Contrasting Bayesian and Frequentist Approaches to Autoregressions: the Role of the Initial Condition*

Marek Jarociński
European Central Bank

Albert Marcet
Institut d'Anàlisi Econòmica CSIC,
ICREA, Barcelona GSE, UAB, MOVE

September 20, 2014

Abstract

The frequentist and the Bayesian approach to the estimation of autoregressions are often contrasted. Under standard assumptions, when the ordinary least squares (OLS) estimate is close to 1, a frequentist adjusts it upwards to counter the small sample bias, while a Bayesian who uses a flat prior considers the OLS estimate to be the best point estimate. This contrast is surprising because a flat prior is often interpreted as the Bayesian approach that is closest to the frequentist approach. We point out that the standard way that inference has been compared is misleading because frequentists and Bayesians tend to use different models, in particular, a different distribution of the initial condition. The contrast between the frequentist and the Bayesian flat prior estimation of the autoregression disappears once we make the same assumption about the initial condition in both approaches.

Keywords: Autoregression, Initial Condition, Bayesian Estimation, Small Sample Distribution, Bias Correction (*JEL codes:* C11, C22, C32)

*We thank Manolo Arellano, Peter Phillips, Marcet acknowledges support from Axa Research Fund, Plan Nacional (Ministerio de Educación y Ciencia), SGR (Generalitat de Catalunya), Programa de Excelencia del Banco de España and the European Research Council under the EU 7th Framework Programme (FP/2007-2013) Advanced Grant Agreement No. 324048. The opinions expressed herein are those of the authors and do not necessarily represent those of the European Central Bank. Contacts: albert.marcet@iae.csic.es and marek.jarocinski@ecb.int.

1 Introduction

From the frequentist point of view the OLS estimator is biased in autoregressions. This has been known since Quenouille (1949) and Hurwicz (1950). Adjustments of the OLS estimator that reduce this bias have been proposed by Marriott and Pope (1954), Andrews (1993), MacKinnon and Smith (1998), and many others. When the autoregressive parameter of an AR(1) model is in the vicinity of 1 the OLS estimate of this parameter is biased downward, therefore a frequentist econometrician facing an OLS estimate close to 1 adjusts it upwards.

But from the Bayesian point of view the OLS estimator is a good summary of sample information. This is because the posterior obtained with a flat prior is centered at the OLS estimate and it is symmetric around it. Bayesian econometricians who just want to characterize the shape of the likelihood often use the flat prior. Such a Bayesian econometrician does not adjust the OLS estimate.

Therefore, the frequentist view is that the OLS estimate should be adjusted and the above Bayesian view is that it should not. This is surprising because a flat prior is often understood as the closest that the Bayesian approach can get to the frequentist approach. The contrast between the above views has been highlighted, for example, in Sims and Uhlig (1991), as a manifestation of fundamental differences between the Bayesian and frequentist approaches to inference: bias is a property of the behavior of an estimator across different samples, so it is relevant to a frequentist but irrelevant to a Bayesian, who conditions only on the observed sample.

We first point out that contrasting the two views about the OLS estimator as is done in the previous paragraph is misleading, since they are based on different models, in particular, a different model of the initial condition. We then show using a Monte Carlo (the "helicopter tour" approach of Sims and Uhlig (1991)) that the sharp contrast disappears when the same model for the initial condition is used. When frequentists and Bayesians use a distribution of the initial condition that depends on the AR parameters in a standard way, they both adjust the OLS estimate upwards (when it is close to 1). When they use a flat density for the initial condition,[1] both frequentists and Bayesians facing an OLS estimate believe it

[1] In other words, when the researcher takes the initial condition as given and therefore uses the so-called "conditional likelihood."

summarizes the sample information optimally and do not adjust it. We discuss the intuition behind these Monte Carlo results.

The implication of our findings is that in most practical situations both frequentist and Bayesian econometricians agree that *an autoregressive process is more persistent than the OLS estimate suggests*. First, because in most practical situations it is reasonable to assume that the initial condition is related to the parameters that apply in the rest of the sample, second, because the OLS estimate is often in the range where it gets adjusted towards higher values, implying a more persistent process.

We believe that this message is important for practitioners. Most applied econometricians do not have strong views on the deep foundations of statistics. It may be unsettling if the judgment on whether the persistence of the OLS estimate is appropriate or not hinged on whether one adopts the Bayesian-flat-prior or the frequentist perspective. Adjusting the OLS estimator of an autoregression often has relevant practical consequences, specially in vector autoregressions (VARs).[2] For example, numerous forecasting studies find that increasing the persistence of a VAR is crucial for improving the forecasting performance, this is in part why the Litterman (1986) prior is so popular. For another example, the persistence of the estimated VAR crucially affects the size of term premia in dynamic term structure models (see Bauer et al. (2012)). Reassuringly, we show that broad lessons that a reasonable econometrician would draw from the data are not sensitive to deep foundations of the inference approach.

Please note some clarifications. First, we do not propose a Bayesian analysis of an autoregression that yields an unbiased point estimate, nor do we study conditions under which Bayesian point estimates are unbiased (on this topic see, e.g., Firth (1993)). Second, we do not contribute to the debate on whether the flat prior is indeed noninformative in autoregressions (see Phillips (1991) and the ensuing debate in the Econometric Theory 1994 special issue). We just take it for granted that our Bayesian econometrician uses the flat prior as a device for reporting the shape of the likelihood. Third, the claims in the paper hold for an AR model where the constant term is unknown, as is usually the case in applied

[2]Vector autoregressions are commonly used in macroeconomics where the samples are small. Moreover, the small sample bias of the OLS estimate towards stationarity increases with dimension. The frequentist proof of this fact is in Abadir et al. (1999) and some Bayesian arguments are in Sims and Zha (1998) p.959.

work, the special case of the known constant term, studied in Sims and Uhlig (1991), is different.[3] Fourth, Sims and Uhlig highlight the fact that a Bayesian posterior is a different object than the frequentist distribution of an estimator; we of course agree with this point. We only highlight that when the constant term is unknown, as is typically the case, both Bayesians and frequentists agree on whether to adjust the OLS estimate upwards or not, provided that they use the same model for the initial condition.

Section 2 explains in more detail the contrasting views of Bayesians and frequentists as they have been described in the literature. Section 3 describes two standard models for the initial condition. Section 4 shows that whether or not the OLS estimate is adjusted depends on the model chosen in section 3.

2 The false contrast

Consider a sample $y \equiv \{y_0, ..., y_T\}$ that is modeled as an AR(1) process with an intercept:

$$y_t = \alpha + \rho y_{t-1} + u_t, \quad u_t \text{ i.i.d. } N(0, \sigma^2) \tag{1}$$

$t = 1, \ldots, T$. y_0 is the initial condition. The OLS estimates of α and ρ are denoted α^{OLS} and ρ^{OLS}.

Throughout the paper, we focus on the small sample inference and not on asymptotics. Therefore, when we say "frequentist" we mean "frequentist, small sample".

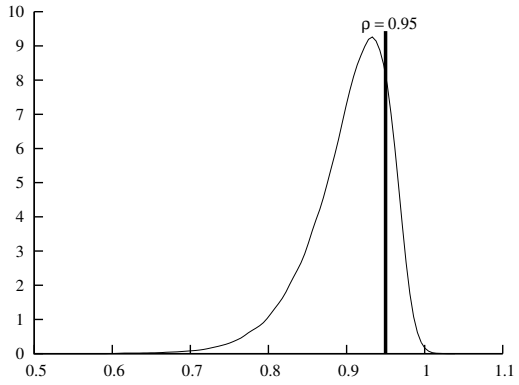
The following two views of ρ^{OLS} are often contrasted.

