GENERATED VARIABLES IN FINANCE: SPECIFICATION, ESTIMATION AND INFERENCE

C.R. McKenzie, Osaka School of International Public Policy, Osaka University
Michael McAleer, Department of Economics, University of Western Australia

Abstract It is now widely recognized in the econometrics literature that the use of generated variables in regression models leads to a measurement error problem, thereby causing the equation error to be serially correlated and heteroscedastic in both small and large samples. The generated variable problem generally leads to a loss of efficiency in estimation. More importantly, the presence of generated variables yields invalid inferences because the formula used to obtain the standard errors is an inconsistent estimate of the true standard errors. In the finance literature, generated variables typically arise in calculating measures of risk and variability, in event studies based on residuals from an estimated market model, and the use of betas estimated from market models as regressors in another model. In addition to the standard generated variable problem associated with inefficiency and invalid inferences, the method of construction of risk and variability measures can also render the ordinary least squares estimates inconsistent. This paper surveys the use of generated variables in finance through a critical examination of papers published in the *Journal of Finance* in 1995. It is found that only one published empirical paper makes an attempt at correcting for the presence of generated variables.

1. INTRODUCTION

When a regression equation (hereafter called the structural equation) contains an unobserved variable, either as the dependent variable or an explanatory variable, one commonly used procedure is to first obtain an estimate of the unobserved variable in some way (when a regression equation is used, this is called the auxiliary equation), and then to estimate the original regression replacing the unobserved variable by its estimate. To highlight the fact that both the dependent and explanatory variables in the structural equation can be generated, this paper refers to the estimate of the unobserved variable as a 'generated variable' rather than the term 'generated regressor' commonly used in the literature. There are three issues that arise for parameter estimates obtained when these generated variables are used at the second stage: consistency; efficiency; and inference. When consistent estimates are obtained at the first stage, the presence of generated variables at the second stage does not affect the consistency of the second stage estimates but generally leads to a loss of efficiency in estimation, and invalid inferences because the formula used to obtain the standard errors is an inconsistent estimate of the true standard errors (see, for example, Pagan (1984, 1986), Newey and McFadden (1994) and McKenzie and McAleer (1997)).

This paper examines the issue of generated variables in the finance literature where generated variables often arise in calculating measures of risk and volatility; in event studies where the dependent variable is obtained as a residual from an estimated market model; and the use of betas estimated from a market model as regressors in another model. One characteristic of quite a few finance papers is that the components of the generated variables are estimated from many different time series models, and the generated variables are then used in a regression model estimated using cross section data. This method of estimation raises some issues not previously considered in the literature.

Section 2 contains details of a survey of generated variable papers published in the *Journal of Finance* in 1995 while section 3 discusses some theoretical econometric issues that arise with the estimation methods and the generated variables used in these papers. A brief conclusion is contained in section 4.

2. SURVEY RESULTS

Table 1 provides a summary of the ten papers appearing in the *Journal of Finance* in 1995 that contain generated variables. In addition, there were several papers that used estimates of unobserved volatilities as variables but the specific model used to generate the volatilities (often a Black–Scholes (1973)–style option pricing model) meant that the estimates of volatility could be obtained as a simple function of observed variables and, therefore, contained no unobserved parameters (see Campa and Chang (1995), Chen *et al.* (1995) and Litzenberger and Rabinowitz (1995)). These papers were excluded from the study. There was one theoretical paper (Kim (1995)) that discussed some of the problems associated with using estimated firm betas as explanatory variables in a regression model from an errors in variables perspective. The reason for choosing empirical papers in finance was that the non–systematic search of the literature reported in McKenzie and McAleer (1997) suggested that generated variables continue to be used quite widely in finance without the standard errors being adjusted appropriately. The *Journal of Finance* was chosen as the subject of the survey as it is the leading journal in the finance literature. The year 1995 was chosen arbitrarily, but it is nevertheless representative.

