Queen's Economics Department Working Paper No. 919

# Approximate Bias Correction in Econometrics

James G. MacKinnon Anthony A. Smith Jr.

Department of Economics
Queen's University
94 University Avenue
Kingston, Ontario, Canada
K7L 3N6

1-1995

Discussion Paper #919

# Approximate Bias Correction in Econometrics

by

James G. MacKinnon
Queen's University

and

Anthony A. Smith, Jr.
Carnegie Mellon University

January 1995

# 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, Fallaw Sowell, Amir Yaron, and Stanley E. Zin.

February, 1995

## Abstract

This paper discusses ways to reduce the bias of consistent estimators that are biased in finite samples. It is necessary that the bias function, which relates parameter values to bias, should be estimable by computer simulation or by some other method. If so, bias can be reduced or, in some cases that may not be unrealistic, even eliminated. In general, several evaluations of the bias function will be required to do this. Unfortunately, reducing bias may increase the variance, or even the mean squared error, of an estimator. Whether or not it does so depends on the shape of the bias function. The techniques 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 estimation of a logit model.

JEL Classification Number: C13

Keywords: bias function, mean squared error, simulation, finite samples

## 1. Introduction

Many econometric estimators are consistent but biased in finite samples. It is natural to try to "correct" 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. Nevertheless, despite the enormous improvements in computers in recent years, bias correction is rarely attempted in practice.

In this paper, we discuss ways in which finite-sample bias can be estimated, reduced, and in some 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, it may even be obtained analytically, at least up to some order of approximation; see Section 6. The paper has two principal results. First of all, bias correction generally seems to do a very good job of reducing bias, even when the bias functions are quite nonlinear. Secondly, although bias correction may often reduce the mean squared error of an estimator, it can, under some circumstances, increase it.

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 $\hat{\theta}^n$ 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**. Then we can always write

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

where $b(\theta_0, n) \equiv E(\hat{\theta}^n) - \theta_0$ and $v(\theta_0, n)$ is defined so that (1) holds. Thus $b(\theta_0, n)$ is the bias of $\hat{\theta}^n$ and $v(\theta_0, n)$ is the random difference between $\hat{\theta}^n$ 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 $y_t$'s. Thus we are able to estimate the bias function by simulation without using the bootstrap (Efron, 1982; Hall, 1992). The key feature of the bias function is that, in general, the bias of $\hat{\theta}$ (henceforth, we will suppress the superscript $n$) 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 nonstationary autoregressive model

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

We choose not to assume stationarity here for two reasons. First of all, the stationarity restriction that $|\rho| < 1$ makes bias correction more complicated. Secondly, because $\rho = 1$ implies the presence of a unit root, we did not want to rule out this interesting case. Because we are not assuming stationarity, we had to make an arbitrary assumption about starting values. We assumed that $y_0 = \mu + u_0$. Alternative assumptions would have resulted in slightly different bias functions.

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 $n = 100$. These functions do not depend on the values of $\mu$ and $\sigma^2$; see Appendix A of Andrews (1993). 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 have been possible to obtain these bias functions analytically (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 $|\rho| = 1$. Thus, in this example, it would never be reasonable to assume that the bias function is constant, but for many values of $\rho$ it might be reasonable to assume that it is linear.

In the next section, we consider what happens when the bias function is linear. This case is simple to deal with and may often be a good approximation. Then, in Section 3, we consider the more general case of a nonlinear bias function. Subsequently, Section 4 extends some 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. Estimation with Constant and Linear Bias Functions

The simplest case to deal with is the one in which the bias function is flat, so that $b(\theta, n) = b(n)$ for all $\theta$. In this case, we could estimate $b(n)$ simply by generating $N$ samples of size $n$ from the model that is hypothesized to have generated the $y_t$'s, using any value of $\theta$ at all. Although it does not matter what value of $\theta$ we use, the obvious one is $\hat{\theta}$. Let the average of the estimates obtained from the $N$ simulated samples be

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

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

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

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 simply be

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

We could then make inferences about $\theta$ by using the estimated asymptotic variance of $\hat{\theta}$. However, it seems more sensible to use the simulation results. We can estimate a confidence interval by using the empirical quantiles of the simulated quantities $\hat{\theta}_j - 2\hat{b}$. Note that we have to subtract $\hat{b}$ twice here, once to allow for the fact that the DGP for the simulation used the biased estimate $\hat{\theta}$ and once to allow for the fact that the $\hat{\theta}_j$'s are biased estimates of $\hat{\theta}$. The ends of a symmetric 95% confidence interval would be the .025 and .975 quantiles of $\hat{\theta}_j - 2\hat{b}$. It would also be possible to obtain non-symmetric confidence intervals, which could be shorter than symmetric ones if the distribution of $\hat{\theta}_j - 2\hat{b}$ were not symmetric.

When the bias function does depend on $\theta_0$, as it normally will, a single simulation will not allow us to obtain an unbiased estimate of $\theta$. When the bias function is linear in $\theta$, however, it is still quite easy to compute unbiased estimators by using simulation. Even if the bias function is not precisely linear, it may often be a reasonable approximation to assume that it is, and we shall make this assumption for the remainder of this section.

If the bias function is linear, we can write it as

$$b(\theta) = \alpha_1 + \alpha_2 \theta, \tag{5}$$

where we have suppressed the explicit dependence of $b(\cdot)$ on $n$. By evaluating (5) at two points, we can solve for $\alpha_1$ and $\alpha_2$. It seems logical that one point should be $\hat{\theta}$. A natural choice for the second point is the CBC estimator $\tilde{\theta}$, which was defined in (4). We need to do a second simulation experiment to obtain $\tilde{b}$. This experiment should be identical to the first one, except that the DGP must be evaluated at $\tilde{\theta}$ instead of $\hat{\theta}$. In order to ensure that the slope of the bias function is estimated accurately, both 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 $\alpha_1$ and $\alpha_2$. The results are

$$\acute{\alpha}_1 = \hat{b} - \left(\frac{\hat{b} - \tilde{b}}{\hat{\theta} - \tilde{\theta}}\right)\hat{\theta} \qquad \text{and} \qquad \acute{\alpha}_2 = \frac{\hat{b} - \tilde{b}}{\hat{\theta} - \tilde{\theta}}\,.$$

Under plausible conditions, $\acute{\alpha}_1$ and $\acute{\alpha}_2$ will converge to $\alpha_1$ and $\alpha_2$ as the number of simulations is increased. Now consider the estimator $\check{\theta}$ that is defined as the solution to the following equation:

$$\check{\theta} = \hat{\theta} - \acute{\alpha}_1 - \acute{\alpha}_2\check{\theta}. \tag{6}$$

Equation (6) simply says that $\check{\theta}$ is equal to $\hat{\theta}$ minus the bias function evaluated at $\check{\theta}$ itself. Solving (6) yields

$$\check{\theta} = \frac{1}{1 + \acute{\alpha}_2}(\hat{\theta} - \acute{\alpha}_1). \tag{7}$$

This is the **linear-bias-correcting**, or **LBC**, estimator.

Unlike the CBC estimator $\tilde{\theta}$, which is unbiased only when the bias function is flat, the LBC estimator $\check{\theta}$ will be unbiased whenever the bias function is linear. To see this, observe that

$$\begin{aligned} E(\check{\theta}) &= E(\hat{\theta} - b(\check{\theta})) \\ &= \theta_0 + b(\theta_0) - b(\theta_0) = 0. \end{aligned} \tag{8}$$

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

As with the CBC estimator, we will generally want to obtain a confidence interval for $\theta$ as well as an unbiased estimate, and there are numerous ways to do so. Since we have already done two simulations to obtain $\check{\theta}$, it seems natural to use simulation again. However, we will need to do one more simulation, because this time we must use $\check{\theta}$ to generate the data. On each replication of this simulation, we calculate

$$\check{\theta}_j = \frac{1}{1 + \acute{\alpha}_2}(\hat{\theta}_j - \acute{\alpha}_1).$$

From the empirical quantiles of the $\check{\theta}_j$'s, we can then calculate whatever confidence interval we are interested in.