Frequentist (View A) For given true values of (α, ρ, σ) and with ρ near 1, the small sample distribution of ρ^{OLS} is skewed to the left and its mean is lower than ρ. Such a distribution for $\rho = 0.95$ and $T = 100$ is presented in the left panel of Figure 1. This is a well known figure, it displays the skewness and the bias, in this case $E(\rho^{OLS}) = 0.91$ and the bias is -0.04. Based on this picture, a frequentist econometrician facing ρ^{OLS} believes that the true ρ is likely to be higher, so to obtain a better estimate of ρ she adjusts ρ^{OLS} upwards. When constructing this picture we draw y_0 from its ergodic distribution as this

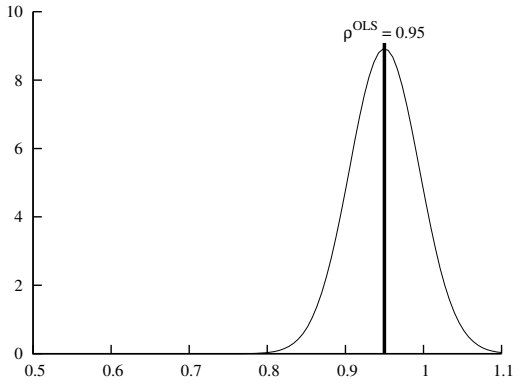
[3] When the value of the constant term is known, the contrast is indeed there, since the posterior of a Bayesian who uses a flat prior is centered at the OLS estimator regardless of the model of the initial condition.

is the standard procedure in the frequentist literature.[4]

Bayesian (View D)[5] When the prior for $(\alpha, \rho, \ln \sigma)$ is flat, the posterior distribution of ρ is symmetric and centered precisely at ρ^{OLS}. An example of this distribution is presented in the right panel of Figure 1 for a particular sample y. Therefore, a Bayesian econometrician who uses a flat prior does not adjust ρ^{OLS}.[6] This holds when we use the likelihood function *conditional* on the initial condition y_0, which is the standard approach.[7]



Frequentist density of ρ^{OLS} when $\rho = 0.95$



Bayesian posterior density of ρ centered at $\rho^{OLS} = 0.95$

Figure 1 – Two views of ρ^{OLS}

The typical interpretation of the contrast between view A and view D is that this is due to the different approaches to statistics. The small sample bias is a frequentist property. Therefore, the frequentist econometrician in A worries about the bias and adjusts ρ^{OLS} in order to correct it. However, the bias is irrelevant to the Bayesian in D, who conditions on the observed sample, and disregards other, unobserved samples, so this Bayesian embraces ρ^{OLS}. This is the point of Sims and Uhlig (1991).

[4]The distribution presented in the left panel of Figure 1 is a smoothed histogram of values of ρ^{OLS} computed from 10,000 samples of 100 observations each, simulated from the process (1). We assume $\alpha = 0, \sigma = 1$.

[5]We use the letter D, because in what follows we introduce also two intermediate views, which we label B and C.

[6]Throughout this paper we use a quadratic loss function for the Bayesian approach. Obviously, nonstandard loss functions would justify other point estimates.

[7]The distribution presented in the right panel of Figure 1 is the marginal posterior $p(\rho|y)$, conditional on a sample y. To obtain this posterior, multiply the prior kernel by the likelihood kernel $p(y_1, ..., y_T|\alpha, \rho, \sigma^2, y_0) \propto \sigma^{-T} \exp\left(-\frac{1}{2}\sum_{t=1}^{T}(y_t - \alpha - \rho y_{t-1})^2/\sigma^2\right)$ and integrate α and σ out. The result is a kernel of the Student-t density. This is a textbook derivation, see e.g. Zellner (1971) Ch.3.1.

Table 1 – Does the OLS estimate need an adjustment?

	Approach to statistics	
model of y_0	Frequentist	Bayesian
depends on α, ρ, σ	A: YES	B: ?
does not depend on α, ρ, σ	C: ?	D: NO

Given that in other contexts Bayesian-flat-prior and frequentist inference are often quite close, this difference is surprising. The implications of the adjustment of OLS estimates in applied work, especially in VARs, can be substantial.

Our point is that there is another difference between views A and D beyond the different approaches to statistics, namely they use different models of y_0. In view A y_0 is drawn from the ergodic distribution, therefore the distribution of y_0 is related to the true value of the parameters α, ρ, σ. View D, in contrast, assumes that the model of y_0 is unrelated to parameters. This is implicit in the use of the standard likelihood function conditional on the observation y_0, $p(y_1, ..., y_T|\alpha, \rho, \sigma, y_0)$. This so-called conditional likelihood function is the one used by flat-prior Bayesians, it has no term relating y_0 to the parameters α, ρ, σ, therefore it is consistent with a flat density for the initial condition, namely $p(y_0|\alpha, \rho, \sigma) \propto 1$. Therefore, in the rest of the paper we interpret that the Bayesian flat-prior approach that uses a conditional likelihood is in effect assuming that y_0 has a flat distribution.

To see what drives the contrast between A and D we need to make the models comparable. We propose to use the same model for the initial condition under each inference approach. In other words, we need to answer the following two questions.

Question B, Bayesian Does a Bayesian econometrician adjust ρ^{OLS} when the model of y_0 depends on parameters α, ρ, σ?

Question C, Frequentist Does a frequentist econometrician adjust ρ^{OLS} when the model of y_0 does not depend on parameters α, ρ, σ?

Table 1 summarizes the above discussion. There are four possible inferences combining two approaches to statistics and two models of the initial condition. The contrast in the literature has focused on cells A and D, in the remainder of the paper we fill cells B and C.

3 Two models of the initial condition y_0

We now posit two models for y_0. Drawing from the literature we assume

$$y_0 \sim \begin{cases} \mathrm{N}\left(\frac{\alpha}{1-\rho},\ \kappa^2 \frac{\sigma^2}{1-\rho^2}\right) & \text{if } |\rho| < 1 \\ 0 & \text{if } \rho = 1 \\ \mathrm{N}\left(\frac{\alpha}{1-\rho},\ \kappa^2 \sigma^2\right) & \text{otherwise.} \end{cases} \tag{2}$$

We consider two cases, $\kappa = 1$ and $\kappa = 100$, which give rise to our two models.

When $\kappa = 1$ our model is the same as, e.g., in Bhargava (1986). The first line states that when $|\rho| < 1$, i.e., when the process is stationary and has an ergodic distribution, then y_0 is drawn from this ergodic distribution, $\mathrm{N}\left(\frac{\alpha}{1-\rho}, \frac{\sigma^2}{1-\rho^2}\right)$. However, the ergodic distribution does not exist when $|\rho| \geq 1$. In this case, one needs a more or less arbitrary assumption. Following Bhargava and many others, we assume $y_0 = 0$ in the unit root case, and we assume $y_0 \sim \mathrm{N}\left(\frac{\alpha}{1-\rho},\ \sigma^2\right)$ in the explosive case.[8]

Now consider the case $\kappa = 100$. As $\kappa \to \infty$, the relation between y_0 and model parameters becomes weaker, and the density $p(y_0|\alpha, \rho, \sigma)$ becomes flat, i.e., proportional to a constant. Consequently, the likelihood function of the whole sample $p(y_0, y_1, ..., y_T|\alpha, \rho, \sigma)$ becomes proportional to the "conditional" likelihood function $p(y_1, ..., y_T|\alpha, \rho, \sigma, y_0)$ and, as we argued before, the flat-prior approach implicitly uses a κ close to ∞ . We use $\kappa = 100$ to approximate this situation in our numerical simulations.

