ANNALS OF ECONOMICS AND FINANCE **16-1**, 163–171 (2015)

Should We Demean the Data?

Yong Bao

Department of Economics, Purdue University, IN 47907, USA
E-mail: ybao@purdue.edu

The sample average is an unbiased estimator of the population mean, so it may seem innocuous that for estimating model parameters that do not involve the population mean, the data can be demeaned first. Using a first-order moving average (MA) model for example, we derive the analytical approximate biases of the quasi maximum likelihood estimators (QMLEs) based on the original and demeaned data. The bias results indicate that the QMLEs can behave quite differently in finite samples and it is not always advisable to demean the data if the MA parameter is of primary interest to estimate.

Key Words: Demean; Moving Average; Bias.
JEL Classification Numbers: C22, C51.

1. INTRODUCTION

For linear models, it is well known that we can first subtract the sample average from the data without affecting estimation of the non-intercept parameters. It is less obvious for nonlinear models.[1] Intuitively, since the sample average is a very good estimator of the population mean, demeaning a univariate time series seems least harmful for estimating other model parameters that do not involve the population mean. In this note, we use the moving average (MA) model of order 1 (MA(1)) as an example to demonstrate that it is not always advisable to take such a view. The MA model is intrinsically nonlinear. Compared with the linear autoregressive model, it can be used to model and forecast economic variables of less persistence and shorter memory. A prominent example is from Stock and Watson (2007), who found that the simple MA(1) model works really well in describing the inflation rate change for the US economy.

[1]When we say a model is nonlinear, we mean that the model parameters are estimated nonlinearly.

1529-7373/2015

Let $\boldsymbol{\beta} = (\mu, \theta, \sigma^2)'$ be the vector of model parameters of the MA(1) model $y_t = \mu + \varepsilon_t + \theta\varepsilon_{t-1}$, where $|\theta| < 1$ and $\varepsilon_t \sim$ i.i.d. $(0, \sigma^2)$, not necessarily normal. Let $\boldsymbol{\beta}_0 = (\mu_0, \theta_0, \sigma_0^2)'$ denote the population parameter vector and $\boldsymbol{y} = (y_1, \cdots, y_T)'$ be the sample observations. Typically, the method of quasi maximum likelihood (QML) is used to estimate the parameters by maximizing a Gaussian sample likelihood function of $\boldsymbol{y}$ even though the data might be nonnormally distributed, see Hamilton (1994). Nevertheless, the sample mean $\bar{y}_T = T^{-1}\sum_{t=1}^{T} y_t$ is an unbiased estimator of the population mean μ_0, so it is quite often that applied economists demean the data first and then estimate $\boldsymbol{\delta} = (\theta, \sigma^2)'$ from $\boldsymbol{x} = (x_1, \cdots, x_T)'$ with $x_t = y_t - \bar{y}_T$. From a theoretical point of view, this amounts to using the combination of two estimation methods, namely, moment estimation (for estimating μ) and quasi maximum likelihood (for estimating $\boldsymbol{\delta}$). Two intriguing questions arise. First, how different is the QML estimator (QMLE) of μ based on $\boldsymbol{y}$ from the simple estimator $\bar{y}_T$? If the true distribution of ε_t is normal, the QMLE becomes the maximum likelihood estimator and it is most efficient, though can still be biased in finite samples, whereas $\bar{y}_T$ is unbiased and asymptotically efficient. Second, will demeaning affect the estimation of (θ, σ^2)? In a typical ordinary least squares framework with cross-section data, demeaning is innocuous, so is with time-series autoregressive models. But for nonlinear time-series models, the answer is far less obvious. We aim to answer the two questions in this note.

Throughout, $\boldsymbol{\iota}$ is a vector of ones, $\boldsymbol{I}$ is the identity matrix, $\boldsymbol{M} = \boldsymbol{I} - T^{-1}\boldsymbol{\iota}\boldsymbol{\iota}'$, and $\boldsymbol{0}$ is a null vector. The dimensions of vectors/matrices are to be read from the context, and thus we suppress the dimension subscripts in our notation.

2. MAIN RESULTS

Conditional on $\varepsilon_0 = 0$, the average Gaussian log likelihood function of the observable data $\boldsymbol{y}$ is

$$L(\boldsymbol{\beta}; \boldsymbol{y}|\varepsilon_0 = 0) = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\sigma^2) - \frac{\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}}{2T\sigma^2}, \tag{1}$$

where $\boldsymbol{\varepsilon} = (\varepsilon_1, \cdots, \varepsilon_T)'$ and ε_t is defined recursively from $\varepsilon_t = y_t - \mu - \theta\varepsilon_{t-1}$ starting with $\varepsilon_0 = 0$. Let $\hat{\boldsymbol{\beta}}_T$ be the QMLE of $\boldsymbol{\beta}_0$ based on the Gaussian likelihood function (1).

Now suppose we work with the demeaned data. By definition, the model should be

$$x_t = u_t + \theta u_{t-1}, \tag{2}$$

where $u_t = \varepsilon_t - \bar{\varepsilon}_T = \varepsilon_t - T^{-1}\sum_{t=1}^{T}\varepsilon_t$. Obviously, u_t is no longer i.i.d. and its variance is $\sigma^2(1 - T^{-1})$. Moreover, the conditional mean of u_t is

not θu_{t-1} and $u_0 \neq 0$. This stands in contrast to the MA(1) model using the original data. Another way to think about this is to write

$$x_t = \mu - \bar{y}_T + \varepsilon_t + \theta\varepsilon_{t-1}, \tag{3}$$

whose constant term $\mu - \bar{y}_T \neq 0$, though $\mathbb{E}(\mu - \bar{y}_T) = 0$. In other words, when using the demeaned data to estimate the model, we are in fact imposing $\mu - \bar{y}_T$ to be zero in the sample.

