D., and Caroll, J. (2004). A new method for dealing with measurement error in explanatory variable of regression models. *Biometrics* **60***, 172-181.*

Freudenheim, L., and Marshall, R. (1988). The problem of profound mismeasurement and the power of epidemiological studies of diet and cancer. *Nutr. Cancer,* **11***, 243-250.*

Fuller, A. (2006). Measurement Error Models. Wiley: New Jersey.

Gaskell, M. (2007). A simple method for making non-linear fits to data sets with no independent error-free coordinate. Submitted to Publications of the Astronomical Society of the Pacific.

Geary, C. (1942). Inherent relations between random variables. *Proc. R. Irish Acad. Sect. A,* **47***, 36-67.*

Geary, C. (1943). Relations between statistics: the general and the sampling problem when the samples are large. *Proceedings of the Royal Irish Academy Section A,* **22***, 177-196.*

Geary, C. (1948). Determination of linear relations between systematic parts of variables with errors of observation the variances of which are unknown. *Econometrica,* **17***, 30-58.*

Geary, C. (1949). Sampling aspects of the problem from the error-in-variable approach. *Econometrica,* **17***, 26-28.*

Gibson, M., and Jowett, H. (1957). Three-group regression analysis. Part 1: Simple regression analysis. *Appplied Statistics,* **6***, 114-122.*

Gleser, J. (1992). The importance of assessing measurement reliability in multivariate regression. *Journal of the American Statistical Association,* **87***, 696-707.*

Goldberger, S. (1972). Structural Equation Methods in the Social Sciences. *Econometrica,* **40***, 979-1001.*

Grewal, I., Nazroo, J., Bajekal, M., Blane, D., and Lewis, J. (2004). Influences on quality of life : a qualitative investigation of ethnic differences among older people in England. *Journal of Ethnic and Migration Studies,* **30**, *4.*

Griliches, Z. (1974). Errors in variables and other unobservables. *Econometrica,* **42***(6), 971-998.*

Grubbs, E. (1973). Errors of measurement, precision, accuracy and the statistical comparison of measuring instruments. *Technometrics* **15***, 53-66.*

Gupta P, Amanullah (1970). A note on the moments of the Wald's estimator. *Statistica Neerlandica,* **24***, 109-123.*

Halfon, E. (1985). Regression method in ecotoxicology: a better formulation using the geometric mean functional regression. *Notes. Environ. Sci. Technol,* **19***, 747-749.*

Halperin, M. (1961). Fitting of straight lines and prediction when both variables are subject to error. *Journal of Amer. Statist. Assoc* **56***, 657-669.*

Hood, K., Nix, J., and Iles, C. (1999). Asymptotic information and variance-covariance matrices for the linear structural model. *The Statistician* **48***(4), 477-493.*

Isobe, T., Feigelson, D., Akritas, G., and Babu, J. (1990). Linear regression in astronomy I. *Astrophys Journal,* **364***, 104-113.*

Johnson, J. (1972). Econometric Methods. McGraw Hill Book Company: New York.

Johnston, S. (1971). Reduction of stratospheric ozone by nitrogen oxide catalysts from supersonic transport exhaust. *Science,* **173***, 517-522.*

Jolicouer, P. (1975). Linear regressions in fishery research: some comments. *Journal of Fish. Res. Board Can,* **32***(8), 1491-1494.*

Gillard, J. (2010). An overview of linear structural models in errors in variables regression. *Revstat Statistical Journal,* **8***(1):5780.*

Gillard, J., and Iles, T. (2009). Methods of fitting straight lines where both variables are subject to measurement error. *Current Clinical Pharmacology* **4***(3), 164-171.*

Kagan, A., and Nagaev, S. (2001). How many moments can be estimated from a large sample?. *Statistics and Probability Letters,* **55***(1), 99-105.*

Kendall, G., and Stuart, A. (1961). The advanced theory of statistics volume two. Charles Griffin and Co Ltd: London.

Kendall, G. (1951). Regression, structure and functional relationship I. *Biometrika,* **38**, *11-25.*

Kendall, G. (1952). Regression, structure and functional relationship II. *Biometrika,* **39**, *96-108.*

Kendall, G., and Stuart, A. (1973). The Advanced Theory of Statistics Volume Two, Third ed. Charles Griffin and Co Ltd: London.

