WORKING PAPER
No. 16. October 1992

SAMPLE BASED PROPORTIONS AS VALUES ON AN INDEPENDENT VARIABLE IN A REGRESSION MODEL

Bo Jonsson

Svensk resumé på sid 34.

Uppsala University
Department of Statistics
P.O. Box 513
S-751 20 Uppsala
Sweden

September 1992

SAMPLE BASED PROPORTIONS AS VALUES ON AN INDEPENDENT VARIABLE IN A REGRESSION MODEL

by

Bo Jonsson
Department of Statistics
Uppsala University

This research reported has been supported by the Swedish Council for Research in the Humanities and Social Sciences (HSFR) and by the National Institute of Economic Research (KI).
Helpful comments from A. Ågren, A Christoffersson and L-E Öller are gratefully acknowledged

ABSTRACT

Proportions estimated from surveys are sometimes used as values of an independent variable in time series regression models. This paper examines the effect of the sample errors in such an independent variable when estimating the slope parameter in a simple linear regression model. It is found that the bias for OLS can be large, especially when the independent variable is on a low level. Some consistent moment estimators have been evaluated in a simulation study. This has been done for three independent variables on different mean levels. For the two variables on the lowest levels these estimators are found to be better or much better than OLS while this mainly was not the case for the third variable. The results indicate that other estimators than OLS should be considered when dealing with independent variables based on survey data.

1. INTRODUCTION

Econometric time series models for macro-economic variables sometimes include independent variables that are based on sample estimates. One example of this is when household survey data are used as explanatory variables in consumption or investment functions. Another example is in connection with what is known as "tests of rationality". Then it is customary to start with a test of so called unbiasedness by estimating the equation $y_t = \alpha + \beta\, y^*_t + \varepsilon_t$, where y_t is inflation and y^*_t expectations on inflation measured through surveys, and testing for the hypothesis $\alpha=0$, $\beta=1$. In both examples sampling error will be source of error in the independent variable (see Jonsson & Ågren, 1991, and Jeong & Maddala, 1991).

The predictive ability of household survey data as concerns economic development and buying plans, has been evaluated in several studies (for a brief survey see Ågren, 1989). The survey variable is sometimes the only explanatory variable, sometimes it appears together with ordinary economic variables. In all the studies the problem of sampling errors is neglected. The problem was pointed out in Jonsson and Ågren (1991). Discussing car expenditures they argue that the effect of sampling errors on the results is not always negligible, at least not for values close to zero of the explanatory variables. They use proportional variables such as the proportion of households that believe their financial situation will improve during the next year (attitude) or the proportion of households that are 100% sure of buying a new car within six months (plan). They found these two variables to perform relatively well (using OLS) in comparison with other indices based on the Swedish household surveys. However, the explanatory power of the plan variable was found to be sensitive to the sample size of the surveys. This was less

pronounced for the attitude variable, probably due to the fact that the plan variable is on a very low level, while the attitude variable is not. The selection of index used for explaining car expenditures will therefore depend on the sample size in the household survey.

When estimating population proportions the sampling variances will not be constant because the proportions vary over time. Another reason for varying error variances is unequal sample sizes. In the Swedish household survey, carried out by Statistics Sweden, the sample size has undergone several changes since the start of the survey in 1973. It is well known that errors in the independent variables will cause inconsistent estimates if OLS is used. Fuller (1987) gives a comprehensive survey of methods for how to correct for this kind of asymptotic bias. The survey includes general methods for handling a situation with unequal error variances for fixed regressors, when the error variances are known. Such methods are of interest here since standard errors for the measurements on the independent variable are sometimes published, sometimes they can easily be estimated under the assumption of simple random sampling.

The aim of this study is to investigate the effect of sampling errors on the OLS-estimator of the slope parameter in a regression model when the population proportion at time point t, x_t, of an event is fixed and the observed number of events is binomially distributed. We also use simulation to evaluate some consistent estimators. This will be done for three levels of x. We only study the case of one explanatory variable.

To solve the problem with errors in variables, Jeong and Maddala use a FIML-estimator based on multiple sources of expectations to correct for the bias. If such information were available this method could of course also be used on household survey data. This is seldom the case but multiple sources of information can be obtained by splitting every survey into two parts and by constructing one index for each part. This will be an easy way to handle the errors in variables problem. In a final section, a two indicator model obtained by splitting the surveys into two parts will be briefly analyzed. This will be done under the binomial assumption, but the method has a much broader applicability.

2. THE ERRORS IN VARIABLES MODEL WHEN THE MEASUREMENTS ON THE EXPLANATORY VARIABLE ARE SAMPLE PROPORTIONS

2.1 The model

The classical errors in variables model with fixed x_t can be presented in the following way:

$$\text{Structural equation:} \quad y_t = \alpha + \beta x_t + q_t \quad \text{where } q_t \text{ is } IN(0,\sigma^2_q).\ t=1,..T. \tag{1}$$

$$\text{Measurement equations:} \quad Y_t = y_t + w_t$$

$$X_t = x_t + u_t,$$

where w_t is $IN(0,\sigma^2_w)$ and u_t is $IN(0,\sigma^2_u)$ and w_t, q_t and u_t are assumed to be independent. IN is an abbreviation for "independently and normally distributed". T denotes the number of observations on a sequence $\{Y_t,X_t\}$.

It is well known that OLS regression of Y_t on X_t will provide an inconsistent estimate b_1 of β (cf. Fuller, 1987). It is the measurement error in x_t that is critical. The probability limit of b_1 is:

$$\text{plim } b_1 = K\beta, \tag{2}$$

where K is the so called reliability ratio and is equal to:

$$K=\sigma^2_x/(\sigma^2_x+\sigma^2_u), \text{ where } \sigma^2_x=\frac{1}{T}\Sigma(x_t-\mu_x)^2 \text{ and } \mu_x=\frac{1}{T}\Sigma x_t,$$

and where

$$b_1=S_{XY}/S_{XX}=\Sigma(X_t-\bar{X})(Y_t-\bar{Y})/\Sigma(X_t-\bar{X})^2 .$$

The population coefficient of determination between the observed variables, R^2_{YX}, is also affected by the size of the measurement error variances and it is easily shown that this coefficient is K times the coefficient of determination for the regression of Y_t on x_t, R^2_{Yx}. When the error variance in X_t is known, a consistent estimator of β can be obtained by the method of moments (See Johnston, 1963, pp. 159-160). The estimator is

$$b = S_{XY}/(S_{XX} - (T-1)\sigma^2_u). \quad (3)$$

In the applications concerning household surveys, mentioned in the introduction, the x values are often population proportions. If the measurement X_t is based on a simple random sample of N_t elements and if the number of elements in the population can be assumed to be large, compared to the sample size, then N_tX_t is approximately binomially distributed and the error variance in X_t, σ^2_{ut}, is equal to $x_t(1-x_t)/N_t$. The measurement error, if any, in the dependent variable is not of interest and will therefore just be added to the error in the equation ($\varepsilon_t=q_t+w_t$). The estimator corresponding to (3) now becomes (cf. Appendix A):

$$b_2 = S_{XY}/(S_{XX} - (1-1/T)\Sigma\sigma^2_{ut}), \text{ where } \sigma^2_{ut}=x_t(1-x_t)/N_t. \quad (4)$$

Replacing x_t by X_t and N_t by N_t-1 provides an unbiased estimator of the error variance in x_t. This estimator of $\Sigma\sigma^2_{ut}$ will have a small standard error if it is based on several surveys, where in each the sample size (N_t) is large (cf. Appendix B).

2.2 The size of the asymptotic bias

We have seen that the size of the asymptotic bias, when using OLS to estimate the slope parameter, is related to the reliability ratio K. Transforming K to match the case of unequal error variances gives:

$$K = \sigma^2_x/[\sigma^2_x+\Sigma x_t(1-x_t)/NT]=N\sigma^2_x/[N\sigma^2_x+\Sigma x_t/T - \Sigma x^2_t/T]= \quad (5)$$

$$=N\sigma^2_x/[(N-1)\sigma^2_x+\mu_x(1-\mu_x)]$$

Here we assume N_tX_t to be binomially distributed and $N_t = N$ for all t. As can be noted, K is decreasing, for given variance in x_t, when the mean level of x_t is getting closer to 0.5. Often it is more realistic to keep the coefficient of variation ($CV=\sigma_x/\mu_x$) constant. Then

$$K=\frac{N(CV)^2}{(N-1)(CV)^2+\frac{1-\mu_x}{\mu_x}}. \qquad (6)$$

The size of K depends on the number of observations (N), the coefficient of variation (CV) and the mean level of x_t (μ_x). Given the two first factors and assuming that $\mu_x<.5$ it is obvious that K is very sensitive to the level of x_t and that it is decreasing, meaning increasing bias, when μ_x decreases. The effect accelerates when μ_x is getting close to zero.

Assume that the x_t values are equally spaced between the minimum and the maximum value of x_t . The squared coefficient of variation can then be shown to be

$$(CV)^2=\frac{T+1}{3(T-1)}(1-\frac{\min(x)}{\mu_x})^2. \qquad (7)$$

Inserting (7) into (6) and replacing N-1 by N in the denominator of (6) gives

$$K=\frac{N^*/3}{\frac{N^*}{3}+\frac{1-\mu_x}{\mu_x}\frac{1}{(1-\frac{\min(x)}{\mu_x})^2}}, \qquad \text{where } N^*=N(T+1)/(T-1). \qquad (8)$$

In order to illustrate the size of the asymptotic bias according to (8), K has been plotted against μ_x for T=50 and N in Figure 1A, where the varying minimum of x is assumed to be half the size of the mean. The figure clearly points out the sensitivity of the bias to the level of the variable. Even when the sample size is as large as 10000 the bias may not be negligible when μ_x is smaller than, say 2%. It can also be seen that the sampling error causes considerably biased estimates for much higher proportions if the sample size is not large enough. The right-hand figure shows how K depends on the variability in x_t measured by the ratio between $\min(x_t)$ and μ_x. The less this variability is the less is K. Even with a sample size of 2400, serious bias occurs for almost all proportions if the variation in x_t is small.

A situation often arising in practice is that a choice has to be made between x-variables on different levels. If the criterion for choosing variable is based on R^2_{YX}, the choice will

FIGURE 1. K as a function of μ_x for different N (A) and different $\min(x)/\mu_x$ (B). T=50.