Several common features of the papers in Table 1 are worth noting: (i) the components of the generated variables are quite often estimated from many different time series models, and the generated variables are then used in a regression model estimated using cross section data (see, for example, Bajaj and Vijh (1995)); (ii) when the auxiliary and structural equations are both estimated using time series data, the sample size at each stage can differ dramatically (see, for example, Ilmanen (1995)); and (iii) generated variables often relate to measures of risk, volatility, a firm's beta and excess returns (see, for example, Billett *et al.* (1995)). Only one of the papers reported in Table 1, Bekaert and Harvey (1995), makes explicit reference to the generated variable problem and Pagan's (1984) classic paper. In Bekaert and Harvey (1995), "to mitigate the generated variable problem" (p. 435) heteroskedasticity–consistent standard errors computed à la White (1980) are reported. An examination of the way the generated regressor was derived suggests that the measurement error will be serially correlated as well. Monte Carlo evidence for the cases reported in Smith and McAleer (1993) suggests that using Newey and West's (1987) heteroscedasticity and autocorrelation consistent covariance matrix to correct for the generated variable problem leads to standard errors that are generally downwardly and, quite often significantly, biased.

3. THEORETICAL ISSUES

This section is devoted to a detailed analysis of several issues raised in the paper by Billett *et al.* (1995) which contains four different variables that are generated from time series models and are then used in a cross section model. For expository purposes, the number of explanatory variables in the structural and auxiliary equations is kept to a bare minimum.

Suppose the structural equation of interest is

$$y_i = a_0 + a_1 X_{1i} + a_2 X_{2i} + \epsilon_i,\ i=1,..,N, \tag{1}$$

where the a_j are unknown parameters, ϵ_i is assumed to be distributed as niid(0, σ^2_ϵ), and in cross sectional analyses i is assumed to refer to the ith firm. Equation (1) is assumed to be correctly specified. The auxiliary equation is assumed to be of the form

$$z_{it} = b_{0i} + b_{1i} W_{1it} + \eta_{it},\ i=1,..N,\ t=1,..,T, \tag{2}$$

where the b_{ji} are unknown parameters, and η_{it} is assumed to be distributed as niid(0, $\sigma^2_{\eta i}$). Equation (2) is also assumed to be correctly specified. Given the structure of (2), ordinary least squares (OLS) can be applied to (2) for each firm i to give unbiased, $\sqrt{T}$ consistent and efficient estimates of the parameters [efficiency in this context only concerns the estimation of (2)]. Denote the OLS estimates of b_{0i}, b_{1i} and $\sigma^2_{\eta i}$ by β_{0i}, β_{1i} and s^2_i, and the predictions (or fitted values) and prediction errors from this model as $\hat{z}_{it} = \beta_{0i} + \beta_{1i}W_{1it}$ and $u_{it} = z_{it} - \hat{z}_{it}$, respectively.

In Billett *et al.* (1995), equation (2) is a market model with z_{it} being the rate of return on firm i's equity at time t, and W_{1it} being the rate of return on the market portfolio at time t (which is not dependent on i). Since this paper is an event study suppose that the event occurs at $T+\tau$, then the excess return (generated variable (a)) is $u_{iT+\tau}$ and the cumulative abnormal return is $C_i = u_{iT+\tau-10} + \dots + u_{iT+\tau-1}$ (generated variable (c)). In Billett *et al.* (1995), τ=50. The parameter estimate β_{1i} is firm i's beta (generated variable (b)), and s_i is the estimate of the standard deviation of the borrower's stock return (generated variable (d)). Implicit in an event study is the assumption that the event has some effect. Thus, at $t=T+\tau$, equation (2) becomes

$$z_{iT+\tau} = \delta_i + b_{0i} + b_{1i}W_{1iT+\tau} + \eta_{iT+\tau},\ i=1,..N, \tag{3}$$

so that δ_i measures the impact of the event on firm i's return. In this case, the excess return, $u_{iT+\tau}$, is given by

$$u_{iT+\tau} = \delta_i + (b_{0i} - \beta_{0i}) + (b_{1i} - \beta_{1i})W_{1iT+\tau} + \eta_{iT+\tau}, \tag{4}$$