Now, that we have specified the two models for y_0, we are ready to answer questions B and C stated above.

4 Joint density of ρ and ρ^{OLS}: A helicopter tour

To study the relation between ρ and ρ^{OLS} we adapt the approach of Sims and Uhlig (1991) to the case of an unknown constant α and simulate the joint distribution of ρ and ρ^{OLS} by

[8] Our reading of the literature is that this is the most popular assumption about the distribution of y_0, but there are alternatives. For example, MacKinnon and Smith (1998), p.2, assume a variance equal to σ^2 both in the stationary and in the explosive case. We obtain qualitatively similar results in all the computations below when we use their assumption. Other approaches to the initial condition can be found, for example, in Uhlig (1994) and Phillips and Magdalinos (2009).

Monte Carlo. The joint distribution of ρ and ρ^{OLS} depends on the sample size T and on the model of the initial condition. We assume $T = 100$ and we generate two joint distributions of ρ and ρ^{OLS}, one with $\kappa = 1$ in (2) and the second one with $\kappa = 100$ in (2). We maintain a flat prior on ρ. The distribution of $\rho^{OLS}|\rho$ is independent of the values of α and σ when a model of the form (2) is used for y_0. This issue as well as the details of the computations are discussed in the Appendix.

Figure 2 presents isoprobability contours for the two joint distributions. The left panel is for $\kappa = 1$ and the right panel is for $\kappa = 100$. In both panels we also plot the $\rho = \rho^{OLS}$ line.[9] Let us make two observations about Figure 2.

First, consider the density for $\kappa = 1$, presented in the left panel. In the plotted range for ρ and ρ^{OLS} more probability mass lays below the $\rho = \rho^{OLS}$ line than above it. Fix a value of ρ, most of the mass of ρ^{OLS} is at values lower than ρ. Fix a value of ρ^{OLS}, most of the mass of ρ is at values higher than ρ^{OLS}.

Second, consider the density for $\kappa = 100$, in the right panel. This density is very concentrated at the $\rho = \rho^{OLS}$ line. There might be more probability mass below the $\rho = \rho^{OLS}$ line than above it, as in the left panel, but this effect is not quantitatively relevant.

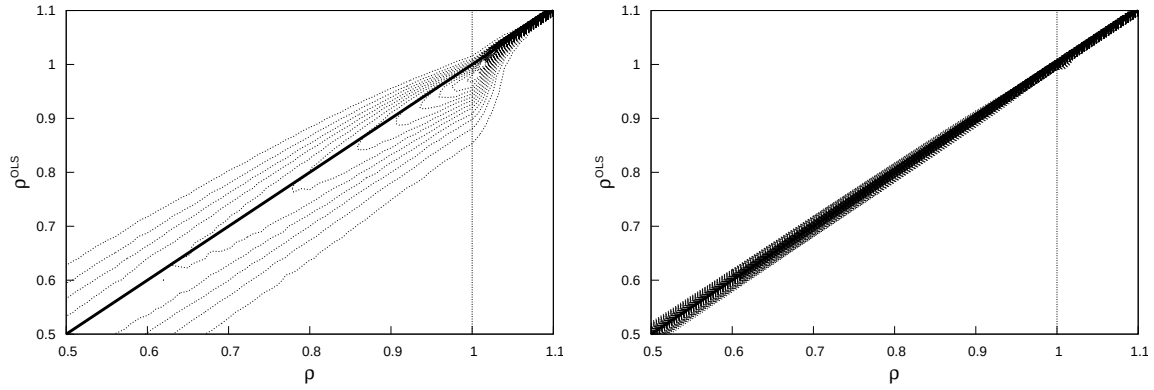
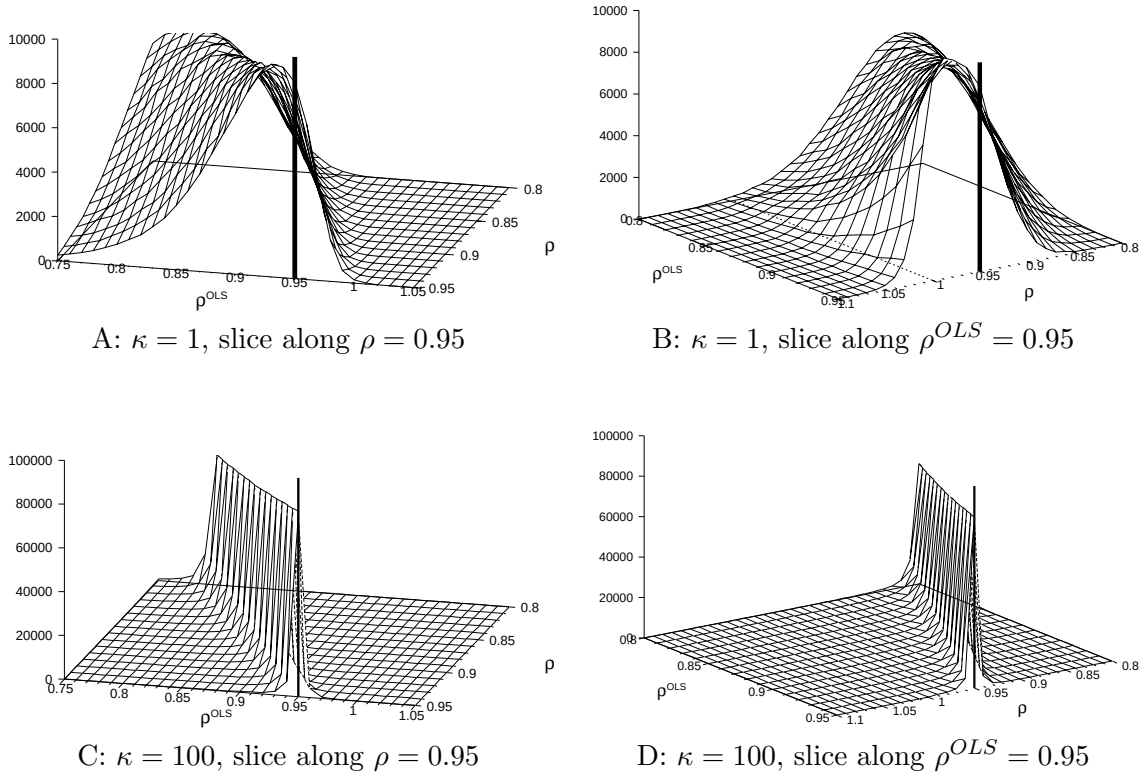


Figure 2 – Two joint densities of ρ and ρ^{OLS} - isoprobability contours.

We now follow in the footsteps of Sims and Uhlig (1991) and take a helicopter tour of

[9]Note also that both distributions have a singularity at $\rho = 1$: at this measure-zero region the distribution is not defined uniquely. This singularity does not matter for our discussion, we discuss this singularity in the appendix.

the two joint distributions of (ρ, ρ^{OLS}). The difference is that our tour is over the land of an AR(1) model *with a constant* α and with two explicit assumptions for y_0. The tour is shown in Figure 3, displaying selected cuts of the two joint densities from Figure 2. The first row of Figure 3 corresponds to the density obtained with $\kappa = 1$ while the second row corresponds to the density obtained with $\kappa = 100$.