Kermack, A., and Haldane S. (1950). Organic correlation and allometry. *Biometrika,* **37**, *30-41.*

Kim, M., and Saleh E. (2002). Preliminary test estimators of the parameters of simple linear model with measurement error. *Metrika,* **57**, *223-251.*

Kim, M., and Saleh E. (2002). Preliminary test prediction of population total under regression models with measurement error. *Pakistan Journal of Statistics* **18**, *335-357.*

Kim, M., and Saleh E. (2003). Improved estimation of regression parameters in measurement error models: Finite sample case. *Calcutta Statistical Association Bulletin* **51**, *215-226.*

Kim, M., and Saleh E. (2005). Improved estimation of regression parameters in measurement error models. *Journal of Multivariate*

Analysis, **95**, *273 300.*

Klepper, S., and Leamer, E. (1984). Consistent sets of estimates for regression with errors in all variables. *Econometrica,* **55**, *163-184.*

Kummel, H. (1879). Reduction of observed equations which contain more than one observed quantity. *Analyst,* **6**, *97-105.*

Lakshminarayanan, Y., and Gunst, F. (1984). Estimation of parameters in linear structural relationships: Sensitivity to the choice of the ratio of error variances. *Biometrika,* **71**, *569-573.*

Leng, L., Zhang, T., Kleinman, L., and Zhu, W. (2007). Ordinary least square regression, orthogonal regression, geometric mean regression and their applications in aerosol science. *Journal of Physics Conference Series,* **78**, *012084-012088.*

Levinton, S., and Allen J. (2005). The paradox of the weakening combatant: trade-off between closing force and gripping speed in a sexually selected combat structure. *Funct Ecol,* **19**, *159-165.*

Lindley, V. (1947). Regression lines and the linear functional relationship. *Suppl. Journal of Roy. Statist. Soc,* **9**, *218-244.*

Linnet, K. (1998). Performance of Deming regression analysis in case of misspecified analytical error ratio in method comparison studies.

Clin Chem, **44***, 1024-1031.*

Ludbrook, J. (2010). Linear regression analysis for comparing two measurers or methods of measurement: But which regression?. *Clinical and Experimental Pharmacology and Physiology,* **37***, 692-699.*

Macdonald, R., and Thompson, J. (1992). Least squares fitting when both variables contain errors: pitfalls and possibilities. *Am Journal Physiol,* **60***, 66-73.*

Madansky, A. (1959). The fitting of straight lines when both variables are subject to error. *Journal of Amer. Statist. Assoc,* **54***, 173-205.*

Maddala, S. (2001). Introduction to econometrics. Second edition. Prentice Hall International, Inc.

Martin, F. (2000). General Deming regression for estimating systematic bias and its confidence interval in method-comparison studies. *Clinical Chemistry,* **46***, 100-104.*

McCartin, J. (2010). Oblique linear least squares approximation. *Applied Mathematical Sciences,* **4***, 2891-2904.*

Mohler, L., Marks. L., and Sprugel, G. (1978). Stand structure and allometry of trees during self-thinning of pure stands. *Journal of Ecol,* **66***, 599-614.*

Nair, R., and Banerjee, S. (1942). A note on fitting of straight lines if both variables are subject to error. *Sankhya,* **6***, 331-343.*

Neyman, J., and Scott, L. (1951). On certain methods of estimating the linear structural relation. *Annals of Mathematical Statistics,* **22***, 352-361.*

Pakes, A. (1982). On the asymptotic bias of the Wald-type estimators of a straight line when both variables are subject to error. *International Economic Review,* **23***, 491-497.*

Pal, M. (1980). Consistent moment estimators of regression coefficients in the presence of errors in variables. *Econometrics,* **14***, 349-364.*

Pearson, K. (1890). Contributions to the mathematical theory of evolution II, skew variation in homogeneous material. *Philos. Trans. Roy. Soc. London Ser. A,* **186***, 343-414.*

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philos. Mag,* **2***, 559-572.*

Prentice, L. (1982). Covariant measurement errors and parameter estimation in a failure time regression model. *Biometrika,* **69***, 331-342.*

Reiersol, O. (1950). Identifiability of a linear relation between variables which are subject to error. *Econometrica,* **18***, 375-89.*

Ricker, E. (1973). Linear regressions in Fishery research. *Journal Fish. Res. Board Can,* **30***, 409-434.*

Riggs, S., Guarnieri, A., and Addelman, S. (1978). Fitting straight line when both variables are subject to error. *Life Sci,* **22***, 1305-1360.*

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

Sarach, S., and Celik, H. (2011). Performance of OLS-bisector regression in method comparison studies. *World Applied Sciences Journal,* **12***(10), 1860-1865.*

Saracli, s., Yilmaz, V., and Dogan, I. (2009). Simple linear regression techniques in measurement error models. *Journal of Science and Technology,* **10***, 335-342.*

Schafer, W. (1986). Combining information on measurement error in errors-in-variables model. *Journal of Am. Statist. Assoc,* **81***, 181-185.*

Schneeweiss, H. (1976). Consistent estimation of a regression with errors in the variables. *Metrika,* **23***, 101-115.*

Scott, L. (1950). Note on consistent estimates of the linear structural relation between two variables. *Anal. Math. Stat,* **21***(2), 284-288.*

Skrondal, A., and Rabe-Hesketh, S. (2004). Generalized latent variable modeling. Interdisciplinary Statistics. Multilevel, longitudinal, and structural equation models. Chapman and Hall/CRC, Boca Raton, FL.

