ESTIMATING REGRESSION MODELS USING SURVEY SAMPLE WEIGHTS

L. Magee, A.L. Robb and J.B. Burbidge

Department of Economics
McMaster University
Hamilton, Ontario L8S 4M4

July, 1996

ABSTRACT

Economists often work with microdata taken from surveys using complex sampling designs. Each record includes a weight variable representing the reciprocal of its probability of getting into the sample. When should these weights be used? If they should be used, what is the best way to use them?

In this paper it is argued that using the weights can be desirable in regression models when the population regression coefficient is of interest. A two-step maximum likelihood estimator is proposed as an alternative to OLS and weighted least squares. Tests for selection bias and misspecification are given. The ML estimator does well in simulations, including several cases where it is based on a misspecified model. Selection bias and misspecification tests are shown to be successful as pretests in selecting the best estimator. As an example, the methods are used to estimate the returns to education using data from the Canadian Survey of Consumer Finances.

We thank Dwayne Benjamin, Roman Viveros-Aguilera, Frank Denton, two anonymous referees, seminar participants at Alberta, Simon Fraser and Lakehead Universities, and the 1993 Canadian Econometrics Study Group in Toronto for helpful comments, and the Social Science and Humanities Research Council of Canada for financial support. A GAUSS program which computes the estimator and tests is available on request: magee@mcmaster.ca
1. INTRODUCTION

Economists often work with microdata taken from surveys that are conducted using complicated sample designs. Population members are not equally likely to be included in the sample. Usually these data sets include a weight variable, \( w \), which is interpreted as the inverse of the probability \( p \) of that member having been selected into the sample.

The weights are often used when estimating simple population quantities such as totals, means, quantiles, variances or inequality measures. For example, the population mean of a random variable \( y \) is estimated from the observed \( y_i \)'s, \( i = 1,...,n \), by \( \sum w_i y_i / \sum w_i \) rather than by \( \sum y_i / n \). Observation \( i \) can be thought of as having been randomly sampled from a subpopulation of size \( w_i \), so it is natural to inflate its contribution in the formula by \( w_i \).

In linear regression, things are more complicated. Some clarifying examples are given in Korn and Graubard (1995b). The best way to proceed depends on what one is trying to estimate. We assume that the goal is to estimate the population regression coefficient vector, which is the vector of regression coefficients from a statistical model assumed to have generated the population. (In the statistics literature, this is sometimes termed the 'superpopulation model'.) Another common choice is the 'census regression coefficient', which is what would be obtained from a regression if the entire population had been sampled. For a thorough presentation of these and other issues to do with regression and survey weights, see Pfefferman (1993).

Section 2 outlines a model in which the coefficient vector of interest is the linear regression coefficient vector \( \beta = (E(xx^T))^{-1}E(xy) \), where \( E \) is expectation according to the density assumed to have generated the population values of the dependent variable \( y \) and the independent variables \( x \). A simple sampling scheme is assumed, reflecting the amount of sampling information
typically available. In this scheme, each population member is assigned a sample inclusion probability \( p \), and the sample inclusion outcome is independent of the other population members.

Section 3 looks at estimators of \( \beta \). OLS inconsistency can arise when \( p \) is related to \( y \) in the sense \( E[y - x^T \beta] | p \neq 0 \). Weighted least squares (WLS) is consistent here under mild assumptions. However, it can have a high variance due to large variation in the weights, much of which may have nothing to do with the bias in OLS. For example, the weights in the Canadian Survey of Consumer Finances range from 1 to about 2500. Motivated by this, a conditional (on \( x \)) maximum likelihood estimator is introduced. It uses testable assumptions about the distributions of \( y \) and \( p \) to achieve a lower variance.

Section 4 examines the performance of these estimators as well as some pretest estimators, both with and without a correct specification for the ML estimator. The pretest estimators use various specification tests as pretests, and the size and power of these tests are investigated. As an example, the return to university education in Canada is estimated in Section 5, while Section 6 concludes. In the Appendices, some covariance estimators are described and the estimators and tests are summarized.

Summarizing the paper from a theoretical perspective, we use a simple iid framework with a stochastic inclusion probability to provide consistency conditions for OLS and WLS as estimators of the population regression coefficient. With extra testable distributional assumptions, a maximum likelihood estimator is given that can be computed from two OLS regressions. Simulations suggest that when these assumptions hold, and even in many cases when they do not, this ML estimator outperforms WLS.

From an applied perspective, the paper points out how easy it is to test for the consistency
of OLS. When there is evidence of OLS bias, or "endogenous sampling", the current estimator of choice among practitioners is WLS. WLS is consistent under weak assumptions, but may have a large variance. We offer an ML alternative to WLS. It requires stronger assumptions for consistency than does WLS. It does not follow, however, that practitioners should always choose WLS over ML, since the variance of WLS may be much larger. We suggest a Hausman type test to help choose between the two.

2. THE MODEL

Assume that the population consists of draws of random variables $x$ and $y$, which are generated according to a joint probability density function $f(x,y)$. (The pdfs also depend on parameters which are suppressed.) $x$ is a covariate vector and $y$ is the dependent variable.

The sampling scheme is modelled as follows. The organization collecting the sample assigns to each population member a value of $p$, the probability that it will be selected into the sample. $p$ may depend on $x$ and $y$ as well as on other factors unknown to the researcher. We model $p$ itself as stochastic, distributed according to the pdf $f(p|x,y)$. Although it is not common to treat $p$ as random, Pfeffermann (1993, p.334) discusses previous work using similar approaches. The sample inclusion decisions are made independently across the population members. This assumes away clustering and stratification, discussed below. In this paper, $E_p$ denotes expectation according to the joint density $f(x,y,p) = f(x,y)f(p|x,y)$.

The parameter of interest is the coefficient vector $\beta$ that satisfies:

$$y = x^T \beta + u, \ E_p(u|x) = 0,$$

(1)
This implies \( \beta = [E_p(xx')]'E_p(xy) \). Since \( p \) does not appear in these expressions, \( E_p \) can be viewed as an expectation according to the population density \( f(x,y) \). \( \beta \) is then the population regression coefficient.

We believe that in many situations where a researcher is interested in modelling a linear relationship between \( y \) and \( x \), if the sample contained the entire 'census' population of draws from \( f(x,y) \), the researcher would regress \( y \) on \( x \) and consider the resulting census regression coefficients to be of interest. \( \beta \) is simply the probability limit of these coefficients as the population itself grows large. If (1) were strengthened to \( E(u|x) = 0 \), then \( E(y|x,\beta) = x'\beta \) and clearly \( \beta \) would be of interest. Even without strengthening (1), \( \beta \) can be a useful descriptive device since it minimizes the unconditional population prediction MSE: \( E(y - x'\beta)'(y - x'\beta) \).

A situation where \( \beta \) may not be of interest is the "mixture model" discussed by DuMouchel and Duncan (1983) and Deaton (1995) among others. Deaton divides the population into two strata. Within each stratum, population members have the same \( \beta \) and \( p \), but these differ across strata. The researcher is interested in the population average of the two \( \beta \)'s, call it \( \beta_A \). In general, \( \beta_A \) does not equal our \( \beta \). Deaton shows that neither WLS nor OLS is consistent for \( \beta_A \).

In this paper, the population consists of \( N \) iid draws from the joint density \( f(x,y,p) \), giving \((x_i,y_i,p_i), i \in P\), where \( P \) is the set of \( N \) population members. The researcher has a sample of \( n \) observations \((x_i,y_i,p_i), i \in S\), where \( S \) is the set of \( n \) sample members. Each \( i \in P \) is included in \( S \) with probability \( p_i \).

