FORECASTING DYNAMIC TIME SERIES IN THE PRESENCE OF DETERMINISTIC COMPONENTS

Serena Ng * Timothy J. Vogelsang†

July 1999

Abstract

This paper studies the error in forecasting a dynamic time series with a deterministic component. We show that when the data are strongly serially correlated, forecasts based on a model which detrends the data before estimating the dynamic parameters are much less precise than those based on an autoregression which includes the deterministic components. The local asymptotic distribution of the forecast errors under the two-step procedure exhibits bimodality, and the forecasts are conditionally median biased in a direction that depends on the order of the deterministic trend function. We explore the conditions under which feasible GLS detrending can lead to forecast error reduction. The finite sample properties of OLS and feasible GLS forecasts are compared with forecasts based on unit root pretesting. The procedures are applied to fifteen macroeconomic time series to obtain real time forecasts. Forecasts based on feasible GLS detrending tend to be more efficient than forecasts based on OLS detrending. Regardless of the detrending method, unit root pretests often improve forecasts.

Keywords: forecasting, trends, unit root, GLS detrending.

JEL Classification: C2,C3,C5

*(Corresponding Author) Department of Economics, Boston College, Chestnut Hill, MA 02467. Tel. (617) 552-2182 Fax (617) 552-2308. Email: Serena.Ng@bc.edu

†Department of Economics, Cornell University Uris Hall, Cornell University, Ithaca, N.Y. 14583-7601 Tel. (607) 255-5108 Fax. (607) 255-2818. Email: tjv2@cornell.edu

1 Introduction

An important use of economic modeling is generating forecasts. If one is interested in forecasting a single time series variable, the starting point is often statistical models of which ARMA models have become most widely used. In many cases ARMA models perform surprisingly well. Alternatively, one could base forecasts on structural models that incorporate economic theory. The usefulness of structural models is often measured by forecast precision compared to forecasts using parsimonious ARMA models. Given the many uses of forecasts from ARMA models, it seems sensible to construct these forecasts using the best methodology possible. We show in this paper that the way in which deterministic components (mean, trend) are treated matters in important ways for ARMA forecasts when the data are strongly serially correlated. In particular, we show that using GLS detrending when estimating AR models can improve forecasts compared to OLS when the errors are highly persistent.

Suppose we are interested in forecasting the h step ahead value of a covariance stationary process with a Wold moving-average representation $y_t = m_t + \psi(L)e_t$, where

$$m_t = \delta_0 + \delta_1 t \ldots + \delta_p t^p = \delta' z_t$$

denotes the deterministic component. Assuming a quadratic loss function, the minimum mean squared error of the h-step ahead forecast conditional upon lags of y_t is given by the Kolmogorov-Wiener (KW) prediction formula:

$$y_{t+h|t} = m_{t+h} + \left[\frac{\psi(L)}{L^h}\right]_+ \frac{1}{\psi(L)}(y_t - m_t),$$

where $[\psi(L)/L^h]_+ = \psi_h + \psi_{h+1}L + \psi_{h+2}L^2 \ldots$. If we specialize to data generated with $\psi(L) = (1-\alpha L)^{-1}$, we have:

$$y_t = m_t + u_t, \tag{1}$$

$$u_t = \alpha u_{t-1} + e_t. \tag{2}$$

By the KW prediction formula, the optimal forecast for this AR(1) model is

$$E(y_{t+h}|y_t, y_{t-1}, \ldots) = y_{t+h|t} = m_{t+h} + \alpha^h(y_t - m_t). \tag{3}$$

But (3) is not the only way to forecast the AR(1) model. It is well known that, given information available at time t, as summarized by some vector x_t, the linear forecast conditional on x_t with the smallest mean squared error (MSE) is provided by the linear projection of y_{t+1} on x_t. That is,

$y_{t+1|t} = \theta' x_t$, where $\theta' = E(y_{t+1}x_t')[E(x_t x_t')]^{-1}$. Therefore, if we write the DGP as:

$$\begin{aligned} y_t &= \sum_{i=0}^{p} \beta_i t^i + \alpha y_{t-1} + e_t \\ &= \mathcal{B}_t + \alpha y_{t-1} + e_t, \end{aligned} \tag{4}$$

and let $x_t = (z_t, y_{t-1})$, the optimal one-period ahead forecast is $y_{t+1|t} = \mathcal{B}_t + \alpha y_t$. By the chain rule of forecasting, $y_{t+h|t} = \mathcal{B}_{t+h} + \alpha\mathcal{B}_{t+h-1} + \ldots + \alpha^{h-1}\mathcal{B}_{t+1} + \alpha^h y_t$. Equation (4) will sometimes be referred to as Durbin's equation below.

If we know α and δ, the two approaches should give the same forecast since one model can be reparameterized as the other exactly. However, α and δ are population parameters which we do not observe[1]. The next best solution is to replace the required population parameters by their unbiased estimates. But because any estimator of α must involve a lagged dependent variable which is not fixed in repeated sampling, an unbiased estimator for α does not exist. Thus, the most that one can do is to replace α by its consistent estimate. To obtain such an estimate when m_t is unknown, we can detrend the data prior to estimating α, or we can estimate $\mathcal{B}_t$ and α at the same time. In this paper, we refer to these as the one-step and the two-steps procedures respectively.

A quick review of textbooks reveals that, although (1) and (2) are always used instead of (4) to present the theory of optimal prediction,[2] the practical recommendation is not unanimous. For example, Pindyck and Rubinfeld (1998, p. 565) and Johnston and Dinardo (1997, p.192, p.232) used (1) and (2) to motivate the theory, but their empirical examples are based upon (4) (see, e.g. Table 6.5). Examples considered in Diebold (1997), on the other hand, are based on an estimated trend function with a correction for serial correlation in the noise component (see, for example, p. 231). This is consistent with using (1) and (2) as the forecasting model.

This paper is motivated by the fact that while $y_{t+h|t}$ is unique, its feasible counterpart is not. Depending on how the parameters are estimated, the mean-squared forecast errors should be expected to be different. A huge body of work exists in the literature which concerns efficient estimation of the trend coefficients when the error term is serially correlated but strictly stationary. More recently, Canjels and Watson (1997) and Vogelsang (1998) considered inference on $\widehat{\delta}_1$ when u_t is highly persistent and possibly has a unit root. Inference on $\widehat{\alpha}$ has also drawn a great deal of attention, but that literature takes the deterministic components as nuisance parameters. Although studies by Stock (1995, 1996, 1997) and Diebold and Kilian (1999) are all motivated by the fact that forecasting time series with a large autoregressive root raises a host of special issues, these

[1] In principle, the optimal forecast is inoperational also because we do not have information on the infinite lags of y_t. For the AR(p) model which is of interest here, we only need information of the past p lags of y_t .

[2] See, for example, Hamilton (1995, p.81) and Box, Jenkins and Reinsel (1994, p. 157). The exception is Clements and Hendry (1994).

analysis have concentrated on the dynamic component of the forecast. But when the objective of the exercise is forecasting, the properties of $\widehat{\delta}$ and $\widehat{\alpha}$ can no longer be considered in isolation. Good forecasts require negatively correlated and precise estimates of both dynamic and deterministic parameters . As will become clear, the error in estimating the deterministic and the dynamic terms interact in ways such that the properties of the forecast error can be very different from those derived under the assumption that the deterministic terms are absent.

In this article, the choice of estimators is assessed from the point of view of forecasting. We focus on two issues. First, do one and two steps detrending by OLS differ in ways that practitioners should care? Second, does efficient estimation of the deterministic components improve forecasts? The answers to both of these questions are yes. Theoretical and empirical properties of the forecast errors under least squares and GLS detrending are presented in Sections 2 and 3. In Section 4, we take the procedures considered to the data. We begin in the next section with forecasts under least squares detrending.

2 Forecasts Under Least Squares Detrending

Throughout our analysis, we assume that the data are generated by (1) and (2). We only consider the two leading cases for m_t. That is, when $p = 0$, $m_t = \delta_0$ and $z_t = 1$. When $p = 1$, $m_t = \delta' \cdot z_t$, where $\delta' = [\delta_0 \ \delta_1]$ and $z_t = [1, t]$. Given $\{y_t\}_{t=1}^T$, we consider the one-step ahead forecast error given information at time T,

$$\begin{aligned} e_{T+1|T} &= y_{T+1} - \widehat{y}_{T+1|T} \\ &= y_{T+1} - y_{T+1|T} + y_{T+1|T} - \widehat{y}_{T+1|T} \\ &= e_{T+1} + \widehat{e}_{T+1|T}. \end{aligned}$$

The innovation e_{T+1} is unforecastable given information at time T and is beyond the control of the forecaster. A forecast is best in a mean-squared sense if $\widehat{y}_{T+1|T}$ is made as close to $y_{T+1|T}$ as possible in a mean squared sense. Throughout, we refer to $\widehat{e}_{T+1|T}$ as the forecast error. Several options are available. If one uses (4) as the forecasting model, one can obtain as a feasible forecast:

$$y_{T+h|T} = \widehat{\mathcal{B}}_{T+h} + \alpha\widehat{\mathcal{B}}_{T+h-1} + \ldots + \widehat{\alpha}^{h-1}\widehat{\mathcal{B}}_{T+1} + \widehat{\alpha}^h y_T, \tag{5}$$

where $\widehat{\mathcal{B}}_t$ and $\widehat{\alpha}$ can be obtained by maximizing the the exact log likelihood corresponding to (5) which, through the first observation, imposes the assumption that $|\alpha| < 1$. If a forecaster so chooses to apply the KW prediction formula, the feasible forecast is:

$$\widehat{y}_{T+h|T} = \widehat{m}_{T+h} + \widehat{\alpha}^h(y_t - \widehat{m}_t), \tag{6}$$

where estimates of $\widehat{\alpha}$ and $\widehat{\delta}$ can be obtained by maximizing the exact log-likelihood associated with (6) which again imposes the assumption that $|\alpha| < 1$.

In practice, forecasts are rarely based upon maximum likelihood estimates. Instead, one relies on least squares estimation of the required parameters. This is valid because any $\widehat{\alpha}$ and $\widehat{\delta}$ that are consistent for α and δ can be used to produce feasible forecasts. Taking this tradition as given, we first seek to compare forecasts based upon (5) and (6) with all parameters estimated by OLS. This leaves two strategies, labelled OLS_1 and OLS_2 hereafter:

1. OLS_1: Estimate (4) by OLS directly in one step.
2. OLS_2: Estimate δ from (1) by OLS to obtain $\widehat{u}_t = y_t - \widehat{m}_t$. Then estimate α from (2) by OLS but replace u_t by $\widehat{u}_t$.

Both procedures yield conditional least squares estimates. Evidently, the dependence of $\widehat{\beta}$ on $\widehat{\alpha}$ is ignored by OLS_1.

2.1 Finite Sample Properties

We first consider the finite sample properties of the forecast errors by monte-carlo experiments. Our main focus is on AR(1) processes. We consider the constant and the linear trend case, each for 10 values of α:

$$\begin{aligned} \delta &= 0 \text{ when p=0 and } \delta = [0\ 0]' \text{ when p=1;} \\ \alpha &= -.4, 0, .4, .8, .9, .95, .975, .99, 1.0, 1.01. \end{aligned}$$

The choice of the parameter set reflects the fact that many macro economic time series are highly and positively autocorrelated. The errors are $N(0,1)$ generated using the rndn() function in Gauss V3.27 with seed=99. For this section, we assume that $u_1 = 0$.

We use $T = 100$ in the estimations to obtain up to $h = 10$ steps ahead forecasts. We use 10,000 replications to obtain the forecast errors $\widehat{e}_{T+h|T} = y_{T+h|T} - \widehat{y}_{T+h|T}$, and then evaluate the root mean-squared error (RMSE) and the mean absolute error (MAE). The MAE and the RMSE provide qualitatively similar information and only the RMSE will be reported.

Table 1a reports results for $h = 1$. As benchmarks, we first consider two infeasible forecasts:- *i)* OLS_2^{α} which assumes α is known and *ii)* OLS_2^{δ} which assumes δ is known. From these, we see that when $p = 0$, the error in estimating α dominates the error in estimating δ_0. But when $p = 1$, the error in $\widehat{\delta}$ dominates unless $\alpha \geq .8$. The RMSE for OLS_2 is smaller than the sum of OLS_2^{α} and OLS_2^{δ}, suggesting a negative covariance between $\widehat{\alpha}$ and $\widehat{\delta}$. The RMSE at $h = 10$ confirms

that the error from estimating α vanishes when α is far away from the unit circle but increases (approximately) linearly with the forecast horizon when α is close to unity. However, the error in estimating δ does not vanish with the forecast horizon even when $\alpha = 0$ as the RMSE for OLS_2^α shows.

The RMSE for OLS_1 and OLS_2 when both parameters have to be estimated are quite similar when $\alpha < .8$, but the similarity ends as the error process becomes more persistent.[3] When $p = 0$, OLS_2 exhibits a sudden increase in RMSE and is sharply inferior to OLS_1 at $\alpha = 1$. When $p = 1$ the RMSE for OLS_1 is always smaller than OLS_2 when $\alpha \geq .8$. The difference is sometimes as large as 20% when α is close to unity. Results for $h = 10$ in Table 1b show a sharper contrast in the two sets of forecast errors. For a given procedure, the forecast errors are much larger when $p = 1$.

From the finite sample simulations, we see that from a RMSE standpoint, the method of detrending is by and large an irrelevant issue when α is small. But for empirically relevant cases when α is large, one step least squares detrending clearly dominates two steps least squares detrending in terms of RMSE. In the next subsection, we report an asymptotic analysis which provides some theoretical explanations for the simulation results. We examine the large sample properties of OLS_1 and OLS_2 when the data are persistent.

