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

Missing Variation in the Great Moderation: Lack of Signal Error and OLS Regression

Jeremy J. Nalewaik

2014-27

Missing Variation in the Great Moderation: Lack of Signal Error and OLS Regression

Jeremy J. Nalewaik*

April 7, 2014

Abstract

This paper studies measurement errors that subtract signal from true variables of interest, labeled lack of signal errors (LoSE). The effect on OLS regression of LoSE is opposite the conventional wisdom about classical measurement errors, with LoSE in the dependent variable, not the explanatory variables, causing attenuation bias under some conditions. The paper provides evidence of LoSE in US GDP growth during the period known as the Great Moderation (roughly the mid-1980s to the mid-2000s), illustrating attenuation bias in regressions of GDP growth on asset prices. These biases may have contributed to conventional macroeconomic analysis missing the severity of the adverse shocks hitting the economy in the Great Recession.

*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, Charles Fleischman, Michael Kiley, David Lebow, Claudia Sahm, Jonathan Millar and Rob Vigfusson for comments. The ideas and analysis in this paper grew out of an earlier one called: "Lack of Signal Error (LoSE) and Implications for OLS Regression: Measurement Error for Macro Data" (FEDS 2008-15). The views expressed in this paper are solely those of the author.

1 Introduction

This paper examines 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*, or classical measurement error (CME); see Klepper and Leamer (1984), Griliches (1986), Fuller (1987), Leamer (1987), Angrist and Krueger (1999), Bound, Brown and Mathiowetz (2001) or virtually any econometrics textbook. The generalization discussed here incorporates measurement error that *subtracts signal* from the true variable of interest, a type of measurement error that this paper calls Lack of Signal Error, or LoSE, for short.

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, in a regression, measurement error in the dependent variable Y poses no real problems for standard estimation and inference, with parameter estimates unbiased and consistent. It is CME in the explanatory variables X that causes the real problems, namely attenuation bias and inconsistency. However with LoSE these results are reversed. For the baseline case considered here, LoSE in the *explanatory variables* X does not lead to bias or inconsistency, similar to CME in Y. It is LoSE in the *dependent variable* Y that introduces an attenuation-type bias and inconsistency into the regression under some circumstances, namely, when the explanatory variables contain some signal missing from the dependent variable.

This point is obvious when we consider the extreme case of maximum LoSE, an error-ridden estimate Y of the true variable $Y^\star$ that is just a constant equal to the unconditional mean of $Y^\star$. Then if $Y^\star = X\beta + U$ for some X with positive variance, a standard OLS regression of Y on X recovers $\widehat{\beta} = \frac{\text{cov}(X,Y)}{\text{var}(X)} = 0$ regardless of the true β. All of the variation in $Y^\star$ from X is missing from the estimate Y, so the parameter estimate is biased all the way to zero. In addition, the LoSE in Y shrinks the variance of the regression residuals U and thus the standard errors, which are zero in this extreme case, raising serious concerns about the robustness of hypothesis tests. Indeed, parameter estimates that have been attenuated and estimated with false precision due to LoSE easily could have led to the rejection of hypotheses that are actually true. The paper derives instrumenting strategies to eliminate bias from LoSE, strategies not derived in the previous literature.

Is LoSE just a curiosity, interesting because it runs contrary to conventional wisdom about the effect of measurement error on regression estimates, but not relevant for the type of work economists actually do? It has long been known that the initial releases of macroeconomic quantities like US gross domestic product (GDP) are contaminated with LoSE; see Mankiw and Shapiro (1986), who show that revisions to GDP growth add news missing from its initial estimates, implying they lack signal. However, it has always been an open question as to whether *all* of the news, or close to all of the news, about true output growth eventually becomes incorporated through revisions. This paper provides evidence that, over the period known as the Great Moderation (roughly the mid-1980s to the mid-2000s), the answer is no: GDP growth still appears to be contaminated with substantial LoSE even after it has passed through all of its revisions.

Regressions of GDP growth and its subcomponents on asset prices are widespread in macroeconomics and finance, and if asset prices capture some of the signal missing from these quantities, the estimated coefficients are biased. The paper examines this hypothesis over the Great Moderation period by regressing different measures of output growth on a fixed set of stock or bond prices. As expected, the regression coefficients increase when we switch the dependent variable from the initial GDP growth estimates based on limited source data to the revised GDP growth estimates reflecting news from more-comprehensive source data, consistent with LoSE in the initial GDP growth estimates. Tellingly, the coefficients increase again when we switch the dependent variable from latest, revised GDP growth to an alternative and likely superior measure of US output growth, GDI growth.[1] This increase in the coefficients is consistent with the hypothesis that LoSE remains in GDP growth *even after it has passed through all of its revisions*. Finally, the paper implements the instrumenting strategies derived here for producing unbiased and consistent parameter estimates when the dependent variable of a regression is contaminated with LoSE. The instrumental variables estimates, which do not employ GDI growth at all, provide independent corroborating evidence of substantial LoSE in latest, revised GDP growth over the Great Moderation period.

The attenuation biases from LoSE discussed here can lead applied macroeconomic analysis astray in ways not fully appreciated by the prior literature. For example, the paper illustrates how these biases may have contributed to conventional analysis underestimating the size of the shocks hitting the economy at the height of the Great Recession, leading to some prominent and widely-discussed forecast errors. A better understanding of the implications of LoSE in GDP growth might help avoid such forecasting mistakes in the future.

Section 2 discusses the relation of the work here to the previous literature. After providing a brief introductory motivation for the generalized measurement error model in section 3, section 4 shows the implications of LoSE for OLS regression and derives

[1]On the superiority of GDI, see Nalewaik (2010) and Aruoba, Diebold, Nalewaik, Schorfheide, and Song (2012, 2013).

valid instruments for dealing with LoSE-induced bias. Section 5 discusses the data and choice of instruments. Section 6 shows regression-based tests and instrumental variables estimates providing evidence for LoSE in GDP growth. Section 7 shows how the attenuation biases from LoSE in GDP growth may have contributed to conventional macroeconomic analysis missing the severity of the shocks hitting the economy at the height of the Great Recession. Section 8 concludes the paper.

2 Relation to Previous Literature

Much of the econometrics literature on non-classical measurement error has focused on binary or categorical response data, for which the classical measurement error assumptions cannot hold; see Card (1996), Bollinger (1996), and Kane, Rouse and Staiger (1999). In a more general linear regression context, Berkson (1950) was an early paper tackling some of the issues addressed here; see the discussion in Durbin (1954), Griliches (1986, section 4), and Fuller (1987, section 1.6.4). Berkson had in mind a regression using "controlled" measurements as the explanatory variable X, readings from a scientific experiment where the unobserved true values of interest $X^\star$ fluctuate around the observed controlled measurements in a random way. Berkson showed that if the unobserved fluctuations $X^\star - X$ are uncorrelated with the measurements X, then regression parameter estimates are unbiased. The literature following Berkson has generally focused on extending his results to regressions employing non-linear functions of X; see Geary (1953), Federov (1974), Carroll and Stefanski (1990), Huwang and Huang (2000), and Wang (2003, 2004). This literature has focused less on the implications of "controlled" measurements of the dependent variable Y.

Use of measurement equations has a long history in economics, with Friedman (1957) being a famous and notable early example, and several papers discuss different LoSE-related estimation issues. These include Sargent (1989), who uses measurement equations in state space models to discuss optimally filter estimates, Bound, Brown and Mathiowetz (2001), Koenig, Dolmas and Piger (2003), and Kimball, Sahm and Shapiro (2008), Kishor and Koenig (2011), Jacobs and van Norden (2011) and Clements and Galvao (2013). Following the pioneering work of Koenig, Dolmas and Piger (2003), many of these papers, such as Clements and Galvao (2013), focus on the implications for forecasting of LoSE in initial macroeconomic estimates (so later revisions yield "news"). In this context, using the LoSE-biased parameter estimates often can yield the most accurate forecasts (especially of the initial estimates), although section 7 below points out some previously-unknown pitfalls of using LoSE-biased parameter estimates in a structural model.

Perhaps the closest paper to this one is Hyslop and Imbens (HI, 2001), which shows some of the major implications of LoSE, while simultaneously considering some other measurement error biases. The results in this paper are distinct from those in HI in at least four ways. First, in defining the LoSE in a variable as the difference between its true value and a conditional expectation of that true value, this paper considers arbitrary conditioning information sets Z, while HI consider more specialized information sets in a univariate regression context.[2] Second, when variables are mismeasured with LoSE, this

[2]A estimate may be informed by many different variables, as government statistical agencies draw on vast information sets in producing their estimates. However, the examples in HI are stylized ones meant to make a point, and they do acknowledge the importance of the conditioning information set: "A crucial ingredient ... is the information set. It may be that the respondent had only a single unbiased measurement of the underlying true variable. Alternatively, other variables, which themselves may enter the econometric model of interest, may be used to produce this estimate."

paper derives precise conditions under which instrumental variables produce consistent estimates. The previous literature has not derived valid instrumenting strategies. Third, the previous literature does not discuss the problems that LoSE-shrunken standard errors pose for hypothesis testing. And fourth, this paper shows evidence of LoSE in US GDP growth even after it has passed through all of its revisions, both by comparing GDP growth with an alternative measure of output, GDI growth, and by implementing the instrumenting strategies derived here. These are important contribution of the paper.

A large body of empirical work has now accumulated on mismeasurement of microeconomic survey data, which generally rejects the CME assumptions and points to negative correlation between the measurement errors and the true variables of interest; see Bound and Krueger (1991), Bound, Brown, Duncan and Rodgers (1994), Pischke (1995), Bollinger (1998), Bound, Brown and Mathiowetz (2001) and the references therein, and Escobal and Laszlo (2008). Such negative correlation is an implication of LoSE, although other measurement error models may generate such a result as well, such as those considered in the appendix of this paper. Some of the problems of imputation in microeconomic surveys are very much related to LoSE as well; see Hirsch and Schumacher (2004) and Bollinger and Hirsch (2006).

