MODELLING WITH DATA DEFICIENCIES

Sri Lanka Journal of
Economic Research
Volume 1(1)
December 2013:143-166
Sri Lanka Forum of
University Economists

Athula Ranasinghe

Professor of Economics and Dean, Faculty of Arts,
University of Colombo, Sri Lanka

Abstract

The availability of numerically measured data without measurement errors is a requirement for any statistical analysis. However, this requirement is not met in many cases, especially in statistical applications in social sciences. In the context of social science applications, two broader types of measurement errors can be identified. Error due to data collection inefficiencies is the first type. The second type is the presence of numerically immeasurable variables. Conceptual Variables, Multi-dimensional Variables and Hidden (latent) Variables are examples of numerically immeasurable variables.

Measurement errors of first type would be minimized if the survey design is done with extra care. Use of proxies is one of the popular methods to overcome problems associated with second type. However, use of proxies is also criticized because it can create another form of measurement error.

Statisticians and econometricians have proposed various statistical methods to overcome the problem of measurement errors. This paper classified them into two as theoretical and technical. Theoretical solutions derive equations with measurable variables to represent conceptual variables. Several technical solutions are reviewed in this paper. Instrumental variable method and orthogonal distance regressions are generally accepted solutions for most forms of measurement error. Discrete choice models with measurable indicator variables and canonical regression methods are used by applied statisticians and econometricians to overcome the problems of latent dependent variable and multi-dimensional variables.

Key Words: *Measurement error problem, Latent Variables, Multi-dimensional variables, Canonical Correlation (Regression), Instrumental Variable Method, Orthogonal Regression*

INTRODUCTION

Economists suffer from more data problems than any other scientists. Therefore, a significant amount of theoretical and technical contributions in econometrics literature are directly or indirectly associated with attempts to solve data and data related problems. Griliches (1986) has identified three inter-related causes of data problems in economic analysis: incorrect or incomplete theories, inappropriate levels of data aggregations and incorrect data (relative to what it purports to measure). Griliches (1986) has also identified one more reason in his article, but it is not listed as one of the "main causes" of data difficulties. The "fourth cause" that Griliches (1986) identified but did not list as a main cause is the gradually increasing complication of economic theories. There are many theories in economics relating to conceptual [Friedman (1957) permanent income] or hidden [McFadden (1980) utility] or multi-dimensional [Vinod (1968) joint-products] variables as dependent or independent variables. Some theories are too complicated for statistical estimation [all productions beyond Cobb-Douglas for example]. Griliches (1986) further elaborated the fact that they (economists) use data collected by others for different purposes and that the data collection itself is done by "non-professional" enumerators. Therefore, data inaccuracy is more likely in economics than any other fields.

First paragraph of Griliches (1986) states;

> *"...Econometricians have an ambivalent attitude about economic data. ..., 'data' are the world they want to explain… At the other level, they are the source of all troubles. Their imperfections make our job more difficult and often impossible… If the data were perfect, … there would be hardly room for a separate field of econometrics.… it is the "badness" of the data that provides us with our living..."*

[Griliches, Z. (1986), p. 1466]

Econometricians have spent more time and effort to developing methods to rectify the data problem.

This paper is an incomplete review of selected theoretical and empirical literature on measurement error problem and presence of immeasurable variables. It is declared as incomplete because it is practically impossible to review all major works relevant to the review. This calls for more comprehensive and more updated reviews on the same subject.

The paper is organised into four sections including the Introduction. In its second section, the paper presents the OLS model, satisfying all assumptions. This is limited to a re-production of the textbook version of OLS method. Purpose of reproduction of the OLS method is to clear the platform for the review in this paper. This section is concluded with highlights of the impacts of the two problems identified above. Next, it directly addresses various sources and forms of measurement error problem, immeasurable variables and remedies for them. Final section is devoted for concluding remarks.

THE OLS METHOD

The history of the Ordinary Least Squared (OLS) Method is reported in Merriman (1877). According to the recorded history, the OLS method is first introduced in 1806 by Lagendre. Merriman (1877) briefly reviews that in the 70 years period from 1806 to 1870 there are thirteen different proofs of the OLS method.

From its inception, the technique is strictly recommended for sciences in which controlled experiments are possible. Assumptions therein are justified in that context. However, most of those assumptions were challenged when the same method was used in the social sciences, in which controlled experiments are not possible. As per common academic consensus, the OLS method was first used in the context of social sciences Jan Tinbergen, the first Nobel laureate in Economics (1969)[1].

Suppose the relationship for estimation is

$$(1) \qquad y_i = \beta x_i + u_i$$

Where, y_i is the dependent variable, x_i is the independent variable, u_i is unexplained or the residual term of regression and β is the regression coefficient to be estimated. The OLS estimator of this equation is derived assuming certain behaviours of y_i, x_i and the u_i. For details of those assumptions, a reader can refer to any standard econometric textbook. [See, *inter-alia* Gujarati (2004)]. Among all other assumptions, for the purpose of this paper it is crucial that measurability of y_i and x_i variables without any error is assumed. For the best utilization of the technique, a continuous dependent variable is assumed. However, in the reality those assumptions are seldom satisfied.

[1] In 1969, Jan Tinbergen and Ragnar Frisch won the first Nobel Prize for Economics. The reason given for the award is "...for having developed and applied dynamic models for the analysis of economic processes: [www.nobelprize.org].

Methods of dealing with non-continuous (discrete) independent variables and such dependent variables are widely discussed, used and generally accepted in econometric analysis[2].

Applied econometricians encounter many more complicated problems about data. Some variables are only conceptual. They have no empirical counterparts. For example, *utility, lifetime income, permanent income* and *human capital* are used in many economic analyses. Obviously these are merely concepts. No measurements are possible. When independent variables cannot be numerically measured only two options are available with OLS: one, to estimate the model without the conceptual variable and the other, to use a "proxy variable" in place of conceptual variable. Use of years of schooling in the human capital production function, use of IQ test scores for ability in earnings functions are only few example. These are generally accepted methods but not without criticisms.