2.2 Asymptotic Properties of OLS Forecasts

Because the difference between OLS_1 and OLS_2 occurs for $0.8 \leq \alpha \leq 1$, we use a local-to-unity framework with non-centrality parameter c to characterize the DGP as:

$$\begin{aligned} y_t &= \delta_0 + \delta_1 t + u_t, \\ u_t &= \alpha_T u_{t-1} + e_t \\ \alpha_T &= 1 + \frac{c}{T}, \end{aligned} \tag{7}$$

where e_t is a martingale difference sequence with $2+d$ moments for some $d > 0$ and $E(e_t^2) = 1$. We assume that $u_0 = 0$, and without loss of generality, let $\delta_0 = \delta_1 = 0$. For a given sample size, u_t is locally stationary when $c < 0$ and locally explosive when $c > 0$, but becomes an integrated process as the sample size increases to infinity. In what follows, we let $\Rightarrow$ denote weak convergence. For $t = [Tr]$ with $0 \leq r \leq 1$, the functional central limit theorem implies that:

$$T^{-1/2} u_{[Tr]} = T^{-1/2} y_{[Tr]} \Rightarrow J_c(r),$$

where $dJ_c(r) = cJ_c(r) + dW(r)$ is a diffusion and $W(r)$ is a standard Brownian motion. The demeaned and detrended variants of $J_c(r)$ are then the limits of the residuals, $\widehat{u}_t$, obtained from

[3] Sampson (1991) showed that under least squares detrending, the deterministic terms have a higher order effect on the forecast errors when u_t is non-stationary.

the projection of u_t on 1 when $p = 0$, and on 1 and t when $p = 1$. These are, respectively,

$$p = 0: \qquad T^{-1/2}\widehat{u}_{[Tr]} \Rightarrow \bar{J}_c(r) = J_c(r) - \int_0^1 J_c(s)ds, \tag{8}$$

$$p = 1: \qquad T^{-1/2}\widehat{u}_{[Tr]} \Rightarrow \widetilde{J}_c(r) = J_c(r) - (4-6r)\int_0^1 J_c(s)ds - (12r-6)\int_0^1 sJ_c(s)ds. \tag{9}$$

If a detrending procedure can remove m_t without leaving asymptotic effects on the data, $\bar{J}_c(r)$ and $\widetilde{J}_c(r)$ would have been identically $J_c(r)$. From the above limits, we see that least squares detrending leaves non-vanishing effects on the detrended data.

The limiting distribution of $T(\widehat{\alpha} - \alpha)$ can be conveniently written using the functional:

$$\Phi(B_c, W) = \frac{\int_0^1 B_c(r)dW(r)}{\int_0^1 B_c(r)^2 dr}$$

for B_c that depends on the method of detrending. Under least squares detrending, $T(\widehat{\alpha} - \alpha) \Rightarrow \Phi(\bar{J}_c, W)$ when $p = 0$ and to $\Phi(\widetilde{J}_c, W)$ when $p = 1$.

We use two theorems to summarize the results.

Theorem 1 *One Step Least Squares Detrending: $OLS_1(p)$*

- *($p = 0$): Let the data be generated by (7). Let $\widehat{\beta}_0$ and $\widehat{\alpha}$ be estimated from $y_t = \beta_0 + \alpha y_{t-1} + e_t$ by OLS. Let $\bar{J}_c(r)$ be defined as in (8). For $h = 1$, we have*

$$\begin{aligned} \widehat{e}_{T+1|T} &= (\beta_0 - \widehat{\beta}_0) + (\alpha - \widehat{\alpha})y_T. \\ \sqrt{T}\widehat{e}_{T+1|T} &\Rightarrow -\bar{J}_c(1)\Phi(\bar{J}_c, W) - W(1) \\ &\equiv \mathcal{P}(0). \end{aligned}$$

- *($p = 1$): Let the data be generated by (7). Let $\widehat{\beta}_0$, $\widehat{\beta}_1$, and $\widehat{\alpha}$ be estimated from $y_t = \beta_0 + \beta_1 t + \alpha y_{t-1} + e_t$ by OLS. Let $\widetilde{J}_c(r)$ be defined as in (9). Let $\mathcal{B}_t = \beta_0 + \beta_1 t$. For $h = 1$ we have*

$$\begin{aligned} \widehat{e}_{T+1|T} &= \mathcal{B}_{T+1} - \widehat{\mathcal{B}}_{T+1} + (\alpha - \widehat{\alpha})y_T, \\ \sqrt{T}\widehat{e}_{T+1|T} &\Rightarrow -\widetilde{J}_c(1)\Phi(\widetilde{J}_c, W) + \int_0^1 (2-6r)dW(r) \\ &\equiv \mathcal{P}(1). \end{aligned}$$

Theorem 2 *Two Step Least Squares Detrending: $OLS_2(p)$*

- $(p=0)$*: Let the data be generated by (7). Let* $\widehat{\delta}_0 = \bar{y} = \widehat{m}_t \ \forall t$*, where* $\bar{y} = T^{-1}\sum_{t-1}^{T} y_t$ *is the least squares estimate of* δ_0*. Let* $\widehat{\alpha}$ *be the least squares estimate from a regression of* $\widehat{u}_t$ *on* $\widehat{u}_{t-1}$ *with* $\widehat{u}_t = y_t - \bar{y}$*. Let* $\bar{J}_c(r)$ *be defined as in (8). For* $h=1$*,*

$$\begin{aligned} \widehat{e}_{T+1|T} &= (\delta_0 - \widehat{\delta}_0)(1-\alpha) + (\alpha - \widehat{\alpha})(y_T - \widehat{m}_T). \\ \sqrt{T}\widehat{e}_{T+1|T} &\Rightarrow -\bar{J}_c(1)\Phi(\bar{J}_c, W) + c\int_0^1 J_c(r)dr \\ &\equiv \mathcal{Q}(0). \end{aligned}$$

- $(p=1)$*: Let the data be generated by (7). Define* $\widehat{m}_t = \widehat{\delta}_0 + \widehat{\delta}_1 t$*, and let* $\widehat{\delta}_0$ *and* $\widehat{\delta}_1$ *be obtained from least squares regression of* y_t *on 1 and* t*, with* $\widehat{u}_t$ *being the estimated residuals. Let* $\widehat{\alpha}$ *be the least squares estimate from a regression of* $\widehat{u}_t$ *on* $\widehat{u}_{t-1}$*. Let* $\widetilde{J}_c(r)$ *be defined as in (9). For* $h=1$*, we have*

$$\begin{aligned} \widehat{e}_{T+1|T} &= (m_{T+1} - \widehat{m}_{T+1})(1-\alpha L) + (\alpha - \widehat{\alpha})(y_T - \widehat{m}_T). \\ \sqrt{T}\widehat{e}_{T+1|T} &\Rightarrow -\widetilde{J}_c(1)\Phi(\widetilde{J}_c, W) + \left[\int_0^1 (6-12r)J_c(r)dr\right] - c\left[\widetilde{J}_c(1) - J_c(1)\right] \\ &\equiv \mathcal{Q}(1). \end{aligned}$$

The main implication of the two theorems is that forecasts based upon one and two step least squares detrending are not asymptotically equivalent. This is in contrast to the often exploited result that one and two steps detrending yields asymptotically equivalent estimates of α. This is because $\widehat{m}_t$ is not invariant to the method of detrending.

To further understand the asymptotic non-equivalence of OLS_1 and OLS_2, let us rewrite the one step ahead forecast error in terms of $\widehat{u}_t$, the residuals from the projection of y_t on $z_t = (1, t)$.[4] Consider also the artificial regression

$$y_t = A_0 + A_1 t + e_t$$

where e_t is white noise. Let $\widehat{A}$ be the least squares estimates of $A = (A_0 \ A_1)'$ so that $\widehat{A} - A = (z'z)^{-1}z'e$. Then the forecast error for OLS_1 can be written as:

$$\widehat{e}_{T+1|T} = (\alpha - \widehat{\alpha})\widehat{u}_T + (A_0 - \widehat{A}_0) + (A_1 - \widehat{A}_1)T + o_p(1). \tag{10}$$

For OLS_2, we have:

$$\widehat{e}_{T+1|T} = (\alpha - \widehat{\alpha})\widehat{u}_T + (1-c)(\delta_1 - \widehat{\delta}_1) - cT^{-1}(\delta_0 - \widehat{\delta}_0). \tag{11}$$

[4] When $p=0$, the expressions simplify with A_1 and $\delta_1 = 0$.

The first term, $(\alpha - \widehat{\alpha})\widehat{u}_T$, which we shall refer to as the dynamic factor, has a limiting distribution that corresponds to the common term in Theorems 1 and 2. It depends on p but not on the method of least squares detrending and is inconsequential as far as a comparison between the two methods of least squares detrending is concerned. The remaining two terms in the forecast error, which we refer to as the deterministic factor, depends not only on p but also on the method of detrending. For a given p, $\widehat{A} - A = (z'z)^{-1}z'e$ under OLS_1, but $\widehat{\delta} - \delta = (z'z)^{-1}z'u$ under OLS_2. Since e_t is a white noise process but u_t is serially correlated, $\widehat{A}$ has smaller variance than $\widehat{\delta}$. Reduced parameter uncertainty translates into smaller forecast errors.

The second reason why two steps least squares detrending is inferior is that δ_0 cannot be identified at $c = 0$ and hence not consistently estimable when $\alpha = 1$. By continuity, $\widehat{\delta}_0$ is imprecisely estimated in the vicinity of $c = 0$. In our local to unity framework, $T^{-1/2}(\widehat{\delta}_0 - \delta_0)$ is $O_p(1)$ and has a non-vanishing effect on OLS_2. In consequence, the forecast error has a larger variance under two steps least squares detrending. Equation (11) also highlights the fact that the deterministic factor does not depend on c under OLS_1 but does so under OLS_2. Theorem 2 confirms that in large samples, c remains in the distribution of OLS_2 both directly, and indirectly through the diffusion process. The forecast errors are thus unstable and sensitive to c around $c = 0$. This result nicely illustrates in the most simple of models the care with which deterministic parameters need to be estimated when constructing forecasts.

The difference between OLS_1 and OLS_2 extends to long horizon forecasts with horizon h satisfying $h/T \to \lambda \in (0, 1)$. It can be shown that under OLS_1,

$$OLS_1(p): \quad \frac{1}{\sqrt{T}}\widehat{e}_{T+h|T} \Rightarrow \lambda \mathcal{P}(p)$$

to a first approximation. Under OLS_2, we have, to a first approximation,

$$OLS_2(p): \quad \frac{1}{\sqrt{T}}\widehat{e}_{T+h|T} \Rightarrow \lambda \mathcal{Q}(p),$$

for $p = 0, 1$. Thus, the long-horizon forecast errors diverge at rate $\sqrt{T}$, the same as the rate reported in Stock (1997) for the case of no deterministic terms. Because the estimation of the deterministic terms does not affect the rate at which the forecast error diverges over long-horizons, the properties of the forecast errors can be understood by concentrating on $h = 1$.[5]

The asymptotic results given by Theorems 1 and 2 can be used to approximate the finite sample behavior of the conditional forecast errors. In particular the mean and variances of the asymptotic

[5]The long horizon forecast errors depend on α^h which takes a limit of $exp(c\lambda)$ as h and T approaches infinity. The stated results make use of the approximation that $exp(c\lambda) - exp(\widehat{c}\lambda) \approx (c - \widehat{c})\lambda$, from which it follows that to a first approximation, the normalized forecast errors increase linearly with the forecast horizon. As in Phillips (1995), $\widehat{\alpha}^h$ is an exponentiated random variable when α is local to unity instead of the usual result that $\widehat{\alpha}^h \to 1$ when h is fixed.

distributions can shed light on the behavior of the bias and MSE of the forecast errors as c varies. We approximate the limiting distributions by approximating $W(r)$ and $B_c(r)$ using using partial sums of $iidN(0,1)$ errors with 500 steps in 10,000 simulations.[6] Five values of c are considered: -15, -5, -2, 0, 1. Densities of the asymptotic distributions are given in Figures 1 and 2 for $p = 0$ and 1 respectively. The result that stands out is that OLS_2 is very sensitive to whether $c \geq 0$.[7] In particular, its dispersion increases and is bimodal when $c \geq 0$. While OLS_1 is also biomodal when $c > 0$ and $p = 0$, there is no apparent increase in the dispersion of the distribution. These asymptotic results are consistent with the finding in finite samples that the forecast errors are much larger for OLS_2 when $c \geq 0$.

The sharp bimodality of OLS_2 when $c \geq 0$ warrants further explanation. Consider $p = 0$. From Theorem 2, the limit of $\widehat{e}_{T+1|T}$ is the limit of $-(y_T - \widehat{m}_T)(\widehat{\alpha} - \alpha)$ when $c = 0$. It is well known that when α is near the unit circle, $T(\widehat{\alpha} - \alpha)$ is skewed to the left. Therefore if the deterministic terms were absent and $y_T > 0$, the forecast errors will be concentrated in the positive range.[8] Stock (1996) evaluated the forecasts conditional on $y_T > 0$ assuming m_t is known and confirms asymptotic median bias. In our setting, $\widehat{e}_{T+1|T}$ will be upward biased at $c = 0$ if $y_T - \widehat{m}_T > 0$. In Figure 3, we present the limiting error distribution conditional on $y_T - m_T > 0$. Although this does not ensure that $y_T - \widehat{m}_T > 0$, this appears to happen sufficiently frequently since the conditional forecast error is still biased upwards. The bimodality in Figure 1 arises because $y_T - \widehat{m}_T$ is unconditional and can take on positive or negative values. In effect, conditional median bias in the dynamic component of the forecast error is necessary for bimodality in the unconditional error distribution, but it is sufficient only for OLS_2 when $c = 0$. Under OLS_1, the deterministic component does not drop out at $c = 0$. The downward biasedness of $\widehat{\alpha}$ is no longer sufficient to determine the sign of the prediction error, even if $y_T - \widehat{m}_T > 0$. For the same reason, bimodality is not observed when $c < 0$ under both methods of least squares detrending.

From Figures 1 and 2, we see that the unconditional forecast errors are roughly median unbiased. However, the conditional forecast errors (conditional on $y_T - m_T > 0$) are upward biased when $p = 0$ and downward biased when $p = 1$.[9] Further examination reveals that while $\widehat{\alpha}$ is downward

[6] The asymptotic MSE are in close range with the finite sample simulations for $T = 500$ which we did not report. In particular, the RMSE for $p = 0$ at $\alpha = 1$ are .078 and .089 respectively. The RMSE based on the asymptotic approximations are .0762 and .0877 respectively. For $p = 1$, the finite sample RMSE are .109 and .143. The asymptotic approximations are .108 and .142 respectively.

