CESPRI

Centro di Ricerca sui Processi di Innovazione e Internazionalizzazione
Università Commerciale "Luigi Bocconi"
Via R. Sarfatti, 25 – 20136 Milano
Tel. 02 58363395/7 – fax 02 58363399
http://www.cespri.unibocconi.it

Giovanni S. F. Bruno

ESTIMATION AND INFERENCE IN DYNAMIC UNBALANCED PANEL DATA MODELS WITH A SMALL NUMBER OF INDIVIDUAL

WP n. 165 **February 2005**

Stampato in proprio da:
Università Commerciale Luigi Bocconi – CESPRI Via Sarfatti, 25 20136 Milano
Anno: 2005

Estimation and inference in dynamic unbalanced panel data models with a small number of individuals

Giovanni S.F. Bruno
Istituto di Economia politica, Bocconi University, Milan
giovanni.bruno@unibocconi.it

February 2005

Abstract. This study describes a new Stata routine that computes bias-corrected LSDV estimators and their bootstrap variance-covariance matrix for dynamic (possibly) unbalanced panel data models. A Monte Carlo analysis is carried out to evaluate the finite-sample performance of the bias corrected LSDV estimators in comparison to the original LSDV estimator and three popular N-consistent estimators: Arellano-Bond, Anderson-Hsiao and Blundell-Bond. Results strongly support the bias-corrected LSDV estimators according to bias and root mean squared error criteria when the number of individuals is small.

Keywords: Bias approximation; Unbalanced panels; Dynamic panel data; LSDV estimator; Monte Carlo experiment; Bootstrap variance-covariance

JEL Codes: C23, C15

1 Introduction

Situations in which past decisions have an impact on current behaviour are ubiquitous in economics. To mention just one of the most familiar cases, in the presence of employment adjustment costs the short-run labour demand of the firm will depend on past employment levels. Another crucial issue in empirical economics, strictly related to the modelling of dynamic relationships, is the presence of unobserved heterogeneity in individual behaviour and characteristics. Panel data sets, where the behaviour of N cross-sectional units is observed over T time periods, provide a solution to accommodating the joint occurrence of dynamics and unobserved individual heteogeneity in the phenomena of interest.

Since the seminal paper by Nickell (1981), where it is shown that the Least Square Dummy Variable estimator (LSDV) is not consistent for finite T in autoregressive panel data models, a number of consistent instrumental variable (IV) and Generalised Method of Moments (GMM) estimators have been proposed in the econometric literature as an alternative to LSDV. Anderson and Hsiao (1982) (AH) suggest two simple IV estimators that, upon transforming the model in first differences to eliminate the unobserved individual heterogeneity, use the second lags of the dependent variable, either differenced or in levels, as an instrument for the differenced one-time lagged dependent variable. Arellano and Bond (1991) (AB) propose a GMM estimator for the first differenced model which, relying on a greater number of internal instruments, is more efficient than AH. Blundell and Bond (1998) (BB) observe that with highly persistent data first-differenced IV or GMM estimators may suffer of a severe small sample bias due to weak instruments. As a solution, they suggest a system GMM estimator with first-differenced instruments for the equation in levels and instrument in levels for the first-differenced equation.

A weakness of IV and GMM estimators is that their properties hold for N large, so they can be severely biased and imprecise in panel data with a small number of cross-sectional units. This is often the case in most macro panels, but also in micro panels where heterogeneity concerns force the researcher not to use all information available, but rather to select a subsample of individuals from the original panel to estimate the parameters of interest. On the other hand, earlier Monte Carlo studies (Arellano and Bond (1991), Kiviet (1995) and Judson and Owen (1999)) demonstrate that LSDV although inconsistent has a relatively small variance compared to IV and GMM estimators.

Moving from the foregoing considerations, an alternative approach based upon the bias-correction of LSDV has recently become popular in the econometric literature. Nickell (1981) derives an expression for the inconsistency of LSDV for $N \to +\infty$, which is bounded of order T^{-1}. Kiviet (1995) uses higher order asymptotic expansion techniques to approximate the small sample bias of the LSDV estimator to include terms of at most order $N^{-1}T^{-1}$. The approximations terms however, all evaluated at the unobserved true parameter values, are of no direct use for estimation, so to make them operational he suggests replacing the true parameters by the estimates from some consistent estimators. Monte Carlo evidence therein shows that the resulting bias-corrected LSDV estimator (LSDVC) often outperforms the IV-GMM estimators in terms of bias and root mean squared error (RMSE). Another piece of Monte Carlo evidence by Judson and Owen (1999) strongly supports LSDVC when N is small as in most macro panels. In Kiviet (1999) the bias expression is more accurate to include terms of at most order $N^{-1}T^{-2}$. Bun and Kiviet (2003), upon simplifying the approximations in Kiviet (1999), carry out Monte Carlo experiments showing that the first order term of the approximation evaluated at the true parameter values is already capable to account for more than 90% of the actual bias.

None of the foregoing procedures to correct the LSDV estimator is feasible for unbalanced panels. This gap is partly filled in Bruno (2005), where the bias approximations in Bun and Kiviet (2003) are extended to accommodate unbalanced panels with a strictly exogenous selection rule. Monte Carlo evidence therein parallels that in Bun and Kiviet (2003).

This paper presents a new *Stata* routine, `xtlsdvc`, which 1) implements LSDVC building upon the theoretical approximation formulae in Bruno (2005) and 2) estimates a bootstrap variance covariance matrix for the corrected estimator. Moreover, the relative performance of LSDVC is evaluated in comparison to LSDV, AB, AH and BB for unbalanced panels with a small N (10 and 20 units) through various Monte Carlo experiments, thus extending the analysis by Judson and Owen (1999). Results show that the three versions of LSDVC computed by `xtlsdvc` outperform all other estimators tried in terms of bias and RMSE.

The paper is laid out as follows. The next section briefly reviews the theoretical results for corrected LSDV estimators. Section 3 describes the `xtlsdvc` routine. Section 4 contains the Monte Carlo analysis and Section 5 concludes. A demonstration of the code in the context of labour demand estimation is offered into an appendix.

2 Bias corrected LSDV estimators

I consider the standard dynamic panel data model

$$y_{it} = \gamma y_{i,t-1} + x'_{it}\beta + \eta_i + \epsilon_{it};\ |\gamma| < 1;\ i = 1, ..., N \ and \ t = 1, ..., T, \quad (1)$$

where y_{it} is the dependent variable; x_{it} is the $((k-1) \times 1)$ vector of strictly exogenous explanatory variables; η_i is an unobserved individual effect; and ϵ_{it} is an unobserved white noise disturbance. Collecting observations over time and across individuals gives

$$y = D\eta + W\delta + \epsilon,$$

where y and $W = \left[y_{-1} \vdots X\right]$ are the $(NT \times 1)$ and $(NT \times k)$ matrices of stacked observations; $D = I_N \otimes \iota_T$ is the $(NT \times N)$ matrix of individual dummies, (ι_T is the $(T \times 1)$ vector of all unity elements); η is the $(N \times 1)$ vector of individual effects; ϵ is the $(NT \times 1)$ vector of disturbances; and $\delta = \left[\gamma \vdots \beta'\right]'$ is the $(k \times 1)$ vector of coefficients.

It has been long recognized that the LSDV estimator for model (1) is not consistent for finite T. Nickell (1981) derives an expression for the inconsistency for $N \to +\infty$, which is $O\left(T^{-1}\right)$. Kiviet (1995) obtains a bias approximation that contains terms of higher order than T^{-1}. In Kiviet (1999) a more accurate bias approximation is derived. Bun and Kiviet (2003) reformulate the approximation in Kiviet (1999) with simpler formulae for each term.

Bruno (2005) extends Bun and Kiviet's (2003) formulae to unbalanced panels with a strictly exogenous selection rule. A more general version of model (1) is considered, which allows missing observations in the interval $[0, T]$ for some individuals. Below, I briefly present the approximation formulae for (possibly) unbalanced data and show their use to obtain LSDVC.

Define a selection indicator r_{it} such that $r_{it} = 1$ if (y_{it}, x_{it}) is observed and $r_{it} = 0$ otherwise. From this define the dynamic selection rule $s\left(r_{it}, r_{i,t-1}\right)$ selecting only the observations that are usable for the dynamic model, namely those for which both current values and one-time lagged values are observable:

$$s_{it} = \begin{cases} 1 & if\,(r_{i,t}, r_{i,t-1}) = (1,1) \\ 0 & otherwise \end{cases} \quad i = 1, ..., N \; and \; t = 1, ..., T.$$

Thus, for any i the number of usable observations is given by $T_i = \sum_{t=1}^{T} s_{it}$. The total number of usable observations is given by $n = \sum_{i=1}^{N} T_i$; and $\overline{T} = n/N$ denotes the average group size. For each i define the $(T \times 1)$-vector $s_i = [s_{i1} ..., s_{iT}]'$ and the $(T \times T)$ diagonal matrix S_i having the vector s_i on its diagonal. Define also the $(NT \times NT)$ block-diagonal matrix $S = diag\left(S_i\right)$. The (possibly) unbalanced dynamic model can then be written as

$$Sy = SD\eta + SW\delta + S\epsilon. \tag{2}$$

The LSDV estimator is given by

$$\delta_{LSDV} = \left(W' M_s W\right)^{-1} W' M_s y,$$

where

$$M_s = S\left(I - D\left(D'SD\right)^{-1} D'\right) S$$

is the symmetric and idempotent $(NT \times NT)$ matrix wiping out individual means and selecting usable observations.