The parameters of the bias function, $\acute{\alpha}_1$ and $\acute{\alpha}_2$, are treated as constants for the purposes of the simulation just described. Similarly, $\hat{b}$ was treated as a constant when confidence intervals for $\tilde{\theta}$ were discussed. Therefore, the confidence intervals that emerge will ignore any variation due to experimental error in the simulations. It would be possible to take account of this error, but only at the cost of considerable complexity. Since experimental error can be made arbitrarily small by making $N$ sufficiently large, there does not seem to be much point in worrying about it. Even when $N$ is only 1000, the standard errors of the simulation errors in estimating $b(\hat{\theta})$ and $b(\tilde{\theta})$ will be only about .032 times the standard error of $\hat{\theta}$. When $N = 10000$, the former will be only .01 times the latter. In many cases, it will be possible to obtain even smaller simulation errors without making $N$ unreasonably large by using control or antithetic variates; see Davidson and MacKinnon (1992). These methods tend to work particularly well when they are used to estimate the mean of a set of parameter estimates.

It is interesting to see how the variance of the LBC estimator $\breve{\theta}$ is related to the variance of the initial biased estimator $\hat{\theta}$. In the remainder of this section, for simplicity, we shall assume that $\acute{\alpha}_2 = \alpha_2$, ignoring the possible effects of experimental randomness. From (7), it is easy to see that

$$V(\breve{\theta}) = \frac{1}{(1+\alpha_2)^2} V(\hat{\theta}), \tag{9}$$

a result which holds whether or not the bias function is actually linear. Thus whether the variance of $\breve{\theta}$ will be greater than or less than that of $\hat{\theta}$ will depend on whether $\alpha_2$ is less than or greater than zero.[1]

If $\alpha_2 < 0$, it is quite possible for the unbiased LBC estimator $\breve{\theta}$ to have greater root mean squared error (RMSE) than the initial estimator $\hat{\theta}$. This will happen whenever

$$\frac{1}{(1+\alpha_2)^2} V(\hat{\theta}) > (\alpha_1 + \alpha_2 \theta_0)^2 + V(\hat{\theta}). \tag{10}$$

Of course, this equation is based on the assumption that the bias function is in fact linear. Notice that, if the variance of $\hat{\theta}$ is small enough or $\alpha_2 > 0$, condition (10) will never be satisfied. Thus bias correction can be expected to work relatively well when the bias function slopes upwards and when the variance of $\hat{\theta}$ is small relative to the bias.

The results (9) and (10) do not make sense if $\alpha_2 = -1$. It seems plausible to assume that $\alpha_2 > -1$, 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, it is certainly conceivable that this assumption could sometimes be false.

[1] Smith, Sowell, and Zin (1993) make a similar point.

It is interesting to look at the behavior of the CBC estimator $\tilde{\theta}$ when the bias function is actually linear. In this case, its bias will be

$$\begin{aligned} b(\theta_0) - E\big(b(\hat{\theta})\big) &= \alpha_1 + \alpha_2\theta_0 - \alpha_1 - \alpha_2 E(\hat{\theta} \,|\, \theta_0) \\ &= \alpha_2\theta_0 - \alpha_2(\theta_0 + \alpha_1 + \alpha_2\theta_0) \\ &= -\alpha_2(\alpha_1 + \alpha_2\theta_0) \\ &= -\alpha_2 b(\theta_0). \end{aligned} \tag{11}$$

Thus the bias of $\tilde{\theta}$ is $-\alpha_2$ times the bias of $\hat{\theta}$. Provided that $|\alpha_2| < 1$, the CBC estimator will be less biased than the initial estimator. It will be biased in the same direction when $\alpha_2 < 0$ and biased in the opposite direction when $\alpha_2 > 0$.

Since $\tilde{\theta} = \hat{\theta} - \alpha_1 - \alpha_2\hat{\theta}$, the variance of the CBC estimator is

$$V(\tilde{\theta}) = (1 - \alpha_2)^2 V(\hat{\theta}). \tag{12}$$

Thus, like the LBC estimator, the CBC estimator will have greater variance than the initial estimator when $\alpha_2 < 0$. If $\alpha_2 \neq 0$ and $|\alpha_2| < \sqrt{2}$, then $(1 - \alpha_2)^2 < 1/(1 + \alpha_2)^2$, which 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 the CBC estimator, causing it to have smaller RMSE than the LBC estimator. The condition for $\tilde{\theta}$ to have smaller RMSE than $\check{\theta}$ is

$$\frac{1}{(1+\alpha_2)^2} V(\hat{\theta}) > (1 - \alpha_2)^2 V(\hat{\theta}) + \alpha_2^2(\alpha_1 + \alpha_2\theta_0)^2. \tag{13}$$

The above results suggest that the CBC and LBC estimators may well have larger RMSE than the initial estimator, and that CBC may have smaller RMSE than LBC even though it is biased and LBC is not. Consider the bias functions in Figure 1, and suppose that $\rho = 0.4$ and $n = 25$. Then the bias of $\hat{\rho}$ is $-0.0869$, its variance is $0.0358$, and the value of $\alpha_2$ is approximately $-0.1257$. Therefore, by (11) and (12), the bias and variance of $\tilde{\rho}$ will be $-0.0109$ and $0.0454$. Similarly, by (9), the variance of $\check{\rho}$ will be $0.0469$. These numbers imply that RMSE($\hat{\rho}$) = 0.2083, RMSE($\tilde{\rho}$) = 0.2134, and RMSE($\check{\rho}$) = 0.2165. These theoretical results will be confirmed in Section 5; see Figure 3.

The results of this section make it 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. These results generalize easily to the case in which there is a vector of parameters to be estimated; see Section 4.

## 3. Estimation with a Nonlinear Bias Function

Bias functions are not always approximately linear. Therefore, the LBC estimator cannot be expected to remove all of the bias in all cases. In this section, we consider techniques for handling arbitrary nonlinear bias functions.

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

$$\ddot{\theta} = \hat{\theta} - b(\ddot{\theta}). \tag{14}$$

If we can solve (14), we can find the **nonlinear-bias-correcting**, or **NBC**, estimator $\ddot{\theta}$. However, there are two problems that do not arise in the linear case. First of all, $\ddot{\theta}$ will not be unbiased. Secondly, we must find some way to solve (14) numerically.

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 (14), as we did in (8), the nonlinearity of $b(\theta)$ implies that $E\big(b(\ddot{\theta})\big) \neq b(\theta_0)$. Although $\ddot{\theta}$ will, in general, be biased, there is reason to hope that the bias will often be small. If we take a second-order Taylor series expansion of (14) around $\theta_0$, we obtain

$$\ddot{\theta} \cong \hat{\theta} - b_0 - b_0'(\ddot{\theta} - \theta_0) - \tfrac{1}{2}b_0''(\ddot{\theta} - \theta_0)^2, \tag{15}$$

where $b_0$ denotes $b(\theta_0)$, and $b_0'$ and $b_0''$ denote the first and second derivatives of $b(\theta)$, evaluated at $\theta_0$. Taking expectations of both sides of (15), dividing through by $1 + b_0'$, and rearranging, we find that

$$E(\ddot{\theta}) - \theta_0 \cong -\frac{1}{2}\frac{b_0''}{1 + b_0'}E(\ddot{\theta} - \theta_0)^2. \tag{16}$$

This suggests, but does not guarantee, since (16) is only an approximation, that there will be no bias if the second derivative of $b(\theta)$ is zero near $\theta_0$. It also suggests that there will be bias if $b_0''$ is not zero, and that the sign of the bias will be opposite to that of $b_0''$. This bias will be of the same order as the square of the difference between $\ddot{\theta}$ and $\theta_0$. Thus, assuming that $\hat{\theta}$, and hence $\ddot{\theta}$, is root-$n$ consistent, the bias will be of order $1/n$.

