Investigation of Non-normality in a Simple Errors-in-variables Model

Lee Chun Yin 3035469140

April 2021

1 Abstract

2 Introduction

Consider the problem of regression through the origin with only one explanatory variable:

$$y = \beta x + \epsilon$$

In real life applications, usually we will first obtain pairs of observations $(\tilde{x}_i, \tilde{y}_i)$, then apply a model by performing linear regression on the data. However, we only have observations on the observed $\tilde{x}_i = x_i + u_i$ and $\tilde{y}_i = y_i + v_i$, which have some additive error compared to the true x and y that thse model assumptions are based on. We further assume that the measurement errors u_i and v_i have mean zero and constant variances, and that the measurement error is uncorrelated and independent to the true values x and y. This gives rise to the *errors-in-variables model.* This imposes a different problem from the classical linear regression model, because the classical model assumes that the observed $\tilde{x}$ is nonrandom, that we have access to the true value of explanatory variable x without any error.

Furthermore, in the classical linear regression model, we often assume that the observed dependent variable y is subject to some ϵ following $N(0, \sigma^2)$. However, the normality assumption often does not hold for real life datasets. For example, when the errors in the dataset have heavy-tails and/or have skewed shapes, then the normal assumption may not be appropiate. For instance, when dealing with datasets with heavy-tailed errors, one of the practices is to assume

t-distributed errors instead of normal-distributed errors, as the light-tailedness of the normal distribution essentially implies that we assume that large errors occur with very low probability, which may not be true in datasets of poorer quality. To make the situation more complicated, in practice non-normality in errors may happen in both the residual ϵ and the observation matter u. Thus, there is a need to investiage the impacts of non-normality on the *errors-in-variables model.*

In this project, we investigate how the non-normality of errors in both the explanatory variable x and the residual of the dependent variable y affects the estimation of the regression coefficient $\hat{beta}$ and the estimation of variance of error σ_ϵ^2. We first perform a literature review on existing results on the *errors-in-variables model.* Then we will describe in detail the methadology computer simulation technique to produce results. Finally, we will present the results and findings from the computer simulations. The code used in this project can be found in the appendix.

3 Literature review

We mainly refer to the lecture notes written by Pischke [1] for the errors-in-variables model.

Suppose we wish to estimate the relationship $y = \beta x + \epsilon$, but we only have data on $\tilde{x} = x + u$. Also, let's further asume that $\sigma_v^2 = 0$, i.e. there is only measurement error in x.

If we substitute $\tilde{x} = x + u$ into $y = \beta x + \epsilon$, we obtain:

$$y_i = \beta(\tilde{x}_i - u_i) + \epsilon_i = \beta\tilde{x} + (\epsilon - \beta u)$$

As the measurement error in x becomes part of the residual error term in the model, the exogenity assumption of the Gauss-Markov theorem is violated as $cov(u, \tilde{x}) \neq 0$. Thus, the ordinary least-squares (OLS) estimator of β may not be the *best linear unbiased estimator.* Unlike the case of with measurement error, the OLS and MME estimators are different. In fact, We will see that the OLS estimators of β and σ_ϵ^2 are biased. In order to obtain unbiased and consistent estimates, we would have to resort to the method of moments (MME) estimators instead.

3.1 OLS and MME for β

Suppose we use the ordinary least-squares (OLS) estimator for β:

$$\hat{\beta} = \frac{cov(\tilde{x,y})}{var(\tilde{x})} = \frac{cov(x+u, \beta x + \epsilon)}{var(x+u)}$$

Because ϵ, u and x are independent to each other, we can obtain

$$\text{plim } \hat{\beta} = \frac{\beta \sigma_x^2}{\sigma_x^2 + \sigma_u^2} = \lambda \beta$$

where

$$\lambda \equiv \frac{\sigma_x^2}{\sigma_x^2 + \sigma_u^2}$$

This λ is also called the reliability or signal-to-variance ratio.

Therefore, we can see that the OLS estimator $\hat{\beta}$ is biased towards zero because $0 < \lambda < 1$. The sign of the bias depends on the sign of the true β.

In order to obtain consistent estimates for β, we can use the MME estimator instead. Suppose we have some prior knowledge on the measurement errors and have obtained the value of σ_x^2, σ_u^2 or λ. Then we can apply the appropiate adjustment for the bias in the OLS $\hat{\beta}$ as $\sigma_{\tilde{x}}^2 = \sigma_x^2 + \sigma_u^2$, and $\sigma_{\tilde{x}}^2$ can be directly measured from the observed data.

