# Approximate Bias Correction in Econometrics

by

**James G. MacKinnon**[*]

**and**

**Anthony A. Smith, Jr.**[**]

[*] Department of Economics
Queen's University
Kingston, Ontario, Canada
K7L 3N6

**jgm@qed.econ.queensu.ca**

[**] Graduate School of Industrial Administration
Carnegie Mellon University
Pittsburgh, PA 15213
U.S.A.

**smithaa+@andrew.cmu.edu**

This research was supported, in part, by the Social Sciences and Humanities Research Council of Canada. We have benefited from discussions with Allan Gregory, Russell Davidson, Paul Ruud, Fallaw Sowell, Amir Yaron, and Stanley E. Zin. We would also like to thank Takeshi Amemiya, two referees, and seminar participants at the University of Guelph, Princeton University, INSEE, the 1996 North American Winter Meetings of the Econometric Society, the University of Toulouse, and University College London for comments on earlier versions.

First version, February, 1995; final revision, May, 1997.

## Abstract

This paper discusses methods for reducing the bias of consistent estimators that are biased in finite samples. These methods are available whenever the bias function, which relates the bias of the parameter estimates to the values of the parameters, can be estimated by computer simulation or by some other method. If so, bias can be reduced by one full order in the sample size and, in some cases that may not be unrealistic, virtually eliminated. Unfortunately, reducing bias may increase the variance, or even the mean squared error, of an estimator. Whether it does so depends on the shape of the bias function. The results of the paper are illustrated by applying them to two problems: estimating the autoregressive parameter in an AR(1) model with a constant term, and estimating a logit model.

JEL Classification Number: C13

Keywords: bias function, mean squared error, simulation, bootstrap, AR(1), logit model

## 1. Introduction

Many econometric estimators are consistent but biased in finite samples. It is natural to try to reduce this bias by using computer simulation, and the idea of doing so is probably very old. For example, in an interview (Phillips, 1988), James Durbin reports that he worked on this idea in the early 1950s but gave up because it was beyond the capabilities of the computers available at that time. In this paper, we discuss some ways in which finite-sample bias can be reduced or, in certain cases, even eliminated. The key concept is that of a "bias function," which relates the bias of some estimator to the parameter value(s). In many cases, this function can be estimated by computer simulation. In some cases, approximations to it may be obtained analytically. In the examples we study, bias correction generally seems to do a very good job of reducing bias. However, bias correction may either increase or decrease the mean squared error of an estimator. The key result of the paper is that how well bias correction works depends in a simple way on the shape of the bias function and the variance of the parameter estimates.

As we shall discuss below, our work is related to some of the extensive literature on the bootstrap; see, among others, Hall (1992) and Efron and Tibshirani (1993). It is also closely related to the literature on indirect inference, which grew out of the work of Smith (1993) and was given that name by Gouriéroux, Monfort, and Renault (1993); see, in particular, Gouriéroux, Renault, and Touzi (1997), who discuss bias correction in the context of indirect inference.

We begin by considering the case of a scalar parameter $\theta$ which can be estimated consistently from data $y_t$, $t = 1, \ldots, n$, by some standard technique such as least squares or maximum likelihood. Let $\hat\theta$ denote a consistent estimator of $\theta$ based on a sample of size $n$, and let $\theta_0$ denote its true value. We shall call $\hat\theta$ the **initial estimator**. Assuming that $E(\hat\theta)$ exists, we can always write

$$\hat\theta = \theta_0 + b(\theta_0, n) + v(\theta_0, n), \tag{1}$$

where $b(\theta_0, n) \equiv E(\hat\theta) - \theta_0$, and $v(\theta_0, n)$ is defined so that (1) holds. Thus $b(\theta_0, n)$ is the bias of $\hat\theta$ and $v(\theta_0, n)$ is the random difference between $\hat\theta$ and its mean. The function $b(\theta_0, n)$ will be called the **bias function**. Except for the parameter $\theta$, we are assumed to know the distribution of the $y_t$. The key feature of the bias function is that, in general, the bias of $\hat\theta$ depends on $\theta_0$. It is this dependence that makes correcting bias difficult and, sometimes, undesirable to do.

As an illustration, Figure 1 plots bias functions for three sample sizes for the OLS estimate of the parameter $\rho$ in the possibly nonstationary autoregressive model

$$y_t = \mu + \rho y_{t-1} + u_t, \quad u_t \sim \mathrm{NID}(0, \sigma^2). \tag{2}$$

We do not assume stationarity here because the stationarity restriction that $|\rho| < 1$ makes bias correction complicated, and because we do not want to rule out the interesting case in which $\rho = 1$. The stationary case will be discussed briefly in

Section 2. Because we are not assuming stationarity, we have to make an assumption about starting values. For the case $\rho \neq 1$, we assume that $y_0 = \mu/(1-\rho) + u_0$. This assumption implies that the bias functions for the OLS estimate of the parameter $\rho$ do not depend on $\mu$ and $\sigma^2$, so that there is effectively only one parameter; see Appendix A of Andrews (1993). For the case $\rho = 1$, we assume that $\mu = 0$. This assumption implies that the bias functions do not depend on $\sigma^2$. Alternative values of $\mu$, however, would cause the bias functions to be different at $\rho = 1$.

The bias functions in Figure 1 were obtained by computer simulation, using 800,000 replications for $n = 25$, 400,000 replications for $n = 50$, and 200,000 replications for $n = 100$. Using the regression technique proposed by Davidson and MacKinnon (1992), the control variate $\sum_{t=2}^{n} u_t y_{t-1}$ was used to reduce the variance of the estimates and, in order to make the graph as smooth as possible, the same seeds were used for all the simulations. In this case, it would probably have been possible to obtain these bias functions analytically by using an approach like that of Sawa (1978), but it was easier to use simulation. We see from the figure that the bias function for $\hat{\rho}$ in this model is nearly linear for $-0.85 \leq \rho \leq 0.85$. However, it is severely nonlinear in the neighborhood of $|\rho| = 1$.

As a first approximation, the bias functions in Figure 1 appear to be $O(n^{-1})$. This is not an illusion. When $\hat{\theta}$ is root-$n$ consistent and asymptotically normal, the bias of $\hat{\theta}$ will generally be $O(n^{-1})$, provided that the second moment of $\hat{\theta}$ is bounded from above. The bias cannot be $O(n^{-1/2})$, because, if it were, the random variable $n^{1/2}(\hat{\theta} - \theta_0)$ could not have mean zero asymptotically.[1] Under standard regularity conditions which allow the density of $\hat{\theta}$ to have an Edgeworth expansion in powers of $n^{-1/2}$, the first term that can contribute to bias is the $O(n^{-1})$ term; see Hall (1992).

In the next section, we consider methods of bias correction that would be appropriate if the bias function were linear. This case is simple to deal with, may often be a good approximation, and yields some intuitively appealing results. Then, in Section 3, we consider the more general case of a nonlinear bias function. Subsequently, Section 4 extends many of the results to the case in which there is a vector of parameters. Finally, in Sections 5 and 6, we present two sets of Monte Carlo results, one for an AR(1) model and one for a logit model.

## 2. Bias Correction with Constant and Linear Bias Functions

The simplest case is the one in which the bias function is flat, so that $b(\theta, n) = b(n)$ for all $\theta$. In this case, if $b(n)$ is not known analytically, we could estimate it simply by generating $N$ samples of size $n$ from the model that is hypothesized to have generated the $y_t$, using any value of $\theta$ at all. Although it does not matter what

[1] Under the bounded moment assumption, Theorem 3.4.1 of Amemiya (1985) ensures that the asymptotic expectation of $\hat{\theta}$ equals the limit of $E(\hat{\theta})$ as $n \to \infty$.

value of $\theta$ we use if the bias function actually is constant, the obvious one is $\hat{\theta}$. Let the average of the estimates obtained from the $N$ simulated samples be

$$\bar{\theta} \equiv \frac{1}{N} \sum_{j=1}^{N} \hat{\theta}_j .$$

Then our estimate of $b(n)$ would be

$$\hat{b} \equiv \hat{b}(n) = \bar{\theta} - \hat{\theta}. \tag{3}$$

Since the simulated samples are assumed to be drawn from the same model as the data, $\hat{b}$ will provide an unbiased estimate of $b(n)$, and as $N \to \infty$ it should, under plausible conditions, converge to $b(n)$.

In this simple situation, then, we can obtain an estimate of $b(n)$ that is as good as we want (or can afford) it to be. The corresponding estimate of $\theta$, which we shall refer to as the **constant-bias-correcting**, or **CBC**, estimator, will be

$$\tilde{\theta} \equiv \hat{\theta} - \hat{b} = 2\hat{\theta} - \bar{\theta}. \tag{4}$$

This estimator is widely used in the bootstrap literature; see Efron and Tibshirani (1993, Chapter 10). In the bootstrap literature, the simulated samples are often obtained by some form of resampling from the data instead of by using the parametric model evaluated at $\hat{\theta}$, but the basic idea is the same. In econometrics, the CBC estimator has sometimes been used in conjunction with approximations to the bias function; see Amemiya (1980) and Kiviet and Phillips (1993, 1994).

There are many ways to obtain confidence intervals for $\theta$. Indeed, much of the bootstrap literature is concerned with ways to do this; see Hall (1992), Efron and Tibshirani (1993, Chapters 12–14), and DiCiccio and Efron (1996). We will not discuss confidence interval estimation in this paper, however. It is an interesting and important topic, but it cannot adequately be treated in a small space, and it is peripheral to the main point of the paper.

It is probably rare for a bias function to be flat. Suppose instead that it is linear in $\theta$, which may often be a reasonable approximation, at least over some range of parameter values. If the bias function is linear, we can write it as

$$b(\theta) = \alpha + b'\theta, \tag{5}$$

where we have suppressed the explicit dependence of $b(\cdot)$ on $n$. The notation $b'$ emphasizes the fact that the coefficient of $\theta$ in (5) is the slope of the bias function. By evaluating (5) at two points, we can solve for $\alpha$ and $b'$. This requires two sets of simulations, which will differ only in the point at which the DGP is evaluated. Natural choices for the two points are $\hat{\theta}$, the initial estimator, and $\tilde{\theta}$, the CBC estimator defined in (4). In order to ensure that the slope of the bias function is estimated accurately, both sets of simulations should use the same sequence of random numbers.

Given $\hat{\theta}$, $\tilde{\theta}$, and the estimates $\hat{b}$ and $\tilde{b}$, it is easy to solve for $\alpha$ and $b'$. The solutions are

$$\acute{\alpha} = \hat{b} - \frac{\hat{b} - \tilde{b}}{\hat{\theta} - \tilde{\theta}}\,\hat{\theta} \qquad \text{and} \qquad \acute{b}' = \frac{\hat{b} - \tilde{b}}{\hat{\theta} - \tilde{\theta}}\;.$$

Under plausible conditions, $\acute{\alpha}$ and $\acute{b}'$ will converge to $\alpha$ and $b'$ as the number of simulations is increased. The bias-correcting estimator $\check{\theta}$ must be equal to $\hat{\theta}$ minus the bias function evaluated at $\check{\theta}$ itself. Thus it is the solution to the equation

$$\check{\theta} = \hat{\theta} - \acute{\alpha} - \acute{b}'\check{\theta}. \tag{6}$$

Solving (6) yields

$$\check{\theta} = \frac{1}{1 + \acute{b}'}(\hat{\theta} - \acute{\alpha}). \tag{7}$$

This is one form of the **linear-bias-correcting**, or **LBC**, estimator. Another form of this estimator will be discussed in the next section.