[7] The finite sample distribution of the forecast errors exhibit the same properties.

[8] Using Edgeworth expansions and assuming α is bounded away from the unit circle, Phillips (1979) showed that the exact distribution of $\widehat{e}_{T+h|T}$ will be skewed to the left if $y_T > 0$ once the dependence of $\widehat{\alpha}$ on y_T is taken into account. His result is for strictly stationary process and applies in finite samples. Thus, the median bias conditional on $y_T > 0$ observed here does not arise for the same reason.

[9] In finite sample results not reported, the unconditional median biases are small but the conditional median biases are noticeable when $c = 0$, in accord with the asymptotic results.

biased and induces a force for under-prediction, $\widehat{m}_t$ is upward biased under both methods of least squares detrending and dominates the error when $p = 1$. The result underscores the fact forecasting with deterministic terms can yield errors with properties that can be quite different than when m_t is absent. The result also indicates that depending on the whether $y_T - m_T > 0$ in the data, median bias can be possible.

Figures 1 and 2 also reveal that irrespective of the method of least squares detrending, the forecast error distributions do not resemble that of $\widehat{\alpha}$. Estimation of the deterministic terms removes much of the asymmetry in $\widehat{\alpha}$ from the forecast error distribution. As well, except when c is negative and far from zero, the error distribution for OLS_2 does not resemble the normal distribution and will pose problems for the construction of prediction intervals.

3 GLS Detrending

The foregoing analysis shows that OLS_1 will dominate OLS_2 when forecasting persistent time series. But does there exist a two step procedure that can improve upon one step least squares detrending? Two options come to mind. One possibility is fix α to remove variability due to $\widehat{\alpha}$. For highly persistent data, imposing a unit root on y_t has been treated as a serious option. We will return to this option subsequently. The alternative route is to search for estimates of δ that have better properties than OLS. Indeed, it is well known that different ways of estimating the serial correlation coefficient could have rather different finite sample implications for the parameters of a regression model with stationary disturbances.[10] The problem is that in the local to unity framework, δ_0 is not consistently estimable by any method.

Consider an alternative decomposition of the forecast error for an arbitrary two-step procedure:

$$\begin{aligned}\sqrt{T}\widehat{e}_{T+1|T} = T(\alpha - \widehat{\alpha})T^{-1/2}u_T + T(\alpha - \widehat{\alpha})T^{-1/2}(\widehat{u}_T - u_T) \\ +\sqrt{T}(\delta_1 - \widehat{\delta}_1) - cT^{-1/2}(\widehat{u}_T - u_T). \qquad (12)\end{aligned}$$

The forecast error has three components: the error from estimating α (the first term), the error from least squares estimation of the deterministic components (the last two terms), and the covariance between the two errors (the second term). While a consistent estimate of δ_0 is not achievable, there may exist an estimator of δ_0 with the property that $T^{-1/2}(\widehat{u}_T - u_T) = o_p(1)$. This property, denoted criterion A, is desirable for forecasting because from (12) we see that $\sqrt{T}\widehat{e}_{T+1|T}$ will then not depend, asymptotically, on the error in estimating δ_0. If criterion A is not satisfied, then the

[10] Using the regression model $y_t = \delta' z_t + u_t$ where u_t is AR(1) with parameter α strictly bounded away from the unit circle and z_t does not include a constant, Rao and Griliches (1969) showed, via monte-carlo experiments, that GLS estimation of δ in conjunction with an initial estimate of α obtained from Durbin's equation [i.e. (4)] is desirable for the mean-squared-error of $\widehat{\delta}$ when $|\alpha| > .3$.

next best is to have an estimator for δ that satisfies criterion B: the estimator should *i)* have a small variance, *ii)* allow for efficient of α, and *iii)* have errors between $\widehat{\alpha}$ and $\widehat{\delta}$ that covary negatively. In the rest of this analysis, we consider the ability of GLS in satisfying these properties.

The usefulness of GLS in forecasting was first analyzed by Goldberger (1962) who considered optimal prediction based upon the model $y_t = \delta' z_t + u_t$ but $E(uu') = \Omega$ is non-spherical. Goldberger showed that the best linear unbiased prediction can be obtained by quasi-differencing the data to obtain GLS estimates of δ, and then exploit the fact that if u_t is serially correlated, the relation between the u_{T+h} and u_T can be used to improve the forecast. When u_t is an AR(1) process with known parameter α, the one-step ahead optimal prediction reduces to $y_{T+1} = \widetilde{\delta}' z_{T+1} + \alpha(y_T - \widetilde{\delta}' z_T)$. This amounts to application of the KW prediction formula using an efficient estimate of δ, assuming α is known.

Estimation by GLS requires quasi-differencing the data. We consider the Prais-Winston (PW) transformation which includes information from the first observation, and the Cochrane-Orcutt (CO) transformation which drops information of the first observation. Specifically, for a given α, the quasi-differenced data y_t^+ and z_t^+, are constructed as follows:

- *PW*: For $t = 2, \ldots T$, $y_t^+ = y_t - \alpha y_{t-1}, z_t^+ = z_t - \alpha z_{t-1}$, with $y_1^+ = y_1$ and $z_1^+ = z_1$,
- *CO*: For $t = 2, \ldots T$, $y_t^+ = y_t - \alpha y_{t-1}, z_t^+ = z_t - \alpha z_{t-1}$.

Then $\widetilde{\delta} = (z^{+\prime} z^+)^{-1}(z^{+\prime} y^+)$ is the GLS estimate of δ, and $\widetilde{u}_t = y_t - \widetilde{\delta}' z_t$ is the GLS detrended data. Our treatment of the first observation under PW is necessitated by the fact that the original PW sets $x_1^+ = \sqrt{1-\alpha} x_1$ which is invalid when $\alpha = 1$. The stated PW transformation is valid if $u_0 = 0$ and was used by Elliott, Rothenberg and Stock (1996) to analyze the power of unit root tests. Phillips and Lee (1996) used the same assumption to show efficiency gains for $\widehat{\delta}_1$ from quasi-differencing.[11]

With either way of quasi-differencing, Goldberger's procedure does not, however, produce a feasible forecast because α is unknown. The estimation of α affects the forecast error not just directly through the dynamic component of the forecast, but also because quasi-differencing is performed at $\widehat{\alpha}$ rather than α. Rao and Griliches (1969) showed, via monte-carlo experiments, that estimating $\widehat{\alpha}$ from Durbin's equation (4) is much more efficient than estimating it from an autoregression in $\widehat{u}_t$ or directly by non-linear least squares when α is positive. We take this result as the starting point. Using $\widehat{\alpha}$ estimated from (4) to quasi-difference the data, we now assess whether

[11] Canjels and Watson (1997) referred to this as conditional GLS. We label this as PW only because it retains information in the first observation, in the same spirit as the Prais-Winston transformation.

GLS detrended data satisfy the criterion A. In the local to unity framework above, this requires that $T^{-1/2}\widetilde{u}_t \Rightarrow J_c(r)$, the limiting distribution of $T^{-1/2}u_t$.

Lemma 1 *Suppose $\widetilde{u}_t$ is the GLS detrended data using the PW transformation at $\widehat{\alpha}$ estimated from Durbin's equation. Let $\widehat{W}(r) = \int_0^1 d\widehat{W}(r)$, and $\widehat{W}(r) = W(r) - (\widehat{c} - c)\int_0^1 J_c(s)ds$.*

1. *$p = 0$: a) If $u_1 = O_p(1)$, $T^{-1/2}\widetilde{u}_t \Rightarrow J_c(r)$. b) If $T^{-1/2}u_1 \Rightarrow J_c^-$, $T^{-1/2}\widetilde{u}_t \Rightarrow J_c(r) - J_c^-$;*
2. *$p = 1$: a) If $u_1 = O_p(1)$, $T^{-1/2}\widetilde{u}_t \Rightarrow J_c(r) - P_3(\widehat{c}, \widehat{W})$. b) If $T^{-1/2}u_1 \Rightarrow J_c^-$, $T^{-1/2}\widetilde{u}_t \Rightarrow J_c(r) - P_4(\widehat{c}, J_c^-, \widehat{W})$.*

The precise expressions for P_3 and P_4 are given in the Appendix. The important thing is that they depend on $\widehat{c}$, and possibly J_c^-. Thus, in general, PW detrending fails criterion A. As Canjels and Watson (1997) noted, the efficiency of $\widehat{\delta}_1$ under GLS detrending can be sensitive to the treatment of the first observation. When $u_1 = O_p(T^{1/2})$ with limit J_c^-, the detrended data are not invariant to this limit both when $p = 0$ and when $p = 1$. But if $u_1 = O_p(1)$, the detrended data does not depend on the initial condition as the Theorem indicates. When $p = 0$, this is sufficient for u_T to be treated as though it were known in large samples. When $p = 1$, this is not sufficient. Elliott et al. (1996) considered quasi-differencing the data at some fixed $\bar{\alpha} = 1 - c/T$ and showed that when there is a linear trend, $\bar{\alpha}$ will remain in the limiting distribution of the detrended data. Not surprisingly, we find the effects of $\widehat{\alpha}$ to persist in the detrended data when $p = 1$. Naturally, $\widehat{\alpha}$ adds variability to the forecast errors via quasi-differencing.

Now consider the case of CO GLS detrending. Although PW and CO yields asymptotically equivalent estimates of δ_0 in a covariance stationary framework, as the following lemma shows this is no longer the case in a nearly integrated setting.

Lemma 2 *Suppose $\widetilde{u}_t$ is the GLS detrended data using the CO transformation. Then for $p = 1, 2$,*

$$T^{-1/2}\widetilde{u}_t \Rightarrow J_c(r) - C_p(\widehat{c}, \widehat{W}),$$

where $C_p(\cdot)$ for $p = 0, 1$ are defined in the Appendix.

By construction, CO ignores the initial observation and therefore the limiting distributions of the detrended data are invariant to the initial condition assumption. This can turn out to be a practical advantage because the initial condition assumption is difficult to validate. However, under CO, $\widehat{\alpha}$ always plays a non-vanishing role in the detrended data. Not only is C_p non-degenerate, they contain terms that are proportional to $\widehat{c}^{-1}$ (see the Appendix). The variance of the detrended data can thus be large when $\widehat{c}$ is close to zero. Thus, the CO fails both criterion A and B.

Lemmas 1 and 2 show that detrending by GLS will, in general, have asymptotic effects on the forecasts, just like OLS. There is only one exception. When $p = 0$ and $u_1 = O_p(1)$, then from Lemma 1 we see that criterion A is satisfied and it follows that $\widehat{e}_{T+1|T} = (\delta_0 - \widetilde{\delta}_0)(1-\widehat{\alpha}) + (\alpha - \widehat{\alpha})(y_T - m_T)$ and

$$\sqrt{T}\widehat{e}_{T+1|T} \quad \Rightarrow \quad -J_c(1)\Phi(\bar{J}_c, W).$$

Evidently, the error distribution depends on least squares detrending only to the extent that $\widehat{\alpha}$ is based on one step least squares detrending. Therefore, if $\widetilde{\alpha}$ were obtained from GLS detrended data, the effects of detrending on the forecast error can be completely removed.

Lemma 3 *When $p = 0$, $u_1 = O_p(1)$ and α is re-estimated from an autoregression using GLS detrended data, $\widetilde{e}_{T+1|T} = (\delta_0 - \widetilde{\delta}_0)(1 - \widetilde{\alpha}) + (\alpha - \widetilde{\alpha})(y_T - m_T)$. Then under PW,*

$$\sqrt{T}\widehat{e}_{T+1|T} \quad \Rightarrow \quad -J_c(1)\Phi(J_c, W).$$

In this very special case, the forecast error has the same properties as if the deterministic terms were known. But note it entails efficient estimation of both δ and α. More generally, the fact that GLS detrending yields more efficient estimates of δ and can improve the power of unit root tests does not imply they will yield more efficient forecasts.

3.1 Finite Sample Properties of Feasible GLS Forecasts

From the above, we see that the efficiency of feasible GLS forecasts cannot be presumed. The desirability of feasible GLS forecasts thus depends on the data under investigation. In this section, we consider six GLS estimators. A QD_n forecast for $QD = PW$ or CO is constructed as follows,

1. Obtain an initial estimate of α by OLS_1.
2. Transform y_t and z_t by QD to obtain y_t^+ and z_t^+. Then $\widetilde{\delta} = (z^{+\prime}z^+)^{-1}(z^{+\prime}y^+)$. The detrended data is $\widetilde{u}_t = y_t - \widetilde{\delta}' z_t$.
3. If $n = 0$, stop. If $n = 1$, then re-estimate α from (2) with u_t replaced by $\widetilde{u}_t$. Denote this estimate by $\widetilde{\alpha}$. For $n > 1$, repeat (a) and (b) until the change in $\widetilde{\alpha}$ between iterations is small.

The objective of iterative estimation ($n > 0$) is to bring the estimates of α and δ closer to being jointly optimal.

The simulations are performed using three sets of assumptions on u_1:

- Assumption A: $u_1 = e_1 \sim N(0, 1)$.

- Assumption B: $u_1 \sim (0, \sigma_e^2/(1-\alpha^2))$ for $\alpha < 1$.
- Assumption C: $u_1 = \sum_{j=0}^{[\kappa T]} \alpha^j e_{1-j}$, $\kappa > 0$.

Under A, $u_1 = O_p(1)$ and does not depend on unknown parameters. Under B, u_1 depends on α. Elliott (1997) showed that unit root tests based on GLS detrending are farther away from the asymptotic power envelope under B than A. Under C, $u_1 = O_p(T^{1/2})$. Canjels and Watson (1997) found that the efficiency of $\widehat{\delta}_1$ under PW-GLS is reduced when $\kappa > 0$. Assumption A is a special case of C with $\kappa = 0$. In the local asymptotic framework, u_1 is $O_p(T^{1/2})$ under both Assumptions B and C.

