Université de Montréal
Faculté des arts et des sciences
Département de sciences économiques

CAHIER 9404

GMM ESTIMATORS FOR LINEAR REGRESSION MODELS WITH ERRORS IN THE VARIABLES

Marcel G. DAGENAIS[1] and Denyse L. DAGENAIS[2]

[1] Département de sciences économiques, Université de Montréal and Centre de recherche et développement en économique (C.R.D.E.).

[2] Institut d'économie appliquée, École des Hautes Études Commerciales.

April 1994

We thank Eric Ghysels, Christian Gouriéroux, Jerry Hausman, Linda Khalaf and James MacKinnon for several useful comments. We also benefited from the research assistance of Christine Lamarre and, in the early phase of the study, from that of Marie-Joséphine Nsengiyumva and Jean Lavoie. Different phases of this research project were supported alternatively by the Natural Sciences and Engineering Research Council of Canada, the Social Sciences and Humanities Research Council of Canada, the Fonds pour la formation de chercheurs et l'aide à la recherche of Québec, the Fonds Marcel-Faribault of the Université de Montréal and the Paradi granted to the C.R.D.E. by the Canadian International Development Agency.

C.P. 6128, succursale A
Montréal (Québec)
H3C 3J7

Télécopieur (FAX): (514) 343-5831
Courrier électronique (E-Mail): econo@tornade.ERE.Umontreal.CA

Ce cahier a également été publié au Centre de recherche et développement en économique (C.R.D.E.) (publication no 0594).

Dépôt légal - 1994
Bibliothèque nationale du Québec
Bibliothèque nationale du Canada

ISSN 0709-9231

RÉSUMÉ

L'objectif de ce rapport est de proposer de nouveaux estimateurs obtenus par la méthode des moments généralisée pour des modèles linéaires de régression multiple avec erreurs sur les variables explicatives. On sait que l'estimateur des moindres carrés ordinaires, qui se base sur les moments échantillonnaux d'ordre deux, est centré lorsqu'il n'y a pas d'erreurs sur les variables, mais qu'il devient biaisé et non convergent en présence de telles erreurs. Les nouveaux estimateurs sont, de leur côté, basés sur des moments échantillonnaux d'ordres supérieurs à deux. Ils demeurent convergents, sous des hypothèses très raisonnables, lorsqu'il y a des erreurs de mesure. Un des estimateurs proposés est également centré dans le cas où les variables explicatives sont mesurées sans erreur. Alors que la plupart des estimateurs convergents basés sur des moments échantillonnaux d'ordres supérieurs à deux, qui ont été suggérés jusqu'à maintenant pour les modèles de régression avec erreurs sur les variables, sont considérés comme étant plutôt erratiques, les estimateurs que l'on propose ici semblent se comporter remarquablement bien dans diverses situations.

Bien que les données comportent la plupart du temps des erreurs de mesure, ce fait est souvent négligé par les analystes qui appliquent généralement des procédures statistiques conçues pour des données mesurées sans erreur. À ce sujet, on démontre que le fait d'ignorer la présence d'erreurs sur les variables peut avoir comme conséquence que les erreurs de type I associées aux tests de Student standards dépassent considérablement le niveau désiré. Par ailleurs, les nouveaux estimateurs proposés ne présentent pas de tels inconvénients. Les résultats de nombreuses expériences suggèrent également que lorsque les erreurs sur les variables ne sont pas négligeables, ces estimateurs dominent l'estimateur des moindres carrés ordinaires en termes d'erreurs quadratiques moyennes.

On propose également des tests d'erreurs sur les variables et on évalue la puissance de ces tests à partir d'expériences de Monte Carlo.

Mots clés : erreurs sur les variables, erreurs de mesure, estimateurs de moments d'ordres supérieurs, estimateurs GMM.

ABSTRACT

This paper proposes new generalized method of moments (GMM) estimators for multiple linear regression models with errors in the explanatory variables. As is very well known, the ordinary least squares estimator (OLS), which is based on the sample moments of order two, is unbiased when there are no errors in the variables, but it becomes biased and inconsistent when there are such errors [Fuller (1987)]. In contrast, the suggested estimators are based on higher sample moments. They are consistent, under quite reasonable assumptions, when there are measurement errors. One of the proposed estimators is also unbiased when the explanatory variables are measured without error. While most consistent estimators based on higher moments (HM) proposed previously in the literature for regressions with errors in the variables seem to be quite erratic, the suggested estimators appear to perform remarkably well in many situations.

Although most data do contain errors of measurement, this fact is often ignored by the analysts, and statistical procedures designed for data measured without error are applied. It is shown that ignoring the presence of measurement errors and using traditional OLS estimators may lead to performing standard Student's t tests with type I errors of considerably higher sizes than intended, while this is not so with the proposed HM estimators. Our experimental findings suggest also that, when the errors in the variables are nonnegligible, our estimators do perform better than the OLS estimators in terms of root mean squared errors.

Tests for the presence of errors in the variables are also described, and the power of the tests are assessed in the Monte Carlo experiments.

Key words : errors in the variables; measurement errors; higher moments estimators, GMM estimators.

1. INTRODUCTION

Most data used in empirical analyses contain errors of measurement. Such errors are probably relatively more important in macroeconomic studies [Morgenstern (1963), Langanskens and Van Rickeghem (1974), Dagenais (1992)], but they are also present in most microeconomic analyses [Rodgers, Brown and Duncan (1993); Duncan and Hill (1985); Altoriji and Siow (1987)]. Although the early econometricians insisted greatly on the presence of errors in the variables, this phenomenon has not been strongly emphasized in the ensuing developments of the discipline [Goldberger (1992); Morgenstern (1963); Griliches and Hausman (1986); Griliches (1986)].[1] In the present state of the art, most econometric textbooks contain a rather short section where it is demonstrated that in linear regression models, errors in the explanatory variables lead to inconsistent ordinary least squares (OLS) estimators. Unless information is available on the variances of these errors, authors suggest essentially the use of instrumental variables [Fuller (1987), Bowden (1984), Aigner et al. (1984)] to obtain consistent estimators. Despite the fact that in applied papers authors often warn the reader that the possible presence of errors in the variables may bias the results, in many cases, no special effort is made to resort to instrumental variable techniques to reduce the possible biases and no special step is taken to test for the presence of errors in the variables (EV) using, for example, Hausman's (1978) instrumental variable test. The attitude of most applied researchers is probably due, in a number of cases, to the fact that it is not always easy to verify that the available instrumental variables satisfy the required conditions to justify their use [Pal (1980)]. In other cases, the eligible instruments may simply not be easily accessible to the researcher [Klepper and Leamer (1984)], and one may feel that the cost of collecting the additional data would be too large in comparison to the benefit derived from the fact of possibly producing somewhat more accurate estimators.

In line with the above considerations, one of the purposes of the present paper is to insist on the perverse effects of the presence of errors of measurement in the independent variables on statistical inference from standard linear regression models. Such errors in the variables lead to inconsistency of the OLS estimators of the regression parameters, to larger mean-squared errors and probably, most importantly, to larger than intended sizes of type I errors of Student tests.

---

1 Adverse effects of the presence of errors in the variables in regression models *with autocorrelated errors* have been underlined in Dagenais (1994) and Grether and Maddala (1973).

