Finance and Economics Discussion Series
Divisions of Research & Statistics and Monetary Affairs
Federal Reserve Board, Washington, D.C.

Lack of Signal Error (LoSE) and Implications for OLS Regression: Measurement Error for Macro Data

Jeremy J. Nalewaik

2008-15

Lack of Signal Error (LoSE) and Implications for OLS Regression: Measurement Error for Macro Data

Jeremy J. Nalewaik*

March 13, 2008

Abstract

This paper proposes a simple generalization of the classical measurement error model, introducing new measurement errors that subtract signal from the true variable of interest, in addition to the usual classical measurement errors (CME) that add noise. The effect on OLS regression of these lack of signal errors (LoSE) is opposite the conventional wisdom about CME: while CME in the explanatory variables causes attenuation bias, LoSE in the dependent variable, not the explanatory variables, causes a similar bias under some conditions. In addition, LoSE in the dependent variable shrinks the variance of the regression residuals, making inference potentially misleading. The paper provides evidence that LoSE is an important source of error in US macroeconomic quantity data such as GDP growth, illustrates downward bias in regressions of GDP growth on asset prices, and provides recommendations for econometric practice.

---

*Board of Governors of the Federal Reserve System, 20th Street and Constitution Avenue NW, Washington, DC 20551. Telephone: 1-202-452-3792. Fax: 1-202-872-4927. E-mail: jeremy.j.nalewaik@frb.gov. Thanks to Katherine Abraham, Miriam Feffer, Charles Fleischman, Michael Kiley, David Lebow, Richard Lyons, Claudia Sahm, Jonathan Millar, Rob Vigfusson, and seminar participants at the Board of Governors of the Federal Reserve System for comments. The views expressed in this paper are solely those of the author.

1 Introduction

This paper proposes a simple generalization of the classical measurement error model and studies its implications for ordinary least squares (OLS) regression. The usual model starts with the true variable of interest and *adds noise*, which we call the classical measurement error (CME) term; see Fuller (1987) or virtually any econometrics textbook. The generalization discussed here incorporates a different kind of measurement error that *subtracts signal* from the true variable; this new error term is called the Lack of Signal Error, or LoSE for short. This additional term adds some much-needed flexibility to the classical measurement error model: it allows the mismeasured variable to have either more or less variance than the true variable of interest, in contrast to the classical model which imposes that the mismeasured variable have greater variance. This restriction does not hold in some important applications in macroeconomics and elsewhere.

The implications of LoSE for OLS regression are opposite the usual intuition about measurement error, which is applicable to CME only. The CME intuition says that measurement error in the dependent variable Y of a regression poses no real problems for standard estimation and inference. Parameter estimates are unbiased and consistent, while hypotheses are more difficult to reject because CME increases the variance of regression residuals and parameter estimates; inference under these circumstances has a cautious slant. CME in the explanatory variables X causes the real problems for OLS regression, namely attenuation bias and inconsistency. However with LoSE these results are reversed. For the baseline case considered here, LoSE in the *explanatory variables* X produces no bias or inconsistency, and similar to CME in Y, LoSE in X boosts the variance of regresson residuals and hence standard errors. And, LoSE in the *dependent variable* Y introduces an attenuation-type bias and inconsistency into the regression under some circumstances (in particular, when the explanatory variables contain some signal missing from the dependent variable). In addition, LoSE in Y shrinks the variance of the regression residuals, thus shrinking parameter standard errors compared to what they would be without this type of mismeasurement. Standard errors are potentially misleadingly small, complicating inferences about the relation between true Y and X. Under these conditions, a more cautious approach to inference than has been taken in the past may be warranted.

Mismeasurement in many types of data used for empirical work in economics and other disciplines may be better described by the generalized measurement error model with LoSE than by the pure classical measurement error model. This paper focuses on US macroeconomic quantity data such as gross domestic product (GDP) and gross domestic income (GDI), which attempt to measure the same underlying concept using different source data; see Fixler and Nalewaik (2007) and Nalewaik (2007a). These data pass through numerous revisions, and the more poorly-measured initial estimates have less variance than the revised estimates, providing a concrete example of measurement error that cannot be CME; see also Mankiw and Shapiro (1986). Section 2 of the paper makes this point while motivating the generalization of the CME model outlined here.

Section 3 of the paper discusses the nature of the source data used to compute US macroeconomic quantity data, and points out some reasons why LoSE may be present in the estimates even after they have passed through all their revisions. The fact that GDP growth should equal GDI growth, but does not in the fully revised quarterly or annual frequency data, proves that some mismeasurement remains in either GDP or GDI growth. The evidence in Fixler and Nalewaik (2007) and Nalewaik (2007a,b) supports the notion that this mismeasurement is largely LoSE, with more LoSE in GDP growth than in GDI growth. Those arguments are recounted briefly in this section, and some simple calculations comparing GDP and GDI growth show that the LoSE in GDP growth is likely substantial: after 1984, at least 30% of the variance of the true growth rate of the economy appears to be missing.

Some of the implications of substantial LoSE in GDP growth are fairly obvious. Realizations of GDP growth are simply less informative about the true growth rate of the economy than many macroeconomists currently believe, given the common but incorrect presumption that fully-revised GDP growth is measured with little error. In a macro forecasting context, true forecast errors are larger, on average, than forecast errors computed using data mismeasured with LoSE. The implications for regression estimation are more complicated, and working out those implications is the main contribution of the paper. This is done in section 4, which illustrates the effect of LoSE in the dependent and explanatory variables under different sets of assumptions. For concreteness, examples of popular regressions from macroeconomics that may conform to each set of assumptions are provided.

In a wide variety of econometric specifications employed in the macroeconomics and finance, variables like GDP growth, investment growth and consumption growth are regressed on asset prices - interest rates, stock price changes, exchange rate changes, etc. These regressions are of particular interest because asset prices potentially capture some signal missing from the mismeasured quantities, implying attenuation-type biases in the coefficients. Section 4 tests for these biases, regressing different output growth measures contaminated with more or less LoSE on a fixed set of stock or bond prices. When the dependent variable is contaminated with more LoSE, the regression coefficients are smaller, and the differences across regressions are often statistically significant. For example, the coefficients increase when we switch the dependent variable from the early GDP growth estimates based on limited source data to later GDP growth estimates based on more-comprehensive data. Tellingly, the coefficients increase again when we switch the dependent variable from GDP growth to GDI growth. The hypothesis that measurement error in the dependent variable does not bias OLS regression coefficients, a core piece of conventional wisdom in the profession, is rejected by the data, just as the paper predicts if the measurement error is LoSE. Section 5 concludes the paper.

2 A Generalization of the Classical Measurement Error Model

Let $Y_t^{\star}$ be the true value of the variable of interest, while Z_t is an $(1 \times l)$ vector of possibly stochastic variables used to construct $Y_t^{\star}$. The mismeasured estimate of the variable of interest is Y_t. In many cases a government statistical agency or some other organization computes Y_t based on information from surveys, administrative records, and other data sources (source data for short); then Z_t will be variables drawn from the source data, possibly including non-linear functions of the original source data.

Under the classical measurement error model,

$$Y_t = Y_t^{\star} + \varepsilon_t.$$

The term ε_t is "noise" or the classical measurement error (CME) in the estimate. In the current context this is taken to imply independence of ε_t and $Y_t^{\star}$, although the weaker assumption $\text{cov}\left(Y_t^{\star}, \varepsilon_t\right) = 0$ suffices for many purposes. The CME may arise from estimation errors or other sources; since many estimates Y_t are based on surveys, survey sampling errors are often thought to be a source of CME.

Under the generalized model of mismeasurement considered here, the mismeasured estimate Y_t is as in Fixler and Nalewaik (2007):

$$(1) \qquad Y_t = E\left(Y_t^{\star}|Z_t\right) + \varepsilon_t.$$

The CME term ε_t is assumed independent of Z_t and $Y_t^{\star}$. It can be seen immediately that the classical measurement error model is a special case of this more general model, where Z_t spans $Y_t^{\star}$ so $E\left(Y_t^{\star}|Z_t\right) = Y_t^{\star}$.

Define the deviation of the variable of interest from its conditional expectation as:

$$(2) \qquad \zeta_t = Y_t^{\star} - E\left(Y_t^{\star}|Z_t\right).$$

This deviation represents the information about $Y_t^{\star}$ not contained in Z_t, and is independent of all functions of Z_t. With $\operatorname{cov}\left(E\left(Y_t^{\star}|Z_t\right), \zeta_t\right) = 0$, the variance of the true variable of interest may be decomposed into the variance of the conditional expectation plus the variance of ζ_t, and: $\operatorname{var}\left(\zeta_t\right) = \operatorname{var}\left(Y_t^{\star}\right) - \operatorname{var}\left(E\left(Y_t^{\star}|Z_t\right)\right)$. The variance of ζ_t represents the variance of the information about $Y_t^{\star}$ missing from the conditional expectation. Substituting into (1):

$$(3) \qquad Y_t = Y_t^{\star} - \zeta_t + \varepsilon_t.$$

Thinking of ε_t as mismeasurement from noise, ζ_t represents an opposite kind of mismeasurement, mismeasurement from lack of signal about $Y_t^{\star}$ in the information used to construct Y_t. As such, ζ_t may be labelled the Lack of Signal Error, or LoSE for short.

Since the CME is independent of $Y_t^\star$, it is naturally independent of the LoSE as well. However the LoSE is clearly correlated with $Y_t^\star$, with $\operatorname{cov}(Y_t^\star, \zeta_t) = \operatorname{var}(\zeta_t)$ in fact. Taking variances of (3):

$$\begin{aligned} \operatorname{var}(Y_t) &= \operatorname{var}(Y_t^\star) + \operatorname{var}(\zeta_t) - 2\operatorname{cov}(Y_t^\star, \zeta_t) + \operatorname{var}(\varepsilon_t) \\ &= \operatorname{var}(Y_t^\star) - \operatorname{var}(\zeta_t) + \operatorname{var}(\varepsilon_t)\,. \end{aligned} \tag{4}$$

Depending on whether the variance of the LoSE is greater than or less than the variance of the CME, the variance of the estimate Y_t may be greater than or less than the variance of the true variable of interest $Y_t^\star$. With CME alone, the variance of the estimate must exceed the variance of the true variable. The key limitation of the CME model is the assumption that $\operatorname{cov}(Y_t - Y_t^\star, Y_t^\star) = 0$; the generalization allows this covariance to range from 0 to a lower bound of minus $\operatorname{var}(Y_t - Y_t^\star)$, in which case all the mismeasurement arises from LoSE.

The restrictions imposed by the CME model on variances and covariances have obvious drawbacks, as it is easy to think of hypothetical and actual counterexamples. As a hypothetical counterexample, assume that $Y_t^\star$ has positive variance, while the estimate Y_t is just a constant for all t. The estimate Y_t is clearly mismeasured, but the classical measurement error model cannot handle this case. The growth rates of US GDP and GDI provide an actual counterexample. These estimates pass through numerous revisions that plausibly reduce measurement error, since they incorporate more comprehensive and higher-quality source data. For example, suitable source data is simply unavailable for many components of the "advance" current quarterly GDP estimate released about a month after the quarter ends. Source data for some of those components is incorporated into the revised "final" current quarterly estimate released about two months later, and higher-quality data are incorporated at subsequent annual and benchmark revisions, likely bringing the estimate closer to its true value.[1] Then an early estimate of GDP growth or GDI growth can be modelled as a later revised estimate (the analog to $Y_t^\star$) plus a measurement error term, which disappears with revision. Table 1 shows that the initial estimates have less variance than the revised estimates, violating the variance restrictions of the CME model.[2] The generalized model here implies that the bulk of the measurement error is LoSE, as noted by Mankiw and Shapiro (1986).