Some MME estimators for β using the first and second moments alone are shown below:

Table 1: Various MME estimators for β

Assumption	Method of Moments Estimator
σ_x^2 known	–
σ_u^2 known	$\frac{\sigma_{\tilde{x}}^2}{\sigma_{\tilde{x}}^2 - \sigma_u^2}$
Reliability ratio λ known	–

For example, in practice, we can obtain the values of σ_u^2 via repeated measurements [2].

3.2 OLS and MME for σ_{ϵ}^{2}

The usual way in OLS to estimate σ_{ϵ}^{2} is to calculate $\hat{\sigma_{\epsilon}^{2}}$ as the sum of squares of the residuals divided by the degrees of freedom $n-d$. To find out what happens to $\hat{\sigma_{\epsilon}^{2}}$, we can first look at what happens to the estimated residual variance first [1]:

$$\begin{aligned}\hat{\epsilon} &= y - \hat{\beta}\tilde{x} \\ &= y - \hat{\beta}(x+u) \\ &= \epsilon - (y - \beta x) + y - \hat{\beta}x - \hat{\beta}u \\ &= \epsilon + (\beta - \hat{\beta})x - \hat{\beta}u\end{aligned} \tag{1}$$

Thus, we can see that the true error term is obfuscated by two additional sources of error.

We can further obtain the OLS estimated variance as we assumed earlier that ϵ, x and u are uncorrelated:

$$\text{plim } \hat{\sigma_{\epsilon}^{2}} = \sigma_{\epsilon}^{2} + (1-\lambda)^{2}\beta^{2}\sigma_{x}^{2} + \lambda^{2}\beta^{2}\sigma_{u}^{2}$$

In order to obtian the MME estimator for σ_{ϵ}^{2}, similar to the case for $\hat{\beta}_{MME}$, if we have some prior knowledge on the observation error, we can obtain consistent estimates for σ_{x}^{2}, σ_{u}, λ and β. Thus, rearranging the expression for $\hat{\sigma_{\epsilon}^{2}}$ gives:

$$\hat{\sigma_{\epsilon}^{2}}_{MME} = \hat{\sigma_{\epsilon}^{2}}_{OLS} - (1-\lambda)^{2}\hat{\beta}_{MME}^{2}\sigma_{x}^{2} - \lambda^{2}\hat{\beta}_{MME}^{2}\sigma_{u}^{2}$$

4 Methadology

4.1 Procedures

The computer simulation method was employed in order to investigate the impacts of non-normal errors on the OLS and MME estimates.

For all simulations, we use a common β_{truth}, n, x_{lo}, xhi. β_{truth} is the underlying ground truth β of the model $y = \beta x + \epsilon$. n is the number of observations used in each simulation trial. x_{lo} and x_{hi} are the lower bound and upper bound of the x being sampled respectively.

In order to produce comparable results when using different error distributions, for each experiment, we first fix a certain σ_u^2 and σ_ϵ^2. Afterwards, for the observation error u, we choose a distribution where the observation errors u_i are sampled from. The model parameters of this distribution is chosen such that it has variance equal to σ_u^2. Similarly, for the residual error ϵ, we choose a distribution where the residual errors ϵ_i are sampled from. The model parameters of this distribution is chosen such that it has variance equal to σ_ϵ^2.

After determining which distributions to use, we perform the observation data generation step. First, we draw independently n instances of explanatory variable x_i from the uniform distribution defined by the bounds x_{lo} and x_{hi}, $U(x_{lo}, x_{hi})$. Afterwards, for each of the instances x_i, we compute $y_i = \beta_{truth} \cdot x_i + \epsilon_i$, where ϵ_i is drawn independently from the distribution for residual errors determined for this simulation trial. For each x_i, we also compute the observed explanatory variable $\tilde{x_i} = x_i + u_i$, where observation error u_i is added for each x_i. u_i is drawn independently from the distribution for observation errors determined for this simulation trial.

After generating the observation data, we perform estimations based on the observation data. We perform both the OLS and MME estimations for β and σ_ϵ^2. For the MME estimations, we use the case where prior information on the observation error σ_u^2 is known. The relevant expressions for the estimators are listed below:

$$
\begin{aligned}
\hat{\beta}_{OLS} &= \frac{cov(x\tilde{,}y)}{var(\tilde{x})} = \frac{cov(x+u, \beta x + \epsilon)}{var(x+u)} \\
\hat{\beta}_{MME} &= \cdots \\
\hat{\sigma_\epsilon^2}_{OLS} &= \cdots \\
\hat{\sigma_\epsilon^2}_{MME} &= \cdots
\end{aligned}
\tag{2}
$$

After computing the estimators, we compute the bias and squared error of each of these estimators when compared to the ground truth β_{truth} and σ_ϵ^2.