Bias approximation terms for unbalanced panels are the following

$$c_1\left(\overline{T}^{-1}\right) = \sigma_\epsilon^2 tr\left(\Pi\right) q_1; \tag{3}$$

$$\begin{aligned} c_2\left(N^{-1}\overline{T}^{-1}\right) &= -\sigma_\epsilon^2 \Big[Q\overline{W}'\Pi M_s \overline{W} + tr\left(Q\overline{W}'\Pi M_s\overline{W}\right) I_{k+1} + \\ & \quad 2\sigma_\epsilon^2 q_{11} tr\left(\Pi'\Pi\Pi\right) I_{k+1}\Big] q_1; \end{aligned}$$

$$\begin{aligned} c_3\left(N^{-1}\overline{T}^{-2}\right) &= \sigma_\epsilon^4 tr\left(\Pi\right) \Big\{2q_{11}Q\overline{W}'\Pi\Pi'\overline{W}q_1 + \Big[\left(q_1'\overline{W}'\Pi\Pi'\overline{W}q_1\right) + \\ & \quad q_{11} tr\left(Q\overline{W}'\Pi\Pi'\overline{W}\right) + 2tr\left(\Pi'\Pi\Pi'\Pi\right) q_{11}^2\Big] q_1\Big\}; \end{aligned}$$

where $Q = [E(W'M_sW)]^{-1} = \left[\overline{W}'M_s\overline{W} + \sigma_\epsilon^2 tr(\Pi'\Pi) e_1e_1'\right]^{-1}$; $\overline{W} = E(W)$; $e_1 = (1, 0, ..., 0)'$ is a $(k \times 1)$ vector; $q_1 = Qe_1$; $q_{11} = e_1'q_1$; L_T is the $(T \times T)$ matrix with unit first lower subdiagonal and all other elements equal to zero; $L = I_N \otimes L_T$; $\Gamma_T = (I_T - \gamma L_T)^{-1}$; $\Gamma = I_N \otimes \Gamma_T$; and $\Pi = M_sL\Gamma$. Clearly, in any balanced design $S \equiv I_{NT}$, so $M_s = I - D(D'D)^{-1}D'$, and the above terms reduce to Bun and Kiviet's (2003).

With an increasing level of accuracy, the following three possible bias approximations emerge

$$B_1 = c_1\left(\overline{T}^{-1}\right); B_2 = B_1 + c_2\left(N^{-1}\overline{T}^{-1}\right); B_3 = B_2 + c_3\left(N^{-1}\overline{T}^{-2}\right). \quad (4)$$

In principle, bias corrected LSDV estimators could be obtained by subtracting any of the above terms from LSDV. In practice, however, depending upon the unknown parameters σ_ϵ^2 and γ, approximations (4) are not feasible for bias correction. Nevertheless, consistent bias corrected estimators can be obtained by finding consistent estimators for σ_ϵ^2 and γ, plugging them into the bias approximations formulae, and then subtracting the resulting bias approximation estimates, $\widehat{B}_i$, from LSDV as follows:

$$LSDVC_i = LSDV - \widehat{B}_i, \ i = 1, \ 2 \ and \ 3. \quad (5)$$

Possible consistent estimators for γ are AH, AB, or BB, for example. Depending on the estimator of choice for γ, say h, a consistent estimator for σ_ϵ^2 is then given by

$$\widehat{\sigma}_h^2 = \frac{e_h'M_se_h}{(N-k-T)}, \quad (6)$$

where $e_h = y - W\delta_h$, and $h = AH, \ AB \ and \ BB$.

3 The xtlsdvc routine

The Stata routine `xtlsdvc` written by the author calculates LSDVC for model (1) using estimates for the bias approximations in (4). The basic syntax of `xtlsdvc` is the following

`xtlsdvc` *depvar* [*varlist*] [`if` *exp*] `, initial(`*estimator*`)` [`level(`*#*`)` `bias(`*#*`)` `vcov(`*#*`)` `first` `lsdv`]

So the routine can estimate the simple autoregressive model with no covariates. The options for `xtlsdvc` are described below.

`level(`*#*`)` specifies the confidence level, in percent, for confidence intervals of the coefficients. The default is `level(95)` or as set by `set level`; see [U] **23.5 Specifying the width of confidence intervals**.

`initial(`*estimator*`)` is required and specifies the consistent estimator chosen to initialize the bias correction.

estimator	description
`ah`	AH estimator, with the dependent variable lagged two times used as an instrument for the first differenced model with no intercept ([R] **ivreg**.)
`ab`	standard one-step AB estimator with no intercept ([XT] **xtabond**.)
`bb`	standard BB estimator with no intercept, as implemented by the user-written Stata routine `xtabond2` by David Roodman (2004).
`my`	a row vector of initial values supplied directly by the user.

To implement the last instance of this option the user has to create a $(1 \times (k+1))$ matrix to be named `my`, the i.th element of which serves as an initial value for the coefficient on the i.th variable in *varlist* and the last, (k+1).th, element as an estimate for the error variance. This may be useful in Monte Carlo simulations or if the user whishes to try initial estimators other than `ah`, `ab` or `bb`.

`bias(#)` determines the accuracy of the approximation: #=1 (default) forces an approximation up to $O(1/T)$; #=2 forces an approximation up to $O(1/NT)$; #=3 forces an approximation up to $O(N^{-1}T^{-2})$.

`vcov(#)` calculates a bootstrap variance-covariance matrix for LSDVC using # repetitions (# may not equal 1). The default is no bootstrap estimation of the variance-covariance matrix and standard errors. Notice that the bootstrap continues to work also in the presence of gaps in the exogenous variables, although in this case bootstrap samples for each unit are truncated to the first missing value encountered. Gaps in the dependent variable, instead, bear no consequence to the bootstrap sample size. This is explained in more detail in Section 3.2. Also consider that bootstrap standard errors are downward biased when values for the unknown parameters are supplied through matrix `my`, since the procedure in this case, keeping the values in `my` fixed over replications, neglects a source of varibility for LSDVC.

`first` requests that the first-stage regression results be displayed.

`lsdv` requests that the original LSDV regression results be displayed.

To work out the approximations `xtlsdvc` invokes the subroutine `xtlsdvc_1` that accomplishes the following tasks. In the first place, `xtlsdvc_1` obtains the uncorrected LSDV estimates via a a call to `xtreg..., fe` ([XT] **xtreg**.)

Second, `xtlsdvc_1` obtains initial estimates for γ and β through one of the following instructions, depending on which *estimator* is specified in `initial`:

```
if "`initial'"=="ah" ivreg D.y D.x (LD.y=L2.y), noconstant
if "`initial'"=="ab" xtabond y x, noconstant
if "`initial'"=="bb" xtabond2 y L.y x, gmm(L.y) iv(x) noconstant.
```

Then $\widehat{\sigma}_h^2$, $h = AH,\ AB\ and\ BB$, is computed as in (6).

Finally, `xtlsdvc_1` computes the bias approximations via the Stata `matrix` commands ([P] **matrix**), and corrects the LSDV estimates as indicated in (5).

3.1 Saved results

`xtlsdvc` saves in `e()`:

Scalars

`e(N)`	number of observations	`e(sigma)`	estimates of σ from the first stage regression
`e(Tbar)`	average number of time periods	`e(N_g)`	number of groups

Macros

`e(cmd)`	`xtlsdvc`	`e(depvar)`	name of dependent variable

Matrices

`e(b)`	`xtlsdvc` estimates	`e(V)`	var–cov matrix of the `xtlsdvc` estimator
`e(b_lsdv)`	`xtreg,fe` estimates	`e(V_lsdv)`	var–cov matrix of the `xtreg,fe` estimator

Functions

`e(sample)`	marks estimation sample

3.2 The bootstrap variance-covariance matrix

Kiviet and Bun (2001) show that LSDVC, however initialized, is asymptotically normal, and derive the analytical expression for the asymptotic variance-covariance matrix of LSDVC in the version initialized by AH. Monte Carlo simulations therein, however, demonstrate that the analytical variance estimator performs poorly for a large γ, perhaps because of the unstable behavior of AH (documented also by the Monte Carlo analysis of this paper, see Section 4). In alternative, therefore, Kiviet and Bun (2001) suggest a parametric bootstrap procedure to estimating the asymptotic variance-covariance matrix of LSDVC, which seems superior to the analytical expression for at least three reasons: 1) it is simpler; 2) it always turns out as relatively accurate; and 3) it can be applied to any version of LSDVC. Thus, `xtlsdvc` adapts Kiviet and Bun's (2001) bootstrap procedure for use with unbalanced panels, as described below.

A first difficulty here is brought about by the dependency in the data implied by the autoregressive data generation process (DGP), which does not permit to adopt any of the official Stata bootstrap instructions, `bootstrap` and `bsample`. A parametric bootstrap is instead followed, which upon maintaining a normal distribution for the disturbances takes full account of the dependency in the DGP.

The subroutine `xtlsdvc_b` is called in `xtlsdvc` by the option `vcov`. It is designed to yield a bootstrap sample and bootstrap LSDVC estimates and is iterated for `vcov(#)` times by `xtlsdvc`.

Let us focus on the generic iteration (*) of `xtlsdvc_b`. It basically goes through the steps below.

1. Upon obtaining LSDVC estimates $\widehat{\gamma}$ and $\widehat{\beta}$ and $\widehat{\sigma}^2$ from `xtlsdvc_1`, it calculates the N-vector of fixed effect estimates $\widehat{\eta} = \overline{y} - \widehat{\gamma} \cdot \overline{y}_{-1} - \widehat{\beta} \cdot \overline{x}$, where $\overline{y}$, $\overline{y}_{-1}$ and $\overline{x}$, indicate N-vectors of group means.

2. It obtains bootstrap errors $\epsilon^{(*)}$ as a draw from $N\left(0, \widehat{\sigma}^2\right)$.

3. Given x, S and y_0, it obtains a bootstrap sample from $s_{it}y_{it}^{(*)} = s_{it}(\widehat{\gamma}\cdot y_{i,t-1}^{(*)} + \widehat{\beta}\cdot x_{it} + \widehat{\eta}_i + \epsilon_{it}^{(*)})$, $i = 1, ..., N$ and $t = 1, ..., T$.
4. It applies LSDVC to $\left(y^{(*)}, S, x\right)$ to yield $\widehat{\gamma}^{(*)}$ and $\widehat{\beta}^{(*)}$.