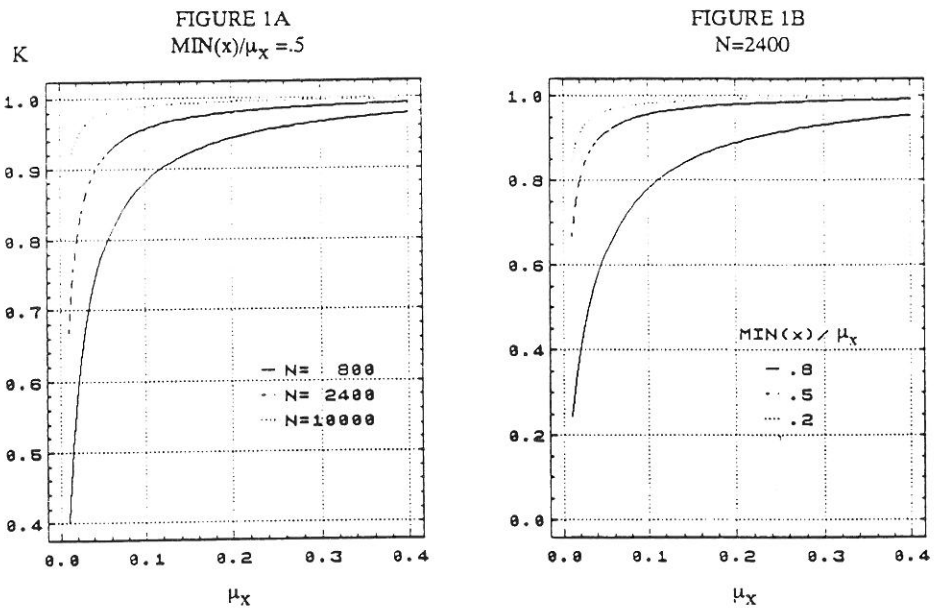
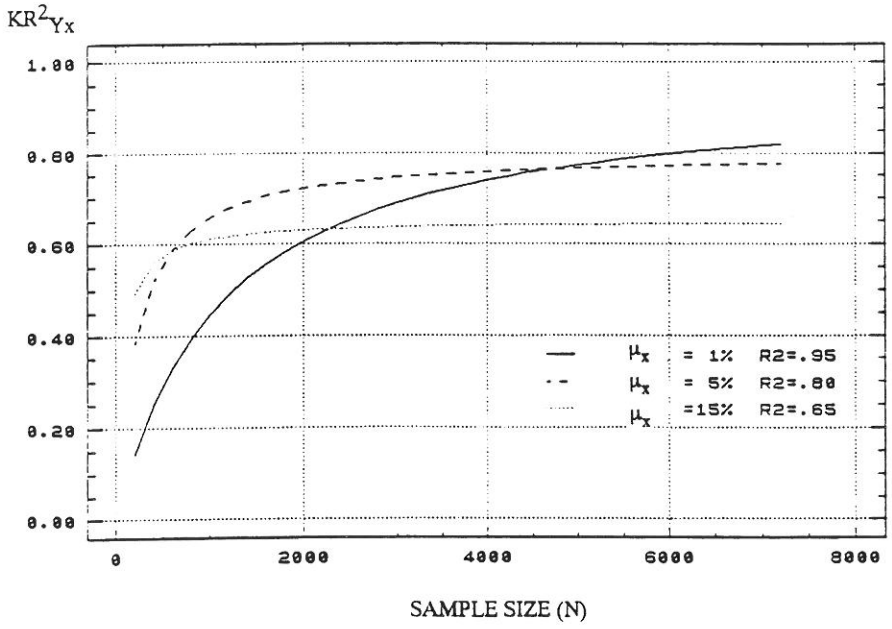


FIGURE 2. KR^2_{YX} against N for different levels of μ_x assuming $\min(x)/\mu_x$ to be .5.



depend on the sample size. This is illustrated in Figure 2. Three x-variables on different mean levels (1%, 5% and 15%) are assumed to have R^2_{Yx}= .95, .80 and .65. The variable with a mean of 1% would be the best one according to R^2 if we know the true x. In Figure 2, $R^2_{YX} = KR^2_{Yx}$ has then been plotted against the sample size (N) using (8) for the calculation of K with $\min(x)/\mu_x$=0.5. This last assumption means that the variable with the lowest mean is rectangularly distributed between .5% and 1.5% and the variable with the highest mean between 10% and 20%. We note that the 1% level variable is to be preferred if N is greater than, say 4800. When N is less than 600 the variable with the highest mean is chosen, while the 5% level variable is "best" when the sample size is between 600 and 4800.

2.3 Estimation of the slope parameter

In (4) a moment estimator b_2 of the slope parameter was given that corrected for the inconsistency caused by the error in the independent variable. One problem with this estimator is that in a sequence of observations one can get a small or even a negative denominator, because σ^2_x might be underestimated due to correlation between x_t and u_t. Fuller (1987, pp. 193) proposes the following estimator in the case of unequal but known error variances:

$$b_3 = \frac{S_{XY}}{S_{XX} - (1-\delta/T)\Sigma\sigma^2_{ut}} = \frac{S_{XY}}{S_{XX} + \frac{\delta}{T}\Sigma\sigma^2_{ut} - \Sigma\sigma^2_{ut}} \qquad \text{if } \lambda^* = \lambda - 1/T \geq 1, \qquad (9a)$$

and

$$b_3 = \frac{S_{XY}}{S_{XX} + \frac{\delta}{T}\Sigma\sigma^2_{ut} - (\lambda - 1/T)\Sigma\sigma^2_{ut}} \qquad \text{if } \lambda^* = \lambda - 1/T < 1, \qquad (9b)$$

where $\lambda = S_{XX}(1-R^2_{XY})/\Sigma\sigma^2_{ut}$.

In (9), δ is a positive constant. If δ=1, (9a) is identical to estimator b_2 in (4). Fuller makes some comments on the choice of δ in the case of normally distributed x_t and errors with constant variance (Fuller 1987, pp. 249). For the case where x is fixed but the

error variance varies no comments are made upon the choice of δ. In this study $\delta=1$ will be used. In (9), λ is the ratio between the residual sum of squares when taking the reverse regression of X_t on Y_t and the sum of error variances in X_t. This ratio is expected to be greater than or equal to one since the error variance in X_t is a part of this residual sum of squares. However, in a sequence of observations, it happens that $\lambda<1$. This is a motivation for the modification of (9a) made in (9b), also excluding the possibility of a negative denominator. Inserting the expression for λ into (9b) gives

$$b = S_{XY}/(S_{XX}R^2_{XY}+2\Sigma\sigma^2_{ut}/T)=(b_{XY} + 2\Sigma\sigma^2_{ut}/TS_{XY})^{-1}, \qquad (10)$$

where the denominator is non-negative (if $S_{XX}R^2_{XY}$ or $\Sigma\sigma^2_{ut}>0$). Note that the inverse of the OLS-estimator, when taking the regression of X on Y (b_{XY}), is approximately the same as (10) if $\Sigma\sigma^2_{ut}/T$ is small compared to $S_{XX}R^2_{XY}$. This inverse is often proposed as an upper limit for β when x is measured with error (cf. Maddala, 1988, pp 381-382).

If the error equations are inserted into the structural model (1) we get the model

$$Y_t = \alpha + \beta X_t - \beta u_t + \varepsilon_t \text{ , where } \varepsilon_t = w_t + q_t. \qquad (11)$$

The error term $v_t=-\beta u_t+\varepsilon_t$ will be heteroscedastic since σ^2_{ut} varies with x_t. Another source of variation in the error variance can be unequal sample sizes (N_t). Fuller (1987, pp. 192) gives as an alternative estimator to (9) a generalized least squares estimator, for which the observations in (9) are weighted by $\hat{t}=1/s^2_{vt}$:

$$b_4 = \frac{S_{\hat{t}XY}}{S_{\hat{t}XX} + \frac{\delta}{T}\Sigma\hat{t}\sigma^2_{ut} - \lambda^*_{\pi}\Sigma\hat{t}\sigma^2_{ut}}, \qquad \lambda^*_{\pi}= 1 \quad \text{if} \quad \lambda_{\pi}-1/T\geq 1 \qquad (12)$$

$$\lambda^*_{\pi}=\lambda_{\pi}-1/T \quad \text{if} \quad \lambda_{\pi}-1/T<1,$$

and

where $\lambda_{\pi}=\dfrac{S_{\hat{t}XX} - (S_{\hat{t}XY})^2/S_{\hat{t}YY}}{\Sigma\hat{t}\sigma^2_{ut}}$, $S_{\hat{t}XY}=\Sigma\hat{t}(X_t-\bar{X})(Y_t-\bar{Y})$, $S_{\hat{t}XX}=\Sigma\hat{t}(X_t-\bar{X})^2$

$S_{\hat{t}YY}=\Sigma\hat{t}(Y_t-\bar{Y})^2$ and

where $\hat{t}=1/s^2_{vt}$ with $s^2_{vt}=s^2_{\varepsilon}+(b_3)^2\sigma^2_{ut}$, $s^2_{\varepsilon}=\Sigma(Y_t-a-b_3X_t)^2/(T-2) - (b_3)^2\Sigma\sigma^2_{ut}/T$ and $a=\Sigma Y_t/T-b_3\Sigma X_t/T$.

s^2_{vt} is obtained by applying (9) in a first step. If λ is less than one the estimator s^2_ε of σ^2_ε is put to zero. According to Fuller (1987, pp 195) the estimator b_4 of β is expected to be superior to (9) in almost all practical situations.

Fuller also gives asymptotic estimators of the variances for (9) and of the weighted estimator (12) in the case of normal errors and for (9) in the case when no assumption of normal errors is made. In this study the measurements on x are assumed to be obtained from a binomial distribution but as the sample sizes (N_t) in the applications are large the errors can be assumed to be approximately normally distributed unless x_t is small. The formulas are presented in Appendix C.