A: $\kappa = 1$, slice along $\rho = 0.95$ B: $\kappa = 1$, slice along $\rho^{OLS} = 0.95$

C: $\kappa = 100$, slice along $\rho = 0.95$ D: $\kappa = 100$, slice along $\rho^{OLS} = 0.95$

Figure 3 – Two bivariate densities of ρ and ρ^{OLS}: various cuts

Note that the panels of Figure 3, labeled A, B, C and D correspond to the cells of Table 1. A cut along a fixed-ρ line (a vertical line in Figure 2) reveals the frequentist distribution of ρ^{OLS} given a fixed value of ρ. Panels A and C present such cuts for $\rho = 0.95$.

A cut along a fixed-ρ^{OLS} line (a horizontal line in Figure 2) reveals the Bayesian posterior distribution of ρ conditional on observing a given value of ρ^{OLS}. The distribution $p(\rho|\rho^{OLS})$ is a convenient summary of posterior distributions of ρ conditional on the data y, $p(\rho|y)$ since it averages out across realizations of the data y. It also answers directly the question

how a Bayesian forms his beliefs about ρ upon observing ρ^{OLS} only.[10] Panels B and D present such cuts for $\rho^{OLS} = 0.95$.

The cut in panel A is familiar. This is the small sample distribution of ρ^{OLS} when $\kappa = 1$, the same as the one plotted in the left panel of Figure 1, to illustrate view A.

The cut in panel D is the Bayesian posterior distribution of ρ given ρ^{OLS}. $\kappa = 100$ approximates the use of the "conditional likelihood." This posterior distribution is different from the posterior in the right panel of Figure 1, because it conditions on ρ^{OLS} and not on y, but of course it confirms the view that ρ^{OLS} should not be adjusted.

4.1 Does the Bayesian econometrician adjust ρ^{OLS} when y_0 is related to parameters α, ρ, σ? (Question B)

We now consider the model given by (1)-(2) with $\kappa = 1$ from the Bayesian point of view and with the flat prior on α, ρ. Bayesian estimation with proper models for initial conditions (using the so-called "exact likelihood") have been studied in Zellner (1971, ch.7.1), Schotman and Van Dijk (1991), Uhlig (1994) and Lubrano (1995). They find, like us, that a Bayesian who assumes a similar distribution of y_0 shifts her posterior distribution of ρ towards values greater than ρ^{OLS}. But these papers used non-flat priors on α, ρ so it is not clear if the higher posterior mean they find is due to the priors or to the model of the initial condition. Therefore they do not answer question B in Table 1. We maintain a flat prior on α, ρ in order to compare with the frequentist approach and to better isolate the effect of modeling the density of y_0.

Panel B of Figure 3 provides the answer to question B, and the answer is YES. When reading this Figure note that, as in Sims and Uhlig (1991), the axis for ρ increases from right to left. The cut presented in panel B exposes the density $p(\rho|\rho^{OLS} = 0.95)$. This density is clearly asymmetric and its mean is higher than the OLS estimate. This is striking, because the analogous figure in Sims and Uhlig (1991) (their Figure 5) for the model without the constant term shows a symmetric density, but in our setup the density becomes asymmetric. In Figure 4 we reproduce the same density again in a two-dimensional graph, but flipping

[10] For a general discussion of posteriors conditional on sample statistics, such as ρ^{OLS}, rather than conditional on the whole sample y, see Kwan (1998).

the horizontal axis, so that the values of ρ are increasing. In the present example, upon observing $\rho^{OLS} = 0.95$ a Bayesian would believe that the true value of ρ is around 0.97, adjusting the OLS estimate upwards, in the same direction as a frequentist who corrects the small sample bias. Figure 4 is simpler to read, but the value added of Figures 2 and 3 is that they show that the densities at other points between, roughly, $\rho^{OLS} = 0.7$ and $\rho^{OLS} = 1$, are qualitatively similar to Figure 4. This illustrates that when the model relates y_0 to the parameters (α, ρ, σ), then in this range of values of ρ^{OLS} Bayesians tend to agree with the frequentists that the OLS estimate is too low and it should be adjusted upwards. Therefore, we answer YES to question B.

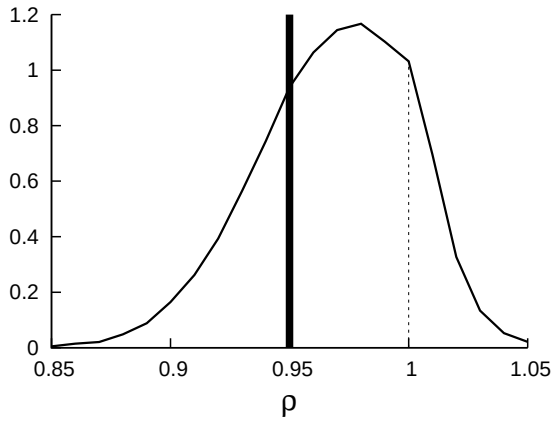


Figure 4 – Bayesian density $p(\rho|\rho^{OLS} = 0.95)$ assuming $\kappa = 1$ (case B). This is a smoothed histogram from the Monte Carlo experiment explained in the appendix. The dotted vertical line at $\rho = 1$ signals that at $\rho = 1$ the density is not uniquely defined.

Intuition. Let us provide an intuition about the results just shown. We will argue that the effect of including the model (2) for y_0 into the likelihood function is similar to the effect of including a data point in the sample as if y_0 would arise from some observation at a notional date. This will be similar to adding an observation consistent with the Random Walk model with $\rho = 1$ and this is likely to cause an adjustment of the posterior mean of ρ towards 1 (which, for $\rho^{OLS} < 1$, means "upwards"). The analytic details follow.

The posterior with initial condition (2) for $\rho \neq 1$ and $\kappa = 1$ is

$$p(\alpha, \rho | y, \sigma^2) \propto \sqrt{1-\rho^2} \exp\left(-\frac{1}{2\sigma^2} \left[\sum_{t=1}^{T} (y_t - \alpha - \rho y_{t-1})^2 + (y_0(1-\rho) - \alpha)^2 \frac{1+\rho}{1-\rho} \right] \right) \text{ for } |\rho| < 1 \tag{3}$$

$$\propto \exp\left(-\frac{1}{2\sigma^2} \left[\sum_{t=1}^{T} (y_t - \alpha - \rho y_{t-1})^2 + (y_0(1-\rho) - \alpha)^2 \frac{1}{(1-\rho)^2} \right] \right) \text{ for } |\rho| > 1$$

We first consider an approximation to the posterior where the terms $\sqrt{1-\rho^2}$, $(1+\rho)/(1-\rho)$ and $(1-\rho)^2$ that appear in this expression are substituted by a constant, so we consider the approximate posterior

$$p^{approx}(\alpha, \rho | y, \sigma^2) \propto \exp\left(-\frac{1}{2\sigma^2} \left[\sum_{t=1}^{T} (y_t - \alpha - \rho y_{t-1})^2 + \frac{\sigma^2}{\sigma_v^2} (y_0 - \alpha - \rho y_0)^2 \right] \right) \tag{4}$$

for a fixed σ_v^2 and for all ρ. The result below will show that in this posterior the term inside the $\exp(\cdot)$ operator pushes an estimator $\rho^{OLS} < 1$ upwards, that is, towards one. The calculations shown in Figure 4 show that this effect actually dominates. We discuss informally the effect of the terms $\sqrt{1-\rho^2}$, $(1+\rho)/(1-\rho)$ and $(1-\rho)^2$ after Result 1.

Note that p^{approx} corresponds exactly to the posterior of a model where the prior on α and ρ is flat but, in addition to the sample, we have an observation for some notional date t'

$$y_{t'} = \alpha + \rho y_{t'-1} + v \tag{5}$$

where $v \sim N(0, \sigma_v^2)$ and we happened to observe on this date $y_{t'} = y_{t'-1} = y_0$.[11] Conversely, notice that plugging y_0 in (5) is quite similar to the model for the initial condition (2).