This sampling assumption can be viewed as stratified sampling with each population member forming its own stratum. It is simpler than the actual sampling methods that produce many of the data sets used by economists; however, those data sets often do not contain enough
information to take the more complex sampling into account. For example, the only published
information available about the sample design in data taken from the Canadian Survey of
Consumer Finances sampling frame is the weight \( w_i = p_i^{-1} \). (For a discussion of this data set see
Bar-Or et al. (1995).) One cannot determine whether two observations came from the same
stratum or cluster, for example. If such information is available, it may be desirable to use it. (For
discussions and examples that use such information, see Pudney (1989, Section 2.5), Deaton
(1993), Selden (1994), Imbens and Lancaster (1991), and Hausman and Wise (1982).) Even in
the absence of stratum/cluster information, it may still be possible to construct covariance
matrices that are consistent under more general sampling scheme assumptions, but we do not
explore this here.

3. ESTIMATION

First we consider the consistency of OLS and weighted least squares (WLS) in estimating
\( \beta \). Our asymptotics hold the population joint density \( f(x,y,p) \) fixed and let the population size \( N \),
and the sample size \( n \), grow large. A two-step ML estimator is proposed later which is suitable
under stronger, but testable, distributional assumptions.

Let \( I_i = 1 \) if \( i \in S \), and \( I_i = 0 \) otherwise, so \( \text{Prob}\{I_i = 1|p_i\} = p_i, \ i \in P \). Let \( \sum_P \) and \( \sum_S \) denote
summation over \( i \) in the population and sample respectively, and let \( g_i = g(x_i,y_i,p_i) \) be an arbitrary
function. Then \( \sum_S g_i = \sum_P I_i g_i \) and \( E_p(Ig) = E_p(E_p(I|p)g) = E_p(pg) \).

3.1 OLS

The OLS estimator is \( \hat{b} = (\sum_S x_i x_i^T)^{-1} \sum_S x_i y_i = (\sum_P I_i x_i x_i^T)^{-1} \sum_P I_i x_i y_i = \beta + (\sum_P I_i x_i x_i^T)^{-1} \sum_P I_i x_i u_i \).
Assuming that \( \text{plim} N^{-1} \sum_P I_i x_i x_i^T \) is finite and positive definite, then \( \text{plim} \ b = \beta \) and OLS is
consistent iff \( \text{plim} \ N^{-1} \sum p_i x_i u_i = E_p(I x u) = E_p(p x u) = 0 \). Since we assume \( E_p(x u) = 0 \), OLS is inconsistent only if \( p \) is related to \( x \) and \( u \) such that \( E_p(x u | p) \neq 0 \).

In Heckman's (1979) selection bias model, for example, this relationship is modelled as \( p = \text{Prob} \{ v > -z^T \omega \} \), where \( v \) is a stochastic term and \( z \) is a vector of explanatory variables. If \( v \) is related to \( u \) and \( z \) is related to \( x \), then in general \( E_p(x u | p) \neq 0 \) and OLS is inconsistent.

### 3.2 Weighted Least Squares

\( p_i \) is observed for \( i \in S \) when \( w_i \) is observed and assumed to equal \( 1/p_i \). (All that is really needed is \( w_i = 1/p_i \).) This makes possible a "weighted least squares" (WLS) estimator \( b_w = \) \[ \left( \sum w_i x_i x_i^T \right)^{-1} \sum w_i x_i y_i. \] This differs from the WLS used for heteroskedastic errors in its motivation for and choice of weights. \( b_w \) is consistent under mild extra conditions since \( b_w - \beta + \) \[ \left( \sum w_i x_i x_i^T \right)^{-1} \sum w_i x_i u_i = \beta + \left( \sum p_i I p_i^{-1} x x_i^T \right)^{-1} \sum p_i I p_i^{-1} x u_i. \] Thus \( \text{plim}(b_w) = \beta + \) \[ \left[ \text{plim} \ N^{-1} \sum p_i I p_i^{-1} x x_i^T \right] \left[ \text{plim} \ N^{-1} \sum p_i I p_i^{-1} x u_i = [E_p(I p^{-1} x x^T)]^{-1} E_p(I p^{-1} x u) = [E_p(p p^{-1} x x^T)]^{-1} E_p(p p^{-1} x u) \right] = [E_p(x x^T)]^{-1} E_p(x u). \] WLS consistency requires only that \( E_p(x x^T) \) be finite and nonsingular. \( E_p(x u) = 0 \) has already been assumed.

The presence of \( p_i^{-1} \) in the above expressions requires that we also assume \( p_i > 0 \) for all \( i \in P \). A weaker assumption \( E_p(x u | p > 0) = 0 \) also would suffice. Otherwise, (1) is not satisfied under the subpopulation density \( f(x, y, p | p > 0) \) and consistent estimation of \( \beta \) is not possible without further assumptions. Thus we rule out the existence of some subpopulation that has been excluded from the sample scheme and that is different in some relevant way from the rest of the population.

Similarly, any non-response is assumed to be exogenous. (Pudney (1989, pp. 80-84) discusses non-response.)
3.3 A Two-Step ML Estimator

Although WLS is consistent under more general conditions than is OLS, it has the disadvantage of having typically a higher variance. This can be especially true when there is large relative variation in the $w_i$'s. In the Canadian Survey of Consumer Finances, for example, $w_i$ ranges from 1 to over 2500. Unless these weights happen to be related to the inverse of the error variance, WLS will inject more heteroskedasticity, possibly a great deal more, into the error term. Remedies include weighting further (Magee (1996)), trimming the weights, and including variables related to the sample design (Korn and Graubard (1995a)).

In this section we consider an estimator that uses more structure on the population density imposed by modelling the process generating the $p_i$'s. Assume that $f(y,p|x)$ can be described as:

\begin{equation}
    y = x^T \beta + u, \quad u \sim N(0, \sigma_u^2)
\end{equation}

\begin{equation}
    q = \ln(p) = g(x)^T \gamma + (yx^T)\gamma + v, \quad v \sim N(0, \sigma_v^2).
\end{equation}

where $u$ and $v$ are independent of each other and of $x$. $g$ is a vector of explanatory variables which are functions of $x$, and includes a constant term set equal to one.

Since $p$ is a probability, then $q \leq 0$, which is not compatible with the normality of $v$ assumed in (3). A similar issue arises when applying the Box-Cox transformation to the dependent variable since this places a bound on the transformed dependent variable (Amemiya (1985, pp.249-252)). Despite this, Box-Cox practitioners typically assume normally distributed errors. In a Monte Carlo study, Spitzer (1978) mentions the issue, and proceeds to assume normality, citing Zarembka's (1974, p.87) observation that "if the probability of such large negative values [of the error term] is quite low, the error term may still be approximately normal." In our application,
discussed in Section 5, the estimated probability of drawing an 'impossible' value of $v$ was extremely small. That is, our estimate of $g(x)^T \gamma_g + (yx)^T \gamma_y$ was always much larger in magnitude than our estimate of $\sigma_q$. The maximum probability of an impossible draw of $v$, given by the maximum over $x$ and $y$ of $\text{Prob}[z > -(g(x)^T \gamma_g + (yx)^T \gamma_y)/\sigma_q]$ where $z \sim N[0,1]$, was always estimated to be less than $\text{Prob}[z > 6.08]$, which is infinitesimal. So while the normality assumption in (3) is, strictly speaking, incompatible with $p < 1$, an adjustment to an appropriately truncated normal distribution for $v$ would have virtually no effect on our empirical results. We expect that this would generally be the case, though an indicator of the severity of the problem can be calculated as above.

In an earlier paper (Magee, Burbidge and Robb (1994)), we described a conditional (on $x$) approach to maximum likelihood (ML) estimation for a more general model defined by densities $f(y|x,\beta)$ and $f(p|y,x,\gamma)$, and gave an ML estimator for the special case (2) and (3). Here we describe a two-step method for obtaining ML estimates. This method is possible because (2) and (3) lead to simple distributions for $y$ and $q$ in the sample.