3 A Generalization of the Classical Measurement Error Model

Let $Y_t^\star$ be the true variable of interest and Y_t be a mismeasured estimate of that variable. The generalized model of mismeasurement considered here is:

$$Y_t = Y_t^\star - \zeta_t + \varepsilon_t. \tag{1}$$

The term ε_t is "noise" or the classical measurement error (CME) in the estimate, with ε_t and $Y_t^\star$ independent. 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. The other measurement error is defined as:

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

Z_t is a $(1 \times l)$ vector of possibly stochastic variables used to construct Y_t, with ε_t and Z_t independent. 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 is functions of the source data. We place no restrictions on Z_t; it may be arbitrarily large, unlike the stylized examples of LoSE studied in HI.[3]

The ζ_t represents the information about $Y_t^\star$ not contained in Z_t—i.e. mismea-

[3]However, Z_t need not be an exhaustive information set - i.e. it need not contain all available relevant pieces of information about unobserved $Y_t^\star$. Resource and other constraints certainly preclude this from being the case, and the sections below considering the implications of LoSE allow for this possibility.

surement 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. The LoSE is uncorrelated with all functions of Z_t, so $\operatorname{cov}\left(E\left(Y_t^\star|Z_t\right), \zeta_t\right) = 0$ and $\operatorname{cov}\left(Y_t^\star, \zeta_t\right) = \operatorname{cov}\left(E\left(Y_t^\star|Z_t\right) + \zeta_t, \zeta_t\right) = \operatorname{var}\left(\zeta_t\right)$, so:

$$\begin{aligned} \operatorname{var}\left(Y_t\right) &= \operatorname{var}\left(Y_t^\star\right) + \operatorname{var}\left(\zeta_t\right) - 2\operatorname{cov}\left(Y_t^\star, \zeta_t\right) + \operatorname{var}\left(\varepsilon_t\right) \\ &= \operatorname{var}\left(Y_t^\star\right) - \operatorname{var}\left(\zeta_t\right) + \operatorname{var}\left(\varepsilon_t\right). \end{aligned} \tag{3}$$

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 true $Y_t^\star$. With CME alone, the variance of the estimate Y_t must exceed the variance of the true variable, but it is easy to think of counterexamples, such as when $Y_t^\star$ has positive variance but the estimate Y_t is just a constant. Note that while the generalized model here is less restrictive than the CME model, some restrictions do remain. In particular, zero covariance between ζ_t and Y_t is a restriction violated by systematic biases in the estimates. Appendix A considers some measurement error models of this form.

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 explanatory variables X_t, using a sample of length T. When stacking together the T observations, time subscripts are dropped for convenience. The results below are for the case in which Y_t follows the generalized model of section 3, and X_t is measured without error, as is the case in the empirical work below. The most interesting empirical results show through to this specialialized case of no measurement error in X_t; the more general case, in which both X_t and Y_t follow the generalized model of section 3, is analyzed in an appendix.

Our full set of assumptions 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 $X_t^\star$, $\forall t, s$. Measured $Y_t = E\left(Y_t^\star | Z_t^y\right) + \varepsilon_t$, with:*

- *The CME ε_t is i.i.d., mean zero, and independent of $X_t^\star$ and Z_t^y, with $\operatorname{var}(\varepsilon_t) = \sigma^2_\varepsilon$.*

- *The LoSE $\zeta_t = \left(X_t^\star - E\left(X_t^\star | Z_t^y\right)\right)\beta + \left(U_t^\star - E\left(U_t^\star | Z_t^y\right)\right) = \zeta_t^{xy}\beta + \zeta_t^u$. ζ_t^u is i.i.d., independent of $X_t^\star$, and mean zero with $\operatorname{var}(\zeta_t^u) = \sigma^2_{\zeta,u}$, while 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.*

Measured $X_t = X_t^\star$, with:

- $\frac{1}{T}\left(X^\star\right)' X^\star \xrightarrow{p} Q_{xx}$

- $\frac{1}{T}\left(E\left(X^\star | Z^y\right)\right)' E\left(X^\star | Z^y\right) \xrightarrow{p} Q_{xx}^{zy} = Q_{xx} - \sigma^2_{\zeta,xy}$

All relevant fourth moments exist.

We impose the i.i.d. assumptions because they are approximately met in the applications below, and because it allows discussion of bias as well as consistency.[4] However, for other time series applications, the i.i.d. assumption will be overly restrictive, and relaxing it could be a topic for future research.

[4] The time series of latest GDP growth estimates over the Great Moderation sample studied here has an AR1 coefficient of only 0.2, and the errors from the GDP growth regressions in section 6 are even less persistent, with Breusch-Godfrey tests not rejecting independence of the errors across various lag lengths.

Given assumption 1, Y_t can be written as:

$$
\begin{aligned}
Y_t &= E\left(X_t|Z_t^y\right)\beta + E\left(U_t^\star|Z_t^y\right) + \varepsilon_t \\
&= X_t\beta + \left(E\left(X_t^\star|Z_t^y\right) - X_t\right)\beta + E\left(U_t^\star|Z_t^y\right) + \varepsilon_t \\
&= X_t\beta - \zeta^{xy}\beta + U_t^\star - \zeta_t^u + \varepsilon_t.
\end{aligned} \tag{4}
$$

The OLS regression estimator is:

$$
\begin{aligned}
\widehat{\beta} &= \left(X'X\right)^{-1} X'Y \\
&= \beta + \left(X'X\right)^{-1} X'\left(-\zeta^{xy}\beta + U^\star - \zeta^u + \varepsilon\right).
\end{aligned} \tag{5}
$$

It is well known that the CME in Y introduces no bias and inconsistency, since ε is independent of X. The LoSE in $U^\star$ introduces no bias or inconsistency either, since it is uncorrelated with X. However, $X = E\left(X|Z^y\right) + \zeta^{xy}$ is clearly not independent of $-\zeta^{xy}\beta$, and:

$$
E\left(\widehat{\beta}\right) = \beta - E\left(\left(X'X\right)^{-1} X'\zeta^{xy}\right)\beta \tag{6}
$$

$$
\widehat{\beta} \xrightarrow{p} \beta - \left(Q_{xx}\right)^{-1}\sigma^2_{\zeta,xy}\beta \tag{7}
$$

The inconsistency of $\widehat{\beta}$ tends towards zero, since some variation in X that appears in $Y^\star$ is missing from mismeasured Y, essentially driving down the covariance between X and Y and 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.

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

The instruments must be uncorrelated with $-\zeta^{xy}$, for example if $W_t \in Z^y_t$, so that W_t is independent of the information about X_t missing from Y_t. With valid instruments, we have:

$$\begin{aligned} \widehat{\beta} &= \left(X'P_WX\right)^{-1}X'P_WY \\ &= \beta + \left(X'P_WX\right)^{-1}X'P_W\left(-\zeta^{xy}\beta + U^\star - \zeta^u + \varepsilon\right), \end{aligned} \tag{8}$$

and $\widehat{\beta} \xrightarrow{p} \beta$.

The asymptotic distribution of the IV estimate $\widehat{\beta}$ is:

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

where $\xrightarrow{d}$ denotes convergence in distribution as $T \longrightarrow \infty$, and $N\left(a, b\right)$ is a Gaussian distribution with mean a and variance b. The usual estimator of the variance of the error term is:

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

The first two terms converge in probability to $\sigma^2_{U^\star} - \sigma^2_{\zeta,u} + \sigma^2_\varepsilon$, and the cross terms converge in probability to zero. The terms involving β and $\widehat{\beta}$ simplify in the limit since $\widehat{\beta} \xrightarrow{p} \beta$, producing a consistent estimate of the asymptotic error variance:

$$s^2 \xrightarrow{p} \sigma^2_{U^\star} - \sigma^2_{\zeta,u} + \sigma^2_{\varepsilon} + \beta' \sigma^2_{\zeta,xy} \beta.$$

The ζ^u component of the LoSE in Y decreases the variance of the regression residuals and standard errors (whether or not the estimator is consistent), and the LoSE in Y can be particularly pernicious when it both attenuates parameter estimates and shrinks the regression standard errors. In these circumstances, the econometrician runs a high risk of rejecting a candidate hypothesis $\beta = \beta^0$, as long as β^0 is non-zero, even when the hypothesis is actually true. Regressions with such LoSE in Y will tend to show an estimated relation between Y and X that is smaller in absolute value than the true relation, with the true size of the relation appearing implausible because of excessively small standard errors.

5 Data: US Macroeconomic Quantities

The decision to test for LoSE in US GDP growth over the Great Moderation period is motivated by several considerations. GDP is estimated in a bottom-up, component-by-component fashion using government survey data to estimate spending for each category (consumption, investment, etc.) and then aggregating.[5] But, goverment survey data at the quarterly frequency is unavailable for many categories comprising a large share of

[5]The BEA does not use the information in stock or bond prices to make any direct or indirect adjustments to this bottom-up estimation procedure, to the author's knowledge.

GDP, including most services categories of personal consumption expenditures.[6] Growth rates for these categories are typically interpolated or extrapolated using related indicators, or estimated as a "trend extrapolation." It is difficult to imagine how this lack of hard information would not introduce some LoSE into GDP growth, and that LoSE may have become more consequential over time as the share of services in US output has increased.

Several comparisons with an alternative measure of output growth, GDI growth, are also consistent with LoSE in GDP growth over the Great Moderation.[7] The output growth estimates are plotted in Figure 1 over this period, from the mid-1980s to the mid-2000s.[8] GDI growth has higher variance than GDP growth over this sample, which, under the generalized measurement error model in section 3, may stem from some combination of: (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 upcoming evidence in section 6 favors placing more weight on the second explanation. Earlier research on revisions—see Fixler and Nalewaik (2007)—supports this notion as well. Briefly, Table 1 shows that the variance of GDI growth becomes relatively large only after the data pass through its sequence of annual revisions (GDI is unavailable when the "advance" estimates are released for each quarter about a month after the quarter ends, but is always available when the "3rd" estimates are released about three

[6]This situation has begun to change with the introduction of the Quarterly Services Survey (QSS) in 2002, but these data have no material affect on the Great Moderation period.