Simulation is not always needed to compute bias-corrected estimators. A particularly simple example of the LBC estimator is the OLS estimator of the error variance of a linear regression model with fixed regressors. Consider the biased estimator $SSR/n$. Its bias function is linear and equal to $-(k/n)\sigma^2$, where $k$ is the number of regressors. Thus $\alpha = 0$ and $b' = -k/n$. Plugging these into (7) yields the familiar unbiased estimator $SSR/(n-k)$.[2]

Even if the bias function is not actually known, an approximation may be available. For the example in Figure 1, it might be reasonable to assume that that the bias function is linear for values of $|\rho| < .85$. In fact, for the stationary version of the autoregressive model (2), Kendall (1954) and Marriott and Pope (1954) derived the approximate bias function

$$b(\rho, n) = -\tfrac{1}{n}(1 + 3\rho) + O(n^{-2}). \tag{8}$$

This function is linear in $\rho$. Orcutt and Winokur (1969) used (8) to obtain the approximately unbiased estimator

$$\check{\rho} = \frac{1}{n-3}(n\hat{\rho} + 1).$$

The true bias function for the stationary case (computed by simulation) and the approximate bias function (8) are plotted in Figure 2 for $n = 25$, $n = 50$, and $n = 100$. For most values of $\rho$, with the notable exception of values near $-1$, (8) provides a fairly good approximation. Note that the bias functions in Figure 2 are not the same as the corresponding ones in Figure 1, even though they apply to the

[2] We are grateful to Fallaw Sowell for suggesting this example.

same least squares estimator. For Figure 2, the simulations imposed the stationarity constraint by treating $y_1$ as stochastic with mean zero and variance $\sigma^2/(1-\rho^2)$.

In writing (7), we ignored the experimental error which arises from the fact that $N$ is finite. Since experimental error can be made arbitrarily small by making $N$ sufficiently large, there is little reason to worry about it. Even when $N$ is only 1000, the standard error of the simulation-based estimate of $b(\hat\theta)$ will be only about .032 times the standard error of $\hat\theta$. Often, it will be feasible either to use much larger values of $N$ or to reduce simulation errors further by using control or antithetic variates; see Davidson and MacKinnon (1992). Henceforth, when we discuss the properties of estimators that may be based on simulation, we implicitly assume that $N$ is infinite.

In the remainder of this section, we examine the bias and variance of the LBC and CBC estimators under the assumption that the bias function is given by (5) with coefficients $\alpha$ and $b'$ that are $O(n^{-1})$. The results are quite simple and, as we shall see in the next section, they generalize easily to the nonlinear case.

The bias of the CBC estimator $\tilde\theta$ is

$$\begin{aligned} b(\theta_0) - E\big(b(\hat\theta)\big) &= \alpha + b'\theta_0 - \alpha - b'E(\hat\theta \,|\, \theta_0) \\ &= b'\theta_0 - b'(\theta_0 + \alpha + b'\theta_0) \\ &= -b'(\alpha + b'\theta_0). \end{aligned} \tag{9}$$

Thus the bias of $\tilde\theta$ is just $-b'$ times the bias of $\hat\theta$. The CBC estimator will be unbiased only when the bias function is flat, that is, when $b' = 0$. Provided that $|b'| < 1$, $\tilde\theta$ will be less biased than the initial estimator. It will be biased in the same direction when $b' < 0$ and biased in the opposite direction when $b' > 0$. The bias of the CBC estimator will evidently be $O(n^{-2})$, since it is the product of two factors, each of which is $O(n^{-1})$.

In contrast, the LBC estimator $\check\theta$ will be unbiased whenever the bias function is linear. To see this, observe that

$$\begin{aligned} E(\check\theta) &= E\big(\hat\theta - b(\check\theta)\big) \\ &= \theta_0 + b(\theta_0) - b(\theta_0) = \theta_0. \end{aligned} \tag{10}$$

The key to this result is that $E\big(b(\check\theta)\big) = b(\theta_0)$, which will be true, in general, only when $b(\theta)$ is linear.

Since $\tilde\theta = \hat\theta - \alpha - b'\hat\theta$, the variance of the CBC estimator is

$$V(\tilde\theta) = (1-b')^2 V(\hat\theta). \tag{11}$$

Similarly, it is easy to see from (7) that the variance of the LBC estimator is

$$V(\check\theta) = \frac{1}{(1+b')^2} V(\hat\theta). \tag{12}$$

Thus, for both bias-correcting estimators, whether their variance will be greater than or less than that of $\hat{\theta}$ will depend on whether $b'$ is less than or greater than zero.[3]

It is interesting to compare (11) with (12). If $b' \neq 0$ and $|b'| < \sqrt{2}$, then $(1 - b')^2 < 1/(1 + b')^2$. This implies that the CBC estimator will have smaller variance than the LBC estimator in almost all cases of interest. It is quite possible that this smaller variance will more than offset the bias of $\tilde{\theta}$, causing it to have smaller mean squared error (MSE) than $\check{\theta}$. The condition for MSE($\tilde{\theta}$) to be smaller than MSE($\check{\theta}$) is

$$\frac{1}{(1+b')^2} V(\hat{\theta}) > (1 - b')^2 V(\hat{\theta}) + (b')^2(\alpha + b'\theta_0)^2. \tag{13}$$

When $b' < 0$, the unbiased LBC estimator $\check{\theta}$ may well have greater MSE than the initial estimator $\hat{\theta}$. If the bias function is linear, this will happen whenever

$$\frac{1}{(1+b')^2} V(\hat{\theta}) > (\alpha + b'\theta_0)^2 + V(\hat{\theta}). \tag{14}$$

If the variance of $\hat{\theta}$ is small enough or $b' > 0$, condition (14) will never be satisfied. Thus bias correction can be expected to work well whenever the bias function slopes upward, or when the variance of $\hat{\theta}$ is small relative to the bias. When the bias function slopes downward, bias correction will only work well if $V(\hat{\theta})$ is small.

The results (12), (13), and (14) do not make sense if $b' = -1$. It seems plausible to assume that $b' > -1$, since otherwise the derivative of $E(\hat{\theta})$ with respect to $\theta_0$ would actually be negative, and it does not seem a very strong requirement for an estimator that its expectation should be positively related to the true parameter value. However, in very small samples, this assumption could sometimes be false.

The above results suggest that the CBC and LBC estimators may well have larger MSE than the initial estimator, and that CBC may have smaller MSE than LBC even though it is biased and LBC is not. We shall encounter an example which displays both these properties in Section 5. Thus it is clear that bias correction is not always a good thing to do. Although bias correction leads to smaller bias in a wide variety of circumstances, it increases mean squared error if the bias function slopes downward and the variance of $\hat{\theta}$ is sufficiently large relative to its bias.

## 3. Bias Correction with a Nonlinear Bias Function

As Figures 1 and 2 make clear, bias functions are not always approximately linear. In this section, we first propose an estimator that can deal with arbitrary nonlinear bias functions. We then discuss how the CBC and LBC estimators are related to the new estimator when the bias function is nonlinear. We obtain several

[3] Smith, Sowell, and Zin (1997) make a similar point.

interesting results. In particular, we show that the LBC estimator can be thought of as an approximation to the new estimator, and that all three estimators are biased only at $O(n^{-2})$.

The key to determining $\check{\theta}$ in the last section was equation (6), in which $\check{\theta}$ was set equal to $\hat{\theta}$ minus an estimate of the bias evaluated at $\check{\theta}$. In the nonlinear case, the analogue of (6) is

$$\ddot{\theta} = \hat{\theta} - b(\ddot{\theta}). \tag{15}$$

If we can solve (15), we can find the **nonlinear-bias-correcting**, or **NBC**, estimator $\ddot{\theta}$. Any technique for finding the roots of an equation in one variable could potentially be used. One *ad hoc* technique that seems to work well is the following. First, find $\hat{b}$ as in (3). Then compute the sequence of estimates

$$\ddot{\theta}^{(j)} = (1-\gamma)\ddot{\theta}^{(j-1)} + \gamma\big(\hat{\theta} - b(\ddot{\theta}^{(j-1)})\big), \tag{16}$$

where $\ddot{\theta}^{(0)} = \hat{\theta}$ and $0 < \gamma \leq 1$, and stop when $|\ddot{\theta}^{(j)} - \ddot{\theta}^{(j-1)}|$ is sufficiently small. It is easy to see that, if this sequence converges, it will converge to a value $\ddot{\theta}$ that satisfies (15). Whether or not it converges will depend on the shape of the bias function and on the value of $\gamma$. Larger values of $\gamma$ are likely to result in a lower probability that the sequence will converge, but faster convergence if it does so. In practice, it may be desirable to try $\gamma = 1$ first and then try lower values of $\gamma$ if the procedure does not seem to be converging. A key advantage of this technique is that it does not require the calculation of any derivatives of the bias function. However, it will require a number of evaluations of that function.

This procedure has recently been used by Smith, Sowell, and Zin (1997) to obtain almost unbiased estimates of the order of integration in a fractionally integrated time-series model. In that application, where the bias function was very flat, it worked well. It also seems to work well for the examples dealt with in Sections 5 and 6. A similar procedure has been used by Gouriéroux, Renault, and Touzi (1997).

It is easy to see that, when $b(\theta)$ is nonlinear, $\ddot{\theta}$ is, in general, biased. When we take expectations of both sides of (15), as we did in (10), the nonlinearity of $b(\theta)$ implies that $E\big(b(\ddot{\theta})\big) \neq b(\theta_0)$. If we take a second-order Taylor series expansion of (15) around $\theta_0$, we obtain

$$\ddot{\theta} \cong \hat{\theta} - b_0 - b_0'(\ddot{\theta} - \theta_0) - \tfrac{1}{2}b_0''(\ddot{\theta} - \theta_0)^2, \tag{17}$$

where "$\cong$" denotes asymptotic equality, which means, in this case, that the probability limit of the difference between the two sides is zero. In (17), $b_0$ denotes $b(\theta_0)$, and $b_0'$ and $b_0''$ denote the first and second derivatives of $b(\theta)$, evaluated at $\theta_0$. The remainder of the Taylor expansion in (17) can be shown to be of lower order than the last term. Thus, provided the remainder has a bounded second moment, we can take expectations of both sides of (17), divide through by $1 + b'_0$, and rearrange to find that

$$E(\ddot{\theta}) - \theta_0 \cong -\tfrac{1}{2}\frac{b''_0}{1 + b'_0}E(\ddot{\theta} - \theta_0)^2. \tag{18}$$

This suggests, but does not guarantee, that the bias will be small if the second derivative of $b(\theta)$ is small near $\theta_0$. It also suggests that the sign of the bias will be opposite to that of $b''_0$.

There is some similarity between the NBC estimator and an estimator that Andrews (1993) recently proposed for a class of autoregressive models. Another way to write (15) is $\ddot{\theta} = h^{-1}(\hat{\theta})$, where $h(\theta) \equiv \theta + b(\theta)$. What we are doing is inverting the "mean function" $h(\theta)$, in much the same way that Andrews inverted the "median function." Because the median of $f(x)$ is equal to $f(m_x)$ for any monotonic function $f(\cdot)$, where $m_x$ is the median of $x$, Andrews was able to obtain median-unbiased estimators. It is because this is not true for expectations that the NBC estimator is, in general, biased. Of course, our technique could easily be used to obtain a median-unbiased estimator. We would simply have to replace the bias function $b(\ddot{\theta})$ in (15) by the difference between the median of $\ddot{\theta}$ and $\theta_0$.

As we remarked in the previous section, there is more than one way to define the LBC estimator. Instead of obtaining a linear approximation to the bias function by computing the value of $b(\theta)$ at $\hat{\theta}$ and $\tilde{\theta}$, we could just as well evaluate the bias function and its slope at $\hat{\theta}$ and use a first-order Taylor expansion. This Taylor expansion is

$$b(\theta) \cong \hat{b} + \hat{b}'(\theta - \hat{\theta}) = \hat{b} - \hat{b}'\hat{\theta} + \hat{b}'\theta,$$

where, of course, $\hat{b}' = b'(\hat{\theta})$. Thus $\hat{b} - \hat{b}'\hat{\theta}$ plays the role of $\alpha$ and $\hat{b}'$ plays the role of $b'$. Substituting these quantities into expression (7) yields

$$\check{\theta} = \frac{1}{1 + \hat{b}'}(\hat{\theta} - \hat{b} + \hat{b}'\hat{\theta}) = \hat{\theta} - \frac{\hat{b}}{1 + \hat{b}'}. \tag{19}$$

If the bias function is linear, the LBC estimator defined in (19) will be identical to the LBC estimator defined in (7). The new definition (19) is more convenient analytically than the earlier one, and it may sometimes be more convenient in practice.