The second problem is how to solve (14) without requiring very many evaluations of $b(\theta)$. 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 in many cases is the following. A key advantage of this technique is that it does not require the calculation of any derivatives of the bias function. First, find $\hat{b}$ as in (3). Then compute the sequence of estimates

$$\ddot{\theta}^{(j)} = (1-\gamma)\ddot{\theta}^{(j-1)} + \gamma(\hat{\theta} - b(\ddot{\theta}^{(j-1)})), \tag{17}$$

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 (14). However, it is certainly not guaranteed to converge. Whether or not it does so will depend on the shape of the bias function and on the parameter $\gamma$. 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.

This procedure has recently been used by Smith, Sowell, and Zin (1993) 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.

There is some similarity between what we are doing here and what Andrews (1993) recently did for a class of autoregressive models. Another way to write (14) is $\ddot{\theta} = h^{-1}(\hat{\theta})$, where $h(\theta) \equiv \theta + b(\theta)$. What we are doing is to invert 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 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 (14) by the difference between the median of $\hat{\theta}$ and $\theta_0$.

## 4. The Vector Case

In the preceding two sections, we have obtained three different bias-correcting estimators. 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, 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 computed in exactly the same way as before. We generate $N$ samples of size $n$ from the DGP with parameters $\hat{\boldsymbol{\theta}}$ and define $\acute{\boldsymbol{\theta}}$ as the mean of the estimates obtained from these samples. Then $\tilde{\boldsymbol{\theta}} \equiv 2\hat{\boldsymbol{\theta}} - \acute{\boldsymbol{\theta}}$.

When the bias function is flat, $\tilde{\boldsymbol{\theta}}$ will be unbiased and will have the same variance as the initial estimator $\hat{\boldsymbol{\theta}}$. More generally, suppose the bias function is linear, so that it can be written as

$$b(\boldsymbol{\theta}) = \boldsymbol{\alpha}_1 + \boldsymbol{\alpha}_2\boldsymbol{\theta}, \tag{18}$$

where $\boldsymbol{\alpha}_1$ is a $k-$vector and $\boldsymbol{\alpha}_2$ is a $k \times k$ matrix. Then, by almost the same algebra as (11), the bias of $\tilde{\boldsymbol{\theta}}$ is seen to be

$$-\boldsymbol{\alpha}_2(\boldsymbol{\alpha}_1 + \boldsymbol{\alpha}_2\boldsymbol{\theta}_0), \tag{19}$$

which is the vector analogue of the scalar result (11). Because, under (18), $\tilde{\boldsymbol{\theta}} = (\mathbf{I} - \boldsymbol{\alpha}_2)\hat{\boldsymbol{\theta}} - \boldsymbol{\alpha}_1$, the covariance matrix of $\tilde{\boldsymbol{\theta}}$ is easily seen to be

$$(\mathbf{I} - \boldsymbol{\alpha}_2)\boldsymbol{V}(\hat{\boldsymbol{\theta}})(\mathbf{I} - \boldsymbol{\alpha}_2)', \tag{20}$$

where $\boldsymbol{V}(\hat{\boldsymbol{\theta}})$ is the covariance matrix of $\hat{\boldsymbol{\theta}}$. This is the vector analogue of the scalar result (12).

The LBC estimator is also fairly easy to obtain when $\boldsymbol{\theta}$ is a vector. We simply have to evaluate $b(\boldsymbol{\theta})$ at $k + 1$ points in order to solve (18) for $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$. Which points are used will not matter if the bias function really is linear, but it may matter when it is not. As before, the LBC estimator will be unbiased if the bias function is linear. Since $\breve{\boldsymbol{\theta}} = (\mathbf{I} + \boldsymbol{\alpha}_2)^{-1}(\hat{\boldsymbol{\theta}} - \boldsymbol{\alpha}_1)$, the covariance matrix of $\breve{\boldsymbol{\theta}}$ is evidently

$$(\mathbf{I} + \boldsymbol{\alpha}_2)^{-1}\boldsymbol{V}(\hat{\boldsymbol{\theta}})\big((\mathbf{I} + \boldsymbol{\alpha}_2)'\big)^{-1}, \tag{21}$$

which is the vector analogue of (9).

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

$$\ddot{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}} - b(\ddot{\boldsymbol{\theta}}), \tag{22}$$

which is the vector version of (14). There are at least two ways to do this. One is to modify the iterative procedure (17) as follows:

$$\ddot{\boldsymbol{\theta}}^{(j)} = (1 - \boldsymbol{\gamma}){*}\ddot{\boldsymbol{\theta}}^{(j-1)} + \boldsymbol{\gamma}{*}\big(\hat{\boldsymbol{\theta}} - b(\ddot{\boldsymbol{\theta}}^{(j-1)})\big), \tag{23}$$

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

$$\ddot{\boldsymbol{\theta}}^{(j+1)} = \ddot{\boldsymbol{\theta}}^{(j)} - \big(\mathbf{I} + \boldsymbol{B}(\ddot{\boldsymbol{\theta}}^{(j)})\big)^{-1}\big(b(\ddot{\boldsymbol{\theta}}^{(j)}) + \ddot{\boldsymbol{\theta}}^{(j)} - \hat{\boldsymbol{\theta}}\big), \tag{24}$$

where $\boldsymbol{B}(\ddot{\boldsymbol{\theta}}^{(j)})$ is a $k \times k$ matrix of the derivatives of $b(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$, evaluated at $\ddot{\boldsymbol{\theta}}^{(j)}$. This matrix of derivatives would have to be evaluated numerically. For the first Newton step, $\ddot{\boldsymbol{\theta}}^{(0)}$ would equal $\hat{\boldsymbol{\theta}}$, and this step would yield an estimator similar to the LBC estimator, although not identical to it when the bias function is nonlinear.

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

The three bias-correcting estimators proposed in Sections 1, 2, and 3 were applied to the estimation of $\rho$ in the 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 $\rho$:

$$\begin{aligned}\rho = &-1.20, -1.18, \ldots, -1.06, -1.05, -1.04, \ldots, -.90, -.85,\\ &\ldots, .90, .91, \ldots, 1.05, 1.06, 1.08, \ldots, 1.20\ .\end{aligned}$$

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

Figures 2 and 4 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. 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 for values of $\rho$ near $\pm 1$, however. Of course, this is where the bias functions are severely nonlinear.

In Section 3, we showed that for a linear bias function, $\tilde{\rho}$ will be less biased than $\hat{\rho}$, provided the condition $|\alpha_2| < 1$ is satisfied. 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 $\hat{\rho}$.

Interestingly, 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. Thus, in this case, there seems to be little to gain by using the NBC estimator rather than the simpler LBC one.

Figures 3 and 5 show the root mean squared errors 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 bias-correcting estimators for $\rho$ between about $-0.9$ and $0.5$. Only for values of $\rho$ greater than about 0.8 do the bias-correcting 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. Since $\breve{\rho}$ and $\ddot{\rho}$ perform almost identically, as the near linearity of the bias function suggests that they should, equations (9) and (12) should be applicable. According to equation (9), the variance of $\check{\rho}$ should be equal to $(1+\alpha_2)^{-2}$ times the variance of $\hat{\rho}$. Similarly, equation (12) implies that the variance of $\tilde{\rho}$ should be equal to $(1-\alpha_2)^2$ times the variance of $\hat{\rho}$. Since $\alpha_2$ is always negative, this implies that the variance of all the bias-correcting estimators will exceed the variance of the OLS estimator. Only for large values of $\rho$, where the bias of $\hat{\rho}$ is large, does the reduced bias of the bias-correcting estimators outweigh their increased variance.