For example, the first method (estimate the regression excluding immeasurable variables) will lead to *exclusion of relevant variable bias.* This is quite common in the earnings function where "ability" is identified as a key determinant of earnings but is excluded due to immeasurability. [Card (1994)].

The second method (use of proxies) is generally criticized for the *measurement error* problem. Friedman (1957) is a good example where according to Friedman (1957) regression of total consumption on total income is incorrect because the true relation is between permanent income and permanent consumption. In this model both variables are not empirically available. It is proved in Friedman (1957) that measurement errors in dependent variable will have no effect on unbiasedness. However, measurement error in independent variables makes OLS coefficients biased.

In modern econometric analysis, specially, in cross section applications of it more complicated problems are identified. Two of the most interesting problems encountered specially in cross section analysis are *latent (hidden) dependent variable* and *multi-dimensional variables.*

[2] In OLS context, dummy variables are used to treat discrete independent variables. For example, categorical variables like sex, ethnicity dummies are used. The coefficients attached to dummy variables measure the effect of the category on average behaviour of continues dependent variable (change of intercept). Use of OLS method with dummy dependent variables is proved to be inefficient due to heteroscedasticity of the error term. [See, *inter-alia*, Maddala (1986)].

Latent dependent variable model is the case when dependent variable is purely conceptual and no empirical observations are available. However, every latent dependent variable is associated with an indicator variable which represents the behavioural changes of the respondents under certain conditions of underlying latent dependent variable. A typical example for the latent variable is *optimal length of schooling.* This will be further elaborated in the next section of this paper. For reference on latent variable models see *inter-alia* Maddala (1986).

There are numerous models with multi-dimensional dependent variables. Output from a joint-product production function [Vinod (1968)], multiple outcomes of education (cognitive and attitudes) and education inputs [Chizmar and Zak (1983), Gyiamah-Brempong and Gyapong (1991), Ranasinghe and Kurukulasooriya (2013)] and demand for education quality [Ranasinghe and Hartog (1998), Ranasinghe (1999)] are only few examples for multi-dimensional output variables.

MEASUREMENT ERROR PROBLEM

The Problem

Manifolds of measurement error problems are encountered in econometric analyses. The simplest among them is the problem of *reporting error*. Due to inefficiencies in survey management some data may be collected inaccurately. Reporting error is a prominent case in this context. For example, respondents may have under reported their income and assets, enumerators have not collected all sources of income from the respondents or all sources of income are not listed in the questionnaire.

Reporting error problem is straightforward and it is assumed that with careful survey designs and efficient survey management this problem can be minimized.

In modern econometric analysis more serious problems of measurement errors are identified. Presence of *conceptual independent variables* is one of the earliest versions of it. For example, Friedman (1957) defined relationship between permanent income and permanent consumption. Ando and Modigliani (1967) described relationship between lifetime income and consumption. Permanent income/consumption and lifetime income are purely conceptual variables. No empirical measures for those variables can be obtained from a field surveys.

Latent dependent variable models [Maddala (1986)] are the next of this sort. Origin of latent variable models is found in biometric studies of effectiveness of insecticides and pesticides. In modelling the effectiveness of insecticides and pesticides it is assumed that each insect (pest) has a *tolerance limit* to insecticides (pesticides). The tolerance limit is a function of various attributes of the insect (pest). Therefore, one can write $d_i = \beta x_i + u_i$ where, d_i is a latent variable measuring the tolerance limit of insect *"i"* and x_i are observable attributes of individual insects.[3] Selected examples for economic applications are optimal length of schooling [Ranasinghe (1999) Ranasinghe and Hartog (2002)], brand loyalty [McFadden (1980)] and product differentiation [Berry (1994)]. Common feature of these models is that the dependent variable is latent or *hidden*[4].

Multi-dimensional dependent variables are the last of this sort. There are numerous economic examples for that. Multi-product production function is the most obvious and earliest example for it. Mutton and Wool are examples quoted in many textbooks. One of the earliest references for joint-product production function is Mundlak (1963). [Mundlak (1963) is referred in Mishra (2008)].

All the remedies proposed in the literature for different types of measurement errors can be classified into two as *theoretical* solutions and *technical* solutions[5].

Theoretical Solutions

Many variables used in econometric analyses are conceptual. There are no direct measures for them. Friedman (1957) proposed his version of consumption theory with permanent consumption and permanent income as dependent and independent variables respectively. In this model both variables are conceptual. No direct measures are available for permanent consumption and permanent income. With further assumptions Friedman (1957) shows that the dependent variable (permanent consumption) can be replaced with total consumption; causing no harm. Friedman showed that the measurement errors in dependent variable add to the error term of the regression. As long as the measurement error is not correlated with independent variables, the regression estimates based on

---

[3] For a detailed discussion on the history of latent dependent variable models see, Amemiya (1975).

[4] All hidden (latent) variables are conceptual variables too. Difference between the two models is that in the former the conceptual variable is appeared in left-hand side whereas in the latter conceptual variables are appeared in right-hand side of the equation. Therefore, the ways that economists deal with the two are entirely different.

[5] The best solution recommended for measurement error problem, specially for its first variety (reporting error) is to go back to the field and recollect accurate data. All the solutions discussed in this paper are only secondary solutions. However, the first best solution is not applicable always.

dependent variable with measurement error is still unbiased. However, efficiency of the OLS estimate is affected due to the increased variance of the new error term (error term of the regression + measurement error in dependent variable).

Measurement errors in independent variables are serious. This is elaborated in detail in Friedman (1957) in the context of consumption function, Card (1994) in the context of earnings function and Hausman (2001) in general. The downward bias of the slope is also called attenuation bias. In general the idea is that due to the measurement errors of independent variables intercept is overestimated and slope is under-estimated[6].