We show below that this approximate posterior, for most samples, will have a mean that is larger than the OLS estimate.

To illustrate the effect of adding this "data point" we show in Figure 5 a scatter plot of data points (y_t, y_{t-1}) for a sample of length 100 simulated from an AR(1) model. In this

[11] In his class-notes, Sims (2006) also points out the similarity between the distribution of the initial condition and a likelihood of a dummy observation for a notional date. Our argument interprets this as an approximation and Result 1 provides sufficient conditions for an upward adjustment of the posterior mean when adding this notional date.

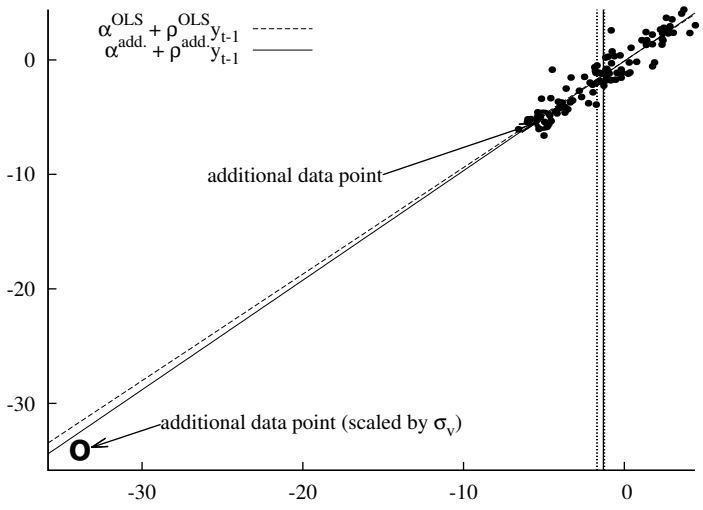


Figure 5 – Scatter plot of y_t against y_{t-1} for a sample of length 100 simulated from an AR(1) in which $\rho^{OLS} = 0.93$. The dashed line shows the OLS line fitted into this sample. The two circles show the "additional data point" implied by the model of the initial condition, the unscaled one and the one scaled by σ_v. The solid line shows the OLS line fitted into the sample that that incorporates this "additional data point" (the scaled one). α^{add} and ρ^{add} denote the coefficients of this line. The solid *vertical* line shows the value of $\alpha^{OLS}/(1-\rho^{OLS})$. The two dotted vertical lines (one of which partly overlaps with the solid vertical line) delimit the range of values of y_0 for which, given the rest of the sample, condition (7) would be violated.

sample $\rho^{OLS} = 0.93$. We also plot the "additional data point" representing the distribution of the initial condition, this point is (y_0, y_0) so it lies exactly on the 45° line. It is intuitive that adding a "data point" on the 45° line tends to tilt the regression line towards the 45° line. This intuition is even stronger if instead of plotting the original "additional data point" (which receives weight $\frac{\sigma^2/\sigma_v^2}{T+\sigma^2/\sigma_v^2}$ in the posterior (4)) we plot a point $(y_0\frac{\sigma}{\sigma_v}, y_0\frac{\sigma}{\sigma_v})$ which we label "additional data point scaled by σ_v". Adding this scaled point with the same weight as the other data points and computing the OLS estimate is an equivalent way of finding the mean of the posterior (4).[12]

The Figure illustrates that adding the scaled additional data point tilts the dashed regression line into the steeper solid line. This is because the estimate of ρ moves upwards relative to the original OLS estimate, while the absolute value of the posterior mean of α is lower. In the sample presented in Figure 5, the posterior mean of ρ is equal to 0.96 while, recall, we found $\rho^{OLS} = 0.93$.[13]

[12] In this data point the 1 corresponding to the constant term is also multiplied by $\frac{\sigma}{\sigma_v}$.

[13] We obtain the posterior mean of 0.96 assuming $\sigma = 1$ and $\sigma_v = 0.15$. This value of σ_v would be justified

Therefore, a Bayesian adjusts the OLS estimate upwards once she takes the distribution of the initial condition into account because the behavior of p^{approx} dominates the behavior of the posterior (3).

We now show an analytic result that gives a necessary and sufficient condition for the upward adjustment due to p^{approx}. Notice that the result is about the posterior mean conditional on the sample y, while the posterior shown in Figures 3, 4 conditions on less information, namely on the value ρ^{OLS}.

Result 1. *Assume p^{approx} is the actual posterior for all ρ. Assume the sample satisfies $\rho^{OLS} \neq 1$. Then we have*

$$E^{p^{approx}}(\rho|y) > \rho^{OLS} \tag{6}$$

if and only if[14]

$$\begin{aligned} &either\ y_0 < \min\left(\frac{\sum_{t=1}^{T} y_{t-1}}{T}, \frac{\alpha^{OLS}}{1-\rho^{OLS}}\right) \\ &or\ y_0 > \max\left(\frac{\sum_{t=1}^{T} y_{t-1}}{T}, \frac{\alpha^{OLS}}{1-\rho^{OLS}}\right). \end{aligned} \tag{7}$$

The proof is in the Appendix.

Condition (7) tends to hold in the common case when y grows more or less as a unit root process. In this case, precisely due to the small sample bias, it is likely that $\rho^{OLS} < 1$, and the initial condition is likely to be lower than both the sample mean and the estimated long run mean.

More generally, the condition tends to hold in the stationary case $\rho^{OLS} < 1$. To see this, note that in stationary datasets the OLS estimate of the steady-state $\frac{\alpha^{OLS}}{1-\rho^{OLS}}$, and the sample mean $\sum_{t=1}^{T} y_{t-1}/T$, are likely to be close to each other, then y_0 is unlikely to be in between $\frac{\alpha^{OLS}}{1-\rho^{OLS}}$ and $\sum_{t=1}^{T} y_{t-1}/T$ since this is a very narrow interval, hence condition (7) is likely to hold. In Figure 5 the range of values of y_0 for which condition (7) is violated is

by assuming $\rho = 0.96$ and noting that model (2) would imply $\sigma_v = \sqrt{(1-\rho)/(1+\rho)}$. Of course, here we treat σ_v as fixed and we use the value of 0.15 only for illustration.

[14] In the zero probability case that $\rho^{OLS} = 1$ the necessary and sufficient condition for the inequality to hold is $\left(y_0 - \frac{\sum_{t=1}^{T} y_{t-1}}{T}\right)\alpha^{OLS} > 0$.

delimited with vertical dotted lines, showing this range is indeed very narrow.[15]

The result also highlights that the upward adjustment does not occur for all the samples y. But since the samples that violate (7) are unlikely given the model for y_0 and $\rho^{OLS} < 1$ this explains why we find in Figures 2, 3 and 4 that $E(\rho|\rho^{OLS}) > \rho^{OLS}$ when ρ^{OLS} is close to, but below 1.

The above intuition and result work for the approximate posterior p^{approx}. Recall that the approximate posterior differs from the actual posterior because the terms $\sqrt{1-\rho^2}$, $(1+\rho)/(1-\rho)$ and $(1-\rho)^2$ in (3) are substituted by constants. Let us discuss how these approximations affect the results.