It is worth pointing out that equations (9) and (12) do a remarkably good job of explaining what we see in Figures 3 and 5 for most values of $\rho$. Figure 6 plots the observed standard errors of $\tilde{\rho}$ and $\check{\rho}$ for $n = 50$. It also plots what those standard errors should be according to equations (9) and (12), which assume that the bias function is linear. 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 bias corrected estimator, and which one, it may often be reasonable to rely on equations (9) and (12).

Another interesting feature of the experiments is that the iterative procedure based on (17) worked extremely well. For most values of $\rho$, the procedure converged in eight 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

The methods of this paper may be particularly attractive in the case of binary response models such as the logit model. The logit model may be written as

$$E(y_t) = P_t(\boldsymbol{X}_t\boldsymbol{\beta}) \equiv \bigl(1+\exp(-\boldsymbol{X}_t\boldsymbol{\beta})\bigr)^{-1}, \tag{25}$$

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 (25), this approximation can be written as

$$\boldsymbol{b}^a(\boldsymbol{\beta}) \equiv \tfrac{1}{2}(\boldsymbol{X}'\boldsymbol{\Omega}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{d}, \tag{26}$$

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

$$d_t \equiv (2P_t - 1)\big[\boldsymbol{\Omega}^{1/2}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{\Omega}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Omega}^{1/2}\big]_{tt}. \tag{27}$$

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 (27) need to be calculated, the approximation (26) 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 $P_t$'s.

It is much more attractive to obtain bias-corrected estimates by using the approximate bias function $\boldsymbol{b}^a(\boldsymbol{\beta})$ defined in (26) and (27) than by using a bias function obtained by simulation. This is true for several reasons. First, simulation would be very much more computationally expensive than evaluating $\boldsymbol{b}^a(\boldsymbol{\beta})$. Second, when simulation is used, the estimated bias function $\acute{\boldsymbol{b}}(\boldsymbol{\theta})$ 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 $y_t$'s in the simulated samples at all. When this happens, the estimates will not change, and the slope of the simulated 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.

Figure 7 shows the actual and approximate bias of the slope coefficient $\beta_1$ as a function of itself in a two-parameter logit model. The constant term $\beta_0$ is equal to either 0 or 2, and the only regressor is distributed as $N(0,1)$. These 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 7, 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 7 has several interesting features. The bias function for the ML estimate of the slope coefficient $\hat{\beta}_1$ slopes upwards, 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 8 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 7. From Figure 8, 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 7 and 8 suggest that bias correction should work very well for this logit model. This is in fact the case, as can be seen from Figures 9, 10, 11, and 12, which are based on 200,000 independent replications. 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 $\check{\beta}$ 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 7 and 8, and thus the results in Figures 9 through 12, 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 13 and 14, which are otherwise similar to Figures 9 and 10. 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 are practically identical. There are two reasons for this. The principal reason is that, as can be seen from (9) 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 7 and 8 that $\boldsymbol{b}^a(\boldsymbol{\beta})$ always underestimates the absolute bias. At the same time, because the bias function slopes upwards for both parameters as functions of themselves, the result (11) 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 (26) in conjunction with the CBC estimator 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

It seems clear that using methods based on evaluating the bias function to reduce bias is feasible and can be effective. Whether such methods will be useful in practice depends on the shape of the bias function and the variance of the initial estimator. 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 down, 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 up, the bias-correcting estimators will have smaller variances than the initial estimators. If the bias function is severely nonlinear, bias correction is harder to do and generally cannot eliminate all bias. Since the problems encountered in the linear and nonlinear cases arise from the variance in 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.

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

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.

Efron, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans*, Philadelphia, Society for Industrial and Applied Mathematics.

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

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., F. Sowell, and S. E. Zin (1993). "Fractional integration with drift: estimation in small samples," manuscript, Carnegie-Mellon University.