Consider the density of $y$ given $x$, in the sample:

$$f(y|x,i \in S) = f(y|x_i) \text{Prob}(i \in S | x,y) / \int \text{Prob}(i \in S | x,y) f(y|x_i) dy.$$  \hspace{1cm} (4)

Since $f(y|x_i) = \sigma_y^{-1} \phi((y-x_i \beta^T / \sigma_y)$, where $\phi$ denotes the standard normal pdf, and $\text{Prob}(i \in S | x,y) = F_y = E_y[\exp(g_i^T \gamma_g + (yx_i)^T \gamma_y + v)] = \exp(g_i^T \gamma_g + (yx_i)^T \gamma_y + \sigma_q^2/2)$, then, ignoring multiplicative terms not involving $y$,
\[ f(y|x_i, i \in S) \propto \phi((y - x_i^T \beta)/\sigma_y) \exp((x_i^T \gamma_y)y). \] (5)

Completing the square yields

\[ (y|x_i, i \in S) \sim N[x_i^T(\beta + \sigma_y^2 \gamma_y), \sigma_y^2]. \] (6)

Similarly, for the density of \( q \) given \( x_i \) and \( y_i \) in the sample:

\[ f(q|x_i, y_i, i \in S) = f(q|y_i) \text{Prob}(i \in S|x_i, y_i, q)/\text{Prob}(i \in S|x_i, y_i). \] (7)

Since \( f(q|x_i, y_i) = \phi((q - g_i^T \gamma_y - (y_i)x_i)^T \gamma_y)/\sigma_q \) and \( \text{Prob}(i \in S|x_i, y_i, q) = \exp(q) \), then completing the square yields

\[ (q|x_i, y_i, i \in S) \sim N[g_i^T \gamma_y + (y_i)x_i^T \gamma_y + \sigma_q^2, \sigma_q^2]. \] (8)

Since \( g_i \) contains a constant, then a consistent estimator of \( \gamma_y \) is obtained from OLS estimation of:

\[ q_i = \ln(p_i) = g_i^T \gamma_y^* + (y_i)x_i \gamma + \text{(error)}, i \in S, \] (9)

where \( \gamma_y^* \) indicates that the intercept term of \( \gamma_y \) is augmented by \( \sigma_q^2 \).

Comparing (6) and (2), we see that the population and sample distributions of \( y \) given \( x_i \)
are the same when $\gamma_y = 0$. Thus testing $\gamma_y = 0$ in (9) is a test for consistency of OLS, and significance tests of individual elements of $\gamma_y$ are tests of the consistency of the OLS estimator of corresponding elements of $\beta$. If $\gamma_y = 0$ is accepted and imposed, then $\beta$ can be consistently estimated by OLS on (2). If not, then using (6) a consistent estimator can be obtained from

$$y_i = x_i^T \beta^* + \text{(error)}, \ i \in S,$$

as the two-step estimator (2STEP):

$$b_{2ST} = b^* - \hat{\alpha}_y^2 \hat{\gamma}_y,$$

where $\hat{\alpha}_y^2$ and $b^* = b$ are from OLS estimation of (10) and $\hat{\gamma}_y$ is from OLS estimation of (9). A covariance estimator is given in Appendix A. If no selection bias is suspected for some elements of $b^*$, those corresponding elements of $\gamma_y$ can be set to zero, $\hat{\gamma}_y$ can be replaced by a restricted estimate obtained by deleting the appropriate $(y_i, x_i)$ regressors from (9), and for covariance estimation, the same elements of $\Psi_{2ST}$ and its derivatives can be deleted in the procedure in Appendix A.

2STEP is a function of least squares estimators of the coefficients and variances of (2) and (3). Because of the recursivity of (2) and (3) and the Gaussian errors, these are ML estimators of the functions of the ‘structural’ parameters $\beta$ and $\gamma$ shown in (6) and (8), and 2STEP is the ML estimator of $\beta$. It is not fully efficient, since it conditions on $x$ - the distribution of $x$ in the sample may contain information about $\beta$, even though it does not in the population. (See Pudney (1989,
p.70) and Imbens and Lancaster (1991).)

In other sample selection models, it may not be possible to identify parameters like $\gamma_y$, which helps to describe the inclusion probability conditional on $x$ and $y$ in the population in (3), using the pdf of $p$ given $x$ and $y$ in the sample, given in (8). This identifiability is not a result of the particular parameterization given in (2) and (3), but rather it comes from the link between the population and sample pdfs of $q$ (hence $p$) given in (7) and the ability to observe $p$ through the weights $w$. Our parameterization was chosen for its tractability, not for identification.

4. SIMULATIONS

Equation (3) imposes strong assumptions on the functional form and distribution of the weights, and (2) imposes normality and sphericity on the errors of the equation of interest. These assumptions are not necessary for consistency of WLS, whereas they are needed for consistency of 2STEP. If the assumptions hold, 2STEP should have a lower RMSE than WLS since 2STEP removes the sampling bias without injecting heteroskedasticity into the estimation problem. In this section, we first compare the performance of these two estimators, their $t$-statistics, and associated pretest estimators, when model (2)-(3) is correct. Then we study the effects of various model misspecifications.

4.1 Well Specified Model

$n=500$ observations were simulated from (6) and (8), with $x = [1,z]^T$, $g = [1,z,z^2]^T$, $z \sim N[0,1]$, $\beta = [0,0]^T$, $\sigma_y^2 = 10$, $\gamma_x = [-5,5,0]^T$, and $\gamma_y = [0,5,5]^T$. Three parameters took on 50 values each: $\gamma_y = .0005,..,0.0495$ in steps of .001; $\sigma_q^2 = .01,..,99$ in steps of .02; and $\gamma_o = .0075,..,7.425$ in steps of .015. One replication was done for each parameter combination, giving $50^3 = 125,000$
replications.

$\gamma_{\hat{y}}$ affects the amount of OLS bias. If $\gamma_{\hat{y}} = 0$, then $\gamma_{\hat{y}} = 0$ and (8) shows there is no bias. As $\gamma_{\hat{y}}$ moves away from zero, the bias in the slope coefficient, but not the intercept term, of $\hat{\beta}^*$ in (10) increases. It was not obvious to us ex ante what its effect would be on the relative performance of WLS and 2STEP, since this bias is removed in the second step (11) of 2STEP.

$\sigma_q^2$ controls the amount of 'noise' in equation (3) that generates the log of the inclusion probability. As $\sigma_q^2$ increases, WLS suffers since there is more variation in $w_i$ and hence a bigger $\text{Var}(\text{WLS})$. A larger $\sigma_q^2$ also increases $\text{Var}(\text{2STEP})$ because of a larger $\text{Var}(\hat{\gamma}_s)$, so this parameter's effect on the relative performance of WLS and 2STEP is uncertain.

$\gamma_g$ controls the amount of variation in $q_i$ that is due to $x_i$ but not related to $u_i$ or $y_i$. As $\gamma_g$ increases in magnitude, $\text{Var}(\text{WLS})$ increases, again because of more variation in the $w_i$'s. 2STEP, however, is not affected by $\gamma_g$. From (8) and (9) it can be seen that $\hat{\gamma}_s$, which is the only statistic from (9) that affects 2STEP, is invariant to $\gamma_g$ as long as $g$ is well specified. So 2STEP should improve relative to WLS for larger values of $\gamma_g$.

g contains a third element, $z^2$, which does not affect $q$ since its coefficient is assigned a value of zero. Exclusion of this term probably would sharpen the estimates of $\gamma_g$ and $\gamma_{\hat{y}}$ and improve the performance of 2STEP. However, we include it since in practice one does not know the correct parameterization. Underspecification of $g$ would bias $\hat{\gamma}_s$ and so would bias 2STEP and the bias test. In practice, it may make sense to specify $g$ liberally to guard against this. However, one should not include variables in $g$ that are not in $x$ unless one is sure that those variables would have zero coefficients in (1). This is because the results in Section 3 depend on the densities $f(y|x)$ and $f(q|y,x)$ being conditional on the same set of exogenous variables $x$. 
The first element of $\gamma_s$ is set to -5. It has no effect on the results in this subsection since all statistics of interest are unaffected by it when the data are generated directly from (6) and (8). However, if $\gamma_s$ is too large, it will result in values of $q$ exceeding zero, that is, $p > 1$, which is impossible since $p$ is a probability. This issue also was discussed after equation (3) in Section 3.