[7]GDI is also estimated in a bottom up fashion, estimating income from various categories (wages and salaries, profits, etc.) and then aggregating. Most of the income-side data is ultimately based on tax and adminstrative records, rather than samples as is the case for GDP.

[8]The "latest" time series employed in this paper are as they appeared close to the end of the Great Moderation, in August 2007. The results in section 6 are similar using later data vintages.

months after the quarter ends, and the variances of the "3rd" GDP and GDI growth estimates are almost equal).[9] Subsequent annual and benchmark revisions incorporate more comprehensive and higher-quality source data, plausibly reducing measurement error in the estimates, either LoSE or CME. If the bulk of the measurement error eliminated by the revisions is LoSE, so the revisions mainly add news to the estimates as in Mankiw and Shapiro (1986), the variance of the estimates should increase as in table 1. Moreover, the revisions increase the variance of GDI growth *more than* the variance of GDP growth, consistent with the revisions adding more news to GDI growth than GDP growth. The implication is that GDP growth is missing some news or signal, and is thus contaminated with LoSE.

The statistics in Table 1 suggest that, of these output growth estimates, "advance" GDP growth is contaminated with the most LoSE, as one would expect since it is based on the least amount of information. Somewhat counterintuitively, it is just such variables that are likely to meet the conditions of Assumption 2 and provide valid instruments W, motivating the instrumenting strategy below. The baseline regressions are of an output growth estimate on current and lagged stock and bond prices, which may reflect some information missing from the output growth estimates.[10] Since stock and bond prices

[9]Each quarterly observation in the "advance" or "3rd" time series is the estimate for that quarter released about one or three months after that quarter ends.

[10]Dynan and Elmendorf (2001) show that asset prices predict revisions to GDP growth, evidence that asset prices contain information missed by the initial estimates of GDP growth. Asset prices may contain information missed by the fully-revised estimates of GDP growth as well, entering through publicly-available information about the state of the economy not fully incorporated into GDP growth, such as the source data used to compute GDI, or as the aggregation of all the private information of asset market participants.

are measured with little error, we have:

$$\begin{aligned}\Delta Y^{\star} &= X\beta + U^{\star} \\ \Delta Y^{i} &= X\widehat{\beta}^{i} + U^{i},\end{aligned}$$

where i indexes output growth estimates. In this case, the instruments must be uncorrelated with $E\left(X|Z^{y^i}\right) - X$, which is the information missing from output growth estimate i that is captured by the asset prices X. Paradoxically, an instrument based on a *smaller* information set, while remaining correlated with X, is more likely to be uncorrelated with this missing information and thus meet the conditions of Assumption 2.

In particular, contemporaneous and lagged "advance" GDP growth rates are presumably in the information sets used to compute the various output growth estimates examined here, all released after the "advance" estimate. The identifying assumption employed here is that the "advance" GDP growth estimate for each quarter, and lagged "advance" estimates, are uncorrelated with whatever information remains missing from later, revised estimates of GDP growth for that quarter. Subcomponents of "advance" GDP growth are likely in the information sets used to compute those later, revised GDP growth estimates as well. Equipment and software (E&S) investment is an appealing subcomponent to use as an instrument because it produces a high first-stage R-square, its growth rate being highly correlated with stock price changes and bond spreads as predicted by Q-theory—see Tobin (1969) and Philippon (2009). For this reason, current and lagged "advance" growth rates of real E&S investment are the main set of instruments W employed in the paper.

6 Regression Evidence for LoSE in GDP growth in the Great Moderation

Under the model of section 4, the OLS β^i estimated in this section are governed by equation (7) with X measured without error, and we have:

$$\widehat{\beta^{GDI}} - \widehat{\beta^{GDP}} \xrightarrow{p} (Q_{xx})^{-1} \left(\sigma^2_{\zeta^{GDP},xy} - \sigma^2_{\zeta^{GDI},xy}\right) \beta, \tag{9}$$

where $\sigma^2_{\zeta^i,xy}$ is the variance of bias-inducing LoSE ζ^{xy} in estimate ΔY^i. When X is univariate, a test of $\widehat{\beta^{GDI}} = \widehat{\beta^{GDP}}$ equivalent to a test of $\sigma^2_{\zeta^{GDP},xy} = \sigma^2_{\zeta^{GDI},xy}$ if Q_{xx} and β are treated as constants. $|\widehat{\beta^{GDI}}| > |\widehat{\beta^{GDP}}|$ is then consistent with positive bias-inducing LoSE variance in ΔY^{GDP}.[11] Furthermore, for the instrumental variables estimates reported below, standard Durbin-Wu-Hausman tests (see Hausman (1978)) are available to test whether the OLS estimates $\widehat{\beta^{GDP}}$ are biased towards zero as in (7).

Table 2 shows estimation results using as the explanatory variable X an average of current and lagged stock price growth; standard errors are in parentheses.[12] Such a specification can be motivated in several ways, but for our purposes, it suffices that a relation between true output growth $\Delta Y^\star$ and stock prices X exists governed by a true parameter vector β.[13] Comparing the first two specifications of Table 2, we see that β

[11] An earlier working paper reported tests of equality between the βs across regressions, which Monte-carlo simulations (of an environment where one dependent variable is more contaminated with LoSE than another) showed have the expected properties of a textbook t-statistic.

[12] The standard errors are corrected for heteroskedasticity and autocorrelation, although there is little evidence of either. Standard errors computed under the assumption of i.i.d. errors are very similar to those reported here.

[13] An earlier working paper examined multivariate specifications of Table 2 that estimated the coefficient on each lag separately, which yielded results very similar to those reported here. The main results

increases when switching the dependent variable from "advance" GDP growth to latest GDP growth, consistent with LoSE in "advance" GDP growth.[14] Switching from latest GDP growth to latest GDI growth, β increases again, consistent with LoSE in not only the "advance" GDP growth estimates, but also latest, revised GDP growth.

Of course, other explanations for this result are possible, but appear less likely. First, alternative measurement error models that do not meet the restrictions of section 3 could hold. Appendix A examines such a model in which GDP and GDI growth are crudely rescaled versions of true output growth, and finds that it is inconsistent with results from reverse regressions. Second, and more obviously, stock prices could be reacting to estimates of corporate profits, which are a component of GDI, more than to output. However, if this were true, β should be particularly large using the initial estimates of GDI (and profits) to which the stock market reacts in real time. The fourth column of the table shows this is not the case. Moreover, the fifth column shows that β actually increases when corporate profits are stripped out of GDI.

The last specification of table 2, the instrumental variables estimate, does not use GDI growth at all, and is consistent with even more LoSE-inducing bias in latest GDP growth than is evident based on the comparison with GDI growth. In particular, this estimate implies attenuation of the OLS β computed using latest GDP growth of about

here are also robust to the inclusion of control variables such as lags of the output growth measures. The stock price changes are quarterly growth rates of the Wilshire 5000 stock price index, while the output growth measures are annualized quarterly growth rates as in table 1, so the effect on the level of output in percentage points of a permanent 1 percent stock price increase is roughly the reported coefficient divided by 4. The stock price index is nominal, and the results change little if the stock price index is deflated.

[14] Substituting either the 2nd or 3rd GDP growth estimates for the "advance" estimates yields a very similar β of 0.13, consistent with these early revisions not adding any of the missing signal that is reflected in stock price movements. The greater signal in the latest estimates is added in subsequent annual revisions.

60 percent.[15] The Durbin-Wu-Hausman test rejects the hypothesis of no bias in that OLS β with a p-value of 0.02.

Table 3 shows similar results using bond spreads—the difference in yield between 10-year and 2-year US treasury notes (TERM), and the difference in yield between corporate bonds and 10-year treasury notes (DEF).[16] Many papers have used similar variables to forecast output growth; see Chen (1991) and Estrella and Hardouvelis (1991), for example. The results (where each pair of βs is from a separate regression) provide almost uniform evidence favoring LoSE-induced attenuation of the OLS coefficients computed using either "advance" or latest, revised GDP growth. All of the β_{DEF} coefficients increase in absolute value when switching the dependent variable from "advance" to latest GDP growth and again when switching from latest GDP growth to latest GDI growth. Similarly, all of the β_{TERM} coefficients increase when switching from latest GDP growth to latest GDI growth except for $k \leq 2$, horizons where the explanatory power of TERM is weakest.

The instrumental variables estimates in Table 3 are generally consistent with even more LoSE-inducing bias in latest GDP growth than is evident from the comparison with GDI growth. The Durbin-Wu-Hausman tests reject the hypothesis of no bias in the OLS βs computed using latest GDP growth, with p-values ranging from 0.07 (for $k = 1$) to 0.002 (for $k = 4$ and $k = 5$). The instruments are highly correlated with DEF at all horizons k, with high first-stage R^2s. The very large instrumental variables

[15] This result is robust to the choice of instruments likely to meet the conditions of Assumption 2. In particular, estimates using only lagged "advance" $E\&S$ growth rates, excluding the contemporaneous growth rate from W, yield a β of 0.47. Substituting "advance" GDP growth for "advance" $E\&S$ growth in W cuts down on the first-stage R^2 considerably, but yields the same β of 0.47.

[16] 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.

estimates of β_{TERM} should be discounted for $k \leq 2$ as the instruments are weak, but for the longer horizons where the instruments have higher first-stage R^2s, the IV β_{TERM}s are consistent with LoSE-induced attenuation of the OLS β_{TERM}s computed using latest GDP growth of between 50 and 70 percent. This degree of attenuation bias, similar to that found using stock prices, could be related to some puzzles regarding the continued forecasting power of the yield curve, as noted in Rudebusch and Williams (2009). In particular, the LoSE in GDP growth may have masked the long-horizon forecasting power of the yield curve to forecasters focused on predicting GDP growth rather than recessions, which are dated based on a broad array of indicators, including income data, that may be less contaminated with LoSE than is GDP growth.