While computational aspects of steps 1 and 2 are straightforward and step 4 only requires a call to `xtlsdvc_1` to calculate the corrected estimates from the generated bootstrap sample, step 3 is instructive and deserves some explanation. One possible way to implement step 3 would be to "manually" generate $y^{(*)}$ by recursion as a function of $\epsilon^{(*)}$, y_0 and x. But this is both computationally cumbersome and unnecessary in Stata. In fact one can exploit the ability of `replace` ([R] **generate, replace**) to work sequentially[1] to obtain $y^{(*)}$ in an effortless way:

```
by ivar:  gen obs=_n
replace y= GAMMA*L.y + BETA*x +THETA +EPSILON if obs>1.
```

Unbalancedness without gaps does not cause any trouble here, since different start-up dates can be dealt with very easily by the time series operators in Stata. The presence of gaps, instead, may cause a specific difficulty as long as they are found in any of the independent variables x's, regardless of the way step 3 is implemented. In fact, since the recursion process generates $y^{(*)}$ from (y_0, S, x), it must stop at the first missing value encountered in the x's, so that eventually a shorter sample is created at each replication. This may deteriorate the accuracy of the estimates or even break down the identification of some coefficients in the shorter bootstrap sample and, consequently, of their standard errors. For example, if for all individuals there is a gap for a given time period, then the coefficients on the time dummies subsequent to the missing period would not be identified in each bootstrap sample, so that their bootstrap standard errors could not be computed too. To the opposite, gaps in the dependent variable are clearly immaterial for the size of the bootstrap samples, since only the start-up values of y are used in the recursion process.

A `simulate` call ([R] **simulate**) in `xtlsdvc` replicates `xtlsdvc_b` for `vcov(#)` times, yielding a data set of bootstrap LSDVC estimates $\widehat{\delta}^*$, of dimension $(\texttt{vcov} \times k)$. Hence, `xtlsdvc` gets the bootstrap variance-covariance matrix V:

$$V = \frac{\widehat{\delta}^{*\prime}\widehat{\delta}^*}{(\texttt{vcov} - 1)}$$

via `matrix accum` ([P] **matrix**.)

The bootstrap variance-covariance matrix V is then used to construct asymptotic t-ratio tests of parameter significance as described in Kiviet and Bun (2001).

[1] I learnt this from the messages by N. J. Cox and D. Kantor to Statalist on May 25, 2004 in response to a question of D. V. Masterov.

Attention should be paid when supplying the initial values through the matrix `my`. In this case, in fact, the bootstrap procedure would not be reliable, since keeping the values in `my` fixed over replications, it neglects a source of varibility for LSDVC, so that the resulting bootstrap standard errors may be severely downward biased.

Finally, users should be warned that the bootstrap procedure may require a considerable amount of time. This tends to increase linearly with the number of replications. Also, the procedure seems slightly faster if LSDVC is initialized by AH. Examples are given in the appendix.

4 Monte Carlo experiments

The Monte Carlo analyses in Kiviet (1995), Kiviet and Bun (2001) and, especially, Judson and Owen (1999) provide support for LSDVC in balanced panels, compared to the traditional IV and GMM estimators. Moreover, Monte Carlo results in Bun and Kiviet (2003) for balanced panels and in Bruno (2005) for unbalanced panels demonstrate that the bias approximations (4), evaluated at the true γ and σ_ϵ^2, account for a significant portion of the bias, never less than 90% and often virtually 100%. The relative merit of LSDVC in unbalanced panels is still to be explored, though. This is exactly what accomplished here, where I evaluate the three versions of LSDVC as implemented by my code in a Monte Carlo study that extends Judson and Owen's (1999) under four respects. First, I evaluate LSDVC in the presence of various unbalanced designs; second, the performance of LSDVC is examined for the three different levels of accuracy; third, initial observations for the simulated data are generated following the procedure by McLeod and Hipel (1978), also adopted in Kiviet (1995) and Bruno (2005), which avoids the waste of random numbers and small sample non-stationary problems; finally, the comparison is extended to BB.

Data for y_{it} are generated by model (1) and for x_{it} by

$$x_{it} = \rho x_{i,t-1} + \xi_{it},\ \xi_{it} \sim N\left(0, \sigma_\xi^2\right),\ i = 1, ..., N \textit{ and } t = 1, ..., T.$$

Initial observations y_{i0} and x_{i0} generated through the McLeod and Hipel (1978) procedure are kept fixed across replications. The long-run coefficient $\beta/(1-\gamma)$ is kept fixed to unity, so $\beta = 1-\gamma$; σ_ϵ^2 is normalized to unity; γ and ρ alternate between 0.2 and 0.8. The individual effects η_i are generated by assuming $\eta_i \sim N\left(0, \sigma_\eta^2\right)$ and $\sigma_\eta = \sigma_\epsilon (1-\gamma)$.

Two different sample sizes are considered, $\left(N, \overline{T}\right) = (20, 20)$ and $\left(N, \overline{T}\right) = (10, 40)$. Then, following Baltagi and Chang (1994), I control for the extent of unbalancedness as measured by the Ahrens and Pincus index: $\omega = N/\left[\overline{T}\sum_{i=1}^{N}(1/T_i)\right]$ ($0 < \omega \leq 1$, $\omega = 1$ when the panel is balanced). For each sample size I analyze a case of mild unbalancedness ($\omega = 0.96$) and a case of severe unbalancedness ($\omega = 0.36$). Individuals are partitioned into two sets of equal dimension: one set contains the first $N/2$ individuals, each with the last h observations discarded, so $T_i = T - h$; the other contains the remaining $N/2$ individuals, each with $T_i = T$. I set T and h so that $\overline{T}$ and ω take on the desired values (the four panel designs are summarized in Table 1).

Table 1
Unbalanced designs

N	$\overline{T}$	T	T_i	ω
20	20	24	16 ($i \leq 10$), 24 ($i > 10$)	0.96
		36	4 ($i \leq 10$), 36 ($i > 10$)	0.36
10	40	48	32 ($i \leq 5$), 48 ($i > 5$)	0.96
		72	8 ($i \leq 5$), 72 ($i > 5$)	0.36

The simple AH estimator is the one chosen to initialize the correction procedure, based on the finding by Kiviet and Bun (2001) that differences in the initial estimators have only a marginal impact on the LSDVC performance. Then, the LSDVC estimator is calculated for each of the three levels of accuracy in the estimated bias approximations.

4.1 Results

Results for γ are presented in figures 1 to 4, while results for β are presented in figures 5 to 8. In each figure the first graph is for $\overline{T} = 20$ and the second for $\overline{T} = 40$. The bias and the RMSE are measured onto the vertical axis, while the points onto the horizontal axis always correspond to the eight possible combinations for γ, ρ and ω (only the combinations with $\omega = 0.36$ are labeled). Since BB is specifically designed for highly persistent series, comparisons involving this estimator are restricted to $\gamma = 0.8$.

As a first general comment on the Monte Carlo results I observe that according to a bias criterion the three versions of LSDVC and, interestingly, AH have the best performances for both γ and β, with virtually zero bias in several cases. Turning to a RMSE criterion, the LSDVC estimators maintain the best performance, while AH shows the worst RMSE levels, also in comparison to LSDV, AB and, for highly persistent series, BB. This evidence highlights LSDVC as the preferred estimator for dynamic panel data models with a small N. These results are in line with what obtained by Kiviet (1995), Judson and Owen (1999) and Kiviet and Bun (2001) in similar Monte Carlo analyses.

This said, some interesting patterns seem to emerge when the behavior of each estimator is examined in more depth.

4.1.1 Estimating γ : bias

$LSDVC_3$ tends to perform slightly better than the other two LSDVC versions, especially when $\overline{T}$ and γ increases. When $\gamma = 0.8$ and $\rho = 0.8$, however, all LSDVC estimators are slightly worse than AH (see Fig. 1).

After noting that the bias of LSDV and AB is always negative, confirming the findings by earlier studies (Kiviet and Bun (2001), Bond (2002), Bun and Kiviet (2003) and Bruno (2005)), I observe that LSDVC estimators, LSDV and AB show similar patterns with respect to the degree of unbalancedness and average group size. As already found for LSDV in Bruno (2005), the biases of such estimators are decreasing in ω. This, always for AB and LSDV and often for LSDVC, brings with it an increase in the bias magnitude. The impact of ω, then, seems particularly strong for AB when $\overline{T} = 20$, always reversing to the worse the relative performance of that estimator with respect to LSDV. When $\overline{T}$ increases, however, besides observing an expected general tendency towards a smaller bias magnitude, I also notice an attenuation of the ω effect for all foregoing estimators. The bias of AH, instead, is always positive and increasing in ω, implying each time a worsening of the bias when unbalancedness reduces. The bias of BB is always positive and expectedely the largest in magnitude with lowly persistent series, but it dramatically improves when the persistence in y and x increases, reaching lower magnitudes than AB and LSDV when $\overline{T} = 20$ and comparable to AB and LSDV when $\overline{T} = 40$ (see Fig. 2).

4.1.2 Estimating γ : RMSE

The RMSE of the LSDVC estimators are almost coincident and always the smallest. To the opposite, AH almost always presents the highest RMSE, which hinders the attractiveness of such estimator in empirical work, despite its simplicity and good bias performance (see Fig. 3).

The RMSE for all estimators but BB is increasing in γ and ρ, especially so for AH, AB and LSDV. For BB, instead, I notice a stable behavior. As already observed discussing bias performances, LSDVC estimators, AB and LSDV all experience a worse RMSE when unbalancedness reduces. Again, this effect is particularly strong for AB and when $\overline{T} = 20$. BB has a satisfactory RMSE in the presence of highly persistent series, performing generally better than AB and LSDV. In particular, when $\overline{T} = 40$ and $\omega = 0.96$ its RMSE gets very close to that of the LSDVC estimators (see Fig. 4).

