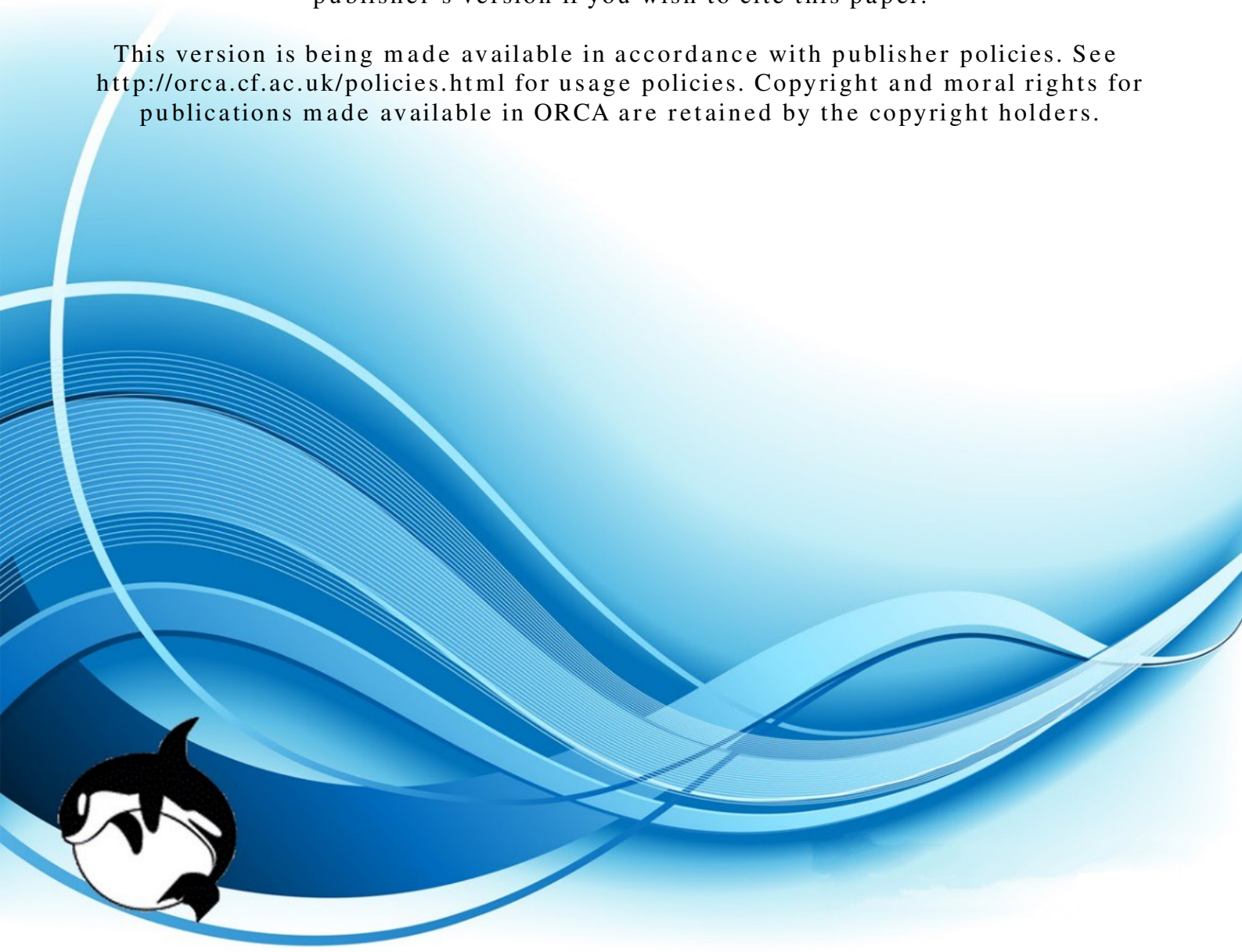


Method of Moments Estimation in Linear Regression with Errors in both Variables

Jonathan Gillard
Cardiff School of Mathematics
Cardiff University
GillardJW@Cardiff.ac.uk

May 25, 2012

Abstract

Recently, in this journal, there has been revised attention on estimating the parameters of the errors in variables, linear structural model. For example, O'Driscoll and Ramirez (2011) used a geometric approach to give insight into the performance of various slope estimators for the linear structural model as introduced by the present author. This paper aims to provide a unified method of moments approach for estimating the parameters in the linear structural model, concentrating attention on estimators using the higher moments, which to date has received only little attention in the literature.

Keywords: Linear structural model, errors in variables, regression, method of moments

1 Introduction

Fitting regression models when there is measurement error in the covariate (as well as the dependent variable) is known as errors in variables regression. Suppose two variables (ξ, η) are linearly related

$$\eta = \alpha + \beta\xi.$$

In the errors in variables model neither variable can be measured directly. They are latent variables and the measurements (x, y) that are made differ from the latent (ξ, η) by additional random components, often called measurement errors. The measurements x and y are assumed to be related to the true values ξ and η by the equations

$$\begin{aligned} x &= \xi + \delta, \\ y &= \eta + \varepsilon = \alpha + \beta\xi + \varepsilon. \end{aligned}$$

In this paper these errors, δ and ε, are assumed to be uncorrelated with each other and with the latent variable ξ. A random sample $\{(x_i, y_i), i = 1, \ldots, n\}$ of paired measurements is available from which parameters of the model are estimated.

The errors δ_i and ε_i are assumed to have zero means and variances that do not change with the suffix i. We define $E[\delta_i] = E[\varepsilon_i] = 0$, $Var[\delta_i] = \sigma_\delta^2$ and $Var[\varepsilon_i] = \sigma_\varepsilon^2$. We also assume that these errors are mutually uncorrelated and that the errors δ_i are uncorrelated with ε_i. Additionally we assume that the variables ξ_i are mutually uncorrelated with the error terms δ and ε.

The errors in variables regression problem is rarely included in statistical texts. There are two texts devoted entirely to the errors in variables regression problem, Fuller [10] and Cheng and van Ness [4]. Casella and Berger [3] has an informative section on the topic, Sprent [29] contains chapters on the problem, as do Kendall and Stuart [22] and Dunn [9]. Draper and Smith [7] on the other hand, in their book on regression analysis, devoted only 7 out of a total of almost 700 pages to errors in variables regression. Carroll et al. [2] described errors in variables models for non-linear regression, and Seber and Wild [27] included a chapter on this topic.

One method of estimation that has been used in errors in variables regression is the method of moments. Geary [11, 12, 14, 13] wrote a series of papers on the method, but using cumulants rather than moments in the later papers. Drion [8], in a paper that is infrequently cited, used the method of moments, and gave some results concerning the variances of the sample moments used in the estimators that he suggested. More recent work using the moments approach has been written by Pal [26], van Montfort et al. [31], van Montfort [30] and Cragg [5]. Much of this work centres on a search for optimal estimators using estimators based on higher moments. Dunn [9] gave formulas for many of the estimators of the slope that we describe later in this paper using a method of moments approach.