7 Application: Underestimating the Depth of the Great Recession

The key macroeconomic forecasting question at the end of 2008 and early 2009 was, given the extraordinary turmoil in financial markets, how sharply would the real economy turn down? Financial markets had already tanked by that time, so the issue was how to translate the information in financial markets into a forecast of real economic activity. Blue Chip Consensus Forecast for the unemployment rate issued in January 2009 (the solid blue line in Figure 2) underpredicted the actual rise in the unemployment rate (the black line) by a wide margin, as did even the average of the top ten Blue Chip forecasts (the dashed blue line). Publicly-available government forecasts, such as the red line, did not do much better (see Romer and Bernstein, 2009). Macroeconomic analysts typically use an Okun's law-type relation to translate GDP forecasts into unemployment rate forecasts, so LoSE in GDP growth may have contributed to these forecast errors. The LoSE in the "advance" 2008Q4 GDP growth estimate is obvious: it badly under-estimated the severity of the downturn, revising down from -3.8 percent (annualized) to -6.3 percent two months later and even more subsequently, and this poor initial estimate may have contributed to the overly-optimistic unemployment rate forecasts. More subtly, LoSE in the latest available GDP growth estimates over the preceding Great Moderation period may have been problematic for forecasting the depth of the Great Recession as well.[17]

Consider the OLS regressions from table 3 using latest GDP growth.[18] Figure 2 plots three additional forecasts of the unemployment rate, the green solid, dashed and dotted lines, using the first difference of the unemployment rate, real GDI growth, and real GDP growth as they appeared in December 2008 as dependent variables in the regression specification in Table 4.[19] The forecasts for GDI growth and GDP growth are translated into unemployment rate forecasts using an Okun's law relation.[20]

[17]Note that while this was an important episode in the history of macroeconomic forecasting, the results in this section are meant to be illustrative only. For more comprehensive out-of-sample forecast analyses, see Koenig, Dolmas and Piger (2003) and Clements and Galvao (2013).

[18]These regressions were posted to the Federal Reserve Board web site in March 2008, prior to massive intensification of financial market turmoil discussed in this section.

[19]Specifically, the forecasts for 2008Q4, 2009Q1, 2009Q2, 2009Q3, and 2009Q4 are predicted values from five regressions as in table 4 (for $k = 0, 1, 2, 3, \text{and } 4$), with the average values for the corporate bond spread and the slope the yield curve in December 2008 used to produce predicted values. The average level of the high-yield corporate bond spread was almost 20 percentage points in December 2008, compared to an average level of around 4 percentage points during expansions. During the previous two recessions, this spread had peaked at around 10 percentage points.

[20]This is estimated by regressing the quarterly change in the unemployment rate on the contemporaneous quarterly output growth measure and two of its lags, using a 1959Q4 to 2008Q3 sample. Note that, if the primary source of measurement error in the output growth measures is LoSE, these regressions yield consistent parameter estimates since the LoSE-ridden variables are explanatory, and the downward biases from the first stage regressions using bond spreads are passed through to the unemployment rate forecasts. In contrast, in the crude rescaling model outlined in Appendix A, the

The unemployment rate forecasts produced directly from bond spreads (the solid green line) track the rise in unemployment almost perfectly over the first three quarters of the projection, before overshooting in the second half of 2009. The Okun's law translation of the GDP growth projection (the dotted green line) undershoots these direct forecasts of the unemployment rate by about a half a percentage point in 2008Q4 and one and a half percentage points in 2009Q4. Interestingly, the Okun's law translation of the GDI growth projection (the dashed green line) also undershoots the direct forecasts of the unemployment rate, and the unemployment rate itself for much of the forecast period. But, since the bond spread coefficients are larger in absolute value, the undershooting is considerably less than using GDP. In particular, in the first half of 2009, more than half of the forecast error from the Okun's law translation of GDP growth disappears when we switch from GDP growth to GDI growth, likely because GDI is less contaminated with LoSE. This suggests that, after financial markets tanked in late 2008, LoSE in GDP growth over the Great Moderation period contributed to the failure of conventional macroeconomic models and analysis to forecast the severity of the Great Recession. Had that analysis employed the information in GDI growth, instead of focusing solely on GDP growth, it might not have misread the signals from financial markets so badly.

8 Conclusions

The canonical classical measurement error (CME) model is too restrictive to handle important types of measurement error, including measurement error in one of the most

bias in the second stage regressions would largely offset the bias in the first stage regressions using bond spreads, which does not appear to be the case empirically.

widely-followed macroeconomic time series, US GDP growth. The paper studies a simple generalization of the CME model that is mathematically tractable, embeds the CME model as a special case, and adds useful flexibility. Instead of just allowing measurement error that *adds noise* to the true variable of interest, the generalization permits measurement errors that *subtract signal* from that variable, called Lack of Signal Errors, or LoSE, for short.

In some ways, this generalization of the CME model is the flip side of the coin regarding the effect of errors in variables on ordinary least squares regression. CME in the dependent variable of a regression Y does not bias parameter estimates and increases standard errors, and, in the baseline case studied here, LoSE in the *explanatory variables* X has the same effect. Of course, CME in the explanatory variables X does bias regression parameter estimates, towards zero in the univariate case; LoSE in the *dependent variable* Y introduces a similar attenuation bias under some circumstances, namely, when some of the signal missing from the dependent variable Y is captured by the explanatory variables X. LoSE in Y also shrinks the variance of the regression residuals, raising concerns about the robustness of hypothesis tests by increasing the probability of type I errors. In the limiting case of maximal LoSE, Y approaches a constant, and in a regression of Y on any non-constant variable X, $\widehat{\beta} = \frac{\text{cov}(X,Y)}{\text{var}(X)}$ and $\text{var}\left(\widehat{\beta}\right)$ approach zero, regardless of the true β. The result is badly attenuated parameter estimates, estimated with false precision. On a positive note, the results derived here provide some clear prescriptions for handling this type of attenuation, in terms of choice of instruments. The previous literature had not developed instrumenting strategies for dealing with bias from LoSE.

The paper provides evidence for LoSE not only in the initial GDP growth estimates based on limited source data, but also the latest, revised GDP growth estimates based on more comprehensive data. In particular, coefficients from regressions of the GDP growth estimates on a fixed set of stock or bond prices are smaller than coefficients from regressions that substitute for GDP an alternative measure of US output, GDI, that is likely more accurate than GDP over the Great Moderation period—see Nalewaik (2010) and Aruoba, Diebold, Nalewaik, Schorfheide, and Song (2012, 2013). These results are consistent with LoSE in GDP growth *even after it has passed through all of its revisions*. The paper shows that some other forms of non-classical measurement error cannot explain the differences in coefficients across these regressions. Furthermore, implementation of the instrumenting strategies derived in this paper, which rely in no way on the information in GDI growth, corroborate and provide independent amplifying evidence of substantial LoSE in latest, revised GDP growth over the Great Moderation period.

Some implications of significant LoSE in latest, revised GDP growth and its major subcomponents follow immediately. Those variables are simply less informative than many macroeconomists currently believe, given the common but incorrect presumption that the fully-revised estimates are measured with little error. And in a macroeconomic forecasting context, the attenuation biases discussed here can lead to serious mistakes. In particular, in late 2008 and early 2009, conventional macroeconomic analysis severly underestimated the size of the shocks that had hit the economy and that were already reflected in the behavior of asset prices. The paper demonstrates that part of that forecast error may have been due to the focus of conventional macroeconomic analysis on GDP growth: LoSE in GDP growth likely biased down the coefficients employed to translate asset prices into forecasts of output and unemployment. A better understanding of the implications of LoSE in GDP growth may help avoid such forecast errors in the future.

References

[1] Angrist, J., and Krueger, A. (1999), "Empirical Strategies in Labor Economics," in *Handbook of Econometrics* (Vol. 5), eds. O. Ashenfelter and D. Card, Amsterdam: Elsevier.

[2] Aruoba, S.B., F.X. Diebold, J. Nalewaik, F Schorfheide, and D. Song (2012), "Improving GDP Measurement: A Forecast Combination Perspective," In X. Chen and N. Swanson (eds.), *Recent Advances and Future Directions in Causality, Prediction, and Specification Analysis: Essays in Honour of Halbert L. White Jr.*, Springer, 1-26.

[3] Aruoba, S.B., F.X. Diebold, J. Nalewaik, F Schorfheide, and D. Song (2013), "Improving GDP Measurement: A Measurement Error Perspective," submitted by invitation, *Journal of Econometrics.*

[4] Berkson, J. (1950), "Are There Two Regressions?" *Journal of the American Statistical Association*, 45, 164-180.

[5] Bernstein, J., and Romer, C. (2009), "The Job Impact of the American Recovery and Reinvestment Plan," Council of Economic Advisers report, January 10, 2009.

[6] Bollinger, C. (1996), "Bounding Mean Regression When a Binary Regressor is Mismeasured," *Journal of Econometrics*, 73, 387-399.

[7] Bollinger, C. (1998), "Measurement Error in the Current Population Survey: A Nonparametric Look," *Journal of Labor Economics*, 16, 576-594.

[8] Bollinger, C., and Hirsch, T. (2006), "Match Bias from Earnings Imputation in the Current Population Survey: The Case of Imperfect Matching" *Journal of Labor Economics*, 24, 483-519.

[9] Bound, J., Brown, C., and Mathiowetz, N. (2001), "Measurement Error in Survey Data," in *Handbook of Econometrics* (Vol. 5), eds. J.J. Heckman and E. Leamer, Amsterdam: Elsevier.