4.1.3 Estimating β : bias

LSDVC estimators and AH continue to show the best bias performance. While for $\rho = 0.2$ also AB and LSDV exhibit a negligible bias magnitude, for $\rho = 0.8$ their bias magnitude dramatically increases. With small $\overline{T}$ I notice a relatively bad perfomance of BB. When $\overline{T} = 40$ and $\omega = 0.36$, however, the bias attains acceptable levels, to worsen back when the degree of unbalancedness decreases (see Figg. 5 and 6).

4.1.4 Estimating β : RMSE

Results here parallel what evidenced for γ, with two differences: 1) There seems to be no clear role for the degree of unbalancedness. For example, when $\overline{T} = 20$ the RMSE of the LSDVC estimators benefits from a decreased unbalancedness, but when $\overline{T} = 40$ exactly the opposite occurs. 2) The RMSE for BB is now markedly increasing in ρ (see Figg. 7 and 8).

The documented evidence for a favourable impact of unbalancedness on bias and RMSE values, apparently surprising, can be explained by the fact that under investigation here is a notion of pure unbalancedness, not involving either gaps or any loss in degrees of freedom and average group size. Although more theoretical work, accompanied by broader Monte Carlo experiments, is needed to reach conclusive results on this issue, there is still a simple lesson to be learnt from my Monte Carlo analysis, that is smoothing unbalancedness at the cost of less time observations for the largest groups may be detrimental for estimation performances in dynamic panel data models, especially if the average group size is small.

5 Conclusion

This paper has presented the new Stata code `xtlsdvc` implementing LSDVC estimators for dynamic (possibly) unbalanced panel data models. The procedure is based upon the bias approximations derived in Bruno (2005), who extends the result by Kiviet (1999) and Bun and Kiviet (2003) to unbalanced panels. The code also computes the bootstrap variance-covariance matrix of the estimators.

Monte Carlo experiments highlight the LSDVC estimators as the preferred ones in comparison to the original LSDV and widely used IV and GMM consistent estimators when the number of individuals is small.

Future improvements of the code will enlarge the class of initial estimators, allowing also more flexibility in the definition of the instrument set for the IV and GMM estimators.

Acknowledgements I wish to thank participants at the 10th UK Stata Users Group meeting and the 1st Ialian Stata Users Group meeting held in London and Rome in 2004 for useful comments and suggestions to earlier versions of this paper.

6 Bibliography

Anderson, T. W. and C. Hsiao. 1982. Formulation and Estimation of Dynamic Models Using Panel Data. *Journal of Econometrics* 18: 570–606.

Arellano, M. and S. Bond. 1991. Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations. *Review of Economic Studies* 58: 277–297.

Blundell, R. and S. Bond. 1998. Initial Conditions and Moment Restrictions in Dynamic Panel Data Models. *Journal of Econometrics* 87: 115–143.

Bond, S. 2002. Dynamic Panel Data Models: A Guide to Micro Data Methods and Practice. *Cemmap working paper CWP09/02* .

Bruno, G. S. F. 2005. Approximating the Bias of the LSDV Estimator for Dynamic Unbalanced Panel Data Models. *Economics Letters, forthcoming* .

Bun, M. J. G. and J. F. Kiviet. 2003. On the diminishing returns of higher order terms in asymptotic expansions of bias. *Economics Letters* 79: 145–152.

Judson, R. A. and A. L. Owen. 1999. Estimating dynamic panel data models: a guide for macroeconomists. *Economics Letters* 65: 9–15.

Kiviet, J. F. 1995. On Bias, Inconsistency and Efficiency of Various Estimators in Dynamic Panel Data Models. *Journal of Econometrics* 68: 53–78.

—. 1999. Expectation of Expansions for Estimators in a Dynamic Panel Data Model; Some Results for Weakly Exogenous Regressors. In *Analysis of Panels and Limited Dependent Variable Models*, eds. L.-F. L. C. Hsiao, K. Lahiri and M. H. Pesaran, 199–225. Cambridge: Cambridge University Press.

Kiviet, J. F. and M. J. G. Bun. 2001. The Accuracy of Inference in Small Samples of Dynamic Panel Data Models. *Tinbergen Institute Discussion Paper TI 2001-006/4* .

McLeod, A. I. and K. W. Hipel. 1978. Simulation Procedures for Box-Jenkins Models. *Water Resources Research* 14: 969–975.

Nickell, S. J. 1981. Biases in Dynamic Models with Fixed Effects. *Econometrica* 49: 1417–1426.

A Appendix: Demonstrating xtlsdvc

I demonstrate the use of `xtlsdvc` in the context of labour demand estimation, using the data set abdata.dta (Arellano and Bond (1991)), a typical micro panel of firm data with a moderately large N (140 firms). The labour demand of the firm is modelled according to specification (1), with the natural log of firm employment, `n`, as the dependent variable and the natural log of the real product wage, `w`, the natural log of the gross capital stock, `k`, and a set of time dummies as explanatory variables. The log of employment lagged one time is also included as a right-hand variable to allow costly employment adjustments.

Differently from the customary approach I do not use all information available to estimate the regression parameters. Instead, I follow a strategy that, exploiting the industry partition of the cross-sectional dimension as defined by the categorical variable `ind`, lets the slopes be industry-specific. This is easily accomplished by restricting the usable data to the panel of firms belonging to a given industry. While such a strategy leads to a less restrictive specification for the firm labour demand, it causes a reduced number of cross-sectional units for use in estimation, so that the researcher must be prepared to deal with a potentially severe small sample bias in any of the industry regressions. Clearly, `xtlsdvc` is the appropriate solution in this case.

The demonstration is kept as simple as possible considering regressions for only one industry panel (`ind=4`). It has been carried out on a pc endowed with a Pentium 4 2.80 GHz CPU and 496 MB of RAM.

The routine is reasonably fast when the bootstrap procedure is not invoked. Otherwise, the waiting time may be considerable, linearly increasing in the number of repetitions. To get an idea of this, a message at the end of each execution displays the amount of time consumed by the code.

```
. sysuse abdata,clear
```

```
.
. * Data description for industry 4
. xtdes if ind==4

      id:  16, 18, ..., 133                                  n =         29
    year:  1976, 1977, ..., 1984                             T =          9
           Delta(year) = 1; (1984-1976)+1 = 9
           (id*year uniquely identifies each observation)

Distribution of T_i:   min      5%     25%       50%       75%     95%     max
                         7       7       7         7         7       8       9

     Freq.  Percent    Cum. |  Pattern
 ---------------------------+-----------
       18     62.07   62.07 |  1111111..
        8     27.59   89.66 |  .1111111.
        1      3.45   93.10 |  ..1111111
        1      3.45   96.55 |  .11111111
        1      3.45  100.00 |  111111111
 ---------------------------+-----------
       29    100.00         |  XXXXXXXXX

.
. set rmsg on
r; t=0.00 17:29:43

.
. * LSDVC initialized by AH.
.
. * Level 1 of accuracy.
. * AH and (uncorrected) LSDV estimates are also displayed.
. xtlsdvc n w k yr1977-yr1984 if ind==4, init(ah) l f


note: yr1983 dropped due to collinearity
      (when the lagged dependent variable  L.n is included)

Bias correction initialized by Anderson and Hsiao estimator

Instrumental variables (2SLS) regression

      Source |       SS       df       MS              Number of obs =     148
-------------+------------------------------           F( 10,   138) =       .
       Model |  1.35967485    10  .135967485           Prob > F      =       .
    Residual |  .933924166   138  .006767566           R-squared     =       .
-------------+------------------------------           Adj R-squared =       .
       Total |  2.29359902   148  .015497291           Root MSE      =  .08227

------------------------------------------------------------------------------
_
D.n          |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
_
n            |
          LD |   .2204939   .4445225     0.50   0.621     -.658462     1.09945
w            |
          D1 |  -.3771841    .134876    -2.80   0.006     -.643875   -.1104933
k            |
          D1 |   .2204505   .0979079     2.25   0.026     .0268569    .4140442
yr1977       |
          D1 |   .1988839   .1633361     1.22   0.225    -.1240813     .521849
yr1978       |
          D1 |   .1719693   .1507414     1.14   0.256    -.1260923    .4700308
yr1979       |
          D1 |   .1489565   .1597643     0.93   0.353     -.166946     .464859
yr1980       |
          D1 |   .0922867    .157367     0.59   0.559    -.2188757    .4034491
```

yr1981						
D1	-.0171367	.1401022	-0.12	0.903	-.2941613	.2598879
yr1982						
D1	-.0650494	.0811056	-0.80	0.424	-.2254199	.095321
yr1984						
D1	.0512528	.0581115	0.88	0.379	-.0636513	.1661569

```
Instrumented:  LD.n
Instruments:   D.w D.k D.yr1977 D.yr1978 D.yr1979 D.yr1980 D.yr1981 D.yr1982
               D.yr1984 L2.n
```

LSDV dynamic regression

n	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
n						
L1	.4056509	.0731424	5.55	0.000	.2622945	.5490074
w	-.3541811	.1315442	-2.69	0.007	-.612003	-.0963593
k	.2541555	.0525718	4.83	0.000	.1511167	.3571944
yr1977	.1590321	.0412702	3.85	0.000	.0781441	.2399201
yr1978	.1480011	.0376338	3.93	0.000	.0742401	.2217621
yr1979	.1166947	.0388859	3.00	0.003	.0404799	.1929096
yr1980	.0615435	.0388128	1.59	0.113	-.0145281	.1376152
yr1981	-.0333848	.036903	-0.90	0.366	-.1057134	.0389438
yr1982	-.0528846	.03139	-1.68	0.092	-.1144079	.0086387
yr1984	.1019097	.0592481	1.72	0.085	-.0142145	.2180339

Bias correction up to order O(1/T)