so that $u_{iT+\tau}$ is implicitly an estimate of δ_i. It should be noted that $E(u_{it+\tau}) = \delta_i$, and $E(\beta_{1i}) = b_{1i}$ but $E(s_i) = \sigma_{\eta i}[2/(T-2)]^{1/2}\Gamma[(T-1)/2]/\Gamma[(T-2)/2] \neq \sigma_{\eta i}$ (see Kawada *et al.* (1980, p. 200) and Johnson *et al.* (1994, section 8.2)). If (2) applies for $t=T+1,..,T+\tau-1$, then $E(C_i) = 0$.

Suppose that X_{2i} is the unobserved variable in (1) and it is replaced by $\hat{X}_{2i}$ to give

$$y_i = a_0 + a_1X_{1i} + a_2\hat{X}_{2i} + \epsilon_i + a_2(X_{2i} - \hat{X}_{2i}),\ i=1,..,N. \tag{5}$$

Define $h_i = \epsilon_i + a_2(X_{2i} - \hat{X}_{2i})$. In these circumstances, whether $E(h_i) = 0$ depends on whether $E(\hat{X}_{2i}) = X_{2i}$. When $\hat{X}_{2i}$ is $u_{iT+\tau}$, C_i or β_i, even though ϵ_i is homoscedastic, h_i will be heteroscedastic because $V(h_i)$ depends on $\sigma^2_{\eta i}$. However, unlike the standard generated variable problem, h_i will not be serially correlated since the errors in (2) are assumed to be independently distributed! This means a White (1980)–style adjustment to the standard errors will be an appropriate remedy for the inconsistency of the OLS standard errors, unlike the cases dealt with in the literature to date (see Pagan (1984, 1986), Newey and McFadden (1994) and McKenzie and McAleer (1997)).

In contrast to the standard generated variable problem, even when $\hat{X}_{2i}$ is an unbiased estimate of X_{2i} there is no guarantee that

$$\Sigma^N_{i=1}\hat{X}_{2i}(X_{2i} - \hat{X}_{2i})/N$$

will converge to zero as $N \to \infty$. Without an additional assumption about how T behaves as $N \to \infty$, we have a standard errors in variables problem, so that OLS applied to (5) is inconsistent. For example, if T is fixed as $N \to \infty$, when $\hat{X}_{2i}$ is a parameter it is not a consistent estimate of X_{2i} or when $\hat{X}_{2i}$ depends on estimated parameters they are not $\sqrt{N}$ consistent estimates of the underlying parameters. If appropriate instruments are available, this inconsistency problem can be overcome by using the instrumental variable technique (see Pagan (1984, Model 6) and McKenzie and McAleer (1997, section 3.5) for a related example, and Kim (1995) for another example). Finding an appropriate instrument for C_i and β_{1i} may be difficult given the assumptions of the model but resort could be made to Durbin's (1954) suggestion to use the rank of the problem variable as an instrument. If $N \to \infty$ and $T \to \infty$, then when $\hat{X}_{2i}$ is an parameter it is a consistent estimate of X_{2i} and when $\hat{X}_{2i}$ depends on estimated parameters they are consistent estimates of the underlying parameters. Thus, estimates of the structural equation will also be consistent.

When $E(\hat{X}_{2i}) \neq X_{2i}$ there is an additional problem, namely that the following two expressions will not converge to zero if T is fixed as $N \to \infty$:

(a) $\Sigma_{i=1}^{N} h_i/N$, (b) $\Sigma_{i=1}^{N} X_{1i} h_i/N$.

Therefore, even instrumental variables estimation with appropriate instruments will not provide consistent estimates. Consistent estimation in this case requires T, $N \to \infty$ so that plim $s_i = \sigma_{\eta i}$ and the consistency of OLS applied to (5) is assured. If interest only focuses on testing the null hypothesis $a_2=0$ (or a joint hypothesis involving $a_2=0$), OLS applied to (5) will be consistent, asymptotically efficient and inference will be valid under the null, since the generated variable effect in the disturbance term disappears (see Pagan (1984) and McKenzie and McAleer (1997) for a discussion of this in the standard case).