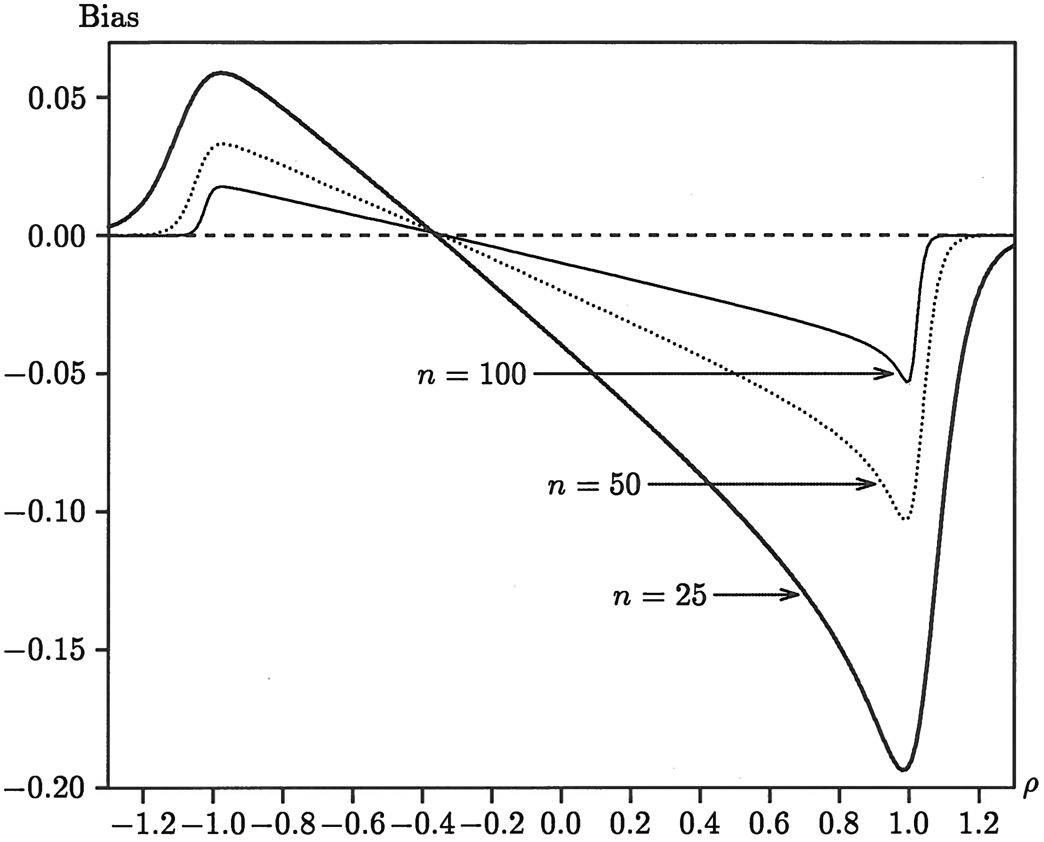

**Figure 1. Bias Functions, AR(1) Coefficient**

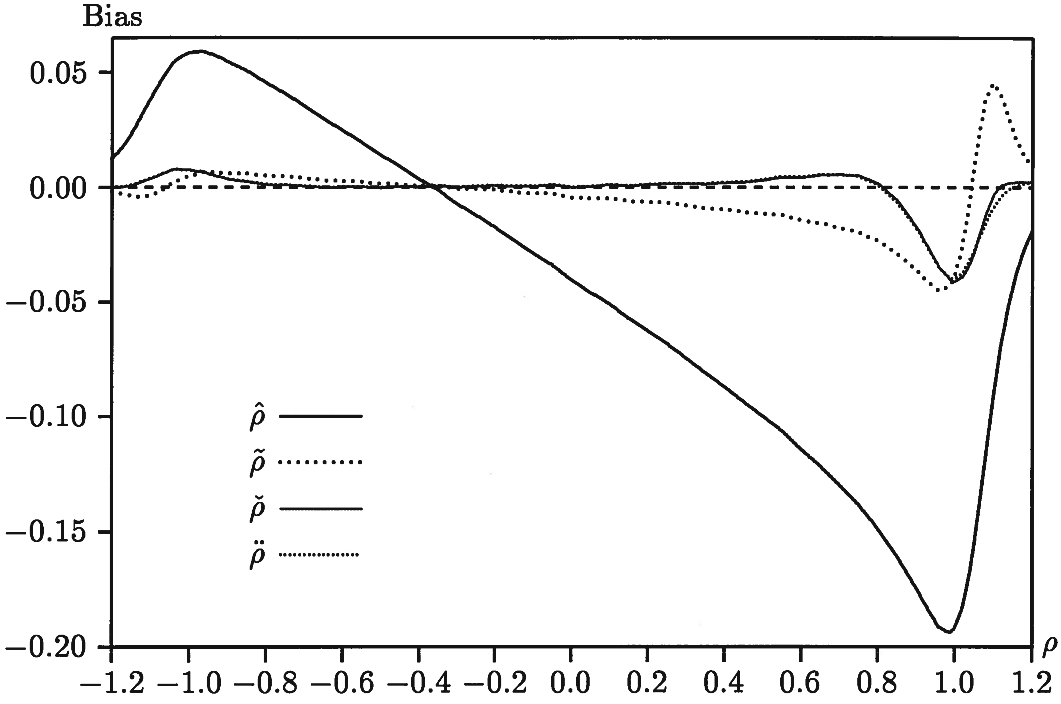

**Figure 2. Bias, $n = 25$**

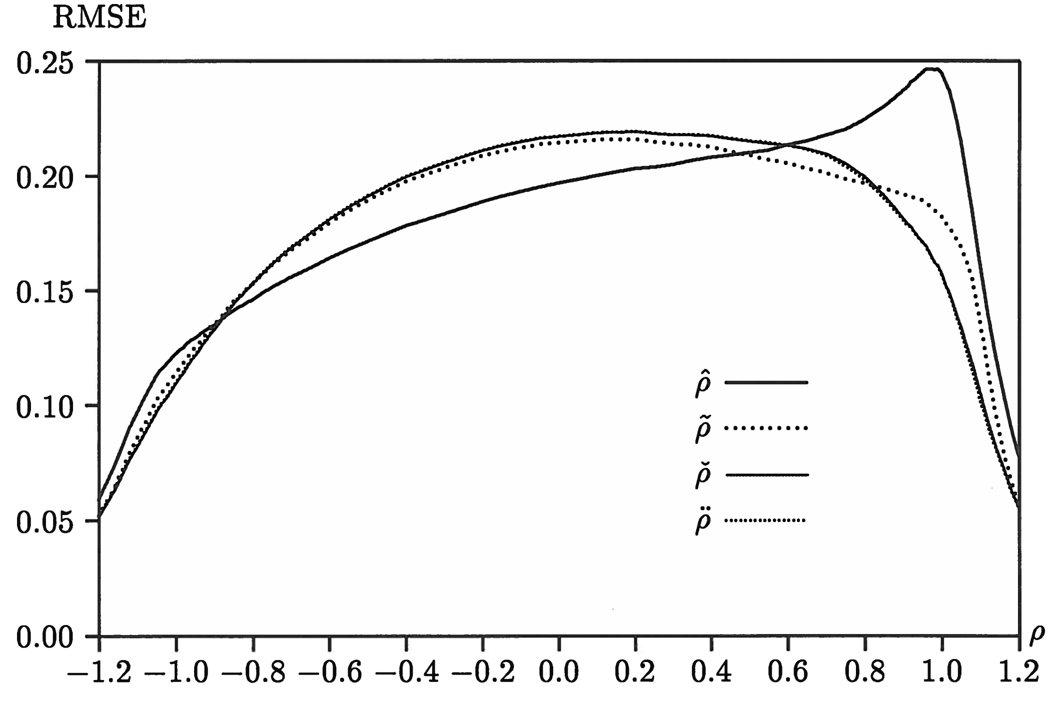

**Figure 3. Root Mean Square Error, $n = 25$**

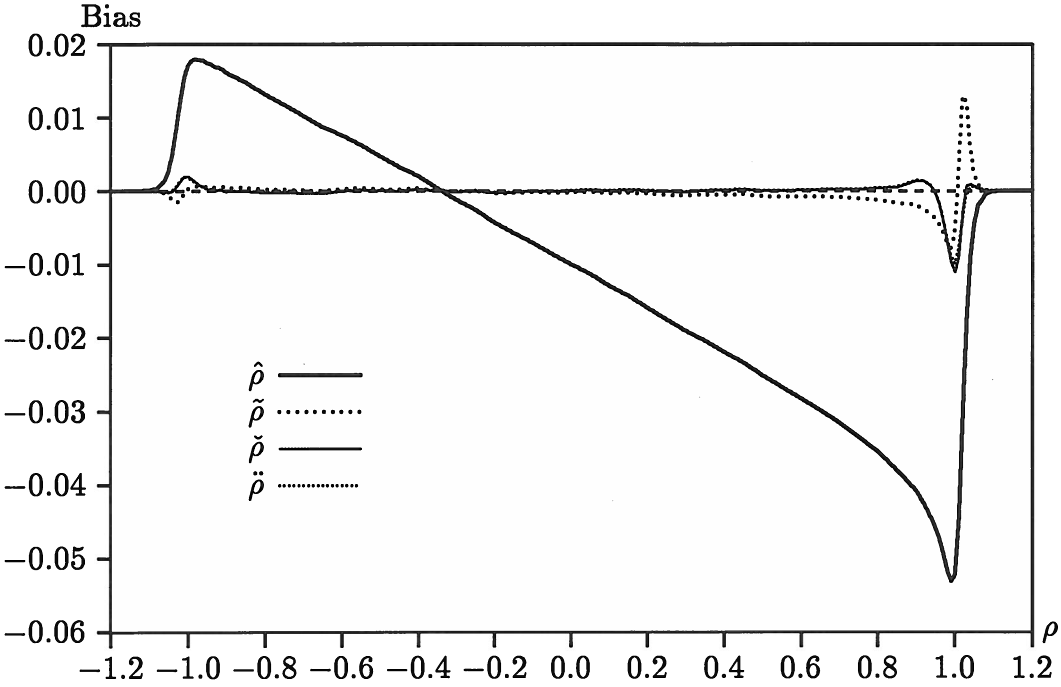

**Figure 4. Bias, $n = 100$**

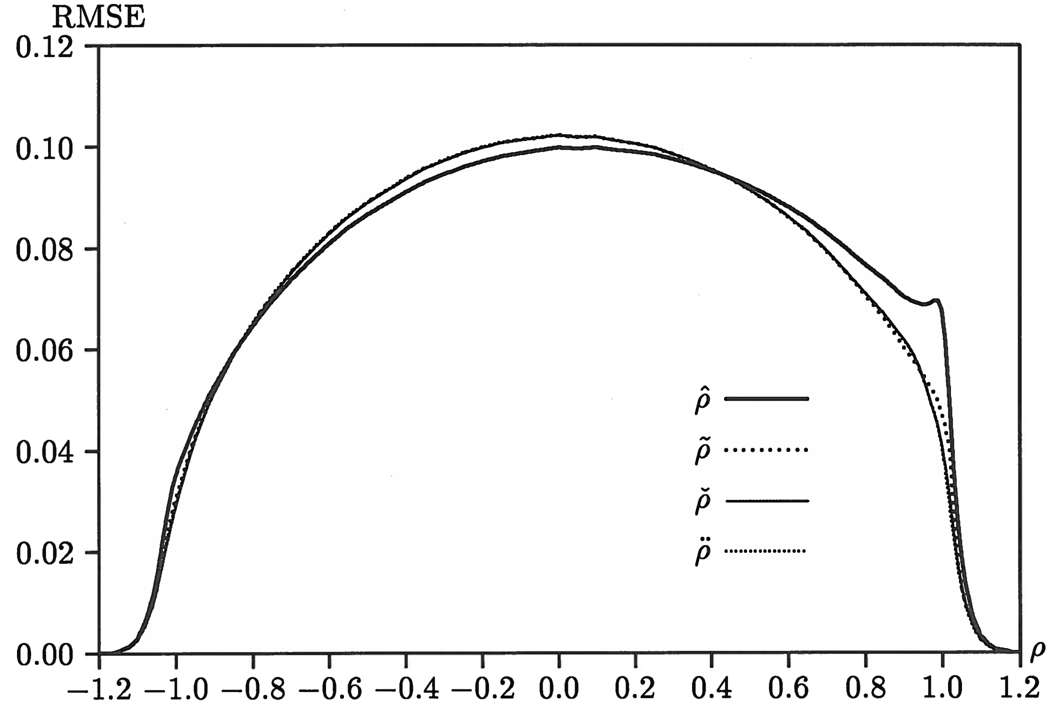