Friedman's permanent income hypothesis of consumption function is

(2) $$c = ky_p + u$$

Where, c is total consumption[7] and y_p is permanent income and u is the error term (OLS residual and difference between actual consumption and permanent consumption). Marginal and Average Propensities of Consumption (MPC and APC) are equal to *k* in this equation.

Permanent income is not empirically observable. Therefore, Friedman (1957) used *adaptive expectation* mechanism to define permanent income. According to Friedman (1957) people form expectations about their income and the expectation is formed according to the adaptive expectation mechanism. [For further details of adaptive expectation, see *inter-alia,* Shaw (1984)].

Adaptive expectation hypothesis can be summarized using following equation.

(3) $$y_{pt} - y_{pt-1} = (1-\lambda)[y_t - y_{pt-1}]\text{; Where, } 0 \le \lambda \le 1 .$$

With small adjustment this can be re-written as $y_{pt} = (1-\lambda)y_t + \lambda y_{pt-1}$. This assumes that present year's expected income (permanent income) is the weighted average of the

---

[6] Friedman (1957) used this to resolve the empirical controversy over the Keynesian consumption theory. According to Friedman (1957) the true consumption goes through origin (no intercept) and it is steeper than the Keynesian consumption function. According to Friedman (1952) this is due to the measurement error of independent variable in Keynesian consumption function.

[7] In its original form, the left-hand side variable is permanent consumption (c_p), not total consumption (c). Based on the argument presented above on the measurement error of the dependent variable allows Friedman (1975) to replace c_p with c.

actual income in present year and the expected (permanent) income of the previous year. The weight factor is λ . With repeated substitutions of lagged values of permanent income y_p, permanent income can be written as a weighted average of actual income in all previous years.

$$(4) \qquad y_{pt} = (1-\lambda)[y_t + \lambda y_{t-1} + \lambda^2 y_{t-2} + \lambda^3 y_{t-3} + ... + \lambda^T y_{t-T}]$$

Equation (4) can be substituted to equation (2) to find "an estimable consumption function" with observable independent variables.

$$(5) \qquad c_t = k(1-\lambda)[y_t + \lambda y_{t-1} + \lambda^2 y_{t-2} + \lambda^3 y_{t-3} + ... + \lambda^T y_{t-T}]$$

Writing equation (5) for previous year (C_{t-1}), multiplying it by λ and taking the difference between it and equation (5);

$$(6) \qquad c_t = k(1-\lambda)y_t + \lambda c_{t-1}\,.$$

This is the equation that Friedman (1957) estimated using US data.

Seater (1982) empirically estimated permanent income hypothesis with slightly different mathematical specification. In Seater (1982) total consumption (one with durables and other without it) is regressed on both permanent and transitory income. Permanent income is calculated using adaptive expectation formulae given above. Seater (1982) also added intercept and transitory income to the model because of the possible measurement errors in this method. In both regressions Seater (1982) found that intercepts and coefficients of transitory income are statistically significant. One may interpret this as the effect of measurement error[8].

Campbell and Mankiw (1990) tested the permanent income hypothesis using instrumental variable technique.

Six years later, Ando and Modigliani (1963) brought similar argument in deriving *lifecycle hypothesis* of consumption. In their version, Ando and Modigliani (1963) defined current consumption as a fraction of the present value of lifetime income (PV_t).

---

[8]This is important. The problem still remains unsolved. This can be a general criticism against any "economic theoretical solutions". It derives an "estimable equation". However, the estimable version is not necessarily free from measurement error problem.

Modigliani's consumption function is then,

$$(7) \qquad c_t = k(PV_t)$$

Empirical estimation of this equation is difficult because PV_t is not empirically observable. In order to derive an estimable version of equation (7) Ando and Modigliani (1963) assume that the present value of lifetime income is approximated by the sum of three elements; current income (y_t), wealth accumulated up to year t (w_t) and present value of expected lifetime non-property income $\left[\sum_{t=T}^{N} \frac{y_t^e}{(1+r)}\right]$. First two components of lifetime income are measurable. Ando and Modigliani (1963) found an "estimable counterpart" for the third component using a very "tricky assumption". They assumed that the average of the present value of lifetime non-property income is proportional to present income.

$$(8) \qquad \bar{y}_t^e = \frac{1}{N-T} \sum_{t=T}^{N} \frac{y_t^e}{(1+r)} = \beta y_t$$

Replacing PV_t with this;

$$(9) \qquad PV_t = w_t + y_t + \beta(N-T)y_t = w_t + [1+\beta(N-T)]y_t$$

In equation (9) present value of lifetime wealth can be measured given the values of β, N and T. In this specification N refers to length of working life and T is the current age. Ando and Modigliani (1963) allowed the econometric model to determine the value of β. However, Ando and Modigliani (1963) expected that β is approximately *1*. Substituting equation (9) to equation (7) we find Ando and Modigliani's "estimable version" of lifecycle consumption function. Equation (10) below reports it.

$$(10) \qquad c_t = kw_t + \gamma y_t \text{; where, } \gamma = k[1+\beta(N-T)].$$

Both the examples show that the estimable versions are derived using very specific assumptions and algebraic manipulations. Therefore, the econometric estimates of these equations must be tested for measurement errors and interpreted carefully. It is further suggested that one can derive entirely different estimable equations using the same initial equations under different assumptions. For example, if *static expectation* ($y_{pt} = y_{t-1}$) is assumed, Friedman's permanent income hypothesis will result $c_t = ky_{t-1}$; current consumption is linearly determined by the income in previous year. If *rational* *expectation* is used instead, a totally different specification is obtained. All the models are equally justifiable and also can be subject to same criticisms.