In practice, application of feasible GLS to persistent data faces the additional problem that $\widehat{\alpha}$ could exceed unity, but quasi-differencing is valid only if $\widehat{\alpha} < 1$. This problem is circumvented in the simulations as follows. If an initial $\widehat{\alpha}$ exceeds one, it is reset to one prior to PW quasi-differencing. From a theoretical perspective, the distribution of the CO detrended data depends on $\widehat{c}^{-1}$ which does not exist when $\widehat{c} = 0$. Numerical problems were indeed encountered if we allow $\widehat{\alpha}$ to be unity. Therefore under CO, we set the upper bound of $\widehat{\alpha}$ to .995. Simulations are performed under the same experimental design described earlier, except that under Assumption B, only cases with $\alpha < 1$ are evaluated. Canjels and Watson (1997) found that for small values of κ, the PW performs well. Here, we report results for $\kappa = 1$, which is considerably large, to put the PW to a challenge.

The results are reported in columns 3 through 8 of Table 2 for $p = 0$ and Table 3 for $p = 1$. Because the CO does not use information in the first observation, it is invariant to the assumption u_1. Differences in the results under alternative assumptions merely reflect sampling variability. When $\alpha < .8$, the gain in GLS estimation over the two OLS procedures is small, irrespective of the assumption on u_1. This is perhaps to be expected since the asymptotic equivalence of OLS and GLS detrending follows from the classic result of Grenander and Rosenblatt (1957) when u_t is strictly stationary. However, as persistence in u_t increases, there are notable differences.

For $p = 0$, first notice that PW_0 displays a sharp increase in RMSE around $\alpha = 1$ just like OLS_2. This shows that GLS estimation of δ alone will not always reduce forecast errors. The best feasible GLS forecast depends on the initial condition. Under Assumption A, PW_∞ outperforms all OLS and GLS estimators at every value of α, sometimes by as much as 20%. Under Assumptions B and C, OLS_1 and CO are best when $.8 \leq \alpha \leq .9$, and PW_∞ has large errors. On the other had, PW_∞ performs very well when the data become more persistent. This situation is problematic because the assumption on u_1 cannot be validated in practice, and no single estimator does well in every case. However, PW_1 has errors similar to OLS_1 when the data are mildly persistent, outperforms PW_∞ when the data are moderately persistent, dominates OLS_1 and is second best

to PW_∞ when the data are extremely persistent. It is perhaps the best feasible GLS forecast when $p = 0$.

It is of some interest to understand why PW_0 performs so poorly, even worse than OLS_2, when α is at or near unity. Because the least squares error in $\widehat{\alpha}$ and $\widehat{\delta}_0$ covary negatively, and both $\widehat{\alpha}$ and $\widehat{\delta}$ enter the limit of OLS_2, the two errors partially offset. But there is no independent effect of $\widehat{\delta}_0$ on PW_0 except through $\widehat{\alpha}$. Thus, PW_0 cannot take advantage of the negative covariation between $\widehat{\alpha}$ and $\widehat{\delta}$. In effect, it fails criterion B. For a feasible GLS forecast to be effective, $\widehat{\delta}$ and $\widehat{\alpha}$ need to be jointly optimal.

Results for $p = 1$ are reported in Table 3. Because the contribution of $\widehat{\delta}$ to the forecast error is large (as can be seen from OLS_2^α in Table 1a), the reduction in forecast error due to efficient estimation of trends is also more substantial. The results in Table 3 show that irrespective of the assumption on the initial condition, the forecast errors are smallest with PW_∞. Even at the one period horizon, the error reduction is 30% over OLS_2. From a RMSE point of view, the choice among the feasible GLS forecasts is clear when $p = 1$.

We also consider two forecasts based on pretesting for a unit root. Setting $\widehat{\alpha} = 1$ will generate the best 1-step ahead forecast if there is indeed a unit root, and in such a case, even long horizon forecasts can be shown to be consistent. Of course, if the unit root is falsely imposed, the forecast will continue to be biased. But one can expect forecast error reduction if we impose a unit root for α close to but not identically one. Campbell and Perron (1991) presented some simulation evidence in this regard for $p = 0$, and Diebold and Kilian (1999) considered the case of $p = 1$.[12] Stock and Watson (1998) considered the usefulness of unit root pretests in empirical applications. However, they forecast using OLS_1 when the hypothesis is rejected, a procedure which we refer to as UP_2. In light of the efficiency of GLS over the OLS, we also consider using PW_1 under the alterative of stationarity. The PW_1 is favored over PW_∞ because it is somewhat more robust to assumptions on u_1. Specifically, we use the DFGLS (based on the PW transformation) with one lag to test for a unit root. If we cannot reject a unit root and $p = 0$, $\widehat{y}_{T+1|T} = y_T$. If $p = 1$, the mean of the first differenced series is estimated. Denoting this by $\overline{\Delta y}$, then $\widehat{y}_{T+1|T} = y_T + \overline{\Delta y}$. If a unit root is rejected and a PW_1 forecast is obtained, the procedure is labelled UP_1 below.

The UP forecast errors are given in the last two columns of Tables 2 and 3. Irrespective of the assumption on u_1, UP_1 has smaller RMSE than UP_2, reflecting the improved efficiency of PW_1 over OLS_1. For both UP procedures, the trade-offs involved are clear: large reduction in RMSE when the data are persistent versus small increase in error when the largest autoregressive root

[12] Diebold and Kilian (1999) found that pretesting is better than always setting $\widehat{\alpha} = 1$ and is often better than always using the OLS estimate of α.

is far from unity. If the unit root test always rejects correctly, the RMSE for $\alpha < 1$ would have coincided with PW_1 or OLS_1. This apparently is not the case and reflects the fact that power of the unit root test is less than one. The increase in RMSE from falsely imposing a unit root is larger when $p = 0$. Nonetheless, the reduction in forecast errors are substantial for values of α very close to or at unity. This arises not just because variability in $\widehat{\alpha}$ is suppressed, but also because first differencing bypasses the need to estimate δ_0, the key source of variability with any two step procedure. Note, however, that although the RMSE are robust to the assumption on u_1, the unconditional median biases (not reported) are quite substantial under Assumption A, but not under B and C.

An overview of the alternatives to OLS_1 is as follows. The two UP procedures clearly yield the minimum RMSE when α is at or ϵ away from one. The problem, of course, is that ϵ is unknown and varies with the data in question. Of the GLS forecasts, PW_∞ performs very well when $p = 1$, but the choice is less clear when $p = 0$ because the results are sensitive to u_1. Nonetheless, feasible GLS and UP forecasts rarely do worse than OLS and should be useful in practice.

4 Empirical Examples

In this section, we take the procedures to fifteen U.S. macroeconomic time series. These are GDP, investment, exports, imports, final sales, personal income, employee compensation, M2 growth rate, unemployment rate, 3 month, 1 year, and 10 year yield on treasury bills, FED funds rate, inflation in the GDP deflator and the CPI. Except for variables already in rates, the logarithm of the data are used. Inflation in the CPI is calculated as the change in the price index between the last month of two consecutive quarters. All data span the sample 1960:1-1998:4 and are taken from FRED. Throughout, we use $k = 4$ lags in the forecasting model. Stock and Watson (1998) found little to gain from using data dependent rules for selecting the lag length in forecasting exercises. Four lags are also used in the unit root tests. We assume a linear time trend for the seven National Account series. Although the unit root test is performed each time the sample is extended, we only keep track of unit root test results for the sample as a whole. Except for investment, the unit root hypothesis cannot be rejected in the full sample for the first seven series. For the remaining variables, we assume $p = 0$. The DFGLS rejects a unit root in M1 growth, unemployment rate and CPI inflation.

Since the proceeding analysis assumes $k = 1$, a discussion on quasi-differencing when $k > 1$ is in order. We continue to obtain $\widehat{\alpha}_i$, $i = 1, \ldots, k$ from Durbin's equation. We experimented with two possibilities. The first is to quasi-difference at $\widehat{\alpha} = \sum_{i=1}^{k} \widehat{\alpha}_i$. The alternative option is to let $x_t^+ = x_t - \sum_{i=1}^{k} \widehat{\alpha}_i x_{t-i}$ for $t = k+1, \ldots, T$. For the CO, we now loose k observations but no

further modification is required. For the PW, we additionally assume $x_i^+ = x_i - \sum_{j=1}^{i} \widehat{\alpha}_j x_{i-j}$ for $i = 1, \ldots, k$. The forecasts are then based on four lags of the quasi-transformed data. Based on our limited experimentation, both approaches give very similar forecast errors and we only report results based on the first procedure. That is, quasi-differencing using the sum of autoregressive parameters.

Our results are based on 100 real time, one period ahead forecasts. Specifically, the first forecast is based on estimation up to 1973:4. The sample is then extended by one period, the models re-estimated, and a new forecast is obtained. Because we do not know the data generating process for the observed data, the forecast errors reflect not only parameter uncertainty, but also potential model misspecification. Procedures sensitive to model misspecification may have larger errors than found in the simulations when the forecasting model is correctly specified.

Our results are summarized in terms of the average RMSE. This is reported in Table 4a. Many of the minimum RMSE is given by UP_1. The feasible GLS procedures rarely come out on top. Half of the worst RMSE is due to the two OLS procedures, with PW_0 and CO_0 taking two of the worst positions each. The impression from these results is that the macroeconomic data considered are sufficiently persistent that we do best by exploiting pretests for a unit root.

Because we only have 100 forecasts on which to average (whereas in the simulations we have 10,000), a procedure may do well very in all but a few periods, but occasional large errors may drag down the ranking. As a final check, we also consider a relative efficiency index, defined for the 10 forecasts label $j = 1$ to 10 as:

$$RE_{jt} = \frac{|\widehat{e}_{t+1,j|t}|}{max_j |\widehat{e}_{t+1,j|t}|}.$$

It measures the one period ahead MAE of model j given information in time t relative to the worst model. The model with a smallest index on average is the best. These are reported in Table 4b. By this measure, OLS accounts for seven of the fifteen worst efficiency, with CO_0 and PW_0 taking six of the remaining eight worst forecasts. While the unit root pretest still comes out on top in three of the sixteen cases, in half the cases, PW_1 and PW_∞ are the best. The average RMSE and the efficiency index differ in how big errors are being weighed. A complete analysis on the appropriate loss function in this situation is beyond the scope of the present analysis, but both measures suggest a role for feasible GLS and unit root pretest in forecasting dynamic time series. [13]

Table 4c reports median bias in the forecasts defined as $\Pr(\widehat{e}_{T+h|T} > 0)$. Median bias is generally larger when a time trend is present. In those seven cases, the forecasts are upward biased, consistent

[13]The real time forecasts could be more formally compared using the simple test proposed by Diebold and Mariano (1995). However, their test is applicable for pairwise comparison of forecasts. Because we are considering 10 forecasts, it is not clear nor obvious how to implement the Diebold and Mariano (1995) test.

with the finite sample simulation results conditional on $y_T - m_T > 0$. Of the remaining series in rate form, only the M2 growth rate exhibits noticeable median bias. Interestingly, forecasts that are more efficient also tend to have larger median bias. While the median bias observed is not large enough to be of serious concern, this bias should be taken into account in the construction of prediction intervals.

Several results come out in both the simulations and the empirical examples. First, feasible GLS can usually do better than OLS. Second, PW_0 is not to be recommended. As explained earlier, this is because $\widehat{\alpha}$ and $\widehat{\delta}$ are not jointly optimal under PW_0. Third, the iterative PW is preferred over iterative CO because the latter could be unstable at $\widehat{\alpha}$ close to unity. Interesting, because $\widehat{\alpha}$ is downward biased, iterative estimation generally leads to upward revision in the estimate, and aggravates the problem. In results not reported, we count the number of times a procedure gives the smallest and largest MAE respectively. The numbers confirm that many of the worst forecasts come from the two OLS procedures and PW_0. But CO_∞ also has its fair share of worst forecasts. Fourth, unit root pretesting is desirable when α is very close to unity. But results one and four together imply that when a unit root is rejected, a feasible GLS (such as PW_1) is a better alternative to OLS_1. The numerical results indeed favor UP_1 over UP_2.

5 Conclusion

In this paper, we show that the forecast errors based upon one step OLS detrending and two steps OLS detrending have rather different empirical and theoretical properties when the autoregressive root is large. This is in sharp contrast to the asymptotic invariance of $\widehat{\alpha}$ to the same two methods of detrending. We then show that efficient estimation of deterministic trend parameters by GLS may improve forecasts under some conditions. Finite sample simulations show that iterative GLS, especially under PW, usually yields smaller forecast errors than one step OLS detrending. In empirical applications to highly persistent data, unit root pretesting yields the lowest average RMSE. However, by a measure of relative efficiency, feasible GLS forecasts are found to be desirable. When forecasting persistent time series with deterministic components, use of PW_1 with or without unit root pretesting dominates least squares detrending.

Appendix

The following limits are used in the evaluation of the limiting distribution of the forecast errors. Its derivations are straightforward and hence omitted.