**Figure 5. Root Mean Square Error, $n = 100$**

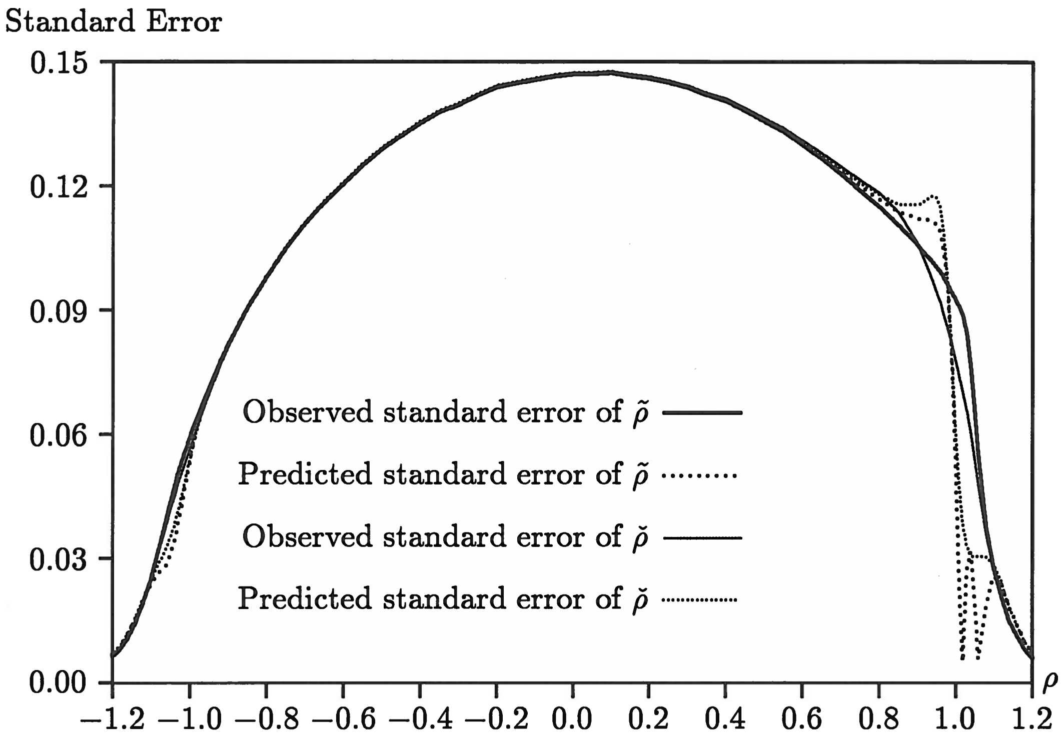

**Figure 6. Observed and Predicted Standard Errors, $n = 50$**

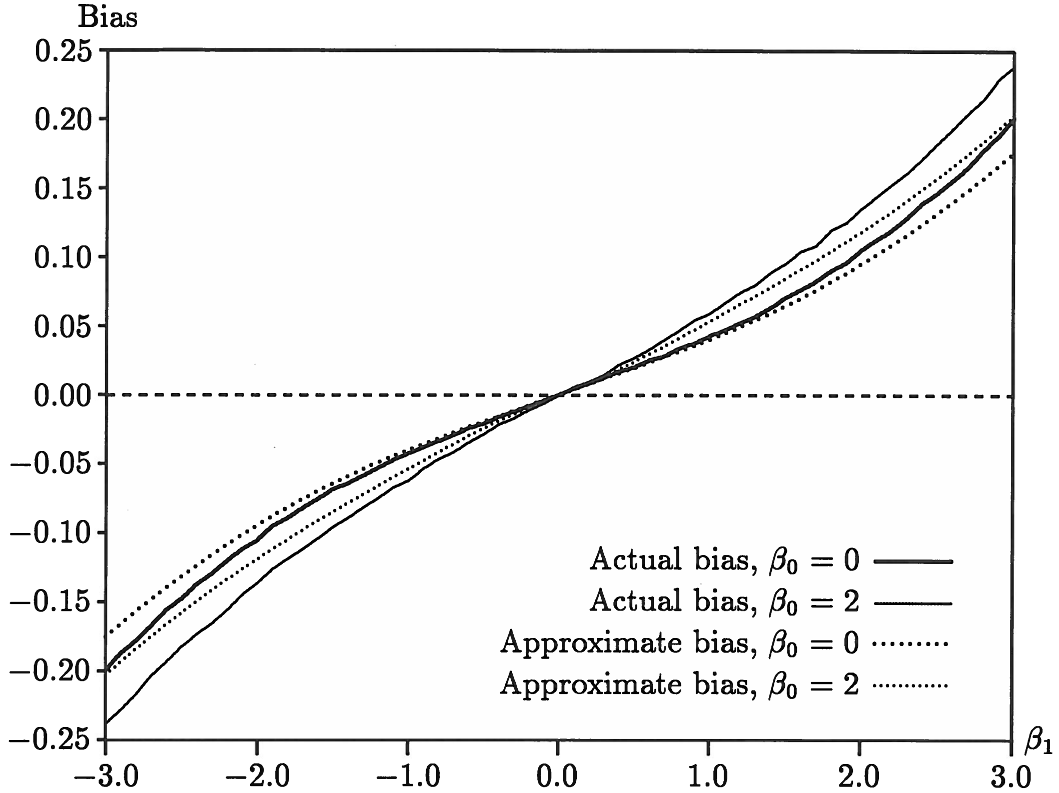

**Figure 7. Bias of Logit Slope Coefficient, $n = 100$**

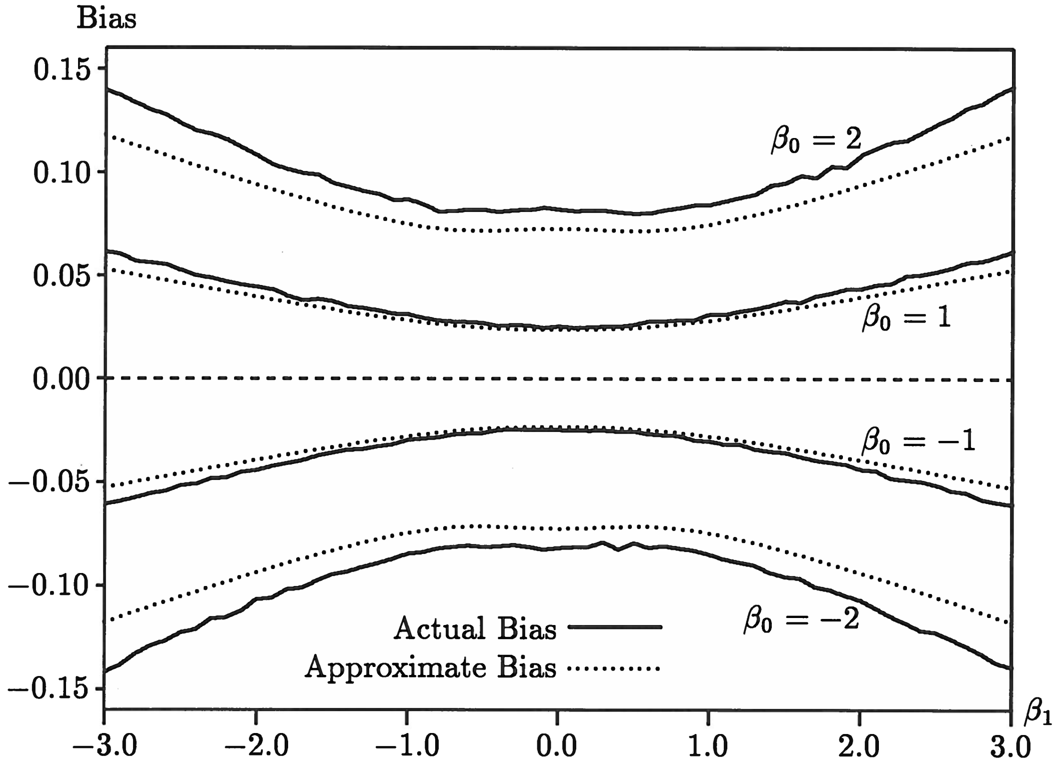

**Figure 8. Bias of Logit Constant Term, $n = 100$**

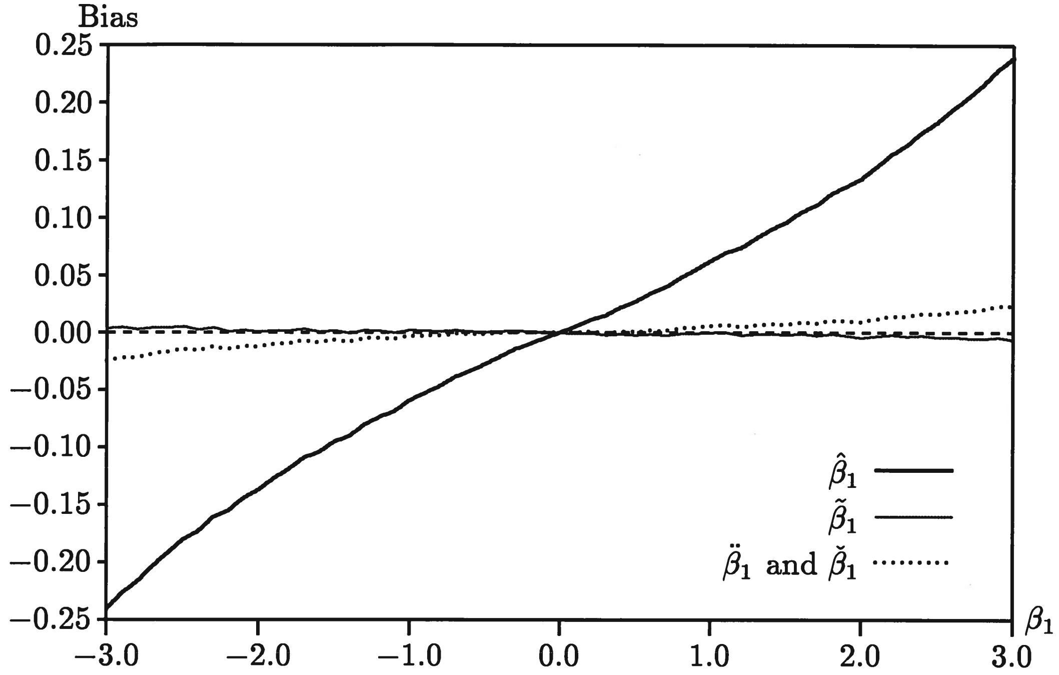