However, if y_i is the unobserved variable in (1) and it is replaced by $\hat{y}_i$ to give

$$\hat{y}_i = a_0 + a_1 X_{1i} + a_2 X_{2i} + \epsilon_i + (\hat{y}_i - y_i), \; i=1,..,N, \tag{6}$$

then it is easily seen that the generated variable problem in (6) persists, regardless of the hypothesis being tested. This substitution does not lead to the errors in variables problem alluded to for a generated variable but if $E(\hat{y}_i - y_i) \neq 0$, then similar results as for the case when $E(\hat{X}_{2i}) \neq X_{2i}$ apply.

The previous discussion is predicated on the assumption that (2) was correctly specified so that applying OLS leads to $\sqrt{T}$ consistent parameter estimates. If (2) is misspecified, for example W_{1it} is not the appropriate regressor or important explanatory variables are omitted, then estimates of the parameters of (2) will be neither unbiased nor $\sqrt{T}$ consistent. The previous discussion suggests this inconsistency will also carry over to the estimation of (1) when the generated variable based on the inconsistent parameter estimates is used (see Pagan (1986) and McKenzie and McAleer (1994) for sufficient conditions for consistency of the estimates of the structural equation). This strong dependence of the consistency of the parameter estimates of (2) provides a powerful case for checking the correctness of the specifications of both (1) and (2).

Even when time series data have been used to estimate both the structural and auxiliary equations, there are a number of studies where different sample sizes have been used to estimate the models at the first and second stages (see, for example McAleer and McKenzie (1991) and the papers they cite). When the errors in the structural and auxiliary equations are uncorrelated, McKenzie and McAleer (1990, section 3.2) show that compared to the case of using an identical number of observations for both the first and second stages, using fewer observations at the first stage will lead to larger variances at the second stage. However, if the difference in the two sample sizes disappears asymptotically, using a different number of observations at the first and second stages will be asymptotically negligible.

An examination of those papers in Table 1 that use time series data at both the first and second stages suggests large differences in the sample sizes used at each stage (see, for example, Ilmanen (1995) and Jorion (1995)). Even this is a little misleading since in Ilmanen (1995), for example, data for the structural equation are *monthly* while for the US bond beta (generated variable (b)) *daily* data over the preceding month are used so that all the data that are ever going to be available are used (assuming different betas for each month). This suggests similar problems will arise to those discussed when cross section data are used for the structural equation and time series data for the auxiliary equation. The local bond beta (generated variable (c)) for each month is computed using *monthly* data for the preceding twelve months. Unlike the standard generated variable problem discussed in Pagan (1984) and McKenzie and McAleer (1997), where the substitution of the generated variable for the unobserved variable causes the structural equation's error at each point to be correlated with every other error, the way in which the local bond variable is calculated using overlapping data will generate a moving average of order eleven in the error of the structural equation!

4. CONCLUSION

This paper has considered some theoretical problems that arise with the type of generated variables commonly used in the finance literature. It remains to be demonstrated that these problems with generated variables are important in practice. However, papers where both unadjusted and adjusted standard errors are presented suggest that the difference can be sufficient to cause the use of unadjusted standard errors to lead to incorrect inferences (see the papers discussed in McKenzie and McAleer (1997, section 1)). As indicated in section 3, further considerations are: whether instrumental variable estimation is required, and the effect of using alternative estimation methods; and what sample sizes are used to estimate the auxiliary and structural equations, and the impact of different sample sizes on the standard errors of the second stage estimates.

5. ACKNOWLEDGEMENTS

The authors wish to thank Kimio Morimune and Kosuke Oya for their helpful suggestions. The first author wishes to acknowledge the financial support of the Asset Management Service Industry Fund, Osaka School of International Public Policy, Osaka University, and the second author wishes to acknowledge the financial support of the Australian Research Council.

6. REFERENCES

Bajaj, M. and A.M. Vijh, Trading behavior and the unbiasedness of the market reaction to dividend announcements, *J. Finance*, 50(1), 255–279, 1995.