Inconsistency. The case of the simple regression model is well known. This model can be expressed as follows :

$$Y = \alpha \iota_N + \beta \bar{X} + u , \tag{1}$$

where ι_N is a $N \times 1$ unit vector, $\bar{X}$ is a nonstochastic vector of values of the explanatory variable, u is a vector of random regression errors with zero mean and finite variance σ_u^2. Y is the vector of the dependent variable and α and β are unknown coefficients. It is further assumed that $\bar{X}$ is observed with a measurement error. The observed variable is in fact :

$$X = \bar{X} + V \tag{2}$$

where V is a vector of random errors uncorrelated with the elements of u, with mean zero and variance σ_v^2. For this simple case, textbooks show that the OLS estimator ($\hat{\beta}$) of β obtained while ignoring the presence of EV is inconsistent. Indeed, one obtains :

$$\underset{N\to\infty}{\text{Plim}}\ \hat{\beta} = \underset{N\to\infty}{\text{Plim}} \sum_{i=1}^{N} x_i y_i \Big/ \sum_{i=1}^{N} x_i^2 = \beta\left[1 - \frac{\lambda}{1+\lambda}\right] \tag{3}$$

where x and y correspond to X and Y in mean deviation form, $\lambda = \sigma_v^2 / \sigma_{\tilde{x}}^2$, $\sigma_{\tilde{x}}^2 = \underset{N\to\infty}{\text{lim}} \sum_{i=1}^{N} \tilde{x}_i^2 / N$ and N is the sample size.

In this case, $\underset{N\to\infty}{\text{Plim}}\ \hat{\beta}$ is always of the same sign as β and is smaller than β in absolute value. Moreover, $\hat{\beta}$ remains a consistent estimator when $\beta = 0$. The situation, however, is not as neat when the model includes more than one regressor. For example, let us take the two regressors case :

$$Y = \alpha \iota_N + \beta_1 \bar{X}_1 + \beta_2 \bar{X}_2 + u , \tag{4}$$

and let us assume that both $\bar{X}_1$ and $\bar{X}_2$ are measured with error. Assuming further that the errors of measurement associated with $\bar{X}_1$ and $\bar{X}_2$ are independent, one obtains the following expression for the probability limit of the OLS estimator :

$$\underset{N\to\infty}{\text{Plim}}\ \hat{\beta} = \underset{N\to\infty}{\text{Plim}} \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} =$$

$$\begin{bmatrix} \beta_1(c_{11} + r_{12}\, c_{12}) + \beta_2(r_{12}\, c_{11} + c_{12})\ \sigma_2/\sigma_1 \\ \beta_1(c_{21} + r_{12}\, c_{22})\ \sigma_1/\sigma_2 + \beta_2(r_{12}\, c_{21} + c_{22}) \end{bmatrix} \tag{5}$$

where

$$\begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix} = \begin{bmatrix} 1+\lambda_1 & r_{12} \\ r_{21} & 1+\lambda_2 \end{bmatrix}^{-1} . \tag{6}$$

$$\sigma_1 = \sqrt{\sigma_{\tilde{x}_1}^2}\ ,\ \sigma_2 = \sqrt{\sigma_{\tilde{x}_2}^2}\ ,$$

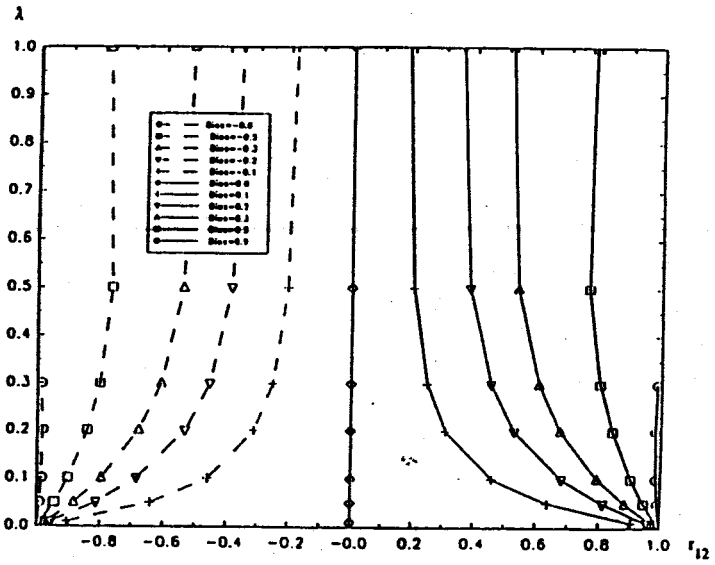
$r_{12} = r_{21}$ = correlation coefficient between $\bar{X}_1$ and $\bar{X}_2$,

$$\lambda_1 = \sigma_{v_1}^2 / \sigma_{\tilde{x}_1}^2\ ,\ \lambda_2 = \sigma_{v_2}^2 / \sigma_{\tilde{x}_2}^2\ .$$

In this case, depending upon the values of r_{12}, λ_1, λ_2 and σ_2/σ_1, the relationship between $\underset{N\to\infty}{\text{Plim}}\ \hat{\beta}_i$ and $\beta_i (i = 1, 2)$ is much more intricate. In particular, when $r_{12} \neq 0$, the OLS estimator of the coefficient associated with one of the variables will generally remain inconsistent even when its true value is equal to zero, if the coefficient associated with the other variable is different from zero and there are errors of measurement in at least one of the explanatory variables. As an illustration, Figure 1 reproduces "iso-inconsistency" curves. These curves show the differences between $\underset{N\to\infty}{\text{Plim}}\ \hat{\beta}_2$ and β_2 for different values of λ and r_{12}, for the case where $\beta_1 = 1$, $\beta_2 = 0$, $\lambda_1 = \lambda_2 = \lambda$, $\sigma_1^2 = 1$ and $\sigma_2^2 = 0.2$. One notes that even if $\beta_2 = 0$, $\underset{N\to\infty}{\text{Plim}}\ \hat{\beta}_2$ differs from β_2 and that this difference can be either positive or negative, depending on the sign of r_{12}. Furthermore, one notes that, for example, the difference between $\underset{N\to\infty}{\text{Plim}}\ \hat{\beta}_2$ and β_2 is the same (namely, 0.2) for $\lambda = 2.5$ % and $r_{12} = 0.9$, and for $\lambda = 75.5$ % and $r_{12} = 0.36$. Therefore, in this case, the same difference could be generated with a measurement error that is thirty times smaller if the correlation between the explanatory variables is 2.5 times larger.

Figure 1

"Iso-Inconsistency" Curves for $\hat{\beta}_2$ as a Function of r_{12} and λ



Mean-squared error.[2] Although OLS estimators have relatively small variances, the fact that the mean-squared error equals variance *plus squared bias* may cause OLS estimators of regression models with errors in the independent variables to have larger mean-squared errors than alternative consistent estimators (with smaller finite sample biases), even if these estimators have larger variances. This will often be the case when a) the variances of the errors of measurement are relatively large since in this case, the biases of the OLS estimators will be important, b) the sample size is relatively large since then, the variances of all estimators are small and the relative importance of the squared biases of the OLS estimators is greater. Situations of these types will be illustrated below, in the results of the Monte Carlo experiments.

Type I error. One of the most perverse effects of ignoring the presence of errors in the variables in regression models, when such errors affect the *independent* variables, concerns the highly misleading determination of the confidence intervals of the regression parameters and, correlatively, of the sizes of the type I errors when

[2] See section 5 for comments on the conditions of existence of biases, variances and mean-squared errors of parameter estimators for regression models with errors in the variables, in a finite sample.

testing hypotheses. Because OLS estimators have relatively small variances but are biased when there are errors in the variables (and therefore their distributions are not centered on the true values of the parameters), intended 95 % confidence intervals obtained while ignoring the presence of errors in the variables may in practice turn out to be almost 0 % confidence intervals, even in cases where the errors of measurement are not exceedingly large, as will be illustrated below! This means that instead of having the computed "95 % confidence interval" include the true value of the parameter in 95 % of the cases as it should, one may be almost sure that this interval *will not* include the true value of the parameter! Similarly, Student t-tests of specific values of the parameters, using the critical value that corresponds normally to type I errors of size equal to 5 %, may in fact correspond to tests with type I errors of size equal to almost 100 %! This may have a dramatic consequence when one tests a theory that implies, for example, that a given coefficient is equal to zero. Because, as we have seen above, in multiple regression models with errors in the variables, the OLS parameter estimator may be relatively strongly biased even when the true parameter is zero, and because OLS estimators have relatively small variances, one may be induced to reject the null hypothesis when this hypothesis is true, with a probability close to 100 %! The paradox of this unfavorable situation is that, contrary to the traditional case, increasing the sample size does not improve the matter but worsens it, since the importance of the bias relative to that of the standard error of the parameter estimator increases.