LSDVC dynamic regression
(SE not computed)

n	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
n						
L1	.5389829	.	.	.	.	.
w	-.3375203	.	.	.	.	.
k	.2218794	.	.	.	.	.
yr1977	.1231041	.	.	.	.	.
yr1978	.1191318	.	.	.	.	.
yr1979	.0871871	.	.	.	.	.
yr1980	.0324267	.	.	.	.	.
yr1981	-.0580636	.	.	.	.	.
yr1982	-.0634494	.	.	.	.	.
yr1984	.0928311	.	.	.	.	.

```
r; t=0.74 17:29:44

.
. * Level 2 of accuracy.
. xtlsdvc n w k yr1977-yr1984 if ind==4, init(ah) bi(2)
```

```
note: yr1983 dropped due to collinearity
      (when the lagged dependent variable  L.n is included)
Bias correction initialized by Anderson and Hsiao estimator

Bias correction up to order O(1/NT)
LSDVC dynamic regression
(SE not computed)
------------------------------------------------------------------------------
           n |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
n            |
          L1 |   .5354691          .        .       .            .           .
w            |  -.3380943          .        .       .            .           .
k            |   .2226967          .        .       .            .           .
yr1977       |   .1238945          .        .       .            .           .
yr1978       |   .1197488          .        .       .            .           .
yr1979       |   .0878223          .        .       .            .           .
yr1980       |   .0330506          .        .       .            .           .
yr1981       |  -.0575616          .        .       .            .           .
yr1982       |  -.0633143          .        .       .            .           .
yr1984       |    .092829          .        .       .            .           .
------------------------------------------------------------------------------
r; t=0.75 17:29:45

.
. * Level 3 of accuracy
. xtlsdvc n w k yr1977-yr1984 if ind==4, init(ah) bi(3)

note: yr1983 dropped due to collinearity
      (when the lagged dependent variable  L.n is included)
Bias correction initialized by Anderson and Hsiao estimator

Bias correction up to order O(1/NT^2)
LSDVC dynamic regression
(SE not computed)
------------------------------------------------------------------------------
           n |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
n            |
          L1 |   .6338054          .        .       .            .           .
w            |  -.3258186          .        .       .            .           .
k            |   .1988694          .        .       .            .           .
yr1977       |   .0973986          .        .       .            .           .
yr1978       |   .0984595          .        .       .            .           .
yr1979       |   .0660618          .        .       .            .           .
yr1980       |   .0115782          .        .       .            .           .
yr1981       |  -.0757634          .        .       .            .           .
yr1982       |  -.0711084          .        .       .            .           .
yr1984       |   .0861093          .        .       .            .           .
------------------------------------------------------------------------------
r; t=0.75 17:29:45
```

```
.
.
. * LSDVC (level 3 of accuracy) initialized by AH, plus bootstrap SE
.
. * 100 replications
. xtlsdvc n w k yr1977-yr1984 if ind==4, init(ah) bi(3) vcov(100)

note: yr1983 dropped due to collinearity
      (when the lagged dependent variable  L.n is included)
Bias correction initialized by Anderson and Hsiao estimator

Bias correction up to order O(1/NT^2)
LSDVC dynamic regression
(bootstrapped SE)
------------------------------------------------------------------------------
-
n            |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
-
n            |
          L1 |   .6338054   .2384333     2.66   0.008     .1664848    1.101126
w            |  -.3258186   .1624866    -2.01   0.045    -.6442865   -.0073507
k            |   .1988694   .0652599     3.05   0.002     .0709623    .3267765
yr1977       |   .0973986   .0709764     1.37   0.170    -.0417126    .2365098
yr1978       |   .0984595   .0686167     1.43   0.151    -.0360267    .2329456
yr1979       |   .0660618   .0732827     0.90   0.367    -.0775695    .2096932
yr1980       |   .0115782    .075663     0.15   0.878    -.1367186     .159875
yr1981       |  -.0757634   .0618594    -1.22   0.221    -.1970056    .0454788
yr1982       |  -.0711084   .0356703    -1.99   0.046     -.141021   -.0011958
yr1984       |   .0861093   .0703664     1.22   0.221    -.0518064     .224025
------------------------------------------------------------------------------
-
r; t=73.50 17:30:59
. * 200 replications
. xtlsdvc n w k yr1977-yr1984 if ind==4, init(ah) bi(3) vcov(200)

note: yr1983 dropped due to collinearity
      (when the lagged dependent variable  L.n is included)
Bias correction initialized by Anderson and Hsiao estimator

Bias correction up to order O(1/NT^2)
LSDVC dynamic regression
(bootstrapped SE)
------------------------------------------------------------------------------
-
n            |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
-
n            |
          L1 |   .6338054   .2366395     2.68   0.007     .1700005     1.09761
w            |  -.3258186   .1740695    -1.87   0.061    -.6669885    .0153514
k            |   .1988694    .082856     2.40   0.016     .0364747    .3612641
yr1977       |   .0973986     .07895     1.23   0.217    -.0573406    .2521377
yr1978       |   .0984595   .0755014     1.30   0.192    -.0495206    .2464395
yr1979       |   .0660618    .080673     0.82   0.413    -.0920542    .2241779
yr1980       |   .0115782   .0812952     0.14   0.887    -.1477575    .1709138
yr1981       |  -.0757634     .06819    -1.11   0.267    -.2094133    .0578865
yr1982       |  -.0711084    .039987    -1.78   0.075    -.1494814    .0072647
```

```
yr1984       |   .0861093   .0714245     1.21   0.228      -.05388    .2260987
------------------------------------------------------------------------------
-
r; t=145.13 17:33:24
.
.
. * LSDVC (level 3 of accuracy) initialized by AB,
. * plus bootstrap SE (100 replications).
. * AB estimates are also displayed.
.
. xtlsdvc n w k yr1977-yr1984 if ind==4, init(ab) f bi(3) vcov(100)


note: yr1983 dropped due to collinearity
      (when the lagged dependent variable  L.n is included)
Bias correction initialized by Arellano and Bond estimator
note: yr1977 dropped due to collinearity
Arellano-Bond dynamic panel-data estimation     Number of obs      =       148
Group variable (i): id                          Number of groups   =        29
                                                Wald chi2(.)       =         .
Time variable (t): year                         Obs per group: min =         5
                                                               avg =  5.103448
                                                               max =         7
One-step results
------------------------------------------------------------------------------
-
D.n          |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
-
n            |
          LD |   .5713301   .0697836     8.19   0.000     .4345568    .7081034
w            |
          D1 |  -.5627737   .1291909    -4.36   0.000    -.8159832   -.3095643
k            |
          D1 |    .149354   .0597419     2.50   0.012     .0322619     .266446
yr1978       |
          D1 |   .0200412    .019743     1.02   0.310    -.0186544    .0587369
yr1979       |
          D1 |   -.008996   .0206913    -0.43   0.664    -.0495503    .0315582
yr1980       |
          D1 |  -.0598638   .0208495    -2.87   0.004    -.1007281   -.0189995
yr1981       |
          D1 |  -.1533807   .0202301    -7.58   0.000    -.1930309   -.1137305
yr1982       |
          D1 |  -.1424881   .0191927    -7.42   0.000    -.1801052   -.1048711
yr1984       |
          D1 |   .0673006   .0514889     1.31   0.191    -.0336159    .1682171
------------------------------------------------------------------------------
-
Sargan test of over-identifying restrictions:
         chi2(27) =    77.04     Prob > chi2 = 0.0000
Arellano-Bond test that average autocovariance in residuals of order 1 is 0:
         H0: no autocorrelation   z =  -2.12   Pr > z = 0.0337
Arellano-Bond test that average autocovariance in residuals of order 2 is 0:
         H0: no autocorrelation   z =  -1.06   Pr > z = 0.2878


Bias correction up to order O(1/NT^2)
LSDVC dynamic regression
```

```
(bootstrapped SE)
------------------------------------------------------------------------------
-
           n |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
-
n            |
          L1 |   .7206262   .1205431     5.98   0.000     .4843659    .9568864
w            |  -.3331545   .1600705    -2.08   0.037     -.646887   -.0194221
k            |   .1844672   .0624622     2.95   0.003     .0620436    .3068907
yr1977       |   .0762851   .0377489     2.02   0.043     .0022987    .1502716
yr1978       |   .0865282   .0361974     2.39   0.017     .0155826    .1574739
yr1979       |   .0516378   .0386154     1.34   0.181    -.0240469    .1273225
yr1980       |  -.0033252   .0384388    -0.09   0.931    -.0786639    .0720134
yr1981       |  -.0895199   .0367064    -2.44   0.015    -.1614631   -.0175767
yr1982       |  -.0766998   .0326953    -2.35   0.019    -.1407815   -.0126182
yr1984       |   .0679823   .0687112     0.99   0.322    -.0666891    .2026537
------------------------------------------------------------------------------
-
r; t=79.67 17:34:44
.
```