While the generalized model here is less restrictive than the CME model, some restrictions do remain. Writing:

$$Y_t + \zeta_t = Y_t^\star + \varepsilon_t, \tag{5}$$

the independence of ζ_t from Y_t is a restriction, implied by the first term in (1) being a conditional expectation. However systematic biases in the estimate, on top of those caused by noise,[3] violate this assumption. For concreteness, assume $E\left(Y_t^\star|Z_t\right) = Z_t\gamma$.

[1] For more on revisions to GDP, see Grimm and Weadock (2006). An estimate of GDI growth is not released at the time of the "advance" GDP estimate because of data limitations, but GDI is always released at the time of the "final" current quarterly estimate. For GDI, annual revisions incorporate information from administrative and tax records that is much more comprehensive than the samples used to compute the "final" current quarterly estimates.

[2] These are annualized quarterly growth rates. Each quarterly observation in the "advance" or "final" time series is the "advance" or "final" estimate for that quarter, i.e. the estimate released one or three months after that quarter closes. We end the sample in 2004 so that all observations in our latest available time series have passed through three annual revisions, ensuring each observation is much more heavily revised than the corresponding "advance" or "final" current quarterly observation.

[3] With the noise term, $E\left(Y_t^\star|Y_t\right) \neq Y_t$; the estimate is biased in this sense.

Consider an estimate of $Y_t^\star$ based on Z_t that misuses the information, so $Y_t = Z_t\widetilde{\gamma} + \varepsilon_t$ with $\widetilde{\gamma} \neq \gamma$. The estimate "misses" in a systematic way.[4] Taking variances, we have:

$$\text{var}\,(Y_t) \;\; = \;\; \text{var}\,(Y_t^\star) + \text{var}\,(Z_t\widetilde{\gamma}) - \text{var}\,(Z_t\gamma) - \text{var}\,(\zeta_t) + \text{var}\,(\varepsilon_t)\,.$$

For estimation and inference about $Y_t^\star$ and its relation to other variables, these systematic biases in $Y_t^\star$ lead to biased and inconsistent estimates, in general. However unless additional information is available about the nature of the "misses" $Z_t\widetilde{\gamma} - Z_t\gamma$, the direction and magnitude of these biases is unclear. In some cases this additional information may be available, but in general, an important goal of all creators of data (government statistical agencies as well as other groups) is to avoid such systematic mismeasurement. Indeed, their ultimate goal is probably to produce estimates Y_t that are as close as possible to $E\,(Y_t^\star|Z_t)$, with as broad an information set Z_t as possible given resource constraints.[5] As such, the generalized model (1) is a useful benchmark, and should approximate well the underlying mismeasurement in many situations. It also has the advantage of being mathematically tractable.

Before concluding this section, it is worth emphasizing that Z_t need not be an exhaustive information set - i.e. it need not contain all available relevant pieces of information about $Y_t^\star$. Resource and other constraints certainly preclude this from being the case, and the sections below considering the implications of LoSE allow for this possibility.

[4] A highly stylized example is $Y_t = \alpha_0 + \alpha_1 Y_t^\star + \varepsilon_t$, with $\alpha_0 \neq 0$ and $\alpha_1 \neq 1$.

[5] If that is not their ultimate goal, it should be.

3 Data

3.1 Discussion of U.S. Macro Quantity Data

Each estimated growth rate of a macro quantity such as gross domestic product (GDP) is an attempt at measuring the change in the value of *all* relevant economic transactions, in the *entire* economy, from one fixed time period to the next. For an entity as large as the U.S. economy, this is a daunting, almost mind-boggling task, as the number of transactors and transactions is typically enormous, with little or no information recorded about many of them at high frequencies. Attempts to measure changes in these macro quantities are much more ambitious than attempts to measure similar changes for a single person, household, or even company. Simply due to their broad, universal nature, estimates of macro quantities are likely to miss more information - i.e. be contaminated with more LoSE - than are estimates of micro quantities (although some micro data sources may be contaminated with LoSE as well).[6]

Of course, the nature of the available source data determines the information content of macro quantity growth rates of interest, and frequency is important in this regard in the case of data from the U.S. National Income and Product Accounts (NIPA).[7] The

[6]For example, it has long been suspected that management of many publicly traded corporations "smooth" quarterly earnings to meet their guidance (prior estimates of what their earnings would be). Such spurious reductions in the variability of measured earnings growth should effectively add LoSE to those measures.

[7]The growth rates of real quantities are of interest in most economics applications. In the NIPAs, real quantities are typically estimated by gathering the appropriate nominal source data and the appropriate price indexes, and then deflating the former with the latter. The discussion of LoSE in source data here focuses on the nominal source data, but there may exist significant LoSE stemming from the price indexes as well. Measured price indexes may miss fluctuations in the quality of goods, from either the introduction of new goods or modifications of existing goods; see Bils and Klenow (2001) and Bils (2004). The length of their time series is short, but Broda and Weinstein (2007) do provide some evidence that product creation (and hence quality improvement embedded in new products) is pro-cyclical, implying

most comprehensive data on GDP and other major NIPA aggregates are only available at the quinquennial frequency (every five years), at the time of the major economic censuses. Even then, resource constraints make true census counts impossible. Many transactions in the underground economy remain unobserved and must be estimated, and some "above-ground" transactions are simply missed by any census.[8] At the annual frequency, the GDP source data are typically samples drawn from the census universe. These samples can be quite large, capturing a sizeable fraction of the relevant value of transactions, but they are typically skewed towards measuring the transactions of larger businesses. As such, they may miss variation arising from the transactions of small companies and from businesses starting, shutting down, and operating in the underground economy. The lack of representation of these segments of the economy may add or substract variance to the official estimates, depending on the relative variance of the non-measured segments and their covariance with the measured segments, but this mismeasurement has the potential to add some LoSE to the data. At the annual frequency, and also at the quarterly frequency to some extent, government tax and administrative records are used as an additional source of information about the value of transactions, especially on the income (GDI) side of the accounts. These data can be informative, but underreporting makes them less than fully comprehensive.

At the quarterly and monthly frequency,[9] reliance on samples is more pronounced,

---

counter-cyclical variation in prices. If standard prices indexes miss this counter-cyclical variation, real quantities deflated by these indexes may not be pro-cyclical enough or variable enough.

[8] In this regard, it should be noted that the Bureau of Labor Statistics and the Census Bureau each maintain a list which attempts to track the entire universe of business establishments in the US, from which each agency draws samples. A 1994 comparison of the two lists found a non-trivial number of non-matches - establishments on one list but not the other.

[9] Treatment of seasonality immediately becomes a major issue when moving to frequencies higher than annual, and identification of the seasonal patterns of interest, the "true" seasonal factors, can be

and the samples are less comprehensive. Smaller samples introduce larger sampling errors, which have traditionally been thought of as introducing CME into the estimates. The samples are typically random, after all, so part of the difference between the population and sample moments is likely random variation uncorrelated with the variation in the population moments. However smaller samples may introduce some LoSE as well, since smaller samples are simply less informative than larger samples: when different segments of the economy behave quite differently, small samples which are not fully representative may miss variation arising from some segments.[10] In addition, usable data on the value of transactions at a frequency higher than annual is simply unavailable for a substantial share of some NIPA aggregates such as GDP; many of the services categories of personal consumption expenditures (PCE) lack usable source data, for example.[11] Quarterly and monthly growth rates are typically interpolated from annual totals, or estimated as "trend extrapolations." The lack of hard information for these categories must introduce some LoSE into the quarterly and monthly estimates.

3.2 Evidence of LoSE from GDP and GDI growth

Some users of quarterly or annual US NIPA data take the view that the variance of CME is negligible after the data have passed through its sequence of revisions, particularly

tenuous; see Watson, 1987. Seasonal adjustment programs are all essentially smoothing algorithms, and as such they risk introducing LoSE into the data.

[10]Samples for which topcodes are binding by definition miss variation arising from the top-coded units. The samples used in the construction of the U.S. NIPA data generally are not top-coded, but analysts at the Bureau of Economic Analysis (BEA) do look at very detailed categories of data and trim outliers, which may have an effect similar to topcoding.

[11]This situation has begun to change, with the introduction of the Quarterly Services Survey (QSS) in 2002, but so far the BEA uses the QSS for a relatively small share of PCE services.

for the more highly aggregated NIPA quantities like GDP growth where any remaining CME variance in its subcomponents may be diminished by averaging. Absent knowledge of the possible existence of LoSE, this view would imply that the variance of overall mismeasurement is close to zero, since $\text{var}\,(Y_t - Y_t^{\star}) = \text{var}\,(\varepsilon_t)$ in the pure CME model.

However there is ample evidence that the variance of overall mismeasurement in the most aggregated NIPA aggregate, GDP, is not close to zero, especially since the mid-1980s. GDI is an alternative estimate of the same quantity, so examining the relation between GDP and GDI provides some direct evidence on mismeasurement. Table 2 shows that, prior to 1984, the variance of each estimate is close to the covariance between the two, for both annual and annualized quarterly growth rates. The two estimates diverge very little, providing little direct evidence of mimeasurement. However after 1984, when the variance of both estimates drops dramatically (see McConnell and Perez-Quiros (2000)), the correlation between the estimates also falls, as the covariance falls relative to the variances on average. This is especially true for the quarterly growth rates, where the correlation falls from 0.93 to 0.68.[12] Interestingly, the variance of GDI growth also increases relative to the variance of GDP growth. Under the generalized CME model of section 2, this relatively large GDI variance may stem from some combination of two possible sources: (1) a relatively large amount of CME in GDI growth, boosting its variance, and (2) a relatively large amount of LoSE in GDP growth, damping its variance. The evidence favors the latter as the more important source of mismeasurement.

First, consider the results in Nalewaik (2007a), who estimates a two-state bivariate

[12]At the annual frequency, the correlation falls from 0.98 to 0.94; the decline in smaller at this frequency primarily because the variance of GDP growth falls below its covariance with GDI growth. This cannot happen in either the pure CME model or the generalization favored here - i.e. the variance of each estimate must be larger than their covariance; see Fixler and Nalewaik (2007). Given that only 20 observations are employed to compute these variances, this may be a small-sample estimation issue.

Markov switching model where mean GDP and GDI growth switch with the state; the low-growth states identified by the model encompass recessions as defined by the NBER. The conditional variance of GDI in that model, conditional on the estimated state of the world, is actually slightly *lower* than the conditional variance of GDP, despite the fact that the unconditional variance of GDI growth is higher. The higher unconditional variance stems from GDI growing faster than GDP in high-growth periods and slower than GDP in slow-growth periods in and around recessions. In other words, GDI growth appears to contain more signal about the state of the world than GDP growth: the larger spread between its high- and low-growth means implies greater informativeness about the state. Greater signal in GDI growth implies some LoSE in GDP growth, relatively more than in GDI growth.