In preliminary simulations, the values of $\sigma_y^2$ and $\beta$ had little effect on the results, so they were held fixed. There is wide variation in population $R^2$s in (3) across replications, ranging from virtually zero (when $\gamma_0$ and $\gamma_Y$ are small) to very close to one (when $\sigma_q^2$ is small). Across the 125,000 replications, the median population $R^2$ in (3) is 0.25 and the mean is 0.3. The weights $w$ have a population coefficient of variation averaging about 1.0 across replications.

The results are summarized in Figures 1 to 6. Rather than trying to describe three-dimensional response surfaces, we report ‘marginal’ results in the following sense. Holding one of the parameters $\gamma_0$, $\gamma_Y$ and $\sigma_q^2$ fixed, there are $50^2 = 2500$ replications corresponding to the full range of values taken by the other two parameters. Results are reported conditional on one of the three parameters averaging over these 2500 replications (i.e. ‘marginalizing’ the other two parameters).

In Figure 1, the ratio of root mean squared errors (RMSEs) of 2STEP to WLS are plotted against each of the three parameters as just described. Since the ratios are less than one, 2STEP outperforms WLS according to this criterion. As the plots get lower, the advantage of 2STEP over WLS gets bigger. $\gamma_Y$ and $\sigma_q^2$ have little effect on this ratio, while it is decreasing against $\gamma_0$. This fits the earlier discussion of the expected effects of these parameters on the relative performance of 2STEP and WLS. 2STEP always has a lower RMSE. Standard errors are not reported, but a good sense of the amount of sampling error in the RMSE ratio estimates is
apparent by inspection since the true functions almost certainly are smooth.

4.2 Well-Specified Model: Pretests and Testing

Figure 2 gives the same plot for pretest versions of these estimators. The pretest 2STEP estimator equals OLS or 2STEP depending on the '2STEP pretest', in which the hypothesis $\gamma = 0$ is accepted (in which case the pretest chooses OLS) or rejected (in which case the pretest chooses 2STEP) at the 5% level using a Wald test from OLS estimation of (9). The pretest WLS estimator equals OLS or WLS depending on the 'OLS/WLS pretest', which is a joint significance test of the differences between the elements of OLS and WLS using a covariance matrix estimator described in Appendix A. If this test accepts, then OLS is chosen, if it rejects, WLS is chosen. The pretest is a Hausman test of the consistency of OLS given WLS consistency, and so it is a test for selection bias in OLS. It is similar to the test proposed by DuMouchel and Duncan (1983), but unlike their test, it does not require homoskedasticity in (10), a drawback of their test noted by Kott (1991). These 2STEP and OLS/WLS pretests are testing two restrictions, only one of which is relevant for the estimation of $\beta_2$. In some experimentation we found that the corresponding bias tests that test only the relevant restriction performed slightly better.

The RMSE ratios in Figure 2 have roughly the same magnitude as those in Figure 1, averaging around 0.8. This time there is a downward slope to the $\gamma_Y$ plot, for which there are two reasons. The 2STEP bias test has a power advantage over the OLS/WLS pretest at higher values of $\gamma_Y$ (see Figure 4). Also, OLS becomes very bad for higher $\gamma_Y$, making the consequences of pretest type II errors increasingly grave.

Figure 3 shows the null rejection rates of t-statistics for the hypothesis $\beta_2 = 0$ resulting from the various estimators and estimated variances given in Appendix A. (Some experimentation
with the finite sample improvements to variance estimators given on p.554 of Davidson and MacKinnon (1993) suggested that they made little difference here, and were not used.) The rejection rates are plotted against $\gamma_\gamma$, the parameter that has the largest effect on them. Both the WLS and 2STEP t-tests reject at close to 5% throughout, with WLS overrejecting slightly. OLS drastically overrejects as the selection bias takes over for larger $\gamma_\gamma$. The pretest estimator plots are hump-shaped due to the higher probability of using the biased OLS (from type II errors in the bias tests) for moderate $\gamma_\gamma$ values than for large ones. The 2STEP pretest estimator does not overreject as much as the WLS one, both because the 2STEP t-test itself rejects slightly less often than the WLS t-test and because the 2STEP pretest selects OLS less often (that is, has higher power) than the OLS/WLS pretest.

Power is examined by plotting the rejection rates of these 2STEP and OLS/WLS pretests against the three parameters in Figures 4, 5 and 6. Figure 4 shows the power advantage of the 2STEP test as just mentioned. Figure 5 shows that the extent of this power advantage is not related to $\sigma_q^2$, although both tests lose power as $\sigma_q^2$ grows. Figure 6 shows that the OLS/WLS test, but not the 2STEP test, loses power for larger values of $\gamma_\sigma$. This accords with the earlier discussion of the effect of $\gamma_\sigma$ on the precision of 2STEP and WLS.

4.3 Misspecified Models

Since the assumptions in (2) and (3) are quite strong, we reexamine the estimators under various misspecifications, which are considered one at a time to keep things simple. First we describe the models.

NOMIS. This benchmark model is taken from the previous section, with the three varying parameters now set at their midpoints: $\gamma_\gamma = .025$, $\sigma_q^2 = .5$, and $\gamma_\sigma = .375$. The remaining
models depart from this in some way.

**HETY1.** Multiplicative heteroskedasticity exists in (2). This is NOMIS but with \( E_p(u^2|x) = k \cdot \exp(x) \). Here, and in the models that follow, \( k \) refers to a constant chosen from a prior simulation to make a relevant unconditional property close to its NOMIS value. This helps to focus on the effect of the misspecification when comparing with NOMIS. Here, for example, \( k = 10/E_p(\exp(x)) \), so that \( E_p(u^2) = 10 \), as in NOMIS.

**HETY5.** Like HETY1 but with less severe heteroskedasticity: \( E_p(u^2|x) = k \cdot \exp(x/2) \).

**NONGY+.** The error in (2) is nonnormal, skewed right: \( u = (10/2)^{1/2}(z^2 - 1) \), where \( z \) is \( N[0,1] \).

So \( E_p u = 0 \) and \( E_p u^2 = 10 \).

**NONGY-.** Like NONGY+ but skewed left: \( u = -5^{1/2}(z^2 - 1) \).

**HETQ1.** Multiplicative heteroskedasticity exists in (3). \( E_p(v^2|x) = k \cdot \exp(x) \). \( k \) is chosen so \( E_p(v^2) = .5 \).

**NONQ+.** The error in (3) is nonnormal, skewed right. \( v = (.5/2)^{1/2}(z^2 - 1) \), where \( z \sim N[0,1] \).

**NONQ-.** Like NONQ+ but skewed left: \( v = -.5^{1/2}(z^2 - 1) \).

**MISO1.** \( x_1 \) and \( y \) enter together in (3) as \( k \cdot \text{sign}(x_1y)(x_1y)^2 \) rather than as \( x_1y \). Both terms have mean zero. \( k \) is chosen to equate their population variances.

**MISO2.** \( x_1 \) and \( y \) enter together in (3) as \( k_1[(x_1y)^2 - k_1] \), rather than as \( x_1y \). \( k_1 \) and \( k_2 \) are chosen so that this term has population mean zero and variance equal to \( \text{Var}(x_1y) \) in NOMIS.

The average coefficient of variation of the weights in the samples from these 10 models ranged from .94 to 1.08. The average ratio of the lowest to the highest weight in the samples ranged from 186 to 210 except for NONQ+ (124) and NONQ- (370). For comparison, the average coefficient of variation of the weights across the 128 cells of data used in the empirical
example of the next section was .81 and the average max/min weight ratio was 217.

Since (2) and (3) no longer hold, the samples could not be generated from (6) and (8). First, a set of population values were generated using the misspecified counterparts to (2) and (3). w was set equal to the nonnegative integer closest to \( \exp(-q) \). Then p was constructed as 1/w. (This introduces another misspecification, since the integer rounding causes \( q \neq \ln(p) \). The w’s in real data sets usually are integers.) The population values were then included in the sample with probability p. This procedure was repeated until \( n=500 \) sample observations were collected.