[10] Bound, J., Brown, C., Duncan, G., and Rogers, W. (1994), "Evidence on the Validity of Cross-sectional and Longitudinal Labor Market Data" *Journal of Labor Economics*, 12, 345-368.

[11] Bound, J., and Krueger, A. (1991), "The Extent of Measurement Error in Longitudinal Earnings Data: Do Two Wrongs Make a Right?," *Journal of Labor Economics*, 9, 1-24.

[12] Campbell, J. and Mankiw, G. (1989), "Consumption, Income, and Interest Rates: Reinterpreting the Time Series Evidence," in *NBER Macroeconomics Annual*, eds. O. Blanchard and S. Fischer, Cambridge, NBER.

[13] Card, D., (1996), "The Effect of Unions on the Structure of Wages: A Longitudinal Analysis." *Econometrica*, 64, 957-979.

[14] Carroll, R., and Stefanski, L. (1990), "Approximate Quasi-likelihood Estimation in Models with Surrogate Predictors," *Journal of the American Statistical Association*, 85, 652-663.

[15] Chen, N. (1991), "Financial Investment Opportunities and the Macroeconomy," *Journal of Finance*, 46, 529-554.

[16] Clements, M. P., and Galvao, A. B. (2013), "Real-Time Forecasting of Inflation and Output Growth with Autoregressive Models in the Presence of Data Revisions," *Journal of Applied Econometrics*, 28, 458-477.

[17] de Leeuw, Frank, and McKelvey, Michael J. (1983), "A 'True' Time Series and Its Indicators" *Journal of the American Statistical Association*, 78, 37-46.

[18] Durbin, J. (1954), "Errors in Variables," *Review of the International Statistical Institute*, 22, 23-32.

[19] Dynan, K. and Elmendorf, D. (2001), "Do Provisional Estimates of Output Miss Economic Turning Points?" Board of Governors of the Federal Reserve System, FEDS working paper 2001-52.

[20] Escobal, J., and Laszlo, S. (2008), "Measurement Error in Access to Markets," *Oxford Bulletin of Economics and Statistics*, 70, 209-243.

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

[22] Fixler, D. and Nalewaik, J. (2007) "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.

[23] Federov, V. V. (1974), "Regression Problems with Controllable Variables Subject to Error," *Biometrika*, 61, 49-56.

[24] Fuller, W. (1987) *Measurement Error Models*, New York: John Wiley and Sons.

[25] Geary, R. C. (1953), "Non-Linear Functional Relationship Between Two Variables When One Variable is Controlled," *Journal of the American Statistical Association*, 48, 94-103.

[26] Griliches, Z. (1986), "Economic Data Issues," in *Handbook of Econometrics* (Vol. 3), eds. Z. Griliches and M.D. Intriligator, Amsterdam: Elsevier.

[27] Hausman, J., (1978), "Specification tests in econometrics," *Econometrica*, 46, 1251-1272.

[28] Hirsch, T., and Schumacher, E. (2004), "Match Bias in Wage Gap Estimates Due to Earnings Imputation" *Journal of Labor Economics*, 22, 689-722.

[29] Huwang, L, and Huang, Y. H. (2000), "On Errors-In-Variables in Polynomial Regression-Berkson Case," *Statistica Sinica*, 10, 923-936.

[30] Hyslop, R., and Imbens, Guido R. (2001). "Bias from Classical and Other Forms of Measurement Error," *Journal of Business and Economic Statistics,* 19, 475-481.

[31] Jacobs, J. and van Norden, S. (2011), "Modelling Data Revisions: Measurement Error and the Dynamics of "True" Values," *Journal of Econometrics*, 161, 101-109.

[32] Kane, T., Rouse, C., and Staiger, D. (1999), "Estimating Returns to Schooling when Schooling is Misreported," working paper 7235, NBER, Cambridge, MA.

[33] Koenig, E., Dolmas, S., and Piger, J. (2003). "The Use and Abuse of Real-Time Data in Economic Forecasting," *The Review of Economics and Statistics,* 85, 618-628.

[34] Kimball, Miles; Sahm, Claudia; and Shapiro, Matthew (2008), "Imputing Risk Tolerance from Survey Responses" *Journal of the American Statistical Association*, 103, 1028-1038.

[35] Kishor, N. K. and Koenig, E. (2011). "VAR Estimation and Forecasting when Data are Subject to Revision," *Journal of Business and Economic Statistics,* 29.

[36] Klepper, S., and Leamer, E., (1984), "Consistent Sets of Estimates for Regressions with Errors in All Variables," *Econometrica*, 52, 163-183.

[37] Leamer, E., (1987), "Errors in Variables in Linear Systems," *Econometrica*, 55, 893-909.

[38] Mankiw, N. and Shapiro, M., (1986), "News or Noise: An Analysis of GNP Revisions" *Survey of Current Business*, 66, 20-25.

[39] Nalewaik, J., (2010), "The Income- and Expenditure-Side Estimates of U.S. Output Growth," *Brookings Papers on Economic Activity*, 1, 71-106.

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

[41] Philippon, T., (2009), "The Bond Market's q," *Quarterly Journal of Economics*, 124, 1011-1056.

[42] Piscke, J-S. (1995). "Measurement Error and Earnings Dynamics: Some Estimates from the PSID Validation Study," *Journal of Business and Economic Statistics,* 13, 305-314.

[43] Rudebusch, G., and Williams, J. (2009). "Forecasting Recessions: The Puzzle of the Enduring Power of the Yield Curve," *Journal of Business and Economic Statistics,* 27, 492-503.

[44] Sargent, T. (1989). "Two Models of Measurement and the Investment Accelerator," *Journal of Political Economy,* 97, 251-287.

[45] Tobin, J., (1969), "A General Equilibrium Approach to Monetary Theory" *Journal of Money, Credit, and Banking*, 1, 15-29.

[46] Wang, L. (2003), "Estimation of Nonlinear Berkson-Type Measurement Error Models," *Statistica Sinica*, 13, 1201-1210.

[47] Wang, L. (2004), "Estimation of Nonlinear Models with Berkson Measurement Errors," *Annals of Statistics*, 32, 2559-2579.

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

Quarterly Data, 1984Q3-2004

Vintage	var $\left(\Delta Y_t^{GDP}\right)$	var $\left(\Delta Y_t^{GDI}\right)$
Current Quarterly, "Advance"	3.1	.
Current Quarterly, "3rd"	4.1	4.0
Latest Vintage Available	4.2	4.8

Note: Each quarterly observation in the "advance" or "3rd" time series is the estimate for that quarter released about one or three months after that quarter ends.

Table 2: Regressions of Different Measures of Quarterly Output Growth on Current and Lagged Stock Price Growth, 1984Q3 to 2004Q4:

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

Measure:	ΔY^{GDP}	ΔY^{GDP}	ΔY^{GDI}	ΔY^{GDI}	ΔY^{GDI-CP}	ΔY^{GDP}
Vintage:	"Advance"	Latest	Latest	"3rd"	Latest	Latest
Estimation:	OLS	OLS	OLS	OLS	OLS	IV
β:	**0.142**	**0.214**	**0.325**	**0.152**	**0.389**	**0.522**
	(0.060)	(0.068)	(0.073)	(0.075)	(0.078)	(0.213)

Note: The instruments are the time t "advance" growth rate of real equipment and software investment, scaled by its share of nominal GDP to approximate contributions to real GDP growth, and 6 of its lags; the first stage R^2 is 0.22.

Table 3: Regressions of Different Measures of Quarterly Output Growth on Lagged Interest Rates Spreads (TERM and DEF), 1988Q3 to 2004Q4:

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

Measure:	ΔY^{GDP}, "Advance"		ΔY^{GDP}, Latest		ΔY^{GDI}, Latest		ΔY^{GDP}, Latest	
Estimation:	OLS		OLS		OLS		IV, E&S	
	β_{TERM}	β_{DEF}	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**	**2.83**	**-1.11**
	(0.26)	(0.13)	(0.26)	(0.13)	(0.29)	(0.10)	(2.68)	(0.36)
k=2	**0.42**	**-0.44**	**0.48**	**-0.53**	**0.43**	**-0.69**	**2.75**	**-0.79**
	(0.26)	(0.12)	(0.31)	(0.12)	(0.33)	(0.13)	(1.18)	(0.27)
k=3	**0.58**	**-0.38**	**0.60**	**-0.40**	**0.68**	**-0.65**	**1.68**	**-0.64**
	(0.30)	(0.12)	(0.36)	(0.15)	(0.37)	(0.15)	(0.59)	(0.20)
k=4	**0.62**	**-0.23**	**0.57**	**-0.28**	**0.70**	**-0.50**	**1.87**	**-0.49**
	(0.32)	(0.15)	(0.39)	(0.17)	(0.40)	(0.17)	(0.55)	(0.21)
k=5	**0.59**	**-0.19**	**0.67**	**-0.29**	**0.75**	**-0.41**	**1.97**	**-0.41**
	(0.35)	(0.14)	(0.38)	(0.14)	(0.44)	(0.19)	(0.63)	(0.23)
k=6	**0.72**	**-0.27**	**0.76**	**-0.32**	**0.92**	**-0.39**	**1.75**	**-0.31**
	(0.35)	(0.10)	(0.38)	(0.13)	(0.41)	(0.16)	(0.63)	(0.24)
k=7	**0.73**	**-0.19**	**0.81**	**-0.20**	**0.96**	**-0.39**	**1.84**	**-0.18**
	(0.35)	(0.10)	(0.36)	(0.13)	(0.38)	(0.15)	(0.73)	(0.22)
k=8	**0.66**	**-0.10**	**0.72**	**-0.15**	**0.94**	**-0.27**	**1.72**	**-0.15**
	(0.34)	(0.13)	(0.36)	(0.14)	(0.37)	(0.15)	(0.71)	(0.21)

Note: The instruments are the time t "advance" growth rate of real equipment and software investment, scaled by its share of nominal GDP to approximate contributions to real GDP growth, and k of its lags. The first stage R^2s for

DEF range from 0.43 to 0.53, depending on k. The first stage R^2s for TERM range from 0.01 ($k = 1$) to 0.26 ($k = 8$).