Recently, in this journal, there has been revised attention on estimating the parameters of the linear errors in variables model. For example, O'Driscoll and Ramirez [25] used a geometric approach to give insight into the performance of various slope estimators for the linear model as originally introduced by the present author. The aim of this present paper is to provide a unified method of moments framework for estimating the parameters of the linear errors in variables regression model. Some guidance as to the variances of the estimators is also provided. We describe in detail estimators of the parameters of the regression model using higher order moments, explaining their derivation and problems inherent in using them.

The structure of the paper is as follows. Section 2 describes method of moments estimators using first and second order moments. Section 3 introduces method of moments estimators using higher order moments, and describes potential problems with these estimators. A simulation study is included in Section 4, and the paper is concluded in Section 5.

2 Fitting the line by restricting the parameter space

In some applications it is assumed that the latent values ξ_i, associated with the measurements x_i, are a sample from a random variable with mean μ and variance σ^2. This is known as the structural model. In the functional model, in contrast, it is assumed that the values ξ_i $(i = 1, \ldots, n)$ are fixed, although unobservable, quantities. For further discussion as to the differences between the structural and functional models the reader is referred to Buonaccorsi [1].

When the method of moments approach is taken the distinction between the structural and functional models is not important in the estimation of the parameters of the model. The distinction needs to be made only if the values ξ_i themselves are to be estimated, and this problem will not be discussed in this paper. All that is needed in the method of moments approach are assumptions about the moments of the random variables δ and ε, and of the latent variable ξ. In this paper the following assumptions are made about these variables:

$$\begin{aligned} E[\delta] &= E[\varepsilon] = 0 \\ E[\xi] &= \mu \\ Var[\xi] &= \sigma^2 \\ Var[\delta] &= \sigma_\delta^2 \\ Var[\varepsilon] &= \sigma_\varepsilon^2 \\ Cov[\delta, \varepsilon] &= Cov[\delta, \xi] = Cov[\xi, \varepsilon] = 0. \end{aligned}$$

If all random variables in this model $(\xi, \delta, \varepsilon)$are assumed to be independent Gaussian, then this model is known as the Gaussian structural model, or Gaussian linear structural model. This terminology will be used throughout this paper. For the functional model, it is usually convenient to replace σ^2 with an alternative representation of the variability in the latent values ξ_i. Notation such as s_ξ^2 has been used by Gillard [19].

The method of moments equations based on the first and second moments have been stated by many previous authors, for example Dunn [9], or Gillard and Iles [18], but are repeated here for reference. Here a tilde is placed over a symbol to denote a method of moments estimator. In these expressions $\bar{x}$ and $\bar{y}$ are the sample means of x and y respectively, s_{xx} and s_{yy} are the sample variances and s_{xy} is the sample covariance.

$$\bar{x} = \tilde{\mu} \tag{1}$$

$$\bar{y} = \tilde{\alpha} + \tilde{\beta}\tilde{\mu} \tag{2}$$

$$s_{xx} = \tilde{\sigma}^2 + \tilde{\sigma}_\delta^2 \tag{3}$$

$$s_{yy} = \tilde{\beta}^2\tilde{\sigma}^2 + \tilde{\sigma}_\varepsilon^2 \tag{4}$$

$$s_{xy} = \tilde{\beta}\tilde{\sigma}^2 \tag{5}$$

One of the main problems in fitting an errors in variables model using the method of moments is that of identifiability of the parameters. It can be seen from equations (1), (2), (3), (4) and (5) that a unique solution cannot be found for the parameters since there are five equations, but six unknown parameters (μ, σ^2, α, β, σ_δ^2 and σ_ε^2).

One way to proceed, and the one adopted in this paper, is to assume that there is some prior knowledge, usually concerning the variances in the model, that enables the parameter space to be restricted so that unique estimators can be found. To estimate parameters such as σ_δ^2 and σ_ε^2 one would usually need repeated measurements (see for example Fuller [10] for full details). It has also been suggested that equations derived from third and fourth moments can be found, but Gillard [19, 15] found that there are limitations in the practical value of these equations, essentially because the data have to be very skewed or very kurtotic for the estimating equations to be reliable. Some estimators of the slope β using the first and second moments alone, with various restrictions on the parameter space, are tabulated below.

	Assumption	Slope estimator
Case 1	Error variance σ_δ^2 known	$\tilde{\beta}_1 = \frac{s_{xy}}{s_{xx}-\sigma_\delta^2}$
Case 2	Error variance σ_ε^2 known	$\tilde{\beta}_2 = \frac{s_{yy}-\sigma_\varepsilon^2}{s_{xy}}$
Case 3	Reliability ratio $\kappa = \frac{\sigma^2}{\sigma^2+\sigma_\delta^2}$ known	$\tilde{\beta}_3 = \frac{s_{xy}}{\kappa s_{xx}}$
Case 4	Ratio $\lambda = \frac{\sigma_\varepsilon^2}{\sigma_\delta^2}$ known	$\tilde{\beta}_4 = \frac{(s_{yy}-\lambda s_{xx})+\sqrt{(s_{yy}-\lambda s_{xx})^2+4\lambda(s_{xy})^2}}{2s_{xy}}$

Case 3 is included as there are methods available to obtain reliability measures such as κ given above. Common methods to estimate reliability include the use of intraclass correlation via an internal replication study; some form of internal validation study where the true values (those not contaminated with measurement error) are observed for a sufficient number of subjects being studied. Alternatively, reliability estimates from previously published studies may be utilised. A comprehensive review on the design and analysis of reliability studies is included in Dunn [9].

Once a slope estimator $\tilde{\beta}$ has been obtained, its value may be substituted into equations (6) to (10) in order to estimate the remaining parameters that have not been assumed known.

$$\tilde{\mu} = \bar{x} \tag{6}$$

$$\tilde{\alpha} = \bar{y} - \tilde{\beta}\bar{x} \tag{7}$$

$$\tilde{\sigma}^2 = \frac{s_{xy}}{\tilde{\beta}} \tag{8}$$

$$\tilde{\sigma}_\delta^2 = s_{xx} - \tilde{\sigma}^2 \tag{9}$$

$$\tilde{\sigma}_\varepsilon^2 = s_{yy} - \tilde{\beta}^2\tilde{\sigma}^2 \tag{10}$$

In order to ensure that the estimators for the variances are non negative, admissibility conditions

must be placed on the equations. The straightforward conditions are included below

$$\begin{aligned} s_{xx} &> \sigma_\delta^2 \\ s_{yy} &> \sigma_\varepsilon^2 \end{aligned}$$

Other admissibility conditions specific to special cases are described later in this Chapter. Admissibility conditions are discussed in detail by Kendall and Stuart [22], Hood [20], Hood et al. [21] and Dunn [9]. Practically speaking, if these admissibility conditions are broken, the choice of a linear structural model must be questioned. More precisely the estimate of the slope must lie between the slopes of the regression lines of y on x and x on y for variance estimates using equations (8), (9) and (10) to be non-negative. This point is demonstrated mathematically here.