Six estimators were compared with WLS: OLS, 2STEP, and four pretest estimators. PT1 is the WLS pretest estimator and PT2 is the 2STEP pretest estimator studied in the previous subsection.

PT3 and PT4 depend on the outcome of a ‘2STEP/WLS’ pretest. This is a test of \( \text{plim}(\text{WLS}) = \text{plim}(\text{2STEP}) \), constructed using the same method as the previous tests (see Appendix A). It can be interpreted as a Hausman test of the assumptions made in (2) and (3) that are required for the consistency of 2STEP that were not required for consistency of WLS. PT3 = 2STEP if the 2STEP/WLS pretest accepts, and = WLS if it rejects. PT4 = OLS if the OLS/WLS pretest accepts, and = PT3 if it rejects. So PT1, PT2 and PT4 all begin with a test for OLS consistency. If it rejects, they move on to choose WLS or 2STEP in some manner. The rejection rates of these pretests are given in Table 3 and discussed later.

Table 1 lists the ratios of the RMSEs of these estimators to RMSE(WLS), based on 10,000 replications for each model. Standard errors were computed using an asymptotic variance given in Appendix A. 2STEP still fares well relative to WLS except for HETY1, HETY.5 and NONGQ+. The effect of heteroskedasticity in u, on 2STEP can be seen from (6). The mean of y
given \( x \) is \( x(\beta + \sigma_y^2 \gamma_y) \) in the sample. If \( \gamma_y \neq 0 \), and \( \sigma_y^2 \) fluctuates in some manner related to \( x \), then ignoring that fact will bias 2STEP. The heteroskedasticity in HETY1 is quite extreme, and while 2STEP is better than OLS, both are much worse than WLS. PT3 performs well, since the 2STEP/WLS test seldom accepts (that is, the pretest seldom selects 2STEP).

The less extreme HETY.5 has 2STEP still faring worse than WLS, but not by nearly as much. However, PT3 and PT4 do worse than in HETY1 because now the pretests reject less often.

All of the pretests in NONGQ+ have low power, and the PT estimators perform about the same as OLS and 2STEP.

In the other seven models, 2STEP has a sizable RMSE advantage over WLS. For several models this advantage is even greater than in the well-specified model NOMIS. PT3 retains most of this advantage because the power of the WLS/2STEP pretest is appropriately low. The other PT estimators do not do nearly as well. They use an initial pretest that involves some risk of selecting OLS, which is worse than WLS except in MISOQ2. This suggests that if one has prior suspicions of a bias in OLS, then it is better to proceed directly to WLS, 2STEP, or PT3 rather than first pretesting for OLS consistency. (This is similar to Nakamura and Nakamura’s (1978) finding on the damaging effects of pretesting for serial correlation in the errors.)

MISOQ2 stands out as the one model where OLS outperforms WLS. \( q \) is related to \( x \) and \( y \), but since it is not linearly correlated with them, OLS is unbiased. 2STEP is only slightly worse than OLS, and all six estimators are better than WLS.

Table 2 gives the null rejection rates of the t-statistics on the slope coefficient of the seven estimators, at the 5% nominal level. The t-statistics for the pretest estimators are those of the
selected component estimator (OLS, WLS or 2STEP). WLS and PT3 reject at below 10% for all models. 2STEP is quite close to 5% except for HETY1 and HETY.5. OLS and the other pretest estimators (which sometimes select OLS) all overreject badly except for MISQ2, for which OLS is not biased.

Table 3 gives the rejection probabilities of the various specification tests. 2STEP and OLS/WLS are testing the null that OLS is unbiased. This null is true only for MISQ2, where they reject at a rate reasonably close to the nominal size of 5%. For the other models, 2STEP has more power than OLS/WLS except for NONGQ+, which is appropriate since this is the one other model where 2STEP does not do much better than OLS.

WLS/2STEP is testing the null hypothesis that 2STEP is consistent, assuming that WLS is consistent. Presumably 2STEP is not consistent except for NOMIS and MISQ2. In most of the other models, though, this test still rejects at close to 5%. As noted earlier, this is desirable because the 2STEP estimator usually has a lower RMSE than WLS despite its inconsistency. In NONGQ+, it rejects at close to 5% despite \text{RMSE}(2STEP) > \text{RMSE}(WLS). In HETY1 and HETY.5, it rejects more often, but as noted earlier, not often enough to prevent PT3 from being poor in HETY.5.

5. **EMPIRICAL EXAMPLE**

The returns to university education were examined using data on males aged 25-64, working full-time, full-year, with earnings not primarily from self-employment, taken from Statistics Canada's Survey of Consumer Finances from 1971 to 1992 (biennial before 1981, annual thereafter except for 1983), totalling 16 surveys. Only those with a university degree and
those with 11-13 years of education ("high school") were selected. This gave a total of 101,725 observations. (See Bar-Or et al. (1995) for more about the data.)

A minimum chi-square approach was used. The data were divided into 16 \times 8 = 128 age/year cells, based on eight 5-year age groups from 25-29 to 60-64. The cell sizes ranged from 106 for age 60-64 in 1971 to 1,791 for age 30-34 in 1990, with an average cell size of 794.7.

(Similar results were obtained when the number of age groups was reduced from 8 to 4.) Within each cell, the log of weekly earnings was regressed on age (to control for within-age-group variation in age), an education dummy (= 1 if university, 0 if high school), and sets of dummies for area size and region (5 area sizes, and five regions: Atlantic, Quebec, Ontario, Prairies, British Columbia). For 2STEP, these same variables were included in $g_i$ in (9). The region and area size variables were included due to their explanatory power in (9), which improved the performance of 2STEP. As discussed in Section 4.1, if they are included in (9) then they must also be in (10) unless one is sure that their coefficients equal 0 in (2).

The coefficient on the education dummy is interpreted as the return to a university education. The OLS, WLS and 2STEP estimates of this coefficient, along with their covariance matrix, were computed for each cell, and saved for use in the next step.

In the minimum chi-square step, these three sets of estimates were regressed separately on eight age dummies, year and year squared, weighting by the inverse of their variance estimates. Higher order and quadratic spline terms in year were not significant, nor were age-year interactions.

Because the correlations between the three estimates were not constrained to be equal across cells, it is asymptotically more efficient to estimate the three equations jointly as a kind of
generalized SURE model, even though the right hand side variables are the same in each equation. This was tried, and the results were almost identical to those reported here.

Letting \( y_j \) be the 128-element vector of coefficient estimates from estimation method \( j \) (\( j = \text{OLS, WLS, 2STEP} \)), \( \Omega_{ij} \) be a diagonal matrix containing their estimated variances, and \( X \) be the 128 \( \times \) 10 matrix of regressors, then we are minimizing \( (y_j - X\alpha_j)^T\Omega_{ij}^{-1}(y_j - X\alpha_j) \) with respect to \( \alpha_j \). The minimized value is a specification test having an asymptotic null distribution of \( \chi^2 \) with 128 \(-\) 10 \( = \) 118 d.f. These statistics were 138.1, 168.9 and 124.6 for OLS, WLS and 2STEP respectively. The WLS test has a p-value of .0015, while the others accept at 5%. Since the WLS test still rejected after adding the extra variables mentioned above, we continued, assuming that the source of this rejection is orthogonal to the variables of interest, i.e. they are specific cell/year "blips". Year- and age-specific specification tests did not reveal any patterns.

The ratios of the estimators' standard errors, averaged across the ten coefficients, are \( \text{St.Err(OLS)}/\text{St.Err(WLS)} = 0.777 \) and \( \text{St.Err(2STEP)}/\text{St.Err(WLS)} = 0.858 \). On that basis, if all three estimators were consistent, OLS beats 2STEP which in turn beats WLS. The consistency of OLS and 2STEP was examined using Hausman tests comparing the coefficient estimates. These were constructed using \( \text{Var}(\hat{\alpha}_j - \hat{\alpha}_k) = A_j^T\Omega_{jj}A_j + A_k^T\Omega_{kk}A_k - A_j^T\Omega_{jk}A_k - A_k^T\Omega_{kj}A_j \), where \( A_j = \Omega_{jj}^{-1}X(X^T\Omega_{jj}^{-1}X)^{-1} \). The results are reported in Table 4. For the full coefficient vectors, both OLS=WLS and WLS=2STEP were rejected, the first indicating inconsistency of OLS and the second indicating inconsistency of 2STEP. However, these rejections appear to be driven by the age variables and not the year ones, since all three tests accept for the year coefficients. These tests lead to the conclusion: OLS and 2STEP are inconsistent for the age effects, and they are both consistent for the year effects. Thus one might use WLS for the age effects and OLS for the
year effects.