Lemma 4 • *When* $p = 0$,

$$T^{-1/2}(\widehat{\delta}_0 - \delta_0) \Rightarrow \int_0^1 J_c(r)dr.$$

• *When* $p = 1$,

$$\begin{aligned} T^{-1/2}(\widehat{\delta}_0 - \delta_0) &\Rightarrow \int_0^1 (4-6r)J_c(r)dr, \\ T^{1/2}(\widehat{\delta}_1 - \delta_1) &\Rightarrow \int_0^1 (12r-6)J_c(r)dr. \end{aligned}$$

Proof of Theorems 1 and 2:

Under OLS_2,

$$\begin{aligned} \widehat{y}_{T+1} &= \widehat{m}_{T+1} + \widehat{\alpha}(y_T - \widehat{m}_T), \\ &= \widehat{\delta}_0 + \widehat{\delta}_1(T+1) + \widehat{\alpha}(y_T - \widehat{\delta}_0 - \widehat{\delta}_1 T). \end{aligned}$$

Then

$$\begin{aligned} \widehat{e}_{T+1|T} &= (m_{T+1} - \widehat{m}_{T+1}) + \alpha(y_T - m_T) - \widehat{\alpha}(y_T - \widehat{m}_T), \\ &= (m_{T+1} - \widehat{m}_{T+1}) + \alpha u_T - \widehat{\alpha}(m_T + u_T - \widehat{m}_T), \\ &= (1 - \widehat{\alpha}L)(m_T - \widehat{m}_T) + (\alpha - \widehat{\alpha})u_T. \end{aligned}$$

When $p = 0$, $m_t - \widehat{m}_t = \delta_0 - \widehat{\delta}_0$ for all t. Since $\alpha = 1 + c/T$, we have, for $p = 0$,

$$\begin{aligned} \widehat{e}_{T+1|T} &= (\alpha - \widehat{\alpha})(u_T + \delta_0 - \widehat{\delta}_0) - cT^{-1}(\delta_0 - \widehat{\delta}_0), \\ T^{1/2}\widehat{e}_{T+1|T} &= T(\alpha - \widehat{\alpha})(T^{-1/2}u_T + T^{-1/2}(\delta_0 - \widehat{\delta}_0)) - cT^{-1/2}(\delta_0 - \widehat{\delta}_0), \\ &= T(\alpha - \widehat{\alpha})T^{-1/2}\widehat{u}_T - cT^{-1/2}(\delta_0 - \widehat{\delta}_0), \\ &\Rightarrow -\phi(\bar{J}_c, W)\bar{J}_c(1) + c\int_0^1 J_c(r)dr. \end{aligned}$$

When $p = 1$, $\widehat{m}_t = \delta_0 + \delta_1 t$ and therefore $(1 - \widehat{\alpha}L)(m_{T+1} - \widehat{m}_{T+1}) = (1 - \widehat{\alpha})(m_T - \widehat{m}_T) + (\delta_1 - \widehat{\delta}_1)$. It follows that

$$\begin{aligned} \widehat{e}_{T+1|T} &= (1 - \widehat{\alpha})(m_T - \widehat{m}_T) + (\delta_1 - \widehat{\delta}_1) + (\alpha - \widehat{\alpha})u_T, \\ &= (\alpha - \widehat{\alpha})(m_T - \widehat{m}_T + u_T) + (\delta_1 - \widehat{\delta}_1) - cT^{-1}(\widehat{u}_T - u_T), \\ T^{1/2}\widehat{e}_{T+1|T} &= T(\alpha - \widehat{\alpha})T^{-1/2}\widehat{u}_T + T^{1/2}(\delta_1 - \widehat{\delta}_1) - cT^{-1/2}(\widehat{u}_T - u_T), \\ &\Rightarrow -\Phi(\widetilde{J}_c(r), W)\widetilde{J}_c(1) - \int_0^1 (12r-6)J_c(r)dr - c\left[\widetilde{J}_c(1) - J_c(1)\right]. \end{aligned}$$

Under OLS_1, let $\beta = (\beta_0\ \beta_1)'$. Then

$$\begin{aligned}
\widehat{y}_{T+1|T} &= \widehat{\beta}_0 + \widehat{\beta}_1(T+1)\widehat{\alpha}y_T, \\
\widehat{e}_{T+1|T} &= (\beta_0 - \widehat{\beta}_0) + (\beta_1 - \widehat{\beta}_1)(T+1) + (\alpha - \widehat{\alpha})(m_T + u_T), \\
&= -[1\ T+1](\beta - \widehat{\beta}) + (\alpha - \widehat{\alpha})(m_T + u_T).
\end{aligned}$$

We first show that the forecast error is invariant to the true values of δ_0 and δ_1. By partitioned regression, recall that $z_t = (1, t)$, and let $y_{-1} = \{y_0, y_1, \ldots, y_{T-1}\}$. Also let D be a $T \times 2$ matrix with 0 in the first column and 1 in the first column. Then

$$\begin{aligned}
\widehat{\beta} - \beta &= (z'z)^{-1}z'e - (z'z)^{-1}z'y_{-1}(\widehat{\alpha} - \alpha), \\
&= (z'z)^{-1}z'e - (z'z)^{-1}z'(z\delta - D\delta + u_{-1})(\widehat{\alpha} - \alpha), \\
&= (z'z)^{-1}z'e - (\delta_0 - \delta_1,\ \ \delta_1)'(\widehat{\alpha} - \alpha) - (z'z)^{-1}z'u_{-1}(\widehat{\alpha} - \alpha).
\end{aligned}$$

Substituting this result into the expression for $\widehat{e}_{T+1|T}$, we have

$$\begin{aligned}
\widehat{e}_{T+1|T} &= -[1\ \ T+1][(z'z)^{-1}z'e - (z'z)^{-1}z'u_{-1}(\widehat{\alpha} - \alpha)] \\
&\quad +(\delta_0 - \delta_1 + \delta_1 T + \delta_1)(\widehat{\alpha} - \alpha) + (\alpha - \widehat{\alpha})(m_T + u_T), \\
&= -[1\ \ T+1][(z'z)^{-1}z'e - (z'z)^{-1}z'u_{-1}(\widehat{\alpha} - \alpha)] + (\alpha - \widehat{\alpha})u_T,
\end{aligned}$$

which does not depend on δ. Therefore, without loss of generality, we let $\delta = 0$ so that $y_T = u_T$. Consider the artificial regression $y_t = A_0 + A_1 t + e_t$ where e_t is white noise. Then $\widehat{\beta} - \beta$ and $\widehat{e}_{T+1|T}$ simplify to

$$\begin{aligned}
\widehat{\beta} - \beta &= (z'z)^{-1}z'e - (z'z)^{-1}z'u_{-1}(\widehat{\alpha} - \alpha), \\
&\equiv \begin{bmatrix} \widehat{A}_0 - A_0 \\ \widehat{A}_1 - A_1 \end{bmatrix} - \begin{bmatrix} \widehat{\delta}_0 - \delta_0 \\ \widehat{\delta}_1 - \delta_1 \end{bmatrix}(\widehat{\alpha} - \alpha), \\
\widehat{e}_{T+1|T} &= (\beta_0 - \widehat{\beta}_0) + (\beta_1 - \widehat{\beta}_1)(T+1) + (\alpha - \widehat{\alpha})u_T.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\widehat{e}_{T+1|T} &= (\delta_0 - \widehat{\delta}_0) - (\widehat{A}_0 - A_0) + (\delta_1 - \widehat{\delta}_1)(\alpha - \widehat{\alpha})(T+1) - (\widehat{A}_1 - A_1)(T+1) + (\alpha - \widehat{\alpha})u_T, \\
&= (\alpha - \widehat{\alpha})(m_T - \widehat{m}_T + u_T) + (\delta_1 - \widehat{\delta}_1)(\alpha - \widehat{\alpha}) - (\widehat{A}_0 - A_0) - (\widehat{A}_1 - A_1)(T+1), \\
&= (\alpha - \widehat{\alpha})\widehat{u}_T + (\delta_1 - \widehat{\delta}_1)(\alpha - \widehat{\alpha}) - (\widehat{A}_0 - A_0) - (\widehat{A}_1 - A_1)(T+1), \\
T^{1/2}\widehat{e}_{T+1|T} &= T(\alpha - \widehat{\alpha})T^{-1/2}\widehat{u}_T - T^{1/2}(\widehat{A}_0 - A_0) - T^{3/2}(\widehat{A}_1 - A_1) + o_p(1).
\end{aligned}$$

Since

$$\begin{bmatrix} T^{1/2} & 0 \\ 0 & T^{3/2} \end{bmatrix}\begin{bmatrix} \widehat{A}_0 - A_0 \\ \widehat{A}_1 - A_1 \end{bmatrix} \Rightarrow \begin{bmatrix} 4 & 6 \\ -6 & 12 \end{bmatrix}\begin{bmatrix} \int_0^1 dW(r) \\ \int_0^1 r dW(r) \end{bmatrix},$$

$$T^{1/2}e_{T+1|T} \Rightarrow -\Phi(\widetilde{J}_c, W)\widetilde{J}_c(1) - \int_0^1 (6r-2)dW(r).$$

For $p=0$, the last term A_1 does not exist and $A_0 = T^{-1}\sum_{t=1}^T e_t$. Thus,

$$\begin{aligned} T^{1/2}\widehat{e}_{T+1|T} &= T(\alpha-\widehat{\alpha})\widehat{u}_T - T^{-1/2}\sum_{t=1}^T e_t \\ &\Rightarrow -\Phi(\bar{J}_c, W)\bar{J}_c(1) - W(1). \end{aligned}$$

For long horizon forecasts, consider OLS_2 when $p=0$.

$$\begin{aligned} \widehat{e}_{T+h|T} &= (\delta_0 - \widehat{\delta}_0)(1-\widehat{\alpha}^h) + (\alpha^h - \widehat{\alpha}^h)(y_T - \delta_0) \\ &= (\delta_0 - \widehat{\delta}_0)(1-\alpha^h) + (\alpha^h - \widehat{\alpha}^h)(y_T - \widehat{\delta}_0) \\ T^{-1/2}\widehat{e}_{T+1|T} &= T^{-1/2}(\delta_0 - \widehat{\delta}_0)(1-\alpha^h) + (\alpha^h - \widehat{\alpha}^h)(y_T - \widehat{\delta}_0) \\ &\Rightarrow c\lambda\int_0^1 J_c(r)dr - \lambda\bar{J}_c(1)\Phi(\bar{J}_c, W) \equiv \lambda\mathcal{Q}(0), \end{aligned}$$

since $\alpha^h \to exp(c\lambda) \approx 1 + c\lambda$. For OLS_1,

$$\widehat{e}_{T+1|T} = (\beta_0 - \widehat{\beta}_0)\left[\frac{1-\widehat{\alpha}^h}{1-\widehat{\alpha}}\right] + (\alpha - \widehat{\alpha}^h)y_T,$$

under the assumption that $\delta_0 = 0$. Since $1-\widehat{\alpha}^h \approx -\lambda\widehat{c}$, and $1-\widehat{\alpha} \approx -\widehat{c}T^{-1}$, $(1-\widehat{\alpha}^h)(1-\widehat{\alpha})^{-1} \approx \lambda T$. Thus,

$$T^{-1/2}\widehat{e}_{T+1|T} \approx \sqrt{T}(\beta_0 - \widehat{\beta}_0)\lambda - (c-\widehat{c})\lambda T^{-1/2}y_T \Rightarrow \lambda\mathcal{P}(0).$$

Results for $p=1$ can be derived analogously.

GLS Detrending

Under PW and the assumption that $u_1 = O_p(1)$,

1. When $p=0$, $T^{1/2}(\widetilde{\delta}_0 - \delta_0 - u_1) = O_p(1)$,

$$T^{-1/2}\widetilde{u}_t = T^{-1/2}u_t - T^{-1/2}(\widetilde{\delta}_0 - \delta_0) \Rightarrow J_c(r).$$

2. When $p=1$, $(\widetilde{\delta}_0 - \delta_0) \Rightarrow J_c^-$. Let $\widehat{\theta} = (1 + \widehat{c}^2/3 - \widehat{c})^{-1}$. Then

$$\begin{aligned} T^{1/2}(\widetilde{\delta}_1 - \delta_1) &\Rightarrow \widehat{\theta}\int_0^1 (1-\widehat{c}s)d\widehat{W}(s), \\ T^{-1/2}\widetilde{u}_t &= T^{-1/2}u_t - T^{-1/2}(\widetilde{\delta}_0 - \delta_0) - \sqrt{T}(\widetilde{\delta}_1 - \delta_1)t/T, \\ &\Rightarrow J_c(r) - r\widehat{\theta}\int_0^1 (1-\widehat{c}s)d\widehat{W}(s), \\ &= J_c(r) - P_3(\widehat{c}, \widehat{W}). \end{aligned}$$

Under PW and the assumption that $T^{-1/2}u_1 = O_p(1)$,

1. When $p = 0$, $T^{-1/2}(\widetilde{\delta}_0 - \delta_0) \Rightarrow J_c^-$. Therefore,

$$T^{-1/2}\widetilde{u}_t = T^{-1/2}u_t + T^{-1/2}(\widetilde{\delta}_0 - \delta_0) \Rightarrow J_c(r) - J^- c.$$

2. When $p = 1$, $T^{-1/2}(\widetilde{\delta}_0 - \delta_0) \Rightarrow J_c^-$,

$$\begin{aligned} T^{1/2}(\widetilde{\delta}_1 - \delta_1) &\Rightarrow .5(\widehat{c} - \widehat{c}^2)J_c^- + \widehat{\theta}\int_0^1 (1 - \widehat{c}s)d\widehat{W}(s), \\ T^{-1/2}\widetilde{u}_t &\Rightarrow J_c(r) - J_c^- - .5r\widehat{\theta}(\widehat{c} - \widehat{c}^2)J^- - r\widehat{\theta}\int_0^1 (1 - \widehat{c}s)d\widehat{W}(s), \\ &\equiv P_4(\widehat{c}, J_c^-, \widehat{W}). \end{aligned}$$

Under CO,

1. When $p = 0$, $\widetilde{\delta}_0 - \delta_0 \Rightarrow -\widehat{c}^{-1}\int_0^1 d\widehat{W}(s)$

$$T^{-1/2}\widehat{u}_t \Rightarrow J_c(r) + \widehat{c}^{-1}\int_0^1 d\widehat{W}(s) \equiv J_c(r) - C_0(\widehat{c}).$$

2. When $p = 1$, $T^{-1/2}(\widetilde{\delta}_0 - \widehat{\delta}_0) \Rightarrow \widehat{c}^{-2}\int_0^1 (6 - 4\widehat{c}) - (12 - 6\widehat{c}^2)sd\widehat{W}(s)$,

$$\begin{aligned} T^{1/2}(\widetilde{\delta}_1 - \delta_1) &\Rightarrow \widehat{c}^{-1}\int_0^1 (6 - 12s)d\widehat{W}(s), \\ T^{-1/2}\widetilde{u}_t &\Rightarrow J_c(r) - \widehat{c}^{-2}\left[\int_0^1 (6 + 2\widehat{c})d\widehat{W}(s) + (12 + 12\widehat{c} - 6\widehat{c}^2)sd\widehat{W}(s)\right] \\ &\equiv J_c(r) - C_1(\widehat{c}, \widehat{W}). \end{aligned}$$

References

Box, G. E. P., Jenkins, G. M. and Reinsel, G. C. (1994), *Time Series Analysis: Forecasting and Control*, Prentice-Hall, Englewood, N.J.

Campbell, J. Y. and Perron, P. (1991), Pitfalls and Opportunities: What Macroeconomists Should Know About Unit Roots, *NBER Macroeconomic Annual*, Vol. 6, M.I.T. Press, pp. 141–201.

Canjels, E. and Watson, M. W. (1997), Estimating Deterministic Trends in the Presence of Serially Correlated Errors, *Review of Economics and Statistics* **May**, 184–200.