$\tilde{\beta}$ and s_{xy} should have the same sign and variances are non-negative. We first deal with the case where $s_{xy} > 0$, hence $\tilde{\beta} > 0$. From equation (3) the condition $\tilde{\sigma}_\delta^2 \geq 0 \Rightarrow s_{xx} \geq \tilde{\sigma}^2$. From equation (5) this gives $\tilde{\beta} s_{xx} \geq \tilde{\beta}\tilde{\sigma}^2 = s_{xy}$ and so $\tilde{\beta} \geq \frac{s_{xy}}{s_{xx}}$. The right hand side is the slope of the simple linear regression of y on x. From equation (4) the condition $\tilde{\sigma_\varepsilon}^2 \geq 0 \Rightarrow s_{yy} \geq \tilde{\beta}^2\tilde{\sigma}^2 = \tilde{\beta} s_{xy}$ from equation (5). Thus $\tilde{\beta} \leq \frac{s_{yy}}{s_{xy}}$. The simple linear regression of x on y gives an estimator for the slope of the equation to predict x with y as $\frac{s_{xy}}{s_{yy}}$. However the slope is usually taken to calculate y with x and comparison should be made with the reciprocal of this estimator which is $\frac{s_{yy}}{s_{xy}}$. Hence the result that the errors in variables slope estimator is between the slopes of y on x and x on y regression is shown. If s_{xy} is negative, all inequalities are reversed. In conclusion for negative s_{xy},

$$\frac{s_{yy}}{s_{xy}} \leq \tilde{\beta} \leq \frac{s_{xy}}{s_{xx}},$$

and for positive s_{xy},

$$\frac{s_{xy}}{s_{xx}} \leq \tilde{\beta} \leq \frac{s_{yy}}{s_{xy}}.$$

All of the above estimating equations (6)-(10) can be written in terms of sample moments and the slope. Unfortunately there is no single errors in variables slope estimator that can be used in all situations. In order to use the first and second moment estimating equations alone, and to avoid the identifiability problem, the practitioner must decide which restriction of the parameter space is likely to suit the purpose best. Various restrictions and their corresponding slope estimates are discussed below. With one exception, these estimators have been described previously; most were given by Kendall and Stuart [22], Hood et al [21] and, in a method of moments context by Dunn [9].

Complete variance covariance matrices for the estimates of the parameters are provided in Gillard [15]. These can be used to estimate standard errors or confidence intervals for any of the estimates (or combinations of them). Examples of computing standard errors and confidence intervals are included in Gillard [19]. For example, the variances of each of the slope estimators for the Gaussian linear structural model are included in the following table (for brevity, the notation $\Sigma = \sigma_\delta^2\sigma_\varepsilon^2 + \beta^2\sigma^2\sigma_\delta^2 + \sigma^2\sigma_\varepsilon^2$ is used).

	Assumption	Slope estimator
Case 1	Error variance σ_δ^2 known	$Var[\tilde{\beta}_1] = \frac{\Sigma + 2\beta^2\sigma_\delta^4}{n\sigma^4}$
Case 2	Error variance σ_ε^2 known	$Var[\tilde{\beta}_2] = \frac{\beta^2\Sigma + 2\sigma_\varepsilon^4}{n\beta^2\sigma^4}$
Case 3	Reliability ratio $\kappa = \frac{\sigma^2}{\sigma^2+\sigma_\delta^2}$ known	$Var[\tilde{\beta}_3] = \frac{\Sigma}{n\sigma^4}$
Case 4	Ratio $\lambda = \frac{\sigma_\varepsilon^2}{\sigma_\delta^2}$ known	$Var[\tilde{\beta}_4] = \frac{\Sigma}{n\sigma^4}$

As derived in Gillard [15, 19], it may be shown that for the Gaussian linear structural model

$$Var[\tilde{\alpha}] = \mu^2 Var[\tilde{\beta}] + \frac{\beta^2\sigma_\delta^2 + \sigma_\varepsilon^2}{n}.$$

General variance covariance matrices, for when the random variables $(\xi, \delta, \varepsilon)$ are not Gaussian distributed are provided in Gillard [15, 19]. Some insight into the derivation of these variances is given in Appendix A.2. An alternative approach to obtain confidence intervals is via the bootstrap (for example). Full details are given in Buonaccorsi [1]. Discussion of some of the results provided in this section are also given by Davidov [6] and McAssey and Hseih [24]. An example of the application of the formulae included in this section is given by Gillard [16], who investigated the construction of time dependent reference intervals.

3 Estimates making use of higher order moments

For the purposes of the present work, we assume that higher order moments exist, and are finite. We introduce the notation:

$$E[(\xi-\mu)^3] = \mu_{\xi,3},\ E[(\xi-\mu)^4] = \mu_{\xi,4}, E[\delta^3] = \mu_{\delta,3},\ E[\delta^4] = \mu_{\delta,4}, E[\varepsilon^3] = \mu_{\varepsilon,3},\ E[\varepsilon^4] = \mu_{\varepsilon,4}.$$

3.1 Estimates using third order moments

The third order moments are written as follows. $s_{xxx} = n^{-1}\sum(x_i - \bar{x})^3$, $s_{xxy} = n^{-1}\sum(x_i-\bar{x})^2(y_i - \bar{y})$, $s_{xyy} = n^{-1}\sum(x_i-\bar{x})(y_i-\bar{y})^2$ and $s_{yyy} = n^{-1}\sum(y_i-\bar{y})^3$.

The four third moment equations take a simple form. Some details on the derivation of these expressions is given in Appendix A.1. The moment equations may be written:

$$\begin{aligned}
s_{xxx} &= \tilde{\mu}_{\xi,3} + \tilde{\mu}_{\delta,3} && (11)\\
s_{xxy} &= \tilde{\beta}\tilde{\mu}_{\xi,3} && (12)\\
s_{xyy} &= \tilde{\beta}^2\tilde{\mu}_{\xi,3} && (13)\\
s_{yyy} &= \tilde{\beta}^3\tilde{\mu}_{\xi,3} + \tilde{\mu}_{\varepsilon,3}. && (14)
\end{aligned}$$

Together with the first and second moment equations (1) to (5) inclusive, there are now nine equations in nine unknown parameters. The additional parameters introduced here are the third moments $\mu_{\xi,3}$, $\mu_{\delta,3}$ and $\mu_{\varepsilon,3}$. There are therefore unique estimators for all nine parameters. However, it is unlikely in practice that there is as much interest in these third moments as there is in the first and second moments, more especially, the slope and intercept of the line. Thus a simpler way of proceeding is probably of more general value.