2.4 Classification of a sequence of observations according to a chi-square test and to λ.

A trivial assumption of regression analyses is that $\sigma^2_X>0$. However, the error in X_t can more or less conceal the true variation in x_t, but the hypothesis $\sigma^2_X>0$ can be tested against $\sigma^2_X=0$. We can simply apply an ordinary chi-square test of equal population proportions with T-1 degrees of freedom. One reasonable strategy will then be to continue the analysis of only those sequences that produce significant chi-square tests. The chi-square statistic is the usual one: $\Sigma\Sigma(O_t-E_t)^2/E_t$, where O_t is the observed and E_t the expected frequency. After some reformulations the statistic becomes:

$$\chi^2=\frac{\Sigma N_t(X_t-\bar{X}^*)^2}{\bar{X}^*(1-\bar{X}^*)} \quad \text{where } \bar{X}^* = \Sigma N_t X_t/\Sigma N_t. \tag{13}$$

For equal N_t the chi-square statistic is

$$\chi^2=NS_{XX}/\bar{X}(1-\bar{X}),$$

which can also be written as

$$N(T-1)(CV_S)^2\bar{X}/(1-\bar{X}),$$

where CV_S is the coefficient of variation of X_t. It is obvious that the mean level of X_t will (for a given coefficient of variation) affect the "power" of the test, which will

decrease with decreasing mean. An estimator k of K is given in Appendix A. If $\Sigma\sigma^2_{ut}$ in this estimator is replaced by $\Sigma X_t(1-X_t)/(N-1)$ we obtain (14):

$$k=\frac{S_{XX}-(1-1/T)\Sigma X_t(1-X_t)/(N-1)}{S_{XX}}=\frac{(N-1/T)S_{XX}-(T-1)\bar{X}(1-\bar{X})}{(N-1)S_{XX}}\ . \qquad (14)$$

Inserting $\chi^2=NS_{XX}/\bar{X}(1-\bar{X})$ into (14) yields:

$$k=\frac{N-1/T}{N-1}-\frac{N}{N-1}\frac{1}{\chi^2/df}\cong 1-\frac{1}{\chi^2/df}\ . \qquad (15)$$

From (15) it is clear that the proposed strategy to only accept a sequence of observations if the chi-square is significant will guarantee positive k:s so the denominator in (4) will stay positive.

The modification based on λ^* made in (9b) also guarantees a positive denominator in the estimator of β. The relation between λ^* and the χ^2-statistic is:

$$\lambda^*=\frac{(N-1)S_{XX}(1-R^2_{YX})}{T\bar{X}(1-\bar{X})\ \Sigma X_t(1-X_t)/\Sigma X_t(1-\bar{X})}-\frac{1}{T}=\frac{N-1}{N}\frac{\chi^2}{df+1}(1-R^2_{YX})\frac{\Sigma X_t(1-\bar{X})}{\Sigma X_t(1-X_t)}-\frac{1}{T}\ . \qquad (16a)$$

If X_t is small $\Sigma X_t(1-X_t)$ will approximately be equal to $\Sigma X_t(1-\bar{X})$. λ^* is then approximately equal to:

$$\lambda^*\cong\frac{\chi^2}{df+1}(1-R^2_{YX})-\frac{1}{T}\ . \qquad (16b)$$

Hence λ^* is a function of the chi-square statistic and R^2_{YX}. When dealing with proportional variables on a low level and when R^2_{YX} is close to zero we note that λ^* is approximately equal to the chi-square statistic divided by its degrees of freedom and increasing with increasing values of the chi-square statistic. The critical value for the λ^* correction in (9) is $\lambda^*=1$. If the inequality $\lambda^* > 1$ is inserted into (16b) this expression becomes

$$\chi^2/(df+2) > 1/(1-R^2_{YX}) \qquad (17)$$

In order to get an insignificant chi-square test statistic for $\lambda^* \geq 1$, R^2_{YX} must be small. From (16) it is seen that it is possible to get $\lambda^* < 1$ even if the chi-square statistic is significant. This holds for low values of the proportion variable if the chi-square statistic is near the critical value of the test and R^2_{YX} is not too low.

Since the values on the chi-square statistic and λ^* can be obtained from a sequence of observations, these sequences can be classified according to the significance of χ^2 and the values of λ^*:

Class 1. Not significant,
Class 2. Significant, $\lambda^* < 1$,
and
Class3. Significant, $\lambda^* \geq 1$

It is then possible to evaluate how the estimators of β work in these classes separately. This will be done next.

3. THE SIMULATION STUDY

3.1 Design

The model to be used in the simulation study to generate Y_t is

$$Y_t = \alpha + \beta x_t + \varepsilon_t \, , \; t=1,.,T.$$

x_t is a fixed proportion, which is estimated on the basis of N_t independent observations. Thus, instead of x_t we observe $X_t = x_t + u_t$ where $N_t X_t$ is binomially distributed and the errors u_t are independent with mean zero and variance $x_t(1-x_t)/N_t$. The variance is estimated by replacing x_t by X_t and N_t by N_t-1. ε_t is assumed to be $IN(0,\sigma^2)$ and independent of u_t.

The simulations will be performed for three x-variables, denoted x1, x2 and x3. The fixed values on these variables have been chosen to approximately correspond to

variables in the Swedish household survey and are on different levels. The first one, x1, represents the proportion of households that expresses a likelihood of 100% that they will buy a new car within six months. The third variable, x3, corresponds to the proportion of households who believe that their financial situation is going to improve within the next year. For these two variables the observed sample proportions, during the period 1976:3 to 1989:2, have been smoothed, and are denoted $\bar{X}_t$. A sequence of T=48 x_t values have then been equally distributed between the minimum and maximum of $\bar{X}_t$. The second variable, x2, has been given the same coefficient of variation as x3 but x2 is on a lower level. The rectangular distribution enables us to get the same distribution for Y_t in the three cases. The mean and the variance of the x-variables are given in Table 1. We note that the variables are on different levels, x1 being on an extremely low level. We also note that this variable has a larger coefficient of variation than the other two. For each x-variable the simulations of Y_t are performed for $R^2_{Yx}=1-\sigma^2_\varepsilon/\sigma^2_y$ set to .65, .80 and .95. The reason for using different R^2 is that the value on λ^* depends on the size of R^2 (cf. (16)). Table 2 shows the parameter values that have to be assumed in order to assure R^2 values of desired magnitude. The simulations of X_t have been performed using sample sizes (N_t) of 800, 1200, 2400, 3600 and 4800 households at each of the T=48 quarters.

TABLE 1. Description of the three x-variables.
CV=Coefficient of variation.

	$100*\min(x)$	$100*\max(x)$	$100\mu_x$	$10000*\sigma^2_x$	$(CV)^2$
x1	.39	1.21	.800	.05842	.0913
x2	3.42	7.38	5.400	1.3617	.0467
x3	10.80	23.30	17.050	13.575	.0467

TABLE 2. Description of the parameters in the three models to be used in the simulation.
($R^2_{Yx} = 1 - \sigma^2_\varepsilon/\sigma^2_Y$)

			$R^2_{Yx}=0.65$		$R^2_{Yx}=0.80$		$R^2_{Yx}=0.95$	
	E(Y)	Var(Y)	α	β	α	β	α	β
x1	10.00	1.000	7.331	333.57	7.040	370.06	6.774	403.26
x2	10.00	1.000	6.269	69.09	5.861	76.65	5.490	83.53
x3	10.00	1.000	6.269	21.88	5.861	24.28	5.490	26.45

The effect of unequal sample sizes is also examined by assuming a sample size of 3600 at the "first and third quarters" each year and a sample size of 1200 "the second and fourth quarters". This approximately corresponds to the actual design in the Swedish household survey. Some investigations have also been done for x1 with N_t=7200 and N_t=10800. The number of replications is set to 2000.

3.2 The classification of the sequences of observations according to the chi-square and the λ-criteria

In Table 3 the outcome of the simulation study on the three classes mentioned in the previous paragraph are presented. Insignificant chi-squares at the 5%-level can only be found for x1 and occurs for the two smallest sample sizes. This is in line with the earlier conclusion that the "power" of the test, for given coefficient of variation, depends on the level of the x variable. When the sample size is 800 nearly 20% of the sequences are recommended to be dropped (cf. 2.4).

The risk of getting $\lambda^* < 1$ is, as expected, increasing with increasing R^2 and is largest for the variable with the smallest mean (x1). For the two smallest sample sizes and for the design with mixed sample sizes we note high frequencies for x1 in the class $\lambda^*<1$, irrespective of R^2. Low frequencies in that class are mainly to be found for x1 when the sample size is large (2400 elements or larger) and R^2 is set to .65 or .80. When R^2 is high (.95). we note for both x1 and x2 and almost all sample sizes relatively high frequencies of $\lambda^*<1$. Some 10 % of the sequences fall into Class 2 when the variable x2 is used with a sample size of 800 and when R^2 is set to .80. For all other combinations of x2 and R^2 $\leq$ .80 the risk is small or very small to get a $\lambda^* <1$. For the variable x3 we get no sequences with $\lambda^*<1$ when R^2=0.65 or .80. When R^2 =.95 we note a high frequency only for the sample size 800. The risk of getting a low λ^* obviously depends on R^2_{Yx} and on the level of the variable, as pointed out in Section 2. R^2_{Yx} is of course unknown, but sometimes it is possible to have some opinion on the size of the explanatory power. And if this is not very high we note from the table that the risk of getting a $\lambda^*<1$ is small for the variables x2 and x3 and, if the sample sizes are large, also for x1.

TABLE 3. Percentage significant chi-square test statistic at the 5% level and percentage of the significant sequences for which $\lambda^* < 1$ for each combination of variable, sample size and R^2. The results are based on 2000 replications for each combination.

	Sample-size	% Sign chi-squares	$R^2_{Yx}=0.65$	$R^2_{Yx}=0.80$	$R^2_{Yx}=0.95$
x1	800	80.95	16.0	29.2	46.4
	1200	95.55	13.7	28.0	49.1
	2400	100	3.1	14.6	48.3
	3600	100	.3	3.4	30.6
	4800	100	.1	1.3	21.9
	7200	100	.0	.1	9.7
	10800	100	.0	0	5.0
	1200/3600	100	7.6	20.8	46.5
x2	800	100	1.4	9.6	45.6
	1200	100	.0	2.2	35.0
	2400	100	.0	.1	21.1
	3600	100	.0	.0	9.2
	4800	100	.0	.0	2.4
	1200/3600	100	.0	.5	26.4
x3	800	100	.0	.0	11.5
	1200	100	.0	.0	3.9
	2400	100	.0	.0	.2
	3600	100	.0	.0	.0
	4800	100	.0	.0	.0
	1200/3600	100	.0	.0	1.2

3.3 Evaluation of the estimators of β

The estimators to be evaluated are the OLS-estimator (b_1), the moment estimator (b_2, cf. (4)), the moment estimator modified according to Fuller (b_3, cf. (9)) and the corresponding weighted estimator (b_4, cf. (12)). If $\lambda^* \geq 1$, so is b_2 equal to b_3. For the sequences of observations in each cell in Table 3 the mean of the relative bias and the root mean square error (RMSE) have been calculated. The results are presented for x1, x2 and x3 in Table D1A, D1B and D1C in Appendix D. No results are given in the appendix in cases where the number of sequences in a cell is less than 50. The results for nonsignificant chi-squares are given in Table D1D.