Now, even if one is convinced that it is important to take account of the possible presence of errors in the variables of one's data set when running linear regressions, one is still left with the problem that it is not easy to identify appropriate instrumental variables and that these variables may often not be readily available [Pal (1980)]. An alternative to the instrumental variable approach in situations involving errors in the variables, that has received little attention in the literature, is to use consistent estimators based on sample moments of higher order than two. Pal (1980) presents a number of such estimators which have the property to remain consistent when there are errors in the explanatory variables, under quite reasonable hypotheses. Pal proposes several estimators based on third-order sample moments for the cases of regression models with one explanatory variable. He mentions that one of these estimators had already been suggested by Durbin (1954), another one by Drion (1951) and a third one by Geary (1942). In particular, Durbin's estimator has the property of also being unbiased when there are no errors in the variables. It also extends readily to the case of models containing more than one regressor. Consistent estimators based on even higher sample moments also exist. For the case of regression

models with one explanatory variable, estimators based on fourth-order moments have been proposed by Geary (1942) as well as Pal (1980). Although these estimators have not been generalized for regression models involving more than one regressor, we will see below that it is possible to do so. One of these latter estimators is also unbiased when there are no errors in the variables.

However, it has long been recognized that regression estimators based on higher moments are notably more erratic than the corresponding least squares estimators [Kendall and Stuart (1963), p. 56; Malinvaud (1978)]. This most probably explains why such estimators have almost never been used in actual applications. The main purpose of this paper is therefore to suggest new higher moment (HM) GMM estimators [Hansen (1982)] which in the examples considered in our numerical applications turn out to have *considerably smaller standard errors* than the HM estimators previously suggested, for reasons which will be explained in section 5.[3]

In our numerical illustrations section, which is based on Monte Carlo experiments, it will be shown that in survey data analyses from samples of several hundred observations, even when the variance of the errors of measurement for a given variable is not very large (10 % of the variance of the variable), usual t tests based on OLS estimators performed at the 5 % intended significance level may, in fact, have a probability of type I error of more than 85 %. Paradoxically, as mentioned above, contrary to what happens in the traditional model without errors in the variables, the results become less and less reliable and precise as the sample grows and the estimated standard errors shrink. If the measurement errors are larger, the same problem may be encountered for much smaller samples. In contrast, tests based on our HM estimators have type I error probabilities of approximately the right size in all situations. In terms of root mean-squared error, which is a traditional measure of performance of estimators, our Monte Carlo experiments suggest that in a number of situations, the OLS estimator beats our HM estimators in small samples, but that in larger samples, the HM estimators are superior, even when the variances of the measurement errors are small. Even in samples of less than 100 observations, however, if the measurement errors have relatively large variances (e.g., 25 % of the variances of the affected variables), our experimental findings indicate that the HM estimators may still, in certain cases, turn out to have smaller mean-squared errors than the OLS estimator.

3 For an alternative approach to obtain consistent regression estimators applicable to a variety of errors-in-variables models with panel data, see Griliches and Hausman (1986).

It is also possible to use our suggested estimators to perform tests of errors in the variables. We shall indeed propose a simple procedure below. It appears from our experiments that the proposed test may be useful to detect the presence of errors in the variables when it really matters. This is the case in large samples, even if the errors are relatively small and if the multiple correlation coefficient is low : a situation often encountered in microeconomic analyses based on survey data. It is also the case in small samples, when the multiple correlation coefficient is high and the measurement errors are relatively important : a typical situation in macroeconomic studies.

2. THE SUGGESTED ESTIMATORS

Let us assume that we have the following regression model :

$$Y = \alpha \iota_N + \bar{X} \beta + u , \tag{7}$$

where $\bar{X}$ is a $N \times K$ matrix of nonstochastic explanatory variables measured without error, with empirical distribution such that $\lim_{N \to \infty} \frac{\bar{X}' \bar{X}}{N} = Q$, where Q is a finite non-singular matrix. The $N \times 1$ vector u is a vector of normal residual errors with covariance matrix $\sigma^2 I_N$ and Y is the $N \times 1$ vector of observations of the dependent variable. The $K \times 1$ vector β and σ^2 are unknown parameters. The scalar α is also an unknown parameter.

We also assume that $\bar{X}$ is unobservable and that the matrix X is observed instead, where

$$X = \bar{X} + V \tag{8}$$

and V is a $N \times K$ matrix of normally distributed errors in the variables. It is further assumed that V is uncorrelated with u and that

$$\mathrm{Var}[\mathrm{Vec}(V)] = \Sigma \otimes I_N , \tag{9}$$

where $\mathrm{Var}[\cdot]$ stands for the covariance matrix, Σ is a $K \times K$ symmetric positive definite matrix and I_N designates the identity matrix of order N. This last assumption implies that the errors in the variables are independent between observations but not between variables. It also implies that for a given variable, the errors of measurement are homoskedastic.

The above model may be rewritten as :

$$Y = \alpha\, \iota_N + X\beta + u - V\beta = \alpha\, \iota_N + X\beta + \eta\,. \tag{10}$$

Our first proposed HM estimator of $\theta = (\alpha, \beta)'$ is obtained as the asymptotically optimal GMM estimator derived from the following orthogonality conditions :

$$\underset{N\to\infty}{E}\ (Z'\eta / N) = 0 \tag{11}$$

where

$$Z = (\iota_N, z_1, \ldots, z_7)\,, \tag{12}$$

$$z_1 = x * x\,,\ z_2 = x * y\,,\ z_3 = y * y\,, \tag{13}$$

$$z_4 = x * x * x - 3\, x\, [\underset{N\to\infty}{Plim}(x'x / N) * I_K]\,, \tag{14}$$

$$z_5 = x * x * y - 2\, x\, [\underset{N\to\infty}{Plim}(x'y / N) * I_k] - y * [\iota_K' \underset{N\to\infty}{Plim}(x' x / N) * I_k]]\,, \tag{15}$$

$$z_6 = x * y * y - x\, [\underset{N\to\infty}{Plim}(y'y / N)] - 2\, y\, [\underset{N\to\infty}{Plim}(y'x / N)]\,, \tag{16}$$

$$z_7 = y * y * y - 3\, y\, [\underset{N\to\infty}{Plim}(y'y / N)]\,, \tag{17}$$

where the symbol * designates the Hadamard element by element matrix multiplication operator.[4]

This HM estimator ($\tilde{\theta}$) minimizes :

$$\varphi = (\eta'Z / N)\, W^{-1} (Z'\eta / N) \tag{18}$$

where

$$W = E(Z'\eta\ \eta'Z / N)\,. \tag{19}$$

[4] The proof that $\underset{N\to\infty}{E}\ (Z'\eta / N) = 0$ is rather straightforward but tedious. Part of the proof [for $\underset{N\to\infty}{E}\ (z_1'\eta / N) = 0$ and $\underset{N\to\infty}{E}\ (z_4'\eta / N) = 0$] can be found in Dagenais and Dagenais (1993), Appendix B.