For the same observation error distribuion and residual error distribution, we perform the whole observation data generation and estimation procedures for 10000 iterations. This is to obtain more reliable conclusions on the mean bias (MBE) and mean squared error (MSE) of the estimators.

Procedure 1: Computer simulation procedure

```
Initialize $\beta_{truth}$, $n$, $x_{lo}$, $x_{hi}$;
forall $\sigma_u^2 \in \{0, 1.5, 2, 2.5, 3.0, 4.5, 6.0, 8.0, 10.0\}$ do
    forall $\sigma_\epsilon^2 \in \{1.5, 2, 2.5\}$ do
        Initialize $MSE_{\hat{\beta}_{OLS}} = 0$, $MSE_{\hat{\beta}_{MME}} = 0$, $MBE_{\hat{\beta}_{OLS}} = 0$,
          $MBE_{\hat{\beta}_{MME}} = 0.$;
        Initialize $MSE_{\hat{\sigma^2_\epsilon}_{OLS}} = 0$, $MSE_{\hat{\sigma^2_\epsilon}_{MME}} = 0$, $MBE_{\hat{\sigma^2_\epsilon}_{OLS}} = 0$,
          $MBE_{\hat{\sigma^2_\epsilon}_{MME}} = 0$;
        for 10000 iterations do
            Draw independently $n$ instances of explanatory variable $x_i$
              from $U(x_{lo}, x_{hi})$;
            Compute $n$ observations of $y_i = \beta_{truth} \cdot x_i + \epsilon_i$, where $\epsilon_i$ is
              drawn independently from a distribution with variance $\sigma_\epsilon^2$;
            Compute $n$ instances of observed explanatory variable
              $\tilde{x}_i = x_i + u_i$, where observation $u_i$ is drawn independently
              from a distribution with variance $\sigma_u^2$;
            Calculate $\hat{\beta}_{OLS}$, $\hat{\beta}_{MME}$, $\hat{\sigma^2_\epsilon}_{OLS}$, $\hat{\sigma^2_\epsilon}_{MSE}$ from the
              observations $\tilde{x}$, $y$, and prior information on observation
              error $\sigma_u^2$;
            Update $MSE_{\hat{\beta}_{OLS}}$, $MSE_{\hat{\beta}_{MME}}$, $MBE_{\hat{\beta}_{OLS}}$, $MBE_{\hat{\beta}_{MME}}$,
              $MSE_{\hat{\sigma^2_\epsilon}_{OLS}}$, $MSE_{\hat{\sigma^2_\epsilon}_{MME}}$, $MBE_{\hat{\sigma^2_\epsilon}_{OLS}}$, $MBE_{\hat{\sigma^2_\epsilon}_{MME}}$;
        end
    end
end
```

The computer simulation procedures are summarized in procedure 1.

4.2 Choice of distribution

In order to investigate the impacts on non-normality on the estimators, there is a need to use non-normal distributions to sample the observation error u and residual error ϵ. In order to satisfy the assumptions of the errors-in-variables model, the distributions used to generate the error terms should have mean zero. In our experiment procedures, we also have to derive the model parameters from a fixed σ_u^2 and σ_ϵ^2. Thus, in order to ease calculation, the distribution chosen should have finite variance and the variance is in a closed form such that the derivation of model parameters from a fixed variance is easy to compute.

Taking the above factors into consideration, we decided to use the following distributions: The normal/Gaussian distribution, the Student's t distribution, and the χ^2 distribution recentered at mean 0. They are chosen because they have different shapes from the normal distribution - the Student's t distribution is known to have heavier tails than the normal distribution, and the χ^2 distribution is known to be skewed. They pose substantial differences from the normal distribution which has light tails and is symmetric at 0.

We will then perform the simulations using these distributions as the underlying sampling distributions of the observation error u and residual error ϵ, as described in the previous subsection.

4.3 Experiment parameters used

4.4 Implementation details

The codes were implemented in Python in a Jupyter notebook. We used the sampling distribution implementations provided in the `numpy.random` package. The detailed implementation and source code can be found in the appendix.

5 Results and discussions

6 Acknowledgements

7 Appendix

References

[1] Steve Pischke. *Lecture Notes on Measurement Error.* 2007.

[2] Jonathan Gillard. *Method of Moments Estimation in Linear Regression with Errors in both Variables.* 2014.