Let $\tilde{\boldsymbol{\beta}}_T = (\bar{y}_T, \tilde{\theta}_T, \tilde{\sigma}_T^2)'$ denote the estimated parameter vector based on the demeaned data, where $(\tilde{\theta}_T, \tilde{\sigma}_T^2)' \equiv \tilde{\boldsymbol{\delta}}_T$ is estimated from (2) based on the log likelihood function

$$L(\theta, \sigma^2; \boldsymbol{x}|u_0 = 0) = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\sigma^2) - \frac{\boldsymbol{u}'\boldsymbol{u}}{2T\sigma^2}, \tag{4}$$

where $\boldsymbol{u} = (u_1, \cdots, u_T)'$ and u_t is defined recursively from $u_t = x_t - \theta u_{t-1}$ starting with $u_0 = 0$.

Using matrix notation, we can write $\boldsymbol{y} = \mu\boldsymbol{\iota} + \boldsymbol{C}\boldsymbol{\varepsilon}$, $\boldsymbol{x} = \boldsymbol{M}\boldsymbol{y} = \boldsymbol{M}\boldsymbol{C}\boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} = \boldsymbol{C}^{-1}(\boldsymbol{y} - \mu\boldsymbol{\iota})$, $\boldsymbol{u} = \boldsymbol{C}^{-1}\boldsymbol{x}$, where $\boldsymbol{C}$ is a $T \times T$ tridiagonal matrix with main diagonal elements 1, super-diagonal elements 0, and sub-diagonal elements θ. Then the score function associated with (1) is

$$\boldsymbol{\psi}_{\boldsymbol{\beta}} = \left(\frac{\boldsymbol{\varepsilon}'\boldsymbol{C}^{-1}\boldsymbol{\iota}}{T\sigma^2}, \quad \frac{\boldsymbol{\varepsilon}'\boldsymbol{A}_1\boldsymbol{\varepsilon}}{T\sigma^2}, \quad \frac{\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}}{2T\sigma^4} - \frac{1}{2\sigma^2}\right)', \tag{5}$$

where $\boldsymbol{A}_1 = \boldsymbol{C}^{-1}\boldsymbol{B}$, and the first-order condition associated with (4) is

$$\boldsymbol{\psi}_{\boldsymbol{\delta}} = \left(\frac{\boldsymbol{\varepsilon}'\boldsymbol{D}'\boldsymbol{A}_1\boldsymbol{D}\boldsymbol{\varepsilon}}{T\sigma^2}, \quad \frac{\boldsymbol{\varepsilon}'\boldsymbol{D}'\boldsymbol{D}\boldsymbol{\varepsilon}}{2T\sigma^4} - \frac{1}{2\sigma^2}\right)', \tag{6}$$

where $\boldsymbol{D} = \boldsymbol{C}^{-1}\boldsymbol{M}\boldsymbol{C}$.

From (5), we note that, if the value of θ is known,

$$\hat{\mu}_T = \frac{\boldsymbol{y}'\boldsymbol{C}^{-1\prime}\boldsymbol{C}^{-1}\boldsymbol{\iota}}{\boldsymbol{\iota}'\boldsymbol{C}^{-1\prime}\boldsymbol{C}^{-1}\boldsymbol{\iota}}, \quad \hat{\sigma}_T^2 = \frac{(\boldsymbol{y} - \frac{\boldsymbol{y}'\boldsymbol{C}^{-1\prime}\boldsymbol{C}^{-1}\boldsymbol{\iota}}{\boldsymbol{\iota}'\boldsymbol{C}^{-1\prime}\boldsymbol{C}^{-1}\boldsymbol{\iota}}\boldsymbol{\iota})'\boldsymbol{C}^{-1\prime}\boldsymbol{C}^{-1}(\boldsymbol{y} - \frac{\boldsymbol{y}'\boldsymbol{C}^{-1\prime}\boldsymbol{C}^{-1}\boldsymbol{\iota}}{\boldsymbol{\iota}'\boldsymbol{C}^{-1\prime}\boldsymbol{C}^{-1}\boldsymbol{\iota}}\boldsymbol{\iota})}{T}.$$

This stands in contrast to the demeaned model:

$$\tilde{\mu}_T = \frac{\boldsymbol{y}'\boldsymbol{\iota}}{\boldsymbol{\iota}'\boldsymbol{\iota}}, \quad \tilde{\sigma}_T^2 = \frac{(\boldsymbol{y} - \frac{\boldsymbol{y}'\boldsymbol{\iota}}{\boldsymbol{\iota}'\boldsymbol{\iota}}\boldsymbol{\iota})'\boldsymbol{M}\boldsymbol{C}^{-1\prime}\boldsymbol{C}^{-1}\boldsymbol{M}(\boldsymbol{y} - \frac{\boldsymbol{y}'\boldsymbol{\iota}}{\boldsymbol{\iota}'\boldsymbol{\iota}}\boldsymbol{\iota})}{T}.$$

So numerically, the original and demeaned models can give different values of parameter estimates.

