

# Omitted Variables Bias in Regime-Switching Models with Slope-Constrained Estimators: Evidence from Monte Carlo Simulations

Ermanno Affuso<sup>1</sup>, Steven B. Caudill<sup>2</sup> and Franklin G. Mixon, Jr.<sup>3</sup>

## Abstract

In a recent study, Beccarini [1] showed that one can eliminate or reduce the bias in OLS regression estimators caused by an omitted polychotomous variable by estimating a regime-switching model. If the missing polychotomous variable assumes  $K$  values, then elimination or reduction of the bias requires the estimation of a  $K$ -component mixture model. In his Monte Carlo simulations, however, the slope of the parameter of interest is estimated once for each of the  $K$  components. After discussing problems associated with multiple estimates of the parameter of interest, this paper extends Beccarini's Monte Carlo analysis to include the slope-constrained estimator obtained by using the EM algorithm of Bartolucci and Scaccia [2]. We find a small gain in efficiency with the slope-constrained estimator and that the weighted-average estimator in Beccarini [1] produces a large number of rejections of the true null hypothesis of a single slope when the components are not widely separated.

**Mathematics Subject Classification:** 62H12, 62H15, 62J05, 65C05

**Keywords:** omitted variables bias, regime-switching models, slope-constrained estimators, Monte Carlo simulations.

## 1 Introduction

In a recent study, Beccarini [1] showed that one can eliminate or reduce the bias in OLS regression estimators caused by an omitted polychotomous variable by estimating a regime-switching model. If the missing polychotomous variable assumes  $K$  values, then

---

<sup>1</sup>University of South Alabama, USA.

<sup>2</sup>Rhodes College, USA.

<sup>3</sup>Columbus State University, USA.

elimination or reduction of the bias requires the estimation of a  $K$ -component mixture model. In his Monte Carlo simulations, however, the slope of the parameter of interest is estimated once for each of the  $K$  components. After discussing problems associated with multiple estimates of the parameter of interest, this paper extends Beccarini's Monte Carlo analysis to include the slope-constrained estimator obtained by using the EM algorithm of Bartolucci and Scaccia [2]. We find a small gain in efficiency with the slope-constrained estimator and that the weighted-average estimator in Beccarini [1] produces a large number of rejections of the true null hypothesis of a single slope when the components are not widely separated.

Beccarini's [1] approach showed that if the missing polychotomous variable assumes  $K$  values, then elimination or reduction of the bias requires the estimation of a  $K$ -component mixture model. Using several Monte Carlo simulations, Beccarini [1] confirms the assertion. However, in each of these cases, the slope of the parameter of interest is estimated once for each of the  $K$  components. These multiple estimates are then averaged by using the mixing weights. Although Beccarini [1, p. 60] acknowledges an efficiency gain from the estimation of a slope-constrained mixture model, the performance of that estimator is not explored in the Monte Carlo experiments.<sup>4</sup> The objective of this paper is to extend Beccarini's [1] Monte Carlo analysis to include the slope-constrained estimator, obtained by using the EM algorithm of Bartolucci and Scaccia [2], in order to assess the efficacy of averaging to obtain a single parameter estimate. Using our approach, a statistically significant difference between the two models indicates that the averaging process in Beccarini [1] is problematic.

In addition to a gain in estimation efficiency, there are at least two other reasons for imposing this constraint in the estimation. First, with multiple estimates of the parameter of interest, it is possible that a likelihood ratio test for the equality of slopes across regimes might be rejected – then what? Perhaps such a rejection would be useful in that it might represent evidence that there are specification issues beyond an omitted polychotomous variable. The other problem associated with multiple estimates of a parameter of interest is that hypothesis testing and confidence intervals for this slope parameter become quite complicated. For example, if a slope estimate is obtained by averaging slope estimates from a ten-regime mixture model, the weighted average needed to estimate the single unknown slope coefficient is a nonlinear function of nineteen parameters. Considerable effort would be required to obtain an approximate standard error for this parameter.

In this paper we conduct new Monte Carlo simulations and use both Beccarini's [1] weighted-average estimator and the constrained estimator following Bartolucci and Scaccia [2]. We expand on Beccarini's [1] simulations in several respects: 1) we include the slope-constrained estimator, 2) we consider samples of sizes 100 and 400, 3) we compare estimator performance using root mean square errors (RMSE), 4) we allow the distance between the components in the mixture model to vary, 5) we conduct 1,000 Monte Carlo trials for each set of parameters, and 6) we keep track of various failures in the simulation. In general, we find that both estimators perform better in terms of RMSE when the sample size is larger and the components are widely separated. We also find that the gain in efficiency of the slope-constrained estimator is small compared to the

---

<sup>4</sup>Beccarini's [1, p. 65] equation (18) indicates a single  $\beta$  parameter which does not change across regimes.

weighted-average estimator. The biggest problem with the use of the weighted-average estimator is the large number of rejections of the true null hypothesis of a single slope when the components are not widely separated.