Starting with the results for $\lambda^* \geq 1$ we note for x1 that OLS, as expected, produces a large negative bias ranging from 23% to 68%, depending on the sample sizes (N_t) used. The moment estimators $b_2=b_3$ and b_4 are always superior to the OLS-estimator, both in bias and in RMSE. Yet, they have non-negligable bias for the smallest sample-sizes, especially when R^2_{Yx} is high (.95). This is related to the risk of getting a sequence in Class 2. The weighted moment estimator (b_4) is often to be preferred to the unweighted one (b_3). However, the gain is substantial only for the case when mixed sample sizes are used and when R^2_{Yx} is high, for the cases based on relatively large samples. Even for the variable x2 (Table D1B) the OLS estimator cannot compete with the other estimators. The biases for the OLS-estimator are much larger and RMSE is always larger. RMSE for the weighted estimator is, with one exception, smaller than those for the unweighted one. However, the differences are negligible, except for the cases with mixed sample sizes. The results for the third variable, x3, also show a larger bias for the OLS-estimator but the bias is less than or close to 5%, except for the two smallest sample sizes. RMSE is in favour of the other estimators only when the sample size is small or when R^2_{Yx} is high. The weighted estimator (b_4) gives almost identical results as the unweighted one (b_3).

When $\lambda^*<1$ the moment estimator b_2 and the modified moment estimator b_3 are not identical. The OLS-estimator underestimates β but not to the same extent as it does for the cases when $\lambda^* \geq 1$. The other estimators in general overestimate β, as expected, and overestimation is often large for x1 and can be large for x2 when small sample sizes are used. For all variables the moment estimator b_2 is always worse than the modified moment estimator (b_3) and the weighted modified moment estimator (b_4), which means that the modification made in λ^* improves the estimator. In all the cases where the estimated risk of getting $\lambda^*<1$ is greater than 15% (M=300) both b_3 and b_4 have smaller bias and RMSE than the OLS estimator, while when there is a relatively small risk of such a λ^*, the OLS-estimator is the best. But, as mentioned before, that risk is unknown.

Looking at the results for Class 1, the cases of non-significant chi-square statistics, it can be noted from Table D1D that all evaluated estimators are generally poor. The decision to drop sequences where there is no evidence of variance in the true x, seems to be a wise one.

Finally, a short comment on the effect of increasing T will be made. A larger T will result in more sequences in Class 3. Furthermore the variance of the estimators will

decrease implying that the importance of the bias part of RMSE will increase. The gain in RMSE is therefore expected to be relatively larger for the consistent estimators than for OLS.

3.4 Evaluation of the estimators of the standard deviations

In Table D2 of Appendix D the standard deviations (s_b), obtained from the M replications of the unweighted modified moment estimator (b_3) and the standard deviations of the corresponding weighted estimator (b_4) are presented. Estimates, $\hat{\sigma}_b$, of the true standard deviations of b_3 and b_4 have been obtained for each sequence of observations using the variance estimators in Appendix C. The ratio between the mean of the estimates $\hat{\sigma}_b$ for b_3 and b_4 and the standard deviations (s_b) are also given in the table. For the variance estimators the results presented are based on the assumption of normal errors. The other estimator that did not require any distributional assumptions concerning the errors (cf. Appendix C) yields results that in all cases are considerably worse for x2 and x3 and in most cases for x1. The only occasions where this variance estimator is better than the one based on normal errors are for the variable x1 when the sample sizes are small (800,1200). However, then the estimator is only marginally better. This variance estimator will therefore not be considered in the following sections.

Starting as before with results for $\lambda^* \geq 1$, we note from Table D2 that the variance estimators work quite well in most cases for the variables x2 and x3. However, when R^2_{Yx} is high (.95) and the sample size is small the standard errors are overestimated by 13% to 35%. The results for the variable x1 are not satisfactory. For the two smallest sample sizes there is an overestimation of between 33% and 86%. When R^2_{Yx} is .95 the overestimation is for all sample sizes larger than 18%. The overestimation is small only when R^2_{Yx} is not so high and the sample size is large (3600 and 4800).

When we estimate the standard errors in Class 2 ($\lambda^*<1$) a question arises whether the variance estimators should be adjusted with regard to the value of λ^* in the same way as was done for b_3 and b_4. Both cases were simulated. The results show that if the adjustment is made the results are better in terms of less average overestimation. Therefore, results only for the adjusted estimated standard errors are presented. Looking at Table D2 we note that for x2 and x3 we get under- or overestimation which in most

cases is less than 10%. The exceptions occur primarily when a sample size of 800 is used. For x1 we often get huge overestimates. This is especially the case for the standard errors of the non-weighted estimator, b_3.

Using the unweighted and the weighted modified moment estimator (b_3 and b_4) and the obtained estimated standard errors, 95% confidence intervals have been calculated as $b \pm 1.96\hat{\sigma}_b$. The number of times these intervals cover, lie totally to the left of or totally to the right of the true β has been counted. The results for $\lambda^* \geq 1$ are presented in Table D3D. As can be noted, the empirical levels of confidence vary from 94% to 98% for the variables x2 and x3. The negative bias in the estimation of β of 5% to 6% for x2 800, when $R^2_{Yx} = .95$, is compensated for by the overestimation of the standard errors. For x1 the results in most cases show empirical confidence levels that lie between 93% and 96% when the sample sizes are between 2400 and 4800 and thus the intervals seem to work satisfactorily for those sample sizes. But this is often due to overestimation of the standard errors, balancing the negative bias. When the sample sizes are small (800 and 1200) the confidence levels vary between 71% and 93%. The lowest value occurs when the R^2_{Yx} is .95 and the sample size is 800 and the highest when R^2_{Yx} takes on the lowest value and the sample size is 1200. One also notes for the variables x1 and x2 that the frequencies of intervals that lie totally to the left of β often are higher, sometimes much higher, than the frequencies of intervals to the right of β. This pattern is most evident when $R^2 = .95$. In fact, for the variable x1 no interval then lies to the right of β.

The empirical level of confidence varies between 87% to 100% in Class 2 (cf. Table D4D). The lowest levels are to be found in cells with relatively few sequences. The highest levels are in some cases caused by huge overestimation of standard errors, mentioned above, leading to very wide intervals of doubtless value.

3.5 A comment on the Swedish household surveys

The x variables used in the simulation study were based on the Swedish household survey of consumer attitudes and buying plans. Historically the sample sizes have been large. Since 1984, a design with 1500 households in the first and third quarters and 4200 households in the second and fourth quarters have been used. The size is reduced by non-response. In the simulations, simple random sampling has been assumed while a stratified

sampling design is used in practice. The latter will probably lead to errors of a smaller size than for simple random sampling, but as we are dealing with proportions the gain with stratified sampling is probably not very large.

A R^2 as high as .95 is often most unlikely in the applications mentioned in the introduction. For the variables x2 and x3 the results for $\lambda^* \geq 1$ are therefore most applicable. Looking at sample sizes of at least 2400 households in each survey or the mixed sample case we have noted that the risk of $\lambda^*<1$ is very small. The moment estimators b_3 and b_4 are then quite satisfactory and are to be preferred to the OLS estimator for the variable x2. However, for x3 the OLS-estimator behaves almost equally well. The weighted estimator (b_4) seems to be somewhat better than or as good as the corresponding unweighted estimator (b_3), for both x2 and x3. The results for x1 are, despite a larger coefficient of variation, not as good as for the two other variables. The moment estimator could well lead to $\lambda^*<1$ and overestimated slopes and standard errors. If $\lambda^* \geq 1$ the results for these estimators are much better than for OLS but we should be careful when interpreting estimates obtained with the moment estimators because of a possible negative bias. In general, the weighted estimator b_4 is to be preferred to the unweighted alternatives. The mixed sample case in general produces poorer results than when a sample size of 2400 households is used. It is possible to show that the asymptotic bias for the mixed sample case approximately corresponds to a bias for a sample size of 1800 households at each time point. For our purposes the design with mixed sample sizes is not optimal.

4. A MODEL WITH TWO INDEPENDENT MEASUREMENTS ON x

Preserving the measurement equation for Y_t, assume that we split each survey into two equally sized subsamples and that:

$$X_{1t}=x_t+u_{1t} \text{ and } X_{2t}=x_t+u_{2t}, \text{ where } E(u_{1t})=E(u_{2t})=E(u_{1t}u_{2t})=0 \quad (18)$$
$$\text{and } Var(u_{1t}) = Var(u_{2t}) = 2\sigma^2_u.$$

X_t will then be the mean of these two measurements. Then, the second moments of the observed variables are:

$$\sigma^2_{X1}=\sigma^2_{X2} = \sigma^2_x + 2\sigma^2_u\,,\ \sigma^2_X = \sigma^2_x + \sigma^2_u\,,\ \sigma_{YX1}=\sigma_{YX2}=\sigma_{YX}=\beta\sigma^2_x \text{ and} \qquad (19)$$

$$\sigma_{X1X2} = \sigma^2_x = \sigma^2_X - \sigma^2_u\,, \text{ where } \sigma^2_u = \Sigma\sigma^2_{ut}/T.$$

Hence

$$\beta = \sigma_{YX}/\sigma_{X1X2} = \sigma_{YX}/(\sigma^2_X - \sigma^2_u) = \sigma_{YX}/(\sigma^2_X - \Sigma\sigma^2_{ut}/T).$$

From (19) one gets the moment estimator (4). The term $\Sigma\sigma^2_{ut}$ was estimated earlier under the assumption that N_tX_t is binomially distributed. Since $\sigma^2_{Xt} = \sigma^2_{ut}$ can be shown to be equal to $E_t(X_{1t}-X_{2t})^2/4$ we also have the possibility of estimating the sum of the error variances by $\Sigma(X_{1t}-X_{2t})^2/4$. No assumption about simple random sampling or that X_t is a proportion need to be added to the assumptions in (18) for this estimator to be appropriate. Of course, here too one can modify the moment estimator according to the value on λ^* using (9). The estimators corresponding to (4) and (9) obtained using $\Sigma(X_{1t}-X_{2t})^2/4$ as an estimator of the sum of the error variances will be denoted b_{22} and b_{23}, respectively. An alternative estimator to b_{22} is $b=S_{YX}/S_{X1X2}$, where $S_{X1X2}/(T-1)$ is the covariance between X1 and X2. However, S_{X1X2} calculated as $\Sigma(X_{1t}-\bar{X})(X_{2t}-\bar{X})$, can be shown to be equal to $S_{XX} - \Sigma(X_{1t}-X_{2t})^2/4$, a good approximation of b_{22}.