Second, table 1 shows that the variance of GDI growth becomes relatively large only after the data pass through annual and benchmark revisions; in the earlier current quarterly estimates, the variance of GDP growth actually exceeds the variance of GDI growth. Since the revisions plausibly bring estimated GDI growth closer to the truth, they must either reduce LoSE, which would increase its variance, or reduce CME, which would decrease its variance. The increase in variance from the revisions, then, is likely a decrease in LoSE, suggesting the relatively large variance of revised GDI stems from relatively less LoSE. The annual and benchmark revisions appear to add more signal to GDI growth than GDP growth, increasing the variance of GDI growth relative to GDP growth.[13] Fixler and Nalewaik (2007) discuss the revisions evidence in more detail,

[13] The results in Nalewaik (2007b) support this interpretation of the revisions. Using the Markov switching model in Nalewaik (2007a), Nalewaik (2007b) shows that the revisions increase mean GDI growth in the high-growth state and reduce mean GDI growth in the low-growth state, effectively increasing its informativeness about the state of the economy. The revisions increase the gap between the high- and low-growth means for GDP growth as well, but the increase is not as large as the increase

testing the hypothesis that the idiosyncratic variation of GDI growth is purely CME and rejecting at conventional significance levels. This again implies some LoSE in GDP growth.

To get a sense of the magnitude of the potential variance missing from GDP growth due to LoSE, assume that the CME variance in each estimate is negligible, so the differences between GDP and GDI growth stem entirely from differential LoSE:

$$\begin{aligned}\Delta Y_t^{GDP} &= E\left(\Delta Y_t^{\star} | Z_t^{GDP}\right) &= \Delta Y_t^{\star} - \zeta_t^{GDP}, \quad \text{and:} \\ \Delta Y_t^{GDI} &= E\left(\Delta Y_t^{\star} | Z_t^{GDI}\right) &= \Delta Y_t^{\star} - \zeta_t^{GDI}.\end{aligned}$$

Taking variances as in (4) yields:

$$\begin{aligned}\text{var}\left(\Delta Y_t^{GDP}\right) &= \text{var}\left(\Delta Y_t^{\star}\right) - \text{var}\left(\zeta_t^{GDP}\right), \\ \text{var}\left(\Delta Y_t^{GDI}\right) &= \text{var}\left(\Delta Y_t^{\star}\right) - \text{var}\left(\zeta_t^{GDI}\right), \quad \text{and the covariance is:} \\ \text{cov}\left(\Delta Y_t^{GDP}, \Delta Y_t^{GDI}\right) &= \text{var}\left(\Delta Y_t^{\star}\right) - \text{var}\left(\zeta_t^{GDP}\right) - \text{var}\left(\zeta_t^{GDI}\right) + \text{cov}\left(\zeta_t^{GDP}, \zeta_t^{GDI}\right).\end{aligned}$$

The idiosyncratic variance of one estimate (its variance minus its covariance with the other estimate) is then proportional to the LoSE in the other estimate:

$$\begin{aligned}\text{var}\left(\Delta Y_t^{GDP}\right) - \text{cov}\left(\Delta Y_t^{GDP}, \Delta Y_t^{GDI}\right) &= \text{var}\left(\zeta_t^{GDI}\right) - \text{cov}\left(\zeta_t^{GDP}, \zeta_t^{GDI}\right), \quad \text{and:} \\ \text{var}\left(\Delta Y_t^{GDI}\right) - \text{cov}\left(\Delta Y_t^{GDP}, \Delta Y_t^{GDI}\right) &= \text{var}\left(\zeta_t^{GDP}\right) - \text{cov}\left(\zeta_t^{GDP}, \zeta_t^{GDI}\right).\end{aligned}$$

The information missed by both estimates is $\text{cov}\left(\zeta_t^{GDP}, \zeta_t^{GDI}\right)$; the idiosyncratic variance

for GDI growth.

of GDI growth is then the variance of all the information about $\Delta Y_t^\star$ missing from measured GDP growth minus the part of that information also absent from GDI growth. Rearranging the covariance provides a lower bound on the variance of true GDP growth $\Delta Y_t^\star$:

$$\begin{aligned}
\mathrm{var}\left(\Delta Y_t^\star\right) &= \mathrm{cov}\left(\Delta Y_t^{GDP}, \Delta Y_t^{GDI}\right) \\
&\quad + \left(\mathrm{var}\left(\zeta_t^{GDI}\right) - \mathrm{cov}\left(\zeta_t^{GDP}, \zeta_t^{GDI}\right)\right) \\
&\quad + \left(\mathrm{var}\left(\zeta_t^{GDP}\right) - \mathrm{cov}\left(\zeta_t^{GDP}, \zeta_t^{GDI}\right)\right) \\
&\quad + \mathrm{cov}\left(\zeta_t^{GDP}, \zeta_t^{GDI}\right), \quad \text{so:} \\
&> \mathrm{cov}\left(\Delta Y_t^{GDP}, \Delta Y_t^{GDI}\right) \\
&\quad + \left(\mathrm{var}\left(\zeta_t^{GDI}\right) - \mathrm{cov}\left(\zeta_t^{GDP}, \zeta_t^{GDI}\right)\right) \\
&\quad + \left(\mathrm{var}\left(\zeta_t^{GDP}\right) - \mathrm{cov}\left(\zeta_t^{GDP}, \zeta_t^{GDI}\right)\right).
\end{aligned}$$

The last column of table 2 uses this equation to set an upper bound on the fraction of variance of $Y_t^\star$ captured by measured GDP growth: $\frac{\mathrm{var}\left(\Delta Y_t^{GDP}\right)}{\mathrm{var}\left(\Delta Y_t^\star\right)}$. Measured GDP growth captures at most 70% of the variation in $\Delta Y_t^\star$ after 1984, under the assumption of negligible CME. Of course the assumption of no noise is an extreme one, particularly for the quarterly estimates. Indeed, the evidence in section 4.3.1 from regressions involving GDP growth, GDI growth, and stock prices indicate about a quarter of the variance of GDP growth is noise, but these results actually tighten the upper bound, decreasing it from 70% to 64%. And this is in fact an upper bound, since it does not account for $\mathrm{cov}\left(\zeta_t^{GDP}, \zeta_t^{GDI}\right)$, the variation in $\Delta Y_t^\star$ missed by both measured GDP and GDI growth. The variation missed by both estimates at the quarterly frequency could be substantial.

Going forward, if these post-1984 variances and covariances are the norm, the implications of a potentially non-trivial amount of LoSE in macro data such as GDP growth should be taken seriously. For estimation and inference, the post-1984 portion of many samples will become increasingly large and important. The next section explores the implications of LoSE for estimation and inference using the most ubiquitous tool in econometrics: OLS regression.

4 Implications for OLS Estimation

Consider ordinary least squares estimation of the relation between a mismeasured variable Y_t and a $(1 \times k)$ set of mismeasured explanatory variables X_t, using a sample of length T. When stacking together the observations, time subscripts are dropped for convenience:

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_T \end{pmatrix}; \quad X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_T \end{pmatrix}.$$

Our full set of assumptions is as follows:

Assumption 1 *$Y_t^\star = X_t^\star \beta + U_t^\star$. $U_t^\star$ is i.i.d., mean zero, with $\operatorname{var}(U_t^\star) = \sigma^2_{U^\star}$ and $U_s^\star$ independent of $X_t^\star$, $\forall t, s$. Y_t follows the generalized measurement error model of section 2: $Y_t = E\left(Y_t^\star | Z_t^y\right) + \varepsilon_t$. The CME ε_t is i.i.d., mean zero, and independent of all conditioning information sets, with $\operatorname{var}(\varepsilon_t) = \sigma^2_\varepsilon$. The LoSE $\zeta_t = \left(X_t^\star - E\left(X_t^\star | Z_t^y\right)\right)\beta + U_t^\star - E\left(U_t^\star | Z_t^y\right) = \zeta_t^{xy}\beta + \zeta_t^u$. ζ_t^u is i.i.d. and mean zero with $\operatorname{var}(\zeta_t^u) = \sigma^2_{\zeta,u}$, and ζ_t^{xy} is i.i.d. and mean zero with $\operatorname{var}(\zeta_t^{xy}) = \sigma^2_{\zeta,xy}$, a $k \times k$ matrix. X_t follows the generalized* *measurement error model of section 2:* $X_t = E(X_t^\star | Z_t^x) + \varepsilon_t^x$. *The CME* ε_t^x *is i.i.d., mean zero, independent of* ε_t *and all conditioning information sets, with* $\operatorname{var}(\varepsilon_t^x) = \sigma_{\varepsilon,x}^2$, *a* $k \times k$ *matrix. The LoSE* $\zeta_t^x = X_t^\star - E(X_t^\star | Z_t^x)$ *is i.i.d. and mean zero with* $\operatorname{var}(\zeta_t) = \sigma_{\zeta,x}^2$, *also a* $k \times k$ *matrix.* $\frac{1}{T}(X^\star)' X^\star \xrightarrow{p} Q_{xx}$, $\frac{1}{T}(E(X^\star|Z^y))' E(X^\star|Z^y) \xrightarrow{p} Q_{xx} - \sigma_{\zeta,xy}^2 = Q_{xx}^{zy}$, $\frac{1}{T}(E(X^\star|Z^x))' E(X^\star|Z^x) \xrightarrow{p} Q_{xx} - \sigma_{\zeta,x}^2 = Q_{xx}^{zx}$, $\frac{1}{T}(E(X^\star|Z^y))' E(X^\star|Z^x) \xrightarrow{p} Q_{xx}^{zb}$, *and* $\frac{1}{T}X'X \xrightarrow{p} = Q_{xx}^{zx} + \sigma_{\varepsilon,x}^2$. *All relevant fourth moments exist.*

Then Y_t can be written as:

$$
\begin{aligned}
Y_t &= E(X_t^\star | Z_t^y)\beta + E(U_t^\star | Z_t^y) + \varepsilon_t \\
&= X_t\beta + (E(X_t^\star | Z_t^y) - X_t)\beta + E(U_t^\star | Z_t^y) + \varepsilon_t \\
&= X_t\beta + (E(X_t^\star | Z_t^y) - E(X_t^\star | Z_t^x) - \varepsilon_t^x)\beta + U_t^\star - \zeta_t^u + \varepsilon_t.
\end{aligned} \tag{6}
$$

The OLS regression estimator is:

$$
\begin{aligned}
\widehat{\beta} &= (X'X)^{-1} X'Y \\
&= \beta + (X'X)^{-1} X' ((E(X^\star|Z^y) - E(X^\star|Z^x) - \varepsilon^x)\beta + U^\star - \zeta^u + \varepsilon).
\end{aligned} \tag{7}
$$

Consider the sources of bias and inconsistency in this estimate. It is well known that the CME in Y introduces no bias and inconsistency, since ε is independent of X. Interestingly, the LoSE in $U^\star$ introduces no bias or inconsistency either: the independence of $U^\star$ from $X^\star$ implies the independence of $E(U^\star|Z^y) = U^\star - \zeta^u$ from $E(X^\star|Z^x)$[14] and

[14] Write the joint distribution of $U^\star$, $X^\star$, Z^y and Z^x as $f(U^\star, X^\star, Z^y, Z^x)$. Since $X^\star$ and $U^\star$ are independent, this distribution can be factored into two marginal distributions, one containing $X^\star$ and the other containing $U^\star$. This implies a parition of Z^y and Z^x between the two marginals: let Z_x^y be the set of variables in Z^y that influence $X^\star$ and are independent of $U^\star$, let Z_u^y be the set of variables in Z^y that influence $U^\star$ and are independent of $X^\star$, and without loss of generality let all variables in

hence $X = E(X^\star|Z^x) + \varepsilon^x$. The other components in the error of (6) do cause bias and inconsistency; taking expectations and probability limits of (6) yields:

$$
\begin{aligned}
(8) \quad E\left(\widehat{\beta}\right) &= \beta + E\left((X'X)^{-1} X' \left(E(X^\star|Z^y) - E(X^\star|Z^x) - \varepsilon^x\right)\right)\beta, \quad \text{and:} \\
(9) \quad \widehat{\beta} &\xrightarrow{p} \beta + \left(Q^{zx}_{xx} + \sigma^2_{\varepsilon,x}\right)^{-1}\left(Q^{zb}_{xx} - Q^{zx}_{xx} - \sigma^2_{\varepsilon,x}\right)\beta.
\end{aligned}
$$

The usual attenuation bias and inconsistency from CME in X is evident here. The additional inconsistency from LoSE depend on the difference between Q^{zb}_{xx} and Q^{zx}_{xx}. Like attenuation bias, these additional biases likely tend towards zero: Q^{zb}_{xx} should be smaller than Q^{zx}_{xx} if Z^y and Z^x contain a substantial amount of non-overlapping (independent) information.