The simplest way of making use of these equations is to make a single further assumption, namely that $\mu_{\xi,3}$ is non zero. There is a practical requirement associated with this assumption, and this is that the sample third moments should be significantly different from 0. It is this requirement that has probably led to the use of third moment estimators receiving relatively little attention in recent literature. It is not always the case that the observed values of x and y are sufficiently skewed to allow these equations to be used with any degree of confidence.

Moreover sample sizes needed to identify third order moments with a practically useful degree of precision are somewhat larger than is the case for first and second order moments. However, if the assumption can be justified from the data then a straightforward estimator for the slope parameter is obtained without assuming anything known a priori about the values taken by any of the parameters. This estimator is obtained by dividing equation (13) by equation (12):

$$\tilde{\beta}_5 = \frac{s_{xyy}}{s_{xxy}}.$$

$\tilde{\beta}_5$ may be substituted into equations (7), (8), (9) and (10) to estimate the intercept α and all three variances σ^2, σ_δ^2 and σ_ε^2. The third order moment $\mu_{\xi,3}$ can be estimated from equation (12). Estimators for $\mu_{\delta,3}$ and $\mu_{\varepsilon,3}$ may be obtained from equations (11) and (14) respectively.

Other simple ways of estimating the slope are obtained if the additional assumptions $\mu_{\delta,3} = \mu_{\varepsilon,3} = 0$ are made. These would be appropriate assumptions to make if the distributions of the error terms δ and ε are symmetric. Note however that this does not imply that the distribution of ξ is symmetric. The observations have to be skewed to allow the use of estimators based on the third moments. With these assumptions the slope β could also be estimated by dividing equations (12) by (11), or by dividing (14) by (13). We do not investigate these estimators further in this report; estimators that make the fewest assumptions are likely to be of the most practical value.

In order to derive formulas for the asymptotic variances and covariances of all method of moments estimators derived in this paper, the variances and covariances of sample moments are needed. Further details on variances and covariances of the estimators are included in the technical paper by Gillard and Iles [17] and Gillard [19, 15]. In these papers, full variance-covariance matrices are provided. For brevity, we now give a formula for the variance of the slope estimator $\tilde{\beta}_5$ where it is assumed that the error terms δ and ε are normally distributed. The expression $Var[\tilde{\beta}_5]$ has been calculated when δ and ε are assumed not to be normally distributed by Gillard and Iles [17].

$$Var[\tilde{\beta}_5] = \frac{1}{n(\mu_{\xi,3})^2}\left[\beta^2\mu_{\xi,4}(\sigma_\varepsilon^2 + \beta^2\sigma_\delta^2) + \frac{3\sigma_\varepsilon^2}{\beta^2}(\sigma^2 + \sigma_\delta^2) + 3\sigma_\delta^2(\beta^2\sigma^2 + \sigma_\varepsilon^2) - 6\sigma^2\sigma_\delta^2\sigma_\varepsilon^2\right]$$

Notice that the formula involves the third and fourth moments of ξ, but not higher moments.

To estimate this variance, all three parameters σ^2, σ_δ^2 and σ_ε^2 have to be estimated using equations (8), (9) and (10) respectively. The third order moment $\mu_{\xi,3}$ is estimated using (12). The fourth order moment can be estimated from one of the equations (16), (17) or (18). The combinations $(\sigma_\varepsilon^2+\beta^2\sigma_\delta^2)$, $(\sigma^2+\sigma_\delta^2)$ and $(\beta^2\sigma^2+\sigma_\varepsilon^2)$ are estimated by $(\tilde{\beta}^2 s_{xx}+s_{yy}-2\tilde{\beta}s_{xy})$, s_{xx} and s_{yy} respectively.

3.2 Estimates making use of fourth order moments

Using the obvious extension of the notation used in this paper, the fourth order moments may be written s_{xxxx}, s_{xxxy}, s_{xxyy}, s_{xyyy} and s_{yyyy}. By using an identical approach to the one adopted in deriving the third order moment estimating equations, the fourth order moment equations are derived thus:

$$
\begin{aligned}
s_{xxxx} &= \tilde{\mu}_{\xi,4} + 6\tilde{\sigma}^2\tilde{\sigma}_\delta^2 + \tilde{\mu}_{\delta,4} && (15)\\
s_{xxxy} &= \tilde{\beta}\tilde{\mu}_{\xi,4} + 3\tilde{\beta}\tilde{\sigma}^2\tilde{\sigma}_\delta^2 && (16)\\
s_{xxyy} &= \tilde{\beta}^2\tilde{\mu}_{\xi,4} + \tilde{\beta}^2\tilde{\sigma}^2\tilde{\sigma}_\delta^2 + \tilde{\sigma}^2\tilde{\sigma}_\varepsilon^2 + \tilde{\sigma}_\delta^2\tilde{\sigma}_\varepsilon^2 && (17)\\
s_{xyyy} &= \tilde{\beta}^3\tilde{\mu}_{\xi,4} + 3\tilde{\beta}\tilde{\sigma}^2\tilde{\sigma}_\varepsilon^2 && (18)\\
s_{yyyy} &= \tilde{\beta}^4\tilde{\mu}_{\xi,4} + 6\tilde{\beta}^2\tilde{\sigma}^2\tilde{\sigma}_\varepsilon^2 + \tilde{\mu}_{\varepsilon,4} && (19)
\end{aligned}
$$

Together with the first and second moment equations (1)-(5) these form a set of ten equations, but there are only nine unknown parameters. The fourth moment equations have introduced three additional parameters $\mu_{\xi,4}$, $\mu_{\delta,4}$ and $\mu_{\varepsilon,4}$, but four new equations. One of the equations is therefore not needed. The easiest practical way of estimating the parameters is to use equations (16) and (18), together with equations (3), (4) and (5).

Equation (16) is multiplied by $\tilde{\beta}^2$ and subtracted from (18) to give $\tilde{\beta}^2 s_{xxxy} - s_{xyyy} = 3\tilde{\beta}\tilde{\sigma}^2(\tilde{\beta}^2\tilde{\sigma}_\delta^2 - \tilde{\sigma}_\varepsilon^2)$. Multiplying (3) by $\tilde{\beta}^2$ and subtracting from (4) gives $\tilde{\beta}^2 s_{xx} - s_{yy} = \tilde{\beta}^2\tilde{\sigma}_\delta^2 - \tilde{\sigma}_\varepsilon^2$. Thus, making use also of equation (5), an estimating equation is obtained for the slope parameter β:

$$
\tilde{\beta}_6 = \sqrt{\frac{s_{xyyy} - 3s_{xy}s_{yy}}{s_{xxxy} - 3s_{xy}s_{xx}}} \qquad (20)
$$