B Appendix: Figures

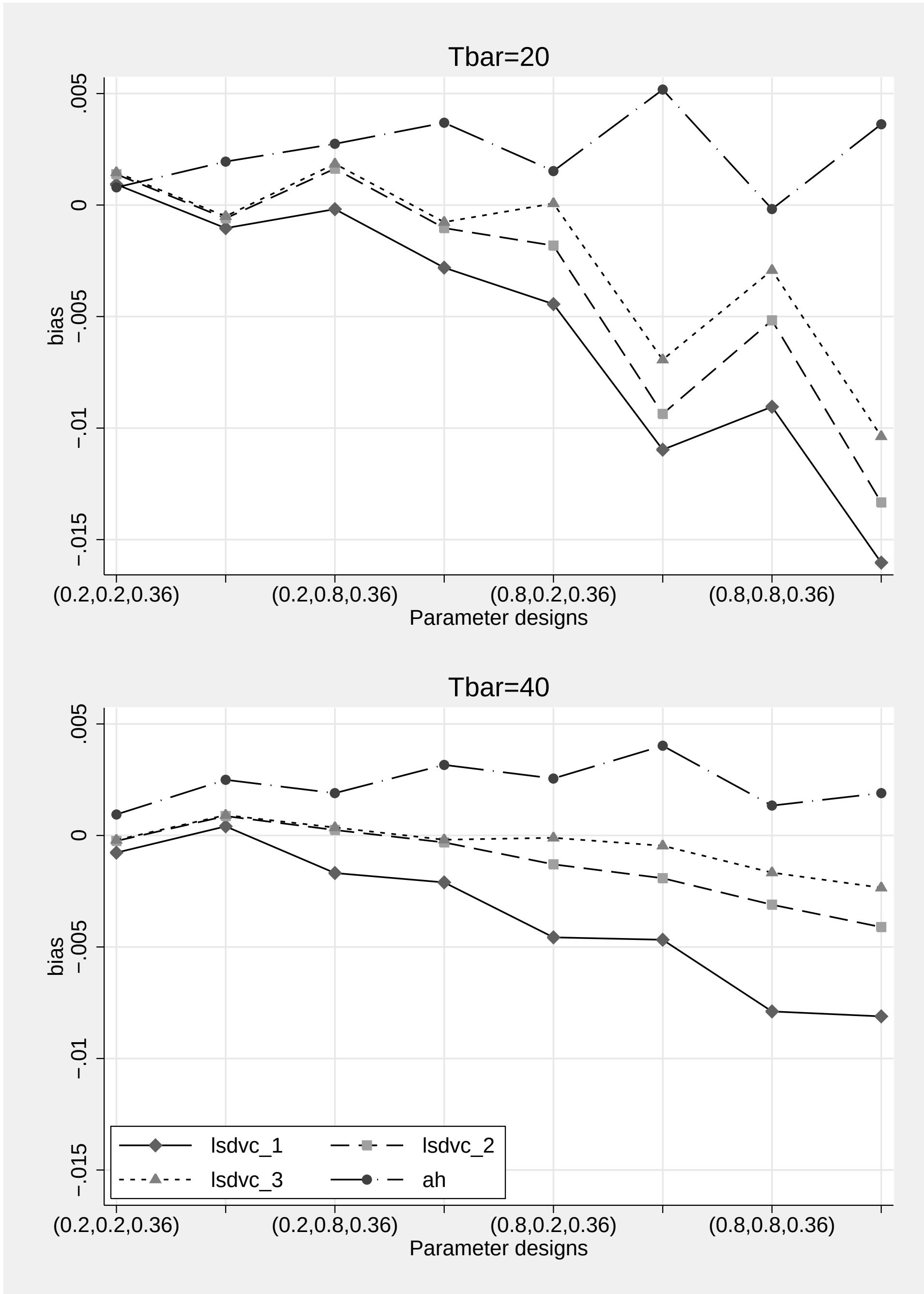

Figure 1: Biases of LSDVC_1, LSDVC_2, LSDVC_3 and AH for γ.

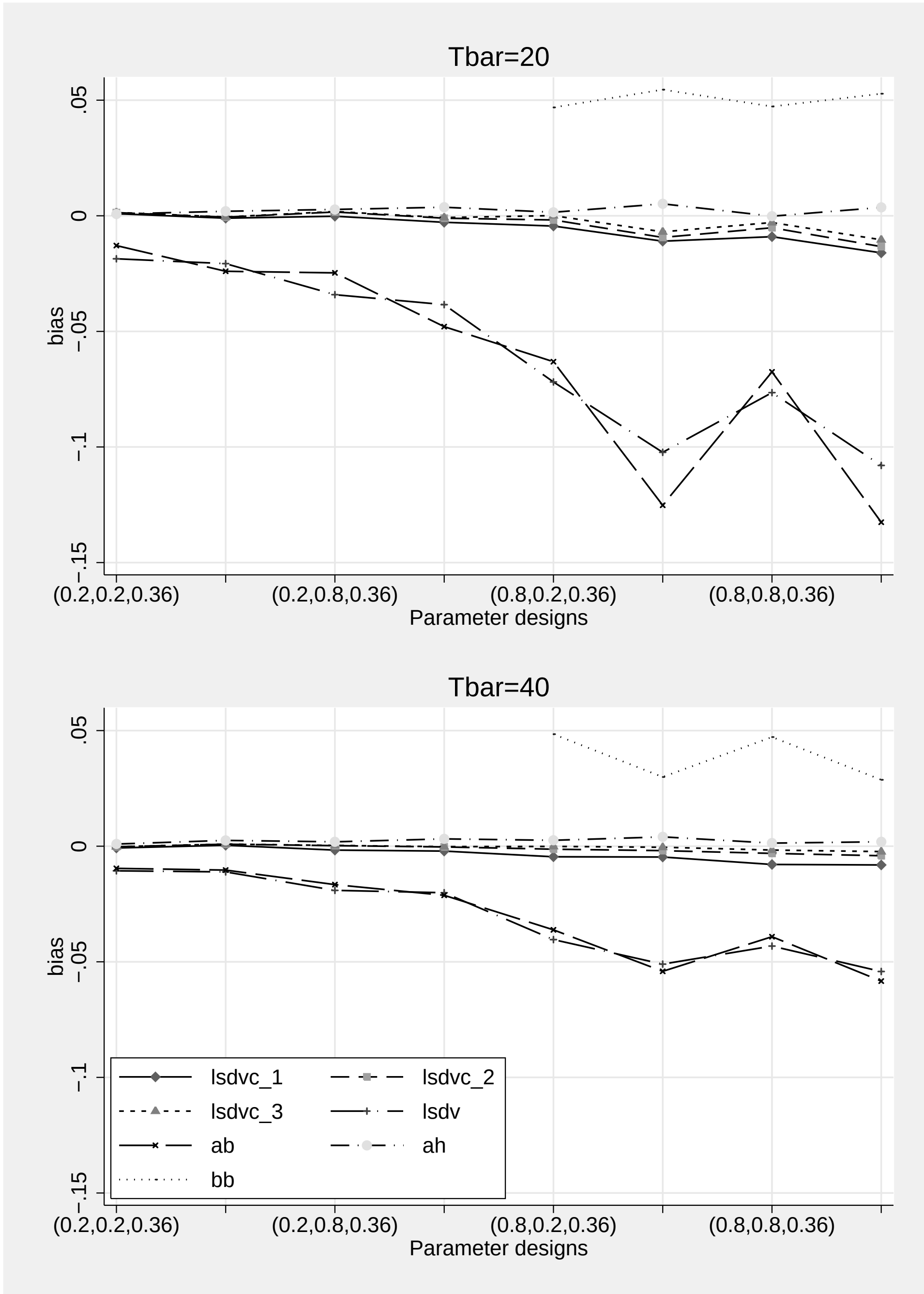

Figure 2: Biases of all estimators for γ.

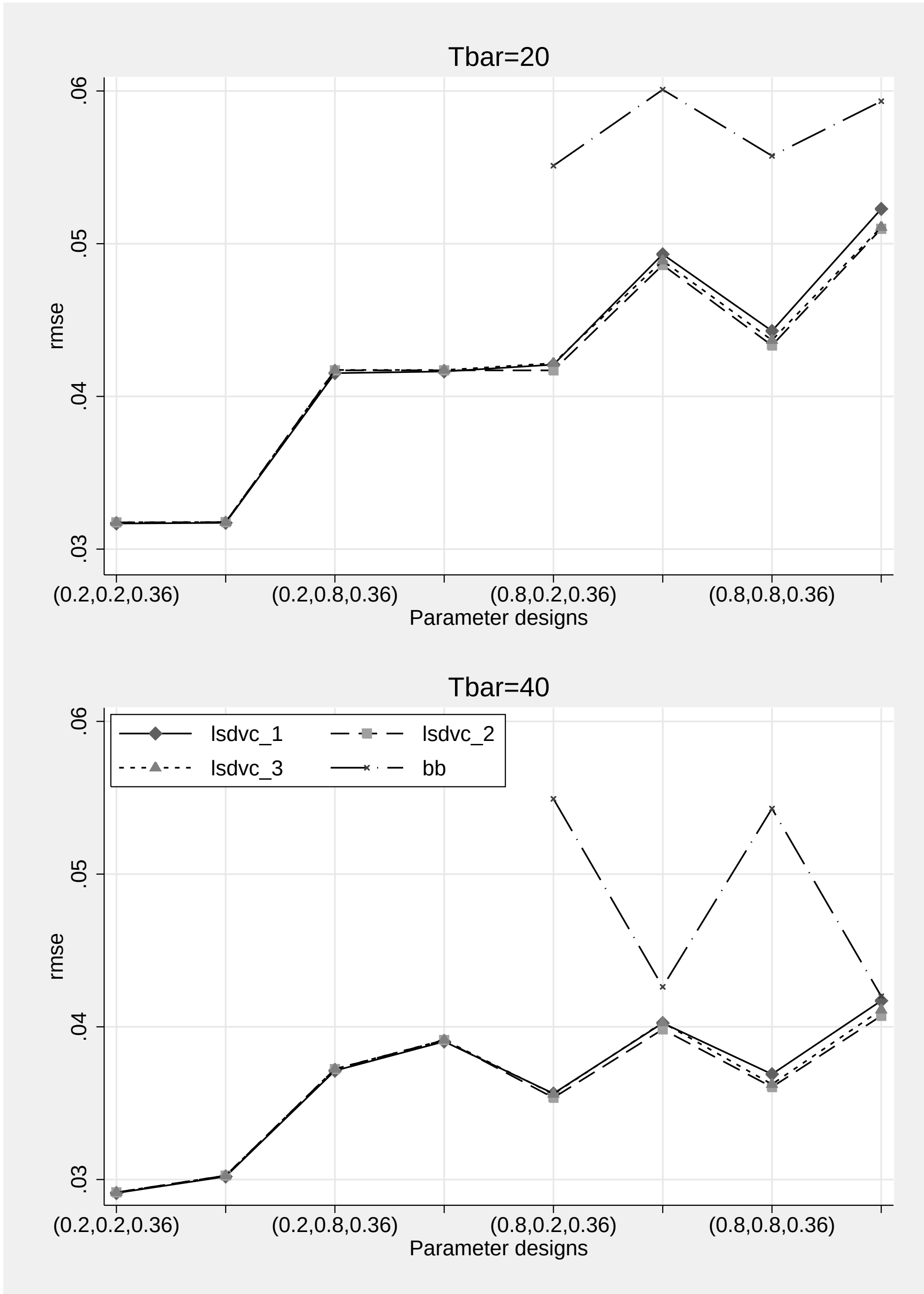

Figure 3: RMSE's of $LSDVC_1$, $LSDVC_2$, $LSDVC_3$ and BB for γ.