Figure 7 shows the predicted return for the age 25-29 group against year for the three estimators. The plots are quadratics in year. OLS and 2STEP have a very similar shape, reflecting the very low statistic testing OLS=2STEP for the year variables. Their levels differ due to differences in the coefficient estimates on the age 25-29 dummy. All three estimates show a decline in the returns to education through the mid-1980's, followed by a smaller rebound. Since consistency of OLS is accepted for the year effect, we might place more stock in the more precise OLS and 2STEP shapes, showing a sharper decline and rebound than does WLS.

Plots for other age groups have the same shape, due to the additivity of the age and year effects, which was supported by the insignificance of age/year interactions. These plots can be compared with those in Figure 6 of Bar-Or et al. (1995), which are much noisier due to a less restrictive model.

Figure 8 plots the predicted returns to education against age for 1981 from the three estimators. OLS and 2STEP are very close, although OLS=2STEP for the age coefficients is rejected at 5%. The specification tests point to WLS as the only consistent estimator of the age effects. It estimates higher returns at higher ages than do OLS and 2STEP.

6. CONCLUSION

We have presented a model that is compatible with the kind of information usually available in data sets drawn from complex samples: sampling weights, but no information on clusters or strata. The parameter of interest is assumed to be the population regression coefficient. Consistency conditions are discussed for OLS and WLS - naturally WLS is consistent under much
weaker assumptions. Another estimator, called 2STEP, is presented which relies on stronger assumptions than WLS, but should perform better if those assumptions hold. Appendix B summarizes the estimators and specification tests.

Simulations support this conjecture. Some specification tests and associated pretest estimators perform acceptably. The methods are used to estimate age and year effects on the returns to a university education using earnings data from the Canadian SCF. OLS appears to suffer from endogenous sampling bias for the age effects but not the year effects.
APPENDIX A

Covariance Matrix Estimation for 2STEP

The 2STEP estimator (11) based on OLS estimation of (9) and (10) is the value of $\beta_{2ST}$ that helps to solve the following set of equations:

$$\sum \psi_{2ST,i} = 0,$$

where

$$\psi_{2ST,i} = \begin{bmatrix} g_i (q_i - g_i^T y_g - y_i x_i^T y_y) \\ y_i x_i (q_i - g_i^T y_g - y_i x_i^T y_y) \\ x_i (y_i - x_i^T \beta_{2ST} - \sigma_y^2 x_i^T y_y) \\ (y_i - x_i^T \beta_{2ST} - \sigma_y^2 x_i^T y_y)^2 - \sigma_y^2 \end{bmatrix}$$

This is a two-step M-estimator of the sort discussed by Duncan (1987), and the regularity conditions described there are satisfied. Letting $\theta = [y_g^T, y_y^T, \beta_{2ST}^T, \sigma_y^2]^T$ be the vector of parameters, then the covariance matrix of $\hat{\theta}$, the solution to (a2), is estimated consistently by

$$\text{VarEst} (\hat{\theta}) = \left[ \sum (\partial \psi_{2ST,i} / \partial \theta^T) \right]^{-1} \left[ \sum \psi_{2ST,i} \psi_{2ST,i}^T \right] \left[ \sum (\partial \psi_{2ST,i} / \partial \theta^T) \right]^{-1},$$

where, after some simplifications stemming from (a1),

$$\sum (\partial \psi_{2ST,i} / \partial \theta^T) = -\sum
\begin{bmatrix}
g_i g_i^T & y_i g_i x_i^T & 0 & 0 \\
y_i x_i g_i^T & y_i^2 x_i x_i^T & 0 & 0 \\
0 & \sigma_y^2 x_i x_i^T & x_i x_i^T & y_i x_i x_i^T \\
0 & 0 & 0 & 1
\end{bmatrix}$$

when evaluated at $\hat{\theta}$.

Covariance Matrix Estimators for OLS, WLS and Covariances
OLS and WLS are also M-estimators, solving
\[ \sum \psi_{OLS,i} = \sum x_i (y_i - x_i \beta) = 0 \]  
(a5)

and
\[ \sum \psi_{WLS,i} = \sum w_i x_i (y_i - x_i \beta_w) = 0 \]  
(a6)
respectively.

Their covariance matrix estimators, and any estimates of covariances across estimators necessary for the specification tests, can be obtained by stacking the first order conditions as:
\[ \sum \psi_i = \sum \psi_{OLS,i} = 0, \quad \psi_{WLS,i} \]
(a7)

and using
\[ \text{VarEst}(\hat{\eta}) = [\sum (\partial \psi_i / \partial \eta^\top)]^{-1} [\sum \psi_i \psi_i^\top] [\sum (\partial \psi_i / \partial \eta^\top)]^{-1} \]

where \( \eta \) contains \( \theta \), \( \beta \) and \( \beta_w \). The matrix \( \sum (\partial \psi_i / \partial \eta^\top) \) is block diagonal, with the blocks consisting of \( (\partial \psi_{2ST,i} / \partial \theta^\top) \) from (a4), \(-\sum x_i x_i^\top\), and \(-\sum w_i x_i x_i^\top\).

Covariance Matrix of Root Mean Square Ratio

Let \( a_i \) and \( b_i \) be the squared errors of two estimators from replication \( i \). Then the estimated MSEs of the two estimators are the sample means \( \bar{a} = \sum a_i/n \) and \( \bar{b} = \sum b_i/n \) and the RMSE ratio estimator is \( \text{EstRMSER} = (\bar{a}/\bar{b})^{1/2} \). The usual linear approximation method gives:

\[ \text{EstVar}(\text{EstRMSER}) = \frac{(\bar{a} + \bar{b})}{4\bar{b}} [\text{EstVar}(\bar{a})/(\bar{a})^2 + \text{EstVar}(\bar{b})/(\bar{b})^2 - 2\text{EstCov}(\bar{a},\bar{b})/(\bar{a}\bar{b})], \]

where \( \text{EstVar}(\bar{a}) = \sum (a_i - \bar{a})^2/n^2 \), etc. are calculated as usual. This variance estimator worked very
well in a small simulation study.

APPENDIX B

Summary of Estimators and Tests

Given the dependent variables $y_i$, explanatory variables $x_i$, and sampling weights $w_i$, $i = 1, \ldots, n$, the goal is to estimate $\beta$ in $y = x^T\beta + u$, where $E(xu) = 0$ in the population. The OLS estimator is

$$b = \left(\sum x_i x_i^T\right)^{-1}\sum x_i y_i.$$

The WLS estimator is $b_w = \left(\sum w_i x_i x_i^T\right)^{-1}\sum w_i x_i y_i$. The 2STEP estimator is $b_{2ST} = b - \hat{\sigma}_y^2 \hat{\gamma}_y$, where $\hat{\sigma}_y^2 = n^{-1} \sum (y_i - x_i^T b)^2$ and $\hat{\gamma}_y$ is from OLS estimation of the equation $-\ln(w_i) = g_i^T \gamma_* + (y_i x_i) \gamma_y + \text{(error)}$, where $g_i$ is some function of $x_i$, $g_i = x_i$ for example. Covariance estimators are outlined in Appendix A.