There may be a practical difficulty associated with the use of equation (20) if the random variable ξ is normally distributed. In this case the fourth moment is equal to 3 times the square of the variance. A random variable for which this property does not hold is said to be kurtotic. A scale invariant measure of kurtosis is given by the following expression $\gamma_2 = \frac{\mu_4}{\sigma^4} - 3$. If the distribution of ξ has zero measure of kurtosis, the average values of the five sample moments used in equation

(20) are as follows:

$$\begin{aligned} E[s_{xyyy}] &= 3\beta^3\sigma^4 + 3\beta\sigma^2\sigma_\varepsilon^2 \\ E[s_{xxx}] &= 3\beta\sigma^4 + 3\beta\sigma^2\sigma_\delta^2 \\ E[s_{xx}] &= \sigma^2 + \sigma_\delta^2 \\ E[s_{yy}] &= \beta^2\sigma^2 + \sigma_\varepsilon^2 \\ E[s_{xy}] &= \beta\sigma^2 \end{aligned}$$

Then it can be seen that the average value of the numerator in equation (20) is approximately equal to zero, as is the average value of the denominator. Thus there is an additional assumption that has to be made for this equation to be reliable as an estimator; $\mu_{\xi,4}$ must be different from $3\sigma^4$. In practical terms, both the numerator and denominator of (20) must be significantly different from zero.

If a reliable estimate of the slope β can be obtained from (20), equations (7) to (10) allow the parameters α, σ^2, σ_δ^2 and σ_ε^2 to be estimated. $\mu_{\xi,4}$ can be estimated from (16) and $\mu_{\delta,4}$ and $\mu_{\varepsilon,4}$ may be estimated from equations (15) and (19) respectively.

Although $\tilde{\beta}_6$ has a closed compact form, its variance is rather cumbersome. Indeed, the variance of $\tilde{\beta}_6$ depends on the sixth central moments of ξ. Since it is likely to be impractical to estimate this moment with any sensible degree of accuracy, there will be no discussion of the asymptotic variance of this estimator. For further details the reader is referred to Gillard [15, 19]. Technical details, and discussions concerning the properties of $\tilde{\beta}_6$ are included in O'Driscoll and Ramirez [25].

4 Comparison study and example

4.1 Comparison study

The aim of this section is to compare the performance of all slope estimators derived in this paper for a particular representation of a structural model. Further extensive simulations are provided in Gillard [19]. Such a comparison will demonstrate the additional variability of using the estimators of the slope based on higher order moments, namely $\tilde{\beta}_5$ and $\tilde{\beta}_6$. 1000 data sets with a sample size of 150 were simulated from a linear structural model with ξ following a chi-square distribution (five degrees of freedom), and normal errors. The remaining parameter settings chosen were $\alpha = 0$, $\beta = 1$, and $\sigma_\delta = \sigma_\varepsilon = 2$. A scatterplot of a typical data set with these parameter settings is included in Figure 1.

Figure 2 contains histograms of the estimators $\tilde{\beta}_1$, $\tilde{\beta}_2$, $\tilde{\beta}_3$, $\tilde{\beta}_4$, $\tilde{\beta}_5$ and $\tilde{\beta}_6$. The scales have deliberately been chosen to be different for each estimator, to demonstrate the differing variation in each estimator. All histograms appear to peak approximately around the true value of the slope $\beta = 1$.

The histograms for $\tilde{\beta}_1$, $\tilde{\beta}_2$, $\tilde{\beta}_3$ and $\tilde{\beta}_4$ are very similar in appearance. As to be expected, $\tilde{\beta}_5$ and $\tilde{\beta}_6$

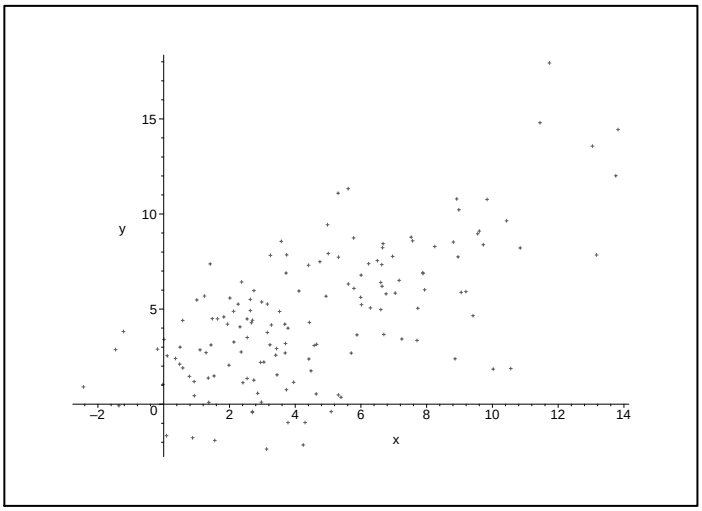


Figure 1: A typical scatterplot with ξ following a chi-square distribution (five degrees of freedom), and normal errors. $\alpha = 0$, $\beta = 1$, and $\sigma_\delta = \sigma_\varepsilon = 2$.

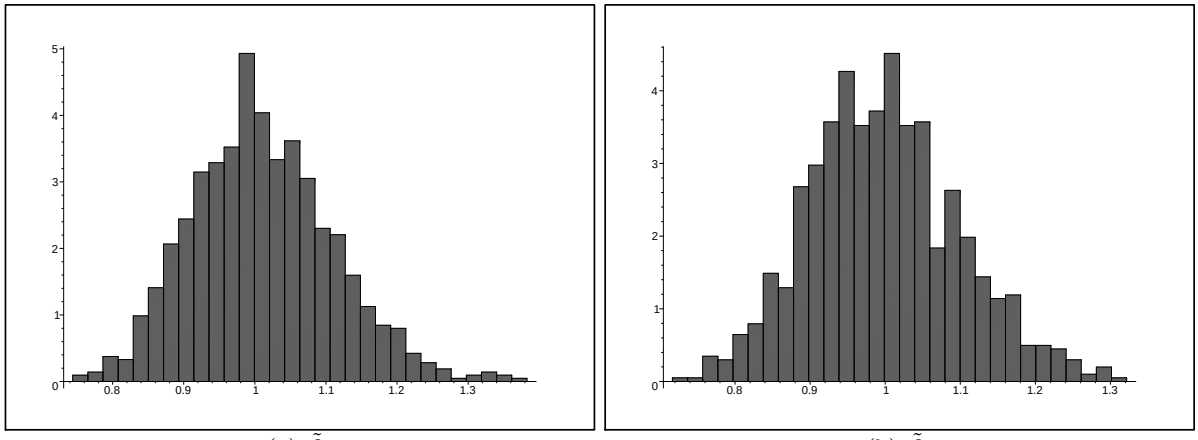
perform least favourably. For both of these slope estimators, the peak of the histogram does appear to approximately lie above the true value of the slope $\beta = 1$, but there is much more spread in both of the histograms. Roughly speaking, the histogram for $\tilde{\beta}_5$ demonstrates that for some samples, the estimate of the slope is as extreme as 2.5, whilst the histogram for $\tilde{\beta}_6$ demonstrates that the slope is estimated as 20 for some samples.