One can easily check that $\mathbb{E}(\boldsymbol{\psi}_{\boldsymbol{\beta}}) = \mathbf{0}$, regardless of the distribution of ε. On the other hand $\mathbb{E}(\boldsymbol{\psi}_{\boldsymbol{\delta}}) \neq \mathbf{0}$; instead, $\lim_{T\to\infty} \mathbb{E}(\boldsymbol{\psi}_{\boldsymbol{\delta}}) = \mathbf{0}$. This implies in finite samples, $\hat{\boldsymbol{\beta}}_T$ and $\tilde{\boldsymbol{\beta}}_T$ can behave quite differently. (And we have already shown that numerically they are always different.) This can be seen more clearly if we follow Bao and Ullah (2007) to implement a stochastic expansion: $\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}_0 = \boldsymbol{\beta}_{-1/2} + \boldsymbol{\beta}_{-1} + o_P(T^{-1})$, where $\boldsymbol{\beta}_{-1/2} = O_P(T^{-1/2}) = \boldsymbol{\Sigma}\boldsymbol{\psi}$, $\boldsymbol{\beta}_{-1} = O_P(T^{-1}) = \boldsymbol{\Sigma}\boldsymbol{V}_1\boldsymbol{\beta}_{-1/2} + \frac{1}{2}\boldsymbol{\Sigma}\mathbb{E}(\boldsymbol{H}_2)(\boldsymbol{\beta}_{-1/2} \otimes \boldsymbol{\beta}_{-1/2})$, $\boldsymbol{H}_i = \bigtriangledown^i\boldsymbol{\psi}$, $\boldsymbol{\Sigma} = -[\mathbb{E}(\boldsymbol{H}_1)]^{-1}$ and $\boldsymbol{V}_1 = \boldsymbol{H}_1 - \mathbb{E}(\boldsymbol{H}_1)$, with all the terms evaluated at the true parameter vector. A similar expansion for $\tilde{\boldsymbol{\delta}}_T$ is $\tilde{\boldsymbol{\delta}}_T - \boldsymbol{\delta}_0 = \boldsymbol{\delta}_{-1/2} + \boldsymbol{\delta}_{-1} + o_P(T^{-1})$. As $\mathbb{E}(\boldsymbol{\psi}_\beta) = \mathbf{0}$, we can see $\mathbb{E}(\boldsymbol{\beta}_{-1/2}) = \mathbf{0}$ and thus the bias of $\hat{\boldsymbol{\beta}}_T$, up to order $O(T^{-1})$, can be approximated by $\mathbb{E}(\boldsymbol{\beta}_{-1})$. But for $\tilde{\boldsymbol{\delta}}_T$**,** we can check that $\mathbb{E}(\boldsymbol{\delta}_{-1/2}) \neq \mathbf{0}$, so the approximate bias needs to be defined as $\mathbb{E}(\boldsymbol{\delta}_{-1/2} + \boldsymbol{\delta}_{-1})$. When ε is normal, the bias of $\hat{\boldsymbol{\beta}}_T$ was derived by Tanaka (1984) and Cordeiro and Klein (1994):

$$\mathbb{E}(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}_0) = \frac{1}{T}\begin{pmatrix} 0 \\ -1 + 2\theta \\ -2\sigma^2 \end{pmatrix} + o(T^{-1}). \tag{7}$$

One can show that the above bias formula is still valid even when ε is nonnormal (see the appendix). On the other hand, the bias of $\tilde{\boldsymbol{\beta}}_T$, with $\bar{y}_T$ being unbiased and the bias of $\tilde{\boldsymbol{\delta}}_T$ given by $\mathbb{E}(\boldsymbol{\delta}_{-1/2} + \boldsymbol{\delta}_{-1})$, is

$$\mathbb{E}(\tilde{\boldsymbol{\beta}}_T - \boldsymbol{\beta}_0) = \frac{1}{T}\begin{pmatrix} 0 \\ \theta \\ -\sigma^2 \end{pmatrix} + o(T^{-1}). \tag{8}$$

The derivation of (8) is given in the appendix.

Comparing (7) and (8), we have several interesting observations regarding $\hat{\boldsymbol{\beta}}_T$ and $\tilde{\boldsymbol{\beta}}_T$.[2] First, both expressions indicate that the bias results from the original and demeaned data are robust, up to the order of approximation, to the distribution of the error term. They hold under both normal and nonnormal distributions. Second, the mean estimator $\hat{\mu}_T$ from the original data is unbiased, up to the order of approximation, whereas $\tilde{\mu}_T(= \bar{y})$ is always unbiased. Third, when θ is positive, the demeaned model tends to overestimate the moving average parameter, but the original model can overestimate (when $1 > \theta > 0.5$) or underestimate (when $0 < \theta < 0.5$). We also see the demeaned model may overestimates θ more than the original model when $1 > \theta > 0.5$. For negative θ, both models tend to underestimate, but the magnitude of the the bias is more severe for the original

[2]Simulations results, not reported here, largely support the described patterns of $\hat{\boldsymbol{\beta}}_T$ and $\tilde{\boldsymbol{\beta}}_T$ in finite samples.

model than the demeaned model. Lastly, both models tend to underestimate the variance, and the approximate downward bias from the original model is two times that from the demeaned model.

3. CONCLUDING REMARKS

We have used the MA(1) model to demonstrate the effects of demeaning the data on the estimation of model parameters in finite samples. By working out the approximate bias results, we see that the QMLEs can behave quite differently in finite samples. In particular, for the MA parameter θ, when $1 > \theta > 0.5$, the degree of overestimation from the demeaned data may be more than that from the original data in finite samples. Thus when the MA parameter is of direct interest, it is not advisable for us to demean the data if its magnitude is moderately large.[3]

APPENDIX: DERIVATION OF BIASES OF $\hat{\boldsymbol{\beta}}_T$ AND $\tilde{\boldsymbol{\delta}}_T$

For notational convenience, define $\boldsymbol{a} = \boldsymbol{C}^{-1}\boldsymbol{\iota}$, $\boldsymbol{A}_1^* = \boldsymbol{A}_1 + \boldsymbol{A}_1'$, $\boldsymbol{A}_2 = 2\boldsymbol{A}_1^2 + \boldsymbol{A}_1'\boldsymbol{A}_1$, $\boldsymbol{A}_3 = \boldsymbol{A}_1^3 + \boldsymbol{A}_1'\boldsymbol{A}_1^2$, $\boldsymbol{A}_4 = \boldsymbol{A}_1^3 + \boldsymbol{A}_1'\boldsymbol{A}_1^2$, and denote $\boldsymbol{D}_0 = \boldsymbol{D}'\boldsymbol{D}$, $\boldsymbol{D}_i = \boldsymbol{D}'\boldsymbol{A}_i\boldsymbol{D}$, $\lambda_j = \boldsymbol{\varepsilon}'\boldsymbol{D}_j\boldsymbol{\varepsilon}$, $d_i = T^{-1}\mathrm{tr}(\boldsymbol{D}_i)$, and $d = (2d_0d_2 - 2d_1^2 - d_2)/(2\sigma_0^4)$. Note that $\mathbb{E}(\lambda_i) = T\sigma_0^2 d_i$.