It is now easy to show that the LBC and NBC estimators are equal through $O_p(n^{-1})$. If we perform a first-order Taylor expansion of the right-hand expression in (19) around $\theta_0$, replace $\hat{\theta}$ by $\theta_0 + b_0 + v_0$, and explicitly retain only terms that are at least $O_p(n^{-1})$, we find that

$$\check{\theta} = \theta_0 + \frac{1}{1 + b'_0}v_0 + O_p(n^{-2}). \tag{20}$$

Now recall from equation (1) that $\hat{\theta} = \theta_0 + b_0 + v_0$, where $v_0 \equiv v(\theta_0)$ is a random variable that is $O_p(n^{-1/2})$. Substituting for $\hat{\theta}$ in (17), we obtain

$$(1 + b_0')\ddot{\theta} \cong (1 + b_0')\theta_0 + v_0 - \tfrac{1}{2}b_0''(\ddot{\theta} - \theta_0)^2.$$

Since the last term here is $O_p(n^{-2})$, we can ignore it and solve for $\ddot{\theta}$ to obtain

$$\ddot{\theta} = \theta_0 + \frac{1}{1 + b_0'}\,v_0 + O_p(n^{-2}). \tag{21}$$

The right-hand side of this expression is identical to the right-hand side of (20). Thus the LBC and NBC estimators are identical through $O_p(n^{-1})$. This result suggests that LBC and NBC should perform very similarly in practice, something that we do indeed find to be the case in the Monte Carlo experiments that will be reported in Sections 5 and 6.

Subtracting $\theta_0$ from both sides of (21) and squaring yields

$$(\ddot{\theta} - \theta_0)^2 = \frac{1}{(1 + b_0')^2}\,v_0^2 + o_p(n^{-2}).$$

Under the assumption that the remainder term here has a bounded second moment, taking expectations yields the variance of the NBC estimator,

$$V(\ddot{\theta}) = \frac{1}{(1 + b_0')^2}V(\hat{\theta}) + o(n^{-2}). \tag{22}$$

The first term in (22) is equal to (12), the variance of the LBC estimator when the bias function is linear. Thus everything that was said in the previous section about how the variance of the LBC estimator depends on the slope of the bias function must be true, as an approximation, for the NBC estimator as well.

The result that the LBC and NBC estimators are the same through $O_p(n^{-1})$ does not imply that they have the same bias to highest order, because their bias is $O(n^{-2})$, a result that we have not yet shown. In a previous version of this paper (MacKinnon and Smith, 1996), we demonstrated that the biases of these two estimators differ only at $O(n^{-3})$ and derived an expression for their bias at $O(n^{-2})$. Because the derivation is tedious, we omit it. The bias of both estimators can be shown to be

$$-\tfrac{1}{2}\frac{b_0''}{1 + b_0'}E(\hat{\theta} - \theta_0)^2 + O(n^{-3}). \tag{23}$$

This expression is, of course, very similar to (18). It implies that both the LBC and NBC estimators will be unbiased through $O(n^{-2})$ if the second derivative of the bias function is zero at the point $\theta_0$. It also implies that, at this order, the sign of the bias will be opposite to the sign of $b_0''$.

We now turn our attention to the properties of the CBC estimator $\tilde{\theta}$ for the case of a nonlinear bias function. Its bias is evidently $b(\theta_0) - E\big(b(\hat{\theta})\big)$, because we construct the estimator by subtracting the estimated bias $b(\hat{\theta})$ instead of the true bias $b(\theta_0)$ from $\hat{\theta}$. Replacing $b(\hat{\theta})$ by a Taylor expansion around $\theta_0$, we obtain

$$\begin{aligned} E(\tilde{\theta}) - \theta_0 &= b_0 - \Big(b_0 + b_0' E(\hat{\theta} - \theta_0) + \frac{1}{2} b_0'' E(\hat{\theta} - \theta_0)^2\Big) + o(n^{-2}) \\ &= -\, b_0' b_0 - \frac{1}{2} b_0'' E(\hat{\theta} - \theta_0)^2 + o(n^{-2}). \end{aligned} \tag{24}$$

The first two terms in (24) are both clearly $O(n^{-2})$. The first term is the bias of the CBC estimator when the bias function is linear but not flat; recall (9). The second term will almost always have the same sign as (23), but it may be either larger or smaller in absolute value, depending on whether $b_0'$ is positive or negative. Thus it is clearly possible for (24) to be either larger or smaller than (23). Therefore, the NBC and LBC estimators may be either less or more biased, through terms of order $n^{-2}$, than the CBC estimator.

Using the fact that $\tilde{\theta} \cong \hat{\theta} - b_0 - b_0'(\hat{\theta} - \theta_0)$, it can easily be shown that the variance of $\tilde{\theta}$ is

$$V(\tilde{\theta}) = (1 - b_0')^2 V(\hat{\theta}) + o(n^{-2}). \tag{25}$$

The leading-order term here is essentially the same as expression (11) for the variance of the CBC estimator in the case of a linear bias function.

The explicit expressions (23) and (24) for the biases of the LBC/NBC and CBC estimators through $O(n^{-2})$ could, in principle, be used to calculate estimators that are unbiased to even higher order. It would simply be necessary to estimate the quantities that appear in these expressions and then subtract the resulting estimates of bias. This would provide an alternative, and potentially much cheaper, technique than the iterated bootstrap (Beran, 1987) to implement higher-order bias correction. This approach will be explored in another paper. The Monte Carlo results that we present below suggest that the first-order bias correction procedures we investigate in this paper may often do an extremely good job of eliminating bias in practice. Higher-order methods are therefore likely to be unnecessary in many cases.

## 4. The Vector Case

In the preceding two sections, we have discussed three different bias-correcting estimators, two of which are equivalent to highest order. These were all based on the assumption that $\theta$ is a scalar. This is not quite as restrictive as it might seem, since our analysis will still be valid when there are other parameters in the model, provided that the bias of $\hat{\theta}$ does not depend on their values. In this section, however, we relax this assumption by considering the case in which $\boldsymbol{\theta}$ is a $k \times 1$ vector.

First of all, the CBC estimator $\tilde{\boldsymbol{\theta}}$ can be calculated exactly as before. We generate $N$ samples of size $n$ from the DGP with parameter vector $\hat{\boldsymbol{\theta}}$ and define

$\bar{\boldsymbol{\theta}}$ as the mean of the estimates obtained from these samples. Then $\tilde{\boldsymbol{\theta}} \equiv 2\hat{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}}$. The covariance matrix of $\tilde{\boldsymbol{\theta}}$ is simply the matrix analogue of expression (25). Since $\tilde{\boldsymbol{\theta}} \cong \hat{\boldsymbol{\theta}} - \boldsymbol{b}_0 - \boldsymbol{B}_0'(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$, it is

$$\boldsymbol{V}(\tilde{\boldsymbol{\theta}}) = (\mathbf{I} - \boldsymbol{B}_0')\boldsymbol{V}(\hat{\boldsymbol{\theta}})(\mathbf{I} - \boldsymbol{B}_0')' + o(n^{-2}). \tag{26}$$

When the bias function is flat, $\tilde{\boldsymbol{\theta}}$ will be unbiased, and it will have the same variance as the initial estimator $\hat{\boldsymbol{\theta}}$. Now let us consider what happens when the bias function is nonlinear. Let $\boldsymbol{b}_0$ denote the bias function, which is a $k$-vector, evaluated at $\boldsymbol{\theta}_0$, $\boldsymbol{B}_0$ denote the $k \times k$ matrix of first derivatives of $\boldsymbol{b}(\boldsymbol{\theta})$, evaluated at $\boldsymbol{\theta}_0$, and $\boldsymbol{b}_0^{ij}$ denote the vector of second derivatives of $\boldsymbol{b}(\boldsymbol{\theta})$ with respect to the parameters $\theta_i$ and $\theta_j$, evaluated at $\boldsymbol{\theta}_0$. Then the bias of the CBC estimator is

$$E(\tilde{\boldsymbol{\theta}}) - \boldsymbol{\theta}_0 = -\boldsymbol{B}_0'\boldsymbol{b}_0 - \tfrac{1}{2}\sum_{i=1}^{k}\sum_{j=1}^{k} E\big((\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)'\big)\boldsymbol{b}_0^{ij} + o(n^{-2}). \tag{27}$$

Expression (27) is the vector analogue of expression (24). The first term is the bias of $\tilde{\boldsymbol{\theta}}$ when the bias function is linear.

The LBC estimator is also fairly easy to obtain. We simply have to evaluate the $k$-vector $\hat{\boldsymbol{b}} \equiv \boldsymbol{b}(\hat{\boldsymbol{\theta}})$ and its $k \times k$ matrix of derivatives $\hat{\boldsymbol{B}} \equiv \boldsymbol{B}(\hat{\theta})$. This will generally require either $k+1$ or $2k+1$ evaluations of the bias function, depending on whether one-sided or two-sided derivatives are used. The LBC estimator is then

$$\check{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}} - (\mathbf{I} + \hat{\boldsymbol{B}}')^{-1}\hat{\boldsymbol{b}}; \tag{28}$$

compare (19). The covariance matrix of $\check{\boldsymbol{\theta}}$ is the matrix analogue of (22):

$$\boldsymbol{V}(\check{\boldsymbol{\theta}}) = (\mathbf{I} + \boldsymbol{B}_0')^{-1}\boldsymbol{V}(\hat{\boldsymbol{\theta}})\big((\mathbf{I} + \boldsymbol{B}_0')'\big)^{-1} + o(n^{-2}). \tag{29}$$

This covariance matrix can be estimated in more than one way, as can (26). We simply have to replace $\boldsymbol{V}(\hat{\boldsymbol{\theta}})$ by a suitable estimate, such as the inverse of the information matrix evaluated at $\check{\boldsymbol{\theta}}$, and replace $\boldsymbol{B}_0$ by $\boldsymbol{B}(\check{\theta})$.

It should be clear from (26) and (29) that either form of bias correction may either increase or decrease the variance of the parameter estimates. It is quite possible that the variance may increase for some parameters and decrease for others. In the special case in which the matrix $\boldsymbol{B}_0$ is diagonal, the results given in Section 2 for the scalar case will still apply. When we apply these results to the $j^{\text{th}}$ parameter, the $j^{\text{th}}$ diagonal element of $\boldsymbol{B}_0$ will play the role of $b'$. Notice that it is easy to estimate (26) and (29). We simply have to replace $\boldsymbol{V}(\hat{\boldsymbol{\theta}})$ by any valid estimate of the covariance matrix of $\hat{\boldsymbol{\theta}}$ and replace $\boldsymbol{B}_0$ by $\hat{\boldsymbol{B}}$.

The NBC estimator $\ddot{\boldsymbol{\theta}}$ can be computed by solving

$$\ddot{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}} - \boldsymbol{b}(\ddot{\boldsymbol{\theta}}), \tag{30}$$

which is the vector version of (15). There are at least two ways to do this. One is to modify the iterative procedure (16) as follows:

$$\ddot{\boldsymbol{\theta}}^{(j)} = (1-\boldsymbol{\gamma}){*}\ddot{\boldsymbol{\theta}}^{(j-1)} + \boldsymbol{\gamma}{*}\big(\hat{\boldsymbol{\theta}} - \boldsymbol{b}(\ddot{\boldsymbol{\theta}}^{(j-1)})\big), \tag{31}$$

where "*" denotes direct product and $\boldsymbol{\gamma}$ is now a $k-$vector, each element of which is between 0 and 1. In practice, of course, it may be easier to make all elements of $\boldsymbol{\gamma}$ the same. As before, this procedure is not guaranteed to converge.