Sladek, V., Berner, M., and Sailer, R. (2006). Mobility in central European late Neolithic and early bronze age: Femoral cross-sectional geometry. *Am Journal Phys Anthropol,* **130***, 320-332.*

Smith, J. (2009). Use and misuse of the reduced major axis for line-fitting. *American Journal Of Physical Anthropology,* **140***, 476-486.*

Sokal, R., and Rohlf, J. (1995). Biometry. 3rd Edn. W. H. Freeman. New York.

Solari, E. (1969). The maximum likelihood solution to the problem of estimating a linear functional relationship. *Journal of Roy. Statist. Soc. Ser. B,* **31***, 372-375.*

Sprent, P. (1970). The saddle-point of the likelihood surface for a linear functional relationship. *J. Roy. Statist. Soc. Ser. B,* **32***, 432-434.*

Sprent, P., and Dolby, R. (1980). Query: the geometric mean functional relationship. *Biometrics,* **36***(3), 547-550.*

Stefanski, A., and Carroll, J. (1985). Covariant measurement error in logistic regression. *Ann. Statist,* **12***, 1335-1351.*

Stuart, A, and Ord, K. (1994). Kendalls advanced theory of statistics, volume 1: Distribution theory, sixth edition. Edward Arnold: London.

Teissier, G. (1948). La relation d'allometrie sa signification statistique et biologique. *Biometrics,* **4***, 14-53.*

Theil, H. (1950). A rank-invariant method of linear and polynomial regression analysis. *Nederlandse Akademie Wetenchappen Series A,* **53***, 386-392.*

Theil, J., and Van Yzeren (1956). The efficiency of Wald's method of fitting straight lines. *Revue de Institut Internationale de Statistique,* **24***, 17-26.*

Vaisman, I. (1997). Analytical Geometry. World Scientific: Singapore.

Van Montfort, K. (1989). Estimating in Structural Models with Non-Normal Distributed Variables: Some Alternative Approaches. DSWO Press: Leiden.

Van Montfort, K., Mooijaart, A., and de Leeuw, J. (1987). Regression with errors in variables: Estimators based on third order moments. *Statist Neerlandica,* **41**, *223-237.*

Vincent E., and Lailvaux P. (2006). Female morphology, web design, and the potential for multiple mating in Nephila clavipes: Do fat-bottomed girls make the spider world go round?. *Biological Journal of the Linnean Society,* **87**, *95-102.*

Wald, A. (1940). Fitting of straight lines if both variables are subject to error. *Ann. Math. Statist,* **11**, *284-300.*

Ware, H. (1972). The fitting of straight lines when both variables are subject to error and the ranks of the means are known. *Journal of the American Statistical Association,* **67**, *891-897.*

Warton, I., Wright, J., Falster, S., and Westoby M. (2006). Bivariate line-fitting methods for allometry. *Biol Rev,* **81**, *259-291.*

Weisberg, S. (1985). Applied Linear Regression. (2nd ed.). John Wiley: New York.

Wong, Y. (1989). Likelihood estimation of a simple linear regression model when both variables have error. *Biometrika* **76***(1), 141-148.*

Zimmermann, F., Breitenmoser-Wursten, C., and Breitenmoser, U. (2005). Natal dispersal of Eurasian lynx (Lynx lynx) in Switzerland. *Journal of Zoology,* **267***, 381-395.*

Associated Research Papers

Refereed Journal Articles and Conference Proceedings

Conference Presentation:

(1) Anwar Saqr. Theories in linear regression model. *Presented in Statistical Society of Canada, 2006 Annual Meeting in London, 28-31 May 2006.*

(2) Anwar Saqr and Shahjahan Khan. Estimation of regression parameters when measurement error in explanatory variable. *Presented in Australian Statistical Conference, Dec 6-10, 2010.*

Conference Proceeding:

(3) Anwar Saqr and Shahjahan Khan (2011). Instrumental variable estimator of regression slope when the explanatory variable is subject to measurement error. *Islamic Countries Society of Statistical Sciences, Proc. ICCS-11, Lahore, Pakistan Dec 19-22, 2011, Vol. 21, pp. 39-53.*

(4) Anwar Saqr and Shahjahan Khan (2012). Weighted reduced major axis method for regression model. *Islamic Countries Society of Statistical Sciences, Proc. ICCS-12, Doha, Qatar, Dec 19-22, 2012, Vol. 23, pp. 61-70.*

(5) Anwar Saqr and Shahjahan Khan (2012). Slope estimator for the linear error-in-variables model. *Islamic Countries Society of Statistical Sciences, Proc. ICCS-12, Doha, Qatar, Dec 19-22, 2012, Vol. 23, pp. 71-80.*

Journal Article:

(6) Anwar Saqr and Shahjahan Khan (2012). Reflection method of estimation for measurement error models. *Journal of Applied Probability and Statistics,* **7***(2), 71-88.*