The inconsistency of $\widehat{\beta}$ can be corrected by instrumenting with a $(1 \times m)$ set of instruments W_t, with $m \geq k$, if the instruments meet the following set of assumptions:

Assumption 2 *With* $P_W = W(W'W)^{-1}W'$, $\frac{1}{T}X'P_W X \xrightarrow{p} Q^w_{xx}$, *a positive semi-definite matrix, and* $\frac{1}{T}X'P_W\left(\left(E(X^\star|Z^y) - E(X^\star|Z^x) - \varepsilon^x\right)\beta + U^\star - \zeta^u + \varepsilon\right) \xrightarrow{p} 0$. *All relevant fourth moments exist.*

Valid instruments must be asymptotically independent of the CME in X, a standard condition. However, an additional condition must be met: the instruments must be

---

Z^x be independent of $U^\star$. Then:

$$
\begin{aligned}
f(U^\star, X^\star, Z^y, Z^x) &= f_{UZ}(U^\star, Z^y_u)\, f_{XZ}(X^\star, Z^y_x, Z^x) \\
&= g_U(U^\star|Z^y_u)\, f_{Z_y}(Z^y_u)\, g_{XZ}(X^\star, Z^y_x|Z^x)\, f_{Z_x}(Z^x) \\
&= g_U(U^\star|Z^y_u)\, f_{Z_y}(Z^y_u)\, g_X(X^\star|Z^x)\, g_{Z_x}(Z^y_x|X^\star, Z^x)\, f_{Z_x}(Z^x),
\end{aligned}
$$

factoring f_{UZ}, f_{XZ} and g_{XZ} into conditional and marginal distributions. $E(U^\star|Z^y)$ is an integral of g_U over $U^\star$, while $E(X^\star|Z^x)$ is an integral of g_X over $X^\star$; since each of these two expectations does not depend on the other variable, and only on variables independent of the other variable, they must be independent.

asymptotically independent of $E(X^{\star}|Z^{y}) - E(X^{\star}|Z^{x})$. This condition is met by instruments W that are common to both information sets (if such information exists), so $W \subset Z^{x}$ and $W \subset Z^{y}$, since $W'E(X^{\star}|Z^{y})$ and $W'E(X^{\star}|Z^{x})$ then have the same probability limit. With valid instruments, we have:

$$\begin{aligned} \widehat{\beta} &= \left(X'P_W X\right)^{-1} X'P_W Y \\ &= \beta + \left(X'P_W X\right)^{-1} X'P_W \left(\left(E(X^{\star}|Z^{y}) - E(X^{\star}|Z^{x}) - \varepsilon^{x}\right)\beta + U^{\star} - \zeta^{u} + \varepsilon\right), \end{aligned} \tag{10}$$

and $\widehat{\beta} \xrightarrow{p} \beta$. The asymptotic distribution of the estimator is:

$$\sqrt{T}\left(\widehat{\beta} - \beta\right) \xrightarrow{d} N\left(0, (Q^{w}_{xx})^{-1}\left(\sigma^2_{U^{\star}} - \sigma^2_{\zeta,u} + \sigma^2_{\varepsilon} + \beta'\left(Q^{zy}_{xx} - 2Q^{zb}_{xx} + Q^{zx}_{xx} + \sigma^2_{\varepsilon,x}\right)\beta\right)\right).$$

where $\xrightarrow{d}$ denotes convergence in distribution as $T \longrightarrow \infty$, and $N(a,b)$ is a Gaussian distribution with mean a and variance b. The usual estimator of the variance of the error term, $s^2 = \frac{1}{T}\left(Y - X\widehat{\beta}\right)'\left(Y - X\widehat{\beta}\right)$, converges to the error variance in this asymptotic distribution:

$$\begin{aligned} s^2 &= \frac{1}{T}\left(E(X|Z^{y})\beta + E(U^{\star}|Z^{y}) + \varepsilon - \left(E(X^{\star}|Z^{x}) + \varepsilon^{x}\right)\widehat{\beta}\right)' \\ &\quad * \left(E(X|Z^{y})\beta + E(U^{\star}|Z^{y}) + \varepsilon - \left(E(X^{\star}|Z^{x}) + \varepsilon^{x}\right)\widehat{\beta}\right) \\ &= \frac{1}{T}E(U^{\star}|Z^{y})'E(U^{\star}|Z^{y}) + \frac{1}{T}\varepsilon'\varepsilon + \frac{1}{T}\beta'E(X^{\star}|Z^{y})'E(X^{\star}|Z^{y})\beta \\ &\quad - \frac{1}{T}\beta'E(X^{\star}|Z^{y})'E(X^{\star}|Z^{x})\widehat{\beta} - \frac{1}{T}\widehat{\beta}'E(X^{\star}|Z^{x})'E(X^{\star}|Z^{y})\beta \\ &\quad + \frac{1}{T}\widehat{\beta}'E(X^{\star}|Z^{x})'E(X^{\star}|Z^{x})\widehat{\beta} + \frac{1}{T}\widehat{\beta}'\varepsilon^{x\prime}\varepsilon^{x}\widehat{\beta} + \frac{1}{T}\text{cross terms.} \end{aligned}$$

The first two terms converge in probability to $\sigma^2_{U^{\star}} - \sigma^2_{\zeta,u} + \sigma^2_{\varepsilon}$; the terms involving β and $\widehat{\beta}$ simplify in the limit since $\widehat{\beta} \xrightarrow{p} \beta$; and the cross terms converge in probability to zero. Then: $s^2 \xrightarrow{p} \sigma^2_{U^{\star}} - \sigma^2_{\zeta,u} + \sigma^2_{\varepsilon} + \beta'\left(Q^{zy}_{xx} - 2Q^{zb}_{xx} + Q^{zx}_{xx} + \sigma^2_{\varepsilon,x}\right)\beta$. The next four subsections discuss the most important implications of LoSE in X and Y for the parameter estimates and standard errors, examining some more specialized examples of this general model that highlight the implications of interest.

4.1 X Mismeasured, Y Not Mismeasured: No LoSE Problems

Given the traditional focus on mismeasurement in X on regression estimation, we begin with this subsection making the following assumption, in addition to assumption 1:

Assumption 3 Y_t *is not mismeasured:* $Y_t = Y_t^\star$.

Then (5) simplifies to:

$$\begin{aligned} Y_t^\star &= X_t^\star \beta + U_t^\star \\ &= X_t \beta + \left(X_t^\star - X_t\right)\beta + U_t^\star \\ &= X_t \beta - \varepsilon_t^x \beta + \zeta_t^x \beta + U_t^\star . \end{aligned}$$

Not all of the true variation in $X_t^\star$ appears in X_t due to LoSE, but all of that variation in $X_t^\star$ does appear in $Y_t^\star$ through $X_t^\star \beta$. That variation in $Y_t^\star$ missing from X_t is relegated to the error term of this equation.

The OLS regression estimator in this case is:

$$\begin{aligned} \widehat{\beta} &= \left(X'X\right)^{-1} X'Y \\ &= \beta + \left(X'X\right)^{-1} X' \left(-\varepsilon^x \beta + \zeta^x \beta + U^\star\right). \end{aligned}$$

Since ζ^x is independent of $E\left(X^\star | Z^x\right) + \varepsilon^x = X$, the LoSE in X introduces no bias into $\widehat{\beta}$ in this case. Given assumption 1, $\frac{1}{T} X' \zeta^x \xrightarrow{p} 0$, and the LoSE introduces no inconsistency either. These results clearly hinge on the assumption that the LoSE is the difference between truth and a conditional expectation, and measurement error of a different form, such as the systematic biases discussed at the end of section 2, would lead to biased and inconsistent parameter estimates. The consistency result here also relies on all k explanatory variables being conditioned on the same information set Z^x. Kimball, Sahm, and Shapiro (2007) discuss a related example, where different elements of X are conditioned on different information sets, causing bias and inconsistency.

Of course, the CME in X produces the usual attenuation bias. By way of review, and for comparison with later results:

$$E\left(\widehat{\beta}\right) = \beta - E\left(\left(X'X\right)^{-1} X'\varepsilon^x\right)\beta, \quad \text{and:} \tag{11}$$

$$\widehat{\beta} \xrightarrow{p} \beta - \left(Q_{xx}^{zx} + \sigma_{\varepsilon,x}^2\right)^{-1} \sigma_{\varepsilon,x}^2 \beta. \tag{12}$$

Instruments uncorrelated with the CME in X yield consistent estimates.

To focus more tightly on the implications of LoSE, the remainder of this subsection considers the case of no CME in X:

Assumption 4 $\operatorname{var}\left(\varepsilon_t^x\right) = 0$.

Then $E\left(\widehat{\beta}\right) = \beta$, and $\widehat{\beta} \xrightarrow{p} \beta$. The variation in $X_t^\star$ that appears in $Y_t^\star$ but is missing from X_t shows up in the error term of the regression, increasing the variance of the parameter estimates. We have $\operatorname{var}\left(\widehat{\beta}\right) = E\left(\operatorname{var}\left(\widehat{\beta}|X\right)\right) + \operatorname{var}\left(E\left(\widehat{\beta}|X\right)\right)$, but $E\left(\widehat{\beta}|X\right) = \beta$ and $\operatorname{var}\left(\beta\right) = 0$, so the second term vanishes. Then since $U^\star$ and ζ^x are independent, with both independent of X,[15] standard manipulations show:

$$
\begin{aligned}
\operatorname{var}\left(\widehat{\beta}\right) &= E\left(\operatorname{var}\left(\widehat{\beta}|X\right)\right) = E\left(E\left(\left(\widehat{\beta}-\beta\right)\left(\widehat{\beta}-\beta\right)'|X\right)\right) \\
&= E\left(E\left((X'X)^{-1}X'\left(U^{\star}+\zeta^{x}\beta\right)\left(U^{\star}+\zeta^{x}\beta\right)'X\left(X'X\right)^{-1}|X\right)\right) \\
&= E\left((X'X)^{-1}X'E\left(\left(U^{\star}U^{\star\prime}+\zeta^{x}\beta\beta'\zeta^{x\prime}\right)|X\right)X\left(X'X\right)^{-1}\right) \\
&= E\left((X'X)^{-1}\right)\left(\sigma^2_{U^{\star}}+\beta'\sigma^2_{\zeta,x}\beta\right).
\end{aligned}
$$

Asymptotically, the analogous distributional results hold, as:

$$
\sqrt{T}\left(\widehat{\beta}-\beta\right) \overset{d}{\longrightarrow} N\left(0,\left(Q^{zx}_{xx}\right)^{-1}\left(\sigma^2_{U^{\star}}+\beta'\sigma^2_{\zeta,x}\beta\right)\right),
$$

and s^2 converges to this error variance $\sigma^2_{U^{\star}}+\beta'\sigma^2_{\zeta,x}\beta$. So the LoSE in X increases the variance of the regression error, and since the power of hypothesis tests is typically decreasing in the variance of the regression error, this decreases power. Then in a regression situation such as that described in this subsection, if a hypothesis test passes at a prescribed level of statistical significance, that is in spite of the dimunition of power from the increased error variance.