Another approach is to use Newton's Method. A typical Newton step would be

$$\ddot{\boldsymbol{\theta}}^{(j+1)} = \ddot{\boldsymbol{\theta}}^{(j)} - \big(\mathbf{I} + \boldsymbol{B}'(\ddot{\boldsymbol{\theta}}^{(j)})\big)^{-1}\big(\boldsymbol{b}(\ddot{\boldsymbol{\theta}}^{(j)}) + \ddot{\boldsymbol{\theta}}^{(j)} - \hat{\boldsymbol{\theta}}\big). \tag{32}$$

The matrix of derivatives $\boldsymbol{B}(\ddot{\boldsymbol{\theta}}^{(j)})$ would have to be evaluated numerically, which could be expensive. For the first Newton step, $\ddot{\boldsymbol{\theta}}^{(0)}$ would equal $\hat{\boldsymbol{\theta}}$, and this step would yield the LBC estimator; compare (28). Notice that, when the iterative procedure (31) works well, it may often require fewer than $k+1$ evaluations of the bias function. In such cases, the NBC estimator will be less expensive to compute than the LBC estimator.

It is obvious that the LBC and NBC estimators are equivalent in the vector case just as they are in the scalar case. Therefore, the right-hand side of expression (29) gives the covariance matrix of the NBC estimator through $O(n^{-2})$. The bias of both estimators can be found by Taylor expanding $\boldsymbol{b}(\ddot{\boldsymbol{\theta}})$ around $\boldsymbol{\theta}_0$, substituting the Taylor expansion into (30), taking expectations, and then solving. The result is

$$-\tfrac{1}{2}(\mathbf{I} + \boldsymbol{B}_0')^{-1}\sum_{i=1}^{k}\sum_{j=1}^{k} E\big((\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)'\big)\boldsymbol{b}_0^{ij} + O(n^{-3}).$$

This is the analogue of expression (23) for the scalar case.

## 5. Monte Carlo Results for an AR(1) Model

The three bias-correcting estimators proposed in Sections 2 and 3 were applied to the estimation of $\rho$ in the nonstationary AR(1) regression model (2). Because the bias function had already been computed numerically for various sample sizes (see Figure 1), it was not necessary to do any simulation to obtain it. This made it feasible to use quite a large number of replications in the Monte Carlo experiments. There were 400,000 replications for each of three sample sizes (25, 50, and 100) and each of the following 83 different values of $\rho$:

$$\rho = -1.20, -1.18, \ldots, -1.06, -1.05, -1.04, \ldots, -.90, -.85, \\ \ldots, .90, .91, \ldots, 1.05, 1.06, 1.08, \ldots, 1.20\ .$$

Different seeds were used for each experiment, and no control variates were employed.

Figures 3 and 5 show the biases of the OLS estimator $\hat{\rho}$, the CBC estimator $\tilde{\rho}$, the LBC estimator $\breve{\rho}$, and the NBC estimator $\ddot{\rho}$ as a function of $\rho$ for $n = 25$ and $n = 100$, respectively. We also obtained results for $n = 50$, but these are not shown. The biases of $\hat{\rho}$ are essentially the same as those in Figure 1. In contrast, for most values of $\rho$, $\tilde{\rho}$ exhibits only a little bias, and the other two estimators exhibit almost no bias. There is a fair amount of bias only for values of $\rho$ near $\pm 1$, which is where the bias functions are severely nonlinear. Except for values of $\rho$ very near $\pm 1$, the magnitude of the bias for the bias-correcting estimators appears to be roughly $O(n^{-2})$, as predicted by the theoretical results of Section 3.

In Section 2, we showed that, for a linear bias function, the CBC estimator $\tilde{\rho}$ will be less biased than $\hat{\rho}$ whenever $|b'| < 1$. This condition is not satisfied for some values of $\rho$ greater than 1. We see from the figures that, for values of $\rho$ in this region, the bias of $\tilde{\rho}$ is opposite in sign and only somewhat smaller in magnitude than the bias of $\hat{\rho}$. As the results of Section 3 suggest, the curves for the LBC estimator $\breve{\rho}$ and the NBC estimator $\ddot{\rho}$ are almost indistinguishable, although the latter does seem to have a bit less bias in the worst cases when $\rho$ is very close to 1.

Figures 4 and 6 show the root mean squared errors (RMSE) of the four estimators as a function of $\rho$ for $n = 25$ and $n = 100$, respectively. Despite the success of the bias-correcting estimators in reducing or eliminating bias, the OLS estimator has lower RMSE than any of the BC estimators for $\rho$ between about $-0.9$ and 0.5. Only for values of $\rho$ greater than about 0.8 do the BC estimators produce a marked reduction in RMSE. The CBC estimator, which performs least well at removing bias, has lower RMSE than the other bias-correcting estimators except for $|\rho| > 0.8$.

The reason why the bias-correcting estimators have larger RMSE than the OLS estimator for most values of $\rho$ is easy to find. According to equation (22), the variance of both $\breve{\rho}$ and $\ddot{\rho}$ should be equal to $(1 + b'_0)^{-2}$ times the variance of $\hat{\rho}$. Similarly, equation (25) implies that the variance of $\tilde{\rho}$ should be equal to $(1 - b'_0)^2$ times the variance of $\hat{\rho}$. Since $b'$ is negative, except for values of $\rho$ near or less than $-1$ and near or greater than $+1$, this implies that the variance of all the BC estimators will generally exceed the variance of the OLS estimator. Only for relatively large values of $\rho$, where the bias of $\hat{\rho}$ is large, does the reduced bias of the BC estimators outweigh their increased variance.

Of course, these results apply only to the model (2). The results of Kiviet and Phillips (1993) suggest that, in more complicated models with additional regressors, the CBC estimator (in their case, based on an approximation to the bias function) performs well over a wider range of values of $\rho$. Their approximate bias function could also be used to obtain approximate LBC or NBC estimators.

Equations (22) and (25) do a remarkably good job of explaining the performance of the bias-correcting estimators. Panels A and B of Figure 7 plot the observed standard errors of $\tilde{\rho}$ and $\ddot{\rho}$ as a function of $\rho$ for $n = 50$, along with the predicted standard errors according to equations (25) and (22), respectively. The predicted standard errors were computed from the observed standard errors of $\hat{\rho}$ and the slope of the bias function. Only for values of $\rho$ between about 0.8 and 1.1 are there substantial discrepancies between what we observe and what the theory predicts. Even for values of $\rho$ near $-1$, where the bias function is decidedly nonlinear, the predicted standard errors are reasonably close to the true ones. This suggests that, when deciding whether to use a BC estimator, and which one, it may often be reasonable to rely on equations (22) and (25) or their matrix analogues (29) and (26).

Another interesting feature of the experiments is that the iterative procedure based on (16) worked extremely well. For most values of $\rho$, the procedure converged in 8 or fewer iterations (using a tolerance of $10^{-5}$), with $\gamma = 1$. Only for values of $\rho$ around 1 was it ever necessary to use values of $\gamma$ less than 1.

## 6. Monte Carlo Results for a Logit Model

Bias correction may be particularly attractive in the case of binary response models such as the logit model. The logit model may be written as

$$E(y_t) = P_t(\boldsymbol{X}_t\boldsymbol{\beta}) \equiv \big(1 + \exp(-\boldsymbol{X}_t\boldsymbol{\beta})\big)^{-1}, \tag{33}$$

where $y_t$ is either 0 or 1, $\boldsymbol{X}_t$ is a $1 \times k$ vector of regressors, and $\boldsymbol{\beta}$ is a $k \times 1$ vector of unknown parameters. Although maximum likelihood estimation of this model is usually quite straightforward, the ML estimates tend to be biased away from zero; see Amemiya (1980). This bias is similar to the bias of the ML estimate of $\sigma^2$ in least squares estimation, which arises because the residuals tend to underestimate the error terms. In a logit model, larger absolute values of $\boldsymbol{\beta}$ correspond to a model that fits better, so the tendency of the ML estimates to overfit the data results in their being biased away from zero.

Amemiya (1980) developed an approximation to the bias of the ML logit estimator that is valid to order $1/n$. For the logit model (33), this approximation can be written as

$$\boldsymbol{b}^a(\boldsymbol{\beta}) \equiv \tfrac{1}{2}(\boldsymbol{X}'\boldsymbol{\Omega}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{d}, \tag{34}$$

where $\boldsymbol{\Omega}$ is an $n \times n$ diagonal matrix which has typical diagonal element $P_t(1 - P_t)$ and $\boldsymbol{d}$ is an $n \times 1$ vector which has typical element

$$d_t \equiv (2P_t - 1)\big[\boldsymbol{\Omega}^{1/2}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{\Omega}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Omega}^{1/2}\big]_{tt}. \tag{35}$$

Here $[\,\cdot\,]_{tt}$ denotes the $t^{\text{th}}$ diagonal element of the matrix within the brackets. Because only the diagonal elements of the $n \times n$ matrix in (35) need to be calculated, the approximation (34) is quite easy to compute. The notation $\boldsymbol{b}^a(\boldsymbol{\beta})$ emphasizes the fact that bias depends on the value of $\boldsymbol{\beta}$ through the $P_t$.

It is much more attractive to obtain bias-corrected estimates by using the approximate bias function $\boldsymbol{b}^a(\boldsymbol{\beta})$ defined in (34) and (35) than by using a bias function obtained by simulation. This is true for several reasons. Firstly, simulation would be very much more computationally expensive than evaluating $\boldsymbol{b}^a(\boldsymbol{\beta})$. Secondly, when simulation is used, the estimated bias function will not be a smooth, or even a monotonic, function of $\boldsymbol{\beta}$. The problem is that a small change in $\boldsymbol{\beta}$ may not change the values of the $y_t$ in the simulated samples at all. When this happens, the estimates will not change, and the slope of the estimated bias function will be precisely $-1$. This will not seriously affect the CBC estimator, but experience has shown that it does cause serious problems for the other two estimators.

A third reason not to use simulation to obtain the bias function is that ML estimation of logit models has a fundamental difficulty which may be encountered during the simulation. The problem is that ML estimates do not exist when every value of $y_t$ in the sample can be predicted correctly. This is especially likely to happen when the sample size is small and the model fits well. Even though this problem is rarely encountered with real data, it might well be encountered during the many ML estimations needed to simulate the bias function. Indeed, it is because we encountered this problem quite often when doing experiments with samples of size 25 and 50 that we used $n = 100$ in the experiments reported here.

Most of our experiments dealt with a two-parameter logit model with $\boldsymbol{X}_t\boldsymbol{\beta} = \beta_0 + \beta_1 x_t$, where the regressor $x_t$ is distributed as $N(0, 1)$. Bias functions were obtained by simulation using 100,000 pairs of antithetic variates for 61 values of $\beta_1$ ranging from $-3.00$ to 3.00 by increments of 0.1. Using antithetic variates yielded somewhat more accurate estimates of bias than simply doing 200,000 independent replications, except for values of $\beta_1$ close to zero when $\beta_0 = 0$, where the efficiency gain was enormous. No smoothing was done, and different random numbers were used for each replication.

During the course of our experiments, there were a few cases in which the ML estimates failed to exist, always for values of $\beta_0$ and/or $\beta_1$ that were relatively large in absolute value. For the experiments that were used to graph the bias functions in Figure 8, there were 8 failures in 12.2 million replications when $\beta_0 = 0$ and 196 failures in 12.2 million replications when $\beta_0 = 2$. Replications for which ML estimates could not be obtained were discarded and replaced. This seems to be the appropriate thing to do, since bias correction will only be used if the original ML estimates exist.