A feasible estimator is obtained by replacing $\underset{N\to\infty}{Plim}\ x'x / N$, $\underset{N\to\infty}{Plim}\ x'y / N$ and $\underset{N\to\infty}{Plim}\ y'y / N$ by x'x / N, x'y / N and y'y / N. Furthermore, $E(Z'\eta\ \eta'Z / N)$ is replaced by $\sum_{i=1}^{N} (Z_i'\, \eta_i^2\, Z_i) / N$ [White (1982)], where Z_i is the i'th row of the Z matrix, $\eta_i = Y_i - \alpha - X_i\, \beta$ and Y_i and X_i correspond also to the i'th rows of Y and X. The solution is found iteratively. To obtain the initial solution ($\tilde{\theta}_0$), one can simply replace $E(Z'\eta\ \eta'Z / N)$ by Z'Z / N, which yields

$$\tilde{\theta}_0 = [\bar{X}'Z(Z'Z)^{-1}\, Z'\bar{X}]^{-1}\, \bar{X}'Z(Z'Z)^{-1}\, Z'Y\,, \tag{20}$$

where $\bar{X} = (\iota_N, X)$.

Under the appropriate hypotheses [Hansen (1982), Gouriéroux and Montfort (1989)], our optimal GMM estimator is asymptotically normally distributed with covariance matrix $N(\bar{X}'Z\, W^{-1}\, Z'\bar{X})^{-1}$. One of these hypotheses is that the rank of $Z'\bar{X} / N$ is equal to K + 1. This implies that the empirical distribution of $\bar{X}$ is not multivariate normal [Reiersol (1950)].

We will also consider, in the numerical illustrations of section 5, the estimator θ^* obtained by replacing Z by $\hat{Z}$ in equation (18), where $\hat{Z} = (\iota_N, z_1, z_4)$. It is interesting to note that if Z is replaced by $\hat{Z}$ and W is replaced by $\hat{Z}'\hat{Z}$, the resulting initial solution θ_0^* is a consistent and *unbiased* estimator of θ when there are no errors in the variables, and it remains consistent when there are errors in the variables.

3. TESTING FOR THE PRESENCE OF ERRORS IN THE VARIABLES

The null hypothesis (H_0) that there are no errors in the variables can be tested by applying a Hausman (1978) type test. This test is most easily performed by the following procedure.

1) Run regressions of the X's as dependent variables on Z, as the matrix of independent variables :

$$X = Z\Gamma + w \tag{21}$$

where Γ is a matrix of parameters and w is the matrix of regression errors.

2) Compute $\hat{X} = Z\hat{\Gamma}$ and $\hat{w} = X - Z\hat{\Gamma}$, where $\hat{\Gamma}$ is the OLS estimator of Γ obtained in step one.

3) Run the following augmented regression :

$$Y = \alpha\iota_N + X\beta + \hat{w}\psi + \varepsilon, \tag{22}$$

where ψ is a vector of parameters and ε is the vector of the regression errors.

4) Test $\psi = 0$, using the usual F test.

If there are no errors in the variables, $X = \bar{X}$ and $Y = \alpha\iota_N + X\beta + u$. Therefore, under H_0, $\varepsilon = u$ and $\psi = 0$. When Z is used in step one, the resulting F test is asymptotic. If $\bar{Z}$ is used instead of Z, the test is exact in finite samples.

4. POSSIBLE EXTENSIONS

Several possible extensions of the HM estimators described in section 2 come to mind. For example, if one is not willing to make the assumption that the nature of the joint density function of the errors in the variables is known, but only that it is symmetric, z_4 should be removed from the definition of Z in equation (12), to preserve the consistency of the estimator. Similarly, if one assumes that the density of the regression errors is unknown but symmetric, z_7 should be removed for the same reason.[5]

If one assumes that the u's have a given *known* symmetric density other than the normal, z_7 could be retained. The factor 3 appearing in the second term of the right-hand side of equation (17), which defines z_7, should then be replaced by K_1, where K_1 is a known quantity equal to the ratio of the fourth centered moment of the density of the u's divided by the square of its second centered moment. Similarly, if one assumes that the joint density of the V's is a given joint symmetric density other than the normal and that this density has the following property :

$$\mu(i, j) = K_2\, \mu(2, 0)\, \mu(2 - j, j), \text{ for } (i, j) = (4,0) \text{ or } (i, j) = (3, 1),$$

where $\mu(i, j)$ designates the centered bivariate cross-moment of order i, j, the factor 3 appearing in the second term of the right-hand side of equation (14), which defines z_4, should be replaced by K_2. Note that if the value of K_2 is unknown, it could be considered as an extra parameter to be estimated. (In the case of the normal distribution, $K_2 = 3$ [Kendall and Stuart (1963, p. 91)].)

One could also devise a pretest estimator by adopting the following procedure : 1) test for the presence of errors in the variables, 2) if H_0 cannot be rejected, use the OLS estimator, otherwise use the HM estimator.

One could introduce on the right-hand side of equation (12), which defines the Z matrix, other "instrumental variables" based on higher sample moments than the third and fourth, or extraneous instrumental variables that are available. If some variables are assumed to be observed without error, they could be introduced directly in the Z matrix, as was done for the ι_N vector. In the same vein, one could make separate EV tests for each of the X variables, that is, make separate tests for each of the elements of ψ in equation (22) and "instrument" only the variables for which the null hypothesis is rejected.

Some of the assumptions underlying the model presented in section 2 could also be relaxed. For example, if the V's are assumed to be heteroskedastic, our estimators $\tilde{\theta}$ and $\tilde{\theta}_0$ would remain consistent, provided the distribution of the variances of the V's is independent of the $\bar{X}$'s, and provided z_4 is excluded from Z. θ^* and θ_0^* would also remain consistent, provided z_4 is replaced by z_5 or z_6 in $\bar{Z}$. Similarly, if the u's are assumed to be heteroskedastic, all four θ estimators would remain consistent if the variances of the u's are distributed independently of the $\bar{X}$'s. In the case of $\tilde{\theta}$ and $\tilde{\theta}_0$, z_7 would also have to be removed from Z. If the V's or the u's are serially correlated, provided they are stationary and ergodic [White (1984)], our estimators would still be consistent. The matrix W would, however, be somewhat more complicated to evaluate.

5. NUMERICAL ILLUSTRATIONS

A) Preliminary remarks

The numerical experiments reported below concern essentially the performance of the $\tilde{\theta}$ and θ^* estimators relative to the OLS estimator $\hat{\theta}$.

The performance criteria generally used in such studies are the bias and the root mean-squared error (RMSE). We also use the discrepancy between the intended and true sizes of the type I errors for tests of null hypotheses, because this criterion appears to be particularly important in the present context for reasons given previously. A major problem arises, however, in the case at hand, because under the hypotheses

underlying the model outlined at the beginning of section 2, none of the estimators considered, including the OLS estimator, have moments in finite samples.[6]

Indeed, in the case of the OLS estimator of the simple regression model, if $V = -\bar{X}$, we have $X = 0$ and x'x has no inverse. Since V is assumed to be normal, its density function at $V = -\bar{X}$ is not strictly equal to 0 and therefore the integrand that appears in the formula of the expected value of $\hat{\theta}$:

$$E(\hat{\theta}) = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} \hat{\theta}\, f(u, V)\, du\, dV\,, \qquad (23)$$

where f(u, V) is the joint density of the elements of u and V, is a discontinuous function at the point where $V = -\bar{X}$. $E(\hat{\theta})$ will therefore only exist if one is ready to assume that the point $V = -\bar{X}$ is excluded from the sample space. We will implicitly be making this type of hypothesis when evaluating "biases" and "root mean-squared errors" in our Monte Carlo experiments and also when using the estimated standard errors of the parameters to test null hypotheses.

The fact that the considered estimators have no moments in finite samples explains why, in practice, the HM estimators proposed hitherto in the literature [Geary (1942), Drion (1951), Durbin (1954), Pal (1900)] have been found to be quite erratic [Kendall and Stuart (1963), Malinvaud (1978)] and have seldom been used in actual applications. It is well known that the distribution of estimators which have no first and second moments may have fat tails and occasionally produce large outliers [Mariano and Sawa (1972)]. This appears to be the case, in fact, for HM estimators that use only *one* of the z_i's (i = 1, ..., 7) defined above,[7] as will be illustrated by our Monte Carlo studies. However, no such outliers were observed with $\hat{\theta}$, nor with our θ^* or $\bar{\theta}$ estimators that use *simultaneously* z_1 and z_4 or all z_i's (i = 1, ..., 7) respectively. Conditions that produce extreme outliers are more likely to be met when only one of

[6] The moments of their asymptotic distributions do exist, however.