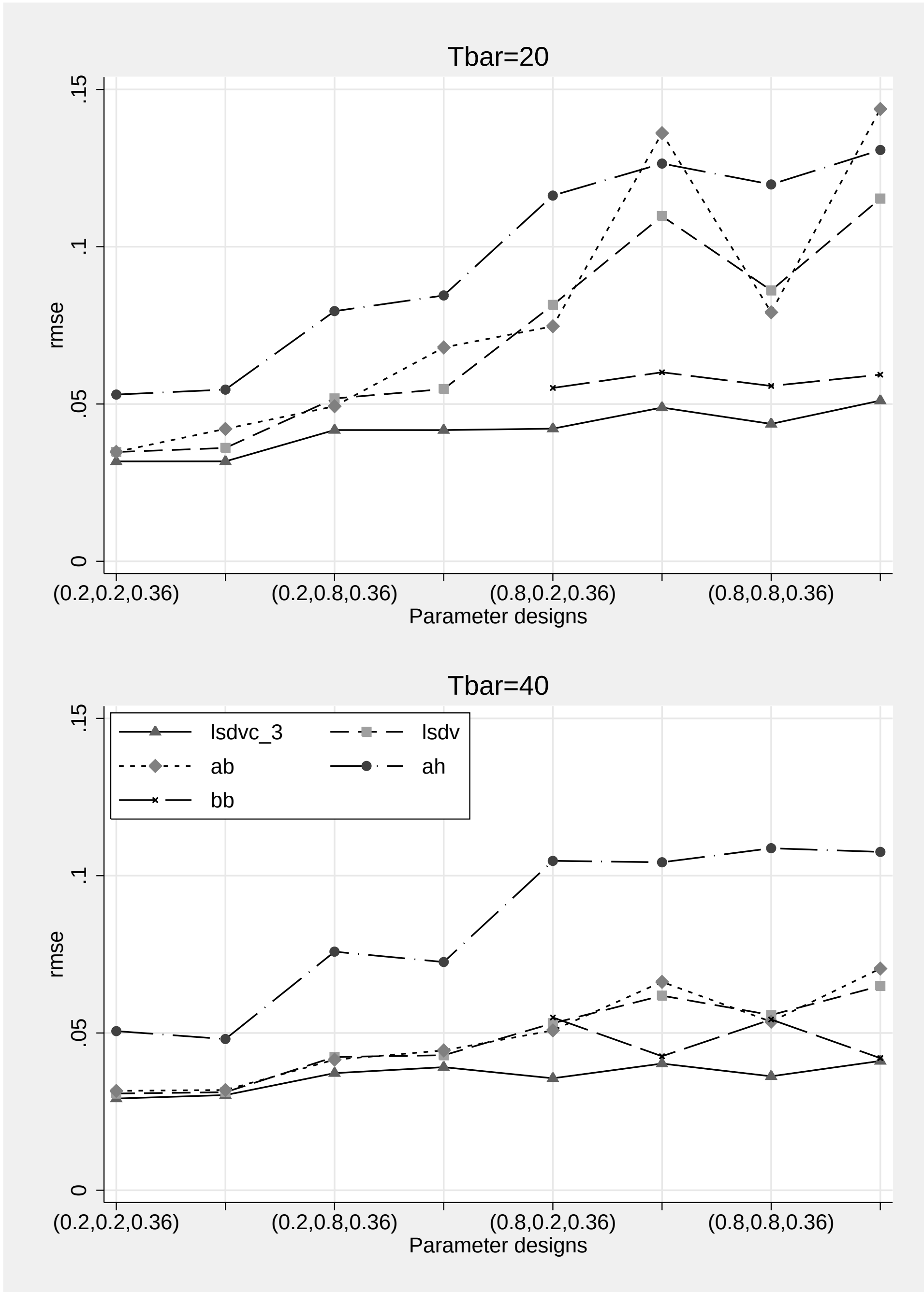

Figure 4: RMSE's of all estimators for γ.

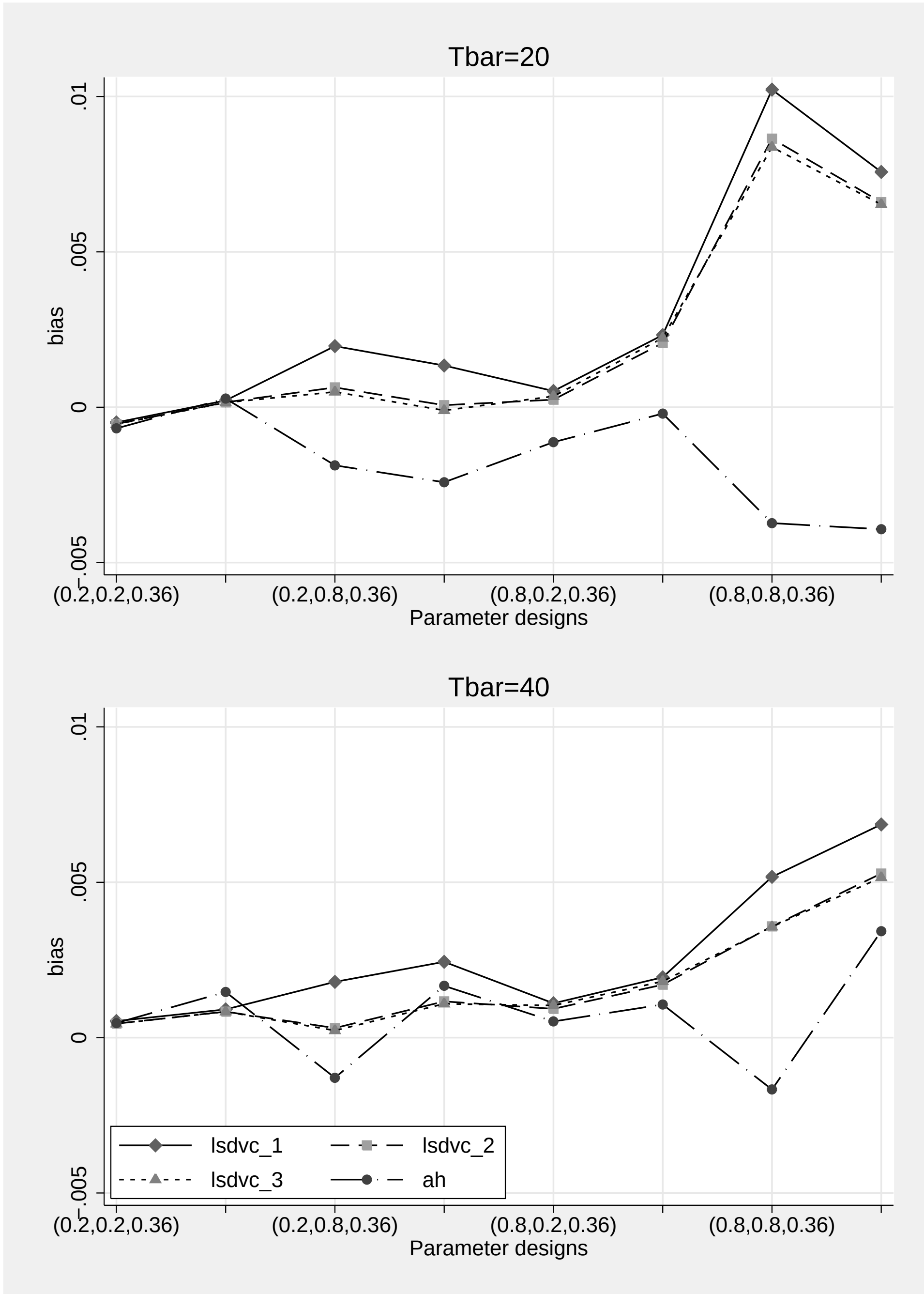

Figure 5: Biases of LSDVC_1, LSDVC_2, LSDVC_3 and AH for β.