Figure 8 shows the actual and approximate bias of the ML estimate $\hat{\beta}_1$ of the slope coefficient as a function of $\beta_1$. This figure has several interesting features. The bias function for $\hat{\beta}_1$ slopes upward, the absolute value of the bias of $\hat{\beta}_1$ increases with the absolute value of $\beta_0$ (the curve for $\beta_0 = -2$ is not shown because it is indistinguishable from the curve for $\beta_0 = 2$), and the approximate bias is always smaller in absolute value than the true bias. Moreover, only a modest amount of nonlinearity is evident in the figure.

Figure 9 shows actual and approximate bias functions for the ML estimate of the constant term $\hat{\beta}_0$ as a function of $\beta_1$ in the same two-parameter logit model. The bias functions for $\beta_0$ as a function of itself are not graphed because they look very similar to the ones in Figure 8. From Figure 9, we see that the bias of $\hat{\beta}_0$ has the same sign as $\beta_1$ and increases in absolute value as the absolute value of $\beta_1$ increases. Once again, the approximate bias is always smaller in absolute value than the true bias, but it does seem to provide a fairly good approximation. Note that, when $\beta_0 = 0$, the bias function for $\hat{\beta}_0$ as a function of $\beta_1$ is essentially flat at zero; this function was not graphed to avoid cluttering the figure.

The bias functions in Figures 8 and 9 suggest that bias correction should work very well for this logit model. This is in fact the case, as can be seen from Figures 10, 11, 12, and 13, which are based on 200,000 independent replications for each value of $\beta_1$. The first two figures show the bias and RMSE of $\hat{\beta}_1$, $\tilde{\beta}_1$, and $\ddot{\beta}_1$ as a function of $\beta_1$ for the case in which $\beta_0 = 2$. The LBC estimator $\breve{\beta}$ here is calculated by taking one Newton step from $\tilde{\beta}_1$, and since it is visually indistinguishable from the NBC estimator $\ddot{\beta}_1$, only the latter is shown. The last two figures show the bias and RMSE of $\hat{\beta}_0$, $\tilde{\beta}_0$, and $\ddot{\beta}_0$, again for $\beta_0 = 2$ as a function of $\beta_1$. The principal impression we obtain from these figures is that bias correction works extremely well, in terms of both bias and RMSE.

The bias functions in Figures 8 and 9, and thus the results in Figures 10 through 13, depend on the distribution of the regressor as well as on the parameters. The theoretical results of Chesher and Peters (1994) and Chesher (1995) suggest that, when regressors are symmetrically distributed, bias functions may have rather special properties. We therefore ran some additional experiments in which the regressor was distributed as $\chi^2(5)$ and then recentered and rescaled to have mean 0 and variance 1. Results are shown in Figures 14 and 15, which are otherwise similar to Figures 10 and 11. The shape of the bias function for $\hat{\beta}_1$ is considerably more complicated than it was previously, but it still slopes upward and it is still not severely nonlinear. Once again, it appears that bias correction works extremely well, in terms of both bias and RMSE.

One aspect of these figures may at first seem a little strange. It is that the mean squared error of the CBC estimator $\tilde{\boldsymbol{\beta}}$ is consistently less than the mean squared errors of the LBC and NBC estimators, which, as the theory of Section 3 predicts, are practically identical. There are two reasons for this. The principal reason is that, as can be seen from equations (11) and (12) for the scalar case, the variance of the CBC estimator is always less than the variance of the other two estimators when the bias function is linear and not flat. This explains most of the difference.

A second, but quantitatively less important, reason for the smaller RMSE of the CBC estimator is that bias correction here is based on the approximate bias function $\boldsymbol{b}^a(\boldsymbol{\beta})$, not on the true bias function $\boldsymbol{b}(\boldsymbol{\beta})$. As a result, the LBC and NBC estimators exhibit somewhat more bias than the CBC one. We saw in Figures 8 and 9 that $\boldsymbol{b}^a(\boldsymbol{\beta})$ always underestimates the absolute bias. At the same time, because the bias function slopes upward for both parameters as functions of themselves, the result (9) suggests that the CBC estimator will tend to subtract an overestimate of the true absolute bias. In the case of the CBC estimator, these two sources of error largely offset each other. In contrast, the LBC and NBC estimators work almost exactly as they should if the true bias function were $\boldsymbol{b}^a(\boldsymbol{\beta})$. The slight bias they exhibit is a result of the discrepancy between $\boldsymbol{b}^a(\boldsymbol{\beta})$ and $\boldsymbol{b}(\boldsymbol{\beta})$.

These results suggest that using Amemiya's approximate bias function (34) in conjunction with the CBC estimator, which is precisely what Amemiya (1980) suggested doing, works very well indeed for the logit model. We are not aware of a similar approximate bias function for the probit model, and so simulation would presumably have to be used if we wished to obtain bias-corrected probit estimates.

## 7. Conclusions

Using methods based on evaluating the bias function to reduce bias is feasible and can be effective. However, reducing bias may increase the variance, or even the mean squared error, of an estimator. Whether such methods will be useful in practice depends on the shape of the bias function and on the variance of the initial estimator. Attempting to correct for bias mechanically, without knowing anything about the shape of the bias function, is clearly not a good idea. It may not always be feasible to investigate the shape of this function in detail, but it should always be possible to estimate the variance, or the covariance matrix, of the bias-corrected estimates, by making use of whichever of expressions (25), (26), (22), or (29) is appropriate.

If the bias function is approximately flat, bias correction is easy to do and should generally work well. If it is approximately linear, bias correction is still fairly easy to do, but it may not work well. In particular, if the bias function slopes downward, the bias-correcting estimators will have larger variances than the initial estimator, and they may therefore have larger mean squared errors. On the other hand, if the bias function slopes upward, the bias-correcting estimators will have smaller variances than the initial estimators. If the bias function is nonlinear, the three methods discussed in this paper can still be used, but they all yield estimates that are biased at $O(n^{-2})$. Since the biases of all the estimators in the nonlinear case depend on the variance of the initial estimator, bias correction is likely to be most effective when the bias is large relative to the variance of that estimator.

## References

Amemiya, T. (1980). "The $n^{-2}$-order mean squared errors of the maximum likelihood and the minimum logit chi-square estimator," *Annals of Statistics*, 8, 488–505.

Amemiya, T. (1985). *Advanced Econometrics*, Cambridge, Mass., Harvard.

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

Beran, R. (1987). "Prepivoting to reduce level error of confidence sets," *Biometrika*, 74, 457–468.

Chesher, A. (1995). “A mirror image invariance for $M$-estimators,” *Econometrica*, 63, 207–211.

Chesher, A., and S. Peters (1994). “Symmetry, regression design, and sampling distributions,” *Econometric Theory*, 10, 116–129.

Davidson, R., and J. G. MacKinnon (1992). “Regression-based methods for using control variates in Monte Carlo experiments,” *Journal of Econometrics*, 54, 203–222.

DiCiccio, T. J., and D. Efron (1996). “Bootstrap confidence intervals,” (with discussion), *Statistical Science*, 11, 189-228.

Efron, B., and R. J, Tibshirani (1993). *An Introduction to the Bootstrap*, New York, Chapman and Hall.

Gouriéroux, C., A. Monfort, and E. Renault (1993). “Indirect Inference,” *Journal of Applied Econometrics*, 8, S85–S118.

Gouriéroux, C., E. Renault and N. Touzi (1997). “Calibration by simulation for small sample bias correction,” in *Simulation-based Inference in Econometrics: Methods and Applications*, eds. R. Mariano, M. Weeks and T. Schuermann, Cambridge, Cambridge University Press.

Hall, P. (1992). *The Bootstrap and Edgeworth Expansion*, New York, Springer-Verlag.

Kendall, M. G. (1954). “Note on the bias in the estimation of autocorrelation,” *Biometrika*, 41, 403–404.

Kiviet, J. F., and G. D. A. Phillips (1993). “Alternative bias approximations in regressions with a lagged-dependent variable,” *Econometric Theory*, 9, 62–80.

Kiviet, J. F., and G. D. A. Phillips (1994). “Bias assessment and reduction in linear error-correction models,” *Journal of Econometrics*, 63, 215–243.

MacKinnon, J. G. and A. A. Smith, Jr. (1996). “Approximate bias correction in econometrics,” GREQAM Document de Travail No. 96A14.

Marriott, F. H. C., and J. A. Pope (1954). “Bias in the estimation of autocorrelations,” *Biometrika*, 41, 393–402.

Orcutt, G. H., and H. S. Winokur, Jr. (1969). “First order autoregression: inference, estimation, and prediction,” *Econometrica*, 37, 1–14.

Phillips, P. C. B. (1988). “The ET interview: Professor James Durbin,” *Econometric Theory*, 4, 125–157.

Sawa, T. (1978). “The exact moments of the least squares estimator for the autoregressive model,” *Journal of Econometrics*, 8, 159–172.

Smith, A. A., Jr. (1993). “Estimating nonlinear time-series models using simulated vector autoregressions,” *Journal of Applied Econometrics*, 8, S63–S84.

Smith, A. A., Jr., F. Sowell, and S. E. Zin (1997). “Fractional integration with drift: estimation in small samples,” *Empirical Economics*, 22, 103–116.