[7] Clearly, z_3 and z_7 can be used alone only in the case of the simple regression model.

the z_i's is used as the set of instrumental variables to estimate θ, than when both z_1 and z_4, or all z_i's, or X itself are utilized.[8]

An alternative to using computed biases as measures of central tendency is to report differences between the medians of the estimates and the true values of the parameters, since the medians do exist for all the considered estimators. Similarly, truncated root mean-squared errors can be computed excluding the two extreme deciles, since again, such truncated RMSE's do exist for the estimators considered. It will be seen, however, that the results using the medians and truncated RMSE's are not very different from those based on traditionally computed biases and RMSE's.

B) Description of data and experiment setup

The data used for the Monte Carlo experiments reported below are drawn from the 1986 survey of Consumer Finances of Statistics Canada (1988). A simple model relating total annual household consumption to the following variables was set up :

$\bar{X}_1$: total annual income of the household;

$\bar{X}_2$: age of the head of the household;

$\bar{X}_3$: number of person-weeks constituting the household during the year.

In order to preserve a certain homogeneity of the sample, we retained only the observations for which the total income of the household ranged between $ 25,000 and $ 55,000. The total sample available included 4,400 observations. We first ran a regression using observed consumption and observed income, age and person-weeks. Then we scaled the explanatory variables so that each of the estimated coefficients became equal to one. We then used the independent variables thus scaled with

[8] This is easier to see in the case of the $\bar{\theta}_0$ estimator shown in equation (20). In that case, θ is estimated by $[\bar{X}'Z(Z'Z)^{-1}Z'\bar{X}]^{-1}\bar{X}'Z(Z'Z)^{-1}Z'Y$. Conditions for the matrix $\bar{X}'Z(Z'Z)^{-1}Z'\bar{X}$ to be nearly singular are likely to be met more often when Z is of the same size as X and $Z'\bar{X}$ / N involves only either the third *or* the fourth sample moments of the independent variables, than when the number of columns in Z is two or more times greater than the number of columns in X and $Z'\bar{X}$ / N involves *simultaneously* third and fourth moments of the variables.

$\theta = (1, 1, 1, 1)'$ to generate the consumption vectors used in our Monte Carlo experiments.[9]

More precisely, the model used was the following :

$$Y_i = X_{0i} + \tilde{X}_{1i} + \tilde{X}_{2i} + \tilde{X}_{3i} + u_i \quad (i = 1, \ldots, N)$$

where $X_{0i} = 1$ for every i, u_i is the normal regression error term, Y_i is household consumption and N is the sample size. All data are expressed in logarithms. Then normal errors in the variables were added to the $\tilde{X}_{1i}$ variable to obtain $X_{1i} = \tilde{X}_{1i} + V_{1i}$. Since no errors were added to $\tilde{X}_{2i}$ and $\tilde{X}_{3i}$, we have $X_{2i} = \tilde{X}_{2i}$ and $X_{3i} = \tilde{X}_{3i}$.

Because the squared multiple correlation coefficient (R^2) that was obtained when we regressed the actual data was equal to approximately 0.40, we set, in all experiments reported in the next subsection, the variance of u_i so as to obtain a theoretical squared multiple correlation coefficient of 0.40 when using our 4,400 available observations. Similarly, since studies made on the accuracy of reported earnings data [Rodgers, Brown and Duncan (1993)] suggest that the ratio of the variance of measurement errors to the variance of declared individual earnings expressed in logarithms is approximately 0.2 for men, and since measurement errors for total household income, which includes nonlabor income, is believed to be greater than for earnings [Altonji and Siow (1987); Radner (1982)], we set, in our first experiment, the variance of V_1 so as to obtain a value of $\lambda_1 = \left[\sigma^2_{v_1} / \sigma^2_{\tilde{x}_1}\right]$ equal to 0.3.[10] Finally, in this same experiment, we set N = 2,000, which is not a very large sample size, according to present standards, for household surveys.

We then made a second experiment where the value of λ_1 was reduced markedly to 0.1, and a third experiment where, in turn, N was reduced to 700. We also made a fourth experiment where we set $\beta_3 = 0$ instead of 1 and verified the size of the probability of rejecting this (true) null hypothesis, using our different estimators.

9 Setting all elements of θ to 1 simplifies the analysis of the tables of results shown below.

10 Rodgers, Brown and Duncan (1993) also suggest that in the case of earnings, the measurement errors are not independent of the true values, but they are negatively correlated ("mean reverting"). It is not known, however, if this is also the case for total income. The possibility that measurement errors are correlated with the true values is not considered in the present paper. Further research is needed to analyze this case.

Finally, since the collinearity between our independent variables is very small, we made a fifth experiment analogous to the fourth one, in which $\tilde{X}_3$ was transformed so as to be much more highly correlated with $\tilde{X}_1$. Prior to transformation, the correlation coefficient between $\tilde{X}_1$ and $\tilde{X}_3$ was 0.15 and after transformation it was equal to 0.5.

Before reporting our results, we would like to insist on the fact that the naïve model presented here is used only for ***illustration*** purposes. It is clearly inadequate for analyzing household consumption. It could be argued, on the one hand, that consumption may be more closely related to *perceived* income than to actual income and that declared income may be closer to perceived income than to true income. This would suggest that the value of λ_1 may be smaller than that used in all our experiments, except possibly in the second experiment. On the other hand, if consumption depends on "permanent" income, the discrepancy between this notion of income and declared annual income might correspond to a much larger value of λ_1. This would most likely still be true even if $\tilde{X}_1$ were replaced by better approximations to "permanent" income than the declared annual income.[11] Finally, it must be pointed out that present consumption might also be influenced by past savings [Avery (1991)] or accumulated wealth [Avery, Elliehausen and Kennickel (1988)], and these variables are likely to contain even much larger errors in the variables than income.

C) The results of the Monte Carlo experiments

We report, in Table 1, section A, from our first experiment,[12] the root mean-squared errors for the $\hat{\theta}$, θ^* and $\bar{\theta}$ estimators. In addition, similar results are given for the estimators based only on z_1 [Durbin (1954)] or z_4 [Pal (1980)] designated respectively as θ_d and θ_p.

We also show the RMSE's for θ_0^* and $\bar{\theta}_0$, since these estimators, which are used as initial solutions to compute θ^* and $\bar{\theta}$ and are much easier to calculate, also perform well.

11 See Jeong and Maddala (1991) about measurement errors in expectations data.

12 All Monte Carlo experiments reported below are based on 400 replications. This number of replications appeared to be sufficient to assure a two-digit accuracy (based on 95 % confidence intervals) for most of the results presented in the tables.