Table 1 has the sample means, sample variances and theoretical variances for the slope estimators $\tilde{\beta}_1$, $\tilde{\beta}_2$, $\tilde{\beta}_3$, $\tilde{\beta}_4$, $\tilde{\beta}_5$, and $\tilde{\beta}_6$ computed for the 1000 simulated data sets: Table 1 confirms the analysis

Slope Estimator	Sample Mean	Sample Variance	Theoretical Variance
$\tilde{\beta}_1$	1.01056	0.00995	0.00853
$\tilde{\beta}_2$	0.99791	0.01000	0.00853
$\tilde{\beta}_3$	0.99492	0.00876	0.00771
$\tilde{\beta}_4$	1.0029	0.00708	0.0064
$\tilde{\beta}_5$	1.0074	0.02922	0.037
$\tilde{\beta}_6$	1.14303	1.07506	0.06982

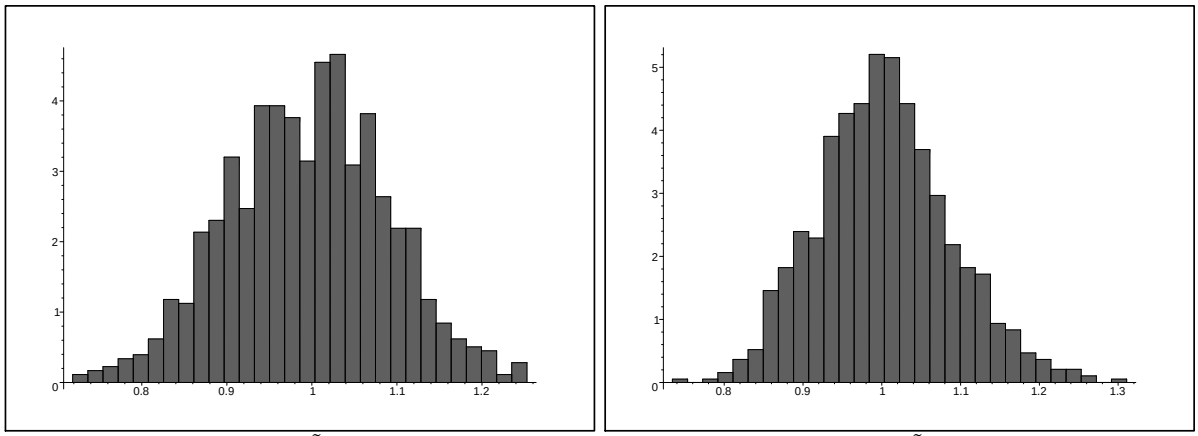
Table 1: Sample means, sample variances and theoretical variances for the slope estimators $\tilde{\beta}_1$ to $\tilde{\beta}_6$

of the histograms conducted earlier. All the estimators of the slope have a sample mean close to the true value of the slope, apart from $\tilde{\beta}_6$ which can be seen to be positively biased. The sample variance for this estimator can be seen to be around 100 times larger than the sample variance for $\tilde{\beta}_1$. For a sample size of 150 however, it is to be expected that $\tilde{\beta}_6$, which is a function of fourth order sample moments will behave more erratically than those estimators based on lower order moments. $\tilde{\beta}_5$ has performed well, with a relatively small variance, although it is still more than double the sample variances for the slope estimators based on first and second order sample moments. $\tilde{\beta}_4$ has the smallest sample variance, followed by $\tilde{\beta}_3$.



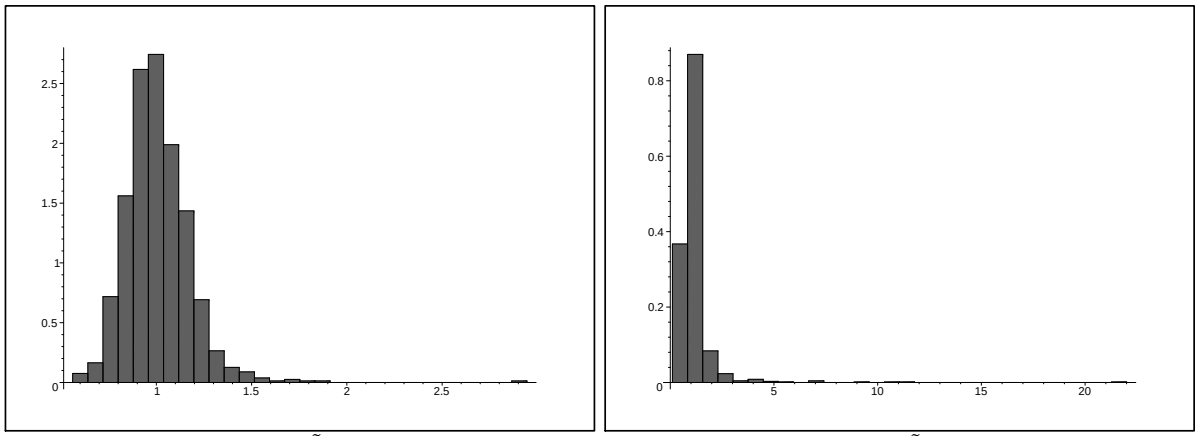
(a) $\tilde{\beta}_1$

(b) $\tilde{\beta}_2$



(c) $\tilde{\beta}_3$

(d) $\tilde{\beta}_4$



(e) $\tilde{\beta}_5$

(f) $\tilde{\beta}_6$

Figure 2: Histograms of different slope estimators for 1000 simulated data sets with a sample size of 150.

There is close agreement with the asymptotic theoretical variances and the sample variances in all cases, apart from $\tilde{\beta}_5$ and $\tilde{\beta}_6$. It seems that sample variation has caused the sample variance of $\tilde{\beta}_6$ to be inflated.

4.2 Example: Alpha Foeto Protein as a Marker for Down's Syndrome

Down's syndrome is an example of a genetic disorder, which is estimated to have an incidence of 1 per 800 births (see for example Selikowitz [28] and the references therein). The disorder however is not only seen in humans, it has been noted in chimpanzees and mice. Down's syndrome is caused by the presence (either in whole, or in part) of an extra twenty-first chromosome, and is typically associated with both physical and cognitive impairments. Examples of the physical impairments include an almond shape to the eyes, shorter limbs and pure muscle tone, whilst cognitive impairments are mainly associated with mild to moderate learning difficulties. The probability of conceiving a child with Down's syndrome increases with maternal age.

In general, pregnant women may receive a number of prenatal screens. Many of the standard screens can aid with the diagnosis of whether the unborn child is likely to have Down's syndrome. The selection of available screens may be split into examples of invasive and non-invasive screens. Examples of invasive screening include amniocentesis (a small amount of amniotic fluid is taken from the amniotic sac surrounding the fetus, and analysed) and chorionic villus sampling (a sample of placental tissue is obtained, and tested). Both of these procedures however do carry some small risk of disrupting the fetus, thus causing potential complications.