**Figure 9. Bias of Logit Slope Coefficient, $n = 100$, $\beta_0 = 2$**

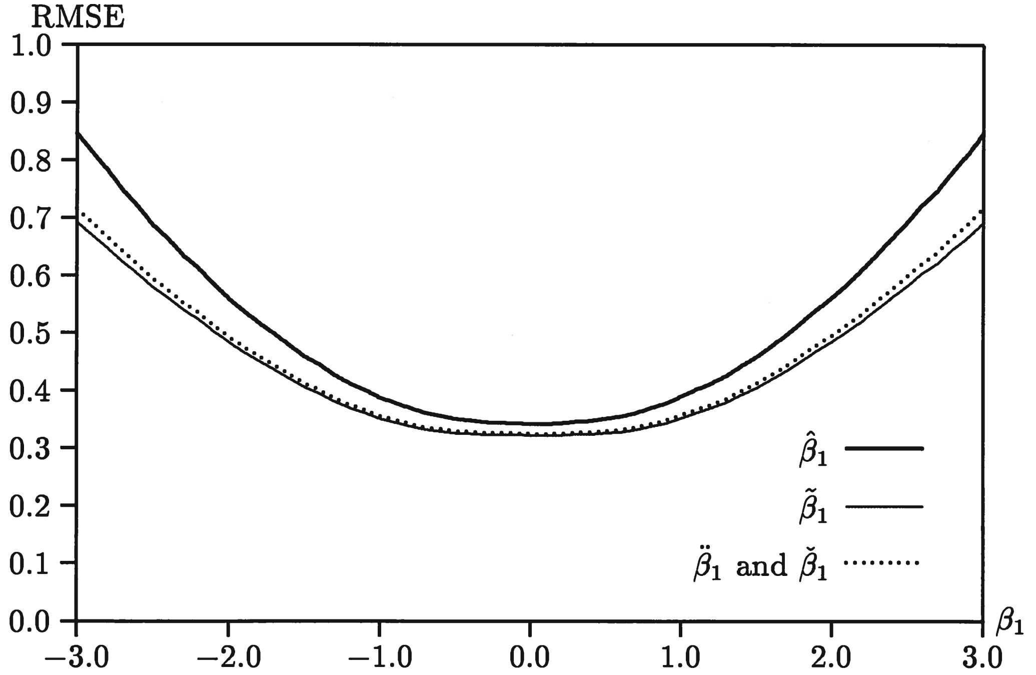

**Figure 10. RMSE of Logit Slope Coefficient, $n = 100$, $\beta_0 = 2$**

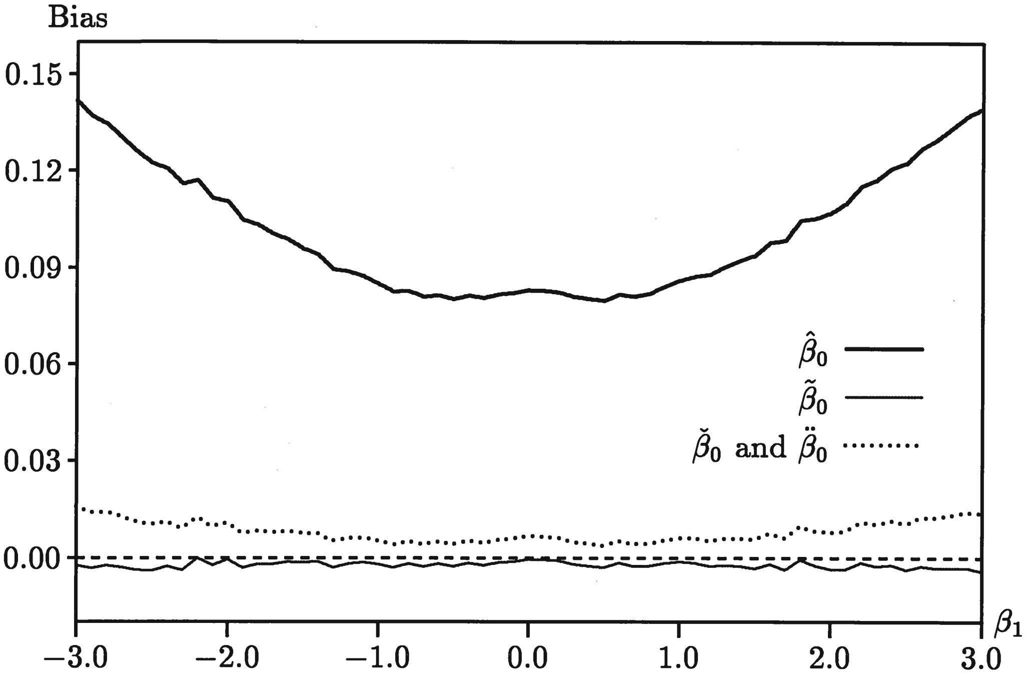

**Figure 11. Bias of Logit Constant Term, $n = 100$, $\beta_0 = 2$**

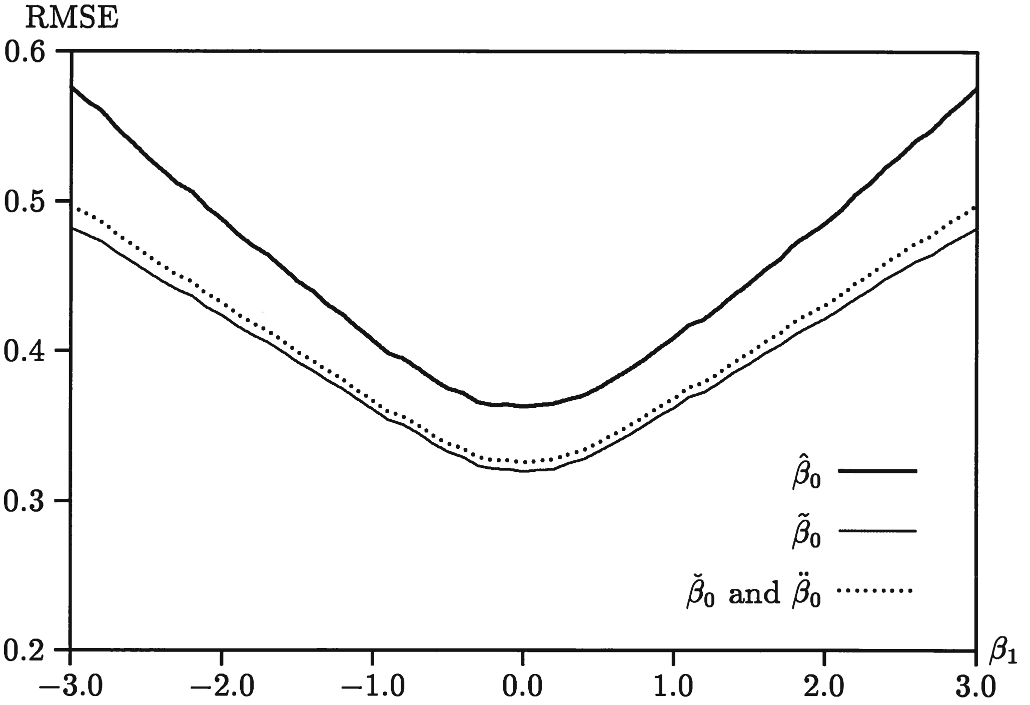