4.1.1 Empirical Examples

Since nominal asset prices are measured with virtually no error, they provide a candidate Y variable meeting the assumptions of this section.[16] Given the evidence from the prior section indicating that macroeconomic quantities are measured with LoSE, a

[15]The independence of $U^{\star}_t$ from $X^{\star}_t$ implies independence of $U^{\star}_t$ from $E\left(X^{\star}_t|Z^{x}_t\right)$, X_t, and ζ^{x}_t.

[16]However real, after-tax asset returns are employed in many regression specifications, and these may be mismeasured via use of inapppropriate marginal tax rates or mismeasured price indexes for deflation.

specification such as the human capital CAPM, essentially a regression of stock prices on labor income growth, may be well represented by the regression model developed here.

4.2 Y Mismeasured, X Not Mismeasured, $X_t \in Z_t^y$: Potentially Misleading Standard Errors

In addition to assumption 1, this subsection makes the following assumptions:

Assumption 5 *X_t is not mismeasured: $X_t = X_t^\star$, and $X_t \in Z_t^y$.*

Then $Y_t^\star = X_t\beta + U_t^\star$. The relation between X_t and the information set Z_t^y has an important effect on the properties of the OLS regression estimates; this subsection considers $X_t \in Z_t^y$, and the next $X_t \notin Z_t^y$.

Since $E\left(X_t|Z_t^y\right) = X_t$, we have: $Y_t = X_t\beta + E\left(U_t^\star|Z_t^y\right) + \varepsilon_t$ in this case. The LoSE impacts only $U_t^\star$, so $\zeta_t = U_t^\star - E\left(U_t^\star|Z_t^y\right)$, and $\operatorname{var}\left(E\left(U_t^\star|Z_t^y\right)\right) = \sigma_{U^\star}^2 - \sigma_\zeta^2$. The OLS regression estimates $\widehat{\beta}$ as:

$$\begin{aligned}
\widehat{\beta} &= \left(X'X\right)^{-1} X'Y \\
&= \beta + \left(X'X\right)^{-1} X'\left(E\left(U^\star|Z^y\right) + \varepsilon\right) \\
&= \beta + \left(X'X\right)^{-1} X'\left(U^\star - \zeta + \varepsilon\right).
\end{aligned}$$

LoSE in $U^\star$ introduces no bias or inconsistency,[17] so the overall measurement error in Y introduces no bias or inconsistency in this case. The assumption that Y is a conditional expectation of $Y^\star$ plus noise again plays a critical role here; consistency and unbiasedness do not follow if the first component of Y is something other than a conditional expectation.

The standard errors around the point estimates are more interesting. For the variance of the point estimates, $\text{var}\left(\widehat{\beta}\right) = E\left(\text{var}\left(\widehat{\beta}|X\right)\right)$, and:

$$
\begin{aligned}
E\left(\text{var}\left(\widehat{\beta}|X\right)\right) &= E\left(E\left(\left(X'X\right)^{-1} X' \left(E\left(U^\star|Z^y\right) + \varepsilon\right)\left(E\left(U^\star|Z^y\right) + \varepsilon\right)' X \left(X'X\right)^{-1}|X\right)\right) \\
&= E\left(\left(X'X\right)^{-1}\right)\left(\sigma^2_{U^\star} - \sigma^2_\zeta + \sigma^2_\varepsilon\right),
\end{aligned}
$$

since $E\left(U^\star|Z^y\right)$ and ε are independent; the analogous asymptotic results hold. Consider first the impact of the CME ε, which increases the variance of the regression residuals and parameter estimates, and reduces the power of hypothesis tests.[18] This is a cost of mismeasurement of this type, as hypothesis tests must overcome this dimunition of power to meet conventional levels of statistical significance. Conducting hypothesis tests in the usual way is actually somewhat cautious, as CME in Y leads to fewer rejections of hypotheses, on average. Perhaps because of this increased caution, this bias against

[17] That $E\left(\left(X'X\right)^{-1} X'E\left(U^\star|Z^y\right)\right)$ is zero can be seen as follows: $E\left(\left(X'X\right)^{-1} X'U^\star\right) = E\left(\left(X'X\right)^{-1} X'\right) E\left(U^\star\right) = 0$, but:

$$
\begin{aligned}
E\left(\left(X'X\right)^{-1} X'U^\star\right) &= E\left(E\left(\left(X'X\right)^{-1} X'U^\star|Z^y\right)\right) \\
&= E\left(\left(X'X\right)^{-1} X'E\left(U^\star|Z^y\right)\right), \quad \text{since } X \in Z^y.
\end{aligned}
$$

[18] Neither LoSE in X nor CME in Y introduce bias, and both increase standard errors compared to the case of no mismeasurement. These two types of mismeasurement are really quite similar, then, in terms of their effect on OLS regressions.

rejecting hypotheses, econometricians have largely concluded that CME in the dependent variable Y_t poses no real problems for standard OLS regression and inference.

LoSE in the dependent variable has an opposite effect, decreasing the variance of the regression residuals and parameter estimates. Measurement error of this type actually *increases* the power of hypothesis tests. Is more power a good thing? I would argue not. If a hypothesis is rejected at a prescribed significance level, an immediate concern is that the rejection was due to LoSE-induced power. If the data were free of mismeasurement, the variance of the parameter estimates would be larger and the rejection may not have occured.

Note that the parameter variances here are correct for the parameters governing the relation between mismeasured Y and X, and are not biased down in the same sense as they would be if, for example, we ignored positive autocorrelation in the residuals of the regression. However in most applications the parameters of interest are those governing the relation between true $Y^\star$ and X, and the econometrician uses the mismeasured data to make inferences about that true relation out of necessity, because the mismeasured data are all that is available. Under the assumptions of this subsection, the parameters estimated using the mismeasured data are unbiased and consistent for the parameters governing the true relation, but the parameter variances are smaller than they would be without the LoSE in Y. Standard errors computed using data mismeasured in this way then give a misleading sense of precision about the relation between true $Y^\star$ and X, and the power of hypothesis tests is misleadingly large.

An example clarifies some of these issues. Imagine a situation where the econometrician has access to true $Y^\star$, X, and a list of other variables Z^y that are orthogonal to X but related to $Y^\star$. Hypotheses about the relation between $Y^\star$ and X are of interest.

Testing hypotheses using parameter estimates and standard errors from a regression of $Y^\star$ on X is the natural and correct way to procede. But consider the following two step procedure: (1) regress $Y^\star$ on X and some subset of Z^y, and compute predicted values which we call Y, and then (2) regress Y on X, testing hypotheses about the relation between $Y^\star$ and X using standard errors from this second regression. The parameter estimates in the second regression are the same as in a regression of $Y^\star$ on X, but the standard errors are smaller and tests have greater power because the first stage generates LoSE in Y. In fact, by making Z^y arbitrarily small, the standard errors can be made arbitrarily small and any hypothesis may be rejected; setting Z^y to the null set drives the standard errors to zero. Artificially generating LoSE in Y in this fashion is clearly ridiculous, and makes hypothesis tests meaningless.

In reality, if the data meet the conditions of this section, the econometrician starts out in stage (2) of this two step procedure, sadly without access to $Y^\star$. Something analogous to step (1) has already taken place with the creation of the data Y, by a government statistical agency or some other entity. The main issue is not materially different if it is a government statistical agency rather than the econometrician who has generated the LoSE in Y.

If the variance of LoSE σ_ζ^2 were known, the most prudent course of action would be to increase the estimated variance of $\widehat{\beta}$ by σ_ζ^2. This essentially brings us back the pure CME case, where econometricians are comfortable using standard procedures for inference.[19] Unfortunately, σ_ζ^2 is typically unknowable; the results in sections 3.2 and

[19]Note that one could make a symmetric argument to the one outlined above for *decreasing* the variance of $\widehat{\beta}$ by the variance of the CME σ_ε^2. However an asymmetric treatment is probably appropriate, where such a downward adjustment to the variance of $\widehat{\beta}$ is not made, on the grounds that it is better to be cautious in rejecting hypotheses.

4.3.1 for measured GDP growth provide only an approximate lower bound on σ_ζ^2. The bottom line is that inferences are fundamentally less definitive when the dependent variable is contaminated with LoSE.

In a forecasting context, it should be noted that the LoSE also shrinks the variance of out-of-sample forecast errors. The actual variance of the out-of-sample forecast error for the true variable of interest, $Y_{t+k}^\star - X_{t+k}\widehat{\beta}$, with $\widehat{\beta}$ estimated using mismeasured Y_t, is $\sigma_{U^\star}^2 + \left(\sigma_{U^\star}^2 - \sigma_\zeta^2 + \sigma_\varepsilon^2\right) X_{t+k} E\left((X'X)^{-1}\right) X_{t+k}'$. However the LoSE reduces this variance by σ_ζ^2, and if the increase in variance from CME σ_ε^2 does not offset this increase, the forecast errors computed using mismeasured Y_{t+k} give a misleading sense of precision: the deviations of $Y_{t+k}^\star$ from the forecasts are larger than those mismeasured forecast errors suggest. For example, table 1 shows that for GDP growth forecasts, forecast errors computed using the "advance" GDP growth estimates will be smaller, on average, than the true forecast errors.

4.2.1 Empirical Examples

Regressions of mismeasured macroeconomic quantities like GDP growth on time trends, dummies, and other deterministic variables certainly meet the conditions outlined in this subsection. These types of regressions are run most frequently in a forecasting context, where the primary goal is to forecast $Y^\star$, not to estimate the deep structural parameters of an economic model relating $Y^\star$ to $X^\star$.[20] Many other types of variables that are mismeasured in the latter context can be treated as measured without error in the former forecasting context, where the explanatory variables X are just forecasting

[20]Of course, if the primary goal is to forecast mismeasured Y instead of true $Y^\star$, then standard OLS results for variables free of measurement error apply.

tools. A good example is source data used to estimate Y (certainly in the relevant information set). The source data X may be a mismeasured estimate of some other variable $X^\star$, but $X^\star$ is not relevant when using X solely as a forecasting tool; X can be treated as an estimate of itself measured without error. Another potential example is lags of the dependent variable;[21] lagged Y is by definition a mismeasured estimate of lagged $Y^\star$, but this is irrelevant for purposes of pure forecasting. The lags are certainly in the information set of the agency estimating Y, although it is not necessarily clear to what extent the estimating agency relies on the lags in computing current period estimates. If the agency uses the lags optimally in constructing the estimates, or if the lags contain no information above and beyond the source data used to compute the current period estimates, then the lags meet this subsection's conditions for X variables in the context of a pure forecasting regression. Any variable that adds no information about the dependent variable above and beyond that contained in its source data would qualify.