Table 1
Root Mean-Squared Errors - Experiment 1
($r = 2{,}000$; $R^2 = 0.4$; $\lambda_1 = 0.3$)

A) Root Mean-Squared Errors

	θ_d	θ_p	$\hat{\theta}$	θ_0^*	θ^*	$\bar{\theta}_0$	$\bar{\theta}$
α	1.888	1.681	0.861	0.347	0.352	0.327	0.349
β_1	0.374	0.342	0.237	0.085	0.087	0.077	0.082
β_2	1.307	1.484	0.127	0.237	0.238	0.224	0.234
β_3	0.451	6.312	0.121	0.155	0.157	0.153	0.163
Average RMSE	1.005	2.455	0.337	0.206	0.209	0.195	0.207

B) Truncated Root Mean-Squared Errors

	θ_d	θ_p	$\hat{\theta}$	θ_0^*	θ^*	$\bar{\theta}_0$	$\bar{\theta}$
α	1.134	1.003	0.849	0.232	0.236	0.219	0.226
β_1	0.209	0.220	0.235	0.052	0.053	0.048	0.052
β_2	0.849	0.984	0.090	0.156	0.157	0.147	0.156
β_3	0.260	4.140	0.104	0.100	0.102	0.098	0.109
Average RMSE	0.612	1.587	0.320	0.135	0.137	0.128	0.136

These results clearly illustrate why HM estimators previously proposed in the literature were almost never used in practice. For example, the averages of their RMSE's for all four parameters are approximately from five to ten times larger than the average RMSE for the elements of θ^*. They are even larger than the average RMSE associated with $\hat{\theta}$, which is itself more than 60 % larger than those of our suggested estimators. Using the averages of the square roots of the *truncated* mean-squared errors improves the performance of all HM estimators relative to that of the OLS estimator, but it still rules out θ_d and θ_p. The findings reported in Table 1 suggest that for all the estimators considered, the additional information obtained from the *truncated* RMSE's is quite consistent with the results of the RMSE's computed without truncation. Therefore, only the results concerning $\hat{\theta}$, θ_0^*, θ^*, $\bar{\theta}_0$ and $\bar{\theta}$ will be reported in the following tables and only the RMSE's computed without truncation.

Table 2 also reports the results of experiment 1. It gives the biases and sizes of type I errors associated with each parameter. The sizes of the type I errors were measured by calculating the percentage of replications for which the true value of the parameter was *not* included in the computed 95 % confidence interval. One notices that the bias of the OLS estimator of β_1 is close to what it would be in the simple regression case, namely : 0.3/1.3 = 0.231. This is not surprising since in this example, the correlation between $\bar{X}_1$ and $\bar{X}_2$ as well as $\bar{X}_3$ is very low. What is *most disturbing*, however, is that the computed sizes of the type I errors for the OLS estimators of α and β_1 are equal to 100 %! The sizes of the type I errors for all the elements of our HM estimators are, on the contrary, much closer to the intended 5 % level. Finally, the powers of the tests based on $\bar{Z}$ and Z are both quite high.

Table 3 gives the results of the second experiment. In this experiment, the relative importance of the measurement errors in $\bar{X}_1$ was *reduced notably*, since λ was set to 0.1. Even then, the performance of our HM estimators is comparable to that of the OLS estimator in terms of the average values of the root mean-squared errors. However, the OLS estimator behaves *rather poorly*, as far as the sizes of type I errors for α and β_1 are considered. As could be anticipated, the power of the tests has decreased notably. Note that these tests are based on 5 % critical values. In the case of the Durbin-Watson autocorrelation test, Fomby and Guilkey (1978) have argued that 50 % critical values should be used instead of the traditional 5 %. A similar strategy would clearly increase the power of our tests.

Table 2
Biases, Type I Errors and EV Tests – Experiment 1
($r = 2{,}000$; $R^2 = 0.4$; $\lambda = 0.3$)

A) Biases

	$\hat{\theta}$	θ_0^*	θ^*	$\bar{\theta}_0$	$\bar{\theta}$
α	0.846	0.020	0.011	0.021	0.005
β_1	-0.235	-0.004	-0.002	-0.006	0.000
β_2	-0.046	0.002	0.002	-0.004	0.003
β_3	0.091	-0.008	-0.009	-0.003	-0.009
Average of absolute values	0.305	0.009	0.006	0.008	0.004

B) Size of Type I Errors, in %

	$\hat{\theta}$	θ_0^*	θ^*	$\bar{\theta}_0$	$\bar{\theta}$
α	100.00	5.00	5.50	7.50	13.25
β_1	100.00	5.50	6.50	7.00	11.25
β_2	8.25	6.25	6.25	5.25	9.00
β_3	18.00	5.00	4.75	7.00	11.25

C) Power of Tests

Test based on $\bar{Z}$: 86 %
Test based on Z : 90 %

Table 3
Results of Experiment 2
($N = 2{,}000$; $R^2 = 0.4$; $\lambda = 0.1$)

A) Biases

	$\hat{\theta}$	θ_0^*	θ^*	$\bar{\theta}_0$	$\bar{\theta}$
α	0.335	-0.004	-0.005	0.005	-0.001
β_1	-0.094	-0.001	-0.000	0.000	0.000
β_2	-0.018	-0.006	-0.006	0.008	0.003
β_3	0.039	0.007	0.007	0.007	0.006
Average of absolute values	0.121	0.004	0.005	0.005	0.003

B) Root Mean-Squared Errors

	$\hat{\theta}$	θ_0^*	θ^*	$\bar{\theta}_0$	$\bar{\theta}$
α	0.369	0.269	0.268	0.262	0.283
β_1	0.099	0.055	0.054	0.053	0.057
β_2	0.116	0.223	0.225	0.217	0.225
β_3	0.089	0.142	0.143	0.138	0.148
Average	0.168	0.172	0.173	0.168	0.178

C) Size of Type I Errors, in %

	$\hat{\theta}$	θ_0^*	θ^*	$\bar{\theta}_0$	$\bar{\theta}$
α	59.75	7.00	7.00	8.25	11.75
β_1	86.50	5.75	5.75	7.75	11.00
β_2	5.25	6.50	6.50	7.50	10.50
β_3	7.25	4.25	5.50	4.75	9.75

D) Power of Tests

Test based on $\bar{Z}$: 49.25 %
Test based on Z : 52.50 %

Table 4
Results of Experiment 3
($N = 700$; $R^2 = 0.4$; $\lambda = 0.30$)

A) Biases

	$\hat{\theta}$	θ_0^*	θ^*	$\bar{\theta}_0$	$\bar{\theta}$
α	0.847	0.011	-0.033	0.049	-0.052
β_1	-0.240	-0.008	0.006	-0.011	0.023
β_2	-0.067	-0.022	-0.010	-0.010	0.041
β_3	0.103	0.004	-0.006	0.010	-0.011
Average of absolute values	0.314	0.011	0.014	0.020	0.032

B) Root Mean-Squared Errors

	$\hat{\theta}$	θ_0^*	θ^*	$\bar{\theta}_0$	$\bar{\theta}$
α	0.884	0.585	0.633	0.536	0.636
β_1	0.246	0.151	0.170	0.136	0.172
β_2	0.210	0.440	0.457	0.428	0.497
β_3	0.177	0.221	0.237	0.214	0.257
Average	0.379	0.349	0.374	0.329	0.390

C) Size of Type I Error, in %

	$\hat{\theta}$	θ_0^*	θ^*	$\bar{\theta}_0$	$\bar{\theta}$
α	91.25	3.75	3.00	4.00	11.75
β_1	99.75	3.25	4.25	6.00	17.25
β_2	7.25	4.25	5.25	8.00	13.50
β_3	12.50	3.75	5.75	4.25	12.00

D) Power of Tests

Test based on $\bar{Z}$: 26.00 %
Test based on Z : 40.75 %

Table 4 presents the results of experiment 3. The purpose of this experiment was to verify whether our HM estimators could still outperform $\hat{\theta}$ with smaller sample sizes. One can see that even for N = 700, the HM estimators are still preferable to $\hat{\theta}$, especially in terms of the sizes of the type I errors for α and β_1. Note, however, that although they are not as catastrophically large as for $\hat{\theta}$, the size of the type I errors for the elements of the $\bar{\theta}$ vector of estimators are notably larger than expected.[13] Considering the results obtained in terms of root mean-squared errors in the three experiments presented above, one notices that the two estimators θ_0^* and $\bar{\theta}_0$ exhibit a slightly better performance. Since these two HM estimators are easier to compute, they therefore appear to be particularly recommendable. Note also that in all three experiments, the tests based on Z were more powerful than those based on $\bar{Z}$. This is particularly the case for experiment 3 in which the sample size was reduced.