First, for the stochastic expansion of $\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}_0$, we take derivatives of (5) and have

$$\boldsymbol{H}_1 = \begin{pmatrix} -\frac{\boldsymbol{a}'\boldsymbol{a}}{T\sigma^2} & -\frac{\boldsymbol{a}'\boldsymbol{A}_1^*\boldsymbol{\varepsilon}}{T\sigma^2} & -\frac{\boldsymbol{a}'\boldsymbol{\varepsilon}}{T\sigma^4} \\ -\frac{\boldsymbol{a}'\boldsymbol{A}_1^*\boldsymbol{\varepsilon}}{T\sigma^2} & -\frac{\boldsymbol{\varepsilon}'\boldsymbol{A}_2\boldsymbol{\varepsilon}}{T\sigma^2} & -\frac{\boldsymbol{\varepsilon}'\boldsymbol{A}_1\boldsymbol{\varepsilon}}{T\sigma^4} \\ -\frac{\boldsymbol{a}'\boldsymbol{\varepsilon}}{T\sigma^4} & -\frac{\boldsymbol{\varepsilon}'\boldsymbol{A}_1\boldsymbol{\varepsilon}}{T\sigma^4} & -\frac{\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}}{T\sigma^6} + \frac{1}{2\sigma^4} \end{pmatrix},$$

$$\boldsymbol{H}_2 = \begin{pmatrix} 0 & \frac{2\boldsymbol{a}'\boldsymbol{A}_1\boldsymbol{a}}{T\sigma^2} & \frac{\boldsymbol{a}'\boldsymbol{a}}{T\sigma^4} & \frac{2\boldsymbol{a}'\boldsymbol{A}_1\boldsymbol{a}}{T\sigma^2} & \frac{2\boldsymbol{a}'\boldsymbol{A}_3\boldsymbol{\varepsilon}}{T\sigma^2} & \frac{\boldsymbol{a}'\boldsymbol{A}_1^*\boldsymbol{\varepsilon}}{T\sigma^4} & \frac{\boldsymbol{a}'\boldsymbol{a}}{T\sigma^4} & \frac{\boldsymbol{a}'\boldsymbol{A}_1^*\boldsymbol{\varepsilon}}{T\sigma^4} & \frac{2\boldsymbol{a}'\boldsymbol{\varepsilon}}{T\sigma^6} \\ \frac{2\boldsymbol{a}'\boldsymbol{A}_1\boldsymbol{a}}{T\sigma^2} & \frac{2\boldsymbol{a}'\boldsymbol{A}_3\boldsymbol{\varepsilon}}{T\sigma^2} & \frac{\boldsymbol{a}'\boldsymbol{A}_1^*\boldsymbol{\varepsilon}}{T\sigma^4} & \frac{2\boldsymbol{a}'\boldsymbol{A}_3\boldsymbol{\varepsilon}}{T\sigma^2} & \frac{6\boldsymbol{\varepsilon}'\boldsymbol{A}_4\boldsymbol{\varepsilon}}{T\sigma^2} & \frac{\boldsymbol{\varepsilon}'\boldsymbol{A}_2\boldsymbol{\varepsilon}}{T\sigma^4} & \frac{\boldsymbol{a}'\boldsymbol{A}_1^*\boldsymbol{\varepsilon}}{T\sigma^4} & \frac{\boldsymbol{\varepsilon}'\boldsymbol{A}_2\boldsymbol{\varepsilon}}{T\sigma^4} & \frac{2\boldsymbol{\varepsilon}'\boldsymbol{A}_1\boldsymbol{\varepsilon}}{T\sigma^6} \\ \frac{\boldsymbol{a}'\boldsymbol{a}}{T\sigma^4} & \frac{\boldsymbol{a}'\boldsymbol{A}_1^*\boldsymbol{\varepsilon}}{T\sigma^4} & \frac{2\boldsymbol{a}'\boldsymbol{\varepsilon}}{T\sigma^6} & \frac{\boldsymbol{a}'\boldsymbol{A}_1^*\boldsymbol{\varepsilon}}{T\sigma^4} & \frac{\boldsymbol{\varepsilon}'\boldsymbol{A}_2\boldsymbol{\varepsilon}}{T\sigma^4} & \frac{2\boldsymbol{\varepsilon}'\boldsymbol{A}_1\boldsymbol{\varepsilon}}{T\sigma^6} & \frac{2\boldsymbol{a}'\boldsymbol{\varepsilon}}{T\sigma^6} & \frac{2\boldsymbol{\varepsilon}'\boldsymbol{A}_1\boldsymbol{\varepsilon}}{T\sigma^6} & \frac{3\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}}{T\sigma^8} - \frac{1}{\sigma^6} \end{pmatrix}.$$

[3] In Stock and Watson (2007), where the simple MA(1) is used in describing the inflation rate change for the post-1984 US economy, the estimated MA parameter is above 0.5.