In practice, one might not know which estimator to use. If there is no (or little) selection bias, then OLS is best. If there is enough selection bias, then either WLS or 2STEP should be better. If the assumptions implied by (2) and (3):

$$y = x^T \beta + u, \quad u \sim N[0, \sigma_y^2]$$

$$q = \ln(p) = g(x)^T \gamma_* + (yx)^T \gamma_y + v, \quad v \sim N[0, \sigma_v^2]$$

that are not already implied by (1):

$$y = x^T \beta + u, \quad E_p(ux) = 0,$$

are sufficiently close to being valid, (refer to these assumptions as A23) then 2STEP is preferred, otherwise WLS is better. Two kinds of tests can help, tests for selection bias and tests for A23.

At least three selection bias tests can be obtained from the above calculations. One is a Wald test of the null hypothesis $\gamma_y = 0$, which is sufficient for OLS consistency if A23 are valid. The other two are Hausman type tests, with null hypotheses $\text{plim } b = \text{plim } b_w$ and $\text{plim } b = \text{plim } b_{2ST}$. 

The $b/b_w$ test is easier to calculate and does not require A23, whereas the $b/b_{2ST}$ test should be more powerful when A23 hold. They can be calculated by testing the joint significance of $(b - b_w)$ and $(b - b_{2ST})$ respectively, using the covariance matrices given in Appendix A. All three selection bias tests can be modified to test for the presence of selection bias in a subvector of $b$, if only certain elements of $\beta$ are of interest.

For testing A23, one might consider tests based on the residuals of (2) and (3). For example, the results of Section 4.2 suggest testing for heteroskedasticity in (2). Another specification check, mentioned in Section 3.2, is to compute the maximum estimated value of $\text{Prob}[z > \{g(\gamma)\gamma_k + (y^\gamma)\gamma_k\}/\sigma_k]$, where $z \sim N[0,1]$. If it is not tiny, then one might worry about the incompatibility of the normality assumption of $v$ in (3) with the fact that $q \leq 0$. In this paper we used a more direct approach, testing the consistency of $b_{2ST}$ by testing the null hypothesis $\text{plim } b_w = \text{plim } b_{2ST}$. It can be computed by testing the joint significance of $(b_w - b_{2ST})$. 
REFERENCES


Table 1. RMSE Ratios for Misspecified Models

<table>
<thead>
<tr>
<th>MODEL</th>
<th>OLS</th>
<th>2STEP</th>
<th>PT1</th>
<th>PT2</th>
<th>PT3</th>
<th>PT4</th>
</tr>
</thead>
<tbody>
<tr>
<td>NOMIS</td>
<td>1.38</td>
<td>0.84</td>
<td>1.27</td>
<td>1.04</td>
<td>0.87</td>
<td>1.22</td>
</tr>
<tr>
<td></td>
<td>(.014)</td>
<td>(.012)</td>
<td>(.010)</td>
<td>(.009)</td>
<td>(.010)</td>
<td>(.011)</td>
</tr>
<tr>
<td>HETY1</td>
<td>3.98</td>
<td>2.67</td>
<td>1.01</td>
<td>2.67</td>
<td>1.05</td>
<td>1.07</td>
</tr>
<tr>
<td></td>
<td>(.241)</td>
<td>(.176)</td>
<td>(.030)</td>
<td>(.176)</td>
<td>(.052)</td>
<td>(.054)</td>
</tr>
<tr>
<td>HETY.5</td>
<td>2.58</td>
<td>1.38</td>
<td>1.27</td>
<td>1.48</td>
<td>1.20</td>
<td>1.40</td>
</tr>
<tr>
<td></td>
<td>(.074)</td>
<td>(.032)</td>
<td>(.027)</td>
<td>(.039)</td>
<td>(.021)</td>
<td>(.032)</td>
</tr>
<tr>
<td>NONGY+</td>
<td>1.38</td>
<td>0.82</td>
<td>1.19</td>
<td>0.92</td>
<td>0.82</td>
<td>1.15</td>
</tr>
<tr>
<td></td>
<td>(.019)</td>
<td>(.033)</td>
<td>(.017)</td>
<td>(.027)</td>
<td>(.033)</td>
<td>(.019)</td>
</tr>
<tr>
<td>NONGY-</td>
<td>1.41</td>
<td>0.87</td>
<td>1.25</td>
<td>1.08</td>
<td>0.89</td>
<td>1.23</td>
</tr>
<tr>
<td></td>
<td>(.016)</td>
<td>(.014)</td>
<td>(.012)</td>
<td>(.011)</td>
<td>(.013)</td>
<td>(.012)</td>
</tr>
<tr>
<td>HETQ1</td>
<td>1.16</td>
<td>0.82</td>
<td>1.18</td>
<td>1.05</td>
<td>0.86</td>
<td>1.12</td>
</tr>
<tr>
<td></td>
<td>(.010)</td>
<td>(.011)</td>
<td>(.007)</td>
<td>(.009)</td>
<td>(.009)</td>
<td>(.008)</td>
</tr>
<tr>
<td>NONGQ+</td>
<td>1.22</td>
<td>1.16</td>
<td>1.21</td>
<td>1.25</td>
<td>1.12</td>
<td>1.20</td>
</tr>
<tr>
<td></td>
<td>(.010)</td>
<td>(.010)</td>
<td>(.008)</td>
<td>(.010)</td>
<td>(.008)</td>
<td>(.009)</td>
</tr>
<tr>
<td>NONGQ-</td>
<td>1.41</td>
<td>0.74</td>
<td>1.20</td>
<td>0.74</td>
<td>0.79</td>
<td>1.12</td>
</tr>
<tr>
<td></td>
<td>(.015)</td>
<td>(.023)</td>
<td>(.010)</td>
<td>(.022)</td>
<td>(.017)</td>
<td>(.011)</td>
</tr>
<tr>
<td>MISQ1</td>
<td>1.31</td>
<td>0.82</td>
<td>1.24</td>
<td>1.02</td>
<td>0.85</td>
<td>1.19</td>
</tr>
<tr>
<td></td>
<td>(.012)</td>
<td>(.013)</td>
<td>(.010)</td>
<td>(.009)</td>
<td>(.011)</td>
<td>(.010)</td>
</tr>
<tr>
<td>MISQ2</td>
<td>0.81</td>
<td>0.85</td>
<td>0.86</td>
<td>0.82</td>
<td>0.88</td>
<td>0.83</td>
</tr>
<tr>
<td></td>
<td>(.008)</td>
<td>(.006)</td>
<td>(.007)</td>
<td>(.008)</td>
<td>(.005)</td>
<td>(.007)</td>
</tr>
</tbody>
</table>

The entries are estimates of RMSE(A)/RMSE(WLS), where ‘A’ refers to the estimator given in the column headings. Standard errors are in parentheses.
Table 2. Null Rejection Rates (in %) of t-tests (5% Nominal Rate)

<table>
<thead>
<tr>
<th>MODEL</th>
<th>OLS</th>
<th>2STEP</th>
<th>PT1</th>
<th>PT2</th>
<th>PT3</th>
<th>PT4</th>
<th>WLS</th>
</tr>
</thead>
<tbody>
<tr>
<td>NOMIS</td>
<td>42.9</td>
<td>5.5</td>
<td>29.9</td>
<td>17.3</td>
<td>6.0</td>
<td>29.5</td>
<td>6.5</td>
</tr>
<tr>
<td>HETY1</td>
<td>94.4</td>
<td>47.1</td>
<td>6.7</td>
<td>47.2</td>
<td>7.3</td>
<td>7.9</td>
<td>6.0</td>
</tr>
<tr>
<td>HETY.5</td>
<td>70.6</td>
<td>11.1</td>
<td>13.7</td>
<td>15.3</td>
<td>8.1</td>
<td>15.9</td>
<td>5.9</td>
</tr>
<tr>
<td>NONGY+</td>
<td>45.3</td>
<td>5.4</td>
<td>28.2</td>
<td>12.0</td>
<td>6.0</td>
<td>27.6</td>
<td>9.4</td>
</tr>
<tr>
<td>NONGY-</td>
<td>38.9</td>
<td>5.3</td>
<td>26.1</td>
<td>17.5</td>
<td>5.4</td>
<td>25.2</td>
<td>5.6</td>
</tr>
<tr>
<td>HETQ1</td>
<td>35.5</td>
<td>5.9</td>
<td>30.5</td>
<td>25.0</td>
<td>6.6</td>
<td>29.6</td>
<td>6.4</td>
</tr>
<tr>
<td>NONGQ+</td>
<td>35.5</td>
<td>6.5</td>
<td>28.9</td>
<td>31.9</td>
<td>5.8</td>
<td>28.0</td>
<td>6.4</td>
</tr>
<tr>
<td>NONGQ-</td>
<td>43.2</td>
<td>5.6</td>
<td>23.5</td>
<td>5.8</td>
<td>6.5</td>
<td>23.6</td>
<td>6.5</td>
</tr>
<tr>
<td>MISQ1</td>
<td>37.7</td>
<td>5.4</td>
<td>29.0</td>
<td>17.5</td>
<td>5.9</td>
<td>28.5</td>
<td>6.8</td>
</tr>
<tr>
<td>MISQ2</td>
<td>5.4</td>
<td>5.4</td>
<td>6.1</td>
<td>5.6</td>
<td>5.9</td>
<td>5.7</td>
<td>5.9</td>
</tr>
</tbody>
</table>
Table 3. Rejection Rates of Specification Tests