Although the estimators b_{22} and b_{23} have a much broader applicability they have been evaluated in the simulation study under the same premises as before, but for fixed sample size (2400) and R^2_{Yx}=.80. The weighted estimator will not be considered. The classification results according to the value of λ^* are given in Table 4. Calculation of λ^* using $\Sigma(X_{1t}-X_{2t})^2/4$ as an estimator of $\Sigma\sigma^2_{ut}$ does not produce the same result as under the binomial assumption in Section 3. We note that the frequencies of $\lambda^* <1$ are now higher compared to those obtained in Section 3. This is of course due to a higher sampling variability in the estimator of the sum of error variances. In Table 5 relative bias and RMSE are given for the moment estimator b_{22} and for the modified moment estimator b_{23}. For the sake of comparison results are also given for the corresponding estimators b_2 and b_3, evaluated in Section 3, and for the OLS-estimator (b_1). From the table we note that RMSE for b_{22} and b_{23}, as expected, is larger than for the corresponding estimators b_2 and b_3. However, the loss in precision is not large for the

variables x2 and x3. The most striking result for the variable x1 is the large effect of the modification made in the value of λ^*. The moment estimators are to be preferred to the OLS estimator for the variables x1 and x2. For the variable x3 the OLS estimator is about as good as the moment estimator.

TABLE 4. Classification into Classes 2 and 3 using λ^* according to Sections 3 and 4, respectively. The sample size is 2400 and R^2_{Yx}=.80. The number of replications is 2000.

	x1		x2		x3	
	Section 3	Section 4	Section 3	Section 4	Section 3	Section 4
$\lambda^*<1$	292	543	2	19	0	0
$\lambda^*\geq 1$	1708	1457	1998	1981	2000	2000

TABLE 5. Mean relative biases and RMSE of the estimators b_1,b_2, b_3, b_{22} and b_{23}. M is the number of sequences on which the calculation is based.

	x1		x1, $\lambda^*<1$		x1, $\lambda^*\geq 1$		x2		x3	
	M=2000		M=543		M=1457		M=2000		M=2000	
	BIAS %	RMSE	BIAS %	RMSE	BIAS %	RMSE	BIAS %	RMSE	BIAS %	RMSE
b_1	-35.1	133.2	-29.5	112.7	-37.2	140.1	-12.3	11.22	-3.2	1.96
b_2	4.0	78.4	17.0	100.8	-.9	68.1	1.5	7.87	.9	1.90
b_3	2.5	69.0	12.3	77.0	-1.2	65.7	1.5	7.87	.9	1.90
b_{22}	10.6	114.1	40.8	193.0	-.7	63.3	2.8	8.64	1.8	1.97
b_{23}	4.9	72.7	19.8	93.3	-.7	63.3	2.7	8.56	1.8	1.97

5. SUMMARY

This study investigates the effect of sample errors in an independent variable when estimating the slope parameter in a simple linear regression model. This is done under the assumption that the true independent variable consists of population proportions of an event and that the observed number of events is binomially distributed. This situation occurs when proportions are estimated from simple random sampling surveys, and used as explanatory variables in a time series regression model. The error variances can then easily be estimated at each time point. Consistent moment estimators can be found in the literature. Here some of these have been evaluated in a simulation study.

The results show that the asymptotic bias for the OLS-estimator depends on the coefficient of variation in the true x_t, the mean of x_t and the sample size (N_t) used to estimate x_t. For a given coefficient of variation the bias is very sensitive to the mean level of the variable when the level becomes small (or close to one).

The simulation study was performed for three variables (x1, x2 and x3) on different levels (0.7%, 5,4% and 17% on average), for various sample sizes and for three different values of the population coefficient of determination between the dependent variable and the true independent variable. The sequences of observations were classified into three classes. The first included sequences for which a chi-square test of no variability in the population proportions was insignificant. Such sequences are recommended to be dropped. Insignificant chi-square statistics could only be found for the variable on a very low level (x1) and for the smallest sample sizes (N_t=800 and N_t=1200 elements). The classification into two other classes was made according to the value of λ^*. If the residual sum of squares for the reverse regression for a sequence of observations is less than the sum of the error variances, λ^* will be less than one (approximately). A modification was then made in two of the studied estimators. The risk of getting $\lambda^* < 1$ was found to be strongly related to the level of the variable and to R^2. High frequencies for such sequences were mainly obtained for the variable on a very low level (x1), and when R^2 was very high for the middle sized variable (x2). The evaluated consistent estimators behaved well when $\lambda^* \geq 1$, except possibly for the variable x1. For that variable the estimators could have a large negative bias for small samples. For the variable on the highest level, x3, the OLS-estimator was about as good as the consistent estimators.

However, for the two other variables the consistent estimators were found to be better or much better than OLS. When $\lambda^*<1$ the results were improved if this was taken into account in the estimators. However the estimators sometimes had serious positive bias. Confidence intervals were also calculated. When $\lambda^*\geq 1$ confidence intervals generally performed well for x2 and x3 and, when the sample sizes were large, also for x1.

Using sample estimates as an independent variable leads to an error in variable situation. Other estimation methods than OLS should then be considered, in particular if the independent variable consists of proportions on a low level or if the variability in the true independent variable can be suspected to be low during the observation period. The error variances can easily be estimated if simple random sampling can be assumed or the standard errors are published. If this is not the case one can split the sample into two parts and then use a model with two independent measurements on x_t. Such a procedure has also been evaluated in a small simulation study producing results that were only slightly worse than those obtained under the binomial assumption for the variables x2 and x3.

REFERENCES

Ågren, A. (1989): A Survey of some Work on the Predictive Value of Attitude Data in Consumption and Saving Models. *Research Report 89-2, Department of Statistics*, Uppsala university, Uppsala.

Fuller, W.A. (1987): *Measurement Error Models*, New York: John Wiley & Sons.

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

Johnston, J. (1963): *Econometric Methods*, New York: McGraw Hill.

Jonsson, B. and Ågren, A (1991): Forecasting Car Expenditures using Household Survey Data. *Research Report 1991:10, Department of Statistics*, Uppsala university, Uppsala.

Maddala, G.S. (1988): *Introduction to Econometrics*, New York: Macmillan.

APPENDIX A.

Derivations of a moment estimator of β when the error variances are known and unequal.

First the observed X_t is replaced by x_t+u_t and Y_t by the structural equation in model (1), where the error term $\varepsilon_t=q_t+w_t$, yielding the following sums of squares and sum of cross products:

$$\Sigma(X_t-\bar{X})^2=\Sigma(x_t-\bar{x})^2+\Sigma(u_t-\bar{u})^2+2\Sigma(x_t-\bar{x})(u_t-\bar{u})$$

$$\Sigma(Y_t-\bar{Y})^2=\beta^2\Sigma(x_t-\bar{x})^2+\Sigma(\varepsilon_t-\bar{\varepsilon})^2+2\beta\Sigma(x_t-\bar{x})(\varepsilon_t-\bar{\varepsilon})$$

$$\Sigma(Y_t-\bar{Y})(X_t-\bar{X})=\Sigma[b(x_t-\bar{x})+(\varepsilon_t-\bar{\varepsilon})][(x_t-\bar{x})+(u_t-\bar{u})].$$

By the assumptions made for the model we get the following expectations for the three sums:

$$E[\Sigma(X_t-\bar{X})^2]=\Sigma(x_t-\bar{x})^2+\Sigma\sigma^2_{ut}-\Sigma\sigma^2_{ut}/T=\Sigma(x_t-\bar{x})^2+(1-1/T)\Sigma\sigma^2_{ut},$$

$$E[\Sigma(Y_t-\bar{Y})^2]=\beta^2\Sigma(x_t-\bar{x})^2+(T-1)\sigma^2_{\varepsilon}, \text{ and}$$

$$E[\Sigma(Y_t-\bar{Y})(X_t-\bar{X})]=\beta\Sigma(x_t-\bar{x})^2.$$

From the first and third expectations we get

$$\beta=E[\Sigma(Y_t-\bar{Y})(X_t-\bar{X})]/\{E[\Sigma(X_t-\bar{X})^2]-(1-1/T)\Sigma\sigma^2_{ut}\}$$

which gives

$$K=\{E[\Sigma(X_t-\bar{X})^2]-(1-1/T)\Sigma\sigma^2_{ut}\}/E[\Sigma(X_t-\bar{X})^2].$$

Replacing the expectations with the corresponding sample moments gives the estimators of β and K as

$$b=\Sigma(Y_t-\bar{Y})(X_t-\bar{X})/\{\Sigma(X_t-\bar{X})^2-(1-1/T)\Sigma\sigma^2_{ut}\}$$

$$k=\{\Sigma(X_t-\bar{X})^2-(1-1/T)\Sigma\sigma^2_{ut}\}/\Sigma(X_t-\bar{X})^2.$$

APPENDIX B

The relative precision of the estimator of the sum of error variances

The variance of the sample variance is (a large population is assumed)

$$VAR(s^2)=\mu_4/N - \sigma^4\frac{N-3}{(N-1)N} ,$$

where μ_4/N is the fourth moment and in the case of a binomial situation

$s^2=X(1-X)$, $\sigma^2=x(1-x)$ and $\mu_4=x(1-x)-3x^2(1-x)^2$.

This gives

$$\begin{aligned} VAR(X(1-X)) &= (x(1-x)-3x^2(1-x)^2)/N - x^2(1-x)^2/(N-1)*(N-3)/N \\ &= VAR(X)[1 - 3x(1-x) - x(1-x)(N-3)/(N-1)] \\ &\cong VAR(X)[1 - 4x(1-x)]. \end{aligned}$$

It is seen that the variance of X is greater than the variance of X(1-X). $VAR(\Sigma s^2_{ut})= VAR\{\Sigma[X_t(1-X_t)/N_t]\}=\Sigma VAR[X_t(1-X_t)/N_t]$ is therefore smaller than $\Sigma VAR(X_t)/N^2_t = \Sigma x_t(1-x_t)/N^3_t$.

Let $N_t=N$, then

$$\Sigma VAR(X_t)/N^2 = \Sigma x_t(1-x_t)/N^3 \leq T\mu_x(1-\mu_x)/N^3 .$$

If x_t is small these expressions will be approximately equal. The right hand side can be estimated by replacing μ_x by the sample mean. Further, if x_t is small, $\Sigma X_t(1-X_t)/N$ is approximately equal to $T\bar{X}(1-\bar{X})/N$. The estimated squared coefficient of variation of the estimator Σs^2_{ut} is then approximately $\{T\bar{X}(1-\bar{X})N\}^{-1}$.

