Good afternoon.

I am Lee Chun Yin

And my project title is Investigation of Non-normality in a Simple Errors-in-variables Model

My supervisor is Dr. Raymond W.L. Wong

First I will give an introduction of the project

In a classical linear regression setting, we often assume that the explanatory variable is nonrandom without any observation error, and that the errors are normally distributed.

However, this may not be the case in real-life applications, where observation errors may exist, and the errors may be heavy-tailed or skewed.

We use the computer simulation technique to demonstrate the impacts of non-normality in the \textit{errors-in-variables} model.

We present numerical results from simulations based on normal, Student's t and χ^2 distributions on the ordinary least squares and method of moments estimation of regression slope parameter β and residual variance σ^2_ϵ.

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

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 $\tilde{x} = x + u$ and $\tilde{y} = y + v$, which have some additive error u v compared to the true x and y that these model assumptions are based on.

We further assume that the measurement errors u and v 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 \textit{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 additive residual ϵ following a normal distribution with mean 0 and constant variance.

The normality assumption is important for statistical inference and maximum likelihood estimation of slope parameter.

However, the normality assumption often does not hold for real life datasets.

For example, when the errors in the dataset have heavy-tails or have skewed shapes, then the normal assumption may not be appropriate.

To make the situation more complicated, in practice non-normality in errors may happen in both the residual ϵ and the measurement error u.

Thus, there is a need to investigate the impacts of non-normality on the \textit{errors-in-variables model}.

In this work, we investigate how non-normality in both the measurement error of explanatory variable x and the residual of the dependent variable y affects the estimation of the regression slope coefficient β and the estimation of variance of error σ^2_ϵ.

We first perform a literature review on existing results on the \textit{errors-in-variables model}.

Then we will describe in detail the methodology, which is the computer simulation technique used to produce results.

Finally, we will present the results and findings from the computer simulation experiments.

We mainly refer to the lecture notes written by Pischke \cite{lecturenotes} 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 variance of v = 0, i.e. there is only measurement error in x.

If we substitute $\tilde{x} = x+u$ into $y = \beta x + \epsilon$, we obtain the following relation.

As the measurement error in x becomes part of the residual error term in the model, the exogeneity assumption of the Gauss-Markov theorem is violated as the estimated residual is correlated with the observed $\tilde{x}$.

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

we can obtain the limit of OLS estimator of beta

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

The OLS estimator is inconsistent.

In order to obtain consistent estimates for β, we can use the method-of-moments estimator instead.

Suppose we have some prior knowledge on the measurement errors and have obtained the value of σ_u^2.

Then we can apply the appropriate adjustment for the bias in the OLS to obtain the MME estimator.

For the estimated variance of residuals,

The usual way in OLS to estimate variance of residuals is to use the sum of squares of the residuals divided by the degrees of freedom $n-1$.

We find the estimated residual contains two additional sources of variation compared to the true error.

We find that the estimated variance of the equation error is biased upwards.

In order to obtain the MME estimator for σ^2_ϵ, similar to the case for beta, if we have some prior knowledge on the observation error, we can obtain a consistent estimate for σ^2_ϵ. As follows:

We use the computer simulation method to produce the results.

For all simulations, we use a common β_{truth}, n, x_{lo}, x_{hi}. β_{truth} is the underlying ground truth β of the model. 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.

To perform the simulation,

we first fix a certain σ^2_u and σ^2_ϵ.

for the observation error u, we choose a distribution that the observation errors u_i are sampled from that has variance σ_u^2.

Similarly, for the residual error ϵ, we choose a distribution that the residual errors ϵ_i are sampled from, that has variance σ_ϵ^2.

we pick n instances of explanatory variable x_i uniformly from the interval $[x_{lo}, x_{hi}]$.

Afterwards, we generate the observation data. For x, we add some additive error u to it. For y, we compute it using the ground truth beta multiplied by the true x, and add additive residual term to it.

OLS and MME estimators for beta and sigma2eps, compute error metrics

We perform this procedure for 10 thousand iterations for each pair of sigma^2_u and sigma^2_\epsilon.

We compute the mean bias error and mean squared error over all iterations.

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 model assumptions, the distributions should have mean zero and finite variance.

We chose the normal distribution, student’s t distribution and chi-squared distribution recentered at mean zero.