The actual posterior that includes these terms is non-symmetric, so that finding the mean analytically is difficult or impossible. For the $|\rho| < 1$ branch, on the one hand, $(1+\rho)/(1-\rho)$ increases in the vicinity of $\rho = 1$, which increases the weight of the "additional data point" in the posterior thus pushing the posterior mean of ρ more strongly towards 1. On the other hand, the term $\sqrt{1-\rho^2}$ downweighs the posterior in the vicinity of $\rho = 1$ hence pushing the posterior mean ρ away from 1. For the $|\rho| > 1$ branch the term $(1-\rho)^2$ also gives more weight to the additional observation. The combined effect on the posterior mean is difficult to judge. This is why we resorted to numerical integration and to summarizing the posteriors $p(\rho|y)$ into $p(\rho|\rho^{OLS})$ in Figures 2, 3 and 4. Our Figures 2, 3 and 4 prove that the adjustment is positive, suggesting that for ρ^{OLS} between, roughly, 0.7 and 1, the intuition provided by Result 1 dominates, since $E(\rho|\rho^{OLS}) > \rho^{OLS}$.[16]

4.2 Does a frequentist econometrician adjust ρ^{OLS} when the model of y_0 is unrelated to the parameters? (Question C)

Consider now the frequentist distribution of ρ^{OLS} when the initial condition is unrelated to parameters ρ, α, σ, modeled as $p(y_0|\rho,\alpha,\sigma) \propto 1$.[17] It is known, but rarely highlighted, that

[15] In the case $\rho^{OLS} > 1$ the condition (7) is likely to be violated and therefore the result implies that it is likely that $E(\rho|y) < \rho^{OLS}$. The condition is likely to be violated because in this case $\alpha^{OLS}/(1-\rho^{OLS})$ is the level from which the series appears to be diverging at an accelerating rate and thus typically y_0 will be between the sample mean and this level.

[16] The inequality might hold for lower values of ρ^{OLS} as well, but there the effect is quantitatively too small to be visible in our plots.

[17] $p(y_0|\rho,\alpha,\sigma) \propto 1$ is, to our minds, the only way to fully capture the idea that the initial condition is unrelated to parameters. Consider e.g. an alternative assumption: $y_0 = 0$. This looks like a natural assumption to a frequentist econometrician who considers a model with $\alpha = 0$, because for a stationary ρ

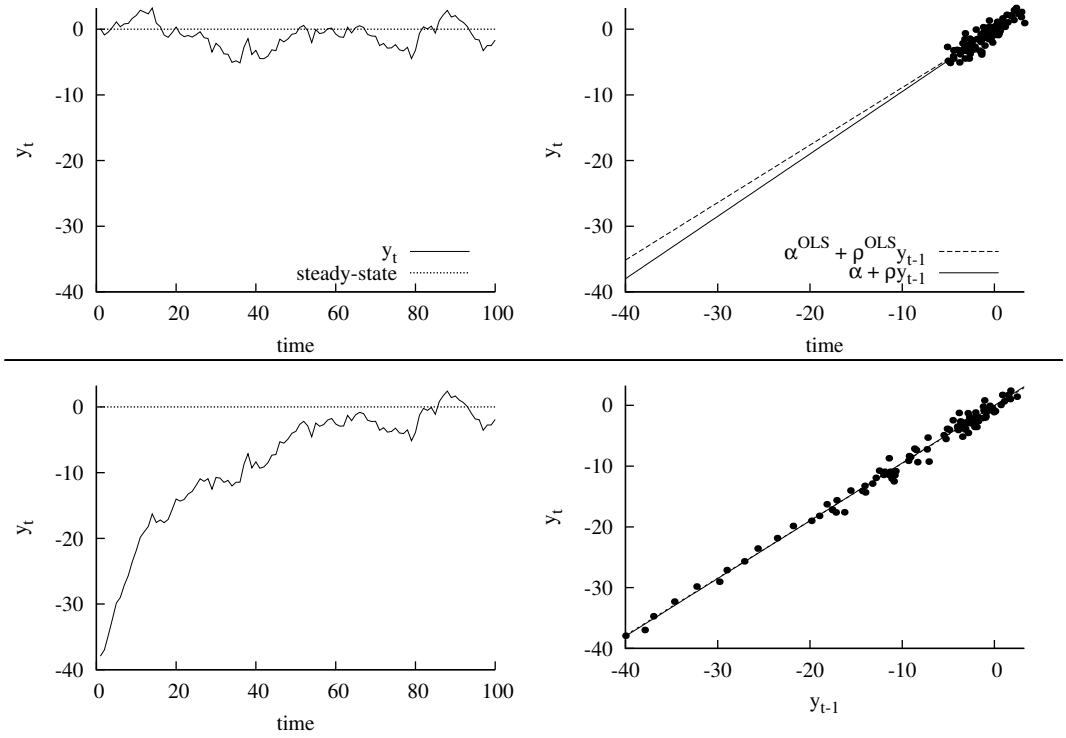


Figure 6 – Two realizations of $y_1, ..., y_T$ and the performance of the OLS estimator. $T = 100, \rho = 0.95$. The axes labeled y_t and y_{t-1} show the distance from the steady state $\alpha/(1-\rho)$ in terms of error standard deviations σ. For example, $y_t = -10$ means that y_t is ten error standard deviations below $\alpha/(1-\rho)$.

in this case the small sample bias vanishes. Analytic results to this effect can be found in Phillips (1987, section 6), and Phillips and Magdalinos (2009). Arellano (2003, p.86) and Chamberlain (2000) show this result for some special cases in the context of panel data. Therefore, we answer NO to question C.

Panel C of Figure 3 illustrates these analytical results. The density in panel C is obtained with $\kappa = 100$, which approximates the situation when y_0 is unrelated to the parameters. It is clear that the bias of ρ^{OLS} in this figure is very small and the small sample distribution of the OLS estimator is very tightly concentrated near the true value.

Intuition. Although in this case we have hard analytic results, it is useful to give some intuition about why the OLS bias shrinks as the distribution of y_0 becomes flat. In

the ergodic mean of the process is $\alpha/(1-\rho) = 0$. So $y_0 = 0$, although with no explicit dependence on α and ρ, actually introduces a dependence of y_0 on α and ρ through the back door. In general, any proper distribution of y_0 introduces some relation between the initial condition and the parameters through the back door.

a nutshell, a large $\text{var}(y_0)$ implies a larger sample variation in the regressors y_{t-1} and, as is well known, this larger variation increases the precision of the OLS estimator. To see this, compare the simulated sequence of $y_1, ..., y_T$ plotted in the top of Figure 6 with the simulated sequence of $y_1, ..., y_T$ plotted in the bottom of Figure 6.

The sequence in the top of Figure 6 starts near its steady state. This is the typical case when we assume $\kappa = 1$. The left graph plots y_t against time and the right graph shows the scatter plot of y_t against y_{t-1}. The values of y_{t-1} in the horizontal axis of the scatter plot are not very dispersed. As a result, when we fit a regression line by OLS into this collection of points we may incur a large error. The small sample bias typically makes this regression line flatter than the true regression line. Compare the dashed line, showing the regression line fitted by OLS with the solid line showing the true regression line: the dashed line is indeed flatter.

The sequence in the bottom of Figure 6 starts far from its steady state. This is the typical case when we assume $\kappa = 100$. The values of y_{t-1} in the horizontal axis are now very dispersed and the regression line fit by OLS into this collection of points is very close to the true regression line. This explains why the small sample bias is negligible when $\kappa = 100$.

5 Conclusion

We have reexamined the classical versus flat-prior-Bayesian controversy about the validity of the OLS estimator in autoregressions. We have shown in detail how the default Bayesian and frequentist analysis of an autoregression assume very different models of the initial condition. Then we have demonstrated how under a natural model of the initial condition the flat-prior-Bayesian posterior tends to suggest that the process is more persistent (has a higher ρ) than the OLS estimate. Finally, we showed how the commonly used "conditional likelihood" is tantamount to a model of the initial condition that implies large sample variation of the data and thus suggests that the OLS estimator is very good, also from the frequentist perspective. The summary of this discussion is found in Table 2.