Appendix A:
Alternative forms of mismeasurement

Start with the conditioning information set Z_t, and assume $E(Y_t^\star|Z_t) = Z_t\gamma$. In an alternative form of mismeasurement, the estimate Y_t misuses Z_t, so $Y_t = Z_t\widetilde{\gamma} + \varepsilon_t$ with $\widetilde{\gamma} \neq \gamma$. The estimate "misses" in a systematic way, inconsistent with the efficiency assumptions of section 2. For estimation and inference about $Y^\star$ (for example in regressions), these systematic "misses" clearly lead to biased and inconsistent estimates. Unless additional information is available about the nature of $Z_t\widetilde{\gamma} - Z_t\gamma$, the direction and magnitude of these biases is unclear, but in highly stylized examples the biases may be derived. One such example is $Y_t = \alpha_0 + \alpha_1 Y_t^\star + \varepsilon_t$, with $\alpha_0 \neq 0$ and $\alpha_1 \neq 1$, and ε_t noise. This model is employed by de Leeuw and McKelvey (1983), Bound, Brown, Duncan and Rodgers (1994), Pischke (1995), and Bound, Brown and Mathiowetz (2001).

In the case of latest GDP and GDI growth, ignoring constants, consider:

$$Y^{GDP} = \alpha^{GDP} Y^\star + \varepsilon^{GDP} \quad \text{and:} \quad Y^{GDI} = \alpha^{GDI} Y^\star + \varepsilon^{GDI},$$

with ε^{GDP} and ε^{GDI} noise. In this model, the regressions in table 2 pin down the relative αs, since:

$$\frac{\widehat{\beta^{GDI}}}{\widehat{\beta^{GDP}}} \xrightarrow{p} \frac{\alpha^{GDI}}{\alpha^{GDP}}. \tag{10}$$

Interestingly, reverse regressions $X = Y\widehat{\beta}_r + U_r$ yield:

$$\frac{\widehat{\beta_r^{GDI}}}{\widehat{\beta_r^{GDP}}} \xrightarrow{p} \frac{\alpha^{GDP}\,\mathrm{var}(Y^\star) + \frac{\sigma^2_{\varepsilon^{GDP}}}{\alpha^{GDP}}}{\alpha^{GDI}\,\mathrm{var}(Y^\star) + \frac{\sigma^2_{\varepsilon^{GDI}}}{\alpha^{GDI}}}. \tag{11}$$

While an increase in α^{GDP}, ceteris paribus, decreases the ratio (10) from the forward regression, if $\mathrm{var}(Y^\star) > \sigma^2_{\varepsilon^{GDP}}$, it increases the ratio (11) from the reverse regression. So, if the variance of true GDP growth exceeds the variance of the noise in measured GDP growth (and GDI growth), which seems plausible, these ratios (10) and (11) move in opposite directions with respect to α^{GDP} (and α^{GDI}) under this crude rescaling model.

Table 2 implies $\alpha^{GDI}/\alpha^{GDP} = 1.5$, so we should observe $\beta_r^{GDI}/\beta_r^{GDP} = 0.66$ in the reverse regression in table 2A with no noise in either estimate. The ratio is very far from that, 1.32. Adding some noise variance to both estimates takes the implied ratio from (11) closer to 1, but it exceeds 1 only if the noise variance in GDP growth exceeds the noise variance in GDI growth by an implausibly large amount. Specifically, a ratio of 1.32 could be consistent with (11) if half the variance of GDP growth were noise uncorrelated with GDI growth, but the estimated covariance between GDP and GDI growth is larger than half the variance of GDP growth, so this is highly unlikely: a test of the hypothesis that this covariance (about 3.1 over the 1984Q3 to 2004Q4 sample) is half the variance of GDP growth (about 4.2) rejects with a p-value of 0.01.[21]

Similarly, univariate specifications similar to table 3 but using only DEF imply $\alpha^{GDI}/\alpha^{GDP}$ ranging from 1.3 to 2.0, as can be seen comparing the third and fourth

[21]These calculations assume no noise in GDI growth, but examination of (11) shows that allowing for noise in GDI growth only increases the already implausibly-large fraction of the variance of GDP growth that must be noise under the crude rescaling model. Allowing a realistic amount of noise in GDI growth, then, only reduces the plausibility of the crude rescaling model.

columns of table 3A. However, comparing the sixth and seventh columns, we see the coefficients using GDI growth as the explanatory variable are once again *larger* in absolute value than the coefficients using GDP growth, inconsistent with the crude rescaling model and plausible assumptions about the noise variances.

By contrast, the generalized LoSE model outlined in section 3 yields the following for the reverse regressions:

$$\widehat{\beta_r^{GDI}} - \widehat{\beta_r^{GDP}} \xrightarrow{p} \left(\sigma^2_{\varepsilon^{GDP}} - \sigma^2_{\varepsilon^{GDI}} \frac{\mathrm{var}(Y^{GDP})}{\mathrm{var}(Y^{GDI})}\right) \frac{\beta_r}{\mathrm{var}(Y^{GDP})}. \tag{12}$$

An increase in $\widehat{\beta^{GDI}}$ relative to $\widehat{\beta^{GDP}}$ implies an increase in the variance of LoSE in Y^{GDP} relative to Y^{GDI}, which *reduces* $\frac{\mathrm{var}(Y^{GDP})}{\mathrm{var}(Y^{GDI})}$ and *increases* $\widehat{\beta_r^{GDI}} - \widehat{\beta_r^{GDP}}$, all else equal. This model is much more consistent with the results from the reverse regressions.

Table 2A: Reverse Regressions of Current and Lagged Stock Price Growth on Different Measures of Quarterly Output Growth, 1984Q3 to 2004Q4:

$$(\Delta p_t + \Delta p_{t-1} + \ldots + \Delta p_{t-6})/7 = \alpha + \beta_r \Delta Y_t^i + U_{r,t}^i$$

Measure:	ΔY^{GDP}	ΔY^{GDP}	ΔY^{GDI}	ΔY^{GDI}	ΔY^{GDI-CP}
Vintage:	"Advance"	Latest	Latest	"3rd"	Latest
β:	**0.411**	**0.454**	**0.600**	**0.338**	**0.497**
	(0.194)	(0.182)	(0.169)	(0.161)	(0.140)

Table 3A: Regressions of Lagged Interest Rates Spreads (DEF) on Different Measures of Quarterly Output Growth, 1988Q3 to 2004Q4:

Forward: $\Delta Y_t^i = \alpha + \beta \left(r_{t-k}^{corp} - r_{t-k}^{10yr} \right) + U_t^i$

Reverse: $\left(r_{t-k}^{corp} - r_{t-k}^{10yr} \right) = \alpha + \beta_r \Delta Y_t^i + U_t^i$

Measure:	Forward βs			Reverse β_rs		
Measure:	ΔY^{GDP}	ΔY^{GDP}	ΔY^{GDI}	ΔY^{GDP}	ΔY^{GDP}	ΔY^{GDI}
Vintage:	"Advance"	Latest	Latest	"Advance"	Latest	Latest
k=1	**-0.47**	**-0.58**	**-0.76**	**-0.48**	**-0.46**	**-0.52**
	(0.13)	(0.13)	(0.11)	(0.12)	(0.11)	(0.10)
k=2	**-0.40**	**-0.48**	**-0.64**	**-0.39**	**-0.38**	**-0.43**
	(0.13)	(0.13)	(0.14)	(0.12)	(0.10)	(0.09)
k=3	**-0.31**	**-0.33**	**-0.57**	**-0.30**	**-0.26**	**-0.38**
	(0.14)	(0.15)	(0.17)	(0.12)	(0.11)	(0.09)
k=4	**-0.15**	**-0.21**	**-0.41**	**-0.15**	**-0.17**	**-0.28**
	(0.16)	(0.16)	(0.17)	(0.15)	(0.12)	(0.10)

Figure 1:
1984 to 2004 Growth Rates of Real GDP and Real GDI,
Latest Available data as of August 2008