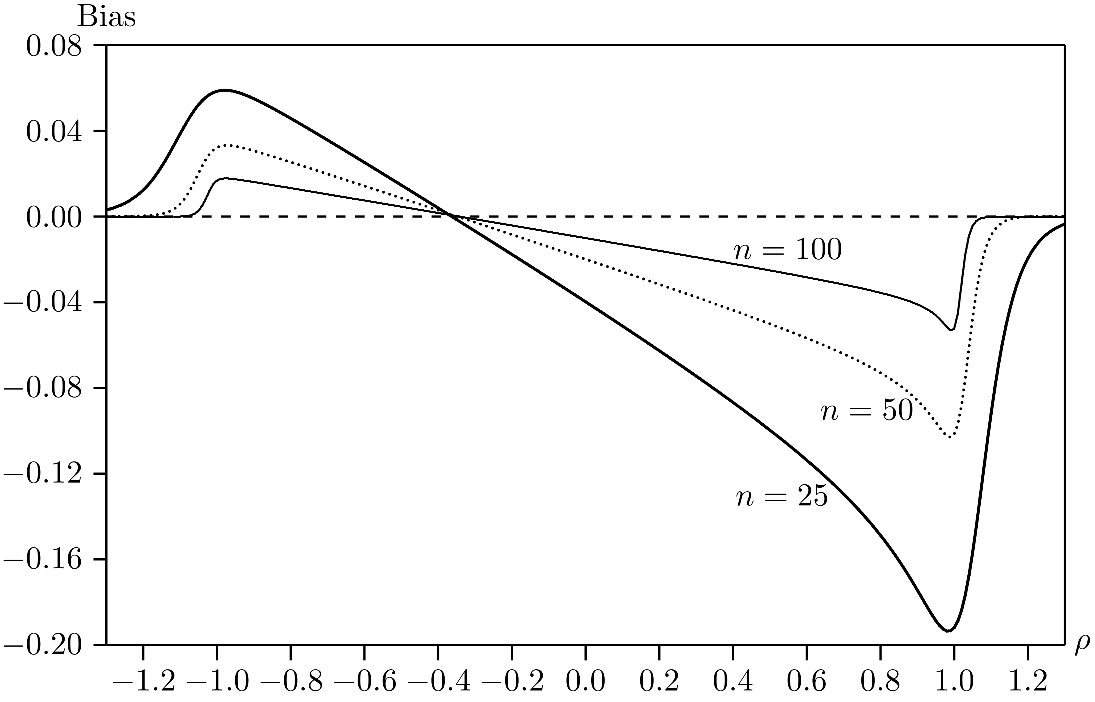


**Figure 1. Bias Functions, AR(1) Coefficient, Nonstationary Case**

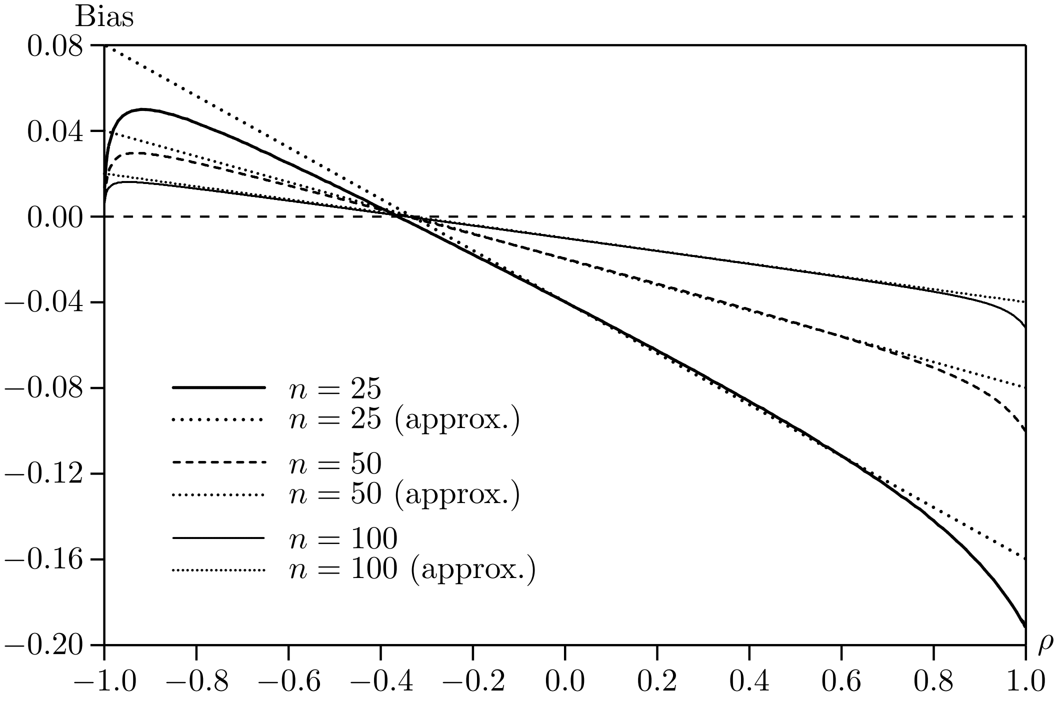


**Figure 2. Bias Functions, AR(1) Coefficient, Stationary Case**

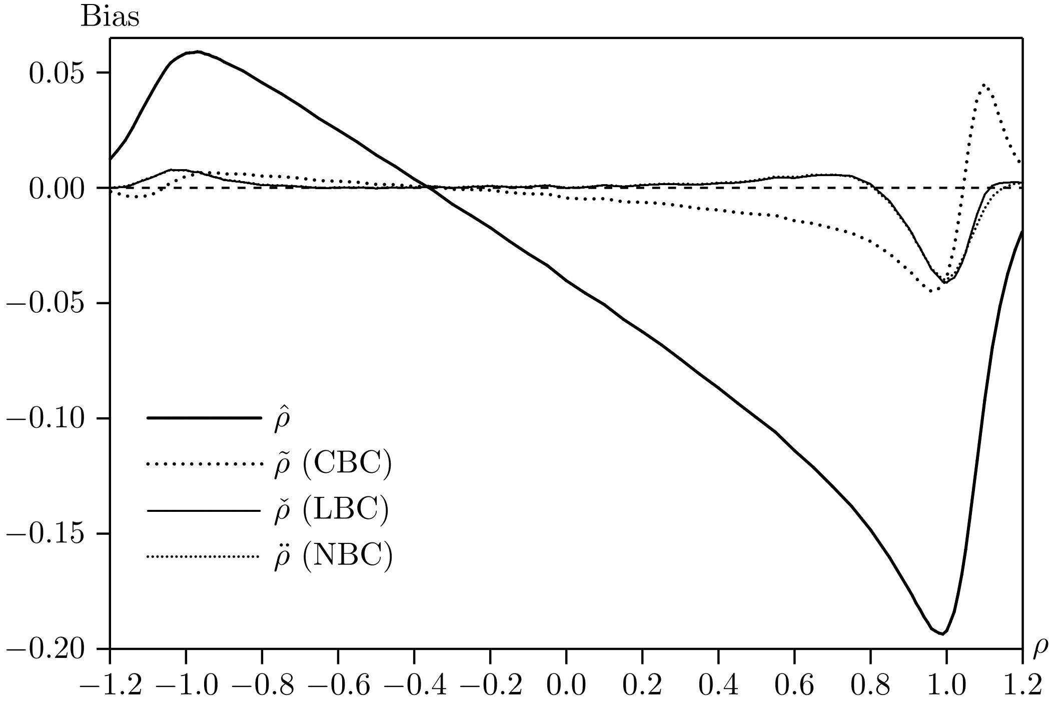


**Figure 3. Bias, $n = 25$**

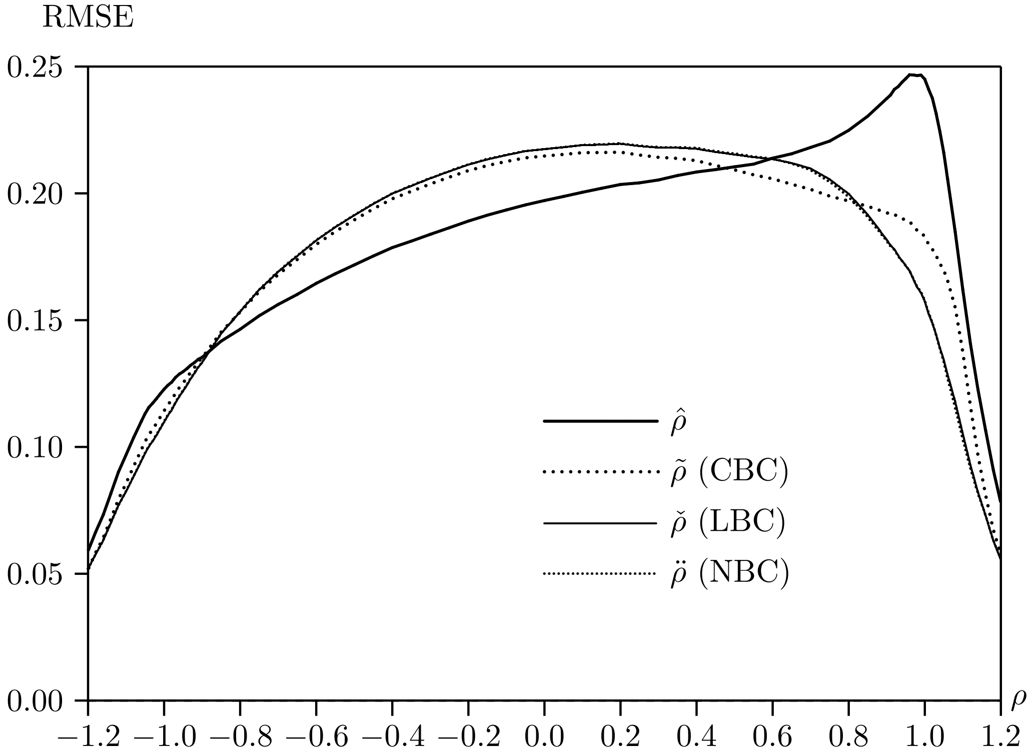


**Figure 4. Root Mean Squared Error, $n = 25$**

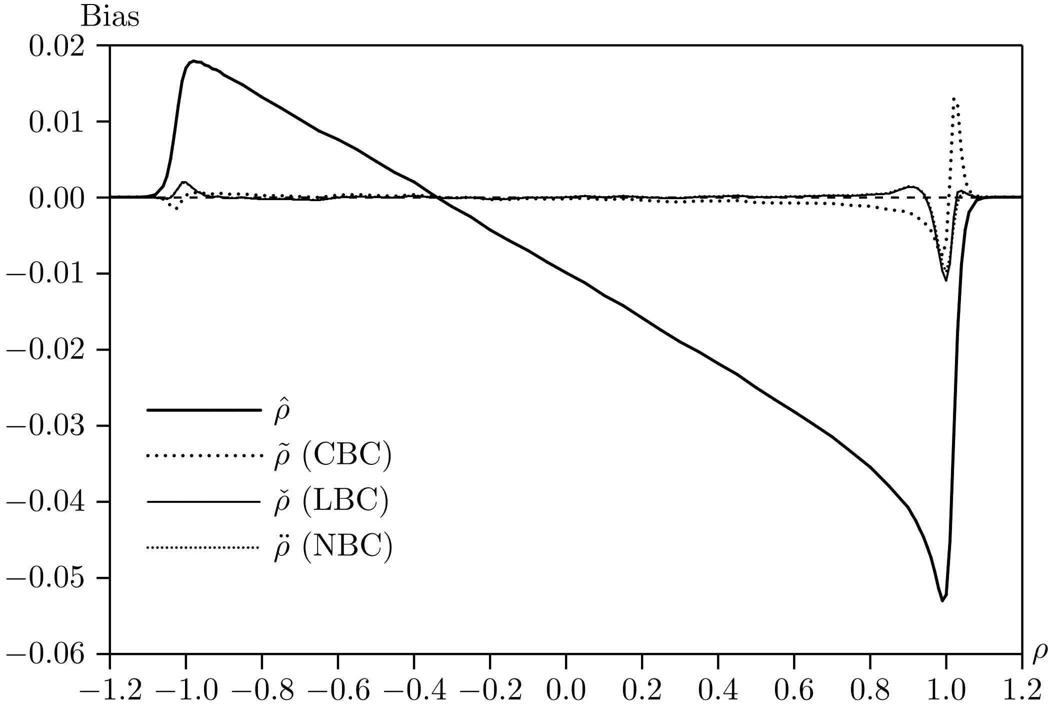


**Figure 5. Bias, $n = 100$**

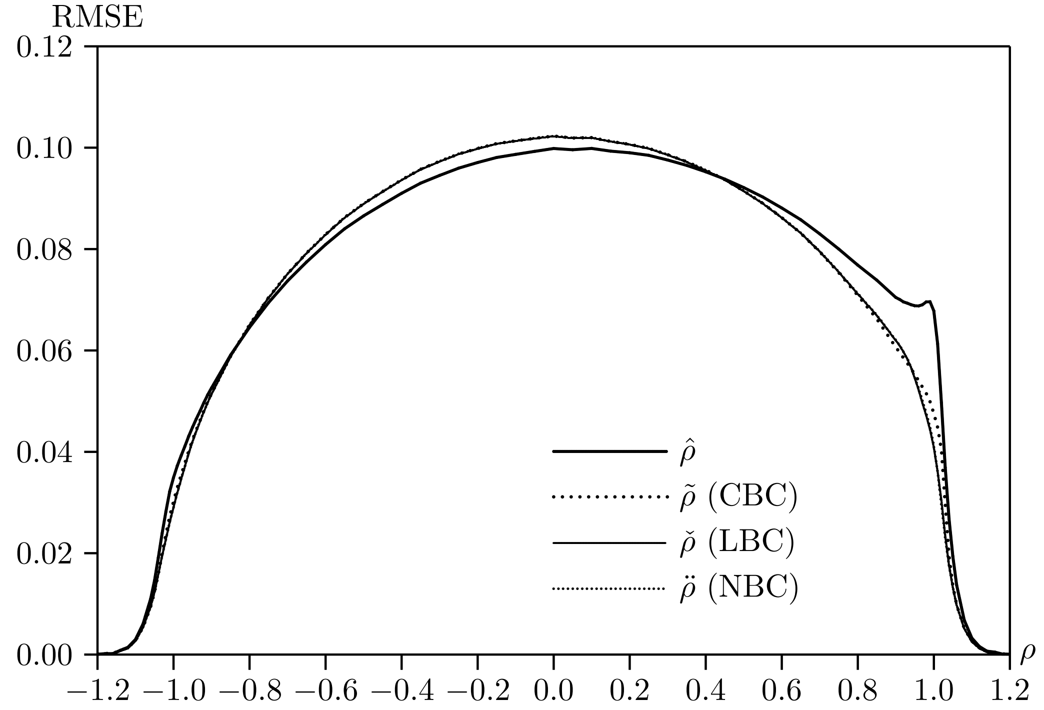
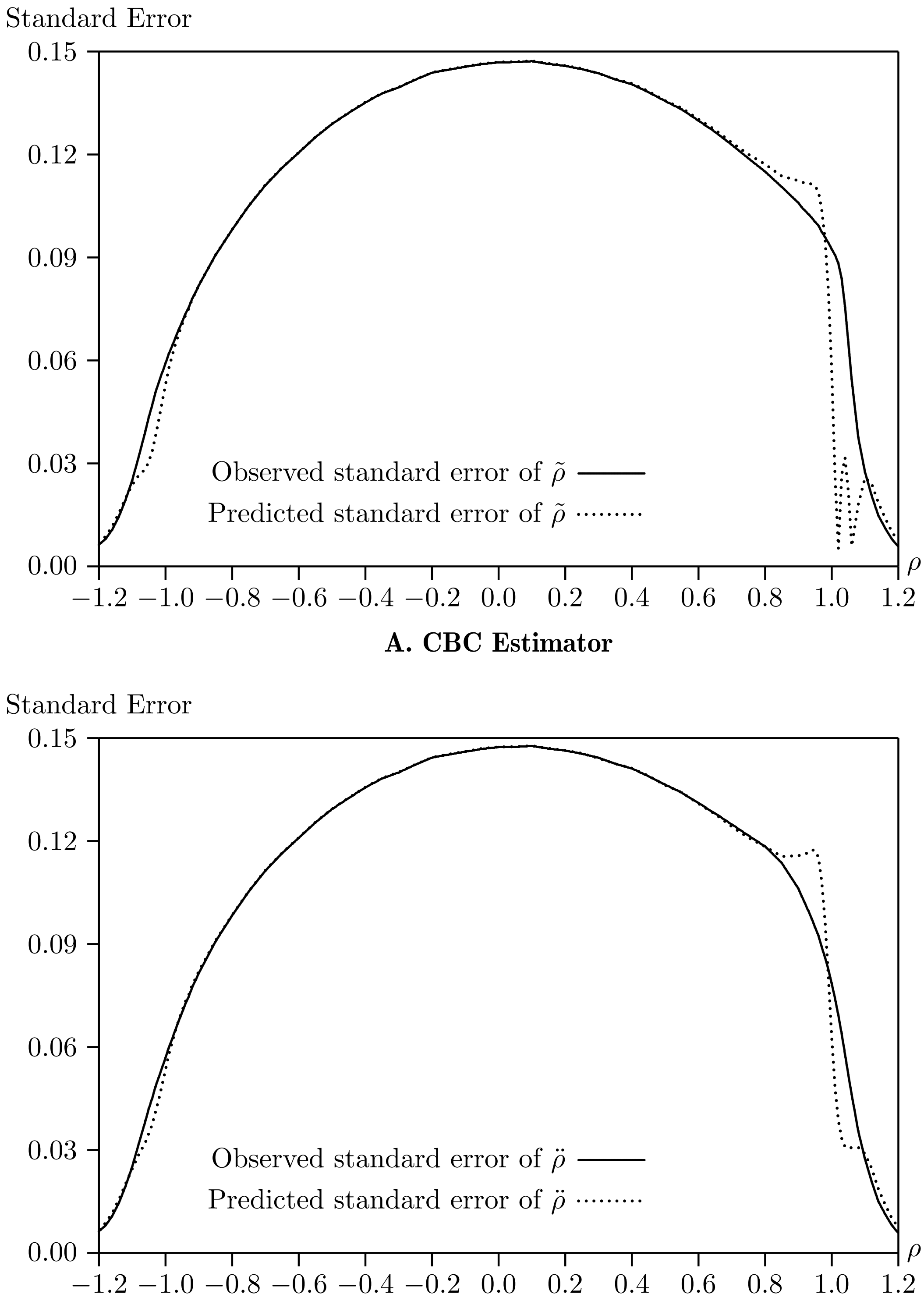