By substitution, the stochastic expansion of $\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}_0$ has

$$\boldsymbol{\beta}_{-1/2} = \begin{pmatrix} \frac{\boldsymbol{a}'\boldsymbol{\varepsilon}}{\boldsymbol{a}'\boldsymbol{a}} \\ \frac{\boldsymbol{\varepsilon}'\boldsymbol{A}_1\boldsymbol{\varepsilon}}{\sigma^2\mathrm{tr}(\boldsymbol{A}_2)} \\ \frac{\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}}{T} - \sigma^2 \end{pmatrix},$$

$$\boldsymbol{\beta}_{-1} = \begin{pmatrix} -\frac{\boldsymbol{a}'\boldsymbol{A}_1^*\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\boldsymbol{A}_1\boldsymbol{\varepsilon}}{\sigma^2\mathrm{tr}(\boldsymbol{A}_2)\boldsymbol{a}'\boldsymbol{a}} + \frac{2\boldsymbol{a}'\boldsymbol{A}_1\boldsymbol{a}\boldsymbol{a}'\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\boldsymbol{A}_1\boldsymbol{\varepsilon}}{\sigma^2\mathrm{tr}(\boldsymbol{A}_2)(\boldsymbol{a}'\boldsymbol{a})^2} \\ \frac{3\mathrm{tr}(\boldsymbol{A}_1^3+\boldsymbol{A}_1'\boldsymbol{A}_1^2)(\boldsymbol{\varepsilon}'\boldsymbol{A}_1\boldsymbol{\varepsilon})^2}{\sigma^4\mathrm{tr}^3(\boldsymbol{A}_2)} - \frac{\boldsymbol{\varepsilon}'\boldsymbol{A}_1\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\boldsymbol{A}_2\boldsymbol{\varepsilon}}{\sigma^4\mathrm{tr}^2(\boldsymbol{A}_2)} - \frac{\boldsymbol{\varepsilon}'\boldsymbol{a}\boldsymbol{a}'\boldsymbol{A}_1^*\boldsymbol{\varepsilon}}{\sigma^2\mathrm{tr}(\boldsymbol{A}_2)\boldsymbol{a}'\boldsymbol{a}} + \frac{\boldsymbol{a}'\boldsymbol{A}_1\boldsymbol{a}\boldsymbol{\varepsilon}'\boldsymbol{a}\boldsymbol{a}'\boldsymbol{\varepsilon}}{\sigma^2\mathrm{tr}(\boldsymbol{A}_2)(\boldsymbol{a}'\boldsymbol{a})^2} + \frac{\boldsymbol{\varepsilon}'\boldsymbol{A}_1\boldsymbol{\varepsilon}}{\sigma^2\mathrm{tr}(\boldsymbol{A}_2)} \\ -\frac{\boldsymbol{\varepsilon}'\boldsymbol{a}\boldsymbol{a}'\boldsymbol{\varepsilon}}{T\boldsymbol{a}'\boldsymbol{a}} - \frac{(\boldsymbol{\varepsilon}'\boldsymbol{A}_1\boldsymbol{\varepsilon})^2}{T\sigma^2\mathrm{tr}(\boldsymbol{A}_2)} \end{pmatrix}.$$

Note that $\boldsymbol{A}_1$ is strictly lower triangular. Then $\mathbb{E}(\boldsymbol{\varepsilon}'\boldsymbol{A}_1\boldsymbol{\varepsilon}) = \sigma^2\mathrm{tr}(\boldsymbol{A}_1) = 0$, so $\mathbb{E}(\boldsymbol{\beta}_{-1/2}) = \boldsymbol{0}$ and the second-order bias of $\hat{\boldsymbol{\beta}}_T$ is given by $\mathbb{E}(\boldsymbol{\beta}_{-1})$. The first element of $\mathbb{E}(\boldsymbol{\beta}_{-1})$ corresponds to

$$\begin{aligned} \mathbb{E}(\hat{\mu} - \mu) &= -\frac{\boldsymbol{a}'\boldsymbol{A}_1^*\mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\boldsymbol{A}_1\boldsymbol{\varepsilon})}{\sigma^2\mathrm{tr}(\boldsymbol{A}_2)\boldsymbol{a}'\boldsymbol{a}} + \frac{2\boldsymbol{a}'\boldsymbol{A}_1\boldsymbol{a}\boldsymbol{a}'\mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\boldsymbol{A}_1\boldsymbol{\varepsilon})}{\sigma^2\mathrm{tr}(\boldsymbol{A}_2)(\boldsymbol{a}'\boldsymbol{a})^2} + o(T^{-1}) \\ &= 0 + o(T^{-1}), \end{aligned}$$

since $\boldsymbol{v}'\mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\boldsymbol{A}_1\boldsymbol{\varepsilon}) = \mathbb{E}(\varepsilon_i^3)\boldsymbol{v}'\mathrm{diag}(\boldsymbol{A}_1) = 0$ for any vector $\boldsymbol{v}$.

For evaluating the biases of $\hat{\theta}$ and $\hat{\sigma}^2$, we need expectations of second-order quadratic forms in $\boldsymbol{\varepsilon}$. From Ullah (2004, p. 187), for any matrices $\boldsymbol{N}_1$ and $\boldsymbol{N}_2$, $\mathbb{E}(\boldsymbol{\varepsilon}'\boldsymbol{N}_1\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\boldsymbol{N}_2\boldsymbol{\varepsilon}) = \sigma^4[\gamma_2\mathrm{tr}(\boldsymbol{N}_1\odot\boldsymbol{N}_2)+\mathrm{tr}(\boldsymbol{N}_1)\mathrm{tr}(\boldsymbol{N}_2)+\mathrm{tr}(\boldsymbol{N}_1\boldsymbol{N}_2) +\mathrm{tr}(\boldsymbol{N}_1'\boldsymbol{N}_2)]$, where $\odot$ is the Hadamard (element by element) product operator and γ_2 is the excess kurtosis coefficient of the distribution of ε_t. Since $\boldsymbol{A}_1$ is strictly lower triangular, $\mathrm{tr}(\boldsymbol{A}_1) = \mathrm{tr}(\boldsymbol{A}_1\boldsymbol{A}_1) = \mathrm{tr}(\boldsymbol{A}_1\odot\boldsymbol{A}_1) = \mathrm{tr}(\boldsymbol{A}_1\odot\boldsymbol{A}_2) = 0$. This leads to