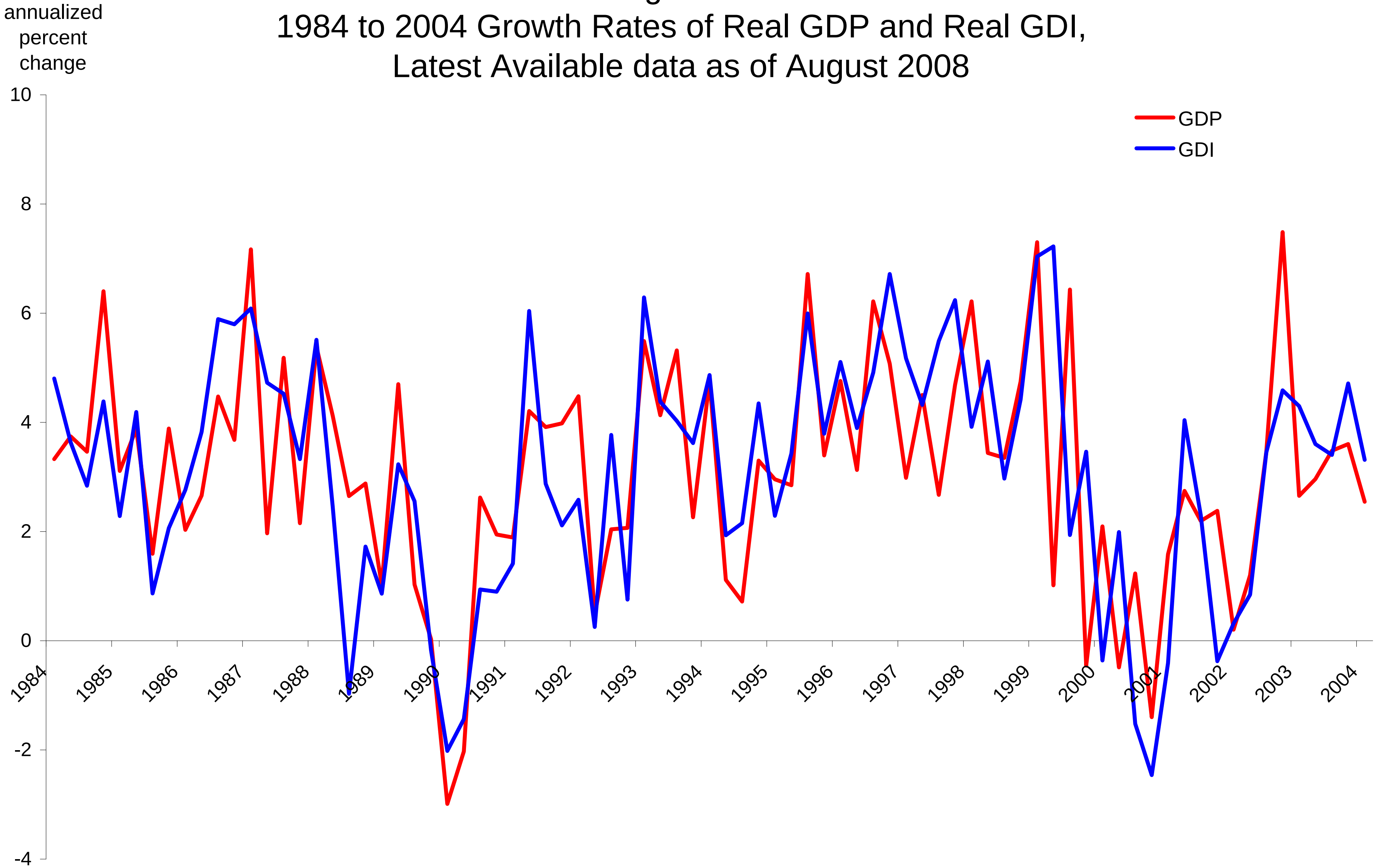
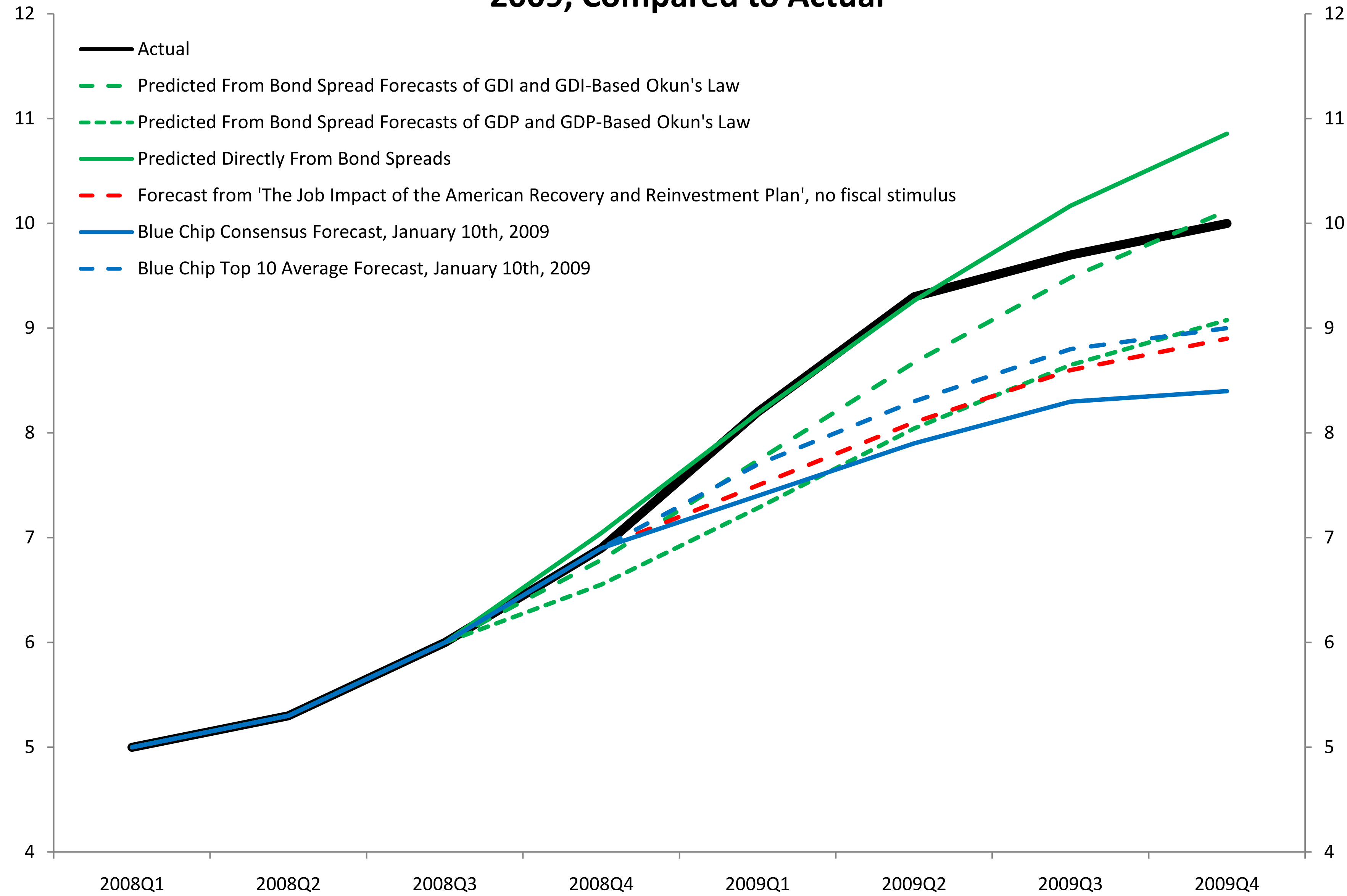


Figure 2: Unemployment Rate Forecasts at the End of 2008 or Early 2009, Compared to Actual



Appendix:
Regression with Measurement Error in both X and Y, and examples highlighting the implications of LoSE

This appendix discusses regression when both X and Y follow the generalized measurement error model of section 3, before examining special cases highlighting the most important implications of LoSE in X and Y for parameter estimates and standard errors. Our full set of assumptions follows:

Assumption 1 *$Y_t^\star = X_t^\star\beta + U_t^\star$. $U_t^\star$ is i.i.d., mean zero, with $\mathrm{var}\left(U_t^\star\right) = \sigma^2_{U^\star}$ and $U_s^\star$ independent of $X_t^\star$, $\forall t, s$. Measured $Y_t = E\left(Y_t^\star | Z_t^y\right) + \varepsilon_t$, with:*

- *The CME ε_t is i.i.d., mean zero, and independent of all conditioning information sets, with $\mathrm{var}\left(\varepsilon_t\right) = \sigma^2_\varepsilon$.*
- *Z^y may be partitioned into two sets of variables, Z_x^y and Z_u^y, with variables in Z_x^y independent of $U^\star$ and Z_u^y, and variables in Z_u^y independent of $X^\star$ and Z_x^y.*
- *The LoSE $\zeta_t = \left(X_t^\star - E\left(X_t^\star | Z_{x,t}^y\right)\right)\beta + \left(U_t^\star - E\left(U_t^\star | Z_{u,t}^y\right)\right) = \zeta_t^{xy}\beta + \zeta_t^u$. ζ_t^u is i.i.d. and mean zero with $\mathrm{var}\left(\zeta_t^u\right) = \sigma^2_{\zeta,u}$, and ζ_t^{xy} is i.i.d. and mean zero with $\mathrm{var}\left(\zeta_t^{xy}\right) = \sigma^2_{\zeta,xy}$, a $k \times k$ matrix.*

Measured $X_t = E\left(X_t^\star | Z_t^x\right) + \varepsilon_t^x$, with:

- *The CME ε_t^x is i.i.d., mean zero, independent of ε_t and all conditioning information sets, with $\mathrm{var}\left(\varepsilon_t^x\right) = \sigma^2_{\varepsilon,x}$, a $k \times k$ matrix.*
- *The variables in Z^x are independent of $U^\star$ and Z_u^y.*

- *The LoSE* $\zeta_t^x = X_t^\star - E\left(X_t^\star | Z_t^x\right)$ *is i.i.d. and mean zero with* $\mathrm{var}\left(\zeta_t\right) = \sigma^2_{\zeta,x}$*, a* $k \times k$ *matrix.*

- *As* $T \longrightarrow \infty$*:*

  - $\frac{1}{T}\left(X^\star\right)' X^\star \xrightarrow{p} Q_{xx}$
  - $\frac{1}{T}\left(E\left(X^\star | Z_x^y\right)\right)' E\left(X^\star | Z_x^y\right) \xrightarrow{p} Q_{xx}^{zy} = Q_{xx} - \sigma^2_{\zeta,xy}$
  - $\frac{1}{T}\left(E\left(X^\star | Z^x\right)\right)' E\left(X^\star | Z^x\right) \xrightarrow{p} Q_{xx}^{zx} = Q_{xx} - \sigma^2_{\zeta,x}$
  - $\frac{1}{T}\left(E\left(X^\star | Z_x^y\right)\right)' E\left(X^\star | Z^x\right) \xrightarrow{p} Q_{xx}^{zb}$
  - $\frac{1}{T}X'X \xrightarrow{p}= Q_{xx}^{zx} + \sigma^2_{\varepsilon,x}.$

All relevant fourth moments exist.