<table>
<thead>
<tr>
<th>MODEL</th>
<th>OLS/WLS</th>
<th>2STEP</th>
<th>WLS/2STEP</th>
</tr>
</thead>
<tbody>
<tr>
<td>NOMIS</td>
<td>38.5</td>
<td>68.0</td>
<td>6.2</td>
</tr>
<tr>
<td>HETY1</td>
<td>98.7</td>
<td>99.3</td>
<td>91.8</td>
</tr>
<tr>
<td>HETY.5</td>
<td>81.6</td>
<td>90.7</td>
<td>38.0</td>
</tr>
<tr>
<td>NONGY+</td>
<td>38.9</td>
<td>80.0</td>
<td>7.1</td>
</tr>
<tr>
<td>NONGY-</td>
<td>41.4</td>
<td>63.7</td>
<td>7.5</td>
</tr>
<tr>
<td>HETQ1</td>
<td>20.5</td>
<td>35.7</td>
<td>5.4</td>
</tr>
<tr>
<td>NONGQ-</td>
<td>25.8</td>
<td>14.7</td>
<td>7.2</td>
</tr>
<tr>
<td>NONGQ-</td>
<td>52.9</td>
<td>99.3</td>
<td>6.5</td>
</tr>
<tr>
<td>MISQ1</td>
<td>30.1</td>
<td>60.4</td>
<td>6.3</td>
</tr>
<tr>
<td>MISQ2</td>
<td>5.8</td>
<td>7.2</td>
<td>5.3</td>
</tr>
</tbody>
</table>
Table 4. Specification Tests for Earnings Data

<table>
<thead>
<tr>
<th></th>
<th>OLS=WLS</th>
<th>OLS=2STEP</th>
<th>WLS=2STEP</th>
</tr>
</thead>
<tbody>
<tr>
<td>all coeff.’s</td>
<td>49.5</td>
<td>17.5</td>
<td>34.1</td>
</tr>
<tr>
<td>(10 d.f.)</td>
<td>(.0000)</td>
<td>(.0649)</td>
<td>(.0002)</td>
</tr>
<tr>
<td>age dummies</td>
<td>43.4</td>
<td>16.5</td>
<td>30.5</td>
</tr>
<tr>
<td>(8 d.f.)</td>
<td>(.0000)</td>
<td>(.0360)</td>
<td>(.0002)</td>
</tr>
<tr>
<td>year &amp; year$^2$</td>
<td>5.07</td>
<td>0.2</td>
<td>5.4</td>
</tr>
<tr>
<td>(2 d.f.)</td>
<td>(.0792)</td>
<td>(.9007)</td>
<td>(.0665)</td>
</tr>
</tbody>
</table>

p-values are reported in parentheses.
Figure 2

RMSE Ratios of Pretest Estimators
RMSE(2step)/RMSE(WLS)
Figure 3

t-stat rejection %
5% nominal size

![Graph showing t-stat rejection percentage with different methods: OLS, WLS-pt, 2step-pt, WLS, and 2step. The x-axis represents gam-Y values ranging from 0 to 0.05, and the y-axis represents % rejections ranging from 0 to 35. The graph illustrates the performance of each method across varying gam-Y values.](image-url)
Figure 4

Bias test rejection rates
5% nominal size

% rejections

0 20 40 60 80 100

0 0.01 0.02 0.03 0.04 0.05

gam-Y

2step

OLS/WLS
Figure 5

Bias test rejection rates
5% nominal size

% rejections

OLS/WLS

2step

sig-Q
Figure 6

bias test rejection rates
5% nominal size

% rejections

2step

OLS/WLS

gam-G
Figure 7

Return to University Education
Age 25-29

Predicted Return (%) vs Year

OLS
2STEP
WLS
Figure 8

Return to University Education
year 1981

predicted return (%)

25 30 35 40 45 50 55 60 65
age
Recent McMaster University Economics Working Papers

(To obtain copies, write to: Secretary, Working Papers, Department of Economics, McMaster University, Hamilton, Ontario, Canada L8S 4M4. A charge of $3 per paper will be levied on orders from institutions that do not have an arrangement for the exchange of working papers. Orders from individuals will be met free of charge, supplies permitting.)

No. 95-09  Emissions Trading with Shares and Coupons when Control over Discharges is Uncertain
            Rob Godby, Stuart Mestelman, R. Andrew Muller and Douglas Welland

No. 95-10  Strategic Manipulation of Pollution Permit Markets: An Experimental Approach
            Jamie Brown-Kruse, Steven R. Elliott and Rob Godby

            Peter Kuhn and A. Leslie Robb

No. 95-12  Children and Demand: Direct and Non-Direct Effects
            Martin Browning and Valerie Lechene

No. 96-01  Canadian Wage Inequality over the Last Two Decades
            J.B. Burbidge, L. Magee, and A.L. Robb

No. 96-02  The Evolution of Welfare Participation Among Canadian Lone Mothers From 1973 - 1991
            Martin D. Dooley

No. 96-03  T.B.A.

No. 96-04  Unemployment Insurance Benefit Levels and Consumption Changes
            Martin Browning and Thomas Crossley

No. 96-05  Markups and the Seasonal Cycle: Testing Collusive Pricing Under Certainty Based on the Seasonal Cycle
            Alok Johri

No. 96-06  Communication in a Common Pool Resource Environment with Probabilistic Destruction
            Andrew Muller and Michelle Vickers

No. 96-07  Estimating Regression Models Using Survey Sample Weights
            L. Magee, A.L. Robb and J.B. Burbidge
<table>
<thead>
<tr>
<th>Number</th>
<th>Title</th>
<th>Author(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>308</td>
<td>Institutional Holding and Trading Volume Reactions to Quarterly Earnings Announcements</td>
<td>J-B. Kim, I. Krinsky, J. Lee</td>
</tr>
<tr>
<td>309</td>
<td>Demographic Change and the Cost of Publicly Funded Health Care</td>
<td>F.T. Denton, B.G. Spencer</td>
</tr>
<tr>
<td>310</td>
<td>Fertility, Age Distribution, and the Production Function</td>
<td>F.T. Denton, D.C. Mountain, B.G. Spencer</td>
</tr>
<tr>
<td>313</td>
<td>Earnings Announcements and the Components of the Bid-Ask Spread</td>
<td>I. Krinsky, J. Lee</td>
</tr>
<tr>
<td>314</td>
<td>Institutional Holdings and Trading Volume Reactions to Quarterly Earnings Announcements</td>
<td>J.-B. Kim, I. Krinsky, J. Lee</td>
</tr>
<tr>
<td>315</td>
<td>A Quadratic Almost Ideal Demand System Estimated with Pooled Regional Time Series: Approximate Aggregation with an Accounting for Age, Cohort, and Trend Effects</td>
<td>F.T. Denton, D.C. Mountain, B.G. Spencer</td>
</tr>
</tbody>
</table>