For simplicity, we also chose the same type of distribution for both the measurement errror u and residual error ϵ.

These are the experiment parameters we used.

Now I will talk about the results for the estimation of beta.

These are the estimators we use.

Under the absence of observation error (σ_u^2=0), the OLS and MME estimators of β are the same.

MBE: no clear trend (centered at zero)

MSE: Increasing with σ_ϵ^2, slightly greater with Student's t errors

Normal

OLS: mean bias is negative and decreasing with σ_u^2, consistent with literature review. Ground truth beta positive, ols beta biased towards zero; MSE is increasing with σ_u^2

MME: mean bias is positive and increasing with σ_u^2; MSE is increasing with σ_u^2

Absolute value of bias and squared error for MME are lower than that of OLS

Expected. MME is consistent estimator, OLS is not.

Now we turn to the student's t case.

Firstly, similar to the normal-distributed errors case, we observe that the bias of OLS estimator of beta is negative and decreasing in σ^2_u

and that the MBE of $\hat{\beta}_{MME}$ is positive and increasing in σ^2_u.

The absolute MBE and MSE are higher than the normal case.

When comparing the OLS and MME estimators,

we find that switching from $\hat{\beta}_{OLS}$ to $\hat{\beta}_{MME}$ sitll provides a decrease in absolute MBE or MSE, but the decrease is not that large when compared to the normal-distributed errors case.

In fact, for larger values of σ^2_u, we observe that the absolute value of the MBE and MSE of $\hat{\beta}_{MME}$ are fact greater than that of $\hat{\beta}_{OLS}$.

Now we turn to the chi2 case.

we have similar observations to the previous 2 cases.

where the MBE of $\hat{\beta}_{OLS}$ is negative and decreasing in σ^2_u,

and the MBE of $\hat{\beta}_{MME}$ is positive and increasing in σ^2_u.

and the MSE for $\hat{\beta}_{MME}$ is smaller than that for $\hat{\beta}_{OLS}$.

The absolute MBE and MSE are both higher than the normal case.

Now I will talk about the results for the estimation of sigma^2_\epsilon.

These are the estimators we use.

Under the absence of observation error ($\sigma_u^2=0$), the OLS and MME estimators of σ_ϵ^2 are the same.

MBE: no clear trend (centered at zero)

MSE: Increasing with σ_ϵ^2, greater with Student's t and χ^2 errors

---

Normal

OLS: MBE is positive and increasing with σ_u^2 , consistent with literature review, MSE is increasing with σ_u^2

MME: MBE is negative and decreasing with σ_u^2, MSE is increasing with σ_u^2

Absolute value of MBE and MSE for MME are lower than that of OLS

Expected. MME is consistent estimator, OLS is not.

Student's t case

OLS:

MBE is positive and increasing with σ_u^2, closer to zero than normal case

MSE higher than normal case

MME:

MBE is negative and decreasing with σ_u^2, (much) farther from zero than normal case

MSE higher than normal case

We find that the improvement in estimation accuracy when using MME instead of using OLS is not as great as compared to the normal case. The improvement decreases as $sigma^2_u$ increases.

Chi2 case

OLS: MBE is positive and increasing with σ_u^2, MSE is higher than normal case

MME: MBE is negative and decreasing with σ_u^2, farther from zero than normal case, MSE is higher than normal case

Absolute value of MBE and MSE for MME are lower than OLS

Give a conclusion

We have presented some numerical results on the case of non-normality under the errors-in-variables model,

and have shown how non-normality in the observation error affects the estimation of the regression parameters β and σ^2_ϵ.

We have compared the errors in different metrics when using different distributions and different estimators for $\hat{\beta}$ and $\hat{\sigma^2_\epsilon}$.

Future work can be extended upon this topic.

We have only presented the numerical results arising from problem of estimation,

and the problem of inference of statistical parameters has not been explored in this work.

Exact inference of statistical parameters under non-normality can also be explored.

The effects of non-normality on multivariate errors-in-variables models and other estimators can also be explored in the future.

I would like to thank Dr. Wong for his support and guidance for this directed studies.

I would also like to thank the HKU Department of Statistics and Actuarial Science for providing me with this opportunity to pursue in this directed studies capstone project.

Thank you very much.