The assumptions imposed on the information sets Z^y and Z^x regarding partitioning and independence allow us to factor the joint distribution of the relevant variables as follows:

$$f\left(U^\star, X^\star, Z^y, Z^x\right) \quad = \quad f_{UZ}\left(U^\star, Z_u^y\right) f_{XZ}\left(X^\star, Z_x^y, Z^x\right).$$

Without these assumptions, the conditioning may introduce correlation between the measurement error in X and the regression residual (which includes the measurement error in Y). For example, assume the information sets Z_x^y and Z_u^y are univariate and $Z^x = Z_u^y + Z_x^y$; then $E\left(X_t^\star | Z_t^x\right)$ and ζ_t^x are correlated with $U^\star$ (as long as Z_u^y captures some variation in $U^\star$), and the above factorization is not valid. Another example is in section 3.1 of HI, and while these biases are likely worthy of further empirical study, they are of a different nature from those introduced by LoSE, and studying them in detail is beyond the scope of this paper.[1]

[1]Most important, these biases are not applicable to regressions using X variables measured without

Given assumption 1, Y_t can be written as:

$$\begin{aligned} (1) \qquad Y_t &= E\left(X_t^\star | Z_{x,t}^y\right)\beta + E\left(U_t^\star | Z_{u,t}^y\right) + \varepsilon_t \\ &= X_t\beta + \left(E\left(X_t^\star | Z_{x,t}^y\right) - X_t\right)\beta + E\left(U_t^\star | Z_{u,t}^y\right) + \varepsilon_t \\ &= X_t\beta + \left(E\left(X_t^\star | Z_{x,t}^y\right) - E\left(X_t^\star | Z_t^x\right) - \varepsilon_t^x\right)\beta + U_t^\star - \zeta_t^u + \varepsilon_t. \end{aligned}$$

The OLS regression estimator is:

$$\begin{aligned} \widehat{\beta} &= \left(X'X\right)^{-1} X'Y \\ (2) \qquad &= \beta + \left(X'X\right)^{-1} X'\left(\left(E\left(X^\star | Z_x^y\right) - E\left(X^\star | Z^x\right) - \varepsilon^x\right)\beta + U^\star - \zeta^u + \varepsilon\right). \end{aligned}$$

Taking expectations and probability limits of (2) yields:

$$\begin{aligned} (3) \qquad E\left(\widehat{\beta}\right) &= \beta + E\left(\left(X'X\right)^{-1} X'\left(E\left(X^\star | Z_x^y\right) - E\left(X^\star | Z^x\right) - \varepsilon^x\right)\right)\beta, \quad \text{and:} \\ (4) \qquad \widehat{\beta} &\xrightarrow{p} \beta + \left(Q_{xx}^{zx} + \sigma_{\varepsilon,x}^2\right)^{-1}\left(Q_{xx}^{zb} - Q_{xx}^{zx} - \sigma_{\varepsilon,x}^2\right)\beta. \end{aligned}$$

The usual attenuation bias and inconsistency from $\sigma_{\varepsilon,x}^2$ is evident. The additional inconsistency from LoSE depend on the difference between Q_{xx}^{zb} and Q_{xx}^{zx}.

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

error—see also section 3.2 of HI. All of the regressions in the empirical sections 6 and 7 of this paper use X variables measured without error.

To correct the biases in OLS, valid instruments must be uncorrelated with ε^x, a standard condition. However, an additional condition must be met: the instruments must be uncorrelated with $E(X^\star|Z_x^y) - E(X^\star|Z^x)$. This condition is met by instruments W that are common to both information sets (if such information exists), so $W \subset Z^x$ and $W \subset Z_x^y$, since $W'E(X^\star|Z_x^y)$ and $W'E(X^\star|Z^x)$ then have the same probability limit. With valid instruments, we have:

$$
\begin{aligned}
\widehat{\beta} &= \left(X'P_W X\right)^{-1} X'P_W Y \\
&= \beta + \left(X'P_W X\right)^{-1} X'P_W \left(\left(E(X^\star|Z_x^y) - E(X^\star|Z^x) - \varepsilon^x\right)\beta + U^\star - \zeta^u + \varepsilon\right), \qquad (5)
\end{aligned}
$$

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

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

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:

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

The first two terms converge in probability to $\sigma_{U^\star}^2 - \sigma_{\zeta,u}^2 + \sigma_\varepsilon^2$; 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: $s^2 \xrightarrow{p} \sigma^2_{U^\star} - \sigma^2_{\zeta,u} + \sigma^2_\varepsilon + \beta' \left(Q^{zy}_{xx} - 2Q^{zb}_{xx} + Q^{zx}_{xx} + \sigma^2_{\varepsilon,x}\right) \beta$.

Several specialized examples of this general measurement error model follow, highlighting the following important implications of LoSE in X and Y for parameter estimates and standard errors.

Example 1:

X Mismeasured, Y Not Mismeasured: No LoSE Problems

The LoSE in X, ζ^x, introduces no bias or inconsistency into the estimates, as long as all k explanatory variables are conditioned on the same information set Z^x. Similar to CME in Y, the only effect of LoSE in X is to increase the variance of the regression residuals.

Given the traditional focus on mismeasurement in X on regression estimation, we begin with this subsection making the following assumption (on top of assumption 1):

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

Then (4) simplifies to:

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

Not all of the true variation in $X_t^\star$ appears in X_t due to LoSE, but all of that variation does appear in $Y_t^\star$ through $X_t^\star \beta$. The 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 uncorrelated with $E\left(X^{\star}|Z^{x}\right) + \varepsilon^{x} = X$, 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 rely on the assumption that the LoSE is the difference between truth and a conditional expectation, and for multivariate regressions, the consistency result also relies on all k explanatory variables being conditioned on the same information set Z^x. Bound, Brown, and Mathiowetz (2001), and Kimball, Sahm, and Shapiro (2008) discuss the case where different elements of X are conditioned on different information sets, causing bias and inconsistency.[2]

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

$$E\left(\widehat{\beta}\right) = \beta - E\left(\left(X'X\right)^{-1} X'\varepsilon^{x}\right)\beta, \quad \text{and:} \tag{6}$$

$$\widehat{\beta} \xrightarrow{p} \beta - \left(Q_{xx}^{zx} + \sigma_{\varepsilon,x}^{2}\right)^{-1} \sigma_{\varepsilon,x}^{2}\beta. \tag{7}$$

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

[2]In that case, consistency may be acheived by instrumenting with variables from the smallest information set only.

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

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

Asymptotically, the analogous distributional results hold, as:

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

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.

Example 2:

Y Mismeasured, X Not Mismeasured, $X_t\in Z^{y}_{x,t}$: Shrunken Standard Errors

The ζ^{u} component of the LoSE in Y introduces no bias or inconsistency into the estimates, but decreases the variance of the regression residuals and standard errors.

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

Assumption 5 *X_t is not mismeasured: $X_t=X^{\star}_t$, and $X_t\in Z^{y}_{x,t}$.*

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

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

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

LoSE in $U^\star$ introduces no bias or inconsistency since Z_u^y is uncorrelated with X, so the overall measurement error in Y introduces no bias or inconsistency in this case.

For the variance of the point estimates, $\operatorname{var}\left(\widehat{\beta}\right) = E\left(\operatorname{var}\left(\widehat{\beta}|X\right)\right)$ since $\operatorname{var}\left(E\left(\widehat{\beta}|X\right)\right) = 0$, and:

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

since $E\left(U^\star|Z_u^y\right)$ and ε are uncorrelated. The analogous asymptotic results hold. The CME in Y increases the variance of the regression residuals and parameter estimates, and reduces the power of hypothesis tests, similar to LoSE in X. *The LoSE in Y has the opposite effect, decreasing the variance of the regression residuals and parameter estimates.*

Such excessively precise standard errors can be a serious problem, especially in the next example where $\widehat{\beta}$ is biased towards zero. As we approach the limiting case of maximal LoSE in Y where Y approaches a constant, $\widehat{\beta}$ and var $\left(\widehat{\beta}\right)$ both approach zero. Under this limiting case, a test of $\beta = \beta^0$ rejects with certainty when β^0 is non-zero, even if the hypothesis is true. The shrunken standard errors increase the risk that the econometrician rejects such true hypotheses.

Example 3:

Y Mismeasured, X Not Mismeasured, $X_t \notin Z^y_{x,t}$: Biased Point Estimates

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

Assumption 6 *X_t is not mismeasured: $X_t = X^\star_t$, and $X_t \notin Z^y_{x,t}$.*

This is the case studied in section 4 of the main paper.

Example 4:

Both X and Y Mismeasured: Illuminating Special Cases

The ζ^{xy} component of the LoSE in Y introduces an attenuation-type bias (i.e. towards zero in the univariate case) and inconsistency into the estimates under some circumstances. In particular, when $Z^x \not\subset Z^y_x$, so measured X contains information about $X^\star$ missed by the information set used to compute measured Y, then the LoSE in Y introduces bias and inconsistency. Put another way, bias and inconsistency occur when the explanatory variables X contain signal missing from the dependent variable Y.

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 $Z^y_x = Z^x$. Then

$E(X^\star|Z_x^y) = E(X^\star|Z^x)$, so their difference in (3) and (4) 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 example 2.

The second illuminating case is where $Z_x^y \subset Z^x$, so Z^x contains all the information about $X^\star$ in Z_x^y, plus additional information. The difference $E(X^\star|Z^x) - E(X^\star|Z_x^y)$ is uncorrelated with Z_x^y; substituting this difference for ζ^{xy} in example 3 then leaves the results of that section unchanged. The estimate $\widehat{\beta}$ is biased and inconsistent, with the bias towards zero in the univariate case; some variation in measured X that appears in $Y^\star$ is missed by measured Y, biasing down the covariance between X and Y. Valid instruments must be drawn from the information set used to compute the more-poorly measured Y.

The last illuminating case is where Z_x^y contains all the information about $X^\star$ in Z^x plus additional information, so $Z_x^y \supset Z^x$. Then $E(X^\star|Z_x^y) - E(X^\star|Z^x)$ is uncorrelated with Z^x and X, and if this difference replaces ζ^x in example 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 LoSE 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 example 1 (augmented with LoSE in $U^\star$), example 2, or example 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.