$$\begin{aligned} \mathbb{E}(\hat{\theta} - \theta) &= \frac{3\mathrm{tr}(\boldsymbol{A}_1^3 + \boldsymbol{A}_1'\boldsymbol{A}_1^2)\mathrm{tr}(\boldsymbol{A}_1'\boldsymbol{A}_1)}{\mathrm{tr}^3(\boldsymbol{A}_2)} - \frac{\mathrm{tr}[\boldsymbol{A}_1^*(\boldsymbol{A}_2 + \boldsymbol{A}_2')]}{2\mathrm{tr}^2(\boldsymbol{A}_2)} \\ &\quad -\frac{\boldsymbol{a}'\boldsymbol{A}_1\boldsymbol{a}}{\mathrm{tr}(\boldsymbol{A}_2)\boldsymbol{a}'\boldsymbol{a}} + o(T^{-1}), \\ \mathbb{E}(\hat{\sigma}^2 - \sigma^2) &= -\frac{\sigma^2}{T} - \frac{\sigma^2\mathrm{tr}(\boldsymbol{A}_1'\boldsymbol{A}_1)}{T\mathrm{tr}(\boldsymbol{A}_2)} + o(T^{-1}), \end{aligned}$$

which suggests that up to order $O(T^{-1})$, $\mathbb{E}(\hat{\theta}-\theta)$ and $\mathbb{E}(\hat{\sigma}^2-\sigma^2)$ are both robust to the distribution of the data. In fact, given special structure of the matrix $\boldsymbol{A}_1$, one can verify that $\boldsymbol{a}'\boldsymbol{a} = T(1+\theta)^{-2} + O(1)$, $\boldsymbol{a}'\boldsymbol{A}_1\boldsymbol{a} = T(1+\theta)^{-3}+O(1)$, $\mathrm{tr}(\boldsymbol{A}_2) = T(1-\theta^2)^{-1}+O(1)$, $\mathrm{tr}(\boldsymbol{A}_1'\boldsymbol{A}_1) = T(1-\theta^2)^{-1}+O(1)$, $\mathrm{tr}[\boldsymbol{A}_1^*(\boldsymbol{A}_2+\boldsymbol{A}_2')] = -8T\theta(1-\theta^2)^{-2}+O(1)$, and $\mathrm{tr}(\boldsymbol{A}_1^3+\boldsymbol{A}_1'\boldsymbol{A}_1^2) = -T\theta(1-\theta^2)^{-2}+O(1)$. Upon substitution, the bias result (7) follows.

Next, for the bias of $\tilde{\boldsymbol{\delta}}_T$, by taking derivatives of (6), we can write

$$\boldsymbol{H}_1 = \begin{pmatrix} -\frac{\lambda_2}{T\sigma^2} & -\frac{\lambda_1}{T\sigma^4} \\ -\frac{\lambda_1}{T\sigma^4} & -\frac{\lambda_0}{T\sigma^6} + \frac{1}{2\sigma^4} \end{pmatrix}, \; \boldsymbol{H}_2 = \begin{pmatrix} \frac{6\lambda_3}{T\sigma^2} & \frac{\lambda_2}{T\sigma^4} & \frac{\lambda_2}{T\sigma^4} & \frac{2\lambda_1}{T\sigma^6} \\ \frac{\lambda_2}{T\sigma^4} & \frac{2\lambda_1}{T\sigma^6} & \frac{2\lambda_1}{T\sigma^6} & \frac{3\lambda_0}{T\sigma^8} - \frac{1}{\sigma^6} \end{pmatrix}.$$

Given this, the stochastic expansion from Bao and Ullah (2007) yields the following:

$$\begin{aligned}
\tilde{\theta}_T - \theta_0 \;=\; & \frac{3d_1^3d_2\lambda_0^2}{8d^3T^2\sigma^{16}} + \frac{d_1d_2^2\lambda_0^2}{8d^3T^2\sigma^{16}} - \frac{3d_0d_1d_2^2\lambda_0^2}{8d^3T^2\sigma^{16}} - \frac{3d_1^2d_3\lambda_0^2}{8d^3T^2\sigma^{16}} + \frac{3d_0d_1^2d_3\lambda_0^2}{4d^3T^2\sigma^{16}} - \frac{d_1^4\lambda_0\lambda_1}{d^3T^2\sigma^{16}} \\
& + \frac{d_0d_1^2d_2\lambda_0\lambda_1}{2d^3T^2\sigma^{16}} + \frac{d_2^2\lambda_0\lambda_1}{8d^3T^2\sigma^{16}} - \frac{d_0d_2^2\lambda_0\lambda_1}{2d^3T^2\sigma^{16}} + \frac{d_0^2d_2^2\lambda_0\lambda_1}{2d^3T^2\sigma^{16}} - \frac{3d_1d_3\lambda_0\lambda_1}{4d^3T^2\sigma^{16}} \\
& + \frac{3d_0d_1d_3\lambda_0\lambda_1}{d^3T^2\sigma^{16}} - \frac{3d_0^2d_1d_3\lambda_0\lambda_1}{d^3T^2\sigma^{16}} - \frac{d_1^3\lambda_1^2}{d^3T^2\sigma^{16}} + \frac{3d_0d_1^3\lambda_1^2}{2d^3T^2\sigma^{16}} - \frac{3d_1d_2\lambda_1^2}{8d^3T^2\sigma^{16}} \\
& + \frac{3d_0d_1d_2\lambda_1^2}{2d^3T^2\sigma^{16}} - \frac{3d_0^2d_1d_2\lambda_1^2}{2d^3T^2\sigma^{16}} - \frac{3d_3\lambda_1^2}{8d^3T^2\sigma^{16}} + \frac{9d_0d_3\lambda_1^2}{4d^3T^2\sigma^{16}} - \frac{9d_0^2d_3\lambda_1^2}{2d^3T^2\sigma^{16}} \\
& + \frac{3d_0^3d_3\lambda_1^2}{d^3T^2\sigma^{16}} - \frac{3d_1^3d_2\lambda_0}{4d^3T\sigma^{14}} - \frac{d_1d_2^2\lambda_0}{4d^3T\sigma^{14}} + \frac{3d_0d_1d_2^2\lambda_0}{4d^3T\sigma^{14}} + \frac{3d_1^2d_3\lambda_0}{4d^3T\sigma^{14}} - \frac{3d_0d_1^2d_3\lambda_0}{2d^3T\sigma^{14}} \\
& + \frac{d_1^4\lambda_1}{d^3T\sigma^{14}} - \frac{d_0d_1^2d_2\lambda_1}{2d^3T\sigma^{14}} - \frac{d_2^2\lambda_1}{8d^3T\sigma^{14}} + \frac{d_0d_2^2\lambda_1}{2d^3T\sigma^{14}} - \frac{d_0^2d_2^2\lambda_1}{2d^3T\sigma^{14}} + \frac{3d_1d_3\lambda_1}{4d^3T\sigma^{14}} \\
& - \frac{3d_0d_1d_3\lambda_1}{d^3T\sigma^{14}} + \frac{3d_0^2d_1d_3\lambda_1}{d^3T\sigma^{14}} + \frac{3d_1^3d_2}{8d^3\sigma^{12}} + \frac{d_1d_2^2}{8d^3\sigma^{12}} - \frac{3d_0d_1d_2^2}{8d^3\sigma^{12}} - \frac{3d_1^2d_3}{8d^3\sigma^{12}} \\
& + \frac{3d_0d_1^2d_3}{4d^3\sigma^{12}} + \frac{d_1d_2\lambda_0^2}{2d^2T^2\sigma^{12}} - \frac{3d_1^2\lambda_0\lambda_1}{2d^2T^2\sigma^{12}} + \frac{d_2\lambda_0\lambda_1}{4d^2T^2\sigma^{12}} - \frac{d_0d_2\lambda_0\lambda_1}{2d^2T^2\sigma^{12}} - \frac{d_1\lambda_1^2}{d^2T^2\sigma^{12}} \\
& + \frac{2d_0d_1\lambda_1^2}{d^2T^2\sigma^{12}} - \frac{d_1\lambda_0\lambda_2}{4d^2T^2\sigma^{12}} + \frac{d_0d_1\lambda_0\lambda_2}{2d^2T^2\sigma^{12}} - \frac{\lambda_1\lambda_2}{4d^2T^2\sigma^{12}} + \frac{d_0\lambda_1\lambda_2}{d^2T^2\sigma^{12}} - \frac{d_0^2\lambda_1\lambda_2}{d^2T^2\sigma^{12}} \\
& + \frac{d_1^3\lambda_0}{2d^2T\sigma^{10}} - \frac{d_1d_2\lambda_0}{2d^2T\sigma^{10}} - \frac{d_0d_1d_2\lambda_0}{2d^2T\sigma^{10}} + \frac{3d_1^2\lambda_1}{2d^2T\sigma^{10}} - \frac{d_0d_1^2\lambda_1}{d^2T\sigma^{10}} - \frac{d_0d_2\lambda_1}{2d^2T\sigma^{10}} \\
& + \frac{d_0^2d_2\lambda_1}{d^2T\sigma^{10}} + \frac{d_1\lambda_2}{4d^2T\sigma^{10}} - \frac{d_0d_1\lambda_2}{2d^2T\sigma^{10}} - \frac{d_1^3}{2d^2\sigma^8} + \frac{d_0d_1d_2}{2d^2\sigma^8} - \frac{d_1\lambda_0}{2dT\sigma^6} - \frac{\lambda_1}{2dT\sigma^6} \\
& + \frac{d_0\lambda_1}{dT\sigma^6} + \frac{d_1}{2d\sigma^4} + o_P(T^{-1}),
\end{aligned}$$