Clements, M. P. and Hendry, D. (1994), Towards a Theory of Economic Forecasting, *in* C. P. Hargreaves (ed.), *Nonstationary Time Series Analysis and Cointegration*, Oxford University Press.

Diebold, F. and Mariano, R. S. (1995), Comparing Predictive Accuracy, *Journal of Business and Economic Statistics* **13:3**, 253–264.

Diebold, F. X. (1997), *Elements of Forecasting*, South Western Publishing, Cincinatti, Ohio.

Diebold, F. X. and Kilian, L. (1999), Unit Root Tests are Useful for Selecting Forecasting Models, mimeo, University of Pennsylvannia.

Elliott, G. (1997), Efficient Tests for a Unit Root when the Initial Observation is Drawn from its Unconditional Distribution, *International Economic Review, forthcoming*.

Elliott, G., Rothenberg, T. J. and Stock, J. H. (1996), Efficient Tests for an Autoregressive Unit Root, *Econometrica* **64**, 813–836.

Goldberger, A. (1962), Best Linear Unbiased Prediction in the Generalized Linear Regression Model, *Journal of the American Statistical Association* **57**, 369–375.

Grenander, U. and Rosenblatt, M. (1957), *Statistical Analysis of Stationary Time Series*, Wiley, New York.

Hamilton, J. D. (1994), *Time Series Analysis*, Princeton University Press, Princeton, N.J.

Phillips, P. and Lee, C. (1996), Efficiency Gains from Quasi-differencing Under Non-stationarity, Cowles Foundation Working Paper.

Phillips, P. C. B. (1995), Impulse Response and Forecast Error Variance Asymptotics in Nonstationary VAR's, manuscript, Cowles Foundation for Research in Economics, Yale University.

Pindyck, R. S. and Rubinfeld, D. (1998), *Econometric Models and Economic Forecasts*, Mcgraw Hill.

Rao, P. and Griliches, Z. (1969), Small Sample Properties of Several Two Stage Regression Methods in the Context of Autocorrelated Errors, *Journal of the American Statistical Association* **64**, 251–72.

Sampson, M. (1991), The Effect of Parameter Uncertainty on Forecast Variances and Confidence Intervals for Unit Root and Trend Stationary Time-Series Models, *Journal of Applied Econometrics* **6**, 67–76.

Stock, J. H. (1995), Point Forecasts and Prediction Intervals for Long Horizon Forecasts, mimeo, Kennedy School of Government, Harvard University.

Stock, J. H. (1996), VAR, Error-Correction and Pretest Forecasts at Long Horizon, *Oxford Bulletin of Economics and Statistics* **58**, 685–701.

Stock, J. H. (1997), Cointegration, Long Run Comovementes and Long Horizon Forecasting, *Advances in Econometrics: Proceedings of the Seventh World Congress of the Econometric Society*, Cambridge University Press.

Stock, J. H. and Watson, M. W. (1998), A Comparison of Linear and Nonlinear Univariate Models for Forecasting Macroeconomic Time Series, mimeo, Harvard University.

Vogelsang, T. J. (1998), Trend Function Hypothesis Testing in the Presence of Serial Correlation Correlation Parameters, *Econometica* **65**, 123–148.

Table 1a: RMSE of OLS Forecast Errors: $h = 1$

α	OLS_1	OLS_2	OLS_2^α	OLS_2^δ	OLS_1	OLS_2	OLS_2^α	OLS_2^δ
	$p = 0$				$p = 1$			
-0.400	0.142	0.141	0.100	0.100	0.228	0.225	0.205	0.100
0.000	0.143	0.141	0.099	0.101	0.228	0.225	0.203	0.101
0.400	0.144	0.143	0.099	0.101	0.230	0.228	0.202	0.101
0.800	0.153	0.152	0.096	0.104	0.242	0.249	0.200	0.104
0.900	0.163	0.162	0.092	0.108	0.253	0.270	0.196	0.108
0.950	0.175	0.173	0.084	0.114	0.263	0.292	0.186	0.114
0.975	0.183	0.179	0.068	0.122	0.264	0.310	0.169	0.122
0.990	0.180	0.180	0.041	0.131	0.257	0.319	0.146	0.131
1.000	0.174	0.196	0.000	0.142	0.244	0.314	0.109	0.142
1.010	0.191	0.292	0.087	0.161	0.220	0.297	0.036	0.161

Table 1b: RMSE of OLS Forecast Errors: $h = 10$

α	OLS_1	OLS_2	OLS_2^α	OLS_2^δ	OLS_1	OLS_2	OLS_2^α	OLS_2^δ
	$p = 0$				$p = 1$			
-0.400	0.071	0.071	0.071	0.001	0.167	0.167	0.167	0.001
0.000	0.100	0.099	0.099	0.000	0.233	0.231	0.231	0.000
0.400	0.166	0.165	0.165	0.001	0.384	0.379	0.379	0.001
0.800	0.471	0.450	0.429	0.159	1.008	0.984	0.972	0.159
0.900	0.767	0.720	0.601	0.410	1.487	1.466	1.365	0.410
0.950	1.054	0.976	0.675	0.669	1.843	1.874	1.569	0.669
0.975	1.244	1.134	0.613	0.897	1.994	2.141	1.572	0.897
0.990	1.305	1.204	0.394	1.123	1.986	2.284	1.427	1.123
1.000	1.315	1.436	0.000	1.345	1.830	2.217	1.090	1.345
1.010	1.715	2.684	0.911	1.681	1.507	1.983	0.364	1.681

Notes:OLS_2^α refers to OLS_2 with α assumed known. OLS_2^δ refers to OLS_2 with δ assumed known.

Table 2a: RMSE of GLS and UP Forecast Errors: $p = 0, h = 1$

u_1 follows Assumption A.

α	OLS_1	OLS_2	CO_0	PW_0	CO_1	PW_1	CO_∞	PW_∞	UP_1	UP_2
0.000	0.143	0.141	0.143	0.142	0.143	0.142	0.143	0.142	0.148	0.148
0.400	0.144	0.143	0.144	0.143	0.144	0.143	0.144	0.143	0.146	0.147
0.800	0.153	0.152	0.153	0.146	0.153	0.146	0.153	0.147	0.151	0.157
0.900	0.163	0.162	0.163	0.145	0.163	0.144	0.163	0.146	0.175	0.187
0.950	0.175	0.173	0.174	0.145	0.175	0.141	0.175	0.139	0.174	0.182
0.975	0.183	0.179	0.179	0.159	0.180	0.143	0.180	0.140	0.142	0.146
0.990	0.180	0.180	0.173	0.205	0.174	0.152	0.175	0.145	0.102	0.105
1.000	0.174	0.196	0.168	0.287	0.163	0.165	0.164	0.153	0.068	0.070

Table 2b: RMSE of GLS and UP Forecast Errors: $p = 0, h = 1$

u_1 follows Assumption B.

α	OLS_1	OLS_2	CO_0	PW_0	CO_1	PW_1	CO_∞	PW_∞	UP_1	UP_2
0.000	0.147	0.146	0.147	0.146	0.147	0.146	0.147	0.146	0.143	0.143
0.400	0.148	0.147	0.148	0.147	0.148	0.147	0.148	0.147	0.144	0.146
0.800	0.154	0.154	0.154	0.155	0.154	0.155	0.154	0.165	0.142	0.155
0.900	0.162	0.163	0.162	0.173	0.162	0.164	0.162	0.177	0.164	0.182
0.950	0.172	0.174	0.172	0.198	0.172	0.165	0.172	0.165	0.172	0.183
0.975	0.180	0.184	0.176	0.225	0.178	0.167	0.178	0.160	0.141	0.146
0.990	0.181	0.190	0.173	0.257	0.175	0.168	0.175	0.158	0.102	0.105

Table 2c: RMSE of GLS and UP Forecast Errors: $p = 0, h = 1$

u_1 follows Assumption C.

α	OLS_1	OLS_2	CO_0	PW_0	CO_1	PW_1	CO_∞	PW_∞	UP_1	UP_2
0.000	0.147	0.145	0.147	0.146	0.147	0.146	0.147	0.146	0.156	0.157
0.400	0.148	0.147	0.148	0.147	0.148	0.148	0.148	0.148	0.155	0.155
0.800	0.153	0.153	0.153	0.155	0.153	0.156	0.153	0.166	0.177	0.176
0.900	0.161	0.161	0.161	0.172	0.161	0.163	0.161	0.175	0.198	0.199
0.950	0.171	0.173	0.170	0.196	0.171	0.165	0.171	0.165	0.175	0.176
0.975	0.179	0.182	0.175	0.222	0.176	0.165	0.177	0.160	0.136	0.137
0.990	0.179	0.187	0.171	0.248	0.172	0.165	0.173	0.156	0.100	0.101
1.000	0.171	0.194	0.165	0.289	0.160	0.164	0.161	0.153	0.069	0.069

Notes: UP_1 is the forecast based on a unit root pretest where PW_1 is used if a unit root is rejected. UP_2 is the forecast based on a unit root pretest where OLS_1 is used if a unit root is rejected.

Table 3a: RMSE of GLS and UP Forecast Errors: $p = 1, h = 1$

u_1 follows Assumption A.

α	OLS_1	OLS_2	CO_0	PW_0	CO_1	PW_1	CO_∞	PW_∞	UP_1	UP_2
0.000	0.228	0.225	0.228	0.227	0.228	0.227	0.228	0.227	0.227	0.228
0.400	0.230	0.228	0.230	0.227	0.230	0.227	0.230	0.227	0.227	0.230
0.800	0.242	0.249	0.242	0.227	0.242	0.226	0.242	0.226	0.251	0.262
0.900	0.253	0.270	0.253	0.231	0.253	0.225	0.253	0.223	0.254	0.259
0.950	0.263	0.292	0.280	0.245	0.263	0.227	0.263	0.222	0.209	0.210
0.975	0.264	0.310	0.298	0.265	0.264	0.232	0.264	0.221	0.175	0.176
0.990	0.257	0.319	0.312	0.279	0.257	0.233	0.257	0.218	0.149	0.150
1.000	0.244	0.314	0.332	0.274	0.242	0.222	0.242	0.204	0.123	0.123

Table 3b: RMSE of GLS and UP Forecast Errors: $p = 1, h = 1$

u_1 follows Assumption B.

α	OLS_1	OLS_2	CO_0	PW_0	CO_1	PW_1	CO_∞	PW_∞	UP_1	UP_2
0.000	0.230	0.227	0.230	0.229	0.230	0.229	0.230	0.229	0.226	0.228
0.400	0.232	0.231	0.232	0.230	0.232	0.230	0.232	0.230	0.226	0.231
0.800	0.241	0.252	0.241	0.240	0.241	0.237	0.241	0.240	0.243	0.258
0.900	0.253	0.276	0.253	0.253	0.253	0.240	0.253	0.238	0.253	0.259
0.950	0.264	0.302	0.273	0.266	0.264	0.240	0.264	0.233	0.209	0.211
0.975	0.266	0.318	0.304	0.277	0.266	0.240	0.266	0.228	0.174	0.175
0.990	0.260	0.324	0.336	0.281	0.259	0.236	0.259	0.221	0.150	0.150

Table 3c: RMSE of GLS and UP Forecast Errors: $p = 1, h = 1$

u_1 follows Assumption C.

α	OLS_1	OLS_2	CO_0	PW_0	CO_1	PW_1	CO_∞	PW_∞	UP_1	UP_2
0.000	0.229	0.227	0.229	0.229	0.229	0.229	0.229	0.229	0.229	0.230
0.400	0.231	0.230	0.231	0.230	0.231	0.230	0.231	0.230	0.230	0.232
0.800	0.241	0.251	0.241	0.238	0.241	0.236	0.241	0.239	0.268	0.272
0.900	0.253	0.275	0.253	0.251	0.253	0.238	0.253	0.237	0.258	0.260
0.950	0.264	0.302	0.273	0.266	0.264	0.240	0.264	0.232	0.211	0.212
0.975	0.267	0.319	0.306	0.278	0.266	0.241	0.266	0.228	0.177	0.177
0.990	0.261	0.325	0.329	0.281	0.260	0.236	0.260	0.221	0.152	0.152
1.000	0.247	0.317	0.349	0.272	0.245	0.223	0.245	0.206	0.123	0.123

Notes: UP_1 is the forecast based on a unit root pretest where PW_1 is used if a unit root is rejected. UP_2 is the forecast based on a unit root pretest where OLS_1 is used if a unit root is rejected.