Bekaert, G. and C.R. Harvey, Time–varying world market integration, *J. Finance*, 50(2), 403–444, 1995.

Billett, M.T., M.J. Flannery, and J.A. Garfinkel, The effect of lender identity on a borrowing firm's equity return, *J. Finance*, 50(2), 699–718, 1995.

Black, F. and M. Scholes, The pricing of options and corporate liabilities, *J. Political Economy*, 81, 637–654, 1973.

Campa, J.M. and P.H.K. Chang, Testing the expectations hypothesis on the term structure of volatilities in foreign exchange options, *J. Finance*, 50(2), 529–547, 1995.

Chen, N.F., C.J. Cuny, and R.A. Haugen, Stock volatility and the levels of the basis and open

interest in futures contracts, *J. Finance*, 50(1), 281–300, 1995.
Dimson, E. and P. Marsh, Capital requirements for securities firms, *J. Finance*, 50(3), 821–851, 1995.
Durbin, J., Errors in variables, *Review of the International Statistical Institute*, 22, 23–32, 1954.
Hamilton, J.D., A new approach to the economic analysis of nonstationary time series and the business cycle, *Econometrica*, 57(2), 357–384, 1989.
Hansen, L., Large sample properties of the generalized method of moments estimators, *Econometrica*, 50, 1029–1054, 1982.
Ilmanen, A. Time–varying expected returns in international bond markets, *J. Finance*, 50(2), 481–506, 1995.
Johnson, N.L., S. Kotz and, N. Balakrishnan, *Continuous Univariate Distributions, Volume 1*, John Wiley & Sons, New York, second edition, 1994
Jorion, P., Predicting volatility in the foreign exchange market, *J. Finance*, 50(2), 507–528, 1995.
Kaplan, S.N. and R.S. Ruback, The valuation of cash flow forecasts: An empirical analysis, *J. Finance*, 50(4), 1059–1093, 1995.
Kawada, Y., F. Maruyama, S. Nabeya and T. Yoshino, *Daigaku Enshu Suri Tokei (Mathematical Statistics for University Seminars)*, Shokabo, Tokyo, 1980.
Kim, D., The errors in the variables problem in the cross–section of expected stock returns, *J. Finance*, 50(5), 1605–1634, 1995.
Kothari, S.P., J. Shanken, and R.G. Sloan, Another look at the cross–section of expected stock returns, *J. Finance*, 50(1), 185–224, 1995.
Litzenberger, R.H. and N. Rabinowitz, Backwardation in the oil futures markets: Theory and empirical evidence, *J. Finance*, 50(5), 1517–1545, 1995.
May, D.O., Do managerial motives influence firm risk reduction strategies?, *J. Finance*, 50(4), 1291–1308, 1995.
Mayhew, S., A. Sarin, and K. Shastri, The allocation of informed trading across related markets: An analysis of the impact of changes in equity–option margin requirements, *J. Finance*, 50(5), 1635–1653, 1995.
McAleer, M. and C.R. McKenzie, Keynesian and new Classical models of unemployment revisited, *Economic Journal*, 101, 359–381, 1991.
McKenzie, C.R. and M. McAleer, On efficient estimation and correct inference in models with generated regressors: A general approach, Working Paper in Economics and Econometrics No. 211, Australian National University, 1990.
McKenzie, C.R. and M. McAleer, On the effects of misspecification errors in models with generated regressors, *Oxford Bulletin of Economics and Statistics*, 56(4), 441–455, 1994.
McKenzie, C.R. and M. McAleer, On efficient estimation and correct inference in models with generated regressors: A general approach, forthcoming *Japanese Economic Review*, 48(4), 1997.
Newey, W.K. and D. McFadden, Large sample estimation and hypothesis testing, in *Handbook of Econometrics, Volume IV*, edited by R.F. Engle and D.L. McFadden, pp. 2111–2245, Elsevier Science, Amsterdam, 1994.
Newey, W.K. and K.D. West, A simple, positive semi–definite, heteroskedasticity and autocorrelation consistent covariance matrix, *Econometrica*, 55(3), 703–708, 1987.
Pagan, A., Econometric issues in the analysis of regressions with generated regressors, *International Economic Review*, 25(1), 221–247, 1984.
Pagan, A., Two stage and related estimators and their applications, *Review of Economic Studies*, 53(4), 517–538, 1986.
Smith, J. and M. McAleer, On the robustness of Barro's new classical unemployment model, *Applied Economics*, 25, 349–360, 1993.
White, H., A heteroscedasticity–consistent covariance matrix estimator and a direct test for heteroscedasticity, *Econometrica*, 48, 817–838, 1980.