## 2 Beccarini's Monte Carlo Simulations

The model used in Beccarini's [1] Monte Carlo simulations is given by,

$$y_t = \beta x_{1t} + \gamma z_t + \varepsilon_t, \quad t = 1, 2, \dots, 100. \quad (1)$$

The dependent variable,  $y$ , depends on  $x$  and  $z$ , where  $x$  is a continuous variable generated as a random normal deviate,  $z$  is a dichotomous variable taking the values  $+1$  and  $-1$ , and  $\varepsilon$  is a normally distributed error term with mean 0 and variance 1. In Beccarini's [1] simulations, the true parameter values are  $\beta = 3$  and  $\gamma = 5$ . The result is a two-regime mixture model with regimes given by,

$$y_t = 3x_{1t} + 5 + \varepsilon_t, \quad (2)$$

and,

$$y_t = 3x_{1t} - 5 + \varepsilon_t. \quad (3)$$

The probability of regime membership is one-half for each. In the simulations, Beccarini [1] considers three cases: Case 1 – omitted variable, dichotomous, zero autocovariance, Case 2 – omitted variable, dichotomous, non-zero autocovariance, and Case 3 – omitted variable, taking its values in an open set of the real line, non-zero autocovariance. Beccarini [1] uses these simulations to show how a finite mixture model provides better estimates in the case of a missing polychotomous variable than OLS using  $x$ , alone. We include the slope-constrained estimator, as proposed by Beccarini [1, p. 60] for the gain in efficiency it provides, and examine only Beccarini's Case 1.

## 3 Our Monte Carlo Simulations

There are some differences between our Monte Carlo simulation and that of Beccarini's Case 1. Beccarini [1] considered a sample size of 100 over 100 Monte Carlo trials, whereas we consider sample sizes of 100 and 400 over 1,000 Monte Carlo trials. Beccarini [1] also considered only the weighted-average estimator, whereas we also estimate the slope-constrained estimator.<sup>5</sup> This study compares estimator performance using mean square error.

Beccarini [1] performed the simulation for  $\gamma = 5$ , resulting in very widely separated components. We consider several values for  $\gamma$ , including: 0.25, 0.50, 1.0, 1.5, 3, and 5. We have done so because we know that the estimation of finite mixture models is complicated by a lack of separation between components. Another difference between our Monte Carlo experiments and those of Beccarini [1] is that we keep track of the number

---

<sup>5</sup>One might point out that the slope-constrained estimator is a special case of the mixture-weighted estimator. While we do not disagree, the comparison chosen in our study is appropriate given that it represents the model used in Beccarini's [1] Monte Carlo experiment.

of estimation problems and the number of nonsensical likelihood ratio test results for the equality of slopes in the two regimes. The first of these problems we call *estimation failures*. One of the problems that may occur when one estimates a mixture model is an estimation failure, which can occur if the EM algorithm locates a singularity and, consequently, the likelihood function increases without bound. In order to check for these estimation failures we have a condition in our program which identifies a failure as having occurred when the standard deviation of either regime is estimated to be less than 0.01. This is circumstantial evidence that one regime is collapsing on a single point.

In our simulations we estimate one model which yields two slope estimates and a second, slope-constrained, model which yields a single slope estimate.<sup>6</sup> Because these two models are nested, we can conduct a likelihood ratio test. Before conducting that test, we check to make certain that the maximized value of the slope-constrained likelihood function is less than the value obtained from the unconstrained estimation. If not, we conclude that we have found a failure associated with a negative chi-square statistic. For these cases the test is nonsensical. For the remaining cases we conduct the usual likelihood ratio test.

## 4 Our Simulation Results

The results of our Monte Carlo simulations are given in Table 1. The top half of the table contains the simulation results when the sample size is 100, and the bottom half of the table contains the results from samples of size 400.

Table 1: Simulation results for 1,000 Monte Carlo trials

Sample Size	$\gamma$	RMSE (simple average)	RMSE (constrained)	RMSE (mixture-weighted)	Estimation Failures	Negative Chi-square	Reject Null of one slope
100	0.25	0.2850	0.1176	0.1086	66	116	362
100	0.50	0.2997	0.1263	0.1230	48	125	364
100	1.00	0.2498	0.1543	0.1508	11	87	293
100	1.50	0.1549	0.1513	0.1553	1	8	136
100	3.00	0.1064	0.1037	0.1053	0	0	135
100	5.00	0.1056	0.1046	0.1054	0	0	100
400	0.25	0.2257	0.0529	0.0531	13	85	349
400	0.50	0.2276	0.0602	0.0599	10	82	366
400	1.00	0.1479	0.0654	0.0656	0	32	301
400	1.50	0.0686	0.0677	0.0691	0	0	102
400	3.00	0.0505	0.0505	0.0505	0	0	91
400	5.00	0.0499	0.0501	0.0500	0	0	106

Note:  $\gamma$  is the sample separation.