APPENDIX C

Variance estimators of (4), (9) and (12).

In the case of normal errors Fuller gives the estimated variances for (4) and (9) as

$$\hat{V}(b)=\frac{1}{T}\hat{M}_{XX}^{-1}\hat{G}\hat{M}_{XX}^{-1}$$

where for the one x variable case:

$$\hat{G}=\frac{1}{T}\begin{bmatrix} \Sigma s_{vt}^2 & \Sigma s_{vt}^2 X_t \\ \Sigma s_{vt}^2 X_t & \Sigma s_{vt}^2 X_t^2 + b^2 \Sigma s_{ut}^4 \end{bmatrix} \quad \text{and} \quad \hat{M}_{XX}=\frac{1}{T}\begin{bmatrix} n & \Sigma X_t \\ \Sigma X_t & \Sigma X_t^2 - \Sigma s_{ut}^2 \end{bmatrix}.$$

If normal errors cannot be assumed Fuller gives the estimated variances as

$$\hat{V}(b)=\frac{1}{T}\hat{M}_{XX}^{-1}\hat{G}\hat{M}_{XX}^{-1}$$

where for the one x variable case:

$$\hat{G}=\frac{1}{T-2}\begin{bmatrix} \Sigma s_{vt}^2 & \Sigma v_t(X_t v_t + b s_{ut}^2) \\ \Sigma v_t(X_t v_t + b s_{ut}^2) & \Sigma (X_t v_t + b s_{ut}^2)^2 \end{bmatrix} \quad \text{and} \quad \hat{M}_{XX}=\frac{1}{T}\begin{bmatrix} T & \Sigma X_t \\ \Sigma X_t & \Sigma X_t^2 - \Sigma s_{ut}^2 \end{bmatrix}.$$

The variance estimator for the weighted least square estimator (12) is, if normal errors can be assumed:

$$\hat{V}(b)=\frac{1}{T^2}\hat{M}_{X\hat{t}X}^{-1}\hat{G}\hat{M}_{X\hat{t}X}^{-1}$$

where

$$\hat{G} = \frac{1}{T}\begin{bmatrix} \Sigma\hat{\pi}_t & \Sigma\hat{\pi}_t X_t \\ \Sigma\hat{\pi}_t X_t & \Sigma\hat{\pi}_t(X_t^2+\hat{\pi}_t b^2\Sigma s_{ut}^4) \end{bmatrix} \quad , \quad \hat{M}_{X\hat{\pi}X} = \frac{1}{T}\begin{bmatrix} \Sigma\hat{\pi}_t & \Sigma\hat{\pi}_t X_t \\ \Sigma\hat{\pi}_t X_t & \Sigma\hat{\pi}_t(X_t^2-\Sigma s_{ut}^2) \end{bmatrix}$$

and

$$\hat{\pi} = \frac{1}{s_{vt}^2}.$$

APPENDIX D

TABLE D 1A. Mean relative biases and RMSE's for the estimators b_1 to b_4 when the variable x1 is used. M is the number of sequences in each cell.

		$\lambda^*<1$						$\lambda^*>1$					
		R^2_{Yx}=0.65 β=333.6		R^2_{Yx}=0.80 β=370.1		R^2_{Yx}=0.95 β=403.3		R^2_{Yx}=0.65 β=333.6		R^2_{Yx}=0.80 β=370.1		R^2_{Yx}=0.95 β=403.3	
		BIAS %	RMSE	BIAS %	RMSE	BIAS %	RMSE	BIAS %	RMSE	BIAS %	RMSE	BIAS %	RMSE
	M	259		473		752		1360		1146		867	
x1	b_1	-52.2	175.1	-56.1	208.2	-58.8	237.5	-65.5	219.9	-66.5	246.8	-67.5	272.8
800	b_2	42.1	163.0	26.7	131.1	15.1	109.6	-15.7	97.2	-20.3	105.2	-25.6	121.9
	b_3	19.8	84.0	4.2	51.8	-8.1	55.6	-15.7	97.2	-20.3	105.2	-25.6	121.9
	b_4	6.2	78.6	-5.4	74.0	-14.7	86.5	-19.0	100.7	-23.3	112.0	-27.7	128.8
	M	262		536		939		1649		1375		972	
x1	b_1	-39.1	132.0	-44.5	165.8	-48.0	194.2	-55.3	186.4	-56.4	209.9	-57.9	234.0
1200	b_2	58.4	220.1	38.2	176.6	23.1	144.2	-6.5	83.6	-11.4	81.8	-17.2	91.4
	b_3	34.4	130.4	14.8	78.1	-.7	48.5	-6.5	83.6	-11.4	81.8	-17.2	91.4
	b_4	20.7	106.1	5.5	74.3	-6.2	64.1	-8.8	78.0	-13.3	79.8	-18.0	91.2
	M	62		292		967		1938		1708		1033	
x1	b_1	-15.5	56.6	-24.3	92.3	-30.5	124.4	-35.7	123.4	-36.9	139.0	-39.3	159.5
2400	b_2	54.8	197.8	30.8	135.6	15.3	94.3	2.3	71.5	-.6	63.6	-6.7	58.6
	b_3	45.1	163.9	20.4	94.5	4.1	52.1	2.3	71.5	-.6	63.6	-6.7	58.6
	b_4	40.7	150.7	17.1	85.1	2.6	50.8	2.3	67.5	-.5	60.2	-6.0	54.6
	M	6		68		613		1994		1932		1387	
x1	b_1			-13.3	53.8	-22.1	91.2	-27.8	99.0	-28.2	108.4	-30.2	123.5
3600	b_2			27.1	109.6	10.7	65.9	-.5	58.7	-1.3	53.7	-5.2	48.0
	b_3			21.7	89.3	4.9	44.8	-.5	58.7	-1.3	53.7	-5.2	48.0
	b_4			20.1	84.4	3.7	42.7	-.1	56.8	-1.0	50.7	-4.5	43.9
	M	2		27		439		1998		1973		1561	
x1	b_1					-16.7	70.4	-22.9	84.2	-23.1	90.3	-24.6	100.9
4800	b_2					8.6	51.8	-1.5	51.2	-1.8	45.4	-4.2	37.3
	b_3					4.4	36.9	-1.5	51.2	-1.8	45.4	-4.2	37.3
	b_4					3.9	35.7	-.9	50.2	-1.1	43.7	-3.3	34.5
	M	153		417		931		1847		1583		1069	
x1	b_1	-26.2	90.3	-32.8	123.1	-37.7	153.2	-44.5	151.5	-45.7	171.0	-47.5	192.5
1200/	b_2	58.9	222.3	37.3	170.5	20.8	129.9	-1.1	78.7	-5.2	71.1	-10.9	72.0
3600	b_3	42.1	155.1	20.5	96.5	3.4	54.4	-1.1	78.7	-5.2	71.1	-10.9	72.0
	b_4	23.3	101.0	11.3	75.0	-.4	50.5	-4.2	65.4	-7.1	62.4	-10.6	65.8

TABLE D 1B. Mean relative biases and RMSE's for the estimators b_1 to b_4 when the variable x2 is used. M is the number of sequences in each cell.

		$\lambda^*<1$						$\lambda^*\geq 1$					
		R^2_{Yx}=0.65 β=69.09		R^2_{Yx}=0.80 β=76.65		R^2_{Yx}=0.95 β=83.53		R^2_{Yx}=0.65 β=69.09		R^2_{Yx}=0.80 β=76.65		R^2_{Yx}=0.95 β=83.53	
		BIAS %	RMSE	BIAS %	RMSE	BIAS %	RMSE	BIAS %	RMSE	BIAS %	RMSE	BIAS %	RMSE
	M	29		192		912		1971		1808		1088	
x2	b_1			-18.6	15.06	-26.6	22.57	-31.2	22.63	-32.1	25.22	-34.3	28.90
800	b_2			31.2	28.69	12.5	17.13	2.2	13.49	.1	11.47	-4.8	9.85
	b_3			22.6	21.03	3.7	9.83	2.2	13.49	.1	11.47	-4.8	9.85
	b_4			21.9	20.55	3.2	9.67	1.8	13.15	-.6	11.13	-5.6	9.86
	M	0		44		701		2000		1956		1299	
x2	b_1					-18.5	16.01	-23.1	17.53	-23.4	18.89	-25.5	21.67
1200	b_2					9.3	12.56	1.3	11.17	.7	9.91	-2.9	7.82
	b_3					4.4	8.96	1.3	11.17	.7	9.91	-2.9	7.82
	b_4					4.1	8.86	1.2	11.10	.6	9.72	-3.1	7.64
	M	0		2		423		2000		1998		1577	
x2	b_1					-8.6	8.31	-12.3	11.29	-12.3	11.22	-13.2	11.78
2400	b_2					6.3	8.05	1.5	9.13	1.5	7.86	.3	5.88
	b_3					4.2	6.73	1.5	9.13	1.5	7.86	.3	5.88
	b_4					4.2	6.62	1.5	9.11	1.5	7.80	.2	5.70
	M	0		0		185		2000		2000		1815	
x2	b_1					-4.6	5.58	-8.8	9.63	-8.8	9.08	-9.2	8.73
3600	b_2					5.7	6.97	.6	8.50	.6	7.12	.2	5.21
	b_3					4.6	6.25	.6	8.50	.6	7.12	.2	5.21
	b_4					4.5	6.11	.6	8.49	.7	7.10	.2	5.13
	M	0		0		48		2000		2000		1952	
x2	b_1							-6.9	8.80	-6.9	7.91	-7.0	6.96
4800	b_2							.2	8.09	.3	6.56	.2	4.47
	b_3							.2	8.09	.3	6.56	.2	4.47
	b_4							.2	8.08	.3	6.54	.2	4.43
	M	0		11		529		2000		1989		1471	
x2	b_1					-12.5	11.23	-17.0	13.81	-17.0	14.33	-18.3	15.82
1200/	b_2					6.6	8.73	.3	9.55	.3	8.42	-1.6	6.69
3600	b_3					3.5	6.69	.3	9.55	.3	8.42	-1.6	6.69
	b_4					3.0	6.37	.1	9.26	.0	7.87	-1.3	5.80

TABLE D 1C. Mean relative biases and RMSE's for the estimators b_1 to b_4 when the variable x3 is used. M is the number of sequences in each cell.