**Figure 6. Root Mean Squared Error, $n = 100$**



**A. CBC Estimator**

**B. NBC Estimator**

**Figure 7. Observed and Predicted Standard Errors, $n = 50$**

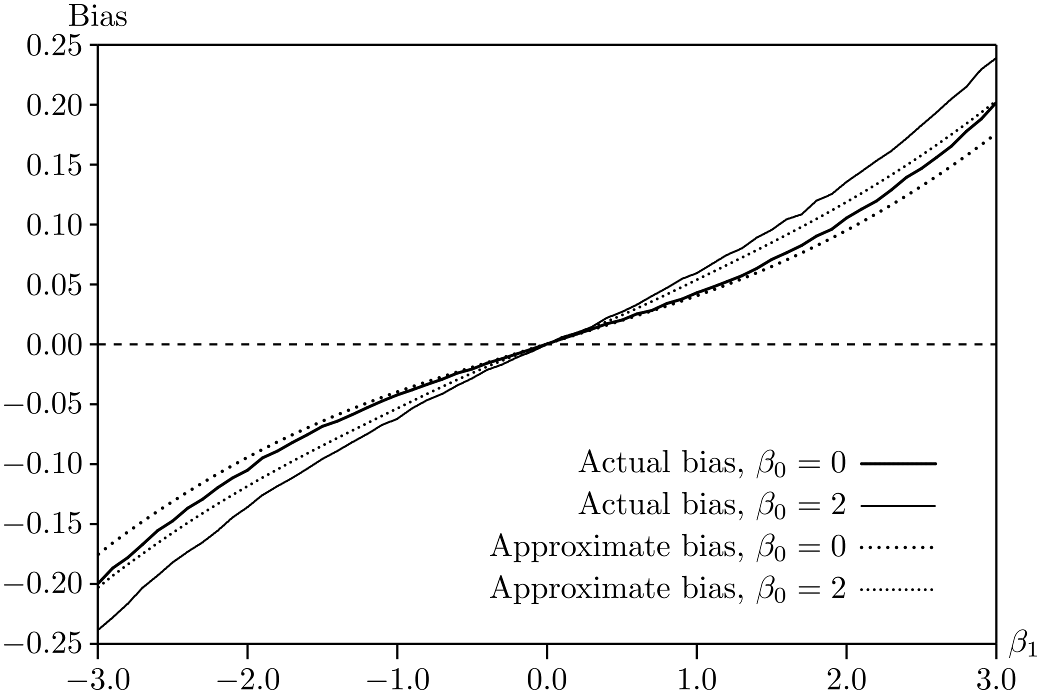


**Figure 8. Bias of Logit Slope Coefficient, $n = 100$**

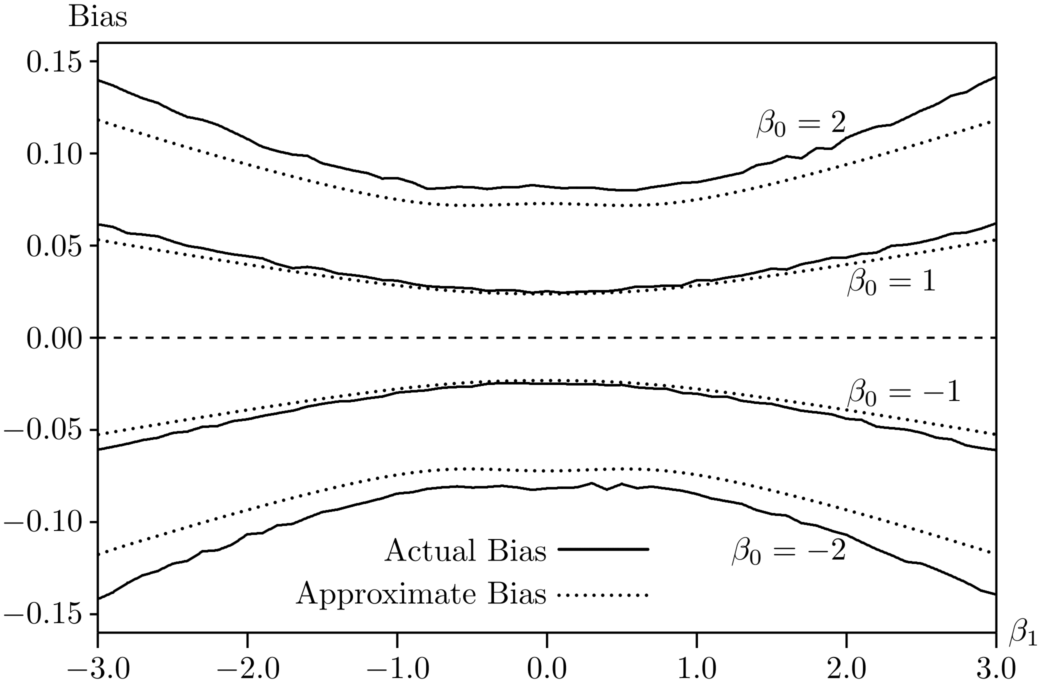


**Figure 9. Bias of Logit Constant Term, $n = 100$**

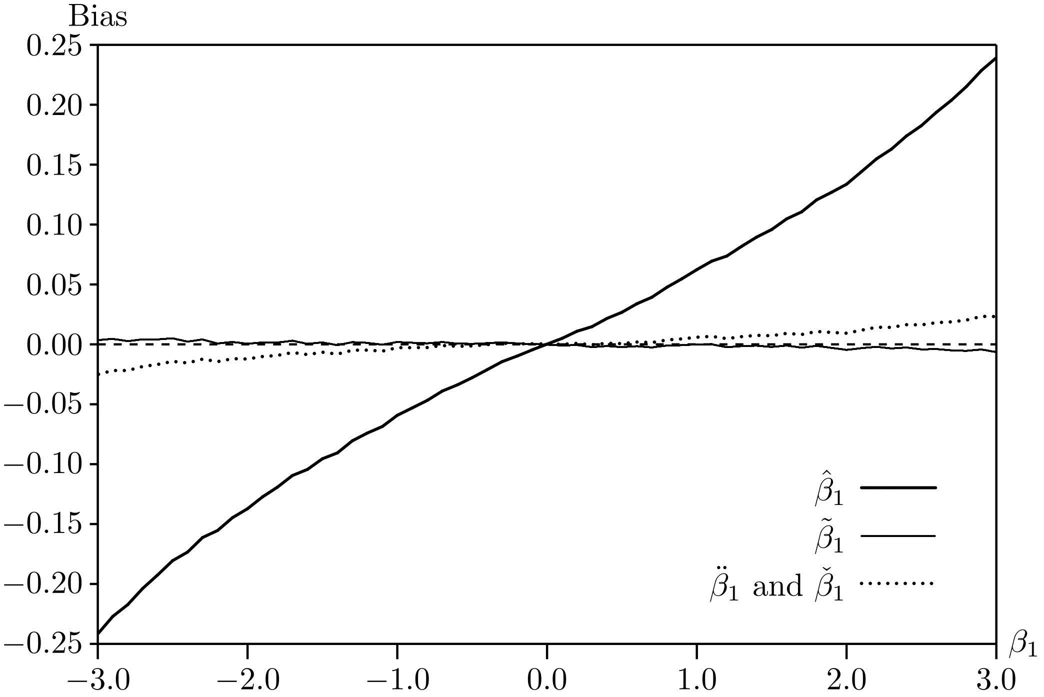


**Figure 10. Bias of Logit Slope Coefficient, $n = 100$, $\beta_0 = 2$**

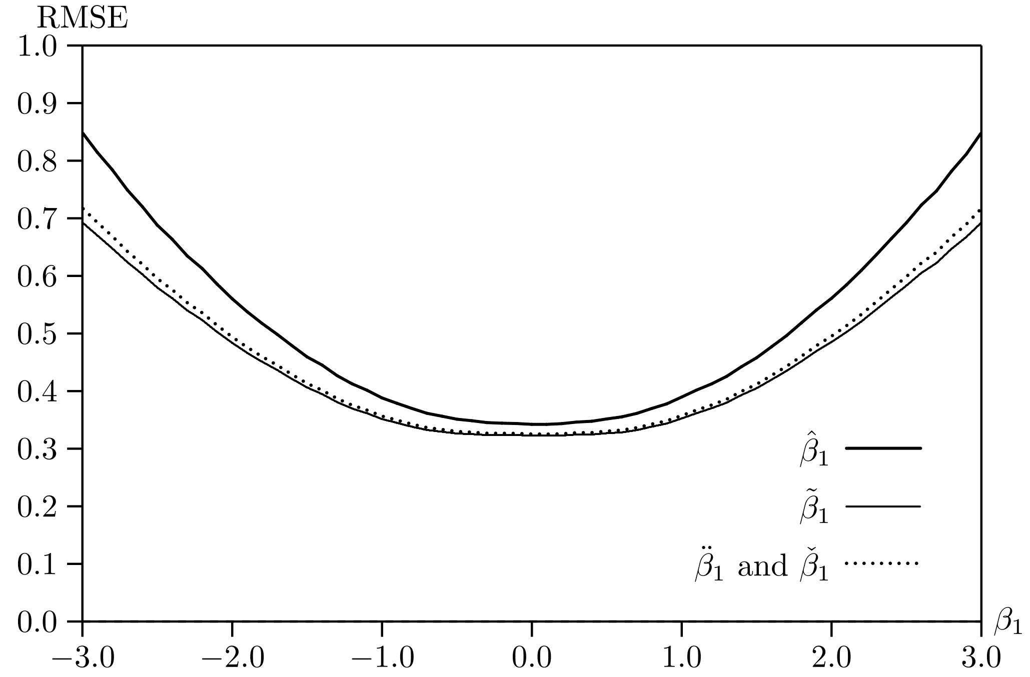