Ando and Modigliani (1967) is also based on same theoretical foundation. But as shown in equations (7) to (10) has resulted entirely different estimable consumption function.

This discussion leads to several inferences. The first is that this is an accepted practice in econometrics literature. Therefore, we can use this method without strong justifications. If the consumption function is estimated, equations (7) or (10) can be used without the requirement of any justification.

Second, it is also obvious that as there can be many estimable versions of the same equation, and that the empirical estimates of the original model are not necessarily consistent. Different versions can result in mutually inconsistent results.

The third issue that concerns an econometrician is that this method does not solve the problem. Instead it changes the problem. The problem is changed from immeasurability to measurement error.

Technical Solutions

This section of the paper describes all the major technical (statistical and econometrics) solutions for all the types of measurement error problems identified above. First the general solutions for measurement error problem are presented and then latent dependent variable models are presented. Problem of multiple dimensional dependent variables is presented thereafter.

General Problem and Solutions

General problem of measurement error refers to a case where, both left-hand and right-hand side variables are measured with some error. Friedman's interpretation to Keynesian consumption theory is a classical example for this. According to Friedman (1957) Keynesian consumption function is with measurement errors in both left-hand side and right-hand side variables. Effect of measurement error problem in left-hand side and right-hand side variables are presented below. First the case with measurement error only in left-hand side is presented. Then the issue of the right-hand side variable is presented.

Equation (11) is reproduction of equation (1) with measurement error problem in dependent variable.

(11) $y_* = \beta x + u$ where, $y_* = y + \varepsilon$ and ε is the measurement error

Equation (11) can be re-written as

(12) $y = \beta x + \lambda$. Where, $\lambda = u + \varepsilon$.

Unbiased estimate for β can be obtained by applying OLS to equation (11) if the measurement error in variable *y* is not related with *x* variable. If we also assume that two error terms are linearly independent, standard error of β from the regression with measurement error is always greater than that of the model without measurement error[9].

Equation (13) reproduces equation (1) when *x* variable is measured with error.

(13) $y = \beta x_* + u$. Where, $x_* = x + \delta$.

When OLS is applied for equation (13) $\hat{\beta} = \frac{\sum x_* y}{\sum x_*^2}$ of which the expected value is $E(\hat{\beta}) = \beta E \frac{\sum x_* x}{\sum x_*^2}$. With some further manipulations $E(\hat{\beta}) = \beta \frac{\sigma_x^2}{\sigma_x^2 + \sigma_\delta^2}$. Therefore, $E(\hat{\beta}) < \beta$.

Two technical solutions are indentified to correct for this bias. The *instrumental variable method* [Hausman (2001)] and the method of *orthogonal distance regression*. [Boggs and Rogers (1990)].

Measurement Error problem and Instrumental Variables

Econometricians have recommended this technique as a remedy for two problems; measurement error problem and endogenity bias. Hausman (2001) is an original and very descriptive assessment of use of *instrumental variables* as a remedy for measurement error problem. Instrumental variable can be any variable which satisfies following conditions. *a.*) It is strongly related with the immeasurable independent variable, *b.*)

---

[9]Standard error of β without measurement error is $\frac{\sigma_{u'}^2}{\sum x^2}$. W2hereas the standard error of β of the model with measurement error is $\frac{\sigma_{\lambda'}^2}{\sum x^2} = \frac{\sigma_u^2 + \sigma_\varepsilon^2}{\sum x^2}$. Measure error has increased the standard error of regression coefficient and therefore the efficiency of OLS is affected.

Instrumental variable is not linearly related with the error term of the original model and also with the measurement error, and *c.)* The instrument does not have any direct association with the dependent variable. This makes sure that the marginal effect of the immeasurable independent variable can be isolated through instrumental variable using *chain rule, i.e.;* $\frac{\partial y}{\partial x} = \frac{\partial y}{\partial z}\frac{\partial z}{\partial x}$ .Once a proper instrument is selected two stage method is used to estimate the instrumental variable method. Choice of appropriate instrumental variable is heavily dependent on several practical conditions. Availability of suitable instrumental variable in the data set is the key concern here. Always, researchers have to work with available data. Instrumental variables are drawn from available data. Before using any variable as an instrument its statistical suitability should be tested. In econometric literature two tests are recommended for this: *a.)* test for relevancy (high correlation between instrument and immeasurable independent variable) and *b.)*exogenity (linear independency between instrument and error term of the original equation).

Examples for using instrumental variable techniques are abundant in econometric literature. Card (1994) has summarized a selected list of literature. Some of the instruments used are given below to show that the choice of instruments is purely experimental basis and repeated use of the same instrument in different context cannot be guaranteed. Angrist and Krueger (1991a) used Year, Quarter and State of Birth as instrumental variables to correct for the endogenity of schooling in their earnings function. Justification of this instrumental variable is that those born in first quarter of the year start schooling with others born in different quarters of the same year. But they reach school leaving age before the others born in the same year. Therefore, quarter of birth will have an effect on schooling length but not on earnings. Angrist and Krueger (1991b) used lottery number assigned during Viet Nam war era as instrument for schooling. The lottery number is based on month and day of birth. Angrist's and Krueger's (1991b) logic behind this choice is that those individuals with a "high risk" of enlisting in Viet Nam war tended to engage further in fulltime education because fulltime enrolees were not enlisted to the war.

Card (1994) has some more examples quoted. All these examples show that the choice of any instrument must be based on experiments and none of the instruments listed above guarantee that the same instrument will be "relevant" for all circumstances.

Orthogonal Distance Regression (ODR)

Orthogonal distance regression is one of the latest techniques developed to resolve the measurement error problem. The advantage of this method is that it addresses the measurement error in independent and dependent variables simultaneously. Boggs and Rogers (1990) is a clear exposition of the method.