		$\lambda^*<1$						$\lambda^*\geq 1$					
		R^2_{Yx}=0.65 β=21.88		R^2_{Yx}=0.80 β=24.28		R^2_{Yx}=0.95 β=26.45		R^2_{Yx}=0.65 β=21.88		R^2_{Yx}=0.80 β=24.28		R^2_{Yx}=0.95 β=26.45	
		BIAS %	RMSE	BIAS %	RMSE	BIAS %	RMSE	BIAS %	RMSE	BIAS %	RMSE	BIAS %	RMSE
	M	0		0		231		2000		2000		1769	
x3	b_1					-6.9	2.17	-11.2	3.37	-11.1	3.28	-11.6	3.30
800	b_2					5.5	2.14	.2	2.71	.2	2.24	-.4	1.57
	b_3					4.0	1.81	.2	2.71	.2	2.24	-.4	1.57
	b_4					3.8	1.79	.2	2.70	.2	2.23	-.5	1.56
	M	0		0		79		2000		2000		1921	
x3	b_1					-4.2	1.63	-7.6	2.82	-7.6	2.58	-7.8	2.35
1200	b_2					4.1	1.77	.2	2.50	.2	2.01	.1	1.35
	b_3					3.2	1.62	.2	2.50	.2	2.01	.1	1.35
	b_4					3.1	1.56	.2	2.50	.2	2.01	.0	1.34
	M	0		0		4		2000		2000		1996	
x3	b_1							-3.3	2.42	-3.2	1.96	-3.2	1.37
2400	b_2							.8	2.42	.9	1.90	.9	1.18
	b_3							.8	2.42	.9	1.90	.9	1.18
	b_4							.8	2.42	.9	1.90	.9	1.17
	M	0		0		0		2000		2000		2000	
x3	b_1							-2.1	2.37	-2.1	1.86	-2.0	1.15
3600	b_2							.7	2.40	.7	1.86	.7	1.08
	b_3							.7	2.40	.7	1.86	.7	1.08
	b_4							.7	2.40	.7	1.86	.7	1.08
	M	0		0		1		2000		2000		1999	
x3	b_1							-1.6	2.34	-1.6	1.83	-1.5	1.08
4800	b_2							.5	2.37	.5	1.83	.6	1.04
	b_3							.5	2.37	.5	1.83	.6	1.04
	b_4							.5	2.37	.5	1.83	.6	1.04
	M	0		0		24		2000		2000		1976	
x3	b_1							-5.1	2.60	-5.1	2.22	-5.1	1.77
1200/	b_2							.3	2.50	.3	1.99	.3	1.29
3600	b_3							.3	2.50	.3	1.99	.3	1.29
	b_4							.3	2.49	.3	1.97	.2	1.21

TABLE D 1D. Mean relative biases and RMSE's for the estimators b1 to b4 for the cases when the chi-squares were insignificant. M is the number of sequencies in each cell.

		R^2_{Yx}=0.65 β=333.6		R^2_{Yx}=0.80 β=370.1		R^2_{Yx}=0.95 β=403.3	
		BIAS %	RMSE	BIAS %	RMSE	BIAS%	RMSE
x1 800	b_1	-59.2	200.7	-59.2	221.7	-59.3	241.0
M=381	b_2	-524.8	44628.7	-500.2	47873.0	-471.9	50103.9
	b_3	51.2	200.7	32.7	148.3	17.2	99.3
	b_4	27.6	173.3	16.3	140.2	6.3	108.5
x1 1200	b_1	-44.6	154.8	-44.8	170.5	-45.0	185.2
M=89	b_2	245.8	1950.2	248.1	2184.3	250.8	2416.4
	b_3	81.1	288.9	55.6	224.8	34.8	160.4
	b_4	56.4	252.5	40.8	215.2	25.2	161.4

TABLE D2. Standard deviations (s_b) of the M estimates of β in each cell and the relations ($\hat{\sigma}/s_b$) between mean estimated standard deviation ($\hat{\sigma}$) for the estimator b and s_b.

		$\lambda^*<1$						$\lambda^*\geq 1$					
		R^2_{Yx}=0.65		R^2_{Yx}=0.80		R^2_{Yx}=0.95		R^2_{Yx}=0.65		R^2_{Yx}=0.80		R^2_{Yx}=0.95	
		s_b	$\hat{\sigma}/s_b$	s_b	$\hat{\sigma}/s_b$	s_b	$\hat{\sigma}/s_b$	s_b	$\hat{\sigma}/s_b$	s_b	$\hat{\sigma}/s_b$	s_b	$\hat{\sigma}/s_b$
x1	b_3	51.96	3.366	49.44	3.204	44.88	3.146	81.84	1.590	73.68	1.727	65.13	1.857
800	b_4	75.95	1.508	71.29	1.485	63.06	1.540	78.20	1.366	71.23	1.439	64.05	1.512
x1	b_3	61.93	2.560	55.65	2.490	48.45	2.418	80.75	1.383	70.09	1.529	59.37	1.699
1200	b_4	80.57	1.405	71.52	1.412	59.08	1.481	72.28	1.327	62.83	1.438	55.19	1.533
x1	b_3	65.91	1.584	56.82	1.488	49.45	1.469	71.11	1.072	63.58	1.160	52.08	1.306
2400	b_4	65.81	1.333	57.13	1.251	49.71	1.246	67.10	1.064	60.19	1.121	49.11	1.241
x1	b_3			38.95	1.591	40.27	1.332	58.71	1.008	53.48	1.048	43.09	1.193
3600	b_4			39.86	1.389	40.08	1.185	56.84	1.009	50.55	1.051	40.04	1.186
x1	b_3					32.35	1.320	50.95	1.021	44.98	1.066	33.14	1.303
4800	b_4					32.18	1.201	50.12	1.020	43.47	1.066	31.83	1.277
x1	b_3	66.22	2.098	59.86	1.969	52.73	1.870	78.65	1.220	68.40	1.361	56.94	1.539
12/36	b_4	64.66	1.180	62.19	1.115	50.51	1.186	63.92	1.147	56.61	1.202	50.12	1.246
x2	b_3			11.91	1.291	9.32	1.345	13.40	1.032	11.47	1.148	9.02	1.350
800	b_4			11.90	1.220	9.29	1.272	13.10	1.029	11.13	1.138	8.71	1.326
x2	b_3					8.16	1.152	11.14	1.016	9.90	1.059	7.44	1.269
1200	b_4					8.17	1.099	11.07	1.011	9.72	1.055	7.18	1.269
x2	b_3					5.72	1.020	9.08	1.005	7.78	1.007	5.87	1.084
2400	b_4					5.64	1.002	9.05	1.004	7.72	1.006	5.70	1.092
x2	b_3					4.95	.926	8.49	.994	7.11	.983	5.21	.999
3600	b_4					4.84	.922	8.48	.993	7.08	.983	5.13	.999
x2	b_3							8.09	1.006	6.56	1.008	4.46	1.032
4800	b_4							8.08	1.005	6.54	1.007	4.43	1.029
x2	b_3					6.03	1.192	9.55	1.041	8.41	1.055	6.57	1.178
12/36	b_4					5.88	.951	9.26	1.033	7.88	1.042	5.69	1.164
x3	b_3					1.48	1.106	2.71	1.025	2.24	1.047	1.57	1.172
800	b_4					1.50	1.065	2.70	1.025	2.23	1.047	1.55	1.169
x3	b_3					1.38	.921	2.50	1.044	2.01	1.062	1.35	1.135
1200	b_4					1.33	.934	2.50	1.044	2.01	1.062	1.34	1.131
x3	b_3							2.42	1.012	1.89	1.013	1.16	1.014
2400	b_4							2.42	1.012	1.89	1.014	1.15	1.017
x3	b_3							2.39	1.000	1.85	.999	1.07	.995
3600	b_4							2.39	1.000	1.85	.999	1.06	.997
x3	b_3							2.37	1.001	1.83	.997	1.03	.983
4800	b_4							2.37	1.001	1.83	.997	1.03	.981
x3	b_3							2.50	1.003	1.99	1.006	1.29	1.026
12/36	b_4							2.49	1.003	1.97	1.007	1.21	1.033

TABLE D3. 95% confidence intervals for β when $\lambda^* \geq 1$. Percentage to the **left** of , percentage that **cover** and percentage to the **right** of β for the estimators b_3 and b_4.

		R^2_{Yx}=0.65			R^2_{Yx}=0.80			R^2_{Yx}=0.95		
		Left	Cover	Right	Left	Cover	Right	Left	Cover	Right
x1	b_3	13.3	86.7	.0	16.1	83.9	.0	21.3	78.7	.0
800	b_4	13.9	86.1	.0	19.8	80.2	.0	29.3	70.7	.0
x1	b_3	7.3	92.7	.0	9.3	90.7	.0	13.4	86.6	.0
1200	b_4	7.3	92.7	.0	10.2	89.8	.0	17.2	82.8	.0
x1	b_3	4.2	95.8	.0	4.7	95.3	.0	7.1	92.9	.0
2400	b_4	4.1	95.8	.1	4.4	95.6	.0	7.7	92.3	.0
x1	b_3	4.7	94.7	.6	6.2	93.7	.1	11.0	89.0	.0
3600	b_4	4.4	94.8	.8	5.3	94.5	.2	8.9	91.1	.0
x1	b_3	4.0	95.0	1.0	4.2	95.5	.3	4.2	95.8	.0
4800	b_4	3.6	95.0	1.4	3.5	95.8	.7	3.7	96.3	.0
x1-	b_3	4.8	95.2	.0	6.4	93.6	.0	10.1	89.9	.0
1200/3600	b_4	4.5	95.5	.0	7.1	92.9	.0	13.7	86.3	.0
x2	b_3	3.5	96.4	.1	3.5	96.5	.0	4.9	95.1	.0
800	b_4	3.7	96.1	.2	3.7	96.3	.0	6.2	93.8	.0
x2	b_3	2.6	96.3	1.1	2.6	96.8	.6	4.1	95.9	.0
1200	b_4	2.5	95.9	1.5	2.5	96.8	.7	3.7	96.3	.0
x2	b_3	2.1	95.1	2.8	2.6	94.8	2.6	2.8	96.6	.6
2400	b_4	2.1	94.7	3.2	2.2	95.0	2.8	2.9	96.8	.3
x2	b_3	2.8	94.3	2.9	2.8	94.3	2.9	3.6	95.0	1.4
3600	b_4	2.7	94.5	2.8	2.8	94.2	3.0	3.7	94.9	1.4
x2	b_3	2.5	95.0	2.5	2.4	94.9	2.7	2.4	95.8	1.8
4800	b_4	2.4	95.0	2.6	2.5	95.0	2.5	2.8	95.3	1.9
x2	b_3	2.7	95.5	1.8	2.6	96.1	1.3	3.0	96.9	.1
1200/3600	b_4	2.7	95.2	2.1	2.2	95.7	2.1	2.4	97.3	.3
x3	b_3	2.3	95.2	2.5	2.1	95.7	2.2	2.0	97.7	0.2
800	b_4	2.2	95.3	2.5	2.2	95.9	1.9	2.7	97.1	0.2
x3	b_3	2.3	95.6	2.1	2.2	95.9	1.9	2.2	96.6	1.2
1200	b_4	2.3	95.6	2.1	2.3	95.7	2.0	2.3	96.6	1.1
x3	b_3	2.4	94.3	3.3	1.9	94.6	3.5	1.6	94.7	3.7
2400	b_4	2.3	94.4	3.3	2.0	94.5	3.5	1.7	94.7	3.5
x3	b_3	2.5	94.6	2.9	2.4	94.6	3.0	1.8	94.6	3.6
3600	b_4	2.5	94.5	3.0	2.4	94.5	3.1	1.9	94.6	3.5
x3	b_3	2.6	94.4	3.0	2.5	94.5	3.0	2.0	94.1	3.9
4800	b_4	2.6	94.4	3.0	2.4	94.6	3.0	2.0	93.9	4.1
x3	b_3	2.5	94.3	3.2	2.4	94.4	3.2	1.9	95.8	2.3
1200/3600	b_4	2.5	94.4	3.1	2.5	94.3	3.2	2.2	95.1	2.7