$$
\begin{aligned}
\tilde{\sigma}_T^2 - \sigma_0^2 \;=\; & -\frac{3d_1^2d_2^2\lambda_0^2}{8d^3T^2\sigma^{14}} - \frac{d_2^3\lambda_0^2}{8d^3T^2\sigma^{14}} + \frac{3d_0d_2^3\lambda_0^2}{8d^3T^2\sigma^{14}} - \frac{3d_1^3d_3\lambda_0^2}{4d^3T^2\sigma^{14}} + \frac{3d_1^3d_2\lambda_0\lambda_1}{2d^3T^2\sigma^{14}} + \frac{d_1d_2^2\lambda_0\lambda_1}{2d^3T^2\sigma^{14}} \\
& - \frac{3d_0d_1d_2^2\lambda_0\lambda_1}{2d^3T^2\sigma^{14}} - \frac{3d_1^2d_3\lambda_0\lambda_1}{2d^3T^2\sigma^{14}} + \frac{3d_0d_1^2d_3\lambda_0\lambda_1}{d^3T^2\sigma^{14}} - \frac{d_1^4\lambda_1^2}{d^3T^2\sigma^{14}} + \frac{d_0d_1^2d_2\lambda_1^2}{2d^3T^2\sigma^{14}} \\
& + \frac{d_2^2\lambda_1^2}{8d^3T^2\sigma^{14}} - \frac{d_0d_2^2\lambda_1^2}{2d^3T^2\sigma^{14}} + \frac{d_0^2d_2^2\lambda_1^2}{2d^3T^2\sigma^{14}} - \frac{3d_1d_3\lambda_1^2}{4d^3T^2\sigma^{14}} + \frac{3d_0d_1d_3\lambda_1^2}{d^3T^2\sigma^{14}} + \frac{3d_1^2d_2^2\lambda_0}{4d^3T\sigma^{12}} \\
& - \frac{3d_0^2d_1d_3\lambda_1^2}{d^3T^2\sigma^{14}} + \frac{d_2^3\lambda_0}{4d^3T\sigma^{12}} - \frac{3d_0d_2^3\lambda_0}{4d^3T\sigma^{12}} + \frac{3d_1^3d_3\lambda_0}{2d^3T\sigma^{12}} - \frac{3d_1^3d_2\lambda_1}{2d^3T\sigma^{12}} - \frac{d_1d_2^2\lambda_1}{2d^3T\sigma^{12}} \\
& + \frac{3d_0d_1d_2^2\lambda_1}{2d^3T\sigma^{12}} + \frac{3d_1^2d_3\lambda_1}{2d^3T\sigma^{12}} - \frac{3d_0d_1^2d_3\lambda_1}{d^3T\sigma^{12}} - \frac{3d_1^2d_2^2}{8d^3\sigma^{10}} - \frac{d_2^3}{8d^3\sigma^{10}} + \frac{3d_0d_2^3}{8d^3\sigma^{10}} \\
& - \frac{3d_1^3d_3}{4d^3\sigma^{10}} - \frac{d_2^2\lambda_0^2}{2d^2T^2\sigma^{10}} + \frac{2d_1d_2\lambda_0\lambda_1}{d^2T^2\sigma^{10}} - \frac{d_1^2\lambda_1^2}{d^2T^2\sigma^{10}} + \frac{d_2\lambda_1^2}{2d^2T^2\sigma^{10}} - \frac{d_0d_2\lambda_1^2}{d^2T^2\sigma^{10}} \\
& - \frac{d_1^2\lambda_0\lambda_2}{2d^2T^2\sigma^{10}} - \frac{d_1\lambda_1\lambda_2}{2d^2T^2\sigma^{10}} + \frac{d_0d_1\lambda_1\lambda_2}{d^2T^2\sigma^{10}} - \frac{d_1^2d_2\lambda_0}{2d^2T\sigma^8} + \frac{d_2^2\lambda_0}{2d^2T\sigma^8} + \frac{d_0d_2^2\lambda_0}{2d^2T\sigma^8} \\
& + \frac{d_1^3\lambda_1}{d^2T\sigma^8} - \frac{d_1d_2\lambda_1}{d^2T\sigma^8} - \frac{d_0d_1d_2\lambda_1}{d^2T\sigma^8} + \frac{d_1^2\lambda_2}{2d^2T\sigma^8} + \frac{d_1^2d_2}{2d^2\sigma^6} - \frac{d_0d_2^2}{2d^2\sigma^6} + \frac{d_2\lambda_0}{2dT\sigma^4} \\
& - \frac{d_1\lambda_1}{dT\sigma^4} - \frac{d_2}{2d\sigma^2} + o_P(T^{-1}).
\end{aligned}
$$

To evaluate the approximate biases of $\tilde{\theta}_T$ and $\tilde{\sigma}_T^2$, we need to work out $\mathbb{E}(\lambda_0^{i_0}\lambda_1^{i_1}\lambda_2^{i_2})$ with $i_0+i_1+i_2 \le 2$. When $i_0+i_1+i_2 = 1$, say, $i_0 = 0$, $i_1 = 1$, $i_2 = 0$, $\mathbb{E}(\lambda_0^{i_0}\lambda_1^{i_1}\lambda_2^{i_2}) = T\sigma_0^2 d_1$. When $i_0 + i_1 + i_2 = 2$, we can use Ullah (2004, p. 187) again on expectations of quadratic forms in $\boldsymbol{\varepsilon}$.

Given the special structures of the matrices involved in the quadratic forms, one can verify the following table, which gives the order and value, up to the order of approximation, of each term needed in taking expectations of $\tilde{\theta}_T - \theta_0$ and $\tilde{\sigma}_T^2 - \sigma_0^2$:

Term	Order	Approximate Value	Term	Order	Approximate Value
d_0	$O(1)$	1	$\mathbb{E}(\lambda_0^2)$	$O(T^2)$	$T^2\sigma^4$
d_1	$O(T^{-1})$	$-T^{-1}(1+\theta)^{-1}$	$\mathbb{E}(\lambda_0\lambda_1)$	$O(T)$	$-T\sigma^4(1+\theta)^{-1}$
d_2	$O(1)$	$(1-\theta^2)^{-1}$	$\mathbb{E}(\lambda_0\lambda_2)$	$O(T^2)$	$T^2\sigma^4(1-\theta^2)^{-1}$
d_3	$O(1)$	$-\theta(1-\theta^2)^{-2}$	$\mathbb{E}(\lambda_1^2)$	$O(T)$	$T\sigma^4(1-\theta^2)^{-1}$
d	$O(1)$	$0.5(1-\theta^2)^{-1}\sigma^{-4}$	$\mathbb{E}(\lambda_1\lambda_2)$	$O(T)$	$-T\sigma^4(1+3\theta)(1-\theta^2)^{-2}$

Substituting the above terms into $\mathbb{E}(\tilde{\theta}_T - \theta_0)$ and $\mathbb{E}(\tilde{\sigma}_T^2 - \sigma_0^2)$ yields (8).

REFERENCES

Bao, Y. and A. Ullah, 2007. The second-order bias and mean squared error of estimators in time series models. *Journal of Econometrics* **140(2)**, 650-669.

Cordeiro, G. M. and R. Klein, 1994. Bias correction in ARMA models. *Statistics and Probability Letters* **19(3)**, 169-176.

Hamilton, J., 1994. *Time Series Analysis*. New Jersey: Princeton University Press.

Stock, J. H., and M. W. Watson, 2007. Why has U.S. inflation become harder to forecast? *Journal of Money, Credit and Banking* **39(s1)**, 3-33.

Tanaka, K., 1984. An asymptotic expansion associated with the maximum likelihood estimators in ARMA models. *Journal of the Royal Statistical Society: Series B* **46(1)**, 58-67.

Ullah, A., 2004. *Finite Sample Econometrics*. New York: Oxford University Press.