

# 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 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 these 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  $\epsilon$  following  $N(0, \sigma^2)$ . However, the normality assumption often does not hold for real life datasets. For example, when dealing with datasets with heavy-tailed errors, one of the practices is to assume  $t$ -distributed errors instead of normal-distributed errors. As due to the light-tailedness of the normal distribution, assuming normal errors essentially implies that we assume that large errors occur with very low probability, which may not be true in datasets of poorer quality and have heavy-tailed errors. Thus, there is a need to investigate 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 noise  $\epsilon$  affects the estimation of the regression coefficient  $\hat{\beta}$  and the estimation of variance of error  $\sigma_\epsilon^2$ . We first perform a literature review on existing results on the *errors-in-variables model*. Then we will describe in detail the 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, and is also available on github online at <https://github.com/jevrii/fyp>.

## 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 assume 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:

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

### 3.1 Ordinary least-squares estimators for $\beta$ and $\sigma_\epsilon^2$

Suppose we use the ordinary least-squares (OLS) estimator for  $\beta$ :

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

Because  $\epsilon$ ,  $u$  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  $\lambda$  is also called the reliability or signal-to-variance ratio.

Thus, 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  $\beta$ .

### 3.2 Method-of-moments estimators for $\beta$ and $\sigma_\epsilon^2$

In order to obtain an unbiased estimate for  $\beta$ , we can use the method-of-moments (MME) estimator instead. Suppose we have the obtained the value of  $\sigma_x^2$ ,  $\sigma_u^2$  or  $\lambda$ , then we can apply the appropriate adjustment for the bias in the OLS  $\hat{\beta}$  as  $\sigma_{\tilde{x}}^2 = \sigma_x^2 + \sigma_u^2$ . Indeed, in some cases we can obtain the value of  $\sigma_u^2$  via repeated measurements.

For example, if  $\sigma_u^2$  is known, then we can obtain the unbiased estimator  $\hat{\beta}_{MME} = \frac{\sigma_{\tilde{x}}^2}{\sigma_{\tilde{x}}^2 - \sigma_u^2}$ .

## 4 Methodology

## 5 Results and discussions

## 6 Acknowledgements

## 7 Appendix

## References

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