This method is based on the assumption that the error term is normally distributed and the MLE is used to find the solution to the optimization problem. The ODR method proposes to find estimates of relevant regression parameters to *minimize the sum of squared measurement errors in both variables.*

Assume that the true regression is one given in equation (1) which links *y* (dependent) with *x* (independent) through regression parameter β. Both variables are measured with errors. Instead of *y* and *x*, y∗ and x∗ are observed. Let, $y_{*i} = y_i + \varepsilon_i$ and $x_{*i} = x_i + \delta_i$. Where, ε_i and δ_i are measurement errors of *y* and *x* respectively. We also assume that both errors are normally distributed around mean error zero and constant variances $\sigma^2_{\varepsilon_i}$ and $\sigma^2_{\delta_i}$ respectively. Subscripts *i* in two variances indicates that variances of the measurement errors are not constant. Then, the ODR optimization problem is to minimize $\sum_{i=1}^{n}\left(\varepsilon_i^2 + \delta_i^2\right)$ subject to the constraint that $y = \beta x$. Note that this is the exact part of the equation (1). In this model it is implicitly assumed that only measurement error contributes to random term. Replacing observable variables (with measurement errors) into the constraint we find $y_{*i} = \beta(x_i + \delta_i) - \varepsilon_i$.

We can use the constraint to eliminate ε_i^2 from the objective function of ODR.

The objective function after substitution is $\sum_{i=1}^{n}\left[\left\{\beta\left(x_i + \delta_i\right) - y_i\right\}^2 + \delta_i^2\right]$. This is further adjusted with weights where weights are the inverse of standard deviations of two error components. Following Boggs and Roger (1990) the weighted ODR objective function is written as

(14) $\sum_{i=1}^{n} w_i^2\left[\left\{\beta\left(x_i + \delta_i\right) - y_i\right\}^2 + d_i^2\delta_i^2\right]$. Where, $w_i = \frac{1}{\sigma_{\varepsilon i}}$ and $d_i = \frac{\sigma_{\varepsilon i}}{\sigma_{\delta_i}}$.

There is no numerical solution to this. Therefore appropriate linear algorithm can be used to find the solution to this function. To estimate the parameter of this function Boggs and Rogers (1990) also recommends open source software. The software recommended is ODRPACK.

Latent Dependent Variables

A clear definition for latent variables is given in Skrongal and Rabe-Hesketh (2007). Skrongal and Rabe-Hesketh (2007) has also referred to an earlier literature review on latent variables by Andersen (1982). According to Andersen (1982) the latent variable method was first introduced in 1950 by Paul Lazersfeld.

The very first paragraph of Skrongal and Rabe-Hesketh (2007) is:

> *"Latent variables are random variables whose realized values are hidden. Their properties must thus be inferred indirectly using a statistical model connecting the latent (unobserved) variables to observed variables.---, latent variables are referred to by different names in different parts of statistics, examples including 'random effects', 'common factors', latent classes', 'underlying variables' and 'frailties'....*"
>
> [Skrongal and Rabe-Hesketh (2007), p. 312]

Table 1: Classification of Latent Variable Models

		Indicator Variable	
		Continuous	Discrete
Latent Variable	Continuous	**Latent:** Standard of Living **Indicator**: GDP	**Latent:** Utility of education. **Indicator:** Whether respondent is in fulltime education or not.
	Discrete	**Latent:** level of economic freedom. 1: Mostly free, 2.) Free, 3.) Moderately free, 4.) Mostly unfree 5.) Unfree 6.) Repressed. **Indicator:** Economic freedom index	**Latent:** Preference of certain brand of product **Indicator:** Consumer's Actual Choice of the brand

This definition clearly presents that latent variable models are associated with a latent variable and an indicator variable. The indicator variable is observable and closely associated with the movements of latent variable. Depending on the nature of latent and indicator variables four types of latent variable models can be identified. Table 1 above summarises the four types with appropriate examples.

Similar classification is also found in Table 1 in Skrongal and Rabe-Hesketh (2007). In this classification the first option (top-left) is similar to the situation described in section 3.2 of this paper except that in section 3.2, conceptual variable is appeared in the right-hand side. In Table 1 it is in the left-hand side.

Top-right case where latent variable is continuous but the indicator variable in discrete is very common in discrete choice regression analysis in econometrics. McFadden (1980) random utility theory provides very convenient interpretation to this. This is demonstrated in McFadden (1980) and in many textbooks [See, *inter-alia* Maddala (1984)]. For the convenience of beginners, the same is reproduced in this paper to model the school enrolment decision. The model is also presented in Ranasinghe (1999) and Ranasinghe and Hartog (2000).

The latent variable in this model is the *optimal length of schooling* (d_i). We assume that the optimal length of schooling is a function of X_i, optimal length is linearly dependent on X and random error term ε_i is normally distributed.

The modelling starts with the assumption that the optimal length is a latent variable and the indicator variable is the discrete choice between schooling and dropping out. Following Becker (1967) it is argued that the rational individual continuous fulltime education as long as the optimal length is greater than the reported length of schooling. Individual leaves school once he/she reaches the optimal length of schooling.

This suggests that when a person from the eligible age group is chosen at random, the probability that he/she is still in school (enrolment probability) is equal to the probability that the optimal length is greater than reported length. Algebraically, $d_i = \beta X_i + \varepsilon_i$ be the optimal length equation. Let S_i be the reported length of schooling. Then,

(15) $$\Pr(E=1) = \Pr(d_i \geq S_i) = \Pr(\beta X_i + \varepsilon_i \geq S_i)$$

Rearranging terms;

(16) $$\Pr(E=1) = \Pr(\varepsilon_i \geq S_i - \beta X_i)$$

Equation (16) shows that the enrolment probability is the area under the Cumulative Distribution Function (CDF) of ε_i over the range specified in equation (16). Then, PROBIT of LOGIT estimators can be used to estimate the coefficient (β). Choice between PROBIT and LOGIT depends on researcher's assumption on the probability distribution of ε_i.