A fourth experiment was made under the same conditions as experiment 1, but the value of β_3 was set equal to zero. Despite the facts that 1) in simple regression models asymptotic biases disappear when the coefficient is equal to zero, 2) X_3 itself did not contain errors of measurement but only X_1 was measured with errors and, finally, 3) X_1 and X_3 were weakly correlated ($r_{13} = 0.153$), the bias of the OLS estimator of β_3 was nonnegligeable (0.102) and the size of the type I error associated with the test that $\beta_3 = 0$ was rather large (24.75 %). This means that in almost 25 % of the cases, the t-test based on an intended 95 % confidence level would have led one to reject the true hypothesis that $\beta_3 = 0$. The other results pertaining to this experiment are similar to those of experiment one and are not reported here. If the correlation between X_1 and X_3 is raised to 0.5, as was done in the last experiment, the bias of the OLS estimator of β_3 increases to 0.366 and the size of the type I error reaches 99 %! The average RMSE of $\hat{\theta}$ also deteriorates much more than those of our HM estimators, as can be verified from Table 5.

D) Summary of experimental findings

The above Monte Carlo experiments, in combination with an extensive set of experiments made in the early phases of this research project on the performance of

13 Given the number of replications used for our Monte Carlo experiments, namely 400, the 95 % confidence interval for the intended 5 % type I errors would extend from 2.86 % to 7.14 %.

Table 5
Results of Experiment 5
(N = 2,000; $R^2 = 0.04$; $\lambda = 0.30$; $r_{13} = 0.5$; $\beta_3 = 0$)

A) Biases

	$\hat{\theta}$	θ_0^*	θ^*	$\bar{\theta}_0$	$\bar{\theta}$
α	1.882	0.077	0.042	0.107	0.013
β_1	-0.286	-0.007	-0.002	-0.011	0.004
β_2	-0.082	0.016	0.019	0.017	0.024
β_3	0.366	0.009	0.002	0.015	-0.008
Average of absolute values	0.654	0.027	0.016	0.037	0.012

B) Root Mean-Squared Errors

	$\hat{\theta}$	θ_0^*	θ^*	$\bar{\theta}_0$	$\bar{\theta}$
α	1.901	0.815	0.841	0.734	0.793
β_1	0.287	0.112	0.116	0.097	0.106
β_2	0.133	0.202	0.203	0.196	0.202
β_3	0.375	0.189	0.193	0.180	0.191
Average	0.674	0.329	0.338	0.302	0.323

C) Size of Type I Error, in %

	$\hat{\theta}$	θ_0^*	θ^*	$\bar{\theta}_0$	$\bar{\theta}$
α	100.00	5.25	5.25	8.00	11.50
β_1	100.00	6.00	6.25	5.25	9.00
β_2	11.75	5.75	6.25	5.00	9.00
β_3	99.00	4.00	4.25	5.75	8.00

D) Power of Tests

Test based on $\bar{Z}$: 69.00 %
Test based on Z : 84.50 %

estimators similar to θ^* [Dagenais and Dagenais (1993)] in which errors of measurement were assumed to affect only one of the explanatory variables (say $\bar{X}_1$), suggest the following general conclusions.[14]

Bias

1) The value of the squared multiple correlation coefficient seems to have no effect on biases.

2) The sample size has no effect on the biases of the elements of the OLS estimator.

3) The biases of the elements of our HM estimators are notably smaller than those of the corresponding elements of the OLS estimator, in small samples.

4) As the sample size grows, the biases of the elements of our HM estimators vanish progressively.

5) The biases of the elements of the OLS estimators increase with λ_1.

6) In small samples, the biases of the elements of the HM estimators are larger for greater values of λ_1. Furthermore, when λ_1 is larger, these biases do not vanish as rapidly, when the sample size grows.

7) When the independent variables are highly collinear, the bias of the OLS estimator of the parameter affecting a variable measured with error may be larger than it would be in the simple regression case. The OLS estimators of the coefficients of the correlated variables may also be strongly biased.

8) The size of the small-sample biases of our HM estimators do not seem to be much affected by the collinearity among the explanatory variables.

[14] In the following paragraphs, we use expressions such as : "small" samples and "small" values of λ or R^2. Although it is difficult to be very precise in such matters, we would say that "small" samples refer roughly to samples smaller than 500 observations. "Small" values of λ or R^2 indicate values of λ smaller than, say, 0.05 and values of R^2 lower than 0.25. In contrast, "large" samples are samples of more than 1,000 observations, "large" values of λ are values greater than 0.25 and "large" R^2s are R^2s greater than 0.75.

Root mean-squared error

1) For small values of N, the RMSE's of the elements of the OLS estimators decrease as R^2 or N increase. For larger values of N, these RMSE's remain almost constant. This is easily explained by the fact that MSE equals squared bias plus variance and that the variance is $O(N^{-1})$ while the bias is O(1). When N gets large, the MSE is essentially equal to the squared bias; hence, the factors affecting the variance no longer have an impact on the MSE.

2) For a given sample size, the RMSE's of the elements of the OLS and HM estimators increase with λ_1.

3) The RMSE's of the elements of the HM estimators are also strongly influenced by the value of R^2 and N. The RMSE's decrease as R^2 or N increase. These results are clearly explained by the fact that the HM estimators have relatively small biases and, hence, the MSE's are merely influenced by the variances. Therefore, the MSE's reduce as the variances reduce.

4) For small values of λ_1, R^2 or N, the OLS estimator may outperform the HM estimators.

5) The HM estimators may outperform the OLS estimator for much smaller sample sizes when collinearity is high.

Size of type I error

1) The relative performance of our HM estimators is always superior to that of the OLS estimator, when there are errors in the variables.

2) In all cases examined, the importance of the type I error of the Student t-tests associated with our HM estimators was always relatively close to the desired 5 % level.

3) The performance of the OLS estimator deteriorates as λ_1, R^2 or N increase. It is very disappointing even, for example, for values of λ_1 as low as 5 %. As discussed earlier, this is explained by the fact that OLS estimators are biased but have relatively small variances.

4) For given values of λ_1, R^2 and N, the sizes of the type I errors of the OLS estimators increase when the data are more collinear.

EV tests

1) The power of the tests increases with λ_1, R^2 and N.

2) The tests have very little power for small samples, unless R^2 and λ_1 are large.

3) When R^2 is low, the tests do not have much power for λ_1 smaller than 10 %, even in very large samples. Even for larger values of λ_1, the power remains fairly low in large samples, when R^2 is small.

4) The performance of the tests improves significantly when R^2 increases.

5) The EV tests are more powerful when the other explanatory variables are strongly correlated with the variables affected by measurement errors.

6. AN ILLUSTRATIVE APPLICATION

As illustrated above, our suggested HM estimators are likely to perform better than the OLS estimator in microeconomic analyses based on survey data where the sample comprises several hundred observations, even if the measurement errors are relatively small. Where analyses are based on smaller samples, only in situations involving more important measurement errors will our HM estimators exhibit a superior performance, in terms of root mean-squared errors. This is likely to be the case, however, in macroeconomic applications, since errors of measurement are known to be important in aggregate data [Morgenstern (1961); Dagenais (1992)]. This is illustrated below by applying our EV tests to the data used by Mankiw, Romer and Weil (1992) to analyze economic growth. Mankiw, Romer and Weil (MRW) estimated a human capital augmented Solow model and tested it with macroeconomic data of 98 countries, using OLS estimators. With the data shown in the appendix of the MRW paper, we have accurately reproduced in our Table 6, the results appearing in the upper part of Table 2 of MRW. Table 6 also gives the p-values of our joint F-tests of errors in the variables. Both versions of the test yield very low p-values. Given that these tests

Table 6.
Human Capital Augmented Solow Model [Mankiw, Romer and Weil (1992)]
Dependent Variable : ℓn GDP Per Capita in 1985.
Observations : Cross Section of 98 Countries