The first row of the table contains Monte Carlo results for samples of size 100, and  $\gamma =$

<sup>6</sup>Of course, the true data-generating process assumes the slopes are the same, which calls for the estimation of the constrained model that we highlight in our study.

0.25. This means that the components are one-half of one standard deviation apart, which is close, thus complicating the estimation. Based on RMSE, the weighted estimator actually performs a bit better than the slope-constrained estimator. This result can also be seen in Figure 1.

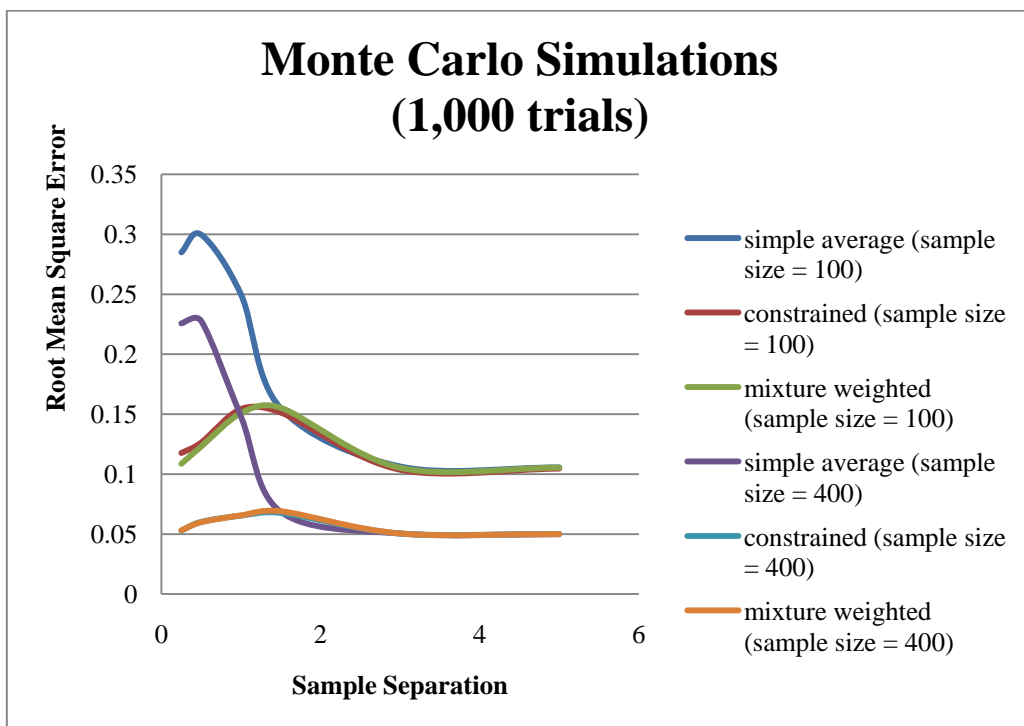


Figure 1: Estimator performance

However, this conclusion must be tempered by an examination of failures and rejections in this Monte Carlo experiment. For example, the table indicates there were 66 estimation failures out of 1,000 trials, and 116 instances of a negative chi-square value. Of the remaining 818 Monte Carlo trials, 362 of led to a rejection of the true null hypothesis (at the  $\alpha = 0.10$  level) of a single underlying slope. There seems to be considerable statistical evidence of the presence of multiple regimes, making the averaging of the two slope estimates problematic. As illustrated in Figure 2, there is an indirect relationship between sample separation and the number of failures. In fact, as Table 1 indicates, if the value of  $\gamma$  increases, the number of estimation failures goes to zero. The same relationship is found between sample separation and negative chi-square values. In particular, when the value of  $\gamma$  reaches five, which is the value used in Beccarini’s simulations, there are 100 rejections of the true null hypothesis out of 1,000 trials, which is exactly what one would predict.