To estimate values of S_i and X_i should be known and they are available from field surveys. Once β is estimated predicted values of optimal length of schooling (d_i) can be estimated using $\hat{d}_i = \hat{\beta} X_i$.

Advantages of this method is that it allows us to estimate the regression coefficients without knowing the dependent variable and it also predicts the probabilities of being in school (enrolment rate) and probability of school leaving (dropout rate).

Bottom-left cell represents the case with discrete latent with continuous indicator variable. Example given in the relevant cell is economic freedom index. Countries are classified into 6 categories on level of economic freedom. The levels are given in the table. However, those levels are not directly observable (latent). The levels are identified using *economic freedom index.* Economic freedom index is the indicator variable. In this case, analysis is simple, we can use OLS with economic freedom index as the dependent variable.

The bottom-right cell shows the case with discrete latent and discrete indicator variables. Most of the *multi-variate* discrete choice models fall into this category. Example presented in the bottom-right cell is about *brand preference* or brand loyalty of consumers.

Similar to the previous case the preference of brand is a latent variable. Theory of consumer behaviour suggests the determinants of brand loyalty. Although the brand preference is not observable we can see which brand is selected. It is the indicator variable of this model.

The advantage of this method is that it provides a framework to estimate and interpret the coefficients of a regression of which the dependent variable is not observable.

A problem of this method would be that the same estimable version of the equation could be drawn from different latent variable models. For example, the enrolment model we derived above could also be derived using utility approach. For example define U_S be the utility of being in school and U_D the utility from dropping out from schooling as a function of X. Then, the discrete decision variable (whether the person selected is in school or not) indicates whether U_S is greater or smaller than U_D. In literature, this is not even identified as a problem because all the applications of latent dependent variable models do not intend to predict the latent variable. All the estimates are limited only to predict probabilities and to interpret regression coefficients. Therefore, prediction of latent variable is not an issue.

Multiple Dimensional Variables

More complicated models with latent dependent variables are also available in economic literature. These are the models with "multi-dimensional" dependent variables. The tradition starts with Vinod (1968) controversial article to *Econometrica* on "joint-product production functions". Vinod (1968) differentiates the joint-products from multiple products. It is also a multiple output production process. In joint products, separate inputs and production process cannot be identified for each product in the multiple-product system. Therefore, Vinod (1968) argued that estimation of separate production functions for each output is not acceptable because they are not yielded through different production processes. Inputs going to each production process cannot be identified separately. As an example, Vinod (1968) takes the mutton and wool industry. One cannot identify inputs for mutton and wool separately. Later on various applied econometricians identified many other production processes with similar characteristics. For example, the education production function is identified as a joint-product production process. Outputs of the education process are multitudinous. In its simplest case, education process improves knowledge, skills and attitudes of students. Similar to the mutton-wool example, no one can differentiate inputs or production processes of the three outputs of the education process. See, *inter-alia* Vinod (1968), (1976), Kuylen and Verhallen (1981), Rijken Van Olst (1981) for early economic applications and Chizmar and Zak (1983), (1984), Chizmar and McCarney (1984), Cohn *et.al* (1989) Gyiamah-Brempong and Gyapong (1991) for applications in education production function. Ranasinghe and Hartog (1998), Ranasinghe (1999) used the same technique to estimate production and demand functions of education simultaneously.

Van Olst (1981), Chetty (1969), Dhrymes*et.al* (1969) and Rao (1969) offer a comprehensive discussion on the Vinod (1968) re-interpretation of canonical correlation in the context of regression.

Canonical correlation is defined as the correlation between *two sets of variables*. Let us define two sets of variables, U and V. The former is a linear function of set Y which consists of k number of variables and the latter is a linear function of another set X. Set X

consists of g number of variables. We assume that both sets consist of standardized variables. We also assume that g is always greater than k.

In matrix form $U = \lambda^{/}Y$ and $V = \beta^{/}X$. Where, λ and β are *canonical weights*. The canonical correlation is the Pearson correlation between U and V. The MLE suggests assigning values for λ and β such that the correlation between U and V (the canonical correlation) is maximized. This maximization is done subject to two constraints that both U and V have unit variance[10].

Canonical correlation a facility is readily available in number of statistical software such as STATA and SPSS. Canonical correlation estimate yields three sets of coefficients; canonical weights of Y-variables ($\hat{\lambda}$), canonical weights of X-variables ($\hat{\beta}$) and canonical correlations ($\hat{r}$). According to Vinod (1968) re-interpretation of canonical correlation suggests $\hat{\lambda}$ and $\hat{r}\hat{\beta}$ are the coefficients of Canonical regression. Interpretation to canonical regression coefficients is similar to that of OLS regression coefficients.

One of the drawbacks of this method is that the canonical regression method cannot be used with discrete data[11].

Three alternative applications of canonical regression method are briefly presented below, especially for beginner' references in their applications. The three cases presented below are Vinod (1968, 1969 and 1976) and critiques of the application and interpretation of the technique in mutton-wool production function - Chizmar and Zak (1983) in the education production function and Ranasinghe and Hartog (1998) and Ranasinghe (1999) - on the use of the technique to estimate demand and production of education in a simultaneous system.

An interpretational limitation of canonical correlation reviewed in this paper is that all tend to interpret only the first canonical correlation whereas the technique itself produces more than one canonical correlation and set of canonical weights for each.

[10]For detailed discussion on the estimation of canonical correlation, see, *inter-alia,* Anderson (1984).

[11]Despite this limitation, Chizmar and Zak (1983) have used several Likert Scale measures in their canonical regression of education production function estimate.