**Figure 12. RMSE of Logit Constant Term, $n = 100$, $\beta_0 = 2$**

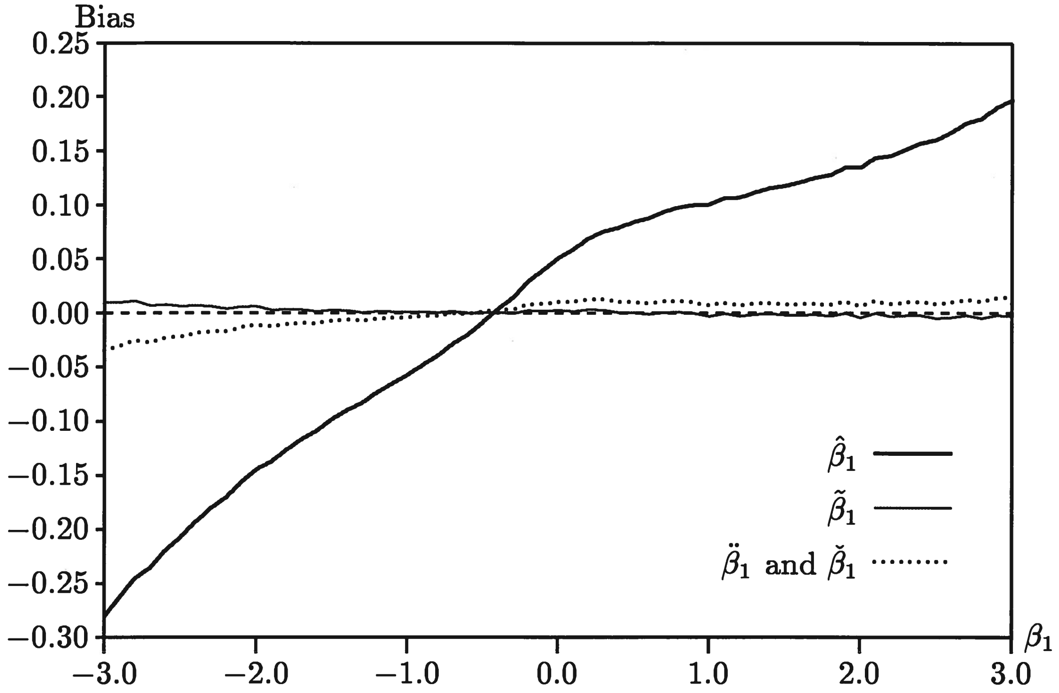

**Figure 13. Bias of Logit Slope Coefficient, $n = 100$, $\beta_0 = 2$**
**Asymmetric Regressor Case**

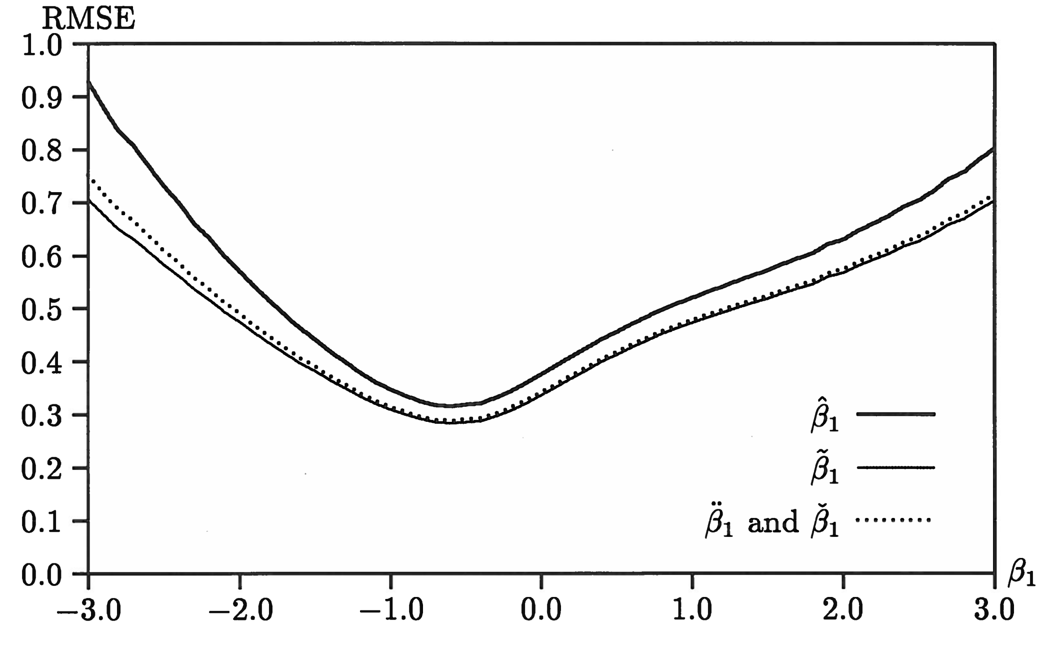

**Figure 14. RMSE of Logit Slope Coefficient, $n = 100$, $\beta_0 = 2$**
**Asymmetric Regressor Case**