TABLE D4. 95% confidence intervals for β when $\lambda^*<1$. Percentage to the **left** of , percentage that **cover** and percentage to the **right** of β for the estimators b_3 and b_4.

		R^2_{Yx}=0.65			R^2_{Yx}=0.80			R^2_{Yx}=0.95		
		Left	Cover	Right	Left	Cover	Right	Left	Cover	Right
x1	b_3	.0	100.0	.0	.0	100.0	.0	.7	99.3	.0
800	b_4	3.1	96.9	.0	7.0	93.0	.0	12.6	87.4	.0
x1	b_3	.0	100.0	.0	.0	100.0	.0	.3	99.7	.0
1200	b_4	1.2	98.0	.8	4.7	95.3	.0	8.6	91.4	.0
x1	b_3	.0	98.4	1.6	.0	100.0	.0	0.8	99.2	.0
2400	b_4	.0	88.7	11.3	.7	99.0	.3	2.8	97.2	.0
x1	b_3				.0	95.6	4.4	2.5	97.5	.0
3600	b_4				.0	89.7	10.3	3.6	96.4	.0
x1	b_3							.2	99.8	.0
4800	b_4							1.6	98.4	.0
x1	b_3	.0	100.0	.0	.0	100.0	.0	.4	99.6	.0
1200/3600	b_4	.0	94.8	5.2	2.4	97.1	.5	5.6	94.4	.0
x2	b_3				.0	99.0	1.0	1.3	98.7	.0
800	b_4				.0	99.0	1.0	1.8	98.2	.0
x2	b_3							.1	99.9	.0
1200	b_4							.4	99.4	.1
x2	b_3							.7	94.6	4.7
2400	b_4							.9	92.9	6.2
x2	b_3							.0	90.3	9.7
3600	b_4							.0	89.7	10.3
x2	b_3									
4800	b_4									
x2	b_3							.6	99.2	.2
1200/3600	b_4							2.1	93.6	4.3
x3	b_3							.4	97.0	2.6
800	b_4							.4	96.5	3.0
x3	b_3							2.5	88.6	8.9
1200	b_4							1.3	89.9	8.9
x3	b_3									
2400	b_4									
x3	b_3									
3600	b_4									
x3	b_3									
4800	b_4									
x3	b_3									
1200/3600	b_4									

Sammanfattning

Ekonometriska modeller innehåller ibland stickprovsbaserade förklarande variabler. Ett exempel är när data från surveyer rörande hushållens attityder till den ekonomiska utvecklingen och deras planer om bilinköp används som förklarande variabler i konsumtions- eller investeringsfunktioner. Dessa variabler är ofta andelar, som exempelvis andel hushåll som vid mättillfället tror att det egna hushållets ekonomi kommer att förbättras under det närmaste året (x3) och andel hushåll som planerar att köpa ny bil inom de närmaste sex månaderna (x1). Urvalsfelet leder till en mätfelssituation. Det är välkänt att vanlig minsta-kvadrat-metod (OLS) då systematiskt underskattar den aktuella parametern. I denna studie undersöks effekten av samplingfel i en oberoende variabel på skattningen av β i en enkel linjär regressionsmodell $Y_t = \alpha + \beta x_t + \varepsilon_t$. Detta görs under antagandet att den sanna variabeln x består av fixa populationsandelar av en händelse och att det observerade antalet händelser är binomialt fördelat. En sådan situation inträffar när andelar estimeras på basis av OSU-surveyer ur stora populationer.

Binomialfördelningsantagandet gör att mätfelsvarianserna lätt kan beräknas eller skattas och att ett enkelt uttryck för den systematiska underskattningen av β kunnat tas fram. Detta uttryck visar att storleken på underskattningen beror på variationskoefficienten för sanna x, nivån på x och den sampelstorlek, N_t, som används för att skatta x_t. Ju större variationskoefficient och ju mindre sampelstorlek desto större är underskattningen. För en given variationskoefficient är underskattningen mycket känslig för den genomsnittliga nivån på x när denna blir låg, som exempelvis är fallet med x1 ovan. För en variabel vars sanna värden är likformigt fördelade mellan 5% och 15%, blir underskattningen (asymptotiskt) vid en sampelstorlek på 2400 hushåll knappt 5%. Om variabelns värden däremot är likformigt fördelade mellan 0.5% och 1.5% blir underskattningen hela 33%.

Eftersom mätfelsvarianserna kan skattas kan OLS-estimatorn modifieras så att den systematiska underskattningen (asymptotiskt) undviks. Några sådana (konsistenta) estimatorer har utvärderats i en simuleringsstudie och jämförts med OLS. Detta har gjorts för tre olika x-variabler varav en har valts att ligga på en låg nivå (0.7% i genomsnitt) och motsvarar x1 ovan. Den tredje variabeln ligger i genomsnitt på nivån 17% och representerar x3. Den andra variabeln (x2) har en genomsnittlig nivå på 5.4%. Olika stickprovstorlekar från 800 hushåll och uppåt har använts vid estimationen av x-värdena. Simuleringen av observerade Y har utförts för tre

förklaringsgrader mellan observerat Y och sant x: .65, .80 och .95. Två test gjordes av egenskaperna hos de simulerade observationssekvenserna och sekvenserna delades in i tre klasser efter resultaten på dessa test. Den första inkluderade sekvenser för vilka icke signifikans erhölls vid test av eventuell variabilitet i sanna x. Analysen visar att sådana sekvenser ej är användbara. Icke-signifikanta chi-två uppträdde emellertid enbart för variabeln på den lägsta nivån (x1) och då för de båda minsta sampelstorlekarna (N_t=800 och N_t=1200). Fördelningen av resterande sekvenser gjordes på basis av värdet på en karaktäristika λ^*, vilken normalt bör vara större eller lika med ett. Simulerade sekvenser från x2 och x3 passerade överlag detta test om förklaringsgraden ej var den högsta (.95). Så var också fallet för x1 om stickprovsstorleken (N_t) var mycket stor. De konsistenta estimatorerna uppförde sig väl när $\lambda^* \geq 1$, utom möjligtvis för variabeln x1. För variabeln på den högsta nivån, x3, var OLS-estimatorn ungefär lika bra som de konsistenta estimatorerna. För de andra två variablerna befanns de konsistenta estimatorerna vara bättre eller avsevärt bättre än OLS.

Att använda stickprovsestimat som oberoende variabel leder till fel i variabeln. Andra estimationsmetoder än OLS bör då övervägas, särskilt om den oberoende variabeln består av andelar på låg nivå eller om variabiliteten i den sanna variabeln kan misstänkas vara liten under observationsperioden. Felvarianserna kan lätt estimeras om OSU kan antas eller om standardfel finns publicerade. Om detta inte är fallet kan man dela upp varje sample i två oberoende delar och sedan använda en modell basered på två oberoende mätningar av x. En sådan procedur har också utvärderats i en liten simuleringsstudie och befunnits ge resultat för variablerna x2 och x3 som är något sämre än de som erhölls under antagande om binomialfördelning.

Prevoius titles in this serie:

No. 1 Current Account and Business Cycles: Stylized Facts for Sweden by Anders Warne and Anders Vredin. December 1989.

No. 2 Change in Technical Structure of the Swedish Economy by Göran Östblom. December 1989.

No. 3 Mamtax. A Dynamic CGE Model for Tax Reform Simulations by Paul Söderlind. December 1989.

No. 4 The Supply Side of the Econometric Model of the NIER by Alfred Kanis and Aleksander Markowski. November 1990.

No. 5 The Financial Sector in the SNEPQ Model by Lennart Berg. February 1991.

No. 6 Consumer Attitudes, Buying Intentions and Consumption Expenditures. An Analysis of the Swedish Household Survey Data by Anders Ågren & Bo Jonsson. April 1991.

No. 7 A Quarterly Consumption Funtion for Sweden 1979-1989 by Lennart Berg & Reinhold Bergström. October 1991.

No. 8 Good Business Cycle Forecasts - A Must for Stabilization Policies by Lars-Erik Öller. February 1992.

No. 9 Forecasting Car Expenditures Using Household Survey Data by Bo Jonsson and Anders Ågren. February 1992.

No. 10 Abstract: Forecasting the Business Cycle Not Using Minimum Autocorrelation Factors by Karl-Gustaf Löfgren, Bo Ranneby & Sara Sjöstedt. February 1992.

No. 11 Current Quarter Forecasts of Swedish GNP Using Monthly Variables by Stefan Gerlach. February 1992.

No. 12 The Relationship Between Manufacturing Production and Different Business Survey Series in Sweden by Reinhold Bergström. February 1992.

No. 13 Forecasting the Swedish Unemployment Rate: V A R vs Transfer Function Modelling by Per-Olov Edlund and Sune Karlsson. March 1992.

No. 14 Business Survey Data in Forecasting the Output of Swedish and Finnish Metal and Engineering Industries: A Kalman Filter Approach by Markku Rahiala and Timo Teräsvirta. March 1992.

No. 15 The Relationship Between Manufacturing and Various BTS Series in Sweden Illuminated By Frequency and Complex Demodulate Methods by Anders Christoffersson, Roland Roberts and Ulla Eriksson. April 1992.