Joint-Product Production Function of Mutton and Wool (Vinod)

Vinod (1968) re-interpreted Hotelling's canonical correlation in a regression context and used US time series data for the period 1951-62 on wool and mutton production process. Inputs of this model are the labour and capital used to produce wool and mutton. Vinod (1968) argued that mutton and wool are joint-products. Labour and capital used for mutton production and wool production cannot be identified separately.

Using the canonical regression estimates Vinod (1968) calculated "ratio elasticity" of the two products with respect to each input; for example, the mutton to wool elasticity of labour. This measures percentage change in the mutton to wool ratio to one percent change in labour. Vinod (1968) further compares canonical results with OLS.

Later, in 1969, several economists criticized Vinod's comparison as incorrect because OLS regression is estimated using original data whereas canonical regression is estimated using standardized variables. [See, Rao (1969), Chetty (1969), Dhrymes and Mitchell (1969)]. Therefore, critiques suggested that canonical regression can be compared with OLS only if the OLS is also estimated using standardized variables in which the regression coefficients are the *beta coefficients*. Vinod (1969) and Vinod (1976) further developed his method considering the criticisms.

Education Production Function (Chizmar and Zak)

Chizmar and Zak (1981) is one of the earliest applications of canonical regression to estimate the education production function. In this study, joint-product education production function is estimated using individual student data from undergraduates of Illinois State University. It is assumed that education produces two outputs, *knowledge* in economic theories (cognitive ability) and *attitude* towards economics. Production of knowledge is measured in terms of end of semester exam scores in economic theory courses. Attitudes are measured using a 14-item Likert-scale instrument. An attitude test was performed twice, one at the beginning of the course and the other at the end of the course. The former was used as an input variable and the latter as one of the output variables.

Ranasinghe and Kurukulasooriya (2013) adopted the same procedure to estimate education production process for first year undergraduates in University of Ruhuna, Sri Lanka. Outputs of this production process were knowledge and attitudes towards economics. It was a replication study of Chizmar and Zak (1981).

Simultaneous Estimate of Demand and Production of Education (Ranasinghe, Hartog and Ranasinghe)

Ranasinghe and Hartog (1998), a discussion paper, and Ranasinghe (1999), an unpublished PhD thesis, report the modelling of education demand and production simultaneously. These two papers use canonical regression method in totally different context. These are attempts to bring education production and demand for education into one specification. Most of the estimates of demand for education use proxy measures for the dependent variable in the demand function. The proxy they choose depends on data availability and purposes of research. Education system produces multi-dimensional output[12]. Ranasinghe and Hartog (1998) and Ranasinghe (1999) assume that whatever the production by schools at equilibrium it is equal to the demand[13].

Researchers cannot measure exactly what is produced by schools. However, they can identify it up to a mathematical equation and known input variables. Similarly demand for education can also be modelled. At the equilibrium, production (production function) is equal to demand (demand function). Then, the canonical regression method is used to estimate both models simultaneously.

SUMMARY

This paper examines the problem of measurement error in econometric analysis. The measurement error problem is defined broadly. Several sources of measurement error are identified.

Incorrect recording of data, presence of conceptual variables, latent dependent variables and multi-dimensional dependent variables are the sources of measurement errors examined in this paper. Applied statisticians and econometricians have identified many solutions to this problem. For the first type, the best solution is to re-collect data with no errors. However, this is not possible always and it is not practical for all types of error. Some economists have invented theoretical solutions to it. This suggests expressing the conceptual variables in terms of known and measurable variables. The instrumental

[12]Various dimensions of education output may be conceptually identified (knowledge, skills and attitudes) but cannot be measured. Ranasinghe and Hartog (1998) and Ranasinghe (1999) do not attempt to measure them.

[13]Kurukulasooriya (on going PhD research) estimates joint-product production function with knowledge and subject related attitude as multi-dimensional output on selected input measures for university education.

variable method is also used quite extensively to correct measurement errors. The idea is to identify an additional variable which is closely related to the variable with measurement error but has no relationship with the dependent variable. Instrumental variable method is used extensively. However, it is difficult to generalize the instrumental variable method, and in many cases, to find suitable instrumental variables.

Orthogonal Distance Regression (ODR) is one of the latest techniques developed for the solution of the measurement error problem. The ODR method is to find regression coefficients to minimize the sum of the squared measurement errors.

Latent dependent variable is the next source of measurement error problem reviewed in this paper. There are plenty of examples in econometric literature with hidden (latent) dependent variables. All discrete choice literature can be identified as the solution to this problem. For that the only requirement is that a discrete indicator variable for the latent variable be identified. This will estimate the regression coefficients on discrete moves of the indicator variable. Any movements within a given range of the latent variable are not captured.

The presence of multi-dimensional variables is identified as the next source of measurement error. Joint-product production function and output of education system are examples cited in literature. Use of Hotelling's Canonical correlation is identified as the solution to that.

Table 2 summarises the contentment of this paper.

Table 2: Summary of Findings

Source	Effect on OLS	Remedy
Inefficiencies in survey management Incorrect definitions	Biased if measurement error is in independent variables. Less efficient.	• Re-collect more accurate data. • Instrumental variables. • Orthogonal Distance Regression.
Conceptual Variables	OLS can be used only with proxy measures of conceptual variables. If proxies are with errors, biased results and less efficient.	• Theoretical solutions. • Instrumental variables • Orthogonal Distance Regression
Latent dependent variables	OLS cannot be used.	Discrete regressions with observable indicator variables.
Multi-dimensional dependent variables	OLS cannot be used.	Canonical Regression

REFERENCES

Amemiya, T. (1975), Qualitative Response Models, *Annals of Economic and Social Measurement*, Vol. 4, No. 3, pp. 363-372.

Andersen, E. B. (1982), Latent structure analysis: a survey. *Scand. J. Statist*. 9, 1–12.

Anderson, T.W. (1984) *An Introduction to Multivariate Statistical Analysis* (2nd edition), Wiles Series in Probability and Mathematical Statistics, John Wiley and Sons, New York.