An example of a non-invasive screening method is the measurement of maternal serum alpha foeto protein (AFP) levels. It is known that AFP levels are markers for Down's syndrome, low values generally being associated with the condition. The level of AFP varies with gestational age, and with the health status of the foetus (see for example Koduah [23]).

The motivation for the use of errors in variables methodology for the use of AFP is clear. There is inherent measurement error in the measurement of gestational age and AFP level. Indeed, Selikowitz [28] has stated that one cause of false positives can be incorrect date of pregnancy. Thus the measurement of gestational age is crucial, and a model that can take into account the error inherent in the measurement of gestational age is desirable.

Figure 3 contains a typical scatterplot of the natural logarithm of AFP against gestational age in days. This particular data set was analysed in detail by Koduah [23]. The usual screening range for AFP is 15 to 18 weeks, and it is known that the standard deviation for the measurements of gestational age is approximately 2.1 days if measured in days, or is approximately 3.4 days if measured in weeks (see references in Koduah [23]). In the notation of the model used in this thesis then, this suggests that $\sigma_\delta = 2.1$. This information concerning the error variance is enough to compute an errors in variables fit to the scatter of data. The slope estimator $\tilde{\beta}_1$ assumes that the error variance σ_δ^2 is known. Table 2 shows values for the estimated slopes and intercepts via x on y regression, y on x regression, and $\tilde{\beta}_1$. It can be seen that $\tilde{\beta}_1$ does lie in between the values of the

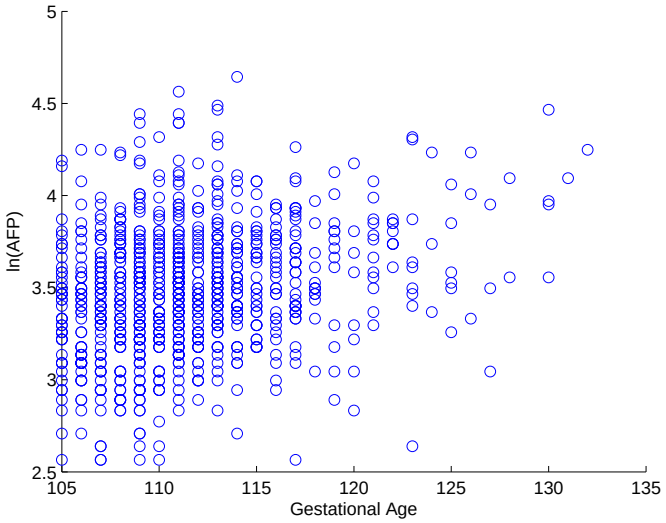


Figure 3: Measurement of the natural logarithm of AFP against gestational age (in days).

Estimator	Estimated Slope	Estimated Intercept
x on y	0.27804	-27.54081
y on x	0.01886	1.38256
$\tilde{\beta}_1$	0.02332	0.88557

Table 2: Estimators of the slope and intercept of the data for this example

slope estimated by x on y and y on x regression. This implies that the estimators of the remaining unknown variance parameters σ^2 and σ_ε^2 are positive. It can be seen that $\tilde{\beta}_1$ does this is close to 1, then it would suggest that the errors in variables estimator of the slope would align closely with the slope estimated by y on x regression.

The remaining parameters with their estimated values from using the solutions to the equations (1) to (5) are:

$$\begin{aligned} \tilde{\mu} &= 111.597 \\ \tilde{\sigma}^2 &= 18.67873 \\ \tilde{\sigma}_\varepsilon^2 &= 0.11094 \end{aligned}$$

The reliability ratio for the natural logarithm of the AFP measurement is estimated as

$$\frac{\tilde{\beta_1}^2 \tilde{\sigma^2}}{\tilde{\beta_1}^2 \tilde{\sigma^2} + \tilde{\sigma_\varepsilon^2}} = 0.08386$$

and it is thus noted that σ_ε^2 is rather large. The range of $\ln(AFP)$ values at any gestational age is approximately 1.3 but the overall range is only approximately 2.1. The slope in this example is also very shallow. It is unlikely that measurements of $\ln(AFP)$ will have such a large error

variance associated with them; there must be considerable natural variation in the $\ln(AFP)$ levels of pregnant women. The problem of this large variability in $\ln(AFP)$ in fitting an errors in variables model is avoided by using an estimator of the slope which does not assume anything concerning the error variance σ_ε^2. As knowledge of the variability in the measurement of gestational age was assumed, the inflated value for σ_ε^2 has no effect upon the estimation of β using $\tilde{\beta}_1$. Using the higher order moments as described in Section 3, we obtain $\tilde{\beta}_6 = 0.0204$. The distribution of $\ln(AFP)$ is too symmetric to allow $\tilde{\beta}_5$ to be reliable. This estimator however, makes no use of the knowledge of the error variation in gestational age.

5 Conclusion

In this paper we have described how the method of moments may be used to estimate the parameters in a linear errors in variables regression model. We have described method of moments based estimators constructed upon assuming certain parameters (or functions thereof) are known. In order to avoid making such assumptions, we have introduced estimators which appeal to the higher order moments, and have described potential problems with these estimators. A simulation study compares these estimators, and confirms that estimators based on the higher moments have larger variance than those based on smaller order moments.

It is potentially useful to offer some recommendations as to when each of the estimators provided in this paper may be used, based on the authors experience. If a practitioner has knowledge as to the value of any of the parameters concerning error variances (such as σ_δ^2 known, σ_ε^2 known, κ known or λ known), then this information has to be utilised, and the appropriate estimator selected from $\tilde{\beta}_1$ to $\tilde{\beta}_4$ must be used. If no *a priori* knowledge is available, but the sample third moments are significantly different from zero, use estimator $\tilde{\beta}_5$. For this estimator to be reliable a sample size of at least 50 is needed. Otherwise, if the coefficients of kurtosis are significantly different from zero, use estimator $\tilde{\beta}_6$. For this estimator to be reliable a sample size of at least 100 is needed.

References

[1] J. P. Buonaccorsi. *Measurement error*. Interdisciplinary Statistics. CRC Press, Boca Raton, FL, 2010. Models, methods, and applications.

[2] R. J. Carroll, D. Ruppert, and L. A. Stefanski. *Measurement error in nonlinear models*, volume 63 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London, 1995.

[3] G. Casella and R. L. Berger. *Statistical Inference*. Wadsworth & Brooks, Pacific Grove, CA, 1990.

[4] C-L. Cheng and J. W. Van Ness. *Statistical Regression with Measurement Error*. Kendall's Library of Statistics 6. Arnold, London, 1999.

[5] J. G. Cragg. Using higher moments to estimate the simple errors-in-variables model. *The RAND Journal of Economics*, 28(0):S71–S91, 1997.