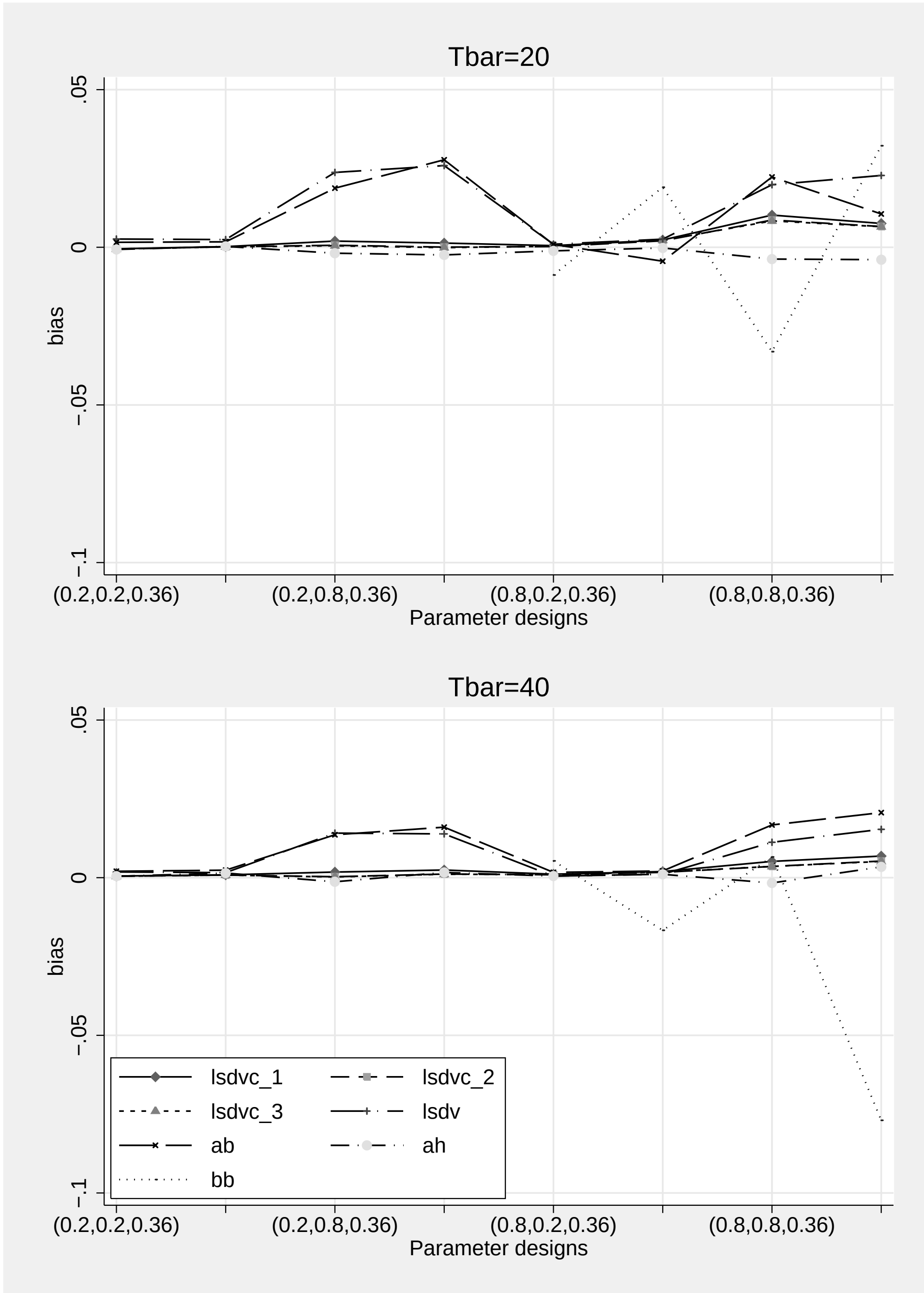

Figure 6: Biases of all estimators for β.

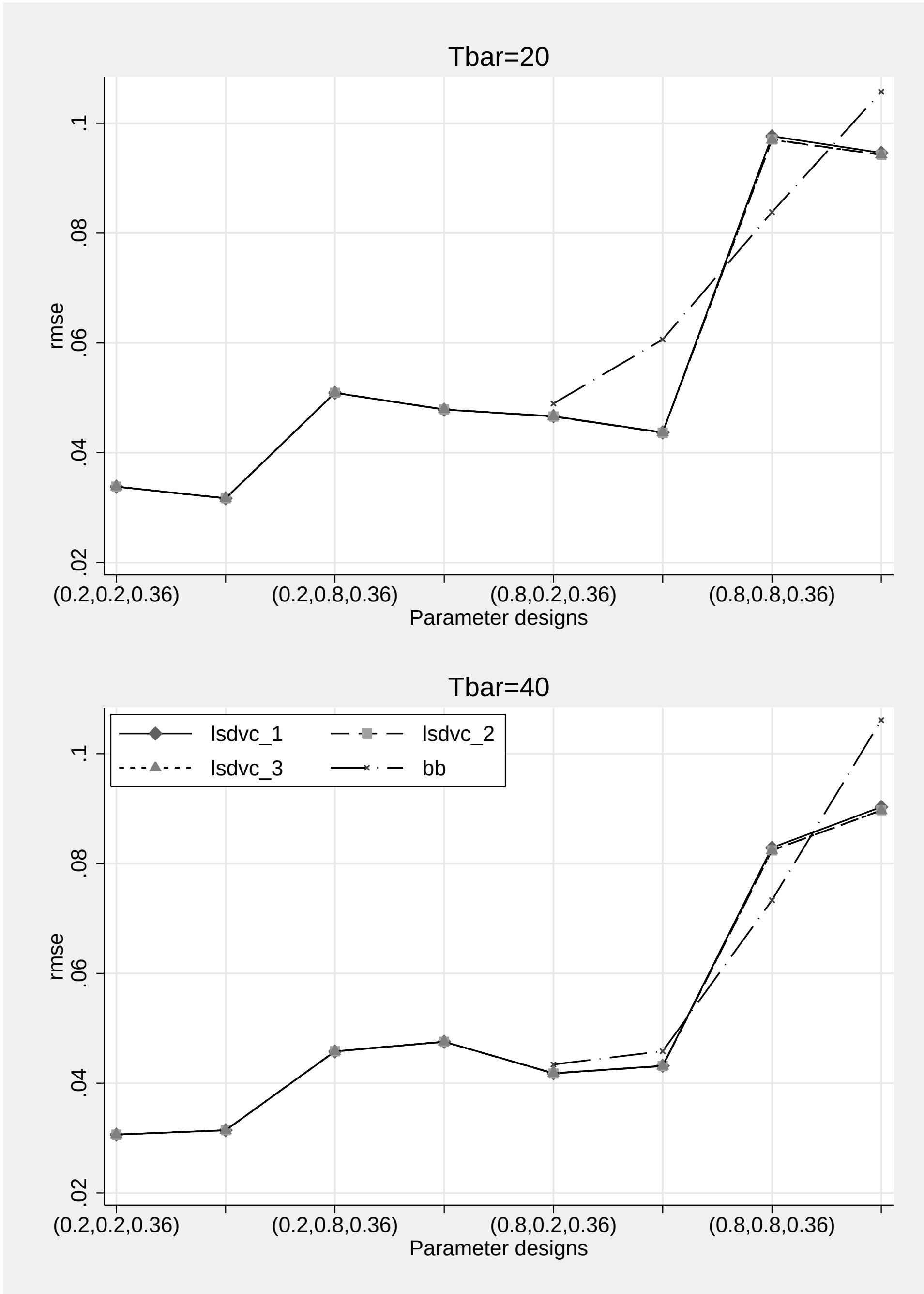

Figure 7: RMSE's of LSDVC_1, LSDVC_2, LSDVC_3 and BB for β.

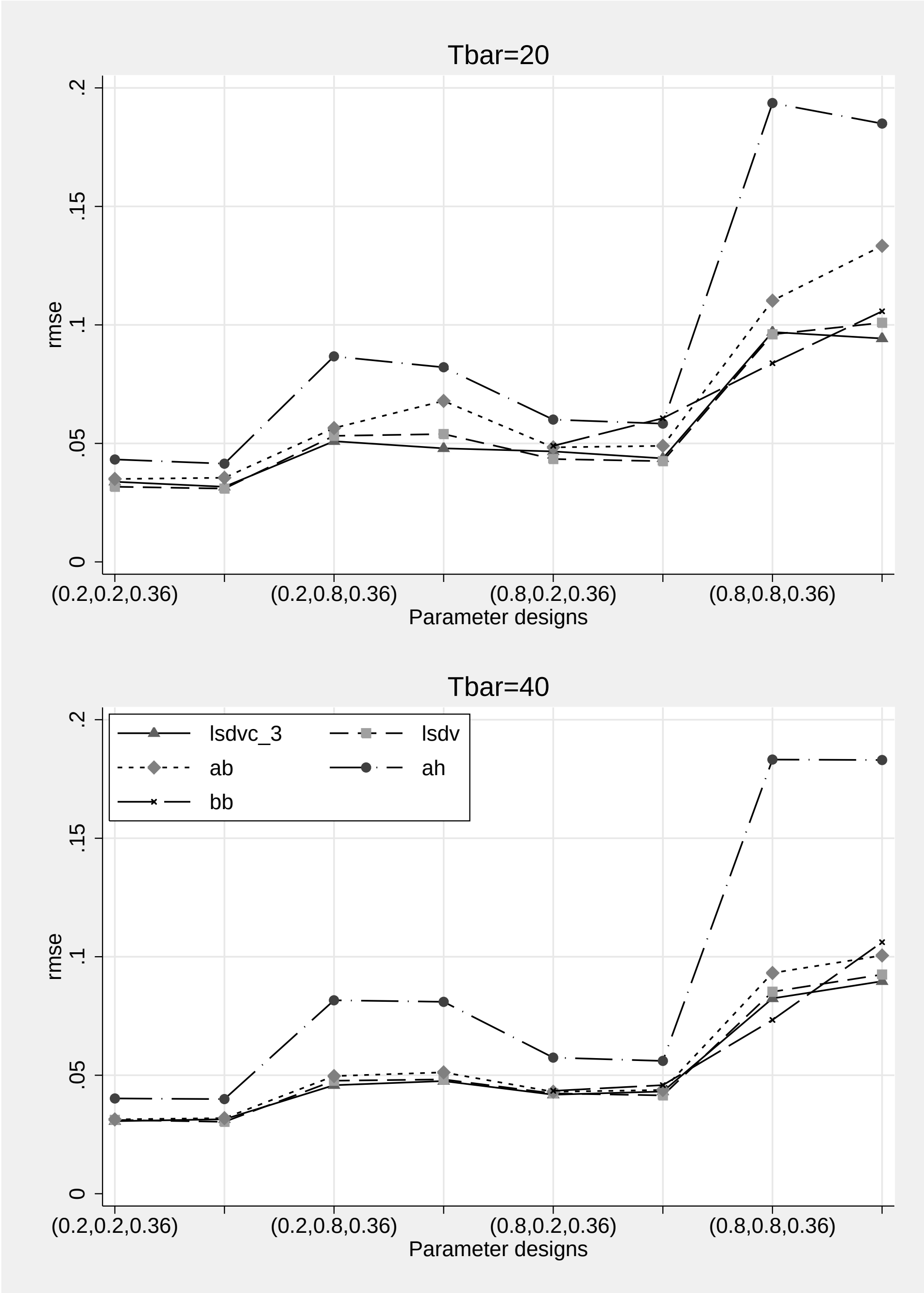

Figure 8: RMSE's of all estimators for β.