Ando, A. and Modigliani, F. (1963), The Lifecycle Hypothesis of Saving: Aggregate information and Tests, *American Economic Review,* Vol. 53, No. 1, pp. 55-84.

Becker, G.S. (1967) Human Capital and the Personal Distribution of Income, *W.S.Woitinsky lecture, no. 1,* University of Michigan, USA.

Berry, S. T. (1994) Estimating Discrete Choice Models of Product differentiation, *RAND Journal of Economics*, Vo. 25, No. 2, pp. 242-262.

Boggs P, T. and Rogers, J. E. (1990) *Orthogonal Distance Regression,* Applied and Computational Mathematics Division, Centre for Computing and Applied Mathematics, U.S. DEPARTMENT OF COMMERCE National Institute of Standards and Technology, Gaithersburg, MD 20899, NISTIR 89–4197

Campbell, J.Y. and Mankiw, N.G. (1990) Permanent Income, Current Income and Consumption, *Journal of Business and Economic Statistics*, Vol. 8, No. 3, pp. 265-279

Card, D. (1994), Earnings, Schooling and Ability Revisited, *NBER Working Paper No. 4832*.

Chetty, V. K. (1969) Econometrics of Joint Production: A Comment, *Econometrica,* Vol. 37, issue 4, pages 731.

Chizmar, J.F and Zak, T.A (1983) Modelling Multiple Outputs in Educational Production Functions, *The American Economic Review*, 73(2), pp. 18-22.

Dhrymes, P. J. and Mitchell, B. M. (1969) Estimation of Joint Production Functions, *Econometrica,*, Vol. 37, issue 4, pages 732-36.

Friedman, M. (1957), A Theory of Consumption Function, Princeton University Press

Griliches, Z. (1986) Economic Data Issues, *Handbook of Econometrics*, (ed.) Z. Griliches and M.D.Intriligator, Vol. III, Ch. 25, pp. 1466-1511.

Gujarati, D. (2004), *Basic Econometrics*, 4th Edition, McGraw-Hill Companies

Gyiamah-Brempong, K. and Gyapong, A.O. (1991) Characteristics of Education Production Functions: An Application of Canonical Regression Analysis, *Economics of Education Review*, Vol. 10, No. I, pp. 7-17

Hausman, J. (2001), Mismeasured Variables in Econometric Analysis: Problems from the Rights and Problems from the Left, *The Journal of Economic Perspectives*, American Economic Association, Vol. 15, No. 4 (Autumn, 2001), pp. 57-67

Hotellings, H. (1936) Relations Between Two Sets of Variables, *Biometrika*, Vol. 28, No. 3/4

Ismail, N.A. (2005) Analysing education production in Malaysia using canonical correlation analysis, *International Education Journal*, 2005, 6(3), 308-315.

Maddala, G. S. (1986) *Limited-dependent and Qualitative Variables in Econometrics*, Cambridge University Press

McFadden, D. (1980) Econometric Models for Probabilistic Choice among Products, *Journal of Business*, vol. 53, No. 3, pp. S13-S29.

Merriman, Mansfield (1877) On the History of the Method of Least Squares, *the Analyst*, Vol. IV, No. 2 (March, 1877)

Mishra, S.K. (2007) Least Squares Estimation of Joint Production Functions by the Differential Evolution method of Global Optimization, *Economics Bulletin*, 3 (51), pp. 1-13. http://economicsbulletin.vanderbilt.edu/2007/volume3/EB-07C10007A.pdf

Mundlak, Y (1963) Specification and Estimation of Multiproduct Production Functions (mimeographed), paper read at the meetings of the Econometric Society and the American Farm Economic Association, Pittsburgh, USA (Dec. 1962), summarized in the *Journal of Farm Economics*, 45 pp. 433-443.

Ranasinghe, A. and Hartog, J. (2002) Free-education in Sri Lanka. Does it eliminate the family effect? *Economics of Education Review* 21 (2002) 623–633

Ranasinghe, A. (1999) *A Pearl of Great Prize: the Free Education System in Sri Lanka*, Unpublished Ph.D. thesis, University of Amsterdam, The Netherlands

Ranasinghe, A. and Kurukulasooriya, N. (2013), Use of Canonical Regression Analysis to Characterize the input-output relationships in the learning process: A case study of a state university in Sri Lanka, *Proceedings of the Annual Research Symposium* 2013, p. 44

Rao, P. (1969) A note on Econometrics of Joint Production, *Econometrica*, Vol. 37, pp. 737-738.

Seater, J.J. (1982) On the Estimation of Permanent Income: Note, *Journal of Money, Credit and Banking* Vol. 14, No. 1 pp. 76-83.

Shaw, G.K. (1984) *Rational Expectation*: *An Elementary Exposition*, St. Martin's Press, New York

Skrongal, A. and Rabe-Hesketh, S. (2007) Latent Variable Modelling: A Survey, *Scandinavian Journal of Statistics*, Vol 34: 712–745, Blackwell

Vinod, H. D. (1968), Econometrics of Joint Production, *Econometrica,* Vol. 36, No. 2, pp.

Vinod H.D. (1969) Econometrics of Joint Production: A Reply, *Econometrica*, Vol. 37, pp 739-740.

Vinod H.D. (1976) Canonical Ridge and Econometrics of Joint Production, *Journal of Econometrics,* Vol. 4, pp. 157-166.

Wansbeek, T. and Meijer, E. (2000), *Measurement Error and Latent Variables in Econometrics*, Volume 37 (Advanced Textbooks in Economics), Elselvier Science B.V.

Weeks, M. (2000) *Models of Unobservable Latent Variables: Advanced Econometrics (Subject: 310)*, Faculty of Economics and Politics, University of Cambridge, www.econometricsociety.org/meetings/wc00/pdf/1689a.pdf