Estimator	$\hat{\theta}$	θ_0^*	θ^*	θ_0	θ	EV Tests Using $\tilde{Z}$. t-Statistics	EV Tests Using Z. t-Statistics
Constant	6.8478 (1.1774)*	3.2946 (1.6466)	1.7803 (1.7951)	5.3805 (1.3571)	5.8337 (1.0751)		
ℓn(I/GDP) Investment/GDP	0.6966 (0.1328)	0.7750 (0.2431)	0.8290 (0.2380)	0.8430 (0.2104)	0.8952 (0.1527)	-0.5339	-1.6787
ℓn(n + g + δ) Growth rates of labor and capital plus depreciation rate	-1.7438 (0.4159)	-3.0535 (0.5732)	-3.6306 (0.6314)	-2.3657 (0.4572)	-2.2571 (0.3605)	3.2778	2.5060
ℓn(school) Percentage of working age population in secondary school	0.6545 (0.0727)	0.5795 (0.1043)	0.5485 (0.1041)	0.6244 (0.0979)	0.6169 (0.0764)	0.9498	-1.4063
t-statistic for zero sum hypothesis	-0.8600	-2.6068	-3.1894	-1.6547	-1.7469		
Joint EV tests on all variables						p-value 0.0092	p-value 0.0017

Standard errors are in parentheses.

do not appear to be very powerful in small samples unless the errors of measurement are very large, there is a very strong presumption that the data used by MRW contain errors of measurement. Student t-tests applied separately to the coefficients associated with each of the three variables suggest that the variable $\ell n(n + g + \delta)$ may be particularly error-ridden. MRW note also that the Student t-statistics based on their OLS estimates *strongly support* the prediction of the augmented model to the effect that the coefficients of the three variables sum to zero. The same statistics based on our more robust HM estimators are, in contrast, notably larger in absolute value and rather indicate that the null hypothesis should be rejected.

In final analysis, the very clear indications of the presence of errors in the variables supplied by our EV tests, together with the results obtained concerning the sum of the coefficients when using the suggested HM estimators, casts very strong doubts on MRW's claim that their data support the human capital augmented Solow model.

REFERENCES

Aigner, D.J., G. Hsiao, A. Kapteyn and T. Wansbeek (1984), "Latent Variable Models in Econometrics", in Z. Griliches and M.D. Intriligator (eds.), *Handbook of Econometrics* 2, Chapter 23, 1321-1393.

Altonji, J.G. and A. Siow (1987), "Testing the Response of Consumption to Income Changes with (Noisy) Parcel Data", *The Quarterly Journal of Economics*, May, 293-328.

Avery, R.B. (1991), "Household Saving in the U.S.", *Review of Income and Wealth* 37(4), 409-432.

Avery, R.B., G.E. Elliehausen and A.B. Kennickell (1988), "Measuring Wealth with Survey Data : An Evaluation of the 1983 Survey of Consumer Finances", *Review of Income and Wealth* 34(4), 339-369.

Bowden, R.J. (1984), *Instrumental Variables*, Cambridge University Press, Cambridge, 277 pages.

Dagenais, M.G. (1992), "Measuring Personal Savings, Consumption and Disposable Income in Canada", *Canadian Journal of Economics* XXV(3), August, 681-707.

Dagenais, M.G. (1994), "Parameter Estimation in Regression Models with Errors in the Variables and Autocorrelated Disturbances", *Journal of Econometrics*, forthcoming.

Dagenais, M.G. and D.L. Dagenais (1993), "Estimation and Testing in Regression Models with Errors in the Variables", Discussion Paper No. IEA-93-03, Institut d'économie appliquée, École des Hautes Études Commerciales, Montréal, February, 39 pages.

Drion, E.F. (1951), "Estimation of the Parameters of a Straight Line and of the Variance of the Variables, if They Are Both Subject to Error", *Indagationes Mathematicae* 13, 256-260.

Duncan, G.J. and D.H. Hill (1985), "An Investigation of the Extent and Consequences of Measurement Errors in Labor-Economic Survey Data", *Journal of Labor Economics* 3(4), 508-532.

Durbin, J. (1954), "Errors in Variables", *International Statistical Review* 22, 23-32.

Fomby, T.B. and D.K. Guilkey (1978), "On Choosing the Optimal Level of Significance for the Durbin-Watson Test and the Bayesian Alternative", *Journal of Econometrics* 8, 203-214.

Fuller, W.A. (1987), *Measurement Error Models*, Wiley, New York, 440 pages.

Geary, R.C. (1942), "Inherent Relations Between Random Variables", *Proc. Royal Irish Academy* 47, 63-76.

Goldberger, A.S. (1972), "Structural Equation Methods in the Social Sciences", *Econometrica* 40(6), 979-1002.

Gouriéroux, C. and A. Montfort (1989), *Statistique et modèles économétriques*, Volume 1, Economica, 565 pages.

Grether, D.M. and G.S. Maddala (1973), "Errors in Variables and Serially Correlated Disturbances in Distributed Lag Models", *Econometrica* 41, 255-262.

Griliches, Z. (1986), "Economic Data Issues", in Z. Griliches and M.D. Intriligator (eds.), *Handbook of Econometrics* 3, Chapter 25, 1465-1514.

Griliches, Z. and J.A. Hausman (1986), "Errors in Variables in Panel Data", *Journal of Econometrics* 31, 93-118.

Hansen, L.P. (1982), "Large Sample Properties of Generalized Method of Moments Estimators", *Econometrica* 50(4), July, 1029-1054.

Hausman, J.D. (1978), "Specification Tests in Econometrics", *Econometrica* 46(6), November, 1251-1271.

Jeong, J. and G.S. Maddala (1991), "Measurement Errors and Tests for Rationality", *Journal of Business and Economic Statistics* 9(4), October, 431-439.

Kendall, M.G. and A. Stuart (1963), *The Advanced Theory of Statistics*, Second edition, Charles Griffin and Company Limited, volume 1.

Klepper, S. and E.E. Leamer (1984), "Consistent Sets of Estimates for Regressions with Errors in All Variables", *Econometrica* 52(1), 163-184.

Langaskens, Y. and M. Van Rickeghem (1974), "A New Method to Estimate Measurement Errors in National Income Account Statistics : The Belgian Case", *International Statistical Review* 42(3), 283-290.

Malinvaud, E. (1978), *Méthodes Statistiques de l'Économétrie*, 3e édition, Dunod.

Mankiw, N.G., D. Romer and D.N. Weil (1992), "A Contribution to the Empirics of Economic Growth", *The Quarterly Journal of Economics*, May, 407-437.

Mariano, R.S. and T. Sawa (1972), "The Exact Finite-Sample Distribution of the Limited-Information Maximum Likelihood Estimator in the Case of Two Included Endogenous Variables", *Journal of the American Statistical Association*, March, 159-163.

Morgenstern, O. (1963), *On the Accuracy of Economic Observations*, Second edition, Princeton University Press, Princeton, N.J.

Nelson, C.R. (1990), "Some Further Results on the Exact Small Sample Properties of the Instrumental Variable Estimator", *Econometrica* 58(4), July, 967-976.

Pal, M. (1980), "Consistent Moment Estimators of Regression Coefficients in the Presence of Errors in Variables", *Journal of Econometrics* 14(3), December, 349-364.

Radner, D.B. (1982), *Social Security Bulletin* 45(7), July, 13-21.

Reiersol, O. (1950), "Identifiability of a Linear Relation Between Variables which are Subject to Error", *Econometrica* 18, 375-389.

Rodgers, W.L., C. Brown and G.J. Duncan (1993), "Errors in Survey Reports of Earnings, Hours Worked and Hourly Wages", *Journal of the American Statistical Association* 88(424), 1208-1218.

Statistics Canada (1988), "Family Expenditure in Canada, 1986", Survey of Family Expenditure, Public Use Microdata file.

White, H. (1982), "Maximum Likelihood Specification of Misspecified Models", *Econometrica*, 1-25.