**Figure 11. RMSE of Logit Slope Coefficient, $n = 100$, $\beta_0 = 2$**

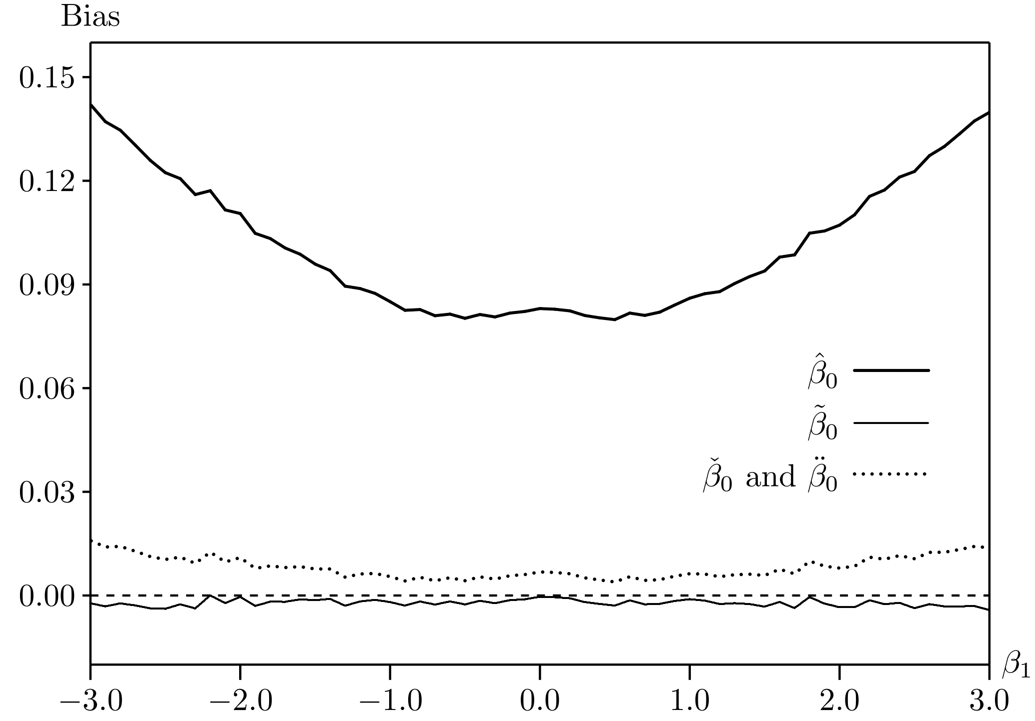


**Figure 12. Bias of Logit Constant Term, $n = 100$, $\beta_0 = 2$**

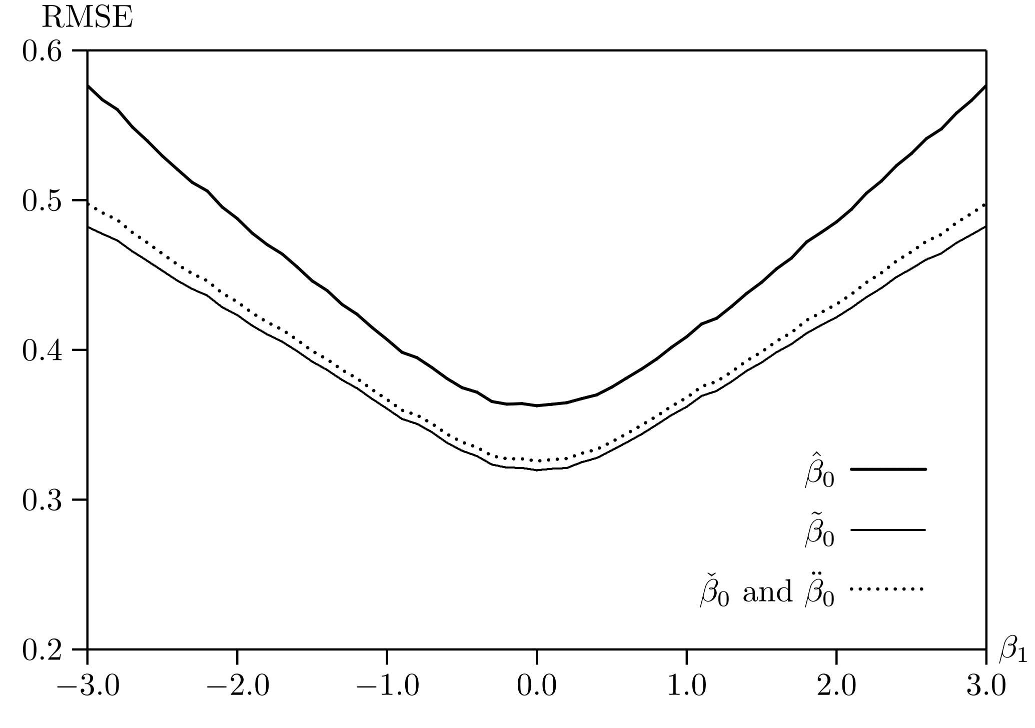


**Figure 13. RMSE of Logit Constant Term, $n = 100$, $\beta_0 = 2$**

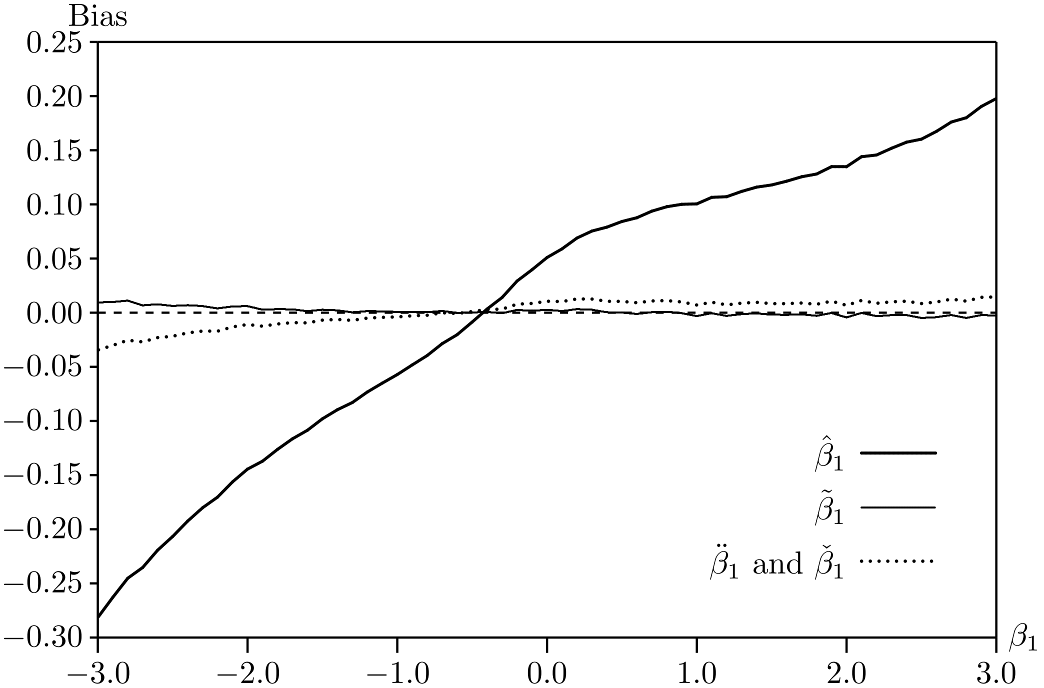


**Figure 14. Bias of Logit Slope Coefficient, $n = 100$, $\beta_0 = 2$**
**Asymmetric Regressor Case**

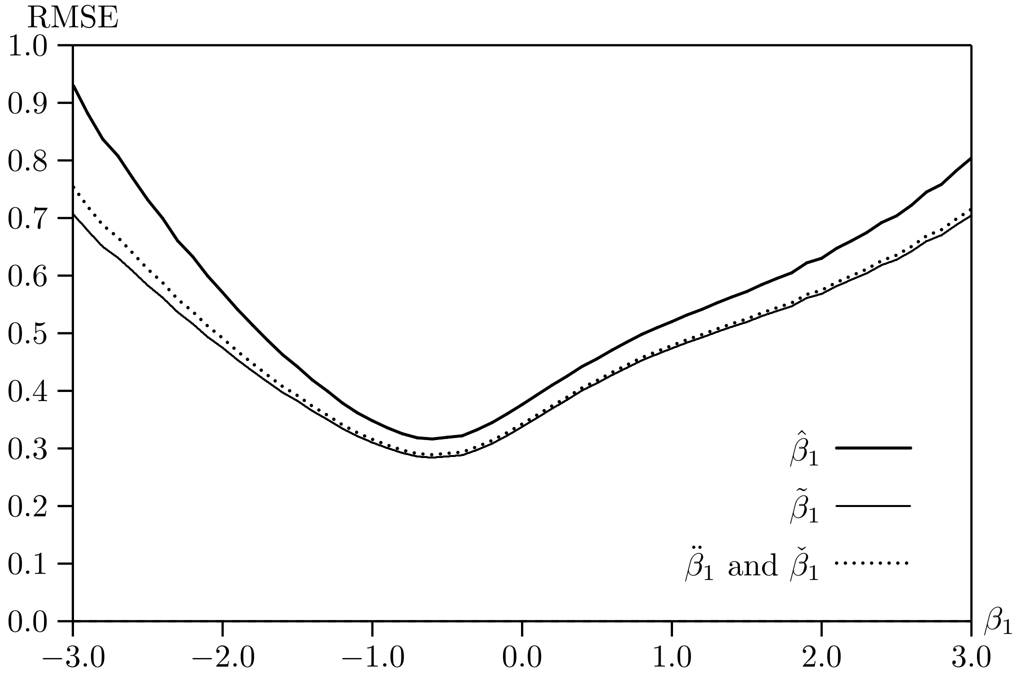


**Figure 15. RMSE of Logit Slope Coefficient, $n = 100$, $\beta_0 = 2$**
**Asymmetric Regressor Case**