Table 4a: Empirical Examples: Average RMSE

	OLS_1	OLS_2	CO_0	PW_0	CO_1	PW_1	CO_∞	PW_∞	UP_1	UP_2	B	W
gdpc92	0.280	0.292	0.279	0.294	0.279	0.286	0.287	0.285	0.278	0.278	9	4
gpdic92	1.587	1.605	1.578	1.599	1.579	1.592	1.581	1.591	1.592	1.587	3	2
expgsc92	0.196	0.204	0.195	0.208	0.193	0.184	0.191	0.182	0.181	0.181	9	4
impgsc92	1.198	1.192	1.195	1.152	1.181	1.145	1.178	1.125	1.123	1.123	9	1
finslc92	1.011	1.033	0.999	1.008	0.995	0.974	0.995	0.957	0.870	0.870	9	2
dpic92	0.356	0.365	0.351	0.356	0.349	0.350	0.352	0.347	0.342	0.342	9	2
wascur	0.239	0.247	0.237	0.247	0.237	0.242	0.241	0.240	0.238	0.238	3	2
m2sl	3.002	2.993	2.992	2.991	2.991	2.989	2.991	2.989	2.989	3.002	8	1
unrate	0.353	0.353	0.353	0.354	0.353	0.354	0.358	0.355	0.354	0.353	2	7
tb3ma	1.591	1.605	1.491	1.601	1.573	1.583	1.577	1.569	1.549	1.549	3	2
gs1	1.474	1.476	1.366	1.475	1.447	1.469	1.458	1.468	1.452	1.452	3	2
gs10	0.864	0.857	0.911	0.863	0.854	0.863	0.855	0.864	0.856	0.856	5	3
fed	1.987	1.985	1.957	1.991	1.976	2.009	1.974	2.000	1.934	1.934	9	6
gdpdef	1.156	1.131	1.213	1.155	1.121	1.148	1.117	1.151	1.115	1.115	9	3
cpiaucs	2.221	2.206	2.206	2.207	2.211	2.217	2.199	2.218	2.217	2.221	7	1

Table 4b: Empirical Examples: Relative Efficiency

	OLS_1	OLS_2	CO_0	PW_0	CO_1	PW_1	CO_∞	PW_∞	UP_1	UP_2	B	W
gdpc92	0.651	0.683	0.658	0.740	0.660	0.669	0.731	0.663	0.655	0.655	1	4
gpdic92	0.845	0.859	0.844	0.859	0.845	0.858	0.846	0.860	0.858	0.845	3	8
expgsc92	0.626	0.672	0.655	0.695	0.597	0.592	0.616	0.580	0.576	0.576	9	4
impgsc92	0.807	0.831	0.797	0.766	0.793	0.763	0.792	0.762	0.771	0.771	8	2
finslc92	0.748	0.762	0.740	0.700	0.740	0.682	0.740	0.676	0.702	0.702	8	2
dpic92	0.701	0.757	0.696	0.718	0.697	0.689	0.767	0.673	0.673	0.673	8	7
wascur	0.688	0.758	0.680	0.744	0.682	0.698	0.724	0.693	0.685	0.685	3	2
m2sl	0.945	0.928	0.927	0.926	0.926	0.925	0.926	0.925	0.925	0.945	8	1
unrate	0.841	0.836	0.846	0.825	0.842	0.818	0.834	0.822	0.818	0.841	6	3
tb3ma	0.787	0.794	0.780	0.783	0.782	0.752	0.783	0.760	0.775	0.775	6	2
gs1	0.827	0.818	0.830	0.797	0.822	0.806	0.829	0.808	0.800	0.800	4	3
gs10	0.804	0.796	0.818	0.842	0.793	0.797	0.794	0.795	0.795	0.795	5	4
fed	0.803	0.793	0.796	0.770	0.797	0.749	0.796	0.759	0.739	0.739	9	1
gdpdef	0.792	0.784	0.803	0.761	0.788	0.749	0.795	0.753	0.755	0.755	6	3
cpiaucs	0.938	0.939	0.927	0.926	0.936	0.926	0.936	0.925	0.926	0.938	8	2

Notes: UP_1 is the forecast based on a unit root pretest where PW_1 is used if a unit root is rejected. UP_2 is the forecast based on a unit root pretest where OLS_1 is used if a unit root is rejected. B and W denote the best and worst forecast where the numbering corresponds to the forecasting procedure from left (1) to right (10).

Table 4c: Median Bias

	OLS_1	OLS_2	CO_0	PW_0	CO_1	PW_1	CO_∞	PW_∞	UP_1	UP_2
gdpc92	0.450	0.350	0.450	0.280	0.450	0.380	0.530	0.410	0.440	0.440
gpdic92	0.420	0.390	0.430	0.400	0.430	0.400	0.430	0.400	0.400	0.420
expgsc92	0.360	0.400	0.310	0.530	0.340	0.470	0.330	0.430	0.430	0.430
impgsc92	0.580	0.540	0.580	0.540	0.580	0.540	0.580	0.550	0.580	0.580
finslc92	0.540	0.530	0.540	0.460	0.540	0.450	0.540	0.430	0.480	0.480
dpic92	0.460	0.320	0.460	0.380	0.460	0.380	0.500	0.400	0.420	0.420
wascur	0.420	0.350	0.430	0.350	0.430	0.410	0.530	0.430	0.430	0.430
m2sl	0.380	0.400	0.400	0.400	0.400	0.400	0.400	0.400	0.400	0.380
unrate	0.500	0.500	0.500	0.580	0.500	0.550	0.500	0.560	0.550	0.500
tb3ma	0.490	0.500	0.500	0.560	0.490	0.540	0.480	0.550	0.490	0.490
gs1	0.470	0.480	0.480	0.510	0.480	0.500	0.480	0.500	0.510	0.510
gs10	0.490	0.530	0.520	0.560	0.490	0.530	0.500	0.530	0.510	0.510
fed	0.460	0.470	0.460	0.550	0.460	0.530	0.450	0.570	0.510	0.510
gdpdef	0.420	0.440	0.440	0.480	0.430	0.450	0.430	0.440	0.450	0.450
cpiaucs	0.470	0.470	0.480	0.470	0.470	0.480	0.470	0.480	0.480	0.470

Table 4c: Empirical Examples: Number of Times (out of 100) for Forecast has Minimum RMSE

	OLS_1	OLS_2	CO_0	PW_0	CO_1	PW_1	CO_∞	PW_∞	UP_1	UP_2
gdpc92	3	24	0	13	2	4	37	1	16	0
gpdic92	7	20	4	17	0	0	38	14	0	0
expgsc92	3	12	7	21	11	3	31	7	5	0
impgsc92	3	29	2	6	0	2	3	3	52	0
finslc92	22	4	1	18	0	1	0	19	35	0
dpic92	1	31	2	8	2	1	30	4	21	0
wascur	6	18	1	18	5	4	47	0	1	0
m2sl	39	5	0	8	0	2	1	45	0	0
unrate	0	5	3	34	1	6	42	9	0	0
tb3ma	14	5	3	27	2	5	1	16	27	0
gs1	10	6	3	30	2	3	1	17	28	0
gs10	14	7	7	34	1	6	3	17	11	0
fed	12	7	2	27	0	2	6	13	31	0
gdpdef	15	4	11	40	1	4	1	19	5	0
cpiaucs	10	13	8	12	4	1	15	37	0	0

Table 4d: Empirical Examples: Number of Times (out of 100) for Forecast has Largest RMSE

	OLS_1	OLS_2	CO_0	PW_0	CO_1	PW_1	CO_∞	PW_∞	UP_1	UP_2
gdpc92	0	16	0	40	0	1	40	0	3	0
gpdic92	9	23	1	25	1	0	27	14	0	0
expgsc92	3	20	16	34	1	1	23	2	0	0
impgsc92	14	42	1	2	1	0	0	1	39	0
finslc92	32	14	0	16	0	0	0	7	31	0
dpic92	2	52	0	3	0	0	36	0	7	0
wascur	1	23	0	33	0	1	42	0	0	0
m2sl	61	7	0	3	0	1	0	28	0	0
unrate	2	7	9	36	2	2	34	8	0	0
tb3ma	19	4	1	35	3	3	0	8	27	0
gs1	23	4	3	31	0	5	2	9	23	0
gs10	19	5	7	44	0	4	1	11	9	0
fed	24	0	0	35	0	2	2	13	24	0
gdpdef	19	0	23	37	0	1	1	15	4	0
cpiaucs	12	13	7	13	11	2	11	31	0	0