[6] O. Davidov. Estimating the slope in measurement error models - a different perspective. *Statistics and Probability Letters.*, 71(3):215–223, 2005.

[7] N. R. Draper and H. Smith. *Applied Regression Analysis.* Wiley-Interscience, Canada, Third edition, 1998.

[8] E. F. Drion. 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, 1951.

[9] G. Dunn. *Statistical Evaluation of Measurement Errors.* Arnold, London, Second edition, 2004.

[10] W. A. Fuller. *Measurement error models.* Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons Inc., New York, 1987.

[11] R. C. Geary. Inherent relations between random variables. *Proc. R. Irish. Acad. Sect. A.*, 47:1541–1546, 1942.

[12] R. C. Geary. Relations between statistics: the general and the sampling problem when the samples are large. *Proc. R. Irish. Acad. Sect. A.*, 22:177–196, 1943.

[13] R. C. Geary. Determination of linear relations between systematic parts of variables with errors of observation the variances of which are unknown. *Econometrica*, 17:30–58, 1949.

[14] R. C. Geary. Sampling aspects of the problem from the error-in-variable approach. *Econometrica*, 17:26–28, 1949.

[15] J. Gillard. Asymptotic variance-covariance matrices for the linear structural model. *Stat. Methodol.*, 8(3):291–303, 2011.

[16] J. Gillard. Linear time-dependent reference intervals where there is measurement error in the time variable - a parametric approach. *Statistical Methods in Medical Research (to appear)*, 2011.

[17] J. Gillard and T. Iles. Variance covariance matrices for linear regression with errors in both variables. *Cardiff School of Mathematics Technical Report*, 2005.

[18] J. Gillard and T. Iles. Methods of fitting straight lines where both variables are subject to measurement error. *Current Clinical Pharmacology*, 4(4):164–171, 2009.

[19] J. W. Gillard. *Errors in variables regression: What is the appropriate model?* PhD thesis, Cardiff University, 2007.

[20] K. Hood. *Some Statistical Aspects of Method Comparison Studies.* Ph.d Thesis, Cardiff University, 1998.

[21] K. Hood, A. B. J. Nix, and T. C. Iles. Asymptotic information and variance-covariance matrices for the linear structural model. *The Statistician*, 48(4):477–493, 1999.

[22] M. G. Kendall and A. Stuart. *The Advanced Theory of Statistics Volume Two*. Charles Griffin and Co Ltd, London, Third edition, 1973.

[23] M. Koduah. *Time-Specific reference intervals when there are errors in both variables*. PhD thesis, Cardiff University, 2004.

[24] M. P. McAssey and F. Hsieh. Slope estimation in structural line-segment heteroscedastic measurement error models. *Stat. Med.*, 29(25):2631–2642, 2010.

[25] D. O'Driscoll and D. E. Ramirez. Geometric view of measurement errors. *Comm. Statist. Simulation Comput.*, 40(9):1373–1382, 2011.

[26] M. Pal. Consistent moment estimators of regression coefficients in the presence of errors in variables. *J. Econometrics*, 14:349–364, 1980.

[27] G. A. F. Seber and C. J. Wild. *Nonlinear Regression*. Wiley, New York, 1989.

[28] M. Selikowitz. *Down Syndrome: The Facts*. Oxford University Press, second edition, 1997.

[29] P. Sprent. *Models in Regression and Related Topics*. Methuen's Statistical Monographs. Matheun & Co Ltd, London, 1969.

[30] K. Van Montfort. *Estimating in Structural Models with Non-Normal Distributed Varibles: Some Alternative Approaches*. DSWO Press, Leiden, 1989.

[31] K. Van Montfort, A. Mooijaart, and J. de Leeuw. Regression with errors in variables: estimators based on third order moments. *Statist. Neerlandica*, 41(4):223–237, 1987.

A Appendix

A.1 Higher moment equations

The moment equations based on the third and fourth moments are slightly more difficult to derive than for the first and second order moment equations. For brevity, we introduce the notation

$$\begin{aligned} \xi_i^* &= \xi_i - \bar{\xi} \\ \delta_i^* &= \delta_i - \bar{\delta} \\ \varepsilon_i^* &= \varepsilon_i - \bar{\varepsilon}. \end{aligned}$$

One example illustrates the general approach:

$$\begin{aligned} E[ns_{xxy}] &= E\left[\sum (x_i - \bar{x})^2 (y_i - \bar{y})\right] \\ &= E\left[\sum \{(\xi_i^*) + (\delta_i^*)\}^2 \{\beta(\xi_i^*) + (\varepsilon_i^*)\}\right] \end{aligned}$$

Terms of order n^{-1} are neglected, so the expectations of all the cross products are zero. Moreover because of the assumptions that ξ, δ and ε are mutually uncorrelated, to order n^{-1} terms such as $E[(\xi_i - \bar{\xi})]$ are also zero. Hence $E[ns_{xxy}] = n\beta\mu_{\xi,3}$, where $\mu_{\xi,3} = E[(\xi - \mu)^3]$.

A.2 Computing variance-covariance matrices

Suppose estimators $\tilde{\theta}$ and $\tilde{\phi}$ of parameters θ and ϕ are calculated from two sample moments, u and v. The formulas in this section can readily be generalised for cases where three or four sample moments are used in the estimator. Let $\tilde{\theta} = f(u, v)$ and $\tilde{\phi} = g(u, v)$.

Let $\overline{\frac{\partial f}{\partial u}} = \frac{\partial f}{\partial u}|_{u=E[u]}$ be the partial derivative evaluated at the expected value of the sample moment u. Then

$$Var[\tilde{\theta}] \approx \left\{\overline{\frac{\partial f}{\partial u}}\right\}^2 Var[u] + \left\{\overline{\frac{\partial f}{\partial v}}\right\}^2 Var[v] + 2\left\{\overline{\frac{\partial f}{\partial u}}\right\}\left\{\overline{\frac{\partial f}{\partial v}}\right\} Cov[u, v]$$

and

$$Cov[\tilde{\theta}, \tilde{\phi}] = \left\{\overline{\frac{\partial f}{\partial u}}\right\}\left\{\overline{\frac{\partial g}{\partial u}}\right\} Var[u] + \left\{\overline{\frac{\partial f}{\partial v}}\right\}\left\{\overline{\frac{\partial g}{\partial v}}\right\} Var[v] + \left(\left\{\overline{\frac{\partial f}{\partial u}}\right\}\left\{\overline{\frac{\partial g}{\partial v}}\right\} + \left\{\overline{\frac{\partial f}{\partial v}}\right\}\left\{\overline{\frac{\partial g}{\partial u}}\right\}\right) Cov[u, v]$$

The algebra needed to work out the variances and covariances is quite lengthy. Full details are included in Gillard [19]. Nevertheless, the resulting formulas are not difficult and estimates of the parameters needed to estimate these variances and covariances are readily obtained from this paper.