We have shown that classical and Bayesian econometricians qualitatively agree about the virtues of OLS estimation when they model the initial condition in the same way. They

Table 2 – Does the OLS estimate need an adjustment? The complete picture

	Approach to statistics	
model of y_0	Frequentist	Bayesian
depends on α, ρ, σ	A: YES	B: YES
does not depend on α, ρ, σ	C: NO	D: NO

both conclude that if ρ^{OLS} is close to, but below 1, the OLS method underestimates the true value of ρ if a proper model for the initial condition is specified, while both of them conclude that the OLS method provides a good point estimate if they model y_0 as unrelated to the parameters that apply in the rest of the sample.

An applied econometrician could then use a certain model of the initial condition and reduce considerably the difference between the standard point estimates coming from the Bayesian and frequentist approaches. The use of a posterior or of confidence intervals still is a different way of summarizing the data in the light of the model, both of them possibly carrying useful information, but the two approaches do not give contradictory recommendations about whether or not OLS is an appropriate point estimate.[18]

Even though our claims are based on simulations displayed in Figures 2-3-4 they are not parameter dependent. However, our claims do depend on the fact that we used an AR(1) model with a constant. We do not know if our results survive in more general models, but it does show that the distribution of the initial condition should be taken into account.

Appendix

[18] Another option for an applied econometrician is to use the priors on observables as in Jarociński and Marcet (2010). That paper motivates using priors on observables as a better way to summarize information that experts actually have about the economy. The effect of their prior on growth rates on the posterior is similar, analytically, to the term introduced by the marginal distribution of the initial condition. Therefore, the prior on growth rates can be understood as another way to connect Bayesian and frequentist approaches.

A Generation of the joint density of ρ and ρ^{OLS} by Monte Carlo

In this appendix we explain how we adapted the Monte Carlo of Sims and Uhlig (1991) to the model in which a constant term α is present. We first describe the Monte Carlo and then argue that the joint density of ρ and ρ^{OLS} obtained with it corresponds to the flat prior on $\alpha, \rho, \ln \sigma$.

A.1 The Monte Carlo

To generate the sample of (ρ, ρ^{OLS}) from their joint distribution we proceed as follows. We use a grid of values of ρ from 0.7 to 1.2 at intervals of 0.01. For each value of ρ we generate 100,000 draws of y, where $y = (y_0, y_1, ..., y_{100})$ is a vector of 101 observations from model (1)-(2). To generate one draw of y we first draw y_0 from (2). Without loss of generality, we fix $\alpha = 0$ and $\sigma = 1$. We explain below why there is no loss of generality in fixing $\alpha = 0$ and $\sigma = 1$. Given y_0 we simulate $y_1, ..., y_{100}$ using (1). For each draw of y obtained this way we compute ρ^{OLS} by regressing vector $(y_1, ..., y_{100})$ on a constant term and vector $(y_0, ..., y_{99})$. Thus, for each value of ρ on the grid we have 100,000 draws of ρ^{OLS}. We assign the draws of ρ^{OLS} to bins $(-\infty, 0.695)$, $[0.695, 0.705)$, $[0.705, 0.715)$, etc. The histogram made from these bins for a given value $\rho = \bar{\rho}$ approximates the distribution of $\rho^{OLS}|\rho = \bar{\rho}$. (The left panel of Figure 1 is such a histogram for $\rho = 0.95$, smoothed and normalized.) The histograms lined up side by side form a surface that approximates the joint distribution of ρ and ρ^{OLS}.

A.2 Flat prior on ρ

A joint distribution of (ρ, ρ^{OLS}) generated as above corresponds to the flat prior for ρ. To see that the underlying prior about ρ is flat notice that the grid of values of ρ is uniform and we make an equal number of draws of ρ^{OLS} for each value of ρ.

A.3 Invariance to α and σ for $\rho \neq 1$

The joint distribution of (ρ, ρ^{OLS}) generated as above is invariant to the values of α and σ everywhere except for the measure-zero set $\rho = 1$. Therefore, outside this set it applies to

any prior on α and σ, including the flat prior on α and $\ln\sigma$. The invariance holds when the first observation comes from (2), and is an implication of the following result.

Result 2. *Assume the model parameterized as*

$$y_t - \mu = \rho\,(y_{t-1} - \mu) + u_t \qquad \textit{for } t = 1 \ldots T \tag{A.1}$$

and assume that the initial condition is given by:

$$y_0 = \mu + \sigma\psi \tag{A.2}$$

where ψ is a random variable. Then, if ψ is independent of the shocks u and its distribution is independent of μ and σ conditionally on ρ, the distribution of the OLS estimator of ρ in (1) is independent of μ and σ conditionally on ρ.

Proof. Define normalized errors: $v \equiv u/\sigma$. (A.2) allows to write:

$$y_t = \mu + \sigma\left(\sum_{i=1}^{t}\rho^{t-i}v_i + \rho^t\psi\right) = \mu + \sigma\tilde{y}_t \tag{A.3}$$

where $\tilde{y}$ is the process with $\mu = 0$, which would obtain from the same realization of errors, but rescaled to have a unit variance. Then it is a matter of simple algebra to show that:

$$\hat{\rho} \equiv \frac{T\sum y_t y_{t-1} - \sum y_{t-1}\sum y_t}{T\sum y_{t-1}^2 - \left(\sum y_{t-1}\right)^2} = \frac{T\sum \tilde{y}_t \tilde{y}_{t-1} - \sum \tilde{y}_{t-1}\sum \tilde{y}_t}{T\sum \tilde{y}_{t-1}^2 - \left(\sum \tilde{y}_{t-1}\right)^2}. \tag{A.4}$$

□

Similar results about invariance of ρ^{OLS} have been used in the literature. Andrews (1993, Appendix A), contains a verbal proof for $|\rho| \leq 1$ and for a particular distribution for ψ. DeJong et al. (1992) contains a similar proof for a fixed initial displacement $y_0 - \mu$. As can be seen, the proof is very simple, but we could not find a formal result focused on giving a general form of the initial condition which guarantees independence of the distribution of ρ^{OLS} from nuisance parameters, so we offer it here for completeness.

To map this result to our case notice that whenever $\rho \neq 1$ the model (A.1)-(A.2) is equivalent to the model (1)-(2), with the mapping $\alpha = (1-\rho)\mu$.

A.4 The singularity at $\rho = 1$

When $\rho = 1$ then, unlike for $\rho \neq 1$, the distribution of ρ^{OLS} does depend on α and σ. The flat prior about α and $\ln \sigma$, that we assume, is a limit of proper priors and the distribution of ρ^{OLS} depends on how we take this limit, since it depends on the relative size of α and σ. Therefore, the distribution in the measure-zero region defined by $\rho = 1$ is not uniquely pinned down.

When α is large relative to σ then the distribution of ρ^{OLS} collapses to a point mass at the true value, 1. This happens because large absolute values of α imply large sample variation in y_t relative to the error standard deviation σ. Note that when $\rho = 1$, α is the growth rate of y_t. Therefore, large α means that y_t changes fast from period to period, and the sample variation in y_t is large. With large sample variation the OLS estimator is very precise (recall Figure 6). When α is small relative to σ then ρ^{OLS} has a proper distribution.