4.3 Y Mismeasured, X Not Mismeasured, $X_t \notin Z_t^y$: Biased Point Estimates

In addition to assumption 1, this subsection makes the following assumptions:

Assumption 6 *X_t is not mismeasured: $X_t = X_t^\star$, and $X_t \notin Z_t^y$.*

This case is applicable when the explanatory variables add information about the dependent variable above and beyond that contained in the source data used to esti-

[21]Some parts of Assumption 1 pertaining to independence will not be applicable with lagged dependent variables, and the small sample results here do not hold in this case. However alternative assumptions may be made for which the analogous asymptotic results do hold.

mate the dependent variable. The mismeasured variable of interest in this case is then $Y_t = E\left(X_t|Z_t^y\right)\beta + E\left(U_t^\star|Z_t^y\right) + \varepsilon_t$. Equation (6) becomes:

$$\begin{aligned}\widehat{\beta} &= \left(X'X\right)^{-1} X'Y \\ &= \beta + \left(X'X\right)^{-1} X'\left(\left(E\left(X|Z^y\right) - X\right)\beta + U^\star - \zeta^u + \varepsilon\right) \\ &= \beta + \left(X'X\right)^{-1} X'\left(-\zeta^{xy}\beta + U^\star - \zeta^u + \varepsilon\right).\end{aligned}$$

Bias and inconsistency are evidently issues here. $X = E\left(X|Z^y\right) + \zeta^{xy}$ is clearly not independent of $-\zeta^{xy}\beta$, and:

$$\begin{aligned}(13)\qquad E\left(\widehat{\beta}\right) &= \beta - E\left(\left(X'X\right)^{-1} X'\zeta^{xy}\right)\beta \\ &= E\left(\left(X'X\right)^{-1} X'E\left(X|Z^y\right)\right)\beta. \\ (14)\qquad \widehat{\beta} &\xrightarrow{p} \beta - \left(Q_{xx}\right)^{-1}\sigma^2_{\zeta,xy}\beta \\ &= \left(Q_{xx}\right)^{-1} Q_{xx}^{zy}\beta.\end{aligned}$$

The inconsistency of $\widehat{\beta}$ tends towards zero, since Q_{xx} equals Q_{xx}^{zy} plus another positive semidefinite matrix $\sigma^2_{\zeta,xy}$. Some variation in X that appears in $Y^\star$ is missing from mismeasured Y, essentially driving down the covariance between X and Y, and driving down the parameter estimates as well since the variance of X is not biased down. If X is univariate, the inconsistency of $\widehat{\beta}$ is unambiguously towards zero, similar to standard attenuation bias from CME in the explanatory variable of a regression. Indeed, comparing these bias and inconsistency results with (10) and (11), it is clear that CME in X and LoSE in Y of the type in this subsection lead to biases that are essentially equivalent algebraically.

Instruments W that meet the conditions of assumption 2 in this case are those for which $X'P_W\zeta^{xy}$ converges in probability to zero, for example if $W_t \in Z_t^y$, so that W_t is independent of the information about $X_t^\star$ missing from Y_t. Instruments typically thought of as valid based on other considerations may not meet this condition; see the discussion of Euler equation estimation at the end of the section. The asymptotic distribution of the IV regression estimates $\widehat{\beta}$ is:

$$\sqrt{T}\left(\widehat{\beta}-\beta\right) \xrightarrow{d} N\left(0,(Q_{xx}^w)^{-1}\left(\sigma_{U^\star}^2-\sigma_{\zeta,u}^2+\sigma_\varepsilon^2+\beta'\sigma_{\zeta,xy}^2\beta\right)\right),$$

with s^2 converging to this asymptotic variance. The concerns about $\sigma_{\zeta,u}^2$ from subsection 4.2 remain, as the LoSE in $U^\star$ reduces the error and parameter variances compared to the case of no mismeasurement. The part of the LoSE that impacts the explanatory variables increases the variance of the error.

4.3.1 Empirical Examples

A wide range of regression specifications may employ explanatory variables that are not mismeasured and contain information about the dependent variable above and beyond that contained in the source data used to estimate the dependent variable. Some such specifications are employed for forecasting $Y^\star$ using X, some are employed to estimate the structural parameters of an economic model relating $Y^\star$ to X, and some are employed for both purposes; the bias results apply in all these situations as long as $Y^\star$ is the object of interest.

Regressions of macroeconomic quantities on asset prices are particularly interesting examples likely to meet the conditions of this subsection. It is entirely possible that some of the variation missed by LoSE-contaminated macroeconomic quantities appears in asset prices. Of course, some variation in asset prices likely arises from misinformation, rational or irrational bubbles, and other factors unrelated to fundamentals, but this does not imply that other more informative types of variation are not present as well. Dynan and Elmendorf (2001) and Fixler and Grimm (2006) show that asset prices predict revisions to economic aggregates like GDP growth, showing that asset prices contain information missed by the earlier vintages of the government estimates. Asset prices may contain information missed by the fully revised, later-vintage estimates as well, and that information may appear in asset prices in at least two ways.

First, there almost surely exists publicly-available information about the state of the economy that is not fully incorporated into GDP growth or its subcomponents, information which is observed by the vast majority of asset market participants and is thus likely incorporated into asset prices. The various pieces of data used to compute GDI may be part of this information, including GDI's source data on corporate profits and employee compensation. Not all of this information appears to be incorporated into measured GDP growth, as section 3.2 illustrates, and financial markets certainly react to this information; see Faust et al (2003), for example.

Second, asset prices aggregate the private information of vast numbers of market participants, private information that is likely correlated with current or future economic activity. For example, the stock price of a company may reflect numerous pieces of private information about that company's cash flow prospects. Aggregating across all firms, the idiosyncratic variation in firms' stock prices averages out, so an aggregate stock market index contains the signal about aggregate economic activity dispersed in all the pieces of private information.[22] Bond prices and exchange rates are fundamentally tied to economic activity, with market participants placing bets with real money about current and future economic prospects; see Evans and Lyons (2005, 2007) for a description of how private information about the economy becomes embedded in exchange rates through the market's filtering of order flow information.

This points to parameter estimates from regressions of macro quantities on asset prices that are biased towards zero. To examine this, consider a regression of several of our quarterly output growth measures on current and lagged growth rates of the Wilshire 5000 stock price index; Fama (1990) studied a similar specification.[23] The relation between economic growth and lagged stock prices can be interpreted in at least two non-mutually-exclusive ways. First, stock prices may be responding to news about current and future economic growth and its effect on expected cash flow to firms.[24] Second, stock price variation may have a causal effect on current and future economic growth, through wealth effects on consumer spending, for example. For our purposes here, differentiating between these two stories is not necessary; it suffices that a relation between true GDP

[22]Even if the aggregate stock price contains useful information about aggregate activity, that does not necessarily imply that any individual holds particularly useful private information - the aggregation of dispersed private information by the market is key - see Hayek (1945). Nalewaik (2006) makes a similar argument about consumption growth.

[23]The stock price changes are quarterly growth rates, while the output growth measures are annualized quarterly growth rates as in tables 1 and 2. The stock price index is nominal. The results change little if the stock price index is deflated with the GDP deflator; deflating introduces some measurement error issues into the explanatory variables and for our purposes here it seems best to avoid that.

[24]Using similar logic, an analogous specification was derived to model the relation between income growth and current and lagged consumption growth by Hansen, Roberds and Sargent (1987), which was employed fruitfully by Nalewaik (2006). Under this interpretation, if the stock price is the present discounted value of expected future profits, the discount rate is constant, and profits are a linear function of true output growth, then in a regression of true output growth on current and lagged stock price growth, the coefficients are proportional to the Wold moving average coefficients from a linear combination of the set of variables governing the market's information about output growth.

growth $Y^\star$ and stock prices X does exist, governed by a true parameter vector β.

The first column of table 3 shows regression results from the post-1984 sample, using as the dependent variable the time series of "advance" GDP growth estimates. Recall that since a substantial amount of data is missing for the "advance" GDP estimate, it is likely contaminated with a particularly large amount of LoSE. The second column shows results using the latest available estimates of GDP growth, which have passed through revisions incorporating more-comprehensive data, reducing LoSE and increasing the variance of the estimates (see table 1).

In comparing these two columns of table 3, first note that each stardard error in the second column exceeds its counterpart in the first (these are Newey-West (1987) corrected for heteroskedasticity and second-order autocorrelation). The LoSE in "advance" GDP growth biases down the standard errors. Second, note that most of the coefficients in the second column exceed those in the first.[25] The last column reports the sum of the coefficients and its standard error, giving a sense of the average downward bias in the coefficients in the first column relative to the second. The difference between the sums in the two columns, 0.072, is statistically significant, with a standard error of 0.036 correcting for cross-correlation and second-order cross-autocorrelation between the two sets of residuals. We reject the hypothesis that the LoSE in "advance" GDP growth does not bias down the regression coefficients. The size of the downward bias is about what we would expect given the summary statistics in table 1: if all the variance in each estimate is signal, the ratio of the variance of the two GDP growth estimates provides

[25] Although intuition about univariate attenuation bias does not necessarily hold for all coefficients in a multivariate setting, $(X'X)^{-1}$ is close to diagonal since Δp_t is approximately serially uncorrelated, and that intuition does hold here.

a reasonable guess as to the expected size of the downward bias.[26] This signal-to-signal ratio is about three-fourths, with the ratio of the sum of the coefficients about two-thirds.

The third column of table 3 reports results using latest GDI growth as the dependent variable. Most striking is the large increase in the sum of the regression coefficients compared to column two, 0.116 with a standard error of 0.040. A straightforward interpretation of this result is that GDI growth contains more information about true output growth than does GDP growth, leading to a greater LoSE-induced attenuation bias in regressions using GDP growth. However, direct measures of corporate profits are included in estimates of GDI growth, opening up other possible interpretations.

One alternate interpretation is that estimates of corporate profits are noisy, and stock prices react to some of that noise. Then ε is positively correlated with X in regressions using GDI growth as the dependent variable, biasing the coefficients up. If this were the case, the problem should be more severe for the early estimates of GDI growth, before profits have been benchmarked to administrative and tax records. Yet the fourth column of table 3 shows that the sum of the coefficients using the early estimates of GDI growth is less than half the sum of the coefficients using the revised estimates released several years later; this noise interpretation does not fit the facts.

A second alternate interpretation is that GDI growth contains more information about corporate profits than does the profits variation implicit in measured consumption, investment, and the other components of GDP growth, but that superior information about profits does not translate into superior information about output growth. This could occur if the profits information is negatively correlated with the remainder of

[26]Of course, it is only signal reflected in stock prices that matters for bias, not overall signal. But if stock prices are a comprehensive information aggregator, this calculation should work well.