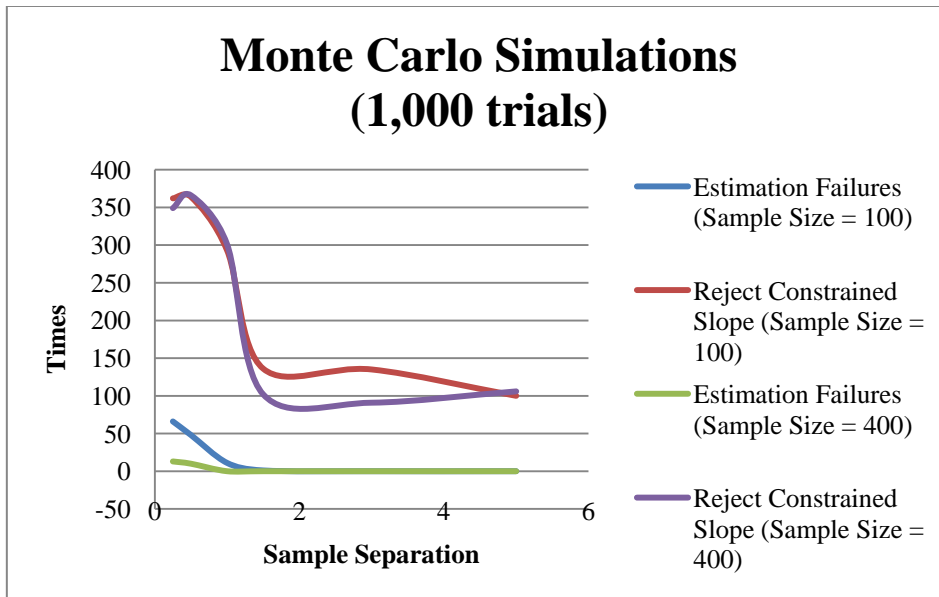


Figure 2: Log-likelihood ratio tests and simulation performance

The bottom half of Table 1 shows the same Monte Carlo experiment for a sample size of 400. As expected, the RMSEs for both estimators are very similar and approximately one-half of their values when the sample size is 100. The number of estimation failures and negative chi-square values diminishes as the sample size increases. However, when the value of  $\gamma$  is 0.25, the number of rejections of the true null hypothesis is still nearly 350. In fact, when the value of  $\gamma$  is 0.50, the number of rejections is 366 (and still 301 for  $\gamma = 1.0$ ). Overall, we see a result consistent with other findings on mixture models – that estimation is less problematic and easier with larger samples and significant sample separation. Those facts are confirmed in our Monte Carlo simulations.

## 5 Conclusion

Statistical models with omitted variables yield biased estimators, unless the omitted variables are orthogonal to the known vectors, or the population parameter of the omitted variable is zero. Beccarini [1] argues that the bias caused by the omission of polychotomous variables can be mitigated by the use of regime-switching models. Beccarini [1] conducted Monte Carlo simulations to estimate a regime-switching model with the number of components equal to the number of values of the omitted polychotomous variable. As the estimator of the slope parameter, Beccarini [1] used the average of the slope estimates from each regime, weighted by the associated mixture weight. We extend the Monte Carlo experiments of Beccarini [1] in several aspects: in addition to Beccarini’s weighted-average estimator, we estimate a slope-constrained mixture model to make an efficiency comparison; in addition to samples of size 100 we also examine samples of size 400, and we compare estimator performance using the RMSE; we allow the distance between the components of the mixture model to vary, and we record various failures of the simulation; finally, we compute a series of log-likelihood

ratio statistics to test our approach (single slope estimator) against Beccarini's [1] approach (variable slope).

We find that that both estimators perform better in terms of RMSE when the sample size is larger and the components are widely separated, and that the gain in efficiency of the slope-constrained estimator is small compared to the weighted-average estimator. The major problem with the use of the weighted-average estimator is the large number of rejections of the true null hypothesis of a single slope when the components are not widely separated. Lastly, the approach developed in Beccarini [1] is, by itself, perhaps quite useful for other, nonlinear regression models. Good examples in recent literature of where Beccarini's [1] approach would be beneficial include Caudill and Long [4] and Caudill [3]. These studies examine grouped-data and censored regression models, respectively.

## References

- [1] A. Beccarini, Eliminating the omitted variable bias by a regime-switching approach, *Journal of Applied Statistics*, **37**(1), (2010), 57-75.
- [2] F. Bartolucci and L. Scaccia, The use of mixtures for dealing with non-normal regression errors, *Computational Statistics and Data Analysis*, **48**(4), (2005), 821-834.
- [3] S.B. Caudill, A partially adaptive estimator for the censored regression model based on a mixture of normal distributions, *Statistical Methods and Applications*, **21**(2), (2012), 121-137.
- [4] S.B. Caudill and J.E. Long, Do former athletes make better managers? Evidence from a partially adaptive grouped-data regression model, *Empirical Economics*, **39**(1), (2010), 275-290.