B Proof of Result 1

As we have argued in the text the posterior mean is obtained by minimizing the sum of square residuals after adding the "additional data point." Simple algebra shows that adding the "additional data point" results in the following adjustment of the previous OLS estimate using only the sample:[19]

$$E^{p^{approx}}\left(\left[\begin{array}{c}\alpha\\ \rho\end{array}\right]\middle| y,\sigma\right) = \left[\begin{array}{c}\alpha^{OLS}\\ \rho^{OLS}\end{array}\right] + R_{T+1}^{-1}\frac{\sigma^2}{\sigma_v^2}\left[\begin{array}{c}1\\ y_0\end{array}\right](y_0 - \alpha^{OLS} - \rho^{OLS}y_0), \quad \text{(B.1)}$$

where the second moment matrices R are given by

$$R_T = \sum_{t=1}^{T}\left[\begin{array}{c}1\\ y_{t-1}\end{array}\right]\left[\begin{array}{c}1\\ y_{t-1}\end{array}\right]',$$

$$R_{T+1} = R_T + \frac{\sigma^2}{\sigma_v^2}\left[\begin{array}{c}1\\ y_0\end{array}\right]\left[\begin{array}{c}1\\ y_0\end{array}\right]'.$$

[19] This type of formula is routinely used in models of least squares learning, see for example equation (4a) in Marcet and Sargent (1989).

In other words, with the "additional data point" the estimate $[\alpha^{OLS}, \rho^{OLS}]$ is updated by a term where the regressors $[1, y_0]$ multiply the forecast error in the new data point and this is weighted by the inverse of the second moment matrix of the regressors R_{T+1}. To arrive at (B.1) note that

$$E^{p^{approx}}\left(\left[\begin{array}{c}\alpha\\ \rho\end{array}\right]\middle| y,\sigma\right) = R_{T+1}^{-1}\left(\sum_{t=1}^{T}\left[\begin{array}{c}1\\ y_{t-1}\end{array}\right]y_t + \frac{\sigma^2}{\sigma_v^2}\left[\begin{array}{c}1\\ y_0\end{array}\right]y_0\right) \tag{B.2}$$

$$= R_{T+1}^{-1}\left(R_{T+1} - \frac{\sigma^2}{\sigma_v^2}\left[\begin{array}{c}1\\ y_0\end{array}\right]\left[\begin{array}{c}1\\ y_0\end{array}\right]'\right)\left[\begin{array}{c}\alpha^{OLS}\\ \rho^{OLS}\end{array}\right] + \frac{\sigma^2}{\sigma_v^2}R_{T+1}^{-1}\left[\begin{array}{c}1\\ y_0\end{array}\right]y_0, \tag{B.3}$$

where the first equality follows from the claim that $E^{p^{approx}}(\alpha, \rho|y)$ is the OLS estimate with the "additional data point" and the second equality follows from simple algebra.

Applying the formula for the inverse matrix in (B.1) we get

$$E^{p^{approx}}(\rho \mid y,\sigma) - \rho^{OLS} =$$

$$= (\det R_{T+1})^{-1}\left[-\sum_{t=1}^{T} y_{t-1} - \frac{\sigma^2}{\sigma_v^2}y_0, T + \frac{\sigma^2}{\sigma_v^2}\right]\frac{\sigma^2}{\sigma_v^2}\left[\begin{array}{c}1\\ y_0\end{array}\right](y_0 - \alpha^{OLS} - \rho^{OLS}y_0) \tag{B.4}$$

$$= (\det R_{T+1})^{-1}\left[\left(T + \frac{\sigma^2}{\sigma_v^2}\right)y_0 - \left(\sum_{t=1}^{T} y_{t-1} + \frac{\sigma^2}{\sigma_v^2}y_0\right)\right]\frac{\sigma^2}{\sigma_v^2}(y_0 - \alpha^{OLS} - \rho^{OLS}y_0). \tag{B.5}$$

The last line is positive if and only if (7) holds hence $E^{p^{approx}}(\rho \mid y,\sigma) > \rho^{OLS}$.

Conditioning both sides of the last inequality on the sample y and applying the law of iterated expectations gives the result. □

References

Abadir, K. M., Hadri, K., and Tzavalis, E. (1999). The influence of VAR dimensions on estimator biases. *Econometrica*, 67(1):163–181.

Andrews, D. W. K. (1993). Exactly median-unbiased estimation of first order autoregressive / unit root models. *Econometrica*, 61(1):139–165.

Arellano, M. (2003). *Panel Data Econometrics*. Oxford University Press, first edition.

Bauer, M. D., Rudebusch, G. D., and Wu, J. C. (2012). Correcting estimation bias in dynamic term structure models. *Journal of Business & Economic Statistics*, 30(3):454–467.

Bhargava, A. (1986). On the theory of testing for unit roots in observed time-series. *Review of Economic Studies*, 53(3):369–384.

Chamberlain, G. (2000). Econometrics and decision theory. *Journal of Econometrics*, 95(2):255 – 283.

DeJong, D. N., Nankervis, J. C., Savin, N. E., and Whiteman, C. H. (1992). Integration versus trend stationary in time series. *Econometrica*, 60(2):423–433.

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, 80(1):27–38.

Hurwicz, L. (1950). Least-squares bias in time series. In Koopmans, T. C., editor, *Statistical Inference in Dynamic Economic Models*. Wiley, New York.

Jarociński, M. and Marcet, A. (2010). Autoregressions in small samples, priors about observables and initial conditions. Working Paper 1263, European Central Bank.

Kwan, Y. K. (1998). Asymptotic bayesian analysis based on a limited information estimator. *Journal of Econometrics*, 88(1):99–121.

Litterman, R. B. (1986). Forecasting with Bayesian vector autoregressions - five years of experience. *Journal of Business and Economic Statistics*, (4):25–38.

Lubrano, M. (1995). Testing for unit roots in a Bayesian framework. *Journal of Econometrics*, 69:81–109.

MacKinnon, J. G. and Smith, A. A. (1998). Approximate bias correction in econometrics. *Journal of Econometrics*, 85:205–230.

Marcet, A. and Sargent, T. J. (1989). Convergence of least squares learning mechanisms in self-referential linear stochastic models. *Journal of Economic Theory*, 48(2):337–368.

Marriott, F. H. C. and Pope, J. A. (1954). Bias in the estimation of autocorrelations. *Biometrika*, XLI:393–403.

Phillips, P. C. B. (1987). Time series regression with a unit root. *Econometrica*, 55(2):277–301.

Phillips, P. C. B. (1991). To criticize the critics: An objective bayesian analysis of stochastic trends. *Journal of Applied Econometrics*, 6(4):333–364.

Phillips, P. C. B. and Magdalinos, T. (2009). Unit root and cointegrating limit theory when initialization is in the infinite past. *Econometric Theory*, 25(06):1682–1715.

Quenouille, M. H. (1949). Approximate tests of correlation in time-series. *Journal of the Royal Statistical Society Series B*, 11:68–84.

Schotman, P. C. and Van Dijk, H. K. (1991). On Bayesian routes to unit roots. *Journal of Applied Econometrics*, 6(4):387–401.

Sims, C. A. (2006). Conjugate dummy observation priors for VAR's. Technical report, Princeton University.

Sims, C. A. and Uhlig, H. (1991). Understanding unit rooters: A helicopter tour. *Econometrica*, 59(6):1591–1599.

Sims, C. A. and Zha, T. (1998). Bayesian methods for dynamic multivariate models. *International Economic Review*, 39(4):949–68.

Uhlig, H. (1994). On Jeffreys prior when using the exact likelihood function. *Econometric Theory*, 10(3-4):633–644.

Zellner, A. (1971). *An Introduction to Bayesian Inference in Econometrics*. Wiley, New York.