GDI growth. Since profits are likely the relevant variable for determining stock prices, relatively large regression coefficients may occur using GDI growth even if it does not contain superior information about output growth. This is unlikely for a number of reasons, but can be examined most directly by regressing the growth rate of GDI minus corporate profits (deflated by the GDP deflator) on the stock price changes. If this alternate interpretation is correct, stripping out profits should reduce the sum of the coefficients, but the last column of the table shows that the sum of the coefficients actually increases.[27]

As tables 1 and 2 show, the ratio of the variances of latest GDP growth over latest GDI growth is 0.87, but the ratio of the sum of the coefficients in table 3 is considerably less, again about two-thirds. One possible explanation is that not all of the variance of GDP growth is signal; some noise in GDP growth would bring down the signal-to-signal ratio of GDP to GDI growth closer the observed downward bias in the regressions. To investigate this, table 3A switches to a univariate regression framework where the explanatory variable is the average stock price change over the current and six previous quarters; table 3B then reverses the regression, using the average stock price change as the dependent variable. The coefficient using GDP growth as the explanatory variable is about three-fourths the size of the coefficient using GDI growth, an attenuation bias that is indicative of noise in GDP growth. While the statistical significance of the difference between the slopes is marginal (the 0.146 difference has a standard error of 0.088), this evidence is suggestive.

Assuming that one-fourth of the variance of GDP growth is noise, then GDP growth captures at most 64 percent of the variance of true GDP growth, recomputing the upper

[27] A similar result occurs using the growth rate of national income minus corporate profits.

bound on $\frac{\text{var}(\Delta Y_t^\star)-\text{var}\left(\zeta_t^{GDP}\right)}{\text{var}(\Delta Y_t^\star)}$ from the end of section 3.2. In this case the variance of signal in GDP growth is about equal to its covariance with GDI growth, and the ratio of this signal variance to the variance of GDI growth gives the upper bound. If all information about output growth is reflected in GDI growth, so $\Delta Y^\star = \Delta Y^{GDI}$, this bound holds and the coefficients in the third column of table 3 are the true parameter vector β. However, if some information about true output growth is missing from GDI growth, these coefficients are themselves biased down, and unfortunately we do not know the size of this downward bias.

These results using stock prices are largely confirmed by regressions of the different output growth measures on bond prices, as shown in table 4. The explanatory variables are TERM, the difference in yield between 10-year and 2-year treasury notes, and DEF, the difference in yield between corporate bonds and 10-year treasury notes.[28] Numerous papers have used similar variables to forecast output growth; see for example Chen (1991) and Estrella and Hardouvelis (1991). The table examines regressions at forecasting horizons ranging from one- to eight-quarters ahead; DEF has substantial explanatory power at shorter horizons, while TERM shows some explanatory power at longer horizons. All of the coefficients except one and all of the standard errors increase when we switch from "advance" GDP as the dependent variable to latest GDP. Switching from latest GDP to latest GDI, the coefficients again all increase, except for TERM at the one- and two-quarter ahead horizons when its statistical significance and marginal explanatory power are weakest. The last column reports p-values from an F-test of equal coefficients from the GDP and GDI regressions; equality is rejected at the three- and

[28]The corporate bond yield measure is the Merrill Lynch High Yield Master II Index. This series extends back only as far as 1986; hence the shorter sample for these regressions.

four- quarter ahead horizons. Similar results obtained from univariate regressions using either TERM or DEF, although the standard errors around the TERM coefficients were larger making definitive statements from those regressions difficult. Using DEF as the dependent variable in reverse regressions, coefficients were smaller using GDP growth as the explanatory variable than using GDI growth, supporting evidence of some noise in GDP growth. The coefficients using GDP growth were between 12 and 42 percent smaller, depending on horizon.

The evidence here and in section 3 shows that GDP growth at the quarterly frequency is contaminated with significant LoSE, implying its major subcomponents are contaminated with significant LoSE as well. As discussed in section 3.1, a large fraction of consumption lacks any real source data at the quarterly frequency, so this component is likely to be particularly contaminated. Consider the implications for Euler equation estimates of the relation between quarterly macro consumption growth Δc_t and interest rates r_t; see Campbell and Mankiw (1989). True consumption growth may have substantial covariance with interest rates, but mismeasured consumption growth misses some of this variation, biasing the OLS regression coefficient towards zero. Lagged variables such as lagged interest rates are almost universally assumed to be valid instruments in estimating the Euler equation, and they may be valid for dealing with expectational errors and some other forms of endogeneity. However if interest rates contain information about actual contemporaneous consumption growth missed by measured consumption growth, lagged interest rates likely contain just as much if not more of this missing information, since interest rates are basically forward-looking. Instrumenting current interest rates with lagged interest rates does not meet Assumption 2, and the instrumental variables parameter estimates remain biased towards zero.

4.4 Both X and Y Mismeasured: Illuminating Special Cases

Again for simplicity, and to focus on the effects of LoSE, this section considers the case of no CME in X, so assumption 4 holds, as well as assumption 1. Three special cases are illuminating. The first is where the information sets used to construct Y and X coincide in the universe of variables correlated with X, so $Z^y_x = Z^x$ (see footnote 14). Then $E(X^\star|Z^y) = E(X^\star|Z^x)$, so their difference in (5) and (6) disappears, leaving unbiased and consistent regression parameter estimates. The variance and asymptotic distribution of $\widehat{\beta}$, and the probability limit of s^2, are as in subsection 4.2. The main concern under these circumstances is the shrinking effect of LoSE on standard errors.

The second illuminating case is where $Z^y_x \subset Z^x$, so Z^x contains all the information about $X^\star$ in Z^y, plus additional information. The difference $E(X^\star|Z^x) - E(X^\star|Z^y)$ is independent of Z^y; substituting this difference for ζ^{xy} in subsection 4.3 then leaves the results of that section unchanged. The estimate $\widehat{\beta}$ is biased and inconsistent, with the bias towards zero; some variation in measured X that appears in $Y^\star$ is missed by measured Y, biasing down the covariance between X and Y relative to the variance of X. Valid instruments must be in the information set used to compute the more-poorly measured Y.

The last illuminating case is where Z^y contains all the information about $X^\star$ in Z^x plus additional information, so $Z^y_x \supset Z^x$. Then $E(X^\star|Z^y) - E(X^\star|Z^x)$ is independent of Z^x and X, and if this difference replaces ζ^x in subsection 4.1, the results in that subsection carry over to this case, except LoSE in $U^\star$ shrinks the error and parameter variances. The estimates are unbiased and consistent.

These cases should help provide some intuition about the potential effects of LoSE in particular regression applications where the econometrician has some knowledge of the relative degree of mismeasurement in the explanatory and dependent variables. For each application, whether $Z_x^y \supset Z^x$, $Z_x^y = Z^x$, or $Z_x^y \subset Z^x$ provides the best description of reality determines which results are most relevant, those from subsection 4.1 (augmented with LoSE in $U^\star$), 4.2, or 4.3. For example, the extent of any bias in the parameter estimates depends on the degree to which the mismeasured explanatory variables contain signal missing from the dependent variable.

4.4.1 Empirical Examples

Regressions of mismeasured macro quantities on other mismeasured macro quantities meet the conditions of this subsection, when the goal is to estimate the relation between $Y^\star$ and $X^\star$, the true values of the quantities. The permanent income hypothesis (PIH), for example, is about the relation between the true values of consumption and income, not mismeasured estimates of consumption and income, so regressions of macro consumption on macro income that attempt to estimate the PIH parameters meet the conditions of this subsection. The evidence in sections 3.2 and 4.3.1 suggests that the $Z_x^y \subset Z^x$ results on bias may be operational in such a PIH regression.[29]

5 Conclusions

This paper has shown that the canonical classical measurement error (CME) model is too restrictive to handle some important cases of mismeasurement. That model assumes that the mismeasured variable of interest is equal to the true variable plus a noise term

[29] Income and consumption do share some source data at the quarterly frequency, so a counterveiling bias here may be correlated CME in the estimates biasing the regression coefficients upwards. Formally, the assumption 1 does not hold since ε_t is positively correlated with ε_t^x.

uncorrelated with the true variable, implying that the variance of the mismeasured variable must exceed the variance of the true variable. It is easy to think of hypothetical examples where this implication is not met, and some important examples of mismeasurement in macroeconomic time series do not meet this implication either. For example, time series such as US GDP growth and GDI growth pass through numerous revisions, and the first-available vintage of a variable can be thought of as a mismeasured estimate of subsequent revised vintages. The variance of the first-available vintage is often less than the variance of better-measured subsequent vintages, a form of mismeasurement that does not fit into the CME box.

The paper proposes a simple generalization of the CME model that is mathematically tractable, embeds the CME model as a special case, and adds useful flexibility, allowing the mismeasured variable to have either more or less variance than the true variable of interest. The key to the generalization is that instead of just allowing mismeasurement that *adds noise* to the true variable of interest, it also permits mismeasurement that *subtracts signal*; I label this reduction of signal from mismeasurement the Lack of Signal Error, or LoSE for short.

In some ways, this generalization of the CME model provides the second half of the story about errors in variables and their effect on ordinary least squares (OLS) regression, as the results here exhibit a symmetry that is intuitively pleasing. CME in the dependent variable of a regression Y does not bias parameter estimates, and its only effect is to increase standard errors compared to what they would be without the mismeasurement. Given this lack of bias, and the caution in rejecting hypotheses introduced by the relatively large standard errors, the consensus in the profession is that CME in the dependent variable Y poses no serious problems for OLS regression and inference. LoSE in the *explanatory variables* X, has the same effect on OLS regression: the LoSE does not introduce bias into the parameter estimates, and it increases their standard errors compared to what they would be without the mismeasurement.

CME in the explanatory variables X, of course, does cause problems for OLS regression, namely bias in the parameter estimates, towards zero in the univariate case. LoSE in the *dependent variable* Y introduces a similar bias under some circumstances: when some of the signal missing from the dependent variable Y is captured by the explanatory variables X, such an attenuation-type bias appears. LoSE in Y is also problematic since it shrinks the variance of the regression residuals, decreasing the size of standard errors compared to what they would be without this type of mismeasurement. This raises the possibility that some rejections of hypotheses may stem from the existence of LoSE in Y, so without this measurement error, the hypotheses would not be rejected. Under such circumstances, additional caution in making inferences about the relation between true Y and X seems appropriate, as conventional standard errors may be misleadingly small.

The paper provides a taxonomy of different types of LoSE and their impact on OLS parameter estimates and standard errors, and provides examples of common regressions in macroeconomics that may fit into the different categories considered. The paper reviews recent evidence in Fixler and Nalewaik (2007) and Nalewaik (2007a,b) that US GDP growth is mismeasured with LoSE even after the data has passed through all its revisions. A comparison with GDI growth yields a lower bound on the variance of LoSE in GDP growth: since the mid-1980s, quarterly or annual GDP growth has captured at most 70% of the variance of the true growth rate of the economy. US GDP growth and its subcomponents like consumption growth have served as the dependent variables in many regression studies in macroeconomics and finance; the potential for biases stemming from mismeasurement of the dependent variable, and misleading standard errors - these issues have not been contemplated in a serious way prior to this paper.

Asset prices are a set of variables that may capture some of the signal missing from GDP growth and its subcomponents, implying attenuation-type biases in regressions of the mismeasured quantities on those prices. The empirical results here confirm that. The government's first estimates of GDP growth are likely contaminated with more LoSE than their revised GDP growth estimates, and those GDP growth estimates are likely contaminated with more LoSE than GDI growth. In regressions of these different measures of output growth on either stock prices or bond prices, the measures of output growth contaminated with more LoSE have smaller coefficients, and the changes in the coefficients across regressions are often statistically significant. The set of explanatory variables is fixed from regression to regression; the only thing changing is the degree of measurement error in the dependent variable. We reject the CME intuition that measurement error in the dependent variable does not bias regression coefficients. It can cause biases if it is LoSE, and does cause biases in this important set of regressions of macro quantities on asset prices.