Table 1: Summary of Journal of Finance Generated Variable Papers

Paper	Structural Equation	Generated Variable(s)
Bajaj–Vijh	Average excess return XS, N= 10, 100, OLS, WLS	Excess return predicted from market model, TS, T=240, OLS, DV
Bekaert–Harvey	Exchange rate TS, T=?, OLS+WHITE	Degree of integration estimated from Hamilton (1989) transition probability model, TS, T=?, X
Billett *et al.*	Abnormal returns XS, N=578, 346, OLS	(a) Excess returns predicted from market model, TS, T=150, OLS, DV (b) Firm's beta from firm's market model, TS, T=150, OLS, X (c) Cumulative abnormal return from market model TS, T=150, OLS, X (d) Standard deviation of borrower's stock return from residual of market model, TS, T=150, OLS, X
Dimson–Marsh	Book volatility XS, N=58, OLS	(a) book volatility estimated as standard deviation of returns on net book, TS, T=60, OLS, DV (b) Position risk requirement estimated as a non–linear function of (i) variance of market returns, (ii) variance of average residual returns for market TS, T=?, DV
Ilmanen	Excess government bond returns TS, T=186, OLS+Newey–West	(a) local bond beta obtained from market model for local bond TS, T=12, OLS, X (b) US bond beta obtained from market model for US bond TS, T≃20, OLS, X
Jorion	Predict volatilities TS, T≃1500, OLS+Hansen–White	(a) Realized volatility computed as moving average of squared returns, TS, T=? DV (b) Volatility estimated from GARCH (1,1) model, TS, T≃1500, X (c) Realized volatility computed as moving average of squared returns, TS, T=20, X

Paper	Structural Equation	Generated Variable(s)
Kaplan–Ruback	Market value of leveraged transactions, XS, N=51, OLS	Cash flow forecast (X) computed using (i) industry based measure of systemic risk as industry equity beta from industry market model, TS, T=480, OLS (ii) firm based measure of systemic risk as industry equity beta from firm market model, TS, T=480, OLS
	Implied risk premium XS, N=51, OLS	(a) industry based measure of systemic risk as industry equity beta from industry market model, TS, T=480, OLS, X (b) firm based measure of systemic risk as industry equity beta from firm market model, TS, T=480, OLS, X
Kothari *et al.*	Portfolio returns XS, N=20, 100, OLS	Portfolio beta estimated from a market model TS, T≃640?, OLS, X
May	Equity return variance XS, N=200, OLS	Equity return variance (DV) modeled as (i) Sample variance of equity returns, TS, T=60 (ii) Variance of residuals from a market model of firm's equity returns, TS, T=60
Mayhew *et al.*	Spread and Depth in Options Market XS, N= 113, 119, OLS+White	Volatility estimated as daily return variance, ?

Notes:
Abbreviations used are as follows: XS = cross section estimate or regression based on N observations; TS = cross section estimate or regression based on T observations; OLS = ordinary least squares estimation; White = White's (1980) heteroscedasticity consistent estimate of the covariance matrix; Newey–West = Newey–West (1987) heteroscedasticity and serial correlation consistent estimate of covariance matrix; Hansen–White = Hansen's (1982) heteroscedasticity and serial correlation consistent estimate of covariance matrix; DV = generated dependent variable; X = generated regressor; ? indicates that the sample size is unclear from the the text.