Figure 1a: OLS_1 p=0

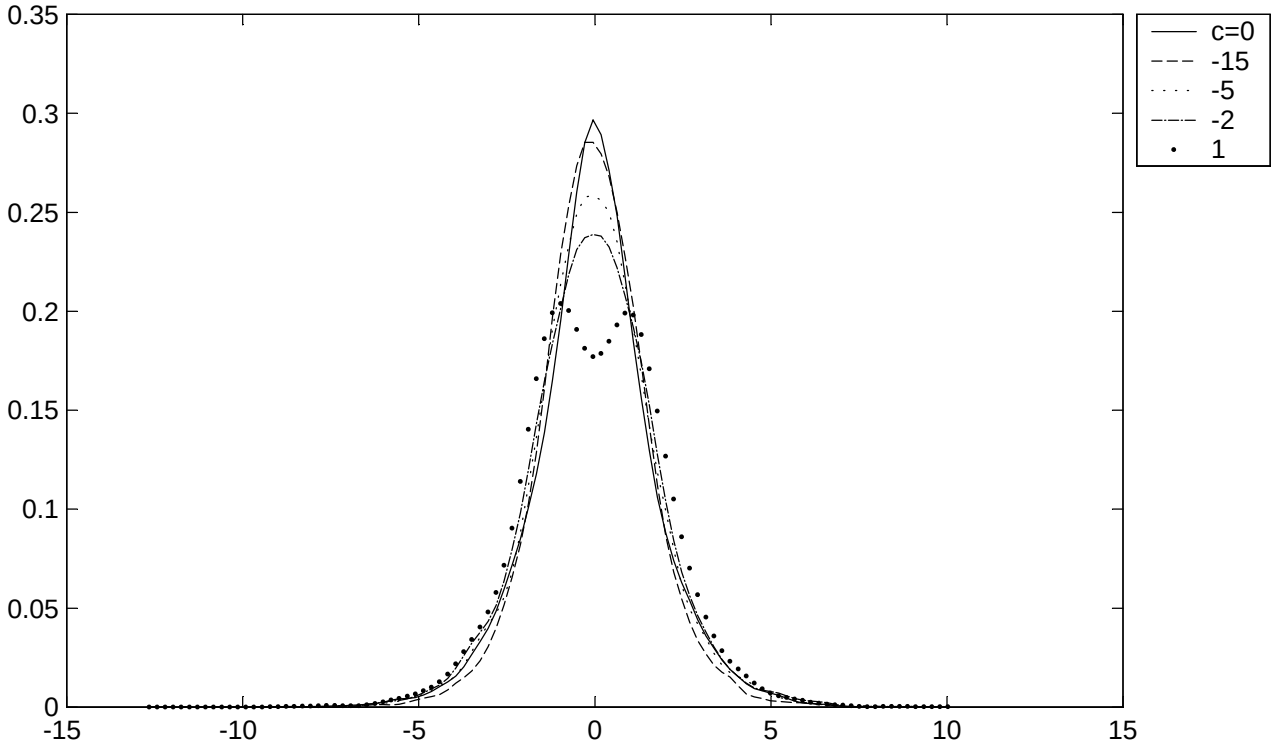

Figure 1b: OLS_2 p=0

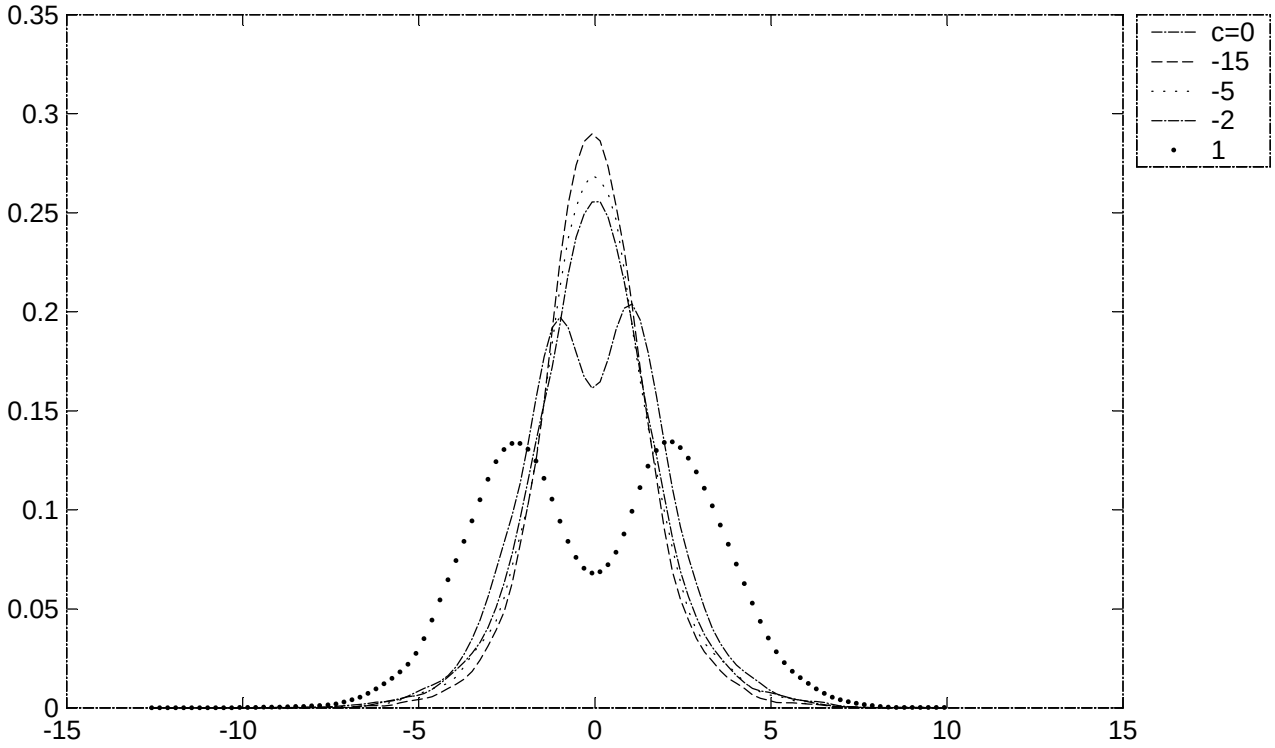

Figure 2a: OLS_1 p=1

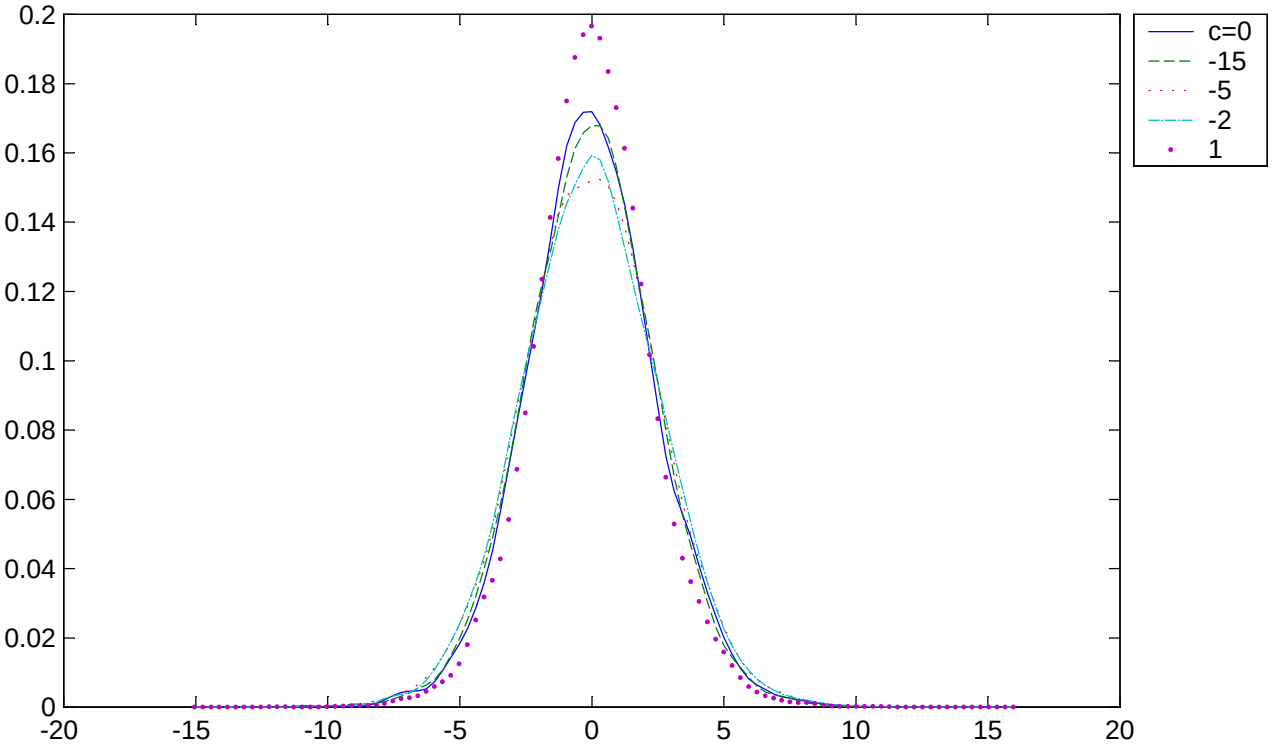

Figure 2b: OLS_2 p=1

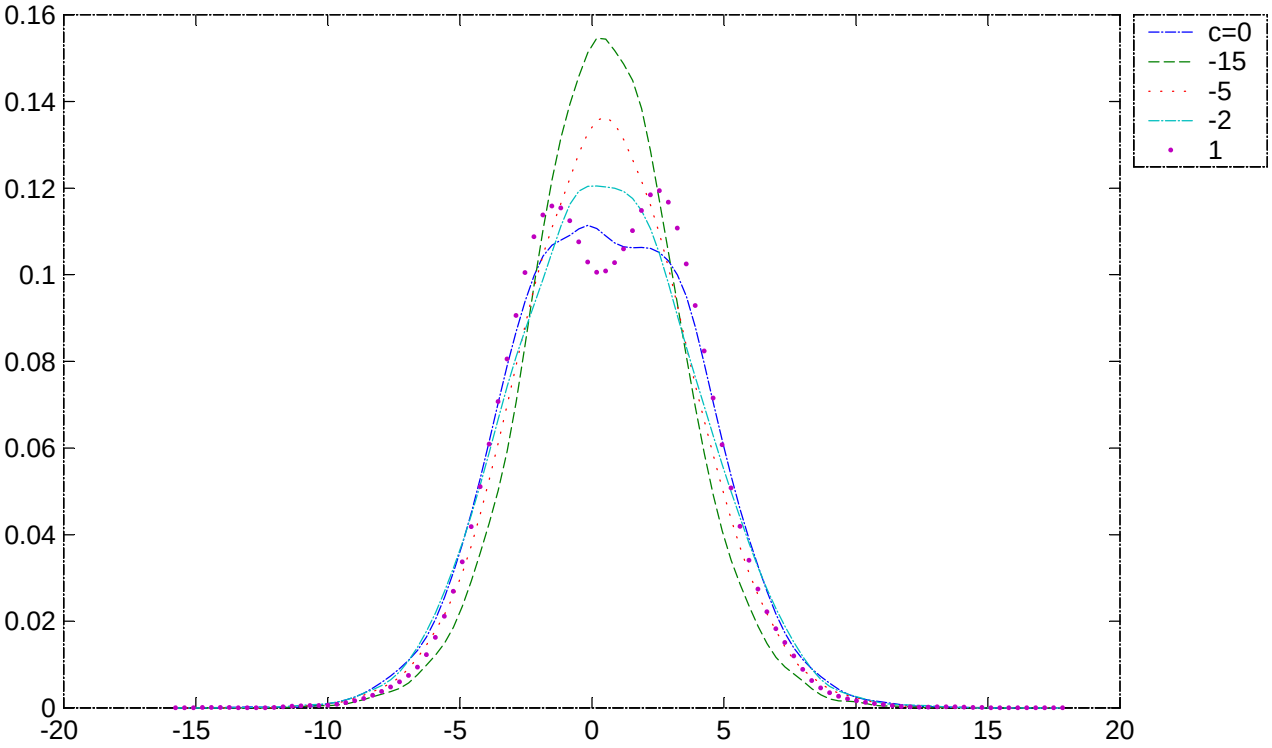

Figure 3a: OLS_1 p=0: Conditional on y_T-m_T>0

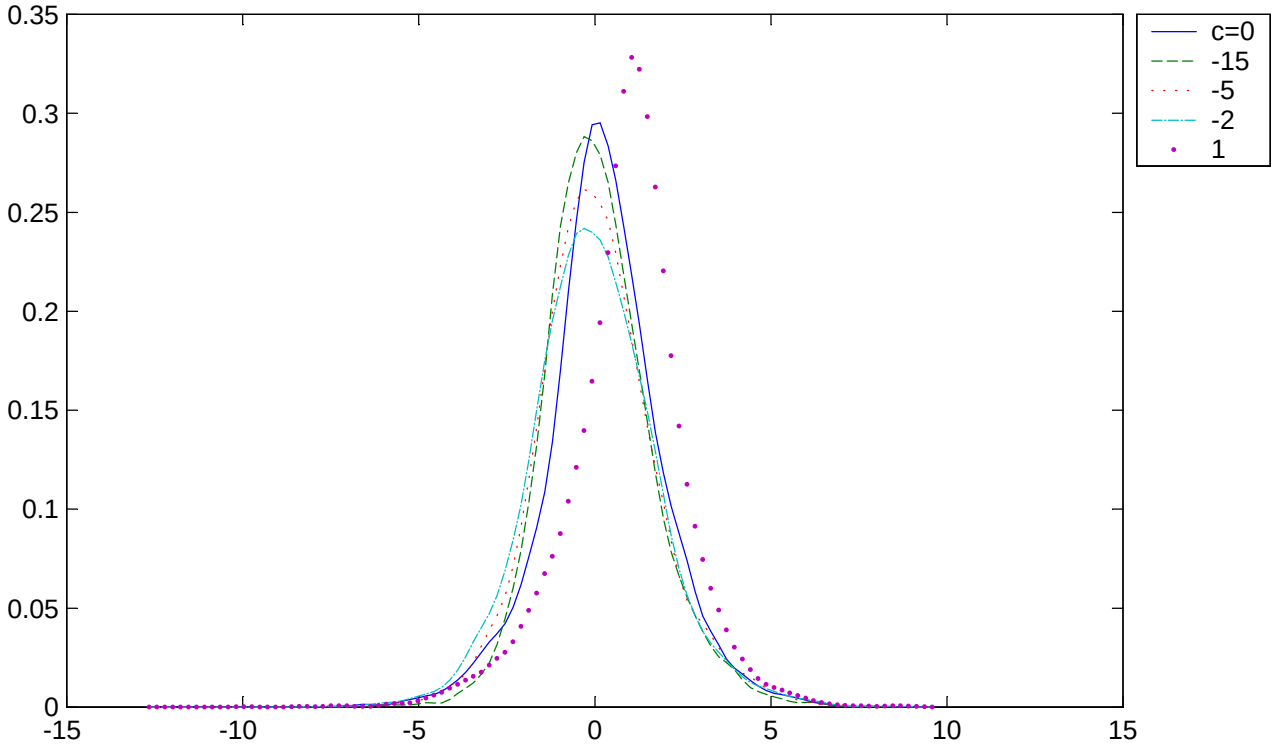

Figure 3b: OLS_2 p=0: Conditional on y_T-m_T>0

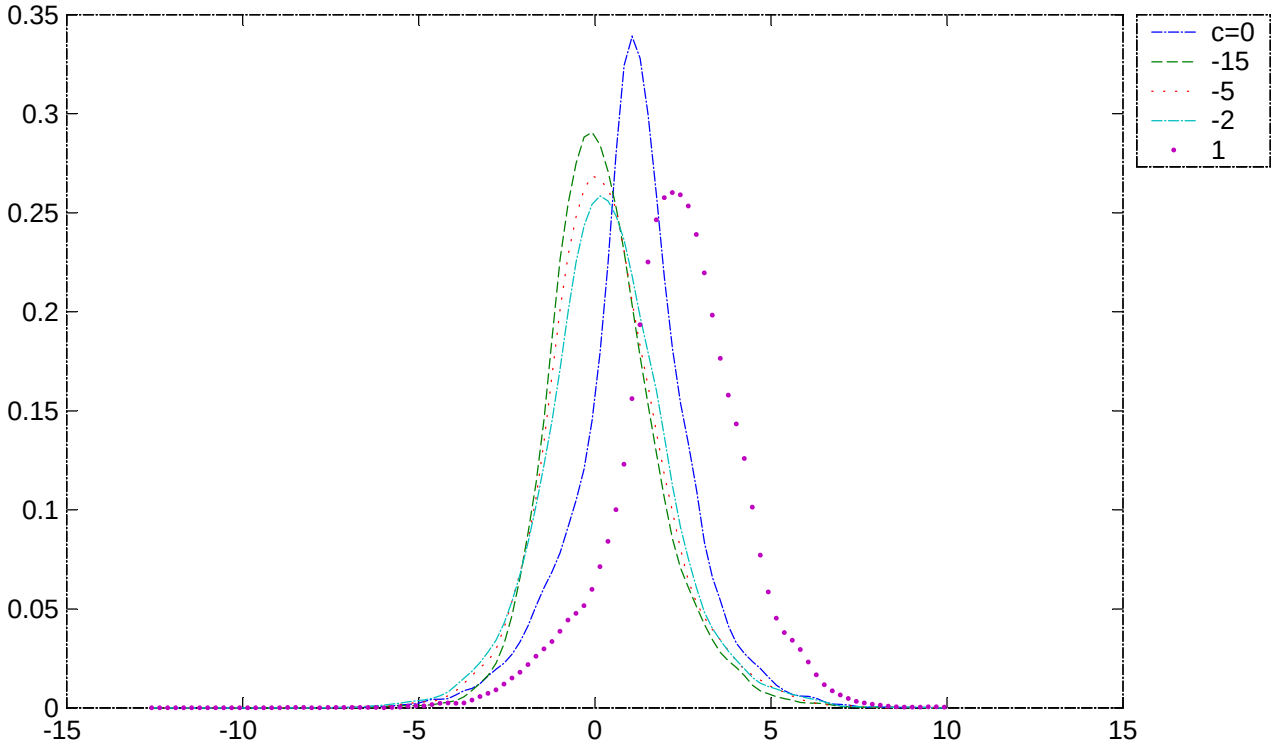

Figure 4a: OLS_1 p=1: Conditional on y_T-m_T>0

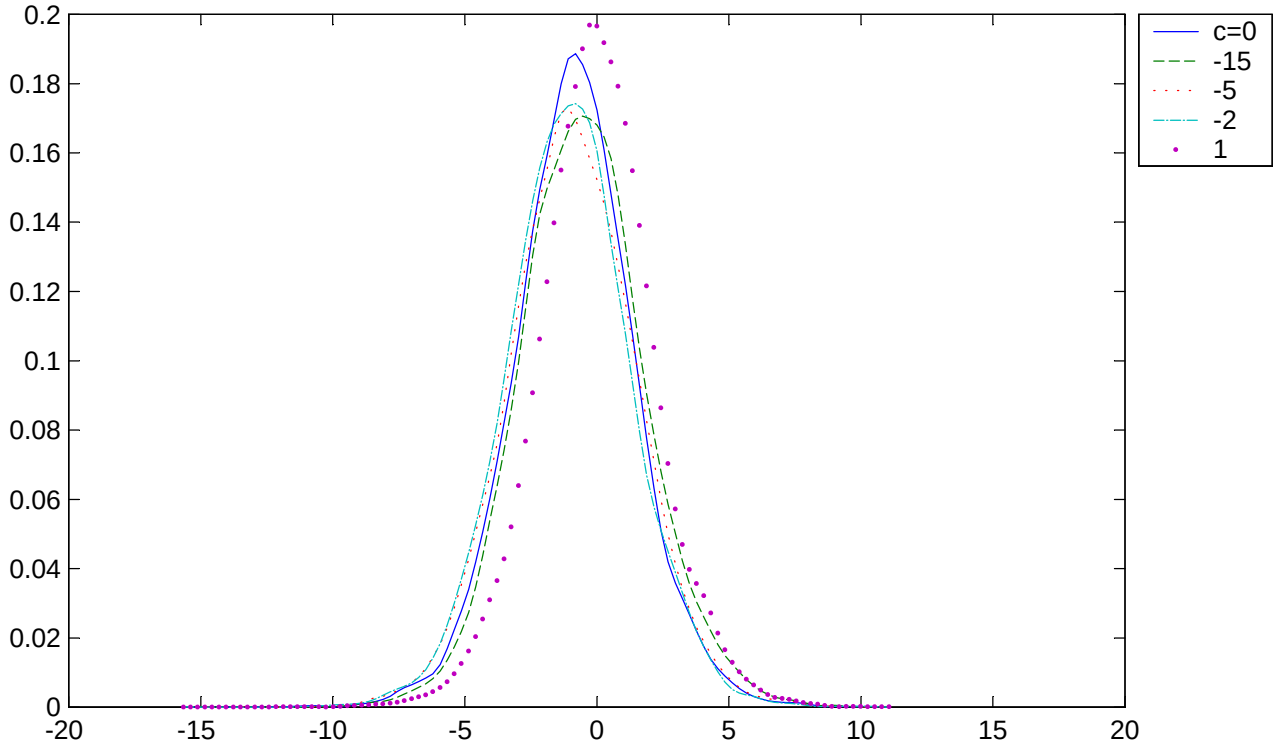

Figure 4b: OLS_2 p=1: Conditional on y_T-m_T>0