On a positive note, the results derived here provide some clear prescriptions for handling different types of mismeasurement, in terms of choice of instruments, and also choice of which variable is dependent Y, and which is explanatory X. As an example, consider estimation of an Euler equation relation between macro data on consumption growth and the interest rate. If consumption growth is mismeasured with LoSE, and the interest rate is free from mismeasurement as it may be under some circumstances, then the results here recommend using the interest rate as the *dependent* variable. The common practice of using consumption growth as the dependent variable, and instrumenting the interest rate with lagged interest rates, does not produce consistent estimates when consumption growth contains LoSE. Such insights are opposite the current conventional wisdom in the profession. The generalized measurement error model with LoSE is likely applicable in a wide variety of econometric specifications beyond the few considered here, and our results should provide helpful insights for making appropriate modifications to econometric practice.

References

[1] Broda, Christian and Weinstein, David. "Product Creation and Destruction: Evidence and Price Implications," working paper, 2007.

[2] Campbell, John Y. and Mankiw, Gregory N. "Consumption, Income, and Interest Rates: Reinterpreting the Time Series Evidence." In Olivier Blanchard and Stanley Fischer, eds., NBER Macroeconomics Annual, (Cambridge, NBER, 1989) pp. 185-245.

[3] Chen, Nai-Fu. "Financial Investment Opportunities and the Macroeconomy." *Journal of Finance*, 1991 (46), pp. 529-554.

[4] Dynan, Karen E. and Elmendorf, Douglas W. "Do Provisional Estimates of Output Miss Economic Turning Points?" *FEDS working paper*, 2001.

[5] Estrella, Arturo and Hardouvelis, Gikas. "The Term Structure as a Predictor of Real Economic Activity." *Journal of Finance*, 1991 (46), pp. 555-576.

[6] Evans, Martin D.D., and Lyons, Richard K. "Meese-Rogoff Redux: Macro-Based Exchange Rate Forecasting." *NBER working paper 11042*, January 2005.

[7] Evans, Martin D.D., and Lyons, Richard K. "Exchange Rate Fundamentals and Order Flow." *NBER working paper 13151*, June 2007.

[8] Fama, Eugene F. "Stock Returns, Expected Returns, and Real Activity." *Journal of Finance*, 1990 (45), pp. 1089-1108.

[9] Faust, Jon; Rogers, John; Wang, Shig-Yi, and Wright, Jonathan. "The High-Frequency Response of Exchange Rates and Interest Rates to Macroeconomic Announcements." Board of Governors of the Federal Reserve System, *International Finance Discussion Paper* 784, 2003.

[10] Fixler, Dennis J. and Grimm, Bruce T. "GDP Estimates: Rationality tests and turning point performance." *Journal of Productivity Analysis*, 2006 (25), pp. 213-229.

[11] Fixler, Dennis J. and Nalewaik, Jeremy J. "News, Noise, and Estimates of the True Unobserved State of the Economy." Board of Governors of the Federal Reserve System, *FEDS working paper* 2007-34.

[12] Fuller, Wayne A. *Measurement Error Models* (New York: John Wiley and Sons, 1987).

[13] Grimm, Bruce T. and Teresa Weadock. "Gross Domestic Product: Revisions and Source Data" *Survey of Current Business*, February 2006, pp. 11-15.

[14] Hayek, F. "The Use of Knowledge in Society." *American Economic Review*, 1945 (35), pp. 519-530.

[15] Kimball, Miles; Sahm, Claudia; and Shapiro, Matthew. "Imputing Risk Tolerance from Survey Responses" *Journal of the American Staistical Association*, forthcoming, 2007.

[16] Mankiw, N. Gregory and Shapiro, Matthew D. "News or Noise: An Analysis of GNP Revisions" *Survey of Current Business*, May 1986, pp. 20-25.

[17] McConnell, Margaret M., and Perez-Quiros, Gabriel. "Output Fluctuations in the United States: What Has Changed Since the Early 1980s?" *American Economic Review*, 2000 (90), pp. 1464-1476.

[18] Nalewaik, Jeremy J. "Current Consumption and Future Income Growth." *Journal of Monetary Economics*, 2006 (53) pp. 2239-66.

[19] Nalewaik, Jeremy J. "Estimating Probabilities of Recession in Real Time Using GDP and GDI." Board of Governors of the Federal Reserve System, *FEDS working paper* 2007-07.

[20] Nalewaik, Jeremy J. "Incorporating Vintage Differences and Forecasts into Markov Switching Models." Board of Governors of the Federal Reserve System, *FEDS working paper* 2007-23.

[21] Newey, W.K. and West, K.D. "A Simple, Positive Semi-Definite Heteroskedasticity and Autocorrelation Consistent Covariance Matrix." *Economtrica* 1987 (55), pp. 703-708.

Table 1: Summary Statistics on Vintages of GDP and GDI Growth

Quarterly Data, 1984Q3-2004

Vintage	$\text{var}\left(\Delta Y_t^{GDP}\right)$	$\text{var}\left(\Delta Y_t^{GDI}\right)$
Current Quarterly, "Advance"	3.1	.
Current Quarterly, "Final"	4.1	4.0
Latest Vintage Available	4.2	4.8

Table 2: Summary Statistics for GDP and GDI Growth

	$\text{var}\left(\Delta Y_t^{GDP}\right)$	$\text{var}\left(\Delta Y_t^{GDI}\right)$	$\text{cov}\left(\Delta Y_t^{GDP}, \Delta Y_t^{GDI}\right)$	Upper bound on: $\frac{\text{var}\left(\Delta Y_t^{GDP}\right)}{\text{var}\left(\Delta Y_t^{\star}\right)}$
Quarterly, 1947-1984Q2	24.1	24.1	22.4	0.93
Annual, 1947-1984	8.2	8.5	8.2	0.97
Quarterly, 1984Q3-2004	4.2	4.8	3.1	0.70
Annual, 1985-2004	1.6	2.6	1.9	0.69

Table 3: Regressions of Different Measures of Quarterly Output Growth on Current and Lagged Stock Price Growth, 1984Q3 to 2004Q4:

$$\Delta Y_t^i = \alpha + \beta_0 \Delta p_t + \beta_1 \Delta p_{t-1} + \ldots + \beta_6 \Delta p_{t-6} + e_t$$

Measure:	ΔY^{GDP}	ΔY^{GDP}	ΔY^{GDI}	ΔY^{GDI}	ΔY^{GDI-CP}
Vintage:	"Advance"	Latest	Latest	"Final"	Latest
β_0:	**0.012**	**0.014**	**0.032**	**0.031**	**0.003**
	(0.019)	(0.027)	(0.026)	(0.023)	(0.025)
β_1:	**0.048**	**0.054**	**0.084**	**0.038**	**0.083**
	(0.018)	(0.024)	(0.023)	(0.023)	(0.030)
β_2:	**0.021**	**0.065**	**0.056**	**0.036**	**0.059**
	(0.024)	(0.027)	(0.027)	(0.027)	(0.028)
β_3:	**0.058**	**0.057**	**0.067**	**0.044**	**0.093**
	(0.017)	(0.022)	(0.023)	(0.020)	(0.031)
β_4:	**0.011**	**0.015**	**0.051**	**0.024**	**0.064**
	(0.023)	(0.028)	(0.029)	(0.024)	(0.028)
β_5:	**-0.003**	**-0.007**	**0.038**	**-0.019**	**0.070**
	(0.025)	(0.026)	(0.022)	(0.022)	(0.028)
β_6:	**0.002**	**0.023**	**0.008**	**0.002**	**0.032**
	(0.016)	(0.017)	(0.027)	(0.018)	(0.030)
$\sum_{k=0}^{6} \beta_k$:	**0.149**	**0.221**	**0.337**	**0.156**	**0.403**
	(0.055)	(0.060)	(0.069)	(0.069)	(0.079)

Table 3A: Regressions of Different Measures of Quarterly Output Growth on Current and Lagged Stock Price Growth, 1984Q3 to 2004Q4:

$$\Delta Y_t^i = \alpha + \beta \left(\Delta p_t + \Delta p_{t-1} + \ldots + \Delta p_{t-6}\right)/7 + e_t$$

Measure:	ΔY^{GDP}	ΔY^{GDP}	ΔY^{GDI}
Vintage:	"Advance"	Latest	Latest
β:	**0.142**	**0.214**	**0.325**
	(0.060)	(0.068)	(0.073)

Table 3B: Reverse Regressions of Current and Lagged Stock Price Growth on Different Measures of Quarterly Output Growth, 1984Q3 to 2004Q4:

$$\left(\Delta p_t + \Delta p_{t-1} + \ldots + \Delta p_{t-6}\right)/7 = \alpha + \beta^r \Delta Y_t^i + e_t$$

Measure:	ΔY^{GDP}	ΔY^{GDP}	ΔY^{GDI}
Vintage:	"Advance"	Latest	Latest
β^r:	**0.411**	**0.454**	**0.600**
	(0.194)	(0.182)	(0.169)

Table 4: Regressions of Different Measures of Quarterly Output Growth on Lagged Interest Rates Spreads (TERM and DEF), 1988Q3 to 2004Q4:

$$\Delta Y_t^i = \alpha + \beta_{TERM}\left(r_{t-k}^{10yr} - r_{t-k}^{2yr}\right) + \beta_{DEF}\left(r_{t-k}^{corp} - r_{t-k}^{10yr}\right) + e_t$$

Measure:	ΔY^{GDP}, "Advance"		ΔY^{GDP}, Latest		ΔY^{GDI}, Latest		p-val., equal βs
	β_{TERM}	β_{DEF}	β_{TERM}	β_{DEF}	β_{TERM}	β_{DEF}	
k=1	**0.20**	**-0.50**	**0.31**	**-0.61**	**0.23**	**-0.79**	0.10
	(0.26)	(0.13)	(0.26)	(0.13)	(0.29)	(0.10)	
k=2	**0.42**	**-0.44**	**0.48**	**-0.53**	**0.43**	**-0.69**	0.13
	(0.26)	(0.12)	(0.31)	(0.12)	(0.33)	(0.13)	
k=3	**0.58**	**-0.38**	**0.60**	**-0.40**	**0.68**	**-0.65**	0.00
	(0.30)	(0.12)	(0.36)	(0.15)	(0.37)	(0.15)	
k=4	**0.62**	**-0.23**	**0.57**	**-0.28**	**0.70**	**-0.50**	0.01
	(0.32)	(0.15)	(0.39)	(0.17)	(0.40)	(0.17)	
k=5	**0.59**	**-0.19**	**0.67**	**-0.29**	**0.75**	**-0.41**	0.40
	(0.35)	(0.14)	(0.38)	(0.14)	(0.44)	(0.19)	
k=6	**0.72**	**-0.27**	**0.76**	**-0.32**	**0.92**	**-0.39**	0.54
	(0.35)	(0.10)	(0.38)	(0.13)	(0.41)	(0.16)	
k=7	**0.73**	**-0.19**	**0.81**	**-0.20**	**0.96**	**-0.39**	0.14
	(0.35)	(0.10)	(0.36)	(0.13)	(0.38)	(0.15)	
k=8	**0.66**	**-0.10**	**0.72**	**-0.15**	**0.94**	**-0.27**	0.27
	(0.34)	(0.13)	(0.36)	(0.14)	(0.37)	(0.15)	